

## CARLEMAN ESTIMATES FOR A CLASS OF DEGENERATE PARABOLIC OPERATORS\*

P. CANNARSA<sup>†</sup>, P. MARTINEZ<sup>‡</sup>, AND J. VANCOSTENOBLE<sup>‡</sup>

**Abstract.** Given  $\alpha \in [0, 2)$  and  $f \in L^2((0, T) \times (0, 1))$ , we derive new Carleman estimates for the degenerate parabolic problem  $w_t + (x^\alpha w_x)_x = f$ , where  $(t, x) \in (0, T) \times (0, 1)$ , associated to the boundary conditions  $w(t, 1) = 0$  and  $w(t, 0) = 0$  if  $0 \leq \alpha < 1$  or  $(x^\alpha w_x)(t, 0) = 0$  if  $1 \leq \alpha < 2$ . The proof is based on the choice of suitable weighted functions and Hardy-type inequalities. As a consequence, for all  $0 \leq \alpha < 2$  and  $\omega \subset\subset (0, 1)$ , we deduce null controllability results for the degenerate one-dimensional heat equation  $u_t - (x^\alpha u_x)_x = h\chi_\omega$  with the same boundary conditions as above.

**Key words.** degenerate parabolic equation, null controllability, Carleman estimates, Hardy-type inequality

**AMS subject classifications.** 35K65, 93B05, 93B07

**DOI.** 10.1137/04062062X

**1. Introduction.** The study of controllability for nondegenerate parabolic equations has attracted the interest of several authors in the past few decades. After the pioneering works [14, 19, 20, 37, 38], there has been substantial progress in understanding the controllability properties of nondegenerate parabolic equations with variable coefficients. In [30], local Carleman estimates for elliptic equations were used to study the null controllability of the heat equation on a manifold. Finally, a powerful new approach, based on global estimates of Carleman type, was developed in [26].

The theory has also been extended to semilinear problems (see, for example, [2, 3, 12, 15, 21, 24, 25]) and to equations in unbounded domains (see, for example, [13, 33, 34]; see also [31, 41]). For the Stokes and Navier–Stokes equations we also refer the reader to [4, 10, 11, 18, 22, 23, 26, 27, 28].

On the contrary, few results are known for degenerate equations, even though many problems that are relevant for applications are described by parabolic equations degenerating at the boundary of the space domain. For instance, in [5, 32, 8, 9] the reader will find a motivating example of a Crocco-type equation coming from the study of the velocity field of a laminar flow on a flat plate.

The goal of this paper is to study the controllability of a simple model of degenerate parabolic equation, namely,

$$u_t - (x^\alpha u_x)_x = h\chi_\omega, \quad x \in (0, 1), t \in (0, T),$$

where the control  $h$  acts on a nonempty subinterval  $\omega$  of  $(0, 1)$ .

---

\*Received by the editors December 10, 2004; accepted for publication (in revised form) May 18, 2007; published electronically January 4, 2008. This work was partially done when the first author was visiting the Université Paul Sabatier Toulouse III, and when the second and third authors were visiting the Università di Roma “Tor Vergata,” invited by the university, and the Istituto Nazionale di Alta Matematica.

<http://www.siam.org/journals/sicon/47-1/62062.html>

<sup>†</sup>Dipartimento di Matematica, Università di Roma “Tor Vergata,” Via della Ricerca Scientifica, 00133 Roma, Italy (cannarsa@mat.uniroma2.it).

<sup>‡</sup>Laboratoire M.I.P., U.M.R. C.N.R.S. 5640, Université Paul Sabatier Toulouse III, 118 route de Narbonne, 31 062 Toulouse Cedex 4, France (martinez@mip.ups-tlse.fr, vancoste@mip.ups-tlse.fr).

## 2. Results.

**2.1. Statement of the controllability problem.** Given  $0 \leq \alpha < 2$ , define

$$\forall x \in [0, 1], \quad a(x) := x^\alpha,$$

and let  $\omega$  be a nonempty subinterval of  $(0, 1)$ . For  $T > 0$ , set

$$Q_T = (0, T) \times (0, 1),$$

and consider the initial-boundary value problem

$$(2.1) \quad \begin{cases} u_t - (au_x)_x = h \chi_\omega, & (t, x) \in Q_T, \\ u(t, 1) = 0, & t \in (0, T), \\ \text{and } \begin{cases} u(t, 0) = 0 & \text{for } 0 \leq \alpha < 1, \\ (au_x)(t, 0) = 0 & \text{for } 1 \leq \alpha < 2, \end{cases} & t \in (0, T), \\ u(0, x) = u_0(x), & x \in (0, 1), \end{cases}$$

where  $u_0$  is given in  $L^2(0, 1)$  and  $h \in L^2(Q_T)$ .

**2.2. Well-posedness.** Let us recall that the above problem is well-posed in appropriate weighted spaces. For  $0 \leq \alpha < 1$ , define the Hilbert space  $H_a^1(0, 1)$  as

$$H_a^1(0, 1) := \{u \in L^2(0, 1) \mid u \text{ absolutely continuous in } [0, 1], \\ \sqrt{a}u_x \in L^2(0, 1) \text{ and } u(0) = u(1) = 0\},$$

and the unbounded operator  $A : D(A) \subset L^2(0, 1) \rightarrow L^2(0, 1)$  by

$$\begin{cases} \forall u \in D(A), & Au := (au_x)_x, \\ D(A) := \{u \in H_a^1(0, 1) \mid au_x \in H^1(0, 1)\}. \end{cases}$$

Notice that if  $u \in D(A)$  (or even  $u \in H_a^1(0, 1)$ ), then  $u$  satisfies the Dirichlet boundary conditions  $u(0) = u(1) = 0$ .

For  $1 \leq \alpha < 2$ , let us change the definition of  $H_a^1(0, 1)$  to

$$H_a^1(0, 1) := \{u \in L^2(0, 1) \mid u \text{ locally absolutely continuous in } (0, 1], \\ \sqrt{a}u_x \in L^2(0, 1) \text{ and } u(1) = 0\}.$$

Then the operator  $A : D(A) \subset L^2(0, 1) \rightarrow L^2(0, 1)$  will be defined by

$$\begin{cases} \forall u \in D(A), & Au := (au_x)_x, \\ D(A) := \{u \in H_a^1(0, 1) \mid au_x \in H^1(0, 1)\} \\ \quad = \{u \in L^2(0, 1) \mid u \text{ locally absolutely continuous in } (0, 1], \\ \quad \quad au \in H_0^1(0, 1), au_x \in H^1(0, 1), \text{ and } (au_x)(0) = 0\}. \end{cases}$$

Notice that if  $u \in D(A)$ , then  $u$  satisfies the Neumann boundary condition  $(au_x)(0) = 0$  at  $x = 0$  and the Dirichlet boundary condition  $u(1) = 0$  at  $x = 1$ .

In both cases, the following results hold (see, e.g., [7] and [9]).

**PROPOSITION 2.1.**  *$A : D(A) \subset L^2(0, 1) \rightarrow L^2(0, 1)$  is a closed self-adjoint negative operator with dense domain.*



Hence,  $A$  is the infinitesimal generator of a strongly continuous semigroup  $e^{tA}$  on  $L^2(0, 1)$ . Consequently, we have the following well-posedness result.

**THEOREM 2.1.** *Let  $h$  be given in  $L^2(Q_T)$ . For all  $u_0 \in L^2(0, 1)$ , problem (2.1) has a unique solution*

$$(2.2) \quad u \in C^0([0, T]; L^2(0, 1)) \cap L^2(0, T; H_a^1(0, 1)).$$

Moreover, if  $u_0 \in D(A)$ , then

$$(2.3) \quad u \in C^0([0, T]; H_a^1(0, 1)) \cap L^2(0, T; D(A)) \cap H^1(0, T; L^2(0, 1)).$$

*Remark 2.1.* Most of the results of this paper hold and will be stated for solutions in the above class (2.2). However, in the proofs, we will assume—often without further notice—that solutions belong to the stronger class (2.3). This can be done without loss of generality, since the general result can always be recovered by a standard density argument.

**2.3. Carleman estimates for degenerate problems.** In order to study the controllability properties of (2.1), we need to derive a Carleman estimate for the adjoint problem. Keeping the notation

$$a(x) := x^\alpha, \quad \text{with } 0 \leq \alpha < 2, \quad \text{and } Q_T = (0, T) \times (0, 1) \quad \text{for } T > 0,$$

let us consider the parabolic problem

$$(2.4) \quad \begin{cases} w_t + (aw_x)_x = f, & (t, x) \in Q_T, \\ w(t, 1) = 0, & t \in (0, T), \\ \text{and } \begin{cases} w(t, 0) = 0 & \text{for } 0 \leq \alpha < 1, \\ (aw_x)(t, 0) = 0 & \text{for } 1 \leq \alpha < 2, \end{cases} & t \in (0, T), \\ w(T, x) = w_T(x), & x \in (0, 1), \end{cases}$$

where  $w_T \in L^2(0, 1)$  and  $f \in L^2(Q_T)$ . Our main result is the following.

**THEOREM 2.2.** *Let  $0 \leq \alpha < 2$  and  $T > 0$  be given. Then there exists  $\sigma : (0, T) \times [0, 1] \rightarrow \mathbb{R}_+^*$  of the form  $\sigma(t, x) = \theta(t)p(x)$ , with*

$$p(x) > 0 \quad \forall x \in [0, 1] \quad \text{and} \quad \theta(t) \rightarrow \infty \text{ as } t \rightarrow 0^+, T^-,$$

*and two positive constants,  $C$  and  $R_0$ , such that, for all  $w_T \in L^2(0, 1)$  and  $f \in L^2(Q_T)$ , the solution  $w$  of (2.4) satisfies, for all  $R \geq R_0$ ,*

$$\begin{aligned} \iint_{Q_T} \left( R\theta x^\alpha w_x^2 + R^3\theta^3 x^{2-\alpha} w^2 \right) e^{-2R\sigma} dx dt \\ \leq C \iint_{Q_T} e^{-2R\sigma} f^2 dx dt + C \int_0^T \left\{ R\theta e^{-2R\sigma} w_x^2 \right\}_{|x=1}. \end{aligned}$$

*Remark 2.2.* The functions  $p$  and  $\theta$  will be explicitly constructed in the proof. As we shall see, the choice of  $\theta$  will be

$$\forall t \in (0, T), \quad \theta(t) = \left( \frac{1}{t(T-t)} \right)^4.$$

This weight function satisfies the following essential properties:

$$\theta(t) \rightarrow +\infty \text{ as } t \rightarrow 0^+ \text{ or } T^- \quad \text{and} \quad |\theta_t| \leq c\theta^{5/4}, \quad |\theta_{tt}| \leq c\theta^{3/2}$$

for some constant  $c > 0$  depending on  $T$ . Moreover, we will take

$$p(x) := \frac{2 - x^{2-\alpha}}{(2 - \alpha)^2} \quad \forall x \in [0, 1].$$

**2.4. Observability inequalities.** As it is well known, very useful tools for studying controllability are provided by observability inequalities for the adjoint problem

$$(2.5) \quad \begin{cases} v_t + (av_x)_x = 0, & (t, x) \in Q_T, \\ v(t, 1) = 0, & t \in (0, T), \\ \text{and } \begin{cases} v(t, 0) = 0 & \text{for } 0 \leq \alpha < 1, \\ (av_x)(t, 0) = 0 & \text{for } 1 \leq \alpha < 2, \end{cases} & t \in (0, T), \\ v(T, x) = v_T(x), & x \in (0, 1), \end{cases}$$

where  $v_T$  is given in  $L^2(0, 1)$ . From the Carleman estimate of Theorem 2.2, we obtain the following observability inequalities for (2.5).

**THEOREM 2.3.** *Let  $0 \leq \alpha < 2$  and  $T > 0$  be given, and let  $\omega$  be a nonempty subinterval of  $(0, 1)$ . Then there exists  $C > 0$  such that, for all  $v_T \in L^2(0, 1)$ , the solution  $v$  of (2.5) satisfies*

$$(2.6) \quad \int_0^1 x^\alpha v_x(0, x)^2 dx \leq C \int_0^T \int_\omega v(t, x)^2 dx dt.$$

**2.5. Application to controllability.** For any  $0 \leq \alpha < 2$ , the following observability inequality follows from Theorem 2.3 and Hardy's inequalities (see the proof in section 5):

$$(2.7) \quad \int_0^1 v(0, x)^2 dx \leq C \int_0^T \int_\omega v(t, x)^2 dx dt.$$

The above inequality is well known in the nondegenerate case ( $\alpha = 0$ ) since it follows, for instance, from classical Carleman estimates for nondegenerate parabolic equations.

For  $\alpha \in [0, 1/2) \cup [5/4, 2)$ , inequality (2.7) was proved in [9] by means of a different Carleman estimate that had been obtained using a different weight function  $p$  but gave no information for  $\alpha \in [1/2, 5/4)$ .

Therefore, inequality (2.7) above fills the gap between  $1/2$  and  $5/4$  which was left open in [9]. Thus, we obtain, by standard arguments (see, e.g., [14, 26]), a null controllability result for degenerate heat equations with initial data in  $L^2(0, 1)$ .

**THEOREM 2.4.** *Let  $0 \leq \alpha < 2$  and  $T > 0$  be given, and let  $\omega$  be a nonempty subinterval of  $(0, 1)$ . Then, for all  $u_0 \in L^2(0, 1)$ , there exists  $h \in L^2((0, T) \times \omega)$  such that the solution of the degenerate problem (2.1) satisfies  $u(T) \equiv 0$  in  $(0, 1)$ .*

**Remark 2.3.** Let us recall that the above result is optimal since, for  $\alpha \geq 2$ , problem (2.1) fails to be null controllable (see [8]). Indeed, a standard change of variable transforms problem (2.1) into the heat equation in the unbounded domain  $]0, +\infty[$ , whereas control supports are still bounded. Then a result by Escauriaza, Seregin, and Šverák [16, 17], which generalizes a result by Micu and Zuazua [33], ensures that null controllability fails for such an equation.

**Remark 2.4.** In [9], inequality (2.7) was applied to a Crocco-type equation to obtain a null controllability result for  $\alpha \in [0, 1/2) \cup [5/4, 2)$ . Thus, the results of the present paper also show the null controllability of this equation for all values of  $\alpha \in [0, 2)$ .

**2.6. Hardy-type inequalities.** A major ingredient for the proofs of Theorems 2.2 and 2.3 is the following well-known lemma (see, for example, [35]; for the reader's convenience, we recall the proof in section 6).

LEMMA 2.1 (Hardy-type inequalities).

- (i) *Let  $0 \leq \alpha^* < 1$ . Then, for all locally absolutely continuous functions  $z$  on  $(0, 1)$  satisfying*

$$z(x) \rightarrow 0 \quad \text{as } x \rightarrow 0^+ \quad \text{and} \quad \int_0^1 x^{\alpha^*} z_x^2 < \infty,$$

*the following inequality holds:*

$$(2.8) \quad \int_0^1 x^{\alpha^*-2} z^2 \leq \frac{4}{(1-\alpha^*)^2} \int_0^1 x^{\alpha^*} z_x^2.$$

- (ii) *Let  $1 < \alpha^* < 2$ . Then the above inequality (2.8) still holds for all locally absolutely continuous functions  $z$  on  $(0, 1)$  satisfying*

$$z(x) \rightarrow 0 \quad \text{as } x \rightarrow 1^- \quad \text{and} \quad \int_0^1 x^{\alpha^*} z_x^2 < +\infty.$$

*Remark 2.5.* Notice that (2.8) is false for  $\alpha^* = 1$ .

**2.7. Further remarks.** In the present paper, we study the case of a degenerate operator of the form  $-(x^\alpha u_x)_x$  with the boundary condition  $u(x=0) = 0$  when  $0 \leq \alpha < 1$  or  $(x^\alpha u_x)(x=0) = 0$  when  $1 \leq \alpha < 2$ . The choice of such an operator in divergence form probably simplifies parts of the computations arising in the proof of Carleman estimates. Of course, it would be interesting to study, in a next step, other operators like  $-x^\alpha u_{xx}$ . On the other hand, the choice of the boundary condition at  $x=0$  ensures a relatively simple framework for the statement of well-posedness. Here again, it would be interesting to study the cases of other boundary conditions. For example, an interesting problem would be the case of Wentzell boundary conditions; see, e.g., [6, 39]. The techniques developed here may be useful to treat such problems. However, both the form of the operator and the boundary conditions play an important role in the computations of the proof of Carleman estimates. For this reason, these other problems have yet to be studied.

On the other hand, let us mention that the ideas of the present paper allow us to prove similar null controllability results for degenerate semilinear problems using a classical fixed point method (see [1]).

Next, instead of a distributed control on  $\omega \subset (0, 1)$ , one could consider a boundary control acting at one extreme point of the domain  $(0, 1)$ . Theorem 2.2 readily implies a boundary null controllability result if the control acts at  $x=1$ . The case of a boundary control at  $x=0$  has not yet been studied.

Finally, another interesting question would be the study of degenerate operators in higher dimensions. Of course, this opens a lot of perspectives since the study will depend on the domain where the operator degenerates and the way it degenerates. This question will be the subject of a forthcoming paper.

### 3. Proof of Theorem 2.2 (Carleman estimates).

**3.1. Notation and reformulation of the problem.** We recall that  $a(x) = x^\alpha$  for all  $x \in [0, 1]$  with  $\alpha \in [0, 2)$  given. Let  $\sigma(t, x) = \theta(t)p(x)$ , where

$$p(x) > 0 \quad \forall x \in [0, 1] \quad \text{and} \quad \theta(t) \rightarrow \infty \text{ as } t \rightarrow 0^+, T^-.$$

For  $R > 0$ , define

$$(3.1) \quad z(t, x) = e^{-R\sigma(t, x)} w(t, x),$$

where  $w$  is a solution of (2.4). Notice that,

$$(3.2) \quad \forall n \in \mathbb{N}, \quad \theta^n z = 0 \quad \text{and} \quad z_x = 0 \quad \text{at time } t = 0 \text{ and } t = T.$$

Moreover  $z$  satisfies

$$(3.3) \quad \begin{cases} (e^{R\sigma} z)_t + (a(e^{R\sigma} z)_x)_x = f, & (t, x) \in Q_T, \\ z(t, 1) = 0, & t \in (0, T), \\ \text{and } \begin{cases} z(t, 0) = 0 & \text{for } 0 \leq \alpha < 1, \\ (az_x)(t, 0) = -R(a\sigma_x z)(t, 0) & \text{for } 1 \leq \alpha < 2, \end{cases} & t \in (0, T). \end{cases}$$

This equation may be recast as follows:

$$P_R z = P_R^+ z + P_R^- z = f e^{-R\sigma},$$

where

$$\begin{aligned} P_R^+ z &:= R\sigma_t z + R^2 a \sigma_x^2 z + (az_x)_x, \\ P_R^- z &:= z_t + R(a\sigma_x z)_x + Ra\sigma_x z_x \\ &= z_t + R(a\sigma_x)_x z + 2Ra\sigma_x z_x. \end{aligned}$$

Moreover, we have

$$(3.4) \quad \|f e^{-R\sigma}\|^2 \geq \|P_R^+ z\|^2 + \|P_R^- z\|^2 + 2\langle P_R^+ z, P_R^- z \rangle \geq 2\langle P_R^+ z, P_R^- z \rangle,$$

where  $\|\cdot\|$  and  $\langle \cdot, \cdot \rangle$  denote the usual norm and scalar product in  $L^2(Q_T)$ .

**3.2. Computation of the scalar product.** We now want to compute the scalar product in  $L^2(Q_T)$  of  $P_R^+ z$  and  $P_R^- z$ . This will be done in two steps.

LEMMA 3.1. *The following identity holds:*

$$\begin{aligned} \langle P_R^+ z, P_R^- z \rangle &= \int_0^T \left[ az_x z_t + R^2 a \sigma_t \sigma_x z^2 + R^3 a^2 \sigma_x^3 z^2 \right. \\ &\quad \left. + Ra(a\sigma_x)_x z z_x + Ra^2 \sigma_x z_x^2 \right]_0^1 \quad (b.t.) \\ &\quad + \left. \int \int_{Q_T} \left( -\frac{1}{2} R \sigma_{tt} - 2R^2 a \sigma_x \sigma_{xt} - R^3 a \sigma_x (a \sigma_x^2)_x \right) z^2 \right. \\ &\quad \left. + \int \int_{Q_T} -R \frac{a}{\sigma_x} (a \sigma_x^2)_x z_x^2 - Ra(a\sigma_x)_{xx} z z_x \right. \quad (d.t.) \end{aligned}$$

Then, using the fact that  $a(x) = x^\alpha$  and  $\sigma(t, x) = \theta(t)p(x)$ , we compute the distributed and boundary terms as follows.

LEMMA 3.2. *For all  $0 \leq \alpha < 2$ , we have*

$$\begin{aligned} (d.t.) = & -\frac{1}{2}R \iint_{Q_T} \theta_{tt} p z^2 - 2R^2 \iint_{Q_T} \theta \theta_t x^\alpha p_x^2 z^2 \\ & - R^3 \iint_{Q_T} \theta^3 x^{2\alpha-1} (2x p_{xx} + \alpha p_x) p_x^2 z^2 \\ & - R \iint_{Q_T} \theta x^{2\alpha-1} (2x p_{xx} + \alpha p_x) z_x^2 - R \iint_{Q_T} \theta x^\alpha (x^\alpha p_x)_{xx} z z_x. \end{aligned}$$

Moreover, for  $0 \leq \alpha < 1$ , the boundary terms (b.t.) are given by

$$(b.t.)_{\text{for } 0 \leq \alpha < 1} = \int_0^T \left\{ R \theta a^2 p_x z_x^2 \right\}_{|x=1} - \int_0^T \left\{ R \theta a^2 p_x z_x^2 \right\}_{|x=0}.$$

For  $1 \leq \alpha < 2$ , the boundary terms (b.t.) become

$$\begin{aligned} (b.t.)_{\text{for } 1 \leq \alpha < 2} = & \int_0^T \left\{ R \theta a^2 p_x z_x^2 \right\}_{|x=1} \\ & + \int_0^T \left\{ -\frac{R}{2} \theta_t a p_x z^2 - R^2 \theta_t \theta a p p_x z^2 - 2R^3 \theta^3 a^2 p_x^3 z^2 + R^2 \theta^2 a p_x (a p_x)_x z^2 \right\}_{|x=0}. \end{aligned}$$

*Proof of Lemma 3.1.* We have

$$\langle P_R^+ z, P_R^- z \rangle = Q_1 + Q_2 + Q_3 + Q_4,$$

where

$$\begin{aligned} Q_1 &:= \langle R \sigma_t z + R^2 a \sigma_x^2 z + (a z_x)_x, z_t \rangle, \\ Q_2 &:= R^2 \langle \sigma_t z, (a \sigma_x)_x z + 2a \sigma_x z_x \rangle, \\ Q_3 &:= R^3 \langle a \sigma_x^2 z, (a \sigma_x)_x z + 2a \sigma_x z_x \rangle, \\ Q_4 &:= R \langle (a z_x)_x, (a \sigma_x)_x z + 2a \sigma_x z_x \rangle. \end{aligned}$$

*First term:*  $Q_1$ .

$$\begin{aligned} Q_1 &= \iint_{Q_T} \left( R \sigma_t z + R^2 a \sigma_x^2 z + (a z_x)_x \right) z_t \\ &= \iint_{Q_T} (R \sigma_t + R^2 a \sigma_x^2) \left( \frac{z^2}{2} \right)_t + \iint_{Q_T} (a z_x)_x z_t \\ &= \left[ \int_0^1 \frac{1}{2} \left( R \sigma_t + R^2 a \sigma_x^2 \right) z^2 \right]_0^T - \iint_{Q_T} \frac{1}{2} \left( R \sigma_t + R^2 a \sigma_x^2 \right)_t z^2 \\ &\quad + \int_0^T \left[ a z_x z_t \right]_0^1 - \iint_{Q_T} a z_x z_{xt} \\ &= \left[ \int_0^1 \left( R \sigma_t + R^2 a \sigma_x^2 \right) \frac{1}{2} z^2 - \frac{1}{2} a z_x^2 \right]_0^T \\ &\quad - \iint_{Q_T} \frac{1}{2} \left( R \sigma_t + R^2 a \sigma_x^2 \right)_t z^2 + \int_0^T \left[ a z_x z_t \right]_0^1. \end{aligned}$$

By (3.2), the terms integrated in time are equal to zero. Hence,

$$(3.5) \quad Q_1 = \int_0^T \left[ az_x z_t \right]_0^1 + \iint_{Q_T} \left( -\frac{1}{2} R \sigma_{tt} - R^2 a \sigma_x \sigma_{xt} \right) z^2.$$

*Second term:  $Q_2$ .*

$$\begin{aligned} Q_2 &= R^2 \iint_{Q_T} \sigma_t z \left( (a\sigma_x)_x z + 2a\sigma_x z_x \right) = R^2 \iint_{Q_T} \sigma_t (a\sigma_x)_x z^2 + a\sigma_t \sigma_x (z^2)_x \\ &= R^2 \iint_{Q_T} \sigma_t (a\sigma_x)_x z^2 + R^2 \int_0^T \left[ a\sigma_t \sigma_x z^2 \right]_0^1 - R^2 \iint_{Q_T} (a\sigma_t \sigma_x)_x z^2. \end{aligned}$$

Therefore,

$$(3.6) \quad Q_2 = R^2 \int_0^T \left[ a\sigma_t \sigma_x z^2 \right]_0^1 - R^2 \iint_{Q_T} a\sigma_x \sigma_{xt} z^2.$$

*Third term:  $Q_3$ .*

$$\begin{aligned} Q_3 &= R^3 \iint_{Q_T} a\sigma_x^2 z \left( (a\sigma_x)_x z + 2a\sigma_x z_x \right) = R^3 \iint_{Q_T} a\sigma_x^2 z \left( (a\sigma_x z)_x + a\sigma_x z_x \right) \\ &= R^3 \int_0^T \left[ a^2 \sigma_x^3 z^2 \right]_0^1 - R^3 \iint_{Q_T} (a\sigma_x^2)_x a\sigma_x z + R^3 \iint_{Q_T} a^2 \sigma_x^3 z z_x. \end{aligned}$$

Thus,

$$(3.7) \quad Q_3 = R^3 \int_0^T \left[ a^2 \sigma_x^3 z^2 \right]_0^1 - R^3 \iint_{Q_T} a\sigma_x (a\sigma_x^2)_x z^2.$$

*Last term:  $Q_4$ .*

$$\begin{aligned} Q_4 &= R \iint_{Q_T} (az_x)_x \left( (a\sigma_x)_x z + 2a\sigma_x z_x \right) \\ &= R \int_0^T \left[ az_x (a\sigma_x)_x z \right]_0^1 - R \iint_{Q_T} az_x \left( (a\sigma_x)_x z \right)_x + R \iint_{Q_T} \sigma_x \left( (az_x)^2 \right)_x \\ &= R \int_0^T \left[ a(a\sigma_x)_x z z_x \right]_0^1 - R \iint_{Q_T} a(a\sigma_x)_x z_x^2 + a(a\sigma_x)_{xx} z z_x \\ &\quad + R \int_0^T \left[ \sigma_x a^2 z_x^2 \right]_0^1 - R \iint_{Q_T} \sigma_{xx} a^2 z_x^2. \end{aligned}$$

Consequently,

$$(3.8) \quad \begin{aligned} Q_4 &= R \int_0^T \left[ a(a\sigma_x)_x z z_x + a^2 \sigma_x z_x^2 \right]_0^1 \\ &\quad - R \iint_{Q_T} \frac{a}{\sigma_x} (a\sigma_x^2)_x z_x^2 - R \iint_{Q_T} a(a\sigma_x)_{xx} z z_x. \end{aligned}$$

Finally, Lemma 3.1 follows from (3.5)–(3.8).  $\square$

*Proof of Lemma 3.2.* With  $a(x) = x^\alpha$  and  $\sigma(t, x) = \theta(t)p(x)$ , the distributed terms (d.t.) can be computed as follows:

$$\begin{aligned}
(d.t.) &= -\frac{1}{2}R \iint_{Q_T} \theta_{tt} p z^2 - 2R^2 \iint_{Q_T} \theta \theta_t x^\alpha p_x^2 z^2 - R^3 \iint_{Q_T} \theta^3 x^\alpha p_x (x^\alpha p_x^2)_x z^2 \\
&\quad - R \iint_{Q_T} \theta \frac{x^\alpha}{p_x} (x^\alpha p_x^2)_x z_x^2 - R \iint_{Q_T} \theta x^\alpha (x^\alpha p_x)_{xx} z z_x \\
&= -\frac{1}{2}R \iint_{Q_T} \theta_{tt} p z^2 - 2R^2 \iint_{Q_T} \theta \theta_t x^\alpha p_x^2 z^2 \\
&\quad - R^3 \iint_{Q_T} \theta^3 x^{2\alpha-1} (2x p_{xx} + \alpha p_x) p_x^2 z^2 \\
&\quad - R \iint_{Q_T} \theta x^{2\alpha-1} (2x p_{xx} + \alpha p_x) z_x^2 - R \iint_{Q_T} \theta x^\alpha (x^\alpha p_x)_{xx} z z_x.
\end{aligned}$$

On the other hand, also taking into account the fact that  $z(t, 1) = 0$ , the boundary terms (b.t.) become

$$\begin{aligned}
(b.t.) &= \int_0^T \left\{ R\theta a^2 p_x z_x^2 \right\}_{|x=1} - \int_0^T \left\{ a z_x z_t + R^2 \theta_t \theta a p p_x z^2 + R^3 \theta^3 a^2 p_x^3 z^2 \right. \\
&\quad \left. + R\theta a (a p_x)_{xx} z z_x + R\theta a^2 p_x z_x^2 \right\}_{|x=0}.
\end{aligned}$$

Now, for  $0 \leq \alpha < 1$ , use the fact that  $z(t, 0) = 0$  to obtain

$$(b.t.)_{\text{for } 0 \leq \alpha < 1} = \int_0^T \left\{ R\theta a^2 p_x z_x^2 \right\}_{|x=1} - \int_0^T \left\{ R\theta a^2 p_x z_x^2 \right\}_{|x=0}.$$

Similarly, for  $1 \leq \alpha < 2$ , recall that  $(a z_x)(t, 0) = -R\theta(t)(a p_x z)(t, 0)$  to conclude that

$$\begin{aligned}
(b.t.)_{\text{for } 1 \leq \alpha < 2} &= \int_0^T \left\{ R\theta a^2 p_x z_x^2 \right\}_{|x=1} + \int_0^T \left\{ R\theta a p_x \left( \frac{z^2}{2} \right)_t \right. \\
&\quad \left. - R^2 \theta_t \theta a p p_x z^2 - R^3 \theta^3 a^2 p_x^3 z^2 + R^2 \theta^2 a p_x (a p_x)_{xx} z^2 - R^3 \theta^3 a^2 p_x^3 z^2 \right\}_{|x=0}.
\end{aligned}$$

Hence

$$\begin{aligned}
(b.t.)_{\text{for } 1 \leq \alpha < 2} &= \int_0^T \left\{ R\theta a^2 p_x z_x^2 \right\}_{|x=1} + \int_0^T \left\{ -\frac{R}{2} \theta_t a p_x z^2 \right. \\
&\quad \left. - R^2 \theta_t \theta a p p_x z^2 - 2R^3 \theta^3 a^2 p_x^3 z^2 + R^2 \theta^2 a p_x (a p_x)_{xx} z^2 \right\}_{|x=0}. \quad \square
\end{aligned}$$

**3.3. Bounds from below.** Let us first define

$$\forall t \in (0, T), \quad \theta(t) := \left( \frac{1}{t(T-t)} \right)^4.$$

Observe that  $\theta$  satisfies the following properties:

$$|\theta_t| \leq c\theta^{5/4} \leq c\theta^2 \quad \text{and} \quad |\theta_{tt}| \leq c\theta^{3/2} \leq c\theta^3.$$

Next, let us recall that  $\alpha \in [0, 2)$  and let us choose

$$\forall x \in [0, 1], \quad p(x) := \frac{2 - x^{2-\alpha}}{(2 - \alpha)^2}.$$

Then

$$p_x(x) = \frac{-x^{1-\alpha}}{2 - \alpha}, \quad p_{xx}(x) = \frac{-(1 - \alpha)}{2 - \alpha} x^{-\alpha}.$$

Hence

$$2xp_{xx} + \alpha p_x = -x^{1-\alpha}$$

and

$$(x^\alpha p_x)_x = \frac{-1}{2 - \alpha}; \quad \text{thus } (x^\alpha p_x)_{xx} = 0.$$

With this choice of  $\theta$  and  $p$ , the distributed and boundary terms can be first computed and then estimated as follows.

LEMMA 3.3. *For all  $\alpha \in [0, 2)$ , the distributed terms (d.t.) become*

$$\begin{aligned} (d.t.) = & -\frac{R}{(2 - \alpha)^2} \iint_{Q_T} \theta_{tt} z^2 + \frac{R}{2(2 - \alpha)^2} \iint_{Q_T} \theta_{tt} x^{2-\alpha} z^2 \\ & - \frac{2R^2}{(2 - \alpha)^2} \iint_{Q_T} \theta \theta_t x^{2-\alpha} z^2 + \frac{R^3}{(2 - \alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2 + R \iint_{Q_T} \theta x^\alpha z_x^2. \end{aligned}$$

For  $0 \leq \alpha < 1$ , the boundary terms (b.t.) become

$$(b.t.)_{\text{for } 0 \leq \alpha < 1} = -\frac{1}{2 - \alpha} \int_0^T \left\{ R\theta z_x^2 \right\}_{|x=1} + \frac{1}{2 - \alpha} \int_0^T \left\{ R\theta x^{1+\alpha} z_x^2 \right\}_{|x=0}.$$

For  $1 \leq \alpha < 2$ , the boundary terms (b.t.) become

$$\begin{aligned} (b.t.)_{\text{for } 1 \leq \alpha < 2} = & -\frac{1}{2 - \alpha} \int_0^T \left\{ R\theta z_x^2 \right\}_{|x=1} + \int_0^T \left\{ \frac{R\theta_t}{2(2 - \alpha)} x z^2 \right. \\ & \left. + \frac{2R^2 \theta_t \theta}{(2 - \alpha)^3} x z^2 - \frac{R^2 \theta_t \theta}{(2 - \alpha)^3} x^{3-\alpha} z^2 + \frac{2R^3 \theta^3}{(2 - \alpha)^3} x^{3-\alpha} z^2 + \frac{R^2 \theta^2}{(2 - \alpha)^2} x z^2 \right\}_{|x=0}. \end{aligned}$$

LEMMA 3.4. *For all  $\alpha \in [0, 2)$ , the distributed terms (d.t.) and the boundary terms (b.t.) satisfy, for  $R$  large enough (depending on  $\alpha$  and  $T$ ),*

$$(d.t.) \geq \frac{1}{4} \frac{R^3}{(2 - \alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2 + \frac{3}{4} R \iint_{Q_T} \theta x^\alpha z_x^2,$$

$$(b.t.) \geq -\frac{1}{2 - \alpha} \int_0^T \left\{ R\theta z_x^2 \right\}_{|x=1}.$$

*Proof of Lemma 3.3.* The conclusion follows from the above choice of  $p$  and the expressions of (d.t.) and (b.t.) given in Lemma 3.2.  $\square$



*Proof of Lemma 3.4.* Let us first analyze the distributed terms. Recall that, owing to Lemma 3.3,

$$\begin{aligned} (d.t.) = & -\frac{R}{(2-\alpha)^2} \iint_{Q_T} \theta_{tt} z^2 + \frac{R}{2(2-\alpha)^2} \iint_{Q_T} \theta_{tt} x^{2-\alpha} z^2 \\ & - \frac{2R^2}{(2-\alpha)^2} \iint_{Q_T} \theta \theta_t x^{2-\alpha} z^2 + \frac{R^3}{(2-\alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2 + R \iint_{Q_T} \theta x^\alpha z_x^2. \end{aligned}$$

Since the two last terms are nonnegative, we only need to estimate the three other terms. We begin with the second term: since  $|\theta_{tt}| \leq c\theta^{3/2} \leq c\theta^3$ , we have

$$\begin{aligned} \left| \frac{R}{2(2-\alpha)^2} \iint_{Q_T} \theta_{tt} x^{2-\alpha} z^2 \right| & \leq \frac{cR}{2(2-\alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2 \\ & \leq \frac{1}{4} \frac{R^3}{(2-\alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2 \end{aligned}$$

for  $R$  large enough. Next, using  $|\theta \theta_t| \leq c\theta^{9/4} \leq c\theta^3$ , we also obtain a bound of the third term for  $R$  large enough:

$$\begin{aligned} \left| \frac{2R^2}{(2-\alpha)^2} \iint_{Q_T} \theta \theta_t x^{2-\alpha} z^2 \right| & \leq \frac{2cR^2}{(2-\alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2 \\ & \leq \frac{1}{4} \frac{R^3}{(2-\alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2. \end{aligned}$$

Therefore,

$$(3.9) \quad (d.t.) \geq -\frac{R}{(2-\alpha)^2} \iint_{Q_T} \theta_{tt} z^2 + \frac{1}{2} \frac{R^3}{(2-\alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2 + R \iint_{Q_T} \theta x^\alpha z_x^2.$$

It remains to bound the first term on the right-hand side above. First let us observe that the solution  $w$  of (2.4) belongs to  $L^2(0, T; H_a^1(0, 1))$  by Theorem 2.1. Since  $z = e^{-R\sigma} w$ , some direct computations imply that  $z$  also belongs to  $L^2(0, T; H_a^1(0, 1))$ . Next we write

$$\begin{aligned} (3.10) \quad \left| \frac{R}{(2-\alpha)^2} \iint_{Q_T} \theta_{tt} z^2 \right| & \leq \frac{cR}{(2-\alpha)^2} \iint_{Q_T} \theta^{3/2} z^2 \\ & = \frac{cR}{(2-\alpha)^2} \iint_{Q_T} \left( \theta x^{(\alpha-2)/3} z^2 \right)^{3/4} \left( \theta^3 x^{2-\alpha} z^2 \right)^{1/4} \\ & \leq \frac{3\epsilon cR}{4(2-\alpha)^2} \iint_{Q_T} \theta x^{(\alpha-2)/3} z^2 + \frac{cR}{4\epsilon^3(2-\alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2. \end{aligned}$$

As this point, we separate the case  $\alpha = 1$  from the other ones. This case is peculiar since Hardy's inequality (Lemma 2.1) does not hold for  $\alpha^* = 1$ .

In the case  $\alpha \neq 1$ , we observe that  $x^{(\alpha-2)/3} \leq x^{\alpha-2}$  (since  $\alpha < 2$ ), and we apply Lemma 2.1 with  $\alpha^* = \alpha \neq 1$  ( $z$  satisfies the assumptions of Lemma 2.1 for almost every  $t$  since it belongs to  $L^2(0, T; H_a^1(0, 1))$ ) to obtain

$$(3.11) \quad \iint_{Q_T} \theta x^{(\alpha-2)/3} z^2 \leq \iint_{Q_T} \theta x^{\alpha-2} z^2 \leq \frac{4}{(\alpha-1)^2} \iint_{Q_T} \theta x^\alpha z_x^2.$$

In the case  $\alpha = 1$ , we apply Lemma 2.1 with  $\alpha^* = 5/3$  and then use the fact that  $x^{5/3} \leq x$  to arrive at a similar conclusion:

$$(3.12) \quad \iint_{Q_T} \theta x^{(\alpha-2)/3} z^2 = \iint_{Q_T} \theta x^{-1/3} z^2 \leq \frac{4}{(\alpha^* - 1)^2} \iint_{Q_T} \theta x^{5/3} z_x^2 \\ \leq 9 \iint_{Q_T} \theta x z_x^2 = 9 \iint_{Q_T} \theta x^\alpha z_x^2.$$

In both cases, combining (3.10) with (3.11) or (3.12), we deduce

$$\left| \frac{R}{(2-\alpha)^2} \iint_{Q_T} \theta_{tt} z^2 \right| \leq \varepsilon c' R \iint_{Q_T} \theta x^\alpha z_x^2 + \frac{cR}{4\varepsilon^3(2-\alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2$$

for some constant  $c' > 0$ . Then, for  $\varepsilon$  small enough and  $R$  large enough, we have

$$(3.13) \quad \left| \frac{R}{(2-\alpha)^2} \iint_{Q_T} \theta_{tt} z^2 \right| \leq \frac{1}{4} R \iint_{Q_T} \theta x^\alpha z_x^2 + \frac{1}{4} \frac{R^3}{(2-\alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2.$$

Summing up, we obtain by (3.9) and (3.13)

$$(d.t.) \geq \frac{1}{4} \frac{R^3}{(2-\alpha)^2} \iint_{Q_T} \theta^3 x^{2-\alpha} z^2 + \frac{3}{4} R \iint_{Q_T} \theta x^\alpha z_x^2 \geq 0.$$

We now turn to the boundary terms. In the case  $0 \leq \alpha < 1$ , there is nothing else to do since, by Lemma 3.3,

$$(b.t.)_{\text{for } 0 \leq \alpha < 1} = -\frac{1}{2-\alpha} \int_0^T \left\{ R \theta z_x^2 \right\}_{|x=1} + \frac{1}{2-\alpha} \int_0^T \left\{ R \theta x^{1+\alpha} z_x^2 \right\}_{|x=0} \\ \geq -\frac{1}{2-\alpha} \int_0^T \left\{ R \theta z_x^2 \right\}_{|x=1}.$$

In the case  $1 \leq \alpha < 2$ , we recall that, by Lemma 3.3,

$$(b.t.)_{\text{for } 1 \leq \alpha < 2} = -\frac{1}{2-\alpha} \int_0^T \left\{ R \theta z_x^2 \right\}_{|x=1} + \int_0^T \left\{ \left( \frac{R \theta_t}{2(2-\alpha)} \right. \right. \\ \left. \left. + \frac{2R^2 \theta_t \theta}{(2-\alpha)^3} - \frac{R^2 \theta_t \theta}{(2-\alpha)^3} x^{2-\alpha} + \frac{2R^3 \theta^3}{(2-\alpha)^3} x^{2-\alpha} + \frac{R^2 \theta^2}{(2-\alpha)^2} \right) x z^2 \right\}_{|x=0}.$$

Thus, applying Lemma 3.5 below (since  $z \in H_a^1(0, 1)$  for almost every  $t$ ), it follows that, for almost every  $t \in (0, T)$ ,

$$x z^2(t, x) \rightarrow 0 \quad \text{as } x \rightarrow 0.$$

Hence,

$$(b.t.)_{\text{for } 1 \leq \alpha < 2} = -\frac{1}{2-\alpha} \int_0^T \left\{ R \theta z_x^2 \right\}_{|x=1}. \quad \square$$

LEMMA 3.5. *Let  $\alpha \in [1, 2)$  be given. Then, for all  $v \in H_a^1(0, 1)$ ,*

$$(3.14) \quad x v^2(x) \rightarrow 0 \quad \text{as } x \rightarrow 0^+.$$

*Proof.* Let  $v$  be given in  $H_a^1(0,1)$ . By the definition of  $H_a^1(0,1)$  in the case  $1 \leq \alpha < 2$ , we know that  $v \in L^2(0,1)$  and  $\sqrt{a}v_x = x^{\alpha/2}v_x \in L^2(0,1)$ . Then  $xv^2 \in L^1(0,1)$ . Moreover,

$$(xv^2)_x = v^2 + 2xvv_x,$$

with  $v^2 \in L^1(0,1)$  and with  $xvv_x = (x^{1-\alpha/2}v)(x^{\alpha/2}v_x) \in L^1(0,1)$ . Hence,  $xv^2 \in W^{1,1}(0,1)$ . Thus,  $xv^2 \rightarrow L \geq 0$  as  $x \rightarrow 0^+$ . Finally,  $L = 0$  since  $L \neq 0$  would imply  $v \notin L^2(0,1)$ . This completes the proof.  $\square$

**3.4. Conclusion.** From Lemmas 3.1 and 3.4 we obtain, for all  $0 \leq \alpha < 2$ ,

$$\begin{aligned} \langle P_R^+ z, P_R^- z \rangle &= (d.t.) + (b.t.) \\ &\geq cR^3 \iint_{Q_T} \theta^3 x^{2-\alpha} z^2 + cR \iint_{Q_T} \theta x^\alpha z_x^2 - c' \int_0^T \left\{ R\theta z_x^2 \right\}_{|x=1} \end{aligned}$$

for some constants  $c, c' > 0$ . By (3.4), we have

$$\begin{aligned} \|f e^{-R\sigma}\|^2 &= \|P_R^+ z\|^2 + \|P_R^- z\|^2 + 2\langle P_R^+ z, P_R^- z \rangle \geq 2\langle P_R^+ z, P_R^- z \rangle \\ &\geq cR^3 \iint_{Q_T} \theta^3 x^{2-\alpha} z^2 + cR \iint_{Q_T} \theta x^\alpha z_x^2 - c' \int_0^T \left\{ R\theta z_x^2 \right\}_{|x=1}. \end{aligned}$$

We recall that  $\sigma(t, x) = \theta(t)p(x)$  and  $p_x(x) = -x^{1-\alpha}/(2-\alpha)$ . Hence,  $x^\alpha \sigma_x^2 = c\theta^2 x^\alpha x^{2-2\alpha} = c\theta^2 x^{2-\alpha}$ . Moreover,  $w = e^{R\sigma} z$ . Thus,  $w_x = R\sigma_x e^{R\sigma} z + e^{R\sigma} z_x$ . Therefore,

$$\begin{aligned} R^3 \theta^3 x^{2-\alpha} w^2 + R\theta x^\alpha w_x^2 &\leq R^3 \theta^3 x^{2-\alpha} e^{2R\sigma} z^2 + R\theta x^\alpha \left( 2R^2 \sigma_x^2 e^{2R\sigma} z^2 + 2e^{2R\sigma} z_x^2 \right) \\ &\leq c \left( R^3 \theta^3 x^{2-\alpha} e^{2R\sigma} z^2 + R\theta x^\alpha e^{2R\sigma} z_x^2 \right). \end{aligned}$$

So,

$$\iint_{Q_T} \left( R^3 \theta^3 x^{2-\alpha} w^2 + R\theta x^\alpha w_x^2 \right) e^{-2R\sigma} \leq c \iint_{Q_T} f^2 e^{-2R\sigma} + c \int_0^T \left\{ R\theta z_x^2 \right\}_{|x=1}.$$

Moreover  $z_x(x=1) = (e^{-R\sigma} w_x)(x=1)$  since  $z(x=1) = 0$ . It follows that

$$\begin{aligned} \iint_{Q_T} \left( R^3 \theta^3 x^{2-\alpha} w^2 + R\theta x^\alpha w_x^2 \right) e^{-2R\sigma} \\ \leq c \iint_{Q_T} f^2 e^{-2R\sigma} + c \int_0^T \left\{ R\theta e^{-2R\sigma} w_x^2 \right\}_{|x=1}. \quad \square \end{aligned}$$

**4. Proof of Theorem 2.3 (observability inequalities).** Theorem 2.2 yields a Carleman estimate for the solutions of (2.5).

LEMMA 4.1. *For all  $0 \leq \alpha < 2$  and all  $T > 0$ , there exist positive constants,  $R_0, C, c > 0$ , such that, for all  $v_T \in L^2(0,1)$ , the solution  $v$  of (2.5) satisfies, for all  $R \geq R_0$ ,*

$$\int_0^T \int_0^1 \left( R\theta x^\alpha v_x^2 + R^3 \theta^3 x^{2-\alpha} v^2 \right) e^{-2cR\theta} dx dt \leq C \int_0^T \int_\omega v^2 dx dt.$$

Let us put off the proof of the above lemma and proceed with the reasoning. Multiplying the equation in (2.5) by  $v_t$  and integrating by parts, we get

$$\begin{aligned} 0 &= \int_0^1 \left( v_t + (x^\alpha v_x)_x \right) v_t dx \\ &= \int_0^1 v_t^2 dx + \left[ x^\alpha v_x v_t \right]_0^1 - \int_0^1 x^\alpha v_x v_{tx} dx \geq -\frac{1}{2} \frac{d}{dt} \int_0^1 x^\alpha v_x^2 dx. \end{aligned}$$

Therefore  $t \mapsto \int_0^1 x^\alpha v_x^2 dx$  is increasing and

$$\int_0^1 x^\alpha v_x(0, x)^2 dx \leq \int_0^1 x^\alpha v_x(t, x)^2 dx \quad \forall t \in [0, T].$$

Integrating over  $[T/4, 3T/4]$ , we have

$$\begin{aligned} \int_0^1 x^\alpha v_x(0, x)^2 dx &\leq \frac{2}{T} \int_{T/4}^{3T/4} \int_0^1 x^\alpha v_x(t, x)^2 dx dt \\ &\leq C \int_{T/4}^{3T/4} \int_0^1 \theta x^\alpha v_x(t, x)^2 e^{-2cR\theta} dx dt. \end{aligned}$$

Hence, owing to Lemma 4.1,

$$(4.1) \quad \int_0^1 x^\alpha v_x(0, x)^2 dx \leq C \int_0^T \int_\omega v^2 dx dt. \quad \square$$

*Proof of Lemma 4.1.* Let  $\omega = (x_0, x_1)$  with  $0 \leq x_0 < x_1 \leq 1$  and consider a smooth cut-off function  $\psi : \mathbb{R} \rightarrow \mathbb{R}$ , such that

$$\begin{cases} 0 \leq \psi(x) \leq 1 & \forall x \in \mathbb{R}, \\ \psi(x) = 1 & \text{for } x \in (0, (2x_0 + x_1)/3), \\ \psi(x) = 0 & \text{for } x \in ((x_0 + 2x_1)/3, 1). \end{cases}$$

We define  $w := \psi v$  where  $v$  is the solution of (2.5). Then  $w$  satisfies

$$(4.2) \quad \begin{cases} w_t + (aw_x)_x = (a\psi_x v)_x + \psi_x a v_x =: f, & (t, x) \in Q_T, \\ w(t, 1) = 0, & t \in (0, T), \\ \text{and } \begin{cases} w(t, 0) = 0 & \text{for } 0 \leq \alpha < 1, \\ (aw_x)(t, 0) = 0 & \text{for } 1 \leq \alpha < 2, \end{cases} & t \in (0, T). \end{cases}$$

Therefore, applying Theorem 2.2 and using the fact that  $w \equiv 0$  in a neighborhood of  $x = 1$  (hence  $w_x(1, t) = 0$ ), we have, for all  $R \geq R_0$ ,

$$\iint_{Q_T} \left( R\theta x^\alpha w_x^2 + R^3 \theta^3 x^{2-\alpha} w^2 \right) e^{-2R\sigma} dx dt \leq C \iint_{Q_T} e^{-2R\sigma} f^2 dx dt.$$

Then using the definition of  $\psi$  and in particular the fact that  $\psi_x$  and  $\psi_{xx}$  are supported in  $\omega' := ((2x_0 + x_1)/3, (x_0 + 2x_1)/3)$ , we can write

$$f^2 = \left( (a\psi_x v)_x + \psi_x a v_x \right)^2 = \left( a_x \psi_x v + 2a\psi_x v_x + a\psi_{xx} v \right)^2 \chi_{\omega'} \leq C(v^2 + v_x^2) \chi_{\omega'},$$

since the function  $a_x$  is bounded on  $\omega'$ . Hence

$$(4.3) \quad \iint_{Q_T} \left( R\theta x^\alpha w_x^2 + R^3\theta^3 x^{2-\alpha} w^2 \right) e^{-2R\sigma} dx dt \leq C \int_0^T \int_{\omega'} e^{-2R\sigma} (v_x^2 + v^2) dx dt,$$

where  $\omega' := ((2x_0 + x_1)/3, (x_0 + 2x_1)/3)$ . At this point, let us apply the following standard estimate, to be proved later on.

LEMMA 4.2 (Caccioppoli's inequality). *For all  $R > 0$ ,*

$$\int_0^T \int_{\omega'} e^{-2R\sigma} v_x^2 dx dt \leq C(R, T) \int_0^T \int_{\omega} v^2 dx dt.$$

Let us continue with the proof of Lemma 4.1. The proof of Lemma 4.2 will be given later. By (4.3) and Lemma 4.2, we obtain a bound for  $v$  on  $(0, (2x_0 + x_1)/3)$  of the form

$$\begin{aligned} & \int_0^T \int_0^{(2x_0+x_1)/3} \left( R\theta x^\alpha v_x^2 + R^3\theta^3 x^{2-\alpha} v^2 \right) e^{-2R\sigma} dx dt \\ &= \int_0^T \int_0^{(2x_0+x_1)/3} \left( R\theta x^\alpha w_x^2 + R^3\theta^3 x^{2-\alpha} w^2 \right) e^{-2R\sigma} dx dt \\ &\leq \iint_{Q_T} \left( R\theta x^\alpha w_x^2 + R^3\theta^3 x^{2-\alpha} w^2 \right) e^{-2R\sigma} dx dt \leq C_R \int_0^T \int_{\omega} v^2 dx dt. \end{aligned}$$

Hence,

$$\int_0^T \int_0^{(2x_0+x_1)/3} \left( R\theta x^\alpha v_x^2 + R^3\theta^3 x^{2-\alpha} v^2 \right) e^{-2c_0 R\theta} dx dt \leq C_R \int_0^T \int_{\omega} v^2 dx dt,$$

where  $c_0 = \max \{p(x); x \in [0, 1]\} = 2/(2 - \alpha)^2$ .

Now, to complete the reasoning, one has to recover a similar inequality on the interval  $((x_0 + 2x_1)/3, 1)$ . But the equation is uniformly parabolic on such a domain. Therefore, the well-known Carleman estimate for the nondegenerate case (see [26]) yields, for  $R$  large enough,

$$\int_0^T \int_{(x_0+2x_1)/3}^1 \left( R\theta v_x^2 + R^3\theta^3 v^2 \right) e^{-2c_1 R\theta} dx dt \leq C_R \int_0^T \int_{\omega} v^2 dx dt$$

for some constant  $c_1 > 0$ . Indeed it is sufficient to apply the classical Carleman estimate to the function  $\tilde{v} = \rho v$  in the space interval  $((2x_0 + x_1)/3, 1)$ , where  $\rho : \mathbb{R} \rightarrow \mathbb{R}$  is some smooth cut-off function, such that

$$\begin{cases} 0 \leq \rho(x) \leq 1 & \forall x \in \mathbb{R}, \\ \rho(x) = 1 & \text{for } x \in ((x_0 + 2x_1)/3, 1), \\ \rho(x) = 0 & \text{for } x \in (0, (2x_0 + x_1)/3). \end{cases}$$

Hence we obtain

$$\int_0^T \int_{(x_0+2x_1)/3}^1 \left( R\theta x^\alpha v_x^2 + R^3\theta^3 x^{2-\alpha} v^2 \right) e^{-2c_1 R\theta} dx dt \leq C_R \int_0^T \int_{\omega} v^2 dx dt.$$

Combining the above estimates and using Lemma 4.2 to bound the integral on the middle interval, we obtain

$$\int_0^T \int_0^1 \left( R\theta x^\alpha v_x^2 + R^3 \theta^3 x^{2-\alpha} v^2 \right) e^{-2c_2 R\theta} dx dt \leq C_R \int_0^T \int_\omega v^2 dx dt,$$

where  $c_2 = \max(c_0, c_1)$ .  $\square$

*Proof of Lemma 4.2.* Consider a smooth function  $\xi : \mathbb{R} \rightarrow \mathbb{R}$  such that

$$\begin{cases} 0 \leq \xi(x) \leq 1 & \forall x \in \mathbb{R}, \\ \xi(x) = 1 & \text{for } x \in \omega', \\ \xi(x) = 0 & \text{for } x \notin \omega. \end{cases}$$

Then, for all  $R > 0$ ,

$$\begin{aligned} 0 &= \int_0^T \frac{d}{dt} \int_0^1 \xi^2 e^{-2R\sigma} v^2 = \iint_{Q_T} -2\xi^2 R\sigma_t e^{-2R\sigma} v^2 + 2\xi^2 e^{-2R\sigma} v v_t \\ &= -2 \iint_{Q_T} \xi^2 R\sigma_t e^{-2R\sigma} v^2 - 2 \iint_{Q_T} \xi^2 e^{-2R\sigma} v (av_x)_x \\ &= -2 \iint_{Q_T} \xi^2 R\sigma_t e^{-2R\sigma} v^2 + 2 \iint_{Q_T} (\xi^2 e^{-2R\sigma})_x a v_x \\ &= -2 \iint_{Q_T} \xi^2 R\sigma_t e^{-2R\sigma} v^2 + 2 \iint_{Q_T} a (\xi^2 e^{-2R\sigma})_x v v_x + \xi^2 e^{-2R\sigma} a v_x^2. \end{aligned}$$

Hence,

$$\begin{aligned} 2 \iint_{Q_T} \xi^2 e^{-2R\sigma} a v_x^2 &= 2 \iint_{Q_T} \xi^2 R\sigma_t e^{-2R\sigma} v^2 - 2 \iint_{Q_T} a (\xi^2 e^{-2R\sigma})_x v v_x \\ &= 2 \iint_{Q_T} \xi^2 R\sigma_t e^{-2R\sigma} v^2 - 2 \iint_{Q_T} \left( \sqrt{a} \xi e^{-R\sigma} v_x \right) \left( \sqrt{a} \frac{(\xi^2 e^{-2R\sigma})_x}{\xi e^{-R\sigma}} v \right) \\ &\leq 2 \iint_{Q_T} \xi^2 R\sigma_t e^{-2R\sigma} v^2 + \iint_{Q_T} \left( \sqrt{a} \xi e^{-R\sigma} v_x \right)^2 + \iint_{Q_T} \left( \sqrt{a} \frac{(\xi^2 e^{-2R\sigma})_x}{\xi e^{-R\sigma}} v \right)^2 \\ &\leq 2 \iint_{Q_T} \xi^2 R\sigma_t e^{-2R\sigma} v^2 + \iint_{Q_T} \left( \sqrt{a} \frac{(\xi^2 e^{-2R\sigma})_x}{\xi e^{-R\sigma}} v \right)^2 + \iint_{Q_T} \xi^2 e^{-2R\sigma} a v_x^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \iint_{Q_T} \xi^2 e^{-2R\sigma} a v_x^2 &\leq 2 \iint_{Q_T} \xi^2 R\sigma_t e^{-2R\sigma} v^2 + \iint_{Q_T} \left( \sqrt{a} \frac{(\xi^2 e^{-2R\sigma})_x}{\xi e^{-R\sigma}} v \right)^2 \\ &\leq C(R, T) \int_0^T \int_\omega v^2. \quad \square \end{aligned}$$

**5. Proof of Theorem 2.4 (controllability result).** By standard arguments, the null controllability result stated in Theorem 2.4 follows from (2.7). Hence it remains to prove (2.7). For  $\alpha \neq 1$ , we apply Hardy's inequality (Lemma 2.1) with  $\alpha^* = \alpha$  to deduce from (2.6) that

$$\int_0^1 x^{\alpha-2} v(0, x)^2 dx \leq C \int_0^1 x^\alpha v_x(0, x)^2 dx \leq C \int_0^T \int_\omega v^2 dx dt.$$

In the case of  $\alpha = 1$ , from (2.6) we deduce that, for all  $0 < \eta < 1$ ,

$$\int_0^1 x^{1+\eta} v_x(0, x)^2 dx \leq \int_0^1 x v_x(0, x)^2 dx \leq C \int_0^T \int_\omega v^2 dx dt.$$

Now, applying Hardy's inequality (Lemma 2.1) with  $\alpha^* = 1 + \eta$ , we obtain

$$\int_0^1 x^{\eta-1} v(0, x)^2 dx \leq C \int_0^1 x^{1+\eta} v_x(0, x)^2 dx \leq C \int_0^T \int_\omega v^2 dx dt.$$

In both cases, (2.7) follows.  $\square$

## 6. Proof of Lemma 2.1 (Hardy's inequalities).

*First case.*  $0 \leq \alpha^* < 1$ . Since  $z$  is absolutely continuous on  $(0, 1)$ , we have

$$\begin{aligned} |z(x) - z(\varepsilon)|^2 &= \left( \int_\varepsilon^x z_x(s) s^{(3-\gamma)/4} s^{(-3+\gamma)/4} ds \right)^2 \\ &\leq \left( \int_\varepsilon^x z_x(s)^2 s^{(3-\gamma)/2} ds \right) \left( \int_\varepsilon^x s^{(-3+\gamma)/2} ds \right), \end{aligned}$$

where we denote  $\gamma := 2 - \alpha^* \in (1, 2]$ . Letting  $\varepsilon \rightarrow 0^+$ , we get

$$|z(x)|^2 \leq \left( \int_0^x z_x(s)^2 s^{(3-\gamma)/2} ds \right) \left( \int_0^x s^{(-3+\gamma)/2} ds \right).$$

Therefore

$$\begin{aligned} \int_0^1 x^{\alpha^*-2} z(x)^2 dx &\leq \int_0^1 x^{-\gamma} \left( \int_0^x z_x(s)^2 s^{(3-\gamma)/2} ds \right) \left( \int_0^x s^{(-3+\gamma)/2} ds \right) dx \\ &= \int_0^1 x^{-\gamma} \left( \int_0^x z_x(s)^2 s^{(3-\gamma)/2} ds \right) \frac{x^{(\gamma-1)/2}}{(\gamma-1)/2} dx \\ &= \frac{2}{\gamma-1} \int_0^1 z_x(s)^2 s^{(3-\gamma)/2} \left( \int_s^1 x^{(-\gamma-1)/2} dx \right) ds \\ &\leq \frac{2}{\gamma-1} \int_0^1 z_x(s)^2 s^{(3-\gamma)/2} \frac{s^{(1-\gamma)/2}}{(\gamma-1)/2} ds = \frac{4}{(1-\alpha^*)^2} \int_0^1 s^{\alpha^*} z_x(s)^2 ds. \end{aligned}$$

*Second case.*  $1 < \alpha^* < 2$ . Denoting  $\gamma := 2 - \alpha^* \in (0, 1)$ , we have

$$\begin{aligned} \int_0^1 x^{\alpha^*-2} z(x)^2 dx &\leq \int_0^1 x^{-\gamma} \left( \int_x^1 z_x(s)^2 s^{(3-\gamma)/2} ds \right) \left( \int_x^1 s^{(-3+\gamma)/2} ds \right) dx \\ &\leq \int_0^1 x^{-\gamma} \left( \int_x^1 z_x(s)^2 s^{(3-\gamma)/2} ds \right) \frac{x^{-(1-\gamma)/2}}{(1-\gamma)/2} dx \\ &= \frac{2}{1-\gamma} \int_0^1 z_x(s)^2 s^{(3-\gamma)/2} \left( \int_0^s x^{(-\gamma-1)/2} dx \right) ds \\ &\leq \frac{4}{(1-\gamma)^2} \int_0^1 z_x(s)^2 s^{2-\gamma} ds = \frac{4}{(\alpha^*-1)^2} \int_0^1 s^{\alpha^*} z_x(s)^2 ds. \quad \square \end{aligned}$$

**Acknowledgments.** The authors wish to thank O. Imanuvilov for many fruitful discussions on the subject and for suggestions concerning the case  $\alpha = 1$  and the referees for their valuable comments that helped improve the presentation of this paper. They also wish to thank the Université Paul Sabatier Toulouse III, the Università di Roma “Tor Vergata,” and the Istituto Nazionale di Alta Matematica for their hospitality and financial support.

## REFERENCES

- [1] F. ALABAU-BOUSSOIRA, P. CANNARSA, AND G. FRAGNELLI, *Carleman estimates for degenerate parabolic operators with applications to null controllability*, J. Evol. Equ., 6 (2006), pp. 161–204.
- [2] S. ANIȚA AND V. BARBU, *Null controllability of nonlinear convective heat equations*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 157–173.
- [3] S. ANIȚA AND D. TATARU, *Null controllability for the dissipative semilinear heat equation*, Appl. Math. Optim., 46 (2002), pp. 97–105.
- [4] V. BARBU, *On local controllability of Navier-Stokes equations*, Adv. Differential Equations, 8 (2003), pp. 1481–1498.
- [5] J.-M. BUCHOT AND J.-P. RAYMOND, *A linearized model for boundary layer equations*, in Optimal Control of Complex Structures (Oberwolfach, 2000), Internat. Ser. Numer. Math. 139, Birkhäuser, Basel, 2002, pp. 31–42.
- [6] M. CAMPITI AND G. METAFUNE, *Ventcel’s boundary conditions and analytic semigroups*, Arch. Math. (Basel), 70 (1998), pp. 377–390.
- [7] M. CAMPITI, G. METAFUNE, AND D. PALLARA, *Degenerate self-adjoint evolution equations on the unit interval*, Semigroup Forum, 57 (1998), pp. 1–36.
- [8] P. CANNARSA, P. MARTINEZ, AND J. VANCOSTENOBLE, *Persistent regional null controllability for a class of degenerate parabolic equations*, Commun. Pure Appl. Anal., 3 (2004), pp. 607–635.
- [9] P. CANNARSA, P. MARTINEZ, AND J. VANCOSTENOBLE, *Null controllability of degenerate heat equations*, Adv. Differential Equations, 10 (2005), pp. 153–190.
- [10] J.-M. CORON, *On the controllability of the 2-D incompressible Navier-Stokes equations with the Navier slip boundary conditions*, ESAIM Contrôle Optim. Calc. Var., 1 (1995/96), pp. 35–75.
- [11] J.-M. CORON AND A. V. FURSIKOV, *Global exact controllability of the 2D Navier-Stokes equations on a manifold without boundary*, Russian J. Math. Phys., 4 (1996), pp. 429–448.
- [12] J.-M. CORON AND E. TRÉLAT, *Global steady-state controllability of one-dimensional semilinear heat equations*, SIAM J. Control Optim., 43 (2004), pp. 549–569.
- [13] L. DE TERESA, *Approximate controllability of a semilinear heat equation in  $\mathbb{R}^N$* , SIAM J. Control Optim., 36 (1998), pp. 2128–2147.
- [14] SZ. DOLECKI AND D. L. RUSSELL, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–220.
- [15] A. DOUBOVA, E. FERNÁNDEZ-CARA, M. GONZÁLEZ-BURGOS, AND E. ZUAZUA, *On the controllability of parabolic systems with a nonlinear term involving the state and the gradient*, SIAM J. Control Optim., 41 (2002), pp. 798–819.
- [16] L. ESCAURIAZA, G. SEREGIN, AND V. ŠVERÁK, *Backward uniqueness for parabolic equations*, Arch. Ration. Mech. Anal., 169 (2003), pp. 147–157.
- [17] L. ESCAURIAZA, G. SEREGIN, AND V. ŠVERÁK, *Backward uniqueness for the heat operator in half-space*, St. Petersburg Math. J., 15 (2004), pp. 139–148.
- [18] C. FABRE AND G. LEBEAU, *Prolongement unique des solutions de l’équation de Stokes*, Comm. Partial Differential Equations, 21 (1996), pp. 573–596.
- [19] H. O. FATTORINI AND D. L. RUSSELL, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rational Mech. Anal., 4 (1971), pp. 272–292.
- [20] H. O. FATTORINI AND D. L. RUSSELL, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. Appl. Math., 32 (1974), pp. 45–69.
- [21] E. FERNÁNDEZ-CARA, *Null controllability of the semilinear heat equation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 87–103.
- [22] E. FERNÁNDEZ-CARA, S. GUERRERO, O. YU. IMANUVILOV, AND J.-P. PUEL, *Local exact controllability of the Navier-Stokes system*, J. Math. Pures Appl. (9), 83 (2004), pp. 1501–1542.
- [23] E. FERNÁNDEZ-CARA, S. GUERRERO, O. YU. IMANUVILOV, AND J.-P. PUEL, *On the controllability*



- ity of the  $N$ -dimensional Navier-Stokes and Boussinesq systems with  $N-1$  scalar controls*, C.R. Math. Acad. Sci. Paris, 340 (2005), pp. 275–280.
- [24] E. FERNÁNDEZ-CARA AND E. ZUAZUA, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.
  - [25] E. FERNÁNDEZ-CARA AND E. ZUAZUA, *Controllability for weakly blowing-up semilinear heat equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 583–616.
  - [26] A. V. FURSIKOV AND O. YU IMANUVILOV, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Seoul, Korea, 1996.
  - [27] O. YU. IMANUVILOV, *Boundary controllability of parabolic equations*, Mat. Sb., 186 (1995), pp. 109–132.
  - [28] O. YU. IMANUVILOV, *Remarks on exact controllability for the Navier-Stokes equations*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 39–72.
  - [29] I. LASIECKA AND R. TRIGGIANI, *Carleman estimates and exact boundary controllability for a system of coupled, non conservative second order hyperbolic equations*, in Partial Differential Equations Methods in Control and Shape Analysis, Lecture Notes in Pure and Appl. Math. 188, Marcel Dekker, New York, 1997, pp. 215–243.
  - [30] G. LEBEAU AND L. ROBBIANO, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.
  - [31] A. LOPEZ, X. ZHANG, AND E. ZUAZUA, *Null controllability of the heat equation as singular limit of the exact controllability of dissipative wave equations*, J. Math. Pures Appl., 79 (2000), pp. 741–808.
  - [32] P. MARTINEZ, J.-P. RAYMOND, AND J. VANCOSTENOBLE, *Regional null controllability of a linearized Crocco-type equation*, SIAM J. Control Optim., 42 (2003), pp. 709–728.
  - [33] S. MICU AND E. ZUAZUA, *On the lack of null controllability of the heat equation on the half-line*, Trans. Amer. Math. Soc., 353 (2001), pp. 1635–1659.
  - [34] L. MILLER, *On the null-controllability of the heat equation in unbounded domains*, Bull. Sci. Math., 129 (2005), pp. 175–185.
  - [35] B. OPIC AND A. KUFNER, *Hardy-Type Inequalities*, Longman Scientific and Technical, Harlow, UK, 1990.
  - [36] L. ROSIER, *Exact boundary controllability for the linear Korteweg–de Vries equation on the half-line*, SIAM J. Control Optim., 39 (2000), pp. 331–351.
  - [37] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–221.
  - [38] T. I. SEIDMAN, *Exact boundary control for some evolution equations*, SIAM J. Control Optim., 16 (1978), pp. 979–999.
  - [39] K. TAIRA, A. FAVINI, AND S. ROMANELLI, *Feller semigroups and degenerate elliptic operators with Wentzell boundary conditions*, Studia Math., 145 (2001), pp. 17–53.
  - [40] D. TATARU, *Carleman estimates and unique continuation near the boundary for P.D.E.'s*, J. Math. Pures Appl. (9), 75 (1996), pp. 367–408.
  - [41] X. ZHANG, *A remark on null exact controllability of the heat equation*, SIAM J. Control Optim., 40 (2001), pp. 39–53.

## OPTIMAL CONTROL OF A CONVECTIVE BOUNDARY CONDITION IN A THERMISTOR PROBLEM\*

VOLODYMYR HRYNKIV<sup>†</sup>, SUZANNE LENHART<sup>‡</sup>, AND VLADIMIR PROTOPODESCU<sup>§</sup>

**Abstract.** We consider the optimal control of a two-dimensional steady-state thermistor. The problem is described by a system of two nonlinear elliptic partial differential equations with appropriate boundary conditions which model the coupling of the thermistor to its surroundings. Based on physical considerations, an objective functional to be minimized is introduced and the convective boundary coefficient is taken as the control. Existence and uniqueness of the optimal control are proven. To characterize this optimal control, the optimality system consisting of the state and adjoint equations is derived.

**Key words.** optimal control, thermistor problem, elliptic systems

**AMS subject classifications.** 49J20, 49K20

**DOI.** 10.1137/06066401X

**1. Introduction.** Thermistor is a generic name for a device made from materials whose electrical conductivity is highly dependent on temperature. The advantages of thermistors as temperature measurement devices include their low cost, high resolution, and flexibility in size and shape. The applications of thermistors include [11, 13]

- (i) temperature sensing and control: thermistors provide inexpensive and reliable temperature sensing for a wide temperature range;
- (ii) thermal relay and switch: voltage regulation, surge protection;
- (iii) indirect measurement of other parameters: when a thermistor is heated its rate of change of temperature depends on its surroundings. This property can be used to monitor other quantities such as liquid level and fluid flow; and
- (iv) long-wavelength detector.

We consider the two-dimensional steady-state thermistor problem

$$\begin{aligned}
 \nabla \cdot (\sigma(u) \nabla \varphi) &= 0 & \text{in } \Omega, \\
 \Delta u + \sigma(u) |\nabla \varphi|^2 &= 0 & \text{in } \Omega, \\
 \frac{\partial u}{\partial n} + \beta u &= 0 & \text{on } \partial\Omega, \\
 \varphi &= \varphi_0 & \text{on } \partial\Omega,
 \end{aligned}
 \tag{1.1}$$

---

\*Received by the editors June 29, 2006; accepted for publication (in revised form) June 4, 2007; published electronically January 4, 2008. The research of the second and third authors was supported in part by the Division of Material Science of the U.S. Department of Energy under contract DE-AC05-00OR22725 with UT-Batelle, LLC. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

<http://www.siam.org/journals/sicon/47-1/66401.html>

<sup>†</sup>Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609 (vhrynkiv@wpi.edu).

<sup>‡</sup>Department of Mathematics, University of Tennessee, Ayres Hall 121, Knoxville, TN 37996 (lenhart@math.utk.edu).

<sup>§</sup>Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 (protopopesva@ornl.gov).

where  $\varphi(x)$  is the electric potential,  $u(x)$  is the temperature, and  $\sigma(u)$  is the temperature-dependent electrical conductivity. Here  $n$  denotes the outward unit normal and  $\partial/\partial n = n \cdot \nabla$  is the normal derivative on  $\partial\Omega$ . The first equation represents the conservation of charge and the second equation describes the steady diffusion of heat in the presence of Joule heating due to the electric current. Boundary conditions show how the thermistor is connected thermally and electrically to its surroundings. Throughout the paper, solutions to system (1.1) are understood in the weak sense. For a more detailed discussion about the physical justification of equations (1.1) the reader is referred to [7, 9, 15].

It is known that large temperature gradients may cause the thermistor to crack. Numerical experiments indicate (see [7, 21]) that low values of the heat transfer coefficient  $\beta$  will result in small temperature variations. On the other hand, low values of the heat transfer coefficient lead to high operating temperatures of a thermistor which are also undesirable. This motivated us to consider the heat transfer coefficient as a control in the optimal control problem of minimizing the heat transfer while keeping the operating temperature of the thermistor reasonably low. These requirements lead us to the following objective functional:

$$J(\beta) = \int_{\Omega} u \, dx + \int_{\partial\Omega} \beta^2 \, ds.$$

Denoting the set of admissible controls by

$$U_M = \{\beta \in L^\infty(\partial\Omega) : 0 < \lambda \leq \beta \leq M\},$$

the optimal control problem is as follows:

$$(1.2) \quad \text{Find } \beta^* \in U_M \text{ such that } J(\beta^*) = \min_{\beta \in U_M} J(\beta).$$

Henceforth we use the standard notation for Sobolev spaces: we denote  $\|\cdot\|_p = \|\cdot\|_{L^p(\Omega)}$  for each  $p \in [1, \infty]$ ; other norms will be clearly marked.

Theoretical analysis of both the steady-state and time-dependent thermistor equations with different types of boundary and initial conditions has received a significant amount of attention. See [1, 4, 6, 5, 7, 9, 15, 17, 18, 19] for existence of weak solutions, uniqueness, and related regularity results in different settings with various assumptions on the coefficients. For example, existence of a weak solution to a stationary problem with Dirichlet boundary conditions was proven by Cimatti and Prodi in [6], whereas the time-dependent case in two dimensions was first considered by Cimatti in [4]. This restriction on the space dimension was eliminated by Shi, Shillor, and Xu in [15]. Asymptotic results for the time-dependent thermistor problem can be found in [7]. So far, the only known optimal control paper on a thermistor problem is a time-dependent case by Lee and Shilkin in [12], where the source is taken to be the control. Thus our work is the first result on optimal control of the thermistor problem for the steady-state case.

In section 2 we derive a priori estimates under the assumption of small boundary data. In section 3 we prove existence of an optimal control. Also, in section 3 we explain why the space dimension is restricted to  $N = 2$ . The optimality system is derived and an optimal control is characterized in section 4. Uniqueness of the optimal control is proven in section 5. For more details on the estimates in this paper, see [10].

**2. A priori estimates.** Throughout we make the following assumptions:

1.  $\Omega \subset R^2$  is a bounded domain with smooth boundary;
2.  $\sigma(s) \in C^1(R)$ ,  $0 < C_1 \leq \sigma(s) \leq C_2$  for all  $s \in R$ ;
3.  $\sigma(s)$  is Lipschitz:  $|\sigma(s_1) - \sigma(s_2)| \leq K |s_1 - s_2|$  for all  $s_1, s_2 \in R$ ;
4.  $\varphi_0|_{\partial\Omega} \in W^{1,\infty}(\partial\Omega)$ ;
5.  $\|\varphi_0\|_{W^{1,\infty}(\Omega)} \leq C_H$  is sufficiently small, where  $C_H$  is derived from formula (30) in [9].

We will need the following result due to Meyers [3, 14, 16].

**THEOREM 2.1.** *Let  $\Omega \subset R^N$  be a bounded domain with a smooth boundary and suppose that  $A : H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$  is the uniformly elliptic operator with bounded coefficients*

$$A = - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left( a_{ij}(x) \frac{\partial}{\partial x_j} \right).$$

Consider the Dirichlet problem

$$(2.1) \quad Av = f, \quad v \in H_0^1(\Omega), f \in H^{-1}(\Omega).$$

Then there exists  $r > 2$  (which depends on  $\alpha$ , the  $L^\infty$  norm of the coefficients  $a_{ij}$ 's, the domain  $\Omega$ , and the dimension) such that if  $f \in W^{-1,r}(\Omega)$ , then  $v \in W_0^{1,r}(\Omega)$ , where  $v$  solves (2.1) and satisfies

$$\|v\|_{W_0^{1,r}(\Omega)} \leq C \|f\|_{W^{-1,r}(\Omega)}$$

and  $C$  depends on the same quantities as  $r$ .

To derive a weak formulation for (1.1) we follow Lemma 1 from [9] and the discussion preceding it. Note that we extend  $\varphi_0$  to the whole domain  $\Omega$  and use the same notation, i.e.,  $\varphi_0 \in W^{1,\infty}(\Omega)$ . We say that  $\{u, \varphi\}$  is a weak solution to (1.1) if

$$(2.2) \quad \begin{aligned} \int_{\Omega} \nabla u \cdot \nabla v \, dx + \int_{\partial\Omega} \beta uv \, ds &= \int_{\Omega} \sigma(u) |\nabla \varphi|^2 v \, dx \quad \forall v \in H^1(\Omega), \\ \int_{\Omega} \sigma(u) \nabla \varphi \cdot \nabla w \, dx &= 0, \quad \varphi_0 - \varphi \in H_0^1(\Omega) \quad \forall w \in H_0^1(\Omega). \end{aligned}$$

The quadratic term on the right-hand side of the first equation in (2.2) creates a difficulty because  $|\nabla \varphi|^2$  is only known to belong to  $L^1(\Omega)$ . Therefore, for an arbitrary  $v \in C^1(\bar{\Omega})$ , take  $w = (\varphi - \varphi_0)v \in H_0^1(\Omega)$  as a test function in the second equation of (2.2). Then

$$\int_{\Omega} \sigma(u) \{ |\nabla \varphi|^2 v - v \nabla \varphi \cdot \nabla \varphi_0 + (\varphi - \varphi_0) \nabla \varphi \cdot \nabla v \} \, dx = 0.$$

By the density argument, the weak formulation of (1.1) is as follows: Find  $u \in H^1(\Omega)$  and  $\varphi \in H^1(\Omega)$  such that

$$(2.3) \quad \begin{aligned} \int_{\Omega} \nabla u \cdot \nabla v \, dx + \int_{\partial\Omega} \beta uv \, ds &= \int_{\Omega} (\varphi_0 - \varphi) \sigma(u) \nabla \varphi \cdot \nabla v \, dx + \int_{\Omega} (\sigma(u) \nabla \varphi \cdot \nabla \varphi_0) v \, dx \\ \int_{\Omega} \sigma(u) \nabla \varphi \cdot \nabla w \, dx &= 0, \quad \varphi_0 - \varphi \in H_0^1(\Omega) \quad \forall w \in H_0^1(\Omega), \forall v \in H^1(\Omega). \end{aligned}$$

Now the terms on the right-hand side of the first equation in (2.3) make sense. Indeed, by Theorem 2.1, there exists  $r > 2$  such that  $\varphi \in W^{1,r}(\Omega)$ . Therefore, since  $\nabla\varphi \in L^r(\Omega)$  and  $\nabla\varphi_0 \in L^2(\Omega)$ , it follows that there exists  $s'$  such that  $\nabla\varphi \cdot \nabla\varphi_0 \in L^{s'}(\Omega)$  and

$$(2.4) \quad \frac{1}{s'} = \frac{1}{2} + \frac{1}{r}.$$

Since  $\Omega \subset \mathbb{R}^2$  it follows that  $v \in H^1(\Omega) \subset L^s(\Omega)$  for  $s \in [1, \infty)$ . This implies that the integral

$$\int_{\Omega} (\sigma(u) \nabla\varphi \cdot \nabla\varphi_0) v \, dx$$

in (2.3) makes sense, since  $\sigma(u)$  is bounded and  $(\sigma(u) \nabla\varphi \cdot \nabla\varphi_0) \in L^{s'}(\Omega)$ , and  $s$  can be chosen such that

$$(2.5) \quad \frac{1}{s'} + \frac{1}{s} = 1.$$

If we denote  $\tilde{M} := \sup_{\partial\Omega} |\varphi_0|$ , then by the weak maximum principle (see formula (16) in [9])

$$(2.6) \quad \sup_{\Omega} |\varphi| \leq \tilde{M}.$$

Therefore the integral  $\int_{\Omega} (\varphi_0 - \varphi) \sigma(u) \nabla\varphi \cdot \nabla v \, dx$  in (2.3) also makes sense. Thus, any solution of (2.2) is a solution to (2.3), and vice versa.

Note that  $r > 2$  is chosen from the Meyers estimate; then  $s'$  and  $s$  are determined and satisfy  $s' < 2 < s$ .

Existence of a solution to (2.3) was proven by Howison, Rodrigues, and Shillor in [9] using Schauder's fixed point theorem. It was also shown in [9] that the solution is unique provided assumption 5 is satisfied. Given  $\beta \in U_M$ , we denote the solution of (2.3) by  $u(\beta), \varphi(\beta)$ , under assumption 5 giving uniqueness.

Here we proceed to the derivation of a priori estimates. Note that as a result of the maximum principle solutions to (1.1) satisfy  $u \geq 0$  on  $\bar{\Omega}$ .

**THEOREM 2.2.** *Let assumptions 1 through 4 hold and let  $\beta \in U_M$  be given. Then  $u$  and  $\varphi$  solving (2.3) satisfy*

$$(2.7) \quad \begin{aligned} \|\varphi\|_{W^{1,r}(\Omega)} &\leq \Phi \text{ for some } r > 2, \\ \|u\|_{H^1(\Omega)} &\leq \tilde{C}, \end{aligned}$$

where  $\Phi$  and  $\tilde{C}$  are some positive constants.

*Proof.* We show the estimate for  $\varphi$  first. Because of assumption 2 we treat  $\sigma(u)$  in (1.1) as a bounded coefficient. Consider the following Dirichlet problem with zero boundary data:

$$(2.8) \quad \begin{aligned} -\nabla \cdot (\sigma(u) \nabla \tilde{\varphi}) &= \nabla \cdot (\sigma(u) \nabla \varphi_0) \text{ in } \Omega, \\ \tilde{\varphi} &= 0 \text{ on } \partial\Omega. \end{aligned}$$

Since the right-hand side of (2.8) is in  $H^{-1}(\Omega)$ , by the standard theory for elliptic equations in divergence form [8], there exists  $\tilde{\varphi} = (\varphi - \varphi_0) \in H_0^1(\Omega)$  that solves (2.8).

By Theorem 2.1, there exists  $r > 2$  such that  $\tilde{\varphi} \in W_0^{1,r}(\Omega)$  and

$$\|\tilde{\varphi}\|_{W_0^{1,r}(\Omega)} \leq C \|\nabla \cdot (\sigma(u) \nabla \varphi_0)\|_{W^{-1,r}(\Omega)} \leq C \|\sigma(u) \nabla \varphi_0\|_r \leq CC_2 \|\nabla \varphi_0\|_r.$$

Since

$$\|\varphi\|_{W^{1,r}(\Omega)} \leq \|\varphi - \varphi_0\|_{W^{1,r}(\Omega)} + \|\varphi_0\|_{W^{1,r}(\Omega)}$$

and

$$\|\varphi - \varphi_0\|_{W^{1,r}(\Omega)} \equiv \|\tilde{\varphi}\|_{W^{1,r}(\Omega)} \leq C' \|\tilde{\varphi}\|_{W_0^{1,r}(\Omega)} \leq C' CC_2 \|\nabla \varphi_0\|_r,$$

we obtain

$$\|\varphi\|_{W^{1,r}(\Omega)} \leq C' CC_2 \|\nabla \varphi_0\|_r + \|\varphi_0\|_{W^{1,r}(\Omega)}.$$

We also have

$$\begin{aligned} \|\nabla \varphi_0\|_r &\leq C_3 \|\nabla \varphi_0\|_\infty, \\ \|\varphi_0\|_{W^{1,r}(\Omega)} &\leq C_4 \|\varphi_0\|_{W^{1,\infty}(\Omega)}, \end{aligned}$$

whence

$$(2.9) \quad \|\varphi\|_{W^{1,r}(\Omega)} \leq \Phi \text{ for some } r > 2,$$

where

$$(2.10) \quad \Phi \stackrel{\text{def}}{=} C' CC_2 C_3 \|\nabla \varphi_0\|_\infty + C_4 \|\varphi_0\|_{W^{1,\infty}(\Omega)}.$$

We show that  $\|u\|_{H^1(\Omega)} \leq \tilde{C}$ . From the weak formulation (2.3) we get

$$\int_\Omega |\nabla u|^2 dx + \int_{\partial\Omega} \beta u^2 ds = \int_\Omega (\varphi_0 - \varphi) \sigma(u) \nabla \varphi \cdot \nabla u dx + \int_\Omega \sigma(u) u \nabla \varphi \cdot \nabla \varphi_0 dx.$$

Taking into account that  $\beta(x) \geq \lambda > 0$  we have

$$\begin{aligned} \int_\Omega |\nabla u|^2 dx + \lambda \int_{\partial\Omega} u^2 ds &\leq \int_\Omega (\varphi_0 - \varphi) \sigma(u) \nabla \varphi \cdot \nabla u dx + \int_\Omega \sigma(u) u \nabla \varphi \cdot \nabla \varphi_0 dx \\ &\leq 2C_2 \tilde{M} \|\nabla \varphi\|_2 \cdot \|\nabla u\|_2 + C_2 \|\nabla \varphi_0\|_\infty \|u\|_2 \cdot \|\nabla \varphi\|_2 \\ &\leq 2C_2 \tilde{M} C_5 \Phi \|\nabla u\|_2 + C_2 \|\nabla \varphi_0\|_\infty C_5 \Phi \|u\|_2 \\ &\leq \tilde{C}_1 \|u\|_{H^1(\Omega)}, \end{aligned}$$

where we used (2.6),  $\|\nabla \varphi\|_2 \leq C_5 \|\nabla \varphi\|_r$ , and  $\tilde{C}_1 \equiv \max(2C_2 \tilde{M} C_5 \Phi, C_2 \|\nabla \varphi_0\|_\infty C_5 \Phi)$ . It can be shown (for example, see [2, 20]) that the quantity

$$\|v\|_*^2 \stackrel{\text{def}}{=} \int_\Omega |\nabla v|^2 dx + \lambda \int_{\partial\Omega} v^2 ds$$

defines a norm on  $H^1(\Omega)$  which is equivalent to the  $\|\cdot\|_{H^1(\Omega)}$  norm; then for some  $k > 0$

$$(2.11) \quad k \|u\|_{H^1(\Omega)}^2 \leq \|u\|_*^2 \leq \tilde{C}_1 \|u\|_{H^1(\Omega)},$$

whence  $\|u\|_{H^1(\Omega)} \leq \tilde{C}_1/k$ .  $\square$

**3. Existence of an optimal control.** Having obtained a priori estimates we proceed to the proof of existence of an optimal control. For the remainder of the paper we assume that assumptions 1 through 5 are satisfied.

**THEOREM 3.1.** *There exists a solution to the optimal control problem (1.2).*

*Proof.* Using the estimate from Theorem 2.2, we can choose a minimizing sequence  $\{\beta_n\}_{n=1}^\infty \subset U_M$  such that

$$\lim_{n \rightarrow \infty} J(\beta_n) = \inf_{\beta \in U_M} J(\beta).$$

Let  $u_n = u(\beta_n)$  and  $\varphi_n = \varphi(\beta_n)$  be the corresponding solutions to (2.3). By Theorem 2.2 we have  $\|u_n\|_{H^1(\Omega)} \leq C$ ,  $\|\varphi_n\|_{W^{1,r}(\Omega)} \leq C$  for all  $n$ , where  $C > 0$  denotes a constant independent of  $n$ . Therefore, there exist  $u^* \in H^1(\Omega)$  and  $\varphi^* \in W^{1,r}(\Omega)$  such that on a subsequence

$$u_n \xrightarrow{w} u^* \text{ in } H^1(\Omega) \quad \text{and} \quad \varphi_n \xrightarrow{w} \varphi^* \text{ in } W^{1,r}(\Omega).$$

Since  $r > 2$  and  $N = 2$ , we have  $W^{1,r}(\Omega) \subset \subset C(\bar{\Omega})$  and  $W^{1,r}(\Omega) \subset H^1(\Omega) \subset \subset L^s(\Omega)$ . Hence, there exists  $\beta^* \in U_M$  such that on a subsequence

$$(3.1) \quad \begin{aligned} \varphi_n &\xrightarrow{s} \varphi^* \quad \text{in } C(\bar{\Omega}), & \nabla \varphi_n &\xrightarrow{w} \nabla \varphi^* \quad \text{in } L^r(\Omega), \\ u_n &\xrightarrow{s} u^* \quad \text{in } L^s(\Omega), & \nabla u_n &\xrightarrow{w} \nabla u^* \quad \text{in } L^2(\Omega), \\ \beta_n &\xrightarrow{w} \beta^* \quad \text{in } L^2(\partial\Omega), & \beta_n &\xrightarrow{w^*} \beta^* \quad \text{in } L^\infty(\partial\Omega), \\ u_n &\xrightarrow{s} u^* \quad \text{in } L^2(\partial\Omega). \end{aligned}$$

Next, we want to show that  $u^* = u(\beta^*)$  and  $\varphi^* = \varphi(\beta^*)$  solve (2.3) with control  $\beta^*$ . We show that for  $v \in H^1(\Omega)$

$$(3.2) \quad \int_{\partial\Omega} \beta_n u_n v \, ds \rightarrow \int_{\partial\Omega} \beta^* u^* v \, ds \text{ as } n \rightarrow \infty.$$

Indeed, by the trace inequality  $u^* \in H^1(\Omega)$  implies  $u^* \in L^2(\partial\Omega)$ , and we obtain

$$\begin{aligned} \left| \int_{\partial\Omega} \beta_n u_n v \, ds - \int_{\partial\Omega} \beta^* u^* v \, ds \right| &\leq \int_{\partial\Omega} |\beta_n u_n v - \beta_n u^* v| \, ds + \left| \int_{\partial\Omega} \beta_n u^* v - \beta^* u^* v \, ds \right| \\ &\leq M \int_{\partial\Omega} |u_n - u^*| \cdot |v| \, ds + \left| \int_{\partial\Omega} (\beta_n - \beta^*) u^* v \, ds \right| \\ &\leq M \|u_n - u^*\|_{L^2(\partial\Omega)} \cdot \|v\|_{L^2(\partial\Omega)} \\ &\quad + \left| \int_{\partial\Omega} (\beta_n - \beta^*) u^* v \, ds \right| \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

This proves (3.2). Next, we show

$$(3.3) \quad \int_{\Omega} \sigma(u_n) v \nabla \varphi_n \cdot \nabla \varphi_0 \, dx \rightarrow \int_{\Omega} \sigma(u^*) v \nabla \varphi^* \cdot \nabla \varphi_0 \, dx \text{ as } n \rightarrow \infty.$$

We have

$$\begin{aligned} &\left| \int_{\Omega} \sigma(u_n) v \nabla \varphi_n \cdot \nabla \varphi_0 \, dx - \int_{\Omega} \sigma(u^*) v \nabla \varphi^* \cdot \nabla \varphi_0 \, dx \right| \\ &\leq K \|\nabla \varphi_0\|_\infty \int_{\Omega} |u_n - u^*| \cdot |v| \cdot |\nabla \varphi_n| \, dx + \left| \int_{\Omega} \sigma(u^*) v (\nabla \varphi_n - \nabla \varphi^*) \cdot \nabla \varphi_0 \, dx \right| \\ &\leq K \|\nabla \varphi_0\|_\infty \|u_n - u^*\|_s \cdot \|v\|_2 \cdot \|\nabla \varphi_n\|_r + \left| \int_{\Omega} \sigma(u^*) v (\nabla \varphi_n - \nabla \varphi^*) \cdot \nabla \varphi_0 \, dx \right| \rightarrow 0 \end{aligned}$$

since  $u_n \xrightarrow{s} u^*$  in  $L^s(\Omega)$ ,  $\nabla \varphi \xrightarrow{w} \nabla \varphi^*$  in  $L^2(\Omega)$ , and  $\|\nabla \varphi_n\|_r \leq C$ . This completes the proof of (3.3). Now we show

$$(3.4) \quad \int_{\Omega} (\varphi_0 - \varphi_n) \sigma(u_n) \nabla \varphi_n \cdot \nabla v \, dx \rightarrow \int_{\Omega} (\varphi_0 - \varphi^*) \sigma(u^*) \nabla \varphi^* \cdot \nabla v \, dx \text{ as } n \rightarrow \infty.$$

Indeed, we can write

$$\begin{aligned} & \left| \int_{\Omega} (\varphi_0 - \varphi_n) \sigma(u_n) \nabla \varphi_n \cdot \nabla v \, dx - \int_{\Omega} (\varphi_0 - \varphi^*) \sigma(u^*) \nabla \varphi^* \cdot \nabla v \, dx \right| \\ & \leq \left| \int_{\Omega} [\sigma(u_n) (\varphi_0 - \varphi_n) \nabla \varphi_n \cdot \nabla v - \sigma(u^*) (\varphi_0 - \varphi_n) \nabla \varphi^* \cdot \nabla v] \, dx \right| \\ & \quad + \left| \int_{\Omega} [\sigma(u^*) (\varphi_0 - \varphi_n) \nabla \varphi^* \cdot \nabla v - \sigma(u^*) (\varphi_0 - \varphi^*) \nabla \varphi^* \cdot \nabla v] \, dx \right| \stackrel{\text{def}}{=} A + B. \end{aligned}$$

We deal with the term  $B$  first:

$$\begin{aligned} B &= \left| \int_{\Omega} \sigma(u^*) \nabla \varphi^* \cdot \nabla v [(\varphi_0 - \varphi_n) - (\varphi_0 - \varphi^*)] \, dx \right| \\ &= \left| \int_{\Omega} \sigma(u^*) \nabla \varphi^* \cdot \nabla v (\varphi^* - \varphi_n) \, dx \right| \\ &\leq C_2 \|\varphi^* - \varphi_n\|_{C(\bar{\Omega})} \cdot \|\nabla v\|_2 \cdot \|\nabla \varphi^*\|_2 \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

Now we look at the term  $A$ :

$$\begin{aligned} A &\leq \left| \int_{\Omega} (\varphi_0 - \varphi_n) (\sigma(u_n) - \sigma(u^*)) \nabla \varphi_n \cdot \nabla v \, dx \right| \\ &\quad + \left| \int_{\Omega} (\varphi_0 - \varphi_n) \sigma(u^*) (\nabla \varphi_n - \nabla \varphi^*) \cdot \nabla v \, dx \right| \stackrel{\text{def}}{=} A_1 + A_2, \end{aligned}$$

where

$$A_1 \leq 2\tilde{M}K \|u_n - u^*\|_s \cdot \|\nabla \varphi_n\|_r \cdot \|\nabla v\|_2 \rightarrow 0 \text{ as } n \rightarrow \infty.$$

For  $A_2$  we get

$$\begin{aligned} A_2 &\leq \left| \int_{\Omega} (\varphi_0 - \varphi^*) \sigma(u^*) (\nabla \varphi_n - \nabla \varphi^*) \cdot \nabla v \, dx \right| \\ &\quad + \left| \int_{\Omega} (\varphi^* - \varphi_n) \sigma(u^*) (\nabla \varphi_n - \nabla \varphi^*) \cdot \nabla v \, dx \right| \stackrel{\text{def}}{=} A_{21} + A_{22}. \end{aligned}$$

It is easy to see that  $A_{21} \rightarrow 0$  as  $n \rightarrow \infty$ . Similarly,

$$A_{22} \leq C_2 \|\varphi^* - \varphi_n\|_{C(\bar{\Omega})} \cdot (\|\nabla \varphi_n\|_2 + \|\nabla \varphi^*\|_2) \cdot \|\nabla v\|_2 \rightarrow 0 \text{ as } n \rightarrow \infty.$$

We can also obtain

$$\int_{\Omega} \sigma(u_n) \nabla \varphi_n \cdot \nabla w \, dx \rightarrow \int_{\Omega} \sigma(u^*) \nabla \varphi^* \cdot \nabla w \, dx \text{ as } n \rightarrow \infty.$$

Therefore  $(u^*, \varphi^*)$  is the weak solution of (2.3) with  $\beta = \beta^*$ :  $u^* = u(\beta^*)$  and  $\varphi^* = \varphi(\beta^*)$ . Using weak lower semicontinuity of  $J(\beta)$  with respect to the  $L^2$  norm, it is easy to show that the infimum is achieved at  $\beta^*$ .  $\square$



**4. Derivation of the optimality system.** In order to characterize an optimal control, we need to derive an optimality system which consists of the original state system coupled with an adjoint system. To obtain the necessary conditions for the optimality system, we differentiate the objective functional with respect to the control. Since the objective functional depends on  $u$ , which is coupled to  $\varphi$  through the PDE, we will need to differentiate  $u$  and  $\varphi$  with respect to control  $\beta$ . Note that to obtain this result,  $\|\varphi_0\|_{W^{1,\infty}}$  may need to be smaller than in assumption 5.

**THEOREM 4.1** (sensitivities). *There exists  $\Lambda_1 > 0$  such that  $\|\varphi_0\|_{W^{1,\infty}} \leq \Lambda_1$ , and the mapping  $\beta \mapsto (u, \varphi)$  is differentiable in the following sense:*

$$(4.1) \quad \begin{aligned} \frac{u(\beta + \varepsilon \ell) - u(\beta)}{\varepsilon} &\xrightarrow{w} \psi_1 \text{ in } H^1(\Omega), \\ \frac{\varphi(\beta + \varepsilon \ell) - \varphi(\beta)}{\varepsilon} &\xrightarrow{w} \psi_2 \text{ in } H_0^1(\Omega) \text{ as } \varepsilon \rightarrow 0 \end{aligned}$$

for any  $\beta, \ell \in U_M$  such that  $(\beta + \varepsilon \ell) \in U_M$  for small  $\varepsilon$ . Moreover, the sensitivities,  $\psi_1 \in H^1(\Omega)$  and  $\psi_2 \in H_0^1(\Omega)$ , satisfy

$$(4.2) \quad \begin{aligned} \Delta \psi_1 + \sigma'(u) |\nabla \varphi|^2 \psi_1 + 2\sigma(u) \nabla \varphi \cdot \nabla \psi_2 &= 0 \text{ in } \Omega, \\ \nabla \cdot [\sigma'(u) \psi_1 \nabla \varphi + \sigma(u) \nabla \psi_2] &= 0 \text{ in } \Omega, \\ \frac{\partial \psi_1}{\partial n} + \beta \psi_1 + \ell u &= 0 \text{ on } \partial\Omega, \\ \psi_2 &= 0 \text{ on } \partial\Omega. \end{aligned}$$

*Proof.* Recall that we denoted  $u = u(\beta)$  and  $\varphi = \varphi(\beta)$ . Now we also denote  $u^\varepsilon = u(\beta^\varepsilon)$ ,  $\varphi^\varepsilon = \varphi(\beta^\varepsilon)$ , where  $\beta^\varepsilon := \beta + \varepsilon \ell$ . We present the proof in three steps.

*Step 1.* First, we derive an estimate for  $(\varphi^\varepsilon - \varphi)/\varepsilon$  in terms of  $(u^\varepsilon - u)/\varepsilon$ . The quotients  $(u^\varepsilon - u)/\varepsilon$  and  $(\varphi^\varepsilon - \varphi)/\varepsilon$  satisfy

$$(4.3) \quad \begin{aligned} &\int_{\Omega} \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \cdot \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) dx + \int_{\partial\Omega} \beta \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \left( \frac{u^\varepsilon - u}{\varepsilon} \right) ds \\ &= - \int_{\partial\Omega} \ell u^\varepsilon \left( \frac{u^\varepsilon - u}{\varepsilon} \right) ds \\ &\quad + \frac{1}{\varepsilon} \int_{\Omega} [(\varphi_0 - \varphi^\varepsilon) \sigma(u^\varepsilon) \nabla \varphi^\varepsilon - (\varphi_0 - \varphi) \sigma(u) \nabla \varphi] \cdot \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) dx \\ &\quad + \frac{1}{\varepsilon} \int_{\Omega} [\sigma(u^\varepsilon) \nabla \varphi^\varepsilon - \sigma(u) \nabla \varphi] \cdot \nabla \varphi_0 \left( \frac{u^\varepsilon - u}{\varepsilon} \right) dx, \\ &\quad \frac{1}{\varepsilon} \int_{\Omega} \left[ \sigma(u^\varepsilon) \nabla \varphi^\varepsilon \cdot \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) - \sigma(u) \nabla \varphi \cdot \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right] dx = 0. \end{aligned}$$

Since  $(\varphi^\varepsilon - \varphi)/\varepsilon \in H_0^1(\Omega)$  it follows from Poincaré's inequality that it is sufficient to have a bound on  $\|\nabla(\varphi^\varepsilon - \varphi)/\varepsilon\|_2$ . The second equation in (4.3) implies

$$(4.4) \quad \int_{\Omega} \sigma(u^\varepsilon) \nabla \left( \frac{\varphi^\varepsilon}{\varepsilon} \right) \cdot \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) dx = \int_{\Omega} \sigma(u) \nabla \left( \frac{\varphi}{\varepsilon} \right) \cdot \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) dx.$$

Taking into account (4.4) we can write

$$(4.5) \quad \int_{\Omega} \sigma(u) \left| \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right|^2 dx = \int_{\Omega} \left( \sigma(u) - \sigma(u^\varepsilon) \right) \nabla \left( \frac{\varphi}{\varepsilon} \right) \cdot \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) dx.$$

Using (4.4) and (4.5) we obtain

$$\begin{aligned}
C_1 \left\| \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right\|_2^2 &\leq \int_{\Omega} \sigma(u) \left| \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right|^2 dx \\
&= \int_{\Omega} \left( \sigma(u) - \sigma(u^\varepsilon) \right) \nabla \left( \frac{\varphi^\varepsilon}{\varepsilon} \right) \cdot \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) dx \\
&\leq K \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_s \cdot \|\nabla \varphi^\varepsilon\|_r \cdot \left\| \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right\|_2,
\end{aligned}$$

whence

$$(4.6) \quad \left\| \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right\|_2 \leq \frac{K}{C_1} \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_{H^1(\Omega)} \cdot \|\nabla \varphi^\varepsilon\|_r.$$

*Step 2.* Now we outline the derivation of an  $H^1$  norm estimate for  $(u^\varepsilon - u)/\varepsilon$ . From the first equation in (4.3) we obtain

$$\begin{aligned}
&\int_{\Omega} \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right|^2 dx + \lambda \int_{\partial\Omega} \left( \frac{u^\varepsilon - u}{\varepsilon} \right)^2 ds \leq \left| - \int_{\partial\Omega} \ell u^\varepsilon \left( \frac{u^\varepsilon - u}{\varepsilon} \right) ds \right. \\
&+ \frac{1}{\varepsilon} \int_{\Omega} \left[ (\varphi_0 - \varphi^\varepsilon) \sigma(u^\varepsilon) \nabla \varphi^\varepsilon - (\varphi_0 - \varphi) \sigma(u) \nabla \varphi \right] \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) dx \\
&+ \left. \frac{1}{\varepsilon} \int_{\Omega} \left[ \sigma(u^\varepsilon) \nabla \varphi^\varepsilon - \sigma(u) \nabla \varphi \right] \nabla \varphi_0 \left( \frac{u^\varepsilon - u}{\varepsilon} \right) dx \right| \\
&\stackrel{\text{def}}{=} \left| - \int_{\partial\Omega} \ell u^\varepsilon \left( \frac{u^\varepsilon - u}{\varepsilon} \right) ds + \mathcal{C} + \mathcal{D} \right|.
\end{aligned}$$

We have

$$\begin{aligned}
|\mathcal{C}| &\leq \int_{\Omega} |\varphi_0 - \varphi| \cdot \left| \frac{\sigma(u^\varepsilon) \nabla \varphi^\varepsilon - \sigma(u) \nabla \varphi}{\varepsilon} \right| \cdot \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx \\
&+ \int_{\Omega} \left| \frac{\varphi - \varphi^\varepsilon}{\varepsilon} \right| \sigma(u^\varepsilon) |\nabla \varphi^\varepsilon| \cdot \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx \\
&\leq 2\tilde{M} \int_{\Omega} \left| \frac{\sigma(u^\varepsilon) - \sigma(u)}{\varepsilon} \right| \cdot |\nabla \varphi^\varepsilon| \cdot \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx \\
&+ 2\tilde{M} \int_{\Omega} \sigma(u) \cdot \left| \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right| \cdot \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx \\
&+ C_2 \int_{\Omega} \left| \frac{\varphi - \varphi^\varepsilon}{\varepsilon} \right| \cdot |\nabla \varphi^\varepsilon| \cdot \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx \\
&\leq 2\tilde{M}K \int_{\Omega} \left| \frac{u^\varepsilon - u}{\varepsilon} \right| \cdot |\nabla \varphi^\varepsilon| \cdot \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx \\
&+ 2\tilde{M}C_2 \int_{\Omega} \left| \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right| \cdot \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx \\
&+ C_2 \int_{\Omega} \left| \frac{\varphi - \varphi^\varepsilon}{\varepsilon} \right| \cdot |\nabla \varphi^\varepsilon| \cdot \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx.
\end{aligned}$$

Similarly, for  $\mathcal{D}$  we obtain

$$\begin{aligned} |\mathcal{D}| &\leq K \|\nabla \varphi_0\|_\infty \int_\Omega \left| \frac{u^\varepsilon - u}{\varepsilon} \right| \cdot |\nabla \varphi^\varepsilon| \cdot \left| \frac{u^\varepsilon - u}{\varepsilon} \right| dx \\ &\quad + C_2 \|\nabla \varphi_0\|_\infty \int_\Omega \left| \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right| \cdot \left| \frac{u^\varepsilon - u}{\varepsilon} \right| dx. \end{aligned}$$

Taking into account the expressions for  $|\mathcal{C}|$ ,  $|\mathcal{D}|$ , and the trace inequality  $\|u^\varepsilon\|_{L^2(\partial\Omega)} \leq M_2 \|u^\varepsilon\|_{H^1(\Omega)}$  we get

$$\begin{aligned} &\int_\Omega \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right|^2 dx + \lambda \int_{\partial\Omega} \left( \frac{u^\varepsilon - u}{\varepsilon} \right)^2 ds \\ &\leq M M_2^2 \|u^\varepsilon\|_{H^1(\Omega)} \cdot \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_{H^1(\Omega)} + 2\tilde{M}K \int_\Omega \left| \frac{u^\varepsilon - u}{\varepsilon} \right| \cdot |\nabla \varphi^\varepsilon| \cdot \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx \\ &\quad + 2\tilde{M}C_2 \int_\Omega \left| \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right| \cdot \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx + C_2 \int_\Omega \left| \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right| |\nabla \varphi^\varepsilon| \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right| dx \\ &\quad + K \|\nabla \varphi_0\|_\infty \int_\Omega \left| \frac{u^\varepsilon - u}{\varepsilon} \right| |\nabla \varphi^\varepsilon| \left| \frac{u^\varepsilon - u}{\varepsilon} \right| dx + C_2 \|\nabla \varphi_0\|_\infty \int_\Omega \left| \nabla \left( \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right) \right| \left| \frac{u^\varepsilon - u}{\varepsilon} \right| dx. \end{aligned}$$

For the rest of this paper, when dealing with constants, we only keep track of explicit dependence on  $\Phi$  and otherwise use generic constants. Using Hölder's inequality and corresponding a priori bounds we have

$$\begin{aligned} \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_*^2 &\equiv \int_\Omega \left| \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \right|^2 dx + \lambda \int_{\partial\Omega} \left( \frac{u^\varepsilon - u}{\varepsilon} \right)^2 ds \leq C_8 \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_{H^1(\Omega)} \\ &\quad + \left( C_9 \Phi + C_{10} \Phi^2 \right) \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_{H^1(\Omega)}^2. \end{aligned}$$

Due to the equivalence of norms expressed by (2.11) we get

$$k \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_{H^1(\Omega)}^2 \leq \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_*^2 \leq C_8 \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_{H^1(\Omega)} + \left( C_9 \Phi + C_{10} \Phi^2 \right) \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_{H^1(\Omega)}^2.$$

Recall that by definition,  $\Phi$  includes  $\|\nabla \varphi_0\|_\infty$  and  $\|\varphi_0\|_{W^{1,\infty}(\Omega)}$  (see (2.10)). Therefore, if  $\|\varphi_0\|_{W^{1,\infty}(\Omega)}$  is chosen sufficiently small such that

$$k_1 \stackrel{\text{def}}{=} k - \left( C_9 \Phi + C_{10} \Phi^2 \right) > 0,$$

then

$$(4.7) \quad \left\| \frac{u^\varepsilon - u}{\varepsilon} \right\|_{H^1(\Omega)} \leq \frac{C_8}{k_1},$$

where the constant in (4.7) does not depend on  $\varepsilon$ . Consequently,

$$(4.8) \quad \left\| \frac{\varphi^\varepsilon - \varphi}{\varepsilon} \right\|_{H^1(\Omega)} \leq C_{11}.$$

*Step 3.* Now we are ready to derive the sensitivity system (4.2). The above

estimates imply that for a subsequence of  $\varepsilon \rightarrow 0$ , there exist  $\psi_1$  and  $\psi_2$  such that

$$\begin{aligned} \frac{u^\varepsilon - u}{\varepsilon} &\xrightarrow{w} \psi_1 \quad \text{in } H^1(\Omega), & \frac{u^\varepsilon - u}{\varepsilon} &\xrightarrow{s} \psi_1 \quad \text{in } L^s(\Omega), \\ \frac{\varphi^\varepsilon - \varphi}{\varepsilon} &\xrightarrow{w} \psi_2 \quad \text{in } H^1(\Omega), & \frac{\varphi^\varepsilon - \varphi}{\varepsilon} &\xrightarrow{s} \psi_2 \quad \text{in } L^s(\Omega), \\ u^\varepsilon &\xrightarrow{s} u \quad \text{in } L^s(\Omega), & \varphi^\varepsilon &\xrightarrow{s} \varphi \quad \text{in } C(\bar{\Omega}), \\ \nabla \varphi^\varepsilon &\xrightarrow{w} \nabla \varphi \quad \text{in } L^r(\Omega), & \frac{u^\varepsilon - u}{\varepsilon} &\xrightarrow{s} \psi_1 \quad \text{in } L^2(\partial\Omega), \\ \beta^\varepsilon &\xrightarrow{w} \beta \quad \text{in } L^2(\partial\Omega), & \beta^\varepsilon &\xrightarrow{w^*} \beta \quad \text{in } L^\infty(\partial\Omega) \text{ as } \varepsilon \rightarrow 0. \end{aligned}$$

Using these convergences we show that the sensitivities satisfy the system (4.2). After subtracting the corresponding weak formulations and dividing by  $\varepsilon$ , we obtain

$$\begin{aligned} (4.9) \quad & \int_{\Omega} \nabla \left( \frac{u^\varepsilon - u}{\varepsilon} \right) \cdot \nabla v \, dx + \int_{\partial\Omega} \beta \left( \frac{u^\varepsilon - u}{\varepsilon} \right) v \, ds + \int_{\partial\Omega} \ell u^\varepsilon \, ds \\ &= \frac{1}{\varepsilon} \int_{\Omega} \left[ (\varphi_0 - \varphi^\varepsilon) \sigma(u^\varepsilon) \nabla \varphi^\varepsilon \cdot \nabla v - (\varphi_0 - \varphi) \sigma(u) \nabla \varphi \cdot \nabla v \right] dx \\ &+ \frac{1}{\varepsilon} \int_{\Omega} \left[ \sigma(u^\varepsilon) \nabla \varphi^\varepsilon - \sigma(u) \nabla \varphi \right] \cdot \nabla \varphi_0 v \, dx \quad \forall v \in H^1(\Omega), \end{aligned}$$

$$(4.10) \quad \frac{1}{\varepsilon} \int_{\Omega} \left[ \sigma(u^\varepsilon) \nabla \varphi^\varepsilon \cdot \nabla w - \sigma(u) \nabla \varphi \cdot \nabla w \right] dx = 0 \quad \forall w \in H_0^1(\Omega).$$

The terms on the left-hand side of (4.9) can be shown to converge by weak convergence. We write the first term on the right-hand side of (4.9) in the form

$$\begin{aligned} & \frac{1}{\varepsilon} \int_{\Omega} \left[ (\varphi_0 - \varphi^\varepsilon) \sigma(u^\varepsilon) \nabla \varphi^\varepsilon \cdot \nabla v - (\varphi_0 - \varphi) \sigma(u) \nabla \varphi \cdot \nabla v \right] dx \\ &= \frac{1}{\varepsilon} \int_{\Omega} (\varphi_0 - \varphi^\varepsilon) \left[ \sigma(u^\varepsilon) \nabla \varphi^\varepsilon - \sigma(u) \nabla \varphi \right] \cdot \nabla v \, dx + \frac{1}{\varepsilon} \int_{\Omega} (\varphi - \varphi^\varepsilon) \sigma(u) \nabla \varphi \cdot \nabla v \, dx \\ &= \mathcal{G} + \mathcal{F}. \end{aligned}$$

We illustrate the type of estimates by considering terms coming from  $\mathcal{G}$ . We write

$$\begin{aligned} \mathcal{G} &= \frac{1}{\varepsilon} \int_{\Omega} (\varphi_0 - \varphi^\varepsilon) \left[ \sigma(u^\varepsilon) \nabla \varphi^\varepsilon - \sigma(u) \nabla \varphi \right] \cdot \nabla v \, dx \\ &= \frac{1}{\varepsilon} \int_{\Omega} (\varphi_0 - \varphi^\varepsilon) \left[ \sigma(u^\varepsilon) - \sigma(u) \right] \nabla \varphi^\varepsilon \cdot \nabla v \, dx \\ &+ \frac{1}{\varepsilon} \int_{\Omega} (\varphi_0 - \varphi^\varepsilon) \sigma(u) \left[ \nabla \varphi^\varepsilon - \nabla \varphi \right] \cdot \nabla v \, dx = \mathcal{G}_1 + \mathcal{G}_2. \end{aligned}$$

One can show

$$(4.11) \quad \mathcal{G}_2 \rightarrow \int_{\Omega} (\varphi_0 - \varphi) \sigma(u) \nabla \psi_2 \cdot \nabla v \, dx \text{ as } \varepsilon \rightarrow 0,$$

and we illustrate terms from  $\mathcal{G}_1$ . Namely, we show the convergence for  $\mathcal{G}_1$ :

$$\begin{aligned}
& \left| \int_{\Omega} (\varphi_0 - \varphi^\varepsilon) \left( \frac{\sigma(u^\varepsilon) - \sigma(u)}{\varepsilon} \right) \nabla \varphi^\varepsilon \cdot \nabla v \, dx - \int_{\Omega} (\varphi_0 - \varphi) \sigma'(u) \psi_1 \nabla \varphi \cdot \nabla v \, dx \right| \\
& \leq \left| \int_{\Omega} (\varphi_0 - \varphi) \left\{ \left( \frac{\sigma(u^\varepsilon) - \sigma(u)}{\varepsilon} \right) \nabla \varphi^\varepsilon - \sigma'(u) \psi_1 \nabla \varphi \right\} \cdot \nabla v \, dx \right| \\
& + \left| \int_{\Omega} (\varphi - \varphi^\varepsilon) \left( \frac{\sigma(u^\varepsilon) - \sigma(u)}{\varepsilon} - \sigma'(u) \psi_1 + \sigma'(u) \psi_1 \right) \nabla \varphi^\varepsilon \cdot \nabla v \, dx \right| \\
& \leq \left| \int_{\Omega} \left\{ (\varphi_0 - \varphi) \left[ \frac{\sigma(u^\varepsilon) - \sigma(u)}{\varepsilon} - \sigma'(u) \psi_1 \right] \nabla \varphi^\varepsilon \cdot \nabla v \right. \right. \\
& + \left. \left. (\varphi_0 - \varphi) \sigma'(u) \psi_1 \nabla (\varphi^\varepsilon - \varphi) \cdot \nabla v \right\} dx \right| \\
& + \left| \int_{\Omega} (\varphi - \varphi^\varepsilon) \left( \frac{\sigma(u^\varepsilon) - \sigma(u)}{\varepsilon} - \sigma'(u) \psi_1 \right) \nabla \varphi^\varepsilon \cdot \nabla v + (\varphi - \varphi^\varepsilon) \sigma'(u) \psi_1 \nabla \varphi^\varepsilon \cdot \nabla v \, dx \right| \\
& \leq \|\varphi_0 - \varphi\|_\infty \cdot \left\| \frac{\sigma(u^\varepsilon) - \sigma(u)}{\varepsilon} - \sigma'(u) \psi_1 \right\|_s \cdot \|\nabla \varphi^\varepsilon\|_r \cdot \|\nabla v\|_2 \\
& + \left| \int_{\Omega} (\varphi_0 - \varphi) \sigma'(u) \psi_1 \nabla (\varphi^\varepsilon - \varphi) \cdot \nabla v \, dx \right| \\
& + \|\varphi - \varphi^\varepsilon\|_{C(\bar{\Omega})} \cdot \left\| \frac{\sigma(u^\varepsilon) - \sigma(u)}{\varepsilon} - \sigma'(u) \psi_1 \right\|_s \cdot \|\nabla \varphi^\varepsilon\|_r \cdot \|\nabla v\|_2 \\
& + \|\varphi - \varphi^\varepsilon\|_{C(\bar{\Omega})} \cdot \|\sigma'(u) \psi_1\|_s \cdot \|\nabla \varphi^\varepsilon\|_r \cdot \|\nabla v\|_2 \rightarrow 0 \text{ as } \varepsilon \rightarrow 0.
\end{aligned}$$

Notice that we need the assumption  $\Omega \subset R^2$  to justify convergence of the integral  $\int_{\Omega} (\varphi - \varphi^\varepsilon) \sigma'(u) \psi_1 \nabla \varphi^\varepsilon \cdot \nabla v \, dx$  to 0 using the imbedding  $W^{1,r}(\Omega) \subset\subset C(\bar{\Omega})$ .

One can show that the second term on the right-hand side of (4.9) is

$$\begin{aligned}
\frac{1}{\varepsilon} \int_{\Omega} \left[ \sigma(u^\varepsilon) \nabla \varphi^\varepsilon - \sigma(u) \nabla \varphi \right] \cdot \nabla \varphi_0 v \, dx & \rightarrow \int_{\Omega} \sigma'(u) \psi_1 \nabla \varphi \cdot \nabla \varphi_0 v \, dx \\
& + \int_{\Omega} \sigma(u) \nabla \psi_2 \cdot \nabla \varphi_0 v \, dx \text{ as } \varepsilon \rightarrow 0.
\end{aligned}$$

Finally, the terms of (4.10) satisfy

$$\begin{aligned}
\frac{1}{\varepsilon} \int_{\Omega} \left[ \sigma(u^\varepsilon) \nabla \varphi^\varepsilon \cdot \nabla w - \sigma(u) \nabla \varphi \cdot \nabla w \right] dx & \rightarrow \int_{\Omega} \sigma'(u) \psi_1 \nabla \varphi \cdot \nabla w \, dx \\
& + \int_{\Omega} \sigma(u) \nabla \psi_2 \cdot \nabla w \, dx \text{ as } \varepsilon \rightarrow 0.
\end{aligned}$$

Letting  $\varepsilon \rightarrow 0$  in (4.9) and (4.10) we obtain

$$\begin{aligned}
& \int_{\Omega} \nabla \psi_1 \cdot \nabla v \, dx + \int_{\partial\Omega} (\beta \psi_1 + \ell u) v \, ds = \int_{\Omega} (\varphi_0 - \varphi) \sigma'(u) \psi_1 \nabla \varphi \cdot \nabla v \, dx \\
& + \int_{\Omega} (\varphi_0 - \varphi) \sigma(u) \nabla \psi_2 \cdot \nabla v \, dx - \int_{\Omega} \psi_2 \sigma(u) \nabla \varphi \cdot \nabla v \, dx \\
(4.12) \quad & + \int_{\Omega} \sigma'(u) \psi_1 \nabla \varphi \cdot \nabla \varphi_0 v \, dx + \int_{\Omega} \sigma(u) \nabla \psi_2 \cdot \nabla \varphi_0 v \, dx \quad \forall v \in H^1(\Omega), \\
& \int_{\Omega} \sigma'(u) \psi_1 \nabla \varphi \cdot \nabla w + \int_{\Omega} \sigma(u) \nabla \psi_2 \cdot \nabla w \, dx = 0 \quad \forall w \in H_0^1(\Omega).
\end{aligned}$$

We can show that the limit of any subsequence satisfies (4.12). By uniqueness of solutions to (4.12), we get convergence of quotients for all  $\varepsilon \rightarrow 0$ . Now the weak formulation for (4.2) is

$$(4.13) \quad \begin{aligned} & \int_{\Omega} \nabla \psi_1 \cdot \nabla v \, dx + \int_{\partial\Omega} (\beta \psi_1 + \ell u) v \, ds = \int_{\Omega} \sigma'(u) \psi_1 |\nabla \varphi|^2 v \, dx \\ & + 2 \int_{\Omega} \sigma(u) \nabla \varphi \cdot \nabla \psi_2 v \, dx \quad \forall v \in H^1(\Omega), \end{aligned}$$

$$(4.14) \quad \int_{\Omega} (\sigma'(u) \psi_1 \nabla \varphi + \sigma(u) \nabla \psi_2) \cdot \nabla w \, dx = 0 \quad \forall w \in H_0^1(\Omega).$$

The quadratic term on the right-hand side of (4.13) is dealt with as before (see [9]). We use (2.4) and (2.5) to conclude

$$(4.15) \quad \frac{1}{s'} = \frac{1}{s} + \frac{1}{r/2}.$$

Since  $\psi_1 \in H^1(\Omega)$  implies  $\psi_1 \in L^s(\Omega)$ , and taking into account that  $|\nabla \varphi|^2 \in L^{r/2}(\Omega)$ , it follows from (4.15) that  $\sigma'(u) |\nabla \varphi|^2 \psi_1 \in L^{s'}(\Omega)$ . Hence,  $\int_{\Omega} \sigma'(u) |\nabla \varphi|^2 \psi_1 v \, dx$  makes sense as  $v \in H^1(\Omega) \subset L^s(\Omega)$ . Eventually, (4.13) can be written as

$$(4.16) \quad \begin{aligned} & \int_{\Omega} \nabla \psi_1 \cdot \nabla v \, dx + \int_{\partial\Omega} (\beta \psi_1 + \ell u) v \, ds = \int_{\Omega} \sigma'(u) \psi_1 \nabla \varphi \cdot \nabla \varphi_0 v \, dx \\ & + \int_{\Omega} \sigma(u) \nabla \psi_2 \cdot \nabla \varphi_0 v \, dx + \int_{\Omega} (\varphi_0 - \varphi) \sigma'(u) \psi_1 \nabla \varphi \cdot \nabla v \, dx \\ & + \int_{\Omega} (\varphi_0 - \varphi) \sigma(u) \nabla \psi_2 \cdot \nabla v \, dx + \int_{\Omega} \sigma(u) \nabla \psi_2 \cdot \nabla \varphi v \, dx \quad \forall v \in H^1(\Omega). \end{aligned}$$

We see that these two weak formulations coincide since  $\int_{\Omega} \sigma(u) v \nabla \psi_2 \cdot \nabla \varphi \, dx = - \int_{\Omega} \sigma(u) \psi_2 \nabla v \cdot \nabla \varphi \, dx$  and  $\nabla \cdot (\sigma(u) \nabla \varphi) = 0$  in  $\Omega$ .  $\square$

In order to characterize the optimal control, we need to introduce adjoint functions  $p$  and  $q$  and the adjoint of the operator  $\mathcal{L}$  in system (4.2). Imposing the boundary conditions  $\frac{\partial p}{\partial n} + \beta^* p = 0$  on  $\partial\Omega$ ,  $q = 0$  on  $\partial\Omega$  gives

$$\begin{aligned} & \int_{\Omega} \begin{pmatrix} p & q \end{pmatrix} \mathcal{L} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} dx = \int_{\Omega} p \Delta \psi_1 \, dx + \int_{\Omega} p \sigma'(u) |\nabla \varphi|^2 \psi_1 \, dx \\ & + 2 \int_{\Omega} p \sigma(u) \nabla \varphi \cdot \nabla \psi_2 \, dx + \int_{\Omega} q \nabla \cdot [\sigma(u) \nabla \psi_2] \, dx + \int_{\Omega} q \nabla \cdot [\sigma'(u) \psi_1 \nabla \varphi] \, dx \\ & = \int_{\Omega} \psi_1 \Delta p \, dx + \int_{\Omega} \psi_1 \sigma'(u) |\nabla \varphi|^2 p \, dx - 2 \int_{\Omega} \psi_2 \nabla \cdot [p \sigma(u) \nabla \varphi] \, dx \\ & - \int_{\Omega} \psi_1 \sigma'(u) \nabla \varphi \cdot \nabla q \, dx + \int_{\Omega} \psi_2 \nabla \cdot [\sigma(u) \nabla q] \, dx = \int_{\Omega} \begin{pmatrix} \psi_1 & \psi_2 \end{pmatrix} \mathcal{L}^* \begin{pmatrix} p \\ q \end{pmatrix} dx, \end{aligned}$$

where

$$\mathcal{L}^* \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} \Delta p + \sigma'(u) |\nabla \varphi|^2 p - \sigma'(u) \nabla \varphi \cdot \nabla q \\ \nabla \cdot [-2p \sigma(u) \nabla \varphi + \sigma(u) \nabla q] \end{pmatrix}.$$

Thus, the adjoint system reads

$$\begin{aligned}
 \Delta p + \sigma'(u)|\nabla\varphi|^2 p - \sigma'(u)\nabla\varphi \cdot \nabla q &= 1 \text{ in } \Omega, \\
 \nabla \cdot [-2p\sigma(u)\nabla\varphi + \sigma(u)\nabla q] &= 0 \text{ in } \Omega, \\
 \frac{\partial p}{\partial n} + \beta^* p &= 0 \text{ on } \partial\Omega, \\
 q &= 0 \text{ on } \partial\Omega,
 \end{aligned}
 \tag{4.17}$$

where the nonhomogeneous term “1” comes from differentiating the integrand of  $J(\beta)$  with respect to the state  $u$ .

**THEOREM 4.2.** *Given an optimal control  $\beta^* \in U_M$  and corresponding states  $u, \varphi$ , there exists  $\Lambda_2 > 0$  such that  $\|\varphi_0\|_{W^{1,\infty}(\Omega)} < \Lambda_2$ , and there exists a solution  $(p, q) \in H^1(\Omega) \times H_0^1(\Omega)$  to the adjoint system (4.17). Furthermore,  $\beta^*$  can be explicitly characterized as*

$$\beta^*(x) = \min \left( \max \left( -\frac{up}{2}, \lambda \right), M \right).
 \tag{4.18}$$

*Proof.* The weak formulation of (4.17) is

$$\begin{aligned}
 & - \int_{\Omega} \nabla p \cdot \nabla v \, dx - \int_{\partial\Omega} \beta^* p v \, ds + \int_{\Omega} \sigma'(u)|\nabla\varphi|^2 p v \, dx \\
 & - \int_{\Omega} \sigma'(u)(\nabla\varphi \cdot \nabla q) v \, dx = \int_{\Omega} v \, dx \quad \forall v \in H^1(\Omega), \\
 & 2 \int_{\Omega} p \sigma(u) \nabla\varphi \cdot \nabla w \, dx - \int_{\Omega} \sigma(u) \nabla q \cdot \nabla w \, dx = 0 \quad \forall w \in H_0^1(\Omega).
 \end{aligned}
 \tag{4.19}$$

*Step 1.* First, we show that the solution to the adjoint system (4.17) exists. Define the map  $F : H^1(\Omega) \times H_0^1(\Omega) \mapsto H^1(\Omega) \times H_0^1(\Omega)$  with  $F(V, W) = (p, q)$  as follows:

$$\begin{aligned}
 \Delta p + \sigma'(u)|\nabla\varphi|^2 V - \sigma'(u)\nabla\varphi \cdot \nabla W &= 1 \text{ in } \Omega, \\
 \nabla \cdot [-2V\sigma(u)\nabla\varphi + \sigma(u)\nabla W] &= 0 \text{ in } \Omega, \\
 \frac{\partial p}{\partial n} + \beta^* p &= 0 \text{ on } \partial\Omega, \\
 q &= 0 \text{ on } \partial\Omega.
 \end{aligned}
 \tag{4.20}$$

Standard  $L^2$ -theory for elliptic problems and the Lax–Milgram lemma imply that this is a well-defined map. The weak formulation of (4.20) is

$$\begin{aligned}
 & - \int_{\Omega} \nabla p \cdot \nabla \Theta \, dx - \int_{\partial\Omega} \beta^* p \Theta \, ds + \int_{\Omega} \sigma'(u)|\nabla\varphi|^2 V \Theta \, dx \\
 & - \int_{\Omega} \sigma'(u)(\nabla\varphi \cdot \nabla W) \Theta \, dx = \int_{\Omega} \Theta \, dx \quad \forall \Theta \in H^1(\Omega), \\
 & 2 \int_{\Omega} V \sigma(u) \nabla\varphi \cdot \nabla \Psi \, dx = \int_{\Omega} \sigma(u) \nabla W \cdot \nabla \Psi \, dx \quad \forall \Psi \in H_0^1(\Omega).
 \end{aligned}$$

We use the Banach fixed point theorem. Let  $F(V_1, W_1) = (p_1, q_1)$  and  $F(V_2, W_2) = (p_2, q_2)$ . We show the contraction property

$$\|p_1 - p_2\|_{H^1} + \|q_1 - q_2\|_{H_0^1} \leq \delta \left( \|V_1 - V_2\|_{H^1} + \|W_1 - W_2\|_{H_0^1} \right)
 \tag{4.21}$$

for some  $0 < \delta < 1$ . Let us denote  $\bar{p} \stackrel{\text{def}}{=} p_1 - p_2$ ,  $\bar{q} \stackrel{\text{def}}{=} q_1 - q_2$ ,  $\bar{W} \stackrel{\text{def}}{=} W_1 - W_2$ , and  $\bar{V} \stackrel{\text{def}}{=} V_1 - V_2$ . Taking  $\Theta = \bar{p}$ , the  $p_1$  and  $p_2$  equations give

$$\int_{\Omega} |\nabla \bar{p}|^2 dx + \int_{\partial\Omega} \beta^* \bar{p}^2 ds = \int_{\Omega} \sigma'(u) |\nabla \varphi|^2 \bar{V} \bar{p} dx - \int_{\Omega} \sigma'(u) \nabla \varphi \cdot \nabla \bar{W} \bar{p} dx.$$

Thus, we obtain

$$\int_{\Omega} |\nabla \bar{p}|^2 dx + \lambda \int_{\partial\Omega} \bar{p}^2 ds \leq \left| \int_{\Omega} \sigma'(u) |\nabla \varphi|^2 \bar{V} \bar{p} dx \right| + \left| \int_{\Omega} \sigma'(u) \nabla \varphi \cdot \nabla \bar{W} \bar{p} dx \right|.$$

Estimating the various terms, one obtains

$$\begin{aligned} \|\bar{p}\|_{H^1(\Omega)} + \|\bar{q}\|_{H_0^1(\Omega)} &\leq (C_{14}\Phi^2 + C_{15}\Phi) \|\bar{V}\|_{H^1(\Omega)} + C_{16}\Phi \|\bar{W}\|_{H^1(\Omega)} \\ &\leq \max \{C_{14}\Phi^2 + C_{15}\Phi, C_{16}\Phi\} (\|\bar{W}\|_{H^1} + \|\bar{V}\|_{H^1}). \end{aligned}$$

Recalling that  $\Phi = c_1 \|\nabla \varphi_0\|_{\infty} + c_2 \|\varphi_0\|_{W^{1,\infty}(\Omega)}$ , where  $c_1, c_2 > 0$  denote generic constants (see (2.10)), and choosing  $\varphi_0$  such that  $\|\varphi_0\|_{W^{1,\infty}(\Omega)}$  is sufficiently small, we obtain

$$\delta = \max \{C_{14}\Phi^2 + C_{15}\Phi, C_{16}\Phi\} < 1.$$

This proves the contraction property (4.21), which gives the desired fixed point of the map and therefore the existence and uniqueness of solutions to the adjoint system.

*Step 2.* We derive the characterization of the optimal control. Recall that for variation  $\ell \in U_M$ , with  $\beta^* + \varepsilon \ell \in U_M$  for  $\varepsilon$  small, the weak formulation of the sensitivity system (4.2) is given by (4.13) and (4.14). Now we are ready to characterize the optimal control. Since the minimum of  $J$  is achieved at  $\beta^*$  and for small  $\varepsilon > 0$ ,  $\beta^* + \varepsilon \ell \in U_M$ , and denoting  $u^\varepsilon = u(\beta^* + \varepsilon \ell)$ ,  $\varphi^\varepsilon = \varphi(\beta^* + \varepsilon \ell)$ , we obtain

$$\begin{aligned} 0 &\leq \lim_{\varepsilon \rightarrow 0^+} \frac{J(\beta^* + \varepsilon \ell) - J(\beta^*)}{\varepsilon} \\ &= \int_{\Omega} \psi_1 dx + \int_{\partial\Omega} 2\beta^* \ell ds = \int_{\Omega} \begin{pmatrix} \psi_1 & \psi_2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} dx + \int_{\partial\Omega} 2\beta^* \ell ds \\ &= \left\{ - \int_{\Omega} \nabla p \nabla \psi_1 dx - \int_{\partial\Omega} \beta^* p \psi_1 ds + \int_{\Omega} \psi_1 \sigma'(u) |\nabla \varphi|^2 p dx \right. \\ &\quad \left. + 2 \int_{\Omega} p \sigma(u) \nabla \varphi \nabla \psi_2 dx \right\} + \left[ - \int_{\Omega} \psi_1 \sigma'(u) \nabla \varphi \cdot \nabla q dx - \int_{\Omega} \sigma(u) \nabla q \cdot \nabla \psi_2 dx \right] \\ &\quad + \int_{\partial\Omega} 2\beta^* \ell ds = \int_{\partial\Omega} \ell (2\beta^* + up) ds, \end{aligned}$$

where we used the sensitivity system (4.13), (4.14) with test functions  $p$  and  $q$ . Using a standard argument on the choice of  $\ell$ , we obtain (4.18).  $\square$

Substituting (4.18) into the state system (1.1) and the adjoint equations (4.17)



we obtain the optimality system

$$\begin{aligned}
 (4.22) \quad & \Delta u^* + \sigma(u^*)|\nabla \varphi^*|^2 = 0 \text{ in } \Omega, \\
 & \nabla \cdot (\sigma(u^*)\nabla \varphi^*) = 0 \text{ in } \Omega, \\
 & \Delta p + \sigma'(u^*)|\nabla \varphi^*|^2 p - \sigma'(u^*)\nabla \varphi^* \cdot \nabla q = 1 \text{ in } \Omega, \\
 & \nabla \cdot [-2p\sigma(u^*)\nabla \varphi^* + \sigma(u^*)\nabla q] = 0 \text{ in } \Omega, \\
 & \frac{\partial p}{\partial n} + \min(\max(-u^*p/2, \lambda), M)p = 0 \text{ on } \partial\Omega, \\
 & q = 0 \text{ on } \partial\Omega, \\
 & \frac{\partial u^*}{\partial n} + \min(\max(-u^*p/2, \lambda), M)u^* = 0 \text{ on } \partial\Omega, \\
 & \varphi^* = \varphi_0 \text{ on } \partial\Omega.
 \end{aligned}$$

Note that existence of solution to the optimality system (4.22) follows from the existence of solution to the state system (1.1) and Theorem 4.2.

*Remarks 1.* If we take into account the second equation in (4.22), then the last equation in (4.22) can be rewritten as

$$-2\sigma(u^*)\nabla \varphi^* \cdot \nabla p + \nabla \cdot [\sigma(u^*)\nabla q] = 0.$$

2. We show that if  $\varphi_0 \equiv \text{constant}$ , then  $\beta^* = \lambda$ . Indeed, the maximum principle for weak solutions implies that  $\varphi^* \equiv \text{constant}$  and therefore  $\nabla \varphi^* = 0$ . This simplifies (4.22) significantly. Namely, we obtain

$$\begin{aligned}
 (4.23) \quad & \Delta u^* = 0 \text{ in } \Omega, \\
 & \Delta p = 1 \text{ in } \Omega, \\
 & \nabla \cdot [\sigma(u^*)\nabla q] = 0 \text{ in } \Omega, \\
 & \frac{\partial u^*}{\partial n} + \beta^* u^* = 0 \text{ on } \partial\Omega, \\
 & \frac{\partial p}{\partial n} + \beta^* p = 0 \text{ on } \partial\Omega, \\
 & q = 0 \text{ on } \partial\Omega.
 \end{aligned}$$

Applying the maximum principle to  $q$  gives  $q \equiv 0$  in  $\bar{\Omega}$ , and therefore only the  $u$  and  $p$  equations remain in (4.23). Now if we integrate the  $u$  equation in (4.23) over  $\Omega$  and integrate by parts, we obtain  $\int_{\partial\Omega} \beta^* u^* ds = 0$ . Since  $\beta^*(x) > 0$  and  $u^*(x) \geq 0$  for all  $x \in \partial\Omega$  it follows that  $u^* = 0$  a.e. on  $\partial\Omega$ . Hence, we can write

$$\begin{aligned}
 \Delta u^* &= 0 \text{ in } \Omega, \\
 u^* &= 0 \text{ a.e. on } \partial\Omega.
 \end{aligned}$$

This implies that  $u^* = 0$  a.e. in  $\Omega$ . Thus, (4.18) implies  $\beta^* = \lambda$  and

$$\begin{aligned}
 (4.24) \quad & \Delta p = 1 \text{ in } \Omega, \\
 & \frac{\partial p}{\partial n} + \lambda p = 0 \text{ on } \partial\Omega.
 \end{aligned}$$

The solution of (4.24) satisfies the equality  $\int_{\partial\Omega} \frac{\partial p}{\partial n} ds = \int_{\Omega} dx$ , which implies that  $-\lambda \int_{\partial\Omega} p ds = \text{meas}(\Omega)$ .

**5. Uniqueness of the optimal control.** In this section we prove that the solution to the optimality system (4.22) is unique under some additional assumptions, in particular, the boundedness of  $u$  and  $p$ . We assume that  $\sigma'$  is Lipschitz with constant  $K_1 > 0$ ,

$$(5.1) \quad |\sigma'(s_1) - \sigma'(s_2)| \leq K_1 |s_1 - s_2| \quad \forall s_1, s_2 \in R.$$

**THEOREM 5.1.** *In addition to all the standard assumptions from section 3, let (5.1) hold. There exists  $\Lambda_3 > 0$  such that  $\|\varphi_0\|_{W^{1,\infty}(\Omega)} \leq \Lambda_3$ , and if in addition  $\lambda$  is large enough, then the solution to the optimality system (4.22), with  $u$  and  $p$  components assumed to belong to  $L^\infty(\Omega)$ , is unique, and therefore the optimal control  $\beta^*$  is unique.*

*Proof.* Let  $(u_1, \varphi_1, p_1, q_1)$  and  $(u_2, \varphi_2, p_2, q_2)$  be two solutions to the optimality system. Using the weak formulation of the state and adjoint equations and taking into account that by the Meyers estimate  $p \in L^\infty(\Omega)$  implies  $\nabla q \in L^r(\Omega)$  we get

$$\begin{aligned} \int_{\Omega} \nabla u_i \cdot \nabla v \, dx + \int_{\partial\Omega} \beta_i u_i v \, ds &= \int_{\Omega} \sigma(u_i) |\nabla \varphi_i|^2 v \, dx \quad \forall v \in H^1(\Omega), \\ \int_{\Omega} \sigma(u_i) \nabla \varphi_i \cdot \nabla w \, dx &= 0, \quad \varphi_0 - \varphi_i \in H_0^1(\Omega) \quad \forall w \in H_0^1(\Omega), \\ \int_{\Omega} \nabla p_i \cdot \nabla \tilde{v} \, dx + \int_{\partial\Omega} \beta_i p_i \tilde{v} \, ds + \int_{\Omega} \tilde{v} \, dx &= \int_{\Omega} \sigma'(u_i) |\nabla \varphi_i|^2 p_i \tilde{v} \, dx \\ &\quad - \int_{\Omega} \sigma'(u_i) \nabla \varphi_i \cdot \nabla q_i \tilde{v} \, dx \quad \forall \tilde{v} \in H^1(\Omega), \\ 2 \int_{\Omega} p_i \sigma(u_i) \nabla \varphi_i \cdot \nabla \tilde{w} \, dx &= \int_{\Omega} \sigma(u_i) \nabla q_i \cdot \nabla \tilde{w} \, dx \quad \forall \tilde{w} \in H_0^1(\Omega), \quad i = 1, 2, \end{aligned}$$

where  $\beta_i = \min(\max(-u_i p_i / 2, \lambda), M)$ .

We set  $\bar{u} = u_1 - u_2$ ,  $\bar{\varphi} = \varphi_1 - \varphi_2$ ,  $\bar{p} = p_1 - p_2$ , and  $\bar{q} = q_1 - q_2$ . Now, taking  $w = \varphi_1 - \varphi_2$  in the equation satisfied by  $u_1$  and  $w = \varphi_1 - \varphi_2$  in the equation satisfied by  $u_2$  yields

$$(5.2) \quad \int_{\Omega} \sigma(u_1) \nabla \varphi_1 \cdot \nabla (\varphi_1 - \varphi_2) \, dx = \int_{\Omega} \sigma(u_2) \nabla \varphi_2 \cdot \nabla (\varphi_1 - \varphi_2) \, dx.$$

Adding  $-\int_{\Omega} \sigma(u_1) \nabla \varphi_2 \cdot \nabla (\varphi_1 - \varphi_2) \, dx$  to both sides of (5.2), one verifies

$$\int_{\Omega} \sigma(u_1) |\nabla \bar{\varphi}|^2 \, dx = \int_{\Omega} (\sigma(u_2) - \sigma(u_1)) \nabla \varphi_2 \cdot \nabla \bar{\varphi} \, dx.$$

Hence,

$$\begin{aligned} C_1 \|\nabla \bar{\varphi}\|_2^2 &= C_1 \int_{\Omega} |\nabla \bar{\varphi}|^2 \, dx \leq \int_{\Omega} \sigma(u_1) |\nabla \bar{\varphi}|^2 \, dx \\ &= \int_{\Omega} (\sigma(u_2) - \sigma(u_1)) \nabla \varphi_2 \cdot \nabla \bar{\varphi} \, dx \\ &\leq K \int_{\Omega} |\bar{u}| |\nabla \varphi_2| |\nabla \bar{\varphi}| \, dx \leq K \Phi \|\bar{u}\|_s \|\nabla \bar{\varphi}\|_2, \end{aligned}$$

which implies

$$(5.3) \quad \|\nabla \bar{\varphi}\|_2 \leq \frac{K\Phi}{C_1} \|\bar{u}\|_s.$$

From PDEs for  $u_1$  and  $u_2$ , we obtain

$$\int_{\Omega} |\nabla \bar{u}|^2 dx + \int_{\partial\Omega} \beta_1 \bar{u}^2 + (\beta_1 - \beta_2) u_2 \bar{u} ds = \int_{\Omega} \left( \sigma(u_1) |\nabla \varphi_1|^2 - \sigma(u_2) |\nabla \varphi_2|^2 \right) \bar{u} dx,$$

which yields

$$\begin{aligned} \int_{\Omega} |\nabla \bar{u}|^2 dx + \lambda \int_{\partial\Omega} \bar{u}^2 ds &\leq \frac{1}{2} \|u_1\|_{\infty} \|u_2\|_{\infty} \|\bar{u}\|_{L^2(\partial\Omega)} \|\bar{p}\|_{L^2(\partial\Omega)} \\ &\quad + \frac{1}{2} \|p_2\|_{\infty} \|u_2\|_{\infty} \|\bar{u}\|_{L^2(\partial\Omega)}^2 \\ &\quad + \int_{\Omega} |\sigma(u_1) |\nabla \varphi_1|^2 - \sigma(u_2) |\nabla \varphi_2|^2| \cdot |\bar{u}| dx, \end{aligned}$$

where we used  $|\beta_1 - \beta_2| \leq \frac{1}{2} |u_1 p_1 - u_2 p_2|$ . Denoting  $\alpha_1 := \|u_1\|_{\infty} \|u_2\|_{\infty}$  and  $\alpha_2 := \|p_2\|_{\infty} \|u_2\|_{\infty}$ , we can write

$$\begin{aligned} \int_{\Omega} |\nabla \bar{u}|^2 dx + \left( \lambda - \frac{\alpha_2}{2} - \frac{\alpha_1}{4} \right) \int_{\partial\Omega} \bar{u}^2 ds &\leq \frac{\alpha_1}{4} \|\bar{p}\|_{L^2(\partial\Omega)}^2 \\ &\quad + \int_{\Omega} |\sigma(u_1) |\nabla \varphi_1|^2 - \sigma(u_2) |\nabla \varphi_2|^2| \cdot |\bar{u}| dx. \end{aligned}$$

We deal with the last term

$$\begin{aligned} \int_{\Omega} |\sigma(u_1) |\nabla \varphi_1|^2 - \sigma(u_2) |\nabla \varphi_2|^2| \cdot |\bar{u}| dx &\leq \|\sigma(u_1) |\nabla \varphi_1|^2 - \sigma(u_2) |\nabla \varphi_2|^2\|_{s'} \|\bar{u}\|_s \\ &\leq \|\sigma(u_1) \nabla \bar{\varphi} \cdot \nabla (\varphi_1 + \varphi_2)\|_{s'} \|\bar{u}\|_s + \|(\sigma(u_1) - \sigma(u_2)) |\nabla \varphi_2|^2\|_{s'} \|\bar{u}\|_s \\ &\leq C_2 \|\nabla \bar{\varphi}\|_2 \cdot \|\nabla (\varphi_1 + \varphi_2)\|_r \cdot \|\bar{u}\|_s + K \|\bar{u}\|_s \|\nabla \varphi_2\|_2^2 \\ &\leq C_{16} \Phi \|\nabla (\varphi_1 + \varphi_2)\|_r \cdot \|\bar{u}\|_s^2 + K \|\nabla \varphi_2\|_2^2 \|\bar{u}\|_s^2 \\ &\leq C_{17} \Phi^2 \|\bar{u}\|_{H^1(\Omega)}^2, \end{aligned}$$

where we have taken into account that  $\|\nabla \varphi\|_r \leq \Phi$ ,  $\|\bar{u}\|_s \leq M_1 \|\bar{u}\|_{H^1(\Omega)}$ , and (5.3). Consequently, we obtain

$$(5.4) \quad \int_{\Omega} |\nabla \bar{u}|^2 dx + \left( \lambda - \frac{\alpha_2}{2} - \frac{\alpha_1}{4} \right) \int_{\partial\Omega} \bar{u}^2 ds \leq \frac{\alpha_1}{4} \|\bar{p}\|_{L^2(\partial\Omega)}^2 + C_{19} \Phi^2 \|\bar{u}\|_{H^1(\Omega)}^2.$$

The estimate for  $\bar{p}$  is derived in a similar manner [10], albeit after rather lengthy calculations, in which condition (5.1) is explicitly used. The result is

$$\begin{aligned} \int_{\Omega} |\nabla \bar{p}|^2 dx + \left( \lambda - \frac{\alpha_3}{2} - \frac{\alpha_4^2}{4} \right) \int_{\partial\Omega} \bar{p}^2 ds &\leq \frac{\alpha_4^2}{4} \|\bar{u}\|_{L^2(\partial\Omega)}^2 + C_{20} \Phi \|\bar{p}\|_{H^1(\Omega)}^2 \\ (5.5) \quad &\quad + C_{21} \Phi \|\bar{p}\|_{H^1} \|\nabla \bar{q}\|_2 + C_{22} \Phi \|\bar{u}\|_{H^1} \|\bar{p}\|_{H^1}, \end{aligned}$$

where  $\alpha_3 := \|u_1\|_{\infty} \|p_2\|_{\infty}$ ,  $\alpha_4 := \|p_2\|_{\infty}$ . Note that the estimate for  $\bar{\varphi}$  is given by

$$(5.6) \quad \|\nabla \bar{\varphi}\|_2 \leq C_{23} \Phi \|\bar{u}\|_{H^1(\Omega)},$$

where we have taken into account that  $\|\bar{u}\|_s \leq M_1 \|\bar{u}\|_{H^1(\Omega)}$ .

Using  $\bar{q} = \bar{w}$  as a test function in the equations for  $q$ , we have

$$\begin{aligned} 2 \int_{\Omega} p_1 \sigma(u_1) \nabla \varphi_1 \cdot \nabla \bar{q} dx &= \int_{\Omega} \sigma(u_1) \nabla q_1 \cdot \nabla \bar{q} dx, \\ 2 \int_{\Omega} p_2 \sigma(u_2) \nabla \varphi_2 \cdot \nabla \bar{q} dx &= \int_{\Omega} \sigma(u_2) \nabla q_2 \cdot \nabla \bar{q} dx. \end{aligned}$$

Upon subtracting and estimating the difference, we obtain

$$(5.7) \quad \|\nabla \bar{q}\|_2 \leq C_{24}\Phi\|\bar{u}\|_{H^1(\Omega)} + C_{25}\Phi\|\bar{p}\|_{H^1(\Omega)}.$$

Now adding (5.4) and (5.5) gives

$$(5.8) \quad \begin{aligned} & \int_{\Omega} |\nabla \bar{u}|^2 dx + \left( \lambda - \frac{\alpha_2}{2} - \frac{\alpha_1}{4} - \frac{\alpha_4^2}{4} \right) \int_{\partial\Omega} \bar{u}^2 ds + \int_{\Omega} |\nabla \bar{p}|^2 dx \\ & + \left( \lambda - \frac{\alpha_3}{2} - \frac{\alpha_4^2}{4} - \frac{\alpha_1}{4} \right) \int_{\partial\Omega} \bar{p}^2 ds \leq C_{19}\Phi^2\|\bar{u}\|_{H^1(\Omega)}^2 + C_{20}\Phi\|\bar{p}\|_{H^1(\Omega)}^2 \\ & + C_{21}\Phi\|\bar{p}\|_{H^1(\Omega)}\|\nabla \bar{q}\|_2 + C_{22}\Phi\|\bar{u}\|_{H^1(\Omega)}\|\bar{p}\|_{H^1(\Omega)}. \end{aligned}$$

Using the estimate for  $\nabla \bar{q}$  allows us to rewrite (5.8) in the following way:

$$\begin{aligned} & \int_{\Omega} |\nabla \bar{u}|^2 dx + \left( \lambda - \frac{\alpha_2}{2} - \frac{\alpha_1}{4} - \frac{\alpha_4^2}{4} \right) \int_{\partial\Omega} \bar{u}^2 ds + \int_{\Omega} |\nabla \bar{p}|^2 dx \\ & + \left( \lambda - \frac{\alpha_3}{2} - \frac{\alpha_4^2}{4} - \frac{\alpha_1}{4} \right) \int_{\partial\Omega} \bar{p}^2 ds \\ & \leq C_{19}\Phi^2\|\bar{u}\|_{H^1(\Omega)}^2 + (C_{20}\Phi + C_{27}\Phi^2)\|\bar{p}\|_{H^1(\Omega)}^2 \\ & + (C_{26}\Phi^2 + C_{22}\Phi)\|\bar{u}\|_{H^1(\Omega)}\|\bar{p}\|_{H^1(\Omega)} \\ & \leq (C_{27}\Phi^2 + C_{28}\Phi)(\|\bar{u}\|_{H^1(\Omega)}^2 + \|\bar{p}\|_{H^1(\Omega)}^2). \end{aligned}$$

Upon properly redefining the constants, we conclude that

$$\mathcal{A}\|\bar{u}\|_{H^1(\Omega)}^2 + \mathcal{B}\|\bar{p}\|_{H^1(\Omega)}^2 \leq 0.$$

Therefore, if  $\lambda$  is large and the boundary data  $\varphi_0$  is sufficiently small, then  $\mathcal{A}$  and  $\mathcal{B}$  are positive and  $\bar{u} = 0, \bar{p} = 0, \nabla \bar{q} = 0$ , and  $\nabla \bar{\varphi} = 0$ . Then it follows that  $\bar{q} = 0$  and  $\bar{\varphi} = 0$  (since  $\nabla \bar{q} = 0, \nabla \bar{\varphi} = 0$  and  $\bar{q} = 0, \bar{\varphi} = 0$  on the boundary), which gives the uniqueness of solutions of the optimality system. This implies the uniqueness of the optimal control.  $\square$

*Remark.* The boundedness of  $\sigma$  and the  $L^\infty$  boundedness of temperature are quite realistic assumptions since physical quantities are bounded. For homogeneous media, the Lipschitz condition is also reasonable; in fact, it can be assumed to hold true even for nonhomogeneous media without sharp interfaces.

## REFERENCES

- [1] S. N. ANTONTSEV AND M. CHIPOT, *The thermistor problem: Existence, smoothness, uniqueness, blowup*, SIAM J. Math. Anal., 25 (1994), pp. 1128–1156.
- [2] K. ATKINSON AND W. HAN, *Theoretical Numerical Analysis. A Functional Analysis Framework*, Springer, New York, 2001.
- [3] A. BENSOUSSAN, J.-L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.
- [4] G. CIMATTI, *Existence of weak solutions for the nonstationary problem of the Joule heating of a conductor*, Ann. Mat. Pura Appl. (4), 162 (1992), pp. 33–42.
- [5] G. CIMATTI, *A bound for the temperature in the thermistor's problem*, IMA J. Appl. Math., 40 (1988), pp. 15–22.
- [6] G. CIMATTI AND G. PRODI, *Existence results for a nonlinear elliptic system modeling a temperature dependent electrical resistor*, Ann. Mat. Pura Appl. (4), 63 (1989), pp. 227–236.
- [7] A. C. FOWLER, I. FRIGAARD, AND S. D. HOWISON, *Temperature surges in current-limiting circuit devices*, SIAM J. Appl. Math., 52 (1992), pp. 998–1011.

- [8] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, Berlin, 1983.
- [9] S. D. HOWISON, J. F. RODRIGUES, AND M. SHILLOR, *Stationary solutions to the thermistor problem*, J. Math. Anal. Appl., 174 (1993), pp. 573–588.
- [10] V. HRYNKIV, *Optimal Control of Partial Differential Equations and Variational Inequalities*, Ph.D. thesis, University of Tennessee, Knoxville, TN, 2006; available online at <http://etd.utk.edu/2006/HrynkivVolodymyr.pdf>.
- [11] K. KWOK, *Complete Guide to Semiconductor Devices*, McGraw-Hill, New York, 1995.
- [12] H.-C. LEE AND T. SHILKIN, *Analysis of optimal control problem for the two-dimensional thermistor system*, SIAM J. Control Optim., 44 (2005), pp. 268–282.
- [13] E. D. MACLEN, *Thermistors*, Electrochemical Publications, Glasgow, 1979.
- [14] N. G. MEYERS, *An  $L^p$  estimate for the gradient of solutions of second order elliptic divergence equations*, Ann. Scuola Norm. Sup. Pisa (3), 17 (1963), pp. 189–206.
- [15] P. SHI, M. SHILLOR, AND X. XU, *Existence of a solution to the Stefan problem with Joule's heating*, J. Differential Equations, 105 (1993), pp. 239–263.
- [16] M. E. TAYLOR, *Partial Differential Equations III. Nonlinear Equations*, Springer, New York, 1996.
- [17] X. XU, *Local regularity theorems for the stationary thermistor problem*, Proc. Roy. Soc. Edinburgh Sect. A, 134 (2004), pp. 773–782.
- [18] X. XU, *Exponential integrability of temperature in the thermistor problem*, Differential Integral Equations, 17 (2004), pp. 571–582.
- [19] G. YUAN, *Regularity of solutions of the nonstationary thermistor problem*, Appl. Anal., 53 (1994), pp. 149–164.
- [20] E. ZEIDLER, *Nonlinear Functional Analysis and Its Applications IV*, Springer, New York, 1988.
- [21] S. ZHOU AND D. R. WESTBROOK, *Numerical solution of the thermistor equations*, J. Comput. Appl. Math., 79 (1997), pp. 101–118.

## OPTIMAL OUTPUT REGULATION FOR DISCRETE-TIME SWITCHED AND MARKOVIAN JUMP LINEAR SYSTEMS\*

JI-WOONG LEE<sup>†</sup> AND PRAMOD P. KHARGONEKAR<sup>‡</sup>

**Abstract.** First, three different but related output regulation performance criteria for the linear time-varying system are defined in the discrete-time domain, namely, the peak impulse response, peak output variance, and average output variance per unit time. Then they are extended for switched linear systems and Markovian jump linear systems, and characterized by an increasing union of finite-dimensional linear matrix inequality conditions. Finally, the infinite-horizon suboptimal LQG control problem, which aims to maintain the average output variance below a given level subject to the uniform exponential stability of the closed-loop system, is solved for both switched linear systems and Markovian jump linear systems; the solution is given by a dynamic linear output feedback controller that not only perfectly observes the present mode but also recalls a finite number of past modes.

**Key words.** discrete linear inclusions, dynamic output feedback,  $\mathcal{H}^2$  control, linear matrix inequalities, LQG control, linear time-varying systems, semidefinite programming

**AMS subject classifications.** 49N10, 93B12, 90C22, 93C55

**DOI.** 10.1137/060662290

**1. Introduction.** Switched systems and Markovian jump systems are abstractions of hybrid (or multimodal) systems; they idealize complex interactions between the continuous state trajectory and the discrete state (or mode) transitions by a finite number of state-space representations, together with a constraint on the admissible switching paths among these representations. If the switching path constraint is specified via a set of autonomous switching sequences, then the resulting abstract model is called a switched system; on the other hand, if the switching sequences are specified to be realizations of a homogeneous Markov chain, then the model is called a Markovian jump system. Many complex control systems today exhibit hybrid and multimodal dynamics due to the presence of digital computers [30, 7], communication networks [31, 9, 47], distributed agents [26], etc. Consequently, the study of switched and Markovian jump systems is becoming increasingly important [36, 45, 38, 12]. In this paper, we focus on the output regulation performance and establish exact convex conditions for the analysis and synthesis of discrete-time switched and Markovian jump linear systems via linear matrix inequalities. We make the restrictive, but standard, assumption that the mode of the system is perfectly observed at each time instant. There are many real-world applications where this assumption is satisfied: e.g., missile autopilot via gain scheduling [1] and networked control subject to randomly delayed but time-stamped measurements [47].

There is a strong link between the properties of switched linear systems and those of standard linear systems. Indeed, exact convex conditions for the uniform exponential stabilization and uniform disturbance attenuation (i.e., root-mean-square gain control) of discrete-time switched linear systems have been obtained [34, 33] using the operator theoretic  $\ell^2$ -induced norm analysis of time-varying systems [24, 19] and

---

\*Received by the editors June 7, 2006; accepted for publication (in revised form) June 15, 2007; published electronically January 4, 2008.

<http://www.siam.org/journals/sicon/47-1/66229.html>

<sup>†</sup>Department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802 (jiwoong@psu.edu).

<sup>‡</sup>College of Engineering, University of Florida, Gainesville, FL 32611 (ppk@ufl.edu).

the semidefinite programming-based  $\mathcal{H}^\infty$  synthesis of time-invariant systems [41, 20]; here, the uniformity is over all admissible switching sequences as well as over all time instants. In this paper, we continue to exploit this link to solve output regulation problems. First, inspired by the operator theoretic  $\ell^2$  seminorm analysis of time-varying systems [25], we define three different performance criteria and establish their relation in terms of observability and reachability gramians. Different performance criteria lead to different output regulation problems, including the infinite-horizon LQG control problem, where the average output variance, and hence the average output root-mean-square value, per unit time is minimized subject to the internal stability of the closed-loop system. These problems all coincide with the  $\mathcal{H}^2$  control problem when the number of modes is equal to one; thus our results build on the semidefinite programming-based  $\mathcal{H}^2$  synthesis of time-invariant systems [43, 39]. Similar results are also established for the output regulation of Markovian jump linear systems subject to the almost sure uniform exponential stability. These results form an “output regulation” counterpart to the “disturbance attenuation” results in [33]; except for stability analysis, key elements of [33] such as a uniformly stabilizing property of the Riccati inequality and an elegant matrix completion-based linear matrix embedding technique do not apply here.

Our approach to the infinite-horizon LQG control of switched systems has three important ingredients. The first is the increasing union of “path-dependent” Lyapunov inequality conditions first introduced in [34]; this enables a control-oriented convex characterization of the uniform exponential stability and stabilizability of switched linear systems. The second ingredient is the “change of variable” argument in [43] that provides a systematic means of turning dynamic output feedback requirements into linear matrix inequalities without any assumption on the system parameters. Finally, the third is the notion of “minimal” switching sequences that we adopt from [35]; if the admissible switching sequences are defined by a directed graph of a row-allowable matrix, then we show that, to determine the average output variance level of the switched linear system, it suffices to check the average output variance level of the periodic systems associated with the switching sequences that are minimal with respect to the given directed graph. Putting these ingredients together leads to a complete solution to the infinite-horizon LQG problem, and other output regulation problems, for switched linear systems. The solution to each sub-optimal infinite-horizon LQG problem is given by a finite-path-dependent dynamic output feedback controller; moreover, in many cases with a reasonable tolerance level, the optimal controller is still finite-path dependent. There are few results in the literature that this result can be compared to, mainly because there has not been a control-oriented convex characterization of the stability, let alone stabilizability, of discrete-time switched linear systems until very recently. Earlier relevant results include stability analysis [15, 2, 21, 16, 3], (conservative) stability synthesis [6, 14], and analysis of problem complexity and decidability [46, 4]; see [5] for the Kalman filtering result for a class of discrete-time switched systems.

In the case of Markovian jump linear systems, the infinite-horizon LQG control problem has already been studied with respect to mean square stability (also known as stochastic stability) in both continuous time [28, 17] and discrete time [27, 13]. Particularly in the discrete-time domain, existing results provide a nice state-space formula for the solution and extend the well-known separation principle to the Markovian jump system. However, a key limitation of these results is that the restriction to mode-dependent controllers, and hence to the Markovian filtering, is suboptimal

compared to the path-dependent controller resulting from the Kalman filtering—see the numerical comparison in [13]; another limitation is that there is no convex characterization of the mean square “stabilizability” of discrete-time Markovian jump linear systems [29, 10]. We show in this paper that if the requirement of mean square stability is replaced with the stronger requirement of almost sure “uniform” stability, the infinite-horizon LQG control result for switched linear systems carries over to Markovian jump linear systems without any of these limitations. As in the case of switched systems, all suboptimal controllers are finite-path dependent, and the optimal controller is often finite-path dependent, too, with a short path length for a reasonable tolerance level. Moreover, our result can be applied to any Markov chain and any controlled plant (e.g., nonergodic chains and plants without the standard orthogonality and rank conditions). To our knowledge, with the exception of [34, 33], the almost sure uniform stability of Markovian jump systems has not been explicitly dealt with in the literature. Almost sure uniform stability is a deterministic notion, and hence is useful when only the sparsity pattern of the underlying transition probability matrix, but not the individual transition probabilities, is exactly known [34]. Moreover, unlike the usual stochastic stability-based approaches to controlling Markovian jump systems, our LQG control framework conforms to the classical stochastic control settings where deterministic uniform stability is the commonly used stability notion.

This paper is organized as follows. In section 2, we consider three output regulation performance criteria for the discrete-time linear time-varying system, namely, the peak impulse response, peak output variance, and average output variance per unit time. In section 3, these performance criteria are extended for switched linear systems and characterized by linear matrix inequality conditions. Section 4 is devoted to establishing the synthesis condition for the LQG control problem where the average output variance is minimized; synthesis conditions for the other two output regulation problems are only briefly noted. These results are extended to Markovian jump linear systems in section 5. Finally, concluding remarks on the results and future research are made in section 6.

*Notation.* For  $x \in \mathbb{R}^n$ , denoted by  $\|x\|$  is the Euclidean vector norm  $\|x\| = \sqrt{x^T x}$  of  $x$ . If  $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times n}$  are symmetric and  $\mathbf{X} - \mathbf{Y}$  is positive definite (resp., nonnegative definite), we write  $\mathbf{X} > \mathbf{Y}$  (resp.,  $\mathbf{X} \geq \mathbf{Y}$ ). The trace of  $\mathbf{X}$  is denoted by  $\text{tr } \mathbf{X}$ , and the spectral radius of  $\mathbf{X}$  by  $\rho \mathbf{X}$ . The identity matrix is denoted by  $\mathbf{I}$  with  $n$  understood. If  $\mathbf{X}$  and  $\mathbf{Y}$  are two matrices, then the direct sum  $\mathbf{X} \oplus \mathbf{Y}$  denotes the block diagonal matrix  $\begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{Y} \end{bmatrix}$ , where  $\mathbf{0}$  denotes the zero matrix of appropriate dimension.

## 2. Analysis of linear time-varying systems. Let

$$(2.1a) \quad \mathcal{G} = \{(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0, \mathbf{D}_0), (\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1, \mathbf{D}_1), \dots\}$$

be an indexed family of matrices  $\mathbf{A}_t \in \mathbb{R}^{n \times n}$ ,  $\mathbf{B}_t \in \mathbb{R}^{n \times m}$ ,  $\mathbf{C}_t \in \mathbb{R}^{l \times n}$ , and  $\mathbf{D}_t \in \mathbb{R}^{l \times m}$  over all  $t = 0, 1, \dots$ . Let a sequence  $\boldsymbol{\theta} = (\theta(0), \theta(1), \dots)$  in  $\{0, 1, \dots\}$  by

$$(2.1b) \quad \theta(t) = t$$

for all  $t = 0, 1, \dots$ . Then we have  $(\mathbf{A}_{\theta(t)}, \mathbf{B}_{\theta(t)}, \mathbf{C}_{\theta(t)}, \mathbf{D}_{\theta(t)}) = (\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t, \mathbf{D}_t)$ , and the pair  $(\mathcal{G}, \boldsymbol{\theta})$  defines the discrete-time linear time-varying system

$$(2.2) \quad \begin{aligned} x(t+1) &= \mathbf{A}_t x(t) + \mathbf{B}_t w(t), \\ z(t) &= \mathbf{C}_t x(t) + \mathbf{D}_t w(t). \end{aligned}$$



Given the initial state  $x(0)$  and disturbance sequence  $\mathbf{w} = (w(0), w(1), \dots)$ , equation (2.2) determines the state sequence  $\mathbf{x} = (x(0), x(1), \dots)$  and output sequence  $\mathbf{z} = (z(0), z(1), \dots)$ .

DEFINITION 2.1. *The system  $(\mathcal{G}, \boldsymbol{\theta})$  is said to be uniformly (exponentially) stable if there exist  $c \geq 1$  and  $\lambda \in (0, 1)$  such that, whenever  $\mathbf{w} = 0$ ,*

$$\|x(t)\| \leq c \lambda^{t-t_0} \|x(t_0)\|$$

for all  $t_0 \geq 0$ ,  $t \geq t_0$ , and  $x(t_0) \in \mathbb{R}^n$ .

The set  $\mathcal{G}$  is said to be bounded if there are a  $b > 0$  and a norm  $|\cdot|$  such that  $|g| \leq b$  for all  $g \in \mathcal{G}$ . The following characterizes the stability of linear time-varying systems in terms of an infinite-dimensional system of Lyapunov inequalities.

LEMMA 2.2. *Let  $\mathcal{G}$  and  $\boldsymbol{\theta}$  be as in (2.1); let  $\mathcal{G}$  be bounded. The following are equivalent:*

- (a) *The system  $(\mathcal{G}, \boldsymbol{\theta})$  is uniformly exponentially stable.*
- (b) *There exist  $\alpha, \beta > 0$  and  $\mathbf{X}_t \in \mathbb{R}^{n \times n}$  such that*

$$\alpha \mathbf{I} \leq \mathbf{X}_t \leq \beta \mathbf{I}; \quad \mathbf{A}_t^T \mathbf{X}_{t+1} \mathbf{A}_t - \mathbf{X}_t \leq -\alpha \mathbf{I}$$

for all  $t \geq 0$ .

- (c) *There exist  $\epsilon, \delta > 0$  and  $\mathbf{Y}_t \in \mathbb{R}^{n \times n}$  such that*

$$\epsilon \mathbf{I} \leq \mathbf{Y}_t \leq \delta \mathbf{I}; \quad \mathbf{A}_t \mathbf{Y}_t \mathbf{A}_t^T - \mathbf{Y}_{t+1} \leq -\epsilon \mathbf{I}$$

for all  $t \geq 0$ .

Moreover, if either (b) or (c) holds, then one may take  $\mathbf{X}_t^{-1} = \mathbf{Y}_t$  for all  $t \geq 0$ .

*Proof.* The equivalence of (a) and (b) is a special case of [19, Theorem 11]. A Schur complement argument shows the equivalence of (b) and (c) via the relation  $\mathbf{X}_t^{-1} = \mathbf{Y}_t$ .  $\square$

Let  $\mathbf{X}_t^{(T)} \geq \mathbf{0}$  be the unique solution to the backward Lyapunov equation

$$\mathbf{A}_t^T \mathbf{X}_{t+1}^{(T)} \mathbf{A}_t - \mathbf{X}_t^{(T)} = -\mathbf{C}_t^T \mathbf{C}_t$$

for  $t = 0, 1, \dots, T$ , with the terminal condition  $\mathbf{X}_{T+1}^{(T)} = \mathbf{0}$ . Similarly, let  $\mathbf{Y}_t^{(t_0)} \geq \mathbf{0}$  be the unique solution to the forward Lyapunov equation

$$\mathbf{A}_t \mathbf{Y}_t^{(t_0)} \mathbf{A}_t^T - \mathbf{Y}_{t+1}^{(t_0)} = -\mathbf{B}_t \mathbf{B}_t^T$$

for  $t = t_0, t_0 + 1, \dots$ , with the initial condition  $\mathbf{Y}_{t_0}^{(t_0)} = \mathbf{0}$ . If  $\Phi(t, t_0)$  is the state transition matrix defined by

$$\Phi(t, t_0) = \begin{cases} \mathbf{I}, & t = t_0; \\ \mathbf{A}_{t-1} \cdots \mathbf{A}_{t_0}, & t > t_0, \end{cases}$$

for  $t \geq t_0 \geq 0$ , then we have

$$(2.3a) \quad \mathbf{X}_t^{(T)} = \sum_{s=t}^T \Phi(s, t)^T \mathbf{C}_s^T \mathbf{C}_s \Phi(s, t)$$

for  $0 \leq t \leq T$ , and

$$(2.3b) \quad \mathbf{Y}_{t+1}^{(t_0)} = \sum_{s=t_0}^t \Phi(t+1, s+1) \mathbf{B}_s \mathbf{B}_s^T \Phi(t+1, s+1)^T$$

for  $t \geq t_0 \geq 0$ .

The symmetric nonnegative definite matrix  $\mathbf{X}_t^{(T)}$ ,  $t \leq T$ , is the observability gramian; that is, under  $\mathbf{w} = 0$ , any state  $x(t) = x_0$  is uniquely determined by the output sequence  $(z(t), \dots, z(T))$  if and only if  $\mathbf{X}_t^{(T)} > \mathbf{0}$ . Similarly, the matrix  $\mathbf{Y}_t^{(t_0)}$ ,  $t_0 < t$ , is the reachability gramian; that is, under  $x(t_0) = 0$ , associated with any state  $x_f \in \mathbb{R}^n$  is a disturbance input sequence  $(w(t_0), \dots, w(t-1))$  such that  $x(t) = x_f$  if and only if  $\mathbf{Y}_t^{(t_0)} > \mathbf{0}$ . The following lemma summarizes some of the simple properties of these matrices.

LEMMA 2.3. *Let  $\mathcal{G}$  be as in (2.1a); let  $\mathbf{X}_t^{(T)}$  and  $\mathbf{Y}_{t+1}^{(t_0)}$  be as in (2.3) whenever  $t_0 \leq t \leq T$ . The following hold:*

(a) *We have*

$$\mathbf{X}_t^{(T)} \leq \mathbf{X}_t^{(T+1)}; \quad \mathbf{Y}_{t+1}^{(t_0+1)} \leq \mathbf{Y}_{t+1}^{(t_0)}$$

*whenever  $t_0 \leq t \leq T$ .*

(b) *We have*

$$(2.4) \quad \sum_{t=t_0}^T \text{tr} \mathbf{B}_t^T \mathbf{X}_{t+1}^{(T)} \mathbf{B}_t = \sum_{t=t_0}^T \text{tr} \mathbf{C}_t \mathbf{Y}_t^{(t_0)} \mathbf{C}_t^T$$

*whenever  $t_0 \leq t \leq T$ .*

(c) *If*

$$\mathbf{A}_t^T \mathbf{X}_{t+1} \mathbf{A}_t - \mathbf{X}_t \leq -\mathbf{C}_t^T \mathbf{C}_t; \quad \mathbf{A}_t \mathbf{Y}_t \mathbf{A}_t^T - \mathbf{Y}_{t+1} \leq -\mathbf{B}_t \mathbf{B}_t^T$$

*for  $t_0 \leq t \leq T$ , with  $\mathbf{X}_{T+1} \geq \mathbf{0}$  and  $\mathbf{Y}_{t_0} \geq \mathbf{0}$ , then*

$$\mathbf{X}_t \geq \mathbf{X}_t^{(T)}; \quad \mathbf{Y}_{t+1} \geq \mathbf{Y}_{t+1}^{(t_0)}$$

*whenever  $t_0 \leq t \leq T$ .*

*Proof.* The proof is immediate from the definitions.  $\square$

Remark 1. If  $\mathcal{G}$  is bounded and the system  $(\mathcal{G}, \theta)$  is uniformly stable, then by (2.4) we have that

$$\limsup_{T \rightarrow \infty} \frac{1}{T - t_0} \sum_{t=t_0}^T \text{tr} \mathbf{B}_t^T \mathbf{X}_{t+1}^{(T)} \mathbf{B}_t = \limsup_{T \rightarrow \infty} \frac{1}{T - t_0} \sum_{t=t_0}^T \text{tr} \mathbf{C}_t \mathbf{Y}_t^{(t_0)} \mathbf{C}_t^T$$

for all  $t_0 \geq 0$ . (Note that the partial sums above are divided by  $T - t_0$ , not by  $T - t_0 + 1$ , because  $\text{tr} \mathbf{B}_T^T \mathbf{X}_{T+1}^{(T)} \mathbf{B}_T = \text{tr} \mathbf{C}_{t_0} \mathbf{Y}_{t_0}^{(t_0)} \mathbf{C}_{t_0}^T = 0$ .) This is a time-varying version of the classical result that if  $\mathbf{A}_t = \mathbf{A}$ ,  $\mathbf{B}_t = \mathbf{B}$ , and  $\mathbf{C}_t = \mathbf{C}$  for all  $t$ , and if  $\rho \mathbf{A} < 1$ , then

$$\text{tr} \mathbf{B}^T \mathbf{X} \mathbf{B} = \text{tr} \mathbf{C} \mathbf{Y} \mathbf{C}^T,$$

where  $\mathbf{X}$ ,  $\mathbf{Y} \geq \mathbf{0}$  uniquely solve the Lyapunov equations

$$\mathbf{A}^T \mathbf{X} \mathbf{A} - \mathbf{X} = -\mathbf{C}^T \mathbf{C}; \quad \mathbf{A} \mathbf{Y} \mathbf{A}^T - \mathbf{Y} = -\mathbf{B} \mathbf{B}^T.$$

A similar relation holds if the time-varying system is represented as a doubly infinite block matrix [25].

In the remainder of this section, three criteria for the output regulation performance of linear time-varying systems will be formulated in terms of the quantities  $\text{tr} \mathbf{B}_t^T \mathbf{X}_{t+1}^{(T)} \mathbf{B}_t$  and  $\text{tr} \mathbf{C}_t \mathbf{Y}_t^{(t_0)} \mathbf{C}_t^T$ . We assume that the family  $\mathcal{G}$  is bounded and the system  $(\mathcal{G}, \theta)$  is uniformly stable. Note that if  $x(0) = 0$  and  $\mathbf{w} = (w(0), w(1), \dots)$ , we have

$$z(t) = \begin{cases} \mathbf{D}_0 w(0), & t = 0; \\ \sum_{s=0}^{t-1} \mathbf{C}_t \Phi(t, s+1) \mathbf{B}_s w(s) + \mathbf{D}_t w(t), & t > 0. \end{cases}$$

Define

$$(2.5) \quad w_k^{(t_0)}(t) = \begin{cases} e_k, & t = t_0; \\ 0, & t \neq t_0, \end{cases}$$

for  $t_0, t \geq 0$  and  $1 \leq k \leq m$ , where  $e_k$  is the  $k$ th standard basis vector in  $\mathbb{R}^m$  (i.e., the  $k$ th column of  $\mathbf{I} \in \mathbb{R}^{m \times m}$ ). Then, with  $x(0) = 0$  and  $\mathbf{w} = (w_k^{(t_0)}(0), w_k^{(t_0)}(1), \dots)$ , the system  $(\mathcal{G}, \theta)$  yields  $z = (z_k^{(t_0)}(0), z_k^{(t_0)}(1), \dots)$ , where

$$(2.6) \quad z_k^{(t_0)}(t) = \begin{cases} \mathbf{D}_{t_0} e_k, & t = t_0; \\ \mathbf{C}_t \Phi(t, t_0+1) \mathbf{B}_{t_0} e_k, & t > t_0. \end{cases}$$

The first performance criterion is based on this, and defined to be the square root of

$$(2.7) \quad \sup_{\{t_0, T: t_0 \leq T\}} \sum_{k=1}^m \sum_{t=t_0}^T \|z_k^{(t_0)}(t)\|^2 = \sup_{\{t_0, T: t_0 \leq T\}} \text{tr} (\mathbf{B}_{t_0}^T \mathbf{X}_{t_0+1}^{(T)} \mathbf{B}_{t_0} + \mathbf{D}_{t_0}^T \mathbf{D}_{t_0}) \\ = \sup_{t_0 \geq 0} \text{tr} (\mathbf{B}_{t_0}^T \mathbf{X}_{t_0+1}^{(\infty)} \mathbf{B}_{t_0} + \mathbf{D}_{t_0}^T \mathbf{D}_{t_0}),$$

where the second equality is due to part (a) of Lemma 2.3; the limit

$$\mathbf{X}_{t_0+1}^{(\infty)} = \lim_{T \rightarrow \infty} \mathbf{X}_{t_0+1}^{(T)}$$

is well-defined for all  $t_0 \geq 0$  (as long as  $\mathbf{G}$  is bounded and  $(\mathcal{G}, \theta)$  is uniformly stable). We shall call the square root of the quantity in (2.7) the *peak impulse response level* of the system  $(\mathcal{G}, \theta)$ . This is a time-varying extension of the deterministic interpretation of the  $\mathcal{H}^2$  norm of linear time-invariant systems, which is defined to be the square root of the sum of the output energies over impulsive inputs.

The second performance criterion extends the stochastic interpretation of the  $\mathcal{H}^2$ -norm as the square root of the output variance under zero-mean white Gaussian disturbance. Define  $(w^{(t_0)}(0), w^{(t_0)}(1), \dots)$  to be an independent identically distributed random sequence such that each  $w^{(t_0)}(t)$  is Gaussian distributed with

$$(2.8) \quad \mathbb{E} [w^{(t_0)}(t)] = 0; \quad \mathbb{E} [w^{(t_0)}(t) w^{(t_0)}(t)^T] = \begin{cases} \mathbf{0}, & t < t_0; \\ \mathbf{I}, & t \geq t_0, \end{cases}$$

for  $t_0, t \geq 0$ , where  $\mathbb{E}[\cdot]$  denotes the expectation over the probability distribution of  $(w^{(t_0)}(0), w^{(t_0)}(1), \dots)$ . Then, with  $x(0) = 0$  and  $\mathbf{w} = (w^{(t_0)}(0), w^{(t_0)}(1), \dots)$ , the

system  $(\mathcal{G}, \theta)$  yields  $\mathbf{z} = (z^{(t_0)}(0), z^{(t_0)}(1), \dots)$ , where

$$(2.9) \quad \mathbb{E} \|z^{(t_0)}(t)\|^2 = \begin{cases} \mathbb{E} \|\mathbf{D}_{t_0} w^{(t_0)}(t_0)\|^2, & t = t_0; \\ \sum_{s=t_0}^{t-1} \mathbb{E} \|\mathbf{C}_t \Phi(t, s+1) \mathbf{B}_s w^{(t_0)}(s)\|^2 + \mathbb{E} \|\mathbf{D}_t w^{(t_0)}(t)\|^2, & t > t_0. \end{cases}$$

From this, we have that

$$(2.10) \quad \sup_{\{t_0, t: t_0 \leq t\}} \mathbb{E} \|z^{(t_0)}(t)\|^2 = \sup_{\{t_0, t: t_0 \leq t\}} \text{tr} (\mathbf{C}_t \mathbf{Y}_t^{(t_0)} \mathbf{C}_t^T + \mathbf{D}_t \mathbf{D}_t^T) \\ = \sup_{t \geq 0} \text{tr} (\mathbf{C}_t \mathbf{Y}_t^{(0)} \mathbf{C}_t^T + \mathbf{D}_t \mathbf{D}_t^T),$$

where the second equality follows from part (a) of Lemma 2.3. The quantity in (2.10) shall be called the *peak output variance level* of the system  $(\mathcal{G}, \theta)$ .

The third performance criterion draws on the interpretation of the  $\mathcal{H}^2$  norm as the square root of the average output variance per unit time. This stochastic interpretation coincides with the deterministic interpretation of the  $\mathcal{H}^2$  norm as the square root of the sum of the average output energies over impulsive inputs. For time-varying systems, we have the following.

PROPOSITION 2.4. *Let  $\mathcal{G}$  and  $\theta$  be as in (2.1); let  $\mathcal{G}$  be bounded. If the system  $(\mathcal{G}, \theta)$  is uniformly exponentially stable, and if  $z_k^{(t_0)}(t)$  and  $z^{(0)}(t)$  are as in (2.6) and (2.9), respectively, then*

$$(2.11) \quad \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t_0=0}^T \sum_{k=1}^m \sum_{t=t_0}^T \|z_k^{(t_0)}(t)\|^2 \\ = \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t_0=0}^T \text{tr} (\mathbf{B}_{t_0}^T \mathbf{X}_{t_0+1}^{(\infty)} \mathbf{B}_{t_0} + \mathbf{D}_{t_0}^T \mathbf{D}_{t_0}) \\ = \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \text{tr} (\mathbf{C}_t \mathbf{Y}_t^{(0)} \mathbf{C}_t^T + \mathbf{D}_t \mathbf{D}_t^T) \\ = \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|z^{(0)}(t)\|^2.$$

*Proof.* Because of (2.7), (2.10), and part (b) of Lemma 2.3, it suffices to show that

$$(2.12) \quad \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t_0=0}^T \text{tr} \mathbf{B}_{t_0}^T \mathbf{X}_{t_0+1}^{(\infty)} \mathbf{B}_{t_0} = \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t_0=0}^T \text{tr} \mathbf{B}_{t_0}^T \mathbf{X}_{t_0+1}^{(T)} \mathbf{B}_{t_0}.$$

Since  $\mathcal{G}$  is bounded and  $(\mathcal{G}, \theta)$  is uniformly stable, there exist  $\tilde{c} \geq 1$  and  $\tilde{\lambda} \in (0, 1)$  such that

$$\text{tr} \mathbf{B}_{t_0}^T \Phi(s, t_0+1)^T \mathbf{C}_s^T \mathbf{C}_s \Phi(s, t_0+1) \mathbf{B}_{t_0} \leq \tilde{c} \tilde{\lambda}^{s-t_0-1},$$

so

$$\begin{aligned}
& \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t_0=0}^T \text{tr} \mathbf{B}_{t_0}^T (\mathbf{X}_{t_0+1}^{(\infty)} - \mathbf{X}_{t_0+1}^{(T)}) \mathbf{B}_{t_0} \\
&= \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t_0=0}^T \sum_{s=T+1}^{\infty} \text{tr} \mathbf{B}_{t_0}^T \Phi(s, t_0+1)^T \mathbf{C}_s^T \mathbf{C}_s \Phi(s, t_0+1) \mathbf{B}_{t_0} \\
&\leq \lim_{T \rightarrow \infty} \frac{1}{T+1} \frac{\tilde{c}(1 - \tilde{\lambda}^{T+1})}{(1 - \tilde{\lambda})^2} = 0.
\end{aligned}$$

This leads to equality (2.12), and hence completes the proof.  $\square$

The quantity in (2.11) shall be called the *average output variance level* of the system  $(\mathcal{G}, \theta)$ . In general, the three performance measures given by (2.7), (2.10), and (2.11) are all different for time-varying systems—see Example 1 below. However, as is well known, they are all equal in the case of time-invariant systems.

*Example 1.* Let  $n = m = l = 1$ . If a linear periodic system is defined by

$$\mathbf{A}_t = 1/\sqrt{2}; \quad \mathbf{B}_t = 1; \quad \mathbf{C}_t = \begin{cases} 1 & \text{if } t \text{ is even;} \\ 0 & \text{if } t \text{ is odd;} \end{cases} \quad \mathbf{D}_t = 0$$

for  $t \geq 0$ , then

$$\mathbf{X}_t^{(\infty)} = \begin{cases} 4/3 & \text{if } t \text{ is even;} \\ 2/3 & \text{if } t \text{ is odd;} \end{cases} \quad \mathbf{Y}_t^{(0)} = \frac{2^t - 1}{2^{t-1}}$$

for  $t \geq 0$ . The quantities in (2.7) and (2.10) are computed as

$$\sup_{t_0 \geq 0} \text{tr} \mathbf{B}_{t_0}^T \mathbf{X}_{t_0+1}^{(\infty)} \mathbf{B}_{t_0} = \max\{4/3, 2/3\} = 4/3,$$

$$\sup_{t \geq 0} \text{tr} \mathbf{C}_t \mathbf{Y}_t^{(0)} \mathbf{C}_t^T = \sup_{t \geq 0} (2^t - 1)/2^{t-1} = 2,$$

and that in (2.11) as

$$\begin{aligned}
\limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t_0=0}^T \text{tr} \mathbf{B}_{t_0}^T \mathbf{X}_{t_0+1}^{(\infty)} \mathbf{B}_{t_0} &= \lim_{T \rightarrow \infty} \frac{1}{T+1} \begin{cases} T + 4/3 & \text{if } T \text{ is even;} \\ T + 1 & \text{if } T \text{ is odd} \end{cases} \\
&= 1
\end{aligned}$$

or

$$\begin{aligned}
\limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \text{tr} \mathbf{C}_t \mathbf{Y}_t^{(0)} \mathbf{C}_t^T &= \lim_{T \rightarrow \infty} \frac{1}{T+1} \begin{cases} T - 2/3 + (2/3)(1/2)^T & \text{if } t \text{ is even;} \\ T - 5/3 + (4/3)(1/2)^T & \text{if } t \text{ is odd} \end{cases} \\
&= 1. \quad \square
\end{aligned}$$

**3. Analysis of switched linear systems.** In this section, we extend the three output regulation properties of time-varying systems to switched systems and derive convex characterizations of the systems that are stable and meet desired performance levels. Given a positive integer  $N$ , let

$$(3.1) \quad \mathcal{G} = \{(\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1, \mathbf{D}_1), \dots, (\mathbf{A}_N, \mathbf{B}_N, \mathbf{C}_N, \mathbf{D}_N)\},$$

with  $\mathbf{A}_i \in \mathbb{R}^{n \times n}$ ,  $\mathbf{B}_i \in \mathbb{R}^{n \times m}$ ,  $\mathbf{C}_i \in \mathbb{R}^{l \times n}$ ,  $\mathbf{D}_i \in \mathbb{R}^{l \times m}$  for  $i = 1, \dots, N$ . Let  $\Omega$  be the set of all infinite sequences in  $\{1, \dots, N\}$ , so that  $\Omega = \{1, \dots, N\}^\infty$ . Each member  $\theta = (\theta(0), \theta(1), \dots)$  of  $\Omega$  is called a *switching sequence*, and  $\theta(t)$  the *mode* at time  $t$ . If  $\Theta$  is a nonempty subset of  $\Omega$ , then the pair  $(\mathcal{G}, \Theta)$  defines the *switched linear system*, which is the collection of linear time-varying systems

$$(3.2) \quad \begin{aligned} x(t+1) &= \mathbf{A}_{\theta(t)}x(t) + \mathbf{B}_{\theta(t)}w(t), \\ z(t) &= \mathbf{C}_{\theta(t)}x(t) + \mathbf{D}_{\theta(t)}w(t) \end{aligned}$$

over all switching sequences  $\theta \in \Theta$ . When the set  $\Theta$  is equal to the entire set  $\Omega$ , the pair  $(\mathcal{G}, \Omega)$  defines the *discrete linear inclusion*, which is the switched linear system without switching path constraint; on the other hand, if  $\Theta = \{\theta\}$  is a singleton, then the pair  $(\mathcal{G}, \Theta)$  is nothing but the linear time-varying system  $(\mathcal{G}, \theta)$  whose parameters vary within the finite set  $\mathcal{G}$  according to  $\theta$ .

Let us first define the stability and performance criteria for switched linear systems. If  $x(0) = 0$ ,  $\mathbf{w} = (w(0), w(1), \dots)$ , and  $\theta = (\theta(0), \theta(1), \dots)$ , then we have  $\mathbf{z} = (z(0), z(1), \dots)$  with

$$z(t) = \begin{cases} \mathbf{D}_{\theta(0)}w(0), & t = 0; \\ \sum_{s=0}^{t-1} \mathbf{C}_{\theta(t)}\Phi_{\theta}(t, s+1)\mathbf{B}_{\theta(s)}w(s) + \mathbf{D}_{\theta(t)}w(t), & t > 0, \end{cases}$$

where

$$\Phi_{\theta}(t, t_0) = \begin{cases} \mathbf{I}, & t = t_0; \\ \mathbf{A}_{\theta(t-1)} \cdots \mathbf{A}_{\theta(t_0)}, & t > t_0. \end{cases}$$

In particular, as in the previous section, if  $\mathbf{w} = (w_k^{(t_0)}(0), w_k^{(t_0)}(1), \dots)$  is an impulsive input sequence where  $w_k^{(t_0)}(t)$  are given by (2.5) for  $t_0, t \geq 0$  and  $1 \leq k \leq m$ , then write  $\mathbf{z} = (z_k^{(t_0)}(0), z_k^{(t_0)}(1), \dots)$ . On the other hand, if  $\mathbf{w} = (w^{(t_0)}(0), w^{(t_0)}(1), \dots)$  is an independent identically distributed random sequence such that  $w^{(t_0)}(t)$  are Gaussian distributed with (2.8) for  $t_0, t \geq 0$ , then write  $\mathbf{z} = (z^{(t_0)}(0), z^{(t_0)}(1), \dots)$ .

**DEFINITION 3.1.** *The system  $(\mathcal{G}, \Theta)$  is said to be uniformly (exponentially) stable if there exist  $c \geq 1$  and  $\lambda \in (0, 1)$  such that, whenever  $\mathbf{w} = 0$ ,*

$$\|x(t)\| \leq c \lambda^{t-t_0} \|x(t_0)\|$$

for all  $t_0 \geq 0, t \geq t_0, x(t_0) \in \mathbb{R}^n$ , and  $\theta \in \Theta$ .

**Remark 2.** By definition, uniformly stable switched linear systems are asymptotically stable (i.e.,  $\|x(t)\| \rightarrow 0$  as  $t \rightarrow \infty$  for each  $\theta \in \Theta$ ). The reverse implication also holds if  $\Theta = \Omega$ : every asymptotically stable discrete linear inclusion is in fact uniformly stable. See, e.g., [34].

DEFINITION 3.2. Let  $\gamma > 0$ . The system  $(\mathcal{G}, \Theta)$  is said to satisfy uniform impulse response level  $\gamma$  if there exists a  $\tilde{\gamma} \in (0, \gamma)$  such that, whenever  $x(0) = 0$ ,

$$\sum_{k=1}^m \sum_{s=t_0}^t \|z_k^{(t_0)}(s)\|^2 \leq \tilde{\gamma}^2$$

for all  $t_0 \geq 0$ ,  $t \geq t_0$ , and  $\theta \in \Theta$ .

DEFINITION 3.3. Let  $\gamma > 0$ . The system  $(\mathcal{G}, \Theta)$  is said to satisfy uniform output regulation level  $\gamma$  if there exists a  $\tilde{\gamma} \in (0, \gamma)$  such that, whenever  $x(0) = 0$ ,

$$(3.3) \quad \mathbb{E} \|z^{(t_0)}(t)\|^2 \leq \tilde{\gamma}^2$$

for all  $t_0 \geq 0$ ,  $t \geq t_0$ , and  $\theta \in \Theta$ , where  $\mathbb{E}$  denotes the expectation with respect to  $\mathbf{w} = (w^{(t_0)}(0), w^{(t_0)}(1), \dots)$ .

DEFINITION 3.4. Let  $\gamma > 0$ . The system  $(\mathcal{G}, \Theta)$  is said to satisfy uniform average output regulation level  $\gamma$  if there exists a  $\tilde{\gamma} \in (0, \gamma)$  such that, whenever  $x(0) = 0$ ,

$$(3.4) \quad \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|z^{(0)}(t)\|^2 \leq \tilde{\gamma}^2$$

for all  $t \geq 0$  and  $\theta \in \Theta$ , where  $\mathbb{E}$  denotes the expectation with respect to  $\mathbf{w} = (w^{(0)}(0), w^{(0)}(1), \dots)$ .

We adopt and extend the notation used in [33] as follows. To help state the analysis results, introduce a dummy mode 0 and think of each  $\theta \in \Theta$  as a two-sided sequence  $(\dots, \theta(-1), \theta(0), \theta(1), \dots)$  by putting  $\theta(t) = 0$  for  $t < 0$ . Given a nonnegative integer  $L$ , each element of the set  $\{0, \dots, N\}^{L+1}$  shall be called a *switching path* of length  $L$ , or simply an *L-path*. For  $\theta \in \Theta$  and  $t \geq 0$ , let

$$\theta_L(t) = (\theta(t-L), \dots, \theta(t)).$$

Each  $\theta \in \Theta$  generates an *L-path switching sequence*  $\theta_L$  defined by

$$\theta_L = (\theta_L(0), \theta_L(1), \dots).$$

Denote the set of *L-paths* occurring in  $\Theta$  by  $\mathcal{L}_L(\Theta)$ , so that

$$\mathcal{L}_L(\Theta) = \{\theta_L(t) : \theta \in \Theta, t \geq 0\}.$$

If  $(i_0, \dots, i_L) \in \mathcal{L}_L(\Theta)$ , then write

$$(i_0, \dots, i_L)_- = (i_0, \dots, i_{L-1}), \quad (i_0, \dots, i_L)_+ = (i_1, \dots, i_L)$$

for  $L > 0$ , and write  $(i_0, \dots, i_L)_- = (i_0, \dots, i_L)_+ = 0$  for  $L = 0$ . If  $L > 0$ , define  $\mathcal{M}_L(\Theta)$  to be the smallest subset of  $\mathcal{L}_L(\Theta)$  such that the following hold:  $\theta_L(t) \in \mathcal{M}_L(\Theta)$  for all  $t \geq L$  and for all  $\theta \in \Theta$ ; and, for each  $j \in \mathcal{L}_0(\Theta)$ , there exists a switching path  $(i_0^j, \dots, i_{L-1}^j) \in \{0, \dots, N\}^L$  such that

$$(3.5) \quad (i_0^{\theta(0)}, \dots, i_{L-1}^{\theta(0)}, \theta(0)), \quad (i_1^{\theta(0)}, \dots, i_{L-1}^{\theta(0)}, \theta(0), \theta(1)), \quad \dots,$$

$$(i_{L-1}^{\theta(0)}, \theta(0), \dots, \theta(L-1)) \in \mathcal{M}_L(\Theta)$$

for all  $\theta \in \Theta$ . If  $L = 0$ , then let  $\mathcal{M}_0(\Theta) = \mathcal{L}_0(\Theta)$ . Define

$$\mathcal{W}_L(\Theta) = \{\theta_L : \theta \in \Theta, t \geq L\} = \mathcal{L}_L(\Theta) \cap \{1, \dots, N\}^{M+1}.$$

Then, in general,

$$\mathcal{W}_L(\Theta) \subset \mathcal{M}_L(\Theta) \subset \mathcal{L}_L(\Theta)$$

for all  $\Theta$  and  $L$ . (While the set  $\mathcal{L}_L(\Theta)$  contains all the  $L$ -paths that occur in  $\Theta$ , the set  $\mathcal{W}_L(\Theta)$  contains only the  $L$ -paths that occur in  $\Theta$  at or after time  $t = L$ . The  $L$ -paths occurring in  $\Theta$  before time  $L$  contain the dummy mode 0, and some of them are redundant in determining the output regulation performance; the set  $\mathcal{M}_L(\Theta)$  is what remains after removing these redundant  $L$ -paths from  $\mathcal{L}_L(\Theta)$ .) Define

$$\mathcal{W}_L^-(\Theta) = \{i_- : i \in \mathcal{W}_L(\Theta)\},$$

$$\mathcal{M}_L^-(\Theta) = \{i_- : i \in \mathcal{M}_L(\Theta)\},$$

$$\mathcal{L}_L^-(\Theta) = \{i_- : i \in \mathcal{L}_L(\Theta)\}.$$

*Example 2.* Let  $N = 3$ . Suppose that  $\Theta$  is the set of all  $\theta \in \Omega$  such that  $\theta(0) \in \{1, 2\}$ ,  $\theta(t+1) = 2$  whenever  $\theta(t) \in \{1, 3\}$ , and  $\theta(t+1) \in \{2, 3\}$  whenever  $\theta(t) = 2$ . Then, writing  $i_0 \cdots i_L$  for  $(i_0, \dots, i_L)$  to simplify notation, we have

$$\mathcal{L}_0(\Theta) = \mathcal{M}_0(\Theta) = \mathcal{W}_0(\Theta) = \{1, 2, 3\}$$

for  $L = 0$ ;

$$\mathcal{L}_1(\Theta) = \{01, 02, 12, 22, 23, 32\},$$

$$\mathcal{M}_1(\Theta) = \{01, 12, 22, 23, 32\},$$

$$\mathcal{W}_1(\Theta) = \{12, 22, 23, 32\}$$

for  $L = 1$ ; and

$$\mathcal{L}_2(\Theta) = \{001, 002, 012, 022, 023, 122, 123, 222, 223, 232, 322, 323\},$$

$$\mathcal{M}_2(\Theta) = \{001, 012, 122, 123, 222, 223, 232, 322, 323\},$$

$$\mathcal{W}_2(\Theta) = \{122, 123, 222, 223, 232, 322, 323\}$$

for  $L = 2$ .  $\square$

**THEOREM 3.5.** *Let  $\mathcal{G}$  be as in (3.1); let  $\Theta \subset \Omega$  be nonempty. The system  $(\mathcal{G}, \Theta)$  is uniformly exponentially stable and satisfies uniform impulse response level  $\gamma > 0$  if and only if there exist a nonnegative integer  $M$  and an indexed family  $\{\mathbf{X}_j : j \in \mathcal{W}_M^-(\Theta)\}$  of symmetric positive definite matrices  $\mathbf{X}_j \in \mathbb{R}^{n \times n}$  such that*

$$(3.6a) \quad \mathbf{A}_{i_0}^T \mathbf{X}_{(i_0, \dots, i_M)_+} \mathbf{A}_{i_0} - \mathbf{X}_{(i_0, \dots, i_M)_-} < -\mathbf{C}_{i_0}^T \mathbf{C}_{i_0},$$

$$(3.6b) \quad \text{tr}(\mathbf{B}_{i_0}^T \mathbf{X}_{(i_0, \dots, i_M)_+} \mathbf{B}_{i_0} + \mathbf{D}_{i_0}^T \mathbf{D}_{i_0}) < \gamma^2$$

for all  $M$ -paths  $(i_0, \dots, i_M) \in \mathcal{W}_M(\Theta)$ .



**THEOREM 3.6.** *Let  $\mathcal{G}$  be as in (3.1); let  $\Theta \subset \Omega$  be nonempty. The system  $(\mathcal{G}, \Theta)$  is uniformly exponentially stable and satisfies uniform output regulation level  $\gamma > 0$  if and only if there exist a nonnegative integer  $M$  and an indexed family  $\{\mathbf{Y}_j : j \in \mathcal{M}_M^-(\Theta)\}$  of symmetric positive definite matrices  $\mathbf{Y}_j \in \mathbb{R}^{n \times n}$  such that*

$$(3.7a) \quad \mathbf{A}_{i_M} \mathbf{Y}_{(i_0, \dots, i_M)_-} \mathbf{A}_{i_M}^T - \mathbf{Y}_{(i_0, \dots, i_M)_+} < -\mathbf{B}_{i_M} \mathbf{B}_{i_M}^T,$$

$$(3.7b) \quad \text{tr}(\mathbf{C}_{i_M} \mathbf{Y}_{(i_0, \dots, i_M)_-} \mathbf{C}_{i_M}^T + \mathbf{D}_{i_M} \mathbf{D}_{i_M}^T) < \gamma^2$$

for all  $M$ -paths  $(i_0, \dots, i_M) \in \mathcal{M}_M(\Theta)$ .

*Proof of Theorems 3.5 and 3.6.* We will first present a complete proof of Theorem 3.6; the proof of Theorem 3.5 will be sketched only briefly. To show the necessity part of Theorem 3.6, suppose that  $(\mathcal{G}, \Theta)$  is uniformly exponentially stable and that there exists a  $\tilde{\gamma} \in (0, \gamma)$  such that (3.3) holds. For  $\varepsilon > 0$ , consider the augmented disturbance signal  $\tilde{w}(t) = [w(t)^T v(t)^T]^T$  with  $v(t) \in \mathbb{R}^n$ ,  $t \geq 0$ , and the perturbed system  $(\mathcal{G}^{(\varepsilon)}, \Theta)$ , where

$$(3.8a) \quad \mathcal{G}^{(\varepsilon)} = \{(\mathbf{A}_1, \mathbf{B}_1^{(\varepsilon)}, \mathbf{C}_1, \mathbf{D}_1^{(\varepsilon)}), \dots, (\mathbf{A}_N, \mathbf{B}_N^{(\varepsilon)}, \mathbf{C}_N, \mathbf{D}_N^{(\varepsilon)})\},$$

$$(3.8b) \quad \mathbf{B}_i^{(\varepsilon)} = [\mathbf{B}_i \sqrt{\varepsilon} \mathbf{I}] \in \mathbb{R}^{n \times (m+n)}, \quad \mathbf{D}_i^{(\varepsilon)} = [\mathbf{D}_i \mathbf{0}] \in \mathbb{R}^{l \times (m+n)}.$$

If  $\mathbf{Y}_{\theta, t}^{(\varepsilon, t_0)}$ ,  $t \geq t_0 \geq 0$ , are such that

$$\mathbf{A}_{\theta(t)} \mathbf{Y}_{\theta, t}^{(\varepsilon, t_0)} \mathbf{A}_{\theta(t)}^T - \mathbf{Y}_{\theta, t+1}^{(\varepsilon, t_0)} = -\mathbf{B}_{\theta(t)}^{(\varepsilon)} \mathbf{B}_{\theta(t)}^{(\varepsilon)T}$$

for  $\theta \in \Theta$  and  $t \geq t_0$ , with  $\mathbf{Y}_{\theta, t_0}^{(\varepsilon, t_0)} = 0$ , then

$$\mathbf{Y}_{\theta, t+1}^{(\varepsilon, t_0)} = \sum_{s=t_0}^t \Phi_{\theta}(t+1, s+1) (\mathbf{B}_{\theta(s)} \mathbf{B}_{\theta(s)}^T + \varepsilon \mathbf{I}) \Phi_{\theta}(t+1, s+1)^T.$$

Hence there exists a sufficiently small  $\varepsilon$ , and a corresponding  $\eta \in (0, \gamma)$  such that

$$\mathbf{A}_{\theta(t)} \mathbf{Y}_{\theta, t}^{(\varepsilon, 0)} \mathbf{A}_{\theta(t)}^T - \mathbf{Y}_{\theta, t+1}^{(\varepsilon, 0)} = -\mathbf{B}_{\theta(t)}^{(\varepsilon)} \mathbf{B}_{\theta(t)}^{(\varepsilon)T},$$

$$\text{tr}(\mathbf{C}_{\theta(t)} \mathbf{Y}_{\theta, t}^{(\varepsilon, 0)} \mathbf{C}_{\theta(t)}^T + \mathbf{D}_{\theta(t)}^{(\varepsilon)} \mathbf{D}_{\theta(t)}^{(\varepsilon)T}) \leq \eta^2$$

for all  $\theta \in \Theta$  and  $t \geq 0$ . Choose a small  $\mathbf{Y} > \mathbf{0}$  such that  $\mathbf{A}_i \mathbf{Y} \mathbf{A}_i^T < \varepsilon \mathbf{I}$  and  $\text{tr}(\mathbf{C}_i \mathbf{Y} \mathbf{C}_i^T) < \gamma^2 - \eta^2$  for all  $i = 1, \dots, N$ , and put

$$\mathbf{Y}_{\theta, t}^{(t_0)} = \begin{cases} \mathbf{Y}, & t = t_0; \\ \mathbf{Y}_{\theta, t}^{(\varepsilon, t_0)}, & t > t_0. \end{cases}$$

Then we have that there are  $\tilde{\varepsilon} > 0$  and  $\tilde{\eta} \in (0, \gamma)$  such that

$$\mathbf{A}_{\theta(t)} \mathbf{Y}_{\theta, t}^{(0)} \mathbf{A}_{\theta(t)}^T - \mathbf{Y}_{\theta, t+1}^{(0)} \leq -\mathbf{B}_{\theta(t)} \mathbf{B}_{\theta(t)}^T - \tilde{\varepsilon} \mathbf{I},$$

$$\text{tr}(\mathbf{C}_{\theta(t)} \mathbf{Y}_{\theta, t}^{(0)} \mathbf{C}_{\theta(t)}^T + \mathbf{D}_{\theta(t)} \mathbf{D}_{\theta(t)}^T) \leq \tilde{\eta}^2$$

for all  $\theta \in \Theta$  and  $t \geq 0$ . Since  $(\mathcal{G}^{(\varepsilon)}, \Theta)$  is uniformly stable, there exist  $\tilde{c} \geq 1$  and  $\tilde{\lambda} \in (0, 1)$ , independent of  $\theta$ , such that

$$\Phi_\theta(t+1, t-M+1)(\mathbf{B}_{\theta(t-M)}\mathbf{B}_{\theta(t-M)}^\top + \varepsilon\mathbf{I})\Phi_\theta(t+1, t-M+1)^\top \leq \tilde{c}\tilde{\lambda}^M\mathbf{I}$$

for  $\theta \in \Theta$  and  $t \geq M \geq 0$ . Choose an  $M > 0$  such that

$$\tilde{c}\tilde{\lambda}^M < \varepsilon.$$

Then

$$\mathbf{A}_{\theta(t)}\mathbf{Y}_{\theta,t}^{(t-M)}\mathbf{A}_{\theta(t)}^\top - \mathbf{Y}_{\theta,t+1}^{(t-M+1)} \leq \tilde{c}\tilde{\lambda}^M - \mathbf{B}_{\theta(t)}\mathbf{B}_{\theta(t)}^\top - \varepsilon\mathbf{I} < -\mathbf{B}_{\theta(t)}\mathbf{B}_{\theta(t)}^\top$$

for  $t \geq M$ . Now, since

$$\mathbf{A}_{\theta(t)}\mathbf{Y}_{\theta,t}^{(0)}\mathbf{A}_{\theta(t)}^\top - \mathbf{Y}_{\theta,t+1}^{(0)} < -\mathbf{B}_{\theta(t)}\mathbf{B}_{\theta(t)}^\top,$$

putting

$$\mathbf{Y}_{(\theta(t-M), \dots, \theta(t-1))} = \begin{cases} \mathbf{Y}_{\theta,t}^{(0)}, & t < M; \\ \mathbf{Y}_{\theta,t}^{(t-M)}, & t \geq M, \end{cases}$$

leads to (3.7a) for all  $(i_0, \dots, i_M) \in \mathcal{L}_M(\Theta)$ , where  $\mathcal{M}_M(\Theta) \subset \mathcal{L}_M(\Theta)$ . Also, it is immediate from part (a) of Lemma 2.3 that (3.7b) holds for all  $(i_0, \dots, i_M) \in \mathcal{M}_M(\Theta)$ .

To show the sufficiency part of Theorem 3.6, suppose that (3.7) holds for some integer  $M \geq 0$  and for all  $(i_0, \dots, i_M) \in \mathcal{M}_M(\Theta)$ . Assume  $M > 0$  without loss of generality. By the definition of  $\mathcal{M}_M(\Theta)$ , one can choose  $(i_0^j, \dots, i_{M-1}^j) \in \{0, \dots, N\}^M$ ,  $j \in \{\theta(0) : \theta \in \Theta\}$ , such that (3.5), with  $L$  replaced by  $M$ , holds for all  $\theta \in \Theta$ . Put

$$\mathbf{Y}_t = \begin{cases} \mathbf{Y}_{(i_0^{\theta(0)}, \dots, i_{M-1}^{\theta(0)})}, & t = 0; \\ \mathbf{Y}_{(i_t^{\theta(0)}, \dots, i_{M-1}^{\theta(0)}, \theta(0), \dots, \theta(t-1))}, & 0 < t < M; \\ \mathbf{Y}_{(\theta(t-M), \dots, \theta(t-1))}, & t \geq M. \end{cases}$$

Then, since  $\mathcal{M}_M(\Theta)$  is finite, one can choose  $\epsilon, \delta > 0$  such that

$$\epsilon\mathbf{I} \leq \mathbf{Y}_t \leq \delta\mathbf{I}; \quad \mathbf{A}_{\theta(t)}\mathbf{Y}_t\mathbf{A}_{\theta(t)}^\top - \mathbf{Y}_{t+1} \leq -\epsilon\mathbf{I}$$

for all  $\theta \in \Theta$  and for all  $t \geq 0$ . Then the system  $(\mathcal{G}, \Theta)$  is uniformly stable due to Lemma 2.2. On the other hand, due to part (c) of Lemma 2.3, we have that

$$\text{tr}(\mathbf{C}_{\theta(t)}\mathbf{Y}_{\theta,t}^{(0)}\mathbf{C}_{\theta(t)}^\top) \leq \text{tr}(\mathbf{C}_{\theta(t)}\mathbf{Y}_t\mathbf{C}_{\theta(t)}^\top)$$

for all  $\theta \in \Theta$ . Hence there exists a  $\tilde{\gamma} \in (0, \gamma)$  such that (3.3) holds. This concludes the proof of Theorem 3.6.

The proof of Theorem 3.5 is analogous to that of Theorem 3.6. In particular, to show sufficiency involves the solution

$$\mathbf{X}_{\theta,t}^{(\varepsilon,T)} = \sum_{s=t}^T \Phi_\theta(s, t)^\top (\mathbf{C}_{\theta(s)}^\top \mathbf{C}_{\theta(s)} + \varepsilon\mathbf{I}) \Phi_\theta(s, t)$$

to

$$\mathbf{A}_{\theta(t)}^T \mathbf{X}_{\theta, t+1}^{(\varepsilon, T)} \mathbf{A}_{\theta(t)} - \mathbf{X}_{\theta, t}^{(\varepsilon, T)} = -\mathbf{C}_{\theta(t)}^T \mathbf{C}_{\theta(t)} - \varepsilon \mathbf{I}$$

for  $\theta \in \Theta$  and  $t \leq T$ , with  $\mathbf{X}_{\theta, T+1}^{(\varepsilon, T)} = 0$ ; for some  $\varepsilon > 0$  and some nonnegative integer  $M$ , we may put

$$\mathbf{X}_{(\theta(t), \dots, \theta(t+M))_+} = \mathbf{X}_{\theta, t+1}^{(\varepsilon, t+M)}$$

for all  $\theta \in \Theta$  and  $t \geq 0$ .  $\square$

To characterize the output regulation performance in terms of the average output variance level, we will consider a special but still quite general case where  $\Theta$  is defined by the directed graph of a matrix  $\mathbf{Q}$  with possible starting nodes given by a row vector  $q$ . Let  $\mathbf{Q} = (q_{ij}) \in \mathbb{R}^{N \times N}$  be a componentwise nonnegative matrix where each row has at least one positive component; that is,  $q_{ij} \geq 0$  for all  $i$  and  $j$ , and  $\sum_{j=1}^N q_{ij} > 0$  for all  $i$ . Similarly, let  $q = (q_i) \in \mathbb{R}^{1 \times N}$  be a componentwise nonnegative row vector with at least one positive component; that is,  $\sum_{i=1}^N q_i > 0$ . Then the pair  $(\mathbf{Q}, q)$  shall be called *row-allowable*. Define

$$\Theta(\mathbf{Q}, q) = \{\theta \in \Omega: q_{\theta(0)} > 0 \text{ and } q_{\theta(t)\theta(t+1)} > 0 \text{ for } t \geq 0\},$$

where  $q_{\theta(0)} = q_i$  if  $\theta(0) = i$  and  $q_{\theta(t)\theta(t+1)} = q_{ij}$  if  $(\theta(t), \theta(t+1)) = (i, j)$ , and write

$$\mathcal{W}_L(\mathbf{Q}, q) = \mathcal{W}_L(\Theta(\mathbf{Q}, q)), \quad \mathcal{W}_L^-(\mathbf{Q}, q) = \mathcal{W}_L^-(\Theta(\mathbf{Q}, q));$$

$$\mathcal{M}_L(\mathbf{Q}, q) = \mathcal{M}_L(\Theta(\mathbf{Q}, q)), \quad \mathcal{M}_L^-(\mathbf{Q}, q) = \mathcal{M}_L^-(\Theta(\mathbf{Q}, q));$$

$$\mathcal{L}_L(\mathbf{Q}, q) = \mathcal{L}_L(\Theta(\mathbf{Q}, q)), \quad \mathcal{L}_L^-(\mathbf{Q}, q) = \mathcal{L}_L^-(\Theta(\mathbf{Q}, q))$$

for nonnegative integers  $L$ .

We need a few definitions adopted from [35]: given a nonnegative integer  $L$  and a row-allowable pair  $(\mathbf{Q}, q)$ , a set  $\mathcal{N} \subset \mathcal{W}_L(\mathbf{Q}, q)$  is said to be a  $(\mathbf{Q}, q)$ -admissible set of  $L$ -paths if, for each  $(i_0, \dots, i_L) \in \mathcal{N}$ , there exist an integer  $K > L$  and a switching path  $(i_{L+1}, \dots, i_K)$  such that  $(i_{K-L}, \dots, i_K) = (i_0, \dots, i_L)$  and  $(i_t, \dots, i_{t+L}) \in \mathcal{N}$  for  $0 \leq t \leq K - L$ . Define  $\mathcal{N}_L(\mathbf{Q}, q)$  to be the union of all  $(\mathbf{Q}, q)$ -admissible sets of  $L$ -paths, and write

$$\mathcal{N}_L^-(\mathbf{Q}, q) = \{i_- : i \in \mathcal{N}_L(\mathbf{Q}, q)\}.$$

If the only  $(\mathbf{Q}, q)$ -admissible set  $\widetilde{\mathcal{N}}$  of  $L$ -paths satisfying  $\widetilde{\mathcal{N}} \subset \mathcal{N}$  is  $\mathcal{N}$  itself, then  $\mathcal{N}$  is said to be a  $(\mathbf{Q}, q)$ -minimal set of  $L$ -paths. It is readily seen that  $\mathcal{N}$  is a  $(\mathbf{Q}, q)$ -admissible set of  $L$ -paths if and only if it is a finite union of  $(\mathbf{Q}, q)$ -minimal sets of  $L$ -paths. Thus the set  $\mathcal{N}_L(\mathbf{Q}, q)$  is obtained by taking the union of all  $(\mathbf{Q}, q)$ -minimal sets. Whenever  $\mathcal{N}$  is a  $(\mathbf{Q}, q)$ -admissible set of  $L$ -paths, there is a periodic switching sequence  $\theta \in \Omega$  such that  $\mathcal{N} = \mathcal{W}_L(\{\theta\})$  and such that the period of  $\theta$  is equal to the cardinality of  $\mathcal{N}$ ; in particular, if  $\mathcal{N}$  is  $(\mathbf{Q}, q)$ -minimal, then such a  $\theta$  is unique up to a time shift and shall be called a  $(\mathbf{Q}, q)$ -minimal switching sequence associated with  $\mathcal{N}$ .

*Example 3.* Let  $N = 3$ . If

$$\mathbf{Q} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad q = [1 \quad 1 \quad 0],$$

then the set  $\Theta(\mathbf{Q}, q)$  is nothing but the set  $\Theta$  considered in Example 2. We will consider the cases of  $L = 0, 1$ , and  $2$ . The  $(\mathbf{Q}, q)$ -minimal sets of zero-paths are  $\{1\}$ ,  $\{2\}$ , and  $\{3\}$ , and the minimal switching sequences associated with them are  $(1, 1, \dots)$ ,  $(2, 2, \dots)$ , and  $(3, 3, \dots)$ , respectively. Note that these switching sequences do not necessarily belong to  $\Theta(\mathbf{Q}, q)$ . The  $(\mathbf{Q}, q)$ -minimal sets of one-paths are  $\{22\}$  and  $\{23, 32\}$ , and associated with them are the minimal switching sequences  $(2, 2, \dots)$  and  $(2, 3, 2, 3, \dots)$ , respectively. Finally, in the case of  $L = 2$ , there are three  $(\mathbf{Q}, q)$ -minimal sets of two-paths, namely,  $\{222\}$ ,  $\{223, 232, 322\}$ , and  $\{232, 323\}$ ; the minimal switching sequences associated with them are  $(2, 2, \dots)$ ,  $(2, 2, 3, 2, 2, 3, \dots)$ , and  $(2, 3, 2, 3, \dots)$ , respectively. We have that

$$\begin{aligned}\mathcal{N}_0(\mathbf{Q}, q) &= \{1, 2, 3\}, \\ \mathcal{N}_1(\mathbf{Q}, q) &= \{22, 23, 32\}, \\ \mathcal{N}_2(\mathbf{Q}, q) &= \{222, 223, 232, 322, 323\}. \quad \square\end{aligned}$$

**THEOREM 3.7.** *Let  $\mathcal{G}$  be as in (3.1); let  $(\mathbf{Q}, q)$  be a row-allowable pair. The system  $(\mathcal{G}, \Theta(\mathbf{Q}, q))$  is uniformly exponentially stable and satisfies uniform average output regulation level  $\gamma > 0$  if and only if there exist a nonnegative integer  $M$  and an indexed family  $\{\mathbf{Y}_j : j \in \mathcal{M}_M^-(\mathbf{Q}, q)\}$  of symmetric positive definite matrices  $\mathbf{Y}_j \in \mathbb{R}^{n \times n}$  such that*

$$(3.9a) \quad \mathbf{A}_{i_M} \mathbf{Y}_{(i_0, \dots, i_M)_-} \mathbf{A}_{i_M}^T - \mathbf{Y}_{(i_0, \dots, i_M)_+} < -\mathbf{B}_{i_M} \mathbf{B}_{i_M}^T$$

for all  $M$ -paths  $(i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{Q}, q)$ , and

$$(3.9b) \quad \frac{1}{N_{\mathcal{N}}} \sum_{(k_0, \dots, k_M) \in \mathcal{N}} \text{tr}(\mathbf{C}_{k_M} \mathbf{Y}_{(k_0, \dots, k_M)_-} \mathbf{C}_{k_M}^T + \mathbf{D}_{k_M} \mathbf{D}_{k_M}^T) < \gamma^2$$

for all  $(\mathbf{Q}, q)$ -minimal sets  $\mathcal{N}$  of  $M$ -paths, where  $N_{\mathcal{N}}$  is the cardinality of  $\mathcal{N}$ .

*Proof.* To show necessity, suppose that  $(\mathcal{G}, \Theta(\mathbf{Q}, q))$  is uniformly stable and that there exists a  $\tilde{\gamma} \in (0, \gamma)$  such that (3.4) holds. For  $\varepsilon > 0$ , consider the perturbed system  $(\mathcal{G}^{(\varepsilon)}, \Theta(\mathbf{Q}, q))$ , where  $\mathcal{G}^{(\varepsilon)}$  is as in (3.8). Write the output that  $(\mathcal{G}^{(\varepsilon)}, \Theta(\mathbf{Q}, q))$  generates under  $x(0) = 0$  and  $\mathbf{w} = (w^{(0)}(0), w^{(0)}(1), \dots)$  as  $\mathbf{z} = (z^{(\varepsilon, 0)}(0), z^{(\varepsilon, 0)}(1), \dots)$ , and let  $\mathbf{Y}_{\theta, t}^{(\varepsilon, 0)} \geq \mathbf{0}$  be the solution to

$$\mathbf{A}_{\theta(t)} \mathbf{Y}_{\theta, t}^{(\varepsilon, 0)} \mathbf{A}_{\theta(t)}^T - \mathbf{Y}_{\theta, t+1}^{(\varepsilon, 0)} = -\mathbf{B}_{\theta(t)}^{(\varepsilon)} \mathbf{B}_{\theta(t)}^{(\varepsilon)T}$$

for  $\theta \in \Theta(\mathbf{Q}, q)$  and  $t \geq 0$ , with the initial condition  $\mathbf{Y}_{\theta, 0}^{(\varepsilon, 0)} = \mathbf{0}$ , so that

$$\limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|z^{(\varepsilon, 0)}(t)\|^2 = \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \text{tr}(\mathbf{C}_{\theta(t)} \mathbf{Y}_{\theta, t}^{(\varepsilon, 0)} \mathbf{C}_{\theta(t)}^T + \mathbf{D}_{\theta(t)}^{(\varepsilon)} \mathbf{D}_{\theta(t)}^{(\varepsilon)T}).$$

Then it follows from the proof of the necessity part of Theorem 3.6 that, for a sufficiently small  $\varepsilon > 0$ , there are a nonnegative integer  $M$  and an indexed family of matrices  $\mathbf{Y}_j > \mathbf{0}$ ,  $j \in \mathcal{L}_M^-(\Theta)$ , such that

$$\mathbf{A}_{i_M} \mathbf{Y}_{(i_0, \dots, i_M)_-} \mathbf{A}_{i_M}^T - \mathbf{Y}_{(i_0, \dots, i_M)_+} < -\mathbf{B}_{i_M} \mathbf{B}_{i_M}^T$$

for  $(i_0, \dots, i_M) \in \mathcal{L}_M(\mathbf{Q}, q)$ , and such that

$$(3.10) \quad \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \text{tr} (\mathbf{C}_{\theta(t)} \mathbf{Y}_{(\theta(t-M), \dots, \theta(t))_-} \mathbf{C}_{\theta(t)}^T + \mathbf{D}_{\theta(t)} \mathbf{D}_{\theta(t)}^T) < \gamma^2$$

for all  $\theta \in \Theta(\mathbf{Q}, q)$ . Since  $\mathcal{M}_M(\mathbf{Q}, q) \subset \mathcal{L}_M(\mathbf{Q}, q)$ , we have that (3.9a) holds for  $(i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{Q}, q)$ . Choose any  $(\mathbf{Q}, q)$ -minimal set  $\mathcal{N}$  of  $M$ -paths and a  $(\mathbf{Q}, q)$ -minimal switching sequence  $\theta$  associated with  $\mathcal{N}$ . If  $\tau_0$  is any integer satisfying  $\tau_0 N_{\mathcal{N}} \geq M$ , where  $N_{\mathcal{N}}$  is the cardinality of  $\mathcal{N}$ , then

$$(3.11) \quad \sum_{t=\tau_0 N_{\mathcal{N}}}^{(\tau_0+1)N_{\mathcal{N}}-1} \text{tr} (\mathbf{C}_{\theta(t)} \mathbf{Y}_{(\theta(t-M), \dots, \theta(t))_-} \mathbf{C}_{\theta(t)}^T + \mathbf{D}_{\theta(t)} \mathbf{D}_{\theta(t)}^T) \\ = \sum_{(k_0, \dots, k_M) \in \mathcal{N}} \text{tr} (\mathbf{C}_{k_M} \mathbf{Y}_{(k_0, \dots, k_M)_-} \mathbf{C}_{k_M}^T + \mathbf{D}_{k_M} \mathbf{D}_{k_M}^T).$$

Since the left-hand side of the established inequality (3.10) is equal to

$$\lim_{\tau \rightarrow \infty} \frac{1}{(\tau - \tau_0)N_{\mathcal{N}}} \sum_{t=\tau_0 N_{\mathcal{N}}}^{\tau N_{\mathcal{N}}-1} \text{tr} (\mathbf{C}_{\theta(t)} \mathbf{Y}_{(\theta(t-M), \dots, \theta(t))_-} \mathbf{C}_{\theta(t)}^T + \mathbf{D}_{\theta(t)} \mathbf{D}_{\theta(t)}^T)$$

for the chosen  $\theta$ , equality (3.11) implies that (3.9b) holds.

Conversely, to show sufficiency, suppose that (3.9) holds for some integer  $M \geq 0$ , for all  $(i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{Q}, q)$ , and for all  $(\mathbf{Q}, q)$ -minimal sets  $\mathcal{N}$  of  $M$ -paths. Then, due to the proof of the sufficiency part of Theorem 3.6, we have that the system  $(\mathcal{G}, \Theta(\mathbf{Q}, q))$  is uniformly stable. Choose any  $\theta \in \Theta(\mathbf{Q}, q)$ . Then, since the set  $\mathcal{L}_M(\mathbf{Q}, q)$  is finite, there exists an  $M$ -path  $(i_0, \dots, i_M) \in \{1, \dots, N\}^{M+1}$  that occurs infinitely many times in  $\theta$ . Let  $(t_1, t_2, \dots)$  be a sequence of time instants with  $t_1 \geq M$  such that  $(\theta(t_j - M), \dots, \theta(t_j)) = (i_0, \dots, i_M)$  for  $j = 1, 2, \dots$ . Then the set of  $M$ -paths

$$(3.12) \quad (\theta(t_j - M), \dots, \theta(t_j)), \quad (\theta(t_j - M + 1), \dots, \theta(t_j + 1)), \quad \dots, \\ (\theta(t_{j+1} - M - 1), \dots, \theta(t_{j+1} - 1))$$

is  $(\mathbf{Q}, q)$ -admissible for each  $j$ . We will show that there exists a  $\tilde{\gamma} \in (0, \gamma)$  such that

$$(3.13) \quad \frac{1}{t_{r+1} - t_1} \sum_{t=t_1}^{t_{r+1}-1} \text{tr} (\mathbf{C}_{\theta(t)} \mathbf{Y}_{(\theta(t-M), \dots, \theta(t))_-} \mathbf{C}_{\theta(t)}^T + \mathbf{D}_{\theta(t)} \mathbf{D}_{\theta(t)}^T) \leq \tilde{\gamma}^2$$

for all  $r = 1, 2, \dots$ , so that part (c) of Lemma 2.3 establishes immediately that  $(\mathcal{G}, \Theta(\mathbf{Q}, q))$  satisfies uniform average output variance level  $\gamma$ . Let  $\mathcal{N}_1, \dots, \mathcal{N}_k$  be all the distinct  $(\mathbf{Q}, q)$ -minimal sets of  $M$ -paths, and denote the cardinality of  $\mathcal{N}_i$  by  $N_i$ . Define

$$c_i = \sum_{(k_0, \dots, k_M) \in \mathcal{N}_i} \text{tr} (\mathbf{C}_{k_M} \mathbf{Y}_{(k_0, \dots, k_M)_-} \mathbf{C}_{k_M}^T + \mathbf{D}_{k_M} \mathbf{D}_{k_M}^T)$$

for  $i = 1, \dots, k$ . For each  $j$  there are nonnegative integers  $\tau_1(j), \dots, \tau_k(j)$  such that the set of  $M$ -paths in (3.12), counting multiplicity, is the union of exactly  $\tau_i(j)$   $\mathcal{N}_i$ 's

over all  $i$ , so that

$$t_{j+1} - t_j = \sum_{i=1}^k \tau_i(j) N_i;$$

otherwise, the condition  $(\theta(t_j - M), \dots, \theta(t_j)) = (\theta(t_{j+1} - M), \dots, \theta(t_{j+1}))$ , which implies that the set of  $M$ -paths in (3.12) is  $(\mathbf{Q}, q)$ -admissible, would not hold. Thus the left-hand side of the desired inequality (3.13) is precisely equal to

$$\frac{1}{\sum_{i=1}^k \sum_{j=1}^r \tau_i(j) N_i} \sum_{i=1}^k \sum_{j=1}^r \tau_i(j) c_i$$

for all  $r = 1, 2, \dots$ . Here, since (3.9b) holds for all  $(\mathbf{Q}, q)$ -minimal sets  $\mathcal{N}$  of  $M$ -paths, there exists a  $\tilde{\gamma} \in (0, \gamma)$  such that  $c_i/N_i \leq \tilde{\gamma}^2$  for all  $i$ . This shows that, indeed, inequality (3.13) holds for all  $r$ .  $\square$

*Remark 3.* The semidefinite program that minimizes  $\gamma^2$  over all path lengths  $M$  subject to the linear matrix inequalities (3.6), (3.7), or (3.9) gives the best performance level that the system satisfies. A potential difficulty is that the computational burden grows exponentially in  $M$  in the worst case. However, the required path length  $M$  that achieves the minimum value of  $\gamma$  up to a reasonable tolerance level is often small. See Examples 5 and 6; see also the remarks and examples in [34, 33].

*Example 4.* Let  $(\mathbf{Q}, q)$ , and hence  $\Theta(\mathbf{Q}, q)$ , be as in Example 3. The condition that (3.9b) holds for all  $(\mathbf{Q}, q)$ -minimal sets  $\mathcal{N}$  of  $M$ -paths, where  $M = 0$  for instance, reads that

$$\text{tr}(\mathbf{C}_i \mathbf{Y}_0 \mathbf{C}_i^T + \mathbf{D}_i \mathbf{D}_i^T) < \gamma^2$$

for all  $i = 1, 2, 3$  and for a single  $\mathbf{Y}_0 > \mathbf{0}$ . If  $M = 1$ , then the condition reads that

$$\text{tr}(\mathbf{C}_2 \mathbf{Y}_2 \mathbf{C}_2^T + \mathbf{D}_2 \mathbf{D}_2^T) < \gamma^2,$$

$$\text{tr}(\mathbf{C}_3 \mathbf{Y}_2 \mathbf{C}_3^T + \mathbf{D}_3 \mathbf{D}_3^T + \mathbf{C}_2 \mathbf{Y}_3 \mathbf{C}_2^T + \mathbf{D}_2 \mathbf{D}_2^T) < 2\gamma^2$$

for some  $\mathbf{Y}_2, \mathbf{Y}_3 > \mathbf{0}$ . Finally, if  $M = 2$ , then the condition becomes that

$$\text{tr}(\mathbf{C}_2 \mathbf{Y}_{(2,2)} \mathbf{C}_2^T + \mathbf{D}_2 \mathbf{D}_2^T) < \gamma^2,$$

$$\text{tr}(\mathbf{C}_3 \mathbf{Y}_{(2,2)} \mathbf{C}_3^T + \mathbf{D}_3 \mathbf{D}_3^T + \mathbf{C}_2 \mathbf{Y}_{(2,3)} \mathbf{C}_2^T + \mathbf{D}_2 \mathbf{D}_2^T + \mathbf{C}_2 \mathbf{Y}_{(3,2)} \mathbf{C}_2^T + \mathbf{D}_2 \mathbf{D}_2^T) < 3\gamma^2,$$

$$\text{tr}(\mathbf{C}_2 \mathbf{Y}_{(2,3)} \mathbf{C}_2^T + \mathbf{D}_2 \mathbf{D}_2^T + \mathbf{C}_3 \mathbf{Y}_{(3,2)} \mathbf{C}_3^T + \mathbf{D}_3 \mathbf{D}_3^T) < 2\gamma^2$$

for some  $\mathbf{Y}_{(2,2)}, \mathbf{Y}_{(2,3)}, \mathbf{Y}_{(3,2)} > \mathbf{0}$ .  $\square$

**4. Control of switched linear systems.** In this section, we will give an exact synthesis condition for controlling the average output variance level. Then we will remark only briefly on the controller synthesis for minimizing the peak impulse response and peak output variance. Consider the set

$$(4.1) \quad \mathcal{T} = \{(\mathbf{A}_i, \mathbf{B}_{1,i}, \mathbf{B}_{2,i}, \mathbf{C}_{1,i}, \mathbf{C}_{2,i}, \mathbf{D}_{11,i}, \mathbf{D}_{12,i}, \mathbf{D}_{21,i}) : i = 1, \dots, N\}$$

with  $\mathbf{A}_i \in \mathbb{R}^{n \times n}$ ,  $\mathbf{B}_{1,i} \in \mathbb{R}^{n \times m_1}$ ,  $\mathbf{B}_{2,i} \in \mathbb{R}^{n \times m_2}$ ,  $\mathbf{C}_{1,i} \in \mathbb{R}^{l_1 \times n}$ ,  $\mathbf{C}_{2,i} \in \mathbb{R}^{l_2 \times n}$ ,  $\mathbf{D}_{11,i} \in \mathbb{R}^{l_1 \times m_1}$ ,  $\mathbf{D}_{12,i} \in \mathbb{R}^{l_1 \times m_2}$ ,  $\mathbf{D}_{21,i} \in \mathbb{R}^{l_2 \times m_1}$  for  $i = 1, \dots, N$ . If  $(\mathbf{Q}, q)$  is a

row-allowable pair, then the pair  $(\mathcal{T}, \Theta(\mathbf{Q}, q))$  defines the *controlled switched linear system* represented by

$$(4.2) \quad \begin{aligned} x(t+1) &= \mathbf{A}_{\theta(t)}x(t) + \mathbf{B}_{1,\theta(t)}w(t) + \mathbf{B}_{2,\theta(t)}u(t), \\ z(t) &= \mathbf{C}_{1,\theta(t)}x(t) + \mathbf{D}_{11,\theta(t)}w(t) + \mathbf{D}_{12,\theta(t)}u(t), \\ y(t) &= \mathbf{C}_{2,\theta(t)}x(t) + \mathbf{D}_{21,\theta(t)}w(t). \end{aligned}$$

Given the initial state  $x(0)$ , disturbance sequence  $\mathbf{w} = (w(0), w(1), \dots)$ , control sequence  $\mathbf{u} = (u(0), u(1), \dots)$ , and switching sequence  $\boldsymbol{\theta} = (\theta(0), \theta(1), \dots) \in \Theta(\mathbf{Q}, q)$ , this system of equations defines the evolution of the state  $x(t)$ , controlled output  $z(t)$ , and measured output  $y(t)$  for  $t \geq 0$ .

We make the standard assumption that *the mode  $\theta(t)$  is perfectly observed at each time instant  $t$* . Also, as in [34, 33], we consider all controllers that have finite memory of past modes and make perfect observation of the present mode. Fix a nonnegative integer  $L$ . Let

$$\Theta_L(\mathbf{Q}, q) = \{\boldsymbol{\theta}_L : \boldsymbol{\theta} \in \Theta(\mathbf{Q}, q)\}$$

be the set of  $L$ -path switching sequences generated by  $(\mathbf{Q}, q)$ ; let

$$\mathcal{K} = \{(\mathbf{A}_{K,i}, \mathbf{B}_{K,i}, \mathbf{C}_{K,i}, \mathbf{D}_{K,i}) : i \in \mathcal{L}_L(\mathbf{Q}, q)\}$$

with  $\mathbf{A}_{K,i} \in \mathbb{R}^{n_K \times n_K}$ ,  $\mathbf{B}_{K,i} \in \mathbb{R}^{n_K \times l_2}$ ,  $\mathbf{C}_{K,i} \in \mathbb{R}^{m_2 \times n_K}$ ,  $\mathbf{D}_{K,i} \in \mathbb{R}^{m_2 \times l_2}$  for  $i \in \mathcal{L}_L(\mathbf{Q}, q)$ . Then the pair  $(\mathcal{K}, \Theta_L(\mathbf{Q}, q))$  defines the  $L$ -path-dependent (linear feedback) controller (of order  $n_K$ ), which determines the control sequence  $\mathbf{u}$  according to

$$(4.3) \quad \begin{aligned} x_K(t+1) &= \mathbf{A}_{K,\theta_L(t)}x_K(t) + \mathbf{B}_{K,\theta_L(t)}y(t), \\ u(t) &= \mathbf{C}_{K,\theta_L(t)}x_K(t) + \mathbf{D}_{K,\theta_L(t)}y(t) \end{aligned}$$

given the initial controller state  $x_K(0)$  and  $L$ -path switching sequence  $\boldsymbol{\theta}_L \in \Theta_L(\mathbf{Q}, q)$ . Controllers that are  $L$ -path dependent for some nonnegative integer  $L$  shall be said to be *finite-path dependent*; zero-path-dependent controllers are called *mode dependent*. The dependence of these controllers on the past measurements  $y(0), \dots, y(t)$  at each time instant  $t$  is encoded in the partition

$$\mathbf{K}_i = \begin{bmatrix} \mathbf{A}_{K,i} & \mathbf{B}_{K,i} \\ \mathbf{C}_{K,i} & \mathbf{D}_{K,i} \end{bmatrix} \in \mathbb{R}^{(n_K+m_2) \times (n_K+l_2)}$$

for  $i \in \mathcal{L}_L(\mathbf{Q}, q)$ .

For our purpose, the set  $\mathcal{K}$  can be replaced by the smaller set

$$\tilde{\mathcal{K}} = \{(\mathbf{A}_{K,i}, \mathbf{B}_{K,i}, \mathbf{C}_{K,i}, \mathbf{D}_{K,i}) : i \in \mathcal{M}_L(\mathbf{Q}, q)\},$$

and the pair  $(\tilde{\mathcal{K}}, \Theta_L(\mathbf{Q}, q))$  shall also be called an  $L$ -path-dependent controller. The reason is that one can always recover  $\mathcal{K}$  from  $\tilde{\mathcal{K}}$ ; this point will become clear later in this section.

Given a finite-path-dependent controller  $(\mathcal{K}, \Theta_L(\mathbf{Q}, q))$ , where  $L$  is the path length, let

$$\begin{aligned} \tilde{\mathbf{A}}_i &= \hat{\mathbf{A}}_{i_L} + \hat{\mathbf{B}}_{2,i_L} \mathbf{K}_i \hat{\mathbf{C}}_{2,i_L}, & \tilde{\mathbf{B}}_i &= \hat{\mathbf{B}}_{1,i_L} + \hat{\mathbf{B}}_{2,i_L} \mathbf{K}_i \hat{\mathbf{D}}_{21,i_L}, \\ \tilde{\mathbf{C}}_i &= \hat{\mathbf{C}}_{1,i_L} + \hat{\mathbf{D}}_{12,i_L} \mathbf{K}_i \hat{\mathbf{C}}_{2,i_L}, & \tilde{\mathbf{D}}_i &= \mathbf{D}_{11,i_L} + \hat{\mathbf{D}}_{12,i_L} \mathbf{K}_i \hat{\mathbf{D}}_{21,i_L} \end{aligned}$$

for  $\mathbf{i} = (i_0, \dots, i_L) \in \mathcal{L}_L(\mathbf{Q}, q)$ , with

$$\begin{aligned}\widehat{\mathbf{A}}_i &= \begin{bmatrix} \mathbf{A}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(n+n_K) \times (n+n_K)}, \\ \widehat{\mathbf{B}}_{1,i} &= \begin{bmatrix} \mathbf{B}_{1,i} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(n+n_K) \times m_1}, \quad \widehat{\mathbf{B}}_{2,i} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{2,i} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(n+n_K) \times (n_K+m_2)}, \\ \widehat{\mathbf{C}}_{1,i} &= [\mathbf{C}_{1,i} \quad \mathbf{0}] \in \mathbb{R}^{l_1 \times (n+n_K)}, \quad \widehat{\mathbf{C}}_{2,i} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{C}_{2,i} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(n_K+l_2) \times (n+n_K)}, \\ \widehat{\mathbf{D}}_{12,i} &= [\mathbf{0} \quad \mathbf{D}_{12,i}] \in \mathbb{R}^{l_1 \times (n_K+m_2)}, \quad \widehat{\mathbf{D}}_{21,i} = \begin{bmatrix} \mathbf{0} \\ \mathbf{D}_{21,i} \end{bmatrix} \in \mathbb{R}^{(n_K+l_2) \times m_1}\end{aligned}$$

for  $i \in \{1, \dots, N\}$ . Let

$$\mathcal{T}_{\mathcal{K}} = \{(\tilde{\mathbf{A}}_{\mathbf{i}}, \tilde{\mathbf{B}}_{\mathbf{i}}, \tilde{\mathbf{C}}_{\mathbf{i}}, \tilde{\mathbf{D}}_{\mathbf{i}}) : \mathbf{i} \in \mathcal{L}_L(\mathbf{Q}, q)\}.$$

If we define the closed-loop state by

$$\tilde{x}(t) = [x(t)^T \ x_K(t)^T]^T \in \mathbb{R}^{n+n_K},$$

then the *closed-loop system*  $(\mathcal{T}_{\mathcal{K}}, \Theta_L(\mathbf{Q}, q))$  has the representation

$$\begin{aligned}(4.4) \quad \tilde{x}(t+1) &= \tilde{\mathbf{A}}_{\theta_L(t)} \tilde{x}(t) + \tilde{\mathbf{B}}_{\theta_L(t)} w(t), \\ z(t) &= \tilde{\mathbf{C}}_{\theta_L(t)} \tilde{x}(t) + \tilde{\mathbf{D}}_{\theta_L(t)} w(t)\end{aligned}$$

for each  $L$ -path switching sequence  $\theta_L \in \Theta_L(\mathbf{Q}, q)$ .

If  $N_L$  is the cardinality of  $\mathcal{L}_L(\mathbf{Q}, q)$ , then there exists a one-to-one correspondence  $f: \mathcal{L}_L(\mathbf{Q}, q) \rightarrow \{1, \dots, N_L\}$ . Label each element of  $\mathcal{L}_L(\mathbf{Q}, q)$  with its image under such an  $f$ . Then each  $L$ -path switching sequence  $\theta_L = (\theta_L(0), \theta_L(1), \dots) \in \Theta_L(\mathbf{Q}, q)$  is mapped to, and hence can be considered to be, a *closed-loop switching sequence* belonging to the set of infinite sequences in  $\{1, \dots, N_L\}$ . On the other hand, if we set  $\theta_L(t) = 0$  for  $t < 0$ , then there is a one-to-one correspondence  $g$  that maps the *closed-loop switching path*  $(\theta_L(t-M), \dots, \theta_L(t))$  to the switching path  $(\theta(t-L-M), \dots, \theta(t))$  of length  $L+M$  for each triple  $(t, L, M)$  of nonnegative integers. In summary, we have the following identities for all integers  $L > 0$  and  $M \geq 0$ :

$$(4.5) \quad \mathcal{M}_M(\Theta_L(\mathbf{Q}, q)) = \mathcal{L}_M(\Theta_L(\mathbf{Q}, q)) = \mathcal{L}_{M+L}(\mathbf{Q}, q).$$

Here, the first identity is due to the mapping  $f$ , and the second identity due to  $g$ . Hence, even if  $L > 0$ , the closed-loop system  $(\mathcal{T}_{\mathcal{K}}, \Theta_L(\mathbf{Q}, q))$  is a switched linear system, where the *closed-loop modes* are the  $L$ -paths in  $\mathcal{L}_L(\mathbf{Q}, q)$ , and the *closed-loop  $M$ -paths* are the  $(M+L)$ -paths in  $\mathcal{L}_{M+L}(\mathbf{Q}, q)$  for each nonnegative integer  $M$ .

**DEFINITION 4.1.** *Let  $\gamma > 0$ . The controller  $(\mathcal{K}, \Theta_L(\mathbf{Q}, q))$  is said to be a  $\gamma$ -admissible ( $L$ -path-dependent) synthesis (of order  $n_K$ ) for the system  $(\mathcal{T}, \Theta(\mathbf{Q}, q))$  if the closed-loop system  $(\mathcal{T}_{\mathcal{K}}, \Theta_L(\mathbf{Q}, q))$  is uniformly exponentially stable and satisfies uniform average output regulation level  $\gamma$ .*



If  $(\mathcal{K}, \Theta_L(\mathbf{Q}, q))$  is an  $\gamma$ -admissible synthesis for  $(\mathcal{T}, \Theta(\mathbf{Q}, q))$ , then it follows from Theorem 3.7 and identity (4.5) that there exist a nonnegative integer  $M \geq L$  and matrices  $\mathbf{Y}_j > \mathbf{0}$  such that

$$\tilde{\mathbf{A}}_{(i_{M-L}, \dots, i_M)} \mathbf{Y}_{(i_0, \dots, i_M)-} \tilde{\mathbf{A}}_{(i_{M-L}, \dots, i_M)}^T - \mathbf{Y}_{(i_0, \dots, i_M)+} < -\tilde{\mathbf{B}}_{(i_{M-L}, \dots, i_M)} \tilde{\mathbf{B}}_{(i_{M-L}, \dots, i_M)}^T$$

for all  $(i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{Q}, q)$ , and

$$\begin{aligned} \frac{1}{N_{\mathcal{N}}} \sum_{(k_0, \dots, k_M) \in \mathcal{N}} \text{tr} \left( \tilde{\mathbf{C}}_{(k_{M-L}, \dots, k_M)} \mathbf{Y}_{(k_0, \dots, k_M)-} \tilde{\mathbf{C}}_{(k_{M-L}, \dots, k_M)}^T \right. \\ \left. + \tilde{\mathbf{D}}_{(k_{M-L}, \dots, k_M)} \tilde{\mathbf{D}}_{(k_{M-L}, \dots, k_M)}^T \right) < \gamma^2 \end{aligned}$$

for all  $(\mathbf{Q}, q)$ -minimal sets  $\mathcal{N}$  of  $M$ -paths, where  $N_{\mathcal{N}}$  is the cardinality of  $\mathcal{N}$ . A Schur complement argument gives that this condition is equivalent to the requirement that

$$(4.6a) \quad \begin{bmatrix} -\mathbf{Y}_{(i_0, \dots, i_M)-}^{-1} & \tilde{\mathbf{A}}_{(i_{M-L}, \dots, i_M)}^T & \mathbf{0} \\ \tilde{\mathbf{A}}_{(i_{M-L}, \dots, i_M)} & -\mathbf{Y}_{(i_0, \dots, i_M)+} & \tilde{\mathbf{B}}_{(i_{M-L}, \dots, i_M)} \\ \mathbf{0} & \tilde{\mathbf{B}}_{(i_{M-L}, \dots, i_M)}^T & -\mathbf{I} \end{bmatrix} < \mathbf{0},$$

$$(4.6b) \quad \begin{bmatrix} -\mathbf{Y}_{(j_0, \dots, j_M)-}^{-1} & \tilde{\mathbf{C}}_{(j_{M-L}, \dots, j_M)}^T & \mathbf{0} \\ \tilde{\mathbf{C}}_{(j_{M-L}, \dots, j_M)} & -\mathbf{Z}_{(j_0, \dots, j_M)} & \tilde{\mathbf{D}}_{(j_{M-L}, \dots, j_M)} \\ \mathbf{0} & \tilde{\mathbf{D}}_{(j_{M-L}, \dots, j_M)}^T & -\mathbf{I} \end{bmatrix} < \mathbf{0},$$

and  $\sum_{(k_0, \dots, k_M) \in \mathcal{N}} \text{tr} \mathbf{Z}_{(k_0, \dots, k_M)} < \gamma^2 N_{\mathcal{N}}$  for all  $(i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{Q}, q)$ , for all  $(j_0, \dots, j_M) \in \mathcal{N}_M(\mathbf{Q}, q)$ , and for all  $(\mathbf{Q}, q)$ -minimal sets  $\mathcal{N}$  of  $M$ -paths. Partition  $\mathbf{Y}_j^{-1}$  and  $\mathbf{Y}_j$ ,  $j \in \mathcal{M}_M(\mathbf{Q}, q)$ , as

$$\mathbf{Y}_j^{-1} = \begin{bmatrix} \mathbf{S}_j & \mathbf{U}_j \\ \mathbf{U}_j^T & * \end{bmatrix}, \quad \mathbf{Y}_j = \begin{bmatrix} \mathbf{R}_j & \mathbf{T}_j \\ \mathbf{T}_j^T & * \end{bmatrix},$$

where  $\mathbf{S}_j, \mathbf{R}_j \in \mathbb{R}^{n \times n}$  and  $\mathbf{U}_j, \mathbf{T}_j \in \mathbb{R}^{n \times n_K}$ . Adopting the change of variable technique in [43], define

$$(4.7) \quad \begin{bmatrix} \mathbf{W}_{11, \mathbf{i}} & \mathbf{W}_{12, \mathbf{i}} \\ \mathbf{W}_{21, \mathbf{i}} & \mathbf{W}_{22, \mathbf{i}} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{\mathbf{i}_+} \mathbf{A}_{i_M} \mathbf{R}_{\mathbf{i}_-} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{U}_{\mathbf{i}_+} & \mathbf{S}_{\mathbf{i}_+} \mathbf{B}_{2, i_M} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \\ \times \begin{bmatrix} \mathbf{A}_{K, (i_{M-L}, \dots, i_M)} & \mathbf{B}_{K, (i_{M-L}, \dots, i_M)} \\ \mathbf{C}_{K, (i_{M-L}, \dots, i_M)} & \mathbf{D}_{K, (i_{M-L}, \dots, i_M)} \end{bmatrix} \begin{bmatrix} \mathbf{T}_{\mathbf{i}_-}^T & \mathbf{0} \\ \mathbf{C}_{2, i_M} \mathbf{R}_{\mathbf{i}_-} & \mathbf{I} \end{bmatrix}$$

for  $\mathbf{i} = (i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{Q}, q)$ . Let

$$\mathbf{M}_{X, \mathbf{i}_+} = \begin{bmatrix} \mathbf{I} & \mathbf{S}_{\mathbf{i}_+} \\ \mathbf{0} & \mathbf{U}_{\mathbf{i}_+}^T \end{bmatrix}, \quad \mathbf{M}_{Y, \mathbf{i}_-} = \begin{bmatrix} \mathbf{I} & \mathbf{R}_{\mathbf{i}_-} \\ \mathbf{0} & \mathbf{T}_{\mathbf{i}_-}^T \end{bmatrix}$$

for all  $\mathbf{i} \in \mathcal{M}_M(\mathbf{Q}, q)$ . If  $n_K = n$ , and if  $\mathbf{U}_{\mathbf{i}_+}$  and  $\mathbf{T}_{\mathbf{i}_-}$  are both invertible, then the change of variable given by (4.7) together with the congruence transformation with

$\mathbf{M}_{Y,i_-} \oplus \mathbf{M}_{X,i_+} \oplus \mathbf{I}$  and  $\mathbf{M}_{Y,i_-} \oplus \mathbf{I} \oplus \mathbf{I}$  on inequalities (4.6a) and (4.6b), respectively, leads to linear matrix inequalities, and we obtain the following result.

**THEOREM 4.2.** *Let  $\mathcal{T}$  be as in (4.1); let  $(\mathbf{Q}, q)$  be a row-allowable pair. Suppose that  $n_K \geq n$  and  $\gamma > 0$ . There exists a  $\gamma$ -admissible finite-path-dependent synthesis of order  $n_K$  for the system  $(\mathcal{T}, \Theta(\mathbf{Q}, q))$  if and only if there exist a nonnegative integer  $M$  and indexed families  $\{(\mathbf{R}_j, \mathbf{S}_j) : j \in \mathcal{M}_M^-(\mathbf{Q}, q)\}$  and  $\{(\mathbf{Z}_i, \mathbf{W}_i) : i \in \mathcal{N}_M(\mathbf{Q}, q)\}$  of symmetric matrices  $\mathbf{R}_j \in \mathbb{R}^{n \times n}$ ,  $\mathbf{S}_j \in \mathbb{R}^{n \times n}$ ,  $\mathbf{Z}_i \in \mathbb{R}^{l_1 \times l_1}$  and rectangular matrices  $\mathbf{W}_i \in \mathbb{R}^{(n+m_2) \times (n+l_2)}$  such that*

$$(4.8a) \quad \mathbf{H}_i + \mathbf{F}_{i_M}^T \mathbf{W}_i \mathbf{G}_{i_M} + \mathbf{G}_{i_M}^T \mathbf{W}_i^T \mathbf{F}_{i_M} < \mathbf{0}$$

for all  $M$ -paths  $i = (i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{Q}, q)$ ,

$$(4.8b) \quad \hat{\mathbf{H}}_i + \hat{\mathbf{F}}_{i_M}^T \mathbf{W}_i \hat{\mathbf{G}}_{i_M} + \hat{\mathbf{G}}_{i_M}^T \mathbf{W}_i^T \hat{\mathbf{F}}_{i_M} < \mathbf{0}$$

for all  $M$ -paths  $i = (i_0, \dots, i_M) \in \mathcal{N}_M(\mathbf{Q}, q)$ , and

$$(4.8c) \quad \frac{1}{N_{\mathcal{N}}} \sum_{(k_0, \dots, k_M) \in \mathcal{N}} \text{tr } \mathbf{Z}_{(k_0, \dots, k_M)} < \gamma^2$$

for all  $(\mathbf{Q}, q)$ -minimal sets  $\mathcal{N}$  of  $M$ -paths, where  $N_{\mathcal{N}}$  is the cardinality of  $\mathcal{N}$ , and

$$\mathbf{H}_i = \begin{bmatrix} -\mathbf{S}_{i_-} & -\mathbf{I} & \mathbf{A}_{i_M}^T & \mathbf{A}_{i_M}^T \mathbf{S}_{i_+} & \mathbf{0} \\ -\mathbf{I} & -\mathbf{R}_{i_-} & \mathbf{R}_{i_-} \mathbf{A}_{i_M}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{A}_{i_M} & \mathbf{A}_{i_M} \mathbf{R}_{i_-} & -\mathbf{R}_{i_+} & -\mathbf{I} & \mathbf{B}_{1,i_M} \\ \mathbf{S}_{i_+} \mathbf{A}_{i_M} & \mathbf{0} & -\mathbf{I} & -\mathbf{S}_{i_+} & \mathbf{S}_{i_+} \mathbf{B}_{1,i_M} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{1,i_M}^T & \mathbf{B}_{1,i_M}^T \mathbf{S}_{i_+} & -\mathbf{I} \end{bmatrix},$$

$$\mathbf{F}_{i_M} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{2,i_M}^T & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{G}_{i_M} = \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{C}_{2,i_M} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{D}_{21,i_M} \end{bmatrix},$$

$$\hat{\mathbf{H}}_i = \begin{bmatrix} -\mathbf{S}_{i_-} & -\mathbf{I} & \mathbf{C}_{1,i_M}^T & \mathbf{0} \\ -\mathbf{I} & -\mathbf{R}_{i_-} & \mathbf{R}_{i_-} \mathbf{C}_{1,i_M}^T & \mathbf{0} \\ \mathbf{C}_{1,i_M} & \mathbf{C}_{1,i_M} \mathbf{R}_{i_-} & -\mathbf{Z}_i & \mathbf{D}_{11,i_M} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_{11,i_M}^T & -\mathbf{I} \end{bmatrix},$$

$$\hat{\mathbf{F}}_{i_M} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_{12,i_M}^T & \mathbf{0} \end{bmatrix}, \quad \hat{\mathbf{G}}_{i_M} = \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{C}_{2,i_M} & \mathbf{0} & \mathbf{0} & \mathbf{D}_{21,i_M} \end{bmatrix}$$

for all  $i = (i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{Q}, q)$ . Moreover, if this condition is satisfied, then given any nonsingular matrices  $\mathbf{T}_j, \mathbf{U}_j \in \mathbb{R}^{n \times n}$  such that

$$(4.9) \quad \mathbf{T}_j \mathbf{U}_j^T = \mathbf{I} - \mathbf{R}_j \mathbf{S}_j$$

for all  $j \in \mathcal{M}_M^-(\mathbf{Q}, q)$ , and

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{W}_{11,i} & \mathbf{W}_{12,i} \\ \mathbf{W}_{21,i} & \mathbf{W}_{22,i} \end{bmatrix}$$

with  $\mathbf{W}_{11,\mathbf{i}} \in \mathbb{R}^{n \times n}$ ,  $\mathbf{W}_{12,\mathbf{i}} \in \mathbb{R}^{n \times l_2}$ ,  $\mathbf{W}_{21,\mathbf{i}} \in \mathbb{R}^{m_2 \times n}$ , and  $\mathbf{W}_{22,\mathbf{i}} \in \mathbb{R}^{m_2 \times l_2}$ , a  $\gamma$ -admissible  $M$ -path-dependent synthesis of order  $n$  is obtained by solving (4.7), with  $n_K = n$  and  $L = M$ , for matrices  $\mathbf{A}_{K,\mathbf{i}}$ ,  $\mathbf{B}_{K,\mathbf{i}}$ ,  $\mathbf{C}_{K,\mathbf{i}}$ ,  $\mathbf{D}_{K,\mathbf{i}}$ ,  $\mathbf{i} \in \mathcal{M}_M(\mathbf{Q}, q)$ .

*Proof.* The proof is essentially the same as the change of variables argument in [43, section IV-B] once we note that

$$\begin{aligned} \mathbf{M}_{X,\mathbf{i}_+}^T \mathbf{Y}_{\mathbf{i}_+} \mathbf{M}_{X,\mathbf{i}_+} &= \begin{bmatrix} \mathbf{R}_{\mathbf{i}_+} & \mathbf{I} \\ \mathbf{I} & \mathbf{S}_{\mathbf{i}_+} \end{bmatrix}, \quad \mathbf{M}_{Y,\mathbf{i}_-}^T \mathbf{Y}_{\mathbf{i}_-}^{-1} \mathbf{M}_{Y,\mathbf{i}_-} = \begin{bmatrix} \mathbf{S}_{\mathbf{i}_-} & \mathbf{I} \\ \mathbf{I} & \mathbf{R}_{\mathbf{i}_-} \end{bmatrix}, \\ \mathbf{M}_{X,\mathbf{i}_+}^T \tilde{\mathbf{A}}_{\mathbf{i}} \mathbf{M}_{Y,\mathbf{i}_-} &= \begin{bmatrix} \mathbf{A}_{i_M} + \mathbf{B}_{2,i_M} \mathbf{W}_{22,\mathbf{i}} \mathbf{C}_{2,i_M} & \mathbf{A}_{i_M} \mathbf{R}_{\mathbf{i}_-} + \mathbf{B}_{2,i_M} \mathbf{W}_{21,\mathbf{i}} \\ \mathbf{S}_{\mathbf{i}_+} \mathbf{A}_{i_M} + \mathbf{W}_{12,\mathbf{i}} \mathbf{C}_{2,i_M} & \mathbf{W}_{11,\mathbf{i}} \end{bmatrix}, \\ \mathbf{M}_{X,\mathbf{i}_+}^T \tilde{\mathbf{B}}_{\mathbf{i}} &= \begin{bmatrix} \mathbf{B}_{1,i_M} + \mathbf{B}_{2,i_M} \mathbf{W}_{22,\mathbf{i}} \mathbf{D}_{21,i_M} \\ \mathbf{S}_{\mathbf{i}_+} \mathbf{B}_{1,i_M} + \mathbf{W}_{12,\mathbf{i}} \mathbf{D}_{21,i_M} \end{bmatrix}, \\ \tilde{\mathbf{C}}_{\mathbf{i}} \mathbf{M}_{Y,\mathbf{i}_-} &= [\mathbf{C}_{1,i_M} + \mathbf{D}_{12,i_M} \mathbf{W}_{22,\mathbf{i}} \mathbf{C}_{2,i_M} \quad \mathbf{C}_{1,i_M} \mathbf{R}_{\mathbf{i}_-} + \mathbf{D}_{12,i_M} \mathbf{W}_{21,\mathbf{i}}], \\ \tilde{\mathbf{D}}_{\mathbf{i}} &= \mathbf{D}_{11,i_M} + \mathbf{D}_{12,i_M} \mathbf{W}_{22,\mathbf{i}} \mathbf{D}_{21,i_M}. \quad \square \end{aligned}$$

*Remark 4.* Since stability and average output regulation performance do not depend on the change in dynamics during any finite number of time instants, the set  $\mathcal{M}_M(\mathbf{Q}, q)$  in Theorem 4.2 can be replaced with  $\mathcal{N}_M(\mathbf{Q}, q)$ . However, using the bigger set  $\mathcal{M}_M(\mathbf{Q}, q)$  is convenient because it allows the control objective to be replaced or mixed with other performance measures such as the uniform output variance level (in Theorem 3.6) and the uniform disturbance attenuation level (considered in [33]), where every single time instant matters.

*Remark 5.* Theorem 3.6 suggests that the synthesis condition for the uniform output variance level control is similar to Theorem 4.2: let  $\gamma > 0$ ,  $n_K \geq n$ , and  $\Theta \subset \Omega$  be nonempty; there exists a finite-path-dependent controller for  $(\mathcal{T}, \Theta)$  such that the closed-loop system is uniformly exponentially stable and satisfies uniform output regulation level  $\gamma$  if and only if there exist a nonnegative integer  $M$  and indexed families  $\{(\mathbf{R}_j, \mathbf{S}_j) : j \in \mathcal{M}_M^-(\Theta)\}$  and  $\{(\mathbf{Z}_i, \mathbf{W}_i) : i \in \mathcal{N}_M(\Theta)\}$  of symmetric matrices  $\mathbf{R}_j$ ,  $\mathbf{S}_j$ ,  $\mathbf{Z}_i$  and rectangular matrices  $\mathbf{W}_i$  such that (4.8a) and  $\text{tr } \mathbf{Z}_i < \gamma^2$  hold for all  $M$ -paths  $\mathbf{i} = (i_0, \dots, i_M) \in \mathcal{M}_M(\Theta)$ , and (4.8b) holds for all  $M$ -paths  $\mathbf{i} = (i_0, \dots, i_M) \in \mathcal{N}_M(\Theta)$ . On the other hand, a condition analogous to this would lead to a synthesis condition for the uniform impulse response level control; however, Theorem 3.5 suggests that if  $M > 0$  in this case, the resulting controller would be required to anticipate future modes as well as observe the present mode.

Suppose that we have obtained a set of controller matrices  $\mathbf{K}_i$ ,  $i \in \mathcal{M}_L(\mathbf{Q}, q)$ , by applying Theorem 4.2. If  $L = 0$ , then it follows from  $\mathcal{M}_0(\mathbf{Q}, q) = \mathcal{L}_0(\mathbf{Q}, q)$  that we have all the matrices  $\mathbf{K}_i$ ,  $i \in \mathcal{L}_0(\mathbf{Q}, q)$ , that define an admissible mode-dependent controller synthesis. If  $L > 0$ , on the other hand, then choose switching paths  $(i_0^j, \dots, i_{L-1}^j) \in \{0, \dots, N\}^L$ ,  $j \in \mathcal{L}_0(\mathbf{Q}, q)$ , such that (3.5) holds for all  $\theta \in \Theta(\mathbf{Q}, q)$ , and put

$$\mathbf{K}_{\theta_L(t)} = \mathbf{K}_{(i_t^{\theta(0)}, \dots, i_{L-1}^{\theta(0)}, \theta(0), \dots, \theta(t))}$$

whenever  $\theta_L \in \Theta_L(\mathbf{Q}, q)$ ,  $t < L$ , and  $\theta_L(t) \notin \mathcal{M}_L(\mathbf{Q}, q)$ ; then we recover all matrices  $\mathbf{K}_i$ ,  $i \in \mathcal{L}_L(\mathbf{Q}, q)$ , that define an admissible  $L$ -path-dependent controller synthesis.

*Example 5.* Let  $\mathcal{T}$  be with  $N = 3$  and

$$\begin{aligned} \mathbf{A}_1 &= 0.5, \quad \mathbf{A}_2 = 1, \quad \mathbf{A}_3 = 0.5; \\ \mathbf{B}_{1,1} &= \mathbf{B}_{1,2} = \mathbf{B}_{1,3} = \begin{bmatrix} 1 & 0 \end{bmatrix}; \quad \mathbf{B}_{2,1} = 0, \quad \mathbf{B}_{2,2} = 1, \quad \mathbf{B}_{2,3} = 0; \\ \mathbf{C}_{1,1} &= \mathbf{C}_{1,2} = \mathbf{C}_{1,3} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}; \quad \mathbf{C}_{2,1} = \mathbf{C}_{2,2} = \mathbf{C}_{2,3} = 1; \\ \mathbf{D}_{11,1} &= \mathbf{D}_{11,2} = \mathbf{D}_{11,3} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}; \quad \mathbf{D}_{12,1} = \mathbf{D}_{12,2} = \mathbf{D}_{12,3} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}; \\ \mathbf{D}_{21,1} &= \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \mathbf{D}_{21,2} = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad \mathbf{D}_{21,3} = \begin{bmatrix} 0 & 1 \end{bmatrix}. \end{aligned}$$

Let

$$\mathbf{Q} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad q = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}.$$

For nonnegative integers  $M$ , denote by  $\gamma_M$  the infimum of  $\gamma > 0$  such that (4.8) holds for all  $M$ -paths  $\mathbf{i} = (i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{Q}, q)$  and for all  $(\mathbf{Q}, q)$ -minimal sets  $\mathcal{N}$  of  $M$ -paths. Then we find by numerical computation that  $1.4953 < \gamma_0 < 1.4954$  and  $1.2395 < \gamma_M < 1.2396$  for all  $M > 0$ . Thus the optimal achievable uniform average output regulation level for the switched system  $(\mathcal{T}, \Theta(\mathbf{Q}, q))$  is  $\gamma^* = 1.2396$  (up to the fourth digit below the decimal point). In particular, if we choose  $M = 1$ , then

$$\mathcal{M}_1(\mathbf{Q}, q) = \{11, 12, 21, 33\},$$

$$\mathcal{N}_1(\mathbf{Q}, q) = \mathcal{N}_1^{(1)} \cup \mathcal{N}_1^{(2)} \cup \mathcal{N}_1^{(3)} = \{11, 12, 21, 33\},$$

where

$$\mathcal{N}_1^{(1)} = \{11\}, \quad \mathcal{N}_1^{(2)} = \{12, 21\}, \quad \mathcal{N}_1^{(3)} = \{33\}$$

are the distinct  $(\mathbf{Q}, q)$ -minimal sets of one-paths; solving (4.8) with  $M = 1$  and  $\gamma = \gamma^*$  yields

$$\begin{aligned} \mathbf{R}_1 &= 1.3335, \quad \mathbf{R}_2 = 1.3339, \quad \mathbf{R}_3 = 1.4086; \\ \mathbf{S}_1 &= 7.5795, \quad \mathbf{S}_2 = 0.89526, \quad \mathbf{S}_3 = 0.83913; \\ \mathbf{Z}_{(1,1)} &= \begin{bmatrix} 1.4012 & 0 \\ 0 & 0.067705 \end{bmatrix}, \quad \mathbf{Z}_{(1,2)} = \begin{bmatrix} 1.3335 & -0.70268 \\ -0.70268 & 0.40576 \end{bmatrix}, \\ \mathbf{Z}_{(2,1)} &= \begin{bmatrix} 1.3339 & 0 \\ 0 & 0.000020022 \end{bmatrix}, \quad \mathbf{Z}_{(3,3)} = \begin{bmatrix} 1.4513 & 0 \\ 0 & 0.042656 \end{bmatrix}; \end{aligned}$$

and

$$\mathbf{W}_{(1,1)} = \begin{bmatrix} -0.0086375 & -6.5753 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{W}_{(1,2)} = \begin{bmatrix} 0.10530 & -0.10426 \\ -0.70268 & -0.067303 \end{bmatrix},$$

$$\mathbf{W}_{(2,1)} = \begin{bmatrix} -1.9995 & -5.5798 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{W}_{(3,3)} = \begin{bmatrix} 0.30236 & -0.18147 \\ 0 & 0 \end{bmatrix}.$$

Plugging  $\mathbf{U}_i = (\mathbf{S}_i - \mathbf{R}_i^{-1})^{1/2}$ ,  $\mathbf{T}_i = -\mathbf{R}_i \mathbf{U}_i$ , and  $L = M$  into (4.7) leads to a  $\gamma^*$ -admissible one-path-dependent synthesis of order one; the resulting controller matrices are

$$\mathbf{K}_{(1,1)} = \begin{bmatrix} -0.40691 & -2.5160 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{K}_{(1,2)} = \begin{bmatrix} 0.30141 & -0.11534 \\ 0.17589 & -0.067303 \end{bmatrix},$$

$$\mathbf{K}_{(2,1)} = \begin{bmatrix} -0.29192 & -2.1351 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{K}_{(3,3)} = \begin{bmatrix} 0.18146 & -0.50482 \\ 0 & 0 \end{bmatrix}.$$

If we choose  $i_0^1 = 1$  and  $i_0^3 = 3$ , then this controller is implemented as follows: at  $t = 0$ , use  $\mathbf{K}_{(1,1)}$  if  $\theta(0) = 1$ , and  $\mathbf{K}_{(3,3)}$  if  $\theta(0) = 3$ ; for  $t > 0$ , use  $\mathbf{K}_{(\theta(t-1), \theta(t))}$ .

If, for some reason, a two-path-dependent controller is desired, then the minimal sets of two-paths are

$$\mathcal{N}_2^{(1)} = \{111\}, \quad \mathcal{N}_2^{(2)} = \{112, 121, 211\}, \quad \mathcal{N}_2^{(3)} = \{121, 212\}, \quad \mathcal{N}_2^{(4)} = \{333\},$$

from which one can obtain controller matrices  $\mathbf{K}_{(1,1,1)}$ ,  $\mathbf{K}_{(1,1,2)}$ ,  $\mathbf{K}_{(1,2,1)}$ ,  $\mathbf{K}_{(2,1,1)}$ ,  $\mathbf{K}_{(2,1,2)}$ , and  $\mathbf{K}_{(3,3,3)}$  proceeding similarly to the case of  $M = 1$ . Choosing  $(i_1^1, i_0^1) = (1, 1)$  and  $(i_1^3, i_0^3) = (3, 3)$  yields the following controller implementation: at  $t = 0$ , use  $\mathbf{K}_{(1,1,1)}$  if  $\theta(0) = 1$ , and  $\mathbf{K}_{(3,3,3)}$  if  $\theta(0) = 3$ ; at  $t = 1$ , use  $\mathbf{K}_{(1,1,1)}$  if  $(\theta(0), \theta(1)) = (1, 1)$ ,  $\mathbf{K}_{(1,1,2)}$  if  $(\theta(0), \theta(1)) = (1, 2)$ , and  $\mathbf{K}_{(3,3,3)}$  if  $(\theta(0), \theta(1)) = (3, 3)$ ; for  $t \geq 2$ , use  $\mathbf{K}_{(\theta(t-2), \theta(t-1), \theta(t))}$ .  $\square$

For the sake of completeness, we conclude this section by presenting a result for the state feedback control under perfect observation of both the state and mode.

**THEOREM 4.3.** *Let  $\mathcal{T}$  be as in (4.1) with  $\mathbf{C}_{2,i} = \mathbf{I}$  and  $\mathbf{D}_{21,i} = \mathbf{0}$  for all  $i$ ; let  $(\mathbf{Q}, q)$  be a row-allowable pair. Suppose that  $\gamma > 0$ . There exists a  $\gamma$ -admissible finite-path-dependent synthesis of order  $n_K = 0$  for the system  $(\mathcal{T}, \Theta(\mathbf{Q}, q))$  if and only if there exist a nonnegative integer  $M$  and indexed families  $\{\mathbf{Y}_j : j \in \mathcal{M}_M^-(\mathbf{Q}, q)\}$  and  $\{(\mathbf{Z}_i, \mathbf{W}_i) : i \in \mathcal{N}_M(\mathbf{Q}, q)\}$  of symmetric matrices  $\mathbf{Y}_j \in \mathbb{R}^{n \times n}$ ,  $\mathbf{Z}_i \in \mathbb{R}^{l_1 \times l_1}$  and rectangular matrices  $\mathbf{W}_i \in \mathbb{R}^{m_2 \times n}$  such that*

$$\begin{bmatrix} -\mathbf{Y}_{i_-} & \mathbf{Y}_{i_-} \mathbf{A}_{i_M}^T + \mathbf{W}_i^T \mathbf{B}_{2,i_M}^T \\ \mathbf{A}_{i_M} \mathbf{Y}_{i_-} + \mathbf{B}_{2,i_M} \mathbf{W}_i & \mathbf{B}_{1,i_M} \mathbf{B}_{1,i_M}^T - \mathbf{Y}_{i_+} \end{bmatrix} < \mathbf{0}$$

for all  $M$ -paths  $i = (i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{Q}, q)$ ,

$$\begin{bmatrix} -\mathbf{Y}_{i_-} & \mathbf{Y}_{i_-} \mathbf{C}_{1,i_M}^T + \mathbf{W}_i^T \mathbf{D}_{12,i_M}^T \\ \mathbf{C}_{1,i_M} \mathbf{Y}_{i_-} + \mathbf{D}_{12,i_M} \mathbf{W}_i & \mathbf{D}_{11,i_M} \mathbf{D}_{11,i_M}^T - \mathbf{Z}_i \end{bmatrix} < \mathbf{0}$$

for all  $M$ -paths  $i = (i_0, \dots, i_M) \in \mathcal{N}_M(\mathbf{Q}, q)$ , and

$$\frac{1}{N_{\mathcal{N}}} \sum_{(k_0, \dots, k_M) \in \mathcal{N}} \text{tr} \mathbf{Z}_{(k_0, \dots, k_M)} < \gamma^2$$

for all  $(\mathbf{Q}, q)$ -minimal sets  $\mathcal{N}$  of  $M$ -paths, where  $N_{\mathcal{N}}$  is the cardinality of  $\mathcal{N}$ . Moreover, if this condition holds, then a  $\gamma$ -admissible  $M$ -path-dependent synthesis of order zero has

$$\mathbf{D}_{K,i} = \mathbf{W}_i \mathbf{Y}_{i_-}^{-1}$$

for all  $i \in \mathcal{M}_M(\mathbf{Q}, q)$ .

*Proof.* Once the inequalities in (4.6) are congruence transformed with  $\mathbf{Y}_{i_-} \oplus \mathbf{I} \oplus \mathbf{I}$ , the result is immediate from the change of variable  $\mathbf{W}_i = \mathbf{D}_{K,i} \mathbf{Y}_{i_-}$  and the Schur complement formula.  $\square$

**5. Control of Markovian jump linear systems.** In this section, we focus on the case where the switching sequences are realizations of a finite-state homogeneous Markov chain. If  $p = (p_i) \in \mathbb{R}^{1 \times N}$  is a row vector whose entries are nonnegative and sum to one, and if  $\mathbf{P} = (p_{ij}) \in \mathbb{R}^{N \times N}$  is a (row) stochastic matrix where each row of  $\mathbf{P}$  has nonnegative entries that sum to one, then the row-allowable pair  $(\mathbf{P}, p)$  defines the Markov chain with transition probability matrix  $\mathbf{P}$  and initial distribution  $p$ . Let  $\mathcal{G}$  be as in (3.1). Then the triple  $(\mathcal{G}, \mathbf{P}, p)$  defines the discrete-time *Markovian jump linear system*, which is the collection of the linear time-varying systems (3.2) over all realizations  $\theta = (\theta(0), \theta(1), \dots)$  of the Markov chain  $(\mathbf{P}, p)$ . The state  $\theta(t)$  of the chain  $(\mathbf{P}, p)$  at time  $t$  defines the mode of the system  $(\mathcal{G}, \mathbf{P}, p)$  at time  $t$ ; the distribution of the mode at time  $t$  is given by  $p\mathbf{P}^t$ . As in the previous section, let  $\Omega$  be the space of all infinite sequences in  $\{1, \dots, N\}$ . Let  $P$  be the unique consistent probability measure [44] on  $\Omega$  such that

$$P\{\theta(t+1) = j \mid \theta(t) = i\} = p_{ij}, \quad P\{\theta(0) = i\} = p_i$$

for all  $i, j$ , and  $t$ .

If  $x(0) = 0$ , and if  $\mathbf{w} = (w^{(t_0)}(0), w^{(t_0)}(1), \dots)$  is an independent identically distributed random sequence independent of  $\theta$  such that  $w^{(t_0)}(t)$  are Gaussian distributed with (2.8) for  $t_0, t \geq 0$ , then write  $\mathbf{z} = (z^{(t_0)}(0), z^{(t_0)}(1), \dots)$ .

**DEFINITION 5.1.** *The system  $(\mathcal{G}, \mathbf{P}, p)$  is said to be almost surely uniformly (exponentially) stable if there exists a set  $\Theta \subset \Omega$  with  $P(\Theta) = 1$  such that the system  $(\mathcal{G}, \Theta)$  is uniformly exponentially stable.*

**Remark 6.** By definition, if a Markovian jump system is almost surely uniformly stable, and if the transition probability matrix  $\mathbf{P}$  is irreducible, then the system is  $\delta$ -moment stable (i.e., the expectation of  $\|x(t)\|^\delta$  converges to zero for all  $x(0)$  and for all  $p$ ) for all  $\delta > 0$ , and hence is mean square stable (i.e., 2-moment stable). Moreover, mean square stable Markovian jump linear systems are almost surely (but not necessarily uniformly) stable [29, 11]. Thus the notion of almost sure uniform stability is conservative; in particular, a mean square stable system can have an unstable mode  $i$  with  $p_{ii} > 0$ , but almost surely uniformly stable systems cannot—see [11, 34] for more details. On the other hand, unlike the usual definitions for the stability of Markovian jump systems, the almost sure uniform stability is defined here with respect to the probability measure  $P$ , and hence to the given pair  $(\mathbf{P}, p)$ , not to the family of  $(\mathbf{P}, p)$  over all  $p$ ; of course, this distinction becomes irrelevant if  $\mathbf{P}$  is irreducible.

**DEFINITION 5.2.** *Let  $\gamma > 0$ . The system  $(\mathcal{G}, \mathbf{P}, p)$  is said to satisfy almost sure impulse response level  $\gamma$  if there exists a set  $\Theta \subset \Omega$  with  $P(\Theta) = 1$  such that the system  $(\mathcal{G}, \Theta)$  satisfies uniform impulse response level  $\gamma$ .*

DEFINITION 5.3. Let  $\gamma > 0$ . The system  $(\mathcal{G}, \mathbf{P}, p)$  is said to satisfy almost sure output regulation level  $\gamma$  if there exists a set  $\Theta \subset \Omega$  with  $P(\Theta) = 1$  such that the system  $(\mathcal{G}, \Theta)$  satisfies uniform output regulation level  $\gamma$ .

DEFINITION 5.4. Let  $\gamma > 0$ . The system  $(\mathcal{G}, \mathbf{P}, p)$  is said to satisfy average output regulation level  $\gamma$  if there exists a  $\tilde{\gamma} \in (0, \gamma)$  such that, whenever  $x(0) = 0$ ,

$$(5.1) \quad \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|z^{(0)}(t)\|^2 \leq \tilde{\gamma}^2,$$

where  $\mathbb{E}$  denotes the expectation with respect to  $\theta$  and  $\mathbf{w} = (w^{(0)}(0), w^{(0)}(1), \dots)$ .

THEOREM 5.5. Let  $\mathcal{G}$  be as in (3.1); let  $(\mathbf{P}, p)$  be a Markov chain. The system  $(\mathcal{G}, \mathbf{P}, p)$  is almost surely uniformly exponentially stable and satisfies almost sure impulse response level  $\gamma > 0$  if and only if the system  $(\mathcal{G}, \Theta(\mathbf{P}, p))$  is uniformly exponentially stable and satisfies uniform impulse response level  $\gamma$ .

THEOREM 5.6. Let  $\mathcal{G}$  be as in (3.1); let  $(\mathbf{P}, p)$  be a Markov chain. The system  $(\mathcal{G}, \mathbf{P}, p)$  is almost surely uniformly exponentially stable and satisfies almost sure output regulation level  $\gamma > 0$  if and only if the system  $(\mathcal{G}, \Theta(\mathbf{P}, p))$  is uniformly exponentially stable and satisfies uniform output regulation level  $\gamma$ .

*Proof of Theorems 5.5 and 5.6.* The results are immediate from Theorems 3.5 and 3.6: necessity follows from the fact that  $\Theta \subset \Omega$  and  $P(\Theta) = 1$  implies  $\mathcal{M}_L(\mathbf{P}, p) \subset \mathcal{M}_L(\Theta)$ , and sufficiency from  $P(\Theta(\mathbf{P}, p)) = 1$ .  $\square$

Label the  $L$ -paths in  $\mathcal{L}_L(\mathbf{P}, p)$  in dictionary order from 1 to  $N_L$ , where  $N_L$  is the cardinality of  $\mathcal{L}_L(\mathbf{P}, p)$ . Define the matrix  $\mathbf{Q}_L(\mathbf{P}, p) = (q_{ij}) \in \mathbb{R}^{N_L \times N_L}$  as follows: whenever  $(i_0, \dots, i_L)$  and  $(j_0, \dots, j_L)$  are  $L$ -paths labeled  $i$  and  $j$ , respectively, set  $q_{ij} = p_{i_L j_L}$  if  $(i_0, \dots, i_L)_+ = (j_0, \dots, j_L)_-$ ; otherwise, set  $q_{ij} = 0$ . Also, define the row vector  $q_L(\mathbf{P}, p) = (q_i) \in \mathbb{R}^{N_L}$  as follows: whenever  $(i_0, \dots, i_L)$  is an  $L$ -path labeled  $i$ , set  $q_i = p_{i_L}$  if  $(i_0, \dots, i_L)_- = (0, \dots, 0)$ ; otherwise, set  $q_i = 0$ . Then the pair  $(\mathbf{Q}_L(\mathbf{P}, p), q_L(\mathbf{P}, p))$  defines the  $L$ -path Markov chain generated by  $(\mathbf{P}, p)$ . If  $L = 0$ , then

$$(5.2a) \quad \mathbf{Q}_0(\mathbf{P}, p) = \mathbf{Q}_{11}, \quad q_0(\mathbf{P}, p) = q$$

for some submatrix  $\mathbf{Q}_{11}$  of  $\mathbf{P}$  and some subvector  $q$  of  $p$ ; if  $L > 0$ , then it is readily seen that  $\mathbf{Q}_L(\mathbf{P}, p)$  is an  $(L+1)$ -by- $(L+1)$  block matrix and  $q_L(\mathbf{P}, p)$  a 1-by- $(L+1)$  block row vector of the form

$$(5.2b) \quad \mathbf{Q}_L(\mathbf{P}, p) = \begin{bmatrix} \mathbf{0} & \mathbf{Q}_{12} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Q}_{23} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_{LL+1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_{L+1L+1} \end{bmatrix}, \quad q_L(\mathbf{P}, p) = [q \quad \mathbf{0} \quad \cdots \quad \mathbf{0} \quad \mathbf{0}].$$

The matrices  $\mathbf{Q}_{ij}$  and the vector  $q$  in general depend on  $L$ . The matrices  $\mathbf{Q}_{ij}$  define the transition probabilities from the  $L$ -paths that contain  $i$  nonzero modes to those that contain  $j$  nonzero modes; in particular,  $\mathbf{Q}_{L+1L+1}$  is stochastic.

LEMMA 5.7. Let  $(\mathbf{P}, p)$  be a Markov chain. If we partition  $\mathbf{Q}_L(\mathbf{P}, p)$  and  $q_L(\mathbf{P}, p)$  as in (5.2), then for each  $L \geq 0$  we have

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T q_L(\mathbf{P}, p) \mathbf{Q}_L(\mathbf{P}, p)^t = \begin{cases} \pi_0(\mathbf{P}, p), & L = 0; \\ \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \pi_L(\mathbf{P}, p) \end{bmatrix}, & L > 0, \end{cases}$$

where

$$(5.3) \quad \pi_L(\mathbf{P}, p) = \begin{cases} q \lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbf{Q}_{11}^t, & L = 0; \\ q \mathbf{Q}_{12} \cdots \mathbf{Q}_{LL+1} \lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbf{Q}_{L+1L+1}^t, & L > 0. \end{cases}$$

*Proof.* Since  $\mathbf{Q}_L(\mathbf{P}, p)$  is a stochastic matrix, the Cesaro limit

$$(5.4) \quad \lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbf{Q}_L(\mathbf{P}, p)^t$$

of the sequence  $(\mathbf{I}, \mathbf{Q}_L(\mathbf{P}, p), \mathbf{Q}_L(\mathbf{P}, p)^2, \dots)$  always exists—see, e.g., [8, Theorem 8.2.2]. The Cesaro limit of the sequence  $(\mathbf{Q}_L(\mathbf{P}, p)^{t_0}, \mathbf{Q}_L(\mathbf{P}, p)^{t_0+1}, \dots)$  is the same as (5.4) for all  $t_0 \geq 0$ . Consider the case of  $t_0 = L$  in particular: it follows from (5.2) that, for all  $t \geq 0$ ,

$$q_L(\mathbf{P}, p) \mathbf{Q}_L(\mathbf{P}, p)^{t+L} = \begin{cases} q \mathbf{Q}_{11}^t, & L = 0; \\ \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & q \mathbf{Q}_{12} \cdots \mathbf{Q}_{LL+1} \mathbf{Q}_{L+1L+1}^t \end{bmatrix}, & L > 0. \end{cases}$$

Since  $\mathbf{Q}_{L+1L+1}$  is stochastic, the limit (5.3) exists, so the result follows.  $\square$

*Remark 7.* If  $\mathbf{P}$  is irreducible, then the stochastic matrix  $\mathbf{Q}_{L+1L+1}$  in (5.2) is irreducible and the limit  $\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbf{Q}_{L+1L+1}^t$  has all rows equal to  $\pi_L(\mathbf{P}, p)$ , which defines the unique stationary distribution of  $\mathbf{Q}_{L+1L+1}$ . Moreover, if  $\mathbf{P}$  is aperiodic as well as irreducible (i.e., if  $\mathbf{P}$  defines an ergodic chain), then  $\pi_L(\mathbf{P}, p)$  defines the unique steady-state distribution and (5.4) can be replaced by the usual limit  $\lim_{t \rightarrow \infty} \mathbf{Q}_L(\mathbf{P}, p)^t$ . See, e.g., [8, 32].

**THEOREM 5.8.** *Let  $\mathcal{G}$  be as in (3.1); let  $(\mathbf{P}, p)$  be a Markov chain. The system  $(\mathcal{G}, \mathbf{P}, p)$  is almost surely uniformly exponentially stable and satisfies average output regulation level  $\gamma > 0$  if and only if there exist a nonnegative integer  $M$  and an indexed family  $\{\mathbf{Y}_j : j \in \mathcal{M}_M^-(\mathbf{P}, p)\}$  of symmetric positive definite matrices  $\mathbf{Y}_j \in \mathbb{R}^{n \times n}$  such that*

$$(5.5a) \quad \mathbf{A}_{i_M} \mathbf{Y}_{(i_0, \dots, i_M)_-} \mathbf{A}_{i_M}^T - \mathbf{Y}_{(i_0, \dots, i_M)_+} < -\mathbf{B}_{i_M} \mathbf{B}_{i_M}^T$$

for all  $M$ -paths  $(i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{P}, p)$ , and

$$(5.5b) \quad \sum_{(k_0, \dots, k_M) \in \mathcal{W}_M(\mathbf{P}, p)} \pi_{(k_0, \dots, k_M)} \text{tr}(\mathbf{C}_{k_M} \mathbf{Y}_{(k_0, \dots, k_M)_-} \mathbf{C}_{k_M}^T + \mathbf{D}_{k_M} \mathbf{D}_{k_M}^T) < \gamma^2,$$

where  $\pi_M(\mathbf{P}, p) = (\pi_{(k_0, \dots, k_M)})$ .

*Proof.* To show necessity, suppose that  $(\mathcal{G}, \mathbf{P}, p)$  is almost surely uniformly stable and that there exists a  $\tilde{\gamma} \in (0, \gamma)$  such that (5.1) holds. For  $\varepsilon > 0$ , consider the perturbed system  $(\mathcal{G}^{(\varepsilon)}, \mathbf{P}, p)$ , where  $\mathcal{G}^{(\varepsilon)}$  is as in (3.8). Choose a  $\Theta \subset \Omega$  such that  $P(\Theta) = 1$  and such that the system  $(\mathcal{G}, \Theta)$  is uniformly stable. Write the output that  $(\mathcal{G}^{(\varepsilon)}, \mathbf{P}, p)$  generates under  $x(0) = 0$  and  $\mathbf{w} = (w^{(0)}(0), w^{(0)}(1), \dots)$  as  $\mathbf{z} = (z^{(\varepsilon, 0)}(0), z^{(\varepsilon, 0)}(1), \dots)$ , and let  $\mathbf{Y}_{\theta, t}^{(\varepsilon, 0)} \geq \mathbf{0}$  be the solution to

$$\mathbf{A}_{\theta(t)} \mathbf{Y}_{\theta, t}^{(\varepsilon, 0)} \mathbf{A}_{\theta(t)}^T - \mathbf{Y}_{\theta, t+1}^{(\varepsilon, 0)} = -\mathbf{B}_{\theta(t)}^{(\varepsilon)} \mathbf{B}_{\theta(t)}^{(\varepsilon)T},$$



with the initial condition  $\mathbf{Y}_{\theta,0}^{(\varepsilon,0)} = 0$ , for  $\theta \in \Theta$  and  $t \geq 0$ , so that

$$\begin{aligned} \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|z^{(\varepsilon,0)}(t)\|^2 \\ = \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \operatorname{tr} (\mathbf{C}_{\theta(t)} \mathbf{Y}_{\theta,t}^{(\varepsilon,0)} \mathbf{C}_{\theta(t)}^T + \mathbf{D}_{\theta(t)}^{(\varepsilon)} \mathbf{D}_{\theta(t)}^{(\varepsilon)T}). \end{aligned}$$

Then it follows from the proof of the necessity part of Theorem 3.6 that, for a sufficiently small  $\varepsilon > 0$ , there are a nonnegative integer  $M$  and an indexed family of matrices  $\mathbf{Y}_{(i_0, \dots, i_M)_-} > \mathbf{0}$ ,  $(i_0, \dots, i_M) \in \mathcal{L}_M(\Theta)$ , such that

$$\mathbf{A}_{i_M} \mathbf{Y}_{(i_0, \dots, i_M)_-} \mathbf{A}_{i_M}^T - \mathbf{Y}_{(i_0, \dots, i_M)_+} < -\mathbf{B}_{i_M} \mathbf{B}_{i_M}^T$$

for  $(i_0, \dots, i_M) \in \mathcal{L}_M(\Theta)$ , and such that

$$(5.6) \quad \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \operatorname{tr} (\mathbf{C}_{\theta(t)} \mathbf{Y}_{(\theta(t-M), \dots, \theta(t))_-} \mathbf{C}_{\theta(t)}^T + \mathbf{D}_{\theta(t)} \mathbf{D}_{\theta(t)}^T) < \gamma^2.$$

Since  $\mathcal{M}_M(\mathbf{P}, p) \subset \mathcal{M}_M(\Theta) \subset \mathcal{L}_M(\Theta)$ , we have that (5.5a) holds for  $(i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{P}, p)$ . Label the  $M$ -paths in  $\mathcal{L}_M(\mathbf{P}, p)$  in dictionary order from 1 to  $N_M$ , where  $N_M$  is the cardinality of  $\mathcal{L}_M(\mathbf{P}, p)$ . Form a column vector  $c \in \mathbb{R}^{N_M}$  such that its  $i$ th component is equal to  $\operatorname{tr} (\mathbf{C}_{k_M} \mathbf{Y}_{(k_0, \dots, k_M)_-} \mathbf{C}_{k_M}^T - \mathbf{D}_{k_M} \mathbf{D}_{k_M}^T)$  whenever  $(k_0, \dots, k_M)$  is the  $i$ th  $M$ -path in  $\mathcal{L}_M(\mathbf{P}, p)$ . Then we have that the left-hand side of inequality (5.6) is equal to

$$\lim_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T q_L(\mathbf{P}, p) \mathbf{Q}_L(\mathbf{P}, p)^t c.$$

Now, by Lemma 5.7, we have that

$$(5.7) \quad \limsup_{T \rightarrow \infty} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|z^{(\varepsilon,0)}(t)\|^2 = \sum_{(k_0, \dots, k_M) \in \mathcal{W}_M(\mathbf{P}, p)} \pi_{(k_0, \dots, k_M)} c_{(k_0, \dots, k_M)} < \gamma^2,$$

where  $c_{(k_0, \dots, k_M)}$  is the component of  $c$  associated with the  $M$ -path  $(k_0, \dots, k_M)$ . Hence (5.5b) holds.

To show sufficiency, suppose that (5.5) holds for some integer  $M \geq 0$  and for all  $(i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{P}, p)$ . Then, by the proof of the sufficiency part of Theorem 3.6, we have that the system  $(\mathcal{G}, \mathbf{P}, p)$  is almost surely uniformly stable because the system  $(\mathcal{G}, \Theta(\mathbf{P}, p))$  is uniformly stable with  $P(\Theta(\mathbf{P}, p)) = 1$ ; also, there exists a  $\tilde{\gamma} \in (0, \gamma)$  satisfying (5.1) because of (5.7).  $\square$

Let  $\mathcal{T}$  be as in (4.1), and let  $(\mathbf{P}, p)$  be a Markov chain. Then the triple  $(\mathcal{T}, \mathbf{P}, p)$  defines the *controlled Markovian jump linear system* represented by (4.2), where  $\theta$  is a realization of  $(\mathbf{P}, p)$ . As in the previous section, we make the standard assumption that *the state  $\theta(t)$  of the chain  $(\mathbf{P}, p)$  is perfectly observed at each time instant  $t$* ; we consider all finite-path-dependent controllers.

Fix a nonnegative integer  $L$ . Let

$$\mathcal{K} = \{(\mathbf{A}_{K,i}, \mathbf{B}_{K,i}, \mathbf{C}_{K,i}, \mathbf{D}_{K,i}) : i \in \mathcal{L}_L(\mathbf{P}, p)\}.$$

Then the pair  $(\mathcal{K}, \Theta_L(\mathbf{P}, p))$ , with

$$\Theta_L(\mathbf{P}, p) = \{\theta_L : \theta \in \Theta(\mathbf{P}, p)\},$$

defines an  $L$ -path-dependent controller, whose state-space representation is given by (4.3). Since the pair  $(\mathbf{Q}_L(\mathbf{P}, p), q_L(\mathbf{P}, p))$  defines the  $L$ -path Markov chain generated by  $(\mathbf{P}, p)$ , the closed-loop system is a Markovian jump linear system given by the triple  $(\mathcal{T}_\mathcal{K}, \mathbf{Q}_L(\mathbf{P}, p), q_L(\mathbf{P}, p))$ , with

$$\mathcal{T}_\mathcal{K} = \{(\tilde{\mathbf{A}}_i, \tilde{\mathbf{B}}_i, \tilde{\mathbf{C}}_i, \tilde{\mathbf{D}}_i) : i \in \mathcal{L}_L(\mathbf{P}, p)\},$$

and represented by (4.4) for each realization  $\theta_L$  of  $(\mathbf{Q}_L(\mathbf{P}, p), q_L(\mathbf{P}, p))$ .

**DEFINITION 5.9.** Let  $\gamma > 0$ . The controller  $(\mathcal{K}, \Theta_L(\mathbf{P}, p))$  is said to be a  $\gamma$ -admissible ( $L$ -path-dependent) synthesis (of order  $n_K$ ) for the system  $(\mathcal{T}, \mathbf{P}, p)$  if the closed-loop system  $(\mathcal{T}_\mathcal{K}, \mathbf{Q}_L(\mathbf{P}, p), q_L(\mathbf{P}, p))$  is almost surely uniformly exponentially stable and satisfies average output regulation level  $\gamma$ .

**THEOREM 5.10.** Let  $\mathcal{T}$  be as in (4.1); let  $(\mathbf{P}, p)$  be a Markov chain. Suppose that  $n_K \geq n$  and  $\gamma > 0$ . There exists a  $\gamma$ -admissible finite-path-dependent synthesis of order  $n_K$  for the system  $(\mathcal{T}, \mathbf{P}, p)$  if and only if there exist a nonnegative integer  $M$  and indexed families  $\{(\mathbf{R}_j, \mathbf{S}_j) : j \in \mathcal{M}_M^-(\mathbf{Q}, q)\}$  and  $\{(\mathbf{Z}_i, \mathbf{W}_i) : i \in \mathcal{W}_M(\mathbf{Q}, q)\}$  of symmetric matrices  $\mathbf{R}_j \in \mathbb{R}^{n \times n}$ ,  $\mathbf{S}_j \in \mathbb{R}^{n \times n}$ ,  $\mathbf{Z}_i \in \mathbb{R}^{l_1 \times l_1}$  and rectangular matrices  $\mathbf{W}_i \in \mathbb{R}^{(n+m_2) \times (n+l_2)}$  such that (4.8a) holds for all  $i = (i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{P}, p)$ , (4.8b) holds for all  $i = (i_0, \dots, i_M) \in \mathcal{W}_M(\mathbf{P}, p)$ , and

$$(5.8) \quad \sum_{(k_0, \dots, k_M) \in \mathcal{W}_M(\mathbf{P}, p)} \pi_{(k_0, \dots, k_M)} \text{tr } \mathbf{Z}_{(k_0, \dots, k_M)} < \gamma^2,$$

where  $\pi_M(\mathbf{P}, p) = (\pi_{(k_0, \dots, k_M)})$ . Moreover, if this condition is satisfied, then given any nonsingular matrices  $\mathbf{T}_j, \mathbf{U}_j \in \mathbb{R}^{n \times n}$  such that (4.9) holds for all  $j \in \mathcal{M}_M^-(\mathbf{P}, p)$ , a  $\gamma$ -admissible  $M$ -path-dependent synthesis of order  $n$  is obtained by solving (4.7), with  $n_K = n$  and  $L = M$ , for matrices  $\mathbf{A}_{K,i}, \mathbf{B}_{K,i}, \mathbf{C}_{K,i}, \mathbf{D}_{K,i}, i \in \mathcal{M}_M(\mathbf{P}, p)$ .

*Proof.* The result follows from Theorem 5.8 and the proof of Theorem 4.2.  $\square$

**Remark 8.** As in Theorem 4.2, it is possible to replace  $\mathcal{M}_M(\mathbf{P}, p)$  and  $\mathcal{W}_M(\mathbf{P}, p)$  with smaller sets. However, using these bigger sets is more convenient, as stated in Remark 4.

**Remark 9.** The condition (5.8) is weaker than the condition that (4.8c) holds for all  $(\mathbf{P}, p)$ -minimal sets  $\mathcal{N}$  of  $M$ -paths. In general, for all  $\gamma$  and  $M$ , there exists a  $\gamma$ -admissible  $M$ -path-dependent synthesis for the Markovian jump system  $(\mathcal{T}, \mathbf{P}, p)$  whenever there exists a  $\gamma$ -admissible  $M$ -path-dependent synthesis for the switched system  $(\mathcal{T}, \Theta(\mathbf{P}, p))$ .

**Example 6.** Let  $\mathcal{T}$  be as in Example 5. Let

$$\mathbf{P} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad p = [1/2 \quad 0 \quad 1/2].$$

Since the sparsity pattern of  $(\mathbf{P}, p)$  is the same as that of  $(\mathbf{Q}, q)$  in Example 5, we have  $\mathcal{M}_M(\mathbf{P}, p) = \mathcal{M}_M(\mathbf{Q}, q)$  for all path lengths  $M$ . Moreover, in this particular example,  $\mathcal{N}_M(\mathbf{P}, p) = \mathcal{W}_M(\mathbf{P}, p) = \mathcal{M}_M(\mathbf{P}, p)$  for all  $M$ . If  $M = 0$ , then  $\mathbf{Q}_{11} = \mathbf{P}$

TABLE 5.1  
In Example 6,  $\gamma_M$  is decreasing in  $M$  and saturates at  $M = 5$ .

$M$	0	1	2	3	4	5	6	7
$\gamma_M$	1.3280	1.1837	1.1815	1.1812	1.1809	1.1808	1.1808	1.1808

and  $q = p$ , so

$$\begin{aligned}\pi_0(\mathbf{P}, p) &= [\pi_1 \quad \pi_2 \quad \pi_3] \\ &= [1/2 \quad 0 \quad 1/2] \begin{bmatrix} 2/3 & 1/3 & 0 \\ 2/3 & 1/3 & 0 \\ 0 & 0 & 1 \end{bmatrix} = [1/3 \quad 1/6 \quad 1/2].\end{aligned}$$

If  $M = 1$ , then

$$\mathbf{Q}_{12} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{Q}_{22} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad q = [1/2 \quad 1/2],$$

so

$$\begin{aligned}\pi_1(\mathbf{P}, p) &= [\pi_{(1,1)} \quad \pi_{(1,2)} \quad \pi_{(2,1)} \quad \pi_{(3,3)}] \\ &= [1/2 \quad 1/2] \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= [1/6 \quad 1/6 \quad 1/6 \quad 1/2].\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}\pi_2(\mathbf{P}, p) &= [\pi_{(1,1,1)} \quad \pi_{(1,1,2)} \quad \pi_{(1,2,1)} \quad \pi_{(2,1,1)} \quad \pi_{(2,1,2)} \quad \pi_{(3,3,3)}] \\ &= [1/12 \quad 1/12 \quad 1/6 \quad 1/12 \quad 1/12 \quad 1/2],\end{aligned}$$

and so on.

For nonnegative integers  $M$ , denote by  $\gamma_M$  the infimum of  $\gamma > 0$  such that (4.8a), (4.8b), and (5.8) hold for all  $M$ -paths  $\mathbf{i} = (i_0, \dots, i_M) \in \mathcal{M}_M(\mathbf{P}, p)$ . Then semidefinite programming gives us Table 5.1; as expected,  $\gamma_M$  decreases as  $M$  gets larger, but it saturates at  $M = 5$ . Thus, we determine that the optimal achievable value of  $\gamma$  is approximately  $\gamma^* = 1.1808$ , and that there exists a  $\gamma^*$ -admissible five-path-dependent synthesis; the resulting 22 controller matrices are denoted  $\mathbf{K}_{(1,1,1,1,1,1)}$ ,  $\mathbf{K}_{(1,1,1,1,1,2)}$ ,  $\mathbf{K}_{(1,1,1,1,2,1)}$ ,  $\dots$ ,  $\mathbf{K}_{(2,1,2,1,2,1)}$ ,  $\mathbf{K}_{(3,3,3,3,3,3)}$ .  $\square$

**6. Conclusion.** A convex characterization of the output regulation performance of switched linear systems and Markovian jump linear systems was given by an increasing union of linear matrix inequality conditions. This characterization gives rise to semidefinite programming-based “offline” algorithms for finding an optimal solution from (almost surely) uniformly stabilizing dynamic output feedback controllers

that have finite memory of past modes. Due to the nature of problem, however, the computational burden can grow drastically in the number of the past modes that the controller recalls. Although this limitation often allows us the complete freedom in controlling the conservatism, or suboptimality, of the resulting controllers, some systems may be better dealt with in practice via, e.g., the dwell-time approaches [23, 48, 22].

The results presented are not only exact and convex, but also very general. One should be able to characterize a large class of performance objectives under which the convexity is preserved in the spirit of [43, 39]. This generality is gained in exchange for giving up the closed-form solutions that standard Riccati equation-based approaches such as those in [18, 42] would yield. However, one could adopt these standard approaches to enhance and complement our results. Possible topics in this direction include whether the noncausality of the solution to the impulse response level minimization problem is due to the nature of problem or not, and a separation principle for the problems formulated in this paper.

The output regulation problems can be easily extended to the simultaneous design of the supervisor-controller pair. As is well known in the context of LQG measurement scheduling [40], the problem of jointly optimizing both the switching sequence and the output feedback controller is separated into two problems: one is to determine the optimal switching rule via an algorithm analogous to that developed in [35], and the other is, given a switching rule, to obtain a path-dependent controller using the results of this paper. For discrete-time switched systems, existing results in this direction seem to be limited to the finite-horizon state feedback problem [37].

#### REFERENCES

- [1] P. APKARIAN, P. GAHINET, AND G. BECKER, *Self-scheduled  $H_\infty$  control of linear parameter-varying systems: A design example*, Automatica, 31 (1995), pp. 1251–1261.
- [2] M. A. BERGER AND Y. WANG, *Bounded semigroups of matrices*, Linear Algebra Appl., 166 (1992), pp. 21–27.
- [3] P.-A. BLIMAN AND G. FERRARI-TRECATE, *Stability analysis of discrete-time switched systems through Lyapunov functions with nonminimal state*, in Proceedings of the IFAC Conference on the Analysis and Design of Hybrid Systems, 2003, pp. 325–330.
- [4] V. D. BLONDEL, J. THEYS, AND A. A. VLADIMIROV, *An elementary counterexample to the finiteness conjecture*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 963–970.
- [5] G. BÖKER AND J. LUNZE, *Stability and performance of switching Kalman filters*, Internat. J. Control, 75 (2002), pp. 1269–1281.
- [6] M. S. BRANICKY, *Multiple Lyapunov functions and other analysis tools for switched and hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 475–482.
- [7] R. W. BROCKETT AND D. LIBERZON, *Quantized feedback stabilization of linear systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1279–1289.
- [8] S. L. CAMPBELL AND C. D. MEYER, *Generalized Inverses of Linear Transformations*, Pitman, San Francisco, CA, 1979.
- [9] H. CHAN AND Ü. ÖZGÜNER, *Closed-loop control of systems over a communication network with queues*, Internat. J. Control, 62 (1995), pp. 493–510.
- [10] O. L. V. COSTA, *Discrete-time coupled Riccati equations for systems with Markov switching parameters*, J. Math. Anal. Appl., 194 (1995), pp. 197–216.
- [11] O. L. V. COSTA AND M. D. FRAGOSO, *Stability results on discrete-time linear systems with Markovian jumping parameters*, J. Math. Anal. Appl., 179 (1993), pp. 154–178.
- [12] O. L. V. COSTA, M. D. FRAGOSO, AND R. P. MARQUES, *Discrete-Time Markov Jump Linear Systems*, Springer-Verlag, London, UK, 2005.
- [13] O. L. V. COSTA AND E. F. TUESTA,  *$H_2$ -control and the separation principle for discrete-time Markovian jump linear systems*, Math. Control Signals Systems, 16 (2004), pp. 320–350.
- [14] J. DAAFOUZ, P. RIEDINGER, AND C. IUNG, *Stability analysis and control synthesis for switched systems: A switched Lyapunov function approach*, IEEE Trans. Automat. Control, 47 (2002), pp. 1883–1887.

- [15] I. DAUBECHIES AND J. C. LAGARIAS, *Sets of matrices all infinite products of which converge*, Linear Algebra Appl., 161 (1992), pp. 227–263.
- [16] I. DAUBECHIES AND J. C. LAGARIAS, *Corrigendum/addendum to: Sets of matrices all infinite products of which converge*, Linear Algebra Appl., 327 (2001), pp. 69–83.
- [17] D. P. DE FARIAS, J. C. GEROMEL, J. B. R. DO VAL, AND O. L. V. COSTA, *Output feedback control of Markov jump linear systems in continuous-time*, IEEE Trans. Automat. Control, 45 (2000), pp. 944–949.
- [18] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State-space solutions to standard  $H_2$  and  $H_\infty$  control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.
- [19] G. E. DULLERUD AND S. LALL, *A new approach for analysis and synthesis of time-varying systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 1486–1497.
- [20] P. GAHINET AND P. APKARIAN, *A linear matrix inequality approach to  $H_\infty$  control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.
- [21] L. GURVITS, *Stability of discrete linear inclusion*, Linear Algebra Appl., 231 (1995), pp. 47–85.
- [22] J. HESPAÑHA, *Root-mean-square gains of switched linear systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 2040–2045.
- [23] J. P. HESPAÑHA AND A. S. MORSE, *Stability of switched systems with average dwell-time*, in Proceedings of the 38th IEEE Conference on Decision and Control, Vol. 3, 1999, pp. 2655–2660.
- [24] P. A. IGLESIAS AND M. A. PETERS, *On the induced norms of discrete-time and hybrid time-varying systems*, Internat. J. Robust Nonlinear Control, 7 (1997), pp. 811–833.
- [25] V. IONESCU AND M. WEISS, *The  $\ell^2$ -control problem for time-varying discrete systems*, Systems Control Lett., 18 (1992), pp. 371–381.
- [26] A. JADBABAIE, J. LIN, AND A. S. MORSE, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.
- [27] Y. JI AND H. J. CHIZECK, *Jump linear quadratic Gaussian control: Steady-state solution and testable conditions*, Control Theory Adv. Tech., 6 (1990), pp. 289–319.
- [28] Y. JI AND H. J. CHIZECK, *Jump linear quadratic Gaussian control in continuous time*, IEEE Trans. Automat. Control, 37 (1992), pp. 1884–1892.
- [29] Y. JI, H. J. CHIZECK, X. FENG, AND K. A. LOPARO, *Stability and control of discrete-time jump linear systems*, Control Theory Adv. Tech., 7 (1991), pp. 247–270.
- [30] X. D. KOUTSOUKOS, P. J. ANTSAKLIS, J. A. STIVER, AND M. D. LEMMON, *Supervisory control of hybrid systems*, Proc. IEEE, 88 (2000), pp. 1026–1046.
- [31] R. KRTOŁICA, Ü. ÖZGÜNER, H. CHAN, H. GÖKTAS, J. WINKELMAN, AND M. LIUBAKKA, *Stability of linear feedback systems with random communication delays*, Internat. J. Control, 59 (1994), pp. 925–953.
- [32] P. R. KUMAR AND P. VARAIYA, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.
- [33] J.-W. LEE AND G. E. DULLERUD, *Optimal disturbance attenuation for discrete-time switched and Markovian jump linear systems*, SIAM J. Control Optim., 45 (2006), pp. 1329–1358.
- [34] J.-W. LEE AND G. E. DULLERUD, *Uniform stabilization of discrete-time switched and Markovian jump linear systems*, Automatica, 42 (2006), pp. 205–218.
- [35] J.-W. LEE AND G. E. DULLERUD, *Uniformly stabilizing sets of switching sequences for switched linear systems*, IEEE Trans. Automat. Control, 52 (2007), pp. 868–874.
- [36] D. LIBERZON, *Switching in Systems and Control*, Birkhäuser, Boston, MA, 2003.
- [37] B. LINCOLN AND B. BERNHARDSSON, *LQR optimization of linear system switching*, IEEE Trans. Automat. Control, 47 (2002), pp. 1701–1705.
- [38] M. MARITON, *Jump Linear Systems in Automatic Control*, Marcel Dekker, New York, 1990.
- [39] I. MASUBUCHI, A. OHARA, AND N. SUDA, *LMI-based controller synthesis: A unified formulation and solution*, Internat. J. Robust Nonlinear Control, 8 (1998), pp. 669–686.
- [40] L. MEIER, III, J. PESCHON, AND R. M. DRESSLER, *Optimal control of measurement subsystems*, IEEE Trans. Automat. Control, 12 (1967), pp. 528–536.
- [41] A. PACKARD, *Gain scheduling via linear fractional transformations*, Systems Control Lett., 22 (1994), pp. 79–92.
- [42] R. RAVI, K. M. NAGPAL, AND P. P. KHARGONEKAR,  *$H^\infty$  control of linear time-varying systems: A state-space approach*, SIAM J. Control Optim., 29 (1991), pp. 1394–1413.
- [43] C. SCHERER, P. GAHINET, AND M. CHILALI, *Multiobjective output-feedback control via LMI optimization*, IEEE Trans. Automat. Control, 42 (1997), pp. 896–911.
- [44] A. N. SHIRYAYEV, *Probability*, Springer-Verlag, New York, 1996.
- [45] Z. SUN AND S. S. GE, *Switched Linear Systems: Control and Design*, Springer-Verlag, London, UK, 2005.

- [46] J. N. TSITSIKLIS AND V. D. BLONDEL, *The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate*, Math. Control Signals Systems, 10 (1997), pp. 31–40.
- [47] L. XIAO, A. HASSIBI, AND J. P. HOW, *Control with random communication delays via a discrete-time jump system approach*, in Proceedings of the American Control Conference, Vol. 3, 2000, pp. 2199–2204.
- [48] G. ZHAI, B. HU, K. YASUDA, AND A. N. MICHEL, *Qualitative analysis of discrete-time switched systems*, in Proceedings of the American Control Conference, Vol. 3, 2002, pp. 1880–1885.

## SINGULARLY PERTURBED PIECEWISE DETERMINISTIC GAMES\*

ALAIN HAURIE<sup>†</sup> AND FRANCESCO MORESINO<sup>‡</sup>

**Abstract.** In this paper we consider a class of hybrid stochastic games with the piecewise open-loop information structure. These games are indexed over a parameter  $\varepsilon$  which represents the time scale ratio between the stochastic (jump process) and the deterministic (differential state equation) parts of the dynamical system. We study the limit behavior of Nash equilibrium solutions to the hybrid stochastic games when the time scale ratio tends to 0. We also establish that an approximate equilibrium can be obtained for the hybrid stochastic games using a Nash equilibrium solution of a reduced order sequential discrete state stochastic game and a family of local deterministic infinite horizon open-loop differential games defined in the stretched out time scale. A numerical illustration of this approximation scheme is also developed.

**Key words.**  $n$ -person games, noncooperative games, stochastic games, differential games, time scale analysis, singular perturbations

**AMS subject classifications.** 91A06, 91A10, 91A15, 91A23, 93C70

**DOI.** 10.1137/050627599

**1. Introduction.** This paper deals with a class of piecewise deterministic stochastic games where the stochastic jump process has a slower time scale than the deterministic continuous time control systems that are defined between successive random jumps. These types of games may occur, for example, in imperfect competition models where the deterministic subsystem describes the productive capital accumulation of the firms competing on a market and where the market conditions are subject to infrequent random switches that are influenced by the actions of the economic agents. Situations where oligopolistic markets can be subject to abrupt modal changes are observed, for example, in the energy sector or in the new technology or telecommunication domains. Another interesting domain where this type of paradigm could be used is the modeling of economic dimensions of climate change. The fast modes would correspond to the competitive economic growth processes of different world economies, whereas the slow modes would be associated with different climate conditions. Indeed the transition from a climate mode to a different one would be influenced by the global emissions of greenhouse gases from all nations.

The information structure that we consider for these games is called *piecewise open-loop*; it has been introduced in [7] and consists in playing open-loop controls between successive jump times; the open-loop controls are adapted to the history of jump times and system states observed at jump times. It has been recognized that these piecewise deterministic games, when played under the piecewise open-loop information structure, are akin to the general class of stochastic sequential games

---

\*Received by the editors March 25, 2005; accepted for publication (in revised form) July 18, 2007; published electronically January 4, 2008. This research was supported by the Swiss NSF NCCR-Climate grant. A preliminary form of this paper without proofs appeared in the Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference 2005 (CDC-ECC'05).

<http://www.siam.org/journals/sicon/47-1/62759.html>

<sup>†</sup>Department of Management Studies, University of Geneva, Switzerland, and GERAD HEC-Montréal, Canada (alain.haurie@hec.unige.ch).

<sup>‡</sup>Haute École de Gestion de Genève, Campus Battelle, Bâtiment F, Route de Drize 7, 1227 Carouge, Switzerland, and REME-EPFL École Polytechnique Fédérale de Lausanne, Switzerland (francesco.moresino@hesge.ch).

with general (Borel) action and state spaces, considered in particular in [17] and [15], where existence and approximation of Nash equilibria have been studied. We also suppose that the players have separated deterministic dynamics and are linked only through the payoff rewards at that level. This means that between two successive modal changes each player  $j = 1, \dots, m$  selects a deterministic trajectory  $x_j(\cdot)$  that has a given initial state determined by the state at the time of the first jump and which will be in force until the second jump occurs. We assume that the  $m$  state variables  $x_j$ ,  $j = 1, \dots, m$ , together influence the jump rates of the discrete modal stochastic process.

The aim of this paper is to develop a theory of approximation for a Nash equilibrium solution to this class of differential games when the time scale ratio between the (continuous) deterministic and (jump) stochastic dynamics tends to 0. This is the realm of *singular perturbation* theory in control. In [6] one can find a rather complete theory of singularly perturbed piecewise deterministic control systems and an illustration of the role played by “turnpikes” (i.e., global attractors for optimal trajectories in infinite horizon control problems) in the definition of the “limit control problem.” The initial objective of this paper was to explore the possibility of extending the results obtained in [6] to the case of piecewise open-loop Nash equilibria. As is often the case when one passes from a single optimizer control formalism to the context of dynamic games with a Nash equilibrium solution, these results cannot be readily generalized, and we obtain substantially different and weaker limit theorems.

In the present paper we consider only the case of infinite horizon hybrid games with discounted payoffs. In brief the contribution of this paper can be stated as follows: (i) We define, in the slow time scale, a limit game problem in the form of a controlled Markov chain. A Nash equilibrium for this limit game, if it exists, will serve to build an approximate equilibrium solution of the original game. (ii) We prove that given a Nash equilibrium for the limit game, defined in terms of attractors for the players  $x_j(\cdot)$ -trajectories, one can construct a  $\zeta$ -equilibrium for any  $G^\varepsilon$  game, where  $\varepsilon$  is small enough by using strategies characterized by a uniform attractor property. (iii) Having solved the limit Nash game, we can use the associated potential function to characterize a set of local infinite horizon open-loop games, whose Nash equilibria satisfy the uniform attractor property. (iv) We thus derive a decomposition principle for this class of games and illustrate it on a numerical example.

The paper is organized as follows: in section 2 we recall the definition of a piecewise deterministic game played with piecewise open-loop strategies; in section 3 we study the limit game when the time scale ratio  $\varepsilon$  between fast and slow modes tends to 0; in section 4 we propose a method to construct a  $\zeta$ -equilibrium solutions for the hybrid game, using uniform attractor policies; in section 5 we study a class of local open-loop games for which the Nash equilibrium strategies satisfy the uniform attractor property; in section 6 we derive from these results a decomposition principle; and in section 7 we provide a numerical illustration of these limit properties and sketch an economic model of climate change policies having this two-time scale structure; in section 8 we summarize what has been achieved.

**2. A class of piecewise deterministic games.** In this section we define the class of dynamic games that are considered in this work. They are particular instances of piecewise deterministic games, as introduced in [7]. We use both a formalism of control systems and a formalism of calculus of variations, very much in the same way as in [4], where the so-called turnpike property for open-loop differential games with decoupled dynamics was established.



**2.1. The dynamics.** Consider  $m$  players, denoted  $j \in M = \{1, \dots, m\}$ , controlling a system that has  $p$  discrete modes denoted  $i \in I = \{0, \dots, p-1\}$ . Each player  $j \in M$  also controls her own dynamical subsystem with mode-dependent dynamics,

$$(2.1) \quad \varepsilon \dot{x}_j(t) = f_j^i(x_j(t), u_j(t)),$$

$$(2.2) \quad u_j(t) \in U_j^i \subset \mathbb{R}^{n_j},$$

$$(2.3) \quad x_j(t) \in X_j \subset \mathbb{R}^{n_j},$$

where the control sets  $U_j^i$  are compacts and the state sets  $X_j$  are bounded. The functions  $f_j^i(x_j, u_j)$  are supposed to satisfy the usual regularity assumptions made in control theory. Here  $\varepsilon$  is a parameter that will eventually be very small. We denote  $\underline{x} = (x_j : j \in M) \in \underline{X}$  and  $\underline{u} = (u_j : j \in M) \in \underline{U}$  to be the state and control vectors, respectively.

The “mode” dynamics is represented by a continuous time jump process  $\xi(\cdot)$ , with state set  $I$  and transition rates

$$q_{k\ell}(\underline{x}(t)) = \lim_{dt \rightarrow 0} \frac{1}{dt} \mathbb{P}[\xi(t+dt) = \ell \mid \xi(t) = k, \underline{x}(t)], \quad k, \ell \in I, \underline{x}(t) \in \underline{X},$$

that depend on the trajectory choice made by all players. Indeed we assume  $q_{k\ell}(\underline{x}) \geq 0$  if  $k \neq \ell$  and  $q_{kk}(\underline{x}) = -\sum_{\ell \neq k} q_{k\ell}(\underline{x})$ . As usual we introduce the notation

$$q^k(\underline{x}) \doteq \sum_{\ell \neq k} q_{k\ell}(\underline{x}).$$

*Remark 1.* The parameter  $\varepsilon$  represents the time scale ratio. Its inverse  $1/\varepsilon$  is therefore a speed of adjustment factor for the deterministic part of the hybrid system; when  $\varepsilon \rightarrow 0$  the deterministic part of the system is allowed to adjust much faster than the stochastic jumps occurrences.

It will be convenient to use a calculus of variations formalism, obtained in the following way: We assume that at time  $t$  the reward rate to Player  $j$ , when the mode is  $i$ , is given by a function  $L_j^i(\underline{x}(t), u_j(t))$  which is  $C^1$  in  $\underline{x}$  and continuous in  $u_j$ . Let  $F_j^i(z_j, x_j) = \{u_j \in U_j^i : z_j = f_j^i(x_j, u_j)\}$  be the set of controls for Player  $j$  that yield a velocity  $z_j$  at state  $x_j$ . We introduce a function  $\mathcal{L}_j^i(\underline{x}, z_j)$  that associates a value in  $\mathbb{R} \cup \{-\infty\}$  with every  $\underline{x} \in \Pi_{l=1}^n X_l \subset \mathbb{R}^n$ ,  $n = \sum_{j \in M} n_j$ , and  $z_j \in \mathbb{R}^{n_j}$  as follows:

$$(2.4) \quad \mathcal{L}_j^i(\underline{x}, z_j) = \begin{cases} -\infty & \text{if } x_j \notin X_j \text{ or } F_j^i(z_j, x_j) = \emptyset; \\ \sup\{L_j^i(\underline{x}, u_j) : u_j \in F_j^i(z_j, x_j)\} & \text{otherwise.} \end{cases}$$

We can now consider a dynamic game where each player  $j \in M = \{1, \dots, m\}$  controls an absolutely continuous trajectory  $x_j(\cdot)$  with state  $x_j(t) \in X_j$  at time  $t \in [0, \infty)$ , where  $X_j$  is a compact subset in  $\mathbb{R}^{n_j}$ .

The game is played as follows: at jump times  $\tau^0 = 0, \tau^1, \dots, \tau^\nu, \dots$  of the  $\xi(\cdot)$  process the players observe the state of the system, i.e., the pair  $s^\nu = (\xi^\nu, \underline{x}^\nu)$ , where  $\xi^\nu = \xi(\tau^\nu)$  and  $\underline{x}^\nu = \underline{x}(\tau^\nu)$ . Then Player  $j$  selects an absolutely continuous function  $y_j : [0, \infty) \rightarrow X_j$ , with initial condition  $y_j(0) = x_j^\nu$ . The trajectory for Player  $j$  will thus be defined, between jump times  $\tau^\nu$  and  $\tau^{\nu+1}$ , by  $x_j(t) = y_j(t - \tau^\nu)$ . We denote by  $\mathcal{X}_j$  the class of admissible functions  $y_j : [0, \infty) \rightarrow X_j$  that serve to define the action set of Player  $j$ . A piecewise open-loop strategy for Player  $j$  is then defined as a mapping  $\gamma_j : (\tau^\nu, s^\nu) \mapsto \mathcal{X}_j$ .

At time  $t$  the reward rate to Player  $j$  is given by  $\mathcal{L}_j^{\xi(t)}(\underline{x}(t), \varepsilon \dot{x}_j(t))$  and depends on the current mode  $\xi(t)$ , the state vector  $\underline{x}(t) = (x_j(t))_{j \in M}$ , and the time derivative  $\dot{x}_j(\cdot)$  of Player  $j$ 's own trajectory multiplied by the time scale ratio  $\varepsilon$ .

**2.2. The hybrid game  $G^\varepsilon$ .** We call *hybrid game*  $G^\varepsilon$  the game in normal form where the players select piecewise open-loop strategies as defined above and obtain payoffs defined as follows:

Let  $\rho_j$  be the discount rate of Player  $j$ . Associated with a strategy  $m$ -tuple  $\underline{\gamma} = \{\gamma_j : j \in M\}$  the payoffs to the players are given by

$$(2.5) \quad V_j^\varepsilon(\underline{\gamma}; i, \underline{x}^o) = \mathbb{E}_{\underline{\gamma}} \left[ \int_0^\infty e^{-\rho_j t} \mathcal{L}_j^{\xi(t)}(\underline{x}(t), \varepsilon \dot{x}_j(t)) dt \mid (\xi(t^0) = i, \underline{x}(t^0) = \underline{x}^o) \right]$$

$$j \in M, \quad (i, \underline{x}^o) \in I \times \underline{X},$$

where  $\mathbb{E}_{\underline{\gamma}}$  is the expectation given the probability measure induced by the strategy vector  $\underline{\gamma}$ .

**DEFINITION 2.1.** (i) A strategy  $m$ -tuple  $\underline{\gamma}^*$  is a  $\varsigma$ -equilibrium, with  $\varsigma \geq 0$  given, if

$$(2.6) \quad V_j^{\varepsilon*}(i, \underline{x}^o) = V_j^\varepsilon(\underline{\gamma}^*; i, \underline{x}^o) \geq V_j^\varepsilon([\underline{\gamma}_{M-j}^*, \gamma_j]; i, \underline{x}^o) - \varsigma \quad \forall \gamma_j \in \Gamma_j$$

$$j \in M, \quad (i, \underline{x}^o) \in I \times \underline{X},$$

where  $[\underline{\gamma}_{M-j}^*, \gamma_j]$  denotes the strategy vector obtained from  $\underline{\gamma}^*$  when only Player  $j$  unilaterally changes her strategy to  $\gamma_j$ .

(ii) A 0-equilibrium is also called a Nash equilibrium.

The reader will note that we distinguish between  $\varepsilon > 0$ , which is the time scale ratio, and  $\varsigma > 0$ , which is the approximation used in the equilibrium conditions.

Indeed, when  $\varepsilon$  becomes very small this game will become ill-conditioned. In the rest of the paper we propose an approach for defining a limit game which is easier to solve and which can be used to construct approximate equilibria of the original game when  $\varepsilon$  is small.

**3. The limit game  $G^0$ .** In this section we introduce the so-called *limit game*  $G^0$ , which is defined as a multiagent controlled Markov chain with states in  $I$  and controls in  $X_j^i$ ,  $i \in I$ ,  $j \in M$ .

**3.1. A discrete-state Markov game.** The limit game  $G^0$  is defined as a controlled Markov chain on the discrete set  $I$  where Player  $j$ 's strategy is defined by a vector  $\tilde{x}_j = (x_j^i : i \in I)$  with  $x_j^i \in X_j$ ,  $j \in M$ . The controlled transition rates of the Markov chain are given by  $q_{k,\ell}(\underline{x}^k)$ , where we use the notation  $\underline{x}^k = (x_j^k : j \in J)$ . The payoff for Player  $j$ , when the game starts in state  $i$  and when the players use the strategy  $m$ -tuple  $\tilde{\underline{x}} = (\tilde{x}_j : j \in M)$ , is defined as follows:

$$(3.1) \quad V_j(\tilde{\underline{x}}; i) = \mathbb{E}_{\tilde{\underline{x}}} \left[ \int_0^\infty e^{-\rho_j t} \mathcal{L}_j^{\xi(t)}(\underline{x}^{\xi(t)}, 0) dt \mid \xi(0) = i \right].$$

**3.2. Nash equilibrium in the limit game.** We assume the following.

**ASSUMPTION 1.** There exists an equilibrium  $\tilde{\underline{x}}^*$  for the limit game. The equilibrium value function for Player  $j$  is given by

$$(3.2) \quad V_j^*(i) = \max_{\tilde{x}_j} \mathbb{E}_{\tilde{\underline{x}}} \left[ \int_0^\infty e^{-\rho_j t} \mathcal{L}_j^{*\xi(t)}([\underline{x}_{M-j}^{*\xi(t)}, \underline{x}_j^{*\xi(t)}], 0) dt \mid (\xi(0) = i) \right],$$

$$i \in I, \quad j \in M,$$

for each player  $j$ .

The Hamilton–Jacobi–Bellman (HJB) system of equations associated with this equilibrium is

$$(3.3) \quad \rho_j V_j^*(i) = \max_{x_j^i \in X_j} \mathcal{L}_j^i([\underline{x}_{M-j}^{*i}, x_j^i], 0) + \sum_{k \in I} q_{ik}([\underline{x}_{M-j}^{*i}, x_j^i]) V_j^*(k), \quad i \in I, j \in M.$$

*Remark 2.* The existence of an equilibrium for a sequential game has been proved in particular in [15] and [17]. A more general theory that covers the class of stochastic games considered here has been proposed in [2]. These theories could be applied to prove that a Nash equilibrium exists for the limit game. However, the assumption is more restrictive since it assumes that the equilibrium can be obtained in pure strategies.

**3.3. Occupation measures.** It will be convenient to use occupation measures to prove the main convergence results in the paper. Introducing the indicator function

$$(3.4) \quad \delta(i, k) = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{otherwise,} \end{cases}$$

we define, for any strategy  $m$ -tuple  $\tilde{x}$  and state  $k$ , given the initial state  $i$ , with  $k, i \in I$ , the occupation measures

$$(3.5) \quad \Pi_j^k(\tilde{x}, i) = \mathbb{E}_{\tilde{x}} \left[ \int_0^\infty e^{-\rho_j t} \delta(k, \xi(t)) dt \mid \xi(0) = i \right].$$

For each  $j \in M$  these occupation measures satisfy the coupled equations

$$(3.6) \quad \rho_j \Pi_j^\ell(\tilde{x}; i) - \sum_{k \in I} q_{ki}(\underline{x}^k) \Pi_j^k(\tilde{x}; i) = \delta(\ell, i), \quad i \in I.$$

We can also rewrite the payoff, given in (3.1), as

$$(3.7) \quad V_j(\tilde{x}; i) = \sum_k \Pi_j^k(\tilde{x}; i) \mathcal{L}_j^k(\underline{x}^k, 0).$$

**4. Uniform attractor policies and  $\varsigma$ -equilibria.** One can make a change of time scale in the dynamical system (2.1)–(2.3) by introducing a *stretched out* time  $\tau = \frac{t}{\varepsilon}$ . We shall assume a uniform reachability condition in this extended time scale.

**ASSUMPTION 2.** *In the stretched out time scale (or when  $\varepsilon = 1$ ), for any  $\eta > 0$  and mode  $i \in I$ , any target state  $\underline{x}^f$  in  $\underline{X}$  can be reached within  $\eta$  by the dynamical system (2.1)–(2.3) from any initial state  $\underline{x}^o \in \underline{X}$  in a uniformly bounded time. We call  $\theta(\eta)$  the uniform bound on the  $\eta$ -reachability time. In summary we assume the following:*

$$\begin{aligned} &\forall \eta > 0, \exists \theta(\eta) > 0 \text{ s.t. } \forall i \in I, \forall \underline{x}^o \text{ and } \underline{x}^f \in \underline{X}, \text{ there exists a} \\ &\text{trajectory } \underline{x}(\cdot) \text{ s.t. } \underline{x}(0) = \underline{x}^o, \mathcal{L}_j^i(\underline{x}(t), \dot{x}_j(t)) < \infty, j \in M \text{ a.e. and} \\ &\forall t > \theta(\eta) \quad \|\underline{x}(t) - \underline{x}^f\| < \eta \text{ holds.} \end{aligned}$$

We shall further assume that this reachability is achieved through the use of “admissible decentralized system behavior.”

**DEFINITION 4.1.** *An admissible decentralized system behavior is a family of mappings  $s_j^i(x_j)$  taking values in  $\mathbb{R}^{n_j}$ ,  $i \in I$ ,  $j \in J$ , and such that the differential equations  $\dot{x}_j(t) = s_j^i(x_j(t))$  admit a uniquely defined solution in  $X_j$  for  $t \in [0, \infty)$ , given  $x(0) = x^o \in X_j$  and such that  $\mathcal{L}_j^i(\underline{x}(t), s_j^i(x_j(t))) > -\infty$  a.e. on  $[0, \infty)$ .*

ASSUMPTION 3. *The uniform reachability Assumption 2 is achieved by a set of admissible decentralized system behavior  $s_j^i(x_j; \underline{x}^o, \underline{x}^f) = \dot{x}_j$  such that*

$$\mathcal{L}_j^i(\underline{x}, s_j^i(x_j; \underline{x}^o, \underline{x}^f)) < \infty \quad \text{and} \quad s_j^i(x_j^f; \underline{x}^o, \underline{x}^f) = 0.$$

*Remark 3.* The notion of admissible decentralized system behavior is closely related to the concept of decentralized feedback control laws in the state equation formulation of the dynamical system.

We also make the following continuity assumption, which should be not too restrictive.

ASSUMPTION 4. *The control laws of admissible decentralized system behavior have the following continuity property:*

$$(4.1) \quad \lim_{\underline{x} \rightarrow \underline{x}^f} s_j^i(x_j; \underline{x}^o, \underline{x}) = s_j^i(x_j^f; \underline{x}^o, \underline{x}^f) = 0.$$

**4.1. A correspondence mapping.** Given a strategy  $\tilde{\underline{x}}$  for the limit game  $G^0$ , we associate a strategy  $\tilde{\underline{\gamma}}^\varepsilon = \sigma^\varepsilon(\tilde{\underline{x}})$  for the  $G^\varepsilon$  game defined as follows: For any discrete state  $i \in I$  and any initial state  $\underline{x}^o$  select the trajectory  $\underline{x}^i(\cdot) : [0, \infty) \rightarrow \underline{X}$ , where each component  $x_j(t)$  is a solution of  $s_j^i(x_j(t); \underline{x}^o, \tilde{\underline{x}}) = \varepsilon \dot{x}_j(t)$  with  $x_j^i(0) = x_j^o$ . This is always possible by Assumption 3, and the following holds:

$$(4.2) \quad \forall \eta > 0, \exists \theta(\eta) \text{ s.t. } \|x_j^i(t) - x_j^i\| < \eta \quad \forall j \in M \quad \forall t \geq \theta(\eta)\varepsilon.$$

Define the occupation measures associated with  $\tilde{\underline{\gamma}}^\varepsilon$  in the  $G^\varepsilon$  game

$$(4.3) \quad \Pi_j^{k\varepsilon}(\tilde{\underline{\gamma}}^\varepsilon; i, x^o) = \mathbb{E}_{\tilde{\underline{\gamma}}^\varepsilon} \left[ \int_0^\infty e^{-\rho_j t} \delta(k, \xi(t)) dt \mid (\xi(0) = i, x(0)) = x^o \right].$$

For any function  $g(\underline{x})$  which is continuous the asymptotic reachability condition (4.2) implies that there exists  $\theta'(\eta)$  such that

$$(4.4) \quad |g(\underline{x}^i(t)) - g(\underline{x}^i)| < \eta \quad \forall t \geq \theta'(\eta)\varepsilon.$$

For the sake of simplifying the notation we shall use simply  $\theta$  instead of  $\theta'(\eta)$  when there is no possibility of confusion. We can now prove the following.

PROPOSITION 4.2. *For any  $i \in I$  and any  $\underline{x}^o \in \underline{X}$  the following convergence holds for the occupation measures:*

$$(4.5) \quad \lim_{\varepsilon \rightarrow 0} |\Pi_j^k(\tilde{\underline{x}}; i) - \Pi_j^{k\varepsilon}(\sigma^\varepsilon(\tilde{\underline{x}}); i, x^o)| = 0.$$

*Proof.* The detailed proof, which is straightforward but lengthy, is given in Appendix A. We summarize here its general development. As  $\delta(i, k)$  is an indicator function it is uniformly bounded and one has for any strategy  $\gamma$  for a game  $G^\varepsilon$

$$(4.6) \quad \mathbb{E}_\gamma \left[ \int_0^\infty e^{-\rho t} \delta(i, \xi(t)) dt \right] = \lim_{T \rightarrow \infty} \mathbb{E}_\gamma \left[ \int_0^T e^{-\rho t} \delta(i, \xi(t)) dt \right],$$

and this convergence is uniform for all  $\gamma$  and  $\varepsilon$ . For any realization, the integral  $\int_0^T e^{-\rho t} \delta(i, \xi(t, \omega)) dt$  can be shown to be a continuous function of the sample path  $\xi(t, \omega)$  in an appropriate norm  $d(\cdot)$  (see Appendix A). Then to establish (4.5) it suffices to show that the weak convergence limit

$$(4.7) \quad P_\varepsilon[\sigma^\varepsilon(\tilde{\underline{x}})] \Rightarrow P[\tilde{\underline{x}}]$$

holds, where  $P_\varepsilon[\sigma^\varepsilon(\underline{x})]$  and  $P[\underline{x}]$  are the probability measures induced on the restriction of the sample space to functions defined over the interval  $[0, T]$  by  $\sigma^\varepsilon(\underline{x})$  and  $\underline{x}$  for the games  $G^\varepsilon$  and  $G^0$ , respectively.

Applying<sup>1</sup> Theorem 15.4 from [1] we can say that the weak convergence property (4.7) holds if, for any finite set of sample times  $t_1, \dots, t_p$ , the probability measures induced on the  $p$  random variables  $\xi(t_1), \dots, \xi(t_p)$  by the strategy  $\sigma^\varepsilon(\underline{x})$  converge weakly to the probability measure induced by the limit game strategy  $\underline{x}$ . We summarize this by the expression

$$(4.8) \quad P_\varepsilon \pi_{t_1, \dots, t_p}^{-1} \Rightarrow P \pi_{t_1, \dots, t_p}^{-1},$$

where  $\pi_{t_1, \dots, t_p}^{-1}$  is the inverse image of the projection of the  $\xi(\cdot)$  process on the  $p$  sample times. This property is shown to hold in the rest of the proof presented in the appendix.  $\square$

The next result will establish convergence for the payoff functionals.

**PROPOSITION 4.3.** *For any strategy  $\underline{x}$  of the limit game  $G^0$  the following holds true:*

$$(4.9) \quad \lim_{\varepsilon \rightarrow 0} |V_j^\varepsilon(\sigma^\varepsilon(\underline{x}); i, x^o) - V_j(\underline{x}; i)| = 0 \quad \forall i \in I.$$

*Proof.* Consider the sample paths of the process  $\{(\xi(\cdot, \omega), \underline{x}(\cdot, \omega)) : [t, \infty) \rightarrow I \times \mathbb{R}^n : \omega \in \Omega\}$  generated by a strategy  $\underline{\gamma}^\varepsilon = \sigma^\varepsilon(\underline{x})$ . Almost surely any sample path has a countable number of jump times denoted  $t_l(\omega)$ ,  $l = 0, \dots, \infty$ . The following holds true:

$$(4.10) \quad \mathbb{E}_{\underline{\gamma}^\varepsilon} \left[ \sum_{l=0}^{\infty} e^{-\rho_j t_l} \right] \leq M.$$

Therefore we can write

$$(4.11) \quad \begin{aligned} V_j^\varepsilon(\underline{\gamma}^\varepsilon; x^o, i) &= \mathbb{E}_{\underline{\gamma}^\varepsilon} \left[ \int_0^\infty e^{-\rho_j t} \mathcal{L}_j^{\xi(t)}(\underline{x}(t), \dot{x}_j(t)) dt \mid \underline{x}(0) = x^o; \xi(0) = i \right] \\ &= \mathbb{E}_{\underline{\gamma}^\varepsilon} \left[ \sum_{l=0}^{\infty} \int_{t_l}^{t_{l+1}} e^{-\rho_j t} \mathcal{L}_j^{\xi_l}(\underline{x}(t), \dot{x}_j(t)) dt \mid \underline{x}(0) = x^o; \xi(0) = i \right] \\ &= \mathbb{E}_{\underline{\gamma}^\varepsilon} \left[ \sum_{l=0}^{\infty} \int_{t_l}^{t_{l+1}} e^{-\rho_j t} \mathcal{L}_j^{\xi_l}(\underline{x}^{\xi_l}, 0) dt \right. \\ &\quad + \sum_{l=0}^{\infty} \int_{t_l}^{\min\{t_l + \varepsilon\theta, t_{l+1}\}} e^{-\rho_j t} \left( \mathcal{L}_j^{\xi_l}(\underline{x}(t), \dot{x}_j(t)) - \mathcal{L}_j^{\xi_l}(\underline{x}^{\xi_l}, 0) \right) dt \\ &\quad + \sum_{l=0}^{\infty} \int_{\min\{t_l + \varepsilon\theta, t_{l+1}\}}^{t_{l+1}} e^{-\rho_j t} \left( \mathcal{L}_j^{\xi_l}(\underline{x}(t), \dot{x}_j(t)) - \mathcal{L}_j^{\xi_l}(\underline{x}^{\xi_l}, 0) \right) dt \\ &\quad \left. \mid \underline{x}(0) = x^o; \xi(0) = i \right]. \end{aligned}$$

<sup>1</sup>For continuous time Markov chains the condition (15.11) in [1] holds as the probability of having more than one jump in an interval  $[t, t + \delta]$  tends to zero as  $\delta$  tends to zero. Furthermore the set of trajectories that have a jump at  $t = T$  has a zero measure. Therefore we can apply Theorem 15.4 from [1].

As  $\mathcal{L}_j^i$  is bounded in  $X \times U_j$ , the second term in (4.11) tends to zero as  $\varepsilon$  tends to zero. As  $\mathcal{L}_j^i$  is continuous, the last term in (4.11) can be made as small as desired as  $\varepsilon$  tends to zero. It remains to show that the first term in (4.11) tends to  $V_j(\tilde{x}; i)$  when  $\varepsilon$  tends to zero, to conclude that (4.9) holds. Since

$$(4.12) \quad \begin{aligned} & \mathbb{E}_{\tilde{\gamma}^\varepsilon} \left[ \sum_{l=0}^{\infty} \int_{t_l}^{t_{l+1}} e^{-\rho_j t} \mathcal{L}_j^{\xi_l}(\underline{x}^l, 0) dt \mid \underline{x}(0) = x^o; \xi(0) = i \right] \\ &= \sum_{k \in I} \Pi^k(\tilde{\gamma}^\varepsilon; i, x^o) \mathcal{L}_j^k(\underline{x}^k, 0), \end{aligned}$$

the result holds true by Proposition 4.2.  $\square$

**4.2. The auxiliary  $\varepsilon$ -control problems.** Given the strategy  $\sigma^\varepsilon(\tilde{x})$  in the game  $G^\varepsilon$  and an initial state  $x^o$ , we define for each player  $j \in M$  an auxiliary  $\varepsilon$ -control problem as follows:

$$(4.13) \quad W_j^\varepsilon(\gamma_j^\varepsilon; i, \underline{x}^o \mid \sigma^\varepsilon(\tilde{x})) = \mathbb{E}_{\gamma_j} \left[ \int_0^\infty e^{-\rho_j t} L_j^{\xi(t)}(\underline{x}(t), u_j(t)) dt \mid \xi(0) = i, \underline{x}(0) = x^o \right],$$

where the Markov jump process  $\xi(\cdot)$  is still characterized by jump rates  $q_{k\ell}(\underline{x}(t))$  and the state equations in mode  $k \in I$  are

$$(4.14) \quad \varepsilon \dot{x}_j(t) = f_j^k(x_j(t), u_j(t)),$$

$$(4.15) \quad \varepsilon \dot{x}_\ell(t) = s_\ell^k(x_\ell(t); \underline{x}^k, \tilde{x}) \quad \text{if } \ell \neq j.$$

In the equations above  $(k, \underline{x}^k) = (\xi(t^{k+}), \underline{x}(t^{k+}))$  is the state of the system right after the last jump time and  $s_\ell^k(x_\ell(t); \underline{x}^k, \tilde{x})$  is the admissible decentralized system behavior associated with strategy  $\sigma^\varepsilon(\tilde{x})$  for Player  $\ell$ . This control problem for Player  $j$  is thus obtained by fixing the dynamics of the other players to their admissible decentralized system behavior, and therefore the following holds:

$$(4.16) \quad W_j^\varepsilon(\gamma_j^\varepsilon; i, \underline{x}^o \mid \sigma^\varepsilon(\tilde{x})) = V_j^\varepsilon([\sigma_{M-j}^\varepsilon(\tilde{x}), \gamma_j^\varepsilon]; i, \underline{x}^o).$$

Using these auxiliary control problems we will be able to exploit existing results from the theory of singularly perturbed systems, in particular those established in [6].

**4.3. The auxiliary limit-control problem.** For a given  $\tilde{x}$  and for a given set of potential vectors  $\tilde{w}_j = (w_j^i)_{i \in I}$  for each player  $j$ , define the Hamiltonians

$$H_j^i(\tilde{w}_j; \tilde{x}) = \max_{x_j, u_j} \left\{ L_j^i([\tilde{x}_{-j}^i, x_j], u_j) + \sum_{k \in I} q_{ik}(\tilde{x}_{-j}^i, x_j) w_j^i \mid \text{s.t. } f_j^i(x_j, u_j) = 0 \right\}.$$

For each player  $j$  we consider then the solution of the algebraic equations

$$(4.17) \quad \rho_j w_j^{i*} = H_j^i(\tilde{w}_j^*; \tilde{x}), \quad i \in I.$$

Note that these problems, for all  $j \in M$ , correspond to the solution of the limit game introduced in section 3, with the HJB equations given in (3.3).

**4.4. Convergence of the auxiliary  $\varepsilon$ -control problem.** These control problems have been studied in [6], where the following convergence result is established.

**THEOREM 4.4.** *Let the assumptions of Theorems 3 and 4 in [6] be satisfied (see Appendix B for a reminder of these assumptions). Let  $\gamma_j^{\varepsilon*}$  be the optimal strategy in the  $\varepsilon$ -control problem and let  $v_j^*$  be the optimal control in the limit-control problem; then there exists a constant  $C$  such that*

$$(4.18) \quad |W_j^\varepsilon(\gamma_j^{\varepsilon*}; i, \underline{x}^o \mid \sigma^\varepsilon(\tilde{\underline{x}}^*)) - w_j^{i*}| \leq C\varepsilon^{\frac{\alpha}{1+\alpha}}.$$

*Remark 4.* Notice that we have  $w_j^{i*} = V_j(\tilde{\underline{x}}^*; i) = V_j^*(i)$ .

**4.5. Convergence of the  $\varepsilon$ -game.** We can now establish the following result.

**THEOREM 4.5.** *Let  $\tilde{\underline{x}}^*$  be an equilibrium in the limit game  $G^0$ . Then for all positive  $\varsigma$  there exists  $\varepsilon_0$  such that for all  $0 < \varepsilon \leq \varepsilon_0$ , the strategy  $m$ -tuple  $\sigma^\varepsilon(\tilde{\underline{x}}^*)$  defines a  $\varsigma$ -Nash equilibrium for the game  $G^\varepsilon$ .*

*Proof.* Let  $\tilde{\underline{x}}^*$  be an equilibrium in the limit game. Let  $\gamma_j^\varepsilon$  be a strategy for Player  $j$  in the game  $G^\varepsilon$ . Given  $\sigma_{M-j}^\varepsilon(\tilde{\underline{x}}^*)$ , let  $\gamma_j^{\varepsilon*}$  be the optimal strategy in the  $\varepsilon$ -control problem.

Given  $\varsigma$ , there exist  $\varepsilon_0$  such that for  $\varepsilon < \varepsilon_0$  we have

$$(4.19) \quad V_j^\varepsilon([\sigma_{M-j}^\varepsilon(\tilde{\underline{x}}^*), \gamma_j^\varepsilon]; i, \underline{x}^o) = W_j^\varepsilon(\gamma_j^\varepsilon; i, \underline{x}^o \mid \sigma^\varepsilon(\tilde{\underline{x}}^*))$$

$$(4.20) \quad \leq W_j^\varepsilon(\gamma_j^{\varepsilon*}; i, \underline{x}^o \mid \sigma^\varepsilon(\tilde{\underline{x}}^*))$$

$$(4.21) \quad \leq V_j(\tilde{\underline{x}}^*; i) + \varsigma/2$$

$$(4.22) \quad \leq V_j^\varepsilon(\sigma^\varepsilon(\tilde{\underline{x}}^*); i, \underline{x}^o) + \varsigma.$$

The first equality is valid by definition. The first inequality comes from the fact that  $\gamma_j^{\varepsilon*}$  is the optimal strategy in the  $\varepsilon$ -control problem. The second inequality comes from Theorem 4.4 and Remark 4. The last inequality comes from Proposition 4.3.  $\square$

We have thus proved that the correspondence mapping introduced in section 4.1 tends to define an approximate Nash equilibrium when the time scale ratio tends to 0.

**5. The local infinite horizon open-loop games.** We have established that a correspondence mapping based on a property of uniform reachability of steady states defines an approximate equilibrium for the  $G^\varepsilon$  game when  $\varepsilon$  is small. We can now go one step further and show that such a correspondence mapping can be obtained from the equilibrium solutions of a class of local infinite horizon open-loop differential games (IHOLDGs), with the overtaking optimality criterion.

Let  $V_j^*(i) : i \in I$  be the potential function associated with Player  $j$  in the equilibrium solution for the limit game. For any discrete state  $i$  and initial continuous state  $\underline{x}(0) = \underline{x}^o$  define in the stretched out time scale the open-loop differential game with rewards over the time interval  $[0, \theta]$  given by

$$(5.1) \quad J_j^\Theta[\underline{x}^o; \underline{x}(\cdot)] = \int_0^\Theta \left\{ \mathcal{L}_j^i(\underline{x}(\tau), \dot{\underline{x}}_j(\tau)) + \sum_{\ell \in I} q_{i\ell}(\underline{x}(\tau)) V_j^*(\ell) \right\} d\tau, \quad j \in M,$$

where each player  $j$  selects an absolutely continuous trajectory  $\{x_j(\tau) : \tau \geq 0\}$  with  $x_j(0) = x_j^o$ .

**DEFINITION 5.1.** *An overtaking equilibrium for the open-loop game defined by the payoff functionals (5.1) is an  $M$ -trajectory  $(\underline{x}^*(\tau), \tau \geq 0)$  such that, for each player  $j \in M$  and trajectory  $(x_j(\tau), \tau \geq 0)$  the following holds:*

$$(5.2) \quad \liminf_{\Theta \rightarrow \infty} (J_j^\Theta[\underline{x}^o; \underline{x}^*(\cdot)] - J_j^\Theta[\underline{x}^o; [\underline{x}^{*-j}(\cdot), x_j(\cdot)]]) \geq 0.$$

ASSUMPTION 5. *The jump rates  $q_{i\ell}(\underline{x})$  are affine in  $\underline{x}$ . The reward rates  $\mathcal{L}_j^i(\underline{x}, z_j)$  are concave in  $x_j$  and  $u_j$  for each  $j \in M$  and satisfy globally the following condition, also called “strict diagonal concavity” in [16] and [4]:*

$$(5.3) \quad \begin{aligned} & \forall \underline{x}^a, \underline{x}^b \in \underline{X}, \underline{x}^a \neq \underline{x}^b \\ & \forall \zeta_j^a \in \partial_{x_j} \mathcal{L}_j^i(\underline{x}^a, z_j), \zeta_j^b \in \partial_{x_j} \mathcal{L}_j^i(\underline{x}^b, z_j), \\ & \sum_{j \in M} (\zeta_j^a - \zeta_j^b)(x_j^a - x_j^b) > 0. \end{aligned}$$

Under this assumption we can establish the following.

- An overtaking equilibrium for this game exists and is unique. It is characterized by a “turnpike,” which is an attractor for all the equilibrium trajectories emanating from different initial states  $\underline{x}^o$ .
- This attractor corresponds to the equilibrium control associated with  $i \in I$  in the limit game.
- From the overtaking equilibrium solutions to these open-loop games defined for all  $i \in I$  and all initial state  $\underline{x}^o$  we can construct an approximate  $(\varsigma)$  equilibrium for the hybrid game problem.

The existence and uniqueness results with the turnpike property have been proved in [4]. The correspondence between the turnpikes and the equilibrium solutions in the limit game is easily obtained in the following lemma.

LEMMA 5.2. *The turnpike attractor for the IHOLDG defined in (5.1) coincides with the equilibrium solution to the limit game.*

The existence and uniqueness results with the turnpike property have been proved in Carlson and Haurie [4]. The correspondence between the turnpikes and the equilibrium solution in the limit game is easily obtained in the following lemma.

LEMMA 5.3. *The turnpike attractor for the open-loop game defined in (5.1) coincides with the equilibrium solution to the limit game.*

*Proof.* Introduce for each player  $j \in M$  the Hamiltonians  $\mathcal{H}^i : \mathbb{R}^n \times \mathbb{R}^{n_j} \rightarrow \mathbb{R} \cup \{-\infty\}$  defined as

$$(5.4) \quad \mathcal{H}^i(\underline{x}, p_j) = \sup_{z_j} \left\{ \mathcal{L}_j^i(\underline{x}, z_j) + \sum_{\ell \in I} q_{i\ell}(\underline{x}) V_j^*(\ell) + p_j z_j \right\}.$$

If  $\underline{x}^*(\cdot)$  is an overtaking equilibrium at  $\underline{x}^o$ , then there exists, for each player  $j \in M$ , an absolutely continuous function  $p_j^*(\cdot)$  such that

$$(5.5) \quad \dot{x}_j(t) \in \partial_{p_j} \mathcal{H}^i(\underline{x}^*(t), p_j^*(t)),$$

$$(5.6) \quad \dot{p}_j(t) \in -\partial_{x_j} \mathcal{H}^i(\underline{x}^*(t), p_j^*(t)).$$

The turnpike is a solution of

$$(5.7) \quad 0 \in \partial_{p_j} \mathcal{H}^i(\bar{\underline{x}}^i, \bar{p}_j^i),$$

$$(5.8) \quad 0 \in -\partial_{x_j} \mathcal{H}^i(\bar{\underline{x}}^i, \bar{p}_j^i),$$

$$(5.9) \quad j \in M.$$

Under the strict diagonal concavity assumption there exists a unique solution to (5.7)–(5.8), and any solution to (5.5)–(5.6) which remains bounded is such that

$$\lim_{t \rightarrow \infty} \underline{x}^*(t) = \bar{\underline{x}}^i, \quad \lim_{t \rightarrow \infty} p_j^*(t) = \bar{p}_j^i, \quad j \in M.$$



Now it is an easy matter to check that the conditions (5.7)–(5.8) correspond to the sufficient optimality conditions for the problem

$$(5.10) \quad \max_{x_j^i} \left[ \mathcal{L}_j^i(\underline{x}_{M-j}^{*i}, x_j^i, 0) + \sum_{\ell \in I} q_{i\ell}(\underline{x}_{M-j}^{*i}, x_j^i) V_j^*(\ell) \right], \quad j \in M,$$

which is also the right-hand side for mode  $i$  of the HJB equations in the limit game. Therefore the unique turnpike defines also an equilibrium in the limit game  $G^0$ .  $\square$

It is established, in the theory of turnpikes for infinite horizon control or open-loop equilibrium problems under the overtaking optimality criterion that the uniform  $\varsigma$ -reachability condition is satisfied by the trajectories converging toward their respective attractors; see the book [5] for a complete discussion of these topics. In order to link these trajectories with a  $\varsigma$ -equilibrium of the  $G^\varepsilon$  game we need this last assumption.

**ASSUMPTION 6.** *In mode  $i$ , the overtaking trajectories of Player  $j$ , emanating from different initial states  $x_j^o$ , can be synthesized, i.e., they are obtained as solutions to a system of state equations*

$$(5.11) \quad \dot{x}_j(t) = f_j^i(x_j(t), \mu_j^{*i}(x_j(t))); \quad x_j(0) = x^o,$$

where  $\mu_j^{*i}(\cdot)$  is an admissible and continuous decentralized feedback law.

We can summarize the developments in the following.

**PROPOSITION 5.4.** *Given the potential functions associated with the Nash equilibrium of the limit game  $G^0$ , one can construct a family of IHOLDGs, with payoffs defined in (5.1). If Assumptions 5 and 6 are satisfied, the Nash equilibria of these IHOLDGs, under the overtaking optimality criterion, define a piecewise open-loop strategy for the  $G^\varepsilon$  game which is a  $\varsigma$ -equilibrium if  $\varepsilon$  is small enough.*

*Proof.* It suffices to apply Theorem 4.5 with a strategy  $m$ -tuple  $\sigma^\varepsilon(\underline{x}^*)$  obtained from the admissible decentralized system behavior (5.11).  $\square$

**6. A decomposition principle.** The result obtained can be interpreted as a decomposition principle for this two-time scale game. At a higher level one solves the limit stochastic game  $G^0$  and one obtains for each player an equilibrium steady state  $\tilde{x}_j$  and an equilibrium potential function  $V_j^*(k) : k \in I$ . These potential functions are transmitted to all players. The  $G^\varepsilon$  game is then played as follows:

*At a jump time  $t^o$  of the process  $\xi(\cdot)$ , the players observe the state  $(\xi(t^{o+}), \underline{x}(t^{o+})) = (i, \underline{x}^i)$ . Making a time translation to get  $t^o = 0$ , the players solve an IHOLDG where for each player  $j \in M$  the payoff is defined for any  $\Theta > 0$  by*

$$J_j^\Theta[\underline{x}^o; \underline{x}(\cdot)] = \int_0^\Theta \left\{ \mathcal{L}_j^i(\underline{x}(\tau), \dot{x}_j(\tau)) + \sum_{\ell \in I} q_{i\ell}(\underline{x}(\tau)) V_j^*(\ell) \right\} d\tau.$$

*The players find the unique Nash overtaking equilibrium for this open-loop game and they follow this trajectory, as long as the jump process remains in state  $i$ .*

This way of playing the game is close to being a piecewise open-loop Nash equilibrium when the time scale ratio  $\varepsilon$  is small.

**7. Examples.** In this section we provide two illustrations of the application of the theory developed above. The first one is a complete numerical computation realized on a dynamic duopoly model. In this example one shows how one can easily solve

the different dynamic games used in this approximation theory. The second illustration is the reformulation as an hybrid stochastic system of a well-known integrated assessment model of climate change with noncooperative behavior of the economic agents (groups of nations). The stochastic jump process represents modifications in climate modes. The complete development and exploitation of this model would take too much space and will be the subject of another article. We nevertheless indicate how the theory of approximation developed herein would permit an important simplification in the type of game to solve.

**7.1. A simple numerical example.** As a complete illustrative example, let us first consider a simple but nontrivial model and compute the equilibrium turnpike values of the limit game and the equilibrium trajectories of the local IHOLDG that exhibit the turnpike property. We consider a duopoly ( $M = \{1, 2\}$ ) with two slow market modes ( $I = \{0, 1\}$ ).

The fast economic dynamics is defined by the state equations that describe the accumulation of production capacities ( $x_j$ ) through investment ( $u_j$ ) by the two firms

$$(7.1) \quad f_j^i(x_j, u_j) = u_j^i - x_j^i, \quad i \in I, j \in M.$$

The slow dynamics is described by the two transition rates between market modes

$$(7.2) \quad q_{01}(x) = x_1^0 + x_2^0,$$

$$(7.3) \quad q_{10}(x) = 1,$$

and  $\xi(0) = 0$ . Typically mode  $\xi = 0$  would represent a “strong” market and mode  $\xi = 1$  would represent a “depressed” market. The transition from strength to depression is influenced by the total supply on the market. The return from depression to strength is random and not controlled.

The reward functions are the firms’ profits expressed as  $L_j^i(x, u) = a_j^i x_j^i - (u_j^i)^2$ . A common discount rate is fixed at  $\rho_j = \rho = 0.05$ . Payoffs are total expected discounted rewards over an infinite time horizon.

This dynamic duopoly model is similar (it has normalized parameter values) to the model proposed in [9], where a theory of stochastic duopoly with modal jumps is developed and a numerical solution method is proposed. We shall solve now the limit game problem and the local IHOLDGs for this singularly perturbed dynamic game.

**7.1.1. Solving the limit game.** In that particular case, it is possible to find an explicit solution to (3.6). Using Maple we obtain the following expressions for the occupation measures:

$$(7.4) \quad \Pi_j^0(0; \tilde{x}) = \frac{(1 + \rho)}{\rho(1 + \rho + x_1^0 + x_2^0)},$$

$$(7.5) \quad \Pi_j^1(0; \tilde{x}) = \frac{x_1^0 + u_2^0}{\rho(1 + \rho + x_1^0 + x_2^0)}.$$

Now using the expression (3.7) for the payoffs associated with the strategy  $\underline{x}$  in the limit game  $G^0$ , we reduce the search for a Nash equilibrium to the solution of a variational inequality which can be solved with an algorithm given by Konnov [11]. The results are displayed in Table 7.1, which provides the equilibrium steady state values for different sets of parameters  $a_j^i$  used in the reward function. These steady state values, provided by the equilibrium solutions of the limit game, indicate the target production capacity to which the duopolists should aim depending on the prevailing market mode.

TABLE 7.1  
Equilibrium policy for different values of  $a_j^i$ .

$a_1^0$	$a_2^0$	$a_1^1$	$a_2^1$	$x_1^0$	$x_2^0$	$x_1^1$	$x_2^1$
2.00	2.00	2.00	2.00	1.0000	1.0000	1.0000	1.0000
2.00	2.00	0.50	0.50	0.8325	0.8325	0.2500	0.2500
2.00	2.00	1.00	0.25	0.8665	0.8261	0.5000	0.1250
4.00	2.00	1.00	0.25	1.4865	0.8580	0.5000	0.1250
4.00	2.00	0.25	1.00	1.4572	0.8914	0.1250	0.5000

**7.1.2. Solving the local IHOLDG.** We also computed an approximation of the equilibrium trajectories of the local IHOLDG. This is done by discretizing the time scale and taking a finite, but large, time horizon. Doing so, we reduce the problem to a variational inequality that can be solved with the same algorithm [11]. Figure 7.1 displays two trajectories with different initial states  $x$ , for the state  $i = 0$  and for the case where  $a_1^0 = 4.00$ ,  $a_2^0 = 2.00$ ,  $a_1^1 = 1.00$ , and  $a_2^1 = 0.25$ . Note that in this case, the potential values (equilibrium payoffs in the limit game) are  $V_1(0) = 26.5682$ ,  $V_2(0) = 6.2776$ ,  $V_1(1) = 25.5411$ ,  $V_2(1) = 5.9936$ , respectively.

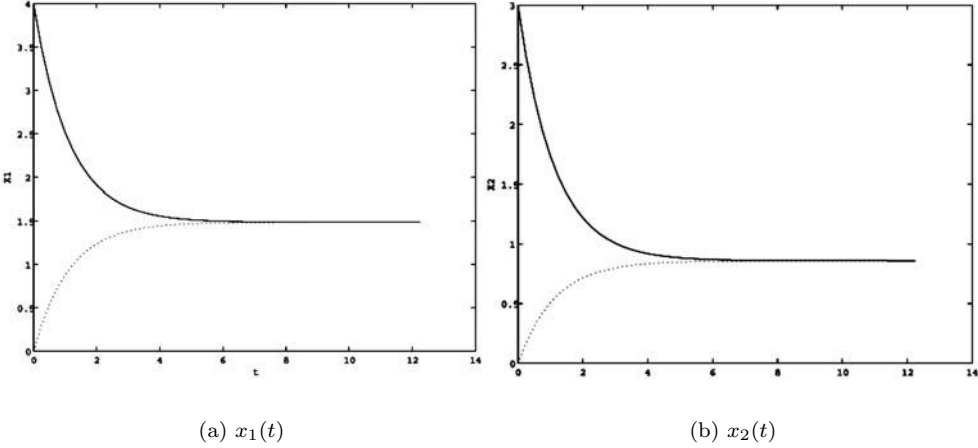


FIG. 7.1. Optimal trajectories for  $i = 0$  when  $a_1^0 = 4$ ,  $a_2^0 = 2$ ,  $a_1^1 = 1$ ,  $a_2^1 = 0.25$ , and  $\rho = 5\%$ . Solid line:  $x_1(0) = 4$  and  $x_2(0) = 3$ . Dotted line:  $x_1(0) = 0$  and  $x_2(0) = 0$ .

As expected, we distinctly see that the trajectories are attracted by the turnpike values given in Table 7.1.

This provides a complete illustration of the results obtained in this paper. The limit game equilibrium tells the player what they should do in the “slow” time scale; the local IHOLDG tells them what they could do in the “fast” time scale in order to be consistent with the long-term equilibrium solution.

**7.2. A model of competitive economic growth with climate change thresholds.** A stochastic control approach to climate change modeling has been advocated in [8] and applied in [10]. These models extended the formalism proposed by Nordhaus [12] by introducing a stochastic jump process representing sudden switches in climate or global environment conditions. Nordhaus and Yang proposed in [14] a deterministic differential game model which represents the noncooperative behavior

of different groups of nations involved in the climate change process. We propose here to extend the model by introducing a description of climate dynamics in the form of discrete modal changes. More precisely let us assume that there exist different climate modes, denoted  $\xi$ , which correspond to different patterns of the general circulation which determines climate dynamics. For example, we distinguish the current pattern ( $\xi = 0$ ) from a second pattern ( $\xi = 1$ ) where the thermo-haline circulation has been stopped and a third pattern ( $\xi = 2$ ) where, in addition, the West-Antarctic ice sheet has collapsed (this type of threshold event was considered in the stochastic control model proposed in [10]).

**7.2.1. The model equations.** Let us denote the following.

**Parameters.**

- $j = 1, \dots, m$  the  $m$  groups of nations, also called players
- $\mu_j$  capital depreciation rate in country  $j$  (typically 10% per year)
- $\nu$  greenhouse gas (GHG) natural elimination rate
- $\rho_j$  long-term discount rate for country  $j$  (typically 0.08% per decade)

**State variables.**

- $K_j$  productive capital stock of country  $j$
- $M$  GHG concentration
- $\xi$  climate mode (discrete value  $\xi \in \{0, 1, 2\}$ )

**Control variables.**

- $I_j$  investment in productive capital stock of country  $j$
- $C_j$  consumption in country  $j$
- $u_j$  abatement effort in country  $j$
- $E_j$  GHG emission in country  $j$

**Production and emission functions.**

- $F_j^{\xi(t)}(K_j, u_j)$  economic output of country; this function is increasing and concave in  $K_j$ , decreasing and concave in  $u_j$ ; its shape depends on  $\xi$
- $G_j^{\xi(t)}(K_j, u_j)$  GHG emissions of country; increasing and convex in  $K_j$ , decreasing and convex in  $u_j$ ; its shape depends on  $\xi$

**State equations.**

- $\dot{K}_j(t) = I_j(t) - \mu_j K_j(t)$  capital accumulation process
- $\dot{M}(t) = \sum_{j=1, \dots, m} E_j(t) - \nu M(t)$  GHG concentration process

**Modal jump rates.**

- $q_{k\ell}(M) = \lim_{dt \rightarrow 0} \frac{P[\xi(t+dt)=\ell | \xi(t)=k]}{dt}$  transition rate from mode  $k$  to mode  $\ell$

**Constraints.**

- $F_j^{\xi(t)}(K_j, u_j) \geq I_j(t) + C_j(t)$  the economic output of country  $j$  can be consumed or saved as an investment
- $G_j^{\xi(t)}(K_j, u_j) \leq E_j(t)$  GHG emissions of country  $j$  are bounded below by the emission function

**Payoffs.**

- $J_j = \int_0^\infty e^{-\rho_j t} U_j(C_j(t)) dt$  the long-term welfare of country  $j$ ;  $U_j(\cdot)$  is a utility function

**Initial conditions.**

$$\begin{aligned}
K_j(0) &= K_j^o && \text{initial physical capital stock of each nation} \\
M(0) &= M^o && \text{initial GHG concentration} \\
\xi(0) &= 0 && \text{initial climate mode}
\end{aligned}$$

**7.2.2. The dynamic game.** This model summarizes the situation of economies where the production activity generates GHG emissions which accumulate and may trigger a climate modal change. An abatement effort can be made by each country at a cost represented by a loss of output in the production function. The climate mode has also a direct influence, also expressed in terms of loss of output, on the economic production function. In this model the  $m$  groups of nations play a noncooperative dynamic game where the objective of the player is to reach an equilibrium for the long-term expected welfare,

$$(7.6) \quad V_j(\gamma; 0, K^o, M^o) = E_\gamma \left[ \int_0^\infty e^{-\rho_j t} U_j(C_j(t)) dt \right],$$

where  $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_m)$  is the vector of piecewise open-loop strategies, and the expectation is taken with respect to the measure induced by  $\gamma$  and the economic and climate dynamics described above. This model retains the cost-benefit analysis framework proposed in [13] and [14] and represents climate change as a stochastic modal switch process.

**7.2.3. The limit game.** Notice that this model combines decoupled dynamics (for the  $K_j$  capital stocks) and a fully coupled state equation (for the GHG concentration  $M$ ). However,  $M$  is a “passive” state variable which influences only the jump rates. It could be shown (see the paper [3] for a discussion of differential games with active and passive variables) that the turnpike property will also hold for differential games having this structure.

The model proposed here introduces the cost of climate change as a modification of the economic production function triggered by a random change of climate mode. The limit game will be a Markov game played over the discrete climate modes and where the controls are the long-term steady state values of the economies. The local IHOLDGs are infinite horizon differential games representing optimal economic growth with an adapted reward and zero discount rate.

**8. Conclusion.** We have considered here a class of hybrid games, under the piecewise open-loop information structure. We have given conditions under which, when the time scale ratio between the stochastic jump process and the deterministic part tend to 0, the Nash equilibrium solutions can be approximated by playing a family of auxiliary infinite horizon open-loop differential games (IHOLDGs). These local games are constructed using the potential functions obtained from the Markov Nash equilibrium of a simplified sequential game. This theory uses the asymptotic stability properties established for IHOLDGs under the strict diagonal concavity assumption. These conditions are the usual ones when one deals with Cournot solutions in dynamic imperfect competition models [16]. The results established in this paper are therefore useful for studying dynamic economic competition, when the market conditions change randomly but relatively seldom, compared with the adjustment speed of the economic decision variables. As indicated in the introduction we can envision such a situation in the framework of competitive economic growth with global environmental impact, like climate change triggered by greenhouse gas emissions, for example.

The evolution of the environmental state, and therefore the evaluation of the environmental damage, is random but evolving slowly. The economies of the world have the possibility to adjust at a much faster speed than the global environment modifications (see [8] for a discussion of control models with different time scales in the climate change modeling domain).

**Appendix A. Continuity and convergence of the occupation measure.**

Let  $D$  be the space of functions on  $[0, 1]$  that are right-continuous and have left-hand limits. Adapting results from [1] we prove that the integral  $\int_0^1 e^{-\rho t} \delta(i, \xi(t, \omega)) dt$  is a continuous function of  $\xi(t, \omega)$  in  $D$ . Let  $\Lambda$  denote the class of strictly increasing, continuous mappings from  $\lambda(\cdot) : [0, 1] \rightarrow [0, 1]$ . For  $\zeta(\cdot)$  and  $\xi(\cdot)$  in  $D$ , define  $d(\zeta(\cdot), \xi(\cdot))$  to be the infimum of the  $\varepsilon > 0$  for which there exists in  $\Lambda$  a  $\lambda(\cdot)$  such that

$$(A.1) \quad \sup_t |\lambda(t) - t| \leq \varepsilon$$

and

$$(A.2) \quad \sup_t |\zeta(t) - \xi(\lambda(t))| \leq \varepsilon.$$

We are now ready to prove that under the norm  $d$ , the function

$$\int_0^1 e^{-\rho t} \delta(i, \xi(t, \omega)) dt$$

is continuous. We know that  $\xi(\cdot, \omega)$  has at most countably many discontinuities. Say that  $\xi(\cdot, \omega)$  has  $N$  discontinuities. Recall that  $\xi(\cdot, \omega)$  takes value in  $I = \{1, 2, \dots, p\}$ . Therefore, given  $\xi(\cdot, \omega) \in D$  for any  $\xi(\cdot, \tilde{\omega}) \in D$  such that  $d(\xi(\cdot, \omega), \xi(\cdot, \tilde{\omega})) < \varepsilon < 1$ , both  $\xi(\cdot, \omega)$  and  $\xi(\cdot, \tilde{\omega})$  must have the same sequence of jumps since otherwise  $d(\xi(\cdot, \omega), \xi(\cdot, \tilde{\omega})) \geq 1$ . Let  $\{t_1, \dots, t_N\}$  (respectively,  $\{\tilde{t}_1, \dots, \tilde{t}_N\}$ ) be the jump times of  $\xi(\cdot, \omega)$  (respectively,  $\xi(\cdot, \tilde{\omega})$ ). By definition of the norm  $|\tilde{t}_n - t_n| < \varepsilon$  for all  $n \in \{1, \dots, N\}$ . We have therefore

$$(A.3) \quad \left| \int_0^1 e^{-\rho t} (\delta(i, \xi(t, \omega)) - \delta(i, \xi(t, \tilde{\omega}))) dt \right| \leq \int_0^1 |\delta(i, \xi(t, \omega)) - \delta(i, \xi(t, \tilde{\omega}))| dt \leq N(p-1)\varepsilon.$$

Since  $p$  and  $N$  are finite and  $\varepsilon$  can be taken as small as desired, the continuity is proved.

*Proof of Proposition 4.2.* For  $\xi(0) = i$ , let us consider in the limit game  $G^0$  and for a game  $G^\varepsilon$ , the probability to have no jump in the interval  $[0, t]$  induced by  $P[\underline{x}]$  and  $P_\varepsilon[\sigma^\varepsilon(\underline{\hat{x}})]$ , respectively. For the limit game  $G^0$  this probability is given by

$$(A.4) \quad \mathcal{P}^0[t, 0, i, i] = e^{-\int_0^t q_{ii}(\underline{x}^i) ds}.$$

For the game  $G^\varepsilon$ , under the strategy  $\sigma^\varepsilon(\underline{\hat{x}})$  this probability is given by

$$(A.5) \quad \mathcal{P}^\varepsilon[t, 0, i, i] = e^{-\int_0^t q_{ii}(\underline{x}(s)) ds},$$

which can be rewritten as

$$\mathcal{P}^\varepsilon[t, 0, i, i] = e^{-\int_0^t q_{ii}(\underline{x}^i) ds - \int_0^{\min(t, \varepsilon\theta)} (q_{ii}(\underline{x}(s)) - q_{ii}(\underline{x}^i)) ds - \int_{\min(t, \varepsilon\theta)}^t (q_{ii}(\underline{x}(s)) - q_{ii}(\underline{x}^i)) ds}.$$

As the jump rates are bounded over  $X$ , the second integral in the expression above converges to 0 when  $\varepsilon \rightarrow 0$ . When  $\varepsilon \rightarrow 0$ , the absolute value of the integral in the third term is bounded by  $\eta t$ , which can be made as small as desired by choosing  $\theta$  sufficiently large. Only the first term remains in the exponent, and this corresponds to (A.5). Therefore this establishes the convergence of  $\mathcal{P}^\varepsilon[t, 0, i, i]$  to  $\mathcal{P}^0[t, 0, i, i]$  when  $\varepsilon \rightarrow 0$ .

We can prove similarly that the probability  $\mathcal{P}^\varepsilon[t, 1, i, k]$  of having  $\xi(t) = k$  and having exactly one jump in the interval  $[0, t]$ , induced by  $\sigma^\varepsilon(\underline{x})$  for the game  $G^\varepsilon$ , converges to the probability  $\mathcal{P}^0[t, 1, i, k]$  of having  $\xi(t) = k$  with exactly one jump in the interval  $[0, t]$ , induced by  $\tilde{x}$  for the game  $G^0$ .  $\mathcal{P}^0[t, 1, i, k]$  is given by

$$(A.6) \quad \mathcal{P}^0[t, 1, i, k] = \int_0^t q_{ik}(\underline{x}^i) e^{-\int_0^s q_{ii}(\underline{x}^i) dv} e^{-\int_s^t q_{kk}(\underline{x}^k) dv} ds,$$

whereas the probability  $\mathcal{P}^\varepsilon[t, 1, i, k]$  is given by

$$(A.7) \quad \mathcal{P}^\varepsilon[t, 1, i, k] = \int_0^t q_{ik}(\underline{x}(s)) e^{-\int_0^s q_{ii}(\underline{x}(v)) dv} e^{-\int_s^t q_{kk}(\underline{x}(v)) dv} ds.$$

For a given  $t$ , if  $\varepsilon$  is small enough we have  $\varepsilon\theta < t/2$  and thus can write

$$\begin{aligned} & |\mathcal{P}^\varepsilon[t, 1, i, k] - \mathcal{P}^0[t, 1, i, k]| \\ & \leq \int_0^{\varepsilon\theta} \left| q_{ik}(\underline{x}(s)) e^{-\int_0^s q_{ii}(\underline{x}(v)) dv} e^{-\int_s^t q_{kk}(\underline{x}(v)) dv} - q_{ik}(\underline{x}^i) e^{-\int_0^s q_{ii}(\underline{x}^i) dv} e^{-\int_s^t q_{kk}(\underline{x}^k) dv} \right| ds \\ & + \int_{t-\varepsilon\theta}^t \left| q_{ik}(\underline{x}(s)) e^{-\int_0^s q_{ii}(\underline{x}(v)) dv} e^{-\int_s^t q_{kk}(\underline{x}(v)) dv} - q_{ik}(\underline{x}^i) e^{-\int_0^s q_{ii}(\underline{x}^i) dv} e^{-\int_s^t q_{kk}(\underline{x}^k) dv} \right| ds \\ & + \int_{\varepsilon\theta}^{t-\varepsilon\theta} \left| q_{ik}(\underline{x}(s)) e^{-\int_0^s q_{ii}(\underline{x}(v)) dv} e^{-\int_s^t q_{kk}(\underline{x}(v)) dv} - q_{ik}(\underline{x}^i) e^{-\int_0^s q_{ii}(\underline{x}^i) dv} e^{-\int_s^t q_{kk}(\underline{x}^k) dv} \right| ds. \end{aligned}$$

The first two terms tend to zero as  $\varepsilon$  tends to zero, whereas the last term can be rewritten as

$$\begin{aligned} & \int_{\varepsilon\theta}^{t-\varepsilon\theta} \left| (q_{ik}(\underline{x}(s)) - q_{ik}(\underline{x}^i)) e^{-\int_0^s q_{ii}(\underline{x}(v)) dv} e^{-\int_s^t q_{kk}(\underline{x}(v)) dv} \right. \\ & \left. - q_{ik}(\underline{x}^i) \left( e^{-\int_0^s q_{ii}(\underline{x}^i) dv} e^{-\int_s^t q_{kk}(\underline{x}^k) dv} - e^{-\int_0^s q_{ii}(\underline{x}(v)) dv} e^{-\int_s^t q_{kk}(\underline{x}(v)) dv} \right) \right| ds \\ & \leq \int_{\varepsilon\theta}^{t-\varepsilon\theta} |q_{ik}(\underline{x}(s)) - q_{ik}(\underline{x}^i)| \left| e^{-\int_0^s q_{ii}(\underline{x}(v)) dv} e^{-\int_s^t q_{kk}(\underline{x}(v)) dv} \right| ds \\ & + \int_{\varepsilon\theta}^{t-\varepsilon\theta} |q_{ik}(\underline{x}^i)| \left| e^{-\int_0^s q_{ii}(\underline{x}^i) dv} e^{-\int_s^t q_{kk}(\underline{x}^k) dv} - e^{-\int_0^s q_{ii}(\underline{x}(v)) dv} e^{-\int_s^t q_{kk}(\underline{x}(v)) dv} \right| ds. \end{aligned}$$

When  $\varepsilon$  tends to zero, the first term on the right-hand side of the inequality above can be made as small as desired, choosing  $\theta$  sufficiently big. The second term can be

rewritten as

$$\begin{aligned}
& \int_{\varepsilon\theta}^{t-\varepsilon\theta} \left| e^{-\int_0^s q_{ii}(\underline{x}^i) dv} e^{-\int_s^t q_{kk}(\underline{x}^k) dv} - e^{-\int_0^s q_{ii}(\underline{x}(v)) dv} e^{-\int_s^t q_{kk}(\underline{x}(v)) dv} \right| ds \\
&= \int_{\varepsilon\theta}^{t-\varepsilon\theta} \left| e^{-\int_0^{\varepsilon\theta} q_{ii}(\underline{x}^i) dv} e^{-\int_{\varepsilon\theta}^s q_{ii}(\underline{x}^i) dv} e^{-\int_s^{s+\varepsilon\theta} q_{kk}(\underline{x}^k) dv} e^{-\int_{s+\varepsilon\theta}^t q_{kk}(\underline{x}^k) dv} \right. \\
&\quad \left. - e^{-\int_0^{\varepsilon\theta} q_{ii}(\underline{x}(v)) dv} e^{-\int_{\varepsilon\theta}^s q_{ii}(\underline{x}(v)) dv} e^{-\int_s^{s+\varepsilon\theta} q_{kk}(\underline{x}(v)) dv} e^{-\int_{s+\varepsilon\theta}^t q_{kk}(\underline{x}(v)) dv} \right| ds \\
&\leq \int_{\varepsilon\theta}^{t-\varepsilon\theta} \left| e^{-\int_0^{\varepsilon\theta} q_{ii}(\underline{x}^i) dv} e^{-\int_s^{s+\varepsilon\theta} q_{kk}(\underline{x}^k) dv} \right| \\
&\quad \left| e^{-\int_{\varepsilon\theta}^s q_{ii}(\underline{x}^i) dv} e^{-\int_{s+\varepsilon\theta}^t q_{kk}(\underline{x}^k) dv} - e^{-\int_{\varepsilon\theta}^s q_{ii}(\underline{x}(v)) dv} e^{-\int_{s+\varepsilon\theta}^t q_{kk}(\underline{x}(v)) dv} \right| ds \\
&\quad + \int_{\varepsilon\theta}^{t-\varepsilon\theta} \left| e^{-\int_{\varepsilon\theta}^s q_{ii}(\underline{x}(v)) dv} e^{-\int_{s+\varepsilon\theta}^t q_{kk}(\underline{x}(v)) dv} \right| \\
&\quad \left| e^{-\int_0^{\varepsilon\theta} q_{ii}(\underline{x}^i) dv} e^{-\int_s^{s+\varepsilon\theta} q_{kk}(\underline{x}^k) dv} - e^{-\int_0^{\varepsilon\theta} q_{ii}(\underline{x}(v)) dv} e^{-\int_s^{s+\varepsilon\theta} q_{kk}(\underline{x}(v)) dv} \right| ds.
\end{aligned}$$

By continuity of the exponential and  $q(\cdot)$  functions and (4.2), this expression can be made as small as desired, when  $\varepsilon$  tends to zero. We can now extend this approach via an induction argument. Suppose that for  $n-1$ ,  $\mathcal{P}^\varepsilon[t, n-1, i, k]$  tends to  $\mathcal{P}^0[t, n-1, i, k]$  when  $\varepsilon$  tends to zero. We have

$$(A.8) \quad \mathcal{P}^\varepsilon[t, n, i, k] = \sum_{l \neq k} \int_0^t \mathcal{P}^\varepsilon[s, n-1, i, l] \mathcal{P}^\varepsilon[t-s, 1, l, k] ds.$$

Using the result proved for  $n=1$ , we can easily conclude that the following also holds:

$$(A.9) \quad \lim_{\varepsilon \rightarrow 0} \mathcal{P}^\varepsilon[t, n, i, k] = \sum_{l \neq k} \int_0^t \mathcal{P}^0[s, n-1, i, l] \mathcal{P}^0[t-s, 1, l, k] ds = \mathcal{P}^0[t, n, i, k].$$

We can thus prove, by induction, that for each finite  $n$ ,  $\mathcal{P}^\varepsilon[t, n, i, k]$  tends to  $\mathcal{P}^0[t, n, i, k]$  when  $\varepsilon$  tends to zero. Knowing that the probability of having more than  $n$  jumps tends to zero as  $n$  tends to infinity, it follows that

$$(A.10) \quad P_\varepsilon \pi_{t_1, \dots, t_k}^{-1} \Rightarrow P \pi_{t_1, \dots, t_k}^{-1},$$

and this establishes the convergence result (4.5), as was indicated at the beginning of the proof.

**Appendix B. Results of [6].** In [6] the single player case is solved and the convergence result used in this paper as Theorem 4.4 is based on the following assumption.

**ASSUMPTION 7.** For any vector  $\underline{v} = \{v(i)\}_{i \in I}$  consider the family of deterministic optimal control problems

$$\begin{aligned}
H^i(\theta, x^o, \underline{v}) &= \inf \frac{1}{\theta} \int_0^\theta \left( L^i(x(t), u(t)) + \sum_{i \in I} q_{ik}(x(t)) v(k) \right), \\
\frac{dx(t)}{dt} &= f^i(x(t), u(t)), \\
u(t) &\in U^i, \\
x(0) &= x^o.
\end{aligned}$$



One assumes that there exist two constants  $A > 0$  and  $\alpha \in (0, 1]$ , and for each  $i \in I$  a function  $H^i(\underline{v})$ , such that for each  $i \in I$ ,  $x^o \in X$ , and  $\underline{v}$  in a bounded set  $\Omega$

$$(B.1) \quad |H^i(\theta, x^o, \underline{v}) - H^i(\underline{v})| \leq \frac{1}{\theta^\alpha}.$$

Under this assumption Theorem 4.4 is established.

**Acknowledgment.** The authors thank Prof. Biancamaria D'Onofrio from the University of Rome La Sapienza for her help in this research.

#### REFERENCES

- [1] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley and Sons, New York, 1968.
- [2] M. BRETON AND P. L'ECUYER, *Noncooperative stochastic games under a local contraction assumption*, *Stochastics*, 26 (1989), pp. 227–245.
- [3] W. BROCK, *Differential games with active and passive variables*, in *Mathematical Economics and Game Theory: Essays in Honor of Oskar Morgenstern*, Springer, Berlin, 1977, pp. 34–52.
- [4] D. CARLSON AND A. HAURIE, *A turnpike theory for infinite horizon open-loop differential games with decoupled dynamics*, in *New Trends in Dynamic Games and Applications*, G. Olsder, ed., *Ann. Internat. Soc. Dynam. Games* 3, Birkhäuser, Boston, 1995, pp. 353–376.
- [5] D. CARLSON, A. HAURIE, AND A. LEIZAROWITZ, *Infinite Horizon Optimal Control: Deterministic and Stochastic Systems*, Springer, Berlin, 1991.
- [6] J. FILAR, V. GAITSGORY, AND A. HAURIE, *Control of singularly perturbed hybrid systems*, *IEEE Trans. Automat. Control*, 46 (2001), pp. 179–190.
- [7] A. HAURIE, *Piecewise deterministic differential games*, in *Differential Games and Applications*, T. Başar and P. Bernhard, eds., *Lecture Notes in Control and Inform. Sci.* 119, Springer, Berlin, 1989, pp. 114–127.
- [8] A. HAURIE, *Integrated assessment modeling for global climate change: An infinite horizon optimization viewpoint*, *Environmental Modeling and Assessment*, 8 (2003), pp. 117–132.
- [9] A. HAURIE AND M. ROCHE, *Turnpikes and computation of piecewise open-loop equilibria in stochastic differential games*, *J. Econom. Dynam. Control*, 18 (1994), pp. 317–344.
- [10] A. HAURIE AND F. MORESINO, *A stochastic control model of economic growth with environmental disaster prevention*, *Automatica*, 42 (2006), pp. 1417–1428.
- [11] I. KONNOV, *Combined relaxation methods for finding equilibrium points and solving related problems*, *Russian Math. (Iz. VUZ)*, 37 (1993), pp. 44–51.
- [12] W. NORDHAUS, *An optimal path for controlling greenhouses gases*, *Science*, 258 (1992), pp. 1315–1319.
- [13] W. NORDHAUS AND J. BOYER, *Warming the World: Economic Models of Global Warming*, MIT Press, Cambridge, MA, 2000.
- [14] W. NORDHAUS AND Z. YANG, *RICE—a regional dynamic general equilibrium model of alternative climate change strategies*, *Amer. Econom. Rev.*, 86 (1996), pp. 741–765.
- [15] A. NOWAK, *Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space*, *J. Optim. Theory Appl.*, 45 (1985), pp. 591–602.
- [16] J. B. ROSEN, *Existence and uniqueness of equilibrium points for concave n-person games*, *Econometrica*, 33 (1965), pp. 520–534.
- [17] W. WHITT, *Representation and approximation of noncooperative sequential games*, *SIAM J. Control Optim.*, 18 (1980), pp. 33–48.

## SUFFICIENT OPTIMALITY CONDITIONS FOR DIRICHLET BOUNDARY CONTROL OF WAVE EQUATIONS\*

ANDRZEJ NOWAKOWSKI†

**Abstract.** We study optimal control problems for wave equations (focusing on the multidimensional wave equation) with control functions in the Dirichlet boundary conditions under pointwise control (and we admit state—by assuming weak hypotheses) constraints.

**Key words.** sufficient optimality condition, wave equations, Dirichlet boundary controls, dual dynamic programming

**AMS subject classifications.** Primary, 49K20, 49J20; Secondary, 93C20, 35L20

**DOI.** 10.1137/050644008

The paper we present is devoted to the study of optimal control problems for state-constrained wave equations with controls as well in state equation as in Dirichlet boundary conditions. We focus our attention to the following problem (P) governed by the multidimensional wave equation:

$$\begin{aligned} \text{minimize } J(x, u, v) = & \int_{[0,T] \times \Omega} L(t, z, x(t, z), u(t, z)) dt dz \\ & + \int_{\Sigma} h(t, z, v(t, z)) dt dz + \int_{\Omega} l(x(T, z)) dz \end{aligned}$$

subject to

$$(0.1) \quad x_{tt}(t, z) - \Delta_z x(t, z) = f(t, z, x(t, z), u(t, z)) \quad \text{a.e. on } (0, T) \times \Omega,$$

$$(0.2) \quad x(0, z) = \varphi(0, z), \quad x_t(0, z) = \psi(0, z) \quad \text{on } \Omega,$$

$$(0.3) \quad x(t, z) = v(t, z) \quad \text{on } (0, T) \times \Gamma,$$

$$(0.4) \quad u(t, z) \in U \quad \text{a.e. on } (0, T) \times \Omega,$$

$$(0.5) \quad v(t, z) \in \mathbf{V} \quad \text{on } (0, T) \times \Gamma,$$

where  $\Omega$  is a given bounded domain of  $R^n$  with boundary  $\Gamma = \partial\Omega$  of  $C^2$ ,  $\Sigma = (0, T) \times \Gamma$ ;  $U \subset R^m$  and  $\mathbf{V} \subset R$  are given nonempty sets,  $\mathbf{V}$ -closed;  $L, f : [0, T] \times \bar{\Omega} \times R \times R^m \rightarrow R$ ,  $l : R \rightarrow R$ ,  $h : [0, T] \times \bar{\Omega} \times R \rightarrow R$ , and  $\varphi, \psi : R^{n+1} \rightarrow R$  are given functions;  $\varphi(0, \cdot) \in L^2(\Omega)$ ,  $\psi(0, \cdot) \in H^{-1}(\Omega)$ ;  $x : [0, T] \times \Omega \rightarrow R$ ,  $x \in W^{2,2}((0, T) \times \Omega) \cap C([0, T]; L^2(\Omega))$ ; and  $u : [0, T] \times \Omega \rightarrow R^m$ ,  $v : [0, T] \times \Gamma \rightarrow R$  are Lebesgue measurable functions in suitable sets. We assume that the functions  $L, f, h, l$  are lower semicontinuous in their domains of definitions. Assuming the lower semicontinuity of these functions only, we

---

\*Received by the editors November 1, 2005; accepted for publication (in revised form) July 23, 2007; published electronically January 4, 2008.

<http://www.siam.org/journals/sicon/47-1/64400.html>

†Faculty of Math, University of Lodz, Banacha 22, 90-238 Lodz, Poland (annowako@math.uni.lodz.pl).

admit that state  $x$  may satisfy some pointwise state constraints, e.g., that  $x(t, z) \in C$  for a.e.  $(t, z) \in [0, T] \times \Omega$  with  $C$  a closed set in  $R$ . We call a trio  $x(t, z)$ ,  $u(t, z)$ ,  $v(t, z)$  admissible if it satisfies (0.1)–(0.5) and  $L(t, z, x(t, z), u(t, z))$ ,  $h(t, z, v(t, z))$  are summable; then the corresponding trajectory  $x(t, z)$  is said to be admissible.

It is well known that optimal control problems with pointwise state constraints belong to one of the most challenging and difficult classes in control theory. Quite recently, growing interest in such problems for parabolic equations has been taken in [1], [2], [11], [13], [25], [26]; see also the references therein. Much less has been done for wave equations. Some control problems for the wave equation in the presence of state constraints are considered in [12], [31], [32], [23], [24] for distributed controls. There are only a few results [23], [24] on boundary control problems for the wave equation and/or for other partial differential equations (PDEs) of the hyperbolic type. Note that there are essential differences between parabolic and hyperbolic systems. Generally, hyperbolic equations exhibit less regularity. This is why in that paper we assume that system (0.1)–(0.5) admits at least one solution belonging to  $W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ . About the existence and regularity problems for that system, see, e.g., [20] and the extensive discussion of our problem in [23], [24].

The aim of the paper is to present sufficient optimality conditions for problem (P) in terms of dynamic programming conditions directly. In the literature, there are no works which study problem (P) directly by a dynamic programming method. The only results known to the author for the parabolic (as well as abstract) case (see, e.g., [7], [8], [9], [10], [11], [12], [13], [14], [15], [22], [27] and the references therein) treat problem (P) first as an abstract problem with an abstract evolution equation (0.1) and then derive from abstract Hamilton–Jacobi equations suitable sufficient optimality conditions for problem (P). We would like to stress that the problem with Dirichlet boundary control is rather difficult to treat by abstract formulation of the problem as then the abstract space of function on  $\Omega$  must depend on control also on  $\partial\Omega$ , i.e., we need to consider the abstract space depending on control. We refer the reader to [12], [20] and their bibliographies for more discussions on important differences between parabolic and hyperbolic systems.

We propose almost a direct method to study (P) by a dual dynamic programming approach following the method described in [28] for the one-dimensional case and in [16] for the multidimensional case. We move all notions of a dynamic programming to a dual space (the space of multipliers) and then develop a dual dynamic approach together with a dual Hamilton–Jacobi equation and as a consequence sufficient optimality conditions for (P). We also define an optimal dual feedback control and formulate sufficient conditions for optimality in terms of it. Such an approach allows us to significantly weaken the assumptions on the data. An approximate minimum in terms of the dual dynamic programming is also investigated.

**1. A dual dynamic programming.** In this section we describe an intuition of a dual dynamic approach to optimal control problems governed by wave equations. Let us recall what is meant by dynamic programming. We have an initial condition  $(t_0, x_0(t_0, z))$ ,  $z \in \Omega$ , and for it assume we have an optimal solution  $(\bar{x}, \bar{u}, \bar{v})$ . Then by necessary optimality conditions (see, e.g., [23]) there exists a function  $\bar{p}(t, z) = (y^0, y(t, z))$  on  $(0, T) \times \Omega$ —conjugate function—that is a solution to the corresponding adjoint system. That  $p = (y^0, y)$  plays the role of multiplier from the classical Lagrange problem with constraints (with multiplier  $y^0$  corresponding to the functional and  $y$  corresponding to the constraint). If we perturb  $(t_0, x_0)$ , then assuming that the optimal solution for each perturbed problem exists we also have

corresponding to it a conjugate function. Therefore making perturbations of our initial conditions, we obtain two sets of functions: optimal trajectories  $\bar{x}$  and corresponding to them conjugate functions  $\bar{p}$ . The graph of those sets of functions cover some sets in state space  $(t, z, x)$ , say a set  $X$  (in the classical calculus of variations it is called the field of extremals), and in conjugate space  $(t, z, p)$ , say a set  $P$  (in classical mechanics it is called the space of momentums). In the classical dynamic programming (see, e.g., [7]) approach, we explore the state space  $(t, z, x)$ , i.e., the set  $X$ , and in the dual dynamic programming (see [28] for the one-dimensional case and [16] for the multidimensional case) approach we explore the conjugate space (the dual space)  $(t, z, p)$ , i.e., the set  $P$ . It is worth noting that in elliptic control optimization problems there is no possibility of perturbing the problems; however, we can still apply dual dynamic programming (see [17]). It is natural that if we want to explore the dual space  $(t, z, p)$ , then we need a mapping between the set  $P$  and the set  $X$ ,  $P \ni (t, z, p) \rightarrow (t, z, \tilde{x}(t, z, p)) \in X$ , to make it possible at the end of some consideration in  $P$  to formulate some conditions about optimality for our original problem as well as on an optimal solution  $\bar{x}$ . (In the classical calculus of variations, when we form a field of extremals, we do have such a mapping  $(t, \beta) \rightarrow (t, x(t, \beta))$  which parametrizes by  $\beta$  the field of extremals.) Of course, such a mapping should have the following property (in calculus of variations  $(t, \beta) \rightarrow (t, x(t, \beta))$  has this property): for each admissible function  $x(t, z)$  lying in  $X$  there must exist a function  $p(t, z)$  lying in  $P$  such that  $x(t, z) = \tilde{x}(t, z, p(t, z))$ . Hence, we do all our investigations in a dual space  $(t, z, p)$  (in Young [33] it is done for fields of extremals of calculus of variations in the space  $(t, \beta)$ ); i.e., most of our notions concerning dynamic programming are defined in the dual space and thus also a dynamic programming equation, which becomes now a dual dynamic programming equation. We would like to stress that Weierstrass sufficient optimality conditions which use the field of extremals are only optimality conditions concerning the relative minimum, i.e.,  $\tilde{x}$  is an argument of minimum relative to all admissible trajectories whose graphs are lying in the set covered by the graphs of extremals of the field  $x(t, \beta)$ . We do exactly the same. Our function corresponding to the field is now  $\tilde{x}(t, z, p)$  and we are able also to investigate only the relative minimum, relative to the set  $X$ . In calculus of variations  $x(t, \beta)$  is defined by a field of extremals; in our case  $\tilde{x}(t, z, p)$  will be defined by derivative of function which satisfies a dual dynamic equation and which to some extent corresponds to a function defining the field of extremals in the classical setting; i.e., if we confine ourselves to the one-dimensional calculus of variations problem, then that function will define exactly a field of extremals.

Therefore let  $P \subset R^{n+3}$  be a set of the variables  $(t, z, p) = (t, z, y^0, y)$ ,  $(t, z) \in [0, T] \times \bar{\Omega}$ ,  $y^0 \leq 0$ ,  $y \in R$ , and let  $\bar{c} = (c^0, c) \in R^2$  be fixed. The constant  $\bar{c}$  is introduced for practical purpose only, in order to make easier calculations of some relations stated below for concrete problems (see section 4). We adopt the convention that  $\bar{c}p = (c^0 y^0, cy)$  for  $(t, z, p) \in P$ . Let  $\tilde{x} : P \rightarrow R$  be such a function of the variables  $(t, z, p)$  that for each admissible trajectory  $x(t, z)$  there exists a function  $p(t, z) = (y^0, y(t, z))$ ,  $p \in W^{2,2}([0, T] \times \bar{\Omega}) \cap C([0, T]; L^2(\Omega))$ ,  $(t, z, p(t, z)) \in P$ , such that

$$(1.1) \quad x(t, z) = \tilde{x}(t, z, p(t, z)) \quad \text{for } (t, z) \in [0, T] \times \bar{\Omega}.$$

Now, let us introduce an auxiliary function  $V(t, z, p) : P \rightarrow R$  being of  $C^2$  such that

the following two conditions are satisfied:

$$(1.2) \quad V(t, z, \bar{c}p) = c^0 y^0 V_{y^0}(t, z, \bar{c}p) + cy V_y(t, z, \bar{c}p) = \bar{c}p V_p(t, z, \bar{c}p) \\ \text{for } (t, z) \in (0, T) \times \Omega, (t, z, \bar{c}p) \in P,$$

$$(1.3) \quad \nabla_z V(t, z, \bar{c}p) \nu(z) = c^0 y^0 \nabla_z V_{y^0}(t, z, \bar{c}p) \nu(z) \\ \text{for } (t, z) \in [0, T] \times \partial\Omega, (t, z, \bar{c}p) \in P,$$

where  $\nu(\cdot)$  is the exterior unit normal vector to  $\partial\Omega$  and  $\nabla V(t, z, p)$  means “ $\nabla$ ” of the function  $z \rightarrow V(t, z, p)$ . The condition (1.2) is a generalization of the transversality condition known in classical mechanics as orthogonality of momentum to the front of the wave. The condition (1.3) has the same meaning but taken on the boundary. Similarly as in classical dynamic programming, define at  $(t, \tilde{p}(\cdot))$ , where  $\tilde{p}(\cdot) = (\tilde{y}^0, \tilde{y}(\cdot))$  is any function  $\tilde{p} \in W^{2,2}(\Omega)$ ,  $(t, z, \tilde{p}(z)) \in P$ ,  $(t, z) \in [0, T] \times \Omega$ , a dual value function  $S_D$  by the formula

$$(1.4) \quad S_D(t, \tilde{p}(\cdot)) := \inf \left\{ -c^0 \tilde{y}^0 \int_{[t, T] \times \Omega} L(\tau, z, x(\tau, z), u(\tau, z)) d\tau dz \right. \\ \left. - c^0 \tilde{y}^0 \int_{\Omega} l(x(T, z)) dz - c^0 \tilde{y}^0 \int_{[t, T] \times \partial\Omega} h(\tau, z, v(\tau, z)) d\tau dz \right\},$$

where the infimum is taken over all admissible trios  $x(\tau, \cdot)$ ,  $u(\tau, \cdot)$ ,  $v(\tau, \cdot)$ ,  $\tau \in [t, T]$ , such that

$$(1.5) \quad x(t, z) = \tilde{x}(t, z, \tilde{p}(z)) \quad \text{for } z \in \Omega,$$

$$(1.6) \quad \tilde{x}(t, z, \tilde{p}(z)) = v(t, z) \quad \text{for } z \in \partial\Omega,$$

i.e., whose trajectories start at  $(t, \tilde{x}(t, \cdot, \tilde{p}(\cdot)))$  and for which there exists such a function  $p(\tau, z) = (\tilde{y}^0, y(\tau, z))$ ,  $p \in W^{2,2}([t, T] \times \bar{\Omega}) \cap C([t, T]; L^2(\Omega))$ ,  $(\tau, z, \bar{c}p(\tau, z)) \in P$ , such that  $x(\tau, z) = \tilde{x}(\tau, z, \bar{c}p(\tau, z))$  for  $(\tau, z) \in (t, T) \times \bar{\Omega}$  and

$$(1.7) \quad y(t, z) = \tilde{y}(z) \quad \text{for } z \in \bar{\Omega}.$$

Then, integrating (1.2) over  $\Omega$ , for any function  $\tilde{p}(\cdot) = (\tilde{y}^0, \tilde{y}(\cdot))$ ,  $\tilde{p} \in W^{2,2}(\Omega)$ ,  $(t, z, \tilde{p}(z)) \in P$ ,  $(t, z, \bar{c}\tilde{p}(z)) \in P$ , such that  $x(\cdot, \cdot)$  satisfying  $x(t, z) = \tilde{x}(t, z, \tilde{p}(z))$  for  $z \in \bar{\Omega}$  is an admissible trajectory, we also have the equalities

$$(1.8) \quad \int_{\Omega} V(t, z, \bar{c}\tilde{p}(z)) dz + \int_{\partial\Omega} \nabla_z V(t, z, \bar{c}\tilde{p}(z)) \nu(z) dz \\ = -c \int_{\Omega} \tilde{y}(z) x(t, z, \tilde{p}(z)) dz - S_D(t, \tilde{p}(\cdot)),$$

with

$$(1.9) \quad \int_{\Omega} c^0 \tilde{y}^0 V_{y^0}(t, z, \bar{c}\tilde{p}(z)) dz + c^0 \tilde{y}^0 \int_{\partial\Omega} \nabla_z V_{y^0}(t, z, \bar{c}\tilde{p}(z)) \nu(z) dz = -S_D(t, \tilde{p}(\cdot))$$

and assuming

$$\tilde{x}(t, z, \tilde{p}(z)) = -V_y(t, z, \bar{c}\tilde{p}(z)) \quad \text{for } (t, z) \in (0, T) \times \bar{\Omega}, (t, z, \bar{c}\tilde{p}(z)) \in P.$$

Denote by the symbol  $\Delta_z h$  the sum of the second partial derivatives of the function  $h : P \longrightarrow R$  with respect to the variable  $z_i$ ,  $i = 1, \dots, n$ , i.e.,

$$(1.10) \quad \Delta_z h(t, z, p) := \sum_{i=1}^n \frac{\partial^2 h}{\partial z_i^2}(t, z, p).$$

It turns out that the function  $V(t, z, p)$  as defined by (1.8), (1.9) satisfies the second order partial differential system

$$(1.11) \quad \begin{aligned} V_{tt}(t, z, \bar{c}p) - \Delta_z V(t, z, \bar{c}p) + H(t, z, -V_y(t, z, \bar{c}p), \bar{c}p) &= 0, \\ (t, z) &\in (0, T) \times \Omega, \quad (t, z, \bar{c}p) \in P, \\ \nabla_z V(t, z, \bar{c}p)\nu(z) + H_\Sigma(t, z, \bar{c}p) &= 0, \quad (t, z) \in (0, T) \times \partial\Omega, \quad (t, z, \bar{c}p) \in P, \end{aligned}$$

where

$$(1.12) \quad \begin{aligned} H(t, z, x, \bar{c}p) &= c^0 y^0 L(t, z, x, u(t, z, p)) + cyf(t, z, x, u(t, z, p)), \\ H_\Sigma(t, z, \bar{c}p) &= c^0 y^0 h(t, z, v(t, z, p)) \end{aligned}$$

and  $u(t, z, p)$ ,  $v(t, z, p)$  are optimal dual feedback controls, respectively, on  $(0, T) \times \Omega$  and  $(0, T) \times \partial\Omega$ , and the dual second order partial differential system of multidimensional dynamic programming (DSPDEMDP)

$$(1.13) \quad \begin{aligned} \sup \{ &V_{tt}(t, z, \bar{c}p) - \Delta_z V(t, z, \bar{c}p) + c^0 y^0 L(t, z, -V_y(t, z, \bar{c}p), u) \\ &+ cyf(t, z, -V_y(t, z, \bar{c}p), u) : u \in U \} = 0, \quad (t, z) \in (0, T) \times \Omega, \quad (t, z, \bar{c}p) \in P, \\ \sup \{ &\nabla_z V(t, z, \bar{c}p)\nu(z) + c^0 y^0 h(t, z, v) : v \in \mathbf{V} \} = 0, \\ &(t, z) \in (0, T) \times \partial\Omega, \quad (t, z, \bar{c}p) \in P. \end{aligned}$$

Let us note that the function  $\tilde{x}(t, z, p)$  was introduced at the beginning of this section a little bit artificially; in fact it is defined by  $-V_y(t, z, p)$ , where  $V$  is a solution to (1.13); i.e., knowing the set  $P$  and  $V_y$  we are able to know the set  $\dot{X}$ , where we need to consider our original problem.

*Remark.* We would like to stress that the duality which is sketched in this section is not a duality in the sense of convex optimization. It is a new nonconvex duality, first described in [28] and next developed in [16], for which we do not have the relation  $\sup(\mathbf{D}) \leq \inf(\mathbf{P})$  ( $\mathbf{D}$  signifies a dual problem,  $\mathbf{P}$  a primal one). But instead we have other relations, namely, (1.2) and (1.8), (1.9), which are generalizations of transversality conditions from classical mechanics. If we find a solution to (1.13), then checking the relation (1.2) for concrete problems is not very difficult.

**2. A verification theorem.** The most important conclusion of a dynamic programming is a verification theorem. We present it in a dual form according to our dual dynamic programming approach described in the previous section.

**THEOREM 2.1.** *Let  $\bar{x}(t, z)$ ,  $\bar{u}(t, z)$ ,  $(t, z) \in (0, T) \times \bar{\Omega}$ ,  $\bar{v}(t, z)$ ,  $(t, z) \in (0, T) \times \partial\Omega$ , be an admissible trio. Assume that there exist  $\bar{c} = (c^0, c) \in R^2$  and a  $C^2$  solution  $V(t, z, p)$  of DSPDEMDP (1.13) on  $P$  such that (1.2), (1.3) hold. Further, let  $\bar{p}(t, z) = (\bar{y}^0, \bar{y}(t, z))$ ,  $\bar{p} \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ ,  $\bar{p} \in L^2([0, T] \times \partial\Omega)$ ,  $(t, z, \bar{c}p(t, z)) \in$*

$P$ , be such a function that  $\bar{x}(t, z) = -V_y(t, z, \bar{c}p(t, z))$  for  $(t, z) \in (0, T) \times \bar{\Omega}$ . Suppose that  $V(t, z, p)$  satisfies the boundary condition for  $(T, z, \bar{c}p) \in P$ ,

$$(2.1) \quad c^0 \bar{y}^0 \int_{\Omega} (d/dt) V_{y^0}(T, z, \bar{c}p) dz = c^0 \bar{y}^0 \int_{\Omega} l(-V_y(T, z, \bar{c}p)) dz.$$

Moreover, assume that

$$(2.2) \quad \begin{aligned} & V_{tt}(t, z, \bar{c}p(t, z)) - \Delta_z V(t, z, \bar{c}p(t, z)) + c^0 \bar{y}^0 L(t, z, -V_y(t, z, \bar{c}p(t, z)), \bar{u}(t, z)) \\ & + c \bar{y}(t, z) f(t, z, -V_y(t, z, \bar{c}p(t, z)), \bar{u}(t, z)) = 0 \quad \text{for } (t, z) \in (0, T) \times \Omega, \end{aligned}$$

$$(\nabla_z) V(t, z, \bar{c}p(t, z)) \nu(z) + c^0 y^0 h(t, z, \bar{v}(t, z)) = 0 \quad \text{for } (t, z) \in (0, T) \times \partial\Omega.$$

Then  $\bar{x}(t, z)$ ,  $\bar{u}(t, z)$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $\bar{v}(t, z)$ ,  $(t, z) \in (0, T) \times \partial\Omega$ , is an optimal trio relative to all admissible trios  $x(t, z)$ ,  $u(t, z)$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $v(t, z)$ ,  $(t, z) \in (0, T) \times \partial\Omega$ , for which there exists such a function  $p(t, z) = (\bar{y}^0, y(t, z))$ ,  $p \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ ,  $p \in L^2([0, T] \times \partial\Omega)$ ,  $(t, z, \bar{c}p(t, z)) \in P$ , such that  $x(t, z) = -V_y(t, z, \bar{c}p(t, z))$  for  $(t, z) \in (0, T) \times \Omega$ ,  $v(t, z) = -V_y(t, z, \bar{c}p(t, z))$  for  $(t, z) \in (0, T) \times \partial\Omega$  and

$$(2.3) \quad y(0, z) = \bar{y}(0, z) \quad \text{for } z \in \Omega.$$

*Proof.* Let  $x(t, z)$ ,  $u(t, z)$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $v(t, z)$ ,  $(t, z) \in (0, T) \times \partial\Omega$ , be an admissible trio for which there exists such a function  $p(t, z) = (\bar{y}^0, y(t, z))$ ,  $p \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ ,  $p \in L^2([0, T] \times \partial\Omega)$ ,  $(t, z, \bar{c}p(t, z)) \in P$ , such that  $x(t, z) = -V_y(t, z, \bar{c}p(t, z))$  for  $(t, z) \in (0, T) \times \Omega$ ,  $v(t, z) = -V_y(t, z, \bar{c}p(t, z))$  for  $(t, z) \in (0, T) \times \partial\Omega$  and (2.3) is satisfied. From transversality condition (1.2), (1.3), we obtain that for  $(t, z) \in (0, T) \times \Omega$ ,

$$(2.4) \quad \begin{aligned} & V_{tt}(t, z, \bar{c}p(t, z)) - \Delta_z V(t, z, \bar{c}p(t, z)) \\ & = c^0 \bar{y}^0 \left[ (d^2/dt^2) V_{y^0}(t, z, \bar{c}p(t, z)) - (\Delta_z) V_{y^0}(t, z, \bar{c}p(t, z)) \right] \\ & + cy(t, z) \left[ (d^2/dt^2) V_y(t, z, \bar{c}p(t, z)) - (\Delta_z) V_y(t, z, \bar{c}p(t, z)) \right] \end{aligned}$$

(since  $V$  is of  $C^2$ ,  $V_y(t, z, \bar{c}p(t, z)) = -x(t, z)$ , and  $x \in W^{2,2}([0, T] \times \Omega)$ , therefore, by (1.2), the above derivatives make sense) and for  $(t, z) \in (0, T) \times \partial\Omega$ ,

$$(2.5) \quad (\nabla_z) V(t, z, \bar{c}p(t, z)) \nu(z) = c^0 \bar{y}^0 (\nabla_z) V_{y^0}(t, z, \bar{c}p(t, z)) \nu(z).$$

Since  $x(t, z) = -V_y(t, z, \bar{c}p(t, z))$ , for  $(t, z) \in (0, T) \times \bar{\Omega}$ , (0.1) shows that for  $(t, z) \in (0, T) \times \Omega$ ,

$$(2.6) \quad \begin{aligned} & (d^2/dt^2) V_y(t, z, \bar{c}p(t, z)) - (\Delta_z) V_y(t, z, \bar{c}p(t, z)) \\ & = -f(t, z, -V_y(t, z, \bar{c}p(t, z)), u(t, z)) \end{aligned}$$

and boundary control (0.3) shows that for  $(t, z) \in (0, T) \times \partial\Omega$ ,

$$-V_y(t, z, \bar{c}p(t, z)) = v(t, z).$$

We conclude from (2.4)–(2.6) that for  $(t, z) \in (0, T) \times \Omega$ ,

$$(2.7) \quad \begin{aligned} & c^0 \bar{y}^0 \left[ (d^2/dt^2) V_{y^0}(t, z, \bar{c}p(t, z)) - (\Delta_z) V_{y^0}(t, z, \bar{c}p(t, z)) \right. \\ & \quad \left. + L(t, z, -V_y(t, z, \bar{c}p(t, z)), u(t, z)) \right] \\ & = V_{tt}(t, z, p(t, z)) - \Delta_z V(t, z, p(t, z)) + c^0 \bar{y}^0 L(t, z, -V_y(t, z, \bar{c}p(t, z)), u(t, z)) \\ & \quad + cy(t, z) f(t, z, -V_y(t, z, \bar{c}p(t, z)), u(t, z)) \end{aligned}$$

and for  $(t, z) \in (0, T) \times \partial\Omega$ ,

$$(2.8) \quad \begin{aligned} & c^0 \bar{y}^0 (\nabla_z) V_{y^0}(t, z, \bar{c}p(t, z)) \nu(z) + c^0 \bar{y}^0 h(t, z, v(t, z)) \\ &= (\nabla_z) V(t, z, \bar{c}p(t, z)) \nu(z) + c^0 \bar{y}^0 h(t, z, v(t, z)). \end{aligned}$$

Hence, by (1.13) and (2.7), we infer that

$$(2.9) \quad \begin{aligned} & c^0 \bar{y}^0 \left[ (d^2/dt^2) V_{y^0}(t, z, \bar{c}p(t, z)) - (\Delta_z) V_{y^0}(t, z, \bar{c}p(t, z)) \right. \\ & \left. + L(t, z, -V_y(t, z, \bar{c}p(t, z)), u(t, z)) \right] \leq 0 \quad \text{for } (t, z) \in (0, T) \times \Omega \end{aligned}$$

and for  $(t, z) \in (0, T) \times \partial\Omega$ ,

$$(2.10) \quad c^0 \bar{y}^0 (\nabla_z) V_{y^0}(t, z, \bar{c}p(t, z)) \nu(z) + c^0 \bar{y}^0 h(t, z, v(t, z)) \leq 0,$$

and finally, after integrating (2.9) and applying (2.10), that

$$(2.11) \quad \begin{aligned} & c^0 \bar{y}^0 \int_{[0, T] \times \Omega} \left[ (d^2/dt^2) V_{y^0}(t, z, \bar{c}p(t, z)) - (\operatorname{div} \nabla_z) V_{y^0}(t, z, \bar{c}p(t, z)) \right] dt dz \\ & \leq -c^0 \bar{y}^0 \int_{[0, T] \times \Omega} L(t, z, x(t, z), u(t, z)) dt dz. \end{aligned}$$

Similarly, in the set  $(0, T) \times \partial\Omega$  we have

$$(2.12) \quad \begin{aligned} & c^0 \bar{y}^0 \int_{[0, T] \times \partial\Omega} (\nabla_z) V_{y^0}(t, z, \bar{c}p(t, z)) \nu(z) dt dz \\ & \leq -c^0 \bar{y}^0 \int_{[0, T] \times \partial\Omega} h(t, z, v(t, z)) dt dz. \end{aligned}$$

Thus from (2.11), (2.12), (2.1), (2.3), and the Green formula it follows that

$$(2.13) \quad \begin{aligned} & c^0 \bar{y}^0 \int_{\Omega} \left[ l(-V_y(T, z, \bar{c}p(T, z))) - (d/dt) V_{y^0}(0, z, c^0 \bar{y}^0, c\bar{y}(0, z)) \right] dz \\ & - c^0 \bar{y}^0 \int_{[0, T]} \left( \int_{\partial\Omega} (\nabla_z) V_{y^0}(t, z, c^0 \bar{y}^0, cy(t, z)) \nu(z) dz \right) dt \\ & \leq -c^0 \bar{y}^0 \int_{[0, T] \times \Omega} L(t, z, x(t, z), u(\tau, z)) dt dz. \end{aligned}$$

So by (2.13) and (2.12) we get

$$(2.14) \quad \begin{aligned} & -c^0 \bar{y}^0 \int_{\Omega} (d/dt) V_{y^0}(0, z, c^0 \bar{y}^0, c\bar{y}(0, z)) dz \\ & \leq -c^0 \bar{y}^0 \int_{[0, T] \times \Omega} L(t, z, x(t, z), u(t, z)) dt dz - c^0 \bar{y}^0 \int_{\Omega} l(x(T, z)) dz \\ & - c^0 \bar{y}^0 \int_{[0, T] \times \partial\Omega} h(t, z, v(t, z)) dt dz. \end{aligned}$$

In the same manner applying (2.2) and (2.7) we have for  $(t, z) \in (0, T) \times \Omega$

$$\begin{aligned} & c^0 \bar{y}^0 \left[ (d^2/dt^2) V_{y^0}(t, z, \bar{c}p(t, z)) - (\Delta_z) V_{y^0}(t, z, \bar{c}p(t, z)) \right. \\ & \left. + L(t, z, -V_y(t, z, \bar{c}p(t, z)), \bar{u}(t, z)) \right] = 0, \end{aligned}$$



and for  $(t, z) \in (0, T) \times \partial\Omega$ ,

$$c^0 \bar{y}^0 (\nabla_z) V_{y^0}(t, z, \bar{c}p(t, z)) \nu(z) + c^0 \bar{y}^0 h(t, z, \bar{v}(t, z)) = 0.$$

Further, we have

$$\begin{aligned} & -c^0 \bar{y}^0 \int_{\Omega} (d/dt) V_{y^0}(0, z, c^0 \bar{y}^0, c\bar{y}(0, z)) dz \\ (2.15) \quad & = -c^0 \bar{y}^0 \int_{[0, T] \times \Omega} L(t, z, \bar{x}(t, z), \bar{u}(t, z)) dt dz - c^0 \bar{y}^0 \int_{\Omega} l(\bar{x}(T, z)) dz \\ & - c^0 \bar{y}^0 \int_{[0, T] \times \partial\Omega} h(t, z, \bar{v}(t, z)) dt dz. \end{aligned}$$

Combining (2.14) with (2.15) gives

$$\begin{aligned} & -c^0 \bar{y}^0 \int_{[0, T] \times \Omega} L(t, z, \bar{x}(t, z), \bar{u}(t, z)) dt dz - c^0 \bar{y}^0 \int_{\Omega} l(\bar{x}(T, z)) dz \\ & - c^0 \bar{y}^0 \int_{[0, T] \times \partial\Omega}^0 h(t, z, \bar{v}(t, z)) dt dz \\ (2.16) \quad & \leq -c^0 \bar{y}^0 \int_{[0, T] \times \Omega} L(t, z, x(t, z), u(t, z)) dt dz \\ & - c^0 \bar{y}^0 \int_{\Omega} l(x(T, z)) dz - c^0 \bar{y}^0 \int_{[0, T] \times \partial\Omega}^0 h(t, z, v(t, z)) dt dz, \end{aligned}$$

which completes the proof.  $\square$

*Remark 2.2.* The notion of relative minimum derives from Weierstrass. His strong relative minimum in the classical calculus of variations is just minimum relative to all admissible trajectories lying in a set covered by trajectories of his “field of extremals.” In our case such a set is defined by  $V_y$ , i.e., this set equals  $X = \{(t, z, x) : x = -V_y(t, z, \bar{c}p), (t, \bar{c}p) \in P\}$ . It is obvious that in general  $X$  is not covered by all admissible trajectories. It depends on the set  $P$ , chosen at the beginning, where the equation DSPDEMDP (1.13) is considered and the solution  $V$  of (1.13) (which generally is not unique). In practice (see the examples below), we choose for  $P = [0, T] \times \bar{\Omega} \times \{(y^0, y) \in R^2 : y^0 \leq 0, y > 0\}$  or even  $P = [0, T] \times \bar{\Omega} \times \{(y^0, y) \in R^2 : y^0 \leq 0\}$ , and we assume  $V_y$  to be of  $C^1$  so then  $X$  is not a thin set except for a degenerate case  $V_y \equiv 0$  (by transversality condition (1.2), then  $V = c^0 y^0 V_{y^0}$ —a case which does not occur in mechanics).

*Remark 2.3.* The requirement in the theorem that  $V(t, z, p)$  is a  $C^2$  solution of DSPDEMDP (1.13) on  $P$  such that (1.2), (1.3), and (2.1) hold looks very complicated and difficult to satisfy. However, if we rewrite it in a better form, it is not much different from known PDE systems. Thus let us assume that  $l \equiv 0$ , put  $w = V_y$ , and rewrite (1.2) to the form  $V_{y^0} = \frac{1}{c^0 y^0} V - \frac{cy}{c^0 y^0} w$  (for  $y^0 < 0$ ). Then we get that  $V$  must satisfy in  $P$  the following system of equations:

$$\begin{aligned} (2.17) \quad & V_{y^0} = \frac{1}{c^0 y^0} V - \frac{cy}{c^0 y^0} w, \\ & V_y = w, \end{aligned}$$

$$V_{tt} - \Delta_z V + H(t, z, -w, \bar{c}p) = 0$$

with initial (end) condition

$$V_{y^0}(T, z, \bar{c}p) = 0$$

and Neumann boundary condition

$$\nabla_z V(t, z, \bar{c}p)\nu(z) + H_\Sigma(t, z, \bar{c}p) = 0,$$

$$(2.18) \quad (t, z) \in (0, T) \times \partial\Omega, (t, z, \bar{c}p) \in P,$$

where  $H$  and  $H_\Sigma$  are defined in the former section. If it happens that both  $H$  and  $H_\Sigma$  are smooth enough functions and  $n \geq 4$  (dimension of  $\Omega$ ), then the existence of continuous and then smooth solutions for (2.17)–(2.18) can be obtained by a standard fixed point method (compare [24] and the smooth case for  $\Omega = R^n$  [30]; see also [21]).

**3. An optimal dual feedback control.** It often occurs that for engineering and practical applications a feedback control is more important than a value function. It turns out that the dual dynamic programming approach allows us also to investigate a kind of feedback control which we call a dual feedback control. Surprisingly it can have better properties than the classical one: now our state equation depends only on parameters and not additionally on state constraints in the feedback function, which made the state equation difficult to solve.

**DEFINITION 3.1.** A pair of functions  $u = \tilde{u}(t, z, p)$  from  $P$  of the points  $(t, z, p) = (t, z, y^0, y)$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $y^0 \leq 0$ ,  $y \in R$ , into  $U$  and  $\tilde{v}(t, z, p)$  from a subset  $P$  of those points  $(t, z, p) = (t, z, y^0, y)$ ,  $(t, z) \in (0, T) \times \partial\Omega$ ,  $(t, z, p) \in P$ , into  $\mathbf{V}$  is called a dual feedback control if there is any solution  $\tilde{x}(t, z, p)$ ,  $P$ , of the PDE

$$(3.1) \quad \tilde{x}_{tt}(t, z, p) - \Delta_z \tilde{x}(t, z, p) = f(t, z, \tilde{x}(t, z, p), \tilde{u}(t, z, p))$$

satisfying boundary condition

$$\tilde{x}(t, z, p) = \tilde{v}(t, z, p) \quad \text{on } (0, T) \times \Gamma, (t, z, p) \in P$$

such that for each admissible trajectory  $x(t, z)$ ,  $(t, z) \in [0, T] \times \Omega$ , there exists such a function  $p(t, z) = (y^0, y(t, z))$ ,  $p \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ ,  $p \in L^2([0, T] \times \partial\Omega)$ ,  $(t, z, p(t, z)) \in P$ , such that (1.1) holds.

**DEFINITION 3.2.** A dual feedback control  $(\bar{u}(t, z, p), \bar{v}(t, z, p))$  is called an optimal dual feedback control if there exist a function  $\bar{x}(t, z, p)$ ,  $(t, z, p) \in P$ , corresponding to  $\bar{u}(t, z, p)$ ,  $\bar{v}(t, z, p)$  as in Definition 3.1, and a function  $\bar{p}(t, z) = (\bar{y}^0, \bar{y}(t, z))$ ,  $\bar{p} \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ ,  $\bar{p} \in L^2([0, T] \times \partial\Omega)$ ,  $(t, z, \bar{p}(t, z)) \in P$ ,  $(t, z, \bar{c}\bar{p}(t, z)) \in P$  with  $\bar{c} = (c^0, c)$ , such that dual value function  $S_D$  (see (1.4)) is defined at  $(t, \bar{p}(t, \cdot))$  by  $\bar{u}(\tau, z, p)$ ,  $\bar{v}(\tau, z, p)$  and corresponding to them  $\bar{x}(\tau, z, p)$ ,  $(\tau, z, p) \in P$ ,  $\tau \in [t, T]$ , i.e.,

$$(3.2) \quad \begin{aligned} S_D(t, \bar{p}(t, \cdot)) &= -c^0 \bar{y}^0 \int_{[t, T] \times \Omega} L(\tau, z, \bar{x}(\tau, z, \bar{p}(\tau, z)), \bar{u}(\tau, z, \bar{p}(\tau, z))) d\tau dz \\ &- c^0 \bar{y}^0 \int_{\Omega} l(\bar{x}(T, z, \bar{p}(T, z))) dz - c^0 y^0 \int_{[t, T] \times \partial\Omega} h(\tau, z, \bar{v}(\tau, z, \bar{p}(\tau, z))) d\tau dz. \end{aligned}$$

Moreover there is  $V(t, z, p)$  satisfying (1.2) and (1.3) for which  $V_{y^0}$  satisfies the equality

$$\int_{\Omega} c^0 y^0 V_{y^0}(t, z, \bar{c}\bar{p}(t, z)) dz + c^0 \bar{y}^0 \int_{\partial\Omega} (\nabla_z) V_{y^0}(t, z, \bar{c}\bar{p}(t, z)) \nu(z) dz = -S_D(t, \bar{p}(t, \cdot))$$

and  $V_y$  satisfies

$$(3.3) \quad V_y(t, z, \bar{c}p) = -\bar{x}(t, z, p) \quad \text{for } (t, z) \in (0, T) \times \bar{\Omega}, \quad (t, z, p) \in P, \quad (t, z, \bar{c}p) \in P.$$

The next theorem is nothing more than the above verification theorem formulated in terms of a dual feedback control.

**THEOREM 3.3.** *Let  $(\bar{u}(t, z, p), \bar{v}(t, z, p))$  be a dual feedback control in  $P$ . Suppose that there exist  $\bar{c} = (c^0, c) \in R^2$  and a  $C^2$  solution  $V(t, z, p)$  of (1.13) on  $P$  such that (1.2) and (2.1) hold. Let  $\bar{p}(t, z) = (\bar{y}^0, \bar{y}(t, z))$ ,  $\bar{p} \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ ,  $\bar{p} \in L^2([0, T] \times \partial\Omega)$ ,  $(t, z, \bar{p}(t, z)) \in P$ ,  $(t, z, \bar{c}\bar{p}(t, z)) \in P$ , be such a function that  $\bar{x}(t, z) = \bar{x}(t, z, \bar{p}(t, z))$ ,  $\bar{u}(t, z) = \bar{u}(t, z, \bar{p}(t, z))$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $\bar{v}(t, z) = \bar{v}(t, z, \bar{p}(t, z))$ ,  $(t, z) \in (0, T) \times \partial\Omega$ , is an admissible trio, where  $\bar{x}(t, z, p)$ ,  $(t, z, p) \in P$ , corresponds to  $\bar{u}(t, z, p)$  and  $\bar{v}(t, z, p)$  as in Definition 3.1. Assume further that  $V_y$  and  $V_{y^0}$  satisfy*

$$(3.4) \quad V_y(t, z, \bar{c}p) = -\bar{x}(t, z, p) \quad \text{for } (t, z) \in [0, T] \times \Omega, \quad (t, z, p) \in P, \quad (t, z, \bar{c}p) \in P,$$

$$(3.5) \quad \begin{aligned} & c^0 \bar{y}^0 \int_{\Omega} V_{y^0}(t, z, \bar{c}\bar{p}(t, z)) dz \\ & + c^0 \bar{y}^0 \int_{[0, T]} \left( \int_{\partial\Omega} (\nabla_z) V_{y^0}(t, z, c^0 \bar{y}^0, c \bar{y}(t, z)) \nu(z) dz \right) dt \\ & = -c^0 \bar{y}^0 \int_{[0, T] \times \Omega} L(t, z, \bar{x}(t, z, \bar{p}(t, z)), \bar{u}(t, z, \bar{p}(t, z))) dt dz \\ & - c^0 \bar{y}^0 \int_{\Omega} l(\bar{x}(T, z, \bar{p}(T, z))) dz - c^0 \bar{y}^0 \int_{[t, T] \times \partial\Omega} h(\tau, s, \bar{v}(\tau, z, \bar{p}(\tau, z))) d\tau ds. \end{aligned}$$

Then  $(\bar{u}(t, z, p), \bar{v}(t, z, p))$  is an optimal dual feedback control.

*Proof.* Take any function

$$p(t, z) = (\bar{y}^0, y(t, z)), \quad p \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega)),$$

$$p \in L^2([0, T] \times \partial\Omega), \quad (t, z, p(t, z)) \in P, \quad (t, z, \bar{c}p(t, z)) \in P,$$

such that  $x(t, z) = \bar{x}(t, z, p(t, z))$ ,  $u(t, z) = \bar{u}(t, z, p(t, z))$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $v(t, z) = \bar{v}(t, z, p(t, z))$ ,  $(t, z) \in [0, T] \times \partial\Omega$ , is an admissible trio and (2.3) holds. By (3.4), it follows that  $x(t, z) = -V_y(t, z, \bar{c}p(t, z))$  for  $(t, z) \in (0, T) \times \Omega$ . As in the proof of Theorem 2.1, (3.5) gives

$$(3.6) \quad \begin{aligned} & -c^0 \bar{y}^0 \int_{[0, T] \times \Omega} L(t, z, \bar{x}(t, z, p(t, z)), \bar{u}(t, z, p(t, z))) dt dz \\ & - c^0 \bar{y}^0 \int_{\Omega} l(\bar{x}(T, z, p(T, z))) dz - c^0 \bar{y}^0 \int_{[t, T] \times \partial\Omega} h(\tau, s, \bar{v}(\tau, z, p(\tau, z))) d\tau ds \\ & \leq -c^0 \bar{y}^0 \int_{[0, T] \times \Omega} L(t, z, \bar{x}(t, z, p(t, z)), \bar{u}(t, z, p(t, z))) dt dz \\ & - c^0 \bar{y}^0 \int_{\Omega} l(\bar{x}(T, z, p(T, z))) dz - c^0 \bar{y}^0 \int_{[t, T] \times \partial\Omega} h(\tau, s, \bar{v}(\tau, z, p(\tau, z))) d\tau ds. \end{aligned}$$

We conclude from (3.6) that

$$(3.7) \quad \begin{aligned} S_D(t, \bar{p}(t, \cdot)) &= -c^0 \bar{y}^0 \int_{[t, T] \times \Omega} L(\tau, z, \bar{x}(\tau, z, \bar{p}(\tau, z)), \bar{u}(\tau, z, \bar{p}(\tau, z))) d\tau dz \\ &\quad - c^0 \bar{y}^0 \int_{\Omega} l(\bar{x}(T, z, \bar{p}(T, z))) dz - c^0 y^0 \int_{[t, T] \times \partial\Omega} h(\tau, s, \bar{v}(\tau, z, \bar{p}(\tau, z))) d\tau ds, \end{aligned}$$

and it is sufficient to show that  $(\bar{u}(t, z, p), \bar{v}(t, z, p))$  is an optimal dual feedback control, by Theorem 2.1 and Definition 3.2.  $\square$

#### 4. Examples.

*Example 1.* Let us denote the following:

$$\begin{aligned} L(t, z, x, u) &:= t^2 x^{-7/6} u^{1/2}, \\ f(t, z, x, u) &:= t^6 x^{1/2} u^{3/2} / n, \\ h(t, z, v) &:= t^4 \left( \sum_{i=1}^{n/2} z_i^2 \right) \left( v - \left( \sum_{i=1}^n z_i \right)^{-2/3} \right)^6, \\ l(x) &= 0, \quad \varphi(0, z) = \left( \sum_{i=1}^n z_i \right)^{-2/3}, \quad \psi(0, z) = 0, \end{aligned}$$

where  $(t, z, x, u) \in [0, T] \times \Omega \times R \times R^+$ ,  $n$ -even,  $\Omega := \{z \in R^n : \sum_{i=1}^n z_i^2 < 1\}$ ,  $Y := \{(y^0, y) \in R^2 : y^0 \leq 0, y > 0\}$ ,  $\bar{c} := (c^0, c) \in R^2$ ,  $\bar{c}p \in Y$ , so  $c^0 > 0$ ,  $c > 0$ ,  $U = R^+$ ,  $\mathbf{V} = [-10, 10]$ .

The Hamiltonian  $H : [0, T] \times \Omega \times R^+ \times Y \rightarrow R$

$$(4.1) \quad H(t, z, x, \bar{c}p) := \max_{u \in R^+} \mathcal{H}(t, z, x, \bar{c}p, u),$$

where  $\mathcal{H} : [0, T] \times \Omega \times R^+ \times Y \times R^+ \rightarrow R$

$$(4.2) \quad \mathcal{H}(t, z, x, \bar{c}p, u) := c^0 y^0 t^2 x^{-7/6} u^{1/2} + c y t^6 x^{1/2} u^{3/2} / n.$$

From (4.1)–(4.2) and taking  $c^0 = 2/9$ ,  $c = 4/9$  we calculate easily that

$$(4.3) \quad H(t, z, x, \bar{c}p) = -2\sqrt{6n}(-y^0)^{3/2} / 27y^{1/2}x^2.$$

Hence the PDE for the same  $c^0 = 2/9$ ,  $c = 4/9$  has the form

$$(4.4) \quad V_{tt}(t, z, \bar{c}p) - \Delta_z V(t, z, \bar{c}p) - 2\sqrt{6n}(-y^0)^{3/2} / \left( 27y^{1/2} (-V_y(t, z, \bar{c}p))^2 \right) = 0.$$

For  $c^0 = 2/9$ ,  $c = 4/9$ , let

$$(4.5) \quad V(t, z, \bar{c}p) := -4/3 y^{3/4} - \sqrt{n} \left( -\frac{2}{3} y^0 \right)^{3/2} \sum_{i=1}^n z_i^2 / 6n.$$

Then the function  $V(t, z, \bar{c}p)$  satisfies on  $P = [0, T] \times \Omega \times Y$  the first equation of DPDEMDP (1.13) and (1.2). Let us take for  $\nu(z)$  the exterior unit normal vector to  $\partial\Omega$  the vector  $(-z_2, z_1, -z_4, z_3, \dots, -z_n, z_{n-1})$ . Then (1.3) is also satisfied as both sides

are simply equal to zero for  $(t, z) \in [0, T] \times \partial\Omega$ ,  $(t, z, \bar{c}p) \in P$ . Moreover for similar reasons the second equation in (1.13) is also satisfied ( $\sup\{c^0 y^0 h(t, z, v) : v \in \mathbf{V}\} = 0$ ). If we take  $y^0 = -(32n/3)^{1/3}$ , then the equation

$$(4.6) \quad x_{tt}(t, z, y) - \Delta_z x(t, z, y) = \sqrt{n} (-y^0)^{3/2} / \left( (6y)^{3/2} (x(t, z, y))^2 \right)$$

has the solution

$$\bar{x}(t, z, y) = y^{-1/2} \left( \sum_{i=1}^n z_i \right)^{2/3}$$

on  $[0, T] \times \Omega$ . We denote that  $y^0$  by  $\bar{y}^0$ . Moreover, (2.2) is fulfilled by  $\bar{p}(t, z) = (\bar{y}^0, \bar{y}(t, z))$ , where  $\bar{y}^0$  comes from (4.6),

$$\bar{y}(t, z) = \left( \sum_{i=1}^n z_i \right)^{8/3},$$

and  $V(t, z, \bar{c}p)$  is given by (4.5). It is also seen that

$$\bar{x}(t, z) = \bar{x}(t, z, \bar{y}(t, z)) = \left( \sum_{i=1}^n z_i \right)^{-2/3}$$

satisfies  $\bar{x}(t, z) = -V_y(t, z, \bar{c}p(t, z))$  for  $(t, z) \in [0, T] \times \Omega$  and

$$\bar{x}(t, z) = \bar{v}(t, z) = \left( \sum_{i=1}^n z_i \right)^{-2/3} \quad \text{for } (t, z) \in [0, T] \times \partial\Omega,$$

and

$$\bar{u}(t, z) = -\bar{y}^0 / \left( 6t^4 \left( \sum_{i=1}^n z_i \right)^{14/9} \right).$$

Moreover, assumptions of Theorem 3.3 hold. Therefore, by Theorem 3.3 we obtain that  $\bar{x}(t, z)$ ,  $\bar{u}(t, z, p)$ ,  $\bar{v}(t, z)$  is an optimal trio.

The next example is linear but with control on the boundary so it cannot be treated by any classical dynamic programming approach. Moreover it is an applied example.

*Example 2.* We shall consider the simple case of a structural acoustic system. Let  $\Omega = \{(z_1, z_2) : -1 \leq z_1 \leq 1, -1 \leq z_2 \leq 1\}$  be the domain occupied by an acoustic medium (air). The boundary  $S$  of the domain consists of two parts,  $S_1$  and  $S_2$ . The part  $S_1$  corresponds to a thin wall (a shell), and  $S_2$  corresponds to a hard wall. An external acoustic eld, through structural acoustic coupling, leads to high interior sound pressure levels in  $\Omega$ . Piezoelectric elements (patches) are used to active control in order to reduce the sound pressure levels in  $\Omega$ . Coupled problems for structural acoustic systems were studied in a series of works; see [3], [4], [5], [6], [18], [19] and the references therein. The acoustic dynamics is described by the equation

$$(4.7) \quad x_{tt}(t, z) - c_0^2 \Delta_z x(t, z) = 0 \quad \text{in } Q = (0, 1) \times \Omega.$$

Here  $c_0^2$  is a positive constant and the pressure function  $q$  in the acoustic medium is defined by  $q(t, z) = \rho_0 x_t(t, z)$ , where  $\rho_0$  is the density of the acoustic medium in

the ground state, and  $\rho_0$  is a positive constant. Let us denote

$$\begin{aligned} L(t, z, x, u) &:= 0, \\ f(t, z, x, u) &:= 0, \\ h(t, z, v) &:= \left( v - \cos\left(t - \frac{\pi}{2}\right) \sin\left(\frac{1}{2c_0}z_1\right) \sin\left(\frac{1}{2c_0}z_2\right) \right)^2, \\ l(x(\cdot)) &= \rho_0 \int_{\Omega} x(1, z) dz, \quad \varphi(0, z) = 0, \end{aligned}$$

where  $(t, z, x, u) \in [0, 1] \times \Omega \times R \times R$ ,  $n = 2$ ,  $Y := \{(y^0, y) \in R^2 : y^0 \leq 0, y \in R\}$ ,  $\bar{c} := (c^0, c) \in R^2$ ,  $\bar{c}p \in Y$ , so  $c^0 > 0$ ,  $c \in R$ ,  $U = R$ ,  $V = R$ . We have chosen the simplest case of  $h$  so as not to concentrate on technicalities which are not related to the described dual method itself.

The Hamiltonian  $H : [0, 1] \times \Omega \times R \times Y \rightarrow R$ ,

$$H(t, z, x, \bar{c}p) = 0.$$

Therefore the first equation of DSPDEMDP (1.13) has the form

$$(4.8) \quad V_{tt}(t, z, \bar{c}p) - c_0^2 \Delta_z V(t, z, \bar{c}p) = 0 \quad \text{in } Q = (0, 1) \times \Omega \times R.$$

The solutions to (4.7), depending on  $\{a_j\}$ , have the form

$$(4.9) \quad \begin{aligned} x(t, z) &= \sum_{j \geq 2} a_j \cos\left(jt - \frac{\pi}{2}\right) \sin\left(\frac{1}{2c_0}jz_1\right) \sin\left(\frac{1}{2c_0}jz_2\right) \\ &\quad + \frac{1}{2} \cos\left(t - \frac{\pi}{2}\right) \sin\left(\frac{1}{2c_0}z_1\right) \sin\left(\frac{1}{2c_0}z_2\right). \end{aligned}$$

We take into account only those solutions (4.9) which belong to  $W^{2,2}((0, 1) \times \Omega) \cap C([0, 1]; L^2(\Omega))$ . By controls on the boundary we take only the functions

$$v = x|_{(0,1) \times \partial\Omega}.$$

Our aim is to minimize the pressure level in  $\Omega$  and the cost of activations of controls on the boundary, i.e., we minimize the functional

$$J(x, v) = \rho_0 \int_{\Omega} x(1, z) dz + \int_{(0,1) \times \partial\Omega} h(t, z, v(t, z)) dt dz.$$

For solution to (4.8) we take

$$(4.10) \quad \begin{aligned} V(t, z, \bar{c}p) &= c^0 y^0 \sin\left(t - \frac{\pi}{2}\right) \sin\left(\frac{1}{2c_0}z_1\right) \sin\left(\frac{1}{2c_0}z_2\right) + c^0 y^0 c y t \\ &\quad + (cy)^2 + tcy - cy \cos\left(t - \frac{\pi}{2}\right) \sin\left(\frac{1}{2c_0}z_1\right) \sin\left(\frac{1}{2c_0}z_2\right). \end{aligned}$$

If we take  $c^0 = 1$ ,  $c = 1/2$ ,  $\bar{y}^0 = -1$ , then we can easily check that  $V$  from (4.10) satisfies transversality conditions (1.2), (1.3) and boundary condition (2.1). Let us observe that

$$-V_y(t, z, \bar{c}p(t, z)) = x(t, z)$$

for

$$y(t, z) = -2 \sum_{j \geq 2} a_j \cos \left( jt - \frac{\pi}{2} \right) \sin \left( \frac{1}{2c_0} j z_1 \right) \sin \left( \frac{1}{2c_0} j z_2 \right), \bar{y}^0 = -1.$$

Let us take for  $\nu(z)$ , the exterior unit normal vector to  $\partial\Omega$ , the vector

$$\nu(z) = \begin{cases} (1, 0) & \text{for } z = (-1, z_2), \quad -1 \leq z_2 \leq 1, \\ (1, 0) & \text{for } z = (1, z_2), \quad -1 \leq z_2 \leq 1, \\ (0, 1) & \text{for } z = (z_1, -1), \quad -1 \leq z_1 \leq 1, \\ (0, 1) & \text{for } z = (z_1, 1), \quad -1 \leq z_1 \leq 1. \end{cases}$$

Then, taking into account (4.10), we see that  $(\nabla_z) V(t, z, \bar{c}p(t, z))\nu(z) = 0$  for  $(t, z) \in (0, 1) \times \partial\Omega$ . Therefore

$$(\nabla_z) V(t, z, \bar{c}p(t, z))\nu(z) + c^0 y^0 h(t, z, \bar{v}(t, z)) = 0 \quad \text{for } (t, z) \in (0, 1) \times \partial\Omega$$

is realized for

$$\bar{v}(t, z) = \cos \left( t - \frac{\pi}{2} \right) \sin \left( \frac{1}{2c_0} z_1 \right) \sin \left( \frac{1}{2c_0} z_2 \right).$$

Hence

$$\bar{x}(t, z) = \cos \left( t - \frac{\pi}{2} \right) \sin \left( \frac{1}{2c_0} z_1 \right) \sin \left( \frac{1}{2c_0} z_2 \right).$$

**5. An  $\varepsilon$ -optimization.** If we want to solve the concrete problem (0.1)–(0.4) for particular data, then usually we are not able to solve it exactly, especially if the problem we consider is nonlinear. Therefore each possibility to approximate our optimal problem (0.1)–(0.4) may turn out very useful. Below we find a certain type of such an approximation.

**DEFINITION 5.1.** Let  $\varepsilon > 0$  and  $c^0 > 0$  be fixed. A function  $S_{\varepsilon D}(t, p(t, \cdot))$  is called an  $\varepsilon$ -dual value function if

$$(5.1) \quad S_D(t, p(t, \cdot)) \leq S_{\varepsilon D}(t, p(t, \cdot)) \leq S_D(t, p(t, \cdot)) - 2\varepsilon c^0 \bar{y}_\varepsilon^0 \text{Vol}(\Omega)$$

for any fixed  $\bar{y}_\varepsilon^0 < 0$ .

**DEFINITION 5.2.** Let  $\varepsilon > 0$  and  $\bar{c} = (c^0, c) \in R^2$ ,  $c^0 > 0$ , be fixed and let  $\tilde{V}(t, z, p)$  be a given  $C^2$  function such that (1.2) and (2.1) hold. Let  $\bar{x}_\varepsilon(t, z)$ ,  $\bar{u}_\varepsilon(t, z)$ ,  $t \in (0, T) \times \Omega$ ,  $\bar{v}_\varepsilon(t, z)$ ,  $t \in (0, T) \times \partial\Omega$ , be an admissible trio and let  $\bar{p}_\varepsilon(t, z) = (\bar{y}_\varepsilon^0, \bar{y}_\varepsilon(t, z))$ ,  $\bar{p}_\varepsilon \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ ,  $\bar{p}_\varepsilon \in L^2([0, T] \times \partial\Omega)$ ,  $(t, z, \bar{c}\bar{p}_\varepsilon(t, z)) \in P$ , be such a function that  $\bar{x}_\varepsilon(t, z) = -\tilde{V}_y(t, z, \bar{c}\bar{p}_\varepsilon(t, z))$  for  $(t, z) \in (0, T) \times \Omega$ . The trio  $\bar{x}_\varepsilon(t, z)$ ,  $\bar{u}_\varepsilon(t, z)$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $\bar{v}_\varepsilon(t, z)$ ,  $t \in (0, T) \times \partial\Omega$ , is called an  $\varepsilon$ -optimal trio if

$$(5.2) \quad \begin{aligned} & -c^0 \bar{y}_\varepsilon^0 \int_{[0, T] \times \Omega} L(t, z, \bar{x}_\varepsilon(t, z), \bar{u}_\varepsilon(t, z)) dt dz - c^0 \bar{y}_\varepsilon^0 \int_{\Omega} l(\bar{x}_\varepsilon(T, z)) dz \\ & - c^0 \bar{y}_\varepsilon^0 \int_{[0, T] \times \partial\Omega} h(t, s, \bar{v}_\varepsilon(t, z)) dt ds \\ & \leq -c^0 \bar{y}_\varepsilon^0 \int_{[0, T] \times \Omega} L(t, z, x(t, z), u(t, z)) dt dz - c^0 \bar{y}_\varepsilon^0 \int_{\Omega} l(x(T, z)) dz \\ & - c^0 \bar{y}_\varepsilon^0 \int_{[0, T] \times \partial\Omega} h(t, s, v(t, z)) dt ds - \varepsilon c^0 \bar{y}_\varepsilon^0 \text{Vol}(\Omega) \end{aligned}$$

for all admissible trios  $x(t, z)$ ,  $u(t, z)$ ,  $t \in (0, T) \times \Omega$ ,  $v(t, z)$ ,  $t \in (0, T) \times \partial\Omega$ , for which there exists such a function  $p(t, z) = (\bar{y}_\varepsilon^0, y(t, z))$ ,  $p \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ ,  $p \in L^2([0, T] \times \partial\Omega)$ ,  $(t, z, \bar{c}p(t, z)) \in P$ , such that

$$x(t, z) = -\tilde{V}_y(t, z, \bar{c}p(t, z)) \quad \text{for } (t, z) \in (0, T) \times \Omega$$

and

$$(5.3) \quad y(0, z) = \bar{y}_\varepsilon(0, z) \quad \text{for } z \in \Omega.$$

**THEOREM 5.3.** *Let  $\bar{x}_\varepsilon(t, z)$ ,  $\bar{u}_\varepsilon(t, z)$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $\bar{v}_\varepsilon(t, z)$ ,  $t \in (0, T) \times \partial\Omega$ , be an admissible trio. Assume that there exist  $\varepsilon > 0$ ,  $\bar{c} = (c^0, c) \in R^2$ ,  $c^0 > 0$ , and a  $C^2$  function  $\tilde{V}(t, z, p)$  such that for  $(t, z, \bar{c}p) \in P$ ,*

$$(5.4) \quad \sup \left\{ \tilde{V}_{tt}(t, z, \bar{c}p) - \Delta_z \tilde{V}(t, z, \bar{c}p) + c^0 y^0 L(t, z, -\tilde{V}_y(t, z, \bar{c}p), u) + cyf(t, z, -\tilde{V}_y(t, z, \bar{c}p), u) : u \in U \right\} \leq -\varepsilon c^0 \bar{y}_\varepsilon^0,$$

$$\sup \left\{ \nabla_z \tilde{V}(t, z, \bar{c}p) \nu(z) + c^0 y^0 h(t, z, v) : v \in \mathbf{V} \right\} \leq -\varepsilon c^0 \bar{y}_\varepsilon^0, \\ (t, z) \in (0, T) \times \partial\Omega, \quad (t, z, \bar{c}p) \in P,$$

$$(5.5) \quad \tilde{V}(t, z, \bar{c}p) = \bar{c}p \tilde{V}_p(t, z, \bar{c}p), \quad (t, z) \in (0, T) \times \Omega, \quad (t, z, \bar{c}p) \in P,$$

$$\nabla_z V(t, z, \bar{c}p) \nu(z) = c^0 y^0 \nabla_z V_{y^0}(t, z, \bar{c}p) \nu(z)$$

$$\text{for } (t, z) \in (0, T) \times \partial\Omega, \quad (t, z, \bar{c}p) \in P.$$

Further, let  $\bar{p}_\varepsilon(t, z) = (\bar{y}_\varepsilon^0, \bar{y}_\varepsilon(t, z))$ ,  $\bar{p}_\varepsilon \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ ,  $\bar{p}_\varepsilon \in L^2([0, T] \times \partial\Omega)$ ,  $(t, z, \bar{c}\bar{p}_\varepsilon(t, z)) \in P$ , be such a function that  $\bar{x}_\varepsilon(t, z) = -\tilde{V}_y(t, z, \bar{c}\bar{p}_\varepsilon(t))$  for  $(t, z) \in (0, T) \times \Omega$ . Suppose that  $\tilde{V}(t, z, p)$  satisfies the boundary condition for  $(T, z, \bar{c}p) \in P$ :

$$(5.6) \quad c^0 \bar{y}_\varepsilon^0 \int_\Omega (d/dt) \tilde{V}_{y^0}(T, z, \bar{c}p) dz = c^0 \bar{y}_\varepsilon^0 \int_\Omega l \left( -\tilde{V}_y(T, z, \bar{c}p) \right) dz.$$

Moreover, suppose that for almost all  $(t, z) \in (0, T) \times \Omega$ ,

$$(5.7) \quad \tilde{V}_{tt}(t, z, \bar{c}\bar{p}_\varepsilon(t, z)) - \Delta_z \tilde{V}(t, z, \bar{c}\bar{p}_\varepsilon(t, z)) + c^0 \bar{y}_\varepsilon^0 L(t, z, -\tilde{V}_y(t, z, \bar{c}\bar{p}_\varepsilon(t, z)), \bar{u}_\varepsilon(t, z)) + \bar{c}\bar{y}_\varepsilon(t, z) f(t, z, -\tilde{V}_y(t, z, \bar{c}\bar{p}_\varepsilon(t, z)), \bar{u}_\varepsilon(t, z)) \geq 0,$$

$(\nabla_z) \tilde{V}(t, z, \bar{c}\bar{p}_\varepsilon(t, z)) \nu(z) + c^0 \bar{y}_\varepsilon^0 h(t, z, \bar{v}_\varepsilon(t, z)) \geq 0$  for almost all  $(t, z) \in (0, T) \times \partial\Omega$ . Then  $\bar{x}_\varepsilon(t, z)$ ,  $\bar{u}_\varepsilon(t, z)$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $\bar{v}_\varepsilon(t, z)$ ,  $t \in (0, T) \times \partial\Omega$ , is an  $\varepsilon$ -optimal trio relative to all admissible trios  $x(t, z)$ ,  $u(t, z)$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $v(t, z)$ ,  $t \in (0, T) \times \partial\Omega$ , for which there exists such a function  $p(t, z) = (\bar{y}_\varepsilon^0, y(t, z))$ ,  $p \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega))$ ,  $p \in L^2([0, T] \times \partial\Omega)$ ,  $(t, z, \bar{c}p(t, z)) \in P$ , such that

$$x(t, z) = -\tilde{V}_y(t, z, \bar{c}p(t, z)) \quad \text{for } (t, z) \in (0, T) \times \Omega$$

and (5.3) is satisfied.



*Proof.* Take any admissible trio  $x(t, z)$ ,  $u(t, z)$ ,  $(t, z) \in (0, T) \times \Omega$ ,  $v(t, z)$ ,  $t \in (0, T) \times \partial\Omega$ , for which there exists such a function

$$p(t, z) = (\bar{y}_\varepsilon^0, y(t, z)), p \in W^{2,2}([0, T] \times \Omega) \cap C([0, T]; L^2(\Omega)),$$

$$p \in L^2([0, T] \times \partial\Omega), (t, z, \bar{c}p(t, z)) \in P,$$

such that

$$x(t, z) = -\tilde{V}_y(t, z, \bar{c}p(t, z)) \quad \text{for } (t, z) \in (0, T) \times \Omega$$

and (5.3) holds. Then, from (5.5), we have, for  $(t, z) \in (0, T) \times \Omega$ ,

$$\begin{aligned} & \tilde{V}_{tt}(t, z, \bar{c}p(t, z)) - \Delta_z \tilde{V}(t, z, \bar{c}p(t, z)) \\ (5.8) \quad &= c^0 \bar{y}_\varepsilon^0 \left[ (d^2/dt^2) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z)) - (\Delta_z) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z)) \right] \\ &+ cy(t, z) \left[ (d^2/dt^2) \tilde{V}_y(t, z, \bar{c}p(t, z)) - (\Delta_z) \tilde{V}_y(t, z, \bar{c}p(t, z)) \right], \end{aligned}$$

and for  $(t, z) \in (0, T) \times \partial\Omega$ ,

$$(5.9) \quad (\nabla_z) \tilde{V}(t, z, \bar{c}p(t, z))\nu(z) = c^0 \bar{y}_\varepsilon^0 (\nabla_z) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z))\nu(z).$$

Since

$$\begin{aligned} & (d^2/dt^2) \tilde{V}_y(t, z, \bar{c}p(t, z)) - (\Delta_z) \tilde{V}_y(t, z, \bar{c}p(t, z)) \\ &= -f(t, z, -\tilde{V}_y(t, z, \bar{c}p(t, z)), u(t, z)) \end{aligned}$$

for  $(t, z) \in (0, T) \times \Omega$ , it follows, by (5.9), that for  $(t, z) \in (0, T) \times \Omega$ ,

$$\begin{aligned} & c^0 \bar{y}_\varepsilon^0 \left[ (d^2/dt^2) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z)) - (\Delta_z) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z)) \right. \\ & \quad \left. + L(t, z, -\tilde{V}_y(t, z, \bar{c}p(t, z)), u(t, z)) \right] \\ (5.10) \quad &= \tilde{V}_{tt}(t, z, \bar{c}p(t, z)) - \Delta_z \tilde{V}(t, z, \bar{c}p(t, z)) \\ &+ c^0 \bar{y}_\varepsilon^0 L(t, z, -\tilde{V}_y(t, z, \bar{c}p(t, z)), u(t, z)) \\ &+ cy(t, z) f(t, z, -\tilde{V}_y(t, z, \bar{c}p(t, z)), u(t, z)) \end{aligned}$$

and for  $(t, z) \in (0, T) \times \partial\Omega$ ,

$$\begin{aligned} & c^0 \bar{y}_\varepsilon^0 (\nabla_z) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z))\nu(z) + c^0 \bar{y}_\varepsilon^0 h(t, z, v(t, z)) \\ (5.11) \quad &= (\nabla_z) \tilde{V}(t, z, \bar{c}p(t, z))\nu(z) + c^0 \bar{y}_\varepsilon^0 h(t, z, v(t, z)). \end{aligned}$$

Thus, by (5.4) and (5.10), we get for  $(t, z) \in (0, T) \times \Omega$

$$\begin{aligned} & c^0 \bar{y}_\varepsilon^0 \left[ (d^2/dt^2) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z)) - (\Delta_z) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z)) \right. \\ (5.12) \quad & \quad \left. + L(t, z, -\tilde{V}_y(t, z, \bar{c}p(t, z)), u(t, z)) \right] \leq -\varepsilon c^0 \bar{y}_\varepsilon^0 \end{aligned}$$

and for  $(t, z) \in (0, T) \times \partial\Omega$

$$(5.13) \quad c^0 \bar{y}_\varepsilon^0 (\nabla_z) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z))\nu(z) + c^0 \bar{y}_\varepsilon^0 h(t, z, v(t, z)) \leq -\varepsilon c^0 \bar{y}_\varepsilon^0,$$

and finally, after integrating (5.12), that

$$\begin{aligned} & c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \Omega} \left[ (d^2/dt^2) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z)) - (\operatorname{div} \nabla_z) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z)) \right] dt dz \\ & \leq -c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \Omega} L(t, z, x(t, z), u(t, z)) dt dz - \varepsilon c^0 \bar{y}_\varepsilon^0 \operatorname{Vol}(\Omega). \end{aligned}$$

Similarly, in the set  $(0, T) \times \partial\Omega$  we have

$$\begin{aligned} & c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \partial\Omega} (\nabla_z) \tilde{V}_{y^0}(t, z, \bar{c}p(t, z)) \nu(z) dt dz \\ & \leq -c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \partial\Omega} h(t, z, v(t, z)) dt dz - \varepsilon c^0 \bar{y}_\varepsilon^0 \operatorname{Vol}(\Omega). \end{aligned}$$

Next we obtain

$$\begin{aligned} & -c^0 \bar{y}_\varepsilon^0 \int_{\Omega} (d/dt) \tilde{V}_{y^0}(0, z, c^0 \bar{y}_\varepsilon^0, c\bar{y}(0, z)) dz \\ (5.14) \quad & \leq -c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \Omega} L(t, z, x(t, z), u(t, z)) dt dz - c^0 \bar{y}_\varepsilon^0 \int_{\Omega} l(x(T, z)) dz \\ & - c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \partial\Omega} h(t, z, v(t, z)) dt dz - 2\varepsilon c^0 \bar{y}_\varepsilon^0 \operatorname{Vol}(\Omega). \end{aligned}$$

Similarly, we obtain

$$(5.15) \quad c^0 \bar{y}_\varepsilon^0 [(d^2/dt^2) \tilde{V}_{y^0}(t, z, \bar{c}\bar{p}_\varepsilon(t, z)) - (\Delta_z) \tilde{V}_{y^0}(t, z, \bar{c}\bar{p}_\varepsilon(t, z)) + L(t, z, -\tilde{V}_y(t, z, \bar{c}\bar{p}_\varepsilon(t, z)), \bar{u}_\varepsilon(t, z))] \geq 0 \text{ for } (t, z) \in (0, T) \times \Omega,$$

$$(5.16) \quad c^0 \bar{y}_\varepsilon^0 (\nabla_z) \tilde{V}_{y^0}(t, z, \bar{c}\bar{p}_\varepsilon(t, z)) \nu(z) + c^0 \bar{y}_\varepsilon^0 h(t, z, \bar{v}_\varepsilon(t, z)) \geq 0 \text{ for } (t, z) \in (0, T) \times \partial\Omega.$$

Now from (5.15), (5.16), (5.6), and the Green formula we have

$$\begin{aligned} & -c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \Omega} L(t, z, \bar{x}_\varepsilon(t, z), \bar{u}_\varepsilon(t, z)) dt dz \\ (5.17) \quad & - c^0 \bar{y}_\varepsilon^0 \int_{\Omega} l(\bar{x}_\varepsilon(T, z)) dz - c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \partial\Omega} h(t, z, \bar{v}_\varepsilon(t, z)) dt dz \\ & \leq -c^0 \bar{y}_\varepsilon^0 \int_{\Omega} (d/dt) \tilde{V}_{y^0}(0, z, c^0 \bar{y}_\varepsilon^0, c\bar{y}(0, z)) dz. \end{aligned}$$

Therefore, combining (5.14) with (5.17) yields

$$\begin{aligned} & -c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \Omega} L(t, z, \bar{x}_\varepsilon(t, z), \bar{u}_\varepsilon(t, z)) dt dz - c^0 \bar{y}_\varepsilon^0 \int_{\Omega} l(\bar{x}_\varepsilon(T, z)) dz \\ & - c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \partial\Omega} h(t, z, \bar{v}_\varepsilon(t, z)) dt dz \\ (5.18) \quad & \leq -c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \Omega} L(t, z, x(t, z), u(t, z)) dt dz - c^0 \bar{y}_\varepsilon^0 \int_{\Omega} l(x(T, z)) dz \\ & - c^0 \bar{y}_\varepsilon^0 \int_{[0,T] \times \partial\Omega} h(t, z, v(t, z)) dt dz - 2\varepsilon c^0 \bar{y}_\varepsilon^0 \operatorname{Vol}(\Omega), \end{aligned}$$

which proves the assertion of the theorem.  $\square$

## REFERENCES

- [1] N. ARADA AND J. P. RAYMOND, *Optimality conditions for state-constrained Dirichlet boundary control problems*, J. Optim. Theory Appl., 102 (1999), pp. 51–68.
- [2] N. ARADA AND J. P. RAYMOND, *Dirichlet boundary control of semilinear parabolic equations, Part 2: Problems with pointwise state constraints*, Appl. Math. Optim., 45 (2002), pp. 145–167.
- [3] G. AVALOS, *The exponential stability of a coupled hyperbolic parabolic system arising in structural acoustic*, Abstr. Appl. Anal., 1 (1996), pp. 203–219.
- [4] G. AVALOS AND I. LASIECKA, *Differential Riccati equation for the active control of a problem in structural acoustics*, J. Optim. Theory Appl., 91 (1996), pp. 695–728.
- [5] H. T. BANKS, R. C. SMITH, AND Y. WANG, *Smart Material Structures: Modeling, Estimation and Control*, Wiley, Chichester, UK, 1996.
- [6] M. CAMURDAN, *Uniform stability of a coupled structural acoustic system by boundary dissipation*, Abstr. Appl. Anal., 3 (1998), pp. 377–400.
- [7] V. BARBU, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, Boston, 1993.
- [8] V. BARBU, *The dynamic programming equation for the time-optimal control problem in infinite dimensions*, SIAM J. Control Optim., 29 (1991), pp. 445–456.
- [9] V. BARBU AND G. DA PRATO, *Hamilton–Jacobi Equations in Hilbert Spaces*, Pitman Advanced Publishing Program, Boston, 1983.
- [10] P. CANNARSA AND O. CÂRJĂ, *On the Bellman equation for the minimum time problem in infinite dimensions*, SIAM J. Control Optim., 43 (2004), pp. 532–548.
- [11] E. CASAS, *Pontryagin’s principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.
- [12] H. O. FATTORINI, *Infinite-Dimensional Optimization and Control Theory*, Cambridge University Press, Cambridge, UK, 1999.
- [13] H. O. FATTORINI AND T. MURPHY, *Optimal control for nonlinear parabolic boundary control systems: The Dirichlet boundary conditions*, Differential Integral Equations, 7 (1994), pp. 1367–1388.
- [14] A. V. FURSIKOV, *Optimal Control of Distributed Systems. Theory and Applications*, AMS, Providence, RI, 2000.
- [15] F. GOZZI AND M. E. TESSITORE, *Optimality conditions for Dirichlet boundary control problems of parabolic type*, J. Math. Systems Estim. Control, 8 (1) (1998).
- [16] E. GALEWSKA AND A. NOWAKOWSKI, *Multidimensional dual dynamic programming*, J. Optim. Theory Appl., 124 (2005), pp. 175–186.
- [17] E. GALEWSKA AND A. NOWAKOWSKI, *A dual dynamic programming for multidimensional elliptic optimal control problems*, Numer. Funct. Anal. Optim., 27 (2006), pp. 279–289.
- [18] I. LASIECKA, *Optimization problems for structural acoustic models with thermo-elasticity and smart materials*, Discuss. Math. Differ. Incl. Control Optim., 20 (2000), pp. 113–140.
- [19] I. LASIECKA, *Mathematical control theory in structural acoustic problems*, Math. Models Methods Appl. Sci., 8 (1998), pp. 1119–1153.
- [20] I. LASIECKA, J. L. LIONS, AND R. TRIGGIANI, *Nonhomogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., 65 (1986), pp. 149–192.
- [21] I. LASIECKA AND R. TRIGGIANI, *Regularity theory of hyperbolic equations with nonhomogeneous Neumann boundary conditions. II. General boundary data*, J. Differential Equations, 94 (1991), pp. 112–164.
- [22] X. LI AND J. YONG, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, 1994.
- [23] B. S. MORDUKHOVICH AND J. P. RAYMOND, *Dirichlet boundary control of hyperbolic equations in the presence of state constraints*, Appl. Math. Optim., 49 (2004), pp. 145–157.
- [24] B. S. MORDUKHOVICH AND J.-P. RAYMOND, *Neumann boundary control of hyperbolic equations with pointwise state constraints*, SIAM J. Control Optim., 43 (2004), pp. 1354–1372.
- [25] B. S. MORDUKHOVICH AND K. ZHANG, *Minimax control of parabolic systems with Dirichlet boundary conditions and state constraints*, Appl. Math. Optim., 36 (1997), pp. 323–360.
- [26] B. S. MORDUKHOVICH AND K. ZHANG, *Dirichlet boundary control of parabolic systems with pointwise state constraints*, in Proceedings of the International Conference on Control and Estimations of Distributed Parameter Systems (Vorau, 1996), Internat. Ser. Numer. Math. 126, Birkhäuser, Basel, 1998, pp. 223–236.
- [27] P. NEITTAANMAKI AND D. TIBA, *Optimal Control of Nonlinear Parabolic Systems*, Marcel Dekker, New York, 1994.
- [28] A. NOWAKOWSKI, *The dual dynamic programming*, Proc. Amer. Math. Soc., 116 (1992), pp. 1089–1096.

- [29] J. P. RAYMOND, *Nonlinear boundary control of semilinear parabolic problems with pointwise state constraints*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 341–370.
- [30] W. S. STRAUSS, *Nonlinear Wave Equations*, CBMS Regional Conf. Ser. in Math. 73, AMS, Providence, RI, 1989.
- [31] L. W. WHITE, *Control of a hyperbolic problem with pointwise stress constraints*, J. Optim. Theory Appl., 41 (1983), pp. 359–369.
- [32] L. W. WHITE, *Distributed control of a hyperbolic problem with control and stress constraints*, J. Math. Anal. Appl., 106 (1985), pp. 41–53.
- [33] L. C. YOUNG, *Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

## LIMIT TIME OPTIMAL SYNTHESIS FOR A CONTROL-AFFINE SYSTEM ON $S^{2*}$

P. MASON<sup>†</sup>, R. SALMONI<sup>‡</sup>, U. BOSCAIN<sup>§</sup>, AND Y. CHITOUR<sup>‡</sup>

**Abstract.** For  $\alpha \in ]0, \pi/2[$ , let  $(\Sigma)_\alpha$  be the control system  $\dot{x} = (F + uG)x$ , where  $x$  belongs to the two-dimensional unit sphere  $S^2$ ,  $u \in [-1, 1]$ , and  $F, G$  are  $3 \times 3$  skew-symmetric matrices generating rotations with perpendicular axes and of respective norms  $\cos(\alpha)$  and  $\sin(\alpha)$ . In this paper, we study the time optimal synthesis (TOS) from the north pole  $(0, 0, 1)^T$  associated to  $(\Sigma)_\alpha$ , as the parameter  $\alpha$  tends to zero; this problem is motivated by specific issues in the control of quantum systems. We first prove that the TOS is characterized by a “two-snakes” configuration on the whole  $S^2$ , except for a neighborhood  $U_\alpha$  of the south pole  $(0, 0, -1)^T$  of diameter at most  $\mathcal{O}(\alpha)$ . We next show that, inside  $U_\alpha$ , the TOS depends on the relationship between  $r(\alpha) := \pi/2\alpha - [\pi/2\alpha]$  and  $\alpha$ . More precisely, we characterize three main relationships by considering sequences  $(\alpha_k)_{k \geq 0}$  satisfying (a)  $r(\alpha_k) = \bar{r}$ , (b)  $r(\alpha_k) = C\alpha_k$ , and (c)  $r(\alpha_k) = 0$ , where  $\bar{r} \in (0, 1)$  and  $C > 0$ . In each case, we describe the TOS and provide, after a suitable rescaling, the limiting behavior, as  $\alpha$  tends to zero, of the corresponding TOS inside  $U_\alpha$ .

**Key words.** control-affine systems, optimal synthesis, control of quantum systems, minimum time, asymptotics

**AMS subject classification.** 49J15

**DOI.** 10.1137/060675988

**1. Introduction.** Let  $\alpha \in ]0, \pi/2[$ . On the unit sphere  $S^2 \subset \mathbb{R}^3$ , consider the control system  $(\Sigma)_\alpha$  defined by

$$(1) \quad (\Sigma)_\alpha \quad \dot{x} = (F + uG)x, \quad x = (x_1, x_2, x_3)^T, \quad \|x\|^2 = 1, \quad |u| \leq 1,$$

where  $F$  and  $G$  are two  $3 \times 3$  skew-symmetric matrices representing two orthogonal rotations with axes of length, respectively,  $\cos(\alpha)$  and  $\sin(\alpha)$ ,  $\alpha \in ]0, \pi/2[$  (for the precise meaning of length, see section 2.3). With no loss of generality, we assume that

$$(2) \quad F := \begin{pmatrix} 0 & -\cos(\alpha) & 0 \\ \cos(\alpha) & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad G := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\sin(\alpha) \\ 0 & \sin(\alpha) & 0 \end{pmatrix}.$$

In this paper, we aim at describing the time optimal synthesis (TOS) from the north pole  $N := (0, 0, 1)^T$  for  $(\Sigma)_\alpha$ ; i.e., for every  $\bar{x} \in S^2$  we want to find the time optimal trajectory steering  $N$  to  $\bar{x}$  in minimum time (see Figure 1).

In particular *we are interested in the qualitative shape of the time optimal synthesis in a neighborhood of the south pole  $S = (0, 0, -1)^T$ , in the limit  $\alpha \rightarrow 0$* . The interest for that problem stems from quantum control issues. Indeed consider the population transfer problem for a two-level quantum system driven by a single external field. This model describes the evolution of the  $z$ -component of the spin of a

\*Received by the editors November 24, 2006; accepted for publication (in revised form) September 14, 2007; published electronically January 4, 2008.

<http://www.siam.org/journals/sicon/47-1/67598.html>

<sup>†</sup>Institut Elie Cartan UMR 7502, Nancy-Université/CNRS/INRIA, POB 239, 54506 Vandoeuvre-lès-Nancy, France (Paolo.Mason@iecn.u-nancy.fr). This author was (partially) supported by IDF—Aide au partage des projets européens.

<sup>‡</sup>Laboratoire des signaux et systèmes, Université Paris-Sud, CNRS, Supélec, 91192 Gif-Sur-Yvette, France (rebecca.salmoni@lss.supelec.fr, yacine.chitour@lss.supelec.fr).

<sup>§</sup>SISSA, via Beirut 2-4, 34014 Trieste, Italy (boscaín@sisssa.it), and Le2i, CNRS UMR 5158, Université de Bourgogne, 9, avenue Alain Savary, BP 47870, 21078 Dijon, France.

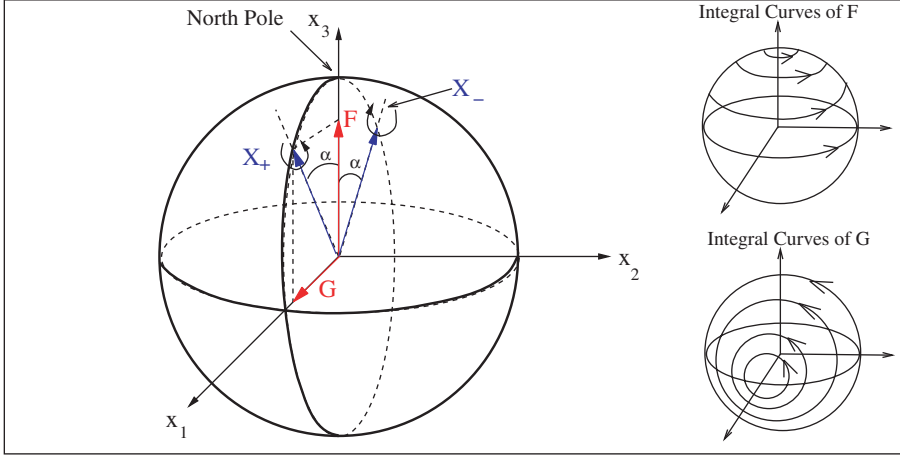


FIG. 1. Geometric interpretation of the system  $(\Sigma)_\alpha$ . The vector fields  $X_+ := F + G$  and  $X_- := F - G$  are two rotations of norm one making an angle  $\alpha$  with the axis  $x_3$ .

(spin-1/2) particle driven by a magnetic field that is constant along the  $z$ -axis and controlled along the  $x$ -axis. Equivalently it describes the first two levels of a molecule driven by an external field without the rotating wave approximation [4, 9]. The dynamics of such a system is governed by the time-dependent Schrödinger equation (in a system of units such that  $\hbar = 1$ ):

$$(3) \quad i \frac{d\psi(t)}{dt} = (H_0 + \Omega(t)H_1)\psi(t).$$

Here  $\psi(\cdot) = (\psi_1(\cdot), \psi_2(\cdot))^T : [0, T] \rightarrow \mathbb{C}^2$  denotes the wave function and verifies  $\sum_{j=1}^2 |\psi_j(t)|^2 = 1$ ; i.e.,  $\psi(t)$  belongs to the sphere  $S^3 \subset \mathbb{C}^2$ . The free Hamiltonian  $H_0$  and the controlled Hamiltonian  $H_1$  are given by

$$(4) \quad H_0 = \begin{pmatrix} -E & 0 \\ 0 & E \end{pmatrix}, \quad H_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

where  $-E$  and  $E$  ( $E > 0$ ) are the two energy levels and the control  $\Omega(\cdot)$  is a real function describing the amplitude of the external field. Here  $|\psi_1(t)|^2$  (respectively,  $|\psi_2(t)|^2$ ) represents the probability of measuring at time  $t$  the energy  $E$  (respectively,  $-E$ ). The control task consists of inducing a transition from the first eigenstate of  $H_0$  (i.e.,  $|\psi_1|^2 = 1$ ) to any other *physical state*. We recall that two states  $\psi$  and  $\psi'$  are physically equivalent if they differ by a factor of phase. More precisely by physical state we mean a point of the two-dimensional sphere (called *Bloch sphere* in this context)  $S^2 = S^3 / \sim$ , where the equivalence relation  $\sim$  is defined as follows:  $\psi \sim \psi'$  (where  $\psi, \psi' \in S^3$ ) if and only if  $\psi = \exp(i\Phi)\psi'$  for some  $\Phi \in [0, 2\pi[$ . The projection  $\Pi : S^3 \rightarrow S^2$  is called a *Hopf projection*. A particularly interesting transition is of course from the first to the second eigenstates of  $H_0$  (i.e., from  $|\psi_1|^2 = 1$  to  $|\psi_2|^2 = 1$ ).

In many applications, the external field should have bounded amplitude  $M$  (i.e.,  $|\Omega(\cdot)| \leq M$ ), and, in order to minimize the unavoidable effects of relaxation and decoherence [3, 13], the transfer should occur as quickly as possible. Therefore we end up addressing a minimum time control problem with one bounded control. Sometimes decoherence is also reduced by taking  $M$  small with respect to  $E$ : This guarantees that the energy injected by the control action into the system is close to the minimal one necessary to induce the transition.

As was shown in [9], the projection of the minimum time control problem for the system (3) on the Bloch sphere gives rise, after time renormalization, to the minimum time control problem for system (1) where (i) the first and second eigenstates of  $H_0$  project respectively onto the north pole  $N$  and the south pole  $S$ , (ii)  $\tan(\alpha) = M/E$  and  $u(t) = \Omega(t)/M$ , and (iii)  $H_0$  project on  $F$  and  $H_1$  on  $G$ . The case  $M \ll E$  corresponds now to the limit  $\alpha \rightarrow 0$ .

Nowadays two-level quantum systems are central in the implementation of the so-called quantum gates (the basic blocks of a quantum computer); see, for instance, [12, 18].

The present paper is actually a continuation of [5, 9] in the sense that it answers questions raised in these papers.

In [5], the purpose was to provide a lower and an upper bound for  $N(\alpha)$ , the maximum number of switchings for time optimal trajectories for the left invariant control system

$$(5) \quad (S)_\alpha \quad \dot{g} = g(F + uG), \quad g \in SO(3), \quad |u| \leq 1,$$

where  $F$  and  $G$  are defined in (2). Recall that, for such control systems, it is known (cf., for instance, [5, 9]) that every time optimal trajectory is a finite concatenation of bang arcs (i.e.,  $u \equiv \pm 1$ ) or singular arcs ( $u = 0$ ). A bang arc is an integral trajectory corresponding to the rotations

$$(6) \quad X_+ := F + G, \quad X_- := F - G$$

and is denoted by  $e^{tX_\varepsilon}x$ ,  $t \in [0, T]$ , where  $\varepsilon = \pm$ ,  $x$  is the starting point of the bang arc and  $T$  is its time duration. Moreover, a switching time—or simply a switching—along a time optimal trajectory is a time  $t_0$  so that the control  $u$  is not constant in any open neighborhood of  $t_0$ .

To estimate  $N(\alpha)$ , a suitable Hopf map  $\Pi : SO(3) \rightarrow S^2$  was introduced to project  $(S)_\alpha$  onto  $(\Sigma)_\alpha$ . In particular, every time optimal trajectory of  $(\Sigma)_\alpha$  is the projection by  $\Pi$  of a time optimal trajectory of  $(S)_\alpha$ . It results that, if a time optimal trajectory on  $S^2$  has a certain number of switchings, then this number is lower than or equal to the maximum number of switchings for the optimal problem on  $SO(3)$ . The construction of time optimal trajectories of  $(\Sigma)_\alpha$  was performed according to the general theory of time optimal synthesis on two-dimensional (2-D) manifolds developed in [6, 7, 10, 11, 14, 15, 19, 20] and recently gathered in the book [8].

The question of studying  $N(\alpha)$  was first addressed in [1], where, using the index theory developed by Agrachev, the authors proved that  $N(\alpha) \leq [\pi/\alpha]$ , where  $[\cdot]$  stands for the integer part. That result was not only an indirect indication that  $N(\alpha)$  would tend to infinity as  $\alpha$  tends to zero, but it also provided a hint on the asymptotic of  $N(\alpha)$  as  $\alpha$  tends to zero. Notice that for  $\alpha = 0$  the systems (1) and (5) are not controllable. With the techniques developed in [5], enough properties for the TOS associated to  $(\Sigma)_\alpha$ ,  $\alpha < \pi/4$ , were identified in order to improve the upper bound of [1] and to actually show that, for  $\alpha$  small,

$$N(\alpha) \leq k_M + 5, \quad \text{where} \quad k_M := \left\lceil \frac{\pi}{2\alpha} \right\rceil.$$

In [5], it is proved that, for  $\alpha < \pi/4$ , the extremals associated to  $(\Sigma)_\alpha$  (i.e., the trajectories candidate for time optimality obtained after using the Pontryagin maximum principle (PMP)), starting from the north pole  $N$  are bang-bang trajectories, i.e.,

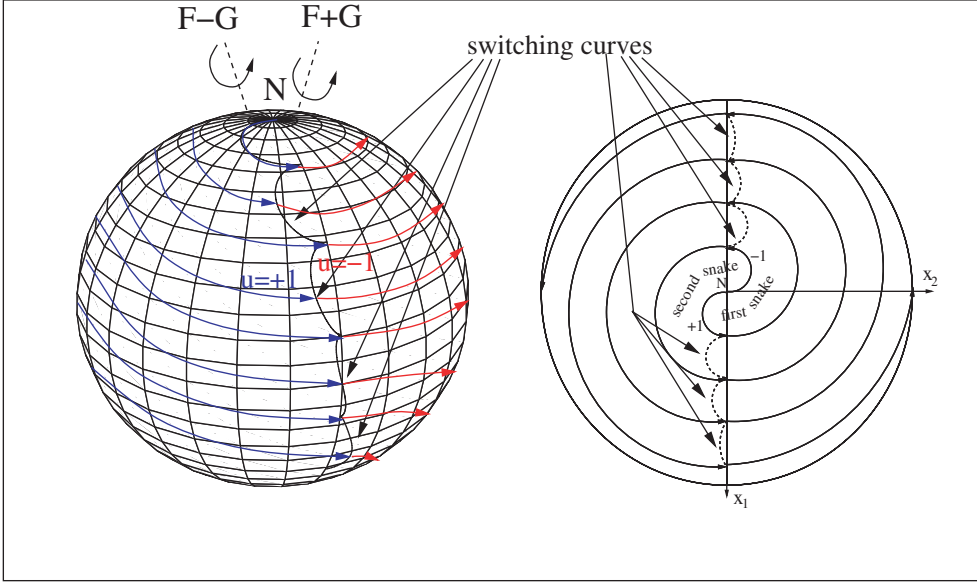


FIG. 2. The two-snakes configuration defined by the extremal flow. Notice that this set of trajectories covers the whole sphere, but in principle not all extremals are optimal, and a point can be reached by more than one trajectory at the same or at different times.

finite concatenations of bang arcs of the type

$$e^{s_f X_{-\varepsilon'}} e^{v(s_i) X_{\varepsilon'}} \dots e^{v(s_i) X_{-\varepsilon}} e^{s_i X_{\varepsilon}} N,$$

where the initial time duration  $s_i$  verifies  $s_i \in (0, \pi]$ , all of the time durations of the interior bang arcs are equal to  $v(s_i)$ , where the function  $v$  is defined in (14) below, and the final time duration  $s_f$  verifies  $s_f \leq v(s_i)$ . Of particular importance for the construction of the TOS are the *switching curves*, i.e., the curves made by points where the control switches from  $+1$  to  $-1$  or vice versa and defined inductively by

$$(7) \quad C_1^\varepsilon(s) = e^{X_\varepsilon v(s)} e^{X_{-\varepsilon} s} N, \quad C_k^\varepsilon(s) = e^{X_\varepsilon v(s)} C_{k-1}^{-\varepsilon}(s) \quad (\text{where } \varepsilon = \pm 1 \text{ and } k = 2, \dots, k_M).$$

Since the PMP gives just a necessary condition for optimality, it is crucial to determine the time after which an extremal is no more optimal. In [5], we showed that the number of bangs must be lower than or equal to  $k_M + 1$  and the extremals cover the sphere  $S^2$  according to the “two-snakes” configuration as depicted in Figure 2. The two “snakes” correspond to extremal trajectories starting with control  $+1$  and  $-1$ , respectively. For more details, see [5].

However, in [5], we were not able to construct the complete TOS associated to  $(\Sigma)_\alpha$ . In particular, we could not show the optimality of all of the extremals up to  $k_M - 1$  bangs arcs, and we could not complete analytically the construction of the synthesis in a neighborhood of the south pole  $S$ . There, the minimum time front develops singularities due to the compactness of  $S^2$ . We provided only numerical simulations describing the evolution of the extremal front in a neighborhood of the south pole. As  $\alpha \rightarrow 0$ , these numerical simulations suggested the emergence of an interesting phenomenon (see Figure 3): Define the remainder

$$(8) \quad r(\alpha) := \pi/2\alpha - [\pi/2\alpha].$$



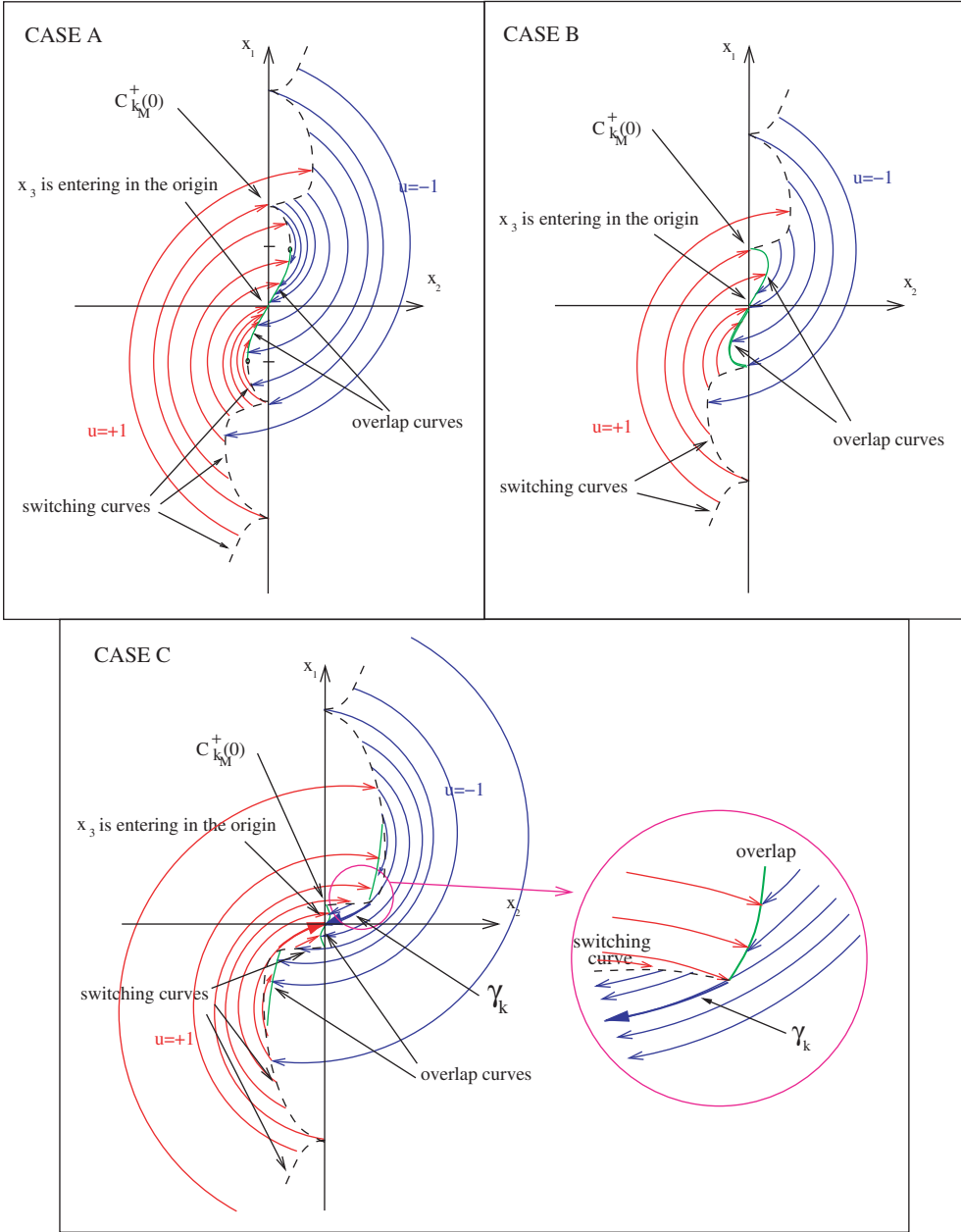


FIG. 3. Conjectured shapes of the synthesis in a neighborhood of the south pole. Switching curves are  $C^1$  curves made by points in which the control switches from  $+1$  to  $-1$  or vice versa. Overlap curves are  $C^1$  curves made by points reached optimally by more than one trajectory. The curve  $\gamma_k$  is a bang arc that is also an overlap curve since trajectories having a different history travel on it at the same time. The singularity appearing at the starting point of  $\gamma_k$  (called  $(C, K)_1$  according to the taxonomy of [8]) is a singularity of the synthesis predicted by the general theory [8], and it is due to a nonlocal phenomenon (see section 7 for more details).

Then there are three possible patterns of TOS in the neighborhood of the south pole  $S$ , each of them depending on a relation between  $r(\alpha)$  and  $\alpha$ . See section 2.2 below, where these relations are formulated as well as conjectures.

In [9], the TOS for  $(\Sigma_\alpha)$  was studied in the context of quantum control as described previously. In that paper, the TOS for  $\alpha \geq \pi/4$  was completed, and, in the case  $\alpha < \pi/4$ , further information was obtained, for what concerns time optimal trajectories steering the north to the south pole (in fact the most interesting trajectories for the quantum mechanical problem). Such optimal trajectories belong to a set  $\Xi$  containing at most eight trajectories, half of them starting with control  $+1$  and the other half starting with control  $-1$ , and switching exactly at the same times. It was also proved that the cardinality of  $\Xi$  depends on the remainder  $r(\alpha)$  defined in (8). For instance, for  $\alpha$  and  $r(\alpha)$  small enough, then  $\Xi$  contains exactly eight trajectories (four of them are optimal) while if  $r(\alpha)$  is close to 1, then  $\Xi$  contains only four trajectories (two of them are optimal).

The purpose of the present paper consists in studying the TOS associated to  $(\Sigma)_\alpha$  as  $\alpha$  tends to zero, focusing in particular on its behavior inside a neighborhood of the south pole. Roughly speaking, we want to determine, as  $\alpha$  tends to zero, what could be a possible limit for the TOS associated to  $(\Sigma)_\alpha$  (as suggested, for instance, by the patterns depicted in Figure 3) and then to prove the convergence (in some suitable sense) of the TOS associated to  $(\Sigma)_\alpha$  to that limit. To proceed, we embark on the study of a geometric object  $\mathcal{F}(\alpha, T)$  called the *extremal front at time  $T$*  along  $(\Sigma)_\alpha$  and defined as the set of points reached at time  $T$  by extremal trajectories starting from  $N$  (see section 3.1 for a precise definition). The extremal front  $\mathcal{F}(\alpha, T)$  contains the *minimum time front*  $OF(\alpha, T)$ , i.e., the set of points reached at time  $T$  by time optimal trajectories. When  $\mathcal{F}(\alpha, T) = OF(\alpha, T)$ , we say that  $\mathcal{F}(\alpha, T)$  is *optimal*.

We first prove, in the case in which  $k_M$  is odd (the other case being analogous), that the extremal front  $\mathcal{F}(\alpha, k_M\pi)$  is made up of the union of two curves  $\mathcal{E}^\varepsilon(\alpha, \cdot) : (0, \pi] \rightarrow S^2$ ,  $\varepsilon = \pm$ , with  $\mathcal{E}^\varepsilon(\alpha, \cdot) = \Pi_{x_3} \mathcal{E}^{-\varepsilon}(\alpha, \cdot)$ , where  $\Pi_{x_3}$  is the orthogonal symmetry with respect to the  $x_3$ -axis. Moreover, for  $\alpha$  small enough,  $\mathcal{E}^\varepsilon(\alpha, \cdot)$  admits a convergent power series of the type  $\sum_{l \geq 0} f_l^\varepsilon(s, r(\alpha))\alpha^l$ , where the  $f_l^\varepsilon(s, r)$  are real-analytic functions of  $(s, r) \in \mathbb{R}^2$ ,  $2\pi$ -periodic in  $s$  with

$$(9) \quad \begin{aligned} f_0^+(s, r) &= \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}, & f_1^+(s, r) &= \begin{pmatrix} -2r\mathbf{c}_s \\ 2r\mathbf{s}_s \\ 0 \end{pmatrix}, \\ f_2^+(s, r) &= \begin{pmatrix} \frac{\pi}{2}(4r + \mathbf{c}_s)\mathbf{s}_s^2 \\ \frac{\pi}{4}(3 + 8r\mathbf{c}_s + \mathbf{c}_{2s})\mathbf{s}_s \\ 2r^2 \end{pmatrix}, & f_l^-(s, r) &= \Pi_{x_3} f_l^+(s, r). \end{aligned}$$

As a trivial consequence, we deduce that, for  $r \in [0, 1]$ ,  $s \in \mathbb{R}$ , and  $\alpha$  small enough, we have

$$(10) \quad \mathcal{E}^\varepsilon(\alpha, s) = f_0^\varepsilon(s, r(\alpha)) + f_1^\varepsilon(s, r(\alpha))\alpha + f_2^\varepsilon(s, r(\alpha))\alpha^2 + \mathcal{O}(\alpha^3)$$

and

$$(11) \quad \frac{\partial}{\partial s} \mathcal{E}^\varepsilon(\alpha, s) = \frac{\partial}{\partial s} f_1^\varepsilon(s, r(\alpha))\alpha + \frac{\partial}{\partial s} f_2^\varepsilon(s, r(\alpha))\alpha^2 + \mathcal{O}(\alpha^3),$$

where  $|\mathcal{O}(\alpha^3)| \leq \bar{C}|\alpha|^3$ , with  $\bar{C} > 0$  constant independent of  $(r, s, \alpha)$ .

Then we show that  $\mathcal{F}(\alpha, T)$  is actually optimal for  $T \leq (k_M - 1)\pi$  and  $\alpha$  small enough (see Remark 7 below). Moreover, we show that  $\mathcal{F}(\alpha, (k_M - 1)\pi)$  is a circle of radius  $2(1 + r(\alpha))\alpha$  up to order  $\alpha^2$  (see Remark 6 below). As a consequence of the optimality of  $\mathcal{F}(\alpha, (k_M - 1)\pi)$ , we get that all of the extremals of the two-snakes configuration depicted in Figure 2 are optimal up to time  $(k_M - 1)\pi$ . In other words, if  $U_\alpha$  is the connected component of  $S^2 \setminus \mathcal{F}(\alpha, (k_M - 1)\pi)$  containing the south pole,

we obtain the optimal synthesis on  $S^2 \setminus U_\alpha$ . Notice that  $U_\alpha$  is a neighborhood of the south pole of size proportional to  $\alpha$ .

In that way we answer the question stated in [5] about optimality of extremals of the two-snake configuration in the case  $\alpha$  small, and it is, of course, of interest for applications to the two-level quantum system.

The expressions (10)–(11) are central tools to understand the possible asymptotic behaviors of the TOS associated to  $(\Sigma)_\alpha$ , as  $\alpha$  tends to zero.

For this purpose we observe that the expressions of  $f_1^+$  and  $f_2^+$  in (9) depend explicitly on the remainder  $r(\alpha)$ . This fact suggests the need to impose particular relationships between  $\alpha$  and  $r(\alpha)$  in order to define any asymptotic behavior. In other words we must let  $\alpha$  go to zero only along certain subsequences  $(\alpha_k)_{k \geq 0}$  where a specific relationship holds between  $\alpha_k$  and  $r(\alpha_k)$ . The analysis of (9) will help us to determine such relationships and to prove that the conjectures made in [5] about the qualitative shape of the synthesis near the south pole were true (see section 2.2 and Figure 3). In particular we will see that there are exactly three qualitatively different asymptotic behaviors of the synthesis as  $\alpha$  goes to zero, described by the following cases.

First, we analyze the case in which  $\alpha$  is arbitrarily small, with  $r(\alpha) \in (0, 1)$  uniformly far from 0 and 1. To further simplify the discussion, it is reasonable to consider the following:

(C1) For  $\bar{r} \in (0, 1)$ , let  $\alpha$  tend to zero along the subsequence  $\alpha_k := \frac{\pi}{2(k + \bar{r})}$ , so that

$$r(\alpha_k) = \bar{r}.$$

In this case  $\mathcal{E}^\varepsilon(\alpha, \cdot)$  is approximated, up to order  $\alpha^2$ , by the expression  $S + f_1^\varepsilon(\cdot, \bar{r})$ . As a consequence  $\mathcal{F}(\alpha, k_M \pi)$  is approximately a circle of radius  $2\bar{r}\alpha$  centered at the south pole. We are then able to give a qualitative description of the optimal synthesis, as stated below in Theorem 1. We then deduce that, if  $\alpha$  is small enough and  $r(\alpha)$  is far enough from 0 and 1, the synthesis in a neighborhood of the south pole is topologically equivalent to the limit synthesis obtained, as  $k$  tends to infinity, along the sequence  $\alpha_k$  above. That synthesis turns out to be exactly the one described in Figure 3 (case B), as predicted in [5].

It remains then to study the cases in which  $r(\alpha)$  can be arbitrarily close to 0 or 1. For this purpose we first consider the case in which  $r(\alpha)/\alpha$  remains bounded above and below by positive constants as  $\alpha$  tends to zero. From (9) it is clear that this is equivalent to saying that  $f_1^\varepsilon(\cdot, r)\alpha$  is comparable to  $f_2^\varepsilon(\cdot, r)\alpha^2$ . For simplicity we consider the following:

(C2) For  $C > 0$ , let  $\alpha$  tend to zero along a subsequence  $(\alpha_k)_{k \geq 0}$  such that

$$r(\alpha_k) = C\alpha_k.$$

In this case  $\mathcal{E}^\varepsilon(\alpha, \cdot)$  is well approximated by  $S + (f_1^\varepsilon(\cdot, C) + f_2^\varepsilon(\cdot, 0))\alpha^2$ . If  $C > \pi/4$ , the synthesis is equivalent to that of the previous case. On the other hand, if  $C < \pi/4$ , the synthesis is more complicated (see section 5), and it turns out to be exactly the one described in Figure 3 (case C), as predicted in [5].

If  $\alpha$  and  $r(\alpha)$  tend to zero, with  $r(\alpha)/\alpha$  tending to infinity (respectively, to zero), it is possible to see that the synthesis is qualitatively equivalent to the one of case (C1) (respectively, (C2)).

The third interesting case is the following:

(C3) Let  $\alpha$  tend to zero along the subsequence  $\alpha_k := \frac{\pi}{2k}$ , so that  $r(\alpha_k) = 0$ .

In this case the extremal front at time  $k_M\pi$  contains the south pole, and the corresponding optimal front reduces to that point. The optimal synthesis is then described starting from the extremal front  $\mathcal{F}(\alpha, (k_M - 1)\pi) = OF(\alpha, (k_M - 1)\pi)$ , and it corresponds to the one described in Figure 3 (case A), as predicted in [5].

With similar arguments, one can see that, in the case in which  $\alpha$  is small and  $r(\alpha)$  is close to 1, the optimal synthesis is qualitatively equivalent either to that of case (C1) or to that of case (C3), and this concludes the description of the possible asymptotic behaviors as  $\alpha$  tends to 0.

*Remark 1.* It is interesting to notice that numerical simulations show that, for  $\alpha$  decreasing to zero continuously, the qualitative shape of the optimal synthesis described in Figure 3 alternates cyclically in the order BCABCA...

Let us describe the results obtained in case (C1) in more details. Since  $\mathcal{F}(\alpha, k_M\pi)$  is approximated, up to  $\mathcal{O}(\alpha^2)$ , by a circle of center  $S$  and radius  $2\bar{r}\alpha$ , we are able to show that it is optimal, so that all of the extremals of the two-snakes configuration depicted in Figure 2 are optimal up to time  $k_M\pi$ . In other words, if  $V_\alpha$  is the connected component of  $S^2 \setminus \mathcal{F}(\alpha, k_M\pi)$  containing the south pole, we obtain the optimal synthesis on  $S^2 \setminus V_\alpha$ .

As  $\alpha$  tends to zero,  $V_\alpha$  collapses on  $S$ . Hence one must rescale the problem by a factor  $1/\alpha$  in order to describe the TOS inside  $V_\alpha$ . Also notice that since we are in a neighborhood of the south pole we can project the problem on the plane  $(x_1, x_2)$ . We are now in a position to define a possible limit behavior for the TOS inside  $V_\alpha$ . Let  $M_\alpha$  be the linear mapping from  $\mathbb{R}^3$  onto  $\mathbb{R}^2$  defined as the composition of the projection  $(x_1, x_2, x_3) \mapsto (x_1, x_2)$  followed by the dilation by  $1/\alpha$ . Denote by  $(\tilde{\Sigma})_\alpha$  (respectively,  $OF(\alpha, k_M\pi)$ ) the image by  $M_\alpha$  of  $(\Sigma)_\alpha$  (respectively,  $OF(\alpha, k_M\pi)$ ). Then  $(\tilde{\Sigma})_\alpha$  is a perturbation by  $\mathcal{O}(\alpha^2)$  of the forced linear pendulum

$$(12) \quad (Pen) : \begin{cases} \dot{z}_1 = -z_2, \\ \dot{z}_2 = z_1 + u, \end{cases} \quad (z_1, z_2) \in \mathbb{R}^2, \quad |u| \leq 1,$$

while  $\widetilde{OF}(\alpha, k_M\pi)$  is a perturbation by  $\mathcal{O}(\alpha^2)$  of  $C(0, 2\bar{r})$ , the planar circle of center  $(0, 0)$  and radius  $2\bar{r}$ . As a consequence, the candidate limit TOS inside  $V_\alpha$  is the one associated to the problem of reaching in minimum time every point of the ball  $B(0, 2\bar{r})$  starting from  $C(0, 2\bar{r})$ , along the dynamics of the standard linearized pendulum. To prove such a result, we first study the above-mentioned optimal control problem and show that the corresponding TOS is characterized by an overlap curve  $\gamma_{pen}^o$ , which is the set of points  $z \in \mathbb{R}^2$ , with  $z_1 z_2 \geq 0$ , and belonging to the locus (see Figure 4)

$$z_1^4 + z_2^4 + 2z_1^2 z_2^2 - 4\bar{r}^2 z_1^2 + (4 - 4\bar{r}^2) z_2^2 = 0.$$

The optimal synthesis inside  $C(0, 2\bar{r})$  is then described by the following feedback, defined on  $B(0, 2\bar{r}) \setminus \gamma_{pen}^o$ : “Above”  $\gamma_{pen}^o$ , the control  $u$  is constantly equal to  $-1$ , and “below”  $\gamma_{pen}^o$ , it is constantly equal to  $1$  (see Figure 5). Finally, the asymptotic result we prove in section 4.2 is the following.

**THEOREM 1.** *For  $\bar{r} \in (0, 1)$ , let  $(\alpha_k)_{k \geq 1}$  be the sequence defined by  $\alpha_k := \frac{\pi}{2(k+\bar{r})}$  for  $k \geq 1$ . Consider  $\gamma_{pen}^o$ , the overlap curve of the TOS for the optimal control problem consisting of starting from  $C(0, 2\bar{r})$ , the planar circle of center  $(0, 0)$  and radius  $2\bar{r}$ , and reaching in minimum time every point of  $B(0, 2\bar{r})$  along the control system (12). Then, for  $k$  large enough, the TOS associated to  $(\tilde{\Sigma})_{\alpha_k}$  inside  $\widetilde{OF}(\alpha_k, k_M\pi)$  is characterized by an overlap curve  $\gamma_{\alpha_k}^o$  so that the optimal feedback takes the value  $-1$  above  $\gamma_{\alpha_k}^o$  and the value  $1$  below  $\gamma_{\alpha_k}^o$ . Moreover,  $\gamma_{\alpha_k}^o$  converges to  $\gamma_{pen}^o$  in the  $C^0$  topology, uniformly with respect to  $\bar{r}$  in any compact interval of  $(0, 1)$ .*

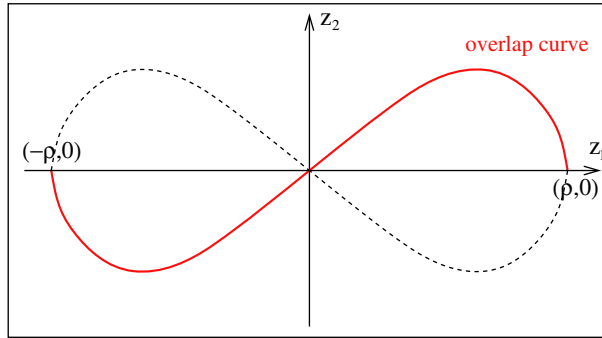


FIG. 4. The overlap curve for the pendulum problem.

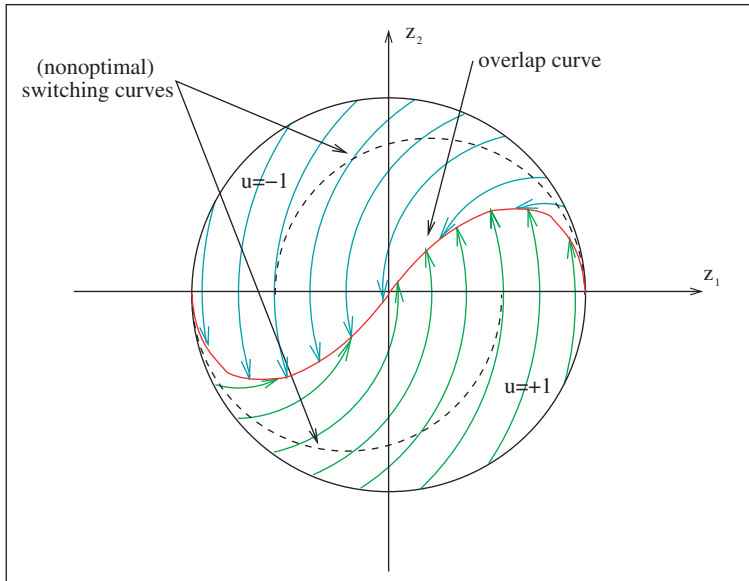


FIG. 5. Optimal synthesis for the linear pendulum.

The results in cases (C2) and (C3) are described in more details in sections 5 and 6.

*Remark 2.* Notice that the sequence  $(\alpha_k)_{k \geq 1}$  defined above has been chosen in order to simplify the previous statement. Indeed the same result could be restated in a more general way by taking an arbitrary sequence  $(\tilde{\alpha}_k)_{k \geq 1}$  converging to zero and such that  $r(\tilde{\alpha}_k)$  converges to  $\bar{r}$  or by letting the remainder vary on a compact subinterval of  $(0, 1)$ .

The paper is organized as follows. In the second section, we collect basic facts, notations, results, and conjectures of [5]. The third section gathers the detailed description of the extremal front and the proof of (9). Sections 4, 5, and 6 treat, respectively, cases (C1), (C2), and (C3). In section 7, we make some final remarks, and we stress the importance of our result in connection with singularity theory for time optimal syntheses on 2-D manifolds. In the appendix, we finally prove a technical result needed throughout the paper.

## 2. Notations and previous results.

### 2.1. Basic facts.

DEFINITION 1. An admissible control  $u(\cdot)$  for the system (1)–(2) is a measurable function  $u(\cdot) : [a, b] \rightarrow [-1, 1]$ , while an admissible trajectory is an absolutely continuous function  $x(\cdot) : [a, b] \rightarrow S^2$  satisfying (1) a.e. for some admissible control  $u(\cdot)$ . If  $x(\cdot)$  is an admissible trajectory and  $u(\cdot)$  the corresponding control, we say that  $(x(\cdot), u(\cdot))$  is an admissible pair.

For every  $\bar{x} \in S^2$ , the minimization problem consists of determining an admissible pair steering the north pole to  $\bar{x}$  in minimum time. More precisely we write the following.

Problem (P). Consider the control system (1)–(2). For every  $\bar{x} \in S^2$ , find an admissible pair  $(x(\cdot), u(\cdot))$  defined on  $[0, T]$  such that  $x(0) = N$ ,  $x(T) = \bar{x}$ , and  $x(\cdot)$  is time optimal.

An optimal synthesis from the north pole (in the following optimal synthesis, for short) is the collection of all of the solutions to the problem (P). More precisely we write the following.

DEFINITION 2 (optimal synthesis). An optimal synthesis for the problem (P) is the collection of all time optimal trajectories  $\Gamma = \{x_{\bar{x}}(\cdot) : [0, b_{\bar{x}}] \mapsto S^2, \bar{x} \in S^2 : x_{\bar{x}}(0) = N, x_{\bar{x}}(b_{\bar{x}}) = \bar{x}\}$ .

For more elaborated definitions of optimal synthesis, see [8, 16], and references therein. The standard tool to look for optimal trajectories is a first order necessary condition for optimality known as the PMP (cf. [2, 17]) as stated below for our minimum time problem on  $S^2$ .

Define the following real-valued map on  $T^*S^2 \times [-1, 1]$ , called the Hamiltonian:

$$\mathcal{H}(\lambda, x, u) := \langle \lambda, (F + uG)x \rangle.$$

Set

$$(13) \quad H(\lambda, x) := \max_{v \in [-1, 1]} \mathcal{H}(\lambda, x, v).$$

The PMP asserts that, if  $\gamma : [a, b] \rightarrow S^2$  is a time optimal trajectory corresponding to a control  $u : [a, b] \rightarrow [-1, 1]$ , there exists a nontrivial field of covectors along  $\gamma$  that is a never vanishing absolutely continuous function  $\lambda : t \in [a, b] \mapsto \lambda(t) \in T_{\gamma(t)}^*S^2$  and a constant  $\lambda_0 \leq 0$  such that, for a.e.  $t \in \text{Dom}(\gamma)$ , we have the following:

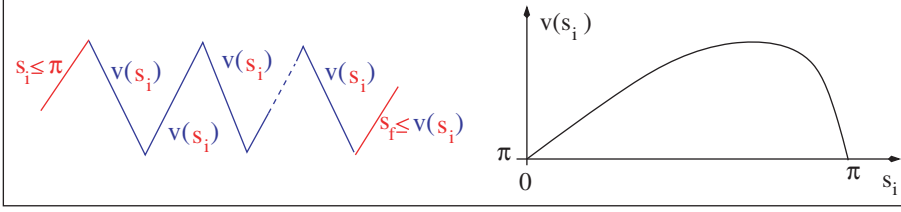
- (i)  $\dot{\lambda}(t) = -\frac{\partial \mathcal{H}}{\partial x}(\lambda(t), \gamma(t), u(t)) = -\lambda(t)(F + u(t)G)$ ,
- (ii)  $\mathcal{H}(\lambda(t), \gamma(t), u(t)) + \lambda_0 = 0$ ,
- (iii)  $\mathcal{H}(\lambda(t), \gamma(t), u(t)) = H(\gamma(t), \lambda(t))$ .

In the more general case in which the target and the initial datum (also called *source*) are two smooth manifolds  $\mathcal{N}_0$  and  $\mathcal{N}_1$ , the previous statement must be modified by adding the so-called *transversality conditions*:

- (iv)  $\langle \lambda(a), v \rangle = 0$  for all  $v \in T_{\gamma(a)}\mathcal{N}_0$ ,  $\langle \lambda(b), w \rangle = 0$  for all  $w \in T_{\gamma(b)}\mathcal{N}_1$ .

Remark 3. A trajectory  $\gamma$  (respectively, a couple  $(\gamma, \lambda)$ ) satisfying the conditions given by the PMP is said to be an *extremal* (respectively, an *extremal pair*). An extremal corresponding to  $\lambda_0 = 0$  is said to be an *abnormal extremal*; otherwise we call it a *normal extremal*.

DEFINITION 3 (bang, singular for the problem (1)–(2)). A control  $u(\cdot) : [a, b] \rightarrow [-1, 1]$  is said to be a bang control if  $u(t) = +1$  a.e. in  $[a, b]$  or  $u(t) = -1$  a.e. in  $[a, b]$ . A control  $u(\cdot) : [a, b] \rightarrow [-1, 1]$  is said to be a singular control if  $u(t) = 0$  a.e. in  $[a, b]$ . A finite concatenation of bang controls is called a bang-bang control. A

FIG. 6. Time optimal trajectories for  $\alpha < \pi/4$ .

switching time of  $u(\cdot)$  is a time  $\bar{t} \in [a, b]$  such that, for every  $\varepsilon > 0$ ,  $u$  is not bang or singular on  $(\bar{t} - \varepsilon, \bar{t} + \varepsilon) \cap [a, b]$ . A trajectory of the control system (1)–(2) is said to be a bang trajectory (or arc), singular trajectory (or arc), or bang-bang trajectory if it corresponds, respectively, to a bang control, singular control, or bang-bang control. If  $\bar{t}$  is a switching time, the corresponding point on the trajectory  $x(\bar{t})$  is called a switching point.

**2.2. Description of previous results.** In [5, 9] it was proved that, for every couple of points, there exists a time optimal trajectory joining them. Moreover it was proved that every time optimal trajectory is a finite concatenation of bang and singular trajectories. More precisely we have the following.

**PROPOSITION 1.** *For the minimum time problem associated to (1)–(2), for each pair of points  $p$  and  $q$  belonging to  $S^2$ , there exists a time optimal trajectory joining  $p$  to  $q$ . Moreover every time optimal trajectory for (1)–(2) is a finite concatenation of bang and singular trajectories.*

Notice that the previous proposition does not apply if  $\alpha = 0$  or  $\alpha = \pi/2$ , since in these cases the controllability property is lost.

In [9] it has been proved that  $\alpha = \pi/4$  is a bifurcation for the qualitative shape of the time optimal synthesis; for instance, the time optimal synthesis contains a singular arc if and only if  $\alpha > \pi/4$ . Since in this paper we are interested in the limit  $\alpha \rightarrow 0$ , in the following we always assume  $\alpha < \pi/4$ . In this case, using the PMP, the following properties characterizing the optimal trajectories were established in [5] (see Figure 6):

- (i)  $x(\cdot)$  is bang bang;
- (ii) the duration  $s_i$  of the first bang arc satisfies  $s_i \in (0, \pi]$ ;
- (iii) the time duration between two consecutive switchings is the same for all interior bang arcs (i.e., excluding the first and the last bangs), and it is equal to  $v(s_i)$ , where  $v(\cdot)$  is the following function:

$$(14) \quad v(s) = \pi + 2 \arctan \left( \frac{\sin(s)}{\cos(s) + \cot^2(\alpha)} \right).$$

One can immediately check that this function satisfies  $v(0) = v(\pi) = \pi$  and  $v(s) > \pi$  for every  $s \in (0, \pi)$ ;

- (iv) the time duration of the last arc is  $s_f \in (0, v(s_i)]$ .

Moreover, thanks to the analysis given in [5], one easily gets (always in the case  $\alpha < \pi/4$ ):

- (v) the number of switchings  $N_x$  of  $x(\cdot)$  satisfies the following inequality:

$$(15) \quad N_x \leq \left\lceil \frac{\pi}{2\alpha} \right\rceil + 1.$$

Conditions (i)–(v) define a set of candidate optimal trajectories. Notice that conditions (i)–(v) are just necessary conditions for optimality, and one is faced with the

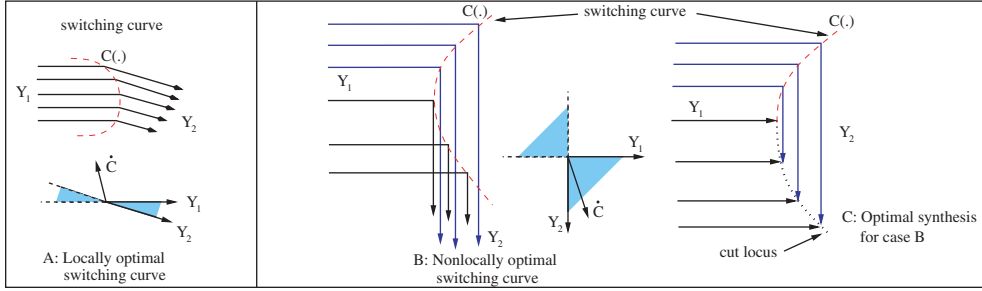


FIG. 7. *Locally optimal switching curves and non locally optimal switching curves with the corresponding synthesis.*

problem of selecting, among them, those that are really optimal. In particular, given a trajectory satisfying conditions (i)–(v), one would like to find the time after which it is no more optimal.

Some questions remained unsolved, in particular questions relative to local optimality of the switching curves defined in (7). Roughly speaking we say that a switching curve is locally optimal if it never “reflects” the trajectories (see Figure 7(A)).<sup>1</sup> When a family of trajectories is reflected by a switching curve, then local optimality is lost, and some *cut loci* appear in the optimal synthesis.

**DEFINITION 4.** *A cut locus is a set of points reached at the same time by two (or more) time optimal trajectories. A subset of a cut locus that is a connected  $\mathcal{C}^1$  manifold is called an overlap curve.*

An example showing how a “reflection” on a switching curve generates a cut locus is portrayed in Figures 7(B) and 7(C). More precisely the following questions were formulated in [5]:

- Question 1.* Are the switching curves  $C_k^\varepsilon(s)$ ,  $s \in (0, \pi]$ , locally optimal? More precisely, one would like to understand how the candidate optimal trajectories described above lose optimality.
- Question 2.* What is the shape of the optimal synthesis in a neighborhood of the south pole?

Numerical simulations suggested some conjectures regarding the above questions. More precisely, in [5] the following conjectures were made:

- C1. The curves  $C_k^\varepsilon(s)$ , ( $k = 1, \dots, k_M$ ) are locally optimal if and only if  $k \leq \left\lceil \frac{\pi - \alpha}{2\alpha} \right\rceil - 1$ .

Analyzing the evolution of the minimum time wave front in a neighborhood of the south pole, it is reasonable to conjecture that:

- C2. The shape of the optimal synthesis in a neighborhood of the south pole depends on the remainder  $r(\alpha)$  defined in (8). Notice that  $r(\alpha)$  belongs to the interval  $[0, 1)$ . More precisely, it was conjectured in [5] that for  $\alpha \in (0, \pi/4)$  there exist two positive numbers  $\alpha_1$  and  $\alpha_2$  such that  $0 < \alpha_1 < \alpha < \alpha_2 < 2\alpha$  and the following:

Case A:  $r(\alpha) \in (\frac{\alpha_2}{2\alpha}, 1)$ . The switching curve  $C_{k_M}^\varepsilon$  glues to an overlap curve that passes through the origin (Figure 3, case A).

<sup>1</sup>More precisely consider a smooth switching curve  $C$  between two smooth vector fields  $Y_1$  and  $Y_2$  on a smooth two-dimensional manifold. Let  $C(s)$  be a smooth parametrization of  $C$ . We say that  $C$  is *locally optimal* if, for every  $s \in \text{Dom}(C)$ , we have  $\dot{C}(s) \neq \alpha_1 Y_1(C(s)) + \alpha_2 Y_2(C(s))$  for every  $\alpha_1, \alpha_2$  such that (s.t.)  $\alpha_1 \alpha_2 \geq 0$ . The points of a switching curve on which this relation is not satisfied are usually called “conjugate points.” See Figure 7.



Case B:  $r(\alpha) \in [\frac{\alpha_1}{2\alpha}, \frac{\alpha_2}{2\alpha}]$ . The switching curve  $C_{k_M}^\varepsilon$  is not reached by optimal trajectories in the interval  $]0, \pi]$ . At the point  $C_{k_M}^\varepsilon(0)$ , an overlap curve starts and passes through the origin (Figure 3, case B).

Case C:  $r(\alpha) \in (0, \frac{\alpha_1}{2\alpha})$ . The situation is more complicated, and it is depicted in the bottom of Figure 3, case C.

For  $r = 0$ , the situation is the same as in Case A but for the switching curve starting at  $C_{k_M-1}^\varepsilon(0)$ .

As explained in the introduction, the presence of several cyclically alternating patterns of optimal synthesis, each of them depending on an arithmetic property of  $\alpha$ , was already confirmed in [9] by counting the number of optimal trajectories reaching the south pole.

*Remark 4.* The first conjecture is implicitly disproved by the results of this paper. More precisely an immediate consequence of our results is that the switching curve  $C_{k_M-2}^\varepsilon$  is always locally optimal, while  $C_{k_M-1}^\varepsilon$  is not, in general. However, for every fixed  $\bar{r} < \frac{1}{2}$  there exists  $\alpha$  small enough with  $\bar{r} \leq r(\alpha) < \frac{1}{2}$  such that  $C_{k_M-1}^\varepsilon$  is locally optimal too, which contradicts the conjecture. On the other hand, conjecture C2 is correct and, in light of our main results, is completely proved and clarified.

**2.3. Notations.** All throughout the paper we use the notation  $\varepsilon = \pm 1$ . The set  $so(3)$  of  $3 \times 3$  skew-symmetric matrices is a three-dimensional vector space on which the bilinear map

$$\langle A, B \rangle = -\text{Tr}(AB), \quad A, B \in so(3),$$

is an inner product. For  $A \in so(3)$ ,  $\|A\| := \sqrt{\langle A, A \rangle}$  is the norm (or length) of  $A$ . With the above notations,  $F$  and  $G$  are perpendicular and normalized so that  $\|F\| = \cos(\alpha)$  and  $\|G\| = \sin(\alpha)$ .

Let  $Id$  be the  $3 \times 3$  identity matrix. We recall that  $N = (0, 0, 1)^T$  and denote the south pole as  $S = (0, 0, -1)^T$ . Set  $\mathbf{c}_t := \cos(t)$  and  $\mathbf{s}_t := \sin(t)$  for  $t \in [0, 2\pi)$ . Recall that  $X_+ := F + G$  and  $X_- := F - G$ , and we have

$$X_+ = \begin{pmatrix} 0 & -\mathbf{c}_\alpha & 0 \\ \mathbf{c}_\alpha & 0 & -\mathbf{s}_\alpha \\ 0 & \mathbf{s}_\alpha & 0 \end{pmatrix}, \quad X_- = \begin{pmatrix} 0 & -\mathbf{c}_\alpha & 0 \\ \mathbf{c}_\alpha & 0 & \mathbf{s}_\alpha \\ 0 & -\mathbf{s}_\alpha & 0 \end{pmatrix}.$$

Let  $\Pi_{x_3}$  be the orthogonal symmetry with respect to the  $x_3$ -axis; i.e.,  $\Pi_{x_3}$  is represented in the canonical basis by  $\text{Diag}(-1, -1, 1)$ . Then we have the following trivial but useful property:

$$(16) \quad \Pi_{x_3} X_\varepsilon = X_{-\varepsilon} \Pi_{x_3}.$$

We next recall standard formulas for a rotation  $e^{tY}$  of  $SO(3)$  in terms of its axis  $Y$  (whose length is equal to one) and its angle  $t$ . We have

$$(17) \quad e^{tY} = Id + \mathbf{s}_t Y + (1 - \mathbf{c}_t) Y^2.$$

Moreover, for  $t \in [0, 2\pi)$ , we have

$$e^{\Theta(t)Z_-(t)} := e^{tX_+} e^{tX_-}, \quad e^{\Theta(t)Z_+(t)} := e^{tX_-} e^{tX_+},$$

where the matrices  $Z_+(t)$  and  $Z_-(t)$  are defined by

$$(18) \quad Z_+(t) = \begin{pmatrix} 0 & -C(t) & -B(t) \\ C(t) & 0 & 0 \\ B(t) & 0 & 0 \end{pmatrix}, \quad Z_-(t) = \begin{pmatrix} 0 & -C(t) & B(t) \\ C(t) & 0 & 0 \\ -B(t) & 0 & 0 \end{pmatrix},$$

respectively, with  $B(t) := \frac{\mathbf{s}_\alpha \mathbf{s}_{t/2}}{\sqrt{\mathbf{s}_{t/2}^2 \mathbf{s}_\alpha^2 + \mathbf{c}_{t/2}^2}}$ ,  $C(t) := -\frac{\mathbf{c}_{t/2}}{\sqrt{\mathbf{s}_{t/2}^2 \mathbf{s}_\alpha^2 + \mathbf{c}_{t/2}^2}}$ , and the angle  $\Theta(t)$  by

$$(19) \quad \Theta(t) = 2 \arccos(\mathbf{s}_{t/2}^2 \mathbf{c}_{2\alpha} - \mathbf{c}_{t/2}^2).$$

### 3. The extremal front.

**3.1. Definition and description.** As said in the introduction,  $\mathcal{F}(\alpha, T)$  the extremal front along  $(\Sigma)_\alpha$  at time  $T$  is the set of points reached at time  $T$  by extremal trajectories starting from  $N$ , i.e.,

$$(20) \quad \mathcal{F}(\alpha, T) := \{\bar{x} \in S^2 : \exists \text{ an extremal pair } (x(\cdot), \lambda(\cdot)) \text{ s.t. } x(0) = N, \ x(T) = \bar{x}\}.$$

Such extremals are parametrized by the length of the first bang arc, the length of the last bang arc, and the number of arcs:

$$(21) \quad \Xi^+(s, t) = \overbrace{e^{X_\varepsilon t} e^{X_{-\varepsilon} v(s)} \dots e^{X_{-v(s)}} e^{X_{+s}}}^{n \text{ terms}} N,$$

$$(22) \quad \Xi^-(s, t) = \overbrace{e^{X_{\varepsilon'} t} e^{X_{-\varepsilon'} v(s)} \dots e^{X_{+v(s)}} e^{X_{-s}}}^{n' \text{ terms}} N,$$

where  $s \in (0, \pi]$ ,  $t \in (0, v(s)]$ , the number of bang arcs ( $n$  and  $n'$ , respectively) is an integer and

- $\varepsilon = +1$  (respectively,  $\varepsilon = -1$ ) if  $n$  is odd (respectively, even), and
- $\varepsilon' = +1$  (respectively,  $\varepsilon' = -1$ ) if  $n'$  is even (respectively, odd).

Roughly speaking, we would like to compute the limit, as  $\alpha \rightarrow 0$ , of  $\mathcal{F}(\alpha, T)$ , when  $T$  is such that the extremal front reaches a neighborhood of the south pole.

The idea is that, once one knows the extremal front  $\mathcal{F}(\alpha, T)$  and if it is optimal, then one can continue to build the synthesis for times bigger than  $T$  using  $\mathcal{F}(\alpha, T)$  as a *source* for the minimization problem.

The identification of the front  $\mathcal{F}(\alpha, T)$  is not easy since it requires the computation of the product of several exponentials of matrices. Moreover, if  $\mathcal{F}(\alpha, T)$  crosses some switching curve, then the number of exponentials in general depends on the point.

This problem is overcome by considering  $\mathcal{F}(\alpha, T)$  only at times equal to multiples of  $\pi$ . Indeed, first notice that, for  $T = \pi \left\lfloor \frac{\pi}{2\alpha} \right\rfloor$ , the extremal front reaches the points  $C_{k_M}^\pm(0)$ , i.e., the points where the last switching curves  $C_{k_M}^\pm$  start. Thanks to Proposition 2 below, at these times, every extremal trajectory has the same number of switchings. The extremal front at times that are not a multiple of  $\pi$  can be obtained a posteriori, continuing the extremal front, as explained above.

From the structure of the extremal trajectories it follows that the time at which the point  $C_k^\pm(s)$  is reached is  $T_k(s) = s + kv(s)$ .

LEMMA 1. *Let  $k$  be an integer satisfying  $1 \leq k \leq \mathcal{N}_{mon} := \left\lfloor \frac{(\cot(\alpha)^2 - 1)^2}{2\cot(\alpha)^2 - 1} \right\rfloor$ , and then  $T_k(s)$  is a strictly increasing function of  $s$ .*

*Proof.* It holds that

$$(23) \quad \frac{d}{ds} T_k(s) = \frac{1 + 2\mathbf{c}_s \cot(\alpha)^2 + \cot(\alpha)^4 + k \left( 2 + 2\mathbf{c}_s \cot(\alpha)^2 \right)}{1 + 2\mathbf{c}_s \cot(\alpha)^2 + \cot(\alpha)^4}.$$

It is clear that the denominator of the above fraction is never vanishing on  $[0, \pi]$  if  $\alpha < \pi/4$ . On the other hand, the numerator, as a function of  $s$ , reaches its minimum

at  $s = \pi$ , where it is equal to  $(\cot(\alpha)^2 - 1)^2 - k(2\cot(\alpha)^2 - 1)$ , and then the conclusion follows easily.  $\square$

As a consequence, we obtain the following important corollary.

**COROLLARY 1.** *Let  $k$  be an integer satisfying  $1 \leq k \leq \mathcal{N}_{\text{mon}}$ . If an extremal trajectory is switching at time  $T = k\pi$ , then the length  $s$  of the first bang arc satisfies  $s = \pi$ .*

Since for  $\alpha$  small  $k_M \leq \mathcal{N}_{\text{mon}}$ , then, for  $T = k\pi$ , where  $k$  is a positive integer such that  $k \leq [\pi/(2\alpha)]$ , we have that all of the extremal trajectories switch exactly  $k$  times (except the trajectories with length of the first switching equal to  $\pi$  that switch  $k - 1$  times). Therefore, the extremal front  $\mathcal{F}(\alpha, k\pi)$  is described by the next proposition.

**PROPOSITION 2.** *Let  $k$  be a positive integer such that  $1 \leq k \leq [\pi/(2\alpha)]$ . Then, if  $\alpha$  is small enough, we have*

$$(24) \quad \mathcal{F}(\alpha, k\pi) = \{\mathcal{E}^+(\alpha, k, s), \quad s \in (0, \pi]\} \cup \{\mathcal{E}^-(\alpha, k, s), \quad s \in (0, \pi]\}, \quad \text{where :}$$

$$(25) \quad \mathcal{E}^+(\alpha, k, s) := \begin{cases} e^{(k\pi - (k-1)v(s) - s)X_-} e^{\frac{k-1}{2}\Theta(v(s))Z_-} e^{(v(s))} e^{sX_+} N & \text{for } k \text{ odd,} \\ e^{(k(\pi - v(s)) - s)X_+} e^{\frac{k}{2}\Theta(v(s))Z_-} e^{(v(s))} e^{sX_+} N & \text{for } k \text{ even.} \end{cases}$$

The expression for  $\mathcal{E}^-$  is the same as the expression for  $\mathcal{E}^+$  after exchanging the subscripts  $+$  and  $-$ . As a consequence,  $\mathcal{E}^{-\varepsilon} = \Pi_{x_3} \mathcal{E}^\varepsilon$ , where  $\Pi_{x_3}$  is the orthogonal symmetry with respect to the  $x_3$ -axis.

*Remark 5.* Notice that  $\mathcal{E}^\varepsilon(\alpha, k, 0) = \mathcal{E}^{-\varepsilon}(\alpha, k, \pi)$ ,  $\varepsilon = \pm$ , so that  $\mathcal{F}$  is described by a continuous closed curve.

**3.2. Description of the extremal front  $\mathcal{F}(\alpha, k_M\pi)$  and consequences.** As sketched in the introduction, we must describe the optimal synthesis on  $S^2$  deprived of a neighborhood of the south pole. For that purpose, we will provide the precise asymptotics of  $\mathcal{F}(\alpha, k_M\pi)$ , as  $\alpha$  tends to zero, and derive, from its topological nature, the minimum time front at time  $k_M\pi$ .

From now on, for simplicity, we drop the dependence of  $\mathcal{E}^\varepsilon$  on  $k_M$ ; i.e., we set  $\mathcal{E}^\varepsilon(\alpha, s) := \mathcal{E}^\varepsilon(\alpha, k_M, s)$ , and we assume that  $k_M$  is odd.

In the following, it will be useful to think of  $\alpha$  and  $r$  as two independent variables. For this purpose, define

$$\begin{aligned} \psi(\alpha, r, s) &:= \left( \frac{\pi}{2\alpha} - r \right) (\pi - v(s)) + v(s) - s, \\ \theta(\alpha, r, s) &:= \left( \frac{\pi}{4\alpha} - \frac{1+r}{2} \right) \Theta(v(s)), \\ \chi^\varepsilon(\alpha, r, s) &:= e^{\psi(\alpha, r, s)X_-} e^{\theta(\alpha, r, s)Z_-} e^{(v(s))} e^{sX_\varepsilon} N. \end{aligned}$$

It is clear from (25) that

$$\mathcal{E}^\varepsilon(\alpha, s) = \chi^\varepsilon(\alpha, r(\alpha), s).$$

The following result is the key point in order to describe the extremal front at time  $k_M\pi$ .

**LEMMA 2.** *There exists  $\alpha_0 > 0$  such that the function  $\chi^\varepsilon$ ,  $\varepsilon = \pm$ , defined above, is real-analytic for  $(r, s, \alpha) \in \mathbb{R}^2 \times I$ , where  $I = (-\alpha_0, \alpha_0)$ . Moreover, it admits a convergent power series*

$$(26) \quad \chi^\varepsilon(\alpha, r, s) = \sum_{l \geq 0} f_l^\varepsilon(s, r) \alpha^l,$$

where the  $f_l^\varepsilon(s, r)$  are real-analytic functions of  $(s, r) \in \mathbb{R}^2$ ,  $2\pi$ -periodic in  $s$  (therefore they are bounded over  $\mathbb{R} \times [0, 1]$ ).

As a consequence, the extremal front  $\mathcal{F}(\alpha, k_M\pi)$ , which is a continuous closed curve, is piecewise analytic with discontinuities at  $s = 0, \pi$  for derivatives of order greater than or equal to one.

*Proof of Lemma 2.* We will prove the proposition only for  $\chi^+$ . Since  $\chi^+$  is  $2\pi$ -periodic in  $s$  and  $r$  enters in an affine way in  $\psi$  and  $\theta$ , the real issue of analyticity revolves around the variable  $\alpha$ . First of all, it is clear that  $v(s)$  is actually a real-analytic function for  $(s, \alpha) \in \mathbb{R} \times I$ , where  $I = (-\alpha_0, \alpha_0)$ , with  $\alpha_0 > 0$  small enough. Therefore, one has only to prove the real-analyticity of  $\tilde{\psi}(\alpha, s) := \frac{v(s) - \pi}{\alpha}$  and  $\frac{\beta(s, \alpha)}{\alpha}$ , where  $\beta(s, \alpha) := \Theta(v(s))$ , for  $(s, \alpha) \in \mathbb{R} \times I$ , where  $I = (-\alpha_0, \alpha_0)$ , for some  $\alpha_0 > 0$ .

Note that

$$\tilde{\psi}(\alpha, s) = \frac{2}{\alpha} \arctan(\mathbf{s}_\alpha^2 \mu(s)), \quad \mu(s) := \frac{\mathbf{s}_s}{\mathbf{c}_\alpha^2 + \mathbf{s}_\alpha^2 \mathbf{c}_s}.$$

The function  $\mu$  is real-analytic for  $(s, \alpha) \in \mathbb{R} \times I$ , with  $I$  an open neighborhood of zero, and thus uniformly bounded over  $\mathbb{R} \times I$ . In addition,  $\arctan(\cdot)$  is real-analytic in a neighborhood of zero. Hence the conclusion for  $\tilde{\psi}(\alpha, s)$ .

As for  $\frac{\beta(s, \alpha)}{\alpha}$ , first rewrite (19) as

$$\cos(\beta(s)) = 1 - G(s, \alpha),$$

with

$$(27) \quad G(s, \alpha) := 2\mathbf{s}_\alpha^2 \left[ 1 + \frac{\mathbf{c}_\alpha^2 \mathbf{s}_\alpha^2 \mu^2(s)}{1 + \mathbf{s}_\alpha^4 \mu^2(s)} + 2\mathbf{s}_\alpha^2 \left( 1 + \frac{\mathbf{c}_\alpha^2 \mathbf{s}_\alpha^2 \mu^2(s)}{1 + \mathbf{s}_\alpha^4 \mu^2(s)} \right)^2 \right].$$

We first need to determine a convergent power series for  $\beta$  from the expression

$$(28) \quad \beta = \arccos(1 - G).$$

Note that  $|G(s, \alpha)| \leq 5\alpha^2$  for  $\alpha$  small enough. We first expand  $\arccos(1 - G)$  in a power series in  $G$ . Starting from the power series

$$(1 - t)^{-1/2} = 1 + \sum_{m \geq 1} s_m t^m,$$

with radius of convergence equal to 1 we get

$$\frac{d}{dG}(\arccos(1 - G)) = -\frac{1}{\sqrt{2G}} \frac{1}{\sqrt{1 - G/2}},$$

and, after simple integration,

$$(29) \quad \arccos(1 - G) = -\sqrt{2G} \left( 1 + \sum_{m \geq 1} \frac{s_m}{2^{m+1}(m + 1/2)} G^m \right).$$

Finally, from (27),  $G$  can be written as  $2\mathbf{s}_\alpha^2(1 + \mathbf{s}_\alpha^2 H(s, \alpha))$ , with  $H(s, \alpha)$  uniformly bounded by 3. Then

$$(30) \quad \sqrt{2G(s, \alpha)} = 2\mathbf{s}_\alpha(1 + \mathbf{s}_\alpha^2 H(s, \alpha))^{1/2}.$$

Gathering (28)–(30), we get the real-analyticity of  $\frac{\beta(s, \alpha)}{\alpha}$  for  $(s, \alpha) \in \mathbb{R} \times I$ , where  $I = (-\alpha_0, \alpha_0)$ , for some  $\alpha_0 > 0$  small enough.  $\square$

We next compute  $f_0^\varepsilon, f_1^\varepsilon$ , and  $f_2^\varepsilon$  and obtain the following proposition.

**PROPOSITION 3.** *For  $\alpha$  small enough, the function  $\chi^\varepsilon$ ,  $\varepsilon = \pm$ , defined above and its derivative with respect to  $s$  have the following expansion:*

$$(31) \quad \chi^\varepsilon(\alpha, r, s) = f_0^\varepsilon(s, r) + f_1^\varepsilon(s, r)\alpha + f_2^\varepsilon(s, r)\alpha^2 + \mathcal{O}(\alpha^3),$$

$$(32) \quad \frac{\partial}{\partial s} \chi^\varepsilon(\alpha, r, s) = \frac{\partial}{\partial s} f_0^\varepsilon(s, r) + \frac{\partial}{\partial s} f_1^\varepsilon(s, r)\alpha + \frac{\partial}{\partial s} f_2^\varepsilon(s, r)\alpha^2 + \mathcal{O}(\alpha^3),$$

where  $f_l^\varepsilon$ ,  $l = 0, 1, 2$ , are defined as in (9) and  $|\mathcal{O}(\alpha^3)| \leq C|\alpha^3|$ , with the constant  $C$  independent of  $s \in \mathbb{R}$  and  $r \in [0, 1)$ .

*Proof of Proposition 3.* We will prove the proposition only for  $\chi^+$ . To proceed, we list the expansions of the form (31) for several quantities, obtained after elementary computations:

$$(33) \quad \psi(\alpha, r, s) = \pi - s - \pi \mathbf{s}_s \alpha + 2(1 - r) \mathbf{s}_s \alpha^2 + \mathcal{O}(\alpha^3),$$

$$(34) \quad \theta(\alpha, r, s) = \pi - 2\alpha(1 + r) + \frac{\pi \mathbf{s}_s^2}{2} \alpha^2 + \mathcal{O}(\alpha^3),$$

$$(35) \quad Z_-(v(s)) = \begin{pmatrix} 0 & -\alpha \mathbf{s}_s & 1 - \frac{\mathbf{s}_s^2}{2} \alpha^2 \\ \alpha \mathbf{s}_s & 0 & 0 \\ -1 + \frac{\mathbf{s}_s^2}{2} \alpha^2 & 0 & 0 \end{pmatrix} + \mathcal{O}(\alpha^3).$$

Using (33) and (34), we get that

$$(36) \quad \sin(\psi(\alpha, r, s)) = \mathbf{s}_s + \pi \mathbf{s}_s \mathbf{c}_s \alpha - \left( 2(1 - r) \mathbf{s}_s \mathbf{c}_s + \frac{\pi}{2} \mathbf{s}_s^3 \right) \alpha^2 + \mathcal{O}(\alpha^3),$$

$$(37) \quad \cos(\psi(\alpha, r, s)) = \mathbf{c}_s - \pi \mathbf{s}_s^2 \alpha + \left( 2(1 - r) \mathbf{s}_s^2 - \frac{\pi^2}{2} \mathbf{s}_s^2 \mathbf{c}_s \right) \alpha^2 + \mathcal{O}(\alpha^3),$$

$$(38) \quad \sin(\theta(\alpha, r, s)) = 2\alpha(1 + r) - \frac{\pi \mathbf{s}_s^2}{2} \alpha^2 + \mathcal{O}(\alpha^3),$$

$$(39) \quad \cos(\theta(\alpha, r, s)) = -1 + 2\alpha^2(1 + r)^2 + \mathcal{O}(\alpha^3).$$

Using (17), (36), and (37), we obtain

$$(40) \quad e^{\psi(\alpha, r, s) X_-} = \begin{pmatrix} -\mathbf{c}_s + \pi \mathbf{s}_s^2 \alpha & -\mathbf{s}_s - \pi \mathbf{c}_s \mathbf{s}_s \alpha & -(1 + \mathbf{c}_s) \alpha \\ \mathbf{s}_s + \pi \mathbf{s}_s \mathbf{c}_s \alpha & \mathbf{c}_s + \pi \mathbf{s}_s^2 \alpha & \mathbf{s}_s \alpha \\ -(1 + \mathbf{c}_s) \alpha & -\mathbf{s}_s \alpha & 1 \end{pmatrix} + \mathcal{R}(s) \alpha^2 + \mathcal{O}(\alpha^3),$$

where

$$\mathcal{R}(s) = \begin{pmatrix} \mathbf{c}_s + \frac{\pi^2}{2} \mathbf{c}_s \mathbf{s}_s^2 + 1 - 2(1 + r) \mathbf{s}_s^2 & \frac{\mathbf{s}_s}{2} + 2\mathbf{c}_s \mathbf{s}_s(1 + r) + \frac{\pi^2}{2} \mathbf{s}_s^3 & \pi \mathbf{s}_s^2 \\ -\frac{\mathbf{s}_s}{2} - 2\mathbf{c}_s \mathbf{s}_s(1 + r) - \frac{\pi^2}{2} \mathbf{s}_s^3 & -2(1 + r) \mathbf{s}_s^2 + \frac{\pi^2}{2} \mathbf{c}_s \mathbf{s}_s^2 & \pi \mathbf{s}_s \mathbf{c}_s \\ \pi \mathbf{s}_s^2 & -\pi \mathbf{s}_s \mathbf{c}_s & -1 - \mathbf{c}_s \end{pmatrix},$$

and, using (17), we have

$$(41) \quad e^{s X_+} N = \begin{pmatrix} \mathbf{s}_\alpha \mathbf{c}_\alpha (1 - \mathbf{c}_s) \\ -\mathbf{s}_\alpha \mathbf{s}_s \\ 1 - \mathbf{s}_\alpha^2 (1 - \mathbf{c}_s) \end{pmatrix} = \begin{pmatrix} \alpha(1 - \mathbf{c}_s) \\ -\alpha \mathbf{s}_s \\ 1 - \alpha^2(1 - \mathbf{c}_s) \end{pmatrix} + \mathcal{O}(\alpha^3).$$

An easy computation yields

$$(42) \quad Z_-^2(v(s)) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -\alpha^2 \mathbf{s}_s^2 & \alpha \mathbf{s}_s \\ 0 & \alpha \mathbf{s}_s & -1 + \alpha^2 \mathbf{s}_s^2 \end{pmatrix} + \mathcal{O}(\alpha^3).$$

Using (17), (38), (39), (41), and the previous equation, we get

$$(43) \quad e^{\theta(\alpha, r, s)Z_-(v(s))} e^{sX_+} N = \begin{pmatrix} \alpha(1 + 2r + \mathbf{c}_s) - \alpha^2 \frac{\pi}{2} \mathbf{s}_s^2 \\ \alpha \mathbf{s}_s \\ -1 + \alpha^2 (1 + 2r + 2r^2 + \mathbf{c}_s + 2r \mathbf{c}_s) \end{pmatrix} + \mathcal{O}(\alpha^3).$$

Applying  $e^{\psi(s)X_-}$  to the previous equation and using (40), we finally get (31) for  $\varepsilon = +$ . The expression of the derivative (32) is then an immediate consequence of the analyticity of the function  $\chi^+$ .

The results for  $\chi^-$  and  $\frac{\partial}{\partial s}\chi^-$  are obtained similarly together with (16).  $\square$

Since the quantities of the form  $\mathcal{O}(\alpha^3)$  in Proposition 3 satisfy  $|\mathcal{O}(\alpha^3)| \leq C|\alpha^3|$  for some  $C$  independent of  $r$ , the expressions (10)–(11) are straightforward consequences. Hence the shape of the extremal front at time  $T = k_M\pi$  is known for  $\alpha$  small. In particular its image with respect to the map  $M_\alpha$  defined in section 1 is approximated, in the  $\mathcal{C}^1$  sense, by a circle of radius  $2r(\alpha)$  centered at the origin.

We finally note that, for  $k_M$  even, Lemma 2 is still valid, while, with computations similar to those made in the proof of Proposition 3, it is easy to see that the formulas for  $f_k^\varepsilon$ ,  $k = 0, 1, 2$ , simply differ, with respect to (9), for the sign of the first two components.

*Remark 6.* Repeating the previous computations, we also obtain series expansions for  $\mathcal{E}^\varepsilon(s, k_M - 1, \alpha)$  and  $\frac{\partial}{\partial s}\mathcal{E}^\varepsilon(s, k_M - 1, \alpha)$ . Indeed, we just have to replace  $r$  by  $1 + r$ . In that case the shape of the extremal front  $\mathcal{F}(\alpha, (k_M - 1)\pi)$ , after applying the map  $M_\alpha$ , is approximated, in the  $\mathcal{C}^1$  sense, by a circle of radius  $2(1 + r(\alpha))$  centered at the origin.

**4. Case  $r(\alpha) = \bar{r} \in (0, 1)$ .** In this section, we study the case in which  $\alpha$  tends to zero, with  $r(\alpha) = \bar{r}$ , for a constant  $\bar{r} \in (0, 1)$ . More precisely we consider the decreasing sequence  $\alpha_k = \frac{\pi}{2(k + \bar{r})}$  for  $k \geq 1$ . We first describe the minimum time front at  $T = k_M\pi$ , then we identify and study the candidate for the limit synthesis, and finally we prove Theorem 1.

**4.1. Description of the minimum time front at  $T = k_M\pi$ .** The purpose of the subsection is to prove the following proposition.

**PROPOSITION 4.** *Fix  $\delta > 0$  small. For  $\alpha$  small enough, with  $r(\alpha) > \delta$ , the extremal front  $\mathcal{F}(\alpha, k_M\pi)$  is homeomorphic to a circle. As a consequence, the switching curves defined inductively in (7) are optimal up to  $k = k_M$ , and  $OF(\alpha, k_M\pi)$ , the minimum time front at time  $k_M\pi$ , coincides with  $\mathcal{F}(\alpha, k_M\pi)$ .*

*Proof of Proposition 4.* From Proposition 3 we get that the extremal front  $\mathcal{F}(\alpha, k_M\pi)$  is the union of two arcs  $\mathcal{E}^+(\alpha, s)$ ,  $s \in [0, \pi]$  and  $\mathcal{E}^-(\alpha, s)$ ,  $s \in [0, \pi]$  so that, for  $\varepsilon = \pm$  and  $s \in [0, \pi]$ ,

$$(44) \quad \mathcal{E}^\varepsilon(\alpha, s) = \begin{pmatrix} -2r(\alpha)\varepsilon\alpha\mathbf{c}_s \\ 2r(\alpha)\varepsilon\alpha\mathbf{s}_s \\ -1 \end{pmatrix} + \mathcal{O}(\alpha^2),$$

and

$$(45) \quad \frac{\partial}{\partial s} \mathcal{E}^\varepsilon(\alpha, s) = 2r(\alpha) \varepsilon \alpha \begin{pmatrix} \mathbf{s}_s \\ \mathbf{c}_s \\ 0 \end{pmatrix} + \mathcal{O}(\alpha^2).$$

Moreover, at  $s = 0$  and  $s = \pi$ , the derivatives of  $\mathcal{E}^\varepsilon(\alpha, s)$  are only one-sided, i.e., as  $s > 0$  tends to zero and  $s < \pi$  tends to  $\pi$ . By a trivial continuity argument, one can parametrize  $\mathcal{F}(\alpha, k_M \pi)$  as a closed continuous curve  $\gamma$  defined on  $[0, 2\pi]$  so that  $\gamma(s) = \mathcal{E}^+(\alpha, s)$  for  $s \in (0, \pi]$  and  $\gamma(s) = \mathcal{E}^-(\alpha, s - \pi)$  for  $s \in (\pi, 2\pi]$ . Moreover, with the previous computations, it is immediate that  $\gamma$  is in fact piecewise  $C^1$  with possible discontinuity jumps for  $\frac{d}{ds}\gamma$  at  $s = 0$  and  $s = \pi$ .

Since the curve  $\gamma$  is in a neighborhood of the south pole of size proportional to  $\alpha$  (thanks to (44)), it is enough to prove that the orthogonal projection  $\gamma_1$  of  $\gamma$  on the  $(x_1, x_2)$ -plane is homeomorphic to the circle  $e^{is}$ ,  $s \in [0, 2\pi]$ . Using (44), we see that  $\|\gamma_1(s)\| = 2r(\alpha)\alpha + \mathcal{O}(\alpha^2)$  on  $[0, 2\pi]$ , which implies that the continuous function  $\|\gamma_1(s)\|$  is always strictly positive for  $\alpha$  small enough. We can therefore parametrize  $\gamma_1$  using polar coordinates  $(\rho, \beta)$ , i.e., for  $s \in [0, 2\pi]$ ,

$$\gamma_1(s) = \rho(s)e^{i\beta(s)},$$

where  $\rho(\cdot) := \|\gamma_1(\cdot)\|$  and the function  $\beta(\cdot)$  are defined on  $[0, 2\pi]$ , continuous, and piecewise  $C^1$ , with possible jumps of discontinuity for their derivatives at  $s = 0$  and  $s = \pi$ .

In addition  $\rho(0) = \rho(2\pi)$ ,  $\beta(0) \equiv \beta(2\pi) \equiv \pi \pmod{2\pi}$ , and, from (44),  $\beta(s) = \pi - s + \mathcal{O}(\alpha)$ . To prove Proposition 4, it suffices now to prove that  $\beta$  is a monotone bijection from  $[0, 2\pi]$  to  $[-\pi, \pi]$ . The latter simply results from (45). Indeed, from that equation, we get that  $\frac{d}{ds}\beta(s) = -1 + \mathcal{O}(\alpha)$ , where  $\beta$  is differentiable and the one-sided derivatives at  $s = 0$  and  $s = \pi$  verify the same equation. We deduce that  $\beta$  is strictly decreasing for  $\alpha$  small enough.

We next show that  $OF(\alpha, k_M \pi)$ , the minimum time front at time  $k_M \pi$ , coincides with  $\mathcal{F}(\alpha, k_M \pi)$ . By the results of [9], we first notice that any time minimal trajectory starting at the north pole reaches the south pole in time  $T > k_M \pi$ . Therefore  $OF(\alpha, k_M \pi)$  is not empty and is included in  $\mathcal{F}(\alpha, k_M \pi)$  according to the PMP. According to Theorem 27 of [8],  $OF(\alpha, k_M \pi)$  is a one-dimensional piecewise  $C^1$  compact embedded submanifold of  $S^2$ . By an easy topological argument, we deduce from the above that  $OF(\alpha, k_M \pi)$  coincides with  $\mathcal{F}(\alpha, k_M \pi)$ .  $\square$

*Remark 7.* Thanks to Remark 6, and with arguments similar to those of the previous proof, one can prove that  $\mathcal{F}(\alpha, (k_M - 1)\pi)$  is optimal for  $\alpha$  small enough, with no assumptions on the remainder  $r$ .

**4.2. Optimal synthesis for the linear pendulum control problem.** Recall that  $M_\alpha : \mathbb{R}^3 \rightarrow \mathbb{R}^2$  is the composition of the projection  $(x_1, x_2, x_3) \mapsto (x_1, x_2)$  followed by the dilation by  $1/\alpha$ . With the results of the previous subsection, it is clear that the original control problem on  $S^2$  can be reduced, near the south pole, to a planar control problem on the neighborhood of the south pole delimited by  $\widetilde{OF}(\alpha, k_M \pi) := M_\alpha(OF(\alpha, k_M \pi))$  along  $(\widetilde{\Sigma})_\alpha$ , the control system obtained as the image of  $(\Sigma)_\alpha$  by  $M_\alpha$ , i.e.,

$$(46) \quad (\widetilde{\Sigma})_\alpha : \begin{cases} \dot{z}_1 = -\cos(\alpha)z_2, \\ \dot{z}_2 = \cos(\alpha)z_1 + u \frac{\sin(\alpha)}{\alpha} \sqrt{1 - (\alpha z_1)^2 - (\alpha z_2)^2}, \end{cases} \quad (z_1, z_2) \in \mathbb{R}^2, \quad |u| \leq 1.$$

It is therefore natural to conjecture (simply set  $\alpha = 0$  in  $\widetilde{OF}(\alpha, k_M\pi)$  and  $(\widetilde{\Sigma})_\alpha$ ) that the limit synthesis should be that of connecting the circle of radius  $2r(\alpha)$ ,  $C(0, 2r(\alpha))$ , to every point of the disk  $B(0, 2r(\alpha))$  along the control system  $(Pen)$  given by (12), which we rewrite as

$$(47) \quad (Pen) \quad \dot{z} = A_0 z + u b_0, \quad \text{with} \quad A_0 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad b_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

where  $z \in \mathbb{R}^2$  and  $u \in [-1, 1]$ . The control system  $(Pen)$  corresponds to a linear pendulum with a forcing term.

Theorem 1 simply states that the conjecture is correct, and, as a first step for an argument, we describe, in more detail in this subsection, the conjectured limit synthesis. Hence we focus on the following problem.

(P). *Given fixed  $\rho \in ]0, 2]$ , for any given  $\bar{y} \in B(0, \rho)$  find a time optimal trajectory connecting the circle of radius  $\rho$  centered at the origin to  $\bar{y}$  along the control system  $(Pen)$ .*

*Remark 8.* The problem of computing the TOS for the linear pendulum, taking the origin as a source, is studied in any textbook of optimal control. Here as the source we take the circle of radius  $\rho$  centered at the origin, which is a level set of the Hamiltonian  $H = \frac{1}{2}(z_1^2 + z_2^2)$  associated to the uncontrolled system.

It is easy to see that the solutions of problem (P) must be bang-bang trajectories. Indeed since  $(Pen)$  is a bidimensional linear control system it is well known that this property is guaranteed by the Kalman controllability condition  $\det(b_0, A_0 b_0) \neq 0$ , which is satisfied by  $(Pen)$ . To determine the TOS, we first look for the switching curves. We know that every extremal trajectory for the problem (P) must satisfy the transversality condition of the PMP stated in section 2.1. Here the source manifold is the circle  $C(0, \rho)$ , and the transversality condition essentially translates into the property that the vector  $\lambda(0)$  (that, without loss of generality, we will assume unitary) is proportional to  $z(0) \in C(0, \rho)$  (identifying the cotangent space with the plane  $\mathbb{R}^2$ ). To determine completely  $\lambda(0)$ , it is enough to observe that a necessary condition for  $z(\cdot)$  to be optimal is that  $\dot{z}(0)$  points inside the disk  $B(0, \rho)$ ; i.e., if we denote by  $u_{opt}$  the corresponding control, then

$$\langle z(0), \dot{z}(0) \rangle \leq 0 \iff \langle z(0), A_0 z(0) + u_{opt} b_0 \rangle \leq 0 \iff \langle z(0), u_{opt} b_0 \rangle \leq 0.$$

Therefore,  $u_{opt} = -\text{sgn}\langle z(0), b_0 \rangle$ . On the other hand, from the maximality condition of the PMP, we have  $u_{opt} = \text{sgn}\langle \lambda(0), b_0 \rangle$ , and, therefore, one can define  $\lambda(0) := -z(0)/\rho$ . Finally  $u_{opt} = -\text{sgn}(z_2(0))$  (except at the points  $\pm(\rho, 0)$ ), while the switching time  $t_{sw}$  must satisfy the condition  $\langle \lambda(t_{sw}), b_0 \rangle = \lambda_2(t_{sw}) = 0$ .

Consider now the adjoint system

$$(48) \quad \begin{cases} \dot{\lambda}_1 = -\lambda_2, \\ \dot{\lambda}_2 = \lambda_1. \end{cases}$$

If we identify  $\mathbb{R}^2$  with the complex plane, so that  $z = z_1 + iz_2$  and  $\lambda = \lambda_1 + i\lambda_2$ , then (47) and (48) become

$$\dot{z} = i(z + u) \quad \text{and} \quad \dot{\lambda} = i\lambda.$$

Moreover we can set  $z(0) = -\rho e^{-i\theta}$  and  $\lambda(0) = e^{-i\theta}$  for some  $\theta \in [0, 2\pi[$ , and the corresponding solutions are

$$\begin{cases} z(t) = (z(0) + u_{opt})e^{it} - u_{opt} = -\rho e^{i(t-\theta)} + u_{opt}(e^{it} - 1), \\ \lambda(t) = \lambda(0)e^{it} = e^{i(t-\theta)}. \end{cases}$$



The switching curves are determined by the relation  $t_{sw} \equiv \theta \pmod{\pi}$ , and this allows one to conclude that the switching curves are the following two semicircles of radius 1:

$$\begin{cases} z(\theta) = 1 - \rho - e^{i\theta} & \theta \in [0, \pi[, \\ z(\theta) = \rho - 1 - e^{i\theta} & \theta \in [\pi, 2\pi[. \end{cases}$$

These switching curves cannot be optimal for  $\rho < 2$  since they are not locally optimal, as can be easily checked using the definition given in section 2.2. We conclude that the optimal trajectories are bang arcs and the corresponding control depends on the sign of the component  $z_2(0)$  of the starting point.

To conclude the description of the synthesis, it is enough to determine the cut locus, i.e., the set of points that are reached by two or more optimal trajectories at the same time. Assume that  $z \in \mathbb{C}$  belongs to the cut locus. Then there exist  $s \in [0, \pi)$ ,  $s' \in [\pi, 2\pi)$  and  $t$  such that

$$(49) \quad \begin{cases} z = -\rho e^{i(t-s)} + 1 - e^{it}, \\ z = -\rho e^{i(t-s')} - 1 + e^{it}. \end{cases}$$

Therefore  $|z - 1 + e^{it}| = |z + 1 - e^{it}| = \rho$ . In particular, denoting by  $\bar{z}$  the complex conjugate to  $z$ , we have

$$(50) \quad \begin{aligned} (z - 1 + e^{it})(\bar{z} - 1 + e^{-it}) - (z + 1 - e^{it})(\bar{z} + 1 - e^{-it}) \\ = -4z_1 + 4z_1 \cos t + 4z_2 \sin t = 0, \end{aligned}$$

$$(51) \quad \begin{aligned} (z - 1 + e^{it})(\bar{z} - 1 + e^{-it}) + (z + 1 - e^{it})(\bar{z} + 1 - e^{-it}) \\ = 2z_1^2 + 2z_2^2 + 4 - 4 \cos t = 2\rho^2. \end{aligned}$$

From (50) we have that  $\cos t = \frac{z_1^2 - z_2^2}{z_1^2 + z_2^2}$ , and, substituting in (51), we find that  $z$  must satisfy the equation

$$(52) \quad z_1^4 + z_2^4 + 2z_1^2 z_2^2 - \rho^2 z_1^2 + (4 - \rho^2) z_2^2 = 0.$$

The previous computations show that the cut locus is a subset of the set of points belonging to the locus defined by (52). Actually it is easy to see that this is the proper subset obtained with the additional condition  $z_1 z_2 \geq 0$  that corresponds to  $t \leq \pi$ . The precise shape of the optimal synthesis, which is now clear, is portrayed in Figure 5 for a particular value of  $\rho < 2$ . Notice that, from the previous computations, we have  $\rho e^{is'} = \rho e^{is} + 2 - 2e^{it}$ , and, since  $\rho e^{is'} + \rho e^{is} = 2\rho e^{is'} - 2 + 2e^{it}$  and  $\rho e^{is'} - \rho e^{is} = 2 - 2e^{it}$  are orthogonal in the complex plane, we find easily the following equation:

$$(2 - \rho \cos s')(\cos t - 1) - \rho \sin s' \sin t = 0.$$

Consequently, for  $t \in [0, 2\pi[$  and  $s' \in [\pi, 2\pi[$ , one has along the overlap curve

$$(53) \quad t = t(s') = -2 \arctan \frac{\rho \sin s'}{2 - \rho \cos s'}.$$

This expression will be useful in the following. Also, notice that by combining (49) and (53) one easily finds a parametrization of the overlap curve in terms of  $s'$  and that in an analogous way it is possible to parametrize it by means of  $s$ . From now on

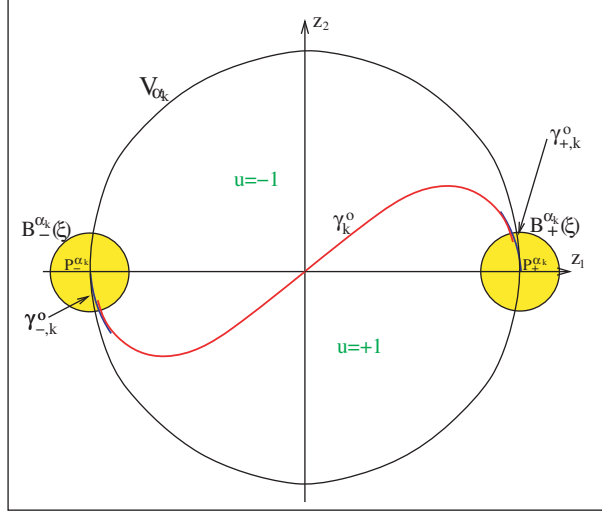


FIG. 8. Propositions 5 and 6.

we will denote by  $\gamma_{pen}^o(\cdot)$  the parametrization of the overlap curve with respect to the parameter  $s$ .

*Remark 9.* If  $\rho = 2$ , the previous reasoning does not apply, and indeed the synthesis is different. In this case the overlap curve coincides with the switching curves and with the trajectories reaching the origin corresponding to  $u = \pm 1$ . A simple way to prove this fact is to study the optimal synthesis starting from the origin with vector fields with opposite signs and to observe that the extremal front at time  $\pi$  is a circle of radius 2.

**4.3. Proof of Theorem 1.** The proof of Theorem 1 is divided in two parts. Roughly speaking, defining  $P_\varepsilon^\alpha := M_\alpha(C_{k_M}^\varepsilon(0))$  for  $\varepsilon = \pm$ , we will look separately at the shape of the synthesis far from  $P_\varepsilon^\alpha$  and inside neighborhoods of  $P_\varepsilon^\alpha$ ,  $\varepsilon = \pm$ . Let us call  $V_\alpha$  the image with respect to  $M_\alpha$  of the neighborhood of the south pole enclosed by  $OF(\alpha, k_M\pi)$  and  $B_\varepsilon^{\alpha_k}(\xi)$  the ball of center  $P_\varepsilon^\alpha$  and radius  $\xi$ .

Then the previous cases correspond to the following two propositions whose meaning is clarified by Figure 8.

**PROPOSITION 5.** Let  $\bar{r} \in (0, 1)$  and  $\alpha_k := \frac{\pi}{2(k+\bar{r})}$ . Then for any  $\xi > 0$  there exist a positive integer  $\bar{k}$  and a compact interval  $I \subset (0, \pi)$  such that it is possible to find a curve  $\gamma_k^o$ , defined on  $I$  for  $k \geq \bar{k}$ , verifying the following:  $\gamma_k^o$  divides  $V_{\alpha_k} \setminus (B_+^{\alpha_k}(\xi) \cup B_-^{\alpha_k}(\xi))$  in two connected components, such that above  $\gamma_k^o$  the optimal feedback associated to the synthesis for  $\alpha = \alpha_k$  takes the value  $-1$ , and below  $\gamma_k^o$  it is equal to  $1$ , and in particular  $\gamma_k^o$  is an overlap curve for  $\alpha = \alpha_k$ . Moreover,  $\gamma_k^o$  converges to  $\gamma_{pen}^o$  in the  $C^0$  topology of  $I$ .

**PROPOSITION 6.** Consider the notations defined above. Then there exist  $\xi > 0$ ,  $\tau_\varepsilon$ ,  $\varepsilon = \pm$ , with  $0 < \tau_- < \tau_+ < \pi$  and a positive integer  $\bar{k}$  such that, for every  $k \geq \bar{k}$ , it is possible to find two curves  $\gamma_{\varepsilon,k}^o$  and  $\gamma_{+\varepsilon,k}^o$ , defined respectively on  $[0, \tau_-]$  and  $[\tau_+, \pi]$ , verifying the following:  $\gamma_{\varepsilon,k}^o$  divides  $V_{\alpha_k} \cap B_\varepsilon^{\alpha_k}(\xi)$  in two connected components, such that above  $\gamma_{\varepsilon,k}^o$  the optimal feedback associated to the synthesis for  $\alpha = \alpha_k$  takes the value  $-1$ , and below  $\gamma_{\varepsilon,k}^o$  it is equal to  $1$ , and in particular the  $\gamma_{\varepsilon,k}^o$  are overlap curves for  $\alpha = \alpha_k$ . Moreover,  $\gamma_{\varepsilon,k}^o$  and  $\gamma_{+\varepsilon,k}^o$  converge to  $\gamma_{pen}^o$  in the  $C^0$  topology, respectively, of  $[0, \tau_-]$  and  $[\tau_+, \pi]$ .

The choice of studying the synthesis separately in neighborhoods of  $P_\varepsilon^\alpha$  and far from  $P_\varepsilon^\alpha$  is justified by the fact that the proofs of the previous propositions rely on different implicit function arguments.

It is clear that, by combining Proposition 5, for an appropriate choice of  $\xi$ , with Proposition 6, one almost completes the proof of Theorem 1. We will not prove explicitly that the convergence of  $\gamma_k^o$  to  $\gamma_{pen}^o$  with respect to the parameter  $\bar{r}$  is uniform in any closed interval of  $(0, 1)$ . As explained in Remark 10, this can be done with the same methods used in the proofs of Propositions 5 and 6.

We will therefore provide only the complete proofs of the propositions. For this purpose, the first step consists in checking whether the switching curves  $C_{k_M}^\varepsilon$ ,  $\varepsilon = \pm 1$ , are optimal or not. In that regard and similarly to the case of the linear pendulum, we have the following result.

**LEMMA 3.** *Let  $\bar{r} \in ]0, 1[$ . Then, if  $\alpha$  is small enough and  $r(\alpha) = \bar{r}$ , the switching curve  $C_{k_M}^\varepsilon$  is nowhere locally optimal; i.e., all of the extremal trajectories switching on  $C_{k_M}^\varepsilon$  lose optimality before reaching it.*

*Proof of Lemma 3.* For simplicity we define  $S(s) := C_{k_M}^+(s)$ , and we assume  $k_M$  odd. As in the proof of Proposition 3, we get the following asymptotic expansions, after applying the map  $M_\alpha$ :

$$(54) \quad S(s) = \begin{pmatrix} 2r - 1 + \mathbf{c}_s \\ \mathbf{s}_s \end{pmatrix} + \mathcal{O}(\alpha), \quad S(0) = \begin{pmatrix} 2r + \mathcal{O}(\alpha) \\ 0 \end{pmatrix},$$

$$(55) \quad S'(s) = \begin{pmatrix} -\mathbf{s}_s \\ \mathbf{c}_s \end{pmatrix} + \mathcal{O}(\alpha), \quad S'(0) = \begin{pmatrix} 0 \\ 1 + \mathcal{O}(\alpha) \end{pmatrix},$$

$$(56) \quad S''(s) = \begin{pmatrix} -\mathbf{c}_s \\ -\mathbf{s}_s \end{pmatrix} + \mathcal{O}(\alpha).$$

Integrating the above equation, we have

$$(57) \quad S'(s) = S'(0) + \int_0^s S''(\tau) d\tau = \begin{pmatrix} -\mathbf{s}_s + \mathcal{O}(s\alpha) \\ \mathbf{c}_s + \mathcal{O}(\alpha) \end{pmatrix},$$

$$(58) \quad S(s) = S(0) + \int_0^s S'(\tau) d\tau = \begin{pmatrix} 2r - 1 + \mathbf{c}_s + \mathcal{O}(\alpha) \\ \mathbf{s}_s + \mathcal{O}(s\alpha) \end{pmatrix},$$

and therefore

$$\begin{aligned} \frac{1}{\mathbf{c}_\alpha} X_\pm(S(s)) &= \begin{pmatrix} -S_2(s) \\ S_1(s) \pm \frac{\tan \alpha}{\alpha} \sqrt{1 - \alpha^2 S_1(s)^2 - \alpha^2 S_2(s)^2} \end{pmatrix} \\ &= \begin{pmatrix} -\mathbf{s}_s + \mathcal{O}(s\alpha) \\ 2r - 1 + \mathbf{c}_s + \mathcal{O}(\alpha) \pm (1 + \mathcal{O}(\alpha^2)) \end{pmatrix}. \end{aligned}$$

Here  $S_i$ ,  $i = 1, 2$ , denotes the  $i$ th component of  $S$ . Dividing the above equation by  $1 + \mathcal{O}(\alpha)$ , we can assume that the first component is identically equal to  $-\mathbf{s}_s$ . The same can be done with the expression (57), so that it is possible to compare the three vectors obtained in this way simply by looking at the second components, which are equal to, respectively,  $2r - 1 + \mathbf{c}_s \pm 1 + \mathcal{O}(\alpha)$  and  $\mathbf{c}_s + \mathcal{O}(\alpha)$ . In particular, the fact that  $S(\cdot)$  is nowhere locally optimal if  $\alpha$  is small enough follows from the inequalities  $2r - 2 + \mathbf{c}_s + \mathcal{O}(\alpha) < \mathbf{c}_s + \mathcal{O}(\alpha) < 2r + \mathbf{c}_s + \mathcal{O}(\alpha)$ .  $\square$

A straightforward consequence of the previous result is the presence of a nontrivial cut locus in the neighborhood of the south pole enclosed by  $F(\alpha, k_M\pi)$ . It remains to clearly define that cut locus, which is the purpose of Propositions 5 and 6.

**4.3.1. Proof of Proposition 5.** As usual, we provide only an argument in the case  $k_M$  odd, and we fix the remainder equal to  $\bar{r} \in (0, 1)$ .

Recall that, according to section 4.1,  $OF(\alpha, k_M\pi)$  is approximately (up to order  $\alpha^2$ ) a circle of radius  $2\bar{r}\alpha$ . To describe the synthesis inside the neighborhood of the south pole enclosed by  $OF(\alpha, k_M\pi)$ , it is more convenient to use the two-dimensional control system  $(\tilde{\Sigma})_\alpha$ , which is rewritten as follows by using (46):

$$\dot{z} = \mathbf{c}_\alpha A_0 z + u \frac{\mathbf{s}_\alpha}{\alpha} \sqrt{1 - \alpha^2 \|z\|^2} b_0.$$

We set  $\tilde{X}_\varepsilon^\alpha(z) := \mathbf{c}_\alpha A_0 z + \varepsilon \frac{\mathbf{s}_\alpha}{\alpha} \sqrt{1 - \alpha^2 \|z\|^2} b_0$  and  $\tilde{X}_\varepsilon^{pen}(z) := A_0 z + \varepsilon b_0$  for  $\varepsilon = \pm$ , and we define  $\widetilde{OF}(\alpha, k_M\pi)$  as the image by  $M_\alpha$  of  $OF(\alpha, k_M\pi)$ . Then we know that, up to order  $\alpha$ ,  $\widetilde{OF}(\alpha, k_M\pi)$  is a circle of radius  $2\bar{r}$ . In particular, as in the proof of Proposition 4, one can construct a piecewise smooth parametrization  $\sigma_\alpha : [0, 2\pi] \rightarrow \tilde{F}(\alpha, k_M\pi)$  so that  $\sigma_\alpha(0) = P_-^\alpha$ ,  $\sigma_\alpha(\pi) = P_+^\alpha$  with a loss of regularity occurring only at  $s = 0, \pi$  (with two-sided differentials at any order). In particular  $\sigma_\alpha(\cdot)$  approximates in the  $\mathcal{C}^0$  sense the function  $\sigma : [0, 2\pi] \rightarrow \mathbb{C} \sim \mathbb{R}^2$ , defined as  $\sigma(s) = 2\bar{r} e^{i(\pi-s)}$ , which is a parametrization of the circle of radius  $2\bar{r}$ .

Taking into account Lemma 3, the cut locus in  $V_\alpha$  is contained inside the set of points  $Q \in \mathbb{R}^2$ , besides  $P_\varepsilon^\alpha$ , such that there exists  $(s, s', t) \in (0, \pi) \times (\pi, 2\pi) \times (0, 2\pi)$  for which  $Q = e^{t\tilde{X}_+^\alpha} \sigma_\alpha(s') = e^{t\tilde{X}_-^\alpha} \sigma_\alpha(s)$ .

In view of applying an inverse function result for characterizing this set, we consider the map  $\Phi$  defined on  $[0, \pi] \times [\pi, 2\pi] \times [0, \pi]$  by

$$\Phi(s, s', t) := (s, e^{t\tilde{X}_+^{pen}} \sigma(s') - e^{t\tilde{X}_-^{pen}} \sigma(s)),$$

which takes values in  $\mathbb{R}^3$ . Similarly, for  $k \geq 1$  and  $\alpha_k$  as in the proposition, we consider the map  $\Phi_k$  defined on  $[0, \pi] \times [\pi, 2\pi] \times [0, \pi]$  by

$$\Phi_k(s, s', t) := (s, e^{t\tilde{X}_+^{\alpha_k}} \sigma_{\alpha_k}(s') - e^{t\tilde{X}_-^{\alpha_k}} \sigma_{\alpha_k}(s)).$$

Note that, since the vector fields  $\tilde{X}_\varepsilon^{\alpha_k}$  converge uniformly to  $\tilde{X}_\varepsilon^{pen}$  on  $V_\alpha$ , it is easy to see that  $\Phi_k$  converges to  $\Phi$  in the  $\mathcal{C}^1$  norm.

For  $(Pen)$ , a point of the overlap curve, besides  $P_\varepsilon^\alpha$ , is then identified with a triple  $(s, s', t) \in (0, \pi) \times (\pi, 2\pi) \times (0, \pi)$  such that  $\Phi(s, s', t) = (s, 0, 0)$ . In other words, the overlap curve can be parametrized by means of the map  $w : [0, \pi] \rightarrow \mathbb{R}^3$  defined implicitly by  $\Phi(w(s)) = (s, 0, 0)$ , while  $\gamma_{pen}^o$  can be obtained as the composition of the two maps  $e^{t\tilde{X}_-^{pen}} \sigma(s)$  and  $w(s)$ .

Similarly, we would like to define the overlap curve corresponding to  $(\Sigma)_{\alpha_k}$ , for  $k$  large enough, by means of the function  $w_k$  defined by  $\Phi_k(w_k(s)) = (s, 0, 0)$ . To proceed, we will apply Theorem 3. The first task consists of computing  $\det D\Phi$  along the overlap curve.

**LEMMA 4.** *Along the set of triples  $(s, s', t) \in (0, \pi) \times (\pi, 2\pi) \times (0, \pi)$  for which  $e^{t\tilde{X}_+^{pen}} \sigma(s') = e^{t\tilde{X}_-^{pen}} \sigma(s)$ , we have*

$$\det D\Phi(s, s', t) = \frac{4\bar{r}(1 - \bar{r}^2) \sin s'}{(1 - \bar{r} \cos s')^2 + (\bar{r} \sin s')^2}.$$

*Proof of Lemma 4.* One has

$$\det D\Phi(s, s', t) = \det \left( (e^{t\tilde{X}_+^{pen}})_* \frac{d\sigma}{ds'}, \tilde{X}_+^{pen} e^{t\tilde{X}_+^{pen}} \sigma(s') - \tilde{X}_-^{pen} e^{t\tilde{X}_-^{pen}} \sigma(s) \right).$$

By taking into account that  $\Phi(s, s', t) = 0$ , the previous determinant is equal to twice the first component of  $(e^{t\tilde{X}_+^{pen}})_* \frac{d\sigma}{ds'}$ , i.e.,  $\det \mathcal{D}\Phi(s, s', t) = 4\bar{r} \sin(s' - t)$ . Using (53), one concludes.  $\square$

Observe that  $\det D\Phi \neq 0$  if  $s' \neq 0, \pi$ . In particular, if we consider a closed interval  $I \subset (0, \pi)$ , then the set  $\text{Im}(w)|_I$  plays the role of the compact set  $\mathcal{K}$  in Theorem 3. All of the assumptions of the theorem are then verified, and therefore we have proved the existence of a map  $w_k$  defined on  $I$ , satisfying  $\Phi_k(w_k(s)) = (s, 0, 0)$  and converging uniformly to  $w$ . If we define  $\gamma_k^\circ$  as the composition of the two maps  $e^{t\tilde{X}_-^{\alpha_k}} \sigma(s)$  and  $w(s)$ , then, since  $I$  was chosen arbitrarily, the proof of the theorem is complete.  $\square$

**4.3.2. Proof of Proposition 6.** With the previous notations, let  $\varphi_k$  be the smooth map defined on  $[0, \pi] \times [\pi, 2\pi] \times [0, 2\pi]$  by

$$\varphi_k(s, s', t) = e^{t\tilde{X}_+^{\alpha_k}} \sigma_k(s') - e^{tX_-^{\alpha_k}} \sigma_k(s).$$

For the rest of this paragraph, we drop the index  $k$  to get lighter notations.

From the Taylor expansion of  $\varphi$  around the points  $(0, 2\pi, 0)$  and  $(\pi, \pi, 0)$ , we derive the asymptotic behaviors of the cut locus close to the points  $P_\varepsilon^\alpha$ ,  $\varepsilon = \pm$ , since that cut locus belongs to the level set  $\varphi = 0$ . We will perform computations only at  $(0, 2\pi, 0)$  since they are entirely similar at  $(\pi, \pi, 0)$ .

Let us call  $\varphi^{(1)}$  and  $\varphi^{(2)}$  the two components of  $\varphi$ . We use  $\varphi_s^{(i)}$  to denote the partial derivative of the component  $\varphi^{(i)}$  with respect to  $s$  evaluated in  $(0, 2\pi, 0)$ , and we define in an analogous way all of the (multiple) partial derivatives evaluated in  $(0, 2\pi, 0)$ . Set  $\tilde{s} := s' - 2\pi$ . Then, after computations, we have  $\varphi_s^{(1)} = \varphi_{\tilde{s}}^{(1)} = \varphi_t^{(1)} = 0$  and

$$\begin{aligned} \varphi_{ss}^{(1)} &= -2\bar{r} + \mathcal{O}(\alpha), & \varphi_{\tilde{s}\tilde{s}}^{(1)} &= 2\bar{r} + \mathcal{O}(\alpha), & \varphi_{tt}^{(1)} &= 2 + \mathcal{O}(\alpha), \\ \varphi_{s\tilde{s}}^{(1)} &= 0, & \varphi_{st}^{(1)} &= -2\bar{r} + \mathcal{O}(\alpha), & \varphi_{\tilde{s}t}^{(1)} &= 2\bar{r} + \mathcal{O}(\alpha), \\ \varphi_s^{(2)} &= 2\bar{r} + \mathcal{O}(\alpha), & \varphi_{\tilde{s}}^{(2)} &= -2\bar{r} + \mathcal{O}(\alpha), & \varphi_t^{(2)} &= 2\bar{r} + \mathcal{O}(\alpha). \end{aligned}$$

We thus get

$$\begin{aligned} \varphi^{(1)}(s, \tilde{s}, t) &= \varphi_{ss}^{(1)} s^2 + \varphi_{\tilde{s}\tilde{s}}^{(1)} \tilde{s}^2 + \varphi_{tt}^{(1)} t^2 + 2\varphi_{st}^{(1)} st + 2\varphi_{\tilde{s}t}^{(1)} \tilde{s}t + \mathcal{O}(|(s, \tilde{s}, t)|^3) \\ (59) \quad &= -2\bar{r}s^2 + 2r\tilde{s}^2 + 2t^2 - 4\bar{r}st + 4r\tilde{s}t + \mathcal{O}(\alpha|(s, \tilde{s}, t)|^2) + \mathcal{O}(|(s, \tilde{s}, t)|^3), \end{aligned}$$

and

$$\begin{aligned} \varphi^{(2)}(s, \tilde{s}, t) &= \varphi_s^{(2)} s + \varphi_{\tilde{s}}^{(2)} \tilde{s} + \varphi_t^{(2)} t + \mathcal{O}(|(s, \tilde{s}, t)|^2) \\ (60) \quad &= 2\bar{r}s - 2r\tilde{s} - 2t + \mathcal{O}(\alpha|(s, \tilde{s}, t)|) + \mathcal{O}(|(s, \tilde{s}, t)|^2), \end{aligned}$$

where, here,  $\mathcal{O}(\cdot)$  is uniform with respect to  $\alpha$ .

Fix  $\xi_0 > 0$  small. We are looking at the cut locus in a neighborhood of  $P_\varepsilon^\alpha$ , and thus we can assume  $|(s, \tilde{s}, t)| < \xi_0$  for some  $\xi_0 > 0$ . The purpose of subsequent computations consists of expressing  $\tilde{s} < 0$  and  $t > 0$  as functions of  $s$ , for  $0 \leq s \leq \xi_0$ , by using the equations  $\varphi^{(1)} = 0$  and  $\varphi^{(2)} = 0$ .

From  $\varphi^{(2)} = 0$ , by applying the implicit function theorem, for  $\xi_0$  small enough and  $|(s, \tilde{s})| < \xi_0$ , we get  $t = h(s, \tilde{s})$ , with  $h \in \mathcal{C}^1$ . Moreover, since  $h(s, \tilde{s}) = \mathcal{O}(|(s, \tilde{s})|)$  we have

$$(61) \quad h(s, \tilde{s}) = \bar{r}s - r\tilde{s} + \mathcal{O}(\alpha|(s, \tilde{s})|) + \mathcal{O}(|(s, \tilde{s})|^2).$$

Consider now the map

$$\phi(s, \tilde{s}) = \frac{\varphi^{(1)}(s, \tilde{s}, h(s, \tilde{s}))}{s - \tilde{s}},$$

which is well defined and  $\mathcal{C}^1$  for  $s > 0, \tilde{s} < 0$ . Again, it is possible to apply the implicit function theorem to the equation  $\phi = 0$ , so that we get  $\tilde{s}$  as a  $\mathcal{C}^1$  function of  $s$ , and this gives the existence of the overlap curve. Moreover, by combining (60) and (61), we get the following:

$$(62) \quad \phi(s, \tilde{s}) = s(1 + \bar{r}) + \tilde{s}(1 - \bar{r}) + \mathcal{O}(\alpha|(s, \tilde{s})|) + \mathcal{O}(|(s, \tilde{s})|^2) = 0,$$

and then  $|\tilde{s}| = \mathcal{O}(|s|)$ . Therefore, from this estimate and the above ones, we immediately obtain that

$$(63) \quad \tilde{s} = -\left(\frac{1 + \bar{r}}{1 - \bar{r}} + \mathcal{O}(\alpha)\right)s + \mathcal{O}(s^2), \quad t = \left(\frac{2\bar{r}}{1 - \bar{r}} + \mathcal{O}(\alpha)\right)s + \mathcal{O}(s^2),$$

from which we get that the overlap curve converges uniformly to  $\gamma_{pen}^o$ .

The proof of Proposition 6 is now complete.  $\square$

*Remark 10.* In order to prove that the overlap curve  $\gamma_k^o$  converges to  $\gamma_{pen}^o$  uniformly with respect to  $\bar{r}$  in any closed interval  $I \subset (0, 1)$ , it is enough to follow the lines of the proofs of Propositions 5 and 6 by considering  $\bar{r}$  as an additional variable. For instance, for Proposition 5, one needs to define the maps  $\tilde{\Phi} : \mathbb{R}^4 \rightarrow \mathbb{R}^4$ ,  $\tilde{\Phi}(\bar{r}, s, s', t) := (\bar{r}, s, e^{t\tilde{X}_+^{pen}}\sigma(s') - e^{t\tilde{X}_-^{pen}}\sigma(s))$  (recall that  $\sigma(\cdot)$  depends on  $\bar{r}$ ) and  $\tilde{\Phi}_k : \mathbb{R}^4 \rightarrow \mathbb{R}^4$ ,  $\tilde{\Phi}_k(\bar{r}, s, s', t) := (\bar{r}, s, e^{t\tilde{X}_+^{\alpha_k}}\sigma_k(s') - e^{t\tilde{X}_-^{\alpha_k}}\sigma_k(s))$  (where  $\alpha_k = \pi/(2(\bar{r} + k))$ ) and  $\sigma_k(\cdot)$  depends on  $\bar{r}$ ).

The uniformity with respect to  $\bar{r}$  is then proved by applying Theorem 3 to  $\tilde{\Phi}, \tilde{\Phi}_k$  with  $\mathcal{K} = \{\tilde{\Phi}^{-1}(\bar{r}, s, 0, 0) : (\bar{r}, s) \in I \times J\}$ , where  $I \times J$  is a compact subset of  $(0, 1) \times (0, \pi)$ .

## 5. Case $r = C\alpha$ .

**5.1. Description of the minimum time front at time  $k_M\pi$ .** Fix  $C > 0$ , and consider the sequence  $(\alpha_k)$  such that  $r(\alpha_k) = C\alpha_k$ ,  $k \geq 0$ . As before, we drop the index  $k$  when possible. For  $\alpha_k$  small enough, one deduces, from the analysis of [9], that the south pole is not reached at time  $k_m\pi = [\frac{\pi}{2\alpha}] \pi$ , so that the optimal front at time  $k_M\pi$  is not empty. The next result provides a description of the extremal front at time  $k_M\pi$ .

LEMMA 5. Define the planar curve  $\mathcal{L} : [0, 2\pi] \rightarrow \mathbb{R}^2$  by

$$(64) \quad \mathcal{L}(s) = \begin{pmatrix} \mathbf{c}_s(-2C + \pi \mathbf{s}_s^2/2) \\ \mathbf{s}_s(\pi + 2C - \pi \mathbf{s}_s^2/2) \end{pmatrix}.$$

Then, for  $s \in [0, \pi]$ , we have

$$(65) \quad \mathcal{E}^+(\alpha, s) = (\alpha^2 \mathcal{L}(s), -1)^T + \mathcal{O}(\alpha^3),$$

and

$$(66) \quad \frac{\partial}{\partial s} \mathcal{E}^+(\alpha, s) = \left( \alpha^2 \frac{d}{ds} \mathcal{L}(s), 0 \right)^T + \mathcal{O}(\alpha^3).$$

At  $s = 0$  and  $s = \pi$ , the derivatives are only one-sided, i.e., as  $s > 0$  tends to zero and  $s < \pi$  tends to  $\pi$ .

Similarly, we have, for  $s \in [0, \pi]$ ,

$$(67) \quad \mathcal{E}^-(\alpha, k_M \pi, s) = (\alpha^2 \mathcal{L}(s + \pi), -1)^T + \mathcal{O}(\alpha^3),$$

and

$$(68) \quad \frac{\partial}{\partial s} \mathcal{E}^-(\alpha, k_M \pi, s) = \left( \alpha^2 \frac{d}{ds} \mathcal{L}(s + \pi), 0 \right)^T + \mathcal{O}(\alpha^3),$$

with one-sided derivatives at  $s = 0$  and  $s = \pi$ .

*Proof of Lemma 5.* The proof is immediate from Proposition 3 applied in the case  $r(\alpha) = C\alpha$ .  $\square$

For  $C < \pi/4$ , consider  $\theta_d \in (0, \pi/2)$ , with  $\sin(\theta_d) = 2\sqrt{C/\pi}$ . The curve  $\mathcal{L}(s)$  has two double points  $D^+ = \mathcal{L}(s_1^+) = \mathcal{L}(s_2^+)$ , with  $s_1^+ = \theta_d$  and  $s_2^+ = \pi - \theta_d$ , and  $D^- = \mathcal{L}(s_1^-) = \mathcal{L}(s_2^-)$ , with  $s_1^- = \pi + \theta_d$  and  $s_2^- = 2\pi - \theta_d$ . It also has four cuspidal points  $Cp_i^\varepsilon$ ,  $i = 1, 2$  and  $\varepsilon = \pm$ , corresponding to the values  $s = s_{cusp,i}^\varepsilon$ , where  $\sin^2 s = \frac{2+4C/\pi}{3}$ .

Finally, let  $\sigma$  be the closed Jordan curve defined as the restriction of  $\mathcal{L}(s)$  to  $[0, s_1^+] \cup [s_2^+, s_1^-] \cup [s_2^-, 2\pi]$ . If  $C > \pi/4$ , we simply define  $\sigma$  to be  $\mathcal{L}$ .

In light of the previous result, we get that  $\mathcal{F}(\alpha, k_M \pi)$ , the extremal front at time  $k_M \pi$ , is contained inside a neighborhood  $W_\alpha$  of the south pole of order  $\mathcal{O}(\alpha^2)$  neighborhood of the south pole. Therefore, in order to understand the shape of the optimal synthesis inside  $W_\alpha$ , we must rescale the whole problem by  $N_\alpha$ , the linear mapping from  $\mathbb{R}^3$  onto  $\mathbb{R}^2$  defined as the composition of the orthogonal projection  $(x_1, x_2, x_3) \mapsto (x_1, x_2)$  followed by the dilation by  $1/\alpha^2$ .

For  $x \in W_\alpha$ , we first consider  $(\Lambda)_\alpha$ , the image of  $(\Sigma)$  by  $N_\alpha$ ; i.e.,  $(\Lambda)_\alpha$  is the planar control system given by

$$(69) \quad (\Lambda)_\alpha : \begin{cases} \dot{z}_1 = -\mathbf{c}_\alpha z_2, \\ \dot{z}_2 = \mathbf{c}_\alpha z_1 + u \frac{\mathbf{s}_\alpha}{\alpha^2} \sqrt{1 - \alpha^4 \|z\|^2}. \end{cases}$$

Let  $\mathcal{L}_\alpha$  be the image of  $\mathcal{F}(\alpha, k_M \pi)$  by  $N_\alpha$ . From Lemma 5,  $\mathcal{L}_\alpha$  converges to  $\mathcal{L}$  in the  $C^1$  topology. It is clear that, for  $C > \pi/4$ ,  $\mathcal{L}_\alpha : [0, 2\pi] \rightarrow \mathbb{R}^2$  is homeomorphic to  $e^{is}$ ,  $s \in [0, 2\pi]$ . In the case where  $C < \pi/4$ , the next lemma shows that, for  $\alpha$  small enough,  $\mathcal{L}_\alpha$  has the same shape as  $\mathcal{L}$ .

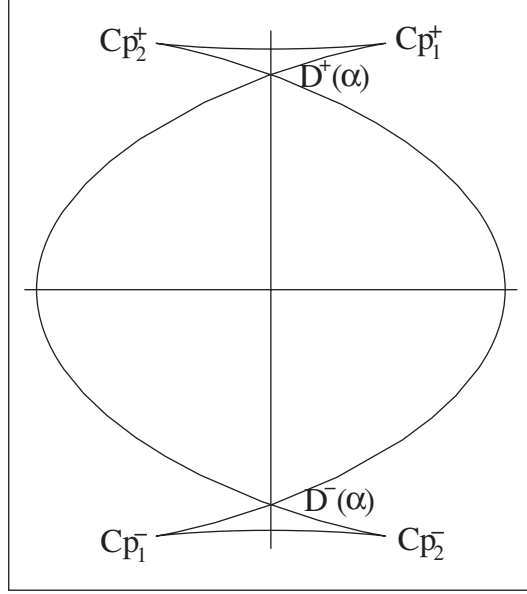
**LEMMA 6.** *If  $C < \pi/4$ , then  $\mathcal{L}_\alpha$  is described by the Figure 9, where  $Cp_i^\varepsilon(\alpha) = \mathcal{L}_\alpha(s_{cusp,i}^\varepsilon(\alpha))$ ,  $i = 1, 2$  and  $\varepsilon = \pm$ , are cuspidal points and  $D^\varepsilon(\alpha)$  are double points, with*

$$(70) \quad D^+(\alpha) = \mathcal{L}_\alpha(s_1^+(\alpha)) = \mathcal{L}_\alpha(s_2^+(\alpha)), \quad D^-(\alpha) = \mathcal{L}_\alpha(s_1^-(\alpha)) = \mathcal{L}_\alpha(s_2^-(\alpha)),$$

where  $s_{cusp,i}^\varepsilon(\alpha)$  and  $s_i^\varepsilon(\alpha)$  tend to, respectively,  $s_{cusp,i}^\varepsilon$  and  $s_i^\varepsilon$  as  $\alpha$  tends to zero for  $i = 1, 2$  and  $\varepsilon = \pm$ . For  $\alpha$  small enough, set  $\sigma_\alpha$ , the closed curve defined as the restriction of  $\mathcal{L}_\alpha(s)$ , to  $[0, s_1^+(\alpha)] \cup [s_2^+(\alpha), s_1^-(\alpha)] \cup [s_2^-(\alpha), 2\pi]$ . Then it is a Jordan curve.

*Proof of Lemma 6.* For  $i = 1, 2$  and  $\varepsilon = \pm$ , the existence of the cuspidal points  $Cp_i^\varepsilon(\alpha)$  is obtained by applying the implicit function theorem to the equation  $DL(s, \alpha) = 0$ , where the function  $DL(s, \alpha)$  is defined by

$$DL(s, \alpha) := \frac{d}{ds} \mathcal{L}_\alpha(s),$$

FIG. 9. Graph of the function  $\mathcal{L}_\alpha$  for  $C < \pi/4$ .

in the neighborhood of each  $(s_{cusp,i}^\varepsilon, 0)$ . We have

$$\partial_s DL(s_{cusp,i}^\varepsilon, 0) = \frac{d^2}{ds^2} \mathcal{L}(s_{cusp,i}^\varepsilon) \neq 0,$$

and we conclude. The uniqueness of these four points, on  $[0, 2\pi]$ , is trivial since  $DL(s, \alpha) = \frac{d}{ds} \mathcal{L}(s) + \mathcal{O}(\alpha)$ .

Similarly, for  $\varepsilon = \pm$ , the existence of the double points  $D^\varepsilon(\alpha)$  follows after applying the implicit function theorem to the equation  $DP(s, s', \alpha) = 0$ , where the function  $DP(s, s', \alpha)$  is defined by

$$DP(s, s', \alpha) = \mathcal{L}_\alpha(s) - \mathcal{L}_\alpha(s'),$$

in the neighborhood of each  $(s_1^\varepsilon, s_2^\varepsilon, 0)$ . For the uniqueness, we proceed as before.  $\square$

In the case  $C > \pi/4$ , we also define  $\sigma_\alpha$  to be equal to  $\mathcal{L}_\alpha$ . As a consequence, we are able to characterize  $OF(\alpha, k_M\pi)$ , the minimum time front at time  $k_M\pi$  when  $C \neq \pi/4$ .

**PROPOSITION 7.** *For  $\alpha$  small enough and  $C \neq \pi/4$ , the minimum time front at time  $k_M\pi$ ,  $OF(\alpha_k, k_M\pi)$  is equal to  $\tilde{\sigma}_\alpha$ , the inverse image on  $S^2$ , by  $N_\alpha$ , of  $\sigma_\alpha$ .*

*Remark 11.* As a consequence, we deduce that, for  $C > \pi/4$  and  $\alpha$  small enough, the optimal synthesis between  $\mathcal{F}(\alpha, (k_M - 1)\pi)$  and  $\mathcal{F}(\alpha, k_M\pi)$  is simply given by the extremal flow, whereas, for  $C < \pi/4$ , there is a loss of optimality along certain extremal curves starting at  $\mathcal{F}(\alpha, (k_M - 1)\pi)$  before reaching  $\mathcal{F}(\alpha, k_M\pi)$ . The values of  $s$  corresponding to such curves can be deduced from the previous characterizations of  $\mathcal{F}(\alpha, k_M\pi)$  and  $OF(\alpha_k, k_M\pi)$ .

*Proof of Proposition 7.* Recall that  $OF(\alpha_k, k_M\pi)$  is a piecewise  $\mathcal{C}^1$  submanifold of  $\mathcal{F}(\alpha, k_M\pi)$ . As in the proof of Proposition 4, the result to establish is a consequence



of the fact that  $\sigma_\alpha = \mathcal{L}_\alpha : [0, 2\pi] \rightarrow \mathbb{R}^2$  is homeomorphic to  $e^{is}$ ,  $s \in [0, 2\pi]$ , and it can be achieved by means of simple topological arguments.

In the case  $C < \pi/4$ ,  $\sigma_\alpha$  is a piecewise  $\mathcal{C}^1$  Jordan curve homeomorphic to  $e^{is}$ ,  $s \in [0, 2\pi]$ . A simple topological argument yields the conclusion.  $\square$

**5.2. Limit of the synthesis.** It remains to describe the limiting dynamics close to the south pole. In order to take the limit, as  $\alpha$  tends to zero, in  $(\Lambda)_\alpha$ , one must reparametrize by the time  $\alpha t$ . The limit is then given by the control system

$$(\Lambda) : \quad \begin{cases} \dot{z}_1 = 0, \\ \dot{z}_2 = u. \end{cases}$$

We now describe the optimal synthesis for the limit problem, i.e., for the problem of reaching in minimum time every point inside  $\sigma$  along  $(\Lambda)$  and starting from  $\sigma$ . Because of the symmetries of  $\sigma$  and because the tangent vector to  $\sigma$  is vertical only at  $s = 0$  and  $s = \pi$ , there exists a unique overlap curve  $(Seg)_C$ , defined as the segment of the  $z_1$ -axis between the points  $(-2C, 0)$  and  $(2C, 0)$ . Above it, the input  $u$  takes the constant value  $-1$ , and, below that overlap curve, the constant value  $1$ . Integral curves are clearly vertical lines.

We next intend to prove that the optimal synthesis consisting of reaching in minimum time every point inside  $\sigma_\alpha$  along  $(\Lambda)_\alpha$  and starting from  $\sigma_\alpha$  converges to the previous synthesis in the following sense.

**THEOREM 2.** *Assume that  $C \neq \pi/4$ . As  $\alpha$  tends to zero, the time optimal synthesis associated to  $(\Lambda)_\alpha$  inside  $\sigma_\alpha$  is characterized by an overlap curve  $(Seg)_C^\alpha$ , converging to  $(Seg)_C$  in the  $C^0$  topology, and, above  $(Seg)_C^\alpha$ , the control  $u$  takes the constant value  $-1$ , and, below  $(Seg)_C^\alpha$ , it is equal to  $1$ . Moreover, there exist only two time optimal trajectories reaching the origin, and, in the case  $C < \pi/4$ , these trajectories start from  $D_\alpha^\varepsilon$ ,  $\varepsilon = \pm$ , the double points of  $\mathcal{L}_\alpha$ .*

*Proof of Theorem 2.* Fix  $C \neq \pi/4$ . We first notice that, for  $\alpha$  small enough, there are not switching curves inside  $\sigma_\alpha$ . Therefore, the cut locus may occur only as images by  $N_\alpha$  of points  $M \in S^2$  such that  $M = e^{\frac{t}{\alpha}X} \tilde{\sigma}(s) = e^{\frac{t}{\alpha}X} \tilde{\sigma}(s')$  for  $t \in [0, \frac{2\pi}{\alpha}]$ ,  $s \in [0, \pi]$ , and  $s' \in [\pi, 2\pi]$ . Proceeding exactly as in the proof of Theorem 1, we apply inverse function arguments first in neighborhoods of  $\sigma_\alpha(0)$  and  $\sigma_\alpha(\pi)$  and second in a region enclosed by  $\sigma_\alpha$  excluding such neighborhoods. It is then easy to determine the values of the input  $u$  in each connected component of the region enclosed by  $\sigma_\alpha$  minus  $(Seg)_C^\alpha$ .

By a continuity argument, it is clear that there exist only two time optimal trajectories reaching the origin: one above  $(Seg)_C^\alpha$  and one below. Finally, suppose that  $C < \pi/4$ . In that case, it was proved in [9] that the only extremals starting at a point  $\mathcal{L}_\alpha(s)$  and reaching the origin from above the overlap curve  $(Seg)_C^\alpha$  correspond to values of  $s$  verifying one of the following three possibilities as  $\alpha$  tends to zero: (a)  $s$  tends to zero, (b)  $s$  tends to  $\pi/2$ , or (c)  $\mathcal{L}_\alpha(s)$  is a double point also associated to  $s' = v(s) - s$ . In view of what precedes, only possibility (c) is allowed for optimality. Theorem 2 is proved.  $\square$

*Remark 12.* As a consequence of the previous argument and from the results of [9], we get that, for  $\alpha$  small enough and  $C < \pi/4$ ,

$$s_2^+(\alpha) = v(s_1^+(\alpha)) - s_1^+(\alpha), \quad s_2^-(\alpha) = 2\pi + v(s_1^-(\alpha) - \pi) - s_1^-(\alpha),$$

where  $s_i^\varepsilon(\alpha)$ ,  $i = 1, 2$   $\varepsilon = \pm$ , were defined in (70).

**6. Case  $r(\alpha) = 0$ .** We assume here that  $r(\alpha) = 0$ , i.e.,  $\alpha_k = \frac{\pi}{2k}$  for  $k \geq 1$ . From Proposition 3, we know that the extremal front at time  $([\frac{\pi}{2\alpha}] - 1)\pi = \frac{\pi}{2\alpha} - \pi$  encloses the south pole, is optimal, and is approximately (in the  $\mathcal{C}^1$  sense) a circle of radius  $2r(\alpha)\alpha$  around the south pole. Moreover, at time  $[\frac{\pi}{2\alpha}]\pi$ , we know that the extremal front must contain the south pole and is equal, up to  $\mathcal{O}(\alpha^3)$ , to  $(\alpha^2\mathcal{L}, -1)^T$  given in (65) and (67) with  $C = 0$ . In that case, the minimum time front reduces to the south pole.

In this case it is interesting to consider the synthesis starting from the extremal front at time  $(k_M - 1)\pi$ , and it is natural to compare it with the synthesis of the linear pendulum studied in section 4.2 and corresponding to  $\rho = 2$ . See Figure 10. Let us first describe briefly that synthesis. Let  $D_2$  and  $C_2$  be the disc and the circle centered at the origin and of radius 2, respectively. The overlap curve inside  $D_2$  coincides with the switching curves and with the trajectories, corresponding to  $u = \pm 1$ , connecting points  $(\pm 2, 0)$  to the origin. In particular, it means that an optimal trajectory of the synthesis starting at any point  $P \in C_2$  reaches the origin, and thus there exists an infinite number of optimal trajectories from  $C_2$  to the origin.

For  $\alpha > 0$  and  $r(\alpha) = 0$ , the situation is rather different. Let us first define  $\tilde{F}(\alpha, (k_M - 1)\pi)$  to be the image of  $\mathcal{F}(\alpha, (k_M - 1)\pi)$  by  $M_\alpha$ . Then, for  $\alpha$  small enough, it was shown in [9] that the only optimal trajectories starting from  $\tilde{F}(\alpha, (k_M - 1)\pi)$  and reaching the origin are those starting at  $P_+^\alpha$  and  $P_-^\alpha$ . Let us refer to them as  $\gamma^+$  and  $\gamma^-$ . Therefore, in the case  $r(\alpha) = 0$ , the synthesis for  $\alpha > 0$  is rather different than the synthesis of the limit candidate when  $\alpha$  tends to zero. It is a clear indication that the case  $r(\alpha) = 0$  is more delicate than the cases  $r(\alpha)$  positive constant or  $r(\alpha) = C\alpha$ . However, we are still able to give a partial description of the limit synthesis as the next proposition shows.

**PROPOSITION 8.** *Assume that  $r(\alpha) = 0$  and  $\alpha$  is small enough. Then the switching curve  $C_{k_M}^+$  (respectively,  $C_{k_M}^-$ ) is optimal for some interval  $[0, s(\alpha)]$ ,  $s(\alpha) < \pi$ , and it is above (respectively, below)  $\gamma^+$  (respectively,  $\gamma^-$ ) as long as it is optimal. Moreover, we have*

$$(71) \quad \lim_{\alpha \rightarrow 0, r(\alpha)=0} s(\alpha) = \bar{s} := \arccos \sqrt{1/3}.$$

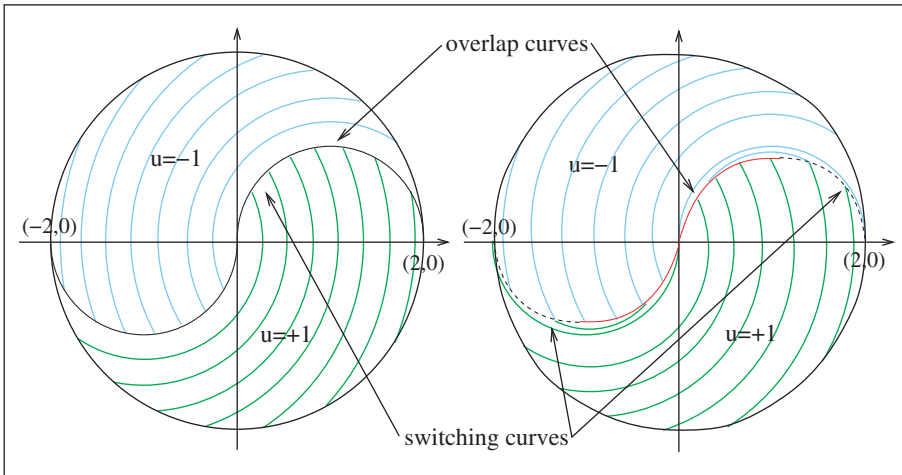


FIG. 10. Comparison between the optimal synthesis for the linear pendulum and the optimal synthesis on the bottom of the sphere in the case  $r(\alpha) = 0$ .

*Proof of Proposition 8.* We provide only an argument for  $C_{k_M}^+$ , the other case being analogous. To prove the first statement of the proposition, we reason by contradiction. If the switching curve is not optimal on any interval  $[0, \tau]$ ,  $\tau > 0$ , we get the existence of an optimal trajectory starting at  $\mathcal{F}(\alpha, (k_M - 1)\pi)$  above  $P_\alpha^+$  and reaching the origin, which is equal to the concatenation of an integral curve of  $X_-$  and a piece of  $\gamma^+$ . Therefore, an optimal integral curve of  $X_-$ , starting above  $\gamma^+$ , must either switch or lose optimality before reaching  $\gamma^+$ . If the second possibility occurs, we must have an overlap; i.e., at that point an optimal integral curve of  $X_+$  arrives. Close to  $P_\alpha^+$ , the latter would imply that the optimal integral curve of  $X_+$  starts at  $\mathcal{F}(\alpha, (k_M - 1)\pi)$  above  $P_\alpha^+$ . This is impossible, because, from every point of  $\mathcal{E}^+(\alpha, (k_M - 1)\pi)$ , the value of the optimal control is  $-1$ . Let  $s(\alpha) \leq \pi$  be the first value of  $s$  for which  $C_{k_M}^+$  ceases to be optimal. Define

$$H(s) := \det \left( X_+(C_{k_M}^+(s)), \frac{dC_{k_M}^+(s)}{ds}, C_{k_M}^+(s) \right)$$

for  $s \in [0, \pi]$ . Then  $s(\alpha)$  is the smallest solution in  $(0, \pi]$  of  $H(s) = 0$ . It is easy to see that  $H$  must take the value zero before  $\pi$ . We deduce that  $s(\alpha) < \pi$ . By taking the asymptotic expansion of the previous expression as  $\alpha$  tends to zero, we get

$$H(s) = \frac{\pi}{4} \mathbf{s}_s \alpha^3 (1 + 3 \cos(2s) + \mathcal{O}(\alpha)).$$

Then  $s(\alpha)$  must converge to  $\bar{s}$  as  $\alpha$  tends to zero, the smallest solution in  $[0, \pi]$  of  $1 + 3 \cos(2s) = 0$ .  $\square$

**7. Conclusion and final remarks.** In this paper, we built the time optimal synthesis for a two-level quantum system driven by an external field, in the case  $M \ll E$ , where  $M$  is the bound on the fields and  $-E$  and  $E$  are the two energy levels. In particular, we answered several questions stated in [5, 9], regarding the locus where extremals lose optimality and the shape of the synthesis at the south pole. To that purpose we characterized a concept of “asymptotic” optimal synthesis in the “noncontrollability” limit  $M/E \rightarrow 0$ , and we described it in detail: There are three main patterns which cyclically alternate as  $M/E \rightarrow 0$ .

Another point of interest of our results lies in the fact that, to the best of our knowledge, we provided here the first nontrivial example of time optimal synthesis with one bounded control on a 2-D compact manifold. Indeed, similarly to what happens in Riemannian geometry, singularities near the south pole appear, as a consequence of the nontrivial topology of  $S^2$ .

As a byproduct of our studies, we obtain the first not “ad hoc” example of the singularity  $(C, K)_1$  predicted by the general theory [8]. Recall that a singularity of an optimal synthesis is the intersection between two special curves (i.e., overlaps, switching curves, singular curves, etc.).

The  $(C, K)_1$  singularity has a “nonlocal” character in the sense that the two optimal trajectories merging at  $(C, K)_1$  are projections of extremals which are “far” in the cotangent bundle (see Figure 3).

**Appendix.** The following version of the inverse function theorem is used in the argument of Proposition 5.

**THEOREM 3.** *Let  $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a  $\mathcal{C}^1$  map and  $\mathcal{K} \subset \mathbb{R}^n$  a compact set such that  $\Phi|_{\mathcal{K}} : \mathcal{K} \rightarrow \Phi(\mathcal{K})$  is bijective and the differential  $D\Phi(x)$  is invertible for  $x \in \mathcal{K}$ . Then there exists an open neighborhood  $U \supset \mathcal{K}$  such that  $\Phi|_U$  is a  $\mathcal{C}^1$  diffeomorphism.*

Let now  $(\Phi_k)_{k \geq 1}$  be a sequence of  $\mathcal{C}^1$  maps converging in the  $\mathcal{C}_{loc}^1$  sense to  $\Phi$ . Then, for every open set  $\tilde{U}$  with closure included in  $U$ , there exists  $\bar{k}$  such that, for every  $k \geq \bar{k}$ ,  $\Phi_{k|_{\tilde{U}}}$  is a  $\mathcal{C}^1$  diffeomorphism and, for every compact subset  $\tilde{\mathcal{K}}$  of  $\tilde{U}$ ,  $\Phi(\tilde{\mathcal{K}}) \subset \Phi_k(\tilde{U})$  and  $\lim_{k \rightarrow \infty} \Phi_k^{-1}(v) = \Phi^{-1}(v)$  uniformly with respect to  $v \in \Phi(\tilde{\mathcal{K}})$ .

*Proof.* Let us define, for  $k \geq 0$ , the following open neighborhoods of  $\mathcal{K}$ :

$$A := \{x \in \mathbb{R}^n : \det D\Phi(x) \neq 0\}, \quad A_k := \cup_{x \in \mathcal{K}} B\left(x, \frac{1}{k}\right) \cap A.$$

In view of the inverse function theorem, in order to conclude the proof of the first part, it is enough to show that for  $k$  large enough the restriction  $\Phi|_{A_k}$  is one-to-one.

We argue by contradiction. Let  $x_k \neq y_k \in A_k$  such that  $\Phi(x_k) = \Phi(y_k)$  for all  $k$ . Then, up to extractions of subsequences, we can assume that the two sequences converge to  $\bar{x}$  and  $\bar{y}$ , respectively. Since  $\bar{x}, \bar{y} \in \cap_k A_k = \mathcal{K}$  and  $\Phi(\bar{x}) = \Phi(\bar{y})$ , we deduce that  $\bar{x} = \bar{y}$ . However, since  $\det D\Phi(\bar{x}) \neq 0$ , we have that  $\Phi$  is bijective in a neighborhood of  $\bar{x}$ , which contradicts the assumption  $\Phi(x_k) = \Phi(y_k)$  for  $k$  large enough.

The proof of the second part is similar. First, fix a subset  $\tilde{U}$  of  $U$ . By the uniform convergence of  $D\phi_k$  to  $D\phi$  on every compact subset of  $U$ , we get  $\det D\Phi_k(x) \neq 0$  for every  $x \in \tilde{U}$  and  $k$  large enough. We also obtain that  $\Phi_k$  is one-to-one with the same argument as above. For the remaining results to establish, they simply follow from the uniform convergence of  $\Phi_k$  to  $\Phi$  on every compact subset of  $U$ .  $\square$

## REFERENCES

- [1] A. A. AGRACHEV AND R. V. GAMKRELIDZE, *Symplectic geometry for optimal control*, in Non-Linear Controllability and Optimal Control, Monogr. Textbooks Pure Appl. Math. 133, Dekker, New York, 1990, pp. 263–277.
- [2] A. A. AGRACHEV AND YU. L. SACHKOV, *Control Theory from the Geometric Viewpoint*, Encyclopaedia Math. Sci., 87, Springer, Berlin, 2004.
- [3] R. ALICKI AND K. LENDI, *Quantum Dynamical Semigroups and Applications*, Lecture Notes in Phys. 286, Springer, Berlin, 1987.
- [4] L. ALLEN AND J. H. EBERLY, *Optical Resonance and Two-Level Atoms*, Wiley, New York, 1975.
- [5] U. BOSCAIN AND Y. CHITOUR, *Time-optimal synthesis for left-invariant control systems on  $SO(3)$* , SIAM J. Control Optim., 44 (2005), pp. 111–139.
- [6] U. BOSCAIN AND B. PICCOLI, *On automaton recognizability of abnormal extremals*, SIAM J. Control Optim., 40 (2002), pp. 1333–1357.
- [7] U. BOSCAIN AND B. PICCOLI, *Extremal syntheses for generic planar systems*, J. Dynam. Control Systems, 7 (2001), pp. 209–258.
- [8] U. BOSCAIN AND B. PICCOLI, *Optimal Synthesis for Control Systems on 2-D Manifolds*, Math. Appl. (Berlin) 43, Springer, Berlin, 2004.
- [9] U. BOSCAIN AND P. MASON, *Time minimal trajectories for a spin 1/2 particle in a magnetic field*, J. Math. Phys., 47 (2006), 062101.
- [10] A. BRESSAN AND B. PICCOLI, *Structural stability for time-optimal planar syntheses*, Dyn. Contin. Discrete Impuls. Syst., 3 (1997), pp. 335–371.
- [11] A. BRESSAN AND B. PICCOLI, *A generic classification of time-optimal planar stabilizing feedbacks*, SIAM J. Control Optim., 36 (1998), pp. 12–32.
- [12] M. A. NIELSEN AND I. L. CHUANG, *Quantum Computation and Quantum Information*, Cambridge University Press, London, 2000.
- [13] G. LINDBLAD, *On the generators of quantum dynamical semigroups*, Comm. Math. Phys., 48 (1976), pp. 119–130.
- [14] B. PICCOLI, *Regular time-optimal syntheses for smooth planar systems*, Rend. Sem. Mat. Univ. Padova, 95 (1996), pp. 59–79.
- [15] B. PICCOLI, *Classification of generic singularities for the planar time-optimal synthesis*, SIAM J. Control Optim., 34 (1996), pp. 1914–1946.
- [16] B. PICCOLI AND H. J. SUSSMANN, *Regular synthesis and sufficiency conditions for optimality*, SIAM J. Control Optim., 39 (2000), pp. 359–410.

- [17] L. S. PONTRYAGIN, V. BOLTJANSKI, R. GAMKRELIDZE, AND E. MITCHTCHENKO, *The Mathematical Theory of Optimal Processes*, Wiley, New York, 1961.
- [18] J. PRESKILL, *Lecture Notes*, <http://www.theory.caltech.edu/people/preskill/ph229/#lecture>.
- [19] H. J. SUSSMANN, *The structure of time-optimal trajectories for single-input systems in the plane: The general real analytic case*, SIAM J. Control Optim., 25 (1987), pp. 868–904.
- [20] H. J. SUSSMANN, *Regular synthesis for time-optimal control of single-input real analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145–1162.

## DUBOVITSKII–MILYUTIN APPROACH IN SET-VALUED OPTIMIZATION\*

G. ISAC<sup>†</sup> AND A. A. KHAN<sup>‡</sup>

**Abstract.** By exploring the ideas around the Dubovitskii–Milyutin approach, necessary optimality conditions are given for various optimality notions in set-valued optimization. These optimality conditions are given by using the contingent derivative and the generalized contingent epiderivative of the objective set-valued map and the set-valued maps defining the constraints. The notions of subgradients and scalarized subgradients for set-valued maps are proposed and used to state some regularity conditions.

**Key words.** set-valued optimization, contingent cone, contingent derivative, generalized contingent epiderivative, Aubin’s property, Lagrange multipliers, optimality conditions, subgradients for set-valued maps

**AMS subject classifications.** 90C26, 90C29, 90C30

**DOI.** 10.1137/S0363012904439684

**1. Introduction.** Let  $\Xi$  be a real normed space and let  $\Sigma_i$ ,  $i = 0, 1, \dots, n+1$ , be nonempty convex cones. Let the cones  $\Sigma_i$ ,  $i = 0, 1, \dots, n$ , be open. The Dubovitskii–Milyutin lemma (DM-lemma) asserts an equivalence between the following two statements (see [9]):

- ( $\alpha$ )  $\bigcap_{i=0}^{n+1} \Sigma_i = \emptyset$ .
- ( $\beta$ ) There exist functionals  $l_i \in \Sigma_i^*$  (dual of  $\Sigma_i$ ),  $i = 0, 1, \dots, n+1$ , which are not all simultaneously equal to zero, such that  $l_0 + l_1 + \dots + l_{n+1} = 0$ .

In optimization theory the relevance of the DM-lemma comes from the fact that the optimality of a point can be conveniently expressed as the disjunction of certain sets. Now, if these sets are locally approximated by using suitable tangent cones so that the disjunction is maintained, then, under suitable convexity assumptions, the functionals appearing in ( $\beta$ ) in fact would give rise to the Lagrange multipliers. In applications of the DM-lemma, the cone  $\Sigma_{n+1}$  is often an approximation of the equality constraints, and the cones  $\Sigma_i$ ,  $i = 0, \dots, n$ , contain some information about the behavior of the derivatives of the objective map and the inequality constraints. The approach based on the DM-lemma is very flexible for many applications and has been proved to be of great use in dealing with scalar and vector optimization problems (see [9], [13], [31]).

In this paper, we present an extension of the Dubovitskii–Milyutin approach to the problems which can be written in the following form:

$$(*) \quad \text{minimize}_C F(x) \quad \text{subject to } x \in Q.$$

---

\*Received by the editors January 14, 2004; accepted for publication (in revised form) May 28, 2007; published electronically January 11, 2008.

<http://www.siam.org/journals/sicon/47-1/43968.html>

<sup>†</sup>Department of Mathematics, Royal Military College of Canada, P.O. Box 17000, STN FORCES, Kingston, ON, Canada, K7K 7B4 (isac-g@rmc.ca).

<sup>‡</sup>Department of Mathematics, University of Wisconsin–Barron County, 1800 College Drive, Rice Lake, WI 54868. Current address: Department of Mathematics and Computer Science, Northern Michigan University, 1001 New Science Facility, Marquette, MI 49855 (akhan@nmu.edu). The work of this author was partially supported by the German Research Foundation (DFG) and was partly carried out during the author’s stay at the Institute of Applied Mathematics, University of Erlangen–Nürnberg, Martensstr. 3, 91058 Erlangen, Germany.

Here  $X$  and  $Y$  are real normed spaces (unless stated otherwise all the spaces here will be real),  $Q \subset X$ ,  $C \subset Y$  is a proper pointed convex cone, the map  $F : X \rightrightarrows Y$  is set-valued, and the minimum is taken in the sense that we seek  $(\bar{x}, \bar{y}) \in X \times Y$  such that  $\bar{y} \in F(\bar{x})$  and  $(\bigcup_{x \in Q} F(x)) \cap (\{\bar{y}\} - C) = \{\bar{y}\}$ . We recall that given a pointed convex cone  $C \subset Y$ , the set of minimal points of  $A \subset Y$ , henceforth denoted by  $\text{Min}(A, C)$ , is defined as  $\text{Min}(A, C) = \{x \in A \mid A \cap (\{x\} - C) = \{x\}\}$ . Being equipped with this terminology, a minimizer to  $(*)$  is a point  $(\bar{x}, \bar{y}) \in X \times Y$  such that  $\bar{y} \in F(\bar{x}) \cap \text{Min}(F(Q), C)$ , where  $F(Q) := \bigcup_{x \in Q} F(x)$ .

The following example explains the above notion of minimality.

*Example 1.1.* Consider a set-valued map  $F : [0, 1] \rightrightarrows \mathbb{R}^2$  defined by

$$F(x) := \{(y_1, y_2) \in \mathbb{R}^2 \mid y_1^2 + y_2^2 \leq x^2\}.$$

Let  $C := \{(x, x) \in \mathbb{R}^2 \mid x \in \mathbb{R}_+\}$  be the ordering cone and let  $Q := [0, 1]$ . Then  $F(Q) = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$  and the minima of  $(*)$  form the set  $\{1\} \times \{(\cos \phi, \sin \phi) \mid \phi \in [\frac{3}{4}\pi, \frac{7}{4}\pi]\}$ . On the other hand, if the ordering cone is

$$C := \{(x, y) \in \mathbb{R}^2 \mid x, y \in \mathbb{R}_+\},$$

then the minima of  $(*)$  form the set  $\{1\} \times \{(\cos \phi, \sin \phi) \mid \phi \in [\pi, \frac{3}{2}\pi]\}$ .

Notice that if the map  $F$  is single-valued, then  $(*)$  collapses to the known vector optimization problem (see [18]). Additionally, if  $Y = \mathbb{R}$  and  $C = \mathbb{R}_+ = \{t \in \mathbb{R} \mid t \geq 0\}$ , then we recover the framework of classical optimization problems (see [9]).

Problems of the type shown above belong to the realm of set-valued optimization, a subject which has attracted a great deal of attention in recent years (cf. [3], [4], [5], [6], [7], [8]). In general, set-valued optimization represents the optimization problems with set-valued objective and/or set-valued constraints. In recent years set-valued optimization has emerged as an important generalization of scalar and vector optimization. Moreover, there are many research areas which directly lead to the type of problem shown above. We illustrate this by depicting the appearance of set-valued optimization problems in connection with the duality principles in vector optimization and the gap functions for vector variational inequalities [12], [17].

*Example 1.2* (duality principles). Consider the following vector optimization problem:

$$(**) \quad \text{minimize}_Q \varphi(x) \quad \text{subject to} \quad x \in \Delta_1 = \{x \in \Delta \mid -\psi(x) \in Q_1\},$$

where  $\Delta \subset \mathbb{R}^n$ ,  $Q \subset \mathbb{R}^p$ , and  $Q_1 \subset \mathbb{R}^m$  are pointed closed convex cones and  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^p$  and  $\psi : \mathbb{R}^n \rightarrow \mathbb{R}^m$  are single-valued mappings. Let  $\mathcal{L}$  be the family of  $p \times m$  matrices  $\Lambda$  such that  $\Lambda Q_1 \subset Q$  and let  $L : \Delta \times \mathcal{L} \rightarrow \mathbb{R}^p$  be the vector-valued Lagrangian given by  $L(x, \Lambda) = \varphi(x) + \Lambda \psi(x)$ . We define a set-valued map  $\Phi : \mathcal{L} \rightrightarrows \mathbb{R}^p$  as follows:

$$\Phi(\Lambda) = \text{Min}(\{L(x, \Lambda) \mid x \in \Delta\}, Q).$$

Then the dual problem associated with  $(**)$  is the following set-valued optimization problem:

$$\text{maximize}_Q \Phi(\Lambda) \quad \text{subject to} \quad \Lambda \in \mathcal{L}.$$

Details on the above model are available in [33]. An infinite-dimensional analogue of this model has been studied by Corley (see the references in [8]).

*Example 1.3* (gap functions). Let  $X$  and  $Y$  be normed spaces and let  $\mathcal{L}(X, Y)$  be the set of all linear continuous operators from  $X$  to  $Y$ . Let  $Q \subset X$  be a nonempty closed convex set, let  $Q_0 \subset X$  be a pointed closed convex cone, and let  $\Gamma : X \rightarrow Y$  be a given map.

Consider the following vector variational inequality (VVI): Find  $\bar{x} \in Q$  such that

$$\langle \Gamma(\bar{x}), x - \bar{x} \rangle \notin -Q_0 \setminus \{0\} \quad \text{for every } x \in Q.$$

We define a set-valued mapping  $\Phi : X \rightrightarrows Y$  by

$$\Phi(x) = \text{Max}(\{\langle \Gamma(\bar{x}), x - Q \rangle, x \in Q\}, Q_0).$$

Then the above VVI is equivalent to the following set-valued optimization problem (cf. [24]):

$$\text{minimize}_{Q_0} \Phi(x) \quad \text{subject to } x \in Q.$$

Another genuine source for such problems is the situation in mathematical programming when the objective and the constraint functions cannot be assigned to some exact values but are allowed to vary in a specified range, hence leading to set-valued data (for example, in stochastic and fuzzy programming). In fuzzy programming this fact is usually characterized by a membership function. However, in the general setting of set-valued optimization the use of the membership function can be avoided as the whole sets are taken into account (cf. [8], [32]). Several other interesting appearances of set-valued optimization are in sensitivity analysis for vector optimization problems (cf. [34]).

In recent years a great deal of attention has been given to set-valued optimization. The starting point for this interesting research domain was the influential paper by Corley [8], where the contingent derivative and the circatangent derivative were used to give general optimality conditions. His results were substantially improved by Luc and Malivert [26] (see also [25]). In the aforementioned works, the derivative notions revolved around the graphs of the involved set-valued maps. Another useful approach based on employing the epigraphs of the involved set-valued maps was initiated by Jahn and Rauh [21], which was further pursued in [3], [11], [14], [19], [20], [23], among others. An interesting approach for set-valued optimization which is based on the notion of coderivatives is given in [10].

As already mentioned above, the main goal of this paper is to present some necessary optimality conditions in set-valued optimization by using the ideas surrounding the Dubovitskii–Milyutin approach. The optimality conditions given here can be divided into two categories. In the first category, we express the optimality of a point as an empty intersection of certain sets in the domain space. Although this result is completely analogous to the results in classical nonlinear programming, it is in contrast with the results obtained so far in set-valued optimization, where the optimality conditions of this nature have always been given in the image space. The second type of optimality conditions which we give here is comparable to the classical KKT optimality conditions. A special emphasis is given to the case when the ordering cone may have an empty interior. As concerns the solvability of set-valued optimization problems, it is known that the notion of a solution in set-valued optimization is not unequivocal and there are many solution concepts. In this paper we present necessary optimality conditions for minimality, proper minimality, weak minimality, and strong minimality. However, keeping in mind the narrow gap among the techniques used



for deriving these optimality conditions for various minimality notions, we present a detailed treatment only for proper minimality. For other optimality notions, we either state the results without a proof or we just mark the differences whenever it becomes necessary.

The contents of this paper are divided into five sections. In section 2, we briefly recall some notation and concepts to be used and formulate the problem. Basic notions which we recall include various definitions of tangent cones and derivatives and epiderivatives for set-valued maps. We devote section 3 to the necessary optimality conditions for proper and weak minimality. In this section we also introduce several subgradients for set-valued maps. However, the use of these subgradients is limited mainly to expressing some regularity conditions. Section 4 is devoted to the necessary optimality conditions for the strong minimality. The final section contains some remarks concerning the approach.

**2. Preliminaries and problem formulation.** First we recall the notions of some tangent cones (cf. [2], [30]). We set  $\mathbb{P} := \{t \in \mathbb{R} \mid t > 0\}$ . In the following  $\text{cl}(S)$  and  $\text{int}(S)$  will represent the closure and the interior of a set  $S$ , respectively.

**DEFINITION 2.1.** *Let  $Z$  be a normed space, let  $S \subset Z$ , and let  $\bar{z} \in S$ .*

- (a) *The contingent cone  $T(S, \bar{z})$  of  $S$  at  $\bar{z}$  is the set of all  $z \in Z$  such that there are sequences  $(\lambda_n) \subset \mathbb{P}$  and  $(z_n) \subset Z$  with  $\lambda_n \downarrow 0$ ,  $z_n \rightarrow z$ , and  $\bar{z} + \lambda_n z_n \in S$  for all  $n \in \mathbb{N}$ .*
- (b) *The interiorly contingent cone  $IT(S, \bar{z})$  of  $S$  at  $\bar{z}$  is the set of all  $z \in Z$  such that for any sequences  $(\lambda_n) \subset \mathbb{P}$  and  $(z_n) \subset Z$  with  $\lambda_n \downarrow 0$  and  $z_n \rightarrow z$  there exists an integer  $m \in \mathbb{N}$  such that  $\bar{z} + \lambda_n z_n \in S$  for all  $n \geq m$ .*
- (c) *The Clarke's tangent cone  $C(S, \bar{z})$  of  $S$  at  $\bar{z}$  is the set of all  $z \in Z$  such that for every sequence  $(z_n) \subset S$  with  $z_n \rightarrow \bar{z}$  and for every  $(\lambda_n) \subset \mathbb{P}$  with  $\lambda_n \downarrow 0$  there exists a sequence  $(\tilde{z}_n)$  with  $\tilde{z}_n \rightarrow z$  such that  $z_n + \lambda_n \tilde{z}_n \in S$  for all  $n \in \mathbb{N}$ .*

**Remark 2.1.** It is known that  $T(S, \bar{z})$  is a closed cone possessing the isotony property; that is, for subsets  $S_1$  and  $S_2$  such that  $S_1 \subset S_2$ , we have  $T(S_1, \bar{z}) \subset T(S_2, \bar{z})$  for every  $\bar{z} \in S_1$ . The Clarke's tangent cone is closed and convex, but it is not isotone. Moreover, the inclusion  $C(S, \bar{z}) \subseteq T(S, \bar{z})$  holds. On the other hand the interiorly contingent cone  $IT(S, \bar{z})$  is an isotone open cone. As concerns the relationship between  $T(S, \bar{z})$  and  $IT(S, \bar{z})$ , we have  $IT(S, \bar{z}) = Z \setminus T(Z \setminus S, \bar{z})$ . As a useful implication of this relationship, the cones  $T(S, \bar{z})$  and  $IT(S, \bar{z})$  form an admissible pair; that is, for every pair of sets  $S_1, S_2 \subset Z$  with  $S_1 \cap S_2 = \emptyset$ , we have  $T(S_1, \bar{z}) \cap IT(S_2, \bar{z}) = \emptyset$  for every  $\bar{z} \in Z$ . Also for arbitrary sets  $S_1, S_2 \subset Z$  we have  $IT(S_1 \cap S_2, \bar{z}) = IT(S_1, \bar{z}) \cap IT(S_2, \bar{z})$  for every  $\bar{z} \in S_1 \cap S_2$ . In general, this property is not shared by the contingent cones. For some  $S \subset Z$ , the identities  $T(S, \bar{z}) = T(\text{cl}(S), \bar{z})$  and  $IT(S, \bar{z}) = IT(\text{int}(S), \bar{z})$  hold. Moreover, for a convex solid set  $S$ , we have  $\text{cl}(IT(S, \bar{z})) = T(S, \bar{z})$  and  $\text{int}(T(S, \bar{z})) = IT(S, \bar{z})$ . For details see [2], [29], [30].

Now we collect some definitions for set-valued maps. Let  $X$  and  $Y$  be normed spaces. Let  $F : X \rightrightarrows Y$  be a set-valued map; that is, for each  $x \in X$ , we have  $F(x) \subset 2^Y$  (power set of  $Y$ ). The (effective) domain and the graph of  $F$  are defined by  $\text{dom}(F) := \{x \in X \mid F(x) \neq \emptyset\}$  and  $\text{gph}(F) := \{(x, y) \in X \times Y \mid y \in F(x)\}$ , respectively. We shall say that  $F$  is *strict* if  $\text{dom}(F) = X$ . We define the *weak-inverse image*  $F[S]^-$  of  $F$  with respect to  $S \subseteq Y$  by  $F[S]^- := \{x \in X \mid F(x) \cap S \neq \emptyset\}$ . If  $Y$  is partially ordered by a convex cone  $C \subset Y$ , the profile map  $F_+ : X \rightrightarrows Y$  is given by  $F_+(x) := F(x) + C$  for every  $x \in \text{dom}(F)$ . In this setting, the epigraph of  $F$

can be defined as the graph of  $F_+$ , that is,  $\text{epi}(F) = \text{gph}(F_+)$ . The map  $F$  is called *convex* if  $\text{gph}(F)$  is convex and *C-convex* if  $\text{epi}(F)$  is convex. We shall say that the map  $F$  is *locally C-convex* at  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  if  $T(\text{epi}(F), (\bar{x}, \bar{y}))$  is convex. Let  $B_Y$  be the unit ball of the space  $Y$ . The map  $F$  is said to have the *Aubin property around*  $(u, v) \in \text{gph}(F)$  if there are a constant  $L \geq 0$  and neighborhoods  $U$  of  $u$  and  $V$  of  $v$  so that

$$F(x_1) \cap V \subseteq F(x_2) + L\|x_1 - x_2\|B_Y \quad \text{for all } x_1, x_2 \in U \cap \text{dom}(F).$$

This concept is due to Aubin. For several useful features of this notion, see [2], [28], [32].

The contingent derivative, given by Aubin [1], is at the heart of our approach. We recall that given normed spaces  $X, Y$  and a set-valued map  $F : X \rightrightarrows Y$ , the contingent derivative of  $F$  at  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  is the set-valued map  $\mathcal{D}F(\bar{x}, \bar{y}) : X \rightrightarrows Y$  defined by

$$(1) \quad \text{gph}(\mathcal{D}F(\bar{x}, \bar{y})) = T(\text{gph}(F), (\bar{x}, \bar{y})).$$

A related notion where the center of attraction, instead of  $\text{gph}(F)$ , is  $\text{epi}(F)$  is of the generalized contingent epiderivative proposed in [7] and [4]. We recall that the generalized contingent epiderivative of  $F$  at  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  is the set-valued map  $D_g F(\bar{x}, \bar{y}) : X \rightrightarrows Y$  given by

$$(2) \quad D_g F(\bar{x}, \bar{y})(x) = \text{Min}(\mathcal{D}F_+(x), C), \quad x \in \text{dom}(\mathcal{D}F_+(\bar{x}, \bar{y})).$$

We say that epiderivative  $D_g F(\bar{x}, \bar{y})$  *dominates* at  $x \in \text{dom}(\mathcal{D}F_+(\bar{x}, \bar{y}))$  if the inclusion  $\mathcal{D}F_+(\bar{x}, \bar{y})(x) \subseteq D_g F(\bar{x}, \bar{y})(x) + C$  holds. It is known that if  $D_g F(\bar{x}, \bar{y})$  dominates at all  $x \in \text{dom}(\mathcal{D}F_+(\bar{x}, \bar{y}))$ , then

$$(3) \quad \text{epi}(D_g F(\bar{x}, \bar{y})) = T(\text{epi}(F), (\bar{x}, \bar{y})).$$

Conditions ensuring the existence and the domination of the above epiderivative are given in [4] and [20]. For example, it is known that if  $C$  is regular (see [16]) and  $\mathcal{D}F_+(\bar{x}, \bar{y})(x)$  is  $C$ -lower-bounded, then  $D_g F(\bar{x}, \bar{y})(x) \neq \emptyset$  and the epiderivative dominates at  $x$  (see [20]). An important difference between the contingent derivative and the generalized contingent epiderivative is that in the latter case we concentrate only on a boundary part of the underlying contingent cone.

*Remark 2.2.* Replacing  $T(\text{gph}(F), (\bar{x}, \bar{y}))$  by  $C(\text{gph}(F), (\bar{x}, \bar{y}))$  in (1), we get circatangent derivative  $\mathcal{C}F(\bar{x}, \bar{y})$  of  $F$  at  $(\bar{x}, \bar{y})$  (cf. [2]). We define generalized circatangent epiderivative  $\mathcal{C}_g F(\bar{x}, \bar{y})$  of  $F$  at  $(\bar{x}, \bar{y})$  by taking  $\mathcal{C}F_+(\bar{x}, \bar{y})(x)$  for  $\mathcal{D}F_+(\bar{x}, \bar{y})$  in (2).

Besides the minimality notion defined above, we also need some other related concepts (cf. [18, 25]). Let  $Y$  be a normed space partially ordered by a proper pointed convex cone  $C \subset Y$ . Given  $C$ , let  $\mathcal{K}$  be the set of all proper pointed closed convex solid cones  $K \subset Y$  such that  $C \setminus \{0_Y\} \subset \text{int}(K)$ . Let  $D \subset Y$  and let  $y \in D$  be arbitrary.

- An element  $y$  is said to be a strongly minimal point of  $D$  if  $D \subseteq \{y\} + C$ .
- Assume that the ordering cone  $C$  is solid, that is, it has a nonempty interior  $\text{int}(C)$ . An element  $y$  is said to be a weakly minimal point of  $D$  if  $D \cap (\{y\} - \text{int}(C)) = \emptyset$ .
- An element  $y$  is said to be a properly minimal point of  $D$  if for some  $K \in \mathcal{K}$  we have  $D \cap (\{y\} - K) = \{y\}$ , that is, the element  $y$  is a minimal point of  $D$  with respect to  $K$ .

The set of all strongly minimal points, weakly minimal points, and properly minimal points of  $D$  with respect to  $C$  will be denoted by  $\text{SMin}(D, C)$ ,  $\text{WMin}(D, C)$ , and  $\text{PMin}(D, C)$ . Moreover, the following chain of inclusions is known to hold:

$$\text{SMin}(D, C) \subseteq \text{PMin}(D, C) \subseteq \text{Min}(D, C) \subseteq \text{WMin}(D, C).$$

We remark that there are a few more notions of minimal points in set optimization which we do not explore in this work (cf. [6], [22], [27]).

Let  $W, X, Y, Z$  be normed spaces and let the spaces  $Y$  and  $Z$  be partially ordered by nontrivial pointed closed convex cones  $C \subset Y$  and  $D \subset Z$ . Let  $Q_0 \subset X$  be nonempty and let  $h \in W$  be a given element. Let  $F : X \rightrightarrows Y$ ,  $G : X \rightrightarrows Z$ , and  $H : X \rightrightarrows W$  be given set-valued maps.

We are concerned with the following set-valued optimization problems:

$$\begin{aligned} (P_0) \quad & \text{Min } F(x) \quad \text{subject to} \quad x \in Q_0. \\ (P_1) \quad & \text{Min } F(x) \quad \text{subject to} \quad x \in Q_1 := \{x \in Q_0 \mid G(x) \cap -D \neq \emptyset\}. \\ (P_2) \quad & \text{Min } F(x) \quad \text{subject to} \quad x \in Q_2 := \{x \in Q_0 \mid G(x) \cap -D \neq \emptyset, h \in H(x)\}. \end{aligned}$$

In the following, we define the optimality for  $(P_2)$ . We set  $F(Q_2) := \bigcup_{x \in Q_2} F(x)$ .

**DEFINITION 2.2.** A pair  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  is called a (a) minimizer of  $(P_2)$  if  $\bar{y} \in \text{Min}(F(Q_2), C)$ ; (b) strong minimizer of  $(P_2)$  if  $\bar{y} \in \text{SMin}(F(Q_2), C)$ ; (c) weak minimizer of  $(P_2)$  if  $\bar{y} \in \text{WMin}(F(Q_2), C)$ ; (d) proper minimizer of  $(P_2)$  if  $\bar{y} \in \text{PMin}(F(Q_2), C)$ .

In view of the definition of the minimal points,  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  is a minimizer if and only if  $F(Q_2) \cap (\bar{y} - C) = \{\bar{y}\}$ . Similar relations are valid for other kinds of minimizers defined above. Observe that  $(P_2)$  reduces to  $(P_0)$  if  $G(x) = 0_Z$  and  $H(x) = h$  uniformly on  $Q_0$ . In this case the set of constraints  $Q_0$  is not explicitly specified. If additionally we have  $Q_0 = Z$ , then  $(P_2)$  is an unconstrained set-valued optimization problem. The optimality notions given in the above definition are global ones; that is, the whole set  $F(Q_2)$  has been taken into account. Their local versions are defined as follows: The point  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  is said to be a local strong minimizer if there exists a neighborhood  $U$  of  $\bar{x}$  such that  $\bar{y} \in \text{SMin}(F(Q_2 \cap U), C)$ . Local minimizers, local weak minimizers, and local proper minimizers are defined analogously by using Definition 2.2.

Notice that in the context of Example 1.1, the proper minima of  $(P_0)$  contain the set  $\{1\} \times \{(\cos \phi, \sin \phi) \mid \phi \in [\frac{3}{4}\pi, \frac{3}{2}\pi]\}$ . There is no strong minimizer.

### 3. Necessary optimality conditions for proper and weak minimality.

**3.1. Necessary optimality conditions for  $(P_1)$ .** We begin with the following necessary optimality condition for  $(P_1)$ .

**THEOREM 3.1.** Let  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  be a local proper minimizer to  $(P_1)$ . Assume that the map  $F_+ (:= F + C)$  possesses the Aubin property around  $(\bar{x}, \bar{y})$ . Then

$$(4) \quad \mathcal{D}F_+(\bar{x}, \bar{y})[-C \setminus \{0_Y\}]^- \bigcap IT(Q_0, \bar{x}) \bigcap T(G[-D]^-, \bar{x}) = \emptyset.$$

*Proof.* Since  $(\bar{x}, \bar{y})$  is a local proper minimizer, there are a neighborhood  $U_1$  of  $\bar{x}$  and a pointed closed convex solid cone  $K \in \mathcal{K}$  such that  $\bar{y} \in \text{Min}(F(Q_1 \cap U_1), K)$ . We claim that

$$(5) \quad \mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^- \bigcap IT(Q_0, \bar{x}) \bigcap T(G[-D]^-, \bar{x}) = \emptyset.$$

Notice that since  $\text{int}(K)$  contains  $C \setminus \{0_Y\}$ , (5) trivially implies (4).

We show that if (5) fails, then a feasible  $u$  can be obtained in a vicinity of  $\bar{x}$  with  $F(u) \cap (\bar{y} - \text{int}(K)) \neq \emptyset$ , hence violating the local proper minimality of  $(\bar{x}, \bar{y})$ . Assume that there exists

$$x \in \mathcal{DF}_+(\bar{x}, \bar{y})[-\text{int}(K)]^- \bigcap IT(Q_0, \bar{x}) \bigcap T(G[-D]^-, \bar{x}).$$

Then in view of the containment  $x \in T(G[-D]^-, \bar{x})$ , there are sequences  $(\lambda_n) \subset \mathbb{P}$  and  $(\tilde{x}_n) \subset X$  such that  $\lambda_n \downarrow 0$ ,  $\tilde{x}_n \rightarrow x$ , and  $G(\bar{x} + \lambda_n \tilde{x}_n) \cap (-D) \neq \emptyset$  for every  $n \in \mathbb{N}$ .

Since we also have  $x \in \mathcal{DF}_+(\bar{x}, \bar{y})[-\text{int}(K)]^-$ , there exists  $y \in -\text{int}(K)$  such that  $(x, y) \in T(\text{epi}(F), (\bar{x}, \bar{y}))$ . Therefore, there are sequences  $(\beta_n) \subset \mathbb{P}$  and  $((\tilde{x}_n, \tilde{y}_n)) \subset X \times Y$  such that  $\beta_n \downarrow 0$ ,  $(\tilde{x}_n, \tilde{y}_n) \rightarrow (x, y)$ , and  $\bar{y} + \beta_n \tilde{y}_n \in F(\bar{x} + \beta_n \tilde{x}_n) + C$  for every  $n \in \mathbb{N}$ .

Because  $(\lambda_n)$  and  $(\beta_n)$  are sequences of strictly positive reals, both converging to zero, there exist increasing sequences of positive integers  $(k_n)$  and  $(m_n)$  such that  $\beta_{m_n}/\lambda_{k_n} \rightarrow 1$  (see [24, Lemma 2.1]). We pick subsequences  $(\beta_{m_n}) \subset (\beta_n)$ ,  $(x_{m_n}) \subset (\tilde{x}_n)$ ,  $(y_{m_n}) \subset (\tilde{y}_n)$  and set  $x_n = (\beta_{m_n} \setminus \lambda_{k_n})x_{m_n}$ ,  $y_n = (\beta_{m_n} \setminus \lambda_{k_n})y_{m_n}$ , and  $\alpha_n = \lambda_{k_n}$ . Clearly  $x_n \rightarrow x$ ,  $y_n \rightarrow y$ ,  $\alpha_n \downarrow 0$  and we have  $\bar{y} + \alpha_n y_n \in F(\bar{x} + \alpha_n x_n) + C$  for every  $n \in \mathbb{N}$ . We also pick a subsequence  $(x_{k_n}) \subset (\tilde{x}_n)$  and set  $\hat{x}_n = x_{k_n}$ . Notice that  $\hat{x}_n \rightarrow x$  and we have  $G(\bar{x} + \alpha_n \hat{x}_n) \cap (-D) \neq \emptyset$  for every  $n \in \mathbb{N}$ . Because  $u_n := (\bar{x} + \alpha_n x_n) \rightarrow \bar{x}$  and  $\tilde{u}_n := (\bar{x} + \alpha_n \hat{x}_n) \rightarrow \bar{x}$ , there exists  $n_1 \in \mathbb{N}$  such that  $u_n, \tilde{u}_n \in U$  for  $n \geq n_1$ . Here  $U := U_1 \cap U_2$  and  $U_2$  is a neighborhood of  $\bar{x}$  which exists, along with a neighborhood  $V$  of  $\bar{y}$ , as a consequence of the Aubin property. Moreover, since  $(\bar{y} + \alpha_n y_n) \rightarrow \bar{y}$ , there exists  $n_2 \in \mathbb{N}$  so that  $\bar{y} + \alpha_n y_n \in V$  for all  $n \geq n_2$ . In view of the Aubin property of  $F_+$  at  $(\bar{x}, \bar{y})$ , we get

$$\begin{aligned} \bar{y} + \alpha_n y_n &\in [F(\bar{x} + \alpha_n x_n) + C] \cap V \quad (\text{for } n \geq n_2) \\ &\subseteq F(\bar{x} + \alpha_n \hat{x}_n) + C + L \alpha_n \|x_n - \hat{x}_n\| B_Y \quad (\text{for } n \geq \max\{n_1, n_2\}), \end{aligned}$$

and hence we can choose a sequence  $b_n \in B_Y$  such that for  $n \geq \max\{n_1, n_2\}$  we have  $\bar{y} + \alpha_n \tilde{y}_n \in F(\bar{x} + \alpha_n \hat{x}_n) + C$ , where  $\tilde{y}_n = (y_n - L b_n \|x_n - \hat{x}_n\|) \rightarrow y$ . Because  $y \in -\text{int}(K)$ ,  $\alpha_n > 0$ , and  $\tilde{y}_n \rightarrow y$  there exists  $n_3 \in \mathbb{N}$  such that  $\alpha_n \tilde{y}_n \in -\text{int}(K)$ . Choose  $w_n \in F(\tilde{u}_n)$  such that  $\bar{y} + \alpha_n \tilde{y}_n \in w_n + C$  for every  $n \in \mathbb{N}$ . We have  $w_n \in \bar{y} + \lambda_n \tilde{y}_n - C \subseteq \bar{y} - \text{int}(K) - C \subset \bar{y} - \text{int}(K)$  and consequently  $w_n \in F(\tilde{u}_n) \cap (\bar{y} - \text{int}(K))$  for  $n \geq \max\{n_1, n_2, n_3\}$ .

Finally we notice that because  $\alpha_n \downarrow 0$  and  $\hat{x}_n \rightarrow x$ , the containment  $x \in IT(Q_0, \bar{x})$  implies that there is an integer  $n_4 \in \mathbb{N}$  such that  $\tilde{u}_n := \bar{x} + \alpha_n \hat{x}_n \in Q_0$  for every  $n \geq n_4$ . Therefore, we have shown that for every  $n \geq \max\{n_1, n_2, n_3, n_4\}$  there are  $\tilde{u}_n \in Q_1 \cap U_1$  such that  $F(\tilde{u}_n) \cap (\bar{y} - \text{int}(K)) \neq \emptyset$ . This, however, contradicts the optimality of  $(\bar{x}, \bar{y})$ . The proof is complete.  $\square$

*Remark 3.1.* If either  $G[-D]^-$  is derivable at  $\bar{x}$  or  $\text{epi}(F)$  is derivable at  $(\bar{x}, \bar{y})$ , then the mechanism of choosing subsequences is unnecessary. Here, by the derivability of  $S \subset Z$  (normed space) at  $\bar{z} \in \text{cl}(S)$ , we mean  $T(S, \bar{z}) = A(S, \bar{z})$ , where  $A(S, \bar{z})$  is the adjacent cone given by  $A(S, \bar{z}) := \{y \in Z \mid \forall (\lambda_n) \subset \mathbb{P}, \lambda_n \downarrow 0 \exists (y_n) \rightarrow y : \bar{z} + \lambda_n y_n \in S \forall n \in \mathbb{N}\}$  (cf. [2]).

In view of the proof of Theorem 3.1 the following useful corollaries are immediate.

**COROLLARY 3.1.** *Let  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  be a local proper minimizer to  $(P_1)$ . Let either  $\text{dom}(F) = \text{dom}(G) = Q_0$  or  $\bar{x} \in \text{int}(Q_0)$  and let  $F_+$  possess the Aubin property around  $(\bar{x}, \bar{y})$ . Then*

$$\mathcal{DF}_+(\bar{x}, \bar{y})[-C \setminus \{0_Y\}]^- \bigcap T(G[-D]^-, \bar{x}) = \emptyset.$$

COROLLARY 3.2. *Let  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  be a local proper minimizer to  $(P_0)$ . Then*

$$IT(Q_0, \bar{x}) \bigcap \text{cl}(\mathcal{D}F_+(\bar{x}, \bar{y})[-C \setminus \{0_Y\}]^-) = \emptyset.$$

Moreover, if either  $\text{dom}(F) = Q_0$  or  $\bar{x} \in \text{int}(Q_0)$ , then

$$\mathcal{D}F_+(\bar{x}, \bar{y})[-C \setminus \{0_Y\}]^- = \emptyset.$$

*Remark 3.2.* Notice that all the above results remain valid if instead of  $\mathcal{D}F_+(\bar{x}, \bar{y})$  we take either  $\mathcal{C}F_+(\bar{x}, \bar{y})$  or  $D_g F(\bar{x}, \bar{y})$ . Moreover, we remark that all the above results in fact will characterize the weak minimality if we replace  $C \setminus \{0_Y\}$  by  $\text{int}(C)$ , provided that  $\text{int}(C) \neq \emptyset$ . Also notice that for weak minimality, Corollary 3.2 states that  $\mathcal{D}F_+(\bar{x}, \bar{y})(x) \subset Y \setminus \text{int}(C)$  for every  $x \in \text{dom}(\mathcal{D}F_+(\bar{x}, \bar{y}))$ . In fact, in most of the available results, this necessary optimality condition is proved by direct arguments, which, primarily because of the definition of the derivative, demands that  $\text{dom}(F) = Q_0$ . See [4], [21] for details.

Our above results are suitable when the cone  $C$  is not solid and when the set-valued maps are not necessarily confined to  $Q_0$ . We illustrate this by slightly modifying Example 1.1.

*Example 3.1.* Consider again the set-valued map  $F : [0, 2] \rightrightarrows \mathbb{R}^2$  defined by

$$F(x) := \{(y_1, y_2) \in \mathbb{R}^2 \mid y_1^2 + y_2^2 \leq x^2\}.$$

Let  $C := \{(x, x) \in \mathbb{R}^2 \mid x \in \mathbb{R}_+\}$  be the ordering cone and let  $Q_0 := [0, 1]$  be the set of implicit constraints. It is easy to verify that the point  $(1, (0, -1)) \in \text{gph}(F)$  is a (global) proper minimizer to  $(P_0)$ . For  $x \in \text{dom}(D_g F(1, (0, -1))) = \mathbb{R}$ , we have

$$D_g F(1, (0, -1))(x) = \{(y_1, y_2) \in \mathbb{R}^2 \mid y_1 \in \mathbb{R}, y_2 = -x\}$$

and hence  $D_g F(1, (0, -1))(x) \cap (-C \setminus \{0_{\mathbb{R}^2}\}) \neq \emptyset$  for every  $x \in \mathbb{P}$ . However, since  $IT(Q_0, \bar{x}) = -\mathbb{P}$  and  $D_g F(1, (0, -1))[-C \setminus \{0_{\mathbb{R}^2}\}]^- = \mathbb{P}$ , the necessary optimality condition announced in Corollary 3.2 holds good as

$$D_g F(1, (0, -1))[-C \setminus \{0_{\mathbb{R}^2}\}]^- \bigcap IT(Q_0, \bar{x}) = \emptyset.$$

Another approach to prove (4), in fact more akin to the classical Dubovitskii–Milyutin approach, is to express the optimality as a disjunction of certain sets and then use some Farkas lemma-type results. In the present setting, the following two results generalize the Dubovitskii–Milyutin approach and give an alternative proof of Theorem 3.1.

LEMMA 3.1. *Let  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  be a local proper minimizer to  $(P_1)$ . Then*

$$(6) \quad IT(Q_0, \bar{x}) \bigcap IT(F[\bar{y} - C \setminus \{0_Y\}]^-, \bar{x}) \bigcap T(G[-D]^-, \bar{x}) = \emptyset.$$

*Proof.* We claim that  $U \bigcap Q_0 \bigcap F[\bar{y} - \text{int}(K)]^- \bigcap G[-D]^- = \emptyset$ , where  $U$  is a neighborhood of  $\bar{x}$  and  $K \in \mathcal{K}$ , both corresponding to the local proper minimality. In fact, if there exists  $x \in U \bigcap Q_0 \bigcap F[\bar{y} - \text{int}(K)]^- \bigcap G[-D]^-$ , then from  $x \in U \bigcap Q_0 \bigcap G[-D]^-$  we notice that  $x$  is feasible and from  $x \in F[\bar{y} - \text{int}(K)]^-$  we obtain  $F(x) \cap (\bar{y} - \text{int}(K)) \neq \emptyset$ , which is a contradiction of the optimality of  $(\bar{x}, \bar{y})$ . The assertion now follows from the properties of the interiorly contingent cones and the contingent cones stated in Remark 2.1. The proof is complete.  $\square$

Besides giving an alternative proof for (4), the following result has its own significance.

**PROPOSITION 3.1.** *Let  $X$  and  $Y$  be normed spaces, and let  $F : X \rightrightarrows Y$  be a set-valued map possessing the Aubin property around  $(\bar{x}, \bar{y}) \in \text{gph}(F)$ . Then for every  $K \in \mathcal{K}$  we have*

$$\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^- \subseteq IT(F[\bar{y} - \text{int}(K)]^-, \bar{x}).$$

*Proof.* The proof is based on the arguments given in the proof of Theorem 3.1.  $\square$

Let  $\text{dom}(F) = \text{dom}(G) = Q_0$ , let  $(\bar{x}, \bar{y}) \in \text{gph}(F)$ , and let  $\bar{z} \in G(\bar{x}) \cap (-D)$ . Then a necessary optimality condition for  $(\bar{x}, \bar{y})$  to be a *weak minimizer* to  $(P_1)$  is that (cf. [8])

$$\mathcal{D}(F, G)(\bar{x}, \bar{y}, \bar{z})(S) \cap (-\text{int}(C) \times (-\text{int}(D) - \bar{z})) = \emptyset,$$

where  $\mathcal{D}(F, G)(\bar{x}, \bar{y}, \bar{z})$  is the contingent derivative of the map  $(F, G) := F \times G$  at  $(\bar{x}, \bar{y}, \bar{z})$  and  $S$  is the domain of this derivative. It is clear that under suitable convexity assumptions these disjoint sets, given in the image space, can be separated and a multiplier rule can be obtained (see [14]). However, the optimality condition (4) is of different nature. First, the disjunction in (4) is taking place in the objective space. Second, in (4) we have not taken into account any derivative of  $G$ . To get a multiplier rule from Theorem 3.1, we need to apply the DM-lemma to (4). For this, among other things, we need to gather some information about the dual of  $\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^-$  and a suitable regularity condition relating the cone  $T(G[-D]^-, \bar{x})$  with some derivative of  $G$  at a feasible point. All this is done in the next section.

**3.2. Inverse images and subgradients of set-valued maps.** Let  $X$  be a normed space, let  $X^*$  be the topological dual of  $X$ , and let  $M \subset X$ . The polar  $M^\circ$  of  $M$  is a subset of  $X^*$  defined by  $M^\circ = \{l \in X^* : l(x) \leq 0 \text{ for every } x \in M\}$ . It is known that if  $M_1 \subseteq M_2$ , then  $M_2^\circ \subseteq M_1^\circ$ . The positive dual  $M^*$  is defined by  $M^* = -M^\circ$ .

Given a set  $\mathcal{A} \subset \mathbb{R}$  and  $b \in \mathbb{R}$ , by the inequality  $\mathcal{A} \geq b$  we understand that  $a \geq b$  for every  $a \in \mathcal{A}$ . Now let  $X$  and  $Y$  be normed spaces, let  $C \subset Y$  be a proper pointed convex cone and let  $F : X \rightrightarrows Y$  be set-valued with  $F_+ = F + C$  as its profile map. Given  $A \subset Y^*$ , we introduce a scalarized subgradient  $\partial_A F(\bar{x}, \bar{y})$  of  $F$  at  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  as follows:

$$\partial_A F(\bar{x}, \bar{y}) = \{L \in X^* \mid \exists y^* \in A : L(x) \leq (y^* \circ \mathcal{D}F_+(\bar{x}, \bar{y}))(x) \forall x \in \text{dom}(\mathcal{D}F_+(\bar{x}, \bar{y}))\}.$$

Further, given  $B \subset Y$ , we define a  $B$ -subgradient  $\partial_B F(\bar{x}, \bar{y})$  of  $F$  at  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  by

$$\partial_B F(\bar{x}, \bar{y}) = \{L \in \mathcal{L}(X, Y) \mid (\mathcal{D}F_+(\bar{x}, \bar{y})(x) - L(x)) \cap B = \emptyset \forall x \in \text{dom}(\mathcal{D}F_+(\bar{x}, \bar{y}))\},$$

where  $\mathcal{L}(X, Y)$  represents the set of all linear and continuous functions defined from  $X$  to  $Y$ . We shall use the terms generalized subgradient, proper subgradient, and weak subgradient if  $B = -C \setminus \{0_Y\}$ ,  $-\text{int}(K)$ , where  $K \subset \mathcal{K}$ , and  $-\text{int}(C)$ , and we denote these variants by  $\partial_g F(\bar{x}, \bar{y})$ ,  $\partial_p F(\bar{x}, \bar{y})$ , and  $\partial_w F(\bar{x}, \bar{y})$ , respectively. If in the above definition of  $B$ -subgradient we replaced  $\mathcal{D}F_+(\bar{x}, \bar{y})$  by  $D_g F(\bar{x}, \bar{y})$ , then we shall replace the term subgradient by subdifferential. In the following the generalized, the proper, and the weak subdifferential of  $F$  at  $(\bar{x}, \bar{y})$  will be denoted by  $\Gamma_g F(\bar{x}, \bar{y})$ ,  $\Gamma_p F(\bar{x}, \bar{y})$ , and  $\Gamma_w F(\bar{x}, \bar{y})$ , respectively. Sticking to our terminology, the notion of weak subdifferential was recently studied by Wong [35]. Other notions are new.

The above notions are motivated by a similar notion due to Baier and Jahn [3]. Recall that, given a set-valued map  $F : X \rightrightarrows Y$  and  $(\bar{x}, \bar{y}) \in \text{gph}(F)$ , the contingent epiderivative  $D_e F(\bar{x}, \bar{y}) : X \rightarrow Y$  is a *single-valued map* satisfying (3). By employing this epiderivative, a notion of subgradients for set-valued maps, given in [3], is as follows:

$$\partial F(\bar{x}, \bar{y}) := \{L \in \mathcal{L}(X, Y) \mid L(x) \leq D_e F(\bar{x}, \bar{y})(x) \ \forall x \in X\}.$$

Though our primary goal is to state some regularity conditions in terms of the scalarized subgradient, we take a pause to give its relationship with the generalized subgradient.

**PROPOSITION 3.2.** *Let  $X$  and  $Y$  be normed spaces, and let  $F : X \rightrightarrows Y$  be a set-valued map locally  $C$ -convex at  $(\bar{x}, \bar{y}) \in \text{gph}(F)$ . Then  $\partial_p F(\bar{x}, \bar{y}) \neq \emptyset$  implies that  $\partial_{C^* \setminus \{0_{Y^*}\}} F(\bar{x}, \bar{y}) \neq \emptyset$ .*

*Proof.* Let  $\partial_p F(\bar{x}, \bar{y}) \neq \emptyset$  and let  $L \in \partial_p F(\bar{x}, \bar{y})$  be arbitrary. By the definition of the proper subgradient, there is a cone  $K \in \mathcal{K}$  such that  $(\mathcal{D}F_+(\bar{x}, \bar{y}) - L)(x) \cap (-\text{int}(K)) = \emptyset$  for every  $x \in S := \text{dom}(\mathcal{D}F_+(\bar{x}, \bar{y}))$ . We claim that the set  $(\mathcal{D}F_+(\bar{x}, \bar{y}) - L)(S)$  is convex. Indeed, for  $i \in \{1, 2\}$ , let  $y_i \in (\mathcal{D}F_+(\bar{x}, \bar{y}) - L)(S)$ . There is  $x_i \in S$  such that  $y_i \in \mathcal{D}F_+(\bar{x}, \bar{y})(x_i) - L(x_i)$  and hence  $y_i + L(x_i) \in \mathcal{D}F_+(\bar{x}, \bar{y})(x_i)$ . This implies that for some  $\lambda \in (0, 1]$  we have  $\lambda y_1 + (1 - \lambda)y_2 \in (\mathcal{D}F(\bar{x}, \bar{y}) - L)(\lambda x_1 + (1 - \lambda)x_2)$ . Since the set  $S$  is convex, we conclude that  $\lambda y_1 + (1 - \lambda)y_2 \in (\mathcal{D}F_+(\bar{x}, \bar{y}) - L)(S)$ . Now, by employing a separation theorem we assure the existence of  $y^* \in K^* \setminus \{0_{Y^*}\} \subseteq C^* \setminus \{0_{Y^*}\}$  such that

$$(y^* \circ \mathcal{D}F_+(\bar{x}, \bar{y}))(x) \geq (y^* \circ L)(x) \quad \text{for all } x \in S.$$

Since  $L \in \mathcal{L}(X, Y)$  and  $y^* \in Y^*$ , we have  $y^* \circ L \in X^*$  and hence  $y^* \circ L \in \partial_{C^* \setminus \{0_{Y^*}\}} F(\bar{x}, \bar{y})$ . The proof is complete.  $\square$

The following is the main result of the section.

**THEOREM 3.2.** *Let  $X$  and  $Y$  be normed spaces, and let  $F : X \rightrightarrows Y$  be a set-valued map locally  $C$ -convex at  $(\bar{x}, \bar{y}) \in \text{gph}(F)$ . Let  $\mathbb{K}$  be a solid closed convex cone such that  $C \subseteq \mathbb{K}$ . Let  $\text{dom}(\mathcal{D}F_+(\bar{x}, \bar{y}))$  be convex. If  $\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(\mathbb{K})]^- \neq \emptyset$ , then*

$$(7) \quad (\mathcal{D}F_+(\bar{x}, \bar{y})[-\mathbb{K}]^-)^\circ \subset (\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(\mathbb{K})]^-)^\circ \subset \partial_{C^*} F(\bar{x}, \bar{y}).$$

*Furthermore, if  $\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(\mathbb{K})]^- = \emptyset$ , then  $0_{X^*} \in \partial_{C^* \setminus \{0_{Y^*}\}} F(\bar{x}, \bar{y})$ .*

We shall deduce the above assertion from the following result.

**PROPOSITION 3.3.** *Let  $X$  and  $Y$  be normed spaces, let  $\Omega \subseteq X$  be convex, and let  $A \subset Y$  be a solid closed convex cone. Let  $\Psi : \Omega \rightrightarrows Y$  be an  $A$ -convex set-valued map. If  $\Psi[-\text{int}(A)]^- \neq \emptyset$ , then for every  $\psi \in (\Psi[-A])^\circ$  there exists  $t \in A^*$  such that*

$$t \circ \Psi(x) \geq \psi(x) \quad \text{for every } x \in \Omega.$$

*If  $\Psi[-\text{int}(A)]^- = \emptyset$ , then there exists  $t \in A^* \setminus \{0_{Y^*}\}$  such that*

$$t \circ \Psi(x) \geq 0 \quad \text{for every } x \in \Omega.$$

*Proof.* We begin with the case when the set  $\Psi[-\text{int}(A)]^-$  is nonempty. Then the polar  $\Phi^\circ$  of  $\Phi := \Psi[-A]^-$  is also nonempty. We choose  $\psi \in \Phi^\circ$  arbitrarily and define a set

$$E := \{(y, \psi(x)) \in Y \times \mathbb{R} \mid y \in \Psi(x) + A, x \in \Omega\}.$$

In view of the assumptions that  $\Omega$  is convex,  $\Psi$  is  $A$ -convex, and  $\psi \in X^*$ , we deduce that  $E$  is a convex set. Indeed, let  $(y_1, z_1), (y_2, z_2) \in E$  be arbitrary. Then by the definition of  $E$ , for  $i = 1, 2$ , there exists  $x_i \in \Omega \subset X$  with  $z_i = \psi(x_i)$  and  $y_i \in \Psi(x_i) + A$ . For  $\lambda \in (0, 1]$ , we have  $\lambda z_1 + (1 - \lambda)z_2 = \psi(\lambda x_1 + (1 - \lambda)x_2)$ . Further, in view of the  $A$ -convexity of  $\Psi$ , we have  $\lambda y_1 + (1 - \lambda)y_2 \in \lambda\Psi(x_1) + (1 - \lambda)\Psi(x_2) + A \subseteq \Psi(\lambda x_1 + (1 - \lambda)x_2) + A$ . This, in view of the convexity of the set  $\Omega$ , implies that  $\lambda(y_1, z_1) + (1 - \lambda)(y_2, z_2) \in E$ .

Next, we claim that  $E \cap (-\text{int}(A) \times \mathbb{P}) = \emptyset$ . In fact, if this is not the case, then there exists  $(x, y) \in X \times Y$  such that  $y \in (\Psi(x) + A) \cap (-\text{int}(A))$  and  $\psi(x) > 0$ . Let  $w \in \Psi(x)$  be such that  $y \in w + A$ . Then  $w \in y - A \subset -\text{int}(A) - A = -\text{int}(A)$ . This, however, contradicts that  $\psi \in \Phi^\circ$ . Therefore  $E \cap (-\text{int}(A) \times \mathbb{P}) = \emptyset$ , and hence by a separation theorem, we get the existence of  $(f, g) \in Y^* \times \mathbb{R} \setminus \{0_{Y^*}, 0\}$  and a real number  $\alpha$  such that we have

$$(8) \quad f(u) + g \cdot v \geq \alpha \quad \text{for every } (u, v) \in E,$$

$$(9) \quad f(c) + g \cdot v < \alpha \quad \text{for every } (c, d) \in -\text{int}(A) \times \mathbb{P}.$$

Since  $A$  is a cone, we can set  $\alpha = 0$  in (8) and (9). By taking  $d \in \mathbb{P}$  arbitrarily close to 0 and  $c \in -\text{int}(A)$  arbitrarily close to  $0_Y$ , we obtain  $f \in A^*$  and  $g \leq 0$ , respectively. We claim that  $g < 0$ . Indeed, if  $g = 0$ , we get  $f(c) < 0$  for every  $c \in -\text{int}(A)$  and  $f(u) \geq 0$  for every  $u \in \Psi(\Omega) + A$ . This, however, is impossible because we have  $(\Psi(\Omega) + A) \cap (-\text{int}(A)) \neq \emptyset$ . Therefore  $g < 0$ . Moreover, from (8), for every  $x \in \Omega$  we have  $f \circ (\Psi + A)(x) \geq -(g \cdot \psi)(x)$ . By setting  $t = (-f/g) \in A^*$  and noticing that  $0_Y \in A$ , we finish the proof of the first part.

For the second part, we notice that if  $\Psi[-\text{int}(A)]^- = \emptyset$ , we have  $\Psi(\Omega) \cap -\text{int}(A) = \emptyset$ , and hence by arguments similar to those given above we can prove the existence of  $t \in A^* \setminus \{0_Y\}$  such that  $t \circ \Psi(x) \geq 0$  for every  $x \in \Omega$ .  $\square$

The following particular case is worth mentioning. Henceforth we set  $(T(\cdot, \cdot))^\circ = N(\cdot, \cdot)$ .

**COROLLARY 3.3.** *If in Proposition 3.3 we assume that  $A = T(B, -\bar{w})$ , where  $B \subset Y$  is a proper solid convex cone with  $\bar{w} \in -B$  and  $\Psi$  is  $B$ -convex, then  $t \in B^*$  and  $t(\bar{w}) = 0$ .*

*Proof.* To apply Proposition 3.3 it suffices to show that the map  $\Psi$  is  $T(B, -\bar{w})$ -convex. For this, we notice that due to  $\bar{w} \in -B$ , we have  $B \subseteq T(B, -\bar{w})$ . Now  $T(B, -\bar{w})$ -convexity of  $\Psi$  follows from its  $B$ -convexity. By invoking Proposition 3.3, for every  $\psi \in (\Psi[-A])^\circ$ , there exists a functional  $t \in -N(B, -\bar{w})$  which satisfies the assertion. Since  $B$  is convex, we have  $T(B, -\bar{w}) \supseteq B + \bar{w}$  and consequently

$$(10) \quad t(b + \bar{w}) \geq 0 \quad \text{for all } b \in B.$$

By setting  $b = 0_Y$  in the above inequality, we obtain  $t(\bar{w}) \geq 0$ . Because  $B$  is a cone and  $\bar{w} \in -B$ , we can substitute  $b = -2\bar{w} \in B$  in (10) to obtain  $t(\bar{w}) \leq 0$ . Therefore, combining the preceding two inequalities, we have  $t(\bar{w}) = 0$  and this, in view of (10), yields  $t \in B^*$ .  $\square$

Now we are ready to give the following proof.

*Proof of Theorem 3.2.* The first inclusion is a consequence of

$$\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(\mathbb{K})]^- \subset \mathcal{D}F_+(\bar{x}, \bar{y})[-\mathbb{K}]^-,$$

and the second inclusion is essentially a restatement of Proposition 3.3 when we notice that the map  $\mathcal{D}F_+(\bar{x}, \bar{y})$  is  $\mathbb{K}$ -convex and set  $\Psi \equiv \mathcal{D}F_+(\bar{x}, \bar{y})$ .  $\square$



We conclude this section by using the scalarized subgradient to express a necessary optimality condition for the local proper minimality in set-valued optimization.

**PROPOSITION 3.4.** *Let  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  be a local proper minimizer to  $(P_0)$ . Let the map  $F : X \rightrightarrows Y$  be locally  $C$ -convex at  $(\bar{x}, \bar{y})$  with  $\text{dom}(F) = Q_0$ . Then  $0_{X^*} \in \partial_{C^* \setminus \{0_{Y^*}\}} F(\bar{x}, \bar{y})$ .*

*Proof.* Since  $(\bar{x}, \bar{y})$  is a local proper minimizer, by arguments similar to those in Theorem 3.1 we can show that  $\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^- = \emptyset$  for some  $K \in \mathcal{K}$ . (In fact this link has led to the proof of Corollary 3.1.) Now by applying Theorem 3.2 we obtain that  $0_{X^*} \in \partial_{C^* \setminus \{0_{Y^*}\}} F(\bar{x}, \bar{y})$ .  $\square$

**3.3. Lagrange multiplier rule.** We begin this section by proposing the following notion of regularity for set-valued maps.

**DEFINITION 3.1.** *Let  $E_1$  and  $E_2$  be normed spaces and let  $\mathcal{C} \subset E_2$  be a convex cone. Let  $\mathcal{F} : E_1 \rightrightarrows E_2$  be a set-valued map and let  $(x_0, y_0) \in \text{gph}(\mathcal{F})$ . The map  $\mathcal{F}$  is called  $\mathcal{C}$ -regular at  $(x_0, y_0)$  if  $(\mathcal{F} + \mathcal{C})$  has the Aubin property around  $(x_0, y_0)$  and  $\mathcal{F}$  is locally  $\mathcal{C}$ -convex at  $(x_0, y_0)$ .*

**Remark 3.3.** In the above definition, if  $E_2$  is finite-dimensional,  $x_0 \in \text{int}(\text{dom}(\mathcal{F}))$ , and  $\text{epi}(\mathcal{F})$  is closed convex, then  $\mathcal{F}$  is  $\mathcal{C}$ -regular at  $(x_0, y_0)$ . This is a consequence of a known fact that if a set-valued map has a closed convex graph and the image space is finite-dimensional, then the map is locally Lipschitz at the interior of its domain. For details see [15, p. 588].

In our next result we give the promised Lagrange multiplier rule. Recall that  $G_+ := G + D$ .

**THEOREM 3.3.** *Assume that  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  is a local proper minimizer to  $(P_1)$ . Assume that  $F$  is locally  $C$ -regular at  $(\bar{x}, \bar{y})$  and there exists an open convex cone  $M \subset IT(Q_0, \bar{x})$ . Assume that for  $K \in \mathcal{K}$  either  $IT(F[\bar{y} - \text{int}(K)]^-, \bar{x})$  is convex or  $\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^-$  is open. Assume that  $\mathcal{D}F_+(\bar{x}, \bar{y})$  and  $\mathcal{D}G_+(\bar{x}, \bar{z})$ , where  $\bar{z} \in G(\bar{x}) \cap (-D)$ , are strict and the following regularity condition (RC) holds: If  $\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_{Z^*}\}]^- = \emptyset$ , then  $0_{X^*} \in \partial_{-N(D, -\bar{z}) \setminus \{0_{Z^*}\}} G(\bar{x}, \bar{z})$ ; otherwise  $T(G[-D]^-, \bar{x})$  is convex and we have*

$$(11) \quad N(G[-D]^-, \bar{x}) \subseteq \partial_{-N(D, -\bar{z})} G(\bar{x}, \bar{z}).$$

*Then there exist  $(s, t, u) \in X^* \times Y^* \times Z^*$ , not all zero, such that  $s \in M^*$ ,  $t \in C^*$ , and  $u \in D^*$ . Moreover the complementary slackness condition  $u(\bar{z}) = 0$  and the following inequality hold:*

$$(12) \quad t \circ \mathcal{D}F_+(\bar{x}, \bar{y})(x) + u \circ \mathcal{D}G_+(\bar{x}, \bar{z})(x) \geq s(x) \quad \text{for every } x \in X.$$

*If either  $\bar{x} \in \text{int}(Q_0)$  or  $\text{dom}(F) = \text{dom}(G) = Q_0$ , then  $s \in Z^*$  can be set to zero. In this particular case, we have  $t \neq 0_{Y^*}$  if the following regularity condition holds:*

$$(13) \quad \mathcal{D}G_+(\bar{x}, \bar{z})(X) + \text{cone}(D + \bar{z}) = Z.$$

*Proof.* We begin by showing that the assertions would hold trivially if we have either  $\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^- = \emptyset$ , where  $K \in \mathcal{K}$ , or  $\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_{Z^*}\}]^- = \emptyset$ . Indeed, if  $\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^- = \emptyset$ , then in view of Theorem 3.2 there exists  $t \in C^* \setminus \{0_{Y^*}\}$  such that for every  $x \in X$  we have  $t \circ \mathcal{D}F_+(\bar{x}, \bar{y})(x) \subseteq \mathbb{R}_+$ . By choosing  $s = 0_{X^*}$  and  $u = 0_{Z^*}$  we get the desired result.

On the other hand, if  $\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_{Z^*}\}]^- = \emptyset$ , then in view of the condition  $0_{X^*} \in \partial_{-N(D, -\bar{z}) \setminus \{0_{Z^*}\}} G(\bar{x}, \bar{z})$  we ensure the existence of  $u \in -N(D, -\bar{z}) \setminus \{0_{Z^*}\}$  such that for every  $x \in X$  we have  $u \circ \mathcal{D}G_+(\bar{x}, \bar{z})(x) \subseteq \mathbb{R}_+$ . We choose  $s = 0_{X^*}$  and

$t = 0_{Y^*}$  to obtain (12). The proof for  $u \in D^*$  and the complementary slackness condition follows from Corollary 3.3.

Therefore, without any loss of generality, we can assume that

$$\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^- \neq \emptyset$$

and

$$\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_Z\}]^- \neq \emptyset.$$

We first consider the case when  $IT(F[\bar{y} - \text{int}(K)]^-, \bar{x})$  is convex. Now, by using Lemma 3.1 and the containment  $M \subset IT(Q_0, \bar{x})$ , we have

$$IT(F[\bar{y} - \text{int}(K)]^-, \bar{x}) \cap M \cap T(G[-D]^-, \bar{x}) = \emptyset.$$

The above disjunction, in view of the DM-lemma and Proposition 3.1, ensures the existence of functionals  $l_1 \in M^\circ$ ,  $l_2 \in (\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^-)^\circ$ , and  $l_3 \in N(G[-D]^-, \bar{x})$  such that

$$(14) \quad l_1 + l_2 + l_3 = 0.$$

Notice that if  $\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^-$  is open, then since it is a convex cone, we can directly obtain the above conclusion. Now, because  $\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_Z\}]^- \neq \emptyset$ , we have  $N(G[-D]^-, \bar{x}) \subseteq \partial_{-N(D, -\bar{z})}G(\bar{x}, \bar{z})$ , and consequently  $l_3 \in \partial_{-N(D, -\bar{z})}G(\bar{x}, \bar{z})$ . This observation and Theorem 3.2 ensure the existence of functionals  $t \in C^*$  and  $u \in -N(D, -\bar{z})$  such that

$$\begin{aligned} t \circ \mathcal{D}F_+(\bar{x}, \bar{y})(x) &\geq l_2(x) && \text{for every } x \in X, \\ u \circ \mathcal{D}G_+(\bar{x}, \bar{z})(x) &\geq l_3(x) && \text{for every } x \in X. \end{aligned}$$

Combining the above two inequalities with (14) yields

$$t \circ \mathcal{D}F_+(\bar{x}, \bar{y})(x) + u \circ \mathcal{D}G_+(\bar{x}, \bar{z})(x) \geq -l_1(x) \quad \text{for every } x \in X.$$

By setting  $s = -l_1 \in M^*$  we finish the proof of (11). The remaining proof for  $u \in D^*$  and for the complementary slackness condition  $u(\bar{z}) = 0$  is the same as that of Corollary 3.3.

Finally, if either  $\bar{x} \in \text{int}(Q_0)$  or  $\text{dom}(F) = \text{dom}(G) = Q_0$ , then we can apply the above arguments to the disjunction given in Corollary 3.1. Notice that this disjunction does not take into account the space  $Z$ . It remains to show that under the regularity condition (13) we have  $t \neq 0_{Y^*}$ . For this assume that  $t = 0_{Y^*}$  and choose  $\tilde{z} \in Z$  arbitrarily. Since

$$\tilde{z} = z + \beta(d + \bar{z}), \quad \text{where } z \in \mathcal{D}G_+(\bar{x}, \bar{z})(X), \quad d \in D, \quad \beta > 0,$$

we deduce that for an arbitrary  $\tilde{z} \in Z$  we have  $u(\tilde{z}) = u(z) + \beta(u(d) + u(\bar{z})) \geq 0$ . This, however, implies that  $u = 0_{Z^*}$ , and we have a contradiction of  $(t, u) \neq (0_{Y^*}, 0_{Z^*})$ . The proof is complete.  $\square$

As is evident from the above proof, the assumption  $(\mathcal{RC})$  is vital to ensure the existence of the multipliers. In the following we justify this assumption by identifying the cases under which it holds. We begin with a result justifying the inclusion  $0_{X^*} \in \partial_{-N(D, -\bar{z}) \setminus \{0_{Z^*}\}}G(\bar{x}, \bar{z})$ .

PROPOSITION 3.5. *Besides the notation of Theorem 3.3, assume that  $T(D, -\bar{z})$  is locally compact and  $Z$  is reflexive Banach. Assume that  $G : X \rightrightarrows Z$  is locally  $D$ -convex at  $(\bar{x}, \bar{z})$ . Then  $0_{X^*} \in \partial_{-N(D, -\bar{z}) \setminus \{0_Z\}} G(\bar{x}, \bar{z})$ , provided that*

$$\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_Z\}]^- = \emptyset.$$

*Proof.* Since  $\mathcal{D}G_+(\bar{x}, \bar{z})[T(-D, \bar{z}) \setminus \{0_Z\}]^- = \emptyset$ , we have  $\mathcal{D}G_+(\bar{x}, \bar{z})(S) \cap T(-D, \bar{z}) = \{0_Z\}$ , where  $S = \text{dom}(\mathcal{D}G_+(\bar{x}, \bar{z}))$ . Now, by employing a cone separation theorem (see [6]), we ensure the existence of  $l \in Z^* \setminus \{0_{Z^*}\}$  such that

$$\begin{aligned} l \circ \mathcal{D}G_+(\bar{x}, \bar{z})(x) &\geq 0 \quad \text{for every } x \in S, \\ l(T(D, -\bar{z})) &\geq 0. \end{aligned}$$

The condition  $0_{X^*} \in \partial_{-N(D, -\bar{z}) \setminus \{0_{Z^*}\}} G(\bar{x}, \bar{z})$  now follows from the above inequalities.  $\square$

As concerns (11), in fact it is motivated by the conclusions in Proposition 3.3 and Corollary 3.3, and by the condition

$$(15) \quad \mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z})]^- \subseteq T(G[-D]^- , \bar{x}) \quad (Q_0 = \text{dom}(G)),$$

which is very comparable to the well-known Guignard constraint qualification (cf. [36]). Since

$$\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_Z\}]^- \subseteq \mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z})]^- ,$$

an obvious implication of (15) is the inclusion

$$N(G[-D]^- , \bar{x}) \subseteq (\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_Z\}]^-)^\circ,$$

provided that  $\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_Z\}]^- \neq \emptyset$ . Hence, if (15) is valid, then in view of  $N(G[-D]^- , \bar{x}) \subseteq (\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_Z\}]^-)^\circ$ , for (11) it suffices to show that

$$(\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_Z\}]^-)^\circ \subseteq \partial_{-N(D, -\bar{z})} G(\bar{x}, \bar{z}).$$

For this, we notice that if we choose  $p \in (\mathcal{D}G_+(\bar{x}, \bar{z})[-T(D, -\bar{z}) \setminus \{0_Z\}]^-)^\circ$  arbitrarily, we have

$$(16) \quad Q \cap (-T(D, -\bar{z}) \setminus \{0_Z\} \times -\mathbb{P}) = \emptyset,$$

where  $Q = \{(y, p(x)) \in Z \times \mathbb{R} \mid y \in \mathcal{D}G_+(\bar{x}, \bar{z})(x), x \in X\}$ . Now, if  $\text{int}(T(D, -\bar{z})) = IT(D, -\bar{z})$  is nonempty, we repeat the above arguments by replacing  $-T(D, -\bar{z}) \setminus \{0_Z\}$  with  $-IT(D, -\bar{z})$  and, as in Proposition 3.3, we will get that  $p \in \partial_{-N(D, -\bar{z})} G(\bar{x}, \bar{z})$ . However, notice that the above nonintersecting cones may also be separated by cone separation theorems or in cases when  $Q$  has a nonempty interior. On the other hand, it can be shown that the condition  $(\mathcal{RC})$  becomes superfluous if the cone  $D$  is solid and some convexity hypothesis holds. To show this we first prove the following result.

PROPOSITION 3.6. *Let  $X$  and  $Y$  be real normed spaces, let  $G : X \rightrightarrows Y$  be set-valued, and let  $(\bar{x}, \bar{z}) \in \text{gph}(G)$ . Let  $D \subset Y$  be a solid convex cone. Then*

$$\text{cl}(\mathcal{D}G_+(\bar{x}, \bar{z})[IT(-D, \bar{z})]^-) \subseteq T(G[-D]^- , \bar{x}).$$

*Proof.* It suffices to show that  $\mathcal{D}G_+(\bar{x}, \bar{z})[IT(-D, \bar{z})]^- \subseteq T(G[-D]^- , \bar{x})$ . Choose  $x \in \mathcal{D}G_+(\bar{x}, \bar{z})[IT(-D, \bar{z})]$  arbitrarily. Then there exists  $z \in Z$  such that  $z \in$

$\mathcal{D}G_+(\bar{x}, \bar{z})(x) \cap IT(-D, \bar{z})$ . Since  $(x, z) \in \text{gph}(\mathcal{D}G(\bar{x}, \bar{z}))$ , there are sequences  $(\lambda_n) \subset \mathbb{P}$ ,  $((x_n, z_n)) \subset X \times Y$  such that  $\lambda_n \downarrow 0$ ,  $x_n \rightarrow x$ ,  $z_n \rightarrow z$ , and  $\bar{z} + \lambda_n z_n \in G(\bar{x} + \lambda_n x_n) + D$  for every  $n \in \mathbb{N}$ . By the definition of  $IT(-D, \bar{z})$ , there exists  $n_1 \in \mathbb{N}$  such that  $\bar{z} + \lambda_n z_n \in -D$  for every  $n \geq n_1$ . Now, let  $w_n \in G(\bar{x} + \lambda_n x_n)$  be such that  $\bar{z} + \lambda_n z_n \in w_n + D$ . Then  $w_n \in -D$  for every  $n \geq n_1$ , implying  $\bar{x} + \lambda_n x_n \in G[-D]^-$  for every  $n \geq n_1$ . Therefore  $x \in T(G[-D]^-, \bar{x})$ .  $\square$

In view of the above result we have the following analogue of Theorem 3.1 where derivatives of both  $F$  and  $G$  are taken into account.

**THEOREM 3.4.** *Assume that  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  is a local proper minimizer to  $(P_1)$  and assume that  $D$  is a solid closed convex cone. Assume that the map  $F_+ (:= F + C)$  possesses the Aubin property around  $(\bar{x}, \bar{y})$ . Then*

$$(17) \quad \mathcal{D}F_+(\bar{x}, \bar{y})[-C \setminus \{0_Y\}]^- \bigcap IT(Q_0, \bar{x}) \bigcap \text{cl}(\mathcal{D}G_+(\bar{x}, \bar{z})[IT(-D, \bar{z})]^-) = \emptyset.$$

Now we are in a position to give the following important particular case of Theorem 3.3.

**COROLLARY 3.4.** *The conclusions of Theorem 3.3 remain valid if instead of  $(\mathcal{RC})$  we assume that the map  $G$  is locally  $D$ -convex at  $(\bar{x}, \bar{z})$  and  $\text{int}(D) \neq \emptyset$ .*

*Proof.* We can consider two cases, namely, when the set  $\mathcal{D}G_+(\bar{x}, \bar{y})[-\text{int}(T(D, -\bar{z}))]^-$  is empty and when it is not empty. In the first case by modifying Proposition 3.5 we get the first part of  $(\mathcal{RC})$ . For the second case we can use (16) and Proposition 3.3. The proof is complete.  $\square$

In fact while dealing with the cases when  $D$  is solid, we can also cope with the situation in which  $IT(Q_0, \bar{x})$  is not defined. For this the following result is instrumental.

**PROPOSITION 3.7.** *Besides the hypothesis of Proposition 3.6, assume that the map  $G$  possess the Aubin property around  $(\bar{x}, \bar{z})$ . Then*

$$\mathcal{D}G_+(\bar{x}, \bar{z})[IT(-D, \bar{z})]^- \subseteq IT(G[-D]^-, \bar{x}).$$

*Proof.* The proof is based on the arguments given in the proof of Proposition 3.6 and the use of the Aubin property depicted in Theorem 3.1.  $\square$

The following is an analogue of Corollary 3.4.

**COROLLARY 3.5.** *The conclusions of Corollary 3.4 remain valid if the hypothesis that there exists some open convex cone  $M \subset IT(Q_0, \bar{x})$  is replaced by the requirement that there exist a closed convex cone  $M \subset T(Q_0, \bar{x})$  and that  $G$  possess the Aubin property around  $(\bar{x}, \bar{z})$ .*

**3.4. Necessary optimality condition for  $(P_2)$ .** We begin this subsection by the following necessary optimality condition for  $(P_2)$ .

**THEOREM 3.5.** *Let  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  be a local proper minimizer to  $(P_2)$ . Let  $D$  be solid and let  $F_+$  have the Aubin property around  $(\bar{x}, \bar{y})$ . Then*

$$(18) \quad IT(Q_0, \bar{x}) \bigcap \mathcal{D}F_+(\bar{x}, \bar{y})[-C \setminus \{0_Y\}]^- \bigcap IT(G[-D]^-, \bar{x}) \bigcap T(H[h]^-, \bar{x}) = \emptyset.$$

*Proof.* Since  $(\bar{x}, \bar{y})$  is a local proper minimizer of  $(P_2)$ , there are a neighborhood  $U_1$  of  $\bar{x}$  and a pointed convex solid cone  $K \in \mathcal{K}$  such that  $\bar{y} \in \text{Min}(F(Q_1 \cap U), K)$ . The disjunction (18) can be obtained from

$$IT(Q_0, \bar{x}) \bigcap \mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^- \bigcap IT(G[-D]^-, \bar{x}) \bigcap T(H[h]^-, \bar{x}) = \emptyset,$$

which can be proved by using arguments similar to those given in the proof of Theorem 3.1.  $\square$

A variant of Theorem 3.5, involving the derivatives of both  $F$  and  $G$ , is as follows.

**THEOREM 3.6.** *Let  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  be a local proper minimizer to  $(P_2)$ . Let  $D$  be solid, let  $F_+$  have the Aubin property around  $(\bar{x}, \bar{y})$ , and let  $G_+$  have the Aubin property around  $(\bar{x}, \bar{z})$  where  $\bar{z} \in G(\bar{x}) \cap -D$ . Then*

$$IT(Q_0, \bar{x}) \cap \mathcal{D}F_+(\bar{x}, \bar{y})[-C \setminus \{0_Y\}]^- \cap \mathcal{D}G_+(\bar{x}, \bar{z})[IT(-D, \bar{z})]^- \cap T(H[h]^- , \bar{x}) = \emptyset.$$

*Proof.* The proof follows by combining Theorem 3.5 and Proposition 3.7.  $\square$

In our next result we give a Lagrange multiplier rule for  $(P_2)$ .

**THEOREM 3.7.** *Assume that  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  is a local proper minimizer to  $(P_2)$  and  $\bar{z} \in G(\bar{x}) \cap (-D)$ . Assume that  $F$  is locally  $C$ -regular at  $(\bar{x}, \bar{y})$  and  $G_+$  is locally  $D$ -regular at  $(\bar{x}, \bar{z})$ . Assume that there are an open convex cone  $M_1 \subset IT(Q_0, \bar{x})$  and a closed convex cone  $M_2 \subset T(H[h]^- , \bar{x})$  and that for  $K \in \mathcal{K}$  either  $IT(F[\bar{y} - \text{int}(K)]^- , \bar{x})$  is convex or  $\mathcal{D}F_+(\bar{x}, \bar{y})[-\text{int}(K)]^-$  is open. Assume that either  $IT(G[-D]^- , \bar{x})$  is convex or  $\mathcal{D}G_+(\bar{x}, \bar{z})[IT(-D, \bar{z})]^-$  is open. Assume that  $\mathcal{D}F_+(\bar{x}, \bar{y})$  and  $\mathcal{D}G_+(\bar{x}, \bar{z})$  are strict. Then there exists  $(s, t, u, v) \in X^* \times Y^* \times Z^* \times W^*$ , not all simultaneously zero, such that  $s \in M_1^*$ ,  $t \in C^*$ ,  $u \in D^*$ , and  $v \in M_2^*$ . Moreover the complementary slackness condition  $u(\bar{z}) = 0$  and the following inequality hold:*

$$t \circ \mathcal{D}F_+(\bar{x}, \bar{y})(x) + u \circ \mathcal{D}G_+(\bar{x}, \bar{z})(x) \geq s(x) + v(x) \quad \text{for every } x \in X.$$

*Proof.* The proof is similar to the proof of Theorem 3.3.  $\square$

It is clear that in the above we have not imposed any differentiability assumption on the map  $H$ . For this, it would be of interest to obtain a variant of the well-known theorem of Lyusternik [18], fitting the present setting, so that the cone  $M_2^o$  contains information about some derivative of  $H$ . In fact, this is completely true if  $H$  is single-valued and sufficiently smooth [31]. To state this particular case, we recall the following known result.

**LEMMA 3.2** (see [31]). *Let  $X$  and  $Y$  be Banach spaces and let  $H : X \rightarrow Y$  be a single-valued map strongly differentiable at  $\bar{x} \in X$ . Let the derivative  $DH(\bar{x}) : X \rightarrow Y$  be a projection. Then*

$$\{h \in X : DH(\bar{x})(h) = 0\} = T(H[H(\bar{x})]^- , \bar{x}).$$

We have the following particular case of Theorem 3.7.

**COROLLARY 3.6.** *Let the hypothesis of Theorem 3.7 be altered in a sense that the map  $H$  satisfies the conditions of Lemma 3.2. Then there exist  $(s, t, u, v) \in X^* \times Y^* \times Z^* \times W^*$ , not all zero, such that  $s \in M_1^*$ ,  $t \in C^*$ , and  $u \in D^*$ . Moreover the complementary slackness condition  $u(\bar{z}) = 0$  and the following inequality hold:*

$$t \circ \mathcal{D}F_+(\bar{x}, \bar{y})(x) + u \circ \mathcal{D}G_+(\bar{x}, \bar{z})(x) + v \circ DH(\bar{x})(x) \geq s(x) \quad \text{for every } x \in X.$$

*Proof.* As in the proof of Theorem 3.3, we can show that there exists  $l_4 \in A^o$ , where  $A := \{x \in X \mid DH(\bar{x})(x) = 0\}$  is a subspace of  $X$ . Clearly, there exists  $v \in W^*$  such that  $v \circ DH(\bar{x}) = l_4$ . This observation then finishes the proof.  $\square$

**4. Optimality conditions for strong minimality.** Recall that a point  $(\bar{x}, \bar{y}) \in \text{gph}(F)$  is called a local strong minimizer of  $(P_2)$  if there exists a neighborhood  $U$  of  $\bar{x}$  such that  $F(Q_2 \cap U) \subseteq \bar{y} + C$ . In the following result, we give necessary optimality

conditions for the local strong minimality. Let the cone  $C$  be closed and convex and let the cone  $D$  be solid.

**THEOREM 4.1.** *Let  $X$  and  $Y$  be normed spaces, let  $F : X \rightrightarrows Y$  be set-valued, and let  $(\bar{x}, \bar{y}) \in \text{gph}(F)$ . Let the map  $F_+$  possess the Aubin property around  $(\bar{x}, \bar{y})$ .*

(a) *If  $(\bar{x}, \bar{y})$  is a strong minimizer to  $(P_1)$ , then the following holds:*

$$\mathcal{D}F_+(\bar{x}, \bar{y})[Y \setminus C]^- \bigcap IT(Q_0, \bar{x}) \bigcap T(G[-D]^-, \bar{x}) = \emptyset.$$

(b) *If  $(\bar{x}, \bar{y})$  is a strong minimizer to  $(P_1)$ , then for  $\bar{z} \in G(\bar{x}) \cap -D$  the following holds:*

$$\mathcal{D}F_+(\bar{x}, \bar{y})[Y \setminus C]^- \bigcap IT(Q_0, \bar{x}) \bigcap \text{cl}(\mathcal{D}G_+(\bar{x}, \bar{z})[IT(-D, \bar{z})]^-) = \emptyset.$$

(c) *If  $(\bar{x}, \bar{y})$  is a strong minimizer to  $(P_2)$ , then the following holds:*

$$\mathcal{D}F_+(\bar{x}, \bar{y})[Y \setminus C]^- \bigcap IT(Q_0, \bar{x}) \bigcap IT(G[-D]^-, \bar{x}) \bigcap T(H[h]^-, \bar{x}) = \emptyset.$$

(d) *Let the map  $G_+$  possess the Aubin property around  $(\bar{x}, \bar{z})$  where  $\bar{z} \in G(\bar{x}) \cap -D$ . If  $(\bar{x}, \bar{y})$  is a strong minimizer to  $(P_2)$ , then the following holds:*

$$\mathcal{D}F_+(\bar{x}, \bar{y})[Y \setminus C]^- \bigcap IT(Q_0, \bar{x}) \bigcap \mathcal{D}G_+[IT(-D), \bar{z}]^- \bigcap T(H[h]^-, \bar{x}) = \emptyset.$$

*Proof.* (a) Assume that there exists

$$x \in \mathcal{D}F_+(\bar{x}, \bar{y})[Y \setminus C]^- \bigcap IT(Q_0, \bar{x}) \bigcap T(G[-D]^-, \bar{x}).$$

In view of the inclusion  $x \in \mathcal{D}F_+(\bar{x}, \bar{y})[Y \setminus C]^-$ , there exists  $y \in Y \setminus C$  such that  $(x, y) \in \text{gph}(\mathcal{D}F_+(\bar{x}, \bar{y}))$ . This implies that there are sequences  $(\lambda_n) \subset \mathbb{P}$  and  $((x_n, y_n)) \subset X \times Y$  such that  $\lambda_n \rightarrow 0$ ,  $(x_n, y_n) \rightarrow (x, y)$ , and  $\bar{y} + \lambda_n y_n \in F(\bar{x} + \lambda_n x_n) + C$  for all  $n \in \mathbb{N}$ . Also, since  $x \in IT(Q_0, \bar{x})$ , there exists  $n_1 \in \mathbb{N}$  such that  $\bar{x} + \lambda_n x_n \in Q_0$  for  $n \geq n_1$ .

Since  $y \notin C$ ,  $y_n \rightarrow y$ , and the cone  $C$  is closed, there exists  $n_2 \in \mathbb{N}$  such that  $y_n \notin C$  for all  $n \geq n_2$ . Because  $\lambda_n > 0$ , we have  $\lambda_n y_n \notin C$ . For  $u_n := \bar{x} + \lambda_n x_n$ , let  $w_n \in F(u_n)$  be such that for  $n \in \mathbb{N}$ ,  $\bar{y} + \lambda_n y_n = w_n + c_n$ , where  $c_n \in C$ . Hence  $w_n + c_n \notin \bar{y} + C$  for  $n \geq n_2$ . Since  $C + C = C$ , we have  $w_n \notin \bar{y} + C$  for  $n \geq n_2$ . As in the proof of Theorem 3.1, we can show that there exists  $n_3 \in \mathbb{N}$  such that  $G(u_n) \cap (-D) \neq \emptyset$  for all  $n \geq n_3$ .

Therefore, we have shown that for sufficiently large  $n \in \mathbb{N}$  there are  $w_n \in F(Q_1)$  such that  $w_n \notin \bar{y} + C$ . This, however, is a contradiction of the assumption that  $(\bar{x}, \bar{y})$  is a strong minimizer. Hence part (a) is true. Parts (b)–(d) can now be shown by unifying the above arguments with those given in the preceding section.  $\square$

**5. Concluding remarks.** New optimality conditions in set-valued optimization are given by employing the ideas around the so-called Dubovitskii–Milyutin approach. For inequality constraints our results appear to be satisfactory; however, the treatment of the equality constraints needs further research. In fact, a generalization of the theorem of Lyusternik is needed which fits the present setting.

It should also be stressed that although our approach is motivated by the Dubovitskii–Milyutin approach, there are significant differences in using the separation schemes for set optimization and for classical optimization. The main challenge is in identifying that the Aubin property enables the disjunction (4). Under suitable

differentiability assumptions such a restriction can be avoided in classical optimization (cf. [13], [31]). Moreover, in classical optimization, the existence of multipliers can also be proved without imposing convexity assumptions [18, section 17.4].

Our main objective in this work is to present a detailed treatment of new necessary optimality conditions. One important question which remained unanswered is whether these conditions are sufficient as well. We strongly believe the conditions stated in Theorem 3.3 become sufficient when stronger convexity conditions are imposed on the set-valued maps involved (see [8], [3]). However, since there are many new concepts of convexity for set-valued maps, a challenging aspect is to choose the most optimal notion for the sufficient optimality conditions. We plan to address this issue in a future work. Another interesting aspect of this study is to apply these results to some concrete problems, for example, those arising in control theory and inverse problems.

**Acknowledgments.** The second author is grateful to Prof. Dr. Johannes Jahn for the stimulating discussion and helpful remarks on this work. The authors are grateful to the referees for their careful reading of the paper and for the suggested improvements.

#### REFERENCES

- [1] J. P. AUBIN, *Contingent derivatives of set-valued maps and existence of solutions to nonlinear inclusions and differential inclusions*, in Mathematical Analysis and Applications, Part A, L. Nachbin, ed., Adv. in Math. Suppl. Stud. 7a, Academic Press, New York, London, 1981, pp. 159–229.
- [2] J. P. AUBIN AND H. FRANKOWSKA, *Set Valued Analysis*, Birkhäuser, Boston, 1990.
- [3] J. BAIER AND J. JAHN, *On subgradients of set-valued maps*, J. Optim. Theory Appl., 100 (1999), pp. 233–240.
- [4] E. M. BEDNARCZUK AND W. SONG, *Contingent epiderivative and its applications to set-valued optimization*, Control Cybernet., 27 (1998), pp. 375–386.
- [5] G. BIGI AND M. CASTELLANI, *K-epiderivatives for set-valued functions and optimization*, Math. Methods Oper. Res., 55 (2002), pp. 401–412.
- [6] J. BORWEIN, *Proper efficient points for maximizations with respect to cones*, SIAM J. Control Optim., 15 (1977), pp. 57–63.
- [7] G. Y. CHEN AND J. JAHN, *Optimality conditions for set-valued optimization problems*, Math. Methods Oper. Res., 48 (1998), pp. 187–200.
- [8] H. W. CORLEY, *Optimality conditions for maximization of set-valued functions*, J. Optim. Theory Appl., 58 (1988), pp. 1–10.
- [9] A. Y. DUBOVITSKII AND A. A. MILYUTIN, *Extremal problems in presence of constraints*, Comput. Math. Math. Phys., 5 (1965), pp. 395–453.
- [10] B. EL ABDOUNI AND L. THIBAUT, *Optimality conditions for problems with set-valued objectives*, J. Appl. Anal., 2 (1996), pp. 183–201.
- [11] F. FLORES-BAZÁN, *Optimality conditions in non-convex set-valued optimization*, Math. Methods Oper. Res., 53 (2001), pp. 403–417.
- [12] F. GIANNESI, ED., *Vector Variational Inequalities and Vector Equilibria*, Nonconvex Optim. Appl. 38, Kluwer Academic, Dordrecht, The Netherlands, 2000.
- [13] I. V. GIRSANOV, *Lectures on Mathematical Theory of Extremum Problems*, Lecture Notes in Econom. and Math. Systems 67, Springer-Verlag, Berlin, New York, 1972.
- [14] A. GÖTZ AND J. JAHN, *The Lagrange multiplier rule in set-valued optimization*, SIAM J. Optim., 10 (1999), pp. 331–344.
- [15] S. HU AND N. S. PAPAGEORGIOU, *Handbook of Multivalued Analysis, Vol. I. Theory*, Math. Appl. 419, Kluwer Academic, Dordrecht, The Netherlands, 1997.
- [16] G. ISAC, *Topological Methods in Complementarity Theory*, Nonconvex Optim. Appl. 41, Kluwer Academic, Dordrecht, The Netherlands, 2000.
- [17] G. ISAC, V. A. BULAVSKI, AND V. V. KALASHNIKOV, *Complementarity, Equilibrium, Efficiency and Economics*, Kluwer Academic, Dordrecht, The Netherlands, 2002.
- [18] J. JAHN, *Vector Optimization. Theory, Applications, and Extensions*, Springer-Verlag, Berlin, 2004.
- [19] J. JAHN AND A. A. KHAN, *Generalized contingent epiderivatives in set-valued optimization*,

- Numer. Funct. Anal. Optim., 27 (2002), pp. 807–831.
- [20] J. JAHN AND A. A. KHAN, *Existence theorems and characterizations of generalized contingent epiderivatives*, J. Nonlinear Convex Anal., 3 (2002), pp. 315–330.
  - [21] J. JAHN AND R. RAUH, *Contingent epiderivatives and set-valued optimization*, Math. Methods Oper. Res., 46 (1997), pp. 193–211.
  - [22] A. JOURANI, *Necessary conditions for extremality and separation theorems with applications to multiobjective optimization*, Optimization, 44 (1998), pp. 327–350.
  - [23] A. A. KHAN AND F. RACITI, *A multiplier rule in set-valued optimization*, Bull. Austral. Math. Soc., 68 (2003), pp. 93–100.
  - [24] S. J. LI, H. YAN, AND G. Y. CHEN, *Differential and sensitivity properties of gap functions for vector variational inequalities*, Math. Methods Oper. Res., 57 (2003), pp. 377–391.
  - [25] D. T. LUC, *Theory of Vector Optimization*, Lecture Notes in Econom. and Math. Systems 319, Springer-Verlag, Berlin, 1988.
  - [26] D. T. LUC AND C. MALIVERT, *Invex optimization problems*, Bull. Austral. Math. Soc., 46 (1992), pp. 47–66.
  - [27] K. MIETTINEN AND M. M. MÄKELÄ, *On cone characterizations of weak, proper and Pareto optimality in multiobjective optimization*, Math. Methods Oper. Res., 53 (2001), pp. 233–245.
  - [28] B. MORDUKHOVICH, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.
  - [29] J. P. PENOT, *On regularity conditions in mathematical programming*, Math. Programming Stud., 19 (1982), pp. 167–199.
  - [30] J.-P. PENOT, *Differentiability of relations and differential stability of perturbed optimization problems*, SIAM J. Control Optim., 22 (1984), pp. 529–551.
  - [31] L. RIGBY, *Contribution to Dubovitsky and Milyutin's optimization formalism*, in Optimization Techniques, Lecture Notes in Comput. Sci. 41, Springer-Verlag, Berlin, New York, 1976, pp. 438–453.
  - [32] R. T. ROCKAFELLAR AND J. B. WETS, *Variational Analysis*, Springer-Verlag, Berlin, 1997.
  - [33] Y. SAWARAGI, H. NAKAYAMA, AND T. TANINO, *Theory of Multiobjective Optimization*, Math. Sci. Engrg. 176, Academic Press, Orlando, FL, 1985.
  - [34] T. TANINO, *Stability and sensitivity analysis in convex vector optimization*, SIAM J. Control Optim., 26 (1988), pp. 521–536.
  - [35] S. WONG, *A note on weak subdifferential of set-valued mappings*, Optimization, 52 (2003), pp. 263–276.
  - [36] J. ZOWE AND S. KURCYUSZ, *Regularity and stability for the mathematical programming problem in normed spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.



## REAL TIME SOLUTION OF THE NONLINEAR FILTERING PROBLEM WITHOUT MEMORY II\*

SHING-TUNG YAU<sup>†</sup> AND STEPHEN S.-T. YAU<sup>‡</sup>

*Dedicated to Professor Tyrone Duncan on the occasion of his 65th birthday*

**Abstract.** It is well known that the nonlinear filtering problem has important applications in both military and commercial industries. The central problem of nonlinear filtering is to solve the Duncan–Mortensen–Zakai (DMZ) equation in real time and in a memoryless manner. The purpose of this paper is to show that, under very mild conditions (which essentially say that the growth of the observation  $|h|$  is greater than the growth of the drift  $|f|$ ), the DMZ equation admits a unique nonnegative weak solution  $u$  which can be approximated by a solution  $u_R$  of the DMZ equation on the ball  $B_R$  with  $u_R|_{\partial B_R} = 0$ . The error of this approximation is bounded by a function of  $R$  which tends to zero as  $R$  goes to infinity. The solution  $u_R$  can in turn be approximated efficiently by an algorithm depending only on solving the observation-independent Kolmogorov equation on  $B_R$ . In theory, our algorithm can solve basically all engineering problems in real time. Specifically, we show that the solution obtained from our algorithms converges to the solution of the DMZ equation in the  $L^1$  sense. Equally important, we have a precise error estimate of this convergence, which is important in numerical computation.

**Key words.** nonlinear filtering, DMZ equation, conditional probability density, Kolmogorov equation

**AMS subject classifications.** 35K15, 60G35, 62M20, 65N15, 90E10, 93E11

**DOI.** 10.1137/050648353

**1. Introduction.** In 1961, Kalman and Bucy [Ka-Bu] first established the finite-dimensional filter for the linear filtering model with Gaussian initial distribution, which is highly influential in modern industry. Since then filtering theory has proved useful in science and engineering, for example, the navigational and guidance systems, radar tracking, sonar ranging, and satellite and airplane orbit determination. Despite its usefulness, however, the Kalman–Bucy filter is not perfect. Its main weakness is that it is restricted to the linear dynamical system with Gaussian initial distribution. Therefore there has been tremendous interest in solving the nonlinear filtering problem which involves the estimation of a stochastic process  $x = \{x_t\}$  (called the signal or state process) that cannot be observed directly. Information containing  $x$  is obtained from observations of a related process  $y = \{y_t\}$  (the observation process). The goal of nonlinear filtering is to determine the conditional density  $\rho(t, x)$  of  $x_t$  given the observation history of  $\{y_s : 0 \leq s \leq t\}$ . In the late 1960s, Duncan [Du], Mortensen [Mo], and Zakai [Za] independently derived the Duncan–Mortensen–Zakai (DMZ) equation for the nonlinear filtering theory, which the conditional probability density  $\rho(t, x)$  has to satisfy. The central problem of nonlinear filtering theory is to solve the DMZ equation in real time and in a memoryless way.

In 2000, we [Ya-Ya] proposed a novel algorithm to do just that. Under the as-

---

\*Received by the editors December 23, 2005; accepted for publication (in revised form) June 15, 2007; published electronically January 11, 2008.

<http://www.siam.org/journals/sicon/47-1/64835.html>

<sup>†</sup>Department of Mathematics, Harvard University, Cambridge, MA 02138 (yau@math.harvard.edu). Research partially supported by the U.S. Army Research Office.

<sup>‡</sup>Institute of Mathematics, East China Normal University, Shanghai, China. Current address: Department of Mathematics, Statistics and Computer Science (MC 249), University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045 (yau@uic.edu).

sumptions that the drift terms  $f_i(x)$ ,  $1 \leq i \leq n$ , and their first and second derivatives, and the observation terms  $h_i(x)$ ,  $1 \leq i \leq m$ , and their first derivatives, have linear growth, we showed that the solution obtained from our algorithms converges to the true solution of the DMZ equation. Although the above approach is quite successful, so far it cannot handle the famous cubic sensor in engineering in which  $f(x) = 0$  and  $h(x) = x^3$ . It is well known that there is no finite-dimensional filter for the cubic sensor [Su].

The purpose of this paper is to show that under very mild conditions (A.2), (A.17), and (C.3) (which essentially say that the growth of  $|h|$  is greater than the growth of  $|f|$ ), the DMZ equation admits a unique nonnegative solution  $u \in W_0^{1,1}((0, T) \times \mathbb{R}^n)$  which can be approximated by solutions  $u_R$  of the DMZ equation on the ball  $B_R$  with  $u_R|_{\partial B_R} = 0$ . The rate of convergence can be efficiently estimated in the  $L^1$  norm. The solution  $u_R$  can in turn be approximated efficiently by an algorithm depending only on solving the time-independent Kolmogorov equation on  $B_R$ . Our algorithm can solve practically all engineering problems, including the cubic sensor problem in real time and in a memoryless fashion. Specifically we show that the solution obtained from our algorithms converges to the solution of the DMZ equation in the  $L^1$  sense. Equally important, we have a precise error estimate of this convergence, which is important in numerical computation.

The filtering problem considered here is based on the signal observation model

$$(1.1) \quad \begin{cases} dx(t) = f(x(t)) dt + dv(t), & x(0) = x_0, \\ dy(t) = h(x(t)) dt + dw(t), & y(0) = 0, \end{cases}$$

in which  $x, v, y$ , and  $w$  are, respectively,  $\mathbb{R}^n$ -,  $\mathbb{R}^n$ -,  $\mathbb{R}^m$ -, and  $\mathbb{R}^m$ -valued processes and  $v$  and  $w$  have components that are independent, standard Brownian processes. We further assume that  $f$  and  $h$  are  $C^\infty$  smooth vector-valued. We shall refer to  $x(t)$  as the state of the system at time  $t$  and  $y(t)$  as the observation at time  $t$ .

Let  $\rho(t, x)$  denote the conditional probability density of the state given the observation  $\{y(s): 0 \leq s \leq t\}$ . It is well known that  $\rho(t, x)$  is given by normalizing a function,  $\sigma(t, x)$ , which satisfies the following DMZ equation:

$$(1.2) \quad d\sigma(t, x) = L_0\sigma(t, x) dt + \sum_{i=1}^n L_i\sigma(t, x) dy_i(t), \quad \sigma(0, x) = \sigma_0,$$

where

$$(1.3) \quad L_0 = \frac{1}{2} \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^n f_i \frac{\partial}{\partial x_i} - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} - \frac{1}{2} \sum_{i=1}^m h_i^2,$$

and for  $i = 1, \dots, m$ ,  $L_i$  is the zero degree differential operator of multiplication by  $h_i$ . (Here we have used the notation  $p_i$  to represent the  $i$ th component of the vector  $p$ .)  $\sigma_0$  is the probability density of the initial point  $x_0$ .

Equation (1.2) is a stochastic partial differential equation. In real applications, we are interested in constructing robust state estimators from observed sample paths with some property of robustness. Davis [Da] studied this problem and proposed some robust algorithms. In our case, his basic idea reduces to defining a new unnormalized density

$$(1.4) \quad u(t, x) = \exp \left( - \sum_{i=1}^m h_i(x) y_i(t) \right) \sigma(t, x).$$

It is easy to show that  $u(t, x)$  satisfies the time-varying partial differential equation

$$(1.5) \quad \begin{cases} \frac{\partial u}{\partial t}(t, x) = L_0 u(t, x) + \sum_{i=1}^m y_i(t) [L_0, L_i] u(t, x) \\ \quad + \frac{1}{2} \sum_{i,j=1}^m y_i(t) y_j(t) [[L_0, L_i], L_j] u(t, x), \\ u(0, x) = \sigma_0, \end{cases}$$

where  $[\cdot, \cdot]$  denotes the Lie bracket. It is shown in [Ya-Ya, p. 236] that the robust DMZ equation (1.5) is of the form

$$(1.6) \quad \begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{1}{2} \Delta u(t, x) + (-f(x) + \nabla K(t, x)) \cdot \nabla u(t, x) \\ \quad + \left( -\operatorname{div} f(x) - \frac{1}{2} |h(x)|^2 + \frac{1}{2} \Delta K(t, x) \right. \\ \quad \left. - f(x) \cdot \nabla K(t, x) + \frac{1}{2} |\nabla K(t, x)|^2 \right) u(t, x), \\ u(0, x) = \sigma_0(x), \end{cases}$$

where  $K = \sum_{j=1}^m y_j(t) h_j(x)$ ,  $f = (f_1, \dots, f_n)$ , and  $h = (h_1, \dots, h_m)$ .

To simplify our presentation, we introduce the following condition.

*Condition (C<sub>1</sub>).*

$$-\frac{1}{2} |h|^2 - \frac{1}{2} \Delta K - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 + |f - \nabla K| \leq c_1 \quad \forall (t, x) \in [0, T] \times \mathbb{R}^n,$$

where  $c_1$  is a constant possibly depending on  $T$ .

Our main theorems are as follows.

**THEOREM A.** *Consider the filtering model (1.1). For any  $T > 0$ , let  $u$  be a solution of the robust DMZ equation (1.6) in  $[0, T] \times \mathbb{R}^n$ . Assume Condition (C<sub>1</sub>) is satisfied.*

*Then*

$$(1.7) \quad \sup_{0 \leq t \leq T} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(t, x) \leq e^{(c_1 + \frac{n+1}{2})T} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(0, x).$$

*In particular,*

$$(1.8) \quad \sup_{0 \leq t \leq T} \int_{|x| \geq R} u(t, x) \leq e^{-\sqrt{1+R^2}} e^{(c_1 + \frac{n+1}{2})T} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(0, x).$$

Theorem A above says that one can choose a ball large enough to capture almost all the density. In fact by (1.8) we have a precise estimate of density lying outside this ball.

**THEOREM B.** *Consider the filtering model (1.1). For any  $T > 0$ , let  $u$  be a solution of the robust DMZ equation (1.6) in  $[0, T] \times \mathbb{R}^n$ . Assume the following:*

(1) *Condition (C<sub>1</sub>) is satisfied.*

(2)  $-\frac{1}{2} |h|^2 - \frac{1}{2} \Delta K - f(x) \cdot \nabla K(t, x) + \frac{1}{2} |\nabla K|^2 + 12 + 2n + 4|f - \nabla K| \leq c_2$  for all  $(t, x) \in [0, T] \times \mathbb{R}^n$ , where  $c_2$  is a constant possibly depending on  $T$ .

(3)  $e^{-\sqrt{1+|x|^2}} [12 + 2n + 4|f - \nabla K|] \leq c_3$  for all  $(t, x) \in [0, T] \times \mathbb{R}^n$ .

Let  $R \geq 1$  and  $u_R$  be the solution of the following DMZ equation on the ball  $B_R$ :

$$(1.9) \quad \begin{cases} \frac{\partial u_R}{\partial t} = \frac{1}{2} \Delta u_R + (-f + \nabla K) \cdot \nabla u_R \\ \quad + \left( -\operatorname{div} f - \frac{1}{2} |h|^2 + \frac{1}{2} \Delta K - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) u_R, \\ u_R(t, x) = 0 \quad \text{for } (t, x) \in [0, T] \times \partial B_R, \\ u_R(0, x) = \sigma_0(x). \end{cases}$$

Let  $v = u - u_R$ . Then  $v \geq 0$  for all  $(t, x) \in [0, T] \times B_R$  and

$$(1.10) \quad \int_{B_R} \phi v(T, x) \leq \frac{e^{c_2 T} - 1}{c_2} c_3 e^{-R} e^{(c_1 + \frac{n+1}{2})T} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(0, x),$$

where  $\phi(x) = e^{\frac{|x|^4}{R^3} - \frac{2|x|^2}{R}} - e^{-R}$ . In particular

$$(1.11) \quad \int_{B_{\frac{R}{2}}} v(T, x) \leq \frac{2(e^{c_2 T} - 1)}{c_2} c_3 e^{-\frac{9}{16}R} e^{(c_1 + \frac{n+1}{2})T} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(0, x).$$

Theorem B above says that we can approximate  $u$  by  $u_R$ . The approximation is good if  $R$  is large enough. In fact we have a precise error estimate of this approximation by (1.11).

**THEOREM C.** Let  $\Omega$  be a bounded domain in  $\mathbb{R}^n$ . Let  $F: [0, T] \times \Omega \rightarrow \mathbb{R}^n$  be a family of vector fields  $C^\infty$  in  $x$  and Hölder continuous in  $t$  with exponent  $\alpha$  and let  $J: [0, T] \times \Omega \rightarrow \mathbb{R}$  be a  $C^\infty$  function in  $x$  and Hölder continuous in  $t$  with exponent  $\alpha$  such that the following properties are satisfied:

$$(1.12) \quad |\operatorname{div} F(t, x)| + 2|J(t, x)| + |F(t, x)| \leq c \quad \text{for } (t, x) \in [0, T] \times \Omega,$$

$$(1.13) \quad |F(t, x) - F(\bar{t}, x)| + |\operatorname{div} F(t, x) - \operatorname{div} F(\bar{t}, x)| + |J(t, x) - J(\bar{t}, x)| \leq c_1 |t - \bar{t}|^\alpha \\ \text{for } (t, x), (\bar{t}, x) \in [0, T] \times \Omega.$$

Let  $u(t, x)$  be the solution on  $[0, T] \times \Omega$  of the equation

$$(1.14) \quad \begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{1}{2} \Delta u(t, x) + F(t, x) \cdot \nabla u(t, x) + J(t, x) u(t, x), \\ u(0, x) = \sigma_0(x), \\ u(t, x)|_{\partial\Omega} = 0. \end{cases}$$

For any  $0 \leq \tau \leq T$ , let  $\mathcal{P}_k = \{0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_k = \tau\}$  be a partition of  $[0, \tau]$ , where  $\tau_i = \frac{i\tau}{k}$ . Let  $u_i(t, x)$  be the solution on  $[\tau_{i-1}, \tau_i] \times \Omega$  of the equation

$$(1.15) \quad \begin{cases} \frac{\partial u_i}{\partial t}(t, x) = \frac{1}{2} \Delta u_i(t, x) + F(\tau_{i-1}, x) \cdot \nabla u_i(t, x) + J(\tau_{i-1}, x) u_i(t, x), \\ u_i(\tau_{i-1}, x) = u_{i-1}(\tau_{i-1}, x), \\ u_i(t, x)|_{\partial\Omega} = 0. \end{cases}$$

Here we use the convention  $u_0(t, x) = \sigma(x)$ . Then the solution  $u(t, x)$  of (1.14) can be computed by means of the solution  $u_i(t, x)$  of (1.15). More specifically,  $u(\tau, x) = \lim_{k \rightarrow \infty} u_k(\tau, x)$  in the  $L^1$  sense on  $\Omega$  and the following estimate holds:

$$(1.16) \quad \int_{\Omega} |u - u_k|(\tau_k, x) \leq \frac{2c_2}{\alpha + 1} \frac{T^{\alpha+1} e^{cT}}{k^\alpha},$$

where

$$(1.17) \quad c_2 = c_1 e^{cT} + c_1 \sqrt{\operatorname{Vol}(\Omega)} e^{c^2 T} \sqrt{2c^2 T \int_{\Omega} u^2(0, x) + \int_{\Omega} |\nabla u(0, x)|^2}.$$

The right-hand side of (1.16) goes to zero as  $k \rightarrow \infty$ .

In case (1.14) and (1.15) are DMZ equations, i.e.,  $F(t, x) = -f(x) + \nabla K$  and  $J(t, x) = -\operatorname{div} f - \frac{1}{2}|h|^2 + \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2$ , by Proposition 2.1 below (which is similar to Proposition 3.1 of [Ya-Ya]),  $u_i(\tau_i, x)$  can be computed by  $\tilde{u}_i(\tau_i, x)$ , where  $\tilde{u}_i(t, x)$  for  $\tau_{i-1} \leq t \leq \tau_i$  satisfies the Kolmogorov equation

(1.18)

$$\begin{cases} \frac{\partial \tilde{u}_i}{\partial t}(t, x) = \frac{1}{2}\Delta \tilde{u}_i(t, x) - \sum_{j=1}^n f_j(x) \frac{\partial \tilde{u}_i}{\partial x_j}(t, x) - \left( \operatorname{div} f(x) + \frac{1}{2} \sum_{j=1}^m h_j^2(x) \right) \tilde{u}_i(t, x), \\ \tilde{u}_i(\tau_{i-1}, x) = \exp \left( \sum_{j=1}^m (y_j(\tau_{i-1}) - y_j(\tau_{i-2})) h_j(x) \right) \tilde{u}_{i-1}(\tau_{i-1}, x). \end{cases}$$

In fact

$$(1.19) \quad u_i(\tau_i, x) = \exp \left( - \sum_{j=1}^m y_j(\tau_{i-1}) h_j(x) \right) \tilde{u}_i(\tau_i, x).$$

Therefore theoretically to solve the DMZ equation in a real time manner, we only need to compute the following Kolmogorov equation off-line:

$$\begin{cases} \frac{\partial \tilde{u}}{\partial t}(t, x) = \frac{1}{2}\Delta \tilde{u}(t, x) - \sum_{j=1}^n f_j(x) \frac{\partial \tilde{u}}{\partial x_j}(t, x) - \left( \operatorname{div} f(x) + \frac{1}{2} \sum_{j=1}^m h_j^2(x) \right) \tilde{u}(t, x), \\ \tilde{u}(0, x) = \phi_i(x), \end{cases}$$

where  $\{\phi_i(x)\}$  is an orthonormal base in  $L^2(\mathbb{R}^n)$ . The only real time computation here is to express arbitrary initial condition  $\phi(x)$  as the linear combination of  $\phi_i(x)$ . But this can be done by means of parallel computation.

The idea of solving the Kolmogorov equation “off-line” for the elements of an orthogonal basis has a substantial history; see, for example, [L-M-R] and the references therein. In the Lototsky–Mikulevicius–Rozovskii [L-M-R] approach, the authors used the Cameron–Martin expansion for the solution of the DMZ equation. Unfortunately, to determine the coefficients of the expansion, they need to consider a system of Kolmogorov-type equations which is a recursive system. The advantage of our method is that we need to deal with only one Kolmogorov equation.

**THEOREM D.** *Let  $u_R$  be the solution of (1.9), the DMZ equation on  $B_R$ . Assume the following:*

- (1)  $f(x)$  and  $h(x)$  have at most polynomial growth.
- (2) For any  $0 \leq t \leq T$ , there exist positive integer  $m$  and positive constants  $c'$  and  $c''$  independent of  $R$  such that the following two inequalities hold on  $\mathbb{R}^n$ :
  - (a)  $\frac{m^2}{2}|x|^{2m-2} - \frac{m}{2}(m+n-2)|x|^{m-2} - m|x|^{m-2}x \cdot (f - \nabla K) - \frac{\Delta K}{2} - \frac{1}{2}|h|^2 - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \geq -c'$ .
  - (b)  $|\frac{m^2|x|^{2m-2}}{2} - \frac{m(m+n-2)}{2}|x|^{m-2} - m|x|^{m-2}(f - \nabla K) \cdot x| \leq \frac{m(m+1)}{2}|x|^{2m-2} + c''$ .
- (3)  $-\frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - \sum_{j=1}^n f_j \frac{\partial K}{\partial x_j} + \frac{1}{2}|\nabla K|^2 \leq c_1$  for all  $(t, x) \in [0, T] \times \mathbb{R}^n$ , where  $c_1$  is a constant possibly depending on  $T$ .

Then for any  $R_0 < R$ ,

$$\begin{aligned} & \int_{B_{R_0}} (e^{-|x|^m} - e^{-R_0^m}) u_R(T, x) \\ & \geq e^{-c'T} \int_{B_{R_0}} (e^{-|x|^m} - e^{-R_0^m}) \sigma_0(x) \\ & \quad + \frac{e^{-R_0^m}}{c'} \left( \frac{m(m+1)}{2} R_0^{2m-2} + c'' \right) (1 - e^{c'T}) \int_{B_R} \sigma_0(x). \end{aligned}$$

In particular, the solution  $u$  of the robust DMZ equation on  $\mathbb{R}^n$  has the estimate

$$\int_{\mathbb{R}^n} e^{-|x|^m} u(T, x) \geq e^{-c'T} \int_{\mathbb{R}^n} e^{-|x|^m} \sigma_0(x).$$

In practical nonlinear filtering computation, it is important to know how much density remains within the given ball. Theorem D provides such a lower estimate. In particular, the solution  $u$  of the DMZ equation in  $\mathbb{R}^n$  obtained by taking  $\lim_{R \rightarrow \infty} u_R$ , where  $u_R$  is the solution of the DMZ equation in the ball  $B_R$ , is a nontrivial solution.

In the appendix, we give a priori estimation of derivatives of the solution of the DMZ equation up to second order. As a consequence we prove the existence of a weak solution of the DMZ equation. The uniqueness of the weak solution is shown in Appendix C.

Existence and uniqueness of solutions to the robust DMZ equation (1.6) have been treated by many well-known authors, including Pardoux [Pa1], [Pa2], Chaleyat-Maurel, Michel, and Pardoux [C-M-P], Rozovskii [Ro], Bensoussan [Be], Fleming and Mitter [Fl-Mi], Sussmann [Su], Michel [Mi], and Baras, Blankenship, and Hopkins [B-B-H]. They all obtained important estimates on the DMZ equation under special conditions. For example, Fleming and Mitter [Fl-Mi] treated the case where  $f$  and  $\nabla f$  are bounded, while Michel [Mi] analyzed regularity properties of solutions to DMZ equations with bounded  $f$  and  $h$ . Pardoux's earlier paper [Pa1] treated the case  $f, h$  bounded using arguments based on coercivity. It also contains many other interesting ideas. Pardoux [Pa2] has also treated nonlinear filtering problems with unbounded coefficients ( $f, h$  have linear growth). Starting with methods somewhat like those used by [Pa3], Baras, Blankenship, and Hopkins also obtained important results on existence, uniqueness, and asymptotic behavior of solutions to a class of DMZ equations with unbounded coefficients. However, they focused on only one spatial dimension and their result cannot cover the linear case. The Sobolev space setup of Appendices B and C in this paper is quite standard in partial differential equations and has been used by many people; see, for example, [Pa1].

The splitting up method has been used extensively by many authors. This technique is like the Trotter product formula from semigroup theory. Hopkins and Wong [Ho-Wo] used the Trotter product formula to study nonlinear filtering. The approximation method proposed for the DMZ equation, that of operator splitting, has a history going back to Bensoussan, Glowinski, and Rascanu [B-G-R1], [B-G-R2]. More recent articles on operator splitting methods in nonlinear filtering are [Gy-Kr], [Na], [It], [It-Ro]. Rates of convergence and “true” numerical schemes are developed in [Fl-Le], [It], and [It-Ro]. As pointed out by Bensoussan, Glowinski, and Rascanu [B-G-R1, section 4.3, p. 1431] the method bears the serious limitation that  $h$  must be bounded. The numerics of the Kushner–Stratonovitch equations were studied by many people. Two highly competitive classes of methods are “particle methods” (see, for example, [D-J-P] and [Cr-Ly]), in which particles move according to the signal dynamics and are weighted, killed, or duplicated according to their likelihood, and “discrete state” approximations (see, for example, [Ku] and [Pa-Ph]). These methods work nicely under the assumption that  $h$  is bounded (cf. [D-J-P, p. 348]).

**2. Some basic results.** In this section, we recall some results from our previous paper. The following proposition plays a fundamental role in our real time solution to the robust DMZ equation (1.6) in a memoryless manner.

PROPOSITION 2.1.  $\tilde{u}(t, x)$  satisfies the Kolmogorov equation

$$(2.1) \quad \frac{\partial \tilde{u}}{\partial t}(t, x) = \frac{1}{2} \Delta \tilde{u}(t, x) - f(x) \cdot \nabla \tilde{u}(t, x) - \left( \operatorname{div} f(x) + \frac{1}{2} \sum_{i=1}^m h_i^2(x) \right) \tilde{u}(t, x)$$

for  $\tau_{\ell-1} \leq t \leq \tau_\ell$  if and only if

$$(2.2) \quad u(t, x) = e^{-\sum_{i=1}^m y_i(\tau_{\ell-1}) h_i(x)} \tilde{u}(t, x)$$

satisfies the robust DMZ equation with observation being frozen at  $y(\tau_{\ell-1})$ :

$$(2.3) \quad \begin{aligned} \frac{\partial u}{\partial t}(t, x) = & \frac{1}{2} \Delta u(t, x) + (-f(x) + \nabla K(\tau_{\ell-1}, x)) \cdot \nabla u(t, x) \\ & + \left( -\operatorname{div} f(x) - \frac{1}{2} |h(x)|^2 + \frac{1}{2} \Delta K(\tau_{\ell-1}, x) \right. \\ & \left. - f(x) \cdot \nabla K(\tau_{\ell-1}, x) + \frac{1}{2} |\nabla K(\tau_{\ell-1}, x)|^2 \right) u(t, x). \end{aligned}$$

*Proof.* Proposition 2.1 is the left-hand version of Proposition 3.1 in [Ya-Ya]. The proof is a straightforward computation.  $\square$

We remark that (2.3) is obtained from the robust DMZ equation by freezing the observation term  $y(t)$  to  $y(\tau_{\ell-1})$ . We shall show that the solution of (2.3) approximates the solution of the robust DMZ equation very well in the  $L^1$  sense.

Suppose that  $u(t, x)$  is the solution of the robust DMZ equation and we want to compute  $u(\tau, x)$ . Let  $\mathcal{P}_k = \{0 = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_k = \tau\}$  be a partition of  $[0, \tau]$ , where  $\tau_i = \frac{i\tau}{k}$ . Let  $u_i(t, x)$  be a solution of the following partial differential equation for  $\tau_{i-1} \leq t \leq \tau_i$ :

$$(2.4) \quad \begin{cases} \frac{\partial u_i}{\partial t}(t, x) = \frac{1}{2} \Delta u_i(t, x) + (-f(x) + \nabla K(\tau_{i-1}, x)) \cdot \nabla u_i(t, x) \\ \quad + \left( -\operatorname{div} f(x) - \frac{1}{2} |h(x)|^2 + \frac{1}{2} \Delta K(\tau_{i-1}, x) \right. \\ \quad \left. - f(x) \cdot \nabla K(\tau_{i-1}, x) + \frac{1}{2} |\nabla K(\tau_{i-1}, x)|^2 \right) u_i(t, x), \\ u_i(\tau_{i-1}, x) = u_{i-1}(\tau_{i-1}, x). \end{cases}$$

In section 4 below we shall show that  $u(\tau, x) = \lim_{k \rightarrow \infty} u_k(\tau_k, x)$  in the  $L^1$  sense. By Proposition 2.1,  $u_1(\tau_1, x)$  can be computed by  $\tilde{u}_1(\tau_1, x)$ , where  $\tilde{u}_1(t, x)$  for  $0 \leq t \leq \tau_1$  satisfies (2.1) with initial condition

$$(2.5) \quad \tilde{u}_1(0, x) = \sigma_0(x).$$

In fact

$$(2.6) \quad u_1(\tau_1, x) = \tilde{u}_1(\tau_1, x).$$

In general Proposition 2.1 tells us that for  $i \geq 2$ ,  $u_i(\tau_i, x)$  can be computed by  $\tilde{u}_i(\tau_i, x)$ , where  $\tilde{u}_i(t, x)$  for  $\tau_{i-1} \leq t \leq \tau_i$  satisfies (2.1) with initial condition

$$(2.7) \quad \tilde{u}_i(\tau_{i-1}, x) = \exp \left[ \sum_{j=1}^m (y_j(\tau_{i-1}) - y_j(\tau_{i-2})) h_j(x) \right] \tilde{u}_{i-1}(\tau_{i-1}, x),$$

where the last initial condition comes from

$$\begin{aligned}
\tilde{u}_i(\tau_{i-1}, x) &= u_i(\tau_{i-1}, x) \exp \left( \sum_{j=1}^m y_j(\tau_{i-1}) h_j(x) \right) \\
&= u_{i-1}(\tau_{i-1}, x) \exp \left( \sum_{j=1}^m y_j(\tau_{i-1}) h_j(x) \right) \\
&= \exp \left( - \sum_{j=1}^m y_j(\tau_{i-2}) h_j(x) \right) \tilde{u}_{i-1}(\tau_{i-1}, x) \exp \left( \sum_{j=1}^m y_j(\tau_{i-1}) h_j(x) \right) \\
&= \exp \left[ \sum_{j=1}^m (y_j(\tau_{i-1}) - y_j(\tau_{i-2})) h_j(x) \right] \tilde{u}_{i-1}(\tau_{i-1}, x).
\end{aligned}$$

In fact,

$$(2.8) \quad u_i(\tau_i, x) = \exp \left( - \sum_{j=1}^m y_j(\tau_{i-1}) h_j(x) \right) \tilde{u}_i(\tau_i, x).$$

**3. Reduction of the problem to the bounded domain case.** In this section, we shall prove that in order to solve the robust DMZ equation (1.6) in  $\mathbb{R}^n$ , it suffices to solve the same equation in a bounded ball  $B_R$  with radius  $R$ . The important points here are that we know how large the  $R$  needs to be and that a precise error estimate is given. These are the essential ingredients for a successful implementation of nonlinear filters.

*Proof of Theorem A.* Let  $\phi$  be a  $C^\infty$  function on  $\mathbb{R}^n$  and  $B_R = \{x \in \mathbb{R}^n : |x| \leq R\}$ .

Let  $u_R$  be the solution of (1.9), the DMZ equation on the ball  $B_R$ :

$$\begin{aligned}
\frac{d}{dt} \int_{B_R} e^\phi u_R &= \frac{1}{2} \int_{B_R} e^\phi \Delta u_R + \int_{B_R} e^\phi (-f + \nabla K) \cdot \nabla u_R \\
&\quad + \int_{B_R} \left( -\operatorname{div} f - \frac{1}{2} |h|^2 + \frac{1}{2} \Delta K - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) e^\phi u_R \\
&= -\frac{1}{2} \int_{B_R} e^\phi \nabla \phi \cdot \nabla u_R + \frac{1}{2} \int_{\partial B_R} e^\phi \frac{\partial u_R}{\partial \nu} - \int_{B_R} \operatorname{div} [e^\phi (-f + \nabla K)] u_R \\
&\quad + \int_{\partial B_R} e^\phi u_R (-f + \nabla K) \cdot \nu \\
&\quad + \int_{B_R} e^\phi u_R \left( -\operatorname{div} f - \frac{1}{2} |h|^2 + \frac{1}{2} \Delta K - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) \\
&= \frac{1}{2} \int_{B_R} \operatorname{div} [e^\phi \nabla \phi] u_R - \frac{1}{2} \int_{\partial B_R} u_R e^\phi \nabla \phi \cdot \nu + \frac{1}{2} \int_{\partial B_R} e^\phi \frac{\partial u_R}{\partial \nu} \\
&\quad - \int_{B_R} e^\phi \nabla \phi \cdot (-f + \nabla K) u_R - \int_{B_R} e^\phi (-\operatorname{div} f + \Delta K) u_R
\end{aligned}$$



$$\begin{aligned}
& + \int_{\partial B_R} u_R e^\phi (-f + \nabla K) \cdot \nu \\
& + \int_{B_R} e^\phi u_R \left( -\operatorname{div} f - \frac{1}{2}|h|^2 + \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right) \\
& = \frac{1}{2} \int_{B_R} e^\phi u_R (\Delta \phi + |\nabla \phi|^2) + \int_{B_R} e^\phi u_R \nabla \phi \cdot (f - \nabla K) \\
& + \int_{B_R} e^\phi u_R \left( -\frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right) \\
& - \frac{1}{2} \int_{\partial B_R} e^\phi u_R \nabla \phi \cdot \nu + \frac{1}{2} \int_{\partial B_R} e^\phi \frac{\partial u_R}{\partial \nu} + \int_{\partial B_R} e^\phi u_R (-f + \nabla K) \cdot \nu,
\end{aligned}$$

where  $\nu$  is the unit outward normal of  $\partial B_R$ . Choose  $\phi = \sqrt{1+|x|^2}$ . Then  $\phi_i = \frac{x_i}{\sqrt{1+|x|^2}}$ ,  $\phi_{ii} = \frac{1}{\sqrt{1+|x|^2}} - \frac{x_i^2}{(1+|x|^2)^{3/2}}$ . Recall that  $u|_{\partial B_R} = 0$  and  $\frac{\partial u_R}{\partial \nu}|_{\partial B_R} \leq 0$ . It follows that

$$\begin{aligned}
\frac{d}{dt} \int_{B_R} e^\phi u_R & \leq \int_{B_R} e^\phi u_R \left[ -\frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right. \\
& \quad \left. + \frac{1}{2}\Delta \phi + \frac{1}{2}|\nabla \phi|^2 + \nabla \phi \cdot (f - \nabla K) \right] \\
& = \int_{B_R} e^\phi u_R \left[ -\frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 + \frac{n}{2\sqrt{1+|x|^2}} \right. \\
& \quad \left. - \frac{|x|^2}{2(1+|x|^2)^{3/2}} + \frac{1}{2} \frac{|x|^2}{1+|x|^2} + \frac{x}{\sqrt{1+|x|^2}} (f - \nabla K) \right] \\
& \leq \int_{B_R} e^\phi u_R \left[ -\frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 + \frac{n+1}{2} + |f - \nabla K| \right] \\
& \leq \left( c_1 + \frac{n+1}{2} \right) \int_{B_R} e^\phi u_R.
\end{aligned}$$

Hence

$$\int_{B_R} e^\phi u_R(t, x) \leq e^{(c_1 + \frac{n+1}{2})t} \int_{B_R} e^\phi u_R(0, x) \quad \forall t \in [0, T].$$

Let  $R$  go to infinity. We have

$$\int_{\mathbb{R}^n} e^\phi u(t, x) \leq e^{(c_1 + \frac{n+1}{2})t} \int_{\mathbb{R}^n} e^\phi u(0, x) \quad \forall t \in [0, T],$$

which implies

$$\sup_{0 \leq t \leq T} \int_{\mathbb{R}^n} e^\phi u(t, x) \leq e^{(c_1 + \frac{n+1}{2})T} \int_{\mathbb{R}^n} e^\phi u(0, x).$$

In particular

$$\begin{aligned}
e^{\sqrt{1+R^2}} \sup_{0 \leq t \leq T} \int_{|x| \geq R} u(t, x) & \leq \sup_{0 \leq t \leq T} \int_{|x| \geq R} e^\phi u(t, x) \leq \sup_{0 \leq t \leq T} \int_{\mathbb{R}^n} e^\phi u(t, x) \\
& \leq e^{(c_1 + \frac{n+1}{2})T} \int_{\mathbb{R}^n} e^\phi u(0, x).
\end{aligned}$$

This implies

$$\sup_{0 \leq t \leq T} \int_{|x| \geq R} u(t, x) \leq e^{-\sqrt{1+R^2} e^{(c_1 + \frac{n+1}{2})T}} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(0, x). \quad \square$$

Theorem A says that we can choose  $R$  large enough so that  $\sup_{0 \leq t \leq T} \int_{|x| \geq R} u(t, x)$  is arbitrarily small. For numerical calculation, we can restrict the DMZ equation to the ball  $B_R$ . In fact we can prove Theorem B, which states that  $u_R$  is a good approximation of  $u$  when  $R$  is large.

*Proof of Theorem B.* By the maximum principle (cf. Theorem 1, p. 34 in Friedman's book [Fr]), we have  $v \geq 0$  for  $(t, x) \in [0, T] \times B_R$  since  $v|_{\partial B_R} \geq 0$  for  $0 \leq t \leq T$ :

$$\begin{aligned} \frac{d}{dt} \int_{B_R} \psi v &= \int_{B_R} \psi \frac{dv}{dt} \\ &= \frac{1}{2} \int_{B_R} \psi \Delta v + \int_{B_R} \psi (-f + \nabla K) \cdot \nabla v \\ &\quad + \int_{B_R} \left( -\operatorname{div} f - \frac{1}{2} |h|^2 + \frac{1}{2} \Delta K - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) \psi v \\ &= -\frac{1}{2} \int_{B_R} \nabla \psi \cdot \nabla v + \frac{1}{2} \int_{\partial B_R} \psi \frac{\partial v}{\partial \nu} - \int_{B_R} \operatorname{div} [\psi (-f + \nabla K)] v \\ &\quad + \int_{\partial B_R} \psi v (-f + \nabla K) \cdot \nu \\ &\quad + \int_{B_R} \left( -\operatorname{div} f - \frac{1}{2} |h|^2 + \frac{1}{2} \Delta K - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) \psi v \\ &= \frac{1}{2} \int_{B_R} (\Delta \psi) v - \frac{1}{2} \int_{\partial B_R} v \frac{\partial \psi}{\partial \nu} + \frac{1}{2} \int_{\partial B_R} \psi \frac{\partial v}{\partial \nu} - \int_{B_R} \nabla \psi \cdot (-f + \nabla K) v \\ &\quad - \int_{B_R} \psi (-\operatorname{div} f + \Delta K) v + \int_{\partial B_R} \psi v (-f + \nabla K) \cdot \nu \\ &\quad + \int_{B_R} \left( -\operatorname{div} f - \frac{1}{2} |h|^2 + \frac{1}{2} \Delta K - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) \psi v. \end{aligned}$$

Let  $\phi$  be a radial symmetric function such that  $\phi|_{\partial B_R} = R$ ,  $\nabla \phi|_{\partial B_R} = 0$ , and  $\phi$  is increasing with  $|x|$ . Let

$$\psi = e^{-\phi(x)} - e^{-R}.$$

Then  $\psi|_{\partial B_R} = 0$  and  $\nabla \psi|_{\partial B_R} = 0$ . Hence

$$\begin{aligned} \frac{d}{dt} \int_{B_R} \psi v &= \frac{1}{2} \int_{B_R} (\Delta \psi) v - \int_{B_R} \nabla \psi \cdot (-f + \nabla K) v \\ &\quad + \int_{B_R} \left( -\frac{1}{2} |h|^2 - \frac{1}{2} \Delta K - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) \psi v \\ &= \frac{1}{2} \int_{B_R} v e^{-\phi} (-\Delta \phi + |\nabla \phi|^2) - \int_{B_R} e^{-\phi} v [\nabla \phi \cdot (f - \nabla K)] \end{aligned}$$

$$\begin{aligned}
 & + \int_{B_R} \left( -\frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right) \psi v \\
 & = \int_{B_R} \psi v \left[ -\frac{1}{2}\Delta\phi + \frac{1}{2}|\nabla\phi|^2 - \nabla\phi \cdot (f - \nabla K) \right. \\
 & \quad \left. - \frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right] \\
 & \quad + e^{-R} \int_{B_R} \left[ \frac{1}{2}(-\Delta\phi + |\nabla\phi|^2) - \nabla\phi \cdot (f - \nabla K) \right] v \\
 & \leq \sup_{B_R} \left[ -\frac{1}{2}\Delta\phi + \frac{1}{2}|\nabla\phi|^2 - \nabla\phi \cdot (f - \nabla K) \right. \\
 & \quad \left. - \frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right] \cdot \int_{B_R} \psi v \\
 & \quad + e^{-R} \sup_{B_R} \left\{ e^{-\sqrt{1+|x|^2}} \left[ \frac{1}{2}(-\Delta\phi + |\nabla\phi|^2) - \nabla\phi \cdot (f - \nabla K) \right] \right\} \\
 & \quad \times \int_{B_R} e^{\sqrt{1+|x|^2}} v.
 \end{aligned}$$

Observe that  $0 \leq v \leq u$  for  $(t, x) \in [0, T] \times B + R$ . Let

$$\chi(x) = 1 - (1 - x)^2 \quad \text{and} \quad \phi(x) = R\chi\left(\frac{|x|^2}{R^2}\right).$$

Then

$$\begin{aligned}
 & \chi'(x) = 2(1 - x), \quad \chi''(x) = -2, \quad \chi(1) = 1, \quad \chi'(1) = 0, \\
 & \nabla\phi(x) = \frac{2x}{R}\chi'\left(\frac{|x|^2}{R^2}\right), \quad \Delta\phi = \frac{4|x|^2}{R^3}\chi''\left(\frac{|x|^2}{R^2}\right) + \frac{2n}{R}\chi'\left(\frac{|x|^2}{R^2}\right), \\
 & \sup_{B_R} \left[ -\frac{1}{2}\Delta\phi + \frac{1}{2}|\nabla\phi|^2 - \nabla\phi \cdot (f - \nabla K) - \frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right] \\
 & = \sup_{B_R} \left\{ -\frac{2|x|^2}{R^3}\chi''\left(\frac{|x|^2}{R^2}\right) - \frac{n}{R}\chi'\left(\frac{|x|^2}{R^2}\right) + 2\frac{|x|^2}{R^2} \left[ \chi'\left(\frac{|x|^2}{R^2}\right) \right]^2 \right. \\
 & \quad \left. - \frac{2}{R}\chi'\left(\frac{|x|^2}{R^2}\right) [x \cdot f - x \cdot \nabla K] - \frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right\} \\
 & \leq \sup_{B_R} \left[ \frac{4|x|^2}{R^3} + \frac{2n}{R} + \frac{8|x|^2}{R} + \frac{4}{R}|x||f - \nabla K| - \frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right] \\
 & \leq 12 + 2n + 4|x||f - \nabla K| - \frac{1}{2}|h|^2 - \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \\
 & \leq c_2.
 \end{aligned}$$

Similarly

$$\begin{aligned} & \sup_{B_R} \left\{ e^{-\sqrt{1+|x|^2}} \left[ \frac{1}{2}(-\Delta\phi + |\nabla\phi|^2) - \nabla\phi \cdot (f - \nabla K) \right] \right\} \\ & \leq \sup_{B_R} \left\{ e^{-\sqrt{1+|x|^2}} [12 + 2n + 4|x||f - \nabla K|] \right\} \leq c_3. \end{aligned}$$

In view of Theorem A, we have

$$\begin{aligned} \frac{d}{dt} \int_{B_R} \psi v & \leq c_2 \int_{B_R} \psi v + e^{-R} c_3 \int_{B_R} e^{\sqrt{1+|x|^2}} u \\ & \leq c_2 \int_{B_R} \psi v + e^{-R} c_3 e^{(c_1 + \frac{n+1}{2})T} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(0, x), \\ \frac{d}{dt} \left[ e^{-c_2 t} \int_{B_R} \psi v \right] & = e^{-c_2 t} \frac{d}{dt} \int_{B_R} \psi v - c_2 e^{-c_2 t} \int_{B_R} \psi v \\ & \leq c_3 e^{-R} e^{-c_2 t} e^{c_1 + \frac{n+1}{2}T} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(0, x). \end{aligned}$$

Recall that  $v(0, x) = 0$  on  $B_R$ . Therefore we have

$$e^{-c_2 T} \int_{B_R} \psi v(T, x) \leq \frac{e^{-c_2 T} - 1}{-c_2} c_3 e^{-R} e^{(c_1 + \frac{n+1}{2})T} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(0, x),$$

which implies

$$\int_{B_R} \psi v(T, x) \leq \frac{e^{c_2 T} - 1}{c_2} c_3 e^{-R} e^{(c_1 + \frac{n+1}{2})T} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(0, x).$$

Notice that  $\psi(x) = e^{\frac{|x|^4}{R^3} - \frac{2|x|^2}{R}} - e^{-R}$ :

$$\begin{aligned} \int_{B_R} \psi v(T, x) & \geq \int_{B_{\frac{R}{2}}} [e^{\frac{|x|^4}{R^3} - \frac{2|x|^2}{R}} - e^{-R}] v(T, x) \\ & \geq (e^{-\frac{7}{16}R} - e^{-R}) \int_{B_{\frac{R}{2}}} v(T, x) \\ & = e^{-\frac{7}{16}R} (1 - e^{-\frac{9}{16}R}) \int_{B_{\frac{R}{2}}} v(T, x) \\ & \geq \frac{1}{2} e^{-\frac{7}{16}R} \int_{B_{\frac{R}{2}}} v(T, x). \end{aligned}$$

Therefore

$$\int_{B_{\frac{R}{2}}} v(T, x) \leq \frac{2(e^{c_2 T} - 1)}{c_2} c_3 e^{-\frac{9}{16}R} e^{(c_1 + \frac{n+1}{2})T} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} u(0, x). \quad \square$$

Theorem B above says that if we replace  $u$  by  $u_R$ , then the error is small when  $R$  goes to infinity. In fact (1.11) gives the precise error estimate.

**4.  $L^1$ -convergence.** In this section, we shall show that our algorithm described in section 2 will yield an  $L^1$ -convergence for bounded domains, i.e.,  $u(\tau, x) = \lim_{k \rightarrow \infty} u_k(\tau_k, x)$  in the  $L^1$  sense for bounded domain. We first begin with the following technical lemma.

LEMMA 4.1. *Let  $\Omega$  be a bounded domain in  $\mathbb{R}^n$  and let  $v: [0, T] \times \overline{\Omega} \rightarrow \mathbb{R}$  be a  $C^1$  function. Assume that  $v(t, x) = 0$  for  $(t, x) \in [0, T] \times \partial\Omega$ . Let  $\Omega_t^+ = \{x \in \Omega: v(t, x) \geq 0\}$ . Then*

$$\frac{d}{dt} \int_{\Omega_t^+} v(t, x) = \int_{\Omega_t^+} \frac{dv}{dt}(t, x) \quad \text{for almost all } t \in [0, t].$$

*Proof.*

$$\begin{aligned} \frac{d}{dt} \int_{\Omega_t^+} v(t, x) &= \lim_{\Delta t \rightarrow 0} \frac{\int_{\Omega_{t+\Delta t}^+} v(t + \Delta t, x) - \int_{\Omega_t^+} v(t, x)}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \int_{\Omega_{t+\Delta t}^+} \frac{v(t + \Delta t, x) - v(t, x)}{\Delta t} + \lim_{\Delta t \rightarrow 0} \frac{\int_{\Omega_{t+\Delta t}^+} v(t, x) - \int_{\Omega_t^+} v(t, x)}{\Delta t} \\ &= \int_{\Omega_t^+} \frac{dv}{dt}(t, x) + \lim_{\Delta t \rightarrow 0} \frac{\int_{\Omega_{t+\Delta t}^+ - \Omega_t^+} v(t, x) - \int_{\Omega_t^+ - \Omega_{t+\Delta t}^+} v(t, x)}{\Delta t} \\ &= \int_{\Omega_t^+} \frac{dv}{dt}(t, x) + \lim_{\Delta t \rightarrow 0} \frac{v(t, \xi_1) \text{Vol}(\Omega_{t+\Delta t}^+ - \Omega_t^+)}{\Delta t} \\ &\quad - \lim_{\Delta t \rightarrow 0} \frac{v(t, \xi_2) \text{Vol}(\Omega_t^+ - \Omega_{t+\Delta t}^+)}{\Delta t}, \end{aligned} \tag{4.1}$$

where  $\xi_1 \in \Omega_{t+\Delta t}^+ - \Omega_t^+$  and  $\xi_2 \in \Omega_t^+ - \Omega_{t+\Delta t}^+$ . Clearly we have  $\lim_{\Delta t \rightarrow 0} v(t, \xi_1) = 0 = \lim_{\Delta t \rightarrow 0} v(t, \xi_2)$ . Therefore it remains to prove that

$$\lim_{\Delta t \rightarrow 0} \frac{\text{Vol}(\Omega_{t+\Delta t}^+ - \Omega_t^+)}{\Delta t} \quad \text{and} \quad \lim_{\Delta t \rightarrow 0} \frac{\text{Vol}(\Omega_t^+ - \Omega_{t+\Delta t}^+)}{\Delta t}$$

are bounded for almost all  $t$ .

Let  $w: A \rightarrow \mathbb{R}$  be a Lipschitz function, where  $A \subset \mathbb{R}^n$  is measurable. The coarea formula for Euclidean space, an important tool of geometric measure theory (cf. [Fe], [Ma]), reads as follows:

$$\int_A h(x) |\nabla w(x)| dx = \int_{\mathbb{R}} \int_{w^{-1}(y)} h(x) \mathcal{H}_{|\cdot|}^{n-1}(x) dy, \tag{4.2}$$

where  $\mathcal{H}_{|\cdot|}^{n-1}$  denotes the Hausdorff measure with respect to the Euclidean distance and  $h: A \rightarrow [-\infty, \infty]$  is a measurable function.

Let  $A = \overline{\Omega_t^+ - \Omega_{t+\Delta t}^+} = \{x \in \Omega_t^+: v(t + \Delta t, x) \leq 0\}$ . Let  $L$  be the Lipschitz constant such that

$$|v(t, x) - v(t + \Delta t, x)| \leq L\Delta t \quad \forall x \in \overline{\Omega}. \tag{4.3}$$

Since  $v(t, x) \geq 0$  for  $x \in A$ , we have

$$v(t + \Delta t, x) \geq v(t, x) - L\Delta t \geq -L\Delta t \quad \text{for } x \in A. \tag{4.4}$$

Let  $h(x) = \frac{1}{|\nabla v(t+\Delta t, x)|}$  and  $w(x) = v(t+\Delta t, x)$  in the coarea formula (4.2). We have

$$(4.5) \quad \begin{aligned} \text{Vol}(\Omega_t^+ - \Omega_{t+\Delta t}^+) &= \text{Vol}(A) \\ &= \int_{-L\Delta t}^0 \int_{\{x \in A: v(t+\Delta t, x)=y\}} \frac{1}{|\nabla v(t+\Delta t, x)|} \mathcal{H}_{|\cdot|}^{n-1}(x) dy. \end{aligned}$$

Consider the map  $\Phi: [0, T] \times \overline{\Omega} \rightarrow [0, T] \times \mathbb{R}$  given by  $\Phi(t, x) = (t, v(t, x))$ . By Sard's theorem, the set of critical values of  $\Phi$  has Lebesgue measure zero. Therefore for almost all  $t$ , almost all  $\Delta t$ , and almost all  $y$ ,  $\nabla v(t+\Delta t, x) \neq 0$  for all  $x \in \{x: v(t+\Delta t, x) = y\}$ . It follows that  $\lim_{\Delta t \rightarrow 0} \frac{\text{Vol}(\Omega_t^+ - \Omega_{t+\Delta t}^+)}{\Delta t}$  is bounded for almost all  $t$ . Similarly one can prove that  $\lim_{\Delta t \rightarrow 0} \frac{\text{Vol}(\Omega_{t+\Delta t}^+ - \Omega_t^+)}{\Delta t}$  is bounded for almost all  $t$ .  $\square$

*Remark 4.2.* Lemma 4.1 is not true for all  $t \in [0, T]$ . As we shall see from the following example, this is because  $\lim_{\Delta t \rightarrow 0} \frac{\text{Vol}(\Omega_t^+ - \Omega_{t+\Delta t}^+)}{\Delta t}$  is not necessarily bounded for all  $t$ .

*Example 4.3.* Let  $0 < a < b < c < d < \infty$  and  $\overline{\Omega} = [a, d]$ . Let  $v(t, x) = (x-a)(x-b-\sqrt{t})(x-c-\sqrt{t})(x-d)$  be defined on  $[0, T] \times \overline{\Omega}$ . Then  $\Omega_t^+ = [b+\sqrt{t}, c+\sqrt{t}]$  and  $\Omega_{t+\Delta t} = [b+\sqrt{t+\Delta t}, c+\sqrt{t+\Delta t}]$ , and

$$\lim_{\Delta t \rightarrow 0} \frac{\text{length}(\Omega_t^+ - \Omega_{t+\Delta t}^+)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{\sqrt{t+\Delta t} - \sqrt{t}}{\Delta t} = \frac{1}{2\sqrt{t}},$$

which is finite except at  $t = 0$ .

*Proof of Theorem C.* Observe that (4.14) and (4.15) imply

$$(4.6) \quad \begin{aligned} \frac{\partial u}{\partial t}(t, x) - \frac{\partial u_i}{\partial t}(t, x) &= \frac{1}{2} \Delta(u - u_i)(t, x) + F(\tau_{i-1}, x) \cdot \nabla(u - u_i)(t, x) \\ &\quad + (F(t, x) - F(\tau_{i-1}, x)) \cdot \nabla u(t, x) + J(\tau_{i-1}, x)(u - u_i)(t, x) \\ &\quad + (J(t, x) - J(\tau_{i-1}, x))u(t, x). \end{aligned}$$

Let  $\Omega_t^+ = \{x \in \Omega: u(t, x) - u_i(t, x) \geq 0\}$ .

$$\begin{aligned} \frac{d}{dt} \int_{\Omega_t^+} (u - u_i)(t, x) &= \frac{1}{2} \int_{\Omega_t^+} \Delta(u - u_i)(t, x) + \int_{\Omega_t^+} F(\tau_{i-1}, x) \cdot \nabla(u - u_i)(t, x) \\ &\quad + \int_{\Omega_t^+} (F(t, x) - F(\tau_{i-1}, x)) \cdot \nabla u(t, x) \\ &\quad + \int_{\Omega_t^+} J(\tau_{i-1}, x)(u - u_i)(t, x) \\ &\quad + \int_{\Omega_t^+} (J(t, x) - J(\tau_{i-1}, x))u(t, x) \\ &= \frac{1}{2} \int_{\partial \Omega_t^+} \frac{\partial(u - u_i)}{\partial \nu}(t, x) - \int_{\Omega_t^+} \text{div} F(\tau_{i-1}, x)(u - u_i)(t, x) \\ &\quad + \int_{\Omega_t^+} (F(t, x) - F(\tau_{i-1}, x)) \cdot \nabla u(t, x) \end{aligned}$$

$$\begin{aligned}
 & + \int_{\Omega_t^+} J(\tau_{i-1}, x)(u - u_i)(t, x) \\
 & + \int_{\Omega_t^+} (J(t, x) - J(\tau_{i-1}, x))u(t, x) \\
 & \leq c \int_{\Omega_t^+} (u - u_i)(t, x) + c_1(t - \tau_{i-1})^\alpha \int_{\Omega_t^+} u(t, x) \\
 (4.7) \quad & + c_1(t - \tau_{i-1})^\alpha \int_{\Omega_t^+} |\nabla u(t, x)|.
 \end{aligned}$$

Notice that

$$\begin{aligned}
 \frac{d}{dt} \int_{\Omega} u(t, x) & = \frac{1}{2} \int_{\Omega} \Delta u + \int_{\Omega} F(t, x) \cdot \nabla u(t, x) + \int_{\Omega} J(t, x)u(t, x) \\
 & = \frac{1}{2} \int_{\partial\Omega} \frac{\partial u}{\partial \nu} - \int_{\Omega} \operatorname{div} F(t, x)u(t, x) + \int_{\Omega} J(t, x)u(t, x) \\
 (4.8) \quad & \leq c \int_{\Omega} u(t, x).
 \end{aligned}$$

This implies, for  $0 \leq t \leq T$ ,

$$(4.9) \quad \int_{\Omega} u(t, x) \leq e^{cT} \int_{\Omega} u(0, x).$$

In order to estimate  $\int_{\Omega} |\nabla u(t, x)|^2$ , we need to estimate the  $L^2$  norm of  $u(t, x)$ .

$$\begin{aligned}
 \frac{d}{dt} \int_{\Omega} u^2(t, x) & = 2 \int_{\Omega} u(t, x) \frac{\partial u}{\partial t}(t, x) \\
 (4.10) \quad & = \int_{\Omega} u(t, x) \Delta u(t, x) + 2 \int_{\Omega} u(t, x) F(t, x) \cdot \nabla u(t, x) + 2 \int_{\Omega} J(t, x)u^2(t, x).
 \end{aligned}$$

Observe that

$$\begin{aligned}
 \int_{\Omega} u(t, x) F(t, x) \cdot \nabla u(t, x) & = - \int_{\Omega} u(t, x) \operatorname{div}[u(t, x) F(t, x)] + \int_{\partial\Omega} u^2(t, x) F(t, x) \cdot \nu \\
 & = - \int_{\Omega} u(t, x) \nabla u(t, x) \cdot F(t, x) - \int_{\Omega} u^2(t, x) \operatorname{div} F(t, x).
 \end{aligned}$$

This implies

$$(4.11) \quad \int_{\Omega} u(t, x) F(t, x) \nabla u(t, x) = -\frac{1}{2} \int_{\Omega} u^2(t, x) \operatorname{div} F(t, x).$$

Putting (4.11) into (4.10), we get

$$\begin{aligned}
 \frac{d}{dt} \int_{\Omega} u^2(t, x) & = - \int_{\Omega} |\nabla u(t, x)|^2 + \int_{\partial\Omega} u(t, x) \frac{\partial u}{\partial \nu}(t, x) - \int_{\Omega} u^2(t, x) \operatorname{div} F(t, x) \\
 & \quad + 2 \int_{\Omega} J(t, x)u^2(t, x) \\
 (4.12) \quad & \leq c \int_{\Omega} u^2(t, x).
 \end{aligned}$$

This implies

$$(4.13) \quad \int_{\Omega} u^2(t, x) \leq e^{ct} \int_{\Omega} u^2(0, x) \leq e^{cT} \int_{\Omega} u^2(0, x).$$

Now we are ready to estimate  $\int_{\Omega} |\nabla u(t, x)|^2$ .

$$\begin{aligned}
 \frac{d}{dt} \int_{\Omega} |\nabla u|^2(t, x) &= \int_{\Omega} 2 \nabla \frac{\partial u}{\partial t}(t, x) \cdot \nabla u(t, x) \\
 &= -2 \int_{\Omega} \frac{\partial u}{\partial t}(t, x) \Delta u(t, x) + 2 \int_{\partial \Omega} \frac{\partial u}{\partial t}(t, x) \frac{\partial u}{\partial \nu}(t, x) \\
 &= - \int_{\Omega} (\Delta u(t, x))^2 - 2 \int_{\Omega} F(t, x) \cdot \nabla u(t, x) \Delta u(t, x) \\
 &\quad - 2 \int_{\Omega} J(t, x) u(t, x) \Delta u(t, x) \\
 &\leq - \int_{\Omega} (\Delta u(t, x))^2 + 2 \int_{\Omega} |F(t, x)|^2 |\nabla u(t, x)|^2 + \frac{1}{2} \int_{\Omega} (\Delta u(t, x))^2 \\
 &\quad + 2 \int_{\Omega} J^2(t, x) u^2(t, x) + \frac{1}{2} \int_{\Omega} (\Delta u(t, x))^2 \\
 &\leq 2c^2 \int_{\Omega} |\nabla u(t, x)|^2 + 2c^2 \int_{\Omega} u^2(t, x) \\
 (4.14) \quad &\leq 2c^2 \int_{\Omega} |\nabla u(t, x)|^2 + 2c^2 e^{ct} \int_{\Omega} u^2(0, x).
 \end{aligned}$$

This implies

$$\begin{aligned}
 \frac{d}{dt} \left[ e^{-2c^2 t} \int_{\Omega} |\nabla u(t, x)|^2 \right] &= e^{-2c^2 t} \left[ \frac{d}{dt} \int_{\Omega} |\nabla u(t, x)|^2 - 2c^2 \int_{\Omega} |\nabla u(t, x)|^2 \right] \\
 (4.15) \quad &\leq 2c^2 e^{-(2c^2 - c)t} \int_{\Omega} u^2(0, x) \leq 2c^2 \int_{\Omega} u^2(0, x).
 \end{aligned}$$

Hence

$$e^{-2c^2 t} \int_{\Omega} |\nabla u(t, x)|^2 - \int_{\Omega} |\nabla u(0, x)|^2 \leq 2c^2 t \int_{\Omega} u^2(0, x)$$

and

$$\begin{aligned}
 \int_{\Omega} |\nabla u(t, x)|^2 &\leq 2c^2 t e^{2c^2 t} \int_{\Omega} u^2(0, x) + e^{2c^2 t} \int_{\Omega} |\nabla u(0, x)|^2 \\
 (4.16) \quad &\leq 2c^2 T e^{2c^2 T} \int_{\Omega} u^2(0, x) + e^{2c^2 T} \int_{\Omega} |\nabla u(0, x)|^2.
 \end{aligned}$$

It follows that

$$\begin{aligned}
 \int_{\Omega} |\nabla u(t, x)| &\leq \sqrt{\text{Vol } (\Omega)} \left[ \int_{\Omega} |\nabla u(t, x)|^2 \right]^{\frac{1}{2}} \\
 (4.17) \quad &\leq \sqrt{\text{Vol } (\Omega)} e^{c^2 T} \sqrt{2c^2 T \int_{\Omega} u^2(0, x) + \int_{\Omega} |\nabla u(0, x)|^2}.
 \end{aligned}$$



Putting (4.9), (4.17) into (4.7), we get

$$\begin{aligned}
 \frac{d}{dt} \int_{\Omega_t^+} (u - u_i)(t, x) &\leq c \int_{\Omega_t^+} (u - u_i)(t, x) + c_1(t - \tau_{i-1})^\alpha \int_{\Omega} u(t, x) \\
 &\quad + c_1(t - \tau_{i-1})^\alpha \int_{\Omega} |\nabla u(t, x)| \\
 &\leq c \int_{\Omega_t^+} (u - u_i)(t, x) + c_1(t - \tau_{i-1})^\alpha e^{cT} \int_{\Omega} u(0, x) \\
 &\quad + c_1(t - \tau_{i-1})^\alpha \sqrt{\text{Vol}(\Omega)} e^{c^2 T} \\
 &\quad \sqrt{2c^2 T \int_{\Omega} u^2(0, x) + \int_{\Omega} |\nabla u(0, x)|^2} \\
 (4.18) \quad &= c \int_{\Omega_t^+} (u - u_i)(t, x) + c_2(t - \tau_{i-1})^\alpha,
 \end{aligned}$$

where

$$(4.19) \quad c_2 = c_1 e^{cT} \int_{\Omega} u(0, x) + c_1 \sqrt{\text{Vol}(\Omega)} e^{c^2 T} \sqrt{2c^2 T \int_{\Omega} u^2(0, x) + \int_{\Omega} |\nabla u(0, x)|^2},$$

$$\begin{aligned}
 &\frac{d}{dt} \left[ e^{-c(t-\tau_{i-1})} \int_{\Omega_t^+} (u - u_i)(t, x) \right] \\
 &= e^{-c(t-\tau_{i-1})} \left[ \frac{d}{dt} \int_{\Omega_t^+} (u - u_i)(t, x) - c \int_{\Omega_t^+} (u - u_i)(t, x) \right] \\
 &\leq c_2(t - \tau_{i-1})^\alpha e^{-c(t-\tau_{i-1})}.
 \end{aligned}$$

This implies

$$\begin{aligned}
 &e^{-c(t-\tau_{i-1})} \int_{\Omega_t^+} (u - u_i)(t, x) - \int_{\Omega_{\tau_{i-1}}^+} (u - u_i)(\tau_{i-1}, x) \\
 (4.20) \quad &\leq c_2 \int_{\tau_{i-1}}^t (s - \tau_{i-1})^\alpha e^{-c(s-\tau_{i-1})} \leq c_2 \frac{(t - \tau_{i-1})^{\alpha+1}}{\alpha + 1}.
 \end{aligned}$$

Hence

$$\begin{aligned}
 \int_{\Omega_t^+} (u - u_i)(t, x) &\leq e^{c(t-\tau_{i-1})} \int_{\Omega_{\tau_{i-1}}^+} (u - u_{i-1})(\tau_{i-1}, x) \\
 (4.21) \quad &+ c_2 \frac{(t - \tau_{i-1})^{\alpha+1}}{\alpha + 1} e^{c(t-\tau_{i-1})}.
 \end{aligned}$$

Similarly one can prove that

$$\begin{aligned}
 \int_{\Omega_t^-} (u - u_i)(t, x) &\leq e^{c(t-\tau_{i-1})} \int_{\Omega_{\tau_{i-1}}^-} (u_{i-1} - u)(\tau_{i-1}, x) \\
 (4.22) \quad &+ c_2 \frac{(t - \tau_{i-1})^{\alpha+1}}{\alpha + 1} e^{c(t-\tau_{i-1})}.
 \end{aligned}$$

Consequently, we have

$$\begin{aligned}
 \int_{\Omega} |u - u_i|(t, x) &= \int_{\Omega_t^+} (u - u_i)(t, x) + \int_{\Omega_t^-} (u_i - u)(t, x) \\
 (4.23) \quad &\leq e^{c(t-\tau_{i-1})} \left[ \int_{\Omega} |u - u_{i-1}|(\tau_{i-1}, x) + 2c_2 \frac{t - \tau_{i-1}}{\alpha + 1} \right]^{\alpha+1}.
 \end{aligned}$$

By applying (4.23) inductively, we have the following estimate:

$$\begin{aligned}
 &\int_{\Omega} |u - u_k|(\tau_k, x) \\
 &\leq e^{c(\tau_k - \tau_{k-1})} \left[ \int_{\Omega} |u - u_{k-1}|(\tau_{k-1}, x) + 2c_2 \frac{(\tau_k - \tau_{k-1})^{\alpha+1}}{\alpha + 1} \right] \\
 &\leq e^{c(\tau_k - \tau_{k-1})} e^{c(\tau_{k-1} - \tau_{k-2})} \left[ \int_{\Omega} |u - u_{k-2}|(\tau_{k-2}, x) + 2c_2 \frac{(\tau_{k-1} - \tau_{k-2})^{\alpha+1}}{\alpha + 1} \right] \\
 &\quad + e^{c(\tau_k - \tau_{k-1})} 2c_2 \frac{(\tau_k - \tau_{k-1})^{\alpha+1}}{\alpha + 1} \\
 &= e^{c(\tau_k - \tau_{k-2})} \int_{\Omega} |u - u_{k-2}|(\tau_{k-2}, x) + \frac{2c_2}{\alpha + 1} [(\tau_{k-1} - \tau_{k-2})^{\alpha+1} e^{c(\tau_k - \tau_{k-2})} \\
 &\quad + (\tau_k - \tau_{k-1})^{\alpha+1} e^{c(\tau_k - \tau_{k-1})}] \\
 &\leq e^{c(\tau_k - \tau_{k-i})} \int_{\Omega} |u - u_{k-i}|(\tau_{k-i}, x) \\
 &\quad + \frac{2c_2}{\alpha + 1} [(\tau_k - \tau_{k-1})^{\alpha+1} e^{c(\tau_k - \tau_{k-1})} + (\tau_{k-1} - \tau_{k-2})^{\alpha+1} e^{c(\tau_k - \tau_{k-2})} \\
 &\quad + \cdots + (\tau_{k-i+1} - \tau_{k-i})^{\alpha+1} e^{c(\tau_k - \tau_{k-i})}] \\
 &\leq e^{cT} \int_{\Omega} |u - u_0|(0, x) + \frac{2c_2}{\alpha + 1} [(\tau_k - \tau_{k-1})^{\alpha+1} e^{c(\tau_k - \tau_{k-1})} \\
 &\quad + (\tau_{k-1} - \tau_{k-2})^{\alpha+1} e^{c(\tau_k - \tau_{k-2})} + \cdots + (\tau_1 - \tau_0)^{\alpha+1} e^{c(\tau_k - \tau_0)}] \\
 &= \frac{2c_2}{\alpha + 1} \frac{T^{\alpha+1}}{k^{\alpha+1}} [e^{c\frac{T}{k}} + e^{c\frac{2T}{k}} + \cdots + e^{c\frac{kT}{k}}] \\
 (4.24) \quad &\leq \frac{2c_2}{\alpha + 1} \frac{T^{\alpha+1} e^{cT}}{k^{\alpha}},
 \end{aligned}$$

which goes to zero as  $k \rightarrow \infty$ .  $\square$

**5. Lower estimate of density function.** In practical nonlinear filtering computation, it is important to know how much density remains within a given ball. In this section, we shall provide such a lower estimate. In particular, the solution  $u$  of the DMZ equation in  $\mathbb{R}^n$  obtained by taking  $\lim_{R \rightarrow \infty} u_R$ , where  $u_R$  is the solution of the DMZ equation in the ball  $B_R$ , is a nontrivial solution.

*Proof of Theorem D.* Let  $\phi = e^{-\rho(x)} - e^{-\rho(R_0)}$ , where  $\rho$  is an increasing function of  $|x|$ . Observe that  $\phi \geq 0$  for  $x \in B_{R_0}$ ,  $\phi = 0$  on  $\partial B_{R_0}$ , and  $\frac{\partial \phi}{\partial \nu}|_{\partial B_{R_0}} \leq 0$ , where  $\nu$  is

the outward normal of  $\partial B_{R_0}$ .

$$\begin{aligned}
\frac{d}{dt} \int_{B_{R_0}} \phi u_R &= \frac{1}{2} \int_{B_{R_0}} \phi \Delta u_R + \int_{B_{R_0}} \phi (-f + \nabla K) \cdot \nabla u_R \\
&\quad + \int_{B_{R_0}} \left( -\operatorname{div} f - \frac{1}{2} |h|^2 + \frac{\Delta K}{2} - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) \phi u_R \\
&= \frac{1}{2} \int_{B_{R_0}} u_R \Delta \phi - \frac{1}{2} \int_{\partial B_{R_0}} \frac{\partial \phi}{\partial \nu} u_R + \frac{1}{2} \int_{\partial B_{R_0}} \phi \frac{\partial u_R}{\partial \nu} \\
&\quad - \int_{B_{R_0}} u_R \operatorname{div} [\phi (-f + \nabla K)] + \int_{\partial B_{R_0}} \phi u_R (-f + \nabla K) \cdot \nu \\
&\quad + \int_{B_{R_0}} \left( -\operatorname{div} f - \frac{1}{2} |h|^2 + \frac{\Delta K}{2} - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) \phi u_R \\
&\geq \frac{1}{2} \int_{B_{R_0}} u_R \Delta \phi + \int_{B_{R_0}} u_R \nabla \phi \cdot (f - \nabla K) \\
&\quad + \int_{B_{R_0}} \left( -\frac{\Delta K}{2} - \frac{1}{2} |h|^2 - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) \phi u_R.
\end{aligned}$$

Notice that  $\nabla \phi = -e^{-\rho(x)} \nabla \rho$  and  $\Delta \phi = e^{-\rho(x)} (|\nabla \rho|^2 - \Delta \rho)$ . Hence

$$\begin{aligned}
\frac{d}{dt} \int_{B_{R_0}} \phi u_R &\geq \int_{B_{R_0}} u_R e^{-\rho(x)} \left[ -\frac{\Delta \rho}{2} + \frac{|\nabla \rho|^2}{2} - \nabla \rho \cdot (f - \nabla K) \right] \\
&\quad + \int_{B_{R_0}} \left( -\frac{\Delta K}{2} - \frac{1}{2} |h|^2 - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right) \phi u_R.
\end{aligned}$$

Let  $r = |x|$ . We have

$$\nabla \rho = \frac{\rho'(\gamma)}{r} x \quad \text{and} \quad \Delta \rho = \rho''(r) + \rho'(r) \frac{n-1}{r}.$$

Hence

$$\begin{aligned}
\frac{d}{dt} \int_{B_{R_0}} \phi u_R &\geq \int_{B_{R_0}} u_R e^{-\rho(R_0)} \left[ -\frac{\Delta \rho}{2} + \frac{|\nabla \rho|^2}{2} - \nabla \rho \cdot (f - \nabla K) \right] \\
&\quad + \int_{B_{R_0}} \left[ -\frac{\Delta \rho}{2} + \frac{|\nabla \rho|^2}{2} - \nabla \rho \cdot (f - \nabla K) \right. \\
&\quad \quad \left. - \frac{\Delta K}{2} - \frac{1}{2} |h|^2 - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right] \phi u_R \\
&= e^{-\rho(R_0)} \int_{B_{R_0}} \left[ \frac{\rho'^2}{2} - \frac{\rho''}{2} - \frac{n-1}{2r} \rho' - \rho' (f - \nabla K) \cdot \frac{x}{r} \right] u_R \\
&\quad + \int_{B_{R_0}} \left[ -\frac{\Delta \rho}{2} + \frac{|\nabla \rho|^2}{2} - \nabla \rho \cdot (f - \nabla K) \right. \\
&\quad \quad \left. - \frac{\Delta K}{2} - \frac{1}{2} |h|^2 - f \cdot \nabla K + \frac{1}{2} |\nabla K|^2 \right] \phi u_R.
\end{aligned} \tag{5.1}$$

We want to choose  $\rho$  such that

$$e^{-\rho(R_0)} \left| \int_{B_{R_0}} \left[ \frac{\rho^{12}}{2} - \frac{\rho''}{2} - \frac{n-1}{2r} \rho' - \rho'(f - \nabla K) \cdot \frac{x}{r} \right] u_R \right| \leq \epsilon(R_0),$$

where  $\epsilon(R_0)$  is small and will be determined later. Then (5.1) implies

$$(5.2) \quad \begin{aligned} \frac{d}{dt} \int_{B_{R_0}} \phi u_R \geq & -\epsilon(R_0) + \int_{B_{R_0}} \left[ -\frac{\Delta \rho}{2} + \frac{|\nabla \rho|^2}{2} - \nabla \rho \cdot (f - \nabla K) \right. \\ & \left. - \frac{\Delta K}{2} - \frac{1}{2}|h|^2 - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right] \phi u_R. \end{aligned}$$

We now take  $\rho = |x|^m$ . Then

$$\begin{aligned} \Delta \rho &= \rho''(r) + \rho'(r) \frac{n-1}{r} = m(m+n-2)r^{m-2}, \\ |\nabla \rho|^2 &= (\rho'(r))^2 = m^2 r^{2m-2}. \end{aligned}$$

Since  $f$  and  $h$  are of polynomial growth, we can choose a positive integer  $m$  large enough such that

$$\begin{aligned} & -\frac{\Delta \rho}{2} + \frac{|\nabla \rho|^2}{2} - \nabla \rho \cdot (f - \nabla K) - \frac{\Delta K}{2} - \frac{1}{2}|h|^2 - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \\ &= -\frac{1}{2}m(m+n-2)r^{m-2} + \frac{1}{2}m^2 r^{2m-2} \\ &\quad - \frac{\rho'(r)}{r} x \cdot (f - \nabla K) - \frac{\Delta K}{2} - \frac{1}{2}|h|^2 - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \\ &\leq -c' \quad \text{on } \mathbb{R}^n, \end{aligned}$$

where  $c'$  is a positive constant independent of  $R$  and  $R_0$ . Hence

$$(5.3) \quad \begin{aligned} \frac{d}{dt} \int_{B_{R_0}} \phi u_R &\geq -\epsilon(R_0) - c' \int_{B_{R_0}} \phi u_R \\ &\Rightarrow \frac{d}{dt} \left[ e^{c't} \int_{B_{R_0}} \phi u_R \right] \geq -\epsilon(R_0) e^{c't} \\ &\Rightarrow e^{c'T} \int_{B_{R_0}} \phi u_R(T, x) - \int_{B_{R_0}} \phi u_R(0, x) \geq \frac{\epsilon(R_0)}{c'} (1 - e^{c'T}). \end{aligned}$$

We are now ready to estimate  $\epsilon(R_0)$ . Observe that  $\frac{\partial u_R}{\partial \nu} \leq 0$  on  $\partial B_R$ , where  $\nu$  is the outward normal of  $\partial B_R$ .

$$\begin{aligned} \frac{d}{dt} \int_{B_R} u_R &= \frac{1}{2} \int_{B_R} \Delta u_R + \int_{B_R} (-f + \nabla K) \cdot \nabla u_R \\ &\quad + \int_{B_R} \left( -\operatorname{div} f - \frac{1}{2}|h|^2 + \frac{1}{2}\Delta K - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right) u_R \\ &= \frac{1}{2} \int_{\partial B_R} \frac{\partial u_R}{\partial \nu} - \int_{B_R} u_R \operatorname{div}(-f + \nabla K) + \int_{\partial B_R} u_R (-f + \nabla K) \cdot \nu \end{aligned}$$

$$\begin{aligned}
 & + \int_{B_R} \left( -\operatorname{div} f - \frac{1}{2}|h|^2 + \frac{\Delta K}{2} - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right) u_R \\
 & \leq \int_{B_R} \left( -\frac{1}{2}|h|^2 - \frac{\Delta K}{2} - f \cdot \nabla K + \frac{1}{2}|\nabla K|^2 \right) u_R \leq c_1 \int_{B_R} u_R.
 \end{aligned}$$

Hence

$$(5.4) \quad \int_{B_R} u_R(t, x) \leq e^{c_1 t} \int_{B_R} u_R(0, x), \quad 0 \leq t \leq T.$$

In order to estimate  $\epsilon(R_0)$ , we need to determine the upper bound of

$$e^{-\rho(R_0)} \left| \int_{B_{R_0}} \left[ \frac{\rho'^2}{2} - \frac{\rho''}{2} - \frac{n-1}{2r} \rho' - \rho'(f - \nabla K) \cdot \frac{x}{r} \right] u_R \right|.$$

Recall that  $\rho = r^m$ . Then for  $m$  large enough,

$$\begin{aligned}
 & \left| \frac{\rho'^2}{2} - \frac{\rho''}{2} - \frac{n-1}{2r} \rho' - \rho'(f - \nabla K) \cdot \frac{x}{r} \right| \\
 & = \left| \frac{m^2 r^{2m-2}}{2} - \frac{m(m+n-2)}{2} r^{m-2} - m r^{m-2} (f - \nabla K) \cdot x \right| \\
 & \leq \frac{m(m+1)}{2} r^{2m-2} + c'',
 \end{aligned}$$

where  $c''$  is independent of  $R$  and  $R_0$ . Therefore

$$\begin{aligned}
 & e^{-\rho(R_0)} \left| \int_{B_{R_0}} \left[ \frac{\rho'^2}{2} - \frac{\rho''}{2} - \frac{n-1}{2r} \rho' - \rho'(f - \nabla K) \cdot \frac{x}{r} \right] u_R(t, x) \right| \\
 & \leq e^{-R_0^m} \left( \frac{m(m+1)}{2} R_0^{2m-2} + c'' \right) \int_{B_{R_0}} u_R(t, x) \\
 & \leq e^{c'T - R_0^m} \left( \frac{m(m+1)}{2} R_0^{2m-2} + c'' \right) \int_{B_R} u_R(0, x),
 \end{aligned}$$

by (5.4)

and we can set

$$\epsilon(R_0) = e^{c'T - R_0^m} \left( \frac{m(m+1)}{2} R_0^{2m-2} + c'' \right) \int_{B_R} u_R(0, x).$$

In view of (5.3), we have

$$\begin{aligned}
 & e^{c'T} \int_{B_{R_0}} \phi u_R(T, x) - \int_{B_{R_0}} \phi u_R(0, x) \\
 & \geq \frac{e^{c'T - R_0^m}}{c'} \left( \frac{m(m+1)}{2} R_0^{2m-2} + c'' \right) (1 - e^{c'T}) \int_{B_R} u_R(0, x),
 \end{aligned}$$

which implies

$$\begin{aligned}
 & \int_{B_{R_0}} \phi u_R(T, x) \geq e^{-c'T} \int_{B_{R_0}} \phi u_R(0, x) \\
 & + \frac{e^{-R_0^m}}{c'} \left( \frac{m(m+1)}{2} R_0^{2m-2} + c'' \right) (1 - e^{c'T}) \int_{B_R} u_R(0, x).
 \end{aligned} \tag{5.5}$$

Observe that the second term on the right-hand side of (5.5) tends to zero as  $R_0 \rightarrow \infty$ . Therefore we have

$$\int_{\mathbb{R}^n} e^{-|x|^m} u(T, x) \geq e^{-c'T} \int_{\mathbb{R}^n} e^{-|x|^m} u(0, x). \quad \square$$

**Appendix A. A priori estimation of derivatives up to second order.** In this section, we shall give a priori estimation of zero, first, and second derivatives of the solution of the robust DMZ equation on  $[0, T] \times B_R$ .

**THEOREM A.1.** *Consider the robust DMZ equation (1.9) on  $[0, T] \times B_R$ , where  $B_R = \{x \in \mathbb{R}^n : |x| \leq R\}$  is a ball of radius  $R$ . Let  $C_1 = \max_{0 \leq t \leq T} [\sum_{i=1}^m |y_i(t)|^2]^{\frac{1}{2}}$  be the smallest constant such that*

$$(A.1) \quad |\nabla K(t, x)| \leq C_1 |\nabla h(x)| \quad \text{for } (t, x) \in [0, T] \times B_R,$$

where  $|\nabla h|^2 = \sum_{i=1}^m |\nabla h_i(x)|^2$ .

Suppose that there exists a constant  $C > 0$  such that for any  $r \geq 0$

$$(A.2) \quad \min_{|x|=r} \frac{|h|^2 + \operatorname{div} f + C}{\sqrt{|f|^2 + |h|^2 + \operatorname{div} f + C + |f|}} - C_1 \max_{|x|=r} |\nabla h| \geq 0.$$

Let  $g(x)$  be a positive radial symmetric function on  $\mathbb{R}^n$  (i.e.,  $g = g(r)$ , where  $r = |x| = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$ ) such that

$$(A.3) \quad |g'(r)| \leq \min_{|x|=r} \frac{|h|^2 + \operatorname{div} f + C}{\sqrt{|f|^2 + |h|^2 + \operatorname{div} f + C + |f|}} - C_1 \max_{|x|=r} |\nabla h|.$$

Then, for  $0 \leq t \leq T$ ,

$$(A.4) \quad \int_{B_R} e^{2g} u_R^2(t, x) \leq e^{ct} \int_{B_R} e^{2g} \sigma^2(x).$$

*Proof.* Let  $\rho$  be any smooth function on  $\mathbb{R} \times \mathbb{R}^n$ . Then

$$\begin{aligned} \frac{d}{dt} \left( \frac{1}{2} \int_{B_R} \rho^2 u_R^2 \right) &= \int_{B_R} \rho \rho_t u_R^2 + \int_{B_R} \rho^2 u_R \frac{\partial u_R}{\partial t} \\ &= \int_{B_R} \rho \rho_t u_R^2 + \int_{B_R} \frac{1}{2} \rho^2 u_R \Delta u_R - \int_{B_R} \rho^2 u_R (f - \nabla K) \cdot \nabla u_R \\ &\quad - \int_{B_R} \rho^2 u_R^2 \left[ \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right] \\ &= \int_{B_R} \rho \rho_t u_R^2 - \int_{B_R} \rho u_R \nabla \rho \cdot \nabla u_R - \frac{1}{2} \int_{B_R} \rho^2 |\nabla u_R|^2 \\ &\quad + \int_{B_R} \rho u_R^2 \nabla \rho \cdot (f - \nabla K) + \frac{1}{2} \int_{B_R} \rho^2 (\operatorname{div} f - \nabla K) u_R^2 \\ &\quad - \int_{B_R} \rho^2 u_R^2 \left[ \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right] \\ &\leq \int_{B_R} \rho \rho_t u_R^2 + \frac{1}{2} \int_{B_R} |\nabla \rho|^2 u_R^2 - \int_{B_R} \rho^2 u_R^2 \left[ \frac{1}{2} |h|^2 + \frac{1}{2} \operatorname{div} f \right. \\ (A.5) \quad &\quad \left. + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 - \sum_{i=1}^n \nabla(\log \rho) \cdot (f - \nabla K) \right]. \end{aligned}$$

If we set  $\rho = e^g$ , then we get

$$\begin{aligned}
 & \frac{d}{dt} \left[ \frac{1}{2} \int_{B_R} \rho^2 u_R^2 \right] \\
 & \leq \int_{B_R} \left[ g_t + \frac{|\nabla g|^2}{2} - \frac{1}{2} |h|^2 - \frac{1}{2} \operatorname{div} f - f \cdot \nabla K \right. \\
 & \quad \left. + \frac{1}{2} |\nabla K|^2 + \nabla g \cdot (f - \nabla K) \right] \rho^2 u_R^2 \\
 & = \int_{B_R} \left[ g_t + \frac{|\nabla g|^2}{2} + \frac{1}{2} |f - \nabla K|^2 - \frac{1}{2} |f|^2 \right. \\
 & \quad \left. - \frac{1}{2} |h|^2 - \frac{1}{2} \operatorname{div} f + \nabla g \cdot (f - \nabla K) \right] g^2 u_R^2 \\
 (A.6) \quad & = \int_{B_R} \left[ g_t + \frac{1}{2} |\nabla g + f - \nabla K|^2 - \frac{1}{2} |f|^2 - \frac{1}{2} |h|^2 - \frac{1}{2} \operatorname{div} f \right] \rho^2 u_R^2.
 \end{aligned}$$

We shall choose  $g$  to be independent of  $t$  and a constant  $C > 0$  so that

$$(A.7) \quad |\nabla g + f - \nabla K| \leq \sqrt{|f|^2 + |h|^2 + \operatorname{div} f + C}.$$

Notice that (A.6) and (A.7) imply

$$(A.8) \quad \frac{d}{dt} \left[ \frac{1}{2} \int_{B_R} \rho^2 u_R^2 \right] \leq C \int_{B_R} \frac{1}{2} \rho^2 u_R^2.$$

Inequality (A.7) can be achieved if we have

$$\begin{aligned}
 (A.9) \quad |\nabla g| & \leq \sqrt{|f|^2 + |h|^2 + \operatorname{div} f + C} - |f| - |\nabla K| \\
 & = \frac{|h|^2 + \operatorname{div} f + C}{\sqrt{|f|^2 + |h|^2 + \operatorname{div} f + C} + |f|} - |\nabla K|.
 \end{aligned}$$

Notice that, for  $0 \leq t \leq T$ ,

$$\begin{aligned}
 |\nabla K(t, x)| & = \left| \sum_{i=1}^m y_i(t) \nabla h_i \right| \leq \sum_{i=1}^m |y_i(t)| |\nabla h_i(x)| \\
 & \leq \left( \sum_{i=1}^m |y_i(t)|^2 \right)^{\frac{1}{2}} \left( \sum_{i=1}^m |\nabla h_i(x)|^2 \right)^{\frac{1}{2}}.
 \end{aligned}$$

Let  $C_1 = \max_{0 \leq t \leq T} \left( \sum_{i=1}^m |y_i(t)|^2 \right)^{\frac{1}{2}}$ . Then we have

$$(A.10) \quad |\nabla K(t, x)| \leq C_1 |\nabla h(x)|.$$

Now we shall choose  $g$  to be radial symmetric so that

$$(A.11) \quad |g'(r)| = |\nabla g| \leq \min_{|x|=r} \frac{|h|^2 + \operatorname{div} f + C}{\sqrt{|f|^2 + |h|^2 + \operatorname{div} f + C} + |f|} - C_1 \max_{|x|=r} |\nabla h|.$$

Inequality (A.9) implies

$$(A.12) \quad \int_{B_R} e^{2g} u_R^2(t, x) \leq e^{ct} \int_{B_R} e^{2g} \sigma(x). \quad \square$$

*Remark A.2.* Notice that from (A.11), if  $h$  grows fast, then we can allow  $g$  to grow fast.

We next give a priori estimation of the first and second derivatives of the solution of the robust DMZ equation on  $[0, T] \times B_R$ . We first observe that for estimation of second derivatives it is sufficient to estimate the Laplacian of the solution.

**LEMMA A.3.** *Let  $\rho$  be a smooth function with compact support in  $B_R$ . Let  $u_R$  be the solution of (1.9). Then*

$$(A.13) \quad \int_{B_R} \sum_{i,j=1}^n \rho^2 (u_R)_{ji}^2 \leq 4 \int_{B_R} \rho^2 (\Delta u_R)^2 + 6 \sup |\nabla \rho|^2 \int_{B_R} |\nabla u_R|^2.$$

*Proof.*

$$\begin{aligned} \int_{B_R} \rho^2 (\Delta u_R)^2 &= - \int_{B_R} 2\rho (\nabla \rho \cdot \nabla u_R) \Delta u_R - \int_{B_R} \rho^2 \nabla (\Delta u_R) \cdot \nabla u_R \\ &= - \int_{B_R} 2\rho \Delta u_R (\nabla u_R \cdot \nabla \rho) - \int_{B_R} \rho^2 \sum_{i,j=1}^n (u_R)_{jji} (u_R)_i \\ &= - \int_{B_R} 2\rho \Delta u_R (\nabla u_R \cdot \nabla \rho) \\ (A.14) \quad &+ \sum_{i,j=1}^n \int_{B_R} 2\rho \rho_j (u_R)_{ji} u_i + \sum_{i,j=1}^n \int_{B_R} \rho^2 (u_R)_{ji}^2. \end{aligned}$$

By the Schwartz inequality, we have

$$\begin{aligned} \int_{B_R} 2\rho \Delta u_R (\nabla u_R \cdot \nabla \rho) &\leq \int_{B_R} 2\rho \Delta u_R |\nabla u_R| |\nabla \rho| \\ (A.15) \quad &\leq \int_{B_R} \rho^2 (\Delta u_R)^2 + \int_{B_R} |\nabla u_R|^2 |\nabla \rho|^2 \end{aligned}$$

$$(A.16) \quad \sum_{i,j=1}^n \int_{B_R} 2\rho \rho_j (u_R)_{ji} (u_R)_i \leq \sum_{i,j=1}^n \int_{B_R} \left[ \frac{1}{2} \rho^2 (u_R)_{ji}^2 + 2\rho_j^2 (u_R)_i^2 \right].$$

Inequalities (A.14), (A.15), and (A.16) imply

$$\begin{aligned} \int_{B_R} \rho^2 (\Delta u_R)^2 &\geq - \int_{B_R} \rho^2 (\Delta u_R)^2 - \int_{B_R} |\nabla u_R|^2 |\nabla \rho|^2 \\ &\quad - \frac{1}{2} \int_{B_R} \sum_{i,j=1}^n \rho^2 (u_R)_{ji}^2 - 2 \sum_{i,j=1}^n \int_{B_R} \rho_j^2 u_i^2 + \sum_{i,j=1}^n \int_{B_R} \rho^2 (u_{ji})^2 \\ &= - \int_{B_R} \rho^2 (\Delta u_R)^2 - \int_{B_R} |\nabla u_R|^2 |\nabla \rho|^2 - 2 \int_{B_R} |\nabla \rho|^2 |\nabla u_R|^2 \\ &\quad + \frac{1}{2} \int_{B_R} \sum_{i,j=1}^n \rho^2 (u_R)_{ji}^2, \end{aligned}$$



which is equivalent to

$$\frac{1}{2} \int_{B_R} \sum_{i,j=1}^n \rho^2(u_R)_{ji}^2 \leq 2 \int_{B_R} \rho^2(\Delta u_R)^2 + 3 \int_{B_R} |\nabla \rho|^2 |\nabla u_R|^2.$$

Hence

$$\int_{B_R} \sum_{i,j=1}^n \rho^2(u_R)_{ji}^2 \leq 4 \int_{B_R} \rho^2(\Delta u_R)^2 + 6 \sup |\nabla \rho|^2 \int_{B_R} |\nabla u_R|^2. \quad \square$$

Now we are ready to give a priori estimation of the first and second derivatives of the solution of the robust DMZ equation on  $[0, T] \times B_R$ .

**THEOREM A.4.** *Consider the robust DMZ equation (1.9) on  $[0, T] \times B_R$ , where  $B_R = \{x \in \mathbb{R}^n : |x| < R\}$  is a ball of radius  $R$ . Assume that*

$$(A.17) \quad \sqrt{\frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 + \frac{C}{2}} - |f| - |\nabla K| \geq 0,$$

where  $C$  is the constant in Theorem A.1. Choose a nonnegative function  $\tilde{g}$  so that

$$(A.18) \quad |\nabla \tilde{g}| \leq \sqrt{\frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 + \frac{C}{2}} - |f| - |\nabla K|$$

and

$$(A.19) \quad e^{2\tilde{g}} \left| \nabla \left( \frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 \right) \right|^2 \leq e^{2g},$$

where  $g$  is chosen as in Theorem A.1. Then

$$(A.20) \quad \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2(T, x) + \frac{1}{2} \int_0^T \int_{B_R} e^{2\tilde{g}} (\Delta u_R)^2(t, x) \\ \leq \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2(0, x) + T \int_{B_R} e^{2g} \sigma^2(x).$$

*Proof.* Recall that  $\frac{\partial u_R}{\partial t} \Big|_{\partial B_R} = 0$ . We have

$$\begin{aligned} & \frac{d}{dt} \left[ \frac{1}{2} \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2 \right] \\ &= \int_{B_R} e^{2\tilde{g}} \nabla u_R \cdot \nabla \frac{\partial u_R}{\partial t} \\ &= \int_{B_R} e^{2\tilde{g}} (-2 \nabla u_R \cdot \nabla g - \Delta u_R) \frac{\partial u_R}{\partial t} \\ &= \int_{B_R} e^{2\tilde{g}} (-2 \nabla u_R \cdot \nabla \tilde{g} - \Delta u_R) \left[ \frac{1}{2} \Delta u_R - (f - \nabla K) \cdot \nabla u_R \right. \\ & \quad \left. - \left( \frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 \right) u_R \right] \end{aligned}$$

$$\begin{aligned}
&= \int_{B_R} e^{2\tilde{g}} \left\{ -\frac{1}{2}(\Delta u_R)^2 + \left[ -\nabla u_R \cdot \nabla \tilde{g} + (f - \nabla K) \cdot \nabla u_R \right] \Delta u_R \right. \\
&\quad + 2(\nabla u_R \cdot \nabla \tilde{g})(f - \nabla K) \cdot \nabla u_R \\
&\quad + u_R \Delta u_R \left( -\frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 \right) \\
&\quad \left. + 2u_R \nabla u_R \cdot \nabla \tilde{g} \left( \frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 \right) \right\}.
\end{aligned}$$

This implies

$$\begin{aligned}
&\frac{d}{dt} \left[ \frac{1}{2} \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2 \right] + \frac{1}{4} \int_{B_R} e^{2\tilde{g}} (\Delta u_R)^2 \\
&= \int_{B_R} e^{2\tilde{g}} \left\{ -\frac{1}{4} \left[ (\Delta u_R)^2 - 4 \left( -\nabla u_R \cdot \nabla \tilde{g} + (f - \nabla K) \cdot \nabla u_R \right) \Delta u_R \right] \right. \\
&\quad + 2(\nabla u_R \cdot \nabla \tilde{g})(f - \nabla K) \cdot \nabla u_R \\
&\quad + u_R \Delta u_R \left( -\frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 \right) \\
&\quad \left. + 2u_R \nabla u_R \cdot \nabla \tilde{g} \left( \frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 \right) \right\} \\
&= \int_{B_R} e^{2\tilde{g}} \left\{ -\frac{1}{4} \left[ \Delta u_R + 2\nabla u_R \cdot \nabla \tilde{g} - 2(f - \nabla K) \cdot \nabla u_R \right]^2 + \frac{1}{4} [2\nabla u_R \cdot \nabla \tilde{g} \right. \\
&\quad - 2(f - \nabla K) \cdot \nabla u_R]^2 + 2(\nabla u_R \cdot \nabla \tilde{g})(f - \nabla K) \cdot \nabla u_R \\
&\quad + u_R \Delta u_R \left( -\frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 \right) \\
&\quad \left. + 2u_R (\nabla u_R \cdot \nabla \tilde{g}) \left( \frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 \right) \right\}.
\end{aligned} \tag{A.21}$$

Notice that

$$\begin{aligned}
&\frac{1}{4} \left[ 2\nabla u_R \cdot \nabla \tilde{g} - 2(f - \nabla K) \cdot \nabla u_R \right]^2 + 2(\nabla u_R \cdot \nabla \tilde{g}) \left[ (f - \nabla K) \cdot \nabla u_R \right] \\
&= \left[ \nabla u_R \cdot \nabla \tilde{g} - (f - \nabla K) \cdot \nabla u_R \right]^2 + 2(\nabla u_R \cdot \nabla \tilde{g}) \left[ (f - \nabla K) \cdot \nabla u_R \right] \\
&= \left[ \nabla u_R \cdot (\nabla \tilde{g} + f - \nabla K) \right]^2
\end{aligned} \tag{A.22}$$

and

$$\begin{aligned}
&\int_{B_R} e^{2\tilde{g}} u_R \Delta u_R \left( \frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 \right) \\
&= - \int_{B_R} e^{2\tilde{g}} \left[ 2u_R \nabla u_R \cdot \nabla \tilde{g} \left( \frac{1}{2}|h|^2 + \operatorname{div} f - \frac{1}{2}\Delta K + f \cdot \nabla K - \frac{1}{2}|\nabla K|^2 \right) \right]
\end{aligned}$$

$$\begin{aligned}
 & - \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2 \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right) \\
 (A.23) \quad & - \int_{B_R} e^{2\tilde{g}} u_R \nabla u_R \cdot \nabla \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right).
 \end{aligned}$$

Putting (A.22), (A.23) in (A.21), we get

$$\begin{aligned}
 & \frac{d}{dt} \left[ \frac{1}{2} \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2 \right] + \frac{1}{4} \int_{B_R} e^{2\tilde{g}} (\Delta u_R)^2 \\
 & \leq \int_{B_R} e^{2\tilde{g}} |u_R|^2 |\nabla \tilde{g} + f - \nabla K|^2 \\
 & \quad - \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2 \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right) \\
 (A.24) \quad & - \int_{B_R} e^{2\tilde{g}} u_R \nabla u_R \cdot \nabla \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right).
 \end{aligned}$$

As before, we look for  $\tilde{g}$  so that

$$|\nabla \tilde{g} + f - \nabla K| \leq \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 + \frac{C}{2} \right)^{\frac{1}{2}}.$$

Hence it suffices to set  $\tilde{g}$  so that

$$|\nabla \tilde{g}| \leq \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 + \frac{C}{2} \right)^{\frac{1}{2}} - |f| - |\nabla K|.$$

For such  $\tilde{g}$ , we have

$$\begin{aligned}
 & \frac{d}{dt} \left[ \frac{1}{2} \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2 \right] + \frac{1}{4} \int_{B_R} e^{2\tilde{g}} (\Delta u_R)^2 \\
 & \leq C \left[ \frac{1}{2} \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2 \right] + \frac{1}{2} \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2 \\
 (A.25) \quad & + \frac{1}{2} \int_{B_R} e^{2\tilde{g}} u_R^2 \left| \nabla \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right) \right|^2.
 \end{aligned}$$

We can choose  $\tilde{g}$  so that

$$(A.26) \quad e^{2\tilde{g}} \left| \nabla \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right) \right|^2 \leq e^{2g}.$$

Inequalities (A.25) and (A.26) imply

$$\begin{aligned}
 & \frac{d}{dt} \left[ \frac{1}{2} \int_{B_R} e^{2\tilde{g}} |\nabla u_R|^2 \right] + \frac{1}{4} \int_{B_R} e^{2\tilde{g}} (\Delta u_R)^2 \\
 & \leq (c+1) \int_{B_R} \frac{1}{2} e^{2\tilde{g}} |\nabla u_R|^2 + \frac{1}{2} e^{ct} \int_{B_R} e^{2g} \sigma^2(x).
 \end{aligned}$$

Hence

$$\begin{aligned}
& \frac{d}{dt} \left[ e^{-(c+1)t} \int_{B_R} \frac{1}{2} e^{2\tilde{g}} |\nabla u_R|^2 \right] \\
&= e^{-(c+1)t} \left[ \frac{d}{dt} \int_{B_R} \frac{1}{2} e^{2\tilde{g}} |\nabla u_R|^2 - (c+1) \int_{B_R} \frac{1}{2} e^{2\tilde{g}} |\nabla u_R|^2 \right] \\
&\leq \frac{1}{2} e^{-t} \int_{B_R} e^{2g} \sigma^2(x) - \frac{1}{4} e^{-(c+1)t} \int_{B_R} e^{2\tilde{g}} (\Delta u_R)^2,
\end{aligned}$$

which implies

$$\begin{aligned}
& e^{-(c+1)T} \int_{B_R} \frac{1}{2} e^{2\tilde{g}} |\nabla u_R|^2(T, x) - \int_{B_R} \frac{1}{2} e^{2\tilde{g}} |\nabla u_R|^2(0, x) \\
&\leq \frac{1}{2} T \int_{B_R} e^{2g} \sigma^2(x) - \frac{1}{4} \int_0^T e^{-(c+1)t} \int_{B_R} e^{2\tilde{g}} (\Delta u_R)^2(t, x).
\end{aligned}$$

It follows that

$$\begin{aligned}
& e^{-(c+1)T} \int_{B_R} \frac{1}{2} e^{2\tilde{g}} |\nabla u_R|^2(T, x) + \frac{1}{4} e^{-(c+1)T} \int_0^T \int_{B_R} e^{2\tilde{g}} (\Delta u_R)^2(t, x) \\
&\leq \int_{B_R} \frac{1}{2} e^{2\tilde{g}} |\nabla u_R|^2(0, x) + \frac{1}{2} T \int_{B_R} e^{2g} \sigma^2(x),
\end{aligned}$$

and (A.20) follows immediately.  $\square$

**Appendix B. Existence of a weak solution for the DMZ equation.** Let  $Q = (0, T) \times \mathbb{R}^n$  and let  $L^2(Q)$  be the space of functions that are square integrable over  $Q$ . The scalar product of two elements  $v_1, v_2$  of  $L^2(Q)$  is defined by the equation

$$(v_1, v_2) = \iint_Q v_1 v_2 \, dx \, dt.$$

The class of  $C^\infty$  functions in  $\overline{Q}$  with compact supports in  $Q$  will be denoted by  $C_0^\infty(Q)$ .

**DEFINITION B.1.** A locally  $L^2$ -integrable function is called a generalized derivative of a locally  $L^2$ -integrable function  $v(t, x)$  in  $Q$  with respect to  $x$  if for each  $\Phi(t, x) \in C_0^\infty(Q)$  the equation

$$(B.1) \quad \iint_Q \left( v \frac{\partial \Phi}{\partial x_k} + w \Phi \right) dx \, dt = 0$$

holds. In this case, we write  $w = \frac{\partial v}{\partial x_k}$ . The generalized derivative with respect to  $t$  and generalized derivatives of higher order are defined similarly (see [So]).

**Remark B.2.** If the sequence of functions  $v_m(t, x)$  weakly tends to  $v(t, x)$  in the space  $L^2(Q)$  as  $m \rightarrow \infty$  and the norms of  $\frac{\partial v_m}{\partial x_k}$  in  $L^2(Q)$  are uniformly bounded with respect to  $m$ , then  $v(t, x)$  has a generalized derivative  $\frac{\partial v}{\partial x_k} \in L^2(Q)$  and  $\frac{\partial v_m}{\partial x_k}$  weakly tends to  $\frac{\partial v}{\partial x_k}$  [So].

DEFINITION B.3. We denote by  $W^1(\mathbb{R}^n)$  the space of functions  $\phi(x)$  such that  $\phi(x) \in L^2(\mathbb{R}^n)$  and  $\frac{\partial \phi}{\partial x_i} \in L^2(\mathbb{R}^n)$  for  $i = 1, \dots, n$ , with the scalar product

$$(B.2) \quad (\phi_1, \phi_2)_1 := \int_{\mathbb{R}^n} \phi_1(x) \phi_2(x) dx + \int_{\mathbb{R}^n} \sum_{i=1}^n \frac{\partial \phi_1}{\partial x_i} \frac{\partial \phi_2}{\partial x_i} dx.$$

We shall denote by  $W^{1,1}(Q)$  the space of functions  $v(t, x)$  for which  $v(t, x) \in L^2(Q)$ ,  $\frac{\partial v(t, x)}{\partial x_i} \in L^2(Q)$  ( $i = 1, \dots, n$ ), and  $\frac{\partial v(t, x)}{\partial t} \in L^2(Q)$ , with the scalar product

$$(B.3) \quad (v_1, v_2)_{1,1} := \iint_Q v_1(t, x) v_2(t, x) dt dx + \iint_Q \left( \sum_{i=1}^n \frac{\partial v_1}{\partial x_i} \frac{\partial v_2}{\partial x_i} + \frac{\partial v_1}{\partial t} \frac{\partial v_2}{\partial t} \right) dx dt.$$

It is known [So] that  $W^1(\mathbb{R}^n)$  and  $W^{1,1}(\mathbb{R}^n)$  are complete. The norms in  $L^2(Q)$ ,  $W^1(\mathbb{R}^n)$ , and  $W^{1,1}(Q)$  will be written  $\|v\|_0$ ,  $\|v\|_1$ , and  $\|v\|_{1,1}$ , respectively.

Remark B.4. It follows from the embedding theorems of Sobolev that a function of  $W^{1,1}(Q)$  can be modified on a set of measure zero in such a way that it is  $L^2$ -integrable on the section of the cylinder  $Q$  by any  $n$ -dimensional plane or  $n$ -dimensional  $C^1$  surface. In particular, such a function is  $L^2$ -integrable on the section of  $Q$  by any plane  $t = \text{constant}$ . Moreover, the values of  $v(t, x) \in W^{1,1}(Q)$  on sufficiently close  $n$ -dimensional planes will differ in mean by as little as we please [So]. In particular, if  $v(t, x) \in W^{1,1}(Q)$  and  $v(x, 0) = \phi(x)$ , then  $\int_Q [v(t, x) - \phi(x)]^2 dx \rightarrow 0$  as  $t \rightarrow 0$ .

DEFINITION B.5. The subspace of  $W^1(\mathbb{R}^n)$  consisting of functions that have compact supports in  $\mathbb{R}^n$  is written  $W_0^1(\mathbb{R}^n)$ , and the subspace of  $W^{1,1}(Q)$  consisting of functions  $v(t, x)$  which have compact supports in  $\mathbb{R}^n$  for any  $t$  is written  $W_0^{1,1}(Q)$ .

DEFINITION B.6. The function  $u(t, x)$  in  $W_0^{1,1}(Q)$  is called a weak solution of the initial value problem

$$(B.4) \quad \begin{cases} \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left( A_{ij}(t, x) \frac{\partial u}{\partial x_j} \right) + \sum_{i=1}^n B_i(t, x) \frac{\partial u}{\partial x_i} + C(t, x) u = \frac{\partial u}{\partial t}, \\ u(0, x) = \phi(x) \end{cases}$$

if for any function  $\Phi(t, x) \in W_0^{1,1}(Q)$  the following relation is valid:

$$(B.5) \quad \iint_Q \left[ \sum_{i,j=1}^n A_{ij} \frac{\partial u}{\partial x_j} \frac{\partial \Phi}{\partial x_i} - \left( \sum_{i=1}^n B_i \frac{\partial u}{\partial x_i} + C u - \frac{\partial u}{\partial t} \right) \Phi \right] dx dt = 0$$

and  $u(0, x) = \phi(x)$ .

We now recall some facts concerning convergence in Hilbert spaces.

Remark B.7. A sequence  $\{u_m\}$ , in a Hilbert space  $H$  with scalar product  $(\cdot, \cdot)$ , is said to be *weakly convergent* (to  $u$ ) if the sequence  $\{(u_m, f)\}$  is convergent (to  $(u, f)$ ) for any  $f \in H$ . A weakly convergent sequence is bounded. From any bounded sequence  $\{u_m\}$  in  $H$  one can extract a weakly convergent subsequence. If  $\{u_m\}$  is weakly convergent to  $u$ , then there exists a subsequence  $\{u_{m'}\}$  whose arithmetic means converge to  $u$  in the  $H$  norm (see [Fr, p. 273]).

THEOREM B.8. Under the hypothesis of Theorem A.4 the robust DMZ equation (1.6) on  $[0, T] \times \mathbb{R}^n$  with initial condition  $\sigma_0(x) \in W_0^1(\mathbb{R}^n)$  has a weak solution.

Proof. Let  $\{R_k\}$  be a sequence of positive number such that  $\lim_{k \rightarrow \infty} R_k = \infty$ . Let  $u_k(x)$  be the solution of the robust DMZ equation (1.9) on  $[0, T] \times B_{R_k}$ , where

$B_{R_k} = \{x \in \mathbb{R}^n : |x| \leq R_k\}$  is a ball of radius  $R_k$ . Let

$$u_k(t, x) \begin{cases} u_{R_k}(t, x) & \text{if } x \in B_{R_k}, \\ 0 & \text{if } x \notin B_{R_k}, \end{cases} \quad \sigma_k(x) \begin{cases} \sigma_0(x) & \text{if } x \in B_{R_k}, \\ 0 & \text{if } x \notin B_{R_k}. \end{cases}$$

In view of Theorems A.1 and A.4, the sequence  $\{u_k\}$  is a bounded set in  $W_0^{1,1}(Q)$ . By Remark B.7, there exists a subsequence  $\{u_{k'}\}$  which is weakly convergent to  $u$ . Moreover,  $u(t, x)$  has generalized derivative  $\frac{\partial u}{\partial x_i}, \frac{\partial^2 u}{\partial x_i^2} \in L^2(Q)$ , and  $\frac{\partial u_{k'}}{\partial x_i}, \frac{\partial^2 u_{k'}}{\partial x_i^2}$  weakly tend to  $\frac{\partial u}{\partial x_i}, \frac{\partial^2 u}{\partial x_i^2}$ , respectively. Now we claim that the weak derivative  $\frac{\partial u}{\partial t}$  exists and is equal to the right-hand side of (B.3). To see this, let  $\Phi(t, x) \in W_0^{1,1}(Q)$ . Then

$$\begin{aligned} & \iint \left[ \frac{1}{2} \Delta u - (f(x) - \nabla K) \cdot \nabla u - \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K \right. \right. \\ & \quad \left. \left. + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right) u \right] \Phi(t, x) dx dt \\ &= \lim_{k' \rightarrow \infty} \iint \left[ \frac{1}{2} \Delta u_{k'} - (f(x) - \nabla K) \cdot \nabla u_{k'} \right. \\ & \quad \left. - \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K \right. \right. \\ & \quad \left. \left. + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right) u_{k'} \right] \Phi(t, x) dx dt \\ &= \lim_{k' \rightarrow \infty} \iint \frac{\partial u_{k'}}{\partial t} \Phi(t, x) dx dt \\ &= - \lim_{k' \rightarrow \infty} \iint u_{k'} \frac{\partial \Phi}{\partial t}(t, x) dx dt \\ &= - \iint u \frac{\partial \Phi}{\partial t}(t, x) dx dt. \end{aligned}$$

Clearly  $u(0, x) = \lim_{k' \rightarrow \infty} u_{k'}(0, x) = \lim_{k' \rightarrow \infty} \sigma_{k'}(x) = \sigma_0(x)$ .  $\square$

**Appendix C. Uniqueness of a weak solution for the DMZ equation.** We are now ready to establish the uniqueness of a weak solution for the DMZ equation. We shall follow the notation in previous sections.

**THEOREM C.1.** *Let  $Q = (0, T) \times \mathbb{R}^n$ . Assume that for some  $c > 0$ ,*

$$(C.1) \quad \sup_{0 \leq t \leq T} \int_{\mathbb{R}^n} e^{cr} u^2(t, x) dx < \infty,$$

$$(C.2) \quad \int_0^T \int_{\mathbb{R}^n} e^{cr} |\nabla u(t, x)|^2 dx dt < \infty,$$

where  $r = \sqrt{x_1^2 + \cdots + x_n^2}$ . Suppose that there exists a finite number  $\alpha$  such that

$$(C.3) \quad \left| \frac{c}{2} \nabla r + f - \nabla K \right|^2 - 2 \left( \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right) \leq \alpha.$$

Then the weak solution  $u(t, x)$  of the robust DMZ equation on  $Q$  is unique.

*Proof.* We only need to prove that  $u(t, x) = 0$  on  $Q$  if  $u(0, x) = 0$ . By iteration, we may assume that  $\alpha T < 1$ . Let  $\Phi \in C_0^\infty(\mathbb{R}^n)$ . According to the definition of weak solution (B.5), we have

$$(C.4) \quad \begin{aligned} \iint_Q \frac{\partial u}{\partial t} \Phi \, dt \, dx &= -\frac{1}{2} \int_0^T \int_{\mathbb{R}^n} \nabla u \cdot \nabla \Phi \, dx \, dt - \int_0^T \int_{\mathbb{R}^n} (f - \nabla K) \cdot \nabla u \Phi \, dx \, dt \\ &\quad - \int_0^T \int_{\mathbb{R}^n} \left[ \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right] u \Phi \, dx \, dt. \end{aligned}$$

Replacing  $\Phi$  by  $\Phi e^{cr}$ , we have

$$(C.5) \quad \begin{aligned} \int_{\mathbb{R}^n} u(T, x) \Phi e^{cr} &= -\frac{1}{2} \int_0^T \int_{\mathbb{R}^n} \nabla u \cdot (e^{cr} \nabla \Phi) - \frac{c}{2} \int_0^T \int_{\mathbb{R}^n} \Phi e^{cr} \nabla r \cdot \nabla u \\ &\quad + \int_0^T \int_{\mathbb{R}^n} e^{cr} \Phi (-f + \nabla K) \cdot \nabla u \\ &\quad - \int_0^T \int_{\mathbb{R}^n} \left[ \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right] \Phi u e^{cr} \\ &\quad + \int_0^T \int_{\mathbb{R}^n} u \frac{\partial \Phi}{\partial t} e^{cr}. \end{aligned}$$

Approximating  $u$  by  $\Phi$  in the  $W^{1,1}(Q)$  norm, we get

$$\begin{aligned} \int_{\mathbb{R}^n} u^2(T, x) e^{cr} &= -\frac{1}{2} \int_0^T \int_{\mathbb{R}^n} e^{cr} |\nabla u|^2 - \frac{c}{2} \int_0^T \int_{\mathbb{R}^n} u e^{cr} \nabla r \cdot \nabla u \\ &\quad + \int_0^T \int_{\mathbb{R}^n} e^{cr} u (-f + \nabla K) \cdot \nabla u \\ &\quad - \int_0^T \int_{\mathbb{R}^n} \left[ \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right] u^2 e^{cr} \\ &\quad + \int_0^T \int_{\mathbb{R}^n} u \frac{\partial u}{\partial t} e^{cr} \\ &= -\frac{1}{2} \int_0^T \int_{\mathbb{R}^n} e^{cr} \left\{ |\nabla u|^2 + [cu \nabla r - 2u(-f + \nabla K)] \cdot \nabla u \right\} \\ &\quad - \int_0^T \int_{\mathbb{R}^n} \left[ \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right] u^2 e^{cr} \\ &\quad + \int_0^T \int_{\mathbb{R}^n} \frac{1}{2} e^{cr} u \Delta u - \int_0^T \int_{\mathbb{R}^n} e^{cr} u (f - \nabla K) \cdot \nabla u \\ &\quad - \int_0^T \int_{\mathbb{R}^n} e^{cr} u^2 \left[ \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right] \\ &= -\int_0^T \int_{\mathbb{R}^n} e^{cr} \left\{ |\nabla u|^2 + [cu \nabla r - 2u(-f + \nabla K)] \cdot \nabla u \right\} \end{aligned}$$

$$\begin{aligned}
& -2 \int_0^T \int_{\mathbb{R}^n} \left[ \frac{1}{2} |h|^2 + \operatorname{div} f - \frac{1}{2} \Delta K + f \cdot \nabla K - \frac{1}{2} |\nabla K|^2 \right] u^2 e^{cr} \\
& = - \int_0^T \int_{\mathbb{R}^n} e^{cr} \left| \nabla u - \frac{cu}{2} \nabla r - uf + u \nabla K \right|^2 \\
& \quad + \int_0^T \int_{\mathbb{R}^n} e^{cr} u^2 \left\{ \left| \frac{c}{2} \nabla r + f - \nabla K \right|^2 \right. \\
& \quad \quad \left. - \left( |h|^2 + 2 \operatorname{div} f - \Delta K + 2f \cdot \nabla K - |\nabla K|^2 \right) \right\} \\
(C.6) \quad & \leq \alpha \int_0^T \int_{\mathbb{R}^n} e^{cr} u^2(t, x).
\end{aligned}$$

By the mean value theorem, there exists  $T_1 \in (0, T)$  such that

$$\int_0^T \int_{\mathbb{R}^n} e^{cr} u^2(t, x) = T \int_{\mathbb{R}^n} e^{cr} u^2(T_1, x).$$

In view of (C.5), we have

$$(C.7) \quad \int_{\mathbb{R}^n} u^2(T, x) e^{cr} \leq \alpha T \int_{\mathbb{R}^n} u^2(T_1, x) e^{cr}.$$

By applying (C.5) successfully, there exists  $T_m \in (0, T)$  such that

$$(C.8) \quad \int_{\mathbb{R}^n} u^2(T, x) e^{cr} \leq (\alpha T)^m \int_{\mathbb{R}^n} u^2(T_m, x) e^{cr}.$$

As  $\alpha T < 1$ , we conclude that  $u = 0$ .  $\square$

**Acknowledgments.** We gratefully acknowledge the long term support from the Army Research Office, which allowed us to answer the challenge proposed by the Naval Research Office ten years ago: Given adequate computational resources, how can one solve the nonlinear filtering problem?

The authors also gratefully acknowledge the referees for many useful suggestions to improve the presentation of this paper. They would particularly like to thank the second referee for providing many useful references.

## REFERENCES

- [B-B-H] J. S. BARAS, G. L. BLANKENSHIP, AND W. E. HOPKINS, *Existence, uniqueness, and asymptotic behavior of solutions to a class of Zakai equations with unbounded coefficients*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 203–214.
- [Be] A. BENSOUSSAN, *Some existence results for stochastic partial differential equations*, in Stochastic Partial Differential Equations and Applications (Trento 1990), Pitman Res. Notes Math. 268, Longman Scientific and Technical, Harlow, UK, 1992, pp. 37–53.
- [B-G-R1] A. BENSOUSSAN, R. GLOWINSKI, AND A. RASCANU, *Approximation of the Zakai equation by the splitting up method*, SIAM J. Control Optim., 28 (1990), pp. 1420–1431.
- [B-G-R2] A. BENSOUSSAN, R. GLOWINSKI, AND A. RASCANU, *Approximation of some stochastic differential equations by the splitting up method*, Appl. Math. Optim., 25 (1992), pp. 81–106.
- [C-M-P] M. CHALEYAT-MAUREL, D. MICHEL, AND E. PARDOUX, *Un théorème d'unicité pour l'équation de Zakai*, Stochastic, 29 (1990), pp. 1–13.
- [Cr-Ly] D. CRISAN AND T. J. LYONS, *A particle approximation of the Kushner-Stratonovich equation*, Probab. Theory Related Fields, 115 (1999), pp. 549–578.



- [Da] M. H. A. DAVIS, *On a multiplicative functional transformation arising in nonlinear filtering theory*, Z. Wahrsch. Verw. Gebiete, 54 (1980), pp. 125–139.
- [D-J-P] P. DEL MORAL, J. JACOD, AND P. PROTTER, *The Monte-Carlo method for filtering with discrete-time observations*, Probab. Theory Related Fields, 120 (2001), pp. 346–368.
- [Du] T. E. DUNCAN, *Probability Density for Diffusion Processes with Applications to Nonlinear Filtering Theory*, Ph.D. thesis, Stanford University, Stanford, CA, 1967.
- [Fe] H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1969.
- [Fl-Mi] W. H. FLEMING AND S. K. MITTER, *Optimal control and nonlinear filtering for nondegenerate diffusion processes*, Stochastics, 8 (1982), pp. 63–77.
- [Fl-Le] P. FLORCHINGER AND F. LE GLAND, *Time discretization of the Zakai equation for diffusion processes observed in correlated noise*, Stoch. Stoch. Rep., 35 (1991), pp. 233–256.
- [Fr] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
- [Gy-Kr] I. GYONGY AND N. KRYLOV, *On the splitting-up method and stochastic partial differential equation*, Ann. Probab., 31 (2003), pp. 564–591.
- [Ho-Wo] W. E. HOPKINS, JR., AND W. S. WONG, *Lie-Trotter product formulas for nonlinear filtering*, Stochastics, 17 (1986), pp. 313–337.
- [I-K-O] A. M. ILIN, A. S. KALASHNIKOV, AND O. A. OLEINIK, *Linear equations of the second order of parabolic type*, Uspehi Math. Nauk SSSR, 17 (1962), pp. 3–146.
- [It] K. ITO, *Approximation of the Zakai equation for nonlinear filtering*, SIAM J. Control Optim., 34 (1996), pp. 620–634.
- [It-Ro] K. ITO AND B. ROZOVSKII, *Approximation of the Kushner equation for nonlinear filtering*, SIAM J. Control Optim., 38 (2000), pp. 893–915.
- [Ka-Bu] R. E. KALMAN AND R. S. BUCY, *New results in linear filtering and prediction theory*, Trans. ASME, 83 (1961), pp. 95–108.
- [Ku] H. J. KUSHNER, *A robust discrete-state approximation to the optimal nonlinear filter for a diffusion*, Stochastics, 3 (1979), pp. 75–83.
- [L-M-R] S. LOTOTSKY, R. MIKULEVICIUS, AND B. L. ROZOVSKII, *Nonlinear filtering revisited: A spectral approach*, SIAM J. Control Optim., 35 (1997), pp. 435–461.
- [Ma] V. MAGNANI, *The Co-area Formula for Real-Valued Lipschitz Maps on Stratified Groups*, GVGMT preprint server, <http://cvgmt.sns.it/papers/mago3/>.
- [Mi] D. MICHEL, *Régularité des lois conditionnelles en théorie du filtrage non-linéaire et calcul des variations stochastiques*, J. Funct. Ann., 41 (1981), pp. 8–36.
- [Mo] R. E. MORTENSEN, *Optional Control of Continuous Time Stochastic Systems*, Ph.D. thesis, University of California, Berkeley, CA, 1966.
- [Na] N. NAGASE, *Remarks on nonlinear stochastic partial differential equations: An application of the splitting-up method*, SIAM J. Control Optim., 33 (1995), pp. 1716–1730.
- [Pa-Ph] G. PAGES AND N. PHAM, *Optimal quantization methods for nonlinear filtering with discrete-time observations*, Bernoulli, 11 (2005), pp. 893–932.
- [Pa1] E. PARDOUX, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics, 3 (1979), pp. 127–128.
- [Pa2] E. PARDOUX, *Filtrage nonlinéaire et équations aux dérivées partielles stochastiques associées*, in École d'Été de Probabilité de St. Flour XIX, Lecture Notes in Math. 1464, Springer-Verlag, Berlin, 1991, pp. 67–163.
- [Pa3] E. PARDOUX, *Équation du filtrage nonlinéaire, de la prédiction et du lissage*, Stochastics, 6 (1982), pp. 193–231.
- [Ro] B. L. ROZOVSKII, *Stochastic Evolution Systems: Linear Theory and Applications to Nonlinear Filtering*, Math. Appl. 35, Kluwer Academic, Dordrecht, The Netherlands, 1990.
- [So] S. L. SOBOLEV, *Applications of Functional Analysis in Mathematical Physics*, Trans. Math. Monographs 7, AMS, Providence, RI, 1963.
- [Su] H. J. SUSSMAN, *Rigorous results on the cubic sensor problem*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., NATO Adv. Study Inst., D. Reidel, Dordrecht, The Netherlands, 1981, pp. 637–648.
- [Ya-Ya] S. T. YAU AND S. S.-T. YAU, *Real time solution of nonlinear filtering problem without memory I*, Math. Res. Lett., 7 (2000), pp. 671–693.
- [Za] M. ZAKAI, *On the optimal filtering of diffusion processes*, Z. Wahrsch. Verw. Gebiete, 11 (1969), pp. 230–243.

# ROBUST SEMIGLOBAL STABILIZATION OF THE SECOND ORDER SYSTEM BY RELAY FEEDBACK WITH AN UNCERTAIN VARIABLE TIME DELAY\*

EUGENII SHUSTIN<sup>†</sup>, LEONID FRIDMAN<sup>‡</sup>, EMILIA FRIDMAN<sup>§</sup>, AND  
FERNANDO CASTAÑOS<sup>¶</sup>

**Abstract.** We present sufficient conditions for robust relay-delayed *semiglobal* stabilization of second order systems, which relate the upper bound to an uncertain time delay and the parameters of the plant. We also suggest an algorithm of delayed relay control gain adaptation for semiglobal stabilization, which is based on *delayed information about the sign of the controlled variable only*. The proposed algorithm suppresses *bounded uncertainties in the time delay*; that is, being designed for the upper bound of uncertainty in the time delay, the control law ensures semiglobal stabilization independently of any variable time delay obeying the given upper bound.

**Key words.** variable structure systems, time delay, robust control

**AMS subject classifications.** 34K20, 34K35, 93B12, 93B51, 93D09

**DOI.** 10.1137/060673333

## 1. Introduction.

**1.1. Statement of the problem.** We study the control problem for the second order system

$$(1) \quad \alpha \ddot{x}(t) = -\beta \dot{x}(t) + F(x(t), t) + u,$$

with positive constants  $\alpha$  and  $\beta$  and some function  $F(x, t)$ , satisfying

$$(2) \quad F \in C^1(R^2), \quad \sup \left| \frac{\partial F}{\partial x} \right| < \infty.$$

The uncontrolled system

$$\alpha \ddot{x} = -\beta \dot{x} + F(x, t)$$

may be unstable, as, for example, in the case  $F(x, t) = kx$ ,  $k > 0$ , and we propose to stabilize it by a negative feedback of relay type:

$$(3) \quad u = -K(t) \cdot \text{sign } x(t - \tau),$$

---

\*Received by the editors October 25, 2006; accepted for publication (in revised form) June 26, 2007; published electronically January 11, 2008.

<http://www.siam.org/journals/sicon/47-1/67333.html>

<sup>†</sup>School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, 69978 Tel Aviv, Israel (shustin@post.tau.ac.il).

<sup>‡</sup>Department of Control, Division of Electrical Engineering, Engineering Faculty, National Autonomous University of Mexico, UNAM, 04510, Mexico, D.F., Mexico (lfridman@servidor.unam.mx, <http://verona.fi-p.unam.mx/~lfridman/>).

<sup>§</sup>Department of Electrical Engineering and Systems, Tel Aviv University, Ramat Aviv, 69978 Tel Aviv, Israel (emilia@eng.tau.ac.il).

<sup>¶</sup>Laboratoire des Signaux et Systèmes, Supélec, 3 rue Joliot Curie, 91192 Gif-sur-Yvette, France (castanos@lss.supelec.fr).

with a controllable bounded magnitude  $K(t) > 0$  and a positive variable uncertain delay  $\tau$ , assumed to be a measurable function of  $t$  obeying the condition

$$(4) \quad 0 < \tau_0(t) \leq \tau(t) \leq h = \text{const}, \quad t \geq 0,$$

where  $\tau_0(t)$  is a positive nonincreasing function.

Our aim is to design a piecewise constant controller  $K(t)$ , which provides a robust semiglobal stabilization of the oscillation magnitude of the solutions to system (1), (3).

**1.2. Motivation.** For the motivation, we point out that time delay in control systems is usually present and must be taken into account. In practice, many systems with time delay naturally admit relay controllers, in particular,

- *systems which can work in switching modes*, for example, power converters (see, for example, [19]);
- *systems with measuring devices that work in the switching mode and have time delay*, for example, controllers of exhausted gas in the fuel injector automotive control systems [15], which act with delay and, moreover, generate relay signals only;
- *sliding mode systems with delayed actuators*, for example, the stabilizers of the fingers for an underwater manipulator [2];
- *mathematical biology systems* as, for example, those considered in [12, 13].

It has been shown in [5, 6] that, in the simplest one-dimensional relay control systems with a constant delay, only oscillatory solutions can occur. Moreover, any such solution becomes periodic after a finite time interval, but only slowly oscillating solutions are stable. The latter property is used to design an algorithm controlling the motion amplitudes.

P.I. (proportional-integral) control algorithms for the amplitude control in one-dimensional relay systems with delay in the input have been suggested in [1]. A Padé approximation of delay that reduces the relay delay output tracking problem to the sliding mode control for a nonminimum phase system was suggested in [16]. Delayed relay control algorithms, suggested in [7, 8], allow one to reach local and nonlocal stabilization of oscillations amplitudes for MIMO systems, respectively, with the use of the delayed value of the magnitude of a current trajectory.

In [11], periodic properties of second order systems via relay-delayed controllers based on the suboptimal control algorithm were investigated, whereas the article [3] studies oscillations in first order systems, containing external forcing in the relay-delayed control element.

**1.3. The main result. Restrictions to the nonlinear element.** Throughout the paper we impose the following bound of the nonlinear term  $F(x, t)$  of (1):

$$(5) \quad 0 \leq \frac{F(x, t) - F(0, t)}{x} \leq k_0, \quad x \neq 0, \quad t \geq 0,$$

with some positive constant  $k_0$ . Furthermore, we separate between the two situations:

$$(6) \quad F(0, t) \equiv 0,$$

and

$$(7) \quad |F(0, t)| \leq \delta, \quad t \geq 0, \quad \delta = \text{const} \in (0, 1),$$

in which the suggested controller and the respective solutions to (1) and (3) behave differently.

**The initial value problem and the definition of the discontinuous element.** For system (1), (3), we state the initial value problem

$$(8) \quad x(t) = \varphi(t), \quad t \in [-h, 0], \quad \varphi \in C_0[-h, 0], \quad \dot{x}(0) = \dot{\varphi}(0),$$

by defining the initial data range to be the space  $C_0[-h, 0]$  of continuous functions  $\varphi : [-h, 0] \rightarrow \mathbb{R}$ , differentiable at the origin. We equip  $C_0[-h, 0]$  with the norm

$$(9) \quad \|\varphi\| = \max_{[-h, 0]} |\varphi(t)| + |\dot{\varphi}(0)|.$$

The fact that the functions  $\varphi \in C_0[-h, 0]$  may vanish along intervals raises an important issue of an appropriate choice of the values of the sign function at vanishing arguments. From the control theory point of view, sign should be a binary sensor or actuator; i.e., it takes the only values  $\pm 1$ . So we will define

$$(10) \quad \text{sign } x(t) = \begin{cases} 1 & \text{if } x(t) > 0, \\ -1 & \text{if } x(t) < 0, \\ \zeta(t) & \text{if } x(t) = 0, \end{cases}$$

where  $\zeta(t)$  is any measurable function with  $|\zeta(t)| = 1$ , and consider the solutions to system (1), (3) in the sense of Carathéodory (see, for example, [9]).<sup>1</sup>

Then we have the following.

LEMMA 1.1. *The equation*

$$\alpha \ddot{x}(t) = -\beta \dot{x}(t) + F(x(t), t) - \text{sign } x(t - \tau),$$

*satisfying (2), (4), and (7), with initial condition (8) supplied with (10), has a unique continuous solution  $x_\varphi(t)$ ,  $t \in [-h, \infty)$ . Moreover,  $x_\varphi$  is differentiable in the interval  $(0, \infty)$ , and its derivative is absolutely continuous and differentiable almost everywhere.*

We omit the proof, which basically coincides with the proof of Lemma 1.1 in [17], and notice only that the lower bound to  $\tau$  in (4) is needed for an accurate justification of the existence and uniqueness of the solution  $x_\varphi$ .

*Remark 1.* (i) The solutions  $x(t)$  to system (1), (3) considered in what follows will satisfy the condition  $|F(x(t), t)| < K(t)$  (cf. the combination of bounds (5), (7), and (20) and of Lemmas A.1, A.2, and A.3 below), and thus, in the same way as in Lemma 1.1, the zero locus of such a solution  $x(t)$  in the interval  $t \geq 0$  will have zero measure whatever the zero locus of the initial function  $\varphi(t) \in C_0[-h, 0]$  is, and hence the results do not depend on the choice of the function  $\zeta(t)$  in (10) for  $t \geq 0$ . In particular, shifting the initial interval to  $[0, h]$ , one obtains the zero locus of zero measure for the (new) initial function, getting rid of any dependence of the function  $\zeta(t)$ . In addition,  $\dot{x}(t)$  turns out to be differentiable almost everywhere.

(ii) The Filippov differential inclusion theory [4] commonly used for nondelayed differential equations with discontinuity and intended to turn solutions into sliding

<sup>1</sup>The sliding modes cannot occur in the considered class of the systems. That is why it is not necessary to use more complicated definitions of the solutions for relay systems with delay (see, for example, [10, 14]).

modes, i.e., motions along the discontinuity locus (see detailed accounts in [10] and [14]), is not quite relevant in our situation. Indeed, the motion along the discontinuity locus should correspond to the zero solution  $x(t) \equiv 0$ , which cannot be stable. The reason is that, in the norm (9), the zero function  $\varphi(t) \equiv 0$  can be approximated by functions with at most one zero, which in turn generate either unbounded solutions or solutions with the sup-norm separated from zero (see, for example, [17]). On the other hand, an attempt to keep the zero level of an eventually vanishing solution leads to the relation

$$(\text{sign } x(t - \tau)) \big|_{x(t-\tau)=0} = F(x(t), t),$$

which is not natural for a controller based on only delayed information.

**The statement.** We provide here a general statement, solving the stated problem, and leave the precise formulation for section 3.

**Main result.** *Given system (1), (3) with  $F(x, t)$  satisfying (5), under certain restrictions to  $\alpha$ ,  $\beta$ ,  $h$ ,  $\delta$ , and  $k_0$ , there exist positive constants  $c, T_0, m$ , and  $\rho < 1$  such that*

(i) *in the case (6), for*

$$(11) \quad K(t) = \rho^n, \quad nT_0 \leq t < (n+1)T_0, \quad n = 0, 1, 2, \dots,$$

*all of the solutions with  $\max\{|x(0)|, |\dot{x}(0)|\} < c$  exponentially decay to zero;*

(ii) *in the case (7), for*

$$(12) \quad K(t) = \begin{cases} \rho^n, & nT_0 \leq t < (n+1)T_0, \quad n = 0, 1, 2, \dots, m-1, \\ \rho^m, & t \geq mT_0, \end{cases}$$

*all of the solutions with  $\max\{|x(0)|, |\dot{x}(0)|\} < c$  come to a neighborhood of zero, whose size is proportional to  $\delta$ .*

In section 3.1, we provide explicit formulas for all of the parameters  $\alpha$ ,  $\beta$ ,  $h$ ,  $\delta$ , and  $k_0$ , and in section 5, we make a numerical simulation.

The meaning of main result is that, whenever the parameters of the system (1) and the controller delay  $\tau$  satisfy some explicitly written restriction, a control presented by a step function  $K(t)$  with a priori fixed switch moments and amplitudes brings solutions to a prescribed neighborhood of zero. In other words, we propose an algorithm for a robust *semiglobal* stabilization of the oscillation magnitude, based on a retarded relay switching of the control gain, which requires only the knowledge of the sign for the controlled variable in the past and allows us to reject uncertainty in the time delay.

**1.4. The ideas behind the main result.** The idea of a piecewise constant control function  $u(t)$  can be traced back to [6], where such a controller, acting with a constant delay, has been used for an exponential stabilization of oscillations in the first order system

$$\dot{x}(t) = F(x(t), t) - \text{sign } x(t - h),$$

with  $F$  satisfying (5) and (6). The key observation was that, if  $k_0 h < \log 2$ , the solutions starting in a small neighborhood of zero cannot reach some critical value  $|x| = M_0$  during the time interval  $h$  and then must return to the zero level, that is, remain bounded and oscillating. Furthermore, for such solutions,  $\sup |x(t)| < M_1 <$

$M_0$ , and hence, switching the magnitude of sign from 1 to  $\rho = M_1/M_0 < 1$  at the moment  $t^*$ , with  $x(t^*) = 0$ , and making change  $x = \rho x^{(1)}$ , we come to an equation

$$\dot{x}^{(1)}(t) = F^{(1)}(x^{(1)}(t), t) - \text{sign } x^{(1)}(t - h),$$

with  $F^{(1)}$  again satisfying (5) and (6), which in turn means  $|x(t)| < \rho M_0$  as  $t \geq t^*$ . Performing inductively the same procedure, one obtains exponentially decreasing solutions. However, that controller was depending on the term  $F(x, t)$  and on the current solution, which made it hard to realize in practice. This difficulty has been resolved in [18], where a similar piecewise constant controller acting with a variable uncertain bounded delay and having a priori fixed switches provided an exponential decay of solutions with sufficiently small initial values.

In a similar way we obtain the main result for the second order system (1), (3). The background property, established in [17], states that, under certain restrictions on the positive parameters  $\alpha, \beta, k, K, h, c$ , the solutions to the equation

$$\alpha \ddot{x} = -\beta \dot{x} + kx - K \cdot \text{sign } x(t - h),$$

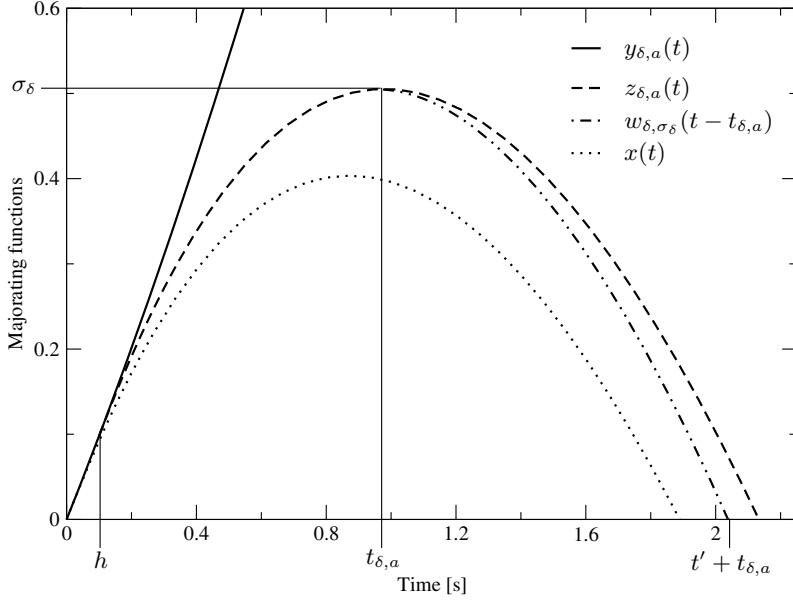
which obey the initial conditions  $x(0) = 0$ ,  $|\dot{x}(0)| < c$ , remain bounded by a constant  $M$ , proportional to  $K$ , and, moreover, the derivatives  $\dot{x}(t^*)$  for all  $t^* > 0$ ,  $x(t^*) = 0$ , belong to a *smaller* range  $(-c_1, c_1)$ , where  $c_1 < c$ . Here we extend this fact to the case of arbitrary functions  $F(x, t)$  with bounded values and derivative and a variable uncertain delay  $\tau(t)$ . So, again after a suitable period of time, we switch the controller magnitude from  $K$  to  $\rho K$ , with some  $\rho \in (c_1/c_0, 1)$ , and make change  $x = \rho x^{(1)}$ , coming to an equation for  $x^{(1)}$ , analogous to (1), (3) and satisfying the hypotheses, which provide  $|x^{(1)}| < M$  and  $|\dot{x}^{(1)}| < c$ , and, in particular,  $|x(t)| < \rho M$  for large  $t > 0$ .

To find suitable bounds to the given data, we model the “worst” behavior of a solution to (1), (3), which means that the absolute value  $|x|$  maximally grows against the negative feedback  $u$  intended to bring the solution to the zero level. That is, if a solution starts at zero with some, say, positive derivative, we assume that  $F(x, t) = \delta + kx$  and  $\tau = h$ , so that the feedback  $u$  remains positive on the largest possible interval of length  $h$ . Then we assume that the value of the control undergoes a switch only after a period of time  $h$  has elapsed, and we know that reversing the sign of the control will eventually force the solution to reach a maximum. When the maximum value is attained, we take  $F(x, t) = -\delta$  and wait until the solution reaches the next zero. We call the pieces of that worst solution *majorating functions*. They are treated in the next section in order to precisely state the sufficient conditions for the existence of the controller proposed in the main result, and these conditions finally reduce to the claim that the absolute value of the derivative of the worst solution at its zero is *strictly greater* than that value at the next zero.

## 2. Majorating functions.

**2.1. Definition of the majorating functions.** First, we point out that restriction (5) on the nonlinearity  $F(x, t)$  comes from the comparison of system (1), (3) with the equation  $\alpha \ddot{x} = -\beta \dot{x} + k_0 x - \text{sign } x(t - \tau)$ , and this makes it natural to introduce the roots  $\lambda_1 > 0 > \lambda_2$  of the characteristic equation  $\alpha \lambda^2 + \beta \lambda - k_0 = 0$ , which will play an important role in the further consideration.

In order to deal with uncertainty, we introduce a majorating function for the actual response  $x(t)$ . This function is intended to model “the worst type of behavior” of the stable solutions to (1), (3). As previously stated, such solutions are periodic

FIG. 1. Actual response  $x(t)$  and the majorating functions.

and slow. In Figure 1 we show (as a dotted line) the upper lobe of one of the periods. Assuming that the distance between the neighboring zeros of  $x(t)$  is greater than  $h$ , we can divide the interval between such zeros into three parts:

1. an interval between the current zero and the (first) control switch,
2. an interval between the control switch and the global extremum,
3. the remaining part from the extremum to the next zero.

For the first interval consider the equation

$$\alpha \ddot{y}(t) = -\beta \dot{y}(t) + k_0 y(t) + 1 + \delta.$$

We assume that the control switch is delayed by  $h$  from the current zero, and, since we are considering the upper lobe, the initial derivative is positive. Hence we impose

$$0 \leq t \leq h, \quad y(0) = 0, \quad \dot{y}(0) = a,$$

where  $a$  is a nonnegative parameter.

The family of functions  $y_{\delta,a}(t)$  (see Figure 1)

$$y_{\delta,a}(t) = \frac{ak_0 - \lambda_2(1 + \delta)}{k_0(\lambda_1 - \lambda_2)} e^{\lambda_1 t} + \frac{\lambda_1(1 + \delta) - ak_0}{k_0(\lambda_1 - \lambda_2)} e^{\lambda_2 t} - \frac{1 + \delta}{k_0}$$

are the solutions of the previous equation.

For the second interval we introduce the solution

$$(13) \quad \begin{aligned} z_{\delta,a}(t) = & -\frac{2e^{-\lambda_1 h} - 1 - \delta - \alpha a \lambda_1}{\alpha \lambda_1(\lambda_1 - \lambda_2)} e^{\lambda_1 t} + \frac{2e^{-\lambda_2 h} - 1 - \delta - \alpha a \lambda_2}{\alpha \lambda_2(\lambda_1 - \lambda_2)} e^{\lambda_2 t} \\ & + \frac{1 - \delta}{k_0} \end{aligned}$$

of the equation

$$\alpha \ddot{z}(t) = -\beta \dot{z}(t) + k_0 z(t) - 1 + \delta, \quad z(h) = y_{\delta,a}(h), \quad \dot{z}(h) = \dot{y}_{\delta,a}(h).$$

Suppose that  $2e^{-\lambda_1 h} - 1 > 0$ ,

$$(14) \quad \delta < 2e^{-\lambda_1 h} - 1,$$

and

$$(15) \quad a < \frac{2e^{-\lambda_1 h} - 1 - \delta}{\alpha \lambda_1}.$$

This means, in particular, that the coefficients of  $e^{\lambda_1 t}$  and  $e^{\lambda_2 t}$  in (13) are negative. Hence  $z_{\delta,a}(t)$  is a concave function, which in view of  $\dot{z}_{\delta,a}(h) = \dot{y}_{\delta,a}(h) > 0$  has a unique maximum in  $(h, \infty)$ . The maximum occurs at the time moment

$$(16) \quad t_{\delta,a} = \frac{1}{\lambda_1 - \lambda_2} \log \frac{2e^{-\lambda_2 h} - 1 - \delta - \alpha a \lambda_2}{2e^{-\lambda_1 h} - 1 - \delta - \alpha a \lambda_1},$$

so for  $z_{\delta,a}(t)$  we add the restriction

$$h \leq t \leq t_{\delta,a}.$$

To keep the notation simple, we will set

$$(17) \quad \sigma_\delta \triangleq z_{\delta,a}(t_{\delta,a}).$$

For the last interval we introduce the equation

$$\alpha \ddot{w}(t) = -\beta \dot{w}(t) - 1 - \delta.$$

In what follows, we will be more concerned about the value of the global extremum  $\sigma_\delta$ , rather than the time of its occurrence  $t_{\delta,a}$ , so we add

$$w(0) = \sigma_\delta, \quad \dot{w}(0) = 0,$$

which defines a majorating function which is shifted in time (see Figure 1). The solution is given by

$$(18) \quad w_{\delta,\sigma_\delta}(t) = \frac{\alpha(1+\delta)}{\beta^2} (1 - e^{-t\beta/\alpha}) - \frac{1+\delta}{\beta} t + \sigma_\delta,$$

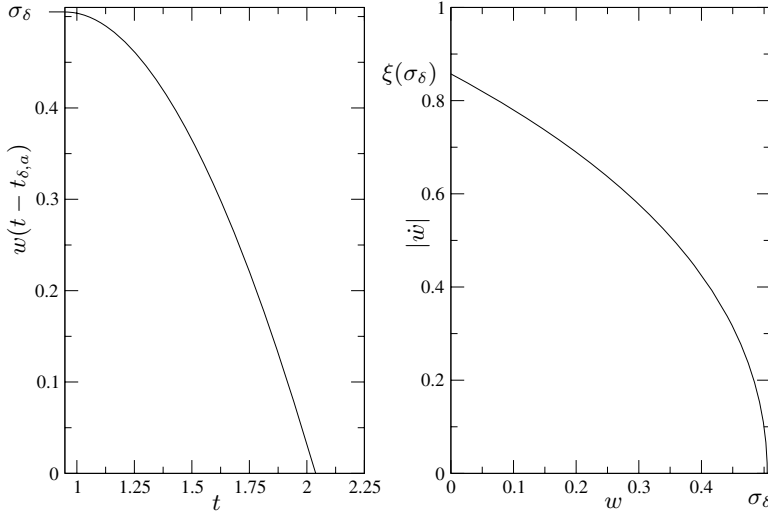
which has a unique positive root  $t'$  (see Figure 2).

Now we can build the majorating function. For any perturbation with a bound  $\delta$  satisfying (14) and an initial derivative  $a$  satisfying (15), the function

$$(19) \quad \phi_{\delta,a}(t) \triangleq \begin{cases} y_{\delta,a}(t), & 0 \leq t \leq h, \\ z_{\delta,a}(t), & h \leq t \leq t_{\delta,a}, \\ w_{\delta,\sigma_\delta}(t - t_{\delta,a}), & t_{\delta,a} \leq t \leq t' + t_{\delta,a}, \end{cases}$$

bounds from above the solutions to (1), (3) and their derivatives. We shall call  $\phi_{\delta,a}(t)$  *the worst solution* to (1), (3).



FIG. 2. Behavior of the last majorating function  $w(t - t_{\delta,a})$ .

**2.2. Properties of the majorating functions.** The next property will play a key role in the argumentation that follows.

**DEFINITION 2.1.** A function  $\phi_a(t) \in C_1$  is said to be differentially contractive (DC) if, whenever it starts at a zero with a derivative belonging to some interval, its derivative at the next zero belongs to a smaller interval.

Notice that  $\phi_{\delta,a}$  is continuous (see Figure 1). Its initial derivative  $a$  is taken from the interval (15), and we want its terminal derivative

$$\xi_\delta(\sigma_\delta) \triangleq |\dot{w}_{\delta,\sigma_\delta}(t')|$$

to belong to a smaller interval.

To fulfill the DC property, we first use (18) to estimate  $t'$

$$\frac{\alpha(1+\delta)}{\beta^2} (1 - e^{-t'/\beta/\alpha}) = \frac{1+\delta}{\beta} t' - \sigma_\delta.$$

Due to  $\dot{z}_{\delta,a}(t_{\delta,a}) = 0$  and  $\ddot{z}_{\delta,a}(t_{\delta,a}) < 0$ , we have

$$(20) \quad \sigma_\delta = \frac{1 - \delta + \beta \dot{z}_{\delta,a}(t_{\delta,a}) + \alpha \ddot{z}_{\delta,a}(t_{\delta,a})}{k_0} < \frac{1 - \delta}{k_0}.$$

Hence  $t' \leq \theta$ , where  $\theta$  is the positive root of the equation

$$(21) \quad \frac{\alpha(1+\delta)}{\beta^2} (1 - e^{-\theta\beta/\alpha}) = \frac{1+\delta}{\beta} \theta - \frac{1-\delta}{k_0}.$$

Notice that, given  $\alpha, \beta, \delta, k_0$ , (21) always has a unique positive root  $\theta$ , since the left-hand side is a positive concave function of  $\theta$  and the right-hand side is an increasing linear function of  $\theta$ , negative at the origin. Next, we have that

$$(22) \quad \xi_\delta(\sigma_\delta) = \frac{1+\delta}{\beta} (1 - e^{-t'/\beta/\alpha}) \leq \frac{1+\delta}{\beta} (1 - e^{-\theta\beta/\alpha}).$$

In view of the last inequality and (15) it is easy to see that  $\phi_{\delta,a}$  fulfills the DC property if

$$(23) \quad \frac{1+\delta}{\beta} (1 - e^{-\theta\beta/\alpha}) < \frac{2e^{-\lambda_1 h} - 1 - \delta}{\alpha\lambda_1}.$$

To understand inequality (23), consider the equality

$$(24) \quad \frac{1+\delta}{\beta} (1 - e^{-\theta(k)\beta/\alpha}) = \frac{2e^{-\lambda_1(k)h} - 1 - \delta}{\alpha\lambda_1(k)}$$

as an equation to the unknown  $k$  with fixed  $\alpha, \beta, \delta$ . Here the left-hand side is a bounded positive function of  $k$ , whereas the right-hand side drops from infinity to negative values as  $k$  grows from zero to infinity. Hence (24) has positive roots,<sup>2</sup> and the minimal one among them we denote by  $k_{\min}$ . So, finally, we reduce (23) to

$$(25) \quad k_0 < k_{\min},$$

which guarantees the DC property.

*Remark 2.* According to (20) the extremum  $\sigma_\delta$  is bounded from above. We shall call that bound  $\sigma_{\max}$ , i.e.,

$$\sigma_\delta < \sigma_{\max} \triangleq \frac{1-\delta}{k_0}.$$

Suppose that the extremum attains the maximum value in the current period; in view of (20), the extremum at the following period satisfies

$$\sigma_\delta^{(1)} \triangleq z_{\delta, \xi_\delta(\sigma_{\max})}(t_{\xi_\delta(\sigma_{\max})}) < \sigma_{\max}.$$

### 3. Main results in detail.

**DEFINITION 3.1.** Denote by  $\Phi_{\delta,a}$  the set of functions  $\varphi \in C_0[-h, 0]$  such that either

$$\varphi^{-1}(0) \neq \emptyset, \quad |\varphi(0)| \leq y_{\delta,a}(-t^*), \quad |\dot{\varphi}(0)| \leq \dot{y}_{\delta,a}(-t^*),$$

where  $t^* = \max \varphi^{-1}(0) > -h$ , or

$$\varphi|_{(-h, 0]} \neq 0, \quad |\varphi(0)| \leq z_{\delta,a}(t^*), \quad |\dot{\varphi}(0)| \leq \dot{z}_{\delta,a}(t^*)$$

for some  $t^* \in [h, t_{\delta,a}]$ .

**3.1. Perturbations that vanish at the origin.** Assume that  $\delta = 0$ . In order to simplify the notation, in this case we always skip the subindex  $\delta$  (i.e., 0) in the notation for  $t, \xi, \sigma, \Phi, x, y, z$ .

Introduce the following parameter. Given

$$0 < a < b < \frac{2e^{-\lambda_1 h} - 1}{\alpha\lambda_1},$$

set

$$(26) \quad \rho(a, b) \triangleq \frac{\alpha a(\lambda_1 e^{\lambda_1 h} - \lambda_2 e^{\lambda_2 h}) + (e^{\lambda_1 h} - e^{\lambda_2 h})}{\alpha b(\lambda_1 e^{\lambda_1 h} - \lambda_2 e^{\lambda_2 h}) + (e^{\lambda_1 h} - e^{\lambda_2 h})}.$$

<sup>2</sup>It is an easy exercise to show that a positive root is unique, but we shall not need this fact.

Clearly,  $\rho(a, b) < 1$ . Then introduce

$$(27) \quad \rho \triangleq \rho(\xi(\sigma^{(1)}), \xi(\sigma_{\max})).$$

Notice that  $\rho$  is defined properly, since  $\xi$  is a strictly increasing function.

**THEOREM 1.** *Assume that  $F(x, t)$  and  $\tau(t)$  satisfy (2), (4), (5), (6), and (25) with  $\delta = 0$ . Let a constant  $c$  satisfy*

$$(28) \quad 0 < c < \frac{2e^{-\lambda_1 h} - 1}{\alpha \lambda_1}.$$

Put

$$K(t) = \begin{cases} 1 & \text{if } 0 \leq t < t_c, \\ \rho^n & \text{if } t_c + nt_{\xi(\sigma_{\max})} \leq t < t_c + (n+1)t_{\xi(\sigma_{\max})}, \\ & n = 0, 1, 2, \dots, \end{cases}$$

where  $\rho$  is defined by (27), and  $t_c$  and  $t_{\xi(\sigma_{\max})}$  are the roots of  $z_c(t)$  and  $z_{\xi(\sigma_{\max})}(t)$ , respectively (see Figure 1).

Then any solution  $x_\varphi(t)$  to (1), (3), (8) with  $\varphi \in \Phi_c$  obeys the restriction

$$(29) \quad |x_\varphi(t)| \leq \frac{1}{k_0} \exp \left( - \left( \log \frac{1}{\rho} \right) \frac{t - t_c - t_{\xi(\sigma_{\max})}}{t_{\xi(\sigma_{\max})}} \right), \quad t \geq t_c + t_{\xi(\sigma_{\max})}.$$

**3.2. Perturbations that do not vanish at the origin.** In realistic models,  $F(0, t)$  does not vanish identically, so we'll consider the case  $\delta \neq 0$ , but we'll maintain restriction (14). In this case, one can drive the system in a finite time to a neighborhood of zero, proportional to  $\delta$ . More precisely, we design a set of controllers, which depend on one continuous and one discrete parameter. The parameters can be chosen in their range according to the initial magnitude, the required rate of convergence, and the size of the target neighborhood of zero. We remark only that one cannot optimize the two latter values simultaneously.

Given  $\delta$  satisfying (14), the range of a positive parameter  $\varepsilon$  is defined by the inequality

$$(30) \quad \frac{1 + \delta}{\beta} (1 - e^{-\theta\beta/\alpha}) + \frac{\delta}{\alpha \lambda_1} < \frac{2e^{-\lambda_1 h} - 1 - \varepsilon}{\alpha \lambda_1}.$$

Observe that (30) defines a nonempty interval, since it turns into (23) for  $\varepsilon = 0$ . Next we choose any natural  $m \geq 1$  and put  $q = q(\varepsilon, m)$  to be the positive root of the equation

$$(31) \quad \frac{1}{q} \left( \frac{1 - e^{-\theta\beta/\alpha}}{\beta} + \frac{(1 - q)(e^{\lambda_1 h} - e^{\lambda_2 h})}{\alpha(\lambda_1 e^{\lambda_1 h} - \lambda_2 e^{\lambda_2 h})} \right) + \frac{\delta}{q^m} \left( \frac{1 - e^{-\theta\beta/\alpha}}{\beta} + \frac{1}{\alpha \lambda_1} \right) = \frac{2e^{-\lambda_1 h} - 1 - \varepsilon}{\alpha \lambda_1}.$$

Such a root does exist; furthermore, it is unique and belongs to the interval  $(0, 1)$ . Indeed, the left-hand side of (31) monotonically decreases from infinity to the left-hand

side of (30), whereas the right-hand sides of (30) and (31) coincide. Furthermore,

$$(32) \quad \frac{1}{q} \left( \frac{1 - e^{-\theta\beta/\alpha}}{\beta} + \frac{(1-q)(e^{\lambda_1 h} - e^{\lambda_2 h})}{\alpha(\lambda_1 e^{\lambda_1 h} - \lambda_2 e^{\lambda_2 h})} \right) + \frac{\delta}{q^{m'}} \left( \frac{1 - e^{-\theta\beta/\alpha}}{\beta} + \frac{1}{\alpha\lambda_1} \right) \leq \frac{2e^{-\lambda_1 h} - 1 - \varepsilon}{\alpha\lambda_1}$$

for all  $m' \leq m$ . At last, put

$$(33) \quad T(\varepsilon) = \frac{1}{\lambda_1} \log \frac{(1-\delta)(\lambda_1 - \lambda_2)}{-\lambda_2 \varepsilon}.$$

**THEOREM 2.** *Under the hypotheses (2), (4), (5), (7), and (25) with  $\delta > 0$  satisfying (14), let  $\varepsilon$  obey (30). Put*

$$(34) \quad K(t) = \begin{cases} q^s & \text{if } sT(\varepsilon) \leq t < (s+1)T(\varepsilon), \quad s = 0, 1, \dots, m-1, \\ q^m & \text{if } t \geq mT(\varepsilon). \end{cases}$$

Then any solution  $x_\varphi(t)$  to (1), (3), (8), with  $\varphi \in \Phi_{\delta,c}$ , where

$$(35) \quad c = \frac{2e^{-\lambda_1 h} - 1 - \delta - \varepsilon}{\alpha\lambda_1},$$

obeys the restriction

$$(36) \quad |x_\varphi(t)| \leq \frac{q^m - \delta}{k_0} \quad \text{as } t \geq mT(\varepsilon).$$

### Comments to Theorem 2.

(i) We point out that the parameters of  $u(t)$  depend only on  $h$ ,  $k_0$ , and the chosen constants  $c$ ,  $\varepsilon$ ,  $m$  and do not depend on the function  $\tau_0(t)$  from (4); i.e., our feedback  $-K(t) \cdot \text{sign } x(t - \tau(t))$  is *robust* with respect to an uncertain variable delay  $\tau(t)$ , as well as a deviation of  $F(x, t)$  in the framework of restrictions (2), (7), (5), (25), (21).

(ii) If  $m \rightarrow \infty$  and  $\varepsilon \rightarrow 0$  in Theorem 2, then

$$q^m \rightarrow \delta \frac{\alpha\lambda_1(1 - e^{-\theta\beta/\alpha}) + \beta}{(2e^{-\lambda_1 h} - 1)\beta - \alpha\lambda_1(1 - e^{-\theta\beta/\alpha})},$$

and hence, the right-hand side of (36) tends to

$$(37) \quad K_0 \delta \triangleq \frac{2(\alpha\lambda_1(1 - e^{-\theta\beta/\alpha}) + \beta(1 - e^{-\lambda_1 h}))}{k_0((2e^{-\lambda_1 h} - 1)\beta - \alpha\lambda_1(1 - e^{-\theta\beta/\alpha}))} \delta.$$

(iii) The hypotheses (14), (25), and (30) in Theorem 2 contain restrictions to the parameters of system (1), (3) in an implicit form. However, one can in principle extract some explicit conditions from them. First of all, given  $\alpha$  and  $\beta$ , the parameter  $k_0$  must satisfy (25) for  $\delta = 0$ . Suppose now that such  $\alpha$ ,  $\beta$ , and  $k_0$  are fixed, and describe  $\delta$  which meet the hypotheses (14), (25), and (30).

Condition (14) is an explicit upper bound to  $\delta$ .

In turn,  $k_{\min} = k_{\min}(\delta)$  in (25) is a strictly decreasing function of  $\delta$  (indeed, when  $\delta$  grows, the allowed range for  $k_0$  must shrink), and this function is given by (21) and (24). That is, (25) also can be written as an upper bound  $\delta < (k_{\min})^{-1}(k_0)$ .

At last, condition (30) after removing  $\varepsilon$  (which can be arbitrarily small positive) reduces to inequality (23), where  $\theta$  comes from (21). It is not difficult to show that the latter equation defines  $\theta$  as a strictly decreasing function of  $\delta$ . Inequality (23) holds for  $\delta = 0$ , since this is equivalent to the above assumption  $k_0 < k_{\min}|_{\delta=0}$ . Thus, the left-hand side of (23) is a positive function of  $\delta$ , whereas the right-hand side drops from positive to negative as  $\delta$  goes from zero to  $\infty$ . Equating both sides of (23), we then obtain the minimal positive root  $\delta_{\min}$  and finally reduce condition (30) to an upper bound  $\delta < \delta_{\min}$ .

**4. Control algorithm.** We shortly describe how to apply Theorems 1 and 2. One begins with a few common initial steps:

1. Given system (1), (3), obeying (2), (4), and (7) with known  $h > 0$  and  $\delta \geq 0$ , we start by solving simultaneously the equations

$$(38a) \quad \frac{\alpha(1+\delta)}{\beta^2} (1 - e^{-\theta_m \beta / \alpha}) = \frac{1+\delta}{\beta} \theta_m - \frac{1-\delta}{k_m},$$

$$(38b) \quad \frac{1+\delta}{\beta} (1 - e^{-\theta_m \beta / \alpha}) = \frac{2e^{-\lambda_1(k_m)h} - 1 - \delta}{\alpha \lambda_1(k_m)}$$

with respect to positive unknowns  $k_m$  and  $\theta_m$ .

2. Take the solution  $(k_m, \theta_m)$ , and verify that the given function  $F(x, t)$  satisfies (5) with certain positive  $k_0 < k_m$ ; then find the positive root  $\theta$  of (21).
3. Compute the roots  $\lambda_1 > 0 > \lambda_2$  of the characteristic equation, and check the validity of (14).

**4.1. Perturbations that vanish at the origin.** Perform steps 1–3 as described above and then do the following.

4. Pick a constant  $c$ , satisfying (28), and compute the values of  $\xi(\sigma_{\max})$ ,  $t_c$ ,  $t_{\xi(\sigma_{\max})}$ , and  $\rho$  using the formulas of Theorem 1. Verify that the initial function  $\varphi$  belongs to  $\Phi_c$  as described in Definition 3.1.

*Remark 3.* It is possible to set a limit  $n^*$  to the maximum number of allowed switches of the controller or to the time interval  $t \leq t^*$ , when switches are allowed. Pick  $n \leq n^*$  or  $n \leq (t^* - t_c)/t_{\xi(\sigma_{\max})}$ , respectively. The solution becomes bounded by  $|x(t)| \leq \rho^n/k_0$  after  $t \geq t^*$ .

**4.2. Perturbations that do not vanish at the origin.** Again perform steps 1–3 as above, and then proceed in the following way.

4. Pick a positive  $\varepsilon$  satisfying (30), and compute  $T(\varepsilon)$  by (33).
5. For the last step there are three possibilities:
  - (a) Choose an upper bound  $m^*$  to the number of allowed switches of the controller, and pick  $m \leq m^*$ .
  - (b) Set the size  $t^*$  of the time interval when switches are allowed, and pick  $m \leq t^*/T(\varepsilon)$ .

In both cases solve (31) with respect to  $q$ . The solution will be bounded according to (36).

- (c) In this case we bring the solution to the  $\delta(K_0 + \kappa)$ -neighborhood of zero, where  $K_0$  is taken from (37), and  $\kappa$  is a (relatively) small prescribed positive parameter. Using (31) we compute

$$q = \frac{B_1 + B_2}{C + B_2 - (1 - \varepsilon)/(k_0(K_0 + \kappa) + 1)},$$

where

$$B_1 = \frac{1 - e^{-\theta\beta/\alpha}}{\beta}, \quad B_2 = \frac{(e^{\lambda_1 h} - e^{\lambda_2 h})}{\alpha(\lambda_1 e^{\lambda_1 h} - \lambda_2 e^{\lambda_2 h})}, \quad C = \frac{2e^{-\lambda_1 h} - 1 - \varepsilon}{\alpha\lambda_1},$$

and, finally, put

$$m = \left\lceil \frac{\log(\delta(k_0(K_0 + \kappa) + 1)/(1 - \varepsilon))}{\log q} \right\rceil + 1.$$

**5. Numerical example: Stabilization of an inverted pendulum.** Consider the stabilization problem of an inverted pendulum via a controller with uncertain delay. The oscillations of an inverted pendulum with unit mass with such a controller are described by

$$(39) \quad \ddot{x} + k\dot{x} - p\sin x + \delta = u(t - \tau(t)),$$

where  $k > 0$  is a friction coefficient,  $p = g/l > 0$ ,  $\delta$  is uncertainty, and  $\tau$  is an uncertain time delay  $0 < \tau_0(t) \leq \tau(t) \leq h$ . Consider the case when  $k = 1$  and  $p = g/l = 1.4$ . In this case (39) takes the form

$$\ddot{x}(t) = -\dot{x}(t) + 1.4\sin(x) + \delta - K(t)\text{sign}(x(t - \tau(t))),$$

with  $\tau = 0.05 + 0.04\sin(t)$ . It is clear that

$$\alpha = \beta = 1 \quad \text{and} \quad F(x, t) = 1.4\sin(x) + \delta$$

and that the bound

$$0 < \tau_0(t) \leq \tau(t) \leq h = 0.1$$

holds.

**5.1. Application of Theorem 1 ( $\delta = 0$ ).** The solution to (38) is  $k_m = 2.69518$ ,  $\theta_m = 1.00498$ . A possible  $k_0 < k_m$  that satisfies (5) is  $k_0 = 1.5$ . For that  $k_0$  we use (21) to obtain  $\theta = 1.42652$ . The roots of the characteristic equation are

$$\lambda_1 = 0.82288 \quad \text{and} \quad \lambda_2 = -1.82288.$$

The validity of (14) is easily verified:  $\delta = 0 < 2e^{-\lambda_1 h} - 1 = 0.84301$ .

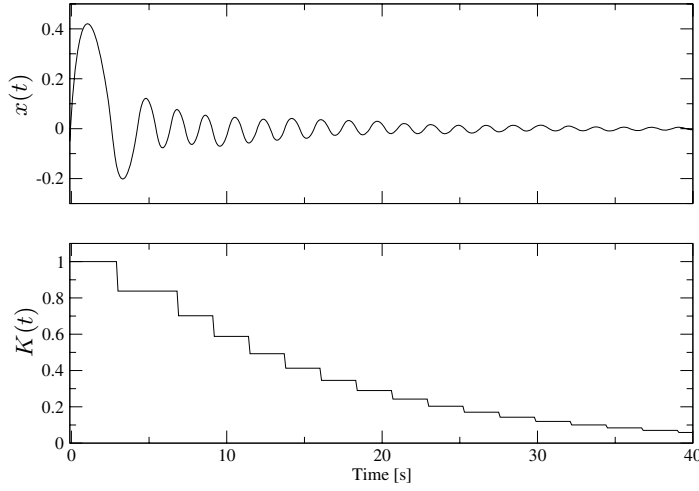
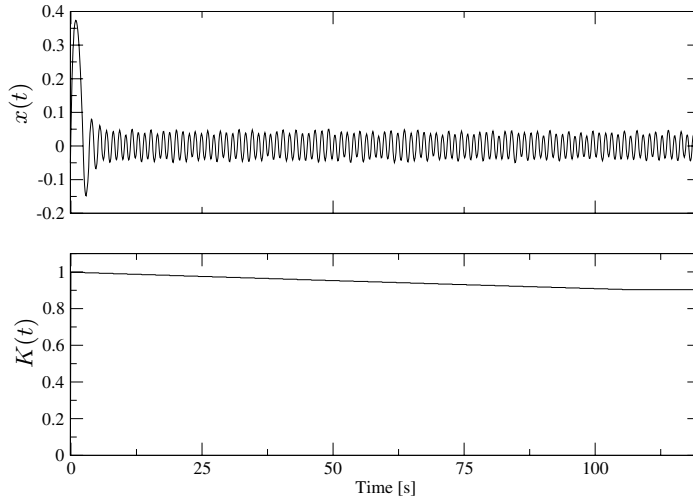
Now we pick a constant  $c$  satisfying (28)

$$c = 1 < \frac{2e^{-\lambda_1 h} - 1}{\alpha\lambda_1} = 1.0233.$$

Next, we have

$$\xi(\sigma_{\max}) = \frac{1 - e^{-\theta\beta/\alpha}}{\beta} = 0.7599, \quad t_c = 5.26, \quad t_{\xi(\sigma_{\max})} = 2.294, \quad \rho = 0.837.$$

Figure 3 shows the response of the system when the initial conditions are set to  $\dot{x}(-h) = 1$  and  $x(-h) = 0.05$ .

FIG. 3. *System's response*,  $\delta = 0$ .FIG. 4. *System's response*,  $\delta = 0.05$ .

**5.2. Application of Theorem 2 ( $\delta = 0.05$ ).** The following parameters were obtained as in the previous section:

$$\begin{aligned} k_m &= 2.2854, & k_0 &= 1.5, & \lambda_1 &= 0.8229, \\ \theta_m &= 1.0438, & \theta &= 1.3418, & \lambda_2 &= -1.8229, \\ \delta &< 2e^{-\lambda_1 h} - 1 &= 0.7930. \end{aligned}$$

Now we pick an  $\varepsilon = 0.15$  satisfying (30) and evaluate  $T(\varepsilon) = 2.696$ . For the last step we choose  $m = m^* = 40$  and obtain  $q = 0.9975$ .

The results are shown in Figure 4. The initial data were  $\dot{x}(-h) = 1$  and  $x(-h) = -0.05$ .

**6. Conclusions.** The dynamics of the second order systems with a delayed relay control is analyzed. Sufficient conditions for robust delayed relay *semiglobal* stabilization of second order systems are found. Such conditions relate to the upper bound of an uncertainty in time delay and the parameters of the plant. An algorithm for a delayed relay control with gain adaptation is suggested. The algorithm *is based on delayed information about the sign of the controlled variable only*. The proposed algorithm suppresses *bounded uncertainties in the time delay*: Once being designed for the upper bound of time delay in the given system, this control law ensures semiglobal stabilization for any constant or variable time delay within the given constraint.

## Appendix A. Proofs.

**A.1. Preliminary estimates.** In what follows we always suppose that  $\tau(t)$  and  $F(x, t)$  satisfy restrictions (4) and (7), respectively.

LEMMA A.1. (i) *Let*

$$0 < a < b < \frac{2e^{-\lambda_1 h} - 1}{\alpha \lambda_1}.$$

*Then*

$$(A.1) \quad y_a(t) \leq \rho(a, b)y_b(t), \quad \dot{y}_{0,a}(t) \leq \rho(a, b)\dot{y}_{0,b}(t), \quad 0 \leq t \leq h,$$

$$(A.2) \quad z_a(t) \leq \rho(a, b)z_b(t), \quad \dot{z}_{0,a}(t) \leq \rho(a, b)\dot{z}_{0,b}(t), \quad h \leq t \leq t_a,$$

*with  $\rho(a, b)$  defined by (26).*

(ii) *Let  $0 < q < 1$  and  $a \geq 0$ . Put*

$$(A.3) \quad \delta_1 = \frac{\delta}{q}, \quad a_1 = \frac{a}{q} + \frac{(1-q)(e^{\lambda_1 h} - e^{\lambda_2 h})}{q\alpha(\lambda_1 e^{\lambda_1 h} - \lambda_2 e^{\lambda_2 h})}.$$

*Then*

$$(A.4) \quad q^{-1}y_{\delta,a}(t) \leq y_{\delta_1,a_1}(t), \quad q^{-1}\dot{y}_{\delta,a}(t) \leq \dot{y}_{\delta_1,a_1}(t), \quad 0 \leq t \leq h,$$

$$(A.5) \quad q^{-1}z_{\delta,a}(t) \leq z_{\delta_1,a_1}(t), \quad q^{-1}\dot{z}_{\delta,a}(t) \leq \dot{z}_{\delta_1,a_1}(t), \quad h \leq t \leq t_{\delta,a}.$$

*Proof.* (i) The first inequality in (A.1) follows from the second one, and, respectively, the first inequality in (A.2) follows from the second inequality and from (A.1).

The second inequality in (A.1) can be rewritten as

$$\rho(a, b) \geq \max_{[0, h]} \frac{\dot{y}_a(t)}{\dot{y}_b(t)}.$$

We have

$$\frac{\dot{y}_a(t)}{\dot{y}_b(t)} = \frac{a\alpha\Sigma(t) + 1}{b\alpha\Sigma(t) + 1}, \quad \Sigma(t) = \frac{\lambda_1 e^{\lambda_1 t} - \lambda_2 e^{\lambda_2 t}}{e^{\lambda_1 t} - e^{\lambda_2 t}}.$$

Since

$$\frac{d}{ds} \left( \frac{a\alpha s + 1}{b\alpha s + 1} \right) = \frac{\alpha(a-b)}{(b\alpha s + 1)^2} < 0, \quad \dot{\Sigma}(t) = -\frac{(\lambda_1 - \lambda_2)^2 e^{-t/\alpha}}{(e^{\lambda_1 t} - e^{\lambda_2 t})^2} < 0,$$



we obtain in view of (26)

$$\max_{[0,h]} \frac{\dot{y}_a(t)}{\dot{y}_b(t)} = \frac{\dot{y}_b(h)}{\dot{y}_b(h)} = \rho(a, b).$$

We start proving the second inequality in (A.2) with the observation that  $t_a < t_b$ . Then, in particular,  $\dot{z}_{0,b}(t) > 0$  as  $h \leq t \leq t_a$ . Again we have to show that

$$\rho(a, b) \geq \max_{h, t_a} \frac{\dot{z}_a(t)}{\dot{z}_b(t)} = \max_{[h, t_a]} \frac{A_1(a)e^{\lambda_1 t} + A_2(a)e^{\lambda_2 t}}{A_1(b)e^{\lambda_1 t} + A_2(b)e^{\lambda_2 t}},$$

where

$$A_1(\sigma) = -(2e^{-\lambda_1 h} - 1 - \alpha\sigma\lambda_1), \quad A_2(\sigma) = 2e^{-\lambda_2 h} - 1 - \alpha\sigma\lambda_2.$$

Since

$$\frac{d}{dt} \left( \frac{A_1(a)e^{\lambda_1 t} + A_2(a)e^{\lambda_2 t}}{A_1(b)e^{\lambda_1 t} + A_2(b)e^{\lambda_2 t}} \right) = \frac{(A_1(a)A_2(b) - A_2(a)A_1(b))(\lambda_1 - \lambda_2)e^{-t/\alpha}}{(A_1(b)e^{\lambda_1 t} + A_2(b)e^{\lambda_2 t})^2}$$

and

$$A_1(a)A_2(b) - A_2(a)A_1(b) = \alpha(b-a)(\lambda_2(2e^{-\lambda_1 h} - 1) - \lambda_1(2e^{-\lambda_2 h} - 1)) < 0,$$

we derive in view of (26) that

$$\max_{h, t_a} \frac{\dot{z}_a(t)}{\dot{z}_b(t)} = \frac{\dot{z}_a(h)}{\dot{z}_b(h)} = \rho(a, b).$$

(ii) It is enough to establish the second inequality both in (A.4) and in (A.5). The second inequality in (A.4) can be rewritten as

$$q \geq \max_{[0,h]} \frac{\dot{y}_{\delta,a}(t)}{\dot{y}_{\delta_1,a_1}(t)} = \max_{[0,h]} \frac{a\alpha\Sigma(t) + 1 + \delta}{a_1\alpha\Sigma(t) + 1 + \delta_1}.$$

Since

$$\frac{d}{ds} \left( \frac{a\alpha s + 1 + \delta}{a_1\alpha s + 1 + \delta_1} \right) = -\frac{1-q}{q} \cdot \frac{a + (1+\delta)(\alpha\Sigma(h))^{-1}}{(a_1\alpha s + 1 + \delta_1)^2} < 0$$

and  $\dot{\Sigma}(t) < 0$ , we obtain

$$\max_{[0,h]} \frac{\dot{y}_{\delta,a}(t)}{\dot{y}_{\delta_1,a_1}(t)} = \frac{\dot{y}_{\delta,a}(h)}{\dot{y}_{\delta_1,a_1}(h)} = q.$$

The second inequality in (A.5) follows, first, from the fact that the function  $q^{-1}z_{\delta,a}(t)$  solves the problem

$$\alpha\ddot{z} = -\beta\dot{z} + k_0z - q^{-1} + \delta_1, \quad z(h) = q^{-1}y_{\delta,a}(h), \quad \dot{z}(h) = q^{-1}\dot{y}_{\delta,a}(h),$$

whereas the function  $z_{\delta_1,a_1}(t)$  solves the problem

$$\alpha\ddot{z} = -\beta\dot{z} + k_0z - 1 + \delta_1, \quad z(h) = y_{\delta_1,a_1}(h), \quad \dot{z}(h) = \dot{y}_{\delta_1,a_1}(h),$$

where  $y_{\delta_1, a_1}(h) \geq q^{-1}y_{\delta, a}(h)$ ,  $\dot{y}_{\delta_1, a_1}(h) = q^{-1}\dot{y}_{\delta, a}(h)$ , and, second, from the inequality  $t_{\delta, a} \leq t_{\delta_1, a_1}$ . In turn, the latter relation is an immediate consequence of (16).

LEMMA A.2. (i) Assume that  $t^* > 0$ ,  $t_0 \in [t^* - h, t^*]$ , and that  $a \geq 0$  satisfies (15). Let a solution  $x_\varphi(t)$  to the equation

$$(A.6) \quad \alpha \ddot{x}(t) = -\beta \dot{x}(t) + F(x(t), t) - \text{sign } x(t - \tau(t))$$

be such that

$$(A.7) \quad 0 \leq x_\varphi(t^*) \leq y_{\delta, a}(t^* - t_0), \quad \dot{x}_\varphi(t^*) \leq \dot{y}_{\delta, a}(t^* - t_0).$$

Then

$$x_\varphi(t) \leq y_{\delta, a}(t - t_0), \quad \dot{x}_\varphi(t) \leq \dot{y}_{\delta, a}(t - t_0), \quad t \in [t^*, t_0 + h].$$

(ii) Assume that  $t_0 \geq 0$  and that  $a \geq 0$  satisfies (15). Let a solution  $x_\varphi(t)$  obey the conditions

$$x_\varphi(t) > 0, \quad t \in (t_0, t_0 + h], \quad x_\varphi(t_0 + h) \leq z_{\delta, a}(\sigma), \quad \dot{x}_\varphi(t_0 + h) \leq \dot{z}_{\delta, a}(\sigma)$$

for some  $\sigma \in [h, t_{\delta, a}]$ . Then

$$x_\varphi(t) \leq z_{\delta, a}(t - \sigma + t_0 + h), \quad \dot{x}_\varphi(t) \leq \dot{z}_{\delta, a}(t - \sigma + t_0 + h)$$

for all  $t \geq t_0 + h$  such that  $x_\varphi(t) \geq 0$ .

*Proof.* (i) Since  $y_{\delta, a}(t)$  is a strictly increasing function, it is sufficient to consider the case when  $\dot{x}_\varphi(t) \geq 0$ ,  $t \in [t^*, t_0 + h]$ . Then, in the interval  $[t^*, t_0 + h]$ , we have

$$\alpha \ddot{x}_\varphi(t) = -\beta \dot{x}_\varphi(t) + F(x_\varphi(t), t) \pm 1 \leq -\beta \dot{x}_\varphi(t) + k_0 x_\varphi(t) + 1 + \delta,$$

which after a double integration turns into

$$(A.8) \quad \dot{x}_\varphi(t) \leq \dot{x}_\varphi(t^*) e^{-\beta(t-t^*)/\alpha} + \frac{1}{\alpha} \int_{t^*}^t (k_0 x_\varphi(\xi) + 1 + \delta) e^{\beta(\xi-t)/\alpha} d\xi,$$

$$(A.9) \quad \begin{aligned} x_\varphi(t) - x_\varphi(t^*) &\leq \frac{1+\delta}{\beta}(t-t^*) + \frac{\alpha}{\beta} \left( \dot{x}_\varphi(t^*) - \frac{1+\delta}{\beta} \right) (1 - e^{-(t-t^*)/\alpha}) \\ &\quad + \frac{k_0}{\beta} \int_{t^*}^t x_\varphi(\xi) (1 - e^{\beta(\xi-t)/\alpha}) d\xi. \end{aligned}$$

Due to the monotonicity of the right-hand sides with respect to  $x_\varphi$  and  $\dot{x}_\varphi$ , inequality (A.7), and the fact that the substitution of  $y_{\delta, a}(t)$  for  $x_\varphi(t)$  turns (A.8) and (A.9) into equalities, we obtain subsequently that  $x_\varphi(t) \leq y_{\delta, a}(t - t_0)$  and  $\dot{x}_\varphi(t) \leq \dot{y}_{\delta, a}(t - t_0)$ ,  $t \in [t^*, t_0 + h]$ .

(ii) Let  $x_\varphi(t) > 0$  in an interval  $[t_0 + h, t_1]$  for some  $t_1 > t_0 + h$ . Since  $x_\varphi(t)$  is positive in  $(t_0, t_0 + h]$ , and  $\tau(t) \leq h$ , we obtain

$$\alpha \ddot{x}_\varphi(t) \leq -\beta \dot{x}_\varphi(t) + k_0 x_\varphi(t) - 1 + \delta, \quad t \in [t_0 + h, t_1],$$

and hence

$$\begin{aligned}
 \dot{x}_\varphi(t) &\leq \dot{x}_\varphi(t_0 + h)e^{-\beta(t-t_0-h)/\alpha} \\
 &+ \frac{1}{\alpha} \int_{t_0+h}^t (k_0 x_\varphi(\xi) - 1 + \delta) e^{\beta(\xi-t)/\alpha} d\xi, \\
 x_\varphi(t) - x_\varphi(t_0 + h) &\leq -\frac{1-\delta}{\beta}(t - t_0 - h) \\
 &+ \frac{\alpha}{\beta} \left( \dot{x}_\varphi(t_0 + h) + \frac{1-\delta}{\beta} \right) (1 - e^{-\beta(t-t_0-h)/\alpha}) \\
 &+ \frac{k_0}{\beta} \int_{t_0+h}^t x_\varphi(\sigma) (1 - e^{\beta(\sigma-t)/\alpha}) d\sigma.
 \end{aligned}
 \tag{A.10}$$

These relations are monotone with respect to  $x_\varphi$  and  $\dot{x}_\varphi$  and thereby imply  $x_\varphi(t) \leq z_a(t - \sigma + t_0 + h)$ ,  $\dot{x}_\varphi(t) \leq \dot{z}_a(t - \sigma + t_0 + h)$ ,  $t \in [t_0 + h, t_1]$ , since the replacement of  $x_\varphi(t)$  by  $z_a(t - \sigma + t_0 + h)$  turns (A.10) and (A.11) into equalities.

LEMMA A.3. *For any nonnegative  $a$  satisfying (15), and any  $\varphi \in C_0[-h, 0]$  such that*

$$\varphi(0) = 0, \quad \dot{\varphi}(0) = a, \tag{A.12}$$

*the solution  $x_\varphi(t)$  to (A.6) satisfies the following conditions:*

- $x_\varphi(t)$  has an unbounded zero locus;
- the first positive zero of  $x_\varphi(t)$  does not exceed  $t_{\delta,a}$ ;
- if  $x_\varphi(t') = 0$  at  $t' > 0$ , then  $x_\varphi$  vanishes at some  $t'' \in (t', t' + t_{\delta,\xi_\delta((1-\delta)/k_0)})$ ;
- $x_\varphi(t)$  obeys the condition

$$|x_\varphi(t)| \leq z_{\delta,\xi_\delta((1-\delta)/k_0)}(t_{\delta,\xi_\delta((1-\delta)/k_0)}) \quad \text{if } t \geq t_{\delta,a}, \tag{A.13}$$

$$|\dot{x}_\varphi(t)| \leq \xi_\delta((1-\delta)/k_0) \quad \text{for all } t \in x_\varphi^{-1}(0) \cap (0, \infty). \tag{A.14}$$

*Proof. Step 1.* We first show that  $x_\varphi(t)$  has a positive root. Assume on the contrary that  $x_\varphi(t) > 0$  for all  $t > 0$ . By Lemma A.2(i),  $x_\varphi(t) \leq y_a(t)$ ,  $\dot{x}_\varphi(t) \leq \dot{y}_a(t)$  as  $t \in [0, h]$ , and hence by Lemma A.2(i),  $x_\varphi(t) \leq z_a(t)$  as  $t \geq h$ . However,  $z_a(t)$  becomes negative for large  $t$ , and so does  $x_\varphi$ . Furthermore, we obtain that  $x(t)$  vanishes at some point  $t_1 \leq t_a$ , where  $t_a$  denotes the positive zero of  $z_a(t)$ . The same argument provides the upper bound  $|x_\varphi(t)| \leq z_a(t_a)$  in the interval  $(0, t_1)$ .

*Step 2.* We intend now to estimate  $|\dot{x}_\varphi(t_1)|$ . Change sign of  $x_\varphi$  so that  $x_\varphi(t) < 0$  as  $t \in (0, t_1)$ . Let  $t' = \max\{t \in (0, t_1) \mid \dot{x}_\varphi(t) = 0\}$ . Since  $|x_\varphi(t')| \leq z_a(t_a) < \frac{1}{k_0}$ , in the interval  $[t', t_1]$ , we have

$$\alpha \ddot{x}_\varphi(t) = -\beta \dot{x}_\varphi(t) + F(x_\varphi(t), t) \pm 1 < -\dot{x}_\varphi(t) + 1, \tag{A.15}$$

which yields

$$\dot{x}_\varphi(t) < 1 - e^{(t'-t)/\alpha} < 1, \quad t \in [t', t_1].$$

Furthermore,  $x_\varphi(t)$  is strongly increasing in  $[t', t_1]$ , and then we can choose it as a variable and rewrite (A.15) in the form

$$\alpha \frac{d\dot{x}_\varphi}{dx_\varphi} \dot{x}_{1,\varphi} + \dot{x}_\varphi < 1 \quad \stackrel{0 \leq \dot{x}_\varphi < 1}{\implies} \int_0^{\dot{x}_\varphi(t_1)} \frac{\alpha \dot{x}_\varphi}{1 - \dot{x}_\varphi} d\dot{x}_{1,\varphi} < -x_\varphi(t').$$

The latter formula turns into the relation for  $\xi(1/k_0)$ , when replacing “ $<$ ” by “ $=$ ” and  $-x_\varphi(t')$  by  $\frac{1}{k_0}$ . Hence  $\dot{x}_\varphi(t_1) < \xi(1/k_0)$ .

*Step 3.* Observe now that (23) is equivalent to

$$\xi\left(\frac{1}{k_0}\right) < \frac{2e^{-\lambda_1 h} - 1}{\alpha\lambda_1}.$$

That is,  $x_\varphi(t - t_1)$  satisfies the hypotheses of Lemma A.3, and one can proceed inductively, proving the statements of the lemma for the whole interval  $[0, \infty)$ .

**A.2. Proof of Theorem 1.** In the interval  $[0, t_c + t_{\xi(1/k_0)}]$ , we have  $x_{u,\varphi}(t) = x_\varphi(t)$ . Hence the conditions imposed on the set  $\Phi_c \subset C_0[-h, 0]$  and Lemmas A.2 and A.3 yield that  $|x_{u,\varphi}(t)| \leq \frac{1}{k_0}$  as  $t \in [0, t_c]$  and  $(x_{u,\varphi})^{-1}(0) \cap (0, t_c] \neq \emptyset$ . Furthermore, for any  $t^* \in (x_{u,\varphi})^{-1}(0) \cap (0, t_c]$ , we have  $|\dot{x}_{u,\varphi}(t^*)| \leq \xi(1/k_0)$ . Next we apply Lemmas A.2 and A.3 to  $x_{u,\varphi}$  restricted to the interval  $[t_c, t_c + t_{\xi(1/k_0)}]$  and obtain that  $|x_{u,\varphi}(t)| \leq z_{\xi(1/k_0)}(t_{\xi(1/k_0)})$  if  $t \in [t_c, t_c + t_{\xi(1/k_0)}]$ , the set  $(x_{u,\varphi})^{-1}(0) \cap (t_c, t_c + t_{\xi(1/k_0)})$  is nonempty, and, for any  $t^* \in (x_{u,\varphi})^{-1}(0) \cap (t_c, t_c + t_{\xi(1/k_0)})$ , it holds that  $|\dot{x}_{u,\varphi}(t^*)| \leq \xi(z_{\xi(1/k_0)}(t_{\xi(1/k_0)}))$ .

In the interval  $[t_c + t_{\xi(1/k_0)}, t_c + 2t_{\xi(1/k_0)}]$ , we have

$$\alpha \ddot{x}_{u,\varphi}(t) = -\beta \dot{x}_{u,\varphi} + F(x_{u,\varphi}(t), t) - \rho \cdot \text{sign } x_{u,\varphi}(t - \tau(t)).$$

The variable change  $x_{u,\varphi}(t) = \rho \cdot x_{u,\varphi}^{(1)}(t)$ ,  $t \in [t_c + t_{\xi(1/k_0)}, t_c + 2t_{\xi(1/k_0)}]$ , leads to the equation

$$\alpha \frac{d^2}{dt^2} x_{u,\varphi}^{(1)}(t) = -\beta \frac{d}{dt} x_{u,\varphi}^{(1)}(t) + \frac{1}{\rho} F(\rho x_{u,\varphi}^{(1)}(t), t) - \text{sign } x_{u,\varphi}^{(1)}(t - \tau(t)).$$

Observe that the function  $F^{(1)}(x, t) := \frac{1}{\rho} F(\rho x, t)$  satisfies restrictions (2) and (5).

Put

$$(A.16) \quad s = \max\left((x_{u,\varphi})^{-1}(0) \cap [0, t_c + t_{\xi(1/k_0)}]\right).$$

Assume, first, that  $s > t_c + t_{\xi(1/k_0)} - h$ . Then by Lemma A.2(i)

$$|x_{u,\varphi}(t_c + t_{\xi(1/k_0)})| \leq y_{\xi(z_{\xi(1/k_0)}(t_{\xi(1/k_0)}))}(t_c + t_{\xi(1/k_0)} - s),$$

$$|\dot{x}_{u,\varphi}(t_c + t_{\xi(1/k_0)})| \leq \dot{y}_{\xi(z_{\xi(1/k_0)}(t_{\xi(1/k_0)}))}(t_c + t_{\xi(1/k_0)} - s).$$

Consequently, by Lemma A.1(i) and by the definition of  $\rho$ ,

$$|x_{u,\varphi}^{(1)}(t_c + t_{\xi(1/k_0)})| \leq y_{\xi(1/k_0)}(t_c + t_{\xi(1/k_0)} - s),$$

$$\left| \frac{d}{dt} x_{u,\varphi}^{(1)}(t_c + t_{\xi(1/k_0)}) \right| \leq \dot{y}_{\xi(1/k_0)}(t_c + t_{\xi(1/k_0)} - s);$$

that is,  $x_{u,\varphi}^{(1)}(t - t_c - t_{\xi(1/k_0)})$  satisfies the conditions of Lemmas A.2(i) and A.3 with  $a = \xi(1/k_0)$ .

Now assume that  $s$  from (A.16) satisfies  $s \leq t_c + t_{\xi(1/k_0)} - h$ . Then by Lemma A.2(ii)

$$|x_{u,\varphi}(t_c + t_{\xi(1/k_0)})| \leq z_{\xi(z_{\xi(1/k_0)}(t_{\xi(1/k_0)}))}(\sigma),$$

$$|\dot{x}_{u,\varphi}(t_c + t_{\xi(1/k_0)})| \leq \dot{z}_{\xi(z_{\xi(1/k_0)}(t_{\xi(1/k_0)}))}(\sigma)$$

for some  $\sigma \in [h, t_{\xi(z_{\xi(1/k_0)}(t_{\xi(1/k_0)}))}]$ . Consequently, by Lemma A.1(i) and by the definition of  $\rho$ ,

$$|x_{u,\varphi}^{(1)}(t_c + t_{\xi(1/k_0)})| \leq z_{\xi(1/k_0)}(\sigma), \quad \left| \frac{d}{dt} x_{u,\varphi}^{(1)}(t_c + t_{\xi(1/k_0)}) \right| \leq \dot{z}_{\xi(1/k_0)}(\sigma),$$

and here  $h \leq \sigma \leq t_{\xi(z_{\xi(1/k_0)}(t_{\xi(1/k_0)}))} < t_{\xi(1/k_0)}$ . Hence  $x_{u,\varphi}^{(1)}(t - t_c - t_{\xi(1/k_0)})$  satisfies the conditions of Lemmas A.2(ii) and A.3 with  $a = \xi(1/k_0)$ .

Under both of the assumptions, Lemmas A.2 and A.3 yield that

$$|x_{u,\varphi}^{(1)}(t)| \leq z_{\xi(1/k_0)}(t_{\xi(1/k_0)}), \quad t \in [t_c + t_{\xi(1/k_0)}, t_c + 2t_{\xi(1/k_0)}],$$

the set  $(x_{u,\varphi}^{(1)})^{-1}(0) \cap (t_c + t_{\xi(1/k_0)}, t_c + 2t_{\xi(1/k_0)})$  is nonempty, and

$$\left| \frac{d}{dt} x_{u,\varphi}^{(1)}(t^*) \right| \leq \xi(z_{\xi(1/k_0)}(t_{\xi(1/k_0)}))$$

as  $t^* \in (x_{u,\varphi}^{(1)})^{-1}(0) \cap (t_c + t_{\xi(1/k_0)}, t_c + 2t_{\xi(1/k_0)})$ . These properties of  $x_{u,\varphi}^{(1)}|_{[t_c + t_{\xi(1/k_0)}, t_c + 2t_{\xi(1/k_0)}]}$  coincide with the aforementioned properties of  $x_{u,\varphi}|_{[t_c, t_c + t_{\xi(1/k_0)}]}$ . Thus, one can proceed inductively, defining

$$x_{u,\varphi}(t) = \rho^n x_{u,\varphi}^{(n)}(t), \quad t \in [t_c + nt_{\xi(1/k_0)}, t_c + (n+1)t_{\xi(1/k_0)}], \quad n = 2, 3, \dots,$$

and deriving

$$|x_{u,\varphi}^{(n)}(t)| \leq z_{\xi(1/k_0)}(t_{\xi(1/k_0)}), \quad t \in [t_c + nt_{\xi(1/k_0)}, t_c + (n+1)t_{\xi(1/k_0)}].$$

The upper bound (29) follows immediately.

**A.3. Proof of Theorem 2.** In the interval  $[0, T(\varepsilon)]$ , we have  $x_{u,\varphi}(t) = x_{1,\varphi}(t)$ . The conditions imposed on the set  $\Phi_c \subset C_0[-h, 0]$  and Lemmas A.2 and A.3 yield that  $|x_{\varphi}(t)| \leq \frac{1-\delta-\varepsilon}{k_0}$  as  $t \geq 0$  and that  $(x_{u,\varphi})^{-1}(0) \cap (0, t_{\delta,c}] \neq \emptyset$ .

Indeed, using (13) for the equation  $z_{\delta,c}(t_{\delta,c}) = 0$ , we obtain

$$\frac{2e^{-\lambda_1 h} - 1 - \delta - \alpha c \lambda_1}{\alpha \lambda_1 (\lambda_1 - \lambda_2)} e^{\lambda_1 t_{\delta,c}} < \frac{1 - \delta}{k_0},$$

$$t_{\delta,c} < \frac{1}{\lambda_1} \log \frac{(1 - \delta) \alpha \lambda_1 (\lambda_1 - \lambda_2)}{k_0 (2e^{-\lambda_1 h} - 1 - \delta - \alpha c \lambda_1)} = \frac{1}{\lambda_1} \log \frac{(1 - \delta) (\lambda_1 - \lambda_2)}{-\lambda_2 \varepsilon} = T(\varepsilon)$$

(cf. (33) and (35)). Furthermore, we have  $|\dot{x}_{\varphi}(t^*)| < \xi_{\delta}((1 - \delta)/k_0)$  for any  $t^* \in (x_{\varphi})^{-1}(0) \cap (0, T(\varepsilon)]$ . That means  $x_{\varphi}(T(\varepsilon))$  and  $\dot{x}_{\varphi}(T(\varepsilon))$  satisfy the hypotheses of Lemma A.2(i) or (ii) with  $a = \xi_{\delta}(\sigma_{\delta})$ ,  $\sigma_{\delta}$  being defined by (17). Furthermore, using (20) and (22), we obtain

$$\begin{aligned} a = \xi_{\delta}(\sigma_{\delta}) &= \frac{1 + \delta}{\beta} (1 - e^{-t' \beta / \alpha}) = \frac{\beta}{\alpha} \left( \frac{1 + \delta}{\beta} t' - \sigma_{\delta} \right) \\ &> \frac{\beta}{\alpha} \left( \frac{1 + \delta}{\beta} t' - \frac{1 - \delta}{k_0} \right) \implies t' < \theta \\ (A.17) \quad &\implies a < \frac{1 + \delta}{\beta} (1 - e^{-\theta \beta / \alpha}). \end{aligned}$$

In the interval  $[T(\varepsilon), 2T(\varepsilon))$ , we have  $K(t) = q$ , and the variable change  $x_{u,\varphi}(t) = qx^{(1)}(t)$  leads to the equation

$$\alpha \frac{d^2}{dt^2} x^{(1)}(t) = -\beta \frac{d}{dt} x^{(1)}(t) + F_1(x^{(1)}(t), t) - \text{sign } x^{(1)}(t - \tau(t)), \quad t \in [T(\varepsilon), 2T(\varepsilon)),$$

where  $F_1(x, t) := q^{-1}F(qx, t)$  obeys restriction (5) and restriction (7) with  $\delta$  replaced by  $\delta_1 = \delta/q$ . In view of (A.17) and by Lemma A.1(ii),  $x^{(1)}(t)$  satisfies the conditions of Lemma A.2 with  $F$ ,  $\delta$ ,  $a$  replaced, respectively, by  $F_1$ ,  $\delta_1$ ,  $a_1$  defined in (A.3). Moreover,

$$\begin{aligned} a_1 &= \frac{a}{q} + \frac{(1-q)(e^{\lambda_1 h} - e^{\lambda_2 h})}{q\alpha(\lambda_1 e^{\lambda_1 h} - \lambda_2 e^{\lambda_2 h})} < \frac{1+\delta}{q\beta}(1 - e^{-\theta\beta/\alpha}) + \frac{(1-q)(e^{\lambda_1 h} - e^{\lambda_2 h})}{q\alpha(\lambda_1 e^{\lambda_1 h} - \lambda_2 e^{\lambda_2 h})} \\ (A.18) \quad &\leq \frac{2e^{-\lambda_1 h} - 1 - \delta/q - \varepsilon}{\alpha\lambda_1} = \frac{2e^{-\lambda_1 h} - 1 - \delta_1 - \varepsilon}{\alpha\lambda_1}, \end{aligned}$$

the first inequality following from (A.17) and the second one from (32). We obtain also that

$$|x_{u,\varphi}(t)| = q|x^{(1)}(t)| \leq q \frac{1 - \delta_1}{k_0} = \frac{q - \delta}{k_0}.$$

Relation (A.18) allows one to continue the procedure inductively by defining  $K(t)$  by (34) and the functions  $x^{(s)}(t)$ ,  $t \geq sT(\varepsilon)$ , by the formula  $x_{u,\varphi}(t) = q^s x^{(s)}(t)$  for  $s = 1, \dots, m$ . Inequality (36) follows immediately.

#### REFERENCES

- [1] M. AKIAN M., P.-A. BLIMAN, AND M. SORINE, *Control of delay systems with relay*, IMA J. Math. Control Inform., 19 (2002), pp. 133–155.
- [2] G. BARTOLINI, G. M. COCOLI, AND A. FERRARA, *Vibration damping and the second-order sliding modes in the control of single finger of the AMADEUS gripper*, Internat. J. Systems Sci., 29 (1998), pp. 497–512.
- [3] D. BARTON, B. KRAUSKOPF, AND R. E. WILSON, *Explicit periodic solutions in a model of a relay controller with delay and forcing*, Nonlinearity, 18 (2005), pp. 2637–2656.
- [4] A. F. FILIPPOV, *Differential Equations with Discontinuous Right-Hand Side*, Kluwer, Dordrecht, 1988.
- [5] E. FRIDMAN, L. FRIDMAN, AND E. SHUSTIN, *Steady modes and stability in discontinuous delay systems with periodic disturbances*, ASME J. Dynam. Syst. Measurement and Control, 122 (2000), pp. 732–737.
- [6] L. FRIDMAN, E. FRIDMAN, AND E. SHUSTIN, *Steady modes and sliding modes in relay control systems with delay*, in *Sliding Mode Control in Engineering*, J. P. Barbot and W. Perruquetti, eds., Marcel Dekker, New York, 2002, pp. 263–293.
- [7] L. FRIDMAN, V. STRYGIN, AND A. POLYAKOV, *Stabilization of oscillations amplitudes via relay delay control*, Internat. J. Control, 76 (2003), pp. 770–780.
- [8] L. FRIDMAN, V. STRYGIN, AND A. POLYAKOV, *Nonlocal stabilization via delayed relay control rejecting uncertainty in a time delay*, Internat. J. Robust Nonlinear Control, 14 (2004), pp. 15–37.
- [9] J. K. HALE AND S. M. VERDUYN LUNEL, *Introduction to Functional Differential Equations*, Springer, New York, 1993.
- [10] L. LEVAGGI, *Infinite dimensional systems sliding motions*, Eur. J. Control, 8 (2003), pp. 508–516.
- [11] L. LEVAGGI AND E. PUNTA, *Analysis of a second-order sliding-mode algorithm in presence of input delays*, IEEE Trans. Automat. Control, 51 (2006), pp. 1325–1332.
- [12] A. LONGTIN AND J. G. MILTON, *Modelling autonomous oscillations in the human pupil light reflex using nonlinear delay-differential equation*, Bull. Math. Biol., 51 (1989), pp. 605–624.

- [13] A. LONGTIN AND J. G. MILTON, *Insight into the transfer function, gain, and oscillation onset for the pupil light reflex using nonlinear delay-differential equations*, Biol. Cybern., 61 (1989), pp. 51–58.
- [14] YU. ORLOV, M. PERRUQUETTI, AND J. P. RICHARD, *Sliding mode control synthesis of uncertain time-delay systems*, Asian J. Control, 5 (2003), pp. 568–577.
- [15] W. B. RIBBENS, *Understanding Automotive Electronics*, 6th ed., Newnes, New York, 2003.
- [16] YU. SHTESSEL, A. ZINOBER, AND I. SHKOLNIKOV, *Sliding mode for nonlinear systems with output delay via method of stable system center*, ASME J. Dynam. Syst. Measurement and Control, 125 (2003), pp. 253–257.
- [17] E. SHUSTIN, E. FRIDMAN, AND L. FRIDMAN, *Oscillations in a second-order discontinuous system with delay*, Discrete Contin. Dyn. Syst., 9 (2003), pp. 339–358.
- [18] E. SHUSTIN, *Exponential decay of oscillations in a multidimensional delay differential system*, in Proceedings of the 4th International Conference on Dynamical Systems and Differential Equations, Wilmington, NC, 2002, American Institute of Mathematical Sciences, Springfield, MO, 2002, pp. 809–816.
- [19] V. I. UTKIN, *Sliding Modes in Control Optimization*, Springer, Berlin, 1992.

# CONTROLLED STOCHASTIC DIFFERENTIAL EQUATIONS UNDER CONSTRAINTS IN INFINITE DIMENSIONAL SPACES\*

RAINER BUCKDAHN<sup>†</sup>, MARC QUINCAMPOIX<sup>†</sup>, AND GIANMARIO TESSITORE<sup>‡</sup>

**Abstract.** In this paper we study the compatibility (or viability) of a given state constraint  $K$  with respect to a controlled stochastic evolution equation in a real Hilbert space  $H$ . We allow the noise to be a cylindrical Wiener process and admit an unbounded linear operator in the state equation. Our assumptions cover, for instance, controlled heat equations with space-time white noise. Our main result is to prove that if  $K$  is  $\varepsilon$ -viable, then the square of the distance from  $K$ :  $d_K^2(x) := \inf_{y \in K} |x - y|^2$  is a viscosity supersolution of a suitable class of fully nonlinear Hamilton–Jacobi–Bellman equations in  $H$ . This extends already obtained results into the finite dimensional case. We use the definition of viscosity supersolutions for “unbounded” elliptic equations in infinite variables that have been recently introduced by Świąch and Kelome. We discuss several cases where the above necessary condition is also sufficient.

**Key words.** viability, stochastic control

**AMS subject classifications.** 93E20, 35R60, 60H15, 49L25

**DOI.** 10.1137/060674284

**1. Introduction.** We investigate the (approximate) compatibility of the controlled stochastic differential equations

$$(1.1) \quad dX_t = (AX_t + F(X_t, u_t)) dt + G(X_t, u_t) dW_t, \quad t \geq 0,$$

with the state constraint

$$X_t \in K \text{ for every } t \geq 0.$$

The constraint set  $K$  is a closed subset of a separable Hilbert space  $H$ . A complete right continuous filtration  $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$  is given on a complete probability space  $(\Omega, \mathcal{F}, P)$ . The process  $W$  is a cylindrical Brownian motion on a separable Hilbert space  $\Xi$ . The control  $u_t$  is an  $\mathbb{F}$ -progressively measurable process from  $u : [0, \infty[ \times \Omega$  into a bounded closed subset  $U$  of a given Banach space. We denote by  $\mathcal{U}$  the set of all controls. The assumptions on the functions  $F : H \times U \rightarrow H$ ,  $G : H \times U \rightarrow L(\Xi, H)$ , and on the linear operator  $A : D(A) \subset H \rightarrow H$  will be discussed later in this section.

A solution  $X_t$  of (1.1) is said to be compatible (or viable) with the constraint  $K$  if and only if for all  $t \geq 0$  we have  $X_t \in K$   $P$  almost surely. Noticing that a viable trajectory  $X_t$  satisfies

$$\mathbb{E} \left[ \int_0^{+\infty} e^{-Ct} d_K^2(X_t) dt \right] = 0$$

---

\*Received by the editors November 7, 2006; accepted for publication (in revised form) June 15, 2007; published electronically January 22, 2008. This research was supported by European Community’s Human Potential Program under contract HPRN-CT-2002-00281, Evolution Equations.

<http://www.siam.org/journals/sicon/47-1/67428.html>

<sup>†</sup>Laboratoire de Mathématiques, UMR CNRS 6205, Université de Bretagne Occidentale, 6 Avenue Victor Le Gorgeu, 29200 Brest, France (Rainer.Buckdahn@univ-brest.fr, Marc.Quincampoix@univ-brest.fr).

<sup>‡</sup>Dipartimento di Matematica e Applicazioni, Università di Milano-Bicocca, Via R. Cozzi 53 - Edificio U5, 20125 Milano, Italy (gianmario.tessitore@unimib.it).



(where  $d_K(x) = \inf_{y \in K} |x - y|$  is the distance to  $K$  and  $C > 0$  a constant), we say that the constraint  $K$  is  $\varepsilon$ -viable (or approximatively compatible) for the control system (1.1) if and only if there exists a constant  $C$  large enough such that, for  $x \in K$ ,

$$(1.2) \quad \inf_{u \in \mathcal{U}} \mathbb{E} \left[ \int_0^{+\infty} e^{-Ct} d_K^2(X_t^{x,u}) dt \right] = 0,$$

where  $X^{x,u}$  denotes the solution of (1.1) associated with the initial condition  $X_0^{x,u} = x$  and with the control  $u \in \mathcal{U}$ . Remark that, obviously, when the infimum in (1.2) is achieved, we obtain a viable trajectory  $X_t$ .

Our main aim consists of characterizing of  $K$  for the control system (1.1).

Since pioneering work of Nagumo [26] for deterministic differential equations, this problem has been solved for deterministic control systems in various cases (see [1] and its bibliography). For finite dimensional stochastic control a characterization of viability has been obtained in [2, 3, 4, 17] through stochastic tangent cones to the set  $K$ . Also in finite dimension, another characterization has been obtained in [7] through the distance function to the set  $K$ . We refer the reader to [8, 9, 10, 28] for various extensions and applications of this method.

The characterization of [7] is crucially based on the Hamilton–Jacobi–Bellman equation

$$(1.3) \quad \Psi(x, V(x), DV(x), D^2V(x)) = 0, \quad x \in \mathbb{R}^n,$$

satisfied by the value function

$$(1.4) \quad x \in \mathbb{R}^n \mapsto V(x) := \inf_{u \in \mathcal{U}} \mathbb{E} \left[ \int_0^{+\infty} e^{-Ct} d_K^2(X_t^{x,u}) dt \right].$$

Indeed, the result of [7] says that  $K$  is viable if and only if the map  $x \mapsto d_K^2(x)$  is a viscosity supersolution to (1.3). The proofs relies on Ito's calculus and on comparison theorem for sub- and supersolutions to (1.3).

In the infinite dimensional case, we are facing several difficulties we wish to point out:

- the meaning of solutions to (1.1) (mild solutions see [12, 13]),
- the relation between  $K$  and the domain of the operator  $A$ ,
- the comparison result for a Hamilton–Jacobi equation is not valid in general (see [24, 30] for a particular case),
- the fact that the Brownian motion is cylindrical forbids a direct application of Ito's calculus,
- the lack of compactness and the consequent difficulty in proving existence of optimal controls for (1.4) (see [18, 19] for some results in the direction of existence of relaxed controls).

For the viability in Hilbert spaces, we mention the approach of [27] in the uncontrolled case and when both  $K$  and the coefficients of the state equation are regular and the Brownian motion is finite dimensional. This approach is based on a tangent characterization of the viability and on a Wong–Zakai approximation of the noise. We notice that, for such an approach, the above mentioned smoothness assumptions seem to be crucial; this precludes, for instance, applicability of the method if the coefficients of the equation are evaluation (Nemitskii) operators even if the noise is finite dimensional.

In contrast our method is concerned with the fully controlled case with a cylindrical Brownian motion. Namely, we propose a new criterion for  $\varepsilon$ -viability of controlled

systems extending results of [2, 3, 4, 7, 17] in both directions: the case where the state variable lies in a Hilbert space, and the case where the Brownian motion is infinite dimensional (cylindrical).

Let us explain how this paper is organized.

In section 1, we set the problem recalling the notion of mild solution and giving hypothesis on the control system.

In section 2, we provide a necessary condition for  $\varepsilon$  viability for a closed convex set  $K$ . This condition says that the function  $x \mapsto d_K^2(x)$  is a viscosity supersolution of a suitable Hamilton–Jacobi–Bellman equation. This relies on two different classes of modification of the original state equation (beside usual approximations obtained replacing unbounded operator  $A$  by its Yosida approximations):

(a) the first modified equation is the stochastic differential equation, which is the equation satisfied by  $\mathbb{E}[X_t^{x,u} | \mathcal{F}_\infty^m]$ , where  $\mathcal{F}_t^m = \sigma\{W_s^m, s \leq t\} \vee \mathcal{N}_P$  is the filtration generated by the projection  $W^m$  of the cylindrical Brownian motion on the  $m$  dimensional space spanned by the  $m$  first elements of an orthonormal base of  $\Xi$ .

(b) the second equation is obtained by replacing  $W$  by  $W^m$  in the original (1.1).

The idea of introducing the previous modified equations is due to the following easy observation: If  $X_t^{x,u}$  is viable in a convex set  $K$ , so is  $\mathbb{E}[X_t^{x,u} | \mathcal{F}_\infty^m]$ . It is worth pointing out that modified equations (a) and (b) are not needed when  $W$  is trace class: in this case we are able to consider (possibly) nonconvex closed sets.

Then we explain how our general necessary condition can be simplified for more smooth sets as linear spaces, balls.

Section 3 is devoted to cases when the necessary condition of section 2 is also sufficient. This can be obtained for instance when  $K$  is a closed ball. We also prove that when  $K$  is a locally compact linear subset then necessarily  $K$  is contained in the domain of the operator  $A$  (we give an alternative proof of the result in Nakayama [27]). In this last case we again prove that the necessary condition of section 2 is also sufficient. Finally, we consider the special case in which the noise is trace class and the diffusion has the restrictive regularity properties required in [24]. In that case we are able to prove that the necessary condition is also sufficient for all sets  $K$  (not necessarily convex) such that the distance  $d_K^2$  is lower semicontinuous with respect to weak topology (this is, for instance, the case when  $K$  is locally compact). The result is obtained by a comparison principle for viscosity solutions with an argument similar to the finite dimensional case; see [30, 24].

Section 4 is concerned with the application of our results to a semilinear heat equation. The characterization of the  $\varepsilon$ -viability is deduced for simple sets like balls and locally compact smooth sets.

**2. Statement of the problem.** We will be dealing with the following state equation:

$$(2.1) \quad \begin{cases} dX_t^{x,u} &= (AX_t^{x,u} + F(X_t^{x,u}, u_t)) dt + G(X_t^{x,u}, u_t) dW_t, \quad t \geq 0, \\ X_0^{x,u} &= x. \end{cases}$$

We work under the following general assumptions and notations:

- $(H, \langle \cdot, \cdot \rangle)$ ,  $(\Xi, \langle \cdot, \cdot \rangle_\Xi)$  are separable real Hilbert spaces. By  $L(\Xi, H)$  we denote the space of bounded linear operators  $\Xi \rightarrow H$ . Moreover, by  $L_2(\Xi, H)$  we denote the subspace of  $L(\Xi, H)$  given by all Hilbert–Schmidt operators. We endow  $L_2(\Xi, H)$  with its natural Hilbertian norm.
- If  $\phi$  is a Fréchet differentiable map  $H \rightarrow \mathbb{R}$  by  $D\phi$ , then we always denote its Fréchet derivative.

- $(\Omega, \mathcal{F}, P)$  is a complete probability space and  $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$  is a filtration defined on  $(\Omega, \mathcal{F}, P)$  that we assume to be complete and right-continuous.
- $W$  is  $\Xi$  valued cylindrical Wiener process with respect to the filtration  $\mathbb{F}$ .
- $A$  is an unbounded linear operator  $A : H \supset D(A) \rightarrow H$  and we assume that  $A$  is an  $m$ -dissipative operator (that is,  $\langle Ax, x \rangle \leq 0$ , for all  $x \in D(A)$ , and  $(I - A)$  is invertible). Consequently,  $A$  is the generator of a  $C_0$  semigroup of contractions  $(T_t)_{t \geq 0}$ .
- $E$  is a real separable and reflexive Banach space and  $U \subset E$  is a bounded closed subset. The space  $U$  is endowed with the Borel  $\sigma$ -field  $\mathcal{B}(U)$  and  $\mathcal{U} = L^0_{\mathbb{F}}(\mathbb{R}_+; U)$  is the set of all  $\mathbb{F}$ -progressively measurable processes  $u : [0, \infty[ \times \Omega \rightarrow U$ .
- $F : H \times U \rightarrow H$ ,  $G : H \times U \rightarrow L(\Xi, H)$  satisfy, for suitable  $C \in \mathbb{R}_+$  and  $\gamma \in (0, 1/2)$  and for all  $t > 0, u \in U, x, y \in H$ ,

$$(2.2) \quad |F(x, u) - F(y, u)| \leq C|x - y|, \quad |F(x, u)| \leq C, \quad |G(x, u)|_{L(\Xi, H)} \leq C,$$

$$(2.3) \quad |T_t G(x, u) - T_t G(y, u)|_{L_2(\Xi, H)} \leq C(1 \wedge t)^{-\gamma}|x - y|,$$

$$(2.4) \quad |T_t G(x, u)|_{L_2(\Xi, H)} \leq C(1 \wedge t)^{-\gamma}.$$

- For all  $x \in H$  and  $t > 0$ ,  $F(x, \cdot) : U \rightarrow H$  and  $T_t G(x, \cdot) : U \rightarrow L_2(\Xi, H)$  are continuous.
- There exists an orthonormal basis  $\{e_i, i \in \mathbb{N}\}$  in  $\Xi$  such that, for all  $i \in \mathbb{N}$  and all  $u \in U, x, y \in H$ ,

$$(2.5) \quad |G(x, u)e_i - G(y, u)e_i| \leq C_i|x - y|,$$

where  $C_i \in \mathbb{R}$  are suitable constants. The basis  $\{e_i, i \in \mathbb{N}\}$  will be fixed throughout this paper.

REMARK 2.1. *We could relax the assumption on dissipativity of  $A$  just requiring that  $A - \lambda I$  is  $m$ -dissipative for a suitable constant  $\lambda$ . This would only slightly complicate the computations.*

We recall that an  $\mathcal{F}$ -progressively measurable process  $X^{x,u}$  such that  $\mathbb{E}[\sup_{s \in [0, T]} |X_s^{x,u}|^2] < +\infty$  is a mild solution of (2.1) if  $P$ -a.s. for all  $t > 0$ :

$$(2.6) \quad X_t^{x,u} = T_t x + \int_0^t T_{t-s} F(X_s^{x,u}, u_s) ds + \int_0^t T_{t-s} G(X_s^{x,u}, u_s) dW_s.$$

The following existence and uniqueness result is standard under the present assumptions (see [12, 16]). The convergence result follows by a straightforward application of a parameter depending contraction argument exactly as in the proof of Proposition 3.2 in [16].

LEMMA 2.2. *For all  $x \in H$  and  $u \in \mathcal{U}$  there exists a unique mild solution  $X^{x,u}$  of (2.1). Moreover, for every  $p \geq 1$  we can find a constant  $C_p \in \mathbb{R}$  such that, for all  $x \in H, u \in \mathcal{U}, T > 0$ ,*

$$\mathbb{E} \sup_{s \in [0, T]} |X_s^{x,u}|^p \leq \exp(C_p T)(1 + |x|^p).$$

Finally, if  $\{u^n : n \geq 1\} \subset \mathcal{U}$  and  $u \in \mathcal{U}$  are such that  $u^n \rightarrow u$  in measure  $ds dP$ , then

$$\mathbb{E} \sup_{s \in [0, T]} |X_s^{x,u^n} - X_s^{x,u}|^p \rightarrow 0.$$

Our objective is to find conditions on  $A, F, G$  under which  $K$  is  $\varepsilon$ -viable for the above stochastic differential equation (SDE).

Let us now recall the definition of  $(\varepsilon)$ -viability. We use the notation in  $d_K(x) := \inf_{y \in K} |x - y|$ ,  $x \in H$ .

DEFINITION 2.3. *A closed subset  $K \subset H$  is  $\varepsilon$ -viable if there exists a constant  $C$  large enough such that, for  $x \in K$ ,*

$$(2.7) \quad \inf_{u \in \mathcal{U}} \mathbb{E} \left[ \int_0^{+\infty} e^{-Ct} d_K^2(X_t^{x,u}) dt \right] = 0.$$

Obviously, if  $K$  is viable for (1.1), then it is also  $\varepsilon$ -viable. The converse is true if the infimum is achieved in (2.7) for every  $x \in K$ . In the infinite dimensional case there are not good conditions on the dynamics ensuring the existence of an optimal control in (2.7); we have only the existence of  $\varepsilon$ -optimal controls. In contrast, in the finite dimensional case ( $H = \mathbb{R}^n$  and  $A \equiv 0$ ), the existence of such an optimal control could be obtained assuming conditions on the right-hand side of the control system (see [15]) and it is then possible to study the viability instead of  $\varepsilon$ -viability [7].

We recall now the following characterization result in finite dimensional space.

THEOREM 2.4 (see [7]). *Suppose that the spaces  $H$  and  $\Xi$  are finite dimensional and  $A \equiv 0$ .*

*Then the closed nonempty set  $K \subset H$  is  $\varepsilon$ -viable w.r.t. (2.1) if and only if there exists  $C > 0$  (large enough) such that  $d_K^2$  is a viscosity supersolution of the following HJB-equation:*

$$(2.8) \quad \Psi(x, V(x), DV(x), D^2V(x)) = 0, \quad x \in H,$$

where

$$(2.9) \quad \Psi(x, v, p, X) = - \inf_{u \in U} \left( \frac{1}{2} \text{Tr} \langle XG(x, u)G(x, u)^* \rangle + \langle F(x, u), p \rangle \right) + Cv - d_K^2(x).$$

In Hilbert spaces, our method will use two different classes of modifications of the state equation which we introduce now.

For the orthonormal basis  $\{(e_i) : i \in \mathbb{N}\}$  in  $\Xi$  fixed above we put  $\beta_t^i := \langle W_t, e_i \rangle$ . By construction  $\{(\beta_i) : i \in \mathbb{N}\}$  is a sequence of mutually independent 1-dimensional Brownian motions and  $W_s^m := \sum_{i=1}^m \beta_t^i e_i$  is an  $m$ -dimensional Brownian motion.

We set  $\mathcal{F}_t^m = \sigma\{W_s^m, 0 \leq s \leq t\} \vee \mathcal{N}_P$ ,  $\mathcal{F}_t^{m,M} = \sigma\{\beta_s^i, 0 \leq s \leq t, m+1 \leq i \leq M\} \vee \mathcal{N}_P$ ,  $t \in [0, +\infty)$  (here  $\mathcal{N}_P$  is the set of all  $P$ -null sets). Moreover,  $\mathcal{F}_\infty^m = \sigma(\bigcup_{t>0} \mathcal{F}_t^m)$ ,  $\mathcal{F}_\infty^{m,M} = \sigma(\bigcup_{t>0} \mathcal{F}_t^{m,M})$ . Finally, by  $\mathbb{F}^m = (\mathcal{F}_t^m)_{t \geq 0}$ ,  $\mathbb{F}^{m,M} = (\mathcal{F}_t^{m,M})_{t \geq 0}$  we denote the corresponding filtrations.

We now define  $Y_t^{x,u,m} := \mathbb{E}(X_t^{x,u} | \mathcal{F}_\infty^m)$ . It is easy to verify that  $Y_t^{x,u,m}$  satisfies

$$(2.10) \quad Y_t^{x,u,m} = T_t x + \int_0^t T_{t-s} \mathbb{E}(F(X_s^{x,u}, u_s) | \mathcal{F}_\infty^m) ds + \int_0^t T_{t-s} \mathbb{E}(G(X_s^{x,u}, u_s) | \mathcal{F}_\infty^m) dW_s^m$$

or, writing formally, we get

$$(2.11) \quad \begin{cases} dY_t^{x,u,m} = AY_t^{x,u,m} dt + \mathbb{E}(F(X_t^{x,u}, u_t) | \mathcal{F}_\infty^m) dt + \mathbb{E}(G(X_t^{x,u}, u_t) | \mathcal{F}_\infty^m) dW_t^m, & t \geq 0, \\ Y_0^{x,u,m} = x. \end{cases}$$

REMARK 2.5. *The idea related to the introduction of the processes  $Y^{x,u,m}$  is the following. Since  $K$  is convex if  $P\{X^{x,u} \in K\} = 1$ , then  $P\{Y^{x,u,m} \in K\} = 1$  for all  $m \in \mathbb{N}$ .*

We shall also consider the equation obtained from (2.1) by replacing  $W$  by its projections  $W^m$ , namely

$$(2.12) \quad \begin{cases} dX_t^{x,u,m} &= AX_t^{x,u,m} dt + F(X_t^{x,u,m}, u_t) dt + G(X_t^{x,u,m}, u_t) dW_t^m, \quad t \geq 0, \\ X_0^{x,u,m} &= x. \end{cases}$$

As in the case of the original (2.1) we have the following lemma.

LEMMA 2.6. *For all  $x \in H$ ,  $m \in \mathbb{N}$ , and  $u \in \mathcal{U}$  there exists a unique mild solution  $X^{x,u,m}$  of (2.12). Moreover, for all  $p \geq 1$  we can find a constant  $C_p \in \mathbb{R}$  such that, for all  $x \in H$ ,  $m \in \mathbb{N}$ ,  $u \in \mathcal{U}$ , and  $T > 0$ ,*

$$\mathbb{E} \sup_{s \in [0, T]} |X_s^{x,u,m}|^p \leq \exp(C_p T) (1 + |x|^p).$$

Finally, if  $\{u^n : n \geq 1\} \subset \mathcal{U}$  and  $u \in \mathcal{U}$  are such that  $u^n \rightarrow u$  in measure  $ds dP$ , then  $ds dP$  then

$$\mathbb{E} \sup_{s \in [0, T]} |X_s^{x,u^n,m} - X_s^{x,u,m}|^p \rightarrow 0.$$

REMARK 2.7. *Observe that if  $u \in \mathcal{U}$  is progressively measurable with respect to  $\mathcal{F}^m$ , then  $X^{x,u,m}$  is progressively measurable with respect to  $\mathcal{F}^m$  as well.*

**3. Necessary conditions.** We will prove our result under two different sets of assumptions. In the second we require much less on the coefficients but a bit more on the semigroup generated by  $A$ .

ASSUMPTION 3.1. For all  $x \in H$  the mapping  $u \rightarrow (F(x, u), G(x, u)) : U \rightarrow H \times L(\Xi, H)$  is the restriction of an affine map  $U \rightarrow H \times L(\Xi, H)$ .

ASSUMPTION 3.2. For all  $x \in H$ , the set  $\{(F(x, u), G(x, u)) : u \in U\}$  is a convex subset of  $H \times L(\Xi, H)$  and, moreover, the semigroup  $\{T_t : t \geq 0\}$  is analytic.

REMARK 3.3. *Clearly, if Assumption 3.1 holds,  $U$  is convex and the semigroup  $\{T_t : t \geq 0\}$  is analytic, then Assumption 3.2 holds.*

Our necessary condition will be formulated in terms of viscosity supersolutions of a suitable Hamilton–Jacobi–Bellman (HJB) type equation. The following definition is taken from [30].

DEFINITION 3.4. Let  $\Sigma(H) := \{L \in L(H) : L \text{ selfadjoint}\}$  and  $\Psi : H \times \mathbb{R} \times H \times \Sigma(H) \rightarrow \mathbb{R}$  be a continuous mapping. Just to fix ideas we also assume that  $\Psi(x, y, p, X) \leq \Psi(x, y, p, Y)$  for all  $x, p \in H$ ,  $y \in \mathbb{R}$ ,  $X, Y \in \Sigma(H)$  with  $X \geq Y$ . A continuous function  $V : H \rightarrow \mathbb{R}$  is called a viscosity supersolution of the partial differential equation (PDE)

$$(3.1) \quad \Psi(x, V(x), DV(x), D^2V(x)) - \langle Ax, DV(x) \rangle = 0, \quad x \in H,$$

if for all  $x \in H$  and  $\varphi \in C^2(H)$  such that

1.  $\varphi(x) = V(x)$ ,  $\varphi(y) \leq V(y)$ ,  $y \in \overline{B_r(x)}$ , for some  $r > 0$ ;
2.  $\varphi(y) = \varphi_0(y) + g(y)$ ,  $y \in H$ , where  $\varphi_0, g \in C^2(H)$  are supposed to be bounded on bounded subsets of  $H$  together with their first order derivatives and to

verify

- (i)  $A^*D\varphi_0 : H \rightarrow H$  is continuous and bounded on bounded sets,
- (ii)  $g(y) = \hat{g}(|y|)$ ,  $y \in H$ , where  $\hat{g} : \mathbb{R} \rightarrow \mathbb{R}$  is a decreasing function,

it holds that

$$\Psi(x, V(x), D\varphi(x), D^2\varphi(x)) - \langle A^*D\varphi_0(x), x \rangle \geq 0.$$

We are now able to formulate the main result of this section.

**THEOREM 3.5.** *Suppose that either Assumption 3.1 or Assumption 3.2 holds and that the closed convex nonempty set  $K \subset H$  is  $\varepsilon$ -viable w.r.t. (2.1). Then there exists  $C > 0$  (large enough) such that for every  $m \geq 1$ ,  $d_K^2$  is a viscosity supersolution of the HJB equation*

$$(3.2) \quad \Psi_m(x, V(x), DV(x), D^2V(x)) - \langle Ax, DV(x) \rangle = 0, \quad x \in H,$$

where for  $X \in \Sigma(H)$ ,  $p \in H$ ,  $v \in \mathbb{R}$ ,  $x \in H$

(3.3)

$$\Psi_m(x, v, p, X) = - \inf_{u \in U} \left( \frac{1}{2} \sum_{i=1}^m \langle XG(x, u)e_i, G(x, u)e_i \rangle + \langle F(x, u), p \rangle \right) + Cv - d_K^2(x).$$

The proof will be based on the following lemma comparing the solution of (2.11) and the solution of (2.12). In order to make this comparison it is crucial to associate for  $m \geq 1$  and  $\delta > 0$  an appropriately chosen  $\delta$ -optimal,  $\mathbb{F}$ -progressively measurable control process  $u$  with an  $\mathbb{F}^m$ -progressively measurable control process  $v$  such that the estimate (3.5) holds.

**LEMMA 3.6.** *Suppose that either Assumption 3.1 or Assumption 3.2 holds and that  $K$  is  $\varepsilon$ -viable. Then there exist two large enough constants  $C, c > 0$  such that for all  $x \in H$ ,  $y \in K$ ,  $m \geq 1$ , and  $\delta > 0$  we can find an  $\mathbb{F}$ -progressively measurable control process  $u \in L_{\mathbb{F}}^0([0, +\infty]; U)$  and an  $\mathbb{F}^m$ -progressively measurable control process  $v \in L_{\mathbb{F}^m}^0([0, 1]; U)$  verifying*

$$(3.4) \quad \mathbb{E} \int_0^{+\infty} e^{-Cs} d_K^2(X_s^{y,u}) ds \leq \delta$$

and

$$(3.5) \quad \mathbb{E} |X_\tau^{x,v,m} - Y_\tau^{y,u,m}|^2 - |x - y|^2 \leq ct|x - y|^2 + ct(t^{1-2\gamma} \vee t^{1/2})(1 + |x|^2 + |y|^2)$$

for all  $t \in ]0, 1]$  and all  $\mathbb{F}^m$ -stopping times  $\tau \leq t$ . We stress that constants  $c$  and  $C$  depend only on the data of the problem and on  $m$  but not on  $x$  and  $y$ .

*Proof.* We start by proving the claim in the case in which Assumption 3.2 holds since it is by far the most difficult. We will comment on the case with Assumption 3.1 at the end of this proof.

Consider for  $M \in \mathbb{N}$ ,  $M > m$  the class  $\mathcal{U}^{m,M}$  given by all processes  $u \in \mathcal{U}$  which are stepwise constant in  $[0, 1]$ , and that verify

$$u|_{[0,1]} = \sum_{i=0}^{N-1} \eta_i I_{(t_i, t_{i+1}]}, \quad 0 = t_0 < \dots < t_N = 1, \quad \eta_i \in \mathcal{L}_{t_i}^{m,M}, \quad 1 \leq i \leq N-1, \quad \eta_0 \in U,$$

where  $\mathcal{L}_t^{m,M}$  denotes the family of  $\mathcal{F}_t$ -measurable simple random variables  $\eta$  of the

form  $\eta = \sum_{i,j=1}^L \alpha_{i,j} I_{A_i \cap B_j}$ , with  $\alpha_{i,j} \in U$ ,  $A_i \in \mathcal{F}_t^m$ ,  $B_j \in \mathcal{F}_t^{m,M}$ ,  $1 \leq i, j \leq L$ , s.t.  $\sum_{i,j=1}^L I_{A_i \cap B_j} = 1$ .

Since, as it can be easily proved,  $\bigcup_{M>m} \mathcal{U}^{m,M}$  is dense in  $\mathcal{U}$  endowed with the norm  $|u|_{\mathcal{U}} = \mathbb{E}[\int_0^{+\infty} e^{-Ct} |u_t| dt]$  and  $K$  is supposed to be  $\varepsilon$ -viable with respect to (2.1), there exist  $M > m$  and  $u \in \mathcal{U}^{m,M}$  such that (3.4) holds.

Fixing  $m \in \mathbb{N}$  and  $u \in \mathcal{U}^{m,M}$  that verifies (3.4) we seek a  $\mathbb{F}^m$ -progressively measurable control process  $v \in L_{\mathbb{F}^m}^0([0, T]; U)$  s.t. (3.5) holds. This will be done in several steps.

*Step 1.* For all  $u \in U$  and  $x \in H$ , let  $G^m(x, u) := (G(x, u)e_1, \dots, G(x, u)e_m) \in H^m$  and  $(G^m, F)(x, u) := (G^m(x, u), F(x, u)) \in H^{m+1}$ . We consider  $\eta \in \mathcal{L}_t^{m,M}$  of the above form. Then

$$\mathbb{E}[(G^m, F)(x, \eta) | \mathcal{F}_t^m] = \sum_{i=1}^L I_{A_i} \left( \sum_{j=1}^L (G^m, F)(x, \alpha_{i,j}) P(B_j) \right).$$

Since  $\sum_{j=1}^L \mathbb{P}(B_j) = 1$  the convexity of  $\{(G^m, F)(x, u) : u \in U\}$  (assumed in Assumption 3.2) gives  $\sum_{j=1}^L (G^m, F)(x, \alpha_{i,j}) P(B_j) \in \{(G^m, F)(x, u) : u \in U\}$ .

From the Jankov-von Neumann measurable selection theorem (see Proposition 7.49 in [6]), there exists a universally measurable mapping  $\ell_i : (H, \mathcal{B}^u(H)) \rightarrow (U, \mathcal{B}(U))$  s.t.

$$\sum_{j=1}^L (G^m, F)(x, \alpha_{i,j}) P(B_j) = (G, F)(x, \ell_i(x)) \quad \text{for all } x \in H.$$

Then, given  $\zeta \in L^2(\Omega, \mathcal{F}_t^m, P; H)$  we put  $\Phi_t^\eta(\zeta) := \sum_{i=1}^L \ell_i(\zeta) I_{A_i}$ .

We notice that  $\Phi_t^\eta(\zeta) \in L^0(\Omega, \mathcal{F}_t^m, P; U)$ . Indeed, since  $\ell_i$  is  $\mathcal{B}^u(H) - \mathcal{B}(U)$ -measurable, it coincides  $P\zeta$ -a.s. with some  $\mathcal{B}(H) - \mathcal{B}(U)$ -measurable mapping  $\theta_i : H \rightarrow U$ . Consequently,  $\ell_i(\zeta) = \theta_i(\zeta)$ ,  $P$ -a.s. But since  $\mathcal{F}_t^m$  contains all  $P$ -null sets, the  $\mathcal{F}_t^m - \mathcal{B}(U)$ -measurability of  $\theta_i(\zeta)$  implies that of  $\ell_i(\zeta)$ .

Moreover, by construction  $\mathbb{E}[(G^m, F)(\zeta, \eta) | \mathcal{F}_\infty^m] = (G^m, F)(\zeta, \Phi_t^\eta(\zeta))$ ,  $P$ -a.s.

*Step 2.* We construct now  $v$  proceeding by iteration. We recall that we have fixed  $u \in \mathcal{U}^{m,M}$ ,  $u = \sum_{i=0}^{N-1} \eta_i I_{(t_i, t_{i+1}]}$ ,  $0 = t_0 < \dots < t_N = 1$ ,  $\eta_i \in \mathcal{L}_{t_i}^{m,M}$ ,  $1 \leq i \leq N-1$ , and  $\eta_0 \in U$  deterministic.

We define, for  $s \in [0, t_1]$ ,  $v_s := u_s = \eta_0$ . Thus  $v$  is deterministic in  $[0, t_1]$  and  $X_s^{x,v,m}$  is  $\mathcal{F}_s^m$ -measurable for all  $s \in [0, t_1]$ .

Then, for  $s \in (t_1, t_2]$ , let  $v_s := \Phi_{t_1}^{\eta_1}(X_{t_1}^{x,v,m})$ . Note that  $(v_s, s \in [0, t_2]) \in L_{\mathbb{F}^m}^0([0, t_2]; U)$  and  $X^{x,v,m} \in L_{\mathbb{F}^m}^2([0, t_2]; H)$ .

We proceed by iteration letting, for  $0 \leq i \leq N-1$  and  $s$  in  $(t_i, t_{i+1}]$ ,  $v_s := \Phi_{t_i}^{\eta_i}(X_{t_i}^{x,v,m})$ . Note that  $v \in L_{\mathbb{F}^m}^0([0, 1]; U)$ ,  $X^{x,v,m} \in L_{\mathbb{F}^m}^2([0, 1]; H)$  and

$$\mathbb{E}[(G^m, F)(X_{t_i}^{x,v,m}, u_s) | \mathcal{F}_\infty^m] = (G^m, F)(X_{t_i}^{x,v,m}, v_s), \quad s \in (t_i, t_{i+1}], \quad 1 \leq i \leq N-1.$$

*Step 3.* We have now to deduce an estimation similar to (3.6) but for  $t$  different from  $t_i$ . This will be done by exploiting the analyticity of the semigroup  $(T_t)_{t \geq 0}$ . To shorten the notation, let  $\hat{X}_t := X_t^{x,v,m}$ ,  $t \in [0, 1]$ .

For  $\alpha \in (0, 1/2)$  we put  $|x|_\alpha^2 = |x|^2 + |(I - A)^\alpha x|^2$ ,  $x \in H$ . Standard estimates yield

$$\mathbb{E}|\hat{X}_s|_\alpha^2 \leq C_\alpha^2 \left( 1 + \left( 1 + \frac{1}{s^{2\alpha}} \right) |y|^2 \right), \quad s \in (0, 1].$$

Moreover, for  $s \in [t_i, t_{i+1}]$ ,

$$\widehat{X}_s = T_{s-t_i} \widehat{X}_{t_i} + \int_{t_i}^s T_{r-t_i} F(\widehat{X}_r, v_r) dr + \int_{t_i}^s T_{r-t_i} G^m(\widehat{X}_r, v_r) dW_r^m,$$

thus

$$\mathbb{E} \left| \widehat{X}_s - T_{s-t_i} \widehat{X}_{t_i} \right|^2 \leq C_m (s - t_i) (1 + |y|^2), \quad s \in (t_i, t_{i+1}].$$

On the other hand,  $|T_t x - x| \leq C_\alpha t^\alpha |x|_\alpha$  and, consequently, for  $c$  depending on  $\alpha$

$$\mathbb{E} \left[ \left| T_{s-t_i} \widehat{X}_{t_i} - \widehat{X}_{t_i} \right|^2 \right] \leq c (s - t_i)^{2\alpha} \mathbb{E} \left[ |\widehat{X}_{t_i}|_\alpha^2 \right] \leq c \left( 1 + \left( 1 + \frac{1}{t_i^{2\alpha}} \right) |x|^2 \right) (s - t_i)^{2\alpha}.$$

This implies for  $\alpha = 1/4$

$$\mathbb{E} \left[ \left| \widehat{X}_s - \widehat{X}_{t_i} \right|^2 \right] \leq c \left( 1 + \left( 1 + \frac{1}{\sqrt{t_i}} \right) |x|^2 \right) (s - t_i)^{1/2}.$$

Hence, for  $s \in (t_i, t_{i+1}]$ ,  $1 \leq i \leq N - 1$ ,

$$\begin{aligned} & \mathbb{E} \left| \mathbb{E} \left[ (G^m, F) \left( \widehat{X}_s, u_s \right) | \mathcal{F}_\infty^m \right] - (G^m, F) \left( \widehat{X}_s, v_s \right) \right|_{H^{m+1}}^2 \\ & \leq 3 \mathbb{E} \left| \mathbb{E} \left[ (G^m, F) \left( \widehat{X}_s, u_{t_{i+1}} \right) - (G^m, F) \left( \widehat{X}_{t_i}, u_{t_{i+1}} \right) | \mathcal{F}_\infty^m \right] \right|_{H^{m+1}}^2 \\ & \quad + 3 \mathbb{E} \left| \mathbb{E} \left[ (G^m, F) \left( \widehat{X}_{t_i}, u_{t_{i+1}} \right) | \mathcal{F}_\infty^m \right] - (G^m, F) \left( \widehat{X}_{t_i}, v_{t_{i+1}} \right) \right|_{H^{m+1}}^2 (= 0) \\ & \quad + 3 \mathbb{E} \left| (G^m, F) \left( \widehat{X}_{t_i}, v_{t_{i+1}} \right) - (G^m, F) \left( \widehat{X}_s, v_{t_{i+1}} \right) \right|_{H^{m+1}}^2 \\ & \leq c \mathbb{E} |\widehat{X}_s - \widehat{X}_{t_i}|^2 \leq c \left( 1 + \left( 1 + \frac{1}{\sqrt{t_i}} \right) |x|^2 \right) (s - t_i)^{1/2}. \end{aligned}$$

We stress the fact that if  $s \in [0, t_1]$ , then  $\mathbb{E}[(G^m, F)(\widehat{X}_s, u_s) | \mathcal{F}_\infty^m] = (G^m, F)(\widehat{X}_s, v_s)$ .

Consequently, for all  $1 \leq i \leq N - 1$ ,  $s \in (t_i, t_{i+1}]$

$$(3.7) \quad \begin{aligned} & \mathbb{E} \int_{t_i}^s \left| \mathbb{E} \left[ (G^m, F) \left( \widehat{X}_r, u_r \right) | \mathcal{F}_\infty^m \right] - (G^m, F) \left( \widehat{X}_r, v_r \right) \right|_{H^{m+1}}^2 dr \\ & \leq c \left( 1 + \left( 1 + \frac{1}{\sqrt{t_i}} \right) |x|^2 \right) (s - t_i)^{3/2}, \end{aligned}$$

while the left-hand side is zero for  $i = 0$ .

*Step 4.* We can now conclude the proof. Recall that to shorten notation we have denoted  $\widehat{X}_t = X_t^{x, v, m}$ . For the same reason we let  $\widehat{Y}_t := Y_t^{y, u, m}$ . Notice that

$$(3.8) \quad \begin{aligned} \widehat{X}_t - \widehat{Y}_t &= T_t(x - y) + \int_0^t T_{t-s} \left( F(\widehat{X}_s, v_s) - \mathbb{E}[F(X_s^{y, u}, u_s) | \mathcal{F}_\infty^m] \right) ds \\ & \quad + \int_0^t T_{t-s} \left( G(\widehat{X}_s, v_s) - \mathbb{E}[G(X_s^{y, u}, u_s) | \mathcal{F}_\infty^m] \right) dW_s^m. \end{aligned}$$

Applying Itô's rule to  $|n(nI - A)^{-1}(\widehat{X}_t - \widehat{Y}_t)|^2$ , and taking into account the dissipativity



of  $A$ , we get that, for every  $\mathbb{F}$ -stopping time  $\tau \leq t(\leq T)$ ,

$$\begin{aligned} & \mathbb{E}|n(nI - A)^{-1}(\widehat{X}_\tau - \widehat{Y}_\tau)|^2 - |n(nI - A)^{-1}(x - y)|^2 \\ & \leq 2\mathbb{E} \int_0^t \left| \left\langle n(nI - A)^{-1} \left( F(\widehat{X}_s, v_s) - \mathbb{E}[F(X_s^{y,u}, u_s) | \mathcal{F}_\infty^m] \right), \right. \right. \\ & \quad \left. \left. n(nI - A)^{-1} (\widehat{X}_s - \widehat{Y}_s) \right\rangle \right| ds \\ & \quad + \sum_{i=1}^m \mathbb{E} \int_0^t \left| n(nI - A)^{-1} \left( G(\widehat{X}_s, v_s) e_i - \mathbb{E}[G(X_s^{y,u}, u_s) e_i | \mathcal{F}_\infty^m] \right) \right|^2 ds, \end{aligned}$$

and letting  $n \rightarrow \infty$ ,

$$\begin{aligned} & \mathbb{E}|\widehat{X}_\tau - \widehat{Y}_\tau|^2 - |x - y|^2 \\ & \leq \mathbb{E} \int_0^t \left| F(\widehat{X}_s, v_s) - \mathbb{E}[F(X_s^{y,u}, u_s) | \mathcal{F}_\infty^m] \right|^2 ds + \mathbb{E} \int_0^t |\widehat{X}_s - \widehat{Y}_s|^2 ds \\ & \quad + \sum_{j=1}^m \mathbb{E} \int_0^t \left| G(\widehat{X}_s, v_s) e_j - \mathbb{E}[G(X_s^{y,u}, u_s) e_j | \mathcal{F}_\infty^m] \right|^2 ds \\ & \leq \mathbb{E} \int_0^t |\widehat{X}_s - \widehat{Y}_s|^2 ds + \mathbb{E} \int_0^t \left| (G^m, F)(\widehat{X}_s, v_s) - \mathbb{E}[(G^m, F)(X_s^{y,u}, u_s) | \mathcal{F}_\infty^m] \right|^2 ds \\ & \leq c_m \mathbb{E} \int_0^t |\widehat{X}_s - \widehat{Y}_s|^2 ds + c_m \sum_{i=1}^{N-1} \left( 1 + \left( 1 + \frac{1}{\sqrt{t_i}} \right) |x|^2 \right) (t_{i+1} \wedge t - t_i \wedge t)^{3/2} \\ & \quad + 2\mathbb{E} \int_0^t \left| (G^m, F)(\widehat{X}_s, u_s) - (G^m, F)(X_s^{y,u}, u_s) \right|^2 ds \end{aligned}$$

for a suitable constant  $c_m$  (from now to the end of the proof its value can change from line to line). Let us come back to (3.8). Since  $F$  and  $G$  are bounded, an application of the standard factorization argument (see [12]) yields

$$\mathbb{E} \sup_{s \in [0, t]} |\widehat{X}_s - X_s^{y,u}|^2 + \mathbb{E} \sup_{s \in [0, t]} |\widehat{X}_s - \widehat{Y}_s|^2 \leq c_m(|x - y|^2 + t^{1-2\gamma}(1 + |x|^2 + |y|^2)),$$

and plugging this into the above equation we obtain

$$\begin{aligned} \mathbb{E}|\widehat{X}_\tau - \widehat{Y}_\tau|^2 - |x - y|^2 & \leq c_m t |x - y|^2 + c_m t^{2-2\gamma}(1 + |x|^2 + |y|^2) \\ & \quad + c_m \sum_{i=1}^{N-1} \left( 1 + \left( 1 + t_i^{-1/2} \right) |x|^2 \right) (t_{i+1} \wedge t - t_i \wedge t)^{3/2}. \end{aligned}$$

Without loss of generality, we can suppose that  $t_{i+1} - t_i \leq t_i^2$ ,  $i = 1, \dots, N-1$  (notice that  $t_i \geq t_1 > 0$  and that we need to add to the partition only points larger than  $t_1$ ; this will only modify the value of  $N$  and the control  $v$  in  $[t_1, 1]$  but not  $t_1$  nor  $v|_{[0, t_1]}$  or  $u$ ). Then

$$\begin{aligned} & \sum_{i=1}^{N-1} \left( 1 + \left( 1 + t_i^{-1/2} \right) |x|^2 \right) (t_{i+1} \wedge t - t_i \wedge t)^{3/2} \\ & \leq \sup_{i=1, \dots, N-1} \left[ (1 + 2t_i^{-1/2} |x|^2) (t_{i+1} \wedge t - t_i \wedge t)^{1/2} \right] \sum_{i=1}^{N-1} (t_{i+1} \wedge t - t_i \wedge t) \\ & \leq 2(1 + |x|^2) \sqrt{t} \sum_{i=1}^{N-1} (t_{i+1} \wedge t - t_i \wedge t) \leq 2(1 + |x|^2) t^{3/2}, \end{aligned}$$

and this completes the proof in the case in which Assumption 3.2 holds.

Finally, let us briefly discuss the case in which Assumption 3.2 is replaced by Assumption 3.1. It turns out that if we set  $v_t = \mathbb{E}[u_t | \mathcal{F}_\infty^m]$ , then instead of (3.7) we get the stronger relation

$$\mathbb{E}[(G^m, F)(\hat{X}_r, u_r) | \mathcal{F}_\infty^m] = (G^m, F)(\hat{X}_r, v_r), \quad \mathbb{P} - \text{a.s. for all } r \in [0, 1].$$

This allows us to conclude as in step 4 previously (indeed, even in an easier way).  $\square$

We are now able to complete the proof that if  $K$  is  $\epsilon$ -viable, then  $d_K^2$  is a viscosity supersolution of (3.2).

*Proof of Theorem 3.5.* We fix arbitrarily  $x \in H$ ,  $m \in \mathbb{N}$ , and a test function  $\varphi \in C^2(H)$  verifying

1.  $\varphi(x) = d_K^2(x)$ ,  $\varphi(z) \leq d_K^2(z)$ ,  $z \in \overline{B_r(x)}$ , for some  $r > 0$ ;
2.  $\varphi(z) = \varphi_0(z) + g(z)$ ,  $z \in H$ , where  $\varphi_0, g \in C^2(H)$  are supposed to be bounded on bounded subsets of  $H$  together with their first order derivatives and to verify
  - (i)  $A^* D\varphi_0 : H \rightarrow H$  is continuous,
  - (ii)  $g(z) = \hat{g}(|z|)$ ,  $z \in H$ , where  $\hat{g} : \mathbb{R} \rightarrow \mathbb{R}$  is a decreasing function.

Let  $y := \pi_K(x) (\in K)$  and  $C > 0$  large enough. Moreover, for any  $\varepsilon > 0$  let  $u^\varepsilon$  and  $v^\varepsilon$  be given by Lemma 3.6 with  $\delta = \varepsilon^4$ .

From the convexity of  $K$  and, hence, that of  $d_K^2$  it follows that

$$\mathbb{E} \left[ \int_0^{+\infty} e^{-Ct} d_K^2 \left( Y_t^{y, u^\varepsilon, m} \right) dt \right] \leq \mathbb{E} \left[ \int_0^{+\infty} e^{-Ct} d_K^2 \left( X_t^{y, v^\varepsilon, m} \right) dt \right] \leq \varepsilon^4.$$

To shorten notation let  $\hat{Y}_t^\varepsilon := Y_t^{y, u^\varepsilon, m} (= \mathbb{E}[X_t^{y, u^\varepsilon} | \mathcal{F}_\infty^m])$ ,  $\hat{X}_t^\varepsilon := X_t^{y, v^\varepsilon, m}$ , and  $t \geq 0$ .

We observe that  $\hat{X}^\varepsilon, \hat{Y}^\varepsilon$  are  $\mathbb{F}^m$ -progressively measurable and define

$$\tau_\varepsilon := \inf\{t > 0 : |\hat{X}_t^\varepsilon - x| > \delta_1\} \wedge \inf\{t > 0 : |\hat{Y}_t^\varepsilon - y| > 1\}$$

for some  $\delta_1 \in (0, r)$  which will be specified later. Let us also remark that, for all  $t \geq 0$ ,

$$\begin{aligned} d_K \left( \hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon \right) &\leq \left| \hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon - \pi_K(\hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon) \right| \leq \left| \hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon - \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right| + d_K \left( \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right), \\ d_K(\hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon) &\leq \left| \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon - y \right| \leq 1, \end{aligned}$$

from where we get

$$\begin{aligned} d_K^2 \left( \hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon \right) &\leq \left| \hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon - \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right|^2 + d_K \left( \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right) \left( d_K \left( \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right) + 2 \left| \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon - \hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon \right| \right) \\ &\leq \left| \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon - \hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon \right|^2 + 2d_K \left( \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right) (2 + r + |y| + |x|) \\ &\leq \left| \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon - \hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon \right|^2 + C_{x, y, r} d_K \left( \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right). \end{aligned}$$

Hence, for all  $t \in [0, T]$ ,

$$\begin{aligned} \varphi(\hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon) - \varphi(x) &\leq d_K^2 \left( \hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon \right) - d_K^2(x) \\ &\leq \left( \left| \hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon - \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right|^2 - |x - y|^2 \right) + C_{x, y, r} d_K \left( \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right), \end{aligned}$$

and from Lemma 3.6,

(3.9)

$$\begin{aligned} E \left[ \varphi(\hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon) - \varphi(x) \right] &\leq \mathbb{E} \left[ \left| \hat{Y}_{\tau \wedge t}^\varepsilon - \hat{X}_{\tau \wedge t}^\varepsilon \right|^2 \right] - |x - y|^2 + C_{x,y,r} \mathbb{E} \left[ d_K \left( \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right) \right] \\ &\leq Ct|x - y|^2 + C_{x,y,r} \mathbb{E} \left[ d_K \left( \hat{Y}_{t \wedge \tau_\varepsilon}^\varepsilon \right) \right] + c_m(t^{2-2\gamma} \vee t^{3/2}). \end{aligned}$$

We claim now that for a suitable  $C_{x,y,\delta_1}$  that depends only on  $x, y, \delta_1$ , and on the data of the problem it holds that

$$P\{\tau_\varepsilon < \varepsilon\} \leq C_{x,y,\delta_1} \varepsilon^4 \quad \text{for all } \varepsilon > 0.$$

To prove the claim we recall that

$$\hat{Y}_t^\varepsilon = T_t y + \int_0^t T_{t-s} \mathbb{E}(F(X_s^{y,u^\varepsilon}, u_s^\varepsilon) | \mathcal{F}_\infty^m) ds + \int_0^t T_{t-s} \mathbb{E}(G(X_s^{y,u^\varepsilon}, u_s^\varepsilon) | \mathcal{F}_\infty^m) dW_s^m$$

and

$$\hat{X}_t^\varepsilon = T_t x + \int_0^t T_{t-s} F(\hat{X}_s^\varepsilon, v_s^\varepsilon) ds + \int_0^t T_{t-s} G(\hat{X}_s^\varepsilon, v_s^\varepsilon) dW_s^m.$$

By standard estimates based on factorization technique (see [12]) we find that for all  $p \geq 1$  there exists a constant  $c_{p,\gamma}$  depending on  $p, \gamma$  and on the data of the problem, but not on  $t$ , such that

$$\mathbb{E} \sup_{s \in [0,t]} \left[ |\hat{Y}_s^\varepsilon - T_s y|^{2p} + |\hat{X}_s^\varepsilon - T_s x|^{2p} \right] \leq c_{p,\gamma} t^{p(1-2\gamma)}.$$

Since the semigroup is strongly continuous, there exists  $s_0 > 0$  depending on  $x, y$ , and  $\delta_1$  such that  $|x - T_s x| \leq \delta_1/2$   $|y - T_s y| \leq 1/2$  for all  $s \in [0, s_0]$ . Thus for all  $\varepsilon \leq s_0$ ,

$$\begin{aligned} \mathbb{P}\{\tau_\varepsilon < \varepsilon\} &\leq \mathbb{P} \left\{ \sup_{s \leq \varepsilon} |\hat{X}_s^\varepsilon - x| \geq \delta_1 \right\} + \mathbb{P} \left\{ \sup_{s \leq \varepsilon} |\hat{Y}_s^\varepsilon - y| \geq 1 \right\} \\ &\leq \mathbb{P} \left\{ \sup_{s \leq \varepsilon} |\hat{X}_s^\varepsilon - T_s x| \geq \delta_1/2 \right\} + \mathbb{P} \left\{ \sup_{s \leq \varepsilon} |\hat{Y}_s^\varepsilon - T_s y| \geq 1/2 \right\} \\ &\leq c_{p,\gamma} 2^{2p} (\delta_1^{-2p} + 1) \varepsilon^{p(1-2\gamma)}, \end{aligned}$$

and the claim is proved choosing  $p$  large enough.

Coming back now to estimate (3.9) and taking into account that  $d_K(\hat{Y}_t^\varepsilon) \leq 1$  for  $0 \leq t \leq \tau_\varepsilon$ , we have

(3.10)

$$\begin{aligned} &\int_0^\varepsilon \mathbb{E} \left[ \varphi(\hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon) - \varphi(x) \right] dt \\ &\leq C\varepsilon^2 |x - y|^2 + C_{x,y,r} \mathbb{E} \left[ \int_0^\varepsilon d_K(\hat{Y}_t^\varepsilon) dt \right] + C_{x,y,r} \varepsilon P\{\tau_\varepsilon < \varepsilon\} + c_m(\varepsilon^{3-2\gamma} \vee \varepsilon^{5/2}) \\ &\leq C\varepsilon^2 |x - y|^2 + C_{x,y,r} e^{C\varepsilon} \mathbb{E} \left[ \int_0^\varepsilon e^{-Ct} d_K(\hat{Y}_t^\varepsilon) dt \right] + c_{m,x,y,\delta_1}(\varepsilon^{3-2\gamma} \vee \varepsilon^{5/2}) \\ &\leq C\varepsilon^2 |x - y|^2 + C_{x,y,r} e^{C\varepsilon} \varepsilon^{1/2} \left( \mathbb{E} \left[ \int_0^{+\infty} e^{-Ct} d_K^2(\hat{Y}_t^\varepsilon) dt \right] \right)^{1/2} \\ &\quad + c_{m,x,y,\delta_1}(\varepsilon^{3-2\gamma} \vee \varepsilon^{5/2}) \\ &\leq C\varepsilon^2 d_K^2(x) + c_{m,x,y,r,\delta_1}(\varepsilon^{3-2\gamma} \vee \varepsilon^{5/2}), \end{aligned}$$

where  $c_{m,x,y,r,\delta_1}$  is a suitable constant depending on the indicated parameters.

We need now to get a lower estimate for  $\int_0^\varepsilon \mathbb{E}[\varphi(\hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon) - \varphi(y)]dt$ .

To this end we consider the usual Yosida approximation  $J_n = n(nI - A)^{-1}$ ,  $n \geq 1$  (recall that  $J_n \in L(H)$ ,  $J_n : H \rightarrow D(A)$ ,  $J_n x \rightarrow x$  for all  $x \in H$ ), and the approximating equation

$$\hat{X}_t^{\varepsilon,n} = T_t(J_n x) + \int_0^t T_{t-s}(J_n F(\hat{X}_s^{\varepsilon,n}, v_s^\varepsilon))ds + \int_0^t T_{t-s}(J_n G(\hat{X}_s^{\varepsilon,n}, v_s^\varepsilon))dW_s^m, \quad t \in [0, T].$$

Exactly as for (2.12) it can be easily shown that the above equation admits a unique mild solution  $\hat{X}^{\varepsilon,n} \in L_{\mathbb{R}^m}^p(\Omega, C([0, T]; D(A)))$ ,  $p \geq 1$ . Moreover,

- (i)  $\mathbb{E}[\sup_{t \in [0, T]} |\hat{X}_t^{\varepsilon,n} - \hat{X}_t^\varepsilon|_H^p] \rightarrow 0$ , as  $n \rightarrow +\infty$ ;
- (ii)  $\hat{X}^{\varepsilon,n}$  is a strong solution of the SDE

$$\begin{cases} d\hat{X}_t^{\varepsilon,n} = A\hat{X}_t^{\varepsilon,n}dt + J_n F(\hat{X}_t^{\varepsilon,n}, v_t^\varepsilon)dt + J_n G(\hat{X}_t^{\varepsilon,n}, v_t^\varepsilon)dW_t^m, & t \geq 0, \\ \hat{X}_0^{\varepsilon,n} = J_n x. \end{cases}$$

Thus we can apply Itô's formula to  $\varphi(\hat{X}_t^\varepsilon)$  and obtain

$$\begin{aligned} & \mathbb{E}[\varphi(\hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon) - \varphi(J_n x)] \\ &= \mathbb{E}\left[\int_0^{t \wedge \tau_\varepsilon} \langle D\varphi(\hat{X}_s^{\varepsilon,n}), A\hat{X}_s^{\varepsilon,n} \rangle ds\right] + \mathbb{E}\left[\int_0^{t \wedge \tau_\varepsilon} \langle D\varphi(\hat{X}_s^{\varepsilon,n}), J_n F(\hat{X}_s^{\varepsilon,n}, v_s^\varepsilon) \rangle ds\right] \\ &+ \frac{1}{2} \sum_{i=1}^m \mathbb{E}\left[\int_0^{t \wedge \tau_\varepsilon} \langle D^2\varphi(\hat{X}_s^{\varepsilon,n})J_n G(\hat{X}_s^{\varepsilon,n}, v_s^\varepsilon)e_i, J_n G(\hat{X}_s^{\varepsilon,n}, v_s^\varepsilon)e_i \rangle ds\right]. \end{aligned}$$

Since

$$\begin{aligned} \langle D\varphi(\hat{X}_s^{\varepsilon,n}), A\hat{X}_s^{\varepsilon,n} \rangle &= \langle A^* D\varphi_0(\hat{X}_s^{\varepsilon,n}), \hat{X}_s^{\varepsilon,n} \rangle + \hat{g}'(|\hat{X}_s^{\varepsilon,n}|)|\hat{X}_s^{\varepsilon,n}| \oplus \langle A\hat{X}_s^{\varepsilon,n}, \hat{X}_s^{\varepsilon,n} \rangle \\ &\geq \langle A^* D\varphi_0(\hat{X}_s^{\varepsilon,n}), \hat{X}_s^{\varepsilon,n} \rangle, \end{aligned}$$

and  $A^* D\varphi_0 : H \rightarrow H$  is continuous, the dominated convergence theorem allows us to take the limit  $n \rightarrow +\infty$  in the preceding relation:

$$\begin{aligned} \mathbb{E}[\varphi(\hat{X}_{t \wedge \tau_\varepsilon}^\varepsilon) - \varphi(x)] &\geq \mathbb{E}\left[\int_0^{t \wedge \tau_\varepsilon} \langle A^* D\varphi_0(\hat{X}_s^\varepsilon), \hat{X}_s^\varepsilon \rangle ds\right] \\ &+ \mathbb{E}\left[\int_0^{t \wedge \tau_\varepsilon} \langle D\varphi(\hat{X}_s^\varepsilon), F(\hat{X}_s^\varepsilon, v_s^\varepsilon) \rangle ds\right] \\ &+ \frac{1}{2} \sum_{i=1}^m \mathbb{E}\left[\int_0^{t \wedge \tau_\varepsilon} \langle D^2\varphi(\hat{X}_s^\varepsilon)G(\hat{X}_s^\varepsilon, v_s^\varepsilon)e_i, G(\hat{X}_s^\varepsilon, v_s^\varepsilon)e_i \rangle ds\right]. \end{aligned}$$

Now let for  $z \in H, \psi \in C^2(H)$ ,

$$\begin{aligned} & L^m \psi(z) \\ &:= \inf_{u \in U} \left\{ \frac{1}{2} \sum_{i=1}^m \langle D^2\psi(z)G(z, u)e_i, G(z, u)e_i \rangle + \langle D\psi(z), F(z, u) \rangle + \langle A^* D\varphi_0(z), z \rangle \right\}. \end{aligned}$$

As the infimum over functions which are uniformly bounded on bounded sets and continuous, uniformly in  $u \in U$ , the function  $L^m \varphi$  is continuous and bounded on

bounded subsets of  $H$ . Given an arbitrarily small  $\rho > 0$  we choose  $\delta_1 > 0$  in the definition of the stopping time  $\tau_\varepsilon$  s.t.  $|L^m\varphi(z) - L^m\varphi(x)| \leq \rho$  for all  $z \in \overline{B_{\delta_1}(x)}$ . Obviously,

$$\begin{aligned} & \int_0^\varepsilon \mathbb{E} \left[ \varphi(\widehat{X}_{t \wedge \tau_\varepsilon}^\varepsilon) - \varphi(x) \right] dt \\ & \geq \mathbb{E} \left[ \int_0^\varepsilon \int_0^{t \wedge \tau_\varepsilon} L^m \varphi(\widehat{X}_s^\varepsilon) ds dt \right] \geq \mathbb{E} \left[ \int_0^\varepsilon \int_0^{t \wedge \tau_\varepsilon} (L^m \varphi(x) - \rho) ds dt \right] \\ & \geq 1/2 \cdot \mathbb{P}\{\tau_\varepsilon \geq \varepsilon\} \varepsilon^2 (L^m \varphi(x) - \rho) - c\varepsilon^2 \mathbb{P}\{\tau_\varepsilon < \varepsilon\}. \end{aligned}$$

Consequently, from the above estimates and (3.10) we have

$$\begin{aligned} & 1/2 \cdot \mathbb{P}\{\tau_\varepsilon \geq \varepsilon\} \varepsilon^2 (L^m \varphi(x) - \rho) - c\varepsilon^2 \mathbb{P}\{\tau_\varepsilon < \varepsilon\} \\ & \leq \int_0^\varepsilon \mathbb{E} \left[ \varphi(\widehat{X}_{t \wedge \tau_\varepsilon}^\varepsilon) - \varphi(x) \right] dt \leq C\varepsilon^2 d_K^2(x) + c_{m,x,y,r,\delta_1} (\varepsilon^{3-2\gamma} \vee \varepsilon^{5/2}). \end{aligned}$$

Since  $\mathbb{P}\{\tau_\varepsilon \geq \varepsilon\} \rightarrow 1$ , by dividing both sides of the inequality by  $1/2 \cdot \varepsilon^2$  and taking the limit as  $\varepsilon \downarrow 0$ , we obtain

$$L^m \varphi(x) - \rho \leq 2C d_K^2(x).$$

Finally, letting  $\rho \downarrow 0$ , we get the wished result.  $\square$

It is worth pointing out that with the above argument (in simplified form) we obtain a more precise result when the noise is trace class: in particular the convexity of  $K$  is not needed. This is done in the following remark.

**REMARK 3.7.** *Suppose that assumptions (2.3), (2.4), and (2.5) are replaced by the stronger*

$$(3.12) \quad |G(x, u) - G(y, u)|_{L_2(\Xi, H)} \leq C|x - y|, \quad |G(x, u)|_{L_2(\Xi, H)} \leq C,$$

and let  $K$  be a general nonempty closed subset of  $H$ .

The above argument yields that if  $K$  is  $\varepsilon$ -viable w.r.t. (2.1), then there exists  $C > 0$  (large enough) such that  $d_K^2$  is a viscosity supersolution of the HJB equation

$$(3.13) \quad \Psi(x, V(x), DV(x), D^2V(x)) - \langle Ax, DV(x) \rangle = 0, \quad x \in H,$$

where  $\Psi(x, v, p, X) = -\inf_{u \in U} \left( \frac{1}{2} \sum_{i=1}^\infty \langle XG(x, u)e_i, G(x, u)e_i \rangle + \langle F(x, u), p \rangle \right) + Cv - d_K^2(x)$  (recall that in this case  $\sum_{i=1}^\infty |G(x, u)e_i|^2 < \infty$ ). We stress the fact that in this case the convexity of  $K$  is not needed and we can drop Assumption 3.1 (or 3.2).

The proof is the simplified version of the above argument obtained replacing  $X^{x,u,m}$  and  $Y^{x,u,m}$  by  $X^{x,u}$  itself (avoiding projections on  $\mathcal{F}^m$  measurable processes and approximations). The reason why this simplification works is that, to obtain the analogue of (3.11), we can directly apply Itô's rule to  $\phi(J_n X^{x,u})$ . Moreover, in the present case in Lemma 3.6, we can just take  $v = u$  and easily prove that

$$\mathbb{E} |X_\tau^{x,u} - X_\tau^{y,u}|^2 - |x - y|^2 \leq ct|x - y|^2.$$

#### 4. Necessary and sufficient conditions.

**4.1. Viability of balls.** Let us consider here the case in which  $K = B(0, R) = \{x \in H : |x| \leq R\}$ . We will derive from the general necessary condition in Theorem 3.5 specific conditions and show that they are also sufficient.

We denote by  $\{J_n : n \in \mathbb{N}\}$  a family of linear bounded operators  $H \rightarrow D(A^*)$  such that  $|J_n|_{L(H)} \leq 1$  and  $J_n x \rightarrow x$ , as  $n \rightarrow \infty$  for all  $x \in H$ . For instance, we can define  $J_n = n(nI - A^*)^{-1}$ , or, if there exists an orthonormal basis  $\{f_n : n \in \mathbb{N}\}$  of eigenvectors of  $A$  (that, consequently, has to be diagonal), then we can define  $J_n$  as the orthogonal projection on the linear subspace generated by  $f_1, \dots, f_n$ .

**PROPOSITION 4.1.** *Assume that  $B(0, R)$  is  $\varepsilon$ -viable for (2.1). Then for all  $x \in D(A^*)$  with  $|x| = R$  and for all  $n \in \mathbb{N}$  there exists a sequence of controls  $\{u_\ell : \ell \in \mathbb{N}\} \subset U$  such that for all  $m \in \mathbb{N}$*

$$(4.1) \quad \begin{aligned} & \lim_{\ell \rightarrow +\infty} \langle G(x, u_\ell) e_i, x \rangle = 0, \quad i \in \mathbb{N}, \\ & \limsup_{\ell \rightarrow +\infty} \left\{ \frac{1}{2} \sum_{i=1}^m |J_n G(x, u_\ell) e_i|^2 + \langle F(x, u_\ell), x \rangle + \langle A^* x, x \rangle \right\} \leq 0. \end{aligned}$$

*Proof.* With our choice of  $K$  we have  $d_K^2(x) = [(|x| - R)^+]^2$ . Let us fix arbitrarily  $x \in D(A^*)$  with  $|x| > R$  and  $\rho < |x| - R$ . Expanding (in Taylor sense)  $d_K^2(x)$  around  $x$  we get that if  $|y - x| \leq \rho$ , then

$$\begin{aligned} d_K^2(y) & \geq d_K^2(x) + 2 \left( 1 - \frac{R}{|x|} \right) \langle y - x, x \rangle \\ & \quad + \frac{R}{(|x| + \rho)^3} \langle y - x, x \rangle^2 + \left( 1 - \frac{R}{|x| - \rho} - \frac{2\rho R|x|}{(|x| + \rho)^3} \right) |x - y|^2. \end{aligned}$$

Choosing  $\rho$  small enough we can assume that  $(1 - \frac{R}{|x| - \rho} - \frac{2\rho R|x|}{(|x| + \rho)^3}) \geq 0$ . Consequently, if we define

$$\begin{aligned} \varphi(y) & := d_K^2(x) + 2 \left( 1 - \frac{R}{|x|} \right) \langle y - x, x \rangle + \frac{R \langle y - x, x \rangle^2}{(|x| + \rho)^3} \\ & \quad + \left( 1 - \frac{R}{|x| - \rho} - \frac{2\rho R|x|}{(|x| + \rho)^3} \right) |J_n^*(x - y)|^2, \end{aligned}$$

then we get  $\varphi(y) \leq d_K^2(y)$  for all  $y \in B(x, \rho)$  (recall that  $|J_n| \leq 1$ ). Moreover, since  $x \in D(A^*)$ , we have that  $A^* D\varphi$  is well defined and continuous (notice that we are considering  $\varphi$  as a function of  $y$ ). Indeed, we have

$$D\varphi(y) := 2 \left( 1 - \frac{R}{|x|} \right) x + \frac{2R \langle y - x, x \rangle}{(|x| + \rho)^3} x + 2 \left( 1 - \frac{R}{|x| - \rho} - \frac{2\rho R|x|}{(|x| + \rho)^3} \right) J_n J_n^*(x - y).$$

We now apply Theorem 3.5 with the test function  $\varphi$  and then let  $\rho \searrow 0$ . This yields

$$\begin{aligned} & \inf_{u \in U} \left\{ \sum_{i=1}^m \left[ \frac{|x| - R}{|x|} |J_n G(x, u) e_i|^2 + \frac{R}{|x|^3} \langle G(x, u) e_i, x \rangle^2 \right] \right. \\ & \quad \left. + 2 \frac{|x| - R}{|x|} \langle F(x, u), x \rangle \right\} + 2 \frac{|x| - R}{|x|} \langle x, A^* x \rangle \leq C_m [(|x| - R)^+]^2. \end{aligned}$$

Now we fix  $x \in D(A^*)$ , with  $|x| = R$ , and apply the previous inequality to  $x_\ell = \ell(\ell - 1)^{-1}x$ . This yields

$$2\frac{1}{\ell}\langle x_\ell, A^*x_\ell \rangle + \inf_{u \in U} \left\{ \sum_{i=1}^m \left[ \frac{1}{\ell} |J_n G(x_\ell, u)e_i|^2 + \frac{(\ell - 1)^3}{R^2 \ell^3} \langle G(x_\ell, u)e_i, x_\ell \rangle^2 \right] + \frac{2}{\ell} \langle F(x_\ell, u), x_\ell \rangle \right\} \leq C_m \frac{R^2}{(\ell - 1)^2}.$$

Thus for all  $\ell \in \mathbb{N}$  large enough, we can find  $u_\ell^m$  such that

$$2\ell \langle x_\ell, A^*x_\ell \rangle + 2\ell \langle F(x_\ell, u_\ell^m), x_\ell \rangle + \sum_{i=1}^m \left[ \ell |J_n G(x_\ell, u_\ell^m)e_i|^2 + \frac{(\ell - 1)^2}{R^2} \langle G(x_\ell, u_\ell^m)e_i, x_\ell \rangle^2 \right] \leq 2C_m R^2.$$

Letting  $\ell \rightarrow \infty$  we then get

$$\lim_{\ell \rightarrow \infty} \sum_{i=1}^m \langle G(x_\ell, u_\ell^m)e_i, x_\ell \rangle^2 = 0, \\ \limsup_{\ell \rightarrow \infty} \left\{ 2\langle x_\ell, A^*x_\ell \rangle + 2\langle F(x_\ell, u_\ell^m), x_\ell \rangle + \sum_{i=1}^m |J_n G(x_\ell, u_\ell^m)e_i|^2 \right\} \leq 0.$$

Since both  $\sum_{i=1}^m \langle G(\cdot, u)e_i, \cdot \rangle^2$  and  $2\langle F(\cdot, u), \cdot \rangle + \sum_{i=1}^m |J_n G(\cdot, u)e_i|^2$  are continuous on  $D(A^*)$ , uniformly in  $u \in U$ , we immediately get that

$$\lim_{\ell \rightarrow \infty} \sum_{i=1}^m \langle G(x, u_\ell^m)e_i, x \rangle^2 = 0, \\ \limsup_{\ell \rightarrow \infty} \left\{ 2\langle x, A^*x \rangle + 2\langle F(x, u_\ell^m), x \rangle + \sum_{i=1}^m |J_n G(x, u_\ell^m)e_i|^2 \right\} \leq 0.$$

Choosing, in a suitable way, the sequence  $\{u_\ell^m : \ell \in \mathbb{N}\}$ , we can assume that

$$\langle G(x, u_\ell^m)e_i, x \rangle^2 \leq 1/\ell, \quad i = 1, \dots, m, \\ 2\langle x, A^*x \rangle + 2\langle F(x, u_\ell^m), x \rangle + \sum_{i=1}^m |J_n G(x, u_\ell^m)e_i|^2 \leq 1/\ell.$$

Then the claim is proved by letting  $u_\ell = u_\ell^\ell$ .  $\square$

**COROLLARY 4.2.** *Assume that  $B(0, R)$  is  $\varepsilon$ -viable for (2.1) and that either  $U$  is compact or the mappings  $u \rightarrow F(x, u)$ ,  $u \rightarrow G(x, u)e_i$ ,  $i \in \mathbb{N}$ , are affine and  $U$  is convex; then for all  $x \in D(A^*)$  with  $|x| = R$  there exists a control  $u \in U$  such that*

$$(4.2) \quad G^*(x, u)x = 0, \quad \frac{1}{2}|G(x, u)e_i|_{L_2(\Xi, H)}^2 + \langle F(x, u), x \rangle + \langle A^*x, x \rangle \leq 0.$$

*Proof.* We prove the claim only in the case in which the mappings  $u \rightarrow F(x, u)$ ,  $u \rightarrow G(x, u)e_i$ ,  $i \in \mathbb{N}$ , are affine since the proof in the other case is identical. Let  $x \in D(A^*)$  be such that  $|x| \geq R$ . We notice that  $U$  is a convex, closed (thus weakly closed), and bounded subset of a reflexive Banach space. Thus we can assume, for the

sequence  $\{u_\ell : \ell \in \mathbb{N}\}$  got by Proposition 4.1, that  $u_\ell \rightharpoonup u_\infty$  (weakly) for some  $u_\infty \in U$ . Since the mappings  $u \rightarrow F(x, u)$  and  $u \rightarrow G(x, u)e_i$  are continuous with respect to the weak topology, we have  $F(x, u_\ell) \rightharpoonup F(x, u_\infty)$  and  $G(x, u_\ell)e_i \rightharpoonup G(x, u_\infty)e_i$ , and passing to the limit as  $\ell \rightarrow \infty$ , relation (4.1) yields

$$\langle G(x, u_\infty)e_i, x \rangle = 0, \quad i \in \mathbb{N}; \quad \frac{1}{2} \sum_{i=1}^m |J_n G(x, u_\infty)e_i|^2 + \langle F(x, u_\infty), x \rangle + \langle A^*x, x \rangle \leq 0.$$

Thus, fixed  $m \in \mathbb{N}$  for all  $n \in \mathbb{N}$ , there exists  $u^n \in U$  such that

$$\langle G(x, u^n)e_i, x \rangle = 0, \quad i \in \mathbb{N}, \quad \frac{1}{2} \sum_{i=1}^m |J_n G(x, u^n)e_i|^2 + \langle F(x, u^n), x \rangle + \langle A^*x, x \rangle \leq 0.$$

Again we can extract a subsequence such that  $u^n \rightharpoonup u^\infty$ . Then  $J_n G(x, u^n)e_i \rightharpoonup G(x, u^\infty)e_i$ ,  $\langle G(x, u^n)e_i, x \rangle \rightarrow \langle G(x, u^\infty)e_i, x \rangle$  for all  $i \in \mathbb{N}$ , and  $\langle F(x, u^n), x \rangle \rightarrow \langle F(x, u^\infty), x \rangle$ . Thus, letting  $n \rightarrow \infty$  we obtain

$$\langle G(x, u^\infty)e_i, x \rangle = 0, \quad i \in \mathbb{N}, \quad \frac{1}{2} \sum_{i=1}^m |G(x, u^\infty)e_i|^2 + \langle F(x, u^\infty), x \rangle + \langle A^*x, x \rangle \leq 0.$$

The claim follows repeating the same argument for  $m \rightarrow \infty$ .  $\square$

REMARK 4.3. Notice that (4.2) immediately implies (4.1).

We now start to show that condition (4.1) implies  $\varepsilon$ -viability of the ball. We will complete the proof under the following additional assumption on  $A$ .

ASSUMPTION 4.4.  $A$  is diagonal and  $\{f_i : i \in \mathbb{N}\}$  is an orthonormal basis of eigenvectors corresponding to the eigenvalues  $\{-\lambda_i : i \in \mathbb{N}\}$  with  $0 \leq \lambda_1 \leq \lambda_2, \dots$ , and  $\lambda_n \nearrow +\infty$ .

By  $J_n$  we denote the orthogonal projection on the space  $H_n$  generated by  $f_1, \dots, f_n$ . Finally, to shorten notation, we assume that  $R = 1$ .

For all  $x \in H$  and  $u \in \mathcal{U}$  we consider

$$(4.3) \quad \begin{cases} dZ_t^{n,m,x,u} = (AZ_t^{n,m,x,u} + J_n F(Z_t^{n,m,x,u}, u_t)) dt + J_n G(Z_t^{n,m,x,u}, u_t) dW_t^m, & t \geq 0, \\ Z_0^{n,m,x,u} = J_n x. \end{cases}$$

We notice that any solution of (4.3) lives in the finite dimensional space  $H_n$  and that its coefficients are Lipschitz (in  $H_n$ ). Thus, by totally standard arguments the above equation has a unique (classical) continuous solution.

LEMMA 4.5. Assume that  $U$  is compact and that  $G^*$  is strongly continuous (that is for all  $h \in H$  the mappings  $(x, u) \rightarrow G^*(x, u)h : H \times U \rightarrow \Xi$  are continuous). Then for all  $a > 0$  and for all  $\varepsilon > 0$  we can find  $n, m \in \mathbb{N}$  verifying

$$\mathbb{E} \int_0^a |X_t^{x,u} - Z_t^{n,m,x,u}|^2 dt \leq \varepsilon \quad \text{for all } u \in \mathcal{U}.$$

*Proof.* For all  $n \in \mathbb{N}$  we consider

$$(4.4) \quad \begin{cases} dZ_t^{n,x,u} = (AZ_t^{n,x,u} + J_n F(Z_t^{n,x,u}, u_t)) dt + J_n G(Z_t^{n,x,u}, u_t) dW_t, & t \geq 0, \\ Z_0^{n,x,u} = J_n x. \end{cases}$$

Exactly as (2.1), the above equation has a unique  $\mathbb{F}$ -progressively measurable mild solution  $Z^{n,x,u}$  satisfying  $\mathbb{E}[\sup_{s \in [0,T]} |Z_s^{n,x,u}|^p] < +\infty$ .



Clearly, it is enough to prove that for an arbitrary  $a > 0$ ,  $x \in H$ , and the arbitrary sequence  $\{u^n : n \in \mathbb{N}\} \subset \mathcal{U}$ , we have

$$(4.5) \quad \lim_{n \rightarrow +\infty} \mathbb{E} \int_0^a |Z_s^{n,x,u^n} - X_s^{x,u^n}|^2 ds = 0$$

and that, for any  $n \in \mathbb{N}$  and any sequence  $\{u^m : m \in \mathbb{N}\} \subset \mathcal{U}$ , we have

$$(4.6) \quad \lim_{m \rightarrow +\infty} \mathbb{E} \int_0^a |Z_s^{n,m,x,u^m} - Z_s^{n,x,u^m}|^2 ds = 0.$$

To prove (4.5) we start by noticing that  $Ax = \sum_{i=1}^{+\infty} (-\lambda_i) \langle x, f_i \rangle f_i$  and  $T_t x = \sum_{i=1}^{+\infty} e^{-\lambda_i t} \langle x, f_i \rangle f_i$ . Thus,  $T_t(x - J_n x) = \sum_{i=n+1}^{+\infty} e^{-\lambda_i t} \langle x, f_i \rangle f_i$  and, consequently,

$$(4.7) \quad |T_t(I - J_n)x| \leq e^{-\lambda_{n+1}t} |x|.$$

Let  $\bar{Z}_t = Z_t^{n,x,u^n} - X_t^{x,u^n}$ . By trivial computations recalling that  $T_t$  and  $J_n$  commute

$$\begin{aligned} \bar{Z}_t &= T_t(J_n - I)x + \int_0^t T_{t-s}(J_n - I)F(X_s^{x,u^n}, u_s^n) ds \\ &\quad + \int_0^t T_{t-s}J_n \left[ F(Z_s^{n,x,u^n}, u_s^n) - F(X_s^{x,u^n}, u_s^n) \right] ds \\ &\quad + \int_0^t T_{t-s}J_n \left[ G(Z_s^{n,x,u^n}, u_s^n) - G(X_s^{x,u^n}, u_s^n) \right] dW_s \\ &\quad + \int_0^t T_{(t-s)/2}(J_n - I)T_{(t-s)/2}G(X_s^{x,u^n}, u_s^n) dW_s. \end{aligned}$$

Thus, exploiting Lipschitzianity and boundedness of  $F$  and  $G$  we get for all  $0 < t \leq a$ , all  $b > 0$ , and a suitable constant  $c$  that may depend on  $a$  but is independent on  $u$ ,  $n$ , and  $b$

$$\begin{aligned} e^{-bt} \mathbb{E} |\bar{Z}_t|^2 &\leq ce^{-2\lambda_{n+1}t} |x|^2 + c \int_0^t (1 + (t-s)^{-2\gamma}) e^{-b(t-s)} e^{-bs} \mathbb{E} |\bar{Z}_s|^2 ds \\ &\quad + c \int_0^t e^{-2\lambda_{n+1}(t-s)} ds + c \int_0^t e^{-\lambda_{n+1}(t-s)} ((t-s)/2)^{-2\gamma} ds. \end{aligned}$$

Therefore,

$$\begin{aligned} \sup_{t \leq a} e^{-bt} \mathbb{E} |\bar{Z}_t|^2 &\leq ce^{-2\lambda_{n+1}t} |x|^2 + c \left[ \int_0^a (1 + \zeta^{-2\gamma}) e^{-b\zeta} d\zeta \right] \left[ \sup_{t \leq a} e^{-bt} \mathbb{E} |\bar{Z}_t|^2 \right] \\ &\quad + c \int_0^a e^{-2\lambda_{n+1}\zeta} d\zeta + c \int_0^a e^{-\lambda_{n+1}\zeta} (\zeta/2)^{-2\gamma} d\zeta. \end{aligned}$$

Choosing  $b$  large enough, we can assume that  $c \int_0^a (1 + \zeta^{-2\gamma}) e^{-b\zeta} d\zeta \leq 1/2$  and deduce that  $\sup_{t \leq a} e^{-bt} \mathbb{E} |\bar{Z}_t|^2 \rightarrow 0$  as  $n \rightarrow \infty$ . We can, consequently, conclude that (4.5) holds.

We now come to (4.6). Let us put  $\bar{Z}_t = Z_t^{n,m,x,u^m} - Z_t^{n,x,u^m}$ . By trivial computations recalling that  $T_t$  and  $J_n$ , commute

$$\begin{aligned} \bar{Z}_t &= \int_0^t T_{t-s} J_n \left[ F(Z_s^{n,m,x,u^m}, u_s^m) - F(Z_s^{n,x,u^m}, u_s^m) \right] ds \\ &\quad + \int_0^t T_{t-s} J_n \left[ G(Z_s^{n,m,x,u^m}, u_s^m) - G(Z_s^{n,x,u^m}, u_s^m) \right] dW_s \\ &\quad - \int_0^t T_{t-s} J_n G(Z_s^{n,x,u^m}, u_s^m) (I - \Pi_m) dW_s, \end{aligned}$$

where  $\Pi_m$  is the orthogonal projection on the linear space generated by  $e_1, \dots, e_m$  (recall that  $\{e_i : i \in \mathbb{N}\}$  is the orthonormal basis in  $\Xi$  we have fixed in (2.5)). Since  $J_n$  is Hilbert–Schmidt we get

$$\mathbb{E} \sup_{s \leq t} |\bar{Z}_s|^2 \leq c_n \mathbb{E} \int_0^t |\bar{Z}_s|^2 ds + c \int_0^t \mathbb{E} |J_n G(Z_s^{n,x,u^m}, u_s^m) (I - \Pi_m)|_{L_2(\Xi, H)}^2 ds,$$

where  $c_n$  is a suitable constant depending on  $n$  but independent on  $u$ ,  $m$ , and  $x$ . Thus, by Gronwall’s lemma, to get (4.6) it is enough to prove that for all fixed  $n$

$$(4.8) \quad \mathbb{E} \int_0^t |J_n G(Z_s^{n,x,u^m}, u_s^m) (I - \Pi_m)|_{L_2(\Xi, H)}^2 ds \rightarrow 0 \text{ as } m \rightarrow +\infty.$$

We start showing that, fixed  $r > 0$ ,  $\mathbb{E} \int_0^t I_{\{|Z_s^{n,x,u^m}| \leq r\}} |J_n G(Z_s^{n,x,u^m}, u_s^m) (I - \Pi_m)|_{L_2(\Xi, H)}^2 ds \rightarrow 0$ . Since  $|J_n G(Z_s^{n,x,u^m}, u_s^m) (I - \Pi_m)|_{L_2(\Xi, H)} \leq |J_n|_{L_2(\Xi, H)} |G(Z_s^{n,x,u^m}, u_s^m)|_{L(\Xi, H)}$ , by the dominated convergence theorem, relation (4.8) follows if we can prove that for arbitrary sequences  $\{z_m \in H_n : m \in \mathbb{N}; |z_m| \leq c\}$  and  $\{v_m \in U : m \in \mathbb{N}\}$  it holds that  $|J_n G(z_m, v_m) (I - \Pi_m)|_{L_2(\Xi, H)}^2 \rightarrow 0$ . Since  $H_n$  is finite dimensional and  $U$  is compact, w.l.o.g. we can assume that  $z_m \rightarrow z_\infty$ ,  $v_m \rightarrow v_\infty$  for a suitable  $z_\infty \in H_n$ ,  $u_\infty \in U$ . Moreover,

$$\begin{aligned} |J_n G(z_m, v_m) (I - \Pi_m)|_{L_2(\Xi, H)}^2 &\leq 2 |J_n [G(z_m, v_m) - G(z_\infty, v_\infty)] (I - \Pi_m)|_{L_2(\Xi, H)}^2 \\ &\quad + 2 |J_n G(z_\infty, v_\infty) (I - \Pi_m)|_{L_2(\Xi, H)}^2 \doteq 2I_m^1 + 2I_m^2. \end{aligned}$$

We notice that

$$I_m^1 = \sum_{i=1}^n \sum_{j=m+1}^\infty \langle f_i, [G(z_m, v_m) - G(z_\infty, v_\infty)] e_j \rangle^2 \leq \sum_{i=1}^n |[G(z_m, v_m) - G(z_\infty, v_\infty)]^* f_i|^2.$$

Thus,  $I_m^1 \rightarrow 0$  by the strong continuity of  $G^*$ . In a similar way

$$I_m^2 = \sum_{i=1}^n \sum_{j=m+1}^\infty \langle G^*(z_\infty, v_\infty) f_i, e_j \rangle^2 \rightarrow 0$$

since  $\sum_{j=1}^\infty \langle G^*(z_\infty, v_\infty) f_i, e_j \rangle^2 = |G^*(z_\infty, v_\infty) f_i|^2 < \infty$ . It remains to be proven that

$$\mathbb{E} \int_0^t I_{\{|Z_s^{n,x,u^m}| > r\}} |J_n G(Z_s^{n,x,u^m}, u_s^m) (I - \Pi_m)|_{L_2(\Xi, H)}^2 ds \rightarrow 0$$

as  $r \nearrow +\infty$  (uniformly with respect to  $m$ ). Since  $J_n$  is Hilbert–Schmidt and  $G$  is bounded, it is enough to show that  $P \otimes \lambda\{|Z_s^{n,x,u^m}| > r : s \in [0, a]\} \rightarrow 0$  as  $r \rightarrow \infty$

(here  $\lambda$  denotes the Lebesgue measure on  $\mathbb{R}$ ). This last statement is obvious since, as in Lemma 2.2, we have  $\mathbb{E} \sup_{s \in [0, a]} |Z_s^{n, x, u^m}|^2 \leq c_a(1 + |x|)$ , where  $c_a$  is a constant independent on  $m$  (and even  $n$ ).  $\square$

We now extend (4.1) holding on the boundary of the ball to its exterior.

LEMMA 4.6. *Assume (4.1) (with  $R = 1$ ) and fix  $m, n \in \mathbb{N}$ . Then there exists a constant  $c$  such that, for all  $\varepsilon > 0$  and  $x \in H_n$  with  $|x| \in [1, 2]$ , there exists an  $\bar{u}(x) \in U$  verifying*

$$(4.9) \quad \begin{aligned} \sum_{i=1}^m \langle G(y, \bar{u}(x))e_i, y \rangle^2 &\leq \varepsilon + c[(1 - |y|)^+]^2, \\ \frac{1}{2} \sum_{i=1}^m |J_n G(y, \bar{u}(x))e_i|^2 &+ \langle F(y, \bar{u}(x)), y \rangle + \langle Ay, y \rangle \leq \varepsilon + c(1 - |y|)^+ \end{aligned}$$

for all  $y \in H_n$  with  $|y - x| \leq \delta_\varepsilon$ , where  $\delta_\varepsilon$  is a suitable constant independent of  $x$  and  $y$ . Moreover, the mapping  $x \rightarrow \bar{u}(x)$  can be chosen to be measurable.

*Proof.* In the following we denote by  $c$  a constant that depends on the data of the problem and on  $n$  and  $m$  but not on  $x$ ,  $y$ , and  $\varepsilon$ ; the value of  $c$  can change from line to line. Let us notice that, clearly,  $H_n \subset D(A) = D(A^*)$ .

If  $x \in H_n$  and  $|x| \in [1, 2]$ , then for all  $u \in U$ ,  $i \in \mathbb{N}$ ,

$$\begin{aligned} |\langle G(x, u)e_i, x \rangle| &\leq |\langle G(x/|x|, u)e_i, x/|x| \rangle| + |\langle G(x, u)e_i, x \rangle - \langle G(x/|x|, u)e_i, x/|x| \rangle| \\ &\leq \langle G(x/|x|, u)e_i, x/|x| \rangle + c|x - x/|x||, \end{aligned}$$

and by (4.1) we immediately have that there exists some  $\bar{u}(x)$  such that

$$\sum_{i=1}^m |\langle G(x, \bar{u}(x))e_i, x \rangle|^2 \leq \varepsilon + c(|x| - 1)^2.$$

Similarly, recalling that  $A$  is a bounded linear operator on  $H_n$ , we get that there exists an  $\tilde{u}(x)$  such that

$$\frac{1}{2} \sum_{i=1}^m |J_n G(x, \tilde{u}(x))e_i|^2 + \langle F(x, \tilde{u}(x)), x \rangle + \langle Ax, x \rangle \leq \varepsilon + c(|x| - 1)^+.$$

From (4.1) we see that we can choose  $\bar{u}(x)$  and  $\tilde{u}(x)$  such that they coincide. Moreover we notice that, for all  $v \in U$ , the mappings

$$y \rightarrow \sum_{i=1}^m |\langle G(y, v)e_i, y \rangle|^2, \quad y \rightarrow \frac{1}{2} \sum_{i=1}^m |J_n G(y, v)e_i|^2 + \langle F(y, v), y \rangle + \langle Ay, y \rangle$$

are Lipschitz on the bounded subsets of  $H_n$  with Lipschitz constant independent on  $v$ . This yields (4.9) for all  $y \in H_n$  with  $|y - x| \leq \delta_\varepsilon$ , where  $\delta_\varepsilon$  can be chosen independently of  $x$  and  $y$ . Finally, the fact that  $\bar{u} : H_n \rightarrow U$  can be gotten as in a measurable function of  $x$  is a direct consequence of well known measurable selection results (see, e.g., [5, Theorem 8.2.10, p. 316]).  $\square$

LEMMA 4.7. *Assume (4.1) (with  $R = 1$ ). For all  $T > 0$ ,  $m, n \in \mathbb{N}$ ,  $\varepsilon > 0$ , and all  $x \in H_n$  with  $|x| \leq 1$  there exists  $u \in \mathcal{U}$  such that*

$$E \int_0^T \left[ (|Z_s^{n, m, x, u}| - 1)^+ \right]^2 ds \leq \varepsilon$$

(recall that the process  $Z^{n,m,x,u}$  has been introduced in (4.3)).

*Proof.* For the sake of simplicity we restrict ourselves to the case  $T = 1$ . The proof for the case  $T > 1$  is clearly identical. Let us extend the definition of  $\bar{u}$  given in Lemma 4.6 by putting  $\bar{u}(x) = u_0$  for all  $x \in H_n$  with  $|x| > 2$  or  $|x| < 1$ . Moreover, we denote by  $Z^{n,m,\tau,\eta,u}$  the solution of (4.3) starting at  $\tau$  from the random variable  $J_n\eta$ . More precisely, for all stopping time  $\tau$ , all  $\eta \in L^2(\Omega, \mathcal{F}_\tau, H)$ , and all  $u \in \mathcal{U}$ ,  $Z^{n,m,\tau,\eta,u}$  is the solution of

$$\begin{cases} dZ_t^{n,m,\tau,\eta,u} = (AZ_t^{n,m,\tau,\eta,u} + J_n F(Z_t^{n,m,\tau,\eta,u}, u_t)) dt \\ \quad + J_n G(Z_t^{n,m,\tau,\eta,u}, u_t) dW_t^m, \quad t \geq \tau, \\ Z_\tau^{n,m,\tau,\eta,u} = J_n \eta. \end{cases}$$

Given  $x \in H_n$ , we let

$$\tilde{\tau}_1 = \left( \inf\{s > 0 : |Z_s^{n,m,0,x,\bar{u}(x)} - x| \geq \delta_\varepsilon\} \vee \inf\{s > 0 : |Z_s^{n,m,0,x,\bar{u}(x)}| \geq 1\} \right) \wedge 1,$$

and we define  $Z_s^\sharp = Z_s^{n,m,0,x,\bar{u}(x)}$ ,  $u_s^\sharp = \bar{u}(x)$  for  $s \in [0, \tilde{\tau}_1[$ . We then proceed iteratively by letting

$$\begin{aligned} \tilde{\tau}_{\ell+1} = & \left( \inf \left\{ s > \tau_\ell : |Z_s^{n,m,\tilde{\tau}_\ell, Z_{\tilde{\tau}_\ell}^\sharp, \bar{u}(Z_{\tilde{\tau}_\ell}^\sharp)} - Z_{\tilde{\tau}_\ell}^\sharp| \geq \delta_\varepsilon \right\} \right. \\ & \left. \vee \inf \left\{ s > \tau_\ell : |Z_s^{n,m,\tilde{\tau}_\ell, Z_{\tilde{\tau}_\ell}^\sharp, \bar{u}(Z_{\tilde{\tau}_\ell}^\sharp)}| \geq 1 \right\} \right) \wedge 1 \end{aligned}$$

and defining  $Z_s^\sharp = Z_s^{n,m,\tilde{\tau}_\ell, Z_{\tilde{\tau}_\ell}^\sharp, \bar{u}(Z_{\tilde{\tau}_\ell}^\sharp)}$ ,  $u_s^\sharp = \bar{u}(Z_{\tilde{\tau}_\ell}^\sharp)$ , for  $s \in [\tilde{\tau}_\ell, \tilde{\tau}_{\ell+1}[$ . By construction  $Z^\sharp$  is the solution of

$$\begin{cases} dZ_t^\sharp = \left( AZ_t^\sharp + J_n F(Z_t^\sharp, u_t^\sharp) \right) dt + J_n G(Z_t^\sharp, u_t^\sharp) dW_t^m, \quad t \in [0, \lim_{\ell \rightarrow \infty} \tilde{\tau}_\ell], \\ Z_0^\sharp = J_n x, \end{cases}$$

or, with the notation introduced by (4.3),  $Z^\sharp = Z^{n,m,x,u^\sharp}$ . Since (4.3) is a SDE on the finite dimensional space  $H_n$  with finite dimensional Brownian noise and uniformly Lipschitz coefficients, standard arguments from the stochastic control theory show that

$$P(\cup_{\ell=1}^\infty \{\tilde{\tau}_\ell = 1\}) = 1 \quad \text{and in particular} \quad \tilde{\tau}_\infty \doteq \lim_{\ell \rightarrow \infty} \tilde{\tau}_\ell = 1, \quad P - \text{a.s..}$$

Indeed, putting

$$\begin{aligned} \eta_1 &= \inf\{s > 0 : |Z_s^{n,m,0,x,u_0^\sharp} - x| \geq \delta_\varepsilon\}, \\ \eta_{\ell+1} &= \inf\{s > \tau_\ell : |Z_s^{n,m,\tilde{\tau}_\ell, Z_{\tilde{\tau}_\ell}^\sharp, u_{\tilde{\tau}_\ell}^\sharp} - Z_{\tilde{\tau}_\ell}^\sharp| \geq \delta_\varepsilon\} \end{aligned}$$

(notice that  $\eta_\ell \leq \tilde{\tau}_\ell$  on  $\{\tilde{\tau}_\ell < 1\}$ ), we have for all  $\ell \in \mathbb{N}$  and all  $h > 0$

$$\begin{aligned} P\{\eta_{\ell+1} - \tilde{\tau}_\ell \leq h | \mathcal{F}_{\tilde{\tau}_\ell}\} &= P\left\{ \sup_{\tilde{\tau}_\ell \leq s \leq \tilde{\tau}_\ell + h} |Z_s^{n,m,\tilde{\tau}_\ell, Z_{\tilde{\tau}_\ell}^\sharp, u_{\tilde{\tau}_\ell}^\sharp} - Z_{\tilde{\tau}_\ell}^\sharp| \geq \delta_\varepsilon \middle| \mathcal{F}_{\tilde{\tau}_\ell} \right\} \\ &\leq c \frac{h^{p/2}}{\delta_\varepsilon^p} (1 + |Z_{\tilde{\tau}_\ell}^\sharp|^p). \end{aligned}$$

Consequently, for  $p > 2$ ,

$$P(\cup_{\ell \geq 0} \{\eta_{\ell+1} - \tilde{\tau}_\ell \leq h/\ell\}) \leq c \frac{h^{p/2}}{\delta_\varepsilon^p} \sum_{\ell=1}^{\infty} \ell^{-p/2} \left( 1 + \mathbb{E} \sup_{s \in [0, \tilde{\tau}_\infty]} \mathbb{E} |Z_{\tilde{\tau}_\ell}^\#|^p \right) \leq c_p \frac{h^{p/2}}{\delta_\varepsilon^p} (1 + |x|^p)$$

from where

$$P(\cup_{\ell=1}^{\infty} \{\tilde{\tau}_\ell = 1\}) \geq \mathbb{P}(\cap_{\ell=1}^{\infty} \{\eta_{\ell+1} - \tilde{\tau}_\ell > h/\ell\}) \geq 1 - c_p \frac{h^{p/2}}{\delta_\varepsilon^p} \rightarrow 1 \text{ as } h \searrow 0.$$

Finally, we let  $\rho = \inf\{t > 0 : |Z_t^\#| \geq 2\} \wedge 1$  and  $\tau_\ell = \tilde{\tau}_\ell \wedge \rho$ . In this way  $P(\cup_{\ell=1}^{\infty} \{\tau_\ell = \rho\}) = 1$ ; moreover,  $|Z_{\tau_\ell}^\#| \in [1, 2]$  and

$$I_{\{|Z_t^\#| \geq 1\}} |Z_t^\# - Z_{\tau_\ell}^\#| \leq \delta_\varepsilon \text{ for all } \ell \in \mathbb{N} \text{ and all } t \in [\tau_\ell, \tau_{\ell+1}[.$$

Finally, by construction, we have  $P$ -a.s. for all  $t \leq \rho$ ,  
(4.10)

$$\begin{aligned} I_{\{|Z_t^\#| \geq 1\}} \left[ \sum_{i=1}^m \langle J_n G(Z_t^\#, u_t^\#) e_i, Z_t^\# \rangle^2 \right] &\leq \varepsilon + c[ (|Z_t^\#| - 1)^+ ]^2, \\ I_{\{|Z_t^\#| \geq 1\}} \left[ \frac{1}{2} \sum_{i=1}^m |J_n G(Z_t^\#, u_t^\#) e_i|^2 + \langle F(Z_t^\#, u_t^\#), Z_t^\# \rangle + \langle A Z_t^\#, Z_t^\# \rangle \right] &\leq \varepsilon + c(|Z_t^\#| - 1)^+. \end{aligned}$$

Now let  $\psi_\varepsilon(r) = 2 \int_0^r \int_0^s [\varepsilon^{-1}(\sigma - 1 + \varepsilon)]^+ \wedge 1 \, d\sigma \, ds$  for  $r \in \mathbb{R}_+$ . We notice that  $\psi_\varepsilon(r) \searrow [(r - 1)^+]^2$ , uniformly on bounded subsets of  $\mathbb{R}_+$ ,  $\psi'_\varepsilon(r) \searrow 2(r - 1)^+$ , uniformly on bounded subsets of  $\mathbb{R}_+$  and  $\psi''_\varepsilon(r) \searrow 2$  if  $r \geq 1$  and  $\psi''_\varepsilon(r) \searrow 0$  if  $r < 1$ . Moreover,  $\psi_\varepsilon(r) \leq c|r|^2$ ,  $\psi'_\varepsilon(r) \leq c|r|$ ,  $\psi''_\varepsilon(r) \leq c$  for a suitable  $c > 0$  and all  $r \in \mathbb{R}_+$ ,  $\varepsilon \in ]0, 1]$ . Finally,  $\psi_\varepsilon(r)$  is of class  $C^2$  and  $\psi_\varepsilon(r) = 0$  for all  $r < 1 - \varepsilon$ .

If we compute by Itô's rule  $d_t \psi_\varepsilon(|Z_t^\#|)$ , integrate between 0 and  $t \wedge \rho$ , and compute the mean value we get, for  $|x| \leq 1$ ,

$$\begin{aligned} \mathbb{E} \psi_\varepsilon(|Z_{t \wedge \rho}^\#|) &= \mathbb{E} \int_0^{t \wedge \rho} \frac{\psi'_\varepsilon(|Z_s^\#|)}{|Z_s^\#|} \\ &\quad \times \left[ \langle Z_s^\#, A Z_s^\# \rangle + \langle Z_s^\#, J_n F(Z_s^\#, u_s^\#) \rangle + \frac{1}{2} \sum_{i=1}^m |J_n G(Z_s^\#, u_s^\#) e_i|^2 \right] ds + \psi_\varepsilon(|x|) \\ &\quad + \frac{1}{2} \mathbb{E} \int_0^{t \wedge \rho} \left[ \frac{\psi''_\varepsilon(|Z_s^\#|)}{|Z_s^\#|^2} - \frac{\psi'_\varepsilon(|Z_s^\#|)}{|Z_s^\#|^3} \right] \left[ \sum_{i=1}^m \langle Z_s^\#, J_n G(Z_s^\#, u_s^\#) e_i \rangle^2 \right] ds. \end{aligned}$$

Note that  $|\psi''_\varepsilon(r)/r^2| + |\psi'_\varepsilon(r)/r^3| \leq c$ ,  $r > 0$ . In the above formula and in the following,  $c$  is a constant independent on  $\varepsilon$  small enough. If we let  $\varepsilon \searrow 0$  in the above equation, then we obtain

$$\begin{aligned} &\mathbb{E}[(|Z_{t \wedge \rho}^\#| - 1)^+]^2 \\ &\leq c \mathbb{E} \int_0^{t \wedge \rho} I_{\{|Z_t^\#| > 1\}} \left[ \sum_{i=1}^m \langle Z_s^\#, J_n G(Z_s^\#, u_s^\#) e_i \rangle^2 \right] ds + c \mathbb{E} \int_0^{t \wedge \rho} (|Z_s^\#| - 1)^+ \\ &\quad \times \left[ \langle Z_s^\#, A Z_s^\# \rangle + \langle Z_s^\#, J_n F(Z_s^\#, u_s^\#) \rangle + \frac{1}{2} \sum_{i=1}^m |J_n G(Z_s^\#, u_s^\#) e_i|^2 \right] ds \end{aligned}$$

and by (4.10)

$$\mathbb{E}[(|Z_{t \wedge \rho}^\#| - 1)^+]^2 \leq c\varepsilon + c\mathbb{E} \int_0^{t \wedge \rho} [(|Z_s^\#| - 1)^+]^2 ds \leq c\varepsilon + c\mathbb{E} \int_0^t [(|Z_{s \wedge \rho}^\#| - 1)^+]^2 ds.$$

Thus, by Gronwall's lemma

$$(4.11) \quad \mathbb{E}[(|Z_{t \wedge \rho}^\#| - 1)^+]^2 \leq c\varepsilon, \quad t \in [0, 1[.$$

Moreover, by dominated convergence, letting  $t \nearrow 1$  the above inequality also yields  $\mathbb{E}[(|Z_\rho^\#| - 1)^+]^2 \leq c\varepsilon$ . On the other side,

$$\mathbb{E} \int_\rho^1 [(|Z_s^\#| - 1)^+]^2 ds \leq (P(\rho < 1))^{1/2} \left( \mathbb{E} \sup_{s \in [0, 1]} |Z_s^\#|^4 \right)^{1/2}.$$

Since, as in Lemma 2.6,  $\mathbb{E} \sup_{s \in [0, 1]} |Z_s^{n, m, x, u}|^4 \leq c(1 + |x|^4)$ , we deduce that

$$\begin{aligned} \mathbb{E} \int_\rho^1 [(|Z_s^\#| - 1)^+]^2 ds &\leq c\sqrt{P(\rho < 1)} \leq c\sqrt{P(|Z_\rho^\#| - 1)^2 = 1)} \\ &\leq c\sqrt{\mathbb{E}[(|Z_\rho^\#| - 1)^+]^2} \leq c\sqrt{\varepsilon}, \end{aligned}$$

and the claim immediately follows from the above relation and (4.11).  $\square$

We can now easily conclude our argument

**COROLLARY 4.8.** *Assume (4.1) (with  $R = 1$ ). Then the ball  $B(0, 1)$  is  $\varepsilon$ -viable.*

*Proof.* By Lemmas 4.5 and 4.7 we have that for arbitrary great  $a > 0$  and all  $x \in H$ ,  $|x| \leq 1$ ,  $\varepsilon > 0$ , the existence of an admissible control  $u$  verifying

$$\mathbb{E} \int_0^a [(|X_s^{x, u}| - 1)^+]^2 \leq \varepsilon.$$

Then it is enough to notice that under our assumptions we have by straightforward computations  $\mathbb{E}|X_t^{x, u}|^2 \leq e^{ct}c(1 + |x|^2)$  with  $c$  independent of  $x$ ,  $u$ , and  $t$ .  $\square$

**4.2. Viability of linear spaces.** We consider here the case in which  $K$  is a finite dimensional linear subspace of  $H$ . As in the previous section we derive from (4.1) specific conditions and show that they are sufficient. To start with we adapt the argument in [27] to show that any  $\varepsilon$ -invariant subset of a finite dimensional linear subspace of  $H$  is a subset of  $D(A)$ .

**LEMMA 4.9.** *Let  $K \subset H$  be an  $\varepsilon$ -viable (not necessarily linear) subset of  $H$ , and let  $\widehat{K}$  be the linear space generated by  $K$ . Then for all  $x \in K$  there exist sequences  $t_n \searrow 0$ ,  $\{u^n : n \in \mathbb{N}\} \subset \mathcal{U}$ ,  $\{k_n : n \in \mathbb{N}\} \subset \widehat{K}$  such that  $|\mathbb{E}X_{t_n}^{x, u^n} - k_n|/t_n \rightarrow 0$ .*

*Proof.* Let  $x \in H$  be arbitrarily chosen and  $\Pi_{\widehat{K}}$  denote its orthogonal projection on  $\widehat{K}$ . Then by the  $\varepsilon$ -viability of  $K$ , since  $d_K(x) \geq |\Pi_{\widehat{K}}x|$ , we know that there exists a sequence  $\{u^n : n \in \mathbb{N}\} \subset \mathcal{U}$  such that

$$\mathbb{E} \int_{n^{-2}}^{2n^{-2}} |X_s^{x, u^n} - \Pi_{\widehat{K}}X_s^{x, u^n}|^2 ds \leq 2^{-2n}, \quad n \geq 1;$$

thus there exists some  $t_n$  with  $n^{-2} \leq t_n \leq 2n^{-2}$  such that

$$\mathbb{E} |X_{t_n}^{x, u^n} - \Pi_{\widehat{K}}X_{t_n}^{x, u^n}|^2 \leq 2^{-2n}n^2.$$

Now let  $k_n = \mathbb{E}\Pi_{\widehat{K}}X_{t_n}^{x,u^n}$ . We notice that  $k_n \in \widehat{K}$ . Finally, by Cauchy inequality,

$$\left| \mathbb{E}X_{t_n}^{x,u^n} - k_n \right| \leq \left( \mathbb{E} \left| X_{t_n}^{x,u^n} - \Pi_{\widehat{K}}X_{t_n}^{x,u^n} \right|^2 \right)^{1/2} \leq 2^{-n}n,$$

and the claim follows.  $\square$

The following result, needed below, is a straightforward consequence of [14, Theorem 1.24]. We report the statement for the reader's convenience.

LEMMA 4.10. *Assume that  $y \in H$  is such that there exists a sequence  $t_n \searrow 0$  for which  $\{|T_{t_n}y - y|/t_n : n \in \mathbb{N}\}$  is bounded. Then  $y \in D(A)$ .*

PROPOSITION 4.11. *If  $K \subset H$  is  $\varepsilon$ -viable and it is included in a finite dimensional linear subspace of  $H$ , then  $K \subset D(A)$ .*

*Proof.* Fix  $x \in K$  and again let  $\widehat{K}$  be the linear space generated by  $K$ . By Lemma 4.9 there exist sequences  $t_n \searrow 0$  and  $\{u^n : n \in \mathbb{N}\} \subset \mathcal{U}$ ,  $\{k_n : n \in \mathbb{N}\} \subset \widehat{K}$  such that

$$\frac{1}{t_n} \left| T_{t_n}x + \int_0^{t_n} \mathbb{E}F(X_s^{x,u^n}, u_s^n) ds - k_n \right| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Notice that, since  $\sup_n |\mathbb{E} \int_0^{t_n} F(X_s^{x,u^n}, u_s^n) ds|/t_n < \infty$ , the above convergence immediately yields

$$(4.12) \quad \sup_n \frac{1}{t_n} |T_{t_n}x - k_n| < \infty.$$

If  $k_n = x$  for infinitely many indices  $n$ , then, along this subsequence,  $\sup_n (t_n)^{-1} |T_{t_n}x - x| < \infty$ , and by Lemma 4.10 we obtain that  $x \in D(A)$ . Hence, w.l.o.g., we can assume that  $k_n \neq x$  for all  $n \in \mathbb{N}$ . Since  $\widehat{K}$  is locally compact, we have, at least along a subsequence (that we still denote by  $\{h_n\}$ ), that there exists some  $v \in \widehat{K}$  with  $(k_n - x)/|k_n - x| \rightarrow v$  and  $|v| = 1$ .

Extracting again a suitable subsequence, if necessary, we can distinguish between two different cases: the one in which  $\sup_n |k_n - x|/t_n < \infty$  and the one in which  $|k_n - x|/t_n \nearrow \infty$ . In the second case

$$\left| \frac{T_{t_n}x - x}{|k_n - x|} - v \right| \leq \frac{t_n}{|k_n - x|} \frac{|T_{t_n}x - k_n|}{t_n} + \left| \frac{k_n - x}{|k_n - x|} - v \right| \rightarrow 0.$$

Now let  $N_n = [t/t_n]$  (where  $[a]$  means the integer part of  $a \in \mathbb{R}_+$ ). By direct computation we get

$$(4.13) \quad \frac{t_n}{|k_n - x|} (T_{N_n t_n}x - x) = t_n \sum_{\ell=1}^{N_n} T_{(\ell-1)t_n} v + t_n \sum_{\ell=1}^{N_n} T_{(\ell-1)t_n} \left[ \frac{T_{t_n}x - x}{|k_n - x|} - v \right].$$

The left-hand side converges to 0 since  $t_n/|k_n - x| \searrow 0$ . Moreover,  $t_n \sum_{\ell=1}^{N_n} T_{(\ell-1)t_n} v \rightarrow \int_0^t T_s v ds$ . Finally since we know that  $[(T_{t_n}x - x)/|k_n - x| - v] \rightarrow 0$ , we have  $t_n \sum_{\ell=1}^{N_n} T_{(\ell-1)t_n} \left[ \frac{T_{t_n}x - x}{|k_n - x|} - v \right] \rightarrow 0$ . Consequently,  $\int_0^t T_s v ds = 0$ . But since  $t$  has been arbitrarily chosen this implies that  $v = 0$ , which is impossible. Therefore we can assume that  $\sup_n |k_n - x|/t_n < \infty$ . Together with (4.12) this implies that  $\sup_n |T_{t_n}x - x|/t_n < +\infty$ , and the claim follows by Lemma 4.10.  $\square$

For the rest of this section we will restrict ourselves to the case in which  $K$  is a finite dimensional linear subspace of  $H$  with  $K \subset D(A)$ .

As in section 4.1 we will derive from Theorem 3.5 specific necessary conditions for the  $\varepsilon$ -viability and show that they are also sufficient.

For this end we make the technical assumption that  $D(A) \subset D(A^*)$  and we denote again by  $\{J_n : n \in \mathbb{N}\}$  a family of linear bounded operators  $H \rightarrow D(A^*)$  such that  $|J_n|_{L(H)} \leq 1$  and  $J_n x \rightarrow x$  as  $n \rightarrow \infty$  for all  $x \in H$ .

**PROPOSITION 4.12.** *Assume that the finite dimensional linear subspace  $K \subset H$  is  $\varepsilon$ -viable for (2.1). We denote by  $\Pi$  the orthogonal projection on  $K^\perp$ . Then, for any  $n \in \mathbb{N}$ , arbitrarily fixed  $x \in K$ , and  $y \in D(A^*) \cap K^\perp$  there exists a sequence of controls  $\{u_\ell : \ell \in \mathbb{N}\} \subset U$  such that*

$$(4.14) \quad \lim_{\ell \rightarrow +\infty} \langle F(x, u_\ell), y \rangle + \langle A^* y, x \rangle \leq 0, \quad \lim_{\ell \rightarrow +\infty} |J_n \Pi G(x, u_\ell) e_i|^2 = 0, \quad i \in \mathbb{N}.$$

*Proof.* Clearly  $d_K^2(z) = |\Pi z|^2$ . Moreover, due to Proposition 4.11 and to our assumption that  $D(A) \subset D(A^*)$ ,  $\Pi = I - \Pi_K$  mappings  $D(A^*)$  into  $K^\perp \cap D(A^*)$ . We fix  $z \in D(A^*)$  and notice that for all  $v \in H$

$$d_K^2(v) = |\Pi(v - z) + \Pi z|^2 \geq d_K^2(z) + |J_n \Pi(v - z)|^2 + 2\langle \Pi z, v - z \rangle := \varphi(v).$$

Since  $D\varphi(v) = 2\Pi z + 2\Pi J_n^* J_n \Pi(v - z) \in D(A^*)$  we can apply Theorem 3.5 to the test function  $\varphi$  obtaining

$$\inf_{u \in U} \left\{ \frac{1}{2} \sum_{i=1}^m |J_n \Pi G(z, u) e_i|^2 + \langle F(z, u), \Pi z \rangle + \langle A^* \Pi z, z \rangle \right\} \leq c |\Pi z|^2.$$

If we now apply the above to  $z = x_\ell \doteq x + \ell^{-1}y$ , by proceeding as in the proof of Proposition 4.1 we can find, for  $\ell \in \mathbb{N}$  large enough, some  $u_\ell^m$  such that

$$\begin{aligned} \lim_{\ell \rightarrow \infty} \sum_{i=1}^m |J_n \Pi G(x_\ell, u_\ell^m) e_i|^2 &= 0, \\ \limsup_{\ell \rightarrow \infty} \{ \langle x_\ell, A^* y \rangle + \langle F(x_\ell, u_\ell^m), y \rangle \} &\leq 0. \end{aligned}$$

Thus, since  $x_\ell \rightarrow x$ , by choosing a suitable subsequence of  $\{u_\ell^m : \ell \in \mathbb{N}\}$  we can assume that

$$\begin{aligned} |J_n \Pi G(x, u_\ell^m) e_i|^2 &\leq \frac{1}{\ell}, \quad i = 1, \dots, m, \\ \langle x, A^* y \rangle + \langle F(x, u_\ell^m), y \rangle &\leq \frac{1}{\ell}. \end{aligned}$$

The claim is proved by putting  $u_\ell = u_\ell^m$ .  $\square$

Exactly as in Corollary 4.2 we also have the following corollary.

**COROLLARY 4.13.** *Assume that  $K$  is  $\varepsilon$ -viable and that either*

1.  *$U$  is compact,*
2.  *$U$  is convex and the mappings  $u \rightarrow F(x, u)$ ,  $u \rightarrow G(x, u) e_i$ , and  $i \in \mathbb{N}$  are affine.*

*Then for all  $x \in K$  and  $y \in D(A^*) \cap K^\perp$  there exists a control  $u \in U$  such that*

$$(4.15) \quad \langle F(x, u), y \rangle + \langle A^* y, x \rangle \leq 0, \quad G(x, u) v \in K, \quad v \in \Xi.$$



As in the previous section we prove that conditions (4.14) are also necessary. Again we assume that  $A$  is diagonal and that  $\{f_i : i \in \mathbb{N}\}$  is an orthonormal basis of eigenvectors corresponding to the eigenvalues  $\{-\lambda_i : i \in \mathbb{N}\}$  with  $0 \leq \lambda_1 \leq \lambda_2, \dots$  and  $\lambda_n \nearrow +\infty$ . Moreover, by  $J_n$  we denote now the orthogonal projection on the space  $H_n$  generated by  $f_1, \dots, f_n$  and  $K$ .

**PROPOSITION 4.14.** *Assume that  $U$  is compact and that  $G^*$  is strongly continuous. Moreover, suppose that (4.14) holds. Then the finite dimensional linear subset  $K \subset H$  is  $\varepsilon$ -viable.*

*Proof.* First of all we notice that  $\Pi J_n = J_n \Pi$  (recall that  $\Pi$  is the projection on  $K^\perp$ ). Indeed  $H_n$  is the space generated by  $K$  and  $\Pi f_1, \dots, \Pi f_n$ . Therefore both  $\Pi J_n$  and  $J_n \Pi$  are equal to the projection on  $\Pi f_1, \dots, \Pi f_n$ .

Moreover, since  $(I - J_n)H$  is included in the space generated by  $\{f_i : i \geq n+1\}$  relation (4.7) still holds.

By (4.14) for all  $x \in H_n$  with  $|\Pi x| \leq 1$  there exists some  $\bar{u}(x) \in U$  verifying

$$\sum_{i=1}^m |J_n \Pi G(\Pi_K x, \bar{u}(x)) e_i|^2 \leq \varepsilon/2, \quad \langle F(\Pi_K x, \bar{u}(x)), \Pi x \rangle + \langle A \Pi_K x, \Pi x \rangle \leq \varepsilon/2$$

(recall that  $\Pi_K$  is the projection on  $K$ ).

Since  $A$  is bounded on  $H_n$  and  $F, G$  are Lipschitz uniformly with respect to  $u \in U$ , we have for all fixed  $m \in \mathbb{N}$

$$(4.16) \quad \sum_{i=1}^m |J_n \Pi G(x, \bar{u}(x)) e_i|^2 \leq \varepsilon/2 + c |\Pi x|^2, \quad \langle F(x, \bar{u}(x)), \Pi x \rangle + \langle Ax, \Pi x \rangle \leq \varepsilon/2 + c |\Pi x|^2.$$

Moreover, the above functions are continuous on bounded subsets of  $H_n$ , uniformly in  $u \in U$ . Thus there exists a constant  $c$  such that for all  $\varepsilon > 0$  and  $x \in H_n$  with  $|\Pi x| \leq 1$  there exists an  $\bar{u}(x) \in U$  verifying

$$(4.17) \quad \sum_{i=1}^m |J_n \Pi G(x, \bar{u}(x)) e_i|^2 \leq \varepsilon + c |\Pi y|^2, \quad \langle F(y, \bar{u}(x)), \Pi y \rangle + \langle Ay, \Pi y \rangle \leq \varepsilon + c |\Pi y|^2$$

for all  $y \in H_n$  with  $|y - x| \leq \delta_\varepsilon$ . Here  $\delta_\varepsilon > 0$  is a suitable constant independent of  $x$  and  $y$ . Moreover, the mapping  $x \rightarrow \bar{u}(x)$  can be chosen to be measurable.

Now let  $a > 0$  and  $x \in K$  be arbitrarily fixed. Using (4.17) and following the argument developed for Lemma 4.7 we can find  $u^\sharp \in \mathcal{U}$  such that denoting by  $Z^\sharp$  the solution to

$$\begin{cases} dZ_t^\sharp = (AZ_t^\sharp + J_n F(Z_t^\sharp, u_t^\sharp)) dt + J_n G(Z_t^\sharp, u_t^\sharp) dW_t^m, & t \geq 0, \\ Z_0^\sharp = J_n x, \end{cases}$$

and by  $\rho$  the stopping time  $\rho = \inf\{t \geq 0 : |\Pi Z_t^\sharp| > 1\} \wedge 1$ , we have,  $P$ -a.s. for all  $t \leq \rho$ ,

$$(4.18) \quad \sum_{i=1}^m |J_n \Pi G(Z_t^\sharp, u_t^\sharp) e_i|^2 \leq \varepsilon + c |\Pi Z_t^\sharp|^2, \\ \langle F(Z_t^\sharp, u_t^\sharp), \Pi Z_t^\sharp \rangle + \langle AZ_t^\sharp, \Pi Z_t^\sharp \rangle \leq \varepsilon + c |\Pi Z_t^\sharp|^2.$$

Again similar to Lemma 4.7, let  $\psi_\varepsilon(r) = 2 \int_0^r \int_0^s [\varepsilon^{-1} \sigma] \wedge 1 d\sigma ds$  for  $r > 0$ ,  $\psi_\varepsilon(r) = 0$  for  $r \leq 0$ . We notice that  $\psi_\varepsilon(r) \nearrow (r^+)^2$  and  $\psi'_\varepsilon(r) \nearrow 2r^+$ , uniformly on bounded

subsets of  $\mathbb{R}_+$  and  $\psi''_\varepsilon(r) \nearrow 2I_{]0,\infty[}(r)$ . Moreover,  $\psi_\varepsilon(r) \leq r^2$ ,  $\psi'_\varepsilon(r) \leq 2|r|$ , and  $\psi''_\varepsilon(r) \leq 2$  for all  $r \in \mathbb{R}, \varepsilon \in ]0, 1]$ . Finally,  $\psi_\varepsilon$  is of class  $C^2$ .

We now compute  $d_t \phi_\varepsilon(|\Pi Z_t^\#|)$ , by Itô's rule, integrate over the interval  $[0, t \wedge \rho]$ , and compute the mean value. Then letting  $\varepsilon \searrow 0$ , we obtain

$$\mathbb{E}(|\Pi Z_{t \wedge \rho}^\#|^2) = \mathbb{E} \int_0^{t \wedge \rho} \left[ \langle 2\Pi Z_s^\#, AZ_s^\# \rangle + 2\langle \Pi Z_s^\#, J_n F(Z_s^\#, u_s^\#) \rangle + \sum_{i=1}^m |\Pi J_n G(Z_s^\#, u_s^\#) e_i|^2 \right] ds$$

and by (4.18)

$$\mathbb{E}|\Pi Z_{t \wedge \rho}^\#|^2 \leq c\varepsilon + c\mathbb{E} \int_0^{t \wedge \rho} |\Pi Z_{t \wedge \rho}^\#|^2 ds \leq c\varepsilon + c\mathbb{E} \int_0^t |\Pi Z_{t \wedge \rho}^\#|^2 ds.$$

Thus, by Gronwall's lemma we get  $\mathbb{E}|\Pi Z_{t \wedge \rho}^\#|^2 \leq c\varepsilon$  and taking the limit as  $t \nearrow \infty$ ,  $\mathbb{E}|\Pi Z_\rho^\#|^2 \leq c\varepsilon$ . Finally, since again as in Lemma 4.7,

$$\mathbb{E} \int_\rho^a |\Pi Z_s^\#|^2 ds \leq c\sqrt{P(\rho < a)} \leq c\sqrt{P(|\Pi Z_\rho^\#| = 1)} \leq c\sqrt{\mathbb{E}|Z_\rho^\#|^2} \leq c\sqrt{\varepsilon},$$

we have  $\mathbb{E} \int_0^a |\Pi Z_s^\#|^2 ds \leq c\sqrt{\varepsilon}$ . Since  $a > 0$   $n$  and  $m$  are arbitrary, the claim follows by Lemma 4.5.  $\square$

### 4.3. Sufficiency of the necessary condition by comparison principle.

In this section we assume that (3.12) holds and, following [24], that there exists a self-adjoint bounded positive operator  $B \in L(H)$  with  $BH \subset D(A^*)$  and a constant  $C_0 > 0$  such that

$$(4.19) \quad -A^*B + C_0B \geq I$$

(for instance, if  $A$  is selfadjoint it is enough to take  $B = (-A^* + I)^{-1}$ ). Condition (4.19) was first introduced in [11] where many examples are given. Moreover a general sufficient condition for (4.19) to hold was given in [29].

We define  $H_{-1}$  as the completion of  $H$  under the norm  $\|x\|_{-1}^2 \doteq \langle Bx, x \rangle = |B^{1/2}x|$ .

**DEFINITION 4.15.** *We say that a continuous function  $u : [0, T] \times H \rightarrow \mathbb{R}$  is B-lower semicontinuous if  $u(t, x) \leq \liminf_{n \rightarrow \infty} u(t_n, x_n)$  whenever  $t_n \rightarrow t$  and  $x_n \rightarrow x$ ,  $Bx_n \rightarrow Bx$ . Clearly,  $u : [0, T] \times H \rightarrow \mathbb{R}$  is B-upper semicontinuous whenever  $-u$  is B-lower semicontinuous.*

In this section we obtain sufficiency of the condition for general  $K$  (in some cases for nonconvex  $K$ ), but the price to pay is that we have to assume, beside the trace class condition included in (3.12), that the two following restrictive requirements hold:

$$(4.20) \quad |G(x, u) - G(y, u)|_{L_2(\Xi, H)} \leq C\|x - y\|_{-1}, \quad \text{for all } x, y, \text{ in } H, \text{ for all } u \in U,$$

$$(4.21) \quad d_K^2 \text{ is B-lower semicontinuous.}$$

**PROPOSITION 4.16.** *Condition (4.21) is satisfied, for instance, in the following cases.*

1.  $K$  is locally compact.
2.  $K$  is convex.
3.  $K$  is a linear subspace of  $H$  with  $K^\perp \subset B^{1/2}H$ .

*Proof.* To prove the first assertion, we assume that  $K$  is locally compact and, by contradiction, that there exist  $x \in K$  and  $\alpha > 0$  such that

$$\liminf_{y \rightarrow x} d_K(y) \leq d_K(x) - 4\alpha < d_K(x).$$

Consider a sequence  $y_n \rightarrow x$  with  $\liminf_{y \rightarrow x} d_K(y) = \lim_n d_K(y_n)$  and observe that there exist  $\Pi_K(y_n) \in K$  and  $\Pi_K(x) \in K$  such that

$$d_K(y_n) = |y_n - \Pi_K(y_n)|, \quad d_K(x) = |x - \Pi_K(x)|.$$

So for any  $n$  large enough

$$(4.22) \quad |y_n - \Pi_K(y_n)| \leq d_K(x) - 3\alpha < d_K(x).$$

The sequence  $\Pi_K(y_n)$  is bounded in the locally compact set  $K$ ; hence one can extract a subsequence (again similarly denoted) which converges to some  $z \in K$ . Hence for any  $n$  large enough

$$(4.23) \quad |z - \Pi_K(y_n)| \leq \alpha.$$

From (4.22) and (4.23), we obtain for  $n$  large enough

$$|y_n - z| \leq d_K(x) - 2\alpha < d_K(x).$$

Namely,  $y_n$  belongs to the closed ball  $B(z, d_K(x) - 2\alpha)$  which is convex and (strongly) closed, hence weakly closed by Mazur's theorem. Consequently,  $x$  belongs to this ball because it is the weak limit of  $y_n$ . Hence

$$|x - z| \leq d_K(x) - 2\alpha < d_K(x) \quad \text{and} \quad z \in K$$

which is a contradiction with the very definition of  $d_K(x)$ . As far as the second assertion is concerned, we remark that if  $K$  is convex, then  $d_K^2$  is convex as well. Thus the epigraph of  $d_K^2$  is convex and closed (since  $d_K^2$  is continuous), thus weakly closed.

Finally if  $K$  is linear with  $K^\perp \subset B^{1/2}H$ , then we get

$$d_K(x) = |\Pi_{K^\perp} x| = \sup_{y \in H, |y|=1} \langle \Pi_{K^\perp} x, y \rangle = \sup_{y \in H, |y|=1} \langle B^{1/2} x, B^{-1/2} \Pi_{K^\perp} y \rangle.$$

Since  $B^{-1/2} \Pi_{K^\perp}$  is bounded it is immediate to verify that  $d_K$  is continuous with respect to the norm  $\|\cdot\|_{-1}$ .  $\square$

We introduce here a weak (in probabilistic sense) notion of  $\varepsilon$ -viability.

**DEFINITION 4.17.** *A closed subset  $K \subset H$  is weakly  $\varepsilon$ -viable if for all  $x \in K$  it holds that*

$$(4.24) \quad \inf_{\mathfrak{U}} \mathbb{E} \left[ \int_0^{+\infty} e^{-Ct} d_K^2(X_t^{x, \mathfrak{U}}) dt \right] = 0,$$

where the infimum is computed over all settings  $\mathfrak{U} = (\Omega, \mathcal{E}, \mathbb{P}, \mathbb{F}, u)$ , where  $(\Omega, \mathcal{E}, \mathbb{P})$  is a probability space;  $\mathbb{F} = \{\mathcal{F}_t\}_{t \geq 0}$  is a filtration in  $\mathcal{E}$  satisfying the usual conditions:  $\{W_t\}_{t \geq 0}$  is a cylindrical,  $\Xi$ -valued  $\mathbb{F}$ -Wiener process, and  $u$  is an  $\mathbb{F}$ -progressively measurable process with values in  $U$ . Moreover, by  $X^{x, \mathfrak{U}}$  we denote the solution of (4.26) with respect to the setting  $\mathfrak{U}$ .

REMARK 4.18. Since, as it is stated in Remark 3.7, under assumptions (3.12), the filtrations  $\mathbb{F}$  and  $\mathbb{F}^m$  do not play any role in the proof of Theorem 3.5. Then the argument of the proof of Theorem 3.5 still yields, taking into account the simplifications stated in Remark 3.7, that if  $K$  is weakly  $\varepsilon$ -viable, then  $d_K^2$  is a viscosity supersolution of the HJB equation (3.13) for a suitable constant  $C > 0$  (large enough).

PROPOSITION 4.19. Assume that (3.12), (4.19), (4.20), and (4.21) hold; then  $K$  is weakly  $\varepsilon$ -viable if and only if there exists  $C > 0$  (large enough) such that  $d_K^2$  is a viscosity supersolution of the HJB equation (3.13).

*Proof.* We start by noticing that it is easy to verify, just by comparing the two definitions, that if a function  $u : H \rightarrow \mathbb{R}$  is a viscosity supersolution of the HJB equation (3.13) in the sense of Definition 3.4 and it is B-lower semicontinuous, then it is a (stationary) viscosity supersolution of the parabolic HJB equation

$$(4.25) \quad \begin{aligned} & -\frac{\partial}{\partial t}V(t, x) + \Psi(x, V(x), D_x V(t, x), D_x^2 V(t, x)) - \langle Ax, D_x V(t, x) \rangle = 0, \\ & \quad \quad \quad t \in [0, T] \ x \in H, \\ & \text{with } \Psi(x, v, p, X) = - \inf_{u \in U} \left( \frac{1}{2} \sum_{i=1}^{\infty} \langle XG(x, u)e_i, G(x, u)e_i \rangle + \langle F(x, u), p \rangle \right) \\ & \quad \quad \quad + Cv - d_K^2(x) \end{aligned}$$

in the sense of in [24, Definition 2.2].

We fix an arbitrary  $T > 0$  and define

$$V(t, x) = \inf_{\mathfrak{U}} \mathbb{E} \int_t^T e^{-Cs} d_K^2(X_s^{t,x,\mathfrak{U}}) ds, \quad x \in H,$$

where  $\mathfrak{U} = (\Omega, \mathcal{E}, \mathbb{P}, \mathbb{F}, u)$  is as in Definition 4.17 and  $X^{t,x,\mathfrak{U}}$  is the solution of

$$(4.26) \quad \begin{cases} d_s X_s = (AX_s + F(X_s, u_s)) ds + G(X_s, u_s) dW_s, & s \in [t, T], \\ X_t = x, \end{cases}$$

corresponding to  $\mathfrak{U}$ .

By [24, Theorem 4.1]  $V$  is the unique  $B$ -continuous viscosity solution of

$$\begin{cases} -\frac{\partial}{\partial t}V(t, x) + \Psi(x, V(x), D_x V(t, x), D_x^2 V(t, x)) - \langle Ax, D_x V(t, x) \rangle = 0, \\ V(T, x) = 0, \end{cases} \quad t \in [0, T] \ x \in H,$$

with quadratic growth. Since  $d_K^2$  is a viscosity supersolution of (4.25) by the comparison result in [24, Theorem 3.3] (see also [24, Remark 3.3], notice that in [24, section 3] it is also shown that the assumptions of [24, Theorem 3.3] are satisfied in the present case); we get  $V(t, x) \leq d_K^2(x)$ ,  $t \in [0, T]$ . In particular  $V(0, x) = 0$  for  $x \in K$ . Due to the arbitrariness of  $T$  we deduce that for all  $\varepsilon > 0$  and all  $T > 0$  there exists some  $\mathfrak{U} = (\Omega, \mathcal{E}, \mathbb{P}, \mathbb{F}, u)$  such that

$$\mathbb{E} \int_0^T e^{-Cs} d_K^2(X_s^{x,\mathfrak{U}}) ds \leq \varepsilon.$$

The claim follows since, by Lemma 2.2, for an arbitrary  $C > C_2$  (see Lemma 2.2),

$$\mathbb{E} \int_T^\infty e^{-Cs} d_K^2(X_s^{x,\mathfrak{U}}) ds \leq e^{-(C-C_2)T} / (C - C_2). \quad \square$$

**5. Application.** Finally, we briefly show how our results can be applied to a general semilinear controlled heat equation with multiplicative white noise. Namely we consider, for  $t \in [0, T]$ ,  $\xi \in [0, 1]$ ,

$$(5.1) \quad \begin{cases} d_t X^u(t, \xi) = \frac{\partial^2}{\partial \xi^2} X^u(t, \xi) dt + f(X^u(t, \xi), u(t, \xi)) dt + \sigma(X^u(t, \xi), u(t, \xi)) \frac{\partial}{\partial t} \mathcal{W}(t, \xi), \\ X^u(t, 0) = X^u(t, 1) = 0, \\ X^u(t_0, \xi) = x_0(\xi), \end{cases}$$

where  $\mathcal{W}$  is a space-time white noise on  $[0, T] \times [0, 1]$ . The functions  $f$  and  $\sigma$  are continuous functions defined on  $\mathbb{R}^2$  with values in  $\mathbb{R}$ . Moreover we assume that  $f$  and  $g$  are Lipschitz in  $x$  and uniformly in  $u$ , if  $u$  varies on a bounded subset of  $\mathbb{R}$ .

Finally, we assume that  $x_0 \in L^2([0, 1])$  and that  $U$  is a subset of  $\{u \in L^2([0, 1]) : |u(\xi)| \leq \delta \text{ for almost all } \xi \in [0, 1]\}$  for a suitable  $\delta > 0$ .

To rewrite the above problem in the abstract way we set  $H = \Xi = L^2([0, 1])$ .

We set  $\{W_t : t \geq 0\}$  to be a cylindrical Wiener process in  $L^2([0, 1])$  (that is formally  $(\partial \mathcal{W} / \partial t)(t, \xi) = \sum_{i=1}^{\infty} \beta_i(t) e_i(\xi)$ , where  $\{\beta_i : i \in \mathbb{N}\}$  is a sequence of independent, real valued Brownian motions and  $\{e_i : i \in \mathbb{N}\}$  is an orthonormal basis in  $H$ ).

Moreover, we define the operator  $A$  with domain  $\mathcal{D}(A)$  by

$$\mathcal{D}(A) = H^2([0, 1]) \cap H_0^1([0, 1]), \quad (Ay)(\xi) = \frac{\partial^2}{\partial \xi^2} y(\xi) \quad \text{for all } y \in \mathcal{D}(A),$$

where  $H^2([0, 1])$  and  $H_0^1([0, 1])$  are the usual Sobolev spaces and

$$F(x, u)(\xi) = f(\xi, x(\xi), u(\xi)), \quad [G(x, u)z](\xi) = \sigma(\xi, x(\xi), u(\xi))z(\xi)$$

for all  $x, z \in L^2([0, 1])$ ,  $u \in L^2([0, 1])$  and a.a.  $\xi \in [0, 1]$ .

With this setting and under the above hypotheses, the general assumptions in section 2 are satisfied; see [13, section 11.2.1] (to ensure that (2.5) holds it is enough to choose a basis in  $L^2([0, 1])$  given by bounded functions). Moreover, it is well known that the operator  $A$  is selfadjoint.

**5.1. Viability of balls.** Let  $K$  be the unit ball in  $H$  and  $U = \{u \in L^2([0, 1]) : |u(\xi)| \leq 1 \text{ for a.e. } \xi \in [0, 1]\}$ . Moreover, we assume that  $f$  is affine,  $f(x, u) = f_0(x)u + f_1(x)$  with  $f_0$  and  $f_1$  continuous and bounded. The coefficient  $\sigma$  is also supposed to be linear,  $\sigma(x, u) = \sigma_0(x)u$ , where  $\sigma_0$  is a continuous and bounded function  $\mathbb{R} \rightarrow \mathbb{R}$ . Finally we suppose that  $\sigma_0(x) \neq 0$  if  $x \neq 0$ .

In this case the first condition in (4.2) takes the form

$$(5.2) \quad \sigma_0(x(\xi))u(\xi)x(\xi) = 0 \text{ for a.e. } \xi \in [0, 1] \text{ and for all } x \in \mathcal{D}(A) \text{ with } |x|_H = 1.$$

Thus  $u(\xi)x(\xi) = 0$ , for a.e.  $\xi \in [0, 1]$  and for all  $x \in \mathcal{D}(A)$  with  $|x|_H = 1$ .

Consequently, the second condition in (4.2) becomes

$$(5.3) \quad \int_0^1 f_1(x(\xi))x(\xi)d\xi - \int_0^1 (x'(\xi))^2 d\xi \leq 0$$

for all  $x \in \mathcal{D}(A)$  with  $|x|_H = 1$ . Resuming, (5.2) and (5.3) give the desired necessary and sufficient condition for the  $\varepsilon$ -viability of  $K$ .

**5.2. Viability of linear spaces.** Now let  $K$  be a finite dimensional linear subspace of  $D(A)$  and

$$U = \{u \in H^1([0, 1]) : |u|_{H^1([0, 1])} \leq r\}$$

for some  $r > 0$  (we notice that  $U$  is a compact subset of  $\{u \in L^2([0, 1]) : |u(\xi)| \leq \delta \text{ for all } \xi \in [0, 1]\}$  for  $\delta$  large enough).

Moreover, we assume that the functions  $f$  and  $g$  are convex in  $u$  (but not necessarily linear), where  $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}$  is assumed to be of class  $C^1(\mathbb{R}^2)$  satisfying  $|\partial\sigma(a, b)/\partial b| > 0$  for all  $a \in \mathbb{R}$  and  $b \in [-\delta, \delta]$ .

In this case the second relation in (4.15) becomes  $\sigma(x(\xi), u(\xi)) = 0$  for all  $\xi \in [0, 1]$ . Indeed, since  $\sigma$ ,  $x$ , and  $u$  are continuous, if we had  $\sigma(x(\xi), u(\xi)) \neq 0$  for some  $\xi \in [0, 1]$ , then the set  $\{\sigma(x, u)v : v \in L^2([0, 1])\}$  would automatically be an infinite dimensional subset of  $L^2([0, 1])$ . Thus, taking into account that  $|\partial\sigma/\partial u(x, u)| > 0$ , the above implies that for all  $a \in \mathbb{R}$  there exists a unique  $b = h(a) \in [-\delta, \delta]$  such that  $\sigma(a, h(a)) = 0$ . Moreover,  $h \in C^1(\mathbb{R}, [-\delta, \delta])$ .

We now observe that the first formula in (4.15) gives

$$(5.4) \quad f(x, h(x)) + x'' \in K \quad \text{for all } x \in K.$$

Resuming  $K$  is  $\varepsilon$ -viable if and only if the following two statements hold:

1. for all  $a \in \mathbb{R}$  there exists a (unique)  $b = h(a) \in [-\delta, \delta]$  such that  $\sigma(a, h(a)) = 0$ ,
2. for all  $x \in K$ ,  $h(x) \in U$  and  $f(x, h(x)) + x'' \in K$ .

**5.3. Nonconvex  $K$ .** We consider here a state equation with more regular diffusion and finite dimensional noise

$$(5.5) \quad \begin{cases} d_t X^u(t, \xi) = \frac{\partial^2}{\partial \xi^2} X^u(t, \xi) dt + f(X^u(t, \xi), u(t, \xi), v(t)) dt \\ \quad + \sum_{i=1}^N \sigma^i \left( \langle X^u(t, \xi), h^i(\xi) \rangle_{L^2([0, 1])}, v(t) \right) d\beta_t^i, \\ X^u(t, 0) = X^u(t, 1) = 0, \\ X^u(t_0, \xi) = x_0(\xi), \end{cases}$$

where  $\beta^1, \dots, \beta^N$  are independent real-valued Brownian. The functions  $f : \mathbb{R}^{2+M} \rightarrow \mathbb{R}$  and  $\sigma : \mathbb{R}^{1+M} \rightarrow \mathbb{R}$  are again continuous and we assume that  $f$  and  $g$  are Lipschitz in  $x$ , uniformly in  $u$  and  $v$ .

The control is now given by  $(u, v)$  with  $u \in L^2([0, 1])$  and  $v \in \mathbb{R}^M$ . We set  $U = \{(u, v) : |u|_{L^2([0, 1])} \leq 1, |v| \leq 1\}$ .

Finally, we assume that  $h_i \in H_0^1([0, 1])$ , and let  $K$  be a (possibly nonconvex) locally compact closed subset of  $L^2([0, 1])$ .

As above we let  $H = L^2([0, 1])$  and  $\mathcal{D}(A) = H^2([0, 1]) \cap H_0^1([0, 1])$ ,  $(Ay)(\xi) = \frac{\partial^2}{\partial \xi^2} y(\xi)$ .

We notice that  $A$  is selfadjoint and we can choose  $B = (I - A)^{-1}$ . Thus  $B^{-1/2} = (I - A)^{1/2}$  with  $D(B^{-1/2}) = H_0^1$ . Moreover, since  $h_i \in D(B^{-1/2})$  the map  $x \rightarrow \langle x, h^i \rangle_{L^2([0, 1])}$  can be rewritten as  $x \rightarrow \langle B^{1/2}x, B^{-1/2}h^i \rangle_{L^2([0, 1])}$ . Thus if  $F$  is defined as in the previous sections and  $G(x, u, v) = (\sigma^1(\langle x, u \rangle_{L^2([0, 1])}, v), \dots, \sigma^N(\langle x, u \rangle_{L^2([0, 1])}, v))$ ,  $x, u \in L^2([0, 1])$ ,  $v \in \mathbb{R}^M$  it is easy to verify that the assumptions of Proposition 4.19 are verified. Thus applying Proposition 4.19 we obtain that the set  $K$  is  $\varepsilon$ -viable if and only if there exists  $C > 0$  (large enough) such that  $d_K^2$  is a viscosity supersolution of the HJB equation (3.13).

## REFERENCES

- [1] J.-P. AUBIN, *Viability Theory*. Birkhäuser Boston, Boston, 1992.
- [2] J.-P. AUBIN AND G. DA PRATO, *Stochastic viability and invariance*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 17 (1990) pp. 595–613.
- [3] J.-P. AUBIN AND G. DA PRATO, *Stochastic Nagumo's viability theorem*, Stochastic Anal. Appl., 13 (1995) pp. 1–11.
- [4] J.-P. AUBIN AND G. DA PRATO, *The viability theorem for stochastic differential inclusion*, Stochastic Anal. Appl., 16 (1998) 1–15.
- [5] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Systems Control Found. Appl. 2, Birkhäuser Boston, Boston, 1990.
- [6] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
- [7] R. BUCKDAHN, S. PENG, M. QUINCAMPOIX, AND C. RAINER, *Existence of stochastic control under state constraints*, C. R. Acad. Sci. Paris Sér. I Math. 327, (1998), pp. 17–22.
- [8] R. BUCKDAHN, P. CARDALIAGUET, AND M. QUINCAMPOIX, *A representation formula for the mean curvature motion*, SIAM J. Math. Anal., 33 (2001), pp. 827–846.
- [9] R. BUCKDAHN, M. QUINCAMPOIX, C. RAINER, AND R. RASCANU, *Viability of moving sets for stochastic differential equation*, Adv. Differential Equations, 7 (2002), pp. 1045–1072.
- [10] R. BUCKDAHN, M. QUINCAMPOIX, C. RAINER, AND R. RASCANU, *Stochastic control with exit time and constraints, application to small time attainability of sets*, Appl. Math. Optim., 49 (2004), pp. 99–112.
- [11] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations in infinite dimensions, IV. Hamiltonians with unbounded linear terms*, J. Funct. Anal., 90 (1990), pp. 237–283.
- [12] G. DA PRATO AND J. ZABCZYK, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.
- [13] G. DA PRATO AND J. ZABCZYK, *Ergodicity for Infinite-Dimensional Systems*, London Math. Soc. Lecture Note Ser. 229, Cambridge University Press, Cambridge, UK, 1996.
- [14] E. B. DAVIES, *One-Parameter Semigroups*, Academic Press, London, 1980.
- [15] N. EL KAROUI, D. HU NGUYEN, AND M. JEANBLANC-PIQUÉ, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics, 20 (1987), pp. 169–219.
- [16] M. FUHRMAN AND G. TESSITORE, *Nonlinear Kolmogorov equations in infinite dimensional spaces: The backward stochastic differential equations approach and applications to optimal control*, Ann. Probab., 30 (2002), pp. 1397–1465.
- [17] S. GAUTIER AND L. THIBAUT, *Viability for constrained stochastic differential equations*, Differential Integral Equations, 6 (1993), pp. 1395–1414.
- [18] D. GATAREK, *Existence of optimal controls for stochastic evolution systems*, in Control of Partial Differential Equations, Lecture Notes in Pure and Appl. Math. 165, G. Da Prato et al., eds., Marcel Dekker, New York, 1994, pp. 81–86.
- [19] D. GATAREK AND J. SOBCZYK, *On the existence of optimal controls of Hilbert space-valued diffusions*, SIAM J. Control Optim., 32 (1994), pp. 170–175.
- [20] H. ISHII, *Viscosity solutions of nonlinear second-order partial differential equations in Hilbert spaces*, Comm. Partial Differential Equations, 18 (1993), pp. 601–650.
- [21] P. L. LIONS, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions, I. The case of bounded stochastic evolutions*, Acta Math., 161 (1988), pp. 243–278.
- [22] P. L. LIONS, *Viscosity solutions of fully nonlinear second order equations and optimal stochastic control in infinite dimensions, II. Optimal control of Zakai's equation*, in Stochastic Partial Differential Equations and Applications, II, Lecture Notes in Math. 1390, G. Da Prato and L. Tubaro eds., Springer, Berlin, 1989, pp. 147–170.
- [23] P. L. LIONS, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions, III. Uniqueness of viscosity solutions for general second-order equations*, J. Funct. Anal., 86 (1989), pp. 1–18.
- [24] D. A. KELOME, *Viscosity Solutions of Second-Order Equations in a Separable Hilbert Space and Applications to Stochastic Optimal Control*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 2002.
- [25] D. A. KELOME AND A. ŚWIĘCH, *Viscosity solutions of an infinite-dimensional Black-Scholes-Barenblatt equation*, Appl. Math. Optim., 47 (2003), 253–278.

- [26] M. NAGUMO, *Über die Lage der Integralkurven gewöhnlicher Differentialgleichungen*. Proc. Phys.-Math. Soc. Japan, 24 (1942), pp. 551–559.
- [27] T. NAKAYAMA, *Viability theorem for SPDE's including HJM framework*, J. Math. Sci. Univ. Tokyo, 11 (2004), pp. 313–324.
- [28] M. QUINCAMPOIX AND C. RAINER, *Stochastic control and compatible subsets of constraints*, Bull. Sci. Math., 129 (2005), pp. 39–55.
- [29] M. RENARDY, *Polar decomposition of positive operators and a problem of Crandall and Lions*, Appl. Anal., 57 (1995), pp. 383–385.
- [30] A. ŚWIECH, *“Unbounded” second order partial differential equations in infinite-dimensional Hilbert spaces*, Comm. Partial Differential Equations, 19 (1994), pp. 1999–2036.



# STOCHASTIC OPTIMAL CONTROL PROBLEMS AND PARABOLIC EQUATIONS IN BANACH SPACES\*

FEDERICA MASIERO†

**Abstract.** We consider stochastic optimal control problems in Banach spaces. These problems are related to nonlinear controlled equations with dissipative nonlinearity and are treated via the backward stochastic differential equation approach, which also allows us to solve, in a mild sense, Hamilton–Jacobi–Bellman equations in Banach spaces. We apply the results to controlled stochastic heat and wave equations with a cost functional that is well defined on continuous functions, and to delay equations in spaces of  $p$ -integrable functions.

**Key words.** stochastic optimal control, infinite dimensional stochastic processes, Banach spaces, backward stochastic differential equations, Hamilton–Jacobi–Bellman equations

**AMS subject classifications.** 93E20, 60H30, 60H15

**DOI.** 10.1137/050632725

**1. Introduction.** In this paper we develop an abstract theory for stochastic optimal control problems with cost functionals that are defined on a Banach space. This allows us to treat, for example, stochastic optimal control problems for heat equations with temperature control in a finite number of points: in an abstract setting this cost functional is well defined on the Banach space of continuous functions, but it is not well defined on the Hilbert space of square integrable functions. In many other situations arising from concrete models, the cost is naturally defined on a Banach space, but not on a Hilbert space; hence, the novelty of this paper is that we can consider such optimal control problems in a general framework and extend the results found in the existing literature. The control problems are treated via backward stochastic differential equations (BSDEs); a similar approach for stochastic optimal control problems in infinite dimensional Hilbert spaces is studied in [16].

First, we consider a stochastic differential equation, with additive noise, in a Banach space  $E$ :

$$(1.1) \quad \begin{cases} dX_\tau = AX_\tau d\tau + F(\tau, X_\tau) d\tau + GdW_\tau, & \tau \in [t, T], \\ X_t = x. \end{cases}$$

$W$  is a cylindrical Wiener process in a separable Hilbert space  $\Xi$ ,  $A$  generates a strongly continuous semigroup of linear operators,  $F : [0, T] \times E \rightarrow E$  is continuous, and there exists  $\eta \in \mathbb{R}$  such that  $A + F(\tau, \cdot) - \eta$  is dissipative on  $E$ , for every  $\tau \in [0, T]$ ;  $G$  is a linear operator from  $\Xi$  to  $H$ ,  $x \in E$  and  $0 \leq t \leq T$ . The Banach space is continuously and densely embedded in another real and separable Hilbert space  $H$ . We make suitable assumptions such that the stochastic convolution

$$W_A(\tau) = \int_0^\tau e^{(\tau-s)A} G dW_s$$

\*Received by the editors May 31, 2005; accepted for publication (in revised form) June 15, 2007; published electronically January 22, 2008.

<http://www.siam.org/journals/sicon/47-1/63272.html>

†Dipartimento di Matematica e Applicazioni, Università di Milano-Bicocca, via Cozzi, 53, 20125 Milano, Italy (federica.masiero@unimib.it).

is well defined as a Gaussian process in  $H$ . Moreover, we assume that the stochastic convolution  $W_A(\tau)$  admits an  $E$ -continuous version. It is well known that under these assumptions, (1.1) admits a unique mild solution, i.e., it satisfies  $\mathbb{P}$ -almost surely ( $\mathbb{P}$ -a.s.)

$$X_\tau = e^{(\tau-t)A}x + \int_t^\tau e^{(\tau-s)A}F(s, X_s)ds + \int_t^\tau e^{(\tau-s)A}GdW_s, \quad \tau \in [t, T];$$

see, e.g., [11], [12], and [13]. We want to remark that  $F$  is well defined only on the Banach space  $E$ , while on the Hilbert space  $H$ ,  $F$  is not even defined: this is a natural situation arising in many evolution equations, and the problem of solving (1.1) has been extensively studied, but not in relation to stochastic optimal control problems. To this end, we have to prove new results on the dependence of the solution  $X$  of (1.1) on the initial datum, in a suitable sense. We prove that if  $F$  is differentiable, then  $X$  is differentiable with respect to the initial datum. This result will be used while proving the so-called fundamental relation via BSDEs.

The controlled stochastic evolution equations that we study have the following form:

$$(1.2) \quad \begin{cases} dX_\tau = AX_\tau d\tau + F(\tau, X_\tau)d\tau + GdW_\tau + GR(\tau, X_\tau, u_\tau)d\tau, & \tau \in [t, T], \\ X_t = x, \end{cases}$$

where  $R$  is defined on  $[0, T] \times E \times U$ , and  $U$  denotes the set of control actions. In (1.2), we notice the occurrence of the operator  $G$  in the control term: this is a restriction imposed by our techniques, but it arises in the abstract formulation of controlled delay and wave equations. In the particular case  $R(\tau, x, u) = u$ , the term  $u_\tau d\tau + dW_\tau$  admits a natural interpretation as a control affected by noise.

Following the approach of [15], we consider the control problem in the weak formulation, that is, we have to minimize a cost functional over all admissible control systems (a.c.s.'s)  $(\Omega, \mathcal{F}, \mathcal{F}_\tau, \mathbb{P}, W, u, X)$ . Namely, to every a.c.s.  $(W, u, X^u)$  we associate a cost  $J$  given by

$$(1.3) \quad J(t, x, (W, u, X)) = \mathbb{E} \int_t^T g(s, X_s, u_s)ds + \mathbb{E}\phi(X_T),$$

where  $g : [0, T] \times E \times U \rightarrow \mathbb{R}$  and  $\phi : E \rightarrow \mathbb{R}$ . The value function of this control problem is defined by

$$(1.4) \quad V(t, x) = \inf_{(W, u, X^u) \text{ a.c.s.}} J(t, x, (W, u, X)).$$

In order to solve the optimal control problem, we introduce the BSDE

$$(1.5) \quad Y_\tau + \int_\tau^T Z_\sigma dW_\sigma = \phi(X_T) + \int_\tau^T \psi(\sigma, X_\sigma, Z_\sigma)d\sigma, \quad \tau \in [t, T];$$

see, e.g., [22]. The solution of this equation is a pair of predictable processes  $(Y, Z)$ , taking values in  $\mathbb{R} \times \Xi^*$ .  $\phi$  is the final cost, and  $\psi : [0, T] \times E \times \Xi^* \rightarrow \mathbb{R}$  is the Hamiltonian function defined in a classical way:

$$(1.6) \quad \psi(\tau, x, z) = \inf_{u \in U} \{g(\tau, x, u) + zR(\tau, x, u)\}.$$

We recall that the process  $X$  takes values in the Banach space  $E$ . It is trivial to deduce existence and uniqueness of the solution of (1.5). The fundamental relation via the BSDE is

$$(1.7) \quad Y_t = J(t, x, u) + \mathbb{E} \int_t^T [\psi(\sigma, X_\sigma, Z_\sigma) - Z_\sigma R(\sigma, X_\sigma, u_\sigma) - g(\sigma, X_\sigma, u_\sigma)] d\sigma.$$

Thus a control  $u$  such that

$$u_\tau \in \Gamma(\tau, x, z) := \{u \in U : g(\tau, x, u) + zR(\tau, x, u) = \psi(\tau, x, z)\}$$

is optimal. In order to characterize the optimal control via a feedback law, we need to identify the process  $Z$  with some function of the process  $X$ . Namely, we define the function  $v(t, x)$  as  $Y(t, t, x)$ , where  $Y(t, t, x)$  is the solution  $Y$  of the BSDE, considered at time  $t$  and depending on the Markov process  $X$ , which is a solution to an equation with the same structure as (1.1), and with initial condition  $x$  given at the initial time  $t$ . We prove that the value function  $V(t, x)$ , defined in (1.4), is equal to  $v(t, x)$ . Our goal is to prove that, when  $v(t, x)$  is differentiable with respect to  $x$ ,  $Z_t = \nabla v(t, X_t)G$ . For such an identification,  $Y(t, t, x)$  has to be differentiable with respect to  $x$ , and this can be proved by requiring the final cost  $\phi$  and the Hamiltonian  $\psi$  to be differentiable with respect to  $x$ , and by using the fact that the process  $X$  is differentiable with respect to the initial datum  $x$ . Differentiability assumptions on the Hamiltonian and on the final cost are a genuine restriction; but the identification of  $Z_t$  with  $\nabla v(t, X_t)G$  allows us to characterize the optimal control via a feedback law. When  $X$  evolves in a Hilbert space  $H$ , in the more general case of multiplicative noise in (1.1) (that is,  $G$  not necessarily constant),  $Z_t$  is identified with  $\nabla v(t, X_t)G(X_t)$  by means of the Malliavin calculus; see [16]. In this paper we deal with equations evolving in a Banach space, so we cannot make direct use of Malliavin calculus. By an idea taken from [2], we can prove that  $Z_t = \nabla v(t, X_t)G$  by avoiding the use of Malliavin calculus.

Moreover, by using the BSDE approach we are able to solve the associated Hamilton–Jacobi–Bellman equation

$$(1.8) \quad \begin{cases} \frac{\partial v}{\partial t}(t, x) = -\mathcal{A}_t v(t, x) - \psi(t, x, \nabla v(t, x)G), & t \in [0, T], x \in E, \\ v(T, x) = \phi(x). \end{cases}$$

We can prove the existence and uniqueness of a mild solution to (1.8), i.e., a function  $v(t, x)$ , which is continuous and Gâteaux differentiable with respect to  $x$  and satisfies

$$v(t, x) = P_{t,T}[\varphi](x) + \int_t^T P_{t,s}[\psi(s, \cdot, \nabla v(s, \cdot)G)](x) ds, \quad t \in [0, T], x \in E.$$

For every  $t \leq \tau \leq T$ , we denote by  $P_{t,\tau}$  the transition semigroup corresponding to the Markov process  $X(\tau, t, x)$ ,  $t \leq \tau \leq T$ , which is solution of (1.1), i.e., for a continuous and bounded function  $\varphi : E \rightarrow \mathbb{R}$ ,  $P_{t,\tau}[\varphi](x) = \mathbb{E}\varphi(X(\tau, t, x))$ . It turns out that the mild solution  $v(t, x)$  of the Hamilton–Jacobi–Bellman equation (1.8) is equal to  $Y(t, t, x)$ .

For the solution of Hamilton–Jacobi–Bellman equations via a BSDE, we cite [26], where the finite dimensional case is treated, and [16], where existence and uniqueness of a mild solution are proved for a Hamilton–Jacobi–Bellman equation on an infinite

dimensional Hilbert space. For Hamilton–Jacobi–Bellman equations on an infinite dimensional Hilbert space, the notion of a viscosity solution has been studied by many authors; we refer the reader to the general reference [9] and to the fundamental papers [6], [19], and [20]. The class of equations that can be treated is much more general than in our case; however, none of the known results guarantee differentiability of the solution. We are not aware of other results in the literature about the solution of stochastic partial differential equations of parabolic type in Banach spaces, apart from the papers [7] and [8], where viscosity solutions are presented for first order partial differential equations. Here we require more regularity on  $\phi$  and  $\psi$ , and we solve the equation in a stronger sense.

We apply the results to some specific models: we treat controlled heat equations in one space dimension with Dirichlet and Neumann boundary conditions, with the cost functional defined on the Banach space of continuous functions. The stochastic heat equation is nonlinear, and the nonlinear term is dissipative in the space of continuous functions, and it cannot be naturally extended to the whole Hilbert space of square integrable functions. Stochastic reaction diffusion equations have been extensively studied; we cite [29] as a general reference. In the papers [3] and [4], where more general reaction diffusion equations are studied in space dimension greater than 1, even when the equation is solved in the space of continuous functions, the related costs are defined in the Hilbert space of square integrable functions. We also cite [18] and [23], where nonlinear reaction diffusion equations with a nonlinear term satisfying polynomial growth conditions are studied. We remark that it is more natural to treat nonlinear terms with polynomial growth conditions in the space of continuous functions rather than in the space of square integrable functions. Indeed, in the first case the nonlinear term is defined on the whole space, while on the contrary it is easy to verify that this is not the case in the space of square integrable functions. We also study stochastic delay equations, with the cost defined on the space of  $p$ -integrable functions, and finally, stochastic wave equations in one space dimension with the cost functional defined on the Banach space of continuous functions; to this end we study the abstract wave equation in a Besov space, which is continuously and densely embedded in the space of continuous functions.

The paper is organized as follows. In section 2 we introduce some notations and recall some results that are needed throughout the paper. Section 3 is devoted to the study of the regularity of the solution of (1.1) with respect to the initial datum and to the identification of  $Z$ . In section 4 we study the regularity of the solution of a forward-backward system with respect to the initial datum in the forward equation. In section 5 we formulate the stochastic optimal control problem and solve it when the final cost and the running cost have polynomial growth with respect to  $x$ . In section 6 we solve, in a mild sense, the Hamilton–Jacobi–Bellman equation, and in section 7 we apply the results to some concrete models.

**2. Preliminaries and notations.** We list some notations that are used in the paper. If  $E$  and  $K$  are Banach spaces,  $L(E, K)$  denotes the space of bounded linear operators from  $E$  to  $K$ , endowed with the usual operator norm. By  $E^*$  we denote the dual space of  $E$ , and the duality product between  $E^*$  and  $E$  is denoted by  $\langle \cdot, \cdot \rangle_{E^*, E}$ . We will always deal with real separable Banach spaces.

The letters  $\Xi$  and  $H$  will always denote real and separable Hilbert spaces with scalar product  $\langle \cdot, \cdot \rangle$ .  $L_2(\Xi, H)$  is the space of Hilbert–Schmidt operators from  $\Xi$  to  $H$ , endowed with the Hilbert–Schmidt norm.

We specify in what sense we consider differentiability between Banach spaces.

Following [16, section 2.2], if  $E$  and  $K$  are two Banach spaces, we say that a function  $f : E \rightarrow K$  belongs to the class  $\mathcal{G}^1(E, K)$  if  $f$  is continuous and Gâteaux differentiable on  $E$  and if the gradient  $\nabla f : E \rightarrow L(E, K)$  is strongly continuous; that is, for every direction  $h \in E$  the map  $\nabla f(\cdot)h : E \rightarrow K$  is continuous. We say that  $f : [0, T] \times E \rightarrow K$  is in  $\mathcal{G}^{0,1}([0, T] \times E, K)$  if  $f$  is continuous and Gâteaux differentiable with respect to every  $e \in E$  on  $[0, T] \times E$  and the gradient  $\nabla f : [0, T] \times E \rightarrow L(E, K)$  is strongly continuous. If  $F$  is another Banach space, we say that  $f : E \times F \rightarrow K$  belongs to  $\mathcal{G}^{1,1}(E \times F, K)$  if it is continuous, if it is Gâteaux differentiable on  $E \times F$ , and if  $\nabla f : E \times F \rightarrow L(E \times F, K)$  is strongly continuous. If  $K = \mathbb{R}$  we write, respectively,  $\mathcal{G}^1(E)$ ,  $\mathcal{G}^{0,1}([0, T] \times E)$ , and  $\mathcal{G}^{1,1}(E \times F)$  instead of  $\mathcal{G}^1(E, \mathbb{R})$ ,  $\mathcal{G}^{0,1}([0, T] \times E\mathbb{R})$ , and  $\mathcal{G}^{1,1}(E \times F, \mathbb{R})$ .

From now on, by  $(\Omega, \mathcal{F}, \mathbb{P})$  we mean a complete probability space endowed with a filtration  $\{\mathcal{F}_t, t \geq 0\}$  satisfying the usual conditions. By a cylindrical Wiener process defined on  $(\Omega, \mathcal{F}, \mathbb{P})$ , and with values in a Hilbert space  $\Xi$ , we mean a family  $\{W_t, t \geq 0\}$  of linear mappings  $\Xi \rightarrow L^2(\Omega)$  such that for every  $\xi, \eta \in \Xi$ ,  $\{W_t\xi, t \geq 0\}$  is a real Wiener process and  $\mathbb{E}(W_t\xi \cdot W_t\eta) = \langle \xi, \eta \rangle_\Xi$ . In the following,  $\{W_t, t \geq 0\}$  is a cylindrical Wiener process adapted to the filtration  $\{\mathcal{F}_t, t \geq 0\}$ . We remember that the natural filtration of a Wiener process, augmented with the family of  $\mathbb{P}$ -null sets, satisfies the usual conditions.

Also, a cylindrical Wiener process can be seen as a  $Q$ -Wiener process with values in another Hilbert space  $\Xi_1$ . Let us recall an explicit construction that will be used in the following. Let  $\Xi_1$  be an arbitrary Hilbert space such that  $\Xi$  is continuously embedded in  $\Xi_1$  and the embedding  $i : \Xi \hookrightarrow \Xi_1$  is of Hilbert–Schmidt type with its image dense in  $\Xi_1$ . Let  $Q = ii^*$ .  $Q_1$  is a positive, symmetric trace class operator. Consider  $(\eta_k)_{k \geq 1}$ , a complete orthonormal system in  $\Xi_1$ , and a sequence of positive real numbers  $(\lambda_k)_{k \geq 1}$  such that

$$Q\eta_k = \lambda_k\eta_k.$$

Fix  $\xi_k = \frac{i^*\eta_k}{\sqrt{\lambda_k}}$  as a complete orthonormal system in  $\Xi$ . We can consider a  $Q$ -Wiener process  $W_t^1$ ,  $t \geq 0$ , taking values in  $\Xi_1$ , which admits the following representation:

$$W_t^1 = \sum_{k=1}^{\infty} \sqrt{\lambda_k} \beta_k(t) \eta_k,$$

where  $\beta_k(t)$ ,  $k \geq 1$ , are real, independent standard Wiener processes. Moreover,

$$\beta_k(t) = \frac{\langle W_t^1, \eta_k \rangle_{\Xi_1}}{\sqrt{\lambda_k}}.$$

For every  $\eta \in \Xi_1$  we set

$$\langle W_t^1, \eta \rangle_{\Xi_1} = W_t^{i^*\eta},$$

and for every  $\xi \in \Xi$ ,

$$W_t^\xi = \sum_{k=1}^{\infty} \beta_k(t) \langle \xi, \xi_k \rangle_\Xi$$

and  $\beta_k(t) = W_t^{\xi_k}$ ,  $k \geq 1$ .

Next we define a class of processes with values in a Banach space  $E$ . For every  $0 \leq t < T$ ,  $\mathcal{H}^p([0, T], E)$  is the space of predictable processes  $(Y_\tau)_{\tau \in [t, T]}$  admitting a continuous version and such that

$$\mathbb{E} \sup_{\tau \in [t, T]} \|Y_\tau\|_E^p < \infty.$$

We recall some results on BSDEs. Consider a BSDE of the form

$$(2.1) \quad Y_\tau + \int_\tau^T Z_\sigma dW_\sigma = \eta + \int_\tau^T \psi_\sigma d\sigma, \quad \tau \in [0, T],$$

in  $(\Omega, \mathcal{F}, \mathbb{P})$ .  $\{W_t, t \geq 0\}$  is a cylindrical Wiener process, and the filtration  $\{\mathcal{F}_t, t \geq 0\}$  is the natural filtration of the Wiener process augmented in the usual way.  $Y$  is a process taking values in  $\mathbb{R}$ , and  $Z$  is a process with values in  $\Xi^*$ , the dual of the separable Hilbert space  $\Xi$ .  $\eta$  is an  $\mathcal{F}_T$ -measurable random variable in  $L^2(\Omega, \mathbb{R})$ , and  $\psi_\sigma$  is an adapted square integrable process satisfying

$$(2.2) \quad \mathbb{E} \int_t^T |\psi_\sigma|^2 d\sigma < \infty.$$

Under the previous assumptions, (2.1) admits a unique solution that is a pair of predictable processes  $(Y_\tau, Z_\tau)$ , taking values in  $\mathbb{R} \times \Xi^*$ , such that  $Y$  has continuous paths and

$$\mathbb{E} \sup_{\tau \in [0, T]} |Y_\tau|^2 + \mathbb{E} \int_t^T \|Z_\sigma\|_{\Xi^*}^2 d\sigma < \infty;$$

see, e.g., [25]. In the following we denote by  $\mathbb{K}_{cont}([0, T])$  the space of such processes.

**3. Stochastic evolution equations in Banach spaces.** In this section we collect some results on the existence and uniqueness of mild solutions for stochastic evolution equations in Banach spaces, and then we investigate the regular dependence of the solution on the initial datum.

In the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , we consider the following stochastic differential equation with values in a Banach space  $E$ :

$$(3.1) \quad \begin{cases} dX_\tau = AX_\tau d\tau + F(\tau, X_\tau) d\tau + GdW_\tau, & \tau \in [t, T], \\ X_t = x, \end{cases}$$

where  $x \in E$  and  $[t, T] \subset [0, T]$ . Throughout the paper, we assume that the Banach space  $E$  is separable and it is continuously and densely embedded in a Hilbert space  $H$ ;  $A$  is a linear operator in  $E$  with domain  $D(A)$ . We will need the following assumptions.

*Hypothesis 3.1.* We assume that either

1.  $A$  generates a  $C_0$ -semigroup in  $E$ , or
2.  $A$  is a sectorial operator in  $E$ .

We cite [21, p. 33] for a definition of sectorial operator. In both cases we denote by  $e^{tA}$ ,  $t \geq 0$ , the semigroup of bounded linear operators on  $E$  generated by  $A$  and we know that for  $t > 0$  the map  $t \mapsto e^{tA}$  is continuous and there exist  $M > 0$  and  $\omega \in \mathbb{R}$  such that  $\|e^{tA}\|_{L(E, E)} \leq Me^{\omega t}$  for all  $t \geq 0$ ; moreover, we assume that  $\|e^{tA}\|_{L(E, E)} \leq e^{\omega t}$  for all  $t \geq 0$  and some  $\omega \in \mathbb{R}$ . Finally, we assume that  $e^{tA}$ ,  $t \geq 0$ , admits an extension to a  $C_0$ -semigroup of bounded linear operators in  $H$ , still denoted by the same symbol,  $e^{tA}$ .

$W$  is a cylindrical Wiener process in another separable Hilbert space  $\Xi$ . We have to make assumptions on  $F$ ,  $G$ , and the stochastic convolution

$$W_A(\tau) = \int_0^\tau e^{(\tau-s)A} G dW_s.$$

Indeed, the Gaussian process  $W_A(\tau)$  is well defined in  $H$  when its covariance operator

$$Q_\tau = \int_0^\tau e^{sA} G G^* e^{sA^*} ds, \quad \tau \geq 0,$$

is a trace class operator on  $H$ .

*Hypothesis 3.2.*

1.  $F : [0, T] \times E \longrightarrow E$  is continuous and is bounded on bounded sets, and there exists  $\eta \in \mathbb{R}$  such that  $A + F(\tau, \cdot) - \eta$  is dissipative on  $E$  for every  $\tau \in [0, T]$ .
2.  $G$  is a linear operator from  $\Xi$  to  $H$  such that the operator  $Q_\tau$ , which turns out to be positive and symmetric, is of trace class for every  $\tau \in [0, T]$ .
3. The stochastic convolution  $W_A(\tau)$  admits an  $E$ -continuous version.

Throughout the paper, for every  $x \in E$ , we denote by  $\partial(\|x\|_E)$ , the subdifferential of the norm at  $x$ , the set of functionals  $x^* \in E^*$  such that  $\langle x^*, x \rangle_{E^*, E} = \|x\|_E$  and  $\|x^*\|_{E^*} = 1$ .

*Remark 3.3.* By [10, Proposition II.7.4], if a map  $f : D(f) \longrightarrow E$  is continuous and dissipative, then it is strongly dissipative, and, again by [10, p. 46], this is equivalent to asking that for every  $x, y \in E$  and  $x^* \in \partial\|x - y\|_E$ , we have  $\langle x^*, f(x) - f(y) \rangle_{E^*, E} \leq 0$ . Thus the map  $A + F(\tau, \cdot) - \eta I : D(A) \longrightarrow E$  is strongly dissipative.

*Remark 3.4.* It is well known that  $W_A$  can be regarded as a Gaussian random variable in  $L^2([0, T], H)$ ; see, e.g., [12, Theorem 5.2.]. We want to remark that it is a Gaussian random variable in  $C([0, T], E)$ , too. Indeed, let us denote  $B = C([0, T], E)$  and  $K = L^2([0, T], H)$ . We have to show that for every  $\Lambda \in B^*$ ,  $\Lambda W_A$  is a real Gaussian random variable. Since  $B$  is continuously and densely embedded in  $K$ , for every  $\Lambda \in B^*$  there exists a net  $\Lambda_\alpha \in K^*$  weakly\* convergent to  $\Lambda$ , i.e., such that  $\Lambda_\alpha x$  converges to  $\Lambda x$ , for every  $x \in B$ . By the Banach Steinhaus theorem the family  $\Lambda_\alpha$  is uniformly bounded in the norm of  $B^*$ . Let us denote  $M = \max(\|\Lambda\|, \sup_\alpha \|\Lambda_\alpha\|)$ . Since  $B$  is separable, in the dual space  $B^*$  the balls of radius  $M$  are weakly\* metrizable (see, e.g., [5, Theorem 5.1]), so there exists a subsequence  $(\Lambda_n)_n$  of  $(\Lambda_\alpha)_\alpha$  such that  $\Lambda_n x$  converges to  $\Lambda x$ . Since  $\Lambda_n W_A$  is a real Gaussian random variable, and  $\Lambda W_A$  is its pointwise limit, we conclude that  $\Lambda W_A$  is a real Gaussian random variable. Thus,  $W_A$  is a Gaussian random variable in  $C([0, T], E)$  and, in particular, it has finite moments of every order.

*Remark 3.5.* In Hypothesis 3.2, point 1, we make dissipative assumptions on  $F$ . The theory applies also to the more restrictive case of  $F$  Lipschitz continuous. In general this is a less interesting case also because, as it will be made clear in section 7.1, when we apply our theory to the heat equation, such nonlinearities are often well defined not only in the Banach space  $E$ , but also in the whole Hilbert space  $H$ , so from the point of view of the evolution equation, a Banach space approach is not so interesting. Nevertheless, when considering the associated control problem, also in the case of linear evolution equations or evolution equations with Lipschitz continuous nonlinearity, it can become necessarily a Banach space approach, e.g., when the cost is well defined on a Banach space but not on a Hilbert space.

We can state the following theorem; see, e.g., [11], [12, Theorem 7.13], and [13, Theorem 5.5.13]. In all these references only the autonomous case is treated, but the

extension to the nonautonomous case is immediate: in [13] Proposition 5.5.6 is applied in the proof of Theorem 5.5.13, and this proposition is about existence and uniqueness of a solution for a deterministic equation in the nonautonomous case. Moreover, the above mentioned theorems deal with a dissipative nonlinearity  $f$ : if  $f$  is Lipschitz continuous, on  $A$  it suffices to ask that it be the generator of a strongly continuous semigroup, but if, as in this paper,  $f$  is only dissipative, in order to guarantee existence and uniqueness of a mild solution of (3.1), we need to take  $A - \omega I$  dissipative; see, e.g., the proof of Theorem 7.13 in [12].

**THEOREM 3.6.** *Assume that Hypotheses 3.1 and 3.2 hold true. Then for every  $x \in E$  (3.1) admits a unique mild solution, that is, an adapted and continuous  $E$ -valued process satisfying  $\mathbb{P}$ -a.s.*

$$X_\tau = e^{(\tau-t)A}x + \int_t^\tau e^{(\tau-s)A}F(s, X_s)ds + \int_t^\tau e^{(\tau-s)A}GdW_s, \quad \tau \in [t, T].$$

If Hypothesis 3.1, point 1, holds true, then the mild solution has paths in  $C([t, T], E)$ ,  $\mathbb{P}$ -a.s. If Hypothesis 3.1, point 2, holds true, then the mild solution has paths in  $C([t, T], E) \cap L^\infty([t, T], E)$ ,  $\mathbb{P}$ -a.s.

We denote the solution by  $X(\tau, t, x)$ ,  $x \in E$ ,  $t \in [0, T]$ , and  $\tau \in [t, T]$ . For  $\tau \in [0, t]$  we define  $X(\tau, t, x) = x$ . By the Markov property it is well known that, for every  $0 \leq t \leq s \leq \sigma \leq T$ ,

$$X(\sigma, s, X(s, t, x)) = X(\sigma, t, x).$$

Now we want to investigate the dependence of the solution on the initial datum.

**PROPOSITION 3.7.** *Let Hypotheses 3.1 and 3.2 hold true. Then the solution  $X(\tau, t, x)$  of (3.1) is Lipschitz in  $x$  uniformly in  $\tau \in [0, T]$ ; more precisely,  $\mathbb{P}$ -a.s.*

$$\begin{aligned} & \|X(\tau, t, x_1) - X(\tau, t, x_2)\|_E \\ & \leq \max\left(1, e^{\eta(T-t)}\right) \|x_1 - x_2\|_E, \quad 0 \leq t \leq \tau \leq T, \quad x_1, x_2 \in E. \end{aligned}$$

*Proof.* Let  $X_1(\tau) = X(\tau, t, x_1)$  and  $X_2(\tau) = X(\tau, t, x_2)$ ,  $x_1, x_2 \in E$ . For  $i = 1, 2$  we set  $X_i^n(\tau) = nR(n, A)X_i(\tau)$ , where  $R(n, A) = (nI - A)^{-1}$  is the resolvent operator of  $A$ . Since  $X_i^n(\tau) \in \mathcal{D}(A)$  for every  $\tau \in [t, T]$ , and

$$\begin{aligned} X_i^n(\tau) &= e^{\tau A}nR(n, A)x_i + \int_t^\tau e^{(\tau-s)A}nR(n, A)F(s, X_i(s))ds \\ &\quad + \int_t^\tau e^{(\tau-s)A}nR(n, A)GdW_s, \end{aligned}$$

we get

$$\begin{aligned} \frac{d}{d\tau}(X_1^n(\tau) - X_2^n(\tau)) &= A(X_1^n(\tau) - X_2^n(\tau)) \\ &\quad + nR(n, A)[F(\tau, X_1(\tau)) - F(\tau, X_2(\tau))]. \end{aligned}$$

Thus, by Proposition II.8.5 in [27],  $\|X_1^n(\tau) - X_2^n(\tau)\|_E$  also admits the lower and upper derivatives with respect to  $\tau$ , and there exists  $x_n^*(\tau) \in \partial(\|X_1^n(\tau) - X_2^n(\tau)\|_E)$  such that the lower derivative of  $\|X_1^n(\tau) - X_2^n(\tau)\|_E$  satisfies

$$\frac{d^-}{d\tau} \|X_1^n(\tau) - X_2^n(\tau)\|_E = \left\langle x_n^*(\tau), \frac{d}{d\tau}(X_1^n(\tau) - X_2^n(\tau)) \right\rangle_{E^*, E}.$$



Thus we have

$$\begin{aligned}
\frac{d^-}{d\tau} \|X_1^n(\tau) - X_2^n(\tau)\|_E &= \langle x_n^*(\tau), A(X_1^n(\tau) - X_2^n(\tau)) \\
&\quad + nR(n, A)[F(\tau, X_1(\tau)) - F(\tau, X_2(\tau))] \rangle_{E^*, E} \\
&= \langle x_n^*(\tau), A(X_1^n(\tau) - X_2^n(\tau)) + F(\tau, X_1^n(\tau)) \\
&\quad - F(\tau, X_2^n(\tau)) \rangle_{E^*, E} \\
&\quad + \langle x_n^*(\tau), nR(n, A)F(\tau, X_1(\tau)) - F(\tau, X_1^n(\tau)) \rangle_{E^*, E} \\
&\quad - \langle x_n^*(\tau), nR(n, A)F(\tau, X_2(\tau)) - F(\tau, X_2^n(\tau)) \rangle_{E^*, E} \\
&\leq \eta \|X_1^n(\tau) - X_2^n(\tau)\|_E + \|\delta_1^n(\tau) - \delta_2^n(\tau)\|_E,
\end{aligned}$$

where for  $i = 1, 2$  we have set  $\delta_i^n(\tau) = nR(n, A)F(\tau, X_i(\tau)) - F(\tau, X_i^n(\tau))$ . We claim that  $\delta_i^n(\tau)$  tends to 0 uniformly in  $\tau \in [t, T]$ . Indeed,

$$\delta_i^n(\tau) = nR(n, A)[F(\tau, X_i(\tau)) - F(\tau, X_i^n(\tau))] + (nR(n, A) - I)F(\tau, X_i^n(\tau)),$$

and the convergence to 0 follows by a classical argument (see, e.g., the proof of Theorem 7.10 in [12]) since  $X_i^n(\tau)$  tends to  $X_i(\tau)$  uniformly in  $\tau \in [t, T]$  and the maps  $\tau \mapsto F(\tau, X_i(\tau))$  and  $\tau \mapsto F(\tau, X_i^n(\tau))$  are continuous with respect to  $\tau$ . By the Gronwall lemma,

$$\begin{aligned}
\|X_1^n(\tau) - X_2^n(\tau)\|_E &\leq \|nR(n, A)(x_1 - x_2)\|_E \max\left(1, e^{\eta(\tau-t)}\right) \\
&\quad + \int_t^\tau e^{(\tau-s)\eta} \|\delta_1^n(s) - \delta_2^n(s)\|_E ds \\
&\leq \|nR(n, A)(x_1 - x_2)\|_E \max\left(1, e^{\eta(T-t)}\right) \\
&\quad + \max\left(1, e^{\eta(T-t)}\right) \int_t^\tau \|\delta_1^n(s) - \delta_2^n(s)\|_E ds.
\end{aligned}$$

Letting  $n \rightarrow \infty$  we get

$$\|X_1(\tau) - X_2(\tau)\|_E \leq \max\left(1, e^{\eta(T-t)}\right) \|x_1 - x_2\|_E. \quad \square$$

*Remark 3.8.* If  $A$ ,  $F$ , and  $G$  satisfy Hypotheses 3.1 and 3.2, then by Proposition 3.7 we get

$$(3.2) \quad \sup_{\tau \in [t, T]} \left\| \frac{X(\tau, t, x + rh) - X(\tau, t, x)}{r} \right\|_E^p \leq \max\left(1, e^{p\eta(T-t)}\right) \|h\|_E^p.$$

The next step is to show that if  $F$  is differentiable with respect to  $x$ , then  $X(\tau, t, x)$  is pathwise differentiable with respect to the initial datum. We make the following assumptions on  $F$  in (3.1).

*Hypothesis 3.9.*  $F \in \mathcal{G}^{0,1}([0, T] \times E, E)$  and its derivative  $\nabla F$  is bounded on bounded sets.

Under this assumption, we can state a result on pathwise differentiability of  $X(\tau, t, x)$  with respect to  $x$ , that is, as a map from  $E$  to  $E$ .

PROPOSITION 3.10. *Assume that Hypotheses 3.1 and 3.2 hold true and that  $F$  satisfies Hypothesis 3.9. Then,  $\mathbb{P}$ -a.s. and for all  $0 \leq t \leq \tau \leq T$ , the map  $x \mapsto X(\tau, t, x)$  belongs to  $\mathcal{G}^1(E, E)$ , and  $\nabla X(\tau, t, x)$  is a mild solution to the equation with values in  $L(E, E)$ :*

$$(3.3) \quad \begin{cases} \frac{dY_\tau}{d\tau} = AY_\tau + \nabla F(\tau, X(\tau, t, x)) Y_\tau, & \tau \in [t, T], \\ Y_t = I. \end{cases}$$

Namely, for every  $h \in E$ ,  $\nabla X(\tau, t, x)h$  satisfies  $\mathbb{P}$ -a.s. the integral equation

$$Y_\tau h = e^{(\tau-t)A}h + \int_t^\tau e^{(\tau-s)A} \nabla F(s, X_s) Y_s h ds.$$

*Proof.* For every  $h \in E$  and for every  $r > 0$ , we consider the difference quotient

$$\begin{aligned} \Delta^r X(\tau, t, x)h &= \frac{X(\tau, t, x + rh) - X(\tau, t, x)}{r} \\ &= e^{(\tau-t)A}h + \int_t^\tau e^{(\tau-s)A} \frac{F(s, X(s, t, x + rh)) - F(s, X(s, t, x))}{r} ds \end{aligned}$$

and we want to show that, as  $r$  tends to 0, its limit exists and it is equal to the mild solution of (3.3) evaluated in  $h$ , that is, to  $Y_\tau h$  satisfying

$$Y_\tau h = e^{(\tau-t)A}h + \int_t^\tau e^{(\tau-s)A} \nabla F(s, X(s, t, x)) Y_s h ds.$$

We evaluate

$$\begin{aligned} &\|\Delta^r X(\tau, t, x)h - Y_\tau h\|_E \\ &= \left\| \int_t^\tau e^{(\tau-s)A} \left[ \frac{F(s, X(s, t, x + rh)) - F(s, X(s, t, x))}{r} - \nabla F(s, X(s, t, x)) Y_s h \right] ds \right\|_E \\ &= \left\| \int_t^\tau e^{(\tau-s)A} \left[ \int_0^1 \nabla F(s, X(s, t, x) + w(X(s, t, x + rh) - X(s, t, x))) \right. \right. \\ &\quad \left. \left. \Delta^r X(s, t, x)h dw - \nabla F(s, X(s, t, x)) Y_s h \right] ds \right\|_E \\ &\leq \int_t^\tau \max\left(1, e^{\omega(\tau-s)}\right) \int_0^1 \|\nabla F(s, X(s, t, x) + w(X(s, t, x + rh) - X(s, t, x))) \\ &\quad - \nabla F(s, X(s, t, x))\| \|\Delta^r X(s, t, x)h\|_E dw ds \\ &\quad + \int_t^\tau \left\| e^{\omega(\tau-s)} \nabla F(s, X(s, t, x)) [\Delta^r X(s, t, x)h - Y_s h] \right\|_E ds \\ &\leq \epsilon(r) + \max\left(1, e^{\omega(T-t)}\right) \int_t^\tau \|\nabla F(s, X(s, t, x))\|_{L(E, E)} \|\Delta^r X(s, t, x)h - Y_s h\|_E ds, \end{aligned}$$

where

$$(3.4) \quad \epsilon(r) = \max \left( 1, e^{\omega(T-t)} \right) \int_t^T \int_0^1 \left\| [\nabla F(s, X(s, t, x) + w(X(s, t, x + rh) - X(s, t, x))) - \nabla F(s, X(s, t, x))] \Delta^r X(s, t, x) h \right\|_E dw ds.$$

By the Gronwall lemma in integral form we deduce that

$$(3.5) \quad \left\| \Delta^r X(\tau, t, x) h - Y_\tau h \right\|_E \leq \epsilon(r) \exp \left\{ \max \left( 1, e^{\omega(T-t)} \right) \int_t^\tau \left\| \nabla F(s, X(s, t, x)) \right\|_{L(E, E)} ds \right\}.$$

By dominated convergence, in definition (3.4) we can pass to the limit as  $r$  goes to 0: indeed, by Hypothesis 3.9  $\nabla F$  is locally bounded, and by estimate (3.2)  $\left\| \Delta^r X(s, t, x) h \right\|_E \leq c \|h\|_E$ , where  $c$  is a constant independent on  $s$  and  $x$ , so the integrand is bounded in  $s$  and  $w$ . It remains to show that  $x \mapsto \nabla X(\tau, t, x)$  is strongly continuous. We note that for every  $x \in E$  the mapping  $h \mapsto \nabla X(\tau, t, x) h$  is continuous from  $E$  to  $E$ : For every  $h, h' \in E$ ,

$$\begin{aligned} \left\| \nabla X(\tau, t, x) h' - \nabla X(\tau, t, x) h \right\|_E &\leq \left\| \nabla X(\tau, t, x) h' - \Delta^r X(\tau, t, x) h' \right\|_E \\ &+ \left\| \Delta^r X(\tau, t, x) h' - \Delta^r X(\tau, t, x) h \right\|_E + \left\| \Delta^r X(\tau, t, x) h - \nabla X(\tau, t, x) h \right\|_E. \end{aligned}$$

For every  $r > 0$  the second term tends to 0 as  $h'$  tends to  $h$ , and as  $r$  tends to 0 the first and third terms in the right-hand side tend to 0. Thus, by Lemma 2.3 in [16], we need only show that the mapping  $\nabla X(\tau, t, \cdot)h : E \rightarrow E$  is continuous. In fact, for every  $x, x', h \in E$ ,

$$\begin{aligned} &\left\| \nabla X(\tau, t, x') h - \nabla X(\tau, t, x) h \right\|_E \\ &= \left\| \int_t^\tau e^{(\tau-s)A} [\nabla F(s, X(s, t, x')) \nabla X(s, t, x') h - \nabla F(s, X(s, t, x)) \nabla X(s, t, x) h] ds \right\|_E \\ &\leq \max \left( 1, e^{\omega(T-t)} \right) \int_t^\tau \left\| \nabla F(s, X(s, t, x')) \nabla X(s, t, x') h - \nabla F(s, X(s, t, x)) \nabla X(s, t, x) h \right\|_E ds \\ &\leq \max \left( 1, e^{\omega(T-t)} \right) \int_t^\tau \left\| \nabla F(s, X(s, t, x')) (\nabla X(s, t, x') h - \nabla X(s, t, x) h) \right\|_E ds \\ &\quad + \max \left( 1, e^{\omega(T-t)} \right) \int_t^\tau \left\| (\nabla F(s, X(s, t, x')) - \nabla F(s, X(s, t, x))) \nabla X(s, t, x) h \right\|_E ds \\ &\leq \max \left( 1, e^{\omega(T-t)} \right) \int_t^\tau \left\| \nabla F(s, X(s, t, x')) \right\|_E \left\| \nabla X(s, t, x') h - \nabla X(s, t, x) h \right\|_E ds + \epsilon(x, x'), \end{aligned}$$

where in the last passage  $\epsilon(x, x')$  tends to 0 as  $x$  tends to  $x'$  in  $E$ : we can apply dominated convergence again since  $\nabla F$  is locally bounded, and by estimate (3.2),

$$\sup_{\tau \in [t, T]} \|\nabla X(\tau, t, x) h\|_E^p \leq \max \left( 1, e^{p\eta(T-t)} \right) \|h\|_E^p.$$

By the Gronwall lemma in integral form we get

$$(3.6) \quad \|\nabla X(\tau, t, x') h - \nabla X(\tau, t, x) h\|_E \leq \epsilon(x, x') \exp \left\{ \max \left( 1, e^{\omega(T-t)} \right) \int_t^\tau \|\nabla F(s, X(s, t, x'))\|_{L(E, E)} ds \right\},$$

and this concludes the proof.  $\square$

We remember that here and in the following,  $(X(\tau, t, x))_\tau$  can be regarded as a process defined in the whole time interval  $[0, T]$  by setting  $X(\tau, t, x) = x$  for  $\tau \in [0, t]$ . Next we make assumptions of polynomial growth on  $F$  in order for the process  $X(\tau, t, x)$  to belong to  $\mathcal{H}^p([0, T], E)$ . Moreover, we will prove that  $X(\tau, t, x)$  is differentiable with respect to  $x$  as a map taking values in  $\mathcal{H}^p([0, T], E)$ .

*Hypothesis 3.11.*  $F \in \mathcal{G}^{0,1}([0, T] \times E, E)$ , its derivative  $\nabla F$  is bounded on bounded sets, and there exist an integer  $k \geq 0$  and a constant  $c > 0$  such that

$$\|F(\tau, x)\|_E \leq c \left( 1 + \|x\|_E^k \right)$$

for every  $\tau \in [0, T]$  and  $x \in E$ .

**LEMMA 3.12.** *Assume that  $A$ ,  $F$ , and  $G$  satisfy Hypotheses 3.1 and 3.2 and that  $F$  satisfies Hypothesis 3.11. Then the process  $(X_\tau)_{\tau \in [0, T]}$  belongs to  $\mathcal{H}^p([0, T], E)$  for every  $1 \leq p < \infty$ .*

*Proof.* Let  $Y_\tau = X_\tau - W_A(\tau)$  so that  $Y_\tau$  satisfies the equation

$$\begin{cases} dY_\tau = AY_\tau d\tau + F(\tau, Y_\tau + W_A(\tau)), \\ Y_t = x. \end{cases}$$

We set  $Y_\tau^n = nR(n, A)Y_\tau$ , where  $R(n, A) = (nI - A)^{-1}$  is the resolvent operator of  $A$ . Since  $Y_\tau^n \in \mathcal{D}(A)$  for every  $\tau \in [t, T]$ , and

$$Y_\tau^n = e^{\tau A} nR(n, A) x + \int_t^\tau e^{(\tau-s)A} nR(n, A) F(s, Y_s + W_A(s)) ds,$$

we get

$$\frac{d}{d\tau} Y_\tau^n = AY_\tau^n + nR(n, A) F(\tau, Y_\tau + W_A(\tau)).$$

Thus, by Proposition II.8.5 in [27],  $\|Y_\tau^n\|_E$  also admits the lower and upper derivatives with respect to  $\tau$  and there exists  $y_{n,\tau}^* \in \partial(\|Y_\tau^n\|_E)$  such that the lower derivative of  $\|Y_\tau^n\|_E$  satisfies

$$\begin{aligned} \frac{d^-}{d\tau} \|Y_\tau^n\|_E &= \left\langle y_{n,\tau}^*, \frac{d}{d\tau} Y_\tau^n \right\rangle_{E^*, E} \\ &= \langle y_\tau^*, AY_\tau^n + F(\tau, Y_\tau^n + W_A(\tau)) - F(\tau, W_A(\tau)) \rangle_{E^*, E} \\ &\quad + \langle y_\tau^*, nR(n, A) F(\tau, Y_\tau + W_A(\tau)) \\ &\quad - F(\tau, Y_\tau^n + W_A(\tau)) + F(\tau, W_A(\tau)) \rangle_{E^*, E} \end{aligned}$$

$$\begin{aligned} &\leq \eta \|Y_\tau\|_E + \|nR(n, A) F(\tau, Y_\tau + W_A(\tau)) \\ &\quad - F(\tau, Y_\tau^n + W_A(\tau))\|_E + c \left(1 + \|W_A(\tau)\|_E^k\right). \end{aligned}$$

By the Gronwall lemma,

$$\|Y_\tau^n\|_E \leq \|x\|_E e^{\eta(\tau-t)} + c \int_t^\tau \left( \|\delta_s^n\|_E + c \left(1 + \|W_A(s)\|_E^k\right) e^{\eta(s-t)} \right) ds,$$

where  $\delta_s^n = nR(n, A)F(s, Y_s + W_A(s)) - F(s, Y_s^n + W_A(s))$ . By the same procedure as in the proof of Proposition 3.7, as  $n \rightarrow \infty$ ,  $\delta_s^n$  tends to 0 uniformly in time. Thus, letting  $n \rightarrow \infty$ , we get

$$\|Y_\tau\|_E \leq \|x\|_E e^{\eta(\tau-t)} + \int_t^\tau c \left(1 + \|W_A(s)\|_E^k\right) e^{\eta(s-t)} ds,$$

and consequently,

$$\begin{aligned} \mathbb{E} \sup_{\tau \in [t, T]} \|Y_\tau\|_E^p &\leq 2^{p-1} \left( \max \left(1, e^{p\eta(T-t)}\right) \|x\|_E^p \right. \\ &\quad \left. + c \mathbb{E} \left( \int_t^T \left(1 + \|W_A(s)\|_E^k\right) e^{\eta(s-t)} ds \right)^p \right) \\ &\leq 2^{p-1} \left( \max \left(1, e^{p\eta(T-t)}\right) \|x\|_E^p + C_{T,k,p} \right). \end{aligned}$$

In the last estimate we refer to Remark 3.4, where we have deduced that as a process with values in  $C([0, T], E)$  the stochastic convolution is a Gaussian process and so it has finite moments of every order. So the process  $(Y_\tau)_{\tau \in [0, T]}$ , and consequently the process  $(X(\tau, t, x))_{\tau \in [0, T]}$ , belongs to  $\mathcal{H}^p([0, T], E)$  for every  $1 \leq p < \infty$ .  $\square$

Now we are ready to prove a result about differentiability of the process  $(X(\tau, t, x))_{\tau \in [0, T]}$ .

**PROPOSITION 3.13.** *Assume that Hypotheses 3.1 and 3.2 hold true and that  $F$  satisfies Hypothesis 3.11. Then for every  $t \in [0, T]$ , the map  $x \mapsto (X(\tau, t, x))_{\tau \in [0, T]} \in \mathcal{G}^1(E, \mathcal{H}^p([0, T], E))$ .*

*Proof.* With the notation of Proposition 3.10 we have to prove that, for every  $h \in E$ , the difference quotient  $\Delta^r X(\tau, t, x)h$  converges to  $\nabla X(\tau, t, x)h$  in  $\mathcal{H}^p([0, T], E)$ : we evaluate

$$\mathbb{E} \sup_{\tau \in [0, T]} \|\Delta^r X(\tau, t, x)h - \nabla X(\tau, t, x)h\|_E^p$$

as  $r$  goes to 0. By estimate (3.2) and by Proposition 3.10 we deduce the following estimate:

$$\sup_{\tau \in [0, T]} (\|\Delta^r X(\tau, t, x)h\|_E^p + \|\nabla X(\tau, t, x)h\|_E^p) \leq 2 \max \left(1, e^{p\eta(T-t)}\right) \|h\|_E^p.$$

Thus by dominated convergence,

$$\begin{aligned} &\lim_{r \rightarrow 0} \mathbb{E} \sup_{\tau \in [0, T]} \|\Delta^r X(\tau, t, x)h - \nabla X(\tau, t, x)h\|_E^p \\ &= \mathbb{E} \lim_{r \rightarrow 0} \sup_{\tau \in [0, T]} \|\Delta^r X(\tau, t, x)h - \nabla X(\tau, t, x)h\|_E^p. \end{aligned}$$

By estimate (3.5), and since  $\nabla F$  is locally bounded, we get

$$\begin{aligned} & \lim_{r \rightarrow 0} \mathbb{E} \sup_{\tau \in [0, T]} \|\Delta^r X(\tau, t, x)h - \nabla X(\tau, t, x)h\|_E^p \\ & \leq \mathbb{E} \lim_{r \rightarrow 0} \left[ \epsilon(r)^p \exp \left( p \max \left( 1, e^{\omega(T-t)} \right) \int_t^T \|\nabla F(s, X(s, t, x))\|_{L(E, E)} ds \right) \right] = 0. \end{aligned}$$

It remains to show that the mapping  $x \mapsto \nabla X(\tau, t, x)h$ , from  $E$  to  $\mathcal{H}^p([0, T], E)$ , is continuous. For every  $x, x', h \in E$ , we evaluate

$$\mathbb{E} \sup_{\tau \in [0, T]} \|\nabla X(\tau, t, x')h - \nabla X(\tau, t, x)h\|_E^p.$$

By (3.2) and by dominated convergence, we get

$$\begin{aligned} & \lim_{\|x-x'\|_E \rightarrow 0} \mathbb{E} \sup_{\tau \in [0, T]} \|\nabla X(\tau, t, x')h - \nabla X(\tau, t, x)h\|_E^p \\ & = \mathbb{E} \lim_{\|x-x'\|_E \rightarrow 0} \sup_{\tau \in [0, T]} \|\nabla X(\tau, t, x')h - \nabla X(\tau, t, x)h\|_E^p \\ & \leq \mathbb{E} \lim_{\|x-x'\|_E \rightarrow 0} \epsilon^p(x, x') \\ & \quad \times \exp \left\{ p \max \left( 1, e^{\omega(T-t)} \right) \int_t^T \|\nabla F(s, X(s, t, x'))\|_{L(E, E)} ds \right\} = 0, \end{aligned}$$

where in the last passages we have used estimate (3.6).  $\square$

**3.1. Representation of the derivative.** In this subsection we consider a function  $v : [0, T] \times E \rightarrow \mathbb{R}$  continuous and such that, for every  $t \in [0, T]$ ,  $v(t, \cdot) \in \mathcal{G}^1(E)$  and the map  $(t, x) \mapsto \nabla v(t, x)$  is Borel measurable. Let  $t \in [0, T]$  and  $x \in E$  be fixed. We are going to prove the following identification for the process  $\nabla v(\tau, X_\tau)$ , where  $X$  is the solution of the forward equation (3.1). For brevity we will write  $X_\tau = X(\tau, t, x)$ . Namely, let  $Z$  be a square integrable process and let  $\psi$  satisfy

$$\mathbb{E} \int_0^T |\psi_\sigma|^2 d\sigma < \infty.$$

Assume that  $Z$  and  $\psi$  depend on  $t$  and  $x$ . If  $v(\tau, X_\tau)$  admits the representation

$$(3.7) \quad v(\tau, X_\tau) = v(T, X_T) + \int_\tau^T \psi_\sigma d\sigma - \int_\tau^T Z_\sigma dW_\sigma, \quad \tau \in [t, T],$$

then  $\nabla v(\tau, X_\tau)G = Z_\tau$  for every  $0 \leq t \leq \tau \leq T$ .

*Remark 3.14.* If  $v$  is sufficiently regular and  $E$  is a Hilbert space, this is an obvious consequence of the Ito rule; see [12, Theorem 4.17].

*Remark 3.15.* The identification of  $Z_s$  with  $\nabla v(s, X_s)G$  is in fact a result of the joint quadratic variation between the process  $v(s, X_s)$ , which turns out to be a semimartingale if  $Z$  is a square integrable process and  $\psi \in L^2(\Omega \times [0, T])$ , and the process  $W^\xi(s) = \int_t^s \xi_\sigma^* dW_\sigma$ , where  $(\xi_\tau)_\tau$  is a bounded predictable process with values in  $\Xi$ . Indeed, by (3.7), the joint quadratic variation is equal to  $\int_t^s Z_\sigma \xi_\sigma d\sigma$ . By the identification  $\nabla v(s, X_s)G = Z_s$ , we can conclude that this joint quadratic variation

between  $v(s, X_s)$  and  $W^\xi(s)$  is equal to  $\int_t^s \nabla v(\sigma, X_\sigma) G \xi_\sigma d\sigma$ . An analogous result about joint quadratic variations is proved in [16], where the process  $X$  evolves in a Hilbert space and  $G$  is not necessarily constant. As far as we know, in the literature there are no results in which  $X$  evolves in a Banach space.

As in the previous section,  $F$  in the forward equation satisfies Hypothesis 3.11, and so  $X$  belongs to  $\mathcal{H}^p$ . On  $A$  and  $G$  we have to require some additional properties: We cannot guarantee that  $G(\Xi) \subset E$ , but we can make the following assumptions, which are verified in most of the applications.

*Hypothesis 3.16.* There exists a Banach subspace  $\Xi_0$  dense in  $\Xi$  such that  $G(\Xi_0) \subset E$  and  $G : \Xi_0 \rightarrow E$  is continuous.

**THEOREM 3.17.** *Assume that Hypotheses 3.1, 3.2, 3.11, and 3.16 hold true. Let  $v : [0, T] \times E \rightarrow \mathbb{R}$  be continuous such that, for every  $t \in [0, T]$ ,  $v(t, \cdot) \in \mathcal{G}^1(E)$ , the map  $(t, x) \mapsto \nabla v(t, x)$  is measurable, and for every  $0 \leq t \leq s \leq T$ ,  $|\nabla v(s, x)h| \leq c(1 + \|x\|_E^j) \|h\|_E$  for some integer  $j \geq 0$  and for every  $x, h \in E$ . Assume that  $v$  verifies equality (3.7), with  $Z$  a square integrable process and  $\psi \in L^2(\Omega \times [0, T])$ , and  $Z$  and  $\psi$  depend on  $t$  and  $x$ . Then, for almost every  $s \in [0, T]$ ,  $Z_s \varsigma = \nabla v(s, X_s) G \varsigma$ ,  $\mathbb{P}$ -almost everywhere and for every  $\varsigma \in \Xi_0$ .*

*Proof.* Let  $\tau = t$  in (3.7). By the definition of  $v$  we can write

$$(3.8) \quad v(t, x) + \int_t^T Z_\sigma dW_\sigma = v(T, X_T) + \int_t^T \psi_\sigma d\sigma.$$

For  $t \leq s \leq T$  we consider the process

$$\begin{aligned} v(s, X(s, t, x)) &= - \int_s^T Z_\sigma dW_\sigma + v(T, X(T, s, X(s, t, x))) + \int_s^T \psi_\sigma d\sigma \\ &= \int_t^s Z_\sigma dW_\sigma - \int_t^s \psi_\sigma d\sigma + v(t, x). \end{aligned}$$

Let  $\mathcal{F}_\tau$  be the natural filtration generated by the Wiener process and augmented in the usual way. We now define a family  $\mathcal{S}$  of predictable processes with real values. For every  $n \in \mathbb{N}$ , we subdivide the interval  $[0, T]$  into subintervals

$$\left[ \frac{kT}{2^n}, \frac{(k+1)T}{2^n} \right), \quad k = 0, \dots, 2^n - 1.$$

A predictable process  $\eta$  belongs to  $\mathcal{S}$  if on

$$\left[ \frac{kT}{2^n}, \frac{(k+1)T}{2^n} \right), \quad k = 0, \dots, 2^n - 1,$$

it has the form

$$(3.9) \quad \eta_t = \eta^k(W_{t_1}, \dots, W_{t_{l_k}}), \quad t \in \left[ \frac{kT}{2^n}, \frac{(k+1)T}{2^n} \right), \quad 0 \leq t_1 \leq \dots \leq t_{l_k} \leq \frac{kT}{2^n},$$

where  $\eta^k$  is a bounded function in  $C^\infty(\mathbb{R}^{l_k}, \mathbb{R})$ , with bounded derivatives of all orders. In the following we will briefly write  $\eta_t = \eta_t(W.)$ , where by  $W.$  we mean the trajectory of  $W$  up to time  $t$ . For  $\varsigma \in \Xi_0$ , we set  $\xi_t = \eta_t \varsigma$ .

Let  $s > t$  and  $\delta > 0$ , small enough such that  $s - \delta > t$ . From now on we identify  $\Xi$  with its dual  $\Xi^*$ , and we write  $\xi$  for  $\xi^*$ . We multiply both sides of (3.8) by  $\int_{s-\delta}^s \xi_\sigma dW_\sigma$

and we take expectation:

$$\begin{aligned}\mathbb{E}\left[v(s, X_s) \int_{s-\delta}^s \xi_\sigma dW_\sigma\right] &= \mathbb{E}\left[\int_t^{s-\delta} \psi_\sigma d\sigma \int_{s-\delta}^s \xi_\sigma dW_\sigma\right] + \mathbb{E}\left[\int_{s-\delta}^s \psi_\sigma d\sigma \int_{s-\delta}^s \xi_\sigma dW_\sigma\right] \\ &\quad + \mathbb{E}\left[\int_t^s Z_\sigma dW_\sigma \int_{s-\delta}^s \xi_\sigma dW_\sigma\right].\end{aligned}$$

It is immediate that

$$\mathbb{E}\left[\int_t^{s-\delta} \psi_\sigma d\sigma \int_{s-\delta}^s \xi_\sigma dW_\sigma\right] = 0, \quad \mathbb{E}\left[\int_t^s Z_\sigma dW_\sigma \int_{s-\delta}^s \xi_\sigma dW_\sigma\right] = \mathbb{E}\left[\int_{s-\delta}^s Z_\sigma \xi_\sigma d\sigma\right].$$

We estimate

$$\begin{aligned}\mathbb{E}\left[\int_{s-\delta}^s \psi_\sigma d\sigma \int_{s-\delta}^s \xi_\sigma dW_\sigma\right] &\leq \left(\mathbb{E}\left|\int_{s-\delta}^s \psi_\sigma d\sigma\right|^2\right)^{1/2} \left(\mathbb{E}\left|\int_{s-\delta}^s \xi_\sigma dW_\sigma\right|^2\right)^{1/2} \\ &\leq \sqrt{\delta} \left(\mathbb{E} \int_{s-\delta}^s |\psi_\sigma|^2 d\sigma\right)^{1/2} \left(\mathbb{E} \int_{s-\delta}^s \|\xi_\sigma\|_\Xi^2 d\sigma\right)^{1/2} \\ &\leq C\delta \left(\mathbb{E} \int_{s-\delta}^s |\psi_\sigma|^2 d\sigma\right)^{1/2}.\end{aligned}$$

If we divide both sides by  $\delta$  and let  $\delta \rightarrow 0$ , then we get, for almost every  $s$ ,

$$\mathbb{E}[Z_s \xi_s] = \lim_{\delta \rightarrow 0} \frac{1}{\delta} \mathbb{E}\left[v(s, X_s) \int_{s-\delta}^s \xi_\sigma dW_\sigma\right].$$

The final step for the identification of  $Z$  is to prove that

$$\lim_{\delta \rightarrow 0} \frac{1}{\delta} \mathbb{E}\left[v(s, X_s) \int_{s-\delta}^s \xi_\sigma dW_\sigma\right] = \mathbb{E}[\nabla v(s, X_s) G \xi_s].$$

We proceed as in [2]. For  $0 \leq t \leq \sigma \leq T$ , we define  $W_\sigma^\varepsilon$  by

$$(3.10) \quad W_\sigma^\varepsilon = W_\sigma - \varepsilon \int_t^\sigma \xi_r(W^\varepsilon) dr,$$

where  $\xi_r(W^\varepsilon)$  depends on the trajectories of  $W^\varepsilon$  up to time  $r$ , and the dependence is given by the definition (3.9) of  $\eta$ . The process  $(W_\sigma^\varepsilon)_\sigma$  is defined as the solution of (3.10), which is not considered as a stochastic differential equation, as specified in [2, p. 476]. Equation (3.10) can be solved step by step in each interval

$$\left[\frac{kT}{2^n}, \frac{(k+1)T}{2^n}\right), \quad k = 0, \dots, 2^n - 1.$$

As we recalled in section 2, cylindrical Wiener processes can be seen as  $Q$ -Wiener processes with values in another Hilbert space  $\Xi_1$ . We solve (3.10) in  $\Xi_1$ . If  $\sigma < t$ ,  $W_\sigma^\varepsilon = W_\sigma$ . Let

$$\bar{k} = \max\left(k : t \geq \frac{kT}{2^n}\right).$$



If

$$\sigma \in \left[ t, \frac{(\bar{k} + 1)T}{2^n} \right),$$

then

$$W_\sigma^\varepsilon = W_\sigma - \varepsilon \int_t^\sigma \xi_{\bar{k}}^\varepsilon(W_{t_1}^\varepsilon, \dots, W_{t_{\bar{k}}}^\varepsilon) dr = W_\sigma - \varepsilon (\sigma - t) \xi_{\bar{k}}^\varepsilon(W_{t_1}, \dots, W_{t_{\bar{k}}}).$$

For

$$\sigma \in \left[ \frac{(\bar{k} + 1)T}{2^n}, \frac{(\bar{k} + 2)T}{2^n} \right),$$

we have

$$\begin{aligned} W_\sigma^\varepsilon &= W_\sigma - \varepsilon \int_t^{\frac{(\bar{k}+1)T}{2^n}} \xi_{\bar{k}}^\varepsilon(W_{t_1}^\varepsilon, \dots, W_{t_{\bar{k}}}^\varepsilon) dr - \varepsilon \int_{\frac{(\bar{k}+1)T}{2^n}}^\sigma \xi_{\bar{k}+1}^\varepsilon(W_{t_1}^\varepsilon, \dots, W_{t_{\bar{k}+1}}}^\varepsilon) dr \\ &= W_\sigma - \varepsilon \left( \frac{(\bar{k} + 1)T}{2^n} - t \right) \xi_{\bar{k}}^\varepsilon(W_{t_1}^\varepsilon, \dots, W_{t_{\bar{k}}}^\varepsilon) \\ &\quad - \varepsilon \left( \sigma - \frac{(\bar{k} + 1)T}{2^n} \right) \xi_{\bar{k}+1}^\varepsilon(W_{t_1}^\varepsilon, \dots, W_{t_{\bar{k}+1}}}^\varepsilon), \end{aligned}$$

and by this procedure,  $(W_\sigma^\varepsilon)_\sigma$  is well defined for every  $0 \leq \sigma \leq T$ . Moreover,  $W_\sigma^\varepsilon$  is a function of the trajectories of  $W$  up to time  $\sigma$ , that is,  $W_\sigma^\varepsilon = W_\sigma^\varepsilon(W)$ . So we can write

$$W_\sigma^\varepsilon = W_\sigma - \varepsilon \int_t^\sigma \xi_r(W_r^\varepsilon(W)) dr, \quad 0 \leq t \leq \sigma \leq T.$$

Now we define a probability measure  $Q_\varepsilon$  such that

$$\frac{dQ_\varepsilon}{d\mathbb{P}} = \exp \left( \varepsilon \int_t^T \xi_\sigma(W_\sigma^\varepsilon(W)) dW_\sigma - \frac{\varepsilon^2}{2} \int_t^T |\xi_\sigma(W_\sigma^\varepsilon(W))|^2 d\sigma \right).$$

By the Girsanov theorem (see, e.g., [12, Theorem 10.14]), under  $Q_\varepsilon$ ,  $W_\sigma^\varepsilon = W_\sigma - \varepsilon \int_t^\sigma \xi_r(W_r^\varepsilon(W)) dr$  is a  $Q$ -Wiener process in  $\Xi_1$  and a cylindrical Wiener process in  $\Xi$ . By this construction of  $(W_\sigma^\varepsilon)_\sigma$ , it is also clear that for every  $0 \leq \sigma \leq T$ ,  $W_\sigma^\varepsilon$  is pathwise differentiable with respect to  $\varepsilon$  and  $\frac{d}{d\varepsilon}|_{\varepsilon=0} W_\sigma^\varepsilon = -\int_t^\sigma \xi_r(W_r) dr$ ; see also [2, p. 476].

By (3.7),  $v(s, X_s)$  is square integrable; indeed,

$$\begin{aligned} \mathbb{E} v^2(s, X_s) &\leq c \left[ 1 + \mathbb{E} \left( \int_t^s \xi_\sigma dW_\sigma \right)^2 + \mathbb{E} \left( \int_t^s \psi_\sigma d\sigma \right)^2 \right] \\ &\leq c \left[ 1 + \mathbb{E} \int_t^s \|\xi_\sigma\|_\Xi^2 d\sigma + T^{1/2} \mathbb{E} \int_t^s \psi_\sigma^2 d\sigma \right] < \infty. \end{aligned}$$

So, by the Cauchy–Schwarz inequality, the expectation of  $v(s, X_s) \int_t^s \xi_\sigma dW_\sigma$  is well defined. We are going to prove that

$$\begin{aligned} & \mathbb{E} \left[ v(s, X_s) \int_{s-\delta}^s \xi_\sigma dW_\sigma \right] \\ &= \frac{d}{d\varepsilon|_{\varepsilon=0}} \mathbb{E} \left[ v(s, X_s) \exp \left( \varepsilon \int_{s-\delta}^s \xi_\sigma (W^\varepsilon) dW_\sigma - \frac{\varepsilon^2}{2} \int_{s-\delta}^s \|\xi_\sigma (W^\varepsilon)\|_\Xi^2 d\sigma \right) \right]. \end{aligned}$$

We evaluate

$$\begin{aligned} & \frac{d}{d\varepsilon|_{\varepsilon=0}} \mathbb{E} \left[ v(s, X_s) \exp \left( \varepsilon \int_{s-\delta}^s \xi_\sigma dW_\sigma - \frac{\varepsilon^2}{2} \int_{s-\delta}^s \|\xi_\sigma\|_\Xi^2 d\sigma \right) \right] \\ &= \lim_{\varepsilon \rightarrow 0} \mathbb{E} \left[ v(s, X_s) \frac{\exp \left( \varepsilon \int_{s-\delta}^s \xi_\sigma dW_\sigma - \frac{\varepsilon^2}{2} \int_{s-\delta}^s \|\xi_\sigma\|_\Xi^2 d\sigma \right) - 1}{\varepsilon} \right] \\ &= \mathbb{E} \left[ v(s, X_s) \int_{s-\delta}^s \xi_\sigma dW_\sigma \right], \end{aligned}$$

where in the last passage the limit goes into the expectation by dominated convergence since  $\xi$  is bounded. So

$$\mathbb{E} \left[ v(s, X_s) \int_{s-\delta}^s \xi_\sigma dW_\sigma \right] = \frac{d}{d\varepsilon|_{\varepsilon=0}} \mathbb{E}_{Q_\varepsilon} [v(s, X_s)].$$

Moreover, in  $(\Omega, \mathcal{F}, Q_\varepsilon)$ ,  $X$  is a mild solution to equation

$$\begin{cases} dX_\tau = AX_\tau d\tau + F(\tau, X_\tau) d\tau + G\varepsilon \xi_\tau (W^\varepsilon) d\tau + GdW_\tau^\varepsilon, & \tau \in [s-\delta, T], \\ X_{s-\delta} = X(s-\delta, t, x). \end{cases}$$

Since  $\xi$  takes values in  $\Xi_0$ , and by Hypothesis 3.16  $G$  maps  $\Xi_0$  in  $E$ , such a mild solution is well defined as a process with values in  $E$ . In  $(\Omega, \mathcal{F}, \mathbb{P})$  we consider the process  $X_\tau^\varepsilon$ , which is a mild solution to the equation

$$\begin{cases} dX_\tau^\varepsilon = AX_\tau^\varepsilon d\tau + F(\tau, X_\tau^\varepsilon) d\tau + G\varepsilon \xi_\tau (W) d\tau + GdW_\tau, & \tau \in [s-\delta, T], \\ X_{s-\delta}^\varepsilon = X(s-\delta, t, x). \end{cases}$$

Then the process  $X$  under  $Q_\varepsilon$  and the process  $X^\varepsilon$  under  $\mathbb{P}$  have the same law, so we get

$$\frac{d}{d\varepsilon|_{\varepsilon=0}} \mathbb{E}_{Q_\varepsilon} [v(s, X_s)] = \frac{d}{d\varepsilon|_{\varepsilon=0}} \mathbb{E} [v(s, X_s^\varepsilon)].$$

Finally, we claim that

$$\frac{d}{d\varepsilon|_{\varepsilon=0}} \mathbb{E} [v(s, X_s^\varepsilon)] = \mathbb{E} \left[ \nabla v(s, X_s) \dot{X}_s \right],$$

where we have set  $\dot{X}_s := \frac{d}{d\varepsilon|_{\varepsilon=0}} X_s^\varepsilon$ ,  $\mathbb{P}$ -a.s. One can easily check that  $\dot{X}_s$  is the unique mild solution to the equation

$$(3.11) \quad \begin{cases} d\dot{X}_\tau = A\dot{X}_\tau d\tau + \nabla F(\tau, X_\tau) \dot{X}_\tau d\tau + G\xi_\tau d\tau, & \tau \in [s-\delta, T], \\ \dot{X}_{s-\delta} = 0. \end{cases}$$

Such a mild solution exists, and it is unique by Hypothesis 3.9 and since (3.11) is a linear equation in  $\dot{X}$ . Let us denote

$$\Delta^\varepsilon X_\tau = \frac{X_\tau^\varepsilon - X_\tau}{\varepsilon}.$$

We want to show that, as  $\varepsilon$  tends to 0, its limit exists and it is equal to the mild solution of (3.11). We set  $X_\tau^n = nR(n, A)X_\tau$ ,  $X_\tau^{n,\varepsilon} = nR(n, A)X_\tau^\varepsilon$ , and  $\Delta^{n,\varepsilon} X_\tau = \frac{X_\tau^{n,\varepsilon} - X_\tau^n}{\varepsilon}$ , where  $R(n, A) = (nI - A)^{-1}$  is the resolvent operator of  $A$ . We note that  $X_\tau^{n,\varepsilon} - X_\tau^n$  is differentiable with respect to  $\tau$ , and so, by Proposition II.8.5 in [27]  $\|X_\tau^{n,\varepsilon} - X_\tau^n\|_E$  also admits the lower and upper derivatives with respect to  $\tau$  and there exists  $x_{n,\tau}^* \in \partial(\|X_\tau^{n,\varepsilon} - X_\tau^n\|_E)$  such that the lower derivative of  $\|X_\tau^{n,\varepsilon} - X_\tau^n\|_E$  satisfies

$$\begin{aligned} \frac{d^-}{d\tau} \|X_\tau^{n,\varepsilon} - X_\tau^n\|_E &= \left\langle x_{n,\tau}^*, \frac{d}{d\tau} (X_\tau^{n,\varepsilon} - X_\tau^n) \right\rangle_{E^*, E} \\ &\leq \eta \|X_\tau^{n,\varepsilon} - X_\tau^n\|_E + |\varepsilon| \|G\xi_\tau\|_E. \end{aligned}$$

Since  $\xi \in \Xi_0$ , by Hypothesis 3.16,  $G\xi \in E$ . By the Gronwall lemma, we get that  $\mathbb{P}$ -a.s.

$$\|X_\tau^{n,\varepsilon} - X_\tau^n\|_E \leq |\varepsilon| \|G\|_{L[\Xi_0, E]} C(T, \xi),$$

where  $C(T, \xi)$  is a constant depending only on the time  $T$  and on the norm of  $\xi(\omega)$ ,  $\omega \in \Omega$ , in  $\Xi_0$ , which is bounded uniformly with respect to  $\omega \in \Omega$ . Thus

$$\|\Delta^{n,\varepsilon} X_\tau\|_E \leq \|G\|_{L[\Xi_0, E]} C(T, \xi),$$

and, letting  $n \rightarrow \infty$ , we get

$$\|\Delta^\varepsilon X_\tau\|_E \leq \|G\|_{L[\Xi_0, E]} C(T, \xi).$$

This estimate is used in order to prove that  $\lim_{\varepsilon \rightarrow 0} \Delta^\varepsilon X_\tau = \dot{X}_\tau$ . We evaluate

$$\begin{aligned} &\left\| \Delta^\varepsilon X_\tau - \dot{X}_\tau \right\|_E \\ &\leq \left\| \int_{s-\delta}^\tau e^{(\tau-r)A} \left[ \frac{F(r, X_r^\varepsilon) - F(r, X_r)}{\varepsilon} - \nabla F(r, X_r) \dot{X}_r \right] dr \right\|_E \\ &= \left\| \int_{s-\delta}^\tau e^{(\tau-r)A} \left[ \int_0^1 \nabla F(r, X_r + w(X_r^\varepsilon - X_r)) \Delta^\varepsilon X_r dw - \nabla F(r, X_r) \dot{X}_r \right] dr \right\|_E \\ &\leq \int_{s-\delta}^\tau e^{\omega(\tau-r)} \left\| \nabla F(r, X_r) \left( \Delta^\varepsilon X_r - \dot{X}_r \right) \right\|_E dr \\ &\quad + \int_{s-\delta}^\tau e^{\omega(\tau-r)} \left\| \int_0^1 [\nabla F(r, X_r + w(X_r^\varepsilon - X_r)) - \nabla F(r, X_r)] \Delta^\varepsilon X_r dw \right\|_E dr. \end{aligned}$$

By the Gronwall lemma and dominated convergence, we get that  $\lim_{\varepsilon \rightarrow 0} \|\Delta^\varepsilon X_\tau - \dot{X}_\tau\|_E = 0$ . Moreover,  $\dot{X}$  is bounded uniformly with respect to time and to  $\omega \in \Omega$ . By an application of the chain rule and of the dominated convergence theorem, which we

can apply since by hypothesis the derivative of  $v$  has polynomial growth with respect to  $x$  and the process  $X$  has finite moments of every order, we can conclude that

$$\frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} \mathbb{E}[v(s, X_s^\varepsilon)] = \mathbb{E}\left[\nabla v(s, X_s) \dot{X}_s\right].$$

We claim that

$$(3.12) \quad \dot{X}_\tau = \int_{s-\delta}^\tau \nabla X(\tau, \sigma, X(\sigma, t, x)) G\xi_\sigma d\sigma.$$

Let relation (3.12) be true. In this case, by the following calculations we conclude that proof:

$$\begin{aligned} \mathbb{E}[Z_s \xi_s] &= \lim_{\delta \rightarrow 0} \frac{1}{\delta} \mathbb{E}\left[\nabla v(s, X_s) \int_{s-\delta}^s \xi_\sigma dW_\sigma\right] \\ &= \lim_{\delta \rightarrow 0} \frac{1}{\delta} \mathbb{E}\left[\nabla v(s, X_s) \dot{X}_s\right] \\ &= \lim_{\delta \rightarrow 0} \frac{1}{\delta} \mathbb{E}\left[\nabla v(s, X_s) \int_{s-\delta}^s \nabla X(s, \sigma, X(\sigma, t, x)) G\xi_\sigma d\sigma\right] \\ &= \mathbb{E}[\nabla v(s, X_s) \nabla X(s, s, X(s, t, x)) G\xi_s] \\ &= \mathbb{E}[\nabla v(s, X_s) G\xi_s]. \end{aligned}$$

So for every  $\eta \in \mathcal{S}$ ,  $\mathbb{E}[Z_s \zeta \eta_s] = \mathbb{E}[\nabla v(s, X_s) G\zeta \eta_s]$  for almost every  $s \in [0, T]$ . By the arbitrariness of  $\eta$  it follows that for almost every  $s \in [0, T]$ ,  $Z_s \zeta = \nabla v(s, X_s) G\zeta$ ,  $\mathbb{P}$ -a.s.

Now it remains to prove (3.12). By Proposition 3.10, for  $0 \leq t \leq s \leq \tau \leq T$ ,

$$\nabla X(\tau, \sigma, y) = e^{(\tau-\sigma)A} + \int_\sigma^\tau e^{(\tau-r)A} \nabla F(r, X(r, \sigma, y)) \nabla X(r, \sigma, y) dr.$$

We multiply both sides by  $G\xi_\sigma$  and integrate. Taking  $y = X(\sigma, t, x)$  we get

$$\begin{aligned} &\int_{s-\delta}^\tau \nabla X(\tau, \sigma, X(\sigma, t, x)) G\xi_\sigma d\sigma \\ &= \int_{s-\delta}^\tau e^{(\tau-\sigma)A} G\xi_\sigma d\sigma + \int_{s-\delta}^\tau e^{(\tau-r)A} \\ &\quad \left\{ \int_\sigma^\tau \nabla F(r, X(r, \sigma, X(\sigma, t, x))) \nabla X(r, \sigma, X(\sigma, t, x)) dr \right\} G\xi_\sigma d\sigma \\ &= \int_{s-\delta}^\tau e^{(\tau-\sigma)A} G\xi_\sigma d\sigma + \int_{s-\delta}^\tau e^{(\tau-r)A} \nabla F(r, X_r) \\ &\quad \left\{ \int_{s-t}^\tau \nabla X(r, \sigma, X(\sigma, t, x)) G\xi_\sigma d\sigma \right\} dr. \end{aligned}$$

This gives the identification

$$\dot{X}_\tau = \int_{s-\delta}^\tau \nabla X(\tau, \sigma, X(\sigma, t, x)) G\xi_\sigma d\sigma$$

by the uniqueness of a mild solution of (3.11).  $\square$

*Remark 3.18.* By the previous theorem, for every  $\varsigma \in \Xi_0$  we have  $Z_s \varsigma = \nabla v(s, X_s)G \varsigma$   $\mathbb{P}$ -a.s. and for almost every  $s \in [0, T]$ . Since  $\Xi_0$  is dense in  $\Xi$ , for every  $\xi \in \Xi$  there exists a sequence  $(\varsigma_n)_n \in \Xi_0$  such that  $\varsigma_n \rightarrow \xi$  in  $\Xi$ . For almost every  $s \in [0, T]$  and almost surely with respect to the law of  $X_s$ , the operator  $\nabla v(s, x)G : \Xi_0 \rightarrow E$  extends to an operator defined in the whole  $\Xi$ . For the sake of brevity, we denote such an extension by  $\nabla v(s, x)G$ . We can conclude that  $Z_s = \nabla v(s, X_s)G$ ,  $\mathbb{P}$ -a.s. and for almost every  $s \in [0, T]$ .

**4. Preliminaries on the forward-backward system.** In this section we study a backward equation like (2.1), with  $\psi$  depending on the Markov process  $X$ , which is solution of (3.1) with initial condition  $x$  given at the initial time  $t$ . Namely, in a complete probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  let  $\{W_\tau, \tau \geq 0\}$  be a cylindrical Wiener process with values in  $\Xi$ , and let  $(\mathcal{F}_\tau)_{\tau \geq 0}$  be its natural filtration, augmented in the usual way. We consider the forward-backward system

$$(4.1) \quad \begin{cases} dX_\tau = AX_\tau d\tau + F(\tau, X_\tau) d\tau + GdW_\tau, & \tau \in [t, T], \\ dY_\tau = -\psi(\tau, X_\tau, Z_\tau) d\tau + Z_\tau dW_\tau, & \tau \in [t, T], \\ X_t = x, \\ Y_T = \phi(X_T), \end{cases}$$

where  $\psi : [0, T] \times E \times \Xi^* \rightarrow \mathbb{R}$  and  $\phi : E \rightarrow \mathbb{R}$  satisfy some suitable assumptions. We want to investigate the existence and uniqueness of a solution of this forward-backward system and its regular dependence on  $x$ . The solution will be denoted by  $(X_\tau, Y_\tau, Z_\tau)$  or, to stress the dependence on the initial condition  $x$  in the forward equation given at the initial time  $t$ , by  $(X(\tau, t, x), Y(\tau, t, x), Z(\tau, t, x))$ . For  $0 \leq \tau \leq t$ , we put  $X(\tau, t, x) = x$ , and the pair  $(Y_\tau, Z_\tau)$  is the solution of the corresponding backward equation.

In section 2, we have already recalled an existence and uniqueness result for a mild solution of the forward equation,

$$\begin{cases} dX_\tau = AX_\tau d\tau + F(\tau, X_\tau) d\tau + GdW_\tau, & \tau \in [t, T], \\ X_t = x, \end{cases}$$

with  $A$ ,  $F$ , and  $G$  satisfying Hypotheses 3.1 and 3.2. Moreover, we have studied the differentiable dependence of this solution on the initial condition  $x$ . We are going to study the dependence of the pair  $(Y_\tau, Z_\tau)$  on  $x \in E$ , where  $(Y_\tau, Z_\tau)$  is a pair of processes which solves the BSDE of the form

$$(4.2) \quad Y_\tau + \int_\tau^T Z_\sigma dW_\sigma = \phi(X_T) + \int_\tau^T \psi(\sigma, X_\sigma, Z_\sigma) d\sigma, \quad \tau \in [t, T].$$

In particular, we recall that  $Y(t, t, x)$  is a deterministic function; let us define  $v(t, x) := Y(t, t, x)$ , where  $v(t, x)$  is a continuous function; we want to prove that  $v(t, \cdot) \in \mathcal{G}^1(E)$  for every  $t \in [0, T]$ . We make suitable assumptions on  $\psi$  such that

$$\mathbb{E} \int_t^T |\psi(\sigma, 0, 0)|^2 d\sigma < \infty.$$

Thus, for the pair of processes  $(Y, Z)$ , which are solution of (4.2), the following a priori estimate holds (see [16, Proposition 4.3]):

$$(4.3) \quad \mathbb{E} \sup_{\tau \in [t, T]} |Y_\tau|^2 + \mathbb{E} \int_t^T \|Z_\sigma\|_{\Xi^*}^2 d\sigma \leq c \mathbb{E} \int_t^T |\psi(\sigma, 0, 0)|^2 d\sigma + c \mathbb{E} |\phi(X_T)|^2.$$

In the following we assume that  $A$ ,  $F$ , and  $G$  satisfy Hypotheses 3.1, 3.2, and 3.11. Moreover, we require that  $\psi$  and  $\phi$  satisfy the following assumptions that guarantee the solvability of the backward equation and the differentiable dependence of the solution  $(Y_\tau, Z_\tau)$  with respect to  $x$ .

*Hypothesis 4.1.* The maps  $\psi : [0, T] \times E \times \Xi^* \longrightarrow \mathbb{R}$  and  $\phi : E \longrightarrow \mathbb{R}$  are Borel measurable and satisfy the following assumptions:

1. There exists  $L > 0$  such that

$$|\psi(\sigma, x, z_1) - \psi(\sigma, x, z_2)| \leq L \|z_1 - z_2\|_{\Xi^*}$$

for every  $\sigma \in [0, T]$ ,  $x \in E$ , and  $z_1, z_2 \in \Xi^*$ ;

2. for every  $\sigma \in [0, T]$ , we have  $\psi(\sigma, \cdot, \cdot) \in \mathcal{G}^{1,1}(E \times \Xi^*)$ ;
3. there exist  $L > 0$  and  $m \geq 0$  such that

$$|\nabla_x \psi(\sigma, x, z) h| \leq L \|h\|_E (1 + \|x\|_E)^m (1 + \|z\|_{\Xi^*})$$

for every  $\sigma \in [0, T]$ ,  $x, h \in E$ , and  $z \in \Xi^*$ ;

4.  $\phi \in \mathcal{G}^1(E)$  and there exists  $L > 0$  such that for every  $x_1, x_2 \in E$ ,

$$|\phi(x_1) - \phi(x_2)| \leq L \|x_1 - x_2\|_E.$$

Under these assumptions we can state a result on the existence and uniqueness of a solution of the BSDE (4.2) and on its regular dependence on  $x$ .

**PROPOSITION 4.2.** *Assume that Hypotheses 3.1, 3.2, and 3.11 hold true and that  $\psi$  and  $\phi$  satisfy Hypothesis 4.1. Then (4.2) admits a unique solution  $(Y_\tau, Z_\tau) \in \mathbb{K}_{\text{cont}}([0, T])$  such that the map  $(t, x) \mapsto (Y(\tau, t, x), Z(\tau, t, x))$  is continuous from  $[0, T] \times E$  with values in  $\mathbb{K}_{\text{cont}}([0, T])$ , and for every  $t \in [0, T]$ , the map  $x \mapsto (Y(\cdot, t, x), Z(\cdot, t, x))$  belongs to  $\mathcal{G}^1(E, \mathbb{K}_{\text{cont}}([0, T]))$ . Moreover, the following estimate holds true: For every  $p \geq 2$ ,*

$$\left[ \mathbb{E} \sup_{\tau \in [0, T]} |\nabla_x Y(\tau, t, x) h|^p \right]^{1/p} \leq C \|h\|_E \left( 1 + \|x\|_E^{(m+1)^2} \right).$$

*Proof.* We give only a sketch of the proof. By Lemma 3.12 and by  $|\psi(\sigma, x, z)| \leq L(1 + \|x\|_E)^m$ , we get that, defining

$$\psi_\sigma = \psi(\sigma, X_\sigma, Z_\sigma),$$

condition (2.2) is satisfied, so existence and uniqueness of the solution of the backward equation follows, e.g., by [16]. The proof of differentiability of the map  $x \mapsto (Y(\cdot, t, x), Z(\cdot, t, x))$  substantially follows the proof of Proposition 5.2 in [16]. In that paper the process  $(X(\tau, t, x))_{\tau \in [0, T]}$  takes its values in a separable Hilbert space  $H$ , while in our context the process  $(X(\tau, t, x))_{\tau \in [0, T]}$  takes its values in the Banach space  $E$ ; nevertheless, the same arguments apply. Indeed, in Lemma 3.12 we have proved that  $(X(\tau, t, x))_{\tau \in [0, T]}$  belongs to  $\mathcal{H}^p([0, T], E)$  for every  $1 \leq p < \infty$ , and in Proposition 3.13 we have proved that the map  $x \mapsto (X(\tau, t, x))_{\tau \in [0, T]} \in \mathcal{G}^1(E, \mathcal{H}^p([0, T], E))$ . By these two facts and following the proof of Proposition 5.2 in [16], we get that for every  $t \in [0, T]$  the map  $x \mapsto (Y(\cdot, t, x), Z(\cdot, t, x))$  belongs to  $\mathcal{G}^1(E, \mathbb{K}_{\text{cont}}([0, T]))$ .  $\square$

*Remark 4.3.* By standard arguments,  $Y(t, t, x)$  is a deterministic function, and we have set  $v(t, x) = Y(t, t, x)$ . We notice that  $v$  depends on  $t, x, A, F, G, \phi$ , and  $\psi$  but not on the probability space and not on the Wiener process.

COROLLARY 4.4. Assume that Hypotheses 3.1, 3.2, and 3.11 hold true and that  $\psi$  and  $\phi$  satisfy Hypothesis 4.1. Then the function  $v(t, x) := Y(t, t, x)$  is continuous and for every  $t \in [0, T]$ ,  $v(t, \cdot)$  belongs to  $\mathcal{G}^1(E, \mathbb{R})$ , and there exists  $C > 0$  such that  $|\nabla_x v(t, x)h| \leq C\|h\|_E(1 + \|x\|_E^{(m+1)^2})$  for all  $t \in [0, T]$ ,  $x, h \in E$ .

*Proof.* It is a direct consequence of the previous proposition and of the properties of the evaluation map.  $\square$

COROLLARY 4.5. Assume that Hypotheses 3.1, 3.2, 3.11, 3.16, and 4.1 hold true and set  $v(t, x) := Y(t, t, x)$ ; then  $Z_s = \nabla v(s, X_s)G$ ,  $\mathbb{P}$ -a.s. and for almost every  $s \in [0, T]$ .

*Proof.* We apply Theorem 3.17 to the case when  $v(t, x) = Y(t, t, x)$ , where  $(Y, Z)$  is solution to (4.2). By Proposition 4.2 and Corollary 4.4,  $v(t, \cdot)$  belongs to  $\mathcal{G}^1(E)$ . Let  $\tau = t$  in (4.2). By the definition of  $v$  we can write

$$v(t, x) + \int_t^T Z_\sigma dW_\sigma = \phi(X_T) + \int_t^T \psi(\sigma, X_\sigma, Z_\sigma) d\sigma.$$

For  $t \leq s \leq T$  we consider the process

$$\begin{aligned} v(s, X(s, t, x)) &= - \int_s^T Z_\sigma dW_\sigma + \phi(X(T, s, X(s, t, x))) \\ &\quad + \int_s^T \psi(\sigma, X(\sigma, s, X(s, t, x)), Z_\sigma) d\sigma \\ &= \int_t^s Z_\sigma dW_\sigma - \int_t^s \psi(\sigma, X(\sigma, s, X(s, t, x)), Z_\sigma) d\sigma + v(t, x). \end{aligned}$$

Thus  $v(s, X(s, t, x))$  admits a representation analogous to the one in (3.7). Then by Theorem 3.17 and Remark 3.18 the corollary is proved.  $\square$

**5. The optimal control problem.** We formulate the optimal control problem in the weak sense following the approach of [15]; see, e.g., Chapter III. The main advantage is that we will be able to solve the closed loop equation in a weak sense, and hence find an optimal control, even if the feedback law is nonsmooth.

As in the previous section, in  $(\Omega, \mathcal{F}, \mathbb{P})$  we consider a cylindrical Wiener process  $\{W_\tau, \tau \geq 0\}$  in  $\Xi$  and  $(\mathcal{F}_\tau)_{\tau \geq 0}$  is its natural filtration, augmented in the usual way. We consider controlled stochastic evolution equations of the form

$$(5.1) \quad \begin{cases} dX_\tau^u = AX_\tau^u d\tau + F(\tau, X_\tau^u) d\tau + GdW_\tau + GR(\tau, X_\tau^u, u_\tau) d\tau, & \tau \in [t, T], \\ X_t^u = x. \end{cases}$$

The control  $u$  is an  $(\mathcal{F}_\tau)$ -predictable process taking values in a normed space  $U$ ; we call such control processes admissible controls and we denote by  $\mathcal{A}_d$  the set of such admissible controls.  $R$  is a map from  $[0, T] \times E \times U$  to  $\Xi$ . The solution of this equation will be denoted by  $X^u(\tau, t, x)$ , or simply by  $X_\tau^u$ .  $X$  denotes the state,  $u$  the control, and  $T > 0$ ,  $t \in [0, T]$  are fixed. We assume that  $R$  satisfies the following.

*Hypothesis 5.1.*  $R : [0, T] \times E \times U \longrightarrow \Xi$  is measurable and  $\|R(\tau, x, u)\|_\Xi \leq K_R$  for a suitable positive constant  $K_R > 0$  and every  $\tau \in [0, T]$ ,  $x \in E$ ,  $u \in U$ .

If  $A$ ,  $F$ , and  $G$  satisfy Hypotheses 3.1 and 3.2, and if  $R$  satisfies Hypothesis 5.1, then (5.1) admits a solution in the weak sense: Since by Hypothesis 5.1  $R$  is measurable and bounded, we can apply the Girsanov theorem. By the Girsanov theorem

the law of this solution depends only on  $x$ ,  $A$ ,  $F$ , and  $G$ , and so the weak solution is unique in law. Notice the occurrence of the operator  $G$  in the control term of (5.1). In the particular case  $R(\tau, x, u) = u$  the term  $u_\tau d\tau + dW_\tau$  admits a natural interpretation as a control affected by noise. In this case Hypothesis 5.1 holds if  $U$  is a bounded subset of  $\Xi$ .

We call  $(\Omega, \mathcal{F}, \mathcal{F}_\tau, \mathbb{P}, W)$  an admissible setup if  $(\Omega, \mathcal{F}, \mathbb{P})$  is a complete probability space with a filtration  $(\mathcal{F}_\tau)_{\tau \geq 0}$  satisfying the usual conditions, and  $\{W_\tau, \tau \geq 0\}$  is a cylindrical Wiener process taking values in a Hilbert space  $\Xi$ , with respect to the filtration  $(\mathcal{F}_\tau)_{\tau \geq 0}$ . We call  $(\Omega, \mathcal{F}, \mathcal{F}_\tau, \mathbb{P}, W, u, X^u)$  an admissible control system (a.c.s.), briefly denoted by  $(W, u, X^u)$ , if

- $(\Omega, \mathcal{F}, \mathcal{F}_\tau, \mathbb{P}, W)$  is an admissible setup;
- the control  $u$  is an  $(\mathcal{F}_\tau)$ -predictable process with values in  $U$ ;
- $X_\tau^u$  is an adapted process in  $E$  that solves (5.1) in the mild sense.

To every a.c.s. we associate the cost  $J$ :

$$(5.2) \quad J(t, x, (W, u, X^u)) = \mathbb{E} \int_t^T g(s, X_s^u, u_s) ds + \mathbb{E} \phi(X_T^u),$$

where  $g : [0, T] \times E \times U \rightarrow \mathbb{R}$  and  $\phi : E \rightarrow \mathbb{R}$  are continuous functions satisfying some growth conditions with respect to  $x \in E$  that will be specified in the following. We define the value function of the control problem as

$$V(t, x) = \inf_{(W, u, X^u) \text{ a.c.s.}} J(t, x, (W, u, X^u)).$$

The control problem in the weak sense is to find an a.c.s.  $(\overline{W}, \overline{u}, \overline{X}^u)$  such that

$$J(t, x, (\overline{W}, \overline{u}, \overline{X}^u)) \leq J(t, x, (W, u, X^u))$$

for every a.c.s.  $(W, u, X^u)$ . Then  $(\overline{W}, \overline{u}, \overline{X}^u)$  is called optimal. In this case the value function is given by  $V(t, x) = J(t, x, (\overline{W}, \overline{u}, \overline{X}^u))$ .

For all  $\tau \in [0, T]$ ,  $x \in E$ ,  $z \in \Xi^*$ , we define in a classical way the Hamiltonian function  $\psi : [0, T] \times E \times \Xi^* \rightarrow \mathbb{R}$ :

$$(5.3) \quad \psi(\tau, x, z) = \inf \{g(\tau, x, u) + zR(\tau, x, u) : u \in U\}.$$

We define the, possibly empty, set

$$(5.4) \quad \Gamma(\tau, x, z) = \{u \in U : g(\tau, x, u) + zR(\tau, x, u) = \psi(\tau, x, z)\}.$$

We remark that by the Filippov theorem (see, e.g., [1, Theorem 8.2.10, p. 316] if  $\Gamma$  is nonempty there exists a Borel measurable map  $\Gamma_0 : [0, T] \times E \times \Xi^* \rightarrow U$  such that, for  $\tau \in [0, T]$ ,  $x \in E$ , and  $z \in \Xi^*$ , we have  $\Gamma_0(\tau, x, z) \in \Gamma(\tau, x, z)$ .

In order to solve the control problem, we will not solve the associated Hamilton–Jacobi–Bellman equation, but we will use the BSDEs, which turns out to be the adequate approach in this situation, since the forward equation evolves in the Banach space  $E$  and the cost functional is well defined in the Banach space  $E$ .

In this paper we treat control problems related to current and final costs which can have polynomial growth with respect to  $x$ . We assume that  $A$ ,  $F$ , and  $G$  in the controlled equation (5.1) satisfy Hypothesis 3.1, 3.2, and 3.11.

On the cost functional defined in (5.2) we make the following assumptions.



*Hypothesis 5.2.* The function  $g : [0, T] \times E \times U \rightarrow \mathbb{R}$  is measurable, and for almost all (a.a.)  $\tau \in [0, T]$ , the map  $g(\tau, \cdot, \cdot) : E \times U \rightarrow \mathbb{R}$  is continuous. Moreover, for some  $j \geq 0$ ,  $|g(\tau, x, u)| \leq c_2(1 + \|x\|_E^j)$ . The function  $\phi : E \rightarrow \mathbb{R}$  is in  $\mathcal{G}^1(E)$  and  $|\phi(x)| \leq c_3(1 + \|x\|_E^j)$ .

We remark that by this hypothesis on  $g$ , the Hamiltonian defined in (5.3) is finite. Moreover, also by Hypothesis 5.1 on  $R$ , there exists  $L > 0$  such that

$$|\psi(\sigma, x, z_1) - \psi(\sigma, x, z_2)| \leq L \|z_1 - z_2\|_{\Xi^*}$$

for every  $\sigma \in [0, T]$ ,  $x \in E$ , and  $z_1, z_2 \in \Xi^*$ . We need that the Hamiltonian  $\psi$  satisfies some regularity assumptions that guarantee differentiable dependence of the solution  $(Y_\tau, Z_\tau)$  with respect to  $x$ , the initial datum of the forward equation (3.1). Namely, the Hamiltonian  $\psi$  defined in (5.3) has to satisfy Hypothesis 4.1, points 2 and 3.

*Remark 5.3.* We clarify why we need the assumptions stated previously on the running and final costs and on the Hamiltonian. By Hypothesis 5.2, namely, by the polynomial growth of  $g$ , together with Lemma 3.12, the cost is well defined. Differentiability assumptions on the Hamiltonian and on the final cost are quite restrictive but are essential in the BSDE approach: The value function is defined, as in Corollary 4.4, by means of the solution of a suitable BSDE. Essential tools are Theorem 3.17 and Corollary 4.5 applied to the value function. Moreover (see also section 6), by the BSDE approach, without requiring any regularizing property of the transition semigroup, we are able to find a unique, Gâteaux differentiable mild solution of the corresponding Hamilton–Jacobi–Bellman equation. We can achieve Gâteaux differentiability of this mild solution by taking  $\psi$  and  $\phi$  Gâteaux differentiable.

*Example 5.4.* Hypothesis 4.1 for the Hamiltonian function can be verified in some concrete cases. We present a situation where we can directly verify that the Hamiltonian  $\psi$  satisfies Hypothesis 4.1, points 2 and 3. We take  $R(t, x, u) = u$ : The space  $U$  coincides with  $\Xi$  and the set of admissible controls is given by  $A_d = \{u \in \Xi : \|u\|_\Xi \leq \delta\}$  for some fixed  $\delta > 0$ . Moreover, we consider a current cost given by  $g(t, x, u) = g_0(\|u\|_\Xi^\alpha) + g_1(t, x)$ , where  $\alpha > 1$ ,  $g_0 \in C^1(\mathbb{R}^+, \mathbb{R}^+)$  convex, and  $g_0'(0) > 0$ , and for every  $t \in [0, T]$ ,  $g_1(t, \cdot) \in \mathcal{G}^1(E)$ , and there exist  $L > 0$  and  $m \geq 0$  such that

$$|\nabla_x g(t, x) h| \leq L \|h\|_E (1 + \|x\|_E)^m$$

for every  $h \in E$ . It turns out also that  $\psi(t, x, z)$  is differentiable with respect to  $z$  and so  $\psi$  satisfies Hypothesis 4.1, points 2 and 3.

Now let us consider a complete probability space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$  and a cylindrical Wiener process  $\{\tilde{W}_\tau, \tau \geq 0\}$  with values in  $\Xi$ . In  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$  we consider the following forward-backward system, where  $(X, Y, Z) \in E \times \mathbb{R} \times \Xi^*$ :

$$\begin{cases} dX_\tau = AX_\tau d\tau + F(\tau, X_\tau) d\tau + G d\tilde{W}_\tau, & \tau \in [t, T], \\ dY_\tau = -\psi(\tau, X_\tau, Z_\tau) d\tau + Z_\tau d\tilde{W}_\tau, & \tau \in [t, T], \\ X_t = x, \\ Y_T = \phi(X_T). \end{cases}$$

By Remark 4.3, if we define

$$(5.5) \quad v : [0, T] \times E \rightarrow \mathbb{R}, \quad v(t, x) = Y(t, t, x),$$

$v$  does not depend either on the space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$  or on the Wiener process  $\{\tilde{W}_\tau, \tau \geq 0\}$ .

In the following proposition we prove the so-called fundamental relation in terms of BSDEs.

**PROPOSITION 5.5.** *Assume Hypotheses 3.1, 3.2, 3.11, 5.1, and 5.2 and that the Hamiltonian  $\psi$  defined in (5.3) satisfies Hypothesis 4.1, points 2 and 3. Let  $v$  be defined in (5.5). Then for every  $t \in [0, T]$  and  $x \in E$ , and for every a.c.s.  $(W, u, X^u)$ , we have  $J(t, x, (W, u, X^u)) \geq v(t, x)$ .*

*Proof.* Let  $X_\tau^u$  be the solution to (5.1) corresponding to the control  $u$ . We define the process

$$W_\tau^u = W_\tau + \int_{t \wedge \tau}^\tau R(\sigma, X_\sigma^u, u_\sigma) d\sigma, \quad \tau \in [0, T],$$

and we note that  $X^u$  solves the equation

$$\begin{cases} dX_\tau^u = AX_\tau^u d\tau + F(\tau, X_\tau^u) d\tau + GdW_\tau^u, & \tau \in [t, T], \\ X_t^u = x. \end{cases}$$

Since  $R$  is bounded, we can apply the Girsanov theorem and deduce that there exists a probability measure  $\mathbb{P}^u$  on  $(\Omega, \mathcal{F})$  such that  $W^u$  is a Wiener process with respect to  $\mathbb{P}^u$ . In  $(\Omega, \mathcal{F}, \mathbb{P}^u)$  let us consider the backward equation for the unknown process  $(Y_\tau^u, Z_\tau^u)$ ,  $\tau \in [t, T]$ :

$$(5.6) \quad Y_\tau^u + \int_\tau^T Z_\sigma^u dW_\sigma^u = \phi(X_T^u) + \int_\tau^T \psi(\sigma, X_\sigma^u, Z_\sigma^u) d\sigma, \quad \tau \in [t, T].$$

Let  $\tau = t$  in (5.6). It turns out that  $Y^u(t, t, x)$ , which is deterministic, depends on only  $t, x, F, G, \phi$ , and  $\psi$  and not on the a.c.s.  $(W, u, X^u)$ ; indeed, in (5.6) the dependence on the control system has disappeared. Thus  $Y^u(t, t, x) = v(t, x)$ , where  $v$  is defined as in (5.5). Now we notice that  $W^u$  is not a  $\mathbb{P}$ -Wiener process, so  $\int_\tau^T Z_\sigma^u dW_\sigma^u$  is not a  $\mathbb{P}$ -martingale; nevertheless, we can state that  $\int_\tau^T Z_\sigma^u dW_\sigma$  is a  $\mathbb{P}$ -martingale. Indeed, we can prove that

$$\mathbb{E} \left( \int_t^T \|Z_\sigma^u\|_{\Xi^*}^2 d\sigma \right)^{1/2} < \infty.$$

We remember that

$$\frac{d\mathbb{P}^u}{d\mathbb{P}} = \exp \left( - \int_t^T R(\sigma, X_\sigma^u, u_\sigma) dW_\sigma - \frac{1}{2} \int_t^T \|R(\sigma, X_\sigma^u, u_\sigma)\|_{\Xi}^2 d\sigma \right).$$

We denote  $\frac{d\mathbb{P}^u}{d\mathbb{P}}$  by  $\rho$ . We estimate

$$\begin{aligned} \mathbb{E} \left( \int_t^T \|Z_\sigma^u\|_{\Xi}^2 d\sigma \right)^{1/2} &= \mathbb{E}^u \left[ \left( \int_t^T \|Z_\sigma^u\|_{\Xi}^2 d\sigma \right)^{1/2} \rho^{-1} \right] \\ &\leq \mathbb{E}^u \left[ \left( \int_t^T \|Z_\sigma^u\|_{\Xi}^2 d\sigma \right)^{1/2} \right] \mathbb{E}^u [\rho^{-2}]^{1/2}. \end{aligned}$$

It remains to prove that  $\mathbb{E}^u [\rho^{-2}]$  is bounded, knowing that

$$\mathbb{E}^u [\rho^{-1}] = \mathbb{E}^u \left[ \exp \left( \int_t^T R(\sigma, X_\sigma^u, u_\sigma) dW_\sigma^u - \frac{1}{2} \int_t^T \|R(\sigma, X_\sigma^u, u_\sigma)\|_\Xi^2 d\sigma \right) \right] = 1.$$

We get

$$\begin{aligned} \mathbb{E}^u [\rho^{-2}] &= \mathbb{E}^u \left[ \exp 2 \left( \int_t^T R(\sigma, X_\sigma^u, u_\sigma) dW_\sigma^u - \frac{1}{2} \int_t^T \|R(\sigma, X_\sigma^u, u_\sigma)\|_\Xi^2 d\sigma \right) \right] \\ &= \mathbb{E}^u \left[ \exp \left( \int_t^T 2R(\sigma, X_\sigma^u, u_\sigma) dW_\sigma^u - \frac{1}{2} \int_t^T 4 \|R(\sigma, X_\sigma^u, u_\sigma)\|_\Xi^2 d\sigma \right) \right. \\ &\quad \left. \times \exp \left( \int_t^T \|R(\sigma, X_\sigma^u, u_\sigma)\|_\Xi^2 d\sigma \right) \right] \\ &\leq \exp \left( T (K_R \mathcal{K}_u)^2 \right). \end{aligned}$$

Then, by applying the Burkholder–Davis–Gundy inequality,  $\int_t^T Z_\sigma dW_\sigma$  is a  $\mathbb{P}$ -martingale, so the stochastic integral in (5.6) has zero expectation with respect to the original probability  $\mathbb{P}$ . If in (5.6) we set  $\tau = t$  and take expectation with respect to the original probability  $\mathbb{P}$ , we obtain

$$(5.7) \quad v(t, x) = \mathbb{E} \phi(X_T^u) + \mathbb{E} \int_t^T [\psi(\sigma, X_\sigma^u, Z_\sigma^u) - Z_\sigma^u R(\sigma, X_\sigma^u, u_\sigma)] d\sigma.$$

Adding and subtracting  $\mathbb{E} \int_t^T g(\sigma, X_\sigma^u, u_\sigma) d\sigma$ , we arrive at

$$(5.8) \quad \begin{aligned} v(t, x) &= J(t, x, (W, u, X^u)) \\ &\quad + \mathbb{E} \int_t^T [\psi(\sigma, X_\sigma^u, Z_\sigma^u) - Z_\sigma^u R(\sigma, X_\sigma^u, u_\sigma) - g(\sigma, X_\sigma^u, u_\sigma)] d\sigma. \end{aligned}$$

By the definition of  $\psi$ , the term in the square brackets is nonpositive, and consequently,

$$J(t, x, (W, u, X^u)) \geq v(t, x). \quad \square$$

Relation (5.8) is a version of the fundamental relation in terms of BSDEs. We immediately deduce the following consequences.

**COROLLARY 5.6.** *Let  $t \in [0, T]$  and  $x \in E$  be fixed. If, for an a.c.s.  $(W, u, X^u)$ , we have  $J(t, x, (W, u, X^u)) = v(t, x)$ , then  $(W, u, X^u)$  is optimal for the control problem starting from  $x$  at time  $t$ . If  $\Gamma$  is nonempty, let us denote by  $\Gamma_0 : [0, T] \times E \times \Xi^* \longrightarrow U$  a Borel measurable map such that, for  $\tau \in [0, T]$ ,  $x \in E$ ,  $z \in \Xi^*$ , we have  $\Gamma_0(\tau, x, z) \in \Gamma(\tau, x, z)$ . If  $u$  is an admissible control satisfying*

$$(5.9) \quad u_\tau = \Gamma_0(\tau, X_\tau^u, Z_\tau^u), \quad \mathbb{P}\text{-a.s. for almost every } \tau \in [t, T],$$

*then  $J(t, x, (W, u, X^u)) = v(t, x)$ , and  $(W, u, X^u)$  is optimal.*

If also Hypothesis 3.16 holds true, by Corollary 4.5 relation (5.8) can be rewritten as

$$(5.10) \quad \begin{aligned} v(t, x) = & J(t, x, (W, u, X^u)) \\ & + \mathbb{E} \int_t^T [\psi(\sigma, X_\sigma^u, \nabla v(\sigma, X_\sigma^u) G) \\ & - \nabla v(\sigma, X_\sigma^u) GR(\sigma, X_\sigma^u, u_\sigma) - g(\sigma, X_\sigma^u, u_\sigma)] d\sigma. \end{aligned}$$

In order to prove the existence of an optimal control, we say that if there exists a measurable selection  $\Gamma_0$  of  $\Gamma$ , then we can study the closed loop equation

$$(5.11) \quad \begin{cases} dX_\tau^u = AX_\tau^u d\tau + F(\tau, X_\tau^u) d\tau + GR(\tau, X_\tau^u, \Gamma_0(\tau, X_\tau^u, \nabla v(\tau, X_\tau^u) G)) d\tau \\ \quad + GdW_\tau, \quad \tau \in [t, T], \\ X_t^u = x. \end{cases}$$

We can prove the main result of this subsection.

**THEOREM 5.7.** *Let  $v$  be defined as in Proposition 5.5. Assume that Hypotheses 3.1, 3.2, 3.11, 3.16, 5.1, and 5.2 hold true and that the Hamiltonian  $\psi$  defined in (5.3) satisfies Hypothesis 4.1, points 2 and 3. For all a.c.s.  $(W, u, X^u)$  we have*

$$J(t, x, (W, u, X^u)) \geq v(t, x),$$

and the equality holds if and only if

$$(5.12) \quad u_\tau \in \Gamma(\tau, X_\tau^u, \nabla v(\tau, X_\tau^u) G), \quad \mathbb{P}\text{-a.s. for a.a. } \tau \in [t, T].$$

Moreover, let us denote by  $\Gamma_0$  a measurable selection of  $\Gamma$ ; then an a.c.s. satisfying the feedback law

$$(5.13) \quad u_\tau = \Gamma_0(\tau, X_\tau^u, \nabla v(\tau, X_\tau^u) G), \quad \mathbb{P}\text{-a.s. for a.a. } \tau \in [t, T],$$

is optimal. The closed loop equation,

$$(5.14) \quad \begin{cases} dX_\tau^* = [AX_\tau^* + F(\tau, X_\tau^*) + GR(\tau, X_\tau^*, \Gamma_0(\tau, X_\tau^*, \nabla v(\tau, X_\tau^*) G))] d\tau + GdW_\tau, \\ X_t^* = x, \quad \tau \in [t, T], \end{cases}$$

admits a weak solution, which is unique in law, and the corresponding a.c.s. is optimal.

*Proof.* The proof follows from the fundamental relation (5.10) and by Corollary 5.6. The closed loop equation can be solved in the weak sense via a Girsanov change of measure. Namely, we fix an arbitrary cylindrical Wiener process  $\{W_\tau, \tau \geq 0\}$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and with values in  $\Xi$ . Let  $\bar{X}_\tau^*$  be the mild solution to

$$(5.15) \quad \begin{cases} d\bar{X}_\tau^* = [A\bar{X}_\tau^* + F(\tau, \bar{X}_\tau^*)] d\tau + G(\tau, \bar{X}_\tau^*) dW_\tau, \\ \bar{X}_t^* = x. \end{cases}$$

Let  $\widehat{\mathbb{P}}$  be the probability on  $\Omega$  under which

$$\widehat{W}_\tau := - \int_t^\tau R(s, X_s^*, \Gamma_0(s, X_s^*, \nabla v(s, X_s^*) G)) ds + W_\tau$$

is a Wiener process. Then  $\overline{X}^*$  is the mild solution to (5.14) relative to the probability  $\widehat{\mathbb{P}}$  and to the Wiener process  $\widehat{W}$ . The law of  $\overline{X}^*$  is unique since it depends on  $A$ ,  $F$ , and  $G$ . The closed loop equation (5.14) always admits a solution in the weak sense. So if  $\Gamma$  is nonempty, then by the Filippov theorem a measurable selection  $\Gamma_0$  of  $\Gamma$  exists, and it is possible to perform the synthesis of the optimal control.  $\square$

**6. Mild solutions of the Hamilton–Jacobi–Bellman equation.** In this section we find mild solutions of the Hamilton–Jacobi–Bellman equation associated with this control problem. Hamilton–Jacobi–Bellman equations associated with the control problem we have already treated are formally given by

$$(6.1) \quad \begin{cases} \frac{\partial v}{\partial t}(t, x) = -\mathcal{A}_t v(t, x) - \psi(t, x, \nabla v(t, x) G), & t \in [0, T], x \in E, \\ v(T, x) = \phi(x). \end{cases}$$

The linear operator  $\mathcal{A}_t$  is formally defined by

$$\mathcal{A}_t f(x) = \frac{1}{2} \text{Trace}_H (GG^* \nabla^2 f(x)) + \langle Ax, \nabla f(x) \rangle_{E, E^*} + \langle F(t, x), \nabla f(x) \rangle_{E, E^*}.$$

We do not look for classical solutions of (6.1); rather, we want to find mild solutions, which we are going to define. In Hilbert spaces, the problem of solving (6.1) has already been treated by an analytical approach in many papers; we cite [17], where  $G$  is such that the Ornstein–Uhlenbeck semigroup is regularizing, namely, in the finite dimensional case; this is equivalent to asking that the second order differential operator  $\mathcal{A}_t$  be hypoelliptic in the sense of Hormander. We also cite [24], where the case of  $G$  not necessarily constant is treated and where, moreover, less restrictive regularizing properties on the associated Ornstein–Uhlenbeck semigroup are required. In [16], (6.1) is studied in the more general case of  $G$  not necessarily constant, and without any nondegeneracy assumptions, via the BSDE approach. As far as we know, apart from the papers [7] and [8], where viscosity solutions are presented for first order partial differential equations, there are no results in the literature concerning mild solutions for (6.1) considered in a Banach space.

In this paper we consider the case when  $F$  in the forward equation has polynomial growth with respect to  $x$ , and thus we also can allow the running cost and the final cost to have polynomial growth with respect to  $x$ .

Let  $X(\tau, t, x)$  be the solution to (3.1) when  $A$ ,  $F$ , and  $G$  satisfy Hypotheses 3.1, 3.2, and 3.11. We remember that this solution is an  $E$ -valued Markov process. We can define the transition semigroup on continuous and bounded functions  $\varphi : E \rightarrow \mathbb{R}$  as

$$P_{t, \tau} [\varphi] (x) = \mathbb{E} \varphi (X (\tau, t, x)).$$

Moreover, the semigroup is also well defined on continuous functions  $\varphi : E \rightarrow \mathbb{R}$  with polynomial growth with respect to  $x$ . Indeed, let  $\varphi : E \rightarrow \mathbb{R}$  such that  $|\varphi(x)| \leq c(1 + \|x\|_E^k)$ . Then

$$\begin{aligned} |P_{t, \tau} [\varphi] (x)| &\leq \mathbb{E} |\varphi (X (\tau, t, x))| \leq \sup_{x \in E} \frac{|\varphi(x)|}{1 + \|x\|_E^k} \mathbb{E} (1 + \|X (\tau, t, x)\|_E^k) \\ &\leq c \sup_{y \in E} \frac{|\varphi(y)|}{1 + \|y\|_E^k}, \end{aligned}$$

where the last passage follows from the fact that, for every  $p \geq 1$ ,  $X \in \mathcal{H}^p([0, T], E)$ .

We look for mild solutions of (6.1), that is, functions  $v(t, x)$  satisfying

$$(6.2) \quad v(t, x) = P_{t,T}[\phi](x) + \int_t^T P_{t,\tau}[\psi(\tau, \cdot, \nabla v(\tau, \cdot)G)](x) d\tau, \quad t \in [0, T], \quad x \in E.$$

We notice that this formula is meaningful if  $v$  has only the first derivative  $\nabla v(t, x)$ , and provided  $\phi$  and  $\psi$  satisfy some growth and measurability conditions. Namely, we give the following definition of a mild solution of the Hamilton–Jacobi–Bellman equation (6.1).

**DEFINITION 6.1.** *A function  $v : [0, T] \times E \rightarrow \mathbb{R}$  is a mild solution of the Hamilton–Jacobi–Bellman equation (6.1) if the following are satisfied:*

1.  *$v$  is continuous, for every  $t \in [0, T]$  we have  $v(t, \cdot) \in \mathcal{G}^1(E)$ , and the map  $(t, x) \mapsto \nabla v(t, x)$  is measurable from  $[0, T] \times E$  with values in  $E^*$ ;*
2. *there exists  $C > 0$  such that  $|v(t, x)| \leq C(1 + \|x\|_E^j)$  and  $|\nabla_x v(t, x)h| \leq C\|h\|_E(1 + \|x\|_E^k)$  for every  $t \in [0, T]$ ,  $x, h \in E$ , and some positive integers  $j$  and  $k$ ;*
3. *equality (6.2) holds.*

In order to solve (6.1) we consider the forward-backward system

$$(6.3) \quad \begin{cases} dX_\tau = AX_\tau d\tau + F(\tau, X_\tau) d\tau + GdW_\tau, & \tau \in [t, T], \\ dY_\tau = -\psi(\tau, X_\tau, Z_\tau) d\tau + Z_\tau dW_\tau, & \tau \in [t, T], \\ X_t = x, \\ Y_T = \phi(X_T). \end{cases}$$

We are ready to prove that there exists a unique solution of (6.3).

**THEOREM 6.2.** *Assume that Hypotheses 3.1, 3.2, 3.11, and 3.16 hold true, and let  $\phi$  and  $\psi$  satisfy Hypothesis 4.1. Then there exists a unique mild solution of the Hamilton–Jacobi–Bellman equation (6.1), given by the formula*

$$v(t, x) = Y(t, t, x),$$

where  $(X, Y, Z)$  is the solution of the forward-backward system (6.3).

*Proof.* The main ideas of the proof are essentially taken from [16, Theorem 6.2] and adapted to this different context.

First, we prove existence. We want to prove that the function  $v(t, x) := Y(t, t, x)$  is in fact a mild solution to (6.1). We have already proved in Proposition 4.2 that for every  $t \in [0, T]$ ,  $v(t, \cdot) \in \mathcal{G}^1(E)$  and that there exists  $C > 0$  such that  $|\nabla_x v(t, x)h| \leq C\|h\|_E(1 + \|x\|_E^j)$  and  $|\nabla_x v(t, x)h| \leq C\|h\|_E(1 + \|x\|_E^k)$  for every  $t \in [0, T]$ ,  $x, h \in E$ , and some positive integers  $j$  and  $k$ . It remains to prove that  $v$  satisfies equality (6.2). To this end we evaluate

$$\begin{aligned} P_{t,\tau}[\psi(\tau, \cdot, \nabla v(\tau, \cdot)G)](x) &= \mathbb{E}[\psi(\tau, X(\tau, t, x), \nabla_x Y(\tau, X(\tau, t, x))G)] \\ &= \mathbb{E}[\psi(\tau, X(\tau, t, x), \nabla_x Y(\tau, t, x)G)] \\ &= \mathbb{E}[\psi(\tau, X(\tau, t, x), Z(\tau, t, x))], \end{aligned}$$

where the last equality is a consequence of Corollary 4.5. In particular, we obtain

$$(6.4) \quad \int_t^T P_{t,\tau}[\psi(\tau, \cdot, \nabla v(\tau, \cdot)G)](x) d\tau = \mathbb{E} \int_t^T \psi(\tau, X(\tau, t, x), Z(\tau, t, x)) d\tau.$$

Next, the pair  $(Y, Z)$  is a solution to the backward equation in the forward-backward system (6.1),

$$Y(t, t, x) + \int_t^T Z(\tau, t, x) dW_\tau = \phi(X(T, t, x)) + \int_t^T \psi(\tau, X(\tau, t, x), Z(\tau, t, x)) d\tau.$$

Taking expectation and applying formula (6.4), we get the integral formula (6.2).

It remains to prove uniqueness. Let  $v$  be a mild solution. Since  $v$  is a mild solution to (6.1), then for every  $s \in [t, T]$ ,

$$v(s, x) = \mathbb{E}\phi(X(T, s, x)) + \mathbb{E} \int_s^T \psi(\tau, X(\tau, s, x), \nabla_x v(\tau, X(\tau, s, x))) G d\tau,$$

and since  $X(\tau, s, x)$  is independent of  $\mathcal{F}_s$ , the expectation coincides with the conditional expectation given  $\mathcal{F}_s$ ; next we can replace  $x$  with  $X(s, t, x)$ , since  $X(s, t, x)$  is  $\mathcal{F}_s$ -measurable. So we obtain

$$\begin{aligned} v(s, X(s, t, x)) &= \mathbb{E}^{\mathcal{F}_s} \phi(X(T, t, x)) \\ &\quad + \mathbb{E}^{\mathcal{F}_s} \int_s^T \psi(\tau, X(\tau, t, x), \nabla_x v(\tau, X(\tau, t, x))) G d\tau \\ &= \mathbb{E}^{\mathcal{F}_s} \eta - \mathbb{E}^{\mathcal{F}_s} \int_t^s \psi(\tau, X(\tau, t, x), \nabla_x v(\tau, X(\tau, t, x))) G d\tau, \end{aligned}$$

where we have defined  $\eta = \phi(X(T, t, x)) + \int_t^T \psi(\tau, X(\tau, t, x), \nabla_x v(\tau, X(\tau, t, x))) G d\tau$ . By the representation theorem of martingales (see, e.g., [12, Theorem 8.2]), there exists a process  $\tilde{Z} \in L^2_{\mathcal{P}}(\Omega \times [0, T], L_2(\Xi, \mathbb{R}))$  such that  $\mathbb{E}^{\mathcal{F}_s} \eta = \int_t^s \tilde{Z}_\tau dW_\tau + v(t, x)$ . Thus

$$(6.5) \quad v(s, X(s, t, x)) = v(t, x) + \int_t^s \tilde{Z}_\tau dW_\tau - \int_t^s \psi(\tau, X(\tau, t, x), \nabla_x v(\tau, X(\tau, t, x))) G d\tau.$$

By Corollary 4.5, equality (6.5) can be rewritten as

$$\begin{aligned} v(s, X(s, t, x)) &= v(t, x) + \int_t^s \nabla_x v(\tau, X(\tau, t, x)) G dW_\tau \\ &\quad - \int_t^s \psi(\tau, X(\tau, t, x), \nabla_x v(\tau, X(\tau, t, x))) G d\tau \\ &= \phi(X(T, t, x)) - \int_s^T \nabla_x v(\tau, X(\tau, t, x)) G dW_\tau \\ &\quad + \int_s^T \psi(\tau, X(\tau, t, x), \nabla_x v(\tau, X(\tau, t, x))) G d\tau. \end{aligned}$$

By comparing with the backward equation in (6.3), we see that the pairs of processes

$$(Y(s, t, x), Z(s, t, x))_{t \leq s \leq T}$$

and

$$(v(s, X(s, t, x)), \nabla_x v(s, X(s, t, x)) G)_{t \leq s \leq T}$$

solve the same equation. By uniqueness of the solution of this backward equation we have, in particular, that  $Y(s, t, x) = v(s, X(s, t, x))$ ,  $t \leq s \leq T$ , and for  $s = t$  we get the desired equality  $Y(t, t, x) = v(t, x)$ .  $\square$

**7. Application to some specific models.** In this section we introduce some stochastic controlled equations which have the structure of (5.1) and can be treated with our techniques. For the sake of simplicity, we consider bounded costs and a coefficient  $F$  in the forward equation with polynomial growth with respect to  $x$ .

**7.1. The controlled semilinear heat equation: Dirichlet boundary conditions, Neumann boundary conditions.** In this subsection we show how our results can be applied to perform the synthesis of the optimal control when the state equation is a general semilinear heat equation with additive noise in one space dimension. As we will see, we consider the space of continuous functions, where the heat semigroup with Dirichlet boundary conditions turns out to be analytic but not strongly continuous. Heat equations, and in general, reaction diffusion equations, arise naturally in applications; for example, reaction diffusion equations with Neumann boundary conditions arise naturally in chemical reactions, and we think that, in view of applications, it is interesting to treat optimal control problems related to them in the Banach space of continuous functions: this allows us, for instance, to control some variable of the state, say, the temperature, in a finite number of points; see below for a mathematical formulation of this problem.

We are given a complete probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  with a filtration  $(\mathcal{F}_\tau)_{\tau \geq 0}$  satisfying the usual conditions. We consider, for  $\tau \in [t, T]$  and  $\xi \in [0, 1]$ , the equation

$$(7.1) \quad \begin{cases} d_\tau X^u(\tau, \xi) = \left[ \frac{\partial^2}{\partial \xi^2} X^u(\tau, \xi) + f(\tau, \xi, X^u(\tau, \xi)) + \sigma(\xi) r(\xi) u(\tau, \xi) \right] d\tau \\ \quad + \sigma(\xi) \dot{W}(\tau, \xi) d\tau, \\ X^u(\tau, 0) = X^u(\tau, 1) = 0, \\ X^u(t, \xi) = x_0(\xi), \end{cases}$$

where  $\dot{W}(\tau, \xi)$  is a space-time white noise on  $[0, T] \times [0, 1]$ .

We introduce the cost functional

$$(7.2) \quad J(t, x, u) = \mathbb{E} \int_t^T \int_0^1 l(s, \xi, X^u(s, \xi), u) \mu(d\xi) ds + \mathbb{E} \int_0^1 k(\xi, X^u(T, \xi)) \mu(d\xi)$$

that we minimize over all admissible setups. Here  $\mu$  is a finite regular measure on  $[0, 1]$ . An admissible control  $u(\tau, \xi)$  is a predictable process such that  $u(\tau, \cdot) \in L^2([0, 1])$ , and  $|u(\tau, \xi)| \leq \delta$ . We denote by  $\mathcal{U}$  the set of such admissible controls. The cost introduced in (7.2) is well defined on the space of continuous function on the interval  $[0, 1]$ , but for an arbitrary  $\mu$  it is not well defined on the Hilbert space of square integrable functions. So this example makes it clear that with this theory we can treat a larger class of stochastic optimal control problems than the class of problems covered by the Hilbert space theory. Stochastic optimal control problems for reaction diffusion equations have been extensively studied in the literature; see, e.g., [18] and [23]. We cite in particular the papers [3] and [4]. In these works, equations with a more general structure than (7.1) are treated, but the costs are defined on the Hilbert space of square integrable functions: In particular, in [3] also the reaction diffusion equation is studied in this Hilbert space, while in [4] the equation is considered in the space of continuous functions, and the cost is defined in the space of square integrable functions.

To satisfy our Hypotheses 3.1, 3.2, 3.11, 5.1, and 5.2 we have to assume the following.



*Hypothesis 7.1.* The functions  $f$ ,  $\sigma$ ,  $r$ ,  $l$ ,  $k$  are all Borel measurable and real valued. Moreover,

1.  $f : [0, T] \times [0, 1] \times \mathbb{R} \longrightarrow \mathbb{R}$  is continuous; for every  $\tau \in [0, T]$  and every  $\xi \in [0, 1]$ , we have  $f(\tau, \xi, \cdot) \in C^1(\mathbb{R})$ ; and there exists  $c_1$  continuous on  $[0, 1]$  such that

$$|f(\tau, \xi, x)| \leq c_1(\xi) (1 + |x|^K), \quad |\nabla_x f(\tau, \xi, x) h| \leq c_1(\xi) |h| (1 + |x|^K)$$

for every  $\tau \in [0, T]$ ,  $\xi \in [0, 1]$ ,  $x, h \in \mathbb{R}$ . For every  $\tau \in [0, T]$ , for every  $\xi \in [0, 1]$ , for every  $x, y \in \mathbb{R}$ , and for every  $\alpha > 0$ ,

$$|x - y| \leq |x - y - \alpha(f(\tau, \xi, x) - f(\tau, \xi, y))|;$$

2.  $\sigma \in C([0, 1], \mathbb{R})$ ;
3.  $r \in L^\infty([0, 1], \mathbb{R})$ ;
4.  $l : [0, T] \times [0, 1] \times \mathbb{R} \times \mathcal{U} \rightarrow \mathbb{R}$  is continuous and bounded.
5.  $k : [0, 1] \times \mathbb{R} \rightarrow \mathbb{R}$  is continuous and bounded, and  $k(\xi, \cdot) \in C_b^1(\mathbb{R})$ ; moreover, there exists  $c_2$  continuous on  $[0, 1]$  such that  $|\nabla_x k(\xi, x)| \leq c_2(\xi)$ ;
6.  $x_0 \in C([0, 1])$ .

A classical example of a map  $f$  satisfying Hypothesis 7.1, point 1, is  $f(\tau, \xi, x) = f(x) = -x^3$ .

In the following we denote  $C_0([0, 1]) = \{f \in C([0, 1]) : f(0) = f(1) = 0\}$ . To rewrite the problem in an abstract way we set  $H = \Xi = L^2([0, 1])$  and  $E = C([0, 1])$ : The generator  $A$  in  $H$  is given by

$$\mathcal{D}(A) = H^2([0, 1]) \cap H_0^1([0, 1]), \quad (Ay)(\xi) = \frac{\partial^2}{\partial \xi^2} y(\xi) \quad \text{for every } y \in \mathcal{D}(A),$$

and its restriction to  $E$  is given by

$$\mathcal{D}(A) = C^2([0, 1]) \cap C_0([0, 1]), \quad (Ay)(\xi) = \frac{\partial^2}{\partial \xi^2} y(\xi) \quad \text{for every } y \in \mathcal{D}(A).$$

We set

$$(7.3) \quad \begin{aligned} F(\tau, x)(\xi) &= f(\tau, \xi, x(\xi)), & (Gz)(\xi) &= \sigma(\xi) z(\xi), \\ (Ru)(\xi) &= r(\xi) u(\xi), \\ g(\tau, x, u)(\xi) &= \int_0^1 l(s, \xi, x(\xi), u) \mu(d\xi), & \phi(x)(\xi) &= \int_0^1 k(\xi, x(\xi)) \mu(d\xi) \end{aligned}$$

for a.a.  $\tau \in [0, T]$  and  $\xi \in [0, 1]$ , for all  $x, z \in C([0, 1])$ , and for all admissible controls  $u$ . It turns out that, with Hypothesis 7.1, the assumptions in Hypotheses 3.1, 3.2, 3.11, 5.1, and 5.2 are satisfied.

Also Hypothesis 3.16 is satisfied by taking  $\Xi_0 = C([0, 1])$ .

With notations (7.3), the heat equation (7.1) can be written in an abstract way in the Banach space  $E$  as

$$(7.4) \quad \begin{cases} dX_\tau^u = [AX_\tau^u + F(\tau, X_\tau^u)] d\tau + GRu_\tau d\tau + GdW_\tau, & \tau \in [t, T], \\ X_t^u = x_0, \end{cases}$$

where  $\{W_\tau, \tau \geq 0\}$  is a cylindrical Wiener process in  $H$  with respect to the filtration  $(\mathcal{F}_\tau)_{\tau \geq 0}$ , and it is such that the stochastic convolution admits a version in  $C([0, T], E)$ ; see, e.g., [12, Chapter 5].  $(\Omega, \mathcal{F}, \mathcal{F}_\tau, \mathbb{P}, W, u, X^u)$  is an a.c.s.

*Remark 7.2.* We remark that, in order to guarantee more generality, in this paper the nonlinear term  $F$  is assumed to be dissipative and not Lipschitz continuous. In the case of the heat equation, if in (7.1)  $f(\tau, \xi, x) = -x^3$ , the nonlinear term in (7.4) is dissipative but not Lipschitz continuous. Moreover, in this case a Lipschitz continuous nonlinearity turns out to be well defined in  $L^2[(0, 1)]$ , while a nonlinear term with polynomial growth is not well defined in  $L^2([0, 1])$  but is well defined in  $C([0, 1])$ .

We consider the Hamilton–Jacobi–Bellman equation relative to (7.4),

$$(7.5) \quad \begin{cases} \frac{\partial v}{\partial t}(t, x) = -\mathcal{A}_t v(t, x) - \psi(t, x, \nabla v(t, x) G), & t \in [0, T], \ x \in H, \\ u(T, x) = \phi(x), \end{cases}$$

where  $\psi$  is the Hamiltonian function defined in (5.3). In order to solve the optimal control problem and to find mild solutions of the Hamilton–Jacobi–Bellman equation, we need  $\psi$  to satisfy Hypothesis 4.1, points 1, 2, and 3. We state direct hypotheses on  $l$  and  $k$  in the definition (7.2) of the concrete cost  $J$  when the measure  $\mu$  is a combination of Dirac measures, that is,

$$\mu = \sum_{i=1}^N \delta_{\xi_i}.$$

$\delta_x$  denotes the Dirac measure at  $x$ , and  $\xi_1, \dots, \xi_N$  belongs to the interval  $[0, 1]$ . In this case the Hamiltonian is given by

$$(7.6) \quad \psi(t, x, z) = \inf_{u \in \mathcal{U}} \left\{ \sum_{i=1}^N l(t, \xi_i, x(\xi_i), u) + \int_0^1 z(\xi) r(\xi) u(\xi) d\xi \right\}.$$

We make the following assumption.

*Hypothesis 7.3.* Let  $\bar{\psi} : [0, T] \times \mathbb{R}^N \times \Xi^* \rightarrow \mathbb{R}$  be given by

$$\bar{\psi}(t, y_1, \dots, y_N, z) = \inf_{u \in \mathcal{U}} \left\{ \sum_{i=1}^N l(t, \xi_i, y_i, u) + \int_0^1 z(\xi) r(\xi) u(\xi) d\xi \right\}.$$

Assume that for every  $t \in [0, T]$ ,  $\bar{\psi}(t, \cdot, \dots, \cdot) : \mathbb{R}^N \times \Xi^* \rightarrow \mathbb{R}$  is differentiable with bounded derivatives.

If Hypothesis 7.3 holds true, then the Hamiltonian defined in (7.6) satisfies Hypothesis 4.1, points 1, 2, and 3. A case when Hypothesis 7.3 can be checked directly on the cost functional is when

$$\sum_{i=1}^N l(t, \xi_i, y_i, u) = \sum_{i=1}^N l(t, \xi_i, y_i) + \int_0^1 \frac{u^2(\xi)}{2} d\xi,$$

where  $l : [0, T] \times [0, 1] \times \mathbb{R} \rightarrow \mathbb{R}$  is continuous and for every  $t \in [0, T]$  and  $\xi_i \in [0, 1]$ ,  $l(t, \xi_i, \cdot)$  is differentiable with bounded derivatives. It turns out that

$$\psi(t, x, z) = \sum_{i=1}^N l(t, \xi_i, x(\xi_i)) + \inf_{u \in \mathcal{U}} \int_0^1 \frac{u^2(\xi)}{2} d\xi + \int_0^1 z(\xi) r(\xi) u(\xi) d\xi.$$

Thus, with our assumptions,  $\psi(t, \cdot, \cdot) \in \mathcal{G}^{1,1}(E \times \Xi^*)$ . Indeed, differentiability with respect to  $x$  is straightforward, and differentiability with respect to  $z$  can be proved

directly by noting that

$$\inf_{u \in \mathcal{U}} \int_0^1 \frac{u^2(\xi)}{2} d\xi + \int_0^1 z(\xi) r(\xi) u(\xi) d\xi = \int_0^1 \gamma(z(\xi) r(\xi)) d\xi,$$

where  $\gamma(y) = \inf_{u \in \mathbb{R}, |u| \leq \delta} (\frac{u^2}{2} + yu)$ . By direct computations,

$$\gamma(y) = \begin{cases} -\frac{y^2}{2} & \text{if } |y| \leq \delta, \\ \frac{\delta^2}{2} - \delta|y| & \text{if } |y| > \delta. \end{cases}$$

**THEOREM 7.4.** *Assume that Hypothesis 7.1 holds true and that the Hamiltonian function satisfies Hypothesis 4.1, points 1, 2, and 3. Then (7.5) has a unique mild solution  $v$ , and for all a.c.s.'s  $(W, u, X^u)$ ,  $J(t, x, (W, u, X^u)) \geq v(t, x)$ . Moreover,  $J(t, x, (W, u, X^u)) = v(t, x)$  if and only if (5.12) holds. If (5.13) holds, there exists an optimal a.c.s.*

Similar arguments apply to stochastic optimal control problems and Hamilton–Jacobi–Bellman equations associated with heat equations with Neumann boundary conditions, that is, equations of the following form:

$$(7.7) \quad \begin{cases} d_\tau X^u(\tau, \xi) = \left[ \frac{\partial^2}{\partial \xi^2} X^u(\tau, \xi) + f(\tau, \xi, X^u(\tau, \xi)) + \sigma(\xi) r(\xi) u(\tau, \xi) \right] d\tau \\ \quad + \sigma(\xi) \dot{W}(\tau, \xi) d\tau, \\ \frac{\partial}{\partial \xi} X^u(\tau, 0) = \frac{\partial}{\partial \xi} X^u(\tau, 1) = 0, \\ X^u(t, \xi) = x(\xi). \end{cases}$$

We consider a cost functional of the form of (7.2), and we assume that Hypothesis 7.1 is satisfied. To rewrite the problem in an abstract way, we set  $H = L^2([0, 1])$  and  $E = C([0, 1])$ ,

$$\mathcal{D}(A) = \{y \in H^2([0, 1]) : y'(0) = y'(1) = 0\},$$

$$(Ay)(\xi) = \frac{\partial^2}{\partial \xi^2} y(\xi) \quad \text{for every } y \in \mathcal{D}(A),$$

and the restriction of  $A$  to  $E$  is given by

$$\mathcal{D}(A) = \{y \in C^2([0, 1]) : y'(0) = y'(1) = 0\},$$

$$(Ay)(\xi) = \frac{\partial^2}{\partial \xi^2} y(\xi) \quad \text{for every } y \in \mathcal{D}(A),$$

and in this case  $A$  is the generator of a strongly continuous semigroup. Then analogous results to the ones stated in Theorem 7.4 can be obtained.

**7.2. A controlled delay equation.** In  $(\Omega, \mathcal{F}, \mathbb{P})$  we consider the following stochastic delay equation with values in  $\mathbb{R}^n$ :

$$(7.8) \quad \begin{cases} dz^u(\tau) = \left[ \int_{-r}^0 a(d\theta) z^u(\tau + \theta) \right] d\tau + u(\tau) d\tau + dW(\tau), & \tau \in [t, T], \\ z^u(t) = h_0, \\ z^u(t + \theta) = h_1(\theta), & \theta \in [-r, 0], \end{cases}$$

where  $a(d\theta)$  is a matrix valued finite measure on  $[-r, 0]$  and  $W$  is a standard Wiener process in  $\mathbb{R}^n$ . The value  $r > 0$  denotes the maximum delay,  $h_1 \in L^p([-r, 0], \mathbb{R}^n)$ , and  $u_\tau$  is an admissible control, that is, an  $(\mathcal{F}_\tau)$ -predictable process taking values in a bounded subset of  $\mathbb{R}^n$ .

We want to rewrite (7.8) in an abstract way. Consider the nonstochastic case and take  $u$  identically zero; let  $z$  be the solution of the problem

$$\begin{cases} dz(\tau) = \left[ \int_{-r}^0 a(d\theta) z(\tau + \theta) \right] d\tau, & \tau \geq t, \\ z(t) = h_0, \\ z(t + \theta) = h_1(\theta), & \theta \in [-r, 0]. \end{cases}$$

Define the Banach space  $E = \mathbb{R}^n \oplus L^p([-r, 0], \mathbb{R}^n)$ . It turns out that

$$e^{\tau A} : E \longrightarrow E, \quad e^{\tau A} \begin{pmatrix} h_0 \\ h_1 \end{pmatrix} = \begin{pmatrix} z(\tau) \\ z_\tau \end{pmatrix}, \quad \tau \geq t \text{ with } z_\tau(\theta) = z(\theta + \tau)$$

defines a  $C_0$ -semigroup; see, e.g., [14] and [30]. The infinitesimal generator  $A$  of  $e^{\tau A}$ ,  $\tau \geq 0$ , is given by

$$\begin{aligned} \mathcal{D}(A) &= \left\{ \begin{pmatrix} h_0 \\ h_1 \end{pmatrix} \in E, h_1 \in W^{1,p}([-r, 0], \mathbb{R}^n), h_1(0) = h_0 \right\}, \\ Ah &= A \begin{pmatrix} h_0 \\ h_1 \end{pmatrix} = \begin{pmatrix} \int_{-r}^0 a(d\theta) h_1(\theta) \\ dh_1/d\theta \end{pmatrix}. \end{aligned}$$

If  $p > 2$ , the Banach space  $E = \mathbb{R}^n \oplus L^p([-r, 0], \mathbb{R}^n)$  is embedded in the Hilbert space  $H = \mathbb{R}^n \oplus L^2([-r, 0], \mathbb{R}^n)$ , where  $e^{\tau A}$  defines a  $C_0$ -semigroup. In  $H$  the infinitesimal generator  $A$  is given by

$$\begin{aligned} \mathcal{D}(A) &= \left\{ \begin{pmatrix} h_0 \\ h_1 \end{pmatrix} \in H, h_1 \in H^1([-r, 0], \mathbb{R}^n), h_1(0) = h_0 \right\}, \\ Ah &= \begin{pmatrix} \int_{-r}^0 a(d\theta) h_1(\theta) \\ dh_1/d\theta \end{pmatrix}. \end{aligned}$$

By setting

$$G : \mathbb{R}^n \longrightarrow E, \quad G = \begin{pmatrix} I \\ 0 \end{pmatrix}, \quad X_\tau^u = \begin{pmatrix} z^u(\tau) \\ z_\tau^u \end{pmatrix}, \quad \text{and} \quad X_t^u = \begin{pmatrix} h_0 \\ h_1 \end{pmatrix},$$

problem (7.8) can be rewritten in an abstract way as a controlled stochastic evolution equation in  $E$ :

$$(7.9) \quad \begin{cases} dX_\tau^u = AX_\tau^u d\tau + Gu_\tau d\tau + GdW_\tau, & \tau \in [t, T], \\ X_t^u = x. \end{cases}$$

The techniques developed in this paper allow us to treat delay equations that evolve in spaces of  $p$ -integrable functions, but not in spaces of continuous functions because in such spaces the operator  $A$  is not, in general, the generator of a strongly continuous semigroup. Thus the treatment of stochastic optimal control problems in such spaces needs substantial technical modification of our result and will be the object of future work. Also when treating delay equations in spaces of  $p$ -integrable functions, which in (7.9) are formulated as evolution equations in  $E = \mathbb{R}^n \oplus L^p([-r, 0], \mathbb{R}^n)$ , we cannot directly apply the results of the previous sections since  $A$ , in general, is not dissipative in  $E$ . It is possible to find  $\gamma \in \mathbb{R}$  such that  $A - \gamma I$  is dissipative if we consider, instead of  $L^p([-r, 0], \mathbb{R}^n)$ , a weighted space, namely,  $L^p([-r, 0], \mathbb{R}^n; \mu)$ ; see [30] for more details. Nevertheless, since in (7.9)  $F = 0$ , in order to apply the results of the previous sections it suffices to prove that  $A$  generates a strongly continuous semigroup  $e^{\tau A}$  in  $E$  which extends to a strongly continuous semigroup in  $H$ . We briefly show that under these assumptions on  $A$ , and if  $G$  satisfies Hypothesis 3.2, the results in section 3 are true for the solution of a linear equation in  $E$ :

$$\begin{cases} dX_\tau = AX_\tau d\tau + GdW_\tau, & \tau \in [t, T], \\ X_t = x. \end{cases}$$

The mild solution of this equation is given by

$$X_\tau = X(\tau, t, x) = e^{(\tau-t)A}x + \int_t^\tau e^{(\tau-s)A}GdW_s,$$

so it is immediate to see that the solution is Lipschitz continuous with respect to the initial datum and is also pathwise differentiable—results that are proved in Proposition 3.7, and in Proposition 3.10 for the more general case of  $F \neq 0$ . Moreover, since we assume that the stochastic convolution admits an  $E$ -continuous version, the process  $(X_\tau)_{\tau \in [0, T]}$  belongs to  $\mathcal{H}^p([0, T], E)$  for every  $1 \leq p < \infty$ , and also the map  $x \mapsto (X(\tau, t, x))_{\tau \in [0, T]}$  belongs to  $\mathcal{G}^1(E, \mathcal{H}^p([0, T], E))$ . These results are the counterpart of Lemma 3.12 and Proposition 3.13.

In the case of the delay semigroup, by simple direct calculations it turns out there exist  $j \geq 0$  and  $C > 0$  such that  $\|e^{\tau A}\|_E \leq C\tau^j$ . In the following lemma we prove that the stochastic convolution admits a continuous version with values in  $E = \mathbb{R}^n \oplus L^p([-r, 0], \mathbb{R}^n)$ .

LEMMA 7.5. *The stochastic convolution  $W_A(\tau) = \int_0^\tau e^{(\tau-s)A}GdW_s$  admits an  $E$ -continuous version.*

*Proof.* By the factorization method (see, e.g., [12, p. 128]), for some  $0 < \alpha < 1/2$ , we can write

$$\begin{aligned} W_A(\tau) &= \int_0^\tau e^{(\tau-s)A}BdW_s = \frac{\sin \pi \alpha}{\pi} \int_0^\tau \left[ \int_s^\tau (\tau - \sigma)^{\alpha-1} (\sigma - s)^{-\alpha} d\sigma \right] e^{(\tau-s)A}GdW_s \\ &= \frac{\sin \pi \alpha}{\pi} \int_0^\tau e^{(\tau-\sigma)A} (\tau - \sigma)^{\alpha-1} Y_\sigma d\sigma, \end{aligned}$$

where we have set  $Y_\sigma = \int_0^\sigma (\sigma - s)^{-\alpha} e^{(\sigma-s)A}GdW_s$ . If we show that for some  $q > 2$ ,  $Y$  is in  $L^q([0, T], E)$ ,  $\mathbb{P}$ -almost everywhere, then the stochastic convolution admits a continuous version in  $E$ . We are going to show the stronger condition

$$\mathbb{E} \int_0^T \|Y_\sigma\|_E^q d\sigma < \infty.$$

The process  $Y$  admits the representation

$$Y_\sigma = \begin{pmatrix} Y_\sigma^1 \\ Y_\sigma^2 \end{pmatrix} = \begin{pmatrix} \int_0^\sigma (\sigma-s)^{-\alpha} \Pi_1 e^{(\sigma-s)A} B dW_s \\ \int_0^\sigma (\sigma-s)^{-\alpha} \Pi_2 e^{(\sigma-s)A} B dW_s \end{pmatrix}$$

and  $Y_\sigma^2(\theta) = Y_{\sigma+\theta}^1 \chi_{(-\sigma,0)}(\theta)$ , where by  $\chi_A$  we mean the indicator function of the set  $A$ . We have recalled that there exist  $j \geq 0$  and  $C > 0$  such that  $\|e^{tA}\|_E \leq C\tau^j$ , so for every  $a \in \mathbb{R}^n$ ,  $\|e^{tA}Ga\|_E \leq C\tau^j|a|_{\mathbb{R}^n}$ . Let  $\Pi_1 : E \rightarrow \mathbb{R}^n$  be the projection on the first component of  $E$ , and  $\Pi_2 : E \rightarrow L^p([-r,0], \mathbb{R}^n)$  be the projection on the second component of  $E$ , that is, for every  $\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} \in E$ ,  $\Pi_1\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = k_1$ , and  $\Pi_2\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = k_2$ . We take  $q = p$ , and so  $q > 2$ , and we estimate

$$\begin{aligned} \mathbb{E} \int_0^T |Y_\sigma^1|_{\mathbb{R}^n}^p d\sigma &= \int_0^T \mathbb{E} \left| \int_0^\sigma (\sigma-s)^{-\alpha} \Pi_1 e^{(\sigma-s)A} dW_s \right|_{\mathbb{R}^n}^p d\sigma \\ &\leq c_p \int_0^T \mathbb{E} \left( \int_0^\sigma (\sigma-s)^{-2\alpha} \left| \Pi_1 e^{(\sigma-s)A} G \right|_{\mathbb{R}^n}^2 ds \right)^{p/2} d\sigma \\ &\leq c \int_0^T \left( \int_0^\sigma (\sigma-s)^{-2\alpha} (\sigma-s)^{2j} ds \right)^{p/2} d\sigma < +\infty, \end{aligned}$$

where  $c$  is a positive constant. So we also get that for every  $\theta \in [-r,0]$ ,

$$\begin{aligned} \mathbb{E} \int_0^T |Y_\sigma^2(\theta)|_{\mathbb{R}^n}^p d\sigma &= \int_0^T \mathbb{E} \left| \int_0^\sigma (\sigma-s)^{-\alpha} \Pi_2 e^{(\sigma-s)A} G dW_s(\theta) \right|_{\mathbb{R}^n}^p d\sigma \\ &= \int_0^T \mathbb{E} \left| \int_0^\sigma (\sigma-s)^{-\alpha} \chi_{(-\sigma+s,0)}(\theta) \Pi_1 e^{(\sigma+\theta-s)A} G dW_s \right|_{\mathbb{R}^n}^p d\sigma \\ &= \int_{-\theta}^T \mathbb{E} \left| \int_0^{\sigma+\theta} (\sigma-s)^{-\alpha} \Pi_1 e^{(\sigma+\theta-s)A} B dW_s \right|_{\mathbb{R}^n}^p d\sigma \\ &\leq c \int_{-\theta}^T \mathbb{E} \left( \int_0^{\sigma+\theta} (\sigma-s)^{-2\alpha} \left| \Pi_1 e^{(\sigma+\theta-s)A} G \right|_{\mathbb{R}^n}^2 ds \right)^{p/2} d\sigma \\ &\leq cT^{pj} \int_{-\theta}^T \left( \int_0^{\sigma+\theta} (\sigma-s)^{-2\alpha} ds \right)^{p/2} d\sigma \\ &\leq cg(\theta). \end{aligned}$$

$c$  is a positive constant depending on  $p$ ,  $C$ , and  $T$ , and  $g : [-r,0] \rightarrow \mathbb{R}^n$  is bounded. We get

$$\mathbb{E} \int_0^T \|Y_\sigma^2\|_{L^p([-r,0], \mathbb{R}^n)}^p d\sigma \leq c \left( \int_{-r}^0 g^p(\theta) d\theta \right)^{1/p} < +\infty,$$

and the proof is concluded.  $\square$

For the delay equation, Hypothesis 3.16 is trivial, since  $\Xi = \mathbb{R}^n$  and  $G$  is the inclusion of  $\mathbb{R}^n$  in the product space  $E = \mathbb{R}^n \times L^p([-r, 0], \mathbb{R}^n)$ .

In relation to the controlled equation (7.8) we want to study the following stochastic control problem, which we are going to state in its weak formulation. We define the cost functional

$$(7.10) \quad J(t, h_0, h_1, u) = \mathbb{E} \int_{-r}^0 z^u(T + \theta) y(\theta) d\theta,$$

where  $y \in L^q([-r, 0], \mathbb{R}^n)$ , with  $q$  the conjugate exponent of  $p$ . The main advantage in treating delay equations in the space of  $p$ -integrable functions, with  $p > 2$ , is that we can treat costs as in (7.10), with  $y$  not necessarily square integrable; indeed, in order to have the cost be well defined, we have only to ask that  $y \in L^q([-r, 0], \mathbb{R}^n)$ , and  $q < 2$  since it is the conjugate exponent of  $p$ .

We have to minimize the cost  $J$  over all admissible controls. An admissible control  $u_\tau$  is a predictable process

$$(\Omega, \mathcal{F}, (\mathcal{F}_\tau)_{\tau \geq 0}, \mathbb{P}) \rightarrow \mathcal{U},$$

where  $\mathcal{U}$  is a bounded subset in  $\mathbb{R}^n$ .

In the abstract setting, the cost functional can be rewritten as

$$J(t, h, u) = \mathbb{E} \phi(X_T^u),$$

where, for  $j \in E$ ,

$$j = \begin{pmatrix} j_1 \\ j_2 \end{pmatrix}, \quad \phi(j) = \int_{-r}^0 j_2(\theta) y(\theta) d\theta, \quad \text{and } \Pi_2 : E \longrightarrow L^p([-r, 0], \mathbb{R}^n)$$

is the projection on the second component of  $E$ .

We can also prove existence and uniqueness of a mild solution of the Hamilton–Jacobi–Bellman equation, associated with a delay equation,

$$(7.11) \quad \begin{cases} \frac{\partial v}{\partial t}(t, x) = -\mathcal{A}_t v(t, x) - \psi(t, x, \nabla v(t, x) G), & t \in [0, T], \quad x \in H, \\ u(T, x) = \phi(x). \end{cases}$$

We summarize the results in the following theorem.

**THEOREM 7.6.** *Equation (7.11) has a unique mild solution  $v$ , and if the cost  $J$  is defined as in (7.10), then for all a.c.s.'s  $(W, u, X^u)$  we have  $J(t, x, (W, u, X^u)) \geq v(t, x)$ , and the equality holds if and only if (5.12) holds. If (5.13) holds, there exists an optimal a.c.s.*

**7.3. A controlled wave equation.** In this section we show how our results can be applied to perform the synthesis of the optimal control when the state equation is a controlled wave equation. Our goal is to solve an optimal control problem for a stochastic wave equation where the cost functional is well defined on continuous functions, so the cost is not necessary well defined on the space of square integrable functions: for such costs, existing results in the literature concerning stochastic optimal control problems in Hilbert spaces do not apply. The Banach space  $E$ , where the wave equation evolves, will be the product of two Besov spaces. The choice of this space will be made clear in the following.

This section is organized as follows. We introduce the wave operator  $A$  with its domain in a Banach space  $E$ , which turns out to be the product of two Besov spaces, and we observe that  $A$  is dissipative in  $E$ . For the stochastic case, we prove that the stochastic convolution admits a continuous version in  $E$ . Finally, we introduce the stochastic controlled wave equation and, by applying the results of the previous sections, we solve a stochastic optimal control problem and the related Hamilton–Jacobi–Bellman equation.

Let us consider a wave equation in one space dimension:

$$(7.12) \quad \begin{cases} \frac{\partial^2}{\partial \tau^2} y(\tau, \xi) = \frac{\partial^2}{\partial \xi^2} y(\tau, \xi), \\ y(\tau, 0) = y(\tau, 1) = 0, \\ y(0, \xi) = x_0(\xi), \\ \frac{\partial y}{\partial \tau}(0, \xi) = x_1(\xi), \end{cases}$$

where  $\tau \in [0, T]$  and  $\xi \in [0, 1]$ . We want to write (7.12) in abstract form. We follow a standard procedure; see, e.g., [12]. We define  $\Lambda$  by

$$\mathcal{D}(\Lambda) = H^2([0, 1]) \cap H_0^1([0, 1]), \quad (\Lambda y)(\xi) = -\frac{\partial^2}{\partial \xi^2} y(\xi)$$

for every  $y \in \mathcal{D}(\Lambda)$ . We introduce the Hilbert space

$$H_1 = H_0^1([0, 1]) \oplus L^2([0, 1]).$$

On  $H_1$  we define the operator  $A$  by

$$\mathcal{D}(A) = H^2([0, 1]) \cap H_0^1([0, 1]) \oplus H_0^1([0, 1]), \quad A \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} 0 & I \\ -\Lambda & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix}$$

for every  $\begin{pmatrix} y \\ z \end{pmatrix} \in \mathcal{D}(A)$ . Next we introduce the Hilbert space

$$H_2 = L^2([0, 1]) \oplus H^{-1}([0, 1]).$$

On  $H_2$  we define the operator  $A$  by

$$\mathcal{D}(A) = H_0^1([0, 1]) \oplus L^2([0, 1]), \quad A \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} 0 & I \\ -\Lambda & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix}$$

for every  $\begin{pmatrix} y \\ z \end{pmatrix} \in \mathcal{D}(A)$ . In both  $H_1$  and  $H_2$ ,  $A$ , with the suitable domain, is the generator of the contractive group

$$e^{tA} \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} \cos \sqrt{\Lambda} t & \frac{1}{\sqrt{\Lambda}} \sin \sqrt{\Lambda} t \\ -\sqrt{\Lambda} \sin \sqrt{\Lambda} t & \cos \sqrt{\Lambda} t \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix}, \quad t \in \mathbb{R}.$$

Equation (7.12) can be rewritten in an abstract way as

$$(7.13) \quad \begin{cases} dX_\tau = AX_\tau d\tau, & \tau \in [0, T], \\ X_0 = x. \end{cases}$$



Next, we consider a stochastic wave equation in  $(\Omega, \mathcal{F}, (\mathcal{F}_\tau)_{\tau \geq 0}, \mathbb{P})$ . We consider, for  $0 \leq t \leq \tau \leq T$  and  $\xi \in [0, 1]$ , the following state equation:

$$(7.14) \quad \begin{cases} \frac{\partial^2}{\partial \tau^2} y(\tau, \xi) = \frac{\partial^2}{\partial \xi^2} y(\tau, \xi) + \dot{W}(\tau, \xi), \\ y(\tau, 0) = y(\tau, 1) = 0, \\ y(0, \xi) = x_0(\xi), \\ \frac{\partial y}{\partial \tau}(0, \xi) = x_1(\xi). \end{cases}$$

$\dot{W}(\tau, \xi)$  is a space-time white noise on  $[0, T] \times [0, 1]$ . This equation can be rewritten in an abstract way in the Hilbert space  $H_2$  in the following form:

$$(7.15) \quad \begin{cases} dX_\tau = AX_\tau d\tau + GdW_\tau, & \tau \in [0, T], \\ X_0 = x, \end{cases}$$

where  $\{W_\tau, \tau \geq 0\}$  is a cylindrical Wiener process in  $L^2([0, 1])$  with respect to the filtration  $(\mathcal{F}_\tau)_{\tau \geq 0}$ . The operator  $G : L^2([0, 1]) \longrightarrow H_2$  is defined by  $Gu = \begin{pmatrix} 0 \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ I \end{pmatrix} u$ , where  $I$  is the embedding of  $L^2([0, 1])$  in  $H^{-1}([0, 1])$ . The solution of (7.15) belongs to  $H_2 = L^2([0, 1]) \oplus H^{-1}([0, 1])$ , since in this space the operators

$$\int_0^t e^{sA} G G^* e^{sA^*} ds$$

are of trace class; see [12, Example 5.8]. So the process

$$W_A(t) = \int_0^t e^{(t-s)A} G dW(s)$$

is well defined in  $H_2$ .

We introduce the Banach space  $E = B_{2,p,\{0\}}^s([0, 1]) \oplus B_{2,p}^{s-1}([0, 1])$ , with  $s \in (0, 1)$  and  $p > 2$ , where  $B_{2,p,\{0\}}^s([0, 1])$  is the Besov space with Dirichlet boundary conditions and  $B_{2,p}^{s-1}([0, 1])$  is a Besov space with negative exponent. The space  $E$  can be obtained by interpolating  $H_1$  and  $H_2$ :

$$\begin{aligned} (H_2, H_1)_{s,p} &= (L^2([0, 1]), H_0^1([0, 1]))_{s,p} \oplus (H^{-1}([0, 1]), L^2([0, 1]))_{s,p} \\ &= B_{2,p,\{0\}}^s([0, 1]) \oplus B_{2,p}^{s-1}([0, 1]). \end{aligned}$$

For more details on real interpolation and Besov spaces, see, e.g., [21, Chapter I] and [28]. The operator  $A$  with domain  $\mathcal{D}(A) = B_{2,p,\{0\}}^{s+1}([0, 1]) \oplus B_{2,p,\{0\}}^s([0, 1])$  is the generator of a group in  $E$ , and, moreover, since  $A$  is dissipative in both  $H_1$  and  $H_2$ , it turns out that  $A$  is dissipative in  $E$ . Equation (7.14) can be written as an evolution equation in the Banach space  $E$ :

$$(7.16) \quad \begin{cases} dX_\tau = AX_\tau d\tau + GdW_\tau, & \tau \in [t, T], \\ X_0 = x, \end{cases}$$

where  $G : L^2([0, 1]) \longrightarrow E$  is defined by  $Gu = \begin{pmatrix} 0 \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ I \end{pmatrix} u$ , and  $I$  is the embedding of  $L^2([0, 1])$  in  $B_{2,p}^{s-1}([0, 1])$ .

*Remark 7.7.* The idea of the wave equation in a Besov space is inspired by the fact that if  $s - 1/2 > 0$ , then  $B_{2,p,\{0\}}^s([0, 1])$  is contained in  $C([0, 1])$ ; see [28,

Theorem 4.6.1]. Moreover, since  $E$  can be obtained by interpolating  $H_1$  and  $H_2$ , dissipativity of the operator is straightforward. In this paper, we take  $s = 1/2 + \beta$ , for some  $0 < \beta < 1/2$ , to be chosen in the proof of the following lemma.

In order to apply our results we note that the Banach space  $E$  is continuously and densely embedded in the Hilbert space  $H_2$ , where the stochastic convolution  $W_A(t)$  takes values. We need to prove that  $W_A(t)$  admits a version in  $C([0, T], E)$ .

LEMMA 7.8. *The stochastic convolution  $W_A(t) = \int_0^t e^{(t-s)A} G dW(s)$  admits a version in  $C([0, T], E)$  for every  $p \geq 6$  and some  $s \in (\frac{1}{2}, 1)$ , i.e.,  $s = \frac{1}{2} + \beta$ , with  $\beta > 0$ .*

*Proof.* Let

$$W_A(t) = \begin{pmatrix} W_A^1(t) \\ W_A^2(t) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{\Lambda}} \int_0^t \sin(\sqrt{\Lambda}(t-s)) dW_s \\ \int_0^t \cos(\sqrt{\Lambda}(t-s)) dW_s \end{pmatrix}.$$

We start by proving that  $W_A^1(t)$  admits a version in  $C([0, T], B_{2,p,\{0\}}^s([0, 1]))$ . Indeed, we prove a stronger result:  $W_A^1(t)$  admits a version in  $C([0, T], B_{p,p,\{0\}}^s([0, 1]))$ . Since  $p > 2$ ,  $B_{p,p,\{0\}}^s([0, 1]) \subset B_{2,p,\{0\}}^s([0, 1])$ . We recall that  $B_{p,p,\{0\}}^s([0, 1])$  is a Sobolev space with fractional exponent and it is usually denoted by  $W_{p,\{0\}}^s([0, 1])$ . With the norm

$$\|f\|_{W_{p,\{0\}}^s}^p = \int_0^1 |f(\xi)|^p d\xi + \int_0^1 \int_0^1 \frac{|f(\xi) - f(\eta)|^p}{|\xi - \eta|^{1+sp}} d\xi d\eta,$$

$W_{p,\{0\}}^s([0, 1])$  is a Banach space. Let  $\beta_k, k \geq 1$ , be standard independent real Wiener processes. Let us first prove that

$$W_A^1(t)(\xi) = \frac{1}{\sqrt{\Lambda}} \int_0^t \sin(\sqrt{\Lambda}(t-s)) dW_s(\xi) = \int_0^t \sum_{k \geq 1} \frac{\sin k\pi(t-s)}{k\pi} \sin k\pi\xi d\beta_k(s)$$

takes values in  $W_{p,\{0\}}^s([0, 1])$ . By its definition,  $W_A^1(t)$  vanishes at the boundary. We prove that, for every  $t \in [0, T]$ ,  $\|W_A^1(t)\|_{W_{p,\{0\}}^s}^p$  is finite a.s.

$$\begin{aligned} & \|W_A^1(t)\|_{W_{p,\{0\}}^s}^p \\ &= \int_0^1 \left| \int_0^t \sum_{k \geq 1} \frac{\sin k\pi(t-s)}{k\pi} \sin k\pi\xi d\beta_k(s) \right|^p d\xi + \int_0^1 \int_0^1 \\ & \quad \times \frac{\left| \int_0^t \sum_{k \geq 1} \frac{\sin k\pi(t-s)}{k\pi} \sin k\pi\xi d\beta_k(s) - \int_0^t \sum_{k \geq 1} \frac{\sin k\pi(t-s)}{k\pi} \sin k\pi\eta d\beta_k(s) \right|^p}{|\xi - \eta|^{1+sp}} d\xi d\eta \\ &= I + II. \end{aligned}$$

In order to prove that  $\|W_A^1(t)\|_{W_{p,\{0\}}^s}^p$  is finite a.s., we prove that  $\mathbb{E}\|W_A^1(t)\|_{W_{p,\{0\}}^s}^p < \infty$ .

We evaluate

$$\begin{aligned}
\mathbb{E}(I) &= \mathbb{E} \int_0^1 \left| \int_0^t \sum_{k \geq 1} \frac{\sin k\pi(t-s)}{k\pi} \sin k\pi \xi d\beta_k(s) \right|^p d\xi \\
&= \int_0^1 \mathbb{E} \left| \int_0^t \sum_{k \geq 1} \frac{\sin k\pi(t-s)}{k\pi} \sin k\pi \xi d\beta_k(s) \right|^p d\xi \\
&\leq c \int_0^1 \left( \mathbb{E} \left| \int_0^t \sum_{k \geq 1} \frac{\sin k\pi(t-s)}{k\pi} \sin k\pi \xi d\beta_k(s) \right|^2 \right)^{p/2} d\xi \\
&= c \int_0^1 \left( \int_0^t \sum_{k \geq 1} \frac{\sin^2 k\pi(t-s)}{k^2 \pi^2} \sin^2 k\pi \xi ds \right)^{p/2} d\xi \\
&\leq c \left( \sum_{k \geq 1} \frac{1}{k^2 \pi^2} \right)^{p/2} \left( \int_0^t \sin^2 k\pi(t-s) ds \right)^{p/2} \\
&\leq c \left( \sum_{k \geq 1} \frac{1}{k^2 \pi^2} \right)^{p/2} t^{p/2},
\end{aligned}$$

and so it is finite for every  $p$ . Then we evaluate  $\mathbb{E}(II)$ . In the following calculations we consider only integer even  $p \geq 6$ ; this suffices to prove the result for every  $p \geq 6$ :

$$\begin{aligned}
\mathbb{E}(II) &= \mathbb{E} \int_0^1 \int_0^1 \frac{1}{|\xi - \eta|^{1+sp}} \\
&\quad \times \left( \int_0^t \sum_{k \geq 1} \frac{\sin k\pi(t-s)}{k\pi} (\sin k\pi \xi - \sin k\pi \eta) d\beta_k(s) \right)^p d\xi d\eta \\
&= \int_0^1 \int_0^1 \frac{1}{|\xi - \eta|^{1+sp}} \mathbb{E} \sum_{k \geq 1} \left( \int_0^t \frac{\sin k\pi(t-s)}{k\pi} (\sin k\pi \xi - \sin k\pi \eta) d\beta_k(s) \right)^p d\xi d\eta \\
&\quad + \int_0^1 \int_0^1 \frac{1}{|\xi - \eta|^{1+sp}} \mathbb{E} \sum_{n=2}^{p-2} c_n \\
&\quad \times \left\{ \sum_{\substack{k, j \geq 1 \\ k \neq j}} \left( \int_0^t \frac{\sin k\pi(t-s)}{k\pi} |\sin k\pi \xi - \sin k\pi \eta| d\beta_k(s) \right)^n \right. \\
&\quad \times \left. \left( \int_0^t \frac{\sin j\pi(t-s)}{j\pi} |\sin j\pi \xi - \sin j\pi \eta| d\beta_j(s) \right)^{p-n} \right\} d\xi d\eta.
\end{aligned}$$

In the last passage the terms with  $n = 1$  or  $n = p - 1$  do not appear since they have null expectation, and  $c_n$  denotes a constant depending only on  $n$ . By applying the Burkholder–Davis–Gundy inequality, we get

$$\begin{aligned}
\mathbb{E}(II) &\leq c_p \int_0^1 \int_0^1 \frac{1}{|\xi - \eta|^{1+sp}} \sum_{k \geq 1} \\
&\quad \left( \int_0^t \frac{\sin^2 k\pi(t-s)}{k^2\pi^2} (\sin k\pi\xi ds - \sin k\pi\eta)^2 ds \right)^{p/2} d\xi d\eta \\
&\quad + c_p \int_0^1 \int_0^1 \frac{1}{|\xi - \eta|^{1+sp}} \sum_{n=2}^{p-2} c_n \\
&\quad \left\{ \sum_{\substack{k,j \geq 1 \\ k \neq j}} \left( \int_0^t \frac{\sin^2 k\pi(t-s)}{k^2\pi^2} (\sin k\pi\xi - \sin k\pi\eta)^2 ds \right)^{n/2} \right. \\
&\quad \left. \left( \int_0^t \frac{\sin^2 j\pi(t-s)}{j^2\pi^2} (\sin j\pi\xi - \sin j\pi\eta)^2 ds \right)^{(p-n)/2} \right\} d\xi d\eta \\
&\leq c_p t^{p/2} \sum_{k \geq 1} \frac{1}{k^p \pi^p} \int_0^1 \int_0^1 \frac{|\sin k\pi\xi - \sin k\pi\eta|^p}{|\xi - \eta|^{1+sp}} d\xi d\eta \\
&\quad + 2c_p t^{p/2} \sum_{n=2}^{p/2} c_n \sum_{\substack{k,j \geq 1 \\ k \neq j}} \frac{1}{k^n \pi^n} \frac{1}{j^{p-n} \pi^{p-n}} \int_0^1 \int_0^1 \\
&\quad \frac{|\sin k\pi\xi - \sin k\pi\eta|^n |\sin j\pi\xi - \sin j\pi\eta|^{p-n}}{|\xi - \eta|^{1+sp}} d\xi d\eta \\
&\leq c_p t^{p/2} \left\{ 2^{p(1-s-\epsilon)} \sum_{k \geq 1} \frac{1}{k^p \pi^p} \int_0^1 \int_0^1 \frac{|\sin k\pi\xi - \sin k\pi\eta|^{p(s+\epsilon)}}{|\xi - \eta|^{1+sp}} d\xi d\eta \right. \\
&\quad + \frac{2^{1+p(1-s-3\beta)}}{\pi^p} * \sum_{n=2}^{p/2} 2^{5n\beta} c_n \sum_{\substack{k,j \geq 1 \\ k \neq j}} \frac{1}{k^n} \frac{1}{j^{p-n}} \int_0^1 \int_0^1 \\
&\quad \left. \frac{|\sin k\pi\xi - \sin k\pi\eta|^{n(s-2\beta)} |\sin j\pi\xi - \sin j\pi\eta|^{(p-n)(s+3\beta)}}{|\xi - \eta|^{1+sp}} d\xi d\eta \right\}
\end{aligned}$$

$$\begin{aligned}
&\leq c_p T^{p/2} \left\{ 2^{p(1-s-\epsilon)} \sum_{k \geq 1} \frac{k^{p(s+\epsilon)} \pi^{p(s+\epsilon)}}{k^p \pi^p} \int_0^1 \int_0^1 \frac{1}{|\xi - \eta|^{1-p\epsilon}} d\xi d\eta \right. \\
&\quad + \frac{2^{1+p(1-s-3\beta)}}{\pi^p} \sum_{n=2}^{p/2} c_n \sum_{\substack{k, j \geq 1 \\ k \neq j}} \frac{k^{n(s-2\beta)} \pi^{n(s-2\beta)}}{k^n} \\
&\quad \left. \frac{k^{(p-n)(s+3\beta)} \pi^{(p-n)(s+3\beta)}}{j^{p-n}} \int_0^1 \int_0^1 \frac{1}{|\xi - \eta|^{1-3\beta p+5\beta n}} d\xi d\eta \right\} \\
&= c_p T^{p/2} \left\{ \frac{2^{p(1-s-\epsilon)}}{\pi^{p(1-s-\epsilon)}} \sum_{k \geq 1} \frac{1}{k^{p(1-s-\epsilon)}} \int_0^1 \int_0^1 \frac{1}{|\xi - \eta|^{1-p\epsilon}} d\xi d\eta \right. \\
&\quad + 2^{1+p(1-s-3\beta)} \sum_{n=2}^{p/2} \frac{c_n}{\pi^{p-5n\beta-ps-3p\beta}} \sum_{\substack{k, j \geq 1 \\ k \neq j}} \frac{1}{k^{n(1-s+2\beta)}} \\
&\quad \left. \frac{1}{j^{(p-n)(1-s-3\beta)}} \int_0^1 \int_0^1 \frac{1}{|\xi - \eta|^{1-3\beta p+5\beta n}} d\xi d\eta \right\}.
\end{aligned}$$

Take  $\epsilon > 0$  such that  $\epsilon < 1 - s$  and  $p > \frac{1}{1-s-\epsilon}$ . Next, remember that we have taken  $s = 1/2 + \beta$ ,  $\beta > 0$ . So for every  $n = 2, \dots, p/2$ ,  $n(1-s+2\beta) > 1$ . Now choose  $\beta$  such that for every  $n = 2, \dots, p/2$ ,  $(p-n)(1-s-3\beta) > 1$ . It is enough to find  $\beta$  such that  $p/2(1-s-3\beta) > 1$ . This leads to taking  $0 < \beta < 1/8 - 1/2p$ , that is,  $1/2 < s < 5/8 - 1/2p$ . With these choices, and with  $C_p$  a constant depending only on  $p$  and  $s$ , we get

$$\mathbb{E}(II) \leq C_p T^{p/2} \left( \sum_{k \geq 1} \frac{1}{k^{p(1-s-\epsilon)}} + \sum_{n=2}^{p/2} c_n \sum_{\substack{k, j \geq 1 \\ k \neq j}} \frac{1}{k^{n(1-s+2\beta)}} \frac{1}{j^{(p-n)(1-s-3\beta)}} \right) < \infty.$$

Thus we conclude that, a.s.,  $W_A^1(t) \in W_{p, \{0\}}^s([0, 1]) \subset B_{2,p, \{0\}}^s([0, 1])$ . Moreover, we claim that, a.s.,  $W_A^2(t) \in B_{2,p}^{s-1}([0, 1])$ . Indeed,

$$W_A^2(t) = \int_0^t \cos(\sqrt{\Lambda}(t-s)) dW_s = \sqrt{\Lambda} \left( \frac{1}{\sqrt{\Lambda}} \int_0^t \cos(\sqrt{\Lambda}(t-s)) dW_s \right).$$

By analogous calculations we have performed in order to prove that  $W_A^1(t) \in B_{2,p, \{0\}}^s([0, 1])$ , we can prove that

$$\left( \frac{1}{\sqrt{\Lambda}} \int_0^t \cos(\sqrt{\Lambda}(t-s)) dW(s) \right) \in B_{2,p}^s([0, 1]);$$

the difference with respect to  $W_A^1(t)$  is that there is  $\cos(\sqrt{\Lambda}(t-s))$  instead of  $\sin(\sqrt{\Lambda}(t-s))$ , so this process does not vanish on the boundary; on the contrary, the calculations are analogous to the ones proving that  $W_A^1(t) \in B_{2,p,\{0\}}^s([0,1])$ . Thus  $W_A^2(t) \in B_{2,p}^{s-1}([0,1])$ .

Next we want to show the existence of a version of the stochastic convolution in  $C([0,T], E)$ . We note that by the factorization method (see, e.g., [12]), for  $0 < \alpha < \frac{1}{2}$  we can write

$$\begin{aligned} W_A(t) &= \int_0^t e^{(t-s)A} G dW_s = \frac{\sin \pi \alpha}{\pi} \int_0^t \left[ \int_s^t (t-\sigma)^{\alpha-1} (\sigma-s)^{-\alpha} d\sigma \right] e^{(t-s)A} G dW_s \\ &= \frac{\sin \pi \alpha}{\pi} \int_0^t e^{(t-\sigma)A} (t-\sigma)^{\alpha-1} Y_\sigma d\sigma, \end{aligned}$$

where we have set  $Y_\sigma = \int_0^\sigma (\sigma-s)^{-\alpha} e^{(\sigma-s)A} G dW_s$ . If we show that for some  $q > 2$ ,  $Y$  is in  $L^q([0,T], E)$ ,  $\mathbb{P}$ -almost everywhere, then the stochastic convolution admits a continuous version in  $E$ . We take  $q = p$ , so that  $q \geq 6$ , and we prove the stronger condition

$$\mathbb{E} \int_0^T \|Y_\sigma\|_E^p d\sigma < \infty.$$

We write the two components of  $Y_\sigma$ ,

$$Y_\sigma = \begin{pmatrix} Y_\sigma^1 \\ Y_\sigma^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{\Lambda}} \int_0^\sigma (\sigma-s)^{-\alpha} \sin(\sqrt{\Lambda}(\sigma-s)) dW_s \\ \int_0^\sigma (\sigma-s)^{-\alpha} \cos(\sqrt{\Lambda}(\sigma-s)) dW_s \end{pmatrix}.$$

We evaluate, at first,

$$\begin{aligned} &\mathbb{E} \int_0^T \|Y_\sigma^1\|_{B_{2,p,\{0\}}^s}^p d\sigma \\ &\leq c \mathbb{E} \int_0^T \|Y_\sigma^1\|_{W_{p,\{0\}}^s}^p d\sigma \\ &= c \mathbb{E} \int_0^T \int_0^1 \left| \int_0^\sigma (\sigma-s)^{-\alpha} \sum_{k \geq 1} \frac{\sin k\pi(\sigma-s)}{k\pi} \sin k\pi\xi d\beta_k(s) \right|^p d\xi d\sigma \\ &\quad + c \mathbb{E} \int_0^T \int_0^1 \int_0^1 \frac{\left| \int_0^\sigma (\sigma-s)^{-\alpha} \sum_{k \geq 1} \frac{\sin k\pi(\sigma-s)}{k\pi} (\sin k\pi\xi - \sin k\pi\eta) d\beta_k(s) \right|^p}{|\xi - \eta|^{1+sp}} d\xi d\eta d\sigma. \end{aligned}$$

By an analogous procedure we have used to evaluate the norm of  $W_A^1(t)$ , we can get that

$$\mathbb{E} \int_0^T \|Y_\sigma^1\|_{B_{2,p,\{0\}}^s}^p d\sigma < \infty.$$

In order to evaluate  $\mathbb{E} \int_0^T \|Y_\sigma^2\|_{B_{2,p}^{s-1}}^p d\sigma$  we note that  $\sqrt{\Lambda} : B_{2,p}^s \longrightarrow B_{2,p}^{s-1}$  continuously. Thus

$$\begin{aligned} \mathbb{E} \int_0^T \|Y_\sigma^2\|_{B_{2,p}^{s-1}}^p d\sigma &= \mathbb{E} \int_0^T \left\| \int_0^\sigma (\sigma-s)^{-\alpha} \cos\left(\sqrt{\Lambda}(\sigma-s)\right) dW_s \right\|_{B_{2,p}^{s-1}}^p d\sigma \\ &= \mathbb{E} \int_0^T \left\| \sqrt{\Lambda} \int_0^\sigma \frac{1}{\sqrt{\Lambda}} (\sigma-s)^{-\alpha} \cos\left(\sqrt{\Lambda}(\sigma-s)\right) dW_s \right\|_{B_{2,p}^{s-1}}^p d\sigma \\ &\leq c \mathbb{E} \int_0^T \left\| \int_0^\sigma \frac{1}{\sqrt{\Lambda}} (\sigma-s)^{-\alpha} \cos\left(\sqrt{\Lambda}(\sigma-s)\right) dW_s \right\|_{B_{2,p}^{s-1}}^p d\sigma < \infty. \end{aligned}$$

We have achieved the desired estimate

$$\mathbb{E} \int_0^T \|Y_\sigma\|_E^p d\sigma < \infty,$$

from which we get that the stochastic convolution  $W_A(t)$  admits a version in  $C([0, T], E)$ .  $\square$

We remark that Hypothesis 3.16 is satisfied; indeed,  $G(\Xi) \subset E$ .

Now, let us consider, for  $0 \leq t \leq \tau \leq T$  and  $\xi \in [0, 1]$ , the following controlled stochastic wave equation:

$$(7.17) \quad \begin{cases} \frac{\partial^2}{\partial \tau^2} y(\tau, \xi) = \frac{\partial^2}{\partial \xi^2} y(\tau, \xi) + u(\tau, \xi) + \dot{W}(\tau, \xi), \\ y(\tau, 0) = y(\tau, 1) = 0, \\ y(t, \xi) = x_0(\xi), \\ \frac{\partial y}{\partial \tau}(t, \xi) = x_1(\xi). \end{cases}$$

$\dot{W}(\tau, \xi)$  is a space-time white noise on  $[0, T] \times [0, 1]$  and  $u(\tau, \cdot)$  is an admissible control, that is, a predictable process

$$\left( \Omega, \mathcal{F}, (\mathcal{F}_\tau)_{\tau \geq 0}, \mathbb{P} \right) \rightarrow L^2(0, 1),$$

taking values in the subset  $\mathcal{U} \subset L^2[0, 1]$ ,  $\mathcal{U} = \{v \in L^2[0, 1] \mid v : [0, 1] \longrightarrow [-\delta, \delta]\}$ . Equation (7.17) can be rewritten in an abstract way in the following form:

$$(7.18) \quad \begin{cases} dX_\tau^u = AX_\tau^u d\tau + Gu_\tau d\tau + GdW_\tau, & \tau \in [t, T], \\ X_t^u = x, \end{cases}$$

where  $\{W_\tau, \tau \geq 0\}$  is a cylindrical Wiener process in  $L^2([0, 1])$  with respect to the filtration  $(\mathcal{F}_\tau)_{\tau \geq 0}$ .

Moreover, we introduce the cost functional

$$\gamma(t, x_0, x_1, u) = \mathbb{E} \int_t^T \left[ \sum_{i=1}^n l(s, \xi_i, y(s, \xi_i)) + \int_0^1 \frac{u^2(s, \xi)}{2} d\xi \right] ds + \mathbb{E} \sum_{i=1}^n k(\xi_i, y(T, \xi_i)).$$

The optimal control problem is to minimize  $\gamma$  over all admissible controls. To satisfy Hypothesis 5.2 we have to assume the following.

*Hypothesis 7.9.* We make the following assumptions:

1. For every  $i = 1, \dots, n$ ,  $l(\cdot, \xi_i, \cdot) : [0, T] \times \mathbb{R} \rightarrow \mathbb{R}$  is continuous and bounded, and the map  $l(\tau, \xi_i, \cdot) : \mathbb{R} \rightarrow \mathbb{R}$  is differentiable and there exists  $c_1$  such that

$$|\nabla_x l(\tau, \xi_i, x) h| \leq c_1 L |h|$$

for every  $\tau \in [0, T]$  and  $x, h \in \mathbb{R}$ .

2. For every  $i = 1, \dots, n$ ,  $k(\xi_i, \cdot) : [0, 1] \times \mathbb{R} \rightarrow \mathbb{R}$  is continuous and  $k(\xi_i, \cdot) \in C_b^1(\mathbb{R})$ ; moreover,  $|\nabla_x k(\xi_i, x)| \leq c_2$ .
3.  $x_0 \in H^1([0, 1])$  and  $x_1 \in L^2([0, 1])$ .

Let  $\Pi_1 : E \rightarrow B_{2,p,\{0\}}^s([0, 1])$  be the projection on the first component of  $E$ . We set

$$g(s, x) + \frac{1}{2} \|u\|_{L^2(0,1)}^2 = \sum_{i=1}^n l(s, \xi_i, \Pi_1 x(\xi_i)) + \int_0^1 \frac{u^2(s, \xi)}{2} d\xi,$$

$$\phi(x) = \sum_{i=1}^n k(\xi_i, \Pi_1 x(\xi_i)) d\xi.$$

In abstract formulation, the cost functional can be written as

$$J(t, x, u) = \mathbb{E} \int_t^T \left[ g(s, X_s^u) + \frac{1}{2} \|u_s\|_{L^2(0,1)}^2 \right] ds + \mathbb{E} \phi(X_T^u).$$

With Hypothesis 7.9,  $\phi$  and  $g$  satisfy Hypothesis 5.2.

The Hamiltonian is given by

$$\psi(t, x, z) = g(t, x) + \inf_u \left\{ \frac{1}{2} \|u_s\|_{L^2(0,1)}^2 + \int_0^1 z(\xi) u(s, \xi) d\xi \right\}.$$

$\psi$  and  $\phi$  satisfy Hypothesis 4.1. In fact, to verify Gâteaux differentiability of  $g$  and  $\phi$  with respect to  $x \in E$ , we note that both  $g$  and  $\phi$  depend only on  $\Pi_1 x$ . So it suffices to show Gâteaux differentiability with respect to  $\Pi_1 x \in B_{2,p,\{0\}}^s([0, 1])$ . Indeed,  $g(s, \cdot, u)$  and  $\phi(\cdot)$  are defined for  $\Pi_1 x \in C([0, 1])$ , and they are Gâteaux differentiable in  $C([0, 1])$ . Thus they are also Gâteaux differentiable with respect to  $\Pi_1 x \in B_{2,p,\{0\}}^s([0, 1])$ . Differentiability with respect to  $z$  can be proved, as we have done while treating the controlled heat equation; see section 7.1.

The process  $u$  turns out to be an admissible control and  $(\Omega, \mathcal{F}, \mathcal{F}_\tau, \mathbb{P}, W, u, X^u)$  turns out to be an a.c.s. We want to study the optimal control problem related to (7.15).

We write the Hamilton–Jacobi–Bellman equation relative to (7.4):

$$(7.19) \quad \begin{cases} \frac{\partial v}{\partial t}(t, x) = -\mathcal{A}_t v(t, x) - \psi(t, x, \nabla v(t, x) G), & t \in [0, T], \ x \in H, \\ u(T, x) = \phi(x). \end{cases}$$

**THEOREM 7.10.** *Assume that Hypothesis 7.9 is satisfied. Then (7.19) has a unique mild solution  $v$ , and, moreover, for all a.c.s.'s  $(W, u, X^u)$  we have  $J(t, x, (W, u, X^u)) \geq v(t, x)$ , and the equality holds if and only if (5.12) holds. If (5.13) holds, there exists an optimal a.c.s.*

**Acknowledgments.** I would like to thank Marco Fuhrman and Gianmario Tesitore for their constant interest in this work, Alessandra Lunardi for suggestions about interpolation theory and Besov spaces, and Susanna Piazzera for bibliographical references on delay equations.



## REFERENCES

- [1] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Systems Control Found. Appl. 2, Birkhäuser Boston, Inc., Boston, MA, 1990.
- [2] J.-M. BISMUT, *Martingales, the Malliavin calculus and hypoellipticity under general Hörmander's conditions*, Z. Wahrsch. Verw. Gebiete, 56 (1981), pp. 469–505.
- [3] S. CERRAI, *Differentiability of Markov semigroups for stochastic reaction-diffusion equations and applications to control*, Stochastic Process Appl., 83 (1999), pp. 15–37.
- [4] S. CERRAI, *Optimal control problems for stochastic reaction-diffusion systems with non-Lipschitz coefficients*, SIAM J. Control Optim., 39 (2001), pp. 1779–1816.
- [5] J. B. CONWAY, *A Course in Functional Analysis*, 2nd ed., Grad. Texts Math. 96, Springer-Verlag, New York, 1990.
- [6] M. G. CRANDALL, M. G. KOCAN, AND A. SWIECH, *On partial sup-convolutions, a lemma of P. L. Lions and viscosity solutions in Hilbert spaces*, Adv. Math. Sci. Appl., 3 (1993/94), pp. 1–15.
- [7] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations in Banach spaces*, in Trends in the Theory and Practice of Nonlinear Analysis (Arlington, TX, 1984), Math. Stud. 110, North-Holland, Amsterdam, 1985, pp. 115–119.
- [8] M. G. CRANDALL AND P.-L. LIONS, *Solutions de viscosité pour les équations de Hamilton-Jacobi dans des espaces de Banach*, C. R. Acad. Sci. Paris Sér. I Math., 300 (1985), pp. 67–70.
- [9] M. G. CRANDALL AND P.-L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.
- [10] G. DA PRATO, *Applications Croissantes et Equations D'évolution Dans les Espaces de Banach*, Academic Press, London, New York, 1976.
- [11] G. DA PRATO AND L. TUBARO, *Some results on semilinear stochastic differential equations in Hilbert spaces*, Stochastics, 15 (1985), pp. 271–281.
- [12] G. DA PRATO AND J. ZABCZYK, *Stochastic Equations in Infinite Dimensions*, Encyclopedia Math. Appl. 44, Cambridge University Press, Cambridge, UK, 1992.
- [13] G. DA PRATO AND J. ZABCZYK, *Ergodicity for Infinite-Dimensional Systems*, London Math. Soc. Lecture Note Ser. 229, Cambridge University Press, Cambridge, UK, 1996.
- [14] M. C. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constant delay*, J. Differential Equations, 12 (1974), pp. 213–235.
- [15] W. H. FLEMING AND H. M. SONER, *Controlled Markov Process and Viscosity Solutions*, Appl. Math. 25, Springer-Verlag, New York, 1993.
- [16] M. FUHRMAN AND G. TESSITORE, *Nonlinear Kolmogorov equations in infinite dimensional spaces: The backward stochastic differential equations approach and applications to optimal control*, Ann. Probab., 30 (2002), pp. 1397–1465.
- [17] F. GOZZI, *Regularity of solutions of second order Hamilton-Jacobi equations in Hilbert spaces and applications to a control problem*, Comm. Partial Differential Equations, 20 (1995), pp. 775–826.
- [18] I. GYÖNGY AND E. PARDOUX, *On quasi-linear stochastic partial differential equations*, Probab. Theory Related Fields, 94 (1993), pp. 413–425.
- [19] P.-L. LIONS, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimension I. The case of bounded stochastic evolutions*, Acta Math., 161 (1988), pp. 243–278.
- [20] P.-L. LIONS, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimension III. Uniqueness of viscosity solutions for general second-order equations*, J. Funct. Anal., 86 (1989), pp. 1–18.
- [21] A. LUNARDI, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Progr. Nonlinear Differential Equations Appl. 16, Birkhäuser-Verlag, Basel, 1995.
- [22] J. MA AND J. YONG, *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Math. 1702, Springer, Berlin, 1999.
- [23] R. MANTHEY, *Existence and uniqueness of a solution of a reaction-diffusion equation with polynomial nonlinearity and white noise disturbance*, Math. Nachr., 125 (1986), pp. 121–133.
- [24] F. MASIERO, *Semilinear Kolmogorov equations and applications to stochastic optimal control*, Appl. Math. Optim., 51 (2005), pp. 201–250.
- [25] E. PARDOUX AND S. PENG, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.
- [26] E. PARDOUX AND S. PENG, *Backward stochastic differential equations and quasilinear parabolic partial differential equations*, in Stochastic Partial Differential Equations and Their Applications, Lecture Notes in Control and Inform. Sci. 176, Springer-Verlag, New York, 1992, pp. 200–217.

- [27] R. E. SHOWALTER, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, Math. Surveys Monographs 49, AMS, Providence, RI, 1997.
- [28] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland Math. Library 18, North-Holland, Amsterdam, New York, 1978.
- [29] J. B. WALSH, *An introduction to stochastic partial differential equations*, École d'été de Probabilités de Saint-Flour, XIV—1984, Lecture Notes in Math. 1180, Springer, Berlin, 1986, pp. 265–439.
- [30] G. F. WEBB, *Functional differential equations and nonlinear semigroups in  $L^p$ -spaces*, J. Differential Equations, 20 (1976), pp. 71–89.

## CONTROL LYAPUNOV FUNCTIONS AND ZUBOV'S METHOD\*

FABIO CAMILLI<sup>†</sup>, LARS GRÜNE<sup>‡</sup>, AND FABIAN WIRTH<sup>§</sup>

**Abstract.** For finite-dimensional nonlinear control systems we study the relation between asymptotic null-controllability and control Lyapunov functions. It is shown that control Lyapunov functions (CLFs) may be constructed on the domain of asymptotic null-controllability as viscosity solutions of a first order PDE that generalizes Zubov's equation. The solution is also given as the value function of an optimal control problem from which several regularity results may be obtained.

**Key words.** asymptotic null-controllability, control Lyapunov functions, Hamilton–Jacobi–Bellman equation, viscosity solutions, Zubov's method

**AMS subject classifications.** 93D10, 35B37, 49L25

**DOI.** 10.1137/06065129X

**1. Introduction.** We consider finite-dimensional control systems of the form

$$(1.1) \quad \dot{x}(t) = f(x(t), u(t)),$$

where  $x \in \mathbb{R}^n$  denotes the state,  $u \in \mathbb{R}^m$  denotes the input, and  $f$  is sufficiently regular with  $f(0, 0) = 0$ . We call a point  $x_0 \in \mathbb{R}^n$  *asymptotically controllable* to 0 if there exists a measurable, essentially bounded function  $u_0 : \mathbb{R}_+ \rightarrow \mathbb{R}^m$  such that the corresponding solution  $\varphi(t, x_0, u_0)$  of (1.1) satisfies  $\varphi(t, x_0, u_0) \rightarrow 0$  for  $t \rightarrow \infty$ . The *domain of asymptotic null-controllability* is the collection of all points that are asymptotically controllable to 0. The main results of this paper are twofold: On the one hand, we provide a converse theorem for maximal *control Lyapunov functions* (CLFs) on the domain of asymptotic null-controllability. On the other hand, we consider a generalized Zubov equation and prove that the maximal CLF is the unique viscosity solution of this equation. Thus beyond the proof of existence of CLFs, a way to their numerical generation is provided.

The construction of CLFs in this paper relies on optimal control methods as they are frequently used in Lyapunov theory. One of the contributions of the paper is to present easily checkable conditions on the running cost that result in an appropriate CLF and give rise to a tractable Hamilton–Jacobi equation.

Converse theorems have a fundamental role in Lyapunov theory, as they state that certain stability properties imply the existence of a Lyapunov function. The direct implication that the existence of a Lyapunov function implies a stability property is usually much easier to prove. Early converse results were obtained by Persidskii (see the discussion in [20, Chapter VI]), Massera [26], and Kurčevič' [21]. In recent times these results have been extended in several directions to cover perturbed systems and differential inclusions [23, 12, 37, 7].

---

\*Received by the editors January 31, 2006; accepted for publication (in revised form) July 3, 2007; published electronically January 22, 2008.

<http://www.siam.org/journals/sicon/47-1/65129.html>

<sup>†</sup>Sez. di Matematica per l'Ingegneria, Dip. di Matematica Pura e Applicata, Università de l'Aquila, 67040 Roio Poggio (AQ), Italy (camilli@ing.univaq.it).

<sup>‡</sup>Mathematisches Institut, Fakultät für Mathematik und Physik, Universität Bayreuth, 95440 Bayreuth, Germany (lars.gruene@uni-bayreuth.de).

<sup>§</sup>Institut für Mathematik, Universität Würzburg, 97074 Würzburg (wirth@mathematik.uni-wuerzburg.de). This author was supported by the Science Foundation Ireland grant 04-IN3-I460.

While for linear systems a constructive procedure to find Lyapunov functions has already been given by Lyapunov, the first general constructive procedure to find Lyapunov functions was obtained by Zubov [39]. Namely, a Lyapunov function on the domain of attraction of an asymptotically stable fixed point  $x^* \in \mathbb{R}^n$  of the system

$$\dot{x}(t) = f(x(t)), \quad t \in \mathbb{R}, x \in \mathbb{R}^n,$$

may be found by solving the first order PDE, called Zubov's equation,

$$Dv(x)f(x) = -h(x)(1 - v(x))\sqrt{1 + \|f\|^2}, \quad x \in \mathbb{R}^n,$$

under the condition that  $v(0) = 0$ . Here  $h$  is an auxiliary function; see [39, 20] for details. This method has been recently extended by the authors to the case of perturbed systems; see [9], where also a discussion of the impact of Zubov's result may be found. Further constructive approaches valid for  $C^2$  systems and based on approximations by radial basis functions (respectively, on linear programming methods) have recently been described in [18, 19].

While for (perturbed) ordinary differential equations the property of interest is stability, for systems with control inputs a basic question concerns the existence of control functions steering the system to a desired target. In contrast to the case of asymptotically stable fixed points of ordinary differential equations, for which smooth Lyapunov functions always exist, it is not reasonable to require too many regularity properties of Lyapunov functions for controllability questions for systems of the form (1.1). For this reason it is now standard to formulate the concept of a CLF in non-differential terms. Recall that a function  $V : \mathbb{R}^n \rightarrow \mathbb{R}$  is called positive definite if  $V(x) \geq 0$  for all  $x \in \mathbb{R}^n$  and  $V(x) = 0$  iff  $x = 0$ . The function  $V$  is proper if preimages of compact sets are compact. A positive definite, proper function  $V$  is called a CLF for (1.1) if there is a positive definite function  $W$  such that for every compact set  $X \subset \mathbb{R}^n$  there is a compact set  $U_X$  of control values so that  $V$  is a continuous viscosity supersolution of

$$(1.2) \quad \max_{u \in U_X} -DV(x)f(x, u) \geq W(x), \quad x \in X.$$

For the definition of viscosity solutions we refer to [3]. In many articles CLFs are defined in terms of proximal subgradients of  $V$ , but the two notions are in fact equivalent [10].

The interest in the theory of control Lyapunov has received widespread attention in recent years, in particular in connection with the design of stabilizing feedbacks. While design techniques using Lyapunov functions have been popular in applied control theory for a long time, the systematic study of converse theorems for CLFs only started with Artstein [1], who proved for the case of systems affine in the control term  $u$  that the existence of a smooth CLF is equivalent to stabilizability by continuous state feedback. For general systems of the form (1.1) the existence of a global continuous CLF is equivalent to global asymptotic null-controllability [31]. Interestingly, the existence of a differentiable CLF is equivalent to the existence of (discontinuous) stabilizing feedbacks that are robust with respect to perturbations in the measurement of the state [22].

Now, in general, asymptotic null-controllability does not imply the existence of continuous stabilizing feedback as there may be topological obstructions to this which even carry over to the case of upper semicontinuous set-valued feedbacks [8, 13, 29]. For this reason discontinuous feedbacks and associated solution concepts have been

one of the focal points of the research on CLFs in recent times starting with [11]. In this context it has been shown by Clarke et al. [10] and Rifford [27, 28] using tools from nonsmooth analysis that semiconcavity of the CLF is an essential tool in order to establish the existence of feedback with nice properties.

Usually, the knowledge of a CLF requests a certain structure of the control system, while a general procedure for its determination is not available. Constructive approaches have therefore received widespread attention in the literature, most notably with techniques known as backstepping and forwarding [17, 30], which, however, rely heavily on the differentiability of the CLF that is obtained. In this article we aim to derive a constructive approach by going back to the original ideas for the construction of CLFs. Here constructive is to be understood in the way that we determine a class of PDEs which have unique solutions in the viscosity sense that are maximal CLFs on the domain of asymptotic null-controllability.

It is a classical approach to the problem to regard CLFs as solutions of steady state Hamilton–Jacobi (HJ) equations. In the uncontrolled case this may be regarded as one of the central elements of the work of Zubov [20]. In [16] the connection between smooth CLFs and HJ equations has been studied in detail. In particular, it is shown in that paper that smooth CLFs may always be interpreted as value functions of an appropriate optimal control problem. This “inverse optimality” property can be exploited in several ways [17]. In a different approach, in [15] a CLF was obtained by a truncating series expansion of analytical solutions of HJ equations in an approach very similar to early studies around Zubov’s equation.

In the present paper we use ideas from [9], where, for the case of a perturbed system, the classical Zubov method was reinterpreted using a suitable notion of weak solution. For controlled or perturbed systems Zubov’s equation becomes a nonlinear first order PDE of HJ type, and it is well known that this class of equations does not admit, in general, classical solutions. Therefore a suitable concept of weak solution has to be introduced, and the one of viscosity solution seems to be appropriate; see [9, 25]. In the construction of the corresponding result for CLFs, several additional technical obstacles have to be overcome, which stem from the possibility of solutions with finite escape time and the unboundedness of the control set, both of which pose no problem in the perturbed case. To this end reparametrization techniques are used—an idea which was introduced in [5] and has been applied by various authors.

A problem similar to the CLF construction in this paper has been studied in [24], in which optimal control and viscosity methods are used in relation to the problem of steering the state of a system to a prescribed target. This leads to an optimal control problem with a positive but vanishing Lagrangian; see [25, 24] for further references. In [24] a “small-time controllability” property is used, which requires in particular that the target can be reached exactly in finite time starting from small neighborhoods of it. In the general context of CLFs, this reachability property is undesirable, so that we have to apply different arguments.

We use this generalization of Zubov’s method to construct a CLF for a finite-dimensional nonlinear control system that is asymptotically null-controllable in a neighborhood of the origin. Our aim is to determine a CLF as (i) an optimal value function of a suitable control problem and (ii) a unique viscosity solution to a suitable HJ equation which is a generalization of the Zubov’s equation.

Concerning the first point, i.e., the connection between CLF and optimal control problems, our procedure can be viewed as an extension of [31], where the equivalence between asymptotic null-controllability and the existence of a CLF has been proved

using an optimal control approach. The significant advantage of the characterization of a CLF as a *unique* viscosity solution of the generalized Zubov equation is that this characterization can be used as the basis for the numerical approximation of the CLF.

From the point of view of the PDE approach, the equation presents some difficulties when attacked using the standard theory of viscosity solution because of the unbounded control set; see [4, 5, 14, 36, 35] for some related papers. In the proof of the necessary comparison result we use the local asymptotic controllability to obtain a local comparison result in a neighborhood of the origin. We then extend the comparison result to all  $\mathbb{R}^n$ , taking advantage, as in the classical Zubov method, of the freedom in the choice of cost function of the associated control problem. For this reason we can make rather general assumptions on the dependence of the dynamics with respect to the control variable compensating them with an appropriate choice of the cost. An example for an explicit construction of the cost function satisfying all requirements is provided.

The comparison result extends results in [36, 35] at the price of studying a much more specific situation. The main difference is that in the setting studied in this paper a uniqueness result is obtained.

We proceed as follows: In the ensuing section 2 the class of systems under consideration is defined, and we prove some preliminary results. In section 3 the optimal control problem that characterizes the domain of asymptotic null-controllability is introduced, and it is shown that under suitable conditions the corresponding value function is continuous, positive definite, and proper on the domain of asymptotic null-controllability. In section 4 we show that the value function of the optimal control problem is the unique viscosity solution of the generalized Zubov equation. In section 5 we discuss an approximation of the problem with an unbounded control set with a sequence of problems with a bounded control set. In the last section we discuss the necessity of our assumptions at the hand of a few examples. It is also shown that for the classical linear quadratic control problem the general equations of this paper reduce to the standard algebraic Riccati equation.

**2. The domain of null-controllability.** We consider nonlinear control systems of the type

$$(2.1) \quad \dot{x}(t) = f(x(t), u(t)),$$

where  $f : \mathbb{R}^n \times U \rightarrow \mathbb{R}^n$  is continuous,  $U \subset \mathbb{R}^m$  is a closed set, and the space of admissible control functions is given by

$$u \in \mathcal{U} := L^\infty([0, \infty), U).$$

Solutions corresponding to an initial value  $x$  and a control  $u \in \mathcal{U}$  at time  $t$  are denoted by  $\varphi(t, x, u)$ , which are defined on a maximal positive interval of definition  $[0, T_{\max}(x, u))$ , where we do not exclude the case that  $T_{\max}(x, u) < \infty$ , i.e., that solutions explode. In the following the open ball of radius  $r$  around a point  $z \in \mathbb{R}^p$  is denoted by  $B(z, r)$ .

Uniqueness of solutions is a consequence of our further standard assumption on  $f$ . These are formulated using comparison functions, a fashionable approach these days.<sup>1</sup>

---

<sup>1</sup>As usual we call a function  $\alpha$  of class  $\mathcal{K}_\infty$  if it is a homeomorphism of  $[0, \infty)$ , and a continuous function  $\beta$  in two real nonnegative arguments is called of class  $\mathcal{KL}$  if it is of class  $\mathcal{K}_\infty$  in the first and decreasing to zero in the second argument.

(H0) There exists  $\gamma \in \mathcal{K}_\infty$  such that for any  $R > 0$  there is  $C_R > 0$ , with

$$\|f(x, u) - f(y, u)\| \leq C_R(1 + \gamma(\|u\|))\|x - y\|,$$

for all  $x, y$  with  $\|x\|, \|y\| \leq R$ .

(H1)  $f(0, 0) = 0$ .

(H2) There exist an open ball  $B(0, r)$ , a constant  $\bar{u} > 0$ , and  $\beta \in \mathcal{KL}$  such that for any  $x \in B(0, r)$  there exists  $u_x \in \mathcal{U}$  with  $\|u_x\|_\infty \leq \bar{u}$ ,  $T_{\max}(x, u_x) = \infty$ , and

$$\|\varphi(t, x, u_x)\| \leq \beta(\|x\|, t) \quad \forall t \geq 0.$$

*Remark 2.1.* The Lipschitz assumption (H0) is weaker than the following assumption: For any  $R > 0$  there exists  $C_R > 0$ , with

$$(2.2) \quad \|f(x, u) - f(y, u)\| \leq C_R(1 + \|u\|)\|x - y\|,$$

for all  $x, y$  with  $\|x\|, \|y\| \leq R$ .

Assumption (2.2) is used in many papers on viscosity solutions, in particular in [35, 36], whose results we will use later. In order to be able to use these results under the weaker assumption (H0), we define the map  $R : \mathbb{R}^m \rightarrow \mathbb{R}^m$  by  $R(u) = \gamma^{-1}(\|u\|)u/\|u\|$  and consider the vector field

$$\hat{f}(x, u) = f(x, R(u)),$$

with  $u \in \tilde{U} := R^{-1}(u)$ . This input-transformed system satisfies

$$\|\hat{f}(x, u) - \hat{f}(y, u)\| \leq C_R(1 + \gamma(\|R(u)\|))\|x - y\| = C_R(1 + \|u\|)\|x - y\|,$$

i.e., (2.2). Hence by applying the results from [35, 36] to  $\hat{f}$ , these immediately carry over to  $f$  under the weaker assumption (H0).

Property (H2) is a local asymptotic controllability property, which ensures that at least from a neighborhood of 0 the system may be steered to 0.

For certain systems it makes sense to strengthen this local asymptotic controllability property (H2) by requiring that  $u_x$  not only is bounded but also converges to 0 as  $t \rightarrow \infty$ . In this case we can strengthen (H2) to the so-called *small control property*:

(H2') There exist an open ball  $B(0, r)$  and  $\beta \in \mathcal{KL}$  such that for any  $x \in B(0, r)$  there exists  $u_x \in \mathcal{U}$  with  $T_{\max}(x, u_x) = \infty$  and

$$\|\varphi(t, x, u_x)\| + \|u_x(t)\| \leq \beta(\|x\|, t), \quad \text{a.e. } t \geq 0.$$

Note that (H2') implies (H2) with  $\bar{u} = \beta(r, 0)$ .

It is known [32] that for any  $\beta \in \mathcal{KL}$  there exist two functions  $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$  such that  $\beta(r, t) \leq \alpha_2(\alpha_1(r)e^{-t})$ . These functions can be computed from  $\beta$ ; e.g., in the case of exponential convergence, i.e.,  $\beta(r, t) = ce^{-\sigma t}r$  for  $c, \sigma > 0$ , one obtains  $\alpha_1(r) = r^{1/\sigma}$  and  $\alpha_2(r) = cr^\sigma$ . Note that (H2) or (H2') immediately implies  $\beta(r, 0) \geq r$  and thus

$$\alpha_2 \circ \alpha_1(r) = \alpha_2(\alpha_1(r)e^{-0}) \geq \beta(r, 0) = r.$$

We note for later use that, by applying  $\alpha_1^{-1} \circ \alpha_2^{-1} = (\alpha_2 \circ \alpha_1)^{-1}$  on both sides of this inequality, we obtain

$$(2.3) \quad r \geq \alpha_1^{-1} \circ \alpha_2^{-1}(r).$$

For ease of presentation we will work with these two functions from now on. Furthermore, we will from now on tacitly assume that  $T_{\max}(x, u) = \infty$  if we write  $\varphi(t, x, u) \rightarrow 0$  as  $t \rightarrow \infty$ .

We define the *domain of null-controllability* by

$$\mathcal{D}_0 := \{x \in \mathbb{R}^n \mid \text{there exists } u \in \mathcal{U} \text{ with } \|\varphi(t, x, u)\| \rightarrow 0 \text{ for } t \rightarrow \infty\}$$

and the *first hitting time* with respect to  $B(0, r)$  by

$$t(x, u) := \inf\{t \geq 0 \mid \varphi(t, x, u) \in B(0, r)\},$$

with the convention  $\inf \emptyset = \infty$ . The following lemma shows how  $\mathcal{D}_0$  and  $t(x, u)$  are related.

LEMMA 2.2. *The set  $\mathcal{D}_0$  is given by*

$$\mathcal{D}_0 = \left\{x \in \mathbb{R}^n \mid \inf_{u \in \mathcal{U}} t(x, u) < \infty\right\}.$$

*Proof.* If we find  $u \in \mathcal{U}$  with  $t(x, u) < \infty$ , then for some  $t(x, u) < t_1$  we have  $\varphi(t_1, x, u) \in B(0, r)$ , and we can concatenate  $u|_{[0, t_1]}$  with the control  $u_{\varphi(t_1, x, u)}$  from (H2), which implies  $\varphi(t, x, u) \rightarrow 0$ . Hence we obtain

$$\mathcal{D}_0 \subseteq \left\{x \in \mathbb{R}^n \mid \inf_{u \in \mathcal{U}} t(x, u) < \infty\right\}.$$

Conversely, if  $x \in \mathcal{D}_0$ , then we have  $\varphi(t, x, u) \rightarrow 0$  for some suitable  $u \in \mathcal{U}$ , which implies  $\varphi(t_1, x, u) \in B(0, r)$  for some  $t_1 > 0$ , and consequently  $t(x, u) \leq t_1 < \infty$ , which implies the converse direction.  $\square$

For the formulation of the next result recall that a set  $M$  is called *viable* (or controlled or weakly invariant) if for every  $x \in M$  there is a  $u \in \mathcal{U}$  such that  $\varphi(t, x, u) \in M$  for all  $t \geq 0$  (see [2]). In the following the convex hull of a set  $M$  is denoted by  $\text{co } M$ .

PROPOSITION 2.3. *Assume (H0), (H1), and (H2) or (H2'). Then the following properties hold:*

- (i)  $\text{cl } B(0, r) \subset \mathcal{D}_0$ ;
- (ii) *the set  $\mathcal{D}_0$  is open, connected, and viable.*

*Proof.* (i) It is clear that  $B(0, r) \subset \mathcal{D}_0$ . In order to show  $\text{cl } B(0, r) \subset \mathcal{D}_0$ , pick  $x \in \partial B(0, r)$  and a sequence  $\{x_n\} \subset B(0, r)$  with  $\lim_{n \rightarrow \infty} x_n = x$ . By assumption for each  $x_n$  there exists a control  $u_n \in \mathcal{U} \cap L^\infty(\mathbb{R}, B(0, \bar{u}))$  such that  $\|\varphi(t, x_n, u_n)\| \leq \alpha_2(\alpha_1(r)e^{-t})$ . This shows that on each compact interval the solutions are bounded uniformly in  $n$ . Since, furthermore, the  $u_n$  are uniformly bounded by continuity, we get  $\lim_{n \rightarrow \infty} \|\varphi(t, x_n, u_n) - \varphi(t, x, u_n)\| = 0$  for each  $t \geq 0$ . Thus, by picking  $t^* > 0$  such that  $\alpha_2(\alpha_1(r)e^{-t^*}) \leq r/2$  for  $n^*$  sufficiently large, we obtain  $\|\varphi(t^*, x, u_{n^*})\| \leq 3r/4$ ; hence,  $\varphi(t^*, x, u_{n^*}) \in B(0, r)$ , and consequently  $t(x, u_{n^*}) < \infty$ . Now Lemma 2.2 yields the assertion.

(ii) Let  $x_0 \in \mathcal{D}_0$  and  $u \in \mathcal{U}$ , with  $\varphi(t, x_0, u) \rightarrow 0$  for  $t \rightarrow \infty$ . Then there exists  $T > 0$  such that  $\varphi(T, x_0, u) \in B(0, r)$ . By continuous dependence on the initial value we obtain

$$\varphi(T, x, u) \in B(0, r)$$

for all  $x$  in a neighborhood of  $x_0$ . Thus  $t(\cdot, u)$  is finite on that neighborhood, which shows that it is contained in  $\mathcal{D}_0$ . As  $x_0$  was arbitrary, this shows the assertion.



Since for any  $x \in \mathcal{D}_0$  there exists a trajectory from  $x$  to  $B(0, r)$ , we obtain that  $\mathcal{D}_0$  is connected.

In order to see viability, consider a point  $x \in \mathcal{D}_0$  and the trajectory  $\varphi(t, x, u) \rightarrow 0$ . Clearly, each point  $x(t) = \varphi(t, x, u)$ ,  $t \geq 0$ , can be controlled to the origin by the control  $u(t + \cdot)$ ; thus  $x(t) \in \mathcal{D}_0$ , and hence  $\varphi(t, x, u) \in \mathcal{D}_0$  for all  $t \geq 0$ ; i.e.,  $\mathcal{D}_0$  is viable.  $\square$

*Remark 2.4.* Note that the domain of null-controllability  $\mathcal{D}_0$  is in general not diffeomorphic to  $\mathbb{R}^n$ . This is in contrast to the theory of domains of attraction of (perturbed) ordinary differential equations, i.e., the set  $\{x_0 \in \mathbb{R}^n : \varphi(t, x_0, u) \rightarrow 0 \text{ as } t \rightarrow +\infty \text{ for any } u \in \mathcal{U}\}$ . In the case of asymptotically stable fixed points the domain of attraction is diffeomorphic to  $\mathbb{R}^n$  even for perturbed systems; see, e.g., [9, 38].

**3. Characterization of  $\mathcal{D}_0$  using optimal control.** In this section we describe how to characterize the domain of asymptotic null-controllability via an optimal control problem and show continuity of the corresponding value function. In order to set up the problem we need a running cost  $g : \mathbb{R}^n \times U \rightarrow \mathbb{R}$ . The assumptions on  $g$  are as follows:

(H3) The function  $g : \mathbb{R}^n \times U \rightarrow \mathbb{R}$  is continuous and satisfies (H0) with the same  $\gamma \in \mathcal{K}_\infty$  as  $f$ . Furthermore, for all  $c > 0$  we have

$$\inf \{g(x, u) \mid \|x\| \geq c, u \in U\} =: g_c > 0.$$

Note that the assumption “with the same  $\gamma \in \mathcal{K}_\infty$  as  $f$ ” can always be met by enlarging the  $\gamma$  from (H0) for  $f$ , if necessary.

We need to ensure convergence of the integral cost that is introduced shortly for the “right” stabilizing solutions. Recall that we use the simplification  $\beta(r, t) \leq \alpha_2(\alpha_1(r)e^{-t})$  for  $\beta$  from (H2) and choose some arbitrary  $\eta > 0$ . We assume that there exists a constant  $C > 0$  such that

$$(H4) \quad g(x, u) \leq C(\alpha_2^{-1}(\|x\|))^\eta \quad \forall (x, u) \in B(0, r) \times B(0, \bar{u}),$$

$$(H5) \quad g(x, u) \geq \|f(x, u)\| + \gamma(\|u\|) \text{ whenever } \|x\| \geq 2r \text{ or } \|u\| \geq 2\bar{u}.$$

*Remark 3.1.* If the small control asymptotic controllability property (H2') holds, then we can weaken assumption (H4) to

$$(H4') \quad g(x, u) \leq C(\alpha_2^{-1}(\|x\| + \|u\|))^\eta \quad \forall (x, u) \in B(0, r) \times B(0, \bar{u}).$$

In what follows we will always assume that either (H2) and (H4) or (H2') and (H4') hold.

We now define the functional

$$(3.1) \quad J(x, u) := \begin{cases} \int_0^\infty g(\varphi(t, x, u), u(t)) dt & \text{if } T_{\max}(x, u) = \infty, \\ \infty & \text{otherwise,} \end{cases}$$

the (extended real-valued) optimal value function

$$(3.2) \quad V(x) := \inf_{u \in \mathcal{U}} J(x, u), \quad x \in \mathbb{R}^n,$$

and the function

$$(3.3) \quad v(x) := 1 - e^{-V(x)}, \quad x \in \mathbb{R}^n.$$

Note that both  $V$  and  $v$  satisfy appropriate dynamic programming principles (see, for example, [35, 36]); i.e., for each  $T > 0$  we have

$$(3.4) \quad V(x) = \inf_{u \in \mathcal{U}} \left\{ \int_0^T g(\varphi(t, x, u), u(t)) dt + V(\varphi(T, x, u)) \right\},$$

and

$$(3.5) \quad v(x) = \inf_{u \in \mathcal{U}} \{1 + G(x, T, u)(v(\varphi(T, x, u)) - 1)\},$$

where

$$G(x, T, u) := \exp \left( - \int_0^T g(\varphi(t, x, u), u(t)) dt \right).$$

We now investigate the properties of  $V$  and  $v$ . For this purpose we need the following observation on the solutions of (2.1). Using the function  $\gamma$  from (H0), we define for  $u \in \mathcal{U}$

$$\|u\|_{\gamma, T} := \int_0^T \gamma(\|u(t)\|) dt.$$

LEMMA 3.2. *Let  $T > 0$ . If  $x \in \mathbb{R}^n$  and  $u \in \mathcal{U}$  are such that  $\|\varphi(t, x, u)\| \geq 2r$ ,  $t \in [0, T]$ , or  $\|u(t)\| \geq 2\bar{u}$  a.e.  $t \in [0, T]$ , then*

$$\int_0^T g(\varphi(t, x, u), u(t)) dt \geq \|\varphi(T, x, u) - x\| + \|u\|_{\gamma, T}.$$

*Proof.* Using (H5), we have that

$$\int_0^T g(\varphi(t, x, u), u(t)) dt \geq \int_0^T \|f(\varphi(t, x, u), u(t))\| dt + \int_0^T \gamma(\|u(t)\|) dt,$$

and the claim follows.  $\square$

PROPOSITION 3.3. *Assume (H0)–(H4) or the respective variants from Remark 3.1. Then*

- (i) *the inequalities  $V(x) < \infty$  and  $v(x) < 1$  hold iff  $x \in \mathcal{D}_0$ , and*
- (ii) *if in addition (H5) holds, then  $V(x) = 0 \Leftrightarrow x = 0$  and  $v(x) = 0 \Leftrightarrow x = 0$ .*

*Proof.* From the definition of  $v$  it immediately follows that the claims for  $V$  and  $v$  are equivalent. We show the statements for  $V$ .

(i) Pick a point  $x \in \mathcal{D}_0$ . Then there exist  $u \in \mathcal{U}$  and  $t_1 > 0$  such that  $\|\varphi(t_1, x, u)\| \leq \alpha_1^{-1} \circ \alpha_2^{-1}(r)$ . (Note that  $\alpha_1^{-1} \circ \alpha_2^{-1}(r) \leq r$  by (2.3).) By assumption (H1) we can assume (by changing  $u$  on  $[t_1, \infty)$  if necessary) that  $\|\varphi(t_1 + t, x, u)\| \leq \alpha_2(\alpha_1(\|\varphi(t_1, x, u)\|)e^{-t}) \leq r$  for all  $t \geq 0$ . Since  $u \in \mathcal{U} = L^\infty([0, \infty), U)$  is essentially bounded, we can find  $\bar{u} > 0$  such that  $\|u(t)\| \leq \bar{u}$  for almost all  $t \geq 0$ . Furthermore, by continuity of  $\varphi(t, x, u)$  in  $t$  we find  $R > 0$  such that  $\|\varphi(t, x, u)\| \leq R$  for all  $t \in [0, t_1]$ . Hence using (H4), we can estimate

$$(3.6) \quad \begin{aligned} V(x) &\leq \int_0^{t_1} g(\varphi(t, x, u), u(t)) dt + \int_{t_1}^\infty g(\varphi(t, x, u), u(t)) dt \\ &\leq t_1 \sup_{x \in B(0, R), u \in B(0, \bar{u})} g(x, u) + \int_{t_1}^\infty C(\alpha_2^{-1}(\|\varphi(t, x, u)\|))^\eta dt \\ &\leq t_1 \sup_{x \in B(0, R), u \in B(0, \bar{u})} g(x, u) + \frac{C}{\eta} \alpha_1(\|\varphi(t_1, x, u)\|)^\eta < \infty. \end{aligned}$$

If (H2') and (H4') hold, then the proof is completely analogous.

Conversely, let  $x \notin \mathcal{D}_0$ . Then we obtain  $t(x, u) = \infty$  for all  $u \in \mathcal{U}$ , which implies

$$J(x, u) = \int_0^\infty g(\varphi(t, x, u), u(t)) dt \geq \int_0^\infty g_r dt = \infty$$

for each  $u \in \mathcal{U}$  and thus also  $V(x) = \inf_{u \in \mathcal{U}} J(x, u) = \infty$ .

(ii) It is clear that  $V(0) = 0$ , so let  $x \neq 0$ . Assume to the contrary that there is a sequence  $\{u_k\} \subset \mathcal{U}$  such that  $J(x, u_k) \rightarrow 0$ . Let  $c := \|x\|/2$ , and denote

$$t_k := \inf\{t \geq 0 \mid \|\varphi(t, x, u_k)\| \leq c\}.$$

By (H3) we have for all  $k$  that  $J(x, u_k) \geq \int_0^{t_k} g(\varphi(s, x, u_k), u_k(s)) ds \geq t_k g_c$ , which implies that  $t_k \rightarrow 0$ . Now  $\|f\|$  is bounded on  $B(0, 2r) \times B(0, 2\bar{u})$  by the constant  $C := C_{2r}(1 + \gamma(2\bar{u}))2r$ . Denote

$$E(k) := \{t \in [0, t_k] \mid (\varphi(t, x, u_k), u(t)) \in B(0, 2r) \times B(0, 2\bar{u})\},$$

which is well defined up to a set of measure zero. Then

$$\int_{E(k)} \|f(\varphi(t, x, u_k), u_k(t))\| dt \leq t_k C.$$

On the other hand, we have for all  $k$  that

$$\int_0^{t_k} \|f(\varphi(t, x, u_k), u_k(t))\| dt \geq \|x - \varphi(t_k, x, u_k)\| \geq c.$$

Using (H5), this implies that

$$\begin{aligned} J(x, u_k) &\geq \int_{[0, t_k] \setminus E(k)} g(\varphi(s, x, u_k), u_k(s)) ds \\ &\geq \int_{[0, t_k] \setminus E(k)} \|f(\varphi(s, x, u_k), u_k(s))\| ds \geq c - t_k C. \end{aligned}$$

As  $t_k \rightarrow 0$  this contradicts  $J(x, u_k) \rightarrow 0$ .  $\square$

Next we turn to the investigation of the regularity properties of the functions  $V$  and  $v$ . We start by proving continuity properties for the trajectories of (2.1).

LEMMA 3.4. Assume (H0), and let  $T > 0$  and  $R > 0$  be arbitrary constants. Then for all  $x, y \in \mathbb{R}^n$  and all  $u \in \mathcal{U}$  satisfying

$$\|\varphi(t, x, u)\| \leq R, \quad \|\varphi(t, y, u)\| \leq R \quad \forall t \in [0, T],$$

we have

$$(3.7) \quad \|\varphi(t, x, u) - \varphi(t, y, u)\| \leq e^{C_R(\|u\|_{\gamma, t} + t)} \|x - y\|$$

for all  $t \in [0, T]$ .

*Proof.* The assumption (H0) yields for almost all  $t \in [0, T]$

$$\begin{aligned} (3.8) \quad &\|f(\varphi(t, x, u), u(t)) - f(\varphi(t, y, u), u(t))\| \\ &\leq C_R(1 + \gamma(\|u(t)\|)) \|\varphi(t, x, u) - \varphi(t, y, u)\|. \end{aligned}$$

Using (3.8), Gronwall's lemma, we then obtain

$$\|\varphi(t, x, u) - \varphi(t, y, u)\| \leq e^{C_R(\int_0^t (1+\gamma(\|u\|))dt)} \|x - y\|,$$

and the assertion follows.  $\square$

Using this lemma, we can prove the following continuity statement.

**PROPOSITION 3.5.** *Assume (H0)–(H5) or their respective variants from Remark 3.1. Then  $V$  and  $v$  are continuous on  $\mathcal{D}_0$ .*

*Proof.* We show the continuity of  $V$ , and then the statement for  $v$  follows immediately from its definition. The proof is performed in several steps. Throughout the proof the constants  $C_R, C$ , etc., are those defined in (H0) and (H4) (respectively, (H4')).

First note that from (3.6) we have

$$(3.9) \quad V(x) \leq \frac{C}{\eta} \alpha_1(\|x\|)^\eta \quad \text{for } x \in B(0, \alpha_1^{-1} \circ \alpha_2^{-1}(r)).$$

(i) (Local boundedness of  $V$  on  $\mathcal{D}_0$ ) Pick an arbitrary  $x_0 \in \mathcal{D}_0$ , and fix  $\varepsilon > 0$ . Then there exists a  $u_0 \in \mathcal{U}$  such that  $J(x_0, u_0) \leq V(x_0) + \varepsilon$ . Since  $J(x_0, u_0)$  is finite, it follows from (H3) that there exists a time  $T_0 > 0$  such that  $\|\varphi(T_0, x_0, u_0)\| \leq \alpha_1^{-1} \circ \alpha_2^{-1}(r)/2$ . By continuity of  $\varphi$  in  $x$  we can pick a ball  $B(x_0, \delta)$  such that

$$(3.10) \quad \|\varphi(T_0, x, u_0)\| \leq \alpha_1^{-1} \circ \alpha_2^{-1}(r) \quad \forall x \in \text{cl } B(x_0, \delta).$$

We define the set

$$K = \{\varphi(t, x, u_0) \mid x \in \text{cl } B(x_0, \delta), t \in [0, T_0]\},$$

which is compact since  $\varphi$  is continuous in  $t$  and  $x$  (recall that  $u_0$  is essentially bounded). Using (3.10), we obtain from Bellman's optimality principle for all  $x \in B(x_0, \delta)$  the inequality

$$\begin{aligned} V(x) &\leq \int_0^{T_0} g(\varphi(t, x, u), u(t))dt + V(\varphi(T_0, x, u)) \\ &\leq \max_{x \in K, u \in B(0, \|u_0\|_\infty)} g(x, u)T_0 + \frac{C}{\eta} \alpha_1(r)^\eta, \end{aligned}$$

where we have used (3.9). This shows that  $\sup_{x \in B(x_0, \delta)} V(x) =: B_V$  is finite.

(ii) (Bounds on  $\varepsilon$ -optimal controls and trajectories) For any  $x \in B(x_0, \delta)$  and any  $\varepsilon \in (0, 1]$  we pick an  $\varepsilon$ -optimal control function  $u_{x, \varepsilon} \in \mathcal{U}$ , i.e.,

$$J(x, u_{x, \varepsilon}) \leq V(x) + \varepsilon.$$

We claim that for any  $\varepsilon, T > 0$  the set

$$K_\varepsilon := \{\varphi(t, x, u_{x, \varepsilon}) \mid t \geq 0, x \in B(x_0, \delta)\}$$

and the sets

$$\{\|u_{x, \varepsilon}\|_{\gamma, T} \mid x \in B(x_0, \delta)\}$$

are bounded. If the first set were unbounded, then there would be an  $x \in B(x_0, \delta)$  and  $t_1 > 0$  such that  $\|\varphi(t_1, x, u_{x,\varepsilon})\| \geq V(x) + 2\varepsilon + 2r$ . If  $t_2 > t_1$  is the first time at which  $\|\varphi(t_2, x, u_{x,\varepsilon})\| = 2r$  again, then we obtain using Lemma 3.2 that

$$\begin{aligned} J(x, u_{x,\varepsilon}) &\geq \int_{t_1}^{t_2} g(\varphi(t, x, u_{x,\varepsilon}), u_{x,\varepsilon}(t)) dt \\ &\geq \|\varphi(t_1, x, u_{x,\varepsilon}) - \varphi(t_2, x, u_{x,\varepsilon})\| \geq V(x) + 2\varepsilon, \end{aligned}$$

a contradiction.

On the other hand, if  $\{\|u_{x,\varepsilon}\|_{\gamma,T} \mid x \in B(x_0, \delta)\}$  is unbounded for a given  $T > 0$ , then there have to be  $x, u_{x,\varepsilon}$  such that  $\|u_{x,\varepsilon}\|_{\gamma,T} \geq V(x) + 2\varepsilon + T\gamma(2\bar{u})$ . This implies that if we integrate over the (measurable) set

$$E := \{t \in [0, T] \mid \|u_{x,\varepsilon}(t)\| \geq 2\bar{u}\},$$

then we obtain

$$\int_E \gamma(\|u_{x,\varepsilon}(t)\|) dt \geq V(x) + 2\varepsilon,$$

as the contribution of the integral over  $[0, T] \setminus E$  to  $\|u_{x,\varepsilon}\|_{\gamma,T}$  can be at most  $T\gamma(2\bar{u})$ . By using an estimate over the set  $E$  and again Lemma 3.2, we obtain again a contradiction to  $J(x, u_{x,\varepsilon}) \leq V(x) + \varepsilon$ .

(iii) (Continuity of trajectories) We denote by  $R_\varepsilon$  an upper bound on the set  $K_\varepsilon$ . By Lemma 3.4 we can conclude that for  $x, y \in B(x_0, \delta)$  and all  $t \geq 0$  such that

$$\|x - y\| \leq R_\varepsilon \exp(-C_{2R_\varepsilon}(\|u_{x,\varepsilon}\|_{\gamma,t} + t))$$

we have

$$(3.11) \quad \|\varphi(t, x, u_{x,\varepsilon}) - \varphi(t, y, u_{x,\varepsilon})\| \leq \exp(C_{2R_\varepsilon}(\|u_{x,\varepsilon}\|_{\gamma,t} + t)) \|x - y\|.$$

(iv) (Continuity of  $V$ ) We show the continuity of  $V$  on  $B(x_0, \delta)$ . Since  $x_0 \in \mathcal{D}_0$  was arbitrary, this proves the proposition. So pick  $\varepsilon > 0$ , and assume without loss of generality that  $\varepsilon < \alpha_2^{-1}(r)C$ .

From the lower bound  $g_c$  on  $g$  in (H3) and the boundedness of  $J(x, u_{x,\varepsilon})$  on  $B(x_0, \delta)$ , it follows that for any  $\rho > 0$  there is a time  $T_\rho$  such that for  $x \in B(x_0, \delta)$  we have  $\varphi(t, x, u_{x,\varepsilon}) \in B(0, \rho)$  for some  $t \leq T_\rho$ . Using (3.9), we may thus assume that the controls  $u_{x,\varepsilon}$  are chosen in such a way that there exists  $T_\varepsilon > 0$  (depending on  $B_V$ ) such that for all  $t \geq T_\varepsilon, x \in B(0, \delta)$  we have

$$\varphi(t, x, u_{x,\varepsilon}) \in B(0, \alpha_1^{-1}(\varepsilon/C)/2) \subset B(0, \alpha_1^{-1} \circ \alpha_2^{-1}(r)/2).$$

Denote

$$m := \exp\left(-C_{2R_\varepsilon}\left(\max_{z \in B(x_0, \delta)} \|u_{z,\varepsilon}\|_{\gamma,T_\varepsilon} + T_\varepsilon\right)\right),$$

and note that the right-hand side is finite by (ii). Choose two points  $x, y \in B(x_0, \delta)$  such that

$$\|x - y\| \leq R_\varepsilon m.$$

Without loss of generality, assume  $V(y) \geq V(x)$ . Abbreviating  $u := u_{x,\varepsilon}$ ,  $T := T_\varepsilon$ , we obtain

$$\begin{aligned} |V(y) - V(x)| &= V(y) - V(x) \\ &\leq V(y) - \int_0^\infty g(\varphi(t, x, u), u(t)) dt + \varepsilon \\ &\leq \int_0^T |g(\varphi(t, y, u), u(t)) - g(\varphi(t, x, u), u(t))| dt + V(\varphi(T, y, u)) + \varepsilon; \end{aligned}$$

using the Lipschitz condition in (H3) and (3.11) we continue

$$\leq \int_0^T C_{2R_\varepsilon} (1 + \gamma(\|u(t)\|)) \quad m \|x - y\| dt + V(\varphi(T, y, u)) + \varepsilon,$$

and we obtain

$$\leq C_{2R_\varepsilon} (T + \|u\|_{\gamma,T}) m \|x - y\| + 2\varepsilon,$$

provided  $\|y - x\| \leq \alpha_1^{-1}(\eta\varepsilon^{1/\eta}/C)/(2m)$ , because in this case we obtain from (3.11) that  $\varphi(T, y, u) \in B(0, \alpha_1^{-1}(\eta\varepsilon^{1/\eta}/C))$ , and thus from (3.9)

$$V(\varphi(T, y, u)) \leq \frac{C}{\eta} \alpha_1(\|\varphi(T, y, u)\|)^\eta \leq \varepsilon.$$

Thus for any  $\varepsilon \in (0, 1]$  and any  $x \in B(x_0, \delta)$  we can find  $\delta_\varepsilon > 0$  such that  $|V(y) - V(x)| \leq 3\varepsilon$  for all  $x, y \in B(x_0, \delta)$ , with  $\|x - y\| \leq \delta_\varepsilon$ . This implies continuity of  $V$  in  $B(x_0, \delta)$  and, since  $x_0 \in \mathcal{D}_0$  was arbitrary, continuity on the whole set  $\mathcal{D}_0$ .  $\square$

The next proposition makes a statement of the behavior of  $V(x)$  near the boundary of  $\mathcal{D}_0$  or at  $\infty$ .

**PROPOSITION 3.6.** *Assume (H0)–(H5) or their respective variants from Remark 3.1. Then for any sequence  $x_k$  which satisfies  $\text{dist}(x_k, \partial\mathcal{D}_0) \rightarrow 0$  or  $\|x_k\| \rightarrow \infty$  we have  $V(x_k) \rightarrow \infty$  and  $v(x_k) \rightarrow 1$ . In particular,  $v$  is continuous on  $\mathbb{R}^n$ .*

*Proof.* If  $\|x_k\| \rightarrow \infty$ , then we have for every  $k$  either that  $x_k \notin \mathcal{D}_0$ , in which case  $V(x_k) = \infty$ , or  $x_k \in \mathcal{D}_0$ . In the latter case we have by Lemma 3.2 that  $V(x_k) \geq \|x_k\| - 2r$  for all  $k$  large enough. This shows the assertion for  $V$ , and the conclusion for  $v$  is immediate from the definition.

To prove the assertion for  $\text{dist}(x_k, \partial\mathcal{D}_0) \rightarrow 0$ , we may now assume that there exist a sequence  $x_k \rightarrow x_0 \in \partial\mathcal{D}_0$  and some  $C > 0$  such that  $V(x_k) \leq C$  holds for all  $k \in \mathbb{N}$ . Pick  $\varepsilon > 0$ , and for each  $k$  choose a control function  $u_k \in \mathcal{U}$  such that we have

$$J(x_k, u_k) \leq V(x_k) + \varepsilon \leq C + \varepsilon.$$

Following step (ii) of the proof of Proposition 3.5, we obtain that  $\{\varphi(t, x_k, u_k) \mid t \geq 0, k \in \mathbb{N}\}$  is bounded and that  $\|u_k\|_{\gamma,t}$  is uniformly bounded in  $k$  for all  $t \geq 0$ . Then we may apply (3.11) as in step (iv) of the proof of Proposition 3.5 to conclude that for every  $t \geq 0$  and every  $\delta > 0$  there is a  $k_0$  such that  $\|\varphi(t, x_k, u_k) - \varphi(t, x_0, u_k)\| < \delta$  for all  $k \geq k_0$ .

Because of the lower bound on  $g$  in (H3) we may assume that there exists  $T > 0$  (independent of  $k$ ) such that

$$\varphi(t, x_k, u_k) \in B(0, r/2) \quad \forall t \geq T, k \in \mathbb{N}.$$

This implies  $\varphi(T, x_0, u_k) \in B(0, r/2)$  for all sufficiently large  $k \in \mathbb{N}$ , which in turn implies  $x_0 \in \mathcal{D}_0$ . This contradicts  $x_0 \in \partial\mathcal{D}_0$  because  $\mathcal{D}_0$  is open.  $\square$

**4. Characterizations of  $V$  and  $v$  by Zubov's method.** The aim of this section is to characterize the functions  $V$  and  $v$  introduced in (3.2) and (3.3) as the (unique) viscosity solutions of the equations

$$(4.1) \quad \sup_{u \in U} \{-DV(x)f(x, u) - g(x, u)\} = 0$$

and

$$(4.2) \quad \sup_{u \in U} \{-Dv(x)f(x, u) - (1 - v(x))g(x, u)\} = 0,$$

respectively (for the definition of viscosity solution we refer to [6, 3]).

Recalling that  $V$  is locally bounded in  $\mathcal{D}_0$  and  $v$  is bounded in  $\mathbb{R}^n$ , our first result follows from a standard application of the dynamic programming principles (3.4) and (3.5); see [3].

**PROPOSITION 4.1.** *Assume (H0)–(H5) or their respective variants from Remark 3.1. Then the functions  $V$  and  $v$  defined in (3.2) and (3.3) are viscosity solutions of (4.1) in  $\mathcal{D}_0$  and of (4.2) in  $\mathbb{R}^n$ , respectively.*

**Remark 4.2.** Note that it follows from these characterizations that  $v$  is a CLF on  $\mathcal{D}_0$  in the usual sense [34]. In fact, a small calculation shows that  $v$  is a viscosity supersolution on  $\mathcal{D}_0$  of

$$\inf_{u \in U} Dv(x)f(x, u) \leq -W(x)g_{\|x\|},$$

where  $0 < W(x) < 1 - v(x)$  for  $x \in \mathcal{D}_0 \setminus \{0\}$  and  $g_{\|x\|}$  denotes the constant from (H3) for  $c = \|x\|$ .

The main result in this section will be a uniqueness statement for (4.1) and (4.2), showing that the above functions are the unique viscosity solutions of these equations.

In order to obtain such a result we make use of the so-called optimality principles developed by Soravia [35, 36]. For the application of the results from these references, we need our system to be defined by a bounded vector field  $f$ . To this end we introduce a standard tool for unbounded control systems which consists in rescaling the coefficients of the equations (see [5, 35])

$$(4.3) \quad \begin{aligned} \tilde{f}(x, u) &= \frac{f(x, u)}{1 + \|f(x, u)\|}, \\ \tilde{g}(x, u) &= \frac{g(x, u)}{1 + \|f(x, u)\|}. \end{aligned}$$

The following summarizes the main properties of the rescaled functions.

**PROPOSITION 4.3.** *Assume (H0)–(H3) and (H5) or their respective variants from Remark 3.1. Then  $\tilde{f}$  and  $\tilde{g}$  satisfy (H0)–(H3) for suitably adjusted  $\mathcal{K}_\infty$  and  $\mathcal{KL}$  functions, and the optimal value functions  $V$  and  $v$  of the original and the rescaled problems coincide.*

*Proof.* First note that (H0), (H1), and the first part of (H3) follow by straightforward computations. In order to prove the second part of (H3) we fix an arbitrary  $c > 0$  and show that

$$\tilde{g}_c := \inf \{\tilde{g}(x, u) \mid \|x\| \geq c, u \in U\}$$

is positive. To this end we pick arbitrary  $x \in \mathbb{R}^n$ ,  $u \in U$ , with  $\|x\| \geq c$ , and distinguish three cases.

*Case 1.*  $\|f(x, u)\| \leq 1$ : In this case from (H3) we get  $\tilde{g}(x, u) \geq g_c/2$ .

*Case 2.*  $\|f(x, u)\| > 1$  and  $(x, u) \in B(0, 2r) \times B(0, 2\bar{u})$ : In this case from (H3) we get  $\tilde{g}(x, u) \geq g_c/(1 + \bar{f})$ , with  $\bar{f} := \max\{\|f(x, u)\| \mid (x, u) \in B(0, 2r) \times B(0, 2\bar{u})\} < \infty$ .

*Case 3.*  $\|f(x, u)\| > 1$  and  $(x, u) \notin B(0, 2r) \times B(0, 2\bar{u})$ : In this case (H5) implies  $\tilde{g}(x, u) \geq \|f(x, u)\|/(1 + \|f(x, u)\|) \geq 1/2$ .

By combining the three cases, we obtain

$$\tilde{g}_c \geq \min\{g_c/2, g_c/(1 + \bar{f}), 1/2\} > 0,$$

which shows the second part of (H3).

In order to show that the optimal value functions coincide, observe that the introduction of the vector field  $\tilde{f}$  and the running cost  $\tilde{g}$  amounts to nothing more than a rescaling of time that does not change trajectories or values associated to a particular control. To see this, let  $x \in \mathbb{R}^n$ ,  $u \in U$  be given. Now introduce a new time variable  $\tau$  through the differential equation

$$\frac{dt(\tau)}{d\tau} = \frac{1}{1 + \|f(\phi(t(\tau), x, u), u(t(\tau)))\|} \quad \text{a.e.},$$

and a control  $\tilde{u}(\tau) := u(t(\tau))$  a.e. Then the function  $\psi(\tau) := \phi(t(\tau), x, u)$  satisfies the differential equation

$$\frac{d\psi(\tau)}{d\tau} = \frac{f(\phi(t(\tau), x, u), u(t(\tau)))}{1 + \|f(\phi(t(\tau), x, u), u(t(\tau)))\|} = \tilde{f}(\psi(\tau), \tilde{u}(\tau)).$$

So if we consider the system

$$(4.4) \quad \dot{x}(t) = \tilde{f}(x(t), u(t)),$$

then using standard transformation of integral formulas, it is also easy to see that if  $T(x, u) = \infty$ , then  $\tilde{J}(x, \tilde{u}) = J(x, u)$ , where  $\tilde{J}$  defines the value along a rescaled trajectory using the running cost  $\tilde{g}$  in (3.1). If the solution explodes, i.e.,  $T(x, u) < \infty$ , then we have so far simply defined the value to be infinity. However, since (H3) holds for  $\tilde{g}$ , the associated integral of the transformed system also diverges because it will never enter  $B(0, r)$ . Thus, the optimal value functions coincide.  $\square$

Note that we do not need (H4) and (H5) for the rescaled problem in order to establish the previous result: (H4) is needed in order to ensure finiteness of  $V$  on  $\mathcal{D}_0$ , while (H5) is needed in order to establish the continuity of  $V$ . Both properties readily carry over to the rescaled problem via the integral transformations.

In order to prove our uniqueness statement we need one final assumption.

(H6) The rescaled function  $\tilde{g}$  satisfies  $\tilde{g}(x, u) \rightarrow \infty$  as  $\|u\| \rightarrow \infty$  for each  $x \in \mathbb{R}^n$ .

To Zubov's equations (4.1) and (4.2) we associate the Hamiltonians

$$H_V : \mathbb{R}^n \times (\mathbb{R}^n)^* \rightarrow \mathbb{R}, \quad H_V(x, p) = \sup_{u \in U} \{-f(x, u)p - g(x, u)\},$$

and

$$H_v : \mathbb{R}^n \times \mathbb{R} \times (\mathbb{R}^n)^* \rightarrow \mathbb{R}, \quad H_v(x, r, p) = \sup_{u \in U} \{-f(x, u)p - (1 - r)g(x, u)\}.$$



From (H5) we obtain that the supremum in these Hamiltonians is attained in a compact subset of  $U$  for  $r < 1$  in the case of  $H_v$ . This implies that the Hamiltonians  $H_V$  and  $H_v$  are locally Lipschitz continuous with respect to their arguments, again for  $r < 1$  in the case of  $H_v$ .

The following Theorem 4.4 and its Corollary 4.5 are the main results of this paper.

**THEOREM 4.4.** *Assume that  $f$  and  $g$  satisfy the assumptions (H0)–(H6) (or their respective variants from Remark 3.1). Then*

- (i) *the function  $v$  from (3.3) is the unique bounded viscosity solution of (4.2) with  $v(0) = 0$ ;*
- (ii) *there exists a unique pair  $(\mathcal{O}, V)$  such that  $\mathcal{O}$  is an open set containing the origin and  $V$  is a locally bounded, nonnegative continuous viscosity solution of (4.1) in  $\mathcal{O}$ , with  $V(0) = 0$  and  $V(x) \rightarrow +\infty$  for  $x \rightarrow \partial\mathcal{O}$  (here  $V$  is the function from (3.2));*
- (iii) *the functions  $v$  and  $V$  characterize the domain of asymptotic controllability via*

$$\mathcal{D}_0 = \{x \in \mathbb{R}^n \mid v(x) < 1\} = \{x \in \mathbb{R}^n \mid V(x) < \infty\};$$

- (iv) *the functions  $v$  and  $V$  satisfy  $v(x_k) \rightarrow 1$  and  $V(x_k) \rightarrow \infty$  for all sequences with  $x_k \rightarrow \partial\mathcal{D}_0$  or  $\|x_k\| \rightarrow \infty$ .*

Before turning to the proof we state the following corollary, whose proof in particular shows how a cost function  $g$  meeting the assumptions of Theorem 4.4 can be constructed.

**COROLLARY 4.5.** *Assume that  $f$  satisfies the conditions (H0)–(H2). Then there exists a continuous function  $v : \mathbb{R}^n \rightarrow [0, 1]$  which is a CLF (in the usual sense; cf. Remark 4.2) on the domain of asymptotic controllability  $\mathcal{D}_0$  and constant equal to 1 on  $\mathbb{R}^n \setminus \mathcal{D}_0$ . Furthermore, this  $v$  is the unique bounded viscosity solution of (4.2), with  $v(0) = 0$  for some suitable  $g : \mathbb{R}^n \times U \rightarrow \mathbb{R}$ .*

*Proof.* In order to prove the theorem it is sufficient to construct a function  $g$  satisfying (H4)–(H6). Then the Lyapunov function property follows immediately from Theorem 4.4 and Remark 4.2.

To this end, consider the Lipschitz function  $\rho : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$  given by

$$\rho(s) := \begin{cases} 0, & s \in [0, 1], \\ s - 1, & s \in [1, 2], \\ 1, & s \in [2, \infty), \end{cases}$$

and set

$$\bar{g}(x, u) := \alpha_2^{-1}(\|x\|) + \rho(s(x, u)) [1 + \|f(x, u)\|] [1 + \|u\|],$$

with

$$s(x, u) = \frac{1}{2} \sqrt{\frac{\|x\|^2}{r^2} + \frac{\|u\|^2}{\bar{u}^2}}.$$

For this function (H3), (H4), and (H6) follow immediately from the construction. We obtain the desired  $g$  by modifying  $\bar{g}$  as follows: Let  $\gamma \in \mathcal{K}_\infty$  such that (H0) and (H3) are satisfied, and set

$$g(x, u) := \bar{g}(x, u) + \rho(s(x, u))\gamma(\|u\|).$$

Then (H5) is satisfied, while straightforward computations show that (H3), (H4), and (H6) carry over from  $\tilde{g}$  to  $g$ .  $\square$

In the proof of Theorem 4.4 we encounter two difficulties: the unbounded dependence of the functions on the control variable and the vanishing of the cost  $g$  at the origin.

To solve the first problem we use the rescaled functions from above. Associated to these functions we introduce two rescaled equations which share with (4.1) and (4.2) the same set of sub- and supersolutions.

LEMMA 4.6. *Assume (H0) and (H3), and consider the equations*

$$(4.5) \quad \sup_{u \in U} \{-D\tilde{V}(x)\tilde{f}(x, u) - \tilde{g}(x, u)\} = 0$$

and

$$(4.6) \quad \sup_{u \in U} \{-D\tilde{v}(x)\tilde{f}(x, u) - (1 - \tilde{v}(x))\tilde{g}(x, u)\} = 0.$$

Then

- (i) *any viscosity subsolution of (4.1) is a viscosity subsolution for (4.5) and vice versa;*
- (ii) *any viscosity supersolution of (4.1) is a viscosity supersolution for (4.5), and, if in addition (H6) holds, then any viscosity supersolution of (4.5) is also a viscosity supersolution for (4.1).*

*The same assertions hold for (4.6) and (4.2).*

The proof of Lemma 4.6 is postponed to the appendix. The following corollary is a simple consequence of the previous lemma.

COROLLARY 4.7. *Assume (H0), (H3), and (H6). Then*

- (i) *any viscosity solution of (4.1) is a viscosity solution of (4.5) in  $\mathcal{D}_0$  and vice versa;*
- (ii) *any viscosity solution of (4.2) is a viscosity solution of (4.6) and vice versa.*

Even if the coefficients of the rescaled equations have a better dependence on the variable  $u$ , there is still the problem of the vanishing of  $\tilde{g}$  at the origin. In order to prove a uniqueness result for (4.5) and (4.6), we use a control theoretic argument and some optimality principles introduced in [35, 36], as stated in the following lemma.

LEMMA 4.8. *Assume (H0), (H3), and (H5), and let  $\tilde{\varphi}(t, x, u)$  be the solution of (4.4). Define*

$$\tilde{G}(x, t, u) := \exp \left( - \int_0^t \tilde{g}(\tilde{\varphi}(\tau, x, u), u(\tau)) d\tau \right).$$

*Then the following properties hold:*

- (i) *Any upper semicontinuous viscosity subsolution  $w^-$  of (4.6) satisfies*

$$(4.7) \quad w^-(x) \leq \inf_{u \in \mathcal{U}} \inf_{t \in [0, T]} \left\{ 1 + \tilde{G}(x, t, u)(w^-(\tilde{\varphi}(t, x, u)) - 1) \right\}$$

*for each  $T > 0$ .*

- (ii) *Consider a continuous viscosity supersolution  $w^+$  of (4.6), and let  $\Omega \subset \mathbb{R}^n$  be an open and bounded set with  $\sup_{x \in \Omega} w^+(x) < 1$ . Consider the first exit time from  $\Omega$  given by*

$$T_{ex}(x, u, \Omega) = \min\{t \geq 0 \mid \varphi(t, x_0, u) \notin \Omega\}.$$

Then  $w^+$  satisfies

$$(4.8) \quad w^+(x) \geq \inf_{u \in \mathcal{U}} \sup_{t \in [0, T_{ex}(x, u, \Omega)]} \left\{ 1 + \tilde{G}(x, t, u)(w^+(\tilde{\varphi}(t, x, u)) - 1) \right\}.$$

*Proof.* Let  $\Omega \subset \mathbb{R}^n$  be an open and bounded set, and let  $\tilde{U}$  be a compact subset of  $U$  with the corresponding space of measurable control functions denoted by  $\tilde{\mathcal{U}}$ . If  $w^-$  is an upper semicontinuous viscosity subsolution of (4.6) in  $\mathbb{R}^n$ , then the restriction of  $w^-$  to  $\Omega$  is also a subsolution of (4.6) on  $\Omega$  with  $\tilde{U}$  instead of  $U$ . For the restricted control value set  $\tilde{U}$ , (4.6) is continuous, and furthermore  $\tilde{f}, \tilde{g}$  are uniformly Lipschitz on  $\Omega$ . Hence we can apply [36, Theorem 3.2(i)], which for each  $u \in \tilde{\mathcal{U}}$  yields

$$w^-(x) \leq \inf_{t \in [0, T_{ex}(x, u, \Omega)]} \left\{ 1 + \tilde{G}(x, t, u)(w^-(\tilde{\varphi}(t, x, u)) - 1) \right\},$$

where  $T_{ex}(x, u, \Omega)$  is the first exit time of  $\tilde{\varphi}(t, x_0, u)$  from the set  $\Omega$  defined in (ii).

Since  $\tilde{f}$  is globally bounded, for any  $x \in \mathbb{R}^n$  and any  $T > 0$  we may find an open and bounded set  $\Omega_{x,T} \subset \mathbb{R}^n$  such that  $T_{ex}(x, u, \Omega_{x,T}) \geq T$  for each  $u \in \mathcal{U}$ . Since each  $u \in \mathcal{U}$  is essentially locally bounded, it lies in  $\tilde{U}$  for an appropriate choice of  $\tilde{U}$ , which shows (i).

The proof of (ii) follows from [36, Theorem 3.2(ii)], observing that (4.6) is continuous on  $\Omega$  since  $w^-(x) < 1$ ; hence, here we do not need to restrict the control value set  $U$ .  $\square$

*Remark 4.9.* Note that the asymmetry of the statements (i) and (ii) is due to the fact that we imposed different conditions in order to obtain continuity of (4.6), which is needed for the application of [36, Theorem 3.2]. In (i) we restrict the set of control values  $U$ , obtaining a result for arbitrary  $\Omega$  (thus for arbitrary  $T$ ) and for upper semicontinuous functions. In (ii) this restriction is not possible because the supersolution property will not persist passing from  $U$  to  $\tilde{U}$ . Thus here we ensure the continuity of (4.6) by considering suitable subsets  $\Omega$  of the state space.

Using these inequalities, we can now prove the following uniqueness results.

LEMMA 4.10. Assume (H0)–(H6), and consider the functions  $V$  and  $v$  defined by (3.2) and (3.3). Then

- (i)  $v$  is the unique bounded continuous viscosity solution of (4.6), with  $v(0) = 0$ ;
- (ii)  $(\mathcal{D}_0, V)$  is the unique pair of an open set containing the origin and a locally bounded, nonnegative continuous viscosity solution of (4.5) in the open set such that  $V(0) = 0$  and  $V(x) \rightarrow +\infty$  for  $x \rightarrow \partial\tilde{\mathcal{O}}$ .

*Proof.* We prove only (i), since the proof of assertion (ii) is similar. Note that by Proposition 4.3 the functions  $v$  and  $V$  can be taken to be defined through (4.4) and the running cost  $\tilde{g}$ . In the following we work with this representation. Again by  $\tilde{\varphi}(t, x, u)$  we denote the solutions of (4.4).

*Claim 1.* If  $w^-$  is a bounded continuous subsolution of (4.6) on  $\mathbb{R}^n$  with  $w^-(0) \leq 0$ , then  $w^- \leq v$ .

By the upper semicontinuity of  $w^-$  and  $w^-(0) \leq 0$  we obtain that for every  $\varepsilon > 0$  there exists a  $\delta > 0$  with  $w^-(x) \leq \varepsilon$  for all  $x \in \mathbb{R}^n$ , with  $\|x\| \leq \delta$ . Now we distinguish two cases:

(i)  $x_0 \in \mathcal{D}_0$ : We choose  $u^* \in \mathcal{U}$  such that  $v(x_0) + \varepsilon > \tilde{J}(x_0, u^*) = 1 - \tilde{G}(x_0, \infty, u^*)$ . Then (H3) (cf. Proposition 4.3) implies that there exists a sequence  $t_k \rightarrow \infty$  such that

$\tilde{\varphi}(t_k, x_0, u^*) \rightarrow 0$  as  $k \rightarrow \infty$ . Thus it follows from the lower optimality principle (4.7) and the definition of  $v$  that

$$\begin{aligned} w^-(x_0) &\leq \limsup_{k \rightarrow \infty} 1 + \tilde{G}(x_0, t_k, u^*)(w^-(\tilde{\varphi}(t_k, x_0, u^*)) - 1) \\ &\leq 1 + \tilde{G}(x_0, \infty, u^*)(\varepsilon - 1) \leq v(x_0) + 2\varepsilon, \end{aligned}$$

which shows the claim as  $\varepsilon > 0$  was arbitrary.

(ii)  $x_0 \notin \mathcal{D}_0$ : In this case by Proposition 3.3 it is sufficient to show that  $w^-(x_0) \leq 1$ . Let  $M$  be a bound on  $|w^-|$ . In this case we have  $\tilde{\varphi}(\tau, x_0, u) \notin B(0, r)$  for all  $\tilde{u} \in \mathcal{U}$  and all  $\tau \geq 0$ , which implies

$$\int_0^\tau \tilde{g}(\tilde{\varphi}(s, x_0, u), \tilde{u}(s)) ds \geq \tilde{g}_r \tau$$

for the constant  $\tilde{g}_r > 0$  from (H3); see Proposition 4.3. Therefore  $\tilde{G}(x_0, \tau, \tilde{u}) \leq \exp(-\tilde{g}_r \tau)$  for all  $\tau \geq 0, \tilde{u} \in \mathcal{U}$ . Hence

$$1 + \tilde{G}(x_0, \tau, \tilde{u})(w^-(\tilde{\varphi}(\tau, x_0, \tilde{u})) - 1) \leq 1 + \exp(-\tilde{g}_r \tau)(M + 1)$$

for all  $\tilde{u} \in \mathcal{U}$ , and the result follows by (4.7) as the right-hand side tends to 1 for  $\tau \rightarrow \infty$ .

Therefore Claim 1 is proved. To conclude the proof we now consider the following.

*Claim 2.* Let  $w^+$  be a bounded continuous supersolution of (4.2) on  $\mathbb{R}^n$  with  $w^+(0) \geq 0$ . Then  $w^+ \geq v$ .

Again we distinguish two cases.

(i)  $x_0 \notin \mathcal{D}_0$ : In this case we know  $v(x_0) = 1$ , and it is sufficient to show  $w^+(x_0) \geq 1$ . In order to prove this inequality by contradiction, we assume  $w^+(x_0) = 1 - \delta$  for some suitable  $\delta > 0$ . Since  $\|\tilde{f}\|$  is bounded by 1 for any ball  $B(R, x_0)$ , we have  $T_{ex}(x_0, u, B(R, x_0)) \geq R$  for all  $u \in \mathcal{U}$ . Furthermore,  $x_0 \notin \mathcal{D}_0$  implies  $\tilde{\varphi}(t, x_0, u) \notin B(r, 0)$  for all  $t \geq 0, u \in \mathcal{U}$ ; hence, by (H3) (cf. Proposition 4.3) we have the inequality  $\tilde{g}(\tilde{\varphi}(t, x_0, u), u(t)) \geq \tilde{g}_r > 0$  for all  $t \geq 0, u \in \mathcal{U}$ . This implies the existence of  $R > 0$  such that  $\tilde{G}(x_0, t, u) \leq \delta/(2(M + 1))$  for all  $t \geq R$  and all  $u \in \mathcal{U}$ , where  $M > 0$  is a bound on  $|w^+|$ . Now pick the set  $\Omega = \{x \in \mathbb{R}^n \mid w^+(x) < 1 - \delta/2\} \cap B(R, x_0)$ . For this set Lemma 4.8(ii) is applicable, and thus by (4.8), for any  $\varepsilon > 0$  we can find  $u_\varepsilon$  with

$$\begin{aligned} w^+(x_0) &\geq \sup_{\tau \in [0, T_{ex}(x_0, u_\varepsilon, \Omega)]} \{1 + \tilde{G}(x_0, \tau, u_\varepsilon)(w^+(\tilde{\varphi}(\tau, x_0, u_\varepsilon)) - 1)\} - \varepsilon \\ &\geq \sup_{\tau \in [0, T_{ex}(x_0, u_\varepsilon, \Omega)]} \{[1 - \exp(-\tau g_r)] - \exp(-\tau g_r)M\} - \varepsilon. \end{aligned}$$

If  $T_{ex}(x_0, u_\varepsilon, \Omega) = \infty$ , then this expression equals  $1 - \varepsilon$ , and hence we obtain a contradiction to  $w^+(x_0) = 1 - \delta < 1$  for  $\varepsilon$  sufficiently small. If  $T_{ex}(x_0, u_\varepsilon, \Omega)$  is finite, then either we have  $\tilde{\varphi}(T_{ex}(x_0, u_\varepsilon, \Omega)) \notin B(R, x_0)$ , in which case we get  $T_{ex}(x_0, u_\varepsilon, \Omega) \geq R$  and thus  $\tilde{G}(x_0, T_{ex}(x_0, u_\varepsilon, \Omega), u_\varepsilon) \leq \delta/(2(M + 1))$ , or we have  $\tilde{\varphi}(T_{ex}(x_0, u_\varepsilon, \Omega)) \in B(R, x_0)$ , which by construction of  $\Omega$  implies  $w^+(\tilde{\varphi}(T_{ex}(x_0, u_\varepsilon, \Omega), x_0, u_\varepsilon)) \geq 1 - \delta/2$ . Since  $\tilde{G}(x_0, t, u) \leq 1$  holds for all  $t \geq 0, u \in \mathcal{U}$ , in both cases we get

$$\begin{aligned} w^+(x_0) &\geq \sup_{\tau \in [0, T_{ex}(x_0, u_\varepsilon, \Omega)]} \{1 + \tilde{G}(x_0, \tau, u_\varepsilon)(w^+(\tilde{\varphi}(\tau, x_0, u_\varepsilon)) - 1)\} - \varepsilon \\ &\geq 1 + \tilde{G}(x_0, T_{ex}(x_0, u_\varepsilon, \Omega), u_\varepsilon)(w^+(\tilde{\varphi}(T_{ex}(x_0, u_\varepsilon, \Omega), x_0, u_\varepsilon)) - 1) - \varepsilon \\ &\geq 1 - \delta/2 - \varepsilon. \end{aligned}$$

This again contradicts our assumption that  $w^+(x_0) = 1 - \delta$  for  $\varepsilon$  sufficiently small.

(ii)  $x_0 \in \mathcal{D}_0$ : In this case we know that  $v(x_0) < 1$ , and hence for  $w^+(x_0) \geq 1$  there is nothing to show. Thus we can assume  $w^+(x_0) = 1 - \delta$  for some suitable  $\delta > 0$  and again consider the set  $\Omega = \{x \in \mathbb{R}^n \mid w^+(x) < 1 - \delta/2\} \cap B(R, x_0)$  from part (i), above. Now fix  $\varepsilon > 0$ , with  $\varepsilon < \delta/2$  implying

$$(4.9) \quad w^+(x_0) + \varepsilon < 1 - \delta/2.$$

Then (4.8) yields the existence of a control function  $u_\varepsilon \in \mathcal{U}$  with

$$(4.10) \quad w^+(x_0) + \varepsilon \geq \sup_{t \in [0, T_{ex}(x_0, u_\varepsilon, \Omega)]} \{1 + \tilde{G}(x_0, t, u_\varepsilon)(w^+(\tilde{\varphi}(t, x_0, u_\varepsilon)) - 1)\}.$$

If  $T_{ex}(x_0, u_\varepsilon, \Omega) < \infty$ , then as in part (i) above, we obtain from (4.9) and (4.10)

$$\begin{aligned} 1 - \delta/2 &> w^+(x_0) + \varepsilon \\ &\geq 1 + \tilde{G}(x_0, T_{ex}(x_0, \tilde{u}, \Omega), \tilde{u})(w^+(\tilde{\varphi}(T_{ex}(x_0, \tilde{u}, \Omega), x_0, \tilde{u})) - 1) \\ &\geq 1 - \delta/2, \end{aligned}$$

which is a contradiction. Thus we obtain  $T_{ex}(x_0, u_\varepsilon, \Omega) = \infty$ .

Now for each  $\eta > 0$  we find  $t$  such that  $\|\tilde{\varphi}(t, x_0, u_\varepsilon)\| \leq \eta$ , because otherwise—as in the first inequality of case (i), above—the right-hand side in (4.10) would be equal to 1, contradicting (4.9). The continuity of  $w^+$  and the assumption  $w^+(0) \geq 0$  imply that there exists a  $\eta_1 > 0$  such that

$$(4.11) \quad w^+(x) \geq -\varepsilon \quad \forall \|x\| \leq \eta_1.$$

On the other hand, since  $v(0) = 0$  and  $v$  is continuous, we find  $\eta_2 > 0$  such that

$$(4.12) \quad v(x) \leq \varepsilon \quad \forall \|x\| \leq \eta_2.$$

Combining these results, we can conclude that for all sufficiently large times  $t > 0$  we have

$$\tilde{w}^+(\tilde{\varphi}(t, x_0, u_\varepsilon)) \geq v(\tilde{\varphi}(t, x_0, u_\varepsilon)) - 2\varepsilon.$$

Thus using (4.10), (3.5), and the inequality  $\tilde{G}(x_0, t_n, u_n) \leq 1$  for sufficiently large  $t > 0$ , we can conclude

$$\begin{aligned} w^+(x_0) &\geq 1 + \tilde{G}(x_0, t, u_\varepsilon)(w^+(\tilde{\varphi}(t, x_0, u_\varepsilon)) - 1) - \varepsilon \\ &\geq 1 + \tilde{G}(x_0, t, u_\varepsilon)(v(\tilde{\varphi}(t, x_0, u_\varepsilon)) - 1) - 3\varepsilon \\ &\geq v(x_0) - 3\varepsilon, \end{aligned}$$

which shows Claim 2, as  $\varepsilon > 0$  is arbitrary.

Finally, since every viscosity solution  $\tilde{w}$  is both a subsolution and a supersolution, the combination of Claims 1 and 2 proves the lemma.  $\square$

*Proof of Theorem 4.4.* All properties follow from the fact that by Lemma 4.10 the functions  $V$  and  $v$  defined by (3.2) and (3.3) are the unique continuous viscosity solutions for (4.6) and (4.5), respectively.

(i) and (ii): By Corollary 4.7 all viscosity solutions to (4.6) and (4.5) are also viscosity solutions of (4.2) and (4.1), respectively, and vice versa. Hence,  $v$  and  $V$  are also the unique viscosity solutions of (4.2) and (4.1), respectively.

(iii): It follows from Proposition 3.3.

(iv): It follows from Proposition 3.6.  $\square$

**5. Approximation with bounded control values.** In this section we consider the bounded approximations  $U_k = U \cap \text{cl } B(0, k)$  of the (possibly) unbounded set  $U$  of control values and the corresponding set  $\mathcal{U}_k := L^\infty([0, \infty), U_k)$  of control functions. Throughout this section we assume that (H0)–(H2) hold, which implies that we can find  $g$ , meeting (H3)–(H6).

PROPOSITION 5.1. *Consider the functions*

$$V_k(x) = \inf_{u \in \mathcal{U}_k} J(x, u) \quad \text{and} \quad v_k(x) = 1 - e^{V_k(x)}.$$

*Then the relations*

$$V(x) = \inf_{k \in \mathbb{N}} V_k(x) \quad \text{and} \quad v(x) = \inf_{k \in \mathbb{N}} v_k(x)$$

*hold.*

*Proof.* Since  $\mathcal{U}_k \subseteq \mathcal{U}$  we obviously have the inequality  $V_k(x) \geq V(x)$ . Now let  $x \in \mathcal{D}_0$  and  $u \in \mathcal{U}$  be such that

$$J(x, u) \leq V(x) + \varepsilon$$

for some  $\varepsilon > 0$ . Since  $u \in \mathcal{U}$ , there exists  $k_0 \in \mathbb{N}$  such that  $\|u\|_\infty \leq k_0$ , and hence  $u \in \mathcal{U}_{k_0}$ . This implies

$$\inf_{k \in \mathbb{N}} V_k(x) \leq V_{k_0}(x) \leq V(x) + \varepsilon.$$

Since  $\varepsilon$  was arbitrary, this shows the claim on  $\mathcal{D}_0$  for both  $V$  and  $v$ . For  $x \notin \mathcal{D}_0$  we have  $V_k(x) = V(x) = \infty$  and  $v_k(x) = v(x) = 1$ , which shows the claim also in this case.  $\square$

*Remark 5.2.* If the assumptions of Proposition 3.6 hold, then, since  $v_k$  is decreasing in  $k$ , Dini's theorem yields that  $v_k$  converges to  $v$  locally uniformly on  $\mathbb{R}^n$ .

For the following proposition recall the definition of set limits, which for a sequence of sets  $X_k$  are given by

$$\limsup_{k \rightarrow \infty} X_k := \bigcap_{k \in \mathbb{N}} \bigcup_{m \geq k} X_m \quad \text{and} \quad \liminf_{k \rightarrow \infty} X_k := \bigcup_{k \in \mathbb{N}} \bigcap_{m \geq k} X_m$$

and, if these two sets coincide,

$$\lim_{k \rightarrow \infty} X_k := \limsup_{k \rightarrow \infty} X_k = \liminf_{k \rightarrow \infty} X_k.$$

PROPOSITION 5.3. *Consider the sets*

$$\mathcal{D}_k := \{x \in \mathbb{R}^n \mid \text{there exists } u \in \mathcal{U}_k \text{ with } \|\varphi(t, x, u)\| \rightarrow 0 \text{ for } t \rightarrow \infty\}.$$

*Then the set limit  $\lim_{k \rightarrow \infty} \mathcal{D}_k$  exists and satisfies*

$$\mathcal{D}_0 = \lim_{k \rightarrow \infty} \mathcal{D}_k.$$

*Proof.* Since we have that  $V \leq \dots \leq V_{k+1} \leq V_k$ , we obtain the inclusion

$$\mathcal{D}_k \subseteq \mathcal{D}_{k+1} \subseteq \dots \subseteq \mathcal{D}_0.$$

It follows that  $\bigcup_{m \geq k} \mathcal{D}_m \subseteq \mathcal{D}_0$  for each  $k$  and hence

$$\limsup_{k \rightarrow \infty} \mathcal{D}_k = \bigcap_{k \in \mathbb{N}} \bigcup_{m \geq k} \mathcal{D}_m \subseteq \mathcal{D}_0.$$

On the other hand, if  $x \in \mathcal{D}_0$ , then for any  $\varepsilon > 0$  there exists  $k_0 \in \mathbb{N}$  with  $V_k(x) \leq V(x) + \varepsilon$  for all  $k \geq k_0$ . This implies that  $x \in \mathcal{D}_k$  for all  $k \geq k_0$  and consequently  $x \in \bigcap_{m \geq k_0} \mathcal{D}_m$ . This implies

$$x \in \bigcup_{k \in \mathbb{N}} \bigcap_{m \geq k} \mathcal{D}_m = \liminf_{k \rightarrow \infty} \mathcal{D}_k,$$

and since  $x \in \mathcal{D}_0$  was arbitrary we obtain

$$\mathcal{D}_0 \subseteq \liminf_{k \rightarrow \infty} \mathcal{D}_k,$$

which shows the claim.  $\square$

*Remark 5.4.* This proposition implies that for any compact set  $K \subset \mathbb{R}^n$  the convergence

$$d_H(K \cap \mathcal{D}_k, K \cap \mathcal{D}_0) \rightarrow 0$$

in the Hausdorff metric holds (see, e.g., [2, Proposition 1.1.5]). In particular, if  $\mathcal{D}_0$  is bounded, then we obtain uniform convergence of  $\mathcal{D}_k$  to  $\mathcal{D}_0$  in the Hausdorff metric.

In particular, this implies that for any compact set  $K \subset \mathcal{D}_0$  we obtain  $K \subset \mathcal{D}_k$  for all sufficiently large  $k$ . Thus, in order to steer the system to 0 from a compact subset  $K \subset \mathcal{D}_0$ , it is sufficient to consider bounded control functions.

**6. Examples.** In this section we discuss the necessity of some of our assumptions. Also it is explained how the classical case of linear quadratic control fits within the present framework.

*Example 6.1.* Consider the one-dimensional dynamics

$$(6.1) \quad \dot{x}(t) = (x(t) - 1)(u(t) + 1) + 1 = x(t)(u(t) + 1) - u(t), \quad t \geq 0,$$

where  $U = \mathbb{R}$ . The origin is an equilibrium point so that (H1) is satisfied, while  $x = 1$  is repulsive, in the sense that any trajectory starting from  $x_0 \geq 1$  cannot reach the origin. With this it is easy to see that (H2) is satisfied and  $\mathcal{D}_0 = (-\infty, 1)$ . Furthermore, (H0) is satisfied with  $\gamma(u) = |u|$ .

Now consider the cost function  $g_1(x, u) = |x|$ , which satisfies (H3) and (H4) but neither (H5) nor (H6). For  $x_0 \in (0, 1)$  and an arbitrary constant  $\alpha > 0$ , choose

$$u(t) = \frac{-\alpha - 1}{\phi(t) - 1} \chi_{[0, x_0/\alpha]}(t),$$

where  $\chi_{[0, x_0/\alpha]}$  denotes the indicator function of the interval  $[0, x_0/\alpha]$ . The corresponding solution of (6.1) is given by

$$\phi(t) = (x_0 - \alpha t) \chi_{[0, x_0/\alpha]}(t).$$

Observe that for  $x_0$  close to 1 we need a very large control to start to move towards the origin. This is because the control  $u$  is multiplied by  $x - 1$ .

By calculating the corresponding cost, we obtain

$$V_1(x_0) \leq \int_0^\infty g_1(\phi(t), u(t)) dt = x_0^2/2\alpha,$$

and therefore, sending  $\alpha \rightarrow +\infty$ , it follows that  $V_1(x_0) = 0$  for any  $x_0 \in (0, 1)$ . Of course,  $V_1(x) = \infty$  for  $x \geq 1$ . Summarizing this shows that  $v_1$  is discontinuous on  $\mathbb{R}$  and not a CLF on  $\mathcal{D}_0$ .

On the other hand, by setting  $g_2(x, u) = \max\{|x| + |u|, (|x| + |u|)^2\}$ , a cost function satisfying (H6) is obtained. To analyze the associated value functions, fix  $x_0 \in (0, 1)$  and choose a control  $u$  such that  $\phi(t) := \phi(t, x, u) \rightarrow 0$ . We will assume that  $\phi$  is strictly decreasing, as otherwise it is clearly not optimal. Now let  $T > 0$  be a time such that  $\phi(T) > 0$ , and then we have

$$\begin{aligned} J_2(x, u) &\geq \int_0^T g_2(\phi(t), u(t)) dt \geq \int_0^T \phi(t) + u(t) dt = \int_0^T \phi(t) + \frac{\phi(t) - \dot{\phi}(t)}{1 - \phi(t)} dt \\ &\geq \int_0^T \frac{-\dot{\phi}(t)}{1 - \phi(t)} dt = \log(1 - \phi(T)) - \log(1 - x_0). \end{aligned}$$

As  $\phi(T)$  approaches 0 (in finite or infinite time) this calculation shows that  $V_2(x_0) \geq -\log(1 - x_0)$  for  $x_0 \in (0, 1)$  so that in particular  $v_2$  is continuous on  $\mathbb{R}$  and a CLF on  $\mathcal{D}_0$  (where we leave the assertion for  $(-\infty, 0)$  to the reader).

Finally note that a combination of the previous examples leads to an intermediate situation. To this end, let  $h : \mathbb{R} \rightarrow [0, 1]$  be a continuous function such that  $h(x) = 1$  if  $x \in (-\infty, 1/2]$ ,  $h(x) = 0$  for  $x \in [3/4, \infty)$ , and let  $g_3(x, u) = |x| + h(x)|u|$ . Then it follows for  $x \in [0, 1/2]$  that  $V_3(x) = V_2(x) \geq -\log(1 - x)$  by the considerations on  $g_2$ , whereas for  $x \in (3/4, 1)$  we have  $V_3(x) = V(3/4)$  using that  $V_1$  is constant on that interval. In this example (H5) and (H6) are not satisfied,  $v_3$  is not continuous, and  $V_3$  is a CLF only on a subset of  $\mathcal{D}_0$ .

*Example 6.2.* Finally we show that the classical linear quadratic control problem fits into our setup. This problem is obtained if we set

$$f(x, u) = Ax + Bu \quad \text{and} \quad g(x, u) = x^T Qx + u^T Ru,$$

where  $A, B, Q, R$  are matrices of appropriate dimensions, with  $Q$  and  $R$  being symmetric and positive definite.

By direct computations one sees that these functions satisfy (H0) for any  $\gamma \in \mathcal{K}_\infty$ , (H1), (H3), and (H5). The linear system also satisfies (H2'), because it is known that local asymptotic controllability implies the existence of a feedback matrix  $F$  such that  $A + BF$  is exponentially stable; i.e., this matrix has all of its eigenvalues in the open left half-plane, which yields (H2') with  $\beta(r, t) = Ke^{-\lambda t}r$  for suitable constants  $K, \lambda > 0$ . Hence we obtain  $\beta(r, t) = \alpha_2(\alpha_1(r)e^{-t})$ , with  $\alpha_2(r) = r^\lambda$ , which implies (H4') for our  $g$  with  $\delta = 2/\lambda$  and  $C = \|Q + R\|$ . Finally, (H6) is satisfied because  $g$  grows quadratically in  $u$  while  $f$  grows only linearly in  $u$ . Thus, the classical linear quadratic problem is a special case of our setup, and the resulting equation (4.1) is given by

$$(6.2) \quad \sup_{u \in U} \{-DV(x)(Ax + Bu) - x^T Qx - u^T Ru\} = 0.$$

For the quadratic ansatz  $V(x) = x^T Px$ , with symmetric matrix  $P$ , we obtain

$$DV(x)(Ax + Bu) = x^T P(Ax + Bu) + (Ax + Bu)^T Px.$$



Assuming  $U = \mathbb{R}^m$ , we can explicitly solve the maximization problem over  $u$  by setting the first derivative of the resulting expression to 0 and obtain

$$u(x) = -R^{-1}B^T Px.$$

Plugging this into (6.2) and multiplying by  $-1$  yields

$$x^T PBR^{-1}B^T Px - x^T PAx - x^T A^T Px - x^T Qx = 0,$$

which is equivalent to

$$PBR^{-1}B^T P - PA - A^T P - Q = 0;$$

i.e., (4.1) reduces to the well-known algebraic Riccati equation from linear optimal control; see [33, section 8.4].

**Appendix.** In this appendix we give the proof of Lemma 4.6.

*Proof.* We prove the lemma for (4.1) and (4.5), and the assertions for (4.2) and (4.6) follow by the same arguments.

(i) If  $V^-$  is a viscosity subsolution of (4.1), then for any supergradient  $p$  of  $V^-$  in  $x$  we have that

$$\sup_{u \in U} \{-f(x, u)p - g(x, u)\} \leq 0.$$

This implies

$$-f(x, u)p - g(x, u) \leq 0 \quad \forall u \in U,$$

and, since  $1 + \|f(x, u)\|$  is positive, this implies

$$-\tilde{f}(x, u)p - \tilde{g}(x, u) = (1 + \|f(x, u)\|)^{-1}(-f(x, u)p - g(x, u)) \leq 0 \quad \forall u \in U,$$

which in turn implies

$$\sup_{u \in U} \{-\tilde{f}(x, u)p - \tilde{g}(x, u)\} \leq 0,$$

and hence  $V^-$  is a viscosity supersolution of (4.5).

The converse direction follows by the same argument, since again we multiply by a positive factor, now  $1 + \|f(x, u)\|$ .

(ii) Let  $V^+$  be a viscosity supersolution of (4.1). Then for any subgradient  $p$  of  $V^+$  in  $x$  we have

$$\sup_{u \in U} \{-f(x, u)p - g(x, u)\} \geq 0.$$

Now we distinguish two cases:

(a) We can find  $u^* \in U$  such that

$$-f(x, u^*)p - g(x, u^*) \geq 0.$$

Since  $1 + \|f(x, u^*)\|$  is positive, we obtain

$$-\tilde{f}(x, u^*)p - \tilde{g}(x, u^*) = (1 + \|f(x, u^*)\|)^{-1}(-f(x, u^*)p - g(x, u^*)) \geq 0.$$

This implies

$$\sup_{u \in U} \{-\tilde{f}(x, u)p - \tilde{g}(x, u)\} \geq 0,$$

and hence  $V^+$  is a viscosity supersolution of (4.5).

(b) For all  $u \in U$  the inequality

$$-f(x, u)p - g(x, u) \leq 0$$

holds. In this case, since  $1 + \|f(x, u)\| \geq 1$ , for all  $u \in U$  we obtain

$$\begin{aligned} -\tilde{f}(x, u)p - \tilde{g}(x, u) &= \underbrace{(1 + \|f(x, u)\|)^{-1}}_{\leq 1} \underbrace{(-f(x, u)p - g(x, u))}_{\leq 0} \\ &\geq -f(x, u)p - g(x, u). \end{aligned}$$

This implies

$$\sup_{u \in U} \{-\tilde{f}(x, u)p - \tilde{g}(x, u)\} \geq \sup_{u \in U} \{-f(x, u)p - g(x, u)\} \geq 0.$$

Thus also in this case  $V^+$  is a viscosity supersolution of (4.5).

Conversely, let  $V^+$  be a viscosity supersolution of (4.5). Then for any subgradient  $p$  of  $V^+$  in  $x$  we have

$$\sup_{u \in U} \{-\tilde{f}(x, u)p - \tilde{g}(x, u)\} \geq 0.$$

Since  $\tilde{f}$  is bounded and  $\tilde{g}$  grows unbounded in  $u$  due to (H6), the supremum over  $u$  is contained in a compact set. Hence by continuity we can find a control value  $u^* \in U$  for which the maximum is attained, i.e.,

$$-\tilde{f}(x, u^*)p - \tilde{g}(x, u^*) \geq 0.$$

Since  $1 + \|f(x, u^*)\|$  is positive, we obtain

$$-f(x, u^*)p - g(x, u^*) = (1 + \|f(x, u^*)\|)(-\tilde{f}(x, u^*)p - \tilde{g}(x, u^*)) \geq 0.$$

This implies

$$\sup_{u \in U} \{-f(x, u)p - g(x, u)\} \geq 0,$$

and hence  $V^+$  is a viscosity supersolution of (4.1).  $\square$

## REFERENCES

- [1] Z. ARTSTEIN, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.
- [2] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Systems Control Found. Appl. 2, Birkhäuser Boston, Boston, MA, 1990.
- [3] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 1997.
- [4] M. BARDI AND F. DA LIO, *On the Bellman equation for some unbounded control problems*, Nonlinear Differential Equations Appl., 4 (1997), pp. 491–510.

- [5] G. BARLES, *An approach of deterministic control problems with unbounded data*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 235–258.
- [6] G. BARLES, *Solutions de Viscosité des Équations de Hamilton-Jacobi*, Math. Appl. (Berlin) 17, Springer-Verlag, Paris, 1994.
- [7] E. N. BARRON AND R. JENSEN, *Lyapunov stability using minimum distance control*, Nonlinear Anal., 43 (2001), pp. 923–936.
- [8] R. W. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory (Houghton, MI, 1982), Progr. Math. 27, Birkhäuser Boston, Boston, MA, 1983, pp. 181–191.
- [9] F. CAMILLI, L. GRÜNE, AND F. WIRTH, *A generalization of Zubov's method to perturbed systems*, SIAM J. Control Optim., 40 (2001), pp. 496–515.
- [10] F. H. CLARKE, YU. S. LEDYAEV, L. RIFFORD, AND R. J. STERN, *Feedback stabilization and Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 25–48.
- [11] F. H. CLARKE, Y. S. LEDYAEV, E. D. SONTAG, AND A. I. SUBBOTIN, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.
- [12] F. H. CLARKE, Y. S. LEDYAEV, AND R. J. STERN, *Asymptotic stability and smooth Lyapunov functions*, J. Differential Equations, 149 (1998), pp. 69–114.
- [13] J.-M. CORON, *A necessary condition for feedback stabilization*, Systems Control Lett., 14 (1990), pp. 227–232.
- [14] F. DA LIO, *On the Bellman equation for infinite horizon problems with unbounded cost functional*, Appl. Math. Optim., 41 (2000), pp. 171–197.
- [15] S. DUBLJEVIĆ AND N. KAZANTSI, *A new Lyapunov design approach for nonlinear systems based on Zubov's method*, Automatica, 38 (2002), pp. 1999–2007.
- [16] R. A. FREEMAN AND P. V. KOKOTOVIĆ, *Inverse optimality in robust stabilization*, SIAM J. Control Optim., 34 (1996), pp. 1365–1391.
- [17] R. A. FREEMAN AND P. V. KOKOTOVIĆ, *Robust Nonlinear Control Design*, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 1996.
- [18] P. GIESL, *Construction of Global Lyapunov Functions Using Radial Basis Functions*, Lecture Notes in Math. 1904, Springer-Verlag, Berlin, 2007.
- [19] S. F. HAFSTEIN, *A constructive converse Lyapunov theorem on exponential stability*, Discrete Contin. Dyn. Syst., 10 (2004), pp. 657–678.
- [20] W. HAHN, *Stability of Motion*, translated from the German manuscript by Arne P. Baartz, Grundlehren Math. Wiss. 138, Springer-Verlag, New York, 1967.
- [21] Y. KURCVEĚL', *On the inversion of the second theorem of Lyapunov on stability of motion*, Czechoslovak Math. J., 6 (1956), pp. 217–259, 455–484 (in Russian).
- [22] Y. S. LEDYAEV AND E. D. SONTAG, *A Lyapunov characterization of robust stabilization*, Nonlinear Anal., 37 (1999), pp. 813–840.
- [23] Y. LIN, E. D. SONTAG, AND Y. WANG, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.
- [24] M. MALISOFF, *Further results on the Bellman equation for optimal control problems with exit times and nonnegative Lagrangians*, Systems Control Lett., 50 (2003), pp. 65–79.
- [25] M. MALISOFF, *Further results on Lyapunov functions and domains of attraction for perturbed asymptotically stable systems*, Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal., 12 (2005), pp. 193–225.
- [26] J. L. MASSERA, *On Liapunov's condition of stability*, Ann. of Math., 50 (1949), pp. 705–721.
- [27] L. RIFFORD, *Existence of Lipschitz and semiconcave control-Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 1043–1064.
- [28] L. RIFFORD, *Semiconcave control-Lyapunov functions and stabilizing feedbacks*, SIAM J. Control Optim., 41 (2002), pp. 659–681.
- [29] E. P. RYAN, *On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback*, SIAM J. Control Optim., 32 (1994), pp. 1597–1604.
- [30] R. SEPULCHRE, M. JANKOVIĆ, AND P. V. KOKOTOVIĆ, *Constructive Nonlinear Control*, Comm. Control Engrg. Ser., Springer-Verlag, Berlin, 1997.
- [31] E. D. SONTAG, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462–471.
- [32] E. D. SONTAG, *Comments on integral variants of ISS*, Systems Control Lett., 34 (1998), pp. 93–100.
- [33] E. D. SONTAG, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, 2nd ed., Texts Appl. Math. 6, Springer-Verlag, New York, 1998.
- [34] E. D. SONTAG, *Stability and stabilization: Discontinuities and the effect of disturbances*, in Nonlinear Analysis, Differential Equations and Control (Montreal, QC, 1998), NATO Sci. Ser. C Math. Phys. Sci. 528, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 551–598.

- [35] P. SORAVIA, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations. I. Equations of unbounded and degenerate control problems without uniqueness*, Adv. Differential Equations, 4 (1999), pp. 275–296.
- [36] P. SORAVIA, *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations. II. Equations of control problems with state constraints*, Differential Integral Equations, 12 (1999), pp. 275–293.
- [37] A. R. TEEL AND L. PRALY, *A smooth Lyapunov function from a class- $\mathcal{KL}$  estimate involving two positive semidefinite functions*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 313–367.
- [38] F. W. WILSON, JR., *The structure of the level surfaces of a Lyapunov function*, J. Differential Equations, 3 (1967), pp. 323–329.
- [39] V. I. ZUBOV, *Methods of A. M. Lyapunov and Their Application*, translation prepared under the auspices of the United States Atomic Energy Commission, Leo F. Boron, ed., P. Noordhoff, Groningen, The Netherlands, 1964.

## A NEW PERSPECTIVE IN THE STABILITY ASSESSMENT OF NEUTRAL SYSTEMS WITH MULTIPLE AND CROSS-TALKING DELAYS\*

NEJAT OLGAC<sup>†</sup>, TOMÁŠ VYHLÍDAL<sup>‡</sup>, AND RIFAT SIPAHI<sup>§</sup>

**Abstract.** The stability of neutral systems with two cross-talking delays is investigated using the method of cluster treatment of characteristic roots (CTCR). There are two main outcomes of this study: (a) we create the “strong stabilizability” (also called the “delay stabilizability”) of the system directly from the CTCR procedure. This is achieved by a small-delay stability treatment while performing the steps of the CTCR. For the “delay-stabilizable systems,” we also arrive at the exact bounds of the stability regions in the domain of the delays. (b) We deploy a point-wise algorithm which computes the rightmost roots of the characteristic quasi polynomial for cross-verification of these stability regions. Several examples are presented. The correspondence between the two methods for all of them is shown to be very strong.

**Key words.** time delay, neutral systems, multiple delay, cross-talking delays, cluster treatment of characteristic roots

**AMS subject classifications.** 15A15, 15A09, 15A23

**DOI.** 10.1137/070679302

**1. Problem statement and an explicit function for stability.** The method of cluster treatment of characteristic roots (CTCR) provides an efficient procedure for determining the complete stability picture of linear time invariant time delay systems (TDS) in the domain of delays. In [20], [21], the method is applied to retarded systems with multiple delays and to neutral systems [16], [17], [19] with single delay only. In this paper, three new aspects are considered within the neutral dynamics: (i) multiple rationally independent delays, (ii) cross-talk among the delays which affect both retarded as well as the neutral parts of the dynamics, and (iii) numerical cross-validation of the results of CTCR. The general dynamic structure is taken as follows:

$$(1) \quad \begin{aligned} \frac{d}{dt}[\mathbf{x}(t) - \mathbf{A}\mathbf{x}(t - \tau_1) - \mathbf{B}\mathbf{x}(t - \tau_2) - \mathbf{C}\mathbf{x}(t - \tau_1 - \tau_2)] \\ = \mathbf{D}\mathbf{x}(t - \tau_1) + \mathbf{F}\mathbf{x}(t - \tau_2) + \mathbf{G}\mathbf{x}(t - \tau_1 - \tau_2) + \mathbf{H}\mathbf{x}(t), \end{aligned}$$

where  $\mathbf{x}(t) \in \mathbb{R}^n$  is the dependent variable,  $\boldsymbol{\tau} = (\tau_1, \tau_2) \in \mathbb{R}^{2+}$  are the delays, and  $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{F}, \mathbf{G}, \mathbf{H}$  are the constant coefficient matrices of compatible dimensions. Notice that the terms with  $\mathbf{C}$  and  $\mathbf{G}$  represent the cross-talk between the delays within the neutral and retarded segments of the equation, respectively. The objective is to

---

\*Received by the editors January 5, 2007; accepted for publication (in revised form) October 2, 2007; published electronically January 25, 2008.

<http://www.siam.org/journals/sicon/47-1/67930.html>

<sup>†</sup>Mechanical Engineering Department, University of Connecticut, Storrs, CT 06269-3139 (olgac@engr.uconn.edu). This author’s research was supported by grants from the DoE (DE-FG02-04ER25656), NSF (CMS-0439980, CMS-0539980, DMI 0522910), and ARO (W911NF-07-1-0557).

<sup>‡</sup>Department of Instrumentation and Control Engineering, Centre for Applied Cybernetics, Faculty of Mechanical Engineering, Czech Technical University, 16607 Prague 6, Czech Republic (Tomas.Vyhldal@fs.cvut.cz). This author’s research was supported by the Ministry of Education of the Czech Republic under project 1M0567 (the work on the presented research started during his visit at the University of Connecticut).

<sup>§</sup>Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA 02115 (rifat@coe.neu.edu). This author’s research was supported by grants from the DoE (DE-FG02-04ER25656), NSF (CMS-0439980, CMS-0539980, DMI 0522910), and ARO (W911NF-07-1-0557).

analyze the stability robustness of the dynamics with respect to the uncertain time delays in the semi-infinite first quadrant of  $(\tau_1, \tau_2) \in \mathbb{R}^{2+}$ .

The characteristic equation of the system (1) is of quasi-polynomial form with infinitely many roots:

$$(2) \quad CE(s, \tau_1, \tau_2) = \det(s(\mathbf{I} - \mathbf{A}e^{-\tau_1 s} - \mathbf{B}e^{-\tau_2 s} - \mathbf{C}e^{-(\tau_1 + \tau_2)s}) - \mathbf{D}e^{-\tau_1 s} - \mathbf{F}e^{-\tau_2 s} - \mathbf{G}e^{-(\tau_1 + \tau_2)s} - \mathbf{H}) = 0.$$

Assuming the system matrices are constant, the stability of this system reduces to finding  $(\tau_1, \tau_2)$  regions where all the roots of (2) remain in  $\mathbb{C}^-$ , i.e., in the open left half of the complex plane. This root set is also commonly referred to as the spectrum of  $CE(s, \tau_1, \tau_2)$ , which we denote with the notation  $\sigma[CE(s, \tau_1, \tau_2)]$ . We will represent the right-half open plane by  $\mathbb{C}^+$  and the imaginary axis by  $\mathbb{C}^0$  in the rest of the text. The entire complex plane becomes the union of all three:  $\mathbb{C} = \mathbb{C}^- \cup \mathbb{C}^0 \cup \mathbb{C}^+$ .

It is also important to note that the nondelayed neutral system may be stable or unstable. The treatment we present here is transparent to this feature. For nonzero delays, however, there may be some intriguing features regarding the root continuity of (2) with respect to the delays [9], [10]. These features are the main discriminators between the neutral and retarded classes of TDS which we will revisit in the text. At this point, however, we wish to present a critical feature: The necessary (but not sufficient) condition for the stability of (1) is the stability of the difference equation

$$(3) \quad \mathbf{x}(t) - \mathbf{A}\mathbf{x}(t - \tau_1) - \mathbf{B}\mathbf{x}(t - \tau_2) - \mathbf{C}\mathbf{x}(t - \tau_1 - \tau_2) = 0.$$

It implies that the spectrum of (3), i.e., the roots of

$$(4) \quad L(s, \tau_1, \tau_2) = \det(\mathbf{I} - \mathbf{A}e^{-\tau_1 s} - \mathbf{B}e^{-\tau_2 s} - \mathbf{C}e^{-(\tau_1 + \tau_2)s}) = 0,$$

must lie in  $\mathbb{C}^-$ . However, it is shown in the cited literature that the stability of (3) at a given  $(\tau_1, \tau_2)$  may be destroyed even for infinitesimal variations of the delays; some root discontinuities and infinitely many unstable roots may appear [1], [9], [11]. To handle such occurrences, a concept called “strong stability” was introduced. The difference equation (3) is strongly stable if it is asymptotically stable for the complete domain of  $(\tau_1, \tau_2) \in \mathbb{R}^{2+}$  [1], [9], [11]; see also [14]. Furthermore, an upper bound of the spectrum which is insensitive to small-delay variations has also been suggested in [15]. If this difference equation (3) is strongly stable, we call the system (1) a “delay-stabilizable” system.

We now look at the main objective of the CTCR procedure, which is the exhaustive determination of the stability robustness picture of (1) against uncertain delays  $(\tau_1, \tau_2)$ , and encounter a surprisingly simple, intermediate result: The strong stability necessary condition, as stated above, becomes evident as a natural by-product of CTCR. This point creates one of the contributory points of the paper.

In the text, we present a cross-validation study for the results of CTCR, utilizing a recent numerical root-finding procedure [24] for computing the roots of a quasi polynomial, such as  $CE(s, \boldsymbol{\tau})$  in (2), at a given point in the delay space  $(\tau_1, \tau_2)$ . The procedure itself is briefly described, and examples are provided to display the concurrence with CTCR.

The paper is organized as follows. In section 2, a review of the CTCR algorithm is provided. The main results are given in section 3, where an original *delay-stabilizability* test is proposed, which is directly obtained from the CTCR approach. A numerical cross-validation work is presented in section 4 over the findings of the CTCR. In section 5, we provide two application examples and in section 6 the conclusions.

**2. Review of CTCR.** The CTCR algorithm was described comprehensively in [16], [17], [20], [21]. We provide a short review here, describing only the main ideas. It is known that for a linear dynamics to switch the stability posture, the characteristic roots should cross the imaginary axis, say, at  $\omega i$ . Thus, for a successful stability analysis, one needs to detect exhaustively all the potential imaginary root crossings for all combinations of delays  $(\tau_1, \tau_2)$ . Let us denote the complete set of such crossing frequencies with  $\Omega$  and the corresponding root set with  $S_\Omega$ :

$$(5) \quad \begin{aligned} \Omega &= \{ \omega | CE(s, \tau) = 0, s = \omega i, \tau \in \mathbb{R}^{2+}, \omega \in \mathbb{R} \}, \\ S_\Omega &= \{ \omega i | \omega \in \Omega \}. \end{aligned}$$

We use  $\langle \tau, \omega \rangle$  notation to indicate this causality relation between  $\tau$  and  $\omega$ . It is trivial to show that an imaginary root  $s = \omega i$  with  $\langle \tau, \omega \rangle$  correspondence will be repeated infinitely many times at the mesh points with equidistant grid size of  $2\pi/\omega$ , as

$$(6) \quad (\tau_{1j}, \tau_{2k}) = \left( \tau_1 + \frac{2\pi}{\omega} j, \tau_2 + \frac{2\pi}{\omega} k \right), \quad j = 0, 1, 2, \dots, \quad k = 0, 1, 2, \dots$$

It has also been shown in earlier studies that the root continuity exists for the retarded class of TDS [1], [10], [11]. That is, small perturbations on  $\tau$  yield small perturbations on  $\omega$ , which can be formalized as

$$(7) \quad \langle \tau + \varepsilon, \omega + \varepsilon_c \rangle, \quad 0 < |\varepsilon| \ll 1, \quad 0 < |\varepsilon_c| \ll 1.$$

Clearly,  $\varepsilon = (\varepsilon_1, \varepsilon_2)$  and  $\varepsilon_c$  are interdependent through the characteristic equation (2). If one single point  $\langle \tau, \omega \rangle$  is known, one can determine infinitely many other points earmarked by the same  $\omega$ , as defined in (6). Obviously, the same remark holds for small perturbations on  $\omega$  as per (7). Therefore, we claim that these infinitely many trajectories generated by the delays,  $\tau$ , for varying  $\omega \in \Omega$  will comply with the rule (6) point by point. According to the D-subdivision method [6], these trajectories continuously partition the  $\tau$  domain into encapsulated regions in which the number of unstable roots,  $NU$ , remains fixed. Consequently, if there is any stability switching, it has to occur at the boundaries of these regions.

The above argument brings us to a complex problem of determining exhaustively the boundaries of these infinitely many regions. Interestingly, however, there is a discipline in this complex picture with two intriguing properties. It is proven in [16], [20], [21] as Proposition I that there is only a *manageably small number* of curves in  $\tau$  space called the “kernel curves”:

$$(8) \quad \wp_0(\tau_1, \tau_2) = \left\{ \tau | \langle \tau, \omega \rangle, \tau \in \mathbb{R}^{2+}, \omega \in \Omega, 0 \leq \tau_k \leq \frac{2\pi}{\omega}, k = 1, 2 \right\},$$

where  $\langle \tau, \omega \rangle$  correspondence creates the complete set of  $\Omega$ . Notice that for all  $\omega \in \Omega$  values  $\wp_0(\tau_1, \tau_2)$  represents the smallest  $\tau_1$  and  $\tau_2$  combination. And all the other curves are created from this set,  $\wp_0(\tau_1, \tau_2)$ , utilizing the point-wise property (6) for  $j, k > 0$ . Obviously, all of these curves will be representing the imaginary root crossings of  $\omega \in \Omega$ . These curves are called the “offspring curves” and are denoted by  $\wp_{jk}(\tau_1, \tau_2)$ , where  $j$  and  $k$  identify the  $j$ th and  $k$ th generation offspring in  $\tau_1$  and  $\tau_2$ , respectively, as per (6). Consequently, the complete set of kernel and offspring curves becomes  $\wp(\tau_1, \tau_2)$ :

$$(9) \quad \wp(\tau_1, \tau_2) = \wp_0(\tau_1, \tau_2) \cup \bigcup_{j=1}^{\infty} \bigcup_{k=1}^{\infty} \wp_{jk}(\tau_1, \tau_2).$$

Any kernel point on the trajectories of  $\wp_0(\tau_1, \tau_2)$  defined by  $j = k = 0$  imposes its  $\omega$  signature identically onto its offspring ( $j > 0$  and  $k > 0$ ). Thus,  $\Omega$  remains invariant from kernel curves to offspring curves. The kernel curves and the offspring constitute the complete (and exhaustive) distribution of  $(\tau_1, \tau_2)$  points where the characteristic equation  $CE(s, \tau_1, \tau_2)$  has root sets containing at least one pair of imaginary roots. And more interestingly, there exists no point in  $(\tau_1, \tau_2) \in \mathbb{R}^{2+}$ , outside the set  $\wp(\tau_1, \tau_2)$ , which renders imaginary characteristic roots.

Another perspective for exhaustive determination of stability switching boundaries in the delay domain is studied by [8] for a retarded class of systems with two delays but without a cross-talking feature (i.e.,  $\mathbf{A} = \mathbf{B} = \mathbf{C} = \mathbf{G} = \mathbf{0}$ ). They cleverly use a geometric triangulation property to capture the potential root crossing frequencies completely. In a separate study [5] and [22] offer a numerical procedure for determining the stability picture based on direct computation of the system roots (the code is publicly available now under the name “Trace-DDE”). We wish to emphasize that all of these findings comply with two fundamental propositions which are the cornerstone of the CTCR paradigm [20], [21].

A crucial property is the directional *root tendency* along the  $\tau_j$ ,  $j = 1, 2$ , axis at the crossing of  $s = \omega i$ , which is defined by

$$(10) \quad RT_{s=\omega i}^{\tau_j} = \operatorname{sgn} [\Re(\partial s / \partial \tau_j) |_{s=\omega i}].$$

$RT$  has a very interesting feature:  $s \in \mathbf{S}_\Omega$  along the  $\tau_1$  (or  $\tau_2$ ) axes across the corresponding points on a kernel curve, and its offspring, the root tendency,  $RT_s^{\tau_j}$ , remains unchanged so long as  $\tau_2$  (or  $\tau_1$ ) is kept fixed. This, which we call the “root tendency invariance” property, is proven in earlier publications; see Proposition II in [17], [20], [21]. This feature, in essence, declares the stabilizing (or destabilizing) transitions along the regional boundaries defined by  $\wp(\tau_1, \tau_2)$ .

Using the two properties above, one can establish the stability robustness picture of the system against delay uncertainties performing the following steps of the CTCR algorithm:

- (1) Determine exhaustively the kernel and offspring curves,  $\wp(\tau_1, \tau_2)$ .
- (2) Start from the nondelayed system,  $\boldsymbol{\tau} = \mathbf{0}$ , and evaluate  $NU(0)$ , which is a trivial task.
- (3) Following line segments in  $\boldsymbol{\tau} \in \mathbb{R}^{2+}$ , which are parallel to the individual coordinates  $\tau_1$  and  $\tau_2$ , connect the origin ( $\boldsymbol{\tau} = \mathbf{0}$ ) to a point of interest  $\boldsymbol{\tau}_0$ .
- (4) As this path crosses the kernel and offspring curves, increase  $NU$  by  $+2$  (or  $-2$ ) for the  $RT = +1$  ( $-1$ ), according to the D-subdivision method of [6].
- (5) Exhaustively identify the regions in  $(\tau_1, \tau_2)$  space with  $NU = 0$  as “stable” and the others ( $NU > 0$ ) as “unstable.”

Step (1) is the most critical one in this procedure, and we present a discussion on it next for clarity. To determine  $\wp(\tau_1, \tau_2)$  exhaustively we utilize Rekasius substitution for the exponential terms [18]

$$(11) \quad e^{-\tau_j s} \Rightarrow \frac{1 - T_j s}{1 + T_j s}, \quad T_j \in \mathbb{R}, \quad j = 1, 2.$$

This representation becomes exact for  $s = \omega i$ , with a constraint between  $T_i$  and  $\tau_i$ :

$$(12) \quad \tau_j = \frac{2}{\omega} [\tan^{-1}(\omega T_j) + k\pi], \quad k = 0, 1, 2, \dots$$



TABLE 1  
Routh's array for  $\overline{CE}(s, T_1, T_2)$ .

$s^{3n}$	$p_{3n}(T_1, T_2)$	$p_{3n-2}(T_1, T_2)$	$\vdots$	$\vdots$	$p_0$
$s^{3n-1}$	$p_{3n-1}(T_1, T_2)$	$p_{3n-3}(T_1, T_2)$	$\vdots$	$\vdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$s^2$	$R_{21}(T_1, T_2)$	$R_{22}(T_1, T_2) = p_0$			
$s^1$	$R_1(T_1, T_2)$				
$s^0$	$R_0(T_1, T_2) = p_0$				

This relation transforms the original characteristic equation  $CE(s, \tau_1, \tau_2) = 0$  given by (2) into a new form  $CE_T(s, T_1, T_2) = 0$  which is of fractional polynomial type. Multiplying this equation by  $(1 + T_1 s)^n(1 + T_2 s)^n$ , one obtains

$$\begin{aligned}
 \overline{CE}(s, T_1, T_2) &= \det[s((1 + T_1 s)(1 + T_2 s)\mathbf{I} - (1 + T_2 s)(1 - T_1 s)\mathbf{A} \\
 &\quad - (1 + T_1 s)(1 - T_2 s)\mathbf{B} - (1 - T_1 s)(1 - T_2 s)\mathbf{C} - (1 + T_2 s)(1 - T_1 s)\mathbf{D} \\
 &\quad - (1 + T_1 s)(1 - T_2 s)\mathbf{F} - (1 - T_1 s)(1 - T_2 s)\mathbf{G} - (1 + T_1 s)(1 + T_2 s)\mathbf{H}] \\
 &= \det(\mathbf{Q}(s, T_1, T_2)) = \sum_{j=0}^{3n} p_j(T_1, T_2) s^j = 0,
 \end{aligned}
 \tag{13}$$

where  $\mathbf{Q}(s, T_1, T_2)$  is a self-evident matrix. An interesting relation between the infinite-dimensional equation (2) and the  $3n$  degree equation (13) is that *they share the same imaginary root sets completely*. That is,

$$\begin{aligned}
 \mathbf{S}_{\Omega} &= \mathbf{S}_{\Omega} [s \mid CE(s, \tau_1, \tau_2) = 0, (\tau_1, \tau_2) \in \mathbb{R}^{2+}] \cap \mathbb{C}^0 \\
 &\equiv \mathbf{S}_{\Omega} [s \mid \overline{CE}(s, T_1, T_2) = 0, (T_1, T_2) \in \mathbb{R}^2] \cap \mathbb{C}^0,
 \end{aligned}
 \tag{14}$$

where  $\mathbf{S}_{\Omega}$  represents the complete topology of root sets of  $\overline{CE}$  for the entire space of  $(T_1, T_2) \in \mathbb{R}^2$ . The most beneficial point in transforming  $CE(s, \tau_1, \tau_2)$  into  $\overline{CE}(s, T_1, T_2)$  is obvious: the parametric equation (13) is much easier to study compared with (2). Thus, the projections of the kernel curves in  $(T_1, T_2)$  space, which are called “core curves,” are obtained.

We now present how to determine the core curve of  $\overline{CE}(s, T_1, T_2)$ , i.e., those  $(T_1, T_2)$  that give rise to  $s = \omega i \in \mathbf{S}_{\Omega} \cap \mathbb{C}^0$ . The easiest procedure to find all the imaginary roots of such characteristic polynomials is the classical Routh–Hurwitz method [12]. From the rules of Routh's array of Table 1, the imaginary roots of (13) are found at

$$R_1(T_1, T_2) = 0, \quad \text{with the condition}$$

$$R_{21}(T_1, T_2)p_0 > 0.$$

At every point  $(T_1, T_2)$  satisfying (15) and (16) there exists a crossing frequency  $\omega = \sqrt{p_0/R_{21}(T_1, T_2)}$ .

The next step in CTCR is to numerically map these core curves into kernel curves via (12) and further to offspring in  $(\tau_1, \tau_2)$  space via (6). This completes the deter-

mination of the entire set of kernel and offspring curves,  $\wp(\tau_1, \tau_2)$ , i.e., the first step in the CTCR procedure. The remaining steps are performed relatively easily.

**3. Delay stabilizability and strong stability concepts.** The “strong stability” concept is an interesting feature of the neutral TDS which was carefully studied in the literature [1], [9], [11]; see also [14], [15]. Following the conceptual discussions in the earlier investigation, we state the following:

(i) “Strong stability” of (3) implies that the system (1) may be asymptotically stable in a two-dimensional region around a point  $\tau_0 \in \mathbb{R}^{2+}$ , defined by  $0 < |\tau - \tau_0| < \varepsilon \ll 1$ .

(ii) There is a necessary condition for (1) to be “delay-stabilizable,” which requires the difference equation (3) to be strongly stable. A system that is not strongly stable cannot be delay-stabilizable. In other words, the spectrum of the respective characteristic equation (4) must be in the left-half open space (excluding the imaginary axis) for  $|\tau - \tau_0| < \varepsilon, \varepsilon \text{ finite}$ , where  $\tau_0 \in \mathbb{R}^{2+}$  is a point in the delay domain, or  $\sigma[L(s, \tau)] \in \mathbb{C}^-$ . Notice that if a system is not “strongly stabilizable,” there exists no point  $\tau_0 \in \mathbb{R}^{2+}$ , for which

$$(17) \quad \sigma[CE(s, \tau)]|_{\forall \tau: |\tau - \tau_0| < \varepsilon, \varepsilon \text{ finite}} \in \mathbb{C}^-.$$

(iii) The “strong-stabilizability” condition is (a) independent of the delays,  $\tau$  [1], [9], [11], and (b) independent of the retarded terms of (1), i.e.,  $\mathbf{D}$ ,  $\mathbf{F}$ ,  $\mathbf{G}$ , and  $\mathbf{H}$ .

(iv) As a direct consequence of (iii(a)), the strong stability condition can be simply tested at  $|\tau| = \varepsilon \ll 1$ , i.e., for small delays. If the system (1) is “strongly stable” for small delays, the necessary condition for strong stability holds.

What motivates this work is that (a) the “strong stability” condition is obtained as a simple by-product of the CTCR procedure, and (b) further deployment of CTCR creates an exhaustive and exact “stability robustness” picture for the system (1) in the entire  $\tau \in \mathbb{R}^{2+}$  domain. We present the details of these claims.

From these discussions, it results that further investigating the stability picture of a neutral system, of which the associated difference equation (3) is not strongly stable, is unnecessary. Even though for some rationally dependent (i.e., commensurate) delays all the roots may lie in the left half of the complex plane (see [14]), when the delays become rationally independent, due to, for example, arbitrarily small changes in the delays, infinitely many destabilizing roots cross the imaginary axis.

### 3.1. Main lemma: Delay stabilizability resulting from CTCR.

**MAIN LEMMA.** *The “delay-stabilizability” (or strong stability) necessary conditions as stated in observations (i)–(iv) above result as a by-product of direct deployment of CTCR.*

*Proof.* The steps of the CTCR procedure are given in section 2. As we perform the second step, a very peculiar feature appears. If for small delays

$$(18) \quad \sigma[CE(s, \tau)|_{\tau=0}] \cap \mathbb{C}^+ \neq \sigma[CE(s, \tau)|_{\tau=0^+}] \cap \mathbb{C}^+$$

it implies that  $\tau : 0 \rightarrow 0^+$  transition introduces some new right-half plane roots. These new roots may be (a) finite, in which case they are the imaginary roots of the nondelayed system, and therefore they are easily identified, or (b) at infinity and with unbounded magnitudes. It is proven that these roots at infinity are infinitely many, and they may show discontinuity complying with the Riemann sphere concept [13]; see also [10]. Earlier investigations state that the mentioned discontinuity as  $\tau : 0 \rightarrow 0^+$  invites infinitely many unbounded unstable roots iff the difference equation (3) is

not strongly stable. Further findings declare that such systems cannot be stabilized anywhere in  $\tau \in \mathbb{R}^+$ , except possibly at those points where  $\tau_1, \tau_2$  are rationally dependent [1], [14], [15]. For single delay cases, [16], [17] study the Riemann sphere concept further. Therefore, we propose imposing the condition that the  $\tau : \mathbf{0} \rightarrow \mathbf{0}^+$  transition does not create any infinite root transition to  $\mathbb{C}^+$  on the Riemann sphere. This feature is referred to as the “no root migration” requirement in the rest of the text.

We now look at this “small-delay” phenomenon under Rekasius transformation and explain how it shapes up if there is an imaginary root  $\omega i$  for  $\tau = \mathbf{0}$ . Even if the migration of roots takes place over the Riemann sphere, since the sphere has only one point at infinity it can be taken as purely imaginary; see [13].

(i) For small delays, (12) becomes

$$(19) \quad \tan \left( \frac{\tau_j \omega}{2} \right) = T_j \omega, \quad j = 1, 2,$$

which implies that

$$(20) \quad \tau_j : 0 \rightarrow \varepsilon_j \implies T_j : 0 \rightarrow \varepsilon_j / 2.$$

Notice that the  $CE(s, \tau)$ ,  $\tau = \mathbf{0}$ , and  $\overline{CE}(s, \mathbf{T})$ ,  $\mathbf{T} = \mathbf{0}$ , are identical, and so are their spectra. In the two-dimensional  $\tau$  domain such a transition can be represented using a parameter

$$(21) \quad m = \frac{\tau_2}{\tau_1} = \frac{T_2}{T_1}, \quad m \in [0, \infty).$$

(ii) The “no root migration” requirement must be independent of  $m$ . Using (21) in (13) it becomes

$$(22) \quad \begin{aligned} \overline{CE}(s, T_1, m T_1) &= \det \mathbf{Q}(s, T_1, m T_1) = \det s [(1 + (1 + m)T_1 s + mT_1^2 s^2)\mathbf{I} \\ &\quad - (1 - (1 - m)T_1 s - mT_1^2 s^2)\mathbf{A} - (1 + (1 - m)T_1 s - mT_1^2 s^2)\mathbf{B} \\ &\quad - (1 - (1 + m)T_1 s + mT_1^2 s^2)\mathbf{C}] - (1 - (1 - m)T_1 s - mT_1^2 s^2)\mathbf{D} \\ &\quad - (1 + (1 - m)T_1 s - mT_1^2 s^2)\mathbf{F} - (1 - (1 + m)T_1 s + mT_1^2 s^2)\mathbf{G} \\ &\quad - (1 + (1 + m)T_1 s + mT_1^2 s^2)\mathbf{H}] = \sum_{j=0}^{3n} p_j(m, T_1) s^j = 0. \end{aligned}$$

Each element of the characteristic matrix  $\mathbf{Q}(s, T_1, m T_1)$  is a third degree polynomial in  $s$  in the generic form

$$(23) \quad \begin{aligned} q_{i,j}(T_1, m, s) &= (e_{i,j} + a_{i,j} + b_{i,j} - c_{i,j})mT_1^2 s^3 \\ &\quad + [(d_{i,j} + f_{i,j} - g_{i,j} - h_{i,j})mT_1^2((1 + m)(e_{i,j} + c_{i,j}) + (1 - m)(a_{i,j} - b_{i,j}))T_1]s^2 \\ &\quad + [(1 - m)(d_{i,j} - f_{i,j}) + (1 + m)(g_{i,j} - h_{i,j}))T_1 + e_{i,j} - a_{i,j} - b_{i,j} - c_{i,j}]s \\ &\quad - d_{i,j} - f_{i,j} - g_{i,j} - h_{i,j} = \alpha_3^{i,j}(m, T_1)s^3 + \alpha_2^{i,j}(m, T_1)s^2 + \alpha_1^{i,j}(m, T_1)s + \alpha_0^{i,j}, \end{aligned}$$

where  $e_{i,j}$  are the elements of the identity matrix  $\mathbf{I}$  ( $e_{i,j} = 1$  for  $i = j$  and  $e_{i,j} = 0$  for  $i \neq j$ ) and lowercase notations represent the elements of the respective uppercase matrices.  $\alpha_k^{i,j}(m, T_1)$  are self-evident expressions. Notice there is no influence of

matrices **A**, **B**, **C** (i.e., the neutral part of the dynamics) in  $\alpha_0^{i,j}$  and no influence of matrices **D**, **F**, **G**, **H** (i.e., the retarded part of the dynamics) in  $\alpha_3^{i,j}(m, T_1)$ . On the other hand, in  $\alpha_1^{i,j}(m, T_1)$  and  $\alpha_2^{i,j}(m, T_1)$  traces of all matrices, **A**,  $\dots$ , **H**, can be found. However, for the limiting case of  $T_1 \rightarrow 0$ , we may drop higher order  $T_1$  terms in favor of lower order ones; i.e., in  $\alpha_2^{i,j}(m, T_1)$  one can eliminate  $T_1^2$  terms but keep the  $T_1$  terms only. This term-wise limit operation on (23) yields

$$\begin{aligned}
 \tilde{q}_{i,j}(T_1, s) &= (e_{i,j} + a_{i,j} + b_{i,j} - c_{i,j})mT_1^2s^3 \\
 &+ [(1+m)(e_{i,j} + c_{i,j}) + (1-m)(a_{i,j} - b_{i,j})]T_1s^2 \\
 &+ (e_{i,j} - a_{i,j} - b_{i,j} - c_{i,j})s - d_{i,j} - f_{i,j} - g_{i,j} - h_{i,j} \\
 &= \tilde{\alpha}_3^{i,j}mT_1^2s^3 + \tilde{\alpha}_2^{i,j}T_1s^2 + \tilde{\alpha}_1^{i,j}s^2 + \alpha_0^{i,j},
 \end{aligned}
 \tag{24}$$

where  $\tilde{\bullet}$  notation is used for limiting formation of  $\bullet$ . Dropping the terms with higher powers of  $T_1$  at each step as we expand the determinant in (22), repeated applications of this step-wise limit operation yield

$$\begin{aligned}
 \widetilde{CE}(s, T_1, mT_1) &= \tilde{p}_{3n}(m)T_1^{2n}s^{3n} + \tilde{p}_{3n-1}(m)T_1^{2n-1}s^{3n-1} \\
 &+ \dots + \tilde{p}_{n+1}(m)T_1s^{n+1} + \tilde{p}_ns^n + \tilde{p}_{n-1}s^{n-1} + \dots + \tilde{p}_1s + \tilde{p}_0 = 0.
 \end{aligned}
 \tag{25}$$

It is trivial to show that  $\tilde{p}_j, j = 0, \dots, 3n$ , coefficients depend on

**A**, **B**, **C** only for  $j = n+1, \dots, 3n$ ,

**A**, **B**, **C**, **D**, **F**, **G**, **H** for  $j = 1, \dots, n$ ,

**D**, **F**, **G**, **H** for  $j = 0$ .

Furthermore,  $\tilde{p}_j, j = 0, \dots, n$ , are constants independent of  $m$ , while  $\tilde{p}_j, j = n+1, \dots, 3n$ , are polynomial functions of  $m$ . One can, therefore, partition the characteristic equation (25) as

$$\widetilde{CE}(s, T_1, mT_1) = \widetilde{CE}_1(s, T_1, mT_1) + \widetilde{CE}_2(s) - \tilde{p}_n(m)s^n = 0,
 \tag{26}$$

where

$$\widetilde{CE}_1(s, T_1, mT_1) = s^n \sum_{j=0}^{2n} \tilde{p}_{j+n}(m)T_1^j s^j
 \tag{27}$$

and

$$\widetilde{CE}_2(s) = \sum_{j=0}^n \tilde{p}_j s^j.
 \tag{28}$$

Please notice two critical features:

(a) Within the  $\widetilde{CE}_1$  expression, the terms given under summation have an important implication:

$$\sum_{j=0}^{2n} \tilde{p}_{j+n}(m)T_1^j s^j = \tilde{L}(s, T_1, mT_1),
 \tag{29}$$

where  $\tilde{L}(s, T_1, mT_1)$  is nothing other than the respective step-wise limiting conversion of the function  $L(s, \tau_1, m\tau_1)$  of (4), following the identical procedures which converted

from  $CE(s, \tau_1, m\tau_1)$  to  $\widetilde{CE}(s, T_1, mT_1)$ . This observation can be stated as follows: the limiting form of the difference equation (3) for small delays is a factor in  $\widetilde{CE}_1$ .

(b)  $\widetilde{CE}_2(s)$  is the characteristic function of the delay free system,  $\tau_1 = \tau_2 = 0$ .

(iii) When we inquire about the imaginary root crossing of (25) using Routh's array, it shapes up as given in Table 2. Note again that the step-wise limit operation is performed in the same manner as described above throughout the formation of the array. The following observations are made on this table:

- $\widetilde{CE}_1$  and  $\widetilde{CE}_2$  (shaded) blocks are clearly separated, except one term of overlap,  $\tilde{p}_n(m)s^n$ .
- $\xi_j$ ,  $j = n+1, \dots, 3n-2$ , are simple constants, and they are functions only of  $m$ .
- Block ① is determined by the coefficients of  $\widetilde{CE}_1(s, T_1)$ , and it is formed only by the elements of the matrices of the difference equation (3).
- Because of the special formation in the array and repeated step-wise limits, the shaded block shifts to the left identically as the array shapes. And ultimately it forms the basis of the nondelayed system's Routh array (see rows marked by  $s^n$  and  $s^{n-1}$ ).
- Following on the previous observation, block ② consists of the elements of  $\widetilde{CE}_2(s, T_1)$  only, the characteristic function of the nondelayed system.

TABLE 2  
Routh's array of (25) with step-wise limit for small  $T_1$  (displayed for  $n$  even).

	$\widetilde{CE}_1$	$\widetilde{CE}_2$
$s^{3n}$	$\tilde{p}_{3n}T_1^{2n}$	$\tilde{p}_{n+4}T_1^4$
$s^{3n-1}$	$\tilde{p}_{3n-1}T_1^{2n-1}$	$\tilde{p}_{n+3}T_1^3$
$s^{3n-2}$	$\xi_{3n-2,1}T_1^{2n-2}$	$\tilde{p}_{n+2}T_1^2$
$s^{3n-3}$	$\xi_{3n-3,1}T_1^{2n-3}$	$\tilde{p}_{n+1}T_1$
$\vdots$		
$s^{n+1}$	$\xi_{n+1,1}T_1$	$\tilde{p}_n$
$s^n$	$\tilde{p}_n$	$\tilde{p}_{n-2}$
$s^{n-1}$	$\tilde{p}_{n-1}$	$\tilde{p}_{n-3}$
$s^{n-2}$	$\tilde{p}_{n-2}$	$\tilde{p}_{n-4}$
$\vdots$		
$s^1$	$\tilde{p}_1$	
$s^0$	$\tilde{p}_0$	

Based on the Routh–Hurwitz criterion, the determining condition for “no root migration” is that the first column should exhibit no additional sign change as  $T_1 : 0 \rightarrow 0^+$  transition occurs. That is,  $T_1 : 0 \rightarrow 0^+$  transition for all  $m \in \mathbb{R}^+$  implies that the number of sign changes in the first column of the array remains unchanged. Block ② is independent of  $m$ , and it declares the number of unstable roots for the nondelayed system. Block ① should not add a sign change; i.e., all the  $2n+1$  elements in this block must be in sign agreement, for all  $m \in \mathbb{R}^+$ , with  $\tilde{p}_n$  (which is independent of  $m$ ). In other words, the number of sign changes in the first column should be coming from block ②, i.e., due to the unstable roots of the delay free system. And the small

delays must not change the number of unstable roots.

Since  $T_1 > 0$ , it plays no role in the signs of these elements. By enforcing the sign agreement for all  $m \in \mathbb{R}^+$ , we declare the independence of the number of sign changes in block ① from  $m$ . Together with (20) and (21), this enforcement becomes independent of the actual values of the small delays,  $\tau_1, \tau_2$ .

It is clear that block ① is created strictly from  $L(s, \tau_1, \tau_2)$  given in (4). Then the function  $L(s, \tau_1, \tau_2)$  must be stable for small delays, in order for the system (1) not to have imaginary root migration for  $\tau : \mathbf{0} \rightarrow \mathbf{0}^+$  transition. This is the “delay-stabilizability” necessary condition, or “strong stability condition.”  $\square$

*Remark 1.* “No sign change in the first column” for  $m \in [0, \infty)$  can be verified easily. The terms corresponding to  $s^j$ ,  $j = n + 1, \dots, 3n$ , in Table 2 are polynomials in  $m$ , while the term corresponding to  $s^n$  is a constant. For no sign change in block ① for  $m \in [0, \infty)$ , the two following conditions have to be satisfied simultaneously: (a) none of the polynomials in block ① has any positive real roots; (b) all the terms in block ① have to agree in sign for arbitrarily chosen positive  $m$ .

**3.2. Application to scalar case.** We next apply the above procedure to a scalar version of system (1), that is, for  $n = 1$ . The limiting form of (25) is found as

$$(30) \quad \begin{aligned} \widetilde{CE}(s, T_1, mT_1) = & (1 + a + b - c)mT_1^2s^3 \\ & + (1 + a - b + c + m(1 - a + b + c))T_1s^2 \\ & + (1 - a - b - c)s - d - f - g - h. \end{aligned}$$

Let us recall that in the coefficients of the powers of  $s$  within  $\widetilde{CE}(s, T_1, mT_1)$ , the terms with lower order of magnitude in  $T_1$ , i.e., with higher powers of  $T_1$ , are omitted in favor of the lowest powered term via earlier explained step-wise limit operations. Routh’s array for (30) is given in Table 3, again with successive deployment of step-wise limit operations.

TABLE 3  
Routh’s array of (30) with step-wise limit for small  $T_1$ .

$s^3$	$(1 + a + b - c) m T_1^2$	$(1 - a - b - c)$
$s^2$	$(1 + a - b + c + m\{1 - a + b + c\}) T_1$	$(-d - f - g - h)$
$s^1$	$(1 - a - b - c)$	①
$s^0$	$(-d - f - g - h)$	
		②

Notice that block ① has no influence on the retarded segments of the equation but only the neutral part. This block is fully determined by the characteristic function of the difference equation corresponding to (4) after Rekasius substitution, i.e.,

$$(31) \quad \begin{aligned} \widetilde{L}(s, T_1, mT_1) = & (1 + T_1s)(1 + mT_1s) - (1 + mT_1s)(1 - T_1s)a \\ & - (1 + T_1s)(1 - mT_1s)b - (1 - T_1s)(1 - mT_1s)c = (1 + a + b - c)mT_1^2s^2 \\ & + (1 + a - b + c + m\{1 - a + b + c\})T_1s + 1 - a - b - c = 0, \end{aligned}$$

while block ② consists of the coefficients of the characteristic equation for the non-delayed system. The strong stability of the difference equation requires that elements

in block ① have the same sign for all  $m \in [0, \infty)$  (i.e., they should all be positive or negative). A closer observation shows that the case with all negative elements results in a contradiction. Therefore, the following inequalities with positive elements are considered:

$$(32) \quad \begin{aligned} 1 + a + b - c &> 0, \\ 1 + a - b + c + m(1 - a + b + c) &> 0, \\ 1 - a - b - c &> 0. \end{aligned}$$

And these conditions can be reduced to

$$(33) \quad 1 + a > |b - c|, \quad 1 - a > |b + c|,$$

determining a tetrahedron as the *strong stability* domain in the coordinates of the parameters  $a$ ,  $b$ , and  $c$ ; see Figure 1 (left). These are identical to the findings of [9] based on a fundamentally different approach. If the scalar difference equation (3) (i.e., with  $n = 1$ ) were to be considered with three independent delays (simply by substituting a third delay,  $\tau_3$ , in place of  $\tau_1 + \tau_2$ ) as studied in [9], the *strong stability* condition becomes

$$(34) \quad |a| + |b| + |c| < 1,$$

which is depicted in Figure 1 (right). As expected, the shaded region in Figure 1 (right) is within that of Figure 1 (left); since the former allows the selection of three delays totally arbitrarily, it results in a more confined strong stability region.

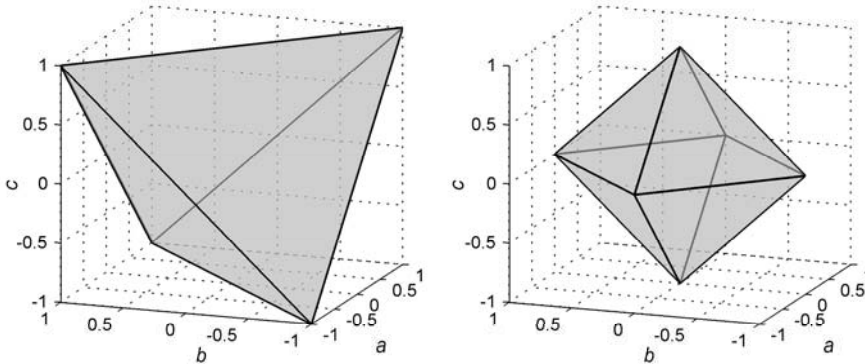


FIG. 1. *Strong stability domain of the scalar difference equation (3): left, with two cross-talking delays; right, with three independent delays.*

Inspired by the remark of an anonymous reviewer we wish to share with the reader the following example dynamics. Take the difference equation

$$(35) \quad x(t) - \exp(1) x(t - \tau) = 0,$$

which represents  $a = -\exp(1)$ ,  $b = c = 0$  in (31). Since inequalities (33) are not satisfied, as well as inequality (34), the difference equation is not strongly stable. The roots of (35) are trivially found as

$$s_1 = \frac{1}{\tau}, \quad s_{2k, 2k+1} = \frac{1}{\tau} \pm i \frac{2k\pi}{\tau}, \quad k = 1, 2, \dots$$

Apparently the transition  $\tau : \mathbf{0}^- \rightarrow \mathbf{0}^+$  brings infinitely many infinite roots *migrating* from the left-half complex plane to the right-half plane. This migration takes place over the single point on the Riemann sphere at infinity, although there does not exist a dynamics for the  $\tau = \mathbf{0}$  value exactly.

**3.3. Application to  $\mathbb{R}^2$  case.** Consider a neutral system (1) for  $n = 2$  this time. After Rekasius substitution we obtain a sixth degree characteristic function  $\widetilde{CE}(s, T_1, T_2)$  as in (13). Substituting  $T_2 = mT_1$  and, again, performing the step-wise limit operation, we arrive at a sixth degree polynomial  $\widetilde{CE}(s, T_1, mT_1)$  which is in the form of (25). Routh’s array corresponding to this polynomial is given in Table 4. We then check the delay-stabilizability condition, i.e., the requirement for no sign changes in the first column for all  $m \in [0, \infty)$ . If this condition is satisfied, the neutral system (1) is strongly stable (or delay-stabilizable). Then the system may have some finite two-dimensional region in delay space  $\tau$ , where it is asymptotically stable. It is critical to state two subtleties here: (a) If the system were not strongly stable, CTCR could declare it from the small-delay analysis shown in Table 4. That means infinitely many unstable characteristic roots of (2) would appear as  $\tau : \mathbf{0} \rightarrow \mathbf{0}^+$  transition occurs. (b) This instability may be reversed at some lower-dimensional regions in  $\tau$  space where  $\tau_1$  and  $\tau_2$  are rationally dependent (i.e., commensurate). Clearly, at these locations there is no finite two-dimensional region of potential stability but isolated one-dimensional lines; i.e.,  $\tau_1/\tau_2 = k, k$  being a rational number. It is obvious that, when the delays are rationally dependent, the system reduces to a single-delay dynamics. Corresponding delay-stabilizability conditions are valid only along the specific one-dimensional space (i.e., lines) designated by  $k$ .

Since the concept of stability robustness in  $\tau \in \mathbb{R}^{2+}$  truly aims at the detection of two-dimensional regions of asymptotic stability, we do not pursue those cases which are not strongly stable. For the strongly stable (i.e., delay-stabilizable) case, however, we continue with the steps of CTCR as given in section 2 and exhaustively determine the stability robustness picture of the system in the two-dimensional space of  $\tau \in \mathbb{R}^{2+}$ . We provide example studies for this in section 5.

TABLE 4  
Routh’s array of (25) for  $n = 2$  with step-wise limit for small  $T_1$ .

	$\widetilde{CE}_1$	$\widetilde{CE}_2$
$s^6$	$\tilde{p}_6 m^2 T_1^4$	$\tilde{p}_4(m) T_1^2$
$s^5$	$\tilde{p}_5(m) T_1^3$	$\tilde{p}_3(m) T_1$
$s^4$	$(\tilde{p}_5(m) \tilde{p}_4(m) - \tilde{p}_6(m) \tilde{p}_3(m)) T_1^2$	$\tilde{p}_2$
$s^3$	$(\tilde{p}_5(m) \tilde{p}_4(m) \tilde{p}_3(m) - \tilde{p}_6(m) \tilde{p}_3^2(m) - \tilde{p}_5(m) \tilde{p}_2(m)) T_1$	$\tilde{p}_1$
$s^2$	$\tilde{p}_2$	$\tilde{p}_0$
$s^1$	$\tilde{p}_1$	
$s^0$	$\tilde{p}_0$	

Please notice that the system (1) does not have to be stable for  $\tau = \mathbf{0}$  for the deployment of CTCR. If, however, this delay-free system is stable, coefficients in block



②, i.e.,  $\tilde{p}_2, \tilde{p}_1, \tilde{p}_0$ , will agree in sign. The CTCR procedure pursues regardless, in order to declare the stable regions exhaustively. An important nuance is that the “delay-stabilizable system” may end up having no stable regions in  $\tau \in \mathbb{R}^{2+}$ . This is clearly the reflection of the “necessary condition” attribute.

**4. Numerical cross-validation of the results of CTCR.** In order to verify the results of CTCR, we use a numerical method. It determines the regions in the delay domain, where  $NU$  remains fixed, by using a point-wise analysis. This method deploys a quasi-polynomial mapping based rootfinder (QPmR) technique [24], and it is based on the direct computation of the rightmost characteristic roots at points of a sufficiently dense grid, which is spread over the domain of interest  $[0, \tau_{1,\max}] \times [0, \tau_{2,\max}]$ . Let us remark that besides the QPmR algorithm, there exist other approaches for approximating the dominant roots of TDS, using the system solution operator [3], [7] or the infinitesimal generator procedure [4]. The QPmR algorithm was chosen here, because it computes all the roots within the same accuracy, and it is utilized in determining the number of unstable roots.

The procedure of the QPmR algorithm is as follows. Substituting  $s = \beta + \omega i$ , the characteristic equation (2) is first split into real and imaginary parts:

$$(36) \quad R(\beta, \omega) = \Re(CE(\beta + \omega i, \tau_1, \tau_2)) = 0,$$

$$(37) \quad I(\beta, \omega) = \Im(CE(\beta + \omega i, \tau_1, \tau_2)) = 0.$$

Intersection points of the curves described by (36) and (37) in space  $(\beta, \omega)$  reveal the characteristic roots of (2). The numerical problem reduces to determine these intersection points exhaustively. Mapping of these curves over a region  $\mathcal{D} = [\beta_{\min}, \beta_{\max}] \times [0, \omega_{\max}]$  is done numerically by applying a contour plotting algorithm; see, e.g., [23] and the references therein (available, e.g., in MATLAB function *contour*).

Since the QPmR performs the rootfinding task over a prescribed search region, it is crucial to determine its boundaries to guarantee that any unstable roots would be captured. For determining the boundaries  $\beta_{\min}, \beta_{\max}$ , we use the following adaptation rule. Consider one of the delays being fixed, e.g.,  $\tau_2$ , and  $s_r$  is the position of the rightmost root for delays  $(\tau_1, \tau_2)$ . When increasing the other delay  $\tau_1$  by  $\Delta\tau_1$ , the boundaries are arranged according to the following rule, which considers the root sensitivities to the delay:

$$(38) \quad \begin{aligned} \beta_{\max} &= \Re \left( s_r + \left[ \frac{\partial s}{\partial \tau_1} \Big|_{s=s_r} \right] \Delta\tau_1 \right) + \Delta\beta, \\ \beta_{\min} &= \min \left[ 0, \Re \left( s_r + \left[ \frac{\partial s}{\partial \tau_1} \Big|_{s=s_r} \right] \Delta\tau_1 \right) - \Delta\beta \right], \end{aligned}$$

where  $\Delta\beta$  determines the width of the region, preferably  $\Delta\beta > |\Re[\frac{\partial s}{\partial \tau_1} \Big|_{s=s_r}] \cdot \Delta\tau_1|$ ,  $\Delta\beta > 0$ . Naturally, the same adaptation rule is applied when  $\tau_1$  is fixed and  $\tau_2$  increases by  $\Delta\tau_2$ . If one starts the computation at the origin,  $\tau = \mathbf{0}$ , the characteristic function  $CE(s, \tau)$  in (2) results in the first  $s_r$ , which iteratively primes (38) for the increased values of delays.

The upper bound  $\omega_{\max}$  of the search region  $\mathcal{D}$  is selected following the root distribution property of [2]

$$(39) \quad \omega_{\max} \gg \min \left( \Omega, \frac{\pi}{n(\tau_1 + \tau_2)} \right),$$

where  $\Omega \gg s_I$  and  $s_I$  is the root of (2) with the largest imaginary part for  $\tau = 0$ .

In this work, we use the QPmR method simply to validate the findings of CTCR. Therefore, the selection of the search region  $\mathcal{D}$ , especially at the expected stability switching locations, is relatively easy to determine. For this, we utilize the information gained from the CTCR algorithm and select sufficiently large  $\omega_{\max} > \omega_c$ , where  $\omega_c$  is the root crossing frequency for a point on either the kernel or an offspring for  $\tau \neq 0$ .

In summary, the QPmR procedure determines the exact position of the rightmost roots for all points of the mesh grid in the delay domain  $[0, \tau_{1,\max}] \times [0, \tau_{2,\max}]$ . This allows us to construct the regions with an equal number of unstable roots ( $NU$ ). Example cases are provided for clarity next.

## 5. Example case studies.

**5.1. Example 1.** Consider a scalar neutral system (1) with  $a = 0.5$ ,  $b = -0.35$ ,  $c = 0.31$ ,  $d = -1.5$ ,  $f = 2$ ,  $g = -2$ ,  $h = -3$ . The first step is to check the strong stability, which is given for the scalar case by inequalities (33). Both inequalities are satisfied; thus the neutral system is strongly stable (i.e., the system is delay-stabilizable). Applying the CTCR algorithm step by step, we obtain the stability picture given in Figure 2. The color of the curves (when viewed in color) distinguishes the root tendency with respect to the delay  $\tau_1$ ; red  $RT^{\tau_1} = 1$ , blue  $RT^{\tau_1} = -1$ . The number of unstable roots in each region,  $NU$ , is also shown sparingly. Obviously, the stable regions are marked by  $NU = 0$ . As can be seen, such a region in Figure 2 consists of two branches; one stable zone covers a narrow region close to axis  $\tau_2$ , while the boundaries of the other one, which is wider, tend to infinity parallel to the line  $\tau_1 = \tau_2$ .

In Figure 3, we present the results obtained by the QPmR algorithm. To create this figure, the algorithm was used at sufficiently dense grid points in the  $(\tau_1, \tau_2)$  plane, here  $150 \times 150$  points. One can see the regions with equal  $NU$  (number of

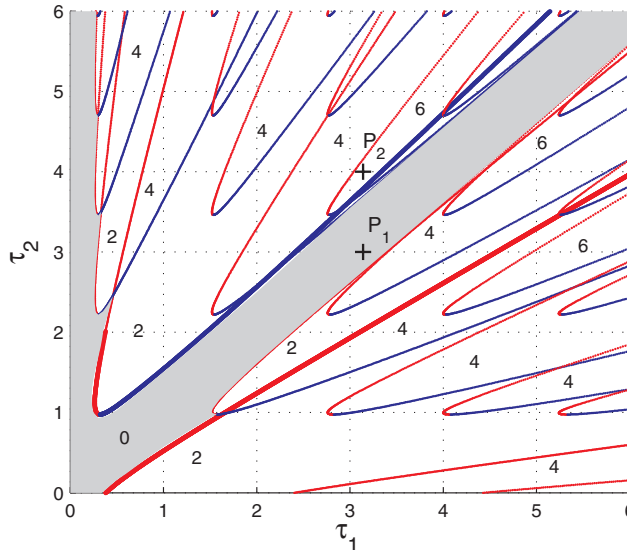


FIG. 2. Stability picture obtained using CTCR for Example 1; kernel (thick), offspring (thin), stable region (shaded).

unstable roots), as they are shaded at different levels. The results obtained by CTCR and QPmR are in complete agreement.

In Figure 4 the spectra of the neutral system and its associated difference equation are computed using QPmR. They are shown for two selected points in the delay domain,  $P_1(\pi, 3)$  and  $P_2(\pi, 4)$ ; see also Figure 2. The number of unstable roots agrees with those found by CTCR (as declared in Figure 2), i.e., for  $P_1$ ,  $NU = 0$  and for  $P_2$ ,  $NU = 6$ .

**5.2. Example 2.** Consider a neutral equation (1), with  $n = 2$ , with the matrices

$$\mathbf{A} = \begin{bmatrix} -0.7 & 0 \\ 0 & -0.1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & -0.4 \\ -0.1 & -0.2 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 0 & -0.9 \\ 0 & -0.3 \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} -0.5 & 0 \\ 0 & -0.3 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} 0 & -0.1 \\ 0.4 & 0 \end{bmatrix},$$

$$\mathbf{G} = \begin{bmatrix} -0.7 & -0.1 \\ 0 & -0.4 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} -0.5 & 0.2 \\ 0.1 & -0.3 \end{bmatrix}.$$

The corresponding characteristic equation is

$$(40) \quad \begin{aligned} CE(s, \tau_1, \tau_2) = & s^2 + 0.8s + 0.13 + (0.8s^2 + 1.06s + 0.3)e^{-\tau_1 s} \\ & + (0.07s^2 + 0.26s + 0.15)e^{-2\tau_1 s} + (0.2s^2 + 0.16s - 0.07)e^{-\tau_2 s} \\ & + (-0.04s^2 + 0.15s + 0.04)e^{-2\tau_2 s} + (0.44s^2 + 1.44s + 0.42)e^{-(\tau_1 + \tau_2)s} \\ & + (0.21s^2 + 0.5s + 0.41)e^{-(2\tau_1 + \tau_2)s} + (0.21s + 0.28)e^{-2(\tau_1 + \tau_2)s} \\ & + (-0.09s^2 + 0.49s + 0.04)e^{-(\tau_1 + 2\tau_2)s} = 0. \end{aligned}$$

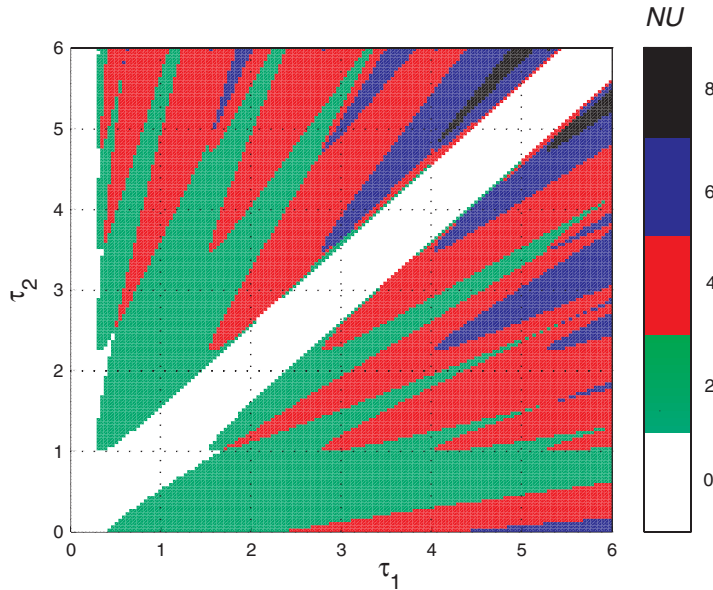


FIG. 3. Number of unstable roots ( $NU$ ) for Example 1 determined using the QPmR algorithm.

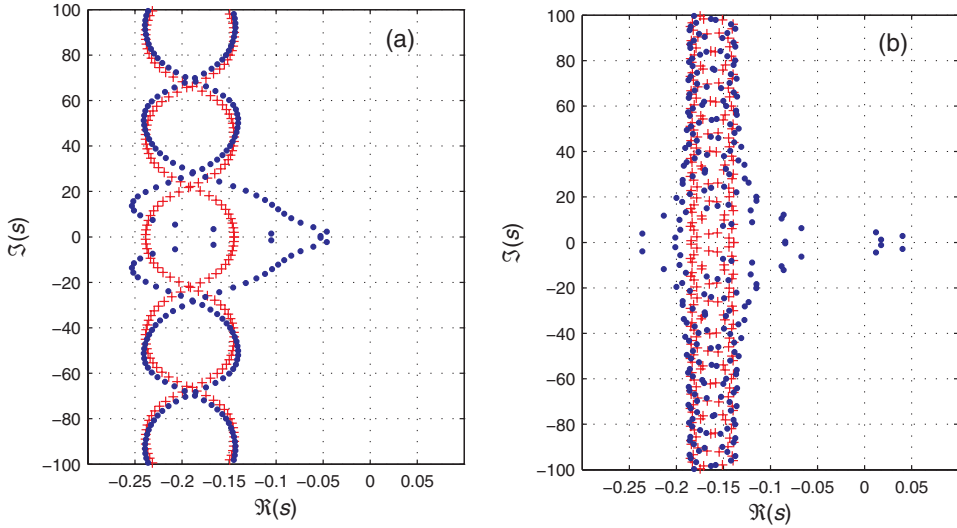


FIG. 4. Spectra of the neutral system for Example 1 (dots) and the associated difference equation (crosses) for (a)  $P_1(\pi, 3)$  and (b)  $P_2(\pi, 4)$ .

The delay stabilizability is checked first as described in subsection 3.1. The difference equation is stable, as small-delay analysis reveals no sign change on the first column of the array described in Table 2. So the difference equation is strongly stable, and the system is delay-stabilizable. CTCR continues to determine the robust stability regions with respect to the uncertain delays. The results of both CTCR and QPmR are shown in Figures 5 and 6, respectively. As can be seen, only one stability region (with  $NU = 0$ ) appears over the selected delay regions. Its shape is determined by the kernel and the offspring curves. QPmR declares the dominant characteristic roots of which the real parts are displayed in Figure 6. The stability boundary is marked by a thick black curve. Both methods again concur.

**6. Conclusions.** A general class of neutral systems with two time delays is studied for the stability robustness against delay uncertainties. The delays appear both in the neutral and retarded parts, and in the cross-talking format, which is a challenging feature. First, the method of cluster treatment of characteristic roots (CTCR) is employed to determine the stability domain in the delays. As the main contribution of the paper, we demonstrate that the CTCR procedure reveals a very critical feature, called the “delay stabilizability,” as a by-product. It is based on small-delay analysis of the dynamics. Even though the criterion is designed for a general system with  $n$ -dimensional matrices, special attention is paid to the scalar and two-dimensional cases for ease of conveyance. Furthermore, an alternative numerical algorithm is deployed for cross-validating the findings of CTCR. This algorithm is based on precisely computing the rightmost characteristic roots of the system for given delay values. Example cases are presented to display the concurrence of the two methods.

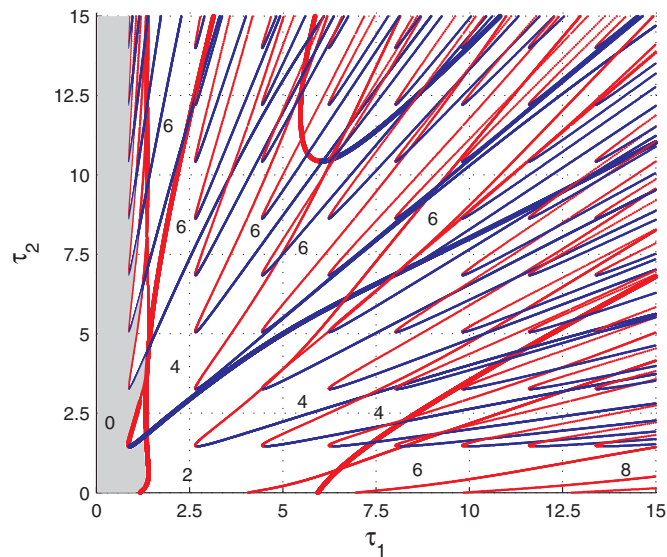


FIG. 5.  $NU$  distribution obtained by CTCR for Example 2; kernel (thick), stable region (shaded).

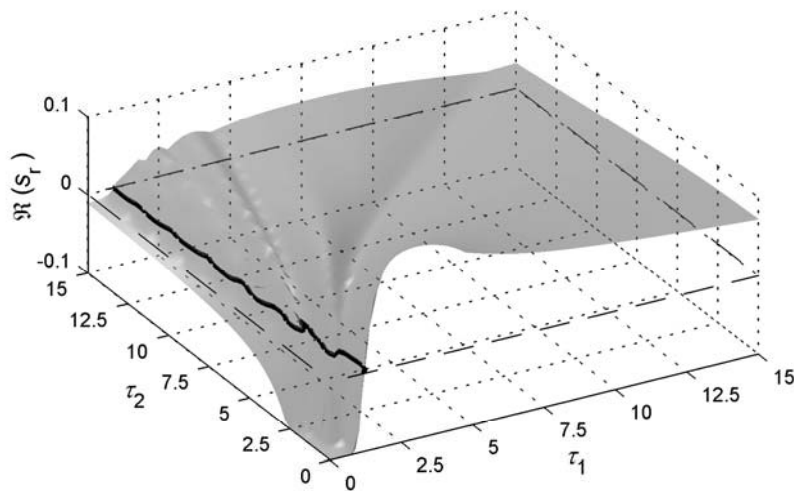


FIG. 6. Real part of the rightmost root  $s_r$  of the spectrum for Example 2 as  $(\tau_1, \tau_2)$  vary. The thick line is the stability boundary in the delay space.

REFERENCES

- [1] C. E. AVELAR AND J. K. HALE, *On the zeros of exponential polynomials*, J. Math. Anal. Appl., 73 (1980), pp. 434–452.
- [2] R. BELLMAN AND K. COOKE, *Differential Difference Equations*, Academic Press, New York, 1963.
- [3] D. BREDÁ, S. MASET, AND R. VERMIGLIO, *Computing the characteristic roots for delay differential equations*, IMA J. Numer. Anal., 24 (2004), pp. 1–19.
- [4] D. BREDÁ, *Solution operator approximation for delay differential equation characteristic roots computation via Runge–Kutta methods*, Appl. Numer. Math., 56 (2006), pp. 305–317.

- [5] D. BREDÁ, S. MASET, AND R. VERMIGLIO, *Efficient computation of stability charts for linear time delay systems*, in Proceedings of ASME-IDETC/CIE 2005, Long Beach, CA, 2005.
- [6] L. E. EL'SGOL'TS AND S. B. NORKIN, *Introduction to the Theory and Application of Differential Equations with Deviating Arguments*, Academic Press, New York, 1973.
- [7] K. ENGELBORGH AND D. ROOSE, *On stability of LMS-methods and characteristic roots of delay differential equations*, SIAM J. Numer. Anal., 40 (2002), pp. 629–650.
- [8] K. GU, S. I. NICULESCU, AND J. CHEN, *On stability crossing curves for general systems with two delays*, J. Math. Anal. Appl., 311 (2005), pp. 231–253.
- [9] J. K. HALE AND S. M. VERDUYN LUNEL, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.
- [10] J. K. HALE AND S. M. VERDUYN LUNEL, *Effects of small delays on stability and control*, in Operator Theory and Analysis, Oper. Theory Adv. Appl. 122, Birkhäuser, Basel, 2001, pp. 275–301.
- [11] J. K. HALE AND S. M. VERDUYN LUNEL, *Strong stabilization of neutral functional differential equations*, IMA J. Math. Control Inform., 19 (2002), pp. 5–23.
- [12] B. C. KUO, *Automatic Control Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
- [13] J. H. MATHEWS, AND R. W. HOWELL, *Complex Analysis for Mathematics and Engineering*, William C. Brown Publishers, Dubuque, IA, 1996.
- [14] W. MICHIELS, K. ENGELBORGH, D. ROOSE, AND D. DOCHAIN, *Sensitivity to infinitesimal delays in neutral equations*, SIAM J. Control Optim., 40 (2001), pp. 1134–1158.
- [15] W. MICHIELS AND T. VYHLÍDAL, *An eigenvalue based approach for the stabilization of linear time-delay systems of neutral type*, Automatica, 41 (2005), pp. 991–998.
- [16] N. OLGAC AND R. SIPAHI, *Direct method for analyzing the stability of neutral type LTI-time delayed systems*, Automatica, 40 (2004), pp. 847–853.
- [17] N. OLGAC AND R. SIPAHI, *The cluster treatment of characteristic roots and the neutral type time-delayed systems*, ASME J. Dynam. Systems Measurement Control, 127 (2005), pp. 88–97.
- [18] Z. V. REKASIUS, *A stability test for systems with delays*, in Proceedings of the Joint Automatic Control Conference, 1980, Paper TP9-A.
- [19] R. SIPAHI AND N. OLGAC, *Degenerate cases in using the direct method*, ASME J. Dynam. Systems Measurement Control, 125 (2003), pp. 194–201.
- [20] R. SIPAHI AND N. OLGAC, *Complete stability map of third order LTI multiple time-delay systems*, Automatica, 41 (2005), pp. 1413–1422.
- [21] R. SIPAHI AND N. OLGAC, *A unique methodology for the stability robustness of multiple time delay systems*, Systems Control Lett., 55 (2006), pp. 819–825.
- [22] R. SIPAHI, N. OLGAC, AND D. BREDÁ, *Complete stability map of neutral type first order - two time delay systems*, in Proceedings of ACC 2007, New York, 2007, pp. 4933–4938.
- [23] W. V. SNYDER, *Algorithm 531: Contour plotting [J6]*, ACM Trans. Math. Software, 4 (1978), pp. 290–294.
- [24] T. VYHLÍDAL AND P. ZÍTEK, *Quasipolynomial mapping based rootfinder for analysis of time delay systems*, in Proceedings of the IFAC Workshop on Time-Delay Systems (TDS'03), Rocquencourt, France, 2003.

## DESIGN OF POSITIVE LINEAR OBSERVERS FOR POSITIVE LINEAR SYSTEMS VIA COORDINATE TRANSFORMATIONS AND POSITIVE REALIZATIONS\*

JUHOON BACK<sup>†</sup> AND ALESSANDRO ASTOLFI<sup>‡</sup>

**Abstract.** Two new approaches to designing positive linear observers for positive linear systems are proposed. The first one employs a coordinates transformation and the second relies on the theory of positive realization. These approaches allow one to enlarge the class of positive systems that admit positive linear observers. New results on compartmental systems are also presented.

**Key words.** positive systems, positive observers, positive realization

**AMS subject classifications.** 15A48, 93B07, 93B51

**DOI.** 10.1137/060663891

**1. Introduction.** A positive linear system is a linear system whose trajectories originating from the nonnegative orthant evolve therein for all positive times. Since there are many physical systems whose state variables represent some quantities that should be nonnegative, such as density and population, a lot of researchers have studied this class of systems; see, for example, [13, 16, 19, 20, 23, 25, 27, 28, 29] and the references therein.

In this paper we are concerned with the observer design problem for positive linear systems. It should be mentioned that for general linear systems the observer design problem has been completely solved; i.e., there exists a (Luenberger) observer if and only if the system is detectable [22]. However, in this case, the state estimate may not belong to the nonnegative orthant for some time interval even if the actual state of the system remains therein. This means that the classical Luenberger observer, if used for estimating the state of a positive system, may generate a meaningless estimate for the actual state, which is always nonnegative. Thus, it is natural to develop observers for positive systems which always produce meaningful information on the system states.

Motivated by this, some results have been published. For example, results on positive linear observers for compartmental systems [32] (see section 2 for the definition) and positive systems [9] have been developed. Although these papers provide checkable conditions for the existence of a positive observer, a system theoretic existence criterion, such as detectability for general linear systems, is not available yet.

The approaches of the aforementioned works share a fundamental restriction: they use the Luenberger (identity) observer to find conditions under which a positive (or compartmental) linear observer exists. This means that the observer is designed in the original coordinates system and is of the same dimension as the system *to be observed*. However, as can be seen from the examples in the paper, this framework severely

---

\*Received by the editors June 28, 2006; accepted for publication (in revised form) April 26, 2007; published electronically January 25, 2008. This work was partially supported by the Leverhulme Trust.

<http://www.siam.org/journals/sicon/47-1/66389.html>

<sup>†</sup>Department of Mechanical Engineering, Korea University, Anam-Dong, Seongbuk-Gu, Seoul, 136-713 Korea (backhoon@korea.ac.kr). This work was performed while this author was at Imperial College.

<sup>‡</sup>Department of Electrical and Electronic Engineering, Imperial College, London SW7 2AZ, UK, and Dipartimento di Informatica, Sistemi e Produzione, Università di Roma Tor Vergata, Via del Politecnico 1, 00133 Roma, Italy (a.astolfi@ic.ac.uk).

restricts the class of systems that admit (Luenberger-type) positive linear observers.

The main objective of this paper is to introduce new approaches for positive linear observers and to provide conditions on their existence. The first idea is to employ a coordinate transformation. It is motivated by the fact that even though a linear system is not a positive system in one coordinate system, it can be positive in another coordinate system. The second idea is to formulate the problem as a positive realization problem [2, 5, 12, 24]. It is based on the fact that a positive linear observer can be regarded as a transfer function which has a positive realization. This method allows us to circumvent the inverse eigenvalue problem [10, 21], which should be solved if we fix the dimension of the observer equal to that of the system. In other words, in the positive realization approach, the dimension of the observer is also a design parameter.

The contribution of this paper (and the organization) is summarized as follows.

1. Luenberger-type observers (section 3): some existing results are examined and corrected to establish basic results, especially on compartmental systems (section 3.2).
2. Positive observers using a coordinates transformation (section 4): two necessary conditions (section 4.2) and one sufficient condition (section 4.3) on the spectrum are provided.
3. Positive observers via positive realization (section 5): a sufficient condition which is less restrictive than other approaches is obtained and applied to compartmental systems (section 5.2).

Several illustrative examples are presented to show the effectiveness of the approaches.

**2. Preliminaries and notation.** Let  $\mathbb{R}$  ( $\mathbb{C}$ , resp.) be the set of real numbers (complex numbers, resp.) and  $\mathbb{R}^n$  the set of  $n$ -tuples with each component belonging to  $\mathbb{R}$ . Define  $\mathbb{R}_+ := [0, \infty)$ ,  $\mathbb{R}_- := (-\infty, 0)$ .  $\mathbb{R}_+^n$  is the set of  $n$ -tuples of  $\mathbb{R}_+$  and  $\text{int } \mathbb{R}_+^n$  is the interior of  $\mathbb{R}_+^n$ . Let  $\mathbb{R}^{n \times m}$  ( $\mathbb{R}_+^{n \times m}$ ) be the set of  $n \times m$  matrices whose elements are in  $\mathbb{R}$  ( $\mathbb{R}_+$ , resp.). Given  $A \in \mathbb{R}^{n \times m}$ ,  $A^t$  is the transpose of  $A$ , and  $\sigma(A)$  the set of all eigenvalues of  $A$ . A matrix  $A$  is said to be singular if  $0 \in \sigma(A)$ .

For  $\lambda \in \mathbb{C}$ ,  $\text{Re } \lambda$  is the real part of  $\lambda$ .  $\mathbb{C}_- := \{\lambda \in \mathbb{C} \mid \text{Re } \lambda < 0\}$ .

Given  $A, B \in \mathbb{R}^{n \times m}$ ,  $A > B$  ( $A \geq B$ , resp.) if and only if  $A_{ij} > B_{ij}$  ( $A_{ij} \geq B_{ij}$ , resp.)  $\forall 1 \leq i \leq n, 1 \leq j \leq m$ .  $A < B$  and  $A \leq B$  are defined similarly.

A matrix  $A \in \mathbb{R}^{n \times n}$  is called a Metzler matrix if  $A_{ij} \geq 0 \forall i \neq j$  (i.e.,  $\exists \alpha \in \mathbb{R}_+$  such that  $A + \alpha I \in \mathbb{R}_+^{n \times n}$ ). A matrix  $A$  is said to be inverse-positive if it is invertible and  $A^{-1} \geq 0$ .

Given  $A_1, \dots, A_n$  where  $A_i \in \mathbb{R}^{n_i \times m_i}$ ,  $\text{diag}\{A_1, \dots, A_n\}$  is a block diagonal matrix whose  $(i, i)$ th block is  $A_i$ , while the nondiagonal blocks are all zero matrices. The matrix  $1_{n \times m} \in \mathbb{R}^{n \times m}$  ( $0_{n \times m}$ , resp.) stands for the matrix whose elements are all 1 (0, resp.).

A compartmental matrix  $A \in \mathbb{R}^{n \times n}$  is a Metzler matrix with the property  $\sum_{i=1}^n a_{ij} \leq 0 \forall j \in \{1, \dots, n\}$ .

A matrix  $A \in \mathbb{R}^{n \times n}$  is said to be reducible if there exists a permutation matrix  $P \in \mathbb{R}^{n \times n}$  such that

$$PAP^t = \begin{bmatrix} \overline{A}_{11} & 0 \\ \overline{A}_{21} & \overline{A}_{22} \end{bmatrix},$$

where  $\overline{A}_{11}$ ,  $\overline{A}_{22}$  are square matrices.  $A$  is said to be irreducible if  $A$  is not reducible.

The following properties are well known.

**THEOREM 2.1** (see [6, 17, 23]). *Let  $A \in \mathbb{R}^{n \times n}$  be a Metzler matrix.*

1. *There exist a real number  $\lambda_{\max}(A)$  and a vector  $v_{\max} \in \mathbb{R}_+^n$  such that (1)  $Av_{\max}$*



$= \lambda_{\max}(A)v_{\max}$  and (2) if  $\lambda \neq \lambda_{\max}(A)$  is any other eigenvalue of  $A$ , then  $\operatorname{Re} \lambda < \lambda_{\max}(A)$ .

2. If  $A$  is irreducible, then the eigenvalue with maximal real part  $\lambda_{\max}(A)$  has (algebraic) multiplicity one and there exists a positive eigenvector  $v_{\max}$  associated to  $\lambda_{\max}(A)$ .
3. If  $A$  is a compartmental matrix, then  $\sigma(A) \subseteq \{\lambda \in \mathbb{C} \mid \operatorname{Re} \lambda < 0 \text{ or } \lambda = 0\}$ .

Recall that  $\lambda_{\max}(A)$  is often called the Perron–Frobenius eigenvalue.

Now we recall the definition of positive linear systems (with positive inputs and positive outputs) and one of their characterizations.

DEFINITION 2.2 (see [23, 32]). Consider a continuous-time linear dynamic system

$$(2.1) \quad \begin{aligned} \dot{x} &= Ax + Bu, & x(t_0) &= x_0, \\ y &= Cx + Du, \end{aligned}$$

where  $x \in \mathbb{R}^n$  is the system state,  $u \in \mathbb{R}^p$  the system input, and  $y \in \mathbb{R}^q$  the system output. This system is called a positive linear system if  $\forall x_0 \in \mathbb{R}_+^n$  and  $\forall u(t) \in \mathbb{R}_+^p$ ,  $\forall t \geq 0$ , we have  $x(t) \in \mathbb{R}_+^n$  and  $y(t) \in \mathbb{R}_+^q$   $\forall t \geq 0$ .

PROPOSITION 2.3 (see [23, 32]). A linear dynamic system of the form (2.1) is a positive linear system if and only if  $B \geq 0$ ,  $C \geq 0$ ,  $D \geq 0$ , and  $A$  is a Metzler matrix.

We close this section with a result on the shift of  $\lambda_{\max}(A)$  which follows trivially from Theorem 2.9 of [7]. (See also [26] for a proof of Brauer’s theorem.)

THEOREM 2.4 (adapted from Theorem 2.9 of [7]). Let  $A \in \mathbb{R}^{n \times n}$  be an irreducible Metzler matrix and let  $v_{\max}$  be the eigenvector associated to  $\lambda_1 := \lambda_{\max}(A)$ . Also, let  $v \in \mathbb{R}^n$ . Then

$$\sigma(A - v_{\max}v^t) = (\sigma(A) \setminus \{\lambda_1\}) \cup \{\lambda_1 - v^t v_{\max}\}.$$

**3. Luenberger-type positive linear observers.** In this section, Luenberger-type positive linear observers are discussed. At first, the problem is defined and a characterization is provided. We also discuss the problem for a special class of positive systems whose maximal real eigenvalue is less than or equal to zero and present a necessary and sufficient condition on the existence of observers.

**3.1. Problem formulation and basic results.** Consider unforced multioutput linear systems,

$$(3.1) \quad \begin{aligned} \dot{x} &= Ax, & x(t_0) &= x_0, \\ y &= Cx, \end{aligned}$$

where  $A \in \mathbb{R}^{n \times n}$ ,  $C \in \mathbb{R}^{q \times n}$ . Throughout the paper, we assume that  $C$  has rank  $q$  and that system (3.1) is a positive linear system (i.e.,  $A$  is Metzler and  $C \geq 0$ ).

As far as the state observation problem is concerned, it is natural to consider the Luenberger-type observer (linear identity observer)

$$(3.2) \quad \dot{\hat{x}} = A\hat{x} + K(y - C\hat{x}),$$

because the observer problem for linear systems has been completely solved in this form [22].

As mentioned in section 1, we would like to design an observer which estimates the states of the system and is itself a positive system whose input is the system’s output  $y$  and whose output is the state estimate  $\hat{x}$ . In other words, we require (3.2) to satisfy

1.  $K \in \mathbb{R}_+^{q \times n}$ ;
2.  $A - KC$  is a Metzler and Hurwitz matrix.

For the single output case ( $q = 1$ ), a necessary and sufficient condition has been stated in [9] whose proof requires the property shown below.

LEMMA 3.1 (adapted from Problem 6.9.1 in [23]). *Let  $A$  be a Metzler matrix. Let  $Z_1, Z_2 \in \mathbb{R}_+^{n \times n}$  be such that  $Z_1 \leq Z_2$ . If  $A - Z_1$  and  $A - Z_2$  are Metzler, then  $\lambda_{\max}(A - Z_1) \geq \lambda_{\max}(A - Z_2)$ .*

With this lemma, it is possible to establish the following result. A slight modification of the proof provided in [9] proves the assertion.

THEOREM 3.2 (see [9]). *Given a Metzler matrix  $A \in \mathbb{R}^{n \times n}$  and a nonzero matrix  $C \in \mathbb{R}_+^{1 \times n}$ , let  $K^* = [k_1^* \ \dots \ k_n^*]^t \in \mathbb{R}_+^n$  be a nonnegative matrix defined as follows. If there exists an index  $i$  such that  $c_i \neq 0$ , while  $c_j = 0$  for  $j \neq i$ , then*

$$k_i^* > \frac{a_{ii}}{c_i},$$

$$k_j^* = \frac{a_{ji}}{c_i} \text{ for } j \neq i.$$

*Otherwise,*

$$k_j^* = \min_{i \neq j, c_i \neq 0} \left\{ \frac{a_{ji}}{c_i} \right\} \quad \forall j.$$

*Then there exists a nonnegative matrix  $K \in \mathbb{R}_+^n$  such that  $A - KC$  is a Hurwitz and Metzler matrix if and only if  $\lambda_{\max}(A - K^*C) < 0$ .*

As mentioned in [9], one can easily check the condition for the existence of positive observers of the form (3.2). Moreover, if  $C$  contains more than two nonzero elements, then Theorem 3.2 gives a gain matrix  $K^*$  such that  $\lambda_{\max}(A - K^*C) \leq \lambda_{\max}(A - KC) \forall K \in \mathbb{R}_+^{n \times 1}$  making  $A - KC$  Metzler; i.e.,  $K^*$  pushes the maximum real eigenvalue toward  $-\infty$  as far as possible.

**3.2. Result on a class of positive systems.** Consider a positive linear system (3.1) with

$$(3.3) \quad 0 \in \sigma(A) \subset \{\lambda \in \mathbb{C} \mid \operatorname{Re}(\lambda) < 0\} \cup \{\lambda = 0\}.$$

Note that the matrix  $A$  is Metzler and its spectrum satisfies the condition (3.3), but it is not in general a compartmental matrix. This class of systems has been dealt with in [19] and contains compartmental systems (see [1, 16] and the references therein). In particular, the paper [19] considers the stabilization problem of single input positive linear systems. As far as the observer design problem is concerned, [32] characterizes systems that admit positive linear observers.

We provide a complement to the main result of [32, Theorems 3.13 and 4.12], especially for the reducible case. Since the class of systems in this subsection covers compartmental systems, the result of this section is a new (and corrected) version of Theorem 3.13 of [32], and its dual is an extension of [19] to the multi-input case.

To remove any notational ambiguity, we use  $[A]_{ij}$  for the  $(i, j)$ th entry of  $A$  (this notation is confined to this subsection). We note that for a compartmental matrix  $A$ , if  $A - B$  is Metzler for some matrix  $B \in \mathbb{R}_+^{n \times n}$ , then  $A - B$  is compartmental. This property is used several times in this subsection.

Following the same procedure of [19] with a transformation  $T = \operatorname{diag}\{t_1, \dots, t_n\}$ ,  $t_i > 0$ , and a suitable permutation matrix  $P$ ,  $A$  is assumed to have the form (3.4)

shown below:

$$(3.4) \quad A := \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{\nu 1} & A_{\nu 2} & \cdots & A_{\nu \nu} \end{bmatrix}$$

for  $\nu = 1$  ( $A$  is irreducible) or  $\nu \in \{2, \dots, n\}$  ( $A$  is reducible), where the block matrices  $A_{ij}$  are such that

1.  $A_{ii}$ ,  $1 \leq i \leq \nu$ , is a square irreducible *compartmental* matrix;
2.  $A_{ij} \geq 0$  if  $i \neq j$ .

Let  $d_i$  be the dimension of the matrix  $A_{ii}$ , and let  $1 \leq s_1 < \dots < s_m$  be such that  $A_{s_i s_i}$  is a singular block in  $A$ . Note that if  $i \notin \{s_1, \dots, s_m\}$ ,  $A_{ii}$  is Hurwitz. Since the matrices  $A_{ii}$  are irreducible, 0 is a simple eigenvalue of each  $A_{s_i s_i}$  and hence  $m$  coincides with the algebraic multiplicity of 0 as an eigenvalue of  $A$ .

The matrices  $C$  and  $K$  (obtained, in turn, by means of the transformation  $T$  and the permutation  $P$ ) can be block-partitioned as follows:

$$(3.5) \quad C := \begin{bmatrix} C_{11} & \cdots & C_{1\nu} \\ \vdots & \ddots & \vdots \\ C_{q1} & \cdots & C_{q\nu} \end{bmatrix},$$

where  $C_{ij} \in \mathbb{R}_+^{1 \times d_j}$ ,  $1 \leq i \leq q$ ,  $1 \leq j \leq \nu$  ( $C_{ij}$  is a row vector with dimension compatible with  $A_{jj}$ ), and

$$(3.6) \quad K := \begin{bmatrix} K_{11} & \cdots & K_{1q} \\ \vdots & \ddots & \vdots \\ K_{\nu 1} & \cdots & K_{\nu q} \end{bmatrix},$$

where  $K_{ij} \in \mathbb{R}_+^{d_i \times 1}$ ,  $1 \leq i \leq \nu$ ,  $1 \leq j \leq q$ .

From this decomposition, we obtain the following result.

**LEMMA 3.3.** *Suppose system (3.1) satisfies (3.3). If the system admits a positive linear observer of the form (3.2), then the algebraic multiplicity  $m$  of the zero eigenvalue of  $A$  is at most  $q$ .*

*Proof.* The existence of a positive linear observer implies that there exists  $K \in \mathbb{R}_+^{n \times q}$ , which renders the matrix  $A - KC$  Metzler (hence, compartmental) and Hurwitz. Since  $A$  is lower block triangular, it is necessary that all blocks  $A_{ij} - \sum_{l=1}^q K_{il} C_{lj}$  for  $j > i$  be zero, namely,

$$(3.7) \quad K_{il} C_{lj} = 0, \quad j > i, \quad l = 1, \dots, q.$$

Then necessarily the diagonal block  $A_{s_i s_i} - \sum_{\mu=1}^q K_{s_i \mu} C_{\mu s_i}$  is compartmental and Hurwitz  $\forall i = 1, \dots, m$ . Thus, there exist  $\eta_1, \dots, \eta_m \in \{1, \dots, q\}$  such that

$$(3.8) \quad K_{s_l \eta_l} C_{\eta_l s_l} \neq 0, \quad l = 1, \dots, m.$$

This relation with  $l = 1$  yields  $K_{s_1 \eta_1} \neq 0$ . This fact and (3.7) imply  $C_{\eta_1 s_2} = 0, \dots, C_{\eta_1 s_m} = 0$ . Thus, it follows from (3.8) that  $\eta_1 \neq \eta_2, \eta_1 \neq \eta_3, \dots, \eta_1 \neq \eta_m$ . Repeating this argument for  $l = 2, \dots, m$ , we have

$$(3.9) \quad \eta_i \neq \eta_j, \quad i \neq j, \quad 1 \leq i, j \leq m,$$

which proves the assertion.  $\square$

Now, we state the main result of this subsection.

**THEOREM 3.4.** *There exists  $K \in \mathbb{R}_+^{n \times q}$  such that system (3.2) is a positive linear observer if and only if there exist  $m$  numbers  $\eta_1, \dots, \eta_m$  with  $\eta_i \leq q$ ,  $i = 1, \dots, m$ , and  $\eta_i \neq \eta_j$  if  $i \neq j$ , and  $m$  numbers  $\xi_1, \dots, \xi_m$  with  $\xi_i \in \{1, \dots, d_{s_i}\}$ ,  $1 \leq i \leq m$ , such that for each  $i \in \{1, \dots, m\}$  the following conditions hold:*

1.  $C_{\eta_i s_i} \neq 0$  and  $\forall \mu \in \{1, \dots, d_{s_i}\}$  with  $\mu \neq \xi_i$  and  $[C_{\eta_i s_i}]_\mu \neq 0$ ; it holds that  $[A_{s_i s_i}]_{\xi_i \mu} \neq 0$ .
2.  $C_{\eta_i s_i+1} = 0, \dots, C_{\eta_i \nu} = 0$ .
3. For every  $\mu \in \{1, \dots, d_{s_i}\}$  and every  $l \in \{1, \dots, s_i - 1\}$ , condition  $[C_{\eta_i l}]_\mu \neq 0$  implies  $[A_{s_i l}]_{\xi_i \mu} \neq 0$ .

*Proof.* Necessity: By Lemma 3.3, it follows that  $m \leq q$ . Moreover, since  $A$  is lower block triangular and  $A - KC$  is Hurwitz and Metzler, one has, as in the proof of Lemma 3.3,

$$(3.10) \quad A_{ij} - \sum_{\eta=1}^q K_{i\eta} C_{\eta j} \geq 0, \quad j < i,$$

$$(3.11) \quad A_{ii} - \sum_{\eta=1}^q K_{i\eta} C_{\eta i} : \text{Hurwitz and compartmental}, \quad 1 \leq i \leq \nu,$$

$$(3.12) \quad \sum_{\eta=1}^q K_{i\eta} C_{\eta j} = 0, \quad j > i.$$

Consider (3.11). Since  $A_{s_i s_i}$  is singular, the matrix  $\sum_{\eta=1}^q K_{s_i \eta} C_{\eta s_i}$  must be nonzero. This means that there exists  $\eta_i \in \{1, \dots, q\}$  such that  $K_{s_i \eta_i} C_{\eta_i s_i} \neq 0$  and  $A_{s_i s_i} - K_{s_i \eta_i} C_{\eta_i s_i}$  is compartmental. Moreover, since  $A_{s_i s_i}$  is an irreducible compartmental matrix and  $A_{s_i s_i} - K_{s_i \eta_i} C_{\eta_i s_i}$  is also a compartmental matrix, it follows from Proposition 3.12 in [32] that  $A_{s_i s_i} - K_{s_i \eta_i} C_{\eta_i s_i}$  is a Hurwitz and compartmental matrix. Since  $K_{s_i \eta_i}$  is a nonzero vector, there exist  $\xi_i \in \{1, \dots, d_{s_i}\}$  such that  $[K_{s_i \eta_i}]_{\xi_i} \neq 0$ , i.e.,  $[K_{s_i \eta_i}]_{\xi_i} > 0$ .

We show that the indices  $\eta_i$  and  $\xi_i$  satisfy Condition 1. Let  $\mu \in \{1, \dots, d_{s_i}\}$  and  $\mu \neq \xi_i$ . Suppose  $[C_{\eta_i s_i}]_\mu \neq 0$  (equivalently,  $[C_{\eta_i s_i}]_\mu > 0$ ). Recalling that  $A_{s_i s_i} - K_{s_i \eta_i} C_{\eta_i s_i}$  is compartmental,  $\mu \neq \xi_i$  implies that

$$0 \leq [A_{s_i s_i} - K_{s_i \eta_i} C_{\eta_i s_i}]_{\xi_i \mu} = [A_{s_i s_i}]_{\xi_i \mu} - [K_{s_i \eta_i}]_{\xi_i} [C_{\eta_i s_i}]_\mu.$$

From the fact that  $[C_{\eta_i s_i}]_\mu > 0$  and  $[K_{s_i \eta_i}]_{\xi_i} > 0$ , it follows that  $[A_{s_i s_i}]_{\xi_i \mu} > 0$ . Therefore, the existence of  $\eta_i$  and  $\xi_i$  satisfying Condition 1 has been proved. Note that the fact  $\eta_i \neq \eta_j$  if  $i \neq j$  has not been proved yet.

We will prove Conditions 2 and 3 for this choice of  $\eta_i$ 's and of  $\xi_i$ 's.

By (3.12) and  $K_{s_i \eta_i} \neq 0$ , we obtain  $C_{\eta_i s_i+1} = 0, \dots, C_{\eta_i \nu} = 0$ , which proves that Condition 2 holds true.

Now, we prove the fact  $\eta_i \neq \eta_j$  if  $i \neq j$ . Note that when  $m = 1$ , there is only one index and hence there is nothing to prove. For the case  $m \geq 2$ , suppose there exist  $i, j \in \{1, \dots, m\}$  such that  $i < j$  and  $\eta_i = \eta_j$ . Then, by Condition 2, we have  $C_{\eta_i s_i+1} = 0, \dots, C_{\eta_i s_j} = 0, \dots, C_{\eta_i \nu} = 0$ . Moreover, the fact  $\eta_i = \eta_j$  results in  $C_{\eta_j s_j} = 0$ . However, this is a contradiction to Condition 1 ( $C_{\eta_i s_i} \neq 0 \forall i$ ).

Moving to Condition 3, let  $\mu \in \{1, \dots, d_{s_i}\}$  and  $l \in \{1, \dots, s_i - 1\}$  and suppose  $[C_{\eta_i l}]_\mu \neq 0$ . Since  $l < s_i$ , relation (3.10) holds, and therefore  $0 \leq A_{s_i l} - \sum_{\eta=1}^q K_{s_i \eta} C_{\eta l}$ .

Hence,  $K_{s_i\eta_i}C_{\eta_i l} \leq \sum_{\eta=1}^q K_{s_i\eta}C_{\eta l} \leq A_{s_i l}$ . Recalling that  $[K_{s_i\eta_i}]_{\xi_i} > 0$ , which has been shown in the first part of this proof, one has  $0 < [K_{s_i\eta_i}]_{\xi_i}[C_{\eta_i l}]_{\mu} = [K_{s_i\eta_i}C_{\eta_i l}]_{\xi_i\mu} \leq [A_{s_i l}]_{\xi_i\mu}$ , which proves that Condition 3 holds true. Therefore the proof of necessity is complete.

Sufficiency: We take  $K_{vw} = 0$  if there is no  $i \in \{1, \dots, m\}$  such that  $v = s_i$  and  $w = \eta_i$ . That is to say, only the vectors  $K_{s_i\eta_i}$ ,  $i = 1, \dots, m$ , will be chosen to be nonzero. Indeed, let  $K_{s_i\eta_i}$ ,  $i = 1, \dots, m$ , be defined as follows.

1. In case  $C_{\eta_i 1} = 0, \dots, C_{\eta_i s_i-1} = 0$ , and  $[C_{\eta_i s_i}]_{\xi} = 0, \xi \neq \xi_i$ ,

$$[K_{s_i\eta_i}]_{\xi} = 0, \quad \xi \neq \xi_i,$$

$$[K_{s_i\eta_i}]_{\xi_i} = \alpha_i, \quad \alpha_i : \text{arbitrary positive number.}$$

2. Otherwise

$$[K_{s_i\eta_i}]_{\xi} = 0, \quad \xi \neq \xi_i,$$

$$[K_{s_i\eta_i}]_{\xi_i} = \min_{\substack{1 \leq \mu \leq n, \mu \neq \mu_i^* \\ [C]_{\eta_i\mu} \neq 0}} \frac{[A_{s_i}]_{\xi_i\mu}}{[C]_{\eta_i\mu}}, \quad \mu_i^* := \xi_i + \sum_{j=1}^{s_i-1} d_j,$$

where  $A_{[s_i]} := [A_{s_i 1} \ \dots \ A_{s_i s_i} \ 0 \ \dots \ 0] \in \mathbb{R}_+^{d_{s_i} \times n}$ .

It is easy to see that the matrix  $K$  constructed above is nonzero and renders  $A - KC$  a Hurwitz and Metzler matrix. Thus the assertion follows.  $\square$

Since our class of systems includes compartmental systems, it is expected that Theorem 3.4 can be refined further. To do this, we assume that (3.1) is a compartmental system ( $A$  is a compartmental matrix and  $C \in \mathbb{R}_+^{q \times n}$ ) with  $0 \in \sigma(A)$  and the algebraic multiplicity of the zero eigenvalue is  $m$ . Without loss of generality it can be assumed that  $A$  has the form (Theorem 3.6 in [32])

$$(3.13) \quad A = \begin{bmatrix} A_{11} & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & 0 & & & \\ A_{\nu-m,1} & & A_{\nu-m,\nu-m} & 0 & & \\ A_{\nu-m+1,1} & & A_{\nu-m+1,\nu-m} & A_{\nu-m+1,\nu-m+1} & & \\ \vdots & & \vdots & 0 & \ddots & 0 \\ A_{\nu 1} & \dots & A_{\nu,\nu-m} & 0 & 0 & A_{\nu\nu} \end{bmatrix},$$

where  $A_{ii} \in \mathbb{R}^{d_i \times d_i}$  is compartmental and irreducible  $\forall i = 1, \dots, \nu$ , and  $0 \in \sigma(A_{ii})$  with multiplicity 1  $\forall i = \nu - m + 1, \dots, \nu$ . Note that  $A_{11}, \dots, A_{\nu-m,\nu-m}$  are Hurwitz and  $A_{ij} \geq 0, i \neq j$ . We also decompose  $C$  as in (3.5). Thus,  $C_{ij} \in \mathbb{R}_+^{1 \times d_j}$ .

The special structure of  $A$  allows us to refine Theorem 3.4 as follows. Note that this result is a correct version of Theorem 3.13 in [32].<sup>1</sup>

<sup>1</sup>Theorem 3.13 in [32]: Consider a linear compartmental system (3.1) with  $A, C$  decomposed as in (3.13) and (3.5). Let  $C_i = [C_{1i}^t \ \dots \ C_{qi}^t]^t$ . Suppose the zero eigenvalue of  $A$  is of multiplicity  $m$ .

1. If  $m = 0$ , then  $(A, C)$  is positively detectable if and only if  $(A, C)$  is positively modifiable.
2. For  $m \geq 1$ ,  $(A, C)$  is positively detectable if and only if  $(A_{ii}, C_i)$  is positively modifiable  $\forall i = \nu - m + 1, \dots, \nu$ .

$(A, C)$  is said to be positively detectable if  $\exists K \geq 0$  such that  $A - KC$  is Hurwitz and compartmental, and  $(A, C)$  is said to be positively modifiable if  $\exists K \geq 0$  such that  $KC \neq 0$  and  $A - KC$  is compartmental.

**COROLLARY 3.5.** *Consider the positive system (3.1) with  $A$  being a compartmental matrix decomposed as in (3.13) and suppose the multiplicity of the zero eigenvalue of  $A$  is  $m \geq 1$ . This system admits a positive linear observer of the form (3.2) with  $K \in \mathbb{R}_+^{n \times q}$  if and only if there exist  $m$  numbers  $\eta_{\nu-m+1}, \dots, \eta_\nu$  with  $\eta_i \leq q$ ,  $\nu - m + 1 \leq i \leq \nu$ , and  $\eta_i \neq \eta_j$  if  $i \neq j$ , and  $m$  numbers  $\xi_{\nu-m+1}, \dots, \xi_\nu$  with  $\xi_i \in \{1, \dots, d_i\}$ ,  $\nu - m + 1 \leq i \leq \nu$ , such that for each  $i \in \{\nu - m + 1, \dots, \nu\}$  the following conditions hold:*

1.  $C_{\eta_i i} \neq 0$ , and  $\forall \mu \in \{1, \dots, d_i\}$  with  $\mu \neq \xi_i$  and  $[C_{\eta_i i}]_\mu \neq 0$ , it holds that  $[A_{ii}]_{\xi_i \mu} \neq 0$ .
2.  $C_{\eta_i k} = 0 \ \forall k \in \{\nu - m + 1, \dots, \nu\} - \{i\}$ .
3. For every  $\mu \in \{1, \dots, d_i\}$  and every  $l \in \{1, \dots, \nu - m\}$ , condition  $[C_{\eta_i l}]_\mu \neq 0$  implies  $[A_{il}]_{\xi_i \mu} \neq 0$ .

Condition 1 of Corollary 3.5 is the detectability condition with a positive gain matrix (in [32], it is called positive detectability). The second condition says that each output associated to the singular block should be decoupled in some sense. The last condition guarantees that  $A - KC$  is a Metzler matrix.

*Example 3.6.* Consider a positive linear system with  $(A, C)$  defined by

$$A = \begin{bmatrix} M & 0 \\ 0 & M \end{bmatrix},$$

$$M := \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix},$$

$$C = [1 \quad 1 \quad 1 \quad 1].$$

Since  $0 \in \sigma(A)$ ,  $m = 2$ , and the pair  $(M, [1 \quad 1])$  is positively modifiable in the sense of [32], Theorem 3.13 of [32] asserts that there exists  $K \in \mathbb{R}_+^{4 \times 1}$  such that  $A - KC$  is a Hurwitz and compartmental matrix. However, any nonzero value of  $k_i$  ( $i$ th component of  $K$ ) yields a negative off-diagonal entry in  $A - KC$ . Thus, this system does not admit a positive linear observer of the form (3.2). (Furthermore, it is not possible to find a  $K \in \mathbb{R}^{4 \times 1}$  rendering  $A - KC$  a Hurwitz matrix, because  $(A, C)$  is not a detectable pair.) Note that this can be verified easily by Corollary 3.5 or Lemma 3.3 since  $m > q = 1$ .

Theorem 3.13 of [32] is not complete even if a detectability condition is added to the conditions. Consider the pair  $(A, C)$  defined by

$$A = \begin{bmatrix} -1 & 0 \\ 0 & M \end{bmatrix}, \quad C = [1 \quad 1 \quad 1].$$

It is easy to see that  $(A, C)$  is a detectable pair, but there is no  $K \in \mathbb{R}_+^3$  that renders  $A - KC$  Hurwitz and compartmental, although this pair satisfies the condition of [32]. According to Corollary 3.5, the problem can be solved if  $(A, C)$  satisfies additional conditions, for example,  $[A]_{2,1} > 0$  or  $[A]_{3,1} > 0$ , or  $C = [0 \quad 1 \quad 1]$ .

*Remark 3.7.* 1. The condition  $m \leq q$  is inspired by the result [19, Proposition 10], which deals with the stabilization problem for single input positive systems (i.e., find a gain vector  $k$  such that  $-k \in R_+^n$  and  $A + gk$  is Hurwitz and Metzler where  $A$  satisfies the condition (3.3)). One can modify Lemma 3.3 and Theorem 3.4 to obtain multi-input extensions of the results in [19].

2. When system (3.1) has an input vector field (i.e.,  $\dot{x} = Ax + Bu$ ,  $y = Cx$ ), the observer becomes  $\dot{\hat{x}} = A\hat{x} + K(y - C\hat{x}) + Bu$ .

#### 4. Positive linear observers using coordinate transformations.

**4.1. Problem formulation and basic results.** In this section, we propose a new design method for positive linear observers. It is based on the fact that even though a linear system is not positive with respect to (w.r.t.) a coordinate system, it can be positive w.r.t. a different coordinate system. With this fact in mind, we propose a new positive observer which employs a coordinate transformation.

**DEFINITION 4.1.** *Consider system (3.1). If there exist  $K \in \mathbb{R}^{n \times q}$  such that  $A - KC$  is Hurwitz and a nonsingular matrix  $T \in \mathbb{R}^{n \times n}$  such that the dynamic system described by*

$$(4.1) \quad \begin{aligned} \dot{\hat{z}} &= T(A - KC)T^{-1}\hat{z} + TKy, \\ \hat{x} &= T^{-1}\hat{z} \end{aligned}$$

*is a positive system w.r.t. input  $y$  and output  $\hat{x}$ , we call this a positive observer for system (3.1).*

Clearly, the system defined above is a positive observer since it guarantees that  $\lim_{t \rightarrow \infty} (\hat{x}(t) - x(t)) = 0$  and  $\hat{x}(t) \in \mathbb{R}_+^n \forall t \geq 0$ . Note that the observers developed in the previous section are special cases of the observer in (4.1) ( $T = I$ ). Before proceeding, we present an example to show the effectiveness of our approach.

*Example 4.2.* Consider the positive system

$$(4.2) \quad \dot{x} = \begin{bmatrix} -1 & 1 \\ 2 & 0 \end{bmatrix} x =: Ax, \quad y = [1 \quad 0] x =: Cx.$$

Theorem 3.2 tells us that it is not possible to design a positive observer with the structure (3.2). (It results in  $k_1^* > -1$  and  $k_2^* = 2$ , which make  $\lambda_{\max}(A - K^*C) \geq 0$ .)

Now, take  $K = [2 \quad 4]^t$ . Then the Luenberger observer becomes

$$\dot{\hat{x}} = \begin{bmatrix} -3 & 1 \\ -2 & 0 \end{bmatrix} \hat{x} + \begin{bmatrix} 2 \\ 4 \end{bmatrix} y,$$

which is not a positive system. However, if we transform this system using  $\hat{z} = T\hat{x}$ , where  $T := \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{2}} & \frac{2}{\sqrt{5}} \end{bmatrix}^{-1}$  with  $T^{-1} \in \mathbb{R}_+^{2 \times 2}$ , one has

$$\begin{aligned} \dot{\hat{z}} &= T(A - KC)T^{-1}\hat{z} + TKy \\ &= \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix} \hat{z} + \begin{bmatrix} 0 \\ 2\sqrt{5} \end{bmatrix} y, \\ \hat{x} &= T^{-1}\hat{z}, \end{aligned}$$

which is a positive system.

Now, we present a preliminary result whose proof is standard.

**THEOREM 4.3.** *Consider a positive linear system (3.1). Suppose  $(A, C)$  is detectable. There exists a positive linear observer (4.1) if and only if there exist  $F, G$ , and  $T$  such that*

1.  $F \in \mathbb{R}^{n \times n}$  is Hurwitz and Metzler, and  $G \in \mathbb{R}_+^{n \times q}$ ;
2.  $T$  is a solution of the Sylvester equation

$$(4.3) \quad TA - FT = GC.$$

3.  $T$  is invertible and  $T^{-1} \in \mathbb{R}_+^{n \times n}$ .

It is well known from linear systems theory that if  $A$  and  $F$  have no common eigenvalues, then a solution  $T$  to (4.3) always exists and is unique. Moreover, if  $A$  and  $F$  have no common eigenvalues, observability of  $(A, C)$  and controllability of  $(F, G)$  are necessary conditions for the existence of a nonsingular solution  $T$  of (4.3) [8]. (It is also sufficient in the single output case.) If  $(A, C)$  is detectable, but not observable, then  $\sigma(A) \cap \sigma(F) \neq \emptyset$ , and hence we will not consider this case, but rather the restrictive one when  $\sigma(A) \cap \sigma(F) = \emptyset$ .

*Remark 4.4.* 1. An inverse-positive matrix  $T$  solves (4.3) if and only if an invertible nonnegative matrix  $S = T^{-1}$  solves the equation

$$(4.4) \quad AS - SF = SGCS.$$

Note that (4.3) is a linear equation w.r.t.  $T$  while (4.4) is a quadratic one w.r.t.  $S$ . Equation (4.4) is called a nonsymmetric algebraic Riccati equation in the literature.

2. To check the existence of positive linear observers, that is to say, to check whether the equation  $TA - FT = GC$  admits a solution such that  $T^{-1} \in \mathbb{R}_+^{n \times n}$ , one should use information on the structure of  $F$  and  $G$ , and in particular on the eigenvalues of  $F$ . However, to the best of the authors' knowledge, the problem of the realization (or of the existence) of a Metzler matrix which has prescribed eigenvalues is not solved. This problem, the so-called inverse eigenvalue problem, has been recognized as a very hard problem (note that a Metzler matrix  $M$  is a nonnegative matrix up to a shift process, i.e.,  $M + \alpha I \in \mathbb{R}_+^{n \times n}$  for some  $\alpha \in \mathbb{R}$ ), and there are very few results on general cases (see [6, 10, 21, 30] and the references therein).

3. For controlled systems ( $\dot{x} = Ax + Bu$ ,  $y = Cx$ ), the matrix  $T$  should satisfy  $TB \geq 0$  in addition to the conditions of Theorem 4.3.

**4.2. Necessary conditions: Single output case.** Two necessary conditions on the system matrix, in particular on the number of real eigenvalues, are provided.

The first necessary condition utilizes the explicit solution  $T$  of (4.3). It also exploits the structure of the matrix  $F$ . As mentioned in Remark 4.4, we do not have sufficient information on the structure the matrix  $F$  may be endowed with in the general case. To make the problem tractable, additional structural conditions on the observers are imposed. We also assume that  $F$  is chosen so that  $\sigma(F) \cap \sigma(A) = \emptyset$  to ensure that the Sylvester equation  $TA - FT = GC$  has a unique solution  $T$  [8].

**THEOREM 4.5.** *Suppose the positive system (3.1) admits a positive linear observer of the form (4.1). Let  $F = T(A - KC)T^{-1}$  and  $G = TK$ . If  $F$  is chosen to satisfy (1)  $\sigma(F) \cap \sigma(A) = \emptyset$ , and (2)  $F = \text{diag}\{\lambda_1, \dots, \lambda_n\}$ ,  $\lambda_i \in \mathbb{R}_-$ , then  $A$  has at least  $n - 1$  distinct negative real eigenvalues.*

*Proof.* Let  $\Delta_A(s)$  be the characteristic polynomial of  $A$ , i.e.,

$$(4.5) \quad \Delta_A(s) = \det(sI - A) = s^n + \alpha_1 s^{n-1} + \alpha_2 s^{n-2} + \dots + \alpha_n.$$

Without loss of generality we assume that  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ . In fact, if this is not the case, we transform  $F$  with a suitable permutation matrix  $P$  such that the diagonal elements of  $\bar{F} = PFP^t$  are ordered as required, since an inverse-positive matrix  $T$  is a solution to  $TA - FT = GC$  if and only if  $\bar{T} = PT$  is an inverse-positive matrix satisfying  $\bar{T}A - \bar{F}\bar{T} = \bar{G}C$ ,  $\bar{G} = PG$ .



Now, following the same derivations as in [8], we have

$$\begin{aligned}
 TI - IT &= 0, \\
 TA - FT &= GC, \\
 TA^2 - F^2T &= GCA + FGC, \\
 &\vdots \\
 TA^n - F^nT &= GCA^{n-1} + FGCA^{n-2} + \cdots + F^{n-2}GCA + F^{n-1}GC,
 \end{aligned}$$

and

$$\begin{aligned}
 T\Delta_A(A) - \Delta_A(F)T &= \begin{bmatrix} G & FG & \cdots & F^{n-1}G \end{bmatrix} \begin{bmatrix} \alpha_{n-1} & \alpha_{n-2} & \cdots & \alpha_1 & 1 \\ \alpha_{n-2} & \alpha_{n-3} & \cdots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \alpha_1 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-2} \\ CA^{n-1} \end{bmatrix} \\
 &=: U_F \Lambda_\alpha V_A.
 \end{aligned}$$

Note that the matrices  $U_F$  and  $V_A$  are the controllability matrix of the pair  $(F, G)$  and the observability matrix of the pair  $(C, A)$ , respectively.

Since  $\Delta_A(A) = 0$ , and the condition  $\sigma(F) \cap \sigma(A) = \emptyset$  ensures that  $\Delta_A(F) = \text{diag}\{\Delta_A(\lambda_1), \dots, \Delta_A(\lambda_n)\}$  is nonsingular, we have

$$(4.6) \quad T = -(\Delta_A(F))^{-1} U_F \Lambda_\alpha V_A.$$

In addition, the existence of a positive linear observer (4.1) implies that there exists  $S \in \mathbb{R}_+^{n \times n}$  such that

$$TS = I.$$

In other words,

$$(4.7) \quad U_F \Lambda_\alpha V_A S = -\Delta_A(F).$$

Now, expand  $U_F$  as follows.

$$U_F = \begin{bmatrix} g_1 & \lambda_1 g_1 & \cdots & \lambda_1^{n-1} g_1 \\ \vdots & \vdots & & \vdots \\ g_n & \lambda_n g_n & \cdots & \lambda_n^{n-1} g_n \end{bmatrix} = \begin{bmatrix} g_1 & & & \\ & \ddots & & \\ & & g_n & \end{bmatrix} \begin{bmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^{n-1} \end{bmatrix}.$$

Notice that since  $\Delta_A(F)$  is nonsingular,  $U_F$  is nonsingular as well, and hence

- $g_i \neq 0$  for every  $i$  (equivalently,  $g_i > 0$  since  $G \geq 0$ );
- the Vandermonde matrix [8] on the right-hand side is nonsingular (equivalently, all  $\lambda_i$ 's are distinct).

Set  $\Theta := U_F \Lambda_\alpha$ , so that (4.7) becomes

$$\Theta V_A S = - \begin{bmatrix} \Delta_A(\lambda_1) & & \\ & \ddots & \\ & & \Delta_A(\lambda_n) \end{bmatrix},$$

and notice that

$$(4.8) \quad \det \Theta = \left( \prod_{i=1}^n g_i \right) \left( \prod_{1 \leq i < j \leq n} (\lambda_i - \lambda_j) \right).$$

If we let  $S_i$  denote the  $i$ th column of  $S$ , we get

$$\Theta V_A S_i = \begin{bmatrix} 0_{(i-1) \times 1} \\ -\Delta_A(\lambda_i) \\ 0_{(n-i) \times 1} \end{bmatrix}.$$

Consequently, by recalling the definition of  $V_A$ , we get for every  $i = 1, \dots, n$

$$\begin{aligned} CS_i &= [1 \quad 0 \quad \cdots \quad 0] V_A S_i = [1 \quad 0 \quad \cdots \quad 0] \Theta^{-1} \begin{bmatrix} 0_{(i-1) \times 1} \\ -\Delta_A(\lambda_i) \\ 0_{(n-i) \times 1} \end{bmatrix} \\ &= (-1)^i \frac{\Delta_A(\lambda_i)}{\det \Theta} \cdot \det \Theta_{[i]}, \end{aligned}$$

where  $\Theta_{[i]}$  is the  $(n-1) \times (n-1)$  submatrix of  $\Theta$  obtained by deleting in  $\Theta$  the first column and the  $i$ th row.

Using (4.8), we compute

$$\begin{aligned} \frac{\det \Theta_{[i]}}{\det \Theta} &= \frac{\prod_{1 \leq u < v \leq n, u \neq i, v \neq i} (\lambda_u - \lambda_v)}{g_i \prod_{1 \leq u < v \leq n} (\lambda_u - \lambda_v)} \\ &= \begin{cases} \frac{1}{g_i \left[ \prod_{1 \leq u < i} (\lambda_u - \lambda_i) \right] \left[ \prod_{i < v \leq n} (\lambda_i - \lambda_v) \right]}, & 2 \leq i \leq n-1, \\ \frac{1}{g_1 (\lambda_1 - \lambda_2) \cdots (\lambda_1 - \lambda_n)}, & i = 1, \\ \frac{1}{g_n (\lambda_1 - \lambda_n) \cdots (\lambda_{n-1} - \lambda_n)}, & i = n. \end{cases} \end{aligned}$$

Since  $\lambda_i < \lambda_{i+1}$ ,  $i = 1, \dots, n-1$ , and  $g_i > 0$ , it follows that

$$CS_i = (-1)^{i+n-1} \Delta_A(\lambda_i) \left| \frac{\det \Theta_{[i]}}{\det \Theta} \right|,$$

which in turn implies that

$$(-1)^{i+n-1} \Delta_A(\lambda_i) > 0, \quad i = 1, \dots, n.$$

Moreover, by the relation

$$(-1)^{i+n-1} \Delta_A(\lambda_i) \cdot (-1)^{i+1+n-1} \Delta_A(\lambda_{i+1}) > 0, \quad i = 1, \dots, n-1,$$

we have

$$\Delta_A(\lambda_i) \Delta_A(\lambda_{i+1}) < 0, \quad i = 1, \dots, n-1.$$

Since  $\Delta_A(\cdot)$  is a continuous function, there exists  $\lambda_{*i}$  such that

$$\Delta_A(\lambda_{*i}) = 0, \quad \lambda_i < \lambda_{*i} < \lambda_{i+1}, \quad i = 1, \dots, n-1.$$

That is to say, there are at least  $n-1$  negative real eigenvalues of  $A$ , which completes the proof.  $\square$

The second result of this subsection is on the number of positive real eigenvalues of the system matrix in the general case, i.e., in the case that no assumption, apart from stability, on the eigenvalues of  $F$  is imposed. To prove the result, we need the following fact.

**LEMMA 4.6.** *Let  $A \in \mathbb{R}^{n \times n}$  be Metzler and  $C \in \mathbb{R}_+^{q \times n}$ . If there exists  $K \in \mathbb{R}_+^{n \times q}$  such that  $A - KC$  is Metzler and  $\lambda_{\max}(A - KC) < \lambda_{\max}(A)$ , then the algebraic multiplicity of  $\lambda_{\max}(A)$  is at most  $q$ .*

*Proof.* Let  $m$  be the multiplicity of  $\lambda_{\max}(A)$  as an eigenvalue of  $A$ . The Metzler matrix  $\bar{A} := A - \lambda_{\max}(A)I_n$  satisfies (3.3) and has 0 as an eigenvalue of multiplicity  $m$ . Also, since  $\lambda_{\max}(\bar{A} - KC) < 0$ , the matrix  $K$  corresponds to the gain matrix of a positive linear observer of the form (3.2) for the system. So, by Lemma 3.3,  $m \leq q$ .  $\square$

This lemma says that for the single output case, we cannot make  $\lambda_{\max}(A - KC)$  less than  $\lambda_{\max}(A)$  if the multiplicity of  $\lambda_{\max}(A)$  is greater than 1.

**THEOREM 4.7.** *Let  $A \in \mathbb{R}^{n \times n}$  be a Metzler matrix and  $C := [c_1, \dots, c_n] \in \mathbb{R}_+^{1 \times n}$ . If there exists a positive linear observer of the form (4.1), then the number of nonnegative real eigenvalues of  $A$  counting the multiplicity is at most 1.*

*Proof.* First recall that the existence of a positive linear observer of the form (4.1) implies that there exists  $G \in \mathbb{R}_+^n$ ,  $S \in \mathbb{R}_+^{n \times n}$  with  $\det S \neq 0$  such that  $F = S^{-1}AS - GCS$  is Metzler and Hurwitz.

Let  $\bar{A} = S^{-1}AS$ ,  $\bar{C} = CS$ . If  $\bar{C}$  is a zero vector, the proof is trivial. Thus,  $\bar{C}$  is assumed to be a nonzero vector.

Let  $G \in \mathbb{R}_+^n$  be such that  $\bar{A} - G\bar{C}$  is Metzler and Hurwitz. Then it is evident that  $\bar{A} - \delta G\bar{C}$ ,  $\delta \in [0, 1]$ , is Metzler. Moreover, since the zeros of a polynomial depend continuously on its coefficients, there exists  $\delta^* \in [0, 1)$  such that  $\bar{A} - \delta G\bar{C}$  is Metzler and Hurwitz  $\forall \delta \in [\delta^*, 1]$ . Let  $\Delta(s, \delta) = \det(sI - \bar{A} + \delta G\bar{C})$ . We will investigate the loci of the roots of  $\Delta(s, \delta)$  changing  $\delta$  from 0 to 1. To do this, the structure of  $\Delta(s, \delta)$  is considered. Note that

$$\begin{aligned} (4.9) \quad \det(sI - \bar{A} + \delta G\bar{C}) &= \det((sI - \bar{A})(I + \delta(sI - \bar{A})^{-1}G\bar{C})) \\ &= \det(sI - \bar{A})(1 + \delta\bar{C}(sI - \bar{A})^{-1}G) \\ &= \det(sI - \bar{A}) + \delta\bar{C}\text{adj}(sI - \bar{A})G \\ &=: D(s) + \delta N(s), \end{aligned}$$

where we used the relation

$$\det(I + xy^t) = 1 + x^t y = 1 + y^t x, \quad x, y \in \mathbb{R}^n.$$

From the structure of (4.9), we can use the classical root locus method to investigate the location of the zeros of  $\Delta(s, \delta)$  parameterized by  $\delta$ . Letting  $G_o(s) := N(s)/D(s)$  ( $n$ : degree of  $D(s)$ ;  $m$ : degree of  $N(s)$ ), we recall basic rules of root locus as follows. For details on the root locus, see, for example, [11, 15, 31].

- RL1. The branches of the root locus are continuous curves that start at each of the  $n$  poles of  $G_o(s)$  for  $\delta = 0$ . As  $\delta \rightarrow \infty$ , the locus branches approach the  $m$  zeros of  $G_o(s)$ . Locus branches for excess poles extend infinitely far from the origin.
- RL2. The root locus includes all points along the real axis to the left of an odd number of poles and zeros of  $G_o(s)$ .
- RL3. If there is a breakaway point (multiple root of  $G_o(s)$ ) on the real axis, the root loci leave the real axis at a gain  $\delta_*$  that is the maximum  $\delta$  in that region of the real axis.

Suppose  $A$  has more than one nonnegative real eigenvalue. Let  $\lambda_M$  ( $\lambda_m$ , resp.) be the largest (the second largest, resp.) nonnegative real eigenvalue of  $A$ . If  $N(\lambda_M) = 0$  or  $N(\lambda_m) = 0$ , then  $\Delta(\lambda_M, \delta) = 0$  or  $\Delta(\lambda_m, \delta) = 0 \forall \delta \in [0, \infty)$ , which means that  $\bar{A} - G\bar{C}$  is not Hurwitz and contradicts the assumption. Thus, we assume  $N(\lambda_M) \neq 0$  and  $N(\lambda_m) \neq 0$ .

We complete the proof by considering three cases.

*Case 1.* The multiplicity of  $\lambda_M$  is greater than 1.

*Case 2.*  $\lambda_M$  is simple and  $(\lambda_M, \infty)$  contains an odd number of zeros of  $N(s)$ .

*Case 3.*  $\lambda_M$  is simple and  $(\lambda_M, \infty)$  contains an even number of zeros of  $N(s)$ .

Case 1 contradicts Lemma 4.6. In Case 2, by RL2,  $\forall \delta \in [0, \infty)$ , there exists  $s^* \in [\lambda_m, \infty)$  (which may depend on  $\delta$ ) such that  $\Delta(s^*, \delta) = 0$ , which means that  $\bar{A} - \delta G\bar{C}$  is not Hurwitz  $\forall \delta \in [0, \infty)$  and contradicts the assumption on the existence of  $G$ .

Consider Case 3. Since  $\lambda_M$  is simple, one has  $\lambda_m < \lambda_M$ . Suppose there exists  $\mu^* \in (\lambda_m, \lambda_M)$  such that  $N(\mu^*) = 0$ . Let  $\mu_M^*$  be the largest one among the  $\mu^*$ 's. Then, again by RL2, there exists  $s^* \in [\mu_M^*, \lambda_M]$  ( $s^*$  may depend on  $\delta$ ) such that  $\Delta(s^*, \delta) = 0 \forall \delta \in [0, \infty)$ , which is, similar to Case 2, a contradiction. Thus, there should be no  $s^* \in (\lambda_m, \lambda_M)$  such that  $N(s^*) = 0$ . This fact implies that the line segment  $(\lambda_m, \lambda_M)$  is a part of the root locus (by RL2) and that there exists  $\delta_* \in (0, \infty)$  such that the equation  $\Delta(s, \delta_*) = 0$  w.r.t.  $s$  has a solution  $s^*$  with multiplicity greater than 1 in the region  $(\lambda_m, \lambda_M)$  where the root locus leaves the real axis (by RL3). If  $\delta_* \geq 1$ , then  $\bar{A} - G\bar{C} = (\bar{A} - \delta_* G\bar{C})|_{\delta=1}$  is not Hurwitz, which contradicts the assumption that  $\bar{A} - G\bar{C}$  is Hurwitz. When  $0 < \delta_* < 1$ ,  $\bar{A} - \delta_* G\bar{C}$  has a nonnegative real eigenvalue  $s^*$  with multiplicity greater than 1. However, since  $\bar{A} - G\bar{C} = \bar{A} - \delta_* G\bar{C} - (1 - \delta_*)G\bar{C}$  ( $\bar{A} - \delta_* G\bar{C}$  is Metzler but not Hurwitz) and  $\bar{A} - G\bar{C}$  is Hurwitz and Metzler, this contradicts Lemma 4.6. Thus, the proof is complete.  $\square$

This result imposes a rather strong condition for the positive linear observer problem solution. According to this theorem, it is not possible to design positive linear observers for single output linear positive systems with more than one nonnegative real eigenvalue.

*Remark 4.8.* The necessary condition given in Theorem 4.5 provides more information on the spectrum of  $A$  (stable eigenvalues as well as unstable eigenvalues) than that of Theorem 4.7, since we fix the structure and the spectrum of the matrix  $F$ . On the contrary, Theorem 4.7 does not rely on a specific structure for  $F$ ; hence it does not allow us to conclude anything on the stable eigenvalues of  $A$ .

**4.3. Sufficient condition: Single output case.** With the necessary conditions developed in mind, we provide a sufficient condition for the existence of positive linear observers. The irreducibility of a matrix plays an important role in this section.

We assume that  $A$  has  $n$  real distinct eigenvalues  $\lambda_1, \dots, \lambda_n$  satisfying  $\lambda_1 > \dots > \lambda_n$ . This spectral assumption enables the transformation of  $A$  into the Jordan form

given by

$$(4.10) \quad A = VJV^{-1},$$

$$(4.11) \quad J := \text{diag}\{\lambda_1, \dots, \lambda_n\},$$

$$(4.12) \quad V := [v_1 \quad \dots \quad v_n],$$

where  $v_i$  satisfies  $Av_i = \lambda_i v_i$ .

**THEOREM 4.9.** *For a single output positive system (3.1), suppose  $A$  is irreducible and  $(A, C)$  is observable. There exists a positive linear observer using coordinates transformation (4.1) if  $A$  has  $n$  distinct real eigenvalues,  $n-1$  of which are negative.*

*Proof.* If  $A$  is Hurwitz, the problem is trivial. Thus, it is assumed that one eigenvalue is nonnegative. Let  $\lambda_i$  (eigenvalues of  $A$ ) and  $v_i$  (eigenvectors of  $A$ ) be such that  $\lambda_1 \geq 0 > \lambda_2 > \dots > \lambda_n$ . Since  $A$  is irreducible, by Theorem 2.1, the eigenvector  $v_1$  corresponding to  $\lambda_1$  satisfies  $v_1 \in \text{int } \mathbb{R}_+^n$ . By the observability assumption, we have  $\forall i = 1, \dots, n$ ,  $Cv_i \neq 0$ . After a suitable scaling, it is assumed that

$$(4.13) \quad Cv_i = 1, \quad i = 1, \dots, n.$$

Let  $\beta, \gamma$  be real numbers in the open interval  $(0, 1)$ . We define

$$(4.14) \quad \bar{S}(\beta, \gamma) = \begin{bmatrix} 1 - \beta & (1 - \gamma)1_{1 \times (n-1)} \\ -\beta 1_{(n-1) \times 1} & \gamma I_{n-1} \end{bmatrix}.$$

Using the properties of determinant, it is easy to see that

$$\det \bar{S} = \det \begin{bmatrix} 1 - n\beta & 1_{1 \times (n-1)} \\ -\beta 1_{(n-1) \times 1} & \gamma I_{n-1} \end{bmatrix} = \det \begin{bmatrix} 1 - n\beta + (n-1)\beta/\gamma & 1_{1 \times (n-1)} \\ 0_{(n-1) \times 1} & \gamma I_{n-1} \end{bmatrix}.$$

Hence,

$$\det \bar{S} = \gamma^{n-2}[\gamma - n\beta\gamma + (n-1)\beta] = \gamma^{n-2}[(1-\beta)\gamma + (n-1)\beta(1-\gamma)].$$

Note that  $\det \bar{S} > 0 \quad \forall \beta, \gamma \in (0, 1)$ , which is obvious from the last relation. During this proof we will use the inverse of  $\bar{S}$ , which can be computed as follows:

$$(4.15) \quad \bar{S}^{-1} = \frac{\gamma^{n-3}}{\det \bar{S}} \begin{bmatrix} \gamma^2 & -\gamma(1-\gamma)1_{1 \times (n-1)} \\ \beta\gamma 1_{(n-1) \times 1} & -\beta(1-\gamma)1_{(n-1) \times (n-1)} + \pi_{\beta, \gamma} I_{n-1} \end{bmatrix},$$

where  $\pi_{\beta, \gamma} := (1-\beta)\gamma + (n-1)\beta(1-\gamma)$ . Notice that  $\pi_{\beta, \gamma} > 0 \quad \forall \beta, \gamma \in (0, 1)$ .

Now, we consider a matrix  $S$  defined by

$$(4.16) \quad S(\beta, \gamma) = V\bar{S}(\beta, \gamma).$$

Let  $\bar{\beta}^* \in (0, 1)$ ,  $\bar{\gamma}^* \in (0, 1)$  such that

$$(4.17) \quad S(\beta, \gamma) \geq 0 \quad \forall \beta \in (0, \bar{\beta}^*), \gamma \in (0, \bar{\gamma}^*).$$

It can be shown that such  $\bar{\beta}^*$  and  $\bar{\gamma}^*$  always exist. In fact, let  $S_i$  be the  $i$ th column of  $S$ . Then

$$(4.18) \quad \begin{aligned} S_1 &= (1-\beta)v_1 + \beta(-v_2 - \dots - v_n), \\ S_i &= (1-\gamma)v_1 + \gamma v_i, \quad i = 2, \dots, n. \end{aligned}$$

Recall that each element of  $v_1$  is positive and note that  $S_i$ 's move continuously from  $v_1$  to  $-(v_2 + \dots + v_n)$  or  $v_i$  as  $\beta$  and  $\gamma$  increase. Thus, by continuity, the existence of  $\bar{\beta}^*$  and  $\bar{\gamma}^*$  is proved.

Using  $S(\beta, \gamma)$ , define a state transformation  $T(\beta, \gamma) = S^{-1}(\beta, \gamma)$  parameterized by  $\beta$  and  $\gamma$ . By construction,  $T^{-1}$  exists and is nonnegative  $\forall \beta \in (0, \bar{\beta}^*), \gamma \in (0, \bar{\gamma}^*)$ . With  $T$ , consider  $F$  given by

$$(4.19) \quad F = TAT^{-1} - GCT^{-1}.$$

Noting that the first column of  $\bar{S}^{-1}$  is elementwise positive, we choose a positive gain matrix  $G$  as

$$(4.20) \quad G = \mu \bar{S}^{-1} e_1,$$

where  $e_1 := [1 \ 0_{1 \times (n-1)}]^t$  and  $\mu$  is a positive number to be chosen later. The transformation constructed so far and the gain  $G$  yield

$$\begin{aligned} F &= S^{-1}VJV^{-1}S - GCS \\ &= \bar{S}^{-1}V^{-1}VJV^{-1}V\bar{S} - \bar{S}^{-1}\mu e_1 CV\bar{S} \\ &= \bar{S}^{-1}[J - \mu e_1 1_{1 \times n}]\bar{S}. \end{aligned}$$

We now need to find  $\mu, \beta, \gamma$  rendering  $F$  Hurwitz and Metzler. Let  $\tilde{F} = \frac{\det \bar{S}}{\gamma^{n-3}} F$  ( $\tilde{F}$  is Metzler if and only if  $F$  is Metzler) and  $\bar{T} := \frac{\det \bar{S}}{\gamma^{n-3}} \bar{S}^{-1}$ . Then

$$\begin{aligned} \tilde{F} &= \bar{T}[J - \mu e_1 1_{1 \times n}]\bar{S} \\ &= \bar{T} \begin{bmatrix} \bar{\pi}_{\beta, \gamma} & [\lambda_1(1 - \gamma) - \mu] 1_{1 \times (n-1)} \\ -\beta [\lambda_2 \ \cdots \ \lambda_n]^t & \gamma \text{diag}\{\lambda_2, \dots, \lambda_n\} \end{bmatrix}, \end{aligned}$$

where  $\bar{\pi}_{\beta, \gamma} := (\lambda_1 - \mu)(1 - \beta) + (n - 1)\mu\beta$ . Thus,

$$\begin{bmatrix} \tilde{F}_{11} \\ \tilde{F}_{21} \\ \vdots \\ \tilde{F}_{n1} \end{bmatrix} = \begin{bmatrix} \gamma^2 \bar{\pi}_{\beta, \gamma} + \beta \gamma (1 - \gamma) \sum_{i=2}^n \lambda_i \\ \beta \gamma \bar{\pi}_{\beta, \gamma} - \beta \pi_{\beta, \gamma} \lambda_2 + \beta^2 (1 - \gamma) \sum_{i=2}^n \lambda_i \\ \vdots \\ \beta \gamma \bar{\pi}_{\beta, \gamma} - \beta \pi_{\beta, \gamma} \lambda_n + \beta^2 (1 - \gamma) \sum_{i=2}^n \lambda_i \end{bmatrix},$$

and for  $k = 2, \dots, n$ ,

$$\begin{bmatrix} \tilde{F}_{1k} \\ \tilde{F}_{2k} \\ \vdots \\ \tilde{F}_{nk} \end{bmatrix} = \begin{bmatrix} \gamma^2 [(\lambda_1 - \lambda_k)(1 - \gamma) - \mu] \\ \beta \gamma [(\lambda_1 - \lambda_k)(1 - \gamma) - \mu] + \delta_{2,k} \gamma \pi_{\beta, \gamma} \lambda_k \\ \vdots \\ \beta \gamma [(\lambda_1 - \lambda_k)(1 - \gamma) - \mu] + \delta_{n,k} \gamma \pi_{\beta, \gamma} \lambda_k \end{bmatrix},$$

where  $\delta_{i,j} = 1$  if  $i = j$  and  $\delta_{i,j} = 0$  otherwise.

From this computation, we derive some properties of  $\tilde{F}$ , which can be easily seen.

P1.  $\forall k = 3, \dots, n, \tilde{F}_{21} \leq \tilde{F}_{k1}$ .

P2.  $\forall k = 2, \dots, n, \tilde{F}_{1k} \geq 0$  if and only if  $\tilde{F}_{jk} \geq 0, j = 2, \dots, n, j \neq k$ .

P3.  $\forall k = 3, \dots, n, \tilde{F}_{12} \leq \tilde{F}_{1k}$ .

Note that the properties P1–P3 ensure that  $\tilde{F}$  is Metzler if and only if  $\tilde{F}_{12} \geq 0$ ,  $\tilde{F}_{21} \geq 0$ . Moreover, since  $F$  is similar to  $J - \mu e_1 1_{1 \times n}$  whose spectrum is  $\{\lambda_1 - \mu, \lambda_2, \dots, \lambda_n\}$ , it follows that  $F$  is Hurwitz and Metzler if there exist  $\mu, \beta \in (0, \bar{\beta}^*)$ , and  $\gamma \in (0, \bar{\gamma}^*)$  such that the following inequalities hold:

$$(4.21) \quad \lambda_1 - \mu < 0,$$

$$(4.22) \quad \gamma \bar{\pi}_{\beta, \gamma} - \pi_{\beta, \gamma} \lambda_2 + \beta(1 - \gamma) \sum_{i=2}^n \lambda_i \geq 0,$$

$$(4.23) \quad (\lambda_1 - \lambda_2)(1 - \gamma) - \mu \geq 0.$$

We choose  $\mu$  in  $(\lambda_1, \lambda_1 - \lambda_2)$  and define

$$\gamma^* = \min \left\{ \bar{\gamma}^*, \frac{\lambda_1 - \lambda_2 - \mu}{\lambda_1 - \lambda_2} \right\}.$$

Then, for each  $\gamma$  in  $(0, \gamma^*)$ , inequalities (4.21) and (4.23) hold. With  $\mu$  and  $\gamma$  chosen above, it remains to choose  $\beta$  for (4.22). To do this, we expand the left-hand side of (4.22) as follows:

$$\begin{aligned} & \gamma \bar{\pi}_{\beta, \gamma} - \pi_{\beta, \gamma} \lambda_2 + \beta(1 - \gamma) \sum_{i=2}^n \lambda_i \\ &= -\beta \gamma (\lambda_1 - \lambda_2 - n\mu) + \beta(1 - \gamma) \sum_{i=2}^n (-\lambda_2 + \lambda_i) \\ & \quad + (\lambda_1 - \lambda_2 - \mu) \gamma. \end{aligned}$$

Thus, inequality (4.22) is equivalent to

$$(4.24) \quad (\lambda_1 - \lambda_2 - \mu) \gamma \geq \Upsilon(\gamma, \mu, \lambda_1, \dots, \lambda_n) \beta,$$

where  $\Upsilon(\gamma, \mu, \lambda_1, \dots, \lambda_n) = \gamma(\lambda_1 - \lambda_2 - n\mu) + (1 - \gamma) \sum_{i=2}^n (\lambda_2 - \lambda_i)$ . This inequality (hence (4.22)) holds for any  $\beta \in (0, \beta^*)$  ( $\beta^*$  may depend on  $\gamma$  and  $\mu$ ), where

$$\beta^* = \begin{cases} \bar{\beta}^* & \text{if } \Upsilon \leq 0, \\ \min \left\{ \bar{\beta}^*, \frac{(\lambda_1 - \lambda_2 - \mu) \gamma}{\Upsilon(\gamma, \mu, \lambda_1, \dots, \lambda_n)} \right\} & \text{if } \Upsilon > 0. \end{cases}$$

Thus, the proof is complete.  $\square$

One novel feature of the matrix  $\bar{S}(\beta, \gamma)$  is that its inverse is fully known, and it is easy to add additional degrees of freedom to this matrix since the parameters are nothing but the convex interpolation coefficients. Moreover, the first column of  $\bar{S}^{-1}(\beta, \gamma)$  is positive, which makes the choice of  $G$  easy.

When the multiplicity of the unstable eigenvalue is greater than 1, it is natural to use the generalized eigenvectors to generalize the idea used here. But this case does not seem to be easy since the scaling involved to make  $Cv_i = 1$  does not hold anymore. This case may require additional degrees of freedom in the scaling matrix, while two parameters are sufficient in the case of distinct eigenvalues.

Similar problems arise when one deals with complex eigenvalues. A basic idea would be to use the real Jordan block representation of a given matrix [18]. When one transforms  $A$  into the real Jordan block, there are two real eigenvectors corresponding

to a pair of conjugate eigenvalues, and they are related in some sense (one of them should be the real part of a complex eigenvector and the other the complex part of it); thus the scaling approach does not apply directly. Moreover, the real Jordan form is not Metzler if a matrix has a complex eigenvalue, which makes the problem more difficult. To illustrate this point, we provide an example where  $A$  has complex eigenvalues. This example shows how to parameterize the matrix  $\bar{S}$  when we use the real Jordan form to solve the problem.

*Example 4.10.* The main idea used in this subsection is applied to the system

$$(4.25) \quad \dot{x} = Ax = \frac{1}{3} \begin{bmatrix} 1 - \sqrt{3} & 1 + \sqrt{3} & 1 \\ 1 & 1 - \sqrt{3} & 1 + \sqrt{3} \\ 1 + \sqrt{3} & 1 & 1 - \sqrt{3} \end{bmatrix} x, \quad y = Cx = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} x,$$

where  $\sigma(A) = \{1, -\frac{1}{2}\sqrt{3} \pm j\frac{1}{2}\}$ . Note that there is no  $K \in \mathbb{R}_+^3$  such that  $A - KC$  is Hurwitz and Metzler by Theorem 3.2.

We consider a real Jordan form  $A_J$  of  $A$ :

$$AV = VA_J, \quad V := \begin{bmatrix} 1 & -1 & \sqrt{3} \\ 1 & 2 & 0 \\ 1 & -1 & -\sqrt{3} \end{bmatrix}, \quad A_J := \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\sqrt{3}/2 & -1/2 \\ 0 & 1/2 & -\sqrt{3}/2 \end{bmatrix}.$$

With three parameters for  $\bar{S}$ ,  $S$  is defined by

$$S = V\bar{S}, \quad \bar{S} := \begin{bmatrix} 1 - \beta & 1 - \gamma_1 & 1 - \gamma_2 \\ -\beta & \gamma_1 & 0 \\ -\beta & 0 & \gamma_2 \end{bmatrix}.$$

Choosing  $\beta = 0.1$ ,  $\gamma_1 = 0.12$ , and  $\gamma_2 = 0.16$  yields

$$S = \begin{bmatrix} 0.8268 & 0.7600 & 1.1171 \\ 0.7000 & 1.1200 & 0.8400 \\ 1.1732 & 0.7600 & 0.5629 \end{bmatrix}, \quad S^{-1}AS = \begin{bmatrix} -0.1362 & 0.6149 & 0.9981 \\ 1.0249 & -0.3536 & 0.1650 \\ 0.1437 & 0.7593 & -0.2422 \end{bmatrix},$$

$$CS = \begin{bmatrix} 0.8268 & 0.7600 & 1.1171 \end{bmatrix}.$$

Thus, if we select  $G = [0.8090 \quad 0.1477 \quad 0.1737]^t$  (from Theorem 3.2), it follows that

$$F = \begin{bmatrix} -0.8051 & 0 & 0.0942 \\ 0.9027 & -0.4659 & 0 \\ 0 & 0.6272 & -0.4363 \end{bmatrix},$$

which is Hurwitz ( $\sigma(F) = \{-0.1626, -0.7724 \pm j0.2863\}$ ) and Metzler, and the positive observer is given by

$$\begin{aligned} \dot{\hat{z}} &= \begin{bmatrix} -0.8051 & 0 & 0.0942 \\ 0.9027 & -0.4659 & 0 \\ 0 & 0.6272 & -0.4363 \end{bmatrix} \hat{z} + \begin{bmatrix} 0.8090 \\ 0.1477 \\ 0.1737 \end{bmatrix} y, \\ \hat{x} &= \begin{bmatrix} 0.8268 & 0.7600 & 1.1171 \\ 0.7000 & 1.1200 & 0.8400 \\ 1.1732 & 0.7600 & 0.5629 \end{bmatrix} \hat{z}. \end{aligned}$$



Note that once  $S$  is chosen,  $G$  can be obtained using Theorem 3.2 with the pair  $(S^{-1}AS, CS)$ . The estimate  $\hat{x}$  of  $x$  resides in the set  $\{S\mu | \mu \in \mathbb{R}_+^3\}$ , which is a proper subset of  $\mathbb{R}_+^3$ .

**5. Positive linear observers via positive realization.** In this section we first formulate the observer design problem as a positive realization problem, and then we propose a novel observer design method.

**5.1. Positive realization approach.** As mentioned in the previous section, the inverse eigenvalue problem is a hard obstacle when we fix the dimension of  $F$ . In this section, we reformulate the problem allowing the dimension of  $F$  to be an additional degree of freedom. To do this, the positive linear observer is defined as follows.

DEFINITION 5.1. *Given a positive linear system (3.1), a linear system*

$$(5.1) \quad \begin{aligned} \dot{\hat{z}} &= F\hat{z} + Gy, \quad \hat{z} \in \mathbb{R}^N, \\ \hat{x} &= H\hat{z}, \quad \hat{x} \in \mathbb{R}^n, \end{aligned}$$

*is called a positive linear observer for (3.1) if  $F$  is Hurwitz and Metzler,  $G \in \mathbb{R}_+^{N \times q}$ ,  $H \in \mathbb{R}_+^{n \times N}$ , and  $\lim_{t \rightarrow \infty} \|H\hat{z}(t) - x(t)\| = 0 \forall \hat{z}(0) \in \mathbb{R}_+^N$  and  $\forall x(0) \in \mathbb{R}_+^n$ .*

Two natural questions arise: (1) Under what conditions is the existence of a positive linear observer guaranteed? (2) How can one construct  $(F, G, H)$ ?

If one recalls the observer design for linear systems, the first thing to do would be to choose the spectrum of  $F$  (if  $N = n$  and  $(A, C)$  is an observable pair, one can assign the spectrum of  $F$  arbitrarily) and compute the appropriate gain matrix  $G$ . This process does not seem to work well in our problem because the requirement that  $F$  be a Metzler matrix restricts the spectrum of  $F$ . As discussed in the previous section, this restriction is a very hard obstacle in our problem.

To circumvent this problem, we do not fix the dimension at the first stage. In other words, the dimension of the observer  $N$  is an additional degree of freedom. To be more precise, note first that  $(F, G, H)$  is a positive realization of some transfer function  $G_o(s)$ . (By a positive realization of a transfer function  $G_o(s)$ , we mean a realization  $(F, G, H)$  such that  $F$  is a Metzler matrix,  $G \geq 0$ , and  $H \geq 0$ . For an outstanding tutorial on this problem, the readers are referred to [5]. See also the references cited in that paper, for example, [2, 3, 12, 24].) Suppose we have designed  $(F_m, G_m, H_m)$  such that (1)  $F_m \in \mathbb{R}^{N \times N}$  is Hurwitz (not necessarily Metzler),  $G_m \in \mathbb{R}^{N \times q}$ , and  $H_m \in \mathbb{R}^{n \times N}$ , (2)  $(F_m, G_m)$  is a controllable pair and  $(F_m, H_m)$  is an observable pair, and (3)  $\dot{\hat{z}} = F_m\hat{z} + G_my$ ,  $\hat{x} = H_m\hat{z}$  is a linear observer (not necessarily a positive system) for (3.1). If  $G_o(s) = H_m(sI - F_m)^{-1}G_m$  admits a positive realization  $(F, G, H)$ , then system (5.1) can be a positive linear observer for system (3.1).

We now recall some results on the positive realization problem.

THEOREM 5.2 (see [12, Theorem 3.1]). *Let  $G_o(s)$  be a rational transfer function.  $G_o(s)$  has a positive realization if and only if*

1. *the impulse response function  $g_o(t)$  is positive, i.e.,  $g_o(t) > 0$  for every  $t > 0$  and  $g_o(0) \geq 0$ ;*
2. *there is a unique (possibly multiple) pole of  $G_o(s)$  with maximal real part.*

Note that this result is for scalar transfer functions. For a general matrix transfer function  $G_o(s)$ , it has been shown in [14, Theorem 2.4] (although it is for discrete time transfer functions, the same argument applies to continuous time ones) that  $G_o(s)$  has a positive realization if and only if each entry of  $G_o(s)$  has a positive realization. Thus, we require each component of  $G_o(s)$ , which is an  $n \times q$  matrix of transfer functions, to have a positive realization.

Based on this observation, the inverse eigenvalue problem does not appear anymore if  $(F_m, G_m, H_m)$  is appropriately chosen such that each component of  $H_m(sI - F_m)^{-1}G_m$  satisfies the conditions of this theorem. It is essential to recall that the positive realization problem is equivalent to finding a finite-dimensional invariant cone  $\mathcal{M}$  generated by the columns of some matrix  $M$  ( $\mathcal{M}$  will be denoted by Cone  $M$ ) and that enlarging the dimension to obtain a positive realization is mandatory in many cases [4], which corresponds to making  $N$  a design freedom in our problem. Recall that once  $M$  is found,  $(F, G, H)$  can be obtained solving the following equations [24, Theorem 5]:

$$(5.2) \quad F_m M = F M, \quad G_m = G M, \quad H = H_m M.$$

One possible way to choose  $(F_m, G_m, H_m)$  is  $(A - KC, K, I)$ , i.e., a matrix  $K \in \mathbb{R}^{n \times q}$  such that  $A - KC$  is Hurwitz (detectability of  $(A, C)$  guarantees this). Note that  $(A - KC, I)$  is observable, while  $(A - KC, K)$  may not be controllable (but it is at least stabilizable).

Now, we state the main result of this section.

**THEOREM 5.3.** *Consider a positive linear system (3.1) with  $(A, C)$  detectable. If there exists  $K \in \mathbb{R}^{n \times q}$  such that  $A - KC$  is a Hurwitz matrix and the transfer function  $G_o(s) = (sI - A + KC)^{-1}K$  admits a positive realization, then there exists a positive realization  $(F, G, H)$ , with  $F$  being Hurwitz, such that the dynamic system (5.1) is a positive linear observer for system (3.1).*

*Proof.* Let  $A_K = A - KC$  and let  $T$  be the transformation matrix decomposing  $(A_K, K, I)$  into controllable and uncontrollable parts [8],

$$TA_K = \begin{bmatrix} \bar{A}_{K,c} & \bar{A}_{12} \\ 0 & \bar{A}_{K,\bar{c}} \end{bmatrix} T, \quad TK = \begin{bmatrix} \bar{K}_c \\ 0 \end{bmatrix}, \quad T^{-1} = \begin{bmatrix} \bar{C}_c & \bar{C}_{\bar{c}} \end{bmatrix},$$

such that  $(\bar{A}_{K,c}, \bar{K}_c)$  is controllable. Then  $G_o(s) = \bar{C}_c(sI - \bar{A}_{K,c})^{-1}\bar{K}_c$ .

Let  $(\bar{F}, \bar{G}, \bar{H})$  be a positive realization of  $G_o(s)$ . Note that this realization may not be Hurwitz (i.e., there may be a positive real eigenvalue which is uncontrollable or unobservable). Following the same procedure of [2, Theorem 3.2], one can exclude these modes to obtain a reduced order positive realization  $(F, G, H)$  such that  $\lambda_{\max}(F)$  is a pole of  $G_o(s)$ , i.e.,  $F$  is Hurwitz.

Now, suppose  $(F, G, H)$ , with  $F$  being Hurwitz, is a positive realization of  $G_o(s)$  with a cone  $\mathcal{M}$  generated by  $M$ , i.e.,

$$(5.3) \quad \bar{A}_{K,c}M = MF, \quad \bar{K}_c = MG, \quad H = \bar{C}_cM.$$

Since  $A_K T^{-1} = A_K \begin{bmatrix} \bar{C}_c & \bar{C}_{\bar{c}} \end{bmatrix} = \begin{bmatrix} \bar{C}_c & \bar{C}_{\bar{c}} \end{bmatrix} \begin{bmatrix} \bar{A}_{K,c} & \bar{A}_{12} \\ 0 & \bar{A}_{K,\bar{c}} \end{bmatrix} = T^{-1} \begin{bmatrix} \bar{A}_{K,c} & \bar{A}_{12} \\ 0 & \bar{A}_{K,\bar{c}} \end{bmatrix}$ , it follows that  $A_K \bar{C}_c = \bar{C}_c \bar{A}_{K,c}$ . Similarly, one has  $\bar{C}_c \bar{K}_c = K$ . Hence,

$$\begin{aligned} AH - HGCH &= A_K H + KCH - HGCH \\ &= A_K \bar{C}_c M + \bar{C}_c \bar{K}_c CH - HGCH \\ &= \bar{C}_c \bar{A}_{K,c} M + \bar{C}_c M GCH - HGCH \\ &= HF. \end{aligned}$$

Let  $e = \hat{x} - x$ . Then

$$\begin{aligned}\dot{e} &= HF\hat{z} + HGCx - Ax \\ &= (AH - HGCH)\hat{z} + HGCx - Ax \\ &= A\hat{x} - HGC\hat{x} + HGCx - Ax \\ &= (A - \bar{C}_c\bar{K}_cC)e \\ &= (A - KC)e.\end{aligned}$$

Thus, the assertion follows.  $\square$

**THEOREM 5.4.** *Consider the positive linear observer (5.1) for system (3.1). Suppose that there exists  $K$  which satisfies the assumptions of Theorem 5.3, and that  $(F, G, H)$  is a positive realization of  $G_o(s) = (sI - A + KC)^{-1}K$ . Let  $n_c$  be the dimension of the controllable part of  $(A - KC, K)$ . Then*

1.  $K \geq 0$ ;
2.  $\text{rank } H = n_c$ ;
3.  $AH - HF = HGCH$ .

*Proof.* The fact  $K \geq 0$  follows from  $g_o(0) = K$ , and the proof of Theorem 5.3 ensures the last statement.

For the second statement, let  $F \in \mathbb{R}^{N \times N}$  and consider the decomposition derived in the proof of Theorem 5.3. We first show that  $\text{rank } M = n_c$ . Since  $M \in \mathbb{R}^{n_c \times N}$ , it follows that  $\text{rank } M \leq n_c$ . Suppose  $\text{rank } M < n_c$ . Then there exists a nonzero row vector  $w$  such that  $wM = 0$ , which results in  $w\bar{K}_c = wMG = 0$ . Moreover, one has  $w\bar{A}_{K,c}\bar{K}_c = 0$  since  $w\bar{A}_{K,c}\bar{K}_c = wMFG = 0$ . Repeating this process results in

$$w \begin{bmatrix} \bar{K}_c & \bar{A}_{K,c}\bar{K}_c & \cdots & \bar{A}_{K,c}^{n_c-1}\bar{K}_c \end{bmatrix} = 0,$$

which contradicts the controllability of  $(\bar{A}_{K,c}, \bar{K}_c)$ . Thus,  $\text{rank } M = n_c$ . From the fact that  $\bar{C}_c$  has full column rank and  $M$  has full row rank, it follows that  $\text{rank } H = n_c$ .  $\square$

From the proof of Theorem 5.3, the following result is straightforward.

**COROLLARY 5.5.** *Consider the positive linear system (3.1). If there exist  $F, G$ , and  $H$  such that (1)  $F$  is Metzler and Hurwitz,  $H \geq 0$ ,  $G \geq 0$ , (2)  $AH - HF = HGCH$ , and (3)  $A - HGC$  is Hurwitz, then the system (5.1) is a positive linear observer for system (3.1).*

**Remark 5.6.** 1. Note that the estimate  $\hat{x}(t)$  will reside in the cone generated by the columns of  $H$ . Thus, the initial condition  $\hat{x}(0)$  of the observer is confined to Cone  $H$ , which may not coincide with  $\mathbb{R}_+^n$ . This reduction of the set of initial conditions  $\hat{x}(0)$  can be thought of as the price for enlarging the class of systems that admit positive linear observers.

2. For controlled systems ( $\dot{x} = Ax + Bu$ ,  $y = Cx$ ), one can reformulate the problem by considering the transfer function matrix  $G_o(s) = (sI - A + KC)^{-1} \begin{bmatrix} K & B \end{bmatrix}$ .

Recall that Cone  $H$  is said to be solid if it contains an open ball in  $\mathbb{R}^n$ , i.e.,  $\text{rank } H = n$ . Theorem 5.4 ensures that if  $K$  is chosen such that  $(A - KC, K)$  is controllable, then for each interior point  $\bar{x}$  of Cone  $H$  there exists a neighborhood  $\mathcal{N}$  centered at  $\bar{x}$  such that  $x(0) \in \mathcal{N}$  implies the existence of  $\hat{x}(0) \in \mathcal{N}$  such that  $x(t) = \hat{x}(t) \forall t \geq 0$ . Note that if Cone  $H$  is not solid, this is not possible. Therefore, one might want Cone  $H$  to be solid. To this end, we provide the following result.

**THEOREM 5.7.** *For the positive linear system (3.1), suppose  $(A, C)$  is an observable pair. Suppose there is  $\bar{K} \in \text{int } \mathbb{R}_+^{n \times q}$  such that the transfer function matrix  $G_o^{\bar{K}}(s) := (sI - A + \bar{K}C)^{-1}\bar{K}$  satisfies the following conditions:*

1.  *$A - \bar{K}C$  has a simple and unique eigenvalue denoted by  $\lambda_{\max}$  whose real part is greater than any other eigenvalues, and  $G_o^{\bar{K}}(s)$  has a unique dominant pole denoted by  $p_{\max}$ . Moreover,  $p_{\max} = \lambda_{\max} < 0$ .*
2. *Let  $G_{o,ij}^{\bar{K}}(s)$  be the  $(i, j)$ th component of  $G_o^{\bar{K}}(s)$ . Then every zero of  $G_{o,ij}^{\bar{K}}(s)$  has real part less than  $p_{\max}$  and each  $G_{o,ij}^{\bar{K}}(s)$  has a simple and unique dominant pole  $p_{\max,ij}$  which is equal to  $p_{\max}$ .*
3.  *$G_o^{\bar{K}}(s)$  admits a positive realization  $(\bar{F}, \bar{G}, \bar{H})$ , which is a positive linear observer.*

*Then there exists a positive linear observer  $(F, G, H)$  of the form (5.1) with Cone  $H$  being solid.*

*Proof.* At first, note that  $\lambda_{\max}$  is real by assumption 1. It is assumed that  $n \geq 2$  since the case  $n = 1$  is trivial.

Consider the space  $\mathbb{R}^{n \times q}$ , equipped with the topology induced by some matrix norm, and let  $\delta > 0$  be such that the neighborhood of radius  $\delta$  of the matrix  $\bar{K}$ ,  $\mathcal{N}_\delta(\bar{K})$ , is included in  $\mathbb{R}_+^{n \times q}$ . Then observability of  $(A, C)$  guarantees that for almost every  $K \in \mathcal{N}_\delta(\bar{K})$ , the pair  $(A, K)$  (and equivalently  $(A - KC, K)$ ) is controllable.

Define  $\epsilon_1 = \min\{|\lambda_i - \lambda_j| \mid \lambda_i, \lambda_j \in \sigma(A - \bar{K}C), \lambda_i \neq \lambda_j\}$ . Recalling that the eigenvalues of  $(A - KC)$  change continuously w.r.t.  $K$ , we can find  $\delta_1 > 0$  such that  $\forall K \in \mathcal{N}_{\delta_1}(\bar{K})$ ,

$$\begin{aligned} & \text{dist}(\sigma(A - KC), \sigma(A - \bar{K}C)) \\ &:= \min\{|\lambda_i - \lambda_j| \mid \lambda_i \in \sigma(A - KC), \lambda_j \in \sigma(A - \bar{K}C)\} < \frac{1}{3}\epsilon_1. \end{aligned}$$

By assumption 2,  $G_{o,ij}^{\bar{K}}(s)$  has the form

$$G_{o,ij}^{\bar{K}}(s) = \frac{\pi_{ij}(\bar{K})(s - z_1) \cdots (s - z_{\bar{n}(i,j)})}{(s - p_1) \cdots (s - p_{\bar{d}(i,j)})}, \quad z_l \neq p_m \quad \forall l, m, \bar{d}(i, j) > \bar{n}(i, j),$$

where  $z_l \in \mathbb{C}$  and  $p_l \in \mathbb{C}$  represent zeros and poles of  $G_{o,ij}^{\bar{K}}(s)$ , respectively,  $\text{Re } z_l < 0$ ,  $\text{Re } p_l < 0$ , and  $\pi_{ij}(\cdot)$  is a continuous function defined on  $\mathcal{N}_{\delta_1}(\bar{K})$  with  $\pi_{ij}(\bar{K}) \neq 0$ . Note that  $z_l$ 's and  $p_l$ 's (and  $\mu_l$ 's and  $\eta_l$ 's used in (5.4)) vary with  $i, j$ . However, for simplicity of notation, we drop the explicit dependence on  $i, j$  whenever not needed.

We fix  $p_1 = p_{\max} \forall G_{o,ij}^{\bar{K}}(s)$  (by assumption 2) and define

$$\epsilon_2 = \min_{1 \leq i \leq n, 1 \leq j \leq q} \{|z_l - p_1| \mid 1 \leq l \leq \bar{n}(i, j)\}.$$

Define a transfer function  $G_o^K(s) = (sI - A + KC)^{-1}K$  parameterized by  $K \in \mathcal{N}_{\delta_1}(\bar{K})$ , and let  $G_{o,ij}^K(s)$  be the  $(i, j)$ th component of  $G_o^K(s)$ , which can be described by

$$(5.4) \quad G_{o,ij}^K(s) = \frac{\pi_{ij}(K)(s - z_1 + \mu_1(K)) \cdots (s - z_{\bar{n}(i,j)} + \mu_{\bar{n}(i,j)}(K))}{(s - p_1 + \eta_1(K)) \cdots (s - p_{\bar{d}(i,j)} + \eta_{\bar{d}(i,j)}(K))} \tilde{G}_{o,ij}(s),$$

where

$$(5.5) \quad \tilde{G}_{o,ij}(s) = \frac{(s - p_{\bar{d}(i,j)+1} + \mu_{\bar{d}(i,j)+1}(K)) \cdots (s - p_{d(i,j)} + \mu_{d(i,j)}(K))}{(s - p_{\bar{d}(i,j)+1} + \eta_{\bar{d}(i,j)+1}(K)) \cdots (s - p_{d(i,j)} + \eta_{d(i,j)}(K))}$$

and  $\eta_l$ 's and  $\mu_l$ 's are (complex valued) continuous functions defined on  $\mathcal{N}_{\delta_1}(\overline{K})$  and vanishing at  $\overline{K}$ . Note that for any real  $K$ ,  $\eta_i$  and  $\mu_i$  render  $G_{o,ij}^K(s)$  a rational function whose coefficients are real numbers.

Since  $p_1$  is real,  $\eta_1(K)$  is real by the way  $\delta_1$  is chosen. The transfer function  $\tilde{G}_{o,ij}(s)$  describes the pole zero cancellation (thus becoming unity) when  $K$  happens to be  $\overline{K}$ .

It is important to note that there is no restriction on  $\mu_i$ 's in (5.4)–(5.5) at this moment; i.e., it might happen that  $p_1 - \eta_1(K) = z_i - \mu_i(K)$ . Thus, we take  $\delta_2 > 0$  such that the poles and zeros of  $G_{o,ij}^K(s)$  do not change significantly. Indeed, by continuity and assumption 2, there exists  $0 < \delta_2 < \delta_1$  such that  $\forall K \in \mathcal{N}_{\delta_2}(\overline{K})$ , it holds that

$$\max \left\{ \max_{l \in I_{\mu,ij}} \{|\mu_l(K)|\}, \max_{l \in I_{\eta,ij}} \{|\eta_l(K)|\} \right\} < \frac{1}{3} \min\{\epsilon_1, \epsilon_2\} \quad \forall 1 \leq i \leq n, 1 \leq j \leq q,$$

$$p_1 - \eta_1^{ij}(K) < 0 \quad \forall 1 \leq i \leq n, 1 \leq j \leq q,$$

where  $I_{\mu,ij} = \{1, \dots, \bar{n}(i, j), \bar{d}(i, j) + 1, \dots, d(i, j)\}$  and  $I_{\eta,ij} = \{1, \dots, d(i, j)\}$ . (Note that  $I_{\mu,ij}$  and  $I_{\eta,ij}$  represent the set of indices appearing in (5.4) and (5.5).) By the way  $\delta_2$  is chosen, it is guaranteed that  $\forall K \in \mathcal{N}_{\delta_2}(\overline{K})$ , the pole  $p_1 - \eta_1(K)$  is the dominant pole of  $G_{o,ij}^K(s)$  and is a negative real number.

Rewrite  $G_{o,ij}^K(s)$  as

$$G_{o,ij}^K(s) = \frac{s - p_1}{s - p_1 + \eta_1(K)} G_{o,ij}^{\overline{K}}(s) \hat{G}_{o,ij}(s) \tilde{G}_{o,ij}(s),$$

where

$$\hat{G}_{o,ij}(s) := \frac{\pi_{ij}(K)}{\pi_{ij}(\overline{K})} \prod_{l=2}^{\bar{d}(i,j)} \frac{s - p_l}{s - p_l + \eta_l(K)} \cdot \prod_{l=1}^{\bar{n}(i,j)} \frac{s - z_l + \mu_l(K)}{s - z_l}.$$

It follows from the structure of  $\hat{G}_{o,ij}(s)$  and  $\tilde{G}_{o,ij}(s)$  that

$$G_{o,ij}^{\overline{K}}(s) \hat{G}_{o,ij}(s) \tilde{G}_{o,ij}(s) = G_{o,ij}^{\overline{K}}(s) + G_{o,ij}^{*K}(s),$$

where  $G_{o,ij}^{*K}(s)$  is a transfer function parameterized by  $K$  in a continuous way. Note that every pole of  $G_{o,ij}^{*K}(s)$  (i.e.,  $p_l - \eta_l(K)$ ,  $l = 2, \dots, d(i, j)$ , and  $z_l$ ,  $l = 1, \dots, \bar{n}(i, j)$ ) resides in  $\mathbb{C}^-$  and the dominant pole of  $G_{o,ij}^{*K}(s)$  is  $p_1$ .

Consider the impulse response functions of  $G_{o,ij}^{\overline{K}}(s)$  and  $G_{o,ij}^{*K}(s)$  denoted by  $g_{o,ij}^{\overline{K}}(t)$  and  $g_{o,ij}^{*K}(t)$ , respectively. Since  $p_1$  is the dominant pole of these two transfer functions, we can rewrite the impulse response functions as

$$\begin{aligned} g_{o,ij}^{\overline{K}}(t) &= e^{p_1 t} \bar{f}_{ij}(t), \quad \bar{f}_{ij}(t) > 0 \quad \forall t, \\ g_{o,ij}^{*K}(t) &= e^{p_1 t} f_{ij}^{*K}(t), \end{aligned}$$

where  $\bar{f}_{ij}(t)$  and  $f_{ij}^{*K}(t)$  can be expressed by the sum of a real constant (the coefficient associated to the dominant pole  $p_1$ ) and exponential functions (multiplied by some power of  $t$ , in general) which decay to zero as  $t \rightarrow \infty$ . Note that the coefficients of the exponential functions associated to  $f_{ij}^{*K}(t)$  are continuous functions of  $K$  and become zero when  $K = \overline{K}$ , i.e.,  $f_{ij}^{*\overline{K}}(t) = 0 \quad \forall t \geq 0$ . It is seen that there exists

$\epsilon_3 > 0$  such that  $\bar{f}_{ij}(t) > \epsilon_3 \ \forall t \geq 0$ . In fact, note that  $\bar{f}_{ij}(t)$  can be decomposed as  $\bar{f}_{ij}(t) = \hat{f}_{ij} + \tilde{f}_{ij}(t)$ , where  $0 < \hat{f}_{ij} \in \mathbb{R}$  and  $\lim_{t \rightarrow \infty} \tilde{f}_{ij}(t) = 0$  (exponential terms). Then there exists  $T > 0$  such that  $|\tilde{f}_{ij}(t)| < \hat{f}_{ij}/2 \ \forall t > T$ . Thus, the claim follows by taking  $\epsilon_3 = \frac{1}{2} \min\{\min_{t \in [0, T]} \bar{f}_{ij}(t), \hat{f}_{ij}\}$ .

By continuity, there exists  $\delta_3 > 0$  ( $\delta_3 < \delta_2$ ) such that  $\forall K \in \mathcal{N}_{\delta_3}(\bar{K})$ ,  $|f_{ij}^{*K}(t)| < \epsilon_3/2$ . This ensures that there exist  $\underline{f}_{ij}, \bar{f}_{ij}$  such that  $0 < \underline{f}_{ij} < \bar{f}_{ij}(t) + f_{ij}^{*K}(t) := f_{ij}(t) < \bar{f}_{ij}$ . Therefore, it follows that  $\forall K \in \mathcal{N}_{\delta_3}(\bar{K})$ ,

$$0 < \underline{f}_{ij} e^{p_1 t} \leq \mathcal{L}^{-1} \left[ G_{o,ij}^{\bar{K}}(s) \hat{G}_{o,ij}(s) \tilde{G}_{o,ij}(s) \right] (t) = e^{p_1 t} f_{ij}(t) \leq \bar{f}_{ij} e^{p_1 t} \quad \forall t \geq 0.$$

To complete the proof we find  $K^* \in \mathcal{N}_{\delta_3}(\bar{K})$  such that  $\eta_1^{ij}(K^*) < 0 \ \forall i, j$ . (Note that  $\eta_1$  of  $G_{o,ij}^K(s)$  is different from that of  $G_{o,\bar{i}\bar{j}}^K(s)$ , in general.) To do this, we write  $A - \bar{K}C$  in the real Jordan form, i.e.,  $A - \bar{K}C = J A_{\bar{K},J} J^{-1}$ . With no loss of generality,  $J_1$  (the first column of  $J$ ) is assumed to be the eigenvector  $v_{\max}$  corresponding to the eigenvalue  $p_{\max}$ .

Choose the  $i^*$ th row of  $C$  such that  $C_{i^*} v_{\max} \neq 0$  (by observability,  $i^*$  exists). Let

$$K^* = -\eta^* v_{\max} e_{i^*}^t, \quad \eta^* := \frac{\xi \delta_3}{C_{i^*} v_{\max}},$$

where  $e_{i^*} \in \mathbb{R}^q$  is the vector whose  $i^*$ th component is one, while the others are zero, and  $\xi > 0$  is chosen sufficiently small such that  $K^* + \bar{K} \in \mathcal{N}_{\delta_3/2}(\bar{K})$ . Thus,  $K^*$  has just one nonzero column (the  $i^*$ th column). From the relation

$$A - \bar{K}C - K^*C = J [A_{\bar{K},J} + \xi \delta_3 e_1 [1 \quad * \quad \cdots \quad *]] J^{-1},$$

it follows that the maximal real eigenvalue of  $A - \bar{K}C - K^*C$  is shifted to  $p_{\max} + \xi \delta_3$ . Thus, by continuity, there exists  $\delta_4 > 0$  such that  $\forall K \in \mathcal{N}_{\delta_4}(\bar{K} + K^*)$ ,  $\eta_1^{ij}(K) < 0 \ \forall 1 \leq i \leq n, 1 \leq j \leq q$  and  $\mathcal{N}_{\delta_4}(\bar{K} + K^*) \subset \mathcal{N}_{\delta_3}(\bar{K})$ .

Pick a  $K \in \mathcal{N}_{\delta_4}(\bar{K} + K^*)$  such that  $(A - KC, K)$  is a controllable pair. From the relation

$$\begin{aligned} g_{o,ij}^K(t) &= \mathcal{L}^{-1} \left[ \frac{s - p_1}{s - p_1 + \eta_1(K)} \right] (t) * \mathcal{L}^{-1} \left[ G_{o,ij}^{\bar{K}}(s) \hat{G}_{o,ij}(s) \tilde{G}_{o,ij}(s) \right] (t) \\ &= e^{p_1 t} f_{ij}(t) - \eta_1(K) \int_0^t e^{(p_1 - \eta_1(K))(t-\tau)} e^{p_1 \tau} f_{ij}(\tau) d\tau > 0 \quad \forall t \geq 0, \end{aligned}$$

where “ $*$ ” represents the convolution integral, and the fact that  $G_{o,ij}^K(s)$  has a simple dominant pole  $p_{\max} - \eta_1(K) < 0$ , we conclude that  $G_{o,ij}^K(s)$  has a positive realization.

Thus, it follows that there exists  $K$  such that  $G_o^K(s)$  has a positive realization  $(F^*, G^*, H^*)$ , and  $(A - KC, K)$  is controllable. By Theorem 5.3, there exists a positive realization  $(F, G, H)$  with  $F$  being Hurwitz such that the realization  $(F, G, H)$  is a positive linear observer. Moreover, by Theorem 5.4 (rank  $H = n$ ), Cone  $H$  is solid. Therefore, the assertion follows.  $\square$

Note that Theorem 5.7 provides only a sufficient condition on the existence of a positive linear observer. It is unclear at this moment whether the converse is true.

Now we present an extension of Theorem 4.9.

**THEOREM 5.8.** *For a single output positive system (3.1), suppose  $A$  is irreducible and  $(A, C)$  is detectable. Let  $\lambda_1 = \lambda_{\max}(A)$ . There exists a positive linear observer*

via positive realization (5.1) if all eigenvalues of  $A$  except  $\lambda_1$  lie in the open left half complex plane.

*Proof.* Since  $A$  is irreducible and  $(A, C)$  is detectable, there exists  $v_{\max} \in \text{int } \mathbb{R}_+^n$  such that  $Av_{\max} = \lambda_{\max}(A)v_{\max}$  and  $Cv_{\max} = 1$ . Let  $K = (\lambda_1 + \epsilon)v_{\max}$ , where  $\epsilon > 0$  is chosen such that  $\lambda^* < -\epsilon < 0$ , where  $\lambda^* = \max\{\text{Re } \lambda \mid \lambda \in \sigma(A), \lambda \neq \lambda_1\}$ . From Theorem 2.4, we have  $\sigma(A - KC) = (\sigma(A) \setminus \{\lambda_1\}) \cup \{-\epsilon\}$  since  $Cv_{\max} = 1$ . That is to say, all the eigenvalues of  $A$  and  $A - KC$  coincide except that  $\lambda_1$  is replaced by  $-\epsilon$ . Thus,  $A - KC$  is Hurwitz and the eigenvalue with maximal real part is  $-\epsilon$ .

Define  $G_o(s) = (sI - A + KC)^{-1}K$ . Then the gain matrix  $K$  chosen above yields

$$\begin{aligned} g_o(t) &= e^{(A-KC)t}K \\ &= e^{(A-(\lambda_1+\epsilon)v_{\max}C)t}(\lambda_1 + \epsilon)v_{\max} \\ &= (\lambda_1 + \epsilon) \left[ \sum_{i=0}^{\infty} \frac{1}{i!} [A - (\lambda_1 + \epsilon)v_{\max}C]^i t^i \right] v_{\max} \\ &= (\lambda_1 + \epsilon) \left[ \sum_{i=0}^{\infty} \frac{1}{i!} (-\epsilon t)^i \right] v_{\max} \\ &= (\lambda_1 + \epsilon)e^{-\epsilon t}v_{\max} > 0, \end{aligned}$$

which implies that  $G_o(s)$  has a positive realization. Therefore, the assertion follows by Theorem 5.3.  $\square$

*Remark 5.9.* 1. The proof of Theorem 5.8 ensures that if  $(A, C)$  is observable, there exists  $\bar{K}$  satisfying all assumptions of Theorem 5.7. Thus, a gain matrix  $K$  sufficiently close to  $\bar{K}$  renders  $\text{rank } H = n$ . Note that if  $(A, C)$  is detectable, this does not hold in general. For example, if  $A = \text{diag}\{1, -1, -1\}$ ,  $C = [1 \ 0 \ 0]$ , then there is no  $K \in \mathbb{R}^3$  such that  $(A, K)$  (equivalently  $(A - KC, K)$ ) is controllable.

2. Theorem 5.8 extends Theorem 4.9. If  $K$  is chosen to be a scalar product of  $v_{\max}$ , then any positive  $\epsilon$  results in a positive observer. In the proof of Theorem 5.8,  $\epsilon$  is selected so that all conditions of Theorem 5.8 are satisfied for the observable case.

3. The observer structure in section 4 is a special case of (5.1). Note that the coordinates transformation is involved during the positive realization procedure.

**5.2. Application to compartmental systems.** In this subsection, we apply the observer design method developed in section 5.1 to compartmental systems. Before proceeding, without loss of generality, assume  $A, C$  are decomposed as in (3.13) and (3.5). Define

$$\begin{aligned} A_- &:= \begin{bmatrix} A_{11} & 0 & 0 \\ \vdots & \ddots & 0 \\ A_{\nu-m,1} & \cdots & A_{\nu-m,\nu-m} \end{bmatrix} \in \mathbb{R}^{(d_1+\cdots+d_{\nu-m}) \times (d_1+\cdots+d_{\nu-m})}, \\ A_+ &:= \text{diag}\{A_{\nu-m+1,\nu-m+1}, \dots, A_{\nu\nu}\}. \end{aligned}$$

Similarly,

$$C_- := \begin{bmatrix} C_{11} & \cdots & C_{1,\nu-m} \\ \vdots & & \vdots \\ C_{q1} & \cdots & C_{q,\nu-m} \end{bmatrix}, \quad C_+ := \begin{bmatrix} C_{1,\nu-m+1} & \cdots & C_{1\nu} \\ \vdots & & \vdots \\ C_{q,\nu-m+1} & \cdots & C_{q\nu} \end{bmatrix}.$$

Then  $A$  and  $C$  can be written as

$$(5.6) \quad A = \begin{bmatrix} A_- & 0 \\ A_* & A_+ \end{bmatrix}, \quad C = [C_- \quad C_+],$$

where  $A_*$  is clearly defined by this. We decompose the state vector  $x$  into  $x_-$  and  $x_+$  according to  $A_-$  and  $A_+$ .

Our observer has the form

$$(5.7) \quad \dot{\hat{z}} = \begin{bmatrix} A_- & 0 \\ 0 & F \end{bmatrix} \hat{z} + \begin{bmatrix} 0 \\ G \end{bmatrix} y, \quad \hat{x} = \begin{bmatrix} I_{d_1+\dots+d_{\nu-m}} & 0 \\ 0 & H \end{bmatrix} \hat{z},$$

where  $F \in \mathbb{R}^{N \times N}$ ,  $G \in \mathbb{R}^{N \times q}$ ,  $H \in \mathbb{R}^{(d_{\nu-m+1}+\dots+d_{\nu}) \times N}$ , and the matrices  $F$ ,  $G$ ,  $H$  and the dimension  $N$  are design parameters. Note that this observer is of the form (5.1).

**THEOREM 5.10.** *Consider the positive system (3.1) where  $A$  is a compartmental matrix and the matrices  $A$  and  $C$  are decomposed as in (5.6). Assume  $q = m$ . If  $C_+ = \text{diag}\{C_{1,\nu-m+1}, \dots, C_{m\nu}\}$  up to renumbering the outputs, then there exists a positive linear observer of the form (5.1) if and only if  $(A, C)$  is detectable.*

*Proof.* Necessity follows trivially.

For sufficiency, it is noted that the pairs  $(A_{\nu-m+1,\nu-m+1}, C_{1,\nu-m+1}), \dots, (A_{\nu\nu}, C_{m\nu})$  are detectable. Let  $v_i$  be the eigenvector associated to the zero eigenvalue of  $A_{ii}$  ( $i = \nu - m + 1, \dots, \nu$ ) such that  $C_{i-\nu+m,i}v_i = 1$ , and choose  $\epsilon_i$  ( $i = \nu - m + 1, \dots, \nu$ ) such that  $\lambda_i^* < -\epsilon_i < 0$ , where  $\lambda_i^* = \max\{\text{Re } \lambda \mid \lambda \in \sigma(A_{ii}), \lambda \neq 0\}$ .

Setting

$$K := \begin{bmatrix} K_- \\ K_+ \end{bmatrix}$$

with  $K_- := 0_{(d_1+\dots+d_{\nu-m}) \times q}$ ,  $K_+ := \text{diag}\{\epsilon_{\nu-m+1}v_{\nu-m+1}, \dots, \epsilon_{\nu}v_{\nu}\}$ , it follows that

$$G_{o+}(s) = (sI - A_+ + K_+C_+)^{-1}K_+ = \text{diag} \left\{ \frac{\epsilon_{\nu-m+1}}{s + \epsilon_{\nu-m+1}}v_{\nu-m+1}, \dots, \frac{\epsilon_{\nu}}{s + \epsilon_{\nu}}v_{\nu} \right\},$$

which has a positive realization

$$F = \text{diag}\{-\epsilon_{\nu-m+1}, \dots, -\epsilon_{\nu}\}, \quad G = \text{diag}\{\epsilon_{\nu-m+1}, \dots, \epsilon_{\nu}\},$$

$$H = \text{diag}\{v_{\nu-m+1}, \dots, v_{\nu}\}.$$

Noting that  $\lim_{t \rightarrow \infty} x_-(t) = 0$ , one can easily prove the convergence of the observer error to zero by modifying the proof of Theorem 5.3.  $\square$

The following result is a direct consequence of Theorem 5.10.

**COROLLARY 5.11.** *Consider the positive system (3.1) where  $A$  is a compartmental matrix and the matrices  $A$  and  $C$  are decomposed as in (5.6). If the multiplicity of the zero eigenvalue of  $A$  is less than or equal to 1, then there exists a positive linear observer of the form (5.1) if and only if  $(A, C)$  is detectable.*

Note that the decoupling condition (block diagonal structure of  $C_+$ ) also plays an important role in Corollary 3.5. Note also that if  $(A, C)$  is observable, each block diagonal component of  $G_{o+}(s)$  satisfies the conditions of Theorem 5.7 ( $A$ ,  $\overline{K}$ , and  $C$  replaced by  $A_{ii}$ ,  $v_i$ , and  $C_{i-\nu+m}$ , respectively, for  $i = \nu - m + 1, \dots, \nu$ ), which implies that one can design a state observer with rank  $H = n$  by changing  $K$  slightly.



*Example 5.12.* Consider a detectable positive linear system described by

$$\dot{x} = Ax = \begin{bmatrix} -2 & 2 & 0 \\ 0 & -2 & 2 \\ 2 & 0 & -2 \end{bmatrix} x, \quad y = Cx = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} x.$$

Simple computation yields that  $\sigma(A) = \{0, -3 \pm j\sqrt{3}\}$  and 0 is the only observable mode. Note that there is no  $K \in \mathbb{R}_+^3$  such that  $A - KC$  is Hurwitz and Metzler (this can be checked by Theorem 3.2).

Now, we apply Theorem 5.10. First note that  $A_+ = A$  and  $C_+ = C$  (we have no  $A_-$  and  $C_-$ ). If we choose  $K_+ = \epsilon \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}^t$ ,  $\epsilon > 0$ , then

$$G_o(s) = G_{o+}(s) = (sI - A_+ + K_+C_+)^{-1}K_+ = \frac{\epsilon}{3(s + \epsilon)} 1_{3 \times 1}.$$

Thus, the positive linear system

$$\begin{aligned} \dot{\hat{z}} &= -\epsilon \hat{z} + \epsilon y, \\ \hat{x} &= \frac{1}{3} 1_{3 \times 1} \hat{z} \end{aligned}$$

is a positive linear observer for the system. Note that  $\epsilon$  can be chosen arbitrarily positive.

To obtain a positive observer whose output matrix generates a solid cone, we select  $K_+ = \begin{bmatrix} \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \end{bmatrix}^t$  (thus,  $(A_+, K_+)$  is controllable), which results in

$$\begin{aligned} G_{o+}(s) &= (sI - A_+ + K_+C_+)^{-1}K_+ \\ &= \begin{bmatrix} s^2 + 8s + 20 \\ 2s^2 + 12s + 20 \\ 2s^2 + 10s + 20 \end{bmatrix} \frac{1}{5(s^2 + 6s + 12)(s + 1)}. \end{aligned}$$

We follow the procedure described in [5] to obtain a positive realization  $(F, G, H)$ . Let  $G_{o+,i}(s)$  be the  $i$ th component of  $G_{o+}(s)$  and let  $(A_{c,i}, B_{c,i}, C_{c,i})$  be the controllable realization of  $G_{o+,i}(s)$ , i.e.,

$$\begin{aligned} A_{c,1} = A_{c,2} = A_{c,3} = A_c &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -12 & -18 & -7 \end{bmatrix}, \\ B_{c,1} = B_{c,2} = B_{c,3} = B_c &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^t, \\ C_{c,1} = \frac{1}{5} \begin{bmatrix} 20 & 8 & 1 \end{bmatrix}, C_{c,2} = \frac{2}{5} \begin{bmatrix} 10 & 6 & 1 \end{bmatrix}, C_{c,3} = \frac{2}{5} \begin{bmatrix} 10 & 5 & 1 \end{bmatrix}. \end{aligned}$$

Let  $\mathcal{P} = \text{Cone } P$ , where

$$P = \begin{bmatrix} B_c & \bar{A}_c B_c & \bar{A}_c^2 B_c & \bar{A}_c^3 B_c \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 1 & -1 & -5 \\ 1 & -4 & -2 & 14 \end{bmatrix}, \quad \bar{A}_c := A_c + 3I.$$

Note that the shift term  $3I$  in  $\overline{A}_c$  corresponds to the real part of the leftmost pole (the pole with the least real part) of  $G_{o+}(s)$ . It is easy to see that Cone  $P$  is a proper invariant cone with respect to  $A_c$ . Solving

$$\overline{A}_c P = P \overline{F}, \quad B_c = P G, \quad H = \begin{bmatrix} C_{c,1} \\ C_{c,2} \\ C_{c,3} \end{bmatrix} P, \quad \overline{F} \geq 0, G \geq 0, H \geq 0,$$

one gets

$$F = \overline{F} - 3I = \begin{bmatrix} -3 & 0 & 0 & 12 \\ 1 & -3 & 0 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 1 & -3 \end{bmatrix}, \quad G = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad H = \frac{1}{5} \begin{bmatrix} 1 & 4 & 10 & 14 \\ 2 & 4 & 4 & 8 \\ 2 & 2 & 6 & 18 \end{bmatrix}.$$

Note that  $\sigma(F) = \{-1, -3 \pm j\sqrt{3}, -5\} \supset \sigma(A - KC)$  and  $\text{rank } H = 3$ . Therefore, the dynamics  $\dot{\hat{z}} = F\hat{z} + Gy, \hat{x} = H\hat{z}$  is the positive linear observer with  $H$  being solid.

*Example 5.13.* Consider the system

$$\dot{x} = Ax = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{bmatrix} x, \quad y = Cx = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} x.$$

Recall that (Example 3.6) it is not possible to design a Luenberger-type positive observer. Corollary 5.11 guarantees that this system admits a positive linear observer of the form (5.7).

**6. Conclusion.** The problem of designing positive linear observers for positive linear systems has been addressed considering the use of coordinates transformations and of observers of higher dimension than the system to be observed. In ascending order of complexity, existence conditions for Luenberger-type positive linear observers, for linear observers using coordinates transformations, and for positive realization-based linear observers are derived. As expected, the class of systems which can be *observed* by a positive linear observer is enlarged in these new frameworks. Although the positive realization approach is the most complex design method, it is a well organized and constructive method, since there are several well-established tools to obtain positive realizations. Future research topics include extension of the work to systems which may have unstable complex eigenvalues, stabilization for positive linear systems, which has some duality to the observation problem, and positive observers for nonlinear positive systems.

**Acknowledgment.** The authors would like to thank the anonymous reviewers for their pertinent comments and suggestions which substantially improved the quality of the paper.

# REFERENCES

- [1] D. H. ANDERSON, *Compartmental Modeling and Tracer Kinetics*, Lecture Notes in Biomath. 50, Springer-Verlag, Berlin, 1983.
- [2] B. D. O. ANDERSON, M. DEISTLER, L. FARINA, AND L. BENVENUTI, *Nonnegative realization of a linear system with nonnegative impulse response*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 43 (1996), pp. 134–142.

- [3] A. ASTOLFI AND P. COLANERI, *A note on the existence of positive realizations*, Linear Algebra Appl., 390 (2004), pp. 329–343.
- [4] L. BENVENUTI AND L. FARINA, *An example of how positivity may force realizations of “large” dimension*, Systems Control Lett., 36 (1999), pp. 261–266.
- [5] L. BENVENUTI AND L. FARINA, *A tutorial on the positive realization problem*, IEEE Trans. Automat. Control, 49 (2004), pp. 651–664.
- [6] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.
- [7] A. BRAUER, *Limits for the characteristic roots of a matrix IV: Applications to stochastic matrices*, Duke Math. J., 19 (1952), pp. 75–91.
- [8] C. T. CHEN, *Linear System Theory and Design*, Holt, Rinehart and Winston, New York, 1984.
- [9] N. DAUTREBANDE AND G. BASTIN, *Positive linear observers for positive linear systems*, in Proceedings of the European Control Conference, Karlsruhe, Germany, 1999, CD-ROM, Session CP-9, paper F371.
- [10] P. D. EGGLESTON, T. D. LENKER, AND S. K. NARAYAN, *The nonnegative inverse eigenvalue problem*, Linear Algebra Appl., 379 (2004), pp. 475–490.
- [11] W. R. EVANS, *Control system synthesis by the root locus method*, Trans. AIEE, 69 (1950), pp. 67–69.
- [12] L. FARINA, *On the existence of a positive realization*, Systems Control Lett., 28 (1996), pp. 219–226.
- [13] L. FARINA AND S. RINALDI, *Positive Linear Systems: Theory and Applications*, Wiley, New York, 2000.
- [14] F.-H. FÖRSTER AND B. NAGY, *Nonnegative realizations of matrix transfer functions*, Linear Algebra Appl., 311 (2000), pp. 107–129.
- [15] G. F. FRANKLIN, J. D. POWELL, AND A. EMAMI-NAEINI, *Feedback Control of Dynamic Systems*, 5th ed., Prentice-Hall, Upper Saddle River, NJ, 2005.
- [16] J. A. JACQUEZ AND C. P. SIMON, *Qualitative theory of compartmental systems*, SIAM Rev., 35 (1993), pp. 43–79.
- [17] P. LANCASTER, *Theory of Matrices*, Academic Press, New York, 1969.
- [18] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices with Applications*, Academic Press, New York, 1985.
- [19] P. D. LEENHEER AND D. AEYELS, *Stabilization of positive linear systems*, Systems Control Lett., 44 (2001), pp. 259–271.
- [20] P. D. LEENHEER AND D. AEYELS, *Stability properties of equilibria of classes of cooperative systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 1996–2001.
- [21] R. LOEWY AND D. LONDON, *A note on an inverse problem for nonnegative matrices*, Linear Multilinear Algebra, 6 (1978), pp. 83–90.
- [22] D. G. LUENBERGER, *An introduction to observers*, IEEE Trans. Automat. Control, 16 (1966), pp. 596–602.
- [23] D. G. LUENBERGER, *Introduction to Dynamic Systems: Theory, Models and Applications*, Wiley, New York, 1979.
- [24] Y. OHTA, H. MAEDA, AND S. KODAMA, *Reachability, observability, and realizability of continuous-time positive systems*, SIAM J. Control Optim., 22 (1984), pp. 171–180.
- [25] C. PICCARDI AND S. RINALDI, *Remarks on excitability, stability and sign of equilibria in cooperative systems*, Systems Control Lett., 46 (2002), pp. 153–163.
- [26] U. G. ROTHBLUM AND C. P. TAN, *Upper bounds on the maximum modulus of subdominant eigenvalues of nonnegative matrices*, Linear Algebra Appl., 66 (1985), pp. 45–86.
- [27] B. SHAFAI AND C. V. HOLLOT, *Robust nonnegative stabilization of interval discrete systems*, in Proceedings of the Conference on Decision and Control, Brighton, UK, 1991, pp. 49–51.
- [28] D. D. ŠILJAK, *Large Scale Dynamic Systems: Stability and Structure*, North-Holland, New York, 1978.
- [29] H. L. SMITH, *Monotone Dynamical Systems—An Introduction to the Theory of Competitive and Cooperative Systems*, AMS, Providence, RI, 1995.
- [30] R. SOTO, A. BOROBIA, AND J. MORO, *On the comparison of some realizability criteria for the real nonnegative inverse eigenvalue problem*, Linear Algebra Appl., 396 (2005), pp. 223–241.
- [31] R. T. STEFANI, B. SHAHIAN, C. J. SAVANT, AND G. H. HOSTETTER, *Design of Feedback Control Systems*, 4th ed., Oxford University Press, New York, 2002.
- [32] J. M. VAN DEN HOF, *Positive linear observers for linear compartmental systems*, SIAM J. Control Optim., 36 (1998), pp. 590–608.

# NONZERO SUM STOCHASTIC DIFFERENTIAL GAMES WITH DISCOUNTED PAYOFF CRITERION: AN APPROXIMATING MARKOV CHAIN APPROACH\*

K. SURESH KUMAR†

**Abstract.** We develop a new constructive method for proving the existence of Nash equilibrium for a class of nonzero sum stochastic differential games. Under certain usual assumptions, we prove the existence of Nash equilibrium for discounted payoff criteria. A novel feature of our method is that it allows us to compute Nash equilibrium for a large class of stochastic differential games.

**Key words.** controlled diffusions, approximating Markov chain, discounted payoff criteria, Nash equilibrium

**AMS subject classifications.** 91A15, 91A10

**DOI.** 10.1137/060650623

**1. Introduction.** We develop a new method for finding Nash equilibria for a class of stochastic differential games where state process  $X(\cdot)$  of the game is given by the solution of the controlled SDE

$$(1.1) \quad \left. \begin{aligned} dX(t) &= [b_1(X(t), v_1(t)) + b_2(X(t), v_2(t))] dt + \sigma(X(t)) dW(t), \\ X(0) &= x \in \mathbb{R}^d, \end{aligned} \right\}$$

where  $v_i$  is the strategy of player  $i, i = 1, 2$ . The payoff criterion of the game for the player  $i = 1, 2$  is given by

$$(1.2) \quad R_{\alpha}^i[v_1, v_2](x) = E \left[ \int_0^{\infty} e^{-\alpha t} [r_{1i}(X(t), v_1(t)) + r_{2i}(X(t), v_2(t))] dt \middle| X(0) = x \right],$$

$i = 1, 2$ . In a two-player game, a Nash equilibrium is a pair of strategies such that unilateral deviation from this pair of strategies by any player is disadvantageous to him. If the first player announces his strategy in advance, then the second player would maximize his payoff. Any strategy of the second player that maximizes his payoff is called his optimal response corresponding to the announced strategy of the first player. Note that there may be several optimal responses of the second player corresponding to each announced strategy of the first player. Similarly, optimal response of the first player is defined. This gives a map which takes each pair of strategies of the players to a corresponding pair of a set of optimal responses. It can be seen that any fixed point of this multivalued map is a Nash equilibrium. Note that this multivalued map may have several fixed points. Thus we may have multiple Nash equilibria. The traditional method of establishing the existence of a Nash equilibrium is to prove the existence of a fixed point of this multivalued map by using a fixed point theorem [8]. The main step in this method lies in establishing the upper semicontinuity of this multivalued map under a suitable metrizable topology on the appropriate set of

\*Received by the editors January 23, 2006; accepted for publication (in revised form) August 31, 2007; published electronically January 30, 2008.

<http://www.siam.org/journals/sicon/47-1/65062.html>

†Department of Mathematics, Indian Institute of Technology, Bombay, Powai, Mumbai 400 076, India (suresh@math.iitb.ac.in).

strategies. Though this methodology is very useful in establishing Nash equilibria, it does not lead to a constructive procedure for computing Nash equilibria.

In this paper we develop a new method based on a discretization procedure which enables us to construct a Nash equilibrium. By suitably discretizing the original stochastic differential game, we construct a parametric family of stochastic games in such a way that the state of the stochastic game is given by a controlled discrete time process  $\{X_n^h | n \geq 0\}$ , defined in section 3, satisfying a certain “local consistency” condition, and the payoff criterion is given by

$$R_\alpha^{ih}[v_1^h, v_2^h](x) = \Delta t^h E \left[ \sum_{n=0}^{\infty} \left( e^{-\alpha \Delta t^h} \right)^n [r_{1i}(X_n^h, v_{1n}^h) + r_{2i}(X_n^h, v_{2n}^h)] \middle| X_0^h = x \right],$$

where  $\Delta t^h$  is the constant step size for time discretization; its explicit form is given in section 3. The local consistency condition enables us to show that, as the parameter  $h$  tends to 0, a Nash equilibrium of the stochastic game converges in an appropriate sense to a Nash equilibrium of the stochastic differential game.

Our convergence analysis is crucially based on the specific structure of the drift vector in (1.1) and the payoff functions in (1.2). Note that the dependence of the variables  $v_1, v_2$  on the drift vector of the process  $X(\cdot)$  as in (1.1) appears in separate functions in an additive manner. The same holds for the payoff functions in (1.2). These separability conditions are typical in stochastic dynamic games in discrete time with general (uncountable) state space as well. In stochastic games in discrete time with uncountable state space, these types of conditions are referred to as AR-AT, which stands for additive reward-additive transition structure; see, for example, [11], [22]. To our knowledge, the problem of establishing the existence of a Nash equilibrium in stationary strategies without the AR-AT structure for discrete time stochastic game in a general state space is still open. The same holds for stochastic differential games as well. In the discrete time case, some progress has been made in this direction. In [25] the authors have established the existence of a Nash equilibrium in stationary strategies without AR-AT structure but under the assumption that the transition is state-independent. Without using the AR-AT structure, in [21] the existence of a subgame perfect equilibrium is established, whereas in [23] the existence of a correlated equilibrium in stationary strategies is established.

The basic idea of approximating SDE (1.1) by using the chain  $\{X_n^h | n \geq 1\}$  was pioneered by Kushner; see [14], [15], [19]. These works deal with numerical procedures for stochastic optimal control problems. The idea has been extended to zero sum stochastic differential games as well; see [16], [17], [26]. A numerical approximation of nonzero sum stochastic differential games is treated in [18]. Our present paper is different from [18] in two aspects. First of all, we consider a different class of stochastic differential games. Second, our main focus lies in using the approximation argument to get a constructive procedure for computing a Nash equilibrium. This point is not addressed in [18].

In the literature of nonzero sum stochastic differential games, not much is known about the computational aspects of Nash equilibrium. Nash equilibria can be explicitly computed only in some specific games; see, for example, [12], [24]. To the best of our knowledge, [20] is the latest in the literature which tries to construct Nash equilibrium for a class of stochastic differential games on the finite horizon. In [20], it is assumed that

$$b_k(x, v_k) = \tilde{b}_k(x)v_k, \quad r_1(x, v_1, v_2) = \tilde{r}_1(x)v_1, \quad r_2(x, v_1, v_2) = \tilde{r}_2(x)v_2,$$

where  $\tilde{b}_k, \tilde{r}_k$  are of  $C^1$  class. Also assume that  $\sigma\sigma'$  and  $(\sigma\sigma')^{-1}$  are Lipschitz continuous. These affine structural conditions are quite crucial in the analysis in [20]. In our paper, we compute Nash equilibrium for a class of stochastic differential games where the conditions in [20] may be treated as a special case. Note that it is our Example 5.6 with  $r_{12} \equiv r_{21} \equiv 0$ .

The rest of this paper is organized as follows: In section 2, we give a detailed description of the stochastic differential game problem. In section 3, we give a description of the approximating chain  $\{X_n^h\}$  and prove two important results (Theorems 3.1 and 3.2), which establish that Markov chains  $\{X_n^h\}$  under Markov strategies are indeed approximations of controlled diffusion given by (1.1) for a suitable pair of admissible strategies of the players. In section 4, we describe a stochastic game with the state equation given by  $\{X_n^h\}$  and payoff criterion (4.1), which is “close” to the payoff criterion (1.2). By using the convergence results proved in section 3, we prove that a Nash equilibrium of the stochastic game converges to a Nash equilibrium of the stochastic differential game along a subsequence. This proves the existence of a Nash equilibrium for the stochastic differential game. In section 5, by using the optimality equation (5.2), which characterizes a Nash equilibrium of the approximate stochastic game described in section 4, we derive an iterative procedure for constructing a Nash equilibrium for the approximate stochastic game. Then by using an additional assumption (A4) we show that iterates indeed converge. The Nash equilibrium we construct by using our procedure is given by the particular selector map  $\bar{u}_x^k(\cdot)$  given in (A4). Hence if multiple selector maps exist, then multiple Nash equilibria can be computed. We also give examples when the selector map is unique. Section 5 deals with numerical results for constructing a Nash equilibrium by using the above iterative procedure. We conclude our paper in section 6 with a few remarks.

**2. Description of the problem.** Let  $U_i$  be a compact subset of  $\mathbb{R}^{n_i}, i = 1, 2$ , and  $V_i = \mathcal{P}(U_i), i = 1, 2$ , the space of probability measures with the topology of weak convergence.

Let  $\bar{b}_k : \mathbb{R}^d \times U_k \rightarrow \mathbb{R}^d$  and  $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}, k = 1, 2$ . Define  $b_k : \mathbb{R}^d \times V_k \rightarrow \mathbb{R}^d, k = 1, 2$ , as follows:

$$b_k(x, v_k) = \int_{U_k} \bar{b}_k(x, u) v_k(du), \quad x \in \mathbb{R}^d, v_k \in V_k.$$

We assume that the following applies.

(A1) (i) The functions  $\sigma, \bar{b}_k, k = 1, 2$ , are bounded, continuous, and Lipschitz continuous in the first uniformly with respect to the second variable; i.e., there exists a constant  $C > 0$  such that

$$\|\bar{b}_k(x, u) - \bar{b}_k(y, u)\| + \|\sigma(x, u) - \sigma(y, u)\| \leq C \|x - y\|$$

for all  $x, y \in \mathbb{R}^d$  and  $u \in U_k, k = 1, 2$ .

(ii) The function  $a := \sigma\sigma'$  is uniformly elliptic; i.e., there exists  $\delta > 0$  such that

$$za(x)z' \geq \delta \|z\|^2 \text{ for all } z, x \in \mathbb{R}^d.$$

Consider the controlled diffusion process  $X(\cdot)$  on  $\mathbb{R}^d$  given by SDE

$$(2.1) \quad \left. \begin{aligned} dX(t) &= [b_1(X(t), v_1(t)) + b_2(X(t), v_2(t))] dt + \sigma(X(t)) dW(t), \\ X(0) &= x \in \mathbb{R}^d, \end{aligned} \right\}$$

where  $W(\cdot)$  is a standard  $d$ -dimensional Wiener process defined on a complete probability space  $(\Omega, \mathcal{F}, P)$ .

A  $V_i$ -valued process  $v_i(\cdot)$  is said to be an admissible strategy for player  $i$  if  $v_i(t) = f_i(t, X(\cdot))$ , for some  $f_i : [0, \infty) \times C([0, \infty); \mathbb{R}^d) \rightarrow V_i$ , is a measurable map and, for each  $t$ ,  $f_i(t, X(\cdot))$  is measurable with respect to  $\sigma(X(s) | s \leq t)$ . The set of all admissible strategies of player  $i$  is denoted by  $A_i$ . By following [6], we have the following interpretation of the admissible strategies of the players. Note that an admissible strategy player  $i$  is a nonanticipative functional of the process  $X(\cdot)$  taking values in the set  $V_i$ . The idea behind this is that whatever extraneous randomization the players might want to incorporate into their controls is already subsumed in the fact that they are choosing  $V_i$ -valued processes rather than  $U_i$ -valued ones. One consequence of this is that the conditional law of  $X(\cdot)$  given  $X(0) = x$  is a.s. the law of a process  $X(x, \cdot)$  controlled by strategies  $f_i(\cdot, X(x, \cdot))$ , with  $X(x, 0) = x$ . Thus we may prescribe the strategies  $(v_1(\cdot), v_2(\cdot))$  for arbitrary initial data by prescribing the  $f_i$ 's. Therefore player 1 chooses the function  $f_1$ , whereas player 2 chooses  $f_2$ . Also note that these choices are made independently of each other. Hence the strict noncooperative nature of the game is maintained at all times.

Suppose that  $(v_1, v_2) \in A_1 \times A_2$  is such that there exist measurable maps  $\bar{v}_i : [0, \infty) \times \mathbb{R}^d \rightarrow V_i, i = 1, 2$ , satisfying  $v_i(t) = \bar{v}_i(t, X(t))$ , where  $X(\cdot)$  solves (2.1) corresponding to the above admissible strategies. Then  $v_i$  is said to be a Markov strategy for player  $i$ . By an abuse of notation we refer to the map  $\bar{v}_i$  itself as a Markov strategy. A Markov strategy  $\bar{v}_i$  is said to be stationary if the map  $\bar{v}_i$  doesn't have any explicit dependence on  $t$ . The set of all stationary Markov strategies for player  $i$  is denoted by  $M_i$ .

Let  $R(U_k \times [0, \infty))$ ,  $k = 1, 2$ , denote the set of all  $V_k$ -valued processes which is nonanticipative w.r.t. an  $\mathbb{R}^d$ -valued Wiener process. The space of strategies  $R(U_k \times [0, \infty))$  is endowed with the weak topology given in [19, pp. 265–267]; i.e.,  $v_k^n$  converges to  $v_k$  in  $R(U_k \times [0, \infty))$  if

$$\int_0^\infty \int_{U_k} \phi(t, u) v_k^n(t)(du) dt \rightarrow \int_0^\infty \int_{U_k} \phi(t, u) v_k(t)(du) dt$$

for all  $\phi \in C_0^\infty((0, \infty) \times U_k)$ . The set  $A_k \subseteq R(U_k \times [0, \infty))$  is endowed with the above weak topology induced on it.

Let  $\bar{r}_{ki} : \mathbb{R}^d \times U_k \rightarrow \mathbb{R}$ ,  $k, i = 1, 2$ , be the payoff functions satisfying the following assumption.

(A2) For each  $k, i$ ,  $\bar{r}_{ki}$  is bounded and continuous, and there exists a constant  $C > 0$  such that

$$|\bar{r}_{ki}(x, u_i) - \bar{r}_{ki}(y, u_i)| \leq C \|x - y\| \text{ for all } x, y \in \mathbb{R}^d, u_i \in U_i.$$

We consider only two-person nonzero sum stochastic differential games for notational simplicity; a general  $N$ -person game follows by mimicking the arguments of two-person games. The description of the two-person nonzero sum game is as follows.

If the state of the system, which is evolving according to the controlled diffusion process given by (2.1), is at  $x \in \mathbb{R}^d$  and the player  $i$  chooses his action  $u_i \in U_i$ , then the player  $i$  receives a payoff  $\bar{r}_{1i}(x, u_1) + \bar{r}_{2i}(x, u_2)$ . We use the relaxed setup, and the randomized payoff is given by

$$r_{ki}(x, v) = \int_{U_k} \bar{r}_{ki}(x, u) v(du) \text{ for } x \in \mathbb{R}^d, v \in V_k;$$

i.e., in the relaxed setup, when the state of the system is at  $x$  and the player  $i$  chooses his mixed action  $v_i \in V_i$ , player  $i$  receives a payoff  $r_{1i}(x, v_1) + r_{2i}(x, v_2)$ .

At time  $t$ , both players have the common knowledge of the state process  $\{X(s) | s \leq t\}$ . With this information players choose their action at  $t$  independent of each other. Each player wants to choose a strategy so that his “cumulative” payoff is maximized. The planning horizon is infinite, and for payoff evaluation criterion we use the discounted payoff given below.

**Discounted payoff criterion.** Let  $\alpha > 0$  be the discount factor. For  $(v_1, v_2) \in A_1 \times A_2$ , the  $\alpha$ -discounted payoff to player  $i$ , for the initial condition  $x \in \mathbb{R}^d$  is given by

$$R_\alpha^i[v_1, v_2](x) = E \left[ \int_0^\infty e^{-\alpha t} [r_{1i}(X(t), v_1(t)) + r_{2i}(X(t), v_2(t))] dt \mid X(0) = x \right],$$

$i = 1, 2$ .

A pair of strategies  $(v_1^*, v_2^*) \in A_1 \times A_2$  is said to be a Nash equilibrium for the initial condition  $x \in \mathbb{R}^d$  if

$$R_\alpha^1[v_1^*, v_2^*](x) \geq R_\alpha^1[v_1, v_2^*](x) \text{ for all } v_1 \in A_1$$

and

$$R_\alpha^2[v_1^*, v_2^*](x) \geq R_\alpha^2[v_1^*, v_2](x) \text{ for all } v_2 \in A_2.$$

**3. Approximating Markov chain.** In this section we derive an approximating Markov chain for the controlled diffusion process (2.1). See [16], [26] for similar approximations. We make the following assumption.

(A3) (i)

$$a_{ii}(x) - \sum_{j:j \neq i} |a_{ij}(x)| \geq 0 \text{ for all } i = 1, 2, \dots, d, x \in \mathbb{R}^d.$$

(ii) There exists  $\delta_0 > 0$  such that

$$\sum_{i=1}^d a_{ii}(x) - \frac{1}{2} \sum_{i \neq j} |a_{ij}(x)| \geq \delta_0 \text{ for all } x \in \mathbb{R}^d.$$

For  $h > 0$ , define the grid

$$\mathbb{R}_h^d = \left\{ x \in \mathbb{R}^d \mid x = \sum_{i=1}^d n_i h e_i, n_i \in \mathbb{Z}, \text{ for all } i \right\},$$

where  $\{e_1, e_2, \dots, e_d\}$  is the usual basis of  $\mathbb{R}^d$ .

For  $x \in \mathbb{R}_h^d$ ,  $v_1 \in V_1$ , and  $v_2 \in V_2$ , set

$$Q^h(x, v_1, v_2) = h \sum_{i=1}^d (|b_{1i}(x, v_1)| + |b_{2i}(x, v_2)|) + \sum_{i=1}^d a_{ii}(x) - \frac{1}{2} \sum_{i \neq j} |a_{ij}(x)|$$

and

$$\bar{Q}^h = \sup_{x \in \mathbb{R}_h^d, (v_1, v_2) \in V_1 \times V_2} Q^h(x, v_1, v_2).$$



Now define transition law  $p^h$  as follows, for  $x \in \mathbb{R}_h^d$ ,  $v_k \in V_k$ :

$$(3.1) \quad \left. \begin{aligned} p^h(x, x, v_1, v_2) &= \frac{1}{\bar{Q}^h} [\bar{Q}^h - Q^h(x, v_1, v_2)], \\ p^h(x, x \pm h e_i, v_1, v_2) &= \frac{1}{\bar{Q}^h} \left[ h (b_{1i}^+(x, v_1) + b_{2i}^+(x, v_2)) \right. \\ &\quad \left. + \frac{1}{2} a_{ii}(x) - \frac{1}{2} \sum_{j:j \neq i} |a_{ij}(x)| \right], \\ p^h(x, x \pm h e_i \pm h e_j, v_1, v_2) &= \frac{1}{4 \bar{Q}^h} a_{ij}^+(x), \\ p^h(x, x \pm h e_i \mp h e_j, v_1, v_2) &= \frac{1}{4 \bar{Q}^h} a_{ij}^-(x), \\ p^h(x, y, v_1, v_2) &= 0 \text{ otherwise,} \end{aligned} \right\}$$

where  $f^+ := \max\{f, 0\}$ ,  $f^- := -\min\{f, 0\}$ .

For a pair of strategies  $(v_1^h, v_2^h)$ , i.e.,  $v_i^h = \{v_{in}^h | n \geq 1\}$ ,  $v_{in}^h : \mathbb{R}_h^d \rightarrow V_i$ , the corresponding controlled chain  $\{X_n^h\}$  is defined such that

$$P\{X_{n+1}^h = y | X_m^h, v_{1m}^h, v_{2m}^h, m \leq n\} = p^h(X_n^h, y, v_{1n}^h(X_n^h), v_{2n}^h(X_n^h)).$$

By a straightforward calculation we can show that, for  $x \in \mathbb{R}_h^d$ ,  $v_1 \in V_1$ ,  $v_2 \in V_2$ ,

$$(3.2) \quad \left. \begin{aligned} E \left[ \Delta X_n^h \mid X_m^h, v_{1m}^h, v_{2m}^h, m \leq n-1, X_n^h = x, v_{1n}^h = v_1, v_{2n}^h = v_2 \right] \\ = [b_1(x, v_1) + b_2(x, v_2)] \Delta t^h, \\ E \left( \left[ \Delta X_n^h - E \left[ \Delta X_n^h \mid X_m^h, v_{1m}^h, v_{2m}^h, m \leq n-1, X_n^h = x, v_{1n}^h = v_1, v_{2n}^h = v_2 \right] \right] \right. \\ \left. \left[ \Delta X_n^h - E \left[ \Delta X_n^h \mid X_m^h, v_{1m}^h, v_{2m}^h, m \leq n-1, X_n^h = x, v_{1n}^h = v_1, v_{2n}^h = v_2 \right] \right]' \right. \\ \left. \mid X_m^h, v_{1m}^h, v_{2m}^h, m \leq n-1, X_n^h = x, v_{1n}^h = v_1, v_{2n}^h = v_2 \right) \\ = a(x) \Delta t^h, \end{aligned} \right\}$$

where

$$\Delta X_n^h = X_{n+1}^h - X_n^h, \quad \Delta t^h = \frac{h^2}{\bar{Q}^h}.$$

Equation (3.2) is called the local consistency conditions; see [19, p. 71] for more details.

Let  $A_i^h$  and  $M_i^h$  denote, respectively, the set of all admissible and the set of all stationary Markov strategies for player  $i$ ,  $i = 1, 2$ . See [7, p. 48] for a description of admissible strategies and stationary Markov strategies. Let  $\{X_n^h | n \geq 0\}$  be the controlled chain corresponding to  $(v_1^h, v_2^h) \in A_1^h \times A_2^h$  and initial condition  $x_0^h \in \mathbb{R}_h^d$ .

Now we define the corresponding interpolated continuous time process as follows:

$$X^h(t) = x_0^h + \sum_{n: t_{n+1}^h \leq t} \Delta X_n^h,$$

where  $t_n^h = n \Delta t^h$ ,  $n = 0, 1, \dots$

For  $v_i = \{v_{in} | n \geq 0\} \in A_i^h$ , the interpolated continuous time strategy for player  $i$  is given by

$$v_i(t) = v_{in} \text{ if } t \in [t_n^h, t_{n+1}^h).$$

We use  $A_i^h$  also to denote the set of all interpolated continuous time admissible strategies. Similarly the set of all interpolated continuous time stationary Markov strategies denoted by  $M_i^h$ ,  $i = 1, 2$ , is defined.

We make use of the following representation for the chain  $\{X_n^h | n \geq 0\}$  corresponding to  $(v_1^h, v_2^h) \in A_1^h \times A_2^h$ :

$$\begin{aligned} X_{n+1}^h &= X_n^h + E\left[\Delta X_n^h \mid X_m^h, v_{1m}^h, v_{2m}^h, m \leq n\right] \\ &\quad + \Delta X_n^h - E\left[\Delta X_n^h \mid X_m^h, v_{1m}^h, v_{2m}^h, m \leq n\right]. \end{aligned}$$

By using the local consistency conditions (3.2), we have

$$X_{n+1}^h = X_n^h + (b_1(X_n^h, v_{1n}^h) + b_2(X_n^h, v_{2n}^h))\Delta t^h + \beta_n^h,$$

where

$$\beta_n^h = \Delta X_n^h - E\left[\Delta X_n^h \mid X_m^h, v_{1m}^h, v_{2m}^h, m \leq n\right].$$

Therefore

$$X^h(t) = x_0^h + \sum_{n:t_{n+1}^h \leq t} (b_1(X_n^h, v_{1n}^h) + b_2(X_n^h, v_{2n}^h)) \Delta t^h + \sum_{n:t_{n+1}^h \leq t} \beta_n^h.$$

Since  $b_i$  is bounded, we can show that

$$\sum_{n:t_{n+1}^h \leq t} (b_1(X_n^h, v_{1n}^h) + b_2(X_n^h, v_{2n}^h)) \Delta t^h = F^h(t) + o(\Delta t^h),$$

where

$$F^h(t) = \int_0^t (b_1(X^h(s), v_1(s)) + b_2(X^h(s), v_2(s))) ds.$$

Define

$$W^h(t) = \sum_{n:t_{n+1}^h \leq t} \Delta W_n^h, \text{ where } \Delta W_n^h = (\sigma(X_n^h))^{-1} \beta_n^h,$$

We can see that  $W^h(\cdot)$  is a martingale with respect to  $\mathcal{F}_n^h = \sigma(X_m^h, v_{1m}^h, v_{2m}^h, m \leq n)$  and

$$\begin{aligned} E[\Delta W_n^h | X_m^h, v_{1m}^h, v_{2m}^h, m \leq n] &= 0, \\ E[(\Delta W_n^h)' \Delta W_n^h] | X_m^h, v_{1m}^h, v_{2m}^h, m \leq n &= \Delta t^h I_{d \times d}, \\ E[(\Delta W_n^h)' \Delta W_m^h] &= 0 \text{ for all } n \neq m, \end{aligned}$$

where  $I_{d \times d}$  is the  $d \times d$  identity matrix. Consider

$$\begin{aligned} \sum_{n: t_{n+1}^h \leq t} \beta_n^h &= \sum_{n: t_{n+1}^h \leq t} \sigma(X_n^h) \Delta W_n^h \\ &= \int_0^t \sigma(X^h(s)) dW^h(s) := B^h(t). \end{aligned}$$

Thus  $X^h(\cdot)$  takes the form

$$(3.3) \quad X^h(t) = x_0^h + B^h(t) + F^h(t).$$

Now we are ready to prove the following theorem.

**THEOREM 3.1.** *Assume (A1) and (A3). Let  $x_0^h \in \mathbb{R}_h^d$  converge to  $x \in \mathbb{R}^d$ ,  $(v_1^h, v_2^h) \in M_h^1 \times M_h^2$ , and  $X^h(\cdot)$  be the process (3.3) corresponding to  $(v_1^h, v_2^h)$  with initial condition  $x_0^h$ . Then the process  $\Phi^h(\cdot) := (X^h(\cdot), B^h(\cdot), F^h(\cdot), W^h(\cdot), v_1^h, v_2^h)$  is tight. Moreover, there exists a process  $\Phi(\cdot) = (X(\cdot), B(\cdot), F(\cdot), W(\cdot), v_1, v_2)$  in  $D^{4d}[0, \infty) \times R(U_1 \times [0, \infty)) \times R(U_2 \times [0, \infty))$  such that the process  $\Phi(\cdot)$  has continuous paths a.s. and  $\Phi^h(\cdot)$  converges weakly along a subsequence to  $\Phi(\cdot)$ .*

*Proof.* Let  $\delta > 0$ . It is easy to see that

$$(3.4) \quad E \left[ \sup_{s \leq t \leq s+\delta} |F^h(t) - F^h(s)| \right] \leq K \delta,$$

where  $K$  is a common bound for  $b_k, r_{ki}, a, i, k = 1, 2$ . Now

$$\begin{aligned} (3.5) \quad E \left[ \sup_{s \leq t \leq s+\delta} |B^h(t) - B^h(s)|^2 \right] &\leq 4E |B^h(s+\delta) - B^h(s)|^2 \\ &= 4E \int_s^{s+\delta} \text{trace}\{a(X^h(s))\} ds \leq 4K\delta, \end{aligned}$$

where  $\bar{v}_i^h(s) = v_i^h(X^h(s))$ ,  $i = 1, 2$ , and

$$\begin{aligned} (3.6) \quad E \left[ \sup_{s \leq t \leq s+\delta} |W^h(t) - W^h(s)|^2 \right] &\leq 4E |W^h(s+\delta) - W^h(s)|^2 \\ &= \sum_{n: s \leq t_n^h < t_{n+1}^h \leq s+\delta} \Delta t^h \leq \delta. \end{aligned}$$

Also, since  $v_i^h, i = 1, 2$ , is taking values in a compact set, the tightness of  $v_i^h$  follows. In view of (3.4), (3.5), and (3.6), from [3, Theorems 15.5 and 8.3] or [14, pp. 31–33], the tightness of  $\Phi^h(\cdot)$  in  $D^{4d}([0, \infty) \times R(U_1 \times [0, \infty)) \times R(U_2 \times [0, \infty))$  follows. The proof of the existence of the process  $\Phi(\cdot)$  follows from standard arguments.  $\square$

**THEOREM 3.2.** *Assume the hypothesis of Theorem 3.1, that the process  $W(\cdot)$  in Theorem 3.1 is a standard Wiener process. Also  $(v_1, v_2) \in A_1 \times A_2$  and*

$$\begin{aligned} B(t) &= \int_0^t \sigma(X(s), v_1(s), v_2(s)) dW(s), \\ F(t) &= \int_0^t (b_1(X(s), v_1(s)) + b_2(X(s), v_2(s))) ds. \end{aligned}$$

*Proof.* By using Skorohod's theorem, there exists a  $P$ -null set  $\mathcal{N}$  such that  $X^h(s, \omega) \rightarrow X(s, \omega)$  and  $W^h(s, \omega) \rightarrow W(s, \omega)$  for all  $s \geq 0$  and  $\omega \in \Omega/\mathcal{N}$ . Then, by following the arguments in the proof of [5, Lemmas 1.2 and 1.3, pp. 24–26], we can show that along the same sequence

$$(3.7) \quad \lim_{h \rightarrow 0} \int_0^t \int_{U_i} f(s, u_i, \omega) \bar{v}_i^h(s, \omega) (du_i) ds = \int_0^t \int_{U_i} f(s, u_i, \omega) v_i(s, \omega) (du_i) ds$$

for all  $f \in C([0, T] \times U_i)$ ,  $T > 0$ ,  $0 \leq t \leq T$ ,  $\omega \in \Omega/\mathcal{N}$ .

Set

$$\bar{F}(t) = \int_0^t [b_1(X(s), v_1(s)) + b_2(X(s), v_2(s))] ds.$$

Now for  $\omega \in \Omega/\mathcal{N}$

$$(3.8) \quad \left. \begin{aligned} & |F^h(t, \omega) - \bar{F}(t, \omega)| \\ & \leq \left| \int_0^t (b_1(X^h(s, \omega), \bar{v}_1^h(s, \omega)) + b_2(X^h(s, \omega), \bar{v}_2^h(s, \omega))) ds \right. \\ & \quad \left. - \int_0^t (b_1(X(s, \omega), \bar{v}_1^h(s, \omega)) + b_2(X(s, \omega), \bar{v}_2^h(s, \omega))) ds \right| \\ & \quad + \left| \int_0^t (b_1(X(s, \omega), \bar{v}_1^h(s, \omega)) + b_2(X(s, \omega), \bar{v}_2^h(s, \omega))) ds \right. \\ & \quad \left. - \int_0^t (b_1(X(s, \omega), v_1(s, \omega)) + b_2(X(s, \omega), v_2(s, \omega))) ds \right|. \end{aligned} \right\}$$

It follows from the Lipschitz continuity of  $b_i$ ,  $i = 1, 2$ , that the first term on the right-hand side of (3.8) converges to zero. The second term on the right-hand side of (3.8) converges to zero by using (3.7). We have, by using Skorohod's theorem and Theorem 3.1, that along the same sequence and same  $\mathcal{N}$  (without the loss of generality)

$$F^h(t, \omega) \rightarrow F(t, \omega), \quad \omega \in \Omega/\mathcal{N}, \quad t > 0.$$

Now by the uniqueness of the limit we have  $\bar{F}(t) = F(t)$ ,  $t > 0$  a.s.

Next we show that  $W(\cdot)$  is a standard Wiener process with respect to  $\mathcal{F}_t = \sigma(X(s), v_1(s), v_2(s) | s \leq t)$ . We have

$$\begin{aligned} E[W^h(t)' W^h(t)] &= E \left( \sum_{n: t_{n+1}^h \leq t} \Delta W_n^h \right)' \left( \sum_{n: t_{n+1}^h \leq t} \Delta W_n^h \right) \\ &= E \left[ \sum_{n, m: t_{n+1}^h, t_{m+1}^h \leq t} \Delta W_{n+1}^h' \Delta W_{m+1}^h \right] \\ &= \left[ \sum_{n: t_{n+1}^h \leq t} \Delta t^h \right] I_{d \times d} = t I_{d \times d} + o(\Delta t^h). \end{aligned}$$

By using similar arguments, we can show that  $E|W^h(t)|^4$  is uniformly bounded in  $h$ . Hence we can see that  $W^h(t)'W^h(t)$  is uniformly integrable. Now, by using the fact that  $W^h(\cdot)$  converges weakly to  $W(\cdot)$ , we have

$$E[W(t)'W(t)] = t I_{d \times d}.$$

Now, by using the arguments in [14, pp. 99–100], we can show that  $W(\cdot)$  is a martingale with respect to  $\mathcal{F}_t$ . Hence, it follows that  $W(\cdot)$  is a standard Wiener process. Consider for  $i = 1, 2$

$$\begin{aligned} & E \left| \int_0^t \sigma(X^h(s)) dW^h(s) - \int_0^t \sigma(X(s)) dW(s) \right|^2 \\ & \leq 8 \left\{ E \left| \int_0^t [\sigma(X^h(s)) - \sigma_\delta(X^h(s))] dW^h(s) \right|^2 \right. \\ & \quad + E \left| \int_0^t \sigma_\delta(X^h(s)) dW^h(s) - \int_0^t \sigma_\delta(X^h(s)) dW(s) \right|^2 \\ & \quad + E \left| \int_0^t [\sigma_\delta(X^h(s)) - \sigma_\delta(X(s))] dW(s) \right|^2 \\ & \quad \left. + E \left| \int_0^t [\sigma_\delta(X(s)) - \sigma(X(s))] dW(s) \right|^2 \right\} \\ & = 8 \left\{ E \sum_{k=0}^{\frac{t}{\delta}} \int_{k\delta}^{(k+1)\delta} \text{trace} [(\sigma(X^h(s)) - \sigma(X^h(k\delta))) \right. \\ & \quad \left. (\sigma(X^h(s)) - \sigma(X^h(k\delta)))'] ds \right. \\ & \quad + E \left| \int_0^t \sigma_\delta(X^h(s)) dW^h(s) - \int_0^t \sigma_\delta(X^h(s)) dW(s) \right|^2 \\ & \quad + E \int_0^t \text{trace} [(\sigma_\delta(X^h(s)) - \sigma_\delta(X(s))) \\ & \quad \left. (\sigma_\delta(X^h(s)) - \sigma_\delta(X(s)))'] ds, \right. \\ & \quad \left. + E \int_0^t \text{trace} [(\sigma_\delta(X(s)) - \sigma(X(s))) \right. \\ & \quad \left. (\sigma_\delta(X(s)) - \sigma(X(s)))'] ds \right\}, \end{aligned}$$

where  $\delta > 0$  is chosen such that  $\frac{t}{\delta}$  is an integer and

$$\int_0^t \sigma_\delta(X^h(s)) dW^h(s) := \sum_{k=0}^{\frac{t}{\delta}} \sigma(X^h(k\delta)) [W^h((k+1)\delta) - W^h(k\delta)].$$

By letting  $h \rightarrow 0$  and then  $\delta \rightarrow 0$  we can see that the right-hand side above converges to zero. Hence along a suitable subsequence (denoted by the same sequence without the loss of generality)

$$\int_0^t \sigma(X^h(s)) dW^h(s) \text{ converges a.s. to } \int_0^t \sigma(X(s)) dW(s) \text{ for all } t.$$

Hence by using the uniqueness of the limit we have

$$B(t) = \int_0^t \sigma(X(s)) dW(s).$$

Note that for any  $f_i \in C_b(\mathbb{R}^d \times V_i)$ ,  $i = 1, 2$ ,  $\int_0^t [f_1(X(s), v_1(s)) + f_2(X(s), v_2(s))] ds$  is nonanticipative with respect to  $\mathcal{F}_t$ . In view of this and by using Theorem 2.2(a) of [5, p. 18], we can assume without any loss of generality that  $(v_1, v_2) \in A_1 \times A_2$ . This completes the proof.  $\square$

**4. Discounted payoff criterion.** In this section we prove the existence of a Nash equilibrium for the stochastic differential game problem with the  $\alpha$ -discounted payoff criterion described in section 2. The main idea is to approximate the nonzero sum stochastic differential game problem by an appropriate nonzero sum stochastic game.

Consider the nonzero sum stochastic game with state space  $\mathbb{R}_h^d$  and  $p^h$  given by (3.1) describing the law of motion. The  $\alpha$ -discounted payoff to player  $i$  for the admissible pair of strategies  $(v_1^h, v_2^h) \in A_1^h \times A_2^h$  for the initial state  $x \in \mathbb{R}_h^d$  is given by (4.1)

$$R_\alpha^{ih}[v_1^h, v_2^h](x) = \Delta t^h E \left[ \sum_{n=0}^{\infty} \left( e^{-\alpha \Delta t^h} \right)^n [r_{1i}(X_n^h, v_{1n}^h) + r_{2i}(X_n^h, v_{2n}^h)] \middle| X_0^h = x \right].$$

The existence of a Nash equilibrium in the class of stationary Markov strategies follows from [9, Theorem 1]. Let  $(v_1^{*h}, v_2^{*h}) \in M_1^h \times M_2^h$  denote a pair of Nash equilibrium for the above approximated stochastic game.

**THEOREM 4.1.** *Assume (A1)–(A3). There exists an  $\alpha$ -discounted Nash equilibrium.*

*Proof.* Fix  $x \in \mathbb{R}^d$ . Let  $x_h \in \mathbb{R}_h^d$  be such that  $x_h \rightarrow x$ . Also let  $\Phi^{*h}(\cdot) = (X^{*h}(\cdot), B^{*h}(\cdot), F^{*h}(\cdot), v_1^{*h}, v_2^{*h})$  be the process given by (3.3), with  $(v_1^{*h}, v_2^{*h}) \in M_1^h \times M_2^h$  given above and for the initial condition  $x_h$ . Then by Theorem 3.2 there exist  $\Phi^*(\cdot) = (X^*(\cdot), B^*(\cdot), F^*(\cdot), v_1^*, v_2^*)$  and a Wiener process  $W(\cdot)$  such that  $X^*(\cdot)$  is the solution of (2.1) for some pair of admissible strategies  $(v_1^*, v_2^*) \in A_1 \times A_2$  and initial condition  $x$ . Also,  $\Phi^{*h}(\cdot)$  converges to  $\Phi^*(\cdot)$  weakly in  $D^{3d+n_1+n_2}[0, \infty)$ . Note that

$$\begin{aligned} & R_\alpha^{1h}[v_1^h, v_2^h](x) \\ &= E \left[ \int_0^\infty e^{-\alpha t} [r_{1i}(X^h(s), v_1^h(s)) + r_{2i}(X^h(s), v_2^h(s))] ds \middle| X^h(0) = x \right] + o(\Delta t^h). \end{aligned}$$

In view of the above, by closely mimicking the arguments of Theorem 3.2, we can show that

$$(4.2) \quad \lim_{h \rightarrow 0} R_\alpha^{1h}[v_1^{*h}, v_2^{*h}](x_h) = R_\alpha^1[v_1^*, v_2^*](x).$$

By using Theorem 2.2(b) of [5, p. 18], for each  $(\bar{v}_1, \bar{v}_2) \in A_1 \times A_2$  and the corresponding solution  $\bar{X}(\cdot)$  of (2.1) with initial condition  $x$ , there exists a process  $X(\cdot)$ , on a possibly augmented probability space, satisfying (2.1) with initial condition  $x$  corresponding to  $(v_1, v_2) \in A_1 \times A_2$  for the prescribed Wiener process  $W(\cdot)$  constructed in the previous theorem. Also  $(\bar{v}_1, \bar{v}_2)$ ,  $(v_1, v_2)$  have the same functional form, and

$$R_\alpha^i[\bar{v}_1, \bar{v}_2](x) = R_\alpha^i[v_1, v_2](x), \quad x \in \mathbb{R}^d, \quad i = 1, 2.$$

Hence we can assume without the loss of generality that  $(v_1, v_2) \in A_1 \times A_2$  is in a common probability space and, in particular,  $(v_1, v_2^*)$  is in a common probability space for all  $v_1 \in A_1$ .

For  $\varepsilon > 0$ , by using Theorem 1.2 of [19, p. 278], there exists a  $\delta > 0$  and  $v_1^\varepsilon \in A_1$  such that  $v_1^\varepsilon$  is constant on intervals  $[n\delta, n\delta + \delta)$ ,  $n = 0, 1, 2, \dots$ , and satisfies

$$(4.3) \quad |R_\alpha^1[v_1^\varepsilon, v_2^*](x) - R_\alpha^1[v_1, v_2^*](x)| < \varepsilon.$$

Now for  $h$  sufficiently small, i.e., for all  $h$  such that  $\Delta t^h < \delta$ , we have  $v_1^\varepsilon \in A_1^h$ . Let  $X^{*\varepsilon h}(\cdot)$  denote the process (3.3) corresponding to  $(v_1^\varepsilon, v_2^{*h})$  for the initial condition  $x_h$  and  $X^{*\varepsilon}(\cdot)$  denote the process (2.1) for the initial condition  $x$  and the pair of admissible strategies  $(v_1^\varepsilon, v_2^*)$ .

Then  $X^{*\varepsilon h}(\cdot)$  converges weakly to  $X^{*\varepsilon}(\cdot)$ . Now, by mimicking the arguments in the proof of Theorem 3.2, we have

$$(4.4) \quad \lim_{h \rightarrow 0} R_\alpha^{1h}[v_1^\varepsilon, v_2^{*h}](x_h) = R_\alpha^1[v_1^\varepsilon, v_2^*](x).$$

Since  $(v_1^{*h}, v_2^{*h})$  is a Nash equilibrium for the stochastic game described above, we have

$$R_\alpha^{1h}[v_1^{*h}, v_2^{*h}](x_h) \geq R_\alpha^{1h}[v_1^{\varepsilon h}, v_2^{*h}](x_h).$$

Let  $h \rightarrow 0$ , and by using (4.2) and (4.4) we have

$$R_\alpha^1[v_1^*, v_2^*](x) \geq R_\alpha^1[v_1^\varepsilon, v_2^*](x).$$

Now, by using (4.3), we get

$$R_\alpha^1[v_1^*, v_2^*](x) \geq R_\alpha^1[v_1, v_2^*](x) - \varepsilon.$$

Since  $\varepsilon > 0$  is arbitrary, we have

$$R_\alpha^1[v_1^*, v_2^*](x) \geq R_\alpha^1[v_1, v_2^*](x).$$

In a similar fashion, we can show that

$$R_\alpha^2[v_1^*, v_2^*](x) \geq R_\alpha^2[v_1^*, v_2](x) \text{ for } v_2 \in A_2.$$

Hence  $(v_1^*, v_2^*)$  is a Nash equilibrium.  $\square$

*Remark 4.2.* We can in fact obtain a Nash equilibrium in the class of stationary Markov strategies. This could be seen as follows: Recall that  $(v_1^{*h}, v_2^{*h}) \subseteq M_1 \times M_2$  and converges to  $(v_1^*, v_2^*)$  according to the convergence criterion in (3.7). Note that  $M_i$ ,  $i = 1, 2$ , is compact in the metric topology defined in [5, p. 30]. Hence there exists  $\tilde{v}_i \in M_i$  such that  $v_i^{*h} \rightarrow \tilde{v}_i$  in  $M_i$ ,  $i = 1, 2$ , along a suitable subsequence. Now, by using [15, Theorem 3.1, p. 151],  $X^h(\cdot)$  converges weakly to  $\tilde{X}(\cdot)$ , where  $\tilde{X}(\cdot)$  is a solution to (2.1) corresponding to  $(\tilde{v}_1, \tilde{v}_2)$ . Since  $X^h(\cdot)$  converges weakly to  $X^*(\cdot)$ , the existence of a Nash equilibrium which is Markov thus follows.

**5. Construction of Nash equilibrium.** In this section, we describe a numerical procedure to construct a Nash equilibrium for the nonzero sum stochastic differential game described in section 2. By using the separable form of the drift coefficient in (2.1), we can write the transition probabilities given in (3.1) in the form

$$(5.1) \quad p^h(x, y, v_1, v_2) = p_1^h(x, y, v_1) + p_2^h(x, y, v_2) \text{ for all } x, y \in \mathbb{R}_h^d, v_i \in V_i,$$

where  $p_1^h, p_2^h$  are continuous functions whose explicit form can be written down in terms of the drift and diffusion coefficients of the SDE (2.1). We omit the details, since we don't require the explicit form for our purpose. Consider the approximate nonzero sum stochastic game described in section 4. We avoid the term  $\Delta t^h$  from the payoff criterion (4.1) of the stochastic game, since it won't affect the analysis once  $h$  is fixed. Now, throughout this section, we fix  $h$ . We use the following lemma, and the proof follows from [4, Theorem 6].

LEMMA 5.1. *A pair of stationary Markov strategies  $(v_1^*, v_2^*) \in M_1^h \times M_2^h$  is a Nash equilibrium iff  $(v_1^*, v_2^*)$  satisfies the optimality equation*

$$(5.2) \quad \left. \begin{aligned} R_\alpha^{1h}[v_1^*, v_2^*](x) &= \sup_{v_1 \in V_1} \left[ r_{11}(x, v_1) + r_{21}(x, v_2^*(x)) \right. \\ &\quad \left. + \beta \sum_{y \in \mathbb{R}_h^d} p^h(x, y, v_1, v_2^*(x)) R_\alpha^{1h}[v_1^*, v_2^*](y) \right], \\ R_\alpha^{2h}[v_1^*, v_2^*](x) &= \sup_{v_2 \in V_2} \left[ r_{12}(x, v_1^*(x)) + r_{22}(x, v_2) \right. \\ &\quad \left. + \beta \sum_{y \in \mathbb{R}_h^d} p^h(x, y, v_1^*(x), v_2) R_\alpha^{2h}[v_1^*, v_2^*](y) \right], \end{aligned} \right\}$$

where  $\beta = e^{-\alpha \Delta t^h}$ .

By using (5.1), the optimality equation becomes

$$(5.3) \quad \left. \begin{aligned} R_\alpha^{1h}[v_1^*, v_2^*](x) &= \sup_{v_1 \in V_1} \left[ r_{11}(x, v_1) + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, v_1) R_\alpha^{1h}[v_1^*, v_2^*](y) \right] \\ &\quad + r_{21}(x, v_2^*(x)) + \beta \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, v_2^*(x)) R_\alpha^{1h}[v_1^*, v_2^*](y), \\ R_\alpha^{2h}[v_1^*, v_2^*](x) &= \sup_{v_2 \in V_2} \left[ r_{22}(x, v_2) + \beta \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, v_2) R_\alpha^{2h}[v_1^*, v_2^*](y) \right] \\ &\quad + r_{12}(x, v_1^*(x)) + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, v_1^*(x)) R_\alpha^{2h}[v_1^*, v_2^*](y). \end{aligned} \right\}$$



Let  $l^\infty(\mathbb{R}_h^d; \mathbb{R}) = \{\phi : \mathbb{R}_h^d \rightarrow \mathbb{R} \mid \phi \text{ is bounded}\}$  with sup norm. Now define  $\pi_k^* : l^\infty(\mathbb{R}_h^d; \mathbb{R}) \rightarrow M_k^h, k = 1, 2$  as follows: For  $\phi \in l^\infty(\mathbb{R}_h^d; \mathbb{R})$ ,  $\pi_k^*(\phi) \in M_k^h$  is such that, for all  $x \in \mathbb{R}_h^d$ ,

$$\begin{aligned} & r_{kk}(x, \pi_k^*(\phi)(x)) + \beta \sum_{y \in \mathbb{R}_h^d} p_k^h(x, y, \pi_k^*(\phi)(x)) \phi(y) \\ &= \sup_{v_k \in V_k} \left[ r_{kk}(x, v_k) + \beta \sum_{y \in \mathbb{R}_h^d} p_k^h(x, y, v_k) \phi(y) \right]. \end{aligned}$$

Now, by using standard stochastic optimal control arguments (see, for example, [2, pp. 225–232]), we can show that if  $(\Phi_1, \Phi_2)$  satisfies the equation

$$(5.4) \quad \left. \begin{aligned} \Phi_1(x) &= r_{11}(x, \pi_1^*(\Phi_1)(x)) + r_{21}(x, \pi_2^*(\Phi_2)(x)) \\ &\quad + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, \pi_1^*(\Phi_1)(x)) \Phi_1(y) + \beta \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\Phi_2)(x)) \Phi_1(y), \\ \Phi_2(x) &= r_{12}(x, \pi_1^*(\Phi_1)(x)) + r_{22}(x, \pi_2^*(\Phi_2)(x)) \\ &\quad + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, \pi_1^*(\Phi_1)(x)) \Phi_2(y) + \beta \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\Phi_2)(x)) \Phi_2(y), \end{aligned} \right\}$$

then  $\Phi_k(x) = R_\alpha^{kh}[\pi_1^*(\Phi_1), \pi_2^*(\Phi_2)]$ . This fact with the help of Lemma 5.1 will imply that, for any solution  $(\Phi_1, \Phi_2)$  of (5.4),  $(\pi_1^*(\Phi_1), \pi_2^*(\Phi_2))$  is a Nash equilibrium. Thus constructing a Nash equilibrium reduces to the construction of the solution to (5.4). To do this we first state the following lemma whose proof follows by closely mimicking the arguments in [9, Theorem 1].

LEMMA 5.2. For each  $x \in \mathbb{R}_h^d$ ,  $k = 1, 2$ , the map  $T_x^k : \mathbb{R}^N \rightarrow 2^{V_k}$  defined by

$$T_x^k(\Phi[x]) = \operatorname{argmax}_{v_k \in V_k} \left[ r_{kk}(x, v_k) + \beta \sum_{y \in \mathbb{R}_h^d} p_k^h(x, y, v_k) \Phi(y) \right],$$

where  $\Phi[x] = (\Phi(y) \mid y = x, x \bar{+} h e_i, x \bar{+} h e_i \bar{+} h e_j)$ ,  $N = \#\{y \mid y = x, x \bar{+} h e_i, x \bar{+} h e_i \bar{+} h e_j\}$ , is upper semicontinuous.

Now we give an iterative procedure to construct a Nash equilibrium. Define  $(\Phi_1^n, \Phi_2^n)$  as follows:  $\Phi_1^0 = 1$ ,  $\Phi_2^0 = 1$ , and for  $n \geq 0$

$$(5.5) \quad \left. \begin{aligned} \Phi_1^{n+1}(x) &= r_{11}(x, \pi_1^*(\Phi_1^n)(x)) + r_{21}(x, \pi_2^*(\Phi_2^n)(x)) \\ &\quad + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, \pi_1^*(\Phi_1^n)(x)) \Phi_1^n(y) + \beta \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\Phi_2^n)(x)) \Phi_1^n(y), \\ \Phi_2^{n+1}(x) &= r_{12}(x, \pi_1^*(\Phi_1^n)(x)) + r_{22}(x, \pi_2^*(\Phi_2^n)(x)) \\ &\quad + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, \pi_1^*(\Phi_1^n)(x)) \Phi_2^n(y) + \beta \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\Phi_2^n)(x)) \Phi_2^n(y). \end{aligned} \right\}$$

THEOREM 5.3. If the iterates  $(\Phi_1^n, \Phi_2^n)$  in (5.5) converge, say, to  $(\Phi_1, \Phi_2)$ , then  $\Phi_k(x) = R_\alpha^{kh}[v_1^*, v_2^*](x)$  for some Nash equilibrium  $(v_1^*, v_2^*)$ .

Proof. From the hypothesis, it follows that  $\Phi_k^n[x] \rightarrow \Phi_k[x]$ . Now since  $V_k$  is compact, along a subsequence, we have, for some  $v_k^* \in M_k^h$ ,

$$(5.6) \quad \Phi_k^n[x] \rightarrow \Phi_k[x], \quad \pi_k^*(\Phi_k^n)(x) \rightarrow v_k^*(x) \text{ for all } x \in \mathbb{R}_h^d.$$

Now, by using Lemma 5.2, we have  $v_k^*(x) \in T_x^k(\Phi_k[x])$  for all  $x \in \mathbb{R}_h^d$ ; i.e.,

$$(5.7) \quad \left. \begin{aligned} & \max_{v_k \in V_k} \left[ r_{kk}(x, v_k) + \beta \sum_{y \in \mathbb{R}_h^d} p_k^h(x, y, v_k) \Phi_k(y) \right] \\ & = \left[ r_{kk}(x, v_k^*(x)) + \beta \sum_{y \in \mathbb{R}_h^d} p_k^h(x, y, v_k^*(x)) \Phi_k(y) \right]. \end{aligned} \right\}$$

Also by using the continuity of  $r_{kl}$ ,  $p_k^h$ ,  $k, l = 1, 2$ , and (5.5) we can see that

$$(5.8) \quad \left. \begin{aligned} \Phi_1(x) &= r_{11}(x, v_1^*(x)) + r_{21}(x, v_2^*(x)) + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, v_1^*(x)) \Phi_1(y) \\ &\quad + \beta \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, v_2^*(x)) \Phi_1(y), \\ \Phi_2(x) &= r_{12}(x, v_1^*(x)) + r_{22}(x, v_2^*(x)) + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, v_1^*(x)) \Phi_2(y) \\ &\quad + \beta \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, v_2^*(x)) \Phi_2(y). \end{aligned} \right\}$$

Now from (5.7) and (5.8) it follows that  $(\Phi_1, \Phi_2), (v_1^*, v_2^*)$  satisfies the optimality equation. Hence  $(v_1^*, v_2^*)$  is a Nash equilibrium.  $\square$

Now we give a sufficient condition for the convergence of the iterates.

(A4) (i) The functions  $\bar{p}_k^h, \bar{r}_{12}, \bar{r}_{21}$  are Lipschitz continuous in the second argument, where  $\bar{p}_k^h(x, u_k) = p_k^h(x, \delta_{u_k})$  for  $u_k \in U_k$  and  $\delta_{u_k}$  denotes the dirac delta at  $u_k$ .

(ii) For each  $x \in \mathbb{R}_h^d$ ,  $k = 1, 2$ , there exists a map  $\bar{u}_x^k : l_M^\infty(\mathbb{R}_h^d) \rightarrow U_k$  such that

$$\bar{u}_x^k(\phi) \in \operatorname{argmax}_{u_k \in U_k} \left[ r_{kk}(x, u_k) + \beta \sum_{y \in \mathbb{R}_h^d} p_k^h(x, y, u_k) \phi(y) \right],$$

which is Lipschitz continuous and the Lipschitz constant is independent of  $x$ . Here  $l_M^\infty(\mathbb{R}_h^d) = \{\phi : \mathbb{R}_h^d \rightarrow \mathbb{R} \mid 0 \leq \phi(y) \leq M\}$  and  $M$  is an upper bound for  $\{R_\alpha^i[v_1, v_2](x) \mid v_i \in A_i, x \in \mathbb{R}_h^d, i = 1, 2\}$ .

*Remark 5.4.* If  $(v_1^*, v_2^*)$  with functional form given by  $(f_1, f_2)$  is a Nash equilibrium of the stochastic differential game given by the state equation

$$dY(t) = [cb(c^{-1}Y(t), v_1(t)) + cb_2(c^{-1}Y(t), v_2(t))]dt + c\sigma(c^{-1}Y(t))dW(t)$$

and payoff criteria

$$R_\alpha^i[v_1, v_2](y) = E \left[ \int_0^\infty ce^{-\alpha t} [r_{1i}(c^{-1}Y(t), v_1(t)) + r_{2i}(c^{-1}Y(t), v_2(t))] dt \mid Y(0) = y \right],$$

$i = 1, 2$ , where  $c > 0$  is a constant, then  $(\tilde{v}_1^*, \tilde{v}_2^*), \tilde{v}_i^*(t) = f_i(t, cX(\cdot))$  is a Nash equilibrium of the original stochastic differential game. Hence by a suitable choice of the constant  $c$ , the Lipschitz coefficient of  $\bar{r}_{12}, \bar{r}_{21}$  in (A4)(i) can be chosen to be any fixed positive constant.

*Example 5.5.* Take  $U_1 = [a, b]$ ,  $U_2 = [c, d]$ . Define  $\bar{b}(x, u_k) = \tilde{b}_k(x)u_k^2$ ,  $\bar{r}_{ki}(x, u_k) = \tilde{r}_{ki}(x)u_k^2$ ,  $k, i = 1, 2$ . The functions  $\tilde{b}_k$ ,  $\tilde{r}_{ki}$  are Lipschitz continuous functions. The function  $\sigma$  is also assumed to be bounded Lipschitz continuous, and  $\tilde{r}_{ki}$  is assumed to be nonnegative. Now, for  $h$  sufficiently small, we can verify that (A4) is satisfied. Moreover, the selector map  $\bar{u}_x^k$  is unique for each  $x$ .

*Example 5.6.* Take  $U_1 = U_2 = [0, 1]$ . Define  $\bar{b}_k(x, u_k) = \tilde{b}_k(x)u_k$ ,  $\bar{r}_{ki}(x, u_k) = \tilde{r}_{ki}(x)u_k$ ,  $k, i = 1, 2$ . The functions  $\tilde{b}_k$ ,  $\tilde{r}_{ki}$  are bounded Lipschitz continuous functions. Then clearly (A4) is satisfied for  $h$  sufficiently small with a unique selector map. This is the setup discussed in [20].

*Example 5.7.* Take  $U_i$ ,  $i = 1, 2$ , to be compact and convex subsets of  $\mathbb{R}^{n_i}$ . Define

$$\bar{b}_k(x, u_k) = b_k(x) f_k(u_k), \quad \bar{r}_{ki}(x, u_k) = \tilde{r}_{ki}(x) g_{ki}(u_k), \quad k, i = 1, 2.$$

For each  $k, i$ , assume that  $b_k, \tilde{r}_{ki}$  are bounded and Lipschitz continuous. Also assume that  $g_{kk}, f_k$  are differentiable,  $f'_k$  is Lipschitz continuous, and  $g'_{kk}$  satisfies

$$K |g'_{kk}(u_1) - g'_{kk}(u_2)| \geq \|u_1 - u_2\|, \quad u_1, u_2 \in U_k, \quad \text{for some } K > 0.$$

Now for  $\phi \in l_M^\infty(\mathbb{R}_h^d)$

$$\begin{aligned} & \max_{0 \leq u_1 \leq 1} \left[ \tilde{r}_{11}(x) g_{11}(u_1) + \beta \sum_{y \in \mathbb{R}_h^d} \bar{p}_1^h(x, y, u_1) \phi(y) \right] \\ &= A(x) + \max_{0 \leq u_1 \leq 1} \left[ \tilde{r}_{11}(x) g_{11}(u_1) + \frac{h\beta}{Q^h} \left\{ \sum_{i=1}^d b_{1i}^+(x) (\phi(x + he_i) - \phi(x)) \right. \right. \\ & \quad \left. \left. + \sum_{i=1}^d b_{1i}^-(x) (\phi(x - he_i) - \phi(x)) \right\} f_1(u_1) \right], \end{aligned}$$

where

$$\begin{aligned} A(x) &= \frac{\beta}{2} \phi(x) + \frac{\beta}{4Q^h} \sum_{i=1}^d a_{ii}(x) [\phi(x + he_i) - 2\phi(x) + \phi(x - he_i)] \\ &+ \frac{\beta}{4Q^h} \sum_{i=1}^d \left( \sum_{j:j \neq i}^d |a_{ij}(x)| \right) [\phi(x + he_i) - 2\phi(x) + \phi(x - he_i)] \\ &+ \frac{\beta}{4Q^h} \sum_{i,j=1}^d a_{ij}^+(x) \phi(x \pm he_i \pm he_j) + \frac{\beta}{4Q^h} \sum_{i,j=1}^d a_{ij}^-(x) \phi(x \pm he_i \mp he_j). \end{aligned}$$

Consider

$$\max_{0 \leq u_1 \leq 1} \left[ \tilde{r}_{11}(x) g_{11}(u_1) + h g(x, \phi, h) f_1(u_1) \right],$$

where

$$g(x, \phi, h) = \frac{\beta}{Q^h} \left\{ \sum_{i=1}^d b_{1i}^+(x) (\phi(x + he_i) - \phi(x)) + \sum_{i=1}^d b_{1i}^-(x) (\phi(x - he_i) - \phi(x)) \right\}.$$

Now by using the assumptions we can see that, for each fixed  $x$ , there exists  $u_1^*$ ,  $u_1^{*h}(\phi)$  (for  $h$  sufficiently small) such that

$$\begin{aligned}\tilde{r}_{11}(x) g'_{11}(u_1^*) &= 0, \\ \tilde{r}_{11}(x) g'_{11}(u_1^{*h}(\phi)) + h g(x, \phi, h) f'_1(u_1^{*h}(\phi)) &= 0 \text{ for all } h, \phi.\end{aligned}$$

Now

$$\begin{aligned}|u_1^{*h}(\phi_1) - u_1^{*h}(\phi_2)| &\leq K |g'_{11}(u_1^{*h}(\phi_1)) - g'_{11}(u_1^{*h}(\phi_2))| \\ &= K_1 |\tilde{r}_{11}(x) g'_{11}(u_1^{*h}(\phi_1)) - \tilde{r}_{11}(x) g'_{11}(u_1^{*h}(\phi_2))| \\ &= h K_1 |g(x, \phi_1, h) f'_1(u_1^{*h}(\phi_1)) - g(x, \phi_2, h) f'_1(u_1^{*h}(\phi_2))|,\end{aligned}$$

where  $K_1 = \frac{K}{|\tilde{r}_{11}(x)|}$ . Therefore, we have

$$\begin{aligned}|u_1^{*h}(\phi_1) - u_1^{*h}(\phi_2)| &\leq h K_1 \left[ |g(x, \phi_1, h) - g(x, \phi_2, h)| |f'_1(u_1^{*h}(\phi_1))| \right. \\ &\quad \left. + |g(x, \phi_2, h)| |f'_1(u_1^{*h}(\phi_1)) - f'_1(u_1^{*h}(\phi_2))| \right] \\ &\leq h K_2 \left[ |g(x, \phi_1, h) - g(x, \phi_2, h)| \right. \\ &\quad \left. + |f'_1(u_1^{*h}(\phi_1)) - f'_1(u_1^{*h}(\phi_2))| \right] \\ &\leq h K_2 |g(x, \phi_1, h) - g(x, \phi_2, h)| \\ &\quad + h K_3 |u_1^{*h}(\phi_1) - u_1^{*h}(\phi_2)|,\end{aligned}$$

where  $K_2, K_3$  are positive constants. Hence, we have

$$|u_1^{*h}(\phi_1) - u_1^{*h}(\phi_2)| \leq \left( \frac{h K_2}{1 - h K_3} \right) |g(x, \phi_1, h) - g(x, \phi_2, h)|.$$

Also a routine calculation shows that  $g$  is Lipschitz continuous and the Lipschitz constant is independent of  $x \in \mathbb{R}_h^d$ . Thus the assumption (A4) is satisfied for  $h$  sufficiently small.

By using (A4) we can see that, for each  $\phi \in l_M^\infty(\mathbb{R}_h^d)$ ,  $\pi_k^*(\phi)(x) = \delta_{\bar{u}_x^k(\phi)}$ . For each fixed  $x$ , define  $T(\phi_1, \phi_2) = (\psi_1(x), \psi_2(x))$ , where

$$\begin{aligned}\psi_1(x) &= \sup_{v_1 \in V_1} \left[ r_{11}(x, v_1) + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, v_1) \phi_1(y) \right] \\ &\quad + r_{21}(x, \pi_2^*(\phi_2)(x)) + \beta \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\phi_2)(x)) \phi_1(y),\end{aligned}$$

and  $\psi_2$  is defined analogously. For  $\phi_i, \tilde{\phi}_i \in l^\infty(\mathbb{R}_h^d; \mathbb{R})$ ,  $i = 1, 2$ , set

$$\begin{aligned}\|(\phi_1, \phi_2)(x) - (\tilde{\phi}_1, \tilde{\phi}_2)(x)\| &= |\phi_1(x) - \tilde{\phi}_1(x)| + |\phi_2(x) - \tilde{\phi}_2(x)|, \quad x \in \mathbb{R}_h^d, \\ \|(\phi_1, \phi_2) - (\tilde{\phi}_1, \tilde{\phi}_2)\|_{l^\infty} &= \|\phi_1 - \tilde{\phi}_1\|_{l^\infty} + \|\phi_2 - \tilde{\phi}_2\|_{l^\infty}.\end{aligned}$$

Then

$$\begin{aligned}
& \|T(\phi_1, \phi_2)(x) - T(\tilde{\phi}_1, \tilde{\phi}_2)(x)\| \\
& \leq \left| \sup_{v_1 \in V_1} \left[ r_{11}(x, v_1) + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, v_1) \phi_1(y) \right] \right. \\
& \quad \left. - \sup_{v_1 \in V_1} \left[ r_{11}(x, v_1) + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, v_1) \tilde{\phi}_1(y) \right] \right| \\
& \quad + \left| r_{21}(x, \pi_2^*(\phi_2)(x)) - r_{21}(x, \pi_2^*(\tilde{\phi}_2)(x)) \right| \\
& \quad + \beta \left| \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\phi_2)(x)) \phi_1(y) - \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\tilde{\phi}_2)(x)) \tilde{\phi}_1(y) \right| \\
& \quad + \left| \sup_{v_2 \in V_2} \left[ r_{22}(x, v_2) + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, v_2) \phi_2(y) \right] \right. \\
& \quad \left. - \sup_{v_2 \in V_2} \left[ r_{22}(x, v_2) + \beta \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, v_2) \tilde{\phi}_2(y) \right] \right| \\
& \quad + \left| r_{12}(x, \pi_1^*(\phi_1)(x)) - r_{12}(x, \pi_1^*(\tilde{\phi}_1)(x)) \right| \\
& \quad + \beta \left| \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, \pi_1^*(\phi_1)(x)) \phi_2(y) - \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, \pi_1^*(\tilde{\phi}_1)(x)) \tilde{\phi}_2(y) \right| \\
& \leq \beta \left| \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, \pi_1^*(\phi_1)(x)) \phi_1(y) - \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, \pi_1^*(\phi_1)(x)) \tilde{\phi}_1(y) \right| \\
& \quad + \left| r_{21}(x, \pi_2^*(\phi_2)(x)) - r_{21}(x, \pi_2^*(\tilde{\phi}_2)(x)) \right| \\
& \quad + \beta \left| \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\phi_2)(x)) \phi_1(y) - \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\tilde{\phi}_2)(x)) \tilde{\phi}_1(y) \right| \\
& \quad + \beta \left| \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\phi_2)(x)) \phi_2(y) - \sum_{y \in \mathbb{R}_h^d} p_2^h(x, y, \pi_2^*(\phi_2)(x)) \tilde{\phi}_2(y) \right| \\
& \quad + \left| r_{12}(x, \pi_1^*(\phi_1)(x)) - r_{12}(x, \pi_1^*(\tilde{\phi}_1)(x)) \right| \\
& \quad + \beta \left| \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, \pi_1^*(\phi_1)(x)) \phi_2(y) - \sum_{y \in \mathbb{R}_h^d} p_1^h(x, y, \pi_1^*(\tilde{\phi}_1)(x)) \tilde{\phi}_2(y) \right| \\
& \leq \beta \|(\phi_1, \phi_2) - (\tilde{\phi}_1, \tilde{\phi}_2)\|_{l^\infty} + K_1 \|\phi_1 - \tilde{\phi}_1\|_{l^\infty} + K_1 \|\phi_2 - \tilde{\phi}_2\|_{l^\infty} \\
& \quad + \beta \left| \sum_{y \in \mathbb{R}_h^d} [p_2^h(x, y, \pi_2^*(\phi_2)(x)) - p_2^h(x, y, \pi_2^*(\tilde{\phi}_2)(x))] \tilde{\phi}_2(y) \right| \\
& \quad + \beta \left| \sum_{y \in \mathbb{R}_h^d} [p_1^h(x, y, \pi_1^*(\phi_1)(x)) - p_1^h(x, y, \pi_1^*(\tilde{\phi}_1)(x))] \tilde{\phi}_1(y) \right|
\end{aligned}$$

for some constant  $K_1 > 0$ . Therefore

$$\begin{aligned} & \|T(\phi_1, \phi_2)(x) - T(\tilde{\phi}_1, \tilde{\phi}_2)(x)\| \\ & \leq \beta \|(\phi_1, \phi_2) - (\tilde{\phi}_1, \tilde{\phi}_2)\|_{l^\infty} + K_2 \|\phi_1 - \tilde{\phi}_1\|_{l^\infty} + K_2 \|\phi_2 - \tilde{\phi}_2\|_{l^\infty}, \end{aligned}$$

where  $K_2 > 0$  is a constant. Hence we have

$$(5.9) \quad \|T(\phi_1, \phi_2) - T(\tilde{\phi}_1, \tilde{\phi}_2)\|_{l^\infty} \leq K_3 \|(\phi_1, \phi_2) - (\tilde{\phi}_1, \tilde{\phi}_2)\|_{l^\infty}.$$

By using Remark 4.2 we can choose  $K_3 < 1$ . Hence from (5.9) it follows that the iterates in (5.5) converge.

**Numerical results.** Consider the stochastic differential game with the following specifications:

$$\begin{aligned} \bar{b}_1(x, u_1) &= b_1(x)u_1^2, \quad \bar{b}_2(x, u_1) = b_2(x)u_2^2, \\ \bar{r}_1(x, u_1, u_2) &= \tilde{r}_{11}(x) \left( \frac{u_1^2}{2} - u_1 \right) + \tilde{r}_{21}(x)u_2, \\ \bar{r}_2(x, u_1, u_2) &= \tilde{r}_{12}(x)u_1 + \tilde{r}_{22}(x) \left( \frac{u_2^2}{2} - u_2 \right), \\ \sigma(x) &= \sigma, \quad x \in \mathbb{R}, u_1, u_2 \in [0, 1]. \end{aligned}$$

A simple calculation shows that

$$\overline{Q}^h = \frac{2h}{\sqrt{e}} + \sigma^2, \quad \Delta t^h = \frac{h^2 \sqrt{e}}{2h + \sigma^2 \sqrt{e}}.$$

With the condition  $\tilde{r}_{kk} \leq 0$ ,  $k = 1, 2$ , we can see that

$$\bar{u}_x^k(\phi) = \begin{cases} \frac{\tilde{r}_{kk}(x)}{\tilde{r}_{kk}(x) + \frac{2h\beta}{\overline{Q}^h} b_k(x)(\phi(x+h) - \phi(x))} & \text{if } \phi(x+h) < \phi(x), \\ 1 & \text{if } \phi(x+h) \geq \phi(x). \end{cases}$$

Now one can see that the assumption (A4) is satisfied. We have run the program for calculating Nash equilibrium and the iterates for the values of  $x$  in  $[-10, 10]$ . Since the input functions  $\tilde{r}_{kj}$  vanish at infinity, we are able to put an artificial boundary for the state, and values of the iterates are set to zero beyond this boundary. By using arguments similar to [1] one can see the following.

Let  $\tau_M = \inf\{t > 0 \mid \|X^h(t)\| \geq M\}$  and  $\epsilon(M)$  denote the error bound  $\|\phi^* - \phi^{M*}\|_{l^\infty}$ , with

$$\phi^* = (R_\alpha^{1h}[v_1^*, v_2^*], R_\alpha^{2h}[v_1^*, v_2^*]), \quad \phi^{M*} = (R_\alpha^{1Mh}[v_1^*, v_2^*], R_\alpha^{2Mh}[v_1^*, v_2^*]),$$

where  $(v_1^*, v_2^*)$  is a Nash equilibrium for the stochastic game ( $h$ -step) and

$$R_\alpha^{iMh}[v_1^*, v_2^*] = E \int_0^\infty e^{-\alpha t} [r_{1i}(X^h(t)), v_1^*(X^h(t)) + r_{2i}(X^h(t)), v_2^*(X^h(t))] I\{\tau_M \geq t\} dt.$$

Then

$$\begin{aligned} (5.10) \quad \epsilon(M) &= E \int_{\tau_M}^\infty e^{-\alpha t} [r_{1i}(X^h(t)), v_1^*(X^h(t)) + r_{2i}(X^h(t)), v_2^*(X^h(t))] dt \\ &= E \int_0^\infty e^{-\alpha t} [r_{1i}(X^h(t)), v_1^*(X^h(t)) + r_{2i}(X^h(t)), v_2^*(X^h(t))] I\{t \geq \tau_M\} dt \\ &\leq K \int_0^\infty e^{-\alpha t} P\{\tau_M \leq t\} dt, \end{aligned}$$

where  $K$  is a bound for the coefficients  $r_1, r_2$ . Consider the process (3.3) corresponding to  $(v_1^*, v_2^*)$  given by

$$X^h(t) = x_0^h + \int_0^t [b_1(X^h(s)), v_1^*(s) + b_2(X^h(s)), v_2^*(s)] ds + \int_0^t \sigma(X^h(s)) dW^h(s). \quad (5.11)$$

$$\begin{aligned} E \sup_{0 \leq t \leq T} \|X^h(t)\|^2 &\leq 5(x_0^h)^2 + 5E \sup_{0 \leq t \leq T} \left\| \int_0^t \sigma(X^h(s)) dW^h(s) \right\|^2 \\ &\quad + 5E \sup_{0 \leq t \leq T} \left\| \int_0^t [b_1(X^h(s)), v_1^*(s) + b_2(X^h(s)), v_2^*(s)] ds \right\|^2 \\ &\leq 5(x_0^h)^2 + 5K_2 \int_0^T E \|\sigma(X^h(t))\|^2 dt + 5K_1 T^2 \\ &\leq 5(x_0^h)^2 + 5K_1 T^2 + 5K_1 K_2 T, \end{aligned}$$

where  $K_1$  is the square of the common bound for  $b_1, b_2$  and  $\sigma$  and  $K_2$  is the constant in the Burkholder–Davis–Gundy inequality. Now

$$\begin{aligned} P\{\tau_M \leq t\} &\leq P\left(\sup_{0 \leq s \leq t} \|X^h(s)\| \geq M\right) \\ &\leq \frac{E \sup_{0 \leq s \leq t} \|X^h(s)\|^2}{M} \\ &\leq \frac{5(x_0^h)^2 + 5K_1 t^2 + 5K_1 K_2 t}{M}. \end{aligned} \quad (5.12)$$

Now from (5.10)–(5.12) we have

$$\begin{aligned} \epsilon(M) &\leq K \int_0^\infty e^{-\alpha t} P\{\tau_M \leq t\} dt \\ &\leq \frac{5K}{\alpha^3 M} [\alpha^2 (x_0^h)^2 + \alpha K_1 K_2 + 2K_1]. \end{aligned}$$

Now for a given accuracy requirement, say, within the limit  $\delta$ , choose  $M$  such that

$$\frac{5K}{\alpha^3 M} [\alpha^2 (x_0^h)^2 + \alpha K_1 K_2 + 2K_1] < \delta;$$

i.e., the boundary  $\partial[-M, M]^d$  depends on the choice of  $h$  and the accuracy requirement. For our example, the choice of boundary was  $x = -20, 20$ . For the calculation, we choose

$$\begin{aligned} b_1(x) &= (10^4 + |x|)e^{-|x|}, & b_2(x) &= |x|e^{-|x|}, \sigma = 1, \\ \tilde{r}_{11}(x) &= -\frac{2}{1 + |x|^3}, & \tilde{r}_{21}(x) &= \frac{1}{1 + 20|x|}, \\ \tilde{r}_{12}(x) &= \frac{2}{2 + x^2}, & \tilde{r}_{22}(x) &= -\frac{1}{1 + 10|x(x+1)|}, \alpha = \frac{1}{2}. \end{aligned}$$

We have run the program for  $h = 0.002, 0.001, 0.0005$ , and  $0.00025$  and performed 1000 iterations. Convergence has been achieved in all cases.

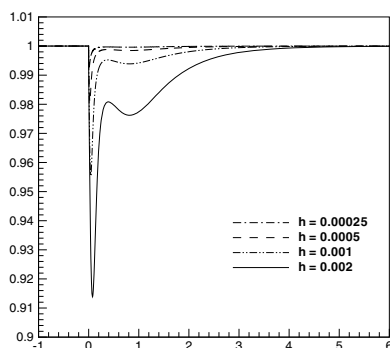


FIG. 1

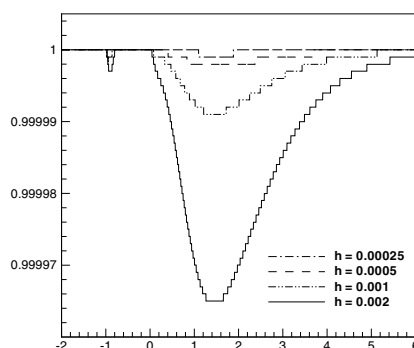


FIG. 2

In Figure 1, the state of the game is the  $x$  component, and the Nash equilibrium component of player 1 is the  $y$  component. Similarly, in Figure 2, the Nash equilibrium component of player 2 against the state of the game is given. The figures display that as  $h$  decreases the Nash equilibrium is approaching the pure strategies  $(1, 1)$ , i.e.,  $v_i^*(x) = 1$ ,  $x \in \mathbb{R}$ ,  $i = 1, 2$ .

**6. Conclusions.** We have developed a new method for obtaining a Nash equilibrium for stochastic differential games based on a discretization procedure. Our method relies mostly on probabilistic arguments, whereas the standard method in the literature involves a direct application of Fan's fixed point theorem. Moreover our method is constructive which can be used to compute Nash equilibria for a large class of stochastic differential games.

**Acknowledgments.** The author thanks M. K. Ghosh for valuable suggestions and S. Baskar for helping him with C programming. He thanks an anonymous referee for constructive comments.

## REFERENCES

- [1] A. BASU AND V. S. BORKAR, *Stochastic Control with Imperfect Models*, preprint.
- [2] D. P. BERTSEKAS, *Dynamic Programming and Stochastic Control*, Math. Sci. Eng. 125, Academic Press, New York, 1976.
- [3] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley and Sons, New York, 1968.
- [4] D. BLACKWELL, *Discounted dynamic programming*, Ann. Math. Stat., 33 (1962), pp. 719–726.
- [5] V. S. BORKAR, *Optimal Control of Diffusion Processes*, Res. Notes Math. 203, Longman, Harlow, 1989.
- [6] V. S. BORKAR AND M. K. GHOSH, *Stochastic differential games: An occupation measure based approach*, J. Optim. Theory Appl., 73 (1992), pp. 359–385; Errata corrige, ibid 88, (1996), pp. 251–252.
- [7] V. S. BORKAR AND M. K. GHOSH, *Denumerable state stochastic games with limiting average payoff*, J. Optim. Theory Appl., 76 (1993), pp. 539–560.
- [8] K. FAN, *Fixed point and minmax theorems in locally compact topological linear spaces*, Proc. Nat. Acad. Sci. USA, 38 (1952), pp. 121–126.
- [9] A. FEDERGRUEN, *On  $N$ -person stochastic games with denumerable state space*, Adv. Appl. Probab., 10 (1978), pp. 452–471.
- [10] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Appl. Math. 25, Springer-Verlag, New York, 1992.
- [11] M. K. GHOSH AND A. BAGCHI, *Stochastic games with average payoff criterion*, Appl. Math. Optim., 38 (1998), pp. 283–301.
- [12] M. K. GHOSH AND K. S. KUMAR, *A nonzero sum stochastic differential game in the orthant*, J. Math. Anal. Appl., 305 (2005), pp. 158–174.



- [13] I. KARATZAS AND S. E. SHREVE, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
- [14] H. J. KUSHNER, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Math. Sci. Eng. 129, Academic Press, New York, 1977.
- [15] H. J. KUSHNER, *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*, Stoch. Model. Appl. Probab. 47, Springer-Verlag, New York, 2001.
- [16] H. J. KUSHNER, *Numerical approximations for stochastic differential games*, SIAM J. Control Optim., 41 (2002), pp. 457–486.
- [17] H. J. KUSHNER, *Numerical approximation for stochastic differential games: The ergodic case*, SIAM J. Control Optim., 42 (2004), pp. 1911–1933.
- [18] H. J. KUSHNER, *Numerical Approximation for Nonzero Sum Stochastic Differential Games*, 2005, preprint.
- [19] H. J. KUSHNER AND P. G. DUPUIS, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York, 1992.
- [20] P. MANNUCCI, *Nonzero-sum stochastic differential games with discontinuous feedback*, SIAM J. Control Optim., 43 (2004), pp. 1222–1233.
- [21] J. F. MERTONS AND T. PARTHASARATHY, *Equilibria for Discounted Stochastic Games*, CORE discussion paper 8750.
- [22] A. S. NOWAK, *Nonrandomized strategy equilibria in noncooperative games with additive transition and reward structure*, J. Optim. Theory Appl., 52 (1987), pp. 429–441.
- [23] A. S. NOWAK AND T. E. S. RAGHAVAN, *Existence of stationary correlated equilibria with symmetric information to discounted stochastic games*, Math. Oper. Res., 17 (1992), pp. 519–526.
- [24] B. OKSENDAL AND K. REIKVAM, *Stochastic differential games with controls—discussion of a specific example*, in Proceedings of the Symposium on Mathematical Finance, E. Lungu, ed., University of Botswana, 1997.
- [25] T. PARTHASARATHY AND S. SINHA, *Existence of stationary equilibrium strategies in nonzero sum stochastic games with uncountable state space and state dependent transitions*, Internat. J. Game Theory, 18 (1989), pp. 189–194.
- [26] K. S. KUMAR, *Numerical analysis of a zero-sum stochastic differential game*, Comput. Appl. Math., 19 (2000), pp. 239–265.

## CONSTRAINED OPTIMAL CONTROL THEORY FOR DIFFERENTIAL LINEAR REPETITIVE PROCESSES\*

M. DYM KOV<sup>†</sup>, E. ROGERS<sup>‡</sup>, S. DYM KOU<sup>§</sup>, AND K. GALKOWSKI<sup>¶</sup>

**Abstract.** Differential repetitive processes are a distinct class of continuous-discrete two-dimensional linear systems of both systems theoretic and applications interest. These processes complete a series of sweeps termed passes through a set of dynamics defined over a finite duration known as the pass length, and once the end is reached the process is reset to its starting position before the next pass begins. Moreover the output or pass profile produced on each pass explicitly contributes to the dynamics of the next one. Applications areas include iterative learning control and iterative solution algorithms, for classes of dynamic nonlinear optimal control problems based on the maximum principle, and the modeling of numerous industrial processes such as metal rolling, long-wall cutting, etc. In this paper we develop substantial new results on optimal control of these processes in the presence of constraints where the cost function and constraints are motivated by practical application of iterative learning control to robotic manipulators and other electromechanical systems. The analysis is based on generalizing the well-known maximum and  $\epsilon$ -maximum principles to them.

**Key words.** two-dimensional systems, optimal control, constraints

**AMS subject classifications.** 93C05, 93C15

**DOI.** 10.1137/060668298

**1. Introduction.** Repetitive processes are a distinct class of two-dimensional (2D) systems of both systems theoretic and applications interest. The unique characteristic of such a process is a series of sweeps, termed passes, through a set of dynamics defined over a fixed finite duration known as the pass length. On each pass an output, termed the pass profile, is produced which acts as a forcing function on, and hence contributes to, the dynamics of the next pass profile. This, in turn, leads to the unique control problem in that the output sequence of pass profiles generated can contain oscillations which increase in amplitude in the pass-to-pass direction.

To introduce a formal definition, let  $\alpha < +\infty$  denote the pass length (assumed constant). Then in a repetitive process the pass profile  $y_k(t)$ ,  $0 \leq t \leq \alpha$ , generated on pass  $k$  acts as a forcing function on, and hence contributes to, the dynamics of the next pass profile  $y_{k+1}(t)$ ,  $0 \leq t \leq \alpha$ ,  $k \geq 0$ .

Physical examples of repetitive processes include long-wall coal cutting and metal-rolling operations (see, for example, the references cited in [17]). Also in recent years applications have arisen where adopting a repetitive process setting for analysis has distinct advantages over alternatives. Examples of these so-called algorithmic applications include classes of iterative learning control (ILC) schemes (see, for example, [13]) and iterative algorithms for solving nonlinear dynamic optimal control problems based on the maximum principle [15]. In the case of iterative learning control for the linear

---

\*Received by the editors August 25, 2006; accepted for publication (in revised form) September 18, 2007; published electronically January 30, 2008.

<http://www.siam.org/journals/sicon/47-1/66829.html>

<sup>†</sup>Belarus State Economic University, Partizanski Ave., 26, Minsk, Belarus (dymkov\_m@bseu.by).

<sup>‡</sup>Corresponding author. School of Electronics and Computer Science, University of Southampton, Southampton SO 17 1BJ, UK (etar@ecs.soton.ac.uk).

<sup>§</sup>Department of Applied Mathematics II, Friedrich-Alexander-University of Erlangen-Nuremberg, Martensstrasse 3, 91058 Erlangen, Germany (dymkou@am2.am.uni-erlangen.de).

<sup>¶</sup>Institute of Control and Computation Engineering, University of Zielona Góra, Podgorna 50, 65-246, Zielona Góra, Poland (K.Galkowski@issi.uz.zgora.pl).

dynamics case, the stability theory for differential (and discrete) linear repetitive processes is one method which can be used to undertake a stability/convergence analysis of a powerful class of such algorithms and thereby produce vital design information concerning the trade-offs required between convergence and transient performance (see, for example, [14]).

Attempts to control these processes using standard (or 1D) systems theory and associated algorithms fail (except in a few very restrictive special cases) precisely because such an approach ignores their inherent 2D systems structure, i.e., information propagation occurs from pass to pass and along a given pass. Also the initial conditions are reset before the start of each new pass, and the structure of these can be somewhat complex. For example, if they are an explicit function of points on the previous pass profile, then this alone can destroy stability. In seeking a rigorous foundation on which to develop a control theory for these processes, it is natural to attempt to exploit structural links which exist between these processes and other classes of 2D linear systems.

The case of 2D discrete linear systems recursive in the positive quadrant  $(i, j) : i \geq 0, j \geq 0$  (where  $i$  and  $j$  denote the directions of information propagation) has been the subject of much research effort over the years using, in the main, the well-known Roesser [16] and Fornasini–Marchesini [8] state-space models. One approach which has been the subject of productive research is optimal control—see, for example, [4, 18]. More recently, productive research has been reported on  $\mathcal{H}_\infty$  and  $\mathcal{H}_2$  approaches to analysis and controller design—see, for example, [19, 7]. In this paper we consider so-called differential linear repetitive processes where information propagation along the pass is governed by a matrix differential equation. The systems theory for 2D discrete linear systems is therefore not applicable. (Also, as noted above, for discrete processes the resetting and structure of the boundary conditions may cause problems which have no Roesser or Fornasini–Marchesini state-space model counterparts.)

In this paper we develop substantial new results on the optimal control of differential linear repetitive processes with constraints which we motivate from the iterative learning control application. The results themselves are obtained by extending the maximum principle and the  $\epsilon$ -maximum principle [11] to them. A sensitivity analysis of the resulting optimal control is also undertaken, and some relevant differentiation properties are established. Finally, a numerical example is given.

**2. Preliminaries.** ILC is a technique for controlling systems which are required to continually repeat the same operation with the requirement that a reference trajectory defined over a finite interval is followed to a high precision. In particular, the system completes a pass (also known as a trial in some literature) and is then reset, the next pass is completed, and so on. The basic idea of ILC is to use information from previous executions of the task in order to improve performance from pass to pass in the sense that the tracking error is sequentially reduced. It is clear therefore that ILC can easily be formulated as a repetitive process, and the stability theory for them can be used to explain why an incorrectly designed ILC scheme can result in nonconvergent behavior which manifests itself as oscillations that increase in amplitude from pass to pass [14].

Since the original work in the mid-1980s [3], the general area of ILC has been the subject of considerable research in terms of the underlying theory (with experimental verification in some cases). Commonly used ILC algorithms construct the input to the plant or process from the input used on the last pass plus an additive increment, which is typically a function of the past values of the measured output error, i.e.,

the difference between the achieved output on the current pass and the desired plant output. Suppose that  $u_k(t)$  denotes the input to the plant on pass  $k$  which is of duration  $\alpha$ , i.e.,  $0 \leq t \leq \alpha < \infty$ . Suppose also that  $e_k(t) = r(t) - y_k(t)$  denotes the current pass error. Then the objective of constructing a sequence of input functions such that the performance achieved is gradually improving with each successive pass can be refined to a convergence condition on the input and error, i.e.,

$$\lim_{k \rightarrow \infty} \|e_k\| = 0, \quad \lim_{k \rightarrow \infty} \|u_k - u_\infty\| = 0,$$

where  $\|\cdot\|$  is a signal norm in a suitably chosen function space with a norm-based topology and  $u_\infty$  is termed the learned control.

A large number of design algorithms have been developed for this general area, some of which have also been experimentally tested. Of these, a good number are based on minimization of a cost function. Given the tracking nature of this general problem in the pass-to-pass direction, it is clearly necessary to penalize control action to prevent a “large” error resulting in the demand for an unacceptably high control input on the next pass in an attempt to minimize the error. One class of such algorithms is termed norm-optimal (with an extension to so-called predictive norm-optimal which is not relevant here). Here (see [1] for full details) on completion of pass  $k$ , the control input for pass  $k+1$  is computed as the solution of the minimum norm optimization problem

$$u_{k+1} = \arg \min_{u_{k+1}} \{J_{k+1}(u_{k+1}) : e_{k+1} = r - y_{k+1}, y_{k+1} = Gu_{k+1}\},$$

where the performance index, or optimality criterion, used is defined to be

$$J_{k+1}(u_{k+1}) = \|e_{k+1}\|_{\mathcal{Y}}^2 + \|u_{k+1} - u_k\|_{\mathcal{U}}^2,$$

where  $\mathcal{Y}$  is a real Hilbert space of output or pass profile signals ( $y_k$ ) and  $\mathcal{U}$  is a real (and possibly distinct) Hilbert space of input signals ( $u_k$ ). Here the initial control  $u_0 \in \mathcal{U}$  can be arbitrary in theory but, in practice, will be a good first guess at the solution of the problem. This problem can be interpreted as the determination of the control input on pass  $k+1$  with the properties that: (i) the tracking error is reduced in an optimal way; and (ii) this new control input does not deviate too much from the control input used on pass  $k$ . The relative weighting of these two objectives can be absorbed into the definitions of the norms in  $\mathcal{Y}$  and  $\mathcal{U}$ . Other approaches to learning control in the presence of input constraints can be found in, for example, [9, 20] (but note that repetitive processes appear in formulations which have no iterative learning control interpretation). Suppose now that the plant dynamics are described by the following matrix differential equation:

$$(1) \quad \frac{dx_k(t)}{dt} = Ax_k(t) + Dx_{k-1}(t) + bu_k(t), \quad 0 \leq t \leq \alpha, \quad k \geq 0,$$

where, on pass  $k$ ,  $x_k(t)$  is the  $n \times 1$  state (equal to the pass profile or output) vector,  $u_k(t)$  is the scalar control input,  $A, D$  are constant  $n \times n$  matrices, and  $b$  is a given  $n \times 1$  vector. (This model is chosen for simplicity of presentation and is easily extended to the case when the pass profile vector is a linear combination of the current pass state, input, and previous pass profile vectors.)

Then it is straightforward to show that the above formulation includes the choice of a linear quadratic cost function as a special case, but the solution has to be modified

slightly to guarantee that the resulting Riccati equation-based solution is causal in the sense that it does not, as the dynamics evolve, require use of information which is not yet available—see [1] for the details here. Algorithms resulting from this approach have been experimentally tested on a chain conveyor system with, on the whole, very encouraging results [2]. However, in some cases it was observed that the computed control input (a scalar variable in this application) was still above the safe operating range of the actuator device and the experiment had to be stopped to prevent damage. Also there was a tendency for the output at the end of each pass to “dip down” in value.

Another feature of repetitive processes which does not appear in the above optimal control problem is that in each practical application only a finite number of passes will actually be completed. Suppose therefore that  $N < \infty$  denotes the number of passes actually completed, introduce the set  $K = \{1, 2, \dots, N\}$ , and let  $T$  denote the finite interval (the pass length)  $[0, \alpha]$ . Then, with the above observations in mind, consider (1) with boundary conditions

$$(2) \quad x_k(0) = d_k, \quad k \in K, \quad x_0(t) = f(t), \quad t \in T,$$

where  $d_k$  is an  $n \times 1$  vector with constant entries and  $f(t)$  is a known function  $t \in T$ . Then the optimal control problem considered is

$$(3) \quad \max_{u_k} J(u), \quad J(u) = \sum_{k \in K} p_k^T x_k(\alpha),$$

where  $p_k$ ,  $k = 1, \dots, N$ , is a given  $n \times 1$  vector subject to an end of pass (or terminal) constraint of the form

$$(4) \quad H_k x_k(\alpha) = o_k, \quad k \in K,$$

where  $o_k$  is an  $m \times 1$  vector and  $H_k$  is an  $m \times n$  matrix, and the control inputs satisfy the following admissibility condition.

**DEFINITION 1.** *For each pass number  $k \in K$  the piecewise continuous function  $u_k : T \rightarrow R$  is termed an admissible control for this pass if it satisfies*

$$(5) \quad |u_k(t)| \leq 1, \quad t \in T,$$

*and the corresponding state vector  $x_k(t)$ ,  $t \in T$ , of (1) satisfies the boundary conditions*

$$x_k(0) = d_k, \quad H_k x_k(\alpha) = o_k.$$

Also, without loss of generality, we assume that the matrix  $A$  has simple eigenvalues  $\lambda_i$ ,  $1 \leq i \leq n$ , and that it is stable in the sense that  $\text{Re } \lambda_i < 0$ ,  $1 \leq i \leq n$ . (Stability of the matrix  $A$  is a necessary condition for so-called stability along the pass (essentially bounded input bounded output stability) independent of the pass length [17].)

**3. Optimality conditions for the supporting control functions.** Consider first (1)–(2) in the absence of the terminal conditions (4). Then it has been shown elsewhere [5] that the solution of these equations can be written as

$$(6) \quad \begin{aligned} x_k(t) = & \sum_{j=1}^k K_j(t) d_{k+1-j} + \int_0^t K_k(t-\tau) D f(\tau) d\tau \\ & + \sum_{j=1}^k \int_0^t K_j(t-\tau) b u_{k+1-j}(\tau) d\tau, \quad k = 1, \dots, N, \end{aligned}$$

where the  $K_i(t)$  are the solutions of the following  $n \times n$  matrix differential equations:

$$(7) \quad \dot{K}_1(t) = AK_1(t), \quad \dot{K}_i(t) = AK_i(t) + DK_{i-1}(t), \quad i = 2, \dots, N,$$

with initial conditions

$$(8) \quad K_1(0) = I_n, \quad K_i(0) = 0, \quad i = 2, \dots, N.$$

Also it is easy to show that these solutions have the following properties:

$$(9) \quad \begin{aligned} K_j(t - \sigma) &= \int_{\sigma}^t K_{j-k}(t - \tau) DK_k(\tau - \sigma) d\tau, \quad 0 \leq \sigma < t \leq \alpha, \quad k = 1, \dots, j-1, \\ K_j(t - \sigma) &= \sum_{s=1}^j K_s(t - \tau) K_{j+1-s}(\tau - \sigma), \quad j = 2, \dots, N-1, \end{aligned}$$

which will be used below.

Now by using (6) we can rewrite the optimal problem considered here in the following integral form:

$$(10) \quad \max_{u_1, \dots, u_N} J(u), \quad J(u) = \sum_{j=1}^N \int_0^{\alpha} c_j(\tau) u_j(\tau) d\tau + \gamma,$$

subject to the terminal conditions (4) and the control constraint (5). Also we can write

$$(11) \quad \begin{aligned} &\int_0^{\alpha} g_{11}(\tau) u_1(\tau) d\tau = h_1, \\ &\int_0^{\alpha} \left[ g_{21}(\tau) u_1(\tau) + g_{22}(\tau) u_2(\tau) \right] d\tau = h_2 \\ &\dots\dots\dots \\ &\int_0^{\alpha} \left[ g_{N1}(\tau) u_1(\tau) + \dots + g_{NN}(\tau) u_N(\tau) \right] d\tau = h_N, \end{aligned}$$

and

$$|u_k(\tau)| \leq 1, \quad \tau \in T, \quad k = 1, \dots, N,$$

where the scalar  $\gamma$  and the scalar functions  $c_j(\tau)$  are defined as follows:

$$\begin{aligned} \gamma &= \sum_{k=1}^N \sum_{j=1}^k p_k^T K_j(\alpha) d_{k+1-j} + \sum_{k=1}^N \int_0^{\alpha} p_k^T K_k(\alpha - \tau) Df(\tau) d\tau, \\ c_j(\tau) &= \sum_{k=j}^N p_k^T K_{k+1-j}(\alpha - \tau) b, \quad j = 1, \dots, N, \quad g_{kj}(\tau) = H_k K_{k+1-j}(\alpha - \tau) b, \quad j \leq k, \\ h_k &= o_k - \sum_{j=1}^k H_k K_j(\alpha) d_{k+1-j} - \int_0^{\alpha} H_k K_k(\alpha - \tau) Df(\tau) d\tau, \quad k = 1, \dots, N. \end{aligned}$$

Also we require the following.







values are less than those on the control constraint boundary and satisfy (11); i.e., the support control function is nonsingular if there exist numbers  $\lambda_0 > 0$ ,  $\mu_0 > 0$ ,  $u_j^k(\lambda)$ ,  $j = 1, \dots, m$ ,  $k = 1, \dots, N$ , such that the following equalities:

$$(17) \quad \sum_{j=1}^k \sum_{i=1}^m u_j^i(\lambda) \int_{\tau_{ij}-\lambda}^{\tau_{ij}+\lambda} g_{kj}(t) dt = \sum_{j=1}^k \sum_{i=1}^m \int_{\tau_{ij}-\lambda}^{\tau_{ij}+\lambda} g_{kj}(t) u_j(t) dt, \\ |u_j^k| \leq 1 - \mu_0, \quad j = 1, \dots, m, \quad k = 1, \dots, N,$$

hold for all  $\lambda$ ,  $0 < \lambda < \lambda_0$ , and  $k$ ,  $1 \leq k \leq N$ . This fact will be used below in the proof of the optimality conditions.

Associate with each supporting time instance  $\tau_{kj}$  a small subinterval  $T_{kj}$  from  $T$  such that the matrix  $G_{gen}^k := \{ \int_{T_{kj}} g_{kk}(\tau) d\tau, j = 1, \dots, m \}$  is nonsingular. Also without loss of generality we can assume that  $\tau_{kj}$  is one or the other of the end points of  $T_{kj}$  and the supporting control functions  $u_k(t) = u_j^k$  for  $t \in T_{kj}$ ,  $j = 1, \dots, N$ , are constant over the segments  $T_{kj}$ . Then we have the following result.

**THEOREM 1.** *A supporting control function  $\{\tau_{sup}^k, u_k^0(t), k = 1, \dots, N\}$  is an optimal solution of the problem (1)–(4) if*

$$(18) \quad u_k^0(t) = -\text{sgn}(\Delta_k(t)), \quad k = 1, \dots, N, \quad t \in T.$$

*Moreover, if this supporting control function is nondegenerate, then the above condition is necessary and sufficient.*

*Proof.* Let  $u_k(t) \neq u_k^0(t)$ ,  $k = 1, \dots, N$ , be an admissible control and  $x_k(t)$  the corresponding trajectory of the system (1)–(2). Then standard transformations yield that the increment  $\Delta J(u) := J(u^0) - J(u)$  of the cost function can be expressed in the form

$$\Delta J(u) = \int_0^{t^*} \sum_{j=1}^N c_j(t) [u_j^0(t) - u_j(t)] dt = - \sum_{j=1}^N \int_0^{t^*} \Delta_j(t) [u_j^0(t) - u_j(t)] dt.$$

Hence by using (18) we have that  $\Delta J(u) \geq 0$  for any admissible control  $u$ ; i.e.,  $\{\tau_{sup}^k, u_k^0\}$  is an optimal supporting control function.

Let  $\{\tau_{sup}^k, u_k^0(t), k = 1, \dots, N\}$  be an optimal nondegenerate control, but there exists  $k_*, 1 \leq k_* \leq N$ , and there exists  $t_* \in T$  such that the theorem is not valid. Suppose also that  $t_* \in [\tau_{k_*j} - \lambda, \tau_{k_*j} + \lambda]$ , where  $\lambda > 0$  is a small number; i.e., the instance  $t_*$  lies in an neighborhood of some supporting time instance  $\tau_{k_*j}$ . Then since the supporting control is nondegenerate there exists a control variation  $\Delta u_{k_*}^0(t)$ , defined on the intervals  $[\tau_{k_*j} - \lambda, \tau_{k_*j} + \lambda]$ , such that  $J(u^0) > 0$ , which contradicts the optimality of  $u_k^0(t)$ .

Given this last fact, we now suppose that  $t_* \notin [\tau_{k_*j} - \lambda, \tau_{k_*j} + \lambda]$  for all  $j = 1, \dots, m$ , for some small  $\lambda > 0$ . Also, without loss of generality, we assume that  $\Delta_{k_*}^0(t_*) > 0$  and  $u_{k_*}(t_*) > 0$ . Then, by continuity of  $\Delta_{k_*}(t)$  and piecewise-continuity of  $u_{k_*}(t)$ , there exists a neighborhood  $T_{k_*}(t_*)$  of  $t_*$ , such that  $\Delta_{k_*}(t) > 0$ ,  $u_{k_*}(t) > -1$  for  $t \in T_{k_*}(t_*)$ . Now we have to construct the admissible control variation such that the corresponding increment of the cost function satisfies  $\Delta J(u) > 0$ , which is impossible for the optimal controls  $u_k^0(t)$ .

Consider now the case of a small real number  $\lambda_0 > 0$  (we see below that the existence of such a number  $\lambda_0$  is guaranteed by the fact that the supporting control is nondegenerate), and for all  $\lambda$ ,  $0 < \lambda < \lambda_0$ , define the control variation  $\Delta u(t) =$

$(\Delta u_1(t), \dots, \Delta u_N(t))$ ,  $t \in T$ , as

$$\Delta u_k(t) = 0, \quad k < k_*, \quad t \in T,$$

$$\Delta u_{k_*}(t) = \begin{cases} \theta(-1 - u_{k_*}(t)), & \theta > 0, \quad t \in T_{k_*}(t), \\ 0, & t \in T \setminus \left( \bigcup_{j=1}^m [\tau_{k_*j} - \lambda, \tau_{k_*j} + \lambda] \cup T_{k_*}(t) \right). \end{cases}$$

We now have that the control variations on the intervals  $[\tau_{k_*j} - \lambda, \tau_{k_*j} + \lambda]$ ,  $j = 1, \dots, m$ , can be chosen as constant functions  $\Delta u_{k_*}(t) \equiv \Delta \vartheta_j^k(\lambda)$ . Those for the remaining passes  $k > k_*$  are defined as

$$\Delta u_k(t) \equiv 0, \quad k = k_* + 1, \dots, N, \quad t \in T \setminus \bigcup_{j=1}^m [\tau_{kj} - \lambda, \tau_{kj} + \lambda],$$

$$\Delta u_k(t) \equiv \Delta \vartheta_j^k(\lambda), \quad t \in [\tau_{kj} - \lambda, \tau_{kj} + \lambda], \quad j = 1, \dots, m, \quad k > k_*,$$

where the  $\Delta \vartheta_j^k(\lambda)$  are unknown constants determined below.

By using (11), it follows that the conditions

$$(19) \quad \int_0^\alpha \sum_{s=1}^k g_{ks}(\tau) \Delta u_s(\tau) d\tau = 0, \quad k = 1, \dots, N,$$

hold for any admissible variation  $\Delta u(t)$ , and these can be rewritten in the form

$$\begin{aligned} \phi_{k_*}(\lambda) &= \sum_{j=1}^m \int_{\tau_{k_*j}-\lambda}^{\tau_{k_*j}+\lambda} g_{k_*k_*}(\tau) \vartheta_j^{k_*}(\lambda) d\tau \\ &= -\theta \int_{T_{k_*}(t_*)} g_{k_*k_*}(\tau) (-1 - u_{k_*}(\tau)) d\tau, \\ \phi_{k_*+1}(\lambda) &:= \sum_{j=1}^m \int_{\tau_{k_*+1j}-\lambda}^{\tau_{k_*+1j}+\lambda} g_{k_*+1k_*+1}(\tau) \vartheta_j^{k_*+1}(\lambda) d\tau \\ &= -\sum_{j=1}^m \int_{\tau_{k_*j}-\lambda}^{\tau_{k_*j}+\lambda} g_{k_*+1k_*}(\tau) \vartheta_j^{k_*}(\lambda) d\tau - \theta \int_{T_{k_*}(t_*)} g_{k_*+1k_*}(\tau) (-1 - u_{k_*}(\tau)) d\tau, \\ &\dots\dots\dots \\ \phi_N(\lambda) &= \sum_{j=1}^m \int_{\tau_{Nj}-\lambda}^{\tau_{Nj}+\lambda} g_{NN}(\tau) \vartheta_j^N(\lambda) d\tau - \sum_{j=1}^m \int_{\tau_{k_*j}-\lambda}^{\tau_{k_*j}+\lambda} g_{Nk_*}(\tau) \vartheta_j^{k_*}(\lambda) d\tau \\ &\quad - \theta \int_{T_{k_*}(t_*)} g_{Nk_*}(\tau) (-1 - u_{k_*}(\tau)) d\tau \\ (20) \quad &- \dots - \sum_{j=1}^m \int_{\tau_{N-1j}-\lambda}^{\tau_{N-1j}+\lambda} g_{NN-1}(\tau) \vartheta_j^{N-1}(\lambda) d\tau. \end{aligned}$$

Expanding the function  $\phi_{k_*}(\lambda)$  of (20) as a Taylor series and setting  $\Delta \vartheta_\lambda^{k_*} = (\Delta \vartheta_1^{k_*}(\lambda), \dots, \Delta \vartheta_m^{k_*}(\lambda))$  now yields

$$\begin{aligned} 2\lambda G_{sup}^{k_*} \Delta \vartheta_\lambda^{k_*} + \frac{\lambda^3}{3} \left\{ \frac{d^2 g_{k_*k_*}(\tau_{k_*j})}{d\tau}, \quad j = 1, \dots, m \right\} \Delta \vartheta_\lambda^{k_*} + o_{k_*}(\lambda^3) \\ = -\theta \int_{T_{k_*}(t_*)} g_{k_*k_*}(\tau) (-1 - u_{k_*}(\tau)) d\tau, \end{aligned}$$

where  $o_{k_*}(\lambda^3)$  denotes terms of degree 3 and above, which are neglected here. Hence the required vector  $\Delta\vartheta_\lambda^{k_*}$  can be represented as

$$\Delta\vartheta_\lambda^{k_*} = \frac{1}{\lambda}\theta\hat{u}_{k_*} + \theta o_{k_*}(\lambda), \quad \text{where} \quad \hat{u}_{k_*} = -\frac{1}{2}G_{sup}^{k_*-1} \int_{T_{k_*}(t_*)} g_{k_*k_*}(\tau)(-1 - u_{k_*}(\tau))d\tau, \quad (21)$$

and  $o_{k_*}(\lambda)$  denotes a residual first order term. Using (17) and (21), it follows that for a small value of  $\lambda \in (0, \lambda_0)$  there exists the real number  $\theta = \theta(\lambda)$  such that  $\theta(\lambda) = \mu_{k_*}\lambda \leq 1$ , where  $\mu_{k_*} > 0$  does not depend on  $\lambda$ , and the following inequalities:

$$|u_j^{k_*}(\lambda) + \Delta\vartheta_j^{k_*}(\lambda)| \leq 1, \quad j = 1, \dots, m,$$

hold. Here we have exploited the fact that the admissible controls are constants  $u_j^k(\lambda)$  over the intervals  $T_j^k$ , containing the supporting points  $\tau_{kj}$ . Hence, the function

$$\bar{u}_{k_*}(t) = \begin{cases} u_j^{k_*}(\lambda) + \Delta\vartheta_j^{k_*}(\lambda), & t \in [\tau_{k_*j} - \lambda, \tau_{k_*j} + \lambda], \\ u_{k_*}(t) + \theta(\lambda)(-1 - u_{k_*}(t)), & t \in T_{k_*}(t_*), \end{cases}$$

is an admissible control function for  $\theta(\lambda) = \mu_{k_*}\lambda \leq 1$  and a sufficiently small  $\mu_{k_*}$ .

In order to find  $\Delta\vartheta_\lambda^{k_*+1}$  and  $\theta(\lambda)$ , expand  $\phi_{k_*+1}(\lambda)$  as a Taylor series to yield

$$\begin{aligned} \sum_{j=1}^m \int_{\tau_{k_*j}-\lambda}^{\tau_{k_*j}+\lambda} g_{k_*+1k_*}(\tau) \Delta\vartheta_j^{k_*}(\lambda) d\tau &= 2\lambda \sum_{j=1}^m g_{k_*+1k_*}(\xi_j) \Delta\vartheta_j^{k_*}(\lambda) \\ &= 2\lambda \tilde{G}_\xi^{k_*+1} \Delta\vartheta_\lambda^{k_*+1} \\ &= 2\lambda \tilde{G}_\xi^{k_*+1} \left( \frac{1}{\lambda} \mu_{k_*} \lambda \hat{u}_{k_*} + \mu_{k_*} \lambda o_{k_*}(\lambda) \right) \\ &= 2\tilde{G}_\xi^{k_*+1} \mu_{k_*} \lambda \hat{u}_{k_*} + \mu_{k_*} \lambda o_{k_*}(\lambda^3). \end{aligned} \quad (22)$$

Here the matrix  $\tilde{G}_\xi^{k_*+1}$  is constructed from the rows  $\{g_{k_*+1k_*}(\xi_j), j = 1, \dots, m\}$ , where  $\xi_j$  are points from the intervals  $[\tau_{k_*j} - \lambda, \tau_{k_*j} + \lambda]$ . Next, set  $\Delta\vartheta_\lambda^{k_*+1} = (\Delta\vartheta_1^{k_*+1}(\lambda), \dots, \Delta\vartheta_m^{k_*+1}(\lambda))$  to obtain

$$\begin{aligned} 2\lambda G_{sup}^{k_*+1} \Delta\vartheta_\lambda^{k_*+1} + \frac{\lambda^3}{3} \left\{ \frac{d^2 g_{k_*+1k_*+1}(\tau_{k_*j})}{d\tau}, j = 1, \dots, m \right\} \Delta\vartheta_\lambda^{k_*+1} + o_{k_*+1}(\lambda^3) \\ = -\mu_{k_*} \lambda \left\{ \tilde{G}_\xi^{k_*+1} \hat{u}_{k_*} + \int_{T_{k_*}(t_*)} g_{k_*+1k_*+1}(\tau)(-1 - u_{k_*}(\tau))d\tau \right\} \\ + \mu_{k_*} \lambda o_{k_*}(\lambda^3), \end{aligned} \quad (23)$$

and hence the required vector  $\Delta\vartheta_\lambda^{k_*+1}$  can be expressed as

$$\begin{aligned} \Delta\vartheta_\lambda^{k_*+1} &= \frac{1}{\lambda} \mu_{k_*} \lambda \hat{u}_{k_*+1} + \mu_{k_*} \lambda o_{k_*+1}(\lambda), \\ \hat{u}_{k_*+1} &= -\frac{1}{2} (G_{sup}^{k_*+1})^{-1} \left\{ \tilde{G}_\xi^{k_*+1} \hat{u}_{k_*} - \int_{T_{k_*}(t_*)} g_{k_*+1k_*+1}(\tau)(-1 - u_{k_*}(\tau))d\tau \right\}. \end{aligned} \quad (24)$$

Now choose  $\Delta\vartheta_\lambda^{k_*+1}$  such that the following inequalities hold:

$$|u_j^{k_*+1}(\lambda) + \Delta\vartheta_j^{k_*+1}(\lambda)| \leq 1, \quad j = 1, \dots, m,$$

and hence the values of  $\mu_{k_*}$  and  $\lambda_0$  can be decreased as required. Continuing this expansion procedure for the remaining equations in (20), we obtain the desired admissible control function in the form

$$\bar{u}(t) = u^0(t) + \Delta u(t) = \left\{ u_1^0(t) + \Delta u_1(t), \dots, u_N^0(t) + \Delta u_N(t) \right\}, \quad t \in T,$$

and note here that  $\Delta u_k(t) = 0$  for all  $k < k_*$ .

At this stage we can calculate the increment of the cost function generated by the designed control function  $\bar{u}(t)$  as

$$\begin{aligned} \Delta J(u) &= J(\bar{u}) - J(u^0) = \sum_{k=1}^N \int_0^\alpha \Delta_k(t) \Delta u_k(t) dt = \sum_{k=k_*}^N \int_0^\alpha \Delta_k(t) \Delta u_k(t) dt \\ &= -\theta \int_{T_{k_*}(t_*)} \Delta_{k_*}(t) (-1 - u_{k_*}(t)) dt \\ &\quad - \sum_{j=1}^m \int_{\tau_{k_*j}-\lambda}^{\tau_{k_*j}+\lambda} \Delta_{k_*}(t) [u_j^{k_*}(\lambda) + \Delta \vartheta_j^{k_*}(\lambda) - u_j^{k_*}(t)] dt \\ (25) \quad &\quad - \sum_{s=k_*+1}^N \sum_{j=1}^m \int_{\tau_{sj}-\lambda}^{\tau_{sj}+\lambda} \Delta_s(t) [u_j^s(\lambda) + \Delta \vartheta_j^s(\lambda) - u_j^s(t)] dt. \end{aligned}$$

Since  $\Delta_k(\tau_{kj}) = 0$ ,  $k = k_*, \dots, N$ ,  $j = 1, \dots, m$ , then, again using the Taylor series expansion in  $\lambda$ , we have the following estimate for the integral components:

$$\begin{aligned} \int_{\tau_{sj}-\lambda}^{\tau_{sj}+\lambda} \Delta_s(t) [u_j^s(\lambda) + \Delta \vartheta_j^s(\lambda) - u_j^s(t)] dt &= \int_{\tau_{sj}}^{\tau_{sj}} \Delta_s(t) [u_j^s(\lambda) + \Delta \vartheta_j^s(\lambda) - u_j^s(t)] dt \\ &\quad + 2\lambda \Delta_s(\tau_{sj}) [u_j^s(\lambda) + \Delta \vartheta_j^s(\lambda) - u_j^s(\tau_{sj})] \\ &\quad + \lambda^2 \frac{d\Delta_s(\tau_{sj})}{dt} [u_j^s(\lambda) + \Delta \vartheta_j^s(\lambda) - u_j^s(\tau_{sj})] \\ (26) \quad &\quad + o_1(\lambda^2) \cong o(\lambda^2). \end{aligned}$$

Hence (25) and (26) yield

$$(27) \quad \Delta J(u) = -\mu_{k_*} \lambda \int_{T_{k_*}(t_*)} \Delta_{k_*}(t) (-1 - u_{k_*}(t)) dt + o(\lambda) > 0$$

for a sufficiently small  $\lambda > 0$ , which contradicts the optimality of control functions  $u_k^0(t)$ ,  $k = 1, \dots, N$ .  $\square$

*Remark 3.* The analysis which now follows shows that the above result can be reformulated in the traditional maximum principle form. In particular, it will be shown that the cocontrol functions  $\Delta_k(t)$ ,  $t \in T$ , here are connected directly to the adjoint (dual) variables  $\psi_k(t)$ ,  $t \in T$ , as  $\Delta_k(t) = -\psi_k^T(t)b$ . Note also that the term  $\psi_k^T(t)b$  is part of the Hamiltonian function which arises in the maximum principle statement of the result here. Moreover, the vectors  $\{\nu^{(k)}, k = 1, \dots, N\}$  (termed Lagrange multipliers in some literature) will be used as the boundary conditions for the corresponding differential equations describing the adjoint (dual) variables  $\psi_k(t)$  (in contrast to the classic maximum principle, where such boundary conditions are not specified).

Let  $\psi_N(t)$  be the solution of

$$(28) \quad \frac{d\psi_N(t)}{dt} = -A^T \psi_N(t), \quad \psi_N(\alpha) = p_N - H_N^T \nu^N, \quad t \in T,$$

or

$$(29) \quad \psi_N(t) = K_1^T(\alpha - t)\psi(\alpha), \quad t \in T.$$

Hence

$$(30) \quad \begin{aligned} \psi_N^T(t)b &= (p_N^T - (\nu^N)^T H_N) K_1(\alpha - t)b = p_N^T K_1(\alpha - t)b \\ &- (\nu^N)^T H_N K_1(\alpha - t)b = c_N(t) - (\nu^N)^T g_{NN}(t) = -\Delta_N(t). \end{aligned}$$

In order to verify the validity of the corresponding conditions for subsequent passes we use (9) for the differential equations (7). Let  $\psi_{N-1}(t)$ ,  $t \in T$ , be a solution of the differential equation

$$(31) \quad \frac{d\psi_{N-1}(t)}{dt} = -A^T \psi_{N-1}(t) - D^T \psi_N(t), \quad \psi_{N-1}(\alpha) = p_{N-1} - H_{N-1}^T \nu^{N-1}, \quad t \in T.$$

Then

$$(32) \quad \begin{aligned} \psi_{N-1}^T(t)b &= (p_{N-1}^T - (\nu^{N-1})^T H_{N-1}) K_1(\alpha - t)b \\ &- (p_N^T - (\nu^N)^T H_N) \int_0^t K_1^T(t - \tau) D^T K_1^T(\alpha - \tau) b d\tau \\ &= p_{N-1}^T K_1(\alpha - t)b - (\nu^{N-1})^T H_{N-1} K_1(\alpha - t)b \\ &- (p_N^T - (\nu^N)^T H_N) K_2(\alpha - t)b \\ &= c_{N-1}(t) - (\nu^{N-1})^T g_{N-1N-1}(t) - (\nu^N)^T g_{NN-1}(t) = -\Delta_{N-1}(t). \end{aligned}$$

By analogy with the case for (31)–(32), we have

$$(33) \quad \psi_k^T(t)b = -\Delta_k(t), \quad k = 2, \dots, N,$$

where  $\psi_k(t)$ ,  $t \in T$ , are the solutions of the following differential equations:

$$(34) \quad \frac{d\psi_k(t)}{dt} = -A^T \psi_k(t) - D^T \psi_{k+1}(t), \quad \psi_k(\alpha) = p_k - H_k^T \nu^k, \quad t \in T.$$

For each  $k = 1, \dots, N$  introduce the associated Hamilton function as

$$(35) \quad H_k(x_{k-1}, x_k, \psi_k, u_k) = \psi_k^T(Ax_k + Dx_{k-1} + bu_k), \quad t \in T.$$

Then the use of (33) yields that the optimality conditions (18) can be reformulated in maximum principle form as the following corollary to Theorem 1.

**COROLLARY 1.** *The admissible supporting control  $\{\tau_{sup}^k, u_k^0(t), k = 1, \dots, N\}$  is optimal if along the corresponding trajectories  $x_k^0(t)$ ,  $\psi_k(t)$  of (1)–(2) and (34) the Hamiltonian function has maximum value, i.e.,*

$$(36) \quad H_k(x_{k-1}^0(t), x_k^0(t), \psi_k, u_k^0(t)) = \max_{|v| \leq 1} H_k(x_{k-1}^0(t), x_k^0(t), \psi_k, v), \quad t \in T,$$

for  $k = 1, \dots, N$ . If the admissible supporting control is nondegenerate, then this condition is necessary and sufficient.

*Remark 4.* In order to further emphasize the relationship between the support elements and the control function, note that the optimality conditions given by Theorem 1 can be equivalently stated in the form

$$(37) \quad \begin{aligned} \Delta_k(t) &> 0 \text{ at } u_k^0(t) = -1, \quad \Delta_k(t) < 0 \text{ at } u_k^0(t) = 1, \\ \Delta_k(t) &= 0 \text{ at } -1 < u_k^0(t) < 1, \quad k = 1, 2, \dots, N, \quad t \in T. \end{aligned}$$

Hence the supporting elements and control function of optimal solution are interconnected such that the supporting instances are the switching moments for optimal bang-bang control functions.

In the next section, the maximum principle for arbitrary admissible control functions of (1)–(4) is established using the suboptimality conditions.

**3.1.  $\epsilon$ -optimality conditions.** Often in the numerical implementation of optimal control algorithms it is beneficial to exploit approximate solutions with corresponding error estimation. Hence it is necessary to introduce the “suboptimality” concept as it is often sufficient to stop the numerical computations when a satisfactory accuracy level has been achieved.

Assume that  $\{u_k^0(t), k \in K\}$  is the optimal control for (1)–(4), and let  $J(u^0)$  denote the corresponding optimal cost function value.

**DEFINITION 5.** We say that the admissible control function  $\{u_k^\epsilon(t), k \in K\}$  is  $\epsilon$ -optimal if the corresponding solution  $\{x_k^\epsilon(t), t \in T, k \in K\}$  of (1)–(4) satisfies  $J(u^0) - J(u^\epsilon) \leq \epsilon$ .

Now we proceed to calculate an estimate of a supporting control function

$$\{u_k, \tau_{sup}^k, k \in K, t \in T\},$$

i.e., a measure of the nonoptimality of the control. Note also that this estimate can be partitioned into two principal parts: one of which evaluates the degree of nonoptimality of the chosen admissible control functions  $u_k(t)$ , and the second the error produced by nonoptimality of the support  $\tau_{sup}^k$ . This partition is a major advantage in the design of numerically applicable solution algorithms.

Introduce an estimate of optimality  $\beta = \beta(\tau_{sup}, u)$  as the value of the maximum increment for the cost function here calculated in the absence of the principal constraints (4); i.e., this estimate is given by the solution of the following relaxed optimization problem:

$$(38) \quad \max_{\Delta u_k} \Delta J(u), \quad |u_k(t) + \Delta_k u(t)| \leq 1, \quad t \in T, \quad k = 1, \dots, N.$$

It is easy to see that

$$(39) \quad \beta = \beta(\tau_{sup}, u) = \sum_{k=1}^N \int_{T_k^+} \Delta_k(t)(u_k(t) + 1)dt + \sum_{k=1}^N \int_{T_k^-} \Delta_k(t)(u_k(t) - 1)dt,$$

where

$$T_k^+ = \{t \in T : \Delta_k(t) > 0\}, \quad T_k^- = \{t \in T : \Delta_k(t) < 0\},$$

and we have the following result.

THEOREM 2 ( $\epsilon$ -maximum principle). *Given any  $\epsilon \geq 0$ , the admissible control  $\{u_k(t), t \in T, k \in K\}$  is  $\epsilon$ -optimal for (1)–(4) if and only if there exists a support  $\{\tau_{sup}^k, k \in K\}$  such that along the solutions  $x_k(t), \psi_k(t), t \in T, k \in K$ , of (1)–(4) and (34) the Hamiltonian attains its  $\epsilon$ -maximum value, i.e.,*

$$H_k(x_{k-1}(t), x_k(t), \psi_k, u_k(t)) = \max_{|v| \leq 1} H_k(x_{k-1}(t), x_k(t), \psi_k, v) - \epsilon_k(t), \quad t \in T, \quad (40)$$

where the functions  $\epsilon_k(t)$ ,  $k \in K$ , satisfy the following inequality:

$$\sum_{k \in K} \int_T \epsilon_k(t) dt \leq \epsilon. \quad (41)$$

*Proof.* Assume that (40) and (41) hold for an admissible control  $\{u_k(t), t \in T, k \in K\}$ . Then by (33) the suboptimal estimate is

$$\begin{aligned} \beta &= \beta(\tau_{sup}, u) = \sum_{k=1}^N \int_{T_k^+} \psi_k^T(t) b(-u_k(t) - 1) dt + \sum_{k=1}^N \int_{T_k^-} \psi_k^T(t) b(1 - u_k(t)) dt \\ &= \sum_{k=1}^N \int_{T_k^+} \psi_k^T(t) (Ax_k(t) + Dx_{k-1}(t) - b) dt \\ &\quad - \sum_{k=1}^N \int_{T_k^+} \psi_k^T(t) (Ax_k(t) + Dx_{k-1}(t) + bu_k(t)) dt \\ &= \sum_{k=1}^N \int_{T_k^-} \psi_k^T(t) (Ax_k(t) + Dx_{k-1}(t) + b) dt \\ &\quad - \sum_{k=1}^N \int_{T_k^-} \psi_k^T(t) (Ax_k(t) + Dx_{k-1}(t) - bu_k(t)) dt \\ &\quad + \sum_{k=1}^N \int_T \left[ \max_{|v| \leq 1} H_k(x_{k-1}(t), x_k(t), \psi_k(t), v) - H_k(x_{k-1}(t), x_k(t), \psi_k(t), u_k(t)) \right] dt \\ &= \sum_{k=1}^N \int_T \epsilon_k(t) dt \leq \epsilon. \end{aligned}$$

Since the suboptimal estimate (38) has been calculated in the absence of constraints (4), then it is obvious that

$$J(u^0) - J(u) \leq \beta(\tau_{sup}, u) \leq \epsilon.$$

This proves the  $\epsilon$ -optimality property of the admissible control  $\{u_k(t), t \in T, k \in K\}$ .

For the converse argument, let  $\{u_k(t), t \in T, k \in K\}$  be an  $\epsilon$ -optimal admissible control, and let  $\{\tau_{sup}^k, k \in K\}$  be an arbitrary support. Then the suboptimal estimate of the control corresponding to the chosen support is given by

$$(42) \quad \beta(\tau_{sup}, u) = \sum_{k=1}^N \int_T \Delta_k(t) u_k(t) dt + \sum_{k=1}^N \int_{T_k^+} \Delta_k(t) dt - \sum_{k=1}^N \int_{T_k^-} \Delta_k(t) dt.$$

Also introduce the following dual optimization problem:

$$(43) \quad \min_{y,v,w} I(y,v,w), \quad I(y,v,w) = \sum_{k \in K} \left[ h_k^T y_k + \int_T v_k(t) dt + \int_T w_k(t) dt \right],$$

subject to

$$(44) \quad \sum_{s=k}^N y_s^T g_{sk}(t) - v_k(t) + w_k(t) = c_k(t), \quad v_k(t) \geq 0, \quad w_k(t) \geq 0, \quad t \in T, \quad k \in K.$$

At this stage we have to check that this dual optimization problem has a nonempty set of admissible variables  $z_k = \{y_k, v_k, w_k, k \in K\}$ . Suppose therefore that we denote the chosen support by  $\tau_{sup}^k, k \in K$ , and then use (18) to construct the vectors  $z_k = \{y_k, v_k, w_k, k \in K\}$  as

$$\begin{aligned} y_k &= \nu_k, \quad \nu_k(t) = \Delta_k(t); \quad w_k(t) = 0 \text{ if } \Delta_k(t) \geq 0, \\ v_k(t) &= 0, \quad w_k(t) = \Delta_k(t) \text{ if } \Delta_k(t) < 0. \end{aligned}$$

Then, by (18), these satisfy the constraint (44) of the dual problem. Also since this dual problem has a nonempty set of feasible variables

$$\{y_k, v_k, w_k, k \in K\},$$

it is routine to show that it has an optimal solution if there exists an optimal control for (1)–(4).

Let  $\{y_k^0, v_k^0(t), w_k^0(t), t \in T, k \in K\}$  denote an optimal solution of (43)–(44). Then (43) and (18) yield

$$\begin{aligned} \beta(\tau_{sup}, u) &= \sum_{k=1}^N \sum_{s=k}^N \int_T \nu_s^T(t) g_{sk}(t) u_k(t) dt - \sum_{k=1}^N \int_T c_k^T(t) u_k(t) dt \\ &\quad + \sum_{k=1}^N \int_T v_k(t) dt - \sum_{k=1}^N \int_T w_k(t) dt \\ &= \left[ \sum_{k=1}^N (\nu^k)^T \sum_{s=1}^k \int_T g_{ks}(t) u_s(t) dt + \sum_{k=1}^N \int_T v_k(t) dt - \sum_{k=1}^N \int_T w_k(t) dt \right] \\ &\quad - \left[ \sum_{k=1}^N \sum_{s=k}^N \int_T (y_s^0)^T g_{sk}(t) u_k^0(t) dt + \sum_{k=1}^N \int_T v_k^0(t) dt - \sum_{k=1}^N \int_T w_k^0(t) dt \right] \\ &\quad + \sum_{k=1}^N \int_T c_k(t) u_k^0(t) dt - \sum_{k=1}^N \int_T c_k(t) u_k(t) dt \\ &= \left[ \sum_{k=1}^N (\nu^k)^T h_k + \sum_{k=1}^N \int_T (v_k(t) - w_k(t)) dt \right] \\ &\quad - \left[ \sum_{k=1}^N (y_k^0)^T h_k + \sum_{k=1}^N \int_T (v_k^0(t) - w_k^0(t)) dt \right] \\ &\quad + \sum_{k=1}^N \int_T c_k(t) u_k^0(t) dt - \sum_{k=1}^N \int_T c_k(t) u_k(t) dt. \end{aligned}$$



Hence the suboptimal estimate can be written in the form

$$(45) \quad \beta(\tau_{sup}, u) = \beta_{sup} + \beta_u,$$

where

$$(46) \quad \beta_{sup} = \sum_{k=1}^N h_k^T(\nu_k - y^{0k}) + \sum_{k=1}^N \int_T \left[ (v_k(t) - v_k^0(t)) - (w_k(t) - w_k^0(t)) \right] dt$$

denotes the nonoptimality measure of the chosen support  $\{\tau_{sup}^k, k \in K\}$  and

$$(47) \quad \beta_u = \sum_{k=1}^N \int_T c_k(t)(u_k(t) - u_k^0(t)) dt$$

denotes the nonoptimality measure of the given control function  $\{u_k(t), t \in T, k \in K\}$ .

Now choose the support  $\tau_{sup}^0 = \{\tilde{\tau}_{sup}^k, k \in K\}$  such that the corresponding collection  $z_k^0 = \{y_k^0, v_k^0, w_k^0, k \in K\}$  of dual variables is an optimal solution of (43)–(44). Then the support  $\tau_{sup}^0 = \{\tilde{\tau}_{sup}^k(\epsilon), k \in K\}$  is the one required for the given  $\epsilon$ -optimal control functions  $\{u_k(t), k \in K\}$ , since  $\beta_{sup} = 0$ , and then  $\beta = \beta(u, \tau_{sup}^0) = \beta_u \leq \epsilon$ . Next set

$$\begin{aligned} \epsilon_k(t) &= \Delta_k(t)(u_k(t) + 1), \quad t \in T_k^+, \\ \epsilon_k(t) &= \Delta_k(t)(u_k(t) - 1), \quad t \in T_k^-, \\ \epsilon_k(t) &= 0 \quad \text{if } \Delta_k(t) = 0, \quad t \in T, \end{aligned}$$

and note from the definition of  $\Delta_k(t)$  that

$$\begin{aligned} \epsilon_k(t) &= -\psi_k^T(t)b(u_k(t) + 1) = \psi_k^T(t)(Ax_k(t) + Dx_{k-1}(t) + b(-1)) \\ &\quad - \psi_k^T(t)(Ax_k(t) + Dx_{k-1}(t) + bu_k(t)) \quad \text{if } \psi_k(t)b < 0, \\ \epsilon_k(t) &= \psi_k^T(t)(Ax_k(t) + Dx_{k-1}(t) + b(+1)) \\ &\quad - \psi_k^T(t)(Ax_k(t) + Dx_{k-1}(t) + bu_k(t)) \quad \text{if } \psi_k(t)b > 0, \\ \epsilon_k(t) &= 0 \quad \text{if } \psi_k(t)b = 0, \quad t \in T, \quad k \in K. \end{aligned}$$

Use of the Hamiltonian (35) now enables these last expressions to be written in the form

$$\epsilon_k(t) = \max_{|v| \leq 1} H_k(x_{k-1}(t), x_k(t), \psi_k, v) - H_k(x_{k-1}(t), x_k(t), \psi_k, u_k(t)), \quad t \in T, \quad k \in K.$$

Finally, noting that  $\{u_k(t)\}$  is a suboptimal control yields

$$\begin{aligned} \sum_{k=1}^N \int_T \epsilon_k(t) dt &= \sum_{k=1}^N \int_{T_k^+} \Delta_k(t)(u_k(t) + 1) dt \\ &\quad + \sum_{k=1}^N \int_{T_k^-} \Delta_k(t)(u_k(t) - 1) dt = \beta(u, \tau_{sup}^0) = \beta_u \leq \epsilon, \end{aligned}$$

and the proof is complete.  $\square$

The maximum principle now follows from this last result on setting  $\epsilon = 0$  as stated formally in the following corollary.

**COROLLARY 2.** *The admissible control  $\{u_k^0(t), k \in K, t \in T\}$  is optimal if and only if there exists a support  $\{\tau_{sup}^{0k}, k \in K\}$  such that the supporting control  $\{u_k^0(t), \tau_{sup}^{0k}, t \in T, k \in K\}$  satisfies the maximum conditions*

$$\max_{|v| \leq 1} H_k(x_{k-1}^0(t), x_k^0(t), \psi_k, v) = H_k(x_{k-1}^0(t), x_k^0(t), \psi_k, u_k^0(t))$$

for all  $k \in K, t \in T$ , where  $\psi_k(t)$  are the corresponding solutions of (34).

**4. Differentiable properties of the optimal solutions.** An important aspect of the optimization theory is sensitivity analysis of optimal controls since, in practice, the system considered can be subject to disturbances or parameters in the available model can easily arise. Mathematically, perturbations can, for example, be described by some parameters in the initial data, boundary conditions, and control and state constraints. Hence it is clearly important to know how a problem solution depends on these parameters, and in this section we aim to characterize the changes in the solutions developed due to “small” perturbations in the parameters. This could, in turn, enable us to design fast and reliable real-time algorithms to correct the solutions for these effects. As shown next, the major advantage of the constructive approach developed in this paper is that the sensitivity analysis and some differential properties of the optimal controls under disturbances can be analyzed.

Suppose that disturbances influence the initial data for (1)–(4). In particular, consider the system (1)–(4) on the interval  $T_s = [s, \alpha]$  with the initial data  $x_k(s) = z_k, z_k \in G_k, k \in K$ , where  $G_k \subset \mathbb{R}^n$  is some neighborhood of the point  $x_k = d_k$  and  $s$  belongs to the neighborhood  $G_0$  of  $t = 0$ . We also assume that the following regularity condition holds: For the given disturbance domain  $G_k, k \in K \cup \{0\}$ , the structure of the optimal control functions for the nondisturbed data is preserved; i.e., the number of switching instances together with their order is constant.

Using Theorem 1, the optimal controls  $\{u_k^0(t, s, z), k \in K\}$  are determined by the supporting time instances  $\tau_{kj} = \tau_{kj}(s, z), k \in K, j = 1, \dots, m$ , which are dependent on the disturbances  $(s, z_k), s \in G_0, z_k \in G_k, k \in K$ . Here we study the differential properties of the functions  $\tau_{kj} = \tau_{kj}(s, z), k \in K, j = 1, \dots, m$ , and for ease of notation we set  $\tau \equiv \tau(s, z) = \{\tau_{kj}(s, z), k \in K, j = 1, \dots, m\}, z = \{z_k, k \in K\}$  in what follows.

**THEOREM 3.** *If (1)–(4) is regular, then for any  $k \in K$  and  $j = 1, \dots, m$  the functions  $\tau_{kj} = \tau_{kj}(s, z)$  are differentiable in the domain  $G_0 \times G_k \subset \mathbb{R} \times \mathbb{R}^n$ .*

*Proof.* Using (10)–(11) and Theorem 1 it follows immediately that the switching instances  $\tau_{kj} = \tau_{kj}(s, z), k \in K, j = 1, \dots, m$ , of the optimal bang-bang control  $\{u_k^0(t, s, z), k \in K\}$  for (1)–(4) in this case are the solutions of the following optimization problem:

$$(48) \quad \max_{\tau_{kj}} \sum_{k \in K} R_k(s, z) \sum_{j=1}^{m+1} (-1)^j \int_{\tau_{kj-1}}^{\tau_{kj}} c_k(t) dt,$$

subject to

$$(49) \quad \sum_{l \in K} R_l(s, z) \sum_{j=1}^{m+1} (-1)^j \int_{\tau_{lj-1}}^{\tau_{lj}} g_{kl}(t) dt = h_k(s, z), \quad k \in K.$$

Here the constant  $R_k(s, z) = \pm 1$  denotes the value ( $u = +1$  or  $u = -1$ ) of the optimal control on pass  $k$  over the first control interval  $t \in [s, \tau_{k1}]$ , and

$$(50) \quad h_k(s, z) = o_k - \sum_{j=1}^k H_k K_j(\alpha) z_{k+1-j} - \int_s^\alpha H_k K_k(\alpha - t) Df(t) dt.$$

Also it is clear that the switching instances  $\tau_{kj} = \tau_{kj}(s, z)$  satisfy

$$\tau_{k0} < \tau_{k1} < \tau_{k2} < \cdots < \tau_{km} < \tau_{km+1}, \quad \tau_{k0} = s, \quad \tau_{km+1} = \alpha.$$

Since  $\{u_k^0, \tau_{sup}^0, k \in K\}$  is the optimal supporting control for (1)–(4) in the absence of disturbances, the optimization problem (48)–(49) has the optimal solution  $\tau_{kj}^0, k \in K, j = 1, \dots, m$  at  $s = 0, z_k = \alpha_k, k \in K, j = 1, \dots, m$ . Hence there exist Lagrange multipliers  $\lambda_k^0 \in \mathbb{R}^m, k \in K$ , which are not simultaneously equal to zero, such that the collection  $\{\lambda_k^0, \tau_{kj}^0\}$  is a stationary point for the following Lagrange function associated with the optimization problem (48)–(49):

$$(51) \quad \begin{aligned} L(\lambda, \tau_{sup}) = & \sum_{k \in K} R_k(s, z) \sum_{j=1}^{m+1} (-1)^j \int_{\tau_{kj-1}}^{\tau_{kj}} c_k(t) dt \\ & + \sum_{k \in K} \lambda_k \left[ \sum_{l \in K} R_l(s, z) \sum_{j=1}^{m+1} (-1)^j \int_{\tau_{lj-1}}^{\tau_{lj}} g_{kl}(t) dt - h_k(s, z) \right]. \end{aligned}$$

The well-known stationarity conditions for a Lagrange function now yield

$$(52) \quad 2R_k(s, z) \left[ c_k(\tau_{kj}) + \sum_{l=k}^N \lambda_l g_{lk}(\tau_{kj}) \right] = 0, \quad j = 1, \dots, m, \quad k \in K,$$

$$(53) \quad \sum_{l=1}^k R_l(s, z) \sum_{j=1}^{m+1} (-1)^j \int_{\tau_{lj-1}}^{\tau_{lj}} g_{kl}(t) dt - h_k(s, z) = 0, \quad k \in K,$$

with respect to the unknown  $\lambda_k$  and  $\tau_k(s, z)$ ,  $k \in K, j = 1, \dots, m$ . Also the Jacobian matrix  $D$  of the mapping (52) with respect to variables  $(\lambda, \tau_{sup})$  calculated at  $s = 0$  and  $z_k = \alpha_k$  can be written in the form

$$(54) \quad D = \prod_{k \in K} 2R_k(0, \alpha) \begin{pmatrix} \hat{G}_{sup} & F \\ 0 & \hat{G}_{sup} \end{pmatrix},$$

where (see also (15) for the notation here) the matrix  $\hat{G}_{sup}$  is given by

$$(55) \quad \hat{G}_{sup} = \begin{pmatrix} g_{kj}(t), & t \in \tau_{sup}^j \\ j \leq k \leq N, & j = 1, \dots, N \end{pmatrix},$$

and the matrix  $F$  is formed from the derivatives of the functions  $c_k(t), g_{kl}(t)$  evaluated at the corresponding points. By the definition of the supporting time instances we have  $\det D \neq 0$ , and by the implicit function theorem there exists a neighborhood of the point  $(0, \alpha_k, k \in K)$  where (52) has a unique solution  $\lambda = \lambda(s, z)$ ,  $\tau_{kj} = \tau_{kj}(s, z)$ , and these functions are also differentiable. This completes the proof.  $\square$

The differential properties of the optimal controls developed above can be used for sensitivity analysis and the solution of the synthesis problem considered here. In

particular, the supporting control approach [10] can be used to produce the differential equations for the switching time functions  $\tau(s, z)$  necessary to design the optimal controllers. In a similar manner to [6] it can be shown that these satisfy the following differential equations:

$$(56) \quad G \frac{\partial \tau}{\partial s} + Q = \frac{\partial h}{\partial s}, \quad P \frac{\partial \tau}{\partial z} = \frac{\partial h}{\partial z},$$

where  $h(s, z) = (h_1(s, z), \dots, h_m(t, s))$  is an  $mN \times 1$ -vector given by (50) and the matrices  $G, Q, P$  are defined (see [6]) by those defining the process dynamics and information associated with the disturbance-free optimal solution. For example (see also (13)),  $G = \Lambda \tilde{G}_{sup}$ , where the compatibly dimensioned block matrix  $\Lambda$  is constructed by the disturbance-free optimal control values  $u_k^0(t); k = 1, \dots, N$  calculated in the supporting moments  $\tau_{kj}$  from  $\tau_{sup}^0$ . Also, by Theorem 1, these values are equivalent to the values of  $\frac{d\Delta_i(\tau_{kj})}{dt}$  evaluated for the corresponding indexes  $i; j; k$ , where the functions  $\Delta_i(t); i = 1, \dots, N$  are designed using the switching times of the basic optimal control function. Note also that analogous differential equations can be established for the optimal values of the cost function (treated as the function  $J(s, z) \equiv J(u(\tau(s, z)))$ ).

*Remark 5.* The equations (56) are (sometimes) termed Pfaff differential equations and model an essentially distinct class of continuous  $nD$  systems. The main characteristic feature of this model is that it is overdetermined (the number of equations exceeds the unknown functions). It can also be shown that if the nondegenerate assumption on the supporting control functions holds, then so do the so-called Frobenius conditions which guarantee the existence and uniqueness of solutions of the Pfaff differential equations.

**5. An example.** In order to demonstrate the advantages of the supporting control function approach, we give an example where, as a preliminary, it is instructive to consider the case of  $N = 1$  and, in particular,

$$(57) \quad \max_{|u| \leq 1} J(u), \quad J(u) := x^{(2)}(1),$$

for

$$(58) \quad \begin{aligned} \frac{dx^{(1)}(t)}{dt} &= x^{(2)}, & x^{(1)}(t), x^{(2)}(t) &\in \mathbb{R}, & t &\in [s, 1], \\ \frac{dx^{(2)}(t)}{dt} &= u(t), & x^{(1)}(s) &= z_1, & x^{(2)}(s) &= z_2, \end{aligned}$$

subject to the following constraints on the control signal and a terminal state constraint, respectively:

$$(59) \quad |u(t)| \leq 1, \quad x^{(1)}(1) = 1/8.$$

Note that here the superscript  $(\cdot)$  is used to denote a particular element in the state vector.

In this case it is easy to verify that for  $s = 0$  and  $x^{(1)}(0) = 0$ ,  $x^{(2)}(0) = 0$  the optimal control signal is given by

$$u^0(t) = -1 \quad \text{for} \quad 0 \leq t \leq 1 - \sqrt{5/8} \quad \text{and} \quad u^0(t) = +1 \quad \text{for} \quad 1 - \sqrt{5/8} < t \leq 1.$$

Synthesis of the optimal control can be realized using the switching instance function  $\tau = \tau(z_1, z_2, s)$ , which has to satisfy the following differential equations:

$$(60) \quad \frac{\partial \tau}{\partial z_1} = \frac{1}{2(1-\tau)},$$

$$\frac{\partial \tau}{\partial z_2} = \frac{1-s}{2(1-\tau)},$$

$$(61) \quad \frac{\partial \tau}{\partial s} = \frac{1-s-z_2}{2(1-\tau)},$$

with the initial condition

$$\tau(0, 0, 0) = 1 - \sqrt{5/8},$$

which is a particular case of (56).

The solution of this Pfaff differential system is given by

$$\tau(z_1, z_2, s) = 1 - \sqrt{5/8 + (s-1)z_2 - z_1 - s + s^2/2}.$$

Also, without loss of generality, assume  $s = 0$ , and then the optimal switching function is

$$\tau(z_1, z_2, 0) = 1 - \sqrt{5/8 - z_1 - z_2}.$$

Figures 1 and 2 illustrate the form of this solution. In particular, Figure 1 shows the state-space variables together with additional variable  $t$ . The optimal trajectories (57)–(59) corresponding to the bang-bang control law lie on the parabolic cylinders  $(Z_1) : x^{(1)} = -\frac{1}{2}(x^{(2)})^2 + C_1 + C_2$  and  $(Z_2) : x^{(1)} = +\frac{1}{2}(x^{(2)})^2 + \tilde{C}_1 + \tilde{C}_2$ , where the constants  $C_i, \tilde{C}_i$ ,  $i = 1, 2$ , are determined by the initial data  $x^{(1)}(0) = z_1$ ,  $x^{(2)}(0) = z_2$ . These cylinders correspond to the solutions of differential equations (58) with  $u \equiv -1$  or  $u \equiv +1$ , respectively. It can also be shown that the admissible initial domain for which the problem can be solved is determined by the inequalities:  $-\frac{3}{8} \leq z_1 + z_2 \leq \frac{5}{8}$ . The switching manifold  $Z_h$  is described in parametric form by

$$\begin{aligned} x^{(1)} &= -\frac{(1 - \sqrt{5/8 - z_2 - z_1})^2}{2} + z_2(1 - \sqrt{5/8 - z_2 - z_1}) + z_1, \\ x^{(2)} &= -1 + \sqrt{5/8 - z_2 - z_1} + z_2, \\ T &= 1 - \sqrt{5/8 - z_2 - z_1} - \frac{3}{8} \leq z_1 + z_2 \leq \frac{5}{8}. \end{aligned}$$

Finally, each optimal trajectory consists of two parts—first it evolves along the vertical parabolic cylinder  $Z_1$  until  $\tau = 1 - \sqrt{5/8 - z_2 - z_1}$ , when it meets the switching manifold  $Z_h$ , and then immediately is switched to continue along the second vertical cylinder  $Z_2$  to meet the target plane  $x^{(1)} = 1/8$ . Figure 1 also shows the optimal trajectory in the space  $R^3$  for zero initial data, and Figure 2 shows the projection of this trajectory onto the  $x^{(1)}, x^{(2)}$  plane.

Consider now the following example where  $N = 2$  (again the superscript  $(\cdot)$  is used to denote a particular element in the state or control vector on any pass):

$$(62) \quad \max_{u_1, u_2} J(u), \quad J(u) := x_1^{(2)}(1) + x_2^{(2)}(1),$$

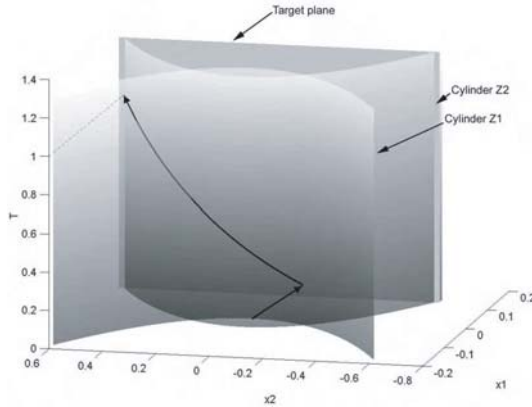


FIG. 1. *Optimal control synthesis.*

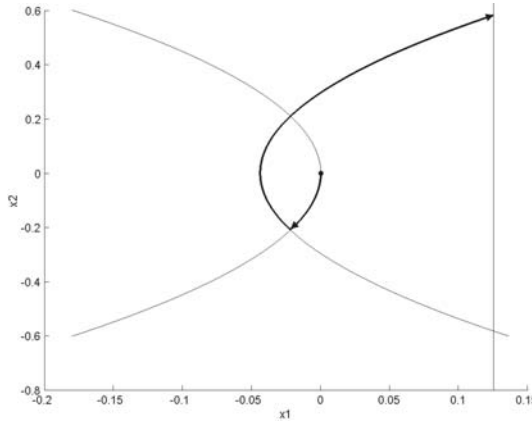


FIG. 2. *Projection on the  $x_1$ - $x_2$  plane.*

for the process

$$(63) \quad \begin{aligned} \frac{dx_1^{(1)}(t)}{dt} &= x_1^{(2)}(t), & \frac{dx_2^{(1)}(t)}{dt} &= x_2^{(2)}(t), \quad t \in [s, 1], \\ \frac{dx_1^{(2)}(t)}{dt} &= u_1(t), & \frac{dx_2^{(2)}(t)}{dt} &= x_1^{(1)}(t) + u_2(t), \end{aligned}$$

with boundary conditions of the form

$$(64) \quad x_1^{(1)}(s) = z_1^{(1)}, \quad x_1^{(2)}(s) = z_1^{(2)}, \quad x_2^{(1)}(s) = z_2^{(1)}, \quad x_2^{(2)}(s) = z_2^{(2)},$$

subject to

$$(65) \quad x_1^{(1)}(1) = 1/8, \quad x_2^{(1)}(1) = 1/384, \quad |u_1(t)| \leq 1, \quad |u_2(t)| \leq 1.$$

Equivalently, we can write the problem here as

$$(66) \quad \begin{bmatrix} \dot{x}_{k+1}^{(1)}(t) \\ \dot{x}_{k+1}^{(2)}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{k+1}^{(1)}(t) \\ x_{k+1}^{(2)}(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_k^{(1)}(t) \\ x_k^{(2)}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_{k+1}(t), \quad k = 0, 1.$$

Without loss of generality we set  $x_0(t) = 0$ ,  $t \in [s, 1]$ .

To apply the results developed here to this example we first rewrite (63)–(65) in the following integral form:

$$\max_{u_1, u_2} \left\{ z_2^{(1)} + z_2^{(2)} + (1-s)z_1^{(1)} + \frac{(1-s)^2}{2}z_2^{(2)} + \int_s^1 \frac{(1-t)^2 + 2}{2}u_1(t)dt + \int_s^1 u_2(t)dt \right\}, \quad (67)$$

subject to

$$\begin{aligned} \int_s^1 (1-t)u_1(t)dt &= \frac{1}{8} - z_1^{(1)} + (1-s)z_1^{(2)}, \\ \int_s^1 \left[ \frac{(1-t)^3}{6}u_1(t) + (1-t)u_2(t) \right] dt &= \frac{1}{384} - z_2^{(1)} - (1-s)z_2^{(2)} - \frac{(1-s)^2}{2}z_1^{(1)} \\ &\quad - \frac{(1-s)^3}{6}z_1^{(2)}. \end{aligned} \quad (68)$$

Hence

$$\begin{aligned} g_{11}(t) &= 1-t, \quad g_{21}(t) = \frac{(1-t)^3}{6}, \quad g_{22}(t) = 1-t, \\ c_1(t) &= \frac{(1-t)^2 + 2}{2}, \quad c_2(t) = 1, \end{aligned} \quad (69)$$

and the multipliers required to design the cocontrol function  $\Delta_i(t)$ ,  $i = 1, 2$ , can, by noting (12), be written as

$$(70) \quad \nu^{(2)}g_{22}(\tau_{2sup}) - c_2(\tau_{2sup}) = 0, \quad \nu^{(1)}g_{11}(\tau_{1sup}) + \nu^{(2)}g_{21}(\tau_{1sup}) - c_1(\tau_{1sup}) = 0.$$

We now have that

$$\begin{aligned} \Delta_1(t) &= (1-t) \left[ \frac{1}{1-\tau_{1sup}} + \frac{1-\tau_{1sup}}{2} - \frac{(1-\tau_{1sup})^2}{6(1-\tau_{2sup})} \right] + \frac{(1-t)^3}{6(1-\tau_{2sup})} - \frac{(1-t)^2}{2} - 1, \\ \Delta_2(t) &= \frac{1-t}{1-\tau_{2sup}} - 1, \end{aligned} \quad (71)$$

and the problem is to find the basic optimal trajectory when all variables in (64) are zero, i.e.,

$$(72) \quad s = 0, \quad x_1^{(1)}(0) = 0, \quad x_1^{(2)}(0) = 0, \quad x_2^{(1)}(0) = 0, \quad x_2^{(2)}(0) = 0.$$

Take the supporting instances as

$$(73) \quad \tau_{1sup} = 1 - \sqrt{\frac{5}{8}}, \quad \tau_{2sup} = 1 - \sqrt{\frac{131}{256}}.$$

Then it follows immediately from Theorem 1 that the optimal control functions for (62)–(65) with initial data (72) are given by

$$\begin{aligned} u_1^0(t) &= \begin{cases} -1, & 0 \leq t < 1 - \sqrt{\frac{5}{8}}, \\ +1, & 1 - \sqrt{\frac{5}{8}} \leq t \leq 1, \end{cases} \\ u_2^0(t) &= \begin{cases} -1, & 0 \leq t < 1 - \sqrt{\frac{131}{256}}, \\ +1, & 1 - \sqrt{\frac{131}{256}} \leq t \leq 1, \end{cases} \end{aligned} \quad (74)$$

and (56) gives the switching functions  $\tau_1 \equiv \tau_1(z_1^{(1)}, z_1^{(2)}, s)$ ,  $\tau_2 \equiv \tau_2(z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s)$  as

$$\begin{aligned}
 & -2 \frac{\partial \tau_2}{\partial s} (1 - \tau_2) - \frac{2(1 - \tau_1)^3}{6} \frac{\partial \tau_1}{\partial s} = \frac{(1 - s)^2}{2} z_1^{(2)} + (1 - s) z_1^{(1)} \\
 & \quad + z_2^{(2)} - \frac{(1 - s)^3}{6} - (1 - s) \\
 & \quad - 2 \frac{\partial \tau_2}{\partial z_1^{(1)}} (1 - \tau_2) - \frac{(1 - \tau_1)^3}{3} \frac{\partial \tau_1}{\partial z_1^{(1)}} \\
 & = -\frac{(1 - s)^2}{2} - 2 \frac{\partial \tau_2}{\partial z_1^{(2)}} (1 - \tau_2) - \frac{(1 - \tau_1)^3}{3} \frac{\partial \tau_1}{\partial z_1^{(2)}} \\
 & = -\frac{(1 - s)^3}{6} - 2 \frac{\partial \tau_2}{\partial z_2^{(1)}} (1 - \tau_2) = -1, \\
 (75) \quad & -2 \frac{\partial \tau_2}{\partial z_2^{(2)}} (1 - \tau_2) = -(1 - s),
 \end{aligned}$$

with initial conditions

$$(76) \quad \tau_1(0, 0, 0) = 1 - \sqrt{\frac{5}{8}}, \quad \tau_2(0, 0, 0, 0, 0) = 1 - \sqrt{\frac{131}{16^2}}.$$

The solutions of this differential system are

$$\begin{aligned}
 & \tau_1(z_1^{(1)}, z_1^{(2)}, s) = 1 - \sqrt{SR_1(z_1^{(1)}, z_1^{(2)}, s)}, \\
 (77) \quad & \tau^{(2)}(z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s) = 1 - \sqrt{SR_2(z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s)},
 \end{aligned}$$

where

$$\begin{aligned}
 & SR_1(z_1^{(1)}, z_1^{(2)}, s) = \frac{5}{8} + (s - 1)z_1^{(2)} - z_1^{(1)} - s + s^2/2, \\
 & SR_2(z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s) = \frac{131}{256} + \frac{2s^4 - 8s^3 + 59s^2 - 102s}{96} + \frac{-20s^2 + 40s - 19}{48} z_1^{(1)} \\
 & \quad - \frac{1}{12} z_1^{(1)2} + \frac{4s^3 - 12s^2 + 11s - 3}{48} z_1^{(2)} + \frac{-s^2 + 2s - 1}{12} z_1^{(2)2} \\
 (78) \quad & \quad + \frac{s z_1^{(1)} z_1^{(2)}}{6} - \frac{z_1^{(1)} z_1^{(2)}}{6} - z_1^{(2)} + (s - 1) z_2^{(2)}.
 \end{aligned}$$

It easy to see that the solution of the differential equations describing the process dynamics with both  $u_1$  and  $u_2$  constant is

$$\begin{aligned}
 & x_1^{(1)}(t) = u_1 \frac{t^2}{2} + tC_1 + C_2, \\
 & x_1^{(2)}(t) = u_1 t + C_1, \\
 & x_2^{(1)}(t) = u_1 \frac{t^4}{24} + C_1 \frac{t^3}{6} + C_2 \frac{t^2}{2} + u_2 \frac{t^2}{2} + tC_3 + C_4, \\
 (79) \quad & x_2^{(2)}(t) = u_1 \frac{t^3}{6} + C_1 \frac{t^2}{2} + tC_2 + tu_2 + C_3
 \end{aligned}$$

and in this case that the optimal control for pass  $k = 1$  coincides with that given earlier in this example.



Now consider disturbances  $\Omega$  such that the optimal control is preserved for the case of zero initial conditions, i.e.,  $u_1 = -1$  for  $t \leq \tau_1(z_1^{(1)}, z_1^{(2)}, s)$ ,  $u_2^0 = -1$  for  $t \leq \tau_2(z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s)$ , and the inequality  $\tau_1(z_1^{(1)}, z_1^{(2)}, s) < \tau_2(z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s)$  holds. Using (77) we have that the domain  $\Omega$  is described by

$$0 \leq \tau_1(z_1^{(1)}, z_1^{(2)}, s) < \tau_2(z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s) \leq 1,$$

$$SR_1(z_1^{(1)}, z_1^{(2)}, s) \geq 0, \quad SR_2(z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s) \geq 0.$$

To construct the solution for pass  $k = 2$ , it is necessary to construct the switching surface  $F$ , which is defined by the vectors

$$x_2^{(1)}(t) \big|_{t=\tau_2} = x_2^{(1)}(\tau_2(z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s), \tau)$$

$$x_2^{(2)}(t) \big|_{t=\tau_2} = x_2^{(2)}(\tau_2(z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s)),$$

when the parameters  $z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s$  are members of the set  $\Omega$ . The parametric description of the switching surface in this case is given by

$$x_2^{(1)}(t) = -\frac{t^4}{24} + C_1 \frac{t^3}{6} + C_2 \frac{t^2}{2} - \frac{t^2}{2} + tC_3 + C_4,$$

$$(80) \quad x_2^{(2)}(t) = -\frac{t^3}{6} + C_1 \frac{t^2}{2} + tC_2 - t + C_3,$$

where the coefficients  $C_i$  are found from the parameters  $z_1^{(1)}, z_2^{(1)}, z_1^{(2)}, z_2^{(2)}, s$ .

**6. Conclusions.** In this paper the supporting control functions approach has been applied to study an optimal control problem for differential linear repetitive processes. The main contribution is the development of constructive necessary and sufficient optimality conditions in forms which can be effectively used for the design of numerical algorithms. The iterative method developed in this work is based on the principle of decrease of the suboptimality estimate; i.e., the iteration  $\{\tau_{sup}^k, u_k(t), k = 1, \dots, N\} \rightarrow \{\hat{\tau}_{sup}^k, \hat{u}_k(t), k = 1, \dots, N\}$  is performed in such a way as to achieve  $\beta(\hat{\tau}_{sup}, \hat{u}) < \beta(\tau_{sup}, u)$ . Also this procedure can be separated into two stages: (i) transformation of the admissible control functions  $\{u_k(t), k = 1, \dots, N\} \rightarrow \{\hat{u}_k(t), k = 1, \dots, N\}$ , which decreases the nonoptimality measure of the admissible controls  $\beta(\hat{u}) < \beta(u)$ , and (ii) variation of the support  $\{\tau_{sup}^k, k = 1, \dots, N\} \rightarrow \{\hat{\tau}_{sup}^k, k = 1, \dots, N\}$  to again decrease the nonoptimality measure of the support, i.e.,  $\beta(\hat{\tau}_{sup}) < \beta(\tau_{sup})$ . These transformations involve, in effect, the duality theory for the problems defined in this work by (1)–(4) and (43)–(44) and exploit the  $\epsilon$ -optimality conditions also developed in this work. These results are the first in this general area, and work is currently proceeding in a number of followup areas. One such area is sensitivity analysis of optimal control in the presence of disturbances, where in the case of the ordinary linear control systems some work on this topic can be found in, for example, [12].

## REFERENCES

- [1] N. AMANN, D. H. OWENS, AND E. ROGERS, *Iterative learning control using optimal feedforward and feedback actions*, Internat. J. Control, 69 (1996), pp. 277–293.
- [2] T. AL-TOWAIM, A. D. BARTON, P. L. LEWIN, E. ROGERS, AND D. H. OWENS, *Iterative learning control — 2D systems from theory to application*, Internat. J. Control, 77 (2004), pp. 877–893.

- [3] S. ARIMOTO, S. KAWAMURA, AND F. MIYAZAKI, *Bettering operations of robots by learning*, J. Robotic Systems, 1 (1984), pp. 123–140.
- [4] M. BISIACCO AND E. FORNASINI, *Optimal control of two-dimensional systems*, SIAM J. Control Optim., 28 (1990), pp. 582–601.
- [5] W. D. COLLINS, *Controllability and canonical forms for multipass systems described by the ordinary differential equations*, IMA J. Math. Control Inform., 1 (1984), pp. 1–25.
- [6] M. P. DYMKOV AND S. GNEVKO, *Continuous linear programming with disturbed parameters*, Reports of the Academy Sciences of Belarus, Series Physics and Mathematics, 4 (1983), pp. 8–14 (in Russian).
- [7] C. DU AND L. XIE,  *$H_\infty$  Control and Filtering of Two-Dimensional Systems*, Lecture Notes in Control and Inform. Sci. 278, Springer-Verlag, Berlin, 2002.
- [8] E. FORNASINI AND G. MARCHESINI, *Doubly indexed dynamical systems: State-space models and structural properties*, Math. Systems Theory, 12 (1978), pp. 59–72.
- [9] K. FUJIMOTO, T. HORIUCHI, AND T. SUGIE, *Optimal control of Hamiltonian systems with input constraints via iterative learning*, in Proceedings of the 42nd IEEE Conference on Decision and Control, 2003, pp. 4387–4392.
- [10] R. GABASOV AND F. M. KIRILLOVA, *Software Optimization*, Plenum Press, New York, 1988.
- [11] R. GABASOV, F. M. KIRILLOVA, AND S. V. PRISCHEPOVA, *Optimal Feedback Control*, Lecture Notes in Control and Inform. Sci. 207, Springer-Verlag, Berlin, 1995.
- [12] O. I. KOSTYUKOVA, *Parametric optimal control problem with variable index*, Comput. Math. Math. Phys., 43 (2003), pp. 26–41; (Zh. Vychis. Mat. i Mat. Fiz., 2003, 43(1), pp. 26–41, in Russian).
- [13] K. L. MOORE, Y. CHEN, AND V. BAHL, *Monotonically convergent iterative learning control for linear discrete-time systems*, Automatica, 41 (2005), pp. 1529–1537.
- [14] D. H. OWENS, N. AMANN, E. ROGERS, AND M. FRENCH, *Analysis of linear iterative learning control schemes - A 2D systems/repetitive processes approach*, Multidimens. Syst. Signal Process., 11 (2000), pp. 125–177.
- [15] P. D. ROBERTS, *Numerical investigations of a stability theorem arising from 2-dimensional analysis of an iterative optimal control algorithm*, Multidimens. Syst. Signal Process., 11 (2000), pp. 109–124.
- [16] R. P. ROESSER, *A discrete state space model for linear image processing*, IEEE Trans. Automat. Control, 20 (1975), pp. 1–10.
- [17] E. ROGERS AND D. H. OWENS, *Stability Analysis for Linear Repetitive Processes*, Lecture Notes in Control and Inform. Sci. 175, Springer-Verlag, Berlin, 1992.
- [18] M. SEBEK AND F. J. KRAUS, *Stochastic LQ-optimal control for 2-D systems*, Multidimens. Syst. Signal Process., 6 (1995), pp. 275–285.
- [19] H. D. TUAN, P. APKARIAN, T. Q. NGUYEN, AND T. NARIKIYO, *Robust Mixed  $H_2/H_\infty$  filtering of 2-D systems*, IEEE Trans. Signal Process., 50 (2002), pp. 1759–1771.
- [20] J.-X. XU, Y. TAN, AND T.-H. LEE, *Iterative learning control design based on composite energy function with input saturation*, Automatica, 40 (2004), pp. 1371–1377.

## CONNECTIONS BETWEEN SINGULAR CONTROL AND OPTIMAL SWITCHING\*

XIN GUO<sup>†</sup> AND PASCAL TOMECEK<sup>‡</sup>

**Abstract.** This paper builds a new theoretical connection between singular control of finite variation and optimal switching problems. This correspondence provides a novel method for solving high-dimensional singular control problems and enables us to extend the theory of reversible investment: Sufficient conditions are derived for the existence of optimal controls and for the regularity of value functions. Consequently, our regularity result links singular controls and Dynkin games through sequential optimal stopping problems.

**Key words.** singular stochastic control, optimal switching, Dynkin games, reversible investment

**AMS subject classifications.** 93E20, 49N60, 91A15, 91A55

**DOI.** 10.1137/060669024

**1. Introduction with a motivating control problem.** Let  $(\Omega, \mathcal{F}, P)$  be a complete probability space and  $\mathbb{F} = \{\mathcal{F}_t; 0 \leq t < \infty\}$  a completed filtration that is right continuous. Consider the following (motivating) singular control problem from [33]:

(1)

$$V(x, y) = \sup_{(\xi^+, \xi^-) \in \mathcal{A}} E \left[ \int_0^\infty e^{-rt} h(X_t, Y_t) dt - K^+ \int_0^\infty e^{-rt} d\xi_t^+ - K^- \int_0^\infty e^{-rt} d\xi_t^- \right],$$

where  $Y_t = y + \xi_t^+ - \xi_t^-$ ,  $(\xi_t^+, \xi_t^-)_{t \geq 0}$  is a pair of  $\mathbb{F}$ -adapted, nondecreasing càglàd processes,  $X_t$  is a diffusion process with  $X_0 = x$ ,  $h(X_t, Y_t)$  is a concave function of  $Y_t$  satisfying appropriate integrability conditions, and  $K^+, K^-, r > 0$  are some constants. Here the supremum is taken over the set  $\mathcal{A}$  of all singular controls with a finite variation.

This multidimensional control problem and its variants have been studied extensively in both the mathematics and the economics literature. For example, taking  $h$  as a concave function with a special additive form  $h(X_t + Y_t)$  and  $K^- + K^+ \geq 0$ , this is the well-known monotone fuel follower problem, for which explicit solutions can be found in [4, 5, 6, 25, 24]. In mathematical economics, (1) is a typical (ir)reversible investment problem in which a company, by adjusting its production capacity through expansion and contraction according to market fluctuations, wishes to maximize its overall expected net profit over an infinite horizon. Under the special additive form (again) of  $h(X_t + Y_t)$ , with  $X_t + Y_t = y + \mu t + \sigma W_t + \xi_t^+$  and  $\xi_t^- = K^- = 0$ , this problem has been investigated by numerous authors (see, for instance, [14, 31, 1, 2, 34, 36, 13, 3, 21]). With another special form of  $h(1 - Y_t + X_t Y_t)$ , where  $h$  is a power function,  $K^- = K^+ = 0$ , and  $Y_t \in [0, 1]$ , the problem was analyzed

\*Received by the editors September 4, 2006; accepted for publication (in revised form) July 18, 2007; published electronically February 1, 2008.

<http://www.siam.org/journals/sicon/47-1/66902.html>

<sup>†</sup>Department of Industrial Engineering and Operations Research, UC Berkeley, Berkeley, CA 94720-1777 (xinguo@ieor.berkeley.edu). This author is on leave from Cornell University.

<sup>‡</sup>School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853-3801 (pascal@orie.cornell.edu).

via a dimension reduction technique in [35]. For a standard reference on irreversible investment, see [15].

Most recently, [33] treated this problem with a more general and genuinely high-dimensional form, where  $X_t$  is a geometric Brownian motion and  $h$  is a function of both  $X_t$  and  $Y_t$ , subject to some technical conditions. They used the traditional approach of the dynamic programming principle: First construct (by ad hoc methods) a solution to the Bellman equation and then validate the optimality of the solution by a sufficient verification theorem for smooth functions.

In this paper, we establish a generic theoretical connection between singular control and optimal switching problems: We define a consistency property for collections of switching controls and prove that there is an exact correspondence between the set of finite variation càglàd processes and the set of consistent collections of switching controls. We then apply this correspondence, in conjunction with direct integration arguments, to obtain an integral representation for the value function of a very general reversible investment problem in terms of the values of corresponding optimal switching problems. Finally, we exploit this representation to study the regularity of the value function and to obtain sufficient conditions on the existence of optimal controls. As a corollary, we are able to represent the value of a Dynkin game as the difference between the values of two related switching problems, thereby linking the general reversible investment problem, the Dynkin game, and the optimal switching problem.

It is worth pointing out that this approach of connecting singular control problems and related optimal stopping problems dates back to the seminal paper of [4] and has since been developed and applied to monotone singular control problems in [25, 26, 27, 28, 29, 17, 18, 19, 2].<sup>1</sup> Indeed, our integral representation theorem for the reversible investment problem is in part inspired by the elegant integration arguments of [2] for irreversible investment. Another closely related body of work is [11, 8, 9, 10]. However, the connections between the singular control problem, the entry-exit problem, and Dynkin's game in their works are established within the framework of forward backward stochastic differential equations and require a finite time horizon with the restrictive assumption that the control has only an additive affect on the diffusion. As such, their results cannot be used to solve the more general reversible investment problems such as (1).

Compared to all previous works and approaches, the correspondence between singular controls and switching controls in our paper *does not depend on the specific form of the control problem*. Thus, our methodology applies to cases for which  $X_t$  can be any diffusion process other than the geometric Brown motion and to cases for which the running payoff function  $h$  is a *general and nonsmooth function* of both the diffusion  $X_t$  and the control process  $Y_t$ . In fact, our method is applicable when the *underlying randomness is not necessarily captured by a diffusion*. This enables us to solve very general reversible investment problems. In particular, when  $h$  is smooth enough, the regularity assumptions for the value function in [33] are recovered.

The organization of the paper is as follows. In section 2, we define consistent collections of switching controls and describe how to obtain such a collection from a singular control and vice versa. We prove that these transformations define a bijection between the set of singular controls and the set of consistent collections of switching controls and prove a change of variable formula. In section 3, we apply this corre-

---

<sup>1</sup>Recently, in [22] it was observed that both the Dynkin game and the two regime optimal switching problem lead to backward stochastic differential equations with two reflecting barriers.

spondence to the problem of reversible investment and show how the value function of the singular control problem can be represented in terms of the value functions of optimal switching problems. Using this representation, we prove the differentiability of the value function and show that, due to the relationship between optimal switching problems and Dynkin games, the derivative can be represented in terms of either one. Last, we give a two-dimensional example with an explicit solution.

## 2. Correspondence between singular controls and switching controls.

In this section, we establish by explicit construction a bijection between admissible singular controls and consistent collections of switching controls with two regimes. Our result is analogous to the well-known correspondence between a nondecreasing,  $\mathbb{F}$ -adapted, càglàd singular control  $(\xi_t)_{t \geq 0}$  and a collection of stopping times  $(\tau^\xi(z))_{z \in \mathbb{R}}$ , given by

$$\tau^\xi(z) = \inf\{t \geq 0 : \xi_t > z\}, \quad \text{and} \quad \xi_t = \sup\{z \in \mathbb{R} : \tau^\xi(z) < t\}.$$

**2.1. Definitions.** Let  $(\Omega, \mathcal{F}, P)$  be a complete probability space and  $\mathbb{F} = \{\mathcal{F}_t; 0 \leq t < \infty\}$  a filtration satisfying the usual hypotheses. Let  $\mathcal{I} \subset \mathbb{R}$  be an open (possibly unbounded) interval and  $\bar{\mathcal{I}}$  be its closure.

Let us first recall the notion of *admissible singular controls*.

**DEFINITION 2.1.** *Given  $y \in \bar{\mathcal{I}}$ , an admissible singular control is a pair  $(\xi_t^+, \xi_t^-)_{t \geq 0}$  of  $\mathbb{F}$ -adapted, nondecreasing càglàd processes such that  $\xi^+(0) = \xi^-(0) = 0$ ,  $Y_t := y + \xi_t^+ - \xi_t^- \in \bar{\mathcal{I}}$ , for all  $t \in [0, \infty)$ , and  $d\xi^+$ ,  $d\xi^-$  are supported on disjoint subsets.*

We denote here  $\mathcal{A}_y$  to be the set of admissible strategies corresponding to an initial capacity level of  $y$ .

Since  $d\xi^+$  and  $d\xi^-$  are supported on disjoint subsets,  $\xi^+$  and  $\xi^-$  are the positive and negative variations of  $Y$ , respectively. By the uniqueness of the variation decomposition, there is a one-to-one correspondence between strategies  $(\xi^+, \xi^-) \in \mathcal{A}_y$  and  $\mathbb{F}$ -adapted càglàd finite variation processes  $Y$ , with  $Y_0 = y$  and  $Y_t \in \bar{\mathcal{I}}$  for all  $t$ .

*Throughout the paper,  $(Y_t)_{t \geq 0}$  is a finite variation control process with  $Y_0 = y$ .*

Next, we introduce *admissible switching controls* (with two regimes).

**DEFINITION 2.2.** *A switching control  $\alpha = (\tau_n, \kappa_n)_{n \geq 0}$  consists of an increasing sequence of stopping times  $(\tau_n)_{n \geq 0}$  and a sequence of new regime values  $(\kappa_n)_{n \geq 0}$  that are assumed immediately after each stopping time.*

When there are only two distinct regimes, an optimal switching problem is often referred to as the starting and stopping problem ([12, 23], etc.) or the entry and exit problem ([9, 16], etc.). Following convention, we label the two regimes 0 and 1.

**DEFINITION 2.3.** *A switching control  $\alpha = (\tau_n, \kappa_n)_{n \geq 0}$  is admissible if the following hold almost surely:  $\tau_0 = 0$ ,  $\tau_{n+1} > \tau_n$  for  $n \geq 1$ ,  $\tau_n \rightarrow \infty$ , and for all  $n \geq 0$ ,  $\kappa_n \in \{0, 1\}$  is  $\mathcal{F}_{\tau_n}$  measurable, with  $\kappa_n = \kappa_0$  for even  $n$  and  $\kappa_n = 1 - \kappa_0$  for odd  $n$ .*

Alternatively, an admissible switching control has a more mathematically convenient representation given by its regime indicator function.

**PROPOSITION 2.4.** *There is a one-to-one correspondence between admissible switching controls and the regime indicator function  $I_t(\omega)$ , which is an  $\mathbb{F}$ -adapted càglàd process of finite variation, so that  $I_t(\omega) : \Omega \times [0, \infty) \rightarrow \{0, 1\}$ , with*

$$(2) \quad I_t := \sum_{n=0}^{\infty} \kappa_n 1_{\{\tau_n < t \leq \tau_{n+1}\}}, \quad I_0 = \kappa_0.$$

LEMMA 2.5. *Given an admissible switching control  $\alpha = (\tau_n, \kappa_n)_{n \geq 0}$ , define the increasing càglàd processes  $I^+$  and  $I^-$  by*

$$I_t^+ := \sum_{n>0, \kappa_n=1}^{\infty} 1_{\{\tau_n < t\}}, \quad I_0^+ = 0 \quad \text{and} \quad I_t^- := \sum_{n>0, \kappa_n=0}^{\infty} 1_{\{\tau_n < t\}}, \quad I_0^- = 0.$$

*Then for all  $t \geq 0$ ,  $I_t^\pm < \infty$  almost surely,  $I_t = \kappa_0 + I_t^+ - I_t^-$ , and  $I_t^+$  ( $I_t^-$ ) is the positive (negative) variation of the corresponding regime indicator function.*

Finally, we define a class of *consistent collections of switching controls*. We shall see later that it is exactly this class of consistent collections of switching controls that corresponds to singular controls of finite variation.

DEFINITION 2.6. *Let  $y \in \bar{\mathcal{I}}$  be given, and for each  $z \in \mathcal{I}$ , let  $\alpha(z) = (\tau_n(z), \kappa_n(z))_{n \geq 0}$  be a switching control. The collection  $(\alpha(z))_{z \in \mathcal{I}}$  is consistent if*

$$(3) \quad \alpha(z) \text{ is admissible for Lebesgue-almost every } z \in \mathcal{I},$$

$$(4) \quad I_0(z) := \kappa_0(z) = 1_{\{z \leq y\}} \text{ for Lebesgue-almost every } z \in \mathcal{I},$$

*and, for all  $t < \infty$ ,*

$$(5) \quad \int_{\mathcal{I}} (I_t^+(z) + I_t^-(z)) dz < \infty, \text{ almost surely, and}$$

$$(6) \quad I_t(z) \text{ is decreasing in } z \text{ for } \mathbb{P} \otimes dz\text{-almost every } (\omega, z).$$

Here  $I_t(z)$ ,  $I_t^+(z)$ , and  $I_t^-(z)$  are defined as in (2) and Lemma 2.5.

For  $I_t(z)$  to be decreasing in  $z$  for  $\mathbb{P} \otimes dz$ -almost every  $(\omega, z)$ , it means there exists a set  $E \subset \Omega \times \bar{\mathcal{I}}$  such that  $\mathbb{P} \otimes dz(E) = 0$ , and if  $(\omega, z_0), (\omega, z_1) \in (\Omega \times \bar{\mathcal{I}}) \setminus E$ , with  $z_0 \leq z_1$ , then  $I_t(\omega, z_0) \geq I_t(\omega, z_1)$ .

**2.2. Bijection.** First, we describe how a consistent collection of switching controls can be obtained from an admissible singular control. To this end, we quote two technical lemmas, the first one adapted from [20, Theorem 5.5.1].

LEMMA 2.7 (Evans and Gariepy). *Let  $f : [0, \infty) \rightarrow \mathbb{R}$  be a function of finite (i.e., locally bounded) variation, and define  $E : [0, \infty) \times \mathbb{R} \rightarrow \mathbb{R}$  by  $E(s, z) = 1_{\{f(s) > z\}}$ . Then*

1. *the function  $E(\cdot, z)$  is of finite variation for almost all  $z \in \mathbb{R}$ , and*
2.  *$\|df\|([0, t)) = \int_{-\infty}^{\infty} \|dE(\cdot, z)\|([0, t)) dz$  for all  $t \in (0, \infty)$ .*

LEMMA 2.8. *Let  $f : [0, \infty) \rightarrow \{0, 1\}$  be of finite variation, and define  $g(t) = \lim_{s \uparrow t} f(s)$ . Then almost surely the paths of  $g$  are càglàd, and, for all  $T < \infty$ ,*

$$\|dg\|([0, T)) \leq \|df\|([0, T)) < \infty.$$

PROPOSITION 2.9 (from singular to switching controls). *Given  $(\xi^+, \xi^-) \in \mathcal{A}_y$ , define a switching control  $\alpha(z) = (\tau_n(z), \kappa_n(z))_{n \geq 0}$  for each  $z \in \mathcal{I}$  through the regime indicator function  $I_t(z) := \lim_{s \uparrow t} 1_{\{Y_s > z\}}$ . Then the resulting collection  $(\alpha(z))_{z \in \mathcal{I}}$  of switching controls is consistent.*

*Proof.* First we show that  $I$  is a regime indicator function as per Proposition 2.4. Let  $Y_t = y + \xi_t^+ - \xi_t^-$ . Since  $Y$  is  $\mathbb{F}$ -adapted, so is  $I(z)$ . Furthermore, since  $Y$  is of finite variation, Lemma 2.7 implies that the function  $s \mapsto 1_{\{Y_s(\omega) > z\}}$  is of finite variation for  $\mathbb{P} \otimes dz$ -almost every  $(\omega, z)$ . Hence  $I_t(\omega, z)$  is càglàd and of finite variation  $(\omega, z)$ -a.e.

by Lemma 2.8. Thus, for almost all  $z$ , there is an admissible switching control  $\alpha(z)$  corresponding to  $I(z)$ .

Moreover, Lemma 2.7, in conjunction with Lemmas 2.8 and 2.5, implies that

$$\begin{aligned} \|dY\|([0, t)) &= \xi_t^+ + \xi_t^- = \int_{-\infty}^{\infty} \|d1_{\{Y > z\}}\|([0, t)) dz \geq \int_{\mathcal{I}} \|dI(z)\|([0, t)) dz \\ &= \int_{\mathcal{I}} I_t^+(z) + I_t^-(z) dz. \end{aligned}$$

Hence  $\int_{\mathcal{I}} I_t^+(z) + I_t^-(z) dz < \infty$ . In addition,  $I_t(z)$  is decreasing in  $z$  and  $I_0(z) = 1_{\{y \geq z\}}$  for all  $z$  (except for  $z = y$ ), so the collection  $(\alpha(z))_{z \in \mathcal{I}}$  is consistent.  $\square$

Next, we construct an admissible singular control  $(\xi^+, \xi^-)$  from a consistent collection of switching controls via their regime indicator functions. Consequently, we give two useful representations of a finite variation process  $Y$ .

**PROPOSITION 2.10** (from switching controls to singular controls). *Given  $y \in \bar{\mathcal{I}}$  and a consistent collection of switching controls  $(\alpha(z))_{z \in \mathcal{I}}$ , define two processes  $\xi^+$  and  $\xi^-$  by setting  $\xi_0^+ = 0$ ,  $\xi_0^- = 0$ , and for  $t > 0$ :  $\xi_t^+ := \int_{\mathcal{I}} I_t^+(z) dz$ ,  $\xi_t^- := \int_{\mathcal{I}} I_t^-(z) dz$ . Then*

1. *the pair  $(\xi^+, \xi^-) \in \mathcal{A}_y$  is an admissible singular control,*
2. *up to indistinguishability,*

$$Y_t = y + \int_y^{\infty} I_t(z) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y (I_t(z) - 1) 1_{\{z \in \mathcal{I}\}} dz, \quad \text{and}$$

3. *for all  $t$ , we almost surely have*

$$Y_t = \text{ess sup}\{z \in \mathcal{I} : I_t(z) = 1\} = \text{ess inf}\{z \in \mathcal{I} : I_t(z) = 0\},$$

where  $\text{ess sup } \emptyset := \inf \mathcal{I}$  and  $\text{ess inf } \emptyset := \sup \mathcal{I}$ .

*Proof.*

1. The proof is obvious from the property of  $(\alpha(z))_{z \in \mathcal{I}}$ , Definition 2.6, and Lemma 2.5.
2. By applying Lemma 2.5, we have a.e. for every  $t \geq 0$ ,

$$\begin{aligned} Y_t &= y + \xi_t^+ - \xi_t^- = y + \int_{\mathcal{I}} (I_t^+(z) - I_t^-(z)) dz \\ &= y + \int_y^{\infty} I_t(z) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y (I_t(z) - 1) 1_{\{z \in \mathcal{I}\}} dz. \end{aligned}$$

3. By fixing  $t \geq 0$  and observing that  $I_t(z) \in \{0, 1\}$  is decreasing in  $z$ ,  $(\omega, z)$ -a.e., we see

$$\begin{aligned} Y_t &= y + \int_y^{\infty} I_t(z) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y (I_t(z) - 1) 1_{\{z \in \mathcal{I}\}} dz \\ &= y + [\text{ess sup}\{z \in \mathcal{I} : I_t(z) = 1\} - y]^+ - [\text{ess sup}\{z \in \mathcal{I} : I_t(z) = 1\} - y]^- \\ &= \text{ess sup}\{z \in \mathcal{I} : I_t(z) = 1\}. \quad \square \end{aligned}$$

PROPOSITION 2.11 (one-to-one mapping). *The mapping from consistent collections of switching controls to singular controls defined by Proposition 2.10 is one-to-one.*

Finally, we shall show that the two mappings defined in Propositions 2.9 and 2.10 are inverses of each other, thus inducing a bijection.

THEOREM 2.12 (bijection). *The mappings in Propositions 2.9 and 2.10 define a bijection between admissible singular controls  $(\xi^+, \xi^-) \in \mathcal{A}_y$  and consistent collections of switching controls (up to equivalence).*

*Proof.* To show that the constructions in Propositions 2.9 and 2.10 are inverses of each other, let us start with a singular control  $(\xi^+, \xi^-) \in \mathcal{A}_y$ . First, applying Proposition 2.9 to  $(\xi^+, \xi^-) \in \mathcal{A}_y$  generates a collection  $(\alpha(z))_{z \in \mathcal{I}}$  of switching controls. Then, applying Proposition 2.10 to  $(\alpha(z))_{z \in \mathcal{I}}$  yields another pair of singular controls  $(\tilde{\xi}^+, \tilde{\xi}^-)$ . We shall show that  $(\xi^+, \xi^-) = (\tilde{\xi}^+, \tilde{\xi}^-)$ .

By Proposition 2.9 we have  $I_t(z) = \lim_{s \uparrow t} 1_{\{Y_s > z\}}$ . Therefore, by Proposition 2.10, the dominated convergence theorem, and with  $Y \in \mathcal{I}$  almost surely, we have

$$\begin{aligned} \tilde{Y}_t &= y + \int_y^\infty I_t(z) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y (I_t(z) - 1) 1_{\{z \in \mathcal{I}\}} dz \\ &= y + \int_y^\infty \lim_{s \uparrow t} 1_{\{Y_s > z\}} 1_{\{z \in \mathcal{I}\}} dz - \int_{-\infty}^y \lim_{s \uparrow t} 1_{\{Y_s \leq z\}} 1_{\{z \in \mathcal{I}\}} dz \\ &= y + \lim_{s \uparrow t} \int_y^\infty 1_{\{Y_s > z\}} 1_{\{z \in \mathcal{I}\}} dz - \lim_{s \uparrow t} \int_{-\infty}^y 1_{\{Y_s \leq z\}} 1_{\{z \in \mathcal{I}\}} dz \\ &= y + [Y_t - y]^+ - [Y_t - y]^- = Y_t. \end{aligned}$$

Thus,  $\tilde{Y}_t$  and  $Y_t$  have the same variation decompositions, and hence  $(\xi^+, \xi^-) = (\tilde{\xi}^+, \tilde{\xi}^-)$  almost surely.

Since the mapping in Proposition 2.10 is one-to-one by Proposition 2.11, this proves that the mappings in Propositions 2.10 and 2.9 are inverses of each other, and hence a bijection exists.  $\square$

Given this correspondence, we shall use the following terminology in what follows. Given a singular control  $(\xi^+, \xi^-) \in \mathcal{A}_y$ , the corresponding collection of switching controls  $(\alpha(z))_{z \in \mathcal{I}}$  refers to the one defined in Proposition 2.9; given a consistent collection of switching controls, the corresponding singular control refers to that in Proposition 2.10.

**2.3. Change of variable formula.** With the bijection established in Theorem 2.12, we are ready to establish a change of variable formula for integration with respect to the variation of a singular control.

LEMMA 2.13. *Let  $(\xi^+, \xi^-) \in \mathcal{A}_y$  be an admissible singular control and  $(\alpha(z))_{z \in \mathcal{I}}$  be the corresponding collection of switching controls. For every càdlàg process  $g : \Omega \times [0, \infty] \rightarrow [0, \infty)$ , with  $g(\infty) \equiv 0$ ,*

$$\begin{aligned} \int_{[0, \infty)} g(t) d\xi_t^+ &= \int_{\mathcal{I}} \sum_{\substack{n > 0 \\ \kappa_n = 1}} g(\tau_n(z)) dz, \quad \text{a.s.}, \\ \text{and } \int_{[0, \infty)} g(t) d\xi_t^- &= \int_{\mathcal{I}} \sum_{\substack{n > 0 \\ \kappa_n = 0}} g(\tau_n(z)) dz, \quad \text{a.s.} \end{aligned}$$



*Proof.* We shall show only the result for  $\xi^+$  as the proof for  $\xi^-$  is almost identical. Suppose  $g$  is a càdlàg process with the representation

$$g(t) = \sum_{i=0}^N g_i 1_{[\sigma_i, \sigma_{i+1})}(t),$$

where  $N$  is finite and constant,  $0 = \sigma_0 \leq \sigma_1 \leq \dots \leq \sigma_{N+1} < \infty$ ,  $g_i \in \mathcal{F}_{\sigma_i}$ , and  $|g_i| < \infty$  almost surely. Then, by Proposition 2.10, the monotone convergence theorem, and Fubini's theorem,

$$\begin{aligned} \int_{[0, \infty)} g(t) d\xi_t^+ &= \int_{[0, \infty)} \sum_{i=0}^N g_i 1_{[\sigma_i, \sigma_{i+1})}(t) d\xi_t^+ = \sum_{i=0}^N g_i \int_{[0, \infty)} 1_{[\sigma_i, \sigma_{i+1})}(t) d\xi_t^+ \\ &= \sum_{i=0}^N g_i (\xi_{\sigma_{i+1}}^+ - \xi_{\sigma_i}^+) \\ &= \sum_{i=0}^N g_i \left( \int_{\mathcal{I}} \sum_{\substack{n>0 \\ \kappa_n=1}} 1_{\{\tau_n(z) < \sigma_{i+1}\}} dz - \int_{\mathcal{I}} \sum_{\substack{n>0 \\ \kappa_n=1}} 1_{\{\tau_n(z) < \sigma_i\}} dz \right) \\ &= \sum_{i=0}^N g_i \left( \int_{\mathcal{I}} \sum_{\substack{n>0 \\ \kappa_n=1}} 1_{\{\sigma_i \leq \tau_n(z) < \sigma_{i+1}\}} dz \right) \\ &= \int_{\mathcal{I}} \sum_{\substack{n>0 \\ \kappa_n=1}} \sum_{i=0}^N g_i 1_{\{\sigma_i \leq \tau_n(z) < \sigma_{i+1}\}} dz \\ &= \int_{\mathcal{I}} \sum_{\substack{n>0 \\ \kappa_n=1}} g(\tau_n(z)) dz. \end{aligned}$$

Since piecewise constant left continuous functions can uniformly approximate càglàd functions, this formula holds for all càglàd processes.  $\square$

In particular, when  $Y$  is nondecreasing (i.e.,  $\xi^- \equiv 0$ ),  $\bar{\mathcal{I}} = [0, \infty)$ , and  $y \geq 0$ , we have  $\tau_n(z) \equiv 0$  for all  $n > 1$  and for  $n = 1$  when  $z \leq y$ . In this case, our change of variable formula reduces to the one for monotone controls in [2], after adjusting for notational differences:

$$\int_{[0, \infty)} g(t) d\xi_t^+ = \int_y^\infty g(\tau_1(z)) dz.$$

**3. Application: Reversible investment.** Having established the correspondence between singular controls and consistent collections of switching controls, we shall illustrate how this theory can be applied to solving singular control problems.

As an example, we return to the aforementioned infinite-horizon, *reversible* investment problem (1): A company adjusts its reversible production capacity (or investment) level by proper controls of expansion and contraction in the presence of a stochastic economic environment. The net profit of such an investment depends on the running production function of the actual capacity, the economic uncertainty

such as price or demand for the product, the benefits of contraction (e.g., via spinning off part of the business), and the cost of expanding and reducing the capital. The company's objective is to maximize the expected profit over an infinite time horizon by controlling expansion and contraction.

**3.1. The singular control problem for reversible investment.** More specifically, the instantaneous operating profit of the company is a function of the production capacity and random variables representing the uncertain economic environment:

$$(7) \quad \Pi(\omega, t, z) : \Omega \times [0, \infty) \times \bar{\mathcal{I}} \rightarrow \mathbb{R}.$$

The unit cost of increasing the capacity at time  $t$  is  $\gamma_+(\omega, t) : \Omega \times [0, \infty) \rightarrow \mathbb{R}$ , and the unit cost of decreasing capacity is  $\gamma_-(\omega, t) : \Omega \times [0, \infty) \rightarrow \mathbb{R}$ , where both  $\gamma_+$  and  $\gamma_-$  are adapted to  $\mathbb{F}$ .<sup>2</sup>

The control of the production capacity  $Y_t$  is represented by a pair  $(\xi_t^+, \xi_t^-)_{t \geq 0}$  of  $\mathbb{F}$ -adapted, nondecreasing càglàd processes such that

$$(8) \quad \xi^+(0) = \xi^-(0) = 0,$$

$$(9) \quad Y_t = y + \xi_t^+ - \xi_t^- \in \bar{\mathcal{I}} \quad \forall t \in [0, \infty).$$

Here  $\xi_t^+$  and  $\xi_t^-$  represent the cumulative expansion and reduction of capital until time  $t$ , respectively. We say the policy  $(\xi^+, \xi^-)$  is integrable if the integrability condition is satisfied for the initial capacity level  $y$ . That is,

$$(10) \quad \mathbb{E} \left[ \int_0^\infty |\Pi(t, Y_t)| dt + \int_{[0, \infty)} |\gamma_+(t)| d\xi_t^+ + \int_{[0, \infty)} |\gamma_-(t)| d\xi_t^- \right] < \infty.$$

We denote  $\mathcal{A}'_y \subset \mathcal{A}_y$  as the set of integrable strategies.

Faced with these profit and cost functions, the company must choose an investment strategy of capacity expansion and reduction which produces the following expected payoff over an infinite horizon:

$$(11) \quad J(y, \xi^+, \xi^-) := \mathbb{E} \left[ \int_0^\infty \Pi(t, Y_t) dt - \int_{[0, \infty)} \gamma_+(t) d\xi_t^+ - \int_{[0, \infty)} \gamma_-(t) d\xi_t^- \right].$$

The objective is to maximize over all integrable policies  $(\xi^+, \xi^-) \in \mathcal{A}'_y$ . Accordingly, the value function is defined as

$$(12) \quad V(y) := \sup_{(\xi^+, \xi^-) \in \mathcal{A}'_y} J(y, \xi^+, \xi^-).$$

Note that, for any  $y \in \bar{\mathcal{I}}$ ,  $\mathcal{A}'_y$  is not empty, as the expected profit of not investing at all (i.e.,  $\xi^+ \equiv 0 \equiv \xi^-$ ) is finite and is given by

$$(13) \quad R(y) := J(y, 0, 0) = \mathbb{E} \left[ \int_0^\infty \Pi(t, y) dt \right].$$

Throughout the remaining section, we impose several conditions.

<sup>2</sup>When there is no risk of ambiguity, we suppress the dependence of the profit and cost functions on  $\omega$ , writing  $\Pi(t, z)$ ,  $\gamma_+(t)$ , and  $\gamma_-(t)$ .

**Standing assumptions.**

- A1.  $\Pi$  is concave in  $y$  and continuous at the boundary of  $\mathcal{I}$ , so that for  $y_1 < y_2 \in \bar{\mathcal{I}}$ ,

$$(14) \quad \Pi(t, y_2) - \Pi(t, y_1) := \int_{y_1}^{y_2} \pi(t, z) dz,$$

where  $\pi$  is decreasing in  $z$  a.s. and adapted to  $\mathbb{F}$ . Furthermore,

$$(15) \quad \mathbb{E} \left[ \int_0^\infty |\Pi(t, z)| dt \right] < \infty \quad \forall z \in \bar{\mathcal{I}},$$

$$(16) \quad \mathbb{E} \left[ \int_0^\infty |\pi(t, z)| dt \right] < \infty \quad \forall z \in \mathcal{I}.$$

This assumption implies that the value function is well-defined, and, although it may take values of  $+\infty$ , it is never  $-\infty$  since  $V(y) \geq R(y) > -\infty$  by (15).

- A2.  $\gamma_+$  and  $\gamma_-$  are adapted to  $\mathbb{F}$ ,  $\gamma^\pm(\infty) := 0$  and

$$(17) \quad \gamma_+(t) + \gamma_-(t) > 0, \quad \text{for all } t, \text{ a.s.}$$

This restriction eliminates the opportunity of making profit by simply switching regimes and immediately switching back.

- A3.
  - If  $\mathcal{I}$  is not bounded above, then  $\gamma_+(t) \geq 0$  for all  $t$  almost surely;
  - if  $\mathcal{I}$  is not bounded below,  $\gamma_-(t) \geq 0$  for all  $t$  almost surely.

This is to ensure that, when the domain is unbounded, an arbitrarily large profit is not obtainable by arbitrarily large changes in the capacity level.

A very special case for the above problem (12) is  $\Pi(\omega, t, z) = e^{-\rho t} (X_t^x(\omega))^\lambda z^\beta$ , where the randomness in the economy is captured by the price process  $X$  of the commodity, and  $X$  is modeled by a geometric Brownian motion  $dX_t^x = bX_t^x dt + \sqrt{2}\sigma X_t^x dW_s$ , with  $X_0 = x > 0$ . The cost functions are  $\gamma_+(\omega, t) = e^{-\rho t} K_1$ ,  $\gamma_-(\omega, t) = e^{-\rho t} K_0$  for some constant  $\rho > 0$ ,  $K_0, K_1$ . We shall provide a detailed analysis and an explicit solution to this case in section 3.4.

**3.2. The corresponding optimal switching problems.** The key to using the connection between singular controls and switching controls to solve problem (12) in section 3.1 is to write the payoff of this problem in terms of the payoffs of its corresponding optimal switching problems. This is accomplished by exploiting the absolute continuity of the running payoff and the change of variable formula for the cost processes.

**3.2.1. Switching controls from singular controls.** First, given the running profit and cost functions from the singular control problem (12), we define a collection of optimal switching problems, indexed by  $z \in \mathcal{I}$ .

DEFINITION 3.1. *The switching cost process  $\gamma : \Omega \times [0, \infty) \times \{0, 1\} \rightarrow \mathbb{R}$  is given by*

$$\gamma(t, \kappa) := \gamma_+(t) 1_{\{\kappa=1\}} + \gamma_-(t) 1_{\{\kappa=0\}}.$$

Here  $\gamma(t, \kappa)$  represents the cost of switching to regime  $\kappa$  at time  $t$ .

The following lemma shows that, for the integrable singular control  $(\xi^+, \xi^-) \in \mathcal{A}'_y$ , the switching controls in the corresponding collection satisfy a certain integrability

condition. It is a simple application of Fubini's theorem, from Lemma 2.13 and condition (10).

LEMMA 3.2. *If  $(\xi^+, \xi^-) \in \mathcal{A}'_y$ , then, for the corresponding consistent collection of switching controls  $(\alpha(z))_{z \in \mathcal{I}}$ , we have  $\alpha(z) \in \mathcal{B}$  for Lebesgue almost every  $z \in \mathcal{I}$ , where  $\mathcal{B}$  is the set of admissible switching controls  $(\tau_n, \kappa_n)_{n \geq 0}$  satisfying*

$$(18) \quad \mathbb{E} \left[ \sum_{n=1}^{\infty} |\gamma(\tau_n, \kappa_n)| \right] < \infty.$$

Note that the converse of the lemma is not true: A consistent collection of switching controls, each of which is integrable, does not necessarily correspond to an integrable singular control.

Next, we establish the following.

PROPOSITION 3.3. *Assume  $(\xi^+, \xi^-) \in \mathcal{A}'_y$ . Let  $(\alpha(z))_{z \in \mathcal{I}}$  be the corresponding consistent collection of switching controls with regime indicator functions  $I(z)$ ; then*

$$J(y, \xi^+, \xi^-) - R(y) = \int_y^{\infty} m_+(z, \alpha(z)) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-(z, \alpha(z)) 1_{\{z \in \mathcal{I}\}} dz,$$

where

$$(19) \quad m_+(z, \alpha) := \mathbb{E} \left[ \int_0^{\infty} \pi(t, z) I_t dt - \sum_{n=1}^{\infty} \gamma(\tau_n, \kappa_n) \right] \in (-\infty, \infty)$$

$$(20) \quad \text{and } m_-(z, \alpha) := \mathbb{E} \left[ \int_0^{\infty} -\pi(t, z) (1 - I_t) dt - \sum_{n=1}^{\infty} \gamma(\tau_n, \kappa_n) \right] \in (-\infty, \infty).$$

Here  $m_+(z, \alpha)$  and  $m_-(z, \alpha)$  are two expected payoffs for the switching controls for each  $z \in \mathcal{I}$  and  $\alpha \in \mathcal{B}$ , with  $\kappa_0 = k \in \{0, 1\}$ .

*Proof of Proposition 3.3.* Since  $(\xi^+, \xi^-) \in \mathcal{A}'_y$ , we have the integrability conditions (10) and (18). By applying Lemma 2.13 to the positive and negative parts of  $\gamma_+$  and  $\gamma_-$ , we see that

$$\int_{[0, \infty)} \gamma_+(t) d\xi_t^+ + \int_{[0, \infty)} \gamma_-(t) d\xi_t^- = \int_{-\infty}^{\infty} \sum_{n=1}^{\infty} \gamma(\tau_n(z), \kappa_n(z)) dz.$$

Moreover, Proposition 2.9 implies that  $I_t(z) = \lim_{s \uparrow t} 1_{\{Y_s > z\}}$ ,  $(\omega, z)$ -a.e. and that  $I(z)$  is of finite variation,  $(\omega, z)$ -a.e. Thus,  $I_t(z) = 1_{\{Y_t > z\}}$ ,  $(\omega, z, t)$ -a.e.

Therefore, by Fubini's theorem and (14), we almost surely have

$$\begin{aligned} \int_0^{\infty} (\Pi(t, Y_t) - \Pi(t, y)) dt &= \int_0^{\infty} \int_y^{\infty} \pi(t, z) 1_{\{Y_t > z\}} dz dt \\ &\quad - \int_0^{\infty} \int_{-\infty}^y \pi(t, z) 1_{\{Y_t \leq z\}} dz dt \\ &= \int_y^{\infty} \int_0^{\infty} \pi(t, z) I_t(z) dt dz \\ &\quad + \int_{-\infty}^y \int_0^{\infty} -\pi(t, z) (1 - I_t(z)) dt dz. \end{aligned}$$

Resorting again to the integrability conditions (10) and (18) and Fubini's theorem yields

$$\begin{aligned}
& J(y, \xi^+, \xi^-) - R(y) \\
&= \mathbb{E} \left[ \int_0^\infty \Pi(t, Y_t) dt - \int_{[0, \infty)} \gamma_+(t) d\xi_t^+ - \int_{[0, \infty)} \gamma_-(t) d\xi_t^- \right] - \mathbb{E} \left[ \int_0^\infty \Pi(t, y) dt \right] \\
&= \mathbb{E} \left[ \int_y^\infty \int_0^\infty \pi(t, z) I_t(z) dt - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) dz \right] \\
&\quad + \mathbb{E} \left[ \int_{-\infty}^y \int_0^\infty -\pi(t, z)(1 - I_t(z)) dt - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) dz \right] \\
&= \int_y^\infty \mathbb{E} \left[ \int_0^\infty \pi(t, z) I_t(z) dt - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) \right] dz \\
&\quad + \int_{-\infty}^y \mathbb{E} \left[ \int_0^\infty -\pi(t, z)(1 - I_t(z)) dt - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) \right] dz \\
&= \int_y^\infty m_+(z; \alpha(z)) dz + \int_{-\infty}^y m_-(z; \alpha(z)) dz.
\end{aligned}$$

The finiteness of the payoff for  $z \in \mathcal{I}$  follows from the assumed integrability of  $\pi$  in (16) and  $|I_t| \leq 1$ .  $\square$

**3.2.2. Representation theorem.** Now, for each  $z \in \mathcal{I}$ , the optimal switching control problem is to maximize the expected payoff over possible switching controls  $\alpha \in \mathcal{B}$  such that  $\kappa_0 = k \in \{0, 1\}$ . This leads to the value functions given by

$$(21) \quad m_+^*(z, k) := \sup_{\substack{\alpha \in \mathcal{B} \\ \kappa_0 = k}} m_+(z, \alpha),$$

$$(22) \quad m_-^*(z, k) := \sup_{\substack{\alpha \in \mathcal{B} \\ \kappa_0 = k}} m_-(z, \alpha),$$

where  $m_+(z, \alpha)$  and  $m_-(z, \alpha)$  are given by (19) and (20).

In fact, these two value functions (21) and (22) are essentially the same as shown in the following lemma.

**LEMMA 3.4.** *The value functions  $m_+^*(z, k)$  and  $m_-^*(z, k)$  in (21) and (22) satisfy, for  $k \in \{0, 1\}$ ,*

$$m_+^*(z, k) - m_-^*(z, k) = \mathbb{E} \left[ \int_0^\infty \pi(t, z) dt \right].$$

*In addition, for fixed  $k \in \{0, 1\}$ , each switching control  $\alpha \in \mathcal{B}$  that is optimal for (21) will also be optimal for (22) and vice versa.*

The proof follows easily by observing that, for any control  $\alpha \in \mathcal{B}$  and any fixed  $z \in \mathcal{I}$ ,

$$m_+(z, \alpha) - m_-(z, \alpha) = \mathbb{E} \left[ \int_0^\infty \pi(t, z) I_t + \pi(t, z)(1 - I_t) dt \right] = \mathbb{E} \left[ \int_0^\infty \pi(t, z) dt \right]. \quad \square$$

Next, we obtain the following lower bounds on the value functions of the switching problems, by considering the no-switching strategies ( $\tau_n = \infty$  for all  $n$ ).

PROPOSITION 3.5. *Given  $m_+^*(z, k)$  and  $m_-^*(z, k)$  in (21) and (22),*

$$\begin{aligned} m_+^*(z, 0) &\geq 0, & m_+^*(z, 1) &\geq \mathbb{E} \left[ \int_0^\infty \pi(t, z) dt \right], \\ m_-^*(z, 0) &\geq -\mathbb{E} \left[ \int_0^\infty \pi(t, z) dt \right], & m_-^*(z, 1) &\geq 0. \end{aligned}$$

Moreover, we have the following upper bound on the value function of the singular control problem.

PROPOSITION 3.6. *Given  $V(y)$  and  $R(y)$  from (12) and (13) and  $m_+^*(z, k)$  and  $m_-^*(z, k)$  in (21) and (22),*

$$(23) \quad V(y) - R(y) \leq \int_y^\infty m_+^*(z, 0) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-^*(z, 1) 1_{\{z \in \mathcal{I}\}} dz.$$

*Proof of Proposition 3.6.* Given any integrable strategy  $(\xi^+, \xi^-) \in \mathcal{A}'_y$ , let  $(\alpha(z))_{z \in \mathcal{I}}$  be the corresponding consistent collection of switching controls. From Proposition 3.3,

$$\begin{aligned} J(y, \xi^+, \xi^-) - R(y) &= \int_y^\infty m_+(z, \alpha(z)) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-(z, \alpha(z)) 1_{\{z \in \mathcal{I}\}} dz \\ &\leq \int_y^\infty m_+^*(z, 0) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-^*(z, 1) 1_{\{z \in \mathcal{I}\}} dz, \end{aligned}$$

since for  $z > y$ ,  $m_+(z, \alpha(z)) \leq m_+^*(z, 0)$  and for  $z \leq y$ ,  $m_-(z, \alpha(z)) \leq m_-^*(z, 1)$ .

(23) follows easily by taking the supremum over all  $(\xi^+, \xi^-) \in \mathcal{A}'_y$ .  $\square$

However, the other direction of the inequality requires additional conditions to guarantee the existence of a consistent collection of optimal (or near-optimal) switching controls and that this consistent collection corresponds to an integrable singular control.

THEOREM 3.7 (representation). *Fix  $y \in \bar{\mathcal{I}}$ , and let  $V(y)$  and  $R(y)$  be given from (12),  $m_+^*(z, k)$  and  $m_-^*(z, k)$  be given by (21) and (22), and  $(\hat{\xi}^{j+}, \hat{\xi}^{j-}) \in \mathcal{A}_y$  be the corresponding singular control as per Proposition 2.10. Assume there is a sequence of consistent collections of switching controls  $(\alpha_j(z))_{z \in \mathbb{R}}$  so that, as  $j \rightarrow \infty$ ,*

$$\begin{aligned} &\int_y^\infty m_+(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz \\ &\rightarrow \int_y^\infty m_+^*(z, 0) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-^*(z, 1) 1_{\{z \in \mathcal{I}\}} dz. \end{aligned}$$

*Assume also that  $(\hat{\xi}^{j+}, \hat{\xi}^{j-}) \in \mathcal{A}'_y$  for all  $j$ . Then*

$$V(y) - R(y) = \int_y^\infty m_+^*(z, 0) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-^*(z, 1) 1_{\{z \in \mathcal{I}\}} dz.$$

*Proof of Theorem 3.7.* Define

$$Q(y) := \int_y^\infty m_+^*(z, 0) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-^*(z, 1) 1_{\{z \in \mathcal{I}\}} dz.$$

First, we treat the case where  $Q(y) = \infty$ . Let  $\epsilon > 0$  be given. From the assumption, find  $j$  so large that  $\frac{1}{\epsilon} < \int_y^\infty m_+(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz$ . By Proposition 3.3 we have

$$\begin{aligned} V(y) - R(y) &\geq J(y, \hat{\xi}^{j+}, \hat{\xi}^{j-}) - R(y) \\ &= \int_y^\infty m_+(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz > \frac{1}{\epsilon}. \end{aligned}$$

Since  $\epsilon$  is arbitrary and  $R(y)$  is finite,  $V(y) = \infty = Q(y)$ .

Next, suppose  $Q(y) < \infty$ , and let  $\epsilon > 0$  be given. Again from the assumption, find  $j$  so large that  $\int_y^\infty m_+(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz > Q(y) - \epsilon$ . By Propositions 3.3 and 3.6,

$$\begin{aligned} V(y) - R(y) &\geq J(y, \hat{\xi}^+, \hat{\xi}^-) - R(y) \\ &= \int_y^\infty m_+(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz \\ &> Q(y) - \epsilon \geq V(y) - R(y) - \epsilon. \end{aligned}$$

Since  $\epsilon$  is arbitrary,  $V(y) - R(y) = Q(y)$  as desired.  $\square$

Moreover, with stronger assumptions, one can further establish the existence of an optimal control strategy, from Propositions 3.3 and 3.6.

*Assumption 3.8.*

1. [Existence of consistent controls]. Fix  $y \in \bar{\mathcal{I}}$ , and let  $m_+^*(z, k)$  and  $m_-^*(z, k)$  be given by (21) and (22). For almost all  $z \in \mathcal{I}$ , there exists an optimal admissible switching control  $\alpha(z) \in \mathcal{B}$  such that

$$\begin{aligned} m_+^*(z, 0) &= m_+(z, \alpha(z)) \quad \text{for } z > y \\ \text{and } m_+^*(z, 1) &= m_+(z, \alpha(z)) \quad \text{for } z \leq y. \end{aligned}$$

Furthermore, the collection  $(\alpha(z))_{z \in \mathbb{R}}$  is consistent.

2. [Integrability of singular control]. Let  $(\hat{\xi}^+, \hat{\xi}^-) \in \mathcal{A}_y$  be the corresponding singular control as per Proposition 2.10, and then  $(\hat{\xi}^+, \hat{\xi}^-) \in \mathcal{A}'_y$ .

**THEOREM 3.9** (representation and existence). *Under Assumption 3.8, the representation Theorem 3.7 holds. Moreover, the strategy  $(\hat{\xi}^+, \hat{\xi}^-)$  is optimal.*

**3.2.3. Remarks on integrability of singular controls.** Although establishing simpler conditions for the consistency of the switching controls requires more structure for the control problem, the equally technical integrability assumption on the singular controls can be reduced to easily verifiable ones when  $\mathcal{I}$  is bounded. These extra assumptions are in line with some of those in [33].

**THEOREM 3.10** (sufficient condition for integrability). *Let  $\mathcal{I}$  be bounded, assume 3.8.1, and let  $(\hat{\xi}^+, \hat{\xi}^-)$  be the corresponding singular control as per Proposition 2.10. Furthermore, suppose*

1.  $\sup_{0 \leq t \leq T} \sup_{z \in \mathcal{I}} |\Pi(\omega, t, z)| < \infty$ , almost surely, for all  $T > 0$ ,
2.  $\limsup_{T \rightarrow \infty} \mathbb{E}[|\gamma_+(T)| + |\gamma_-(T)|] < \infty$ , and
3. for every strategy  $(\xi^+, \xi^-) \in \mathcal{A}_y$ , either  $(\xi^+, \xi^-) \in \mathcal{A}'_y$  or there exists an  $\mathbb{F}$ -adapted process  $Z$  such that  $U. \leq Z$ . almost surely,  $\mathbb{E}[|Z_T|] < \infty$  for all  $T \geq 0$ , and  $\limsup_{T \rightarrow \infty} \mathbb{E}[Z_T] = -\infty$ , where

$$(24) \quad U_T(y, \xi^+, \xi^-) := \int_0^T \Pi(t, Y_t) dt - \int_{[0, T)} \gamma_+(t) d\xi_t^+ - \int_{[0, T)} \gamma_-(t) d\xi_t^-.$$

Then  $(\hat{\xi}^+, \hat{\xi}^-) \in \mathcal{A}'_y$ . Hence Assumption 3.8 holds, yielding Theorem 3.9.

The proof is somewhat technical and thus is given in the appendix.

Note that, when  $\mathcal{I}$  is unbounded, integrable consistent controls may not exist under these extra conditions. Nevertheless, we see the following.

**COROLLARY 3.11.** *If  $\mathcal{I}$  is unbounded, the assumptions of Theorem 3.10 yield Theorem 3.7.*

*Proof.* Let  $\mathcal{I}$  be unbounded and  $(\alpha(z))_{z \in \mathcal{I}}$  be the optimal consistent collection from Assumption 3.8.1. For each  $j \geq 1$ , define  $\alpha_j(z) = \alpha(z)$  for  $z \in \mathcal{I} \cap (-j, j)$ . For  $z \notin \mathcal{I} \cap (-j, j)$ , define  $\alpha_j(z)$  to be the no action switching control (corresponding to the regime indicator function  $I_t^j(z) = 1_{\{z \leq y\}}$ ).

The resulting collection  $(\alpha_j(z))_{z \in \mathcal{I}}$  is clearly consistent, so we let  $(\hat{\xi}^{j+}, \hat{\xi}^{j-})$  be the corresponding singular controls. Furthermore, by considering the control problem restricted to  $\bar{\mathcal{I}} \cap [-j, j]$ , Theorem 3.10 implies that  $(\hat{\xi}^{j+}, \hat{\xi}^{j-}) \in \mathcal{A}'_y$ . Last, by the monotone convergence theorem (since  $m_+^*(z, 0)$  and  $m_-^*(z, 1)$  are nonnegative), we have, as  $j \rightarrow \infty$ ,

$$\begin{aligned} & \int_y^\infty m_+(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-(z, \alpha_j(z)) 1_{\{z \in \mathcal{I}\}} dz \\ &= \int_y^\infty m_+^*(z, 0) 1_{\{z \in \mathcal{I} \cap (-j, j)\}} dz + \int_{-\infty}^y m_-^*(z, 1) 1_{\{z \in \mathcal{I} \cap (-j, j)\}} dz \\ &\rightarrow \int_y^\infty m_+^*(z, 0) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-^*(z, 1) 1_{\{z \in \mathcal{I}\}} dz. \quad \square \end{aligned}$$

**3.3. Regularity of the value function.** In this section, we provide conditions under which the value function of the switching controls is not only continuous but also continuously differentiable. As a result, we prove directly the smooth fit condition assumed a priori in [33].

**PROPOSITION 3.12.** *Suppose that, for some  $y \in \mathcal{I}$ ,*

$$(25) \quad \lim_{z \rightarrow y} \mathbb{E} \left[ \int_0^\infty |\pi(t, z) - \pi(t, y)| dt \right] = 0.$$

*Then for  $k \in \{0, 1\}$ ,  $m_+^*(\cdot, k)$  and  $m_-^*(\cdot, k)$  (from (21) and (22)) are continuous at  $y$ .*

*Proof of Proposition 3.12.* Let  $y \in \mathcal{I}$  and  $k \in \{0, 1\}$  be given, and consider any admissible strategy  $\alpha \in \mathcal{B}$ . By (18), (19), and (25),

$$\lim_{z \rightarrow y} |m_+(z, \alpha) - m_+(y, \alpha)| \leq \lim_{z \rightarrow y} \mathbb{E} \left[ \int_0^\infty |\pi(t, z) - \pi(t, y)| dt \right] = 0.$$

Note that convergence to zero is uniform across all strategies  $\alpha \in \mathcal{B}$ .



Let  $\epsilon > 0$  be given. There exists  $\delta > 0$  so that, for any strategy  $\alpha \in \mathcal{B}$ ,  $|m_+(z, \alpha) - m_+(y, \alpha)| < \frac{\epsilon}{2}$  for all  $z \in \mathcal{I}$  such that  $|z - y| < \delta$ .

Now there exists a strategy  $\hat{\alpha} \in \mathcal{B}$  with  $\kappa_0 = k$  such that

$$m_+^*(y, k) \leq m_+(y, \hat{\alpha}) + \frac{\epsilon}{2}.$$

So for all  $z \in \mathcal{I}$  such that  $|z - y| < \delta$ ,

$$m_+^*(y, k) \leq m_+(y, \hat{\alpha}) + \frac{\epsilon}{2} \leq m_+(z, \hat{\alpha}) + \epsilon \leq m_+^*(z, k) + \epsilon.$$

Furthermore, for any such  $z$ , there exists a switching control  $\alpha_z \in \mathcal{B}$  with  $\kappa_0 = 0$  such that

$$m_+^*(z, k) \leq m_+(z, \alpha_z) + \frac{\epsilon}{2} \leq m_+(y, \alpha_z) + \epsilon \leq m_+^*(y, k) + \epsilon.$$

Hence for all  $z \in \mathcal{I}$  such that  $|z - y| < \delta$ ,

$$m_+^*(y, k) - \epsilon \leq m_+^*(z, k) \leq m_+^*(y, k) + \epsilon.$$

Thus,  $\lim_{z \rightarrow y} m_+^*(z, k) = m_+^*(y, k)$ . Moreover,  $\lim_{z \rightarrow y} m_-^*(z, k) = m_-^*(y, k)$  follows from Lemma 3.4.  $\square$

**THEOREM 3.13** (regularity). *Assume the conditions in Proposition 3.12 on an open interval  $\mathcal{J} \subset \mathcal{I}$ . Suppose that, on  $\mathcal{J}$ , the value function has the representation*

$$V(y) - R(y) = \int_y^\infty m_+^*(z, 0) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y m_-^*(z, 1) 1_{\{z \in \mathcal{I}\}} dz.$$

Then  $V$  is  $C^1$  on  $\mathcal{J}$ , and, for any  $y \in \mathcal{J}$ ,

$$V'(y) = \mathbb{E} \left[ \int_0^\infty \pi(t, y) dt \right] + m_-^*(y, 1) - m_+^*(y, 0) = m_+^*(y, 1) - m_+^*(y, 0).$$

*Proof of Theorem 3.13.* By Proposition 3.12, it remains to show that  $R'(y) = \mathbb{E} \left[ \int_0^\infty \pi(t, y) dt \right]$ . Fixing  $z_0 \in \mathcal{I}$ , this follows easily from (14) and (15), and

$$\begin{aligned} R(y) - R(z_0) &= \mathbb{E} \left[ \int_0^\infty (\Pi(t, y) - \Pi(t, z_0)) dt \right] \\ &= \mathbb{E} \left[ \int_0^\infty \int_{z_0}^y \pi(t, z) dz dt \right] = \int_{z_0}^y \mathbb{E} \left[ \int_0^\infty \pi(t, z) dz dt \right]. \quad \square \end{aligned}$$

Note that previous results of [30, 8, 9, 10] on the differentiability of the value function for the (ir)reversible investment problem are special cases of ours. Another major difference is that the derivative in their work is in terms of the value of a Dynkin game, whereas the derivative here is the difference between the value functions of an optimal switching problem.

In the remainder of this section, we shall show that, under very mild assumptions, the value of a Dynkin game exists and is equal to the difference of the value functions for the optimal switching problem defined by (19) and (20); thereby we demonstrate that optimal switching problems provide a “missing link” between Dynkin games and singular control problems.

For simplicity, we consider an infinite-horizon Dynkin game with no terminal payoff. With a slight modification, our arguments can be adapted for the finite-horizon case.

**3.3.1. Dynkin games.** A Dynkin game is a game of timing between two players, whom we call MAX and MIN, following [10]. We fix some level  $z \in \mathcal{I}$ . While the game is in progress, MIN pays MAX at rate  $\pi(t, z)$ , and the game ends when one player chooses to stop. Thus, MAX and MIN each choose strategies on when to exit the game (the stopping times  $\sigma_-$  and  $\sigma_+$ , respectively). The player to exit first pays an amount to her opponent equal to  $\gamma_-(\sigma_-)$  if MAX exits first and  $\gamma_+(\sigma_+)$  if MIN exits first. If both players exit at the same time, we treat it as though MIN exited first. Furthermore, each player may choose never to exit, i.e.,  $\sigma = \infty$ . MAX chooses her strategy  $\sigma_-$  to maximize her payoff, and MIN chooses  $\sigma_+$  in order to minimize MAX's payoff.

This game is formally described below. To ensure that the payoff of the game is well-defined, we assume in this section that, for every stopping time  $\sigma$ ,  $\mathbb{E}[|\gamma_-(\sigma)|] < \infty$  and  $\mathbb{E}[|\gamma_+(\sigma)|] < \infty$ .

**DEFINITION 3.14.** *Given  $z \in \mathcal{I}$  and  $\mathbb{F}$ -stopping times  $\sigma_-$  and  $\sigma_+$ , the payoff of the Dynkin game is*

$$D(\sigma_-, \sigma_+; z) = \int_0^{\sigma_- \wedge \sigma_+} \pi(t, z) dt + \gamma_+(\sigma_+) 1_{\{\sigma_+ \leq \sigma_-\}} - \gamma_-(\sigma_-) 1_{\{\sigma_- < \sigma_+\}}.$$

*The game has a value if*

$$\sup_{\sigma_-} \inf_{\sigma_+} \mathbb{E}[D(\sigma_-, \sigma_+; z)] = \inf_{\sigma_+} \sup_{\sigma_-} \mathbb{E}[D(\sigma_-, \sigma_+; z)].$$

It is easy to see that  $\sup_{\sigma_-} \inf_{\sigma_+} \mathbb{E}[D(\sigma_-, \sigma_+; z)] \leq \inf_{\sigma_+} \sup_{\sigma_-} \mathbb{E}[D(\sigma_-, \sigma_+; z)]$ . Moreover, we have the following theorem.

**THEOREM 3.15.** *Given any  $z \in \mathcal{I}$  such that conditions (16) and (17) hold, the value of the Dynkin game exists and is equal to*

$$m_+^*(z, 1) - m_+^*(z, 0) = \sup_{\sigma_-} \inf_{\sigma_+} \mathbb{E}[D(\sigma_-, \sigma_+; z)] = \inf_{\sigma_+} \sup_{\sigma_-} \mathbb{E}[D(\sigma_-, \sigma_+; z)].$$

*Proof of Theorem 3.15.* It suffices to show  $m_+^*(z, 1) - m_+^*(z, 0) \leq \sup_{\sigma_-} \inf_{\sigma_+} \mathbb{E}[D(\sigma_-, \sigma_+; z)]$ , since it follows similarly for  $m_+^*(z, 1) - m_+^*(z, 0) \geq \inf_{\sigma_+} \sup_{\sigma_-} \mathbb{E}[D(\sigma_-, \sigma_+; z)]$ .

Note that  $m_+^*(z, 1) - m_+^*(z, 0) \leq \sup_{\sigma_-} \inf_{\sigma_+} \mathbb{E}[D(\sigma_-, \sigma_+; z)]$  if and only if, for all  $\epsilon > 0$ , there exists  $\hat{\sigma}_-$  such that, for all  $\sigma_+$ ,  $\mathbb{E}[D(\hat{\sigma}_-, \sigma_+; z)] + \epsilon \geq m_+^*(z, 1) - m_+^*(z, 0)$ .

Let  $\epsilon > 0$  be given, let  $\alpha^1 \in \mathcal{B}$  be a switching control with  $\kappa_0 = 1$  such that

$$m_+(z, \alpha^1) + \epsilon \geq m_+^*(z, 1),$$

and define  $\hat{\sigma}_- = \tau_1 = \inf\{t : I_t^1 = 0\}$ .

Let  $\sigma_+$  be an arbitrary stopping time, and define  $\alpha^0$  by taking  $I_t^0 = 0$  for  $t \leq \hat{\sigma}_- \wedge \sigma_+$  and  $I_t^0 = I_t^1$  for  $t > \hat{\sigma}_- \wedge \sigma_+$ . Thus  $I^0$  is a regime indicator function, and hence  $\alpha^0$  is an admissible switching control. In fact, since  $\alpha^1 \in \mathcal{B}$ , we also have  $\alpha^0 \in \mathcal{B}$ .

Thus, for any  $\sigma_+$ ,

$$\begin{aligned}
m_+^*(z, 1) - m_+^*(z, 0) &\leq m_+^*(z, 1) - m_+(z, \alpha^0) \leq m_+(z, \alpha^1) - m_+(z, \alpha^0) + \epsilon \\
&= \mathbb{E} \left[ \int_0^\infty \pi(t, z)(I_t^1 - I_t^0) dt + \gamma_+(\sigma_+) 1_{\{\sigma_+ < \hat{\sigma}_-\}} - \gamma_-(\hat{\sigma}_-) 1_{\{\hat{\sigma}_- \leq \sigma_+\}} \right] + \epsilon \\
&= \mathbb{E} \left[ \int_0^{\hat{\sigma}_- \wedge \sigma_+} \pi(t, z) dt + \gamma_+(\sigma_+) 1_{\{\sigma_+ < \hat{\sigma}_-\}} - \gamma_-(\hat{\sigma}_-) 1_{\{\hat{\sigma}_- \leq \sigma_+\}} \right] + \epsilon \\
&= \mathbb{E} \left[ \int_0^{\hat{\sigma}_- \wedge \sigma_+} \pi(t, z) dt + \gamma_+(\sigma_+) 1_{\{\sigma_+ < \hat{\sigma}_-\}} \right. \\
&\quad \left. - \gamma_-(\hat{\sigma}_-) 1_{\{\hat{\sigma}_- < \sigma_+\}} - \gamma_-(\sigma_+) 1_{\{\hat{\sigma}_- = \sigma_+\}} \right] + \epsilon \\
&\leq \mathbb{E} \left[ \int_0^{\hat{\sigma}_- \wedge \sigma_+} \pi(t, z) dt + \gamma_+(\sigma_+) 1_{\{\sigma_+ \leq \hat{\sigma}_-\}} - \gamma_-(\hat{\sigma}_-) 1_{\{\hat{\sigma}_- < \sigma_+\}} \right] + \epsilon \\
&= \mathbb{E} [D(\hat{\sigma}_-, \sigma_+; z)] + \epsilon,
\end{aligned}$$

where the last inequality follows from (17). Thus,

$$m_+^*(z, 1) - m_+^*(z, 0) \leq \sup_{\sigma_-} \inf_{\sigma_+} \mathbb{E} [D(\sigma_-, \sigma_+; z)]. \quad \square$$

Furthermore, we see the following.

**COROLLARY 3.16.** *If (16) and (17) hold and  $\pi(t, z)$  is decreasing in  $z$ , then  $m_+^*(z, 1) - m_+^*(z, 0)$  is decreasing in  $z$ .*

That is, when the marginal payoff is decreasing in the capacity level  $z$ , the added benefit of being invested in the project at level  $z$  is also decreasing in  $z$ . The economic interpretation is that there are decreasing returns to scale.

**3.4. Examples with explicit solutions.** We now illustrate how our methodology can be used to solve a reversible investment problem with a Cobb–Douglas production function. This is a special case of the problem solved in [33]. Note that, although our method can handle the general problem in [33] among others, we nevertheless have selected this simple case to illustrate our techniques: Unlike [33], we solve *without* assuming a priori the continuous differentiability of the value function or any assumptions on the structure of the switching regions.

*Singular control problem.*

$$(26) \quad V(x, y) := \sup_{(\xi^+, \xi^-) \in \mathcal{A}_y} \mathbb{E} \left[ \int_0^\infty \Pi(t, Y_t) dt - \int_{[0, \infty)} \gamma_+(t) d\xi_t^+ - \int_{[0, \infty)} \gamma_-(t) d\xi_t^- \right],$$

subject to

$$\begin{aligned}\xi^+(0) &= \xi^-(0) := 0, \\ Y_t &:= y + \xi_t^+ - \xi_t^- \in \bar{\mathcal{I}} \quad \forall t \in [0, \infty), \\ \Pi(\omega, t, y) &:= e^{-\rho t} (X_t^x(\omega))^\lambda y^\beta, \\ dX_t^x &:= bX_t^x dt + \sqrt{2}\sigma X_t^x dW_t, \quad X_0 := x > 0, \\ \text{and } \gamma_+(t) &:= e^{-\rho t} K_1, \gamma_-(t) := e^{-\rho t} K_0.\end{aligned}$$

For simplicity, we assume  $\bar{\mathcal{I}} = [A, B] \subset [0, \infty)$  is a bounded interval,  $\rho > 0$ ,  $\lambda \in (0, n)$ ,  $\beta \in (0, 1]$ , and  $K_0 < 0$ ,  $K_1 > 0$ , with  $K_0 + K_1 > 0$ . Here  $n = \frac{-(b-\sigma^2) + \sqrt{(b-\sigma^2)^2 + 4\sigma^2\rho}}{2\sigma^2}$ . This formulation is from [33].

This problem is solved in several steps.

*Step 1: Corresponding optimal switching problem.* First, one can check that standing assumptions A1, A2, and A3 and the assumptions in Theorem 3.10 hold for this problem, with

$$\pi(\omega, t, y) = \beta e^{-\rho t} (X_t^x(\omega))^\lambda y^{-(1-\beta)}.$$

Also,  $R(x, y)$  is differentiable in  $y$ , so that

$$R(x, y) = \mathbb{E} \left[ \int_0^\infty \Pi(t, y) dt \right] = \frac{-x^\lambda y^\beta}{\sigma^2 \lambda^2 + (b - \sigma^2) \lambda - \rho},$$

and

$$r(x, y) = R_y(x, y) = \mathbb{E} \left[ \int_0^\infty \pi(t, y) dt \right] = \frac{-\beta x^\lambda y^{-(1-\beta)}}{\sigma^2 \lambda^2 + (b - \sigma^2) \lambda - \rho}.$$

By Lemma 3.4, it suffices to solve for the optimal switching problem defined by

(27)

$$v_k(x, z) := m_+^*(x, z, k) = \sup_{\substack{\alpha \in \mathcal{B} \\ \kappa_0 = k}} \mathbb{E} \left[ \int_0^\infty e^{-\rho t} \beta z^{-(1-\beta)} (X_t^x)^\lambda I_t dt - \sum_{n=1}^\infty e^{-\rho \tau_n} K_{\kappa_n} \right].$$

*Step 2: Consistent collection of optimal switching controls.* Applying the regularity results of [32] for switching controls to problem (27), we see that, for any given  $z \in (A, B)$  and  $k \in \{0, 1\}$ , the value function  $v_k(\cdot, z)$  is continuously differentiable. Moreover, for each  $z \in (A, B)$ , an optimal switching control exists and can be described in terms of the switching regions: For each  $z \in (A, B)$ , there exist  $0 < F(z) < G(z) < \infty$  such that it is optimal to switch from regime 0 to regime 1 (to invest in the project at level  $z$ ) when  $X_t^x \in [G(z), \infty)$  and to switch from regime 1 to regime 0 (disinvest at level  $z$ ) when  $X_t^x \in [0, F(z)]$ .

Therefore, given any initial value  $y \in [A, B]$  for the singular control problem, a collection of optimal switching controls can be defined as follows.

For  $(x, z) \in \mathcal{X} \times (A, B)$ , define the switching control  $\hat{\alpha}(x, z) = (\hat{\tau}_n, \hat{\kappa}_n)_{n \geq 0}$ , starting from  $\hat{\tau}_0 = 0$  and  $\kappa_0 = 1_{\{z \leq y\}}$  by setting  $\hat{\kappa}_n := 1 - \kappa_{n-1}$  for all  $n \geq 1$  and

- if  $\kappa_{n-1} = 0$ ,  $\hat{\tau}_n := \inf\{t > \tau_{n-1} : X_t^x \geq G(z)\}$ , or

- if  $\kappa_{n-1} = 1$ ,  $\hat{\tau}_n := \inf\{t > \tau_{n-1} : X_t^x \leq F(z)\}$ .

Moreover, by the regularity of the value functions, we solve for  $F(z)$  and  $G(z)$  explicitly in our case, obtaining  $F(z) = \kappa z^{\frac{1-\beta}{\lambda}}$  and  $G(z) = \nu z^{\frac{1-\beta}{\lambda}}$ , where  $\kappa$  and  $\nu$  are unique solutions to

$$\frac{\beta}{\lambda - m} [\nu^{\lambda-m} - \kappa^{\lambda-m}] = -\frac{\rho}{m} [K_1 \nu^{-m} + K_0 \kappa^{-m}],$$

$$\frac{\beta}{n - \lambda} [\nu^{\lambda-n} - \kappa^{\lambda-n}] = \frac{\rho}{n} [K_1 \nu^{-n} + K_0 \kappa^{-n}].$$

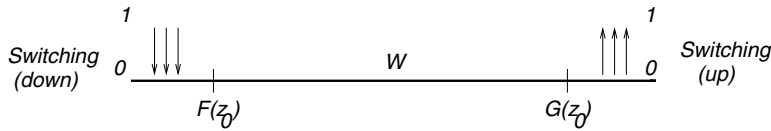
Here  $m < 0 < n$ , and  $n, m = \frac{-(b-\sigma^2) \pm \sqrt{(b-\sigma^2)^2 + 4\sigma^2 r}}{2\sigma^2}$ .

Finally, by checking the appropriate integrability conditions, and by noting that  $F$  and  $G$  are increasing in  $z$ , it is not hard to verify that the above collection of optimal switching controls is consistent. (See Figure 1.)

*Step 3: Optimal singular control and value functions.* By Proposition 2.10, this consistent collection of optimal switching control corresponds to an admissible singular control  $(\hat{\xi}^+, \hat{\xi}^-) \in \mathcal{A}_y$ . Moreover, since  $\mathcal{I}$  is bounded, it is integrable following Theorem 3.10.

Put together, the investment region is given by  $\{(x, z) : x \geq G(z)\}$  and the disinvestment region by  $\{(x, z) : x \leq F(z)\}$ .  $Y_t$  is constant when  $(X_t, Y_t)$  is in the wait region, given by  $\{(x, z) : F(z) < x < G(z)\}$ . If  $(x, y)$  is in the investment (or disinvestment) region, then a jump is exerted at time zero to make  $Y_{0+} = G^{-1}(x)$  (or  $Y_{0+} = F^{-1}(x)$ ).

(1) For fixed  $z_0$  switching control



(2) For general  $z$ , consistent switching controls

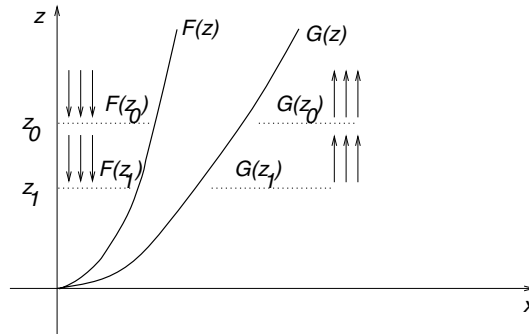


FIG. 1. Illustration of optimal consistent switching control from optimal singular control.

Finally, by Lemma 3.4 and Theorem 3.9 the value function has the following representation:

$$\begin{aligned} V(x, y) &= R(x, y) + \int_y^B v_0(x, z) dz + \int_A^y (v_1(x, z) - r(x, z)) dz \\ &= R(x, A) + \int_y^B v_0(x, z) dz + \int_A^y v_1(x, z) dz, \end{aligned}$$

where  $v_0$  and  $v_1$  are given by [32, Theorem 4.2.3]

$$\begin{aligned} v_0(x, z) &= \begin{cases} A(z)x^n, & x < G(z), \\ B(z)x^m + r(x, z) - K_1, & x \geq G(z), \end{cases} \\ v_1(x, z) &= \begin{cases} A(z)x^n - K_0, & x \leq F(z), \\ B(z)x^m + r(x, z), & x > F(z), \end{cases} \end{aligned}$$

where  $A(z) = \frac{\nu^{-n}}{n-m} \left( \frac{\beta\nu^\lambda}{\sigma^2(n-\lambda)} + mK_1 \right) z^{\frac{-n(1-\beta)}{\lambda}}$ ,  $B(z) = -\frac{\kappa^{-m}}{n-m} \left( \frac{\beta\kappa^\lambda}{\sigma^2(\lambda-m)} + nK_0 \right) z^{\frac{-m(1-\beta)}{\lambda}}$ .

The above results are consistent with equations (117)–(120) and Remark 4 in [33].

**4. Conclusions.** This paper builds a generic connection between singular controls of finite variation and sequential optimal stopping problems. This correspondence is independent of any particular formulation of control problems and provides a novel method for solving explicitly high-dimensional singular control problems where randomness may not be necessarily captured by a diffusion and where payoff functions can be nonsmooth. It also enables us to derive sufficient conditions for the existence of optimal controls, for the smooth fit principle, and for the regularity of value functions. Consequently, this regularity result links singular controls and Dynkin games through sequential optimal stopping problems.

**Appendix A. Proof of Theorem 3.10.** The proof is built on the following proposition.

PROPOSITION A.1. *Under Assumption 3.8.1, for almost all  $z \in \mathcal{I}$  and any  $T > 0$ ,*

$$\mathbb{E}[u_T(z, \alpha(z))] \geq \begin{cases} -\mathbb{E}[\gamma_+(T)I_{T+}(z)] & \text{when } z > y, \\ -\mathbb{E}[\gamma_-(T)(1 - I_{T+}(z))] & \text{when } z \leq y, \end{cases}$$

where  $\alpha(z) \in \mathcal{B}$  is given by Assumption 3.8.1 and

$$u_T(z, \alpha(z)) = \begin{cases} \int_0^T \pi(t, z) I_t(z) dt - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) 1_{\{\tau_n(z) < T\}}, & z > y, \\ \int_0^T -\pi(t, z) (1 - I_t(z)) dt - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) 1_{\{\tau_n(z) < T\}}, & z \leq y. \end{cases}$$

*Proof.* By Assumption 3.8.1,  $\alpha(z)$  is an optimal admissible switching control for almost all  $z \in \mathcal{I}$ . Fix such a  $z \in \mathcal{I}$ . Consider the admissible switching control  $\tilde{\alpha}_T(z)$  defined by the regime indicator function

$$\tilde{I}_t(z) = I_0(z) 1_{\{t \leq T\}} + I_t(z) 1_{\{t > T\}}.$$

Assume for now  $z > y$ . In the new switching control we have defined, we may have to switch at time  $T$  from regime  $\kappa_0 = 0$  to regime  $I_{T+}(z)$ , if  $I_{T+}(z) = 1$ . Hence,

the cost of the possible switch at  $T$  is given by  $-\gamma_+(T)I_{T+}(z)$ . After time  $T$ , the switching costs are the same for both strategies.

Since the switching control  $\alpha(z) \in \mathcal{B}$  is optimal,  $m_+(z, \alpha(z)) - m_+(z, \tilde{\alpha}_T(z)) \geq 0$ . This means that

$$\begin{aligned} 0 &\leq m_+(z, \alpha(z)) - m_+(z, \tilde{\alpha}_T(z)) = \mathbb{E} \left[ \int_0^\infty \pi(t, z) I_t(z) dt - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) \right] \\ &\quad - \mathbb{E} \left[ \int_0^\infty \pi(t, z) \tilde{I}_t(z) dt - \sum_{n=1}^\infty \gamma(\tilde{\tau}_n(z), \tilde{\kappa}_n(z)) \right] \\ &= \mathbb{E} \left[ \int_0^\infty \pi(t, z) I_t(z) dt - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) \right] \\ &\quad - \mathbb{E} \left[ \int_T^\infty \pi(t, z) I_t(z) dt - \gamma_+(T) I_{T+}(z) - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) 1_{\{T < \tau_n(z)\}} \right] \\ &= \mathbb{E} [\gamma_+(T) I_{T+}(z)] + \mathbb{E} \left[ \int_0^T \pi(t, z) I_t(z) dt - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) 1_{\{\tau_n(z) < T\}} \right]. \end{aligned}$$

Thus, by Assumption 3.8,  $\mathbb{E}[u_T(z)] \geq -\mathbb{E}[\gamma_+(T) I_{T+}(z)]$  for almost all  $z > y$ .

Similarly, for almost all  $z \leq y$ ,

$$\begin{aligned} 0 &\leq \mathbb{E} [\gamma_-(T)(1 - I_{T+}(z))] \\ &\quad + \mathbb{E} \left[ \int_0^T -\pi(t, z)(1 - I_t(z)) dt - \sum_{n=1}^\infty \gamma(\tau_n(z), \kappa_n(z)) 1_{\{\tau_n(z) < T\}} \right]. \end{aligned}$$

Hence the claim.  $\square$

*Proof of Theorem 3.10.* Suppose  $\mathcal{I}$  is bounded, and let  $(\alpha(z))_{z \in \mathcal{I}}$  be the collection of optimal, consistent switching controls given by Assumption 3.8.1 and  $(\hat{\xi}^+, \hat{\xi}^-) \in \mathcal{A}_y$  be the corresponding singular control. The idea of the proof is to show that, for any  $\mathbb{F}$ -adapted process  $Z$  with  $U_T \leq Z_T$  and  $\mathbb{E}[|Z_T|] < \infty$  almost surely for each  $T > 0$ , we have  $\limsup_{T \rightarrow \infty} \mathbb{E}[Z_T] > -\infty$ . It then follows from the assumptions of the theorem that  $(\hat{\xi}^+, \hat{\xi}^-) \in \mathcal{A}'_y$ .

By the assumptions of the theorem, for any fixed  $T < \infty$ ,  $|U_T(y, \hat{\xi}^+, \hat{\xi}^-)| < \infty$  almost surely. Furthermore, by applying the same arguments as in the proof of Proposition 3.3 we get

$$U_T(y, \hat{\xi}^+, \hat{\xi}^-) - \int_0^T \Pi(t, y) dt = \int_{\mathcal{I}} u_T(z) dz,$$

where  $u_T$  is defined in Proposition A.1.

Let  $Z$  be any  $\mathbb{F}$ -adapted process with  $U_T \leq Z_T$  and  $\mathbb{E}[|Z_T|] < \infty$  almost surely for each  $T > 0$ . Thus,

$$Z_T \geq U_T(y, \hat{\xi}^+, \hat{\xi}^-) = \int_{\mathcal{I}} u_T(z, \alpha(z)) dz.$$

Since  $\mathcal{I}$  is bounded and  $|u_T|$  has finite expectation,  $\int_{\mathcal{I}} \mathbb{E}[|u_T(z, \alpha(z))|] dz < \infty$ . Hence by Fubini's theorem and Proposition A.1,

$$\begin{aligned}
 \mathbb{E}[Z_T] &\geq \mathbb{E} \left[ \int_{\mathcal{I}} u_T(z, \alpha(z)) dz + \int_0^T \Pi(t, y) dt \right] \\
 &\geq \int_{\mathcal{I}} \mathbb{E}[u_T(z, \alpha(z))] dz - \mathbb{E} \left[ \int_0^\infty |\Pi(t, y)| dt \right] \\
 &\geq - \int_y^\infty \mathbb{E}[\gamma_+(T) I_{T+}(z)] 1_{\{z \in \mathcal{I}\}} dz - \int_{-\infty}^y \mathbb{E}[\gamma_-(T)(1 - I_{T+}(z))] 1_{\{z \in \mathcal{I}\}} dz \\
 &\quad - \mathbb{E} \left[ \int_0^\infty |\Pi(t, y)| dt \right] \\
 &= \mathbb{E}[-\gamma_+(T)[Y_{T+} - y]^+ - \gamma_-(T)[Y_{T+} - y]^-] - \mathbb{E} \left[ \int_0^\infty |\Pi(t, y)| dt \right] \\
 &\geq -C\mathbb{E}[|\gamma_+(T)| + |\gamma_-(T)|] - \mathbb{E} \left[ \int_0^\infty |\Pi(t, y)| dt \right],
 \end{aligned}$$

where  $C = \sup \mathcal{I} - \inf \mathcal{I} < \infty$ .

Thus, by (15) and the assumptions of the theorem,  $\limsup_{T \rightarrow \infty} \mathbb{E}[Z_T] > -\infty$ .  $\square$

**Acknowledgments.** The authors thank the Associate Editor and the two anonymous referees for their constructive and detailed suggestions and remarks, which led to a substantial improvement of the paper. Generous support from the OpenLink Fund at the Coleman Fung Risk Management Center at UC Berkeley is gratefully acknowledged.

#### REFERENCES

- [1] A. B. ABEL AND J. C. EBERLY, *An exact solution for the investment and value of a firm facing uncertainty, adjustment costs, and irreversibility*, J. Econom. Dynam. Control, 21 (1997), pp. 831–852.
- [2] F. M. BALDURSSON AND I. KARATZAS, *Irreversible investment and industry equilibrium*, Finance Stoch., 1 (1997), pp. 69–89.
- [3] P. BANK, *Optimal control under a dynamic fuel constraint*, SIAM J. Control Optim., 44 (2005), pp. 1529–1541.
- [4] J. BATHER AND H. CHERNOFF, *Sequential decisions in the control of a spaceship*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Berkeley, California, 1965/66), Vol. III: Physical Sciences, University of California Press, Berkeley, CA, 1967, pp. 181–207.
- [5] J. BATHER AND H. CHERNOFF, *Sequential decisions in the control of a spaceship (finite fuel)*, J. Appl. Probab., 4 (1967), pp. 584–604.
- [6] V. E. BENEŠ, L. A. SHEPP, AND H. S. WITSENHAUSEN, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.
- [7] A. BENSOUSSAN AND J.-L. LIONS, *Impulse Control and Quasivariational Inequalities*, Gauthier-Villars, Montrouge, 1984.
- [8] F. BOETIUS, *Bounded variation singular stochastic control and associated Dynkin game*, in Mathematical Finance (Konstanz, 2000), Trends Math., Birkhäuser, Basel, 2001, pp. 111–120.
- [9] F. BOETIUS, *Singular Stochastic Control and its Relations to Dynkin Game and Entry-Exit Problems*, Ph.D. dissertation, Universität Konstanz, Konstanz, Germany, 2003.



- [10] F. BOETIUS, *Bounded variation singular stochastic control and Dynkin game*, SIAM J. Control Optim., 44 (2005), pp. 1289–1321.
- [11] F. BOETIUS AND M. KOHLMANN, *Connections between optimal stopping and singular stochastic control*, Stochastic Process. Appl., 77 (1998), pp. 253–281.
- [12] K. A. BREKKE AND B. ØKSENDAL, *Optimal switching in an economic activity under uncertainty*, SIAM J. Control Optim., 32 (1994), pp. 1021–1036.
- [13] M. B. CHIAROLLA AND U. G. HAUSSMANN, *Explicit solution of a stochastic, irreversible investment problem and its moving threshold*, Math. Oper. Res., 30 (2005), pp. 91–108.
- [14] M. H. A. DAVIS, M. A. H. DEMPSTER, S. P. SETHI, AND D. VERMES, *Optimal capacity expansion under uncertainty*, Adv. in Appl. Probab., 19 (1987), pp. 156–176.
- [15] A. K. DIXIT AND R. S. PINDYCK, *Investment under Uncertainty*, Princeton University Press, Princeton, NJ, 1994.
- [16] J. K. DUCKWORTH AND M. ZERVOS, *An investment model with entry and exit decisions*, J. Appl. Probab., 37 (2000), pp. 547–559.
- [17] N. EL KAROUI AND I. KARATZAS, *Probabilistic aspects of finite-fuel, reflected follower problems*, Acta Appl. Math., 11 (1988), pp. 223–258.
- [18] N. EL KAROUI AND I. KARATZAS, *Integration of the optimal risk in a stopping problem with absorption*, in Séminaire de Probabilités, XXIII, Lecture Notes in Math. 1372, Springer, Berlin, 1989, pp. 405–420.
- [19] N. EL KAROUI AND I. KARATZAS, *A new approach to the Skorohod problem, and its applications*, Stoch. Stoch. Rep., 34 (1991), pp. 57–82.
- [20] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.
- [21] X. GUO AND H. PHAM, *Optimal partially reversible investment with entry decision and general production function*, Stochastic Process. Appl., 115 (2005), pp. 705–736.
- [22] S. HAMADÈNE AND M. HASSANI, *BSDEs with two reflecting barriers driven by a Brownian and a Poisson noise and related Dynkin game*, Electron. J. Probab., 11 (2006), pp. 121–145.
- [23] S. HAMADÈNE AND M. JEANBLANC, *On the starting and stopping problem: Application in reversible investments*, Math. Oper. Res., 32 (2007), pp. 182–192.
- [24] J. M. HARRISON AND M. I. TAKSAR, *Instantaneous control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 439–453.
- [25] I. KARATZAS, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.
- [26] I. KARATZAS, *Probabilistic aspects of finite-fuel stochastic control*, Proc. Natl. Acad. Sci. USA, 82 (1985), pp. 5579–5581.
- [27] I. KARATZAS AND S. E. SHREVE, *Connections between optimal stopping and singular stochastic control I. Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.
- [28] I. KARATZAS AND S. E. SHREVE, *Connections between optimal stopping and singular stochastic control II. Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–451.
- [29] I. KARATZAS AND S. E. SHREVE, *Equivalent models for finite-fuel stochastic control*, Stochastics, 18 (1986), pp. 245–276.
- [30] I. KARATZAS AND H. WANG, *Connections between bounded variation control and Dynkin games*, in Optimal Control and Partial Differential Equations (volume in honor of A. Bensoussan), J. Menaldi, E. Rofman, and A. Sulem, eds., Institute of Science Press, Amsterdam, 2001, pp. 363–373.
- [31] T. Ø. KOBILA, *A class of solvable stochastic investment problems involving singular controls*, Stoch. Stoch. Rep., 43 (1993), pp. 29–63.
- [32] V. LY VATH AND H. PHAM, *Explicit solution to an optimal switching problem in the two-regime case*, SIAM J. Control Optim., 46 (2007), pp. 395–426.
- [33] A. MERHI AND M. ZERVOS, *A model for reversible investment capacity expansion*, SIAM J. Control Optim., 46 (2007), pp. 839–876.
- [34] A. ØKSENDAL, *Irreversible investment problems*, Finance Stoch., 4 (2000), pp. 223–250.
- [35] J. A. SCHEINKMAN AND T. ZARIPHPOULOU, *Optimal environmental management in the presence of irreversibilities*, J. Econom. Theory, 96 (2001), pp. 180–207.
- [36] H. WANG, *Capacity expansion with exponential jump diffusion processes*, Stoch. Stoch. Rep., 75 (2003), pp. 259–274.

# STOCHASTIC DIFFERENTIAL GAMES AND VISCOSITY SOLUTIONS OF HAMILTON–JACOBI–BELLMAN–ISAACS EQUATIONS\*

RAINER BUCKDAHN<sup>†</sup> AND JUAN LI<sup>‡</sup>

**Abstract.** In this paper we study zero-sum two-player stochastic differential games with the help of the theory of backward stochastic differential equations (BSDEs). More precisely, we generalize the results of the pioneering work of Fleming and Souganidis [*Indiana Univ. Math. J.*, 38 (1989), pp. 293–314] by considering cost functionals defined by controlled BSDEs and by allowing the admissible control processes to depend on events occurring before the beginning of the game. This extension of the class of admissible control processes has the consequence that the cost functionals become random variables. However, by making use of a Girsanov transformation argument, which is new in this context, we prove that the upper and the lower value functions of the game remain deterministic. Apart from the fact that this extension of the class of admissible control processes is quite natural and reflects the behavior of the players who always use the maximum of available information, its combination with BSDE methods, in particular that of the notion of stochastic “backward semigroups” introduced by Peng [*BSDE and stochastic optimizations*, in *Topics in Stochastic Analysis*, Science Press, Beijing, 1997], allows us then to prove a dynamic programming principle for both the upper and the lower value functions of the game in a straightforward way. The upper and the lower value functions are then shown to be the unique viscosity solutions of the upper and the lower Hamilton–Jacobi–Bellman–Isaacs equations, respectively. For this Peng’s BSDE method is extended from the framework of stochastic control theory into that of stochastic differential games.

**Key words.** stochastic differential games, value function, backward stochastic differential equations, dynamic programming principle, viscosity solution

**AMS subject classifications.** 93E05, 90C39

**DOI.** 10.1137/060671954

**1. Introduction.** With their pioneering paper of 1989 Fleming and Souganidis [14] were the first to study in a rigorous manner two-player zero-sum stochastic differential games and to prove that the lower and the upper value functions of such games satisfy the dynamic programming principle and that they are the unique viscosity solutions of the associated Bellman–Isaacs equations and coincide under the Isaacs condition. Their work has translated former results on differential games by Isaacs [23], Friedman [15], and, in particular, Evans and Souganidis [13] from the purely deterministic into the stochastic framework and has given an important impulse for the research in the theory of stochastic differential games. The paper of Fleming and Souganidis [14] has been the starting point for a lot of works using their approach and translating it into new contexts. So, for instance, Buckdahn, Cardaliaguet, and Rainer [6] adapted the methods of [14] in order to prove the existence of Nash equilibrium points for stochastic nonzero-sum differential games and to characterize them.

---

\*Received by the editors October 10, 2006; accepted for publication (in revised form) July 18, 2007; published electronically February 1, 2008.

<http://www.siam.org/journals/sicon/47-1/67195/html>

<sup>†</sup>Département de Mathématiques, Université de Bretagne Occidentale, 6, avenue Victor-le-Gorgeu, B.P. 809, 29285 Brest cedex, France (Rainer.Buckdahn@univ-brest.fr).

<sup>‡</sup>School of Mathematical Sciences, Fudan University, Shanghai 200433, People’s Republic of China, and Department of Mathematics, Shandong University at Weihai, Weihai 264200, People’s Republic of China (juanli@sdu.edu.cn). The work of this author has been supported by a one-year fellowship awarded by the General Council of Finistère, France, the NSF of the People’s Republic of China (10426022; 10371067), Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT), and National Basic Research Program of China (973 Program (2007CB814904)).

Another direction for generalization was chosen by Bayraktar and Poor [4]: They studied stochastic differential games whose modulating process is a fractional Brownian motion, while Browne [5] considered stochastic dynamic investment games in continuous time, provided conditions under which a general payoff function has an achievable value, and gave, under these conditions, an explicit representation for the value and resulting equilibrium strategies. Concerning optimal stopping games the interested reader is referred to the work of Ekström and Peskir [11], of Karatzas and Sudderth [21], as well as of Karatzas and Zamfirescu [22]. Finally, the reader more interested in the subject of stochastic differential games is also referred to the references given in [14].

The present work also investigates two-player zero-sum stochastic differential games, but with two main differences to the setting chosen by Fleming and Souganidis [14] and the other papers mentioned above: On the one hand, we allow our admissible control processes to depend on the full past of the trajectories of the driving Brownian motion; this means, in particular, that they can also depend on information occurring before the beginning of the game (which has the consequence that the cost functionals become random variables); on the other hand, we consider a more general running cost functional, which implies that the cost functionals will be given by a backward stochastic differential equation (BSDE). Both of these extensions of the framework in [14] are crucial because they allow one to harmonize the setting for stochastic differential games with that for the stochastic control theory and to simplify considerably the approach in [14] by using BSDE methods.

BSDEs in their general nonlinear form were introduced by Pardoux and Peng [24] in 1990. They have been studied since then by a lot of authors and have found various applications, namely, in stochastic control, finance, and the second order PDE theory. BSDE methods, originally developed by Peng [26], [27] for the stochastic control theory, have been introduced in the theory of stochastic differential games by Hamadene and Lepeltier [16] and Hamadene, Lepeltier, and Peng [17] to study games with a dynamics whose diffusion coefficient is strictly elliptic and does not depend on the controls. In our present work there is not any such restriction on the diffusion coefficient and the application of BSDE methods; in particular, the notion of stochastic backward semigroups (Peng [26]) allows us to prove the dynamic programming principle for the upper and lower value functions of the game in a very straightforward way (i.e., in particular without making use of  $r$ -strategies (see Definition 1.7 in [14]) and  $\pi$ -admissible strategies (see Definition 2.2 in [14]) playing an essential role in [14]) and to derive from it with the help of Peng's method (see [26], [27]) the associated Bellman–Isaacs equations.

The dynamics of the stochastic differential game we investigate is given by the controlled stochastic differential equation

$$(1.1) \quad \begin{cases} dX_s^{t,x;u,v} &= b(s, X_s^{t,x;u,v}, u_s, v_s)ds + \sigma(s, X_s^{t,x;u,v}, u_s, v_s)dB_s, \\ X_t^{t,x;u,v} &= x (\in \mathbb{R}^n), \end{cases} \quad s \in [t, T],$$

where  $T > 0$  is an arbitrarily fixed finite time horizon,  $B = (B_s)_{s \in [0, T]}$  is a  $d$ -dimensional standard Brownian motion, and  $u = (u_s)_{s \in [t, T]}$ ,  $v = (v_s)_{s \in [t, T]}$  are progressively measurable with respect to the Brownian filtration and take their values in some compact metric spaces  $U$  and  $V$ , respectively (we will say that  $u \in \mathcal{U}_{t, T}$ ,  $v \in \mathcal{V}_{t, T}$ ). Precise assumptions on the coefficients  $b : [0, T] \times \mathbb{R}^n \times U \times V \rightarrow \mathbb{R}^n$  and  $\sigma : [0, T] \times \mathbb{R}^n \times U \times V \rightarrow \mathbb{R}^{n \times d}$  are given in the next section.

The cost functional (interpreted as a payoff for player I and as a cost for player

II) is introduced by a BSDE:

$$(1.2) \quad \begin{cases} -dY_s^{t,x;u,v} &= f(s, X_s^{t,x;u,v}, Y_s^{t,x;u,v}, Z_s^{t,x;u,v}, u_s, v_s)ds - Z_s^{t,x;u,v}dB_s, \\ Y_T^{t,x;u,v} &= \Phi(X_T^{t,x;u,v}), \end{cases} \quad s \in [t, T],$$

where the driver  $f : [0, T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \times U \times V \rightarrow \mathbb{R}$  describes the running cost and  $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$  the terminal cost. Under the assumptions on  $f$  and  $\Phi$  that will be introduced in the next section, the above BSDE has a unique solution  $(Y_s^{t,x;u,v}, Z_s^{t,x;u,v})_{s \in [t, T]}$ , and the cost functional is given by

$$(1.3) \quad J(t, x; u, v) = Y_t^{t,x;u,v}.$$

As usual in the differential game theory, the players cannot restrict to play only control processes, and one player has to fix a strategy while the other player chooses the best answer to this strategy in the form of a control process. A strategy admissible for player I (resp., player II) is a nonanticipating mapping  $\alpha : \mathcal{V}_{t,T} \rightarrow \mathcal{U}_{t,T}$  (resp.,  $\beta : \mathcal{U}_{t,T} \rightarrow \mathcal{V}_{t,T}$ ) which associates every admissible control of the other player with one of his own admissible controls (we write:  $\alpha \in \mathcal{A}_{t,T}, \beta \in \mathcal{B}_{t,T}$ ; the precise definitions can be found in section 3). We define the lower value function of our stochastic differential game as follows:

$$(1.4) \quad W(t, x) := \operatorname{essinf}_{\beta \in \mathcal{B}_{t,T}} \operatorname{esssup}_{u \in \mathcal{U}_{t,T}} J(t, x; u, \beta(u)),$$

and the upper value function is given by

$$(1.5) \quad U(t, x) := \operatorname{esssup}_{\alpha \in \mathcal{A}_{t,T}} \operatorname{essinf}_{v \in \mathcal{V}_{t,T}} J(t, x; \alpha(v), v).$$

The objective of our paper is to investigate these lower and upper value functions. The main results of the paper state that  $W$  and  $U$  are deterministic (Proposition 3.3) continuous viscosity solutions of the Bellman–Isaacs equations (Theorem 4.2)

$$(1.6) \quad \begin{cases} \frac{\partial}{\partial t} W(t, x) + H^-(t, x, W, DW, D^2W) = 0, & (t, x) \in [0, T] \times \mathbb{R}^n, \\ W(T, x) = \Phi(x), & x \in \mathbb{R}^n, \end{cases}$$

and

$$(1.7) \quad \begin{cases} \frac{\partial}{\partial t} U(t, x) + H^+(t, x, U, DU, D^2U) = 0, & (t, x) \in [0, T] \times \mathbb{R}^n, \\ U(T, x) = \Phi(x), & x \in \mathbb{R}^n, \end{cases}$$

respectively, associated with the Hamiltonians

$$H^-(t, x, y, p, X) = \sup_{u \in U} \inf_{v \in V} H(t, x, y, p, X, u, v),$$

$$H^+(t, x, y, p, X) = \inf_{v \in V} \sup_{u \in U} H(t, x, y, p, X, u, v),$$

$(t, x, y, p, X) \in [0, T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times S^n$  (recall that  $S^n$  denotes the set of all  $n \times n$  symmetric matrices), where

$$(1.8) \quad \begin{aligned} H(t, x, y, p, X, u, v) &= 1/2 \cdot \operatorname{tr}(\sigma \sigma^T(t, x, u, v)X) \\ &+ p \cdot b(t, x, u, v) + f(t, x, y, p \cdot \sigma(t, x, u, v), u, v). \end{aligned}$$

Moreover, we prove the uniqueness (Theorem 5.3) in a class of continuous functions with a growth condition which was introduced by Barles, Buckdahn, and Pardoux [3] and is weaker than the polynomial growth assumption. Fleming and Souganidis [14] obtained the above equations for stochastic differential games with the running cost  $f(t, x, y, z, u, v) = f(t, x, u, v)$ . Let us also mention that Cheridito et al. [7] discussed Bellman–Isaacs equations with  $f(t, x, y, z, u, v) = \alpha(t, x, u, v) + \beta(t, x, u, v)y$  as an example of nonlinear PDEs to which they gave a stochastic control interpretation in terms of so-called 2-BSDEs and auxiliary stochastic target problems.

Notice the fact that  $W$  and  $U$ , introduced as a combination of essential infimum and essential supremum over a class of random variables, are deterministic is far from being trivial. The method developed by Peng [26], [27] (see also Theorem A.2 of the present paper) for value functions involving only control processes but not strategies does not apply here since the strategies from  $\mathcal{A}_{t,T}$  and  $\mathcal{B}_{t,T}$  do not have, in general, any continuity property. To overcome this difficulty we show in Proposition 3.3 and Lemma 3.4 that  $W$  and  $U$  are invariant under Girsanov transformation and use the fact that a functional of the Brownian motion which is invariant under Girsanov transformation into all directions of the Cameron–Martin space must be deterministic. We emphasize that the proof of Lemma 3.4 does not use BSDE methods which makes this method also applicable to other situations, such as standard stochastic control problems.

Our paper is organized as follows. Section 2 and Appendix A recall some elements of the theory of backward SDEs and forward-backward SDEs which will be needed in what follows. Section 3 introduces the setting of the stochastic differential game and, in particular, the notion of the lower and the upper value functions (see (1.4) and (1.5)). Both functions  $W$  and  $U$ , a priori random fields, turn out to be deterministic functions (Proposition 3.3) which are Lipschitz in  $x$  (Lemma 3.5) and  $\frac{1}{2}$ -Hölder continuous in  $t$  (Theorem 3.10). Moreover, they satisfy the dynamic programming principle (DPP) (Theorem 3.6). The DPP allows us to derive in section 4 with the help of Peng’s method that  $W$  and  $U$  are viscosity solutions of the associated Bellman–Isaacs equations (Theorem 4.2). The uniqueness is studied in section 5. The main result of this section is a comparison result which compares viscosity sub- and supersolutions of Bellman–Isaacs equations (Theorem 5.3). This theorem not only allows us to characterize  $W$  and  $U$  as unique viscosity solutions of associated Bellman–Isaacs equations but also to show that, under the Isaacs condition,  $W$  and  $U$  coincide (one says that the game has a value; see Remark 5.2), while, in general, without the Isaacs condition we have only  $W \leq U$  (see Remark 5.1). Finally, under the more restrictive assumptions of Fleming and Souganidis, we identify  $W$  and  $U$  with the value functions defined by them in [14]; see Remark 5.3.

**2. Preliminaries.** Let us first introduce the setting in which we want to study stochastic differential games. The probability space on which we work is the classical Wiener space  $(\Omega, \mathcal{F}, P)$ , and the driving Brownian motion  $B$  will be the coordinate process on  $\Omega$ . Let us be more precise:  $\Omega$  is the set of continuous functions from  $[0, T]$  to  $\mathbb{R}^d$  starting from 0 ( $\Omega = C_0([0, T]; \mathbb{R}^d)$ ),  $\mathcal{F}$  is the Borel  $\sigma$ -algebra over  $\Omega$ , completed with respect to the Wiener measure  $P$  on this space, and  $B$  denotes the coordinate process:  $B_s(\omega) = \omega_s$ ,  $s \in [0, T]$ ,  $\omega \in \Omega$ . By  $\mathbb{F} = \{\mathcal{F}_s, 0 \leq s \leq T\}$  we denote the natural filtration generated by  $\{B_s\}_{0 \leq s \leq T}$  and augmented by all  $P$ -null sets, i.e.,

$$\mathcal{F}_s = \sigma\{B_r, r \leq s\} \vee \mathcal{N}_P, \quad s \in [0, T],$$

where  $\mathcal{N}_P$  is the set of all  $P$ -null subsets and  $T > 0$  a fixed real time horizon. For any  $n \geq 1$ ,  $|z|$  denotes the Euclidean norm of  $z \in \mathbb{R}^n$ . We also shall introduce the

following spaces of processes which will be used frequently in what follows:

$$\mathcal{S}^2(0, T; \mathbb{R}) := \{(\psi_t)_{0 \leq t \leq T} \text{ real-valued adapted càdlàg process} :$$

$$E[\sup_{0 \leq t \leq T} |\psi_t|^2] < +\infty\};$$

$$\mathcal{H}^2(0, T; \mathbb{R}^n) := \{(\psi_t)_{0 \leq t \leq T} \text{ } \mathbb{R}^n\text{-valued progressively measurable process} :$$

$$\|\psi\|_2^2 = E[\int_0^T |\psi_t|^2 dt] < +\infty\}.$$

Let us now consider a function  $g : \Omega \times [0, T] \times \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$  with the property that  $(g(t, y, z))_{t \in [0, T]}$  is progressively measurable for each  $(y, z)$  in  $\mathbb{R} \times \mathbb{R}^d$ , and we also make the following assumptions on  $g$  throughout the paper:

(A1) There exists a constant  $C \geq 0$  such that,  $P$ -a.s., for all  $t \in [0, T]$ ,  $y_1, y_2 \in \mathbb{R}$ ,  $z_1, z_2 \in \mathbb{R}^d$ ,

$$|g(t, y_1, z_1) - g(t, y_2, z_2)| \leq C(|y_1 - y_2| + |z_1 - z_2|).$$

(A2)  $g(\cdot, 0, 0) \in \mathcal{H}^2(0, T; \mathbb{R})$ .

The following result on BSDEs is by now well known; for its proof the reader is referred to Pardoux and Peng [24].

LEMMA 2.1. *Under the assumptions (A1) and (A2), for any random variable  $\xi \in L^2(\Omega, \mathcal{F}_T, P)$ , the BSDE*

$$(2.1) \quad y_t = \xi + \int_t^T g(s, y_s, z_s) ds - \int_t^T z_s dB_s, \quad 0 \leq t \leq T,$$

has a unique adapted solution

$$(y_t^{T, g, \xi}, z_t^{T, g, \xi})_{t \in [0, T]} \in \mathcal{S}^2(0, T; \mathbb{R}) \times \mathcal{H}^2(0, T; \mathbb{R}^d).$$

In what follows, we always assume that the driving coefficient  $g$  of a BSDE satisfies (A1) and (A2).

Let us remark that Lemma 2.1 remains true when assumption (A1) is replaced by weaker assumptions, for instance those studied in Bahlali [1], Bahlali et al. [2], or Pardoux and Peng [25]. However, here, for the sake of simplicity of the calculus we prefer to work with the Lipschitz assumption.

We also shall recall the following basic results on BSDEs. We begin with the well-known comparison theorem (see Theorem 2.2 in El Karoui, Peng, and Quenez [12]).

LEMMA 2.2 (comparison theorem). *Given two coefficients  $g_1$  and  $g_2$  satisfying (A1) and (A2) and two terminal values  $\xi_1, \xi_2 \in L^2(\Omega, \mathcal{F}_T, P)$ , we denote by  $(y^1, z^1)$  and  $(y^2, z^2)$  the solution of BSDE with the data  $(\xi_1, g_1)$  and  $(\xi_2, g_2)$ , respectively. Then we have:*

(i) (Monotonicity). *If  $\xi_1 \geq \xi_2$  and  $g_1 \geq g_2$ , a.s., then  $y_t^1 \geq y_t^2$ , a.s., for all  $t \in [0, T]$ .*

(ii) (Strict monotonicity). *If, in addition to (i), we also assume that  $P(\xi_1 > \xi_2) > 0$ , then  $P\{y_t^1 > y_t^2\} > 0$ ,  $0 \leq t \leq T$ , and, in particular,  $y_0^1 > y_0^2$ .*

Using the notation introduced in Lemma 2.2 we now suppose that, for some  $g : \Omega \times [0, T] \times \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$  satisfying (A1) and (A2) and for some  $i \in \{1, 2\}$ , the drivers  $g_i$ ,  $i = 1, 2$ , are of the form

$$g_i(s, y_s^i, z_s^i) = g(s, y_s^i, z_s^i) + \varphi_i(s), \quad \text{dsdP-a.e., } i = 1, 2,$$

where  $\varphi_i \in \mathcal{H}^2(0, T; \mathbb{R})$ ,  $i = 1, 2$ . Then, for terminal values  $\xi_1, \xi_2$  belonging to  $L^2(\Omega, \mathcal{F}_T, P)$  we have the following.

LEMMA 2.3. *The difference of the solutions  $(y^1, z^1)$  and  $(y^2, z^2)$  of BSDE with the data  $(\xi_1, g_1)$  and  $(\xi_2, g_2)$ , respectively, satisfies the following estimate:*

$$\begin{aligned} & |y_t^1 - y_t^2|^2 + \frac{1}{2} E \left[ \int_t^T e^{\beta(s-t)} [|y_s^1 - y_s^2|^2 + |z_s^1 - z_s^2|^2] ds | \mathcal{F}_t \right] \\ & \leq E[e^{\beta(T-t)} |\xi_1 - \xi_2|^2 | \mathcal{F}_t] + E \left[ \int_t^T e^{\beta(s-t)} |\varphi_1(s) - \varphi_2(s)|^2 ds | \mathcal{F}_t \right], \\ & \quad P\text{-a.s., for all } 0 \leq t \leq T, \end{aligned}$$

where  $\beta = 16(1 + C^2)$ .

*Proof.* For the proof the reader is referred to Proposition 2.1 in El Karoui, Peng, and Quenez [12] or Theorem 2.3 in Peng [26].  $\square$

**3. Stochastic differential games and associated dynamic programming principles.** Now we can study our stochastic differential games. The set of admissible control processes  $\mathcal{U}$  (resp.,  $\mathcal{V}$ ) for the first (resp., second) player is the set of all  $U$  (resp.,  $V$ )-valued  $\mathcal{F}_t$ -progressively measurable processes. The control state spaces  $U$  and  $V$  are supposed to be compact metric spaces.

For given admissible controls  $u(\cdot) \in \mathcal{U}$  and  $v(\cdot) \in \mathcal{V}$ , the according orbit which regards  $t$  as the initial time and  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$  as the initial state is defined by the solution of the following SDE:

$$(3.1) \quad \begin{cases} dX_s^{t, \zeta; u, v} &= b(s, X_s^{t, \zeta; u, v}, u_s, v_s) ds + \sigma(s, X_s^{t, \zeta; u, v}, u_s, v_s) dB_s, \quad s \in [t, T], \\ X_t^{t, \zeta; u, v} &= \zeta, \end{cases}$$

where the mappings

$$b : [0, T] \times \mathbb{R}^n \times U \times V \rightarrow \mathbb{R}^n \quad \text{and} \quad \sigma : [0, T] \times \mathbb{R}^n \times U \times V \rightarrow \mathbb{R}^{n \times d}$$

satisfy the following conditions:

(H3.1)

- (i) For every fixed  $x \in \mathbb{R}^n$ ,  $b(\cdot, x, \cdot, \cdot)$  and  $\sigma(\cdot, x, \cdot, \cdot)$  are continuous in  $(t, u, v)$ ;
- (ii) there exists a  $C > 0$  such that, for all  $t \in [0, T]$ ,  $x, x' \in \mathbb{R}^n$ ,  $u \in U$ ,  $v \in V$ ,  $|b(t, x, u, v) - b(t, x', u, v)| + |\sigma(t, x, u, v) - \sigma(t, x', u, v)| \leq C|x - x'|$ .

From (H3.1) we can get the global linear growth conditions of  $b$  and  $\sigma$ , i.e., the existence of some  $C > 0$  such that, for all  $0 \leq t \leq T$ ,  $u \in U$ ,  $v \in V$ ,  $x \in \mathbb{R}^n$ ,

$$(3.2) \quad |b(t, x, u, v)| + |\sigma(t, x, u, v)| \leq C(1 + |x|).$$

It follows from (A.2) in the Appendix that, under the above assumptions, for any  $u(\cdot) \in \mathcal{U}$  and  $v(\cdot) \in \mathcal{V}$ , SDE (3.1) has a unique strong solution. Moreover, for any  $p \geq 2$ , there exists  $C_p \in \mathbb{R}$  such that, for any  $t \in [0, T]$ ,  $u(\cdot) \in \mathcal{U}$ ,  $v(\cdot) \in \mathcal{V}$ , and  $\zeta, \zeta' \in L^p(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , we also have the following estimates,  $P$ -a.s.:

$$(3.3) \quad \begin{aligned} E \left[ \sup_{s \in [t, T]} |X_s^{t, \zeta; u, v} - X_s^{t, \zeta'; u, v}|^p | \mathcal{F}_t \right] &\leq C_p |\zeta - \zeta'|^p, \\ E \left[ \sup_{s \in [t, T]} |X_s^{t, \zeta; u, v}|^p | \mathcal{F}_t \right] &\leq C_p (1 + |\zeta|^p). \end{aligned}$$

The constant  $C_p$  depends only on the Lipschitz and the linear growth constants of  $b$  and  $\sigma$  with respect to  $x$ .

Let now be given two functions

$$\Phi : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f : [0, T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \times U \times V \rightarrow \mathbb{R}$$

that satisfy the following conditions:

(H3.2)

- (i) For every fixed  $(x, y, z) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d$ ,  $f(\cdot, x, y, z, \cdot, \cdot)$  is continuous in  $(t, u, v)$ , and there exists a constant  $C > 0$  such that, for all  $t \in [0, T]$ ,  $x, x' \in \mathbb{R}^n$ ,  $y, y' \in \mathbb{R}$ ,  $z, z' \in \mathbb{R}^d$ ,  $u \in U$ , and  $v \in V$ ,

$$|f(t, x, y, z, u, v) - f(t, x', y', z', u, v)| \leq C(|x - x'| + |y - y'| + |z - z'|);$$

- (ii) there is a constant  $C > 0$  such that, for all  $x, x' \in \mathbb{R}^n$ ,

$$|\Phi(x) - \Phi(x')| \leq C|x - x'|.$$

From (H3.2) we see that  $f$  and  $\Phi$  also satisfy the global linear growth condition in  $x$ , i.e., there exists some  $C > 0$  such that, for all  $0 \leq t \leq T$ ,  $u \in U$ ,  $v \in V$ , and  $x \in \mathbb{R}^n$ ,

$$(3.4) \quad |f(t, x, 0, 0, u, v)| + |\Phi(x)| \leq C(1 + |x|).$$

For any  $u(\cdot) \in \mathcal{U}$ ,  $v(\cdot) \in \mathcal{V}$ , and  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , the mappings  $\xi := \Phi(X_T^{t, \zeta; u, v})$  and  $g(s, y, z) := f(s, X_s^{t, \zeta; u, v}, y, z, u_s, v_s)$  satisfy the conditions of Lemma 2.1 on the interval  $[t, T]$ . Therefore, there exists a unique solution to the following BSDE:

$$(3.5) \quad \begin{cases} -dY_s^{t, \zeta; u, v} &= f(s, X_s^{t, \zeta; u, v}, Y_s^{t, \zeta; u, v}, Z_s^{t, \zeta; u, v}, u_s, v_s)ds - Z_s^{t, \zeta; u, v}dB_s, \\ Y_T^{t, \zeta; u, v} &= \Phi(X_T^{t, \zeta; u, v}), \end{cases}$$

where  $X^{t, \zeta; u, v}$  is introduced by (3.1).

Moreover, in analogy to Proposition A.1, we can see that there exists some constant  $C > 0$  such that, for all  $0 \leq t \leq T$ ,  $\zeta, \zeta' \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ ,  $u(\cdot) \in \mathcal{U}$ , and  $v(\cdot) \in \mathcal{V}$ ,  $P$ -a.s.,

$$(3.6) \quad \begin{aligned} \text{(i)} \quad & |Y_t^{t, \zeta; u, v} - Y_t^{t, \zeta'; u, v}| \leq C|\zeta - \zeta'|; \\ \text{(ii)} \quad & |Y_t^{t, \zeta; u, v}| \leq C(1 + |\zeta|). \end{aligned}$$

We now introduce the following subspaces of admissible controls.

**DEFINITION 3.1.** *An admissible control process  $u = \{u_r, r \in [t, s]\}$  (resp.,  $v = \{v_r, r \in [t, s]\}$ ) for player I (resp., II) on  $[t, s]$  ( $t < s \leq T$ ) is an  $\mathcal{F}_r$ -progressively measurable process taking values in  $U$  (resp.,  $V$ ). The set of all admissible controls for player I (resp., II) on  $[t, s]$  is denoted by  $\mathcal{U}_{t,s}$  (resp.,  $\mathcal{V}_{t,s}$ ). We identify two processes  $u$  and  $\bar{u}$  in  $\mathcal{U}_{t,s}$  and write  $u \equiv \bar{u}$  on  $[t, s]$ , if  $P\{u = \bar{u} \text{ a.e. in } [t, s]\} = 1$ . Similarly we interpret  $v \equiv \bar{v}$  on  $[t, s]$  in  $\mathcal{V}_{t,s}$ .*

Finally, we have still to define the admissible strategies for the game.

**DEFINITION 3.2.** *A nonanticipative strategy for player I on  $[t, s]$  ( $t < s \leq T$ ) is a mapping  $\alpha : \mathcal{V}_{t,s} \rightarrow \mathcal{U}_{t,s}$  such that, for any  $\mathbb{F}$ -stopping time  $S : \Omega \rightarrow [t, s]$  and any  $v_1, v_2 \in \mathcal{V}_{t,s}$ , with  $v_1 \equiv v_2$  on  $\llbracket t, S \rrbracket$ , it holds that  $\alpha(v_1) \equiv \alpha(v_2)$  on  $\llbracket t, S \rrbracket$ . Nonanticipative strategies for player II on  $[t, s]$ ,  $\beta : \mathcal{U}_{t,s} \rightarrow \mathcal{V}_{t,s}$ , are defined similarly. The set of all nonanticipative strategies  $\alpha : \mathcal{V}_{t,s} \rightarrow \mathcal{U}_{t,s}$  for player I on  $[t, s]$  is denoted by  $\mathcal{A}_{t,s}$ . The set of all nonanticipative strategies  $\beta : \mathcal{U}_{t,s} \rightarrow \mathcal{V}_{t,s}$  for player II on  $[t, s]$  is denoted by  $\mathcal{B}_{t,s}$ . (Recall that  $\llbracket t, S \rrbracket = \{(r, \omega) \in [0, T] \times \Omega, t \leq r \leq S(\omega)\}$ .)*



Given the control processes  $u(\cdot) \in \mathcal{U}_{t,T}$  and  $v(\cdot) \in \mathcal{V}_{t,T}$  we introduce the following associated cost functional:

$$(3.7) \quad J(t, x; u, v) := Y_t^{t,x;u,v}, \quad (t, x) \in [0, T] \times \mathbb{R}^n,$$

where the process  $Y^{t,x;u,v}$  is defined by BSDE (3.5).

Similarly to the proof of Theorem A.2 we can get that, for any  $t \in [0, T]$  and  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ ,

$$(3.8) \quad J(t, \zeta; u, v) = Y_t^{t,\zeta;u,v}, \quad P\text{-a.s.}$$

Being particularly interested in the case of a deterministic  $\zeta$ , i.e.,  $\zeta = x \in \mathbb{R}^n$ , we define the lower value function of our stochastic differential game

$$(3.9) \quad W(t, x) := \operatorname{ess\,inf}_{\beta \in \mathcal{B}_{t,T}} \operatorname{ess\,sup}_{u \in \mathcal{U}_{t,T}} J(t, x; u, \beta(u))$$

and its upper value function

$$(3.10) \quad U(t, x) := \operatorname{ess\,sup}_{\alpha \in \mathcal{A}_{t,T}} \operatorname{ess\,inf}_{v \in \mathcal{V}_{t,T}} J(t, x; \alpha(v), v).$$

The names “lower value function” and “upper value function” for  $W$  and  $U$ , respectively, are justified later by Remark 5.1.

*Remark 3.1.* (1) Here the essential infimum and the essential supremum should be understood as one with respect to indexed families of random variables (see, e.g., Dunford and Schwartz [10], Dellacherie [9], or the appendix in Karatzas and Shreve [20] for detailed discussions). For the convenience of the reader we recall the notion of  $\operatorname{ess\,inf}$  of processes. Given a family of real-valued random variables  $\eta_\alpha$ ,  $\alpha \in I$ , a random variable  $\eta$  is said to be  $\operatorname{ess\,inf}_{\alpha \in I} \eta_\alpha$ , if

- (i)  $\eta \leq \eta_\alpha$ ,  $P$ -a.s., for any  $\alpha \in I$ ;
- (ii) if there is another random variable  $\xi$  such that  $\xi \leq \eta_\alpha$ ,  $P$ -a.s., for any  $\alpha \in I$ , then  $\xi \leq \eta$ ,  $P$ -a.s.

The random variable  $\operatorname{ess\,sup}_{\alpha \in I} \eta_\alpha$  can be introduced now by the relation

$$\operatorname{ess\,sup}_{\alpha \in I} \eta_\alpha = -\operatorname{ess\,inf}_{\alpha \in I} (-\eta_\alpha).$$

Finally, recall that  $\operatorname{ess\,inf}_{\alpha \in I} \eta_\alpha = \inf_{n \geq 1} \eta_{\alpha_n}$  for some countable family  $(\alpha_n) \subset I$ ;  $\operatorname{ess\,sup}_{\alpha \in I} \eta_\alpha$  has the same property.

(2) Obviously, under the assumptions (H3.1)–(H3.2), the lower value function  $W(t, x)$  as well as the upper value function  $U(t, x)$  are well-defined, and a priori they both are bounded  $\mathcal{F}_t$ -measurable random variables. But it turns out that  $W(t, x)$  and  $U(t, x)$  are even deterministic. Indeed, concentrating on the study of the properties of  $W(t, x)$  (the function  $U(t, x)$  can be analyzed in the same manner), we can state the following,

**PROPOSITION 3.3.** *For any  $(t, x) \in [0, T] \times \mathbb{R}^n$ , we have  $W(t, x) = E[W(t, x)]$ ,  $P$ -a.s. By identifying  $W(t, x)$  with its deterministic version  $E[W(t, x)]$  we can consider  $W : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  as a deterministic function.*

*Remark 3.2.* Recall that the fact that the lower and the upper value functions defined by Fleming and Souganidis [14] are deterministic is an immediate consequence of their definition. Indeed, for a game over the time interval  $[t, T]$  only control processes which are independent of the past  $\mathcal{F}_t$  are considered as admissible, and since the admissible strategies are supposed to associate admissible control processes of

one player with those of the other player, all of the associated cost functionals are independent of  $\mathcal{F}_t$  and hence deterministic.

*Proof.* Let  $H$  denote the Cameron–Martin space of all absolutely continuous elements  $h \in \Omega$  whose derivative  $\dot{h}$  belongs to  $L^2([0, T], \mathbb{R}^d)$ .

For any  $h \in H$ , we define the mapping  $\tau_h \omega := \omega + h$ ,  $\omega \in \Omega$ . Obviously,  $\tau_h : \Omega \rightarrow \Omega$  is a bijection, and its law is given by  $P \circ [\tau_h]^{-1} = \exp\{\int_0^T \dot{h}_s dB_s - \frac{1}{2} \int_0^T |\dot{h}_s|^2 ds\} P$ . Let  $(t, x) \in [0, T] \times \mathbb{R}^n$  be arbitrarily fixed, and put  $H_t = \{h \in H | h(\cdot) = h(\cdot \wedge t)\}$ . We split now the proof in the following steps:

First step: For any  $u \in \mathcal{U}_{t,T}$ ,  $v \in \mathcal{V}_{t,T}$ , and  $h \in H_t$ ,  $J(t, x; u, v)(\tau_h) = J(t, x; u(\tau_h), v(\tau_h))$ ,  $P$ -a.s.

Indeed, we apply the Girsanov transformation to SDE (3.1) (with  $\zeta = x$ ) and compare the obtained equation with the SDE obtained from (3.1) by substituting the transformed control processes  $u(\tau_h)$  and  $v(\tau_h)$  for  $u$  and  $v$ . Then from the uniqueness of the solution of (3.1) we get  $X_s^{t,x;u,v}(\tau_h) = X_s^{t,x;u(\tau_h),v(\tau_h)}$  for any  $s \in [t, T]$ ,  $P$ -a.s. Furthermore, by a similar Girsanov transformation argument we get from the uniqueness of the solution of BSDE (3.5)

$$Y_s^{t,x;u,v}(\tau_h) = Y_s^{t,x;u(\tau_h),v(\tau_h)} \text{ for any } s \in [t, T], \text{ } P\text{-a.s.,}$$

$$Z_s^{t,x;u,v}(\tau_h) = Z_s^{t,x;u(\tau_h),v(\tau_h)}, \text{ dsd } P\text{-a.e. on } [t, T] \times \Omega.$$

That means

$$J(t, x; u, v)(\tau_h) = J(t, x; u(\tau_h), v(\tau_h)), \text{ } P\text{-a.s.}$$

Second step: For  $\beta \in \mathcal{B}_{t,T}$ ,  $h \in H_t$ , let  $\beta^h(u) := \beta(u(\tau_{-h}))(\tau_h)$ ,  $u \in \mathcal{U}_{t,T}$ . Then  $\beta^h \in \mathcal{B}_{t,T}$ .

Obviously,  $\beta^h$  maps  $\mathcal{U}_{t,T}$  into  $\mathcal{V}_{t,T}$ . Moreover, this mapping is nonanticipating. Indeed, let  $S : \Omega \rightarrow [t, T]$  be an  $\mathbb{F}$ -stopping time and  $u_1, u_2 \in \mathcal{U}_{t,T}$ , with  $u_1 \equiv u_2$  on  $\llbracket t, S \rrbracket$ . Then, obviously,  $u_1(\tau_{-h}) \equiv u_2(\tau_{-h})$  on  $\llbracket t, S(\tau_{-h}) \rrbracket$  (notice that  $S(\tau_{-h})$  is still a stopping time), and because  $\beta \in \mathcal{B}_{t,T}$  we have  $\beta(u_1(\tau_{-h})) \equiv \beta(u_2(\tau_{-h}))$  on  $\llbracket t, S(\tau_{-h}) \rrbracket$ . Therefore,

$$\beta^h(u_1) = \beta(u_1(\tau_{-h}))(\tau_h) \equiv \beta(u_2(\tau_{-h}))(\tau_h) = \beta^h(u_2) \text{ on } \llbracket t, S \rrbracket.$$

Third step: For all  $h \in H_t$  and  $\beta \in \mathcal{B}_{t,T}$  we have

$$\{\text{esssup}_{u \in \mathcal{U}_{t,T}} J(t, x; u, \beta(u))\}(\tau_h) = \text{esssup}_{u \in \mathcal{U}_{t,T}} \{J(t, x; u, \beta(u))(\tau_h)\}, \text{ } P\text{-a.s.}$$

Indeed, with the notation  $I(t, x, \beta) := \text{esssup}_{u \in \mathcal{U}_{t,T}} J(t, x; u, \beta(u))$ ,  $\beta \in \mathcal{B}_{t,T}$ , we have  $I(t, x, \beta) \geq J(t, x; u, \beta(u))$ , and thus  $I(t, x, \beta)(\tau_h) \geq J(t, x; u, \beta(u))(\tau_h)$ ,  $P$ -a.s., for all  $u \in \mathcal{U}_{t,T}$ . On the other hand, for any random variable  $\zeta$  satisfying  $\zeta \geq J(t, x; u, \beta(u))(\tau_h)$ , and hence also  $\zeta(\tau_{-h}) \geq J(t, x; u, \beta(u))$ ,  $P$ -a.s., for all  $u \in \mathcal{U}_{t,T}$ , we have  $\zeta(\tau_{-h}) \geq I(t, x, \beta)$ ,  $P$ -a.s., i.e.,  $\zeta \geq I(t, x, \beta)(\tau_h)$ ,  $P$ -a.s. Consequently,

$$I(t, x, \beta)(\tau_h) = \text{esssup}_{u \in \mathcal{U}_{t,T}} \{J(t, x; u, \beta(u))(\tau_h)\}, \text{ } P\text{-a.s.}$$

Fourth step:  $W(t, x)$  is invariant with respect to the Girsanov transformation  $\tau_h$ , i.e.,

$$W(t, x)(\tau_h) = W(t, x), \text{ } P\text{-a.s., for any } h \in H.$$

Indeed, similarly to the third step we can show that for all  $h \in H_t$

$$\{\text{essinf}_{\beta \in \mathcal{B}_{t,T}} I(t, x; \beta)\}(\tau_h) = \text{essinf}_{\beta \in \mathcal{B}_{t,T}} \{I(t, x; \beta)(\tau_h)\}, \text{ } P\text{-a.s.}$$

Then, from the first step to the third step we have, for any  $h \in H_t$ ,

$$\begin{aligned} W(t, x)(\tau_h) &= \text{essinf}_{\beta \in \mathcal{B}_{t,T}} \text{esssup}_{u \in \mathcal{U}_{t,T}} \{J(t, x; u, \beta(u))(\tau_h)\} \\ &= \text{essinf}_{\beta \in \mathcal{B}_{t,T}} \text{esssup}_{u \in \mathcal{U}_{t,T}} J(t, x; u(\tau_h), \beta^h(u(\tau_h))) \\ &= \text{essinf}_{\beta \in \mathcal{B}_{t,T}} \text{esssup}_{u \in \mathcal{U}_{t,T}} J(t, x; u, \beta^h(u)) \\ &= \text{essinf}_{\beta \in \mathcal{B}_{t,T}} \text{esssup}_{u \in \mathcal{U}_{t,T}} J(t, x; u, \beta(u)) \\ &= W(t, x), \text{ } P\text{-a.s.}, \end{aligned}$$

where we have used  $\{u(\tau_h)|u(\cdot) \in \mathcal{U}_{t,T}\} = \mathcal{U}_{t,T}$  and  $\{\beta^h|\beta \in \mathcal{B}_{t,T}\} = \mathcal{B}_{t,T}$  in order to obtain both latter equalities. Therefore, for any  $h \in H_t$ ,  $W(t, x)(\tau_h) = W(t, x)$ ,  $P$ -a.s., and since  $W(t, x)$  is  $\mathcal{F}_t$ -measurable, we have this relation even for all  $h \in H$ . Indeed, recall that our underlying fundamental space is  $\Omega = C_0([0, T]; \mathbb{R}^d)$  and that, due to the definition of the filtration, the  $\mathcal{F}_t$ -measurable random variable  $W(t, x)(\omega)$ ,  $\omega \in \Omega$ , depends only on the restriction of  $\omega$  to the time interval  $[0, t]$ .

The result of the fourth step combined with the following auxiliary Lemma 3.4 completes the proof.  $\square$

LEMMA 3.4. *Let  $\zeta$  be a random variable defined over our classical Wiener space  $(\Omega, \mathcal{F}_T, P)$ , such that  $\zeta(\tau_h) = \zeta$ ,  $P$ -a.s., for any  $h \in H$ . Then  $\zeta = E\zeta$ ,  $P$ -a.s.*

*Proof.* Let  $h \in H$  and  $A \in \mathcal{B}(\mathbb{R})$ . Then

$$\begin{aligned} &E \left[ \mathbf{1}_{\{\zeta \in A\}} \exp \left\{ \int_0^T \dot{h}_s dB_s - \frac{1}{2} \int_0^T |\dot{h}_s|^2 ds \right\} \right] \\ &= E \left[ \mathbf{1}_{\{\zeta(\tau_{-h}) \in A\}} \exp \left\{ \int_0^T \dot{h}_s dB_s - \frac{1}{2} \int_0^T |\dot{h}_s|^2 ds \right\} \right] \\ &= E[\mathbf{1}_{\{\zeta \in A\}}], \end{aligned}$$

from which we deduce that

$$E \left[ \mathbf{1}_{\{\zeta \in A\}} \exp \left\{ \int_0^T \dot{h}_s dB_s \right\} \right] = E[\mathbf{1}_{\{\zeta \in A\}}] E \left[ \exp \left\{ \int_0^T \dot{h}_s dB_s \right\} \right],$$

i.e., for any  $\varphi \in L^2([0, T]; \mathbb{R}^d)$ ,

$$(3.11) \quad E \left[ \mathbf{1}_{\{\zeta \in A\}} \exp \left\{ \int_0^T \varphi_s dB_s \right\} \right] = E[\mathbf{1}_{\{\zeta \in A\}}] E \left[ \exp \left\{ \int_0^T \varphi_s dB_s \right\} \right].$$

Consequently, taking into consideration the arbitrariness of  $A \in \mathcal{B}(\mathbb{R})$  and of  $\varphi \in L^2([0, T]; \mathbb{R}^d)$ , the independence of  $\zeta$  of  $B$  and hence of  $\mathcal{F}_T$  follows, but this is possible only for deterministic  $\zeta$ .  $\square$

The first property of the lower value function  $W(t, x)$  which we present is an immediate consequence of (3.6) and (3.9).

LEMMA 3.5. *There exists a constant  $C > 0$  such that, for all  $0 \leq t \leq T$ ,  $x, x' \in \mathbb{R}^n$ ,*

$$(3.12) \quad \begin{aligned} \text{(i)} \quad &|W(t, x) - W(t, x')| \leq C|x - x'|; \\ \text{(ii)} \quad &|W(t, x)| \leq C(1 + |x|). \end{aligned}$$

We now discuss (the generalized) DPP for our stochastic differential game (3.1), (3.5), and (3.9). For this end we have to define the family of (backward) semigroups associated with BSDE (3.5). This notion of stochastic backward semigroups was first introduced by Peng [26] which was applied to study the DPP for stochastic control problems. Our approach adapts Peng's ideas to the framework of stochastic differential games.

Given the initial data  $(t, x)$ , a positive number  $\delta \leq T - t$ , admissible control processes  $u(\cdot) \in \mathcal{U}_{t, t+\delta}$  and  $v(\cdot) \in \mathcal{V}_{t, t+\delta}$ , and a real-valued random variable  $\eta \in L^2(\Omega, \mathcal{F}_{t+\delta}, P; \mathbb{R})$ , we put

$$(3.13) \quad G_{s, t+\delta}^{t, x; u, v}[\eta] := \tilde{Y}_s^{t, x; u, v}, \quad s \in [t, t + \delta],$$

where the couple  $(\tilde{Y}_s^{t, x; u, v}, \tilde{Z}_s^{t, x; u, v})_{t \leq s \leq t+\delta}$  is the solution of the following BSDE with the time horizon  $t + \delta$ :

$$\begin{cases} -d\tilde{Y}_s^{t, x; u, v} = f(s, X_s^{t, x; u, v}, \tilde{Y}_s^{t, x; u, v}, \tilde{Z}_s^{t, x; u, v}, u_s, v_s)ds \\ \quad - \tilde{Z}_s^{t, x; u, v} dB_s, & s \in [t, t + \delta], \\ \tilde{Y}_{t+\delta}^{t, x; u, v} = \eta, \end{cases}$$

and  $X^{t, x; u, v}$  is the solution of SDE (3.1). Then, obviously, for the solution  $(Y^{t, x; u, v}, Z^{t, x; u, v})$  of BSDE (3.5) we have

$$(3.14) \quad G_{t, T}^{t, x; u, v}[\Phi(X_T^{t, x; u, v})] = G_{t, t+\delta}^{t, x; u, v}[Y_{t+\delta}^{t, x; u, v}].$$

Moreover,

$$\begin{aligned} J(t, x; u, v) &= Y_t^{t, x; u, v} = G_{t, T}^{t, x; u, v}[\Phi(X_T^{t, x; u, v})] = G_{t, t+\delta}^{t, x; u, v}[Y_{t+\delta}^{t, x; u, v}] \\ &= G_{t, t+\delta}^{t, x; u, v}[J(t + \delta, X_{t+\delta}^{t, x; u, v}; u, v)]. \end{aligned}$$

*Remark 3.3.* When  $f$  is independent of  $(y, z)$  it holds that

$$G_{s, t+\delta}^{t, x; u, v}[\eta] = E \left[ \eta + \int_s^{t+\delta} f(r, X_r^{t, x; u, v}, u_r, v_r) dr \middle| \mathcal{F}_s \right], \quad s \in [t, t + \delta].$$

**THEOREM 3.6.** *Under the assumptions (H3.1) and (H3.2), the lower value function  $W(t, x)$  obeys the following DPP: For any  $0 \leq t < t + \delta \leq T$ ,  $x \in \mathbb{R}^n$ ,*

$$(3.15) \quad W(t, x) = \operatorname{essinf}_{\beta \in \mathcal{B}_{t, t+\delta}} \operatorname{esssup}_{u \in \mathcal{U}_{t, t+\delta}} G_{t, t+\delta}^{t, x; u, \beta(u)}[W(t + \delta, X_{t+\delta}^{t, x; u, \beta(u)})].$$

*Proof.* To simplify notations we put

$$W_\delta(t, x) = \operatorname{essinf}_{\beta \in \mathcal{B}_{t, t+\delta}} \operatorname{esssup}_{u \in \mathcal{U}_{t, t+\delta}} G_{t, t+\delta}^{t, x; u, \beta(u)}[W(t + \delta, X_{t+\delta}^{t, x; u, \beta(u)})].$$

The proof that  $W_\delta(t, x)$  coincides with  $W(t, x)$  will be split into the following lemmas, which all are supposed to satisfy (H3.1) and (H3.2).  $\square$

**LEMMA 3.7.**  *$W_\delta(t, x)$  is deterministic.*

*Proof.* The proof of this lemma uses the same ideas as that of Proposition 3.3, so it can be omitted here.  $\square$

**LEMMA 3.8.**  *$W_\delta(t, x) \leq W(t, x)$ .*

*Proof.* Let  $\beta \in \mathcal{B}_{t,T}$  be arbitrarily fixed. Then, given a  $u_2(\cdot) \in \mathcal{U}_{t+\delta,T}$ , we define as follows the restriction  $\beta_1$  of  $\beta$  to  $\mathcal{U}_{t,t+\delta}$ :

$$\beta_1(u_1) := \beta(u_1 \oplus u_2)|_{[t,t+\delta]}, \quad u_1(\cdot) \in \mathcal{U}_{t,t+\delta},$$

where  $u_1 \oplus u_2 := u_1 \mathbf{1}_{[t,t+\delta]} + u_2 \mathbf{1}_{(t+\delta,T]}$  extends  $u_1(\cdot)$  to an element of  $\mathcal{U}_{t,T}$ . It is easy to check that  $\beta_1 \in \mathcal{B}_{t,t+\delta}$ . Moreover, from the nonanticipativity property of  $\beta$  we deduce that  $\beta_1$  is independent of the special choice of  $u_2(\cdot) \in \mathcal{U}_{t+\delta,T}$ . Consequently, from the definition of  $W_\delta(t, x)$ ,

$$(3.16) \quad W_\delta(t, x) \leq \text{esssup}_{u_1 \in \mathcal{U}_{t,t+\delta}} G_{t,t+\delta}^{t,x;u_1,\beta_1(u_1)}[W(t+\delta, X_{t+\delta}^{t,x;u_1,\beta_1(u_1)})], \quad P\text{-a.s.}$$

We use the notation  $I_\delta(t, x, u, v) := G_{t,t+\delta}^{t,x;u,v}[W(t+\delta, X_{t+\delta}^{t,x;u,v})]$  and notice that there exists a sequence  $\{u_i^1, i \geq 1\} \subset \mathcal{U}_{t,t+\delta}$  such that

$$I_\delta(t, x, \beta_1) := \text{esssup}_{u_1 \in \mathcal{U}_{t,t+\delta}} I_\delta(t, x, u_1, \beta_1(u_1)) = \sup_{i \geq 1} I_\delta(t, x, u_i^1, \beta_1(u_i^1)), \quad P\text{-a.s.}$$

For any  $\varepsilon > 0$ , we put  $\tilde{\Gamma}_i := \{I_\delta(t, x, \beta_1) \leq I_\delta(t, x, u_i^1, \beta_1(u_i^1)) + \varepsilon\} \in \mathcal{F}_t$ ,  $i \geq 1$ . Then  $\Gamma_1 := \tilde{\Gamma}_1$ ,  $\Gamma_i := \tilde{\Gamma}_i \setminus (\cup_{l=1}^{i-1} \tilde{\Gamma}_l) \in \mathcal{F}_t$ ,  $i \geq 2$ , form an  $(\Omega, \mathcal{F}_t)$ -partition, and  $u_1^\varepsilon := \sum_{i \geq 1} \mathbf{1}_{\Gamma_i} u_i^1$  belongs obviously to  $\mathcal{U}_{t,t+\delta}$ . Moreover, from the nonanticipativity of  $\beta_1$  we have  $\beta_1(u_1^\varepsilon) = \sum_{i \geq 1} \mathbf{1}_{\Gamma_i} \beta_1(u_i^1)$ , and from the uniqueness of the solution of the forward-backward SDE (FBSDE), we deduce that  $I_\delta(t, x, u_1^\varepsilon, \beta_1(u_1^\varepsilon)) = \sum_{i \geq 1} \mathbf{1}_{\Gamma_i} I_\delta(t, x, u_i^1, \beta_1(u_i^1))$ ,  $P$ -a.s. Hence,

$$(3.17) \quad \begin{aligned} W_\delta(t, x) &\leq I_\delta(t, x, \beta_1) \leq \sum_{i \geq 1} \mathbf{1}_{\Gamma_i} I_\delta(t, x, u_i^1, \beta_1(u_i^1)) + \varepsilon = I_\delta(t, x, u_1^\varepsilon, \beta_1(u_1^\varepsilon)) + \varepsilon \\ &= G_{t,t+\delta}^{t,x;u_1^\varepsilon,\beta_1(u_1^\varepsilon)}[W(t+\delta, X_{t+\delta}^{t,x;u_1^\varepsilon,\beta_1(u_1^\varepsilon)})] + \varepsilon, \quad P\text{-a.s.} \end{aligned}$$

On the other hand, using the fact that  $\beta_1(\cdot) := \beta(\cdot \oplus u_2) \in \mathcal{B}_{t,t+\delta}$  does not depend on  $u_2(\cdot) \in \mathcal{U}_{t+\delta,T}$ , we can define  $\beta_2(u_2) := \beta(u_1^\varepsilon \oplus u_2)|_{[t+\delta,T]}$ , for all  $u_2(\cdot) \in \mathcal{U}_{t+\delta,T}$ . The such defined  $\beta_2 : \mathcal{U}_{t+\delta,T} \rightarrow \mathcal{V}_{t+\delta,T}$  belongs to  $\mathcal{B}_{t+\delta,T}$  since  $\beta \in \mathcal{B}_{t,T}$ . Therefore, from the definition of  $W(t+\delta, y)$  we have, for any  $y \in \mathbb{R}^n$ ,

$$W(t+\delta, y) \leq \text{esssup}_{u_2 \in \mathcal{U}_{t+\delta,T}} J(t+\delta, y; u_2, \beta_2(u_2)), \quad P\text{-a.s.}$$

Finally, because there exists a constant  $C \in \mathbb{R}$  such that

$$(3.18) \quad \begin{aligned} &\text{(i)} \quad |W(t+\delta, y) - W(t+\delta, y')| \leq C|y - y'| \text{ for any } y, y' \in \mathbb{R}^n; \\ &\text{(ii)} \quad |J(t+\delta, y, u_2, \beta_2(u_2)) - J(t+\delta, y', u_2, \beta_2(u_2))| \leq C|y - y'|, \quad P\text{-a.s.,} \\ &\quad \text{for any } u_2 \in \mathcal{U}_{t+\delta,T}, \end{aligned}$$

(see Lemma 3.5(i) and (3.6)(i)) we can show by approximating  $X_{t+\delta}^{t,x;u_1^\varepsilon,\beta_1(u_1^\varepsilon)}$  that

$$W(t+\delta, X_{t+\delta}^{t,x;u_1^\varepsilon,\beta_1(u_1^\varepsilon)}) \leq \text{esssup}_{u_2 \in \mathcal{U}_{t+\delta,T}} J(t+\delta, X_{t+\delta}^{t,x;u_1^\varepsilon,\beta_1(u_1^\varepsilon)}; u_2, \beta_2(u_2)), \quad P\text{-a.s.}$$

To estimate the right side of the latter inequality we note that there exists some sequence  $\{u_j^2, j \geq 1\} \subset \mathcal{U}_{t+\delta,T}$  such that

$$\begin{aligned} &\text{esssup}_{u_2 \in \mathcal{U}_{t+\delta,T}} J(t+\delta, X_{t+\delta}^{t,x;u_1^\varepsilon,\beta_1(u_1^\varepsilon)}; u_2, \beta_2(u_2)) \\ &= \sup_{j \geq 1} J(t+\delta, X_{t+\delta}^{t,x;u_1^\varepsilon,\beta_1(u_1^\varepsilon)}; u_j^2, \beta_2(u_j^2)), \quad P\text{-a.s.} \end{aligned}$$

Then, putting  $\tilde{\Delta}_j := \{\text{esssup}_{u_2 \in \mathcal{U}_{t+\delta, T}} J(t + \delta, X_{t+\delta}^{t, x; u_1^\varepsilon, \beta_1(u_1^\varepsilon)}; u_2, \beta_2(u_2)) \leq J(t + \delta, X_{t+\delta}^{t, x; u_1^\varepsilon, \beta_1(u_1^\varepsilon)}; u_j^2, \beta_2(u_j^2)) + \varepsilon\} \in \mathcal{F}_{t+\delta}$ ,  $j \geq 1$ ; we have with  $\Delta_1 := \tilde{\Delta}_1$ ,  $\Delta_j := \tilde{\Delta}_j \setminus (\cup_{l=1}^{j-1} \tilde{\Delta}_l) \in \mathcal{F}_{t+\delta}$ ,  $j \geq 2$ , an  $(\Omega, \mathcal{F}_{t+\delta})$ -partition and  $u_2^\varepsilon := \sum_{j \geq 1} \mathbf{1}_{\Delta_j} u_j^2 \in \mathcal{U}_{t+\delta, T}$ . From the nonanticipativity of  $\beta_2$  we have  $\beta_2(u_2^\varepsilon) = \sum_{j \geq 1} \mathbf{1}_{\Delta_j} \beta_2(u_j^2)$ , and from the definition of  $\beta_1$ ,  $\beta_2$  we know that  $\beta(u_1^\varepsilon \oplus u_2^\varepsilon) = \beta_1(u_1^\varepsilon) \oplus \beta_2(u_2^\varepsilon)$ . Thus, again from the uniqueness of the solution of our FBSDE, we get

$$\begin{aligned} J(t + \delta, X_{t+\delta}^{t, x; u_1^\varepsilon, \beta_1(u_1^\varepsilon)}; u_2^\varepsilon, \beta_2(u_2^\varepsilon)) &= Y_{t+\delta}^{t+\delta, X_{t+\delta}^{t, x; u_1^\varepsilon, \beta_1(u_1^\varepsilon)}; u_2^\varepsilon, \beta_2(u_2^\varepsilon)} \quad (\text{see (3.8)}) \\ &= \sum_{j \geq 1} \mathbf{1}_{\Delta_j} Y_{t+\delta}^{t+\delta, X_{t+\delta}^{t, x; u_1^\varepsilon, \beta_1(u_1^\varepsilon)}; u_j^2, \beta_2(u_j^2)} \\ &= \sum_{j \geq 1} \mathbf{1}_{\Delta_j} J(t + \delta, X_{t+\delta}^{t, x; u_1^\varepsilon, \beta_1(u_1^\varepsilon)}; u_j^2, \beta_2(u_j^2)), \quad P\text{-a.s.} \end{aligned}$$

Consequently,

$$\begin{aligned} (3.19) \quad W(t + \delta, X_{t+\delta}^{t, x; u_1^\varepsilon, \beta_1(u_1^\varepsilon)}) &\leq \text{esssup}_{u_2 \in \mathcal{U}_{t+\delta, T}} J(t + \delta, X_{t+\delta}^{t, x; u_1^\varepsilon, \beta_1(u_1^\varepsilon)}; u_2, \beta_2(u_2)) \\ &\leq \sum_{j \geq 1} \mathbf{1}_{\Delta_j} Y_{t+\delta}^{t, x; u_1^\varepsilon \oplus u_j^2, \beta(u_1^\varepsilon \oplus u_j^2)} + \varepsilon \\ &= Y_{t+\delta}^{t, x; u_1^\varepsilon \oplus u_2^\varepsilon, \beta(u_1^\varepsilon \oplus u_2^\varepsilon)} + \varepsilon \\ &= Y_{t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)} + \varepsilon, \quad P\text{-a.s.}, \end{aligned}$$

where  $u^\varepsilon := u_1^\varepsilon \oplus u_2^\varepsilon \in \mathcal{U}_{t, T}$ . From (3.17) and (3.19) and Lemmas 2.2 (comparison theorem for BSDEs) and 2.3, we have

$$\begin{aligned} (3.20) \quad W_\delta(t, x) &\leq G_{t, t+\delta}^{t, x; u_1^\varepsilon, \beta_1(u_1^\varepsilon)} [Y_{t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)} + \varepsilon] + \varepsilon \\ &\leq G_{t, t+\delta}^{t, x; u_1^\varepsilon, \beta_1(u_1^\varepsilon)} [Y_{t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)}] + (C + 1)\varepsilon \\ &= G_{t, t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)} [Y_{t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)}] + (C + 1)\varepsilon \\ &= Y_t^{t, x; u^\varepsilon, \beta(u^\varepsilon)} + (C + 1)\varepsilon \\ &\leq \text{esssup}_{u \in \mathcal{U}_{t, T}} Y_t^{t, x; u, \beta(u)} + (C + 1)\varepsilon, \quad P\text{-a.s.} \end{aligned}$$

Since  $\beta \in \mathcal{B}_{t, T}$  has been arbitrarily chosen we have (3.20) for all  $\beta \in \mathcal{B}_{t, T}$ . Therefore,

$$(3.21) \quad W_\delta(t, x) \leq \text{essinf}_{\beta \in \mathcal{B}_{t, T}} \text{esssup}_{u \in \mathcal{U}_{t, T}} Y_t^{t, x; u, \beta(u)} + (C + 1)\varepsilon = W(t, x) + (C + 1)\varepsilon.$$

Finally, letting  $\varepsilon \downarrow 0$ , we get  $W_\delta(t, x) \leq W(t, x)$ .  $\square$

LEMMA 3.9.  $W(t, x) \leq W_\delta(t, x)$ .

*Proof.* We continue to use the notations introduced above, and from the definition of  $W_\delta(t, x)$  we have

$$\begin{aligned} W_\delta(t, x) &= \text{essinf}_{\beta_1 \in \mathcal{B}_{t, t+\delta}} \text{esssup}_{u_1 \in \mathcal{U}_{t, t+\delta}} G_{t, t+\delta}^{t, x; u_1, \beta_1(u_1)} [W(t + \delta, X_{t+\delta}^{t, x; u_1, \beta_1(u_1)})] \\ &= \text{essinf}_{\beta_1 \in \mathcal{B}_{t, t+\delta}} I_\delta(t, x, \beta_1), \end{aligned}$$

and, for some sequence  $\{\beta_i^1, i \geq 1\} \subset \mathcal{B}_{t, t+\delta}$ ,

$$W_\delta(t, x) = \inf_{i \geq 1} I_\delta(t, x, \beta_i^1), \quad P\text{-a.s.}$$

For any  $\varepsilon > 0$ , we let  $\tilde{\Lambda}_i := \{I_\delta(t, x, \beta_i^1) - \varepsilon \leq W_\delta(t, x)\} \in \mathcal{F}_t$ ,  $i \geq 1$ ,  $\Lambda_1 := \tilde{\Lambda}_1$  and  $\Lambda_i := \tilde{\Lambda}_i \setminus (\cup_{l=1}^{i-1} \tilde{\Lambda}_l) \in \mathcal{F}_t$ ,  $i \geq 2$ . Then  $\{\Lambda_i, i \geq 1\}$  is an  $(\Omega, \mathcal{F}_t)$ -partition,  $\beta_1^\varepsilon := \sum_{i \geq 1} \mathbf{1}_{\Lambda_i} \beta_i^1$  belongs to  $\mathcal{B}_{t, t+\delta}$ , and from the uniqueness of the solution of our FBSDE we conclude that  $I_\delta(t, x, u_1, \beta_1^\varepsilon(u_1)) = \sum_{i \geq 1} \mathbf{1}_{\Lambda_i} I_\delta(t, x, u_1, \beta_i^1(u_1))$ ,  $P$ -a.s., for all  $u_1(\cdot) \in \mathcal{U}_{t, t+\delta}$ . Hence,

(3.22)

$$\begin{aligned} W_\delta(t, x) &\geq \sum_{i \geq 1} \mathbf{1}_{\Lambda_i} I_\delta(t, x, \beta_i^1) - \varepsilon \\ &\geq \sum_{i \geq 1} \mathbf{1}_{\Lambda_i} I_\delta(t, x, u_1, \beta_i^1(u_1)) - \varepsilon \\ &= I_\delta(t, x, u_1, \beta_1^\varepsilon(u_1)) - \varepsilon \\ &= G_{t, t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)}[W(t + \delta, X_{t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)})] - \varepsilon, \quad P\text{-a.s., for all } u_1 \in \mathcal{U}_{t, t+\delta}. \end{aligned}$$

On the other hand, from the definition of  $W(t + \delta, y)$ , with the same technique as before, we deduce that, for any  $y \in \mathbb{R}^n$ , there exists  $\beta_y^\varepsilon \in \mathcal{B}_{t+\delta, T}$  such that

$$(3.23) \quad W(t + \delta, y) \geq \text{esssup}_{u_2 \in \mathcal{U}_{t+\delta, T}} J(t + \delta, y; u_2, \beta_y^\varepsilon(u_2)) - \varepsilon, \quad P\text{-a.s.}$$

Let  $\{O_i\}_{i \geq 1} \subset \mathcal{B}(\mathbb{R}^n)$  be a decomposition of  $\mathbb{R}^n$  such that  $\sum_{i \geq 1} O_i = \mathbb{R}^n$  and  $\text{diam}(O_i) \leq \varepsilon$ ,  $i \geq 1$ . Let  $y_i$  be an arbitrarily fixed element of  $O_i$ ,  $i \geq 1$ . Defining  $[X_{t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)}] := \sum_{i \geq 1} y_i \mathbf{1}_{\{X_{t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)} \in O_i\}}$ , we have

$$(3.24) \quad |X_{t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)} - [X_{t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)}]| \leq \varepsilon, \quad \text{everywhere on } \Omega, \text{ for all } u_1 \in \mathcal{U}_{t, t+\delta}.$$

Moreover, for each  $y_i$ , there exists some  $\beta_{y_i}^\varepsilon \in \mathcal{B}_{t+\delta, T}$  such that (3.23) holds, and, clearly,  $\beta_{u_1}^\varepsilon := \sum_{i \geq 1} \mathbf{1}_{\{X_{t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)} \in O_i\}} \beta_{y_i}^\varepsilon \in \mathcal{B}_{t+\delta, T}$ .

Now we can define the new strategy  $\beta^\varepsilon(u) := \beta_1^\varepsilon(u_1) \oplus \beta_{u_1}^\varepsilon(u_2)$ ,  $u \in \mathcal{U}_{t, T}$ , where  $u_1 = u|_{[t, t+\delta]}$ ,  $u_2 = u|_{(t+\delta, T]}$  (restriction of  $u$  to  $[t, t+\delta] \times \Omega$  and  $(t+\delta, T] \times \Omega$ , resp.). Obviously,  $\beta^\varepsilon$  maps  $\mathcal{U}_{t, T}$  into  $\mathcal{V}_{t, T}$ . Moreover,  $\beta^\varepsilon$  is nonanticipating: Indeed, let  $S : \Omega \rightarrow [t, T]$  be an  $\mathbb{F}$ -stopping time and  $u, u' \in \mathcal{U}_{t, T}$  be such that  $u \equiv u'$  on  $\llbracket t, S \rrbracket$ . Decomposing  $u, u'$  into  $u_1, u'_1 \in \mathcal{U}_{t, t+\delta}$ ,  $u_2, u'_2 \in \mathcal{U}_{t+\delta, T}$  such that  $u = u_1 \oplus u_2$  and  $u' = u'_1 \oplus u'_2$ . We have  $u_1 \equiv u'_1$  on  $\llbracket t, S \wedge (t+\delta) \rrbracket$  from which we get  $\beta_1^\varepsilon(u_1) \equiv \beta_1^\varepsilon(u'_1)$  on  $\llbracket t, S \wedge (t+\delta) \rrbracket$  (recall that  $\beta_1^\varepsilon$  is nonanticipating). On the other hand,  $u_2 \equiv u'_2$  on  $\llbracket t+\delta, S \vee (t+\delta) \rrbracket \subset (t+\delta, T] \times \{S > t+\delta\}$ , and on  $\{S > t+\delta\}$  we have  $X_{t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)} = X_{t+\delta}^{t, x; u'_1, \beta_1^\varepsilon(u'_1)}$ . Consequently, from our definition,  $\beta_{u_1}^\varepsilon = \beta_{u'_1}^\varepsilon$  on  $\{S > t+\delta\}$  and  $\beta_{u_1}^\varepsilon(u_2) \equiv \beta_{u'_1}^\varepsilon(u'_2)$  on  $\llbracket t+\delta, S \vee (t+\delta) \rrbracket$ . This yields  $\beta^\varepsilon(u) = \beta_1^\varepsilon(u_1) \oplus \beta_{u_1}^\varepsilon(u_2) \equiv \beta_1^\varepsilon(u'_1) \oplus \beta_{u'_1}^\varepsilon(u'_2) = \beta^\varepsilon(u')$  on  $\llbracket t, S \rrbracket$ , from which it follows that  $\beta^\varepsilon \in \mathcal{B}_{t, T}$ .

Let now  $u \in \mathcal{U}_{t, T}$  be arbitrarily chosen and decomposed into  $u_1 = u|_{[t, t+\delta]} \in \mathcal{U}_{t, t+\delta}$  and  $u_2 = u|_{(t+\delta, T]} \in \mathcal{U}_{t+\delta, T}$ . Then, from (3.22), (3.18)(i), (3.24), and Lemmas 2.2 (comparison theorem) and 2.3, we obtain

(3.25)

$$\begin{aligned} W_\delta(t, x) &\geq G_{t, t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)}[W(t + \delta, X_{t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)})] - \varepsilon \\ &\geq G_{t, t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)}[W(t + \delta, [X_{t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)}]) - C\varepsilon] - \varepsilon \\ &\geq G_{t, t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)}[W(t + \delta, [X_{t+\delta}^{t, x; u_1, \beta_1^\varepsilon(u_1)}])] - C\varepsilon \end{aligned}$$

$$= G_{t,t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)} \left[ \sum_{i \geq 1} \mathbf{1}_{\{X_{t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)} \in O_i\}} W(t+\delta, y_i) \right] - C\varepsilon, \quad P\text{-a.s.}$$

Furthermore, from (3.23), (3.18)(ii), (3.24), and Lemmas 2.2 and 2.3, we have

(3.26)

$$\begin{aligned} W_\delta(t, x) &\geq G_{t,t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)} \left[ \sum_{i \geq 1} \mathbf{1}_{\{X_{t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)} \in O_i\}} J(t+\delta, y_i; u_2, \beta_{y_i}^\varepsilon(u_2)) - \varepsilon \right] - C\varepsilon \\ &\geq G_{t,t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)} \left[ \sum_{i \geq 1} \mathbf{1}_{\{X_{t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)} \in O_i\}} J(t+\delta, y_i; u_2, \beta_{y_i}^\varepsilon(u_2)) \right] - C\varepsilon \\ &= G_{t,t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)} [J(t+\delta, [X_{t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)}]; u_2, \beta_{u_1}^\varepsilon(u_2))] - C\varepsilon \\ &\geq G_{t,t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)} [J(t+\delta, X_{t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)}; u_2, \beta_{u_1}^\varepsilon(u_2)) - C\varepsilon] - C\varepsilon \\ &\geq G_{t,t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)} [J(t+\delta, X_{t+\delta}^{t,x;u_1,\beta_1^\varepsilon(u_1)}; u_2, \beta_{u_1}^\varepsilon(u_2))] - C\varepsilon \\ &= G_{t,t+\delta}^{t,x;u,\beta^\varepsilon(u)} [Y_{t+\delta}^{t,x,u,\beta^\varepsilon(u)}] - C\varepsilon \\ &= Y_t^{t,x;u,\beta^\varepsilon(u)} - C\varepsilon, \quad P\text{-a.s., for any } u \in \mathcal{U}_{t,T}. \end{aligned}$$

Consequently,

$$\begin{aligned} W_\delta(t, x) &\geq \operatorname{esssup}_{u \in \mathcal{U}_{t,T}} J(t, x; u, \beta^\varepsilon(u)) - C\varepsilon \\ (3.27) \quad &\geq \operatorname{essinf}_{\beta \in \mathcal{B}_{t,T}} \operatorname{esssup}_{u \in \mathcal{U}_{t,T}} J(t, x; u, \beta(u)) - C\varepsilon \\ &= W(t, x) - C\varepsilon, \quad P\text{-a.s.} \end{aligned}$$

Finally, letting  $\varepsilon \downarrow 0$  we get  $W_\delta(t, x) \geq W(t, x)$ . The proof is complete.  $\square$

*Remark 3.4.* (i) From the inequalities (3.17) and (3.22) we see that for all  $(t, x) \in [0, T] \times \mathbb{R}^n$ ,  $\delta > 0$ , with  $0 < \delta \leq T - t$  and  $\varepsilon > 0$ , the following hold: (a) For every  $\beta \in \mathcal{B}_{t,t+\delta}$ , there exists some  $u^\varepsilon(\cdot) \in \mathcal{U}_{t,t+\delta}$  such that

$$(3.28) \quad W(t, x) (= W_\delta(t, x)) \leq G_{t,t+\delta}^{t,x;u^\varepsilon,\beta(u^\varepsilon)} [W(t+\delta, X_{t+\delta}^{t,x;u^\varepsilon,\beta(u^\varepsilon)})] + \varepsilon, \quad P\text{-a.s.}$$

(b) There exists some  $\beta^\varepsilon \in \mathcal{B}_{t,t+\delta}$  such that, for all  $u \in \mathcal{U}_{t,t+\delta}$ ,

$$(3.29) \quad W(t, x) (= W_\delta(t, x)) \geq G_{t,t+\delta}^{t,x;u,\beta^\varepsilon(u)} [W(t+\delta, X_{t+\delta}^{t,x;u,\beta^\varepsilon(u)})] - \varepsilon, \quad P\text{-a.s.}$$

(ii) Recall that the lower value function  $W$  is deterministic. Thus, by choosing  $\delta = T - t$  and taking the expectation on both sides of (3.28) and (3.29) we can show that

$$W(t, x) = \inf_{\beta \in \mathcal{B}_{t,T}} \sup_{u \in \mathcal{U}_{t,T}} E[J(t, x; u, \beta(u))].$$

In analogy we also have

$$U(t, x) = \sup_{\alpha \in \mathcal{A}_{t,T}} \inf_{v \in \mathcal{V}_{t,T}} E[J(t, x; \alpha(v), v)].$$

The above formulas look similar to the definitions of the lower and the upper value functions defined by Fleming and Souganidis [14] for the case of  $f$  being independent



of  $(y, z)$ . However, they consider only control processes which are independent of the past  $\mathcal{F}_t$ . In Remark 5.3 we will come back to this comparison and identify their value functions with ours for such a coefficient  $f$ .

In Lemma 3.5 we have already seen that the lower value function  $W(t, x)$  is Lipschitz continuous in  $x$ , uniformly in  $t$ . With the help of Theorem 3.6 we can now also study the continuity property of  $W(t, x)$  in  $t$ .

**THEOREM 3.10.** *Let us suppose that the assumptions (H3.1) and (H3.2) hold. Then the lower value function  $W(t, x)$  is  $\frac{1}{2}$ -Hölder continuous in  $t$ : There exists a constant  $C$  such that, for every  $x \in \mathbb{R}^n$ ,  $t, t' \in [0, T]$ ,*

$$|W(t, x) - W(t', x)| \leq C(1 + |x|)|t - t'|^{\frac{1}{2}}.$$

*Proof.* Let  $(t, x) \in [0, T] \times \mathbb{R}^n$  and  $\delta > 0$  be arbitrarily given such that  $0 < \delta \leq T - t$ . Our objective is to prove the following inequality by using (3.28) and (3.29):

$$(3.30) \quad -C(1 + |x|)\delta^{\frac{1}{2}} \leq W(t, x) - W(t + \delta, x) \leq C(1 + |x|)\delta^{\frac{1}{2}}.$$

From it we obtain immediately that  $W$  is  $\frac{1}{2}$ -Hölder continuous in  $t$ . We will check only the second inequality in (3.30); the first one can be shown in a similar way. To this end we note that due to (3.28), for an arbitrarily small  $\varepsilon > 0$ ,

$$(3.31) \quad W(t, x) - W(t + \delta, x) \leq I_\delta^1 + I_\delta^2 + \varepsilon,$$

where

$$\begin{aligned} I_\delta^1 &:= G_{t, t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)}[W(t + \delta, X_{t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)})] - G_{t, t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)}[W(t + \delta, x)], \\ I_\delta^2 &:= G_{t, t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)}[W(t + \delta, x)] - W(t + \delta, x) \end{aligned}$$

for arbitrarily chosen  $\beta \in \mathcal{B}_{t, t+\delta}$  and  $u^\varepsilon \in \mathcal{U}_{t, t+\delta}$  such that (3.28) holds. From Lemma 2.3 and the estimate (3.12) we obtain that, for some constant  $C$  independent of the controls  $u^\varepsilon$  and  $\beta(u^\varepsilon)$ ,

$$\begin{aligned} |I_\delta^1| &\leq [CE(|W(t + \delta, X_{t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)}) - W(t + \delta, x)|^2 | \mathcal{F}_t)]^{\frac{1}{2}} \\ &\leq [CE(|X_{t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)} - x|^2 | \mathcal{F}_t)]^{\frac{1}{2}}, \end{aligned}$$

and since  $E[|X_{t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)} - x|^2 | \mathcal{F}_t] \leq C(1 + |x|^2)\delta$  we deduce that  $|I_\delta^1| \leq C(1 + |x|)\delta^{\frac{1}{2}}$ . From the definition of  $G_{t, t+\delta}^{t, x; u^\varepsilon, \beta(u^\varepsilon)}[\cdot]$  (see (3.13)) we know that the second term  $I_\delta^2$  can be written as

$$\begin{aligned} I_\delta^2 &= E \left[ W(t + \delta, x) + \int_t^{t+\delta} f(s, X_s^{t, x; u^\varepsilon, \beta(u^\varepsilon)}, \tilde{Y}_s^{t, x; u^\varepsilon, \beta(u^\varepsilon)}, \tilde{Z}_s^{t, x; u^\varepsilon, \beta(u^\varepsilon)}, u_s^\varepsilon, \beta_s(u_s^\varepsilon)) ds \right. \\ &\quad \left. - \int_t^{t+\delta} \tilde{Z}_s^{t, x; u^\varepsilon, \beta(u^\varepsilon)} dB_s | \mathcal{F}_t \right] - W(t + \delta, x) \\ &= E \left[ \int_t^{t+\delta} f(s, X_s^{t, x; u^\varepsilon, \beta(u^\varepsilon)}, \tilde{Y}_s^{t, x; u^\varepsilon, \beta(u^\varepsilon)}, \tilde{Z}_s^{t, x; u^\varepsilon, \beta(u^\varepsilon)}, u_s^\varepsilon, \beta_s(u_s^\varepsilon)) ds | \mathcal{F}_t \right]. \end{aligned}$$

With the help of the Schwartz inequality and the estimates (3.3) and (A.4)(i), we

then have

$$\begin{aligned}
 |I_\delta^2| &\leq \delta^{\frac{1}{2}} E \left[ \int_t^{t+\delta} |f(s, X_s^{t,x;u^\varepsilon, \beta(u^\varepsilon)}, \tilde{Y}_s^{t,x;u^\varepsilon, \beta(u^\varepsilon)}, \tilde{Z}_s^{t,x;u^\varepsilon, \beta(u^\varepsilon)}, u_s^\varepsilon, \beta_s(u^\varepsilon))|^2 ds | \mathcal{F}_t \right]^{\frac{1}{2}} \\
 &\leq \delta^{\frac{1}{2}} E \left[ \int_t^{t+\delta} (|f(s, X_s^{t,x;u^\varepsilon, \beta(u^\varepsilon)}, 0, 0, u_s^\varepsilon, \beta_s(u^\varepsilon))| + C|\tilde{Y}_s^{t,x;u^\varepsilon, \beta(u^\varepsilon)}| \right. \\
 &\quad \left. + C|\tilde{Z}_s^{t,x;u^\varepsilon, \beta(u^\varepsilon)}|)^2 ds | \mathcal{F}_t \right]^{\frac{1}{2}} \\
 &\leq C\delta^{\frac{1}{2}} E \left[ \int_t^{t+\delta} (|1 + |X_s^{t,x;u^\varepsilon, \beta(u^\varepsilon)}| + |\tilde{Y}_s^{t,x;u^\varepsilon, \beta(u^\varepsilon)}| + |\tilde{Z}_s^{t,x;u^\varepsilon, \beta(u^\varepsilon)}|)^2 ds | \mathcal{F}_t \right]^{\frac{1}{2}} \\
 &\leq C(1 + |x|)\delta^{\frac{1}{2}}.
 \end{aligned}$$

Hence, from (3.31),

$$W(t, x) - W(t + \delta, x) \leq C(1 + |x|)\delta^{\frac{1}{2}} + \varepsilon,$$

and letting  $\varepsilon \downarrow 0$  we get the second inequality of (3.30). The proof is complete.  $\square$

**4. Viscosity solution of Isaacs' equation: Existence theorem.** In this section we consider the following Isaacs' equations:

$$(4.1) \quad \begin{cases} \frac{\partial}{\partial t} W(t, x) + H^-(t, x, W, DW, D^2W) = 0, & (t, x) \in [0, T] \times \mathbb{R}^n, \\ W(T, x) = \Phi(x), & x \in \mathbb{R}^n, \end{cases}$$

and

$$(4.2) \quad \begin{cases} \frac{\partial}{\partial t} U(t, x) + H^+(t, x, U, DU, D^2U) = 0, & (t, x) \in [0, T] \times \mathbb{R}^n, \\ U(T, x) = \Phi(x), & x \in \mathbb{R}^n, \end{cases}$$

associated with the Hamiltonians

$$H^-(t, x, y, p, X) = \sup_{u \in U} \inf_{v \in V} \left\{ \frac{1}{2} \text{tr}(\sigma \sigma^T(t, x, u, v) X) + p \cdot b(t, x, u, v) + f(t, x, y, p \cdot \sigma, u, v) \right\}$$

and

$$H^+(t, x, y, p, X) = \inf_{v \in V} \sup_{u \in U} \left\{ \frac{1}{2} \text{tr}(\sigma \sigma^T(t, x, u, v) X) + p \cdot b(t, x, u, v) + f(t, x, y, p \cdot \sigma, u, v) \right\},$$

respectively, where  $t \in [0, T]$ ,  $x \in \mathbb{R}^n$ ,  $y \in \mathbb{R}$ ,  $p \in \mathbb{R}^n$ , and  $X \in \mathbf{S}^n$  (recall that  $\mathbf{S}^n$  denotes the set of  $n \times n$  symmetric matrices). Here the functions  $b, \sigma, f$ , and  $\Phi$  are supposed to satisfy (H3.1) and (H3.2), respectively.

In this section we want to prove that the lower value function  $W(t, x)$  introduced by (3.9) is the viscosity solution of (4.1), while the upper value function  $U(t, x)$  defined by (3.10) is the viscosity solution of (4.2). For this we extend Peng's BSDE approach [26] developed in the framework of stochastic control theory into that of stochastic differential games. The difficulties related with this extension come from the fact that now, contrarily to the framework of stochastic control theory studied by Peng, we have to do with stochastic differential games in which strategies are played versus controls. In order to overcome these difficulties in the proof that  $W$  is a viscosity supersolution, we have, in particular, to enrich Peng's BSDE method by Lemma

4.6 and to adapt heavily the proof of Lemma 4.5. On the other hand, the proof that  $W$  is a viscosity subsolution is not covered by Peng's BSDE method and requires a quite new approach. The uniqueness of the viscosity solution will be shown in the next section for the class of continuous functions satisfying some growth assumption which is weaker than the polynomial growth condition. We first recall the definition of a viscosity solution of (4.1) and similarly for (4.2). The reader more interested in viscosity solutions is referred to Crandall, Ishii, and Lions [8].

DEFINITION 4.1. A real-valued continuous function  $W \in C([0, T] \times \mathbb{R}^n)$  is called  
(i) a viscosity subsolution of (4.1) if  $W(T, x) \leq \Phi(x)$  for all  $x \in \mathbb{R}^n$  and if for all functions  $\varphi \in C_{l,b}^3([0, T] \times \mathbb{R}^n)$  and  $(t, x) \in [0, T) \times \mathbb{R}^n$  such that  $W - \varphi$  attains its local maximum at  $(t, x)$ :

$$\frac{\partial \varphi}{\partial t}(t, x) + H^-(t, x, \varphi, D\varphi, D^2\varphi) \geq 0;$$

(ii) a viscosity supersolution of (4.1) if  $W(T, x) \geq \Phi(x)$  for all  $x \in \mathbb{R}^n$  and if for all functions  $\varphi \in C_{l,b}^3([0, T] \times \mathbb{R}^n)$  and  $(t, x) \in [0, T) \times \mathbb{R}^n$  such that  $W - \varphi$  attains its local minimum at  $(t, x)$ :

$$\frac{\partial \varphi}{\partial t}(t, x) + H^-(t, x, \varphi, D\varphi, D^2\varphi) \leq 0;$$

(iii) a viscosity solution of (4.1) if it is both a viscosity sub- and a supersolution of (4.1).

Remark 4.1.  $C_{l,b}^3([0, T] \times \mathbb{R}^n)$  denotes the set of the real-valued functions that are continuously differentiable up to the third order and whose derivatives of order from 1 to 3 are bounded.

We first prove that the lower value function  $W(t, x)$  is a viscosity solution of (4.1).

THEOREM 4.2. Under the assumptions (H3.1) and (H3.2) the lower value function  $W(t, x)$  is a viscosity solution of (4.1).

For the proof of this theorem we need four auxiliary lemmas. To abbreviate notations we put, for some arbitrarily chosen but fixed  $\varphi \in C_{l,b}^3([0, T] \times \mathbb{R}^n)$ ,

$$(4.3) \quad F(s, x, y, z, u, v) = \frac{\partial}{\partial s} \varphi(s, x) + \frac{1}{2} \text{tr}(\sigma \sigma^T(s, x, u, v) D^2 \varphi) + D\varphi \cdot b(s, x, u, v) \\ + f(s, x, y + \varphi(s, x), z + D\varphi(s, x) \cdot \sigma(s, x, u, v), u, v),$$

$(s, x, y, z, u, v) \in [0, T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \times U \times V$ , and we consider the following BSDE defined on the interval  $[t, t + \delta]$  ( $0 < \delta \leq T - t$ ):

$$(4.4) \quad \begin{cases} -dY_s^{1,u,v} = F(s, X_s^{t,x;u,v}, Y_s^{1,u,v}, Z_s^{1,u,v}, u_s, v_s) ds - Z_s^{1,u,v} dB_s, \\ Y_{t+\delta}^{1,u,v} = 0, \end{cases}$$

where the process  $X^{t,x,u,v}$  has been introduced by (3.1) and  $u(\cdot) \in \mathcal{U}_{t,t+\delta}$ ,  $v(\cdot) \in \mathcal{V}_{t,t+\delta}$ .

Remark 4.2. It is not hard to check that  $F(s, X_s^{t,x;u,v}, y, z, u_s, v_s)$  satisfies (A1) and (A2). Thus, due to Lemma 2.1, (4.4) has a unique solution.

We can characterize the solution process  $Y^{1,u,v}$  as follows.

LEMMA 4.3. For every  $s \in [t, t + \delta]$ , we have the following relationship:

$$(4.5) \quad Y_s^{1,u,v} = G_{s,t+\delta}^{t,x;u,v}[\varphi(t + \delta, X_{t+\delta}^{t,x;u,v})] - \varphi(s, X_s^{t,x;u,v}), \quad P\text{-a.s.}$$

*Proof.* We recall that  $G_{s,t+\delta}^{t,x;u,v}[\varphi(t+\delta, X_{t+\delta}^{t,x;u,v})]$  is defined with the help of the solution of the BSDE

$$\begin{cases} -dY_s^{u,v} = f(s, X_s^{t,x;u,v}, Y_s^{u,v}, Z_s^{u,v}, u_s, v_s)ds - Z_s^{u,v}dB_s, & s \in [t, t+\delta], \\ Y_{t+\delta}^{u,v} = \varphi(t+\delta, X_{t+\delta}^{t,x;u,v}), \end{cases}$$

by the following formula:

$$(4.6) \quad G_{s,t+\delta}^{t,x;u,v}[\varphi(t+\delta, X_{t+\delta}^{t,x;u,v})] = Y_s^{u,v}, \quad s \in [t, t+\delta]$$

(see (3.13)). Therefore we only need to prove that  $Y_s^{u,v} - \varphi(s, X_s^{t,x;u,v}) \equiv Y_s^{1,u,v}$ . This result can be obtained easily by applying Itô's formula to  $\varphi(s, X_s^{t,x;u,v})$ . Indeed, we get that the stochastic differentials of  $Y_s^{u,v} - \varphi(s, X_s^{t,x;u,v})$  and  $Y_s^{1,u,v}$  coincide, while at the terminal time  $t+\delta$ ,  $Y_{t+\delta}^{u,v} - \varphi(t+\delta, X_{t+\delta}^{t,x;u,v}) = 0 = Y_{t+\delta}^{1,u,v}$ . So the proof is complete.  $\square$

Now we consider the following simple BSDE in which the driving process  $X^{t,x;u,v}$  is replaced by its deterministic initial value  $x$ :

$$(4.7) \quad \begin{cases} -dY_s^{2,u,v} = F(s, x, Y_s^{2,u,v}, Z_s^{2,u,v}, u_s, v_s)ds - Z_s^{2,u,v}dB_s, \\ Y_{t+\delta}^{2,u,v} = 0, \quad s \in [t, t+\delta], \end{cases}$$

where  $u(\cdot) \in \mathcal{U}_{t,t+\delta}$ ,  $v(\cdot) \in \mathcal{V}_{t,t+\delta}$ . The following lemma will allow us to neglect the difference  $|Y_t^{1,u,v} - Y_t^{2,u,v}|$  for sufficiently small  $\delta > 0$ .

LEMMA 4.4. *For every  $u \in \mathcal{U}_{t,t+\delta}$ ,  $v \in \mathcal{V}_{t,t+\delta}$ , we have*

$$(4.8) \quad |Y_t^{1,u,v} - Y_t^{2,u,v}| \leq C\delta^{\frac{3}{2}}, \quad P\text{-a.s.},$$

where  $C$  is independent of the control processes  $u$  and  $v$ .

*Proof.* From (3.3) we have for all  $p \geq 2$  the existence of some  $C_p \in \mathbb{R}_+$  such that

$$E \left[ \sup_{t \leq s \leq T} |X_s^{t,x;u,v}|^p | \mathcal{F}_t \right] \leq C_p(1 + |x|^p), \quad P\text{-a.s.}, \quad \text{uniformly in } u \in \mathcal{U}_{t,t+\delta}, v \in \mathcal{V}_{t,t+\delta}.$$

This combined with the estimate

$$\begin{aligned} E \left[ \sup_{t \leq s \leq t+\delta} |X_s^{t,x;u,v} - x|^p | \mathcal{F}_t \right] &\leq 2^{p-1} E \left[ \sup_{t \leq s \leq t+\delta} \left| \int_t^s b(r, X_r^{t,x;u,v}, u_r, v_r) dr \right|^p | \mathcal{F}_t \right] \\ &\quad + 2^{p-1} E \left[ \sup_{t \leq s \leq t+\delta} \left| \int_t^s \sigma(r, X_r^{t,x;u,v}, u_r, v_r) dB_r \right|^p | \mathcal{F}_t \right] \end{aligned}$$

yields

$$(4.9) \quad E \left[ \sup_{t \leq s \leq t+\delta} |X_s^{t,x;u,v} - x|^p | \mathcal{F}_t \right] \leq C_p \delta^{\frac{p}{2}}, \quad P\text{-a.s.}, \quad \text{uniformly in } u \in \mathcal{U}_{t,t+\delta}, v \in \mathcal{V}_{t,t+\delta}.$$

We now apply Lemma 2.3 combined with (4.9) to (4.4) and (4.7). For this we set in Lemma 2.3:

$$\xi_1 = \xi_2 = 0, \quad g(s, y, z) = F(s, X_s^{t,x;u,v}, y, z, u_s, v_s),$$

$$\varphi_1(s) = 0, \quad \varphi_2(s) = F(s, x, Y_s^{2,u,v}, Z_s^{2,u,v}, u_s, v_s) - F(s, X_s^{t,x;u,v}, Y_s^{2,u,v}, Z_s^{2,u,v}, u_s, v_s).$$

Obviously, the function  $g$  is Lipschitz with respect to  $(y, z)$ , and  $|\varphi_2(s)| \leq C(1 + |x|^2)(|X_s^{t,x;u,v} - x| + |X_s^{t,x;u,v} - x|^3)$  for  $s \in [t, t + \delta]$ ,  $(t, x) \in [0, T] \times \mathbb{R}^n$ ,  $u \in \mathcal{U}_{t,t+\delta}$ ,  $v \in \mathcal{V}_{t,t+\delta}$ . Thus, with the notation  $\rho_0(r) = (1 + |x|^2)(r + r^3)$ ,  $r \geq 0$ , we have

$$\begin{aligned} & E \left[ \int_t^{t+\delta} (|Y_s^{1,u,v} - Y_s^{2,u,v}|^2 + |Z_s^{1,u,v} - Z_s^{2,u,v}|^2) ds | \mathcal{F}_t \right] \\ & \leq CE \left[ \int_t^{t+\delta} \rho_0^2(|X_s^{t,x,u,v} - x|) ds | \mathcal{F}_t \right] \\ & \leq C\delta E \left[ \sup_{t \leq s \leq t+\delta} \rho_0^2(|X_s^{t,x,u,v} - x|) | \mathcal{F}_t \right] \\ & \leq C\delta^2. \end{aligned}$$

Therefore,

$$\begin{aligned} & |Y_t^{1,u,v} - Y_t^{2,u,v}| = |E[(Y_t^{1,u,v} - Y_t^{2,u,v}) | \mathcal{F}_t]| \\ & = \left| E \left[ \int_t^{t+\delta} (F(s, X_s^{t,x,u,v}, Y_s^{1,u,v}, Z_s^{1,u,v}, u_s, v_s) - F(s, x, Y_s^{2,u,v}, Z_s^{2,u,v}, u_s, v_s)) ds | \mathcal{F}_t \right] \right| \\ & \leq CE \left[ \int_t^{t+\delta} [\rho_0(|X_s^{t,x,u,v} - x|) + |Y_s^{1,u,v} - Y_s^{2,u,v}| + |Z_s^{1,u,v} - Z_s^{2,u,v}|] ds | \mathcal{F}_t \right] \\ & \leq CE \left[ \int_t^{t+\delta} \rho_0(|X_s^{t,x,u,v} - x|) ds | \mathcal{F}_t \right] + C\delta^{\frac{1}{2}} \left\{ E \left[ \int_t^{t+\delta} |Y_s^{1,u,v} - Y_s^{2,u,v}|^2 ds | \mathcal{F}_t \right]^{\frac{1}{2}} \right. \\ & \quad \left. + E \left[ \int_t^{t+\delta} |Z_s^{1,u,v} - Z_s^{2,u,v}|^2 ds | \mathcal{F}_t \right]^{\frac{1}{2}} \right\} \\ & \leq C\delta^{\frac{3}{2}}. \end{aligned}$$

Thus, the proof is complete.  $\square$

LEMMA 4.5. *Let  $Y_0(\cdot)$  be the solution of the following ordinary differential equation:*

$$(4.10) \quad \begin{cases} -\dot{Y}_0(s) &= F_0(s, x, Y_0(s), 0), \quad s \in [t, t + \delta], \\ Y_0(t + \delta) &= 0, \end{cases}$$

where the function  $F_0$  is defined by

$$(4.11) \quad F_0(s, x, y, z) = \sup_{u \in U} \inf_{v \in V} F(s, x, y, z, u, v).$$

Then,  $P$ -a.s.,

$$(4.12) \quad \text{esssup}_{u \in \mathcal{U}_{t,t+\delta}} \text{essinf}_{v \in \mathcal{V}_{t,t+\delta}} Y_t^{2,u,v} = Y_0(t).$$

*Proof.* Obviously,  $F_0(s, x, y, z)$  is Lipschitz in  $(y, z)$ , uniformly with respect to  $(s, x)$ . This guarantees existence and uniqueness for (4.10). We first introduce the function

$$(4.13) \quad F_1(s, x, y, z, u) = \inf_{v \in V} F(s, x, y, z, u, v), \quad (s, x, y, z, u) \in [0, T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \times U$$

and consider the BSDE

$$(4.14) \quad \begin{cases} -dY^{3,u}(s) &= F_1(s, x, Y^{3,u}(s), Z^{3,u}(s), u_s)ds - Z^{3,u}(s)dB_s, \\ Y^{3,u}(t+\delta) &= 0, \quad s \in [t, t+\delta], \end{cases}$$

for  $u \in \mathcal{U}_{t,t+\delta}$ . We notice that since  $F_1(s, x, y, z, u_s)$  is Lipschitz in  $(y, z)$ , for every  $u \in \mathcal{U}_{t,t+\delta}$ , there exists a unique solution  $(Y^{3,u}, Z^{3,u})$  to the BSDE (4.14). Moreover,

$$Y^{3,u}(t) = \operatorname{essinf}_{v(\cdot) \in \mathcal{V}_{t,t+\delta}} Y_t^{2,u,v}, \quad P\text{-a.s.}, \quad \text{for any } u \in \mathcal{U}_{t,t+\delta}.$$

Indeed, from the definition of  $F_1$  and Lemma 2.2 (comparison theorem) we have

$$Y^{3,u}(t) \leq \operatorname{essinf}_{v(\cdot) \in \mathcal{V}_{t,t+\delta}} Y_t^{2,u,v}, \quad P\text{-a.s.}, \quad \text{for all } u \in \mathcal{U}_{t,t+\delta}.$$

On the other hand, there exists a measurable function  $v^3 : [t, T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \times U \rightarrow V$  such that

$$F_1(s, x, y, z, u) = F(s, x, y, z, u, v^3(s, x, y, z, u)) \quad \text{for any } s, x, y, z, u.$$

We then put

$$\tilde{v}_s^3 := v^3(s, x, Y_s^{3,u}, Z_s^{3,u}, u_s), \quad s \in [t, t+\delta],$$

and we observe that  $\tilde{v}^3 \in \mathcal{V}_{t,t+\delta}$ , and

$$F_1(s, x, Y_s^{3,u}, Z_s^{3,u}, u_s) = F(s, x, Y_s^{3,u}, Z_s^{3,u}, u_s, \tilde{v}_s^3), \quad s \in [t, t+\delta].$$

Consequently, from the uniqueness of the solution of the BSDE it follows that  $(Y^{3,u}, Z^{3,u}) = (Y^{2,u,\tilde{v}^3}, Z^{2,u,\tilde{v}^3})$  and, in particular,  $Y_t^{3,u} = Y_t^{2,u,\tilde{v}^3}$ ,  $P$ -a.s., for any  $u \in \mathcal{U}_{t,t+\delta}$ . This proves that

$$Y^{3,u}(t) = \operatorname{essinf}_{v \in \mathcal{V}_{t,t+\delta}} Y_t^{2,u,v}, \quad P\text{-a.s.}, \quad \text{for all } u \in \mathcal{U}_{t,t+\delta}.$$

Finally, since  $F_0(s, x, y, z) = \sup_{u \in U} F_1(s, x, y, z, u)$ , an argument similar to that developed above yields

$$Y_0(t) = \operatorname{esssup}_{u \in \mathcal{U}_{t,t+\delta}} Y^{3,u}(t) (= \operatorname{esssup}_{u \in \mathcal{U}_{t,t+\delta}} \operatorname{essinf}_{v \in \mathcal{V}_{t,t+\delta}} Y_t^{2,u,v}), \quad P\text{-a.s.}$$

It uses the fact that (4.10) can be considered as a BSDE with the solution  $(Y_s, Z_s) = (Y_0(s), 0)$ . The proof is complete.  $\square$

LEMMA 4.6. *For every  $u \in \mathcal{U}_{t,t+\delta}$ ,  $v \in \mathcal{V}_{t,t+\delta}$ , we have*

$$(4.15) \quad E \left[ \int_t^{t+\delta} |Y_s^{2,u,v}| ds | \mathcal{F}_t \right] + E \left[ \int_t^{t+\delta} |Z_s^{2,u,v}| ds | \mathcal{F}_t \right] \leq C\delta^{\frac{3}{2}}, \quad P\text{-a.s.},$$

where the constant  $C$  is independent of the controls  $u, v$ .

*Proof.* Since  $F(s, x, \cdot, \cdot, u, v)$  has a linear growth in  $(y, z)$ , uniformly in  $(u, v)$ , we get from Lemma 2.3 that, for some constant  $C$  independent of  $\delta$  and the control processes  $u, v$ ,

$$|Y_s^{2,u,v}|^2 \leq C\delta, \quad E \left[ \int_s^{t+\delta} |Z_r^{2,u,v}|^2 dr | \mathcal{F}_s \right] \leq C\delta, \quad s \in [t, t+\delta].$$

On the other hand, from (4.7),

$$\begin{aligned} |Y_s^{2,u,v}| &\leq E \left[ \int_s^{t+\delta} |F(r, x, Y_r^{2,u,v}, Z_r^{2,u,v}, u_r, v_r)| dr | \mathcal{F}_s \right] \\ &\leq CE \left[ \int_s^{t+\delta} (1 + |x|^2 + |Y_r^{2,u,v}| + |Z_r^{2,u,v}|) dr | \mathcal{F}_s \right] \\ &\leq C\delta + C\sqrt{\delta} \left( E \left[ \int_s^{t+\delta} |Z_r^{2,u,v}|^2 dr | \mathcal{F}_s \right] \right)^{\frac{1}{2}} \leq C\delta, \quad P\text{-a.s.}, s \in [t, t + \delta], \end{aligned}$$

and, since

$$\int_t^{t+\delta} Z_s^{2,u,v} dB_s = \int_t^{t+\delta} F(s, x, Y_s^{2,u,v}, Z_s^{2,u,v}, u_s, v_s) ds - Y_t^{2,u,v},$$

we can get  $E[\int_t^{t+\delta} |Z_s^{2,u,v}|^2 ds | \mathcal{F}_t] \leq C\delta^2$ . Finally,

$$\begin{aligned} E \left[ \int_t^{t+\delta} |Y_s^{2,u,v}|^2 ds | \mathcal{F}_t \right] + E \left[ \int_t^{t+\delta} |Z_s^{2,u,v}|^2 ds | \mathcal{F}_t \right] \\ \leq C\delta^2 + \delta^{\frac{1}{2}} \left\{ E \left[ \int_t^{t+\delta} |Z_s^{2,u,v}|^2 ds | \mathcal{F}_t \right] \right\}^{\frac{1}{2}} \leq C\delta^{\frac{3}{2}}. \end{aligned}$$

The proof is complete.  $\square$

Now we are able to give the proof of Theorem 4.2.

*Proof.* (1) Obviously,  $W(T, x) = \Phi(x)$ ,  $x \in \mathbb{R}^n$ . Let us show in a first step that  $W$  is a viscosity supersolution. For this we suppose that  $\varphi \in C_{l,b}^3([0, T] \times \mathbb{R}^n)$  and  $(t, x) \in [0, T] \times \mathbb{R}^n$  are such that  $W - \varphi$  attains its minimum at  $(t, x)$ . Notice that we can replace the condition of a local minimum by that of a global one in the definition of the viscosity supersolution since  $W$  is continuous and of at most linear growth. Without loss of generality we may also suppose that  $\varphi(t, x) = W(t, x)$ . Then, due to the DPP (see Theorem 3.6),

$$\begin{aligned} \varphi(t, x) = W(t, x) &= \text{essinf}_{\beta \in \mathcal{B}_{t,t+\delta}} \text{esssup}_{u \in \mathcal{U}_{t,t+\delta}} G_{t,t+\delta}^{t,x;u,\beta(u)} [W(t + \delta, X_{t+\delta}^{t,x;u,\beta(u)})], \\ &0 \leq \delta \leq T - t, \end{aligned}$$

and from  $W \geq \varphi$  and the monotonicity property of  $G_{t,t+\delta}^{t,x;u,\beta(u)}[\cdot]$  (see Lemma 2.2) we obtain

$$\text{essinf}_{\beta \in \mathcal{B}_{t,t+\delta}} \text{esssup}_{u \in \mathcal{U}_{t,t+\delta}} \{G_{t,t+\delta}^{t,x;u,\beta(u)}[\varphi(t + \delta, X_{t+\delta}^{t,x;u,\beta(u)})] - \varphi(t, x)\} \leq 0, \quad P\text{-a.s.}$$

Thus, from Lemma 4.3,

$$\text{essinf}_{\beta \in \mathcal{B}_{t,t+\delta}} \text{esssup}_{u \in \mathcal{U}_{t,t+\delta}} Y_t^{1,u,\beta(u)} \leq 0, \quad P\text{-a.s.},$$

and further, from Lemma 4.4 we have

$$\text{essinf}_{\beta \in \mathcal{B}_{t,t+\delta}} \text{esssup}_{u \in \mathcal{U}_{t,t+\delta}} Y_t^{2,u,\beta(u)} \leq C\delta^{\frac{3}{2}}, \quad P\text{-a.s.}$$

Consequently, since  $\operatorname{ess\,inf}_{v \in \mathcal{V}_{t,t+\delta}} Y_t^{2,u,v} \leq Y_t^{2,u,\beta(u)}$ ,  $\beta \in \mathcal{B}_{t,t+\delta}$ , we get

$$\operatorname{ess\,sup}_{u \in \mathcal{U}_{t,t+\delta}} \operatorname{ess\,inf}_{v \in \mathcal{V}_{t,t+\delta}} Y_t^{2,u,v} \leq \operatorname{ess\,inf}_{\beta \in \mathcal{B}_{t,t+\delta}} \operatorname{ess\,sup}_{u \in \mathcal{U}_{t,t+\delta}} Y_t^{2,u,\beta(u)} \leq C\delta^{\frac{3}{2}}, \text{ } P\text{-a.s.},$$

and Lemma 4.5 implies

$$Y_0(t) \leq C\delta^{\frac{3}{2}}, \text{ } P\text{-a.s.},$$

where  $Y_0$  is the unique solution of (4.10). It then follows easily that

$$\sup_{u \in U} \inf_{v \in V} F(t, x, 0, 0, u, v) = F_0(t, x, 0, 0) \leq 0,$$

and from the definition of  $F$  we see that  $W$  is a viscosity supersolution of (4.1).

(2) The second step is devoted to the proof that  $W$  is a viscosity subsolution. For this we suppose that  $\varphi \in C_{l,b}^3([0, T] \times \mathbb{R}^n)$  and  $(t, x) \in [0, T] \times \mathbb{R}^n$  are such that  $W - \varphi$  attains its maximum at  $(t, x)$ . Without loss of generality we suppose again that  $\varphi(t, x) = W(t, x)$ . We must prove that

$$\sup_{u \in U} \inf_{v \in V} F(t, x, 0, 0, u, v) = F_0(t, x, 0, 0) \geq 0.$$

Let us suppose that this is not true. Then there exists some  $\theta > 0$  such that

$$(4.16) \quad F_0(t, x, 0, 0) = \sup_{u \in U} \inf_{v \in V} F(t, x, 0, 0, u, v) \leq -\theta < 0,$$

and we can find a measurable function  $\psi : U \rightarrow V$  such that

$$F(t, x, 0, 0, u, \psi(u)) \leq -\frac{3}{4}\theta \text{ for all } u \in U.$$

Moreover, since  $F(\cdot, x, 0, 0, \cdot, \cdot)$  is uniformly continuous on  $[0, T] \times U \times V$ , there exists some  $T - t \geq R > 0$  such that

$$(4.17) \quad F(s, x, 0, 0, u, \psi(u)) \leq -\frac{1}{2}\theta \text{ for all } u \in U \text{ and } |s - t| \leq R.$$

On the other hand, due to the DPP (see Theorem 3.6), for every  $\delta \in (0, R]$ ,

$$\varphi(t, x) = W(t, x) = \operatorname{ess\,inf}_{\beta \in \mathcal{B}_{t,t+\delta}} \operatorname{ess\,sup}_{u \in \mathcal{U}_{t,t+\delta}} G_{t,t+\delta}^{t,x;u,\beta(u)}[W(t + \delta, X_{t+\delta}^{t,x;u,\beta(u)})],$$

and from  $W \leq \varphi$  and the monotonicity property of  $G_{t,t+\delta}^{t,x;u,\beta(u)}[\cdot]$  (see Lemma 2.2) we obtain

$$\operatorname{ess\,inf}_{\beta \in \mathcal{B}_{t,t+\delta}} \operatorname{ess\,sup}_{u \in \mathcal{U}_{t,t+\delta}} \{G_{t,t+\delta}^{t,x;u,\beta(u)}[\varphi(t + \delta, X_{t+\delta}^{t,x;u,\beta(u)})] - \varphi(t, x)\} \geq 0, \text{ } P\text{-a.s.}$$

Thus, from Lemma 4.3,

$$\operatorname{ess\,inf}_{\beta \in \mathcal{B}_{t,t+\delta}} \operatorname{ess\,sup}_{u \in \mathcal{U}_{t,t+\delta}} Y_t^{1,u,\beta(u)} \geq 0, \text{ } P\text{-a.s.},$$

and, in particular,

$$\operatorname{ess\,sup}_{u \in \mathcal{U}_{t,t+\delta}} Y_t^{1,u,\psi(u)} \geq 0, \text{ } P\text{-a.s.}$$



Here, by putting  $\psi_s(u)(\omega) = \psi(u_s(\omega))$ ,  $(s, \omega) \in [t, T] \times \Omega$ , we identify  $\psi$  as an element of  $\mathcal{B}_{t, t+\delta}$ . Given an arbitrarily  $\varepsilon > 0$  we can choose  $u^\varepsilon \in \mathcal{U}_{t, t+\delta}$  such that  $Y_t^{1, u^\varepsilon, \psi(u^\varepsilon)} \geq -\varepsilon\delta$ . From Lemma 4.4 we further have

$$(4.18) \quad Y_t^{2, u^\varepsilon, \psi(u^\varepsilon)} \geq -C\delta^{\frac{3}{2}} - \varepsilon\delta, \quad P\text{-a.s.}$$

Taking into account that

$$Y_t^{2, u^\varepsilon, \psi(u^\varepsilon)} = E \left[ \int_t^{t+\delta} F(s, x, Y_s^{2, u^\varepsilon, \psi(u^\varepsilon)}, Z_s^{2, u^\varepsilon, \psi(u^\varepsilon)}, u_s^\varepsilon, \psi_s(u^\varepsilon)) ds | \mathcal{F}_t \right]$$

we get from the Lipschitz property of  $F$  in  $(y, z)$ , (4.17), and Lemma 4.6 that

$$(4.19) \quad \begin{aligned} Y_t^{2, u^\varepsilon, \psi(u^\varepsilon)} &\leq E \left[ \int_t^{t+\delta} (C|Y_s^{2, u^\varepsilon, \psi(u^\varepsilon)}| + C|Z_s^{2, u^\varepsilon, \psi(u^\varepsilon)}| + F(s, x, 0, 0, u_s^\varepsilon, \psi_s(u^\varepsilon))) ds | \mathcal{F}_t \right] \\ &\leq C\delta^{\frac{3}{2}} - \frac{1}{2}\theta\delta, \quad P\text{-a.s.} \end{aligned}$$

From (4.18) and (4.19),  $-C\delta^{\frac{1}{2}} - \varepsilon \leq C\delta^{\frac{1}{2}} - \frac{1}{2}\theta$ ,  $P$ -a.s. Letting  $\delta \downarrow 0$ , and then  $\varepsilon \downarrow 0$ , we deduce that  $\theta \leq 0$ , which induces a contradiction. Therefore,

$$F_0(t, x, 0, 0) = \sup_{u \in U} \inf_{v \in V} F(t, x, 0, 0, u, v) \geq 0,$$

and from the definition of  $F$ , we know that  $W$  is a viscosity subsolution of (4.1). Finally, the results from the first and the second steps prove that  $W$  is a viscosity solution of (4.1).  $\square$

*Remark 4.3.* Similarly, we can prove that  $U$  is a viscosity solution of (4.2).

**5. Viscosity solution of Isaacs' equation: Uniqueness theorem.** The objective of this section is to study the uniqueness of the viscosity solution of Isaacs' equation (4.1):

$$(5.1) \quad \begin{cases} \frac{\partial}{\partial t} \omega(t, x) + H^-(t, x, \omega, D\omega, D^2\omega) = 0, & (t, x) \in [0, T] \times \mathbb{R}^n, \\ \omega(T, x) = \Phi(x), & x \in \mathbb{R}^n. \end{cases}$$

Recall that

$$H^-(t, x, y, p, X) = \sup_{u \in U} \inf_{v \in V} \left\{ \frac{1}{2} \text{tr}(\sigma \sigma^T(t, x, u, v) X) + p \cdot b(t, x, u, v) + f(t, x, y, p \cdot \sigma, u, v) \right\},$$

$t \in [0, T]$ ,  $x \in \mathbb{R}^n$ ,  $y \in \mathbb{R}$ ,  $p \in \mathbb{R}^n$ , and  $X \in \mathbf{S}^n$ . The functions  $b, \sigma, f$ , and  $\Phi$  are still supposed to satisfy (H3.1) and (H3.2), respectively.

We will prove the uniqueness for (5.1) in the following space of continuous functions:

$$\Theta = \{ \varphi \in C([0, T] \times \mathbb{R}^n) : \exists \tilde{A} > 0 \text{ such that } \lim_{|x| \rightarrow \infty} \varphi(t, x) \exp\{-\tilde{A}[\log(|x|^2 + 1)^{\frac{1}{2}}]^2\} = 0, \text{ uniformly in } t \in [0, T] \}.$$

This space of continuous functions is endowed with a growth condition which is slightly weaker than the assumption of polynomial growth but more restrictive than that of exponential growth. This growth condition was introduced by Barles, Buckdahn, and

Pardoux [3] to prove the uniqueness of the viscosity solution of an integro-partial differential equation associated with a decoupled FBSDE with jumps. It was shown in [3] that this kind of growth condition is optimal for the uniqueness and can, in general, not be weakened. We adapt the ideas developed in [3] to Isaacs' equation (5.1) to prove the uniqueness of the viscosity solution in  $\Theta$ . Since the proof of the uniqueness in  $\Theta$  for (4.2) is the same, we will restrict ourselves to only that of (5.1). Before stating the main result of this section, let us begin with two auxiliary lemmas. Denoting by  $K$  a Lipschitz constant of  $f(t, x, \cdot, \cdot, \cdot, \cdot)$ , that is, uniformly in  $(t, x)$ , we have the following.

LEMMA 5.1. *Let  $u_1 \in \Theta$  be a viscosity subsolution and  $u_2 \in \Theta$  be a viscosity supersolution of (5.1). Then the function  $\omega := u_1 - u_2$  is a viscosity subsolution of the equation*

$$(5.2) \quad \begin{cases} \frac{\partial}{\partial t} \omega(t, x) + \sup_{u \in U, v \in V} \left\{ \frac{1}{2} \text{tr}(\sigma \sigma^T(t, x, u, v) D^2 \omega) + D\omega \cdot b(t, x, u, v) + K|\omega| \right. \\ \quad \left. + K|D\omega \cdot \sigma(t, x, u, v)| \right\} = 0, & (t, x) \in [0, T) \times \mathbb{R}^n, \\ \omega(T, x) = 0, & x \in \mathbb{R}^n. \end{cases}$$

The proof of this lemma follows directly that of Lemma 3.7 in [3], and it is even simpler because, contrary to Lemma 3.7 in [3], we do not have any integral part here in (5.1). In analogy to [3] we also have the following.

LEMMA 5.2. *For any  $\tilde{A} > 0$ , there exists  $C_1 > 0$  such that the function*

$$\chi(t, x) = \exp[(C_1(T - t) + \tilde{A})\psi(x)],$$

with

$$\psi(x) = [\log((|x|^2 + 1)^{\frac{1}{2}}) + 1]^2, \quad x \in \mathbb{R}^n,$$

satisfies

$$(5.3) \quad \begin{aligned} & \frac{\partial}{\partial t} \chi(t, x) + \sup_{u \in U, v \in V} \left\{ \frac{1}{2} \text{tr}(\sigma \sigma^T(t, x, u, v) D^2 \chi) + D\chi \cdot b(t, x, u, v) + K\chi(t, x) \right. \\ & \quad \left. + K|D\chi(t, x) \cdot \sigma(t, x, u, v)| \right\} < 0 \quad \text{in } [t_1, T] \times \mathbb{R}^n, \quad \text{where } t_1 = T - \frac{\tilde{A}}{C_1}. \end{aligned}$$

*Proof.* By direct calculus we first deduce the following estimates for the first and second derivatives of  $\psi$ :

$$|D\psi(x)| \leq \frac{2[\psi(x)]^{\frac{1}{2}}}{(|x|^2 + 1)^{\frac{1}{2}}} \leq 4, \quad |D^2\psi(x)| \leq \frac{C(1 + [\psi(x)]^{\frac{1}{2}})}{|x|^2 + 1}, \quad x \in \mathbb{R}^n.$$

These estimates imply that, if  $t \in [t_1, T]$ ,

$$\begin{aligned} |D\chi(t, x)| & \leq (C_1(T - t) + \tilde{A})\chi(t, x)|D\psi(x)| \\ & \leq C\chi(t, x) \frac{[\psi(x)]^{\frac{1}{2}}}{(|x|^2 + 1)^{\frac{1}{2}}}, \end{aligned}$$

and, similarly,

$$|D^2\chi(t, x)| \leq C\chi(t, x) \frac{\psi(x)}{|x|^2 + 1}.$$

We should notice that the above estimates do not depend on  $C_1$  because of the definition of  $t_1$ . By virtue with the above estimates we have

$$\begin{aligned} & \frac{\partial}{\partial t} \chi(t, x) + \sup_{u \in U, v \in V} \left\{ \frac{1}{2} \text{tr}(\sigma \sigma^T(t, x, u, v) D^2 \chi) + D\chi \cdot b(t, x, u, v) + K\chi(t, x) \right. \\ & \quad \left. + K|D\chi(t, x) \cdot \sigma(t, x, u, v)| \right\} \\ & \leq -\chi(t, x) \{C_1 \psi(x) - C\psi(x) - C[\psi(x)]^{\frac{1}{2}} - K\} \\ & < -\chi(t, x) \{C_1 - [2C + K]\} \psi(x) < 0, \text{ if } C_1 > 2C + K \text{ large enough.} \quad \square \end{aligned}$$

Now we can prove the uniqueness theorem.

**THEOREM 5.3.** *We assume that (H3.1) and (H3.2) hold. Let  $u_1$  (resp.,  $u_2$ )  $\in \Theta$  be a viscosity subsolution (resp., supersolution) of (5.1). Then we have*

$$(5.4) \quad u_1(t, x) \leq u_2(t, x) \quad \text{for all } (t, x) \in [0, T] \times \mathbb{R}^n.$$

*Proof.* Let us put  $\omega := u_1 - u_2$ . Then we have, for some  $\tilde{A} > 0$ ,

$$\lim_{|x| \rightarrow \infty} \omega(t, x) e^{-\tilde{A}[\log((|x|^2 + 1))^{\frac{1}{2}}]^2} = 0,$$

uniformly with respect to  $t \in [0, T]$ . This implies, in particular, that, for any  $\alpha > 0$ ,  $\omega(t, x) - \alpha\chi(t, x)$  is bounded from above in  $[t_1, T] \times \mathbb{R}^n$  and that

$$M := \max_{[t_1, T] \times \mathbb{R}^n} (\omega - \alpha\chi)(t, x) e^{-K(T-t)}$$

is achieved at some point  $(t_0, x_0) \in [t_1, T] \times \mathbb{R}^n$  (depending on  $\alpha$ ). We now have to distinguish between two cases.

For the first case we suppose that:  $\omega(t_0, x_0) \leq 0$ , for any  $\alpha > 0$ .

Then, obviously  $M \leq 0$  and  $u_1(t, x) - u_2(t, x) \leq \alpha\chi(t, x)$  in  $[t_1, T] \times \mathbb{R}^n$ . Consequently, letting  $\alpha$  tend to zero we obtain

$$u_1(t, x) \leq u_2(t, x) \quad \text{for all } (t, x) \in [t_1, T] \times \mathbb{R}^n.$$

For the second case we assume that there exists some  $\alpha > 0$  such that  $\omega(t_0, x_0) > 0$ .

We notice that  $\omega(t, x) - \alpha\chi(t, x) \leq (\omega(t_0, x_0) - \alpha\chi(t_0, x_0))e^{-K(t-t_0)}$  in  $[t_1, T] \times \mathbb{R}^n$ . Then, putting

$$\varphi(t, x) = \alpha\chi(t, x) + (\omega - \alpha\chi)(t_0, x_0)e^{-K(t-t_0)}$$

we get  $\omega - \varphi \leq 0 = (\omega - \varphi)(t_0, x_0)$  in  $[t_1, T] \times \mathbb{R}^n$ . Consequently, since  $\omega$  is a viscosity subsolution of (5.2), from Lemma 5.1 we have

$$\begin{aligned} & \frac{\partial}{\partial t} \varphi(t_0, x_0) + \sup_{u \in U, v \in V} \left\{ \frac{1}{2} \text{tr}(\sigma \sigma^T(t_0, x_0, u, v) D^2 \varphi(t_0, x_0)) + D\varphi(t_0, x_0) \cdot b(t_0, x_0, u, v) \right. \\ & \quad \left. + K|\varphi(t_0, x_0)| + K|D\varphi(t_0, x_0) \cdot \sigma(t_0, x_0, u, v)| \right\} \geq 0. \end{aligned}$$

Moreover, due to our assumption that  $\omega(t_0, x_0) > 0$  and since  $\omega(t_0, x_0) = \varphi(t_0, x_0)$ , we can replace  $K|\varphi(t_0, x_0)|$  by  $K\varphi(t_0, x_0)$  in the above formula. Then, from the definition of  $\varphi$  and Lemma 5.2,

$$0 \leq \alpha \left\{ \frac{\partial \chi}{\partial t}(t_0, x_0) + \sup_{u \in U, v \in V} \left\{ \frac{1}{2} \text{tr}(\sigma \sigma^T(t_0, x_0, u, v) D^2 \chi(t_0, x_0)) + D\chi(t_0, x_0) \cdot b(t_0, x_0, u, v) \right. \right. \\ \left. \left. + K\chi(t_0, x_0) + K|D\chi(t_0, x_0) \cdot \sigma(t_0, x_0, u, v)| \right\} \right\} < 0,$$

which is a contradiction. Finally, by applying successively the same argument on the interval  $[t_2, t_1]$ , with  $t_2 = (t_1 - \frac{\tilde{A}}{C_1})^+$ , and then, if  $t_2 > 0$ , on  $[t_3, t_2]$ , with  $t_3 = (t_2 - \frac{\tilde{A}}{C_1})^+$ , etc., we get

$$u_1(t, x) \leq u_2(t, x), \quad (t, x) \in [0, T] \times \mathbb{R}^n.$$

Thus, the proof is complete.  $\square$

*Remark 5.1.* Obviously, since the lower value function  $W(t, x)$  is of at most linear growth, it belongs to  $\Theta$ , and so  $W(t, x)$  is the unique viscosity solution in  $\Theta$  of (5.1). Similarly we get that the upper value function  $U(t, x)$  is the unique viscosity solution in  $\Theta$  of (4.2). On the other hand, since  $H^- \leq H^+$ , any viscosity solution of (4.2) is a supersolution of (5.1). Then, again from Theorem 5.3, it follows that  $W \leq U$ . This justifies calling  $W$  the lower value function and  $U$  the upper value function.

*Remark 5.2.* If the Isaacs' condition holds, that is, if for all  $(t, x, y, p, X) \in [0, T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbf{S}^n$

$$H^-(t, x, y, p, X) = H^+(t, x, y, p, X),$$

then (5.1) and (4.2) coincide, and from the uniqueness in  $\Theta$  of viscosity solution it follows that the lower value function  $W(t, x)$  equals the upper value function  $U(t, x)$  which means that the associated stochastic differential game has a value.

*Remark 5.3.* Let us assume that the coefficient of BSDE (3.5)  $f(t, x, y, z, u, v) \equiv f(t, x, u, v)$  is independent of  $(y, z)$  and denote by  $\tilde{W}(t, x)$  (resp.,  $\tilde{U}(t, x)$ ) the lower value function (resp., the upper value function) defined by Fleming and Souganidis [14]; see Remark 3.4. It is shown in [14] that  $\tilde{W}(t, x)$  is a viscosity solution in  $\Theta$  of (5.1) and  $\tilde{U}(t, x)$  a viscosity solution in  $\Theta$  of (4.2). Then, due to Theorem 5.3,  $W(t, x) = \tilde{W}(t, x)$  and  $U(t, x) = \tilde{U}(t, x)$ ,  $(t, x) \in [0, T] \times \mathbb{R}^n$ . Moreover, if the Isaacs' condition holds, then  $W(t, x) = \tilde{W}(t, x) = \tilde{U}(t, x) = U(t, x)$ .

### Appendix A. Forward-backward SDEs (FBSDEs).

In this section we give an overview over basic results which are necessary for us on BSDEs associated with forward SDEs. We consider measurable functions  $b : [0, T] \times \Omega \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  and  $\sigma : [0, T] \times \Omega \times \mathbb{R}^n \rightarrow \mathbb{R}^{n \times d}$ , which are supposed to satisfy the following conditions:

(H6.1)

- (i)  $b(\cdot, 0)$  and  $\sigma(\cdot, 0)$  are  $\mathcal{F}_t$ -adapted processes, and there exists some constant  $C > 0$  such that  $|b(t, x)| + |\sigma(t, x)| \leq C(1 + |x|)$ , a.s., for all  $0 \leq t \leq T$ ,  $x \in \mathbb{R}^n$ ;
- (ii)  $b$  and  $\sigma$  are Lipschitz in  $x$ ; i.e., there is some constant  $C > 0$  such that

$$|b(t, x) - b(t, x')| + |\sigma(t, x) - \sigma(t, x')| \leq C|x - x'|, \text{ a.s.,} \\ \text{for all } 0 \leq t \leq T, \ x, \ x' \in \mathbb{R}^n.$$

We now consider the following SDE parameterized by the initial condition  $(t, \zeta) \in [0, T] \times L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ :

$$(A.1) \quad \begin{cases} dX_s^{t, \zeta} &= b(s, X_s^{t, \zeta})ds + \sigma(s, X_s^{t, \zeta})dB_s, \quad s \in [t, T], \\ X_t^{t, \zeta} &= \zeta. \end{cases}$$

Under the assumption (H6.1), SDE (A.1) has a unique strong solution, and, for any  $p \geq 2$ , there exists  $C_p \in \mathbb{R}$  such that, for any  $t \in [0, T]$  and  $\zeta, \zeta' \in L^p(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ ,

$$(A.2) \quad \begin{aligned} E \left[ \sup_{t \leq s \leq T} |X_s^{t, \zeta} - X_s^{t, \zeta'}|^p | \mathcal{F}_t \right] &\leq C_p |\zeta - \zeta'|^p, \quad a.s., \\ E \left[ \sup_{t \leq s \leq T} |X_s^{t, \zeta}|^p | \mathcal{F}_t \right] &\leq C_p (1 + |\zeta|^p), \quad a.s. \end{aligned}$$

These well-known standard estimates can be consulted, for instance, in Ikeda and Watanabe [18, pp. 166–168] and also in Karatzas and Shreve [19, pp. 289–290].

We emphasize that the constant  $C_p$  in (A.2) depends only on the Lipschitz and the growth constants of  $b$  and  $\sigma$ . Let now be given two real-valued functions  $f(t, x, y, z)$  and  $\Phi(x)$  which shall satisfy the following conditions:

(H6.2)

- (i)  $\Phi : \Omega \times \mathbb{R}^n \rightarrow \mathbb{R}$  is an  $\mathcal{F}_T \otimes \mathcal{B}(\mathbb{R}^n)$ -measurable random variable and  $f : [0, T] \times \Omega \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$  is a measurable process such that  $f(\cdot, x, y, z)$  is  $\mathcal{F}_t$ -adapted, for all  $(x, y, z) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d$ ;
- (ii) There exists a constant  $C > 0$  such that  $|f(t, x, y, z) - f(t, x', y', z')| + |\Phi(x) - \Phi(x')| \leq C(|x - x'| + |y - y'| + |z - z'|)$ , *a.s.*, for all  $0 \leq t \leq T$ ,  $x, x' \in \mathbb{R}^n$ ,  $y, y' \in \mathbb{R}$  and  $z, z' \in \mathbb{R}^d$ ;
- (iii)  $f$  and  $\Phi$  satisfy a linear growth condition; i.e., there exists some  $C > 0$  such that,  $dt \times dP$ -a.e., for all  $x \in \mathbb{R}^n$ ,  $|f(t, x, 0, 0)| + |\Phi(x)| \leq C(1 + |x|)$ .

With the help of the above assumptions we can verify that the coefficient  $f(s, X_s^{t, \zeta}, y, z)$  satisfies the hypotheses (A.1) and (A.2) and  $\xi = \Phi(X_T^{t, \zeta}) \in L^2(\Omega, \mathcal{F}_T, P; \mathbb{R})$ . Therefore, the following BSDE possesses a unique solution:

$$(A.3) \quad \begin{cases} -dY_s^{t, \zeta} &= f(s, X_s^{t, \zeta}, Y_s^{t, \zeta}, Z_s^{t, \zeta})ds - Z_s^{t, \zeta}dB_s, \quad s \in [t, T], \\ Y_T^{t, \zeta} &= \Phi(X_T^{t, \zeta}). \end{cases}$$

**PROPOSITION A.1.** *We suppose that the hypotheses (H6.1) and (H6.2) hold. Then, for any  $0 \leq t \leq T$  and the associated initial conditions  $\zeta, \zeta' \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , we have the following estimates:*

- (i)  $E \left[ \sup_{t \leq s \leq T} |Y_s^{t, \zeta}|^2 + \int_t^T |Z_s^{t, \zeta}|^2 ds | \mathcal{F}_t \right] \leq C(1 + |\zeta|^2), \quad a.s.;$
- (ii)  $E \left[ \sup_{t \leq s \leq T} |Y_s^{t, \zeta} - Y_s^{t, \zeta'}|^2 + \int_t^T |Z_s^{t, \zeta} - Z_s^{t, \zeta'}|^2 ds | \mathcal{F}_t \right] \leq C|\zeta - \zeta'|^2, \quad a.s.$

*In particular,*

$$(A.4) \quad \text{(iii) } |Y_t^{t, \zeta}| \leq C(1 + |\zeta|), \quad a.s.; \quad \text{(iv) } |Y_t^{t, \zeta} - Y_t^{t, \zeta'}| \leq C|\zeta - \zeta'|, \quad a.s.,$$

where the constant  $C > 0$  depends only on the Lipschitz and the growth constants of  $b$ ,  $\sigma$ ,  $f$ , and  $\Phi$ .

For the proof the reader is referred to Proposition 4.1 of Peng [26]; a similar result can be found in El Karoui, Peng, and Quenez [12, Proposition 4.1].

Let us now introduce the random field:

$$(A.5) \quad u(t, x) = Y_s^{t,x}|_{s=t}, \quad (t, x) \in [0, T] \times \mathbb{R}^n,$$

where  $Y^{t,x}$  is the solution of BSDE (A.3) with  $x \in \mathbb{R}^n$  at the place of  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ .

As a consequence of Proposition A.1 we have that, for all  $t \in [0, T]$ ,  $P$ -a.s.,

$$(A.6) \quad \begin{aligned} (i) \quad & |u(t, x) - u(t, y)| \leq C|x - y| \text{ for all } x, y \in \mathbb{R}^n; \\ (ii) \quad & |u(t, x)| \leq C(1 + |x|) \text{ for all } x \in \mathbb{R}^n. \end{aligned}$$

*Remark A.1.* In the general situation  $u$  is an adapted random function; that is, for any  $x \in \mathbb{R}^n$ ,  $u(\cdot, x)$  is an  $\mathcal{F}_t$ -adapted real-valued process. Indeed, recall that  $b$ ,  $\sigma$ , and  $f$  all are  $\mathbb{F}$ -adapted random functions while  $\Phi$  is  $\mathcal{F}_T$ -measurable. On the other hand, it is well known that, under the additional assumption that the functions

$$(H6.3) \quad b, \sigma, f, \text{ and } \Phi \text{ are deterministic,}$$

$u$  is also a deterministic function of  $(t, x)$ .

The random field  $u$  and  $Y^{t,\zeta}$ ,  $(t, \zeta) \in [0, T] \times L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , are related by the following theorem.

**THEOREM A.2.** *Under the assumptions (H6.1) and (H6.2), for any  $t \in [0, T]$  and  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , we have*

$$(A.7) \quad u(t, \zeta) = Y_t^{t,\zeta}, \quad P\text{-a.s.}$$

The proof of Theorem A.2 can be found in Peng [26]; we give it for the reader's convenience. It makes use of the following definition.

**DEFINITION A.3.** *For any  $t \in [0, T]$ , a sequence  $\{A_i\}_{i=1}^N \subset \mathcal{F}_t$  (with  $1 \leq N \leq \infty$ ) is called a partition of  $(\Omega, \mathcal{F}_t)$  if  $\cup_{i=1}^N A_i = \Omega$  and  $A_i \cap A_j = \emptyset$ , whenever  $i \neq j$ .*

*Proof of Theorem A.2.* We first consider the case where  $\zeta$  is a simple random variable of the form

$$(A.8) \quad \zeta = \sum_{i=1}^N x_i \mathbf{1}_{A_i},$$

where  $\{A_i\}_{i=1}^N$  is a finite partition of  $(\Omega, \mathcal{F}_t)$  and  $x_i \in \mathbb{R}^n$ , for  $1 \leq i \leq N$ .

For each  $i$ , we put  $(X_s^i, Y_s^i, Z_s^i) \equiv (X_s^{t,x_i}, Y_s^{t,x_i}, Z_s^{t,x_i})$ . Then  $X^i$  is the solution of the SDE

$$X_s^i = x_i + \int_t^s b(r, X_r^i) dr + \int_t^s \sigma(r, X_r^i) dB_r, \quad s \in [t, T],$$

and  $(Y^i, Z^i)$  is the solution of the associated BSDE

$$Y_s^i = \Phi(X_T^i) + \int_s^T f(r, X_r^i, Y_r^i, Z_r^i) dr - \int_s^T Z_r^i dB_r, \quad s \in [t, T].$$

The above two equations are multiplied by  $\mathbf{1}_{A_i}$  and summed up with respect to  $i$ . Thus, taking into account that  $\sum_i \varphi(x_i) \mathbf{1}_{A_i} = \varphi(\sum_i x_i \mathbf{1}_{A_i})$ , we get

$$\sum_{i=1}^N \mathbf{1}_{A_i} X_s^i = \sum_{i=1}^N x_i \mathbf{1}_{A_i} + \int_t^s b\left(r, \sum_{i=1}^N \mathbf{1}_{A_i} X_r^i\right) dr + \int_t^s \sigma\left(r, \sum_{i=1}^N \mathbf{1}_{A_i} X_r^i\right) dB_r$$

and

$$\begin{aligned} \sum_{i=1}^N \mathbf{1}_{A_i} Y_s^i &= \Phi \left( \sum_{i=1}^N \mathbf{1}_{A_i} X_T^i \right) + \int_s^T f \left( r, \sum_{i=1}^N \mathbf{1}_{A_i} X_r^i, \sum_{i=1}^N \mathbf{1}_{A_i} Y_r^i, \sum_{i=1}^N \mathbf{1}_{A_i} Z_r^i \right) dr \\ &\quad - \int_s^T \sum_{i=1}^N \mathbf{1}_{A_i} Z_r^i dB_r. \end{aligned}$$

Then the strong uniqueness property of the solution of the SDE and the BSDE yields

$$X_s^{t,\zeta} = \sum_{i=1}^N X_s^i \mathbf{1}_{A_i}, \quad (Y_s^{t,\zeta}, Z_s^{t,\zeta}) = \left( \sum_{i=1}^N \mathbf{1}_{A_i} Y_s^i, \sum_{i=1}^N \mathbf{1}_{A_i} Z_s^i \right), \quad s \in [t, T].$$

Finally, from  $u(t, x_i) = Y_t^i$ ,  $1 \leq i \leq N$ , we deduce that

$$Y_t^{t,\zeta} = \sum_{i=1}^N Y_t^i \mathbf{1}_{A_i} = \sum_{i=1}^N u(t, x_i) \mathbf{1}_{A_i} = u \left( t, \sum_{i=1}^N x_i \mathbf{1}_{A_i} \right) = u(t, \zeta).$$

Therefore, for simple random variables, we have the desired result.

Given a general  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$  we can choose a sequence of simple random variables  $\{\zeta_i\}$  which converges to  $\zeta$  in  $L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ . Consequently, from the estimates (A.4) and (A.6) and the first step of the proof, we have

$$\begin{aligned} E|Y_t^{t,\zeta_i} - Y_t^{t,\zeta}|^2 &\leq CE|\zeta_i - \zeta|^2 \rightarrow 0, \quad i \rightarrow \infty, \\ E|u(t, \zeta_i) - u(t, \zeta)|^2 &\leq CE|\zeta_i - \zeta|^2 \rightarrow 0, \quad i \rightarrow \infty, \end{aligned}$$

and

$$Y_t^{t,\zeta_i} = u(t, \zeta_i), \quad i \geq 1.$$

Then the proof is complete.  $\square$

*Remark A.2.* Under (H6.1), (H6.2), and (H6.3) we know  $u(t, x)$  is  $\frac{1}{2}$ -Hölder continuous in  $t$ : There exists a constant  $C$  such that, for every  $x \in \mathbb{R}^n$ ,  $t, t' \in [0, T]$ ,

$$|u(t, x) - u(t', x)| \leq C(1 + |x|)|t - t'|^{\frac{1}{2}}.$$

This inequality can be proved with the help of Theorem A.2. Since, on the other hand, a similar result but in a more general setting is proved (see Theorem 3.10), we do not give the proof here.

Although it is not used in our paper, let us also mention that, in the case of random coefficients  $b, \sigma, f$ , and  $\Phi$  the following holds true.

*Remark A.3.* Let us suppose in addition to the assumptions (H6.1) and (H6.2) that  $\sigma(\omega, t, \cdot)$  and  $b(\omega, t, \cdot)$  are continuously differentiable with a Lipschitz derivative such that, for some constant  $C$ ,

$$\begin{aligned} |D_x \sigma(\omega, t, x)| + |D_x b(\omega, t, x)| &\leq C, \quad \text{dtd}P\text{-a.e., for all } x \in \mathbb{R}^n; \\ D_x \sigma(\omega, t, \cdot), D_x b(\omega, t, \cdot) &\text{ are Lipschitz, uniformly in } (\omega, t). \end{aligned}$$

Then the random field  $u(\omega, t, x) : \Omega \times [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  possesses a continuous version.

The proof of the above property uses a standard argument for BSDEs based on the fact that  $u(s, x)$  can be written as  $Y_s^{t, \overline{X}_s^{t,x}}$  for  $0 < s - t$  small enough, where  $\overline{X}_s^{t,x}$  denotes the local inversion of the stochastic flow generated by (A.1):  $\overline{X}_s^{t,x} = (X_s^{t,\cdot})^{-1}(x)$ .

**Acknowledgments.** Juan Li thanks the Laboratory of Mathematics of University of Brest, France, for its hospitality during her one-year stay in France and expresses her gratitude to the General Council of Finistère for financing this stay in France.

## REFERENCES

- [1] K. BAHLALI, *Backward stochastic differential equations with locally Lipschitz coefficient*, C. R. Acad. Sci. Paris Ser. I Math., 333 (2001), pp. 481–486.
- [2] K. BAHLALI, E. H. ESSAKY, M. HASSANI, AND E. PARDOUX, *Existence, uniqueness and stability of backward stochastic differential equations with locally monotone coefficient*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 757–762.
- [3] G. BARLES, R. BUCKDAHN, AND E. PARDOUX, *Backward stochastic differential equations and integral-partial differential equations*, Stoch. Stoch. Rep., 60 (1997), pp. 57–83.
- [4] E. BAYRAKTAR AND H. V. POOR, *Stochastic differential games in a non-Markovian setting*, SIAM J. Control Optim., 43 (2005), pp. 1737–1756.
- [5] S. BROWNE, *Stochastic differential portfolio games*, J. Appl. Probab., 37 (2000), pp. 126–147.
- [6] R. BUCKDAHN, P. CARDALIAGUET, AND C. RAINER, *Nash equilibrium payoffs for nonzero-sum stochastic differential games*, SIAM J. Control Optim., 43 (2004), pp. 624–642.
- [7] P. CHERIDITO, H. M. SONER, N. TOUZI, AND N. VICTOIR, *Second-order backward stochastic differential equations and fully nonlinear parabolic PDEs*, Comm. Pure Appl. Math., 60 (2007), pp. 1081–1110.
- [8] M. G. CRANDALL, H. ISHII, AND P. L. LIONS, *User’s guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.
- [9] C. DELLACHERIE, *Sur l’existence de certains essinf et essup de familles de processus mesurables*, Sem. Probab. XII, Lecture Notes in Math. 649, Springer-Verlag, Berlin, New York, 1977.
- [10] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. Part I: General Theory*, Wiley-Interscience, New York, 1957.
- [11] E. EKSTRÖM AND G. PESKIR, *Optimal stopping games for Markov processes*, SIAM J. Control Optim., to appear.
- [12] N. EL KAROUI, S. PENG, AND M. C. QUENEZ, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.
- [13] L. C. EVANS AND P. E. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton–Jacobi–Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.
- [14] W. H. FLEMING AND P. E. SOUGANIDIS, *On the existence of value functions of two-player, zero-sum stochastic differential games*, Indiana Univ. Math. J., 38 (1989), pp. 293–314.
- [15] A. FRIEDMAN, *Differential Games*, Wiley, New York, 1971.
- [16] S. HAMADENE AND J. P. LEPELTIER, *Zero-sum stochastic differential games and backward equations*, Systems Control Lett., 24 (1995), pp. 259–263.
- [17] S. HAMADENE, J. P. LEPELTIER, AND S. PENG, *BSDEs with continuous coefficients and stochastic differential games*, in Backward Stochastic Differential Equations, Pitman Res. Notes Math. Ser. 364, N. El Karoui and L. Mazliak, eds., Longman, Harlow, UK, 1997, pp. 115–128.
- [18] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland-Kodansha, Amsterdam, Tokyo, 1989.
- [19] I. KARATZAS AND S. E. SHREVE, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1987.
- [20] I. KARATZAS AND S. E. SHREVE, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.
- [21] I. KARATZAS AND W. SUDDERTH, *The controller-and-stopper game for a linear diffusion*, Ann. Probab., 29 (2001), pp. 1111–1127.
- [22] I. KARATZAS AND M. ZAMFIRESCU, *Martingale approach to stochastic control with discretionary stopping*, Appl. Math. Optim., 53 (2006), pp. 163–184.
- [23] R. ISAACS, *Differential Games*, Wiley, New York, 1965.
- [24] E. PARDOUX AND S. PENG, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 61–74.



- [25] E. PARDOUX AND S. PENG, *Backward stochastic differential equations and quasilinear parabolic partial differential equations*, in Stochastic Partial Differential Equations and their Applications, Proceedings of the IFIP International Conference, Charlotte, NC 1991, Lect. Notes Control Inf. Sci. 176, Springer, Berlin, 1992, pp. 200–217.
- [26] S. PENG, *BSDE and stochastic optimizations*, in Topics in Stochastic Analysis, J. Yan, S. Peng, S. Fang, and L. Wu, eds., Science Press, Beijing, 1997 (in Chinese).
- [27] S. PENG, *A generalized dynamic programming principle and Hamilton–Jacobi–Bellman equation*, Stoch. Stoch. Rep., 38 (1992), pp. 119–134.

## FINITE HORIZON ROBUST STATE ESTIMATION FOR UNCERTAIN FINITE-ALPHABET HIDDEN MARKOV MODELS WITH CONDITIONAL RELATIVE ENTROPY CONSTRAINTS\*

LI XIE<sup>†</sup>, VALERY A. UGRINOVSKII<sup>‡</sup>, AND IAN R. PETERSEN<sup>‡</sup>

**Abstract.** We consider a robust state estimation problem for time-varying uncertain discrete-time, homogeneous, first-order, finite-state finite-alphabet hidden Markov models (HMMs). A class of time-varying uncertain HMMs is considered in which the uncertainty is sequentially described by a conditional relative entropy constraint on perturbed conditional probability measures given a realized observation sequence. For this class of uncertain HMMs, the robust state estimation problem is formulated as a constrained optimization problem. Using a Lagrange multiplier technique and a variational formula for conditional relative entropy, the above problem is converted into an unconstrained optimization problem and a problem related to partial information risk-sensitive filtering. A measure transformation technique and an information state method are employed to solve this equivalent problem related to risk-sensitive filtering. A characterization of the solution to the robust state estimation problem is also presented.

**Key words.** finite-alphabet hidden Markov models, robust state estimation, conditional relative entropy constraints, finite horizon

**AMS subject classifications.** 93E03, 93E10, 93E11, 62M05

**DOI.** 10.1137/040611379

**1. Introduction.** First state estimation algorithms for hidden Markov models (HMMs) are traditionally associated with problems of speech processing [15], and over the years many variations of the HMM state estimation problem were considered in the literature; we refer the reader to the comprehensive survey [9]. In loose terms, the problem is to estimate the state of a Markov process based on the observed data, and the optimization under various criteria is commonly regarded as one way to obtain such an estimate. The most common optimality criteria are the minimum symbol error probability criterion and minimum sequence error probability criterion. The first criterion results in the maximum a posteriori (MAP) symbol decision rule. Estimation of the state sequence using the second criterion results in a sequence estimate which is obtained by using the celebrated Viterbi algorithm.

In the control literature, the paradigm of optimal HMM state estimation has received much attention. For example, in [4, 8, 16] the HMM state estimation problem was formulated as an optimization problem in which a state estimate was sought to attain the minimum of a conditional mean square error cost or an exponential risk-sensitive cost, given a realized observation sequence. To solve such a risk-sensitive optimization problem, a probability measure transformation technique [8], an information state method, and dynamic programming were used [4, 16]. It was also observed using simulations [4] that, when the risk-sensitive parameter is within a

---

\*Received by the editors July 12, 2004; accepted for publication (in revised form) July 23, 2007; published electronically February 1, 2008. This work was supported by the Australian Research Council.

<http://www.siam.org/journals/sicon/47-1/61137.html>

<sup>†</sup>Department of Automatic Control, Beijing Institute of Technology, Beijing 100081, China (xieli.lixie@gmail.com).

<sup>‡</sup>School of Information Technology and Electrical Engineering, University of New South Wales at the Australian Defence Force Academy, ADFA, Canberra, ACT, 2600 Australia (valu@ee.adfa.edu.au, i.petersen@adfa.edu.au).

certain range, risk-sensitive filters exhibited useful robustness properties in an uncertain noise environment. Other types of the HMM estimation problem considered in the literature concern the estimation of states of discrete and continuous range partially observed Markov processes subject to Gaussian or Poisson measurement noises; e.g., see [5, 6, 12]. Another related line of research concerns applications to filtering and control of hybrid stochastic systems whose dynamics are modulated by Markov chains [26]. These problems, however, are substantially different from the one considered in this paper, both in the structure of the process being estimated and controlled and the information about the statistics of the process and noise disturbances.

Recently, the notion of relative entropy has been introduced into the areas of robust control and estimation for stochastic uncertain systems to describe uncertainties arising in stochastic control systems [1, 14, 19]. Petersen, James, and Dupuis [14] and Ugrinovskii and Petersen [19] introduced a new description of stochastic uncertain systems which uses the relative entropy between noise distributions to measure the uncertainty in uncertain systems; see also more recent papers [24, 25]. As demonstrated in these papers, the quantity of relative entropy allows one to quantify the amount of uncertainty in the underlying stochastic system and also allows for this quantitative information to be translated into a tractable robust control or robust filter design. The robust state estimation problem for a class of uncertain HMMs considered in this paper is defined under similar uncertain information constraints. We introduce a direct description for the uncertainty in HMMs which uses an a posteriori probability distance between HMMs. Then we show that the corresponding robust state estimation problem for the class of uncertain hidden Markov models under consideration can be solved using a framework similar to the one developed in [14, 19] since an HMM is essentially a probabilistic model. Techniques developed in [4, 8] are instrumental in this approach to robust state estimation.

For an HMM under consideration, we suppose that the true parameter set which defines the true HMM probability distribution is unknown, but a margin on the divergence from some design (reference) probability distribution is available. This is a key difference of our problem formulation from the one considered in [4, 8, 16]. Based on the relationship between HMM parameter sets and Kolmogorov measures of HMMs, this also means that the true Kolmogorov measure of the HMM under consideration is unknown (see [17] for further details concerning the Kolmogorov measure of a process and [23]). Instead, the estimation algorithm is constructed using a different reference parameter set satisfying some conditions. In our approach to the state estimation problem, since the observation sequence is available up to time step  $k$ , the difference between the true and reference parameter sets can be characterized using a posteriori probability distributions, and it is natural to compare the true and reference HMMs using related conditional probability measures. From this viewpoint, given an observation sequence, the true conditional probability measure related to the true Kolmogorov measure can be viewed as a perturbation of the reference conditional probability measure.

The relative entropy between conditional probability measures, or for short, the conditional relative entropy, can be viewed as an a posteriori measure of discrepancy between HMMs after a realized observation sequence is observed. In our robust state estimation problem, we will use the conditional relative entropy between the perturbed and reference conditional probability measures to define a constraint on the mismatch between the two HMMs, given measurements up to time step  $k$ . As a result, the uncertainty about the true conditional probability measure is sequentially described in terms of a conditional relative entropy constraint. More specifically, we define

a feasible uncertainty set for our robust state estimation problem as follows. Let  $y_{1,k} \triangleq \{y_1, y_2, \dots, y_k\}$  denote a realization of the observation process up to time step  $k$  which has positive probability to occur under the reference probability measure. For a given  $y_{1,k}$ , the feasible set of probability measures consists of all conditional probability measures given  $y_{1,k}$  whose conditional relative entropy with respect to the reference conditional probability measure satisfies a certain bound. The feasible set is therefore determined by the reference probability measure and certain design parameters. It is also a function of realizations of the observation process and a time step. We seek worst-case optimal estimates, and the robust state estimation problem under consideration is defined as a minimax optimization problem. In this problem the infimum with respect to the state estimate is taken over the state space. Also, the supremum with respect to conditional probability measures is taken over the feasible set.

The state estimator presented in this paper is based on processing realized observation sequences generated by the HMM with its true parameter set. The estimation error computed under the true conditional probability measure is guaranteed not to exceed a certain error cost value for the estimator. The upper bound on the estimation error is uniform, and it does not depend on the true probability distribution of the HMM, provided that the true probability distribution satisfies the specified conditional relative entropy constraint at each time step. This property indicates robustness of the estimator. In general, the robust state estimator which is presented is a finite horizon state estimator. The only exception to this is in the special case in which the only uncertainty is in the initial distribution. For this case, the robust state estimator can be generalized to the infinite horizon case if the state transition matrix satisfies an extra condition.

The organization of this paper is as follows. In section 2, we first introduce the notion of observability of states. Then the class of uncertain HMMs under consideration is defined and the conditional relative entropy uncertainty description is presented based on a calculation of the relative entropy between HMMs. The robust state estimation problem for this class of uncertain HMMs is introduced in section 3. Furthermore, we define an unconstrained optimization problem related to the underlying robust state estimation problem. This involves the use of a Lagrange multiplier. Using a variational formula for the conditional relative entropy, we show that the above unconstrained optimization problem with a positive Lagrange multiplier is equivalent to a problem related to partial information, risk-sensitive filtering. Furthermore, using a Lagrange multiplier technique, we convert the underlying robust state estimation problem into an unconstrained optimization problem which depends on a parameter which is optimized over  $[0, \infty)$ . It is interesting to note that in this associated risk-sensitive optimization problem, the Lagrange multiplier plays a role similar to that of a risk-sensitivity parameter. However, the resulting risk-sensitive optimization problem is different from the problem considered in [4], since in our case the running cost of the problem is also dependent on the Lagrange multiplier. Also, our solution method involves optimization over the Lagrange multiplier at each step of the algorithm. In contrast in [4], the risk-sensitivity parameter is kept constant once selected. In section 4, an information state method is used to compute the cost values of the above unconstrained optimization problem with a positive Lagrange multiplier. In section 5, a mathematical characterization of the solution to the robust state estimation problem under consideration is presented. In section 6, an illustrative example is given.

Throughout the paper, we use the standard conventions  $0 \log 0 = 0$ ,  $0 \cdot \infty = 0$ ,

$\frac{0}{0} = 1$ , and  $\infty \pm x = \infty$ , where  $|x| < \infty$ . We use upper case letters to denote random variables and lower case letters to denote values, i.e., realizations, of random variables.

**2. Finite-alphabet HMMs.** Consider an HMM with the parameter set  $\zeta = (A, C, p)$ . Here  $A = (a_{ji})$  is the one-step state transition matrix,  $C = (c_{qi})$  is the one-step state-observation transition matrix, and  $p = (p_i)$  is the initial distribution of the state process. Without loss of generality, we assume that the state space is  $\mathbb{E}_X = \{e_1, \dots, e_N\}$ , where  $e_i = (0, \dots, 1, \dots, 0)' \in \mathbb{R}^N$  with a 1 in the  $i$ th position and  $N$  is the size of  $\mathbb{E}_X$ . Also, we assume that the observation space is  $\mathbb{E}_Y = \{f_1, \dots, f_M\}$ , where  $f_j = (0, \dots, 1, \dots, 0)' \in \mathbb{R}^M$  with a 1 in the  $j$ th position and  $M$  is the size of  $\mathbb{E}_Y$ ; see [8].

Let  $(\Omega, \mathcal{B})$  be the canonical measurable space related to the HMM under consideration. That is,  $\Omega$  consists of all infinite sequences  $(x_0, \dots, x_k, \dots; y_1, \dots, y_k, \dots)$ , where  $x_0, \dots, x_k \in \mathbb{E}_X$  and  $y_1, \dots, y_k \in \mathbb{E}_Y$ , and  $\mathcal{B}$  is the smallest  $\sigma$ -algebra generated by cylinder sets with finite-dimensional measurable bases. Let a probability measure  $\mathbf{P}$  on  $(\Omega, \mathcal{B})$  be defined by extending finite-dimensional probability distribution functions of the HMM with the parameter set  $\zeta = (A, C, p)$  onto  $(\Omega, \mathcal{B})$  using Kolmogorov's extension theorem; see [17, 18]. The probability measure  $\mathbf{P}$  is called the Kolmogorov measure of the HMM with the parameter set  $\zeta$ . Obviously under  $\mathbf{P}$ , the pair of coordinate processes  $\{X_k, Y_{k+1}\}_{k \geq 0}$  is an HMM and  $\zeta$  is the parameter set of this HMM. The probability space  $(\Omega, \mathcal{B}, \mathbf{P})$  and the coordinate processes  $\{X_k, Y_{k+1}\}_{k \geq 0}$  are called the Kolmogorov model of the HMM with the parameter set  $\zeta$  and will be used to denote the HMM under consideration. The marginal distributions of the coordinate processes of this HMM can be expressed by

$$\mathbf{P}(X_0 = x_0, \dots, X_k = x_k; Y_1 = y_1, \dots, Y_k = y_k) = a_{x_k x_{k-1}} c_{y_k x_{k-1}} \dots a_{x_1 x_0} c_{y_1 x_0} p_{x_0},$$

where  $a_{x_k x_{k-1}} = \mathbf{P}(X_k = x_k | X_{k-1} = x_{k-1})$ ,  $c_{y_k x_{k-1}} = \mathbf{P}(Y_k = y_k | X_{k-1} = x_{k-1})$ , and  $p_{x_0} = \mathbf{P}(X_0 = x_0)$ .

Let  $\mathcal{G}_k = \sigma(X_0, \dots, X_k, Y_1, \dots, Y_k)$ . The sub- $\sigma$ -algebra  $\mathcal{Y}_k = \sigma\{Y_1, \dots, Y_k\}$  is generated by a finite partition of  $\Omega$ :

$$\mathcal{Y}_k = \sigma\{\{\omega : Y_1(\omega) = y_1, \dots, Y_k(\omega) = y_k | y_1 \in \mathbb{E}_Y, \dots, y_k \in \mathbb{E}_Y\}\}.$$

In accordance with this partitioning, each observation path  $y_{1,k}$  corresponds to an event  $\{\omega : Y_1(\omega) = y_1, \dots, Y_k(\omega) = y_k\}$ . In what follows we will be concerned with observation sequences and corresponding events of  $\mathcal{Y}_k$  which occur with positive probability under  $\mathbf{P}$ . Such sequences will be referred to as *feasible* sequences.

Let  $\mathbf{P}_k$  and  $\mathbf{P}_{\mathcal{Y}_k}$  denote the restrictions of  $\mathbf{P}$  to  $\mathcal{G}_k$  and  $\mathcal{Y}_k$ , respectively.

**2.1. The observability of states.** We now introduce the notion of observability of states for HMMs under consideration.

**DEFINITION 2.1.** *Under the probability measure  $\mathbf{P}$ , consider an HMM with  $\mathbb{E}_X$  and  $\mathbb{E}_Y$  as its state and observation spaces. We say a state  $x_k \in \mathbb{E}_X$  at time step  $k$  is observable in the sense of probability if for any feasible observation path  $y_{1,k}$ , we have  $\mathbf{P}(X_k = x_k; Y_1 = y_1, \dots, Y_k = y_k) > 0$ . We also say that the HMM is observable at time step  $k$  under the probability measure  $\mathbf{P}$  if each of its states is observable at time step  $k$  under  $\mathbf{P}$ .*

Next, we make an assumption about the observability of the HMMs.

**Assumption 2.1.** Under the probability measure  $\mathbf{P}$ , the HMM with the parameter set  $\zeta$  is observable at any time step  $k > 0$ .

Assumption 2.1 is a technical assumption which will be used later in the paper to guarantee that the supremum of a certain unconstrained optimization problem with zero Lagrange multiplier can be attained and a function related to an unconstrained optimization problem is strictly convex almost surely. We next state a necessary and sufficient condition in terms of  $\xi = (A, C, p)$  under which any state is observable at any time step  $k > 0$ .

LEMMA 2.1. *An HMM is observable at any time step  $k > 0$  if and only if the following conditions are satisfied:*

1. *For any  $z \in \mathbb{E}_X$  and for any  $y \in \mathbb{E}_Y$  satisfying  $D_y = \{x : c_{yx}p_x > 0\} \neq \emptyset$ , there exists an  $x \in D_y$  such that  $a_{zx} > 0$ . Here  $x$  is dependent on  $y, z$ .*
2. *Let  $\bar{D} = \{y : c_{yx} = 0 \forall x \in \mathbb{E}_X\}$ . For any  $y \in \mathbb{E}_Y - \bar{D}$  and  $z \in \mathbb{E}_X$ , there exists an  $x \in \mathbb{E}_X$  such that  $a_{zx}c_{yx} > 0$ . Here  $x$  is dependent on  $y, z$ .*

The proof of this lemma as well as proofs of other auxiliary lemmas used in what follows are given in the appendices. Since the sets  $\{\omega : Y_1(\omega) = y_1, \dots, Y_k(\omega) = y_k | y_1 \in \mathbb{E}_Y, \dots, y_k \in \mathbb{E}_Y\}$  form a finite partition of the sample space  $\Omega$  (that is,  $\Omega = \cup_{y_1 \in \mathbb{E}_Y, \dots, y_k \in \mathbb{E}_Y} \{\omega : Y_1(\omega) = y_1, \dots, Y_k(\omega) = y_k\}$ ), then there exists at least one feasible observation path  $(y_1, \dots, y_k)$ . It is obvious that the observability of a state  $x_k$  at time step  $k > 0$  implies that  $\mathbf{P}(X_k = x_k) > 0$ . This can be seen from

$$P(X_k = x_k) = \sum_{y_1 \in \mathbb{E}_Y} \cdots \sum_{y_k \in \mathbb{E}_Y} P(X_k = x_k; Y_1 = y_1, \dots, Y_k = y_k).$$

This fact will be used in the proof of Theorem 3.1. From now on we suppose Assumption 2.1 holds.

**2.2. Conditional relative entropy constraints.** Consider another HMM with  $\bar{\zeta} = (\bar{A}, \bar{C}, \bar{p})$  as its parameter set. We say that  $\bar{\zeta}$  is absolutely continuous with respect to  $\zeta$  and denote this fact by  $\bar{\zeta} \ll \zeta$ , provided that all of the following implications hold:  $\bar{a}_{ji} = 0$  if  $a_{ji} = 0$ ,  $\bar{c}_{qi} = 0$  if  $c_{qi} = 0$ , and  $\bar{p}_i = 0$  if  $p_i = 0$ . Let  $\bar{\mathbf{P}}$  be the Kolmogorov measure of the HMM with the parameter set  $\bar{\zeta}$ . Also, let  $\bar{\mathbf{P}}_k$  and  $\bar{\mathbf{P}}_{y_k}$  be the restrictions of  $\bar{\mathbf{P}}$  to  $\mathcal{G}_k$  and  $\mathcal{Y}_k$ , respectively.

In the robust state estimation problem considered in this paper the parameter sets  $\zeta$  and  $\bar{\zeta}$  will be regarded as the reference or nominal parameter set and the true (or perturbed) parameter set, respectively. This section introduces a bound on the discrepancy between these parameter sets and corresponding HMMs, given a realized feasible  $k$  step long observation sequence  $y_{1,k}$ . Note that since the conditional probability measures  $\mathbf{P}(\cdot | y_{1,k})$  and  $\bar{\mathbf{P}}(\cdot | y_{1,k})$  are probability measures on  $\mathcal{G}_k$ , the relative entropy between these probability measures, or conditional relative entropy,  $\mathcal{R}(\bar{\mathbf{P}}_k(\cdot | y_{1,k}) \| \mathbf{P}_k(\cdot | y_{1,k}))$  can be considered. The conditional relative entropy between two arbitrary conditional probability measures  $\mathbf{P}(\cdot | y_{1,k})$  and  $\bar{\mathbf{P}}(\cdot | y_{1,k})$  is defined by

$$(2.1) \quad \mathcal{R}(\bar{\mathbf{P}}_k(\cdot | y_{1,k}) \| \mathbf{P}_k(\cdot | y_{1,k})) \triangleq \begin{cases} \mathbf{E}^{\bar{\mathbf{P}}_k(\cdot | y_{1,k})} [\log \frac{d\bar{\mathbf{P}}_k(\cdot | y_{1,k})}{d\mathbf{P}_k(\cdot | y_{1,k})}] & \text{if } \bar{\mathbf{P}}_k(\cdot | y_{1,k}) \ll \mathbf{P}_k(\cdot | y_{1,k}); \\ +\infty & \text{otherwise,} \end{cases}$$

where  $\frac{d\bar{\mathbf{P}}_k(\cdot | y_{1,k})}{d\mathbf{P}_k(\cdot | y_{1,k})}$  is the Radon–Nikodym derivative of  $\bar{\mathbf{P}}_k(\cdot | y_{1,k})$  with respect to  $\mathbf{P}_k(\cdot | y_{1,k})$ .

The conditional relative entropy describes the discrepancy between the two HMMs with the parameter sets  $\bar{\zeta}$  and  $\zeta$  after a path of measurement process is realized.

Hence the conditional relative entropy  $\mathcal{R}(\bar{\mathbf{P}}_k(\cdot|y_{1,k})\|\mathbf{P}_k(\cdot|y_{1,k}))$  can be used as an a posteriori probability distance to describe the amount of uncertainty contained in the true (perturbed) HMM with the parameter set  $\bar{\zeta}$  relative to the nominal HMM with the parameter set  $\zeta$ . Using this a posteriori probability distance, given  $y_{1,k}$ , we define a set of conditional probability measures for the robust state estimation problem to be considered in the next section as follows:

$$(2.2) \quad \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k}) \triangleq \{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) : \bar{\mathbf{P}}_k(y_{1,k}) > 0, \mathcal{R}(\bar{\mathbf{P}}_k(\cdot|y_{1,k})\|\mathbf{P}_k(\cdot|y_{1,k})) < \infty\}.$$

In (2.2),  $\bar{\mathbf{P}}_k$  is an arbitrary probability measure defined on  $(\Omega, \mathcal{G}_k)$  and satisfying the condition  $\bar{\mathbf{P}}_k(y_{1,k}) > 0$ . Note that since  $\mathcal{G}_k$  is generated by a finite partition of  $\Omega$ , the condition  $\mathcal{R}(\bar{\mathbf{P}}_k(\cdot|y_{1,k})\|\mathbf{P}_k(\cdot|y_{1,k})) < \infty$  in (2.2) is equivalent to  $\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \ll \mathbf{P}_k(\cdot|y_{1,k})$ .

In general, the set (2.2) is larger than the set of Kolmogorov measures corresponding to HMM parameter sets  $\bar{\zeta}$  such that  $\bar{\zeta} \ll \zeta$ . If there exists a probability measure  $\bar{\mathbf{P}}$  on  $(\Omega, \mathcal{B})$  whose restriction on  $\mathcal{G}_k$  is  $\bar{\mathbf{P}}_k$ , then under  $\bar{\mathbf{P}}$ , the pair of the coordinate processes  $\{X_k, Y_{k+1}\}_{k \geq 0}$  may not be an HMM.

It was shown in Lemma 2.2 of [21] that for the two HMMs under consideration, if  $\bar{\zeta} \ll \zeta$ , then the relative entropy between the corresponding regular conditional probability measures  $\bar{\mathbf{P}}_k(\cdot|\mathcal{Y}_k)(\omega)$  and  $\mathbf{P}_k(\cdot|\mathcal{Y}_k)(\omega)$  (or “regular conditional relative entropy”) satisfies the following inequality:

$$(2.3) \quad \mathbf{E}^{\bar{\mathbf{P}}_k}[\mathcal{R}(\bar{\mathbf{P}}_k(\cdot|\mathcal{Y}_k)\|\mathbf{P}_k(\cdot|\mathcal{Y}_k))] \leq \bar{d}_0 + \sum_{l=1}^k \sum_{r=1}^N \bar{d}_r \mathbf{E}^{\bar{\mathbf{P}}_k}[\langle X_{l-1}, e_r \rangle].$$

In (2.3),  $\mathbf{E}^{\bar{\mathbf{P}}_k}[\cdot]$  denotes the expectation with respect to  $\bar{\mathbf{P}}_k$ ,  $\bar{d}_r = \sum_{s=1}^N \bar{a}_{sr} \log \frac{\bar{a}_{sr}}{\bar{a}_{sr}} + \sum_{q=1}^M \bar{c}_{qr} \log \frac{\bar{c}_{qr}}{\bar{c}_{qr}}$ , and  $\bar{d}_0 = \sum_{r=1}^N \bar{p}_r \log \frac{\bar{p}_r}{p_r}$ . The notation  $\langle X_{l-1}, e_r \rangle$  denotes the inner product satisfying  $\langle X_{l-1}, e_r \rangle = 0$  if  $X_{l-1} \neq e_r$ , otherwise 1, where  $e_r \in \mathbb{E}_X$ . For a feasible observation path  $y_{1,k}$ ,  $\mathcal{R}(\bar{\mathbf{P}}_k(\cdot|\mathcal{Y}_k)\|\mathbf{P}_k(\cdot|\mathcal{Y}_k))(\omega) = \mathcal{R}(\bar{\mathbf{P}}_k(\cdot|y_{1,k})\|\mathbf{P}_k(\cdot|y_{1,k}))$ . Inequality (2.3) provides a useful insight into how a conditional relative entropy constraint is to be defined to describe sequentially the mismatch between a perturbed conditional probability measure and the reference conditional probability measure given  $y_{1,k}$ . We now give a formal definition.

**DEFINITION 2.2.** Let  $d > 0$ ,  $d_r \geq 0$ ,  $r = 1, \dots, N$ , be given constants. These constants will be the design parameters in our robust state estimator design. Given a feasible sequence  $y_{1,k}$ , the conditional relative entropy of the conditional probability measure  $\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})$  with respect to  $\mathbf{P}_k(\cdot|y_{1,k})$  defines an admissible (perturbed) conditional probability measure if

$$(2.4) \quad \mathcal{R}(\bar{\mathbf{P}}_k(\cdot|y_{1,k})\|\mathbf{P}_k(\cdot|y_{1,k})) \leq d + \sum_{l=1}^k \sum_{r=1}^N d_r \mathbf{E}^{\bar{\mathbf{P}}_k}[\langle X_{l-1}, e_r \rangle | y_{1,k}].$$

That is, given  $y_{1,k}$ , if  $\bar{\mathbf{P}}_k(\cdot|y_{1,k})$  satisfies inequality (2.4), then  $\bar{\mathbf{P}}_k(\cdot|y_{1,k})$  is called an admissible conditional probability measure for this  $y_{1,k}$ .

A set of admissible conditional probability measures  $\bar{\mathbf{P}}_k(\cdot|y_{1,k})$  for the robust state estimation problem, which will be defined in the next section, consists of all conditional probability measures  $\bar{\mathbf{P}}_k(\cdot|y_{1,k})$  satisfying the above definition:

$$(2.5) \quad \Xi_k(y_{1,k}) \triangleq \{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) : \bar{\mathbf{P}}_k(y_{1,k}) > 0, \bar{\mathbf{P}}_k(\cdot|y_{1,k}) \text{ satisfies (2.4)}\}.$$

Obviously, it follows from (2.2) and (2.5) that  $\Xi_k(y_{1,k}) \subset \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})$ .

Inequality (2.4) is called the conditional relative entropy constraint for a feasible observation path  $y_{1,k}$ . The conditional relative entropy constraint (2.4) is introduced based on the insights gained from our study on probability distance problems for the HMMs (see [22, 23]), though inequality (2.3) does not imply that (2.4) holds for all feasible observation sequences. From Definition 2.2, a conditional probability measure may be admissible for a feasible observation path, but it may be inadmissible for another feasible observation path. In other words, the set consisting of all admissible conditional probability measures is a function of feasible observation sequences.

Observing inequality (2.4), the right-hand side describes a bound on the conditional relative entropy at time step  $k$ . This bound represents a weighted sum of the occupation probability estimates of the states  $e_r$  given the measurements up to time step  $k$  with respect to  $\bar{\mathbf{P}}_k(\cdot|y_{1,k})$ . The more often the state  $e_r$  occurs, the more discrepancy is allowed by the state  $e_r$ . The parameters  $d_r$  weight the contribution of each state into the bound. Condition (2.4) defines a constraint on the perturbed conditional probability measure in terms of the conditional relative entropy. Note that (2.4) defines a time-varying constraint. The reason for considering time-varying constraints is that the conditional relative entropy between HMMs may increase as the time step increases; see Theorem 3.2 in [22].

*Remark 2.1.* As a special case of the conditional relative entropy constraint defined by (2.4), the conditional relative entropy constraint (2.4) with zero  $d_r$  for  $r = 1, \dots, N$  describes the uncertainty in the initial distribution of the perturbed HMM only. Indeed, in this special case the conditional relative entropy between an admissible perturbed measure  $\bar{\mathbf{P}}_k(\cdot|y_{1,k})$  and the reference measure  $\mathbf{P}_k(\cdot|y_{1,k})$  is bounded uniformly in  $k$ . Since the conditional relative entropy generally increases with time, this can be only when the uncertainty does not affect the state-to-state and state-to-measurement transition probability matrices  $A$  and  $C$ .

**3. A robust state estimation problem.** In this section, we will present our main results. Consider an HMM on the canonical measurable space  $(\Omega, \mathcal{B})$ . We do not know the exact parameter set of this HMM. Instead, we choose the reference parameter set  $\zeta = (A, C, p)$  as its parameter set. Then under the reference Kolmogorov measure  $\mathbf{P}$ , the pair of the coordinate processes  $\{X_k, Y_{k+1}\}_{k \geq 0}$  is an HMM with  $\zeta = (A, C, p)$  as its parameter set.

Let  $\Psi_k(x, \xi) = (x - \xi)'Q_k(x - \xi)$ , where  $x$  and  $\xi$  take values in the state space  $\mathbb{E}_X$ , and  $Q_k$  is a positive definite matrix. The sequence  $Q_k$  is assumed to be bounded in  $k$ . Let  $\Xi_k(y_{1,k})$  be defined by (2.5). Given a feasible observation sequence  $y_{1,k}$ , we address the following robust state estimation problem: Find an estimate  $\hat{X}_k$  of the HMM state process  $X_k$  which can be computed in a recursive fashion, where  $\hat{X}_k$  takes values in  $\mathbb{E}_X$  and is a function of  $y_{1,k}$  (and hence  $\hat{X}_k$  is a  $\mathcal{Y}_k$ -measurable estimate of  $X_k$ ). This estimate is required to be such that given  $y_{1,k}$ , the estimate  $\hat{X}_k$  solves the following minimax problem subject to the constraint (2.4):

$$(3.1) \quad \hat{X}_k = \arg \inf_{\xi \in \mathbb{E}_X} \sup_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \Xi_k(y_{1,k})} \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)|y_{1,k}].$$

The optimization problem (3.1) is a constrained optimization problem in which the maximizer  $\bar{\mathbf{P}}_k(\cdot|y_{1,k})$  is constrained to take values from the set  $\Xi_k(y_{1,k})$ . As in [14], a Lagrange multiplier technique will be used to convert the constrained optimization problem (3.1) into an unconstrained optimization problem. This conversion will be considered in section 3.2. In the next subsection, we introduce such an unconstrained



optimization problem and present some properties of a cost functional related to this problem.

**3.1. An unconstrained optimization problem.** Given a  $\tau \in \mathbb{R}$  and a  $\xi \in \mathbb{E}_X$ , for a feasible observation sequence  $y_{1,k}$ , we define an augmented cost functional:

$$(3.2) \quad \begin{aligned} F(\bar{\mathbf{P}}_k(\cdot|y_{1,k}), \xi, \tau) &= \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)|y_{1,k}] - \tau \left( \mathcal{R}(\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \|\mathbf{P}_k(\cdot|y_{1,k})) \right. \\ &\quad \left. - d - \sum_{l=1}^k \sum_{r=1}^N d_r \mathbf{E}^{\bar{\mathbf{P}}_k}[\langle X_{l-1}, e_r \rangle | y_{1,k}] \right). \end{aligned}$$

The following optimization problem is instrumental in the derivation of a solution to the robust state estimation problem (3.1):

$$(3.3) \quad V_k(\tau, \xi, y_{1,k}) = \sup_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})} F(\bar{\mathbf{P}}_k(\cdot|y_{1,k}), \xi, \tau).$$

Unlike (3.1), here the supremum is taken over the entire set  $\mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})$  of feasible perturbation conditional probability measures.

For  $\tau > 0$ , we write

$$(3.4) \quad V_k(\tau, \xi, y_{1,k}) = \tau(W_k(\tau, \xi, y_{1,k}) + d),$$

where

$$(3.5) \quad \begin{aligned} W_k(\tau, \xi, y_{1,k}) &= \sup_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})} \left( \mathbf{E}^{\bar{\mathbf{P}}_k} \left[ \tau^{-1} \Psi_k(X_k, \xi) + \sum_{l=1}^k q(X_{l-1}) | y_{1,k} \right] \right. \\ &\quad \left. - \mathcal{R}(\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \|\mathbf{P}_k(\cdot|y_{1,k})) \right) \end{aligned}$$

and

$$(3.6) \quad q(x) \triangleq \sum_{r=1}^N d_r \langle x, e_r \rangle, \quad x \in \mathbb{E}_X.$$

In order to establish our results, a conditional version of the variational formula for probability measures will be used which is presented in Lemma 3.1. This lemma is a direct extension of Proposition 1.4.2 presented in [7, p. 33] for the standard relative entropy.

**LEMMA 3.1.** *Let  $g(\omega)$  be a bounded measurable function. With the above definitions, we have the following variational formula:*

$$(3.7) \quad \begin{aligned} -\log \mathbf{E}^{\mathbf{P}_k}[e^{-g}|y_{1,k}] &= \inf_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})} \left\{ \mathcal{R}(\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \|\mathbf{P}_k(\cdot|y_{1,k})) \right. \\ &\quad \left. + \mathbf{E}^{\bar{\mathbf{P}}_k}[g|y_{1,k}] \right\}. \end{aligned}$$

The variational formula for conditional relative entropy (3.7) will be used to calculate  $V_k(\tau, \xi, y_{1,k})$  defined by (3.3). The next theorem presents this result.

**THEOREM 3.1.** *The cost functional  $V_k(\tau, \xi, y_{1,k})$  defined by (3.3) can be calculated in two cases as follows:*

(i) When  $\tau = 0$ ,

$$(3.8) \quad V_k(0, \xi, y_{1,k}) = \sup_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})} \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)|y_{1,k}] = \max_{x \in \mathbb{E}_X} \Psi_k(x, \xi).$$

(ii) For any given  $\tau \in (0, \infty)$ ,

$$(3.9) \quad V_k(\tau, \xi, y_{1,k}) = \tau \left( \log \mathbf{E}^{\mathbf{P}_k} \left[ \exp \left( \tau^{-1} \Psi_k(X_k, \xi) + \sum_{l=1}^k q(X_{l-1}) \right) | y_{1,k} \right] + d \right).$$

*Proof.* We first consider the case  $\tau = 0$ . Since the state space  $\mathbb{E}_X$  of the process  $\{X_k\}_{k \geq 0}$  is finite,  $\Psi_k(x, \xi)$  is a bounded function, where  $\xi \in \mathbb{E}_X$ . Let  $x_k^* \in \mathbb{E}_X$  satisfy

$$\Psi_k(x_k^*, \xi) = \max_{x \in \mathbb{E}_X} \Psi_k(x, \xi).$$

Clearly, such an  $x_k^*$  exists and is a function of  $\xi$ . Then for all  $\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})$ ,

$$(3.10) \quad \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)|y_{1,k}] \leq \Psi_k(x_k^*, \xi).$$

Hence we conclude that  $V_k(0, \xi, y_{1,k}) \leq \Psi_k(x_k^*, \xi)$ . We now show that there exists a probability measure on  $(\Omega, \mathcal{G}_k)$  which attains the equality in (3.10). Let

$$(3.11) \quad D_\xi = \{\omega : X_k(\omega) = x_k^*\}.$$

Note that under Assumption 2.1, we have  $\mathbf{P}_k(D_\xi) > 0$ , and hence we can define the required probability measure  $\check{\mathbf{P}}_k$  on  $\mathcal{G}_k$  by letting  $\check{\mathbf{P}}_k(D) = \mathbf{P}_k(D|D_\xi)$  for any  $D \in \mathcal{G}_k$  and noting that  $\check{\mathbf{P}}_k(D|y_{1,k}) = \mathbf{P}_k(D|D_\xi, y_{1,k})$ ; see also Remark 3.1. Hence  $\check{\mathbf{P}}_k(D_\xi) = 1$  and  $\check{\mathbf{P}}_k(D_\xi|y_{1,k}) = 1$ . It follows from this definition that  $\check{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})$ . Furthermore, it can be shown using a direct calculation that

$$\mathbf{E}^{\check{\mathbf{P}}_k}[\Psi_k(X_k, \xi)|y_{1,k}] = \Psi_k(x_k^*, \xi) = \max_{x \in \mathbb{E}_X} \Psi_k(x, \xi)$$

from which claim (i) follows.<sup>1</sup>

To prove claim (ii), we note that for each given  $\tau \in (0, \infty)$ ,  $\xi$ , and  $k$ ,  $\tau^{-1} \Psi_k(X_k, \xi) + \sum_{l=1}^k q(X_{l-1})$  is a bounded measurable random variable. Then by applying (3.7) in Lemma 3.1 to (3.5), we have

$$(3.12) \quad W_k(\tau, \xi, y_{1,k}) = \log \mathbf{E}^{\mathbf{P}_k} \left[ \exp \left( \tau^{-1} \Psi_k(X_k, \xi) + \sum_{l=1}^k q(X_{l-1}) \right) | y_{1,k} \right].$$

Furthermore, (3.12) and (3.4) yield (3.9). This completes the proof of the theorem.  $\square$

*Remark 3.1.* Without Assumption 2.1, when  $\tau = 0$  the value of  $\max_{x \in \mathbb{E}_X} \Psi_k(x, \xi)$  may not be attained in the unconstrained optimization problem. To illustrate this fact, suppose there exists a unique  $x^*$  such that  $\Psi_k(x^*, \xi) = \max_{x \in \mathbb{E}_X} \Psi_k(x, \xi)$ . For any  $\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})$ , it follows from  $\mathcal{R}(\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \| \mathbf{P}_k(\cdot|y_{1,k})) < \infty$  and

<sup>1</sup>The above result can also be derived from a more general setting considered in Theorem 5.2 of [21].

Definition 2.1 that we must have  $\bar{\mathbf{P}}(\cdot|y_{1,k}) \ll \mathbf{P}(\cdot|y_{1,k})$ . Now suppose Assumption 2.1 does not hold, and therefore the set  $D_\xi$  defined by (3.11) can have zero probability under  $\mathbf{P}(\cdot|y_{1,k})$ . If  $\mathbf{P}(D_\xi|y_{1,k}) = 0$ , then  $\bar{\mathbf{P}}(D_\xi|y_{1,k}) = 0$ . As a result, for any  $\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})$ , we have

$$\begin{aligned} \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)|y_{1,k}] &= \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)I_{D_\xi}|y_{1,k}] + \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)I_{D_\xi^c}|y_{1,k}] \\ &= \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)I_{D_\xi^c}|y_{1,k}] < \max_{x \in \mathbb{E}_X} \Psi_k(x, \xi), \end{aligned}$$

where  $D_\xi^c$  is the complement of  $D_\xi$ . Thus Assumption 2.1 on the reference parameter set, which guarantees  $\mathbf{P}(D_\xi|y_{1,k}) > 0$ , also guarantees that the unconstrained optimization problem with  $\tau = 0$  attains the maximum.

**3.2. A solution to the constrained optimization problem.** Next, we will use the following standard result on constrained optimization to convert the constrained optimization problem (3.1) into the unconstrained optimization problem (3.3) dependent on a parameter  $\tau$ . The result and its proof can be found in [11, p. 217]. In [14] this result is presented as follows.

**LEMMA 3.2.** *Let  $\mathbb{X}$  be a linear vector space, and let  $\tilde{\Omega}$  be a convex subset of  $\mathbb{X}$ . Also, let  $f$  be a real-valued concave functional on  $\tilde{\Omega}$ , and let  $g$  be a real-valued convex functional on  $\tilde{\Omega}$ . Assume a point exists  $x_1 \in \tilde{\Omega}$  such that  $g(x_1) < 0$  (this is a constraint qualification condition), and let*

$$(3.13) \quad \mu_0 = \sup f(x) \quad \text{subject to } x \in \tilde{\Omega}, \quad g(x) \leq 0.$$

1. *If  $\mu_0$  is finite, then there exists a  $\tau \geq 0$  such that*

$$(3.14) \quad \mu_0 = \sup_{x \in \tilde{\Omega}} \{f(x) - \tau g(x)\}.$$

2. *If the supremum in (3.13) is achieved by an  $x_0 \in \tilde{\Omega}$ ,  $g(x_0) \leq 0$ , it is achieved by  $x_0$  in (3.14) and  $\tau g(x_0) = 0$ .*

Theorem 2.1 of [14] makes it possible to convert the constrained optimization problem (3.1) into the unconstrained optimization problem (3.3) dependent on a parameter  $\tau$  which is then optimized over  $[0, \infty)$ . We now use a similar argument to the one used in the proof of Theorem 2.1 of [14] to present the main result of this paper.

**THEOREM 3.2.** *Consider the robust state estimation problem defined by (3.1). For each  $k$ , we have*

$$(3.15) \quad \sup_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \Xi_k(y_{1,k})} \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)|y_{1,k}] = \inf_{\tau \in [0, \infty)} V_k(\tau, \xi, y_{1,k}).$$

Furthermore, the robust state estimate  $\hat{X}_k$  is given by

$$(3.16) \quad \hat{X}_k = \arg \inf_{\xi} \inf_{\xi \in \mathbb{E}_X} \inf_{\tau \in [0, \infty)} V_k(\tau, \xi, y_{1,k}).$$

*Proof.* Let

$$L_k(\xi, y_{1,k}) \triangleq \sup_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \Xi_k(y_{1,k})} \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)|y_{1,k}].$$

Since  $\Xi_k(y_{1,k}) \subset \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})$  and  $\tau \geq 0$ , it follows from (3.3) and (3.2) that

$$\begin{aligned} V_k(\tau, \xi, y_{1,k}) &= \sup_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})} F(\bar{\mathbf{P}}_k(\cdot|y_{1,k}), \xi, \tau) \\ (3.17) \quad &\geq \sup_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \Xi_k(y_{1,k})} F(\bar{\mathbf{P}}_k(\cdot|y_{1,k}), \xi, \tau) \geq L_k(\xi, y_{1,k}). \end{aligned}$$

Note that (3.17) holds for any  $\tau \in [0, \infty)$ . Hence

$$(3.18) \quad L_k(\xi, y_{1,k}) \leq \inf_{\tau \in [0, \infty)} V_k(\tau, \xi, y_{1,k}) \leq V_k(0, \xi, y_{1,k}) = \max_{x \in \mathbb{E}_X} \Psi_k(x, \xi).$$

Thus  $L_k(\xi, y_{1,k})$  is finite. Then for any  $\xi \in \mathbb{E}_X$ , applying part 1 of Lemma 3.2 to  $L_k(\xi, y_{1,k})$  with respect to the conditional probability measures  $\bar{\mathbf{P}}_k(\cdot|y_{1,k})$  since all required conditions are satisfied, it follows that a  $\tau_k^*(\xi, y_{1,k}) \in [0, \infty)$  exists such that

$$\begin{aligned} L_k(\xi, y_{1,k}) &= \sup_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k|y_{1,k})} F(\bar{\mathbf{P}}_k(\cdot|y_{1,k}), \xi, \tau_k^*(\xi, y_{1,k})) \\ (3.19) \quad &= V_k(\tau_k^*(\xi, y_{1,k}), \xi, y_{1,k}). \end{aligned}$$

Furthermore, for each  $\tau_k^*(\xi, y_{1,k}) \in [0, \infty)$ ,

$$\inf_{\tau \in [0, \infty)} V_k(\tau, \xi, y_{1,k}) \leq V_k(\tau_k^*(\xi, y_{1,k}), \xi, y_{1,k}).$$

Hence, the last line in (3.19) implies that

$$(3.20) \quad \inf_{\tau \in [0, \infty)} V_k(\tau, \xi, y_{1,k}) \leq L_k(\xi, y_{1,k}).$$

Combining the first inequality of (3.18) and (3.20) yields (3.15). The robust state estimate  $\hat{X}_k$  is given by (3.16). Obviously,  $\hat{X}_k$  is a function of  $y_{1,k}$ . Note that the robust state estimate may not be unique and  $\inf_{\tau \in [0, \infty)} V_k(\tau, \xi, y_{1,k})$  attains its infimum on the finite set  $\mathbb{E}_X$ . This completes the proof of the theorem.  $\square$

*Remark 3.2.* In what follows, we will observe that in general, the robust state estimator presented in this paper applies only to finite horizon problems except for the special case of  $d_r = 0$ , for  $r = 1, \dots, N$ , and under some additional special conditions. In section 5, we will discuss how to specify the maximum allowable time interval  $K$ , based on a necessary and sufficient condition for the existence of a certain stationary point in the optimization of the cost.

**4. Calculation of the unconstrained cost functional.** Theorem 3.2 shows that the solution to the constrained optimization problem (3.1) can be obtained from the solution to the unconstrained optimization problem on the right-hand side of (3.15). This requires that the quantity  $V_k(\tau, \xi, y_{1,k})$  be readily computable for all  $\tau \in [0, \infty)$ . Part (i) of Theorem 3.1 shows how this quantity can be computed for the case  $\tau = 0$ . In this section we consider the second case and calculate  $V_k(\tau, \xi, y_{1,k})$  for the case  $\tau > 0$  using (3.9). This is done by using the information state approach to risk-sensitive control and filtering.

**4.1. A change of probability measure.** Define a new probability measure  $\tilde{\mathbf{P}}$ , under which the pair of coordinate processes  $\{X_k, Y_{k+1}\}_{k \geq 0}$  has the parameter set  $(A, (1/M)_{M \times N}, p)$  and the  $\sigma$ -algebra  $\sigma\{Y_k\}$  is independent of  $\sigma\{X_0, \dots, X_k, Y_1, \dots, Y_{k-1}\}$  by letting

$$\tilde{\mathbf{P}}(X_0 = x_0, \dots, X_k = x_k; Y_1 = y_1, \dots, Y_k = y_k) = \frac{1}{M^k} a_{x_k x_{k-1}} \dots a_{x_1 x_0} p_{x_0}.$$

This probability measure is related to the reference probability measure  $\mathbf{P}$  via the Radon–Nikodym derivative  $\frac{d\tilde{\mathbf{P}}_k}{d\mathbf{P}_k} = \bar{\Lambda}_k$ , where

$$\bar{\Lambda}_k = \prod_{l=0}^k \bar{\lambda}_l, \quad \bar{\lambda}_l = \prod_{r=1}^N \prod_{q=1}^M (M c_{qr})^{\langle X_{l-1}, e_r \rangle \langle Y_l, f_q \rangle}, \quad \text{and} \quad \bar{\lambda}_0 = 1;$$

see [8]. For a feasible observation path  $y_{1,k}$ , using the conditional Bayes theorem [8], the cost functional  $V_k(\tau, \xi, y_{1,k})$  given by (3.9) can be reformulated under the new probability measure  $\tilde{\mathbf{P}}_k$  as

$$V_k(\tau, \xi, y_{1,k}) = \tau \left( \log \left( \frac{\mathbf{E}^{\tilde{\mathbf{P}}_k} [\bar{\Lambda}_k \exp(\tau^{-1} \Psi_k(X_k, \xi) + \sum_{l=1}^k q(X_{l-1})) | y_{1,k}]}{\mathbf{E}^{\tilde{\mathbf{P}}_k} [\bar{\Lambda}_k | y_{1,k}]} \right) + d \right).$$

**4.2. Information states.** We now apply the information state method for the HMMs of [8] and [4] to evaluate the expression on the right-hand side of the above equation.

**DEFINITION 4.1.** Define  $\alpha_k(y_{1,k}) \triangleq (\alpha_k(e_1, y_{1,k}), \dots, \alpha_k(e_n, y_{1,k}))'$  and  $\beta_k(y_{1,k}) \triangleq (\beta(e_1, y_{1,k}), \dots, \beta(e_n, y_{1,k}))'$  to be unnormalized information states such that

$$(4.1) \quad \alpha_k(e_i, y_{1,k}) = \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \bar{\Lambda}_k \exp \left( \sum_{l=1}^k q(X_{l-1}) \right) \langle X_k, e_i \rangle | y_{1,k} \right],$$

$$(4.2) \quad \beta_k(e_i, y_{1,k}) = \mathbf{E}^{\tilde{\mathbf{P}}_k} [\bar{\Lambda}_k \langle X_k, e_i \rangle | y_{1,k}].$$

*Remark 4.1.* The above definition defines the information states  $\alpha_k$  and  $\beta_k$  as  $\mathcal{Y}_k$  measurable random variables. Indeed the identities (4.1) and (4.2) are equivalent to

$$(4.3) \quad \alpha_k(e_i) = \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \bar{\Lambda}_k \exp \left( \sum_{l=1}^k q(X_{l-1}) \right) \langle X_k, e_i \rangle | \mathcal{Y}_k \right],$$

$$(4.4) \quad \beta_k(e_i) = \mathbf{E}^{\tilde{\mathbf{P}}_k} [\bar{\Lambda}_k \langle X_k, e_i \rangle | \mathcal{Y}_k].$$

Using the statistical independence of  $\{Y_{k+1}\}_{k \geq 0}$  under  $\tilde{\mathbf{P}}$ , the information states can be calculated recursively.

**LEMMA 4.1.** The information states  $\alpha_k(y_{1,k})$  and  $\beta_k(y_{1,k})$  obey the following recursions:

$$(4.5) \quad \alpha_k(y_{1,k}) = A D_k(y_{1,k}) \alpha_{k-1}(y_{1,k}),$$

$$(4.6) \quad \beta_k(y_{1,k}) = A \bar{D}_k(y_{1,k}) \beta_{k-1}(y_{1,k}),$$

where  $\alpha_0(y_{1,k}) = \beta_0(y_{1,k}) = p$ , and

$$D_k(y_{1,k}) = \text{diag} \left( \prod_{j=1}^M (Mc_{j1})^{y_k^j} \exp(d_1), \dots, \prod_{j=1}^M (Mc_{jN})^{y_k^j} \exp(d_N) \right),$$

$$\overline{D}_k(y_{1,k}) = \text{diag} \left( \prod_{j=1}^M (Mc_{j1})^{y_k^j}, \dots, \prod_{j=1}^M (Mc_{jN})^{y_k^j} \right),$$

where  $y_k^j$  denotes the  $j$ th component of  $y_k$ .

The next lemma shows that the information states are positive for a feasible observation sequence.

LEMMA 4.2. *The information states  $\alpha_k$  and  $\beta_k$  satisfy the conditions*

$$(4.7) \quad \alpha_k(e_i, y_{1,k}) > 0, \quad \beta_k(e_i, y_{1,k}) > 0$$

for any  $e_i \in \mathbb{E}_X$ .

Let

$$J_k(\xi, \tau, y_{1,k}) = \frac{\mathbf{E}^{\tilde{\mathbf{P}}_k}[\overline{\Lambda}_k \exp(\tau^{-1} \Psi_k(X_k, \xi) + \sum_{l=1}^k q(X_{l-1})) | y_{1,k}]}{\mathbf{E}^{\tilde{\mathbf{P}}_k}[\overline{\Lambda}_k | y_{1,k}]}.$$

Then we can express  $J_k(\xi, \tau, y_{1,k})$  in terms of the information states  $\alpha_k$  and  $\beta_k$ .

THEOREM 4.1. *The cost functional  $V_k(\tau, \xi, y_{1,k})$ ,  $\tau > 0$ , given by (3.9) can be written as*

$$V_k(\tau, \xi, y_{1,k}) = \tau(\log J_k(\xi, \tau, y_{1,k}) + d),$$

where

$$(4.8) \quad J_k(\xi, \tau, y_{1,k}) = \frac{H_k \alpha_k(y_{1,k})}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})} = \frac{H_k A D_k(y_{1,k}) \alpha_{k-1}(y_{1,k})}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})},$$

$$H_k = \left[ \exp\left(\frac{1}{\tau}(e_1 - \xi)' Q_k(e_1 - \xi)\right), \dots, \exp\left(\frac{1}{\tau}(e_N - \xi)' Q_k(e_N - \xi)\right) \right].$$

Remark 4.2. A special case in which the underlying HMM is not subject to uncertainty can be recovered by letting  $d \rightarrow 0$ . In this case,

$$\lim_{\tau \rightarrow \infty} \tau \log J_k(\xi, \tau, y_{1,k}) = R_k \beta_k(y_{1,k}) / \sum_{i=1}^N \beta_k(e_i, y_{1,k}),$$

where  $R_k = [(e_1 - \xi)' Q_k(e_1 - \xi), \dots, (e_N - \xi)' Q_k(e_N - \xi)]$ . That is, the limit of the robust state estimate is the HMM risk-neutral recursive filter; see [8].

**5. Characterization of the robust state estimator.** The aim in this section is to derive a mathematical characterization of the solution of the robust state estimation problem. From Theorems 3.1 and 4.1, the cost functional  $V_k(\tau, \xi, y_{1,k})$  can be written as

$$(5.1) \quad V_k(\tau, \xi, y_{1,k}) = \begin{cases} \bar{b}_k(\xi), & \tau = 0; \\ \tau(\log J_k(\xi, \tau, y_{1,k}) + d), & \tau \in (0, \infty), \end{cases}$$

where  $\bar{b}_k(\xi) = \max_{x \in \mathbb{E}_X} \Psi_k(x, \xi)$  and  $J_k(\xi, \tau, y_{1,k})$  is given by (4.8).

DEFINITION 5.1. We say  $\tau_k^* \in (0, \infty)$  is a stationary point of  $V_k(\tau, \xi, y_{1,k})$  if

$$\left. \frac{d(\tau(\log J_k(\xi, \tau, y_{1,k}) + d))}{d\tau} \right|_{\tau=\tau_k^*} = 0.$$

A stationary point, if it exists, is a function of  $\xi$  and  $y_{1,k}$  and can be found from the above equation in which

$$\begin{aligned} \frac{d(\tau(\log J_k(\xi, \tau, y_{1,k}) + d))}{d\tau} &= \log \left( \left( \alpha_k(\xi, y_{1,k}) + \sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau} \right) \right. \right. \\ &\quad \times \alpha_k(e_i, y_{1,k}) \Big) / \sum_{i=1}^N \beta_k(e_i, y_{1,k}) \Big) + d \\ &\quad - \frac{1}{\tau} \frac{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau} \right) \alpha_k(e_i, y_{1,k}) b_i}{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau} \right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k})}, \end{aligned} \quad (5.2)$$

where for  $\xi = e_j$ ,  $\mathcal{S} = \{1, \dots, j-1, j+1, \dots, N\}$  and  $b_i = \Psi_k(e_i, \xi)$ ,  $i \in \mathcal{S}$ , is a function of  $\xi$  and  $k$ . Clearly,  $\mathcal{S}$  is also a function of  $\xi$ . An explicit expression for a stationary point in terms of the information states  $\alpha_k$  and  $\beta_k$  and  $b_i$ ,  $d$  does not seem to be readily obtainable from the above equation. The stationary point over  $(0, \infty)$  can, however, be found numerically if it exists.

**5.1. The existence of a stationary point.** We first establish the continuity of  $V_k(\tau, \xi, y_{1,k})$  and its first derivative and the strict convexity of  $V_k(\tau, \xi, y_{1,k})$  with respect to  $\tau$ .

LEMMA 5.1. As functions of  $\tau$ ,  $V_k(\tau, \xi, y_{1,k})$  and  $\frac{dV_k(\tau, \xi, y_{1,k})}{d\tau}$  are continuous with respect to  $\tau$  over  $[0, \infty)$  for each  $\xi$ . Also  $V_k(\tau, \xi, y_{1,k})$  is a strictly convex function of  $\tau$  on  $(0, \infty)$  for each  $\xi$ .

We now present a necessary and sufficient condition for the existence of a stationary point. This condition is expressed in terms of the value of the first derivative of  $V_k(\tau, \xi, y_{1,k})$  with respect to  $\tau$  at  $\tau = 0$ .

THEOREM 5.1. For a given  $\xi$ , a stationary point exists on  $(0, \infty)$  if and only if

$$d < -\log \frac{\sum_{i \in \mathcal{W}} \alpha_k(e_i, y_{1,k})}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})}, \quad (5.3)$$

where  $\mathcal{W}$  denotes the set of all  $i \in \mathcal{S}$  satisfying  $(e_i - \xi)' Q_k(e_i - \xi) = \bar{b}_k(\xi)$  for this given  $\xi \in \mathbb{E}_X$ .

*Proof.* We first prove the necessity part of the theorem. By Lemma 5.1, since  $\frac{d^2 V_k(\tau, \xi, y_{1,k})}{d\tau^2} > 0$  for  $\tau > 0$ , and  $\frac{dV_k(\tau, \xi, y_{1,k})}{d\tau}$  is continuous with respect to  $\tau$  over  $[0, \infty)$ , we have that  $\frac{dV_k(\tau, \xi, y_{1,k})}{d\tau}$  is strictly monotonically increasing with  $\tau$  increasing. Also, in this part of the theorem, the stationary point is assumed to exist on  $(0, \infty)$ . This then implies that

$$\left. \frac{dV_k(\tau, \xi, y_{1,k})}{d\tau} \right|_{\tau=0^+} = d + \log \frac{\sum_{i \in \mathcal{W}} \alpha_k(e_i, y_{1,k})}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})} < 0. \quad (5.4)$$

The necessity part of the theorem now follows directly from (5.4).

Conversely, from (5.2), we have

$$(5.5) \quad \lim_{\tau \rightarrow \infty} \frac{d\tau(\log J_k(\xi, \tau, y_{1,k}) + d)}{d\tau} = \log \frac{\sum_{i=1}^N \alpha_k(e_i, y_{1,k})}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})} + d > 0$$

since  $\alpha_k(e_i, y_{1,k}) \geq \beta_k(e_i, y_{1,k})$  and  $d > 0$ . Hence, the sufficiency part of this theorem follows from (5.3), the equality in (5.4), and the continuity of  $\frac{dV_k(\tau, \xi, y_{1,k})}{d\tau}$  with respect to  $\tau$ .  $\square$

We now consider the special case of the conditional relative entropy constraint (2.4) in which  $d_r = 0$  for  $r = 1, \dots, N$ .

**COROLLARY 5.1.** *For a given  $\xi$ , if  $d_r = 0$  for  $r = 1, \dots, N$ , then a stationary point of  $V_k(\tau, \xi, y_{1,k})$  exists if and only if*

$$(5.6) \quad d < -\log \sum_{i \in \mathcal{W}} \mathbf{P}(X_k = e_i | y_{1,k}).$$

## 5.2. The maximum robust state estimator horizon. Let

$$(5.7) \quad d_c = -\log \frac{\sum_{i \in \mathcal{W}} \alpha_k(e_i, y_{1,k})}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})}.$$

Note that  $d_c$  is a function of  $k, \xi$  and is a  $\mathcal{Y}_k$  measurable random variable. The next theorem shows the behavior of  $d_c$  as  $k \rightarrow \infty$  in the case where at least one of the design parameters  $d_r$  in the definition of feasible uncertain probabilities is positive and all entries of the state transition matrix  $A$  are greater than zero.

**THEOREM 5.2.** *Assume all entries of the state transition matrix  $A$  are greater than zero. Then the following conclusions hold:*

(i) *If not all of  $d_r, r = 1, \dots, N$ , equal zero, then for any  $\xi$ ,  $d_c$  tends to  $-\infty$  almost surely with respect to the reference probability measure  $\mathbf{P}$  as  $k$  approaches  $\infty$ .*

(ii) *If  $d_r > 0$  for  $r = 1, \dots, N$ , then for any  $\xi$ , there exists a time step  $K$  such that for  $k \geq K$ , inequality (5.3) does not hold.*

*Proof.* We first prove the first part of this theorem. It follows from (4.1), (4.2), and the conditional Bayes theorem that

$$d_c = -\log \sum_{i \in \mathcal{W}} \frac{\alpha_k(e_i, y_{1,k})}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})} = -\log \mathbf{E}^{\mathbf{P}^k} \left[ \exp \left( \sum_{l=1}^k q(X_{l-1}) \right) \sum_{i \in \mathcal{W}} \langle X_k, e_i \rangle | y_{1,k} \right].$$



Furthermore, since  $\exp(\sum_{l=1}^k q(X_{l-1}))$  is  $\mathcal{G}_{k-1}$ -measurable and given that  $\mathcal{G}_{k-1}$ ,  $Y_k$ , and  $X_k$  are conditionally independent, using part (iv) of Theorem 7.3.1 in [2], we have

$$\begin{aligned}
 d_c &= -\log \mathbf{E}^{\mathbf{P}^k} \left[ \mathbf{E}^{\mathbf{P}^k} \left[ \exp \left( \sum_{l=1}^k q(X_{l-1}) \right) \sum_{i \in \mathcal{W}} \langle X_k, e_i \rangle | \mathcal{G}_{k-1}, Y_k \right] | y_{1,k} \right] \\
 &= -\log \mathbf{E}^{\mathbf{P}^k} \left[ \exp \left( \sum_{l=1}^k q(X_{l-1}) \right) \mathbf{E}^{\mathbf{P}^k} \left[ \sum_{i \in \mathcal{W}} \langle X_k, e_i \rangle | \mathcal{G}_{k-1} \right] | y_{1,k} \right] \\
 &= -\log \mathbf{E}^{\mathbf{P}^k} \left[ \exp \left( \sum_{l=1}^k q(X_{l-1}) \right) \sum_{i \in \mathcal{W}} A_i X_{k-1} | y_{1,k} \right] \\
 &\leq -\mathbf{E}^{\mathbf{P}^k} \left[ \log \left( \exp \left( \sum_{l=1}^k q(X_{l-1}) \right) \sum_{i \in \mathcal{W}} A_i X_{k-1} \right) | y_{1,k} \right]
 \end{aligned}$$

(by the convexity of  $-\log$  and Jensen's inequality; cf. Theorem 9.1.4 in [2])

$$(5.8) \quad = -\mathbf{E}^{\mathbf{P}^k} \left[ \sum_{l=1}^k q(X_{l-1}) | y_{1,k} \right] - \mathbf{E}^{\mathbf{P}^k} \left[ \log \sum_{i \in \mathcal{W}} A_i X_{k-1} | y_{1,k} \right],$$

where  $A_i$  is the  $i$ th row of the state transition matrix  $A$ . It follows from the assumption in this theorem that the state transition matrix  $A$  is irreducible and all states are positive recurrent. Then from the ergodic theorem in [13] and (3.6), we have that

$$\frac{1}{k} \sum_{l=1}^k q(X_{l-1}) = \sum_{r=1}^N d_r \left( \frac{1}{k} \sum_{l=1}^k \langle X_{l-1}, e_r \rangle \right) \rightarrow \sum_{r=1}^N d_r \pi_r \text{ as } k \rightarrow \infty, \text{ } \mathbf{P}\text{-a.s.},$$

where  $\pi = [\pi_1, \dots, \pi_N]'$  is the stationary distribution of  $A$  and  $\pi_r > 0$  for  $r = 1, \dots, N$ . Furthermore, from Corollary 11.4 in [2] or Theorem 9.4.8 in [3], the following equality holds (almost surely under  $\mathbf{P}$ ):

$$\lim_{k \rightarrow \infty} \frac{1}{k} \mathbf{E}^{\mathbf{P}^k} \left[ \sum_{l=1}^k q(X_{l-1}) | y_{1,k} \right] = \mathbf{E}^{\mathbf{P}} \left[ \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{l=1}^k q(X_{l-1}) | y_{1,\infty} \right] = \sum_{r=1}^N d_r \pi_r$$

since  $\frac{1}{k} \sum_{l=1}^k q(X_{l-1}) \leq \sum_{r=1}^N d_r$  and  $\sum_{r=1}^N d_r$  is a nonrandom constant. Hence, for feasible observation sequences,

$$(5.9) \quad \lim_{k \rightarrow \infty} \mathbf{E}^{\mathbf{P}^k} \left[ \sum_{l=1}^k q(X_{l-1}) | y_{1,k} \right] \rightarrow \infty \text{ as } k \rightarrow \infty$$

holds almost surely. Let  $\underline{a}$  be the minimum entry of  $A$ . Since we assume  $a_{ji} > 0$  for all  $i, j = 1, \dots, N$ , we have  $1 > \underline{a} > 0$ . Then the second term of (5.8) satisfies

$$(5.10) \quad -\mathbf{E}^{\mathbf{P}^k} \left[ \log \sum_{i \in \mathcal{W}} A_i X_{k-1} | y_{1,k} \right] \leq -\log \underline{a}.$$

This implies the second term of (5.8) is bounded. Hence, as time step  $k$  increases, it follows from (5.8)–(5.10) that  $d_c$  tends to  $-\infty$ . This completes the proof of the first part of the theorem.

We now prove the second part of this theorem. Since

$$(5.11) \quad \sum_{l=1}^k q(X_{l-1}) = \sum_{l=1}^k \sum_{r=1}^N d_r \langle X_{l-1}, e_r \rangle \geq \sum_{l=1}^k \underline{d}_r \sum_{r=1}^N \langle X_{l-1}, e_r \rangle = k\underline{d},$$

where  $\underline{d} = \min\{d_r, r = 1, \dots, N\} > 0$ , we have, for each  $\xi$ ,

$$d_c \leq -\mathbf{E}^{\mathbf{P}^k} \left[ \sum_{l=1}^k q(X_{l-1}) | y_{1,k} \right] - \mathbf{E}^{\mathbf{P}^k} \left[ \log \sum_{i \in \mathcal{W}} A_i X_{k-1} | y_{1,k} \right] \leq -k\underline{d} + \log \underline{a}^{-1},$$

where  $\log \underline{a}^{-1} > 0$  is finite. Hence it follows that there exists a finite time step at which  $d_c$  will be less than  $d$ . From this, the second part of the theorem follows.  $\square$

By Theorem 5.2, if  $d_r > 0$  for  $r = 1, \dots, N$  and  $a_{ji} > 0$  for  $i, j = 1, \dots, N$ , then there exists a  $K$  such that if  $k \geq K$ , then inequality (5.3) fails for any  $\xi$ . For a given  $\xi$  and a feasible observation path  $y_{1,k}$ , for simplicity, let us assume that inequality (5.3) holds for all  $k$  such that  $0 < k < K$  and fails for all  $k \geq K$ . Then we have  $d_c > d$  and the cost functional  $V_k(\tau, \xi, y_{1,k})$  has a stationary point  $\tau_k^* \in (0, \infty)$ . This implies that

$$(5.12) \quad V_k(\tau_k^*, \xi, y_{1,k}) = \min_{\tau \in [0, \infty)} V_k(\tau, \xi, y_{1,k}) = \min_{\tau \in (0, \infty)} V_k(\tau, \xi, y_{1,k}) < V_k(0, \xi, y_{1,k}),$$

where  $V_k(0, \xi, y_{1,k}) = \max_{x \in \mathbb{E}_X} \Psi_k(x, \xi) = \bar{b}_k(\xi)$  is a constant. On the contrary, for  $k \geq K$ , we have  $d \geq d_c$ . Then  $V_k(\tau, \xi, y_{1,k}) \geq V_k(0, \xi, y_{1,k})$  and

$$\min_{\tau \in [0, \infty)} V_k(\tau, \xi, y_{1,k}) = \min \left\{ V_k(0, \xi, y_{1,k}), \min_{\tau \in (0, \infty)} V_k(\tau, \xi, y_{1,k}) \right\} = V_k(0, \xi, y_{1,k}).$$

This implies that for this case,  $\tau_k^* = 0$ .

Since  $V_k(0, \xi, y_{1,k})$  is independent of the measurements, then the state estimate obtained from taking  $\min_{\xi \in \mathbb{E}_X} V_k(0, \xi, y_{1,k})$  is also independent of the measurements. In order to obtain an estimate which depends on the measurements, it is necessary to exclude the case  $\tau_k^* = 0$ . This is done by defining the robust state estimator under consideration as a finite horizon estimator.

**DEFINITION 5.2.** *For a feasible observation path  $y_{1,k}$ , we say the robust state estimator under consideration is a finite horizon state estimator if there exists a finite time step such that at this time step, inequality (5.3) does not hold for some  $\xi$ . The maximum of time steps at which inequality (5.3) holds for any  $\xi$ ,  $K$  is referred to as the maximum horizon of the robust state estimator under consideration and is characterized by inequality (5.3).*

Note that in Definition 5.2, we do not make any additional assumptions on the design parameters  $d_r$  and the state transition matrix  $A$ . That is, as long as inequality (5.3) does not hold for some finite time step  $k$ , we define the robust state estimator as a finite horizon estimator.

When the parameters  $d_r$ ,  $r = 1, \dots, N$ , all equal zero, it is possible that we can find a number  $\bar{d}$  such that if  $d < \bar{d}$ , then the robust state estimator can be applied on an infinite horizon. Note that here an infinite horizon implies that the time step can take any finite value.

LEMMA 5.2. *Consider the uncertainty in the initial distribution. Let  $d_r = 0$ ,  $r = 1, \dots, N$ . For  $j = 1, \dots, N$ , let  $\bar{a}_j$  denote the maximum of the  $j$ th row of the state transition matrix  $A$ . Let  $\bar{a}$  be  $\max_{i=1, \dots, N} \{\sum_{j=1}^N \bar{a}_j - \bar{a}_i\}$  and  $\bar{d} = -\log \bar{a}$ . If  $\bar{a} < 1$  and  $d < \bar{d}$ , then inequality (5.6) holds for any time step  $k$ .*

Note that the condition  $\bar{a} < 1$  of Lemma 5.2 implies that  $a_{i,j} > 0$  for all  $i, j = 1, \dots, N$ . From (3.15) in Theorem 3.2, for a feasible observation path  $y_{1,k}$  and  $0 < k \leq K$ , let  $x_k$  denote the robust state estimate at time step  $k$ . Then for any  $\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \Xi_k(y_{1,k})$ , we have

$$\begin{aligned}
 \min_{\xi \in \mathbb{E}_X} \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi)|y_{1,k}] &\leq \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, x_k)|y_{1,k}] \\
 &\leq \sup_{\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \Xi_k(y_{1,k})} \mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, x_k)|y_{1,k}] \\
 (5.13) \qquad &= \min_{\tau \in [0, \infty)} V_k(\tau, x_k, y_{1,k}) = \min_{\xi \in \mathbb{E}_X} \min_{\tau \in [0, \infty)} V_k(\tau, \xi, y_{1,k}).
 \end{aligned}$$

This means that for a feasible observation path  $y_{1,k}$ , at each time step  $k$ , the estimation error with the optimal state estimate or robust state estimate under  $\bar{\mathbf{P}}_k(\cdot|y_{1,k})$  does not exceed the cost value of the robust state estimator.

**5.3. Monotonicity results and the dynamic behavior of the optimal parameter and minimum cost value.** Let  $K$  be the maximum horizon of the robust state estimator defined in Definition 5.2. For a fixed  $k$ ,  $k \leq K$ , the following lemma gives monotonicity results for the optimal parameter  $\tau_k^*$  and the minimum cost value as  $d$  is varied. It also gives a monotonicity result for the maximum horizon as one of the design parameters is varied. As before let  $\bar{b}_k(\xi) = \max_{x \in \mathbb{E}_X} \Psi_k(x, \xi)$  and  $b_i(\xi) = \Psi_k(e_i, \xi)$  for  $i \in \mathcal{S}$ , where  $\mathcal{S} = \{1, \dots, j-1, j+1, \dots, N\}$  whenever  $\xi = e_j$ . Note that  $b_i(\xi)$  is a function of  $k$ . Sometimes for simplicity we will omit the variable  $\xi$ .

LEMMA 5.3. *For a given  $\xi$  and a feasible observation path  $y_{1,k}$ , at time step  $k \leq K$ ,  $\tau_k^*$  is strictly decreasing as  $d$  increases. Also, the minimum cost value is strictly monotone increasing with  $\tau_k^*$  decreasing and with  $d$  increasing, respectively. In addition, the maximum horizon  $K$  is decreasing as any of the design parameters increases.*

Next, we will consider the dynamic behavior of the optimal parameter  $\tau_k^*$  and the minimum cost value as  $k$  increases based on an upper bound of the optimal parameter  $\tau_k^*$ . For simplicity, let  $Q_k = Q > 0$ , where  $Q$  is a finite positive definite matrix. Then for each  $\xi$ ,  $\bar{b}_k(\xi)$  and  $b_i(\xi)$  are constant. We use  $\bar{b}(\xi)$  to denote  $\bar{b}_k(\xi)$ .

LEMMA 5.4. *Assume all of the design parameters  $d_r$  are greater than zero. Then for a given feasible observation sequence  $y_{1,k}$ , the optimal parameter  $\tau_k^*$  approaches zero as  $k$  increases for each  $\xi$ .*

*Proof.* We first give an upper bound on the optimal parameter  $\tau_k^*$ . Let

$$d_{y_{1,k}} \triangleq \log \frac{\sum_{i=1}^N \alpha_k(e_i, y_{1,k})}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})}.$$

By its definition and (5.2),  $\tau_k^* \in (0, \infty)$  satisfies the following equation:

$$(5.14) \quad d = -\log \left( \left( \alpha_k(\xi, y_{1,k}) + \sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) \right) / \sum_{i=1}^N \beta_k(e_i, y_{1,k}) \right) \\ + \frac{1}{\tau_k^*} \frac{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) b_i}{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k})}.$$

Then since  $\exp(\frac{b_i}{\tau_k^*}) > 1$  and  $\log$  is a strictly monotone increasing function, it follows from (5.14) that

$$\begin{aligned} & \frac{1}{\tau_k^*} \frac{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) b_i}{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k})} \\ &= d + \log \left( \left( \alpha_k(\xi, y_{1,k}) + \sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) \right) / \sum_{i=1}^N \beta_k(e_i, y_{1,k}) \right) \\ &> d + \log \left( \left( \alpha_k(\xi, y_{1,k}) + \sum_{i \in \mathcal{S}} \alpha_k(e_i, y_{1,k}) \right) / \sum_{i=1}^N \beta_k(e_i, y_{1,k}) \right) \\ &= d + d_{y_{1,k}}. \end{aligned}$$

Note that  $d_{y_{1,k}} \geq 0$  since  $\alpha_k(e_i, y_{1,k}) \geq \beta_k(e_i, y_{1,k})$ . Hence,

$$(5.15) \quad \begin{aligned} \tau_k^* &< \frac{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) b_i}{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k})} \frac{1}{d + d_{y_{1,k}}} \\ &< \frac{\bar{b}(\xi) \sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k})}{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k})} \frac{1}{d + d_{y_{1,k}}} < \frac{\bar{b}(\xi)}{d + d_{y_{1,k}}}. \end{aligned}$$

Furthermore, it follows from (5.11) that

$$(5.16) \quad d_{y_{1,k}} = \log \mathbf{E}^{\mathbf{P}_k} \left[ \exp \left( \sum_{l=1}^k q(X_{l-1}) \right) | y_{1,k} \right] \geq \mathbf{E}^{\mathbf{P}_k} \left[ \sum_{l=1}^k q(X_{l-1}) | y_{1,k} \right] \geq k \underline{d},$$

where  $\underline{d} = \min\{d_r, r = 1, \dots, N\} > 0$ . Hence, it follows from (5.15) and (5.16) that

$$\tau_k^* < \frac{\bar{b}(\xi)}{d + k \underline{d}}.$$

This implies that  $\tau_k^*$  approaches zero as  $k$  increases. This completes the proof of the lemma.  $\square$

**THEOREM 5.3.** *Assume all of the design parameters  $d_r$  are greater than zero. Then for a given feasible observation sequence  $y_{1,k}$ , the minimum cost value approaches  $\min_{\xi \in \mathbb{E}_X} \bar{b}(\xi)$  as  $k$  increases. In particular, for each  $\xi$ ,  $\min_{\tau \in (0, \infty)} V_k(\tau, \xi, y_{1,k})$  approaches  $\bar{b}(\xi)$  as  $k$  increases.*

*Proof.* By (5.14), the stationary point  $\tau_k^*$  must satisfy

$$(5.17) \quad \tau_k^* \left( \log \left( \alpha_k(\xi, y_{1,k}) + \sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) / \sum_{i=1}^N \beta_k(e_i, y_{1,k}) \right) + d \right) = \frac{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) b_i}{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k})}.$$

The left-hand side of (5.17) is the minimum of the cost functional; see (4.8). Then

$$(5.18) \quad \min_{\tau \in (0, \infty)} \tau (\log J_k(\xi, \tau, y_{1,k}) + d) = \frac{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) b_i}{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k})}.$$

Dividing the numerator and denominator of the right-hand side of (5.18) by  $\sum_{i=1}^N \alpha_k(e_i, y_{1,k}) = \alpha_k(y_{1,k}) > 0$  and  $\exp(\frac{\bar{b}}{\tau_k^*})$ , simultaneously, we have

$$(5.19) \quad \begin{aligned} & \min_{\tau \in (0, \infty)} \tau (\log J_k(\xi, \tau, y_{1,k}) + d) \\ &= \frac{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) b_i}{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i}{\tau_k^*} \right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k})} \\ &= \frac{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i - \bar{b}}{\tau_k^*} \right) \frac{\alpha_k(e_i, y_{1,k})}{\alpha_k(y_{1,k})} b_i}{\sum_{i \in \mathcal{S}} \exp \left( \frac{b_i - \bar{b}}{\tau_k^*} \right) \frac{\alpha_k(e_i, y_{1,k})}{\alpha_k(y_{1,k})} + \exp \left( \frac{-\bar{b}}{\tau_k^*} \right) \frac{\alpha_k(\xi, y_{1,k})}{\alpha_k(y_{1,k})}} \\ &= \frac{\sum_{i \in \mathcal{S} - \mathcal{W}} \exp \left( \frac{b_i - \bar{b}}{\tau_k^*} \right) \frac{\alpha_k(e_i, y_{1,k})}{\alpha_k(y_{1,k})} b_i + \bar{b} \sum_{i \in \mathcal{W}} \frac{\alpha_k(e_i, y_{1,k})}{\alpha_k(y_{1,k})}}{\sum_{i \in \mathcal{S} - \mathcal{W}} \exp \left( \frac{b_i - \bar{b}}{\tau_k^*} \right) \frac{\alpha_k(e_i, y_{1,k})}{\alpha_k(y_{1,k})} + \sum_{i \in \mathcal{W}} \frac{\alpha_k(e_i, y_{1,k})}{\alpha_k(y_{1,k})} + \exp \left( \frac{-\bar{b}}{\tau_k^*} \right) \frac{\alpha_k(\xi, y_{1,k})}{\alpha_k(y_{1,k})}}, \end{aligned}$$

where  $\mathcal{W} \subseteq \mathcal{S}$  denotes the set of all  $i \in \mathcal{S}$  satisfying  $b_i(\xi) = \bar{b}(\xi)$ . It follows from  $\frac{\alpha_k(e_i, y_{1,k})}{\alpha_k(y_{1,k})} < 1$  that

$$\exp \left( \frac{b_i - \bar{b}}{\tau_k^*} \right) \frac{\alpha_k(e_i, y_{1,k})}{\alpha_k(y_{1,k})} < \exp \left( \frac{b_i - \bar{b}}{\tau_k^*} \right) \text{ for } i \in \mathcal{S} - \mathcal{W}$$

and

$$\exp\left(\frac{-\bar{b}}{\tau_k^*}\right) \frac{\alpha_k(\xi, y_{1,k})}{\alpha_k(y_{1,k})} < \exp\left(\frac{-\bar{b}}{\tau_k^*}\right).$$

Hence from Lemma 5.4, the following statements hold:

1.  $\exp\left(\frac{b_i - \bar{b}}{\tau_k^*}\right) \frac{\alpha_k(e_i, y_{1,k})}{\alpha_k(y_{1,k})}$  approaches zero as  $k$  increases for  $i \in \mathcal{S} - \mathcal{W}$ .
2.  $\exp\left(\frac{-\bar{b}}{\tau_k^*}\right) \frac{\alpha_k(\xi, y_{1,k})}{\alpha_k(y_{1,k})}$  approaches zero as  $k$  increases.

Here  $-\bar{b} < 0$ ,  $b_i - \bar{b} < 0$  for  $i \in \mathcal{S} - \mathcal{W}$ . Since for each  $\xi$ ,  $\min_{\tau \in (0, \infty)} \tau(\log J_k(\xi, \tau, y_{1,k}) + d)$  is a continuous function of  $\tau_k^*$  and  $\xi$  belongs to a finite value space, the theorem follows from (5.19) and statements 1 and 2 as above.  $\square$

**5.4. Connection with a maximum information state estimator.** In the next theorem, we will establish a connection between the robust state estimator presented in this paper and a maximum a posteriori (MAP) probability or a conditional expectation estimator. We also point out that the MAP probability estimator (cf. section 2.B in [16]) is a special case of the robust state estimator presented in this paper as  $d \rightarrow 0$  and  $d_r = 0$  for  $r = 1, \dots, N$ .

**THEOREM 5.4.** *Let  $K$  be the maximum horizon of the robust state estimator for a feasible observation sequence  $y_{1,k}$ . At time step  $k$  with  $0 < k \leq K$ , if for any  $\xi \in \mathbb{E}_X$ , for all  $i \in \mathcal{S}$ ,  $b_i(\xi)$  is a constant, then the robust state estimate under consideration equals the maximum information state estimate. Specifically, the robust state estimate equals the MAP probability or conditional expectation estimate under the reference probability measure  $\mathbf{P}$ .*

*Proof.* For simplicity, we prove this theorem only for the case  $N = 3$ ; i.e.,  $\mathbb{E}_X = \{e_1, e_2, e_3\}$ . Then it follows from  $0 < k \leq K$  and (4.8) that

$$(5.20) \quad \min_{\tau \in [0, \infty)} V_k(\tau, \xi, y_{1,k}) = \begin{cases} \tau_k^*(e_1, y_{1,k}) \left( \log \left( \bar{\alpha}_k(e_1, y_{1,k}) + \exp\left(\frac{b_2(e_1)}{\tau_k^*(e_1, y_{1,k})}\right) \bar{\alpha}_k(e_2, y_{1,k}) \right. \right. \\ \quad \left. \left. + \exp\left(\frac{b_3(e_1)}{\tau_k^*(e_1, y_{1,k})}\right) \bar{\alpha}_k(e_3, y_{1,k}) \right) + d \right), & \xi = e_1; \\ \tau_k^*(e_2, y_{1,k}) \left( \log \left( \exp\left(\frac{b_1(e_2)}{\tau_k^*(e_2, y_{1,k})}\right) \bar{\alpha}_k(e_1, y_{1,k}) + \bar{\alpha}_k(e_2, y_{1,k}) \right. \right. \\ \quad \left. \left. + \exp\left(\frac{b_3(e_2)}{\tau_k^*(e_2, y_{1,k})}\right) \bar{\alpha}_k(e_3, y_{1,k}) \right) + d \right), & \xi = e_2; \\ \tau_k^*(e_3, y_{1,k}) \left( \log \left( \exp\left(\frac{b_1(e_3)}{\tau_k^*(e_3, y_{1,k})}\right) \bar{\alpha}_k(e_1, y_{1,k}) + \exp\left(\frac{b_2(e_3)}{\tau_k^*(e_3, y_{1,k})}\right) \bar{\alpha}_k(e_2, y_{1,k}) \right. \right. \\ \quad \left. \left. + \bar{\alpha}_k(e_3, y_{1,k}) \right) + d \right), & \xi = e_3, \end{cases}$$

where  $\bar{\alpha}_k(e_i, y_{1,k}) = \frac{\alpha_k(e_i, y_{1,k})}{\sum_{j=1}^N \beta_k(e_j, y_{1,k})}$  and  $b_i(e_j) = (e_i - e_j)' Q_k(e_i - e_j)$  for  $i \neq j$ ,  $1 \leq i, j \leq N$ . Also note that  $b_i(e_j) = b_j(e_i)$ . Suppose  $\bar{\alpha}_k(e_3, y_{1,k}) = \max_{i=1,2,3} \bar{\alpha}_k(e_i, y_{1,k})$ . By the assumption in this theorem,  $b_2(e_3) = b_2(e_1) = b_1(e_3)$ ; i.e., for any  $\xi \in \mathbb{E}_X$ , for all  $i \in \mathcal{S}$ ,  $b_i$  is a constant. Then we claim that

$$(5.21) \quad V_k(\tau_k^*, e_3, y_{1,k}) = \min_{i=1,2,3} V_k(\tau_k^*, e_i, y_{1,k}).$$

That is,  $e_3$  is the robust state estimate at time step  $k$ . We next prove this claim. It follows from (5.20) that

$$\begin{aligned}
 V_k(\tau_k^*, e_3, y_{1,k}) &= \tau_k^*(e_3, y_{1,k}) \left( \log \left( \exp \left( \frac{b_1(e_3)}{\tau_k^*(e_3, y_{1,k})} \right) \bar{\alpha}_k(e_1, y_{1,k}) + \exp \left( \frac{b_2(e_3)}{\tau_k^*(e_3, y_{1,k})} \right) \right. \right. \\
 &\quad \left. \left. \times \bar{\alpha}_k(e_2, y_{1,k}) + \bar{\alpha}_k(e_3, y_{1,k}) \right) + d \right) \\
 &\leq \tau_k^*(e_1, y_{1,k}) \left( \log \left( \exp \left( \frac{b_1(e_3)}{\tau_k^*(e_1, y_{1,k})} \right) \bar{\alpha}_k(e_1, y_{1,k}) + \exp \left( \frac{b_2(e_3)}{\tau_k^*(e_1, y_{1,k})} \right) \right. \right. \\
 &\quad \left. \left. \bar{\alpha}_k(e_2, y_{1,k}) + \bar{\alpha}_k(e_3, y_{1,k}) \right) + d \right) \\
 &\quad (\text{by } V_k(\tau_k^*, e_3, y_{1,k}) \text{ is the minimum at } \tau_k^*(e_3, y_{1,k})) \\
 &\leq \tau_k^*(e_1, y_{1,k}) \left( \log \left( \bar{\alpha}_k(e_1, y_{1,k}) + \exp \left( \frac{b_2(e_1)}{\tau_k^*(e_1, y_{1,k})} \right) \bar{\alpha}_k(e_2, y_{1,k}) \right. \right. \\
 &\quad \left. \left. + \exp \left( \frac{b_3(e_1)}{\tau_k^*(e_1, y_{1,k})} \right) \bar{\alpha}_k(e_3, y_{1,k}) \right) + d \right) \\
 &= V_k(\tau_k^*, e_1, y_{1,k}).
 \end{aligned}$$

Here the last inequality follows from the facts that  $\bar{\alpha}_k(e_1, y_{1,k}) \leq \bar{\alpha}_k(e_3, y_{1,k})$ ,  $b_2(e_3) = b_2(e_1)$ , and

$$\begin{aligned}
 &\bar{\alpha}_k(e_3, y_{1,k}) + \exp \left( \frac{b_1(e_3)}{\tau_k^*(e_1, y_{1,k})} \right) \bar{\alpha}_k(e_1, y_{1,k}) \\
 &\leq \bar{\alpha}_k(e_1, y_{1,k}) + \exp \left( \frac{b_3(e_1)}{\tau_k^*(e_1, y_{1,k})} \right) \bar{\alpha}_k(e_3, y_{1,k}).
 \end{aligned}$$

Similarly, we have  $V_k(\tau_k^*, e_3, y_{1,k}) \leq V_k(\tau_k^*, e_2, y_{1,k})$ . Hence, (5.21) holds. Note that here we have assumed that  $\bar{\alpha}_k(e_3, y_{1,k})$  is the maximum among  $\{\bar{\alpha}_k(e_1, y_{1,k}), \bar{\alpha}_k(e_2, y_{1,k}), \bar{\alpha}_k(e_3, y_{1,k})\}$ . Also since

$$\bar{\alpha}_k(e_i, y_{1,k}) = \frac{\alpha_k(e_i, y_{1,k})}{\sum_{j=1}^N \beta_k(e_j, y_{1,k})}$$

and  $\sum_{i=1}^N \beta_k(e_j, y_{1,k})$  is independent of  $e_i$ , maximizing  $\bar{\alpha}_k(y_{1,k})$  is equivalent to maximizing the information state  $\alpha_k(e_i, y_{1,k})$ . Hence, the robust state estimate equals the maximum information state estimate. Furthermore, under the reference probability measure  $\mathbf{P}$ ,

(5.22)

$$\bar{\alpha}_k(e_i, y_{1,k}) = \begin{cases} \mathbf{P}(X_k = e_i | y_{1,k}), & d_r = 0, r = 1, \dots, N; \\ \mathbf{E}^{\mathbf{P}}[\exp(\sum_{l=1}^k q(X_{l-1})) \langle X_k, e_i \rangle | y_{1,k}], & \text{otherwise.} \end{cases}$$

Hence, the robust state estimate equals the MAP probability or conditional expectation estimate. The above proof can be extended in a straightforward manner to the

general case when the dimension  $N$  of  $\mathbb{E}_X$  is finite and arbitrary. This completes the proof of the theorem.  $\square$

*Remark 5.1.* If  $Q_k = \alpha I$  for  $0 < k \leq K$ , where  $\alpha > 0$  is constant, then the robust state estimator is equivalent to the maximum information state estimator (i.e., the MAP probability or conditional expectation estimator) given the measurements under the reference probability measure  $\mathbf{P}$ . Note that for  $d_r = 0$  for  $r = 1, \dots, N$ , by (5.22), the robust state estimates are independent of the parameter  $d$  and the optimal cost value is a function of  $d$ .

*Remark 5.2.* We consider an HMM with an unknown parameter set and choose the triple  $(A, C, p)$  as the reference parameter set satisfying Assumption 2.1. Let  $\mathbf{P}$  be the reference probability measure. Then under  $\mathbf{P}$ , the pair of the coordinate processes  $\{X_k, Y_{k+1}\}_{k \geq 0}$  on  $(\Omega, \mathcal{B})$  is an HMM with  $(A, C, p)$  as its parameter set. By Theorem 5.4, in the case of  $d_r = 0$  for  $r = 1, \dots, N$ , it is interesting to note that the MAP probability estimator under  $\mathbf{P}$  can be equivalent to a robust state estimator with uncertainty in the initial distribution if we can find the design parameter  $d$  such that inequality (5.6) holds. Hence, we can say the MAP estimator has a natural level of robustness. A similar argument can be used in the more general case, i.e., the case in which there exists a  $d_r > 0$ .

*Remark 5.3.* Under the same assumption as in Theorem 5.4 (that is, for any  $\xi \in \mathbb{E}_X$ ,  $b_i(\xi)$  is a constant for all  $i \in \mathcal{S}$ ), then as  $d \rightarrow 0$ , it is clear that inequality (5.6) holds for  $k > 0$ , any  $\xi$ , and feasible observation sequence  $y_{1,k}$ . This follows since  $\mathcal{W} \subseteq \mathcal{S} \subset \{1, \dots, N\}$ ,  $\mathbf{P}(X_k = \xi | y_{1,k}) > 0$ , and  $\sum_{i \in \mathcal{W}} \mathbf{P}(X_k = e_i | y_{1,k}) < 1$ . Hence the MAP probability estimator (cf. section 2.B in [16]) is a special case of the robust state estimator presented in this paper in the limit, when  $d_r = 0$  for  $r = 1, \dots, N$  and the true parameter set equals the reference one.

**6. Illustrative example.** In this section, we present an example to illustrate the main results of this paper. Consider an HMM whose true parameter set  $\bar{\zeta} = (\bar{A}, \bar{C}, \bar{p})$  is given by

$$\bar{A} = \begin{bmatrix} 0.695 & 0.3 & 0 \\ 0.105 & 0.18 & 0.71 \\ 0.2 & 0.52 & 0.29 \end{bmatrix}, \bar{C} = \begin{bmatrix} 0.49 & 0.21 & 0.49 \\ 0 & 0.51 & 0 \\ 0.51 & 0.28 & 0.51 \end{bmatrix}, \text{ and } \bar{p} = \begin{bmatrix} 0.35 \\ 0 \\ 0.65 \end{bmatrix}.$$

Although we will use this true parameter set in computer experiments to generate realized sequences of the observation process  $\{Y_{k+1}\}_{k \geq 0}$ , we suppose that the true parameter set is not known exactly and select the following reference parameter set  $\zeta = (A, C, p)$  for the design of a state estimator for this HMM:

$$A = \begin{bmatrix} 0.7 & 0.3 & 0 \\ 0.1 & 0.2 & 0.7 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}, C = \begin{bmatrix} 0.5 & 0.2 & 0.5 \\ 0 & 0.5 & 0 \\ 0.5 & 0.3 & 0.5 \end{bmatrix}, \text{ and } p = \begin{bmatrix} 0.4 \\ 0 \\ 0.6 \end{bmatrix}.$$

In this example at any finite time step  $k$ , the marginal probability measures  $\bar{\mathbf{P}}_k$  and  $\mathbf{P}_k$  corresponding to the true and reference parameter sets  $\bar{\zeta}$  and  $\zeta$  are equivalent,  $\bar{\mathbf{P}}_k \sim \mathbf{P}_k$ , since the true parameter set  $\bar{\zeta}$  is equivalent to the reference parameter set  $\zeta$  in the sense of the absolute continuity of parameter sets. Hence, for any feasible  $y_{1,k}$ , we have  $\bar{\mathbf{P}}_k(\cdot | y_{1,k}) \in \mathcal{P}(\Omega, \mathcal{G}_k, \mathbf{P}_k | y_{1,k})$ . Also under  $\mathbf{P}$ , Assumption 2.1 holds. Meanwhile, the event  $\{\omega : Y_1(\omega) = f_2\}$ , where  $f_2 = (0 \ 1 \ 0)'$ , has zero probability under both  $\mathbf{P}$  and  $\bar{\mathbf{P}}$ , and hence this event does not occur almost surely.

In every experiment only a path of the measurement process  $\{Y_{k+1}\}_{k \geq 0}$  is observed up to time step  $k$ . That is, in each experiment we obtain sequentially the



realized sequence of the observation process,  $Y_1(\omega) = y_1, \dots, Y_k(\omega) = y_k$ . This observation path defines an event which consists of all sample  $\omega \in \Omega$  compatible with the realized observation path  $y_{1,k}$ . However, the corresponding realized sequence of the state process,  $X_1(\omega) = x_1, \dots, X_k(\omega) = x_k$ , cannot be observed directly. We will use the robust state estimator presented in this paper to construct robust state estimates of this realized sequence of the state process  $\{X_{k+1}\}_{k \geq 0}$  based on the realized sequence of the observation process.

In our numerical experiments, we generated a sequence  $y_{1,k}$  using the true parameter set  $\bar{\zeta}$ . For this realized observation path, we verified numerically that the true probability measure  $\bar{\mathbf{P}}_k$  satisfied the conditional relative entropy constraint (2.4) in which the design parameters were selected as follows:

$$(6.1) \quad d = 0.006, \quad d_1 = 0.001, \quad d_2 = 0.0015, \quad d_3 = 0.0005.$$

Hence, we ensured that the true probability distribution satisfied the constrained condition  $\bar{\mathbf{P}}_k(\cdot|y_{1,k}) \in \Xi_k(y_{1,k})$ ; this fact is illustrated in Figure 6.1 in which data2 and data1 represent the conditional relative entropy and the bound on the right-hand side of (2.4), respectively.

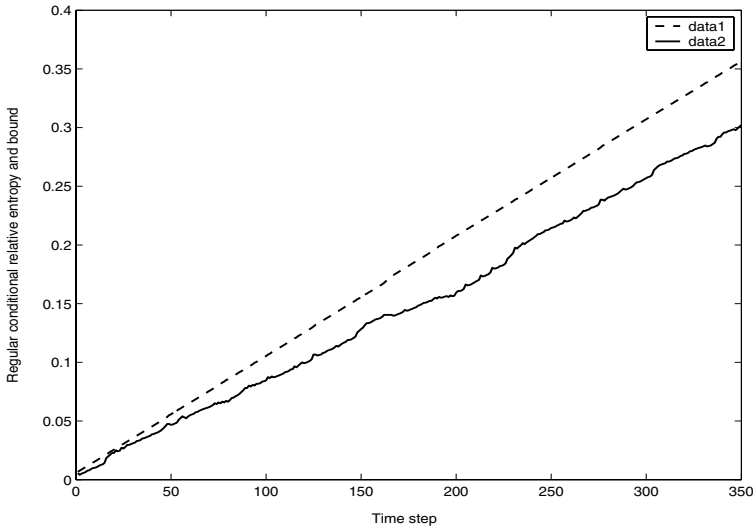


FIG. 6.1. The conditional relative entropy and the bound on the right-hand side of (2.4) vs. time step.

Next, we considered the robust state estimation problem (3.1) in which the weighting matrices in the cost functional (3.1) were chosen to be

$$Q_k = \text{diag}(0.07, 0.08, 0.07).$$

Computer simulation showed that for this given observation sequence, the maximum horizon  $K$  equaled 331. Figure 6.2 shows the cost value  $V_k(\tau, \xi, y_{1,k})$  corresponding to  $k = 17$  and  $\xi = e_1$ . The optimal parameter  $\tau_k^*$  for each  $k = 1, \dots, K$  is shown in Figure 6.3. In Figure 6.4, data1 represents the cost values  $V_k(\tau_k^*, \xi_k^*, y_{1,k})$ , where  $\xi_k^*$  and  $\tau_k^*$  are the robust state estimate and the value of  $\tau$  at which  $V_k(\tau, \xi_k^*, y_{1,k})$  attains its minimum at time step  $k$ . Also, the conditional expectation values  $\mathbf{E}^{\bar{\mathbf{P}}_k}[\Psi_k(X_k, \xi_k^*)|y_{1,k}]$

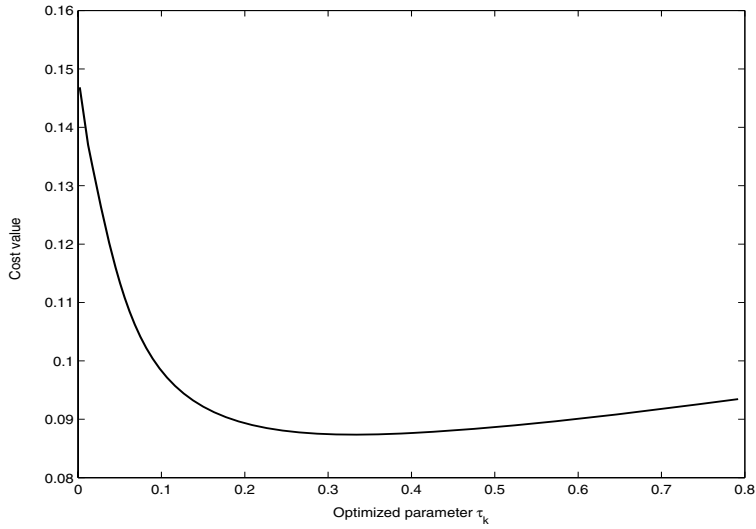


FIG. 6.2. The cost value vs. parameter  $\tau_k$  at time step  $k = 17$ .

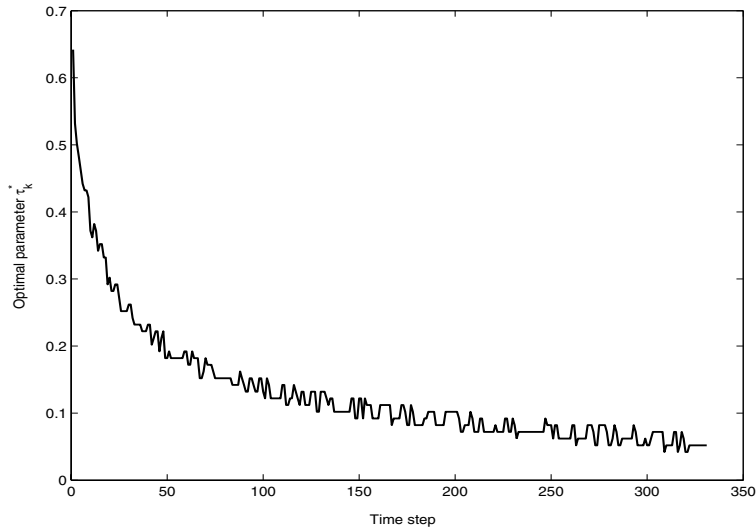


FIG. 6.3. The optimal  $\tau_k^*$  at each time step.

are given by data2, and data3 shows the cost values  $V_k(0, \xi, y_{1,k})$  corresponding to the unconstrained estimation problem. Figure 6.4 shows that with the robust state estimates  $\xi_k^*$  at each time step  $k$ , the cost values  $\mathbf{E}^{\mathbf{P}^k}[\Psi_k(X_k, \xi_k^*)|y_{1,k}]$  are less than the guaranteed cost values of the robust state estimator.

We also considered another HMM with the true parameter set  $\hat{\zeta} = (\hat{A}, \hat{C}, \hat{p})$ , where

$$\hat{A} = \begin{bmatrix} 0.695 & 0.6 & 0 \\ 0.105 & 0.18 & 0.5 \\ 0.2 & 0.22 & 0.5 \end{bmatrix},$$

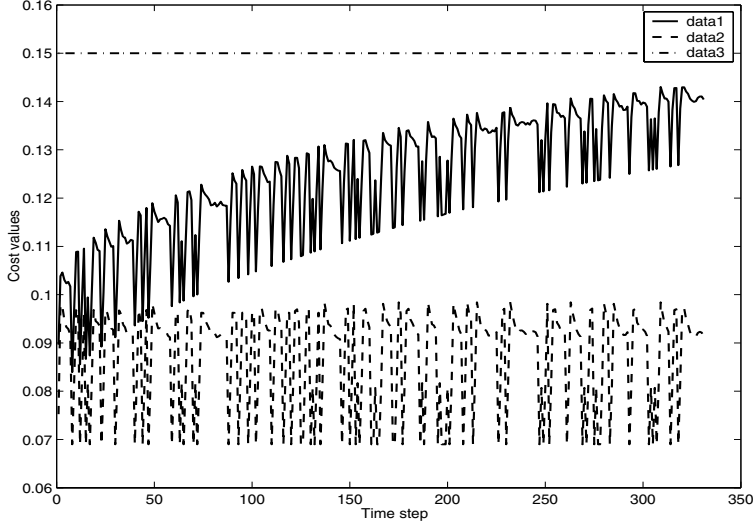


FIG. 6.4. Cost values vs time step.

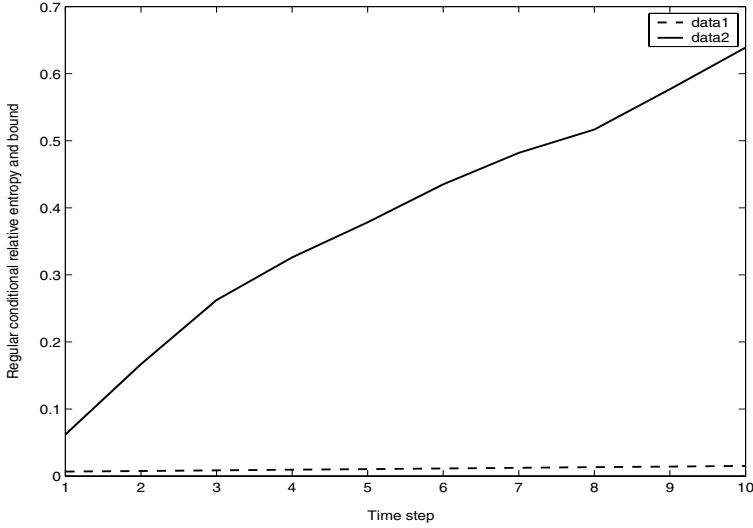


FIG. 6.5. Conditional relative entropy and bound vs. time step.

$\hat{C} = \bar{C}$ , and  $\hat{p} = \bar{p}$ ; let  $\hat{\mathbf{P}}$  denote the Kolmogorov measure of this HMM. For this HMM we also generated a realized observation sequence  $y_{1,k}$  using  $\hat{\zeta}$  and checked whether  $\hat{\mathbf{P}}_k(\cdot|y_{1,k})$  satisfied the constraint (2.4) with the same design parameters given in (6.1). As Figure 6.5 shows, the conditional relative entropy  $\mathcal{R}(\hat{\mathbf{P}}_k(\cdot|y_{1,k})\|\mathbf{P}_k(\cdot|y_{1,k}))$  does not satisfy the constraint (2.4) with the design parameters (6.1) for any  $k$ , and therefore  $\hat{\mathbf{P}}_k(\cdot|y_{1,k})(y_{1,k}) \notin \Xi_k(y_{1,k})$ . Hence even though we were using the same algorithm for computing robust state estimates, the cost values  $\mathbf{E}^{\hat{\mathbf{P}}_k}[\Psi_k(X_k, \xi_k^*)|y_{1,k}]$  were not guaranteed to be less than the predicted bounds on the cost values of the robust state estimator. In fact, the cost values  $\mathbf{E}^{\hat{\mathbf{P}}_k}[\Psi_k(X_k, \xi_k^*)|y_{1,k}]$  exceeded the predicted bounds; this can be seen in Figure 6.6.

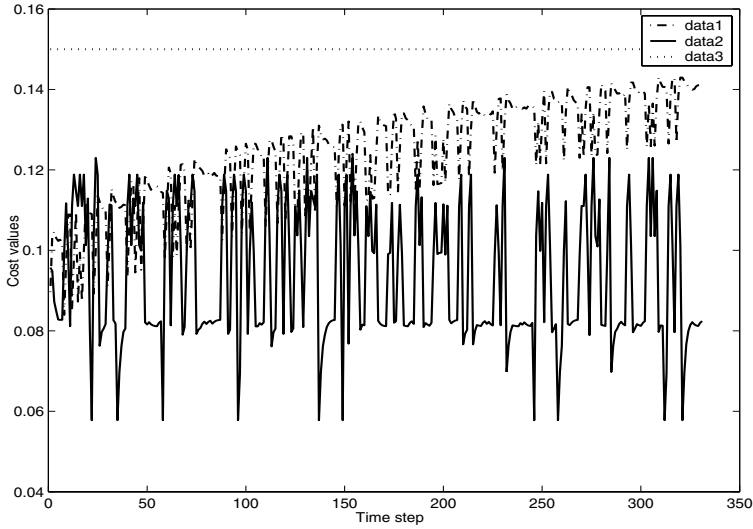


FIG. 6.6. Cost values vs. time step.

**7. Conclusion.** In this paper, a robust state estimator for uncertain HMMs has been introduced in which the uncertainty is characterized in terms of a conditional relative entropy constraint. A finite horizon robust state estimator has been derived that guarantees a certain bound on the state estimation error, provided that the true conditional probability measure satisfies the conditional relative entropy constraint. Also it is pointed out that under certain conditions, our robust state estimator is equivalent to the MAP probability or a conditional expectation estimator given the measurements under the reference probability measure.

We have presented a necessary and sufficient condition for the existence of a stationary point, under which the conditional relative entropy constraint is active. This condition allows one to determine the maximum time horizon for the estimator. A characterization of the solution to the robust state estimation problem is also presented. In many HMM applications, however, a robust state estimator is most desirable which is capable of producing estimates over an infinite interval of time, as in [20]. For this further work, Lemma 1 in [10] deserves attention.

**Appendix A. Proof of Lemma 2.1.** We now establish the sufficiency part by induction. We first consider the case  $k = 1$ . Consider a state  $x_1 \in \mathbb{E}_X$  and let  $y_1$  be an arbitrary feasible 1-step observed sequence. Since  $\mathbf{P}(Y_1 = y_1) = \sum_{x_0 \in \mathbb{E}_X} c_{y_1 x_0} p_{x_0} > 0$ , then clearly there must exist  $x_0 \in \mathbb{E}_X$  such that  $c_{y_1 x_0} p_{x_0} > 0$ , and hence the set  $D_{y_1}$  is not empty. Therefore it follows from condition 1 of the lemma that  $a_{x_1 x_0} > 0$  and thus

$$\mathbf{P}(X_1 = x_1, Y_1 = y_1) = \sum_{x_0 \in \mathbb{E}_X} a_{x_1 x_0} c_{y_1 x_0} p_{x_0} > 0.$$

That is, at  $k = 1$ , the state  $x_1$  is observable.

Furthermore, we consider an arbitrary  $k$ -step long feasible observed sequence  $\{y_1, \dots, y_k\}$ ; that is,  $\mathbf{P}(Y_1 = y_1, \dots, Y_{k-1} = y_{k-1}, Y_k = y_k) > 0$ . Clearly,  $y_k \notin \bar{D}$

since for any  $y \in \bar{D}$ ,

$$\mathbf{P}(Y_k = y) = \sum_{x_0, \dots, x_{k-1} \in \mathbb{E}_X} c_{yx_{k-1}} a_{x_{k-1}x_{k-2}} \cdots a_{x_1x_0} p_{x_0} = 0.$$

Also  $\mathbf{P}(Y_1 = y_1, \dots, Y_{k-1} = y_{k-1}) > 0$ . We next assume the conclusion in the lemma holds at time step  $k-1$ . Hence  $\mathbf{P}(X_{k-1} = x_{k-1}, Y_1 = y_1, \dots, Y_{k-1} = y_{k-1}) > 0$  for any  $x_{k-1}$ . It follows from

$$\begin{aligned} & \mathbf{P}(X_k = x_k, Y_1 = y_1, \dots, Y_k = y_k) \\ &= \sum_{x_{k-1} \in \mathbb{E}_X} a_{x_k x_{k-1}} c_{y_k x_{k-1}} \mathbf{P}(X_{k-1} = x_{k-1}, Y_1 = y_1, \dots, Y_{k-1} = y_{k-1}) > 0 \end{aligned}$$

that any state is observable at time step  $k$ .

We now prove the necessity part. We first consider part 1. For any  $y_1$  satisfying  $D_{y_1} \neq \emptyset$ , it follows that

$$\mathbf{P}(Y_1 = y_1) = \sum_{x_0 \in \mathbb{E}_X} c_{y_1 x_0} p_{x_0} > 0;$$

that is, the 1-step sequence  $y_1$  is feasible. Since in this part of the proof we assume that any state is observable, it follows that for any  $x_1$ ,

$$\mathbf{P}(X_1 = x_1, Y_1 = y_1) = \sum_{x_0 \in \mathbb{E}_X} a_{x_1 x_0} c_{y_1 x_0} p_{x_0} > 0.$$

Hence for any  $x_1 \in \mathbb{E}_X$ , there must exist an  $x_0 \in D_{y_1}$  such that  $a_{x_1 x_0} c_{y_1 x_0} p_{x_0} > 0$ . Since the condition  $x_0 \in D_{y_1}$  implies that  $c_{y_1 x_0} p_{x_0} > 0$ , then  $a_{x_1 x_0} > 0$ , and condition 1 of the lemma holds. It is clear from this proof that  $x_0$  depends on  $y_1$  and  $x_1$ .

We now establish condition 2. Since the sets  $\{Y_1 = y_1 | y_1 \in \mathbb{E}_Y\}$  form a finite partition of the sample space, there exists a  $y_1$  such that  $\mathbf{P}(Y_1 = y_1) > 0$ . Then, since any state of the HMM is assumed to be observable, we have that for any  $x_1$ ,

$$\mathbf{P}(X_1 = x_1, Y_1 = y_1) = \sum_{x_0 \in \mathbb{E}_X} a_{x_1 x_0} c_{y_1 x_0} p_{x_0} > 0.$$

Moreover, it follows from the definition of the set  $\bar{D}$  given in condition 2 of this lemma that for any  $y_2 \in \mathbb{E}_Y - \bar{D}$ , there exists an  $x_1$  such that  $c_{y_2 x_1} > 0$ , and hence

$$\mathbf{P}(Y_1 = y_1, Y_2 = y_2) = \sum_{x_1 \in \mathbb{E}_X} c_{y_2 x_1} \mathbf{P}(X_1 = x_1, Y_1 = y_1) > 0.$$

Note that  $\bar{D} \neq \mathbb{E}_Y$  since  $\sum_{y_2 \in \mathbb{E}_Y} c_{y_2 x_1} = 1$ . That is,  $\mathbb{E}_Y - \bar{D} \neq \emptyset$ . Finally, it follows from the fact that any state is observable that

$$\mathbf{P}(X_2 = x_2, Y_1 = y_1, Y_2 = y_2) = \sum_{x_1 \in \mathbb{E}_X} c_{y_2 x_1} a_{x_2 x_1} \mathbf{P}(X_1 = x_1, Y_1 = y_1) > 0.$$

This implies that for any  $y_2 \in \mathbb{E}_Y - \bar{D}$  and any  $x_2 \in \mathbb{E}_X$ , there exists an  $x_1 \in \mathbb{E}_X$  such that  $a_{x_2 x_1} c_{y_2 x_1} > 0$ . That is, condition 2 holds. Here  $x_1$  is a function of  $y_2$  and  $x_2$ . This completes the proof of the necessity of conditions 1 and 2 in this lemma.

**Appendix B. Proof of Lemma 4.1.** Since under  $\tilde{\mathbf{P}}_k$ ,  $\sigma\{Y_k\}$  is independent of  $\sigma\{X_0, \dots, X_k, Y_1, \dots, Y_{k-1}\}$ , we have

$$\begin{aligned}
& \alpha_k(e_i, y_{1,k}) \\
&= \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \bar{\lambda}_k \exp(q(X_{k-1})) \bar{\Lambda}_{k-1} \exp \left( \sum_{l=1}^{k-1} q(X_{l-1}) \right) \langle X_k, e_i \rangle | \mathcal{G}_{k-1}, Y_k \right] | y_{1,k} \right] \\
&= \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \bar{\lambda}_k \exp(q(X_{k-1})) \bar{\Lambda}_{k-1} \exp \left( \sum_{l=1}^{k-1} q(X_{l-1}) \right) \mathbf{E}^{\tilde{\mathbf{P}}_k} [\langle X_k, e_i \rangle | \mathcal{G}_{k-1}, Y_k] | y_{1,k} \right] \\
&= \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \bar{\lambda}_k \exp(q(X_{k-1})) \bar{\Lambda}_{k-1} \exp \left( \sum_{l=1}^{k-1} q(X_{l-1}) \right) \langle AX_{k-1}, e_i \rangle | y_{1,k} \right] \\
&= \sum_{n=1}^N \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \bar{\lambda}_k \exp(q(X_{k-1})) \bar{\Lambda}_{k-1} \exp \left( \sum_{l=1}^{k-1} q(X_{l-1}) \right) \langle AX_{k-1}, e_i \rangle \right. \\
&\quad \left. \times \langle X_{k-1}, e_n \rangle | y_{1,k} \right] \\
&= \sum_{n=1}^N \prod_{j=1}^M (M c_{jn})^{y_k^j} \exp(d_n) \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \bar{\Lambda}_{k-1} \exp \left( \sum_{l=1}^{k-1} q(X_{l-1}) \right) \langle AX_{k-1}, e_i \rangle \right. \\
&\quad \left. \times \langle X_{k-1}, e_n \rangle | y_{1,k-1} \right]
\end{aligned}$$

(B.1)

$$= \sum_{n=1}^N \prod_{j=1}^M (M c_{jn})^{y_k^j} \exp(d_n) a_{in} \alpha_{k-1}(e_n).$$

Using the matrix notation on the right-hand side of (B.1) yields (4.5). Furthermore, (4.6) follows from (4.5) as a special case in which  $q(X_{l-1}) = 0, l = 1, \dots, k$ . Note that (4.6) defines a risk-neutral recursive filter for HMMs [8].

**Appendix C. Proof of Theorem 4.1.** Using arguments similar to those used in the proof of Lemma 4.1, we have

$$\begin{aligned}
& \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \bar{\Lambda}_k \exp \left( \tau^{-1} \Psi_k(X_k, \xi) + \sum_{l=1}^k q(X_{l-1}) \right) | y_{1,k} \right] \\
&= \sum_{i=1}^N \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \bar{\Lambda}_k \exp \left( \tau^{-1} \Psi_k(X_k, \xi) + \sum_{l=1}^k q(X_{l-1}) \right) \langle X_k, e_i \rangle | y_{1,k} \right] \\
&= \sum_{i=1}^N \exp \left( \frac{1}{\tau} (e_i - \xi)' Q_k (e_i - \xi) \right) \mathbf{E}^{\tilde{\mathbf{P}}_k} \left[ \bar{\Lambda}_k \exp \left( \sum_{l=1}^k q(X_{l-1}) \right) \langle X_k, e_i \rangle | y_{1,k} \right] \\
&= H_k \alpha_k(y_{1,k})
\end{aligned}$$

and

$$\mathbf{E}^{\tilde{\mathbf{P}}_k}[\bar{\Lambda}_k|y_{1,k}] = \sum_{i=1}^N \beta_k(e_i, y_{1,k}).$$

Then (4.8) follows.

**Appendix D. Proof of Lemma 4.2.** It is obvious that we need only establish  $\beta_k(e_i, y_{1,k}) > 0$  for any  $e_i$  since  $\alpha_k(e_i, y_{1,k}) > \beta_k(e_i, y_{1,k})$ . By the conditional Bayes theorem,

$$(D.1) \quad \mathbf{E}^{\mathbf{P}_k}[\langle X_k, e_i \rangle | y_{1,k}] = \frac{\beta(e_i, y_{1,k})}{\sum_{i=1}^N \beta(e_i, y_{1,k})}.$$

Under Assumption 2.1, any state is observable at any  $k$ . Hence,

$$\mathbf{E}^{\mathbf{P}_k}[\langle X_k, e_i \rangle | y_{1,k}] = \mathbf{P}_k(X_k = e_i | y_{1,k}) > 0;$$

from this and (D.1), then  $\beta(e_i, y_{1,k}) > 0$  follows.

**Appendix E. Proof of Lemma 5.1.** To establish that  $V_k(\tau, \xi, y_{1,k})$  is continuous with respect to  $\tau$  over  $[0, \infty)$ , we need only prove that  $V_k(\tau, \xi, y_{1,k})$  is right continuous at  $\tau = 0$ . This can be seen from  $\lim_{\tau \rightarrow 0^+} V_k(\tau, \xi, y_{1,k}) = V_k(0, \xi, y_{1,k}) = \bar{b}_k$ . Furthermore, after a tedious computation, we have

$$\frac{dV_k(\tau, \xi, y_{1,k})}{d\tau} \Big|_{\tau=0^+} = \lim_{\tau \rightarrow 0^+} \frac{d\tau(\log J_k(\xi, \tau, y_{1,k}) + d)}{d\tau}.$$

This means that at  $\tau = 0$ ,  $\frac{dV_k(\tau, \xi, y_{1,k})}{d\tau}$  is right continuous. Thus,  $\frac{dV_k(\tau, \xi, y_{1,k})}{d\tau}$  is continuous with respect to  $\tau$  over  $[0, \infty)$ . We now calculate the second derivative of  $\tau(\log J_k(\xi, \tau, y_{1,k}) + d)$  with respect to  $\tau$ . Let  $a_i = \exp(\frac{1}{\tau}(e_i - \xi))' Q_k(e_i - \xi) \alpha_k(e_i, y_{1,k})$ ,  $i \in \mathcal{S}$ , and  $a_j = \alpha_k(\xi, y_{1,k})$ . By (5.2), we have

$$\begin{aligned} & \frac{d^2 \tau(\log J_k(\xi, \tau, y_{1,k}) + d)}{d\tau^2} \\ &= \frac{1}{\tau^3} \left( \frac{\sum_{i \in \mathcal{S}} a_i b_i^2 \sum_{i \in \mathcal{S}} a_i - (\sum_{i \in \mathcal{S}} a_i b_i)^2}{(\sum_{i=1}^N a_i)^2} + \frac{a_j \sum_{i \in \mathcal{S}} a_i b_i^2}{(\sum_{i=1}^N a_i)^2} \right) \\ &= \frac{1}{\tau^3} \left( \frac{\sum_{i \in \mathcal{S}} (b_i \sqrt{a_i})^2 \sum_{i \in \mathcal{S}} (\sqrt{a_i})^2 - (\sum_{i \in \mathcal{S}} b_i \sqrt{a_i} \sqrt{a_i})^2}{(\sum_{i=1}^N a_i)^2} + \frac{a_j \sum_{i \in \mathcal{S}} a_i b_i^2}{(\sum_{i=1}^N a_i)^2} \right) > 0. \end{aligned}$$

The last inequality follows from the Cauchy inequality  $(\sum_{i=1}^n a_i b_i)^2 \leq \sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2$  and the fact that  $a_j \sum_{i \in \mathcal{S}} a_i b_i^2 > 0$  since  $b_l > 0$ , for all  $l \in \mathcal{S}$ , and  $a_i > 0$  for any  $i$ ; see Lemma 4.2. Thus,  $\tau(\log J_k(\xi, \tau, y_{1,k}) + d)$  is a strictly convex function of  $\tau$  on  $(0, \infty)$ .

**Appendix F. Proof of Corollary 5.1.** When  $d_r = 0, r = 1, \dots, N$ , we have  $\alpha_k(y_{1,k}) = \beta_k(y_{1,k})$ . Then from inequality (5.3), the definition of  $\alpha_k(y_{1,k})$  and  $\beta_k(y_{1,k})$ , and the conditional Bayes theorem, we have

$$\begin{aligned} \log \frac{\sum_{i \in \mathcal{W}} \alpha_k(e_i, y_{1,k})}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})} &= \log \frac{\sum_{i \in \mathcal{W}} \mathbf{E}^{\tilde{\mathbf{P}}_k}[\bar{\Lambda}_k \langle X_k, e_i \rangle | y_{1,k}]}{\mathbf{E}^{\tilde{\mathbf{P}}_k}[\bar{\Lambda}_k | y_{1,k}]} \\ &= \log \sum_{i \in \mathcal{W}} \mathbf{E}^{\mathbf{P}}[\langle X_k, e_i \rangle | y_{1,k}] = \log \sum_{i \in \mathcal{W}} \mathbf{P}(X_k = e_i | y_{1,k}) \end{aligned}$$

from which (5.6) follows.

**Appendix G. Proof of Lemma 5.2.** From (B.1),

$$\begin{aligned}
 -\log \sum_{i \in \mathcal{W}} \mathbf{P}(X_k = e_i | y_{1,k}) &= -\log \frac{\sum_{i \in \mathcal{W}} \alpha_k(e_i, y_{1,k})}{\sum_{i=1}^N \alpha_k(e_i, y_{1,k})} \\
 &= -\log \frac{\sum_{i \in \mathcal{W}} \sum_{n=1}^N \prod_{j=1}^M (Mc_{jn})^{y_k^j} a_{in} \alpha_{k-1}(e_n, y_{1,k})}{\sum_{n=1}^N \prod_{j=1}^M (Mc_{jn})^{y_k^j} \alpha_{k-1}(e_n, y_{1,k})} \\
 (G.1) \quad &\geq -\log \left( \sum_{i \in \mathcal{W}} \max_{j=1, \dots, N} a_{ij} \right).
 \end{aligned}$$

In order to derive the last inequality, we use the assumption  $\bar{a} < 1$  in this lemma, which implies that all entries of the state transition matrix  $A$  are greater than zero. Furthermore, under the assumption in this lemma and from (G.1), we have

$$d < -\log \bar{a} \leq -\log \left( \sum_{i \in \mathcal{W}} \max_{j=1, \dots, N} a_{ij} \right) \leq -\log \sum_{i \in \mathcal{W}} \mathbf{P}(X_k = e_i | y_{1,k}).$$

That is, inequality (5.6) holds such that a stationary point always exists for any  $\xi$  and  $k$ . Note that any such  $\bar{d}$  is independent of the measurements. Also the condition  $\bar{a} < 1$ , and  $d < \bar{d}$  is a sufficient condition under which inequality (5.6) holds.

**Appendix H. Proof of Lemma 5.3.** We first prove the first statement of this lemma. From the proof of the second part of Lemma 5.1, the first derivative of the right-hand side of the above equation with respect to  $\tau_k^*$  is strictly less than zero. Hence,  $\tau_k^*$  is strictly decreasing with  $d$  increasing if the inequality (5.3) holds.

We now prove the second statement of this lemma. Using the same argument as used in Lemma 5.1, we can calculate the first derivative of the right-hand side of (5.18) with respect to  $\tau_k^*$  as follows:

$$\begin{aligned}
 &-\frac{1}{(\tau_k^*)^2} \left( \frac{\left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) b_i^2 \right) \left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k}) \right)}{\left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k}) \right)^2} \right. \\
 &\quad \left. - \frac{\left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) b_i \right)^2}{\left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k}) \right)^2} \right) \\
 &= \frac{-1}{(\tau_k^*)^2} \frac{\left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) b_i^2 \right) \left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) \right)}{\left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k}) \right)^2} \\
 &\quad - \frac{1}{(\tau_k^*)^2} \frac{-\left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) b_i \right)^2 + \alpha_k(\xi, y_{1,k}) \left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) b_i^2 \right)}{\left( \sum_{i \in \mathcal{S}} \exp\left(\frac{b_i}{\tau_k^*}\right) \alpha_k(e_i, y_{1,k}) + \alpha_k(\xi, y_{1,k}) \right)^2} \\
 &< 0.
 \end{aligned}$$



Hence, the minimum cost value is strictly monotone increasing with  $\tau_k^*$  decreasing. From the first statement of this lemma, we can now conclude the minimum cost value is strictly monotone increasing with  $d$  increasing.

Finally, we assume that  $d_j, 1 \leq j \leq N$ , increases. Let  $\bar{d}_j = d_j + \delta, \delta > 0$ . Then from (5.7), for any  $k$ , we have

$$\begin{aligned} \bar{d}_c &= -\log \frac{\sum_{i \in \mathcal{W}} \alpha_k(e_i, y_{1,k})}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})} \\ &= -\log \frac{\sum_{i \in \mathcal{W}} \mathbf{E}^{\bar{\mathbf{P}}_k} [\bar{\Lambda}_k \exp(\sum_{l=1}^k q(X_{l-1})) \exp(\sum_{l=1}^k \delta \langle X_j, e_r \rangle) \langle X_k, e_i \rangle | y_{1,k}]}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})} \\ &\leq -\log \frac{\sum_{i \in \mathcal{W}} \mathbf{E}^{\bar{\mathbf{P}}_k} [\bar{\Lambda}_k \exp(\sum_{l=1}^k q(X_{l-1})) \langle X_k, e_i \rangle | y_{1,k}]}{\sum_{i=1}^N \beta_k(e_i, y_{1,k})} = d_c. \end{aligned}$$

Hence from the definition of the maximum horizon, we have  $\bar{K} \leq K$ . It follows directly from (5.3) that  $K$  is decreasing as  $d$  increases. This completes the proof of the lemma.

#### REFERENCES

- [1] R. K. BOEL, M. R. JAMES, AND I. R. PETERSEN, *Robustness and risk-sensitive filtering*, IEEE Trans. Automat. Control, 47 (2002), pp. 451–461.
- [2] Y. S. CHOW AND H. TEICHER, *Probability Theory, Independence, Interchangeability, Martingales*, 2nd ed., Springer-Verlag, New York, 1988.
- [3] K. L. CHUNG, *A Course in Probability Theory*, 2nd ed., Academic Press, New York, London, 1974.
- [4] S. DEY AND J. B. MOORE, *Risk-sensitive filtering and smoothing for hidden Markov models*, Systems Control Lett., 25 (1995), pp. 361–366.
- [5] F. DUFOUR AND R. J. ELLIOTT, *Filtering with discrete state observations*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997.
- [6] F. DUFOUR AND R. J. ELLIOTT, *Adaptive control of linear systems with Markov perturbations*, IEEE Trans. Automat. Control, 44 (1999), pp. 2271–2282.
- [7] P. DUPUIS AND E. ELLIS, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, New York, 1997.
- [8] R. J. ELLIOTT, L. AGGOUN, AND J. B. MOORE, *Hidden Markov Models: Estimation and Control*, Springer-Verlag, New York, 1994.
- [9] Y. E. EPHRAIM AND N. MERHAV, *Hidden Markov processes*, IEEE Trans. Inform. Theory, 48 (2002), pp. 1518–1569.
- [10] B. C. LEVY AND R. NIKOUKHAH, *Robust least-squares estimation with a relative entropy constraint*, IEEE Trans. Inform. Theory, 50 (2004), pp. 89–104.
- [11] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1997.
- [12] W. P. MALCOLM, R. J. ELLIOTT, F. DUFOUR, AND M. R. ARULAMPALAM, *An algorithmic estimation scheme for hybrid stochastic systems*, in Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference, Seville, Spain, 2005.
- [13] J. R. NORRIS, *Markov Chains*, Cambridge University Press, Cambridge, UK, 1997.
- [14] I. R. PETERSEN, M. R. JAMES, AND P. DUPUIS, *Minimax optimal control of stochastic uncertain systems with relative entropy constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 398–412.
- [15] L. R. RABINER, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proc. IEEE, 77 (1989), pp. 257–286.
- [16] V. R. RAMEZANI AND S. I. MARCUS, *Estimation of hidden Markov models: Risk-sensitive filter banks and qualitative analysis of their sample paths*, IEEE Trans. Automat. Control, 47 (2002), pp. 1999–2009.
- [17] P. C. SHIELDS, *The Ergodic Theory of Discrete Sample Paths*, Grad. Stud. Math. 13, AMS, Providence, RI, 1996.

- [18] A. N. SHIRYAYEV, *Probability*, Springer-Verlag, New York, 1984.
- [19] V. A. UGRINOVSKII AND I. R. PETERSEN, *Finite horizon minimax optimal control of stochastic partially observed time varying uncertain systems*, Math. Control Signals Systems, 12 (1999), pp. 1–23.
- [20] V. A. UGRINOVSKII AND I. R. PETERSEN, *Robust filtering of stochastic uncertain systems on an infinite time horizon*, Internat. J. Control, 75 (2002), pp. 614–626.
- [21] L. XIE, V. A. UGRINOVSKII, AND I. R. PETERSEN, *A duality relationship for regular conditional relative entropy*, in Proceedings of the 16th IFAC World Congress, Prague, 2005. A version of this paper has been submitted to Ann. Appl. Probab.
- [22] L. XIE, V. A. UGRINOVSKII, AND I. R. PETERSEN, *A posteriori probabilistic distances between finite-alphabet hidden Markov models*, IEEE Trans. Inform. Theory, 53 (2007), pp. 783–793.
- [23] L. XIE, V. A. UGRINOVSKII, AND I. R. PETERSEN, *Probabilistic distances between finite-state finite-alphabet hidden Markov models*, IEEE Trans. Automat. Control, 50 (2005), pp. 505–511.
- [24] M. G. YOON, V. A. UGRINOVSKII, AND I. R. PETERSEN, *Robust finite horizon minimax filtering for discrete time stochastic uncertain systems*, Systems Control Lett., 52 (2004), pp. 99–112.
- [25] M. G. YOON, V. A. UGRINOVSKII, AND I. R. PETERSEN, *On the worst-case disturbance of minimax optimal control*, Automatica, 41 (2005), pp. 847–855.
- [26] Q. ZHANG AND G. YIN, *On nearly optimal controls of hybrid LQG problems*, IEEE Trans. Automat. Control, 44 (1999), pp. 2271–2282.

## ADAPTIVE FINITE ELEMENTS FOR ELLIPTIC OPTIMIZATION PROBLEMS WITH CONTROL CONSTRAINTS\*

B. VEXLER<sup>†</sup> AND W. WOLLNER<sup>‡</sup>

**Abstract.** In this paper we develop a posteriori error estimates for finite element discretization of elliptic optimization problems with pointwise inequality constraints on the control variable. We derive error estimators for assessing the discretization error with respect to the cost functional as well as with respect to a given quantity of interest. These error estimators provide quantitative information about the discretization error and guide an adaptive mesh refinement algorithm allowing for substantial saving in degrees of freedom. The behavior of the method is demonstrated on numerical examples.

**Key words.** mesh adaptivity, optimal control, a posteriori error estimates, finite element method, quantity of interest, pointwise inequality constraints

**AMS subject classifications.** 65N50, 65N30, 65K10

**DOI.** 10.1137/070683416

**1. Introduction.** In this paper we develop a posteriori error estimates for finite element approximations of optimization problems governed by elliptic partial differential equations. We discuss this question in a general manner, including the consideration of optimal control and parameter identification problems with control constraints given through a closed convex admissible set. The derived error estimates have the goal of guiding an adaptive mesh refinement algorithm for finding economical meshes for the optimization problem under consideration.

The use of adaptive techniques based on a posteriori error estimation is well accepted in the context of finite element discretization of partial differential equations; see, e.g., [6, 13, 35]. To our knowledge there are only a few results published on adaptive finite elements for optimization problems; see [2, 17, 20, 23, 25, 27, 4, 7, 8, 30].

In articles [17, 20, 23, 25, 27] the authors provide a posteriori error estimates for elliptic optimal control problems with distributed or Neumann control subject to box constraints. These estimates assess the error in the control, state, and the adjoint variable with respect to the natural norms of the corresponding spaces. In [2] another approach for the estimation of the error with respect to the norm of the control space is presented. In [17] convergence of an adaptive algorithm for a control constrained optimal control problem is shown.

However, in many applications, the error in global norms does not provide a useful error bound for the error in the quantity of physical interest. The a posteriori estimators derived in this paper grant access to the error with respect to given functionals.

---

\*Received by the editors February 22, 2007; accepted for publication (in revised form) August 10, 2007; published electronically February 1, 2008.

<http://www.siam.org/journals/sicon/47-1/68341.html>

<sup>†</sup>Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenberger Straße 69, 4040 Linz, Austria (boris.vexler@oeaw.ac.at). This author's research was partially supported by the Austrian Science Fund FWF project P18971-N18 "Numerical analysis and discretization strategies for optimal control problems with singularities."

<sup>‡</sup>Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, INF 294, 69120 Heidelberg, Germany (winnifried.wollner@iwr.uni-heidelberg.de). This author's research was supported by the DFG priority program "Optimization with Partial Differential Equations."

In [4, 6] the authors present a general concept for a posteriori estimation of the discretization error with respect to the cost functional in the context of optimal control problems. In articles [7, 8] the authors have extended this approach to the estimation of the discretization error with respect to an arbitrary functional depending on both the control and the state variable, so-called *quantity of interest*. This allowed, among other things, the treatment of parameter identification and model calibration problems. However, in all these publications, the control variable was searched for in a Hilbert space  $Q$  without additional (inequality) constraints. Therefore the main contribution of this work is the extension of these techniques to the case of optimization problems with additional control constraints given through a closed convex admissible set  $Q_{\text{ad}} \subset Q$ . In the majority of practical cases this admissible set is described by inequality control constraints of box type  $q_- \leq q(x) \leq q_+$ . Therefore we will concentrate on this case, although our techniques may also be extended to the consideration of more general admissible sets  $Q_{\text{ad}}$ .

In this paper we consider optimization problems governed by (nonlinear) partial differential equations. The aim is to minimize a given cost functional  $J(q, u)$  which depends on the state variable  $u \in V$  and the control variable  $q \in Q$ , with Hilbert spaces  $V$  and  $Q$ . These variables have to satisfy the state equation

$$(1.1) \quad A(q, u) = f,$$

where  $A$  denotes a (nonlinear) differential operator and  $f$  represents the given data. The optimization problem is then formulated as follows:

$$(1.2) \quad \begin{cases} \text{Minimize } J(q, u), & u \in V, q \in Q_{\text{ad}}, \\ A(q, u) = f. \end{cases}$$

Constraints on the control are incorporated via the definition of the closed and convex set  $Q_{\text{ad}}$  representing the set of admissible controls.

For numerical treatment this infinite dimensional optimization problem is discretized in virtue of finite element methods; see the discussion in section 3. Let the solution to the discretized problem be denoted by  $(q_h, u_h)$ . Our aim is to derive a posteriori error estimates for the error between the solutions to the continuous and the discrete problem. A crucial point for our error analysis is the choice of a quantity, which describes the goal of the computation. If this quantity coincides with the cost functional, we have to estimate the error

$$J(q, u) - J(q_h, u_h).$$

In a more general case, we suppose  $I: Q \times V \rightarrow \mathbb{R}$  to be a given functional describing the quantity of interest. Then the error to be estimated is

$$I(q, u) - I(q_h, u_h).$$

The consideration of quantities of interest is important, for instance, in the context of parameter identification and model calibration problem; see [8] for an application of this concept to an optimization problem from computational fluid dynamics.

To the authors' knowledge this is the first article providing a posteriori error estimates with respect to a given functional for optimization problems with partial differential equations and subject to control constraints.

The paper is organized as follows. In the next section we describe the optimization problem under consideration, discuss necessary optimality conditions, and sketch the

solution algorithm on the continuous level. In section 3 we describe the discretization of the optimization problem in virtue of finite element methods. Section 4 is devoted to a posteriori error estimation. In sections 4.1 and 4.2 we derive two different error estimates for the error with respect to the cost functional  $J$ . The first error estimator is based on the optimality system involving a variational inequality, whereas the second one exploits Lagrange multipliers for the treatment of inequality constraints. Due to the fact that the optimal control  $q$  is not expected to be sufficiently smooth (due to inequality constraints), the approximation of (interpolation) weights involved in the error estimator cannot be treated in a usual way. To overcome this difficulty we exploit the projection formula (2.9) from the optimality conditions and propose an approximation on the (interpolation) weights using a postprocessing step (4.7), which is motivated by the considerations in [31]. In section 4.3 we provide an error estimator with respect to a given quantity of interest. To this end we utilize an additional (dual) linear-quadratic optimal control problem describing the sensitivity with respect to the quantity of interest. In the last section we present numerical examples to illustrate the behavior of our method.

**2. Optimization problem.** In this section we give a precise formulation of the optimization problem under consideration and describe necessary optimality conditions and the solution algorithm.

In order to deal with different types of optimization problems simultaneously, we seek the control variable  $q$  in the Hilbert space  $Q = L^2(\omega)$  with scalar product  $(\cdot, \cdot)$  and norm  $\|\cdot\|$ . Typically,  $\omega$  is a subset of the computational domain  $\Omega$  or a subset of its boundary  $\partial\Omega$ . The case of finite dimensional controls is realized by choosing  $\omega = \{1, 2, \dots, n\}$  resulting in  $Q \cong \mathbb{R}^n$ .

Throughout this paper we suppose that the state equation (1.1) for  $u \in V$  is given in a weak form:

$$(2.1) \quad a(q, u)(\varphi) = f(\varphi) \quad \forall \varphi \in V,$$

where  $a: Q \times V \times V \rightarrow \mathbb{R}$  is a four times directional differentiable form which is linear in the third argument and  $f$  is in the dual space  $V'$ . A possible choice for this space is  $V = H^1(\Omega)$ , or  $V = H_0^1(\Omega)$ , or a direct product of such spaces. In the presence of inhomogeneous Dirichlet boundary conditions, one seeks the state variable  $u$  in  $\hat{u} + V$ , where  $\hat{u}$  represents the boundary data. However, for clarity of notation, we assume throughout that  $\hat{u} = 0$ .

*Remark 2.1.* Throughout this paper we use two pairs of parentheses after a form to indicate that the form is linear in all variables enclosed by the second pair of parentheses, as seen in (2.1) for  $a(\cdot, \cdot)(\cdot)$ .

The cost function is given by

$$(2.2) \quad J(q, u) = J_1(u) + \frac{\alpha}{2} \|q\|^2,$$

where  $J_1$  is a four times directionally differentiable operator on  $V$  and  $\alpha > 0$ . Let the admissible set  $Q_{\text{ad}}$  be given through box constraints on  $q$ , i.e.,

$$(2.3) \quad Q_{\text{ad}} = \{q \in Q \mid q_- \leq q(x) \leq q_+ \text{ a.e. on } \omega\},$$

with bounds  $q_-, q_+ \in \mathbb{R} \cup \{\pm\infty\}$  and  $q_- < q_+$ .

Now we are able to formulate the optimization problem as

$$(2.4) \quad \text{Minimize } J(q, u), \quad u \in V, q \in Q_{\text{ad}}, \quad \text{subject to (2.1).}$$

*Remark 2.2.* The choice of constant bounds  $q_-, q_+ \in \mathbb{R} \cup \{\pm\infty\}$  is not a limitation, since one can transform an optimal control problem with bounds  $q_-, q_+ \in Q$  into an equivalent one with constant bounds for the control.

To shorten notation we introduce the space  $\mathcal{X}$  and the admissible set  $\mathcal{X}_{\text{ad}}$  by

$$(2.5) \quad \mathcal{X} = Q \times V \times V,$$

$$(2.6) \quad \mathcal{X}_{\text{ad}} = Q_{\text{ad}} \times V \times V.$$

In addition we shall write  $\xi = (q, u, z)$  for a vector in  $\mathcal{X}$  or  $\mathcal{X}_{\text{ad}}$ , where  $z$  will denote an adjoint state.

Throughout the paper we assume that the problem (2.4) admits a solution. Conditions ensuring the existence of solutions to optimal control problems may, for instance, be found in [16, 26, 34]. We shall especially assume that the primal and dual equations associated with (2.4) are solvable for every given  $q \in Q$ .

To establish an optimality system, we introduce the Lagrangian  $\mathcal{L}: \mathcal{X} \rightarrow \mathbb{R}$  as follows:

$$\mathcal{L}(\xi) = J_1(u) + \frac{\alpha}{2} \|q\|^2 + f(z) - a(q, u)(z),$$

where  $z$  denotes the dual variable. Due to the convexity of the admissible set  $Q_{\text{ad}}$ , the first-order necessary optimality condition for  $(q, u) \in Q_{\text{ad}} \times V$  reads as follows:

There exists  $z \in V$  such that the triple  $\xi = (q, u, z) \in \mathcal{X}_{\text{ad}}$  satisfies

$$(2.7a) \quad \mathcal{L}'_u(\xi)(\delta u) = 0 \quad \forall \delta u \in V,$$

$$(2.7b) \quad \mathcal{L}'_q(\xi)(\delta q - q) \geq 0 \quad \forall \delta q \in Q_{\text{ad}},$$

$$(2.7c) \quad \mathcal{L}'_z(\xi)(\delta z) = 0 \quad \forall \delta z \in V.$$

This system can be stated explicitly in the following form:

$$(2.8a) \quad J'_1(u)(\delta u) - a'_u(q, u)(\delta u, z) = 0 \quad \forall \delta u \in V,$$

$$(2.8b) \quad \alpha(q, \delta q - q) - a'_q(q, u)(\delta q - q, z) \geq 0 \quad \forall \delta q \in Q_{\text{ad}},$$

$$(2.8c) \quad f(\delta z) - a(q, u)(\delta z) = 0 \quad \forall \delta z \in V.$$

We introduce a projection operator  $\mathcal{P}_{Q_{\text{ad}}}: Q \rightarrow Q_{\text{ad}}$  by

$$\mathcal{P}_{Q_{\text{ad}}}(p) = \max(q_-, \min(p, q_+))$$

pointwise a.e. This allows us to rewrite variational inequality (2.8b) (see, e.g., [34]) as

$$(2.9) \quad q = \mathcal{P}_{Q_{\text{ad}}} \left( \frac{1}{\alpha} a'_q(q, u)(\cdot, z) \right),$$

where  $a'_q(u, q)(\cdot, z)$  is understood as a Riesz representative of a linear functional on  $Q$ .

For a solution  $(q, u)$  of (2.4) we introduce active sets  $\omega_-$  and  $\omega_+$  as follows:

$$(2.10) \quad \omega_- = \{x \in \omega \mid q(x) = q_-\},$$

$$(2.11) \quad \omega_+ = \{x \in \omega \mid q(x) = q_+\}.$$

Let  $\xi \in \mathcal{X}$  be a solution to (2.7); then we introduce an additional Lagrange multiplier  $\mu \in Q$  by the following identification:

$$(2.12) \quad (\mu, \delta q) = -\alpha(q, \delta q) + a'_q(q, u)(\delta q, z) = -\mathcal{L}'_q(\xi)(\delta q) \quad \forall \delta q \in Q.$$

The variational inequality (2.8b) or the projection formula (2.9) are known to be equivalent to the following conditions:

$$(2.13a) \quad \mu(x) \leq 0 \quad \text{a.e. on } \omega_-,$$

$$(2.13b) \quad \mu(x) \geq 0 \quad \text{a.e. on } \omega_+,$$

$$(2.13c) \quad \mu(x) = 0 \quad \text{a.e. on } \omega \setminus (\omega_- \cup \omega_+).$$

Using this representation of the optimality condition (2.8b) we apply nonlinear primal dual active set strategy (see, e.g., [9, 24]) to solve (2.4). In the following we sketch the corresponding algorithm on the continuous level.

**Nonlinear primal-dual active set strategy**

1. Choose initial guess  $q^0, \mu^0$  and  $c > 0$  and set  $n = 1$ .
2. While not converged
3. Determine the active sets  $\omega_+^n$  and  $\omega_-^n$ :

$$\omega_-^n = \{x \in \omega \mid q^{n-1}(x) + \mu^{n-1}(x)/c - q_- \leq 0\},$$

$$\omega_+^n = \{x \in \omega \mid q^{n-1}(x) + \mu^{n-1}(x)/c - q_+ \geq 0\}.$$

4. Solve the equality-constrained optimization problem

$$\text{Minimize} \quad J_1(u^n) + \frac{\alpha}{2} \|q^n\|^2, \quad u^n \in V, q^n \in Q,$$

subject to (2.1) and

$$q^n(x) = q_- \text{ on } \omega_-^n, \quad q^n(x) = q_+ \text{ on } \omega_+^n.$$

5. Set

$$\mu^n = -\alpha q^n + a'_q(q^n, u^n)(\cdot, z^n)$$

with adjoint variable  $z^n$ .

6. Set  $n = n + 1$  and go to 2.

*Remark 2.3.* The convergence in step 2 can be determined conveniently from agreement of the active sets in two consecutive iterations.

*Remark 2.4.* The algorithm above is known to be globally convergent for a class of optimal control problems if  $\alpha$  is sufficiently large; see, e.g., [9, 24]. Moreover, local superlinear convergence can be shown; see, e.g., [21].

In our practical realization, the equality-constrained optimization problem in step 4 is solved by Newton's method on the control space without assembling the Hessian. The finite element discretization of the optimization problem, described in the next section, allows us to directly translate these algorithms onto the discrete level.

As we will encounter some trouble with the variational inequality in the necessary optimality condition (2.8) due to missing Galerkin orthogonality, we consider in addition the full Lagrangian  $\tilde{\mathcal{L}}: \mathcal{X} \times Q \times Q \rightarrow \mathbb{R}$  which is given by

$$\tilde{\mathcal{L}}(\chi) = \mathcal{L}(\xi) + (\mu^-, q_- - q) + (\mu^+, q - q_+),$$

with  $\chi = (\xi, \mu^-, \mu^+) = (q, u, z, \mu^-, \mu^+) \in \mathcal{X} \times Q \times Q$ , where  $\mu^-$  and  $\mu^+$  denote the variables corresponding to Lagrange multipliers for the inequality constraints. To shorten notation we introduce the abbreviation

$$(2.14) \quad \mathcal{Y} = \mathcal{X} \times Q \times Q.$$

Using the subspaces

$$\begin{aligned} Q_- &= \{r \in Q \mid r = 0 \text{ a.e. on } \omega \setminus \omega_-\}, \\ Q_+ &= \{r \in Q \mid r = 0 \text{ a.e. on } \omega \setminus \omega_+\}, \end{aligned}$$

we introduce

$$(2.15) \quad \mathcal{Y}_{\text{ad}} = \mathcal{X}_{\text{ad}} \times Q_- \times Q_+,$$

$$(2.16) \quad \tilde{\mathcal{Y}}_{\text{ad}} = \mathcal{X} \times Q_- \times Q_+$$

and see that the following equality holds for all  $\chi \in \mathcal{Y}_{\text{ad}}$ :

$$(2.17) \quad \mathcal{L}(\xi) = \tilde{\mathcal{L}}(\chi).$$

We can rewrite the first-order necessary optimality condition for  $(q, u) \in Q_{\text{ad}} \times V$  equivalently as follows (cf. [34]):

There exist  $z \in V$ ,  $\mu^- \in Q_-$ ,  $\mu^+ \in Q_+$  such that the following conditions hold for  $\chi = (q, u, z, \mu^-, \mu^+) \in \mathcal{Y}_{\text{ad}}$ :

$$(2.18a) \quad \tilde{\mathcal{L}}'_u(\chi)(\delta u) = 0 \quad \forall \delta u \in V,$$

$$(2.18b) \quad \tilde{\mathcal{L}}'_q(\chi)(\delta q) = 0 \quad \forall \delta q \in Q,$$

$$(2.18c) \quad \tilde{\mathcal{L}}'_z(\chi)(\delta z) = 0 \quad \forall \delta z \in V,$$

$$(2.18d) \quad \tilde{\mathcal{L}}'_{\mu^-}(\chi)(\delta \mu^-) = 0 \quad \forall \delta \mu^- \in Q_-,$$

$$(2.18e) \quad \tilde{\mathcal{L}}'_{\mu^+}(\chi)(\delta \mu^+) = 0 \quad \forall \delta \mu^+ \in Q_+,$$

$$(2.18f) \quad \mu^+, \mu^- \geq 0 \quad \text{a.e. on } \omega.$$

It is easy to verify that the Lagrange multipliers  $\mu^+$  and  $\mu^-$  are given as the positive and negative part of the Lagrange multiplier  $\mu$  from (2.12); cf. [34].

Note that (2.18d), (2.18e) are equivalent to the complementarity conditions

$$(2.19) \quad \mu^-(q_- - q) = \mu^+(q - q_+) = 0 \quad \text{a.e. on } \omega.$$

For later use we recall a second-order sufficient optimality condition.

**LEMMA 2.1** (sufficient optimality condition). *Let  $\xi = (q, u, z) \in \mathcal{X}_{\text{ad}}$  satisfy the first-order necessary condition (2.7a)–(2.7c) of optimization problem (2.4). Moreover, let  $z \mapsto a'_u(q, u)(\cdot, z) : V \rightarrow V'$  be surjective. If there exists  $\rho > 0$  such that*

$$(2.20) \quad (\delta q, \delta u) \begin{bmatrix} \mathcal{L}''_{qq}(\xi)(\cdot, \cdot) & \mathcal{L}''_{qu}(\xi)(\cdot, \cdot) \\ \mathcal{L}''_{uq}(\xi)(\cdot, \cdot) & \mathcal{L}''_{uu}(\xi)(\cdot, \cdot) \end{bmatrix} \begin{pmatrix} \delta q \\ \delta u \end{pmatrix} \geq \rho (\|\delta u\|_V^2 + \|\delta q\|_Q^2)$$

*holds for all  $(\delta q, \delta u)$  satisfying the linear (tangent) partial differential equation*

$$(2.21) \quad a'_u(q, u)(\delta u, \varphi) + a'_q(q, u)(\delta q, \varphi) = 0 \quad \forall \varphi \in V,$$

*then  $(q, u)$  is a (strict) local solution to the optimization problem (2.4).*



We refer the reader to [29] for the proof.

*Remark 2.5.* Throughout the paper we exploit only first-order information. This means that the error estimators proposed in section 4 are applicable to all solutions of the optimality system (2.8) or (2.18), respectively.

For the convenience of the reader we list the assumptions made in the preceding section.

*Assumption 1.* The optimization problem (2.4) possesses a solution  $(q, u)$ . In addition there exists  $z \in V$  such that the first-order necessary conditions (2.8) are fulfilled by the triple  $(q, u, z)$ .

*Remark 2.6.* It is sufficient for the existence of  $z$  in the preceding assumption if the mapping  $z \mapsto a'_u(q, u)(\cdot, z)$  is surjective onto  $V'$ . This is one of the requirements in Lemma 2.1 and is fulfilled by all examples given in this article.

*Assumption 2.* The functional  $a(\cdot, \cdot)(\cdot) : Q \times V \times V \rightarrow \mathbb{R}$  defined in (2.1) is assumed to be four times directional differentiable.

*Assumption 3.* The functional  $J(\cdot, \cdot) : Q \times V \rightarrow \mathbb{R}$  defined in (2.2) is assumed to be four times directional differentiable.

*Assumption 4.* The functional  $I(\cdot, \cdot) : Q \times V \rightarrow \mathbb{R}$  mentioned in the introduction (see also (4.20)) is assumed to be three times directional differentiable.

**3. Finite element discretization.** In this section we discuss finite element discretization of the optimization problem (2.4).

To keep the following sections simple we restrain ourselves to the case of problems where  $H^1$ -conforming finite elements are satisfactory. However, the ideas can be adapted to other problems.

Let  $\mathcal{T}_h$  be a triangulation (mesh) of the computational domain  $\Omega$  consisting of closed cells  $K$  which are either triangles or quadrilaterals. The straight parts which make up the boundary  $\partial K$  of a cell  $K$  are called *faces*. The mesh parameter  $h$  is defined as a cellwise constant function by setting  $h|_K = h_K$ , and  $h_K$  is the diameter of  $K$ . The mesh  $\mathcal{T}_h$  is assumed to be shape regular. In order to ease the mesh refinement we allow the cells to have nodes, which lie on midpoints of faces of neighboring cells. But at most one of such *hanging nodes* is permitted per face.

On the mesh  $\mathcal{T}_h$  we define a finite element space  $V_h \subset V$  consisting of linear or bilinear shape functions; see, e.g., [14] or [10]. The case of hanging nodes requires some additional remarks. There are no degrees of freedom corresponding to these irregular nodes, and therefore the value of the finite element function is determined by pointwise interpolation. This implies continuity and therefore global conformity.

For the discretization of the optimization problem (2.4) we introduce an additional finite dimensional subspace  $Q_h \subset Q$  of the control space. Depending on the concrete situation there are different possible ways to choose the space  $Q_h$ . It is reasonable to set  $Q_h = Q$  if  $Q$  is finite dimensional. In the case where the control variable is a distributed function on the computational domain  $\Omega$ , i.e.,  $Q = L^2(\Omega)$ , one may choose  $Q_h$  as an analogue to  $V_h$  or consider  $Q_h$  as a space of cellwise constant functions on the mesh  $\mathcal{T}_h$ . A priori error analysis for the last two choices in the context of distributed (or boundary) elliptic optimal control problems can be found, e.g., in [1, 11, 15, 18, 28] for cellwise constant control or in [12, 32, 33] for continuous cellwise linear control. An approach without discretization of the control variable is presented in [22].

We denote a basis of  $Q_h$  by

$$(3.1) \quad \mathcal{B} = \{\psi_i\}, \text{ with } \psi_i \geq 0, \quad \sum_i \psi_i = 1, \quad \max_{x \in \omega} \psi_i(x) = 1.$$

*Remark 3.1.* It might be desirable to use different meshes for the control and the state variable in the case of distributed control. The error estimator presented below can provide information for separate refinement of the control and state meshes. One can split the error estimator into two parts, one containing the functionals on the space  $V$  which give information for the refinement of the state mesh and one part consisting of the functionals defined on the control space  $Q$  which give information for the refinement of the control mesh. The refinement then follows an equilibration strategy for both estimators; cf. [30].

The discrete admissible set  $Q_{\text{ad},h}$  is defined as

$$Q_{\text{ad},h} = Q_h \cap Q_{\text{ad}},$$

and the discretized optimization problem is formulated as follows:

$$(3.2) \quad \text{Minimize } J(q_h, u_h), \quad u_h \in V_h, \quad q_h \in Q_{\text{ad},h},$$

subject to

$$(3.3) \quad a(q_h, u_h)(v_h) = f(v_h) \quad \forall v_h \in V_h.$$

We introduce the discretized versions of (2.5) and (2.6) by

$$(3.4) \quad \mathcal{X}_h = Q_h \times V_h \times V_h,$$

$$(3.5) \quad \mathcal{X}_{\text{ad},h} = Q_{\text{ad},h} \times V_h \times V_h$$

and denote a vector from these sets by  $\xi_h = (q_h, u_h, z_h)$ . The optimality system for the discretized optimization problem is formulated as follows:

$$(3.6a) \quad J'_1(u_h)(\delta u_h) - a'_u(q_h, u_h)(\delta u_h, z_h) = 0 \quad \forall \delta u_h \in V_h,$$

$$(3.6b) \quad \alpha(q_h, \delta q_h - q_h) - a'_q(q_h, u_h)(\delta q_h - q_h, z_h) \geq 0 \quad \forall \delta q_h \in Q_{\text{ad},h},$$

$$(3.6c) \quad f(\delta z_h) - a(q_h, u_h)(\delta z_h) = 0 \quad \forall \delta z_h \in V_h.$$

The nonlinear primal dual active set strategy, described in the previous section, can be translated directly into the discrete level to solve (3.6a)–(3.6c).

In order to formulate the analog system to (2.18a)–(2.18f) we introduce discrete active sets  $\omega_{-,h}$  and  $\omega_{+,h}$  for a solution  $(q_h, u_h)$  to (3.2), (3.3) by

$$(3.7) \quad \omega_{-,h} = \{x \in \omega \mid q_h(x) = q_-\},$$

$$(3.8) \quad \omega_{+,h} = \{x \in \omega \mid q_h(x) = q_+\}$$

and define a Lagrange multiplier  $\mu_h \in Q_h$  via

$$(3.9) \quad (\mu_h, \delta q_h) = -\mathcal{L}'_q(q_h, u_h, z_h)(\delta q_h) \quad \forall \delta q_h \in Q_h.$$

Moreover, we introduce  $\mu_h^- \in Q_h$  and  $\mu_h^+ \in Q_h$  by

$$(3.10) \quad \mu_h^+ - \mu_h^- = \mu_h, \quad (\mu_h^-, \psi_i) \geq 0, \quad (\mu_h^+, \psi_i) \geq 0 \quad \forall \psi_i \in \mathcal{B}$$

by which  $\mu_h^\pm$  are uniquely determined if in addition the following complementarity conditions hold:

$$(3.11) \quad (\mu_h^-, q_h - q_-) = (\mu_h^+, q_+ - q_h) = 0.$$

*Remark 3.2.* This definition corresponds to the Lagrange multipliers obtained for the inequality constraints if the discrete optimization problem (3.2), (3.3) is considered a finite dimensional optimization problem for  $q_h = \sum_i q_i \psi_i \in Q_h$  with the following restrictions:

$$q_- \leq q_i \leq q_+ \quad \forall i.$$

Note that due to the choice of the basis  $\mathcal{B}$  in (3.1) this is equivalent to  $q_- \leq q_h(x) \leq q_+$  for all  $x \in \omega$ . Utilizing this fact, the discrete active sets  $\omega_{-,h}$ ,  $\omega_{+,h}$  are completely determined by the values of the coordinate vector of  $q_h$ . In particular they consist only of whole cells, edges, and nodes.

To obtain the complementarity conditions with respect to the  $Q = L^2(\omega)$ -inner product (3.11) one requires

$$(\mu_h^+, \psi_i) = 0 \text{ if } q_i < q_+ \text{ and } (\mu_h^-, \psi_i) = 0 \text{ if } q_i > q_-.$$

We now define the discretized versions of (2.14), (2.16), and (2.15) by

$$(3.12) \quad \mathcal{Y}_h = \mathcal{X}_h \times Q_h \times Q_h,$$

$$(3.13) \quad \mathcal{Y}_{\text{ad},h} = \mathcal{X}_{\text{ad},h} \times Q_{-,h} \times Q_{+,h},$$

$$(3.14) \quad \tilde{\mathcal{Y}}_{\text{ad},h} = \mathcal{X}_h \times Q_{-,h} \times Q_{+,h},$$

where

$$Q_{-,h} = \{r \in Q_h \mid r(x) = 0 \text{ a.e. on } \omega \setminus \omega_h^-\},$$

$$Q_{+,h} = \{r \in Q_h \mid r(x) = 0 \text{ a.e. on } \omega \setminus \omega_h^+\}.$$

A vector from these spaces will be abbreviated by  $\chi_h = (q_h, u_h, z_h, \mu_h^-, \mu_h^+)$ .

Using the definitions above we have the first-order necessary optimality condition for  $(q_h, u_h) \in Q_{\text{ad},h} \times V_h$ :

There exist  $z_h \in V_h$ ,  $\mu_h^- \in Q_{-,h}$ ,  $\mu_h^+ \in Q_{+,h}$  such that for  $\chi_h = (q_h, u_h, z_h, \mu_h^-, \mu_h^+) \in \mathcal{Y}_{\text{ad}}$  the following conditions hold:

$$(3.15a) \quad \tilde{\mathcal{L}}'_u(\chi_h)(\delta u) = 0 \quad \forall \delta u \in V_h,$$

$$(3.15b) \quad \tilde{\mathcal{L}}'_q(\chi_h)(\delta q) = 0 \quad \forall \delta q \in Q_h,$$

$$(3.15c) \quad \tilde{\mathcal{L}}'_z(\chi_h)(\delta z) = 0 \quad \forall \delta z \in V_h,$$

$$(3.15d) \quad \tilde{\mathcal{L}}'_{\mu^-}(\chi_h)(\delta \mu^-) = 0 \quad \forall \delta \mu^- \in Q_{-,h},$$

$$(3.15e) \quad \tilde{\mathcal{L}}'_{\mu^+}(\chi_h)(\delta \mu^+) = 0 \quad \forall \delta \mu^+ \in Q_{+,h},$$

$$(3.15f) \quad \mu_h^+ - \mu_h^- = \mu_h, \quad (\mu_h^-, \psi_i) \geq 0, \quad (\mu_h^+, \psi_i) \geq 0 \quad \forall \psi_i \in \mathcal{B}.$$

Here again (3.15d), (3.15e) are equivalent to the complementarity condition

$$(3.16) \quad (\mu_h^-, q_- - q_h) = (\mu_h^+, q_h - q_+) = 0.$$

Finally we state the following assumption concerning our discretization which is the analogue to Assumption 1.

*Assumption 5.* The optimization problem (3.2), (3.3) possesses a solution  $(q_h, u_h)$ . In addition there exists  $z_h \in V_h$  such the first-order necessary conditions (3.6) are fulfilled by the triple  $(q_h, u_h, z_h)$ .

**4. A posteriori error estimation.** The aim of this section is to derive a posteriori error estimates for the error with respect to the cost functional and to an arbitrary quantity of interest. These error estimates extend the results from [4, 6, 7, 8] to the case of optimization problems with control constraints. The provided estimators will be used within the following adaptive algorithm for error control and mesh refinement: We start on a coarse mesh, solve the discretized optimization problem, and evaluate the error estimator. Thereafter we refine the current mesh using local information obtained from the error estimator, allowing for efficient reduction of the discretization error with respect to the quantity of interest. This procedure is iterated until the value of the error estimator is below a given tolerance; see, e.g., [7] for a detailed description of this algorithm.

The section is structured as follows: First we will derive two a posteriori error estimators for the error with respect to the cost functional. The first one is based on the first-order necessary condition (2.8), which involves a variational inequality, and the second estimator uses the information obtained from the Lagrange multipliers for the inequality constraints. Both estimators can be evaluated in terms of the solution to the discretized optimization problem (3.2), (3.3). Then we will proceed with the error estimator with respect to an arbitrary quantity of interest, which requires the solution to an auxiliary linear-quadratic optimization problem. Even though the idea behind the estimators remains unchanged, the latter estimators require a more technical discussion.

Throughout this section we shall denote a solution to the optimization problem (2.4) by  $(q, u)$  and the corresponding solution to the optimality system (2.7) by  $\xi = (q, u, z) \in \mathcal{X}_{\text{ad}}$  and its discrete counterpart (3.6) by  $\xi_h = (q_h, u_h, z_h) \in \mathcal{X}_{\text{ad},h}$ . The corresponding solution to (2.18) and its discrete counterpart (3.15) will be abbreviated as  $\chi = (q, u, z, \mu^-, \mu^+) \in \mathcal{Y}_{\text{ad}}$  and  $\chi_h = (q_h, u_h, z_h, \mu_h^-, \mu_h^+) \in \mathcal{Y}_{\text{ad},h}$ .

**4.1. Error in the cost functional.** For the derivation of the error estimator with respect to the cost functional, we introduce the residual functionals  $\rho_u(\xi_h)(\cdot)$ ,  $\rho_z(\xi_h)(\cdot) \in V'$ , and  $\rho_q(\xi_h)(\cdot) \in Q'$  by

$$(4.1) \quad \rho_u(\xi_h)(\cdot) = f(\cdot) - a(q_h, u_h)(\cdot),$$

$$(4.2) \quad \rho_z(\xi_h)(\cdot) = J'_1(u_h)(\cdot) - a'_u(q_h, u_h)(\cdot, z_h),$$

$$(4.3) \quad \rho_q(\xi_h)(\cdot) = \alpha(q_h, \cdot) - a'_q(u_h, q_h)(\cdot, z_h).$$

The following theorem is an extension of the result from [6].

**THEOREM 4.1.** *Let  $\xi \in \mathcal{X}_{\text{ad}}$  be a solution to the first-order necessary system (2.7) and  $\xi_h \in \mathcal{X}_{\text{ad},h}$  be its Galerkin approximation (3.6). Then the following estimate holds:*

$$(4.4) \quad J(q, u) - J(q_h, u_h) \leq \frac{1}{2} \rho_u(\xi_h)(z - \tilde{z}_h) + \frac{1}{2} \rho_z(\xi_h)(u - \tilde{u}_h) + \frac{1}{2} \rho_q(\xi_h)(q - q_h) + R_1,$$

where  $\tilde{u}_h, \tilde{z}_h \in V_h$  are arbitrarily chosen and  $R_1$  is a remainder term given by

$$(4.5) \quad R_1 = \frac{1}{2} \int_0^1 \mathcal{L}'''(\xi_h + s(\xi - \xi_h))(\xi - \xi_h, \xi - \xi_h, \xi - \xi_h) s(s-1) ds.$$

*Proof.* From optimality system (2.7a)–(2.7c) we obtain that

$$J(q, u) = \mathcal{L}(\xi).$$

A similar equality holds on the discrete level. Therefore we have

$$J(q, u) - J(q_h, u_h) = \mathcal{L}(\xi) - \mathcal{L}(\xi_h) = \int_0^1 \mathcal{L}'(\xi_h + s(\xi - \xi_h))(\xi - \xi_h) ds.$$

We approximate this integral by the trapezoidal rule and obtain

$$(4.6) \quad J(q, u) - J(q_h, u_h) = \frac{1}{2} \mathcal{L}'(\xi)(\xi - \xi_h) + \frac{1}{2} \mathcal{L}'(\xi_h)(\xi - \xi_h) + R_1,$$

with the reminder term  $R_1$  as in (4.5). For the first term we have

$$\mathcal{L}'(\xi)(\xi - \xi_h) = \mathcal{L}'_u(\xi)(u - u_h) + \mathcal{L}'_z(\xi)(z - z_h) + \mathcal{L}'_q(\xi)(q - q_h).$$

Using optimality system (2.7a)–(2.7c) and the fact that  $q_h \in Q_{\text{ad},h} \subset Q_{\text{ad}}$ , we deduce that

$$\mathcal{L}'(\xi)(\xi - \xi_h) = -\mathcal{L}'_q(\xi)(q_h - q) \leq 0.$$

Rewriting the second term in (4.6) we obtain

$$\mathcal{L}'(\xi_h)(\xi - \xi_h) = \rho_u(\xi_h)(z - z_h) + \rho_z(\xi_h)(u - u_h) + \rho_q(\xi_h)(q - q_h).$$

Due to the Galerkin orthogonality for the state and adjoint equations, we have for arbitrary  $\tilde{u}_h, \tilde{z}_h \in V_h$

$$\rho_u(\xi_h)(z - z_h) = \rho_u(\xi_h)(z - \tilde{z}_h) \quad \text{and} \quad \rho_z(\xi_h)(u - u_h) = \rho_z(\xi_h)(u - \tilde{u}_h).$$

This completes the proof.  $\square$

*Remark 4.1.* We note that, in contrast to the terms involving the residuals of state and the adjoint equations, the error  $q - q_h$  in the term  $\rho_q(\xi_h)(q - q_h)$  in (4.4) cannot be replaced by  $q - \tilde{q}_h$  with an arbitrary  $\tilde{q}_h \in Q_{\text{ad},h}$ . This fact is caused by the control constraints. However, we may replace  $\rho_q(\xi_h)(q - q_h)$  by  $\rho_q(\xi_h)(q - q_h + \tilde{q}_h)$  with arbitrary  $\tilde{q}_h$  fulfilling  $\text{supp}(\tilde{q}_h) \subset \omega \setminus (\omega_{-,h} \cup \omega_{+,h})$  due to the structure of  $\rho_q(\xi_h)(\cdot)$ .

In order to use the estimate from the theorem above for computable error estimation we proceed as follows: First we choose  $\tilde{u}_h = i_h u$ ,  $\tilde{z}_h = i_h z$ , with an interpolation operator  $i_h: V \rightarrow V_h$ ; then we have to approximate the corresponding interpolation errors  $u - i_h u$  and  $z - i_h z$ . There are several heuristic techniques to do this; see, for instance, [6, 7]. Assume we have an operator  $\pi: V_h \rightarrow \tilde{V}_h$ , with  $\tilde{V}_h \neq V_h$ , such that  $u - \pi u_h$  has a better local asymptotical behavior as  $u - i_h u$ . Then we approximate

$$\rho_u(\xi_h)(z - i_h z) \approx \rho_u(\xi_h)(\pi z_h - z_h) \quad \text{and} \quad \rho_z(\xi_h)(u - i_h u) \approx \rho_z(\xi_h)(\pi u_h - u_h).$$

Such an operator can be constructed, for example, by the interpolation of the computed bilinear finite element solution in the space of biquadratic finite elements on patches of cells. For this operator the improved approximation property relies on local smoothness of  $u$  and superconvergence properties of the approximation  $u_h$ . The use of such “local higher-order approximation” is observed to work very successfully in the context of a posteriori error estimation; see, e.g., [6, 7].

The approximation of the term  $\rho_q(\xi_h)(q - q_h)$  requires more care. In contrast to the state  $u$  and the adjoint state  $z$ , the control variable  $q$  can generally not be approximated by “local higher-order approximation” for the following reasons:

- In the case of finite dimensional control space  $Q$ , there is no “patch-like” structure allowing for “local higher-order approximation.”

• If  $q$  is a distributed control, it typically does not possess sufficient smoothness (due to the inequality constraints) for the improved approximation property. We therefore suggest another approximation of  $\rho_q(\xi_h)(q - q_h)$  based on the projection formula (2.9). To this end we introduce  $\tilde{q} \in Q_{\text{ad}}$  by

$$(4.7) \quad \tilde{q} = \mathcal{P}_{Q_{\text{ad}}} \left( \frac{1}{\alpha} a'_q(q_h, \pi u_h)(\cdot, \pi z_h) \right).$$

In some cases one can show better approximation behavior of  $q - \tilde{q}$  in comparison with  $q - q_h$ ; see [31] and [22] for similar considerations in the context of a priori error analysis.

This construction results in the following computable a posteriori error estimator:

$$\eta_1 = \frac{1}{2} (\rho_u(\xi_h)(\pi z_h - z_h) + \rho_z(\xi_h)(\pi u_h - u_h) + \rho_q(\xi_h)(\tilde{q} - q_h)).$$

*Remark 4.2.* In order to use this error estimator as an indicator for mesh refinement, we have to localize it to cellwise or nodewise contributions. A direct localization of the terms like  $\rho_u(\xi_h)(\pi z_h - z_h)$  leads, in general, to the local contributions of wrong order (overestimation) due to oscillatory behavior of the residual terms. To overcome this, one may integrate the residual terms by part (see, e.g., [6]) or use a filtering operator; see [36] for details.

We should note that (4.4) does not provide an estimate for the absolute value of  $J(q, u) - J(q_h, u_h)$ , which is due to the inequality sign in (4.4). In the next section we will overcome this difficulty utilizing the alternative optimality system (2.18a)–(2.18f).

**4.2. Error in the cost functional reviewed.** In order to derive an error estimator for the absolute value of  $J(q, u) - J(q_h, u_h)$  we introduce the additional residual functionals  $\tilde{\rho}_q(\chi_h)(\cdot)$ ,  $\tilde{\rho}_{\mu^-}(\chi_h)(\cdot)$ ,  $\tilde{\rho}_{\mu^+}(\chi_h)(\cdot) \in Q'$  by

$$(4.8) \quad \tilde{\rho}_q(\chi_h)(\cdot) = \alpha(q_h, \cdot) - a'_q(q_h, u_h)(\cdot, z_h) + (\mu_h^+ - \mu_h^-, \cdot),$$

$$(4.9) \quad \tilde{\rho}_{\mu^-}(\chi_h)(\cdot) = (\cdot, q_- - q_h),$$

$$(4.10) \quad \tilde{\rho}_{\mu^+}(\chi_h)(\cdot) = (\cdot, q_h - q_+).$$

In what follows, the last two residual functional will also be evaluated in the point  $\chi$  where they read as follows:

$$\tilde{\rho}_{\mu^-}(\chi)(\cdot) = (\cdot, q_- - q), \quad \tilde{\rho}_{\mu^+}(\chi)(\cdot) = (\cdot, q - q_+).$$

Analogous to Theorem 4.1 we obtain the following theorem.

**THEOREM 4.2.** *Let  $\chi \in \mathcal{Y}_{\text{ad}}$  be a solution to the first-order necessary condition (2.18a)–(2.18f) and  $\chi_h \in \mathcal{Y}_{\text{ad},h}$  be its Galerkin approximation (3.15a)–(3.16). Then the following estimate holds:*

$$(4.11) \quad \begin{aligned} J(q, u) - J(q_h, u_h) = & \frac{1}{2} \rho_u(\chi_h)(z - \tilde{z}_h) + \frac{1}{2} \rho_z(\chi_h)(u - \tilde{u}_h) + \frac{1}{2} \tilde{\rho}_q(\chi_h)(q - \tilde{q}_h) \\ & + \frac{1}{2} \tilde{\rho}_{\mu^-}(\chi_h)(\mu^- - \tilde{\mu}_h^-) + \frac{1}{2} \tilde{\rho}_{\mu^+}(\chi_h)(\mu^+ - \tilde{\mu}_h^+) \\ & + \frac{1}{2} \tilde{\rho}_{\mu^-}(\chi)(\tilde{\mu}^- - \mu_h^-) + \frac{1}{2} \tilde{\rho}_{\mu^+}(\chi)(\tilde{\mu}^+ - \mu_h^+) + R_2, \end{aligned}$$

where  $\tilde{u}_h, \tilde{z}_h \in V_h$ ,  $\tilde{q}_h \in Q_h$ ,  $\tilde{\mu}_h^- \in Q_{-,h}$ ,  $\tilde{\mu}_h^+ \in Q_{+,h}$ ,  $\tilde{\mu}^- \in Q_-$ ,  $\tilde{\mu}^+ \in Q_+$  are arbitrarily chosen and  $R_2$  is a remainder term given by

$$(4.12) \quad R_2 = \frac{1}{2} \int_0^1 \tilde{\mathcal{L}}'''(\chi_h + s(\chi - \chi_h))(\chi - \chi_h, \chi - \chi_h, \chi - \chi_h) s(s-1) ds.$$

*Proof.* From (2.17) and optimality system (2.8a)–(2.8c) we obtain

$$J(q, u) = \mathcal{L}(\xi) = \tilde{\mathcal{L}}(\chi).$$

The analog result holds on the discrete level. We therefore have

$$J(q, u) - J(q_h, u_h) = \tilde{\mathcal{L}}(\chi) - \tilde{\mathcal{L}}(\chi_h) = \int_0^1 \tilde{\mathcal{L}}'(\chi_h + s(\chi - \chi_h))(\chi - \chi_h) ds.$$

As in the proof of Theorem 4.1 we approximate this integral by the trapezoidal rule and obtain

$$(4.13) \quad J(q, u) - J(q_h, u_h) = \frac{1}{2} \tilde{\mathcal{L}}'(\chi)(\chi - \chi_h) + \frac{1}{2} \tilde{\mathcal{L}}'(\chi_h)(\chi - \chi_h) + R_2,$$

with the remainder term  $R_2$  as in (4.12). For the first term we have

$$\begin{aligned} \tilde{\mathcal{L}}'(\chi)(\chi - \chi_h) &= \tilde{\mathcal{L}}'_u(\chi)(u - u_h) + \tilde{\mathcal{L}}'_z(\chi)(z - z_h) + \tilde{\mathcal{L}}'_q(\chi)(q - q_h) \\ &\quad + \tilde{\mathcal{L}}'_{\mu^-}(\chi)(\mu^- - \mu_h^-) + \tilde{\mathcal{L}}'_{\mu^+}(\chi)(\mu^+ - \mu_h^+). \end{aligned}$$

Using optimality system (2.18a)–(2.18f) we deduce that

$$\tilde{\mathcal{L}}'(\chi)(\chi - \chi_h) = \tilde{\mathcal{L}}'_{\mu^-}(\chi)(\mu^- - \mu_h^-) + \tilde{\mathcal{L}}'_{\mu^+}(\chi)(\mu^+ - \mu_h^+).$$

From (2.18d) and (2.18e) together with linearity of  $\tilde{\mathcal{L}}'_{\mu^-}(\chi)(\cdot)$  and  $\tilde{\mathcal{L}}'_{\mu^+}(\chi)(\cdot)$  we obtain that for arbitrary  $\tilde{\mu}^- \in Q_-$  and  $\tilde{\mu}^+ \in Q_+$

$$\tilde{\mathcal{L}}'_{\mu^-}(\chi)(\mu^- - \mu_h^-) = \tilde{\mathcal{L}}'_{\mu^-}(\chi)(\tilde{\mu}^- - \mu_h^-), \quad \tilde{\mathcal{L}}'_{\mu^+}(\chi)(\mu^+ - \mu_h^+) = \tilde{\mathcal{L}}'_{\mu^+}(\chi)(\tilde{\mu}^+ - \mu_h^+)$$

holds, and thus we obtain

$$\tilde{\mathcal{L}}'(\chi)(\chi - \chi_h) = \tilde{\rho}_{\mu^-}(\chi)(\tilde{\mu}^- - \mu_h^-) + \tilde{\rho}_{\mu^+}(\chi)(\tilde{\mu}^+ - \mu_h^+).$$

Rewriting the second term in (4.13) we obtain

$$\begin{aligned} \tilde{\mathcal{L}}'(\chi_h)(\chi - \chi_h) &= \rho_u(\chi_h)(u - u_h) + \rho_z(\chi_h)(z - z_h) + \tilde{\rho}_q(\chi_h)(q - q_h) \\ &\quad + \tilde{\rho}_{\mu^-}(\chi_h)(\mu^- - \mu_h^-) + \tilde{\rho}_{\mu^+}(\chi_h)(\mu^+ - \mu_h^+), \end{aligned}$$

where we can use linearity of the residual functionals in the second argument and (3.15a)–(3.15c) to obtain the following equalities:

$$(4.14) \quad \rho_u(\chi_h)(u - u_h) = \rho_u(\chi_h)(u - \tilde{u}_h),$$

$$(4.15) \quad \rho_z(\chi_h)(z - z_h) = \rho_z(\chi_h)(z - \tilde{z}_h),$$

$$(4.16) \quad \tilde{\rho}_q(\chi_h)(q - q_h) = \tilde{\rho}_q(\chi_h)(q - \tilde{q}_h)$$

for arbitrary  $\tilde{u}_h, \tilde{z}_h \in V_h$ ,  $\tilde{q}_h \in Q_h$ . Additionally we gain from (3.15d) and (3.15e) that for arbitrary  $\tilde{\mu}_h^- \in Q_{-,h}$  and  $\tilde{\mu}_h^+ \in Q_{+,h}$

$$(4.17) \quad \tilde{\rho}_{\mu^-}(\chi_h)(\mu^- - \mu_h^-) = \tilde{\rho}_{\mu^-}(\chi_h)(\mu^- - \tilde{\mu}_h^-),$$

$$(4.18) \quad \tilde{\rho}_{\mu^+}(\chi_h)(\mu^+ - \mu_h^+) = \tilde{\rho}_{\mu^+}(\chi_h)(\mu^+ - \tilde{\mu}_h^+)$$

holds. This completes the proof.  $\square$

To gain a computable error estimator we proceed as in the previous section. In order to deal with the new residual functionals we utilize (2.12) and construct an approximation for  $\mu$  by

$$(4.19) \quad \tilde{\mu} = -\alpha\tilde{q} + a'_q(\tilde{q}, \pi u_h)(\cdot, \pi z_h),$$

where  $\tilde{q}$  is given by (4.7). This leads to a computable a posteriori error estimator:

$$\begin{aligned} \eta_2 = & \frac{1}{2}(\rho_u(\chi_h)(\pi z_h - z_h) + \rho_z(\chi_h)(\pi u_h - u_h) + \tilde{\rho}_q(\chi_h)(\tilde{q} - q_h), \\ & \tilde{\rho}_{\mu^-}(\chi_h)(\tilde{\mu}^- - \mu_h^-) + \tilde{\rho}_{\mu^+}(\chi_h)(\tilde{\mu}^+ - \mu_h^+), \\ & \tilde{\rho}_{\mu^-}(\tilde{\chi})(\tilde{\mu}^- - \mu_h^-) + \tilde{\rho}_{\mu^+}(\tilde{\chi})(\tilde{\mu}^+ - \mu_h^+)). \end{aligned}$$

*Remark 4.3.* We note that the a posteriori error estimates derived in Theorems 4.1 and 4.2 coincide if the control constraints are inactive, e.g., if  $Q_{\text{ad}} = Q$ . Moreover, if the active sets are approximated from outside, i.e.,  $\omega_- \subset \omega_{-,h}$  and  $\omega_+ \subset \omega_{+,h}$ , these error estimators coincide as well.

**4.3. Error in the quantity of interest.** The aim of this section is the derivation of an error estimator for the error

$$(4.20) \quad I(q, u) - I(q_h, u_h)$$

with a given functional  $I : Q \times V \rightarrow \mathbb{R}$  describing the quantity of interest which we require to be three times directional differentiable. To this end we consider an additional Lagrangian  $\mathcal{M} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$  defined by

$$(4.21) \quad \mathcal{M}(\chi)(\psi) = I(q, u) + \tilde{\mathcal{L}}'(\chi)(\psi),$$

where we abbreviate  $\chi = (q, u, z, \mu^-, \mu^+)$  and  $\psi = (p, v, y, \nu^-, \nu^+)$ . Here  $(p, v, y, \nu^-, \nu^+)$  will be variables dual to  $(q, u, z, \mu^-, \mu^+)$ . Note that for the solution  $\chi$  to the optimality system (2.18a)–(2.18f) of the optimization problem (2.4) the identity

$$(4.22) \quad \mathcal{M}(\chi)(\psi) = I(q, u)$$

holds for all  $\psi \in \tilde{\mathcal{Y}}_{\text{ad}}$ . To proceed as in the proof of Theorem 4.2 it remains to find  $\psi \in \tilde{\mathcal{Y}}_{\text{ad}}$  such that  $(\chi, \psi)$  is a stationary point of  $\mathcal{M}$  on  $\tilde{\mathcal{Y}}_{\text{ad}} \times \tilde{\mathcal{Y}}_{\text{ad}}$ .

Therefore we consider the auxiliary (linear-quadratic) optimization problem

$$(4.23) \quad \text{Minimize } K(\chi, p, v), \quad p \in P_{\text{ad}}, v \in V,$$

$$(4.24) \quad \text{subject to } \tilde{\mathcal{L}}''_{uz}(\chi)(v, \varphi) + \tilde{\mathcal{L}}''_{qz}(\chi)(p, \varphi) = 0 \quad \forall \varphi \in V$$

for given  $\chi \in \mathcal{Y}$ . The admissible set  $P_{\text{ad}}$  is given as

$$(4.25) \quad P_{\text{ad}} = \{p \in Q \mid p_-(x) \leq p(x) \leq p_+(x) \text{ a.e. on } \omega\},$$

with the bounds

$$\begin{aligned} p_-(x) &= \begin{cases} 0, & \mu(x) \neq 0 \text{ or } q(x) = q_-(x), \\ -\infty & \text{else,} \end{cases} \\ p_+(x) &= \begin{cases} 0, & \mu(x) \neq 0 \text{ or } q(x) = q_+(x), \\ +\infty & \text{else,} \end{cases} \end{aligned}$$



and the cost functional  $K : \mathcal{Y} \times Q \times V \rightarrow \mathbb{R}$  is defined via

$$(4.26) \quad K(\chi, p, v) = I'_u(q, u)(v) + I'_q(q, u)(p) + \tilde{\mathcal{L}}''_{uq}(\chi)(v, p) + \frac{1}{2}\tilde{\mathcal{L}}''_{uu}(\chi)(v, v) + \frac{1}{2}\tilde{\mathcal{L}}''_{qq}(\chi)(p, p).$$

We introduce the following abbreviation for later use:

$$(4.27) \quad \bar{\mathcal{Y}}_{\text{ad}} = P_{\text{ad}} \times V \times V \times Q_- \times Q_+.$$

*Remark 4.4.* Consideration of the auxiliary optimization problem (4.23), (4.24) is motivated by the unconstrained case  $Q_{\text{ad}} = Q$ . There the stationary point of  $\mathcal{M}$  is given as the solution to (4.23), (4.24) with  $P_{\text{ad}} = Q$ . A similar linear-quadratic optimization problem is considered in [19] in the context of sensitivity analysis.

*Remark 4.5.* If we assume that the second-order sufficient condition from Lemma 2.1 holds, the linear-quadratic optimization problem (4.23) possesses a solution. This is the case, as the quadratic part  $\tilde{\mathcal{L}}''_{uq}(\chi)(v, p) + \frac{1}{2}\tilde{\mathcal{L}}''_{uu}(\chi)(v, v) + \frac{1}{2}\tilde{\mathcal{L}}''_{qq}(\chi)(p, p)$  of  $K(p, v)$  is positive definite (see (2.20)) for all solutions to the linear equation (2.21), which is exactly the same as (4.24).

We introduce an auxiliary Lagrangian  $\mathcal{N} : \mathcal{Y} \times \mathcal{X} \rightarrow \mathbb{R}$  for (4.23), (4.24) by

$$(4.28) \quad \mathcal{N}(\chi, p, v, y) = K(\chi, p, v) + \tilde{\mathcal{L}}''_{uz}(\chi)(v, y) + \tilde{\mathcal{L}}''_{qz}(\chi)(p, y).$$

For a solution  $(p, v)$  to (4.23), (4.24) the following first-order necessary condition holds:

There exists  $y \in V$  such that

$$(4.29a) \quad \mathcal{N}'_y(\chi, p, v, y)(\delta y) = 0 \quad \forall \delta y \in V,$$

$$(4.29b) \quad \mathcal{N}'_v(\chi, p, v, y)(\delta v) = 0 \quad \forall \delta v \in V,$$

$$(4.29c) \quad \mathcal{N}'_p(\chi, p, v, y)(\delta p - p) \geq 0 \quad \forall \delta p \in P_{\text{ad}}$$

or, if written more explicitly,

$$(4.30a) \quad \tilde{\mathcal{L}}''_{uz}(\chi)(v, \delta y) + \tilde{\mathcal{L}}''_{qz}(\chi)(p, \delta y) = 0 \quad \forall \delta y \in V,$$

$$(4.30b) \quad I'_u(q, u)(\delta v) + \tilde{\mathcal{L}}''_{uq}(\chi)(\delta v, p) + \tilde{\mathcal{L}}''_{uu}(\chi)(\delta v, v) + \tilde{\mathcal{L}}''_{uz}(\chi)(\delta v, y) = 0 \quad \forall \delta v \in V,$$

$$(4.30c) \quad I'_q(q, u)(\delta p) + \tilde{\mathcal{L}}''_{uq}(\chi)(v, \delta p) + \tilde{\mathcal{L}}''_{qq}(\chi)(\delta p, p) + \tilde{\mathcal{L}}''_{qz}(\chi)(\delta p, y) \geq 0 \quad \forall \delta p \in P_{\text{ad}} - p.$$

Again we can introduce the full Lagrangian  $\tilde{\mathcal{N}} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$  by

$$(4.31) \quad \tilde{\mathcal{N}}(\chi, \psi) = \mathcal{N}(\chi, p, v, y) + (\nu^-, p_- - p) + (\nu^+, p - p_+).$$

As in (2.18a)–(2.18f) we can rewrite the necessary optimality condition for  $\psi \in \bar{\mathcal{Y}}_{\text{ad}}$  as

$$(4.32a) \quad \tilde{\mathcal{N}}'_v(\chi, \psi)(\delta v) = 0 \quad \forall \delta v \in V,$$

$$(4.32b) \quad \tilde{\mathcal{N}}'_p(\chi, \psi)(\delta p) = 0 \quad \forall \delta p \in Q,$$

$$(4.32c) \quad \tilde{\mathcal{N}}'_y(\chi, \psi)(\delta y) = 0 \quad \forall \delta y \in V,$$

$$(4.32d) \quad \tilde{\mathcal{N}}'_{\nu^-}(\chi, \psi)(\delta \nu^-) = 0 \quad \forall \delta \nu^- \in Q_-,$$

$$(4.32e) \quad \tilde{\mathcal{N}}'_{\nu^+}(\chi, \psi)(\delta \nu^+) = 0 \quad \forall \delta \nu^+ \in Q_+,$$

$$(4.32f) \quad \nu^+ - \nu^- = \nu, \quad \nu^-(p_- - p) = \nu^+(p - p_+) = 0 \quad \text{a.e. on } \omega,$$

$$(4.32g) \quad \text{supp } \nu^+ \subseteq \omega \setminus \{x \in \omega \mid q = q_- \text{ and } \mu \neq 0\}, \quad \nu^+ \geq 0, \quad \text{a.e. where } \mu = 0,$$

$$(4.32h) \quad \text{supp } \nu^- \subseteq \omega \setminus \{x \in \omega \mid q = q_+ \text{ and } \mu \neq 0\}, \quad \nu^- \geq 0, \quad \text{a.e. where } \mu = 0,$$

where  $\nu^-$  and  $\nu^+$  are given by the following relations depending on  $\nu = -\mathcal{N}'_p(\chi, p, v, y)(\cdot)$ :

$$\nu^+(x) = \begin{cases} \nu, & q(x) = q_+ \text{ and } \mu(x) \neq 0, \\ 0, & q(x) = q_- \text{ and } \mu(x) \neq 0, \\ \max(0, \nu) & \text{else,} \end{cases}$$

$$\nu^-(x) = \begin{cases} \nu, & q(x) = q_- \text{ and } \mu(x) \neq 0, \\ 0, & q(x) = q_+ \text{ and } \mu(x) \neq 0, \\ \max(0, -\nu) & \text{else.} \end{cases}$$

Note that due to the choice of  $p_-$  and  $p_+$  the Lagrange multipliers are contained in the desired spaces, e.g.,  $\nu^- \in Q_-$  and  $\nu^+ \in Q_+$ .

*Remark 4.6.* It should be noted that we use the convention  $\pm\infty \cdot 0 = 0$  in (4.31), (4.32f) to ease notation. The same convention will be used throughout this section.

*Remark 4.7.* The condition (4.32g) arises naturally, as  $\nu^+$  is the Lagrange multiplier which corresponds to the equality and inequality constraints for  $p$  that are induced by the active upper control bound  $q_+$ . Similarly (4.32h) arises from the active lower control bound  $q_-$ .

We introduce

$$(4.33) \quad \bar{\mathcal{Y}}_{\text{ad},h} = P_{\text{ad},h} \times V_h \times V_h \times Q_{-,h} \times Q_{+,h}$$

to shorten notation. This is discretized using the discretized admissible set

$$(4.34) \quad P_{\text{ad},h} = \{p \in Q_h \mid p_{h,-}(x) \leq p(x) \leq p_{h,+}(x) \text{ a.e. on } \omega\},$$

with the bounds

$$p_{h,-}(x) = \begin{cases} 0, & \mu_h(x) \neq 0 \text{ or } q_h(x) = q_-(x), \\ -\infty & \text{else,} \end{cases}$$

$$p_{h,+}(x) = \begin{cases} 0, & \mu_h(x) \neq 0 \text{ or } q_h(x) = q_+(x), \\ +\infty & \text{else.} \end{cases}$$

Then the following first-order condition holds with the discretized full Lagrangian:

$$\tilde{\mathcal{N}}_h(\chi, \psi) = \mathcal{N}(\chi, p, v, y) + (\nu^-, p_{h,-} - p) + (\nu^+, p - p_{h,+}),$$

where  $\tilde{\mathcal{N}}_h : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ . There exist  $y_h \in V_h$ ,  $\nu_h^+, \nu_h^- \in Q_h$  such that for  $\psi_h = (p_h, v_h, y_h, \nu_h^-, \nu_h^+) \in \bar{\mathcal{Y}}_{ad,h}$  the following hold:

$$(4.35a) \quad \tilde{\mathcal{N}}'_{h,v}(\chi_h, \psi_h)(\delta v) = 0 \quad \forall \delta v \in V_h,$$

$$(4.35b) \quad \tilde{\mathcal{N}}'_{h,p}(\chi_h, \psi_h)(\delta p) = 0 \quad \forall \delta p \in Q_h,$$

$$(4.35c) \quad \tilde{\mathcal{N}}'_{h,y}(\chi_h, \psi_h)(\delta y) = 0 \quad \forall \delta y \in V_h,$$

$$(4.35d) \quad \tilde{\mathcal{N}}'_{h,\nu^-}(\chi_h, \psi_h)(\delta \nu^-) = 0 \quad \forall \delta \nu^- \in Q_{-,h},$$

$$(4.35e) \quad \tilde{\mathcal{N}}'_{h,\nu^+}(\chi_h, \psi_h)(\delta \nu^+) = 0 \quad \forall \delta \nu^+ \in Q_{+,h},$$

$$(4.35f) \quad \nu_h^+ - \nu_h^- = \nu_h \quad (\nu_h^-, p_{h,-} - p_h) = (\nu_h^+, p_h - p_{h,+}) = 0,$$

$$(4.35g) \quad (\nu_h^+, \psi_i) = 0 \quad \forall i : (\mu_h, \psi_i) \neq 0 \text{ and } q_i = q_-,$$

$$(4.35h) \quad (\nu_h^+, \psi_i) \geq 0 \quad \forall i : (\mu_h, \psi_i) = 0,$$

$$(4.35i) \quad (\nu_h^-, \psi_i) = 0 \quad \forall i : (\mu_h, \psi_i) \neq 0 \text{ and } q_i = q_+,$$

$$(4.35j) \quad (\nu_h^-, \psi_i) \geq 0 \quad \forall i : (\mu_h, \psi_i) = 0.$$

For the error estimator with respect to the quantity of interest we introduce the residual functionals  $\tilde{\rho}_v(\chi_h, \psi_h)(\cdot)$ ,  $\tilde{\rho}_y(\chi_h, \psi_h)(\cdot) \in V'$  and  $\tilde{\rho}_p(\chi_h, \psi_h)(\cdot)$ ,  $\tilde{\rho}_{\nu^-}(\chi_h, \psi_h)(\cdot)$ ,  $\tilde{\rho}_{\nu^+}(\chi_h, \psi_h)(\cdot) \in Q'$  by

$$(4.36) \quad \tilde{\rho}_v(\chi_h, \psi_h)(\cdot) = \tilde{\mathcal{L}}''_{zu}(\chi_h)(\cdot, v_h) + \tilde{\mathcal{L}}''_{zq}(\chi_h)(\cdot, p_h),$$

$$(4.37) \quad \begin{aligned} \tilde{\rho}_y(\chi_h, \psi_h)(\cdot) &= I'_u(q_h, u_h)(\cdot) + \tilde{\mathcal{L}}''_{uu}(\chi_h)(\cdot, v_h) + \tilde{\mathcal{L}}''_{uz}(\chi_h)(\cdot, y_h) \\ &\quad + \tilde{\mathcal{L}}''_{uq}(\chi_h)(\cdot, p_h), \end{aligned}$$

$$(4.38) \quad \begin{aligned} \tilde{\rho}_p(\chi_h, \psi_h)(\cdot) &= I'_q(q_h, u_h)(\cdot) + \tilde{\mathcal{L}}''_{qu}(\chi_h)(\cdot, v_h) + \tilde{\mathcal{L}}''_{qz}(\chi_h)(\cdot, y_h) \\ &\quad + \tilde{\mathcal{L}}''_{qq}(\chi_h)(\cdot, p_h) + (\cdot, \nu_h), \end{aligned}$$

$$(4.39) \quad \tilde{\rho}_{\nu^-}(\chi_h, \psi_h)(\cdot) = -(\cdot, p_h),$$

$$(4.40) \quad \tilde{\rho}_{\nu^+}(\chi_h, \psi_h)(\cdot) = (\cdot, p_h),$$

in addition to the already defined residual functionals (4.1)–(4.10). Again the last two residual functionals also have to be evaluated in the point  $(\chi, \psi)$  where they read as follows:

$$\tilde{\rho}_{\nu^-}(\chi, \psi)(\cdot) = -(\cdot, p), \quad \tilde{\rho}_{\nu^+}(\chi, \psi)(\cdot) = (\cdot, p).$$

**THEOREM 4.3.** *Let  $\chi \in \mathcal{Y}_{ad}$  be a solution to the necessary optimality condition (2.18) and  $\chi_h \in \mathcal{Y}_{ad,h}$  be its Galerkin approximation (3.15). In addition let  $\psi \in \bar{\mathcal{Y}}_{ad}$  be a solution to the necessary optimality condition (4.32) of the auxiliary optimization problem (4.23), (4.24) and  $\psi_h \in \bar{\mathcal{Y}}_{ad,h}$  be its discrete approximation (4.35). Then the*

following estimate holds:

(4.41)

$$\begin{aligned}
 I(q, u) - I(q_h, u_h) = & \frac{1}{2} \rho_u(\chi_h)(y - \tilde{y}_h) + \frac{1}{2} \rho_z(\chi_h)(v - \tilde{v}_h) + \frac{1}{2} \tilde{\rho}_q(\chi_h)(p - \tilde{p}_h) \\
 & + \frac{1}{2} \tilde{\rho}_{\mu^-}(\chi_h)(\nu^- - \tilde{\nu}_h^-) + \frac{1}{2} \tilde{\rho}_{\mu^+}(\chi_h)(\nu^+ - \tilde{\nu}_h^+) \\
 & + \frac{1}{2} \tilde{\rho}_v(\chi_h, \psi_h)(z - \tilde{z}_h) + \frac{1}{2} \tilde{\rho}_y(\chi_h, \psi_h)(u - \tilde{u}_h) + \frac{1}{2} \tilde{\rho}_p(\chi_h, \psi_h)(q - \tilde{q}_h) \\
 & + \frac{1}{2} \tilde{\rho}_{\nu^-}(\chi_h, \psi_h)(\mu^- - \tilde{\mu}_h^-) + \frac{1}{2} \tilde{\rho}_{\nu^+}(\chi_h, \psi_h)(\mu^+ - \tilde{\mu}_h^+) \\
 & + \frac{1}{2} \tilde{\rho}_{\mu^-}(\chi)(\tilde{\nu}^- - \nu_h^-) + \frac{1}{2} \tilde{\rho}_{\mu^+}(\chi)(\tilde{\nu}^+ - \nu_h^+) \\
 & + \frac{1}{2} \tilde{\rho}_{\nu^-}(\chi, \psi)(\tilde{\mu}^- - \mu_h^-) + \frac{1}{2} \tilde{\rho}_{\nu^+}(\chi, \psi)(\tilde{\mu}^+ - \mu_h^+) + R_3,
 \end{aligned}$$

where  $\tilde{u}_h, \tilde{v}_h, \tilde{z}_h, \tilde{y}_h \in V_h, \tilde{q}_h, \tilde{p}_h \in Q_h, \tilde{\mu}_h^-, \tilde{\nu}_h^- \in Q_{-,h}, \tilde{\mu}_h^+, \tilde{\nu}_h^+ \in Q_{+,h}$  as well as  $\tilde{\mu}^-, \tilde{\nu}^- \in Q_-, \tilde{\mu}^+, \tilde{\nu}^+ \in Q_+$  are arbitrarily chosen and  $R_3$  is a remainder term given by

$$(4.42) \quad R_3 = \frac{1}{2} \int_0^1 \mathcal{M}'''((\chi_h, \psi_h) + se)(e, e, e)s(s-1) ds,$$

with  $e = (\chi - \chi_h, \psi - \psi_h)$ .

*Proof.* From (4.22) and the analog discrete result we obtain

$$I(q, u) - I(q_h, u_h) = \mathcal{M}(\chi, \psi) - \mathcal{M}(\chi_h, \psi_h) = \int_0^1 \mathcal{M}'((\chi_h, \psi_h) + se)(e) ds.$$

Approximation by the trapezoidal rule gives

$$(4.43) \quad I(q, u) - I(q_h, u_h) = \frac{1}{2} \mathcal{M}'(\chi, \psi)(e) + \frac{1}{2} \mathcal{M}'(\chi_h, \psi_h)(e) + R_3,$$

with the remainder term  $R_3$  as in (4.42). For the first term we have

$$\begin{aligned}
 \mathcal{M}'(\chi, \psi)(e) = & \mathcal{M}'_u(\chi, \psi)(u - u_h) + \mathcal{M}'_v(\chi, \psi)(v - v_h) \\
 & + \mathcal{M}'_z(\chi, \psi)(z - z_h) + \mathcal{M}'_y(\chi, \psi)(y - y_h) \\
 & + \mathcal{M}'_q(\chi, \psi)(q - q_h) + \mathcal{M}'_p(\chi, \psi)(p - p_h) \\
 & + \mathcal{M}'_{\mu^-}(\chi, \psi)(\mu^- - \mu_h^-) + \mathcal{M}'_{\nu^-}(\chi, \psi)(\nu^- - \nu_h^-) \\
 & + \mathcal{M}'_{\mu^+}(\chi, \psi)(\mu^+ - \mu_h^+) + \mathcal{M}'_{\nu^+}(\chi, \psi)(\nu^+ - \nu_h^+).
 \end{aligned}$$

Using the identities

$$\begin{aligned}
 \mathcal{M}'_u(\chi, \psi)(\cdot) &= \tilde{\mathcal{N}}'_v(\chi, p, v, y)(\cdot), & \mathcal{M}'_v(\chi, \psi)(\cdot) &= \tilde{\mathcal{L}}'_u(\chi)(\cdot), \\
 \mathcal{M}'_z(\chi, \psi)(\cdot) &= \tilde{\mathcal{N}}'_y(\chi, p, v, y)(\cdot), & \mathcal{M}'_y(\chi, \psi)(\cdot) &= \tilde{\mathcal{L}}'_z(\chi)(\cdot), \\
 \mathcal{M}'_q(\chi, \psi)(\cdot) &= \tilde{\mathcal{N}}'_p(\chi, p, v, y)(\cdot), & \mathcal{M}'_p(\chi, \psi)(\cdot) &= \tilde{\mathcal{L}}'_q(\chi)(\cdot),
 \end{aligned}$$

we see that the first six terms on the right-hand side vanish due to (2.18a)–(2.18c) and (4.32a)–(4.32c). Furthermore we see from (2.18d), (2.18e) and (4.32d), (4.32e)

that with arbitrary  $\tilde{\mu}^-, \tilde{\nu}^- \in Q_-$  and  $\tilde{\mu}^+, \tilde{\nu}^+ \in Q_+$  the following identities hold:

$$(4.44) \quad \mathcal{M}'_{\mu^-}(\chi, \psi)(\mu^- - \mu_h^-) = \mathcal{M}'_{\mu^-}(\chi, \psi)(\tilde{\mu}^- - \mu_h^-) = \tilde{\rho}_{\nu^-}(\chi, \psi)(\tilde{\mu}^- - \mu_h^-),$$

$$(4.45) \quad \mathcal{M}'_{\nu^-}(\chi, \psi)(\nu^- - \nu_h^-) = \mathcal{M}'_{\nu^-}(\chi, \psi)(\tilde{\nu}^- - \nu_h^-) = \tilde{\rho}_{\mu^-}(\chi)(\tilde{\nu}^- - \nu_h^-),$$

$$(4.46) \quad \mathcal{M}'_{\mu^+}(\chi, \psi)(\mu^+ - \mu_h^+) = \mathcal{M}'_{\mu^+}(\chi, \psi)(\tilde{\mu}^+ - \mu_h^+) = \tilde{\rho}_{\nu^+}(\chi, \psi)(\tilde{\mu}^+ - \mu_h^+),$$

$$(4.47) \quad \mathcal{M}'_{\nu^+}(\chi, \psi)(\nu^+ - \nu_h^+) = \mathcal{M}'_{\nu^+}(\chi, \psi)(\tilde{\nu}^+ - \nu_h^+) = \tilde{\rho}_{\mu^+}(\chi)(\tilde{\nu}^+ - \nu_h^+).$$

Thus we obtain

$$\begin{aligned} \mathcal{M}'(\chi, \psi)(e) &= \tilde{\rho}_{\mu^-}(\chi)(\tilde{\nu}^- - \nu_h^-) + \tilde{\rho}_{\mu^+}(\chi)(\tilde{\nu}^+ - \nu_h^+) \\ &\quad + \tilde{\rho}_{\nu^-}(\chi, \psi)(\tilde{\mu}^- - \mu_h^-) + \tilde{\rho}_{\nu^+}(\chi, \psi)(\tilde{\mu}^+ - \mu_h^+). \end{aligned}$$

For the second term we obtain from (3.15a)–(3.15e) and (4.32a)–(4.32e) that

$$\mathcal{M}'(\chi_h, \psi_h)(e) = \mathcal{M}'(\chi_h, \psi_h)(\chi - \tilde{\chi}_h, \psi - \tilde{\psi}_h)$$

for each  $\tilde{\chi}_h, \tilde{\psi}_h \in \tilde{\mathcal{Y}}_{\text{ad},h}$ , which completes the proof.  $\square$

*Remark 4.8.* Note that in the case  $I = J$  the solution  $(p, v, y)$  to (4.29) is given by  $(0, 0, z)$ , which can be seen after some calculations. Using this, one obtains that for  $I = J$  the estimates in Theorems 4.2 and 4.3 coincide.

We define the projection onto the admissible set by

$$\mathcal{P}_{P_{\text{ad},h}}(p) = \max(p_{h,-}, \min(p, p_{h,+})).$$

To obtain a computable error estimator we introduce  $\tilde{p} \in P_{\text{ad}}$  as an approximation to  $p$  by

$$(4.48) \quad \tilde{p} = \mathcal{P}_{P_{\text{ad},h}} \left( \frac{1}{\alpha} (a'_q(\cdot, \pi y_h) + a''_{qu}(\cdot, \pi v_h, \pi z_h) + a''_{qq}(\cdot, p_h, \pi z_h) - I'_q(\tilde{q}, \pi u_h)(\cdot)) \right),$$

where  $()$  is an abbreviation for  $(\tilde{q}, \pi u_h)$ , and  $\tilde{\nu}$  is introduced as an approximation to  $\nu$  by

$$(4.49) \quad \tilde{\nu} = -\alpha \tilde{p} + a'_q(\cdot, \pi y_h) + a''_{qu}(\cdot, \pi v_h, \pi z_h) + a''_{qq}(\cdot, p_h, \pi z_h) - I'_q(\tilde{q}, \pi u_h)(\cdot),$$

which is an analogue to the construction of the approximations  $\tilde{q}$  and  $\tilde{\mu}$  in (4.7) and (4.19).

Using these approximations we obtain the following computable error estimator:

$$\begin{aligned} \eta_{\text{QI}} &= \frac{1}{2} \rho_u(\chi_h)(\pi y - y_h) + \frac{1}{2} \rho_z(\chi_h)(\pi v - v_h) + \frac{1}{2} \tilde{\rho}_q(\chi_h)(\tilde{p} - p_h) \\ &\quad + \frac{1}{2} \tilde{\rho}_{\mu^-}(\chi_h)(\tilde{\nu}^- - \nu_h^-) + \frac{1}{2} \tilde{\rho}_{\mu^+}(\chi_h)(\tilde{\nu}^+ - \nu_h^+) \\ &\quad + \frac{1}{2} \tilde{\rho}_v(\chi_h, \psi_h)(\pi z - z_h) + \frac{1}{2} \tilde{\rho}_y(\chi_h, \psi_h)(\pi u - u_h) + \frac{1}{2} \tilde{\rho}_p(\chi_h, \psi_h)(\tilde{q} - q_h) \\ &\quad + \frac{1}{2} \tilde{\rho}_{\nu^-}(\chi_h, \psi_h)(\tilde{\mu}^- - \mu_h^-) + \frac{1}{2} \tilde{\rho}_{\nu^+}(\chi_h, \psi_h)(\tilde{\mu}^+ - \mu_h^+) \\ &\quad + \frac{1}{2} \tilde{\rho}_{\mu^-}(\tilde{\chi})(\tilde{\nu}^- - \nu_h^-) + \frac{1}{2} \tilde{\rho}_{\mu^+}(\tilde{\chi})(\tilde{\nu}^+ - \nu_h^+) \\ &\quad + \frac{1}{2} \tilde{\rho}_{\nu^-}(\tilde{\chi}, \tilde{\psi})(\tilde{\mu}^- - \mu_h^-) + \frac{1}{2} \tilde{\rho}_{\nu^+}(\tilde{\chi}, \tilde{\psi})(\tilde{\mu}^+ - \mu_h^+), \end{aligned}$$

where  $\tilde{\chi} = (\tilde{q}, \pi u, \pi z, \tilde{\mu}^-, \tilde{\mu}^+)$  and  $\tilde{\psi} = (\tilde{p}, \pi v, \pi y, \tilde{\nu}^-, \tilde{\nu}^+)$ .

*Remark 4.9.* We would like to point out that in case of strict complementarity, e.g., if the set

$$\{x \in \omega \mid q(x) = q_-(x) \text{ or } q(x) = q_+(x)\} \setminus \{x \in \omega \mid \mu(x) \neq 0\}$$

has zero measure, the auxiliary problem (4.23), (4.24) does not involve inequality constraints for the controls. In that case the set  $P_{\text{ad}}$  is not only convex but in fact a real subspace of  $Q$ .

*Remark 4.10.* The constrained linear-quadratic optimization problem (4.23), (4.24) can be solved using primal-dual active set strategy. In the case of strict complementarity the algorithm will converge in one step due to the fact that  $P_{\text{ad}}$  is a linear subspace of  $Q$  in this case.

*Remark 4.11.* Due to the definition of  $P_{\text{ad}}$  (4.25), the solution  $p \in Q$  of auxiliary optimization problem (4.23)–(4.24) is usually discontinuous. Therefore, a cellwise constant discretization of the control space  $Q$  seems to be more suitable than a discretization with continuous trial functions if the error with respect to a quantity of interest is estimated.

**5. Numerical examples.** In this section we discuss two numerical examples illustrating the behavior of our method. For both examples we use bilinear ( $H^1$ -conforming) finite elements for the discretization of the state variable and cellwise constant discretization of the control space. The optimization problems are solved by primal-dual active set strategy as sketched in section 2, where the equality-constrained problems in the inner loop are solved using Newton's method for the reduced cost functional.

All examples have been computed using the optimization library RoDoBo [5] and the finite element toolkit Gascoigne [3].

**5.1. Example 1.** We consider the following nonlinear optimization problem:

$$(5.1) \quad \text{Minimize} \quad \frac{1}{2} \|u - u^d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|q\|_{L^2(\Omega)}^2, \quad u \in V, \quad q \in Q_{\text{ad}},$$

subject to

$$(5.2) \quad \begin{aligned} -\Delta u + 30u^3 + u &= f + q && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where  $\Omega = \omega = (0, 1)^2 \setminus [0.4, 0.6]^2$ ,  $V = H_0^1(\Omega)$ ,  $Q = L^2(\Omega)$ , and the admissible set  $Q_{\text{ad}}$  is given by

$$Q_{\text{ad}} = \{q \in Q \mid -7 \leq q(x) \leq 20 \text{ a.e. on } \Omega\}.$$

The desired state  $u^d$  and the right-hand side  $f$  are defined as

$$u^d(x) = x_1 \cdot x_2, \quad f(x) = ((x_1 - 0.5)^2 + (x_2 - 0.5)^2)^{-1},$$

and the regularization parameter is chosen as  $\alpha = 10^{-4}$ . We note that the state equation (5.2) is a monotone semilinear equation, which possesses a unique solution  $u \in V$  for each  $q \in Q$ . The proof of the existence of a global solution as well as derivation of necessary and sufficient optimality conditions for the corresponding optimization problem (5.1)–(5.2) can be found, e.g., in [34].

In section 4 we derived two different error estimators for the error with respect to the cost functional and one error estimator with respect to a quantity of interest. In this example, we choose the quantity of interest as

$$(5.3) \quad I(q, u) = \frac{1}{2} \int_{(0.7, 0.8)^2} |\nabla u(x)|^2 dx + \int_{(0.2, 0.3)^2} q(x) dx.$$

In order to check the quality of the error estimators, we define the following effectivity indices:

$$(5.4) \quad I_{\text{eff}}(\eta_1) = \frac{J(u) - J(u_h)}{\eta_1}, \quad I_{\text{eff}}(\eta_2) = \frac{J(u) - J(u_h)}{\eta_2}, \quad I_{\text{eff}}(\eta_{\text{QI}}) = \frac{I(q, u) - I(q_h, u_h)}{\eta_{\text{QI}}}.$$

In Table 5.1 these effectivity indices are listed for different types of mesh refinement: random refinement and refinement based on the error estimator  $\eta_{\text{QI}}$  for the quantity of interest.

TABLE 5.1  
*Effectivity indices.*

N	$I_{\text{eff}}(\eta_1)$	$I_{\text{eff}}(\eta_2)$	$I_{\text{eff}}(\eta_{\text{QI}})$	N	$I_{\text{eff}}(\eta_1)$	$I_{\text{eff}}(\eta_2)$	$I_{\text{eff}}(\eta_{\text{QI}})$
432	1.1	1.1	1.2	432	1.1	1.1	1.1
906	1.1	1.1	1.1	824	1.1	1.1	1.4
2328	1.3	1.2	2.3	1692	1.0	1.0	0.3
5752	1.2	1.2	1.4	3992	1.0	1.0	0.2
13872	1.3	1.3	1.5	11396	1.0	1.0	0.5
33964	1.3	1.3	1.4	30604	1.0	1.0	1.0
83832	1.2	1.2	1.5	80354	1.0	1.0	1.3

(a) Random refinement

(b) Refinement according to  $\eta_{\text{QI}}$

We observe that the error estimators provide quantitative information about the discretization error. We note that the results for  $\eta_1$  and  $\eta_2$  are very close to each other in this example; cf. Remark 4.3.

In addition, our results show that the local mesh refinement based on error estimators derived above leads to substantial saving in degrees of freedom for achieving a given level of the discretization error. In Figure 5.1 the dependence of discretization error on the number of degrees of freedom is shown for different refinement criteria: global (uniform) refinement, refinement based on the error estimator  $\eta_1$  for the cost functional, and refinement based on the error estimator  $\eta_{\text{QI}}$  for the quantity of interest. In Figure 5.1(a) the error with respect to the cost functional (5.1) and in Figure 5.1(b) the error with respect to the quantity of interest (5.3) are considered, respectively.

We observe the best behavior of error with respect to the cost functional if the mesh is refined based on  $\eta_1$  and the best behavior of error with respect to the quantity of interest for the refinement based on  $\eta_{\text{QI}}$ .

A series of meshes generated according to the information obtained from the error estimators are shown in Figure 5.2 together with the optimal control  $q$  and the corresponding state  $u$ .

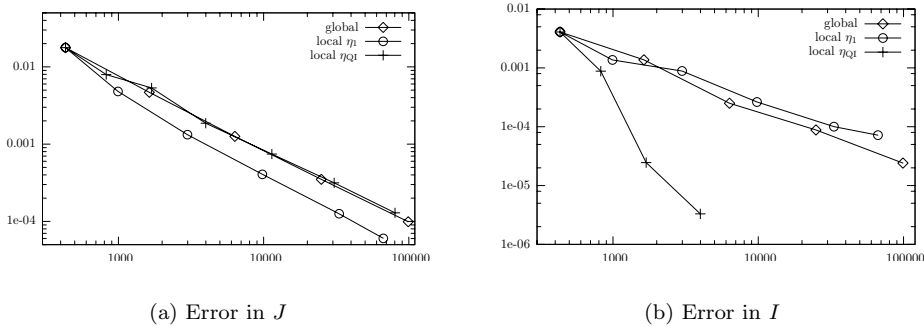


FIG. 5.1. Discretization error for different refinement criteria.

**5.2. Example 2.** Our second example is motivated by a parameter identification problem. The minimization problem is given by

(5.5)      Minimize       $\frac{1}{2}\|u - u^d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|q\|_{L^2(\Omega)}^2, \quad u \in V, \ q \in Q_{\text{ad}},$

subject to

(5.6)      
$$\begin{aligned} -\Delta u + qu &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where  $\Omega = \omega = (0, 0.5) \times (0, 1) \cup (0, 1) \times (0.5, 1)$ ,  $V = H_0^1(\Omega)$ ,  $Q = L^2(\Omega)$ , and the admissible set  $Q_{\text{ad}}$  is given by

$$Q_{\text{ad}} = \{q \in Q \mid q_-(x) \leq q(x) \leq q_+(x) \text{ a.e. on } \Omega\}, \text{ with } q_-(x) = 0, \quad q_+(x) = 0.3.$$

The desired state  $u^d$  and the right-hand side  $f$  are defined as

$$u^d(x) = \frac{1}{8\pi^2} \sin(2\pi x_1) \sin(2\pi x_2), \quad f(x) = 1,$$

and the regularization parameter is chosen  $\alpha = 10^{-4}$ . Note that for any given  $q \in Q_{\text{ad}}$  the state equation (5.6) possesses a unique solution  $u \in V$  due to  $q \geq 0$ .

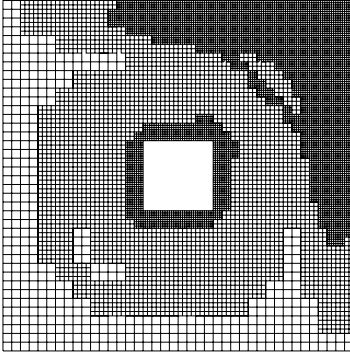
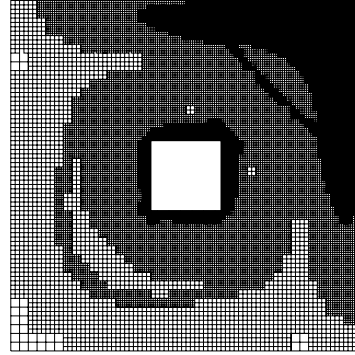
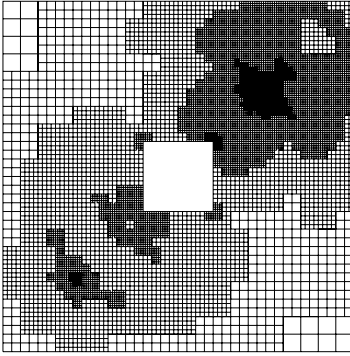
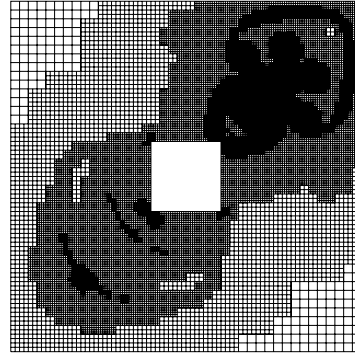
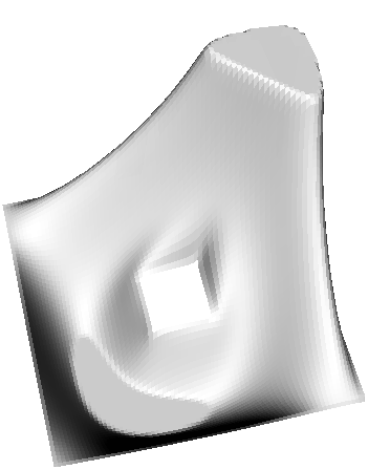
We are interested in the error in the unknown parameter, and thus we choose

$$I(q, u) = \int_{\Omega_O} q(x) \, dx,$$

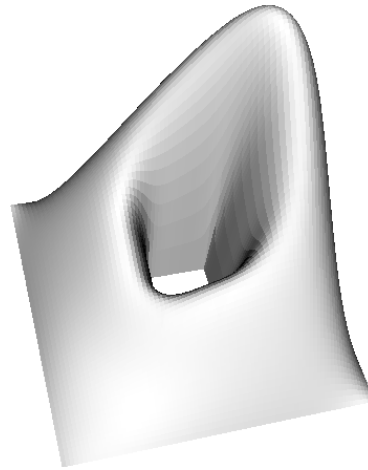
where  $\Omega_O = (0, 0.25) \times (0.75, 1)$ .

In Table 5.2 the effectivity indices, defined as in (5.4), are listed for different types of mesh refinement: global (uniform) refinement, random refinement, refinement based on the error estimator  $\eta_1$  for the cost functional, and refinement based on the error estimator  $\eta_{QI}$  for the quantity of interest. As in the first example we observe that the error estimators provide quantitative information on the discretization errors.



(a) Mesh 3 from  $\eta_1$ (b) Mesh 4 from  $\eta_1$ (c) Mesh 4 from  $\eta_{Q1}$ (d) Mesh 5 from  $\eta_{Q1}$ 

(e) Optimal control



(f) State

FIG. 5.2. *Locally refined meshes and solution.*

TABLE 5.2  
*Effectivity indices.*

N	$I_{\text{eff}}(\eta_1)$	$I_{\text{eff}}(\eta_2)$	$I_{\text{eff}}(\eta_{QI})$
65	1.2	1.2	2.0
225	1.3	1.2	1.9
833	1.4	1.4	1.5
3201	1.5	1.5	1.7

(a) Global refinement

N	$I_{\text{eff}}(\eta_1)$	$I_{\text{eff}}(\eta_2)$	$I_{\text{eff}}(\eta_{QI})$
65	1.2	1.2	2.0
225	1.3	1.3	1.9
785	1.4	1.4	1.6
2705	1.5	1.5	1.7

(b) Refinement according to  $\eta_1$

N	$I_{\text{eff}}(\eta_1)$	$I_{\text{eff}}(\eta_2)$	$I_{\text{eff}}(\eta_{QI})$
65	1.2	1.2	2.0
141	1.2	1.2	2.0
307	1.2	1.2	0.5
763	1.4	1.4	2.0

(c) Random refinement

N	$I_{\text{eff}}(\eta_1)$	$I_{\text{eff}}(\eta_2)$	$I_{\text{eff}}(\eta_{QI})$
65	1.2	1.2	2.0
173	1.2	1.2	1.8
509	1.2	1.2	1.3
1317	1.2	1.2	1.3

(d) Refinement according to  $\eta_{QI}$

From Figure 5.3(a), where the discretization error with respect to the quantity of interest is plotted for different refinement criteria, we again observe that the local mesh refinement based on the appropriate error estimator leads to a certain saving in degrees of freedom for achieving a given tolerance for the discretization error. A typical mesh generated using the information obtained from  $\eta_{QI}$  is shown in Figure 5.3(b).

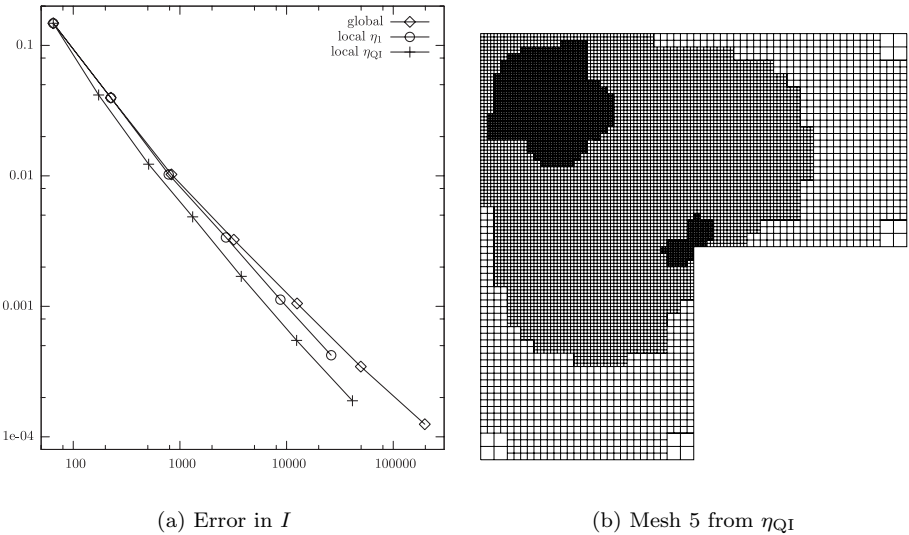


FIG. 5.3. *Discretization error and mesh.*

REFERENCES

[1] N. ARADA, E. CASAS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Comput. Optim. Appl., 23 (2002), pp. 201–229.

- [2] R. BECKER, *Estimating the control error in the discretized PDE-constrained optimization*, J. Numer. Math., 14 (2006), pp. 163–185.
- [3] *The Finite Element Toolkit* GASCOINGE; <http://www.gascoigne.uni-hd.de/>.
- [4] R. BECKER, H. KAPP, AND R. RANNACHER, *Adaptive finite element methods for optimal control of partial differential equations: Basic concept*, SIAM J. Control Optim., 39 (2000), pp. 113–132.
- [5] *A C++ Library for Optimization with Stationary and Nonstationary PDEs*; <http://www.rodobo.uni-hd.de/>.
- [6] R. BECKER AND R. RANNACHER, *An optimal control approach to a posteriori error estimation*, in Acta Numerica 2001, A. Iserles, ed., Cambridge University Press, Cambridge, UK, 2001, pp. 1–102.
- [7] R. BECKER AND B. VEXLER, *A posteriori error estimation for finite element discretizations of parameter identification problems*, Numer. Math., 96 (2004), pp. 435–459.
- [8] R. BECKER AND B. VEXLER, *Mesh refinement and numerical sensitivity analysis for parameter calibration of partial differential equations*, J. Comput. Phys., 206 (2005), pp. 95–110.
- [9] M. BERGOUNIOUX, K. ITO, AND K. KUNISCH, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.
- [10] S. BRENNER AND R.L. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, Heidelberg, New York, 1994.
- [11] E. CASAS, M. MATEOS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl., 31 (2005), pp. 193–220.
- [12] E. CASAS AND F. TRÖLTZSCH, *Error estimates for linear-quadratic elliptic control problems*, in Analysis and Optimization of Differential Systems (Constanta, 2002), Kluwer Academic Publishers, Boston, MA, 2003, pp. 89–100.
- [13] K. ERIKSSON, D. ESTEP, P. HANSBO, AND C. JOHNSON, *Introduction to adaptive methods for differential equations*, in Acta Numerica 1995, A. Iserles, ed., Cambridge University Press, Cambridge, UK, 1995, pp. 105–158.
- [14] K. ERIKSSON, D. ESTEP, P. HANSBO, AND C. JOHNSON, *Computational Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.
- [15] R. S. FALK, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.
- [16] A. V. FURSIKOV, *Optimal Control of Distributed Systems: Theory and Applications*, Transl. Math. Monogr. 187, AMS, Providence, RI, 2000.
- [17] A. GAEVSKAYA, R. H. W. HOPPE, Y. ILIASH, AND M. KIEWEG, *Convergence analysis of an adaptive finite element method for distributed control problems with control constraints*, in Control of Coupled Partial Differential Equations, Birkhäuser, Basel, 2007, pp. 47–68.
- [18] T. GEVECI, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, RAIRO Anal. Numér., 13 (1979), pp. 313–328.
- [19] R. GRIESSE AND B. VEXLER, *Numerical sensitivity analysis for the quantity of interest in PDE-constrained optimization*, SIAM J. Sci. Comput., 29 (2007), pp. 22–48.
- [20] M. HINTERMÜLLER, R. H. W. HOPPE, Y. ILIASH, AND M. KIEWEG, *An a posteriori error analysis of adaptive finite element methods for distributed elliptic control problems with control constraints*, ESAIM Control Optim. Calc. Var., to appear.
- [21] M. HINTERMÜLLER, K. ITO, AND K. KUNISCH, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.
- [22] M. HINZE, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–61.
- [23] R. H. W. HOPPE, Y. ILIASH, C. IYUNNI, AND N. H. SWEILAM, *A posteriori error estimates for adaptive finite element discretizations of boundary control problems*, J. Numer. Math., 14 (2006), pp. 57–82.
- [24] K. KUNISCH AND A. RÖSCH, *Primal-dual active set strategy for a general class of constrained optimal control problems*, SIAM J. Optim., 13 (2002), pp. 321–334.
- [25] R. LI, W. LIU, H. MA, AND T. TANG, *Adaptive finite element approximation for distributed elliptic optimal control problems*, SIAM J. Control Optim., 41 (2002), pp. 1321–1349.
- [26] J.-L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Grundlehren Math. Wiss. 170, Springer-Verlag, Berlin, 1971.
- [27] W. LIU AND N. YAN, *A posteriori error estimates for distributed convex optimal control problems*, Adv. Comput. Math., 15 (2001), pp. 285–309.
- [28] K. MALANOWSKI, *Convergence of approximations versus regularity of solutions for convex, control-constrained optimal-control problems*, Appl. Math. Optim., 8 (1982), pp. 69–95.
- [29] H. MAURER AND J. ZOWE, *First and second-order necessary and sufficient optimality conditions*

- for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.
- [30] D. MEIDNER AND B. VEXLER, *Adaptive space-time finite element methods for parabolic optimization problems*, SIAM J. Control Optim., 46 (2007), pp. 116–142.
  - [31] C. MEYER AND A. RÖSCH, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim., 43 (2004), pp. 970–985.
  - [32] C. MEYER AND A. RÖSCH,  *$L^\infty$ -estimates for approximated optimal control problems*, SIAM J. Control Optim., 44 (2005), pp. 1636–1649.
  - [33] A. RÖSCH, *Error estimates for linear-quadratic control problems with control constraints*, Optim. Methods Softw., 21 (2006), pp. 121–134.
  - [34] F. TRÖLTZSCH, *Optimale Steuerung partieller Differentialgleichungen*, Vieweg, Wiesbaden, Germany, 2005.
  - [35] R. VERFÜRTH, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley/Teubner, New York, Stuttgart, 1996.
  - [36] B. VEXLER, *Adaptive Finite Elements for Parameter Identification Problems*, Ph.D. thesis, Institut für Angewandte Mathematik, Universität Heidelberg, Heidelberg, Germany, 2004.

## NECESSARY CONDITIONS FOR CONSTRAINED PROBLEMS UNDER MANGASARIAN–FROMOWITZ CONDITIONS\*

MARIA DO ROSÁRIO DE PINHO<sup>†</sup> AND JAVIER F. ROSENBLUETH<sup>‡</sup>

**Abstract.** The focus of this paper is on first order necessary conditions for optimal control problems with mixed state-control equality and inequality constraints. We consider the case when the cost and dynamics are nonsmooth, and the constraints satisfy Mangasarian–Fromowitz-type assumptions that weaken the commonly used hypothesis that the Jacobian of the active constraints, with respect to the free variable, is of full rank. The results are formulated as unmaximized Hamiltonian inclusion-type conditions involving not the customary product of partial subdifferentials but the joint subdifferential with respect to the state and control variables.

**Key words.** optimal control, nonsmooth analysis, mixed constraints

**AMS subject classification.** 49K15

**DOI.** 10.1137/060663623

**1. Introduction.** Consider the following optimal control problem with mixed constraints:

$$(P) \quad \left\{ \begin{array}{ll} \text{Minimize } l(x(0), x(1)) \text{ subject to} \\ \dot{x}(t) = f(t, x(t), u(t), v(t)) & \text{a.e. in } T, \\ 0 = b(t, x(t), u(t), v(t)) & \text{a.e. in } T, \\ 0 \geq g(t, x(t), u(t), v(t)) & \text{a.e. in } T, \\ v(t) \in V(t) & \text{a.e. in } T, \\ (x(0), x(1)) \in C, \end{array} \right.$$

where  $T = [0, 1]$ , and we are given functions

$$l: \mathbf{R}^n \times \mathbf{R}^n \rightarrow \mathbf{R}, \quad (f, b, g): T \times \mathbf{R}^n \times \mathbf{R}^{k_u} \times \mathbf{R}^{k_v} \rightarrow \mathbf{R}^n \times \mathbf{R}^{m_b} \times \mathbf{R}^{m_g},$$

$V(t) \subset \mathbf{R}^{k_v}$  for all  $t \in T$ , and  $C \subset \mathbf{R}^n \times \mathbf{R}^n$ . Let  $m := m_b + m_g$ , let  $k := k_u + k_v$ , and assume that  $k \geq m$ . Usually one has  $m_b \geq 1$  and  $m_g \geq 1$ , but we allow for  $m_b = 0$  (no equality constraints) or  $m_g = 0$  (no inequality constraints).

For (P), a *process* is a triple  $(x, u, v)$  comprising a function  $x \in W^{1,1}(T; \mathbf{R}^n)$  and measurable functions  $u: T \rightarrow \mathbf{R}^{k_u}$  and  $v: T \rightarrow \mathbf{R}^{k_v}$  satisfying the constraints. Here  $W^{1,1}(T; \mathbf{R}^n)$  denotes the space of absolutely continuous functions mapping  $T$  to  $\mathbf{R}^n$ . A process  $(\bar{x}, \bar{u}, \bar{v})$  is a *strong minimizer* for (P) if there exists  $\epsilon > 0$  such that it minimizes the cost over all processes  $(x, u, v)$  satisfying  $|x(t) - \bar{x}(t)| \leq \epsilon$  for all  $t \in T$ , and it is a *weak minimizer* if, for some  $\epsilon > 0$ , it minimizes the cost over all processes  $(x, u, v)$  satisfying  $|x(t) - \bar{x}(t)| \leq \epsilon$  for all  $t \in T$  and

$$|u(t) - \bar{u}(t)| \leq \epsilon, \quad |v(t) - \bar{v}(t)| \leq \epsilon \quad \text{a.e. in } T.$$

\*Received by the editors June 23, 2006; accepted for publication (in revised form) August 16, 2007; published electronically February 1, 2008. This work was supported by FEDER and FCT, Projecto POSC/EEA-SRI/61831/2004.

<http://www.siam.org/journals/sicon/47-1/66362.html>

<sup>†</sup>ISR and DEEC, Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal (mrpinho@fe.up.pt).

<sup>‡</sup>IIMAS-UNAM, Universidad Nacional Autónoma de México, Apartado Postal 20-726, México DF 01000, México (jfrl@servidor.unam.mx).

The set  $\mathcal{I}_a(t)$  denotes the set of indexes of the *active constraints*, that is,

$$\mathcal{I}_a(t) = \{i \in \{1, \dots, m_g\} \mid g_i(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) = 0\},$$

and  $q_a(t)$  denotes the cardinality of  $\mathcal{I}_a(t)$ . Also

$$\nabla_u g^{\mathcal{I}_a(t)}(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) \in \mathbf{R}^{q_a(t) \times k_u}$$

(if  $q_a(t) = 0$ , the latter holds vacuously) is the matrix we obtain after removing from  $\nabla_u g(t, \bar{x}(t), \bar{u}(t), \bar{v}(t))$  all the rows of index  $i \notin \mathcal{I}_a(t)$ .

In what follows the set of necessary conditions for optimal control problems with nonsmooth data in the spirit of [3] is referred to as *nonsmooth maximum principle*.

This paper focuses on necessary conditions of optimality for problem (P) with cost and dynamics possibly nonsmooth and with smooth mixed constraints. The distinct aspects of (P) are the equality and inequality mixed constraints and the fact that the control variable comprises two components,  $u$  (unconstrained) and  $v$  (pointwise constrained). This splitting decomposition of the control variable is common in the literature (see [2, 13] for an explanation, as well as [1, 12, 14], where the statement of the problem is proposed as a canonical one and there is a deep investigation on the subject). We keep such division because optimal control problems with mixed constraints are relevant to optimal control problems involving differential and algebraic equations (for a discussion on this topic we refer the reader to [9, 10]).

Although it is well known that the maximum principle is not in general valid for (P), weak and strong forms of the maximum principle have been shown to hold when some regularity assumptions are imposed on the mixed constraints (see, for example, [2, 13, 16, 17, 20, 24] to name but a few). Derivation of optimality conditions for problems with nonregular mixed constraints remains a largely unexplored area (see [13] and the references therein).

For problems with smooth data including continuity with respect to  $t$ , regularity assumptions commonly involve the full rank condition

$$(1.1) \quad \det M(t)M(t)^T \neq 0 \quad \text{for all } t \in T$$

for a certain matrix  $M$ , the more general being

$$(1.2) \quad F(t) := \begin{pmatrix} \nabla_u b(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) \\ \nabla_u g^{\mathcal{I}_a(t)}(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) \end{pmatrix}.$$

Exceptions are to be found in [2, 13], where strong versions of the maximum principle for smooth problems with data *measurable* with respect to  $t$  are validated under assumptions that may be viewed in analogy to Mangasarian–Fromowitz constraint qualifications (to be defined below), also known as *positive linear independence* conditions.

Strong forms of the nonsmooth maximum principle for (P), which apply to strong minimizers, are derived in [10, 11]. In the aforementioned papers, regularity assumptions on the mixed constraints involve convexity. Weak nonsmooth maximum principles for (P), which in turn apply to weak minimizers, have also been derived in [7, 21], where convexity assumptions are replaced by differentiability with respect to  $(x, u, v)$  of the functions defining the mixed constraints. Because the data of (P) in [7, 21] are assumed to be merely measurable with respect to  $t$ , the full rank condition (1.1) is replaced by a uniform full rank condition of the form

$$(1.3) \quad \det M(t)M(t)^T \geq K \quad \text{a.e. in } T$$

for some  $K > 0$ , applied to the matrix

$$\Upsilon(t) := \begin{pmatrix} \nabla_u b(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) & 0 \\ \nabla_u g(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) & \text{diag} \{ \sqrt{-g_i(t, \bar{x}(t), \bar{u}(t), \bar{v}(t))} \} \end{pmatrix}.$$

A weak version of the nonsmooth maximum principle, derived in [7] under the uniform full rank condition imposed on  $\Upsilon$  (that is,  $M = \Upsilon$  in (1.3)), is extended in [5] to cover problems for which (1.3) is imposed on  $F$ . Remarkably, the uniform full rank condition (1.3) on  $\Upsilon$  implies (1.3) on  $F$ , but the converse may not hold (see [7]). Notice that condition (1.1) on  $\Upsilon$  is in fact equivalent to (1.1) on  $F$ .

Here we establish the validity of a weak nonsmooth maximum principle for (P) under regularity assumptions on the mixed constraints that may be viewed in analogy to Mangasarian–Fromowitz constraint qualifications in mathematical programming. The conditions we consider are an adaptation of [15, condition (23)]. Regularity assumptions of similar nature are imposed in [4, 13], where strong versions of the maximum principle are derived for smooth problems. Also, they are used in [22] to derive first and second order optimality conditions for problems involving more general constraints. However, in contrast to [4, 13, 22], we treat problems with possibly nonsmooth dynamics and cost. Moreover, as we illustrate through an example, the Mangasarian–Fromowitz-type conditions we impose on (P) are in general less restrictive constraint qualifications than the uniform full rank condition (1.3) imposed on  $F$  (though both sets of conditions coincide when  $m_g = 0$ ).

The novelty of this paper concerns both the nature of the regularity assumptions and the analytic techniques used to derive necessary conditions for (P) with possibly nonsmooth data. Although we impose some degree of smoothness on the mixed constraints, the proof of our main result (Theorem 3.1) relies heavily on techniques developed for nonsmooth analysis. The key idea behind the proof is based on a technique introduced in [3] for problems with pure state constraints (see also [25]). A crucial step in the proof is the definition of a sequence of optimization problems leading to a sequence of optimal control problems with mixed constraints to which optimality necessary conditions under the uniform full rank condition (1.3), previously obtained in [5], apply.

Similarly to [5, 7], the necessary conditions for (P) validated in this paper involve  $\partial_{x,u,v}H$ , the joint subdifferential of the function

$$H(t, x, p, q, r, u, v) = p \cdot f(t, x, u, v) + q \cdot b(t, x, u, v) + r \cdot g(t, x, u, v)$$

in the state and control variables. The necessary conditions we deal with in this paper were first derived in [8] for “standard” optimal control problems (that is, when the mixed constraints defined by  $g$  and  $b$  are absent), and they were later extended to cover more general problems as, for example, problems with pure state constraints (see [6] and the references therein). In [8] such conditions are designated as Euler–Lagrange inclusion conditions. This designation was soon abandoned to avoid confusion with the same name used in nonsmooth optimal control for conditions involving generalized normals to the graph of the admissible velocity mapping. Since, for standard nonsmooth optimal control problems, our conditions involve the *unmaximized Hamiltonian* (also known as pseudo-Hamiltonian or Pontryagin’s Hamiltonian), they were coined *unmaximized Hamiltonian inclusion (UHI)-type conditions*.

UHI-type conditions are distinct from the conventional weak nonsmooth maximum principle which involves  $\partial_x H \times \partial_u H \times \partial_v H$ , the product of the partial subdifferentials of  $H$  in the state and control variables. As shown in [5, 6], the main features of

UHI-type conditions are retained in the presence of mixed or pure state constraints. Namely, these conditions are optimality sufficient conditions in the linear-convex normal case, whereas the nonsmooth maximum principle is not. Moreover, for standard optimal control problems and for problems with pure state constraints, UHI-type conditions can eliminate processes that satisfy strong nonsmooth maximum principles and yet are not locally optimal (see [6, 8]), a feature which is retained for UHI-type conditions for (P).

**2. Preliminaries.** The notation  $r \geq 0$  for  $r \in \mathbf{R}^p$  ( $p \in \mathbf{N}$ ) means that each component of  $r$  is nonnegative. Throughout,  $|\cdot|$  denotes the Euclidean norm, and  $B$  represents the closed unit ball centered at the origin regardless of the dimension of the underlined space. For much of the analysis we shall denote by  $(\bar{x}, \bar{u}, \bar{v})$  a local minimizer of (P) and by  $\bar{\phi}(t)$  the evaluation of a function  $\phi$  at  $(t, \bar{x}(t), \bar{u}(t), \bar{v}(t))$ . The set  $\Omega_\epsilon(t)$  is defined as

$$\Omega_\epsilon(t) := (\bar{x}(t) + \epsilon B) \times (\bar{u}(t) + \epsilon B) \times ((\bar{v}(t) + \epsilon B) \cap V(t)).$$

The linear space  $W^{1,1}(T; \mathbf{R}^p)$  denotes the space of absolutely continuous functions,  $L^1(T; \mathbf{R}^p)$  the space of integrable functions, and  $L^\infty(T; \mathbf{R}^p)$  the space of essentially bounded functions mapping  $T$  to  $\mathbf{R}^p$ , respectively. For  $\delta \in L^\infty(T; \mathbf{R}^{m_a})$ ,  $\delta^{\mathcal{I}_a(t)}(t) \in \mathbf{R}^{q_a(t)}$  denotes the vector we get after removing from  $\delta(t)$  all the components  $\delta_i(t)$  such that  $i \notin \mathcal{I}_a(t)$ .

Since we consider (P) with possibly nonsmooth data, we make use of standard constructs from nonsmooth analysis. The full calculus for some basic constructions in nonsmooth analysis can be found, for example, in [3, 18, 19, 23, 25]. The following definition will be essential in our setup.

**DEFINITION 2.1.** Let  $A \subset \mathbf{R}^k$  be a closed set and  $x \in A$ . We say that  $p \in \mathbf{R}^k$  is a limiting normal to  $A$  at  $x$  if there exist  $p_i \rightarrow p$  and  $x_i \rightarrow x$  in  $A$  and a sequence of positive numbers  $\{M_i\}$  such that, for each  $i \in \mathbf{N}$ ,

$$p_i \cdot (y - x_i) \leq M_i |y - x_i|^2 \quad \text{for all } y \in A.$$

The limiting normal cone to  $A$  at  $x$ , written  $N_A(x)$ , is the set of all limiting normals to  $A$  at  $x$ .

Given a lower semicontinuous function  $f: \mathbf{R}^k \rightarrow \mathbf{R} \cup \{+\infty\}$  and a point  $x \in \mathbf{R}^k$  such that  $f(x) < \infty$ , the limiting subdifferential of  $f$  at  $x$  (also referred to as Mordukhovich subdifferential) is the set

$$\partial f(x) := \{\zeta \mid (-1, \zeta) \in N_{\text{epi}\{f\}}(f(x), x)\},$$

where  $\text{epi}\{f\} := \{(\eta, x) \mid \eta \geq f(x)\}$  denotes the epigraph set (when  $f$  is Lipschitz continuous near  $x$ , the convex hull of the limiting subdifferential,  $\text{co } \partial f(x)$ , coincides with the Clarke subdifferential (see [3])).

For completeness we state two results that will be of importance in what follows. The first one corresponds to a uniform implicit function theorem [9, Corollary 4.2].

**PROPOSITION 2.2.** Consider a set  $A \subset \mathbf{R}^k$ , a number  $\alpha > 0$ , a family of functions  $\{\psi_a: \mathbf{R}^m \times \mathbf{R}^n \rightarrow \mathbf{R}^n\}_{a \in A}$ , and a point  $(u_0, v_0) \in \mathbf{R}^m \times \mathbf{R}^n$  such that  $\psi_a(u_0, v_0) = 0$  for all  $a \in A$ . Assume that

- (i)  $\psi_a$  is continuously differentiable on  $(u_0, v_0) + \alpha B$  for all  $a \in A$ ,
- (ii) there exists a monotone increasing function  $\theta: (0, \infty) \rightarrow (0, \infty)$  with  $\theta(s) \downarrow 0$  as  $s \downarrow 0$  such that, for all  $a \in A$  and  $(u, v) \neq (u', v') \in (u_0, v_0) + \alpha B$ ,

$$|\nabla \psi_a(u, v) - \nabla \psi_a(u', v')| \leq \theta(|(u, v) - (u', v')|),$$



(iii)  $\nabla_v \psi_a(u_0, v_0)$  is nonsingular for each  $a \in A$  and there exists  $c > 0$  such that, for all  $a \in A$ ,  $|\nabla_v \psi_a(u_0, v_0)|^{-1} \leq c$ .

Then there exist  $\delta \geq 0$  and a family of continuously differentiable functions  $\{\phi_a: u_0 + \delta B \rightarrow v_0 + \alpha B\}_{a \in A}$  which are Lipschitz continuous with common Lipschitz constant  $k$  such that, for all  $a \in A$ ,

- (a)  $v_0 = \phi_a(u_0)$ ,
- (b)  $\psi_a(u, \phi_a(u)) = 0$  for all  $u \in u_0 + \delta B$ ,
- (c)  $\nabla_u \phi(u_0) = -[\nabla_v \psi_a(u_0, v_0)]^{-1} \nabla_u \psi_a(u_0, v_0)$ .

The numbers  $\delta$  and  $k$  depend on  $\theta$ ,  $c$ , and  $\alpha$  only. Furthermore, if  $A$  is a Borel set and  $a \mapsto \psi_a(u, v)$  is a Borel measurable function for each  $(u, v) \in (u_0, v_0) + \alpha B$ , then  $a \mapsto \phi_a(u)$  is a Borel measurable function for each  $u \in u_0 + \delta B$ .

The second result we state (see [5, Theorem 3.1]) corresponds to UHI-type conditions for (P) under full rank assumptions. We invoke the following hypotheses on (P), which make reference to a parameter  $\epsilon > 0$  and a process  $(\bar{x}, \bar{u}, \bar{v})$ . The notation  $(b, g)(\cdot)$  means  $(b(\cdot), g(\cdot))$ .

(H1) The function  $t \mapsto f(t, x, u, v)$  is Lebesgue measurable for each  $(x, u, v)$  and there exists an integrable function  $k_f$  such that, for almost every  $t \in T$ ,

$$|f(t, x, u, v) - f(t, x', u', v')| \leq k_f(t) |(x, u, v) - (x', u', v')|$$

for all  $(x, u, v), (x', u', v') \in \Omega_\epsilon(t)$ .

(H2) Graph  $V$  is a Borel measurable set and  $V_\epsilon(t) := (\bar{v}(t) + \epsilon B) \cap V(t)$  is closed for almost every  $t \in T$ .

(H3)  $C$  is closed and  $l$  is locally Lipschitz in a neighborhood of  $(\bar{x}(0), \bar{x}(1))$ .

(H4)  $(b, g)(\cdot, x, u, v)$  is measurable for each  $(x, u, v)$  and  $t \mapsto g(t, \bar{x}(t), \bar{u}(t), \bar{v}(t))$  is  $L^\infty(T, \mathbf{R}^{m_g})$ .

(H5)  $(b, g)(t, \cdot, \cdot, \cdot)$  is continuously differentiable on  $(\bar{x}(t), \bar{u}(t), \bar{v}(t)) + \epsilon B$  a.e. in  $T$  and there exists an integrable function  $L_{b,g}$  such that, for almost every  $t \in T$ ,

$$|(b, g)(t, x, u, v) - (b, g)(t, x', u', v')| \leq L_{b,g}(t) |(x, u, v) - (x', u', v')|$$

for all  $(x, u, v), (x', u', v') \in \Omega_\epsilon(t)$ .

(H6) There exist  $K_{b,g} > 0$  and an increasing function  $\tilde{\theta}: (0, \infty) \rightarrow (0, \infty)$  with  $\tilde{\theta}(s) \downarrow 0$  as  $s \downarrow 0$  such that, for almost every  $t \in T$ ,

$$|\nabla_x(\bar{b}, \bar{g})(t)| + |\nabla_u(\bar{b}, \bar{g})(t)| + |\nabla_v(\bar{b}, \bar{g})(t)| \leq K_{b,g},$$

$$|\nabla_{x,u,v}(b, g)(t, x, u, v) - \nabla_{x,u,v}(b, g)(t, x', u', v')| \leq \tilde{\theta}(|(x, u, v) - (x', u', v')|)$$

for all  $(x, u, v) \neq (x', u', v')$  belonging to  $\Omega_\epsilon(t)$ .

(H\*) There exists  $K > 0$  such that  $\det F(t)F(t)^T \geq K$  a.e. in  $T$ , where  $F$  is defined in (1.2).

**THEOREM 2.3.** *Let  $(\bar{x}, \bar{u}, \bar{v})$  be a weak minimizer for (P). Set*

$$H(t, x, p, q, r, u, v) := p \cdot f(t, x, u, v) + q \cdot b(t, x, u, v) + r \cdot g(t, x, u, v).$$

*If, for some  $\epsilon > 0$ , (H1)–(H6) and (H\*) are satisfied, then there exist  $p \in W^{1,1}(T; \mathbf{R}^n)$ ,  $q \in L^1(T; \mathbf{R}^{m_b})$ ,  $r \in L^1(T; \mathbf{R}^{m_g})$ ,  $\zeta \in L^1(T; \mathbf{R}^{k_v})$ , and  $\lambda \geq 0$  such that*

- (i)  $\|p\|_\infty + \lambda \neq 0$ ,
- (ii)  $(-\dot{p}(t), 0, \zeta(t)) \in \text{co } \partial_{x,u,v} H(t, \bar{x}(t), p(t), q(t), r(t), \bar{u}(t), \bar{v}(t))$  a.e. in  $T$ ,
- (iii)  $\zeta(t) \in \text{co } N_{V(t)}(\bar{v}(t))$  a.e. in  $T$ ,
- (iv)  $r(t) \cdot g(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) = 0$  and  $r(t) \leq 0$  a.e. in  $T$ ,
- (v)  $(p(0), -p(1)) \in N_C(\bar{x}(0), \bar{x}(1)) + \lambda \partial l(\bar{x}(0), \bar{x}(1))$ .

*Furthermore, for some integrable function  $K_m$ ,  $|(q(t), r(t))| \leq K_m(t)|p(t)|$  a.e. in  $T$ .*

**3. Main results.** In this section we validate the UHI-type conditions given in Theorem 2.3 to cover problems with data satisfying merely Mangasarian–Fromowitz-type assumptions. In this respect, Theorem 2.3 and Proposition 2.2 play a crucial role.

We shall invoke the following additional hypothesis which, like (H1)–(H6), makes reference to some process  $(\bar{x}, \bar{u}, \bar{v})$  and a parameter  $\epsilon > 0$ .

(H7) There exist constants  $K_1, K_2 > 0$  and functions  $h \in L^\infty(T; \mathbf{R}^{k_u})$  and  $a \in L^\infty(T; \mathbf{R}^{m_g})$  with  $|h(t)| = 1$  a.e. in  $T$  such that, for almost every  $t \in T$ ,

- (i)  $a_i(t) > K_2$  for  $i \in \mathcal{I}_a(t)$ ,
- (ii)  $\nabla_u \bar{g}^{\mathcal{I}_a(t)}(t) \cdot h(t) = a^{\mathcal{I}_a(t)}(t)$ ,
- (iii)  $\nabla_u \bar{b}(t) \cdot h(t) = 0$ ,
- (iv)  $\det \nabla_u \bar{b}(t) \nabla_u \bar{b}(t)^T \geq K_1$ .

The difference between (H7) and assumptions previously considered in [13] resides in the fact that (H7) needs to be checked only along the optimal solution. On the other hand, (H7) coincides with (R3) in [22] (in this respect we refer the reader to Theorem 4.1 in [22]).

Hypothesis (H7) is a uniform Mangasarian–Fromowitz-type condition, the term “uniform” being used to emphasize the fact that assumptions (H7)(i) and (H7)(iv) are bounded away from the origin, uniformly in  $t$ .

Let us first state UHI-type conditions for (P) without *equality* constraints; that is, let us concentrate on the problem

$$(Q) \quad \begin{cases} \text{Minimize } l(x(0), x(1)) \text{ subject to} \\ \dot{x}(t) = f(t, x(t), u(t), v(t)) & \text{a.e. in } T, \\ 0 \geq g(t, x(t), u(t), v(t)) & \text{a.e. in } T, \\ (u(t), v(t)) \in \mathbf{R}^{k_u} \times V(t) & \text{a.e. in } T, \\ (x(0), x(1)) \in C. \end{cases}$$

Note that, for this case, hypotheses (H7)(iii) and (H7)(iv) are ignored.

**THEOREM 3.1.** *Let  $(\bar{x}, \bar{u}, \bar{v})$  be a weak minimizer for (Q) with  $m_g \geq 1$ . Set*

$$H(t, x, p, r, u, v) := p \cdot f(t, x, u, v) + r \cdot g(t, x, u, v).$$

*If, for some  $\epsilon > 0$ , (H1)–(H7) are satisfied, then there exist  $p \in W^{1,1}(T; \mathbf{R}^n)$ ,  $r \in L^1(T; \mathbf{R}^{m_g})$ ,  $\zeta \in L^1(T; \mathbf{R}^{k_v})$ , and  $\lambda \geq 0$  such that*

- (i)  $\|p\|_\infty + \lambda \neq 0$ ,
- (ii)  $(-\dot{p}(t), 0, \zeta(t)) \in \text{co } \partial_{x,u,v} H(t, \bar{x}(t), p(t), r(t), \bar{u}(t), \bar{v}(t))$  a.e. in  $T$ ,
- (iii)  $\zeta(t) \in \text{co } N_{V(t)}(\bar{v}(t))$  a.e. in  $T$ ,
- (iv)  $r(t) \cdot g(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) = 0$  and  $r(t) \leq 0$  a.e. in  $T$ ,
- (v)  $(p(0), -p(1)) \in N_C(\bar{x}(0), \bar{x}(1)) + \lambda \partial l(\bar{x}(0), \bar{x}(1))$ .

*Furthermore, for some integrable function  $R$ ,  $|r(t)| \leq R(t)|p(t)|$  a.e. in  $T$ .*

We now turn to problem (P) with mixed constraints in the form of equalities and inequalities.

**THEOREM 3.2.** *Theorem 2.3 remains valid if we replace (H\*) with (H7).*

Let us point out that, for problems *without* inequality constraints, Theorem 3.2 coincides with [7, Theorem 3.1]. The novelty of the above theorem is that, in the presence of inequality constraints, it subsumes previous UHI-type conditions for (P) validated under full rank conditions (Theorem 2.3). We shall prove this new result by associating (P) with an auxiliary problem in the form of (Q) to which Theorem 3.1 applies.

As mentioned in the introduction, it has been shown that UHI-type conditions in the normal form, for standard optimal control problems, are also sufficient for linear convex problems (see [8]). This feature is retained when UHI-type conditions are extended to cover problems with mixed constraints under the uniform full rank condition  $(H^*)$  as proved in [5, Proposition 2.1]. However, that proposition does not depend on  $(H^*)$  but only on the conclusions (ii)–(v) of [5, Theorem 3.1] which coincide with conclusions (ii)–(v) of Theorem 2.3 above. Consequently, UHI-type conditions for (P) in the normal form, validated under assumption (H7) as in Theorem 3.2, are also sufficient for linear convex problems.

We end this section with two examples. Let us first compare the two hypotheses  $(H^*)$  and (H7). It is a simple matter to see that  $(H^*)$  implies (H7), but the opposite implication may not hold, as the following example shows.

*Example 3.3.* Consider the problem of minimizing  $x(1)$  subject to

$$\dot{x}(t) = u_1^2(t) + u_3^2(t), \quad u_1(t) + u_2^2(t) + u_3^2(t) \leq 0, \quad u_1(t) - u_2^3(t) \leq 0, \quad x(0) = 0.$$

Here the control variable is  $(u_1, u_2, u_3)$ . This is a problem without equality constraints. A minimizer is  $(0, 0, 0, 0)$  and  $\mathcal{I}_a(t) = \{1, 2\}$  for all  $t \in T$ . It is easy to check that Theorem 3.1 holds with  $p(t) = -1$  and  $\lambda = 1$ . For this problem, the matrix

$$F(t) = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

does satisfy (H7) but not  $(H^*)$ : set  $h(t) = (1, 0, 0)^T$ ,  $a(t) = (1, 1)^T$ , and  $K_2 = 1/2$ .

In situations when  $m_g \neq 0$  and  $m_b \neq 0$ , it is crucial for  $h \in L^\infty(T; \mathbf{R}^{k_u})$  to satisfy simultaneously (H7)(ii) and (H7)(iii). The following simple example, for which there is no such  $h$ , illustrates this fact.

*Example 3.4.* Consider the problem of minimizing  $x(1)$  subject to

$$\dot{x}(t) = u_1(t), \quad u_2(t) - x^2(t) = 0, \quad u_2(t) - u_1^3(t) \leq 0, \quad x(0) = 0.$$

It is easy to see that  $(\bar{x}, \bar{u}_1, \bar{u}_2) \equiv (0, 0, 0)$  is a minimizer. Along this minimizer the inequality constraint is active for all  $t \in T$ . For this problem we have

$$H(t, x, p, q, r, u_1, u_2) = pu_1 + q(u_2 - x^2) + r(u_2 - u_1^3).$$

Theorem 3.2 does not hold for this problem. Indeed, by applying the conclusions of this theorem, we get

$$(-\dot{p}(t), 0, 0) = (0, p(t), q(t) + r(t))$$

and  $p(1) = -\lambda$ . It follows that  $p \equiv 0$ . Consequently  $\lambda = 0$ , contradicting the nontriviality condition (i) of Theorem 3.2. The pathological aspect of this problem is that (H7) is not satisfied. Indeed, since

$$\nabla_{u_1, u_2} b(t, 0, 0, 0) = \nabla_{u_1, u_2} g(t, 0, 0, 0) = (0, 1) \quad \text{for all } t \in T,$$

it is impossible to define a vector  $h \in \mathbf{R}^2$  for which

$$\nabla_{u_1, u_2} b(t, 0, 0, 0) \cdot h = 0 \quad \text{and} \quad \nabla_{u_1, u_2} g(t, 0, 0, 0) \cdot h \neq 0.$$

**4. Proof of Theorem 3.1.** The proof technique we shall use consists in defining a sequence of optimization problems  $(R_k)$ , and, by applying Ekeland's variational principle to such a sequence, we obtain a sequence of optimal control problems with mixed constraints, satisfying full rank conditions, to which Theorem 2.3 applies. Taking limits, we get necessary conditions for (Q) satisfying (ii)–(v) of Theorem 3.1. Finally, by appealing to (H7), we prove that the nontriviality condition (i) of Theorem 3.1 is also verified. We proceed in six steps.

*Step 1.* Introduction of new sets and functions.

We shall find it convenient to introduce auxiliary functions  $\bar{\alpha}, \bar{z}, \bar{w}: T \rightarrow \mathbf{R}^{m_g}$  such that  $\bar{\alpha}(t) = \bar{z}(t) = \bar{w}(t) = 0$  a.e. in  $T$ . Also, set

$$U_\epsilon(t) := \bar{u}(t) + \epsilon B, \quad V_\epsilon(t) := V(t) \cap (\bar{v}(t) + \epsilon B), \quad \mathcal{A}_\epsilon(t) := \bar{\alpha}(t) + \epsilon B,$$

and consider the functions

$$F_z(w, \alpha) := w - \alpha, \quad F_w(\alpha) := \alpha^2, \quad G(t, x, z, u, v, \alpha) := g(t, x, u, v) + z + \alpha$$

which take values in  $\mathbf{R}^{m_g}$ . Componentwise, these functions are defined as

$$F_{z_i}(w, \alpha) = w_i - \alpha_i, \quad F_{w_i}(\alpha) = \alpha_i^2(t), \quad G_i(t, x, z, u, v, \alpha) = g_i(t, x, u, v) + z_i + \alpha_i$$

for all  $i \in \{1, \dots, m_g\}$ . Under the hypotheses,  $F_z$ ,  $F_w$ , and  $G$  are measurable with respect to  $t$  and Lipschitz continuous with respect to the remaining variables near

$$(\bar{x}(t), \bar{z}(t), \bar{w}(t), \bar{u}(t), \bar{v}(t), \bar{\alpha}(t)).$$

Also, as it can be easily proved, there exist integrable functions  $C_f, C_{F_z}, C_{F_w}$ , and  $C_G$  such that  $|f(t, x, u, v)| \leq C_f(t)$ ,  $|F_z(w, \alpha)| \leq C_{F_z}(t)$ ,  $|F_w(\alpha)| \leq C_{F_w}(t)$ , and

$$|G(t, x, z, u, v, \alpha)| \leq C_G(t)$$

for all  $(x, z, w, u, v, \alpha) \in (\bar{x}(t), \bar{z}(t), \bar{w}(t), \bar{u}(t), \bar{v}(t), \bar{\alpha}(t)) + \epsilon B$  a.e. in  $T$ .

*Step 2.* Definition of a sequence of optimization problems and verification that Ekeland's variational principle applies to such a sequence.

Define  $W$  as the set of all measurable functions  $(u, v, \alpha)$  and all vectors  $(a, b) \in \mathbf{R}^n \times \mathbf{R}^n$  such that, for almost every  $t \in T$ ,  $(u(t), v(t), \alpha(t)) \in U_\epsilon(t) \times V_\epsilon(t) \times \mathcal{A}_\epsilon(t)$  and  $(a, b) \in C$ , and for which there exist absolutely continuous functions  $x, y, z$ , and  $w$  such that

$$\begin{cases} \dot{x}(t) = f(t, x(t), u(t), v(t)), \quad \dot{y}(t) = 0, \quad \dot{z}(t) = F_z(w(t), \alpha(t)), \quad \dot{w}(t) = F_w(\alpha(t)) & \text{a.e.}, \\ 0 \geq G(t, x(t), z(t), u(t), v(t), \alpha(t)) & \text{a.e.}, \\ (x(t), y(t), z(t), w(t)) \in (\bar{x}(t), \bar{x}(1), \bar{z}(t), \bar{w}(t)) + \epsilon B & \text{a.e.}, \\ (x(0), y(0), z(0), w(0)) = (a, b, 0, 0). \end{cases}$$

Here  $z$  and  $y$  are two additional state variables and  $\alpha$  an additional control.

Let  $\{\epsilon_k\}_{k \in \mathbf{N}}$  be a sequence of positive scalars such that  $\epsilon_k \rightarrow 0$  and  $k \rightarrow \infty$ , and define

$$\Psi_k(x, y, x', y', z, w) := \max\{l(x, y) - l(\bar{x}(0), \bar{x}(1)) + \epsilon_k^2, \epsilon_k^2|w|, |w - z|, |x' - y'|\}.$$

To simplify the notation set  $E = (a, b) \in \mathbf{R}^{2n}$ . Let  $|E - E'| = |a - a'| + |b - b'|$  and

$$\nu((u, v, \alpha), (u', v', \alpha')) := \int_0^1 |u(t) - u'(t)| dt + \int_0^1 |v(t) - v'(t)| dt + \int_0^1 |\alpha(t) - \alpha'(t)| dt.$$

Define  $\delta_W: W \times W \rightarrow \mathbf{R}$  by

$$\delta_W((u, v, \alpha, E), (u', v', \alpha', E')) = \nu((u, v, \alpha), (u', v', \alpha')) + |E - E'|.$$

Consider the sequence of optimization problems

$$(R_k) \quad \text{Minimize } J_k(u, v, \alpha, E) \text{ subject to } (u, v, \alpha, E) \in W,$$

where

$$J_k(u, v, \alpha, E) = \Psi_k(x(0), y(0), x(1), y(1), z(1), w(1)).$$

Set  $\bar{E} = (\bar{x}(0), \bar{x}(1))$ . Since  $(\bar{u}, \bar{v}, \bar{\alpha}, \bar{E}) \in W$ ,  $W$  is nonempty. Moreover, as one readily verifies,  $\delta_W$  defines a metric in  $W$ , the set  $W$  is a complete metric space with respect to this metric, and the function  $(u, v, \alpha, E) \mapsto J_k(u, v, \alpha, E)$  is continuous on  $(W, \delta_W)$ . Now, for all  $k \in \mathbf{N}$ , we have

$$J_k(\bar{u}, \bar{v}, \bar{\alpha}, \bar{E}) = \Psi_k(\bar{x}(0), \bar{x}(1), \bar{x}(1), \bar{x}(1), \bar{z}(1), \bar{w}(1)) = \epsilon_k^2.$$

Since, for all  $k \in \mathbf{N}$ ,  $J_k(u, v, \alpha, E) \geq 0$ ,  $(\bar{u}, \bar{v}, \bar{\alpha}, \bar{E})$  is an “ $\epsilon_k^2$ -minimizer” for  $(R_k)$ . By Ekeland’s variational principle (see [25]), there exists a sequence  $(u_k, v_k, \alpha_k, E_k) \in W$  such that, for each  $k \in \mathbf{N}$ ,

$$(4.1) \quad \delta_W((u_k, v_k, \alpha_k, E_k), (\bar{u}, \bar{v}, \bar{\alpha}, \bar{E})) \leq \epsilon_k$$

and  $(u_k, v_k, \alpha_k, E_k)$  minimizes the perturbed cost function

$$J_k(u, v, \alpha, E) + \epsilon_k \delta_W((u_k, v_k, \alpha_k, E_k), (u, v, \alpha, E))$$

for all  $(u, v, \alpha, E) \in W$ .

*Step 3.* Derivation of a sequence of standard optimal control problems by rewriting the conclusions of Ekeland’s theorem in control theoretic terms.

Let  $(x_k, y_k, z_k, w_k)$  be the trajectory corresponding to  $(u_k, v_k, \alpha_k, E_k)$ . For each  $k \in \mathbf{N}$ , the process

$$(x_k, y_k, z_k, w_k, \omega_1, \omega_2, \omega_3, u_k, v_k, \alpha_k) \quad \text{with} \quad (\omega_1, \omega_2, \omega_3) \equiv (0, 0, 0)$$

solves the control problem  $(C_k)$  of minimizing

$$\begin{aligned} \Phi_k(\gamma(0), \gamma(1)) &:= \Psi_k(x(0), y(0), x(1), y(1), z(1), w(1)) \\ &\quad + \epsilon_k |x(0) - x_k(0)| + \epsilon_k |y(0) - y_k(0)| + \sum_{i=1}^3 \epsilon_k \omega_i(1), \end{aligned}$$

where  $\gamma(t) = (x(t), y(t), z(t), w(t), \omega_1(t), \omega_2(t), \omega_3(t))$ , subject to

$$\left\{ \begin{array}{l} \dot{x}(t) = f(t, x(t), u(t), v(t)), \quad \dot{y}(t) = 0, \quad \dot{z}(t) = F_z(w(t), \alpha(t)), \quad \dot{w}(t) = F_w(\alpha(t)), \\ \dot{\omega}_1(t) = |u(t) - u_k(t)|, \quad \dot{\omega}_2(t) = |v(t) - v_k(t)|, \quad \dot{\omega}_3(t) = |\alpha(t) - \alpha_k(t)|, \\ 0 \geq G(t, x(t), z(t), u(t), v(t), \alpha(t)), \\ (x(t), y(t), z(t), w(t)) \in (\bar{x}(t), \bar{x}(1), \bar{z}(t), \bar{w}(t)) + \epsilon B, \\ (u(t), v(t), \alpha(t)) \in U_\epsilon(t) \times V_\epsilon(t) \times \mathcal{A}_\epsilon(t), \\ (x(0), y(0), z(0), w(0)) \in C \times \{(0, 0)\}, \quad (\omega_1(0), \omega_2(0), \omega_3(0)) = (0, 0, 0), \end{array} \right.$$

all the relations but the last two being understood in an a.e. sense. Since  $\epsilon_k \downarrow 0$ , we can arrange by subsequence extraction, if necessary, that  $\sum \epsilon_k < \infty$ . By (4.1) we deduce that  $(u_k, v_k, \alpha_k) \rightarrow (\bar{u}, \bar{v}, \bar{\alpha})$  strongly in  $L^1$  and  $E_k \rightarrow (\bar{x}(0), \bar{x}(1))$ . Subsequence extraction yields  $(u_k(t), v_k(t), \alpha_k(t)) \rightarrow (\bar{u}(t), \bar{v}(t), \bar{\alpha}(t))$  a.e. We deduce from the above that  $(x_k, y_k, z_k, w_k) \rightarrow (\bar{x}, \bar{x}(1), \bar{z}, \bar{w})$  *uniformly*. By discarding initial terms of the sequence, if necessary, we have

$$(x_k(t), y_k(t), z_k(t), w_k(t)) \in (\bar{x}(t), \bar{x}(1), \bar{z}(t), \bar{w}(t)) + (\epsilon/2)B \quad \text{for all } k.$$

Then  $S_k := (x_k, y_k, z_k, w_k, 0, 0, 0, u_k, v_k, \alpha_k)$  is a weak minimizer of  $(\tilde{C}_k)$ , a variant of  $(C_k)$  obtained by dropping the state constraint

$$(x(t), y(t), z(t), w(t)) \in (\bar{x}(t), \bar{x}(1), \bar{z}(t), \bar{w}(t)) + \epsilon B.$$

*Step 4.* Derivation of necessary conditions for  $(\tilde{C}_k)$ .

Let  $\mathcal{R} := \{(\rho_1, \rho_2, \rho_3, \rho_4) \in \mathbf{R}^4 \mid \rho_i \geq 0, \rho_1 + \dots + \rho_4 = 1\}$ , and define  $\tilde{H}$  as

$$\begin{aligned} \tilde{H}(t, x, z, w, p, q_z, q_w, r, u, v, \alpha) \\ := p \cdot f(t, x, u, v) + q_z \cdot F_z(w, \alpha) + q_w \cdot F_w(\alpha) + r \cdot G(t, x, z, u, v, \alpha). \end{aligned}$$

LEMMA 4.1. *There exist scalars  $\lambda_k$ ,  $\rho_{1_k}$ ,  $\rho_{2_k}$ ,  $\rho_{3_k}$ , and  $\rho_{4_k}$ , vectors  $e_{2_k}$ ,  $e_{3_k} \in \mathbf{R}^{m_g}$ , and  $e_{4_k} \in \mathbf{R}^n$ , integrable functions  $\zeta_k: T \rightarrow \mathbf{R}^{k_v}$  and  $r_k: T \rightarrow \mathbf{R}^{m_g}$ , and absolutely continuous functions  $p_k$ ,  $q_{z_k}$ , and  $q_{w_k}$ , such that*

- (a)  $\lambda_k \rho_{1_k} + \|p_k\|_\infty + \|q_{z_k}\|_\infty + \|q_{w_k}\|_\infty + 3\lambda_k \epsilon_k = 1$ ,
- (b)  $\lambda_k \geq 0$ ,  $|e_{i_k}| = 1$  for  $i = 2, 3, 4$  and  $\rho_k = (\rho_{1_k}, \rho_{2_k}, \rho_{3_k}, \rho_{4_k}) \in \mathcal{R}$ ,
- (c)  $p_k(1) = -\lambda_k \rho_{4_k} e_{4_k}$ ,  $q_{w_k}(1) = -q_{z_k}(1) + \lambda_k \rho_{2_k} \epsilon_k^2 e_{2_k}$ ,  $q_{z_k}(1) = -\lambda_k \rho_{3_k} e_{3_k}$ ,
- (d)  $(p_k(0), -p_k(1)) \in N_C(x_k(0), y_k(0)) + \lambda_k \rho_{1_k} \partial l(x_k(0), y_k(0)) + \lambda_k \epsilon_k (B \times B)$ ,
- (e)  $\zeta_k(t) \in \text{co } N_{V(t)}(v_k(t))$  a.e.,
- (f)  $(-\dot{p}_k(t), -\dot{q}_{z_k}(t), -\dot{q}_{w_k}(t), 0, \zeta_k(t), 0) \in \text{co } \partial \tilde{H}(t, x_k(t), z_k(t), w_k(t), p_k(t), q_{z_k}(t), q_{w_k}(t), r_k(t), u_k(t), v_k(t), \alpha_k(t)) + \lambda_k \epsilon_k (\{(0, 0, 0)\} \times B \times B \times B)$  a.e.,

where  $\partial \tilde{H}$  refers to the subdifferential of  $\tilde{H}$  in the  $(x, z, w, u, v, \alpha)$  variables,

- (g)  $r_k(t) \cdot G(t, x_k(t), z_k(t), u_k(t), v_k(t), \alpha_k(t)) = 0$  and  $r_k(t) \leq 0$  a.e.

*Proof.* Without loss of generality assume that  $|w_k(1)| < 1$  for all  $k \in \mathbf{N}$ . Define

$$\begin{aligned} h(t, x, y, z, w, \omega_1, \omega_2, \omega_3, p, s, q_z, q_w, r, \Pi, u, v, \alpha) \\ := p \cdot f(t, x, u, v) + s \cdot 0 + q_z \cdot F_z(w, \alpha) + q_w \cdot F_w(\alpha) + r \cdot G(t, x, z, u, v, \alpha) \\ + \pi_1 |u - u_k(t)| + \pi_2 |v - v_k(t)| + \pi_3 |\alpha - \alpha_k(t)|, \end{aligned}$$

where  $\Pi = (\pi_1, \pi_2, \pi_3)$ . To simplify the notation, set  $P_k := (p_k, s_k, q_{z_k}, q_{w_k})$ . It is a simple matter to see that

$$\frac{\partial}{\partial \alpha} G(t, x_k(t), z_k(t), u_k(t), v_k(t), \alpha_k(t)) = I \quad \text{a.e. in } T,$$

where  $I$  is the identity matrix. Thus problem  $(\tilde{C}_k)$  satisfies the conditions under which Theorem 2.3 applies. Observing that  $u_k$  and  $\alpha_k$  take values in the interior of the appropriated control sets, an application of Theorem 2.3 yields absolutely continuous functions  $P_k(t)$ ,  $\Pi_k(t)$ , a scalar  $\lambda_k \geq 0$ , and integrable functions  $r_k$ ,  $\zeta_k$  such that

$$(A) \quad \lambda_k + \|p_k\|_\infty + \|s_k\|_\infty + \|q_{z_k}\|_\infty + \|q_{w_k}\|_\infty + \sum_{i=1}^3 \|\pi_{i_k}\|_\infty > 0,$$

- (B)  $(-\dot{P}_k(t), -\dot{\Pi}_k(t), 0, \zeta_k(t), 0) \in \text{co } \partial h_k(t)$  a.e.,  
 (C)  $\zeta_k(t) \in \text{co } N_{V(t)}(v_k(t))$  a.e.,  
 (D)  $(P_k(0), \Pi_k(0), -P_k(1), -\Pi_k(1))$   
 $\in (N_C(x_k(0), y_k(0)) \times \mathbf{R}^{m_g} \times \mathbf{R}^{m_g}) \times \mathbf{R}^3 \times \{(0, 0, 0, 0)\} \times \{(0, 0, 0)\}$   
 $+ \lambda_k \partial \Phi_k(\gamma_k(0), \gamma_k(1)),$   
 (E)  $r_k(t) \cdot G(t, x_k(t), z_k(t), u_k(t), v_k(t), \alpha_k(t)) = 0$  and  $r_k(t) \leq 0$  a.e.,  
 where  $\partial h_k(t)$  denotes the subdifferential of  $h$  with respect to

$$(x, y, z, w, \omega_1, \omega_2, \omega_3, u, v, \alpha)$$

evaluated at  $S_k$ . Since  $h$  does not depend on  $y, \omega_1, \omega_2$ , and  $\omega_3$ , we deduce from (B) that

$$(4.2) \quad -\dot{P}_k(t) = (-\dot{p}_k(t), 0, -\dot{q}_{z_k}(t), -\dot{q}_{w_k}(t)) \quad \text{and} \quad -\dot{\Pi}_k(t) = (0, 0, 0).$$

We now turn to  $\Psi_k$ . Observe that, by (H3),  $\Psi_k(x, y, x', y', z, w)$  is Lipschitz around

$$(\bar{x}(0), \bar{x}(1), \bar{x}(1), \bar{x}(1), \bar{z}(1), \bar{w}(1)).$$

We claim that, for  $k$  sufficiently large,

$$(4.3) \quad \Psi_k(x_k(0), y_k, x_k(1), y_k, z_k(1), w_k(1)) > 0.$$

By definition this relation is nonnegative. Suppose it equals zero. Then

$$(4.4) \quad y_k(1) = x_k(1), \quad l(x_k(0), y_k(0)) < l(\bar{x}(0), \bar{x}(1)),$$

$$(4.5) \quad (x_k(0), x_k(1)) \in C, \quad z_k(1) = 0, \quad w_k(1) = 0.$$

Since  $\dot{w}_k(t) \geq 0$  a.e. and  $w_k(0) = w_k(1) = 0$  we have  $w_k(t) = 0$  and  $\alpha_k(t) = 0$  a.e. Thus  $\dot{z}_k(t) = 0$  a.e., which, together with  $z_k(0) = 0$ , implies that  $z_k(t) = 0$  a.e. It follows from the above that

$$(4.6) \quad G_i(t, x_k(t), z_k(t), u_k(t), v_k(t), \alpha_k(t)) = g_i(t, x_k(t), u_k(t), v_k(t)) \leq 0 \quad \text{a.e.}$$

for all  $i \in \{1, \dots, m_g\}$ . Thus, if  $\Psi_k(x_k(0), y_k, x_k(1), y_k, z_k(1), w_k(1)) = 0$ , we deduce from (4.4)–(4.6) that  $(x_k, u_k, v_k)$  is an admissible process for (Q), contradicting the optimality of  $(\bar{x}, \bar{u}, \bar{v})$ . This proves the claim.

Let us now prove that the limiting subdifferential calculus (Max Rule; see [25, Theorem 5.2.2]) and (4.3) yield the following estimation of the subdifferential of  $\Psi_k$  with respect to  $(x, y, x', y', z, w)$ :

$$(4.7) \quad \partial \Psi_k(x_k(0), y_k(0), x_k(1), y_k(1), z_k(1), w_k(1)) \subset \{\rho_1(\theta_1, \theta_2, 0, 0, 0, 0) \\ + \rho_2(0, 0, 0, 0, \epsilon_k^2 e_2) + \rho_3(0, 0, 0, 0, e_3, -e_3) + \rho_4(0, 0, e_4, -e_4, 0, 0) \mid \\ (\theta_1, \theta_2) \in \partial l(x_k(0), y_k(0)), |e_2| = |e_3| = |e_4| = 1, \rho \in \mathcal{R}, \rho_i = 0 \text{ if } \bar{\psi}_{i_k} < \bar{\Psi}_k\},$$

where  $e_2, e_3$ , and  $e_4$  are vectors such that  $e_2, e_3 \in \mathbf{R}^{m_g}, e_4 \in \mathbf{R}^n$ ,

$$\Psi_k = \max\{\psi_{1_k}, \psi_{2_k}, \psi_{3_k}, \psi_{4_k}\},$$

$$\psi_{1_k} = l(x, y) - l(\bar{x}(0), \bar{x}(1)) + \epsilon_k^2, \quad \psi_{2_k} = \epsilon_k^2 |w|, \quad \psi_{3_k} = |w - z|, \quad \psi_{4_k} = |x' - y'|,$$

and  $\bar{\Psi}_k$  and  $\bar{\psi}_{i_k}$  denote, respectively, the functions  $\Psi_k$  and  $\psi_{i_k}$  evaluated at

$$(x_k(0), y_k(0), x_k(1), y_k(1), z_k(1), w_k(1)).$$

Suppose that  $x_k(1) \neq y_k(1)$ . Then inclusion (4.7) holds with  $e_4 \in \mathbf{R}^n$  and  $|e_4| = 1$ . Analogously, if  $w_k(1) \neq z_k(1)$ , then (4.7) holds with  $e_3 \in \mathbf{R}^{m_g}$ ,  $|e_3| = 1$ . Finally, if  $w_k(1) \neq 0$ , then (4.7) also holds with  $e_2 \in \mathbf{R}^{m_g}$ ,  $|e_2| = 1$ . On the other hand, if  $\bar{\psi}_{i_k} = 0$  for some  $i \in \{2, 3, 4\}$ , from (4.3) we have  $\rho_{i_k} = 0$ . Thus (4.7) holds.

From (D), (4.2), and (4.7) we deduce the existence of vectors

$$\rho_k = (\rho_{1_k}, \rho_{2_k}, \rho_{3_k}, \rho_{4_k}), \quad e_{2_k}, e_{3_k} \in \mathbf{R}^{m_g}, \quad e_{4_k} \in \mathbf{R}^n$$

such that  $\rho_k \in \mathcal{R}$ ,  $\rho_{i_k} = 0$  if  $\bar{\psi}_{i_k} < \bar{\Psi}_k$ ,

$$(4.8) \quad |e_{2_k}| = |e_{3_k}| = |e_{4_k}| = 1,$$

$$(4.9) \quad \Pi_k(t) \equiv \lambda_k \epsilon_k (1, 1, 1).$$

Also,  $s_k$  is constant, and

$$(4.10) \quad q_{z_k}(1) = -\lambda_k \rho_{3_k} e_{3_k},$$

$$(4.11) \quad q_{w_k}(1) = -q_{z_k}(1) + \lambda_k \rho_{2_k} \epsilon_k^2 e_{2_k},$$

$$(4.12) \quad p_k(1) = -s_k = -\lambda_k \rho_{4_k} e_{4_k},$$

$$(4.13) \quad |p_k(1)| = \lambda_k \rho_{4_k},$$

$$(4.14) \quad (p_k(0), s_k) \in N_C(x_k(0), y_k(0)) + \lambda_k \rho_{1_k} \partial l(x_k(0), y_k(0)) + \lambda_k \epsilon_k (B \times B).$$

Equations (4.8), (4.12), and (4.13) yield  $|s_k| = |p_k(1)|$ . Since  $\rho_k \in \mathcal{R}$  we have  $|s_k| = \lambda_k(1 - \rho_{1_k} - \rho_{2_k} - \rho_{3_k})$ , which, together with (4.10), implies that  $\lambda_k = |s_k| + |q_{z_k}(1)| + \lambda_k \rho_{1_k} + \lambda_k \rho_{2_k}$ . It follows from (A) and the above that

$$(4.15) \quad \lambda_k \rho_{1_k} + \lambda_k \rho_{2_k} + \|p_k\|_\infty + 2|p_k(1)| + \|q_{z_k}\|_\infty + |q_{z_k}(1)| + \|q_{w_k}\|_\infty + 3\lambda_k \epsilon_k > 0.$$

We claim that this last inequality implies that

$$(4.16) \quad \lambda_k \rho_{1_k} + \|p_k\|_\infty + 2|p_k(1)| + \|q_{z_k}\|_\infty + |q_{z_k}(1)| + \|q_{w_k}\|_\infty + 3\lambda_k \epsilon_k > 0.$$

Seeking a contradiction suppose that (4.15) holds and

$$(4.17) \quad \lambda_k \rho_{1_k} + \|p_k\|_\infty + 2|p_k(1)| + \|q_{z_k}\|_\infty + |q_{z_k}(1)| + \|q_{w_k}\|_\infty + 3\lambda_k \epsilon_k = 0.$$

Then  $\lambda_k \rho_{2_k} \neq 0$ , i.e.,  $\lambda_k \neq 0$  and  $\rho_{2_k} \neq 0$ . It follows from (4.17) that  $\rho_{1_k} = \rho_{3_k} = \rho_{4_k} = 0$  and, consequently,  $\rho_{2_k} = 1$ . By (4.17) we also have  $q_{z_k}(1) = 0$  and  $q_{w_k}(t) \equiv 0$ . However, since  $\lambda_k \rho_{2_k} \neq 0$ , it follows from (4.11) that  $q_{w_k}(1) \neq 0$ , a contradiction. Thus (4.16) holds, a condition which ensures that

$$\lambda_k \rho_{1_k} + \|p_k\|_\infty + \|q_{z_k}\|_\infty + \|q_{w_k}\|_\infty + 3\lambda_k \epsilon_k > 0.$$

Equations (4.10)–(4.13) yield (c). On the other hand, the requirement that  $\lambda_k \geq 0$ ,  $\rho_k \in \mathcal{R}$ , and (4.8) correspond to (b). Inclusion (4.14), together with (4.12), implies (d), and (C) and (E) are, respectively, (e) and (g). To prove (f), note that

$$\begin{aligned} h(t, x, y, z, w, \omega_1, \omega_2, \omega_3, p, s, q_z, q_w, r_k, \Pi, u, v, \alpha) \\ = \tilde{H}(t, x, z, w, p, q_z, q_w, r_k, u, v, \alpha) + \pi_1 |u - u_k| + \pi_2 |v - v_k| + \pi_3 |\alpha - \alpha_k|. \end{aligned}$$

Estimating the subdifferential  $\text{co } \partial h_k$  with the help of the sum rule and (4.9), we have

$$\begin{aligned} \text{co } \partial h_k \subset \{(\varrho_1, 0, \varrho_2, \varrho_3, 0, 0, 0, \varrho_4, \varrho_5, \varrho_6) \mid (\varrho_1, \varrho_2, \varrho_3, \varrho_4, \varrho_5, \varrho_6) \\ \in \text{co } \partial_{x, z, w, u, v, \alpha} \tilde{H}(t, x_k, z_k, w_k, p_k, q_{z_k}, q_{w_k}, r_k, u_k, v_k, \alpha_k)\} \\ + \lambda_k \epsilon_k (\{(0, 0, 0, 0, 0, 0)\} \times B \times B \times B), \end{aligned}$$



and so (f) holds in virtue of the above, (B), and (4.2). Finally, by a simple scaling argument, we can normalize to one the sum of the multipliers norms without affecting the other conclusions of the lemma, implying that (a) also holds. This completes the proof of the lemma.  $\square$

*Step 5.* Derivation of necessary conditions for (P) by considering  $\epsilon_k \rightarrow 0$  and taking limits.

The sequences  $\{e_{2_k}\}$ ,  $\{e_{3_k}\}$ ,  $\{e_{4_k}\}$ , and  $\{\rho_k\}$  are uniformly bounded by Lemma 4.1(b). From Lemma 4.1(a), the sequences  $\{\lambda_k\}$ ,  $\{\|p_k\|_\infty\}$ ,  $\{\|q_{z_k}\|_\infty\}$ , and  $\{\|q_{w_k}\|_\infty\}$  are also uniformly bounded. We can therefore arrange, by subsequence extraction, if necessary, that

$$e_{2_k} \rightarrow e_2, \quad e_{3_k} \rightarrow e_3, \quad e_{4_k} \rightarrow e_4, \quad \lambda_k \rightarrow \hat{\lambda}, \quad \text{and} \quad \rho_k \rightarrow \rho = (\rho_1, \rho_2, \rho_3, \rho_4),$$

where  $|e_2| = |e_3| = |e_4| = 1$ ,  $\hat{\lambda} \geq 0$ , and  $\rho \in \mathcal{R}$ .

Now, by appealing to measurable selection theorems and taking into account the differentiability properties of  $g$ ,  $F_z$ , and  $F_w$ , we deduce, in view of Lemma 4.1(f), the existence of an integrable function  $K_r$  such that  $|r_k(t)| \leq K_r(t)$  a.e. in  $T$ . We also deduce that the sequences  $\{\dot{p}_k\}$ ,  $\{\dot{q}_{z_k}\}$ , and  $\{\dot{q}_{w_k}\}$  are uniformly integrably bounded and there exists an integrable function  $K_\zeta$  such that  $|\zeta_k(t)| \leq K_\zeta(t)$  a.e. in  $T$ .

We shall find it convenient to introduce the following scaled version  $\tilde{r}_k$  of  $r_k$ , given by  $\tilde{r}_k(t) = (1 + K_r(t))^{-1} r_k(t)$ . Note that the sequence  $\{\|\tilde{r}_k\|_\infty\}$  is uniformly bounded and  $\{t \mapsto \int_0^t \zeta_k ds\}$  is equicontinuous and uniformly bounded. Following extraction of subsequences we have that, for some absolutely continuous functions  $p$ ,  $q_z$ ,  $q_w$ , an integrable function  $\zeta$ , and  $\tilde{r} \in L^\infty$ ,

$$p_k \rightarrow p, \quad q_{z_k} \rightarrow q_z, \quad q_{w_k} \rightarrow q_w, \quad \int_0^t \zeta_k ds \rightarrow \int_0^t \zeta ds \quad \text{uniformly,}$$

$\dot{p}_k \rightarrow \dot{p}$ ,  $\dot{q}_{z_k} \rightarrow \dot{q}_z$ ,  $\dot{q}_{w_k} \rightarrow \dot{q}_w$ ,  $\zeta_k \rightarrow \zeta$  weakly in  $L^1$ , and  $\tilde{r}_k \rightarrow \tilde{r}$  weakly\* in  $L^\infty$ . Taking into account Lemma 4.1(g) we deduce that, for any measurable set  $B \subset T$ ,

$$0 = \int_B \tilde{r}_k(t) \cdot G_k(t) dt = \int_B \tilde{r}_k(t) \cdot \{G_k(t) - \bar{G}(t)\} dt + \int_B \tilde{r}_k(t) \cdot \bar{G}(t) dt$$

and  $\int_B \tilde{r}_k(t) dt \leq 0$ , where

$$G_k(t) = G(t, x_k(t), z_k(t), u_k(t), v_k(t), \alpha_k(t)), \quad \bar{G}(t) = G(t, \bar{x}(t), \bar{z}(t), \bar{u}(t), \bar{v}(t), \bar{\alpha}(t)).$$

By taking limits, we conclude that  $\tilde{r}(t) \leq 0$  and  $\tilde{r}(t) \cdot \bar{G}(t) = 0$  a.e. in  $T$ . Also, a straightforward modification of the proof of [3, Theorem 3.1.7] and an appeal to the upper semicontinuity properties of limiting normal cones and subdifferentials allow us to pass to the limit in the relationships (a)–(f) of Lemma 4.1. There results

$$(A') \quad \lambda + \|p\|_\infty + \|q_z\|_\infty + \|q_w\|_\infty > 0,$$

$$(B') \quad (-\dot{p}(t), -\dot{q}_z(t), -\dot{q}_w(t), 0, \zeta(t), 0)$$

$$\in \text{co } \partial \bar{H}(t, \bar{x}(t), \bar{z}(t), \bar{w}(t), p(t), q_z(t), q_w(t), r(t), \bar{u}(t), \bar{v}(t), \bar{\alpha}(t)) \quad \text{a.e.,}$$

$$(C') \quad \zeta(t) \in \text{co } N_{V(t)}(\bar{v}(t)) \quad \text{a.e.,}$$

$$(D') \quad (p(0), -p(1)) \in N_C(\bar{x}(0), \bar{x}(1)) + \lambda \partial l(\bar{x}(0), \bar{x}(1)),$$

$$(E') \quad |q_w(1)| = |q_z(1)|, \quad r(t) \leq 0, \quad \text{and} \quad r(t) \cdot G(t, \bar{x}(t), \bar{z}(t), \bar{u}(t), \bar{v}(t), \bar{\alpha}(t)) = 0 \quad \text{a.e.,}$$

where  $r(t) = (1 + K_r(t))\tilde{r}(t)$ .

*Step 6.* Rewriting relations (A')–(E') in the required form.

Recall that

$$\begin{aligned} \tilde{H}(t, x, z, w, p, q_z, q_w, r, u, v, \alpha) \\ = p \cdot f(t, x, u, v) + q_z \cdot (w - \alpha) + q_w \cdot \alpha^2 + r \cdot (g(t, x, u, v) + z + \alpha). \end{aligned}$$

We are now in a position to get an estimation for  $\text{co } \partial \tilde{H}$ . We have

$$\begin{aligned} \text{co } \partial \tilde{H}(t, x, z, w, p, q_z, q_w, r, u, v, \alpha) \subset \{(\theta_1 + r \nabla_x g, r, q_z, \theta_2 + r \nabla_u g, \\ \theta_3 + r \nabla_v g, -q_z + 2q_w \alpha + r) \mid (\theta_1, \theta_2, \theta_3) \in \text{co } \partial_{x,u,v} p \cdot f\}. \end{aligned}$$

Since  $\bar{\alpha}(t) = 0$  and  $\bar{w}(t) = 0$  a.e. in  $T$ , by appealing to an appropriate selection theorem, we deduce the existence of measurable functions  $\theta_1$ ,  $\theta_2$ ,  $\theta_3$ , and  $\zeta$  satisfying

$$(4.18) \quad (\theta_1(t), \theta_2(t), \theta_3(t)) \in \text{co } \partial_{x,u,v} p(t) \cdot f(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) \quad \text{a.e. in } T, \\ \zeta(t) \in N_{V(t)}(\bar{v}(t)) \quad \text{a.e. in } T.$$

We conclude from the above that

$$(4.19) \quad (-\dot{p}(t), -\dot{q}_z(t), -\dot{q}_w(t), 0, \zeta(t), 0) = (\theta_1(t) + r(t) \nabla_x \bar{g}(t), r(t), q_z(t), \\ \theta_2(t) + r(t) \nabla_u \bar{g}(t), \theta_3(t) + r(t) \nabla_v \bar{g}(t), r(t) - q_z(t)).$$

From (4.19) and (E') we deduce that, a.e. in  $T$ ,

$$(4.20) \quad q_z(t) = r(t), \quad \dot{q}_z(t) = -r(t), \quad \dot{q}_w(t) = -q_z(t)$$

and  $|q_w(1)| = |q_z(1)|$ . Set

$$H(t, x, p, r, u, v) := p \cdot f(t, x, u, v) + r \cdot g(t, x, u, v).$$

From the above and (4.19) we have

$$(4.21) \quad (-\dot{p}(t), 0, \zeta(t)) = (\theta_1(t) + r(t) \nabla_x \bar{g}(t), \theta_2(t) + r(t) \nabla_u \bar{g}(t), \theta_3(t) + r(t) \nabla_v \bar{g}(t))$$

for some  $(\theta_1(t), \theta_2(t), \theta_3(t)) \in \text{co } \partial_{x,u,v} p \cdot f(t, \bar{x}(t), \bar{u}(t), \bar{v}(t))$  a.e. in  $T$ . It follows that

$$(4.22) \quad (-\dot{p}(t), 0, \zeta(t)) \in \text{co } \partial_{x,u,v} H(t, \bar{x}(t), p(t), r(t), \bar{u}(t), \bar{v}(t)) \quad \text{a.e. in } T.$$

Taking into account (E'), we also have

$$(4.23) \quad r(t) \cdot g(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) = 0 \quad \text{and} \quad r(t) \leq 0 \quad \text{a.e. in } T.$$

We deduce that (4.22), (C'), (4.23), and (D') are, respectively, (ii), (iii), (iv), and (v) of the theorem.

It remains to prove (i). We claim that

$$(4.24) \quad \lambda + |p(t)| > 0 \quad \text{for all } t \in T.$$

First observe that if  $q_z \equiv 0$  and  $q_w \equiv 0$ , then (4.24) holds. Seeking a contradiction assume that  $\lambda = 0$  and  $p(\tau) = 0$  for some  $\tau \in T$ . There are two cases to consider.

*Case 1.* Suppose that  $r(t) = 0$  a.e. Then by (4.20) we have  $q_z \equiv 0$ ,  $q_w \equiv 0$ . We deduce from (4.21) and (H1) the existence of  $K_1$  integrable with  $K_1(t) \geq 0$  a.e. such that  $|\dot{p}(t)| \leq K_1(t)|p(t)|$  a.e. Gronwall's inequality (see [25]) and the fact that  $p(\tau) = 0$

for some  $\tau \in T$  imply that  $p \equiv 0$ . But this, together with the fact that  $\lambda = 0$  and  $q_z \equiv 0$ ,  $q_w \equiv 0$ , and  $p \equiv 0$ , contradicts (A').

*Case 2.* Suppose now that  $r(t) \neq 0$  on a subset of  $T$  of Lebesgue measure different from zero. We deduce from (4.21) that  $0 = \theta_2(t) + r(t) \cdot \nabla_u \bar{g}(t)$  a.e. in  $T$ , where  $\theta_2(t)$  is defined in (4.18). Taking the inner product of the left-hand side of the previous equation and the vector  $h$  (defined in (H7)) we obtain, from (H7) and (iv) of the theorem, that  $\theta_2(t) \cdot h(t) = -r(t) \cdot a(t)$  a.e. in  $T$ . By hypotheses, there exists  $K_2$  integrable with  $K_2(t) \geq 0$  a.e. such that

$$(4.25) \quad |r(t)| \leq K_2(t)|p(t)| \quad \text{a.e. in } T.$$

We deduce from (4.21), (4.25), and (H1) the existence of  $K_3$  integrable with  $K_3(t) \geq 0$  a.e. such that  $|\dot{p}(t)| \leq K_3(t)|p(t)|$  a.e. in  $T$ . Again Gronwall's inequality and the fact that  $p(\tau) = 0$  for some  $\tau \in T$  imply that  $p \equiv 0$ . But then, by (4.25), we have  $r(t) = 0$  a.e., a contradiction.

We proved that  $\lambda + |p(t)| > 0$  for all  $t \in T$ , a condition which ensures that  $\lambda + \|p\|_\infty > 0$ . The proof is complete.  $\square$

**5. Proof of Theorem 3.2.** Consider the function  $\mu: T \times (\mathbf{R}^n \times \mathbf{R}^{k_u} \times \mathbf{R}^{k_v}) \times \mathbf{R}^{m_b} \rightarrow \mathbf{R}^{m_b}$  given by

$$\mu(t, (\xi, u, v), \eta) := b(t, \bar{x}(t) + \xi, \bar{u}(t) + u + \nabla_u \bar{b}(t)^T \eta, \bar{v}(t) + v).$$

We have  $\mu(t, (0, 0, 0), 0) = 0$  a.e. in  $T$ . Choose  $S_0 \subset T$  to be the largest subset such that this relation and each of the hypotheses do not hold for every  $t \in S_0$ . By assumption,  $S_0$  is of Lebesgue measure zero. Thus, there exists a Borel set  $S_1$ , which is the intersection of a countable collection of open sets, such that  $S_0 \subset S_1$  and  $S_1 \setminus S_0$  has measure zero. Thus  $S_1$  is a Borel set of measure zero. We define  $S := T \setminus S_1$ , a Borel set of full measure. We have

$$\frac{\partial \mu}{\partial \eta}(t, (0, 0, 0), 0) = \Gamma(t) := \nabla_u \bar{b}(t) \nabla_u \bar{b}(t)^T \quad \text{for all } t \in S.$$

By (H6) and (H7)(iv) there exists a constant  $M > 0$  such that  $|\Gamma(t)|^{-1} \leq M$  for all  $t \in S$ . Thus Proposition 2.2 applies to the function  $\mu$ , and it asserts the existence of  $\epsilon_1 \in (0, \epsilon)$ ,  $\delta_1 \in (0, \epsilon)$ , and an implicit map  $d: T \times \epsilon_1 B \times \epsilon_1 B \times \epsilon_1 B \rightarrow \delta_1 B$  such that  $d(\cdot, \xi, u, v)$  is a measurable function for fixed  $(\xi, u, v)$ , the functions  $\{d(t, \cdot, \cdot, \cdot) \mid t \in S\}$  are Lipschitz continuous with common Lipschitz constant  $K_d > 0$ , and  $d(t, \cdot, \cdot, \cdot)$  is continuously differentiable for fixed  $t \in S$ . Choose  $\sigma_1, \delta > 0$  such that

$$(5.1) \quad \sigma_1 \in (0, \min\{\epsilon_1, \epsilon/2\}), \quad \delta \in (0, \min\{\delta_1, \epsilon/2\}), \quad \sigma_1 + K_{b,g}\delta \in (0, \epsilon/2),$$

where  $\epsilon_1$  and  $\delta_1$  (which do not depend on  $t$ ) are as above and  $K_{b,g}$  is given by (H6).

In what follows, and without loss of generality, we consider the implicit function  $d$  defined on  $T \times \sigma_1 B \times \sigma_1 B \times \sigma_1 B$  and taking values in  $\delta B$ . The function  $d(t, \cdot, \cdot, \cdot)$  is continuously differentiable and Lipschitz continuous on  $\sigma_1 B \times \sigma_1 B \times \sigma_1 B$  with Lipschitz constant  $K_d$ , and it satisfies  $d(t, 0, 0, 0) = 0$  a.e. in  $T$ :

$$\mu(t, (\xi, u, v), d(t, \xi, u, v)) = 0 \quad \text{a.e. in } T, \quad \text{for all } (\xi, u, v) \in \sigma_1 B \times \sigma_1 B \times \sigma_1 B,$$

$$(d_\xi, d_u, d_v)(t, 0, 0, 0) = -\Gamma(t)^{-1}(\nabla_x \bar{b}(t), \nabla_u \bar{b}(t), \nabla_v \bar{b}(t)) \quad \text{a.e. in } T.$$

Define the functions

$$\begin{aligned} D(t, x, u, v) &:= d(t, x - \bar{x}(t), u - \bar{u}(t), v - \bar{v}(t)), \\ F(t, x, u, v) &:= f(t, x, u + \nabla_u \bar{b}(t)^T D(t, x, u, v), v), \\ G(t, x, u, v) &:= g(t, x, u + \nabla_u \bar{b}(t)^T D(t, x, u, v), v) \end{aligned}$$

and the sets  $\mathcal{U}(t) := \bar{u}(t) + \sigma_1 B$ ,  $\mathcal{V}(t) := V(t) \cap (\bar{v}(t) + \sigma_1 B)$ . Consider the problem

$$(P_{\text{aux}}) \quad \begin{cases} \text{Minimize } l(x(0), x(1)) \text{ subject to} \\ \dot{x}(t) = F(t, x(t), u(t), v(t)) & \text{a.e. in } T, \\ 0 \geq G(t, x(t), u(t), v(t)) & \text{a.e. in } T, \\ (u(t), v(t)) \in \mathcal{U}(t) \times \mathcal{V}(t) & \text{a.e. in } T, \\ (x(0), x(1)) \in C. \end{cases}$$

The process  $(\bar{x}, \bar{u}, \bar{v})$  is admissible for  $(P_{\text{aux}})$ . Assume that  $(\tilde{x}, \tilde{u}, \tilde{v})$  is a solution with lesser cost. Set

$$\begin{aligned} \hat{u}(t) &:= \tilde{u}(t) + \nabla_u \bar{b}(t)^T D(t, \tilde{x}(t), \tilde{u}(t), \tilde{v}(t)), \\ \xi(t) &:= \tilde{x}(t) - \bar{x}(t), \quad u_1(t) := \tilde{u}(t) - \bar{u}(t), \quad v_1(t) := \tilde{v}(t) - \bar{v}(t). \end{aligned}$$

From (5.1) and the definition of  $d$  it follows that

$$|\hat{u}(t) - \bar{u}(t)| \leq |\tilde{u}(t) - \bar{u}(t)| + K_{b,g} \delta \leq \sigma_1 + K_{b,g} \delta < \epsilon, \quad |\tilde{v}(t) - \bar{v}(t)| \leq \sigma_1 < \epsilon.$$

By definition of  $d$ , for almost all  $t \in T$  we have

$$\mu(t, (\xi(t), u_1(t), v_1(t)), d(t, \xi(t), u_1(t), v_1(t))) = b(t, \tilde{x}(t), \hat{u}(t), \tilde{v}(t)) = 0.$$

We conclude that  $(\tilde{x}, \hat{u}, \tilde{v})$  is a solution to  $(P)$  with lesser cost, contradicting the optimality of  $(\bar{x}, \bar{u}, \bar{v})$ . It follows that  $(\bar{x}, \bar{u}, \bar{v})$  is a minimizer for  $(P_{\text{aux}})$ .

Let us now check that  $(P_{\text{aux}})$  satisfies the conditions under which Theorem 3.1 holds. We need only verify that

$$(5.2) \quad \nabla_u \bar{G}^{\mathcal{I}_a(t)}(t) \cdot h(t) = a^{\mathcal{I}_a(t)}(t) \quad \text{a.e. in } T,$$

where  $h$  and  $a$  are the functions whose existence is postulated in (H7) and satisfy (H7)(i) and (H7)(ii). Observe that

$$\nabla_u G(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) = \nabla_u \bar{g}(t) - \nabla_u \bar{g}(t) \nabla_u \bar{b}(t)^T \Gamma(t)^{-1} \nabla_u \bar{b}(t).$$

Taking into account (H7), the inner product of the left-hand side of the previous equation and vector  $h(t)$  leads to

$$\nabla_u \bar{G}^{\mathcal{I}_a(t)}(t) \cdot h(t) = \nabla_u \bar{g}^{\mathcal{I}_a(t)}(t) \cdot h(t) = a^{\mathcal{I}_a(t)}(t) \quad \text{a.e. in } T$$

proving (5.2). We now apply Theorem 3.1 to  $(P_{\text{aux}})$ . It asserts the existence of  $\lambda \geq 0$ ,  $p \in W^{1,1}(T; \mathbf{R}^n)$ ,  $r \in L^1(T; \mathbf{R}^{m_g})$ , and  $\zeta \in L^1(T; \mathbf{R}^{k_v})$  such that

- (i)  $\|p\|_\infty + \lambda \neq 0$ ,
- (ii)  $(-\dot{p}(t), 0, \zeta(t)) \in \text{co } \partial_{x,u,v} \tilde{H}(t, \bar{x}(t), p(t), r(t), \bar{u}(t), \bar{v}(t))$  a.e. in  $T$ ,
- (iii)  $\zeta(t) \in \text{co } N_{V(t)}(\bar{v}(t))$  a.e. in  $T$ ,
- (iv)  $r(t) \cdot G(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)) = 0$  and  $r(t) \leq 0$  a.e. in  $T$ ,
- (v)  $(p(0), -p(1)) \in N_C(\bar{x}(0), \bar{x}(1)) + \lambda \partial l(\bar{x}(0), \bar{x}(1))$ ,

where  $\tilde{H}(t, x, p, r, u, v) = p \cdot F(t, x, u, v) + r \cdot G(t, x, u, v)$ . From the nonsmooth chain rule (see [25]), the differentiability properties of  $d$ , and an appropriate selection theorem, there exist measurable functions  $\theta_1$ ,  $\theta_2$ ,  $\theta_3$ , and  $\zeta$  satisfying, a.e. in  $T$ ,

$$\begin{aligned} (\theta_1(t), \theta_2(t), \theta_3(t)) &\in \text{co } \partial_{x,u,v} p(t) \cdot f(t, \bar{x}(t), \bar{u}(t), \bar{v}(t)), \quad \zeta(t) \in \text{co } N_{V(t)}(\bar{v}(t)), \\ (-\dot{p}(t), 0, \zeta(t)) &= (\theta_1(t) + q(t) \nabla_x \bar{b}(t) + r(t) \nabla_x \bar{g}(t), \\ \theta_2(t) + q(t) \nabla_u \bar{b}(t) + r(t) \nabla_u \bar{g}(t) \theta_3(t) + q(t) \nabla_v \bar{b}(t) + r(t) \nabla_v \bar{g}(t)), \end{aligned}$$

where

$$(5.3) \quad q(t) = -(\theta_2(t) + r(t)\nabla_u \bar{g}(t))\nabla_u \bar{b}(t)^T \Gamma(t)^{-1}.$$

Under the hypotheses,  $\theta_1$ ,  $\theta_2$ ,  $\theta_3$ ,  $r$ , and  $\zeta$  are all integrable functions, and so is  $q$ . This proves that  $\lambda$ ,  $p$ ,  $q$ ,  $r$ , and  $\zeta$  satisfy (i)–(iv) of the theorem. Since there exists  $R$  integrable such that  $|r(t)| \leq R(t)|p(t)|$  a.e. in  $T$ , we deduce from (5.3) the existence of  $K_m$  integrable such that the last conclusion of the theorem holds. This completes the proof.  $\square$

**Acknowledgments.** The authors thank Franco Rampazzo for calling their attention to [15], Richard B Vinter for many discussions on the subject, and Vera Zeidan for discussions on the regularity assumption (R3) in [22]. Additionally, the authors express their gratitude to the anonymous referees for useful comments and suggestions.

#### REFERENCES

- [1] A. P. AFANAS'EV, V. V. DIKUSAR, A. A. MILYUTIN, AND S. A. CHUKANOV, *A Necessary Condition in Optimal Control*, Nauka, Moscow, 1990 (in Russian).
- [2] A. V. ARUTYUNOV, *Optimality Conditions. Abnormal and Degenerate Problems*, Math. Appl. 526, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.
- [3] F. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983. (Reprinted as Classics Applied Math. 5, SIAM, Philadelphia, 1990.)
- [4] F. CLARKE, *The maximum principle in optimal control, then and now*, Control Cybernet., 34 (2005), pp. 709–722.
- [5] M. R. DE PINHO, *Mixed constrained control problems*, J. Math. Anal. Appl., 278 (2003), pp. 293–307.
- [6] M. R. DE PINHO, M. M. A. FERREIRA, AND F. A. C. C. FONTES, *Unmaximized inclusion necessary conditions for nonconvex constrained optimal control problems*, ESAIM Control Optim. Calc. Var., 11 (2005), pp. 614–632.
- [7] M. R. DE PINHO AND A. ILCHMANN, *Weak maximum principle for optimal control problems with mixed constraints*, Nonlinear Anal., 48 (2002), pp. 1179–1196.
- [8] M. R. DE PINHO AND R. B. VINTER, *An Euler-Lagrange inclusion for optimal control problems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1191–1198.
- [9] M. R. DE PINHO AND R. B. VINTER, *Necessary conditions for optimal control problems involving nonlinear differential algebraic equations*, J. Math. Anal. Appl., 212 (1997), pp. 493–516.
- [10] M. R. DE PINHO, R. B. VINTER, AND H. ZHENG, *A maximum principle for optimal control problems with mixed constraints*, IMA J. Math. Control Inform., 18 (2001), pp. 189–205.
- [11] E. N. DEVDARIANI AND Y. S. LEDYAEV, *Maximum principle for implicit control systems*, Appl. Math. Optim., 40 (1999), pp. 79–103.
- [12] V. V. DIKUSAR AND A. A. MILYUTIN, *Qualitative and Numerical Methods in the Maximum Principle*, Nauka, Moscow, 1989 (in Russian).
- [13] A. V. DMITRUK, *Maximum principle for the general optimal control problem with phase and regular mixed constraints*, Comput. Math. Model., 4 (1993), pp. 364–377.
- [14] A. Y. DUBOVITSKIĬ AND A. A. MILYUTIN, *Theory of the principle of the maximum*, in Methods of the Theory of Extremal Problems in Economics, Nauka, Moscow, 1981, pp. 7–47 (in Russian).
- [15] H. FRANKOWSKA AND F. RAMPAZZO, *Relaxation of control systems under state constraints*, SIAM J. Control Optim., 37 (1999), pp. 1291–1309.
- [16] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [17] A. A. MILYUTIN AND N. P. OSMOLOVSKIĬ, *Calculus of Variations and Optimal Control*, Transl. Math. Monogr. 180, AMS, Providence, RI, 1998.
- [18] B. S. MORDUKHOVICH, *Variational Analysis and Generalized Differentiation I. Basic Theory*, Grundlehren Math. Wiss. 330, Springer-Verlag, Berlin, 2006.
- [19] B. S. MORDUKHOVICH, *Variational Analysis and Generalized Differentiation II. Applications*, Grundlehren Math. Wiss. 331, Springer-Verlag, Berlin, 2006.
- [20] L. W. NEUSTADT, *Optimization. A Theory of Necessary Conditions*, Princeton University Press, Princeton, NJ, 1976.

- [21] Z. PÁLES AND V. ZEIDAN, *First- and second-order necessary conditions for control problems with constraints*, Trans. Amer. Math. Soc., 346 (1994), pp. 421–453.
- [22] Z. PÁLES AND V. ZEIDAN, *Optimal control problems with set-valued control and state constraints*, SIAM J. Optim., 14 (2003), pp. 334–358.
- [23] R. T. ROCKAFELLAR AND R. J-B. WETS, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1998.
- [24] G. STEFANI AND P. ZEZZA, *Optimality conditions for a constrained control problem*, SIAM J. Control Optim., 34 (1996), pp. 635–659.
- [25] R. B. VINTER, *Optimal Control*, Systems Control Found. Appl., Birkhäuser, Boston, 2000.

## CONTROL OF A TIP-FORCE DESTABILIZED SHEAR BEAM BY OBSERVER-BASED BOUNDARY FEEDBACK\*

MIROSLAV KRSTIC<sup>†</sup>, BAO-ZHU GUO<sup>‡</sup>, ANDRAS BALOGH<sup>§</sup>, AND  
ANDREY SMYSHLYAEV<sup>¶</sup>

**Abstract.** We consider a model of the undamped shear beam with a destabilizing boundary condition. The motivation for this model comes from atomic force microscopy, where the tip of the cantilever beam is destabilized by van der Waals forces acting between the tip and the material surface. Previous research efforts relied on collocated actuation and sensing at the tip, exploiting the passivity property between the corresponding input and output in the beam model. In this paper we design a stabilizing output-feedback controller in a noncollocated setting, with measurements at the free end (tip) of the beam and actuation at the beam base. Our control design is a novel combination of the classical “damping boundary feedback” idea with a recently developed backstepping approach. A change of variables is constructed which converts the beam model into a wave equation (for a very short string) with boundary damping. This approach is physically intuitive and allows both an elegant stability analysis and an easy selection of design parameters for achieving desired performance. Our observer design is a dual of the similar ideas, combining the damping feedback with backstepping, adapted to the observer error system. Both stability and well-posedness of the closed-loop system are proved. The simulation results are presented.

**Key words.** distributed parameter systems, shear beam, backstepping, stabilization, boundary control

**AMS subject classifications.** 35J05, 93B07, 93D15, 93B52, 93B60

**DOI.** 10.1137/060676969

**1. Introduction.** Flexible beams constitute an important benchmark problem in many application areas ranging from aerospace to civil structures. In some of the exciting modern fields such as atomic force microscopy, the cantilever beam is more than just a prototype problem and constitutes an important application topic in its own right.

In this paper we consider a model of the undamped shear beam [3] with a destabilizing boundary condition. It consists of a wave equation coupled with a second-order-in-space ODE or can be alternatively represented as a fourth-order-in-space/second-order-in-time PDE. This makes it more complex than the Euler–Bernoulli model [3], similar in structure to the Rayleigh beam model [3], and slightly simpler than the Timoshenko model [3]. The destabilizing boundary condition is motivated by the physics of the atomic force microscopy (AFM), where the tip of the cantilever beam is destabilized by van der Waals forces acting between the tip and the material surface [19].

---

\*Received by the editors December 7, 2006; accepted for publication (in revised form) July 1, 2007; published electronically February 6, 2008. This work was supported by the National Science Foundation and by the Los Alamos National Laboratory.

<http://www.siam.org/journals/sicon/47-2/67696.html>

<sup>†</sup>Corresponding author. Department of Mechanical and Aerospace Engineering, University of California at San Diego, La Jolla, CA 92093 (krstic@ucsd.edu).

<sup>‡</sup>Academy of Mathematics and Systems Science, Academia Sinica, Beijing 100080, People’s Republic of China, and School of Computational and Applied Mathematics, University of the Witwatersrand, Wits 2050, Johannesburg, South Africa (bzguo@iss.ac.cn).

<sup>§</sup>Department of Mathematics, University of Texas – Pan American, Edinburg, TX 78541 (abalogh@utpa.edu).

<sup>¶</sup>Department of Mechanical and Aerospace Engineering, University of California at San Diego, La Jolla, CA 92093 (asmyshly@ucsd.edu).

Extensive literature exists on control of beam models [1, 2, 8, 4, 7, 9, 11, 17, 20, 21]. However, previous research efforts all relied on collocated actuation and sensing at the tip, exploiting in an elegant way the passivity property between the corresponding input and output in the beam model. The main drawback of this approach is that the tip of the beam is not a very convenient place to put an actuator. Therefore such feedbacks are usually implemented via passive dampers or through rather elaborate ways, such as electromagnets or small airjets at the tip of the beam.

Our objective is different—to design controllers implementable through noncollocated architecture, with actuation only at the base and sensing only at the tip of the beam. This architecture makes active control more readily implementable to several applications; for example, in AFM this allows a natural use of piezo actuation at the base of the beam. Our control design is a novel combination of the classical “damping boundary feedback” idea with a recently developed backstepping approach, which has been used to design boundary controllers [12, 14] and observers [13] for parabolic equations and for the Timoshenko beam model with a small amount of Kelvin–Voigt damping [6]. A change of variables is constructed which converts the beam model into a wave equation (for a very short string) with boundary damping. This approach is physically intuitive and allows both an elegant stability analysis and an easy selection of design parameters for achieving desired performance. Our observer design is a dual of the similar ideas, combining the damping feedback with backstepping, adapted to the observer error system.

In addition to rigorous stability and well-posedness analysis of the closed-loop system, we also present the results of simulations that illustrate the performance of the controller.

**2. Model.** The shear beam model can be represented in a number of equivalent ways [3]. One often used form is a single second-order-in-time fourth-order-in-space PDE

$$(2.1) \quad aw_{tt}(x, t) - \delta w_{xxtt}(x, t) + w_{xxxx}(x, t) = 0, \quad 0 < x < 1, t > 0,$$

with  $a, \delta > 0$  and two “free end” boundary conditions  $\delta w_{tt}(0, t) = w_{xx}(0, t)$  and  $\delta w_{xtt}(0, t) = w_{xxx}(0, t) - q_\delta w(0, t)$ .

We, however, will use another common form of this model, consisting of a wave equation coupled with a second-order ODE:

$$(2.2) \quad \left\{ \begin{array}{l} \delta w_{tt}(x, t) = w_{xx}(x, t) - \alpha_x(x, t), \quad 0 < x < 1, t > 0, \\ w_x(0, t) = \alpha(0, t) - qw(0, t), \quad t \geq 0, \\ w(1, t) = u_1(t), \quad t \geq 0, \\ 0 = \alpha_{xx}(x, t) - b^2 \alpha(x, t) + b^2 w_x(x, t), \quad 0 < x < 1, t > 0, \\ \alpha_x(0, t) = 0, \quad t \geq 0, \\ \alpha(1, t) = u_2(t), \quad t \geq 0, \\ y(t) = (w(0, t), \alpha(0, t)), \quad t \geq 0, \end{array} \right.$$

which is obtained from (2.1) by introducing a new state  $\alpha_x = w_{xx} - \delta w_{tt}$  and denoting  $b = \sqrt{a/\delta}$ . The state  $w$  represents the transversal displacement of the beam, and  $\alpha$  is the angle due to bending. The objective is to use the control input  $u(t) = (u_1(t), u_2(t))$  at the base of the beam to stabilize the tip of the beam with the measurement  $y(t)$  available only at the free end.



It is important to note the term  $-qw(0, t)$  in the boundary condition (2.2). This type of boundary condition corresponds to situations where the tip of the beam is subject to an external force which depends on the displacement. Such a force arises in AFM as a van der Waals force acting between the atoms on the material surface and the beam tip. The term  $-qw(0, t)$  is the linearized model of that force; the original nonlinear model has cubic nonlinearity [19]. Typically  $q > 0$  has a destabilizing effect (in that case, one can think of this parameter as “antistiffness”), whereas  $q < 0$  has a stabilizing effect. We stress that this force occurs on the opposite end of the beam from where the actuator is located. If the actuator were at the tip, canceling the effect of this force would be trivial. In the configuration that we pursue here, stabilization, and even vibration suppression when  $q = 0$ , is a nontrivial problem.

We will first present control and observer designs separately to make the ideas clear and then prove the certainty equivalence principle and well-posedness of the closed-loop system in sections 5–7.

**3. Controller.** In order to proceed with the control design we first need to write the model (2.2) in yet another form. To this end, we solve the ODE part of (2.2) as a two point boundary value problem for  $\alpha$  with boundary condition  $\alpha_x(0, t) = 0$ :

$$(3.1) \quad \alpha(x, t) = \cosh(bx)\alpha(0, t) - b \int_0^x \sinh(b(x-s))w_x(s, t) ds.$$

Setting  $x = 1$  in (3.1) and using the boundary condition  $\alpha(1, t) = u_2(t)$ , we can express  $\alpha(0, t)$  in terms of  $w$  and  $u_2$ :

$$(3.2) \quad \alpha(0, t) = \frac{1}{\cosh(b)}u_2(t) + \frac{b}{\cosh(b)} \int_0^1 \sinh(b(1-s))w_x(s, t) ds.$$

Next, we differentiate (3.1) in  $x$  and substitute the result into the first equation of (2.2). This way, instead of a wave equation coupled with a second-order ODE, we obtain a single hyperbolic partial integrodifferential equation for  $w$ :

$$(3.3) \quad \left\{ \begin{array}{l} \delta w_{tt}(x, t) = w_{xx}(x, t) - b^2 \cosh(bx)w(0, t) + b^3 \int_0^x \sinh(b(x-y))w(y, t) dy \\ \quad + b^2 w(x, t) - \frac{b \sinh(bx)}{\cosh(b)} \left[ u_2(t) + b \int_0^1 \sinh(b(1-s))w_x(s, t) ds \right], \\ w_x(0, t) = \frac{1}{\cosh(b)} \left[ u_2(t) + b \int_0^1 \sinh(b(1-s))w_x(s, t) ds \right] - qw(0, t), \\ w(1, t) = u_1(t). \end{array} \right.$$

Since the backstepping control design [12] needs the PDE to be in a “strict-feedback” form (in other words, its right-hand side must be “causal” in  $x$ ), we are going to use the control  $u_2(t)$  to cancel the definite integral both in the domain and in the boundary condition:

$$(3.4) \quad u_2(t) = -b \int_0^1 \sinh(b(1-s))w_x(s, t) ds.$$

We get the following PDE:

$$(3.5) \quad \begin{cases} \delta w_{tt}(x, t) = w_{xx}(x, t) + b^2 w(x, t) - b^2 \cosh(bx)w(0, t) \\ \quad + b^3 \int_0^x \sinh(b(x-y))w(y, t)dy, \\ w_x(0, t) = -qw(0, t), \\ w(1, t) = u_1(t). \end{cases}$$

The basic idea of the backstepping design is to use the transformation

$$(3.6) \quad \bar{w}(x, t) = w(x, t) - \int_0^x k(x, y)w(y, t) dy,$$

with specially designed control kernel  $k(x, y)$  along with the boundary feedback law

$$(3.7) \quad \begin{cases} u_1(t) = w(1, t), \\ w_x(1, t) = k(1, 1)w(1, t) - c_1 w_t(1, t) \\ \quad + c_1 \int_0^1 k(1, y)w_t(y, t) dy + \int_0^1 k_x(1, y)w(y, t) dy \end{cases}$$

to map (3.5) into the exponentially stable target system

$$(3.8) \quad \begin{cases} \delta \bar{w}_{tt}(x, t) = \bar{w}_{xx}(x, t), \\ \bar{w}_x(0, t) = c_0 \bar{w}(0, t), \\ \bar{w}_x(1, t) = -c_1 \bar{w}_t(1, t), \end{cases}$$

where  $c_0 > 0$  and  $c_1 > 0$  are design parameters. The system (3.8) is exponentially stable at the origin iff  $c_0$  and  $c_1$  are positive. Note the crucial difference between the second equations of (3.5) and (3.8)—the destabilizing negative sign in the former and the stabilizing positive sign in the latter. The gain kernel  $k(x, y)$  is given by the following PDE:

$$(3.9) \quad \begin{cases} k_{xx}(x, y) = k_{yy}(x, y) + b^2 k(x, y) - b^3 \sinh(b(x-y)) \\ \quad + b^3 \int_y^x k(x, \xi) \sinh(b(\xi-y))d\xi, \\ k(x, x) = -\frac{b^2}{2}x - c_0 - q, \\ k_y(x, 0) = -b^2 \left[ \cosh(bx) - \int_0^x k(x, y) \cosh(by)dy \right] - qk(x, 0), \end{cases}$$

which is obtained by substituting (3.6) into (3.8) and matching the terms. Incidentally, this equation for  $k(x, y)$  is in the same class as the one obtained in the control design for parabolic PDEs [12]. As shown in [12], the PDE (3.9) has a unique solution  $k \in C^2(\Omega)$ .

It can be solved either numerically or by using the following symbolic recursion:

$$(3.10) \quad \left\{ \begin{array}{l} k(x, y) = \lim_{n \rightarrow \infty} k_n(x, y), \\ k_0 = -\frac{b}{2}[-\sinh(b(x-y)) + by \cosh(b(x-y))] - c_0 - q, \\ k_{n+1} = k_0 + b^2 \int_{\frac{x-y}{2}}^{\frac{x+y}{2}} \int_0^{\frac{x-y}{2}} k_n(\sigma+s, \sigma-s) ds d\sigma + q \int_0^{x-y} k_n(\sigma, 0) d\sigma \\ \quad + b^2 \int_0^{\frac{x-y}{2}} \int_0^\sigma [2k_n(\sigma+s, \sigma-s) - k_n(\sigma, s) \cosh(bs)] ds d\sigma \\ \quad + b^3 \int_{\frac{x-y}{2}}^{\frac{x+y}{2}} \int_0^{\frac{x-y}{2}} \int_{\sigma-s}^{\sigma+s} k_n(\sigma+s, \xi) \sinh(b(\xi-\sigma+s)) d\xi ds d\sigma \\ \quad + 2b^3 \int_0^{\frac{x-y}{2}} \int_0^\sigma \int_{\sigma-s}^{\sigma+s} k_n(\sigma+s, \xi) \sinh(b(\xi-\sigma+s)) d\xi ds d\sigma. \end{array} \right.$$

The first step of this recursion provides approximate control gain kernels, which are explicit:

$$(3.11) \quad \left\{ \begin{array}{l} k_0(1, y) = -\frac{b}{2}[-\sinh(b(1-y)) + by \cosh(b(1-y))] - c_0 - q, \\ k_{0x}(1, y) = -\frac{b}{2}[-\cosh(b(1-y)) + by \sinh(b(1-y))], \\ k_0(1, 1) = k(1, 1) = -\frac{b^2}{2} - c_0 - q. \end{array} \right.$$

Since  $k \in C^2(\Omega)$ , the transformation (3.6) is bounded invertible, and therefore the system (3.5) with the controller (3.7) dynamically behaves as (3.8). The important question is why we chose our system's "target" behavior as in (3.8). This PDE with a homogeneous Dirichlet boundary condition at  $x = 0$  (i.e., for  $c_0 = \infty$ ) has been studied in many papers on control of wave equations by "boundary damper" feedback. For a large positive  $c_0$ , our "target" system has a similar behavior to those well-studied problems. Obviously, the most desirable behavior would be with  $c_0 = 0$ ; however, such behavior is achievable only if one could put an actuator at the tip. In that case, the end  $x = 0$  would be clamped, and the end  $x = 1$  would be actuated with a "boundary damper." Since we are pursuing the opposite problem, where the tip end  $x = 0$  is free and the actuator is at the opposite end  $x = 1$ , it is only through the very sophisticated construction that we presented above that a behavior similar to the boundary damper feedback is achievable. The plant boundary condition at  $x = 0$  is of Robin type, and no state transformation can change it into Dirichlet. However, we can change it into a Robin condition of favorable sign ( $c_0 > 0$ ) and make it behave similar to a Dirichlet condition (with large  $c_0$ ). To achieve all of this, we construct the change of variable (3.6) which starts at  $x = 0$  and goes towards  $x = 1$ , collecting all of the terms in the shear beam model and converting them into a wave equation model. But it is ultimately the boundary feedback (3.7) that absorbs the effects of the transformation and results in a damping boundary condition at the end  $x = 1$ . Clearly such feedback has to be rather complicated because it achieves a similar effect as the boundary damper but from the opposite end. In addition to the first two terms on the right-hand side of (3.7), which arise in boundary dampers and essentially amount to PD control, our feedback law incorporates the two integral operators acting on

the displacement and velocity fields (as we will show in the next section, the direct measurement of  $w(x, t)$  and  $w_t(x, t)$  along the whole beam is not necessary).

The control law (3.7) has to be implemented by solving for  $w(1, t)$ . In the frequency domain it is equivalent to employing a low pass filter acting on  $w_x(1, s)$  and the integral operator:

$$(3.12) \quad u_1(s) = \frac{1}{c_1 s + \frac{b^2}{2} + c_0 + q} \left[ -w_x(1, s) + \int_0^1 (c_1 s k(1, y) + k_x(1, y)) w(y, s) dy \right].$$

In AFM  $u_1$  is implemented via a piezo actuator which actuates the beam base displacement. Implementation of  $u_2$  would involve two piezo actuators to produce a commanded  $u_2$ .

**4. Observer.** Before we start with the observer design, we write the model (3.3) in a slightly different form using (3.2):

$$(4.1) \quad \begin{cases} \delta w_{tt}(x, t) = w_{xx}(x, t) + b^2 w(x, t) + b^3 \int_0^x \sinh(b(x-y)) w(y, t) dy \\ \quad - b^2 \cosh(bx) w(0, t) - b \sinh(bx) \alpha(0, t), \\ w_x(0, t) = \alpha(0, t) - q w(0, t), \\ w(1, t) = u_1(t). \end{cases}$$

Note that in this form  $u_2$  is not selected as in (3.4) and the observer is designed for arbitrary inputs  $u_1$  and  $u_2$ .

We assume that the only available measurements are of the tip displacement  $w(0, t)$  and of the tip angle due to bending  $\alpha(0, t)$ . In AFM the displacement and the slope of the tip are routinely measured using a laser and a photodiode.

The observer is designed along the lines of the design presented in [13] for parabolic systems and follows a standard finite-dimensional approach “copy of the plant plus output injection terms”:

$$(4.2) \quad \begin{cases} \delta \hat{w}_{tt}(x, t) = \hat{w}_{xx}(x, t) + b^2 \hat{w}(x, t) + b^3 \int_0^x \sinh(b(x-y)) \hat{w}(y, t) dy \\ \quad - b^2 \cosh(bx) w(0, t) - b \sinh(bx) \alpha(0, t) \\ \quad + p_y(x, 0)[w(0, t) - \hat{w}(0, t)] - c_2 p(x, 0)[w_t(0, t) - \hat{w}_t(0, t)], \\ \hat{w}_x(0, t) = \alpha(0, t) - q w(0, t) + p(0, 0)[w(0, t) - \hat{w}(0, t)] \\ \quad - c_2 [w_t(0, t) - \hat{w}_t(0, t)], \\ \hat{w}(1, t) = u_1(t). \end{cases}$$

The constant  $c_2 > 0$  is the design parameter that sets the convergence rate of the observer. Note that the output error terms are injected both in the domain and in the boundary condition. The observer gains  $p(x, 0)$ ,  $p_y(x, 0)$ , and  $p(0, 0)$  in (4.2) are determined by solving the following PDE in  $\Omega = \{(x, y) | 0 \leq y \leq x \leq 1\}$ :

$$(4.3) \quad \begin{cases} p_{yy}(x, y) = p_{xx}(x, y) + b^2 p(x, y) - b^3 \sinh(b(x-y)) \\ \quad + b^3 \int_y^x p(\xi, y) \sinh(b(x-\xi)) d\xi, \\ p(x, x) = \frac{b^2}{2}(x-1), \\ p(1, y) = 0. \end{cases}$$

It has been shown in [13] that this equation has a unique solution  $p \in C^2(\Omega)$ . One can see the similarity between this PDE and the one for the control kernel. This is due to the duality between observer and control designs, a concept well known in finite-dimensional control. One can think of the gains  $p(x, 0)$ ,  $p_y(x, 0)$ , and  $p(0, 0)$  as dual counterparts to the control gains  $k(1, y)$ ,  $k_x(1, y)$ , and  $k(1, 1)$ .

(4.3) can be solved numerically or symbolically using the following recursive procedure [13]:

$$(4.4) \quad \left\{ \begin{array}{l} p(x, y) = \lim_{n \rightarrow \infty} p_n(x, y), \\ p_0 = -\frac{b}{2}[-\sinh(b(x-y)) + b(1-x)\cosh(b(x-y))], \\ p_{n+1} = p_0 + b^2 \int_{\frac{x-y}{2}}^{\frac{2-x-y}{2}} \int_0^{\frac{x-y}{2}} p_n(\sigma+s, \sigma-s) ds d\sigma \\ \quad + 2b^2 \int_0^{\frac{x-y}{2}} \int_0^\sigma p_n(\sigma+s, \sigma-s) ds d\sigma \\ \quad + b^3 \int_{\frac{x-y}{2}}^{\frac{2-x-y}{2}} \int_0^{\frac{x-y}{2}} \int_{\sigma-s}^{\sigma+s} p_n(\sigma+s, \xi) \sinh(b(\xi-\sigma+s)) d\xi ds d\sigma \\ \quad + 2b^3 \int_0^{\frac{x-y}{2}} \int_0^\sigma \int_{\sigma-s}^{\sigma+s} p_n(\sigma+s, \xi) \sinh(b(\xi-\sigma+s)) d\xi ds d\sigma. \end{array} \right.$$

The observer gain in the boundary condition (4.2) is known exactly:

$$(4.5) \quad p(0, 0) = -\frac{b^2}{2}.$$

Let us denote the observer error by  $\varepsilon(x, t) = w(x, t) - \hat{w}(x, t)$ . Using (4.2) and (4.1) we obtain the observer error dynamics

$$(4.6) \quad \left\{ \begin{array}{l} \delta \varepsilon_{tt}(x, t) = \varepsilon_{xx}(x, t) + b^2 \varepsilon(x, t) + b^3 \int_0^x \sinh(b(x-y)) \varepsilon(y, t) dy \\ \quad - p_y(x, 0) \varepsilon(0, t) + c_2 p(x, 0) \varepsilon_t(0, t), \\ \varepsilon_x(0, t) = -p(0, 0) \varepsilon(0, t) + c_2 \varepsilon_t(0, t), \\ \varepsilon(1, t) = 0. \end{array} \right.$$

The convergence of the observer is established by the following lemma.

LEMMA 4.1. *Suppose the classical solution of (4.6) exists. Then the invertible transformation*

$$(4.7) \quad \left\{ \begin{array}{l} \varepsilon(x, t) = \tilde{\varepsilon}(x, t) - \int_0^x p(x, y) \tilde{\varepsilon}(y, t) dy = [(I - \mathbb{P}_1) \tilde{\varepsilon}](x, t), \\ \tilde{\varepsilon}(x, t) = [(I - \mathbb{P}_1)^{-1} \varepsilon](x, t) = \varepsilon(x, t) - \int_0^x p^\ominus(x, y) \varepsilon(y, t) dy \end{array} \right.$$

converts the error system (4.6) into the exponentially stable system

$$(4.8) \quad \left\{ \begin{array}{l} \delta \tilde{\varepsilon}_{tt}(x, t) = \tilde{\varepsilon}_{xx}(x, t), \\ \tilde{\varepsilon}_x(0, t) = c_2 \tilde{\varepsilon}_t(0, t), \\ \tilde{\varepsilon}(1, t) = 0. \end{array} \right.$$

*Proof.* We differentiate the transformation (4.7) with respect to  $t$  and  $x$ :

$$\begin{aligned}
 \delta \varepsilon_{tt}(x, t) &= \delta \tilde{\varepsilon}_{tt}(x, t) - \int_0^x p(x, y) \delta \tilde{\varepsilon}_{tt}(y, t) dy - \varepsilon_{xx}(x, t) + \varepsilon_{xx}(x, t) \\
 &= \delta \tilde{\varepsilon}_{tt}(x, t) - \int_0^x p_{yy}(x, y) \tilde{\varepsilon}(y, t) dy - p(x, x) \tilde{\varepsilon}_x(x, t) + p(x, 0) \tilde{\varepsilon}_x(0, t) \\
 &\quad + p_y(x, x) \tilde{\varepsilon}(x, t) - p_y(x, 0) \tilde{\varepsilon}(0, t) - \tilde{\varepsilon}_{xx}(x, t) + [2p_x(x, x) + p_y(x, x)] \tilde{\varepsilon}(x, t) \\
 &\quad + p(x, x) \tilde{\varepsilon}_x(x, t) + \int_0^x p_{xx}(x, y) \tilde{\varepsilon}(y, t) dy + \varepsilon_{xx}(x, t) \\
 &= \varepsilon_{xx}(x, t) + b^2 \varepsilon(x, t) + c_2 p(x, 0) \varepsilon_t(0, t) - p_y(x, 0) \varepsilon(0, t) \\
 &\quad + \int_0^x (p_{xx}(x, y) - p_{yy}(x, y) + b^2 p(x, y)) \tilde{\varepsilon}(y, t) dy.
 \end{aligned}$$

Using the observer gain PDE (4.3) in the above equation we get the governing equation of (4.8).

Next we differentiate the transformation (4.7) with respect to  $x$  and set  $x = 0$ :

$$\varepsilon_x(0, t) = \tilde{\varepsilon}_x(0) - p(0, 0) \tilde{\varepsilon}(0, t).$$

Comparing this with the boundary condition of (4.6), which can be written as

$$\varepsilon_x(0, t) = -p(0, 0) \tilde{\varepsilon}(0, t) + c_2 \tilde{\varepsilon}_t(0, t),$$

we get the boundary condition of (4.8) at  $x = 0$ . Finally, the boundary condition at  $x = 1$  is obviously satisfied because  $p(1, y) = 0$ .  $\square$

**5. Output feedback.** Consider the observer (4.2) and the control (3.4), (3.7) with the observer state instead of the unmeasured plant state:

$$(5.1) \quad \left\{ \begin{array}{l} u_1(t) = \hat{w}(1, t), \\ \hat{w}_x(1, t) = k(1, 1) \hat{w}(1, t) - c_1 \hat{w}_t(1, t) \\ \quad + c_1 \int_0^1 k(1, y) \hat{w}_t(y, t) dy + \int_0^1 k_x(1, y) \hat{w}(y, t) dy, \\ u_2(t) = -b \int_0^1 \sinh(b(1 - y)) \hat{w}_x(y, t) dy. \end{array} \right.$$

We employ an invertible state transformation

$$(5.2) \quad \left\{ \begin{array}{l} \tilde{w}(x, t) = \hat{w}(x, t) - \int_0^x k(x, y) \hat{w}(y, t) dy = [(I - \mathbb{P}_2) \hat{w}](x, y), \\ \hat{w}(x, t) = [(I - \mathbb{P}_2)^{-1} \tilde{w}](x, y) = \tilde{w}(x, t) - \int_0^x k^\ominus(x, y) \tilde{w}(x, y) dy, \end{array} \right.$$

where  $k(x, y)$  is given by (3.10) and both  $k(x, y)$  and  $k^\ominus(x, y)$  are of  $C^2$  in  $\Omega$  [12].

LEMMA 5.1. *Suppose the classical solution of (4.2) with the control (5.1) exists. Then the transformation (5.2) converts (4.2) and (5.1) into*

$$(5.3) \quad \left\{ \begin{array}{l} \delta \tilde{w}_{tt}(x, t) = \tilde{w}_{xx}(x, t) - b \sinh(bx) \alpha(0, t) + k(x, 0) \alpha(0, t) + p_y(x, 0) \varepsilon(0, t) \\ \quad + k(x, 0) p(0, 0) \varepsilon(0, t) + k_y(x, 0) \varepsilon(0, t) - c_2 k(x, 0) \varepsilon_t(0, t) \\ \quad + b \int_0^x k(x, y) \sinh(by) dy \alpha(0, t) - \int_0^x k(x, y) p_y(y, 0) dy \varepsilon(0, t) \\ \quad - c_2 p(x, 0) \varepsilon_t(0, t) + c_2 \int_0^x k(x, y) p(y, 0) dy \varepsilon_t(0, t), \\ \tilde{w}_x(0, t) = c_0 \tilde{w}(0, t) + \alpha(0, t) + p(0, 0) \varepsilon(0, t) - c_2 \varepsilon_t(0, t), \\ \tilde{w}_x(1, t) = -c_1 \tilde{w}_t(1, t). \end{array} \right.$$

*Proof.* First we compute the second spatial derivative of the transformation (5.2):

$$(5.4) \quad \begin{aligned} \tilde{w}_{xx}(x, t) &= \widehat{w}_{xx}(x, t) - [2k_x(x, x) + k_y(x, x)] \widehat{w}(x, t) - k(x, x) \widehat{w}_x(x, t) \\ &\quad - \int_0^x k_{xx}(x, y) \widehat{w}(y, t) dy. \end{aligned}$$

The next step is to compute  $\tilde{w}_{tt}$ :

$$(5.5) \quad \begin{aligned} \delta \tilde{w}_{tt}(x, t) &= \delta \widehat{w}_{tt}(x, t) - \int_0^x k(x, y) \delta \widehat{w}_{tt}(y, t) dy \\ &= \widehat{w}_{xx}(x, t) + b^2 \widehat{w}(x, t) + b^3 \int_0^x \sinh(b(x-y)) \widehat{w}(y, t) dy \\ &\quad - b^2 \cosh(bx) w(0, t) - b \sinh(bx) \alpha(0, t) + p_y(x, 0) \varepsilon(0, t) \\ &\quad - c_2 p(x, 0) \varepsilon_t(0, t) - \int_0^x k(x, y) \widehat{w}_{yy}(y, t) dy - b^2 \int_0^x k(x, y) \widehat{w}(y, t) dy \\ &\quad - b^3 \int_0^x \int_y^x k(x, \xi) \sinh(b(\xi-y)) \widehat{w}(y, t) dy \\ &\quad + b^2 \int_0^x k(x, y) \cosh(by) dy w(0, t) + b \int_0^x k(x, y) \sinh(by) dy \alpha(0, t) \\ &\quad - \int_0^x k(x, y) p_y(y, 0) dy \varepsilon(0, t) + c_2 \int_0^x k(x, y) p(y, 0) dy \varepsilon_t(0, t). \end{aligned}$$

We notice that

$$(5.6) \quad \begin{aligned} \int_0^x k(x, y) \widehat{w}_{yy}(y, t) dy &= \int_0^x k_{yy}(x, y) \widehat{w}(y, t) dy \\ &\quad + k(x, x) \widehat{w}_x(x, t) - k(x, 0) \widehat{w}_x(0, t) \\ &\quad - k_y(x, x) \widehat{w}(x, t) + k_y(x, 0) \widehat{w}(0, t) \\ &= \int_0^x k_{yy}(x, y) \widehat{w}(y, t) dy + k(x, x) \widehat{w}_x(x, t) - k(x, 0) \alpha(0, t) \\ &\quad + qk(x, 0) w(0, t) + k(x, 0) p(0, 0) \varepsilon(0, t) + c_2 k(x, 0) \varepsilon_t(0, t) \\ &\quad - k_y(x, x) \widehat{w}(x, t) - k_y(x, 0) \varepsilon(0, t) + k_y(x, 0) w(0, t). \end{aligned}$$

Subtracting (5.4) from (5.5) and using (5.6), we get (5.3).

The boundary condition at  $x = 1$  is verified in the following way:

$$\begin{aligned} 0 &= \tilde{w}_x(1, t) + c_1 \tilde{w}_t(1, t) \\ &= \hat{w}_x(1, t) - k(1, 1) \hat{w}(1, t) - \int_0^1 k_x(1, y) \hat{w}(y, t) \\ &\quad + c_1 \hat{w}_t(1, t) - c_1 \int_0^1 k(1, y) \hat{w}_t(y, t) dy, \end{aligned}$$

which gives exactly the controller (5.1). Finally, for the boundary condition at  $x = 0$  we have

$$\begin{aligned} \tilde{w}_x(0, t) &= \hat{w}_x(0, t) - k(0, 0) \hat{w}(0, t) = \hat{w}_x(0, t) - k(0, 0) \tilde{w}(0, t) \\ &= \alpha(0, t) - qw(0, t) + p(0, 0) \varepsilon(0, t) - c_2 \varepsilon_t(0, t) - k(0, 0) \tilde{w}(0, t) \\ &= c_0 \tilde{w}(0, t) + \alpha(0, t) + p(0, 0) \varepsilon(0, t) - c_2 \varepsilon_t(0, t). \end{aligned}$$

The proof is complete.  $\square$

**6. Well-posedness and stability of the transformed system.** Lemmas 4.1 and 5.1 establish the following transformed system  $(\tilde{\varepsilon}, \tilde{w})$  which is a cascade of two wave equations (with additional integral terms):

$$(6.1) \quad \left\{ \begin{array}{l} \delta \tilde{\varepsilon}_{tt}(x, t) = \tilde{\varepsilon}_{xx}(x, t), \\ \tilde{\varepsilon}_x(0, t) = c_2 \tilde{\varepsilon}_t(0, t), \\ \tilde{\varepsilon}(1, t) = 0, \\ \delta \tilde{w}_{tt}(x, t) = \tilde{w}_{xx}(x, t) - b \sinh(bx) \alpha(0, t) + k(x, 0) \alpha(0, t) - c_2 k(x, 0) \tilde{\varepsilon}_t(0, t) \\ \quad + p_y(x, 0) \tilde{\varepsilon}(0, t) + k(x, 0) p(0, 0) \tilde{\varepsilon}(0, t) + k_y(x, 0) \tilde{\varepsilon}(0, t) \\ \quad + b \int_0^x k(x, y) \sinh(by) dy \alpha(0, t) - \int_0^x k(x, y) p_y(y, 0) dy \tilde{\varepsilon}(0, t) \\ \quad - c_2 p(x, 0) \tilde{\varepsilon}_t(0, t) + c_2 \int_0^x k(x, y) p(y, 0) dy \tilde{\varepsilon}_t(0, t), \\ \tilde{w}_x(0, t) = c_0 \tilde{w}(0, t) + \alpha(0, t) + p(0, 0) \tilde{\varepsilon}_t(0, t) - c_2 \tilde{\varepsilon}_t(0, t), \\ \tilde{w}_x(1, t) = -c_1 \tilde{w}_t(1, t). \end{array} \right.$$

Here  $\alpha(0, t)$  is expressed in terms of  $\tilde{\varepsilon}$  using (3.2) and (4.7):

$$(6.2) \quad \begin{aligned} \alpha(0, t) &= \frac{b}{\cosh(b)} \int_0^1 \sinh(b(1-x)) [\tilde{\varepsilon}_x(x, t) - p(x, x) \tilde{\varepsilon}(x, t)] dx \\ &\quad - \frac{b}{\cosh(b)} \int_0^1 \int_x^1 \sinh(b(1-y)) p_x(y, x) dy \tilde{\varepsilon}(x, t) dx. \end{aligned}$$

From (6.2) and the fact that  $\tilde{\varepsilon}(1, t) = 0$ , we know that there exists a constant  $C_1 > 0$  such that

$$(6.3) \quad |\alpha(0, t)|^2 \leq C_1 \int_0^1 \tilde{\varepsilon}_x^2(x, t) dx.$$

We consider the system (6.1) in the space  $H = H_R^1(0, 1) \times L^2(0, 1) \times H^1(0, 1) \times L^2(0, 1)$ ,



$H_R^1(0, 1) = \{f \mid f \in H^1(0, 1) \mid f(1) = 0\}$ , with the inner product

$$\begin{aligned} & \langle (f_1, g_1, \phi_1, \psi_1), (f_2, g_2, \phi_2, \psi_2) \rangle \\ &= K \int_0^1 \left[ f_1'(x) \overline{f_2'(x)} + \frac{1}{\delta} g_1(x) \overline{g_2(x)} + \delta_0(x-1)(f_1'(x) \overline{g_2(x)} + g_1(x) \overline{f_2'(x)}) \right] dx \\ &+ \int_0^1 \left[ \phi_1'(x) \overline{\phi_2'(x)} + \frac{1}{\delta} \psi_1(x) \overline{\psi_2(x)} + \delta_0(x+1)(\phi_1'(x) \overline{\psi_2(x)} + \psi_1(x) \overline{\phi_2'(x)}) \right] dx \\ &+ c_1 \phi_1(0) \overline{\phi_2(0)} \quad \forall (f, g, \phi, \psi) \in H, \end{aligned}$$

where  $\delta_0 > 0$  is sufficiently small so that above inner product is well-defined and  $K > 0$  is large enough so that  $A$  is dissipative in  $H$  as in the proof of Lemma 5.1 below. Define the system operator  $A : D(A) \subset H \rightarrow H$  as follows:

$$(6.4) \quad \left\{ \begin{aligned} D(A) &= \left\{ (f, g, \phi, \psi) \in (H^2(0, 1) \cap H_R^1(0, 1)) \times H_R^1(0, 1) \times H^2(0, 1) \right. \\ &\quad \times H^1(0, 1) \mid f'(0) = \frac{c_2}{\delta} g(0), \phi'(1) = -\frac{c_1}{\delta} \psi(1) \\ &\quad \left. \phi'(0) = c_0 \phi(0) + \frac{1}{\delta} [p(0, 0) - c_2] g(0) + \alpha(0) \right\}, \\ A(f, g, \phi, \psi) &= \left( \frac{g}{\delta}, f'', \frac{\psi}{\delta}, \phi'' - b \sinh(bx) \alpha(0) + [p_y(x, 0) + k_y(x, 0)] f(0) \right. \\ &\quad + k(x, 0) [\alpha(0) + p(0, 0) f(0)] - \int_0^x k(x, y) p_y(y, 0) dy f(0) \\ &\quad + b \int_0^x k(x, y) \sinh(by) dy \alpha(0) - \frac{c_2}{\delta} k(x, 0) g(0) \\ &\quad \left. + \frac{c_2}{\delta} \left[ -p(x, 0) + \int_0^x k(x, y) p(y, 0) dy \right] g(0) \right), \\ \alpha(0) &= \frac{b}{\cosh(b)} \int_0^1 \sinh(b(1-x)) [f'(x) - p(x, x) f(x)] dx \\ &\quad - \frac{b}{\cosh(b)} \int_0^1 \int_x^1 \sinh(b(1-y)) p_x(y, x) dy f(x) dx \\ &\quad \forall (f, g, \phi, \psi) \in D(A). \end{aligned} \right.$$

Then the system (6.1) can be written as

$$(6.5) \quad \frac{d}{dt} (\tilde{\varepsilon}(\cdot, t), \delta \tilde{\varepsilon}_t(\cdot, t), \tilde{w}(\cdot, t), \delta \tilde{w}_t(\cdot, t)) = A(\tilde{\varepsilon}(\cdot, t), \delta \tilde{\varepsilon}_t(\cdot, t), \tilde{w}(\cdot, t), \delta \tilde{w}_t(\cdot, t)).$$

**THEOREM 6.1.** *Let  $A$  be defined by (6.4). Then  $A$  generates an exponential stable  $C_0$ -semigroup on  $H$ . For any initial value  $(\tilde{\varepsilon}(\cdot, 0), \delta \tilde{\varepsilon}_t(\cdot, 0), \tilde{w}(\cdot, 0), \delta \tilde{w}_t(\cdot, 0)) \in H$ , there exists a unique (mild) solution to (6.1) such that  $(\tilde{\varepsilon}(\cdot, t), \delta \tilde{\varepsilon}_t(\cdot, t), \tilde{w}(\cdot, t), \delta \tilde{w}_t(\cdot, t)) \in C([0, \infty); H)$ , and there exists a positive constant  $\omega$  such that*

$$(6.6) \quad \begin{aligned} & \|(\tilde{\varepsilon}(\cdot, t), \delta \tilde{\varepsilon}_t(\cdot, t), \tilde{w}(\cdot, t), \delta \tilde{w}_t(\cdot, t))\|_H \\ & \leq e^{-\omega t} \|(\tilde{\varepsilon}(\cdot, 0), \delta \tilde{\varepsilon}_t(\cdot, 0), \tilde{w}(\cdot, 0), \delta \tilde{w}_t(\cdot, 0))\|_H. \end{aligned}$$

Moreover, if  $(\tilde{\varepsilon}(\cdot, 0), \delta \tilde{\varepsilon}_t(\cdot, 0), \tilde{w}(\cdot, 0), \delta \tilde{w}_t(\cdot, 0)) \in D(A)$ , then

$$(6.7) \quad (\tilde{\varepsilon}(\cdot, t), \delta \tilde{\varepsilon}_t(\cdot, t), \tilde{w}(\cdot, t), \delta \tilde{w}_t(\cdot, t)) \in C^1([0, \infty); H)$$

is the classical solution of (6.1).

*Proof.* Define the Lyapunov functions

$$(6.8) \quad E_{\tilde{\varepsilon}}(t) = \frac{1}{2} \int_0^1 [\tilde{\varepsilon}_x^2(x, t) + \delta \tilde{\varepsilon}_t^2(x, t)] dx + \delta_0 \int_0^1 (x-1) \tilde{\varepsilon}_x(x, t) \delta \tilde{\varepsilon}_t(x, t) dx$$

and

$$(6.9) \quad E_{\tilde{w}}(t) = \frac{1}{2} \int_0^1 [\tilde{w}_x^2(x, t) + \delta \tilde{w}_t^2(x, t)] dx + \frac{\tilde{\delta}_0}{2} \tilde{w}^2(0, t) + \delta_0 \int_0^1 (1+x) \tilde{w}_x(x, t) \delta \tilde{w}_t(x, t) dx.$$

Both of them are positive definite for small  $\delta_0, \tilde{\delta}_0 > 0$ . The time derivatives of  $E_{\tilde{\varepsilon}}$  and  $E_{\tilde{w}}$  along the trajectory of (6.1) are, respectively,

$$(6.10) \quad \dot{E}_{\tilde{\varepsilon}}(t) = - \left[ c_2 - \frac{\delta_0}{2} (1 + c_2^2) \right] \tilde{\varepsilon}_t^2(0, t) - \frac{\delta_0}{2} \int_0^1 [\tilde{\varepsilon}_x^2(x, t) + \delta \tilde{\varepsilon}_t^2(x, t)] dx,$$

$$(6.11) \quad \begin{aligned} \dot{E}_{\tilde{w}}(t) = & -c_1 \tilde{w}_t^2(1, t) - c_0 \tilde{w}^2(0, t) - \alpha(0, t) \tilde{w}_t(0, t) - p(0, 0) \tilde{w}_t(0, t) \tilde{\varepsilon}_t(0, t) \\ & + c_2 \tilde{w}_t(0, t) \tilde{\varepsilon}_t(0, t) + \int_0^1 [k(x, 0) - b \sinh(bx)] \tilde{w}_t(x, t) dx \alpha(0, t) \\ & + \int_0^1 \tilde{w}_t(x, t) p_y(x, 0) dx \tilde{\varepsilon}(0, t) + b \int_0^1 \tilde{w}_t(x, t) dx \int_0^x k(x, y) \sinh(by) dy \alpha(0, t) \\ & + \int_0^1 \tilde{w}_t(x, t) k(x, 0) dx p(0, 0) \tilde{\varepsilon}(0, t) + \int_0^1 \tilde{w}_t(x, t) k_y(x, 0) dx \tilde{\varepsilon}(0, t) \\ & - \int_0^1 \tilde{w}_t(x, t) dx \int_0^x k(x, y) p_y(y, 0) dy \tilde{\varepsilon}(0, t) - c_2 \int_0^1 \tilde{w}_t(x, t) k(x, 0) dx \tilde{\varepsilon}_t(0, t) \\ & - c_2 \int_0^1 \tilde{w}_t(x, t) p(x, 0) dx \tilde{\varepsilon}_t(0, t) + c_2 \int_0^1 \tilde{w}_t(x, t) dx \int_0^x k(x, y) p(y, 0) dy \tilde{\varepsilon}_t(0, t) \\ & - \frac{\delta_0}{2} \int_0^1 [\tilde{w}_x^2(x, t) + \delta \tilde{w}_t^2(x, t)] dx - \frac{\delta \delta_0}{2} \tilde{w}_t^2(0, t) + \delta_0 (\delta + c_1^2) \tilde{w}_t^2(1, t) \\ & - \frac{\delta_0}{2} [c_0 \tilde{w}(0, t) + \alpha(0, t) + p(0, 0) \tilde{\varepsilon}_t(0, t) - c_2 \tilde{\varepsilon}_t(0, t)]^2 \\ & + \delta_0 \int_0^1 (1+x) \tilde{w}_x(x, t) (k(x, 0) - b \sinh(bx)) dx \alpha(0, t) \\ & + \delta_0 \int_0^1 (1+x) \tilde{w}_x(x, t) (p_y(x, 0) + k_y(x, 0) + p(0, 0) k(x, 0)) dx \tilde{\varepsilon}(0, t) \\ & + b \delta_0 \int_0^1 (1+x) \tilde{w}_x(x, t) dx \int_0^x k(x, y) \sinh(by) dy \alpha(0, t) \\ & - \delta_0 \int_0^1 (1+x) \tilde{w}_x(x, t) dx \int_0^x k(x, y) p_y(y, 0) dy \tilde{\varepsilon}(0, t) \\ & - c_2 \delta_0 \int_0^1 (1+x) \tilde{w}_x(x, t) (p(x, 0) + k(x, 0)) dx \tilde{\varepsilon}_t(0, t) \\ & + c_2 \delta_0 \int_0^1 (1+x) \tilde{w}_x(x, t) dx \int_0^x k(x, y) p(y, 0) dy \tilde{\varepsilon}_t(0, t) + \tilde{\delta}_0 \tilde{w}(0, t) \tilde{w}_t(0, t). \end{aligned}$$

Using (6.3) and the fact that  $k, p \in C^2(\Omega)$ , we obtain

$$\begin{aligned} \dot{E}_{\tilde{w}}(t) \leq & \left[ \frac{\delta_0}{2} - \delta_1 \right] \int_0^1 [\tilde{w}_x^2(x, t) + \delta \tilde{w}_t^2(x, t)] dx - \left[ c_0 + \frac{\delta c_0^2}{2} - \delta_2 \right] \tilde{w}^2(0, t) \\ & - [c_1 - \delta_0(\delta + c_1^2)] \tilde{w}_t^2(1, t) - \left[ \frac{\delta \delta_0}{2} - \delta_3 \right] \tilde{w}_t^2(0, t) \\ & + C_2 \alpha^2(0, t) + C_2 \tilde{\varepsilon}^2(0, t) + C_2 \tilde{\varepsilon}_t^2(0, t), \end{aligned}$$

where  $C_2$  and  $\delta_i, i = 1, 2, 3$ , are some positive constants satisfying

$$(6.12) \quad \delta_1 < \frac{\delta_0}{2}, \quad \delta_2 < c_0 + \frac{\delta c_0^2}{2}, \quad \delta_3 < \frac{\delta \delta_0}{2}.$$

Now for large  $K > 0$ , we take the overall Laypunov function as

$$(6.13) \quad E(t) = E_{\tilde{w}}(t) + K E_{\tilde{\varepsilon}}(t).$$

Since from (6.3),  $\alpha^2(0, t) + \tilde{\varepsilon}^2(0, t) \leq (1 + C_1) \|\tilde{\varepsilon}_x(\cdot, t)\|_{L^2(0,1)}^2$ , we obtain its derivative along the solution of (6.1) that

$$\begin{aligned} \dot{E}(t) \leq & -K \left[ c_2 - \frac{\delta_0}{2} (1 + c_2^2) \right] \tilde{\varepsilon}_t^2(0, t) - K \frac{\delta_0}{2} \int_0^1 [\tilde{\varepsilon}_x^2(x, t) + \delta \tilde{\varepsilon}_t^2(x, t)] dx \\ (6.14) \quad & - \left[ \frac{\delta_0}{2} - \delta_1 \right] \int_0^1 [\tilde{w}_x^2(x, t) + \delta \tilde{w}_t^2(x, t)] dx - \left[ c_0 + \frac{\delta c_0^2}{2} - \delta_2 \right] \tilde{w}^2(0, t) \\ & + C_2 \alpha^2(0, t) + C_2 \tilde{\varepsilon}^2(0, t) + C_2 \tilde{\varepsilon}_t^2(0, t), \end{aligned}$$

where we assumed that

$$(6.15) \quad \delta_0(\delta + c_1^2) < c_1.$$

Hence

$$\begin{aligned} \dot{E}(t) \leq & - \left[ K \left( c_2 - \frac{\delta_0}{2} (1 + c_2^2) \right) - C_2 \right] \tilde{\varepsilon}_t^2(0, t) \\ (6.16) \quad & - \left[ K \frac{\delta_0}{2} - C_2(1 + C_1) \right] \int_0^1 [\tilde{\varepsilon}_x^2(x, t) + \delta \tilde{\varepsilon}_t^2(x, t)] dx \\ & - \left[ \frac{\delta_0}{2} - \delta_1 \right] \int_0^1 [\tilde{w}_x^2(x, t) + \delta \tilde{w}_t^2(x, t)] dx - \left[ c_0 + \frac{\delta c_0^2}{2} - \delta_2 \right] \tilde{w}^2(0, t). \end{aligned}$$

Choosing  $K > 0$  sufficiently large, it follows from (6.16) that there exists an  $\omega > 0$  such that

$$(6.17) \quad \dot{E}(t) \leq -\omega E(t).$$

The above procedure also gives the following estimate:

$$\operatorname{Re} \langle A(f, g, \phi, \psi), (f, g, \phi, \psi) \rangle_H \leq -\omega \|(f, g, \phi, \psi)\|_H^2 \quad \forall (f, g, \phi, \psi) \in D(A).$$

So  $A$  is dissipative in  $H$  ([10]), and if  $A$  generates a  $C_0$ -semigroup, this semigroup must be exponentially stable. By the Lumer–Phillips theorem (Theorem 4.3, p. 14 in [10]), the proof will be accomplished if we can show that  $A^{-1}$  exists and is bounded on  $H$ . Actually, a simple computation shows that

$$A^{-1}(f, g, \phi, \psi) = (f^*, g^*, \phi^*, \psi^*) \quad \forall (f, g, \phi, \psi) \in H,$$

where  $g^* = \delta f$ ,  $\psi^* = \delta\phi$  and

$$\begin{aligned}
 f^*(x) &= c_2 f(0)(x-1) + \int_0^x (x-\tau)g(\tau)d\tau - \int_0^1 (1-\tau)g(\tau)d\tau, \\
 \phi^*(x) &= \int_1^x (x-\tau)\psi(\tau)d\tau + \int_1^x (x-\tau)F(\tau)d\tau - c_1\phi(1)x \\
 &\quad - \int_0^1 \tau\psi(\tau)d\tau - \int_0^1 \tau F(\tau)d\tau + \phi^*(0), \\
 \phi^*(0) &= -\frac{1}{c_0} \int_0^1 \psi(\tau)d\tau - \frac{1}{c_0} \int_0^1 F(\tau)d\tau - \frac{c_1}{c_0}\phi(1) - \frac{\alpha^*(0)}{c_0} - \frac{p(0,0) - c_2}{c_0}f(0), \\
 F(x) &= b \sinh(bx)\alpha^*(0) - k(x,0)\alpha^*(0) - [p_y(x,0) + k(x,0)p(0,0) + k_y(x,0)]f^*(0) \\
 &\quad - b \int_0^x k(x,y) \sinh(by)dy\alpha^*(0) + \int_0^x k(x,y)p_y(y,0)dyf^*(0) \\
 &\quad + \left[ c_2k(x,0) + c_2p(x,0) - c_2 \int_0^x k(x,y)p(y,0)dy \right] f(0), \\
 \alpha^*(0) &= \frac{b}{\cosh(b)} \int_0^1 \sinh(b(1-x))f^{*'}(x)dx \\
 &\quad - \frac{b}{\cosh(b)} \int_0^1 \left[ \sinh(b(1-x))p(x,x) + \int_x^1 \sinh(b(1-y))p_x(y,x)dy \right] f^*(x)dx.
 \end{aligned}$$

The proof is complete.  $\square$

**7. Well-posedness and stability of the closed-loop system.** The closed-loop system consists of the plant (3.3), the observer (4.2), and the feedback controller (5.1):

$$(7.1) \quad \left\{ \begin{aligned}
 \delta w_{tt}(x,t) &= w_{xx}(x,t) + b^2 w(x,t) + b^3 \int_0^x \sinh(b(x-y))w(y,t)dy \\
 &\quad - b^2 \cosh(bx)w(0,t) - b \sinh(bx)\alpha(0,t), \\
 w_x(0,t) &= \alpha(0,t) - qw(0,t), \\
 w(1,t) &= \hat{w}(1,t), \\
 \delta \hat{w}_{tt}(x,t) &= \hat{w}_{xx}(x,t) + b^2 \hat{w}(x,t) + b^3 \int_0^x \sinh(b(x-y))\hat{w}(y,t)dy \\
 &\quad - b^2 \cosh(bx)w(0,t) - b \sinh(bx)\alpha(0,t) \\
 &\quad + p_y(x,0)[w(0,t) - \hat{w}(0,t)] - c_2 p(x,0)[w_t(0,t) - \hat{w}_t(0,t)], \\
 \hat{w}_x(0,t) &= \alpha(0,t) - qw(0,t) + p(0,0)[w(0,t) - \hat{w}(0,t)] \\
 &\quad - c_2[w_t(0,t) - \hat{w}_t(0,t)], \\
 \hat{w}_x(1,t) &= -c_1 \hat{w}_t(1,t) + k(1,1)\hat{w}(1,t) \\
 &\quad + c_1 \int_0^1 k(1,y)\hat{w}_t(y,t)dy + \int_0^1 k_x(1,y)\hat{w}(y,t)dy, \\
 \alpha(0,t) &= \frac{b}{\cosh(b)} \int_0^1 \sinh(b(1-s))[w_x(x,t) - \hat{w}_x(x,t)]dx,
 \end{aligned} \right.$$

and

$$(7.2) \quad \alpha(x, t) = \cosh(bx)\alpha(0, t) - b \int_0^x \sinh(b(x-s))w_x(s, t)ds.$$

We consider the system (7.1) in the state space  $\mathcal{H} = \{(f, g, \phi, \psi) \in (H^1(0, 1) \times L^2(0, 1))^2 \mid f(1) = \phi(1)\}$ . Define the system operator

$$(7.3) \quad \left\{ \begin{array}{l} D(\mathcal{A}) = \left\{ (f, g, \phi, \psi) \in \mathcal{H} \mid \mathcal{A}(f, g, \phi, \psi) \in \mathcal{H}, f'(0) = \alpha(0) - qf(0), \right. \\ \quad \phi'(0) = \alpha(0) - qf(0) + p(0, 0)[f(0) - \phi(0)] - \frac{c_2}{\delta}[g(0) - \psi(0)], \\ \quad \phi'(1) = k(1, 1)\phi(1) - \frac{c_1}{\delta}\psi(1) + \frac{c_1}{\delta} \int_0^1 k(1, x)\psi(x)dx + \int_0^1 k_x(1, x)\phi(x)dx \Big\}, \\ \quad [\mathcal{A}(f, g, \phi, \psi)](x) = \left( \frac{g(x)}{\delta}, f''(x) + b^2f + b^3 \int_0^x \sinh(b(x-y))f(y)dy \right. \\ \quad \left. - b^2 \cosh(bx)f(0) - b \sinh(bx)\alpha(0), \frac{\psi(x)}{\delta}, \phi''(x) + b^2\phi(x) \right. \\ \quad \left. + b^3 \int_0^x \sinh(b(x-y))\phi(y)dy - b^2 \cosh(bx)f(0) - b \sinh(bx)\alpha(0) \right. \\ \quad \left. + p_y(x, 0)[f(0) - \phi(0)] - \frac{c_2}{\delta}p(x, 0)[g(0) - \psi(0)] \right), \\ \quad \alpha(0) = \frac{b}{\cosh(b)} \int_0^1 \sinh(b(1-s))[f'(x) - \phi'(x)]dx \quad \forall (f, g, \phi, \psi) \in D(\mathcal{A}). \end{array} \right.$$

Then the system (7.3) can be written as an evolution equation in  $\mathcal{H}$ :

$$(7.4) \quad \frac{d}{dt}(w(\cdot, t), \delta w_t(\cdot, t), \widehat{w}(\cdot, t), \delta \widehat{w}_t(\cdot, t)) = \mathcal{A}(w(\cdot, t), \delta w_t(\cdot, t), \delta \widehat{w}_t(\cdot, t), \widehat{w}_t(\cdot, t)).$$

**THEOREM 7.1.** *Let  $\mathcal{A}$  be defined by (7.3). Then  $\mathcal{A}$  generates a  $C_0$ -semigroup  $e^{\mathcal{A}t}$  on  $\mathcal{H}$ , which is exponentially stable:*

$$\|e^{\mathcal{A}t}\|_{\mathcal{H}} \leq Me^{-\omega t} \quad \forall t \geq 0$$

for some positive constants  $M$  and  $\omega$  independent of  $t$ . In particular,

$$(7.5) \quad E_o(t) \leq Ce^{-\omega t}E_o(0)$$

for some  $C > 0$ , where

$$(7.6) \quad E_o(t) = \int_0^1 [w_x^2(x, t) + \delta w_t^2(x, t) + \widehat{w}_x^2(x, t) + \delta \widehat{w}_t^2(x, t) + \alpha^2(x, t)] dx.$$

*Proof.* For any initial value  $(w(\cdot, 0), \delta w_t(\cdot, 0), \widehat{w}(\cdot, 0), \delta \widehat{w}_t(\cdot, 0)) \in D(\mathcal{A})$ , let

$$(7.7) \quad \left\{ \begin{array}{l} \widetilde{\varepsilon}(x, 0) = [(I - \mathbb{P}_1)^{-1}(w(\cdot, 0) - \widehat{w}(\cdot, 0))](x, 0), \\ \delta \widetilde{\varepsilon}_t(x, 0) = [(I - \mathbb{P}_1)^{-1}(\delta w_t(\cdot, 0) - \delta \widehat{w}_t(\cdot, 0))](x, 0), \\ \widetilde{w}(x, 0) = \widehat{w}(x, 0) - \int_0^x k(x, y)\widehat{w}(y, 0)dy, \\ \delta \widetilde{w}_t(x, 0) = \delta \widehat{w}_t(x, 0) - \int_0^x k(x, y)\delta \widehat{w}_t(y, 0)dy. \end{array} \right.$$

A direct computation shows that  $(\tilde{\varepsilon}(\cdot, 0), \delta\tilde{\varepsilon}_t(\cdot, 0), \tilde{w}(\cdot, 0), \delta\tilde{w}_t(\cdot, 0)) \in D(A)$ . So there exists a unique classical solution to (6.1) with this initial value. Let

$$(7.8) \quad \begin{cases} w(x, t) = \hat{w}(x, t) + \tilde{\varepsilon}(x, t) - \int_0^x p(x, y)\tilde{\varepsilon}(y, t)dy, \\ \hat{w}(x, t) = [(I - \mathbb{P}_2)^{-1}\tilde{w}](x, t). \end{cases}$$

Similarly to (5.5), one can show that  $(w, \hat{w})$  defined in this way satisfies (7.1) with initial value  $(w(\cdot, 0), \delta w_t(\cdot, 0), \hat{w}(\cdot, 0), \delta\hat{w}_t(\cdot, 0))$ . This solution is unique by the invertible transformation and the uniqueness of the classical solution to (6.1), where  $\mathbb{T}$  is a one to one

$$(7.9) \quad \begin{pmatrix} \tilde{\varepsilon} \\ \delta\tilde{\varepsilon}_t \\ \tilde{w} \\ \delta\tilde{w}_t \end{pmatrix} = \begin{pmatrix} I - \mathbb{P}_1 & 0 & -I + \mathbb{P}_1 & 0 \\ 0 & I - \mathbb{P}_1 & 0 & -I + \mathbb{P}_1 \\ 0 & 0 & I - \mathbb{P}_2 & 0 \\ 0 & 0 & 0 & I - \mathbb{P}_2 \end{pmatrix} \begin{pmatrix} w \\ \delta w_t \\ \hat{w} \\ \delta\hat{w}_t \end{pmatrix},$$

$$\begin{pmatrix} w \\ \delta w_t \\ \hat{w} \\ \delta\hat{w}_t \end{pmatrix} = \begin{pmatrix} I - \mathbb{P}_1 & 0 & (I - \mathbb{P}_2)^{-1} & 0 \\ 0 & I - \mathbb{P}_1 & 0 & (I - \mathbb{P}_1)^{-1} \\ 0 & 0 & (I - \mathbb{P}_2)^{-1} & 0 \\ 0 & 0 & 0 & (I - \mathbb{P}_2)^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\varepsilon} \\ \delta\tilde{\varepsilon}_t \\ \tilde{w} \\ \delta\tilde{w}_t \end{pmatrix}$$

and onto operator from  $\mathcal{H}$  to  $H$ . Moreover, this solution is exponentially stable by (6.6) and (7.9):

$$(7.10) \quad \begin{aligned} & \| (w(\cdot, t), \delta w_t(\cdot, t), \hat{w}(\cdot, t), \delta\hat{w}_t(\cdot, t)) \|_{\mathcal{H}} \\ & \leq M e^{-\omega t} \| (w(\cdot, 0), \delta w_t(\cdot, 0), \hat{w}(\cdot, 0), \delta\hat{w}_t(\cdot, 0)) \|_{\mathcal{H}} \end{aligned}$$

for some positive constant  $M$  independent of  $t$ . From transformation (7.9), we know that  $\mathcal{A} = \mathbb{T}^{-1}A\mathbb{T}$ , and hence  $\mathcal{A} = \mathbb{T}A\mathbb{T}^{-1}$ , where  $A$  is the operator defined by (6.4). Hence  $\mathcal{A}^{-1}$  exists and is bounded on  $\mathcal{H}$ , which implies that  $\rho(\mathcal{A})$ , the resolvent set of  $\mathcal{A}$ , is not empty. Since obviously  $D(\mathcal{A})$  is dense in  $\mathcal{H}$ , it follows from Theorem 1.3 on page 102 of [10] that  $\mathcal{A}$  generates a  $C_0$ -semigroup  $e^{\mathcal{A}t}$  on  $\mathcal{H}$ . (7.10) shows that  $e^{\mathcal{A}t}$  is exponentially stable, with  $\|e^{\mathcal{A}t}\| \leq M e^{-\omega t}$  for all  $t \geq 0$ . Finally, it follows from (7.2) that  $\|\alpha(\cdot, t)\|_{L^2(0,1)} \leq \tilde{C}[\|w_x(\cdot, t)\|_{L^2(0,1)} + \|\hat{w}_x(\cdot, t)\|_{L^2(0,1)}]$  for some constant  $\tilde{C} > 0$  independent of  $t$ . This together with (7.10) gives (7.5). The proof is complete.  $\square$

**8. Simulation results.** In this section we demonstrate through numerical simulations the effectiveness of the control and the observer.

We use the backward Euler method in the time domain and the Chebyshev spectral method in space. For this purpose the second-order-in-time equations are first converted into first-order-in-time (evolution-type) systems of equations. The control kernel  $k$  is first approximated on a uniform grid using the iterative scheme (3.10), and then linear interpolation was used to obtain values on the nonuniform Chebyshev grid. The boundary conditions were implemented using second-order explicit discretization. In the numerical simulations we used grid size  $N = 40$  in space and time step  $dt = 10^{-4}$ . The convergence of the numerical method was checked by varying  $N$  between  $N = 30$  and  $N = 70$  and varying  $dt$  between  $dt = 10^{-2}$  and  $dt = 10^{-5}$ . The maximum variation of the solution did not exceed  $10^{-3}$  over the whole time and space domain. The numerical code was programed in MATLAB (see, e.g., [18]).

The main system parameters are set to  $\delta = 1$ ,  $b = 0.6$ , and  $q = 0.9$ . The design parameters are set to  $c_0 = 10$ ,  $c_1 = 1$ , and  $c_2 = 1$ . The initial conditions are

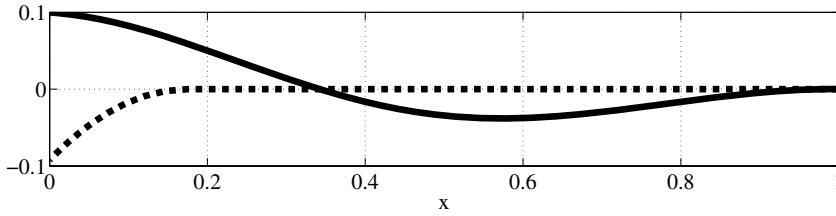


FIG. 8.1. Initial conditions of the beam. Solid line,  $w(0, x)$ ; dashed line,  $w_t(0, x)$ .

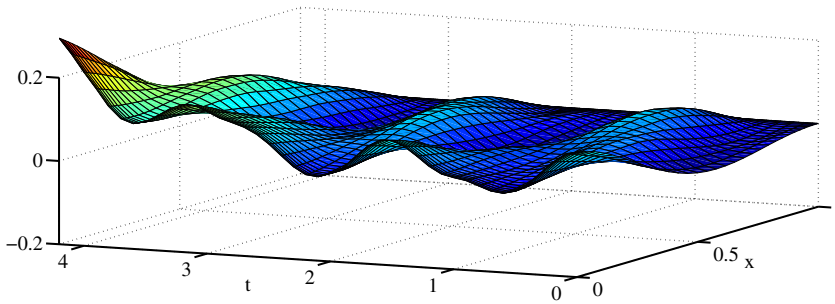


FIG. 8.2. Beam response  $w(x, t)$ . Uncontrolled case, clamped at  $x = 1$ . Note the instability that results from  $q = 0.9$ .

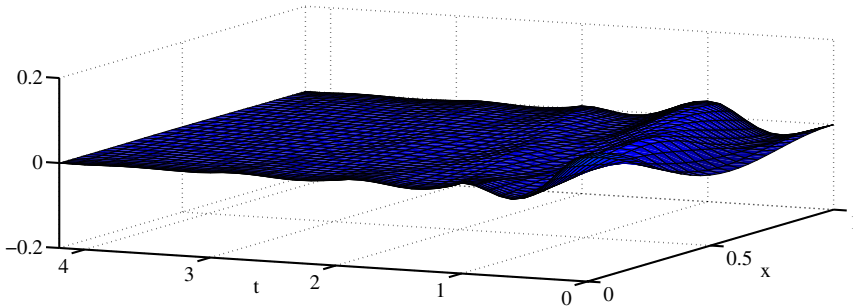


FIG. 8.3. Beam response  $w(x, t)$  with control  $w_x(1, t)$  applied at  $x = 1$ . Note that the instability at  $x = 0$  is stabilized.

$w(x, 0) = -0.1(1 - x) \sin(1.52\pi(1 - x))$  for  $x \in [0, 1]$  and

$$w_t(x, 0) = \begin{cases} -3(x - x_0)^2 & \text{if } x \in [0, x_0], \\ 0 & \text{if } x \in [x_0, 1], \end{cases}$$

with  $x_0 = 0.1753$ . These initial conditions correspond to hitting the tip part of an already bent beam (Figure 8.1).

**8.1. No observer, full state feedback.** First we consider the full state feedback case. This is equivalent to assuming that the observer starts from the same initial conditions as the plant itself, and hence it is identical with it for all time. Figures 8.2 and 8.3 show the results of our simulation for the shear beam with a zero

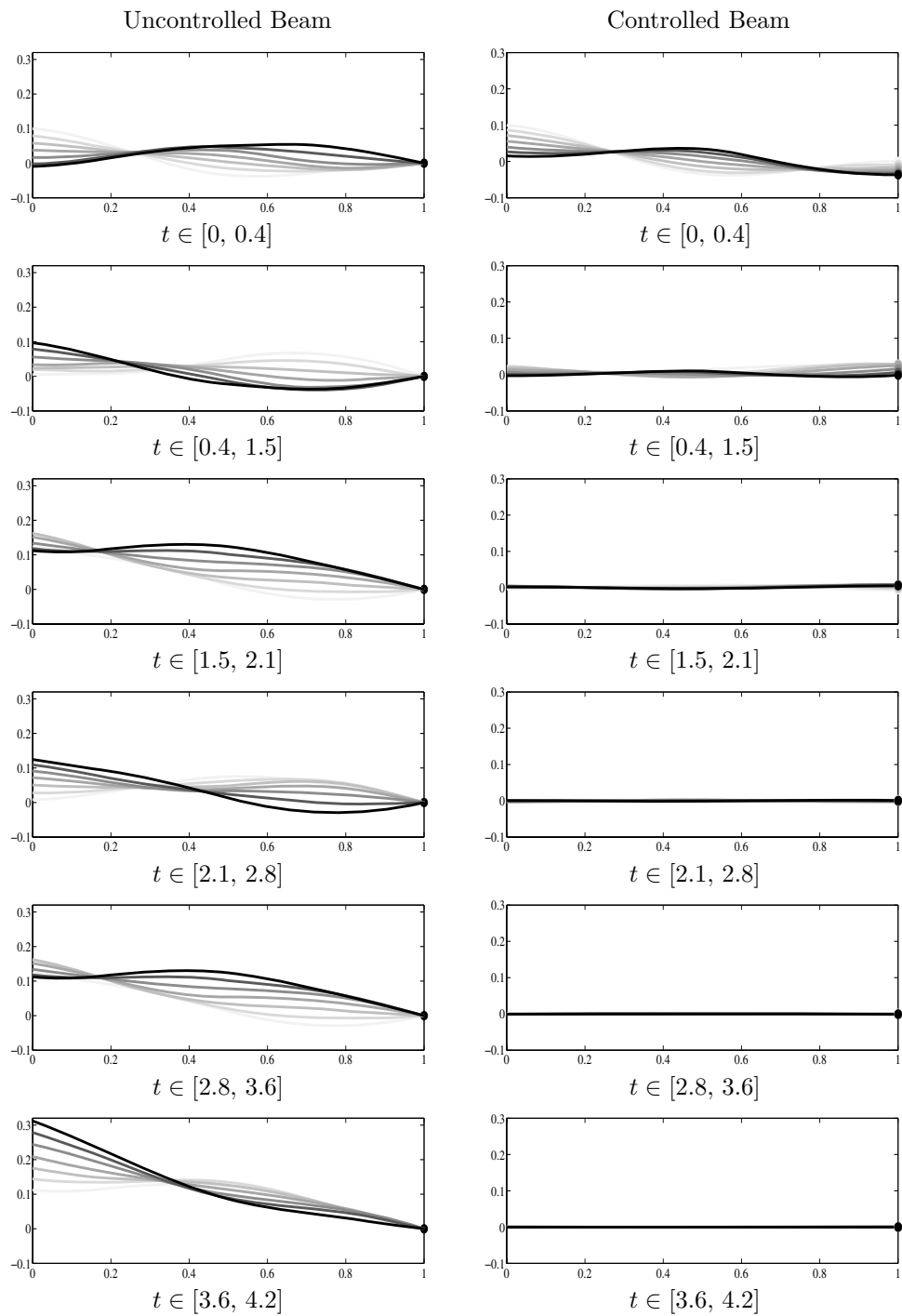


FIG. 8.4. Snapshots of the beam movements with increasing darkness denoting increasing time in the sequences.



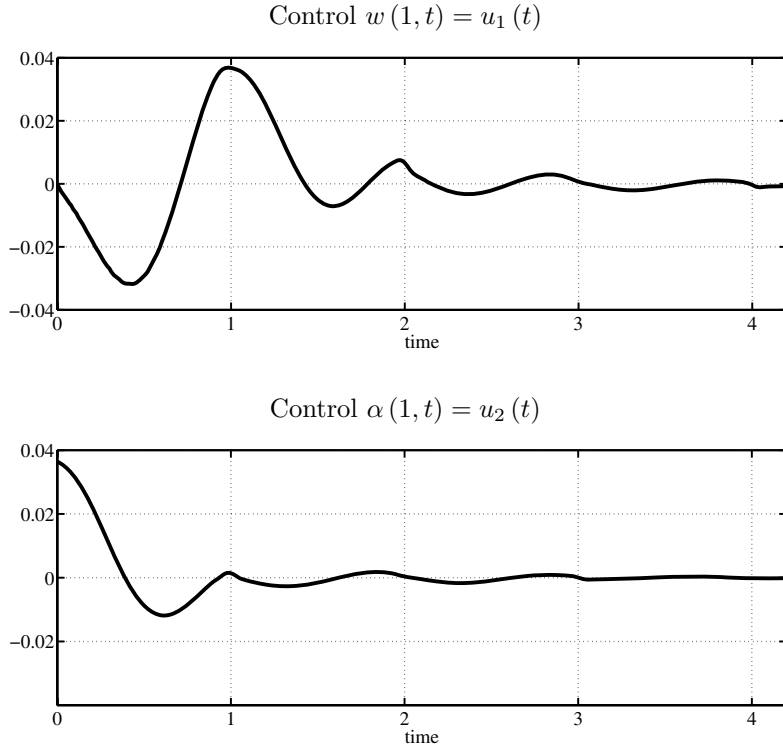


FIG. 8.5. Time trace of controls  $w(1, t) = u_1(t)$  and  $\alpha(1, t) = u_2(t)$ .

Dirichlet boundary condition at  $x = 1$  in the uncontrolled case and with control (5.1) in the controlled case. The uncontrolled case shows that the relatively large value  $q = 0.9$  destabilizes the trivial zero solution. The controlled case shows asymptotic stability with small control effort. Snapshots of beam movements (uncontrolled and controlled) are depicted in Figure 8.4, where vibrations are shown in sequences of time intervals. In each sequence the time evolution is represented by increasing darkness. The control effort is shown explicitly as function of time in Figure 8.5. Notice that the maximum control effort is about one magnitude smaller than the maximum of the uncontrolled solution. Gain kernels  $k(1, y)$  and  $k_x(1, y)$  of the first control law (4.3) are shown in Figure 8.6 for  $y \in [0, 1]$ . These kernel functions have small spatial variations and can be easily approximated by low order polynomials.

**8.2. Observer design.** We now introduce the observer (4.2) in the simulations. We assume no knowledge of the initial state of the beam, which means that the observer is started with zero initial conditions  $\hat{w}(x, 0) = \hat{w}_t(x, 0)$  for  $x \in [0, 1]$ . Figure 8.7 shows that in the uncontrolled case, although the observer starts far from the state, the observer error quickly converges to zero over a time period of  $t \in [0, 4]$ . As expected in the controlled case (see Figure 8.8) the convergence takes place over a longer time period  $t \in [0, 6]$ . The reason for this short delay in the convergence is that the observer has to compensate for the additional error introduced by the control feedback of the observer into the plant. Nevertheless, the closed-loop state quickly converges to zero (Figure 8.9). Finally, the observer gains  $p(x, 0)$  and  $p_y(x, 0)$  can be seen in Figure 8.10.

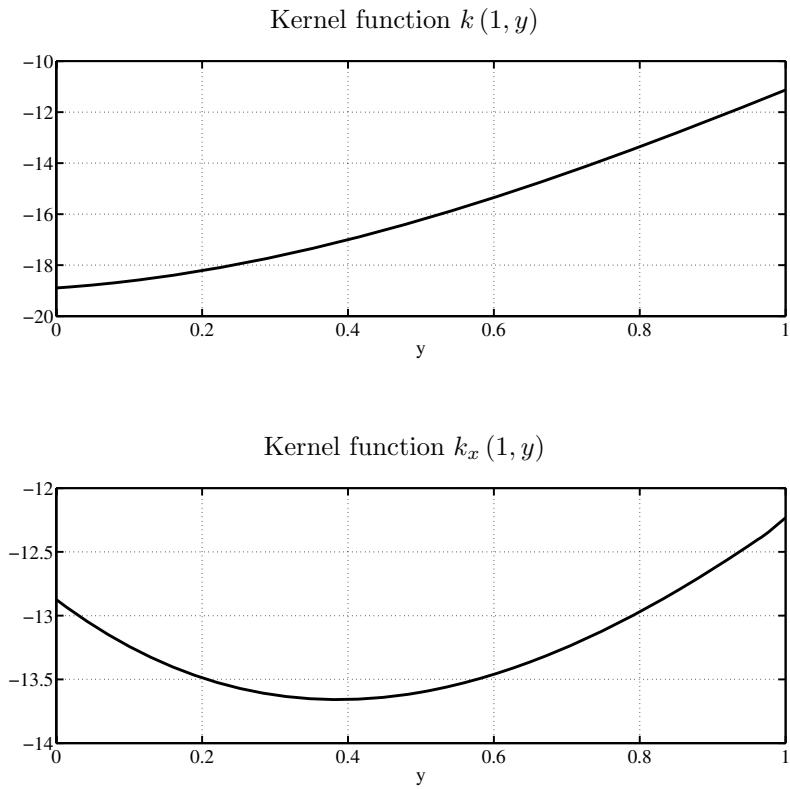


FIG. 8.6. Kernel functions  $k(1, y)$  and  $k_x(1, y)$  of control law (5.1) for  $y \in [0, 1]$ .

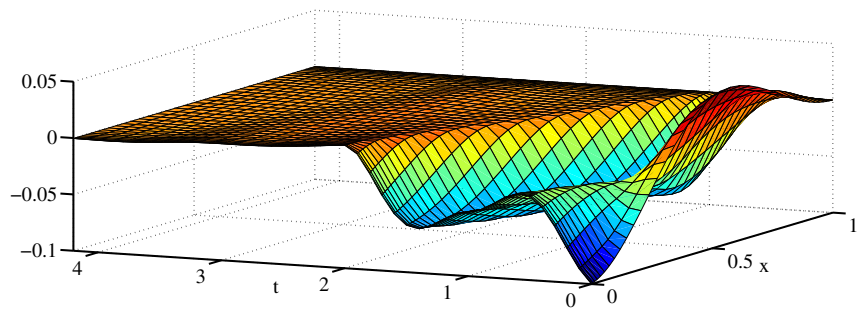


FIG. 8.7. Observer error  $\hat{w} - w$  in the uncontrolled case.

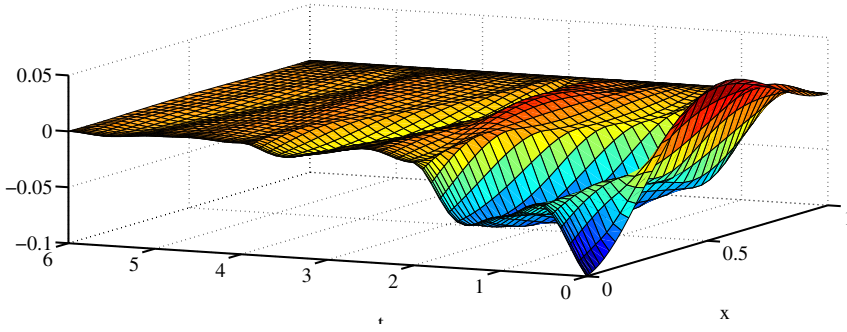
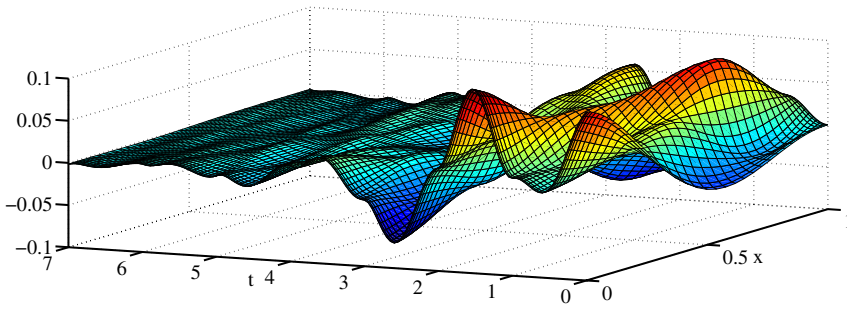
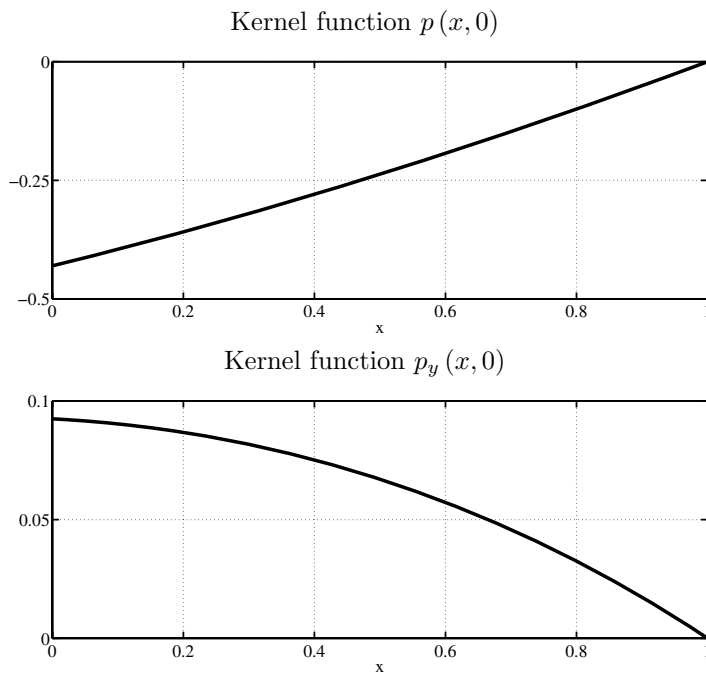

 FIG. 8.8. Observer error  $\hat{w} - w$  in the controlled case.


FIG. 8.9. Plant controlled using the observer.


 FIG. 8.10. Kernel functions  $p(x, 0)$  and  $p_y(x, 0)$  of the observer (4.2) for  $x \in [0, 1]$ .

**9. Conclusion.** In this paper we presented the output-feedback controller for an undamped shear beam. Future efforts will be concentrated on developing the controllers for higher-dimensional flexible structures such as plates and shells. Another interesting avenue of research is the control of beams (plates, shells) in the presence of parametric uncertainties, such as unknown structural damage. Successful backstepping boundary adaptive controllers for parabolic PDEs were recently developed in [5, 15, 16], and similar ideas could be applied to the hyperbolic equations.

## REFERENCES

- [1] M. S. DE QUEIROZ, D. M. DAWSON, M. AGARWAL, AND F. ZHANG, *Adaptive nonlinear boundary control of a flexible link robot arm*, IEEE Trans. Robotics and Automation, 15 (1999), pp. 779–787.
- [2] M. S. DE QUEIROZ, D. M. DAWSON, S. P. NAGARKATTI, AND F. ZHANG, *Lyapunov-based Control of Mechanical Systems*, Birkhauser, Boston, 2000.
- [3] S. M. HAN, H. BENAROYA, AND T. WEI, *Dynamics of transversely vibrating beams using four engineering theories*, J. Sound Vibration, 225 (1999), pp. 935–988.
- [4] J. U. KIM AND Y. RENARDY, *Boundary control of the Timoshenko beam*, SIAM J. Control Optim., 25 (1987), pp. 1417–1429.
- [5] M. KRSTIC AND A. SMYSHLYAEV, *Adaptive boundary control for unstable parabolic PDEs—Part I: Lyapunov design*, IEEE Trans. Automat. Control, to appear.
- [6] M. KRSTIC, A. SMYSHLYAEV, AND A. SIRANOSIAN, *Backstepping boundary controllers and observers for the slender Timoshenko beam: Part I—Design*, in Proceedings of the American Control Conference, IEEE, Piscataway, NJ, 2006.
- [7] A. MACCHELLI AND C. MELCHIORRI, *Modeling and control of the Timoshenko beam. The distributed port Hamiltonian approach*, SIAM J. Control Optim., 43 (2004), pp. 743–767.
- [8] Z. H. LUO, B. Z. GUO, AND O. MORGUL, *Stability and Stabilization of Infinite Dimensional Systems with Applications*, Springer-Verlag, Berlin, 1999.
- [9] O. MORGUL, *Dynamic boundary control of the Timoshenko beam*, Automatica, 28 (1992), pp. 1255–1260.
- [10] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, 1983.
- [11] D. H. SHI, S. H. HOU, AND D.-X. FENG, *Feedback stabilization of a Timoshenko beam with an end mass*, Internat. J. Control, 69 (1998), pp. 285–300.
- [12] A. SMYSHLYAEV AND M. KRSTIC, *Closed form boundary state feedbacks for a class of 1-D partial integro-differential equations*, IEEE Trans. Automat. Control, 49 (2004), pp. 2185–2202.
- [13] A. SMYSHLYAEV AND M. KRSTIC, *Backstepping observers for a class of parabolic PDEs*, Systems Control Lett., 54 (2005), pp. 613–625.
- [14] A. SMYSHLYAEV AND M. KRSTIC, *On control design for PDEs with space-dependent diffusivity or time-dependent reactivity*, Automatica, 41 (2005), pp. 1601–1608.
- [15] A. SMYSHLYAEV AND M. KRSTIC, *Adaptive boundary control for unstable parabolic PDEs—Part II: Estimation-based designs*, Automatica, 43 (2007), pp. 1543–1556.
- [16] A. SMYSHLYAEV AND M. KRSTIC, *Adaptive boundary control for unstable parabolic PDEs—Part III: Output feedback examples with swapping identifiers*, Automatica, 43 (2007), pp. 1557–1564.
- [17] S. W. TAYLOR AND S. C. B. YAU, *Boundary control of a rotating Timoshenko beam*, ANZIAM J., 44 (2003), pp. E143–E184.
- [18] L. N. TREFETHEN, *Spectral Methods in MATLAB*, SIAM, Philadelphia, 2000.
- [19] R. VAZQUEZ, F. J. RUBIO-SIERRA, AND R. W. STARK, *Multimodal analysis of force spectroscopy based on a transfer function study of micro-cantilevers*, Nanotechnology, 18 (2007), 185504.
- [20] F. ZHANG, D. M. DAWSON, M. S. DE QUEIROZ, AND P. VEDAGARBHA, *Boundary control of the Timoshenko beam with free-end mass/inertia*, in Proceedings of the IEEE Conference on Decision and Control, 1997.
- [21] H. L. ZHAO, K. S. LIU, AND C. G. ZHANG, *Stability for the Timoshenko beam system with local Kelvin-Voigt damping*, Acta Math. Sin. (Engl. Ser.), 21 (2005), pp. 655–666.

## REACHING A CONSENSUS IN A DYNAMICALLY CHANGING ENVIRONMENT: A GRAPHICAL APPROACH\*

MING CAO<sup>†</sup>, A. STEPHEN MORSE<sup>‡</sup>, AND BRIAN D. O. ANDERSON<sup>‡</sup>

**Abstract.** This paper presents new graph-theoretic results appropriate for the analysis of a variety of consensus problems cast in dynamically changing environments. The concepts of rooted, strongly rooted, and neighbor-shared are defined, and conditions are derived for compositions of sequences of directed graphs to be of these types. The graph of a stochastic matrix is defined, and it is shown that under certain conditions the graph of a Sarymsakov matrix and a rooted graph are one and the same. As an illustration of the use of the concepts developed in this paper, graph-theoretic conditions are obtained which address the convergence question for the leaderless version of the widely studied Vicsek consensus problem.

**Key words.** cooperative control, graph theory, switched systems, multiagent systems

**AMS subject classifications.** 93C05, 05C50, 05C75, 15A51, 40A20, 68W15

**DOI.** 10.1137/060657005

**1. Introduction.** Current interest in cooperative control of groups of mobile autonomous agents has led to the rapid increase in the application of graph-theoretic ideas to problems of analyzing and synthesizing a variety of desired group behaviors such as maintaining a formation, swarming, rendezvousing, or reaching a consensus. While this in-depth assault on group coordination using a combination of graph theory and system theory is in its early stages, it is likely to significantly expand in the years to come. One line of research which illustrates the combined use of these concepts is the recent theoretical work by a number of individuals [17, 19, 22, 1, 3, 26] which successfully explains the heading synchronization phenomenon observed in simulation by Vicsek et al. [29], Reynolds [23], and others more than a decade ago. Vicsek and coauthors consider a simple discrete-time model consisting of  $n$  autonomous agents or particles all moving in the plane with the same speed but with different headings. Each agent's heading is updated using a local rule based on the average of its own heading plus the current headings of its "neighbors." Agent  $i$ 's *neighbors* at time  $t$  are those agents which are either in or on a circle of prespecified radius centered at agent  $i$ 's current position. In their paper, Vicsek et al. provide a variety of interesting simulation results which demonstrate that the nearest neighbor rule they are studying can cause all agents to eventually move in the same direction despite the absence of centralized coordination and despite the fact that each agent's set of nearest neighbors can change with time. A theoretical explanation for this observed behavior

---

\*Received by the editors April 11, 2006; accepted for publication (in revised form) August 16, 2007; published electronically February 6, 2008. A preliminary version of this work can be found in A. S. Morse, *Logically switched dynamical systems*, in Nonlinear and Optimal Control Theory, Springer-Verlag, Berlin, 2008, pp. 1–84.

<http://www.siam.org/journals/sicon/47-2/65700.html>

<sup>†</sup>Electrical Engineering, Yale University, P.O. Box 208267, New Haven, CT 06520 (m.cao@yale.edu, morse@sysc.eng.yale.edu). The research of these authors was supported by the U.S. Army Research Office, the U.S. National Science Foundation, and a gift from the Xerox Corporation.

<sup>‡</sup>Australian National University and National ICT Australia Ltd., Locked bag 8001, Canberra ACT 2601, Australia (brian.anderson@nicta.com.au). The research of this author was supported by National ICT Australia, which is funded by the Australian Government's Department of Communications, Information Technology, and the Arts, and the Australian Research Council through the Backing Australia's Ability initiative and the ICT Centre of Excellence Program.

has recently been given in [17]. The explanation exploits ideas from graph theory [13] and from the theory of nonhomogeneous Markov chains [25, 30, 15]. Experience has shown that it is more the graph theory than the Markov chains which is key to this line of research. An illustration of this is the recent extension of the findings of [17] which explain the behavior of Reynolds' full nonlinear "boid" system [26].

Mathematically Vicsek's problem is what in statistics and computer science is called a "consensus problem" [10] or an "agreement problem" [21], although in computer science the issues tend to be concerned more with fault tolerance [12] rather than convergence. Roughly speaking, one has a group of agents which are all trying to agree on a specific value of some quantity. Each agent initially has only limited information available. The agents then try to reach a consensus by communicating what they know to their neighbors either just once or repeatedly, depending on the specific problem of interest. For the Vicsek problem, each agent knows only its own heading and the headings of its current neighbors. One feature of the Vicsek problem which sharply distinguishes it from other consensus problems is that each agent's neighbors can change with time, because all agents are in motion. The theoretical consequence of this is profound: it renders essentially useless, without elaboration, a large body of literature appropriate to the convergence analysis of "nearest neighbor" algorithms with fixed neighbor relationships. Said differently, for the linear heading update rules considered in this paper, understanding the difference between fixed neighbor relationships and changing neighbor relationships is much the same as understanding the difference between the stability of time-invariant linear systems and time-varying linear systems. Various mathematically similar versions of Vicsek's problem have been addressed in the literature [17, 19, 22, 1, 3]; some it turns out well before Vicsek's own paper was published [10, 9, 27, 28, 2].

The central aim of this paper is to establish a number of basic properties of "compositions" of sequences of directed graphs which, as shown in [7], are useful in explaining how a consensus is achieved in various settings. To motivate the graph-theoretic questions addressed and to demonstrate the utility of the answers obtained, we reconsider the version of the Vicsek consensus problem studied by Moreau [19] and Ren and Beard [22]. We derive a condition for agents to reach a consensus exponentially fast which is slightly different than but equivalent to the condition established in [19]. What this paper contributes, then, is a different approach to the understanding of the consensus phenomenon, one in which graphs and their compositions are at center stage. Of course if the consensus problem studied in [19, 22] were the only problem to which this approach were applicable, its development would have hardly been worth the effort. In a sequel to this paper [7] and elsewhere [4, 8, 6, 5] it is demonstrated that in fact the graph-theoretic approach we are advocating is applicable to a broad range of consensus problems which have so far either been only partially resolved or not studied at all.

To the best of our knowledge, all of the statements in this paper about graph compositions are original. However, because the literature on nonhomogeneous Markov chains is vast, some of these statements can undoubtedly be shown to be equivalent to statements about stochastic matrix product in the existing literature [25, 15]. The main convergence result on leaderless flocking, namely Theorem 3, is equivalent to one of the main results of [19]. Corollary 1 is in essence the main result of [17].

In section 2 we reconsider the leaderless coordination problem studied in [17] but without the assumption that the agents all have the same sensing radii. Agents are labelled 1 to  $n$  and are represented by correspondingly labelled vertices in a directed

graph  $\mathbb{N}$  whose arcs represent current neighbor relationships. We define the concept of a “strongly rooted graph” and show by an elementary argument that convergence to a common heading is achieved if the neighbor graphs encountered along a system trajectory are all strongly rooted. We also derive a worst case convergence rate for these types of trajectories. We next define the concept of a “rooted graph” and the operation of “graph composition.” The directed graphs appropriate to the Vicsek model have self-arcs at all vertices. We prove that any composition of  $(n-1)^2$  such rooted graphs is strongly rooted. Armed with this fact, we establish conditions under which consensus is achieved which are different than but equivalent to those obtained in [19, 22]. We then turn to a more in-depth study of rooted graphs. We prove that a so-called neighbor-shared graph is a special type of rooted graph and in so doing make a connection between the consensus problem under consideration and the elegant theory of “scrambling matrices” found in the literature on nonhomogeneous Markov chains [25, 15]. By exploiting this connection in [7], we are able to derive worst case convergence rate results for several versions of the Vicsek problem. The nonhomogeneous Markov chain literature also contains interesting convergence results for a class of stochastic matrices studied by Sarymsakov [24]. The class of Sarymsakov matrices is bigger than the class of all stochastic scrambling matrices. We make contact with this literature by proving that the graph of any Sarymsakov matrix is rooted and also that any stochastic matrix with a rooted graph whose vertices all have self-arcs is a Sarymsakov matrix.

**2. Leaderless coordination.** The system to be studied consists of  $n$  autonomous agents, labelled 1 through  $n$ , all moving in the plane with the same speed but with different headings. Each agent’s heading is updated using a simple local rule based on the average of its own heading plus the headings of its “neighbors.” Agent  $i$ ’s *neighbors* at time  $t$  are those agents, including itself, which are in a closed disk of prespecified radius  $r_i$  centered at agent  $i$ ’s current position. In what follows  $\mathcal{N}_i(t)$  denotes the set of labels of those agents which are neighbors of agent  $i$  at time  $t$ . Agent  $i$ ’s heading, written  $\theta_i$ , evolves in discrete time in accordance with a model of the form

$$(1) \quad \theta_i(t+1) = \frac{1}{n_i(t)} \left( \sum_{j \in \mathcal{N}_i(t)} \theta_j(t) \right),$$

where  $t$  is a discrete-time index taking values in the nonnegative integers  $\{0, 1, 2, \dots\}$ , and  $n_i(t)$  is the number of neighbors of agent  $i$  at time  $t$ .

**2.1. Neighbor graph.** The explicit form of the update equations determined by (1) depends on the relationships between neighbors which exist at time  $t$ . These relationships can be conveniently described by a directed graph  $\mathbb{N}(t)$  with vertex set  $\mathcal{V} = \{1, 2, \dots, n\}$  and “arc set”  $\mathcal{A}(\mathbb{N}(t)) \subset \mathcal{V} \times \mathcal{V}$  which is defined so that  $(i, j)$  is an *arc* or directed edge from  $i$  to  $j$  just in case agent  $i$  is a neighbor of agent  $j$ . Thus  $\mathbb{N}(t)$  is a directed graph on  $n$  vertices with at most one arc connecting each ordered pair of distinct vertices and with exactly one self-arc at each vertex. We write  $\mathcal{G}_{sa}$  for the set of all such graphs and  $\mathcal{G}$  for the set of all directed graphs with vertex set  $\mathcal{V}$ . It is natural to call a vertex  $i$  a *neighbor* of vertex  $j$  in  $\mathbb{G} \in \mathcal{G}$  if  $(i, j)$  is an arc in  $\mathbb{G}$ . In addition we sometimes refer to a vertex  $k$  as an *observer* of vertex  $j$  in  $\mathbb{G}$  if  $(j, k)$  is an arc in  $\mathbb{G}$ . Thus every vertex of  $\mathbb{G}$  can *observe* its neighbors, which with the interpretation of vertices as agents is precisely the kind of relationship  $\mathbb{G}$  is supposed to represent.

**2.2. State equation.** The set of agent heading update rules defined by (1) can be written in state form. Towards this end, for each graph  $\mathbb{N} \in \mathcal{G}_{sa}$ , define the *flocking matrix*

$$(2) \quad F = D^{-1}A',$$

where  $A'$  is the transpose of the “adjacency matrix” of  $\mathbb{N}$  and  $D$  the diagonal matrix whose  $j$ th diagonal element is the “in-degree” of vertex  $j$  within the graph.<sup>1</sup> The function  $\mathbb{N} \mapsto F$  is bijective. Then

$$(3) \quad \theta(t+1) = F(t)\theta(t), \quad t \in \{0, 1, 2, \dots\},$$

where  $\theta$  is the heading vector  $\theta = [\theta_1 \ \theta_2 \ \dots \ \theta_n]'$  and  $F(t)$  is the flocking matrix of the *neighbor graph*  $\mathbb{N}(t)$  which represents the neighbor relationships of (1) at time  $t$ . A complete description of this system would have to include a model which explains how  $\mathbb{N}(t)$  changes over time as a function of the positions of the  $n$  agents in the plane. While such a model is easy to derive and is essential for simulation purposes, it would be difficult to take into account in a convergence analysis. To avoid this difficulty, we shall adopt a more conservative approach which ignores how  $\mathbb{N}(t)$  depends on the agent positions in the plane and assumes instead that  $t \mapsto \mathbb{N}(t)$  might be any signal in some suitably defined set of interest.

Our ultimate goal is to show for a large class of signals  $t \mapsto \mathbb{N}(t)$  and for any initial set of agent headings that the headings of all  $n$  agents will converge to the same steady state value  $\theta_{ss}$ . Convergence of the  $\theta_i$  to  $\theta_{ss}$  is equivalent to the state vector  $\theta$  converging to a vector of the form  $\theta_{ss}\mathbf{1}$ , where  $\mathbf{1} \triangleq [1 \ 1 \ \dots \ 1]_{n \times 1}'$ . Naturally there are situations where convergence to a common heading cannot occur. The most obvious of these is when one agent—say the  $i$ th—starts so far away from the rest that it never acquires any neighbors. Mathematically this would mean not only that  $\mathbb{N}(t)$  is never strongly connected<sup>2</sup> at any time  $t$  but also that vertex  $i$  remains an isolated vertex of  $\mathbb{N}(t)$  for all  $t$  in the sense that within each  $\mathbb{N}(t)$ , vertex  $i$  has no incoming arcs other than its own self-arc. This situation is likely to be encountered if the  $r_i$  are very small. At the other extreme, which is likely if the  $r_i$  are very large, all agents might remain neighbors of all others for all time. In this case,  $\mathbb{N}(t)$  would remain fixed along such a trajectory as the complete graph. Convergence of  $\theta$  to  $\theta_{ss}\mathbf{1}$  can easily be established in this special case because with  $\mathbb{N}(t)$  so fixed, (3) is a linear, time-invariant, discrete-time system. The situation of perhaps the greatest interest is between these two extremes when  $\mathbb{N}(t)$  is not necessarily complete or even strongly connected for any  $t \geq 0$  but when no strictly proper subset of  $\mathbb{N}(t)$ ’s vertices is isolated from the rest for all time. Establishing convergence in this case is challenging because  $F(t)$  changes with time and (3) is not time-invariant. It is this case which we intend to study.

**2.3. Strongly rooted graphs.** In what follows we will call a vertex  $i$  of a directed graph  $\mathbb{G}$  a *root* of  $\mathbb{G}$  if for each other vertex  $j$  of  $\mathbb{G}$ , there is a path from  $i$  to

<sup>1</sup>By the *adjacency matrix* of a directed graph  $\mathbb{G} \in \mathcal{G}$  we mean an  $n \times n$  matrix whose  $ij$ th entry is 1 if  $(i, j)$  is an arc in  $\mathcal{A}(\mathbb{G})$  and 0 if it is not. The *in-degree* of vertex  $j$  in  $\mathbb{G}$  is the number of arcs in  $\mathcal{A}(\mathbb{G})$  of the form  $(i, j)$ ; thus  $j$ ’s in-degree is the number of *incoming* arcs to vertex  $j$ .

<sup>2</sup>A directed graph  $\mathbb{G} \in \mathcal{G}$  with arc set  $\mathcal{A}$  is *strongly connected* if it has a “path” between each distinct pair of its vertices  $i$  and  $j$ ; by a *path* (of *length*  $m$ ) between vertices  $i$  and  $j$  we mean a sequence of arcs in  $\mathcal{A}$  of the form  $(i, k_1), (k_1, k_2), \dots, (k_{m-1}, k_m)$ , where  $k_m = j$  and, if  $m > 1$ ,  $i, k_1, \dots, k_{m-1}$  are distinct vertices.  $\mathbb{G}$  is *complete* if it has a path of length one (i.e., an arc) between each distinct pair of its vertices.



$j$ . Thus  $i$  is a root of  $\mathbb{G}$  if it is the root of a directed spanning tree of  $\mathbb{G}$ . We will say that  $\mathbb{G}$  is *rooted at  $i$*  if  $i$  is in fact a root. Thus  $\mathbb{G}$  is rooted at  $i$  just in case each other vertex of  $\mathbb{G}$  is *reachable* from vertex  $i$  along a path within the graph.  $\mathbb{G}$  is *strongly rooted at  $i$*  if each other vertex of  $\mathbb{G}$  is reachable from vertex  $i$  along a path of length 1. Thus  $\mathbb{G}$  is strongly rooted at  $i$  if  $i$  is a neighbor of every other vertex in the graph. A *rooted graph*  $\mathbb{G}$  is a directed graph which possesses at least one root. Finally, a *strongly rooted graph* is a graph which has at least one vertex at which it is strongly rooted. It is now possible to state the following elementary convergence result which illustrates, under a restrictive assumption, the more general types of results to be derived later in the paper.

**THEOREM 1.** *Let  $\theta(0)$  be fixed. For any trajectory of the system (3) along which each graph in the sequence of neighbor graphs  $\mathbb{N}(0), \mathbb{N}(1), \dots$  is strongly rooted, there is a constant steady state heading  $\theta_{ss}$  for which*

$$(4) \quad \lim_{t \rightarrow \infty} \theta(t) = \theta_{ss} \mathbf{1},$$

where the limit is approached exponentially fast.

**2.3.1. Stochastic matrices.** In order to explain why Theorem 1 is true, we will make use of certain structural properties of the flocking matrices determined by the neighbor graphs in  $\mathcal{G}_{sa}$ . As defined, each flocking matrix  $F$  is square and nonnegative, where by a *nonnegative* matrix we mean a matrix whose entries are all nonnegative. Each  $F$  also has the property that its row sums all equal 1 (i.e.,  $F\mathbf{1} = \mathbf{1}$ ). Matrices with these two properties are called (row) *stochastic* [16]. It is easy to verify that the class of all  $n \times n$  stochastic matrices is closed under multiplication. It is worth noting that because the vertices of the graphs in  $\mathcal{G}_{sa}$  all have self-arcs, the  $F$  also have the property that their diagonal elements are positive. While the proof of Theorem 1 does not exploit this property, the more general results derived later in the paper depend crucially on it.

In what follows we write  $M \geq N$  whenever  $M - N$  is a nonnegative matrix. We also write  $M > N$  whenever  $M - N$  is a positive matrix where by a *positive matrix* we mean a matrix with all positive entries.

**2.3.2. Products of stochastic matrices.** Stochastic matrices have been extensively studied in the literature for a long time largely because of their connection with Markov chains [25, 30, 14]. One problem studied which is of particular relevance here is to describe the asymptotic behavior of products of  $n \times n$  stochastic matrices of the form

$$S_j S_{j-1} \cdots S_1$$

as  $j$  tends to infinity. This is equivalent to looking at the asymptotic behavior of all solutions to the recursion equation

$$(5) \quad x(j+1) = S_j x(j)$$

since any solution  $x(j)$  can be written as

$$x(j) = (S_j S_{j-1} \cdots S_1) x(1), \quad j \geq 1.$$

One especially useful idea, which goes back at least to [11] and has been extensively used [27], is to consider the behavior of the scalar-valued nonnegative function  $V(x) =$

$\lceil x \rceil - \lfloor x \rfloor$  along solutions to (5), where  $x = [x_1 \ x_2 \ \dots \ x_n]'$  is a nonnegative  $n$  vector and  $\lceil x \rceil$  and  $\lfloor x \rfloor$  are its largest and smallest elements, respectively. The key observation is that for any  $n \times n$  stochastic matrix  $S$ , the  $i$ th entry of  $Sx$  satisfies

$$\sum_{j=1}^n s_{ij}x_j \geq \sum_{j=1}^n s_{ij}\lfloor x \rfloor = \lfloor x \rfloor$$

and

$$\sum_{j=1}^n s_{ij}x_j \leq \sum_{j=1}^n s_{ij}\lceil x \rceil = \lceil x \rceil.$$

Since these inequalities hold for all rows of  $Sx$ , it must be true that  $\lfloor Sx \rfloor \geq \lfloor x \rfloor$ , that  $\lceil Sx \rceil \leq \lceil x \rceil$ , and, as a consequence, that  $V(Sx) \leq V(x)$ . These inequalities and (5) imply that the sequences

$$\lfloor x(1) \rfloor, \lfloor x(2) \rfloor, \dots, \quad \lceil x(1) \rceil, \lceil x(2) \rceil, \dots, \quad V(x(1)), V(x(2)), \dots$$

are each monotone. Thus because each of these sequences is also bounded, the limits

$$\lim_{j \rightarrow \infty} \lfloor x(j) \rfloor, \quad \lim_{j \rightarrow \infty} \lceil x(j) \rceil, \quad \lim_{j \rightarrow \infty} V(x(j))$$

each exist. Note that whenever the limit of  $V(x(j))$  is zero, all components of  $x(j)$  together with  $\lfloor x(j) \rfloor$  and  $\lceil x(j) \rceil$  must tend to the same constant value.

There are various different ways in which one might approach the problem of developing conditions under which  $x(j)$  converges to some scalar multiple of  $\mathbf{1}$  or equivalently  $S_j S_{j-1} \cdots S_1$  converges to a constant matrix of the form  $\mathbf{1}c$  for some constant row vector  $c$ . For example, since for any  $n \times n$  stochastic matrix  $S$ ,  $S\mathbf{1} = \mathbf{1}$ , it must be true that  $\text{span}\{\mathbf{1}\}$  is an  $S$ -invariant subspace for any such  $S$ . From this and standard existence conditions for solutions to linear algebraic equation, it follows that for any  $(n-1) \times n$  matrix  $P$  with kernel spanned by  $\mathbf{1}$ , the equation  $PS = \tilde{S}P$  has unique solutions  $\tilde{S}$ , and, moreover, that

$$(6) \quad \text{spectrum } S = \{1\} \cup \text{spectrum } \tilde{S}.$$

As a consequence of the equation  $PS_j = \tilde{S}_j P$ ,  $j \geq 1$ , it can easily be seen that

$$\tilde{S}_j \tilde{S}_{j-1} \cdots \tilde{S}_1 P = PS_j S_{j-1} \cdots S_1.$$

Since  $P$  has full row rank and  $P\mathbf{1} = 0$ , the convergence of a product  $S_j S_{j-1} \cdots S_1$  to a matrix of the form  $\mathbf{1}c$  is equivalent to convergence of the corresponding product  $\tilde{S}_j \tilde{S}_{j-1} \cdots \tilde{S}_1$  to the zero matrix. There are two problems with this approach. First, since  $P$  is not unique, neither are the  $\tilde{S}_i$ . Second, it is not so clear how to go about picking  $P$  to make tractable the problem of proving that the resulting product  $\tilde{S}_j \tilde{S}_{j-1} \cdots \tilde{S}_1$  tends to zero. Tractability of the latter problem generally boils down to choosing a norm for which the  $\tilde{S}_i$  are all contractive. For example, one might seek to choose a suitably weighted 2-norm. This is in essence the same thing as choosing a common quadratic Lyapunov function. Although each  $\tilde{S}_i$  can easily be shown to be discrete-time stable with all eigenvalues of magnitude less than 1, it is known that there are classes of  $S_i$  which give rise to  $\tilde{S}_i$  for which no such common Lyapunov

matrix exists [18] regardless of the choice of  $P$ . Of course there are many other possible norms to choose from other than 2-norms. In the end, success with this approach requires one to simultaneously choose *both* a suitable  $P$  and an appropriate norm with respect to which the  $\tilde{S}_i$  are all contractive. In what follows we adopt a slightly different but closely related approach which ensures that we can work with what is perhaps the most natural norm for this type of convergence problem, the infinity norm.

To proceed, we need a few more ideas concerned with nonnegative matrices. For any nonnegative matrix  $R$  of any size, we write  $\|R\|$  for the largest of the row sums of  $R$ . Note that  $\|R\|$  is the induced infinity norm of  $R$  and consequently is submultiplicative. Note in addition that  $\|x\| = \lceil x \rceil$  for any nonnegative  $n$  vector  $x$ . Moreover,  $\|M_1\| \leq \|M_2\|$  if  $M_1 \leq M_2$ . Observe that for any  $n \times n$  stochastic matrix  $S$ ,  $\|S\| = 1$  because the row sums of a stochastic matrix all equal 1. We extend the domain of definitions of  $\lfloor \cdot \rfloor$  and  $\lceil \cdot \rceil$  to the class of all nonnegative  $n \times m$  matrix  $M$  by letting  $\lfloor M \rfloor$  and  $\lceil M \rceil$  now denote the  $1 \times m$  row vectors whose  $j$ th entries are the smallest and largest elements, respectively, of the  $j$ th column of  $M$ . Note that  $\lfloor M \rfloor$  is the largest  $1 \times m$  nonnegative row vector  $c$  for which  $M - \mathbf{1}c$  is nonnegative and that  $\lceil M \rceil$  is the smallest nonnegative row vector  $c$  for which  $\mathbf{1}c - M$  is nonnegative. Note in addition that for any  $n \times n$  stochastic matrix  $S$ , one can write

$$(7) \quad S = \mathbf{1} \lfloor S \rfloor + \lceil S \rceil \quad \text{and} \quad S = \mathbf{1} \lceil S \rceil - \lfloor S \rfloor,$$

where  $\lfloor S \rfloor$  and  $\lceil S \rceil$  are nonnegative matrices defined by the equations

$$(8) \quad \lfloor S \rfloor = S - \mathbf{1} \lfloor S \rfloor \quad \text{and} \quad \lceil S \rceil = \mathbf{1} \lceil S \rceil - S,$$

respectively. Moreover, the row sums of  $\lfloor S \rfloor$  are all equal to  $1 - \lfloor S \rfloor \mathbf{1}$  and the row sums of  $\lceil S \rceil$  are all equal to  $\lceil S \rceil \mathbf{1} - 1$ , and so

$$(9) \quad \|\lfloor S \rfloor\| = 1 - \lfloor S \rfloor \mathbf{1} \quad \text{and} \quad \|\lceil S \rceil\| = \lceil S \rceil \mathbf{1} - 1.$$

In what follows we will also be interested in the matrix

$$(10) \quad \llbracket S \rrbracket = \lfloor S \rfloor + \lceil S \rceil.$$

This matrix satisfies

$$(11) \quad \llbracket S \rrbracket = \mathbf{1}(\lceil S \rceil - \lfloor S \rfloor)$$

because of (7).

For any infinite sequence of  $n \times n$  stochastic matrices  $S_1, S_2, \dots$ , we henceforth use the symbol  $\lfloor \dots S_j \dots S_1 \rfloor$  to denote the limit

$$(12) \quad \lfloor \dots S_j \dots S_2 S_1 \rfloor = \lim_{j \rightarrow \infty} \lfloor S_j \dots S_2 S_1 \rfloor.$$

From the preceding discussion it is clear that for  $i \in \{1, 2, \dots, n\}$ , the limit  $\lfloor \dots S_j \dots S_1 \rfloor e_i$  exists, where  $e_i$  is the  $i$ th unit  $n$ -vector. Thus the limit  $\lfloor \dots S_j \dots S_1 \rfloor$  always exists, and this is true even if the product  $S_j \dots S_2 S_1$  itself does not have a limit. Two situations can occur. Either the product  $S_j \dots S_2 S_1$  converges to a rank one matrix or it does not. In fact, even if  $S_j \dots S_2 S_1$  does converge, it is quite possible that the limit is not a rank one matrix. An example of this would be a sequence in which  $S_1$  is any stochastic matrix of rank greater than 1 and for all  $i > 1$ ,  $S_i = I_{n \times n}$ . In what follows we will develop sufficient conditions for  $S_j \dots S_2 S_1$  to converge to a rank one

matrix as  $j \rightarrow \infty$ . Note that if this occurs, then the limit must be of the form  $\mathbf{1}c$ , where  $c\mathbf{1} = 1$  because stochastic matrices are closed under multiplication.

In what follows we will say that a matrix product  $S_j S_{j-1} \cdots S_1$  converges to  $\mathbf{1}[\cdots S_j \cdots S_1]$  exponentially fast at a rate no slower than  $\lambda$  if there are nonnegative constants  $b$  and  $\lambda$  with  $\lambda < 1$ , such that

$$(13) \quad \|(S_j \cdots S_1) - \mathbf{1}[\cdots S_j \cdots S_2 S_1]\| \leq b\lambda^j, \quad j \geq 1.$$

The following proposition implies that such a stochastic matrix product will so converge if  $\|S_j \cdots S_1\|$  converges to 0.

**PROPOSITION 1.** *Let  $\bar{b}$  and  $\lambda$  be nonnegative numbers with  $\lambda < 1$ . Suppose that  $S_1, S_2, \dots$  is an infinite sequence of  $n \times n$  stochastic matrices for which*

$$(14) \quad \|[S_j \cdots S_1]\| \leq \bar{b}\lambda^j, \quad j \geq 0.$$

*Then the matrix product  $S_j \cdots S_2 S_1$  converges to  $\mathbf{1}[\cdots S_j \cdots S_1]$  exponentially fast at a rate no slower than  $\lambda$ .*

The proof of Proposition 1 makes use of the first of the two inequalities which follow.

**LEMMA 1.** *For any two  $n \times n$  stochastic matrices  $S_1$  and  $S_2$ ,*

$$(15) \quad \lfloor S_2 S_1 \rfloor - \lfloor S_1 \rfloor \leq \lceil S_2 \rceil \lfloor S_1 \rfloor,$$

$$(16) \quad \lceil S_2 S_1 \rceil \leq \lceil S_2 \rceil \lceil S_1 \rceil.$$

*Proof of Lemma 1.* Since  $S_2 S_1 = S_2(\mathbf{1}\lfloor S_1 \rfloor + \lceil S_1 \rceil) = \mathbf{1}\lfloor S_1 \rfloor + S_2 \lceil S_1 \rceil$  and  $S_2 = \mathbf{1}\lceil S_2 \rceil - \lfloor S_2 \rfloor$ , it must be true that  $S_2 S_1 = \mathbf{1}(\lfloor S_1 \rfloor + \lceil S_2 \rceil \lfloor S_1 \rfloor) - \lfloor S_2 \rfloor \lceil S_1 \rceil$ . Thus  $\mathbf{1}(\lfloor S_1 \rfloor + \lceil S_2 \rceil \lfloor S_1 \rfloor) - \lfloor S_2 \rfloor \lceil S_1 \rceil$  is nonnegative. But  $\lceil S_2 S_1 \rceil$  is the smallest nonnegative row vector  $c$  for which  $\mathbf{1}c - S_2 S_1$  is nonnegative. Therefore

$$(17) \quad \lceil S_2 S_1 \rceil \leq \lfloor S_1 \rfloor + \lceil S_2 \rceil \lfloor S_1 \rfloor.$$

Moreover,  $\lfloor S_2 S_1 \rfloor \leq \lceil S_2 S_1 \rceil$  because of the definitions of  $\lfloor \cdot \rfloor$  and  $\lceil \cdot \rceil$ . This and (17) imply that  $\lfloor S_2 S_1 \rfloor \leq \lfloor S_1 \rfloor + \lceil S_2 \rceil \lfloor S_1 \rfloor$  and thus that (15) is true.

Since  $S_2 S_1 = S_2(\mathbf{1}\lfloor S_1 \rfloor + \lceil S_1 \rceil) = \mathbf{1}\lfloor S_1 \rfloor + S_2 \lceil S_1 \rceil$  and  $S_2 = \lfloor S_2 \rfloor + \lceil S_2 \rceil$ , it must be true that  $S_2 S_1 = \mathbf{1}(\lfloor S_1 \rfloor + \lfloor S_2 \rfloor \lceil S_1 \rceil) + \lceil S_2 \rceil \lceil S_1 \rceil$ . Thus  $S_2 S_1 - \mathbf{1}(\lfloor S_1 \rfloor + \lfloor S_2 \rfloor \lceil S_1 \rceil)$  is nonnegative. But  $\lfloor S_2 S_1 \rfloor$  is the largest nonnegative row vector  $c$  for which  $S_2 S_1 - \mathbf{1}c$  is nonnegative, and so

$$(18) \quad S_2 S_1 \leq \mathbf{1}\lfloor S_2 S_1 \rfloor + \lceil S_2 \rceil \lceil S_1 \rceil.$$

Now it is also true that  $S_2 S_1 = \mathbf{1}\lfloor S_2 S_1 \rfloor + \lceil S_2 S_1 \rceil$ . From this and (18) it follows that (16) is true.  $\square$

*Proof of Proposition 1.* Set  $X_j = S_j \cdots S_1, j \geq 1$ , and note that each  $X_j$  is a stochastic matrix. In view of (15),

$$\lfloor X_{j+1} \rfloor - \lfloor X_j \rfloor \leq \lceil S_{j+1} \rceil \lfloor X_j \rfloor, \quad j \geq 1.$$

By hypothesis,  $\|[X_j]\| \leq \bar{b}\lambda^j, j \geq 1$ . Moreover,  $\|\lceil S_{j+1} \rceil\| \leq n$  because all entries in  $S_{j+1}$  are bounded above by 1. Therefore

$$(19) \quad \|\lfloor X_{j+1} \rfloor - \lfloor X_j \rfloor\| \leq n\bar{b}\lambda^j, \quad j \geq 1.$$

Clearly

$$\lfloor X_{j+i} \rfloor - \lfloor X_j \rfloor = \sum_{k=1}^i (\lfloor X_{i+j+1-k} \rfloor - \lfloor X_{i+j-k} \rfloor), \quad i, j \geq 1.$$

Thus, by the triangle inequality

$$||\lfloor X_{j+i} \rfloor - \lfloor X_j \rfloor|| \leq \sum_{k=1}^i ||\lfloor X_{i+j+1-k} \rfloor - \lfloor X_{i+j-k} \rfloor||, \quad i, j \geq 1.$$

This and (19) imply that

$$||\lfloor X_{j+i} \rfloor - \lfloor X_j \rfloor|| \leq n\bar{b} \sum_{k=1}^i \lambda^{(i+j-k)}, \quad i, j \geq 1.$$

Now

$$\sum_{k=1}^i \lambda^{(i+j-k)} = \lambda^j \sum_{k=1}^i \lambda^{(i-k)} = \lambda^j \sum_{q=1}^i \lambda^{q-1} \leq \lambda^j \sum_{q=1}^{\infty} \lambda^{q-1}.$$

But  $\lambda < 1$ , and so

$$\sum_{q=1}^{\infty} \lambda^{q-1} = \frac{1}{1-\lambda}.$$

Therefore

$$(20) \quad ||\lfloor X_{i+j} \rfloor - \lfloor X_j \rfloor|| \leq n\bar{b} \frac{\lambda^j}{1-\lambda}, \quad i, j \geq 1.$$

Set  $c = \lfloor \cdots S_j \cdots S_1 \rfloor$  and note that

$$||\lfloor X_j \rfloor - c|| = ||\lfloor X_j \rfloor - \lfloor X_{i+j} \rfloor + \lfloor X_{i+j} \rfloor - c|| \leq ||\lfloor X_j \rfloor - \lfloor X_{i+j} \rfloor|| + ||\lfloor X_{i+j} \rfloor - c||, \quad i, j \geq 1.$$

In view of (20)

$$||\lfloor X_j \rfloor - c|| \leq n\bar{b} \frac{\lambda^j}{1-\lambda} + ||\lfloor X_{i+j} \rfloor - c||, \quad i, j \geq 1.$$

Since

$$\lim_{i \rightarrow \infty} ||\lfloor X_{i+j} \rfloor - c|| = 0$$

it must be true that

$$||\lfloor X_j \rfloor - c|| \leq n\bar{b} \frac{\lambda^j}{1-\lambda}, \quad j \geq 1.$$

But  $||\mathbf{1}(\lfloor X_j \rfloor - c)|| = ||\lfloor X_j \rfloor - c||$  and  $X_j = S_j \cdots S_1$ . Therefore

$$(21) \quad ||\mathbf{1}(\lfloor S_j \cdots S_1 \rfloor - c)|| \leq n\bar{b} \frac{\lambda^j}{1-\lambda}, \quad j \geq 1.$$

In view of (7)

$$S_j \cdots S_1 = \mathbf{1} \lfloor S_j \cdots S_1 \rfloor + \lVert S_j \cdots S_1 \rVert, \quad j \geq 1.$$

Therefore

$$\begin{aligned} \|(S_j \cdots S_1) - \mathbf{1}c\| &= \|\mathbf{1} \lfloor S_j \cdots S_1 \rfloor + \lVert S_j \cdots S_1 \rVert - \mathbf{1}c\| \\ &\leq \|\mathbf{1} \lfloor S_j \cdots S_1 \rfloor - \mathbf{1}c\| + \|\lVert S_j \cdots S_1 \rVert\|, \quad j \geq 1. \end{aligned}$$

From this, (14), and (21) it follows that

$$\|S_j \cdots S_1 - \mathbf{1}c\| \leq \bar{b} \left(1 + \frac{n}{1-\lambda}\right) \lambda^j, \quad j \geq 1,$$

and thus that (13) holds with  $b = \bar{b}(1 + \frac{n}{1-\lambda})$ .  $\square$

**2.3.3. Convergence.** We are now in a position to make some statements about the asymptotic behavior of a product of  $n \times n$  stochastic matrices of the form  $S_j S_{j-1} \cdots S_1$  as  $j$  tends to infinity. Note first that (16) generalizes to sequences of stochastic matrices of any length. Thus

$$(22) \quad \lfloor S_j S_{j-1} \cdots S_2 S_1 \rfloor \leq \lfloor S_j \rfloor \lfloor S_{j-1} \rfloor \cdots \lfloor S_1 \rfloor.$$

It is therefore clear that condition (14) of Proposition 1 will hold with  $\bar{b} = 1$  if

$$(23) \quad \|\lVert S_j \rVert\| \cdots \lVert S_1 \rVert\| \leq \lambda^j$$

for some nonnegative number  $\lambda < 1$ . Because  $\|\cdot\|$  is submultiplicative, this means that a product of stochastic matrices  $S_j \cdots S_1$  will converge to a limit of the form  $\mathbf{1}c$  for some constant row vector  $c$  if each of the matrices  $S_i$  in the sequence  $S_1, S_2, \dots$  satisfies the norm bound  $\|\lVert S_i \rVert\| \leq \lambda$ . We now develop a condition, tailored to our application, for this to be so.

As a first step it is useful to characterize those stochastic matrices  $S$  for which  $\|\lVert S \rVert\| < 1$ . Note that this condition is equivalent to the requirement that the row sums of  $\lfloor S \rfloor$  are less than 1. This, in turn, is equivalent to the requirement that  $\mathbf{1} \lfloor S \rfloor \neq 0$  since  $\lVert S \rVert = S - \mathbf{1} \lfloor S \rfloor$ . Now  $\mathbf{1} \lfloor S \rfloor \neq 0$  if and only if  $S$  has at least one nonzero column since the indices of the nonzero columns of  $S$  are the same as the indices of the nonzero columns of  $\lfloor S \rfloor$ . Thus  $\|\lVert S \rVert\| < 1$  if and only if  $S$  has at least one nonzero column. For our purposes it proves to be especially useful to restate this condition in equivalent graph theoretic terms. For this we need the following definition.

*The graph of a stochastic matrix.* For any  $n \times n$  stochastic matrix  $S$ , let  $\gamma(S)$  denote the graph  $\mathbb{G} \in \mathcal{G}$  whose adjacency matrix is the transpose of the matrix obtained by replacing all of  $S$ 's nonzero entries with 1's. The graph-theoretic condition is as follows.

LEMMA 2. *A stochastic matrix  $S$  has a strongly rooted graph  $\gamma(S)$  if and only if*

$$(24) \quad \|\lVert S \rVert\| < 1.$$

*Proof.* Let  $A$  be the adjacency matrix of  $\gamma(S)$ . Since the positions of the nonzero entries of  $S$  and  $A$  are the same, the  $i$ th column of  $S$  will be positive if and only if  $A$ 's  $i$ th row is positive. Thus (23) will hold just in case  $A$  has a positive row. But strongly

rooted graphs in  $\mathcal{G}$  are precisely those graphs whose adjacency matrices have at least one positive row. Therefore (23) will hold if and only if  $\gamma(S)$  is strongly rooted.  $\square$

Lemma 2 can be used to prove the following.

**PROPOSITION 2.** *Let  $\mathcal{S}_{sr}$  be any closed set of stochastic matrices which are all the same size and whose graphs  $\gamma(S)$ ,  $S \in \mathcal{S}_{sr}$ , are all strongly rooted. Then as  $j \rightarrow \infty$ , any product  $S_j \cdots S_1$  of matrices from  $\mathcal{S}_{sr}$  converges exponentially fast to  $\mathbf{1}[\cdots S_j \cdots S_1]$  at a rate no slower than*

$$\lambda = \max_{S \in \mathcal{S}_{sr}} \|\|S\|\|,$$

where  $\lambda$  is a nonnegative constant satisfying  $\lambda < 1$ .

*Proof of Proposition 2.* In view of Lemma 2,  $\|\|S\|\| < 1$ ,  $S \in \mathcal{S}_{sr}$ . Because  $\mathcal{S}_{sr}$  is closed and bounded and  $\|\|\cdot\|\|$  is continuous,  $\lambda < 1$ . Clearly  $\|\|S_i\|\| \leq \lambda$ ,  $i \geq 1$ , and so (23) must hold for any sequence of matrices  $S_1, S_2, \dots$  from  $\mathcal{S}_{sr}$ . Therefore for any such sequence  $\|\|S_j \cdots S_1\|\| \leq \lambda^j$ ,  $j \geq 0$ . Thus by Proposition 1, the product  $\Pi(j) = S_j S_{j-1} \cdots S_1$  converges to  $\mathbf{1}[\cdots S_j \cdots S_1]$  exponentially fast at a rate no slower than  $\lambda$ .  $\square$

*Proof of Theorem 1.* Let  $\mathcal{F}_{sr}$  denote the set of flocking matrices with strongly rooted graphs. Since  $\mathcal{S}_{sa}$  is a finite set, so is the set of strongly rooted graphs in  $\mathcal{G}_{sa}$ . Therefore  $\mathcal{F}_{sr}$  is closed. By assumption,  $F(t) \in \mathcal{F}_{sr}$ ,  $t \geq 0$ . In view of Proposition 2, the product  $F(t) \cdots F(0)$  converges exponentially fast to  $\mathbf{1}[\cdots F(t) \cdots F(0)]$  at a rate no slower than

$$\lambda = \max_{F \in \mathcal{F}_{sr}} \|\|F\|\|.$$

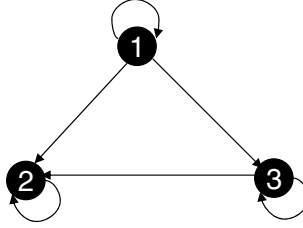
But it is clear from (3) that  $\theta(t) = F(t-1) \cdots F(1)F(0)\theta(0)$ ,  $t \geq 1$ . Therefore (4) holds with  $\theta_{ss} = \mathbf{1}[\cdots F(t) \cdots F(0)]\theta(0)$  and the convergence is exponential.  $\square$

**2.3.4. Convergence rate.** Using (9) it is possible to calculate a worst case value for the convergence rate  $\lambda$  used in the proof of Theorem 1. Fix  $F \in \mathcal{F}_{sr}$ . Because  $\gamma(F)$  is strongly rooted, at least one vertex—say the  $k$ th—must be a root with arcs to each other vertex. In the context of (1), this means that agent  $k$  must be a neighbor of every agent. Thus  $\theta_k$  must be in each sum in (1). Since each  $n_i$  in (1) is bounded above by  $n$ , this means that the smallest element in column  $k$  of  $F$  is bounded below by  $\frac{1}{n}$ . Since (9) asserts that  $\|\|F\|\| = 1 - \lfloor F \rfloor \mathbf{1}$ , it must be true that  $\|\|F\|\| \leq 1 - \frac{1}{n}$ . This holds for all  $F \in \mathcal{F}_{sr}$ . Moreover, in the worst case when  $\mathbb{F}$  is strongly rooted at just one vertex and all vertices are neighbors of at least one common vertex,  $\|\|F\|\| = 1 - \frac{1}{n}$ . It follows that the worst case convergence rate is

$$(25) \quad \max_{F \in \mathcal{F}_{sr}} \|\|F\|\| = 1 - \frac{1}{n}.$$

An example of a graph of a flocking matrix for which (25) holds is shown in Figure 1.

**2.4. Rooted graphs.** The proof of Theorem 1 depends crucially on the fact that the graphs encountered along a trajectory of (3) are all strongly rooted. It is natural to ask if this requirement can be relaxed and still have all agents' headings converge to a common value. The aim of this section is to show that this can indeed be accomplished. To do this we need to have a meaningful way of “combining” sequences of graphs so that only the combined graph need be strongly rooted but not necessarily the individual graphs making up the combination. One possible notion of combination

FIG. 1. *Example.*

of a sequence  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_k$  with the same vertex set  $\mathcal{V}$  would be the graph with vertex set  $\mathcal{V}$  whose arc set is the union of the arc sets of the graphs in the sequence. It turns out that because we are interested in *sequences* of graphs rather than mere *sets* of graphs, a simple union is not quite the appropriate notion for our purposes because a union does not take into account the order in which the graphs are encountered along a trajectory. What is appropriate is a slightly more general notion which we now define.

**2.4.1. Composition of graphs.** By the *composition* of a directed graph  $\mathbb{G}_p \in \mathcal{G}$  with a directed graph  $\mathbb{G}_q \in \mathcal{G}$ , written  $\mathbb{G}_q \circ \mathbb{G}_p$ , we mean the directed graph with the vertex set  $\{1, 2, \dots, n\}$  and arc set defined in such a way so that  $(i, j)$  is an arc of the composition just in case there is a vertex  $k$  such that  $(i, k)$  is an arc of  $\mathbb{G}_p$  and  $(k, j)$  is an arc of  $\mathbb{G}_q$ . Thus  $(i, j)$  is an arc in  $\mathbb{G}_q \circ \mathbb{G}_p$  if and only if  $i$  has an observer in  $\mathbb{G}_p$  which is also a neighbor of  $j$  in  $\mathbb{G}_q$ . Note that  $\mathcal{G}$  is closed under composition and that composition is an associative binary operation; because of this, the definition extends unambiguously to any finite sequence of directed graphs  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_k$ .

If we focus exclusively on graphs with self-arcs at all vertices, namely the graphs in  $\mathcal{G}_{sa}$ , more can be said. In this case the definition of composition implies that the arcs of  $\mathbb{G}_p$  and  $\mathbb{G}_q$  are arcs of  $\mathbb{G}_q \circ \mathbb{G}_p$ . The definition also implies in this case that if  $\mathbb{G}_p$  has a directed path from  $i$  to  $k$  and  $\mathbb{G}_q$  has a directed path from  $k$  to  $j$ , then  $\mathbb{G}_q \circ \mathbb{G}_p$  has a directed path from  $i$  to  $j$ . Both of these implications are consequences of the requirement that the vertices of the graphs in  $\mathcal{G}_{sa}$  all have self-arcs. Note in addition that  $\mathcal{G}_{sa}$  is closed under composition. It is worth emphasizing that the union of the arc sets of a sequence of graphs  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_k$  in  $\mathcal{G}_{sa}$  must be contained in the arc set of their composition. However, the converse is not true in general, and it is for this reason that composition rather than union proves to be the more useful concept for our purposes.

Suppose that  $A_p = [a_{ij}(p)]$  and  $A_q = [a_{ij}(q)]$  are the adjacency matrices of  $\mathbb{G}_p \in \mathcal{G}$  and  $\mathbb{G}_q \in \mathcal{G}$ , respectively. Then the adjacency matrix of the composition  $\mathbb{G}_q \circ \mathbb{G}_p$  must be the matrix obtained by replacing all nonzero elements in  $A_p A_q$  with ones. This is because the  $ij$ th entry of  $A_p A_q$ , namely

$$\sum_{k=1}^n a_{ik}(p) a_{kj}(q),$$

will be nonzero just in case there is at least one value of  $k$  for which both  $a_{ik}(p)$  and  $a_{kj}(q)$  are nonzero. This of course is exactly the condition for the  $ij$ th element of the adjacency matrix of the composition  $\mathbb{G}_q \circ \mathbb{G}_p$  to be nonzero. Note that if  $S_1$  and  $S_2$  are  $n \times n$  stochastic matrices for which  $\gamma(S_1) = \mathbb{G}_p$  and  $\gamma(S_2) = \mathbb{G}_q$ , then the matrix which results by replacing by ones all nonzero entries in the stochastic



matrix  $S_2 S_1$  must be the transpose of the adjacency matrix of  $\mathbb{G}_q \circ \mathbb{G}_p$ . In view of the definition of  $\gamma(\cdot)$ , it therefore must be true that  $\gamma(S_2 S_1) = \gamma(S_2) \circ \gamma(S_1)$ . This obviously generalizes to finite products of stochastic matrices.

LEMMA 3. *For any sequence of stochastic matrices  $S_1, S_2, \dots, S_j$  which are all the same size,*

$$\gamma(S_j \cdots S_1) = \gamma(S_j) \circ \cdots \circ \gamma(S_1).$$

**2.4.2. Compositions of rooted graphs.** We now give several different conditions under which the composition of a sequence of graphs is strongly rooted.

PROPOSITION 3. *Suppose  $n > 1$  and let  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_m}$  be a finite sequence of rooted graphs in  $\mathcal{G}_{sa}$ .*

1. *If  $m \geq (n-1)^2$ , then  $\mathbb{G}_{p_m} \circ \mathbb{G}_{p_{m-1}} \circ \cdots \circ \mathbb{G}_{p_1}$  is strongly rooted.*
2. *If  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_m}$  are all rooted at  $v$  and  $m \geq n-1$ , then  $\mathbb{G}_{p_m} \circ \mathbb{G}_{p_{m-1}} \circ \cdots \circ \mathbb{G}_{p_1}$  is strongly rooted at  $v$ .*

The requirement of assertion 2 above that all the graphs in the sequence be rooted at a single vertex  $v$  is obviously more restrictive than the requirement of assertion 1 that all the graphs be rooted but not necessarily at the same vertex. The price for the less restrictive assumption is that the bound on the number of graphs needed in the more general case is much higher than the bound given in the case in which all the graphs are rooted at  $v$ . It is probably true that the bound  $(n-1)^2$  for the more general case is too conservative, but this remains to be shown. The more special case when all graphs share a common root is relevant to the leader-follower version of the problem which will be discussed later in the paper. Proposition 3 will be proved shortly.

Note that a strongly connected graph is the same as a graph which is rooted at every vertex and that a complete graph is the same as a graph which is strongly rooted at every vertex. In view of these observations and Proposition 3 we can state the following proposition.

PROPOSITION 4. *Suppose  $n > 1$  and let  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_m}$  be a finite sequence of strongly connected graphs in  $\mathcal{G}_{sa}$ . If  $m \geq n-1$ , then  $\mathbb{G}_{p_m} \circ \mathbb{G}_{p_{m-1}} \circ \cdots \circ \mathbb{G}_{p_1}$  is complete.*

To prove Proposition 3 we will need some more ideas. We say that a vertex  $v \in \mathcal{V}$  is an *observer* of a subset  $\mathcal{S} \subset \mathcal{V}$  in a graph  $\mathbb{G} \in \mathcal{G}$  if  $v$  is an observer of at least one vertex in  $\mathcal{S}$ . By the *observer function* of a graph  $\mathbb{G} \in \mathcal{G}$ , written  $\alpha(\mathbb{G}, \cdot)$ , we mean the function  $\alpha(\mathbb{G}, \cdot) : 2^{\mathcal{V}} \rightarrow 2^{\mathcal{V}}$  which assigns to each subset  $\mathcal{S} \subset \mathcal{V}$  the subset of vertices in  $\mathcal{V}$  which are observers of  $\mathcal{S}$  in  $\mathbb{G}$ . Thus  $j \in \alpha(\mathbb{G}, i)$  just in case  $(i, j) \in \mathcal{A}(\mathbb{G})$ . Note that if  $\mathbb{G}_p \in \mathcal{G}$  and  $\mathbb{G}_q$  in  $\mathcal{G}_{sa}$ , then

$$(26) \quad \alpha(\mathbb{G}_p, \mathcal{S}) \subset \alpha(\mathbb{G}_q \circ \mathbb{G}_p, \mathcal{S}), \quad \mathcal{S} \in 2^{\mathcal{V}},$$

because  $\mathbb{G}_q \in \mathcal{G}_{sa}$  implies that the arcs in  $\mathbb{G}_p$  are all arcs in  $\mathbb{G}_q \circ \mathbb{G}_p$ . Observer functions have the following important and easily proved property.

LEMMA 4. *For all  $\mathbb{G}_p, \mathbb{G}_q \in \mathcal{G}$  and any nonempty subset  $\mathcal{S} \subset \mathcal{V}$ ,*

$$(27) \quad \alpha(\mathbb{G}_q, \alpha(\mathbb{G}_p, \mathcal{S})) = \alpha(\mathbb{G}_q \circ \mathbb{G}_p, \mathcal{S}).$$

*Proof.* Suppose first that  $i \in \alpha(\mathbb{G}_q, \alpha(\mathbb{G}_p, \mathcal{S}))$ . Then  $(j, i)$  is an arc in  $\mathbb{G}_q$  for some  $j \in \alpha(\mathbb{G}_p, \mathcal{S})$ . Hence  $(k, j)$  is an arc in  $\mathbb{G}_p$  for some  $k \in \mathcal{S}$ . In view of the definition of composition,  $(k, i)$  is an arc in  $\mathbb{G}_q \circ \mathbb{G}_p$ , and so  $i \in \alpha(\mathbb{G}_q \circ \mathbb{G}_p, \mathcal{S})$ . Since this holds for all  $i \in \mathcal{V}$ ,  $\alpha(\mathbb{G}_q, \alpha(\mathbb{G}_p, \mathcal{S})) \subset \alpha(\mathbb{G}_q \circ \mathbb{G}_p, \mathcal{S})$ .

For the reverse inclusion, fix  $i \in \alpha(\mathbb{G}_q \circ \mathbb{G}_p, \mathcal{S})$  in which case  $(k, i)$  is an arc in  $\mathbb{G}_q \circ \mathbb{G}_p$  for some  $k \in \mathcal{S}$ . By definition of composition, there exists an  $j \in \mathcal{V}$  such that  $(k, j)$  is an arc in  $\mathbb{G}_p$  and  $(j, i)$  is an arc in  $\mathbb{G}_q$ . Thus  $j \in \alpha(\mathbb{G}_p, \mathcal{S})$ . Therefore  $i \in \alpha(\mathbb{G}_q, \alpha(\mathbb{G}_p, \mathcal{S}))$ . Since this holds for all  $i \in \mathcal{V}$ ,  $\alpha(\mathbb{G}_q, \alpha(\mathbb{G}_p, \mathcal{S})) \supset \alpha(\mathbb{G}_q \circ \mathbb{G}_p, \mathcal{S})$ . Therefore (27) is true.  $\square$

To proceed, let us note that each subset  $\mathcal{S} \subset \mathcal{V}$  induces a unique subgraph of  $\mathbb{G}$  with vertex set  $\mathcal{S}$  and arc set  $\mathcal{A}$  consisting of those arcs  $(i, j)$  of  $\mathbb{G}$  for which both  $i$  and  $j$  are vertices of  $\mathcal{S}$ . This, together with the natural partial ordering of  $\mathcal{V}$  by inclusion, provides a corresponding partial ordering of  $\mathcal{G}$ . Thus if  $\mathcal{S}_1$  and  $\mathcal{S}_2$  are subsets of  $\mathcal{V}$  and  $\mathcal{S}_1 \subset \mathcal{S}_2$ , then  $\mathbb{G}_1 \subset \mathbb{G}_2$ , where, for  $i \in \{1, 2\}$ ,  $\mathbb{G}_i$  is the subgraph of  $\mathbb{G}$  induced by  $\mathcal{S}_i$ . For any  $v \in \mathcal{V}$ , there is a unique largest subgraph rooted at  $v$ , namely the graph induced by the vertex set  $\mathcal{V}(v) = \{v\} \cup \alpha(\mathbb{G}, v) \cup \dots \cup \alpha^{n-1}(\mathbb{G}, v)$ , where  $\alpha^i(\mathbb{G}, \cdot)$  denotes the composition of  $\alpha(\mathbb{G}, \cdot)$  with itself  $i$  times. We call this graph the *rooted graph generated by  $v$* . It is clear that  $\mathcal{V}(v)$  is the smallest  $\alpha(\mathbb{G}, \cdot)$ -invariant subset of  $\mathcal{V}$  which contains  $v$ .

The proof of Proposition 3 depends on the following lemma.

LEMMA 5. *Let  $\mathbb{G}_p$  and  $\mathbb{G}_q$  be graphs in  $\mathcal{G}_{sa}$ . If  $\mathbb{G}_q$  is rooted at  $v$  and  $\alpha(\mathbb{G}_p, v)$  is a strictly proper subset of  $\mathcal{V}$ , then  $\alpha(\mathbb{G}_p, v)$  is also a strictly proper subset of  $\alpha(\mathbb{G}_q \circ \mathbb{G}_p, v)$ .*

*Proof of Lemma 5.* In general  $\alpha(\mathbb{G}_p, v) \subset \alpha(\mathbb{G}_q \circ \mathbb{G}_p, v)$  because of (26). Thus if  $\alpha(\mathbb{G}_p, v)$  is not a strictly proper subset of  $\alpha(\mathbb{G}_q \circ \mathbb{G}_p, v)$ , then  $\alpha(\mathbb{G}_p, v) = \alpha(\mathbb{G}_q \circ \mathbb{G}_p, v)$ , and so  $\alpha(\mathbb{G}_q \circ \mathbb{G}_p, v) \subset \alpha(\mathbb{G}_p, v)$ . In view of (27),  $\alpha(\mathbb{G}_q \circ \mathbb{G}_p, v) = \alpha(\mathbb{G}_q, \alpha(\mathbb{G}_p, v))$ . Therefore  $\alpha(\mathbb{G}_q, \alpha(\mathbb{G}_p, v)) \subset \alpha(\mathbb{G}_p, v)$ . Moreover,  $v \in \alpha(\mathbb{G}_p, v)$  because  $v$  has a self-arc in  $\mathbb{G}_p$ . Thus  $\alpha(\mathbb{G}_p, v)$  is a strictly proper subset of  $\mathcal{V}$  which contains  $v$  and is  $\alpha(\mathbb{G}_q, \cdot)$ -invariant. But this is impossible because  $\mathbb{G}_q$  is rooted at  $v$ .  $\square$

*Proof of Proposition 3.* Assertion 2 will be proved first. Suppose that  $m \geq n - 1$  and that  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_m}$  are all rooted at  $v$ . In view of (26),  $\mathcal{A}(\mathbb{G}_{p_k} \circ \mathbb{G}_{p_{k-1}} \circ \dots \circ \mathbb{G}_{p_1}) \subset \mathcal{A}(\mathbb{G}_{p_m} \circ \mathbb{G}_{p_{m-1}} \circ \dots \circ \mathbb{G}_{p_1})$  for any positive integer  $k \leq m$ . Thus  $\mathbb{G}_{p_m} \circ \mathbb{G}_{p_{m-1}} \circ \dots \circ \mathbb{G}_{p_1}$  will be strongly rooted at  $v$  if there exists an integer  $k \leq n - 1$  such that

$$(28) \quad \alpha(\mathbb{G}_{p_k} \circ \mathbb{G}_{p_{k-1}} \circ \dots \circ \mathbb{G}_{p_1}, v) = \mathcal{V}.$$

It will now be shown that such an integer exists.

If  $\alpha(\mathbb{G}_{p_1}, v) = \mathcal{V}$ , set  $k = 1$ , in which case (28) clearly holds. If  $\alpha(\mathbb{G}_{p_1}, v) \neq \mathcal{V}$ , then let  $i > 1$  be the greatest positive integer not exceeding  $n - 1$  for which  $\alpha(\mathbb{G}_{p_{i-1}} \circ \dots \circ \mathbb{G}_{p_1}, v)$  is a strictly proper subset of  $\mathcal{V}$ . If  $i < n - 1$ , set  $k = i$ , in which case (28) is clearly true. Therefore suppose  $i = n - 1$ ; we will prove that this cannot be so. Assuming that it is,  $\alpha(\mathbb{G}_{p_{j-1}} \circ \dots \circ \mathbb{G}_{p_1}, v)$  must be a strictly proper subset of  $\mathcal{V}$  for  $j \in \{2, 3, \dots, n - 1\}$ ; by Lemma 5,  $\alpha(\mathbb{G}_{p_{j-1}} \circ \dots \circ \mathbb{G}_{p_1}, v)$  is also a strictly proper subset of  $\alpha(\mathbb{G}_{p_j} \circ \dots \circ \mathbb{G}_{p_1}, v)$  for  $j \in \{2, 3, \dots, n - 1\}$ . In view of this and (26), each containment in the ascending chain

$$\alpha(\mathbb{G}_{p_1}, v) \subset \alpha(\mathbb{G}_{p_2} \circ \mathbb{G}_{p_1}, v) \subset \dots \subset \alpha(\mathbb{G}_{p_{n-1}} \circ \dots \circ \mathbb{G}_{p_1}, v)$$

is strict. Since  $\alpha(\mathbb{G}_{p_1}, v)$  has at least two vertices in it, and there are  $n$  vertices in  $\mathcal{V}$ , (28) must hold with  $k = n - 1$ . Thus assertion 2 is true.

To prove assertion 1, suppose that  $m \geq (n - 1)^2$ . Since there are  $n$  vertices in  $\mathcal{V}$ , the sequence  $p_1, p_2, \dots, p_m$  must contain a subsequence  $q_1, q_2, \dots, q_{n-1}$  for which the graphs  $\mathbb{G}_{q_1}, \mathbb{G}_{q_2}, \dots, \mathbb{G}_{q_{n-1}}$  all have a common root. By assertion 2,  $\mathbb{G}_{q_{n-1}} \circ \dots \circ \mathbb{G}_{q_1}$  must be strongly rooted. But  $\mathcal{A}(\mathbb{G}_{q_{n-1}} \circ \dots \circ \mathbb{G}_{q_1}) \subset \mathcal{A}(\mathbb{G}_{p_m} \circ \mathbb{G}_{p_{m-1}} \circ \dots \circ \mathbb{G}_{p_1})$

because  $\mathbb{G}_{q_1}, \mathbb{G}_{q_2}, \dots, \mathbb{G}_{q_{n-1}}$  is a subsequence of  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_m}$  and all graphs in the sequence  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_m}$  have self-arcs. Therefore  $\mathbb{G}_{p_m} \circ \mathbb{G}_{p_{m-1}} \circ \dots \circ \mathbb{G}_{p_1}$  must be strongly rooted.  $\square$

Proposition 3 implies that every sufficiently long composition of graphs from a given subset  $\widehat{\mathcal{G}} \subset \mathcal{G}_{sa}$  will be strongly rooted if each graph in  $\widehat{\mathcal{G}}$  is rooted. The converse is also true. To understand why, suppose to the contrary that it is not. In this case there would have to be a graph  $\mathbb{G} \in \widehat{\mathcal{G}}$ , which is not rooted but for which  $\mathbb{G}^m$  is strongly rooted for  $m$  sufficiently large, where  $\mathbb{G}^m$  is the  $m$ -fold composition of  $\mathbb{G}$  with itself. Thus  $\alpha(\mathbb{G}^m, v) = \mathcal{V}$ , where  $v$  is a root of  $\mathbb{G}^m$ . But via repeated application of (27),  $\alpha(\mathbb{G}^m, v) = \alpha^m(\mathbb{G}, v)$ , where  $\alpha^m(\mathbb{G}, \cdot)$  is the  $m$ -fold composition of  $\alpha(\mathbb{G}, \cdot)$  with itself. Thus  $\alpha^m(\mathbb{G}, v) = \mathcal{V}$ . But this can occur only if  $\mathbb{G}$  is rooted at  $v$  because  $\alpha^m(\mathbb{G}, v)$  is the set of vertices reachable from  $v$  along paths of length  $m$ . Since this is a contradiction,  $\mathbb{G}$  must be rooted. We summarize.

**PROPOSITION 5.** *Every possible sufficiently long composition of graphs from a given subset  $\widehat{\mathcal{G}} \subset \mathcal{G}_{sa}$  is strongly rooted if and only if every graph in  $\widehat{\mathcal{G}}$  is rooted.*

**2.4.3. Sarymsakov graphs.** We now briefly discuss a class of graphs in  $\mathcal{G}$ , namely “Sarymsakov graphs,” whose corresponding stochastic matrices form products which are known to converge to rank one matrices [25] even though the graphs in question need not have self-arcs at all vertices. Sarymsakov graphs are defined as follows.

First, let us agree to say that a vertex  $v \in \mathcal{V}$  is a *neighbor of a subset*  $\mathcal{S} \subset \mathcal{V}$  in a graph  $\mathbb{G} \in \mathcal{G}$  if  $v$  is a neighbor of at least one vertex in  $\mathcal{S}$ . By a *Sarymsakov graph* we mean a graph  $\mathbb{G} \in \mathcal{G}$  with the property that for each pair of nonempty subsets  $\mathcal{S}_1$  and  $\mathcal{S}_2$  in  $\mathcal{V}$  which have no neighbors in common,  $\mathcal{S}_1 \cup \mathcal{S}_2$  contains a smaller number of vertices than does the set of neighbors of  $\mathcal{S}_1 \cup \mathcal{S}_2$ . Such seemingly obscure graphs are so named because they are the graphs of an important class of nonnegative matrices studied by Sarymsakov in [24]. In what follows we will prove that Sarymsakov graphs are in fact rooted graphs. We will also prove that the class of rooted graphs we are primarily interested in, namely those in  $\mathcal{G}_{sa}$ , are Sarymsakov graphs.

It is possible to characterize Sarymsakov graphs a little more concisely using the following concept. By the *neighbor function* of a graph  $\mathbb{G} \in \mathcal{G}$ , written  $\beta(\mathbb{G}, \cdot)$ , we mean the function  $\beta(\mathbb{G}, \cdot) : 2^{\mathcal{V}} \rightarrow 2^{\mathcal{V}}$  which assigns to each subset  $\mathcal{S} \subset \mathcal{V}$  the subset of vertices in  $\mathcal{V}$  which are neighbors of  $\mathcal{S}$  in  $\mathbb{G}$ . Thus in terms of  $\beta$ , a Sarymsakov graph is a graph  $\mathbb{G} \in \mathcal{G}$  with the property that for each pair of nonempty subsets  $\mathcal{S}_1$  and  $\mathcal{S}_2$  in  $\mathcal{V}$  which have no neighbors in common,  $\mathcal{S}_1 \cup \mathcal{S}_2$  contains fewer vertices than does the set  $\beta(\mathbb{G}, \mathcal{S}_1 \cup \mathcal{S}_2)$ . Note that if  $\mathbb{G} \in \mathcal{G}_{sa}$ , the requirement that  $\mathcal{S}_1 \cup \mathcal{S}_2$  contain fewer vertices than  $\beta(\mathbb{G}, \mathcal{S}_1 \cup \mathcal{S}_2)$  simplifies to the equivalent requirement that  $\mathcal{S}_1 \cup \mathcal{S}_2$  be a strictly proper subset of  $\beta(\mathbb{G}, \mathcal{S}_1 \cup \mathcal{S}_2)$ . This is because every vertex in  $\mathbb{G}$  is a neighbor of itself if  $\mathbb{G} \in \mathcal{G}_{sa}$ .

**PROPOSITION 6.**

1. *Each Sarymsakov graph in  $\mathcal{G}$  is rooted.*
2. *Each rooted graph in  $\mathcal{G}_{sa}$  is a Sarymsakov graph.*

It follows that if we restrict attention exclusively to graphs in  $\mathcal{G}_{sa}$ , then rooted graphs and Sarymsakov graphs are one and the same.

In what follows  $\beta^m(\mathbb{G}, \cdot)$  denotes the  $m$ -fold composition of  $\beta(\mathbb{G}, \cdot)$  with itself. The proof of Proposition 6 depends on the following ideas.

**LEMMA 6.** *Let  $\mathbb{G} \in \mathcal{G}$  be a Sarymsakov graph. Let  $\mathcal{S}$  be a nonempty subset of  $\mathcal{V}$  such that  $\beta(\mathbb{G}, \mathcal{S}) \subset \mathcal{S}$ . Let  $v$  be any vertex in  $\mathcal{V}$ . Then there exists a nonnegative integer  $m \leq n$  such that  $\beta^m(\mathbb{G}, v) \cap \mathcal{S}$  is nonempty.*

*Proof.* If  $v \in \mathcal{S}$ , set  $m = 0$ . Suppose next that  $v \notin \mathcal{S}$ . Set  $\mathcal{T} = \{v\} \cup \beta(\mathbb{G}, v) \cup \cdots \cup \beta^{n-1}(\mathbb{G}, v)$  and note that  $\beta^n(\mathbb{G}, v) \subset \mathcal{T}$  because  $\mathbb{G}$  has  $n$  vertices. Since  $\beta(\mathbb{G}, \mathcal{T}) = \beta(\mathbb{G}, v) \cup \beta^2(\mathbb{G}, v) \cup \cdots \cup \beta^n(\mathbb{G}, v)$ , it must be true that  $\beta(\mathbb{G}, \mathcal{T}) \subset \mathcal{T}$ . Therefore

$$(29) \quad \beta(\mathbb{G}, \mathcal{T} \cup \mathcal{S}) \subset \mathcal{T} \cup \mathcal{S}.$$

Suppose  $\beta(\mathbb{G}, \mathcal{T}) \cap \beta(\mathbb{G}, \mathcal{S})$  is empty. Then because  $\mathbb{G}$  is a Sarymsakov graph,  $\mathcal{T} \cup \mathcal{S}$  contains fewer vertices than  $\beta(\mathbb{G}, \mathcal{T} \cup \mathcal{S})$ . This contradicts (29), and so  $\beta(\mathbb{G}, \mathcal{T}) \cap \beta(\mathbb{G}, \mathcal{S})$  is not empty. In view of the fact that  $\beta(\mathbb{G}, \mathcal{T}) = \beta(\mathbb{G}, v) \cup \beta^2(\mathbb{G}, v) \cup \cdots \cup \beta^n(\mathbb{G}, v)$ , it must therefore be true for some positive integer  $m \leq n$  that  $\beta^m(\mathbb{G}, v) \cap \beta(\mathbb{G}, \mathcal{S})$  is nonempty. But by assumption  $\beta(\mathbb{G}, \mathcal{S}) \subset \mathcal{S}$ , and so  $\beta^m(\mathbb{G}, v) \cap \mathcal{S}$  is nonempty.  $\square$

LEMMA 7. *Let  $\mathbb{G} \in \mathcal{G}$  be rooted at  $r$ . Each nonempty subset  $\mathcal{S} \subset \mathcal{V}$  not containing  $r$  is a strictly proper subset of  $\mathcal{S} \cup \beta(\mathbb{G}, \mathcal{S})$ .*

*Proof of Lemma 7.* Let  $\mathcal{S} \subset \mathcal{V}$  be nonempty and not containing  $r$ . Pick  $v \in \mathcal{S}$ . Since  $\mathbb{G}$  is rooted at  $r$ , there must be a path in  $\mathbb{G}$  from  $r$  to  $v$ . Since  $r \notin \mathcal{S}$  there must be a vertex  $x \in \mathcal{S}$  which has a neighbor which is not in  $\mathcal{S}$ . Thus there is a vertex  $y \in \beta(\mathbb{G}, \mathcal{S})$  which is not in  $\mathcal{S}$ . This implies that  $\mathcal{S}$  is a strictly proper subset of  $\mathcal{S} \cup \beta(\mathbb{G}, \mathcal{S})$ .  $\square$

By a *maximal rooted subgraph* of  $\mathbb{G}$  we mean a subgraph  $\mathbb{G}^*$  of  $\mathbb{G}$  which is rooted and which is not contained in any rooted subgraph of  $\mathbb{G}$  other than itself. Graphs in  $\mathcal{G}$  may have one or more maximal rooted subgraphs. Clearly  $\mathbb{G}^* = \mathbb{G}$  just in case  $\mathbb{G}$  is rooted. Note that if  $\widehat{\mathcal{R}}$  is the set of all roots of a maximal rooted subgraph  $\widehat{\mathbb{G}}$ , then  $\beta(\mathbb{G}, \widehat{\mathcal{R}}) \subset \widehat{\mathcal{R}}$ . For if this were not so, then it would be possible to find a vertex  $x \in \beta(\mathbb{G}, \widehat{\mathcal{R}})$  which is not in  $\widehat{\mathcal{R}}$ . This would imply the existence of a path from  $x$  to some root  $\widehat{v} \in \widehat{\mathcal{R}}$ ; consequently the graph induced by the set of vertices along this path together with  $\widehat{\mathcal{R}}$  would be rooted at  $x \notin \widehat{\mathcal{R}}$  and would contain  $\widehat{\mathbb{G}}$  as a strictly proper subgraph. But this contradicts the hypothesis that  $\widehat{\mathbb{G}}$  is maximal. Therefore  $\beta(\mathbb{G}, \widehat{\mathcal{R}}) \subset \widehat{\mathcal{R}}$ . Now suppose that  $\widehat{\mathbb{G}}$  is any rooted subgraph in  $\mathcal{G}$ . Suppose that  $\widehat{\mathbb{G}}$ 's set of roots  $\widehat{\mathcal{R}}$  satisfies  $\beta(\mathbb{G}, \widehat{\mathcal{R}}) \subset \widehat{\mathcal{R}}$ . We claim that  $\widehat{\mathbb{G}}$  must then be maximal. For if this were not so, there would have to be a rooted graph  $\mathbb{G}^*$  containing  $\widehat{\mathbb{G}}$  as a strictly proper subset. This, in turn, would imply the existence of a path from a root  $x^*$  of  $\mathbb{G}^*$  to a root  $v$  of  $\widehat{\mathbb{G}}$ ; consequently  $x^* \in \beta^i(\mathbb{G}, \widehat{\mathcal{R}})$  for some  $i \geq 1$ . But this is impossible because  $\widehat{\mathcal{R}}$  is  $\beta(\mathbb{G}, \cdot)$  invariant. Thus  $\widehat{\mathbb{G}}$  is maximal. We summarize.

LEMMA 8. *A rooted subgraph of a graph  $\mathbb{G}$  generated by any vertex  $v \in \mathcal{V}$  is maximal if and only if its set of roots is  $\beta(\mathbb{G}, \cdot)$ -invariant.*

*Proof of Proposition 6.* Write  $\beta(\cdot)$  for  $\beta(\mathbb{G}, \cdot)$ . To prove assertion 1, pick  $\mathbb{G} \in \mathcal{G}$ . Let  $\mathbb{G}^*$  be any maximal rooted subgraph of  $\mathbb{G}$  and write  $\mathcal{R}$  for its root set; in view of Lemma 8,  $\beta(\mathcal{R}) \subset \mathcal{R}$ . Pick any  $v \in \mathcal{V}$ . Then by Lemma 6, for some positive integer  $m \leq n$ ,  $\beta^m(v) \cap \mathcal{R}$  is nonempty. Pick  $z \in \beta^m(v) \cap \mathcal{R}$ . Then there is a path from  $z$  to  $v$  and  $z$  is a root of  $\mathbb{G}^*$ . But  $\mathbb{G}^*$  is maximal, and so  $v$  must be a vertex of  $\mathbb{G}^*$ . Therefore every vertex of  $\mathbb{G}$  is a vertex of  $\mathbb{G}^*$ , which implies that  $\mathbb{G}$  is rooted.

To prove assertion 2, let  $\mathbb{G} \in \mathcal{G}_{sa}$  be rooted at  $r$ . Pick any two nonempty subsets  $\mathcal{S}_1, \mathcal{S}_2$  of  $\mathcal{V}$  which have no neighbors in common. If  $r \notin \mathcal{S}_1 \cup \mathcal{S}_2$ , then  $\mathcal{S}_1 \cup \mathcal{S}_2$  must be a strictly proper subset of  $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \beta(\mathcal{S}_1 \cup \mathcal{S}_2)$  because of Lemma 7.

Suppose next that  $r \in \mathcal{S}_1 \cup \mathcal{S}_2$ . Since  $\mathbb{G} \in \mathcal{G}_{sa}$ ,  $\mathcal{S}_i \subset \beta(\mathcal{S}_i)$ ,  $i \in \{1, 2\}$ . Thus  $\mathcal{S}_1$  and  $\mathcal{S}_2$  must be disjoint because  $\beta(\mathcal{S}_1)$  and  $\beta(\mathcal{S}_2)$  are. Therefore  $r$  must be in either  $\mathcal{S}_1$  or  $\mathcal{S}_2$  but not both. Suppose that  $r \notin \mathcal{S}_1$ . Then  $\mathcal{S}_1$  must be a strictly proper subset of  $\beta(\mathcal{S}_1)$  because of Lemma 7. Since  $\beta(\mathcal{S}_1)$  and  $\beta(\mathcal{S}_2)$  are disjoint,  $\mathcal{S}_1 \cup \mathcal{S}_2$  must be a strictly proper subset of  $\beta(\mathcal{S}_1 \cup \mathcal{S}_2)$ . By the same reasoning,  $\mathcal{S}_1 \cup \mathcal{S}_2$  must be a strictly proper subset of  $\beta(\mathcal{S}_1 \cup \mathcal{S}_2)$  if  $r \notin \mathcal{S}_2$ . Thus in conclusion  $\mathcal{S}_1 \cup \mathcal{S}_2$  must be a strictly

proper subset of  $\beta(\mathcal{S}_1 \cup \mathcal{S}_2)$  whether or not  $r$  is in  $\mathcal{S}_1 \cup \mathcal{S}_2$ . Since this conclusion holds for all such  $\mathcal{S}_1$  and  $\mathcal{S}_2$  and  $\mathbb{G} \in \mathcal{G}_{sa}$ ,  $\mathbb{G}$  must be a Sarymsakov graph.  $\square$

**2.5. Neighbor-shared graphs.** There is a different assumption which one can make about a sequence of graphs from  $\mathcal{G}$  which also ensures that the sequence's composition is strongly rooted. For this we need the concept of a “neighbor-shared graph.” Let us call  $\mathbb{G} \in \mathcal{G}$  *neighbor-shared* if each set of two distinct vertices shares a common neighbor. Suppose that  $\mathbb{G}$  is neighbor-shared. Then both vertices in any given pair of vertices are clearly reachable from a single vertex along directed paths. Suppose that for some integer  $k \in \{2, 3, \dots, n-1\}$ , each subset of  $k$  vertices  $\{v_1, v_2, \dots, v_k\}$  has the property that every vertex in  $\{v_1, v_2, \dots, v_k\}$  is reachable from a single vertex. Let  $\{v_1, v_2, \dots, v_k\}$  be any such set and  $v$  be a vertex from which all  $k$  vertices in the set can be reached. Let  $w$  be any vertex not in  $\{v_1, v_2, \dots, v_k\}$ . Since  $v$  and  $w$  can be reached from a common vertex  $y$ , every vertex in  $\{v_1, v_2, \dots, v_k, w\}$  can be reached from  $y$ . This proves that each subset of  $k+1$  vertices has the property that every vertex in the subset is reachable from a single vertex. By induction we can therefore conclude that every vertex in  $\mathbb{G}$  is reachable from a single vertex. We have proved the following proposition.

PROPOSITION 7. *Each neighbor-shared graph in  $\mathcal{G}$  is rooted.*

It is worth noting that although neighbor-shared graphs are rooted, the converse is not necessarily true. The reader may wish to construct a three-vertex example which illustrates this. Although rooted graphs in  $\mathcal{G}_{sa}$  need not be neighbor-shared, it turns out that the composition of any  $n-1$  rooted graphs in  $\mathcal{G}_{sa}$  is.

PROPOSITION 8. *The composition of any set of  $m \geq n-1$  rooted graphs in  $\mathcal{G}_{sa}$  is neighbor-shared.*

This result is equivalent to Theorem 5.1 of [31], which was independently derived.

To prove Proposition 8 we need some more ideas. By the *reverse graph* of  $\mathbb{G} \in \mathcal{G}$ , written  $\mathbb{G}'$ , we mean the graph in  $\mathcal{G}$  which results when the directions of all arcs in  $\mathbb{G}$  are reversed. It is clear that  $\mathcal{G}_{sa}$  is closed under the reverse operation and that if  $A$  is the adjacency matrix of  $\mathbb{G}$ , then  $A'$  is the adjacency matrix of  $\mathbb{G}'$ . It is also clear that  $(\mathbb{G}_p \circ \mathbb{G}_q)' = \mathbb{G}'_q \circ \mathbb{G}'_p$ ,  $p, q \in \mathcal{P}$ , and that

$$(30) \quad \alpha(\mathbb{G}', \mathcal{S}) = \beta(\mathbb{G}, \mathcal{S}), \quad \mathcal{S} \in 2^{\mathcal{V}}.$$

LEMMA 9. *For all  $\mathbb{G}_p, \mathbb{G}_q \in \mathcal{G}$  and any nonempty subset  $\mathcal{S} \subset \mathcal{V}$ ,*

$$(31) \quad \beta(\mathbb{G}_q, \beta(\mathbb{G}_p, \mathcal{S})) = \beta(\mathbb{G}_p \circ \mathbb{G}_q, \mathcal{S}).$$

*Proof of Lemma 9.* In view of (27),  $\alpha(\mathbb{G}'_p, \alpha(\mathbb{G}'_q, \mathcal{S})) = \alpha(\mathbb{G}'_p \circ \mathbb{G}'_q, \mathcal{S})$ . But  $\mathbb{G}'_p \circ \mathbb{G}'_q = (\mathbb{G}_q \circ \mathbb{G}_p)'$ , and so  $\alpha(\mathbb{G}'_p, \alpha(\mathbb{G}'_q, \mathcal{S})) = \alpha((\mathbb{G}_q \circ \mathbb{G}_p)', \mathcal{S})$ . Therefore  $\beta(\mathbb{G}_p, \beta(\mathbb{G}_q, \mathcal{S})) = \beta(\mathbb{G}_q \circ \mathbb{G}_p, \mathcal{S})$  because of (30).  $\square$

LEMMA 10. *Let  $\mathbb{G}_p$  and  $\mathbb{G}_q$  be rooted graphs in  $\mathcal{G}_{sa}$ . If  $u$  and  $v$  are distinct vertices in  $\mathcal{V}$  for which*

$$(32) \quad \beta(\mathbb{G}_q, \{u, v\}) = \beta(\mathbb{G}_q \circ \mathbb{G}_p, \{u, v\}),$$

*then  $u$  and  $v$  have a common neighbor in  $\mathbb{G}_q \circ \mathbb{G}_p$ .*

*Proof.*  $\beta(\mathbb{G}_q, u)$  and  $\beta(\mathbb{G}_q, v)$  are nonempty because  $u$  and  $v$  are neighbors of themselves. Suppose  $u$  and  $v$  do not have a common neighbor in  $\mathbb{G}_q \circ \mathbb{G}_p$ . Then  $\beta(\mathbb{G}_q \circ \mathbb{G}_p, u)$  and  $\beta(\mathbb{G}_q \circ \mathbb{G}_p, v)$  are disjoint. But  $\beta(\mathbb{G}_q \circ \mathbb{G}_p, u) = \beta(\mathbb{G}_p, \beta(\mathbb{G}_q, u))$  and  $\beta(\mathbb{G}_q \circ \mathbb{G}_p, v) = \beta(\mathbb{G}_p, \beta(\mathbb{G}_q, v))$  because of (31). Therefore  $\beta(\mathbb{G}_p, \beta(\mathbb{G}_q, u))$  and

$\beta(\mathbb{G}_p, \beta(\mathbb{G}_q, v))$  are disjoint. But  $\mathbb{G}_p$  is rooted and thus a Sarymsakov graph because of Proposition 6. Thus  $\beta(\mathbb{G}_q, \{u, v\})$  is a strictly proper subset of  $\beta(\mathbb{G}_q, \{u, v\}) \cup \beta(\mathbb{G}_p, \beta(\mathbb{G}_q, \{u, v\}))$ . But  $\beta(\mathbb{G}_q, \{u, v\}) \subset \beta(\mathbb{G}_p, \beta(\mathbb{G}_q, \{u, v\}))$  because all vertices in  $\mathbb{G}_q$  are neighbors of themselves and  $\beta(\mathbb{G}_p, \beta(\mathbb{G}_q, \{u, v\})) = \beta(\mathbb{G}_q \circ \mathbb{G}_p, \{u, v\})$  because of (31). Therefore  $\beta(\mathbb{G}_q, \{u, v\})$  is a strictly proper subset of  $\beta(\mathbb{G}_q \circ \mathbb{G}_p, \{u, v\})$ . This contradicts (32), and so  $u$  and  $v$  have a common neighbor in  $\mathbb{G}_q \circ \mathbb{G}_p$ .  $\square$

*Proof of Proposition 8.* Let  $u$  and  $v$  be distinct vertices in  $\mathcal{V}$ . Let  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_{n-1}}$  be a sequence of rooted graphs in  $\mathcal{G}_{sa}$ . Since  $\mathcal{A}(\mathbb{G}_{p_{n-1}} \circ \dots \circ \mathbb{G}_{p_{n-i}}) \subset \mathcal{A}(\mathbb{G}_{p_{n-1}} \circ \dots \circ \mathbb{G}_{p_{n-(i+1)}})$  for  $i \in \{1, 2, \dots, n-2\}$ , it must be true that the  $\mathbb{G}_p$  yield the ascending chain

$$\beta(\mathbb{G}_{n-1}, \{u, v\}) \subset \beta(\mathbb{G}_{p_{n-1}} \circ \mathbb{G}_{p_{n-2}}, \{u, v\}) \subset \dots \subset \beta(\mathbb{G}_{p_{n-1}} \circ \dots \circ \mathbb{G}_{p_2} \circ \mathbb{G}_{p_1}, \{u, v\}).$$

Because there are  $n$  vertices in  $\mathcal{V}$ , this chain must converge for some  $i < n-1$ , which means that

$$\beta(\mathbb{G}_{p_{n-1}} \circ \dots \circ \mathbb{G}_{p_{n-i}}, \{u, v\}) = \beta(\mathbb{G}_{p_{n-1}} \circ \dots \circ \mathbb{G}_{p_{n-i}} \circ \mathbb{G}_{p_{n-(i+1)}}, \{u, v\}).$$

This and Lemma 10 imply that  $u$  and  $v$  have a common neighbor in  $\mathbb{G}_{p_{n-1}} \circ \dots \circ \mathbb{G}_{p_{n-i}}$  and thus in  $\mathbb{G}_{p_{n-1}} \circ \dots \circ \mathbb{G}_{p_2} \circ \mathbb{G}_{p_1}$ . Since this is true for all distinct  $u$  and  $v$ ,  $\mathbb{G}_{p_{n-1}} \circ \dots \circ \mathbb{G}_{p_2} \circ \mathbb{G}_{p_1}$  is a neighbor-shared graph.  $\square$

If we restrict attention to those rooted graphs in  $\mathcal{G}_{sa}$  which are strongly connected, we can obtain a neighbor-shared graph by composing a smaller number of rooted graphs than the one claimed in Proposition 8.

**PROPOSITION 9.** *Let  $q$  be the integer quotient of  $n$  divided by 2. The composition of any set of  $m \geq q$  strongly connected graphs in  $\mathcal{G}_{sa}$  is neighbor-shared.*

*Proof of Proposition 9.* Let  $k < n$  be a positive integer and let  $v$  be any vertex in  $\mathcal{V}$ . Let  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_k}$  be a sequence of strongly connected graphs in  $\mathcal{G}_{sa}$ . Since each vertex of a strongly connected graph must be a root,  $v$  must be a root of each  $\mathbb{G}_{p_i}$ . Note that the  $\mathbb{G}_{p_i}$  yield the ascending chain

$$\{v\} \subset \beta(\mathbb{G}_{p_k}, \{v\}) \subset \beta(\mathbb{G}_{p_k} \circ \mathbb{G}_{p_{k-1}}, \{v\}) \subset \dots \subset \beta(\mathbb{G}_{p_k} \circ \dots \circ \mathbb{G}_{p_2} \circ \mathbb{G}_{p_1}, \{v\})$$

because  $\mathcal{A}(\mathbb{G}_{p_k} \circ \dots \circ \mathbb{G}_{p_{k-(i-1)}}) \subset \mathcal{A}(\mathbb{G}_{p_k} \circ \dots \circ \mathbb{G}_{p_{k-i}})$  for  $i \in \{1, 2, \dots, k-1\}$ . Moreover, since  $k < n$  and  $v$  is a root of each  $\mathbb{G}_{p_k} \circ \dots \circ \mathbb{G}_{p_{k-(i-1)}}$ ,  $i \in \{1, 2, \dots, k\}$ , it must be true for each such  $i$  that  $\beta(\mathbb{G}_{p_k} \circ \dots \circ \mathbb{G}_{p_{k-(i-1)}}, v)$  contains at least  $i+1$  vertices. In particular  $\beta(\mathbb{G}_{p_k} \circ \dots \circ \mathbb{G}_{p_1}, v)$  contains at least  $k+1$  vertices.

Set  $k = q$  and let  $v_1$  and  $v_2$  be any pair of distinct vertices in  $\mathcal{V}$ . Then there must be at least  $q+1$  vertices in  $\beta(\mathbb{G}_{p_q} \circ \dots \circ \mathbb{G}_{p_2} \circ \mathbb{G}_{p_1}, \{v_1\})$  and  $q+1$  vertices in  $\beta(\mathbb{G}_{p_q} \circ \dots \circ \mathbb{G}_{p_2} \circ \mathbb{G}_{p_1}, \{v_2\})$ . But  $2(q+1) > n$  because of the definition of  $q$ , and so  $\beta(\mathbb{G}_{p_q} \circ \dots \circ \mathbb{G}_{p_2} \circ \mathbb{G}_{p_1}, \{v_1\})$  and  $\beta(\mathbb{G}_{p_q} \circ \dots \circ \mathbb{G}_{p_2} \circ \mathbb{G}_{p_1}, \{v_2\})$  must have at least one vertex in common. Since this is true for each pair of distinct vertices  $v_1, v_2 \in \mathcal{V}$ ,  $\mathbb{G}_{p_q} \circ \dots \circ \mathbb{G}_2 \circ \mathbb{G}_1$  must be neighbor-shared.  $\square$

Lemma 7 and Proposition 3 imply that any composition of  $(n-1)^2$  neighbor-shared graphs in  $\mathcal{G}_{sa}$  is strongly rooted. The following proposition asserts that the composition need only consist of  $(n-1)$  neighbor-shared graphs and, moreover, that the graphs need only be in  $\mathcal{G}$  and not necessarily in  $\mathcal{G}_{sa}$ .

**PROPOSITION 10.** *The composition of any set of  $m \geq n-1$  neighbor-shared graphs in  $\mathcal{G}$  is strongly rooted.*

Note that Propositions 8 and 10 imply the first assertion of Proposition 3.

To prove Proposition 10 we need a few more ideas. For any integer  $1 < k \leq n$ , we say that a graph  $\mathbb{G} \in \mathcal{G}$  is  $k$  neighbor-shared if each set of  $k$  distinct vertices shares

a common neighbor. Thus a neighbor-shared graph and a 2 neighbor-shared graph are one and the same. Clearly an  $n$  neighbor-shared graph is strongly rooted at the common neighbor of all  $n$  vertices.

LEMMA 11. *If  $\mathbb{G}_p \in \mathcal{G}$  is a neighbor-shared graph and  $\mathbb{G}_q \in \mathcal{G}$  is a  $k$  neighbor-shared graph with  $k < n$ , then  $\mathbb{G}_q \circ \mathbb{G}_p$  is a  $(k + 1)$  neighbor-shared graph.*

*Proof.* Let  $v_1, v_2, \dots, v_{k+1}$  be any distinct vertices in  $\mathcal{V}$ . Since  $\mathbb{G}_q$  is a  $k$  neighbor-shared graph, the vertices  $v_1, v_2, \dots, v_k$  share a common neighbor  $u_1$  in  $\mathbb{G}_q$  and the vertices  $v_2, v_3, \dots, v_{k+1}$  share a common neighbor  $u_2$  in  $\mathbb{G}_q$  as well. Moreover, since  $\mathbb{G}_p$  is a neighbor-shared graph,  $u_1$  and  $u_2$  share a common neighbor  $w$  in  $\mathbb{G}_p$ . It follows from the definition of composition that  $v_1, v_2, \dots, v_k$  have  $w$  as a neighbor in  $\mathbb{G}_q \circ \mathbb{G}_p$  as do  $v_2, v_3, \dots, v_{k+1}$ . Therefore  $v_1, v_2, \dots, v_{k+1}$  have  $w$  as a neighbor in  $\mathbb{G}_q \circ \mathbb{G}_p$ . Since this must be true for any set of  $k + 1$  vertices in  $\mathbb{G}_q \circ \mathbb{G}_p$ ,  $\mathbb{G}_q \circ \mathbb{G}_p$  must be a  $(k + 1)$  neighbor-shared graph as claimed.  $\square$

*Proof of Proposition 10.* The preceding lemma implies that the composition of any 2 neighbor-shared graphs is 3 neighbor-shared. From this and induction it follows that for  $m < n$ , the composition of  $m$  neighbor-shared graphs is  $(m + 1)$  neighbor-shared. Thus the composition of  $(n - 1)$  neighbor-shared graphs is  $n$  neighbor-shared and consequently strongly rooted.  $\square$

**2.6. Convergence.** We are now in a position to significantly relax the conditions under which the conclusion of Theorem 1 holds. Towards this end, recall that each flocking matrix  $F$  is row stochastic. Moreover, because each vertex of each  $F$ 's graph  $\gamma(F)$  has a self-arc, the  $F$  have the additional property that their diagonal elements are all nonzero. Let  $\mathcal{S}$  denote the set of all  $n \times n$  row stochastic matrices whose diagonal elements are all positive.  $\mathcal{S}$  is closed under multiplication because the class of all  $n \times n$  stochastic matrices is closed under multiplication and because the class of  $n \times n$  nonnegative matrices with positive diagonals is also.

THEOREM 2. *Let  $\theta(0)$  be fixed. For any trajectory of the system (3) along which each graph in the sequence of neighbor graphs  $\mathbb{N}(0), \mathbb{N}(1), \dots$  is rooted, there is a constant steady state heading  $\theta_{ss}$  for which*

$$(33) \quad \lim_{t \rightarrow \infty} \theta(t) = \theta_{ss} \mathbf{1},$$

where the limit is approached exponentially fast.

The theorem says that a unique heading is achieved asymptotically along any trajectory on which all neighbor graphs are rooted. It is possible to deduce an explicit convergence rate for the situation addressed by this theorem [8, 7]. The theorem's proof relies on the following generalization of Proposition 2. The proposition exploits the fact that any composition of sufficiently many rooted graphs in  $\mathcal{G}_{sa}$  is strongly rooted (cf. Proposition 3).

PROPOSITION 11. *Let  $\mathcal{S}_r$  be any closed set of  $n \times n$  stochastic matrices with rooted graphs in  $\mathcal{G}_{sa}$ . There exists an integer  $m$  such that the graph of the product of every set of  $m$  matrices from  $\mathcal{S}_r$  is strongly rooted. Let  $m$  be any such integer and write  $\mathcal{S}_r^m$  for the set of all such matrix products. Then as  $j \rightarrow \infty$ , any product  $S_j \cdots S_1$  of matrices from  $\mathcal{S}_r$  converges exponentially fast to  $\mathbf{1}[\cdots S_j \cdots S_1]$  at a rate no slower than*

$$\lambda = \left( \max_{S \in \mathcal{S}_r^m} |||S||| \right)^{\frac{1}{m}},$$

where  $\lambda < 1$ .

*Proof of Proposition 11.* By assumption, each graph  $\gamma(S)$ ,  $S \in \mathcal{S}_r$ , is in  $\mathcal{G}_{sa}$  and is rooted. In view of Proposition 3,  $\gamma(S_q) \circ \cdots \circ \gamma(S_1)$  is strongly rooted for every list of  $q$  matrices  $\{S_1, S_2, \dots, S_q\}$  from  $\mathcal{S}_r$ , provided  $q \geq (n-1)^2$ . But  $\gamma(S_q) \circ \cdots \circ \gamma(S_1) = \gamma(S_q \cdots S_1)$  because of Lemma 3. Therefore  $\gamma(S_q \cdots S_1)$  is strongly rooted for all products  $S_q \cdots S_1$ , where each  $S_i \in \mathcal{S}_r$ . Thus  $m$  could be taken as  $q$ , which establishes the existence of such an integer.

Now any product  $S_j \cdots S_1$  of matrices in  $\mathcal{S}_r$  can be written as  $S_j \cdots S_1 = \bar{S}(j) \bar{S}_k \cdots \bar{S}_1$ , where  $\bar{S}_i = S_{im} \cdots S_{(i-1)m+1}$ ,  $1 \leq i \leq k$ , is a product in  $\mathcal{S}_r^m$ ,  $\bar{S}(j) = S_j \cdots S_{(km+1)}$ , and  $k$  is the integer quotient of  $j$  divided by  $m$ . In view of Proposition 2,  $\bar{S}_k \cdots \bar{S}_1$  must converge to  $\mathbf{1}[\cdots \bar{S}_k \cdots \bar{S}_1]$  exponentially fast as  $k \rightarrow \infty$  at a rate no slower than  $\bar{\lambda}$ , where

$$\bar{\lambda} = \max_{\bar{S} \in \mathcal{S}_r^m} \|\bar{S}\|.$$

But  $\bar{S}(j)$  is a product of at most  $m$  stochastic matrices, and so it is a bounded function of  $j$ . It follows that the product  $S_j S_{j-1} \cdots S_1$  must converge to  $\mathbf{1}[\cdots S_j \cdots S_1]$  exponentially fast at a rate no slower than  $\lambda = \bar{\lambda}^{\frac{1}{m}}$ .  $\square$

The proof of Proposition 11 can also be applied to any closed subset  $\mathcal{S}_{ns} \subset \mathcal{S}$  of stochastic matrices with neighbor-shared graphs. In this case, one would define  $m = n-1$  because of Proposition 10. Similarly, the proof also applies to any closed subset of stochastic matrices whose graphs share a common root; in this case one would define  $m = n-1$  because of the first assertion of Proposition 3.

*Proof of Theorem 2.* Let  $\mathcal{F}_r$  denote the set of flocking matrices with rooted graphs. Since  $\mathcal{G}_{sa}$  is a finite set, so is the set of rooted graphs in  $\mathcal{G}_{sa}$ . Therefore  $\mathcal{F}_r$  is closed. By assumption,  $F(t) \in \mathcal{F}_r$ ,  $t \geq 0$ . In view of Proposition 11, the product  $F(t) \cdots F(0)$  converges exponentially fast to  $\mathbf{1}[\cdots F(t) \cdots F(0)]$  at a rate no slower than

$$\lambda = \left( \max_{S \in \mathcal{F}_r^m} \|S\| \right)^{\frac{1}{m}},$$

where  $m = (n-1)^2$  and  $\mathcal{F}_r^m$  is the finite set of all  $m$ -term flocking matrix products of the form  $F_m \cdots F_1$  with each  $F_i \in \mathcal{F}_r$ . But it is clear from (3) that  $\theta(t) = F(t-1) \cdots F(1)F(0)\theta(0)$ ,  $t \geq 1$ . Therefore (33) holds with  $\theta_{ss} = \mathbf{1}[\cdots F(t) \cdots F(0)]\theta(0)$  and the convergence is exponential.  $\square$

The proof of Theorem 2 also applies to the case when all of the  $N(t)$ ,  $t \geq 0$ , are neighbor-shared. In this case, one would define  $m = n-1$  because of Proposition 10. By similar reasoning, the proof also applies to the case when all of the  $N(t)$ ,  $t \geq 0$ , shared a common root; one would also define  $m = n-1$  for this case because of the first assertion of Proposition 3.

**2.7. Jointly rooted sets of graphs.** It is possible to relax further still the conditions under which the conclusion of Theorem 1 holds. Towards this end, let us agree to say that a finite sequence of directed graphs  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_k}$  in  $\mathcal{G}$  is *jointly rooted* if the composition  $\mathbb{G}_{p_k} \circ \mathbb{G}_{p_{k-1}} \circ \cdots \circ \mathbb{G}_{p_1}$  is rooted.

Note that since the arc sets of any graphs  $\mathbb{G}_p, \mathbb{G}_q \in \mathcal{G}_{sa}$  are contained in the arc set of any composed graph  $\mathbb{G}_q \circ \mathbb{G}_p$ ,  $\mathbb{G} \in \mathcal{G}_{sa}$ , it must be true that if  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_k}$  is a jointly rooted sequence in  $\mathcal{G}_{sa}$ , then so is  $\mathbb{G}_q, \mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_k}, \mathbb{G}_p$ . In other words, a jointly rooted sequence of graphs in  $\mathcal{G}_{sa}$  remain jointly rooted if additional graphs from  $\mathcal{G}_{sa}$  are added to either end of the sequence.

There is an analogous concept for neighbor-shared graphs. We say that a finite sequence of directed graphs  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_k}$  from  $\mathcal{G}$  is *jointly neighbor-shared* if the



composition  $\mathbb{G}_{p_k} \circ \mathbb{G}_{p_{k-1}} \circ \cdots \circ \mathbb{G}_{p_1}$  is a neighbor-shared graph. Jointly neighbor-shared sequences of graphs from  $\mathcal{G}_{sa}$  remain jointly neighbor-shared if additional graphs from  $\mathcal{G}_{sa}$  are added to either end of the sequence. The reason for this is the same as for the case of jointly rooted sequences. Although the discussion which follows is just for the case of jointly rooted graphs, the material covered extends in the obvious way to the case of jointly neighbor-shared graphs.

In what follows we will say that an infinite sequence of graphs  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots$  in  $\mathcal{G}$  is *repeatedly jointly rooted* if there is a positive integer  $q$  for which each finite sequence  $\mathbb{G}_{p_{q(k-1)+1}}, \dots, \mathbb{G}_{p_{qk}}, k \geq 1$ , is jointly rooted. If such an integer exists, we sometimes say that  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots$  is repeatedly jointly rooted *by subsequences of length  $q$* . We are now in a position to generalize Proposition 11.

**PROPOSITION 12.** *Let  $\bar{S}$  be any closed set of stochastic matrices with graphs in  $\mathcal{G}_{sa}$ . Suppose that  $S_1, S_2, \dots$  is an infinite sequence of matrices from  $\bar{S}$  whose corresponding sequence of graphs  $\gamma(S_1), \gamma(S_2), \dots$  is repeatedly jointly rooted by subsequences of length  $q$ . Suppose that the set of all products of  $q$  matrices from  $\bar{S}$  with rooted graphs, written  $\bar{S}(q)$ , is closed. There exists an integer  $m$  such that the product of every set of  $m$  matrices from  $\bar{S}(q)$  is strongly rooted. Let  $m$  be any such integer and write  $(\bar{S}(q))^m$  for the set of all such matrix products. Then as  $j \rightarrow \infty$ , the product  $S_j \cdots S_1$  converges exponentially fast to  $\mathbf{1}[\cdots S_j \cdots S_1]$  at a rate no slower than*

$$\lambda = \left( \max_{S \in (\bar{S}(q))^m} |||S||| \right)^{\frac{1}{mq}},$$

where  $\lambda < 1$ .

It is worth pointing out that the assumption that  $\bar{S}(q)$  is closed is not necessarily implied by the assumption that  $\bar{S}$  is closed. For example, if  $\bar{S}$  is the set of all  $2 \times 2$  stochastic matrices whose diagonal elements are no smaller than some positive number  $\alpha < 1$ , then  $\bar{S}(2)$  cannot be closed even though  $\bar{S}$  is; this is because there are matrices in  $\bar{S}(2)$  which are arbitrarily close (in the induced infinity norm) to the  $2 \times 2$  identity which, in turn, is not in  $\bar{S}(2)$ . There are at least three different situations where  $\bar{S}(q)$  turns out to be closed. The first is when  $\bar{S}$  is a finite set, as is the case when  $\bar{S}$  is all  $n \times n$  flocking matrices; in this case it is obvious that for any  $q \geq 1$ ,  $\bar{S}(q)$  is closed because it is also a finite set.

The second situation arises when the simple average rule (1) is replaced by a convex combination rule as was done in [3]. In this case, the set  $\bar{S}$  turns out to be all  $n \times n$  stochastic matrices whose diagonal entries are nonzero and whose nonzero entries (on the diagonal or not) are all underbounded by a positive number  $\alpha < 1$ . In this case it is easy to see that for each graph  $\mathbb{G} \in \mathcal{G}_{sa}$ , the subset  $\bar{S}(\mathbb{G})$  of  $S \in \bar{S}$  for which  $\gamma(S) = \mathbb{G}$  is closed. Thus for any pair of graphs  $\mathbb{G}_1, \mathbb{G}_2 \in \mathcal{G}_{sa}$ , the subset of products  $S_2 S_1$  such that  $S_1 \in \bar{S}(\mathbb{G}_1)$  and  $S_2 \in \bar{S}(\mathbb{G}_2)$  is also closed. Since  $\bar{S}(2)$  is the union of a finite number of sets of products of this type, namely those for which the pairs  $(\mathbb{G}_1, \mathbb{G}_2)$  have rooted compositions  $\mathbb{G}_2 \circ \mathbb{G}_1$ , it must be that  $\bar{S}(2)$  is closed. Continuing this reasoning, one can conclude that for any integer  $q > 0$ ,  $\bar{S}(q)$  is closed as well.

The third situation in which  $\bar{S}(q)$  turns out to be compact is considerably more complicated and arises in connection with an asynchronous version of the flocking problem we have been studying. In this case, the graphs of the matrices in  $\bar{S}$  do not have self-arcs at all vertices. We refer the reader to [6] for details.

*Proof of Proposition 12.* Since  $\gamma(S_1), \gamma(S_2), \dots$  is repeatedly jointly rooted by subsequences of length  $q$ , for each  $k \geq 1$ , the subsequence  $\gamma(S_{q(k-1)+1}), \dots, \gamma(S_{qk})$

is jointly rooted. For  $k \geq 1$  define  $\bar{S}_k = S_{qk} \cdots S_{q(k-1)+1}$ . By Lemma 3,  $\gamma(S_{qk} \cdots S_{q(k-1)+1}) = \gamma(S_{qk}) \circ \cdots \circ \gamma(S_{q(k-1)+1})$ ,  $k \geq 1$ . Therefore  $\gamma(\bar{S}_k)$  is rooted for  $k \geq 1$ . Thus each such  $\bar{S}_k$  is in the closed set  $\bar{\mathcal{S}}(q)$ .

By Proposition 11, there exists an integer  $m$  such that the graph of the product of every set of  $m$  matrices from  $\bar{\mathcal{S}}(q)$  is strongly rooted. Moreover, since each  $\bar{S}_k \in \bar{\mathcal{S}}(q)$ , Proposition 11 also implies that  $k \rightarrow \infty$ , and the product  $\bar{S}_k \cdots \bar{S}_1$  converges exponentially fast to  $\mathbf{1}[\cdots \bar{S}_k \cdots \bar{S}_1]$  at a rate no slower than

$$\bar{\lambda} = \left( \max_{S \in (\bar{\mathcal{S}}(q))^m} |||S||| \right)^{\frac{1}{m}},$$

where  $\bar{\lambda} < 1$ .

Now the product  $S_j \cdots S_1$  can be written as

$$S_j \cdots S_1 = \hat{S}(j) \bar{S}_k \cdots \bar{S}_1,$$

where  $k$  is the integer quotient of  $j$  divided by  $mq$  and  $\hat{S}(j)$  is the identity if  $mq$  is a factor of  $j$  or  $\hat{S}(j) = S_j \cdots S_{(kmq+1)}$  if it is not. But  $\hat{S}(j)$  is a product of at most  $mq$  stochastic matrices, and so it is a bounded function of  $j$ . It follows that the product  $S_j S_{j-1} \cdots S_1$  must converge to  $\mathbf{1}[\cdots S_j \cdots S_1]$  exponentially fast at a rate no slower than  $\lambda = \bar{\lambda}^{\frac{1}{mq}}$ .  $\square$

We are now in a position to apply Proposition 12 to leaderless coordination.

**THEOREM 3.** *Let  $\theta(0)$  be fixed. For any trajectory of the system (3) along which each graph in the sequence of neighbor graphs  $\mathbb{N}(0), \mathbb{N}(1), \dots$  is repeatedly jointly rooted, there is a constant steady state heading  $\theta_{ss}$  for which*

$$(34) \quad \lim_{t \rightarrow \infty} \theta(t) = \theta_{ss} \mathbf{1},$$

where the limit is approached exponentially fast.

*Proof of Theorem 3.* By hypothesis, the sequence of graphs  $\gamma(F(0)), \gamma(F(1)), \dots$  is repeatedly jointly rooted. Thus there is an integer  $q$  for which the sequence is repeatedly jointly rooted by subsequences of length  $q$ . Since the set of  $n \times n$  flocking matrices  $\mathcal{F}$  is finite, so is the set of all products of  $q$  flocking matrices with rooted graphs, namely  $\mathcal{F}(q)$ . Therefore  $\mathcal{F}(q)$  is closed. Moreover, if  $m = (n-1)^2$ , every product of  $m$  matrices from  $\mathcal{F}(q)$  is strongly rooted. It follows from Proposition 12 that the product  $F(t) \cdots F(1)F(0)$  converges to  $\mathbf{1}[\cdots F(t) \cdots F(1)F(0)]$  exponentially fast as  $t \rightarrow \infty$  at a rate no slower than

$$\lambda = \left( \max_{S \in (\mathcal{F}(q))^m} |||S||| \right)^{\frac{1}{mq}},$$

where  $m = (n-1)^2$ ,  $\lambda < 1$ , and  $(\mathcal{F}(q))^m$  is the closed set of all products of  $m$  matrices from  $\mathcal{F}(q)$ . But it is clear from (3) that

$$\theta(t) = F(t-1) \cdots F(1)F(0)\theta(0), \quad t \geq 1.$$

Therefore (34) holds with  $\theta_{ss} = [\cdots F_{\sigma(t)} \cdots F_{\sigma(0)}]\theta(0)$  and the convergence is exponential.  $\square$

It is possible to compare Theorem 3 with similar results derived in [19, 22]. To do this it is necessary to introduce a few concepts. By the *union*  $\mathbb{G}_1 \cup \mathbb{G}_2$  of two directed

graphs  $\mathbb{G}_1$  and  $\mathbb{G}_2$  with the same vertex set  $\mathcal{V}$  we mean the graph whose vertex set is  $\mathcal{V}$  and whose arc set is the union of the arc sets of  $\mathbb{G}_1$  and  $\mathbb{G}_2$ . The definition extends in the obvious way to finite sets of directed graphs with the same vertex set. Let us agree to say that a finite set of graphs  $\{\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_k}\}$  with the same vertex set is *collectively rooted* if the *union* of the graphs in the set is a rooted graph. In parallel with the notion of repeatedly jointly rooted, we say that an infinite sequence of graphs  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots$  in  $\mathcal{G}_{sa}$  is *repeatedly collectively rooted* if there is a positive integer  $q$  for which each finite set  $\mathbb{G}_{p_{q(k-1)+1}}, \dots, \mathbb{G}_{p_{qk}}$ ,  $k \geq 1$  is collectively rooted. One of the main contributions of [19] is to prove that the conclusions of Theorem 3 hold if the theorem's hypothesis is replaced by the hypothesis that the sequence of graphs  $\mathbb{G}_{\sigma(0)}, \mathbb{G}_{\sigma(1)}, \dots$  is repeatedly collectively rooted. The two hypotheses prove to be equivalent. The reason this is so can be explained as follows.

Note first that because all graphs in  $\mathcal{G}_{sa}$  have self-arcs, each arc  $(i, j)$  in the union  $\mathbb{G}_2 \cup \mathbb{G}_1$  of two graphs  $\mathbb{G}_1, \mathbb{G}_2$  in  $\mathcal{G}_{sa}$  is an arc in the composition  $\mathbb{G}_2 \circ \mathbb{G}_1$ . While the converse is not true, the definition of composition does imply that for each arc  $(i, j)$  in the composition  $\mathbb{G}_2 \circ \mathbb{G}_1$  there is a *path* in the union  $\mathbb{G}_2 \cup \mathbb{G}_1$  of length at most two between  $i$  and  $j$ . More generally, simple induction proves that if  $(i, j)$  is an arc in the composition of  $q$  graphs from  $\mathcal{G}_{sa}$ , then the union of the same  $q$  graphs must contain a path of length at most  $q$  from  $i$  to  $j$ . These observations clearly imply that a sequence of  $q$  graphs  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_q}$  in  $\mathcal{G}_{sa}$  is jointly rooted if and only if the set of graphs  $\{\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots, \mathbb{G}_{p_q}\}$  is collectively rooted. It follows that a sequence of graphs in  $\mathcal{G}_{sa}$  is repeatedly jointly rooted if and only if the set of graphs in the sequence is collectively jointly rooted.

Although Theorem 3 and the main result of [19] are equivalent, the difference between results based on unions and results based on compositions begins to emerge, when one looks deeper into the convergence question, especially when issues of convergence rate are taken into consideration. For example, if  $\pi_u(m, n)$  were the number of  $m$  term sequences of graphs in  $\mathcal{G}_{sa}$  whose unions are strongly rooted, and  $\pi_c(m, n)$  were the number of  $m$ -term sequences of graphs in  $\mathcal{G}_{sa}$  whose compositions are strongly rooted, then it is easy to see that the ratio  $\rho(m, n) = \pi_c(m, n)/\pi_u(m, n)$  would always be greater than 1. In fact,  $\rho(2, 3) = 1.04$  and  $\rho(2, 4) = 1.96$ . Moreover, probabilistic experiments suggest that this ratio can be as large as 18,181 for  $m = 2$  and  $n = 50$ . One would expect  $\rho(m, n)$  to increase not only with increasing  $n$  but also with increasing  $m$ . One would also expect similar comparisons for neighbor-shared graphs rather than strongly rooted graphs. Interestingly, preliminary experimental results suggest that this is not the case, but more work needs to be done to understand why this is so. Like strongly rooted graphs, neighbor-shared graphs also play a key role in determining convergence rates [7].

**3. Symmetric neighbor relations.** It is natural to call a graph in  $\mathcal{G}$  *symmetric* if for each pair of vertices  $i$  and  $j$  for which  $j$  is a neighbor of  $i$ ,  $i$  is also a neighbor of  $j$ . Note that  $\mathbb{G}$  is symmetric if and only if its adjacency matrix is symmetric. It is worth noting that for symmetric graphs, the properties of rooted and rooted at  $v$  are both equivalent to the property that the graph is strongly connected. Within the class of symmetric graphs, neighbor-shared graphs and strongly rooted graphs are also strongly connected graphs, but in neither case is the converse true. It is possible to represent a symmetric directed graph  $\mathbb{G}$  with an undirected graph  $\mathbb{G}^s$  in which each self-arc is replaced with an undirected edge and each pair of directed arcs  $(i, j)$  and  $(j, i)$  for distinct vertices is replaced with an undirected edge between  $i$  and  $j$ . Notions of strongly rooted and neighbor-shared extend in the obvious way

to undirected graphs. An undirected graph is said to be *connected* if there is an undirected path between each pair of vertices. Thus a strongly connected, directed graph which is symmetric is in essence the same as a connected, undirected graph. Undirected graphs are applicable when the sensing radii  $r_i$  of all agents are the same. It was the symmetric version of the flocking problem which Vicsek addressed in [29] and which was analyzed in [17] using undirected graphs.

Let  $\mathcal{G}^s$  and  $\mathcal{G}_{sa}^s$  denote the subsets of symmetric graphs in  $\mathcal{G}$  and  $\mathcal{G}_{sa}$ , respectively. Simple examples show that neither  $\mathcal{G}^s$  nor  $\mathcal{G}_{sa}^s$  is closed under composition. In particular, composition of two symmetric directed graphs in  $\mathcal{G}$  or  $\mathcal{G}_{sa}$  is not typically symmetric. On the other hand, the union is. It is clear that both  $\mathcal{G}^s$  and  $\mathcal{G}_{sa}^s$  are closed under the union operation. It is worth emphasizing that union and composition are really quite different operations. For example, as we have already seen with Proposition 4, the composition of any  $n - 1$  strongly connected graphs, symmetric or not, is always complete. On the other hand, the union of  $n - 1$  strongly connected graphs is not necessarily complete. In terms of undirected graphs, it is simply not true that the union of  $n - 1$  undirected graphs with vertex set  $\mathcal{V}$  is complete, even if each graph in the union has self-loops at each vertex. As noted before, the root cause of the difference between union and composition stems from the fact that the union and composition of two graphs in  $\mathcal{G}$  have different arc sets—and in the case of graphs from  $\mathcal{G}_{sa}$ , the arc set of the union is always contained in the arc set of the composition but not conversely.

In [17] use is made of the notion of a “jointly connected set of graphs.” Specifically, a set of undirected graphs with vertex set  $\mathcal{V}$  is *jointly connected* if the union of the graphs in the collection is a connected graph. The notion of jointly connected also applies to directed graphs in which case the collection is jointly connected if the union is strongly connected. In what follows we will say that an infinite sequence of graphs  $\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \dots$  in  $\mathcal{G}_{sa}$  is *repeatedly jointly connected* if there is a positive integer  $m$  for which each finite sequence  $\mathbb{G}_{p_{m(k-1)+1}}, \dots, \mathbb{G}_{p_{mk}}, k \geq 1$ , is jointly connected. The main result of [17] is, in essence, a corollary to Theorem 3.

**COROLLARY 1.** *Let  $\theta(0)$  be fixed. For any trajectory of the system (3) along which each graph in the sequence of symmetric neighbor graphs  $\mathbb{N}(0), \mathbb{N}(1), \dots$  is repeatedly jointly connected, there is a constant steady state heading  $\theta_{ss}$  for which*

$$(35) \quad \lim_{t \rightarrow \infty} \theta(t) = \theta_{ss} \mathbf{1},$$

where the limit is approached exponentially fast.

**4. Concluding remarks.** The main goal of this paper has been to establish a number of basic properties of compositions of directed graphs which are useful in explaining how a consensus is achieved under various conditions in a dynamically changing environment. The paper brings together in one place a number of results scattered throughout the literature and at the same time presents new results concerned with compositions of graphs as well as graphical interpretations of several specially structured stochastic matrices appropriate to nonhomogeneous Markov chains.

In a sequel to this paper [7], we consider a modified version of the Vicsek consensus problem in which integer-valued delays occur in sensing the values of headings which are available to agents. In keeping with our thesis that such problems can be conveniently formulated and solved using graphs and graph operations, we analyze the sensing delay problem from mainly a graph-theoretic point of view using the tools developed in this paper. In [7] we also consider another modified version of the Vicsek

problem in which each agent independently updates its heading at times determined by its own clock. We do not assume that the groups' clocks are synchronized together or that the times any one agent updates its heading are evenly spaced. Using graph-theoretic concepts from this paper we show in [7] that for both versions of the problem considered, the conditions under which a consensus is achieved are essentially the same as in the synchronized, delay-free case addressed here.

A number of questions are suggested by this work. For example, it would be interesting to have a complete characterization of those rooted graphs which are of Sarymaskov type. It would also be of interest to have convergence results for more general versions of the asynchronous consensus problem in which heading transitions occur continuously. Extensions of these results to more realistic settings such as the one considered in [26] would also be useful.

## REFERENCES

- [1] D. ANGELI AND P. A. BLIMAN, *Extension of a result by Moreau on stability of leaderless multi-agent systems*, in Proceedings of the 2005 IEEE CDC, 2005, pp. 759–764.
- [2] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
- [3] V. D. BLONDEL, J. M. HENDRICHX, A. OLSHEVSKY, AND J. N. TSITSIKLIS, *Convergence in multiagent coordination, consensus, and flocking*, in Proceedings of the 2005 IEEE CDC, 2005, pp. 2996–3000.
- [4] M. CAO, A. S. MORSE, AND B. D. O. ANDERSON, *Coordination of an asynchronous multi-agent system via averaging*, in Proceedings of the 2005 IFAC Congress, 2005.
- [5] M. CAO, A. S. MORSE, AND B. D. O. ANDERSON, *Reaching a consensus in the face of measurement delays*, in Proceedings of the 2006 Symposium on Mathematical Theory of Networks and Systems, 2006.
- [6] M. CAO, A. S. MORSE, AND B. D. O. ANDERSON, *Agreeing asynchronously*, IEEE Trans. Automat. Control, to appear.
- [7] M. CAO, A. S. MORSE, AND B. D. O. ANDERSON, *Reaching a consensus in a dynamically changing environment: Convergence rates, measurement delays, and asynchronous events*, SIAM J. Control Optim., 47 (2008), pp. 601–623.
- [8] M. CAO, D. A. SPIELMAN, AND A. S. MORSE, *A lower bound on convergence of a distributed network consensus algorithm*, in Proceedings of the 2005 IEEE CDC, 2005, pp. 2356–2361.
- [9] S. CHATTERJEE AND E. SENETA, *Towards consensus: Some convergence theorems on repeated averaging*, J. Appl. Probability, 14 (1977), pp. 89–97.
- [10] M. H. DE GROOT, *Reaching a consensus*, J. Amer. Statist. Assoc., 69 (1974), pp. 118–121.
- [11] J. L. DOOB, *Markov processes—discrete parameter*, in Stochastic Processes, John Wiley and Sons, New York, 1953.
- [12] M. J. FISCHER, N. A. LYNCH, AND M. S. PATERSON, *Impossibility of distributed consensus with one faulty process*, J. Assoc. Comput. Mach., 32 (1985), pp. 347–382.
- [13] C. GODSIL AND G. ROYLE, *Algebraic Graph Theory*, Grad. Texts in Math. 207, Springer-Verlag, New York, 2001.
- [14] D. J. HARTFIEL, *Markov Set-Chains*, Springer-Verlag, Berlin, New York, 1998.
- [15] D. J. HARTFIEL, *Nonhomogeneous Matrix Products*, World Scientific, Singapore, 2002.
- [16] R. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.
- [17] A. JADBABAIE, J. LIN, AND A. S. MORSE, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.
- [18] C. F. MARTIN AND W. P. DAYAWANSA, *On the existence of a Lyapunov function for a family of switching systems*, in Proceedings of the 1996 IEEE CDC, 1996, pp. 1820–1823.
- [19] L. MOREAU, *Stability of multi-agent systems with time-dependent communication links*, IEEE Trans. Automat. Control, 50 (2005), pp. 169–182.
- [20] A. S. MORSE, *Logically switched dynamical systems*, in Nonlinear and Optimal Control Theory, Springer-Verlag, Berlin, 2008, pp. 1–84.
- [21] M. PEASE, R. SHOSTAK, AND L. LAMPORT, *Reaching agreement in the presence of faults*, J. Assoc. Comput. Mach., 27 (1980), pp. 228–234.
- [22] W. REN AND R. BEARD, *Consensus seeking in multiagent systems under dynamically changing interaction topologies*, IEEE Trans. Automat. Control, 50 (2005), pp. 655–661.

- [23] C. REYNOLDS, *Flocks, birds, and schools: A distributed behavioral model*, Comput. Graphics, 21 (1987), pp. 25–34.
- [24] T. A. SARYMSAKOV, *Inhomogeneous Markov chains*, Theor. Probability Appl., 6 (1961), pp. 178–185.
- [25] E. SENETA, *Non-negative Matrices and Markov Chains*, Springer-Verlag, New York, 1981.
- [26] H. TANNER, A. JADBABAIE, AND G. PAPPAS, *Flocking in fixed and switching networks*, IEEE Trans. Automat. Control, 52 (2007), pp. 863–868.
- [27] J. N. TSISIKLIS, *Problems in Decentralized Decision Making and Computation*, Ph.D thesis, MIT, Cambridge, MA, 1984.
- [28] J. N. TSISIKLIS, D. P. BERTSEKAS, AND M. ATHANS, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, 3 (1986), pp. 803–812.
- [29] T. VICSEK, A. CZIRÓK, E. BEN-JACOB, I. COHEN, AND O. SHOCHET, *Novel type of phase transition in a system of self-driven particles*, Phys. Rev. Lett., 75 (1995), pp. 1226–1229.
- [30] J. WOLFOWITZ, *Products of indecomposable, aperiodic, stochastic matrices*, Proc. Amer. Math. Soc., 14 (1963), pp. 733–737.
- [31] C. W. WU, *Synchronization and convergence of linear dynamics in random directed networks*, IEEE Trans. Automat. Control, 51 (2006), pp. 1207–1210.

## REACHING A CONSENSUS IN A DYNAMICALLY CHANGING ENVIRONMENT: CONVERGENCE RATES, MEASUREMENT DELAYS, AND ASYNCHRONOUS EVENTS\*

MING CAO<sup>†</sup>, A. STEPHEN MORSE<sup>‡</sup>, AND BRIAN D. O. ANDERSON<sup>‡</sup>

**Abstract.** This paper uses recently established properties of compositions of directed graphs together with results from the theory of nonhomogeneous Markov chains to derive worst case convergence rates for the headings of a group of mobile autonomous agents which arise in connection with the widely studied Vicsek consensus problem. The paper also uses graph-theoretic constructions to solve modified versions of the Vicsek problem in which there are measurement delays, asynchronous events, or a group leader. In all three cases the conditions under which consensus is achieved prove to be almost the same as the conditions under which consensus is achieved in the synchronous, delay-free, leaderless case.

**Key words.** cooperative control, graph theory, switched systems, convergence rates, delays, asynchronism

**AMS subject classifications.** 93C05, 05C50, 05C75, 15A51, 40A20, 68W15

**DOI.** 10.1137/060657029

**1. Introduction.** In a recent paper [6] the present authors defined the notion of “graph composition” and established a number of basic properties of compositions of directed graphs which are useful in explaining how a consensus might be reached by a group of mobile autonomous agents in a dynamically changing environment. The aim of this paper is to use the graph-theoretic findings of [6] to address several issues related to the well-known Vicsek consensus problem [20] which have either not been considered before or have been only partially resolved.

The paper begins with a brief review in section 2 of the basic leaderless consensus problem treated in [6, 14, 16]. Section 3 exploits the connection between “neighbor-shared” graphs and the elegant theory of “scrambling matrices” found in the literature on nonhomogeneous Markov chains [17, 9] to help in the derivation of worst case agent heading convergence rates for the leaderless version of the Vicsek problem. Section 4 addresses a modified version of the consensus problem in which integer-valued delays occur in the values of the headings which agents measure. In keeping with the overall theme of this paper, the effect of measurement delays is analyzed from a mainly graph-theoretic point of view. This enables us to significantly relax previously derived conditions [18, 19, 3] under which consensus can be achieved in the face of measurement delays. A comparison is made between the results of [3] and

---

\*Received by the editors April 11, 2006; accepted for publication (in revised form) August 16, 2007; published electronically February 6, 2008. A preliminary version of this work can be found in A. S. Morse, *Logically switched dynamical systems*, in *Nonlinear and Optimal Control Theory*, Springer-Verlag, Berlin, 2008, pp. 1–84.

<http://www.siam.org/journals/sicon/47-2/65702.html>

<sup>†</sup>Electrical Engineering, Yale University, P.O. Box 208267, New Haven, CT 06520 (m.cao@yale.edu, morse@sysc.eng.yale.edu). The research of these authors was supported by the U.S. Army Research Office, the U.S. National Science Foundation, and a gift from the Xerox Corporation.

<sup>‡</sup>Australian National University and National ICT Australia Ltd., Locked bag 8001, Canberra ACT 2601, Australia (brian.anderson@nicta.com.au). The research of this author was supported by National ICT Australia, which is funded by the Australian Government’s Department of Communications, Information Technology, and the Arts, and the Australian Research Council through the Backing Australia’s Ability initiative and the ICT Centre of Excellence Program.

the main result of this paper on measurement delays, namely Theorem 2. To model dynamics when delays are present requires a somewhat different type of stochastic “flocking matrix” than the one which is appropriate in the delay-free case. The graphs of the type of matrices to which we are referring are directed, just as in the delay-free case, but do not have self-arcs at every vertex. As a result, the set of such graphs, denoted by  $\mathcal{D}$ , is not closed under composition. The smallest set of directed graphs which contains  $\mathcal{D}$  and which is closed under composition is called the set of “extended delay graphs.” This class is explicitly characterized. Section 4 then develops the requisite properties of extended delay graphs needed to prove Theorem 2.

Section 5 considers a modified version of the flocking problem in which each agent independently updates its heading at times determined by its own clock. It is not assumed that the groups’ clocks are synchronized together or that the times any one agent updates its heading are evenly spaced. In this case, the deriving of conditions under which all agents eventually move with the same heading requires the analysis of the asymptotic behavior of an overall *asynchronous* process which models the  $n$ -agent system. The analysis is carried out by first embedding this process in a suitably defined *synchronous* discrete-time, hybrid dynamical system  $\mathbb{S}$ . This is accomplished using the concept of *analytic synchronization* outlined previously in [12, 13]. This enables us to bring to bear results derived earlier in [6] to characterize a rich class of system trajectories under which consensus is achieved.

In section 6 we briefly consider a modified version of the consensus problem for the same group of  $n$  agents as before but now with one of the group’s members (say agent 1) acting as the group’s *leader*. The remaining agents, called *followers* and labelled 2 through  $n$ , do not know who the leader is or even if there is a leader. Accordingly they continue to function as if there was no leader using the same update rules as are used in the leaderless case. The leader, on the other hand, acting on its own, ignores these update rules and moves with a constant heading. Using the main result on leaderless consensus summarized in section 2, we then develop conditions under which all follower agents eventually move in the same direction as the leader. These conditions correct prior findings on leader following in [11] which are in error.

**2. Background.** As in [6], the system of interest consists of  $n$  autonomous agents, labelled 1 through  $n$ , all moving in the plane with the same speed but with different headings. Each agent’s heading is updated using a simple local rule based on the average of its own heading plus the headings of its “neighbors.” Agent  $i$ ’s *neighbors* at time  $t$  are those agents, including itself, which are in a closed disk of prespecified radius  $r_i$  centered at agent  $i$ ’s current position. In what follows  $\mathcal{N}_i(t)$  denotes the set of labels of those agents which are neighbors of agent  $i$  at time  $t$ . Agent  $i$ ’s heading, written  $\theta_i$ , evolves in discrete time in accordance with a model of the form

$$(1) \quad \theta_i(t+1) = \frac{1}{n_i(t)} \left( \sum_{j \in \mathcal{N}_i(t)} \theta_j(t) \right),$$

where  $t$  is a discrete-time index taking values in the nonnegative integers  $\{0, 1, 2, \dots\}$ , and  $n_i(t)$  is the number of neighbors of agent  $i$  at time  $t$ .

**2.1. Neighbor graph.** The explicit form of the update equations determined by (1) depends on the relationships between neighbors which exist at time  $t$ . These relationships can be conveniently described by a directed graph  $\mathbb{N}(t)$  with vertex set



$\mathcal{V} = \{1, 2, \dots, n\}$  and arc set  $\mathcal{A}(\mathbb{N}(t)) \subset \mathcal{V} \times \mathcal{V}$  which is defined so that  $(i, j)$  is an arc or directed edge from  $i$  to  $j$  just in case agent  $i$  is a neighbor of agent  $j$  at time  $t$ . Thus  $\mathbb{N}(t)$  is a directed graph on  $n$  vertices with at most one arc connecting each ordered pair of distinct vertices and with exactly one self-arc at each vertex. We write  $\mathcal{G}_{sa}$  for the set of all such graphs and  $\mathcal{G}$  for the set of all directed graphs with vertex set  $\mathcal{V}$ . It is natural to call a vertex  $i$  a *neighbor* of vertex  $j$  in a graph  $\mathbb{G} \in \mathcal{G}$  if  $(i, j)$  is an arc in  $\mathbb{G}$ .

**2.2. Heading update rule.** The set of agent heading update rules defined by (1) can be written in state form. Towards this end, for each graph  $\mathbb{N} \in \mathcal{G}_{sa}$  define the *flocking matrix*

$$(2) \quad F = D^{-1}A',$$

where  $A'$  is the transpose of the adjacency matrix of  $\mathbb{N}$  and  $D$  the diagonal matrix whose  $j$ th diagonal element is the in-degree of vertex  $j$  within  $\mathbb{N}$ . Then

$$(3) \quad \theta(t+1) = F(t)\theta(t), \quad t \in \{0, 1, 2, \dots\},$$

where  $\theta$  is the heading vector  $\theta = [\theta_1 \ \theta_2 \ \dots \ \theta_n]'$  and  $F(t)$  is the flocking matrix of the *neighbor graph*  $\mathbb{N}(t)$ .

**2.3. Leaderless consensus.** To proceed, we need to recall a few definitions from [6]. We call a vertex  $i$  of a directed graph  $\mathbb{G}$  a *root* of  $\mathbb{G}$  if for each other vertex  $j$  of  $\mathbb{G}$ , there is a path from  $i$  to  $j$ . Thus  $i$  is a root of  $\mathbb{G}$  if it is the root of a directed spanning tree of  $\mathbb{G}$ . We say that  $\mathbb{G}$  is *rooted at  $i$*  if  $i$  is in fact a root. Thus  $\mathbb{G}$  is rooted at  $i$  just in case each other vertex of  $\mathbb{G}$  is *reachable* from vertex  $i$  along a path within the graph.  $\mathbb{G}$  is *strongly rooted at  $i$*  if each other vertex of  $\mathbb{G}$  is reachable from vertex  $i$  along a path of length 1. Thus  $\mathbb{G}$  is strongly rooted at  $i$  if  $i$  is a neighbor of every other vertex in the graph. A *rooted graph*  $\mathbb{G}$  is a graph which possesses at least one root. Finally, a *strongly rooted graph* is a graph which has at least one vertex at which it is strongly rooted.

By the *composition* of two directed graphs  $\mathbb{G}_p, \mathbb{G}_q$  with the same vertex set  $\mathcal{V}$  we mean the graph  $\mathbb{G}_q \circ \mathbb{G}_p$  with the same vertex set  $\mathcal{V}$  and arc set defined such that  $(i, j)$  is an arc of  $\mathbb{G}_q \circ \mathbb{G}_p$  if for some vertex  $k$ ,  $(i, k)$  is an arc of  $\mathbb{G}_p$  and  $(k, j)$  is an arc of  $\mathbb{G}_q$ . A finite sequence of directed graphs  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_q$  with the same vertex set is *jointly rooted* if the composition  $\mathbb{G}_q \circ \mathbb{G}_{q-1} \circ \dots \circ \mathbb{G}_1$  is rooted. An infinite sequence of graphs  $\mathbb{G}_1, \mathbb{G}_2, \dots$  with the same vertex set is *repeatedly jointly rooted by subsequences of length  $q$*  if there is a positive integer  $q$  for which each finite sequence  $\mathbb{G}_{qk+1}, \dots, \mathbb{G}_{q(k+1)}$ ,  $k \geq 0$ , is jointly rooted. The main result on leaderless consensus [14, 16] is equivalent to the following result from [6].

**THEOREM 1.** *Let  $\theta(0)$  be fixed. For any trajectory of the system determined by (1) along which the sequence of neighbor graphs  $\mathbb{N}(0), \mathbb{N}(1), \dots$  is repeatedly jointly rooted by sequences of length  $q$ , there is a constant  $\theta_{ss}$ , depending only on  $\theta(0)$ , for which*

$$(4) \quad \lim_{t \rightarrow \infty} \theta(t) = \theta_{ss} \mathbf{1},$$

where the limit is approached exponentially fast.

**3. Convergence rates.** The aim of this section is to derive a bound on the rate at which  $\theta$  converges.<sup>1</sup> There are two distinct ways to go about this, and below we

<sup>1</sup>This section summarizes and extends some of the key findings of [7].

describe both. To do this we will make use of certain structural properties of the  $F$ . As defined, each  $F$  is square and nonnegative, where by a *nonnegative* matrix we mean a matrix whose entries are all nonnegative. Each  $F$  also has the property that its row sums all equal 1 (i.e.,  $F\mathbf{1} = \mathbf{1}$ ). Matrices with these two properties are called (row) *stochastic* [10]. It is easy to verify that the class of all  $n \times n$  stochastic matrices is closed under multiplication. It is worth noting that because the vertices of the graphs in  $\mathcal{G}_{sa}$  all have self-arcs, the  $F$  also have the property that their diagonal elements are positive.

In what follows we write  $M \geq N$  whenever  $M - N$  is a nonnegative matrix. We also write  $M > N$  whenever  $M - N$  is a positive matrix, where by a *positive matrix* we mean a matrix with all positive entries. For any nonnegative matrix  $R$  of any size, we write  $\|R\|$  for the largest of the row sums of  $R$ . Note that  $\|R\|$  is the induced infinity norm of  $R$  and consequently is submultiplicative. Moreover,  $\|M_1\| \leq \|M_2\|$  if  $M_1 \leq M_2$ . Observe that for any  $n \times n$  stochastic matrix  $S$ ,  $\|S\| = 1$  because the row sums of a stochastic matrix all equal 1. As in [6] we write  $\lfloor M \rfloor$  and  $\lceil M \rceil$  for the  $1 \times m$  row vectors whose  $j$ th entries are the smallest and largest elements, respectively, of the  $j$ th column of  $M$ . Note that  $\lfloor M \rfloor$  is the largest  $1 \times m$  nonnegative row vector  $c$  for which  $M - \mathbf{1}c$  is nonnegative and that  $\lceil M \rceil$  is the smallest nonnegative row vector  $c$  for which  $\mathbf{1}c - M$  is nonnegative. Note in addition that for any  $n \times n$  stochastic matrix  $S$ , one can write

$$(5) \quad S = \mathbf{1}\lfloor S \rfloor + \lceil S \rceil \quad \text{and} \quad S = \mathbf{1}\lceil S \rceil - \lfloor S \rfloor,$$

where  $\lfloor S \rfloor$  and  $\lceil S \rceil$  are the nonnegative matrices defined by the equations

$$(6) \quad \lfloor S \rfloor = S - \mathbf{1}\lfloor S \rfloor \quad \text{and} \quad \lceil S \rceil = \mathbf{1}\lceil S \rceil - S,$$

respectively. Moreover, the row sums of  $\lfloor S \rfloor$  are all equal to  $1 - \lfloor S \rfloor \mathbf{1}$  and the row sums of  $\lceil S \rceil$  are all equal to  $\lceil S \rceil \mathbf{1} - 1$ , and so

$$(7) \quad \|\lfloor S \rfloor\| = 1 - \lfloor S \rfloor \mathbf{1} \quad \text{and} \quad \|\lceil S \rceil\| = \lceil S \rceil \mathbf{1} - 1.$$

In what follows we will also be interested in the matrix

$$(8) \quad \llbracket S \rrbracket = \lfloor S \rfloor + \lceil S \rceil.$$

This matrix satisfies

$$(9) \quad \llbracket S \rrbracket = \mathbf{1}(\lceil S \rceil - \lfloor S \rfloor)$$

because of (5).

To prove that all  $\theta_i$  converge to a common heading, it is necessary to prove that  $\theta$  converges to a vector of the form  $\theta_{ss}\mathbf{1}$ , where  $\mathbf{1}$  is the  $n \times 1$  vector of 1's. It is clear from (3) that  $\theta$  will converge to such a vector just in case, as  $t \rightarrow \infty$ , the matrix product  $F(t) \cdots F(0)$  converges to a rank one matrix of the form  $\mathbf{1}c$  for some  $n \times 1$  row vector  $c$ . Thus to study how such matrix products converge it is sufficient to study how products of stochastic matrices of the form  $S_j \cdots S_1$  converge as  $j \rightarrow \infty$ . As in [6], we say that a matrix product  $S_j S_{j-1} \cdots S_1$  *converges exponentially fast at a rate no slower than  $\lambda$*  to a matrix of the form  $\mathbf{1}c$  if there are nonnegative constants  $b$  and  $\lambda$  with  $\lambda < 1$ , such that

$$(10) \quad \|(S_j \cdots S_1) - \mathbf{1}c\| \leq b\lambda^j, \quad j \geq 1.$$

The following fact is proved in [6].

PROPOSITION 1. *If an infinite sequence of stochastic matrices  $S_1, S_2, \dots$  satisfies*

$$(11) \quad |||S_j \cdots S_1||| \leq \bar{b}\lambda^j, \quad j \geq 0,$$

*for some nonnegative constants  $\bar{b}$  and  $\lambda < 1$ , then the product  $S_j S_{j-1} \cdots S_1$  converges exponentially fast at a rate no slower than  $\lambda$  to a matrix of the form 1c.*

We will exploit this inequality in deriving specific convergence rates.

Any  $n \times n$  stochastic matrix  $S$  determines a directed graph  $\gamma(S)$  with the vertex set  $\{1, 2, \dots, n\}$  and arc set defined in such a way so that  $(i, j)$  is an arc of  $\gamma(S)$  from  $i$  to  $j$  just in case the  $j$ th entry of  $S$  is nonzero. Note that the graph of any stochastic matrix with positive diagonal elements must be in  $\mathcal{S}_{sa}$ . Since flocking matrices have this property, their graphs must be in  $\mathcal{G}_{sa}$ . It is known [6] that for the set of  $n \times n$  stochastic matrices  $S_1, S_2, \dots, S_p$

$$(12) \quad \gamma(S_p \cdots S_2 S_1) = \gamma(S_p) \circ \cdots \circ \gamma(S_2) \circ \gamma(S_1).$$

We will make use of the fact that for any two  $n \times n$  stochastic matrices  $S_1$  and  $S_2$ ,

$$(13) \quad \phi(S_2 S_1) \geq \phi(S_2)\phi(S_1),$$

where for any nonnegative matrix  $M$ ,  $\phi(M)$  denotes the smallest nonzero element of  $M$ . To prove that this is so, note first that any stochastic matrix  $S$  can be written as  $S = \phi(S)\bar{S}$ , where  $\bar{S}$  is a nonzero matrix whose nonzero entries are all bounded below by 1; moreover, if  $S = \hat{\phi}(S)\hat{S}$ , where  $\hat{\phi}(S)$  is a number and  $\hat{S}$  is also a nonzero matrix whose nonzero entries are all bounded below by 1, then  $\phi(S) \geq \hat{\phi}(S)$ . Accordingly, write  $S_i = \phi(S_i)\bar{S}_i$ ,  $i \in \{1, 2\}$ , where each  $\bar{S}_i$  is a nonzero matrix whose nonzero entries are all bounded below by 1. Since  $S_2 S_1 = \phi(S_2)\phi(S_1)\bar{S}_2\bar{S}_1$  and  $S_2 S_1$  is nonzero,  $\bar{S}_2\bar{S}_1$  must be nonzero as well. Moreover, the nonzero entries of  $\bar{S}_2\bar{S}_1$  must be bounded below by 1 because the product of any two  $n \times n$  matrices with all nonzero entries bounded below by 1 must be a matrix with the same property. Therefore  $\phi(S_2 S_1) \geq \phi(S_2)\phi(S_1)$  as claimed. An important consequence of (13) is that for any set of stochastic matrices  $S_1, S_2, \dots, S_m$  for which each  $\phi(S_i)$  is bounded below by a positive number  $b$ ,

$$(14) \quad \phi(S_m \cdots S_1) \geq b^m.$$

Our goal is now to use these facts to derive an explicit convergence rate for the situation considered by Theorem 1. We will do this in two different ways. The first way is based on properties of stochastic matrices with strongly rooted graphs.

**3.1. Strongly rooted graphs.** Let  $\mathcal{F}(q)$  denote the set of all products of  $q$  flocking matrices whose corresponding sequences of  $q$  graphs are each jointly rooted. In view of (12), each matrix in  $\mathcal{F}(q)$  must have a rooted graph in  $\mathcal{G}_{sa}$ . In other words, each matrix in  $\mathcal{F}(q)$  has a rooted graph and is a product of  $q$  flocking matrices. Since the set of all flocking matrices is finite, so is  $\mathcal{F}(q)$ . It is shown in [6] that the composition of any set of at least  $(n-1)^2$  rooted graphs in  $\mathcal{G}_{sa}$  is strongly rooted. This and (12) imply that the product of any  $(n-1)^2$  matrices in  $\mathcal{F}(q)$  must have a strongly rooted graph in  $\mathcal{G}_{sa}$ . Thus if we set  $m = (n-1)^2$  and write  $(\mathcal{F}(q))^m$  for the set of all products of  $m$  matrices from  $\mathcal{F}(q)$ , then each matrix in  $(\mathcal{F}(q))^m$  must have a

strongly rooted graph. Moreover,  $(\mathcal{F}(q))^m$  must be a finite set because  $\mathcal{F}(q)$  is. It is shown in [6] that convergence of the  $\theta_i$  in Theorem 1 occurs at a rate no slower than

$$\lambda = \left( \max_{S \in (\mathcal{F}(q))^m} |||S||| \right)^{\frac{1}{mq}}.$$

Our goal is to derive an explicit bound for  $\lambda$ .

As a first step towards this end, let  $S$  be any stochastic matrix with a strongly rooted graph. We claim that

$$(15) \quad |||S||| \leq 1 - \phi(S).$$

To understand why this is so, note first that because  $\gamma(S)$  is strongly rooted, at least one vertex—say the  $k$ th—must be a root with arcs to every other vertex. This means that the  $k$ th column of  $S$  must be positive. Since  $\phi(S)$  is a lower bound on all nonzero elements in  $S$ , the smallest element in the  $k$ th column of  $S$  is bounded below by  $\phi(S)$ . Therefore  $[S]\mathbf{1} \geq \phi(S)\mathbf{1}$ . But from (7)  $|||S||| = 1 - [S]\mathbf{1}$ . This implies that (15) is true.

As a second step, let us note that the definition of a flocking matrix implies that all nonzero entries are bounded below by  $\frac{1}{n}$ . In other words,  $F \in \mathcal{F}$  implies that  $\phi(F) \geq \frac{1}{n}$ . But the flocking matrix  $F = \frac{1}{n}\mathbf{1}\mathbf{1}'$  is in  $\mathcal{F}$ , and for this matrix  $\phi(F) = \frac{1}{n}$ . Therefore

$$(16) \quad \min_{F \in \mathcal{F}} \phi(F) = \frac{1}{n}.$$

Now suppose that  $S \in \mathcal{F}(q)$ . Thus  $S$  is the product of  $q$  matrices from  $\mathcal{F}$ . From this, (14), and (16) it follows that  $\phi(S) \geq \frac{1}{n^q}$ . Therefore

$$(17) \quad \min_{S \in \mathcal{F}(q)} \phi(S) \geq \frac{1}{n^q}.$$

Next suppose that  $S \in (\mathcal{F}(q))^m$ . Thus  $S$  is the product of  $m$  matrices from  $\mathcal{F}(q)$ . From this, (14), and (17) it follows that  $\phi(S) \geq \frac{1}{n^{qm}}$ . This and (15) thus imply that  $|||S||| \leq 1 - \frac{1}{n^{qm}}$  and thus that

$$\max_{S \in (\mathcal{F}(q))^{(n-1)^2}} |||S||| \leq 1 - \frac{1}{n^{qm}}.$$

Therefore, since  $m = (n-1)^2$ ,

$$(18) \quad \lambda \leq \left( 1 - \frac{1}{n^{q(n-1)^2}} \right)^{\frac{1}{q(n-1)^2}}.$$

The derivation of this particular upper bound on the rate at which the  $\theta_i$  converge to  $\theta_{ss}$  ultimately depends on two facts established in [6]. First, as we said before, the composition of at most  $(n-1)^2$  rooted graphs is strongly rooted. Second, for any infinite sequence of stochastic matrices  $S_1, S_2, \dots$ , with strongly rooted graphs which come from a compact set  $\mathcal{S}_{sr}$ , the product  $S_j \cdots S_1$  converges exponentially fast, as  $j \rightarrow \infty$ , to a rank one matrix  $\mathbf{1}c$  at a rate no slower than

$$\max_{S \in \mathcal{S}_{sr}} |||S|||.$$

It turns out that by exploiting two different but corresponding facts about stochastic matrices with “neighbor-shared” graphs we can obtain a significantly smaller bound than the one given by (18).

**3.2. Neighbor-shared graphs.** By a neighbor-shared graph we mean any graph with two or more vertices with the property that each pair of vertices in the graph shares a common neighbor. Every neighbor-shared graph is rooted, but the converse is false [6]. The convergence rate bounds we are about to derive depend on two facts. First, the composition of at most  $(n - 1)$  rooted graphs is neighbor shared [6]. Second, for any infinite sequence of stochastic matrices  $S_1, S_2, \dots$ , with neighbor-shared graphs which come from a compact set  $\mathcal{S}_{ns}$ , the product  $S_j \cdots S_1$  converges exponentially fast, as  $j \rightarrow \infty$ , to a rank one matrix  $\mathbf{1}c$  at a rate no slower than

$$\max_{S \in \mathcal{S}_{ns}} \mu(S),$$

where  $\mu(S)$  is a positive number, called a *scrambling constant*, which is defined by the formula

$$(19) \quad \mu(S) = \max_{i,j} \left( 1 - \sum_{k=1}^n \min\{s_{ik}, s_{jk}\} \right).$$

In what follows we will make use of some well-known ideas and constructions from the theory of nonhomogeneous Markov chains [17] to explain why this second statement is true.

**Scrambling constants.** Let  $S$  be any  $n \times n$  stochastic matrix. Observe that for any nonnegative  $n$ -vector  $x$ , the  $i$ th minus the  $j$ th entries of  $Sx$  can be written as

$$\sum_{k=1}^n (s_{ik} - s_{jk})x_k = \sum_{k \in \mathcal{K}} (s_{ik} - s_{jk})x_k + \sum_{k \in \bar{\mathcal{K}}} (s_{ik} - s_{jk})x_k,$$

where

$$\mathcal{K} = \{k : s_{ik} - s_{jk} \geq 0, k \in \{1, 2, \dots, n\}\}$$

and

$$\bar{\mathcal{K}} = \{k : s_{ik} - s_{jk} < 0, k \in \{1, 2, \dots, n\}\}.$$

Therefore

$$\sum_{k=1}^n (s_{ik} - s_{jk})x_k \leq \left( \sum_{k \in \mathcal{K}} (s_{ik} - s_{jk}) \right) \lceil x \rceil + \left( \sum_{k \in \bar{\mathcal{K}}} (s_{ik} - s_{jk}) \right) \lfloor x \rfloor.$$

But

$$\sum_{k \in \mathcal{K} \cup \bar{\mathcal{K}}} (s_{ik} - s_{jk}) = 0,$$

and so

$$\sum_{k \in \bar{\mathcal{K}}} (s_{ik} - s_{jk}) = - \sum_{k \in \mathcal{K}} (s_{ik} - s_{jk}).$$

Thus

$$\sum_{k=1}^n (s_{ik} - s_{jk})x_k \leq \left( \sum_{k \in \mathcal{K}} (s_{ik} - s_{jk}) \right) (\lceil x \rceil - \lfloor x \rfloor).$$

Now

$$\sum_{k \in \mathcal{K}} (s_{ik} - s_{jk}) = 1 - \sum_{k \in \bar{\mathcal{K}}} s_{ik} - \sum_{k \in \mathcal{K}} s_{jk}$$

because the row sums of  $S$  are all one. Moreover,

$$s_{ik} = \min\{s_{ik}, s_{jk}\}, \quad k \in \bar{\mathcal{K}},$$

$$s_{jk} = \min\{s_{ik}, s_{jk}\}, \quad k \in \mathcal{K},$$

and so

$$\sum_{k \in \mathcal{K}} (s_{ik} - s_{jk}) = 1 - \sum_{k=1}^n \min\{s_{ik}, s_{jk}\}.$$

It follows that

$$\sum_{k=1}^n (s_{ik} - s_{jk}) x_k \leq \left(1 - \sum_{k=1}^n \min\{s_{ik}, s_{jk}\}\right) (\lceil x \rceil - \lfloor x \rfloor).$$

Hence with  $\mu$  as defined by (19),

$$\sum_{k=1}^n (s_{ik} - s_{jk}) x_k \leq \mu(S)(\lceil x \rceil - \lfloor x \rfloor).$$

Since this holds for all  $i, j$ , it must hold for the  $i$  and  $j$  for which

$$\sum_{k=1}^n s_{ik} x_k = \lceil Sx \rceil \quad \text{and} \quad \sum_{k=1}^n s_{jk} x_k = \lfloor Sx \rfloor.$$

Therefore

$$(20) \quad \lceil Sx \rceil - \lfloor Sx \rfloor \leq \mu(S)(\lceil x \rceil - \lfloor x \rfloor).$$

Now let  $S_1$  and  $S_2$  be any two  $n \times n$  stochastic matrices and let  $e_i$  be the  $i$ th unit  $n$ -vector. Then from (20),

$$(21) \quad \lceil S_2 S_1 e_i \rceil - \lfloor S_2 S_1 e_i \rfloor \leq \mu(S_2)(\lceil S_1 e_i \rceil - \lfloor S_1 e_i \rfloor).$$

Meanwhile, from (9),

$$\llbracket S_2 S_1 \rrbracket e_i = \mathbf{1}(\lceil S_2 S_1 \rceil - \lfloor S_2 S_1 \rfloor) e_i$$

and

$$\llbracket S_1 \rrbracket e_i = \mathbf{1}(\lceil S_1 \rceil - \lfloor S_1 \rfloor) e_i.$$

But for any nonnegative matrix  $M$ ,  $\lceil M \rceil e_i = \lceil M e_i \rceil$  and  $\lfloor M \rfloor e_i = \lfloor M e_i \rfloor$ , and so

$$\llbracket S_2 S_1 \rrbracket e_i = \mathbf{1}(\lceil S_2 S_1 e_i \rceil - \lfloor S_2 S_1 e_i \rfloor)$$

and

$$\llbracket S_1 \rrbracket e_i = \mathbf{1}(\lceil S_1 e_i \rceil - \lfloor S_1 e_i \rfloor).$$

From these expressions and (21) it follows that

$$\llbracket S_2 S_1 \rrbracket e_i \leq \mu(S_2) \llbracket S_1 \rrbracket e_i.$$

Since this is true for all  $i$ , we arrive at the following fact.

LEMMA 1. *For any two stochastic matrices in  $\mathcal{S}$ ,*

$$(22) \quad \llbracket S_2 S_1 \rrbracket \leq \mu(S_2) \llbracket S_1 \rrbracket.$$

Note that since the row sums of  $S$  all equal 1,  $\mu(S)$  is nonnegative. It is easy to see that  $\mu(S) = 0$  just in case all the rows of  $S$  are equal. Let us note that for fixed  $i$  and  $j$ , the  $k$ th term in the sum appearing in (19) will be positive just in case both  $s_{ik}$  and  $s_{jk}$  are positive. It follows that the sum will be positive if and only if for at least one  $k$ ,  $s_{ik}$  and  $s_{jk}$  are both positive. Thus  $\mu(S) < 1$  if and only if for each distinct  $i$  and  $j$ , there is at least one  $k$  for which  $s_{ik}$  and  $s_{jk}$  are both positive. Matrices with this property have been widely studied and are called *scrambling matrices* [17]. Thus a stochastic matrix  $S$  is a scrambling matrix if and only if  $\mu(S) < 1$ . It is easy to see that the definition of a scrambling matrix also implies that  $S$  is scrambling if and only if its graph  $\gamma(S)$  is neighbor-shared.

As before, let  $\mathcal{S}_{ns}$  be a closed subset consisting of stochastic matrices whose graphs are all neighbor-shared. Then the scrambling constant  $\mu(S)$  defined in (19) satisfies  $\mu(S) < 1$ ,  $S \in \mathcal{S}_{ns}$  because each such  $S$  is a scrambling matrix. Let

$$\bar{\mu} = \max_{S \in \mathcal{S}_{ns}} \mu(S).$$

Then  $\bar{\mu} < 1$  because  $\mathcal{S}_{ns}$  is closed and bounded and because  $\mu(\cdot)$  is continuous. In view of Lemma 1,

$$\|\llbracket S_2 S_1 \rrbracket\| \leq \bar{\mu} \|\llbracket S_1 \rrbracket\|, \quad S_1, S_2 \in \mathcal{S}_{ns}.$$

Hence by induction, for any sequence of matrices  $S_1, S_2, \dots$  in  $\mathcal{S}_{ns}$

$$\|\llbracket S_j \cdots S_1 \rrbracket\| \leq \bar{\mu}^{j-1} \|\llbracket S_1 \rrbracket\|, \quad S_i \in \mathcal{S}_{ns}.$$

But from (8),  $\llbracket S \rrbracket \leq \llbracket S \rrbracket$ ,  $S \in \mathcal{S}$ , and so  $\|\llbracket S \rrbracket\| \leq \|\llbracket S \rrbracket\|$ ,  $S \in \mathcal{S}$ . Therefore for any sequence of stochastic matrices  $S_1, S_2, \dots$  with neighbor-shared graphs

$$(23) \quad \|\llbracket S_j \cdots S_1 \rrbracket\| \leq \bar{\mu}^{j-1} \|\llbracket S_1 \rrbracket\|.$$

Therefore from Proposition 1, any such product  $S_j \cdots S_1$  converges exponentially at a rate no slower than  $\bar{\mu}$  as  $j \rightarrow \infty$ . This establishes the validity of the statement about convergence of products of stochastic matrices made at the beginning of section 3.2.

Suppose now that  $F$  is a flocking matrix for which  $\gamma(F)$  is neighbor-shared. In view of the definition of a flocking matrix, any nonzero entry in  $F$  must be bounded below by  $\frac{1}{n}$ . Fix distinct  $i$  and  $j$  and suppose that  $k$  is a neighbor that  $i$  and  $j$  share. Then  $f_{ik}$  and  $f_{jk}$  are both nonzero, and so  $\min\{f_{ik}, f_{jk}\} \geq \frac{1}{n}$ . This implies that the sum in (19) must be bounded below by  $\frac{1}{n}$  and consequently that  $\mu(F) \leq 1 - \frac{1}{n}$ .

Now let  $F$  be that flocking matrix whose graph  $\gamma(F)$  is such that vertex 1 has no neighbors other than itself, vertex 2 has every vertex as a neighbor, and vertices 3 through  $n$  have only themselves and agent 1 as neighbors. Since vertex 1 has no neighbors other than itself,  $f_{1k} = 0$  for all  $k > 1$ . Thus for all  $i, j$ , it must be true that  $\sum_{k=1}^n \min\{f_{ik}, f_{jk}\} = \min\{f_{i1}, f_{j1}\}$ . Now vertex 2 has  $n$  neighbors, and so  $f_{2,1} = \frac{1}{n}$ .

Thus  $\min\{f_{i1}, f_{j1}\}$  attains its lower bound of  $\frac{1}{n}$  when either  $i = 2$  or  $j = 2$ . It thus follows that with this  $F$ ,  $\mu(F)$  attains its upper bound of  $1 - \frac{1}{n}$ . We summarize.

LEMMA 2. *Let  $\mathcal{F}_{ns}$  be the set of  $n \times n$  flocking matrices with neighbor-shared graphs. Then*

$$(24) \quad \max_{F \in \mathcal{F}_{ns}} \mu(F) = 1 - \frac{1}{n}.$$

Thus  $1 - \frac{1}{n}$  is a tight bound on the convergence rate for an infinite product of flocking matrices with neighbor-shared graphs. In [6] it is shown that

$$\max_{F \in \mathcal{F}_{sr}} \mu(F) = 1 - \frac{1}{n},$$

where  $\mathcal{F}_{sr}$  is the set of flocking matrices with strongly rooted graphs. Thus  $1 - \frac{1}{n}$  is also a tight bound on the convergence rate for an infinite product of flocking matrices with strongly rooted graphs. Of course a strongly rooted graph is a more special type of graph than a neighbor-shared graph because strongly rooted graphs are neighbor shared but not conversely.

We now use the preceding to derive a better convergence rate bound than the one in (18) for the type of trajectory addressed by Theorem 1. As a first step towards this end, we exploit the fact that for any  $n \times n$  stochastic scrambling matrix  $S$ , the scrambling constant of  $\mu(S)$  satisfies the inequality

$$(25) \quad \mu(S) \leq 1 - \phi(S).$$

To understand why this is so, assume that  $S$  is any given scrambling matrix. Note that for any distinct  $i$  and  $j$ , there must be a  $k$  for which  $\min\{s_{ik}, s_{jk}\}$  is nonzero and bounded below by  $\phi(S)$ . Thus

$$\sum_{k=1}^n \min\{s_{ik}, s_{jk}\} \geq \phi(S),$$

and so

$$1 - \sum_{k=1}^n \min\{s_{ik}, s_{jk}\} \leq 1 - \phi(S).$$

But this holds for all distinct  $i$  and  $j$ . In view of the definition of  $\mu(S)$  in (19), (25) must therefore be true.

As before, let  $\mathcal{F}(q)$  denote the set of products of  $q$  flocking matrices  $F_1, F_2, \dots, F_q$  from  $\mathcal{F}$  for which  $\{\gamma(F_1), \gamma(F_2), \dots, \gamma(F_q)\}$  is a jointly rooted set. Then as noted before, each matrix in  $\mathcal{F}(q)$  is rooted. Set  $p = (n-1)$  and let  $(\mathcal{F}(q))^p$  now denote the set of all products of  $p$  matrices from  $\mathcal{F}(q)$ . Then each matrix in  $(\mathcal{F}(q))^p$  is neighbor-shared. Let  $S$  be any such matrix. Then  $S$  is a product of  $qp$  flocking matrices. But each such flocking matrix  $F$  satisfies (16). Because of this and (14), it must be true that  $\phi(S) \geq \frac{1}{n^{qp}}$ . Therefore  $\mu(S) \leq 1 - \frac{1}{n^{qp}}$  because of (25). Since this is true for all  $S \in (\mathcal{F}(q))^p$ ,  $1 - \frac{1}{n^{qp}}$  must be a convergence rate upper bound for all infinite products of matrices from  $(\mathcal{F}(q))^p$ . Therefore, since  $p = n-1$ ,

$$(26) \quad \left(1 - \frac{1}{n^{q(n-1)}}\right)^{\frac{1}{q(n-1)}}$$



must be an upper bound on the convergence rate for all infinite products of flocking matrices  $F_1, F_2, \dots$  which have the property that the sequence of graphs  $\gamma(F_1), \gamma(F_2), \dots$  is repeatedly jointly rooted by subsequences of length  $q$ . Since this is precisely the type of sequence of flocking matrices which arise under the assumptions of Theorem 1, (26) is a convergence rate bound for the type of trajectory addressed by the theorem. Note that this convergence rate upper bound is much smaller (i.e., faster) than the one given by (18). We refer the reader to [7] for additional convergence rate calculations along these lines.

**4. Measurement delays.** In this section we consider a modified version of the flocking problem in which integer-valued delays occur in sensing the values of headings which are available to agents. More precisely we suppose that at each time  $t \in \{0, 1, 2, \dots\}$ , the value of neighboring agent  $j$ 's headings which agent  $i$  may sense is  $\theta_j(t - d_{ij}(t))$ , where  $d_{ij}(t)$  is a delay whose value at  $t$  is some integer between 0 and  $m_j - 1$ ; here  $m_j$  is a prespecified positive integer. While well-established principles of feedback control would suggest that delays should be dealt with using dynamic compensation, in this paper we will consider the situation in which the delayed value of agent  $j$ 's heading sensed by agent  $i$  at time  $t$  is the value which will be used in the heading update law for agent  $i$ . Thus

$$(27) \quad \theta_i(t+1) = \frac{1}{n_i(t)} \left( \sum_{j \in \mathcal{N}_i(t)} \theta_j(t - d_{ij}(t)) \right),$$

where  $d_{ij}(t) \in \{0, 1, \dots, (m_j - 1)\}$  if  $j \neq i$  and  $d_{ij}(t) = 0$  if  $i = j$ . Our main result is the following theorem, which states in essence that the conclusions of Theorem 1 continue to hold for the update model described by (27).

**THEOREM 2.** *Let  $\theta(0)$  be fixed. For any trajectory of the system determined by (27) along which the sequence of neighbor graphs  $\mathbb{N}(0), \mathbb{N}(1), \dots$  is repeatedly jointly rooted, there is a constant  $\theta_{ss}$ , depending only on  $\theta(0)$ , for which*

$$(28) \quad \lim_{t \rightarrow \infty} \theta(t) = \theta_{ss} \mathbf{1},$$

where the limit is approached exponentially fast.

As noted in the introduction, the consensus problem with measurement delays we have been discussing has been considered previously in [3]. It is possible to compare the hypotheses of Theorem 2 with the corresponding hypotheses for exponential convergence stated in [3], namely assumptions 2 and 3 of that paper. To do this, let us agree, as before, to say that the *union* of a set of graphs  $\mathbb{G}_{r_1}, \mathbb{G}_{r_2}, \dots, \mathbb{G}_{r_k}$  with vertex set  $\mathcal{V}$  is the graph with the vertex set  $\mathcal{V}$  and arc set consisting of the union of the arcs of all of the graphs  $\mathbb{G}_{r_1}, \mathbb{G}_{r_2}, \dots, \mathbb{G}_{r_k}$ . Taken together, assumptions 2 and 3 of [3] are essentially equivalent to assuming that there are finite positive integers  $q$  and  $s$  such that the *union*

$$\mathbb{G}(k) \triangleq \mathbb{N}((k+1)q-1) \cup \mathbb{N}((k+1)q-2) \cup \dots \cup \mathbb{N}(kq)$$

is strongly connected and independent of  $k$  for  $k \geq s$ . By way of comparison, the hypothesis of Theorem 2 is equivalent to assuming that there is a finite positive integer  $q$  such that the *composition*

$$\bar{\mathbb{G}}(k) \triangleq \mathbb{N}((k+1)q-1) \circ \mathbb{N}((k+1)q-2) \circ \dots \circ \mathbb{N}(kq)$$

is rooted for  $k \geq 0$ . The latter assumption is weaker than the former for several reasons. First, the arc set of  $\mathbb{G}(k)$  is always a subset of the arc set of  $\bar{\mathbb{G}}(k)$ , and in some cases the containment may be strict. Second,  $\bar{\mathbb{G}}(k)$  is not assumed to be independent of  $k$ , even for  $k$  sufficiently large, whereas  $\mathbb{G}(k)$  is; in other words,  $\bar{\mathbb{G}}(k)$  is not assumed to converge, whereas  $\mathbb{G}(k)$  is. Third, each  $\mathbb{G}(k)$  is assumed to be strongly connected, whereas each  $\bar{\mathbb{G}}(k)$  need only be rooted; note that a strongly connected graph is a special type of rooted graph in which every vertex is a root. Perhaps what is most important about Theorem 2 and the development which justifies it is that the underlying structural properties of the graphs involved required for consensus are explicitly determined.

**4.1. State space system.** Using standard lifting techniques for dealing with delays in discrete-time systems, it is possible to represent the agent system defined by (27) as a state space model similar to the model discussed earlier for the delay-free case. Our first objective is to characterize the class of graphs  $\mathcal{D}$  of the stochastic matrices which result from this lifting process. Towards this end, let  $\bar{\mathcal{G}}$  denote the set of all directed graphs with vertex set  $\bar{\mathcal{V}} = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \dots \cup \mathcal{V}_n$ , where  $\mathcal{V}_i = \{v_{i1} \dots, v_{im_i}\}$ . Here vertex  $v_{ij}$  labels the  $j$ th possible delay value of agent  $i$ , namely  $j - 1$ . We sometimes write  $i$  for  $v_{i1}$ ,  $i \in \{1, 2, \dots, n\}$ , write  $\mathcal{V}$  for the subset of vertices  $\{v_{11}, v_{21}, \dots, v_{n1}\}$ , and think of  $v_{i1}$  as an alternative label of agent  $i$ .

To take account of the fact that each agent can use its own current heading in its update formula (27), we will utilize those graphs in  $\bar{\mathcal{G}}$  which have self-arcs at each vertex in  $\mathcal{V}$ . We will also require the arc set of each such graph to have, for  $i \in \{1, 2, \dots, n\}$ , an arc from each vertex  $v_{ij} \in \mathcal{V}_i$  except the last to its successor  $v_{i(j+1)} \in \mathcal{V}_i$ . Finally we stipulate that for each  $i \in \{1, 2, \dots, n\}$ , each vertex  $v_{ij}$  with  $j > 1$  has in-degree of exactly 1. In what follows we call any such graph a *delay graph* and write  $\mathcal{D}$  for the subset of all such graphs. Note that unlike the class of graphs  $\mathcal{G}_{sa}$  considered before, there are graphs in  $\mathcal{D}$  possessing vertices without self-arcs. Nonetheless each vertex of each graph in  $\mathcal{D}$  has positive in-degree. An example of a delay graph for a three-agent system is shown in Figure 1.

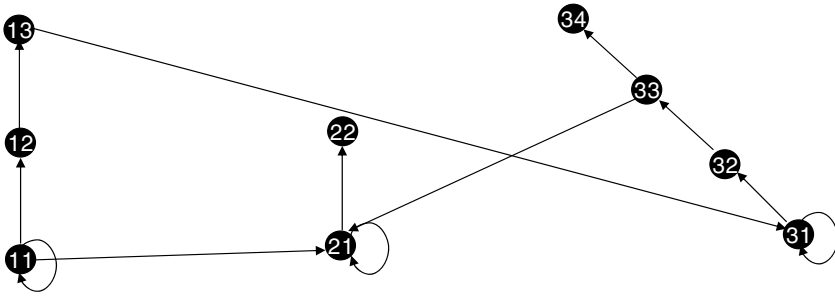


FIG. 1. Delay graph.

The specific delay graph representing the sensed headings the agents use at time  $t$  to update their own headings according to (27) is the graph  $\mathbb{D}(t) \in \mathcal{D}$  whose arc set contains an arc from  $v_{ik} \in \mathcal{V}_i$  to  $v_{j1} \in \mathcal{V}$  if agent  $j$  uses  $\theta_i(t + 1 - k)$  to update. There is a simple relationship between  $\mathbb{D}(t)$  and the neighbor graph  $\mathbb{N}(t)$  defined earlier. In particular,

$$(29) \quad \mathbb{N}(t) = Q(\mathbb{D}(t)),$$

where  $Q(\mathbb{D}(t))$  is the “quotient graph” of  $\mathbb{D}(t)$ . By the *quotient graph* of any  $\mathbb{G} \in \bar{\mathcal{G}}$ , written  $Q(\mathbb{G})$ , we mean the directed graph in  $\bar{\mathcal{G}}$  with vertex set  $\mathcal{V}$  whose arc set consists of those arcs  $(i, j)$  for which  $\mathbb{G}$  has an arc from some vertex in  $\mathcal{V}_i$  to some vertex in  $\mathcal{V}_j$ . The quotient graph of  $\mathbb{D}(t)$  thus models which headings are being used by each agent in updates at time  $t$  without describing the specific delayed headings actually being used. The quotient graph of the delay graph in Figure 1 is shown in Figure 2.

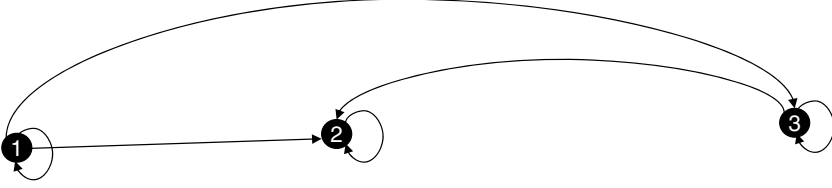


FIG. 2. Quotient graph.

The set of agent heading update rules defined by (27) can be written in state form. Towards this end, define  $\theta(t)$  to be the  $(m_1 + m_2 + \cdots + m_i)$  vector whose first  $m_1$  elements are  $\theta_1(t)$  to  $\theta_1(t+1-m_1)$ , whose next  $m_2$  elements are  $\theta_2(t)$  to  $\theta_2(t+1-m_2)$ , and so on. Order the vertices of  $\bar{\mathcal{V}}$  as  $v_{11}, \dots, v_{1m_1}, v_{21}, \dots, v_{2m_2}, \dots, v_{n1}, \dots, v_{nm_n}$ , and with respect to this ordering define for each graph  $\mathbb{D} \in \mathcal{D}$  the flocking matrix

$$(30) \quad F = D^{-1}A',$$

where  $A'$  is the transpose of the adjacency matrix of  $\mathbb{D}$  and  $D$  the diagonal matrix whose  $ij$ th diagonal element is the in-degree of vertex  $v_{ij}$  within the graph. Then  $\gamma(F) = \mathbb{D}$  and

$$(31) \quad \theta(t+1) = F(t)\theta(t), \quad t \in \{0, 1, 2, \dots\}.$$

Let  $\bar{\mathcal{F}}$  denote the set of all such  $F$ . As before our goal is to characterize the sequences of neighbor graphs  $N(0), N(1), \dots$  for which all entries of  $\theta(t)$  converge to a common steady state value.

There are a number of similarities and a number of differences between the situation under consideration here and the delay-free situation considered in [6]. For example, the notion of graph composition defined earlier can be defined in the obvious way for graphs in  $\bar{\mathcal{G}}$ . On the other hand, unlike the situation in the delay-free case, the set of graphs used to model the system under consideration, namely the set of delay graphs  $\mathcal{D}$ , is not closed under composition except in the special case when all of the delays are at most 1, i.e., when all of the  $m_i \leq 2$ . In order to characterize the smallest subset of  $\bar{\mathcal{G}}$  containing  $\mathcal{D}$  which is closed under composition, we will need several new concepts.

**4.2. Hierarchical graphs.** As before, let  $\mathcal{G}$  be the set of all directed graphs with vertex set  $\mathcal{V} = \{1, 2, \dots, n\}$ . Let us agree to say that a rooted graph  $\mathbb{G} \in \mathcal{G}$  is a *hierarchical graph* with *hierarchy*  $\{v_1, v_2, \dots, v_n\}$  if it is possible to relabel the vertices in  $\mathcal{V}$  as  $v_1, v_2, \dots, v_n$  in such a way so that  $v_1$  is a root of  $\mathbb{G}$  with a self-arc and for  $i > 1$ ,  $v_i$  has a neighbor  $v_j$  “lower” in the hierarchy where by *lower* we mean  $j < i$ . It is clear that any graph in  $\mathcal{G}$  with a root possessing a self-arc is hierarchical. Note that a graph may have more than one hierarchy and two graphs with the same hierarchy need not be equal. Note also that even though rooted graphs with the same hierarchy

share a common root, examples show that the composition of hierarchical graphs in  $\mathcal{G}$  need not be hierarchical or even rooted. On the other hand, the composition of two rooted graphs in  $\mathcal{G}$  with the same hierarchy is always a graph with the same hierarchy. To understand why this is so, consider two graphs  $\mathbb{G}_1$  and  $\mathbb{G}_2$  in  $\mathcal{G}$  with the same hierarchy  $\{v_1, v_2, \dots, v_n\}$ . Note first that  $v_1$  has a self-arc in  $\mathbb{G}_2 \circ \mathbb{G}_1$  because  $v_1$  has self-arcs in  $\mathbb{G}_1$  and  $\mathbb{G}_2$ . Next pick any vertex  $v_i$  in  $\mathcal{V}$  other than  $v_1$ . By definition, there must exist vertex  $v_j$  lower in the hierarchy than  $v_i$  such that  $(v_j, v_i)$  is an arc of  $\mathbb{G}_2$ . If  $v_j = v_1$ , then  $(v_1, v_i)$  is an arc in  $\mathbb{G}_2 \circ \mathbb{G}_1$  because  $v_1$  has a self-arc in  $\mathbb{G}_1$ . On the other hand, if  $v_j \neq v_1$ , then there must exist a vertex  $v_k$  lower in the hierarchy than  $v_j$  such that  $(v_k, v_j)$  is an arc of  $\mathbb{G}_1$ . It follows from the definition of composition that in this case  $(v_k, v_i)$  is an arc in  $\mathbb{G}_2 \circ \mathbb{G}_1$ . Thus  $v_i$  has a neighbor in  $\mathbb{G}_2 \circ \mathbb{G}_1$  which is lower in the hierarchy than  $v_i$ . Since this is true for all  $v_i$ ,  $\mathbb{G}_2 \circ \mathbb{G}_1$  must have the same hierarchy as  $\mathbb{G}_1$  and  $\mathbb{G}_2$ . This proves the claim that composition of two rooted graphs with the same hierarchy is a graph with the same hierarchy.

Our objective is to show that the composition of a sufficiently large number of graphs in  $\mathcal{G}$  with the same hierarchy is strongly rooted. Note that the fact that the composition of  $(n-1)^2$  graphs in  $\mathcal{G}_{sa}$  is rooted [6] cannot be used to reach this conclusion because the  $v_i$  in the graphs under consideration here do not all necessarily have self-arcs.

As before, let  $\mathbb{G}_1$  and  $\mathbb{G}_2$  be two graphs in  $\mathcal{G}$  with the same hierarchy  $\{v_1, v_2, \dots, v_n\}$ . Let  $v_i$  be any vertex in the hierarchy and suppose that  $v_j$  is a neighbor vertex of  $v_i$  in  $\mathbb{G}_2$ . If  $v_j = v_1$ , then  $v_i$  retains  $v_1$  as a neighbor in the composition  $\mathbb{G}_2 \circ \mathbb{G}_1$  because  $v_1$  has a self-arc in  $\mathbb{G}_1$ . On the other hand, if  $v_j \neq v_1$ , then  $v_j$  has a neighboring vertex  $v_k$  in  $\mathbb{G}_1$  which is lower in the hierarchy than  $v_j$ . Since  $v_k$  is a neighbor of  $v_i$  in the composition  $\mathbb{G}_2 \circ \mathbb{G}_1$ , we see that in this case  $v_i$  has acquired a neighbor in  $\mathbb{G}_2 \circ \mathbb{G}_1$  lower in the hierarchy than a neighbor it had in  $\mathbb{G}_2$ . In summary, any vertex  $v_i \in \mathcal{V}$  either has  $v_1$  as neighbor in  $\mathbb{G}_2 \circ \mathbb{G}_1$  or has a neighbor in  $\mathbb{G}_2 \circ \mathbb{G}_1$  which is at least one vertex lower in the hierarchy than any neighbor it had in  $\mathbb{G}_2$ .

Now consider three graphs  $\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3$  in  $\mathcal{G}$  with the same hierarchy. By the same reasoning as above, any vertex  $v_i \in \mathcal{V}$  either has  $v_1$  as neighbor in  $\mathbb{G}_3 \circ \mathbb{G}_2 \circ \mathbb{G}_1$  or has a neighbor in  $\mathbb{G}_3 \circ \mathbb{G}_2 \circ \mathbb{G}_1$  which is at least one vertex lower in the hierarchy than any neighbor it had in  $\mathbb{G}_3 \circ \mathbb{G}_2$ . Similarly  $v_i$  either has  $v_1$  as neighbor in  $\mathbb{G}_3 \circ \mathbb{G}_2$  or has a neighbor in  $\mathbb{G}_3 \circ \mathbb{G}_2$  which is at least one vertex lower in the hierarchy than any neighbor it had in  $\mathbb{G}_3$ . Combining these two observations we see that any vertex  $v_i \in \mathcal{V}$  either has  $v_1$  as neighbor in  $\mathbb{G}_3 \circ \mathbb{G}_2 \circ \mathbb{G}_1$  or has a neighbor in  $\mathbb{G}_3 \circ \mathbb{G}_2 \circ \mathbb{G}_1$  which is at least two vertices lower in the hierarchy than any neighbor it had in  $\mathbb{G}_3$ . This clearly generalizes, and so after the composition of  $m$  such graphs  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_m$ ,  $v_i$  either has  $v_1$  as neighbor in  $\mathbb{G}_m \circ \dots \circ \mathbb{G}_2 \circ \mathbb{G}_1$  or has a neighbor in  $\mathbb{G}_m \circ \dots \circ \mathbb{G}_2 \circ \mathbb{G}_1$  which is at least  $m-1$  vertices lower in the hierarchy than any neighbor it had in  $\mathbb{G}_m$ . It follows that if  $m \geq n$ , then  $v_i$  must be a neighbor of  $v_1$ . Since this is true for all vertices, we have proved the following.

**PROPOSITION 2.** *Let  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_m$  denote a set of rooted graphs in  $\mathcal{G}$  which all have the same hierarchy. If  $m \geq n-1$ , then  $\mathbb{G}_m \circ \dots \circ \mathbb{G}_2 \circ \mathbb{G}_1$  is strongly rooted.*

**4.3. The Closure of  $\mathcal{D}$ .** We now return to the study of the graphs in  $\mathcal{D}$ . As before  $\mathcal{D}$  is the subset of  $\bar{\mathcal{G}}$  consisting of those graphs which (i) have self-arcs at each vertex in  $\mathcal{V} = \{v_{11}, v_{21}, \dots, v_{n1}\}$ , (ii) for each  $i \in \{1, 2, \dots, n\}$  have an arc from each vertex  $v_{ij} \in \mathcal{V}_i$  except the last to its successor  $v_{i(j+1)} \in \mathcal{V}_i$ , and (iii) for each  $i \in \{1, 2, \dots, n\}$ , each vertex  $v_{ij}$  with  $j > 1$  has in-degree of exactly 1. It can easily be shown by example that  $\mathcal{D}$  is not closed under composition. We deal with this

problem as follows. First, let us agree to say that a vertex  $v$  in a graph  $\mathbb{G} \in \bar{\mathcal{G}}$  is a *neighbor* of a subset of  $\mathbb{G}$ 's vertices  $\mathcal{U}$  if  $v$  is a neighbor of at least one vertex in  $\mathcal{U}$ . Next we say that a graph  $\mathbb{G} \in \bar{\mathcal{G}}$  is an *extended delay graph* if for each  $i \in \{1, 2, \dots, n\}$ , (i) every neighbor of  $\mathcal{V}_i$  which is not in  $\mathcal{V}_i$  is a neighbor of  $v_{i1}$  and (ii) the subgraph of  $\mathbb{G}$  induced by  $\mathcal{V}_i$  has  $\{v_{i1}, \dots, v_{im_i}\}$  as a hierarchy. We write  $\bar{\mathcal{D}}$  for the set of all extended delay graphs in  $\bar{\mathcal{G}}$ . It is easy to see that every delay graph is an extended delay graph. The converse, however, is not true. The set of extended delay graphs has the following property.

PROPOSITION 3.  $\bar{\mathcal{D}}$  is closed under composition.

In light of this proposition it is natural to call  $\bar{\mathcal{D}}$  the *closure* of  $\mathcal{D}$ . To prove the proposition, we will need the following fact.

LEMMA 3. Let  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_q$  be any sequence of  $q > 1$  directed graphs with vertex set  $\mathcal{V}$ . For  $i \in \{1, 2, \dots, q\}$ , let  $\bar{\mathbb{G}}_i$  be the subgraph of  $\mathbb{G}_i$  induced by  $\mathcal{U} \subset \mathcal{V}$ . Then  $\bar{\mathbb{G}}_q \circ \dots \circ \bar{\mathbb{G}}_2 \circ \bar{\mathbb{G}}_1$  is contained in the subgraph of  $\mathbb{G}_q \circ \dots \circ \mathbb{G}_2 \circ \mathbb{G}_1$  induced by  $\mathcal{U}$ .

*Proof of Lemma 3.* It will be enough to prove the lemma for  $q = 2$ , since the proof for  $q > 2$  would then directly follow by induction. Suppose  $q = 2$ . Let  $(i, j)$  be in  $\mathcal{A}(\bar{\mathbb{G}}_2 \circ \bar{\mathbb{G}}_1)$ . Then  $i, j \in \mathcal{U}$  and there exists an integer  $k \in \mathcal{U}$  such that  $(i, k) \in \mathcal{A}(\bar{\mathbb{G}}_1)$  and  $(k, j) \in \mathcal{A}(\bar{\mathbb{G}}_2)$ . Therefore  $(i, k) \in \mathcal{A}(\mathbb{G}_1)$  and  $(k, j) \in \mathcal{A}(\mathbb{G}_2)$ . Thus  $(i, j) \in \mathcal{A}(\mathbb{G}_2 \circ \mathbb{G}_1)$ . But  $i, j \in \mathcal{S}$ , and so  $(i, j)$  must be an arc in the subgraph of  $\mathbb{G}_2 \circ \mathbb{G}_1$  induced by  $\mathcal{U}$ . Since this clearly is true for all arcs in  $\mathcal{A}(\bar{\mathbb{G}}_2 \circ \bar{\mathbb{G}}_1)$ , the proof is complete.  $\square$

*Proof of Proposition 3.* Let  $\mathbb{G}_1$  and  $\mathbb{G}_2$  be two extended delay graphs in  $\bar{\mathcal{D}}$ . It will first be shown that for each  $i \in \{1, 2, \dots, n\}$ , every neighbor of  $\mathcal{V}_i$  which is not in  $\mathcal{V}_i$  is a neighbor of  $v_{i1}$  in  $\mathbb{G}_2 \circ \mathbb{G}_1$ . Fix  $i \in \{1, 2, \dots, n\}$  and let  $v$  be a neighbor of  $\mathcal{V}_i$  in  $\mathbb{G}_2 \circ \mathbb{G}_1$  which is not in  $\mathcal{V}_i$ . Then  $(v, k) \in \mathcal{A}(\mathbb{G}_2 \circ \mathbb{G}_1)$  for some  $k \in \mathcal{V}_i$ . Thus there is a  $s \in \bar{\mathcal{V}}$  such that  $(v, s) \in \mathcal{A}(\mathbb{G}_1)$  and  $(s, k) \in \mathcal{A}(\mathbb{G}_2)$ . If  $s \notin \mathcal{V}_i$ , then  $(s, v_{i1}) \in \mathcal{A}(\mathbb{G}_2)$  because  $\mathbb{G}_2$  is an extended delay graph. Thus in this case  $(v, v_{i1}) \in \mathcal{A}(\mathbb{G}_2 \circ \mathbb{G}_1)$  because of the definition of composition. If, on the other hand,  $s \in \mathcal{V}_i$ , then  $(v, v_{i1}) \in \mathcal{A}(\mathbb{G}_1)$  because  $\mathbb{G}_1$  is an extended delay graph. Thus in this case  $(v, v_{i1}) \in \mathcal{A}(\mathbb{G}_2 \circ \mathbb{G}_1)$  because  $v_{i1}$  has a self-arc in  $\mathbb{G}_2$ . This proves that every neighbor of  $\mathcal{V}_i$  which is not in  $\mathcal{V}_i$  is a neighbor of  $v_{i1}$  in  $\mathbb{G}_2 \circ \mathbb{G}_1$ . Since this must be true for each  $i \in \{1, 2, \dots, n\}$ ,  $\mathbb{G}_2 \circ \mathbb{G}_1$  has the first property defining extended delay graphs in  $\bar{\mathcal{D}}$ .

To establish the second property, we exploit the fact that the composition of two graphs with the same hierarchy is a graph with the same hierarchy. Thus for any integer  $i \in \{1, 2, \dots, n\}$ , the composition of the subgraphs of  $\mathbb{G}_1$  and  $\mathbb{G}_2$ , respectively, induced by  $\mathcal{V}_i$  must have the hierarchy  $\{v_{i1}, v_{i2}, \dots, v_{im_i}\}$ . But by Lemma 3, for any integer  $i \in \{1, 2, \dots, n\}$ , the composition of the subgraphs of  $\mathbb{G}_1$  and  $\mathbb{G}_2$ , respectively, induced by  $\mathcal{V}_i$  is contained in the subgraph of the composition of  $\mathbb{G}_1$  and  $\mathbb{G}_2$  induced by  $\mathcal{V}_i$ . This implies that for  $i \in \{1, 2, \dots, n\}$ , the subgraph of the composition of  $\mathbb{G}_1$  and  $\mathbb{G}_2$  induced by  $\mathcal{V}_i$  has  $\{v_{i1}, v_{i2}, \dots, v_{im_i}\}$  as a hierarchy.  $\square$

Our main result regarding extended delay graphs is as follows.

PROPOSITION 4. Let  $m$  be the largest integer in the set  $\{m_1, m_2, \dots, m_n\}$ . The composition of any set of at least  $m(n-1)^2 + m - 1$  extended delay graphs will be strongly rooted if the quotient graph of each of the graphs in the composition is rooted.

To prove this proposition we will need several more concepts. Let us agree to say that a extended delay graph  $\mathbb{G} \in \bar{\mathcal{D}}$  has *strongly rooted hierarchies* if for each  $i \in \mathcal{V}$ , the subgraph of  $\mathbb{G}$  induced by  $\mathcal{V}_i$  is strongly rooted. Proposition 2 states that a hierarchical graph on  $m_i$  vertices will be strongly rooted if it is the composition of at least  $m_i - 1$  rooted graphs with the same hierarchy. This and Lemma 3 imply that

the subgraph of the composition of at least  $m_i - 1$  extended delay graphs induced by  $\mathcal{V}_i$  will be strongly rooted. We are led to the following lemma.

LEMMA 4. *Any composition of at least  $m - 1$  extended delay graphs in  $\bar{\mathcal{D}}$  has strongly rooted hierarchies.*

To proceed we will need one more type of graph which is uniquely determined by a given graph in  $\bar{\mathcal{G}}$ . By the *agent subgraph* of  $\mathbb{G} \in \bar{\mathcal{G}}$  we mean the subgraph of  $\mathbb{G}$  induced by  $\mathcal{V}$ . Note that while the quotient graph of  $\mathbb{G}$  describes relations between distinct agent hierarchies, the agent subgraph of  $\mathbb{G}$  captures only the relationships between the roots of the hierarchies. Note in addition that both the agent subgraph of  $\mathbb{G}$  and the quotient graph of  $\mathbb{G}$  are graphs in  $\mathcal{G}_{sa}$  because all  $n$  vertices of  $\mathbb{G}$  in  $\mathcal{V}$  have self-arcs. The agent subgraph of the graph in Figure 1 is shown in Figure 3.



FIG. 3. Agent subgraph.

LEMMA 5. *Let  $\mathbb{G}_p$  and  $\mathbb{G}_q$  be extended delay graphs in  $\bar{\mathcal{D}}$ . If  $\mathbb{G}_p$  has a strongly rooted agent subgraph and  $\mathbb{G}_q$  has strongly rooted hierarchies, then the composition  $\mathbb{G}_q \circ \mathbb{G}_p$  is strongly rooted.*

*Proof of Lemma 5.* Let  $v_{i1}$  be a root of the agent subgraph of  $\mathbb{G}_p$  and let  $v_{jk}$  be any vertex in  $\bar{\mathcal{V}}$ . Then  $(v_{i1}, v_{j1}) \in \mathcal{A}(\mathbb{G}_p)$  because the agent subgraph of  $\mathbb{G}_p$  is strongly rooted. Moreover,  $(v_{j1}, v_{jk}) \in \mathcal{A}(\mathbb{G}_q)$  because  $\mathbb{G}_q$  has strongly rooted hierarchies. Therefore, in view of the definition of graph composition,  $(v_{i1}, v_{jk}) \in \mathcal{A}(\mathbb{G}_q \circ \mathbb{G}_p)$ . Since this must be true for every vertex  $v_{jk} \in \bar{\mathcal{V}}$ ,  $\mathbb{G}_q \circ \mathbb{G}_p$  is strongly rooted.  $\square$

LEMMA 6. *The agent subgraph of any composition of at least  $(n - 1)^2$  extended delay graphs in  $\bar{\mathcal{D}}$  will be strongly rooted if the agent subgraph of each of the graphs in the composition is rooted.*

*Proof of Lemma 6.* Let  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_q$  be any sequence of  $q \geq (n - 1)^2$  extended delay graphs in  $\bar{\mathcal{D}}$  whose agent subgraphs,  $\bar{\mathbb{G}}_i$ ,  $i \in \{1, 2, \dots, q\}$ , are all rooted. Since the  $\bar{\mathbb{G}}_i$  are in  $\mathcal{G}_{sa}$ , Proposition 3 of [6] applies, and it can therefore be concluded that  $\bar{\mathbb{G}}_q \circ \dots \circ \bar{\mathbb{G}}_2 \circ \bar{\mathbb{G}}_1$  is strongly rooted. But  $\bar{\mathbb{G}}_q \circ \dots \circ \bar{\mathbb{G}}_2 \circ \bar{\mathbb{G}}_1$  is contained in the agent subgraph of  $\mathbb{G}_q \circ \dots \circ \mathbb{G}_2 \circ \mathbb{G}_1$  because of Lemma 3. Therefore the agent subgraph of  $\mathbb{G}_q \circ \dots \circ \mathbb{G}_2 \circ \mathbb{G}_1$  is strongly rooted.  $\square$

LEMMA 7. *Let  $\mathbb{G}_p$  and  $\mathbb{G}_q$  be extended delay graphs in  $\bar{\mathcal{D}}$ . If  $\mathbb{G}_p$  has strongly rooted hierarchies and  $\mathbb{G}_q$  has a rooted quotient graph, then the agent subgraph of the composition  $\mathbb{G}_q \circ \mathbb{G}_p$  is rooted.*

*Proof of Lemma 7.* Let  $(i, j)$  be any arc in the quotient graph of  $\mathbb{G}_q$  with  $i \neq j$ . This means that  $(v_{ik}, v_{js}) \in \mathcal{A}(\mathbb{G}_q)$  for some  $v_{ik} \in \mathcal{V}_i$  and  $v_{js} \in \mathcal{V}_j$ . Clearly  $(v_{i1}, v_{ik}) \in \mathcal{A}(\mathbb{G}_p)$  because  $\mathbb{G}_p$  has strongly rooted hierarchies. Moreover, since  $i \neq j$ ,  $v_{ik}$  is a neighbor of  $\mathcal{V}_j$  which is not in  $\mathcal{V}_j$ . From this and the definition of an extended delay graph, it follows that  $v_{ik}$  is a neighbor of  $v_{j1}$ . Therefore  $(v_{ik}, v_{j1}) \in \mathcal{A}(\mathbb{G}_q)$ . Thus  $(v_{i1}, v_{j1}) \in \mathcal{A}(\mathbb{G}_q \circ \mathbb{G}_p)$ . We have therefore proved that for any path of length one between any two distinct vertices  $i, j$  in the quotient graph of  $\mathbb{G}_q$ , there is a corresponding path between vertices  $v_{i1}$  and  $v_{j1}$  in the agent subgraph of  $\mathbb{G}_q \circ \mathbb{G}_p$ . This implies that for any path of any length between any two distinct vertices  $i, j$  in the quotient graph of  $\mathbb{G}_q$ , there is a corresponding path between vertices  $v_{i1}$  and  $v_{j1}$  in the agent subgraph of  $\mathbb{G}_q \circ \mathbb{G}_p$ . Since by assumption the quotient graph of  $\mathbb{G}_q$  is rooted, the agent subgraph of  $\mathbb{G}_q \circ \mathbb{G}_p$  must be rooted as well.  $\square$

*Proof of Proposition 4.* Let  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_s$  be a sequence of at least  $m(n-1)^2 + m - 1$  extended delay graphs with rooted quotient graphs. The graph  $\mathbb{G}_s \circ \dots \circ \mathbb{G}_{(m(n-1)^2+1)}$  is composed of at least  $m - 1$  extended delay graphs. Therefore  $\mathbb{G}_s \circ \dots \circ \mathbb{G}_{(m(n-1)^2+1)}$  must have strongly rooted hierarchies because of Lemma 4. In view of Lemma 5, to complete the proof it is enough to show that  $\mathbb{G}_{m(n-1)^2} \circ \dots \circ \mathbb{G}_1$  has a strongly rooted agent subgraph. But  $\mathbb{G}_{m(n-1)^2} \circ \dots \circ \mathbb{G}_1$  is the composition of  $(n-1)^2$  graphs, each itself a composition of  $m$  extended delay graphs with rooted quotient graphs. In view of Lemma 6, to complete the proof it is enough to show that the agent subgraph of any composition of  $m$  extended delay graphs is rooted if each quotient graph of each extended delay graph in the composition is rooted. Let  $\mathbb{H}_1, \mathbb{H}_2, \dots, \mathbb{H}_m$  be such a family of extended delay graphs. By assumption,  $\mathbb{H}_m$  has a rooted quotient graph. In view of Lemma 7, the agent subgraph of  $\mathbb{H}_m \circ \mathbb{H}_{m-1} \circ \dots \circ \mathbb{H}_1$  will be rooted if  $\mathbb{H}_{m-1} \circ \dots \circ \mathbb{H}_1$  has strongly rooted hierarchies. But  $\mathbb{H}_{m-1} \circ \dots \circ \mathbb{H}_1$  has this property because of Lemma 4.  $\square$

Finally we will need the following fact.

**PROPOSITION 5.** *Let  $\mathbb{G}_1, \dots, \mathbb{G}_r$  be a sequence of extended delay graphs in  $\bar{\mathcal{D}}$ . If the composition  $Q(\mathbb{G}_r) \circ \dots \circ Q(\mathbb{G}_1)$  is rooted, then so is the quotient graph  $Q(\mathbb{G}_r \circ \dots \circ \mathbb{G}_1)$ .*

This proposition is a direct consequence of the following lemma.

**LEMMA 8.** *Let  $\mathbb{G}_p, \mathbb{G}_q$  be two extended delay graphs in  $\bar{\mathcal{D}}$ . For each arc  $(i, j)$  in the composition  $Q(\mathbb{G}_q) \circ Q(\mathbb{G}_p)$ , there is a path from  $i$  to  $j$  in the quotient graph  $Q(\mathbb{G}_q \circ \mathbb{G}_p)$ .*

*Proof of Lemma 8.* Fix  $(i, j) \in \mathcal{A}(Q(\mathbb{G}_q) \circ Q(\mathbb{G}_p))$ . If  $i = j$ , then  $(i, j) \in \mathcal{A}(Q(\mathbb{G}_q \circ \mathbb{G}_p))$  because  $Q(\mathbb{G}_q \circ \mathbb{G}_p) \in \mathcal{G}_{sa}$ . Thus in this case there is a path of length 1 from  $i$  to  $j$  in  $Q(\mathbb{G}_q \circ \mathbb{G}_p)$ .

Suppose  $i \neq j$ . Since  $(i, j) \in \mathcal{A}(Q(\mathbb{G}_q) \circ Q(\mathbb{G}_p))$ , there exists an integer  $k \in \mathcal{V}$  such that  $(i, k) \in \mathcal{A}(Q(\mathbb{G}_p))$  and  $(k, j) \in \mathcal{A}(Q(\mathbb{G}_q))$ . Thus there are integers  $v_{is}, v_{kt}, v_{ku}, v_{jw} \in \mathcal{V}$  such that  $(v_{is}, v_{kt}) \in \mathcal{A}(\mathbb{G}_p)$  and  $(v_{ku}, v_{jw}) \in \mathcal{A}(\mathbb{G}_q)$ . Since  $\mathbb{G}_p \in \bar{\mathcal{D}}$ ,  $\mathbb{G}_p$  has a hierarchy rooted at  $v_{k1}$ . This means that there must be a vertex  $v_{kx}$  no higher in this hierarchy than  $v_{ku}$  such that  $(v_{kx}, v_{ku}) \in \mathcal{A}(\mathbb{G}_p)$ . Therefore  $(v_{kx}, v_{jw}) \in \mathcal{A}(\mathbb{G}_q \circ \mathbb{G}_p)$ . If  $k = i$ , then  $(v_{ix}, v_{jw}) \in \mathcal{A}(\mathbb{G}_q \circ \mathbb{G}_p)$ , and so  $(i, j) \in \mathcal{A}(Q(\mathbb{G}_q \circ \mathbb{G}_p))$ . Thus in this case there is a path of length 1 from  $i$  to  $j$  in  $Q(\mathbb{G}_q \circ \mathbb{G}_p)$ .

Suppose  $k \neq i$ . Since  $(v_{is}, v_{kt}) \in \mathcal{A}(\mathbb{G}_p)$  and  $\mathbb{G}_p \in \bar{\mathcal{D}}$ ,  $(v_{is}, v_{k1}) \in \mathcal{A}(\mathbb{G}_p)$ . But  $\mathbb{G}_q$  must have a self-arc at  $v_{k1}$  because  $\mathbb{G}_q \in \bar{\mathcal{D}}$ . Therefore  $(v_{is}, v_{k1}) \in \mathcal{A}(\mathbb{G}_q \circ \mathbb{G}_p)$ . Moreover, there must be a path in  $\mathbb{G}_q \circ \mathbb{G}_p$  from  $v_{k1}$  to  $v_{kx}$  because  $v_{kx}$  is in the hierarchy rooted at  $v_{k1}$ . But both  $(v_{is}, v_{k1})$  and  $(v_{kx}, v_{jw})$  are arcs in  $\mathbb{G}_q \circ \mathbb{G}_p$ , and so there must be a path in  $\mathbb{G}_q \circ \mathbb{G}_p$  from  $v_{is}$  to  $v_{jw}$ . This implies that there must be a path in  $Q(\mathbb{G}_q \circ \mathbb{G}_p)$  from  $i$  to  $j$ .  $\square$

*Proof of Proposition 5.* To prove the proposition it is enough to show that if  $Q(\mathbb{G}_r) \circ \dots \circ Q(\mathbb{G}_1)$  contains a path from some  $i \in \mathcal{V}$  to some  $j \in \mathcal{V}$ , then  $Q(\mathbb{G}_r \circ \dots \circ \mathbb{G}_1)$  also contains a path from  $i$  to  $j$ . As a first step towards this end, we claim that if  $\mathbb{G}_p, \mathbb{G}_q$  are graphs in  $\bar{\mathcal{D}}$  for which  $Q(\mathbb{G}_q) \circ Q(\mathbb{G}_p)$  contains a path from  $u$  to  $v$ , for some  $u, v \in \mathcal{V}$ , then  $Q(\mathbb{G}_q \circ \mathbb{G}_p)$  also contains a path from  $u$  to  $v$ . To prove that this is so, fix  $u, v \in \mathcal{V}$  and  $\mathbb{G}_p, \mathbb{G}_q \in \bar{\mathcal{D}}$  and suppose that  $Q(\mathbb{G}_q) \circ Q(\mathbb{G}_p)$  contains a path from  $u$  to  $v$ . Then there must be a positive integer  $s$  and vertices  $k_1, k_2, \dots, k_s$  ending at  $k_s = v$  for which  $(u, k_1), (k_1, k_2), \dots, (k_{s-1}, k_s)$  are arcs in  $Q(\mathbb{G}_q) \circ Q(\mathbb{G}_p)$ . In view of Lemma 8, there must be paths in  $Q(\mathbb{G}_q \circ \mathbb{G}_p)$  from  $i$  to  $k_1$ ,  $k_1$  to  $k_2$ ,  $\dots$ , and  $k_{s-1}$  to  $k_s$ . It follows that there must be a path in  $Q(\mathbb{G}_q \circ \mathbb{G}_p)$  from  $i$  to  $j$ . Thus the claim is established.

It will now be shown by induction for each  $s \in \{2, \dots, m\}$  that if  $Q(\mathbb{G}_s) \circ \dots \circ Q(\mathbb{G}_1)$  contains a path from  $i$  to some  $j_s \in \mathcal{V}$ , then  $Q(\mathbb{G}_r \circ \dots \circ \mathbb{G}_1)$  also contains a path from  $i$  to  $j_s$ . In view of the claim just proved above, the assertion is true if  $s = 2$ . Suppose the assertion is true for all  $s \in \{2, 3, \dots, t\}$ , where  $t$  is some integer in  $\{2, \dots, r-1\}$ . Suppose that  $Q(\mathbb{G}_{t+1}) \circ \dots \circ Q(\mathbb{G}_1)$  contains a path from  $i$  to  $j_{t+1}$ . Then there must be an integer  $k$  such that  $Q(\mathbb{G}_t) \circ \dots \circ Q(\mathbb{G}_1)$  contains a path from  $i$  to  $k$  and  $Q(\mathbb{G}_{t+1})$  contains a path from  $k$  to  $j_{t+1}$ . In view of the inductive hypothesis,  $Q(\mathbb{G}_t \circ \dots \circ \mathbb{G}_1)$  contains a path from  $i$  to  $k$ . Therefore  $Q(\mathbb{G}_{t+1}) \circ Q(\mathbb{G}_t \circ \dots \circ \mathbb{G}_1)$  has a path from  $i$  to  $j_{t+1}$ . Hence the claim established at the beginning of this proof applies, and it can be concluded that  $Q(\mathbb{G}_{t+1} \circ \mathbb{G}_t \circ \dots \circ \mathbb{G}_1)$  has a path from  $i$  to  $j_{t+1}$ . Therefore by induction the aforementioned assertion is true.  $\square$

**4.4. Proof of convergence.** Our aim is to make use of the properties of extended delay graphs just derived to prove Theorem 2. We will also need the following result from [6].

**PROPOSITION 6.** *Let  $\mathcal{S}_{sr}$  be any closed set of stochastic matrices which are all of the same size and whose graphs  $\gamma(S)$ ,  $S \in \mathcal{S}_{sr}$ , are all strongly rooted. As  $j \rightarrow \infty$ , any product  $S_j \cdots S_1$  of matrices from  $\mathcal{S}_{sr}$  converges exponentially fast to a matrix of the form  $\mathbf{1}c$  at a rate no slower than  $\lambda$ , where  $c$  is a nonnegative row vector depending on the sequence and  $\lambda$  is a nonnegative constant less than 1 depending only on  $\mathcal{S}_{sr}$ .*

*Proof of Theorem 2.* In view of (31),  $\theta(t) = F(t-1) \cdots F(0)\theta(0)$ . Thus to prove the theorem it suffices to prove that as  $t \rightarrow \infty$  the matrix product  $F(t) \cdots F(0)$  converges exponentially fast to a matrix of the form  $\mathbf{1}c$ .

By hypothesis, the sequence of neighbor graphs  $\mathbb{N}(0), \mathbb{N}(1), \dots$  is repeatedly jointly rooted by subsequences of length  $q$ . This means that each of the sequences  $\mathbb{N}(kq), \dots, \mathbb{N}((k+1)q-1)$ ,  $k \geq 0$ , is jointly rooted. Let  $\mathbb{D}(t) = \gamma(F(t))$ ,  $t \geq 0$ . In view of (29),  $\mathbb{N}(t) = Q(\mathbb{D}(t))$ ,  $t \geq 0$ . Thus each of the sequences  $Q(\mathbb{D}(kq)), \dots, Q(\mathbb{D}((k+1)q-1))$ ,  $k \geq 0$ , is jointly rooted, and so each composition  $Q(\mathbb{D}((k+1)q-1)) \circ \dots \circ Q(\mathbb{D}(kq))$  is a rooted graph. In view of Proposition 5, each graph  $Q(\mathbb{D}((k+1)q-1)) \circ \dots \circ \mathbb{D}(kq)$ ,  $k \geq 0$ , is also rooted.

Set  $p = (m(n-1)^2 + m - 1)q$ , where  $m$  is the largest integer in the set  $\{m_1, m_2, \dots, m_n\}$ . In view of Proposition 4, each of the graphs  $\mathbb{D}((k+1)p-1) \circ \dots \circ \mathbb{D}(kp)$ ,  $k \geq 0$ , is strongly rooted. Let  $\mathcal{F}(p)$  denote the set of all products of  $p$  matrices from  $\bar{\mathcal{F}}$  which have the additional property that each such product has a strongly rooted graph. Then  $\mathcal{F}(p)$  is finite and therefore compact, because  $\bar{\mathcal{F}}$  is.

For  $k \geq 0$ , define

$$(32) \quad S(k) = F((k+1)p-1) \cdots F(kp).$$

In view of (12) and the fact that  $\gamma(F(t)) = \mathbb{D}(t)$ ,  $t \geq 0$ , it must be true that  $\gamma(S(k)) = \mathbb{D}((k+1)p-1) \circ \dots \circ \mathbb{D}(kp)$ ,  $k \geq 0$ . Thus each  $S(k)$  has a strongly rooted graph. Moreover, each such  $S(k)$  is the product of  $p$  matrices from  $\bar{\mathcal{F}}$ . Therefore  $S(k) \in \mathcal{F}(p)$ ,  $k \geq 0$ . Therefore Proposition 6 applies with  $\mathcal{S}_{sr} = \mathcal{F}(p)$ , and so it can be concluded that the matrix product  $S(k) \cdots S(0)$  converges exponentially fast as  $k \rightarrow \infty$  to a matrix of the form  $\mathbf{1}c$  as  $k \rightarrow \infty$ .

In view of the definition of  $S(k)$  it is clear that for any  $t$ , there is an integer  $k(t)$  and a stochastic matrix  $\hat{S}(t)$  composed of the product of at most  $p-1$  matrices from  $\bar{\mathcal{F}}$  such that

$$F(t) \cdots F(1) = \hat{S}(t)S(k(t)) \cdots S(0).$$



Moreover,  $t \mapsto k(t)$  must be an unbounded, strictly increasing function; because of this the product  $S(k(t)) \cdots S(0)$  must converge exponentially fast as  $t \rightarrow \infty$  to a limit of the form  $\mathbf{1}c$ . Since  $\widehat{S}(t)\mathbf{1}c = \mathbf{1}c$ ,  $t \geq 0$ , the product  $F(t) \cdots F(1)$  must also converge exponentially fast as  $t \rightarrow \infty$  to the same limit  $\mathbf{1}c$ .  $\square$

**5. Asynchronous flocking.** In this section we consider a modified version of the consensus problem treated in [6] in which each agent independently updates its heading at times determined by its own clock.<sup>2</sup> We do not assume that the groups' clocks are synchronized or that the times any one agent updates its heading are evenly spaced. Updating of agent  $i$ 's heading is done as follows. At its  $k$ th *sensing event time*  $t_{ik}$ , agent  $i$  senses the headings  $\theta_j(t_{ik})$ ,  $j \in \mathcal{N}_i(t_{ik})$ , of its current neighbors (which includes itself) and from this data computes its  $k$ th "way-point"  $w_i(t_{ik})$ . In what follows we will consider way-points based on averaging. In particular, agent  $i$ 's  $k$ th *way-point* is defined by the rule

$$(33) \quad w_i(t_{ik}) = \frac{1}{n_i(t_{ik})} \left( \sum_{j \in \mathcal{N}_i(t_{ik})} \theta_j(t_{ik}) \right), \quad i \in \{1, 2, \dots, n\},$$

where  $n_i(t_{ik})$  is the number of neighbor elements in the neighbor index set  $\mathcal{N}_i(t_{ik})$ . After computing  $w_i(t_{ik})$ , agent  $i$  changes its heading from  $\theta_i(t_{ik})$  to  $w_i(t_{ik})$  on the interval  $(t_{ik}, t_{i(k+1)}]$ . In this paper we will consider the case when each agent updates its heading instantaneously at its own event times and holds its heading fixed between event times. More precisely, we will assume that agent  $i$ 's heading  $\theta_i(t)$  takes on its agent  $i$ 's  $k$ th way-point value  $w_i(t_{ik})$  immediately after its  $k$ th event time  $t_{ik}$  and that  $\theta_i(t)$  is constant on each continuous-time interval  $(t_{i(k-1)}, t_{ik}]$ ,  $k \geq 1$ , where  $t_{i0} = 0$  is agent  $i$ 's zeroth event time. In other words for  $k \geq 0$ , agent  $i$ 's heading satisfies

$$(34) \quad \theta_i(t_{i(k+1)}) = \frac{1}{n_i(t_{ik})} \left( \sum_{j \in \mathcal{N}_i(t_{ik})} \theta_j(t_{ik}) \right),$$

$$(35) \quad \theta_i(t) = \theta_i(t_{ik}), \quad t_{i(k-1)} < t \leq t_{ik}.$$

**5.1. Analytic synchronization.** To develop conditions under which all agents eventually move with the same heading requires the analysis of the asymptotic behavior of the *asynchronous* process which the  $2n$  heading equations of the form (34), (35) define. Despite the apparent complexity of this process, it is possible to capture its salient features using a suitably defined *synchronous* discrete-time, hybrid dynamical system  $\mathbb{S}$ . The sequence of steps involved in defining  $\mathbb{S}$  has been discussed before and is called *analytic synchronization* [12, 13]. Analytic synchronization is applicable to any finite family of continuous or discrete-time dynamical processes  $\{\mathbb{P}_1, \mathbb{P}_2, \dots, \mathbb{P}_n\}$  under the following conditions. First, each process  $\mathbb{P}_i$  must be a dynamical system whose inputs consist of functions of the states of the other processes as well as signals which are exogenous to the entire family. Second, each process  $\mathbb{P}_i$  must have associated with it an ordered sequence of event times  $\{t_{i1}, t_{i2}, \dots\}$  defined in such a way so that the state of  $\mathbb{P}_i$  at event time  $t_{i(k+1)}$  is uniquely determined by values of the exogenous signals and states of the  $\mathbb{P}_j$ ,  $j \in \{1, 2, \dots, n\}$ , at event times  $t_{jkj}$  which occur

<sup>2</sup>A preliminary version of the material in this section was presented at the 2005 IFAC congress [4].

prior to  $t_{i(k_i+1)}$  but in the finite past. Event time sequences for different processes need not be synchronized. Analytic synchronization is a procedure for creating a single synchronous process for purposes of analysis which captures the salient features of the original  $n$  asynchronously functioning processes. As a first step, all  $n$  event time sequences are merged into a single ordered sequence of event times  $\mathcal{T}$ . (This clever idea has been used before in [2] to study the convergence of totally asynchronous iterative algorithms.) The “synchronized” state of  $\mathbb{P}_i$  is then defined to be the original state of  $\mathbb{P}_i$  at  $\mathbb{P}_i$ ’s event times  $\{t_{i1}, t_{i2}, \dots\}$  plus possibly some additional variables; at values of  $t \in \mathcal{T}$  between event times  $t_{ik_i}$  and  $t_{i(k_i+1)}$ , the synchronized state of  $\mathbb{P}_i$  is taken to be the same as the value of its original state at time  $t_{ik}$ . Although it is not always possible to carry out all of these steps, when it is what ultimately results is a synchronous dynamical system  $\mathbb{S}$  evolving on the index set of  $\mathcal{T}$ , with the state composed of the synchronized states of the  $n$  individual processes under consideration. We now use these ideas to develop such a synchronous system  $\mathbb{S}$  for the asynchronous process under consideration.

**5.2. Definition of  $\mathbb{S}$ .** As a first step, let  $\mathcal{T}$  denote the set of all event times of all  $n$  agents. Relabel the elements of  $\mathcal{T}$  as  $t_0, t_1, t_2, \dots$  in such a way so that  $t_j < t_{j+1}$ ,  $j \in \{1, 2, \dots\}$ . Next define

$$(36) \quad \bar{\theta}_i(\tau) = \theta_i(t_\tau), \quad \tau \geq 0, \quad i \in \{1, 2, \dots, n\}.$$

In view of (34), it must be true that if  $t_\tau$  is an event time of agent  $i$ , then

$$\bar{\theta}_i(\tau') = \frac{1}{\bar{n}_i(\tau)} \left( \sum_{j \in \bar{\mathcal{N}}_i(\tau)} \bar{\theta}_j(\tau) \right),$$

where  $\bar{\mathcal{N}}_i(\tau) = \mathcal{N}_i(t_\tau)$ ,  $\bar{n}_i(\tau) = n_i(t_\tau)$ , and  $t_{\tau'}$  is the next event time of agent  $i$  after  $t_\tau$ . But  $\bar{\theta}_i(\tau') = \bar{\theta}_i(\tau + 1)$  because  $\theta_i(t)$  is constant for  $t_\tau < t \leq t_{\tau'}$  (approximately (35)). Therefore

$$(37) \quad \bar{\theta}_i(\tau + 1) = \frac{1}{\bar{n}_i(\tau)} \left( \sum_{j \in \bar{\mathcal{N}}_i(\tau)} \bar{\theta}_j(\tau) \right)$$

if  $t_\tau$  is an event time of agent  $i$ . Meanwhile if  $t_\tau$  is not an event time of agent  $i$ , then

$$(38) \quad \bar{\theta}_i(\tau + 1) = \bar{\theta}_i(\tau),$$

again because  $\theta_i(t)$  is constant between event times. Note that if we define  $\bar{\mathcal{N}}_i(\tau) = \{i\}$  and  $\bar{n}_i(\tau) = 1$  for every value of  $\tau$  for which  $t_\tau$  is not an event time of agent  $i$ , then (38) can be written as

$$(39) \quad \bar{\theta}_i(\tau + 1) = \frac{1}{\bar{n}_i(\tau)} \left( \sum_{j \in \bar{\mathcal{N}}_i(\tau)} \bar{\theta}_j(\tau) \right).$$

Doing this enables us to combine (37) and (39) into a single formula valid for all  $\tau \geq 0$ . In other words, agent  $i$ ’s heading satisfies

$$(40) \quad \bar{\theta}_i(\tau + 1) = \frac{1}{\bar{n}_i(\tau)} \left( \sum_{j \in \bar{\mathcal{N}}_i(\tau)} \bar{\theta}_j(\tau) \right), \quad \tau \geq 0,$$

where

$$(41) \quad \bar{\mathcal{N}}_i(\tau) = \begin{cases} \mathcal{N}_i(t_\tau) & \text{if } t_\tau \text{ is an event time of agent } i \\ \{i\} & \text{if } t_\tau \text{ is not an event time of agent } i \end{cases}$$

and  $\bar{n}_i(\tau) = 1$  if  $t_\tau$  is not an event time of agent  $i$ . Thus for all  $\tau$ ,  $\bar{n}_i(\tau)$  is the number of indices in  $\bar{\mathcal{N}}_i(\tau)$ . For purposes of analysis, it is useful to interpret (41) as meaning that between agent  $i$ 's event times, its only neighbor is itself. There are  $n$  equations of the form in (40), and together they define a synchronous system  $\mathbb{S}$  which models the evolutions of the  $n$  agents' headings at event times.

**5.3. State space model.** As before, we can represent the neighbor relationships associated with (41) using a directed graph  $\mathbb{N}$  with vertex set  $\mathcal{V} = \{1, 2, \dots, n\}$  and arc set  $\mathcal{A}(\mathbb{N}) \subset \mathcal{V} \times \mathcal{V}$  which is defined in such a way so that  $(i, j)$  is an arc from  $i$  to  $j$  just in case agent  $i$  is a neighbor of agent  $j$ . Thus as before,  $\mathbb{N}$  is a directed graph on  $n$  vertices with at most one arc from any vertex to another and with exactly one self-arc at each vertex. We continue to write  $\mathcal{G}_{sa}$  for the set of all such graphs.

For each graph  $\mathbb{N} \in \mathcal{G}_{sa}$  let  $F = D^{-1}A'$ , where  $A'$  is the transpose of the adjacency matrix of  $\mathbb{N}$  and  $D$  the diagonal matrix whose  $j$ th diagonal element is the in-degree of vertex  $j$  within the graph. The set of agent heading update rules defined by (41) can be written in state form as

$$(42) \quad \bar{\theta}(\tau + 1) = F(\tau)\bar{\theta}(\tau), \quad \tau \in \{0, 1, 2, \dots\},$$

where  $\bar{\theta}$  is the heading vector  $\bar{\theta} = [\bar{\theta}_1 \quad \bar{\theta}_2 \quad \dots \quad \bar{\theta}_n]'$ , and  $F(\tau)$  is the flocking matrix determined by neighbor graph  $\mathbb{N}(\tau)$  at event time  $t_\tau$ .

Up to this point the development is essentially the same as in the leaderless consensus problem discussed in section 2. But when one considers the type of graphs in  $\mathcal{G}_{sa}$  which are likely to be encountered along a given trajectory, things are quite different. Note, for example, that the only vertices of  $\mathbb{N}(\tau)$  which can have more than one incoming arc are those of agents for whom  $t_\tau$  is an event time. Thus in the most likely situation when distinct agents have only distinct event times, there will be at most one vertex in each graph  $\mathbb{N}(\tau)$  which has more than one incoming arc. It is this situation we want to explore further. Towards this end, let  $\mathcal{G}_{sa}^* \subset \mathcal{G}_{sa}$  denote the subclass of all graphs which have at most one vertex with more than one incoming arc. Note that for  $n > 2$ , there is no rooted graph in  $\mathcal{G}_{sa}^*$ . Nonetheless, in light of Theorem 1 it is clear that convergence to a common steady state heading will occur if the infinite sequence of graphs  $\mathbb{N}(0), \mathbb{N}(1), \dots$  is repeatedly jointly rooted. This of course would require that there exist a jointly rooted sequence of graphs from  $\mathcal{G}_{sa}^*$ . We will now explain why such sequences do in fact exist.

Let us agree to call a graph  $\mathbb{G} \in \mathcal{G}_{sa}$  an *all neighbor graph centered at  $v$*  if every vertex of  $\mathbb{G}$  is a neighbor of  $v$ . Note that every all neighbor graph in  $\mathcal{G}_{sa}$  is also in  $\mathcal{G}_{sa}^*$ . Note also that all neighbor graphs are maximal in  $\mathcal{G}_{sa}^*$  with respect to the partial ordering of  $\mathcal{G}_{sa}^*$  by inclusion. Note also the composition of any all neighbor graph with itself is itself. On the other hand, because the union of two graphs in  $\mathcal{G}_{sa}$  is always contained in the composition of the two graphs, the composition of  $n$  all neighbor graphs with distinct centers must be a graph in which each vertex is a neighbor of every other, i.e., the complete graph. Thus the composition of  $n$  all neighbor graphs with distinct centers is strongly rooted. In summary, the hypothesis of Theorem 1 is not vacuous for the asynchronous problem under consideration. When that hypothesis is satisfied, convergence to a common steady state heading will occur.

**6. Leader following.** In this section we consider a modified version of the flocking problem for the same group of  $n$  agents as before but now with one of the group's members (say agent 1) acting as the group's *leader* [11, 8]. The remaining agents, henceforth called *followers* and labelled 2 through  $n$ , do not know who the leader is or even if there is a leader. Accordingly they continue to use the same heading update rule (1) as before. The leader, on the other hand, acting on its own, ignores update rule (1) and moves with a constant heading  $\theta_1(0)$ . Thus

$$(43) \quad \theta_1(t+1) = \theta_1(t).$$

The situation just described can be modelled as a state space system

$$(44) \quad \theta(t+1) = F(t)\theta(t), \quad t \geq 0,$$

just as before, except now agent 1 is constrained to have no neighbors other than itself. The neighbor graphs  $\mathbb{N}$  which model neighbor relations accordingly all have a distinguished *leader vertex* which has no incoming arcs other than its own.

Much like before, our goal here is to show for a large class of switching signals and for any initial set of follower agent headings that the headings of all  $n$  followers converge to the heading of the leader. Convergence in the leaderless case under the most general conditions required the sequence of neighbor graphs  $\mathbb{N}(0), \mathbb{N}(1), \dots$  encountered along a trajectory to be repeatedly jointly rooted. For the leader-follower case now under consideration, what is required is exactly the same. However, since the leader vertex has only one incoming arc which is a self-arc, the only way  $\mathbb{N}(0), \mathbb{N}(1), \dots$  can be repeatedly jointly rooted is that the sequence be “rooted at the leader vertex  $v = 1$ .” More precisely, an infinite sequence of graphs  $\mathbb{G}_1, \mathbb{G}_2$  in  $\mathcal{G}_{sa}$  is *repeatedly jointly rooted at  $v$*  if there is a positive integer  $m$  for which each finite sequence  $\mathbb{G}_{m(k-1)+1}, \dots, \mathbb{G}_{mk}, k \geq 1$ , is “jointly rooted at  $v$ ”; a finite sequence of directed graphs  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_k$  is *jointly rooted at  $v$*  if the composition  $\mathbb{G}_k \circ \mathbb{G}_{k-1} \circ \dots \circ \mathbb{G}_1$  is rooted at  $v$ . Our main result on discrete-time leader following is next.

**THEOREM 3.** *Let  $\theta(0)$  be fixed. For any trajectory of the system determined by (1) along which the sequence of neighbor graphs  $\mathbb{N}(0), \mathbb{N}(1), \dots$  is repeatedly jointly rooted at vertex 1, there is a constant  $\theta_{ss}$ , depending only on  $\theta(0)$ , for which*

$$\lim_{t \rightarrow \infty} \theta(t) = \theta_1(0)\mathbf{1},$$

where the limit is approached exponentially fast.

*Proof of Theorem 3.* Since any sequence which is repeatedly jointly rooted at  $v$  is repeatedly jointly rooted, Theorem 1 is applicable. Therefore the headings of all  $n$  agents converge exponentially fast to a single common steady state heading  $\theta_{ss}$ . But since the heading of the leader is fixed,  $\theta_{ss}$  must be the leader's heading  $\theta_1(0)$ .  $\square$

**7. Concluding remarks.** The main goal of this paper has been to study various versions the flocking problem considered in [14, 16, 3, 11, 1] and elsewhere from a single point of view which emphasizes the underlying graphical structures for which consensus can be reached. The paper brings together in one place a number of results scattered throughout the literature and at the same time presents new results concerned with convergence rates, asynchronous operation, sensing delays, and graphical interpretations of several specially structured stochastic matrices appropriate to nonhomogeneous Markov chains.

The approach taken in this paper to analyze consensus in the face of measurement delays first goes through a lifting process and then focuses on the resulting state space

model. As we have explained, the lifting process determines stochastic matrices whose graphs do not have self-arcs at all vertices. Nonetheless the graphs which result, namely delay graphs, have special structure, which we have exploited. One is able to associate with each such graph two special graphs, namely a quotient graph and an agent subgraph. These graphs play roles in the analysis of the consensus problem with delays which are similar to the roles played by corresponding quotient graphs and the “injected subgraph” used in the analysis of the asynchronous flocking problem treated in [5]. Although the corresponding graphs which arise in the two problems are completely different, there does seem to be a general pattern of use appropriate to both problems. This suggests that quotient graphs and graphs similar to injected graphs or agent subgraphs may be useful in analyzing other problems as well.

## REFERENCES

- [1] D. ANGELI AND P. A. BLIMAN, *Extension of a result by Moreau on stability of leaderless multi-agent systems*, in Proceedings of the 2005 IEEE CDC, 2005, pp. 759–764.
- [2] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [3] V. D. BLONDEL, J. M. HENDRICHX, A. OLSHEVSKY, AND J. N. TSITSIKLIS, *Convergence in multiagent coordination, consensus, and flocking*, in Proceedings of the 2005 IEEE CDC, 2005, pp. 2996–3000.
- [4] M. CAO, A. S. MORSE, AND B. D. O. ANDERSON, *Coordination of an asynchronous multi-agent system via averaging*, in Proceedings of the 2005 IFAC Congress, 2005.
- [5] M. CAO, A. S. MORSE, AND B. D. O. ANDERSON, *Agreeing asynchronously*, IEEE Trans. Automat. Control, to appear.
- [6] M. CAO, A. S. MORSE, AND B. D. O. ANDERSON, *Reaching a consensus in a dynamically changing environment: A graphical approach*, SIAM J. Control Optim., 47 (2008), pp. 575–600.
- [7] M. CAO, D. A. SPIELMAN, AND A. S. MORSE, *A lower bound on convergence of a distributed network consensus algorithm*, in Proceedings of the 2005 IEEE CDC, 2005, pp. 2356–2361.
- [8] L. GAO AND D. CHENG, *Comment on: “Coordination of groups of mobile autonomous agents using nearest neighbor rules,”* IEEE Trans. Automat. Control, 50 (2005), pp. 1913–1916.
- [9] D. J. HARTFIEL, *Nonhomogeneous Matrix Products*, World Scientific, Singapore, 2002.
- [10] R. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.
- [11] A. JADBABAIE, J. LIN, AND A. S. MORSE, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.
- [12] J. LIN, A. S. MORSE, AND B. D. O. ANDERSON, *The multi-agent rendezvous problem—the asynchronous case*, in Proceedings of the 2004 IEEE CDC, 2004, pp. 1926–1931.
- [13] J. LIN, A. S. MORSE, AND B. D. O. ANDERSON, *The multi-agent rendezvous problem. Part 2: The asynchronous case*, SIAM J. Control Optim., 46 (2007), pp. 2120–2147.
- [14] L. MOREAU, *Stability of multi-agent systems with time-dependent communication links*, IEEE Trans. Automat. Control, 50 (2005), pp. 169–182.
- [15] A. S. MORSE, *Logically switched dynamical systems*, in Nonlinear and Optimal Control Theory, Springer-Verlag, Berlin, 2008, pp. 1–84.
- [16] W. REN AND R. BEARD, *Consensus seeking in multiagent systems under dynamically changing interaction topologies*, IEEE Trans. Automat. Control, 50 (2005), pp. 655–661.
- [17] E. SENETA, *Non-negative Matrices and Markov Chains*, Springer-Verlag, New York, 1981.
- [18] J. N. TSITSIKLIS, *Problems in Decentralized Decision Making and Computation*, Ph.D. thesis, MIT, Cambridge, MA, 1984.
- [19] J. N. TSITSIKLIS, D. P. BERTSEKAS, AND M. ATHANS, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, 31 (1986), pp. 803–812.
- [20] T. VICSEK, A. CZIRÓK, E. BEN-JACOB, I. COHEN, AND O. SHOCHET, *Novel type of phase transition in a system of self-driven particles*, Phys. Rev. Lett., 75 (1995), pp. 1226–1229.

## MIXING ENHANCEMENT BY OPTIMAL FLOW ADVECTION\*

WEIJIU LIU†

**Abstract.** We consider the problem of optimal mixing control. Our objective is to best enhance mixing by flow advection while the flow is optimized in the sense that it is almost steady and is of the least magnitude and the least rotation. For this we define a mixing efficiency functional by penalizing the average of variance of a diffusive scalar, the average of the flow, and the average of its acceleration and strain tensor. By variational principles, we prove the existence of an optimal flow and derive optimality conditions that consist of a system of nonlinear advection-diffusion equations, wave equations, and Laplace's equation.

**Key words.** mixing enhancement, optimal advection, optimal mixing control, advection-diffusion equation

**AMS subject classifications.** 76F25, 49J20

**DOI.** 10.1137/050647888

**1. Introduction.** A fluid mixture consists of diffusive physical quantities and a fluid in which the physical quantities are immersed. Typical examples of such a mixture include fuel and air in a combustor and chemical pollutants and water in the environment. These physical quantities can be mathematically regarded as scalars. If a scalar such as the fuel does not significantly influence the fluid motion, it is called a passive scalar. If chemical reactions can be neglected, then the scalar usually undergoes two processes: molecular diffusion and flow advection. These two processes can be mathematically modeled by the advection-diffusion equation

$$(1) \quad \frac{\partial c}{\partial t} + (\mathbf{v} \cdot \nabla)c = \kappa \nabla^2 c, \quad c(\mathbf{x}, 0) = c^0(\mathbf{x}) \quad \text{in } \Omega, \quad \text{and} \quad \frac{\partial c}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega$$

in the absence of a source or sink. In the above equation,  $c = c(\mathbf{x}, t)$  denotes the concentration of the scalar,  $c^0(\mathbf{x})$  is an initial concentration,  $\kappa > 0$  denotes the molecular diffusivity of the scalar,  $\Omega$  is a bounded domain in  $\mathbb{R}^n$ ,  $\frac{\partial}{\partial \mathbf{n}}$  denotes the normal derivative along the boundary  $\partial\Omega$ ,  $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$  denotes an incompressible velocity field ( $\nabla \cdot \mathbf{v} = 0$ ),  $\nabla = (\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n})$ , and  $\nabla^2 = \frac{\partial^2}{\partial x_1^2} + \dots + \frac{\partial^2}{\partial x_n^2}$ . We assume that  $\mathbf{v}$  satisfies no-penetration boundary conditions on the boundary  $\partial\Omega$  ( $\mathbf{n} \cdot \mathbf{v} = 0$  with  $\mathbf{n}$  denoting the unit normal on the boundary).

Often a certain level of homogeneity of a mixture is desired. For instance, before fuel is burned in a combustor, it is required to be well mixed so that the combustor can achieve its best efficiency. Hence, it is important to design efficient and practical mixing enhancement techniques.

Because a turbulent flow can greatly enhance mixing [4, 10, 11, 12, 13, 15, 22, 26], a useful mixing enhancement technique is to destabilize a flow so that it becomes as turbulent or chaotic as possible. In the design of a stainless cylindrical microcombustor,

---

\*Received by the editors December 17, 2005; accepted for publication (in revised form) August 21, 2007; published electronically February 6, 2008. The preliminary result of this paper appeared in *Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, CA, 2006, pp. 5323–5328. This work was supported by the start-up fund and the University Research Council Fund of the University of Central Arkansas.

<http://www.siam.org/journals/sicon/47-2/64788.html>

†Department of Mathematics, University of Central Arkansas, 201 Donaghey Avenue, Conway, AR 72035 (weijiul@uca.edu).

a critical component for micropower systems using hydrogen and hydrocarbon fuels as an energy source, Yang et al. [28] used the backward facing step to provide a simple yet effective solution to enhance the mixing of the fuel mixture, prolong the residence time, control the position of the flame, and widen the operational range of the flow rate and  $H_2$ /air ratio. Charyulu et al. [7] studied mixing enhancement with two-dimensional (2D) lobed nozzles in a dual stream supersonic flow facility, and their results indicated an enormous enhancement in mixing when a 2D lobed nozzle was employed in comparison with a conventional plain 2D nozzle. The enhanced mixing performance could be attributed to the large-scale axial vortices observed in the flow field. In addition to the use of these passive control devices, a flow can be destabilized by open-loop active excitations through flaps, wall-jets, or other devices [14] so that the flow is separated and large-scale coherent structures are developed in the flow. Active feedback controllers were developed by Aamo, Krstic, and Bewley [1] for destabilization of 2D channel flows; by Balogh, Aamo, and Krstic [5] for destabilization of three-dimensional (3D) pipe flows; by Yuan, Krstic, and Bewley [29] for destabilization of jet nozzle flows; and by Wang et al. [27] for generation of flow separation in bluff body shear flows.

In these control designs, the optimization of control efforts is ignored. While we try to enhance mixing by destabilizing a flow, it is desirable to minimize this destabilization effort. For instance, after mixing has been enhanced, the destabilization should be stopped to save the control efforts. The goal of this paper is to characterize a flow that best enhances mixing and is optimized in some sense.

Mixing will be best enhanced if the scalar variance  $\|c(t; \mathbf{v}) - \langle c(t; \mathbf{v}) \rangle\|$  is made as small as possible [19], where  $c(\mathbf{x}, t; \mathbf{v})$  is the solution of (1) corresponding to the velocity  $\mathbf{v}$  and  $\langle c(t; \mathbf{v}) \rangle$  denotes the mean concentration. As for the control efforts, we could say that the flow velocity  $\mathbf{v}$  is optimal if the flow is almost steady and irrotational. This implies that the flow, its acceleration, and its strain tensor need to be minimized. Therefore we define a mixing efficiency functional by penalizing the average of variance of the scalar, the average of the flow velocity, the average of the strain tensor, and the average of the acceleration. We show that the functional is weakly lower semicontinuous and then it attains its minimum. The minimizer of the functional is called an optimal flow. By variational principles, we then derive optimality conditions that consist of a system of nonlinear partial differential equations.

There are different measures for mixing efficiency such as Lagrangian and Eulerian time averages of a flow [3], the mixing variance coefficient [6], and the Mix-Norm defined by Mathew, Mezić, and Petzold [20]. For the convenience of treatment of our optimal control problem, we use the  $L_2$  norm of a scalar variance as the mixing efficiency measurement.

The optimal mixing problem has been studied in the literature. Using the entropy of automorphisms of dynamical systems as the measure of mixing efficiency, D'Alessandro, Dahleh, and Mezic [2] formulated an optimal mixing problem by maximizing the entropy among all permissible periodic sequences composed of two shear flows orthogonal to each other. They derived the form of the protocol which maximizes the entropy by developing appropriate ergodic-theoretic tools. Another optimal mixing problem was defined by Noack et al. [21], who used the flux across a recirculation region as the measure of mixing efficiency and then maximized the flux among all permissible controlled vortex motions. These optimal mixing problems are different from the one discussed here. First, in our case, the advection-diffusion equation is used to describe the fluid mixing, while a system of ordinary differential equations was used in their studies. Second, the measures of mixing efficiency are different. Third,

the optimal objectives are different. While mixing and flow are both optimized in our case, the optimization of flow was not considered in their cases. Finally, because of these differences, the characterizations of the optimal flow are different. In our case, the optimal flow is characterized by a system of nonlinear partial differential equations, while, in their cases, the optimal sequence of flow is given by the sequence of period 2 of two matrices in [2], and the optimal vortex motion was identified in [21] by finding the optimal flat output trajectory which maximizes the flux.

The paper is organized as follows. We define a mixing efficiency functional in section 2 and prove the existence of an optimal flow in section 3. Optimality conditions are presented in section 4 and are proved in sections 5 and 6.

**2. Mixing efficiency functionals.** Throughout this paper,  $H^s(\Omega)$  denotes the usual Sobolev space [9] for any  $s \in \mathbb{R}$ . For  $s \geq 0$ ,  $H_0^s(\Omega)$  denotes the completion of  $C_0^\infty(\Omega)$  in  $H^s(\Omega)$ , where  $C_0^\infty(\Omega)$  denotes the space of all infinitely differentiable functions on  $\Omega$  with a compact support in  $\Omega$ .

We will need the following vector function spaces:

$$\begin{aligned}\mathbf{L}^2(\Omega) &= \{L^2(\Omega)\}^n, \\ \mathbf{H}^1(\Omega) &= \{H^1(\Omega)\}^n, \\ \mathbf{H}^2(\Omega) &= \{H^2(\Omega)\}^n, \\ \mathbf{H}_{div}^1(\Omega) &= \{\mathbf{v} \in \mathbf{H}^1(\Omega) : \operatorname{div}(\mathbf{v}) = 0 \text{ in } \Omega\}, \\ \mathbf{L}_{div}^2(\Omega) &= \text{the closure of } \mathbf{H}_{div}^1(\Omega) \text{ in } \mathbf{L}^2(\Omega).\end{aligned}$$

The  $\mathbf{L}^2$  norm of a function  $\mathbf{f}(\mathbf{x}) \in \mathbf{L}^2(\Omega)$  is denoted by

$$\|\mathbf{f}\| = \left( \int_{\Omega} |\mathbf{f}(\mathbf{x})|^2 dV \right)^{1/2}.$$

We will also need spaces involving time. Let  $X$  denote a Banach space with a norm  $\|\cdot\|$  and  $0 < T$ . The space  $L^2(0, T; X)$  consists of all measurable functions  $\mathbf{v} : [0, T] \rightarrow X$  with

$$\|\mathbf{v}\|_{L^2(0, T; X)} = \left( \int_0^T \|\mathbf{v}(t)\|^2 dt \right)^{1/2} < \infty.$$

The Sobolev space  $H^1(0, T; X)$  consists of all functions  $\mathbf{v} \in L^2(0, T; X)$  such that  $\mathbf{v}'$  exists in the weak sense and belongs to  $L^2(0, T; X)$ . The norm is defined by

$$\|\mathbf{v}\|_{H^1(0, T; X)} = \left( \int_0^T (\|\mathbf{v}(t)\|^2 + \|\mathbf{v}'(t)\|^2) dt \right)^{1/2}.$$

We denote

$$H_0^1(0, T; X) = \{\mathbf{v} \in H^1(0, T; X) \mid \mathbf{v}(0) = \mathbf{v}(T) = 0\}.$$

The space  $C([0, T]; X)$  consists of all continuous functions  $\mathbf{v} : [0, T] \rightarrow X$  with

$$\|\mathbf{v}\|_{C([0, T]; X)} = \max_{0 \leq t \leq T} \|\mathbf{v}(t)\| < \infty.$$



The strain tensor of the velocity  $\mathbf{v} = (v_1, v_2, v_3)$  is denoted by

$$\nabla \mathbf{v} = \begin{pmatrix} \frac{\partial v_1}{\partial x_1} & \frac{\partial v_1}{\partial x_2} & \frac{\partial v_1}{\partial x_3} \\ \frac{\partial v_2}{\partial x_1} & \frac{\partial v_2}{\partial x_2} & \frac{\partial v_2}{\partial x_3} \\ \frac{\partial v_3}{\partial x_1} & \frac{\partial v_3}{\partial x_2} & \frac{\partial v_3}{\partial x_3} \end{pmatrix}.$$

The mean concentration of  $c(\mathbf{x}, t; \mathbf{v})$  is defined by

$$\langle c(t; \mathbf{v}) \rangle = \frac{1}{\text{mes}(\Omega)} \int_{\Omega} c(\mathbf{x}, t; \mathbf{v}) dV.$$

Mixing will be best enhanced if the scalar variance  $\|c(t; \mathbf{v}) - \langle c(t; \mathbf{v}) \rangle\|$  is made as small as possible. While mixing is best enhanced, the velocity  $\mathbf{v}$  is desired to be optimized in the sense that it is almost steady and is of the least magnitude  $\|\mathbf{v}\|$  and the least rotation. The least magnitude ensures that the cost of generating the flow will be lowest, and the least rotation guarantees that the flow is not too turbulent and chaotic. A development of rotation in a flow requires shear stress to be present on a fluid particle surface. The shear stress depends on the strain tensor  $\nabla \mathbf{v}$  of the velocity  $\mathbf{v}$  [23]. Thus, to have the least rotation, the magnitude  $\|\nabla \mathbf{v}\|$  needs to be minimized. To make the flow almost steady, the acceleration magnitude  $\|\frac{\partial \mathbf{v}(t)}{\partial t}\|$  needs to be minimized. This motivates us to define the following mixing efficiency functional:

$$\begin{aligned} J(\mathbf{v}) = & \int_0^T \left( \|c(t; \mathbf{v}) - \langle c(t; \mathbf{v}) \rangle\|^2 + \alpha \|\mathbf{v}(t)\|^2 + \beta \|\nabla \mathbf{v}(t)\|^2 + \gamma \left\| \frac{\partial \mathbf{v}(t)}{\partial t} \right\|^2 \right) dt \\ (2) \quad & + \mu \|c(T; \mathbf{v}) - \langle c(T; \mathbf{v}) \rangle\|^2, \end{aligned}$$

where  $T > 0$  is some desired time, and  $\alpha > 0, \beta, \gamma, \mu \geq 0$  are weight constants. In optimal control theory, the first two terms  $\|c(t; \mathbf{v}) - \langle c(t; \mathbf{v}) \rangle\|^2, \alpha \|\mathbf{v}(t)\|^2$ , which penalize the averages of the scalar variance and the controlling cost, are standard [8, 16]. The variance  $\|c(T; \mathbf{v}) - \langle c(T; \mathbf{v}) \rangle\|^2$  at the final time is optional but included in this functional to ensure that the highest level of homogenization of the scalar will be achieved. Another standard functional in optimal control theory is the one over an infinite time interval for a regulator problem. For the problem of mixing enhancement, this functional is not appropriate since mixing needs to be enhanced in a desired finite time.

The weight constants in (2) play an important role in determining the control strength. For small values of  $\alpha, \beta, \gamma$ , the functional will result in an optimal solution with a small variance of the scalar but with big magnitudes of the velocity  $\mathbf{v}$ , of the strain tensor  $\nabla \mathbf{v}$ , and of the acceleration  $\frac{\partial \mathbf{v}}{\partial t}$ . This implies that the smaller the weights, the more turbulent the optimal flow, and then the better the mixing enhancement.

We note that the mean is conserved. In fact, integrating (1) over  $\Omega$  gives

$$\frac{d}{dt} \langle c \rangle = \frac{\kappa}{\text{mes}(\Omega)} \int_{\Omega} \nabla^2 c dV = 0,$$

where we have used the boundary conditions on  $\mathbf{v}$  and  $c$ . Therefore we can assume zero mean without loss of generality. With the zero-mean assumption, the cost functional reduces to

$$(3) \quad J(\mathbf{v}) = \int_0^T \left( \|c(t; \mathbf{v})\|^2 + \alpha \|\mathbf{v}(t)\|^2 + \beta \|\nabla \mathbf{v}(t)\|^2 + \gamma \left\| \frac{\partial \mathbf{v}(t)}{\partial t} \right\|^2 \right) dt + \mu \|c(T; \mathbf{v})\|^2.$$

Then the optimal control problem is to minimize  $J$  in an admissible velocity space  $\mathcal{V} = H_0^1(0, T; \mathbf{H}_{div}^1(\Omega))$ :

$$(4) \quad J(\mathbf{v}^*) = \min_{\mathbf{v} \in \mathcal{V}} J(\mathbf{v}).$$

The minimizer  $\mathbf{v}^*$  is called an *optimal flow*.

In deriving optimality conditions for the optimal flow below, control flows are required to satisfy the condition  $\mathbf{v}(0) = \mathbf{v}(T) = 0$ . This mathematical condition, in fact, is quite realistic because the control flows should start from the rest and return back to the rest at the final time when the mixing has been enhanced. For instance, before coffee is stirred, the flow is at rest.

In this theoretical study, we assume that an arbitrary unsteady flow can be generated. This may not be realistic. In a future work, we will consider specific velocity fields such as  $\mathbf{v} = \sum_{i=1}^N \mathbf{v}_i(\mathbf{x}) u_i(t)$ , where  $\mathbf{v}_i(\mathbf{x})$  ( $i = 1, \dots, N$ ) are given steady flows which prescribe how the control action is distributed in the flow field.

**3. Existence of optimal flows.** For convenience, we state a well-known estimate about the solution of (1) as follows.

**LEMMA 3.1.** *Let  $\mathbf{v} \in L^2(0, T; \mathbf{L}_{div}^2(\Omega))$ . Then the solution  $c$  of (1) satisfies the following estimate:*

$$(5) \quad \|c(t)\|^2 + 2\kappa \int_0^t \|\nabla c(s)\|^2 ds = \|c^0\|^2.$$

*Proof.* Multiplying (1) by  $c$  and using the boundary conditions, we obtain the equation

$$(6) \quad \frac{1}{2} \frac{d}{dt} \|c\|^2 = -\kappa \|\nabla c\|^2.$$

Integrating over  $[t_0, t]$  gives (5).  $\square$

To prove the existence of an optimal flow, we need the following weakly lower semicontinuity of the function  $J$ .

**LEMMA 3.2.** *The functional  $J$  defined by (3) is weakly lower semicontinuous. That is, if  $\mathbf{v}_n$  converges weakly to  $\mathbf{v}_0$  in  $H^1(0, T; \mathbf{H}_{div}^1(\Omega))$ , then*

$$J(\mathbf{v}_0) \leq \liminf_{n \rightarrow \infty} J(\mathbf{v}_n).$$

*Proof.* Let  $\mathbf{v}_n$  converge weakly to  $\mathbf{v}_0$  in  $H^1(0, T; \mathbf{H}_{div}^1(\Omega))$  and let  $c_n(\mathbf{x}, t; \mathbf{v}_n)$  be the solution of

$$(7) \quad \frac{\partial c_n}{\partial t} + (\mathbf{v}_n \cdot \nabla) c_n = \kappa \nabla^2 c_n, \quad c_n(\mathbf{x}, 0) = c^0(\mathbf{x}) \quad \text{in } \Omega, \quad \text{and} \quad \frac{\partial c_n}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega.$$

Then  $\mathbf{v}_n$  converges strongly to  $\mathbf{v}_0$  in  $L^2(0, T; \mathbf{L}_{div}^2(\Omega))$ . Moreover, it follows from (5) that there exists a subsequence of  $c_n(\mathbf{x}, t; \mathbf{v}_n)$ , still denoted by itself for convenience, that converges weakly to  $c_0^*$  in  $L^2(0, T; H^1(\Omega))$ . Therefore we can pass to the limit in (7) and obtain

$$\frac{\partial c_0^*}{\partial t} + (\mathbf{v}_0 \cdot \nabla) c_0^* = \kappa \nabla^2 c_0^*, \quad c_0^*(\mathbf{x}, 0) = c^0(\mathbf{x}) \quad \text{in } \Omega, \quad \text{and} \quad \frac{\partial c_0^*}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega.$$

Since any norm of a Banach space is weakly lower semicontinuous [17], it therefore follows that

$$\begin{aligned} \liminf_{n \rightarrow \infty} J(\mathbf{v}_n) &\geq \liminf_{n \rightarrow \infty} \int_0^T \left( \|c_n(t; \mathbf{v}_n)\|^2 + \alpha \|\mathbf{v}_n(t)\|^2 \right. \\ &\quad \left. + \beta \|\nabla \mathbf{v}_n(t)\|^2 + \gamma \left\| \frac{\partial \mathbf{v}_n(t)}{\partial t} \right\|^2 \right) dt \\ &\quad + \mu \liminf_{n \rightarrow \infty} \|c_n(T; \mathbf{v}_n)\|^2 \\ &\geq \int_0^T \left( \|c_0^*(t; \mathbf{v}_0)\|^2 + \alpha \|\mathbf{v}_0(t)\|^2 + \beta \|\nabla \mathbf{v}_0(t)\|^2 + \gamma \left\| \frac{\partial \mathbf{v}_0(t)}{\partial t} \right\|^2 \right) dt \\ &\quad + \mu \|c_0^*(T; \mathbf{v}_0)\|^2 \\ &= J(\mathbf{v}_0). \end{aligned}$$

So the functional  $J$  is weakly lower semicontinuous.  $\square$

From this lemma, we can readily prove the following existence theorem.

**THEOREM 3.1.** *If  $\beta, \gamma > 0$ , then there exists an optimal flow  $\mathbf{v}^* \in \mathcal{V} = H^1(0, T; \mathbf{H}_{div}^1(\Omega))$  such that*

$$(8) \quad J(\mathbf{v}^*) = \min_{\mathbf{v} \in \mathcal{V}} J(\mathbf{v}).$$

*Proof.* Let  $\mathbf{v}_n$  be the minimizing sequence in  $H^1(0, T; \mathbf{H}_{div}^1(\Omega))$ . That is,

$$\lim_{n \rightarrow \infty} J(\mathbf{v}_n) = \min_{\mathbf{v} \in \mathcal{V}} J(\mathbf{v}).$$

Then  $\mathbf{v}_n$  is bounded in  $H^1(0, T; \mathbf{H}_{div}^1(\Omega))$ . This implies that there exists a subsequence, still denoted by  $\mathbf{v}_n$ , that converges weakly to  $\mathbf{v}^*$  in  $H^1(0, T; \mathbf{H}_{div}^1(\Omega))$ . It therefore follows from Lemma 3.2 that

$$J(\mathbf{v}^*) \leq \lim_{n \rightarrow \infty} J(\mathbf{v}_n) = \min_{\mathbf{v} \in \mathcal{V}} J(\mathbf{v}),$$

which implies (8).  $\square$

If  $\beta = \gamma = 0$ , the existence is open. In this case, the minimizing sequence  $\mathbf{v}_n$  is bounded only in  $L^2(0, T; \mathbf{L}_{div}^2(\Omega))$  and then may not converge strongly to  $\mathbf{v}_0$  in  $L^2(0, T; \mathbf{L}_{div}^2(\Omega))$ . Thus passing to the limit in (7) cannot be guaranteed. Also the uniqueness of the optimal flow is open because we could not prove that the functional  $J$  is convex.

#### 4. Optimality conditions.

THEOREM 4.1. *If  $\mathbf{v}^*$  is an optimal flow under the cost functional  $J$  defined by (3), then it satisfies the following equations:*

$$(9) \quad \frac{\partial c}{\partial t} + (\mathbf{v}^* \cdot \nabla)c = \kappa \nabla^2 c,$$

$$(10) \quad \frac{\partial g}{\partial t} + (\mathbf{v}^* \cdot \nabla)g = -\kappa \nabla^2 g + c(\mathbf{v}^*),$$

$$(11) \quad \nabla^2 p(\mathbf{x}, t) = \nabla g(\mathbf{x}, t; \mathbf{v}^*) \cdot \nabla c(\mathbf{x}, t; \mathbf{v}^*) + g(\mathbf{x}, t; \mathbf{v}^*) \nabla^2 c(\mathbf{x}, t; \mathbf{v}^*),$$

$$(12) \quad -\alpha \mathbf{v}^* + \beta \nabla^2 \mathbf{v}^* + \gamma \frac{\partial^2 \mathbf{v}^*}{\partial t^2} = g(\mathbf{x}, t; \mathbf{v}^*) \nabla c(\mathbf{x}, t; \mathbf{v}^*) - \nabla p,$$

$$(13) \quad \frac{\partial c}{\partial \mathbf{n}} = \frac{\partial g}{\partial \mathbf{n}} = \frac{\partial p}{\partial \mathbf{n}} = 0, \quad \mathbf{v}^* = 0 \quad \text{on } \partial\Omega,$$

$$(14) \quad \mathbf{v}^*(\mathbf{x}, 0) = \mathbf{v}^*(\mathbf{x}, T) = 0 \quad \text{in } \Omega,$$

$$(15) \quad c(\mathbf{x}, 0) = c^0(\mathbf{x}), \quad g(\mathbf{x}, T) = -\mu c(\mathbf{x}, T; \mathbf{v}^*) \quad \text{in } \Omega.$$

We will prove this theorem in the next two sections.

If  $\beta = 0$ , we can solve (12) to obtain

$$\begin{aligned} \mathbf{v}^* &= \frac{1}{2} \sqrt{\frac{1}{\alpha\gamma}} \int_0^t (\nabla p(\mathbf{x}, s) - g(\mathbf{x}, s) \nabla c(\mathbf{x}, s)) \left( e^{\sqrt{\alpha/\gamma}(s-t)} - e^{\sqrt{\alpha/\gamma}(t-s)} \right) ds \\ &\quad + \frac{1}{2} \sqrt{\frac{1}{\alpha\gamma}} \frac{e^{-t\sqrt{\alpha/\gamma}} - e^{t\sqrt{\alpha/\gamma}}}{e^{-T\sqrt{\alpha/\gamma}} - e^{T\sqrt{\alpha/\gamma}}} \\ &\quad \times \int_0^T (\nabla p(\mathbf{x}, s) - g(\mathbf{x}, s) \nabla c(\mathbf{x}, s)) \left( e^{\sqrt{\alpha/\gamma}(T-s)} - e^{\sqrt{\alpha/\gamma}(s-T)} \right) ds. \end{aligned}$$

This control law shows that the optimal control flow depends on all the concentration gradients during the whole time period from 0 to  $T$ .

Solving the system (9)–(15) numerically or analytically is a challenging problem since it is highly nonlinear. As an initial attempt, we give a preliminary numerical result.

One potential method for solving (9)–(15) could be the iteration method. We first solve the advection-diffusion (9) with a given velocity  $\mathbf{v}_1$ . Then with this solution  $c(\mathbf{v}_1)$ , we solve (10) and then (11). Through (12), we obtain a new velocity  $\mathbf{v}_2$ . With this  $\mathbf{v}_2$ , we repeat the above procedure, and so on.

To test whether or not this iteration method works, we consider a simplified case where  $\beta = \gamma = \mu = 0$  and the system is considered in a 2D domain. Under this simplification, this testing could be purely numerical, as real mixing may take place only in the 3D space. Since  $\beta = \gamma = 0$ , the boundary condition (14) is not needed, and then  $\mathbf{v}^*(\mathbf{x}, 0)$  may not be equal to zero in this case.

In our computations, the domain  $\Omega = (0, 1) \times (0, 1)$ , the initial condition  $c^0(x, y) = \sin(2\pi x) \sin(2\pi y)$ , the diffusivity  $\kappa = 0.01$ , the starting velocity  $\mathbf{v}_1 = \mathbf{0}$ ,  $T = 2$ ,  $\alpha = 0.5$ , and  $\beta = \gamma = \mu = 0$ . All equations are solved by the finite element method developed in [25] (with some modifications for this particular problem).

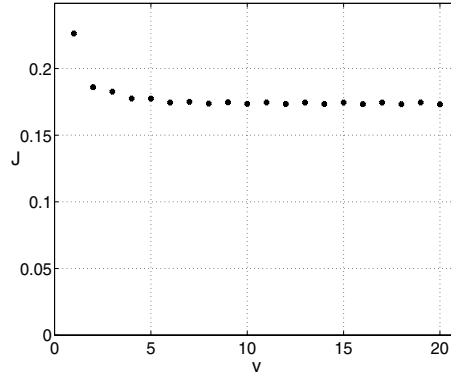


FIG. 1. Starting with the velocity  $\mathbf{v}_1 = 0$ , the functional  $J$  reaches its minimum after a number of iterations and then stays there.

Figure 1 shows that the functional  $J$  reaches its minimum after a number of iterations and then stays there. The approximate optimal flow  $\mathbf{v}^*$  obtained via 20 iterations in this numerical experiment is shown in Figure 2. From this figure it can be seen that the flow is decreasing to zero. This could imply that after the scalar is well advected, no further advection is needed to save the control efforts.

To see how this optimal flow enhances mixing, we compare the variance decay in the case of  $\mathbf{v}_1 = 0$  with the variance decay in the case of the optimal flow, where the variance is defined by

$$V(t) = \|c(t; \mathbf{v}) - \langle c(t; \mathbf{v}) \rangle\|^2.$$

Figure 3 shows that the variance of the scalar advected by the optimal flow decays much faster than the one without advection.

To further test whether or not the functional  $J(\mathbf{v})$  really attains the minimum at the optimal flow obtained above, we consider a couple of other model flows. One of them is the following time-periodic velocity [18], denoted by  $\mathbf{v}_{p_1}$ :

$$(16) \quad \begin{aligned} v_1(x, y, t) &= \begin{cases} \sin(\pi x) \cos(\pi y) & \text{if } n \leq t < n + 0.5; \\ -\sin(2\pi x) \cos(\pi y) & \text{if } n + 0.5 \leq t < n + 1; \end{cases} \\ v_2(x, y, t) &= \begin{cases} -\cos(\pi x) \sin(\pi y) & \text{if } n \leq t < n + 0.5; \\ 2 \cos(2\pi x) \sin(\pi y) & \text{if } n + 0.5 \leq t < n + 1. \end{cases} \end{aligned}$$

As above, we can compute the value of the functional  $J(\mathbf{v}_{p_1})$  at this flow and obtain

$$J(\mathbf{v}_{p_1}) = 1.0429,$$

which is greater than the value of the functional at the above optimal flow:

$$J(\mathbf{v}^*) = 0.1731.$$

Another flow is the simplified model flow, denoted by  $\mathbf{v}_{p_2}$ , of time-aperiodic Rayleigh–Bénard convection. The velocity field of the flow is derived from the stream function

$$(17) \quad \Psi = \frac{A}{n} \sin(2\pi x) \sin\{n[x + B \sin(\omega t)]\} W(y),$$

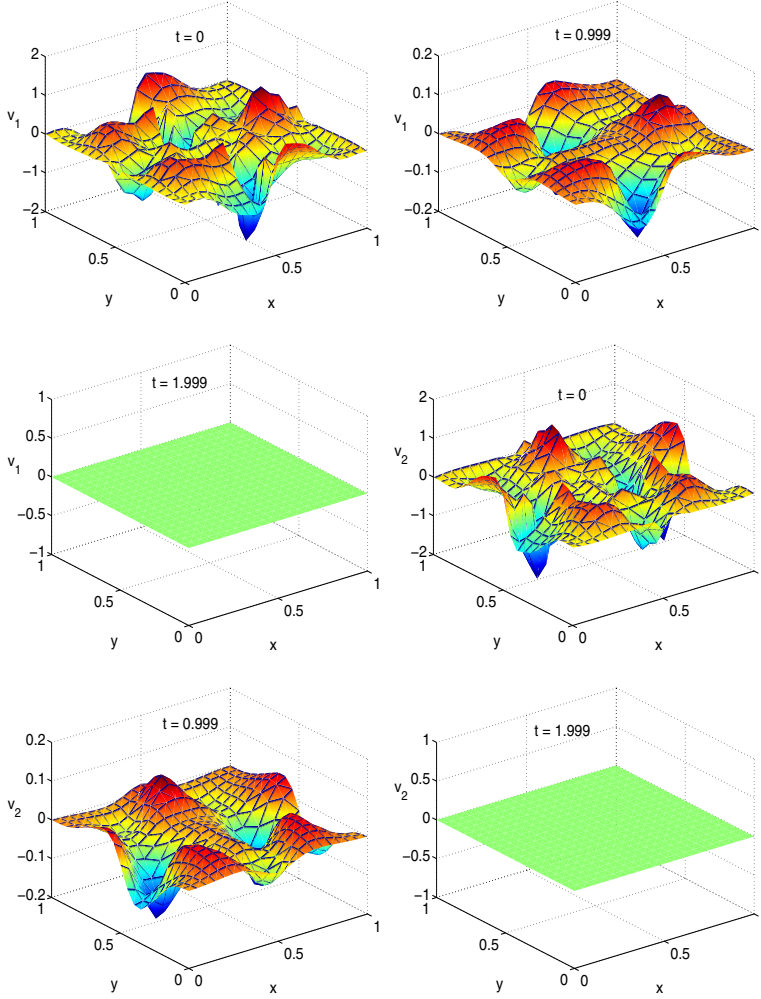


FIG. 2. The  $x$ -component  $v_1$  (top row and middle row (left)) and  $y$ -component  $v_2$  (middle row (right) and bottom row) of the optimal flow  $\mathbf{v}^*$  at  $t = 0, 0.999, 1.999$ , obtained via 20 iterations starting with the velocity  $\mathbf{v}_1 = 0$ .

where  $A$  is a positive constant,  $n$  is the wave number  $2\pi/\lambda$  with a constant  $\lambda$ , and  $W(y)$  is a function that satisfies the rigid boundary conditions at the top and bottom surfaces. Here we use the following function  $W(y)$ :

$$W(y) = (1 - y)y.$$

This stream function is obtained by adding the factor  $\sin(2\pi x)$  to a stream function used in [24] to make it satisfy the no-penetration boundary condition. In this computation,  $A = 1.8, B = 0.06, \omega = 2\pi$ , and  $\lambda = 2$ . The value of the functional at this flow is

$$J(\mathbf{v}_{p_2}) = 0.2801.$$

As before, it is also greater than  $J(\mathbf{v}^*) = 0.1731$ .

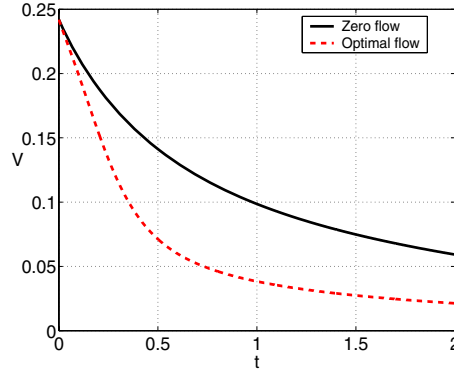


FIG. 3. The variance of a scalar advected by the optimal flow decays much faster than the one without advection.

The above is a very preliminary attempt. The complete resolution of the problem could require an extensive regularity analysis of solutions and applications of fixed point theorems.

The asymptotic behavior of solutions of the system (9)–(15) in the zero limit of the diffusivity  $\kappa$  is a singular perturbation problem. This problem is interesting but difficult. The resolution of the problem will require extensive asymptotic analysis such as the decomposition of inner and outer solutions and boundary layer analysis. This is beyond the reach of the paper.

**5. Gâteaux differentials.** The space  $C(\Omega)$  consists of all continuous functions  $f$  on  $\Omega$  with

$$\|f\|_{\infty} = \max_{\mathbf{x} \in \Omega} |f(\mathbf{x})| < \infty.$$

The function vector space  $\mathbf{C}(\Omega) = \{C(\Omega)\}^n$ .

**THEOREM 5.1.** *Let the functional  $J$  be defined by (3). If  $\mathbf{v} = (v_1, v_2, v_3) \in H^1(0, T; \mathbf{H}_{div}^1(\Omega))$  and  $\mathbf{u} = (u_1, u_2, u_3) \in H^1(0, T; \mathbf{H}_{div}^1(\Omega) \cap \mathbf{C}(\Omega))$ , then the Gâteaux differential of  $J$  is given by*

$$\begin{aligned} \langle J'(\mathbf{v}), \mathbf{u} \rangle &= \lim_{\varepsilon \rightarrow 0} \frac{J(\mathbf{v} + \varepsilon \mathbf{u}) - J(\mathbf{v})}{\varepsilon} \\ &= 2 \int_0^T \int_{\Omega} c(\mathbf{v}) h(\mathbf{v}, \mathbf{u}) dV dt + 2\mu \int_{\Omega} c(T; \mathbf{v}) h(T; \mathbf{v}, \mathbf{u}) dV \\ &\quad + 2 \int_0^T \int_{\Omega} \left( \alpha \mathbf{v} \cdot \mathbf{u} + \beta \nabla \mathbf{v} \cdot \nabla \mathbf{u} + \gamma \frac{\partial \mathbf{v}}{\partial t} \cdot \frac{\partial \mathbf{u}}{\partial t} \right) dV dt, \end{aligned} \quad (18)$$

where  $\nabla \mathbf{v} \cdot \nabla \mathbf{u} = \sum_{i,j=1}^3 \frac{\partial v_i}{\partial x_j} \frac{\partial u_i}{\partial x_j}$  and  $h$  is the solution of

$$\frac{\partial h}{\partial t} + (\mathbf{v} \cdot \nabla) h = \kappa \nabla^2 h - (\mathbf{u} \cdot \nabla) c(\mathbf{v}) \quad \text{in } \Omega, \quad (19)$$

$$h(\mathbf{x}, 0) = 0 \quad \text{in } \Omega, \quad \text{and} \quad \frac{\partial h}{\partial \mathbf{n}} = 0 \quad \text{on } \partial \Omega.$$

To prove this theorem, we need the following lemma. For a positive constant  $\varepsilon$  and  $\mathbf{v}, \mathbf{u} \in L^2(0, T; \mathbf{L}_{div}^2(\Omega))$ , we denote by  $c(\mathbf{v})$  and  $c_\varepsilon(\mathbf{v}, \mathbf{u})$  the solutions of (1) corresponding the velocities  $\mathbf{v}$  and  $\mathbf{v} + \varepsilon\mathbf{u}$ , respectively.

LEMMA 5.1. *Let  $\mathbf{u} \in L^2(0, T, \mathbf{L}_{div}^2(\Omega) \cap \mathbf{C}(\Omega))$  and denote  $h_\varepsilon(\mathbf{v}, \mathbf{u}) = c_\varepsilon(\mathbf{v}, \mathbf{u}) - c(\mathbf{v})$ . Then the  $h_\varepsilon$  satisfies the following estimates:*

$$(20) \quad \max_{0 \leq s \leq t} \|h_\varepsilon(s)\|^2 \leq \frac{2\varepsilon^2}{\kappa} \|c^0\|^2 \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds,$$

$$(21) \quad \frac{1}{2} \|h_\varepsilon(t)\|^2 + \kappa \int_0^t \|\nabla h_\varepsilon(s)\|^2 ds \leq \frac{\varepsilon^2}{\kappa} \|c^0\|^2 \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds,$$

$$(22) \quad \max_{0 \leq s \leq t} \left\| \frac{h_\varepsilon(s)}{\varepsilon} - h(s) \right\|^2 \leq \frac{4\varepsilon^2}{\kappa^2} \|c^0\|^2 \left( \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds \right)^2.$$

*Proof.* A direct calculation shows that  $h_\varepsilon$  satisfies

$$(23) \quad \begin{aligned} \frac{\partial h_\varepsilon}{\partial t} + (\mathbf{v} \cdot \nabla) h_\varepsilon &= \kappa \nabla^2 h_\varepsilon - \varepsilon (\mathbf{u} \cdot \nabla) c_\varepsilon(\mathbf{v}, \mathbf{u}) \quad \text{in } \Omega, \\ h_\varepsilon(\mathbf{x}, 0) &= 0 \quad \text{in } \Omega, \quad \text{and} \quad \frac{\partial h_\varepsilon}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Multiplying (23) by  $h_\varepsilon$  and using the boundary conditions, we obtain the equation

$$(24) \quad \begin{aligned} \frac{1}{2} \frac{d}{dt} \|h_\varepsilon(t)\|^2 &= -\kappa \|\nabla h_\varepsilon\|^2 - \varepsilon \int_\Omega h_\varepsilon (\mathbf{u} \cdot \nabla) c_\varepsilon dV \\ &\leq \varepsilon \|\mathbf{u}(t)\|_\infty \|h_\varepsilon(t)\| \|\nabla c_\varepsilon(t)\|. \end{aligned}$$

Integrating over  $[t_0, t]$  gives

$$\begin{aligned} \|h_\varepsilon(t)\|^2 &\leq 2\varepsilon \max_{0 \leq s \leq t} \|h_\varepsilon(s)\| \int_0^t \|\mathbf{u}(s)\|_\infty \|\nabla c_\varepsilon(s)\| ds \\ &\leq 2\varepsilon \max_{0 \leq s \leq t} \|h_\varepsilon(s)\| \left( \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds \right)^{1/2} \left( \int_0^t \|\nabla c_\varepsilon(s)\|^2 ds \right)^{1/2}, \end{aligned}$$

which implies that

$$\begin{aligned} \max_{0 \leq s \leq t} \|h_\varepsilon(s)\|^2 &\leq 2\varepsilon \max_{0 \leq s \leq t} \|h_\varepsilon(s)\| \left( \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds \right)^{1/2} \left( \int_0^t \|\nabla c_\varepsilon(s)\|^2 ds \right)^{1/2} \\ &\leq \frac{1}{2} \left( \max_{0 \leq s \leq t} \|h_\varepsilon(s)\| \right)^2 + 2\varepsilon^2 \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds \int_0^t \|\nabla c_\varepsilon(s)\|^2 ds. \end{aligned}$$

It then follows from (5) that

$$(25) \quad \begin{aligned} \max_{0 \leq s \leq t} \|h_\varepsilon(s)\|^2 &\leq 4\varepsilon^2 \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds \int_0^t \|\nabla c_\varepsilon(s)\|^2 ds \\ &\leq \frac{2\varepsilon^2}{\kappa} \|c^0\|^2 \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds. \end{aligned}$$



This proves (20).

To prove (21), we use (24) again and derive that

$$\begin{aligned}
 & \frac{1}{2} \|h_\varepsilon(t)\|^2 + \kappa \int_0^t \|\nabla h_\varepsilon(s)\|^2 ds \\
 & \leq \varepsilon \max_{0 \leq s \leq t} \|h_\varepsilon(s)\| \int_0^t \|\mathbf{u}(s)\|_\infty \|\nabla c_\varepsilon(s)\| ds \\
 & \leq \varepsilon \max_{0 \leq s \leq t} \|h_\varepsilon(s)\| \left( \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds \right)^{1/2} \left( \int_0^t \|\nabla c_\varepsilon(s)\|^2 ds \right)^{1/2}.
 \end{aligned}$$

It then follows from (5) and (20) that

$$\begin{aligned}
 \frac{1}{2} \|h_\varepsilon(t)\|^2 + \kappa \int_0^t \|\nabla h_\varepsilon(s)\|^2 ds & \leq \varepsilon^2 \frac{\sqrt{2}}{\sqrt{\kappa}} \|c^0\| \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds \frac{\|c^0\|}{\sqrt{2\kappa}} \\
 (26) \qquad \qquad \qquad & \leq \frac{\varepsilon^2}{\kappa} \|c^0\|^2 \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds.
 \end{aligned}$$

To prove (22), we denote  $f_\varepsilon = \frac{h_\varepsilon}{\varepsilon} - h$ . A direct calculation shows that

$$\begin{aligned}
 (27) \qquad \frac{\partial f_\varepsilon}{\partial t} + (\mathbf{v} \cdot \nabla) f_\varepsilon & = \kappa \nabla^2 f_\varepsilon - (\mathbf{u} \cdot \nabla) h_\varepsilon \quad \text{in } \Omega, \\
 f_\varepsilon(\mathbf{x}, 0) & = 0 \quad \text{in } \Omega, \quad \text{and} \quad \frac{\partial f_\varepsilon}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega.
 \end{aligned}$$

In the same way as above, we can derive that

$$\begin{aligned}
 \max_{0 \leq s \leq t} \|f_\varepsilon(s)\|^2 & \leq 4 \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds \int_0^t \|\nabla h_\varepsilon(s)\|^2 ds \\
 & \leq \frac{4\varepsilon^2}{\kappa^2} \|c^0\|^2 \left( \int_0^t \|\mathbf{u}(s)\|_\infty^2 ds \right)^2. \quad \square
 \end{aligned}$$

We are now ready to prove Theorem 5.1. Using the estimates (20) and (22), we derive that

$$\begin{aligned}
 & \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \int_0^T \int_\Omega [|c(\mathbf{v} + \varepsilon \mathbf{u})|^2 - |c(\mathbf{v})|^2] dV dt \\
 & = \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \int_0^T \int_\Omega [|c(\mathbf{v}) + h_\varepsilon(\mathbf{v}, \mathbf{u})|^2 - |c(\mathbf{v})|^2] dV dt \\
 & = \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \int_0^T \int_\Omega [2c(\mathbf{v})h_\varepsilon(\mathbf{v}, \mathbf{u}) + (h_\varepsilon(\mathbf{v}))^2] dV dt \\
 (28) \qquad \qquad \qquad & = 2 \int_0^T \int_\Omega c(\mathbf{v})h(\mathbf{v}, \mathbf{u}) dV dt.
 \end{aligned}$$

In the same way, we can show that

$$\begin{aligned}
& \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \int_{\Omega} [|c(T; \mathbf{v} + \varepsilon \mathbf{u})|^2 - |c(T; \mathbf{v})|^2] dV \\
&= \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \int_{\Omega} [|c(T; \mathbf{v}) + h_{\varepsilon}(T; \mathbf{v}, \mathbf{u})|^2 - |c(\mathbf{v})|^2] dV \\
&= \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \int_{\Omega} [2c(T; \mathbf{v})h_{\varepsilon}(T; \mathbf{v}, \mathbf{u}) + (h_{\varepsilon}(T; \mathbf{v}, \mathbf{u}))^2] dV \\
&= 2 \int_{\Omega} c(T; \mathbf{v})h(T; \mathbf{v}, \mathbf{u}) dV.
\end{aligned}$$

Using these limits, we then readily prove Theorem 5.1.

**6. Proof of Theorem 4.1.** If the flow  $\mathbf{v}^*$  is an optimal flow under the functional  $J$  defined by (3), then it satisfies

$$\langle J'(\mathbf{v}), \mathbf{u} \rangle = 0$$

for all  $\mathbf{u} \in H_0^1(0, T; \mathbf{H}_{div}^1(\Omega))$ . It then follows from (18) that

$$\begin{aligned}
(29) \quad & \int_0^T \int_{\Omega} c(\mathbf{x}, t; \mathbf{v}^*)h(\mathbf{x}, t; \mathbf{u}, \mathbf{v}^*) dV dt + \mu \int_{\Omega} c(\mathbf{x}, T; \mathbf{v}^*)h(\mathbf{x}, T; \mathbf{u}, \mathbf{v}^*) dV \\
& + \int_0^T \int_{\Omega} \left( \alpha \mathbf{v}^* \cdot \mathbf{u} + \beta \nabla \mathbf{v}^* \cdot \nabla \mathbf{u} + \gamma \frac{\partial \mathbf{v}^*}{\partial t} \cdot \frac{\partial \mathbf{u}}{\partial t} \right) dV dt = 0
\end{aligned}$$

for all  $\mathbf{u} \in H_0^1(0, T; \mathbf{H}_{div}^1(\Omega) \cap \mathbf{C}(\Omega))$ . Consider the adjoint equation

$$(30) \quad \frac{\partial g}{\partial t} + (\mathbf{v}^* \cdot \nabla)g = -\kappa \nabla^2 g + c(\mathbf{v}^*) \quad \text{in } \Omega,$$

$$g(\mathbf{x}, T) = -\mu c(\mathbf{x}, T; \mathbf{v}^*) \quad \text{in } \Omega, \quad \text{and} \quad \frac{\partial g}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega.$$

Multiplying (30) by  $h$  and (19) by  $g$  and integrating over  $\Omega \times [0, T]$ , we obtain

$$\begin{aligned}
(31) \quad & \int_0^T \int_{\Omega} c(\mathbf{x}, t; \mathbf{v}^*)h(\mathbf{x}, t; \mathbf{u}, \mathbf{v}^*) dV dt + \mu \int_{\Omega} c(\mathbf{x}, T; \mathbf{v}^*)h(\mathbf{x}, T; \mathbf{u}, \mathbf{v}^*) dV \\
& = \int_0^T \int_{\Omega} (\mathbf{u} \cdot \nabla c(\mathbf{x}, t; \mathbf{v}^*)) g(\mathbf{x}, t; \mathbf{v}^*) dV dt.
\end{aligned}$$

After integration by parts with respect to  $t$ , we deduce from (29) and (31) that

$$(32) \quad \int_0^T \int_{\Omega} \mathbf{u} \cdot \left( \alpha \mathbf{v}^* - \beta \nabla^2 \mathbf{v}^* - \gamma \frac{\partial^2 \mathbf{v}^*}{\partial t^2} + g(\mathbf{x}, t; \mathbf{v}^*) \nabla c(\mathbf{x}, t; \mathbf{v}^*) \right) dV dt = 0,$$

which implies that there exists a potential function  $p$  such that

$$(33) \quad \alpha \mathbf{v}^* - \beta \nabla^2 \mathbf{v}^* - \gamma \frac{\partial^2 \mathbf{v}^*}{\partial t^2} + g(\mathbf{x}, t; \mathbf{v}^*) \nabla c(\mathbf{x}, t; \mathbf{v}^*) = \nabla p.$$

To determine  $p$ , we apply the divergence operation to the above equation and then obtain

$$(34) \quad \nabla^2 p = \nabla g(\mathbf{x}, t; \mathbf{v}^*) \cdot \nabla c(\mathbf{x}, t; \mathbf{v}^*) + g(\mathbf{x}, t; \mathbf{v}^*) \nabla^2 c(\mathbf{x}, t; \mathbf{v}^*).$$

Thus we have proved Theorem 4.1.

**7. Conclusions.** We have studied the problem of optimal mixing control whose objective is to best enhance mixing by flow advection while the flow is optimized in the sense that it is almost steady and is of the least magnitude and the least rotation. In solving this problem, we defined a mixing efficiency functional by penalizing the average of variance of the scalar, the average of the flow, and the average of the strain tensor and acceleration of the flow. We showed that the functional is weakly lower semicontinuous and then it attains its minimum. By variational principles, we then derived optimality conditions that consist of a system of nonlinear partial differential equations.

A number of issues are left open. Solving the optimality partial differential equations numerically or analytically is challenging since nonlinearity is presented in the advection term, which, like in the study of Navier–Stokes equations, is difficult to estimate. The resolution of the problem could require the regularity analysis of solutions and applications of fixed point theorems. The uniqueness of the optimal flow is open because we could not prove that the efficiency functional  $J$  is convex. Another interesting problem is the singular perturbation problem for the optimality partial differential equations in the zero limit of diffusivity. The resolution of the problem will require an extensive asymptotic analysis such as the decomposition of inner and outer solutions and boundary layer analysis.

Results presented in this paper could have potential applications in aerospace engineering and mixing-related industry. Often a certain level of homogeneity of a fluid mixture is desired. For instance, before fuel is burned in a combustor, it is required to be well mixed so that the combustor has its best efficiency. Hence, optimality conditions derived in this paper could serve as guidelines in implementing an efficient and practical control technique for mixing enhancement.

**Acknowledgments.** The author thanks George Haller for introducing the mixing problem to the author and for constant discussions, thanks Enrique Zuazua for valuable comments on the weakly lower semicontinuity of the efficiency functional, and thanks the referees for their critical comments. One of the referees provided especially sharp insights into the mixing physics and English language comments which greatly improve the quality of presentation of the paper and make the author's Chinese style English close to native English. The author is grateful to his son, Xin Liu, for proofreading the text part of the manuscript.

#### REFERENCES

- [1] O. M. AAMO, M. KRSTIC, AND T. R. BEWLEY, *Control of mixing by boundary feedback in 2D channel flow*, Automatica, 39 (2003), pp. 1597–1606.
- [2] D. D'ALESSANDRO, M. DAHLEH, AND I. MEZIĆ, *Control of mixing in fluid flow: A maximum entropy approach*, IEEE Trans. Automat. Control, 44 (1999), pp. 1852–1863.
- [3] D. D'ALESSANDRO, I. MEZIĆ, AND M. DAHLEH, *Statistical properties of controlled fluid flows with applications to control of mixing*, Systems Control Lett., 45 (2002), pp. 249–256.
- [4] T. M. ANTONSEN, Z. FAN, E. OTT, AND E. GARCIA-LOPEZ, *The role of passive scalars in the determination of power spectra of passive scalars*, Phys. Fluids, 8 (1996), pp. 3094–3104.
- [5] A. BALOGH, O. M. AAMO, AND M. KRSTIC, *Optimal mixing enhancement in 3D pipe flow*, IEEE Trans. Control Systems Technology, 13 (2005), pp. 27–41.
- [6] R. BOTTAUSCI, I. MEZIĆ, C. D. MEINHART, AND C. CARDONNE, *Mixing in the shear superposition micromixer: Three-dimensional analysis*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 1001–1018.
- [7] B. V. N. CHARYULU, J. KURIAN, P. VENUGOPALAN, AND V. SRIRAMULU, *Experimental study on mixing enhancement in two dimensional supersonic flow*, Exp. Fluids, 24 (1998), pp. 340–346.

- [8] R. F. CURTAIN AND H. ZWART, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.
- [9] L. C. EVANS, *Partial Differential Equations*, AMS, Providence, RI, 1998.
- [10] A. FANNJIANG AND G. PAPANICOLAOU, *Convection enhanced diffusion for periodic flows*, SIAM J. Appl. Math., 54 (1994), pp. 333–408.
- [11] D. R. FEREDAY, P. H. HAYNES, AND A. WONHAS, *Scalar variance decay in chaotic advection and the Batchelor–regime turbulence*, Phys. Rev. E (3), 65 (2002), 035301.
- [12] M. GIONA, S. CERBELLI, AND V. VITACOLONNA, *Universality and imaginary potentials in advection-diffusion equations in closed flows*, J. Fluid Mech., 513 (2004), pp. 221–237.
- [13] M. GIONA, V. VITACOLONNA, S. CERBELLI, AND A. ADROVER, *Advection diffusion in nonchaotic closed flows: Non-Hermitian operators, universality, and localization*, Phys. Rev. E (3), 70 (2004), 046224.
- [14] D. GREENBLATT AND I. J. WYGNANSKI, *The control of flow separation by periodic excitation*, Progr. Aerospace Sci., 36 (2000), p. 487.
- [15] P. H. HAYNES AND J. VANNESTE, *What controls the decay of passive scalars in smooth flows?*, Phys. Fluids, 17 (2005), 097103.
- [16] I. LASIECKA AND R. TRIGGIANI, *Control Theory for Partial Differential Equations, Volume 1: Abstract Parabolic Systems*, Cambridge University Press, Cambridge, UK, 2000.
- [17] X. LI AND J. YONG, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, 1995.
- [18] W. LIU AND G. HALLER, *Strange eigenmodes and decay of variance in the mixing of diffusive tracers*, Phys. D, 188 (2004), pp. 1–39.
- [19] W. LIU, *Does a fast mixer really exist?*, Phys. Rev. E (3), 72 (2005), 016312.
- [20] G. MATHEW, I. MEZÍČ, AND L. PETZOLD, *A multiscale measure for mixing*, Phys. D, 211 (2005), pp. 23–46.
- [21] B. R. NOACK, I. MEZÍČ, G. TADMOR, AND A. BANASZUK, *Optimal mixing in recirculation zones*, Phys. Fluids, 16 (2004), pp. 867–888.
- [22] A. PIKOVSKY AND O. POPOVYCH, *Persistent patterns in deterministic mixing flows*, Europhys. Lett., 61 (2003), pp. 625–631.
- [23] I. H. SHAMES, *Mechanics of Fluids*, 4th ed., McGraw-Hill, New York, 2003.
- [24] T. H. SOLOMON AND J. P. GOLLUB, *Chaotic particle transport in time-dependent Rayleigh-Bénard convection*, Phys. Rev. A (3), 38 (1988), pp. 6280–6286.
- [25] E. G. THOMPSON, *Introduction to the Finite Element Method: Theory, Programming and Applications*, John Wiley and Sons, Hoboken, NJ, 2005.
- [26] Y.-K. TSANG, T. M. ANTONSEN, JR., AND E. OTT, *Exponential decay of chaotically advected passive scalars in the zero diffusivity limit*, Phys. Rev. E (3), 71 (2005), 066301.
- [27] Y. WANG, G. HALLER, A. BANASZUK, AND G. TADMOR, *Closed-loop Lagrangian separation control in a bluff body shear flow model*, Phys. Fluids, 15 (2003), pp. 2251–2266.
- [28] W. M. YANG, S. K. CHOU, C. SHU, Z. W. LI, AND H. XUE, *Combustion in micro-cylindrical combustors with and without a backward facing step*, Appl. Thermal Eng., 22 (2002), pp. 1777–1787.
- [29] C. C. YUAN, M. KRSTIC, AND T. BEWLEY, *Active control of jet mixing*, IEE Proc. Control Theory Appl., 151 (2004), pp. 763–772.

## EFFECT OF INPUT NOISE ON A MAGNETOMETER WITH QUANTUM FEEDBACK\*

ZHIGANG ZHANG<sup>†</sup>

**Abstract.** This article investigates the effects of input noise on a magnetometer with quantum feedback. Previous research has shown that feedback makes the measurement robust to an unknown parameter, the number of atoms involved, with the assumption that the feedback is noise free. To evaluate the effects of the feedback noise, we extend the original model by an input noise term. We then analyze the steady state performance of the Kalman filter for both the closed-loop and open-loop cases and retrieve the second moment of the estimation error. The results are compared and criteria for evaluating the effects of input noise are obtained. Robust and optimal designs are also discussed. Computations and simulations show how quantitatively the input noise increases the estimation error and changes the region where the closed-loop case behaves better than the open-loop case.

**Key words.** quantum control, magnetometry, Kalman filter, LQG controller

**AMS subject classifications.** 93E10, 93E20, 81V45, 62K25

**DOI.** 10.1137/060667189

**1. Introduction.** Several forces are driving the research of quantum control. The rapid development of optics and atomic physics has made it possible to now observe and manipulate small-scale quantum systems in laboratories, approaching the limitation allowed by quantum mechanics. The idea of controlling a quantum system is no longer impractical, and a steadily growing number of quantum control systems are now experimentally accessible. A particular area of interest in which quantum control is important is *quantum computing*. For many laser-driven, microscopically small quantum computing components, special control technologies are needed to make the quantum gates robust against uncertainties and reliable even under small disturbances. Quantum control also finds its applications in chemistry, metrology, and other fields. For example, substantial improvement can be made by quantum control in protein structure determination via nuclear magnetic resonance (NMR) [7].

To describe a quantum system, one requires knowledge of quantum mechanics. The model is usually a Schrödinger equation and may contain random factors to constitute a *quantum stochastic differential equation* (SDE) [19] or a *stochastic master equation* (SME) [18, 21, 20]. These models are mostly nonlinear and do not appear in the setting of classical control theory. Fortunately, a simplified model for control purposes is available in most cases, and the design of the controller and filter is analogous to that of a classical system [2]. Using the separation principle, we can normally separate the control system into a filter and a controller. The filter, combined with a continuous *quantum nondemolition measurement* (QND), covers all the quantum details and estimates the system state. This makes the controller design much easier. It is not surprising that many classical control methods find their way into the control of quantum systems. Examples include, but are not limited to, Bayesian estimation, the Kalman filter, bang-bang control, the linear quadratic Gaussian (LQG)

---

\*Received by the editors August 9, 2006; accepted for publication (in revised form) October 19, 2007; published electronically February 8, 2008.

<http://www.siam.org/journals/sicon/47-2/66718.html>

<sup>†</sup>Department of Mathematics, University of Houston, Houston, TX 77004-3008 (zgzhang@math.uh.edu).

controller, and optimal control. Realization of nonclassical states [12], such as squeezed states [18, 21, 3] and Dicke-states [17], which had been difficult to achieve stably in the laboratory, now occurs frequently in the literature. Other applications, such as tracking a single molecule [1, 8], estimating the phase of a continuous beam of light [11], or even cooling a single atom with a laser [15], are also reported.

Measurement of basic physics quantities, such as a magnetic field, is fundamental for physics and for applications such as quantum computation. The system discussed in this article is a *magnetometer* with an ensemble of atomic spins as its probe. When put in a magnetic field, the collective spin rotates perpendicularly in the direction of the field. A light beam passing through experiences polarization changes due to the collective spin, and this polarization is then measured continuously to estimate the magnetic field. Feedback can be applied by adding another magnetic field controlled by a computer. The same system has been used to produce a squeezed spin state [3]. With the assumption that the applied feedback magnetic field has infinite accuracy, Stockton et al. [16] built a model for the system and showed that the measurement with feedback was robust to an unknown parameter in the model, which is the number of atoms in the ensemble. Molmer and Madsen [9] also present a theory for the estimation. Petersen, Madsen, and Molmer [10] extend the system setup to several ensembles to measure a vector field and gain precision, but the input noise has not yet been taken into account.

Our study here focuses on the effects of the input noise on the measurement. We begin from the SME model and prove that it is equivalent to a simple linear SDE model [16]. This model is extended with an input noise. We then proceed to find the steady state performances of the measurement and the second moments of the estimation error in both closed-loop and open-loop cases. As was done previously, we use the *Kalman filter* to estimate the system state and the *LQG controller* to find the compensative field. We also assume that the observer does not know the exact number of atoms.

This article is organized as follows. In section 2, we introduce the system setup and find the simplification of the SME model. In section 3, we determine the steady state performance when no feedback is applied (open-loop case). Similarly, the steady state performance with feedback (closed-loop case) is determined in section 4. In section 5, we compare the results between the open-loop and closed-loop cases, and in section 6, we look into the optimal and robust designs. Finally, in section 7, a conclusion is given.

**2. System setup and its model.** The system is designed to measure an unknown and possibly fluctuating magnetic field  $b(t)$  oriented along the  $y$ -axis. A schematic is given in Figure 2.1. It consists of a cloud of atoms, a linearly polarized off-resonant light beam, and a polarization detector formed by two photon detectors and one amplifier. When the control loop is closed, a compensative magnetic field  $u(t)$  is applied by a computer via a group of coils. The computer collects information from the polarization detector, estimates the state of the system, and decides the current through the coils. The atomic nuclei have nonzero spin. Originally initialized along the  $x$ -axis, the collective spin of the nuclei (hereafter we just call it the collective spin) will rotate in the  $x - z$  plane if  $b(t)$  is not completely compensated by  $u(t)$ . The  $z$ -component of the collective spin is measured with the light beam, which is linearly polarized and off-resonant to the atoms. While passing through the atomic cloud, the light beam experiences a Faraday rotation proportional to the magnetic field along its propagation direction, the  $z$ -axis. Because both  $u(t)$  and  $b(t)$  are along the  $y$ -axis,

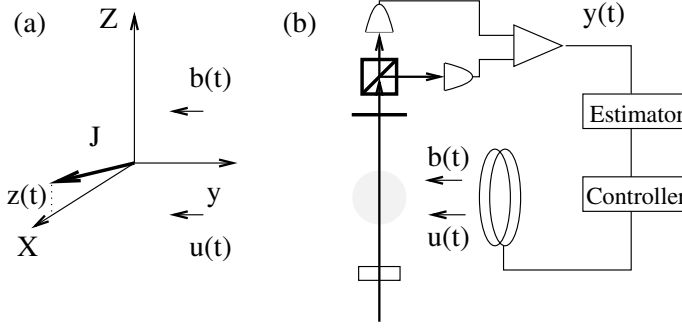


FIG. 2.1. The setup of a magnetometer. The external magnetic field  $b(t)$  is directed along the  $y$ -axis, which is unknown and possibly fluctuating. To measure it, an ensemble of atoms with nonzero nuclear spin is put in the field and initially polarized along the  $x$ -axis. A linearly polarized light beam travels through the atomic cloud along the  $z$ -axis. The signal of the polarization change is picked up at the end by a pair of photodetectors. A spin ensemble is shown in (a) with spins polarized among a small angle of  $x$ -axis. The Kalman filter (estimator) and controller are shown in (b). The picture is excerpted from [16].

the rotation is purely an effect of the collective spin.

Let the state of a quantum system be given by its wavefunction  $|\psi\rangle$ , which lives in a finite dimensional  $\mathcal{L}^2$  space. According to quantum mechanics, every observable physics quantity,  $O$ , is associated with an operator on the  $\mathcal{L}^2$  space, denoted by  $\hat{O}$ . Let  $|\xi_i\rangle$  be an orthonormal basis of the space; the expectation of the quantity, denoted by  $\langle\hat{O}\rangle$ , is given by

$$E[O] = \langle\hat{O}\rangle = \langle\psi|\hat{O}|\psi\rangle = \text{Tr}(\hat{O}|\psi\rangle\langle\psi|) = \sum_i \langle\xi_i|\hat{O}|\psi\rangle\langle\psi|\xi_i\rangle,$$

where we have used the Dirac notation. For an ensemble of many identical quantum systems which are prepared similarly, we may not know the state of each system. Instead, as for the ensemble of atomic spins, only the probability distribution of possible states  $|\psi_j\rangle$ ,  $P_j$ , is available. The expectation of the quantity is then the average over all possible  $|\psi_j\rangle$  and is given by

$$(2.1) \quad \langle\hat{O}\rangle = \text{Tr}(\hat{O}\rho),$$

where  $\rho = \sum_j P_j |\psi_j\rangle\langle\psi_j|$  is called the *density matrix* or *density operator*. Similarly, the variance of  $O$  is given by

$$\text{Var}[O] = E[(O - E[O])^2] = \langle(\hat{O} - E[O])^2\rangle = \langle\Delta\hat{O}^2\rangle,$$

where  $\Delta\hat{O} = \hat{O} - \langle\hat{O}\rangle$ . When there is only one operator involved, we use  $\langle O \rangle$  and  $\langle\Delta O^2\rangle$  instead of  $\langle\hat{O}\rangle$  and  $\langle\Delta\hat{O}^2\rangle$ , respectively, without confusion.

The evolution of  $\rho$  can be derived using the Schrödinger equation of  $|\psi\rangle$ ,

$$(2.2) \quad \dot{\rho} = -\frac{i}{\hbar}[\hat{H}, \rho] = -\frac{i}{\hbar}(\hat{H}\rho - \rho\hat{H}),$$

where  $\hat{H}$  is the system Hamiltonian without interaction with its environment. This equation is called the Liouville or Von Neumann equation [13] or master equation. For the ensemble of spins we consider in this article, we have to take into account

the decay of the system due to its interaction with its environment (bath). Moreover, information from the measurement improves our knowledge about the system, and a density matrix conditioned on the measurement history fits our needs better. The evolution of this conditioned density matrix satisfies a more complicated SME [13, 18, 21]. In the system setup of this article, it is given by

$$(2.3) \quad d\rho_c(t) = -idt[\hat{H}(t), \rho_c(t)] + \mathcal{D}[\sqrt{M}\hat{J}_z]\rho_c(t)dt + \sqrt{\eta}\mathcal{H}[\sqrt{M}\hat{J}_z]\rho_c(t) \cdot 2\sqrt{M\eta}(y(t)dt - \langle\hat{J}_z\rangle(t)dt),$$

where  $\hat{H} = \gamma h \hat{J}_z$ ,  $\rho_c$  is the conditioned density matrix, and  $\hat{J}_z$  and  $\hat{J}_y$  are the operators associated with the  $z$ -component and  $y$ -component of the collective spin, respectively. Other symbols are defined below:  $\gamma$  is a constant called the gyromagnetic ratio,  $h = b(t) + u(t)$  is the total magnetic field,  $M$  is the measurement rate,  $\eta$  is the quantum efficiency of the detection, and  $y(t)$  is the photocurrent detected. Operators  $\mathcal{H}$  and  $\mathcal{D}$  are two superoperators defined by

$$(2.4) \quad \begin{aligned} \mathcal{D}(\hat{c})\rho &= \hat{c}\rho\hat{c}^\dagger - (\hat{c}^\dagger\hat{c}\rho + \rho\hat{c}^\dagger\hat{c})/2, \\ \mathcal{H}(\hat{c})\rho &= \hat{c}\rho + \rho\hat{c}^\dagger - \text{Tr}[(\hat{c} + \hat{c}^\dagger)\rho]\rho, \end{aligned}$$

where  $\hat{c}$  is an operator.

We define  $d\bar{W}(t) = \frac{1}{\sqrt{\sigma_m}}(y(t)dt - \langle J_z \rangle(t)dt)$ , where  $\sigma_m = \frac{1}{4M\eta}$ . The process  $\bar{W}$  is a stochastic quantity representing the shot noise in the photodetection process, and it is a Brownian motion. The SME model is first derived for an ensemble of atoms interacting with a single mode far-off resonant cavity [18]. However, Silberfarb and Deutsch lately show that it also applies to the free-space system we have [14]. When  $\eta = 0$ , we discard the measurement result and (2.3) changes into a deterministic master equation. In terms of  $d\bar{W}$ , the measurement result  $y(t)$  is also a stochastic process:

$$(2.5) \quad y(t)dt = \langle J_z \rangle(t)dt + \sqrt{\sigma_m}d\bar{W}(t).$$

According to its definition, the density matrix is another form describing the system state besides the wavefunction. When the measurement result is taken into account, the conditioned density matrix,  $\rho_c$ , gives us the best estimation of the system state. Its evolution is described by the SME models of (2.3) and (2.5). In this setup, to estimate the magnetic field we are concerned only with the current through the detector, and thus the  $z$ -component of the collective spin. It will be good enough for us to know the various moments of  $J_z$ , especially the first two.

Given  $\rho_c$ , the expectation and variance of  $J_z$ , denoted as  $\langle J_z \rangle$  and  $\langle \Delta J_z^2 \rangle$ , respectively, can be computed using (2.1):

$$(2.6) \quad \begin{cases} \langle J_z \rangle = \text{Tr}(\rho_c \hat{J}_z), \\ \langle \Delta J_z^2 \rangle = \text{Tr}(\rho_c (\hat{J}_z - \langle \hat{J}_z \rangle)^2). \end{cases}$$

The SDE of  $\langle J_z \rangle$  and its variance  $\langle \Delta J_z^2 \rangle$  can be derived from the SME [16]:

$$(2.7) \quad \begin{aligned} d\langle J_z \rangle(t) &= \gamma \langle J_x \rangle(t) h(t) dt + \frac{\langle \Delta J_z^2 \rangle(t)}{\sqrt{\sigma_m}} d\bar{W}(t), \\ d\langle \Delta J_z^2 \rangle(t) &= -\frac{\langle \Delta J_z^2 \rangle^2}{\sigma_m} dt - i\gamma \langle [\Delta \hat{J}_z^2, \hat{J}_y] \rangle(t) h(t) dt + \frac{\langle \Delta \hat{J}_z^3 \rangle(t)}{\sqrt{\sigma_m}} d\bar{W}(t). \end{aligned}$$



Clearly, both  $\langle J_z \rangle$  and  $\langle \Delta J_z^2 \rangle$  are scalar stochastic processes driven by  $d\bar{W}$  and  $h(t)$ . According to (2.6), it is actually the conditional expectation of  $J_z$  based on the measurement history, and thus the best estimation.

The above SDE, (2.7), is nonlinear and intractable. However, it can be simplified under certain conditions. First, we assume that the collective spin vector,  $\mathbf{J} = [J_x, J_y, J_z]$ , is a stochastic Gaussian process, which is called the *Gaussian assumption*; second, we assume that  $\mathbf{J}$  is kept around the  $x$ -axis, which is called the *small angle assumption*. Then a one-dimensional linear stochastic process whose optimal estimation is equivalent to  $\langle J_z \rangle$  can be found under these two assumptions.

**THEOREM 2.1.** *A linear stochastic process defined by*

$$(2.8) \quad \begin{cases} dz(t) &= \gamma J h(t) dt, \\ y(t) dt &= z(t) dt + \sqrt{\sigma_m} dw_2(t), \end{cases}$$

where  $w_2(t)$  is a Brownian motion, is equivalent to the system described by (2.3) and (2.5) under the small angle and Gaussian assumptions. The equivalence is in the meaning that its optimal estimator (Kalman filter) is the same as (2.7) by replacing  $\langle J_z \rangle$  by the optimal estimation of  $z(t)$  and replacing  $d\bar{W}$  by a multiple of the innovation process. Specifically, the optimal estimation of  $z(t)$  and  $\langle J_z \rangle$  satisfies the same SDE when  $d\bar{W}$  is replaced by  $\frac{N_t}{\sqrt{\sigma_m}}$ , where  $N_t$  is the innovation process of the Kalman filter.

*Proof.* Under the Gaussian assumption, both  $\langle \Delta \hat{J}_z^3 \rangle$ , the third order moment of  $J_z$ , and  $\langle [\Delta \hat{J}_z^2, \hat{J}_y] \rangle$  vanish. If the spin angle  $\langle J_z \rangle / \langle J_x \rangle$  is kept small (the small angle assumption),  $\langle J_x \rangle(t) \cong J e^{-Mt/2} \cong J$  when  $t \ll 1/M$ , where  $J$  is the amplitude of the vector  $\mathbf{J}$  or the number of atoms. Thus (2.7) can be simplified to

$$(2.9) \quad \begin{aligned} d\langle J_z \rangle(t) &= \gamma J h(t) dt + \frac{\langle \Delta J_z^2 \rangle(t)}{\sqrt{\sigma_m}} d\bar{W}(t), \\ d\langle \Delta J_z^2 \rangle(t) &= -\frac{\langle \Delta J_z^2 \rangle^2(t)}{\sigma_m} dt. \end{aligned}$$

The last differential equation can be solved analytically:

$$\langle \Delta J_z^2 \rangle(t) = \frac{\langle \Delta J_z^2 \rangle(0) \sigma_m}{\sigma_m + \langle \Delta J_z^2 \rangle(0) t},$$

where  $\langle \Delta J_z^2 \rangle(0) = J/2$  for an initially coherent spin state.

Now, if we think  $\langle J_z \rangle$  is the optimal estimation of  $J_z$  and  $\bar{W}(t)$  is a multiple of the innovation process, then (2.9) has the structure of an optimal estimator.

The following shows the optimal estimation  $\tilde{z}(t)$  of the linear system described by (2.8) satisfying an SDE:

$$(2.10) \quad \begin{aligned} d\tilde{z}(t) &= \gamma J h(t) dt + \frac{\Sigma}{\sigma_m} (y(t) dt - \tilde{z}(t) dt), \\ \frac{d}{dt} \Sigma(t) &= -\frac{\Sigma^2}{\sigma_m}. \end{aligned}$$

We note that it is a Kalman filter. Comparing (2.10) and (2.9), we obtain what we want.  $\square$

We have two notes about the above theorem. First, although the sample used is small, the number of atoms involved is still large. Thus the Gaussian assumption is practical. The term  $\langle [\Delta \hat{J}_z^2, \hat{J}_z] h(t) \rangle$  can be ignored also because  $h(t)$  is close to

zero when feedback control is applied. Second, the equivalence of these two models is based on the fact that we are interested only in  $\langle J_z \rangle$ , which connects  $h(t)$  and  $y(t)$ .

The above theorem shows that we can work on a simpler linear SDE model instead of the formidable SME in (2.3) and (2.5). Both  $b(t)$  and  $u(t)$  are classical. We assume that  $b(t)$  is fluctuating and can be described by a stochastic process:

$$(2.11) \quad db(t) = -r_b b(t)dt + \sqrt{\sigma_{bf}}dw_1,$$

where  $w_1$  is another Brownian motion independent of  $w_2$  and  $r_b$  defines the bandwidth of  $b(t)$ . Then a two-dimensional process combining (2.8) and (2.11) can be obtained:

$$(2.12) \quad d\mathbf{x}(t) = A\mathbf{x}(t)dt + \begin{bmatrix} 0 \\ \sqrt{\sigma_{bf}} \end{bmatrix} dw_1,$$

$$(2.13) \quad y(t)dt = C\mathbf{x}(t)dt + \sqrt{\sigma_m}dw_2,$$

where we use  $\mathbf{x}$  to denote the state,

$$\mathbf{x} = \begin{bmatrix} z(t) \\ b(t) \end{bmatrix}.$$

The matrices are given as

$$(2.14) \quad A = \begin{bmatrix} 0 & \gamma J \\ 0 & -r_b \end{bmatrix}, \quad C = [1, 0].$$

The initial variance matrix of the state is

$$(2.15) \quad \Sigma_0 = \begin{bmatrix} \sigma_{z0} & 0 \\ 0 & \sigma_{b0} \end{bmatrix}.$$

We also define  $\Sigma_1$  and  $\Sigma_2$  as follows:

$$(2.16) \quad \Sigma_1 = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{bf} \end{bmatrix}, \quad \Sigma_2 = \sigma_m = \frac{1}{4}M\eta.$$

### 3. Steady state performance without feedback.

**3.1. Methods without using the Kalman filter.** The magnetic field can be measured without using the Kalman filter. If  $b(t)$  is not compensated by another field, the collective spin will rotate around the  $y$ -axis. If the magnetic field is strong and constant, an oscillating signal can be obtained before the spin damps significantly. Since the oscillation frequency (Larmor frequency) is proportional to the magnetic field, the amplitude of the field can be found using the signal's Fourier transform. But this is not always the case. If the field is weak, only a noisy sloped line of a small angle rotation can be recorded. By fitting the curve with a line, we can find an estimation of its slope, denoted as  $s$ . If we know  $J$ , the field can be retrieved through identity  $\tilde{b} = b = s/\gamma J$ , where  $\tilde{b}$  is the estimation.

However, the above slope method doesn't work well when  $J$  is unknown. If a guess  $J'$  is used instead of the real  $J$ , the actual estimation satisfies  $\tilde{b} = s/\gamma J' = (J/J')b$ .

The second moment and expectation of the estimation error now depend on  $J'$ :

$$(3.1) \quad E[(b - \tilde{b})^2] = \left(1 - \frac{1}{f}\right)^2 E[b^2],$$

$$(3.2) \quad E[b - \tilde{b}] = \left(1 - \frac{J}{J'}\right) E[b] = \left(1 - \frac{1}{f}\right) E[b],$$

where  $f = J'/J$ . If the magnetic field is constant, the error is systematic and can be calibrated away. If both  $b$  and  $J$  are random, the error is no longer systematic and the slope does not make much sense.

**3.2. A method using the Kalman filter.** The Kalman filter gives us the best estimation for a linear stochastic process with Gaussian-type noise. In this section, we will find the steady state performance of the Kalman filter for any  $J'$  picked, using the SDE model formed by (2.12) and (2.13). Because the linearization requires a small angle assumption, we constrain  $b(t)$  to be weak and possibly fluctuating so that  $\langle J_z \rangle / \langle J_x \rangle$  is small. We also assume that the time scale is short enough, so that the damping effect can be ignored, and long enough so that the estimation reaches its steady state. The result sets a standard for future comparison.

Finding the Kalman gain vector involves solving a matrix Riccati equation

$$(3.3) \quad \frac{d}{dt} \Sigma(t) = A' \Sigma + \Sigma A'^T - \Sigma C^T (\Sigma_2)^{-1} C \Sigma + \Sigma_1$$

with initial condition

$$\Sigma(0) = E[(\mathbf{x}(0) - \tilde{\mathbf{x}}(0))(\mathbf{x}(0) - \tilde{\mathbf{x}}(0))^T] = \Sigma_0,$$

where  $A'$  is the same as  $A$  except that  $J$  is replaced by  $J'$ . The Kalman gain vector then is obtained from  $\Sigma$  using identity

$$(3.4) \quad K_o(t) = \Sigma C^T (\Sigma_2)^{-1}.$$

In this article, we are interested in only the steady state performance of the estimation. The above differential equation can be replaced by an algebraic equation,

$$(3.5) \quad A' \Sigma + \Sigma A'^T - \Sigma C^T (\Sigma_2)^{-1} C \Sigma + \Sigma_1 = 0,$$

and  $K_o$  becomes constant. Denoting the two entries of  $K_o$  as  $K_z$  and  $K_b$  and solving the above equation, we have

$$(3.6) \quad K_o = \begin{bmatrix} K_z \\ K_b \end{bmatrix} = \begin{bmatrix} -r_b + \sqrt{r_b^2 + 2\gamma J' \sqrt{\frac{\sigma_{bf}}{\sigma_m}}} \\ \sqrt{\frac{\sigma_{bf}}{\sigma_m}} - \frac{r_b}{\gamma J'} \left( \sqrt{r_b^2 + 2\gamma J' \sqrt{\frac{\sigma_{bf}}{\sigma_m}}} - r_b \right) \end{bmatrix} \approx \begin{bmatrix} \sqrt{2J'} \sigma_{bf}^{1/4} \sigma_m^{-1/4} \\ \sigma_{bf}^{1/2} \sigma_m^{-1/2} \end{bmatrix}.$$

Note that  $J'$  instead of  $J$  appears in (3.6). The observer actually chooses  $J'$  in the design of the Kalman filter. The approximation is based on the *large  $J$  assumption*:  $r_b^2 \ll \gamma J \sqrt{\sigma_{bf}/\sigma_m}$ . Although  $J'$  may be different from  $J$ , it is reasonable to assume that they have the same order because the observer does have some idea about  $J$ .

For simplicity, we define  $\mathbb{J} = \gamma J$  and  $\mathbb{J}' = \gamma J'$  because only the products,  $\gamma J$  and  $\gamma J'$ , appear in our matrices.

The combination of the linear stochastic system and its Kalman filter forms a four-dimensional system. By defining  $\Delta z = z - \tilde{z}$  and  $\Delta b = b - \tilde{b}$ , where  $\tilde{b}$  and  $\tilde{z}$  are the optimal estimations, we obtain a three-dimensional system:

$$(3.7) \quad d\boldsymbol{\theta} = G\boldsymbol{\theta}dt + \beta \begin{bmatrix} dw_1 \\ dw_2 \end{bmatrix}.$$

The matrices and vectors are defined as

$$(3.8) \quad \boldsymbol{\theta} = \begin{bmatrix} \Delta z \\ \Delta b \\ b \end{bmatrix}, \quad G = \begin{bmatrix} -K_z & \mathbb{J}' & \mathbb{J} - \mathbb{J}' \\ -K_b & -r_b & 0 \\ 0 & 0 & -r_b \end{bmatrix},$$

$$\beta = \begin{bmatrix} 0 & -K_z\sqrt{\sigma_m} \\ \sqrt{\sigma_{bf}} & -K_b\sqrt{\sigma_m} \\ \sqrt{\sigma_{bf}} & 0 \end{bmatrix}.$$

A total state covariance matrix  $\Theta$  is defined as

$$(3.9) \quad \Theta = E[\boldsymbol{\theta}\boldsymbol{\theta}^T] = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{12} & \theta_{22} & \theta_{23} \\ \theta_{13} & \theta_{23} & \theta_{33} \end{bmatrix},$$

and its evolution satisfies

$$(3.10) \quad \frac{d\Theta}{dt} = G\Theta + \Theta G^T + \beta\beta^T.$$

The steady solution of  $\Theta$  satisfies an algebraic Riccati equation:

$$(3.11) \quad G\Theta + \Theta G^T + \beta\beta^T = 0.$$

The second moments of the estimation error can then be retrieved from  $\Theta$ :

$$(3.12) \quad \begin{aligned} \sigma_{bo} &= E[\Delta b^2] = \theta_{22}, \\ \sigma_{zo} &= E[\Delta z^2] = \theta_{11}. \end{aligned}$$

We use subscripts  $o$  and  $c$  to distinguish the open-loop case from the closed-loop case.

The solution of (3.11) gives us

$$(3.13) \quad \begin{aligned} \sigma_{zo} &= E[(z - \tilde{z})^2] = \theta_{11} \\ &= \frac{1}{4} \frac{4\sigma_{bf}\mathbb{J}^2 + K_z^2(3K_z^2 + 8K_z + 4r_b^2)\sigma_m}{K_z(K_z + r_b)(K_z + 2r_b)} \end{aligned}$$

and

$$(3.14) \quad \begin{aligned} \sigma_{bo} &= E[(b - \tilde{b})^2] = \theta_{22} \\ &= \frac{1}{8} \frac{K_z^4\sigma_m}{(K_z + r_b)\mathbb{J}'^2} + \frac{1}{2}\sigma_{bf} \left( \frac{K_z(K_z(\mathbb{J} - \mathbb{J}') - r_b\mathbb{J}')^2}{(K_z + r_b)(K_z + 2r_b)^2r_b\mathbb{J}'^2} + \frac{3K_z + 4r_b}{(K_z + 2r_b)^2} \right). \end{aligned}$$

We have used  $K_b = \frac{1}{2\mathbb{J}'} K_z^2$ .

The large  $J$  assumption implies that  $K_z \gg r_b$ , and this leads to an approximation of  $\sigma_{zo}$ :

$$\begin{aligned}
 \sigma_{zo} &= \frac{1}{4} \frac{4\sigma_m K_z^2 (K_z + r_b)^2 - \sigma_m K_z^4 + 4\sigma_{bf} \mathbb{J}^2}{K_z (K_z + 2r_b) (K_z + r_b)} \\
 (3.15) \quad &\approx \frac{3}{4} K_z \sigma_m + \sigma_{bf} \frac{\mathbb{J}^2}{K_z^3} \\
 &\approx \sigma_m^{3/4} \sigma_{bf}^{1/4} \left( \frac{3}{4} \sqrt{2\mathbb{J}'} + \frac{\mathbb{J}^2}{2\mathbb{J}' \sqrt{2\mathbb{J}'}} \right).
 \end{aligned}$$

THEOREM 3.1.  $\sigma_{zo}$  in (3.15) reaches its minimum when  $J = J'$ .

*Proof.* Let  $v = \sqrt{2\mathbb{J}'}$  and define  $g(v) = \frac{3}{4}v + \frac{\mathbb{J}^2}{v^3}$ . The first two derivatives of  $g$  can be computed as

$$(3.16) \quad g'(v) = \frac{3}{4} - \frac{3\mathbb{J}^2}{v^4}, \quad g''(v) = \frac{12\mathbb{J}^2}{v^5} > 0.$$

Equation  $g'(v) = 0$  has two solutions:  $v = \sqrt{2\mathbb{J}}$  and  $v = -\sqrt{2\mathbb{J}}$ . The last one can be discarded since  $v > 0$  according to its definition. Thus  $f$  obtains its minimum when  $v = \sqrt{2\mathbb{J}}$  since  $g''(v)$  is always positive. Identity  $v = \sqrt{2\mathbb{J}}$  implies  $\mathbb{J} = \mathbb{J}'$  and  $\sigma_z = \sqrt{2\mathbb{J}} \sigma_m^{3/4} \sigma_{bf}^{1/4}$ .  $\square$

Approximation of  $\sigma_{bo}$  can also benefit from the large  $J$  assumption:

$$(3.17) \quad \sigma_{bo} \approx \frac{K_z^3}{8\mathbb{J}'^2} \sigma_m + \frac{1}{2} \sigma_{bf} \left( \frac{3}{K_z} + \frac{(K_z(\mathbb{J} - \mathbb{J}') - r_b \mathbb{J}')^2}{K_z^2 \mathbb{J}'^2 r_b} \right).$$

By substituting  $K_z$  in terms of  $\sigma_{bf}$  and  $\sigma_m$ , we can further write  $\sigma_{bo}$  as

$$\sigma_{bo} \approx \sqrt{\frac{2}{\mathbb{J}'}} \sigma_{bf}^{3/4} \sigma_m^{1/4} + \frac{\sigma_{bf}}{2r_b} \left( \frac{1-f}{f} \right)^2 - \frac{\sigma_{bf}^{3/4} \sigma_m^{1/4}}{\sqrt{2\mathbb{J}'}} \left( \frac{1-f}{f} \right) + \frac{\sigma_{bf} r_b}{2K_z^2}.$$

The last term can be discarded as  $K_z \gg r_b$ , and the above expression becomes

$$(3.18) \quad \sigma_{bo} \approx \frac{\sigma_{bf}^{3/4} \sigma_m^{1/4}}{\sqrt{2\mathbb{J}'}} \left( \frac{3}{\sqrt{f}} - \frac{1}{f^{3/2}} \right) + \frac{\sigma_{bf}}{2r_b} \left( \frac{1-f}{f} \right)^2.$$

THEOREM 3.2. When  $J'$  is used in the design of the estimator (Kalman filter) and the control loop is open,  $\sigma_{bo}$  obtained for the linear system (2.12) can be approximated under the large  $J$  assumption by

$$(3.19) \quad \frac{1}{\sqrt{2\mathbb{J}'}} \sigma_{bf}^{3/4} \sigma_m^{1/4} \left( \frac{3}{\sqrt{f}} - \frac{1}{f^{3/2}} \right) + \frac{\sigma_{bf}}{2r_b} \left( \frac{1-f}{f} \right)^2,$$

where  $f = \mathbb{J}'/\mathbb{J}$ . Specifically, when  $\mathbb{J} = \mathbb{J}'$ ,

$$\begin{aligned}
 \sigma_{zo} &= \sqrt{2\mathbb{J}} \sigma_{bf}^{1/4} \sigma_m^{3/4}, \\
 (3.20) \quad \sigma_{bo} &= \sqrt{\frac{2}{\mathbb{J}}} \sigma_{bf}^{3/4} \sigma_m^{1/4}.
 \end{aligned}$$

*Proof.* We have already shown most of the theorem. For the specific case of  $\sigma_{bo}$  when  $\mathbb{J} = \mathbb{J}'$ , replacing  $f$  by 1 in (3.18), we obtain

$$\sigma_{bo} = \sqrt{\frac{2}{\mathbb{J}}} \sigma_{bf}^{3/4} \sigma_m^{1/4}. \quad \square$$

Because of the large  $J$  assumption,  $r_b^2 \ll \mathbb{J}\sqrt{\sigma_{bf}/\sigma_m}$ , the second term in (3.19) quickly dominates the first one when  $f$  is away from 1, which coincides with the result of the Caltech group. It vanishes only when  $f = 1$ , so  $\sigma_{bo}$  obtains its minimum near  $f = 1$ .

**4. Steady state performance with noised feedback.** When the control loop is closed and  $u(t)$  has infinite accuracy, feedback makes the measurement robust to an unknown parameter, the number of atoms involved. Choosing smaller  $J'$  in the design of filter and controller seems beneficial because the measurement is more robust with smaller  $J'$ .

Dropping the noise-free assumption raises concerns about the measurement. We need to know how much the input noise affects the measurement quantitatively and to find when the noise can be ignored. To investigate its effect, we extend the previous system by an input noise term, use a Kalman filter to estimate the magnetic field  $b(t)$ , and choose an LQG controller to find the compensative field  $u(t)$ . Then we proceed to derive the steady state performance of the magnetometer, i.e., the second moments of the estimation errors. As before, we reduce the system dimension, by defining  $\Delta z = z - \tilde{z}$  and  $\Delta b = b - \tilde{b}$ , and simplify the derivation by invoking the large  $J$  assumption and the large  $\lambda$  assumption. The last one will be explained later.

**4.1. The system, the Kalman filter, and the LQG controller.** Besides perturbation from external electromagnetic radiation, input noise arises also from inside the electric circuit in such forms as shot noise, thermal noise, truncation error in the computation, and quantization noise in the digital-to-analog (DA) converters. In practice, the input signal is always accompanied by uncertainty. This becomes serious especially when the input signal is small.

Magnitude varies from one kind of noise to another and highly depends on the specific equipment used. Thermal noise, the fluctuating charges across a resistor, for example, can be modeled as a combination of a noiseless resistor and a shunt current source connected in parallel. The mean square value of the current, a measurement of the magnitude of the noise [6], is

$$E[i^2] = 4KT\Delta fG,$$

where  $T$  is the absolute temperature in Kelvin,  $K$  is the Boltzmann's constant,  $G = 1/R$ , and  $\Delta f$  is the bandwidth of the circuit.

The driver circuit which provides current through the coil usually is a cascade of several amplifiers. The mean square value of the shot noise across a semiconductor junction is proportional to the direct current passing through it,  $I$ ,

$$E[i^2] = 2eI\Delta f,$$

where  $e$  is the electron charge. The shot noise in the first amplifier clearly is the most important since it is amplified in the following stages.

When discrete values are used to replace continuous values, truncation errors bring extra noise to the algorithm. Besides this, quantization error arises in the analog-to-digital (AD) and DA converters. It is nonlinear and signal dependent. The signal-to-noise ratio is low for low-resolution DA converters. Even for high-resolution converters, the signal-to-noise ratio deteriorates when the signal is small. This happens when the measured field is weak and the control loop maintains only a very small current.

There are also other sources of noise. Detailed evaluation of the noise level is time consuming and heavily depends on the specific equipment and its environment. Unfortunately, since this is a new research area and a new experiment setup is used, there is no detailed noise analysis available yet to our knowledge. Experiments published have shown that the signal is strongly polluted by noise, and deeper investigation is necessary.

The input noise appears in the  $h(t)$  term of (2.3) as a stochastic process. When the bandwidth of the noise is wide compared with that of the system, it is equivalent to an additional white noise added to  $u(t)$  in (2.12). Thus we introduce an input noise  $\sqrt{\sigma_{uf}}dw_3$  into the process where  $w_3$  is a Brownian motion independent of  $w_1$  and  $w_2$ , and this changes the original system described by (2.12) and (2.11) to

$$(4.1) \quad \begin{aligned} d\mathbf{x}(t) &= A\mathbf{x}(t)dt + Bu(t)dt + \begin{bmatrix} \sqrt{\sigma_{uf}} & 0 \\ 0 & \sqrt{\sigma_{bf}} \end{bmatrix} \begin{bmatrix} dw_3 \\ dw_1 \end{bmatrix}, \\ y(t)dt &= C\mathbf{x}(t)dt + \sqrt{\sigma_m}dw_2(t). \end{aligned}$$

The matrices are listed as

$$(4.2) \quad A = \begin{bmatrix} 0 & \mathbb{J} \\ 0 & -r_b \end{bmatrix}, \quad B = \begin{bmatrix} \mathbb{J} \\ 0 \end{bmatrix}, \quad C = [1, 0].$$

A constant feedback gain vector is used in our model:

$$(4.3) \quad u(t)dt = -K_c\tilde{x}(t)dt,$$

where  $K_c$  is the controller gain vector. Similarly, only the steady solution of the filter gain vector is used.

LEMMA 4.1. *The steady solution of the Kalman filter gain vector  $K_o$  of the linear system defined by (4.1) is*

$$(4.4) \quad \begin{aligned} K_z &= -r_b + \sqrt{r_b^2 + \sigma_{uf}\sigma_m^{-1} + 2\sqrt{\mathbb{J}'^2\sigma_{bf}\sigma_m^{-1} + r_b^2\sigma_{uf}\sigma_m^{-1}}}, \\ K_b &= \frac{1}{2\mathbb{J}'}(-\sigma_{uf}\sigma_m^{-1} + K_z^2) \end{aligned}$$

when  $J'$  instead of  $J$  is used in the design of the Kalman filter. Under the large  $J$  and  $\lambda$  assumption, the LQG controller gain vector  $K_c$  can be approximated by

$$(4.5) \quad K_c \approx [\lambda, 1],$$

where  $\lambda = \sqrt{q/r}$  for a cost function defined by

$$\mathcal{J} = \int_0^\infty (qz(t)^2 + ru(t)^2)dt.$$

*Proof.* We first find  $K_c$  [5, 4]. It satisfies

$$K_c \equiv r^{-1}B^TV,$$

where  $V$  is the solution of

$$(4.6) \quad P + A'^TV + VA' - \frac{1}{r}VB'B'^TV = 0.$$

The matrices  $A'$  and  $B'$  are the same as  $A$  and  $B$  except that  $J$  is replaced by  $J'$ . Matrix  $P$  is defined from the cost function:

$$P = \begin{bmatrix} q & 0 \\ 0 & 0 \end{bmatrix}.$$

The solution of (4.6) leads to

$$(4.7) \quad K_c = \left[ \lambda, \frac{\mathbb{J}'\lambda}{r_b + \mathbb{J}'\lambda} \right].$$

While the original large  $\lambda$  assumption in [16] implies that  $\lambda^2 \gg \sqrt{\sqrt{\sigma_{bf}/\sigma_m}/(2\mathbb{J})}$ , in this article, we change it to  $\lambda \gg \sqrt{\sqrt{\sigma_{bf}/\sigma_m}/(2\mathbb{J})}$ . This does not make a significant difference in the experiment but makes our derivation much easier. Combined with the large  $J$  assumption, our large  $\lambda$  assumption leads to  $\lambda\mathbb{J} \gg \sqrt{\sqrt{\sigma_{bf}/\sigma_m}\mathbb{J}} \gg r_b$ . Thus  $K_c$  can be approximated by

$$(4.8) \quad K_c \approx [\lambda, 1]$$

as we wanted.

Using the formulae of the Kalman filter, we obtain

$$K_o = \Sigma(t)C^T\Sigma_2^{-1}$$

and

$$\frac{d}{dt}\Sigma(t) = \Sigma_1 + A'\Sigma(t) + \Sigma(t)A'^T - \Sigma(t)C^T\Sigma_2^{-1}C\Sigma(t),$$

where  $\Sigma_2$  is defined in (2.16) as before, but  $\Sigma_1$  changes to

$$\Sigma_1 = \begin{bmatrix} \sigma_{uf} & 0 \\ 0 & \sigma_{bf} \end{bmatrix}.$$

Other matrices are defined as before. Let  $\Sigma(t)$  be constant and denoted as

$$\Sigma = \begin{bmatrix} \sigma_{zs} & \sigma_{cs} \\ \sigma_{cs} & \sigma_{bs} \end{bmatrix}.$$

We find a set of equations

$$(4.9) \quad \begin{cases} \sigma_{uf} + 2\mathbb{J}'\sigma_{cs} - \sigma_{zs}^2\sigma_m^{-1} = 0, \\ \mathbb{J}'\sigma_{bs} - r_b\sigma_{cs} - \sigma_{cs}\sigma_{zs}\sigma_m^{-1} = 0, \\ \sigma_{bf} - 2r_b\sigma_{bs} - \sigma_{cs}^2\sigma_m^{-1} = 0. \end{cases}$$

From among the three unknown entries, we need only  $\sigma_{zs}$  and  $\sigma_{cs}$ :

$$(4.10) \quad \begin{aligned} \sigma_{zs} &= -r_b\sigma_m + \sigma_m\sqrt{r_b^2 + \sigma_{uf}\sigma_m^{-1} + 2\sqrt{\mathbb{J}'^2\sigma_{bf}\sigma_m^{-1} + r_b^2\sigma_{uf}\sigma_m^{-1}}}, \\ \sigma_{cs} &= \frac{1}{2\mathbb{J}'}(-\sigma_{uf} + \sigma_{zs}^2\sigma_m^{-1}). \end{aligned}$$

The entries of  $K_o$  are then obtained from  $\sigma_{zs}$  and  $\sigma_{cs}$ :

$$(4.11) \quad \begin{aligned} K_z = \frac{\sigma_{zs}}{\sigma_m} &= -r_b + \sqrt{r_b^2 + \sigma_{uf}\sigma_m^{-1} + 2\sqrt{\mathbb{J}'^2\sigma_{bf}\sigma_m^{-1} + r_b^2\sigma_{uf}\sigma_m^{-1}}}, \\ K_b = \frac{\sigma_{cs}}{\sigma_m} &= \frac{1}{2\mathbb{J}'}(-\sigma_{uf}\sigma_m^{-1} + K_z^2). \quad \square \end{aligned}$$

There are several useful identities about  $K_z$  and  $K_b$ , and we list them in the following lemmas.



LEMMA 4.2.

$$(4.12) \quad \sigma_{uf} = K_z^2 \sigma_m - 2\mathbb{J}' K_b \sigma_m,$$

$$(4.13) \quad \mathbb{J}' \sigma_{bf} = (\mathbb{J}' K_b^2 + 2r_b^2 K_b + 2r_b K_b K_z) \sigma_m,$$

$$(4.14) \quad K_z r_b + K_b \mathbb{J}' = \sqrt{\mathbb{J}'^2 \sigma_{bf} \sigma_m^{-1} + r_b^2 \sigma_{uf} \sigma_m^{-1}}.$$

*Proof.* Identity (4.12) is just the first equation in (4.9) in terms of  $K_z$  and  $K_b$ . We need only prove (4.13) and (4.14). Using (4.12) and (4.11), we have

$$(4.15) \quad \begin{aligned} K_z r_b + K_b \mathbb{J}' &= \frac{1}{2} (2K_z r_b + K_z^2 - \sigma_{uf} \sigma_m^{-1}) \\ &= \frac{1}{2} ((K_z + r_b)^2 - r_b^2 - \sigma_{uf} \sigma_m^{-1}) \\ &= \sqrt{\mathbb{J}'^2 \sigma_{bf} \sigma_m^{-1} + r_b^2 \sigma_{uf} \sigma_m^{-1}}, \end{aligned}$$

and this proves (4.14). For (4.13), look into identity

$$(4.16) \quad \begin{aligned} \sigma_{bf} &= \sigma_{cs}^2 \sigma_m^{-1} + 2r_b \sigma_{bs} \\ &= \sigma_{cs}^2 \sigma_m^{-1} + 2r_b \frac{1}{\mathbb{J}'} (r_b \sigma_{cs} + \sigma_{cs} \sigma_{zs} \sigma_m^{-1}), \end{aligned}$$

which is equivalent to

$$\mathbb{J}' \sigma_{bf} = \mathbb{J}' K_b^2 \sigma_m + 2r_b^2 K_b \sigma_m + 2r_b K_b K_z \sigma_m. \quad \square$$

LEMMA 4.3.

$$(4.17) \quad \frac{(K_b^2 \sigma_m + \sigma_{bf})(r_b K_z + \mathbb{J}' K_b) + K_b^2 \sigma_{uf} + K_z^2 \sigma_{bf}}{2(K_z r_b + K_b \mathbb{J}')(K_z + r_b)} = \frac{K_b(K_z + r_b) \sigma_m}{\mathbb{J}'}.$$

*Proof.* Using Lemma 4.2, we have

$$(4.18) \quad \begin{aligned} &2(K_z r_b + K_b \mathbb{J}')(K_z + r_b)^2 K_b \sigma_m - \mathbb{J}'(\sigma_m K_b^2 + \sigma_{bf})(r_b K_z + \mathbb{J}' K_b) \\ &= (K_z r_b + \mathbb{J}' K_b)(2K_b(K_z + r_b)^2 \sigma_m - \mathbb{J}'(K_b^2 \sigma_m + \sigma_{bf})) \\ &= (K_z r_b + \mathbb{J}' K_b)(2K_b(K_z + r_b)^2 - 2(\mathbb{J}' K_b^2 + r_b^2 K_b + r_b K_b K_z)) \sigma_m \\ &= 2(K_z r_b + \mathbb{J}' K_b)(K_b(K_z^2 + K_z r_b - K_b \mathbb{J}')) \sigma_m \\ &= 2K_b(r_b K_z^3 + r_b^2 K_z^2 + \mathbb{J}' K_b K_z^2 - \mathbb{J}'^2 K_b^2) \sigma_m \end{aligned}$$

and

$$(4.19) \quad \begin{aligned} &K_b^2 \mathbb{J}' \sigma_{uf} + K_z^2 \mathbb{J}' \sigma_{bf} \\ &= \sigma_m (K_b^2 \mathbb{J}' (K_z^2 - 2\mathbb{J}' K_b) + K_z^2 (K_b^2 \mathbb{J}' + 2r_b^2 K_b + 2r_b K_b K_z)) \\ &= \sigma_m (\mathbb{J}' K_z^2 K_b^2 - 2\mathbb{J}'^2 K_b^3 + \mathbb{J}' K_z^2 K_b^2 + 2r_b^2 K_b K_z^2 + 2r_b K_b K_z^3) \\ &= 2K_b(r_b K_z^3 + r_b^2 K_z^2 + \mathbb{J}' K_b K_z^2 - \mathbb{J}'^2 K_b^2) \sigma_m. \end{aligned}$$

Thus

$$(4.20) \quad \begin{aligned} &2(K_z r_b + K_b \mathbb{J}')(K_z + r_b)^2 K_b \sigma_m - \mathbb{J}'(\sigma_m K_b^2 + \sigma_{bf})(r_b K_z + \mathbb{J}' K_b) \\ &= K_b^2 \mathbb{J}' \sigma_{uf} + K_z^2 \mathbb{J}' \sigma_{bf}, \end{aligned}$$

and that is equivalent to (4.17).  $\square$

LEMMA 4.4. If  $\mathbb{J} \gg r_b^2 \sqrt{\frac{\sigma_m}{\sigma_{bf}}}$  (large  $J$  assumption),  $K_z \gg r_b$ .

*Proof.* If  $\beta \gg \alpha > 0$ , then

$$\sqrt{\alpha + \beta} - \sqrt{\alpha} = \sqrt{\alpha}(\sqrt{1 + \beta/\alpha} - 1) \approx \sqrt{\beta} \gg \sqrt{\alpha}.$$

Letting  $\alpha = r_b^2$  and  $\beta = 2\sqrt{\mathbb{J}'^2 \sigma_{bf}/\sigma_m + r_b^2 \sigma_{uf}/\sigma_m} + \sigma_{uf}/\sigma_m$ , we obtain what we want.  $\square$

The large  $J$  assumption implies that  $r_b^2 \ll \sqrt{\mathbb{J}' \sigma_{bf} \sigma_m^{-1} + r_b^2 \sigma_{uf} \sigma_m^{-1}}$ , leading to an approximation of  $K_z$ :

$$(4.21) \quad K_z \approx \sqrt{\sigma_{uf} \sigma_m + 2\sqrt{\mathbb{J}'^2 \sigma_{bf} \sigma_m^3 + r_b^2 \sigma_{uf} \sigma_m^3}} \frac{1}{\sigma_m}$$

and an approximation of  $K_b$ :

$$(4.22) \quad K_b \approx \sqrt{\sigma_{bf} \sigma_m + r_b^2 \mathbb{J}'^{-2} \sigma_{uf} \sigma_m} \frac{1}{\sigma_m}.$$

**4.2. Derivation of the steady state performance.** In this subsection, we will find the steady state performance of the measurement under the influence of the input noise and simplify the results by invoking the large  $\lambda$  and  $J$  assumptions. We start with combining the system, the estimator, and the controller together to form a four-dimensional system:

$$(4.23) \quad d \begin{bmatrix} z(t) \\ b(t) \\ \tilde{z}(t) \\ \tilde{b}(t) \end{bmatrix} = \begin{bmatrix} 0 & \mathbb{J} & -\mathbb{J}\lambda & -\mathbb{J} \\ 0 & -r_b & 0 & 0 \\ K_z & 0 & -\mathbb{J}'\lambda - K_z & 0 \\ K_b & 0 & -K_b & -r_b \end{bmatrix} \begin{bmatrix} z(t) \\ b(t) \\ \tilde{z}(t) \\ \tilde{b}(t) \end{bmatrix} dt + \begin{bmatrix} \sqrt{\sigma_{uf}} & & & \\ & \sqrt{\sigma_{bf}} & & \\ & & K_z \sqrt{\sigma_m} & \\ & & K_b \sqrt{\sigma_m} & \end{bmatrix} \begin{bmatrix} dw_3 \\ dw_2 \\ dw_1 \end{bmatrix}.$$

As before, we define  $\Delta z = z - \tilde{z}$  and  $\Delta b = b - \tilde{b}$ , and a three-dimensional system is obtained as

$$(4.24) \quad d \begin{bmatrix} \Delta z \\ \Delta b \\ \tilde{z} \end{bmatrix} = \begin{bmatrix} -K_z & \mathbb{J} & \lambda(\mathbb{J}' - \mathbb{J}) \\ -K_b & -r_b & 0 \\ K_z & 0 & -\lambda\mathbb{J}' \end{bmatrix} \begin{bmatrix} \Delta z dt \\ \Delta b dt \\ \tilde{z} dt \end{bmatrix} + \begin{bmatrix} \sqrt{\sigma_{uf}} & 0 & -K_z \sqrt{\sigma_m} \\ 0 & \sqrt{\sigma_{bf}} & -K_b \sqrt{\sigma_m} \\ 0 & 0 & K_z \sqrt{\sigma_m} \end{bmatrix} \begin{bmatrix} dw_3 \\ dw_1 \\ dw_2 \end{bmatrix}.$$

The vector and matrices are defined as

$$(4.25) \quad \theta = \begin{bmatrix} \Delta z \\ \Delta b \\ \tilde{z} \end{bmatrix}, \quad \alpha = \begin{bmatrix} -K_z & \mathbb{J} & \lambda(\mathbb{J}' - \mathbb{J}) \\ -K_b & -r_b & 0 \\ K_z & 0 & -\lambda\mathbb{J}' \end{bmatrix},$$

$$\beta = \begin{bmatrix} \sqrt{\sigma_{uf}} & 0 & -K_z \sqrt{\sigma_m} \\ 0 & \sqrt{\sigma_{bf}} & -K_b \sqrt{\sigma_m} \\ 0 & 0 & K_z \sqrt{\sigma_m} \end{bmatrix}.$$

As before, the dynamic equation of the total state covariance satisfies

$$(4.26) \quad \frac{d\Theta}{dt} = \alpha\Theta + \Theta\alpha^T + \beta\beta^T,$$

where  $\Theta = E[\theta\theta^T]$ . Its steady solution can be found by solving

$$(4.27) \quad 0 = \alpha\Theta + \Theta\alpha^T + \beta\beta^T.$$

Because  $\Theta$  is symmetric, we can assume

$$\Theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{12} & \theta_{22} & \theta_{23} \\ \theta_{13} & \theta_{23} & \theta_{33} \end{bmatrix},$$

and (4.27) is transformed into six linear equations:

$$(4.28) \quad \begin{cases} -K_z\theta_{11} + \mathbb{J}\theta_{12} + \lambda(\mathbb{J}' - \mathbb{J})\theta_{13} = -(\sigma_{uf} + K_z^2\sigma_m)/2, \\ -K_z\theta_{12} + \mathbb{J}\theta_{22} + \lambda(\mathbb{J}' - \mathbb{J})\theta_{23} - K_b\theta_{11} - r_b\theta_{12} = -K_zK_b\sigma_m, \\ -K_z\theta_{13} + \mathbb{J}\theta_{23} + \lambda(\mathbb{J}' - \mathbb{J})\theta_{33} + K_z\theta_{11} - \lambda\mathbb{J}'\theta_{13} = K_z^2\sigma_m, \\ K_b\theta_{12} + r_b\theta_{22} = (\sigma_{bf} + K_b^2\sigma_m)/2, \\ -K_b\theta_{13} - r_b\theta_{23} + K_z\theta_{12} - \lambda\mathbb{J}'\theta_{23} = K_zK_b\sigma_m, \\ K_z\theta_{13} - \lambda\mathbb{J}'\theta_{33} = -K_z^2\sigma_m/2. \end{cases}$$

We need only  $\sigma_{bc} = E[(\Delta b)^2] = \theta_{22}$ . Its solution can be written as a rational function:

$$(4.29) \quad \theta_{22} = \frac{Poly_N}{Poly_D},$$

where

$$(4.30) \quad Poly_D = 2\mathbb{J}(r_bK_z + \mathbb{J}'K_b)((\mathbb{J}'K_b + (K_z + \lambda\mathbb{J}')(r_b + \lambda\mathbb{J}'))(K_z + r_b) \\ + (\mathbb{J} - \mathbb{J}')(\lambda K_z(\lambda\mathbb{J}' + K_z) + K_b(r_b + K_z))),$$

and

$$(4.31) \quad \begin{aligned} Poly_N &= \mathbb{J}((\sigma_m K_b^2 + \sigma_{bf})(r_b K_z + \mathbb{J}'K_b) + K_b^2\sigma_{uf} + K_z^2\sigma_{bf}) \\ &\quad (\mathbb{J}'K_b + (K_z + \lambda\mathbb{J}')(r_b + \lambda\mathbb{J}')) \\ &\quad + (\mathbb{J} - \mathbb{J}')(\lambda\mathbb{J}\sigma_{bf}K_z^3 + \lambda\mathbb{J}\sigma_m r_b K_z^2 K_b^2 + \lambda^2\mathbb{J}'\sigma_{bf}K_z^2 \\ &\quad + r_b\mathbb{J}\sigma_m K_z K_b^3 + \lambda\mathbb{J}\mathbb{J}'\sigma_m K_z K_b^3 - \lambda\mathbb{J}'\sigma_{uf}K_z K_b^2 \\ &\quad + r_b\mathbb{J}\sigma_{bf}K_z K_b - \lambda\mathbb{J}\mathbb{J}'\sigma_{bf}K_z K_b + \mathbb{J}\mathbb{J}'\sigma_m K_b^4 + \mathbb{J}\mathbb{J}'\sigma_{bf}K_b^2 \\ &\quad - \lambda\mathbb{J}'\sigma_{uf}r_b K_b^2 - \lambda^2\mathbb{J}'^2\sigma_{uf}K_b^2) \\ &= \mathbb{J}(\dots) \\ &\quad + (\mathbb{J} - \mathbb{J}')(\mathbb{J}(K_b^2\sigma_m + \sigma_{bf})K_b(r_b K_z + \mathbb{J}'K_b) \\ &\quad - 2\lambda\frac{\mathbb{J}}{\mathbb{J}'}K_z K_b r_b (K_z + r_b)(r_b K_z + \mathbb{J}'K_b)\sigma_m \\ &\quad + \lambda(K_z + r_b + \lambda\mathbb{J}')(\mathbb{J}K_z^2\sigma_{bf} - \mathbb{J}'K_b^2\sigma_{uf})). \end{aligned}$$

Observing that

$$\frac{\alpha\mathcal{A} + \beta\mathcal{B}}{\alpha + \beta\mathcal{C}} = \mathcal{A} + \frac{\beta(\mathcal{B} - \mathcal{A}\mathcal{C})}{\alpha + \beta\mathcal{C}},$$

comparing it with (4.29)–(4.31), and using Lemma 4.3, we find

$$(4.32) \quad \theta_{22} = \frac{K_b(K_z + r_b)}{\mathbb{J}'} \sigma_m + \frac{(\mathbb{J} - \mathbb{J}') \text{Poly}_N^2}{\text{Poly}_D},$$

where

$$(4.33) \quad \begin{aligned} \text{Poly}_N^2 = & \mathbb{J}(K_b^2 \sigma_m + \sigma_{bf}) K_b(r_b K_z + \mathbb{J}' K_b) \\ & - 2\lambda \frac{\mathbb{J}}{\mathbb{J}'} K_z K_b r_b (K_z + r_b)(r_b K_z + K_b \mathbb{J}') \sigma_m \\ & + \lambda(K_z + r_b + \lambda \mathbb{J}')(\mathbb{J} K_z^2 \sigma_{bf} - \mathbb{J}' K_b^2 \sigma_{uf}) \\ & - \frac{K_b(K_z + r_b)}{\mathbb{J}'} \sigma_m (K_b r_b + K_b K_z + \lambda^2 \mathbb{J}' K_z + \lambda K_z^2) 2\mathbb{J}(K_z r_b + \mathbb{J}' K_b). \end{aligned}$$

Since

$$(4.34) \quad \begin{aligned} \mathbb{J}(K_b^2 \sigma_m + \sigma_{bf}) &= \frac{\mathbb{J}}{\mathbb{J}'} (K_b^2 \mathbb{J}' \sigma_m + \mathbb{J}' \sigma_{bf}) \\ &= \frac{2\mathbb{J}}{\mathbb{J}'} (K_b^2 \mathbb{J}' + r_b^2 K_b + r_b K_b K_z) \sigma_m, \end{aligned}$$

the first term in (4.33) can be rewritten in terms of  $\sigma_m$ . By collecting all  $\sigma_m$  terms together, we can rewrite  $\text{Poly}_N^2$  as

$$(4.35) \quad \begin{aligned} \text{Poly}_N^2 = & 2 \frac{\mathbb{J}}{\mathbb{J}'} \sigma_m K_b(r_b K_z + \mathbb{J}' K_b) (\lambda(K_z + r_b) K_z (-r_b - K_z - \lambda \mathbb{J}') \\ & + K_b^2 \mathbb{J}' - K_b K_z^2 - K_b K_z r_b) \\ & + \frac{\mathbb{J}}{\mathbb{J}'} (K_z + r_b + \lambda \mathbb{J}') \lambda \left( \mathbb{J}' K_z^2 \sigma_{bf} - \frac{\mathbb{J}^2}{\mathbb{J}} K_b^2 \sigma_{uf} \right). \end{aligned}$$

Using Lemma 4.2, we can further write  $\text{Poly}_N^2$  as a multiple of  $\sigma_m$ :

$$(4.36) \quad \begin{aligned} \frac{\text{Poly}_N^2}{\sigma_m} = & \lambda K_b^2 (K_z + r_b + \lambda \mathbb{J}') (2\mathbb{J}'^2 K_b - \mathbb{J}' K_z^2 - 2r_b \mathbb{J} K_z - \mathbb{J} K_z^2) \\ & + 2 \frac{\mathbb{J}}{\mathbb{J}'} K_b (K_b \mathbb{J}' + r_b K_z) (K_b^2 \mathbb{J}' - K_b K_z^2 - r_b K_z K_b). \end{aligned}$$

We can do the same for  $\theta_{22}$ :

$$(4.37) \quad \frac{\theta_{22}}{\sigma_m} = \frac{K_b(K_z + r_b)}{\mathbb{J}'} + \frac{(\mathbb{J} - \mathbb{J}') \text{Poly}_N^2}{\text{Poly}_D \sigma_m}.$$

Before we use the large  $\lambda$  and  $J$  assumptions to simplify (4.37), we first look into two identities:

$$(4.38) \quad \begin{aligned} & 2\mathbb{J}'^2 K_b - \mathbb{J}' K_z^2 - \mathbb{J} K_z^2 - 2r_b \mathbb{J} K_z \\ &= \mathbb{J}' (2\mathbb{J}' K_b - K_z^2) - \mathbb{J} K_z (K_z + 2r_b) \\ &= -\mathbb{J}' \sigma_{uf} \sigma_m^{-1} - \mathbb{J} K_z (K_z + 2r_b) \\ &= -(\mathbb{J} + \mathbb{J}') \sigma_{uf} \sigma_M^{-1} - 2\mathbb{J} \sqrt{T} \end{aligned}$$

and

$$\begin{aligned}
 & K_b^2 \mathbb{J}' - K_b K_z^2 - r_b K_z K_b \\
 &= K_b (\mathbb{J}' K_b - K_z^2 - r_b K_z) \\
 &= K_b \left( \frac{1}{2} (K_z^2 - \sigma_{uf} \sigma_m^{-1}) - K_z^2 - r_b K_z \right) \\
 (4.39) \quad &= K_b \left( -\frac{1}{2} (K_z^2 + 2r_b K_z + r_b^2) + \frac{r_b^2}{2} - \frac{\sigma_{uf}}{2\sigma_m} \right) \\
 &= \frac{K_b}{2} \left( -2 \frac{\sigma_{uf}}{\sigma_m} - 2\sqrt{T} \right),
 \end{aligned}$$

where  $T = \mathbb{J}'^2 \sigma_{bf} \sigma_m^{-1} + r_b^2 \sigma_{uf} \sigma_m^{-1}$ . Now (4.36) can be rewritten as

$$\begin{aligned}
 \frac{Poly_N^2}{\sigma_m} &= - \left( \lambda K_b^2 \mathbb{J} \left( \left( 1 + \frac{\mathbb{J}'}{\mathbb{J}} \right) \frac{\sigma_{uf}}{\sigma_m} + 2\sqrt{T} \right) (K_z + r_b + \lambda \mathbb{J}') \right. \\
 (4.40) \quad &\quad \left. + \frac{\mathbb{J}}{\mathbb{J}'} K_b^2 (r_b K_z + K_b \mathbb{J}') \left( 2 \frac{\sigma_{uf}}{\sigma_m} + 2\sqrt{T} \right) \right) \\
 &= -2K_b^2 \left( \frac{\sigma_{uf}}{\sigma_m} + \sqrt{T} \right) \left( \frac{1}{f} \sqrt{T} + \lambda \mathbb{J} (K_z + r_b + \lambda \mathbb{J}') \right) \\
 &\quad + (1-f) \lambda K_b^2 \mathbb{J} (K_z + r_b + \lambda \mathbb{J}') \frac{\sigma_{uf}}{\sigma_m},
 \end{aligned}$$

where  $f = \mathbb{J}'/\mathbb{J}$ .

The denominator  $Poly_D$  can be simplified by reorganizing its terms first:

$$\begin{aligned}
 Poly_D &= 2\mathbb{J} (K_z r_b + \mathbb{J}' K_b) ((K_z + r_b)(K_z r_b + \mathbb{J}' K_b) \\
 &\quad + \lambda \mathbb{J}' (K_z + r_b + \lambda \mathbb{J}') (K_z + r_b) \\
 &\quad + (\mathbb{J} - \mathbb{J}') (\lambda K_z (\lambda \mathbb{J}' + K_z) + K_b (r_b + K_z))) \\
 (4.41) \quad &= 2\mathbb{J} \sqrt{T} ((K_z + r_b) \sqrt{T} + \lambda \mathbb{J}' (K_z + r_b)^2 + \lambda^2 \mathbb{J}'^2 (K_z + r_b) \\
 &\quad + (\mathbb{J} - \mathbb{J}') (\lambda K_z (\lambda \mathbb{J}' + K_z) + K_b (r_b + K_z))) \\
 &= 2\mathbb{J} \sqrt{T} ((K_z + r_b) (\sqrt{T} + (\mathbb{J} - \mathbb{J}') K_b) \\
 &\quad + \lambda (\mathbb{J} K_z^2 + \lambda \mathbb{J}' (\mathbb{J} K_z + \mathbb{J}' r_b) + \mathbb{J}' r_b^2 + 2\mathbb{J}' r_b K_z)) \\
 &\approx 2\mathbb{J} \sqrt{T} \left( K_z \left( \sqrt{T} + \left( \frac{1}{f} - 1 \right) \sqrt{T} \right) + \lambda \mathbb{J} K_z (K_z + \lambda \mathbb{J}') \right) \\
 &\approx 2\mathbb{J} \sqrt{T} K_z \left( \frac{\mathbb{J}}{\mathbb{J}'} \sqrt{T} + \lambda \mathbb{J} (K_z + \lambda \mathbb{J}') \right).
 \end{aligned}$$

We have used (4.21) and the large  $J$  assumption in the above approximation.

Combining (4.40) and (4.41), we obtain

$$(4.42) \quad \frac{Poly_N^2}{Poly_D} = - \frac{K_b^2 (\sigma_{uf} + \sigma_m \sqrt{T})}{\mathbb{J} \sqrt{T} K_z} + (f-1) \frac{\lambda K_b^2 (K_z + r_b + \lambda \mathbb{J}') \sigma_{uf}}{2\mathbb{J} \sqrt{T} K_z (\frac{\sqrt{T}}{\mathbb{J}'} + \lambda (K_z + \lambda \mathbb{J}'))}.$$

The second term in the above equation can be made simpler if  $\lambda (K_z + \lambda \mathbb{J}') \gg \frac{\sqrt{T}}{\mathbb{J}'}$ :

$$(4.43) \quad \frac{\lambda K_b^2 (K_z + r_b + \lambda \mathbb{J}') \sigma_{uf}}{2\mathbb{J} \sqrt{T} K_z (\frac{\sqrt{T}}{\mathbb{J}'} + \lambda (K_z + \lambda \mathbb{J}'))} \approx \frac{K_b^2 \sigma_{uf}}{2\mathbb{J} \sqrt{T} K_z},$$

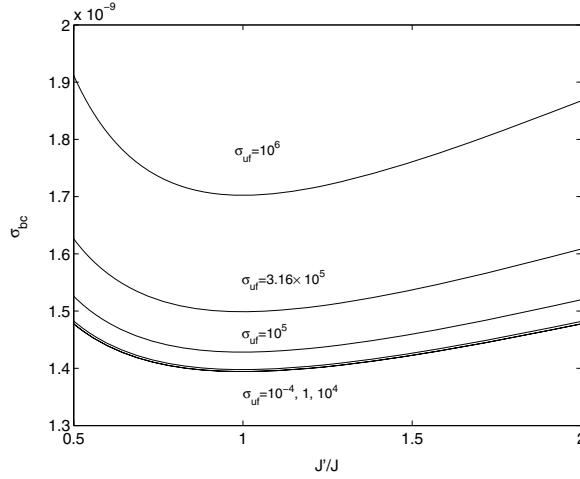


FIG. 5.1. Curves of  $\sigma_{bc}$  with respect to  $f$ , when  $\sigma_{uf}$  increases from  $10^{-4}$  to  $10^6$ . Curves are almost the same when  $\sigma_{uf} \leq 10^4$ . Parameters are chosen as in (5.1).

and  $\sigma_{bc}$  can be found by substituting (4.42) and (4.43) back into (4.33). We state this in a theorem.

**THEOREM 4.5.** *When  $J'$  is used in the design of the estimator (Kalman filter) and the controller,  $\sigma_{bc}$  obtained for the linear system (4.23) is approximated by*

$$(4.44) \quad \sigma_{bc}/\sigma_m \approx \frac{(K_z + r_b)K_b}{J'} - (1-f)\frac{K_b^2}{K_z} + \frac{K_b^2\sigma_{uf}\sigma_m^{-1}}{2\sqrt{T}K_z}(f^2 - 1).$$

The large  $\lambda$  and  $J$  assumptions are assumed to hold and so is  $\lambda(K_z + \lambda J') \gg \frac{\sqrt{T}}{J'}$ . Entries  $K_z$  and  $K_b$  are defined as before.

The inequality  $\lambda(K_z + \lambda J') \gg \frac{\sqrt{T}}{J'}$  is an extension of the large  $\lambda$  and  $J$  assumptions. According to the large  $\lambda$  assumption,  $\lambda^2$  is far larger than  $\sqrt{\frac{\sigma_{bf}}{\sigma_m} \frac{1}{2J'}}$ , and  $\frac{\sqrt{T}}{J'}$ , which is equal to  $\sqrt{\frac{\sigma_{bf}}{\sigma_m} + \frac{r_b^2\sigma_{uf}}{J'^2\sigma_m}}$ , is in the order of  $\sqrt{\frac{\sigma_{bf}}{\sigma_m}}$  if  $\sigma_{uf}$  is not too large. The condition about  $\sigma_{uf}$  is very reasonable if one considers the large  $J$  assumption.

**5. Effect of the input noise.** Equipped with Theorems 4.5 and 3.2, we can now compare  $\sigma_{bo}$  and  $\sigma_{bc}$ . When the compensative field is not applied with infinite accuracy, additional uncertainty is brought into the system. It influences and finally damages the measurement.

An example is useful. Figure 5.1 illustrates different  $\sigma_{bc}$  curves with respect to  $f$  for different  $\sigma_{uf}$ . Numbers are chosen as

$$(5.1) \quad \sigma_m = \sigma_{bf} = 10^{-4}, \quad J = 10^{10}, \quad \lambda = 0.5, \quad r_b = 10^3.$$

Figure 5.2 shows the three-dimensional mesh of  $\sigma_{bc}$  with respect to  $f$  and  $\sigma_{uf}$ . Additional input noise raises the estimation error, but its effect is not obvious when  $\sigma_{uf} \leq 10^5$ . For most  $\sigma_{uf}$ ,  $\sigma_{bc}$  obtains its minimum near  $f = 1$ .

A comparison between the open-loop and closed-loop results is shown in Figure 5.3 by drawing the ratio  $\sigma_{bc}/\sigma_{bo}$ . When  $\sigma_{uf}$  is small,  $\sigma_{bc}/\sigma_{bo} \leq 1$  for most  $J'$ . When  $\sigma_{uf}$  increases,  $\sigma_{bc}/\sigma_{bo}$  first becomes larger than 1 near  $f = 1$  and then the range of  $J'$ ,

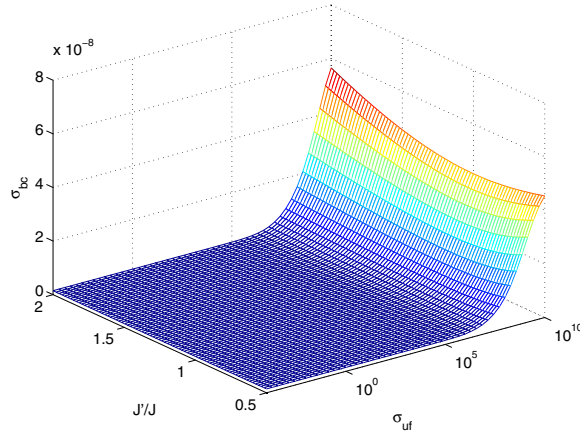


FIG. 5.2. Three-dimensional mesh of  $\sigma_{bc}$  with respect to  $f$  and  $\sigma_{uf}$ . Parameters are chosen as in (5.1).

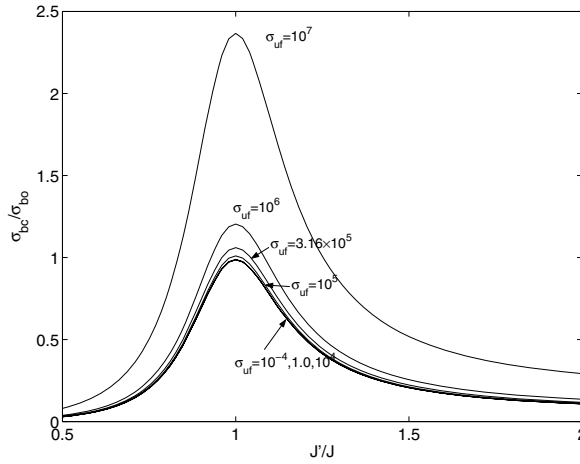


FIG. 5.3. Curves of  $\sigma_{bc}/\sigma_{bo}$  with respect to  $f$ , when  $\sigma_{uf}$  increases from  $10^{-4}$  to  $10^6$ . Curves are almost the same when  $\sigma_{uf} < 10^6$ . Parameters are chosen as in (5.1).

where  $\sigma_{bc}/\sigma_{bo} > 1$  expands towards both sides. Note that the figures are not accurate near  $f = 1$  because of simplifications in the computation and round-off errors.

Theorem 4.5 can be used to find a threshold of  $\sigma_{uf}$ . If we can assume that  $\frac{\sigma_{uf}}{\sigma_m} \ll \sqrt{T}$ ,  $K_z$  and  $K_b$  have a simpler form:

$$(5.2) \quad K_z \approx \sqrt{2\sqrt{T}}, \quad K_b \approx \frac{1}{\mathbb{J}'} \sqrt{T}.$$

Substituting (5.2) back into (4.44), we find an approximation of  $\sigma_{bc}$ :

$$(5.3) \quad \begin{aligned} \frac{\sigma_{bc}}{\sigma_m} &\approx \frac{K_z K_b}{\mathbb{J}'} + \frac{K_b^2}{K_z} (f - 1) \\ &\approx \left( \frac{\sigma_{bf}}{\sigma_m} + \frac{r_b^2 \sigma_{uf}}{\mathbb{J}'^2 \sigma_m} \right)^{3/4} \frac{\sqrt{2}}{2} \left( \frac{1}{\sqrt{\mathbb{J}'}} + \frac{\sqrt{\mathbb{J}'}}{\mathbb{J}} \right). \end{aligned}$$

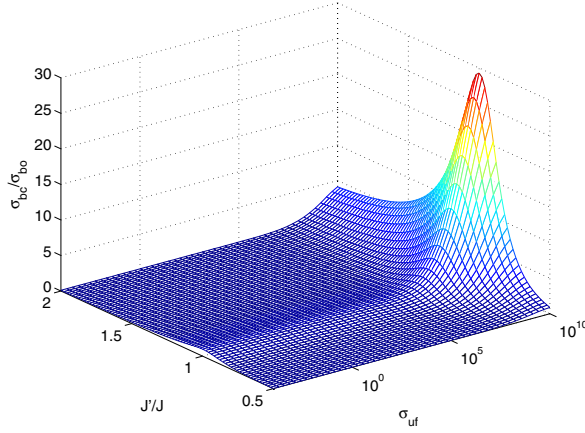


FIG. 5.4. Three-dimensional mesh of  $\sigma_{bc}/\sigma_{bo}$  with respect to  $f$  and  $\sigma_{uf}$ . Parameters are chosen as in (5.1).

The last term in (4.44) is discarded because  $\frac{\sigma_{uf}}{\sigma_m} \ll \sqrt{T}$ .

When

$$\frac{r_b^2 \sigma_{uf}}{\mathbb{J}'^2 \sigma_m} \ll \frac{\sigma_{bf}}{\sigma_m} \text{ or } \sigma_{uf} \ll \frac{\mathbb{J}'^2 \sigma_{bf}}{r_b^2},$$

the  $\sigma_{uf}$  term can be further discarded and

$$(5.4) \quad \frac{\sigma_{bc}}{\sigma_m} \approx \left( \frac{\sigma_{bf}}{\sigma_m} \right)^{3/4} \frac{\sqrt{2}}{2} \left( \frac{1}{\sqrt{\mathbb{J}'}} + \frac{\sqrt{\mathbb{J}'}}{\mathbb{J}} \right),$$

which is the same as the noise-free case. Thus  $\sigma_{bc}$  obtains its minimum near  $\mathbb{J} = \mathbb{J}'$ . In our example,

$$\frac{\mathbb{J}'^2 \sigma_{bf}}{r_b^2} = 10^{10} \text{ and } \sigma_m \sqrt{\mathbb{J}'^2 \sigma_{bf} \sigma_m^{-1}} = 10^6,$$

which is a good approximation of  $\sigma_m \sqrt{T}$ . Thus when  $\sigma_{uf} \ll 10^6$ , the input noise can be ignored. This result coincides with the simulation in Figures 5.2 and 5.4, where  $10^5$  is a point below which  $\sigma_{uf}$  has no obvious effect.

**6. Robust and optimal choices of  $J'$  when  $\sigma_{uf}$  is negligible.** In this section, besides giving an explanation of the robust choice of  $J'$  by the Caltech group, we will further discuss its optimal choice when  $J$  is a random number. We assume that  $\sigma_{uf}$  is not too large so that we can use (5.4).

When  $\sigma_{uf}$  is small enough so that it can be neglected, let  $\mathbb{J} = a\mathbb{J}_0$  and  $\mathbb{J}' = b\mathbb{J}_0$ , where  $0.5 \leq a \leq 2.0$  and  $\mathbb{J}_0$  is some constant. We can rewrite (5.4) as

$$(6.1) \quad \sigma_{bc} \sigma_m^{-1} = c \left( \frac{1}{\sqrt{b}} + \frac{\sqrt{b}}{a} \right),$$

where

$$c = \left( \frac{\sigma_{bf}}{\sigma_m} \right)^{3/4} \cdot \frac{\sqrt{2}}{2} \cdot \frac{1}{\sqrt{\mathbb{J}_0}}.$$



Denote the minimum of  $a$  as  $\min(a)$ ; the worst result happens for a fixed  $b$  when  $a = \min(a)$ :

$$(6.2) \quad \sigma_b \sigma_m^{-1} \leq c \left( \frac{1}{\sqrt{b}} + \frac{\sqrt{b}}{\min(a)} \right).$$

The minimum of the right-hand side is obtained when  $b = \min(a)$  or  $J' = \min(a) \cdot J_0$ , corresponding to the result by the Caltech group [16]. Thus the Caltech design is robust in the maximum-minimum meaning.

When  $J$  is fixed, simulation in the previous section already shows that the optimal choice of  $J'$  to minimize  $\sigma_{bc}$  is still  $J$ . Thus the best  $J'$  is still the real  $J$ . If  $J$  is random, an optimal design can be found by choosing the optimal objective function as

$$\mathcal{J} = E[\sigma_{bc}],$$

where the expectation is taken with respect to  $J$ . From (6.1), we have

$$(6.3) \quad E[\sigma_{bc}] = \sigma_m c \left( \frac{1}{\sqrt{b}} + \sqrt{b} E \left[ \frac{1}{a} \right] \right) \geq 2c\sigma_m \sqrt{E \left[ \frac{1}{a} \right]},$$

and the identity is obtained only when  $J' = \frac{1}{E[1/J]}$ .

**7. Conclusion.** Extending the model proposed by the Caltech group by an input noise term, we have evaluated its effects on the measurement of a magnetic field using feedback. Large  $J$  and  $\lambda$  assumptions are repeatedly used in simplifying the results, especially the large  $J$  assumption. This implies that the number of atoms involved is relatively large. In our simulation, it is about  $10^6$ .

Both the closed-loop case and open-loop case are studied, but we have more interest in the closed-loop case and use the open-loop one for comparison. The input noise can be ignored if

$$\sigma_{uf} \ll \min \left( \sigma_m \sqrt{T}, \frac{J'^2 \sigma_{bc}}{r_b^2} \right).$$

When  $\sigma_{uf}$  is large,  $\sigma_{bc}$  is not always less than  $\sigma_{bo}$ , as it is when  $\sigma_{uf}$  is small enough. The region where  $\sigma_{bc} < \sigma_{bo}$  shrinks when  $\sigma_{uf}$  gets large. Although the computation is too tedious for us to give every detail, we have given all the main results. The quantitative results are helpful for explaining data collected in the experiment [16].

The best choice of  $J'$  is the actual  $J$  if we know  $J$  exactly. When we do not know  $J$  exactly or it is a random number, the optimal  $J'$  satisfies  $J' = 1/E[1/J]$  instead of  $J' = E[J]$ .

**Acknowledgments.** The author would like to thank Dr. Goong Chen for support and advice, and to acknowledge useful discussions with Dr. Zijian Diao.

#### REFERENCES

- [1] A. J. BERGLUND AND H. MABUCHI, *Feedback controller design for tracking a single fluorescent molecule*, Applied Physics B: Lasers and Optics, 78 (2004), pp. 653–659.
- [2] A. C. DOHERTY AND K. JACOBS, *Feedback control of quantum systems using continuous state estimation*, Phys. Rev. A, 60 (1999), pp. 2700–2711.

- [3] J. M. GEREMIA, J. K. STOCKTON, AND H. MABUCHI, *Real-time quantum feedback control of atomic spin-squeezing*, Science, 304 (2004), pp. 270–273.
- [4] G. C. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction and Control*, Prentice–Hall, Englewood, NJ, 1984.
- [5] M. GRIMBLE AND M. JOHNSON, *Optimal Control and Stochastic Estimation: Theory and Applications*, Vol. 1, John Wiley & Sons, Chichester, 1988.
- [6] S. A. MAAS, *Noise in Linear and Nonlinear Circuits*, Artech House, Norwood, MA, 2005.
- [7] H. MABUCHI AND N. KHANEJA, *Principles and applications of control in quantum systems*, Internat. J. Robust Nonlinear Control, 15 (2005), pp. 647–667.
- [8] K. MCHALE, A. J. BERGLUND, AND H. MABUCHI, *Bayesian estimation for species identification in single-molecule fluorescence microscopy*, Biophys. J., 86 (2004), pp. 3409–3422.
- [9] K. MOLMER AND L. B. MADSEN, *Estimation of a classical parameter with Gaussian probes: Magnetometry with collective atomic spins*, Phys. Rev. A, 70 (2004), 052102.
- [10] V. PETERSEN, L. B. MADSEN, AND K. MOLMER, *Magnetometry with entangled atomic samples*, Phys. Rev. A, 71 (2005), 012312.
- [11] D. T. POPE, H. M. WISEMAN, AND N. K. LANGFORD, *Adaptive phase estimation is more accurate than nonadaptive phase estimation for continuous beams of light*, Phys. Rev. A, 70 (2004), 043812.
- [12] J. E. REINER, H. M. WISEMAN, AND H. MABUCHI, *Quantum jumps between dressed states: A proposed cavity-qed test using feedback*, Phys. Rev. A, 67 (2003), 042106.
- [13] M. SCULLY AND M. S. ZUBAIRY, *Quantum Optics*, Cambridge University Press, Cambridge, UK, 1997.
- [14] A. SILBERFARB AND I. H. DEUTSCH, *Continuous measurement with traveling-wave probes*, Phys. Rev. A, 68 (2003), 013817.
- [15] D. A. STECK, K. JACOBS, H. MABUCHI, T. BHATTACHARYA, AND S. HABIB, *Quantum feedback control of atomic motion in an optical cavity*, Phys. Rev. Lett., 92 (2004), 223004.
- [16] J. K. STOCKTON, J. M. GEREMIA, A. C. DOHERTY, AND H. MABUCHI, *Robust quantum parameter estimation: Coherent magnetometry with feedback*, Phys. Rev. A, 69 (2004), 032109.
- [17] J. K. STOCKTON, R. VAN HANDEL, AND H. MABUCHI, *Deterministic Dicke-state preparation with continuous measurement and control*, Phys. Rev. A, 70 (2004), 022106.
- [18] L. K. THOMSEN, S. MANCINI, AND H. M. WISEMAN, *Spin squeezing via quantum feedback*, Phys. Rev. A, 65 (2002), 061801.
- [19] R. VAN HANDEL, J. K. STOCKTON, AND H. MABUCHI, *Feedback control of quantum state reduction*, IEEE Trans. Automat. Control, 50 (2005), pp. 768–780.
- [20] H. M. WISEMAN, *Quantum theory of continuous feedback*, Phys. Rev. A, 49 (1994), pp. 2133–2150.
- [21] H. M. WISEMAN AND G. J. MILBURN, *Squeezing via feedback*, Phys. Rev. A, 49 (1994), pp. 1350–1366.

## STABLE SYNCHRONIZATION OF MECHANICAL SYSTEM NETWORKS\*

SUJIT NAIR<sup>†</sup> AND NAOMI EHRICH LEONARD<sup>†</sup>

**Abstract.** In this paper we address stabilization of a network of underactuated mechanical systems with unstable dynamics. The coordinating control law stabilizes the unstable dynamics with a term derived from the method of controlled Lagrangians and synchronizes the dynamics across the network with potential shaping designed to couple the mechanical systems. The coupled system is Lagrangian with symmetry, and energy methods are used to prove stability and coordinated behavior. Two cases of asymptotic stabilization are discussed; one yields convergence to synchronized motion staying on a constant momentum surface, and the other yields convergence to a relative equilibrium. We illustrate the results in the case of synchronization of  $n$  carts, each balancing an inverted pendulum.

**Key words.** coordinated control, mechanical systems, networks, synchronization, stabilization, energy shaping

**AMS subject classifications.** 70Q05, 70H33, 93D15

**DOI.** 10.1137/050646639

**1. Introduction.** Coordinated motion and cooperative control have become important topics of late because of growing interest in the possibility of faster data processing and more efficient decision-making by a network of autonomous systems. For example, mobile sensor networks are expected to provide better data about a distributed environment if the sensors can be made to cooperate towards optimal coverage and efficient coordination.

Much of the recent work explores coordination and cooperative control with very simple dynamical systems, e.g., single or double integrator models (see, e.g., [10, 17, 18]) or nonholonomic models (see, e.g., [4]). For example, in some of these and related works, stabilization of coordinated group dynamics is studied in the case of limited, time-varying communication topologies. These authors deliberately choose to focus on the coordination issues independently of issues in the stabilization of individual dynamics.

However, for networks of autonomous systems such as unmanned helicopters or underwater vehicles, stability of individual dynamics can be important and challenging, and it may not always be possible (or desirable) to decouple the stabilization problem of individual dynamics from the coordination problem. In [22] the authors consider stability of a group with dynamics that satisfy a leader-to-formation stability (LFS) condition based on input-to-state stability [20]. Examples include linear dynamical systems and kinematic nonholonomic robots; in the latter case feedback linearization is used for stabilization. Using the LFS property, the authors are able to quantify how leader inputs and disturbances affect group stability. In [6], an extension to the previous work of [5] on unmanned aerial vehicle motion planning is presented

---

\*Received by the editors December 2, 2005; accepted for publication (in revised form) October 31, 2007; published electronically February 8, 2008. Research on this paper was partially supported by the Office of Naval Research under grants N00014-02-1-0826 and N00014-04-1-0534. A preliminary version of some parts of this paper appeared in [16].

<http://www.siam.org/journals/sicon/47-2/64663.html>

<sup>†</sup>Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544 (nair@alumni.princeton.edu, naomi@princeton.edu).

for identical multiple-vehicle stabilization and coordination. The single-vehicle motion planning is based on the interconnection of a finite number of suitably defined motion primitives. The problem is set in such a way that multiple-vehicle motion coordination primitives are obtained from the single-vehicle primitives. The technique is applied to motion planning for a group of small model helicopters.

Networks of rigid bodies are addressed in [8]. Reduction theory is applied in the case that control inputs depend only on relative configuration (relative orientation or position). The reduction results are used to study coordinated behavior of satellite and underwater vehicle network dynamics. Stability of a network of rotating rigid satellites and a network of coordinated underwater vehicles is proved in [14, 15].

In this paper, we investigate the problem of coordination of a network of underactuated mechanical systems with unstable dynamics. As a first step we make use of the method of controlled Lagrangians to stabilize the unstable dynamics of each mechanical system. The method of controlled Lagrangians and the equivalent interconnection and damping assignment passivity-based control (IDA-PBC) method use energy shaping for stabilization of underactuated mechanical systems (see [1, 19] and references therein). The method of controlled Lagrangians provides a control law for underactuated mechanical systems such that the closed-loop dynamics derive from a Lagrangian. The approach is to choose the control law to shape the controlled kinetic and potential energy for stability.

The class of underactuated mechanical systems we consider in this paper satisfies the *simplified matching conditions* (SMC) defined in [2, 1]. This class includes the planar or spherical inverted pendulum on a (controlled) cart. The goal of the development in this paper is to stabilize unstable dynamics for each individual mechanical system in the network and stably synchronize the actuated configuration variables across the network. For example, for a network of pendulum/cart systems, the problem is to stabilize each pendulum in the upright position while synchronizing the motion of the carts.

For stabilization of individual unstable dynamics we use the approach in [1]. To simultaneously synchronize the dynamics across the network, we show that potentials that couple the individual systems can be prescribed so that the complete coupled system still satisfies the SMC. Accordingly, we can choose potentials, find a Lagrangian for the coupled system, and prove Lyapunov stability of the stabilized and synchronized network. Since the controlled Lagrangian has a symmetry, we use Routh reduction and Routh criteria to prove stability.

We then design additional dissipative control terms and prove asymptotic stability. We show, on the one hand, how to apply a dissipative control term that yields convergence to synchronization staying on a constant momentum surface. In the pendulum/cart system example, this corresponds to a synchronized motion of the carts such that all the carts move together with a common velocity that is the sum of a constant plus an oscillation. Likewise, the pendula synchronize and oscillate at the same frequency as the carts. The oscillation frequency for the carts and pendula is determined by the control parameters. On the other hand, we show how to apply a dissipative control term that yields convergence to a relative equilibrium. In the example, this corresponds to steady, synchronized motion of  $n$  carts, each balancing its inverted pendulum.

In this paper we consider a homogeneous group of mechanical systems, i.e., no leaders, and a fixed, bidirectional, connected communication topology. Possibilities for extension include integration of the results with prior works cited above that

address time-varying and directed communication topologies and/or the presence of leaders in the group.

The organization of the paper is as follows. In section 2, we define notation and the different kinds of stabilization studied. In section 3, we give a brief background on the class of underactuated mechanical systems that satisfy the SMC defined in [2, 1]. We discuss how unstable dynamics are stabilized with feedback control that preserves Lagrangian structure. In section 4, we study a network of  $n$  systems, each of which satisfies the SMC. We choose coupling potentials in section 5, and we prove stability and coordination of the network. Asymptotic stabilization is investigated in sections 6 and 7. We illustrate the theory with the example of  $n$  planar, inverted pendulum/cart systems in section 8. In section 9 we conclude with a few remarks.

**2. Definitions.** In [1] the method of controlled Lagrangians is used to derive a control law that asymptotically stabilizes a class of underactuated mechanical systems with otherwise unstable dynamics. This class of systems satisfies a set of “simplified matching conditions” and we denote such systems as *SMC systems*. SMC systems lack gyroscopic forces; the planar inverted pendulum on a cart and the spherical inverted pendulum on a 2D cart are two such systems.

Consider an underactuated mechanical system with an  $(m + r)$ -dimensional configuration space. Let  $x^\alpha$  denote the coordinates for the unactuated directions with index  $\alpha$  going from 1 to  $m$ .  $\theta^a$  denotes the coordinates for the actuated directions with index  $a$  going from 1 to  $r$ . In the case of a network of  $n$  mechanical systems, each with the same  $(m + r)$ -dimensional configuration space,  $x_i^\alpha$  and  $\theta_i^a$  are the corresponding coordinates for the  $i$ th mechanical system,  $i = 1, \dots, n$ . Beginning in section 5, we will assume that the configuration space for the actuated variables for each individual system is  $\mathbb{R}^r$ . Note that we only require the configuration space for the individual mechanical systems to be the same and do not require that each system be identical, e.g., the individual systems can have different mass and inertia values. We will need to make the assumption of individual systems being identical only in section 6.

The goal of coordination is to synchronize the actuated variables  $\theta_i^a$  with the variables  $\theta_j^a$  for all  $i, j = 1, \dots, n$ . We define stable synchronization of these variables as stabilization of  $\theta_i^a - \theta_j^a = 0$  for all  $i \neq j$ .

We define the following stability notions for the mechanical system network.

**DEFINITION 2.1 (SSRE).** *A relative equilibrium of the mechanical system network dynamics is a stable synchronized relative equilibrium (SSRE) if it is defined by  $\theta_i^a - \theta_j^a = 0$  for all  $i \neq j$ ,  $x_i^\alpha = 0$  for all  $i$ , and if it is Lyapunov stable. This implies that the unactuated dynamics are stable and the actuated dynamics are stably synchronized.*

**DEFINITION 2.2 (ASSRE).** *A relative equilibrium of the mechanical system network dynamics is an asymptotically stable synchronized relative equilibrium (ASSRE) if it is SSRE and asymptotically stable.*

**DEFINITION 2.3 (ASSM).** *An asymptotically stable solution of the mechanical system network dynamics is an asymptotically stable synchronized motion (ASSM) if it is defined by  $x_i^\alpha - x_j^\alpha = 0$  and  $\theta_i^a - \theta_j^a = 0$  for all  $i \neq j$  and the dynamics of the network evolve on a constant momentum surface.*

We note that an ASSRE is a special case of an ASSM. In the example of the network of pendulum/cart systems, the relative equilibrium of interest corresponds to the carts moving together at the same constant speed with each pendulum at rest in the upright position. In section 8 we asymptotically stabilize this synchronized relative equilibrium as well as a family of synchronized motions that exhibit a synchronized steady motion plus an oscillation of the carts and pendula.

**3. SMC.** Let the Lagrangian for an individual mechanical system be given by

$$L(x^\alpha, \theta^a, \dot{x}^\beta, \dot{\theta}^b) = \frac{1}{2}g_{\alpha\beta}\dot{x}^\alpha\dot{x}^\beta + g_{\alpha a}\dot{x}^\alpha\dot{\theta}^a + \frac{1}{2}g_{ab}\dot{\theta}^a\dot{\theta}^b - V(x^\alpha, \theta^a),$$

where summation over indices is implied,  $g$  is the kinetic energy metric, and  $V$  is the potential energy. It is assumed that the actuated directions are symmetry directions for the kinetic energy; that is, we assume  $g_{\alpha\beta}$ ,  $g_{\alpha a}$ ,  $g_{ab}$  are all independent of  $\theta^a$ . The equations of motion for the mechanical system with control inputs  $u_a$  are given by

$$\begin{aligned}\mathcal{E}_{x^\alpha}(L) &= 0, \\ \mathcal{E}_{\theta^a}(L) &= u_a,\end{aligned}$$

where  $\mathcal{E}_q(L)$  denotes the Euler–Lagrange expression corresponding to a Lagrangian  $L$  and generalized coordinates  $q$ , i.e.,

$$(3.1) \quad \mathcal{E}_q(L) = \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q}.$$

For such a system, following [1], the SMC are

- $g_{ab} = \text{constant}$ ;
- $\frac{\partial g_{\alpha a}}{\partial x^\beta} = \frac{\partial g_{\beta a}}{\partial x^\alpha}$ ;
- $\frac{\partial^2 V}{\partial x^\alpha \partial \theta^a} g^{ad} g_{\beta d} = \frac{\partial^2 V}{\partial x^\beta \partial \theta^a} g^{ad} g_{\alpha d}$ .

Satisfaction of these SMC allows for a structured feedback shaping of kinetic and potential energy. In particular, a control law  $u_a = u_a^{\text{cons}}$  is given in [1] such that the closed-loop system is a Lagrangian system. The controlled Lagrangian  $L_c$ , parametrized by constant parameters  $\kappa$  and  $\rho$  and by a potential term  $V_\epsilon$ , is given by

$$\begin{aligned}L_c(x^\alpha, \theta^a, \dot{x}^\beta, \dot{\theta}^b) &= \frac{1}{2} \left( g_{\alpha\beta} + \rho(\kappa + 1) \left( \kappa + \frac{\rho - 1}{\rho} \right) g_{\alpha a} g^{ab} g_{b\beta} \right) \dot{x}^\alpha \dot{x}^\beta + \rho(\kappa + 1) g_{\alpha a} \dot{x}^\alpha \dot{\theta}^a \\ &\quad + \frac{1}{2} \rho g_{ab} \dot{\theta}^a \dot{\theta}^b - V(x^\alpha, \theta^b) - V_\epsilon(x^\alpha, \theta^b),\end{aligned}$$

where  $V_\epsilon$  must satisfy

$$(3.2) \quad - \left( \frac{\partial V}{\partial \theta^a} + \frac{\partial V_\epsilon}{\partial \theta^a} \right) \left( \kappa + \frac{\rho - 1}{\rho} \right) g^{ad} g_{\alpha d} + \frac{\partial V_\epsilon}{\partial x^\alpha} = 0.$$

The results in [1] further give conditions on  $\rho$ ,  $\kappa$ , and  $V_\epsilon$  that ensure stability of the equilibrium in the full state space. Without loss of generality, we assume that the equilibrium of interest is the origin. We further assume that it is a *maximum* of the original potential energy  $V$  (the case when the origin is a minimum can be handled similarly). The inverted pendulum systems fall into this category. In this case,  $\kappa > 0$  and  $\rho < 0$  and the potential  $V_\epsilon$  can be chosen such that the energy function  $E_c$  for the controlled Lagrangian has a maximum at the origin of the full state space. Asymptotic stability is obtained by adding a dissipative term  $u_a^{\text{diss}}$  to the control law, i.e.,

$$u_a = u_a^{\text{cons}} + \frac{1}{\rho} u_a^{\text{diss}},$$

which drives the controlled system to the maximum value of the energy  $E_c$ .

In [1], it is also shown how to select new, useful coordinates  $(x^\alpha, y^a, \dot{x}^\alpha, \dot{y}^a)$ . In particular, for any SMC system, there exists a function  $h^a(x^\alpha)$  defined on an open subset of the configuration space of the unactuated variables such that

$$\frac{\partial h^a}{\partial x^\alpha} = \left( \kappa + \frac{\rho - 1}{\rho} \right) g^{\alpha c} g_{\alpha c}, \quad h^a(0) = 0.$$

The new coordinates are defined as

$$(x^\alpha, y^a) = (x^\alpha, \theta^a + h^a(x^\alpha)).$$

Note that if the origin is an equilibrium in the original coordinates, it is also an equilibrium in the new coordinates. In these coordinates, the closed-loop Lagrangian takes the form

$$(3.3) \quad L_c = \frac{1}{2} \left( g_{\alpha\beta} - \left( \kappa + \frac{\rho - 1}{\rho} \right) g_{\alpha a} g^{ab} g_{b\beta} \right) \dot{x}^\alpha \dot{x}^\beta + g_{\alpha a} \dot{x}^\alpha \dot{y}^a + \frac{1}{2} \rho g_{ab} \dot{y}^a \dot{y}^b \\ - V(x^\alpha, y^a - h^a(x^\alpha)) - V_\epsilon(y^a)$$

$$(3.4) \quad = \frac{1}{2} \tilde{g}_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta + \tilde{g}_{\alpha a} \dot{x}^\alpha \dot{y}^a + \frac{1}{2} \tilde{g}_{ab} \dot{y}^a \dot{y}^b - V(x^\alpha, y^a - h^a(x^\alpha)) - V_\epsilon(y^a),$$

where

$$(3.5) \quad \begin{aligned} \tilde{g}_{\alpha\beta} &= \left( g_{\alpha\beta} - \left( \kappa + \frac{\rho - 1}{\rho} \right) g_{\alpha a} g^{ab} g_{b\beta} \right), \\ \tilde{g}_{\alpha a} &= g_{\alpha a}, \\ \tilde{g}_{ab} &= \rho g_{ab}. \end{aligned}$$

Further, after adding dissipation  $u_a^{\text{diss}}$ , the Euler–Lagrange equations in the new coordinates become

$$\begin{aligned} \mathcal{E}_{x^\alpha}(L_c) &= 0, \\ \mathcal{E}_{y^a}(L_c) &= u_a^{\text{diss}}. \end{aligned}$$

**4. Matching for a network of SMC systems.** In this section we examine a network of  $n$  systems, each of which satisfies the SMC. We determine what control design freedom remains under the constraint that the complete network dynamics are Lagrangian and satisfy the simplified matching conditions.

Consider  $n$  SMC systems and let the  $i$ th system have dynamics described by Lagrangian  $L_i$ , where

$$(4.1) \quad L_i(x_i^\alpha, \theta_i^a, \dot{x}_i^\beta, \dot{\theta}_i^b) = \frac{1}{2} g_{\alpha\beta}^i \dot{x}_i^\alpha \dot{x}_i^\beta + g_{\alpha a}^i \dot{x}_i^\alpha \dot{\theta}_i^a + \frac{1}{2} g_{ab}^i \dot{\theta}_i^a \dot{\theta}_i^b - V_i(x_i^\alpha, \theta_i^a),$$

and the index  $i$  on every variable refers to the  $i$ th system.

The Lagrangian for the total (uncontrolled, uncoupled) system is  $L = \sum_{i=1}^n L_i = \frac{1}{2} \dot{\mathbf{x}}^T M \dot{\mathbf{x}} - \sum_{i=1}^n V_i(x_i^\alpha, \theta_i^a)$ , where  $\mathbf{x} = (x_1^\alpha, \dots, x_n^\beta, \theta_1^a, \dots, \theta_n^b)^T$ , and

$$M = \left( \begin{array}{cc|cc} g_{\alpha\beta}^1 & 0 & g_{\alpha a}^1 & 0 \\ & \ddots & & \ddots \\ 0 & g_{\alpha\beta}^n & 0 & g_{\alpha a}^n \\ \hline g_{a\alpha}^1 & 0 & g_{ab}^1 & 0 \\ & \ddots & & \ddots \\ 0 & g_{a\alpha}^n & 0 & g_{ab}^n \end{array} \right).$$

Since each system satisfies the SMC,  $g_{ab}^i = \text{constant}$  for each  $i = 1, \dots, n$ . It can be easily verified that the SMC are satisfied for the total system  $L$ , since they are satisfied for each individual system.

For the total system, the symmetry coordinates are  $(\theta_1^a, \dots, \theta_n^b)$ . As in [1], we can find a control law and a change of coordinates  $\mathbf{x} = (x_1^\alpha, \dots, x_n^\beta, \theta_1^a, \dots, \theta_n^b) \mapsto \mathbf{x}' = (x_1^\alpha, \dots, x_n^\beta, y_1^a, \dots, y_n^b)$  such that the closed-loop system is equivalent to another Lagrangian system with

$$(4.2) \quad L'_c = \frac{1}{2}(\dot{\mathbf{x}}')^T M_c \dot{\mathbf{x}}' - V'_\epsilon(\mathbf{x}')$$

and

$$(4.3) \quad M_c = \left( \begin{array}{cc|cc} \tilde{g}_{\alpha\beta}^1 & 0 & \tilde{g}_{\alpha a}^1 & 0 \\ & \ddots & & \ddots \\ 0 & \tilde{g}_{\alpha\beta}^n & 0 & \tilde{g}_{\alpha a}^n \\ \hline \tilde{g}_{a\alpha}^1 & 0 & \tilde{g}_{ab}^1 & 0 \\ & \ddots & & \ddots \\ 0 & \tilde{g}_{a\alpha}^n & 0 & \tilde{g}_{ab}^n \end{array} \right) := \left( \begin{array}{c|c} M_{11} & M_{12} \\ \hline M_{12}^T & M_{22} \end{array} \right),$$

$$V'_\epsilon = \sum_{i=1}^n \left( V_i(x_i^\alpha, y_i^a - h_i^a(x_i^\alpha)) + V_{\epsilon i}(x_i^\alpha, y_i^a) \right).$$

Here,  $\tilde{g}_{\alpha\beta}^i$ ,  $\tilde{g}_{\alpha a}^i$ , and  $\tilde{g}_{ab}^i$  are defined as in (3.5) with all variables replaced with those corresponding to the  $i$ th system, e.g.,  $\tilde{g}_{ab}^i = \rho_i g_{ab}^i$ , etc.

The control gains  $\kappa_i$  and  $\rho_i$  and control potentials  $V_{\epsilon i}$  can be chosen such that the mass matrix  $M_c$  is negative definite and the potential  $V'_\epsilon$  has a maximum when the configuration of each system, i.e.,  $(x_i^\alpha, \theta_i^a)$ , is at the origin. This means the control law brings each system independently to the origin without coordination.

To determine what additional freedom exists in the choice of the control, notably in the choice of control potentials  $V_{\epsilon i}$ , such that the network system satisfies the SMC, we specialize to a network of SMC systems which each satisfy the following condition.

**AS1.** *The potential energy for each system in the original coordinates satisfies  $V_i(x_i^\alpha, \theta_i^a) = V_{1i}(x_i^\alpha) + V_{2i}(\theta_i^a)$ .*

The inverted pendulum examples satisfy this assumption in the general case that the cart moves on an inclined plane. In the case that the cart moves in the horizontal plane,  $V_2 = 0$ .

As shown in [1], given the assumption **AS1**,  $V_{\epsilon i}$  in the new coordinates for  $i = 1, \dots, n$  can be chosen to take the form

$$V_{\epsilon i}(x_i^\alpha, y_i^a) = -V_{2i}(y_i^a - h_i^a(x_i^\alpha)) + \bar{V}_{\epsilon i}(y_i^a),$$

where  $\bar{V}_{\epsilon i}$  is an arbitrary function and  $h_i^a(x_i^\alpha)$  satisfies

$$(4.4) \quad \frac{\partial h_i^a}{\partial x_i^\alpha} = \left( \kappa_i + \frac{\rho_i - 1}{\rho_i} \right) g_i^{ac} g_{\alpha c}^i, \quad h_i^a(0) = 0.$$

We show next that a more general potential  $V_\epsilon$  can be used in  $V'_\epsilon$  in place of the sum of potentials  $V_{\epsilon i}(x_i^\alpha, y_i^a)$ .



PROPOSITION 4.1. *Under assumption **AS1**, the potential  $V'_\epsilon = V + V_\epsilon$  satisfies the SMC with*

$$(4.5) \quad \begin{aligned} V &= \sum_{i=1}^n (V_{1i}(x_i^\alpha)) + V_{2i}(y_i^a - h_i^a(x_i^\alpha)), \\ V_\epsilon &= - \left( \sum_{i=1}^n V_{2i}(y_i^a - h_i^a(x_i^\alpha)) \right) + \tilde{V}_\epsilon(y_1^a, \dots, y_n^a) \end{aligned}$$

and  $\tilde{V}_\epsilon$  an arbitrary function.

*Proof.* Recall that the potential  $V'_\epsilon = V + V_\epsilon$  given by (4.5) satisfies the SMC if (3.2) holds. Following [1], we can use the definition of  $h_i^a(x_i^\alpha)$  given by (4.4) to write the SMC (3.2) for the potential as

$$(4.6) \quad \frac{\partial V_\epsilon}{\partial x_i^\alpha} = \frac{\partial V}{\partial y_i^a} \frac{\partial h_i^a(x_i^\alpha)}{\partial x_i^\alpha}, \quad i = 1, \dots, n.$$

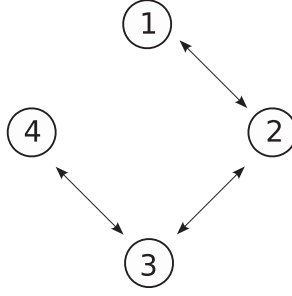
By a direct computation, one can check that each side of (4.6) is equal to  $\frac{\partial V_{2i}}{\partial v_i^a} \frac{\partial v_i^a}{\partial x_i^\alpha}$ , where  $v_i^a = y_i^a - h_i^a(x_i^\alpha)$ .  $\square$

Proposition 4.1 implies that *we can couple the  $n$  vehicles in the network using the freedom in our choice of  $\tilde{V}_\epsilon = \tilde{V}_\epsilon(y_1^a, \dots, y_n^a)$ , and the network dynamics will still satisfy the simplified matching conditions.* This result is completely independent of the degree of coupling; i.e., it extends from a network of uncoupled systems to a network of completely connected systems.

**5. Stable coordination of an SMC network.** In this section we make use of Proposition 4.1 to design coupling potentials  $\tilde{V}_\epsilon$  for stable coordination of the network of SMC systems. We prove that the relative equilibrium of interest is an SSRE. Recall from section 2 that to be an SSRE, a relative equilibrium should be defined by  $\theta_i^a - \theta_j^a = 0$  for all  $i \neq j$  and  $x_i^\alpha = 0$  for all  $i$  and should be Lyapunov stable. We note that this is equivalent to showing that  $y_i^a - y_j^a = 0$  for all  $i \neq j$  and  $x_i^\alpha = 0$  for all  $i$  is Lyapunov stable. In the remainder of the paper we assume that the configuration space for the actuated variables for each individual system is  $\mathbb{R}^r$ .

To synchronize the actuated variables we use the results of Proposition 4.1 and design coupling potentials for stabilization of  $y_i^a - y_j^a = 0$  for all  $i \neq j$ . Note that the condition  $y_i^a - y_j^a = 0$  for all  $i \neq j$  by itself is necessary but not sufficient for  $\theta_i^a - \theta_j^a = 0$  for all  $i \neq j$  and  $x_i^\alpha = 0$  for all  $i$ . We have  $y_i^a - y_j^a = 0$  for all  $i \neq j$  under more general conditions, e.g., if  $\theta_i^a - \theta_j^a = 0$  for all  $i \neq j$  and  $h_i(x_i^\alpha) = h_j(x_j^\alpha) \neq 0$ ,  $i \neq j$ . This more general case makes possible interesting synchronized dynamics, when we add dissipation for asymptotic stability, as will be discussed in section 6.

We choose  $\tilde{V}_\epsilon$  such that the closed-loop potential  $V'_\epsilon$ , defined in Proposition 4.1, has a maximum when  $x_i^\alpha = 0$  and  $y_i^a - y_j^a = 0$  for all  $i \neq j$ . This is possible since from (4.5), the closed-loop potential is  $V'_\epsilon = \sum_{i=1}^n (V_{1i}(x_i)) + \tilde{V}_\epsilon(y_1^a, \dots, y_n^a)$  and the  $V_{1i}$  are assumed to already be maximized at  $x_i^\alpha = 0$ . We choose in this paper  $\tilde{V}_\epsilon$  to be quadratic in  $(y_i^a - y_j^a)$  with a maximum at  $y_i^a - y_j^a = 0$  for all  $i \neq j$ . In this case, consider a graph with one node corresponding to each individual system in the network. There is an (undirected) edge between nodes  $k$  and  $l$  if the term  $(y_k^a - y_l^a)$  appears in the quadratic function  $\tilde{V}_\epsilon$ . Then,  $V'_\epsilon$  has a strict maximum when  $x_i^\alpha = 0$  and  $y_i^a - y_j^a = 0$  for all  $i \neq j$  if the (undirected) graph is connected. Figure 5.1 illustrates an example of a connected, undirected communication graph for four vehicles.

FIG. 5.1. *Connected, undirected communication graph for four vehicles.*

With coupling of the individual systems using terms that depend only on  $y_i^a - y_j^a$ , the network system has a translational symmetry. Specifically, the system dynamics are invariant under translation of the center of mass of the network. Consider a new set of coordinates given by

$$(5.1) \quad \mathbf{x}_c = (x_1^\alpha, \dots, x_n^\beta, z_1^a, \dots, z_n^b)^T,$$

where

$$\begin{aligned} z_i^a &= y_1^a - y_{i+1}^a, \quad i = 1, \dots, n-1, \\ z_n^b &= y_1^b + \dots + y_n^b. \end{aligned}$$

In this coordinate system, the controlled Lagrangian for the total system (with abuse of notation for  $V'_\epsilon$ ) is

$$(5.2) \quad \tilde{L}_c = \frac{1}{2} \dot{\mathbf{x}}_c^T \tilde{M}_c \dot{\mathbf{x}}_c - V'_\epsilon(\mathbf{x}_r),$$

where  $\mathbf{x}_r = (x_1^\alpha, \dots, x_n^\beta, z_1^a, \dots, z_{n-1}^b)^T$  and

$$(5.3) \quad \tilde{M}_c = \begin{pmatrix} \tilde{M}_{11} & \tilde{M}_{12} \\ \tilde{M}_{12}^T & \tilde{M}_{22} \end{pmatrix}.$$

The transformation which takes the coordinates  $\mathbf{x}_c$  to the coordinates  $\mathbf{x}' = (x_1^\alpha, \dots, x_n^\beta, y_1^c, \dots, y_n^d)$  is given by the matrix

$$(5.4) \quad B = \begin{bmatrix} I_{mn \times mn} & 0 \\ 0 & B_{22} \end{bmatrix},$$

where

$$(5.5) \quad B_{22} = \frac{1}{n} \begin{bmatrix} I_{r \times r} & I_{r \times r} & \dots & I_{r \times r} \\ (1-n)I_{r \times r} & I_{r \times r} & \dots & I_{r \times r} \\ \vdots & \vdots & \dots & \vdots \\ I_{r \times r} & \dots & (1-n)I_{r \times r} & I_{r \times r} \end{bmatrix}$$

and  $I_{l \times l}$  denotes an  $l \times l$  identity matrix and  $B_{22}$  is an  $rn \times rn$  matrix. The expression for  $\tilde{M}_c$  in terms of  $M_c$  from (4.3) is

$$(5.6) \quad \tilde{M}_c = B^T M_c B.$$

We can compute the block elements in  $\tilde{M}_c$  to be

$$(5.7) \quad \tilde{M}_{11} = M_{11},$$

$$(5.8) \quad \tilde{M}_{12} = \frac{1}{n} \begin{pmatrix} \tilde{g}_{\alpha a}^1 & \tilde{g}_{\alpha a}^1 & \cdots & \tilde{g}_{\alpha a}^1 & \tilde{g}_{\alpha a}^1 \\ (1-n)\tilde{g}_{\alpha a}^2 & \tilde{g}_{\alpha a}^2 & \cdots & \tilde{g}_{\alpha a}^2 & \tilde{g}_{\alpha a}^2 \\ & & \ddots & & \\ \tilde{g}_{\alpha a}^{n-1} & \tilde{g}_{\alpha a}^{n-1} & \cdots & \tilde{g}_{\alpha a}^{n-1} & \tilde{g}_{\alpha a}^{n-1} \\ \tilde{g}_{\alpha a}^n & \tilde{g}_{\alpha a}^n & \cdots & (1-n)\tilde{g}_{\alpha a}^n & \tilde{g}_{\alpha a}^n \end{pmatrix},$$

$$(5.9) \quad \tilde{M}_{22} = \frac{1}{n^2} B_{22}^T M_{22} B_{22},$$

where  $M_{11}$  and  $M_{22}$  are as defined in (4.3). From (5.5) and (4.3), we can calculate the lowermost diagonal  $r \times r$  block of  $\tilde{M}_{22}$  to be

$$(5.10) \quad \tilde{g}_{ab} = \frac{1}{n^2} \sum_{i=1}^n (\tilde{g}_{ab}^i).$$

Thus, we can define  $\bar{M}_{22} = \tilde{g}_{ab}$  and  $\bar{M}_{11}$  and  $\bar{M}_{12}$  in terms of  $\tilde{M}_c$  such that

$$\begin{pmatrix} \bar{M}_{11} & \bar{M}_{12} \\ \bar{M}_{12}^T & \bar{M}_{22} \end{pmatrix} = \tilde{M}_c.$$

Then, we can rewrite (5.2) as

$$\tilde{L}_c = \frac{1}{2} \begin{pmatrix} \dot{\mathbf{x}}_r^T & \dot{\mathbf{z}}_n^T \end{pmatrix} \begin{pmatrix} \bar{M}_{11} & \bar{M}_{12} \\ \bar{M}_{12}^T & \bar{M}_{22} \end{pmatrix} \begin{pmatrix} \dot{\mathbf{x}}_r \\ \dot{\mathbf{z}}_n \end{pmatrix} - V'_\epsilon(\mathbf{x}_r),$$

where  $\mathbf{z}_n = (z_n^a)^T$ .

Note that in these coordinates  $z_n^a$  is the symmetry variable. We are interested in the relative equilibria given by

$$(5.11) \quad \mathbf{v}_{RE} := \begin{pmatrix} \mathbf{x}_r \\ \dot{\mathbf{x}}_r \\ \dot{\mathbf{z}}_n \end{pmatrix},$$

where

$$\mathbf{x}_r = \mathbf{0}, \quad \dot{\mathbf{x}}_r = \mathbf{0}, \quad \dot{z}_n^d = \zeta^d,$$

and  $\zeta^d$  corresponds to ( $n$  times) the constant velocity of the center of mass of the network.

**DEFINITION 5.1** (amended potential [13]). *The amended potential for the Lagrangian system with Lagrangian (5.2) is defined by*

$$V_\mu(\mathbf{x}_r) = V'_\epsilon(\mathbf{x}_r) + \frac{1}{2} \tilde{g}^{cd} \mu_c \mu_d,$$

where  $V'_\epsilon$  is given by (4.5) and  $\tilde{g}_{ab}$  is given by (5.10). If  $J_a$  is the momentum conjugate to  $z_n^a$ , then  $\mu_a$  is  $J_a$  evaluated at the relative equilibrium corresponding to  $\dot{z}_n^a = \zeta^a$ , i.e.,

$$(5.12) \quad J_a = \frac{\partial \tilde{L}_c}{\partial \dot{z}_n^a} = (\bar{M}_{12}^T \dot{\mathbf{x}}_r + \bar{M}_{22} \dot{\mathbf{z}}_n)_a, \quad \mu_a = \left. \frac{\partial \tilde{L}_c}{\partial \dot{z}_n^a} \right|_{\mathbf{x}_r=0, \dot{\mathbf{x}}_r=0, \dot{z}_n^a=\zeta^a} = \tilde{g}_{ab} \zeta^b.$$

By the Routh criteria, the relative equilibrium is stable if the second variation of

$$(5.13) \quad E_\mu := \frac{1}{2} \dot{\mathbf{x}}_r^T (\bar{M}_{11} - \bar{M}_{12} \bar{M}_{22}^{-1} \bar{M}_{12}^T) \dot{\mathbf{x}}_r + V_\mu(\mathbf{x}_r),$$

evaluated at the origin, is definite. Also, if  $R^\mu(\mathbf{x}_r, \dot{\mathbf{x}}_r)$  is defined as

$$(5.14) \quad R^\mu := \frac{1}{2} \dot{\mathbf{x}}_r^T (\bar{M}_{11} - \bar{M}_{12} \bar{M}_{22}^{-1} \bar{M}_{12}^T) \dot{\mathbf{x}}_r - V_\mu(\mathbf{x}_r),$$

then the reduced Euler–Lagrange equations can be written as

$$\mathcal{E}_{x_r^\alpha} R^\mu = 0.$$

The Routhian  $R^\mu$  plays the role of a Lagrangian for the reduced system in variables  $(\mathbf{x}_r, \dot{\mathbf{x}}_r)$ . Since  $\tilde{g}_{ab}^i$  is a constant for each  $i \in \{1, 2, \dots, n\}$ , the second term in the amended potential  $V_\mu$  does not contribute to the second variation. It follows that the relative equilibrium with momentum  $\mu_a$  is stable if the matrix  $(\bar{M}_{11} - \bar{M}_{12} \bar{M}_{22}^{-1} \bar{M}_{12}^T)$  evaluated at the origin is negative definite, since the potential  $V_\epsilon'$  is already maximum at the equilibrium. But  $(\bar{M}_{11} - \bar{M}_{12} \bar{M}_{22}^{-1} \bar{M}_{12}^T)$  is negative definite because it is the Schur complement of the negative definite matrix  $\bar{M}_c$  [9].

**THEOREM 5.2 (SSRE).** *Consider a network of  $n$  SMC systems, each satisfying assumption **AS1**. Suppose for each system that the origin is an equilibrium and that the original potential energy is maximum at the origin. Consider the kinetic energy shaping defined in section 4 and potential energy coupling defined above with a connected graph so that the closed-loop dynamics derive from the Lagrangian  $\tilde{L}_c$  given by (5.2) and the potential energy  $V_\epsilon'$  is maximized at the relative equilibrium (5.11). The corresponding control law for the  $i$ th mechanical system is*

$$(5.15) \quad \begin{aligned} u_{a,i} = u_{a,i}^{\text{cons}} = & -\kappa_i \left\{ g_{\beta a, \gamma}^i - g_{\delta a}^i A_i^{\delta \alpha} \left[ g_{\alpha \beta, \gamma}^i - \frac{1}{2} g_{\beta \gamma, \alpha}^i - (1 + \kappa_i) g_{\alpha d}^i g_{\beta a, \gamma}^i \right] \right\} \dot{x}_i^\beta \dot{x}_i^\gamma \\ & + \kappa_i g_{\delta a}^i A_i^{\delta \alpha} \frac{\partial V_i}{\partial x_i^\alpha} + \frac{\partial V_i}{\partial \theta_i^\alpha} - \frac{1}{\rho_i} (1 + \kappa_i g_{\delta a}^i A_i^{\delta \alpha} g_{\alpha d}^i g_{\beta a}^{db}) \frac{\partial V_\epsilon'}{\partial \theta_i^\alpha}, \end{aligned}$$

where  $A_{\alpha \beta}^i = g_{\alpha \beta}^i - (1 + \kappa_i) g_{\alpha d}^i g_{\beta a}^{da} g_{\beta a}^i$ ,  $\rho_i < 0$ , and

$$\kappa_i + 1 > \max \{ \lambda | \det (g_{\alpha \beta}^i - \lambda g_{\alpha a}^i g_{\beta b}^{ab} g_{\beta b}^i) |_{x_i^\alpha=0} = 0 \}.$$

Then, the relative equilibrium (5.11) is an SSRE for any  $\zeta^d$ .

*Proof.* Since  $(\bar{M}_{11} - \bar{M}_{12} \bar{M}_{22}^{-1} \bar{M}_{12}^T)$  evaluated at the origin is negative definite, the second variation of  $E_\mu$  evaluated at the origin is definite. Hence, the relative equilibrium (5.11) is stable for the total network system independent of momentum value  $\mu_a$ .  $\square$

**6. Asymptotic stability of the constant momentum solution.** In this section we investigate asymptotic stabilization of the coordinated network to a solution corresponding to a constant momentum  $J_a = \mu_a$ . We prove that the solution is an ASSM. Recall from section 2 that an ASSM is an asymptotically stable solution of the mechanical system network defined by  $x_i^\alpha = x_j^\alpha$  and  $\theta_i^a = \theta_j^a$  for all  $i \neq j$  and dynamics that evolve on a constant momentum surface. An ASSM describes a fully synchronized motion, i.e., one in which each degree of freedom is synchronized across

the whole network. If  $x_i^\alpha = x_j^\alpha = 0$ , then the solution is a relative equilibrium. However, in general, an ASSM is *not* a relative equilibrium. For example, in the case of a network of pendulum/cart systems presented in section 8, the ASSM corresponds to periodic solutions (synchronized oscillations of pendula and carts). In this section we prove a control law that yields ASSM where the constant value of the momentum is given by the initial conditions. Equivalently, given an arbitrary momentum value  $\mu_a$ , initial conditions on the corresponding momentum surface converge to the ASSM on the same momentum surface. In the pendulum/cart example of section 8, we show that control gains can be used to determine the frequency of the periodic solution (ASSM). We discuss at the end of the section a second case in which a momentum value is prescribed and a control term is added to drive the ASSM to the prescribed constant momentum surface.

In this section we apply no dissipative control in the  $x_i^\alpha$  directions for all  $i$  and as our Case I below we use no control in the  $z_n^a$  direction. Recall that for our closed-loop system,  $z_n^a$  is the symmetry direction. If there is no control applied in this direction,  $J_a$  remains a constant; i.e., the system evolves on a constant momentum surface. On this surface,  $E_\mu$  as defined in (5.13) is a conserved quantity and can be chosen as a Lyapunov function to prove stability. By choosing appropriate dissipation in the nonsymmetry directions  $z_1^a, \dots, z_{n-1}^b$ , we prove that solutions on a constant momentum surface, corresponding to  $x_i^\alpha - x_j^\alpha = 0$  and  $\theta_i^a - \theta_j^a = 0$  for all  $i \neq j$ , are asymptotically stable, i.e., they are ASSM.

Let the control input for the  $i$ th mechanical system be

$$(6.1) \quad u_{a,i} = u_{a,i}^{\text{cons}} + \frac{1}{\rho_i} u_{a,i}^{\text{diss}},$$

where  $u_{a,i}^{\text{cons}}$  is the “conservative” control term given by (5.15) and  $u_{a,i}^{\text{diss}}$  is the dissipative control term to be designed. The Euler–Lagrange equations in the original coordinates for the  $i$ th uncontrolled system are

$$\mathcal{E}_{x_i^\alpha}(L_i) = 0; \quad \mathcal{E}_{\theta_i^a}(L_i) = u_{a,i}^{\text{cons}} + \frac{1}{\rho_i} u_{a,i}^{\text{diss}},$$

where  $L_i$  is given by (4.1).

In the new coordinates given by (5.1), we have for  $i = 1, \dots, n$ ,

$$(6.2) \quad \mathcal{E}_{x_i^\alpha}(\tilde{L}_c) = 0; \quad \mathcal{E}_{z_i^a}(\tilde{L}_c) = \frac{1}{n} \tilde{u}_{a,i}^{\text{diss}},$$

where  $\tilde{L}_c$  is given by (5.2) and

$$\begin{aligned} \tilde{u}_{a,i}^{\text{diss}} &= \sum_{j=1, j \neq i+1}^n u_{a,j}^{\text{diss}} - (n-1)u_{a,i+1}^{\text{diss}}, \quad i = 1, \dots, n-1, \\ \tilde{u}_{a,n}^{\text{diss}} &= \sum_{j=1}^n u_{a,j}^{\text{diss}}. \end{aligned}$$

*Case I.*  $\tilde{u}_{a,n}^{\text{diss}} = 0$ .

Let  $\tilde{E}_c$  be the energy function for the Lagrangian  $\tilde{L}_c$ . Given momentum value  $\mu_a$ , let  $\xi^b = \tilde{g}^{ab} \mu_a$ . Then, the function  $\tilde{E}_c^\xi$  defined by

$$\tilde{E}_c^\xi = \tilde{E}_c - J_a \xi^a$$

has the property that its restriction to the level set  $J_a = \mu_a = \tilde{g}_{ab}\xi^b$  of the momentum gives  $E_\mu$  (5.13). We can use this fact to calculate the time derivative of  $E_\mu$  as follows. From (6.2), we get

$$(6.3) \quad \frac{d}{dt}\tilde{E}_c = \frac{1}{n} \sum_{i=1}^n (\dot{z}_i^a \tilde{u}_{a,i}^{\text{diss}}).$$

Using (6.3) and the fact that  $\frac{d}{dt}J_a = \frac{1}{n}\tilde{u}_{a,n}^{\text{diss}}$ , we get

$$(6.4) \quad \frac{d}{dt}\tilde{E}_c^\xi = \frac{1}{n} \sum_{i=1}^n (\dot{z}_i^a \tilde{u}_{a,i}^{\text{diss}}) - \left( \frac{1}{n} \tilde{u}_{a,n}^{\text{diss}} \xi^a \right).$$

The expression for the time derivative of  $E_\mu$  is obtained by restricting  $\frac{d}{dt}\tilde{E}_c^\xi$  to the set  $J_a = \mu_a$ . This and (5.12) give us

$$\begin{aligned} \frac{d}{dt}E_\mu &= \frac{1}{n} \sum_{i=1}^{n-1} (\dot{z}_i^a \tilde{u}_{a,i}^{\text{diss}}) + \frac{1}{n} \tilde{u}_{a,n}^{\text{diss}} (\dot{z}_n^a|_{J_b=\mu_b} - \xi^a) \\ &= \frac{1}{n} \sum_{i=1}^{n-1} (\dot{z}_i^a \tilde{u}_{a,i}^{\text{diss}}) + \frac{1}{n} \tilde{u}_{a,n}^{\text{diss}} (\tilde{g}^{ab}(\mu_b - (\bar{M}_{12}^T \dot{\mathbf{x}}_r)_b) - \xi^a) \\ &= \frac{1}{n} \sum_{i=1}^{n-1} (\dot{z}_i^a \tilde{u}_{a,i}^{\text{diss}}) + \frac{1}{n} \tilde{u}_{a,n}^{\text{diss}} (-\tilde{g}^{ab}(\bar{M}_{12}^T \dot{\mathbf{x}}_r)_b). \end{aligned}$$

Here,  $\bar{M}_{12}^T \dot{\mathbf{x}}_r$  is a covariant vector just like a momentum. Hence, its components are denoted by subscripts. Since  $\tilde{u}_{a,n}^{\text{diss}}$  is chosen to be zero, we get

$$(6.5) \quad \frac{d}{dt}E_\mu = \frac{1}{n} \sum_{i=1}^{n-1} (\dot{z}_i^a \tilde{u}_{a,i}^{\text{diss}}).$$

Expressing  $\tilde{u}_{a,i}^{\text{diss}}$  in terms of  $u_{a,i}^{\text{diss}}$ , we can write the expression for  $\dot{E}_\mu$  as

$$(6.6) \quad n \frac{d}{dt}E_\mu = u_{a,1}^{\text{diss}} \left( \sum_{j=1}^{n-1} \dot{z}_j^a \right) + \sum_{j=2}^{n-1} u_{a,j}^{\text{diss}} \left( -(n-1)\dot{z}_{j-1}^a + \sum_{k=1, k \neq j-1}^{n-1} \dot{z}_k^a \right)$$

and choose

$$\begin{aligned} u_{a,1}^{\text{diss}} &= d_{ab} \left( \sum_{j=1}^{n-1} \dot{z}_j^b \right), \\ u_{a,j}^{\text{diss}} &= d_{ab} \left( -(n-1)\dot{z}_{j-1}^b + \sum_{k=1, k \neq j-1}^{n-1} \dot{z}_k^b \right), \\ (6.7) \quad &j = 2, \dots, n-1, \end{aligned}$$

where  $d_{ab}$  is a positive definite control gain matrix, possibly dependent on  $x_i^\alpha$ ,  $i = 1, \dots, n$ , and  $z_i^a$ ,  $j = 1, \dots, n-1$ . With the dissipative control term (6.7),  $\frac{d}{dt}E_\mu \geq 0$ .

We note that this dissipative control term requires that each individual system can measure the variables  $\dot{z}_i^a$  of all other vehicles. Recall that for Lyapunov stability

the interconnection among individual systems need only be *connected* for the coupling potential  $\tilde{V}_\epsilon$  which is a function of the  $y_k^a$ ,  $k = 1, \dots, n$ . That is, for Lyapunov stability, each individual system need only measure its relative position with respect to some subset of the other individual systems. However, for ASSM we require *complete* interconnection in the dissipative control term which is a function of the variables  $\dot{z}_n$ . That is, each individual system feedbacks relative velocity with respect to every other individual system. Figure 6.1 illustrates a complete interconnected graph for the case of four vehicles. Complete interconnection is not needed for stabilization of group dynamics in the simpler dynamical models used more typically in the literature, as described in section 1. It is hoped that the interconnection limitation here in stabilization of networks of underactuated mechanical systems can likewise be overcome in future work.

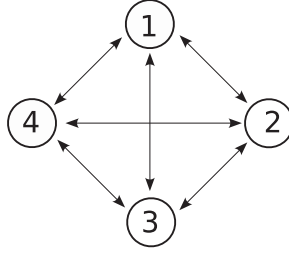


FIG. 6.1. Complete interconnected communication graph for four vehicles.

We next study convergence of the system using the LaSalle invariance principle [11]. For  $c > 0$ , let  $\Omega_c = \{(\mathbf{x}_r, \dot{\mathbf{x}}_r) | E_\mu \geq c\}$ .  $\Omega_c$  is a compact and positive invariant set with integral curves starting in  $\Omega_c$  and staying in  $\Omega_c$  for all  $t \geq 0$ . Define the LaSalle surface

$$\mathcal{E} = \left\{ (\mathbf{x}_r, \dot{\mathbf{x}}_r) \left| \frac{d}{dt} E_\mu = 0 \right. \right\}.$$

On this surface,  $u_{a,j}^{\text{diss}} = 0$ ,  $i = 1, \dots, n$ , which implies that  $\dot{z}_i^a = 0$  for  $i = 1, \dots, n-1$ . Let  $\mathcal{M}$  be the largest invariant set contained in  $\mathcal{E}$ . By the LaSalle invariance principle, solutions that start in  $\Omega_c$  approach  $\mathcal{M}$ . The relative equilibrium (5.11) is contained in  $\mathcal{M}$ ; however, there are other solutions in this set.

We now proceed to analyze in more detail the structure of solutions on the LaSalle surface  $\mathcal{E}$ . Using the condition  $\dot{z}_i^a = 0$  for  $i = 1, \dots, n-1$ , we get  $\dot{y}_i^a = \dot{y}_j^a$  for all  $i, j \in \{1, \dots, n\}$ . This gives  $y_i^a - y_j^a = \text{constant}$ . Since we have chosen  $\tilde{V}_\epsilon$  to be a quadratic function of the terms  $y_i^a - y_j^a$ , we get  $\frac{\partial \tilde{V}_\epsilon}{\partial y_i^a} = \text{constant} =: \Delta_a^i$ . The equations of motion for the  $y_i^a$  restricted to the LaSalle surface are  $\mathcal{E}_{y_i^a}(L'_c) = 0$ , where  $L'_c$  is given by (4.2). Equivalently,

$$(6.8) \quad \ddot{y}_i^a + \frac{d}{dt} (\tilde{g}_i^{ab} g_{ab}^i \dot{x}_i^a) = -\tilde{g}_i^{ab} \frac{\partial \tilde{V}_\epsilon}{\partial y_i^b} = -\tilde{g}_i^{ab} \Delta_b^i.$$

As illustrated in [1], for SMC systems, there is a function  $l_i^a(x_i^\alpha)$  for each vehicle  $i$  defined on an open set of the configuration space for the  $i$ th vehicle's unactuated variables such that

$$(6.9) \quad \frac{\partial l_i^a}{\partial x_i^\alpha} = \tilde{g}_i^{ac} g_{ac}^i.$$

We can assume, by shrinking  $\Omega_c$  if necessary, that (6.9) holds in  $\Omega_c$ .

Let  $K_c$  be the projection of  $\Omega_c$  onto the coordinates  $(\mathbf{x}_n, \dot{\mathbf{x}}_n)$ , where  $\mathbf{x}_n = (x_1^\alpha, \dots, x_n^\alpha)$ . Then, since  $l_i^a$  is continuous and  $K_c$  is compact, there exist constants  $m_i$  and  $n_i$  such that

$$(6.10) \quad m_i \leq \|l_i(x_i)\| \leq n_i$$

for all  $x_i^\alpha$  such that  $\mathbf{x}_n \in \Omega_c$ . Using (6.8), (6.9), and the condition  $\dot{y}_i^a = \dot{y}_j^a$  on  $\mathcal{E}$ , we get

$$(6.11) \quad \frac{d}{dt}(l_i^a - l_j^a) = \tilde{g}_j^{ab} \Delta_b^j - \tilde{g}_i^{ab} \Delta_b^i.$$

Therefore, on  $\mathcal{E}$ ,

$$(6.12) \quad l_i^a - l_j^a = \frac{1}{2}(\tilde{g}_j^{ab} \Delta_b^j - \tilde{g}_i^{ab} \Delta_b^i)t^2 + \nu_1^a t + \nu_2^a$$

for some constant vectors  $\nu_1^a$  and  $\nu_2^a$ . The only way (6.10) can also be satisfied is if  $\tilde{g}_j^{ab} \Delta_b^j - \tilde{g}_i^{ab} \Delta_b^i = 0$  and  $\nu_1^a = 0$ .

To simplify our calculations, we assume that the  $n$  individual mechanical systems are identical. In this case,  $\tilde{g}_j^{ab} = \tilde{g}_i^{ab}$  for any  $i, j \in \{1, \dots, n\}$ . This gives  $\Delta_a^i = \Delta_a^j$  for any  $i, j \in \{1, \dots, n\}$ , and so for a connected network with potential  $V'_\epsilon$  having a maximum at  $x_i^\alpha = 0$  and  $y_i^a = y_j^a$  for all  $i \neq j$ , we get that  $y_i^a = y_j^a$  on  $\mathcal{E}$  for all  $i, j \in \{1, \dots, n\}$ .

Using the definition (6.9) and the assumption that the individual systems are identical, the fact that  $\dot{l}_i^a - \dot{l}_j^a = 0$  on  $\mathcal{E}$  yields

$$(6.13) \quad g_{\alpha b}^i \dot{x}_i^\alpha = g_{\alpha b}^j \dot{x}_j^\alpha,$$

where  $g_{\alpha b}^k = g_{\alpha b}(x_k^\alpha)$  for all  $k = 1, \dots, n$ . Therefore, on the LaSalle surface  $\mathcal{E}$ , we see that solutions are of the form  $(\mathbf{x}_n(t), \dot{\mathbf{x}}_n(t), y_1^a(t), \dots, y_n^b(t), \dot{y}_1^a(t), \dots, \dot{y}_n^b(t))$ , where  $y_i^a(t) = y_j^a(t)$  for any  $i, j \in \{1, \dots, n\}$ ,  $J_a = \mu_a$ , and condition (6.13) holds. Since  $z_n^a = \sum_{i=1}^n y_i^a$  and the individual systems are identical, we have

$$\begin{aligned} J_a &= \frac{\partial \tilde{L}_c}{\partial \dot{z}_n^a} = \sum_{i=1}^n (g_{\alpha a}^i \dot{x}_i^\alpha + \tilde{g}_{ab} \dot{y}_i^b) \\ &= \tilde{g}_{ab} \sum_{i=1}^n (\tilde{g}^{bc} g_{\alpha c}^i \dot{x}_i^\alpha + \dot{y}_i^b) \\ &= n \tilde{g}_{ab} (\tilde{g}^{bc} g_{\alpha c}^i \dot{x}_i^\alpha + \dot{y}_i^b) \end{aligned}$$

for any  $i \in \{1, \dots, n\}$ , where we have used the facts that  $\dot{y}_i^a = \dot{y}_j^a$  and (6.13) holds on  $\mathcal{E}$ . Therefore, for each  $i$  we get

$$(6.14) \quad \dot{y}_i^a = \frac{1}{n} \tilde{g}^{ab} \mu_b - \tilde{g}^{ab} g_{\alpha b}^i \dot{x}_i^\alpha.$$

Substituting (6.14) into the closed-loop equations for the Lagrangian  $L'_c$  (4.2), we get the following equations for the  $x_i^\alpha$  variables:

$$(6.15) \quad \frac{d}{dt} \frac{\partial L^\mu}{\partial \dot{x}_i^\alpha} = \frac{\partial L^\mu}{\partial x_i^\alpha},$$



where

$$\begin{aligned}
 L^\mu &= \sum_{i=1}^n \left( \frac{1}{2} (\tilde{g}_{\alpha\beta}^i - \tilde{g}^{ab} g_{\alpha a}^i g_{\beta b}^i) \dot{x}_i^\alpha \dot{x}_i^\beta - V_{1i}(x_i^\alpha) \right) \\
 (6.16) \quad &= \sum_{i=1}^n \left( \frac{1}{2} (g_{\alpha\beta}^i - (\kappa + 1) g^{ab} g_{\alpha a}^i g_{\beta b}^i) \dot{x}_i^\alpha \dot{x}_i^\beta - V_{1i}(x_i^\alpha) \right),
 \end{aligned}$$

and  $V_{1i}$  is defined by assumption **AS1**. Here,  $\kappa_i = \kappa$  for all  $i = 1, \dots, n$ .

$L^\mu$  is just the Routhian  $R^\mu$  for a mechanical system with abelian symmetry variables without a linear term in velocity and without the amended part of the potential. This follows because, for SMC systems, these latter terms do not contribute to the dynamics of the reduced system. We also see that the  $x_i^\alpha$  dynamics completely decouple from the  $x_j^\alpha$  dynamics on the LaSalle surface  $\mathcal{E}$  for all  $i$  and  $j$ . The  $y_i^a$  dynamics given by (6.14) can be thought of as a reconstruction of dynamics in the symmetry variables, obtained after solving the reduced dynamics in the  $x_i^\alpha$  variables. We now make the following assumption.

**AS2.** Consider two solutions  $(x^\alpha(t), y^a(t))$  and  $(\tilde{x}^\alpha(t), \tilde{y}^a(t))$  of the Euler–Lagrange equations corresponding to the Lagrangian given by (3.3). If  $y^a(t) = \tilde{y}^a(t)$  and  $g_{\alpha a}(x^\alpha(t))\dot{x}^\alpha(t) = g_{\alpha a}(\tilde{x}^\alpha(t))\dot{\tilde{x}}^\alpha(t)$ , then  $x^\alpha(t) = \tilde{x}^\alpha(t)$ .

Note that checking this condition does not require extensive computation since we already know the expression for the closed-loop Lagrangian. Consider two solutions  $x^\alpha(t)$  and  $\tilde{x}^\alpha(t)$  such that  $g_{\alpha a}(x^\alpha(t))\dot{x}^\alpha(t) = g_{\alpha a}(\tilde{x}^\alpha(t))\dot{\tilde{x}}^\alpha(t)$ . This is equivalent to  $l^a(x^\alpha) = l^a(\tilde{x}^\alpha) + c^a$ , where  $l^a$  is defined by (6.9) and  $c^a$  is a constant; i.e.,  $x^\alpha(t)$  and  $\tilde{x}^\alpha(t)$  are two solutions in  $(x^\alpha, \dot{x}^\alpha)$  space satisfying the Euler–Lagrange equation corresponding to the Lagrangian  $L^\mu$  given by (6.16) and differing by a constant. For mechanical systems with symmetries, it may be possible to prove that  $c^a$  is zero, as is done for the pendulum/cart case in section 8. Then, **AS2** is equivalent to assuming that the function  $l^a$  is injective, i.e.,  $g_{\alpha a}$  is one-to-one in a neighborhood about the equilibrium. For the pendulum/cart example in section 8, this holds in the neighborhood defined by pendulum angles which are above the horizontal plane. As mentioned in [1], **AS2** is equivalent to the (local) strong inertial coupling assumption in [21] and internal/external convertible system in [7].

Using (6.13) and the fact that  $y_i^a = y_j^a$  on the LaSalle surface, we get from **AS2** that  $x_i^\alpha = x_j^\alpha$  and  $\theta_i^a = \theta_j^a$  for all  $i, j \in \{1, \dots, n\}$ . So we get that the dissipation control law given by (6.7) yields asymptotic convergence to synchronized motion on a constant momentum surface.

**THEOREM 6.1 (ASSM).** Consider a network of  $n$  identical SMC systems that each satisfy **AS1** and **AS2**. Suppose for each individual system that the origin is an equilibrium and that the original potential energy is maximum at the origin. Consider the kinetic energy shaping defined in section 4 and potential energy coupling  $\tilde{V}_\epsilon$  defined in section 5, where the terms in  $\tilde{V}_\epsilon$  are quadratic in  $y_i^a - y_j^a$  and the corresponding interconnection graph is connected. The closed-loop dynamics (6.2) derive from the Lagrangian  $\tilde{L}_c$  given by (5.2), and the potential energy  $V'_\epsilon$  is maximized at the relative equilibrium (5.11). The control input takes the form (6.1), where  $u_{a,i}^{\text{cons}}$  is given by (5.15) and  $\rho_i = \rho$ ,  $\kappa_i = \kappa$ . The dissipative control term given by (6.7) asymptotically stabilizes the solution in which all the vehicles have synchronized dynamics such that  $\theta_i^a = \theta_j^a$  and  $x_i^\alpha = x_j^\alpha$  for all  $i$  and  $j$ , and each has the same constant momentum in the  $\theta_i^a$  direction. The system stays on the constant momentum surface determined by the initial conditions.

*Remark 6.2.* Consider a Case II in which we choose  $\tilde{u}_{a,n}^{\text{diss}} = -\lambda(J_a - \mu_a)$  and  $u_{a,i}$  for  $i = 1, \dots, n-1$  as in Case I. Then  $J_a = (J_a(0) - \mu_a) \exp(-\lambda t) + \mu_a$  and we can rewrite the reduced system in  $(\mathbf{x}_r, \dot{\mathbf{x}}_r)$  coordinates as follows:

$$(6.17) \quad \mathcal{E}_{\mathbf{x}_r}(R^\mu) = \begin{pmatrix} 0 \\ \frac{1}{n} \tilde{\mathbf{u}}^{\text{diss}} \end{pmatrix} + \lambda \bar{M}_{12} \bar{M}^{22} (\mathbf{J}(0) - \boldsymbol{\mu}) \exp(-\lambda t).$$

Here,  $\tilde{\mathbf{u}}^{\text{diss}} = (\tilde{u}_{a,1}^{\text{diss}}, \dots, \tilde{u}_{a,n-1}^{\text{diss}})$  is an  $rn$ -dimensional vector, and  $\mathbf{J}$  and  $\boldsymbol{\mu}$  are  $r$ -dimensional vectors with components  $J_a$  and  $\mu_b$ , respectively. When  $\lambda = 0$ , we get Case I. When  $\lambda \neq 0$ , the momentum  $J_a$  is no longer a conserved quantity. This case needs to be analyzed more carefully since we are pumping energy into the system now to drive it to a particular momentum value. Equation (6.17) can be considered to be a parameter dependent differential equation with the parameter being  $\lambda$ . When  $\lambda = 0$ , we already know the solution from Case I. From the continuity of dependence of solutions upon parameters, we get that when  $0 < \lambda < \delta$ , the solution stays within an  $\epsilon$ -tube of the solution in Case I for time  $t \in [0, t_1]$  for some  $t_1$  if the initial conditions are in a  $\delta$ -neighborhood. Our simulations for pendulum/cart systems suggest that this holds true for the infinite time interval. We plan to investigate this case further in our future work.

*Remark 6.3.* The simplifying requirement for Theorem 6.1 that all systems be identical is a weakness of the result and motivates the question of robustness to uncertainty in system parameters. Simulations suggest that the stability of Theorem 6.1 is robust to model parameter uncertainty, but a formal robustness analysis is warranted.

In section 8 we illustrate the result of Theorem 6.1 and the dynamics of (6.16) in more detail in the case of a network of inverted pendulum/cart systems. Solutions for this example correspond to synchronized balanced pendula on synchronized moving carts, where the motion of the carts is the sum of a constant velocity plus an oscillation and the motion of the pendula is oscillatory with the same frequency as the carts.

**7. Asymptotic stabilization of relative equilibria.** In the previous section, we proved asymptotic stability of the coordinated network in the case when the network asymptotically converges to the momentum surface  $J_a = \mu_a$ . This can lead to nontrivial and interesting synchronized group dynamics, as is discussed in section 8. Stabilization was proved using  $E_\mu$  as a Lyapunov function on the reduced space. The dynamics after adding a dissipative control term are given by  $\theta_i^a = \theta_j^a$  and  $x_i^\alpha = x_j^\alpha$  for all  $i, j = 1, \dots, n$ . The dissipative terms are chosen such that the momentum is preserved.

In this section, we demonstrate how to isolate and asymptotically stabilize the particular synchronized and constant momentum solutions corresponding to the relative equilibria given by (5.11). The value of the momentum  $\mu_a$  can be chosen arbitrarily. We use a different Lyapunov function from that used in section 6. We note that in the example of a network of inverted pendulum/cart systems, the relative equilibrium corresponds to the synchronized motion of all carts moving in unison at a steady speed with all pendula at rest in the upright position; i.e., it is the special case of the motion proved in Theorem 6.1 without the oscillation.

Consider the following function:

$$(7.1) \quad E_{RE} = \frac{1}{2} (\dot{\mathbf{x}}_c - \mathbf{v}_{RE})^T \tilde{M}_c (\dot{\mathbf{x}}_c - \mathbf{v}_{RE}) + V'_\epsilon,$$

where  $\mathbf{v}_{RE}$  is defined by (5.11).  $E_{RE}$  is a Lyapunov function in directions transverse to the group orbit of the relative equilibrium, i.e.,  $E_{RE} > 0$  in a neighborhood of the Euler–Lagrange solution given by  $(\mathbf{x}_r, \mathbf{z}_n, \dot{\mathbf{x}}_r, \dot{\mathbf{z}}_n)$ , where  $\mathbf{x}_r = \mathbf{0}$ ,  $\dot{\mathbf{x}}_r = \mathbf{0}$ ,  $\mathbf{z}_n^d = \zeta^d t$ ,  $\dot{\mathbf{z}}_n^d = \zeta^d$ , and  $\zeta^d$  corresponds to ( $n$  times) the constant velocity of the center of mass of the network.

The time derivative of  $E_{RE}$  along the flow given by (6.2) can be computed to be

$$\frac{d}{dt}E_{RE} = \frac{1}{n}(\dot{\mathbf{x}}_c - \mathbf{v}_{RE}) \cdot \begin{pmatrix} 0 \\ \tilde{\mathbf{u}}^{\text{diss}} \end{pmatrix}.$$

See [3] for the steps involved in proving this identity. Choose

$$(7.2) \quad \tilde{u}_{a,i}^{\text{diss}} = \begin{cases} n\sigma_i \dot{z}_i^a & \text{for } i = 1, \dots, n-1, \\ n\sigma_n(\dot{z}_n^a - \zeta^a) & \text{for } i = n, \end{cases}$$

where control parameters  $\sigma_i$  are positive constants. Then,

$$\frac{d}{dt}E_{RE} = \sum_{j=1}^{n-1} \sigma_j (\dot{z}_j^a)^2 + \sigma_n (\dot{z}_n^a - \zeta^a)^2 \geq 0.$$

We note here that, unlike the case of asymptotic stabilization in the previous section, where a complete interconnection was required to realize the dissipative control term (6.7), the dissipative control term (7.2) requires only a connected interconnection graph.

Let  $\Omega_c^{RE} = \{(\mathbf{x}_r, \dot{\mathbf{x}}_r, \dot{\mathbf{z}}_n^a) | E_{RE} \geq c\}$  for  $c > 0$ .  $\Omega_c^{RE}$  is a compact set, i.e.,  $E_{RE}$  is a proper Lyapunov function. Assume that the Euler–Lagrange system (6.2) satisfies the following controllability condition.

**AS3.** *The system (6.2) is linearly controllable at each point in a neighborhood of the relative equilibrium solution manifold.*

Note that checking this condition does not require extensive computation since we already know the expression for the closed-loop Lagrangian.

We now use a result from nonlinear control theory, which is stated in [3] as Lemma 2.1 and the remark following it, to conclude that the system (6.2) with dissipative control terms given by (7.2) converges exponentially to the set

$$\mathcal{E}_{RE} = \{(\mathbf{x}_r, \dot{\mathbf{x}}_r, \dot{\mathbf{z}}_n^a) | E_{RE} = 0\}.$$

On this set, the solution is given by (5.11). Thus, we have shown that the solutions of the controlled system will exponentially converge to  $(x_i^\alpha, \theta_i^\alpha, \dot{x}_i^\beta, \dot{\theta}_i^\beta) = (0, \frac{1}{n}\zeta^a t + \gamma^a, 0, \frac{1}{n}\zeta^b)$ , with  $\gamma^a$  constant.

**THEOREM 7.1 (ASSRE).** *Consider a network of  $n$  (not necessarily identical) individual SMC systems that each satisfy assumption **AS1**. Suppose for each individual system that the origin is an equilibrium and that the original potential energy is maximum at the origin. Consider the kinetic energy shaping defined in section 4 and potential energy coupling  $\tilde{V}_\epsilon$  defined in section 5, where the terms in  $\tilde{V}_\epsilon$  are quadratic in  $y_i^a - y_j^a$  and the corresponding interconnection graph is connected. The closed-loop dynamics (6.2) derive from the Lagrangian  $\tilde{L}_c$  given by (5.2) and the potential energy  $V'_\epsilon$  is maximized at the relative equilibrium (5.11). The control input takes the form (6.1), where  $u_{a,i}^{\text{cons}}$  is given by (5.15) and  $\rho_i = \rho$ . If (6.2) satisfies **AS3**, then the dissipative control term given by (7.2) exponentially stabilizes the relative equilibrium given by (5.11) in which  $x_i^\alpha = \dot{x}_i^\alpha = 0$  for all  $i = 1, \dots, n$  and  $\theta_i^a = \theta_j^a$  and  $\dot{\theta}_i^a = \dot{\theta}_j^a = \frac{1}{n}\zeta^a$  for all  $i$  and  $j$ .*

**8. Coordination of multiple inverted pendulum/cart systems.** As an illustration, we now consider the coordination of  $n$  identical planar inverted pendulum/cart systems. For the  $i$ th system, the pendulum angle relative to the vertical is  $x_i$  and the position of the cart is  $\theta_i$ . Let the Lagrangian for each system shown in Figure 8.1 be

$$L_i = \frac{1}{2}\alpha\dot{x}_i^2 + \beta\cos(x_i)\dot{x}_i\dot{\theta}_i + \frac{1}{2}\gamma\dot{\theta}_i^2 + D\cos(x_i); \quad i = 1, \dots, n,$$

where  $l, m, M$  are the pendulum length, pendulum bob mass, and cart mass, respectively.  $g$  is the acceleration due to gravity. The quantities  $\alpha, \beta, \gamma$ , and  $D$  are expressed in terms of  $l, m, M, g$  as follows:

$$\alpha = ml^2, \quad \beta = ml, \quad \gamma = m + M, \quad D = -mgl.$$

The equations of motion for the  $i$ th system are

$$\begin{aligned} \mathcal{E}_{x_i}(L_i) &= 0, \\ \mathcal{E}_{\theta_i}(L_i) &= u_i, \end{aligned}$$

where  $u_i$  is the control force applied to the  $i$ th cart.

One can see that  $\theta_i$  is a symmetry variable. Further, it can be easily verified that each pendulum/cart system satisfies the simplified matching conditions [1, 2]. The  $n$  inverted planar pendulum/cart systems lie on  $n$  parallel tracks corresponding to the  $\theta_i$  directions. The coordination problem is to prescribe control forces  $u_i$ ,  $i = 1, \dots, n$ , that asymptotically stabilize the solution where each pendulum is in the vertical upright position (in the case of ASSRE) or moving synchronously (in the case of ASSM) and the carts are moving at the same position along their respective tracks with the same common velocity. The relative equilibrium  $\mathbf{v}_{RE}$  (5.11) corresponds to  $x_i = \dot{x}_i = 0$  for all  $i$ ,  $\theta_i = \theta_j$  for all  $i \neq j$ , and  $\dot{\theta}_i = \frac{1}{n}\zeta$  for some constant scalar velocity  $\zeta$ .

Following (5.2), the closed-loop Lagrangian for the total system in the coordinates  $\mathbf{x}_c = (x_1, \dots, x_n, z_1, \dots, z_n)$ , where  $z_i = y_1 - y_{i+1}$  for  $i = 1, \dots, n-1$ ,  $z_n = y_1 + \dots + y_n$ ,  $y_i = \theta_i + p \sin x_i$ , and  $p = \frac{\beta}{\gamma}(\kappa + 1 - \frac{1}{\rho})$ , is

$$(8.1) \quad \tilde{L}_c = \frac{1}{2}\dot{\mathbf{x}}^T \tilde{M}_c \dot{\mathbf{x}} - V'_\epsilon(x_1, \dots, x_n, z_1, \dots, z_{n-1}).$$

$\tilde{M}_c$  is as in (5.6) and  $M_c$  is as in (4.3),

$$(8.2) \quad \begin{aligned} \tilde{g}_{\alpha\beta}^i &= \alpha - \left(\kappa + 1 - \frac{1}{\rho}\right) \frac{\beta^2}{\gamma} \cos^2(x_i), & \tilde{g}_{\alpha\alpha}^i &= \beta \cos(x_i), \\ \tilde{g}_{ab}^i &= \rho\gamma, & V'_\epsilon &= -D \sum_{i=1}^{n-1} \left( \cos(x_i) - \frac{1}{2}\epsilon \frac{\gamma^2}{\beta^2} z_i^2 \right) - D \cos(x_n) \end{aligned}$$

with  $\epsilon > 0$ . The control law (6.1) for the  $i$ th system is

$$(8.3) \quad u_i = \frac{\kappa\beta \left( \sin x_i \left( \alpha\dot{x}_i^2 + \cos(x_i)D \right) - B_i \left( \frac{\partial V'_\epsilon}{\partial \theta_i} - u_i^{\text{diss}} \right) \right)}{\alpha - \frac{\beta^2}{\gamma} (1 + \kappa) \cos^2(x_i)},$$

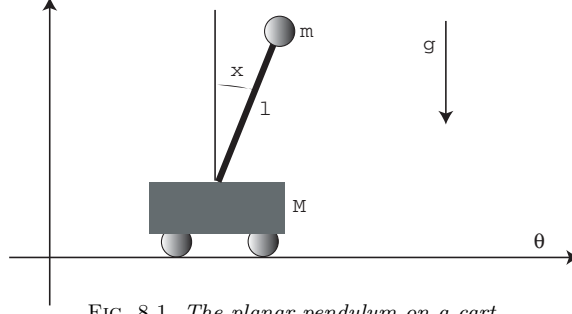


FIG. 8.1. The planar pendulum on a cart.

where

$$B_i = \frac{1}{\rho} \left( \alpha - \frac{\beta^2 \cos^2(x_i)}{\gamma} \right).$$

Note that we have chosen  $\rho_i = \rho$  and  $\kappa_i = \kappa$ . In the case  $u_i^{\text{diss}} = 0$ , by Theorem 5.2, we get stability of the relative equilibrium  $\mathbf{v}_{RE}$  (SSRE) if we choose  $\rho < 0$ ,  $\epsilon > 0$ , and  $\kappa$  such that  $m_\kappa := \alpha - (\kappa + 1) \frac{\beta^2}{\gamma} < 0$ . The choice of  $u_i^{\text{diss}}$  depends upon what kind of asymptotic stability we want, i.e., convergence to a synchronized constant momentum solution or to a relative equilibrium.

The dependence of  $V'_\epsilon$  on  $z_i^2$  in (8.2) implies that coupling between the pendulum/cart systems introduced by the control is a function of terms  $y_i - y_j$  rather than  $\theta_i - \theta_j$ . That is, our approach to simultaneous stabilization and synchronization of a network of planar pendulum/cart systems yields coupling not simply as a function of relative cart positions but, rather more subtly, as a function of the horizontal component of relative positions of pendulum bobs (where pendulum length is scaled by  $p$ ). Numerical simulations show that naively coupling the positions of the carts for the purpose of synchronization in fact destabilizes the network. This particular example illustrates the need to integrate synchronization and stabilization tasks.

### 8.1. Asymptotic stability on a constant momentum surface (ASSM).

Following (6.7), we let  $u_1^{\text{diss}}$  be

$$u_1^{\text{diss}} = d_1 \left( \sum_{k=1}^{n-1} (\dot{z}_k) \right)$$

and  $u_i^{\text{diss}}$  for  $i = 2, \dots, n$  be

$$u_i^{\text{diss}} = d_i \left( -(n-1)\dot{z}_{i-1} + \sum_{k=1, k \neq i-1}^{n-1} \dot{z}_k \right),$$

where coefficients  $d_i$  are constant positive scalars.

We now analyze the dynamics on the LaSalle surface. On this surface, we have  $\dot{y}_i = \dot{y}_j$  for all  $i, j \in \{1, \dots, n\}$  and  $J = \mu$ , where momentum  $\mu$  is determined by the initial conditions. From the calculations made in section 6, we also get  $y_i = y_j$  and  $\cos(x_i)\dot{x}_i = \cos(x_j)\dot{x}_j$ . The  $x_i$  dynamics are given by (6.15) with

$$(8.4) \quad L^\mu = \sum_{i=1}^n \left( \frac{1}{2} \left( \alpha - (\kappa + 1) \frac{\beta^2}{\gamma} \cos^2(x_i) \right) \dot{x}_i^2 + D \cos(x_i) \right).$$

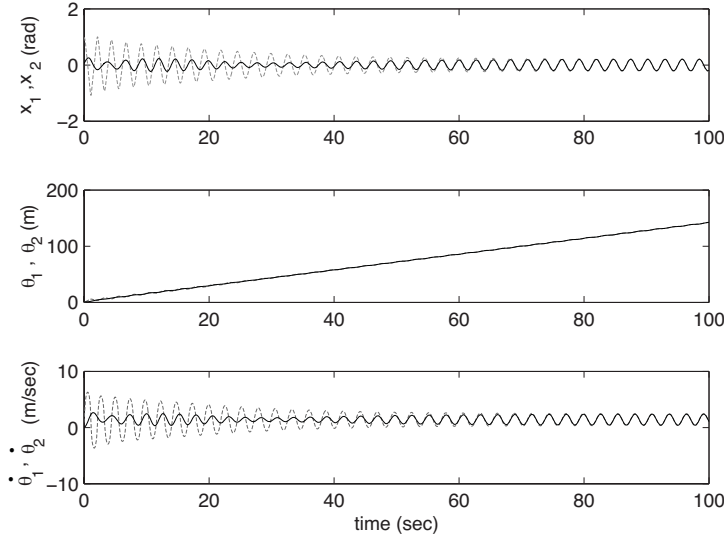


FIG. 8.2. Simulation of a controlled network of pendulum/cart systems with dissipation designed for ASSM. The pendulum angle, cart position, and cart velocity are plotted as a function of time for each of the two pendulum/cart systems in the network.

To verify **AS2** we need to check that if  $\cos(x_i)\dot{x}_i = \cos(x_j)\dot{x}_j$  about the origin for a system corresponding to the Lagrangian  $L^\mu$ , then  $x_i = x_j$  identically. This condition can also be written as  $\sin(x_i) = \sin(x_j) + c$ , where  $c$  is a constant. Note that if  $x_i(t)$  is an Euler–Lagrange solution corresponding to  $L^\mu$  for the  $i$ th vehicle, then  $-x_i(t)$  is also a solution. Since we have a stable pendulum oscillation about the upright position,  $x_i(t)$  and therefore  $|\sin(x_i(t))|$  oscillates with mean zero for all  $i$ . This can also be concluded from the fact that the solution curves are closed level curves in the  $(x_i, \dot{x}_i)$  plane of  $L^\mu$  given by (8.4) and  $L^\mu$  is invariant under the sign change  $(x_i, \dot{x}_i) \mapsto -(x_i, \dot{x}_i)$ . Since  $|\sin(x_i)|$  oscillates with zero mean for all  $i$ , the constant  $c$  must be zero. Hence,  $x_i(t) = x_j(t)$  for all  $i, j$  identically and **AS2** is verified. Thus, by Theorem 6.1 the pendulum network asymptotically goes to an ASSM.

From (8.4), it can be seen that on the LaSalle surface, the dynamics of  $x_i$  are decoupled from the dynamics of  $x_j$  for all  $i \neq j$ . For small  $x_i$ , the dynamics of each individual term in  $L^\mu$  corresponds to the stable dynamics of a spring-mass system with a  $\kappa$ -dependent mass  $-m_\kappa > 0$  and spring constant  $-D > 0$ . The mass  $-m_\kappa$ , which determines the oscillation frequency of the pendulum for each individual cart, can be controlled by the choice of  $\kappa$ . For the nonlinear system also, constant energy curves are closed curves in the  $(x_i, \dot{x}_i)$  plane. Hence, we have a periodic orbit for the angle made by each pendulum with the vertical line with a  $\kappa$ -dependent frequency. On the LaSalle surface,  $J = \rho\gamma\dot{\theta}_i + (\beta + p\rho\gamma)\cos(x_i)\dot{x}_i = \text{constant}$ . Therefore, the velocity of the cart  $\dot{\theta}_i$  oscillates about a constant velocity with the same frequency as the pendulum oscillation.

Figure 8.2 shows the results of a MATLAB simulation for the controlled network of pendulum/cart systems using the following values for the system parameters. The pendulum/cart systems have identical pendulum bob masses, lengths, and cart masses. The pendulum bob mass is chosen to be  $m = 0.14$  kg, cart mass is  $M = 0.44$  kg, and pendulum length is  $l = 0.215$  m. The control gains are  $\rho = -0.27$ ,  $\kappa = 40$ ,  $d_i = d = 0.2$ , and  $\epsilon = 0.0005$ . We compute  $m_\kappa = -0.058 \text{ kgm}^2 < 0$  as required for

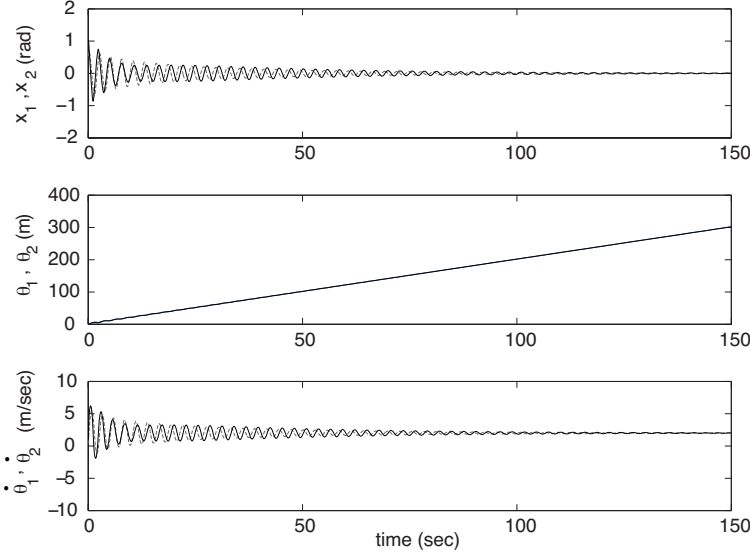


FIG. 8.3. Simulation of a controlled network of pendulum/cart systems with dissipation designed for ASSRE. The pendulum angle, cart position, and cart velocity are plotted as a function of time for each of the two pendulum/cart systems in the network.

stability. The initial conditions for the two systems shown are

$$\begin{pmatrix} x_1(0) & \dot{x}_1(0) & \theta_1(0) & \dot{\theta}_1(0) & x_2(0) & \dot{x}_2(0) & \theta_2(0) & \dot{\theta}_2(0) \end{pmatrix} \\ = ( 0.48 \quad 0.99 \quad 0.37 \quad 0.53 \quad 0.18 \quad 0.50 \quad 0.42 \quad 0.66 ).$$

Figure 8.2 shows plots of the pendulum angle, cart position, and cart velocity as a function of time for two of the coupled pendulum/cart systems. Convergence to an ASSM is evident. The frequency of oscillation of the pendula can be observed to be the same as the frequency of oscillation in the cart velocities. This frequency of oscillation can be computed as  $\omega = \sqrt{D/m_\kappa}$  and the period of oscillation as  $T = 2\pi/\omega = 2.8$  s, which is precisely the period of the oscillations observed in Figure 8.2.

**8.2. Asymptotic stability of relative equilibria (ASSRE).** In this case, we want to asymptotically stabilize the relative equilibrium  $\mathbf{v}_{RE}$ , i.e.,  $x_i = \dot{x}_i = 0$  for all  $i$ ,  $\theta_i = \theta_j$  for all  $i \neq j$ , and  $\dot{\theta}_i = \frac{1}{n}\zeta$  for all  $i$ , and any constant scalar velocity  $\zeta$ . Recall that this corresponds to each pendulum angle at rest in the upright position and all carts aligned and moving together with the same constant velocity  $\frac{1}{n}\zeta$ . Following (7.2), we let

$$u_i^{\text{diss}} = nd_i \dot{z}_i$$

for  $i = 1, \dots, n-1$  and

$$u_n^{\text{diss}} = nd_n(\dot{z}_n - \zeta),$$

where the control parameters  $d_i$  are positive constants.

Figure 8.3 shows the results of a MATLAB simulation for the controlled network of pendulum/cart systems with this dissipative control. We choose  $\zeta = 2n$  m/s, and

the remaining system and control parameters are as above in the ASSM case. The initial conditions for the two systems shown are

$$\begin{pmatrix} x_1(0) & \dot{x}_1(0) & \theta_1(0) & \dot{\theta}_1(0) & x_2(0) & \dot{x}_2(0) & \theta_2(0) & \dot{\theta}_2(0) \end{pmatrix} \\ = ( 0.53 \quad 1.12 \quad 0.56 \quad 0.50 \quad 1.02 \quad 0.63 \quad 0.24 \quad 0.81 ).$$

Figure 8.3 shows convergence to the relative equilibrium; the pendula are stabilized in the upright position, the cart positions become synchronized, and the cart velocities converge to 2 m/s.

**9. Final remarks.** We have derived control laws to stabilize and stably synchronize a network of mechanical systems with otherwise unstable dynamics. We have proved stability of relative equilibria corresponding to synchronization in all variables and common steady motion in the actuated directions. Using two different choices of a dissipative term in the control law, we prove two different kinds of asymptotic stability. In the first case of dissipation, we show how to drive the network to a synchronized motion on the constant momentum surface determined by the initial conditions. Such a synchronized motion can be interesting when examined in physical space. In our example of a network of planar pendulum/cart systems, we show that the synchronized motion is periodic and the period of the oscillation can be controlled with a control parameter. In the second case of dissipation, we show how to isolate and asymptotically stabilize the relative equilibrium for any choice of constant momentum. We illustrate all of our results for a network of pendulum/cart systems. For this example, our approach yields a subtle choice in the coupling variables: The coupling that leads to stable synchronization is a function of relative positions of pendulum bobs, not simply relative positions of carts. Indeed, coupling as a function of relative cart positions destabilizes the network.

For asymptotic stabilization of the relative equilibrium, we assume that the interconnection graph for the network is connected. However, for asymptotic stabilization of a synchronized motion on the constant momentum surface, we assume that the interconnection graph for the dissipative control is completely connected. It is of interest in future work to determine whether this latter condition can be relaxed.

In Theorem 6.1 we prove asymptotic stabilization of a synchronized motion on the constant momentum surface; however, we cannot select the value of the momentum because it is determined by the initial conditions. In Remark 6.2 we propose a control law to simultaneously drive the momentum to a desired value. This control law appears to work in simulation; however, the stability analysis is more subtle. It raises a number of interesting questions. For example, suppose we have a dynamical system depending upon a parameter  $\lambda$ , i.e., the Lagrangian is given by a function  $L(q, \dot{q}, \lambda)$ , where  $q$  is the state variable. Assume that for each  $\lambda \in [0, \epsilon]$ , the (controlled) system is Lyapunov stable. If we now let  $\lambda$  evolve in time such that it “slowly” goes to a value  $\bar{\epsilon} \in (0, \epsilon)$ , can we still conclude that the system is Lyapunov stable in the infinite time domain? See [12] for results in the case when the unperturbed system has a uniformly asymptotically stable equilibrium. We plan to build on these tools to study our parameter dependency problem in future work.

Another future direction is the inclusion of collision avoidance in our framework. For instance, in our example, the carts move on parallel tracks, and hence collision avoidance is not an issue. However, it is interesting to consider the case in which all of the carts are on the same track and the pendulum/cart systems can be controlled without collisions for stable synchronization.



## REFERENCES

- [1] A. M. BLOCH, D. E. CHANG, N. E. LEONARD, AND J. E. MARSDEN, *Controlled Lagrangians and the stabilization of mechanical systems II: Potential shaping*, IEEE Trans. Automat. Control, 46 (2001), pp. 1556–1571.
- [2] A. M. BLOCH, N. E. LEONARD, AND J. E. MARSDEN, *Controlled Lagrangians and the stabilization of mechanical systems I: The first matching theorem*, IEEE Trans. Automat. Control, 45 (2000), pp. 2253–2270.
- [3] F. BULLO, *Stabilization of relative equilibria for underactuated systems on Riemannian manifolds*, Automatica, 36 (2000), pp. 1819–1834.
- [4] J. P. DESAI, J. P. OSTROWSKI, AND V. KUMAR, *Modeling and control of formations of non-holonomic mobile robots*, IEEE Trans. Robotics and Automation, 17 (2001), pp. 905–908.
- [5] E. FRAZZOLI, *Robust Hybrid Control for Autonomous Vehicle Motion Planning*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001.
- [6] E. FRAZZOLI, *Maneuver-based motion planning and coordination for multiple UAVS*, in Proceedings of the 21st AIAA/IEEE Digital Avionics Systems Conference, Vol. 2, 2002, IEEE Press, Piscataway, NJ, pp. 1–12.
- [7] N. H. GETZ, *Dynamic Inversion of Nonlinear Maps with Applications to Nonlinear Control and Robotics*, Ph.D. thesis, University of California at Berkeley, Berkeley, CA, 1996.
- [8] H. HANSSMANN, N. E. LEONARD, AND T. R. SMITH, *Symmetry and reduction for coordinated rigid bodies*, European J. Control, 12 (2006), pp. 176–194.
- [9] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
- [10] A. JADBABAIE, J. LIN, AND A. S. MORSE, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.
- [11] H. KHALIL, *Nonlinear Systems*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.
- [12] H. K. KHALIL AND P. V. KOKOTOVIC, *On stability properties of nonlinear systems with slowly varying inputs*, IEEE Trans. Automat. Control, 36 (1991), pp. 229–229.
- [13] J. E. MARSDEN, *Lectures on Mechanics*, Cambridge University Press, Cambridge, UK, 1992.
- [14] S. NAIR AND N. E. LEONARD, *Stabilization of a coordinated network of rotating rigid bodies*, in Proceedings of the 43rd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2004, pp. 4690–4695.
- [15] S. NAIR AND N. E. LEONARD, *Stabilization synchronization of rigid body networks*, Networks and Heterogeneous Media, 2 (2007), pp. 595–624.
- [16] S. NAIR, N. E. LEONARD, AND L. MOREAU, *Coordinated control of networked mechanical systems with unstable dynamics*, in Proceedings of the 42nd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2003, pp. 550–555.
- [17] P. ÖGREN, E. FIORELLI, AND N. E. LEONARD, *Cooperative control of mobile sensor networks: Adaptive gradient climbing in a distributed environment*, IEEE Trans. Automat. Control, 49 (2004), pp. 1292–1302.
- [18] R. OLFATI-SABER AND R. M. MURRAY, *Graph rigidity and distributed formation stabilization of multi-vehicle systems*, in Proceedings of the 41st IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2002, pp. 2965–2971.
- [19] R. ORTEGA, M. W. SPONG, F. GÓMEZ-ESTERN, AND G. BLANKENSTEIN, *Stabilization of underactuated mechanical systems via interconnection and damping assignment*, IEEE Trans. Automat. Control, 47 (2002), pp. 1281–1233.
- [20] E. D. SONTAG AND Y. WANG, *On characterizations of the input-to-state stability properties*, Systems Control Lett., 24 (1995), pp. 351–359.
- [21] M. W. SPONG, *The control of underactuated mechanical systems*, in Proceedings of the 1st International Conference on Mechatronics, Mexico City, Mexico, 1994, pp. 26–29.
- [22] H. G. TANNER, G. J. PAPPAS, AND V. KUMAR, *Leader-to-formation stability*, IEEE Trans. Robotics and Automation, 20 (2004), pp. 443–455.

## OPTIMAL STOPPING GAMES FOR MARKOV PROCESSES\*

ERIK EKSTRÖM<sup>†</sup> AND GORAN PESKIR<sup>†</sup>

**Abstract.** Let  $X = (X_t)_{t \geq 0}$  be a strong Markov process, and let  $G_1$ ,  $G_2$ , and  $G_3$  be continuous functions satisfying  $G_1 \leq G_3 \leq G_2$  and  $\mathbb{E}_x \sup_t |G_i(X_t)| < \infty$  for  $i = 1, 2, 3$ . Consider the optimal stopping game where the sup-player chooses a stopping time  $\tau$  to maximize, and the inf-player chooses a stopping time  $\sigma$  to minimize, the expected payoff  $M_x(\tau, \sigma) = \mathbb{E}_x[G_1(X_\tau)I(\tau < \sigma) + G_2(X_\sigma)I(\sigma < \tau) + G_3(X_\tau)I(\tau = \sigma)]$ , where  $X_0 = x$  under  $\mathbb{P}_x$ . Define the upper value and the lower value of the game by  $V^*(x) = \inf_\sigma \sup_\tau M_x(\tau, \sigma)$  and  $V_*(x) = \sup_\tau \inf_\sigma M_x(\tau, \sigma)$ , respectively, where the horizon  $T$  (the upper bound for  $\tau$  and  $\sigma$  above) may be either finite or infinite (it is assumed that  $G_1(X_T) = G_2(X_T)$  if  $T$  is finite and  $\liminf_{t \rightarrow \infty} G_2(X_t) \leq \limsup_{t \rightarrow \infty} G_1(X_t)$  if  $T$  is infinite). If  $X$  is right-continuous, then the *Stackelberg equilibrium* holds, in the sense that  $V^*(x) = V_*(x)$  for all  $x$  with  $V := V^* = V_*$  defining a measurable function. If  $X$  is right-continuous and left-continuous over stopping times (quasi-left-continuous), then the *Nash equilibrium* holds, in the sense that there exist stopping times  $\tau_*$  and  $\sigma_*$  such that  $M_x(\tau, \sigma_*) \leq M_x(\tau_*, \sigma_*) \leq M_x(\tau_*, \sigma)$  for all stopping times  $\tau$  and  $\sigma$ , implying also that  $V(x) = M_x(\tau_*, \sigma_*)$  for all  $x$ . Further properties of the value function  $V$  and the optimal stopping times  $\tau_*$  and  $\sigma_*$  are exhibited in the proof.

**Key words.** optimal stopping game, Stackelberg equilibrium, Nash equilibrium, saddle point, optimal stopping, Snell envelope, Markov process, martingale

**AMS subject classifications.** Primary, 91A15, 60G40; Secondary, 60J25, 60G44

**DOI.** 10.1137/060673916

**1. Introduction.** Let  $X = (X_t)_{t \geq 0}$  be a strong Markov process, and let  $G_1$ ,  $G_2$ , and  $G_3$  be continuous functions satisfying  $G_1 \leq G_3 \leq G_2$  (for further details see section 2 below). Consider the optimal stopping game where the sup-player chooses a stopping time  $\tau$  to maximize, and the inf-player chooses a stopping time  $\sigma$  to minimize, the expected payoff

$$(1.1) \quad M_x(\tau, \sigma) = \mathbb{E}_x[G_1(X_\tau)I(\tau < \sigma) + G_2(X_\sigma)I(\sigma < \tau) + G_3(X_\tau)I(\tau = \sigma)],$$

where  $X_0 = x$  under  $\mathbb{P}_x$ .

Define the upper value and the lower value of the game by

$$(1.2) \quad V^*(x) = \inf_\sigma \sup_\tau M_x(\tau, \sigma) \quad \text{and} \quad V_*(x) = \sup_\tau \inf_\sigma M_x(\tau, \sigma),$$

respectively, where the horizon  $T$  (the upper bound for  $\tau$  and  $\sigma$  above) may be either finite or infinite (it is assumed that  $G_1(X_T) = G_2(X_T)$  if  $T$  is finite and  $\liminf_{t \rightarrow \infty} G_2(X_t) \leq \limsup_{t \rightarrow \infty} G_1(X_t)$  if  $T$  is infinite). Note that  $V_*(x) \leq V^*(x)$  for all  $x$ .

In this context one distinguishes: (i) the *Stackelberg equilibrium*, meaning that

$$(1.3) \quad V^*(x) = V_*(x)$$

for all  $x$  (in this case  $V := V^* = V_*$  unambiguously defines the value of the game), and (ii) the *Nash equilibrium*, meaning that there exist stopping times  $\tau_*$  and  $\sigma_*$  such

\*Received by the editors November 1, 2006; accepted for publication (in revised form) June 6, 2007; published electronically February 15, 2008.

<http://www.siam.org/journals/sicon/47-2/67391.html>

<sup>†</sup>School of Mathematics, The University of Manchester, Oxford Road, Manchester M13 9PL, UK (ekstrom@maths.man.ac.uk, [www.maths.man.ac.uk/ekstrom](http://www.maths.man.ac.uk/ekstrom), goran@maths.man.ac.uk, [www.maths.man.ac.uk/goran](http://www.maths.man.ac.uk/goran)).

that

$$(1.4) \quad M_x(\tau, \sigma_*) \leq M_x(\tau_*, \sigma_*) \leq M_x(\tau_*, \sigma)$$

for all stopping times  $\tau$  and  $\sigma$  and for all  $x$  (in other words  $(\tau_*, \sigma_*)$  is a saddle point). It is easily seen that the Nash equilibrium implies the Stackelberg equilibrium with  $V(x) = M_x(\tau_*, \sigma_*)$  for all  $x$ .

A variant of the problem above was first studied by Dynkin [5] using martingale methods similar to those of Snell [21]. Specific examples of the same problem were studied in [9] and [12] using Markovian methods (see also [13] for martingale methods). In parallel to that, Bensoussan and Friedman (cf. [10], [2], [3]) developed an analytic approach (for diffusions) based on variational inequalities. Martingale methods were further advanced in [18] (see also [23]), and Markovian setting was studied in [8] (via Wald–Bellman equations) and [22] (via penalty equations). More recent papers on optimal stopping games include [14], [16], [1], [11], [6], [7], and [15]. These papers study specific problems and often lead to explicit solutions. For optimal stopping games with randomized stopping times, see [17] and the references therein. For connections with singular stochastic control (forward/backward SDE), see [4] and the references therein.

The most general martingale result known to date assumes an upper/lower semi-continuity from the left (cf. [18, Theorem 15, p. 42]) so that it does not cover the case of Lévy processes, for example. The most general Markovian result known to date assumes an asymptotic condition uniformly over initial points (cf. [22, Condition (A3), p. 2]) so that it is not always easily verifiable. The present paper aims at closing these gaps.

The main result of the paper (Theorem 2.1) may be summarized as follows. If  $X$  is *right-continuous*, then the *Stackelberg equilibrium* holds with a measurable value function. If  $X$  is right-continuous and *left-continuous over stopping times* (quasi-left-continuous), then the *Nash equilibrium* holds (see also Example 3.1 and Theorem 3.2). These two sufficient conditions are known to be most general in optimal stopping theory (see, e.g., [19] and [20]). Further properties of the value function  $V$  and the optimal stopping times  $\tau_*$  and  $\sigma_*$  are exhibited in the proof.

**2. Result and proof.** 1. Throughout we will consider a strong Markov process  $X = (X_t)_{t \geq 0}$  defined on a filtered probability space  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P}_x)$  and taking values in a measurable space  $(E, \mathcal{B})$ , where  $E$  is a locally compact Hausdorff space with a countable base and  $\mathcal{B}$  is the Borel  $\sigma$ -algebra on  $E$ . It will be assumed that the process  $X$  starts at  $x$  under  $\mathbb{P}_x$  for  $x \in E$  and that the sample paths of  $X$  are (first) right-continuous and (then) left-continuous over stopping times. The latter condition is often referred to as quasi-left-continuity and means that  $X_{\tau_n} \rightarrow X_\tau$   $\mathbb{P}_x$ -a.s. whenever  $\tau_n$  and  $\tau$  are stopping times such that  $\tau_n \uparrow \tau$  as  $n \rightarrow \infty$ . (Stopping times are always referred to with respect to the filtration  $(\mathcal{F}_t)_{t \geq 0}$  given above.) It is also assumed that the filtration  $(\mathcal{F}_t)_{t \geq 0}$  is right-continuous (implying that the first entry times to open and closed sets are stopping times) and that  $\mathcal{F}_0$  contains all  $\mathbb{P}_x$ -null sets from  $\mathcal{F}_\infty^X = \sigma(X_t : t \geq 0)$  (implying also that the first entry times to Borel sets are stopping times). The main example we have in mind is when  $\mathcal{F}_t = \sigma(\mathcal{F}_t^X \cup \mathcal{N})$ , where  $\mathcal{F}_t^X = \sigma(X_s : 0 \leq s \leq t)$  and  $\mathcal{N} = \{A \subseteq \Omega : \exists B \in \mathcal{F}_\infty^X, A \subseteq B, \mathbb{P}_x(B) = 0\}$  for  $t \geq 0$  with  $\mathcal{F} = \mathcal{F}_\infty$ . In addition, it is assumed that the mapping  $x \mapsto \mathbb{P}_x(F)$  is (universally) measurable for each  $F \in \mathcal{F}$ . It follows that the mapping  $x \mapsto \mathbb{E}_x(Z)$  is (universally) measurable for each (integrable) random variable  $Z$ . Finally, without loss of generality we will assume that  $\Omega$  equals the canonical space  $E^{[0, \infty)}$  with  $X_t(\omega) = \omega(t)$  for  $\omega \in \Omega$ .

and  $t \geq 0$ , so that the shift operator  $\theta_t : \Omega \rightarrow \Omega$  is well defined by  $\theta_t(\omega)(s) = \omega(t+s)$  for  $\omega \in \Omega$  and  $t, s \geq 0$ .

2. Given continuous functions  $G_1, G_2, G_3 : E \rightarrow \mathbb{R}$  satisfying  $G_1 \leq G_3 \leq G_2$  and the following integrability condition:

$$(2.1) \quad \mathbb{E}_x \sup_t |G_i(X_t)| < \infty \quad (i = 1, 2, 3)$$

for all  $x \in E$ , we consider the *optimal stopping game* where the sup-player chooses a stopping time  $\tau$  to maximize, and the inf-player chooses a stopping time  $\sigma$  to minimize, the expected payoff

$$(2.2) \quad \mathbb{M}_x(\tau, \sigma) = \mathbb{E}_x[G_1(X_\tau)I(\tau < \sigma) + G_2(X_\sigma)I(\sigma < \tau) + G_3(X_\tau)I(\tau = \sigma)],$$

where  $X_0 = x$  under  $\mathbb{P}_x$ .

Define the upper value and the lower value of the game by

$$(2.3) \quad V^*(x) = \inf_{\sigma} \sup_{\tau} \mathbb{M}_x(\tau, \sigma) \quad \text{and} \quad V_*(x) = \sup_{\tau} \inf_{\sigma} \mathbb{M}_x(\tau, \sigma),$$

respectively, where the horizon  $T$  (the upper bound for  $\tau$  and  $\sigma$  above) may be either finite or infinite. If  $T < \infty$ , then it will be assumed that  $G_1(X_T) = G_2(X_T) = G_3(X_T)$ . In this case it is most interesting to assume that  $X$  is a time-space process  $(t, Y_t)$  for  $t \in [0, T]$ , so that  $G_i = G_i(t, y)$  will be functions of both time and space for  $i = 1, 2, 3$ . If  $T = \infty$ , then it will be assumed that  $\liminf_{t \rightarrow \infty} G_2(X_t) \leq \limsup_{t \rightarrow \infty} G_1(X_t)$ , and the common value for  $G_3(X_\infty)$  could formally be assigned as either of the preceding two values (if  $\tau$  and  $\sigma$  are allowed to take the value  $\infty$ ) yielding the same results as in Theorem 2.1 below. For simplicity of the exposition, however, we will assume that  $\tau$  and  $\sigma$  in (2.2) are finite valued.

3. The main result of the paper may now be stated as follows.

**THEOREM 2.1.** *Consider the optimal stopping game (2.3). If  $X$  is right-continuous, then the Stackelberg equilibrium (1.3) holds with  $V := V^* = V_*$  defining a measurable function. If  $X$  is right-continuous and left-continuous over stopping times, then the Nash equilibrium (1.4) holds with*

$$(2.4) \quad \tau_* = \inf \{t : X_t \in D_1\} \quad \text{and} \quad \sigma_* = \inf \{t : X_t \in D_2\},$$

where  $D_1 = \{V = G_1\}$  and  $D_2 = \{V = G_2\}$ .

*Proof.* Both finite and infinite horizons can be treated by slight modifications of the same method which we will therefore present without referring to the horizon.

(I) In the first part of the proof we will assume that  $X$  is right-continuous, and we will show that this hypothesis implies the Stackelberg equilibrium with  $V := V^* = V_*$  defining a measurable function. This will be done in a number of steps as follows.

1. Given  $\varepsilon > 0$  set

$$(2.5) \quad D_1^\varepsilon = \{V^* \leq G_1 + \varepsilon\} \quad \text{and} \quad D_2^\varepsilon = \{V_* \geq G_2 - \varepsilon\},$$

and consider the stopping times

$$(2.6) \quad \tau_\varepsilon = \inf \{t : X_t \in D_1^\varepsilon\} \quad \text{and} \quad \sigma_\varepsilon = \inf \{t : X_t \in D_2^\varepsilon\}.$$

The key is to show that

$$(2.7) \quad \mathbb{M}_x(\tau, \sigma_\varepsilon) - \varepsilon \leq V_*(x) \leq V^*(x) \leq \mathbb{M}_x(\tau_\varepsilon, \sigma) + \varepsilon$$

for all  $\tau, \sigma, x$ , and  $\varepsilon > 0$ . Indeed, suppose that (2.7) is valid. Then

$$(2.8) \quad V^*(x) \leq \inf_{\sigma} M_x(\tau_{\varepsilon}, \sigma) + \varepsilon \leq \sup_{\tau} \inf_{\sigma} M_x(\tau, \sigma) + \varepsilon = V_*(x) + \varepsilon$$

for all  $\varepsilon > 0$ . Letting  $\varepsilon \downarrow 0$  we see that  $V^* = V_*$ , and the claim follows (up to measurability which will be derived below).

Since the first inequality in (2.7) is analogous to the third one, and since the second inequality holds generally, we focus on establishing the third one, which states that

$$(2.9) \quad V^*(x) \leq M_x(\tau_{\varepsilon}, \sigma) + \varepsilon$$

for all  $\sigma, x$ , and  $\varepsilon > 0$ .

2. To prove (2.9) take any stopping time  $\sigma$ , and consider the optimal stopping problem

$$(2.10) \quad \hat{V}_{\sigma}^*(x) = \sup_{\tau} \hat{M}_x(\tau, \sigma),$$

where we set

$$(2.11) \quad \hat{M}_x(\tau, \sigma) = E_x[G_1(X_{\tau})I(\tau < \sigma) + G_2(X_{\sigma})I(\sigma \leq \tau)].$$

Note that the gain process  $G^{\sigma}$  in (2.10) is given by

$$(2.12) \quad G_t^{\sigma} = G_1(X_t)I(t < \sigma) + G_2(X_{\sigma})I(\sigma \leq t),$$

from which we see that  $G^{\sigma}$  is right-continuous and adapted (satisfying also a sufficient integrability condition which can be derived using (2.1)). Thus general optimal stopping results of the martingale approach (cf. [19]) are applicable to the problem (2.10). In order to make use of these results in the Markovian setting of the present theorem (where  $P_x$  forms a family of probability measures when  $x$  runs through  $E$ ), we will first verify a regularity property of the value function  $\hat{V}_{\sigma}^*$ .

3. We show that the function  $x \mapsto \hat{V}_{\sigma}^*(x)$  defined in (2.10) is measurable. The basic idea of the proof is to embed the problem (2.10) into a setting of the Wald–Bellman equations (cf. [19]) and then exploit the underlying Markovian structure in this context.

For this, let us first assume that the stopping times  $\tau$  in (2.10) take values in a finite set, and without loss of generality let us assume that this set equals  $\{0, 1, \dots, N\}$ . Introduce the auxiliary optimal stopping problems

$$(2.13) \quad V_n^N(x) = \sup_{n \leq \tau \leq N} E_x G_{\tau}^{\sigma}$$

for  $n = N, \dots, 1, 0$ , and recall that the Wald–Bellman equations in this setting read:

$$(2.14) \quad S_n^N = G_N^{\sigma} \quad \text{for } n = N,$$

$$(2.15) \quad S_n^N = G_n^{\sigma} \vee E_x(S_{n+1}^N | \mathcal{F}_n) \quad \text{for } n = N-1, \dots, 1, 0,$$

with  $V_n^N(x) = E_x S_n^N$  for  $n = N, \dots, 1, 0$  (see, e.g., [19, pp. 3–6]). In particular, since  $V_0^N = \hat{V}_{\sigma}^*$  we see that

$$(2.16) \quad \hat{V}_{\sigma}^*(x) = E_x S_0^N$$

for all  $x$ . Thus the problem is reduced to showing that  $x \mapsto \mathbf{E}_x S_0^N$  is measurable.

If  $\sigma$  is a hitting time, then by the strong Markov property of  $X$  it follows using (2.14)–(2.15) inductively that the following identity holds:

$$(2.17) \quad S_0^N = G_1(x)I(0 < \sigma) + F_N(x) + G_2(x)I(\sigma = 0)$$

under  $\mathbf{P}_x$ , where  $x \mapsto F_N(x)$  is a measurable function obtained by means of the following recursive relations:

$$(2.18) \quad F_n(x) = \mathbf{E}_x[G_1(X_1)I(1 < \sigma) \vee F_{n-1}(x)] + \mathbf{E}_x[G_2(X_\sigma)I(0 < \sigma \leq 1)]$$

for  $n = 1, 2, \dots, N$ , where  $F_0 \equiv -\infty$ . Taking  $\mathbf{E}_x$  in (2.17) and using (2.16) we get

$$(2.19) \quad \hat{V}_\sigma^*(x) = G_1(x)\mathbf{P}_x(0 < \sigma) + F_N(x) + G_2(x)\mathbf{P}_x(\sigma = 0)$$

for all  $x$ . Hence we see that  $x \mapsto \hat{V}_\sigma^*(x)$  is measurable as claimed. In the case of a general stopping time  $\sigma$  one can make use of the extended Radon–Nikodym theorem, which states that  $(x, \omega) \mapsto \mathbf{E}_x(Z_x | \mathcal{G})(\omega)$  is measurable when  $(x, \omega) \mapsto Z_x(\omega)$  is measurable and  $\mathcal{G} \subseteq \mathcal{F}$  is a  $\sigma$ -algebra. Applying this fact inductively in (2.15) and using (2.16) it follows that  $x \mapsto \hat{V}_\sigma^*(x)$  is measurable as claimed. [Note that this argument also applies when  $\sigma$  is a hitting time; however, the explicit formula (2.19) is no longer available if  $\sigma$  is a general stopping time.]

Let us now consider the general case when the stopping times  $\tau$  from (2.10) can take arbitrary values. Setting  $\tau_n = k/2^n$  on  $\{(k-1)/2^n < \tau \leq k/2^n\}$  one knows that each  $\tau_n$  is a stopping time with values in the set  $Q_n$  of dyadic rationals of the form  $k/2^n$ , and  $\tau_n \downarrow \tau$  as  $n \rightarrow \infty$ . Hence by right-continuity of  $G^\sigma$  and Fatou's lemma (using a needed integrability condition which is derived by means of (2.1) above), one gets

$$(2.20) \quad \mathbf{E}_x G_\tau^\sigma = \mathbf{E}_x \left( \lim_{n \rightarrow \infty} G_{\tau_n}^\sigma \right) \leq \liminf_{n \rightarrow \infty} \mathbf{E}_x G_{\tau_n}^\sigma \leq \sup_{n \geq 1} V_n(x),$$

where we set

$$(2.21) \quad V_n(x) = \sup_{\tau \in Q_n} \mathbf{E}_x G_\tau^\sigma.$$

Taking the supremum in (2.20) over all  $\tau$ , and using that  $V_n \leq \hat{V}_\sigma^*$  for all  $n \geq 1$ , it follows that

$$(2.22) \quad \hat{V}_\sigma^*(x) = \sup_{n \geq 1} V_n(x)$$

for all  $x$ . By the first part of the proof above we know that each function  $x \mapsto V_n(x)$  is measurable, so it follows from (2.22) that  $x \mapsto \hat{V}_\sigma^*(x)$  is measurable as claimed.

4. Since the function  $x \mapsto \hat{V}_\sigma^*(x)$  is measurable, it follows that

$$(2.23) \quad \hat{V}_\sigma^*(X_\rho) = \sup_{\tau} \hat{\mathbf{M}}_{X_\rho}(\tau, \sigma)$$

defines a random variable for any stopping time  $\rho$  which is given and fixed. On the other hand, by the strong Markov property we have

$$\begin{aligned}
 (2.24) \quad \hat{M}_{X_\rho}(\tau, \sigma) &= E_{X_\rho} [G_1(X_\tau) I(\tau < \sigma) + G_2(X_\sigma) I(\sigma \leq \tau)] \\
 &= E_x [G_1(X_{\rho+\tau \circ \theta_\rho}) I(\rho + \tau \circ \theta_\rho < \rho + \sigma \circ \theta_\rho) \\
 &\quad + G_2(X_{\rho+\sigma \circ \theta_\rho}) I(\rho + \sigma \circ \theta_\rho \leq \rho + \tau \circ \theta_\rho) | \mathcal{F}_\rho].
 \end{aligned}$$

From (2.23) and (2.24) we see that

$$(2.25) \quad \hat{V}_\sigma^*(X_\rho) = \operatorname{ess\,sup}_\tau \hat{M}_x(\rho + \tau \circ \theta_\rho, \rho + \sigma \circ \theta_\rho | \mathcal{F}_\rho),$$

where we set

$$(2.26) \quad \hat{M}_x(\tau_\rho, \sigma_\rho | \mathcal{F}_\rho) = E_x [G_1(X_{\tau_\rho}) I(\tau_\rho < \sigma_\rho) + G_2(X_{\sigma_\rho}) I(\sigma_\rho \leq \tau_\rho) | \mathcal{F}_\rho],$$

with  $\tau_\rho = \rho + \tau \circ \theta_\rho$  and  $\sigma_\rho = \rho + \sigma \circ \theta_\rho$  (being stopping times).

5. By general optimal-stopping results of the martingale approach (cf. [19]), we know that the supermartingale

$$(2.27) \quad \hat{S}_t^\sigma = \operatorname{ess\,sup}_{\tau \geq t} \hat{M}_x(\tau, \sigma | \mathcal{F}_t)$$

admits a right-continuous modification (the Snell envelope) such that

$$(2.28) \quad \hat{V}_\sigma^*(x) = E_x \hat{S}_\rho^\sigma$$

for every stopping time  $\rho \leq \tau_\varepsilon^\sigma$ , where

$$(2.29) \quad \tau_\varepsilon^\sigma = \inf \{ t : \hat{S}_t^\sigma \leq G_t^\sigma + \varepsilon \}.$$

Moreover, by the well-known properties of the Snell envelope (stating that equality between the essential supremum and its right-continuous modification is preserved at stopping times and that the essential supremum is attained over hitting times), we see upon recalling (2.25) above that the following identity holds:

$$(2.30) \quad \hat{V}_\sigma^*(X_\rho) = \hat{S}_\rho^\sigma \quad \text{P}_x\text{-a.s.}$$

for every stopping time  $\rho \leq \sigma$ . The precise meaning of (2.30) is

$$(2.30') \quad \hat{V}_{\sigma^\rho(\omega, \cdot)}^*(X_\rho(\omega)) = \hat{S}_\rho^\sigma(\omega)$$

for  $\omega \in \Omega \setminus N$ , with  $P_x(N) = 0$ , where  $\sigma(\omega) = \rho(\omega) + \sigma^\rho(\omega, \theta_\rho(\omega))$  for a mapping  $(\omega, \omega') \mapsto \sigma^\rho(\omega, \omega')$  which is  $\mathcal{F}_\rho \otimes \mathcal{F}_\infty$ -measurable and  $\omega' \mapsto \sigma^\rho(\omega, \omega')$  is a stopping time for each  $\omega$  given and fixed. We will simplify the notation in what follows by dropping  $\rho$  and  $\omega$  from  $\sigma^\rho(\omega, \cdot)$  in (2.30') and simply writing  $\sigma$  instead. This also applies to the expression on the right-hand side of (2.23) above. [Note that, if  $\sigma$  is a hitting time, then  $\sigma^\rho(\omega, \omega') = \sigma(\omega')$  for all  $\omega$  and  $\omega'$ , and this simplification is exact.] In particular, using (2.30) with  $\rho = t$  and the fact that  $\hat{S}^\sigma$  and  $G^\sigma$  are right-continuous, we see that  $\tau_\varepsilon^\sigma$  from (2.29) can be equivalently defined as

$$(2.31) \quad \tau_\varepsilon^\sigma = \inf \{ t \in Q : \hat{V}_\sigma^*(X_t) \leq G_t^\sigma + \varepsilon \},$$

where  $Q$  is any (given and fixed) countable dense subset of the time set.

6. Setting

$$(2.32) \quad \hat{V}^*(x) = \inf_\sigma \sup_\tau \hat{M}_x(\tau, \sigma)$$

for  $x \in E$ , let us assume that the function  $x \mapsto \hat{V}^*(x)$  is measurable, and let us consider the stopping time  $\tau_\varepsilon$  from (2.6) above but defined over  $Q$  for  $\hat{V}^*$ , i.e.,

$$(2.33) \quad \tilde{\tau}_\varepsilon = \inf \{ t \in Q : X_t \in \hat{D}_1^\varepsilon \},$$

where  $\hat{D}_1^\varepsilon = \{ \hat{V}^* \leq G_1 + \varepsilon \}$ . Let a stopping time  $\beta$  be given and fixed, and let  $\sigma$  be any stopping time satisfying  $\sigma \geq \beta \wedge \tilde{\tau}_\varepsilon$ . Then, for any  $t \in Q$  such that  $t < \beta \wedge \tilde{\tau}_\varepsilon$ , we have

$$(2.34) \quad \begin{aligned} \hat{V}_\sigma^*(X_t) &\geq \hat{V}^*(X_t) > G_1(X_t) + \varepsilon \\ &= G_1(X_t) I(t < \sigma) + G_2(X_\sigma) I(\sigma \leq t) + \varepsilon \\ &= G_t^\sigma + \varepsilon, \end{aligned}$$

since  $t < \sigma$  so that  $I(\sigma \leq t) = 0$ . Hence we see by (2.31) that  $\beta \wedge \tilde{\tau}_\varepsilon \leq \tau_\varepsilon^\sigma$ . By (2.28) and (2.30) we can conclude that

$$(2.35) \quad \hat{V}_\sigma^*(x) = \mathbf{E}_x \hat{V}_\sigma^*(X_{\beta \wedge \tilde{\tau}_\varepsilon})$$

for any  $\sigma \geq \beta \wedge \tilde{\tau}_\varepsilon$ . Taking the infimum over all such  $\sigma$  we obtain

$$(2.36) \quad V^*(x) \leq \hat{V}^*(x) \leq \inf_{\sigma \geq \beta \wedge \tilde{\tau}_\varepsilon} \hat{V}_\sigma^*(x) = \inf_{\sigma \geq \beta \wedge \tilde{\tau}_\varepsilon} \mathbf{E}_x \hat{V}_\sigma^*(X_{\beta \wedge \tilde{\tau}_\varepsilon})$$

for every stopping time  $\beta$ . In the next step we will show that the infimum and the expectation in (2.36) can be interchanged.

7. We show that the family of random variables

$$(2.37) \quad \left\{ \sup_\tau \hat{M}_{X_\rho}(\tau, \sigma) : \sigma \text{ is a stopping time} \right\}$$

is downwards directed. Recall that a family of random variables  $\{Z_\sigma : \sigma \in I\}$  is downwards directed if for all  $\sigma_1, \sigma_2 \in I$  there exists  $\sigma_3 \in I$  such that  $Z_{\sigma_3} \leq Z_{\sigma_1} \wedge Z_{\sigma_2}$   $\mathbf{P}_x$ -a.s. for all  $x$ .

To prove the claim, recall that by the strong Markov property we have

$$(2.38) \quad \hat{M}_{X_\rho}(\tau, \sigma) = \hat{M}_x(\rho + \tau \circ \theta_\rho, \rho + \sigma \circ \theta_\rho | \mathcal{F}_\rho).$$

If  $\sigma_1$  and  $\sigma_2$  are two stopping times given and fixed, set  $\tau_\rho = \rho + \tau \circ \theta_\rho$ ,  $\sigma'_1 = \rho + \sigma_1 \circ \theta_\rho$ , and  $\sigma'_2 = \rho + \sigma_2 \circ \theta_\rho$ , and define

$$(2.39) \quad B = \left\{ \sup_\tau \hat{M}_x(\tau_\rho, \sigma'_1 | \mathcal{F}_\rho) \leq \sup_\tau \hat{M}_x(\tau_\rho, \sigma'_2 | \mathcal{F}_\rho) \right\}.$$

Then  $B \in \mathcal{F}_\rho$ , and the random variable

$$(2.40) \quad \sigma' := \sigma'_1 I_B + \sigma'_2 I_{B^c}$$

is a stopping time. For this, note that  $\{\sigma' \leq t\} = (\{\sigma'_1 \leq t\} \cap B) \cup (\{\sigma'_2 \leq t\} \cap B^c) = (\{\sigma'_1 \leq t\} \cap B \cap \{\rho \leq t\}) \cup (\{\sigma'_2 \leq t\} \cap B^c \cap \{\rho \leq t\}) \in \mathcal{F}_t$ , since  $B$  and  $B^c$  belong to  $\mathcal{F}_\rho$ , which verifies the claim.

Moreover, the stopping time  $\sigma'$  can be written as

$$(2.41) \quad \sigma' = \rho + \sigma \circ \theta_\rho$$



for some stopping time  $\sigma$ . Indeed, setting

$$(2.42) \quad A = \left\{ \sup_{\tau} \hat{M}_{X_0}(\tau, \sigma_1) \leq \sup_{\tau} \hat{M}_{X_0}(\tau, \sigma_2) \right\},$$

we see that  $A \in \mathcal{F}_0$  and  $B = \theta_{\rho}^{-1}(A)$  upon recalling (2.38). Hence from (2.40) we get

$$(2.43) \quad \begin{aligned} \sigma' &= (\rho + \sigma_1 \circ \theta_{\rho}) I_B + (\rho + \sigma_2 \circ \theta_{\rho}) I_{B^c} \\ &= \rho + [(\sigma_1 \circ \theta_{\rho})(I_A \circ \theta_{\rho}) + (\sigma_2 \circ \theta_{\rho})(I_{A^c} \circ \theta_{\rho})] \\ &= \rho + (\sigma_1 I_A + \sigma_2 I_{A^c}) \circ \theta_{\rho}, \end{aligned}$$

which implies that (2.41) holds with the stopping time  $\sigma = \sigma_1 I_A + \sigma_2 I_{A^c}$ . (The latter is a stopping time since  $\{\sigma \leq t\} = (\{\sigma_1 \leq t\} \cap A) \cup (\{\sigma_2 \leq t\} \cap A^c) \in \mathcal{F}_t$  due to the fact that  $A \in \mathcal{F}_0 \subseteq \mathcal{F}_t$  for all  $t$ .)

Finally, we have

$$(2.44) \quad \begin{aligned} \sup_{\tau} \hat{M}_{X_{\rho}}(\tau, \sigma) &= \sup_{\tau} \hat{M}_x(\tau_{\rho}, \sigma'_1 | \mathcal{F}_{\rho}) I_B + \sup_{\tau} \hat{M}_x(\tau_{\rho}, \sigma'_2 | \mathcal{F}_{\rho}) I_{B^c} \\ &= \sup_{\tau} \hat{M}_x(\tau_{\rho}, \sigma'_1 | \mathcal{F}_{\rho}) \wedge \sup_{\tau} \hat{M}_x(\tau_{\rho}, \sigma'_2 | \mathcal{F}_{\rho}) \\ &= \sup_{\tau} \hat{M}_{X_{\rho}}(\tau, \sigma_1) \wedge \sup_{\tau} \hat{M}_{X_{\rho}}(\tau, \sigma_2), \end{aligned}$$

which proves that the family (2.37) is downwards directed as claimed.

8. It is well-known (see, e.g., [19, pp. 6–7]) that, if a family  $\{Z_{\sigma} : \sigma \in I\}$  of random variables is downwards directed, there exists a countable subset  $J = \{\sigma_n : n \geq 1\}$  of  $I$  such that

$$(2.45) \quad \operatorname{ess\,inf}_{\sigma \in I} Z_{\sigma} = \lim_{n \rightarrow \infty} Z_{\sigma_n} \quad \mathbb{P}_x\text{-a.s.},$$

where  $Z_{\sigma_1} \geq Z_{\sigma_2} \geq \dots$   $\mathbb{P}_x$ -a.s. In particular, if there exists a random variable  $Z$  such that  $\mathbb{E}_x Z < \infty$  and  $Z_{\sigma} \leq Z$  for all  $\sigma \in I$ , then

$$(2.46) \quad \mathbb{E}_x \operatorname{ess\,inf}_{\sigma \in I} Z_{\sigma} = \lim_{n \rightarrow \infty} \mathbb{E}_x Z_{\sigma_n} = \inf_{\sigma \in I} \mathbb{E}_x Z_{\sigma};$$

i.e., the order of the infimum and the expectation can be interchanged.

Applying the preceding general fact to the family in (2.37) upon returning to (2.36) we can conclude that

$$(2.47) \quad V^*(x) \leq \hat{V}^*(x) \leq \inf_{\sigma \geq \beta \wedge \tilde{\tau}_{\varepsilon}} \mathbb{E}_x \hat{V}_{\sigma}^*(X_{\beta \wedge \tilde{\tau}_{\varepsilon}}) = \mathbb{E}_x \hat{V}^*(X_{\beta \wedge \tilde{\tau}_{\varepsilon}}).$$

In the next step we will relate the process  $\hat{V}^*(X_{\tilde{\tau}_{\varepsilon}})$  to yet another right-continuous modification which will play a useful role in what follows.

9. We show that the process

$$(2.48) \quad \hat{S}_{t \wedge \tilde{\tau}_{\varepsilon}} = \operatorname{ess\,inf}_{\sigma \geq t \wedge \tilde{\tau}_{\varepsilon}} \hat{S}_{\sigma}^{\sigma}$$

admits a right-continuous modification. For this, simplify the notation by setting

$$(2.49) \quad \hat{S}_t^{\varepsilon} = \hat{S}_{t \wedge \tilde{\tau}_{\varepsilon}} \quad \text{and} \quad M_t^{\sigma} = \hat{S}_{t \wedge \tilde{\tau}_{\varepsilon}}^{\sigma},$$

and note that the (stopped) process  $M^{\sigma}$  is a martingale. Indeed, recalling the conclusion in relation to (2.34) above that if  $\sigma \geq t \wedge \tilde{\tau}_{\varepsilon}$ , then  $t \wedge \tilde{\tau}_{\varepsilon} \leq \tau_{\varepsilon}^{\sigma}$ , we see that

the martingale property follows by (2.28) above (this is a well-known property of the Snell envelope).

Moreover, since by (2.30) we have

$$(2.50) \quad \hat{S}_{t \wedge \tilde{\tau}_\varepsilon}^\sigma = \sup_\tau \hat{M}_{X_{t \wedge \tilde{\tau}_\varepsilon}}(\tau, \sigma)$$

when  $\sigma \geq t \wedge \tilde{\tau}_\varepsilon$ , it follows by (2.37) and (2.48) that there exists a sequence of stopping times  $\{\sigma_n : n \geq 1\}$  satisfying  $\sigma_n \geq t \wedge \tilde{\tau}_\varepsilon$  such that

$$(2.51) \quad \hat{S}_t^\varepsilon = \lim_{n \rightarrow \infty} M_t^{\sigma_n} \quad \mathbb{P}_x\text{-a.s.},$$

where  $M_t^{\sigma_1} \geq M_t^{\sigma_2} \geq \dots$   $\mathbb{P}_x$ -a.s. Hence by the conditional monotone convergence theorem (using the integrability condition (2.1) above), we find for  $s < t$  that

$$(2.52) \quad \mathbb{E}_x(\hat{S}_t^\varepsilon | \mathcal{F}_s) = \lim_{n \rightarrow \infty} \mathbb{E}_x(M_t^{\sigma_n} | \mathcal{F}_s) = \lim_{n \rightarrow \infty} M_s^{\sigma_n} \geq \hat{S}_s^\varepsilon,$$

where the martingale property of  $M^{\sigma_n}$  and the definition of  $\hat{S}_s^\varepsilon$  are used. This shows that  $\hat{S}^\varepsilon$  is a submartingale.

A well-known result in martingale theory states that the submartingale  $\hat{S}^\varepsilon$  admits a right-continuous modification if and only if

$$(2.53) \quad t \mapsto \mathbb{E}_x \hat{S}_t^\varepsilon \text{ is right-continuous.}$$

To verify (2.53) note that by the submartingale property of  $\hat{S}^\varepsilon$  we have  $\mathbb{E}_x \hat{S}_t^\varepsilon \leq \dots \leq \mathbb{E}_x \hat{S}_{t_2}^\varepsilon \leq \mathbb{E}_x \hat{S}_{t_1}^\varepsilon$ , so that  $L := \lim_{n \rightarrow \infty} \mathbb{E}_x \hat{S}_{t_n}^\varepsilon$  exists and  $\mathbb{E}_x \hat{S}_t^\varepsilon \leq L$  whenever  $t_n \downarrow t$  as  $n \rightarrow \infty$  is given and fixed. To prove the reverse inequality, fix  $N \geq 1$ , and by means of (2.51) and the monotone convergence theorem choose  $\sigma \geq t \wedge \tilde{\tau}_\varepsilon$  such that

$$(2.54) \quad \mathbb{E}_x M_t^\sigma \leq \mathbb{E}_x \hat{S}_t^\varepsilon + 1/N.$$

Fix  $\delta > 0$ , and note that there is no restriction to assume that  $t_n \in [t, t + \delta]$  for all  $n \geq 1$ . Define a stopping time  $\sigma_n$  by setting

$$(2.55) \quad \sigma_n = \begin{cases} \sigma & \text{if } \sigma > t_n \wedge \tilde{\tau}_\varepsilon, \\ t \wedge \tilde{\tau}_\varepsilon + \delta & \text{if } \sigma \leq t_n \wedge \tilde{\tau}_\varepsilon \end{cases}$$

for  $n \geq 1$ . Then for all  $n \geq 1$  we have

$$(2.56) \quad \mathbb{E}_x M_t^{\sigma_n} = \mathbb{E}_x M_{t_n}^{\sigma_n} \geq \mathbb{E}_x \hat{S}_{t_n}^\varepsilon$$

by the martingale property of  $M^{\sigma_n}$  and the definition of  $\hat{S}_{t_n}^\varepsilon$  using that  $\sigma_n \geq t_n \wedge \tilde{\tau}_\varepsilon \geq t \wedge \tilde{\tau}_\varepsilon$ .

Since  $\{\sigma > t_n \wedge \tilde{\tau}_\varepsilon\}$  and  $\{\sigma \leq t_n \wedge \tilde{\tau}_\varepsilon\}$  belong to  $\mathcal{F}_{t_n \wedge \tilde{\tau}_\varepsilon}$ , it is easily verified using (2.30) above that  $M_{t_n}^{\sigma_n} I(\sigma > t_n \wedge \tilde{\tau}_\varepsilon) = M_{t_n}^\sigma I(\sigma > t_n \wedge \tilde{\tau}_\varepsilon)$  and  $M_{t_n}^{\sigma_n} I(\sigma \leq t_n \wedge \tilde{\tau}_\varepsilon) = M_{t_n}^{t \wedge \tilde{\tau}_\varepsilon + \delta} I(\sigma \leq t_n \wedge \tilde{\tau}_\varepsilon)$  for all  $n \geq 1$ . Hence

$$(2.57) \quad \mathbb{E}_x M_{t_n}^{\sigma_n} = \mathbb{E}_x [M_{t_n}^\sigma I(\sigma > t_n \wedge \tilde{\tau}_\varepsilon) + M_{t_n}^{t \wedge \tilde{\tau}_\varepsilon + \delta} I(\sigma \leq t_n \wedge \tilde{\tau}_\varepsilon)]$$

for all  $n \geq 1$ . Letting  $n \rightarrow \infty$  in (2.56) and using (2.57) we get

$$(2.58) \quad \mathbb{E}_x [M_t^\sigma I(\sigma > t \wedge \tilde{\tau}_\varepsilon) + M_t^{t \wedge \tilde{\tau}_\varepsilon + \delta} I(\sigma \leq t \wedge \tilde{\tau}_\varepsilon)] \geq L$$

for all  $\delta > 0$ .

By (2.30) (recall also (2.30')) we have

$$\begin{aligned}
 (2.59) \quad M_t^{t \wedge \tilde{\tau}_\varepsilon + \delta} &= \hat{S}_{t \wedge \tilde{\tau}_\varepsilon}^{t \wedge \tilde{\tau}_\varepsilon + \delta} = \hat{V}_{t \wedge \tilde{\tau}_\varepsilon + \delta}^*(X_{t \wedge \tilde{\tau}_\varepsilon}) = \sup_{\tau} M_{X_{t \wedge \tilde{\tau}_\varepsilon}}(\tau, \delta) \\
 &= \sup_{\tau} \mathbb{E}_{X_{t \wedge \tilde{\tau}_\varepsilon}} [G_1(X_\tau) I(\tau < \delta) + G_2(X_\delta) I(\delta \leq \tau)] \\
 &\leq \sup_{\tau} \mathbb{E}_{X_{t \wedge \tilde{\tau}_\varepsilon}} [G_2(X_{\tau \wedge \delta})] \rightarrow G_2(X_{t \wedge \tilde{\tau}_\varepsilon}) = M_t^{t \wedge \tilde{\tau}_\varepsilon},
 \end{aligned}$$

where the convergence relation follows by

$$(2.60) \quad \left| \sup_{\tau} \mathbb{E}_x G_2(X_{\tau \wedge \delta}) - G_2(x) \right| \leq \mathbb{E}_x \sup_{0 \leq t \leq \delta} |G_2(X_t) - G_2(x)| \rightarrow 0$$

as  $\delta \downarrow 0$  upon using that  $X$  is right-continuous (at zero) and that the integrability condition (2.1) holds. Inserting (2.59) in (2.58) and using that  $\sigma \geq t \wedge \tilde{\tau}_\varepsilon$ , it follows that

$$(2.61) \quad \mathbb{E}_x M_t^\sigma \geq L.$$

Combining this with (2.54) we see that  $L \leq \mathbb{E}_x \hat{S}_t^\varepsilon$  and thus  $L = \mathbb{E}_x \hat{S}_t^\varepsilon$ . This establishes (2.53), and hence  $\hat{S}^\varepsilon$  admits a right-continuous modification (denoted by the same symbol) as claimed.

Moreover, from (2.48) and (2.50) upon using (2.37) it is easily verified that equality between the process in (2.48) and its right-continuous modification extends from deterministic times to all stopping times (via discrete stopping times upon using that each stopping time is the limit of a decreasing sequence of discrete stopping times). Hence by (2.30)+(2.32) and (2.48)+(2.49) we find that

$$(2.62) \quad \hat{V}^*(X_{\beta \wedge \tilde{\tau}_\varepsilon}) = \hat{S}_\beta^\varepsilon \quad \text{P}_x\text{-a.s.}$$

for every stopping time  $\beta$ .

10. We claim that

$$(2.63) \quad \hat{S}_{\tilde{\tau}_\varepsilon} \leq G_1(X_{\tilde{\tau}_\varepsilon}) + \varepsilon \quad \text{P}_x\text{-a.s.}$$

To verify this note first that  $\tilde{\tau}_{\varepsilon_2} \leq \tilde{\tau}_{\varepsilon_1}$  for  $\varepsilon_1 < \varepsilon_2$ , so that the right-continuous modification of (2.48) extends by letting  $\varepsilon \downarrow 0$  to become a right-continuous modification of the process

$$(2.64) \quad \hat{S}_{t \wedge \tilde{\tau}_{0-}} = \operatorname{ess\,inf}_{\sigma \geq t \wedge \tilde{\tau}_{0-}} \hat{S}_\sigma^\sigma,$$

where  $\tilde{\tau}_{0-} = \lim_{\varepsilon \downarrow 0} \tilde{\tau}_\varepsilon$  is a stopping time. But then by right-continuity of  $\hat{S}$  and  $G_1(X)$  on  $[0, \tau_{0-})$  it follows that on  $\{\tilde{\tau}_\varepsilon < \tilde{\tau}_{0-}\}$  we have the inequality (2.63) satisfied. Note that  $\tilde{\tau}_{0-} \leq \tilde{\tau}_0$ , where  $\tilde{\tau}_0$  is defined as in (2.33) with  $\varepsilon = 0$ .

To see what happens on  $\{\tilde{\tau}_\varepsilon = \tilde{\tau}_{0-}\}$ , let us consider the process

$$(2.65) \quad \hat{S}_t = \operatorname{ess\,inf}_{\sigma \geq t} \hat{S}_t^\sigma.$$

We claim that, if  $\rho_n$  and  $\rho$  are stopping times such that  $\rho_n \downarrow \rho$  as  $n \rightarrow \infty$ , then

$$(2.66) \quad \mathbb{E}_x \hat{S}_\rho \leq \liminf_{n \rightarrow \infty} \mathbb{E}_x \hat{S}_{\rho_n}.$$

Indeed, for this note first (since the families are downwards and upwards directed) that

$$(2.67) \quad \mathbb{E}_x \hat{S}_\rho = \inf_{\sigma \geq \rho} \sup_{\tau \geq \rho} \hat{M}_x(\tau, \sigma) \leq \inf_{\sigma > \rho_n} \sup_{\tau \geq \rho} \hat{M}_x(\tau, \sigma).$$

Taking  $\sigma > \rho_n$  we find that

$$(2.68) \quad \begin{aligned} \hat{M}_x(\tau, \sigma) &= \mathbb{E}_x [(G_1(X_\tau) I(\tau < \sigma) + G_2(X_\sigma) I(\sigma \leq \tau)) I(\tau < \rho_n)] \\ &\quad + \mathbb{E}_x [(G_1(X_{\tau \vee \rho_n}) I(\tau \vee \rho_n < \sigma) + G_2(X_\sigma) I(\sigma \leq \tau \vee \rho_n)) I(\tau \geq \rho_n)] \\ &= \mathbb{E}_x [G_1(X_\tau) I(\tau < \rho_n)] \\ &\quad + \mathbb{E}_x [G_1(X_{\tau \vee \rho_n}) I(\tau \vee \rho_n < \sigma) + G_2(X_\sigma) I(\sigma \leq \tau \vee \rho_n)] \\ &\quad - \mathbb{E}_x [(G_1(X_{\tau \vee \rho_n}) I(\tau \vee \rho_n < \sigma) + G_2(X_\sigma) I(\sigma \leq \tau \vee \rho_n)) I(\tau < \rho_n)] \\ &= \mathbb{E}_x [G_1(X_\tau) I(\tau < \rho_n) - G_1(X_{\rho_n}) I(\tau < \rho_n)] \\ &\quad + \mathbb{E}_x [G_1(X_{\tau \vee \rho_n}) I(\tau \vee \rho_n < \sigma) + G_2(X_\sigma) I(\sigma \leq \tau \vee \rho_n)] \\ &= \mathbb{E}_x [G_1(X_{\tau \wedge \rho_n}) - G_1(X_{\rho_n})] + \hat{M}_x(\tau \vee \rho_n, \sigma). \end{aligned}$$

From (2.67) and (2.68) we get

$$(2.69) \quad \begin{aligned} \mathbb{E}_x \hat{S}_\rho &\leq \mathbb{E}_x \sup_{\rho \leq t \leq \rho_n} |G_1(X_t) - G_1(X_{\rho_n})| + \inf_{\sigma > \rho_n} \sup_{\tau \geq \rho_n} \hat{M}_x(\tau, \sigma) \\ &= \mathbb{E}_x \sup_{\rho \leq t \leq \rho_n} |G_1(X_t) - G_1(X_{\rho_n})| + \inf_{\sigma \geq \rho_n} \sup_{\tau \geq \rho_n} \hat{M}_x(\tau, \sigma) \\ &= \mathbb{E}_x \sup_{\rho \leq t \leq \rho_n} |G_1(X_t) - G_1(X_{\rho_n})| + \mathbb{E}_x \hat{S}_{\rho_n}, \end{aligned}$$

where the first equality can easily be justified by using that each  $\sigma$  is the limit of a strictly decreasing sequence of discrete stopping times  $\sigma_m$  as  $m \rightarrow \infty$  yielding

$$(2.70) \quad \sup_{\tau \geq \rho_n} \hat{M}_x(\tau, \sigma) \geq \limsup_{m \rightarrow \infty} \sup_{\tau \geq \rho_n} \hat{M}_x(\tau, \sigma_m),$$

which is obtained directly from (2.93) below. Letting  $n \rightarrow \infty$  in (2.69) and using that the second-last expectation tends to zero since  $G_1(X)$  is right-continuous and the integrability condition (2.1) holds, we get (2.66) as claimed.

Returning to the question of  $\{\tilde{\tau}_\varepsilon = \tilde{\tau}_0 -\}$ , consider the Borel set  $\hat{D}_1^0 = \{\hat{V}^* = G_1\}$ , and choose compact sets  $K_1 \subseteq K_2 \subseteq \dots \subseteq \hat{D}_1^0$  such that  $\tau_n := \inf\{t : X_t \in K_n\}$  satisfy  $\tau_n \downarrow \tilde{\tau}_0$   $\mathbb{P}_x$ -a.s. as  $n \rightarrow \infty$ . (The latter is a well-known consequence of the fact that each probability measure on  $E$  is tight.) Since each  $K_n$  is closed, we have  $\hat{S}_{\tau_n} = \hat{V}^*(X_{\tau_n}) = G_1(X_{\tau_n})$  by right-continuity of  $X$  for all  $n \geq 1$ . Hence by (2.66) we find

$$(2.71) \quad \mathbb{E}_x \hat{S}_{\tilde{\tau}_0} \leq \liminf_{n \rightarrow \infty} \mathbb{E}_x \hat{S}_{\tau_n} = \liminf_{n \rightarrow \infty} \mathbb{E}_x G_1(X_{\tau_n}) = \mathbb{E}_x G_1(X_{\tilde{\tau}_0})$$

by right-continuity of  $G_1(X)$  using also the integrability condition (2.1) above. Since  $\hat{S}_{\tilde{\tau}_0} \geq G_1(X_{\tilde{\tau}_0})$   $\mathbb{P}_x$ -a.s. by definition, we see from (2.71) that  $\hat{S}_{\tilde{\tau}_0} = G_1(X_{\tilde{\tau}_0})$   $\mathbb{P}_x$ -a.s. Moreover, if we consider the Borel set  $\hat{D}_1^\varepsilon = \{\hat{V}^* \leq G_1 + \varepsilon\}$  and likewise choose stopping times  $\tau_n^\varepsilon$  satisfying  $\tau_n^\varepsilon \downarrow \tilde{\tau}_\varepsilon$   $\mathbb{P}_x$ -a.s., then the same arguments as in (2.71) show that

$$\begin{aligned}
(2.72) \quad \mathbb{E}_x G_1(X_{\tilde{\tau}_0-}) &= \mathbb{E}_x \lim_{\varepsilon \downarrow 0} G_1(X_{\tilde{\tau}_\varepsilon}) \leq \mathbb{E}_x \liminf_{\varepsilon \downarrow 0} \hat{S}_{\tilde{\tau}_\varepsilon} \leq \liminf_{\varepsilon \downarrow 0} \mathbb{E}_x \hat{S}_{\tilde{\tau}_\varepsilon} \\
&\leq \liminf_{\varepsilon \downarrow 0} \left( \liminf_{n \rightarrow \infty} \mathbb{E}_x \hat{S}_{\tau_n^\varepsilon} \right) \leq \liminf_{\varepsilon \downarrow 0} \left( \liminf_{n \rightarrow \infty} \mathbb{E}_x G_1(X_{\tau_n^\varepsilon}) + \varepsilon \right) \\
&= \mathbb{E}_x G_1(X_{\tilde{\tau}_0-})
\end{aligned}$$

upon using that  $G_1(X_{\tilde{\tau}_\varepsilon}) \leq \hat{S}_{\tilde{\tau}_\varepsilon}$   $\mathbb{P}_x$ -a.s. and applying Fatou's lemma. Hence all of the inequalities in (2.72) are equalities, and thus

$$(2.73) \quad G_1(X_{\tilde{\tau}_0-}) = \liminf_{\varepsilon \downarrow 0} \hat{S}_{\tilde{\tau}_\varepsilon} \quad \mathbb{P}_x\text{-a.s.}$$

Since  $\tilde{\tau}_\varepsilon \uparrow \tilde{\tau}_0-$  as  $\varepsilon \downarrow 0$ , we see from (2.73) that  $G_1(X_{\tilde{\tau}_0-}) = \hat{S}_{\tilde{\tau}_0-}$   $\mathbb{P}_x$ -a.s. on  $\{\tilde{\tau}_\varepsilon = \tilde{\tau}_0-\}$ . This implies that  $\tilde{\tau}_0 \leq \tilde{\tau}_0-$  and thus  $\tilde{\tau}_0 = \tilde{\tau}_0-$  both  $\mathbb{P}_x$ -a.s. on  $\{\tilde{\tau}_\varepsilon = \tilde{\tau}_0-\}$ . Recalling also that  $\hat{S}_{\tilde{\tau}_0} = G_1(X_{\tilde{\tau}_0})$   $\mathbb{P}_x$ -a.s. we finally see that on  $\{\tilde{\tau}_\varepsilon = \tilde{\tau}_0-\}$  one has  $\hat{S}_{\tilde{\tau}_\varepsilon} = \hat{S}_{\tilde{\tau}_0} = G_1(X_{\tilde{\tau}_0}) = G_1(X_{\tilde{\tau}_\varepsilon}) \leq G_1(X_{\tilde{\tau}_\varepsilon}) + \varepsilon$   $\mathbb{P}_x$ -a.s. so that (2.63) holds as claimed. [Note that (2.63) can also be obtained by showing that  $\hat{S}$  defined in (2.65) admits a right-continuous modification. This proof can be used instead of parts 9 and 10 above, which focused on exploiting the submartingale characterization (2.53) above.]

11. Inserting (2.62) into (2.47) and using (2.63) we get

$$\begin{aligned}
(2.74) \quad V^*(x) &\leq \hat{V}^*(x) \leq \mathbb{E}_x \hat{S}_\beta^\varepsilon = \mathbb{E}_x [\hat{S}_{\tilde{\tau}_\varepsilon} I(\tilde{\tau}_\varepsilon \leq \beta) + \hat{S}_\beta I(\beta < \tilde{\tau}_\varepsilon)] \\
&\leq \mathbb{E}_x [(G_1(X_{\tilde{\tau}_\varepsilon}) + \varepsilon) I(\tilde{\tau}_\varepsilon < \beta) + G_2(X_\beta) I(\beta < \tilde{\tau}_\varepsilon) \\
&\quad + (G_3(X_{\tilde{\tau}_\varepsilon}) + \varepsilon) I(\tilde{\tau}_\varepsilon = \beta)] \\
&\leq M_x(\tilde{\tau}_\varepsilon, \beta) + \varepsilon
\end{aligned}$$

for every stopping time  $\beta$ . Proceeding as in (2.8) above we find that  $\hat{V}^* = V^* = V_*$ , and thus (2.63) yields (2.9) with  $\tilde{\tau}_\varepsilon$  in place of  $\tau_\varepsilon$ .

12. To derive (2.9) with  $\tau_\varepsilon$  from (2.6), first note that  $\tau_\varepsilon \leq \tilde{\tau}_\varepsilon$ , and recall from (2.47) that

$$(2.75) \quad V^*(x) \leq \mathbb{E}_x V^*(X_{\beta \wedge \tilde{\tau}_\varepsilon})$$

for every stopping time  $\beta$ . From general theory of Markov processes (upon using that  $t \mapsto X_{t \wedge \tilde{\tau}_\varepsilon}$  is right-continuous and adapted) it is known that (2.75) implies that  $V^*$  is finely lower-semicontinuous up to  $\tilde{\tau}_\varepsilon$  in the sense that

$$(2.76) \quad V^*(x) \leq \liminf_{t \downarrow 0} V^*(X_{t \wedge \tilde{\tau}_\varepsilon}) \quad \mathbb{P}_x\text{-a.s.}$$

This in particular implies (since  $X^{\tilde{\tau}_\varepsilon}$  is a strong Markov process) that

$$(2.77) \quad V^*(X_\tau) \leq \liminf_{t \downarrow 0} V^*(X_{\tau+t}) \quad \mathbb{P}_x\text{-a.s. on } \{\tau < \tilde{\tau}_\varepsilon\}$$

for every stopping time  $\tau$ . Indeed, setting  $Y_t = X_{t \wedge \tilde{\tau}_\varepsilon}$ ,

$$(2.78) \quad A = \{V^*(x) \leq \liminf_{t \downarrow 0} V^*(Y_t)\} \quad \text{and} \quad B = \{V^*(Y_\tau) \leq \liminf_{t \downarrow 0} V^*(Y_{\tau+t})\}$$

we see that  $B = \theta_\tau^{-1}(A)$  and  $B^c = \theta_\tau^{-1}(A^c)$  so that the strong Markov property of  $Y$  implies

$$(2.79) \quad \begin{aligned} P_x(B^c) &= P_x(\theta_\tau^{-1}(A^c)) = E_x[E_x(I_{A^c} \circ \theta_\tau | \mathcal{F}_\tau)] \\ &= E_x E_{X_\tau}(I_{A^c}) = E_x P_{X_\tau}(A^c) = 0, \end{aligned}$$

since  $P_y(A^c) = 0$  for all  $y$ . Hence (2.77) holds as claimed. In particular, if (2.77) is applied to  $\tau_\varepsilon$ , we get

$$(2.80) \quad V^*(X_{\tau_\varepsilon}) \leq G_1(X_{\tau_\varepsilon}) + \varepsilon \quad P_x\text{-a.s. on } \{\tau_\varepsilon < \tilde{\tau}_\varepsilon\}.$$

With this new information we can now revisit (2.74) via (2.75) upon using (2.63) and (2.80). This gives

$$(2.81) \quad \begin{aligned} V^*(x) &\leq E_x V^*(X_{\beta \wedge \tau_\varepsilon}) = E_x[V^*(X_{\tau_\varepsilon})I(\tau_\varepsilon \leq \beta) + V^*(X_\beta)I(\beta < \tau_\varepsilon)] \\ &= E_x[V^*(X_{\tau_\varepsilon})I(\tau_\varepsilon \leq \beta, \tau_\varepsilon < \tilde{\tau}_\varepsilon) + \hat{S}_{\tilde{\tau}_\varepsilon} V^*(X_{\tau_\varepsilon})I(\tau_\varepsilon \leq \beta, \tau_\varepsilon = \tilde{\tau}_\varepsilon) \\ &\quad + V^*(X_\beta)I(\beta < \tau_\varepsilon)] \\ &\leq E_x[(G_1(X_{\tau_\varepsilon}) + \varepsilon)I(\tau_\varepsilon \leq \beta, \tau_\varepsilon < \tilde{\tau}_\varepsilon) + (G_1(X_{\tilde{\tau}_\varepsilon}) + \varepsilon)I(\tau_\varepsilon \leq \beta, \tau_\varepsilon = \tilde{\tau}_\varepsilon) \\ &\quad + G_2(X_\beta)I(\beta < \tau_\varepsilon)] \\ &\leq E_x[(G_1(X_{\tau_\varepsilon}) + \varepsilon)I(\tau_\varepsilon < \beta) + G_2(X_\beta)I(\beta < \tau_\varepsilon) \\ &\quad + (G_3(X_{\tau_\varepsilon}) + \varepsilon)I(\tau_\varepsilon = \beta)] \\ &\leq M_x(\tau_\varepsilon, \beta) + \varepsilon \end{aligned}$$

for every stopping time  $\beta$ . This completes the proof of (2.9) when the function  $x \mapsto \hat{V}^*(x)$  from (2.32) is assumed to be measurable.

13. If  $x \mapsto \hat{V}^*(x)$  is not assumed to be measurable, then the proof above can be repeated with reference only to  $\hat{S}^\sigma$  and  $\hat{S}$  under  $P_x$  with  $x$  given and fixed. In exactly the same way as above, this gives the identity  $\hat{V}^*(x) = V^*(x) = V_*(x)$  for this particular and thus all  $x$ . But then the measurability follows from the following general fact: If  $V^* = V_*$ , then  $V := V^* = V_*$  defines a measurable function.

To derive this fact consider the optimal stopping game (2.2)+(2.3) when  $X$  is a discrete-time Markov chain, so that  $\tau$  and  $\sigma$  (without loss of generality) take values in  $\{0, 1, 2, \dots\}$ . The horizon  $N$  (the upper bound for  $\tau$  and  $\sigma$  in (2.2)+(2.3) above) can be either finite or infinite. When  $N$  is finite, the most interesting case is when  $G_i = G_i(x, n)$  for  $i = 1, 2, 3$ , with  $G_1(x, N) = G_2(x, N) = G_3(x, N)$  for all  $x$ . When  $N$  is infinite, then

$$(2.82) \quad \liminf_{n \rightarrow \infty} G_2(X_n) \leq \limsup_{n \rightarrow \infty} G_1(X_n)$$

as stipulated following (2.3) above, and the common value for  $G_3(X_\infty)$  could formally be assigned as either of the two values in (2.82) (if  $\tau$  and  $\sigma$  are allowed to take the value  $\infty$ ).

Then the following Wald–Bellman equations are valid:

$$(2.83) \quad V_n(x) = G_1(x) \vee TV_{n-1}(x) \wedge G_2(x)$$

for  $n = 1, 2, \dots$ , where  $V_0$  is set to be either  $G_1$  or  $G_2$ . This yields  $V_N = V^* = V_*$ , with  $V_\infty = \lim_{n \rightarrow \infty} V_n$  if  $N = \infty$  (see [8] for details).

Recalling that  $T$  denotes the transition operator defined by

$$(2.84) \quad TF(x) = E_x F(X_1),$$

one sees that  $x \mapsto TF(x)$  is measurable whenever  $F$  is so (and  $E_x F(X_1)$  is well defined for all  $x$ ). Applying this argument inductively in (2.83) we see that  $x \mapsto V_N(x)$  is a measurable function. Thus, optimal stopping games for discrete-time Markov chains always lead to measurable value functions.

To treat the case of general  $X$ , let  $Q_n$  denote the set of all dyadic rationals  $k/2^n$  in the time set, and for a given stopping time  $\tau$  let  $\tau_n$  be defined by setting  $\tau_n = k/2^n$  on  $\{(k-1)/2^n < \tau \leq k/2^n\}$ . Then each  $\tau_n$  is a stopping time taking values in  $Q_n$ , and the following inequality is valid:

$$(2.85) \quad M_x(\tau, \sigma) \leq M_x(\tau_n, \sigma) + E_x |G_1(X_\tau) - G_1(X_{\tau_n})|$$

for every stopping time  $\sigma \in Q_n$  (meaning that  $\sigma$  takes values in  $Q_n$ ). Indeed, this can be derived as follows:

$$\begin{aligned} (2.86) \quad & M_x(\tau, \sigma) - M_x(\tau_n, \sigma) \\ &= E_x [G_1(X_\tau) I(\tau < \sigma) + G_2(X_\sigma) I(\sigma < \tau) + G_3(X_\tau) I(\tau = \sigma, \tau \neq \tau_n) \\ &\quad - G_1(X_{\tau_n}) I(\tau_n < \sigma) - G_2(X_\sigma) I(\sigma < \tau_n) - G_3(X_{\tau_n}) I(\tau_n = \sigma, \tau_n \neq \tau)] \\ &\leq E_x [(G_1(X_\tau) - G_1(X_{\tau_n})) I(\tau < \sigma) \\ &\quad + G_1(X_{\tau_n}) (I(\tau < \sigma) - I(\tau_n < \sigma) - I(\tau_n = \sigma, \tau_n \neq \tau)) \\ &\quad + G_2(X_\sigma) (I(\sigma < \tau) + I(\tau = \sigma, \tau \neq \tau_n) - I(\sigma < \tau_n))] \\ &= E_x [(G_1(X_\tau) - G_1(X_{\tau_n})) I(\tau < \sigma) + (G_1(X_{\tau_n}) - G_2(X_\sigma)) I(\tau < \sigma < \tau_n)] \end{aligned}$$

being true for any stopping times  $\tau, \sigma$ , and  $\tau_n$  such that  $\tau \leq \tau_n$ . In particular, if  $\sigma \in Q_n$  (and  $\tau_n$  is defined as above), then  $\{\tau < \sigma < \tau_n\} = \emptyset$ , so that (2.86) becomes

$$\begin{aligned} (2.87) \quad & M_x(\tau, \sigma) \leq M_x(\tau_n, \sigma) + E_x [(G_1(X_\tau) - G_1(X_{\tau_n})) I(\tau < \sigma)] \\ &\leq M_x(\tau_n, \sigma) + E_x |G_1(X_\tau) - G_1(X_{\tau_n})| \end{aligned}$$

as claimed in (2.85) above.

Let  $\tau_n^*$  and  $\sigma_n^*$  denote the optimal stopping times (in the Nash sense) for the optimal stopping game (2.2)+(2.3) with the time set  $Q_n$ , and let  $V_n(x)$  denote the corresponding value of the game, i.e.,

$$(2.88) \quad V_n(x) = M_x(\tau_n^*, \sigma_n^*)$$

for all  $x$ . (From (2.83) one sees that such optimal stopping times always exist in the discrete-time setting.) By the first part above (applied to the Markov chain  $(X_t)_{t \in Q_n}$ ) we know that  $x \mapsto V_n(x)$  is measurable.

Setting  $\varepsilon_n(x, \tau) = E_x |G_1(X_\tau) - G_1(X_{\tau_n})|$  we see that (2.85) reads

$$(2.89) \quad M_x(\tau, \sigma) \leq M_x(\tau_n, \sigma) + \varepsilon_n(x, \tau)$$

for every  $\tau$  and every  $\sigma \in Q_n$ . Hence we find that

$$\begin{aligned} (2.90) \quad & M_x(\tau, \sigma_n^*) \leq M_x(\tau_n, \sigma_n^*) + \varepsilon_n(x, \tau) \\ &\leq M_x(\tau_n^*, \sigma_n^*) + \varepsilon_n(x, \tau) = V_n(x) + \varepsilon_n(x, \tau). \end{aligned}$$

This implies that

$$(2.91) \quad \inf_{\sigma} M_x(\tau, \sigma) \leq \liminf_{n \rightarrow \infty} V_n(x),$$

since  $\varepsilon_n(x, \tau) \rightarrow 0$  by right-continuity of  $X$  and the fact that  $\tau_n \downarrow \tau$  as  $n \rightarrow \infty$  (using also the integrability condition (2.1) above). Taking the supremum over all  $\tau$  we conclude that

$$(2.92) \quad V_*(x) \leq \liminf_{n \rightarrow \infty} V_n(x)$$

for all  $x$ .

On the other hand, similarly to (2.86) one finds that

$$(2.93) \quad \begin{aligned} M_x(\tau, \sigma) - M_x(\tau, \sigma_n) &\geq E_x[(G_2(X_\sigma) - G_2(X_{\sigma_n})) I(\sigma < \tau) \\ &\quad + (G_2(X_{\sigma_n}) - G_1(X_\tau)) I(\sigma < \tau < \sigma_n)] \end{aligned}$$

for any stopping times  $\tau, \sigma$ , and  $\sigma_n$  such that  $\sigma \leq \sigma_n$ . If  $\sigma_n$  is defined analogously to  $\tau_n$  above (with  $\sigma$  in place of  $\tau$ ), then (2.93) yields the following analogue of (2.89) above:

$$(2.94) \quad M_x(\tau, \sigma) \geq M_x(\tau, \sigma_n) - \delta_n(x, \sigma),$$

where  $\delta_n(x, \sigma) = E_x|G_2(X_\sigma) - G_2(X_{\sigma_n})| \rightarrow 0$  as  $n \rightarrow \infty$  for the same reasons as above. This analogously yields

$$(2.95) \quad \limsup_{n \rightarrow \infty} V_n(x) \leq V^*(x)$$

for all  $x$ . Thus, if  $V^* = V_*$ , then by (2.92) and (2.95) we see that  $V := V^* = V_*$  satisfies

$$(2.96) \quad V(x) = \lim_{n \rightarrow \infty} V_n(x)$$

for all  $x$ . Since each  $V_n$  is measurable, we see that  $V$  is measurable as claimed. This completes the first part of the proof.

(II) In the second part of the proof we will assume that  $X$  is right-continuous and left-continuous over stopping times, and we will show that these hypotheses imply the Nash equilibrium (1.4) with  $\tau_*$  and  $\sigma_*$  from (2.4).

1. Since  $X$  is right-continuous we know by the first part of the proof above that  $V^* = V_*$ , with  $V := V^* = V_*$  defining a measurable function which by (2.7) satisfies

$$(2.97) \quad M_x(\tau, \sigma_\varepsilon) - \varepsilon \leq V(x) \leq M_x(\tau_\varepsilon, \sigma) + \varepsilon$$

for all  $\tau, \sigma, x$ , and  $\varepsilon > 0$ . Recalling from (2.5)+(2.6) that

$$(2.98) \quad \tau_\varepsilon = \inf \{ t : X_t \in D_1^\varepsilon \},$$

where  $D_1^\varepsilon = \{ V \leq G_1 + \varepsilon \}$ , we will now show that the second inequality in (2.97) implies

$$(2.99) \quad V(x) \leq M_x(\tau_0, \sigma)$$

for all  $\sigma$  and  $x$ , where  $\tau_0 = \inf \{ t : X_t \in D_1^0 \}$ , with  $D_1^0 = \{ V = G_1 \}$ . (Note that  $\tau_0$  coincides with  $\tau_*$  in the notation above.)

2. It is clear from the definitions that  $\tau_\varepsilon \uparrow \tau_{0-}$  as  $\varepsilon \downarrow 0$ , where  $\tau_{0-}$  is a stopping time satisfying  $\tau_{0-} \leq \tau_0$ . We will now show that  $\tau_{0-} = \tau_0$  P<sub>x</sub>-a.s. For this, let us first



establish the following general fact: If  $\rho_n$  and  $\rho$  are stopping times such that  $\rho_n \uparrow \rho$  as  $n \rightarrow \infty$ , then

$$(2.100) \quad \mathbb{E}_x V(X_\rho) \leq \liminf_{n \rightarrow \infty} \mathbb{E}_x V(X_{\rho_n}).$$

To see this recall from the first part of the proof above that  $V(X_\beta) = \hat{V}(X_\beta) = \hat{S}_\beta = \check{V}(X_\beta) = \check{S}_\beta$  for every stopping time  $\beta$ , where  $\check{V}$  and  $\check{S}$  are defined analogously to  $\hat{V}$  and  $\hat{S}$  but with  $\hat{M}_x(\tau, \sigma) = \mathbb{E}_x[G_1(X_\tau) I(\tau \leq \sigma) + G_2(X_\sigma) I(\sigma < \tau)]$  in place of  $\hat{M}_x(\tau, \sigma)$  and with the order of the supremum and the infimum being interchanged. Hence we find that

$$(2.101) \quad \mathbb{E}_x V(X_{\rho_n}) = \mathbb{E}_x \hat{S}_{\rho_n} = \sup_{\tau \geq \rho_n} \inf_{\sigma \geq \rho_n} \hat{M}_x(\tau, \sigma) \geq \sup_{\tau \geq \rho} \inf_{\sigma \geq \rho_n} \hat{M}_x(\tau, \sigma).$$

Taking  $\tau \geq \rho$  we find that

$$(2.102) \quad \begin{aligned} \hat{M}_x(\tau, \sigma) &= \mathbb{E}_x[(G_1(X_\tau) I(\tau < \sigma) + G_2(X_\sigma) I(\sigma \leq \tau)) I(\sigma \leq \rho)] \\ &\quad + \mathbb{E}_x[(G_1(X_\tau) I(\tau < \sigma \vee \rho) + G_2(X_{\sigma \vee \rho}) I(\sigma \vee \rho \leq \tau)) I(\sigma > \rho)] \\ &= \mathbb{E}_x[G_2(X_\sigma) I(\sigma \leq \rho)] \\ &\quad + \mathbb{E}_x[G_1(X_\tau) I(\tau < \sigma \vee \rho) + G_2(X_{\sigma \vee \rho}) I(\sigma \vee \rho \leq \tau)] \\ &\quad - \mathbb{E}_x[(G_1(X_\tau) I(\tau < \sigma \vee \rho) + G_2(X_{\sigma \vee \rho}) I(\sigma \vee \rho \leq \tau)) I(\sigma \leq \rho)] \\ &= \mathbb{E}_x[G_2(X_\sigma) I(\sigma \leq \rho) - G_2(X_\rho) I(\sigma \leq \rho)] \\ &\quad + \mathbb{E}_x[G_1(X_\tau) I(\tau < \sigma \vee \rho) + G_2(X_{\sigma \vee \rho}) I(\sigma \vee \rho \leq \tau)] \\ &= \mathbb{E}_x[G_2(X_{\sigma \wedge \rho}) - G_2(X_\rho)] + \hat{M}_x(\tau, \sigma \vee \rho). \end{aligned}$$

From (2.101) and (2.102) we get

$$(2.103) \quad \begin{aligned} \mathbb{E}_x V(X_{\rho_n}) &\geq \inf_{\sigma \geq \rho_n} \mathbb{E}_x[G_2(X_{\sigma \wedge \rho}) - G_2(X_\rho)] + \sup_{\tau \geq \rho} \inf_{\sigma \geq \rho} \hat{M}_x(\tau, \sigma) \\ &= \mathbb{E}_x V(X_\rho) + \inf_{\rho_n \leq \sigma \leq \rho} \mathbb{E}_x[G_2(X_\sigma) - G_2(X_\rho)]. \end{aligned}$$

Letting  $n \rightarrow \infty$  and using that the final expectation tends to zero, since  $X$  is left-continuous over stopping times and the integrability condition (2.1) holds, we get (2.100) as claimed.

Applying (2.100) to  $\tau_\varepsilon$  and  $\tau_{0-}$ , and recalling from the first part of the proof above that  $V(X_{\tau_\varepsilon}) \leq G_1(X_{\tau_\varepsilon}) + \varepsilon$   $\mathbb{P}_x$ -a.s., it follows that

$$(2.104) \quad \mathbb{E}_x V(X_{\tau_{0-}}) \leq \liminf_{\varepsilon \downarrow 0} \mathbb{E}_x V(X_{\tau_\varepsilon}) \leq \liminf_{\varepsilon \downarrow 0} \mathbb{E}_x[G_1(X_{\tau_\varepsilon}) + \varepsilon] = \mathbb{E}_x G_1(X_{\tau_{0-}})$$

upon using that  $G_1(X)$  is left-continuous over stopping times (as well as the integrability condition (2.1) above). Since, on the other hand, we have  $V(X_{\tau_{0-}}) \geq G_1(X_{\tau_{0-}})$ , we see from (2.104) that  $V(X_{\tau_{0-}}) = G_1(X_{\tau_{0-}})$  and thus  $\tau_0 \leq \tau_{0-}$   $\mathbb{P}_x$ -a.s., proving that  $\tau_0 = \tau_{0-}$   $\mathbb{P}_x$ -a.s. as claimed.

3. Motivated by passing to the limit in (2.97) for  $\varepsilon \downarrow 0$ , we will now establish the following general fact: If  $\tau_n$  and  $\tau$  are stopping times such that  $\tau_n \uparrow \tau$ , then

$$(2.105) \quad \limsup_{n \rightarrow \infty} M_x(\tau_n, \sigma) \leq M_x(\tau, \sigma)$$

for every stopping time  $\sigma$  given and fixed. To see this, note that

$$\begin{aligned}
 (2.106) \quad & M_x(\tau, \sigma) - M_x(\tau_n, \sigma) \\
 &= \mathbb{E}_x [G_1(X_\tau) I(\tau < \sigma) + G_2(X_\sigma) I(\sigma < \tau) + G_3(X_\tau) I(\tau = \sigma, \tau \neq \tau_n) \\
 &\quad - G_1(X_{\tau_n}) I(\tau_n < \sigma) - G_2(X_\sigma) I(\sigma < \tau_n) \\
 &\quad - G_3(X_{\tau_n}) I(\tau_n = \sigma, \tau_n \neq \tau)] \\
 &\geq \mathbb{E}_x [(G_1(X_\tau) - G_1(X_{\tau_n})) I(\tau < \sigma) \\
 &\quad + G_1(X_{\tau_n}) (I(\tau < \sigma) + I(\tau = \sigma, \tau \neq \tau_n) - I(\tau_n < \sigma)) \\
 &\quad + G_2(X_\sigma) (I(\sigma < \tau) - I(\sigma < \tau_n) - I(\tau_n = \sigma, \tau_n \neq \tau))] \\
 &= \mathbb{E}_x [(G_1(X_\tau) - G_1(X_{\tau_n})) I(\tau < \sigma) \\
 &\quad + (G_2(X_\sigma) - G_1(X_{\tau_n})) I(\tau_n < \sigma < \tau)] \\
 &\geq -\mathbb{E}_x |G_1(X_\tau) - G_1(X_{\tau_n})| \\
 &\quad - \mathbb{E}_x [(\sup_t |G_2(X_t)| + \sup_t |G_1(X_t)|) I(\tau_n < \sigma < \tau)].
 \end{aligned}$$

Letting  $n \rightarrow \infty$  and using the fact that the final two expectations tend to zero, since  $G_1(X)$  is left-continuous over stopping times and the integrability condition (2.1) holds, we see that (2.105) follows as claimed.

Applying (2.105) to  $\tau_\varepsilon$  and  $\tau_0$  upon letting  $\varepsilon \downarrow 0$  in (2.97), we get (2.99). The inequality

$$(2.107) \quad M_x(\tau, \sigma_0) \leq V(x)$$

can be established analogously. Combining (2.99) and (2.107), we get (1.4), and the proof is complete.  $\square$

**3. Concluding remarks.** The following example shows that the Nash equilibrium (1.4) may fail when  $X$  is right-continuous but not left-continuous over stopping times.

*Example 3.1.* Let the state space  $E$  of the process  $X$  be  $[-1, 1]$ . If  $X$  starts at  $x \in (-1, 1)$ , let  $X$  be a standard Brownian motion until it hits either  $-1$  or  $1$ ; at this time let  $X$  start afresh from  $0$  as an independent copy of  $B$  until it hits either  $-1$  or  $1$ ; and so on. If  $X$  starts at  $x \in \{-1, 1\}$ , let  $X$  stay at the same  $x$  for the rest of time.

It follows that  $X$  is a right-continuous strong Markov process which is not left-continuous over stopping times. Indeed, if we consider the first hitting time  $\rho_\varepsilon$  of  $X$  to  $b_\varepsilon$  under  $\mathbb{P}_x$  for  $x \in (-1, 1)$  given and fixed, where  $b_\varepsilon$  equals either  $-1 + \varepsilon$  or  $1 - \varepsilon$  for all  $\varepsilon > 0$  sufficiently small, then  $\rho_\varepsilon \uparrow \rho$  as  $\varepsilon \downarrow 0$  so that  $\rho$  is a stopping time; however, the value  $X_{\rho_\varepsilon} = b_\varepsilon$  does not converge to  $X_\rho = 0$  as  $\varepsilon \downarrow 0$ , implying the claim.

Let  $G_1(x) = x(x+1)-1$  and  $G_2(x) = -x(x-1)+1$  for  $x \in [-1, 1]$ , and let  $G_3$  be equal to  $G_1$  on  $[-1, 1]$ . Note that  $G_i(-1) = -1$  and  $G_i(1) = 1$  for  $i = 1, 2, 3$ . To include stopping times  $\tau$  and  $\sigma$  which are allowed to take the value  $\infty$  below, let us set  $G_3(X_\infty) = \limsup_{t \rightarrow \infty} G_1(X_t)$ . Note that  $G_3(X_\infty) \equiv 1$  under  $\mathbb{P}_x$  when  $x \in (-1, 1]$  and  $G_3(X_\infty) \equiv -1$  under  $\mathbb{P}_x$  when  $x = -1$ .

It is then easily seen (using the first part of Theorem 2.1 above) that  $V^*(x) = V_*(x) = x$  for all  $x \in [-1, 1]$  with  $\tau_\varepsilon = \inf \{t : X_t \leq a_\varepsilon^1 \text{ or } X_t \geq b_\varepsilon^1\}$  (where  $a_\varepsilon^1 < b_\varepsilon^1$  satisfy  $G_1(a_\varepsilon^1) = a_\varepsilon^1 - \varepsilon$  and  $G_1(b_\varepsilon^1) = b_\varepsilon^1 - \varepsilon$ ) and  $\sigma_\varepsilon = \inf \{t : X_t \leq a_\varepsilon^2 \text{ or } X_t \geq b_\varepsilon^2\}$  (where  $a_\varepsilon^2 < b_\varepsilon^2$  satisfy  $G_2(a_\varepsilon^2) = a_\varepsilon^2 + \varepsilon$  and  $G_2(b_\varepsilon^2) = b_\varepsilon^2 + \varepsilon$ ), being approximate stopping times satisfying (2.7) above. (Note that  $a_\varepsilon^i \downarrow -1$  and  $b_\varepsilon^i \uparrow 1$  as  $\varepsilon \downarrow 0$  for  $i = 1, 2$ .)

Thus the Stackelberg equilibrium (1.3) holds with  $V(x) = x$  for all  $x \in [-1, 1]$ . It is clear, however, that the Nash equilibrium fails as it is impossible to find stopping times  $\tau_*$  and  $\sigma_*$  satisfying (1.4) above. [Note that the natural candidates  $\tau \equiv \infty$  and  $\sigma \equiv \infty$  are ruled out, since  $M_x(\infty, \infty) = 1$  for  $x \in (-1, 1]$  and  $M_x(\infty, \infty) = -1$  for  $x = -1$ .]

The methodology used in the proof of Theorem 2.1 above (second part) extends from the Markovian approach to the martingale approach for optimal stopping games. For the sake of completeness we will formulate the analogous results of the martingale approach.

Let  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$  be a filtered probability space such that  $(\mathcal{F}_t)_{t \geq 0}$  is right-continuous and  $\mathcal{F}_0$  contains all  $\mathbb{P}$ -null sets from  $\mathcal{F}$ . Given adapted stochastic processes  $G^1, G^2, G^3$  on  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$  satisfying  $G_t^1 \leq G_t^2 \leq G_t^3$  for all  $t$  and the integrability condition

$$(3.1) \quad \mathbb{E} \sup_t |G_t^i| < \infty \quad (i = 1, 2, 3),$$

consider the optimal stopping game where the sup-player chooses a stopping time  $\tau$  to maximize, and the inf-player chooses a stopping time  $\sigma$  to minimize, the expected payoff

$$(3.2) \quad M(\tau, \sigma | \mathcal{F}_t) = \mathbb{E}[G_\tau^1 I(\tau < \sigma) + G_\sigma^2 I(\sigma < \tau) + G_\tau^3 I(\tau = \sigma) | \mathcal{F}_t]$$

for each  $t$  given and fixed. Note that, if  $\mathcal{F}_0$  is trivial (in the sense that  $\mathbb{P}(F)$  equals either 0 or 1 for all  $F \in \mathcal{F}_0$ ), then  $M(\tau, \sigma | \mathcal{F}_0)$  equals  $\mathbb{E}[G_\tau^1 I(\tau < \sigma) + G_\sigma^2 I(\sigma < \tau) + G_\tau^3 I(\tau = \sigma)]$ , and this expression is then denoted by  $M(\tau, \sigma)$  for all  $\tau$  and  $\sigma$ .

Define the upper value and the lower value of the game by

$$(3.3) \quad V_t^* = \operatorname{ess\,inf}_{\sigma \geq t} \operatorname{ess\,sup}_{\tau \geq t} M(\tau, \sigma | \mathcal{F}_t) \quad \text{and} \quad V_t^t = \operatorname{ess\,sup}_{\tau \geq t} \operatorname{ess\,inf}_{\sigma \geq t} M(\tau, \sigma | \mathcal{F}_t),$$

respectively, where the horizon  $T$  (the upper bound for  $\tau$  and  $\sigma$  above) may be either finite or infinite. If  $T < \infty$ , then it is assumed that  $G_T^1 = G_T^2 = G_T^3$ . If  $T = \infty$ , then it is assumed that  $\liminf_{t \rightarrow \infty} G_t^2 \leq \limsup_{t \rightarrow \infty} G_t^1$ , and the common value for  $G_\infty^3$  could formally be assigned as either of the preceding two values (if  $\tau$  and  $\sigma$  are allowed to take the value  $\infty$ ).

**THEOREM 3.2.** *Consider the optimal stopping game (3.3). If  $G_i$  is right-continuous for  $i = 1, 2, 3$ , then the Stackelberg equilibrium holds in the sense that*

$$(3.4) \quad V_t^* = V_t^t \quad \mathbb{P}\text{-a.s.},$$

with  $V_t := V_t^* = V_t^t$  defining a right-continuous process (modification) for  $t \geq 0$ . Moreover, the stopping times

$$(3.5) \quad \tau_\varepsilon = \inf \{ t : V_t \leq G_t^1 + \varepsilon \} \quad \text{and} \quad \sigma_\varepsilon = \inf \{ t : V_t \geq G_t^2 - \varepsilon \}$$

satisfy the following inequalities:

$$(3.6) \quad M(\tau, \sigma_\varepsilon | \mathcal{F}_t) - \varepsilon \leq M(\tau_\varepsilon, \sigma_\varepsilon | \mathcal{F}_t) \leq M(\tau_\varepsilon, \sigma | \mathcal{F}_t) + \varepsilon$$

for each  $t$  and every  $\varepsilon > 0$ . If  $G_i$  is right-continuous and left-continuous over stopping times for  $i = 1, 2, 3$ , then the Nash equilibrium holds in the sense that the stopping times

$$(3.7) \quad \tau_* = \inf \{ t : V_t = G_t^1 \} \quad \text{and} \quad \sigma_* = \inf \{ t : V_t = G_t^2 \}$$

satisfy the following inequalities:

$$(3.8) \quad M(\tau, \sigma_* | \mathcal{F}_t) \leq M(\tau_*, \sigma_* | \mathcal{F}_t) \leq M(\tau_*, \sigma | \mathcal{F}_t)$$

for each  $t$  and all stopping times  $\tau$  and  $\sigma$ .

*Proof.* The first part of the theorem (Stackelberg equilibrium) was established in [18] (under slightly more restrictive conditions on integrability and the common value at the end of time but the same method extends to cover the present case without major changes). The second part of the theorem (Nash equilibrium) can be derived using the same arguments as in the second part of the proof of Theorem 2.1 above.  $\square$

Note that the second part of Theorem 3.2 (Nash equilibrium) is applicable to all Lévy processes (without additional hypotheses on the jump structure).

#### REFERENCES

- [1] E. J. BAURDOUX AND A. E. KYPRIANOU, *Further calculations for Israeli options*, Stoch. Stoch. Rep., 76 (2004), pp. 549–569.
- [2] A. BENSOUSSAN AND A. FRIEDMAN, *Nonlinear variational inequalities and differential games with stopping times*, J. Funct. Anal., 16 (1974), pp. 305–352.
- [3] A. BENSOUSSAN AND A. FRIEDMAN, *Nonzero-sum stochastic differential games with stopping times and free boundary problems*, Trans. Amer. Math. Soc., 231 (1977), pp. 275–327.
- [4] F. BOETIUS, *Bounded variation singular stochastic control and Dynkin game*, SIAM J. Control Optim., 44 (2005), pp. 1289–1321.
- [5] E. B. DYNKIN, *Game variant of a problem of optimal stopping*, Soviet Math. Dokl., 10 (1969), pp. 16–19.
- [6] E. EKSTRÖM, *Properties of game options*, Math. Methods Oper. Res., 63 (2006), pp. 221–238.
- [7] E. EKSTRÖM AND S. VILLENEUVE, *On the value of optimal stopping games*, Ann. Appl. Probab., 16 (2006), pp. 1576–1596.
- [8] N. V. ELBAKIDZE, *Construction of the cost and optimal policies in a game problem of stopping a Markov process*, Theory Probab. Appl., 21 (1976), pp. 163–168.
- [9] E. B. FRID, *The optimal stopping rule for a two-person Markov chain with opposing interests*, Theory Probab. Appl., 14 (1969), pp. 713–716.
- [10] A. FRIEDMAN, *Stochastic games and variational inequalities*, Arch. Ration. Mech. Anal., 51 (1973), pp. 321–346.
- [11] P. V. GAPEEV, *The Spread Option Optimal Stopping Game, Exotic Option Pricing and Advanced Lévy Models*, John Wiley, New York, 2005, pp. 293–305.
- [12] S. M. GUSEIN-ZADE, *On a game connected with a Wiener process*, Theory Probab. Appl., 14 (1969), pp. 701–704.
- [13] Y. KIFER, *Optimal stopping in games with continuous time*, Theory Probab. Appl., 16 (1971), pp. 545–550.
- [14] Y. KIFER, *Game options*, Finance Stoch., 4 (2000), pp. 443–463.
- [15] C. KÜHN AND A. E. KYPRIANOU, *Callable puts as composite exotic options*, Math. Finance, 17 (2007), pp. 487–502.
- [16] A. E. KYPRIANOU, *Some calculations for Israeli options*, Finance Stoch., 8 (2004), pp. 73–86.
- [17] R. LARAKI AND E. SOLAN, *The value of zero-sum stopping games in continuous time*, SIAM J. Control Optim., 43 (2005), pp. 1913–1922.
- [18] J. P. LEPELTIER AND M. A. MAINGUENEAU, *Le jeu de Dynkin en théorie générale sans l'hypothèse de Mokobodski*, Stochastics, 13 (1984), pp. 25–44.
- [19] G. PESKIR AND A. N. SHIRYAEV, *Optimal Stopping and Free-Boundary Problems*, Lectures Math. ETH Zürich, Birkhäuser, Basel, 2006.
- [20] A. N. SHIRYAEV, *Optimal Stopping Rules*, Springer-Verlag, Berlin, 1978.
- [21] J. L. SNELL, *Applications of martingale system theorems*, Trans. Amer. Math. Soc., 73 (1952), pp. 293–312.
- [22] L. STETTNER, *Zero-sum Markov games with stopping and impulsive strategies*, Appl. Math. Optim., 9 (1982), pp. 1–24.
- [23] L. STETTNER, *On closedness of general zero-sum stopping game*, Bull. Poli. Acad. Sci. Math., 32 (1984), pp. 351–361.

## ON THE OPTIMAL STOCHASTIC IMPULSE CONTROL OF LINEAR DIFFUSIONS\*

LUIS H. R. ALVAREZ<sup>†</sup> AND JUKKA LEMPA<sup>†</sup>

**Abstract.** We consider a class of stochastic impulse control problems of linear diffusions arising in studies considering the determination of optimal dividend policies. This class of problems appears also in studies analyzing the optimal management of renewable resources. We state a set of weak conditions guaranteeing both existence and uniqueness of the boundary characterizing the optimal policy and its value. We also analyze two associated stochastic control problems and establish a general ordering for both the values and the marginal values of the considered stochastic control problems. In this way we extend previous findings obtained by relying on linear payoff characterizations.

**Key words.** stochastic impulse and singular control, optimal stopping, diffusions

**AMS subject classifications.** 93E20, 60G40, 49J10, 49K10

**DOI.** 10.1137/060659375

**1. Introduction.** A stochastic impulse control policy can be characterized by two factors, namely, by the sequence of random dates at which the policy is exercised and by the sequence of impulses describing the magnitude of the applied policies. Thus, solving an impulse control problem typically involves the consideration of two endogenously determined variables: the timing and size of an impulse policy. For example, in most forest economic applications of stochastic impulse control the implemented impulse size is constrained by an exogenously determined generic initial state at which the underlying stochastic process is restarted after the forest has been harvested (see, for example, [3], [4], [5], [6], [37], and [39]). Hence, in those models the only endogenous variable determining the size of the optimal policy is the single boundary at which the irreversible policy is optimally exercised. On the other hand, most capital theoretic and cash flow management applications of impulse control are based on models where both the exercise boundary at which the impulse policy is exercised and the generic initial state at which the controlled process is restarted after the irreversible policy has been exerted have to be simultaneously determined (see, for example, [7], [9], [10], [18], [29], [34]; see also [26] for an excellent survey on stochastic impulse control applications in finance). Given the general applicability of stochastic impulse control models in various fields, it is not surprising that the mathematical analysis of such problems is well established (see, for example, [16], [17], [21], [28], [30], [32]; see also [12] for a seminal textbook on quasi-variational inequalities and impulse control). In most cases the impulse control problem is studied by relying on a combination of variational and quasi-variational inequalities. Even though that approach is general and applies in the multidimensional setting as well, it typically results into functional inequalities which, depending naturally on the explicit form of

---

\*Received by the editors May 9, 2006; accepted for publication (in revised form) July 18, 2007; published electronically February 15, 2008.

<http://www.siam.org/journals/sicon/47-2/65937.html>

<sup>†</sup>Department of Economics, Quantitative Methods in Management, Turku School of Economics and Business Administration, FIN-20500 Turku, Finland (luis.alvarez@tukkk.fi, jukka.lempa@tukkk.fi). The first author acknowledges the financial support from the Foundation for the Promotion of the Actuarial Profession, the Finnish Insurance Society, the Yrjö Jahnsson Foundation, and the Research Unit of Economic Structures and Growth (RUESG) at the University of Helsinki.

the considered problem, may be relatively difficult to analyze and in that way difficult to interpret in terms of the particular application.

Given the arguments mentioned above, we consider in this study a class of stochastic impulse control problems of linear time-homogenous diffusion processes arising, among others, in various cash flow management applications and in studies on the rational management of renewable resources. As usual, we assume that the decision maker has to choose both the timing and the size of the optimal policy affecting the dynamics of the underlying diffusion. We generalize the analysis of the study [7] in two ways. First, instead of relying on a simple linear and state-independent exercise payoff, we introduce a state-dependent and potentially nonlinear cash flow term measuring the revenue flow accrued from continuing operation (in capital theoretic applications of impulse control this flow can be interpreted either as the short-run profit flow or as a continuous dividend stream, and in forest economics this flow term is typically interpreted as the flow of revenues accrued from amenity services; cf. [5]). This extension is of interest, since, as our analysis clearly demonstrates, in the presence of a state-dependent and potentially nonlinear cash flow, no strong concavity requirements are needed in order to guarantee both existence and uniqueness of an optimal policy. This is a result which is in sharp contrast with the findings of the linear state-independent exercise payoff case studied in [7]. Second, in order to model the potential imperfect controllability of the underlying stochastic dynamics, we also consider situations where an arbitrary admissible impulse may result in a jump discontinuity which is either greater or smaller than the size of the actual impulse (such configurations typically arise in models considering the effects of taxation or other financial frictions on rational cash flow management). We model the imperfect controllability as scalar multiplication of the applied impulse control policy. Our analysis shows that small (linear) changes in the controllability of the system result in nonlinear and possibly dramatic changes in the required rate of return and, consequently, in both the optimal policy and its value.

Instead of analyzing the considered class of stochastic impulse control problems directly by relying on the ordinary Hamilton–Jacobi–Bellman approach, we follow the approach introduced in [3] and [4] and first derive the value accrued from applying a potentially suboptimal stochastic impulse control policy characterized by a sequence of constant-sized impulses exerted every time the underlying diffusion hits a predetermined and constant exercise threshold. By relying on standard nonlinear programming techniques we then state the ordinary first order necessary conditions characterizing both the exercise threshold and the impulse size of a potentially optimal policy maximizing the value of the associated class of Markovian functionals (for a recent study utilizing a similar idea, see [10]). The advantage of this approach is that it simplifies the economic analysis of the optimal policy and its value by admitting the application of standard marginalistic interpretations familiar from ordinary microeconomic theory. We present a set of relatively weak sufficient conditions under which a unique pair satisfying the necessary first order conditions exists. We establish that this pair constitutes the optimal impulse control policy in terms of both the size of the optimal impulse and the threshold at which the irreversible policy should be optimally exerted. In accordance with these observations, we then find that given the policy mentioned above the iteratively defined Markovian functional actually constitutes the value of the optimal stochastic impulse control policy.

We also consider two associated stochastic control problems (namely, a singular stochastic control and an optimal stopping problem) and study the boundary value

problem connecting the values (including the value of the stochastic impulse control problem). Both of these associated classical problems have been studied extensively (for singular control cf., e.g., [11], [22], [8], [25], [3], and [38]; for optimal stopping we refer the reader to the recent textbook [35]), and it is well known from this literature that singular stochastic control problems and optimal stopping problems are closely connected. More precisely, for a large class of singular stochastic control problems the derivative of the optimal value (i.e., the marginal value) coincides with the value of an associated optimal stopping problem of a certain transformed diffusion (cf. [23], [24], [14], [13]; see also [2]). The results of the current study connecting the problems differ from the previous characterizations in two ways. First, in all three of the considered separate stochastic control problems, the underlying dynamics are the same in the absence of control. Second, we study how, on the one hand, the optimal values and, on the other hand, the marginal values are interrelated. As is intuitively clear, we find that the value of the associated singular stochastic control problem dominates the value of the stochastic impulse control problem, which, in turn, dominates the value of the associated optimal stopping problem. Somewhat surprisingly, we also find that the same ordering is satisfied by the marginal values of the optimal policies as well. More precisely, we establish that the marginal value of the associated singular stochastic control problem dominates the marginal value of the stochastic impulse control problem, which, in turn, dominates the marginal value of the associated optimal stopping problem. This finding is important from the point of view of economic and financial applications, since our results demonstrate that increased policy flexibility unambiguously increases the *Tobin's marginal  $q$*  (i.e., marginal value) associated with the considered stochastic control problems as well. In this way our results extend the findings of [7] by demonstrating that the positivity of the relationship between the (marginal) value and the flexibility of the admissible policy is satisfied in the presence of a state-dependent and potentially nonlinear cash flow as well.

The contents of this study are as follows. In section 2 we present the considered class of stochastic impulse control problems. In section 3 we then state a set of auxiliary results and analyze the two associated stochastic control problems. In section 4 we then analyze the considered stochastic impulse control problem and state our main results. Finally, our results are explicitly illustrated in section 5 in a model based on geometric Brownian motion.

## 2. The impulse control problem.

**2.1. General setup.** It is our purpose in this study to analyze a class of stochastic impulse control problems of linear diffusions arising in many financial and economical applications of stochastic control theory. In order to accomplish this task, let  $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$  denote a complete filtered probability space satisfying the usual conditions and assume that the dynamics of the underlying controlled diffusion process are given by the generalized Itô equation

$$(2.1) \quad X_t^\nu = x + \int_0^t \mu(X_s^\nu) ds + \int_0^t \sigma(X_s^\nu) dW_s - \sum_{\tau_k \leq t} \beta \zeta_k, \quad 0 \leq t \leq H_0^\nu,$$

where  $\beta > 0$  is an exogenously given constant,  $H_0^\nu = \inf\{t \geq 0 : X_t^\nu \leq 0\}$  denotes the possibly finite first exit time of the controlled diffusion  $X^\nu$  from the state-space  $\mathbb{R}_+$ , and  $\mu : \mathbb{R}_+ \rightarrow \mathbb{R}$  and  $\sigma : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  are known sufficiently smooth mappings (at least continuous) guaranteeing the existence of a solution for the stochastic differential

equation

$$(2.2) \quad dX_t = \mu(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = x,$$

characterizing the dynamics of the underlying diffusion in the absence of interventions (cf. [15, pp. 46–47]). In a rational cash flow management application, the uncontrolled diffusion  $X$  represents the evolution of cash reserve in the absence of dividend payments and the impulse  $\zeta_k$  represent the amount paid out as dividends to stock holders at the corresponding time  $\tau_k$ ; hence the sum  $\sum_{\tau_k \leq t} \beta \zeta_k$  represents the total amount of dividends paid out until time  $t$  (or the cumulative portfolio wealth consumed by the decision maker up to time  $t$ ). The parameter  $\beta$  can be interpreted as a measure of the *imperfect controllability* of the underlying stochastic dynamics, since whenever  $\beta \neq 1$  an arbitrary admissible impulse results in a jump discontinuity which is either greater or smaller than the size of the actual impulse  $\zeta_k$ . As in [32], an admissible impulse control policy for the system (2.1) is a potentially infinite joint sequence  $\nu = \{(\tau_k, \zeta_k)\}_{k=1}^N$ ,  $N \leq \infty$ , where  $\{\tau_k\}_{k=1}^N$  denotes an increasing sequence of  $\mathcal{F}_t$ -stopping times for which  $\tau_1 \geq 0$  and  $\{\zeta_k\}_{k=1}^N$  denotes a sequence of nonnegative,  $\mathcal{F}_{\tau_k}$ -measurable impulses exerted at the corresponding intervention dates  $\{\tau_k\}_{k=1}^N$ , respectively. We denote as  $\mathcal{V}$  the class of admissible impulse controls  $\nu$  and assume that  $\tau_k \rightarrow H_0^\nu$  almost surely for all  $\nu \in \mathcal{V}$  and  $x \in \mathbb{R}_+$ . This convergence should be understood as follows: If  $N = \infty$ , then the convergence is typical almost sure convergence. However, if  $N < \infty$ , then we augment the control  $\nu$  with a pair  $(\tau_{N+1}, \zeta_{N+1}) := (H_0^\nu, 0)$ . As usual, we denote as  $\mathcal{A} = \frac{1}{2}\sigma^2(x)\frac{d^2}{dx^2} + \mu(x)\frac{d}{dx}$  the differential operator associated with  $X$ . For the diffusion  $X$ , the densities of scale function  $S$  and the speed measure  $m$  are defined as  $S'(x) = \exp(-\int^x \frac{2\mu(y)}{\sigma^2(y)} dy)$  and  $m'(x) = 2/(\sigma^2(x)S'(x))$ , respectively, for all  $x \in \mathbb{R}$ .

Denote as  $\mathcal{L}_1$  the class of measurable mappings  $f: \mathbb{R}_+ \rightarrow \mathbb{R}$  satisfying the condition

$$\mathbf{E}_x \left[ \int_0^{H_0} e^{-rs} |f(X_s)| ds \right] < \infty,$$

where  $r > 0$  is a given constant and  $H_0 = \inf\{t \geq 0 : X_t \leq 0\}$  is the first, potentially infinite, exit time for the uncontrolled diffusion  $X$  from  $\mathbb{R}_+$ . For a given  $f \in \mathcal{L}_1$ , define the resolvent  $(R_r f): \mathbb{R}_+ \rightarrow \mathbb{R}$  as

$$(R_r f)(x) = \mathbf{E}_x \left[ \int_0^{H_0} e^{-rs} f(X_s) ds \right].$$

The resolvent  $(R_r f)$  measures the expected cumulative present value of the cash flow  $f(X_t)$  from the present up to  $H_0$ . It is well known from the literature on linear diffusions that  $(R_r f)$  can be rewritten as (cf. [31])

$$(2.3) \quad (R_r f)(x) = B^{-1}\varphi(x) \int_0^x \psi(y)f(y)m'(y)dy + B^{-1}\psi(x) \int_x^\infty \varphi(y)f(y)m'(y)dy,$$

where  $\psi$  denotes the increasing and  $\varphi$  the decreasing fundamental solutions of the linear differential equation  $\mathcal{A}u = ru$  and  $B = (\psi'(x)\varphi(x) - \varphi'(x)\psi(x))/S'(x)$  denotes the Wronskian determinant of  $X$  (for a characterization of the fundamental solutions and the Green function of a linear diffusion, see [15, pp. 18–20]). The fundamental solutions  $\psi$  and  $\varphi$  constitute the *minimal  $r$ -excessive functions* for  $X$  since any non-trivial  $r$ -excessive function for  $X$  can be expressed in terms of  $\psi$  and  $\varphi$  via an integral expression (the so-called Martin integral representation theorem; cf. [15, p. 33]).



**2.2. Formulation of the impulse control problem.** Given the stochastic dynamics in (2.1) and the assumptions presented above on the dynamics of the controlled system, define the *expected cumulative net present value of the revenues from the present up to a potentially infinite future* as

$$(2.4) \quad J_c^\nu(x) = \mathbf{E}_x \left[ \int_0^{H_0^\nu} e^{-rs} \pi(X_s^\nu) ds + \sum_{k=1}^N e^{-r\tau_k} (\lambda \zeta_k - c) \right],$$

where  $r > 0$  is the discount rate,  $\lambda > 0$  is an exogenously given constant,  $c > 0$  is a known constant measuring a lump-sum sunk cost associated with the irreversible policy, and  $\pi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is a given continuous, nondecreasing and nonnegative mapping measuring the *revenue flow accrued from continuing the operation*. This type of objective functional arise frequently in studies considering rational cash flow management (optimal dividend policy) and in studies considering the rational harvesting of renewable resources. In cash flow management application, the right-hand side of (2.4) represents the expected cumulative present value of the revenues accrued from the present up to the potentially infinite horizon at which the firm is liquidated. Along the lines of our interpretation of  $\beta$ , the parameter  $\lambda$  can be interpreted as another measure of imperfect controllability, since whenever  $\lambda \neq 1$  the realized revenue is strictly smaller or larger than the size of the actual impulse  $\zeta_k$  (in most economic and financial applications this parameter arises due to either taxes or subsidies).

Given the definition of  $J_c^\nu$  we plan to study the stochastic impulse control problem

$$(2.5) \quad V_c(x) = \sup_{\nu \in \mathcal{V}} J_c^\nu(x), \quad x \in \mathbb{R}_+,$$

and to determine an admissible impulse control  $\nu^*$  for which the maximum  $J_c^{\nu^*}(x) = V_c(x)$  is attained for all  $x \in \mathbb{R}_+$ . We will analyze the problem (2.5) under the following standing assumptions.

ASSUMPTION 2.1. (1) *We assume that the upper boundary  $\infty$  is natural and that the lower boundary 0 is natural, exit, or regular for the uncontrolled diffusion  $X$  in the absence of interventions. In the case when the origin is regular, we assume that it is killing.*

(2) *Define the mapping  $\theta : \mathbb{R}_+ \rightarrow \mathbb{R}$  as*

$$(2.6) \quad \theta(x) = \beta\pi(x) + \lambda\rho(x),$$

where  $\rho(x) = \mu(x) - rx$ . Throughout the study, we assume that  $\theta \in \mathcal{L}_1(\mathbb{R}_+)$  and that there is a unique state  $x^* \geq 0$  for which  $\theta$  is increasing on  $(0, x^*)$  and decreasing on  $(x^*, \infty)$ . Moreover, we assume that  $0 \leq \lim_{x \rightarrow 0+} \theta(x) < \infty$  and that  $\lim_{x \rightarrow \infty} \theta(x) < 0$ .

We prove all our main results under Assumption 2.1. First, we prove that under Assumption 2.1 the impulse control problem (2.5) has a unique solution. We characterize the optimal admissible impulse control policy explicitly as a threshold policy with a constant threshold and a constant impulse and derive a closed form expression of the optimal value function (Theorem 4.2). Moreover, we consider two associated control problems, namely, the associated singular control problem (3.6) and the optimal stopping problem (3.12), where the uncontrolled underlying dynamics follow the same diffusion  $X$  given by (2.2) as in the impulse control problem. This allows us to study the effect of flexibility of the control on the optimal value. We also show that these associated problems are solvable under Assumption 2.1 (Lemmas 3.4 and 3.6). We establish that the value functions of all three problems can be presented as

solutions of a certain free-boundary problem (see (4.16)). Finally, we prove a strong ordering of both the values and the marginal values of the control problems (Theorem 4.6). Informally, this ordering can be expressed as follows: *Increased flexibility of the control increases not only the value but also the rate at which the value grows.*

The assumption (1) characterizes the boundary behavior of the underlying diffusion. In line with most financial and economical applications, the upper boundary is assumed to be natural. Hence, even though the underlying controlled diffusion process may be expected to drift toward infinity, it is never expected to attain it in finite time. In the context of cash flow management applications, this assumption can be interpreted as a requirement that the retained profits from which dividends are paid out to the shareholders cannot become infinitely large in finite time. On the other hand, the assumed behavior of the underlying diffusion at the lower boundary is in line with the concept of liquidation and essentially guarantees that no dividends can be paid from negative reserves (since the time horizon is defined up to  $H_0^v$ ). From a mathematical point of view, the minimal  $r$ -harmonic functions  $\psi$  and  $\varphi$  satisfy useful limiting conditions depending on the boundary behavior of  $X$  (see [15, p. 19]). More precisely, if the lower boundary 0 is natural, then  $\lim_{x \rightarrow 0+} \psi(x) = 0$ ,  $\lim_{x \rightarrow 0+} \psi'(x)/S'(x) = 0$ ,  $\lim_{x \rightarrow 0+} \varphi(x) = \infty$ , and  $\lim_{x \rightarrow 0+} \varphi'(x)/S'(x) = -\infty$ . Analogous conditions hold also for the upper boundary  $\infty$  (which was assumed to be natural). If the origin is an exit boundary, then the second and third conditions are replaced by  $\lim_{x \rightarrow 0+} \psi'(x)/S'(x) > 0$  and  $\lim_{x \rightarrow 0+} \varphi(x) < \infty$ . In the case of a killing boundary, we have a limiting condition only for  $\psi$ , namely, that  $\lim_{x \rightarrow 0+} \psi(x) = 0$ .

The assumption (2) is also quite reasonable from the point of view of financial and economical applications. In a cash flow management application, the function  $\theta$  measures the expected net return (the sum of the continuous dividend flow and the expected capital gain) accrued from postponing the dividend payment into the future instead of paying out dividends instantaneously. Hence, the assumed limiting behavior of the net return guarantees that the rate of return earned from a retained unit dominates its opportunity cost (i.e., the return from a safe investment) when the reserves are low and that the opposite argument is valid when the reserves are large. In this respect our assumption (2) characterizes a set of sufficient conditions under which the decision maker has incentives to distribute part of the reserves when they become sufficiently large without liquidating the corporation instantaneously. The absence of speculative bubbles condition  $\theta \in \mathcal{L}_1$  guarantees that the expected cumulative present value of the net returns accrued from the present up to the liquidation date is finite.

In order to proceed in the analysis of the considered class of stochastic control problems, we first establish the following verification theorem.

LEMMA 2.1. *Assume that there is a mapping  $F : \mathbb{R}_+ \mapsto \mathbb{R}_+$  satisfying the following conditions.*

- (a) *The function  $F(x) - (R_r\pi)(x)$  is nonnegative and  $r$ -superharmonic for  $X_t$ .*
- (b)  *$F$  satisfies the inequality*

$$(2.7) \quad F(x) \geq \sup_{\beta\zeta \in [0, x]} [\lambda\zeta - c + F(x - \beta\zeta)]$$

for all  $x \in \mathbb{R}_+$ . Then,  $F(x) \geq V_c(x)$  for all  $x \in \mathbb{R}_+$ .

*Proof.* Let  $\nu \in \mathcal{V}$  be an admissible stochastic impulse control. Since  $\{\tau_j\}_{j=1}^N$  is an increasing sequence of stopping times, we first observe that the assumed

$r$ -superharmonicity of the function  $F(x) - (R_r\pi)(x)$  implies that (cf. [33, Lemma 10.1.3, p. 207])

$$(2.8) \quad \mathbf{E}_{\mathcal{F}_{\tau_j}} \left[ e^{-r\tau_{j+1}} (F(X_{\tau_{j+1}-}^\nu) - (R_r\pi)(X_{\tau_{j+1}-}^\nu)) \right] \leq e^{-r\tau_j} (F(X_{\tau_j}^\nu) - (R_r\pi)(X_{\tau_j}^\nu)).$$

Since  $((\mathcal{A} - r)(R_r\pi))(x) = -\pi(x)$  for all  $x \in \mathbb{R}_+$ , application of Dynkin's theorem to  $(R_r\pi)$  yields

$$\mathbf{E}_{\mathcal{F}_{\tau_j}} \left[ e^{-r\tau_{j+1}} (R_r\pi)(X_{\tau_{j+1}-}^\nu) \right] = e^{-r\tau_j} (R_r\pi)(X_{\tau_j}^\nu) - \mathbf{E}_{\mathcal{F}_{\tau_j}} \left[ \int_{\tau_j}^{\tau_{j+1}-} e^{-rs} \pi(X_s^\nu) ds \right],$$

implying that inequality (2.8) can be re-expressed as

$$e^{-r\tau_j} F(X_{\tau_j}^\nu) - \mathbf{E}_{\mathcal{F}_{\tau_j}} \left[ e^{-r\tau_{j+1}} F(X_{\tau_{j+1}-}^\nu) \right] \geq \mathbf{E}_{\mathcal{F}_{\tau_j}} \left[ \int_{\tau_j}^{\tau_{j+1}-} e^{-rs} \pi(X_s^\nu) ds \right].$$

Taking expectations and invoking the tower property of conditional expectations then yield

$$\mathbf{E}_x \left[ e^{-r\tau_j} F(X_{\tau_j}^\nu) \right] - \mathbf{E}_x \left[ e^{-r\tau_{j+1}} F(X_{\tau_{j+1}-}^\nu) \right] \geq \mathbf{E}_x \left[ \int_{\tau_j}^{\tau_{j+1}-} e^{-rs} \pi(X_s^\nu) ds \right].$$

Letting  $\tau_0 = 0$ , summing terms from  $j = 0$  to  $j = n \wedge N$ , and applying the nonnegativity of the mapping  $F(x)$  result in

$$F(x) \geq \sum_{j=1}^{n \wedge N} \mathbf{E}_x \left[ e^{-r\tau_j} F(X_{\tau_j-}^\nu) - F(X_{\tau_j}^\nu) \right] + \mathbf{E}_x \left[ \int_0^{\tau_{n \wedge N+1}-} e^{-rs} \pi(X_s^\nu) ds \right].$$

Since  $X_{\tau_j} = X_{\tau_j-} - \beta\zeta_j$  for any admissible strategy and  $F$  satisfies the quasi-variational inequality  $F(x) \geq \sup_{\beta\zeta \in [0, x]} [\lambda\zeta - c + F(x - \beta\zeta)]$  for all  $x \in \mathbb{R}_+$ , we find that

$$F(x) \geq \mathbf{E}_x \left[ \int_0^{\tau_{n \wedge N+1}-} e^{-rs} \pi(X_s^\nu) ds + \sum_{j=1}^{n \wedge N} e^{-r\tau_j} (\lambda\zeta_j - c) \right].$$

Letting  $n \rightarrow \infty$  and invoking dominated convergence then finally imply that

$$F(x) \geq \mathbf{E}_x \left[ \int_0^{H_0^\nu} e^{-rs} \pi(X_s^\nu) ds + \sum_{j=1}^N e^{-r\tau_j} (\lambda\zeta_j - c) \right].$$

Since this inequality is valid for any admissible impulse control, it has to be valid for the optimal as well, from which the alleged result follows.  $\square$

Lemma 2.1 states a set of considerably weak sufficient conditions which can be applied in the verification of the optimality of a value attained by applying an admissible policy. An interesting implication of Lemma 2.1 stating a set of more easily applicable sufficient conditions is now summarized in the following.

**COROLLARY 2.2.** *Assume that the mapping  $F: \mathbb{R}_+ \mapsto \mathbb{R}_+$  satisfies the conditions  $F \in C^1(\mathbb{R}_+) \cap C^2(\mathbb{R}_+ \setminus \mathcal{D})$ , where  $\mathcal{D}$  is a set of measure zero and  $F''(x \pm) < \infty$  for all  $x \in \mathcal{D}$ . Assume also that  $F$  satisfies the quasi-variational inequality (2.7) for all*

$x \in \mathbb{R}_+$  and the variational inequality  $(\mathcal{A}F)(x) - rF(x) + \pi(x) \leq 0$  for all  $x \notin \mathcal{D}$ . Then,  $F(x) \geq V_c(x)$  for all  $x \in \mathbb{R}_+$ .

*Proof.* As was established in Theorem D.1. in [33, pp. 315–318] the conditions of our corollary guarantee that there exists a sequence  $\{F_n\}_{n=1}^\infty$  of mappings  $F_n \in C^2(\mathbb{R}_+)$  such that

- (i)  $F_n \rightarrow F$  uniformly on compact subsets of  $\mathbb{R}_+$  as  $n \rightarrow \infty$ ;
- (ii)  $(\mathcal{A}F_n) - rF_n \rightarrow (\mathcal{A}F) - rF$  uniformly on compact subsets of  $\mathbb{R}_+ \setminus \mathcal{D}$  as  $n \rightarrow \infty$ ;
- and
- (iii)  $\{(\mathcal{A}F_n) - rF_n\}_{n=1}^\infty$  is locally bounded on  $\mathbb{R}_+$ .

Applying Itô's theorem to the mapping  $(t, x) \mapsto e^{-rt} \Delta_n(x)$ , where  $\Delta_n(x) = F_n(x) - (R_r \pi)(x)$ , taking expectations, and reordering terms yield

$$\begin{aligned} e^{-r\tau_j} \Delta_n(X_{\tau_j}^\nu) &= \mathbf{E}_{\mathcal{F}_{\tau_j}} \left[ e^{-r\tau_{j+1}} \Delta_n(X_{\tau_{j+1}-}^\nu) \right] \\ &\quad - \mathbf{E}_{\mathcal{F}_{\tau_j}} \left[ \int_{\tau_j}^{\tau_{j+1}-} e^{-rs} ((\mathcal{A}F_n)(X_s^\nu) - rF_n(X_s^\nu) + \pi(X_s^\nu)) ds \right]. \end{aligned}$$

Letting  $n \rightarrow \infty$ , applying Fatou's theorem, and invoking the variational inequality  $(\mathcal{A}F)(x) - rF(x) + \pi(x) \leq 0$  then result in inequality (2.8). The alleged result now follows from Lemma 2.1.  $\square$

### 3. Auxiliary results.

**3.1. Some associated functionals.** In subsection 2.2 we made a standing assumption that the upper boundary  $\infty$  is natural and the lower boundary 0 is natural, exit, or killing. It is important to point out that in all of these cases we have that  $\lim_{x \rightarrow 0+} \psi(x) = \lim_{x \rightarrow \infty} \varphi(x) = 0$  (cf. [15, p. 19]). These conditions will be used in this subsection without explicit indication. Recall the definition (2.6) of  $\theta$  and consider the expected cumulative present value  $(R_r \theta)$ . By invoking the representation (2.3), differentiating the equation sidewise, and dividing the resulting identity with  $\psi'(x)$ , we find that

$$(3.1) \quad \frac{(R_r \theta)'(x)}{\psi'(x)} = B^{-1} \frac{\varphi'(x)}{\psi'(x)} \int_0^x \psi(y) \theta(y) m'(y) dy + B^{-1} \int_x^\infty \varphi(y) \theta(y) m'(y) dy.$$

Differentiating (3.1) and noticing that  $\varphi''(x)\psi'(x) - \varphi'(x)\psi''(x) = 2rBS'(x)/\sigma^2(x)$  now yield

$$\frac{d}{dx} \left[ \frac{(R_r \theta)'(x)}{\psi'(x)} \right] = \frac{2S'(x)}{\sigma^2(x)\psi'^2(x)} L(x),$$

where the functional  $L : \mathbb{R}_+ \rightarrow \mathbb{R}$  is defined as

$$L(x) = r \int_0^x \psi(y) \theta(y) m'(y) dy - \theta(x) \frac{\psi'(x)}{S'(x)}.$$

The functional  $L$  will prove to be the principal determinant of the optimal policies in all the considered stochastic control problems. Our first auxiliary result is now summarized in the following.

**LEMMA 3.1.** *Let Assumption 2.1 be satisfied. Then there is a unique state  $\hat{x} = \operatorname{argmin}\{\frac{(R_r \theta)'(x)}{\psi'(x)}\} \in (x^*, \infty)$  satisfying the condition  $L(\hat{x}) = 0$ .*

*Proof.* Assume first that  $x^* < x < z$ . Since the mapping  $\theta$  is decreasing on  $(x^*, \infty)$  and

$$r \int_a^b \psi(y) m'(y) dy = \frac{\psi'(b)}{S'(b)} - \frac{\psi'(a)}{S'(a)}$$

for any  $0 < a < b < \infty$ , we have that

$$\begin{aligned} \frac{1}{r} [L(z) - L(x)] &= \int_x^z \psi(y) \theta(y) m'(y) dy - \frac{\theta(z)}{r} \frac{\psi'(z)}{S'(z)} + \frac{\theta(x)}{r} \frac{\psi'(x)}{S'(x)} \\ &> \frac{\theta(z)}{r} \left[ \frac{\psi'(z)}{S'(z)} - \frac{\psi'(x)}{S'(x)} \right] - \frac{\theta(z)}{r} \frac{\psi'(z)}{S'(z)} + \frac{\theta(x)}{r} \frac{\psi'(x)}{S'(x)} \\ &= \frac{[\theta(x) - \theta(z)]}{r} \frac{\psi'(x)}{S'(x)} > 0, \end{aligned}$$

proving that  $L$  is monotonically increasing on  $(x^*, \infty)$ . Analogously, we find that whenever  $z < x < x^*$

$$\begin{aligned} \frac{1}{r} [L(x) - L(z)] &= \int_z^x \psi(y) \theta(y) m'(y) dy - \frac{\theta(x)}{r} \frac{\psi'(x)}{S'(x)} + \frac{\theta(z)}{r} \frac{\psi'(z)}{S'(z)} \\ &< \frac{[\theta(z) - \theta(x)]}{r} \frac{\psi'(z)}{S'(z)} < 0 \end{aligned}$$

showing that  $L$  is monotonically decreasing on  $(0, x^*)$ .

Since the boundary 0 is assumed to be natural, exit, or killing, we find that  $\lim_{x \rightarrow 0+} L(x) \leq 0$ . Moreover, since  $x^*$  is the global maximum of  $\theta$ , we find that

$$\begin{aligned} L(x^*) &= r \int_0^{x^*} \psi(y) \theta(y) m'(y) dy - \theta(x^*) \frac{\psi'(x^*)}{S'(x^*)} \\ &< \theta(x^*) \left[ \frac{\psi'(x^*)}{S'(x^*)} - \frac{\psi'(0)}{S'(0)} \right] - \theta(x^*) \frac{\psi'(x^*)}{S'(x^*)} \leq 0. \end{aligned}$$

Assumption 2.1 implies that there is a unique state  $x_0 \in (x^*, \infty)$  such that  $\theta(x_0) = 0$  and that  $\theta(x) > 0$  on  $(0, x_0)$ . Hence,

$$L(x_0) = r \int_0^{x_0} \psi(y) \theta(y) m'(y) dy > 0.$$

Since  $L$  is increasing on  $(x^*, \infty)$ , we find there is a unique state  $\hat{x} \in (x^*, \infty)$  such that  $L(\hat{x}) = 0$ .  $\square$

Along with the functional  $L$ , two additional functionals will be important in the subsequent study of the considered stochastic control problems. These functionals, which are denoted as  $I : \mathbb{R}_+ \rightarrow \mathbb{R}$  and  $J : \mathbb{R}_+ \rightarrow \mathbb{R}$ , are defined as

$$(3.2) \quad I(x) = \frac{\beta(R_r \pi)'(x) - \lambda}{\psi'(x)}$$

and as

$$(3.3) \quad J(x) = \beta(R_r \pi)(x) - \lambda x - I(x) \psi(x).$$

Note that  $J'(x) = -\psi(x)I'(x)$  for all  $x \in \mathbb{R}_+$ . The functional  $J$  has an alternative representation which will be used later. More precisely, since

$$\frac{\psi(x)}{\psi'(x)} - x = \frac{S'(x)}{\psi'(x)} \int_0^x \psi(y)\rho(y)m'(y)dy$$

(cf. [7, Lemma 3.3]), we find by using the representation (2.3) for  $(R_r\pi)$  that

$$(3.4) \quad J(x) = \frac{S'(x)}{\psi'(x)} \int_0^x \psi(y)\theta(y)m'(y)dy.$$

We will now demonstrate that the behavior of  $L$  dictates the monotonicity properties of the associated functionals  $I$  and  $J$ . This is accomplished in the following.

LEMMA 3.2. *Let Assumption 2.1 be satisfied. Then  $I'(x) \gtrless 0$  and  $J'(x) \lesseqgtr 0$  when  $x \gtrless \hat{x}$ .*

*Proof.* First note that since  $\rho \in \mathcal{L}_1$ , the expression

$$\frac{d}{dx} \left[ \frac{(R_r\rho)'(x)}{\psi'(x)} \right] = \frac{2S'(x)}{\sigma^2(x)\psi'^2(x)} \left[ r \int_0^x \psi(y)\rho(y)m'(y)dy - \rho(x) \frac{\psi'(x)}{S'(x)} \right]$$

holds. On the other hand, we have that

$$\psi''(x) = \frac{2S'(x)}{\sigma^2(x)} \left[ r \int_0^x \psi(y)\rho(y)m'(y)dy - \rho(x) \frac{\psi'(x)}{S'(x)} \right]$$

(cf. [3, Lemma 2.1]), which in turn implies that

$$\frac{d}{dx} \left[ \frac{(R_r\rho)'(x)}{\psi'(x)} \right] = \frac{\psi''(x)}{\psi'^2(x)}.$$

Combining this observation with the findings of Lemma 3.1 now yields

$$I'(x) = \frac{d}{dx} \left[ \frac{\beta(R_r\pi)'(x) - \lambda}{\psi'(x)} \right] = \frac{d}{dx} \left[ \frac{(R_r\theta)'(x)}{\psi'(x)} \right] = \frac{2S'(x)}{\sigma^2(x)\psi'^2(x)} L(x),$$

from which the first alleged inequality follows. The remaining part of the proof follows now from the equation  $J'(x) = -\psi(x)I'(x)$ .  $\square$

Lemma 3.2 essentially demonstrates that under our Assumption 2.1 the threshold  $\hat{x}$  constitutes the global minimum of the functional  $I(x)$  and the global maximum of the functional  $J(x)$ . To close the subsection, we present a lemma determining the limiting properties of  $I$  and  $J$ .

LEMMA 3.3. *Let Assumption 2.1 be satisfied. Then*

$$\lim_{x \rightarrow 0+} J(x) \geq 0, \lim_{x \rightarrow 0+} I(x) \geq 0, \lim_{x \rightarrow \infty} I(x) \leq 0, \text{ and } \lim_{x \rightarrow \infty} J(x) = -\infty.$$

*Proof.* If the origin is attainable, then the representation (3.4) implies that  $J(0) = 0$ , since  $\lim_{x \rightarrow 0+} \frac{\psi'(x)}{S'(x)} > 0$ . On the other hand, if the origin is unattainable, then  $\lim_{x \rightarrow 0+} \frac{\psi'(x)}{S'(x)} = 0$ . In this case l'Hôpital's rule yields

$$\lim_{x \rightarrow 0+} J(x) = \lim_{x \rightarrow 0+} \frac{\psi(x)\theta(x)m'(x)}{\frac{d}{dx} \left[ \frac{\psi'(x)}{S'(x)} \right]} = \lim_{x \rightarrow 0+} \frac{\psi(x)\theta(x)m'(x)}{r\psi(x)m'(x)} = \lim_{x \rightarrow 0+} \frac{\theta(x)}{r} \geq 0.$$

To prove the alleged behavior of  $J$  at infinity, note that

$$\lim_{x \rightarrow \infty} \int_0^x \psi(y) \theta(y) m'(y) dy = -\infty$$

and that  $\lim_{x \rightarrow \infty} \frac{\psi'(x)}{S'(x)} = \infty$ . Thus  $\lim_{x \rightarrow \infty} J(x) = \lim_{x \rightarrow \infty} \frac{\theta(x)}{r} = -\infty$ .

We showed in Lemma 3.1 that the state  $\hat{x} = \operatorname{argmax}\{J(x)\} = \operatorname{argmin}\{I(x)\}$  lies in the interval  $(x^*, \infty)$ , i.e., where  $\theta$  is monotonically decreasing. Hence for  $x > \hat{x}$ , we have that

$$\begin{aligned} I(x) &= B^{-1} \frac{\varphi'(x)}{\psi'(x)} \left[ \int_0^{\hat{x}} \psi(y) \theta(y) m'(y) dy + \int_{\hat{x}}^x \psi(y) \theta(y) m'(y) dy \right] \\ &\quad + B^{-1} \int_x^\infty \varphi(y) \theta(y) m'(y) dy \\ &= B^{-1} \frac{\varphi'(x)}{\psi'(x)} \left[ \frac{\theta(\hat{x})}{r} \frac{\psi'(\hat{x})}{S'(\hat{x})} + \int_{\hat{x}}^x \psi(y) \theta(y) m'(y) dy \right] + B^{-1} \int_x^\infty \varphi(y) \theta(y) m'(y) dy \\ &\leq B^{-1} \frac{\varphi'(x)}{\psi'(x)} \left[ \frac{\theta(\hat{x})}{r} \frac{\psi'(\hat{x})}{S'(\hat{x})} + \frac{\theta(\hat{x})}{r} \left( \frac{\psi'(x)}{S'(x)} - \frac{\psi'(\hat{x})}{S'(\hat{x})} \right) \right] + B^{-1} \int_x^\infty \varphi(y) \theta(y) m'(y) dy \\ &= B^{-1} \frac{\varphi'(x)}{S'(x)} \frac{\theta(\hat{x})}{r} + B^{-1} \int_x^\infty \varphi(y) \theta(y) m'(y) dy. \end{aligned}$$

By letting  $x$  tend to infinity in the inequality above, we discover that  $\lim_{x \rightarrow \infty} I(x) \leq 0$ , since  $\lim_{x \rightarrow \infty} \frac{\varphi'(x)}{S'(x)} = 0$ . The property  $\lim_{x \rightarrow 0+} I(x) \geq 0$  is still left to prove. To prove this, observe that the condition  $I(x) \geq I(\hat{x})$  implies that

$$\lim_{x \rightarrow 0+} \frac{(R_r \theta)'(x)}{S'(x)} = \lim_{x \rightarrow 0+} I(x) \frac{\psi'(x)}{S'(x)} \geq \lim_{x \rightarrow 0+} I(\hat{x}) \frac{\psi'(x)}{S'(x)} = 0;$$

hence  $\lim_{x \rightarrow 0+} (R_r \theta)'(x) \geq 0$ . Now the desired result

$$\lim_{x \rightarrow 0+} I(x) = \lim_{x \rightarrow 0+} \frac{(R_r \theta)'(x)}{\psi'(x)} \geq 0$$

follows, since  $\psi'(x) > 0$ .  $\square$

**3.2. The associated singular control problem.** Before proceeding to the analysis of the stochastic impulse control problem, we first consider an associated singular stochastic control problem. To this end, consider the associated controlled diffusion process  $X^Z$  on  $\mathbb{R}_+$  given by the generalized Itô stochastic differential equation

$$(3.5) \quad dX_t^Z = \mu(X_t^Z) dt + \sigma(X_t^Z) dW_t - \beta dZ_t, \quad X_0^Z = x,$$

where the process  $Z_t$  is an *admissible control*, meaning a nonnegative, nondecreasing, right-continuous, and  $\{\mathcal{F}_t\}$ -adapted process. We denote the class of such processes as  $\Lambda$  and assume that  $\mu$  and  $\sigma$  satisfy the same regularity conditions as in the impulse control case. Given these assumptions, we will consider the associated singular control problem

$$(3.6) \quad K(x) = \sup_{Z \in \Lambda} \mathbf{E}_x \left[ \int_0^{H_0^Z} e^{-rs} (\pi(X_s^Z) ds + \lambda dZ_s) \right],$$

where  $H_0^Z = \inf\{t \geq 0 : X_t^Z \leq 0\}$  denotes the first exit time of the controlled diffusion  $X^Z$  from  $\mathbb{R}_+$ . It is important to emphasize that since the value accumulates only up to the first exit time  $H_0^Z$ , we are actually considering only such controls  $Z$  that keep the process  $X^Z$  positive. Moreover, it is worth observing that applying the generalized Itô theorem to the linear mapping  $x \mapsto \lambda x/\beta$  yields

$$\mathbf{E}_x \int_0^{\tau_N} e^{-rs} \lambda dZ_s = \frac{\lambda x}{\beta} + \mathbf{E}_x \int_0^{\tau_N} e^{-rs} \frac{\lambda}{\beta} \rho(X_s^Z) ds - \mathbf{E}_x \left[ e^{-r\tau_N} \frac{\lambda}{\beta} X_{\tau_N}^Z \right],$$

where  $\rho(x) = \mu(x) - rx$  and  $\tau_N = H_0^Z \wedge N \wedge \inf\{t \geq 0 : X_t^Z \geq N\}$  is an almost surely finite stopping time. The nonnegativity of the controlled process then results by letting  $N$  tend to infinity and invoking monotone convergence to the inequality

$$(3.7) \quad K(x) \leq \beta^{-1} \left[ \lambda x + \sup_{Z \in \Lambda} \mathbf{E}_x \int_0^{H_0^Z} e^{-rs} \theta(X_s^Z) ds \right].$$

It is clear that if the implemented admissible policy is such that

$$\lim_{N \rightarrow \infty} \mathbf{E}_x [e^{-r\tau_N} X_{\tau_N}^Z] = 0,$$

then the inequality (3.7) becomes an equality. In that case the value of the optimal policy can be decomposed into a part measuring the value of the instantaneous liquidation policy and the expected cumulative present value of the excess return accrued from following the optimal policy and postponing the immediate liquidation of the underlying process.

The Hamilton–Jacobi–Bellman equation for (3.6) can be written as

$$(3.8) \quad \max \{ \mathcal{A}K(x) - rK(x) + \pi(x), K'(x) - \lambda/\beta \} = 0$$

for all  $x \in \mathbb{R}_+$ . In the next lemma we will establish the value and the optimal policy for the problem (3.6). These results will later turn out to be useful in the analysis of the impulse control problem (2.5) as well.

LEMMA 3.4. *Let Assumption 2.1 be satisfied. Then the optimal singular stochastic control exists and is given by*

$$(3.9) \quad Z_t^* = \begin{cases} (x - \hat{x})^+, & t = 0, \\ \mathcal{L}(t, \hat{x}), & t > 0, \end{cases}$$

where the threshold  $\hat{x} \in (x^*, \infty)$  is the unique root of the first order condition  $L(\hat{x}) = 0$  and  $\mathcal{L}(t, \hat{x})$  is the local time of  $X$  at  $\hat{x}$  (cf. [15, pp. 21–24]). Moreover, the value function  $K$  reads as

$$(3.10) \quad K(x) = \begin{cases} \beta^{-1} \left( \lambda x + \frac{\theta(\hat{x})}{r} \right), & x \geq \hat{x}, \\ (R_r \pi)(x) - \beta^{-1} I(\hat{x}) \psi(x), & x < \hat{x}, \end{cases}$$

implying that the marginal value  $K'$  can be expressed as

$$(3.11) \quad \begin{aligned} K'(x) &= (R_r \pi)'(x) + \beta^{-1} \psi'(x) \sup_{y \geq x} \left[ \frac{\lambda - \beta (R_r \pi)'(y)}{\psi'(y)} \right] \\ &= \begin{cases} \lambda \beta^{-1}, & x \geq \hat{x}, \\ (R_r \pi)'(x) - \beta^{-1} I(\hat{x}) \psi'(x), & x < \hat{x}. \end{cases} \end{aligned}$$



The value function  $K$  satisfies also the smooth pasting condition  $\lim_{x \rightarrow \hat{x}} K''(x) = 0$ .

*Proof.* Denote the function defined in (3.10) as  $\hat{K}(x)$ . We will now demonstrate that  $K(x) = \hat{K}(x)$  for all  $x \in \mathbb{R}_+$ . Since  $\hat{K}(x)$  is attained by the admissible local time push policy (3.9) (i.e., reflection at  $\hat{x}$ ), it is clear that  $\hat{K}(x) \leq K(x)$  for all  $x \in \mathbb{R}_+$  (cf. section 1.6 in [19]). In order to establish the opposite inequality, we first observe by ordinary differentiation that

$$\hat{K}'(x) = \begin{cases} \lambda\beta^{-1}, & x \geq \hat{x}, \\ \beta^{-1} [\lambda + \psi'(x) (I(x) - I(\hat{x}))], & x < \hat{x}. \end{cases}$$

Since the state  $\hat{x}$  is the global minimum of the functional  $I$ , we find that  $\hat{K}'(x) \geq \lambda\beta^{-1}$  for all  $x \in \mathbb{R}_+$ . Moreover,

$$(\mathcal{A}\hat{K})(x) - r\hat{K}(x) + \pi(x) = \begin{cases} \beta^{-1} (\theta(x) - \theta(\hat{x})), & x \geq \hat{x}, \\ 0, & x < \hat{x}. \end{cases}$$

Since the state  $\hat{x}$  is on the set where  $\theta$  is strictly decreasing, we find that  $(\mathcal{A}\hat{K})(x) - r\hat{K}(x) + \pi(x) \leq 0$  for all  $x \in \mathbb{R}_+$ . Finally, since  $\hat{x}$  minimizes  $I$ , first order optimality conditions imply that  $\psi''(\hat{x})\beta(R_r\pi)''(\hat{x}) = \psi''(\hat{x})(\beta(R_r\pi)'(\hat{x}) - \lambda)$ . Therefore,  $\lim_{x \rightarrow \hat{x}} \hat{K}''(x) = 0$ , implying that  $\hat{K} \in C^2(\mathbb{R}_+)$ . The function  $K$  now satisfies the conditions of Lemma 1 in [1]. Thus  $\hat{K}(x) \geq K(x)$  for all  $x \in \mathbb{R}_+$ .  $\square$

Lemma 3.4 demonstrates that under Assumption 2.1 the associated singular control problem (3.6) is solvable and that the optimal value is attained by utilizing a local time push policy at the threshold  $\hat{x}$  (cf. [20]). The optimal control policy  $Z_t^*$  does not exhibit jumps at time  $H_0^Z$  and, therefore, does not produce interior singularities into the value function  $K$  (see (3.6)). Moreover, Lemma 3.4 also shows that the smooth pasting property holds for (3.6), and we notice from the proof that it is an implication of our approach to the problem. In particular, Lemma 3.4 shows that the value function  $K$  defined in (3.10) is a  $C^2$ -solution of the Hamilton–Jacobi–Bellman equation (3.8). A set of interesting comparative static results implied by Lemma 3.4 are now summarized in the following.

**COROLLARY 3.5.** *Let Assumption 2.1 be satisfied. Then the following hold.*

- (i) *The value  $K$  and the marginal value  $K'$  of the optimal policy are decreasing functions of the parameter  $\beta$ .*
- (ii) *The value  $K$  and the marginal value  $K'$  of the optimal policy are increasing functions of the parameter  $\lambda$ .*
- (iii) *The optimal exercise threshold  $\hat{x}$  is an increasing mapping of the parameter  $\beta$  and a decreasing mapping of the parameter  $\lambda$ .*
- (iv) *If  $\lambda = \beta$ , then the optimal exercise threshold  $\hat{x}$  is independent of  $\lambda$  and  $\beta$ .*

*Proof.* (i) Denote the value associated with the parameter  $\beta_i$  as  $K_i$ ,  $i = 1, 2$ . It is now clear from the proof of Lemma 3.4 that  $K_2$  satisfies the sufficient variational inequalities  $(\mathcal{A}K_2)(x) - rK_2(x) + \pi(x) \leq 0$  and  $K_2'(x) \geq \lambda/\beta_2 > \lambda/\beta_1$  for all  $x \in \mathbb{R}_+$ . Hence,  $K_2(x) \geq K_1(x)$  for all  $x \in \mathbb{R}_+$ . In order to establish that  $K_2'(x) \geq K_1'(x)$  for all  $x \in \mathbb{R}_+$ , we observe that the mapping  $x \mapsto (\lambda - \beta(R_r\pi)'(x))/(\beta\psi'(x))$  is a decreasing function of the parameter  $\beta$  from which the alleged result follows by invoking the representation (3.11). Proving part (ii) is entirely analogous. It remains to consider the sensitivity of the optimal exercise threshold  $\hat{x}$  with respect to parametric changes. To this end, consider the mapping

$$\bar{L}(x, \lambda, \beta) = r \int_0^x \psi(y)(\beta\pi(y) + \lambda\rho(y))m'(y)dy - (\beta\pi(x) + \lambda\rho(x)) \frac{\psi'(x)}{S'(x)}.$$

If  $\beta_1 > \beta_2$ , then

$$\bar{L}(x, \lambda, \beta_1) - \bar{L}(x, \lambda, \beta_2) = (\beta_1 - \beta_2) \left[ r \int_0^x \psi(y) \pi(y) m'(y) dy - \pi(x) \frac{\psi'(x)}{S'(x)} \right] \leq 0$$

since

$$r \int_0^x \psi(y) \pi(y) m'(y) dy - \pi(x) \frac{\psi'(x)}{S'(x)} \leq -\pi(x) \frac{\psi'(0)}{S'(0)} \leq 0$$

by the assumed monotonicity and nonnegativity of  $\pi$ . Therefore, if  $\hat{x}_i$  denotes the optimal exercise threshold associated with  $\beta_i$ ,  $i = 1, 2$ , we observe that  $0 = \bar{L}(\hat{x}_1, \lambda, \beta_1) \leq \bar{L}(\hat{x}_1, \lambda, \beta_2)$ , which, in turn, implies that  $\hat{x}_1 \geq \hat{x}_2$ . Establishing that  $\hat{x}$  is a decreasing mapping of the parameter  $\lambda$  is entirely analogous. Finally, if  $\lambda = \beta$ , then

$$L(x) = \beta \left[ r \int_0^x \psi(y) (\pi(y) + \rho(y)) m'(y) dy - (\pi(x) + \rho(x)) \frac{\psi'(x)}{S'(x)} \right],$$

from which the alleged result follows.  $\square$

Corollary 3.5 characterizes the impact of parametric changes on the value, the marginal value, and the optimal exercise threshold of the irreversible policy. We observe that an increase in  $\beta$  decreases both the value and the marginal value of the optimal policy and, therefore, postpones exercise by increasing the optimal exercise threshold and, therefore, expanding the continuation region where waiting is optimal. The contrary happens when the parameter  $\lambda$  increases. An interesting implication of these comparative static results is that parametric changes are neutral (i.e., do not affect the optimal exercise threshold  $\hat{x}$ ) as long as the ratio  $\lambda/\beta$  is held constant. As usual in models considering cash flow management in the presence of taxation, Corollary 3.5 shows that when  $\lambda = \beta$  the optimal policy is independent of the parameters  $\lambda$  and  $\beta$  (i.e., the *harmonization of tax rates implies the tax neutrality of the optimal policy*).

**3.3. The associated optimal stopping problem.** In this subsection we consider another associated control problem, namely, an optimal stopping problem. Let  $X$  be the diffusion evolving on  $\mathbb{R}_+$  according to the ordinary Itô stochastic differential equation (2.2), and assume that the infinitesimal coefficients  $\mu$  and  $\sigma$  satisfy the same regularity conditions as in the impulse control case. Given these assumptions, consider the corresponding optimal stopping problem

$$(3.12) \quad G_c(x) = \sup_{\tau < H_0} \mathbf{E}_x \left[ \int_0^\tau e^{-rs} \pi(X_s) ds + e^{-r\tau} (\lambda \beta^{-1} X_\tau - c) \right],$$

where  $c \geq 0$  is an arbitrary constant and  $\tau$  is an arbitrary  $\mathcal{F}_t$ -stopping time satisfying the constraint  $\tau < H_0$  stating that the stopping time problem is defined up to the first exit time from  $\mathbb{R}_+$ . Following the reasoning of (3.7), we find by applying Dynkin's theorem to the mapping  $x \mapsto \lambda x/\beta - c$  (or by applying Itô's theorem to the process  $t \mapsto e^{-rt}(\lambda \beta^{-1} X_t - c)$ ) that

$$G_c(x) = \frac{\lambda}{\beta} x - c + \frac{1}{\beta} \sup_{\tau < H_0} \mathbf{E}_x \int_0^\tau e^{-rs} (\theta(X_s) + \beta c r) ds,$$

demonstrating how the value of the optimal policy can in this case be decomposed into the sum of the exercise payoff and the early exercise premium. Our main findings on this associated stopping problem are now summarized in the following.

LEMMA 3.6. *Let Assumption 2.1 be satisfied. Then an optimal stopping policy is to stop at the Markov time  $\tau_{\bar{x}_c} = \inf\{t \geq 0 : X_t \geq \bar{x}_c\}$ , where  $\bar{x}_c$ , denoting the optimal stopping threshold, is the unique root of the equation  $J(\bar{x}_c) = -\beta c$ , where  $J$  is as defined in (3.3). Moreover, the value can be written as*

$$(3.13) \quad \begin{aligned} G_c(x) &= (R_r\pi)(x) + \beta^{-1}\psi(x) \sup_{y \geq x} \left[ \frac{\lambda y - \beta(R_r\pi)(y) - \beta c}{\psi(y)} \right] \\ &= \begin{cases} \beta^{-1}\lambda x - c, & x \geq \bar{x}_c, \\ (R_r\pi)(x) - \beta^{-1}I(\bar{x}_c)\psi(x), & x < \bar{x}_c, \end{cases} \end{aligned}$$

and it satisfies the smooth-pasting condition  $\lim_{x \rightarrow \bar{x}_c-} G'_c(x) = \beta^{-1}\lambda$ .

*Proof.* In order to establish (3.13), denote as  $x_0$  the unique interior state at which  $\theta(x_0) = 0$ . The expression (3.4) implies that

$$\frac{d}{dx} \left[ \frac{\psi'(x)}{S'(x)} J(x) \right] = \psi(x)\theta(x)m'(x) \geq 0, \quad x \leq x_0.$$

Thus  $J(x) > 0$  for all  $x \in (0, x_0)$ , since  $\lim_{x \rightarrow 0+} \frac{\psi'(x)}{S'(x)} J(x) \geq 0$ . On the other hand, we proved in Lemma 3.3 that  $\lim_{x \rightarrow \infty} J(x) = -\infty$ . Together with the monotonicity properties of  $J$ , this implies that there is a unique state  $\bar{x}_c \in \theta^{-1}(\mathbb{R}_-)$  at which the condition  $J(\bar{x}_c) = -\beta c$  is satisfied. Since

$$\frac{d}{dx} \left[ \frac{\lambda x - \beta(R_r\pi)(x) - \beta c}{\psi(x)} \right] = \frac{\psi'(x)}{\psi^2(x)} (J(x) + \beta c),$$

we find that  $\bar{x}_c = \operatorname{argmax} \{(\lambda x - \beta(R_r\pi)(x) - \beta c)/\psi(x)\}$ . The first order optimality condition now implies that

$$\frac{\lambda \bar{x}_c - \beta(R_r\pi)(\bar{x}_c) - \beta c}{\psi(\bar{x}_c)} = \frac{\lambda - \beta(R_r\pi)'(\bar{x}_c)}{\psi'(\bar{x}_c)} = -I(\bar{x}_c),$$

which gives us the expression (3.13) and proves the smooth-pasting condition

$$\lim_{x \rightarrow \bar{x}_c} G'_c(x) = \beta^{-1}\lambda.$$

Given these observations, denote the function defined in (3.13) as  $\hat{G}_c(x)$ . Since

$$\hat{G}_c(x) = \mathbf{E}_x \left[ \int_0^{\tau_{\bar{x}_c}} e^{-rs} \pi(X_s) ds + e^{-r\tau_{\bar{x}_c}} (\lambda \beta^{-1} X_{\tau_{\bar{x}_c}} - c) \right],$$

where  $\tau_{\bar{x}_c} = \inf\{t \geq 0 : X_t \geq \bar{x}_c\}$ , we find that  $\hat{G}_c(x) \leq G_c(x)$  for all  $x \in \mathbb{R}_+$ . On the other hand, we also observe that  $\hat{G}_c$  is continuously differentiable on  $\mathbb{R}_+$ , is twice continuously differentiable on  $\mathbb{R}_+ \setminus \{\bar{x}_c\}$ , satisfies the inequalities  $|\hat{G}_c''(\bar{x}_c \pm)| < \infty$ , and satisfies the variational inequality  $\min\{r\hat{G}_c(x) - (\mathcal{A}\hat{G}_c)(x) - \pi(x), \hat{G}_c(x) - \lambda\beta^{-1}x + c\} = 0$ . Thus  $\hat{G}_c(x) \geq G_c(x)$  for all  $x \in \mathbb{R}_+$ .  $\square$

Lemma 3.6 establishes that under Assumption 2.1 the optimal stopping problem (3.12) is solvable and that an optimal stopping policy is a threshold policy requiring that the underlying process should be stopped once it hits the constant boundary  $\bar{x}_c$  at which the expected present value of the exercise payoff is maximized. At the optimal exercise threshold  $\bar{x}_c$  the standard balance identity holds and the value of

the project  $\bar{x}_c$  coincides with the sum of the sunk cost  $c$  and the lost option value  $G_c(\bar{x}_c)$ . Moreover, following the reasoning of our findings on the associated singular stochastic control problem, we find that the smooth-pasting principle holds for the problem (3.12) as well. We also note that  $G_c''(\bar{x}_c-) \neq 0$ ; in other words,  $G_c$  is not twice continuously differentiable over the boundary  $\bar{x}_c$ . A set of interesting comparative static results implied by Lemma 3.6 are now summarized in the following.

**COROLLARY 3.7.** *Let Assumption 2.1 be satisfied. Then the following hold.*

- (i) *The value  $G_c$  is a decreasing function of both the parameter  $\beta$  and the sunk cost  $c$  and an increasing function of the parameter  $\lambda$ .*
- (ii) *The optimal exercise threshold  $\bar{x}_c$  is an increasing mapping of both the parameter  $\beta$  and the sunk cost  $c$  and a decreasing mapping of the parameter  $\lambda$ .*
- (iii) *If  $\lambda = \beta$ , then the optimal exercise threshold  $\bar{x}_c$  is independent of  $\lambda$  and  $\beta$ .*

*Proof.* The claim of part (i) of our corollary follows directly from the definition of the exercise payoff. Thus, it is sufficient to consider the sensitivity of the optimal threshold  $\bar{x}_c$  to changes in  $\lambda$ ,  $\beta$ , or  $c$ . To this end, consider the mapping

$$\tilde{L}(x, \lambda, \beta, c) = \int_0^x \psi(y)(\beta\pi(y) + \lambda\rho(y))m'(y)dy + \beta c \frac{\psi'(x)}{S'(x)}$$

and denote as  $\bar{x}_c(\beta_i)$  the optimal exercise threshold associated with the parameter  $\beta_i$ . If  $\beta_1 > \beta_2$ , then

$$\tilde{L}(x, \lambda, \beta_1, c) - \tilde{L}(x, \lambda, \beta_2, c) = (\beta_1 - \beta_2) \left[ \int_0^x \psi(y)\pi(y)m'(y)dy + c \frac{\psi'(x)}{S'(x)} \right] > 0,$$

which implies that  $0 = \tilde{L}(\bar{x}_c(\beta_1), \lambda, \beta_1, c) > \tilde{L}(\bar{x}_c(\beta_1), \lambda, \beta_2, c)$  and, therefore, that  $\bar{x}_c(\beta_1) > \bar{x}_c(\beta_2)$ . The analysis of the impact of changes in either  $\lambda$  or  $c$  on the optimal exercise threshold is entirely analogous. Finally, if  $\lambda = \beta$ , then

$$\bar{x}_c = \operatorname{argmax} \{ (x - (R_r\pi)(x) - c)/\psi(x) \},$$

from which the alleged result follows.  $\square$

Corollary 3.7 extends the findings of Corollary 3.5 to the present example. More precisely, we observe that an increase in  $\beta$  postpones rational exercise by expanding the continuation region where stopping is suboptimal. The opposite is shown to happen when  $\lambda$  increases. Interestingly, we again find that parametric changes are neutral (i.e., do not affect the optimal exercise threshold  $\bar{x}_c$ ) as long as the ratio  $\lambda/\beta$  is held constant. Following the reasoning of our findings on the associated singular control problem, we again find that if  $\lambda = \beta$ , the optimal exercise strategy is independent of the parameters  $\lambda$  and  $\beta$  (i.e., *harmonization results in neutrality*). Moreover, as is intuitively clear, our findings indicate that increased sunk costs decrease the value and postpone rational exercise by expanding the continuation region.

#### 4. Optimal impulse control policy.

**4.1. Necessary conditions.** The stochastic impulse control problems of type (2.5) are typically tackled by relying on a combination of the classical Hamilton–Jacobi–Bellman approach and quasi-variational inequalities. In this study, we plan to adopt an alternative approach which results in more easily interpretable conditions characterizing a potentially optimal policy. Instead of considering all the admissible impulse controls at once, we restrict our attention to the subclass  $\nu_{(\zeta, y)}$  of impulse

controls characterized by the sequence of intervention times  $\tau_0^y = 0$ ,  $\tau_k^y = \inf\{t \geq \tau_{k-1}^y : X_t^{\nu(\zeta, y)} \geq y\}$  and the sequence of interventions  $\zeta_k^y = \zeta + (x - y)^+$  for all  $k \geq 1$ . That is, we restrict our attention to control policies consisting of sequence of constant-sized impulses (with the exception of the initial impulse, which depends on the initial state) exerted every time the underlying diffusion hits a predetermined, constant exercise threshold  $y$ . Given this class of admissible impulse controls, define the value  $F_c : \mathbb{R}_+ \rightarrow \mathbb{R}$  accrued from applying the impulse control  $\nu_{(\zeta, y)}$  as  $F_c(x) = J_c^{(\zeta, y)}(x)$ . Since  $X_{\tau_k^y+}^{\nu(\zeta, y)} = X_{\tau_k^y-}^{\nu(\zeta, y)} - \beta\zeta = y - \beta\zeta$  for all  $k$  and the controlled diffusion evolves as the linear diffusion  $X$  between any two successive intervention dates, we observe that for all  $x < y$  the value satisfies the functional relation (a so-called *running present value formulation*)

$$(4.1) \quad F_c(x) = \mathbf{E}_x \left[ \int_0^{\tau_y} e^{-rs} \pi(X_s) ds + e^{-r\tau_y} (\lambda(X_{\tau_y} - (y - \zeta)) - c + F_c(y - \beta\zeta)) \right],$$

where  $\tau_y = \inf\{t \geq 0 : X_t \geq y\}$ . Invoking the strong Markov property of diffusions now implies that the value  $F_c(x)$  can be represented as

$$(4.2) \quad F_c(x) = \begin{cases} F_c(y - \beta\zeta) + \lambda(x - y + \zeta) - c, & x \geq y, \\ (R_r\pi)(x) + (\lambda\zeta - c - (R_r\pi)(y) + F_c(y - \beta\zeta)) \frac{\psi(x)}{\psi(y)}, & x < y. \end{cases}$$

First, note that letting  $x$  tend to  $y$  in (4.2) yields the *value-matching condition*  $F_c(y) = F_c(y - \beta\zeta) + \lambda\zeta - c$ , which can be re-expressed in the more familiar form  $F_c(y - \beta\zeta) + \lambda\zeta = F_c(y) + c$  stating that *the value of the investment opportunity has to coincide with its full costs (lost option value + sunk cost)*. On the other hand, letting  $x$  tend to  $y - \beta\zeta$  yields

$$(4.3) \quad F_c(y - \beta\zeta) = \frac{\psi(y)(R_r\pi)(y - \beta\zeta) + [\lambda\zeta - c - (R_r\pi)(y)]\psi(y - \beta\zeta)}{\psi(y) - \psi(y - \beta\zeta)}.$$

Now, inserting (4.3) into (4.2) implies that the value can be expressed as

$$(4.4) \quad F_c(x) = \begin{cases} (R_r\pi)(y - \beta\zeta) + h(\zeta, y)\psi(y - \beta\zeta) + \lambda(x - y + \zeta) - c, & x \geq y, \\ (R_r\pi)(x) + h(\zeta, y)\psi(x), & x < y, \end{cases}$$

where the mapping  $h : \mathbb{R}_+^2 \rightarrow \mathbb{R}$  is defined as

$$(4.5) \quad h(\zeta, y) = \frac{(R_r\pi)(y - \beta\zeta) - (R_r\pi)(y) + \lambda\zeta - c}{\psi(y) - \psi(y - \beta\zeta)}.$$

In order to prove the existence and uniqueness of the optimal impulse control policy, we will consider the constrained nonlinear programming problem

$$(4.6) \quad \sup_{\substack{\beta\zeta \in [0, y], \\ y \in \mathbb{R}_+}} \frac{(R_r\pi)(y - \beta\zeta) - (R_r\pi)(y) + \lambda\zeta - c}{\psi(y) - \psi(y - \beta\zeta)}.$$

To ease the subsequent analysis, introduce a linear change of variables  $z := y - \beta\zeta$ . Thus  $\zeta = \beta^{-1}(y - z)$ . Since the parameter  $\beta$  is assumed to be positive, the programming problem (4.6) can now be rewritten as

$$(4.7) \quad \sup_{\substack{z \in [0, y], \\ y \in \mathbb{R}_+}} \frac{(R_r\pi)(z) - (R_r\pi)(y) + \lambda\beta^{-1}(y - z) - c}{\psi(y) - \psi(z)}.$$

If an interior pair maximizing the mapping  $h$  exists, denote the associated mapping of the form (4.1) as  $F_c^*$ . More precisely, if an interior pair  $(z_c^*, y_c^*)$  satisfying the problem (4.7) exists, define the mapping  $F_c^* : \mathbb{R} \rightarrow \mathbb{R}$  as

$$(4.8) \quad F_c^*(x) = \begin{cases} (R_r\pi)(z_c^*) + h(z_c^*, y_c^*)\psi(z_c^*) + \gamma(x) - c, & x \geq y_c^*, \\ (R_r\pi)(x) + h(z_c^*, y_c^*)\psi(x), & x < y_c^*, \end{cases}$$

where  $\gamma(x) = \lambda(x - y_c^* - \beta^{-1}(y_c^* - z_c^*))$ . Since  $h$  is differentiable, it is clear that if an interior pair  $(z_c^*, y_c^*)$  satisfying the problem (4.7) exists, then this pair satisfies the ordinary necessary first order conditions  $\frac{\partial h}{\partial z}(z_c^*, y_c^*) = \frac{\partial h}{\partial y}(z_c^*, y_c^*) = 0$ . More precisely, if an optimal pair exists, it must satisfy the conditions

$$(4.9) \quad \begin{cases} (\psi(y_c^*) - \psi(z_c^*)) (\lambda\beta^{-1} - (R_r\pi)'(y_c^*)) = r(z_c^*, y_c^*)\psi'(y_c^*), \\ (\psi(y_c^*) - \psi(z_c^*)) (\lambda\beta^{-1} - (R_r\pi)'(z_c^*)) = r(z_c^*, y_c^*)\psi'(z_c^*), \end{cases}$$

where  $r(z, y) = (R_r\pi)(z) - (R_r\pi)(y) + \lambda\beta^{-1}(y - z) - c$ . This immediately yields the condition

$$\frac{\lambda - \beta(R_r\pi)'(z_c^*)}{\psi'(z_c^*)} = \frac{\lambda - \beta(R_r\pi)'(y_c^*)}{\psi'(y_c^*)}.$$

Using the definition (3.2), this can be rewritten as

$$(4.10) \quad I(y_c^*) - I(z_c^*) = 0.$$

On the other hand, since

$$\frac{\psi'(z_c^*)}{\psi(y_c^*) - \psi(z_c^*)} = \frac{(R_r\pi)'(z_c^*) - \lambda\beta^{-1}}{r(z_c^*, y_c^*)},$$

we find by invoking condition (4.10) and reordering the terms that

$$[\beta(R_r\pi)(y_c^*) - I(y_c^*)\psi(y_c^*) - \lambda y_c^*] - [\beta(R_r\pi)(z_c^*) - I(z_c^*)\psi(z_c^*) - \lambda z_c^*] = -\beta c.$$

Using the definition (3.3), this can be expressed as

$$(4.11) \quad J(y_c^*) - J(z_c^*) = -\beta c.$$

Conditions (4.10) and (4.11) are standard necessary first order conditions for the existence of the solution of the problem (4.7). In a recent paper [10], a similar idea is used to solve another impulse control problem, where first order optimality conditions are derived for an associated functional reminiscent of (4.8) (see [10, equation 2.17]). However, our control problem differs from the problem of [10] on a fundamental level (we will comment on this in the next subsection). Moreover, the actual solution methods are also different.

**4.2. Existence and sufficiency.** Having the necessary conditions (4.10) and (4.11) at our disposal, we will now study their solvability under Assumption 2.1.

**LEMMA 4.1.** *Let Assumption 2.1 be satisfied. Then there exists a unique interior pair  $(z_c^*, y_c^*)$  for which the necessary conditions (4.10) and (4.11) are satisfied.*

*Proof. Existence.* Define the mappings  $\tilde{J} : (0, \hat{x}) \rightarrow (\tilde{J}(0), \tilde{J}(\hat{x}))$  and  $\hat{J} : [\hat{x}, \infty) \rightarrow (-\infty, \hat{J}(\hat{x}))$  as restrictions of the mapping  $J$  and the mapping  $k : \mathbb{R} \rightarrow \mathbb{R}$  as

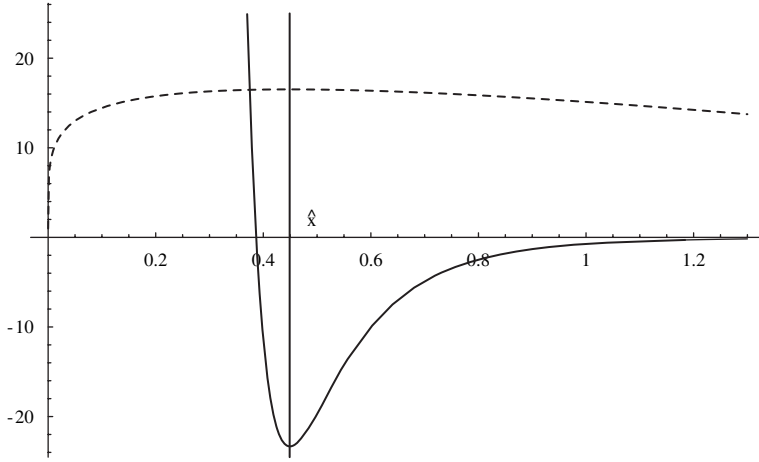


FIG. 4.1. A prototype figure of the functions  $I$  (solid curve) and  $J$  (dashed curve). The vertical solid line marks the state  $\hat{x}$ .

$k(x) = x - \beta c$ . It is clear that both  $\check{J}$  and  $\hat{J}$  are continuous and monotonic. Define now the mapping  $\hat{y} : (0, \hat{x}) \rightarrow (\hat{y}(\hat{x}), \hat{y}(0))$  as  $\hat{y}(x) = (\hat{J}^{-1} \circ k \circ \check{J})(x)$ . By the definitions of  $\check{J}$ ,  $\hat{J}$ , and  $k$ , we observe that  $\hat{y}$  is well defined. Moreover, the mapping  $\hat{y}$  is continuous as a composition of continuous mappings. Finally, since  $\hat{y}(z) = \hat{J}^{-1}(\check{J}(z) - \beta c)$ , we find that the equation  $J(\hat{y}(z)) - J(z) = -\beta c$  holds for all  $z \in (0, \hat{x})$ .

Analogously, define the mappings  $\check{I} : (0, \hat{x}) \rightarrow (\check{I}(\hat{x}), \check{I}(0))$  and  $\hat{I} : [\hat{x}, \infty) \rightarrow [\hat{I}(\hat{x}), \hat{I}(0))$  as restrictions of the mapping  $I$ . It is clear that both  $\check{I}$  and  $\hat{I}$  are continuous and monotonic. Define the mapping  $Y : (0, \hat{x}) \rightarrow (\check{I}^{-1}(\hat{I}(\hat{y}(\hat{x}))), \check{I}^{-1}(\hat{I}(\hat{y}(0))))$  as  $Y(x) = (\check{I}^{-1} \circ \hat{I} \circ \hat{y})(x)$ . Since  $0 < \beta c < \infty$  and  $\hat{J}$  is decreasing, we find that  $\hat{y}(\hat{x}) = \hat{J}^{-1}(\check{J}(\hat{x}) - \beta c) > \hat{J}^{-1}(\check{J}(\hat{x})) = \hat{x}$  and  $\hat{y}(0) = \hat{J}^{-1}(\check{J}(0) - \beta c) \leq \hat{J}^{-1}(-\beta c) < \infty$ . First, these inequalities together with monotonicity of  $\check{I}$  and  $\hat{I}$  guarantee that  $Y$  is well defined. Moreover, coupled with Lemma 3.3 they imply that

$$\left( \check{I}^{-1}(\hat{I}(\hat{y}(\hat{x}))), \check{I}^{-1}(\hat{I}(\hat{y}(0))) \right) \subsetneq \left( \check{I}^{-1}(\hat{I}(\hat{x})), \check{I}^{-1}(\hat{I}(x)) \right) \Big|_{x=\infty} \subseteq (0, \hat{x}).$$

In other words, we find that the image of  $Y$  is strictly included in the domain of  $Y$ . This observation coupled with the fact that  $Y$  is continuous implies that  $Y$  has a fixed point. In other words, there is a state  $z_c^* \in (0, \hat{x})$  for which  $I(z_c^*) = \check{I}(z_c^*) = \hat{I}(\hat{y}(z_c^*)) = I(\hat{y}(z_c^*))$ . Moreover, since  $z_c^* \in (0, \hat{x})$  the equation  $J(\hat{y}(z_c^*)) - J(z_c^*) = -\beta c$  also holds.

*Uniqueness.* Assume that  $z^*$  is a fixed point for the mapping  $Y$ . Since  $J'(x) = -\psi(x)I'(x)$ , we find that

$$Y'(z^*) = \frac{J'(z^*)}{I'(z^*)} \frac{I'(\hat{y}(z^*))}{J'(\hat{y}(z^*))} = \frac{\psi(z^*)}{\psi(\hat{y}(z^*))} < 1.$$

In other words we find that the curve  $Y(x)$  intersects the diagonal always from above. Continuity of  $Y$  yields now the desired uniqueness.  $\square$

Lemma 4.1 demonstrates that Assumption 2.1 is sufficient for both the existence and uniqueness of a solution for the typically highly nonlinear (cf. Figure 4.1) necessary conditions (4.10) and (4.11). It is worth pointing out that since the existence result of Lemma 4.1 is based on a fixed point argument, the existence of a potentially

optimal pair is guaranteed for a considerably broad class of problems. In comparison to the recent paper [10], Lemma 4.1 is analogous to the main result of [10], where the original impulse control problem is defined over threshold policies described by single open interval. We will now proceed by proving that the optimal threshold policy determined by Lemma 4.1 is optimal over the class  $\mathcal{V}$  defined in (2.5), which is much wider than the class of threshold policies.

**THEOREM 4.2.** *Let Assumption 2.1 be satisfied. Then the optimal impulse control policy is to instantaneously take the controlled diffusion  $X_t^\nu$  to the state  $y_c^* - \beta\zeta_c^*$  whenever it hits the state  $y_c^*$  (i.e., the size of the impulse is  $\beta\zeta_c^*$ ). If the initial state  $x \geq y_c^*$ , then  $\tau_1 = 0$  and  $\zeta_1 = \beta^{-1}(x - (y_c^* - \beta\zeta_c^*))$ . Moreover, the value of the optimal impulse control policy reads as*

$$(4.12) \quad V_c(x) = F_c^*(x) = \begin{cases} \beta^{-1}(\lambda x + J(y_c^*)), & x \geq y_c^*, \\ (R_r\pi)(x) - \beta^{-1}I(y_c^*)\psi(x), & x < y_c^*, \end{cases}$$

and the optimal intervention times are given by  $\tau_{i+1} = \inf\{t \geq \tau_i : X_t^\nu \geq y_c^*\}$ ,  $i \geq 1$ .

*Proof.* Since the policy described above is admissible, it is clear that  $F_c^*(x) \leq V_c(x)$  for all  $x \in \mathbb{R}_+$ . To prove the opposite inequality, we first observe that  $F_c^* \in C^1(\mathbb{R}_+) \cap C^2(\mathbb{R}_+ \setminus \{y_c^*\})$  and that  $F_c^{*''}(y_c^*+) = 0 \leq |(R_r\pi)''(y_c^*) - \beta^{-1}I(y_c^*)\psi''(y_c^*)| = |F_c^{*''}(y_c^*-)| < \infty$  implying that  $F_c^*$  is stochastically  $C^2(\mathbb{R}_+)$ . Moreover, we find that  $((\mathcal{A} - r)F_c^*)(x) + \pi(x) = 0$  on  $(0, y_c^*)$  and that  $((\mathcal{A} - r)F_c^*)(x) + \pi(x) = \beta^{-1}(\theta(x) - rJ(y_c^*))$  on  $(y_c^*, \infty)$ . On the other hand, (3.4) implies that

$$I'(x) = \frac{2(rJ(x) - \theta(x))}{\sigma^2(x)\psi'(x)}.$$

Since  $I$  is nondecreasing on  $(\hat{x}, \infty)$  and  $\hat{x} < y_c^*$ , we find that  $\theta(x) \leq rJ(x)$  on  $(y_c^*, \infty)$ . This implies that  $((\mathcal{A} - r)F_c^*)(x) + \pi(x) \leq r\beta^{-1}(J(x) - J(y_c^*)) \leq 0$  for all  $x \in (y_c^*, \infty)$ , since  $J$  is nonincreasing on  $(y_c^*, \infty)$ . Hence  $((\mathcal{A} - r)F_c^*)(x) + \pi(x) \leq 0$  for all  $x \in \mathbb{R}_+ \setminus \{y_c^*\}$ .

Our next task is to show that  $F_c^*$  satisfies the quasi-variational inequality

$$F_c^*(x) \geq \sup_{\beta\zeta \in [0, x]} [F_c^*(x - \beta\zeta) + \lambda\zeta - c]$$

for all  $x \in \mathbb{R}_+$ . Note that this quasi-variational inequality can also be written in the form  $F_c^*(x) \geq \beta^{-1}(\lambda x - \beta c) + \sup_{y \in [0, x]} [F_c^*(y) - \lambda\beta^{-1}y]$ . Define now the mapping  $A : \mathbb{R}_+ \rightarrow \mathbb{R}$  as

$$A(x) = F_c^*(x) - \beta^{-1}(\lambda x - \beta c) - \sup_{y \in [0, x]} [F_c^*(y) - \lambda\beta^{-1}y].$$

Utilizing (4.10) we find that

$$F_c^{*'}(x) = \begin{cases} \lambda\beta^{-1}, & x \geq y_c^*, \\ \beta^{-1}(\lambda + \psi'(x)(I(x) - I(y_c^* - \beta\zeta_c^*))), & x < y_c^*. \end{cases}$$

Since  $I$  is nonincreasing on  $(0, \hat{x})$  and  $y_c^* - \beta\zeta_c^* < \hat{x}$ , we find that  $F_c^{*'}(x) > \lambda\beta^{-1}$  on  $(0, y_c^* - \beta\zeta_c^*)$ . Moreover, the condition (4.10) implies that  $I(x) - I(y_c^*) < 0$  for all  $x \in (y_c^* - \beta\zeta_c^*, y_c^*)$ . Therefore,  $F_c^{*'}(x) \leq \lambda\beta^{-1}$  on  $(y_c^* - \beta\zeta_c^*, y_c^*)$ . Finally, since



$F_c^{*'}(x) = \lambda\beta^{-1}$  on  $(y_c^*, \infty)$ , we find that the function  $x \mapsto F_c^*(x) - \lambda\beta^{-1}x$  attains a global maximum at  $y_c^* - \beta\zeta_c^*$ . Therefore,

$$\sup_{y \in [0, x]} [F_c^*(y) - \lambda\beta^{-1}y] = \begin{cases} F_c^*(y_c^* - \beta\zeta_c^*) - \lambda\beta^{-1}(y_c^* - \beta\zeta_c^*), & x > y_c^* - \beta\zeta_c^*, \\ F_c^*(x) - \lambda\beta^{-1}x, & x \leq y_c^* - \beta\zeta_c^*. \end{cases}$$

Using (4.11) the mapping  $A$  can now be written in the form

$$A(x) = \begin{cases} 0, & x \geq y_c^*, \\ (R_r\pi)(x) - I(y_c^*)\psi(x) - \lambda\beta^{-1}x - J(y_c^*), & x \in (y_c^* - \beta\zeta_c^*, y_c^*), \\ c, & x \leq y_c^* - \beta\zeta_c^*. \end{cases}$$

Since  $\lim_{x \rightarrow y_c^* -} A(x) = 0$  and  $A'(x) = (R_r\pi)(x) - I(y_c^*)\psi'(x) - \lambda\beta^{-1} = \psi'(x)(I(x) - I(y_c^*)) < 0$  for all  $x \in (y_c^* - \beta\zeta_c^*, y_c^*)$ , we find that  $A(x) \geq 0$  on  $(y_c^* - \beta\zeta_c^*, y_c^*)$ ; hence  $A(x) \geq 0$  for all  $x \in \mathbb{R}_+$ .

Finally, given the continuity of  $F_c^*$  and the fact that the state-space  $(0, y_c^*)$  of the controlled diffusion  $X_t^\nu$  is bounded, we observe that  $\mathbf{E}_x[e^{-rt}F_c^*(X_t^\nu)] \rightarrow 0$  for all  $x \in \mathbb{R}_+$  as  $t \rightarrow \infty$ . Thus  $F_c^*(x) \geq V_c(x)$  for all  $x \in \mathbb{R}_+$  and  $\nu^* = \nu_{(\zeta_c^*, y_c^*)}$ .  $\square$

Theorem 4.2 demonstrates that the admissible policy  $\nu^* = \nu_{(\zeta_c^*, y_c^*)}$  is optimal and  $F_c^*$  is the value of the optimal policy under Assumption 2.1. This observation is of interest since it emphasizes the role of the mapping  $\theta$  as the principal determinant of both the existence and uniqueness of an optimal policy. In comparison to [7], it is worth emphasizing that the conditions of Theorem 4.2 are relatively weak since no concavity assumptions are needed and only the monotonicity and continuity properties of the mapping  $\theta$  are required for guaranteeing the validity of our results.

Having studied the existence and uniqueness of an optimal impulse control policy, we now plan to analyze the comparative static properties of the optimal policy and its value. In accordance with our earlier findings in Corollaries 3.5 and 3.7 we can now establish the following corollary.

**COROLLARY 4.3.** *Let Assumption 2.1 be satisfied. Then the value  $V_c$  is a decreasing function of both the parameter  $\beta$  and the sunk cost  $c$  and an increasing function of the parameter  $\lambda$ . In particular, if  $\lambda = \beta$ , then  $\partial\zeta_c^*/\partial\lambda = -\zeta_c^*/\lambda < 0$  and both the optimal exercise boundary  $y_c^*$  and the optimal generic initial state  $z_c^* = y_c^* - \lambda\zeta_c^*$  are independent of  $\lambda$  and  $\beta$ .*

*Proof.* Denote as  $V_{c, \lambda_i}$  the value of the optimal policy associated with the parameter  $\lambda_i$ ,  $i = 1, 2$ , and assume that  $\lambda_1 > \lambda_2$ . It is now clear from the proof of Theorem 4.2 that the value  $V_{c, \lambda_1}$  satisfies the variational inequality  $(\mathcal{A}V_{c, \lambda_1})(x) - rV_{c, \lambda_1}(x) + \pi(x) \leq 0$  for all  $x \in \mathbb{R}_+ \setminus \{y_{c, \lambda_1}^*\}$ , where  $y_{c, \lambda_1}^*$  denotes the optimal exercise threshold associated with the parameter  $\lambda_1$ . Moreover, since  $V_{c, \lambda_1}$  also satisfies the sufficient quasi-variational inequality

$$V_{c, \lambda_1}(x) \geq \sup_{y \in [0, x]} \left[ V_{c, \lambda_1}(y) + \frac{\lambda_1}{\beta}(x - y) \right] - c \geq \sup_{y \in [0, x]} \left[ V_{c, \lambda_1}(y) + \frac{\lambda_2}{\beta}(x - y) \right] - c,$$

we find that  $V_{c, \lambda_1}(x) \geq V_{c, \lambda_2}(x)$ . Proving that  $V_c$  is a decreasing function of both the parameter  $\beta$  and the sunk cost  $c$  is entirely analogous. Finally, if  $\lambda = \beta$ , then the necessary conditions (4.9) imply that  $y_c^*$  and  $z_c^*$  are independent of  $\lambda$  and  $\beta$ . However, since  $z_c^* = y_c^* - \lambda\zeta_c^*$ , we find that  $\partial\zeta_c^*/\partial\lambda = -\zeta_c^*/\lambda < 0$  by partial differentiation.  $\square$

Corollary 4.3 summarizes the impact of parametric changes on the value of the optimal policy. Interestingly, and in contrast to our findings on the associated control

problems, Corollary 4.3 proves that even though the optimal exercise boundary and generic initial state are independent of the parameters  $\lambda$  and  $\beta$  whenever  $\lambda = \beta$ , the optimal impulse  $\zeta_c^*$  is a decreasing function of  $\lambda$ . Thus, the harmonization of the parameters  $\lambda$  and  $\beta$  does not result in the neutrality of the optimal policy. Unfortunately, it is difficult to explicitly characterize the impact of parametric changes in either  $\lambda$  or  $\beta$  on the optimal exercise boundary  $y_c^*$  and the optimal generic initial state  $y_c^* - \beta\zeta_c^*$ . Fortunately, the impact of changes in the sunk cost  $c$  can be explicitly characterized by studying the behavior of the implicit curves  $I(y_c^*) - I(y_c^* - \beta\zeta_c^*) = 0$  and  $J(y_c^*) - J(y_c^* - \beta\zeta_c^*) = -\beta c$ . Implicit differentiation of these curves with respect to  $c$  together with the relation  $J'(x) = -\psi(x)I'(x)$  yield the conditions

$$(4.13) \quad \frac{d(y_c^* - \beta\zeta_c^*)}{dc} = -\frac{\beta}{I'(y_c^* - \beta\zeta_c^*)[\psi(y_c^* - \beta\zeta_c^*) - \psi(y_c^*)]} < 0$$

and

$$(4.14) \quad \frac{dy_c^*}{dc} = -\frac{\beta}{I'(y_c^*)[\psi(y_c^* - \beta\zeta_c^*) - \psi(y_c^*)]} > 0.$$

In other words, the optimal threshold  $y_c^*$  decreases, the regeneration state  $y_c^* - \beta\zeta_c^*$  increases, and, therefore, the optimal impulse  $\zeta_c^*$  decreases as the fixed intervention cost  $c$  decreases. This observation is intuitively clear, since the proof of Lemma 4.1 implies that  $\lim_{c \rightarrow 0+} y_c^* = \hat{x}$  and  $\lim_{c \rightarrow 0+} \zeta_c^* = 0$ . Moreover, by continuity of the increasing fundamental solution  $\psi$ , we discover that  $\lim_{c \rightarrow 0+} \frac{dy_c^*}{dc} = \infty$  and  $\lim_{c \rightarrow 0+} \frac{d\zeta_c^*}{dc} = -\infty$ . Finally, by ordinary differentiation we find that

$$\frac{dV_c}{dc}(x) = \begin{cases} \beta^{-1} J'(y_c^*) \frac{dy_c^*}{dc}, & x \geq y_c^* \\ -\beta^{-1} \psi(x) I'(y_c^*) \frac{dy_c^*}{dc}, & x < y_c^* \end{cases} < 0.$$

Summarizing, we formulate the following lemma characterizing the impact of the transaction cost  $c$  on the value of the optimal policy (see, for example, [32] for a similar observation).

LEMMA 4.4. *Let Assumption 2.1 be satisfied. Then,  $d(y_c^* - \beta\zeta_c^*)/dc < 0$ ,  $dy_c^*/dc > 0$ , and  $d\zeta_c^*/dc > 0$ . Moreover,*

$$\lim_{c \rightarrow 0+} y_c^* = \hat{x}, \lim_{c \rightarrow 0+} \zeta_c^* = 0, \lim_{c \rightarrow 0+} \frac{dy_c^*}{dc} = \infty, \lim_{c \rightarrow 0+} \frac{d\zeta_c^*}{dc} = -\infty, \text{ and } \lim_{c \rightarrow 0+} \frac{dV_c}{dc}(x) = -\infty$$

for all  $x \in \mathbb{R}_+$ .

In light of our general findings and the explicit characterization of the value of the optimal impulse policy, it would be of interest to analyze how increased volatility affects the optimal boundary  $y_c^*$  and the optimal impulse  $\zeta_c^*$ . Unfortunately, as our results indicated, the value function is neither concave nor convex on the entire state-space of the controlled process. Thus, presenting a set of easily verifiable general conditions under which the sign of the relationship between increased volatility and the optimal policy could be unambiguously characterized is extremely difficult, if possible at all.

**4.3. Ordering of the values.** We have established in Lemmas 3.4 and 3.6 and in Theorem 4.2 that all three control problems (3.6), (3.12), and (2.5) are solvable under Assumption 2.1. In this subsection we study how the value functions as well as the marginal values of these associated stochastic control problems can be ordered. To

this end, recall first the expression (3.10) for the value of the singular control problem (3.6). This value can be rewritten as

$$(4.15) \quad K(x) = \begin{cases} \beta^{-1}(\lambda x + J(\hat{x})), & x \geq \hat{x}, \\ (R_r\pi)(x) - \beta^{-1}I(\hat{x})\psi(x), & x < \hat{x}. \end{cases}$$

In the next lemma we show that the value (4.15) has an interesting maximality property. This lemma extends the results obtained in [7] to a model subject to a linear exercise payoff.

LEMMA 4.5. *Define the continuously differentiable mapping  $H : \mathbb{R}_+^2 \rightarrow \mathbb{R}_+$  as*

$$H(x, y) = \begin{cases} \beta^{-1}(\lambda x + J(y)), & x \geq y, \\ (R_r\pi)(x) - \beta^{-1}I(y)\psi(x), & x < y, \end{cases}$$

*and let Assumption 2.1 be satisfied. Then  $K(x) = H(x, \hat{x}) > H(x, y)$  and  $K'(x) = H_x(x, \hat{x}) > H_x(x, y)$  for all  $(x, y) \in \mathbb{R}_+ \times \mathbb{R} \setminus \{\hat{x}\}$ . Moreover,  $H_y(x, y) < 0$ , for all  $(x, y) \in \mathbb{R}_+ \times (\hat{x}, \infty)$ .*

*Proof.* We first find that

$$H_y(x, y) = \begin{cases} -\beta^{-1}I'(y)\psi(y), & x \geq y, \\ -\beta^{-1}I'(y)\psi(x), & x < y. \end{cases}$$

By the monotonicity properties of the function  $I$ , this implies that  $H_y(x, y) \geq 0$ ,  $y \leq \hat{x}$ . This observation coupled with the identity  $K(x) = H(x, \hat{x})$  proves that  $K(x) = H(x, \hat{x}) > H(x, y)$  for all  $(x, y) \in \mathbb{R}_+ \times \mathbb{R} \setminus \{\hat{x}\}$ . Moreover, since

$$H_{xy}(x, y) = \begin{cases} 0, & x \geq y, \\ -\beta^{-1}I'(y)\psi'(x), & x < y, \end{cases}$$

the monotonicity properties of  $I$  imply that  $K'(x) = H_x(x, \hat{x}) > H_x(x, y)$  for all  $(x, y) \in \mathbb{R}_+ \times \mathbb{R} \setminus \{\hat{x}\}$ .  $\square$

Lemma 4.5 shows that the value of the associated singular stochastic control problem not only dominates but also grows faster than any other solution of the associated free-boundary value problem

$$(4.16) \quad \begin{cases} (\mathcal{A}u)(x) - ru(x) + \pi(x) = 0, & x < y, \\ u'(x) = \lambda/\beta, & x \geq y. \end{cases}$$

This result is of interest since it emphasizes the role of the flexibility of the admissible policy as the main determinant of both the actual value and its growth rate. As we will later observe, it is these variational inequalities which relate the considered stochastic impulse control problem to both the associated singular stochastic control problem and to the associated optimal stopping problem.

Our main characterization of the impact of the flexibility of the applied policy on the values and the marginal values of the considered stochastic control problems is now summarized in the following (cf. [7] for a similar observation in the linear payoff case).

THEOREM 4.6. *Let Assumption 2.1 be satisfied. Then*

$$K(x) \geq V_c(x) \geq G_c(x) \quad \text{and} \quad K'(x) \geq V'_c(x) \geq G'_c(x)$$

*for all  $x \in \mathbb{R}_+$ . Moreover,  $\bar{x}_c > y_c^* > \hat{x}$  for all  $c > 0$ .*

*Proof.* In order to prove that  $K(x) \geq G_c(x)$  for all  $x \in \mathbb{R}_+$ , observe that  $K$  satisfies the variational inequality  $(\mathcal{A}K)(x) - rK(x) + \pi(x) \leq 0$  for all  $x \in \mathbb{R}_0$ . Moreover, since  $I$  is decreasing on  $(0, \hat{x})$ , we find that the inequality  $K(x) - (\lambda\beta^{-1}x - c) \geq \beta^{-1}J(\min(x, \hat{x})) \geq 0$  holds for all  $x \in \mathbb{R}_+$ . Thus  $K$  satisfies the sufficient variational inequalities, guaranteeing that  $K(x) \geq G_c(x)$  for all  $x \in \mathbb{R}_+$ . The inequality  $K'(x) \geq G'_c(x)$  for all  $x \in \mathbb{R}_+$  is now a straightforward consequence of the representation (3.11) and Lemma 4.5.

Inequality  $K(x) \geq V_c(x)$  for all  $x \in \mathbb{R}_+$  follows directly from Lemma 4.5 and the representation (4.12). On the other hand, as was established in the proof of Theorem 4.2, the value function  $V_c$  is continuously differentiable on the whole of  $\mathbb{R}_+$ , is twice continuously differentiable on  $\mathbb{R}_+ \setminus \{y_c^*\}$ , and satisfies the variational inequality  $(\mathcal{A}V_c)(x) - rV_c(x) + \pi(x) \leq 0$  for all  $x \in \mathbb{R}_+ \setminus \{y_c^*\}$ . Moreover, since

$$V_c(x) \geq \sup_{\beta\zeta \leq x} [\lambda\zeta - c + V_c(x - \beta\zeta)] \geq \lambda\beta^{-1}x - c$$

for all  $x \in \mathbb{R}_+$ , we observe that  $V_c$  satisfies the sufficient quasi-variational inequalities guaranteeing that  $V_c(x) \geq G_c(x)$  for all  $x \in \mathbb{R}_+$ .

It is clear from the proof of Lemma 4.1 that  $y_c^* > \hat{x}$ . Moreover, since

$$0 \leq V_c(x) - G_c(x) = \beta^{-1}\psi(x)(I(\bar{x}_c) - I(y_c^*))$$

for all  $x \in (0, \min(y_c^*, \bar{x}_c))$  and both of the thresholds  $\bar{x}_c$  and  $y_c^*$  are attained on the set where  $I(x)$  is nondecreasing, we find that  $\bar{x}_c \geq y_c^*$ .

It remains to establish that  $K'(x) \geq V'_c(x) \geq G'_c(x)$  for all  $x \in \mathbb{R}_+$ . Again, the inequality  $K'(x) \geq V'_c(x)$  for all  $x \in \mathbb{R}_+$  follows directly from Lemma 4.5. Since  $\bar{x}_c \geq y_c^* \geq \hat{x}$ , we find that

$$V'_c(x) - G'_c(x) \geq \begin{cases} \beta^{-1}J(y_c^* - \beta\zeta_c^*), & y_c^* < \bar{x}_c \leq x, \\ \beta^{-1}\psi(x)[I(\bar{x}_c) - I(x)], & y_c^* \leq x < \bar{x}_c \text{ or } x < y_c^* < \bar{x}_c. \end{cases}$$

Consequently, we find that  $V'_c(x) - G'_c(x) \geq 0$  for all  $x \in \mathbb{R}_+$  since both of the thresholds  $y_c^*$  and  $\bar{x}_c$  are attained on the set where  $I$  is nondecreasing.  $\square$

Theorem 4.6 states a strong ordering for the stochastic control problems in terms of their values, marginal values, and continuation regions. Interestingly, Theorem 4.6 proves that increased policy flexibility increases not only the value of the optimal stochastic control but also its marginal value. This observation is interesting from the point of view of financial and economical applications since essentially it implies that increased cash flow management flexibility increases not only the value of a rationally managed corporation but also the rate at which this value grows and, therefore, *Tobin's marginal q* associated to the particular cash flow management problem.

The proof of Theorem 4.6 relies on Lemma 4.5. However, it is of interest to notice that the ordering of the optimal values can be justified by a direct inclusion argument. More precisely, it is clear that within our problem specifications, an admissible stopping policy is also an admissible impulse control policy corresponding to a single impulse  $\zeta = (X_\tau/\beta)\chi_{\tau < H_0}$  which takes the underlying diffusion to 0 at the stopping time. On the other hand, since an admissible impulse control policy is nonnegative, nondecreasing, right-continuous, and  $\{\mathcal{F}_t\}$ -adapted, we find that since the value of the associated singular stochastic control problem constitutes the largest attainable value within this class of policies, it, in turn, dominates the value of an admissible impulse control policy, thus resulting in the desired ordering of the values. However, we

emphasize that this simple argument cannot be used for the ordering of the marginal values, which is an essentially more delicate result.

**5. Illustration: Controlled geometric Brownian motion.** In order to illustrate our results explicitly, we now assume that the underlying controlled geometric Brownian motion evolves according to the dynamics characterized by the stochastic differential equation

$$(5.1) \quad X_t^\nu = x + \int_0^t \mu X_s^\nu ds + \int_0^t \sigma X_s^\nu dW_s - \sum_{\tau_k \leq t} \beta \zeta_k, \quad 0 \leq t \leq \tau_0^\nu,$$

where  $\mu > 0$  and  $\sigma > 0$  are exogenously determined known parameters. For the sake of the finiteness of the value of the considered stochastic control problems, we assume that  $r > \mu$ , that is, that the discount rate dominates the expected per capita growth rate of the controlled GBM. It is well known that in this case the fundamental solutions read as  $\psi(x) = x^\kappa$  and  $\varphi(x) = x^\phi$ , where

$$\kappa = \frac{1}{2} - \frac{\mu}{\sigma^2} + \sqrt{\left(\frac{1}{2} - \frac{\mu}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}} > 1$$

and

$$\phi = \frac{1}{2} - \frac{\mu}{\sigma^2} - \sqrt{\left(\frac{1}{2} - \frac{\mu}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}} < 0.$$

Given the considered controlled process, we now assume that the revenue flow accrued from continuing operation reads as  $\pi(x) = x^\alpha$ , where  $\alpha \in (0, 1)$ . Hence, we observe that  $\theta(x) = \beta x^\alpha - (r - \mu)\lambda x$ , implying that the conditions of Lemma 3.1 are satisfied and that

$$x^* = \operatorname{argmax}\{\theta(x)\} = \left(\frac{\alpha\beta}{(r - \mu)\lambda}\right)^{1/(1-\alpha)}.$$

Moreover, standard integration implies that  $(R_r\pi)(x) = x^\alpha/(r - \delta(\alpha))$ , where  $\delta(\alpha) = \alpha\mu + \sigma^2\alpha(\alpha - 1)/2$ .

The value of the optimal singular stochastic control policy reads as

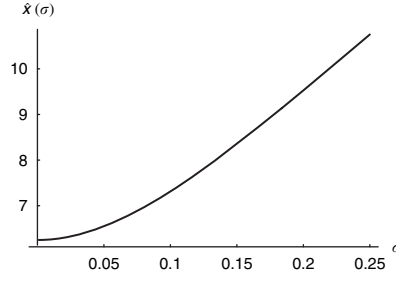
$$(5.2) \quad K(x) = \begin{cases} \frac{\lambda}{\beta}(x - \hat{x}) + \frac{1}{r}\left(\hat{x}^\alpha + \frac{\lambda\mu}{\beta}\hat{x}\right), & x \geq \hat{x}, \\ \frac{x^\alpha}{(r - \delta(\alpha))} + \frac{1}{\kappa}\left(\frac{\lambda}{\beta}\hat{x} - \frac{\alpha\hat{x}^\alpha}{(r - \delta(\alpha))}\right)\left(\frac{x}{\hat{x}}\right)^\kappa, & x < \hat{x}, \end{cases}$$

where the optimal threshold  $\hat{x}$  reads as

$$\hat{x} = \left(\frac{\alpha\beta(\kappa - \alpha)}{(r - \delta(\alpha))(\kappa - 1)\lambda}\right)^{1/(1-\alpha)} = \left(1 + \frac{1 - \alpha}{\alpha - \phi}\right)^{1/(1-\alpha)}.$$

Since

$$\frac{\partial\phi}{\partial\sigma} = \frac{2\phi(\phi - 1)}{\sigma(\kappa - \phi)} > 0,$$

FIG. 5.1. The optimal exercise boundary  $\hat{x}(\sigma)$ .

we immediately find that

$$\frac{\partial \hat{x}}{\partial \sigma} = \left( \frac{1 - \phi}{\alpha - \phi} \right)^{\alpha/(1-\alpha)} \frac{x^*}{(\alpha - \phi)^2} \frac{\partial \phi}{\partial \sigma} > 0.$$

Hence, we find that increased volatility increases the optimal threshold at which the irreversible policy should be exercised. Moreover, standard differentiation also yields that

$$\frac{\partial \hat{x}}{\partial \beta} = \frac{\hat{x}}{(1 - \alpha)\beta} > 0 \quad \text{and} \quad \frac{\partial \hat{x}}{\partial \lambda} = -\frac{\hat{x}}{(1 - \alpha)\lambda} < 0,$$

demonstrating along the lines of our Corollary 3.5 that the optimal exercise threshold is an increasing function of the parameter  $\beta$  and a decreasing function of the parameter  $\lambda$ . The optimal exercise boundary is illustrated as a function of the underlying volatility in Figure 5.1 for  $\beta = 0.9, 1, 1.1$  under the assumption that  $r = 0.045, \mu = 0.025, \alpha = 0.5$ , and  $\lambda = 10$ .

The value of the associated optimal stopping problem reads as

$$(5.3) \quad G_c(x) = \begin{cases} \frac{\lambda}{\beta}x - c, & x \geq \bar{x}_c, \\ \frac{x^\alpha}{(r - \delta(\alpha))} + \left( \frac{\lambda}{\beta}\bar{x}_c - \frac{\bar{x}_c^\alpha}{(r - \delta(\alpha))} - c \right) \left( \frac{x}{\bar{x}_c} \right)^\kappa, & x < \bar{x}_c, \end{cases}$$

where the optimal stopping boundary  $\bar{x}_c > \hat{x}$  is the unique root of the equation

$$\bar{x}_c^\alpha - \frac{(\kappa - 1)(r - \delta(\alpha))\lambda}{\beta(\kappa - \alpha)}\bar{x}_c + \frac{\kappa c(r - \delta(\alpha))}{\kappa - \alpha} = 0.$$

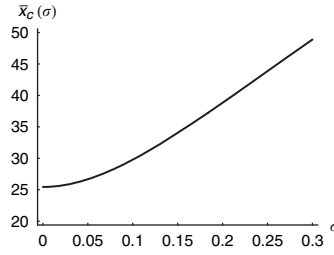
The optimal exercise boundary  $\bar{x}_c$  is illustrated as a function of the underlying volatility in Figure 5.2 for  $\beta = 0.9, 1, 1.1$  under the assumption that  $r = 0.045, \mu = 0.025, \alpha = 0.5, c = 1$ , and  $\lambda = 10$ .

The value of the considered stochastic impulse control problem reads as

$$(5.4) \quad V_c(x) = \begin{cases} \frac{\lambda}{\beta}(x - y_c^*) + \frac{1}{\kappa} \left( \frac{(\kappa - \alpha)y_c^{*\alpha}}{(r - \delta(\alpha))} + \frac{\lambda}{\beta}y_c^* \right), & x \geq y_c^*, \\ \frac{x^\alpha}{(r - \delta(\alpha))} - \frac{1}{\kappa} \left( \frac{\alpha y_c^{*\alpha}}{r - \delta(\alpha)} - \frac{\lambda}{\beta}y_c^* \right) \left( \frac{x}{y_c^*} \right)^\kappa, & x < y_c^*, \end{cases}$$

where the optimal impulse threshold  $y_c^*$  and generic initial state  $z_c^* = y_c^* - \beta\zeta_c^*$  are the unique roots of the optimality conditions

$$\alpha\beta(y_c^{*\alpha-\kappa} - z_c^{*\alpha-\kappa}) = (r - \delta(\alpha))\lambda(y_c^{*1-\kappa} - z_c^{*1-\kappa})$$

FIG. 5.2. The optimal exercise boundary  $\bar{x}_c(\sigma)$ .TABLE 1  
The impact of increased volatility.

$\sigma$	0.01	0.05	0.1	0.15	0.2	0.25
$y_c^*$	9.697	10.168	11.443	13.196	15.222	17.399
$\zeta_c^*$	5.215	5.511	6.319	7.447	8.780	10.252
$\zeta_c^*/y_c^*$	0.5377	0.5420	0.5522	0.5643	0.5768	0.5892
$y_c^* - \zeta_c^*$	4.482	4.656	5.124	5.749	6.443	7.147

TABLE 2  
The impact of increased volatility.

$\sigma$	0.01	0.05	0.1	0.15	0.2	0.25
$y_c^*$	11.562	12.122	13.641	15.727	18.138	20.725
$\zeta_c^*$	6.091	6.437	7.382	8.699	10.257	11.977
$\zeta_c^*/y_c^*$	0.5268	0.5310	0.5412	0.5532	0.5655	0.5779
$y_c^* - \zeta_c^*$	5.471	5.685	6.259	7.027	7.881	8.747

and

$$\beta(\kappa - \alpha)(y_c^{*\alpha} - z_c^{*\alpha}) - \lambda(r - \delta(\alpha))(\kappa - 1)(y_c^* - z_c^*) = -\kappa\beta(r - \delta(\alpha))c.$$

Unfortunately, solving these nonlinear equations explicitly is difficult (if possible at all). Hence, we numerically illustrate the optimal exercise threshold  $y_c^*$ , the optimal impulse  $\zeta_c^*$ , the ratio  $\zeta_c^*/y_c^*$ , and the optimal generic initial state  $y_c^* - \zeta_c^*$  in Table 1 under the assumption that  $r = 0.045$ ,  $\mu = 0.025$ ,  $\alpha = 0.5$ ,  $c = 1$ ,  $\beta = 1$ , and  $\lambda = 10$ .

Table 1 clearly indicates that increased volatility not only increases the optimal threshold at which the impulse policy is irreversibly exercised but also simultaneously increases both the size of the optimal policy and the optimal generic initial state. This result is of interest from the point of view of risk management since it clearly demonstrates that increased volatility will result in both larger but less frequent dividends and a larger generic initial capital protecting the rationally managed corporation from future unfavorable yet uncertain events (i.e., a larger capital buffer). It is also worth emphasizing that our results indicate that the optimal dividend-capital-ratio  $\zeta_c^*/y_c^*$  is also an increasing function of volatility. Consequently, even though increased volatility increases the generic initial state  $y_c^* - \zeta_c^*$ , it simultaneously decreases the ratio between the buffers and the optimal capital.

The impact of a change in the parameter  $\beta$  on the risk sensitivity of the optimal impulse control policy is numerically illustrated in Table 2 under the assumption that  $r = 0.045$ ,  $\mu = 0.025$ ,  $\alpha = 0.5$ ,  $c = 1$ ,  $\beta = 1.1$ , and  $\lambda = 10$  and in Table 3 under the assumption that  $r = 0.045$ ,  $\mu = 0.025$ ,  $\alpha = 0.5$ ,  $c = 1$ ,  $\beta = 0.9$ , and  $\lambda = 10$ . Along the lines of our previous findings on both the associated optimal stopping problem and the associated singular stochastic control problem, our numerical illustrations seem

TABLE 3  
*The impact of increased volatility.*

$\sigma$	0.01	0.05	0.1	0.15	0.2	0.25
$y_c^*$	7.989	8.377	9.429	10.876	12.55	14.35
$\zeta_c^*$	4.394	4.643	5.323	6.272	7.394	8.635
$\zeta_c^*/y_c^*$	0.55	0.5542	0.5646	0.5767	0.5892	0.6017
$y_c^* - \zeta_c^*$	3.595	3.734	4.106	4.604	5.155	5.715

to indicate that the optimal variables are increasing as functions of the parameter  $\beta$  and decreasing as functions of the parameter  $\lambda$ .

**6. Concluding comments.** In this article, we considered a broad class of stochastic impulse control problems arising in the literature on rational cash flow management and optimal harvesting. We presented a set of weak conditions guaranteeing the existence and uniqueness of an optimal pair characterizing the state at which the impulse policy should be exerted and the size of the optimal impulse policy. We derived the value of the optimal policy and characterized its sensitivity with respect to changes in the cost parameters. We also studied two associated stochastic control problems and presented a general ordering for the values as well as for the marginal values of the considered problems. In line with economic intuition, our findings supported the view according to which increased policy flexibility should increase the value of a rationally managed corporation. Moreover, our results indicated that the sign of the relationship between policy flexibility and the rate at which the value of the optimal policy is increasing as a function of the controlled diffusion is unambiguously positive as well.

Even though our results are relatively general in the sense that the controlled diffusion was assumed only to be one-dimensional, they are based on the idea that the admissible bounded variation control policy is one-sided. It would, therefore, be of interest to study whether our findings could be extended to a more general setting where the applied impulse control policy can drive the underlying diffusion both upward and downward (as in the recent study by Weerasinghe [38]). Unfortunately, such extension is outside the scope of the present study and, therefore, left for future research.

**Acknowledgment.** The authors are grateful to two anonymous referees for their constructive comments and suggested improvements on an earlier version of this study.

#### REFERENCES

- [1] L. H. R. ALVAREZ, *A class of solvable singular stochastic control problems*, Stochastics Stochastics Rep., 67 (1999), pp. 83–122.
- [2] L. H. R. ALVAREZ, *Singular stochastic control, linear diffusions, and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.
- [3] L. H. R. ALVAREZ, *Stochastic forest stand value and optimal timber harvesting*, SIAM J. Control Optim., 42 (2004), pp. 1972–1993.
- [4] L. H. R. ALVAREZ, *A class of solvable impulse control problems*, Appl. Math. Optim., 49 (2004), pp. 265–295.
- [5] L. H. R. ALVAREZ AND E. KOSKELA, *The forest rotation problem with stochastic harvest and amenity value*, Natur. Resource Modeling, 20 (2007), pp. 477–509.



- [6] L. H. R. ALVAREZ AND E. KOSKELA, *Taxation and rotation age under stochastic forest stand value*, Journal of Environmental Economics and Management, 54 (2007), pp. 113–127.
- [7] L. H. R. ALVAREZ AND J. VIRTANEN, *A class of solvable stochastic dividend optimization problems: On the general impact of flexibility on valuation*, Econom. Theory, 28 (2006), pp. 373–398.
- [8] F. M. BALDURSSON, *Singular stochastic control and optimal stopping*, Stochastics, 21 (1987), pp. 1–40.
- [9] A. BAR-ILAN, D. PERRY, AND W. STADJE, *A generalized impulse control model of cash management*, J. Econom. Dynam. Control, 28 (2004), pp. 1013–1033.
- [10] E. BAYRAKTAR AND M. EGAMI, *The effects of implementation delay on decision-making under uncertainty*, Stochastic Process. Appl., 117 (2007), pp. 333–358.
- [11] V. BENEŠ, L. A. SHEPP, AND H. S. WITSENHAUSEN, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.
- [12] A. BENSOUSSAN AND J.-L. LIONS, *Impulse Control and Quasivariational Inequalities*, Gauthier-Villars, Montrouge, France, 1984.
- [13] F. E. BENTH AND K. REIKVAM, *A connection between singular stochastic control and optimal stopping*, Appl. Math. Optim., 49 (2004), pp. 27–41.
- [14] F. BOETIUS AND M. KOHLMANN, *Connections between optimal stopping and singular stochastic control*, Stochastic Process. Appl., 77 (1998), pp. 253–281.
- [15] A. BORODIN AND P. SALMINEN, *Handbook on Brownian Motion: Facts and Formulae*, 2nd ed., Birkhäuser, Basel, 2002.
- [16] K. A. BREKKE AND B. ØKSENDAL, *A verification theorem for combined stochastic control and impulse control*, in Stochastic Analysis and Related Topics VI (Geilo, 1996), Birkhäuser Boston, Boston, 1998, pp. 211–220.
- [17] K. A. BREKKE AND B. ØKSENDAL, *Optimal switching in an economic activity under uncertainty*, SIAM J. Control Optim., 32 (1994), pp. 1021–1036.
- [18] A. CADENILLAS, S. SARKAR, AND F. ZAPATERO, *Optimal dividend policy with mean reverting cash reservoir*, Math. Finance, 17 (2007), pp. 81–110.
- [19] M. FREIDLIN, *Functional Integration and Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1985.
- [20] J. M. HARRISON, *Brownian Motion and Stochastic Flow Systems*, Wiley, New York, 1985.
- [21] J. M. HARRISON, T. M. SELLKE, AND A. J. TAYLOR, *Impulse control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 454–466.
- [22] I. KARATZAS, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.
- [23] I. KARATZAS AND S. E. SHREVE, *Connections between optimal stopping and singular stochastic control I. Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.
- [24] I. KARATZAS AND S. E. SHREVE, *Connections between optimal stopping and singular stochastic control II. Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–451.
- [25] T. Ø. KOBILA, *A class of solvable stochastic investment problems involving singular controls*, Stochastics Stochastics Rep., 43 (1993), pp. 29–63.
- [26] R. KORN, *Some applications of impulse control in mathematical finance*, Math. Methods Oper. Res., 50 (1999), pp. 493–518.
- [27] J. L. MENALDI AND M. ROBIN, *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278 (1983), pp. 771–802.
- [28] J. L. MENALDI AND E. ROFMAN, *On stochastic control problems with impulse cost vanishing*, in Semi-infinite Programming and Applications, Lecture Notes in Econom. and Math. Systems 215, A. V. Fiacco and K. O. Kortanek, eds., Springer-Verlag, Berlin, 1983, pp. 281–294.
- [29] G. MUNDACA AND B. ØKSENDAL, *Optimal stochastic intervention control with application to the exchange rate*, J. Math. Econom., 29 (1998), pp. 225–243.
- [30] A. ØKSENDAL, *A Semi-group Approach to Impulse Control Problems*, Department of Mathematics, University of Oslo, Oslo, Norway, Preprint series #14, 2000.
- [31] A. ØKSENDAL, *Irreversible investment problems*, Finance Stoch., 4 (2000), pp. 223–250.
- [32] B. ØKSENDAL, *Stochastic control problems where small intervention costs have big effects*, Appl. Math. Optim., 40 (1999), pp. 355–375.
- [33] B. ØKSENDAL, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., Springer-Verlag, Berlin, 2003.
- [34] S. PEURA AND J. S. KEPPO, *Optimal bank capital with costly recapitalization*, J. Business, 79 (2006), pp. 2163–2201.

- [35] G. PESKIR AND A. SHIRYAEV *Optimal Stopping and Free-boundary Problems*, Birkhäuser, Basel, 2006.
- [36] P. PROTTER, *Stochastic Integration and Differential Equations*, Springer-Verlag, New York, 1990.
- [37] S. SØDAL, *The stochastic rotation problem: A comment*, J. Econom. Dynam. Control, 26 (2002), pp. 509–515.
- [38] A. WEERASINGHE, *A bounded variation control problem for diffusion processes*, SIAM J. Control Optim., 44 (2005), pp. 389–417.
- [39] Y. WILLASSEN, *The stochastic rotation problem: A generalization of Faustmann's formula to stochastic forest growth*, J. Econom. Dynam. Control, 22 (1998), pp. 573–596.

# AN INVERSE PROBLEM FOR A PARABOLIC VARIATIONAL INEQUALITY WITH AN INTEGRO-DIFFERENTIAL OPERATOR\*

YVES ACHDOU†

**Abstract.** We consider the calibration of a Lévy process with American vanilla options. The price of an American vanilla option as a function of the maturity and the strike satisfies a forward in time linear complementarity problem involving a partial integro-differential operator. It leads to a variational inequality in a suitable weighted Sobolev space. Calibrating the Lévy process amounts to solving an inverse problem where the state variable satisfies the previously mentioned variational inequality. We propose a regularized least square method. After studying the variational inequality carefully, we find necessary optimality conditions for the least square problem. In this work, we focus on the case when the volatility is bounded away from zero.

**Key words.** Lévy process, American options, model calibration, inverse problem, variational inequality, optimality conditions

**AMS subject classifications.** 35R30, 35R45, 35R35, 45K05, 47G20, 49K20, 49K22, 49N45, 91B24

**DOI.** 10.1137/060660692

**1. Introduction.** Consider an arbitrage-free market described by a probability measure  $\mathbb{P}$  on a scenario space  $(\Omega, \mathcal{A})$ . There is a risk-free asset whose price at time  $\tau$  is  $e^{r\tau}$ ,  $r \geq 0$ , and a risky asset whose price at time  $\tau$  is  $S_\tau$ . Specifying an arbitrage-free option pricing model necessitates the choice of a risk-neutral measure, i.e., a probability  $\mathbb{P}^*$  equivalent to  $\mathbb{P}$  such that the discounted price  $(e^{-r\tau}S_\tau)_{\tau \in [0, T]}$  is a martingale under  $\mathbb{P}^*$ . Such a probability measure  $\mathbb{P}^*$  allows for the pricing of European options; consider a European option with payoff  $\overline{P}_o$  at maturity  $t \leq T$ : its price at time  $\tau \leq t$  is  $P_\tau = e^{-r(t-\tau)}\mathbb{E}^{\mathbb{P}^*}(\overline{P}_o(S_t)|\mathcal{F}_\tau)$ , where  $(\mathcal{F}_\tau)_{\tau \in [0, T]}$  is the natural filtration. Similarly, consider an American option with payoff  $\overline{P}_o$  and maturity  $t \leq T$ : the price of this option at time  $\tau$  is

$$(1.1) \quad P_\tau = \sup_{s \in \mathcal{T}_{\tau, t}} \mathbb{E}^{\mathbb{P}^*} \left( e^{-r(s-\tau)} \overline{P}_o(S_s) \mid \mathcal{F}_\tau \right),$$

where  $\mathcal{T}_{\tau, t}$  denotes the set of stopping times in  $[\tau, t]$ .

The pricing model  $\mathbb{P}^*$  must be compatible with the prices of the options observed on the market, whose number may be large. *Model calibration* consists of finding  $\mathbb{P}^*$  such that the discounted price  $(e^{-r\tau}S_\tau)_{\tau \in [0, T]}$  is a martingale and such that the option prices computed by, e.g., (1.1) in the case of American options coincide with the observed option prices. This is an *inverse problem*. We focus on the case when the observed prices  $(\bar{p}_i)_{i \in I}$  are those of a family of American vanilla put options indexed by  $i \in I$ , with maturities  $t_i$  (assuming for simplicity that  $T = \max_{i \in I} t_i$ ) and strikes  $x_i$ .

The Black–Scholes model assumes that  $(S_\tau)_{\tau \in [0, T]}$  is a geometric Brownian motion under  $\mathbb{P}^*$ :  $dS_\tau = S_\tau(r d\tau + \sigma dW_\tau)$ , where the volatility  $\sigma$  is a constant. Unfortu-

\*Received by the editors May 24, 2006; accepted for publication (in revised form) July 18, 2007; published electronically February 15, 2008.

<http://www.siam.org/journals/sicon/47-2/66069.html>

†UFR Mathématiques, Université Paris 7, Case 7012, 75251 Paris Cedex 05, France, and Laboratoire Jacques-Louis Lions, Université Paris 6, France (achdou@math.jussieu.fr).

nately, this model is often too simple to match the observed option prices and must be replaced by more involved models.

(1) Black–Scholes models with local volatility. The volatility is assumed to be a function of time and of the price of the underlying asset. This volatility function is calibrated by observing the option prices available on the markets and solving inverse problems involving either partial differential equations or inequalities; see [3, 7, 24] for volatility calibration with European options and [2, 5] with American options;

(2) Models where the volatility is also a stochastic process; see, e.g., [21]. The option price is then found as a function of time, the price of the underlying asset, and the volatility. These models also lead to parabolic partial differential equations or inequalities with possible degeneracies when the volatility vanishes; stochastic volatility calibration has been performed in [32].

(3) Models with Lévy driven underlying assets. Lévy processes are processes with stationary and independent increments which are continuous in probability; see the book by Cont and Tankov [13] and the references therein—for example, [19, 20]. The option price is found by solving partial integro-differential (PID) equations or inequalities. Calibration of Lévy models with European options has been discussed in [14, 15]. The present work is devoted to the calibration of Lévy processes with American options. At this stage, it is not yet necessary to discuss Lévy processes in detail. For the moment, we just assume that the model is characterized by parameters  $\theta$  in a suitable class  $\Theta$ .

The last two classes of models describe incomplete markets: the knowledge of the historical price process alone does not allow us to compute the option prices in a unique manner. When the option prices do not determine the model completely, additional information may be introduced into the problem by specifying a *prior* model. If the historical price process has been estimated statistically from the time series of the underlying asset, this knowledge has to be injected in the inverse problem; calling  $\mathbb{P}_0$  the prior probability measure obtained as an estimation of  $\mathbb{P}$ , we are going to focus on least square formulations of the following type: find  $\theta \in \Theta$  which minimizes

$$(1.2) \quad \sum_{i \in I} \omega_i \left( P^\theta(0, S_o, t_i, x_i) - \bar{p}_i \right)^2 + \rho J_2(\mathbb{P}^\theta, \mathbb{P}_0),$$

where

- $\omega_i$  are suitable positive weights,
- $S_o$  is the price of the underlying asset today,
- $P^\theta(0, S_o, t_i, x_i)$  is the price of the option with maturity  $t_i$  strike  $x_i$ , computed with the pricing model associated with  $\theta$ , and
- $\rho J_2(\mathbb{P}^\theta, \mathbb{P}_0)$  is a regularization term which measures the closedness of the model  $\mathbb{P}^\theta$  to the prior. The number  $\rho > 0$  is called the regularization parameter. This functional has two roles: (1) it stabilizes the inverse problem; for that,  $\rho$  should be large enough and  $J_2$  should be convex or at least convex in a large enough region; (2) it guarantees that  $\mathbb{P}^\theta$  remains close to  $\mathbb{P}_0$  in some sense. The choice of  $J_2$  is very important:  $J_2(\mathbb{P}^\theta, \mathbb{P}_0)$  is often chosen as the relative entropy of the pricing measure  $\mathbb{P}^\theta$  w.r.t. the prior model  $\mathbb{P}_0$  (see [8]) because the relative entropy becomes infinite if  $\mathbb{P}^\theta$  is not equivalent to  $\mathbb{P}_0$ . Some authors have argued that such a choice may be too conservative in some cases for two reasons: (a) the historical data which determine the prior may be missing or partially available; (b) in the context of, e.g., volatility calibration, once the volatility is specified under  $\mathbb{P}_0$ , then the volatility under

$\mathbb{P}^\theta$  must be the same for the relative entropy to be finite. In [9] a different approach was considered which allowed for volatility calibration.

Note that evaluating the functional in (1.2) requires solving  $\#I$  linear complementarity problems (LCPs) involving PID operators in the variables  $\tau$  and  $S$ ; see section 2.1 below. This approach was chosen in [2, 5] for calibrating local volatility with American options.

In the present case (Lévy driven assets), we show that there is a better approach which consists of computing the prices  $P^\theta(0, S_o, t_i, x_i)$ ,  $i \in I$ , by using a single forward in time LCP with a PID operator in the variables maturity  $t$  and strike  $x$ . This LCP is introduced in section 2.2; see (2.9)–(2.11) below. It is reminiscent of the forward equation (known as Dupire’s equation in the finance community) which is often used for local volatility calibration with vanilla European options; see [4, 18].

We then find a new least square problem where the functional is evaluated by solving a single LCP involving a PID operator. The main goal of the paper is to study this least square problem theoretically for a rather general parameterization of the Lévy density  $k$  (see (2.14) below), with the volatility  $\sigma$  bounded away from 0, and to give necessary optimality conditions. This problem has connections with some optimal control problems for variational inequalities studied in [11, 23, 31]. The article of Hintermüller [22] on an inverse problem for an elliptic variational inequality has inspired [2] and the present work.

As far as we know, this is the first attempt at calibrating Lévy processes with American options, so comparison with other methods is difficult. The results below can be used in practice because they have their discrete counterparts when finite elements or finite differences are used. The accuracy is expected to be similar to that observed in [5].

The paper is organized as follows. In section 2, we obtain the forward LCP (2.9)–(2.11) and make some assumptions on the Lévy density. In section 3, we introduce a family of fractional weighted Sobolev spaces and give preliminary results on the nonlocal operator in (2.9). In section 4 we carefully study the variational inequality stemming from (2.9)–(2.11). For the analysis, we must first study a regularized nonlinear problem posed in a bounded domain and then let the regularization parameter tend to 0 and the domain’s boundary tend to infinity. The sensitivity of the solution to variations of  $\sigma$  and  $k$  is discussed in section 5. Finally, the inverse problem is studied in section 6: necessary optimality conditions are given. Some technical proofs are postponed in Appendices A and B.

For the reader’s convenience, let us point out the main results of this work:

- The forward complementarity problem is written in (2.9)–(2.11), and the assumptions on  $k$  are described in section 2.3.
- Theorem 4.9 contains a result of existence and uniqueness for the variational inequality associated to (2.9)–(2.11) in suitable Sobolev spaces. It is also proved that the related free boundary stays in a bounded region. Note that, by using the theory presented in [10], it is possible to study the variational inequality in Sobolev spaces with decaying weights as  $x \rightarrow 0$  and  $x \rightarrow +\infty$  (actually the variable  $\log(x)$  was used instead of  $x$  in [10]). Here, we show that these weights can be avoided. Another advantage of the present analysis is that it can be extended to the case when  $\sigma = 0$  by singular perturbation arguments if the Lévy measure is chosen to keep the problem parabolic. This will be done in a forthcoming work [1].
- The sensitivity of solutions w.r.t. variations of the Lévy process is studied in section 5.

- Theorem 6.6 contains the necessary optimality conditions for the least square inverse problem. These conditions are obtained by first studying a modified inverse problem whose state variable satisfies the above-mentioned regularized nonlinear problem and then by passing to the limit as the regularization parameter tends to zero.

## 2. Description of the model.

**2.1. The backward LCP.** For a Lévy process  $(X_\tau)_{\tau \geq 0}$  on a filtered probability space, the Lévy–Khintchine formula says that there exists a function  $\chi : \mathbb{R} \rightarrow \mathbb{C}$  such that  $\mathbb{E}(e^{iuX_\tau}) = e^{\tau\chi(u)}$ , with

$$\chi(u) = -\frac{\sigma^2 u^2}{2} + i\beta u + \int_{|z| < 1} (e^{iuz} - 1 - iuz)\nu(dz) + \int_{|z| > 1} (e^{iuz} - 1)\nu(dz)$$

for  $\sigma \geq 0$ ,  $\beta \in \mathbb{R}$ , and a positive measure  $\nu$  on  $\mathbb{R} \setminus \{0\}$  such that  $\int_{\mathbb{R}} \min(1, z^2)\nu(dz) < +\infty$ . The measure  $\nu$  is called the Lévy measure of  $(X_\tau)_{\tau \geq 0}$ .

We assume that the discounted price of the risky asset is a martingale obtained as the exponential of a Lévy process:  $e^{-r\tau}S_\tau = S_0 e^{\tilde{X}_\tau}$ . The fact that the discounted price is a martingale is equivalent to

$$\int_{|z| > 1} e^z \nu(dz) < \infty \quad \text{and} \quad \beta = -\frac{\sigma^2}{2} - \int_{\mathbb{R}} (e^z - 1 - z1_{|z| \leq 1})\nu(dz).$$

We also assume that  $\int_{|z| > 1} e^{2z}\nu(dz) < \infty$ , so the discounted price is a square integrable martingale.

In what follows, we assume that the Lévy measure has a density,  $\nu(dz) = k(z)dz$ , with  $k$  possibly singular at  $z = 0$ . Doing so, we exclude the simplest Lévy processes obtained as the sum of Brownian motions and Poisson processes. This is not a fundamental restriction in the sense that the methods proposed below could be extended (and even simplified) to calibrate the previously mentioned processes. The restriction is mainly done in order to focus on the difficulties posed by the possible singularities of  $k$  at  $z = 0$ .

We denote  $\bar{B}$  the integral operator:

$$(\bar{B}v)(S) = \int_{\mathbb{R}} \left( v(Se^z) - v(S) - S(e^z - 1) \frac{\partial}{\partial S} v(S) \right) k(z) dz.$$

Consider an American option with payoff  $\bar{P}_o$  and maturity  $t$ . In [10], Bensoussan and Lions assume  $\sigma > 0$  and study the variational inequality stemming from the LCP:  $P(t, S) = \bar{P}_o(S)$ , and for  $\tau < t$  and  $S > 0$ ,

$$(2.1) \quad \frac{\partial P}{\partial \tau}(\tau, S) + \frac{\sigma^2 S^2}{2} \frac{\partial^2 P}{\partial S^2}(\tau, S) + rS \frac{\partial P}{\partial S}(\tau, S) - rP(\tau, S) + (\bar{B}P)(\tau, S) \leq 0,$$

$$(2.2) \quad P(\tau, S) \geq \bar{P}_o(S),$$

$$(2.3) \quad \left( \frac{\partial P}{\partial \tau}(\tau, S) + \frac{\sigma^2 S^2}{2} \frac{\partial^2 P}{\partial S^2}(\tau, S) + rS \frac{\partial P}{\partial S}(\tau, S) - rP(\tau, S) + (\bar{B}P)(\tau, S) \right) (P(\tau, S) - \bar{P}_o(S)) = 0$$

in suitable Sobolev spaces with decaying weights near  $+\infty$  and 0. They prove that the price of the American option is  $P_\tau = P(\tau, S_\tau)$ . Other approaches with viscosity solutions are possible (see [34]), especially in the case  $\sigma = 0$ . One advantage of the variational methods is that they provide stability estimates. For numerical methods for options on Lévy driven assets; see [4, 16, 17, 28, 29, 30].

**2.2. The forward LCP.** As already explained, we aim at finding a forward LCP in the variables maturity/strike; a single solution of this problem will be needed for evaluating the cost function in (1.2).

Hereafter, since the observed prices are those of vanilla American put options, we use the notation

$$(2.4) \quad P_{\circ}(x) = (x - S)_{+}.$$

If  $\bar{P}_{\circ}(S) = (x - S)_{+}$ , it can be seen that the solution of (2.1)–(2.3) is of the form

$$(2.5) \quad P(\tau, S, t, x) = xg(\xi, y), \quad y = S/x \in \mathbb{R}_{+}, \quad \xi = t - \tau \in (0, t),$$

where  $g$  is the solution of the complementarity problem independent of  $x$ ,  $g(0, y) = (1 - y)_{+}$ , and for  $0 < \xi \leq t$ ,  $y \in \mathbb{R}_{+}$ ,

$$(2.6) \quad -\frac{\partial g}{\partial \xi}(\xi, y) + \frac{\sigma^2 y^2}{2} \frac{\partial^2 g}{\partial y^2}(\xi, y) + ry \frac{\partial g}{\partial y}(\xi, y) - rg(\xi, y) + (\check{B}g)(\xi, y) \leq 0,$$

$$(2.7) \quad g(\xi, y) \geq (1 - y)_{+},$$

$$(2.8) \quad \left( \begin{array}{c} -\frac{\partial g}{\partial \xi}(\xi, y) + \frac{\sigma^2 y^2}{2} \frac{\partial^2 g}{\partial y^2}(\xi, y) + ry \frac{\partial g}{\partial y}(\xi, y) \\ -rg(\xi, y) + (\check{B}g)(\xi, y) \end{array} \right) (g(\xi, y) - (1 - y)_{+}) = 0,$$

where  $(\check{B}v)(y) = \int_{\mathbb{R}} (v(ye^z) - v(y) - y(e^z - 1) \frac{\partial}{\partial y} v(y)) k(z) dz$ . From this observation and the identities  $x \frac{\partial g}{\partial \xi} = -\frac{\partial P}{\partial t}$ ,  $xy \frac{\partial g}{\partial y} = -x \frac{\partial P}{\partial x} + P$ , and  $xy^2 \frac{\partial^2 g}{\partial y^2} = x^2 \frac{\partial^2 P}{\partial x^2}$ , we deduce that, as a function of  $t$  and  $x$ ,  $P(0, S, t, x)$  satisfies the following forward problem:  $P(t = 0) = P_{\circ}$  and for  $t \in (0, T]$  and  $x > 0$ ,

$$(2.9) \quad \left( \frac{\partial P}{\partial t} - \frac{\sigma^2 x^2}{2} \frac{\partial^2 P}{\partial x^2} + rx \frac{\partial P}{\partial x} + BP \right) \geq 0,$$

$$(2.10) \quad P(t, x) \geq P_{\circ}(x),$$

$$(2.11) \quad \left( \frac{\partial P}{\partial t} - \frac{\sigma^2 x^2}{2} \frac{\partial^2 P}{\partial x^2} + rx \frac{\partial P}{\partial x} + BP \right) (P - P_{\circ}) = 0,$$

where the integral operator  $B$  is defined by

$$(2.12) \quad (Bu)(x) = - \int_{\mathbb{R}} k(z) \left( x(e^z - 1) \frac{\partial u}{\partial x}(x) + e^z (u(xe^{-z}) - u(x)) \right) dz.$$

Note that the arguments yielding (2.9)–(2.11) are much easier than those used for getting Dupire's equation (see [4, 18]), because (2.5) does not hold with local volatility. Problem (2.9)–(2.11) can also be obtained by probabilistic arguments. Note also that finding a forward LCP in the variables  $t$  and  $x$  is not possible in the case of American options with local volatility, because the arguments in [4, 18] do not apply to nonlinear problems. This explains why, in [2, 5], the evaluation of the least square cost functional necessitates the solution of  $\#I$  LCP instead of one here. In this respect, we may say that with American options, the calibration of Lévy processes is easier than the calibration of local volatility.

**2.3. Choice of the Lévy process.** We have already discussed our choice to take  $\nu(dz) = k(z)dz$ , with

$$(2.13) \quad \max \left( \int_{\mathbb{R}} \min(1, z^2) k(z) dz, \int_1^{+\infty} e^{2z} k(z) dz \right) < \infty.$$

We need to make further restrictions on the Lévy process for several reasons:

1. In practice, we need to specify a class of Lévy densities  $k$  in order to define the inverse problem.
2. The analysis below will need problem (2.16)–(2.19) to be parabolic. This implies restrictions on the pair  $(\sigma, k)$ .

As will appear in section 3.2 below, the restrictions in order to have a parabolic problem are

1. either  $\sigma > 0$  and  $k$  satisfies (2.13),
2. or  $\sigma = 0$  and  $k$  satisfies (2.13) and is sufficiently singular near  $z = 0$ . The result in section 3.2 will imply that choosing  $k(z) \sim |z|^{-1-2\alpha}$  with  $1/2 < \alpha < 1$  yields a parabolic problem. To keep the length of this article reasonable, this case will be discussed elsewhere.

ASSUMPTION 1. *For this reason, we assume that  $k$  is of the form*

$$(2.14) \quad k(z) = \psi(z)|z|^{-(1+2\alpha)},$$

where  $\psi$  is a nonnegative function in  $L^\infty(\mathbb{R})$  such that  $\psi(z) \geq \underline{\psi} > 0$  in a fixed neighborhood of  $z = 0$ , and  $\alpha$  is such that  $-1/2 \leq \alpha < 1$ . We assume furthermore that (2.13) is satisfied.

For practical purposes, one can impose further restrictions on  $\psi$ , for example, let  $\psi$  belong to a finite dimensional function space, but this need not be discussed at this stage.

Assumption 1 holds for models of jump-diffusion type, for example, the Merton model ( $\sigma > 0$  and the jumps in the log-price have a Gaussian distribution) or some Kou models ( $\sigma > 0$  and the distribution of jumps is an asymmetric exponential with a fast enough decay at infinity); see [13, p. 111]. Indeed, these models can be obtained by taking  $\alpha = -1/2$  and choosing  $\psi$  properly. Assumption 1 also holds for some variance gamma processes ( $\sigma > 0$ ,  $\alpha = 0$ ) and normal inverse Gaussian processes ( $\sigma > 0$ ,  $\alpha = 1/2$ ) (see [13, p. 117]), with a fast enough decay of the jump density at infinity. It also holds for some tempered stable processes (see [13, p. 119]) or some parabolic CGMY models discussed by Carr et al. [12]. These last two models usually take  $\sigma = 0$ . Allowing  $\sigma > 0$  in the analysis can be seen as a step toward  $\sigma = 0$ .

REMARK 1. *The assumption  $\psi(z) \geq \underline{\psi} > 0$  near 0 avoids ambiguities in the definition of the singularity of  $k$  at  $z = 0$ . It is a bit restrictive since, for example, a logarithmic singularity of  $k$  at  $z = 0$  is ruled out. However, this assumption is unessential and most of the results below hold without it.*

**2.4. Change of unknown function in the forward problem.** In order to have a datum with a compact support in  $x$ , it is helpful to change the unknown function: we set

$$(2.15) \quad u_o(x) = (S - x)_+; \quad u(t, x) = P(t, x) - x + S.$$

The function  $u$  satisfies, for  $t \in (0, T]$  and  $x > 0$ ,

$$(2.16) \quad \frac{\partial u}{\partial t} - \frac{\sigma^2 x^2}{2} \frac{\partial^2 u}{\partial x^2} + rx \frac{\partial u}{\partial x} + Bu \geq -rx,$$

$$(2.17) \quad u(t, x) \geq u_o(x),$$

$$(2.18) \quad \left( \frac{\partial u}{\partial t} - \frac{\sigma^2 x^2}{2} \frac{\partial^2 u}{\partial x^2} + rx \frac{\partial u}{\partial x} + Bu + rx \right) (u - u_o) = 0.$$



The initial condition for  $u$  is

$$(2.19) \quad u(t = 0, x) = u_o(x), \quad x > 0.$$

For writing the variational inequalities stemming from (2.16)–(2.19), we need to introduce suitable weighted Sobolev spaces. In particular, fractional order weighted Sobolev spaces will be useful for studying the nonlocal part of the operator.

### 3. Preliminary results.

#### 3.1. Functional setting.

**3.1.1. Sobolev spaces on  $\mathbb{R}$ .** For a real number  $s$ , let the Sobolev space  $H^s(\mathbb{R})$  be defined as follows: the distribution  $w$  defined on  $\mathbb{R}$  belongs to  $H^s(\mathbb{R})$  if and only if its Fourier transform  $\widehat{w}$  satisfies  $\int_{\mathbb{R}} (1 + \xi^2)^s |\widehat{w}(\xi)|^2 d\xi < +\infty$ . The spaces  $H^s(\mathbb{R})$  are Hilbert spaces, with the inner product and norm:

$$(w_1, w_2)_{H^s(\mathbb{R})} = \int_{\mathbb{R}} (1 + \xi^2)^s \widehat{w}_1(\xi) \overline{\widehat{w}_2(\xi)} d\xi, \quad \|w\|_{H^s(\mathbb{R})} = \sqrt{(w, w)_{H^s(\mathbb{R})}}.$$

For two real numbers  $s_1, s_2$ ,  $s_1 \leq s_2$ ,  $H^{s_2}(\mathbb{R}) \subset H^{s_1}(\mathbb{R})$  with a continuous injection. It can be seen that  $H^0(\mathbb{R}) = L^2(\mathbb{R})$  and that if  $s$  is a positive integer,  $H^s(\mathbb{R})$  is the space of all the functions whose derivatives up to order  $s$  are square integrable. If  $s$  is a non-negative integer, the norm  $\|\cdot\|_{H^s(\mathbb{R})}$  is equivalent to the norm  $v \mapsto \sqrt{\sum_{\ell=0}^s \|\frac{d^\ell v}{dy^\ell}\|_{L^2(\mathbb{R})}^2}$ . If  $s > 0$  is not an integer, the norm  $\|\cdot\|_{H^s(\mathbb{R})}$  is equivalent to

$$(3.1) \quad v \mapsto \sqrt{\sum_{\ell=0}^m \left\| \frac{d^\ell v}{dy^\ell} \right\|_{L^2(\mathbb{R})}^2 + \int_{\mathbb{R}} \int_{\mathbb{R}} |y - z|^{2(m-s)-1} \left( \frac{d^m v}{dy^m}(y) - \frac{d^m v}{dy^m}(z) \right)^2 dy dz},$$

where  $m$  is the integer part of  $s$ . For  $s \geq 0$ , the space  $\mathcal{D}(\mathbb{R})$  is dense in  $H^s(\mathbb{R})$ .

It is well known (see [27, 6]) that if  $0 < s < 1$ , then  $H^s(\mathbb{R})$  can be obtained by real or complex interpolation between the spaces  $H^1(\mathbb{R})$  and  $L^2(\mathbb{R})$  (the parameter for the real interpolation is  $\nu = 1/2 - s$ ; see [6, p. 204]), and that the norm obtained by the interpolation process is equivalent to the one defined in (3.1).

For  $s \geq 0$ ,  $H^{-s}(\mathbb{R})$  is the dual of  $H^s(\mathbb{R})$ , and for  $s > 0$ , the norm  $\|\cdot\|_{H^{-s}(\mathbb{R})}$  is equivalent to the norm  $v \mapsto \sup_{w \in H^s(\mathbb{R}), w \neq 0} \frac{|\langle v, w \rangle|}{\|w\|_{H^s(\mathbb{R})}}$ . If  $s$  is a nonnegative integer, we define the seminorm  $|v|_{H^s(\mathbb{R})} = \sqrt{\sum_{\ell=1}^s \|\frac{d^\ell v}{dy^\ell}\|_{L^2(\mathbb{R})}^2}$ . If  $s > 0$  is not an integer, we define  $|v|_{H^s(\mathbb{R})}$  by  $|v|_{H^s(\mathbb{R})}^2 = \sum_{\ell=1}^m \|\frac{d^\ell v}{dy^\ell}\|_{L^2(\mathbb{R})}^2 + \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{(\frac{d^m v}{dy^m}(y) - \frac{d^m v}{dy^m}(z))^2}{|y - z|^{1+2s}}$ , where  $m$  is the integer part of  $s$ .

**3.1.2. Some weighted Sobolev spaces on  $\mathbb{R}_+$ .** Let  $V^1$  be the weighted Sobolev space

$$V^1 = \left\{ v \in L^2(\mathbb{R}_+), \quad x \frac{\partial v}{\partial x} \in L^2(\mathbb{R}_+) \right\},$$

which is a Hilbert space with the norm  $\|v\|_{V^1} = \sqrt{\|v\|_{L^2(\mathbb{R}_+)}^2 + \|x \frac{\partial v}{\partial x}\|_{L^2(\mathbb{R}_+)}^2}$ . It is proved in [4] that  $\mathcal{D}(\mathbb{R}_+)$  is a dense subspace of  $V^1$ , and that the following Poincaré inequality is true: for all  $v \in V^1$ ,

$$(3.2) \quad \|v\|_{L^2(\mathbb{R}_+)} \leq 2 \|x \frac{dv}{dx}\|_{L^2(\mathbb{R}_+)}.$$

Therefore, the seminorm  $|\cdot|_{V^1}: |v|_{V^1} = \|x \frac{dv}{dx}\|_{L^2(\mathbb{R}_+)}$  is a norm equivalent to  $\|\cdot\|_{V^1}$ .

For a function  $v$  defined on  $\mathbb{R}_+$ , call  $\tilde{v}$  the function defined on  $\mathbb{R}$  by

$$(3.3) \quad \tilde{v}(y) = v(\exp(y)) \exp(y/2).$$

By using the change of variable  $y = \log(x)$ , it can be seen that the mapping  $v \mapsto \tilde{v}$  is a topological isomorphism from  $L^2(\mathbb{R}_+)$  onto  $L^2(\mathbb{R})$ , and from  $V^1$  onto  $H^1(\mathbb{R})$ . This leads to defining the space  $V^s$  for  $s \in \mathbb{R}$  by

$$V^s = \{v : \tilde{v} \in H^s(\mathbb{R})\},$$

which is a Hilbert space with the norm  $\|v\|_{V^s} = \|\tilde{v}\|_{H^s(\mathbb{R})}$ . Using the interpolation theorem given in, e.g., [6, Theorem 7.17], one can prove that if  $0 < s < 1$ , then  $V^s$  can be obtained by real interpolation between the spaces  $V^1$  and  $L^2(\mathbb{R}_+)$  (the parameter for the real interpolation is  $\nu = 1/2 - s$ ), and that the norm obtained by the interpolation process is equivalent to the one defined above.

For  $s > 0$ , the space  $V^{-s}$  is the topological dual of  $V^s$ .

For  $s > 0$ , we introduce the seminorm  $|v|_{V^s} = |\tilde{v}|_{H^s(\mathbb{R})}$ .

LEMMA 3.1. *Let  $s$  be a real number such that  $1/2 < s \leq 1$ . Then for all  $v \in V^s$ ,  $v$  is continuous on  $(0, +\infty)$  and there exists a constant  $C > 0$  such that*

$$(3.4) \quad \sqrt{x}|v(x)| \leq C\|v\|_{V^s} \quad \forall x \in [1, +\infty).$$

*Proof.* From the Sobolev continuous imbedding  $H^s(\mathbb{R}) \subset L^\infty(\mathbb{R}) \cap C^0(\mathbb{R})$  for  $s > 1/2$ , we see that  $V^s \subset C^0((0, +\infty))$  and there exists a constant  $C$  such that  $|v(x)| = |\tilde{v}(\log(x))|/\sqrt{x} \leq C\|\tilde{v}\|_{H^s(\mathbb{R})}/\sqrt{x} = C\|v\|_{V^s}/\sqrt{x}$  for all  $v \in V^s$ , for all  $x \geq 1$ .  $\square$

For a continuous and nonnegative function  $\phi$  defined on  $\mathbb{R}$  and a measurable function  $v$  on  $\mathbb{R}_+$ , consider

$$|v|_{\phi,s}^2 = \int_{\mathbb{R}_+} dx \int_{\mathbb{R}} \frac{\phi(z)}{|z|^{1+2s}} (v(xe^{-z}) - v(x))^2 dz, \text{ and } \|v\|_{\phi,s} = \sqrt{|v|_{\phi,s}^2 + \|v\|_{L^2(\mathbb{R}_+)}^2}.$$

LEMMA 3.2. *Let  $\phi$  be a continuous and nonnegative function defined on  $\mathbb{R}$ . If  $\phi(0) > 0$  and if the function  $z \mapsto \phi(z) \max(e^z, 1)$  is bounded, then for any  $s \in (0, 1)$ ,  $\|\cdot\|_{\phi,s}$  is a norm on  $V^s$  equivalent to the norm  $\|\cdot\|_{V^s}$ .*

*Proof.* For the reader's ease, the proof is postponed to Appendix A.  $\square$

REMARK 2. *Lemma 3.2 remains true if  $\phi$  is a function in  $L^\infty(\mathbb{R})$  and if for a given positive constant  $\underline{\phi}$ ,  $\phi \geq \underline{\phi} > 0$  a.e. in a neighborhood of 0.*

REMARK 3. *If the assumption  $\phi(0) > 0$  is not satisfied, then the conclusion of Lemma 3.2 becomes  $\exists C > 0$  such that  $|u|_{\phi,s} \leq C\|u\|_{V^s}$  for all  $u \in V^s$ .*

## 3.2. The integro-differential operator.

### 3.2.1. The integral operator.

We study the operator  $B$  defined in (2.12).

LEMMA 3.3. *Let  $(\alpha, \psi)$  satisfy Assumption 1. For each  $s \in \mathbb{R}$ , the following hold:*

- if  $\alpha > 1/2$ , then the operator  $B$  is continuous from  $V^s$  to  $V^{s-2\alpha}$ ,
- if  $\alpha < 1/2$ , then the operator  $B$  is continuous from  $V^s$  to  $V^{s-1}$ , and
- if  $\alpha = 1/2$ , then the operator  $B$  is continuous from  $V^s$  to  $V^{s-1-\epsilon}$  for any  $\epsilon > 0$ .

*Proof.* See Appendix A.  $\square$

COROLLARY 3.4. *If  $(\alpha, \psi)$  satisfy Assumption 1 and if  $1/2 < \alpha < 1$ , then the operator  $B$  is continuous from  $V^\alpha$  to  $V^{-\alpha}$ .*

LEMMA 3.5. *If  $(\alpha, \psi)$  satisfy Assumption 1 and  $1/2 < \alpha < 1$ , then for all  $v, w \in V^\alpha$ ,*

$$(3.5) \quad \langle Bu, v \rangle + \langle Bv, u \rangle = \begin{cases} \int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z (u(x) - u(xe^{-z})) (v(x) - v(xe^{-z})) dx dz \\ + (\int_{\mathbb{R}} k(z) (2e^z - e^{2z} - 1) dz) \int_{\mathbb{R}_+} u(x) v(x) dx, \end{cases}$$

where  $\langle \cdot, \cdot \rangle$  stands for the duality pairing between  $V^{-\alpha}$  and  $V^\alpha$ .

If  $-1/2 \leq \alpha \leq 1/2$ , then (3.5) is true for  $u, v \in V^s$ ,  $s > 1/2$ , defining  $\langle \cdot, \cdot \rangle$  as the duality pairing between  $V^{-s}$  and  $V^s$ .

*Proof.* See Appendix A.  $\square$

REMARK 4. *If  $(\alpha, \psi)$  satisfy Assumption 1, the operator  $B^T$  defined by*

$$(3.6) \quad B^T u(x) = \int_{\mathbb{R}} k(z) \left( x(e^z - 1) \frac{\partial u}{\partial x}(x) - e^{2z} u(xe^z) + (2e^z - 1) u(x) \right) dz$$

*is a continuous operator from  $V^s$  to  $V^{s-2\alpha}$  if  $\alpha > 1/2$ , is a continuous operator from  $V^s$  to  $V^{s-1}$  if  $\alpha < 1/2$ , and is a continuous operator from  $V^s$  to  $V^{s-1-\epsilon}$  for any  $\epsilon > 0$  if  $\alpha = 1/2$ .*

*If  $\alpha > 1/2$ , then for all  $u, v \in V^\alpha$ ,  $\langle B^T u, v \rangle = \langle Bv, u \rangle$ . This identity holds for all  $u, v \in V^s$  with  $s > 1/2$  if  $\alpha \leq 1/2$ .*

LEMMA 3.6. *If  $(\alpha, \psi)$  satisfy Assumption 1 and if*

- *either  $\alpha < 1/2$ ,*
- *or  $\psi$  is continuous near 0 and there exists a bounded function  $\omega : \mathbb{R} \rightarrow \mathbb{R}$  and two positive numbers  $\zeta$  and  $C$  such that  $\psi(z)e^{3/2z} - \psi(0)e^{-3/2z} = z\omega(z)$ , with  $|\omega(z)| \leq C|z|e^{-\zeta|z|}$ , for all  $z \in \mathbb{R}$ ,*

*then for any  $s \in \mathbb{R}$ , the operator  $B - B^T$  is continuous from  $V^s$  to  $V^{s-1}$ .*

*Proof.* See Appendix A.  $\square$

PROPOSITION 3.7 (Gårding inequality). *Let  $(\alpha, \psi)$  satisfy Assumption 1. If  $1/2 < \alpha < 1$ , there exist two constants  $\underline{C} > 0$  and  $\lambda \geq 0$  such that, for all  $v \in V^\alpha$ ,*

$$(3.7) \quad \langle Bv, v \rangle \geq \underline{C} \|v\|_{V^\alpha}^2 - \lambda \|v\|_{L^2(\mathbb{R}_+)}^2.$$

*If  $\alpha \leq 1/2$ , then (3.7) holds for any  $v \in V^s$ ,  $s > 1/2$  ( $\langle \cdot, \cdot \rangle$  standing for the duality pairing between  $V^{-s}$  and  $V^s$ ), with  $\underline{C} = 0$  if  $\alpha < 0$ .*

*Proof.* If  $0 < \alpha < 1$ , the function  $\phi : z \mapsto e^z \psi(z)$  satisfies the assumptions of Remark 2. Therefore,  $u \mapsto \sqrt{\|u\|_{L^2(\mathbb{R}_+)}^2 + \int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z (u(x) - u(xe^{-z}))^2 dx dz}$  is a norm on  $V^\alpha$  equivalent to the norm  $\|\cdot\|_{V^\alpha}$ . From this and (3.5), we deduce (3.7).  $\square$

Consider the following two situations:

- $1/2 < \alpha < 1$ ,  $\psi$  and  $u \in V^\alpha$ : It can be shown (using the interpolation theorem 7.17 in [6]) that the functions  $u_+$  and  $u_-$  belong to  $V^\alpha$ .
- $\alpha \leq 1/2$  and  $u \in V^s$ ,  $s > 1/2$ .

In both cases,  $\int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z u_-(xe^{-z}) u_+(x) dx dz$  is well defined because

$$\begin{aligned} \int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z u_-(xe^{-z}) u_+(x) dx dz &= \int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z (u_-(xe^{-z}) - u_-(x)) u_+(x) dx dz \\ &\leq C \|u_+\|_{L^2(\mathbb{R}_+)} \sqrt{\int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z (u_-(xe^{-z}) - u_-(x))^2 dx dz} \end{aligned}$$

and is nonnegative. Therefore,

$$\begin{aligned}\langle Bu, u_+ \rangle &= \langle Bu_+, u_+ \rangle - \int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z (u(xe^{-z}) - u_+(xe^{-z})) u_+(x) dx dz \\ &= \langle Bu_+, u_+ \rangle + \int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z u_-(xe^{-z}) u_+(x) dx dz \geq \langle Bu_+, u_+ \rangle.\end{aligned}$$

We have proved the following lemma.

LEMMA 3.8. *If  $(\alpha, \psi)$  satisfy Assumption 1, then there exist two constants  $\underline{C} > 0$  and  $\lambda \geq 0$  such that, for all  $u \in V^\alpha$ , if  $\alpha > 1/2$ , or for all  $u \in V^s$ ,  $s > 1/2$ , if  $\alpha \leq 1/2$ ,*

$$(3.8) \quad \langle Bu, u_+ \rangle \geq \underline{C} |u_+|_{V^\alpha}^2 - \lambda \|u_+\|_{L^2(\mathbb{R}_+)}^2,$$

with  $\underline{C} = 0$  if  $\alpha < 0$ .

**3.2.2. The integro-differential operator.** With  $B$  defined in (2.12), we introduce the integro-differential operator  $A$ ,

$$(3.9) \quad Av = -\frac{\sigma^2 x^2}{2} \frac{\partial^2 v}{\partial x^2} + rx \frac{\partial v}{\partial x} + Bv,$$

where  $\sigma$  and  $r$  are nonnegative real numbers. In this work, we limit ourselves to the case  $\sigma > 0$ . The case  $\sigma = 0$ ,  $\alpha > 1/2$  requires working in the fractional Sobolev spaces described above and will be treated in [1]. Since the space  $V^1$  will play a special role, we use the shorter notation  $V = V^1$ .

If  $\sigma > 0$  and if  $(\alpha, \psi)$  satisfy Assumption 1, then the following hold:

- $A$  is a continuous operator from  $V$  to  $V^{-1}$ .
- We have the Gårding inequalities: there exist  $\underline{c} > 0$  and  $\lambda \geq 0$  such that

$$(3.10) \quad \langle Av, v \rangle \geq \underline{c} |v|_V^2 - \lambda \|v\|_{L^2(\mathbb{R}_+)}^2 \quad \forall v \in V,$$

$$(3.11) \quad \langle Av, v_+ \rangle \geq \underline{c} |v_+|_V^2 - \lambda \|v_+\|_{L^2(\mathbb{R}_+)}^2 \quad \forall v \in V.$$

- The operator  $A + \lambda I$  is one-to-one and continuous from  $V^2$  onto  $L^2(\mathbb{R}_+)$ , with a continuous inverse.

REMARK 5. *The assumption that  $\psi > 0$  near  $z = 0$  is not necessary for  $A$  to have the above properties. Its role is to allow a clear identification of the kernel's singularity at  $z = 0$ .*

**3.3. The variational inequalities.** We are ready to write the variational inequalities corresponding to the LCP (2.16)–(2.19).

We introduce the closed subspace of  $V$ :

$$(3.12) \quad K = \{v \in V, v(x) \geq u_\circ(x) \text{ in } \mathbb{R}_+\}.$$

The variational problem will consist of looking for  $u \in L^2(0, T; V) \cap C^0([0, T]; L^2(\mathbb{R}_+))$ , with  $\frac{\partial u}{\partial t} \in L^2((0, T) \times \mathbb{R}_+)$ , such that the following hold:

1. There exists a constant  $X_T > S$  such that  $u(t, x) = 0$  for all  $t \in [0, T]$ , for all  $x \geq X_T$ .
2.  $u(t) \in K$  for almost every  $t \in (0, T)$ .
3. For almost every  $t \in (0, T)$ , for any  $v \in K$  with bounded support,

$$(3.13) \quad \left\langle \frac{\partial u}{\partial t} + Au + rx, v - u \right\rangle \geq 0,$$

where  $\langle \cdot, \cdot \rangle$  stands for the duality pairing between  $V'$  (the dual of  $V$ ) and  $V$ .

4.  $u(t = 0) = u_\circ$ .

Hereafter, this problem will be referred to as (VIP). The goal of section 4 below is to prove that (VIP) has a unique solution and to study its properties.

#### 4. Analysis of the variational inequalities.

**4.1. Orientation.** Hereafter, we assume that  $\sigma > 0$ . Problem (2.16)–(2.19) is posed in an unbounded domain. This is a technical difficulty in order to use variational methods, and we first have to replace this problem by a similar one posed in a bounded domain. Therefore, the program is to

1. approximate (2.16)–(2.19) by a similar problem posed in  $[0, T] \times [0, X]$ , for some given positive parameter  $X > S$ , and write the related variational problem, which will be called  $(VIP_X)$  below;
2. solve first a penalized version of  $(VIP_X)$  by introducing a semilinear monotone operator, and pass to the limit as the penalty parameter tends to zero;
3. prove that the free boundary of  $(VIP_X)$  stays in a bounded domain as  $X$  tends to infinity; this will show that for  $X$  large enough a solution of  $(VIP_X)$  is actually a solution of (VIP); and
4. obtain estimates for the solution of (VIP) independent of the parameters  $(\sigma, \alpha, \psi)$ , when these parameters vary in a suitably defined set.

**4.2. Approximation of (VIP) in a bounded domain.** Let  $X$  be a positive number greater than  $S$ . Hereafter, for a function  $v \in L^2((0, X))$  we call  $\mathcal{E}_X(v)$  the function in  $L^2(\mathbb{R}_+)$  obtained by extending  $v$  by 0 outside  $(0, X)$ . We introduce the Sobolev spaces  $W_X^1 = \{v \in L^2((0, X)), x \frac{\partial v}{\partial x} \in L^2((0, X))\}$  and  $W_X^2 = \{v \in W_X^1, x^2 \frac{\partial^2 v}{\partial x^2} \in L^2((0, X))\}$ . For  $\beta, 0 < \beta < 1$ ,  $W_X^\beta$  is the space obtained by real interpolation between  $W_X^1$  and  $L^2(0, X)$  with parameter  $\nu = 1/2 - \beta$  (see [6, p. 204], [27]), and  $W_X^{1+\beta} = \{v \in W_X^1, x \frac{\partial v}{\partial x} \in W_X^\beta\}$ .

For  $\beta, 0 \leq \beta < 3/2$ , we introduce  $V_X^\beta = \{v \in L^2(0, X), \mathcal{E}(v) \in V^\beta\}$ , endowed with the norm  $\|v\|_{V_X^\beta} = \|\mathcal{E}_X(v)\|_{V^\beta}$ . Note that for  $\beta, 0 \leq \beta < 1/2$ ,  $V_X^\beta = W_X^\beta$ . Let  $V_X^{-\beta}$  be the dual of  $V_X^\beta$ . Thanks to Lemma 3.1, we know that for  $\beta > 1/2$ , a function  $v \in V_X^\beta$  is continuous in  $[0, X]$  and vanishes at  $X$ .

Since the space  $V_X^1$  will often be used, we introduce the special notation

$$(4.1) \quad V_X = \{v \in L^2(0, X), \mathcal{E}_X(v) \in V\},$$

and  $\|v\|_{V_X} = \|\mathcal{E}_X(v)\|_V$ . We define the operators  $A_X$  and  $B_X, V_X \rightarrow V_X'$ :

$$(4.2) \quad \langle A_X v, w \rangle = \langle A \mathcal{E}_X(v), \mathcal{E}_X(w) \rangle \quad \text{and} \quad \langle B_X v, w \rangle = \langle B \mathcal{E}_X(v), \mathcal{E}_X(w) \rangle.$$

A Gårding inequality for  $A_X$  is deduced from (3.10), with constants independent of  $X$ . We define

$$(4.3) \quad D_X = \{v \in V_X : A_X v \in L^2((0, X))\}.$$

It follows from the Gårding inequality that  $(A_X, D_X)$  is the infinitesimal generator of an analytic semigroup [33]. Proposition 4.1 below contains information on  $D_X$ .

**PROPOSITION 4.1.** *If  $v \in D_X$ , then for any number  $X' < X$ ,  $v|_{(0, X')} \in W_{X'}^2$ .*

*For  $\alpha, 0 < \alpha < 3/4$ ,  $D_X = W_X^2 \cap V_X$ . For  $\alpha, 3/4 \leq \alpha < 1$ , there exists  $\epsilon > 0$  such that  $D_X \subset W_X^{3/2+\epsilon} \cap V_X$ . In any case, if  $v \in D_X$ , then  $\frac{\partial v}{\partial x} \in C^0((0, X])$ .*

*Proof.* See Appendix B.  $\square$

REMARK 6. *It can be proved by lengthy calculations that if  $v(x) = X - x$  (note that  $v \in W_X^2 \cap V_X$ ), then  $A_X v$  behaves like  $(X - x)^{1-2\alpha}$  near  $x = X$ , so  $A_X v \notin L^2((0, X))$  if  $\alpha > 3/4$ .*

We introduce

$$(4.4) \quad K_X = \{v \in V_X, v(x) \geq u_o(x) \text{ in } (0, X)\}.$$

We are going to look for  $u_X \in L^2(0, T; V_X) \cap C^0([0, T]; L^2((0, X)))$ , with  $\frac{\partial u_X}{\partial t} \in L^2((0, T) \times (0, X))$ , such that the following hold:

1.  $u_X(t) \in K_X$  for almost every  $t \in (0, T)$ .
2. For almost every  $t \in (0, T)$ ,

$$(4.5) \quad \left\langle \frac{\partial u_X}{\partial t} + A_X u_X + rx, v - u_X \right\rangle \geq 0$$

for any  $v \in K_X$ . Here  $\langle \cdot, \cdot \rangle$  stands for the duality pairing between  $V_X'$  (the dual of  $V_X$ ) and  $V_X$ .

3.  $\mathcal{E}_X(u_X)(t=0) = u_o$ .

Hereafter, this problem will be referred to as  $(VIP_X)$ . In order to prove that  $(VIP_X)$  has a unique solution, we follow [25] and introduce first a sequence of monotone problems which can be seen as penalized versions of (4.5): find  $u_{X,\epsilon}$  such that

$$(4.6) \quad \begin{aligned} \frac{\partial u_{X,\epsilon}}{\partial t} + A_X u_{X,\epsilon} + rx(1 - 1_{\{x>S\}} \mathcal{V}_\epsilon(u_{X,\epsilon})) &= 0, \quad t \in (0, T], \quad 0 < x < X, \\ u_{X,\epsilon}(t=0, x) &= u_o(x), \quad 0 < x < X, \\ u_{X,\epsilon}(t, X) &= 0, \quad t \in (0, T]. \end{aligned}$$

where  $\mathcal{V}_\epsilon(u) = \mathcal{V}(u/x\epsilon)$  and  $\mathcal{V}$  is a smooth nonincreasing convex function such that

$$\mathcal{V}(0) = 1, \quad \mathcal{V}(u) = 0 \quad \text{for } u \geq 1, \quad 0 \geq \mathcal{V}'(u) \geq -2 \quad \text{for } 0 \leq u \leq 1.$$

In what follows, we call  $u_X^{(E)}$  and  $\underline{u}_X^{(E)}$  the solutions to the linear problems:

$$\begin{aligned} \frac{\partial u_X^{(E)}}{\partial t} + A_X u_X^{(E)} &= 0, \quad \frac{\partial \underline{u}_X^{(E)}}{\partial t} + A_X \underline{u}_X^{(E)} = -rx, \quad t \in (0, T], \quad 0 < x < X, \\ u_X^{(E)}(t=0, x) &= \underline{u}_X^{(E)}(t=0, x) = u_o(x), \quad 0 < x < X, \\ u_X^{(E)}(t, X) &= \underline{u}_X^{(E)}(t, X) = 0, \quad t \in (0, T]. \end{aligned}$$

It can be seen that  $\underline{u}_X^{(E)}(t, 0) = S$  for all  $t \in [0, T]$  and that

$$(4.7) \quad \underline{u}_X^{(E)}(t, x) > S - x \quad \forall (t, x) \in (0, T] \times (0, X].$$

Let  $u^{(E)}$  be the solution of the linear problem:

$$\frac{\partial u^{(E)}}{\partial t} + Au^{(E)} = 0, \quad t \in (0, T], \quad x > 0, \quad u^{(E)}(t=0, x) = u_o(x), \quad x > 0.$$

The function  $u^{(E)}$  is smooth near  $x = 0$  and  $\frac{\partial u^{(E)}}{\partial x}(t, 0) = -1$  for all  $t \geq 0$ .

THEOREM 4.2. *If  $(\alpha, \psi)$  satisfy Assumption 1 and if  $\sigma > 0$ , then (4.6) has a unique weak solution  $u_{X,\epsilon} \in L^2(0, T; V_X) \cap C^0([0, T]; L^2(0, X))$ . It satisfies*

$$(4.8) \quad \underline{u}_X^{(E)} \leq u_{X,\epsilon} \leq u_X^{(E)} \leq u^{(E)}.$$

The function  $u_{X,\epsilon}$  belongs to  $C^0([0, T]; K_X) \cap L^2(0, T; D_X)$  and is continuous and nondecreasing w.r.t.  $t$ . For two positive numbers  $\epsilon' < \epsilon$ , we have

$$(4.9) \quad u_{X,\epsilon'} \leq u_{X,\epsilon} \leq u_{X,\epsilon'} + \epsilon.$$

The quantities  $\|u_{X,\epsilon}\|_{L^\infty(0,T;V_X)}$ ,  $\|u_{X,\epsilon}\|_{L^2(0,T;D_X)}$ , and  $\|\frac{\partial u_{X,\epsilon}}{\partial t}\|_{L^2((0,T)\times(0,X))}$  are bounded independently of  $\epsilon$ . The quantities  $\|u_{X,\epsilon}\|_{L^\infty(0,T;L^2(0,X))}$  and  $\|u_{X,\epsilon}\|_{L^2(0,T;V_X)}$  are bounded independently of  $X$ .

*Proof.* See Appendix B.  $\square$

**THEOREM 4.3.** The function  $x \frac{\partial u_{X,\epsilon}}{\partial x}$  is the sum of  $\tilde{z}_{X,\epsilon} \in C^0([0, T]; L^2(0, X))$  and of  $\hat{z}_{X,\epsilon} \in L^2(0, T; V_X)$  such that  $\tilde{z}_{X,\epsilon} \leq 0$  and  $\lim_{\epsilon \rightarrow 0} \|\hat{z}_{X,\epsilon}\|_{L^2(0,T;V_X)} = 0$ .

Finally, for two numbers  $X$  and  $X'$  such that  $S < X < X'$ , for any  $\epsilon > 0$ ,

$$(4.10) \quad \mathcal{E}_X(u_{X,\epsilon}) \leq \mathcal{E}_{X'}(u_{X',\epsilon}).$$

*Proof.* See Appendix B.  $\square$

**THEOREM 4.4.** If  $(\alpha, \psi)$  satisfy Assumption 1 and if  $\sigma > 0$ ,  $(VIP_X)$  has a unique solution  $u_X \in C^0([0, T]; K_X) \cap L^2(0, T; D_X)$ , with  $\frac{\partial u_X}{\partial t} \in L^2((0, T) \times (0, X))$ .

The function  $u_X$  is continuous in  $[0, T] \times [0, X]$ , with  $u_X(t, 0) = S$  for all  $t \in [0, T]$ ,  $\underline{u}_X^{(E)} \leq u_X \leq u_X^{(E)} \leq u^{(E)}$ , and  $u_X(t, x) > u_o(x)$  for  $0 < t \leq T$  and  $0 < x \leq S$ .

The function  $u_X$  is nondecreasing w.r.t.  $t$  and nonincreasing w.r.t.  $x$ . The quantities  $\|\mathcal{E}_X(u_X)\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$  and  $\|\mathcal{E}_X(u_X)\|_{L^2(0,T;V)}$  are bounded independently of  $X$ .

For  $\epsilon > 0$ , we have the bounds

$$(4.11) \quad u_X \leq u_{X,\epsilon} \leq u_X + \epsilon,$$

and the sequence  $u_{X,\epsilon}$  converges to  $u_X$  uniformly as  $\epsilon \rightarrow 0$ .

For two numbers  $X$  and  $X'$  such that  $S < X < X'$ ,  $\mathcal{E}_X(u_X) \leq \mathcal{E}_{X'}(u_{X'})$ .

*Proof.* The proof consists mainly of passing to the limit in (4.6) as  $\epsilon \rightarrow 0$ . It uses the Minty trick; see [25]. We skip it since it is rather classical.  $\square$

**LEMMA 4.5.** If  $(\alpha, \psi)$  satisfy Assumption 1 and if  $\sigma > 0$ , there exists a nondecreasing function  $\gamma_X : (0, T] \rightarrow (S, X]$ , such that the set  $\{(t, x) : u_X(t, x) = u_o(x)\}$  coincides with the set  $\{(t, x) : x \geq \gamma_X(t)\}$ . Calling

$$(4.12) \quad \mu_X = \frac{\partial u_X}{\partial t} + A_X u_X + r x,$$

we have a.e.

$$(4.13) \quad 0 \leq \mu_X \leq r x 1_{\{u_X=0\}} = r x 1_{\{x \geq \gamma_X(t)\}}.$$

*Proof.* We know that for all  $t \in [0, T]$ ,  $u_X(t, X) = u_o(X) = 0$ . Thus, at each time  $t$ , the set where  $u_X(t, x)$  coincides with  $u_o$  is nonempty. It is closed since  $u_X$  and  $u_o$  are continuous. We also know that  $u_X(t, x) > u_o(x)$  for  $t > 0$  and  $x \leq S$ ; thus,  $\{x > 0$  such that  $u_X(t, x) = u_o(x)\} \subset (S, X]$  for  $t > 0$ . On the other hand, for all  $t \in (0, T]$ , the function  $u_X(t)$  is nonincreasing w.r.t.  $x$ , so  $\{x > 0$  such that  $u_X(t, x) = u_o(x)\}$  is an interval  $[\gamma_X(t), X]$ , with  $\gamma_X(t) > S$ . Since  $u_X$  is nondecreasing w.r.t.  $t$ , the function  $\gamma_X$  is nondecreasing.

With  $\mu_X \in L^2((0, T) \times \mathbb{R}_+)$  given by (4.12), we have  $\mu_X = 0$  a.e. in the open region where  $u_X > 0$ . Now,  $\mu_X$  is the weak limit of  $r x 1_{x>S} \mathcal{V}_\epsilon(u_{X,\epsilon})$  in  $L^2((0, T) \times (0, X))$ . From (4.11), we deduce that  $r x 1_{x>S} \mathcal{V}_\epsilon(u_{X,\epsilon}) \leq r x 1_{x>S} \mathcal{V}_\epsilon(u_X)$ , and  $1_{x>S} \mathcal{V}_\epsilon(u_X)$  converges pointwise to  $1_{\{u_X=0\}}$ . Therefore,  $\mu_X \leq r x 1_{\{u_X=0\}}$ .  $\square$

PROPOSITION 4.6. *If  $(\alpha, \psi)$  satisfy Assumption 1 and if  $\sigma > 0$ , the function  $\gamma_X$  is nondecreasing and lower semicontinuous. The graph of  $\gamma_X$  has measure 0 (Lebesgue measure in  $\mathbb{R}^2$ ) and*

$$(4.14) \quad \begin{aligned} \mu_X(t, x) &= 1_{\{u_X(t, x)=0\}}(rx + B_X u_X(t, x)) \\ &= 1_{\{u_X(t, x)=0\}} \left( rx - \int_{\mathbb{R}} k(z) e^z u_X(t, x e^{-z}) dz \right) \quad \text{for a.a. } t, x. \end{aligned}$$

*Proof.* We have already seen that  $\gamma_X$  is nondecreasing. The epigraph of  $\gamma_X$  is the set where  $u_X$  vanishes. This region is closed since  $u_X$  is continuous.

Since  $\gamma_X$  has left and right limits at each point  $t$ , the graph of  $\gamma_X$  has measure 0 (Lebesgue measure in  $\mathbb{R}^2$ ); see Theorem 3.7 in [2] for the proof. As a consequence, the boundary of the coincidence set  $\{u_X = 0\}$  has measure 0 (Lebesgue measure in  $\mathbb{R}^2$ ). From this and since the identity  $\mu_X(t, x) = rx - \int_{\mathbb{R}} k(z) e^z u_X(t, x e^{-z}) dz$  is true in the set  $\{x > \gamma_X(t)\}$ , we obtain (4.14).  $\square$

REMARK 7. *We will not try to obtain further regularity results on  $\gamma_X$ . Yet, this is certainly an interesting topic on which little seems to be known.*

Let  $T_X$  be defined by

$$(4.15) \quad T_X = \sup\{t, 0 < t \leq T, \gamma_X(t) < X\}.$$

Since  $\gamma_X$  is nondecreasing, we know that if  $T_X < T$ , then for all  $t \in [T_X, T]$ ,  $\gamma_X(t) = X$ .

Note that  $\mathcal{E}_X(u_X)$  is a solution of (2.16)–(2.19) in  $(0, T_X) \times \mathbb{R}_+$ , so for all  $X' > X$ ,  $\mathcal{E}_X(u_X)$  coincides with  $\mathcal{E}_{X'}(u_{X'})$  for  $0 < t < T_X$ . In particular, this implies that  $X \mapsto T_X$  is a nondecreasing function.

LEMMA 4.7. *If  $(\alpha, \psi)$  satisfy Assumption 1 and if  $\sigma > 0$ , there exists  $X_T > S$  such that for all  $X \geq X_T$ ,  $T_X = T$ . For  $X > X_T$ ,  $u_X \in L^2(0, T; W_X^2)$ .*

*Proof.* The proof is by contradiction: if  $X_T$  does not exist, then  $\lim_{X \rightarrow \infty} T_X = \underline{T} < T$ . We have  $\frac{\partial u_X}{\partial t} + A_X u_X = -rx$  in  $[\underline{T}, T] \times (0, X)$  for all  $X > S$ . We choose a smooth and nonnegative function  $\phi$  defined on  $\mathbb{R}$  with compact support, and for  $y > 0$ , we call  $\phi_y$  the function  $\phi_y(x) = \phi(x - y)/\sqrt{y}$ . Then we take  $\phi_y$  as a test function in (4.12). We have

$$\int_0^X (u_X(T, x) - u_X(\underline{T}, x)) \phi_y(x) + \int_{\underline{T}}^T \langle A_X u_X(t), \phi_y \rangle dt = -r \int_{\underline{T}}^T \int_0^X x \phi_y(x) dx.$$

Take  $y = X/2$  and let  $X$  tend to  $\infty$ . From the bounds on  $u_X$ , the left-hand side in the identity above remains bounded, whereas the right-hand side tends to infinity. We have obtained the desired contradiction. The last statement of Lemma 4.7 follows easily from the first statement of Proposition 4.1.  $\square$

PROPOSITION 4.8. *If  $(\alpha, \psi)$  satisfy Assumption 1 and if  $\sigma > 0$ , the function  $\mu_{X, \epsilon} = rx 1_{\{x > S\}} \mathcal{V}_{\epsilon}(u_{X, \epsilon})$ , converges to  $\mu_X$  in  $L^p((0, T) \times (0, X))$  for  $p, 1 \leq p < +\infty$ . The sequence  $u_{X, \epsilon}$  converges to  $u_X$  strongly in  $L^2(0, T; D_X)$  and in  $L^\infty(0, T; V_X)$ .*

*Proof.* See Appendix B.  $\square$

**4.3. The problem (VIP).** From Theorem 4.4, Proposition 4.6, and Lemma 4.7, we can pass to the limit as  $X \rightarrow \infty$ .

THEOREM 4.9. *If  $(\alpha, \psi)$  satisfy Assumption 1 and if  $\sigma > 0$ , there exists a unique solution of problem (VIP), i.e., a function  $u \in C^0([0, T]; K) \cap L^2(0, T; V^2)$ , with  $\frac{\partial u}{\partial t} \in L^2((0, T) \times \mathbb{R}_+)$ , such that  $u(t = 0) = u_\circ$ ,*

$$(4.16) \quad u(t, x) = 0 \quad \forall t \in [0, T], x \geq X_T,$$



where  $X_T$  is defined in Lemma 4.7, and satisfying the variational inequality (3.13) for all  $v \in K$  with bounded support in  $x$ . The function  $u$  coincides with  $u_X$  for  $X \geq X_T$ . There exists a nondecreasing and lower semicontinuous function  $\gamma : (0, T] \rightarrow (S, X_T)$ , such that for all  $t \in (0, T)$ ,  $\{x > 0 \text{ such that } u(t, x) = u_o(x)\} = [\gamma(t), +\infty)$ . Calling

$$(4.17) \quad \mu = \frac{\partial u}{\partial t} + Au + rx,$$

we have a.e.  $0 \leq \mu \leq rx1_{\{u=0\}} = rx1_{\{x \geq \gamma(t)\}}$ .

**PROPOSITION 4.10.** *The function  $\mu$  defined in (4.17) is nondecreasing w.r.t.  $x$  (i.e., the distribution  $\frac{\partial \mu}{\partial x}$  is negative) and nonincreasing w.r.t.  $t$ , (i.e., the distribution  $\frac{\partial \mu}{\partial t}$  is positive). For any  $X > X_T$ , the total variation of  $\mu$  in  $(0, T) \times (0, X)$  is bounded by  $rX(T + X)$ .*

*Proof.* Consider  $X > X_T$ . The function  $\mu$  coincides with  $\mu_X$  on  $(0, T) \times (0, X)$ . The monotone character of  $\mu$  w.r.t. the two variables stems from (4.14) and from the fact that  $u$  is nonincreasing w.r.t.  $x$  and nondecreasing w.r.t.  $t$ .

The same result can be proved by observing that  $\mu_X$  is the weak limit in  $L^2((0, T) \times (0, X))$  of the sequence  $rx1_{x>S}\mathcal{V}_\epsilon(u_{X,\epsilon})$  and using the properties of  $rx1_{x>S}\mathcal{V}_\epsilon(u_{X,\epsilon})$ .

The bound on the total variation of  $\mu$  on  $(0, T) \times (0, X)$  comes from the fact that  $\mu$  is nondecreasing w.r.t.  $x$  and nonincreasing w.r.t.  $t$  and that  $0 \leq \mu \leq rX$  a.e. in  $(0, T) \times (0, X)$ .  $\square$

**PROPOSITION 4.11.** *A.e. in the coincidence set  $\{(t, x) : u(t, x) = 0\}$ ,  $\mu > 0$ .*

*Proof.* We know from Proposition 4.6 that the boundary of the coincidence set has measure 0 (Lebesgue measure in  $\mathbb{R}^2$ ). Assume that  $\mu = 0$  in some subset of  $x > \gamma(t)$  with positive measure. In view of the monotone behavior of  $\mu$ , this implies that  $\mu = 0$  in a rectangle contained in the set  $x > \gamma(t)$ . From Proposition 4.6, this implies that  $\int_{\mathbb{R}} k(z)e^z u(t, xe^{-z}) dz = rx$  in this rectangle. Taking the derivative w.r.t.  $x$ , we obtain that  $\int_{\mathbb{R}} k(z) \frac{\partial u}{\partial x}(t, xe^{-z}) dz = r$  in the rectangle. But this is impossible, since  $u(t, x)$  is nonincreasing w.r.t.  $x$  and nonidentically 0.  $\square$

**REMARK 8.** *Proposition 4.11 tells us that there is a.e. strict complementarity: the reaction term  $\mu$  is positive at almost every point where  $u = 0$ .*

**4.4. Further bounds.** Let us choose some constants  $\underline{\sigma}, \bar{\sigma}, \underline{\alpha}, b_1, b_2, \underline{\psi}, \bar{\psi}$ , and  $\bar{z}$  such that  $0 < \underline{\sigma} \leq \bar{\sigma}$ ,  $0 < \underline{\alpha} < 1/2$ ,  $b_1 > 1$ ,  $b_2 > 1$ ,  $\bar{\psi} \geq \underline{\psi} > 0$ , and  $\bar{z} > 0$ . Let us define the subset  $\mathcal{F}$  of  $\mathbb{R}_+ \times \mathbb{R} \times L^\infty(\mathbb{R})$  by

$$(4.18) \quad \mathcal{F} = [\underline{\sigma}, \bar{\sigma}] \times [-1/2, 1 - \underline{\alpha}] \times \left\{ \psi : \begin{array}{l} \|\max(e^{2b_1 z}, |z|^{b_2}, 1)\psi\|_{L^\infty(\mathbb{R})} \leq \bar{\psi}; \\ \psi \geq 0, \psi \geq \underline{\psi} \text{ a.e. in } [-\bar{z}, \bar{z}] \end{array} \right\}.$$

We can make the following three observations:

1. The norm of  $A$  as an operator from  $V$  to  $V'$  is bounded independently of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ .
2. The constants in (3.10)–(3.11) can be taken independent of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ .
3. With  $\lambda$  in (3.10) independent of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ , the operator  $A + \lambda I$  is one-to-one and continuous from  $V^2$  onto  $L^2(\mathbb{R}_+)$  and  $(A + \lambda I)^{-1} : L^2(\mathbb{R}_+) \mapsto V^2$  is bounded with constants independent of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ .

By carefully inspecting the proofs of Theorems 4.2, 4.4, and 4.9, we see that

1. the quantities  $\|u_{X,\epsilon}\|_{L^\infty(0,T;L^2(0,X))}$  and  $\|u_{X,\epsilon}\|_{L^2(0,T;V_X)}$  are bounded independently of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ ,
2. the quantities  $\|\mathcal{E}_X(u_X)\|_{L^\infty(0,T;L^2(\mathbb{R}_+)})$  and  $\|\mathcal{E}_X(u_X)\|_{L^2(0,T;V)}$  are bounded independently of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ , and

3. the quantities  $\|u\|_{L^\infty(0,T;L^2(\mathbb{R}_+))}$  and  $\|u\|_{L^2(0,T;V)}$  are bounded independently of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ .

**PROPOSITION 4.12.** *The function  $\gamma$  is bounded in  $[0, T]$  by some constant  $\bar{X}$  independent of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ . The quantities  $\|u\|_{L^\infty(0,T;V)}$ ,  $\|u\|_{L^2(0,T;V^2)}$ , and  $\|\frac{\partial u}{\partial t}\|_{L^2((0,T)\times\mathbb{R}_+)}$  are bounded independently of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ .*

*Proof.* For a sequence  $(\sigma_n, \alpha_n, \psi_n)$  in  $\mathcal{F}$ , let us call  $u_n$  the corresponding solution of problem (VIP) and  $\gamma_n$  the function such that  $u_n(t, x) = 0 \Leftrightarrow x \geq \gamma_n(t)$ . Assume that  $\lim_{n \rightarrow \infty} \gamma_n(T/2) = +\infty$ . Then, we can use the same arguments as in the proof of Lemma 4.7 and reach a contradiction. Therefore,  $\gamma|_{[0,T/2]}$  is bounded independently of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ . Since (VIP) can always be solved in  $(0, 2T) \times \mathbb{R}_+$  instead of  $(0, T) \times \mathbb{R}_+$ , and since for the solution  $u$ ,  $\|u\|_{L^2(0,2T;V)} + \|\frac{\partial u}{\partial t}\|_{L^2(0,2T;V')}$  is bounded independently of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ , we can use the same arguments and prove that  $\gamma|_{[0,T]}$  is bounded independently of  $(\sigma, \alpha, \psi)$  in  $\mathcal{F}$ .

Therefore, it is possible to choose  $\bar{X}$  such that, for any  $(\sigma, \alpha, \psi) \in \mathcal{F}$ ,  $\gamma < \bar{X}$ , and  $u$  coincides with  $\mathcal{E}_{\bar{X}}(u_{\bar{X}})$ , where  $u_{\bar{X}}$  is the solution of (VIP) $_{\bar{X}}$ . For  $x > \bar{X}$ ,

$$\mu(t, x) = rx - \int_{z > \log(x/\bar{X})} k(z) e^z u(xe^{-z}) dz.$$

Thus, for  $x$  large enough such that, for example,  $\log(x/\bar{X}) > 1$ ,

$$\begin{aligned} 0 \leq rx - \mu(t, x) &\leq S \int_{z > \log(x/\bar{X})} k(z) e^z dz \leq S \int_{z > \log(x/\bar{X})} \psi(z) e^{2z} e^{-z} dz \\ (4.19) \quad &\leq S \bar{\psi} \int_{z > \log(x/\bar{X})} e^{-z} dz = \bar{\psi} S \frac{\bar{X}}{x}. \end{aligned}$$

Therefore,  $\|\frac{\partial u}{\partial t} + Au\|_{L^2(0,T;L^2(\mathbb{R}_+))}$  is bounded by a constant independent of  $(\sigma, \alpha, \psi)$ . This implies that the quantities  $\|u\|_{L^\infty(0,T;V)}$ ,  $\|u\|_{L^2(0,T;V^2)}$ , and  $\|\frac{\partial u}{\partial t}\|_{L^2((0,T)\times\mathbb{R})}$  are bounded independently of  $(\sigma, \alpha, \psi) \in \mathcal{F}$ .  $\square$

**REMARK 9.** *It may be possible to impose weaker conditions on  $\psi$ , and other choices of  $\mathcal{F}$  could be made.*

**5. Sensitivity analysis.** Here, we aim at understanding the sensitivity of the solution  $u$  of (VIP) and of  $\mu$  given by (4.17) to the variations of  $(\sigma, \alpha, \psi) \in \mathcal{F}$ . Let us introduce  $\mathcal{B} = \{f : z \mapsto f(z) \max(1, |z|^{b_2}, e^{2b_1 z}) \in L^\infty(\mathbb{R})\}$  endowed with the norm  $\|f\|_{\mathcal{B}} = \|f(\cdot) \max(1, |\cdot|^{b_2}, e^{2b_1 \cdot})\|_{L^\infty(\mathbb{R})}$ . For  $(\sigma, \alpha, \psi) \in \mathcal{F}$ , let  $u(\sigma, \alpha, \psi)$  be the corresponding solution of (VIP). Accordingly, let  $\mu(\sigma, \alpha, \psi)$  be given by (4.17) and  $\gamma(\sigma, \alpha, \psi)$  be the function defining the free boundary.

**PROPOSITION 5.1.** *There exists  $C > 0$  such that  $(\sigma, \alpha, \psi) \in \mathcal{F}$ ,  $(\tilde{\sigma}, \tilde{\alpha}, \tilde{\psi}) \in \mathcal{F}$ ,*

$$(5.1) \quad \|u - \tilde{u}\|_{L^2(0,T;V)} + \|u - \tilde{u}\|_{L^\infty(0,T;L^2(\mathbb{R}_+))} \leq C \left( |\sigma - \tilde{\sigma}| + |\alpha - \tilde{\alpha}| + \|\psi - \tilde{\psi}\|_{\mathcal{B}} \right),$$

$$(5.2) \quad \int_0^T \int_{\mathbb{R}} (\mu(\tilde{u} - u_o) + \tilde{\mu}(u - u_o)) \leq C \left( |\sigma - \tilde{\sigma}| + |\alpha - \tilde{\alpha}| + \|\psi - \tilde{\psi}\|_{\mathcal{B}} \right)^2,$$

calling  $u = u(\sigma, \alpha, \psi)$ ,  $\mu = \mu(\sigma, \alpha, \psi)$ ,  $\tilde{u} = u(\tilde{\sigma}, \tilde{\alpha}, \tilde{\psi})$ , and  $\tilde{\mu} = \mu(\tilde{\sigma}, \tilde{\alpha}, \tilde{\psi})$ .

*Proof.* We skip the proof since its arguments are well known.  $\square$

**PROPOSITION 5.2.** *Consider  $(\sigma, \alpha, \psi) \in \mathcal{F}$  and let  $(\sigma_n, \alpha_n, \psi_n)_{n \in \mathbb{N}}$  be a sequence of coefficients in  $\mathcal{F}$  such that  $\lim_{n \rightarrow \infty} (|\sigma - \sigma_n| + |\alpha - \alpha_n| + \|\psi - \psi_n\|_{\mathcal{B}}) = 0$ . Calling*

$$(5.3) \quad \lim_{n \rightarrow +\infty} \|u_n - u\|_{L^\infty((0,T) \times \mathbb{R}_+)} = 0, \quad \lim_{n \rightarrow +\infty} \|\mu_n - \mu\|_{L^p((0,T) \times \mathbb{R}_+)} = 0$$

for all  $p$ ,  $1 < p < +\infty$ , and

$$(5.4) \quad \|u_n - u\|_{L^\infty(0,T;V^1)} + \|u_n - u\|_{L^2(0,T;V^2)} + \left\| \frac{\partial(u_n - u)}{\partial t} \right\|_{L^2((0,T) \times \mathbb{R}_+)} \rightarrow 0.$$

*Proof.* From the facts that

- $u(t, x) = u_n(t, x) = 0$  for  $x > \bar{X}$  (where  $\bar{X}$  is given in Proposition 4.12 and does not depend of  $(\sigma, \alpha, \psi) \in \mathcal{F}$ ), and
- for all  $n$ ,  $S - x \leq u_n(t, x) \leq S$  and  $S - x \leq u(t, x) \leq S$ , which implies that  $u - u_n$  is arbitrarily small as  $x \rightarrow 0$  uniformly w.r.t.  $n$ ,

it is enough to prove that for all  $\epsilon > 0$ ,

$$(5.5) \quad \lim_{n \rightarrow +\infty} \|u_n - u\|_{L^\infty((0,T) \times (\epsilon, \bar{X}))} = 0.$$

From (5.1), we see that  $\lim_{n \rightarrow \infty} \|u_n - u\|_{L^\infty(0,T;L^2(\mathbb{R}_+))} = 0$ . On the other hand, we know that  $\|u_n - u\|_{L^\infty(0,T;V)}$  is bounded independently of  $n$ . These two observations imply (5.5), and the first part of (5.3) is proved.

Let us prove the second part of (5.3): from the fact that  $\mu_n - rx$  is bounded  $L^2((0, T) \times \mathbb{R}_+)$ , one can extract a subsequence converging weakly in  $L^2((0, T) \times \mathbb{R}_+)$ . The limit is nothing else but  $\mu - rx$ , and the whole sequence  $\mu_n - rx$  converges to  $\mu - rx$  weakly in  $L^2((0, T) \times \mathbb{R}_+)$ . Thanks to (4.19) with  $\bar{X}$  independent of  $(\sigma, \alpha, \psi) \in \mathcal{F}$ , it is enough to prove that  $\mu_n$  strongly converges to  $\mu$  in  $L^p((0, T) \times (0, X))$ , for any  $X > S$ , and  $1 < p < +\infty$ . But, from Proposition 4.10, we know that the sequence  $(\mu_n)_n$  is bounded in  $BV((0, T) \times (0, X))$  and in  $L^\infty((0, T) \times (0, X))$ , and therefore relatively compact in  $L^p((0, T) \times (0, X))$ ,  $1 \leq p < +\infty$ . Therefore, a subsequence of  $(\mu_n)_n$  converges in  $L^p((0, T) \times (0, X))$ , and the limit is nothing but  $\mu$ , from the observation above. The whole sequence  $(\mu_n)_n$  converges to  $\mu$  in  $L^p((0, T) \times (0, X))$ . We have proved that  $\mu_n - rx$  converges to  $\mu - rx$  in  $L^p((0, T) \times \mathbb{R}_+)$ ,  $1 < p < +\infty$ .

Finally, (5.4) follows from (5.3).  $\square$

**6. The least square inverse problem.** Let us introduce a Hilbert space  $H_\psi$  endowed with the norm  $\|\cdot\|_{H_\psi}$  and such that the bounded subsets of  $H_\psi$  are relatively compact in  $\mathcal{B}$ .

Consider  $\mathcal{H}_\psi$  a closed and convex subset of  $H_\psi$ . We assume that  $\mathcal{H}_\psi$  is contained in  $\{\psi : \|\psi\|_{\mathcal{B}} \leq \bar{\psi}; \psi \geq 0\}$  and that (a) the functions  $\psi \in \mathcal{H}_\psi$  are continuous near 0, (b) there exist two positive constants  $\underline{\psi}$  and  $\bar{z}$  such that  $\psi(z) \geq \underline{\psi}$  for all  $z$  such that  $|z| \leq \bar{z}$ , and (c) there exist two constants  $\zeta > 0$  and  $C \geq 0$  such that for all  $\psi \in \mathcal{H}_\psi$ ,  $\psi(z)e^{3/2z} - \psi(0)e^{-3/2z} = z\omega(z)$ , with  $|\omega(z)| \leq C|z|e^{-\zeta|z|}$ , for all  $z \in \mathbb{R}$ . This choice of  $\mathcal{H}_\psi$  will allow us to use the results stated in Lemma 3.6.

Finally, consider the set  $\mathcal{H} = [\underline{\sigma}, \bar{\sigma}] \times [-1/2, 1 - \underline{\alpha}] \times \mathcal{H}_\psi$ . Let  $J_R$  be a convex, coercive, and  $\mathcal{C}^1$  function defined on  $[\underline{\sigma}, \bar{\sigma}] \times [-1/2, 1 - \underline{\alpha}] \times H_\psi$ . It is well known that  $J_R$  is also weakly lower semicontinuous. The functional  $J_R$  may depend on suitable prior parameters  $\sigma_0$ ,  $\alpha_0$ , and  $\psi_0$ . It is the analogue of the function  $\rho J_2$  discussed in section 1.

### 6.1. Toward the calibration problem.

**6.1.1. Orientation.** For calibrating the Lévy process, one observes the spot price  $S$  and the prices  $(\bar{p}_i)_{i \in I}$  of a family of American put options with maturities/ strikes given by  $(T_i, x_i)$ , and we call  $\bar{u}_i = \bar{p}_i - x_i + S$ ,  $i \in I$ . The parameters of the

Lévy process, i.e., the volatility  $\sigma$ , the exponent  $\alpha$ , and the function  $\psi$  will be found as solutions of a least square problem, where the functional to be minimized is the sum of a suitable Tychonoff regularization functional and of

$$J(u) = \sum_{i \in I} \omega_i (u(T_i, x_i) - \bar{u}_i)^2,$$

where  $\omega_i$  are positive weights, and  $u = u(\sigma, \alpha, \psi)$  is a solution of (VIP).

We aim at finding some necessary optimality conditions satisfied by the solutions of the least square problem. The main difficulty comes from the fact that the derivability of the functional  $J(u)$  w.r.t. the parameter  $(\sigma, \alpha, \psi)$  is not guaranteed. To obtain some necessary optimality conditions, we shall consider first a least square problem where  $u$  is the solution of the penalized problem (4.6) rather than (VIP), obtain the necessary optimality conditions for this new problem, and then have the penalty parameter  $\epsilon$  tend to 0 and pass to the limit in the optimality conditions. Such a program has already been applied in [2] for calibrating the local volatility with American options; see also [4, 5] for a related numerical method and results. The idea originally comes from Hintermüller [22] and Ito and Kunisch [23], who applied a similar program for elliptic variational inequalities. Let us also mention Mignot and Puel [31], who applied a nice method for finding the optimality conditions of a special control problem with a parabolic variational inequality.

In order to simplify the notation, we are going to consider first a toy problem where only one price is observed. Of course, observing only one price is not enough. However, finding the optimality conditions for this simplified calibration problem presents the same difficulties as for the original one.

**6.1.2. The least square problem and its penalized version.** A first step toward the calibration problem is to consider the functional  $J$ ,

$$(6.1) \quad J : \mathcal{C}^0([0, T] \times \mathbb{R}_+) \rightarrow \mathbb{R}, \quad J(u) = (u(T, x_{ob}) - \bar{u})^2,$$

where  $x_{ob}$  and  $\bar{u}$  are positive numbers. We fix  $\bar{X}$  (independent of  $(\sigma, \alpha, \psi) \in \mathcal{H}$ ) as in Proposition 4.12 and assume that  $x_{ob} < \bar{X}$ . Consider the least square problem:

$$(6.2) \quad \text{Minimize } J(u) + J_R(\sigma, \alpha, \psi) \quad \left| \quad (\sigma, \alpha, \psi) \in \mathcal{H}, u = u(\sigma, \alpha, \psi) \text{ satisfies (VIP)} \right|.$$

Fixing  $X \geq \bar{X}$ , we know that  $u|_{[0, T] \times [0, X]} = u_X$ , where  $u_X$  is the solution of (VIP<sub>X</sub>). Therefore, (6.2) is equivalent to the least square problem:

$$(6.3) \quad \text{Minimize } J(u) + J_R(\sigma, \alpha, \psi) \quad \left| \quad (\sigma, \alpha, \psi) \in \mathcal{H}, u \text{ satisfies (VIP}_X\text{)} \right|.$$

We will also consider the least square problem related to the penalized problem:

$$(6.4) \quad \text{Minimize } J(u_\epsilon) + J_R(\sigma, \alpha, \psi) \quad \left| \quad (\sigma, \alpha, \psi) \in \mathcal{H}, u_\epsilon \text{ satisfies (4.6)} \right|.$$

**LEMMA 6.1.** *Let  $(\epsilon_n)_n$  be a sequence of penalty parameters such that  $\epsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ , and let  $(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*), u_{\epsilon_n}^*$  be a solution of problem (6.4). Consider a subsequence such that  $(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*)$  converges to  $(\sigma^*, \alpha^*, \psi^*)$  in  $\mathcal{F}$ ,  $\psi_{\epsilon_n}^*$  weakly converges to  $\psi^*$  in  $H_\psi$ , and  $u_{\epsilon_n}^* \rightarrow u^*$  weakly in  $L^2(0, T; V_X)$ , where  $V_X$  is defined in (4.1). Then  $(\sigma^*, \alpha^*, \psi^*), u^*$  is a solution of (6.3). We have that*

- $u_{\epsilon_n}^*$  converges to  $u^*$  uniformly in  $[0, T] \times [0, X]$ , and in  $L^2(0, T; V_X)$ ,
- $1_{\{x>S\}}rx\mathcal{V}_{\epsilon_n}(u_{\epsilon_n}^*)$  converges to  $\mu^*$  strongly in  $L^2((0, T) \times (0, X))$ , and
- for all smooth functions  $\chi$  with compact support contained in  $[0, X]$ ,  $\chi\mathcal{E}_X(u_{\epsilon_n}^*)$  converges to  $\chi\mathcal{E}_X(u^*)$  strongly in  $L^2(0, T; V^2)$  and in  $L^\infty(0, T; V)$ .

*Proof.* For brevity, the proof is outlined only. We skip the proof that  $u^*$  satisfies (VIP<sub>X</sub>) with  $(\sigma, \alpha, \psi) = (\sigma^*, \alpha^*, \psi^*)$  and the proofs of the first two points above, since they are in the same spirit as the proofs of Theorem 4.4 and Proposition 4.8. The third point above is proved by writing the boundary value problems satisfied by  $y_n = \chi\mathcal{E}_X(u_{\epsilon_n}^*)$  and  $y = \chi\mathcal{E}_X(u^*)$ , with the PID equations

$$\frac{\partial y_n}{\partial t} + A_n y_n = f_n, \quad \frac{\partial y}{\partial t} + Ay = f,$$

where  $A$  (resp.,  $A_n$ ) is given by (3.9) and (2.12) with  $(\sigma, \alpha, \psi) = (\sigma^*, \alpha^*, \psi^*)$  (resp.,  $(\sigma, \alpha, \psi) = (\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*)$ ) and where the right-hand side  $f$  (resp.,  $f_n$ ) can be written in terms of  $\chi$ ,  $u^*$ , and  $\mu^*$  (resp.,  $\chi$  and  $u_{\epsilon_n}^*$ ). By using the first two points above and the same arguments as in the proofs of Propositions 5.1 and 5.2, it can be proved that  $f_n$  converges to  $f$  in  $L^2((0, T) \times \mathbb{R}_+)$  and that  $y_n$  converges to  $y$  in  $L^2(0, T; V^2)$  and in  $L^\infty(0, T; V)$ .

As a consequence of the first point above,  $J(u_{\epsilon_n}^*) \rightarrow J(u^*)$ . Moreover, from the assumptions on  $J_R$ ,  $J_R(\sigma^*, \alpha^*, \psi^*) \leq \liminf_{n \rightarrow \infty} J_R(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*)$ .

Since  $(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*)$  is a solution of (6.4),

$$J(u_{\epsilon_n}^*) + J_R(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*) \leq J(u_{\epsilon_n}(\sigma, \alpha, \psi)) + J_R(\sigma, \alpha, \psi) \quad \forall (\sigma, \alpha, \psi) \in \mathcal{H},$$

where  $u_{\epsilon_n}(\sigma, \alpha, \psi)$  is the solution of (4.6) with  $\epsilon = \epsilon_n$ . This implies that

$$J(u^*) + J_R(\sigma^*, \alpha^*, \psi^*) \leq J(u(\sigma, \alpha, \psi)) + J_R(\sigma, \alpha, \psi) \quad \forall (\sigma, \alpha, \psi) \in \mathcal{H},$$

where  $u(\sigma, \alpha, \psi)$  satisfies (VIP<sub>X</sub>) and  $(\sigma^*, \alpha^*, \psi^*), u^*$  is a solution of (6.3).  $\square$

**REMARK 10.** Let  $(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*), u_{\epsilon_n}^*$  be a subsequence converging to  $(\sigma^*, \alpha^*, \psi^*), u^*$  as in Lemma 6.1. It is clear from the continuity of  $u^*$  and from the uniform convergence of  $u_{\epsilon_n}^*$  that if  $u^*(T, x_{ob}) > u_o(x_{ob})$ , then there exist a constant  $a > 0$  and an integer  $N$  such that for  $n > N$ ,  $u_{\epsilon_n}^*(t, x) > u_o(x) + \epsilon_n$  for all  $(t, x)$  with  $|x - x_{ob}| < a$  and  $t > T - a$ .

**6.1.3. First order necessary optimality conditions for (6.4).** We take  $(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*), u_{\epsilon_n}^*$  and  $(\sigma^*, \alpha^*, \psi^*), u^*$  as in Lemma 6.1. We assume that  $u^*(T, x_{ob}) > u_o(x_{ob})$ , and we take  $N$  and  $a$  as in Remark 10. For  $n > N$ , we wish to find necessary optimality conditions for the solution  $(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*), u_{\epsilon_n}^*$  of (6.4). In order to simplify the notation, we drop the index  $n$ ; below,  $\epsilon$  means  $\epsilon_n$ .

We shall need to solve an adjoint problem. Since the cost functional involves pointwise values of  $u$ , the adjoint problem will have a singular data. In that context, the notion of very weak solution of boundary value problems will be relevant: we introduce the space  $Z_\epsilon = \{v \in \tilde{Z}_\epsilon; v(t = 0) = 0\}$ , where

$$\tilde{Z}_\epsilon = \left\{ v \in L^2(0, T; V_X); \frac{\partial v}{\partial t} + A_{\epsilon, X} v - rx1_{\{x>S\}}\mathcal{V}'(u_\epsilon^*)v \in L^2((0, T) \times (0, X)) \right\}$$

and  $A_{\epsilon, X}$  is the operator defined by (4.2), with  $(\sigma, \alpha, \psi) = (\sigma_\epsilon^*, \alpha_\epsilon^*, \psi_\epsilon^*)$ . The space  $Z_\epsilon$  endowed with the graph norm is a Banach space.

LEMMA 6.2. Assume that  $u^*(T, x_{ob}) > u_o(x_{ob})$  and take  $N$  and  $a$  as in Remark 10. There exists a unique  $p_\epsilon^* \in L^2((0, T) \times (0, X))$  such that, for all  $v \in Z_\epsilon$ ,

$$(6.5) \quad \int_0^T \int_0^X \left( \frac{\partial v}{\partial t} + A_{\epsilon, X} v - r x 1_{\{x > S\}} \mathcal{V}'_\epsilon(u_\epsilon^*) v \right) p_\epsilon^* = 2(u_\epsilon^*(T, x_{ob}) - \bar{u}) v(T, x_{ob}),$$

and  $\|p_\epsilon^*\|_{L^2((0, T) \times (0, X))}$  is bounded by a constant independent of  $\epsilon$  in the subsequence. For a fixed smooth function  $\phi$  taking the value 1 for  $|x - x_{ob}| \geq a/2$ ,  $T - t \geq a/2$  and vanishing in a neighborhood of  $(T, x_{ob})$ , we have that  $\phi p_\epsilon^* \in L^2(0, T; V_X) \cap C^0([0, T]; L^2((0, X)))$ , with norms bounded independently of  $\epsilon$ .

*Proof.* See Appendix B.  $\square$

REMARK 11. Problem (6.5) is a very weak formulation of

$$(6.6) \quad \begin{aligned} \frac{\partial p_\epsilon^*}{\partial t} - A_{\epsilon, X}^T p_\epsilon^* + r x 1_{\{x > S\}} \mathcal{V}'_\epsilon(u_\epsilon^*) p_\epsilon^* &= 0, \quad (t, x) \in [0, T) \times (0, X), \\ p_\epsilon^*(t, X) &= 0, \quad t \in (0, T), \\ p_\epsilon^*(T) &= -2(u_\epsilon^*(T, x_{ob}) - \bar{u}) \delta_{x_{ob}}, \end{aligned}$$

where  $A_{\epsilon, X}^T v(x) = -\frac{(\sigma_\epsilon^*)^2}{2} \frac{\partial^2}{\partial x^2} (x^2 v) + B_{\epsilon, X}^T v(x) - \frac{\partial}{\partial x} (r x v)$  with

$$B_{\epsilon, X}^T v(x) = \int_{\mathbb{R}} k_\epsilon^*(z) \left( x(e^z - 1) \frac{\partial v}{\partial x}(x) - 1_{z < \log \frac{x}{x}} e^{2z} v(xe^z) + (2e^z - 1)v(x) \right) dz,$$

and  $k_\epsilon^*(z) = |z|^{-(2\alpha_\epsilon^* + 1)} \psi_\epsilon^*(z)$ .

REMARK 12. Similarly,  $\left\| \frac{\partial(\phi p_\epsilon^*)}{\partial t} - A_{\epsilon, X}^T(\phi p_\epsilon^*) + r x 1_{\{x > S\}} \mathcal{V}'_\epsilon(u_\epsilon^*)(\phi p_\epsilon^*) \right\|_{L^2((0, T) \times (0, X))}$  is bounded independently of  $\epsilon$ .

From Lemma 6.2, we see that  $\int_0^T \langle x^2 \frac{\partial^2 u_\epsilon^*}{\partial x^2}, \phi p_\epsilon^* \rangle$  is well defined, where  $\langle \cdot, \cdot \rangle$  is the duality pairing between  $(V_X)'$  and  $V_X$ . On the other hand,  $\int_0^T \int_0^X ((1 - \phi) x^2 \frac{\partial^2 u_\epsilon^*}{\partial x^2}) p_\epsilon^*$  is well defined since both  $((1 - \phi) x^2 \frac{\partial^2 u_\epsilon^*}{\partial x^2})$  and  $p_\epsilon^*$  are square integrable. Moreover, the sum  $\int_0^T \langle x^2 \frac{\partial^2 u_\epsilon^*}{\partial x^2}, \phi p_\epsilon^* \rangle + \int_0^T \int_0^X ((1 - \phi) x^2 \frac{\partial^2 u_\epsilon^*}{\partial x^2}) p_\epsilon^*$  does not depend on the choice of  $\phi$ . Therefore, we call  $\mathcal{G}^{(\sigma)}(u_\epsilon^*, p_\epsilon^*)$  the quantity

$$(6.7) \quad \mathcal{G}^{(\sigma)}(u_\epsilon^*, p_\epsilon^*) = \int_0^T \left\langle x^2 \frac{\partial^2 u_\epsilon^*}{\partial x^2}, \phi p_\epsilon^* \right\rangle + \int_0^T \int_0^X \left( (1 - \phi) x^2 \frac{\partial^2 u_\epsilon^*}{\partial x^2} \right) p_\epsilon^*.$$

Let us introduce the operator  $B_{\epsilon, X}^{(\alpha)}$ :

$$(6.8) \quad \begin{aligned} B_{\epsilon, X}^{(\alpha)} v(x) &= - \int_{\mathbb{R}} k_\epsilon^*(z) \log(|z|) \left( x(e^z - 1) \frac{\partial v}{\partial x}(x) \right. \\ &\quad \left. + e^z (1_{\{z > -\log(\frac{x}{x})\}} v(xe^{-z}) - v(x)) \right) dz, \end{aligned}$$

where  $k_\epsilon^*(z) = |z|^{-2\alpha_\epsilon^* - 1} \psi_\epsilon^*(z)$ . From Lemma 6.2, the quantity  $\int_0^T \langle B_{\epsilon, X}^{(\alpha)} u_\epsilon^*, \phi p_\epsilon^* \rangle$  is well defined, where  $\langle \cdot, \cdot \rangle$  is the duality pairing between  $(V_X)'$  and  $V_X$ . On the other hand, the quantity  $\int_0^T \int_0^X ((1 - \phi) B_{\epsilon, X}^{(\alpha)} u_\epsilon^*) p_\epsilon^*$  is well defined since both  $p_\epsilon^*$  and  $((1 - \phi) B_{\epsilon, X}^{(\alpha)} u_\epsilon^*)$  are square integrable. Moreover, the sum  $\int_0^T \langle B_{\epsilon, X}^{(\alpha)} u_\epsilon^*, \phi p_\epsilon^* \rangle + \int_0^T \int_0^X ((1 - \phi) B_{\epsilon, X}^{(\alpha)} u_\epsilon^*) p_\epsilon^*$  does not depend on  $\phi$ . Therefore, we denote  $\mathcal{G}_\epsilon^{(\alpha)}(u_\epsilon^*, p_\epsilon^*)$

the quantity

$$(6.9) \quad \mathcal{G}_\epsilon^{(\alpha)}(u_\epsilon^*, p_\epsilon^*) = \int_0^T \left\langle B_{\epsilon, X}^{(\alpha)} u_\epsilon^*, \phi p_\epsilon^* \right\rangle + \int_0^T \int_0^X \left( (1 - \phi) B_{\epsilon, X}^{(\alpha)} u_\epsilon^* \right) p_\epsilon^*.$$

Similarly, for  $\kappa \in H_\psi$ , we introduce the operator  $B_{\epsilon, X}^{(\psi, \kappa)}$ :

$$B_{\epsilon, X}^{(\psi, \kappa)} v(x) = \int_{\mathbb{R}} \frac{\kappa(z)}{|z|^{1+2\alpha_\epsilon^*}} \left( x(e^z - 1) \frac{\partial v}{\partial x}(x) + e^z (1_{\{z > -\log(\frac{x}{\epsilon})\}} v(xe^{-z}) - v(x)) \right) dz,$$

and the quantity

$$(6.10) \quad \left\langle \mathcal{G}_\epsilon^{(\psi)}(u_\epsilon^*, p_\epsilon^*), \kappa \right\rangle = \int_0^T \left\langle B_{\epsilon, X}^{(\psi, \kappa)} u_\epsilon^*, \phi p_\epsilon^* \right\rangle + \int_0^T \int_0^X \left( (1 - \phi) B_{\epsilon, X}^{(\psi, \kappa)} u_\epsilon^* \right) p_\epsilon^*,$$

which does not depend on  $\phi$ . We are now ready to give necessary optimality for the least square problem (6.4).

PROPOSITION 6.3. *The optimality conditions for problem (6.4) are as follows: for all  $(\sigma, \alpha, \psi) \in \mathcal{H}$ ,*

$$(6.11) \quad (\sigma - \sigma_\epsilon^*) \left( D_\sigma J_R(\sigma_\epsilon^*, \alpha_\epsilon^*, \psi_\epsilon^*) + \sigma_\epsilon^* \mathcal{G}^{(\sigma)}(u_\epsilon^*, p_\epsilon^*) \right) \geq 0,$$

$$(6.12) \quad (\alpha - \alpha_\epsilon^*) \left( D_\alpha J_R(\sigma_\epsilon^*, \alpha_\epsilon^*, \psi_\epsilon^*) + 2\mathcal{G}_\epsilon^{(\alpha)}(u_\epsilon^*, p_\epsilon^*) \right) \geq 0,$$

$$(6.13) \quad \langle D_\psi J_R(\sigma_\epsilon^*, \alpha_\epsilon^*, \psi_\epsilon^*), \psi - \psi_\epsilon^* \rangle + \left\langle \mathcal{G}_\epsilon^{(\psi)}(u_\epsilon^*, p_\epsilon^*), \psi - \psi_\epsilon^* \right\rangle \geq 0.$$

*Proof.* The proof is quite standard. It is omitted for brevity.  $\square$

**6.1.4. First order necessary optimality conditions for (6.3).** In order to obtain optimality conditions for (6.3), we wish to pass to the limit in the optimality conditions for (6.4). Let  $\epsilon_n$  be sequence of penalty parameters converging to zero, and let  $(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*, u_{\epsilon_n}^*)$  be a sequence of solutions to (6.4) converging to  $(\sigma^*, \alpha^*, \psi^*, u^*)$  as in Lemma 6.1. Assume that there exists a positive number  $a$  such that  $u_{\epsilon_n}^*(t, x) > u_o(x) + \epsilon_n$  for all  $(t, x)$  with  $|x - x_{ob}| \leq a$  and  $T - t \leq a$ . Let  $p_{\epsilon_n}^*$  be the adjoint state defined by Lemma 6.2. There exists a subsequence denoted  $n_k$  such that  $p_{\epsilon_{n_k}}^*$  weakly converges to  $p^*$  in  $L^2((0, T) \times (0, X))$  and  $\phi p_{\epsilon_{n_k}}^*$  weakly converges to  $\phi p^*$  in  $L^2(0, T; V_X)$ , where  $\phi$  is given in Lemma 6.2.

We call  $\tilde{Z}$  and  $Z$  the spaces

$$(6.14) \quad \begin{aligned} \tilde{Z} &= \left\{ v \in L^2(0, T; V_X); \frac{\partial v}{\partial t} + A_X v \in L^2((0, T) \times (0, X)) \right\}, \\ Z &= \left\{ v \in \tilde{Z}; v(t = 0) = 0 \right\}, \end{aligned}$$

where  $A_X$  is the operator given by (4.2), (3.9), and (2.12), with the parameter  $(\sigma^*, \alpha^*, \psi^*)$ . These spaces, endowed with the graph norm, are Banach spaces.

PROPOSITION 6.4. *There exists a Radon measure  $\xi^*$  such that for all  $v \in Z$ ,*

$$(6.15) \quad \int_0^T \int_0^X \left( \frac{\partial v}{\partial t} + A_X v \right) p^* + \langle \xi^*, v \rangle = 2(u^*(T, x_{ob}) - \bar{u})v((T, x_{ob})).$$

*The function  $p^*$  satisfies*

$$(6.16) \quad \frac{\partial p^*}{\partial t} - A_X^T p^* - \xi^* = 0$$

in the sense of distributions. Furthermore, with  $u^*, \mu^*$  defined as in Lemma 6.1,

$$(6.17) \quad \mu^*|p^*| = 0,$$

$$(6.18) \quad |u^*|\xi^* = 0.$$

*Proof.* For simplicity, we drop the index  $n$  in  $\epsilon_n$ . In what follows,  $\epsilon$  means  $\epsilon_n$ . For a positive parameter  $\delta$ , we introduce the nondecreasing function  $\rho_\delta : \mathbb{R} \rightarrow \mathbb{R}$ ,

$$\rho_\delta(p) = -1 \text{ for } p \leq -\delta, \quad \rho_\delta(p) = p/\delta \text{ for } -\delta \leq p \leq \delta, \quad \rho_\delta(p) = 1 \text{ for } p \geq \delta,$$

and the nonnegative function  $R_\delta(p) = \int_0^p \rho_\delta(q) dq$ .

In what follows,  $\delta$  will be the generic term of a decreasing sequence of positive parameters which converges to 0.

For  $\phi$  introduced in Lemma 6.2, we use Remark 12: there exists a function  $g_\epsilon \in L^2((0, T) \times (0, X))$  with a norm bounded independently of  $\epsilon$  such that  $\phi p_\epsilon^*$  is the weak solution to

$$\frac{\partial(\phi p_\epsilon^*)}{\partial t} - A_{\epsilon, X}^T(\phi p_\epsilon^*) + r x 1_{\{x > S\}} \mathcal{V}'_\epsilon(u_\epsilon^*)(\phi p_\epsilon^*) = g_\epsilon$$

with the Cauchy condition  $(\phi p_\epsilon^*)(T, \cdot) = 0$  and the boundary condition  $(\phi p_\epsilon^*)(\cdot, X) = 0$ . Therefore,  $\|\phi p_\epsilon^*\|_{L^2(0, T; V_X)}$  is bounded uniformly in  $\epsilon$ . Moreover, from the properties of  $\phi$  (see Lemma 6.2), we have  $\mathcal{V}'_\epsilon(u_\epsilon^*)(\phi p_\epsilon^*) = \mathcal{V}'_\epsilon(u_\epsilon^*)p_\epsilon^*$ .

Multiplying the last equation by  $\rho_\delta(\phi p_\epsilon^*)$ , we obtain that there exists a constant  $C$  independent of  $\delta$  and  $\epsilon$  such that

$$(6.19) \quad \begin{aligned} & \int_0^X R_\delta(\phi p_\epsilon^*)(0, x) dx + \int_0^T \int_0^X \frac{(\sigma_\epsilon^*)^2 x^2}{2} \rho'_\delta(\phi p_\epsilon^*) \left( \frac{\partial(\phi p_\epsilon^*)}{\partial x} \right)^2 \\ & - r \int_0^T \int_S^X x \mathcal{V}'_\epsilon(u_\epsilon^*) p_\epsilon^* \rho_\delta(p_\epsilon^*) + \int_0^T \langle (B_{\epsilon, X}^T(\phi p_\epsilon^*)), \rho_\delta(\phi p_\epsilon^*) \rangle \leq C. \end{aligned}$$

Let us focus on the last term in the sum above: we can write it as

$$(6.20) \quad \begin{aligned} & \int_0^T \int_{\mathbb{R}_+} (B_\epsilon^T(\mathcal{E}_X(\phi p_\epsilon^*))) \rho_\delta(\mathcal{E}_X(\phi p_\epsilon^*)) = \frac{1}{2} \int_0^T \langle (B_\epsilon + B_\epsilon^T)(\mathcal{E}_X(\phi p_\epsilon^*)), \rho_\delta(\mathcal{E}_X(\phi p_\epsilon^*)) \rangle \\ & - \frac{1}{2} \int_0^T \langle (B_\epsilon - B_\epsilon^T)(\mathcal{E}_X(\phi p_\epsilon^*)), \rho_\delta(\mathcal{E}_X(\phi p_\epsilon^*)) \rangle. \end{aligned}$$

From Lemma 3.6 and the choice of  $\mathcal{H}_\psi$ , there exists a constant  $C$  independent of  $(\sigma, \alpha, \psi) \in \mathcal{H}$  and of  $\delta$  such that

$$(6.21) \quad \begin{aligned} & \left| \int_0^T \langle (B_\epsilon - B_\epsilon^T)(\mathcal{E}_X(\phi p_\epsilon^*)), \mathcal{E}_X(\rho_\delta(\phi p_\epsilon^*)) \rangle \right| \\ & \lesssim \|\phi p_\epsilon^*\|_{L^2((0, T); V_X)} \|\rho_\delta(\phi p_\epsilon^*)\|_{L^2((0, T) \times (0, X))} \leq C. \end{aligned}$$

On the other hand, from Lemma 3.5,

$$(6.22) \quad \begin{aligned} & \left| \int_0^T \langle (B_\epsilon + B_\epsilon^T)(\mathcal{E}_X(\phi p_\epsilon^*)), \mathcal{E}_X(\rho_\delta(\phi p_\epsilon^*)) \rangle \right. \\ & \left. - \int_0^T \int_0^X \int_{\mathbb{R}} k_\epsilon(z) e^z ((\phi p_\epsilon^*)(x) - (\phi p_\epsilon^*)(x e^{-z})) (\rho_\delta(\phi p_\epsilon^*)(x) - \rho_\delta(\phi p_\epsilon^*)(x e^{-z})) \right| \\ & \lesssim \|\phi p_\epsilon^*\|_{L^2((0, T) \times (0, X))} \|\rho_\delta(\phi p_\epsilon^*)\|_{L^2((0, T) \times (0, X))} \leq C. \end{aligned}$$



From (6.19), (6.20), (6.21), and (6.22), we see that

$$\begin{aligned} & \int_0^X R_\delta(\phi p_\epsilon^*)(0, x) dx + \int_0^T \int_0^X \int_{\mathbb{R}} k_\epsilon(z) e^z ((\phi p_\epsilon^*)(x) - (\phi p_\epsilon^*)(x e^{-z})) \\ & \quad \cdot (\rho_\delta(\phi p_\epsilon^*)(x) - \rho_\delta(\phi p_\epsilon^*)(x e^{-z})) \\ & + \int_0^T \int_0^X \frac{(\sigma_\epsilon^*)^2 x^2}{2} \rho'_\delta(\phi p_\epsilon^*) \left( \frac{\partial(\phi p_\epsilon^*)}{\partial x} \right)^2 - r \int_0^T \int_S^X x \mathcal{V}'_\epsilon(u_\epsilon^*) p_\epsilon^* \rho_\delta(p_\epsilon^*) \leq C. \end{aligned}$$

Since  $p \mapsto p \rho_\delta(p)$  is a nonnegative function and since  $\rho_\delta$  is nondecreasing, all the terms in the sum above are nonnegative. Therefore, for a constant  $C$  independent of the parameters,  $-r \int_0^T \int_S^X x \mathcal{V}'_\epsilon(u_\epsilon^*) p_\epsilon^* \rho_\delta(p_\epsilon^*) \leq C$ .

On the other hand we know that  $-x \mathcal{V}'_\epsilon(u_\epsilon^*) p_\epsilon^* \rho_\delta(p_\epsilon^*)$  defines an increasing (as  $\delta$  decreases) sequence of nonnegative functions, which converges almost everywhere to  $x |\mathcal{V}'_\epsilon(u_\epsilon^*) p_\epsilon^*|$  as  $\delta$  tends to 0. Thus, Beppo Levi's theorem tells us that  $-r \int_0^T \int_S^X x \mathcal{V}'_\epsilon(u_\epsilon^*) p_\epsilon^* \rho_\delta(p_\epsilon^*)$  tends to  $r \int_0^T \int_S^X x |\mathcal{V}'_\epsilon(u_\epsilon^*) p_\epsilon^*|$  as  $\delta \rightarrow 0$ . Therefore, for a positive constant  $C$ ,

$$(6.23) \quad r \int_0^T \int_S^X x |\mathcal{V}'_\epsilon(u_\epsilon^*) p_\epsilon^*| \leq C.$$

It is thus possible to extract a subsequence  $\epsilon_{n_k}$ , such that  $p_{\epsilon_{n_k}}^* \rightarrow p^*$  weakly in  $L^2((0, T) \times (0, X))$ ,  $\phi p_{\epsilon_{n_k}}^* \rightarrow \phi p^*$  weakly in  $L^2(0, T; V_X)$ , and  $-r x 1_{\{x > S\}} \mathcal{V}'_{\epsilon_{n_k}}(u_{\epsilon_{n_k}}^*) p_{\epsilon_{n_k}}^*$  converges to  $\xi^*$  weakly\* in  $(L^\infty((0, T) \times (0, X)))^*$ . In order to simplify the notation, we omit the indexes  $n_k$ ; now  $\epsilon$  means  $\epsilon_{n_k}$ .

From this, (6.15) is obtained as well by passing to the limit in (6.5), and (6.16) is satisfied in the sense of distributions.

For proving (6.17), we use the convexity of  $\mathcal{V}_\epsilon$  (still dropping the index  $n_k$  in  $\epsilon_{n_k}$ ): since  $\mathcal{V}_\epsilon(\epsilon) = 0$ , we have that for all  $u \in [0, \epsilon]$ ,  $\mathcal{V}_\epsilon(u) \leq -\mathcal{V}'_\epsilon(u)(\epsilon - u) \leq -\epsilon \mathcal{V}'_\epsilon(u)$ . This implies that  $\mathcal{V}_\epsilon(u_\epsilon^*) \leq -\epsilon \mathcal{V}'_\epsilon(u_\epsilon^*)$  because we also know that  $\mathcal{V}_\epsilon(u_\epsilon^*) = 0$  if  $u_\epsilon^* \geq \epsilon$ . Thus, calling  $\mu_\epsilon^* = r x 1_{\{x > S\}} \mathcal{V}'_\epsilon(u_\epsilon^*)$ , (6.23) implies that

$$(6.24) \quad 0 \leq \int_0^T \int_0^X \mu_\epsilon^* |p_\epsilon^*| \leq -\epsilon r \int_0^T \int_S^X x \mathcal{V}'_\epsilon(u_\epsilon^*) |p_\epsilon^*| \rightarrow 0.$$

But we also know that  $p_\epsilon^* \rightarrow p^*$  weakly in  $L^2((0, T) \times (0, X))$  and that  $\mu_\epsilon^* \rightarrow \mu^*$  strongly in  $L^2((0, T) \times (0, X))$  from Lemma 6.1. Hence,  $\int_0^T \int_0^X \mu_\epsilon^* |p_\epsilon^*| \rightarrow \int_0^T \int_0^X \mu^* |p^*|$ , and (6.17) is proved.

Let us call  $\xi_\epsilon^* = -r x 1_{\{x > S\}} \mathcal{V}'_\epsilon(u_\epsilon^*) p_\epsilon^* = -r x 1_{\{x > S\}} \mathcal{V}'_\epsilon(u_\epsilon^*) \phi p_\epsilon^*$ ; for  $\chi$  a continuous function in  $[0, T] \times [0, X]$ , we have

$$\int_0^T \int_0^X |\xi_\epsilon^*| \chi u_\epsilon^* \leq r \left( \int_0^T \int_S^X |x \mathcal{V}'_\epsilon(u_\epsilon^*)| |\phi p_\epsilon^*|^2 \right)^{\frac{1}{2}} \left( \int_0^T \int_S^X |x \mathcal{V}'_\epsilon(u_\epsilon^*)| |\chi u_\epsilon^*|^2 \right)^{\frac{1}{2}}$$

from the Cauchy-Schwarz inequality. But it can be checked that  $|\mathcal{V}'_\epsilon(u_\epsilon^*)| u_\epsilon^* \chi^2 \leq C \epsilon$ , which yields  $\int_0^T \int_S^X |x \mathcal{V}'_\epsilon(u_\epsilon^*)| u_\epsilon^* \chi^2 \leq C \epsilon$ . On the other hand, it is easy to check that  $\int_0^T \int_S^X |x \mathcal{V}'_\epsilon(u_\epsilon^*)| |\phi p_\epsilon^*|^2 \leq C$ . Therefore,  $\int_0^T \int_{\mathbb{R}_+} \xi_\epsilon^* |u_\epsilon^*| \chi \rightarrow 0$  as  $\epsilon \rightarrow 0$ . We know that  $\xi_\epsilon^* \rightarrow \xi^*$  weakly in  $(L^\infty)^*$  and that  $|u_\epsilon^*| \chi \rightarrow |u^*| \chi$  in  $C^0([0, T] \times [0, X])$  from Lemma 6.1. We can pass to the limit as  $\epsilon \rightarrow 0$  and (6.18) is proved.  $\square$

Proceeding as in (6.7), (6.9), and (6.10), we introduce the quantities

$$(6.25) \quad \mathcal{G}^{(\sigma)}(u^*, p^*) = \int_0^T \left\langle x^2 \frac{\partial^2 u^*}{\partial x^2}, \phi p^* \right\rangle + \int_0^T \int_0^X \left( (1 - \phi) x^2 \frac{\partial^2 u^*}{\partial x^2} \right) p^*,$$

$$(6.26) \quad \mathcal{G}^{(\alpha)}(u^*, p^*) = \int_0^T \left\langle B_X^{(\alpha)} u^*, \phi p^* \right\rangle + \int_0^T \int_0^X \left( (1 - \phi) B_X^{(\alpha)} u^* \right) p^*,$$

$$(6.27) \quad \left\langle \mathcal{G}^{(\psi)}(u^*, p^*), \kappa \right\rangle = \int_0^T \left\langle B_X^{(\psi, \kappa)} u^*, \phi p^* \right\rangle + \int_0^T \int_0^X \left( (1 - \phi) B_X^{(\psi, \kappa)} u^* \right) p^*,$$

where  $\phi$  is chosen as in Lemma 6.2, and where

$$B_X^{(\alpha)} v(x) = - \int_{\mathbb{R}} k^*(z) \log(|z|) \left( x(e^z - 1) \frac{\partial v}{\partial x}(x) + e^z (1_{\{z > -\log(\frac{x}{x_0})\}} v(xe^{-z}) - v(x)) \right),$$

$$B_X^{(\psi, \kappa)} v(x) = \int_{\mathbb{R}} \frac{\kappa(z)}{|z|^{1+2\alpha^*}} \left( x(e^z - 1) \frac{\partial v}{\partial x}(x) + e^z (1_{\{z > -\log(\frac{x}{x_0})\}} v(xe^{-z}) - v(x)) \right) dz.$$

One can check exactly as above that  $\mathcal{G}^{(\sigma)}(u^*, p^*)$ ,  $\mathcal{G}^{(\alpha)}(u^*, p^*)$ , and  $\langle \mathcal{G}^{(\psi)}(u^*, p^*), \kappa \rangle$  are well defined and do not depend of the particular choice of  $\phi$ . We are now ready to give necessary optimality for the least square problem (6.4).

**PROPOSITION 6.5.** *Let  $(\sigma^*, \alpha^*, \psi^*, u^*)$  be a solution to problem (6.3) obtained in Lemma 6.1. Assume that  $u^*(T, x_{ob}) > u_0(x_{ob})$  and take  $a$  as in Remark 10. There exist  $p^* \in L^2((0, T) \times (0, X))$  and a Radon measure  $\xi^*$  satisfying (6.15), (6.17), (6.18), and such that, for all  $(\sigma, \alpha, \psi) \in \mathcal{H}$ ,*

$$(6.28) \quad (\sigma - \sigma^*) \left( D_\sigma J_R(\sigma^*, \alpha^*, \psi^*) + \sigma^* \mathcal{G}^{(\sigma)}(u^*, p^*) \right) \geq 0,$$

$$(6.29) \quad (\alpha - \alpha^*) \left( D_\alpha J_R(\sigma_\epsilon^*, \alpha_\epsilon^*, \psi_\epsilon^*) + 2\mathcal{G}^{(\alpha)}(u^*, p^*) \right) \geq 0,$$

$$(6.30) \quad \langle D_\psi J_R(\sigma_\epsilon^*, \alpha_\epsilon^*, \psi_\epsilon^*), \psi - \psi^* \rangle + \langle \mathcal{G}^{(\psi)}(u^*, p^*), \psi - \psi^* \rangle \geq 0.$$

*Proof.* We consider a sequence of parameters  $\epsilon_n$  such that

- (1)  $(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*, u_{\epsilon_n}^*)$  is a sequence of solutions to (6.4) converging to  $(\sigma^*, \alpha^*, \psi^*, u^*)$  as in Lemma 6.1,
- (2)  $u_{\epsilon_n}^*(t, x) > u_0(x) + \epsilon_n$  for all  $(t, x)$  with  $|x - x_{ob}| \leq a$  and  $T - t \leq a$ ,
- (3) for the adjoint states  $p_\epsilon^*$  defined by Lemma 6.2,  $p_{\epsilon_n}^*$  weakly converges to  $p^*$  in  $L^2((0, T) \times (0, X))$  and  $\phi p_{\epsilon_n}^*$  weakly converges to  $\phi p^*$  in  $L^2(0, T; V_X)$ , where  $\phi$  is given in Lemma 6.2.

We drop the index  $n$  in  $\epsilon_n$ . We have to prove that  $\lim_{\epsilon \rightarrow 0} \mathcal{G}^{(\sigma)}(u_\epsilon^*, p_\epsilon^*) = \mathcal{G}^{(\sigma)}(u^*, p^*)$ . Since  $u_\epsilon^* \rightarrow u^*$  strongly in  $L^2(0, T; V_X)$  (see Lemma 6.1) and  $\phi p_\epsilon^* \rightarrow \phi p^*$  weakly in  $L^2(0, T; V_X)$ , we deduce that

$$\lim_{\epsilon \rightarrow 0} \int_0^T \left\langle \frac{\partial^2 u_\epsilon^*}{\partial x^2}, \phi p_\epsilon^* \right\rangle = \int_0^T \left\langle \frac{\partial^2 u^*}{\partial x^2}, \phi p^* \right\rangle.$$

On the other hand,  $(1 - \phi) \frac{\partial^2 u_\epsilon^*}{\partial x^2}$  strongly converges to  $(1 - \phi) \frac{\partial^2 u^*}{\partial x^2}$  in  $L^2((0, T) \times (0, X))$  from Lemma 6.1, and  $p_\epsilon^*$  weakly converges to  $p^*$  in  $L^2((0, T) \times (0, X))$ . Thus

$$\lim_{\epsilon \rightarrow 0} \int_0^T \int_0^X \left( (1 - \phi) \frac{\partial^2 u_\epsilon^*}{\partial x^2} \right) p_\epsilon^* = \int_0^T \int_0^X \left( (1 - \phi) \frac{\partial^2 u^*}{\partial x^2} \right) p^*.$$

From the two points above, we see that  $\lim_{\epsilon \rightarrow 0} \mathcal{G}^{(\sigma)}(u_\epsilon^*, p_\epsilon^*) = \mathcal{G}^{(\sigma)}(u^*, p^*)$ ; we can pass to the limit in (6.11) and obtain (6.28).

We have to prove that  $\lim_{\epsilon \rightarrow 0} \mathcal{G}_\epsilon^{(\alpha)}(u_\epsilon^*, p_\epsilon^*) = \mathcal{G}^{(\alpha)}(u^*, p^*)$ . The fact that  $u_\epsilon^* \rightarrow u^*$  strongly in  $L^2(0, T; V_X)$  (see Lemma 6.1) implies that  $B_{\epsilon, X}^{(\alpha)} u_\epsilon^*$  converges to  $B_X^{(\alpha)} u^*$  in  $L^2(0, T, (V_X)')$ . On the other hand,  $\phi p_\epsilon^* \rightarrow \phi p^*$  weakly in  $L^2(0, T; V_X)$ . This implies that

$$\lim_{\epsilon \rightarrow 0} \int_0^T \left\langle B_{\epsilon, X}^{(\alpha)} u_\epsilon^*, \phi p_\epsilon^* \right\rangle = \int_0^T \left\langle B_X^{(\alpha)} u^*, \phi p^* \right\rangle.$$

It can also be checked that  $(1 - \phi)B_{\epsilon, X}^{(\alpha)} u_\epsilon^*$  strongly converges to  $(1 - \phi)B_X^{(\alpha)} u^*$  in  $L^2((0, T) \times (0, X))$ . From the weak convergence of  $p_\epsilon^*$  to  $p^*$  in  $L^2((0, T) \times (0, X))$ , we deduce that

$$\lim_{\epsilon \rightarrow 0} \int_0^T \int_0^X \left( (1 - \phi)B_{\epsilon, X}^{(\alpha)} u_\epsilon^* \right) p_\epsilon^* = \int_0^T \int_0^X \left( (1 - \phi)B_X^{(\alpha)} u^* \right) p^*.$$

The two points above yield that  $\lim_{\epsilon \rightarrow 0} \mathcal{G}_\epsilon^{(\alpha)}(u_\epsilon^*, p_\epsilon^*) = \mathcal{G}^{(\alpha)}(u^*, p^*)$ . This and (6.12) yield (6.29). The last condition (6.30) is obtained in the same manner.  $\square$

**6.2. Conclusion: Optimality condition for the calibration problem.** For calibrating the Lévy process, one observes the spot price  $S$  and the prices  $(\bar{p}_i)_{i \in I}$  of a family of American put options with maturities/strikes given by  $(T_i, x_i)$ , and we call  $\bar{u}_i = \bar{p}_i - x_i + S$ ,  $i \in I$ . We assume that

$$\bar{u}_i > u_o(x_i) \quad \text{for all } i \in I.$$

Call  $T = \max_{i \in I} T_i$ . Let  $\bar{X}$  be such that for all  $(\sigma, \alpha, \psi) \in \mathcal{H}$ , the exercise price  $\gamma(t)$  is smaller than  $\bar{X}$  for all  $t \leq T$ , and take  $X \geq \bar{X}$ . The calibration problem has the form (6.3) with the new definition of  $J$ :

$$J(u) = \sum_{i \in I} \omega_i (u(T_i, x_i) - \bar{u}_i)^2,$$

where  $\omega_i$  are positive weights.

As above, we can also define the modified least square problem (6.4) and have  $\epsilon$  tend to 0. Let a subsequence  $(\sigma_{\epsilon_n}^*, \alpha_{\epsilon_n}^*, \psi_{\epsilon_n}^*, u_{\epsilon_n}^*)$  of solutions of (6.4) converge to  $(\sigma^*, \alpha^*, \psi^*, u^*)$  as in Lemma 6.1; then  $(\sigma^*, \alpha^*, \psi^*, u^*)$  is a solution of (6.3).

We assume that  $u^*(T_i, x_i) > u_o(x_i)$  for all  $i \in I$ . It is clear from the continuity of  $u^*$  and from the uniform convergence of  $u_{\epsilon_n}^*$  that there exist a positive real number  $a$  and an integer  $N$  such that for  $n > N$ ,  $u_{\epsilon_n}^*(t, x) > u_o(x) + \epsilon_n$  for all  $(t, x)$  such that  $|t - T_i| < a$  and  $|x - x_i| < a$  for some  $i \in I$ . We may fix a smooth function  $\phi$  taking the value 1 for all  $x$  such that  $|x - x_i| \geq a/2$ ,  $|T_i - t| \geq a/2$  for all  $i \in I$ , and vanishing in neighborhoods of  $(T_i, x_i)$ ,  $i \in I$ .

Calling  $A_X$  the operator defined by (4.2), (3.9), and (2.12) with the parameters  $(\sigma, \alpha, \psi) = (\sigma^*, \alpha^*, \psi^*)$ , we obtain the optimality conditions exactly as in section 6.1.4.

**THEOREM 6.6.** *Under the assumptions made at the beginning of section 6.2, there exist a function  $p^* \in L^2((0, T) \times (0, X))$  and a Radon measure  $\xi^*$  satisfying (6.17), (6.18) and for all  $v \in Z$  ( $Z$  is defined by (6.14)),*

$$(6.31) \quad \int_0^T \int_0^X \left( \frac{\partial v}{\partial t} + A_X v \right) p^* + \langle \xi^*, v \rangle = 2 \sum_{i \in I} \omega_i (u^*(T_i, x_i) - \bar{u}_i) v((T_i, x_i)),$$

equation (6.31) which remains unchanged and which I do not write and such that (6.28), (6.29), and (6.30) hold for all  $(\sigma, \alpha, \psi) \in \mathcal{H}$ , with  $\mathcal{G}^{(\sigma)}$ ,  $\mathcal{G}^{(\alpha)}$ , and  $\mathcal{G}^{(\psi)}$  defined, respectively, by (6.25), (6.26), and (6.27) (with the new choice of  $\phi$ ).

*Proof.* The proof follows exactly the same lines as that of Proposition 6.5.  $\square$

### Appendix A.

*Proof of Lemma 3.2.* By the change of variable  $y = \log(x)$ ,

$$\begin{aligned} |v|_{\phi,s}^2 &= \int_{\mathbb{R}} e^y dy \int_{\mathbb{R}} \frac{\phi(z)}{|z|^{1+2s}} (v(e^{y-z}) - v(e^y))^2 dz \\ &= \int_{\mathbb{R}} dy \int_{\mathbb{R}} \frac{\phi(z)}{|z|^{1+2s}} (e^{\frac{z}{2}} \tilde{v}(y-z) - \tilde{v}(y))^2 dz. \end{aligned}$$

By Fubini's theorem,  $|v|_{\phi,s}^2 = \int_{\mathbb{R}} dz \phi(z) |z|^{-(1+2s)} \int_{\mathbb{R}} (e^{\frac{z}{2}} \tilde{v}(y-z) - \tilde{v}(y))^2 dy$ ; after a Fourier transform w.r.t. the variable  $y$ ,

$$\begin{aligned} |v|_{\phi,s}^2 &= \int_{\mathbb{R}} dz \phi(z) |z|^{-(1+2s)} \int_{\mathbb{R}} \left| e^{z(\frac{1}{2}-i\xi)} \widehat{\tilde{v}}(\xi) - \widehat{\tilde{v}}(\xi) \right|^2 d\xi \\ &= \int_{\mathbb{R}} dz \phi(z) |z|^{-(1+2s)} \int_{\mathbb{R}} (e^z + 1 - 2e^{\frac{z}{2}} \cos(\xi z)) |\widehat{\tilde{v}}(\xi)|^2 d\xi \\ &= \int_{\mathbb{R}} dz \phi(z) |z|^{-(1+2s)} \int_{\mathbb{R}} \left( (e^{\frac{z}{2}} - 1)^2 + 4e^{\frac{z}{2}} \sin^2\left(\frac{\xi z}{2}\right) \right) |\widehat{\tilde{v}}(\xi)|^2 d\xi \\ &= \left( \int_{\mathbb{R}} \frac{\phi(z)}{|z|^{1+2s}} (e^{\frac{z}{2}} - 1)^2 dz \right) \|v\|_{L^2(\mathbb{R}_+)}^2 + 4 \int_{\mathbb{R}} dz \frac{\phi(z) e^{\frac{z}{2}}}{|z|^{1+2s}} \int_{\mathbb{R}} \sin^2\left(\frac{\xi z}{2}\right) |\widehat{\tilde{v}}(\xi)|^2 d\xi. \end{aligned}$$

But  $4 \int_{\mathbb{R}} dz \frac{\phi(z) e^{\frac{z}{2}}}{|z|^{1+2s}} \int_{\mathbb{R}} \sin^2\left(\frac{\xi z}{2}\right) |\widehat{\tilde{v}}(\xi)|^2 d\xi = 4 \int_{\mathbb{R}} |\xi|^{2s} |\widehat{\tilde{v}}(\xi)|^2 \int_{\mathbb{R}} \frac{\phi(z) e^{\frac{z}{2}}}{|\xi z|^{1+2s}} \sin^2\left(\frac{\xi z}{2}\right) |\xi| dz$ . Define

$$C_1 = \int_{\mathbb{R}} \phi(z) |z|^{-(1+2s)} (e^{\frac{z}{2}} - 1)^2 dz,$$

which is a real number since  $z \mapsto \phi(z) \max(e^z, 1)$  is bounded. Similarly, there exists a constant  $\beta > 0$  such that

$$4 \int_{\mathbb{R}} \frac{\phi(z) e^{\frac{z}{2}}}{|\xi z|^{1+2s}} \sin^2\left(\frac{\xi z}{2}\right) |\xi| dz \leq 4\beta \int_{\mathbb{R}} \frac{\sin^2\left(\frac{\xi z}{2}\right)}{|\xi z|^{1+2s}} |\xi| dz = 4\beta \int_{\mathbb{R}} \frac{\sin^2\left(\frac{u}{2}\right)}{|u|^{1+2s}} du,$$

and we introduce  $C_2 = 4\beta \int_{\mathbb{R}} \frac{\sin^2\left(\frac{u}{2}\right)}{|u|^{1+2s}} du$ . We have obtained that

$$|v|_{\phi,s}^2 \leq C_1 \|v\|_{L^2(\mathbb{R}_+)}^2 + C_2 \int_{\mathbb{R}} |\xi|^{2s} |\widehat{\tilde{v}}(\xi)|^2 d\xi.$$

On the other hand,

$$\begin{aligned} \int_{\mathbb{R}} \frac{\phi(z) e^{\frac{z}{2}}}{|\xi z|^{1+2s}} \sin^2\left(\frac{\xi z}{2}\right) |\xi| dz &= \int_{\mathbb{R}} \phi\left(\frac{u}{|\xi|}\right) e^{\frac{u}{2|\xi|}} \frac{\sin^2\left(\frac{u}{2}\right)}{|u|^{1+2s}} du \\ &\geq \int_{-1}^1 \phi\left(\frac{u}{|\xi|}\right) e^{\frac{u}{2|\xi|}} \frac{\sin^2\left(\frac{u}{2}\right)}{|u|^{1+2s}} du \end{aligned}$$

implies that

$$\liminf_{|\xi| \rightarrow \infty} \int_{\mathbb{R}} \phi(z) e^{\frac{z}{2}} |\xi z|^{-(1+2s)} \sin^2\left(\frac{\xi z}{2}\right) |\xi| dz \geq \phi(0) \int_{-1}^1 \sin^2\left(\frac{u}{2}\right) |u|^{-(1+2s)} du,$$

which shows that there exists a constant  $M > 0$  such that for  $|\xi| \geq M$ ,

$$4 \int_{\mathbb{R}} \phi(z) e^{\frac{z}{2}} |\xi z|^{-(1+2s)} \sin^2 \left( \frac{\xi z}{2} \right) |\xi| dz \geq 2\phi(0) \int_{-1}^1 \sin^2 \left( \frac{u}{2} \right) |u|^{-(1+2s)} du,$$

and we introduce  $C_4 = 2\phi(0) \int_{-1}^1 \frac{\sin^2(\frac{u}{2})}{|u|^{1+2s}} du$ . Thus

$$|v|_{\phi,s}^2 \geq C_1 \|v\|_{L^2(\mathbb{R}_+)}^2 + C_4 \int_{|\xi| > M} |\xi|^{2s} |\widehat{v}(\xi)|^2 d\xi \geq \frac{C_1}{2} \|v\|_{L^2(\mathbb{R}_+)}^2 + C_3 \int_{\mathbb{R}} |\xi|^{2s} |\widehat{v}(\xi)|^2 d\xi,$$

with  $C_3 = \min(C_1/(2M^{2s}), C_4)$ .  $\square$

*Proof of Lemma 3.3.* It is enough to prove that  $B|_{\mathcal{D}(\mathbb{R}_+)}$  is continuous from  $\mathcal{D}(\mathbb{R}_+)$  endowed with the norm of  $V^s$  to  $V^{s-\max(2\alpha,1)}$  if  $\alpha \neq 1/2$ , and to  $V^{s-1-\epsilon}$  if  $\alpha = 1/2$ . For that, we use the change of variable  $y = \log(x)$  and call  $\tilde{u}$  the function defined on  $\mathbb{R}$  by  $\tilde{u}(y) = u(e^y) e^{\frac{y}{2}}$ . This yields that  $\langle Bu, v \rangle = \langle \tilde{B}\tilde{u}, \tilde{v} \rangle$ , where

$$\begin{aligned} \tilde{B}\tilde{u}(y) &= - \int_{\mathbb{R}} k(z) e^z \left( (1 - e^{-z}) \left( \frac{\partial \tilde{u}}{\partial y}(y) - \frac{1}{2} \tilde{u}(y) \right) + (e^{\frac{z}{2}} \tilde{u}(y - z) - \tilde{u}(y)) \right) dz \\ &= - \int_{\mathbb{R}} k(z) e^z \left( (1 - e^{-z}) \frac{\partial \tilde{u}}{\partial y}(y) + \left( \frac{1}{2} e^{-z} - \frac{3}{2} \right) \tilde{u}(y) + e^{\frac{z}{2}} \tilde{u}(y - z) \right) dz. \end{aligned}$$

The Fourier transform of  $\tilde{B}\tilde{u}$  is

$$\hat{b} = -\widehat{\tilde{u}}(\xi) \int_{\mathbb{R}} k(z) e^z \left( (1 - e^{-z}) i\xi + \frac{1}{2} e^{-z} - \frac{3}{2} + e^{\frac{z}{2}} e^{-i\xi z} \right) dz.$$

We make out three cases.

(1)  $\alpha > 1/2$ . In this case, one sees that

$$\begin{aligned} (1 - e^{-z}) i\xi + \frac{1}{2} e^{-z} - \frac{3}{2} + e^{\frac{z}{2}} e^{-i\xi z} &= \frac{1}{2} (e^{-z} + 2e^{\frac{z}{2}} - 3) + e^{\frac{z}{2}} (e^{-i\xi z} - 1 + iz\xi) \\ &\quad - i\xi (e^{-z} - 1 + ze^{\frac{z}{2}}). \end{aligned}$$

Therefore,

$$(A.1) \quad \hat{b} = -\widehat{\tilde{u}}(\xi) \left( \begin{aligned} &\int_{\mathbb{R}} \frac{k(z)}{2} (1 + 2e^{\frac{3z}{2}} - 3e^z) + \int_{\mathbb{R}} k(z) e^{\frac{3z}{2}} (e^{-i\xi z} - 1 + iz\xi) \\ &- i\xi \int_{\mathbb{R}} k(z) e^z (e^{-z} - 1 + ze^{\frac{z}{2}}) \end{aligned} \right).$$

From the assumption on  $\psi$ , the first integral is a real number independent of  $\xi$ .

As in Lemma 3.2, by introducing  $\theta = \xi/|\xi|$  for  $\xi \neq 0$ , and writing that

$$\int_{\mathbb{R}} k(z) e^{\frac{3z}{2}} (e^{-i\xi z} - 1 + iz\xi) dz = |\xi|^{2\alpha} \int_{\mathbb{R}} |y|^{-(1+2\alpha)} \psi \left( \frac{y}{|\xi|} \right) e^{\frac{3y}{2|\xi|}} (e^{-iy\theta} - 1 + iy\theta) dy,$$

we see from the assumptions on  $\psi$  that there exists a positive constant  $C_1$  such that  $|\int_{\mathbb{R}} k(z) e^{\frac{3z}{2}} (e^{-i\xi z} - 1 + iz\xi) dz| \leq C_1 |\xi|^{2\alpha}$ , because  $y \mapsto \psi(y) \exp(3y/2)$  is a real bounded function and  $\int_{\mathbb{R}} |e^{-iy\theta} - 1 + iy\theta| |y|^{-(1+2\alpha)} dy$  can be bounded independently of  $\xi$ . The third integral in (A.1) is a real number independent of  $\xi$ . Therefore,

$$|\hat{b}| \lesssim (1 + |\xi| + |\xi|^{2\alpha}) |\widehat{\tilde{u}}(\xi)| \lesssim (1 + |\xi|^{2\alpha}) |\widehat{\tilde{u}}(\xi)|.$$

(2)  $\alpha < 1/2$ . We can still split  $\hat{b}$  as in (A.1). From the assumption on  $\psi$ , the first integral is a real number independent of  $\xi$ . The second integral can be split into the sum of  $\int_{\mathbb{R}} k(z) e^{\frac{3z}{2}} (e^{-i\xi z} - 1) dz$  and  $i\xi \int_{\mathbb{R}} k(z) e^{\frac{3z}{2}} z dz$ , which are both bounded by  $C_1|\xi|$ . The third integral is a real number independent of  $\xi$ , so

$$|\hat{b}| \lesssim (1 + |\xi|)|\hat{u}(\xi)|.$$

(3)  $\alpha = 1/2$ . The only change w.r.t. the previous two cases concerns the second integral: we write it as

$$(A.2) \quad |\xi| \int_{\mathbb{R}} \psi \left( \frac{y}{|\xi|} \right) e^{\frac{3y}{2|\xi|}} \frac{(e^{-iy\theta} - 1 + iy1_{|y|<1}\theta)}{|y|^2} dy + i\theta|\xi| \int_{|z\xi|>1} \psi(z) e^{\frac{3z}{2}} \frac{z}{|z|^2} dz.$$

The first integral in (A.2) is bounded by a constant independent of  $\xi$ , because  $z \mapsto \psi(z) \exp(3z/2)$  is a bounded function and  $\int_{\mathbb{R}} |y|^{-2} |e^{-iy\theta} - 1 + iy1_{|y|<1}\theta|$  can be bounded independently of  $\xi$ . For the second integral in (A.2), we have, if  $\xi > 1$ ,

$$\left| \int_{|z\xi|>1} \psi(z) e^{\frac{3z}{2}} \frac{z}{|z|^2} dz \right| \leq \int_{|z|>1} \frac{\psi(z) e^{\frac{3z}{2}}}{|z|} dz + \int_{|\xi|^{-1} \leq |z| \leq 1} \frac{\psi(z) e^{\frac{3z}{2}}}{|z|} dz \lesssim (1 + \log(\xi))$$

$$|\hat{b}| \lesssim (1 + |\xi| + |\xi \log(|\xi|)|)|\hat{u}(\xi)|. \quad \square$$

*Proof of Lemma 3.5.* It is enough to prove the result for  $u, v \in \mathcal{D}(\mathbb{R}_+)$ :

$$\langle Bu, v \rangle = - \int_{\mathbb{R}_+} dx \int_{\mathbb{R}} k(z) \left( x(e^z - 1) \frac{\partial u}{\partial x}(x) + e^z (u(xe^{-z}) - u(x)) \right) v(x) dz = I + II + III,$$

where

$$\begin{aligned} I &= \int_{\mathbb{R}_+} dx \int_{\mathbb{R}} k(z) e^z (u(x) - u(xe^{-z})) (v(x) - v(xe^{-z})) dz, \\ II &= - \int_{\mathbb{R}_+} dx \int_{\mathbb{R}} k(z) x(e^z - 1) \frac{\partial u}{\partial x}(x) v(x) dz, \\ III &= \int_{\mathbb{R}_+} dx \int_{\mathbb{R}} k(z) e^z (u(x) - u(xe^{-z})) v(xe^{-z}) dz. \end{aligned}$$

But

$$II = \int_{\mathbb{R}_+} dx \int_{\mathbb{R}} k(z) x(e^z - 1) \frac{\partial v}{\partial x}(x) u(x) dz + \int_{\mathbb{R}_+} dx \int_{\mathbb{R}} k(z) (e^z - 1) u(x) v(x) dz.$$

From this,

$$\begin{aligned} II + III &= \int_{\mathbb{R}_+} \left( \int_{\mathbb{R}} k(z) \left( x(e^z - 1) \frac{\partial v}{\partial x}(x) + e^z (v(xe^{-z}) - v(x)) \right) dz \right) u(x) dx \\ &\quad + \left( \int_{\mathbb{R}} k(z) (2e^z - e^{2z} - 1) dz \right) \int_{\mathbb{R}_+} u(x) v(x) dx. \end{aligned}$$

The desired result is obtained.  $\square$

*Proof of Lemma 3.6.* The assertion is already proved in the case  $\alpha < 1/2$ , thanks to Lemma 3.3 and Remark 4. Thus, let us focus on the case when  $\alpha \geq 1/2$ : after a few calculations, one sees that

$$(B^T u - Bu)(x) = \int_{\mathbb{R}} k(z) \left( x(e^z - 1) \left( 2 \frac{\partial u}{\partial x}(x) + u(x) \right) + e^z u(xe^{-z}) - e^{2z} u(xe^{-z}) \right) dz.$$

The same change of variables as in the proof of Lemma 3.3 leads to  $\langle (B - B^T)u, v \rangle = \langle (\tilde{B} - \tilde{B}^T)\tilde{u}, \tilde{v} \rangle$ , where

$$(\tilde{B}\tilde{u} - \tilde{B}^T\tilde{u})(y) = - \int_{\mathbb{R}} k(z) \left( 2(e^z - 1) \frac{\partial \tilde{u}}{\partial y}(y) + e^{\frac{3z}{2}} (\tilde{u}(y - z) - \tilde{u}(y + z)) \right) dz.$$

The Fourier transform of  $(\tilde{B} - \tilde{B}^T)\tilde{u}$  is

$$\begin{aligned} & -\widehat{\tilde{u}}(\xi) \int_{\mathbb{R}} k(z) \left( 2i\xi(e^z - 1) - e^{\frac{3z}{2}} (e^{i\xi z} - e^{-i\xi z}) \right) dz \\ &= -2i\xi \widehat{\tilde{u}}(\xi) \int_{\mathbb{R}} k(z) (e^z - 1 - ze^{\frac{3z}{2}}) dz + \widehat{\tilde{u}}(\xi) \int_{\mathbb{R}} k(z) e^{\frac{3z}{2}} (e^{i\xi z} - e^{-i\xi z} - 2i\xi z) dz. \end{aligned}$$

From the assumptions, the first integral in the sum above is a real number. Let us focus on the second integral: since the function  $z \mapsto e^{i\xi z} - e^{-i\xi z} - 2i\xi z$  is odd,  $\psi(0) \int_{\mathbb{R}} |z|^{-(1+2\alpha)} e^{\frac{-3|z|}{2}} (e^{i\xi z} - e^{-i\xi z} - 2i\xi z) dz = 0$ , and

$$\begin{aligned} & \int_{\mathbb{R}} k(z) e^{\frac{3z}{2}} (e^{i\xi z} - e^{-i\xi z} - 2i\xi z) dz = \int_{\mathbb{R}} |z|^{-(1+2\alpha)} z \omega(z) (e^{i\xi z} - e^{-i\xi z} - 2i\xi z) dz \\ & \lesssim \int_{\mathbb{R}} |z|^{-2\alpha} e^{-\zeta|z|} (e^{i\xi z} - e^{-i\xi z} - 2i\xi z) dz \lesssim |\xi|^{2\alpha-1} \lesssim (1 + |\xi|), \end{aligned}$$

where, in the case  $\alpha = 1/2$ , we have used the fact that  $|\sin(\xi z)|/|z| \leq |\xi|$ . This concludes the proof.  $\square$

## Appendix B.

*Proof of Proposition 4.1.* Consider  $X'$ ,  $0 < X' < X$  and let  $\phi$  be a smooth cut-off function taking the value 1 in  $[0, 3/4X' + \frac{1}{4}X]$  and 0 in  $[3/4X + \frac{1}{4}X', X]$ . It is possible to prove that  $A_X(\mathcal{E}_X(\phi v)) \in L^2((\mathbb{R}_+))$ , which yields that  $\mathcal{E}_X(\phi v) \in V^2$  and  $\phi v \in W_X^2$ . This yields the first statement of Proposition 4.1.

Assume that  $v \in V_X$  is such that  $A_X v \in L^2((0, X))$ . Then there exists  $f \in L^2((0, X))$  such that

$$(B.1) \quad -\sigma x^2 \frac{\partial^2 v}{\partial x^2} = f - B_X v.$$

If  $0 \leq \alpha < 1/2$ , then, from Lemma 3.3,  $B_X v \in L^2((0, X))$ , and (B.1) implies that  $v \in W_X^2 \cap V_X$ .

If  $\alpha > 1/2$ , then, from Lemma 3.3,  $B_X v \in V_X^{1-2\alpha}$ . From this and (B.1), one immediately deduces that  $v \in V_X \cap W_X^{3-2\alpha}$ . A boot-strap argument is needed for improving this result.

If  $1/2 < \alpha < 3/4$ , then for all  $\epsilon > 0$ ,  $v \in W_X^{3/2-\epsilon}$ , and  $\mathcal{E}_X(v) \in V^{3/2-\epsilon}$ . Note that we cannot give a better regularity result for  $\mathcal{E}_X(v)$  (for example,  $\mathcal{E}_X(v) \in V^{3/2+\epsilon}$ ), because this would require the condition  $\frac{\partial v}{\partial x}(x = X) = 0$ , which is not proved. Then Lemma 3.3 yields that  $B_X v \in L^2((0, X))$  and that  $v \in W_X^2 \cap V_X$  from (B.1). In the

case  $\alpha = 1/2$ , we obtain from Lemma 3.3 that  $v \in W_X^2 \cap V_X$  as well. If  $\alpha = 3/4$ , the same argument shows that  $v \in W_X^{2-\epsilon} \cap V_X$  for all  $\epsilon > 0$ .

On the contrary, if  $3/4 < \alpha < 1$ , we have to keep on boot-strapping:  $v \in V_X \cap W_X^{3-2\alpha}$  implies that  $B_X v \in V_X^{3-4\alpha}$ , and from (B.1),  $v \in W_X^{5-4\alpha}$ . Either  $3/4 < \alpha < 7/8$ , and we see that there exists  $\epsilon > 0$  such that  $v \in W_X^{3/2+\epsilon}$ , or  $7/8 \leq \alpha < 1$ , and we keep on boot-strapping. After a finite number of steps, we obtain the first two statements of Proposition 4.1.

Then we obtain that  $\frac{\partial v}{\partial x} \in \mathcal{C}^0((0, X))$  from Sobolev imbeddings.  $\square$

*Proof of Theorem 4.2.* For brevity, and since the proof uses rather classical arguments, we shall omit some details. By using results on parabolic equations with monotone operators [26, p. 156], it is possible to prove that (4.6) has a unique weak solution in  $L^2(0, T; V_X) \cap C^0([0, T]; L^2(0, X))$ , with  $\frac{\partial u_{X,\epsilon}}{\partial t} \in L^2(0, T; V'_X)$ . Note that for all  $t_0$ ,  $0 < t_0 < T$ ,  $u_{X,\epsilon}$  is smooth in  $(t_0, T] \times [a, b]$ , where  $[a, b]$  is any interval strictly contained in  $(0, S)$  or in  $(S, X)$ . From (3.11), the weak maximum principle may be used. It yields that, a.e.,  $u_{X,\epsilon}$  is nonnegative on the one hand and greater than or equal to  $x \mapsto S - x$  on the other hand. Therefore, for almost every time  $t$ ,  $u_{X,\epsilon}(t) \in K_X$ . This implies that  $0 \leq rx(1 - 1_{\{x>S\}}) \leq rx(1 - 1_{\{x>S\}}) \mathcal{V}_\epsilon(u_{X,\epsilon}) \leq rx$ .

From this and (4.6),  $u_{X,\epsilon}$  belongs to  $C^0([0, T]; V_X) \cap L^2(0, T; D_X)$ ,  $\frac{\partial u_{X,\epsilon}}{\partial t} \in L^2((0, T) \times (0, X))$ , and the norms  $\|u_{X,\epsilon}\|_{L^\infty(0, T; V_X)}$ ,  $\|u_{X,\epsilon}\|_{L^2(0, T; D_X)}$ , and  $\|\frac{\partial}{\partial t} u_{X,\epsilon}\|_{L^2((0, T) \times (0, X))}$  are bounded independently of  $\epsilon$ .

Since  $V \subset \mathcal{C}^0((0, +\infty))$  and since for any  $t$ ,  $\lim_{x \rightarrow 0} u_{X,\epsilon}(t, x) = S$  (because  $S - x \leq u_X(t, x) \leq S$ ), we see that  $\mathcal{E}_X(u_{X,\epsilon}) \in \mathcal{C}^0([0, T] \times [0, +\infty))$ .

The maximum principle yields (4.8) and (4.9). From the bounds  $u_o(x) \leq u_{X,\epsilon}(t, x) \leq u^{(E)}(t, x)$ , and from the fact that  $\frac{\partial u^{(E)}}{\partial x}(t, 0) = -1$ , we see that  $u_{X,\epsilon}(t, x)$  has a derivative w.r.t.  $x$  at  $x = 0$  and that  $\frac{\partial u_{X,\epsilon}}{\partial x}(t, 0) = -1$  for all  $t \geq 0$ .

By calling  $y_{X,\epsilon}$  the time derivative of  $u_{X,\epsilon}$ , we see that

$$(B.2) \quad \begin{aligned} \frac{\partial y_{X,\epsilon}}{\partial t} + A_X y_{X,\epsilon} - rx 1_{\{x>S\}} \mathcal{V}'_\epsilon(u_{X,\epsilon}) y_{X,\epsilon} &= 0, \quad t \in (0, T], \quad 0 < x < X, \\ y_{X,\epsilon}(t, X) &= 0 \quad t \in (0, T]. \end{aligned}$$

Note that

$$(B.3) \quad -rx 1_{\{x>S\}} \mathcal{V}'_\epsilon(u_{X,\epsilon}) \geq 0.$$

Since  $y_{X,\epsilon} \in \mathcal{C}^0([0, T], V'_X)$ , we have that  $y_{X,\epsilon}(t = 0) = -A_X u_o|_{(0, X)} + rx 1_{x < S} = \frac{\sigma^2 S^2}{2} \delta_{x=S} - B_X u_o|_{(0, X)}$ . It can be seen that  $-B_X u_o|_{(0, X)}$  is a positive distribution in  $V'_X$ , because  $u_o$  is convex: to prove it, one can approximate  $u_o$  in  $V_X$  by a sequence  $u_{o,n}$  of smooth convex functions with bounded support such that  $-B_X u_{o,n} \geq 0$ , and pass to the limit. Therefore,  $y_{X,\epsilon}(t = 0) \geq 0$  in  $V'_X$ . From this, and from (B.2), (B.3), we deduce that  $y_{X,\epsilon} \geq 0$  a.e.. Therefore,  $u_{X,\epsilon}$  is nondecreasing w.r.t.  $t$ .

Finally, the quantities  $\|u_{X,\epsilon}\|_{L^\infty(0, T; L^2(0, X))}$  and  $\|u_{X,\epsilon}\|_{L^2(0, T; V_X)}$  can be bounded independently of  $X$  by taking  $u_{X,\epsilon}$  as a test function in the weak formulation of (4.6) and by observing that the constants in Gårding's inequality for  $A_X$  do not depend of  $X$ .  $\square$

*Proof of Theorem 4.3.* We know that  $u_{X,\epsilon}$  belongs to  $\mathcal{C}^0([\tau, T]; D_X)$  for all  $\tau$ ,  $0 < \tau < T$ . Therefore, from Proposition 4.1,  $u_{X,\epsilon} \in \mathcal{C}^0([\tau, T]; W_X^{3/2+\epsilon})$  for some positive  $\epsilon$ . This yields that for each time  $t > 0$ ,  $u_{X,\epsilon} \in \mathcal{C}^1((0, X])$ . On the other hand, we know that  $u_{X,\epsilon}(t, X) = 0$  for  $t \in [0, T]$ , and  $u_{X,\epsilon} \geq 0$  in  $[0, T] \times [0, X]$ . From the last three observations, we see that for all  $t$ ,  $0 < t \leq T$ ,  $\frac{\partial u_{X,\epsilon}}{\partial x}(t, X) \leq 0$ .



We aim at proving that for each  $t > 0$  there exists a number  $\xi(t)$ ,  $0 \leq \xi(t) < X$ , such that  $\frac{\partial u_{X,\epsilon}}{\partial x}(t, x) \leq 0$  if  $\xi(t) < x < X$ . Indeed, if this were not the case, we would be in one of the following two situations.

(1)  $\frac{\partial u_{X,\epsilon}}{\partial x}(t, x) > 0$  in some interval  $[y(t), X)$ ,  $y(t) < X$ . This implies that  $u_{X,\epsilon}(t, x) < 0$  in  $(y(t), X)$ , which is impossible since  $u(t, \cdot) \geq u_0$ .

(2) There exists a strictly increasing sequence of numbers  $y_n$ ,  $0 < y_n < y_{n+1} < X$ , such that  $\lim_{n \rightarrow \infty} y_n = X$  and  $\frac{\partial u_{X,\epsilon}}{\partial x}(t, y_n) = 0$ , and  $\frac{\partial u_{X,\epsilon}}{\partial x}(t, x)$  is positive for  $x$  in  $(y_{2n}, y_{2n+1})$  and negative for  $x$  in  $(y_{2n+1}, y_{2n+2})$ . The numbers  $y_{2n}$ ,  $n \in \mathbb{N}$ , are local minima of  $u_{X,\epsilon}(t, \cdot)$ . Let us consider the terms entering (4.6) at  $x = y_{2n}$ : we have  $\frac{\partial u_{X,\epsilon}}{\partial t}(t, y_{2n}) \geq 0$  and  $\lim_{n \rightarrow \infty} \frac{\partial u_{X,\epsilon}}{\partial t}(t, y_{2n}) = 0$ . It is clear that  $-\frac{\sigma^2 y_{2n}^2}{2} \frac{\partial^2 u_{X,\epsilon}}{\partial x^2}(t, y_{2n}) \leq 0$  and  $ry_{2n} \frac{\partial u_{X,\epsilon}}{\partial x}(t, y_{2n}) = 0$  because  $y_{2n}$  is a local minimum. We also know that  $ry_{2n}(1 - 1_{\{y_{2n} > S\}}) \mathcal{V}_\epsilon(u_{X,\epsilon}(t, y_{2n})) \geq 0$  and that  $\lim_{n \rightarrow \infty} ry_{2n}(1 - 1_{\{y_{2n} > S\}}) \mathcal{V}_\epsilon(u_{X,\epsilon}(t, y_{2n})) = 0$ . Therefore,

$$\liminf_{n \rightarrow \infty} B_X u_{X,\epsilon}(t, y_{2n}) \geq 0,$$

and, since  $\frac{\partial u_{X,\epsilon}}{\partial x}(t, y_{2n}) = 0$ ,

$$\limsup_{n \rightarrow \infty} \int_{\mathbb{R}} k(z) \left( e^z (1_{z > \log \frac{y_{2n}}{X}} u_{X,\epsilon}(t, y_{2n} e^{-z}) - u_{X,\epsilon}(t, y_{2n})) \right) dz \leq 0.$$

This yields  $\int_{z > 0} k(z) e^z u_{X,\epsilon}(t, X e^{-z}) dz \leq 0$ , which is impossible since  $u_{X,\epsilon} \geq u_0$ .

Therefore, for all  $t > 0$ , the function  $x \mapsto \frac{\partial u_{X,\epsilon}}{\partial x}(t, x)$  is nonpositive in a neighborhood of  $X$ , and  $(\frac{\partial u_{X,\epsilon}}{\partial x}(t, \cdot))_+$  is zero near  $X$ .

Moreover, since  $u_{X,\epsilon}$  is nondecreasing and  $u_{X,\epsilon}(\cdot, X) = 0$ , the function  $t \mapsto \frac{\partial u_{X,\epsilon}}{\partial x}(t, X)$  is nonincreasing. Therefore, there exists  $\tau_0$ ,  $0 \leq \tau_0 \leq T$ , such that  $\frac{\partial u_{X,\epsilon}}{\partial x}(t, X) < 0$  for  $\tau_0 < t \leq T$  and  $\frac{\partial u_{X,\epsilon}}{\partial x}(t, X) = 0$  for  $0 \leq t \leq \tau_0$ .

By taking the derivative of (4.6) w.r.t.  $x$  and multiplying by  $x$ , we see that  $z_{X,\epsilon} = x \frac{\partial u_{X,\epsilon}}{\partial x}$  satisfies

$$\begin{aligned} (B.4) \quad & \frac{\partial z_{X,\epsilon}}{\partial t} + A_X z_{X,\epsilon} - rx 1_{\{x > S\}} \mathcal{V}'_\epsilon(u_{X,\epsilon}) z_{X,\epsilon} \\ & = -rx(1 - 1_{x > S} \mathcal{V}_\epsilon(u_{X,\epsilon})) + rS \mathcal{V}_\epsilon(u_{X,\epsilon}) \delta_{x=S}, \quad t \in (0, T], \quad 0 < x < X, \\ & z_{X,\epsilon}(t = 0, x) = -x 1_{0 < x < S}, \quad 0 < x < X. \end{aligned}$$

Since  $z_{X,\epsilon}(t, X) = 0$  for  $t \in [0, \tau_0]$  ( $\tau_0$  is defined above), the function  $z_{X,\epsilon}|_{t \in (0, \tau_0)} \in L^2(0, \tau_0; V_X)$ . On the other hand,  $z_{X,\epsilon}(t, \cdot) \notin V_X$  for  $t \in (\tau_0, T]$ . In (B.4), for  $t > \tau_0$ ,  $A_X z_{X,\epsilon}(t)$ , i.e.,

$$\begin{aligned} & A_X z_{X,\epsilon}(t, x) = -\frac{\sigma^2 x^2}{2} \frac{\partial^2 z_{X,\epsilon}}{\partial x^2}(t, x) + rx \frac{\partial z_{X,\epsilon}}{\partial x}(t, x) \\ & - \int_{\mathbb{R}} k(z) \left( x(e^z - 1) \frac{\partial z_{X,\epsilon}}{\partial x}(t, x) + e^z (1_{\{z > -\log(X/x)\}} z_{X,\epsilon}(t, x e^{-z}) - z_{X,\epsilon}(t, x)) \right) dz, \end{aligned}$$

has a sense as a distribution and for all  $X' < X$ , belongs to the dual of  $\{v \in V_X, v = 0 \text{ in } (X', X)\}$ .

We split the function  $z_{X,\epsilon}$  into the sum of two functions  $\tilde{z}_{X,\epsilon} \in \mathcal{C}^0([0, T]; L^2(0, X))$

and  $\hat{z}_{X,\epsilon} \in L^2(0, T; V_X)$  which satisfy

$$(B.5) \quad \begin{aligned} & \frac{\partial \hat{z}_{X,\epsilon}}{\partial t} + A_X \hat{z}_{X,\epsilon} - rx 1_{\{x>S\}} \mathcal{V}'_\epsilon(u_{X,\epsilon}) \hat{z}_{X,\epsilon} = rS \mathcal{V}_\epsilon(u_{X,\epsilon}) \delta_{x=S}, \\ & \hat{z}_{X,\epsilon}(t=0, x) = 0, \quad 0 < x < X, \\ & \hat{z}_{X,\epsilon}(t, X) = 0, \quad 0 < t < T, \end{aligned}$$

and

$$(B.6) \quad \begin{aligned} & \frac{\partial \tilde{z}_{X,\epsilon}}{\partial t} + A_X \tilde{z}_{X,\epsilon} - rx 1_{\{x>S\}} \mathcal{V}'_\epsilon(u_{X,\epsilon}) \tilde{z}_{X,\epsilon} = -rx(1 - 1_{x>S} \mathcal{V}_\epsilon(u_{X,\epsilon})), \\ & \tilde{z}_{X,\epsilon}(t=0, x) = -x 1_{0<x<S}, \quad 0 < x < X, \\ & \tilde{z}_{X,\epsilon}(t, X) \leq 0, \quad 0 < t < T. \end{aligned}$$

From the fact that  $u_{X,\epsilon} \geq \underline{u}_X^{(E)}$  and from (4.7), we know that

$$\lim_{\epsilon \rightarrow 0} \|\mathcal{V}_\epsilon(u_{X,\epsilon}(S)) \delta_{x=S}\|_{L^2(0,T;V')} = 0.$$

Thus,  $\lim_{\epsilon \rightarrow 0} \|\hat{z}_{X,\epsilon}\|_{L^2(0,T;V)} = 0$ . One can also prove that  $\hat{z}_{X,\epsilon}(t, 0) = 0$  and that  $\hat{z}_{X,\epsilon} \geq 0$ .

From the last observation and since  $z_{X,\epsilon}(t, 0) = 0$ , we see that for all  $t \in [0, T]$ ,  $\tilde{z}_{X,\epsilon}(t, 0) = 0$ .

We know that  $\tilde{z}_{X,\epsilon}|_{(0,\tau_0)} \in L^2(0, \tau_0, V_X)$ : in  $(0, \tau_0) \times (0, X)$ , we can take  $e^{-Mt}(\tilde{z}_{X,\epsilon})_+$  as a test-function in the equation satisfied by  $\tilde{z}_{X,\epsilon}$ . From Gårding's inequality, choosing  $M$  large enough yields that  $\tilde{z}_{X,\epsilon}(t, \cdot)_+ = 0$  for  $t \in [0, \tau_0]$ .

On the other hand, for  $\tau_1 > \tau_0$ , there exists a constant  $\underline{z} > 0$  such that  $\tilde{z}_{X,\epsilon}(t, X) \leq -\underline{z}$  for  $t \in [\tau_1, T]$ . This and the continuity of  $\tilde{z}_{X,\epsilon}$  imply that there exists  $\underline{X}_{\tau_1}$ ,  $S < \underline{X}_{\tau_1} < X$ , such that  $\tilde{z}_{X,\epsilon} \leq 0$  in  $[\tau_1, T] \times [\underline{X}_{\tau_1}, X]$ . Therefore, in the time interval  $[\tau_1, T]$ , we can take  $(\tilde{z}_{X,\epsilon}(t, x))_+ e^{-Mt}$  as a test-function in the equation satisfied by  $\tilde{z}_{X,\epsilon}$ , even if  $\tilde{z}_{X,\epsilon}$  does not belong to  $V_X$ , (indeed  $(\tilde{z}_{X,\epsilon}(t, \cdot))_+$  does not see the singular behavior of  $z_{X,\epsilon}(t, \cdot)$  near  $X$ ). From Gårding's inequality, we have that for  $M$  large enough,  $t \mapsto e^{-Mt} \int_0^X ((\tilde{z}_{X,\epsilon})_+)^2(t, x) dx$  is nonincreasing in  $(\tau_1, T)$ . We can have  $\tau_1$  tend to  $\tau_0$ . This yields that  $(\tilde{z}_{X,\epsilon})_+ = 0$  in  $(\tau_0, T) \times (0, X)$ . We have proved that  $\tilde{z}_{X,\epsilon} \leq 0$  in  $(0, T) \times (0, X)$ .

Finally, let  $X$  and  $X'$  be two numbers such that  $S < X < X'$ . Call  $\tilde{u}_{X,\epsilon}$  the function obtained by extending  $u_{X,\epsilon}$  by 0 in  $[0, T] \times [X, X']$ . Clearly,  $\tilde{u}_{X,\epsilon} \in C^0([0, T]; K_{X'})$ . It can be seen from (4.6) and from  $\frac{\partial u_{X,\epsilon}}{\partial x}(x = X) \leq 0$  that  $\frac{\partial \tilde{u}_{X,\epsilon}}{\partial t} + A_{X'} \tilde{u}_{X,\epsilon} + rx(1 - 1_{\{x>S\}} \mathcal{V}_\epsilon(\tilde{u}_{X,\epsilon}))$  is a negative distribution in  $(0, T) \times (0, X')$ . This and the maximum principle imply (4.10).  $\square$

*Proof of Proposition 4.8.* It is enough to prove that  $\mu_{X,\epsilon}$  converges to  $\mu_X$  in  $L^1((0, T) \times (0, X))$  because  $0 \leq \mu_{X,\epsilon} \leq rx$ . For that, we make two observations.

- (a) Since  $u_{X,\epsilon}$  is nondecreasing w.r.t.  $t$ ,  $\mu_{X,\epsilon}$  is nonincreasing w.r.t.  $t$ .
- (b)

$$\begin{aligned} \frac{\partial \mu_{X,\epsilon}}{\partial x} &= r 1_{x>S} \mathcal{V}_\epsilon(u_{X,\epsilon}) + rS \mathcal{V}_\epsilon(u_{S,\epsilon}) \delta_{x=S} + r 1_{x>S} \mathcal{V}'_\epsilon(u_{X,\epsilon}) \tilde{z}_{X,\epsilon} \\ &\quad + r 1_{x>S} \mathcal{V}'_\epsilon(u_{X,\epsilon}) \hat{z}_{X,\epsilon}, \end{aligned}$$

where  $\tilde{z}_{X,\epsilon}$  and  $\hat{z}_{X,\epsilon}$  are respectively defined in (B.6) and (B.5). The first three terms in the right-hand side are positive distributions. Let us study the

last term more carefully: we call  $g_\epsilon = r1_{\{x>S\}}\mathcal{V}'_\epsilon(u_{X,\epsilon})\hat{z}_{X,\epsilon}$ . We know that  $\hat{z}_{X,\epsilon}$  is nonnegative and tends to 0 in  $L^2(0,T;V_X)$ . Hence,  $g_\epsilon$  is a nonpositive function. Moreover, let  $\phi_\eta$  be a smooth function defined on  $[0,X]$  such that  $0 \leq \phi_\eta \leq 1$ ,  $\phi_\eta = 1$  for  $0 \leq x \leq X - \eta$ , and  $\phi_\eta(x) = 0$  for  $X - \eta/2 \leq x \leq X$ . Taking  $\phi_\eta$  as a test function in (B.5) yields

$$\lim_{\epsilon \rightarrow 0} \left( \int_0^X \hat{z}_{X,\epsilon}(T,x)\phi_\eta(x)dx + \int_0^T \langle A_X \hat{z}_{X,\epsilon}, \phi_\eta \rangle - \int_0^T \int_0^X g_\epsilon(t,x)\phi_\eta(x)dxdt \right) = 0.$$

This proves that  $\lim_{\epsilon \rightarrow 0} \|g_\epsilon\|_{L^1((0,T) \times (0,X-\eta))} = 0$ .

To summarize,  $\mu_{X,\epsilon}|_{\{x < X-\eta\}}$  is the sum of a nondecreasing function and of  $\tilde{\mu}_{X,\epsilon} = \int_0^x g_\epsilon(t,y)dy$ , and  $\tilde{\mu}_{X,\epsilon}$  and its derivative w.r.t.  $x$  tend to 0 in  $L^1((0,T) \times (0,X-\eta))$ .

From (a) and (b), one sees that the total variation of  $\mu_{X,\epsilon}$  on  $(0,T) \times (0,X-\eta)$  is bounded. Therefore, we can extract a subsequence of  $\mu_{X,\epsilon}|_{\{x < X-\eta\}}$  converging strongly in  $L^1((0,T) \times (0,X-\eta))$ . The limit cannot be anything but  $\mu_X|_{\{x < X-\eta\}}$ , so the whole sequence converges to  $\mu_X|_{\{x < X-\eta\}}$ . Since  $\eta$  is arbitrarily small and  $\mu_{X,\epsilon}$  is bounded, we have that  $\lim_{\epsilon \rightarrow 0} \|\mu_{X,\epsilon} - \mu_X\|_{L^1((0,T) \times (0,X))} = 0$ .

The convergence results for  $u_{X,\epsilon}$  are an easy consequence of the strong convergence of  $\mu_{X,\epsilon}$  to  $\mu_X$ .  $\square$

*Proof of Lemma 6.2.* The proof is similar to an argument given in [2]. For brevity, we shall omit some details. We call  $\mathcal{Q}_\epsilon$  the bilinear form on  $L^2((0,T) \times (0,X)) \times Z_\epsilon$ :

$$\mathcal{Q}_\epsilon(q,v) = \int_0^T \int_0^X \left( \frac{\partial v}{\partial t} + A_{\epsilon,X}v - rx1_{\{x>S\}}\mathcal{V}'_\epsilon(u_\epsilon^*)v \right) q.$$

It is clear that  $\mathcal{Q}_\epsilon$  is continuous. Moreover, there exists a positive constant  $c$ , independent of  $\epsilon$ , such that

$$\inf_{q \in L^2((0,T) \times (0,X))} \sup_{v \in Z_\epsilon} \frac{\mathcal{Q}_\epsilon(q,v)}{\|q\|_{L^2((0,T) \times (0,X))} \|v\|_{Z_\epsilon}} \geq c.$$

To prove this inf-sup condition, take  $v \in L^2(0,T;V_X) \cap H^1(0,T;L^2((0,X)))$  as the weak solution of

$$\frac{\partial v}{\partial t} + A_{\epsilon,X}v - rx1_{\{x>S\}}\mathcal{V}'_\epsilon(u_\epsilon^*)v = q \quad t > 0, \quad v(0,\cdot) = 0.$$

and observe that  $\|v\|_{Z_\epsilon} \leq C\|q\|_{L^2((0,T) \times (0,X))}$  for a constant  $C$  independent of  $\epsilon$ .

Therefore, calling  $Q_\epsilon$  the linear and continuous operator from  $L^2((0,T) \times (0,X))$  to the dual of  $Z_\epsilon$  defined by  $\langle Q_\epsilon p, v \rangle = \mathcal{Q}_\epsilon(p,v)$ , the range of  $Q_\epsilon$  is closed and  $Q_\epsilon$  is injective. On the other hand,  $\mathcal{Q}_\epsilon(q,v)$  for all  $q \in L^2((0,T) \times (0,X))$  implies that  $v = 0$ . Therefore,  $Q_\epsilon^T$  is injective. We have proved that  $Q_\epsilon$  is an isomorphism from  $L^2((0,T) \times (0,X))$  onto the dual of  $Z_\epsilon$  and that its inverse is continuous with a norm independent of  $\epsilon$ .

From this and since  $z \mapsto 2(u_\epsilon^*(T,x_{ob}) - \bar{u})z((T,x_{ob}))$  is a continuous linear form on  $Z_\epsilon$  with a continuity constant independent of  $\epsilon$ , there exists a unique  $p_\epsilon^* \in L^2((0,T) \times (0,X))$  such that for all  $v \in Z_\epsilon$ ,  $\mathcal{Q}_\epsilon(p_\epsilon^*,v) = 2(u_\epsilon^*(T,x_{ob}) - \bar{u})v((T,x_{ob}))$  and  $\|p_\epsilon^*\|_{L^2((0,T) \times (0,X))}$  is bounded independently of  $\epsilon$ . The first part of the lemma is proved.

To prove the second part of the lemma, consider  $G_\epsilon \in L^2((0, T) \times \mathbb{R}_+)$  the solution of the backward Cauchy problem:

$$(B.7) \quad \frac{\partial G_\epsilon}{\partial t} + \frac{(\sigma_\epsilon^*)^2}{2} \frac{\partial^2}{\partial x^2} (x^2 G_\epsilon) - B_\epsilon^T G_\epsilon + \frac{\partial}{\partial x} (rx G_\epsilon) = 0, \quad (t, x) \in [0, T] \times \mathbb{R}_+,$$

$$G_\epsilon(t = T) = -2(u_\epsilon^*(T, x_{ob}) - \bar{u})\delta_{x=x_{ob}},$$

where  $B_\epsilon^T$  is given by (3.6). One can check that  $G_\epsilon$  is smooth for  $t < T$  and that for any integer  $k$  and for any compact  $\omega$  in  $[0, T] \times [0, +\infty)$  which does not contain  $(T, x_{ob})$ , the norm of  $G_\epsilon$  in  $\mathcal{C}^k(\omega)$  is bounded independently of  $\epsilon$ . Also, for the function  $\phi$  defined in Lemma 6.2,  $\phi G_\epsilon \in L^2(0, T; V)$ , with a norm bounded by a constant independent of  $\epsilon$ .

Let  $\chi$  be a smooth function with a compact support contained in  $(0, T] \times [0, X)$ , taking the value 1 in a neighborhood of  $(T, x_{ob})$ , and whose support does not intersect the support of  $\mathcal{V}_\epsilon(u_\epsilon^*)$  for all  $\epsilon$ . For example,  $\chi = 1 - \phi$  can be chosen. With  $A_{\epsilon, X}^T$  defined in Remark 11, it may be checked that  $\|\chi A_\epsilon^T G_\epsilon - A_{\epsilon, X}^T(\chi G_\epsilon)\|_{L^2((0, T) \times (0, X))}$  is bounded independently of  $\epsilon$ . The reason for that is that  $\chi$  is constant near the point where  $G_\epsilon$  is singular.

One sees that  $q_\epsilon^* = p_\epsilon^* - \chi G_\epsilon$  is the unique solution (in the very weak sense defined above, i.e. by duality with the functions in  $Z^\epsilon$ ) of a boundary value problem in  $(0, T] \times (0, X)$ , of the form

$$\begin{aligned} \frac{\partial q_\epsilon^*}{\partial t} - A_{\epsilon, X}^T q_\epsilon^* + rx 1_{\{x > S\}} \mathcal{V}_\epsilon'(u_\epsilon^*) q_\epsilon^* &= g_\epsilon^*, & (t, x) \in [0, T] \times (0, X), \\ q_\epsilon^*(t, X) &= 0, & t \in (0, T), \\ q_\epsilon^*(T, x) &= 0, & x \in (0, X), \end{aligned}$$

where  $g^* = \frac{\partial \chi}{\partial t} G_\epsilon + \chi A_\epsilon^T G_\epsilon - A_{\epsilon, X}^T(\chi G_\epsilon) \in L^2((0, T) \times (0, X))$ . This last boundary value problem has a unique weak solution in  $L^2(0, T; V_X)$ , with a norm bounded independently of  $\epsilon$ . The weak and the very weak solutions coincide. Therefore,  $p_\epsilon^* - \chi G_\epsilon \in L^2(0, T; V_X)$  and  $\|p_\epsilon^* - \chi G_\epsilon\|_{L^2(0, T; V_X)}$  is bounded by a constant independent of  $\epsilon$ . Therefore,  $\phi p_\epsilon^* \in L^2(0, T; V_X)$ , with a norm bounded independently of  $\epsilon$ .  $\square$

**Acknowledgment.** It is a pleasure to thank Rama Cont for introducing me to the topic and for helpful discussions.

#### REFERENCES

- [1] Y. ACHDOU, *Calibration of Lévy Processes with American Options*, in preparation.
- [2] Y. ACHDOU, *An inverse problem for a parabolic variational inequality arising in volatility calibration with American options*, SIAM J. Control Optim., 43 (2005), pp. 1583–1615.
- [3] Y. ACHDOU AND O. PIRONNEAU, *Volatility smile by multilevel least squares*, Int. J. Theor. Appl. Finance, 5 (2002), pp. 619–643.
- [4] Y. ACHDOU AND O. PIRONNEAU, *Computational Methods for Option Pricing*, Frontiers in Applied Mathematics 30, SIAM, Philadelphia, 2005.
- [5] Y. ACHDOU AND O. PIRONNEAU, *Numerical procedure for calibration of volatility with American options*, Appl. Math. Finance, 12 (2005), pp. 201–241.
- [6] R. A. ADAMS, *Sobolev Spaces*, Pure and Applied Mathematics 65, Academic Press, New York, London, 1975.
- [7] L. B. G. ANDERSEN AND R. BROTHERTON-RATCLIFFE, *The equity option volatility smile: An implicit finite difference approach*, J. Comput. Finance, 1 (1998), pp. 5–32.
- [8] M. AVELLANEDA, *Minimum entropy calibration of asset pricing models*, Int. J. Theor. Appl. Finance, 1 (1998), pp. 5–37.

- [9] M. AVELLANEDA, M. FRIEDMAN, C. HOLMES, AND D. SAMPERI, *Calibrating volatility surfaces via relative entropy minimization*, Appl. Math. Finance, 4 (1997), pp. 37–64.
- [10] A. BENSOUSSAN AND J.-L. LIONS, *Impulse Control and Quasivariational Inequalities*,  $\mu$ , translated from the French by J. M. Cole, Gauthier-Villars, Montrouge, 1984.
- [11] M. BERGOUNIOUX AND F. MIGNOT, *Optimal control of obstacle problems: Existence of Lagrange multipliers*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 45–70.
- [12] P. CARR, H. GEMAN, D. B. MADAN, AND M. YOR, *Stochastic volatility for Lévy processes*, Math. Finance, 13 (2003), pp. 345–382.
- [13] R. CONT AND P. TANKOV, *Financial Modelling with Jump Processes*, Chapman & Hall/CRC Financial Mathematics Series, Chapman & Hall/CRC, Boca Raton, FL, 2004.
- [14] R. CONT AND P. TANKOV, *Nonparametric calibration of jump-diffusion option pricing models*, J. Comput. Finance, 7 (2004), pp. 1–49.
- [15] R. CONT AND P. TANKOV, *Retrieving Lévy processes from option prices: Regularization of an ill-posed inverse problem*, SIAM J. Control Optim., 45 (2006), pp. 1–25.
- [16] R. CONT AND E. VOLTCHKOVA, *Finite Difference Methods for Option Pricing in Jump-Diffusion and Exponential Lévy Models*, Rapport Interne 513, CMAP, Ecole Polytechnique, Palaiseau, France, 2003.
- [17] R. CONT AND E. VOLTCHKOVA, *Integro-Differential Equations for Option Prices in Exponential Lévy Models*, Rapport Interne 547, CMAP, Ecole Polytechnique, Palaiseau, France, 2004.
- [18] B. DUPIRE, *Pricing and hedging with smiles*, in Mathematics of Derivative Securities (Cambridge, 1995), Cambridge University Press, Cambridge, UK, 1997, pp. 103–111.
- [19] E. EBERLEIN, *Application of generalized hyperbolic Lévy motions to finance*, in Lévy Processes, Birkhäuser Boston, Boston, 2001, pp. 319–336.
- [20] E. EBERLEIN AND S. RAIBLE, *Term structure models driven by general Lévy processes*, Math. Finance, 9 (1999), pp. 31–53.
- [21] J.-P. FOUQUE, G. PAPANICOLAOU, AND K. R. SIRCAR, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, Cambridge, UK, 2000.
- [22] M. HINTERMÜLLER, *Inverse coefficient problems for variational inequalities: Optimality conditions and numerical realization*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 129–152.
- [23] K. ITO AND K. KUNISCH, *Optimal control of elliptic variational inequalities*, Appl. Math. Optim., 41 (2000), pp. 343–364.
- [24] N. JACKSON, E. SÜLI, AND S. HOWISON, *Computation of deterministic volatility surfaces*, Appl. Math. Finance, 2 (1998), pp. 5–37.
- [25] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Classics in Applied Mathematics 31, SIAM, Philadelphia, 2000.
- [26] J.-L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.
- [27] J.-L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vol. 1, Travaux et Recherches Mathématiques 17, Dunod, Paris, 1968.
- [28] A.-M. MATACHE, P.-A. NITSCHKE, AND C. SCHWAB, *Wavelet Galerkin pricing of American options on Lévy driven assets*, Quant. Finance, 5 (2005), pp. 403–424.
- [29] A.-M. MATACHE, C. SCHWAB, AND T. P. WIHLE, *Fast Numerical Solution of Parabolic Integro-Differential Equations with Applications in Finance*, Research report 1954, IMA, University of Minnesota, Minneapolis, MN, 2004.
- [30] A. M. MATACHE, T. VON PETERSDOFF, AND C. SCHWAB, *Fast deterministic pricing of Lévy driven assets*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 37–71.
- [31] F. MIGNOT AND J.-P. PUEL, *Contrôle optimal d'un système gouverné par une inéquation variationnelle parabolique*, C. R. Acad. Sci. Paris Sér. I Math., 298 (1984), pp. 277–280.
- [32] S. NAYAK AND G. PAPANICOLAOU, *Stochastic Volatility Surface Estimation*, preprint, Stanford University, Stanford, CA, 2006, <http://georgep.stanford.edu/~papanico/pubs.html>.
- [33] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.
- [34] H. PHAM, *Optimal stopping of controlled jump-diffusion processes: A viscosity solution approach*, J. Math. Systems Estim. Control, 8 (1) (1998).

## A QUADRATIC REGULATOR PROBLEM RELATED TO IDENTIFICATION PROBLEMS AND SINGULAR SYSTEMS\*

A. FAVINI<sup>†</sup> AND L. PANDOLFI<sup>‡</sup>

**Abstract.** In this paper we study a new form of the quadratic regulator problem which is suggested by recent applications to singular systems and to identification problems. The new feature of the quadratic regulator problem under study is the penalization of the values taken by the control at individual instants of time.

**Key words.** quadratic regulator problems, identification problems, singular systems

**AMS subject classification.** 49J27

**DOI.** 10.1137/050637248

**1. Introduction.** The quadratic regulator problem has a key role in control theory and it has been studied from many different points of view. We study here a new version which is encountered in the study of singular systems and in the solution of an inverse problem. The system that we study is described by

$$(1) \quad \dot{x} = Ax + Bu, \quad y(t) = Gu(t), \quad s < t \leq T, \quad x(s) = x_0.$$

We note that  $y(t)$  depends on  $u(t)$ .

The operators  $A$ ,  $B$ , and  $G$  in (1) satisfy the following:

- The operator  $A$  generates a  $C_0$ -semigroup on a Hilbert space  $X$ ;
- the operators  $B$  and  $G$  are linear and continuous,  $B \in \mathcal{L}(U, X)$ ,  $G \in \mathcal{L}(U, Y)$ , where  $U$  and  $Y$  are Hilbert spaces;
- we assume  $\ker G^* = 0$ . This is usually not restrictive, but see Remark 2.

The quadratic cost we consider is nonstandard. It has the following form:

$$(2) \quad J_s(x_0; u) = \int_s^T F(x(t) - \xi(t), u(t)) dt + \left\langle \begin{bmatrix} x(\tau_0) - \xi_0 \\ y(\tau_0) \end{bmatrix}, \tilde{M} \begin{bmatrix} x(\tau_0) - \xi_0 \\ y(\tau_0) \end{bmatrix} \right\rangle + \left\langle \begin{bmatrix} x(T) - \xi_T \\ y(T) \end{bmatrix}, M \begin{bmatrix} x(T) - \xi_T \\ y(T) \end{bmatrix} \right\rangle.$$

Here  $\tau_0 \leq T$  and

$$(3) \quad F(x, u) = \langle x, Qx \rangle + \|u\|^2,$$

while  $\xi$ ,  $\xi_0$ ,  $\xi_T$  are given reference signals. We *do not* assume that  $\xi_0$  and  $\xi_T$  are the values of the function  $\xi$  at the corresponding time instants.

The “initial time”  $s$  belongs to an interval on the left of  $T$ ,  $s \in [0, T)$  for definiteness. Note that  $\tau_0 < s$  is possible. In this case the penalization at  $\tau_0$  has to be ignored.

\*Received by the editors August 1, 2005; accepted for publication (in revised form) October 9, 2007; published electronically February 15, 2008. This paper fits the research programs of GNAMPA-INDAM.

<http://www.siam.org/journals/sicon/47-2/63724.html>

<sup>†</sup>Department of Mathematics, University of Bologna, Piazza di Porta San Donato, 5, Bologna, Italy (favini@dm.unibo.it).

<sup>‡</sup>Department of Mathematics, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy (luciano.pandolfi@polito.it).

The goal of this paper is the study of the properties of the infimum, possibly the minimum, of the quadratic cost when the input  $u$  belongs to a subspace of  $L^2(s, T; U)$  which is specified below. However, we mention here the fact that we shall also consider the case that the functions  $u(t)$  are forced to be zero on a subinterval  $(T_0, T)$  of the interval  $(s, T)$  on which the quadratic cost is computed. This is required by the application described in section 2.1.

The novelty of this problem stems from the intermediate and final penalization of the control. Due to these terms, the quadratic functional is neither continuous nor closed (according to [11, p. 313]) on  $L^2(0, T; U)$ .

In order to simplify the notation we shall put

$$(4) \quad \begin{aligned} \Phi_{\tau_0}^1(x, y) &= \left\langle \begin{bmatrix} x - \xi_0 \\ y \end{bmatrix}, \tilde{M} \begin{bmatrix} x - \xi_0 \\ y \end{bmatrix} \right\rangle, \quad \Phi_T^1(x, y) = \left\langle \begin{bmatrix} x - \xi_T \\ y \end{bmatrix}, M \begin{bmatrix} x - \xi_T \\ y \end{bmatrix} \right\rangle, \\ \Phi_{\tau_0}(x, u) &= \Phi_{\tau_0}^1(x, Gu), \quad \Phi_T(x, u) = \Phi_T^1(x, Gu), \\ J_{s, \text{int}}(x_0; u) &= \int_s^T F(x(t) - \xi(t), u(t)) \, dt, \quad J_{\text{fin}}(x_0; u) = \begin{cases} \Phi_{\tau_0}(x, u) + \Phi_T(x, u) & \text{if } s \leq \tau_0, \\ \Phi_T(x, u) & \text{if } s > \tau_0. \end{cases} \end{aligned}$$

Standing assumptions:  $Q$  and  $M$  and  $\tilde{M}$  are symmetric nonnegative continuous linear operators. We assume furthermore that

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^* & M_{22} \end{bmatrix}, \quad M_{22} > cI > 0, \quad \tilde{M} = \begin{bmatrix} \tilde{M}_{11} & \tilde{M}_{12} \\ \tilde{M}_{12}^* & \tilde{M}_{22} \end{bmatrix}, \quad \tilde{M}_{22} > cI > 0.$$

This setup is encountered in particular in the study of important classes of singular control systems, as documented in [8]. Moreover, recently in [1, 10] similar quadratic control problems have been studied for the solution of identification (inverse) problems. See section 2 for details.

*Remark 1.* As we said, the novelty of this problem consists in the presence of the penalization of the final and intermediate values of the control. We have been stimulated to study this problem by identification problems and by the theory of singular control systems. These applications require that we study the problem in an infinite dimensional Hilbert space. This kind of problem has rarely been studied even for finite dimensional systems; see [4]. The problem with  $s = 0$ ,  $\tau_0 = 0$ , and null matrices  $\tilde{M}_{12}$ ,  $\tilde{M}_{22}$  is studied in [13], in the context of Kalman filtering for lumped systems.

Now we comment on apparently more general versions of our problem.

It may seem that  $y = Hx + Gu$  gives a more general problem than (1) and (2). However, it is easily seen that this case can be treated as below. This is left to the reader.

Finally, let us consider the general case

$$F(x, u) = F_0(x, u) + \langle u, u \rangle, \quad F_0(x, u) = \langle x, Qx \rangle + 2\Re \langle x, Q_{12}Gu \rangle + \langle Gu, Q_{22}Gu \rangle \geq 0$$

(see section 2.2 concerning the application to singular systems). This can be reduced to the form (3) as follows: First we absorb  $\langle y, Q_{22}y \rangle = \langle u, G^*Q_{22}Gu \rangle$  into the last term which takes the form  $\langle u, (I + G^*Q_{22}G)u \rangle$ . A coordinate transformation in  $U$  reduces this to  $\langle u, u \rangle$  again. We then apply a feedback  $u = -Q_{12}x + v$  in order to absorb the mixed term. This changes the operator  $A$  to  $A - BQ_{12}$  and does not

change the value of the infimum. We assume that these transformations have been already performed.  $\square$

We are going to study our quadratic regulator problem with  $u \in L^2(s, T; U)$ . Clearly, the quadratic cost is not well defined for every  $u \in L^2(s, T; U)$ . So, we introduce a suitable domain over which the quadratic cost makes sense.

Let

$$\operatorname{ess} \lim_{t \rightarrow \tau_0} u(t) = l$$

when for every  $\epsilon > 0$  there exists  $\delta > 0$  such that the following set has zero Lebesgue measure:

$$\{t, \text{ such that } \|u(t) - l\| > \epsilon, |t - \tau_0| < \delta\}.$$

Analogous definition for the left limits.

If two functions  $u$  and  $u'$  belong to the same equivalence class  $[u] \in L^2(s, T; U)$ , then the  $\operatorname{ess} \lim$  exists for the first if it exists for the second, and the limit itself is the same.

We introduce the linear space  $\mathcal{U}_{\operatorname{ess}}$  of those equivalent classes in  $L^2(s, T; U)$  identified by a representative  $u$  such that the essential limits for  $t \rightarrow T-$  and  $t \rightarrow \tau_0$  exist, and we define for  $u \in \mathcal{U}_{\operatorname{ess}}$ ,

$$u(\tau_0) = \operatorname{ess} \lim_{t \rightarrow \tau_0} u(t), \quad u(T) = \operatorname{ess} \lim_{t \rightarrow T-} u(t).$$

The subspace over which we study our quadratic regulator problem is

$$\mathcal{U} = \mathcal{U}_{\operatorname{ess}} \bigcap \mathcal{U}_0,$$

where  $\mathcal{U}_0 \subseteq L^2(s, T; U)$  is

$$\mathcal{U}_0 = \{u \in L^2(s, T; U), u(t) = 0 \text{ for } t > T_0\}, \quad T_0 > 0.$$

Here  $T_0 > 0$  is fixed. Of course, if  $T_0 < s$ , then  $\mathcal{U}_0 = \{0\}$ .

As already noted, the introduction of the space  $\mathcal{U}_0$  is required by the identification problem described in section 2, while the definition of the subspace  $\mathcal{U}_{\operatorname{ess}}$  is suggested in [12, Chap. 1].

If  $T_0 > T$ , then we intend that  $\mathcal{U} = \mathcal{U}_{\operatorname{ess}}$ . If  $\tau_0 = T$  or if  $\tau_0 < s$ , then we simply ignore the term  $\Phi_{\tau_0}$ .

In general, the quadratic cost we are studying does not admit an optimal control. The optimal control might exist for special initial conditions  $x_0$ . An initial condition which admits an optimal control will be called "optimizable" and  $\mathcal{O}$  is the set of all the optimizable initial conditions.

We already said that, in general, the optimal control does not exist. We define

$$I_s(x_0; \xi_0, \xi_T, \xi) = \inf_{u \in \mathcal{U}} J_s(x_0; u).$$

The most important case in the applications to singular control systems is when the reference signals are equal to zero. In this case we use the simpler notation  $I_s(x_0)$ .

It is convenient to introduce the following additional notation:

- The solution of (1) (initial time is  $s$  and initial value  $x_0$ ) is denoted  $x(t; s, x_0, u)$ .



- When it exists, the optimal control on  $[s, T]$  which corresponds to the initial condition  $x_0$  (at the initial time  $s$ ) is denoted  $u_s^+(\cdot; x_0)$ . The corresponding optimal trajectory is  $x_s^+(\cdot; x_0)$ . We shall see uniqueness of the optimal control, so that the notation is unambiguous.
- An *optimal pair* is the pair of an optimal control and the corresponding trajectory for a given initial condition  $x_0$  at the initial time  $s$ .
- In order to use lighter notation, the index  $s$  is omitted in those sections in which  $s$  is kept fixed, i.e.,  $s = 0$  without restriction; hence we shall write  $J(x_0; u)$  instead of  $J_s(x_0; u)$ , and similarly we shall write  $x^+(\cdot; x_0)$ ,  $u^+(\cdot; x_0)$ ,  $I(x_0; \xi_0, \xi_T, \xi)$ ,  $I(x_0)$ .

The organization of the paper is as follows. In section 2 we present two applications of the problem we are studying and we derive a preliminary characterization of the optimal controls. We shall see that optimal controls will exist only for special initial conditions. But, the infimum of the cost is always finite (and nonnegative). The properties of the infimum are studied in section 3, where it is proved that the infimum is always a continuous quadratic form of the parameters  $x_0, \xi_0, \xi_T$  in  $X$  and  $\xi(\cdot)$  in  $L^2(s, T; X)$ . The infimum of the cost as a function of the initial time  $s \geq 0$  is studied in section 3.2. If, in particular,  $T_0 \geq T = \tau_0$ , we prove that the infimum of the cost is a continuous function of the initial time. With the application to singular systems in mind, this should be contrasted with the result in [3], where it is proved that when the cost is singular but the system is regular, the infimum of the cost is not a continuous function of the “initial time”  $s$ ; it is an upper semicontinuous function of  $s$ .

In section 4 we concentrate on the problem which is most important for the applications to singular systems, that is, the problem with  $\tau_0 = T < T_0$  and reference signals equal to zero. Also in this case the optimal control exists only for suitable initial conditions. The set  $\mathcal{O}$  of the optimizable initial conditions and the corresponding optimal controls are characterized in section 4, where we also clarify the relation of the optimal control with the Riccati equation. A noteworthy result is that the set of the optimizable initial conditions is a closed subspace of  $X$  which is characterized as the kernel of a continuous linear operator.

**2. Examples, applications, and the optimal control.** We present two examples in this section which justify our study. We then characterize the optimal controls and we see that, for a given  $x_0$ , the optimal control, now characterized by a “multipoint problem,” in general does not exist. This justifies our analysis of the properties of the infimum of the cost.

**2.1. An identification problem.** This example is taken from [1]. In applications, delay systems are often used as simple models of more complex distributed parameter systems. Let a signal  $\tilde{y}$  be measured and let us assume that we guessed a delay system

$$(5) \quad y' = Ay + By(t - h)$$

( $y \in \mathbb{R}^n$ ) which could be used as an approximation of the device which produces  $\tilde{y}$ . The solution  $y$  depends on the initial condition  $\phi(\cdot)$ ,  $y(t) = \phi(t)$ ,  $t \in [-h, 0)$  and the approximation is “tuned” by choosing  $\phi$  in the “best possible” way. In [1] the index

to be minimized for the choice of  $\phi$  is a quadratic index,

$$(6) \quad \frac{\alpha}{2} \int_{-h}^0 \|\phi(t) - \tilde{\phi}(t)\|^2 dt + \frac{1}{2} \int_0^T \|y(t) - \tilde{y}(t)\|^2 dt \\ + \frac{\beta}{2} \|\phi(0) - \tilde{\phi}(0)\|^2 + \frac{\gamma}{2} \|y(0) - \tilde{y}(0)\|^2.$$

We show that the previous problem fits into the framework described in section 1. First of all we rewrite (5) as

$$(7) \quad y' = Ay + By(t-h) + Bu(t), \quad y(t) = 0 \text{ if } t < 0, \quad y(0) = y_0.$$

Note that  $y(t) = 0$  for  $t < 0$  since the action of the initial condition  $\phi(\cdot)$  on  $[-h, 0)$  is taken into account by the “control”  $u$  which is

$$u(t) = \begin{cases} \phi(t-h) & \text{if } t \leq T_0 = h, \\ 0 & \text{if } t > T_0 = h. \end{cases}$$

Note that the introduction of  $T_0$  is now required by the very essence of the problem: The initial condition  $\phi$  acts only on a time interval of length  $h$ :  $T_0 = h$ .

After that, we use a standard idea in order to represent the system as a semigroup system on the Hilbert space  $M^2 = \mathbb{R}^n \times L^2(-h, 0; \mathbb{R}^n)$ . We refer to [2, 5] for the technical details and we informally present the idea which leads to this representation: We note that  $y(t-h) = \theta(t, -1)$  if  $\theta(t, s)$  solves

$$\theta_t = \theta_s, \quad \theta(t, 0) = y(t).$$

Combining this equation and (7) suggests the introduction of the following operators:

$$\mathcal{B} : \mathbb{R}^n \rightarrow M^2, \quad \mathcal{B}u = \begin{bmatrix} B \\ 0 \end{bmatrix} u,$$

and  $\mathcal{A}$  from  $M^2$  to itself,

$$\text{dom } \mathcal{A} = \left\{ \begin{bmatrix} y \\ \theta \end{bmatrix}, \theta \in W^{1,2}(-h, 0; \mathbb{R}^n), \theta(0) = y \right\}, \quad \mathcal{A} \begin{bmatrix} y \\ \theta \end{bmatrix} = \begin{bmatrix} Ay + B\theta(-h) \\ \theta'(s) \end{bmatrix}.$$

It is possible to prove that the operator  $\mathcal{A}$  generates a  $C_0$ -semigroup on  $M^2$  and that a solution  $Y(t) = \text{col} \begin{bmatrix} y(t) & \theta(t, \cdot) \end{bmatrix}$  of

$$(8) \quad Y' = \mathcal{A}Y + \mathcal{B}u, \quad Y(0) = \begin{bmatrix} y_0 \\ 0 \end{bmatrix}$$

has the form  $Y(t) = \text{col} \begin{bmatrix} y(t) & y(t+s) \end{bmatrix}$ ,  $s \in (-h, 0)$ , if and only if  $y(t)$  solves (7). Hence, the pair of the quadratic cost (6) and system (8) gives an optimization problem of the same form as described in section 1, when the reference signal  $\tilde{\phi}(t)$  is zero. The presence of the reference signal  $\tilde{\phi}(t)$  to the control  $u$ , as in the cost (6), is not really a more general problem than that described in section 1 since if we replace  $v(t) = u(t) - \tilde{\phi}(t)$ , then  $x(t)$  is replaced by

$$x(t) + \xi_1(t), \quad \xi_1(t) = \int_0^t e^{A(t-s)} B \tilde{\phi}(s) ds,$$

and  $\xi_1(t)$  can be absorbed by the reference signals  $\tilde{y}(t)$  for the state function.

A more interesting observation is in the next remark.

*Remark 2.* Let us consider the case that there are signals that  $y(\tau_0)$  and  $y(T)$  should track, which do not belong to  $\text{im } G$ ; i.e., for example, we want the intermediate penalization at  $\tau_0$  to have the form  $\|Gu(\tau_0) - \omega\|^2$  and  $\omega$  to not be in the closure of the image of  $G$ . In this case it is not possible to replace  $Y$  with the closure of  $\text{im } G$ ; i.e., the condition  $\ker G^* = 0$  is restrictive in this case.

**2.2. Application to singular systems.** Let  $m(x)$  be a continuous nonnegative function defined on a Jordan region  $\Omega$  bounded by a smooth curve, for example, of class  $C^2$  (as we are giving an example, we don't need to use the most general assumptions).

We explicitly assume that  $m(x)$  is zero on a subset of  $\Omega$  of positive measure, so that the following equation is degenerate:

$$(9) \quad \frac{\partial}{\partial t}[m(x)\eta(t, x)] = \mathcal{A}\eta + \mathcal{B}u.$$

The operator  $\mathcal{A}$  is the Laplacian on  $\Omega$ , with Dirichlet homogeneous conditions, and  $\mathcal{B}u$  is given by

$$\mathcal{B}u = \sum_{j=1}^m b_j u_j, \quad b_j \in H^{-1}(\Omega).$$

Hence, the control is finite dimensional,

$$u = \text{col} \begin{bmatrix} u_1 & u_2 & \dots & u_m \end{bmatrix}, \quad m \geq 1.$$

A regularity assumption on the vectors  $b_j$  is shown below.

We impose the initial condition

$$\lim_{t \rightarrow 0+} m(x)\eta(t, x) = m(x)\eta_0(x), \quad \eta_0 \in H_0^1(\Omega).$$

The cost functional most often associated to the problem under study is the cost (2) with  $\tau_0 = T < T_0$  and reference signals equal to zero, i.e., a functional of the form

$$(10) \quad \int_0^T [\|Q\eta(t)\|^2 + \|u(t)\|_{\mathbb{R}^m}^2] dt + \|\eta(T)\|^2$$

(see below for the norms used).

Let  $M$  be the multiplication operator by  $m(x)$ . It is proved in [7, 8] that

$$(11) \quad \|M(\lambda M - \mathcal{A})^{-1}\|_{\mathcal{L}(H^{-1}(\Omega))} \leq \frac{C}{1 + |\lambda|}, \quad \Re \lambda \geq 0.$$

This inequality and the fact that  $\mathcal{A}^{-1} \in \mathcal{L}(H^{-1}(\Omega), H_0^1(\Omega))$  suggest that we choose  $H^{-1}(\Omega)$  as a space for the solution  $\eta$  and for the vectors  $b_j$ . Hence,  $Q$  must be a continuous function on  $H^{-1}(\Omega)$  and, for definiteness, we assume that it takes values in  $H^{-1}(\Omega)$ . Hence, the norms of  $Q\eta(t)$  and  $\eta(T)$  in (10) are the norms of  $H^{-1}(\Omega)$ .

Let us introduce the operator  $\mathcal{T} = M\mathcal{A}^{-1}$  and let us split

$$X = (\ker \mathcal{T}) \oplus \left( \overline{(\text{im } \mathcal{T})} \right).$$

Note that this decomposition is nontrivial when the multiplication operator  $M$  has a nontrivial kernel, i.e., in the case that  $m(x)$  is zero on a subset of  $\Omega$  of positive measure, as we assumed.

We shall assume the following regularity assumption on the vectors  $b_j$ : The vectors  $b_j$  belong to  $(\ker \mathcal{T}) \oplus (\operatorname{im} \mathcal{T})$ .

Using condition (11), it is possible to prove that the restriction  $\tilde{\mathcal{T}}$  of  $\mathcal{T}$  to  $[(\operatorname{im} \mathcal{T})]$  has an inverse (let it be denoted  $\tilde{\mathcal{T}}^{-1}$ ), which generates a holomorphic semigroup.

Let  $P$  denote the projection of  $X$  onto  $\ker \mathcal{T}$  along  $[(\operatorname{im} \mathcal{T})]$ . Then, it turns out that the system can be split as

$$\begin{cases} x' = \tilde{\mathcal{T}}^{-1}x + \tilde{\mathcal{T}}^{-1}(I - P)\mathcal{B}u, \\ y(t) = -P\mathcal{B}u. \end{cases}$$

In order to see this, we perform the following transformation: We define

$$\tilde{x} = \mathcal{A}\eta$$

so that our equation takes the form

$$\frac{d}{dt}(\mathcal{T}\tilde{x}) = \tilde{x} + \mathcal{B}u, \quad \text{i.e.,} \quad \begin{cases} \frac{d}{dt}(\tilde{\mathcal{T}}(I - P))\tilde{x} &= (I - P)\tilde{x} + (I - P)\mathcal{B}u, \\ 0 &= P\tilde{x} + P\mathcal{B}u. \end{cases}$$

We now introduce

$$x = (I - P)\tilde{x}, \quad y = P\tilde{x}.$$

We get a (nontrivial) system of the form (1) when  $P\mathcal{B} \neq 0$ .

Now we choose the space  $Y$ . We note that the equations force  $y = P\tilde{x} = -P\mathcal{B}u$  to belong to the finite dimensional subspace  $(\operatorname{im} P\mathcal{B}) \subseteq \operatorname{im} P$  of  $H^{-1}(\Omega)$ . So, we shall take this finite dimensional subspace as the output space  $Y$ . The reason for this will appear below.

We examine the effects of these transformations on the cost functional (10).

We replace  $\eta = \mathcal{A}^{-1}[x + y]$ ,  $y = -P\mathcal{B}u$ , in the quadratic cost. Norms and inner products being taken in  $H^{-1}(\Omega)$ , we obtain

$$\begin{aligned} \|\eta(T)\|^2 &= \|\mathcal{A}^{-1}x(T)\|^2 + 2\langle \mathcal{A}^{-2}x(T), y(T) \rangle + \|\mathcal{A}^{-1}y(T)\|^2 \\ &= \|\mathcal{A}^{-1}x(T)\|^2 - 2\langle \mathcal{A}^{-2}x(T), P\mathcal{B}u(T) \rangle + \langle \mathcal{A}^{-2}P\mathcal{B}u(T), P\mathcal{B}u(T) \rangle. \end{aligned}$$

Analogously we find that  $\|Q\eta(t)\|^2$  takes the following form, where  $Z = (Q\mathcal{A}^{-1})^*Q\mathcal{A}^{-1}$ :

$$\langle Zx, x \rangle - 2\langle Z\mathcal{B}u, x \rangle + \|Q\mathcal{A}^{-1}\mathcal{B}u\|^2.$$

We get the form of the functionals examined in Remark 1, where we have seen that this seemingly more general functional can be reduced to the form (2).

The system in the form we have obtained does not seem to fit the assumptions in this paper since  $M_{22} = P^*\mathcal{A}^{-2}P$  is not a coercive operator. However, we did not yet use the fact that  $y$  is forced to belong to the finite dimensional space  $Y$ . Taking this into account, coercivity can be recovered. Let  $\mathbf{R}$  denote restriction to  $Y$ ,  $H\mathbf{R}y = Hy$  for every  $y \in Y$ , and every operator  $H$  defined on a linear space which contains  $Y$ . Then

$$\langle P^*\mathcal{A}^{-2}Py(T), y(T) \rangle = \langle \mathbf{R}^*P^*\mathcal{A}^{-2}P\mathbf{R}y(T), y(T) \rangle$$

(an analogous operation is performed on the mixed term of the inner product which defines  $\|\eta(T)\|^2$ ). Hence,

$$M_{22} = \mathbf{R}^* P^* \mathcal{A}^{-2} P \mathbf{R}.$$

This is now a selfadjoint positive operator on a finite dimensional space, whose kernel is 0, because  $y \in Y$  takes values in  $\text{im } P$ ,  $y = Ph$ , so that

$$P \mathbf{R} y = P P h = P h = y$$

and of course  $A^{-2}$  is boundedly invertible. Hence,  $M_{22}$  is coercive, as wanted.

**2.3. The optimal control.** In this section,  $s$  is a fixed value so that we put  $s = 0$  without restriction and we don't explicitly indicate the dependence on  $s$ . An optimality condition is easily obtained, provided that the optimal control exists. We present this condition now and we deduce that in general the optimal control *does not exist*. This justifies our interest in the properties of the infimum of the cost.

Let the optimal control  $u^+(\cdot; x_0)$  exist for a fixed initial condition  $x_0$ . As  $x_0$  is now fixed, we shall simply denote it as  $u^+$ , and  $x^+$  is the corresponding trajectory. Moreover, we introduce the error  $e^+(t) = x^+(t) - \xi(t)$ . If  $t = \tau_0$  or  $t = T$ , we introduce  $e_{\tau_0}^+ = x^+(\tau_0) - \xi_0$ ,  $e_T^+ = x^+(T) - \xi_T$ . In general, these vectors are not the values of  $e^+(t)$  for  $t = \tau_0$  or  $t = T$ .

We compute the Gâteaux derivative of the cost at  $u^+$ . Standard computations give that the following conditions must hold:

$$(12) \quad \begin{aligned} & \int_0^T \langle u^+(t) + B^* p_0(t) + B^* e^{A^*(T-s)} [M_{11} e_T^+ + M_{12} G u^+(T)], v(s) \rangle \, ds \\ & + \int_0^{\tau_0} \langle B^* e^{A^*(\tau_0-s)} [\tilde{M}_{11} e_{\tau_0}^+ + \tilde{M}_{12} G u^+(\tau_0)], v(s) \rangle \, ds = 0 \end{aligned}$$

and

$$(13) \quad \begin{cases} \langle G^* \tilde{M}_{22} G u^+(\tau_0) + G^* \tilde{M}_{12}^* e_{\tau_0}^+, v(\tau_0) \rangle = 0, \\ \langle G^* M_{22} G u^+(T) + G^* M_{12}^* e_T^+, v(T) \rangle = 0, \end{cases}$$

where

$$p_0(t) = \int_t^T e^{A^*(s-t)} Q e^+(s) \, ds.$$

Here  $v$  is *every admissible* input, i.e., every  $v \in \mathcal{U}$ . From (13) we obtain that the following conditions must hold (here we use  $\ker G^* = 0$ . If this condition does not hold, then the equalities below are replaced by inclusions in  $\ker G^*$  of the values taken by suitable functions):

$$(14) \quad \begin{cases} \tilde{M}_{22} G u^+(\tau_0) = -\tilde{M}_{12}^* e_{\tau_0}^+ & \text{if } T_0 > \tau_0, \\ M_{22} G u^+(T) = -M_{12}^* e_T^+ & \text{if } T_0 > T. \end{cases}$$

Condition (12) shows that the optimal control might be discontinuous at  $\tau_0$  and the following relations must hold:

$$(15) \quad u^+(t) = -B^* p(t), \quad 0 \leq t \leq T_0, \quad u(t) = 0, \quad t > T_0,$$

where

$$(16) \quad \begin{aligned} p(t) = & p_0(t) + \mathbf{1}(\tau_0 - t)e^{A^*(\tau_0 - t)} \left\{ \tilde{M}_{11}e_{\tau_0}^+ + \tilde{M}_{12}Gu^+(\tau_0) \right\} \\ & + e^{A^*(T-t)} \left\{ M_{11}e_T^+ + M_{12}Gu^+(T) \right\}. \end{aligned}$$

The function  $\mathbf{1}(t)$  is the Heaviside function, equal to 1 for  $t \geq 0$ , and equal to 0 otherwise (used formally to say that the corresponding term does not appear if the argument of  $\mathbf{1}$  is negative; strictly speaking, in this case this term might be meaningless).

We can use the compatibility conditions (14) in order to replace  $Gu^+(\tau_0)$  and  $Gu^+(T)$  in this last expression, and we find

$$\begin{aligned} p(t) = & p_0(t) + \mathbf{1}(\tau_0 - t)e^{A^*(\tau_0 - t)} \left\{ \tilde{M}_{11} - \tilde{M}_{12}\tilde{M}_{22}^{-1}\tilde{M}_{12}^* \right\} e_{\tau_0}^+ \\ & + e^{A^*(T-t)} \left\{ M_{11} - M_{12}M_{22}^{-1}M_{12}^* \right\} e_T^+. \end{aligned}$$

It is easily seen that when  $G = 0$  and  $T_0 > \tau_0 = T$  these conditions are the usual two-point problem for the proposed cost.

It is now possible to see that the optimal control does not exist, even in the simplest cases. Let  $X = \mathbb{R}$ ,  $A = 0$ ,  $B = 1$ ,  $T_0 = \tau_0 = T = 1$ . Let  $M_{22} = 1$ ,  $M_{11} = 1$ ,  $M_{12} = 0$ ,  $Q = 0$ ,  $G = 1$ . Let the reference signals be zero. The optimality system takes the form

$$\dot{x} = -p, \quad x(0) = 1, \quad \dot{p} = 0, \quad p(1) = x(1)$$

together with the “compatibility condition”

$$0 = M_{22}Gu^+(1) = -M_{22}GB^*p(1), \quad \text{i.e., } p(1) = 0,$$

which cannot be met since if  $p(1) = 0$ , then  $p(t) \equiv 0$  and  $x(t) \equiv 1$ , in contrast with the compatibility condition.

The previous relations have been derived as conditions which an optimal control must satisfy. The converse holds; see the next theorem.

**THEOREM 3.** *If  $x_0$  is an optimizable initial condition, then the optimal control is unique.*

*Let the functions  $x$  and  $p$  solve (1) and (16) when  $u$  is given by (15), and furthermore let conditions (14) be satisfied. Then,  $x_0$  is an optimizable initial condition and the control  $u$  in (15) is the optimal control.*

*Proof.* We prove uniqueness first. Let  $u_1$  and  $u_2$  be different optimal controls for the same initial condition  $x_0$  and let  $v = [u_1 + u_2]/2$ . Then,  $J_{\text{int}}(x_0; v) < J_{\text{int}}(x_0; u_i)$  while  $J_{\text{fin}}(x_0; v) \leq J_{\text{fin}}(x_0; u_i)$  so that  $v$  gives a smaller value of the cost if  $u_1 \neq u_2$ . This is not possible, and thus  $u_1 = u_2$ .

The second part follows because if we replace  $u$  with  $u + v$  and the functions  $x$ ,  $p$ , and  $u$  satisfy the stated condition, we get a quadratic form in  $v$ , which has a minimum when  $v = 0$ .  $\square$

As we said already, this uniqueness result justifies the notation  $u^+(t; x_0)$  chosen to denote the unique optimal control.

The infimum of the cost is always nonnegative. As the optimal control does not exist, we construct a minimizing sequence  $\{\hat{u}_n\}$ , i.e., a sequence such that  $J(x_0; \hat{u}_n)$  converges to the infimum. We define

$$(17) \quad \tilde{u} = \arg \min \{ J_{\text{int}}(x_0; u) + \langle M_{11}[x(T) - \xi_T], x(T) - \xi_T \rangle \}$$

( $\tilde{x}$  is the corresponding solution). Note that this functional is coercive in  $u(\cdot) \in L^2(0, T; U)$  so that the minimum point exists and is unique.

Consider now

$$\tilde{y}_{\tau_0} = \arg \min \Phi_{\tau_0}^1(\tilde{x}(\tau_0), y), \quad \tilde{y}_T = \arg \min \Phi_T^1(\tilde{x}(T), y).$$

The existence of the minima is clear since  $\tilde{M}_{22}$  and  $M_{22}$  are coercive, but the minima do not belong to  $\text{im } G$  in general. As we assumed  $\ker G^* = 0$ , we can find sequences  $\{\hat{u}_{\tau_0}^n\}$  and  $\{\hat{u}_T^n\}$  such that  $G\hat{u}_{\tau_0}^n \rightarrow \tilde{y}_{\tau_0}$  and  $G\hat{u}_T^n \rightarrow \tilde{y}_T$ . These sequences need not be bounded. We choose a sequence  $\{\sigma_n\}$  such that  $\sigma_n \|\hat{u}_{\tau_0}^n\|^2 \rightarrow 0$ , and  $\sigma_n \|\hat{u}_T^n\|^2 \rightarrow 0$  and we define

$$(18) \quad \hat{u}_n(t) = \begin{cases} \tilde{u}(t), & t \notin [\tau_0 - 1/\sigma_n, \tau_0 + 1/\sigma_n] \cup [T - 1/\sigma_n, T], \\ \hat{u}_{\tau_0}^n, & t \in (\tau_0 - 1/\sigma_n, \tau_0 + 1/\sigma_n), \\ \hat{u}_T^n, & t \in [T - 1/\sigma_n, T]. \end{cases}$$

It is easy to check that  $\{\hat{u}_n\}$  is a minimizing sequence.

The limit of  $\{\hat{u}_n\}$  exists in  $L^2(0, T; U)$  but in general is not the optimal control.

A robustness property of this minimizing sequence will be given in section 3.1.

**3. The infimum of the quadratic cost.** In this section the initial time  $s$  is kept fixed,  $s = 0$  for definiteness, and we study the dependence of  $I(x_0; \xi_0, \xi_T, \xi) = I_{s=0}(x_0; \xi_0, \xi_T, \xi)$  on the parameters  $x_0, \xi_0, \xi_T, \xi(\cdot)$ . We shall prove that  $I(x_0; \xi_0, \xi_T, \xi)$  is a continuous quadratic form of its arguments, even in the case  $\ker G \neq 0$ ; see Remark 2 for the interest of this case.

We easily see that the following inequalities hold:

$$(19) \quad \begin{aligned} J(x_0; u) &\geq \|u\|_2^2, \\ \inf_{u \in \mathcal{U}} J(x_0; u) &\leq J(x_0; 0) \leq K \{ \|x_0\|^2 + \|\xi_0\|^2 + \|\xi_T\|^2 + \|\xi(\cdot)\|_2^2 \}. \end{aligned}$$

Here and below,  $\|\cdot\|$  denotes the norms in the spaces  $X, U$ , or  $Y$ , while  $\|\cdot\|_2$  will be used to denote the norm in either  $L^2(0, T; X)$  (as above) or in  $L^2(0, T; U)$ . Moreover, we have the following lemma.

**LEMMA 4.** *Let  $x_0$  be fixed and let  $\{u_n\} \in \mathcal{U}$  be a minimizing sequence. The sequences  $\{Gu_n(\tau_0)\}$  and  $\{Gu_n(T)\}$  are bounded.*

*Proof.* The first inequality (19) shows that  $\{u_n\}$  is bounded in  $L^2(0, T; U)$  so that the sequences  $\{x(\cdot; x_0, u_n)\}$ ,  $\{x(T; x_0, u_n)\}$  are bounded, respectively, in  $C(0, T; X)$  and in  $X$ .

We prove boundedness of the second sequence  $\{Gu_n(T)\}$ . The first sequence is treated analogously.

By contradiction, let  $\lim \|Gu_n(T)\| = +\infty$ . In this case,

$$2\Re \langle x_n(T) - \xi_T, M_{12}Gu_n(T) \rangle + \langle Gu_n(T), M_{22}Gu_n(T) \rangle$$

is unbounded since  $\{x_n(T)\}$  is bounded and  $M_{22}$  is coercive. Hence,  $\{u_n\}$  cannot be a minimizing sequence, which is a contradiction.  $\square$

So, every minimizing sequence  $\{u_n(\cdot)\}$  is bounded in  $L^2(0, T; U)$ , and the sequences  $\{Gu_n(\tau_0)\}$ ,  $\{Gu_n(T)\}$  are bounded in  $U$ . Weak compactness follows. However,  $\mathcal{U}$  is not closed and we cannot use this observation to deduce the existence of the minimum. In spite of this we have the following result, which shows the dependence of the infimum on  $(x_0, \xi_0, \xi_T) \in X^3$  and on  $\xi \in L^2(0, T; X)$ .

THEOREM 5. *The nonnegative functional  $I(x_0; \xi_0, \xi_T, \xi)$  is continuous and quadratic; i.e., there exists  $\mathbf{P} = \mathbf{P}^* \geq 0$ ,  $\mathbf{P} \in \mathcal{L}(X^3 \times L^2(0, T; X))$  such that*

$$I(x_0; \xi_0, \xi_T, \xi) = \langle (x_0, \xi_0, \xi_T, \xi), \mathbf{P}(x_0, \xi_0, \xi_T, \xi) \rangle.$$

*Proof.* For simplicity of notation, we give the proof in the case  $\xi_0 = \xi_T = 0$  and  $\xi = 0$ . Hence, the infimum will be denoted  $I(x_0)$ . The proof can easily be repeated in general; see also Remark 6.

In order to prove the theorem we must show that

1. the transformation  $x \rightarrow I(x)$  is continuous;
2. the parallelogram identity holds for  $I(x)$ ;

see [9, section 9.2].

We prove continuity first. By contradiction, let there exist a sequence  $\{x_n\}$  such that

$$\lim x_n = x_0, \quad \lim I(x_n) \neq I(x_0).$$

This means that there exists  $\epsilon > 0$  such that

$$\text{either (1a) } \limsup I(x_n) > I(x_0) + \epsilon \quad \text{or (1b) } \liminf I(x_n) < I(x_0) - \epsilon.$$

We consider separately (1a) and (1b).

(1a) Let  $\tilde{u}$  be such that

$$|I(x_0) - J(x_0; \tilde{u})| < \epsilon$$

so that

$$\limsup I(x_n) > J(x_0; \tilde{u}) + \epsilon/2.$$

Hence, for a suitable subsequence still denoted  $\{x_n\}$ , we have for every  $n$

$$I(x_n) > J(x_0; \tilde{u}) + \epsilon/3$$

and for every  $u \in \mathcal{U}$  we have

$$J(x_n; u) > J(x_0; \tilde{u}) + \epsilon/3.$$

In particular when  $u = \tilde{u}$ ,

$$J(x_n; \tilde{u}) > J(x_0; \tilde{u}) + \epsilon/3.$$

This cannot be since, with  $\tilde{u}$  fixed,

$$\lim J(x_n; \tilde{u}) = J(x_0; \tilde{u}).$$

Hence, case (1a) is impossible.

If (1b) holds, then we can find  $\{u_n\}$  such that

$$J(x_n; u_n) < I(x_0) - \epsilon/2 < J(x_0; u_n) - \epsilon/2.$$

As already noted, the sequences  $\{u_n\}$ ,  $\{Gu_n(\tau_0)\}$ , and  $\{Gu_n(T)\}$  are bounded. Hence, it is not restrictive to assume

$$u_n \rightharpoonup \tilde{u}, \quad Gu_n(\tau_0) \rightharpoonup \tilde{\eta}, \quad Gu_n(T) \rightharpoonup \eta.$$



Now we use the representation

$$(20) \quad J(x_0; u) = J_{\text{int}}(x_0; u) + J_{\text{fin}}(x_0; u)$$

as in (4). We have

$$J(x_n; u_n) - J(x_0; u_n) < -\epsilon/2$$

so that (with  $x_n(t) = x(t; x_0, u_n)$ )

$$(21) \quad -\frac{\epsilon}{2} > \{J_{\text{int}}(x_n; u_n) - J_{\text{int}}(x_0; u_n)\} \\ + \{[\langle e^{A\tau_0} x_n, \tilde{M}_{11} e^{A\tau_0} x_n \rangle - \langle e^{A\tau_0} x_0, \tilde{M}_{11} e^{A\tau_0} x_0 \rangle] + 2\langle e^{A\tau_0} (x_n - x_0), \tilde{M}_{11} \Lambda_{\tau_0} u_n \rangle \\ (22) \quad + 2\langle e^{A\tau_0} (x_n - x_0), \tilde{M}_{12} G u_n(\tau_0) \rangle\} \\ + \{[\langle e^{AT} x_n, M_{11} e^{AT} x_n \rangle - \langle e^{AT} x_0, M_{11} e^{AT} x_0 \rangle] + 2\langle e^{AT} (x_n - x_0), M_{11} \Lambda_T u_n \rangle \\ (23) \quad + 2\langle e^{AT} (x_n - x_0), M_{12} G u_n(T) \rangle\},$$

where

$$\Lambda_t u = \int_0^t e^{A(t-s)} B u(s) \, ds.$$

The brace in (23) converges to zero because the sequences  $\{u_n\}$  and  $\{G u_n(T)\}$  are bounded and

$$\lim x_n = x_0.$$

Analogously it is seen that the brace in (22) converges to zero.

The terms which contain solely  $u_n$  in the brace (21) cancel out and, as we noted,  $x_n \rightarrow x_0$ ,  $u_n(\cdot) \rightarrow u_0(\cdot)$  so that the limit of the brace in (21) is zero. This is a contradiction and proves continuity of  $I(x)$ .

We now prove the parallelogram identity

$$I(x+y) + I(x-y) = 2[I(x) + I(y)].$$

It is known that the parallelogram identity holds for  $J$ :

$$J(x+y; u+v) + J(x-y; u-v) = 2[J(x; u) + J(y; v)]$$

for every  $x, y$  in  $X$  and  $u, v$  in  $\mathcal{U}$ .

We fix  $x$  and  $y$  and  $\epsilon > 0$  and choose  $u_x$  and  $u_y$  such that

$$J(x; u_x) < I(x) + \epsilon/2, \quad J(y; u_y) < I(y) + \epsilon/2$$

so that

$$J(x+y; u_x + u_y) + J(x-y; u_x - u_y) = 2J(x; u_x) + 2J(y; u_y) < 2[I(x) + I(y)] + 2\epsilon.$$

This proves the inequality

$$I(x+y) + I(x-y) \leq 2[I(x) + I(y)].$$

We prove that the inequality cannot be strict. We prove that if  $\epsilon$  satisfies

$$(24) \quad I(x+y) + I(x-y) \leq 2[I(x) + I(y)] - \epsilon,$$

then  $\epsilon = 0$ .

If (24) holds, then we can find  $\tilde{u}$  and  $\tilde{v}$  such that

$$J(x + y; \tilde{u}) + J(x - y; \tilde{v}) \leq 2[I(x) + I(y)] - \epsilon/2.$$

We define

$$u_0 = \frac{\tilde{u} + \tilde{v}}{2}, \quad v_0 = \frac{\tilde{u} - \tilde{v}}{2}$$

so that

$$2[J(x; u_0) + J(y; v_0)] = J(x + y; u_0 + v_0) + J(x - y; u_0 - v_0) \leq 2[I(x) + I(y)] - \epsilon/2.$$

However, we also have

$$I(x) + I(y) \leq J(x; u_0) + J(y; v_0) \leq [I(x) + I(y)] - \epsilon/2.$$

This shows  $\epsilon = 0$  so that the parallelogram identity holds.

The operator  $\mathbf{P}$  is now constructed by polarization,

$$(25) \quad 4\langle x, \mathbf{P}y \rangle = I(x + y) - I(x - y).$$

*Remark 6.* The previous proof can be repeated verbatim in the case when  $\xi(\cdot)$ ,  $\xi_0$ , and  $\xi_T$  are not zero, but this is not really needed since the cost  $J$  does not depend on these quantities separately but on the triple of the differences  $(\Gamma x_0)(\cdot) - \xi(\cdot)$ ,  $(\Gamma x_0)(\tau_0) - \xi_0$ ,  $(\Gamma x_0)(T) - \xi_T$ , where  $(\Gamma x_0)(t) = e^{At}x_0$ . This triple can be chosen as the “new parameter  $x_0$ ” in the previous proof. See Corollary 11 for a use of this fact.

**3.1. Robustness of the minimizing sequence.** Also in this section the value of the initial time  $s$  is kept fixed,  $s = 0$ , without restriction.

The nonexistence of the optimal control forces us in general to relay on minimizing sequences, i.e., sequences  $\{u_n\}$  such that  $\{J(x_0; u_n)\}$  converges to the infimum. Note that, in general, a minimizing sequence is not convergent.

In this section we prove that the minimizing sequence constructed in (18) is robust under perturbations, in the sense that we explain now.

In practice the parameters which define the cost, i.e., the reference signals and the time  $\tau_0$ , are never exactly equal to their nominal value. The discrepancy between the nominal and actual values of the parameters is usually modeled as follows: Let  $\epsilon \geq 0$  be a parameter whose nominal value should be 0, and let  $\xi_{\tau_0}^\epsilon$ ,  $\xi_T^\epsilon$ ,  $\xi^\epsilon$ ,  $\tau_0^\epsilon$  depend continuously on  $\epsilon$  (respectively, in  $X$ ,  $L^2(0, T; X)$ , and  $[0, T]$ ), with the nominal values taken for  $\epsilon = 0$  (so, the exponent  $\epsilon$  denotes functional dependence and not a power).

We recall the definition of  $\hat{u}_n$  from (18) and denote by  $\{\hat{u}_n^\epsilon\}$  the minimizing sequence constructed as in (18), but for the case when the reference signals and intermediate penalization time are those which correspond to a certain value  $\epsilon > 0$ . Let  $J^\epsilon$ ,  $J_{\text{int}}^\epsilon$ ,  $\Phi_{\tau_0}^\epsilon$ ,  $\Phi_T^\epsilon$  be the functionals in (4), computed with the parameters which correspond to the value  $\epsilon \geq 0$ .

We compare the values of  $J(x_0; \hat{u}_n)$  and  $J(x_0; \hat{u}_n^\epsilon)$  and prove that  $J(x_0; \hat{u}_n^\epsilon)$  approximates  $I(x_0; \xi_0, \xi_T, \xi)$  when  $n$  is large and  $\epsilon$  sufficiently close to zero.

In this sense we say that the construction of the minimizing sequence is *robust* under the action of the perturbations.

We prove the following.

THEOREM 7. *Let  $\delta > 0$ . There exists  $\eta_\delta > 0$  such that if  $0 \leq \epsilon < \eta_\delta$ , then we have*

$$(26) \quad \|J(x_0; \hat{u}_n) - J^\epsilon(x_0; \hat{u}_n^\epsilon)\| < \delta.$$

*Proof.* As a preliminary observation we note that if  $\eta^\epsilon$  depends continuously on  $\epsilon$  and  $M$  is fixed and coercive, the minimum point  $v^\epsilon$  of

$$\langle Mv, v \rangle + \langle \eta^\epsilon, v \rangle, \quad \text{i.e.,} \quad v^\epsilon = -M^{-1}\eta^\epsilon,$$

is a continuous function of  $\epsilon$ . Also the minimum value is a continuous function of  $\epsilon$ .

We apply to the perturbed data the construction for the minimizing sequence described in section 2.3. The construction is in two steps. In the first step we use (17) and construct  $\tilde{u}^\epsilon$ . This does not depend on  $n$ ; it is a continuous function of  $\epsilon$ , and also the corresponding solution  $\tilde{x}^\epsilon(\cdot)$  is a  $C(0, T; X)$  continuous function of  $\epsilon$  thanks to the observation above. Moreover,

$$\{J_{\text{int}}^\epsilon(x_0; \tilde{u}^\epsilon) + \langle [\tilde{x}^\epsilon(T) - \xi_T^\epsilon], M_{11}[\tilde{x}^\epsilon(T) - \xi_T^\epsilon] \rangle\}$$

is a continuous function of  $\epsilon$ . Hence, in order to prove the theorem it is sufficient to consider the remaining terms.

We consider sequences  $\{\hat{u}_{\tau_0}^{\epsilon, n}\}$  and  $\{\hat{u}_T^{\epsilon, n}\}$  and  $\sigma_n^\epsilon$  such that for every  $\epsilon$ , we have

$$\lim_n G\hat{u}_{\tau_0}^{\epsilon, n} = \tilde{y}_{\tau_0}^\epsilon, \quad \lim_n G\hat{u}_T^{\epsilon, n} = \tilde{y}_T^\epsilon,$$

and  $\sigma_n^\epsilon$  such that for each  $\epsilon$  fixed, we have

$$\lim_n \sigma_n^\epsilon \|\hat{u}_{\tau_0}^{\epsilon, n}\|^2 = 0, \quad \lim_n \sigma_n^\epsilon \|\hat{u}_T^{\epsilon, n}\|^2 = 0.$$

We then define  $\hat{u}_n^\epsilon(t)$  with a formula analogous to (18). Thanks to the choice of  $\sigma_n^\epsilon$ , the contribution to the cost of the restriction of  $\hat{u}_n^\epsilon(t)$  to the intervals  $(\tau_0^\epsilon - \sigma_n^\epsilon, \tau_0^\epsilon + \sigma_n^\epsilon)$  and  $(T - \sigma_n^\epsilon, T]$  tends to zero uniformly with respect to  $\epsilon$  so that we have (26).  $\square$

**3.2. The function  $s \rightarrow I_s(x_0; \xi_0, \xi_T, \xi)$ .** In this section the parameters  $x_0, \xi_0, \xi_T, \xi$  are kept fixed while  $s$  varies on the interval  $[0, T]$ . We are going to study the regularity properties of the function  $s \rightarrow I_s(x_0; \xi_0, \xi_T, \xi)$ .

We recall that  $x(t; s, x_0, u)$  is the solution of (1).

We need the following lemma, which is analogous to Lemma 4.

LEMMA 8. *Let  $s_n \rightarrow \hat{\chi}$  and let  $\{u_n\}$  be a sequence in  $L^2(s_n, T; U)$  such that  $J_{s_n}(x_0; u_n) < \beta$  ( $\beta$  is a fixed number). Then the sequences  $\{Gu_n(\tau_0)\}$  and  $\{Gu_n(T)\}$  are bounded (of course the sequence  $\{Gu_n(\tau_0)\}$  is considered only if  $\hat{\chi} \leq \tau_0$ ).*

*Proof.* We see from (19) that  $\{u_n\}$  is a bounded sequence in  $L^2(0, T; U)$  so that the sequence  $\{x(T; s_n, x_0, u_n)\}$  is bounded in  $X$ . The result now follows as in Lemma 4.  $\square$

THEOREM 9. *The function  $s \rightarrow I_s(x_0; \xi_0, \xi_T, \xi)$  is continuous for every  $x_0$  and  $s \neq \tau_0$ . For  $s = \tau_0$ , it is left continuous and upper semicontinuous from the right.*

*Proof.* As above, we prove the result for  $I_s(x_0)$ . The general case when  $\xi_0, \xi_T$ , and  $\xi(\cdot)$  are nonzero is completely analogous. We prove separately upper and lower semicontinuity at every  $\hat{\chi} \in [0, T]$ ,  $\hat{\chi} \neq \tau_0$ . The arguments we present hold also for  $\hat{\chi} = \tau_0$  but only from the left, while only the argument concerning upper semicontinuity holds at  $\tau_0$  from the right.

We prove first

$$\limsup_{s \rightarrow \hat{\chi}} I_s(x_0) \leq I_{\hat{\chi}}(x_0).$$

Let us fix  $\alpha > I_{\hat{\chi}}(x_0)$ . We prove

$$\limsup_{s \rightarrow \hat{\chi}} I_s(x_0) \leq \alpha.$$

Let  $u$  satisfy

$$I_{\hat{\chi}}(x_0) < J_{\hat{\chi}}(x_0; u) < \alpha.$$

Now we distinguish the two limits for  $s \rightarrow \hat{\chi}+$  and  $s \rightarrow \hat{\chi}-$ .

If  $s \rightarrow \hat{\chi}+$ , we use

$$x(t; s, x_0, u) = \left\{ e^{A(t-s)} x_0 + \int_s^t e^{A(t-r)} B u(r) \, dr \right\} \longrightarrow x(t; \hat{\chi}, x_0, u)$$

uniformly on every  $[r, T]$ ,  $r > \hat{\chi}$ , and it remains bounded. Hence,

$$\lim_{s \rightarrow \hat{\chi}+} J_s(x_0; u|_{[s, T]}) = \begin{cases} J_{\hat{\chi}}(x_0; u), & \hat{\chi} \neq \tau_0, \\ J_{\tau_0}(x_0; u) - \Phi_{\tau_0}(x(\tau_0), u(\tau_0)), & \hat{\chi} = \tau_0. \end{cases}$$

In both cases, the expression on the right-hand side is less than or equal to  $J_{\hat{\chi}}(x_0; u) \leq \alpha$ .

For every  $s$ , we have

$$I_s(x_0) \leq J_s(x_0; u|_{[s, T]}).$$

Hence, we have

$$\limsup_{s \rightarrow \hat{\chi}+} I_s(x_0) \leq \alpha.$$

Analogously, let  $s \rightarrow \hat{\chi}-$ . The input  $u$  is as above, and for every  $s$  we introduce  $u_s$ :

$$u_s(t) = \begin{cases} u(t) & \text{if } t > \hat{\chi}, \\ u(\tau_0) & \text{if } s < t < \hat{\chi} \end{cases}$$

so that we still have

$$\lim_{s \rightarrow \hat{\chi}-} x(t; s, x_0, u) = x(t; \hat{\chi}, x_0, u).$$

The limit is uniform on  $[\hat{\chi}, T]$ . The result follows as above (and we don't need to single out the case  $\hat{\chi} = \tau_0$ ).

We prove now lower semicontinuity,

$$\liminf_{s \rightarrow \hat{\chi}} I_s(x_0) \geq I_{\hat{\chi}}(x_0).$$

We prove this for  $\hat{\chi} \neq \tau_0$ , and when  $\hat{\chi} = \tau_0$  we prove the result only for the left limit. We choose any  $\beta$  and  $\beta'$  such that  $\beta < \beta' < I_{\hat{\chi}}(x_0)$  and we prove

$$\liminf_{s \rightarrow \hat{\chi}} I_s(x_0) \geq \beta.$$

By contradiction, let  $\liminf_{s \rightarrow \hat{\chi}} I_s(x_0) < \beta$  and let  $s_n \rightarrow \hat{\chi}$ ,  $\{u_n\}$  be sequences such that

$$J_{s_n}(x_0; u_n) < \beta'.$$

If  $\hat{\chi} = \tau_0$ , we assume  $s_n < \tau_0$ . Lemma 8 shows that  $\{u_n(\cdot)\}$  is bounded in  $L^2(0, T; U)$  and that  $\{Gu_n(\tau_0)\}$ ,  $\{Gu_n(T)\}$  are bounded. We first consider the case  $s \rightarrow \hat{\chi}-$ . We shall prove below that

$$(27) \quad |J_{\hat{\chi}}(x_0; u_n) - J_{s_n}(x_0; u_n)| \longrightarrow 0.$$

Accepting this, there exists  $N_\epsilon$  such that for  $n > N_\epsilon$  we have

$$I_{\hat{\chi}}(x_0) \leq J_{\hat{\chi}}(x_0; u_n) < \beta'$$

and this is a contradiction because we assumed  $\beta' < I_{\hat{\chi}}(x_0)$ .

In order to complete this argument, we give an estimate of the absolute value in (27) (recall  $s \rightarrow \hat{\chi}-$ ). We represent

$$(28) \quad J_{\hat{\chi}}(x_0; u_n) - J_{s_n}(x_0; u_n)$$

$$(29) \quad = \int_{\hat{\chi}}^T \left\{ \left\langle \begin{bmatrix} x(t; \hat{\chi}, x_0, u_n) \\ u_n(t) \end{bmatrix}, Q \begin{bmatrix} x(t; \hat{\chi}, x_0, u_n) \\ u_n(t) \end{bmatrix} \right\rangle + |u_n(t)|^2 \right\} dt$$

$$(30) \quad + \Phi_{\tau_0}(x(\tau_0; \hat{\chi}, x_0, u), u(\tau_0)) + \Phi_T(x(T; \hat{\chi}, x_0, u), u(T))$$

$$- \int_{s_n}^T \left\{ \left\langle \begin{bmatrix} x(t; s_n, x_0, u_n) \\ u_n(t) \end{bmatrix}, Q \begin{bmatrix} x(t; s_n, x_0, u_n) \\ u_n(t) \end{bmatrix} \right\rangle + |u_n(t)|^2 \right\} dt$$

$$- \Phi_{\tau_0}(x(\tau_0; s_n, x_0, u), u(\tau_0)) - \Phi_T(x(T; s_n, x_0, u), u(T)).$$

Now we represent the sum of the integrals in (29) and in (30) as

$$\int_{\hat{\chi}}^T = \int_{s_n}^{\hat{\chi}} + \int_{\hat{\chi}}^T.$$

We note that the quadratic terms which contain only  $u_n$  cancel out. Boundedness of the integrand (in the square norm) proves that the first integral on the right-hand side converges to zero. The second integral converges to zero because  $\{u_n\}$  is bounded in  $L^2(0, T; U)$ , while

$$\|x(\cdot; \hat{\chi}, x_0, u_n) - x(\cdot; s, x_0, u_n)\|_2$$

converges to zero uniformly.

Analogous arguments prove that the contributions of the final and intermediate costs, represented by the  $\Phi$ 's, tend to zero. We stress the fact that the purely quadratic terms in  $u(\cdot)$ ,  $Gu(\tau_0)$ , and  $Gu(T)$  cancel out.

If  $s_n \rightarrow \hat{\chi}_+$ , an analogous argument holds, provided that  $\hat{\chi} \neq \tau_0$ , by using

$$\tilde{u}_n(t) = \begin{cases} u_n(t) & \text{if } s_n \leq t \leq T, \\ 0 & \text{if } \hat{\chi} \leq t \leq s_n. \end{cases}$$

We cannot repeat this last argument for  $\hat{\chi} = \tau_0$  because if  $s > \tau_0$ , the contribution of  $\Phi_{\tau_0}$  is not accounted for.  $\square$

Note that the condition  $\ker G^* = 0$  has not been used in the previous proof and that the previous result holds for every  $T_0$ . Of course, if  $T_0 < \tau_0$ , the control is not penalized at  $\tau_0$  and we have continuity also at  $\tau_0$ .

Moreover, we have the following.

COROLLARY 10. *Let  $\tau_0 = 0$ . The function  $s \rightarrow I_s(x_0)$  is continuous on  $(0, T]$ .*

COROLLARY 11. *Let  $\tau_0 < T$ . We have that  $I_T(x_0)$  is a quadratic function of  $(x_0 - \xi_T)$ ; i.e., there exists  $N = N^* \geq 0$  in  $\mathcal{L}(X)$  such that*

$$(31) \quad I_T(x_0) = \inf_u \left\langle \begin{bmatrix} x_0 - \xi_T \\ Gu \end{bmatrix}, M \begin{bmatrix} x_0 - \xi_T \\ Gu \end{bmatrix} \right\rangle = \langle (x_0 - \xi_T), N(x_0 - \xi_T) \rangle.$$

We note that  $G^*M_{22}G$  is not assumed coercive, so that the infimum in (31) is not generally a minimum or realized by a unique minimum point. However, we note the following.

LEMMA 12. *Let  $x_0$  be an optimizable initial condition. Then, the optimal control is unique. Furthermore, if the optimal control  $u^+(\cdot; x_0) \in \mathcal{U}$  exists for the initial condition  $x_0$ , then we have*

$$(32) \quad \begin{cases} u^+(\tau_0; x_0) = \arg \min_{u \in \mathcal{U}} \left\langle \begin{bmatrix} x^+(\tau_0; x_0) \\ Gu \end{bmatrix}, \tilde{M} \begin{bmatrix} x^+(\tau_0; x_0) \\ Gu \end{bmatrix} \right\rangle, \\ u^+(T; x_0) = \arg \min_{u \in \mathcal{U}} \left\langle \begin{bmatrix} x^+(T; x_0) \\ Gu \end{bmatrix}, M \begin{bmatrix} x^+(T; x_0) \\ Gu \end{bmatrix} \right\rangle. \end{cases}$$

*Proof.* Unicity was already proved in Theorem 3. Conditions (32) might be proved from (13). Instead, we give an independent proof. We consider the optimal control and  $t = T$ . We recall that  $u(\cdot) \rightarrow x(\cdot; x_0, u)$  is a linear continuous transformation from  $L^2(0, T; U)$  to  $C(0, T; X)$ . By contradiction,  $u(\cdot)$  does not satisfy the second condition in (32). Then, we can find  $\tilde{u}_0 \in U$  such that

$$\left\langle \begin{bmatrix} x^+(T; x_0) \\ G\tilde{u}_0 \end{bmatrix}, M \begin{bmatrix} x^+(T; x_0) \\ G\tilde{u}_0 \end{bmatrix} \right\rangle < \left\langle \begin{bmatrix} x^+(T; x_0) \\ Gu^+(T) \end{bmatrix}, M \begin{bmatrix} x^+(T; x_0) \\ Gu^+(T) \end{bmatrix} \right\rangle - \epsilon_0, \quad \epsilon_0 > 0.$$

We change the definition of  $u^+(\cdot)$  in a short time interval  $[T - \sigma, T]$ ,

$$\tilde{u}(t) = \begin{cases} u^+(t; x_0) & \text{if } t < T - \sigma, \\ \tilde{u}_0 & \text{if } T - \sigma \leq t \leq T. \end{cases}$$

Clearly we can find  $\sigma$  so small that

$$|J_{\text{int}}(x_0, \tilde{u}) - J_{\text{int}}(x_0, u^+)| < \epsilon_0/8,$$

$$\left| \left\langle \begin{bmatrix} x^+(T; x_0) \\ \tilde{u}_0 \end{bmatrix}, M \begin{bmatrix} x^+(T; x_0) \\ \tilde{u}_0 \end{bmatrix} \right\rangle - \left\langle \begin{bmatrix} x(T; x_0, \tilde{u}) \\ \tilde{u}_0 \end{bmatrix}, M \begin{bmatrix} x(T; x_0, \tilde{u}) \\ \tilde{u}_0 \end{bmatrix} \right\rangle \right| < \epsilon_0/8$$

so that

$$J(x_0; \tilde{u}) < J(x_0; u) - \epsilon_0/2.$$

Hence,  $u(\cdot)$  is not the optimal control of  $x_0$ . The proof of the first equality in (32) is similar.  $\square$

**4. Optimal control, Riccati equation, and dissipation inequality.** In this section we consider the case which is most important for the applications to singular systems,  $T_0 > \tau_0 = T$ , and zero reference signals. The initial time of interest is  $s = 0$ , but we shall use dynamic programming arguments so that we have to consider also the general case  $s \in [0, T]$ : If  $x_0$  is optimizable with initial time 0, then dynamic programming shows that the restriction of  $u^+(t; x_0)$  to  $[s, T]$  is optimal for the “initial condition”  $x^+(s; x_0)$  assigned at the initial time  $s$ . Moreover, we recall from section 3.2:

$$(33) \quad I_s(x_0) = \langle P(s)x_0, x_0 \rangle \quad \text{and} \quad P(s) = P^*(s) \in \mathcal{L}(\mathcal{X}), \quad P(s) \geq 0.$$

Due to the fact that reference signals are now equal to zero, we can derive the properties of the functional  $I_s(x_0)$  using an auxiliary (standard) quadratic regulator problem, and we can characterize the optimizable initial conditions and the optimal controls.

We rewrite the characterization of the optimal controls in the special case we are studying now: Theorem 3 takes the following form (the solutions of the differential equations must be intended in the weak sense. In particular,  $p(t)$  is now a continuous function).

**THEOREM 13.** *Let  $x_0 \in X$  and let us consider the following two-point problem:*

$$(34) \quad \begin{cases} x(0) = x_0, \\ \dot{x} = Ax(t) + Bu, & u(t) = -B^*p(t), \\ \dot{p} = -A^*p - Qx, \\ p(T) = \{M_{11} - M_{12}M_{22}^{-1}M_{12}^*\}x(T), \end{cases}$$

*and let us consider the compatibility condition*

$$(35) \quad Gu(T; x_0) = -M_{22}^{-1}M_{12}^*x(T; x_0).$$

*The initial condition  $x_0$  is optimizable if and only if problem (34) has a solution  $(x(\cdot), p(\cdot))$  such that the function  $u(t) = -B^*p(t)$  satisfies the compatibility condition (35). In this case, the function  $-B^*p(t)$  is the optimal control.*

Now we note that the two-point problem (34) (without the compatibility condition (35)) is always solvable as follows.

**THEOREM 14.** *We have*

$$(36) \quad M_{11} - M_{12}M_{22}^{-1}M_{12}^* \geq 0$$

*so that the two-point problem (34) is solvable. Moreover, the vector  $x(T)$  is a linear and continuous function of  $x_0$ .*

*Proof.* The positivity condition (36) follows from  $M \geq 0$ .

The two-point problem (34) (without the compatibility condition (35)) is the two-point problem of the standard quadratic cost

$$(37) \quad \tilde{J}_s(x_0; u) = \int_s^T \{ \langle Qx(t), x(t) \rangle + \langle u(t), u(t) \rangle \} dt + \langle \mathcal{M}x(T), x(T) \rangle,$$

where (see the first line in (34)) the initial time is  $s = 0$  and  $\mathcal{M} = M_{11} - M_{12}M_{22}^{-1}M_{12}^* \geq 0$  (below we shall write  $\tilde{J}(x_0; u)$  when  $s = 0$ ). This is our “auxiliary” cost function.

The functional (37) admits a unique optimal control  $\tilde{u}(t; x_0)$  for every  $x_0$ . Furthermore, if  $\tilde{x}(t; x_0)$  is the optimal trajectory of  $x_0$ , then the transformation  $x_0 \rightarrow \tilde{x}(t; x_0)$  is linear and continuous.  $\square$

We use the last statement of the theorem as follows: The compatibility condition (35) can be written as

$$(38) \quad \begin{aligned} \mathcal{L}x(T) &= 0, \\ \mathcal{L} &= -M_{22}GB^* [M_{11} - M_{12}M_{22}^{-1}M_{12}^*] - M_{12}^*. \end{aligned}$$

Hence, we have the following.

**THEOREM 15.** *The initial condition  $x_0$  is optimizable if and only if  $\mathcal{L}x(T) = 0$ . In particular, the set  $\mathcal{O}$  of the optimizable initial conditions is a closed subspace of  $X$ .*

The second assertion follows since we already noted continuity of the transformation  $x_0 \rightarrow x(T)$  when  $(x(\cdot), p(\cdot))$  solves (34).

Of course, the optimal control  $\tilde{u}(t; x_0)$  for  $\tilde{J}(x_0; u)$  is given by

$$\tilde{u}(\cdot; x_0) = -B^*p(\cdot).$$

It always exists, while  $u^+(\cdot; x_0)$  exists only if  $x_0$  is optimizable. The previous considerations show the following.

**THEOREM 16.** *If  $x_0$  is optimizable, then  $\tilde{u}(\cdot; x_0) = u^+(\cdot; x_0)$ .*

The Riccati equation is an important tool in the quadratic regulator problem. Hence we now relate our problem to a differential Riccati equation. It is known that the component  $p$  of the two-point problem (34) is expressed as

$$p(t) = \mathcal{P}(t)x(t),$$

where  $\mathcal{P}(t)$  solves the Riccati equation

$$\begin{aligned} \frac{d}{dt} \langle \mathcal{P}(t)x, y \rangle &= -\langle Ax, \mathcal{P}(t)y \rangle - \langle \mathcal{P}(t)x, Ay \rangle - \langle Qx, y \rangle \\ &\quad + \langle B^*\mathcal{P}(t)x, B^*\mathcal{P}(t)y \rangle \quad \forall x, y \in \text{dom} A; \quad \mathcal{P}(T) = \mathcal{M}, \end{aligned}$$

so that the optimal control of the auxiliary functional  $\tilde{J}(x_0; u)$  in (37) is

$$(39) \quad \tilde{u}(t) = -B^*\mathcal{P}(t)x(t),$$

where  $x$  solves the closed loop equation

$$(40) \quad \dot{x} = (A - BB^*\mathcal{P}(t))x, \quad x(0) = x_0.$$

As we noted, this is also the optimal control of  $J(x_0; u)$  when  $x_0$  is optimizable. Hence, we have the next theorem.

**THEOREM 17.** *Let  $x_0$  be optimizable. Then, the optimal control has the feedback form*

$$u^+(t; x_0) = -B^*\mathcal{P}(t)x^+(t; x_0).$$

We recall that

$$(41) \quad \min_{u \in L^2(s, T)} \tilde{J}_s(x_0; u) = \langle x_0, \mathcal{P}(s)x_0 \rangle, \quad \min_{u \in L^2(0, T)} \tilde{J}(x_0; u) = \langle x_0, \mathcal{P}(0)x_0 \rangle.$$



Using this equality and the previous formulas which do depend on the assumption  $\ker G^* = 0$ , we can now prove the next theorem.

**THEOREM 18.** *Let the reference signals  $\xi(\cdot)$ ,  $\xi_0$ ,  $\xi_T$  be zero and let  $\ker G^* = 0$ . For every  $s \in [0, T]$ , we have*

$$(42) \quad P(s) = \mathcal{P}(s),$$

where  $P(s)$  is the operator of the quadratic form  $I_s(x_0)$ .

*Proof.* We note that

$$(43) \quad J(x_0; u) = \tilde{J}(x_0; u) + \{ \Phi_T(x(T; x_0, u), u(T)) - \langle x(T; x_0, u), \mathcal{M}x(T; x_0, u) \rangle \},$$

$$\mathcal{M} = [M_{11} - M_{12}M_{22}^{-1}M_{12}^*].$$

The cost  $\tilde{J}(x_0; u)$  is the auxiliary cost defined in (37) and the brace is nonnegative. Hence,

$$P(s) \geq \mathcal{P}(s).$$

We use the same idea as that used to derive the minimizing sequence (18) in order to prove that the inequality cannot be strict. We consider the optimal control  $\tilde{u}(t)$  in (39) and, for every  $n$ , we define the sequence

$$u_n(t) = \begin{cases} \tilde{u}(t) & \text{if } t < T_n, \\ u_{n,T} & \text{if } T_n < t < T, \end{cases}$$

where  $u_{n,T}$  satisfies

$$\|Gu_{n,T} - \tilde{y}\| < \frac{1}{n}, \quad \text{where } \tilde{y} = \arg \min \left\langle \begin{bmatrix} \tilde{x}(T; s, x_0) \\ y \end{bmatrix}, M \begin{bmatrix} \tilde{x}(T; s, x_0) \\ y \end{bmatrix} \right\rangle.$$

It is easily seen that letting  $T_n \rightarrow T$  we have constructed a minimizing sequence for  $J$ , so that the inequality cannot be strict.  $\square$

*Remark 19.* The referees stimulated us to try to understand whether

$$(44) \quad \min_{L^2(0,T;U)} \tilde{J}(x_0; u) = \inf_{\mathcal{U}} J(x_0; u)$$

in every case. Combining equalities (33), (41), and (42), we see that (44) holds if  $\ker G^* = 0$ . Taking into account the observation in Remark 2, it is interesting to see that equality (44) does not hold in general if  $\ker G^* \neq 0$ .

We present the following finite dimensional counterexample:  $T = 1 = \tau_0 < T_0$ ,  $X = U = \mathbb{R}$ , while  $Y = \mathbb{R}^2$ . The system is

$$\dot{x} = u, \quad y = \begin{bmatrix} 0 \\ 1 \end{bmatrix} u,$$

and the cost is

$$J(x_0; u) = \int_0^1 u^2(s) \, ds + \left\langle \begin{bmatrix} x(1) \\ y(1) \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x(1) \\ y(1) \end{bmatrix} \right\rangle$$

(no intermediate penalization and  $F(x, u) = u^2$ ) so that  $M_{11} - M_{12}M_{22}^{-1}M_{12}^* = 0$  and the auxiliary cost is

$$\tilde{J}(x_0; u) = \int_0^1 u^2(s) \, ds.$$

Of course the minimum of  $\tilde{J}$  is zero for every initial condition  $x_0$ . The original cost is

$$\begin{aligned} J(x_0; u) &= \int_0^1 u^2(s) \, ds + x(1)^2 + u(1)^2 \\ &= \int_0^1 u^2(s) \, ds + \left[ x_0 + \int_0^1 u(s) \, ds \right]^2 + u(1)^2. \end{aligned}$$

If the infimum is zero, then we must find a sequence  $\{u_n\}$  in  $L^2(0, 1)$  such that

$$\|u_n\|_{L^2(0,1)} \rightarrow 0, \quad \int_0^1 u_n(s) \, ds \rightarrow -x_0.$$

Of course this is possible only if  $x_0 = 0$ . So the two functionals do not have the same infimum if  $x_0 \neq 0$ .

It is interesting to note that  $x_0 = 0$  is the unique optimizable initial condition for this problem.

We note that in this counterexample  $\tau_0 = T < T_0$ ; hence, the counterexample is valid even in the case that the optimal control is characterized by the “usual” two-point problem (plus the compatibility conditions at  $T$ ). In general the minimum point of the functional  $J$  is characterized by the “multipoint” problem in section 2.3. This problem is nonstandard, and we can interpret the results in this paper as a first step toward a variational investigation of this nonstandard multipoint problem. In this regard we note that *the condition  $\ker G^* = 0$  has not been used in the proofs of Theorems 5 and 9.*

The counterexample just presented shows that if  $\ker G^* \neq 0$ , then the functional  $\tilde{J}(x_0; u)$  is not the relaxed functional of  $J(x_0; u)$  (see [6] for the definition). Further insight into our problem is given by the observation that if, instead,  $\ker G^* = 0$ , then the relaxed functional of  $J(x_0; u)$  is  $\tilde{J}(x_0; u)$ . In order to see this, it is sufficient to note that the brace in (43) is nonnegative since the second addendum in that brace is

$$\min \left\{ \left\langle \begin{bmatrix} x(T) \\ v \end{bmatrix}, M \begin{bmatrix} x(T) \\ v \end{bmatrix} \right\rangle, \quad v \in U \right\}.$$

The minimum is realized by  $v = [-M_{22}^{-1}M_{12}^*x(T)]$ . Now let  $\hat{u}$  be any element in  $L^2(0, T; U)$  and let  $\hat{x} = x(T; \hat{u})$ . Using the fact that  $\ker G^* = 0$  we can construct a sequence  $\{u_n\}$  in  $\mathcal{U}$  which converges to  $\hat{u}$  in  $L^2(0, T; U)$  and such that  $Gu_n(T) \rightarrow [-M_{22}^{-1}M_{12}^*\hat{x}]$  so that when  $n \rightarrow +\infty$  the brace, with this particular sequence, converges to 0.

Finally, we prove that  $P(s)$ , the operator of the quadratic form  $I_s(x_0)$ , and any input  $u$  are related by the usual dissipation inequality.

**THEOREM 20.** *For every  $s > s_0$  and every  $u \in \mathcal{U}$ , we have*  
(45)

$$\langle x(s; s_0, x_0, u), P(s)x(s; s_0, x_0, u) \rangle - \langle x_0, P(s_0)x_0 \rangle + \int_{s_0}^s F(x(r; s_0, x_0, u), u(r)) \, dr \geq 0.$$

Let  $x_0$  be an initial condition which admits the optimal control  $u^+(\cdot; x_0)$ . If  $u^+(\cdot; x_0)$  is replaced with  $u$  in (45), then equality holds for every  $s_0$  and every  $s > s_0$ .

Conversely, let  $x_0$  be given and  $P(s)$  be a solution of the dissipation inequality, which satisfies condition

$$\langle x_0, P(T)x_0 \rangle = \inf_v \left\langle \begin{bmatrix} x_0 \\ Gv \end{bmatrix}, M \begin{bmatrix} x_0 \\ Gv \end{bmatrix} \right\rangle = \langle x_0, Nx_0 \rangle$$

(the operator  $N$  is the same one appearing in (31)). Let  $u(\cdot)$  be an input such that the dissipation inequality (with  $s_0 = 0$ ) holds as an equality along  $x(\cdot; x_0, u)$  and  $u(\cdot)$  and, furthermore, let  $u(T)$  satisfy

$$(46) \quad u(T) = \arg \min_{v \in U} \left\langle \begin{bmatrix} x(T; x_0, u) \\ Gv \end{bmatrix}, M \begin{bmatrix} x(T; x_0, u) \\ Gv \end{bmatrix} \right\rangle = \langle x(T; x_0, u), Nx(T; x_0, u) \rangle$$

(i.e., the second condition in (32)). Then,  $u(\cdot) = u^+(\cdot; x_0)$  is the optimal control of  $x_0$ .

This can be repeated for each initial time  $s_0$ .

*Proof.* In fact, it is known from the theory of the (standard) quadratic regulator problem that  $\mathcal{P}(s)$  solves the dissipation inequality and that  $\tilde{u}$  is characterized as the control which realizes equality in the dissipation inequality. Hence, the statement follows from Theorems 18 and 16.  $\square$

It is known that  $\mathcal{P}(s)$  is maximal among the selfadjoint nonnegative operator functions which solve the dissipation inequality. Hence, we have the following.

**THEOREM 21.** *Let  $P(s)$  be the operator of the quadratic form  $I_s$ . Then,  $P(s)$  is maximal among the nonnegative selfadjoint solutions of the dissipation inequality which satisfy the final condition (31).*

*Remark 22.* Following the suggestion of one referee, we derived Theorems 20 and 21 from the known properties of  $\mathcal{P}(s)$  in order to stress the role of the auxiliary cost  $\tilde{J}$ . In view of the counterexample presented in Remark 19, we note that it is possible to prove Theorems 20 and 21 from first principles, using arguments which resemble those in [3], thanks to the fact that if  $x_0$  is optimizable, then the restriction of  $u^+(t; x_0)$  to  $[s, T]$  is optimal for the initial condition at  $s$  given by  $x^+(s; x_0)$ . This proof holds even if  $\ker G^* \neq 0$ .

**5. Conclusions.** In this paper we have shown two applications which force us to study a quadratic control problem which penalizes the value taken by the control at the final time  $T$  and at an intermediate time  $\tau_0$  (of course similar arguments can be repeated if the values of the control at a finite number of intermediate instants are penalized). An obvious approach to this problem is to assume that  $u \in W^{1,2}(0, T)$  and to take as a “new control” the derivative  $u' \in L^2(0, T; U)$ . The examples we have presented, however, show that the regularity assumption on  $u$  is not natural, and that we must assume solely that  $u$  is square integrable.

The optimal control for the problem under study in general does not exist. We characterized the initial conditions which admit an optimal control (which is unique) and we characterized the optimal control in terms of a two-point problem and compatibility conditions. In the important case that the reference signal is zero, we expressed the optimal control in terms of a suitable Riccati equation.

The optimal control, however, in general does not exist. So, we studied the value function and we proved that it is a continuous function of the initial datum and the reference signal, and also a continuous function of the initial time, except at the intermediate points at which the control is penalized. Moreover, we gave a construction for a minimizing sequence and we proved that this construction is robust with respect to the reference signal and the time at which the value of the control is penalized.

**Acknowledgments.** We thank the referees for the careful reading of the paper, which led to an improved presentation and, in particular, to a more precise analysis of the role of the auxiliary cost function.

## REFERENCES

- [1] C. T. H. BAKER AND E. I. PARMUZIN, *Analysis via integral equations of an identification problem for delay differential equations*, J. Integral Equations Appl., 16 (2004), pp. 111–135.
- [2] A. BENSOUSSAN, G. DA PRATO, M. DELFOUR, AND S. MITTER, *Representation and Control of Infinite Dimensional Systems*, Birkhäuser Boston, Boston, MA, 1992.
- [3] F. BUCCI AND L. PANDOLFI, *The value function of the singular quadratic regulator problem with distributed control action*, SIAM J. Control Optim., 36 (1998), pp. 115–136.
- [4] D. J. CLEMENTS AND B. D. O. ANDERSON, *Singular optimal control: The linear-quadratic problem*, Lecture Notes in Control and Inform. Sci. 5, Springer-Verlag, Berlin, New York, 1978.
- [5] R. F. CURTAIN AND H. ZWART, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.
- [6] G. DAL MASO, *Semicontinuit  e rilassamento*, in Equazioni Differenziali e Calcolo delle Variazioni, G. Buttazzo, A. Marino, and M. K. V. Murthy, eds., Quaderni dell’Unione Matematica Italiana 39, Pitagora editrice, Bologna, 1995.
- [7] A. FAVINI AND A. YAGI, *Multivalued linear operators and degenerate evolution equations*, Ann. Mat. Pura Appl. (4), 163 (1993), pp. 353–380.
- [8] A. FAVINI AND A. YAGI, *Degenerate Differential Equations in Banach Spaces*, Marcel Dekker, New York, 1999.
- [9] W. GREUB, *Linear Algebra*, 3rd ed., Grundlehren Math. Wiss., Band 97, Springer-Verlag, Berlin, 1967.
- [10] H. R. JOSHI AND S. LENHART, *Solving a parabolic identification problem by optimal control methods*, Houston J. Math., 30 (2004), pp. 1219–1242.
- [11] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, Berlin, 1980.
- [12] I. LASIECKA AND R. TRIGGIANI, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories. I. Abstract Parabolic Systems*, Encyclopedia Math. Appl. 74, Cambridge University Press, Cambridge, UK, 2000.
- [13] E. SONTAG, *Mathematical Control Theory*, Springer-Verlag, New York, 1998.

## GLOBAL OPTIMIZATION OF LINEAR HYBRID SYSTEMS WITH VARYING TRANSITION TIMES\*

C. K. LEE<sup>†</sup> AND P. I. BARTON<sup>†</sup>

**Abstract.** Open loop optimal control problems with linear hybrid (discrete/continuous) systems embedded are often approximated as dynamic optimization problems. We propose a deterministic global optimization algorithm for linear hybrid systems with varying transition times. First, the control parametrization enhancing transform is used to transform the problem from a linear hybrid system with scaled discontinuities and varying transition times into a nonlinear one with stationary discontinuities and fixed transition times. Next, a theory is developed for constructing convex relaxations of arbitrary Bolza-type functionals subject to the transformed hybrid system. Finally, the convex relaxations are utilized in a branch-and-bound framework to obtain the solution to  $\epsilon$  global optimality within a finite number of iterations.

**Key words.** multistage dynamic optimization, nonconvex dynamic optimization, hybrid optimal control

**AMS subject classifications.** 34A30, 90C26

**DOI.** 10.1137/050625539

**1. Introduction.** Hybrid systems exhibit both discrete state and continuous state dynamics and have become indispensable for modeling systems exhibiting discontinuities in their dynamics [4]. This article focuses on the global solution of a specific class of dynamic optimization problems with linear time varying (LTV) hybrid systems embedded: problems in which the temporal sequence of modes is fixed, but the transition times between modes are allowed to vary. This is motivated by the fact that many practical problems can be expressed as open loop optimal control problems with hybrid systems embedded, which in turn can be approximated by dynamic optimization problems with hybrid systems embedded [4]. The latter transformation is carried out via control parametrization [29, 35], a partial discretization method where the controls are approximated by a finite series of piecewise continuous basis functions over the time horizon. The numerical solution of the resulting parameter optimization problem can then be obtained to local optimality by iterating finitely between the following subproblems:

1. An initial value (IVP) subproblem in which the hybrid system model is simulated for given values of the real valued variables parameterizing the controls using robust hybrid simulation technology, e.g., DAEPACK [37].
2. A nonlinear programming (NLP) Master problem that searches in the Euclidean parameter space using function and gradient information furnished by the IVP subproblem. Smoothness of the objective and constraint functionals are crucial for the application of existing gradient based algorithms, e.g., SNOPT [16].

Clearly, the practicality of such a method hinges on the existence and uniqueness of the parametric sensitivities of the hybrid system (or the related adjoints), which are

---

\*Received by the editors February 28, 2005; accepted for publication (in revised form) October 30, 2007; published electronically February 15, 2008. This material is based upon work supported by the National Science Foundation under grant CCR-0208956.

<http://www.siam.org/journals/sicon/47-2/62553.html>

<sup>†</sup>Process Systems Engineering Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 (chakun@alum.mit.edu, pib@mit.edu).

employed to calculate the gradients of the objective and constraint functionals used by the Master problem [10, 38]. Sufficient conditions for the existence and uniqueness of these sensitivities have been developed in [14], and these results indicate that the sensitivity trajectories of a hybrid system will usually exist a.e. in the parameter space. Subject to the key restriction that the temporal sequence of modes visited by executions of the hybrid system is unchanged throughout the parameter space (the timing of switches and jumps may still vary), the resulting Master NLP is smooth under mild assumptions and existing gradient based methods may be used to find local solutions [13, 3]. On the other hand, if the temporal sequence of modes varies as a function of the optimization parameters, then most resulting Master NLPs will exhibit some degree of nonsmoothness [13, 3]. These critical observations explicate the requirement of fixing the sequence of modes, a key issue. The restriction to a fixed sequence of modes reduces the embedded hybrid system to a *multistage* dynamic system. The extension of classical control parametrization of single-stage dynamic systems [29, 35] to multistage systems has been previously studied in [25, 39], where gradient based solvers were used to obtain local solutions of the NLP Master problem.

More recent research in the hybrid systems community has focused on optimal control problems where the switching times of an embedded hybrid system with a fixed sequence of modes are allowed to vary. In [41], the timings of the transitions are parameterized, and the gradients to the local NLP solver are obtained by solving the Hamilton–Jacobi–Bellman (HJB) equations using a dynamic programming approach. This problem (of determining the optimal switching times) constitutes the Stage 1 subproblem of a two stage optimization algorithm described in [42]. One of the methods presented in [42] involves obtaining the gradients of the participating functionals through formulating the costate equations, and this is further expanded upon by a different group of authors in [8], in which they derive the gradient of the cost functional for an especially simple form (special structure on the costate equations; the problem considered has no controls and has only the switching times as variables). However, this body of research has only focused on obtaining local solutions to the optimization problem (with no guarantees to the global optimality of the transition times). In this respect this body of research proves to be very similar in vein to the aforementioned work on multistage dynamic optimization.

In the context of global optimization, a deterministic branch-and-bound framework has been developed for the global optimization of linear hybrid systems with fixed transition times and a fixed sequence of modes [19]. This work has been extended to determine the optimal mode sequence using a mixed-integer optimization approach [5]. The final, missing piece of the puzzle for linear hybrid systems is to obtain the globally optimal switching times for a fixed sequence of modes. This problem is difficult because it is inherently nonconvex (see, e.g., [5]).

Many current deterministic approaches for global optimization rely on the ability to construct convex relaxations of the participating functionals, e.g., branch-and-bound algorithms [17, 28] and decomposition algorithms (outer approximation and generalized Benders decomposition) [12, 18, 15]. For static problems in mathematical programming, these functionals are provided by the convex envelope, McCormick’s composition technique [23], or  $\alpha$ BB by Adjiman, et al. [1]. The extension to systems for optimal control problems is nontrivial and has only been recently developed for both linear and nonlinear ODE systems [32, 34]. In addition, the extension to linear hybrid systems has also only recently been proposed [19, 5]. Unfortunately, these dynamic extensions can handle only problems for which the switching times are fixed in the optimal control problem. Currently, no satisfactory method exists for

constructing convex relaxations when the switching times are allowed to vary.

Bearing the aforementioned issues in mind, it is thus very desirable to have a method for transforming a problem with variable switching times into one with fixed switching times, which can then be solved using standard control parametrization algorithms. The control parametrization enhancing transform (CPET) [20] is a natural transform to use for this purpose. Unfortunately, this transformation comes with an associated difficulty: the right-hand sides of the differential equations become multiplied by the enhancing control. Thus, the resulting dynamic system no longer has the special structure exploited by methods specific to linear systems. While this is not significant for local optimization, it poses a considerable obstacle for global optimization because global optimization of nonlinear dynamic systems is much more challenging than that for linear systems. In this article, we will present an extension of the methods proposed in [34, 19] to solve the transformed problem, establishing a relaxation theory for the global optimization of general, nonlinear hybrid systems with a fixed sequence of modes and fixed transition times, i.e., nonlinear multistage systems with fixed switching times.

The rest of this article is organized as follows. Section 2 introduces the hybrid system modeling framework. Section 3 presents the formulation of the problem. It also includes a discussion on nonsmoothness of the problem when the objective function or constraints are evaluated at fixed points in time, which has profound implications on the type of discontinuities permitted in the dynamics of each mode. The transformed problem is constructed in section 4, while the theory for solving the transformed problem is presented in section 5. Finally, section 6 contains some examples illustrating the proposed approach, and section 7 concludes the article.

**2. Hybrid systems: Notation.** The modeling framework of [4] is used as a basis to define the following hybrid system of interest.

DEFINITION 2.1. *The hybrid system considered is the 10-tuple  $\mathcal{H} = (\mathcal{M}, \mathcal{E}, T_\mu, \sigma_1, \delta, \mathbf{p}, \mathbf{x}, \mathcal{F}, \mathcal{T}^0, \mathcal{T})$ , where*

- $\mathcal{M} = \{1, \dots, n_m\}$ ,  $1 \leq n_m < +\infty$ ,
- $\mathcal{E} = \{1, \dots, n_e\}$ ,  $1 \leq n_e < +\infty$ ,
- $T_\mu = \{m_i\}_{i \in \mathcal{E}}$ ,  $m_i \in \mathcal{M} \ \forall i \in \mathcal{E}$ ,
- $\sigma_1 \in \mathbb{R}$ ,
- $\delta \in \Delta \subset \mathbb{R}_+^{n_e}$ ,
- $\mathbf{p} \in P \subset \mathbb{R}^{n_p}$ ,
- $\mathbf{x} : \mathcal{E} \times P \times \Delta \times \mathbb{R} \rightarrow \mathbb{R}^{n_x}$ ,
- $\mathcal{F} : \mathcal{M} \times \mathbb{R}^{n_x} \times P \times \Delta \times \mathbb{R} \rightarrow \mathbb{R}^{n_x}$ ,
- $\mathcal{T}^0 : P \times \Delta \rightarrow \mathbb{R}^{n_x}$ , and
- $\mathcal{T} : \mathcal{E} \setminus \{n_e\} \times \mathbb{R}^{n_x} \times P \times \Delta \rightarrow \mathbb{R}^{n_x}$ .

The elements of  $\mathcal{M}$  are called the *modes* of  $\mathcal{H}$ .  $T_\mu$  is called the *hybrid mode trajectory* and is the discrete state variable of  $\mathcal{H}$ .  $\sigma_1$  is the initial time.  $\delta$  is the vector of nonnegative durations, and  $\mathbf{p}$  is the vector of parameters.  $\mathbf{x}$  is the vector of continuous state variables, and  $\mathcal{F}$  is the vector field for  $\mathbf{x}$ .  $\mathcal{T}^0$  are the initial conditions, and  $\mathcal{T}$  are the *transition functions*.  $\mathcal{E}$  is the index set for the *epochs*, which are defined in the following definition.

DEFINITION 2.2. *The hybrid time trajectory of  $\mathcal{H}$  is a finite sequence of intervals  $T_\tau = \{I_i\}_{i \in \mathcal{E}}$ , where  $I_i = [\sigma_i, \tau_i]$ ,  $\tau_i = \sigma_i + \delta_i$  for  $i \in \mathcal{E}$  and  $\sigma_i = \tau_{i-1}$  for  $i = 2, \dots, n_e$ . The  $I_i$  are called *epochs*.*

DEFINITION 2.3. *Consider the epoch  $I_i = [\sigma_i, \tau_i]$  and its corresponding scaled time interval  $\hat{I}_i = [\hat{\sigma}_i, \hat{\tau}_i] = [i - 1, i]$ . A scaled simple discontinuity occurring at time*

$t \in I_i$  is one that occurs at a fixed (stationary) point  $s \in \hat{I}_i$  such that

$$\frac{s - \hat{\sigma}_i}{\hat{\tau}_i - \hat{\sigma}_i} = s - i + 1 = \frac{t - \sigma_i}{\tau_i - \sigma_i}.$$

It is clear from Definition 2.3 that there is a stationary simple discontinuity [32, Definition 2.1] at  $s^*$  in  $\hat{I}_i$  iff there is a scaled simple discontinuity at  $t^*$  in  $I_i$ .

We will impose the following assumptions to make the optimization problem (introduced later) well posed:

A1.  $\Delta = [\delta^L, \delta^U]$  and  $P = [\mathbf{p}^L, \mathbf{p}^U]$  are nondegenerate interval vectors. The vector field  $\mathcal{F}(m, \cdot)$  for each  $m \in \mathcal{M}$  is affine in the continuous state variables  $\mathbf{x}$  and the parameters  $\mathbf{p}$  so that for each epoch  $I_{i \in \mathcal{E}}$  the continuous state variables evolve according to the LTV ODE system

$$\dot{\mathbf{x}}(i, \mathbf{p}, \delta, t) \equiv \left. \frac{d\mathbf{x}}{dt} \right|_{i, \mathbf{p}, \delta, t} = \mathbf{A}(m, \delta, t)\mathbf{x}(i, \mathbf{p}, \delta, t) + \mathbf{B}(m, \delta, t)\mathbf{p} + \mathbf{q}(m, \delta, t) \quad \forall t \in (\sigma_i, \tau_i].$$

Moreover, for all  $(m, i, \delta) \in \mathcal{M} \times \mathcal{E} \times \Delta$ ,  $\mathbf{A}(m, \delta, \cdot)$ ,  $\mathbf{B}(m, \delta, \cdot)$ , and  $\mathbf{q}(m, \delta, \cdot)$  are piecewise continuous on epoch  $I_i$  with a finite number of scaled simple discontinuities and are defined at any point of discontinuity.

A2. The initial conditions  $\mathcal{T}^0$  are affine functions so that the initial conditions for epoch  $I_1$  are given by

$$\mathbf{x}(1, \mathbf{p}, \delta, \sigma_1) = \mathbf{E}(0)\mathbf{p} + \mathbf{J}(0)\delta + \mathbf{k}(0).$$

A3. The transition functions  $\mathcal{T}(i, \cdot)$  for each  $i \in \mathcal{E} \setminus \{n_e\}$  are affine functions so that the initial conditions for epochs  $I_{i \in \mathcal{E} \setminus \{1\}}$  are given by

$$(2.1) \quad \mathbf{x}(i, \mathbf{p}, \delta, \sigma_i) = \mathbf{D}(i-1)\mathbf{x}(i-1, \mathbf{p}, \delta, \tau_{i-1}) + \mathbf{E}(i-1)\mathbf{p} + \mathbf{J}(i-1)\delta + \mathbf{k}(i-1).$$

DEFINITION 2.4. *Given values for  $\mathbf{p}$  and  $\delta$ , the solution, or execution, of a hybrid system  $\mathcal{H}$  subject to assumptions A1–A3 is  $\mathbf{x}(i, \mathbf{p}, \delta, t)$ ,  $t \in I_i$ ,  $i \in \mathcal{E}$ , where  $\mathbf{x}(i, \mathbf{p}, \delta, t)$  is the solution of the ODE system in A1 with initial conditions A2 if  $i = 1$ , and A3 otherwise.*

We shall now describe an execution of the hybrid system in time. The finite time horizon is partitioned into contiguous intervals called epochs. Starting from the initial conditions given by  $\mathcal{T}^0$ , the continuous state variables  $\mathbf{x}(1, \mathbf{p}, \delta, \cdot)$  evolve in time,  $t$ , according to the differential equations defined by the vector field  $\mathcal{F}(m_1, \cdot)$  for a (possibly trivially zero) duration of  $\delta_1$ . At time  $\tau_1$ , a *transition* is made from mode  $m_1$  to mode  $m_2$ . The transition functions in (2.1) map the value of the continuous state at  $\tau_1$  in epoch  $I_1$  to an initial condition for epoch  $I_2$  at time  $\sigma_2$ . The hybrid system then evolves according to the differential equations defined by the vector field  $\mathcal{F}(m_2, \cdot)$  for a duration of  $\delta_2$ , and so on and so forth. Note that for epoch  $I_i$ , the system evolves continuously in time if  $\delta_i > 0$ , and it evolves discretely by making an instantaneous transition if  $\delta_i = 0$ . We would like to emphasize that throughout this article, the hybrid mode trajectory  $T_\mu$  is known a priori and is fixed in the sense that it is not a decision variable in the optimization problem.

THEOREM 2.5. *A solution  $\mathbf{x}(i, \mathbf{p}, \delta, t)$ ,  $t \in I_i$ ,  $i \in \mathcal{E}$  to the hybrid system defined by Definition 2.1 and assumptions A1–A3 exists and is unique for each  $(\mathbf{p}, \delta) \in P \times \Delta$ .*

*Proof.* Consider any arbitrary  $(\mathbf{p}^*, \delta^*) \in P \times \Delta$  and the first epoch  $I_1$ . Since  $(\mathbf{p}^*, \delta^*)$  is fixed, the form of the LTV ODE system in the first epoch satisfies the



nonhomogenous linear system in [6, Chapter 3, p. 74]. Hence, there exists a unique solution of the hybrid system in the first epoch. At the transition to the second epoch, the initial conditions for the second epoch are clearly bounded by (2.1). Thus, the form of the LTV ODE system in the second epoch satisfies the nonhomogenous linear system in [6, Chapter 3, p. 74]. Therefore, a unique solution of the hybrid system in the second epoch exists. By induction, a unique solution of the hybrid system exists for all epochs  $I_{i \in \mathcal{E}}$ . Since  $(\mathbf{p}^*, \delta^*)$  was arbitrary, we obtain the desired result.  $\square$

*Remark 2.6.* It is possible to transcribe a problem in which  $\sigma_1$  is a decision variable bounded by the interval  $[\sigma_1^L, \sigma_1^U]$  into one where the initial time is fixed by prepending an additional mode to the hybrid system:

1. Introduce a new mode  $m_0$ .
2. Increase the number of durations (and epochs) by one,  $\delta^\dagger = (\delta_0, \delta)$ , where  $\delta_0 \in [0, \sigma_1^U - \sigma_1^L]$ .
3. The new mode trajectory becomes  $T_\mu^\dagger = m_0, T_\mu$ .
4. The dynamics of the initial mode  $m_0$  is given by  $\dot{\mathbf{x}}(t) = \mathbf{0}$  with initial conditions  $\mathbf{x}(0, \mathbf{p}, \delta^\dagger, \sigma_1^L) = \mathbf{E}(0)\mathbf{p} + \mathbf{J}(0)\delta + \mathbf{k}(0)$ .
5. The transition from mode  $m_0$  to  $m_1$  occurs at  $\sigma_1^L + \delta_0$  with state continuity as the transition function,  $\mathbf{x}(1, \mathbf{p}, \delta^\dagger, \sigma_1) = \mathbf{x}(0, \mathbf{p}, \delta^\dagger, \sigma_1^L + \delta_0)$  (*state continuity* means that the state variables are continuous at the transition, i.e., the transition function is the identity mapping).

*Remark 2.7.* In general, it is very difficult to characterize the exact image of  $P \times \Delta$  in the solution of the hybrid system (the *implied state bounds* in [32, 19]) when the transition times are varying (i.e., the durations of the epochs are decision variables in the optimization problem), thus we will work with relaxations of the image set.

**DEFINITION 2.8** (implied state bounds). *Define the following convex sets for all  $i \in \mathcal{E}$ , where  $\mathcal{S}_i \equiv [\sigma_1 + \sum_{j=1}^i \delta_j^L, \sigma_1 + \sum_{j=1}^i \delta_j^U]$ . For any fixed  $\underline{t} \in \mathcal{S}_i$ ,*

$$X(i, \underline{t}; P, \Delta) \equiv [\mathbf{x}^L(\underline{t}), \mathbf{x}^U(\underline{t})] \mid \mathbf{x}^L(\underline{t}) \leq \mathbf{x}(i, \mathbf{p}, \delta, \underline{t}) \leq \mathbf{x}^U(\underline{t}) \quad \forall (\mathbf{p}, \delta) \in P \times \Delta.$$

In addition,  $X(i, P, \Delta) \equiv [\mathbf{x}^L, \mathbf{x}^U] \mid X(i, \underline{t}; P, \Delta) \subset [\mathbf{x}^L, \mathbf{x}^U] \quad \forall \underline{t} \in \mathcal{S}_i$ .

**3. Problem formulation.** Consider the following problem.

**PROBLEM 3.1.**

$$\min_{\mathbf{p} \in P, \delta \in \Delta} F(\mathbf{p}, \delta) = \sum_{i=1}^{n_e} \left\{ \sum_{j=1}^{n_{\phi i}} \phi_{ij}(\mathbf{x}(i, \mathbf{p}, \delta, \alpha_{ij}(\delta)), \mathbf{p}, \delta) + \int_{\sigma_i(\delta)}^{\tau_i(\delta)} f_i(\mathbf{x}, \mathbf{p}, \delta, t) dt \right\},$$

subject to the following point and isoperimetric constraints,

$$\mathbf{G}(\mathbf{p}, \delta) = \sum_{i=1}^{n_e} \left\{ \sum_{j=1}^{n_{\eta i}} \eta_{ij}(\mathbf{x}(i, \mathbf{p}, \delta, \beta_{ij}(\delta)), \mathbf{p}, \delta) + \int_{\sigma_i(\delta)}^{\tau_i(\delta)} \mathbf{g}_i(\mathbf{x}, \mathbf{p}, \delta, t) dt \right\} \leq \mathbf{0},$$

where  $\mathbf{x}(i, \mathbf{p}, \delta, t)$  is given by the solution of the embedded hybrid system in Definition 2.1 subject to assumptions A1–A3;  $f_i$  and  $\mathbf{g}_i$  are piecewise continuous mappings  $f_i : X(i, P, \Delta) \times P \times \Delta \times \mathcal{S}_i \rightarrow \mathbb{R}$  and  $\mathbf{g}_i : X(i, P, \Delta) \times P \times \Delta \times \mathcal{S}_i \rightarrow \mathbb{R}^{n_c}$  for all  $i \in \mathcal{E}$ , where only a finite number of scaled simple discontinuities are allowed;  $n_{\phi i}$  is an arbitrary number of scaled point objectives in epoch  $I_i$ ,  $\alpha_{ij}(\delta) \in I_i$  such that  $\alpha_{ij}(\delta) = \sigma_i + \delta_i(\hat{\alpha}_{ij} - i + 1)$  for some fixed  $\hat{\alpha}_{ij} \in \hat{I}_i$ , and  $\phi_{ij}$  is a continuous mapping  $\phi_{ij} : X(i, P, \Delta) \times P \times \Delta \rightarrow \mathbb{R}$  for all  $j = 1, \dots, n_{\phi i}$  and  $i \in \mathcal{E}$ ; and  $n_{\eta i}$  is an

arbitrary number of scaled point constraints in epoch  $I_i$ ,  $\beta_{ij} \in I_i$  such that  $\beta_{ij}(\delta) = \sigma_i + \delta_i(\hat{\beta}_{ij} - i + 1)$  for some fixed  $\hat{\beta}_{ij} \in \hat{I}_i$ , and  $\eta_{ij}$  is a continuous mapping  $\eta_{ij} : X(i, P, \Delta) \times P \times \Delta \rightarrow \mathbb{R}^{n_c}$  for all  $j = 1, \dots, n_{\eta_i}$  and  $i \in \mathcal{E}$ . Additionally, we require that the set  $G = \{(\mathbf{p}, \delta) \in P \times \Delta \mid \mathbf{G}(\mathbf{p}, \delta) \leq \mathbf{0}\}$  is nonempty.

*Remark 3.2.* Problem 3.1 is a finite dimensional optimization problem in the Euclidean space  $P \times \Delta$ . Under the control parametrization framework (see, e.g., [29, 35]) the function and gradient information for the objective and constraint functionals are furnished by the solution of the embedded hybrid system as an IVP. More specifically, it is easy to show that Problem 3.1 encompasses open loop optimal control problems where the controls,  $\mathbf{u}$ , appear linearly in the dynamics of the system

$$\dot{\mathbf{x}}(i, \mathbf{p}, \delta, t) = \tilde{\mathbf{A}}(m, \delta, t)\mathbf{x}(\mathbf{p}, \delta, t) + \tilde{\mathbf{B}}(m, \delta, t)\mathbf{p} + \tilde{\mathbf{C}}(m, \delta, t)\mathbf{u}(\mathbf{p}, \delta, t) + \tilde{\mathbf{q}}(m, \delta, t)$$

for all  $i \in \mathcal{E}$ , where  $\tilde{\mathbf{A}}(m, \delta, \cdot)$ ,  $\tilde{\mathbf{B}}(m, \delta, \cdot)$ ,  $\tilde{\mathbf{C}}(m, \delta, \cdot)$ , and  $\tilde{\mathbf{q}}(m, \delta, \cdot)$  are piecewise continuous on epoch  $I_i$  with a finite number of scaled simple discontinuities and defined at any point of discontinuity for all  $(m, i) \in \mathcal{M} \times \mathcal{E}$ ; and the bounded real valued controls are parameterized linearly:

$$\begin{aligned} \mathbf{u}(\mathbf{p}, \delta, t) &= \tilde{\mathbf{S}}(m, \delta, t)\mathbf{p} + \tilde{\mathbf{v}}(m, \delta, t), \\ \mathbf{u}^L(t) \leq \mathbf{u}(\mathbf{p}, \delta, t) \leq \mathbf{u}^U(t) \quad \forall t \in \left[ \sigma_1, \sigma_1 + \sum_{j=1}^{n_e} \delta_j^U \right], \end{aligned}$$

where  $\mathbf{u}^L(t)$  and  $\mathbf{u}^U(t)$  are known bounds on the control variables  $\mathbf{u}(\mathbf{p}, \delta, t)$  that define the set  $U$ , and  $\tilde{\mathbf{S}}(m, \delta, \cdot)$ ,  $\tilde{\mathbf{v}}(m, \delta, \cdot)$  are piecewise continuous on epoch  $I_i$  with a finite number of scaled simple discontinuities and defined at any point of discontinuity for all  $(m, i) \in \mathcal{M} \times \mathcal{E}$ . Note that the latter parametrization of the control variables includes the cases where the controls are approximated by piecewise Lagrange polynomials, e.g., piecewise constant, linear, or quadratic approximations.

*Remark 3.3.* It is possible to cast the optimization decision variables as the transition times  $\tau$  instead of the epoch durations  $\delta$ . The equivalence between the two is established by the following:

$$\tau_i = \sigma_1 + \sum_{j=1}^i \delta_j \quad \forall i \in \mathcal{E}.$$

However, it is advantageous to work in terms of the epoch durations for the following reasons: (a) it is the natural formulation that facilitates the application of the CPET; and (b) the implicit constraints for feasible simulation trajectories using transition times,

$$\tau_{i-1} \leq \tau_i \quad \forall i = 2, \dots, n_e$$

need to be added explicitly to the Master NLP problem, whereas the same constraints with the duration formulation are handled by the simple bound constraints  $\delta^L \geq \mathbf{0}$ . This subtle difference is important in the control parametrization framework, as the decision variables passed to the IVP solver have to effect feasible simulations. For the majority of NLP solvers, this is handled much more robustly as simple bound constraints rather than as explicit constraints (i.e., simple bound constraints are satisfied throughout the solution process).

Comparing this problem formulation with that in [19], the main difference lies in the type of discontinuities allowed: the participating functionals and hybrid system now include scaled simple discontinuities rather than stationary simple discontinuities. The following examples illustrate how fixed-time point objectives, stationary simple discontinuities in the integrand of the objective function and stationary simple discontinuities in the dynamics of the hybrid system, can cause nonsmoothness in the problem. A discussion of sufficient conditions for the smoothness of the problem is deferred to the next section.

*Example 3.4* (fixed-time point objective). Consider the problem

$$\min_{\delta_1 \in [0.5, 1.5]} F(\delta_1) = x(1, \delta_1, 1) + x(2, \delta_1, 2),$$

subject to the hybrid system

$$\text{Mode 1 : } \dot{x}(i, \delta_1, t) = 0, \quad \text{Mode 2 : } \dot{x}(i, \delta_1, t) = 1$$

for all  $i \in \mathcal{E} = \{1, 2\}$ , with  $\sigma_1 = 0$ ,  $x(1, \delta_1, 0) = 0$ ,  $\delta_2 = 2 - \delta_1$ ,  $t \in [0, 2]$ ,  $T_\mu = 1, 2$ , and state continuity as the transition function  $x(2, \delta_1, \sigma_2) = x(1, \delta_1, \tau_1)$ . The nonsmooth objective function is given by

$$F(\delta_1) = \begin{cases} 3 - 2\delta_1 & \text{if } \delta_1 < 1, \\ 2 - \delta_1 & \text{if } \delta_1 \geq 1. \end{cases}$$

*Example 3.5* (integrand with stationary simple discontinuity). Consider the problem

$$\min_{\delta_1 \in [0.5, 1.5]} F(\delta_1) = \int_{\sigma_1}^{\tau_1(\delta_1)} f_1(x(1, \delta_1, t)) \, dt + \int_{\sigma_2(\delta_1)}^{\tau_2(\delta_1)} f_2(x(2, \delta_1, t)) \, dt,$$

where

$$f_1(x(1, \delta_1, t)) = \begin{cases} 0 & \text{if } t < 1, \\ x(1, \delta_1, t) + 1 & \text{if } t \geq 1, \end{cases}$$

$$f_2(x(2, \delta_1, t)) = \begin{cases} 0 & \text{if } t < 1, \\ x(2, \delta_1, t) + 1 & \text{if } t \geq 1, \end{cases}$$

subject to the hybrid system

$$\text{Mode 1 : } \dot{x}(i, \delta_1, t) = 0,$$

for all  $i \in \mathcal{E} = \{1, 2\}$ , with  $\sigma_1 = 0$ ,  $x(1, \delta_1, 0) = 0$ ,  $\delta_2 = 2 - \delta_1$ ,  $t \in [0, 2]$ ,  $T_\mu = 1, 1$ , and the transition function  $x(2, \delta_1, \sigma_2) = x(1, \delta_1, \tau_1) + 1$ . The nonsmooth objective function is given by

$$F(\delta_1) = \begin{cases} 2 & \text{if } \delta_1 < 1, \\ 4 - 2\delta_1 & \text{if } \delta_1 \geq 1. \end{cases}$$

*Example 3.6* (piecewise continuous dynamic system with stationary simple discontinuity). Consider the problem

$$\min_{\delta_1 \in [0.5, 1.5]} F(\delta_1) = x(2, \delta_1, 2),$$

subject to the hybrid system

$$\text{Mode 1 : } \dot{x}(i, \delta_1, t) = \begin{cases} 0 & \text{if } t < 1, \\ 1 & \text{if } t \geq 1, \end{cases} \quad \text{Mode 2 : } \dot{x}(i, \delta_1, t) = 0$$

for all  $i \in \mathcal{E} = \{1, 2\}$ , with  $\sigma_1 = 0$ ,  $x(1, \delta_1, 0) = 0$ ,  $\delta_2 = 2 - \delta_1$ ,  $t \in [0, 2]$ ,  $T_\mu = 1, 2$ , and state continuity as the transition function. The nonsmooth objective function is given by

$$F(\delta_1) = \begin{cases} 0 & \text{if } \delta_1 < 1, \\ \delta_1 - 1 & \text{if } \delta_1 \geq 1. \end{cases}$$

**4. The transform.** The CPET (see, e.g., [20, 36, 21] for details) is implemented as follows. Consider the original independent variable time ( $t$ ) in Problem 3.1. We now wish to construct a new time scale in which the varying epoch durations (transition times) are fixed,  $s \in [0, n_e]$ . The transformation (CPET) from  $t \in [\sigma_1, \sigma_1 + \sum_{i=1}^{n_e} \delta_i^U]$  to  $s \in [0, n_e]$  is defined by

$$(4.1) \quad \frac{dt}{ds} = v(\boldsymbol{\delta}, s), \quad t(\boldsymbol{\delta}, 0) = \sigma_1,$$

where the function  $v : \Delta \times [0, n_e] \rightarrow \mathbb{R}$  is called the *enhancing control*. It is a piecewise constant function with possible simple discontinuities at the prefixed knots  $1, \dots, n_e - 1$ ,

$$v(\boldsymbol{\delta}, s) = \sum_{i=1}^{n_e} \delta_i \chi_i(s),$$

where  $\chi_i(s)$  is the indicator function defined by

$$\chi_i(s) = \begin{cases} 1 & \text{if } s \in [i-1, i], \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$(4.2) \quad t(\boldsymbol{\delta}, s) = \sigma_1 + \int_0^s v(\boldsymbol{\delta}, z) dz = \sigma_1 + \delta_i(s - (i-1)) + \sum_{j=1}^{i-1} \delta_j = (s - i + 1)\delta_i + \sigma_i$$

for  $s \in [i-1, i]$ ,  $i \in \mathcal{E}$ , where the value of the enhancing control in the transformed time interval  $(i-1, i)$  corresponds to the value of the duration of epoch  $I_i$  in the original time scale. In addition, the scaled simple discontinuities, point objectives, and point constraints in Problem 3.1 become stationary simple discontinuities, point objectives, and point constraints in the new time scale, according to Definition 2.3. Finally, let  $\mathbf{x}' \equiv \frac{d\mathbf{x}}{ds}$ . It follows from the CPET that

$$\begin{aligned} \frac{\mathbf{x}'(i, \mathbf{p}, \boldsymbol{\delta}, t(\boldsymbol{\delta}, s))}{v(\boldsymbol{\delta}, s)} = & \left( \mathbf{A}(m, \boldsymbol{\delta}, t(\boldsymbol{\delta}, s)) \mathbf{x}(i, \mathbf{p}, \boldsymbol{\delta}, t(\boldsymbol{\delta}, s)) + \mathbf{B}(m, \boldsymbol{\delta}, t(\boldsymbol{\delta}, s)) \mathbf{p} \right. \\ & \left. + \mathbf{q}(m, \boldsymbol{\delta}, t(\boldsymbol{\delta}, s)) \right), \end{aligned}$$

where  $t$  is an additional differential state variable that has to satisfy (4.1). We can substitute for the explicit form of  $t(\boldsymbol{\delta}, s)$  to obtain

$$(4.3) \quad \hat{\mathbf{x}}'(i, \mathbf{p}, \boldsymbol{\delta}, s) = v(\boldsymbol{\delta}, s) \left( \hat{\mathbf{A}}(m, \boldsymbol{\delta}, s) \hat{\mathbf{x}}(\mathbf{p}, \boldsymbol{\delta}, s) + \hat{\mathbf{B}}(m, \boldsymbol{\delta}, s) \mathbf{p} + \hat{\mathbf{q}}(m, \boldsymbol{\delta}, s) \right),$$

where  $\hat{\mathbf{x}}(i, \mathbf{p}, \delta, s) \equiv \mathbf{x}(i, \mathbf{p}, \delta, t(\delta, s))$ ,  $\hat{\mathbf{x}}' \equiv \frac{d\hat{\mathbf{x}}}{ds}$ ,  $\hat{\mathbf{A}}(m, \delta, s) \equiv \mathbf{A}(m, \delta, t(\delta, s))$ ,  $\hat{\mathbf{B}}(m, \delta, s) \equiv \mathbf{B}(m, \delta, t(\delta, s))$ ,  $\hat{\mathbf{q}}(m, \delta, s) \equiv \mathbf{q}(m, \delta, t(\delta, s))$ , and  $t(\delta, s)$  is given by (4.2). Consider now any  $i \in \mathcal{E}$ . It is clear that  $v = \delta_i$  is continuous on  $\Delta \times (i-1, i)$  and defined at the points of discontinuity  $s = i-1$  and  $s = i$ . Also,  $t$  is continuous on  $\Delta \times [i-1, i]$  from (4.2). Let the epoch  $I_i$  be split into a finite number of contiguous intervals (*subepochs*) where  $\mathbf{A}(m, \delta, \cdot)$ ,  $\mathbf{B}(m, \delta, \cdot)$ , and  $\mathbf{q}(m, \delta, \cdot)$  are continuous internal to each subepoch. From Definition 2.3, the scaled simple discontinuities (in  $t$ ) for  $\mathbf{A}(m, \delta, \cdot)$ ,  $\mathbf{B}(m, \delta, \cdot)$ , and  $\mathbf{q}(m, \delta, \cdot)$  become stationary simple discontinuities (in  $s$ ) for  $\hat{\mathbf{A}}(m, \delta, \cdot)$ ,  $\hat{\mathbf{B}}(m, \delta, \cdot)$ , and  $\hat{\mathbf{q}}(m, \delta, \cdot)$ . Internal to any arbitrary, transformed subepoch in  $\hat{I}_i$ ,  $\hat{\mathbf{A}}(m, \delta, \cdot)$ ,  $\hat{\mathbf{B}}(m, \delta, \cdot)$ , and  $\hat{\mathbf{q}}(m, \delta, \cdot)$  are continuous (this follows from the fact that the composition of continuous functions is also continuous [27, Theorems 9.15 and 4.7]). The right-hand side of (4.3) is thus piecewise continuous in  $s$  with a finite number of stationary simple discontinuities, defined at each point of discontinuity. The objective function and constraints after the CPET are given by

$$(4.4) \quad \hat{F}(\mathbf{p}, \delta) = \sum_{i=1}^{n_e} \left\{ \sum_{j=1}^{n_{\phi i}} \phi_{ij}(\hat{\mathbf{x}}(i, \mathbf{p}, \delta, \hat{\alpha}_{ij}), \mathbf{p}, \delta) + \int_{i-1}^i f_i(\hat{\mathbf{x}}, \mathbf{p}, \delta, t(\delta, s)) v(\delta, s) ds \right\},$$

$$(4.5) \quad \hat{\mathbf{G}}(\mathbf{p}, \delta) = \sum_{i=1}^{n_e} \left\{ \sum_{j=1}^{n_{\eta i}} \eta_{ij}(\hat{\mathbf{x}}(i, \mathbf{p}, \delta, \hat{\beta}_{ij}), \mathbf{p}, \delta) + \int_{i-1}^i \mathbf{g}_i(\hat{\mathbf{x}}, \mathbf{p}, \delta, t(\delta, s)) v(\delta, s) ds \right\}.$$

Note that  $\hat{\alpha}_{ij}$  and  $\hat{\beta}_{ij}$  are no longer a function of  $\delta$ . Henceforth, we shall use the superscript prime notation to denote the transformed time derivative, i.e.,  $' \equiv \frac{d}{ds}$ . We are now able to formally state the transformed hybrid system and problem.

DEFINITION 4.1. A CPET hybrid system is the 8-tuple  $\hat{\mathcal{H}} = (\mathcal{M}, \mathcal{E}, T_\mu, \hat{\mathbf{p}}, \hat{\mathbf{x}}, \mathcal{F}, \mathcal{T}^0, \mathcal{T})$ , where  $\mathcal{M}$ ,  $\mathcal{E}$ , and  $T_\mu$  are as defined in Definition 2.1, and

- $\hat{\mathbf{p}} = (\mathbf{p}, \delta) \in \hat{P} = P \times \Delta \subset \mathbb{R}^{n_p + n_e}$ ,
- $\hat{\mathbf{x}} : \mathcal{E} \times \hat{P} \times \mathbb{R} \rightarrow \mathbb{R}^{n_x}$ ,
- $\mathcal{F} : \mathcal{M} \times \mathbb{R}^{n_x} \times \hat{P} \times \mathbb{R} \rightarrow \mathbb{R}^{n_x}$ ,
- $\mathcal{T}^0 : \hat{P} \rightarrow \mathbb{R}^{n_x}$ , and
- $\mathcal{T} : \mathcal{E} \setminus \{n_e\} \times \mathbb{R}^{n_x} \times \hat{P} \rightarrow \mathbb{R}^{n_x}$ .

As before, the elements of  $\mathcal{M}$  are called the modes of  $\hat{\mathcal{H}}$ .  $T_\mu$  is called the hybrid mode trajectory and is the discrete state variable of  $\hat{\mathcal{H}}$ .  $\hat{\mathbf{p}}$  is the vector of parameters.  $\hat{\mathbf{x}}$  is the vector of continuous state variables, and  $\mathcal{F}$  is the vector field for  $\mathbf{x}$ .  $\mathcal{T}^0$  are the initial conditions, and  $\mathcal{T}$  are the transition functions.  $\mathcal{E}$  remains the index set for the epochs, which are defined in the following definition.

DEFINITION 4.2. The hybrid time trajectory of  $\hat{\mathcal{H}}$  is a finite sequence of intervals  $T_\tau = \{\hat{I}_i\}_{i \in \mathcal{E}}$ , where  $\hat{I}_i = [\hat{\sigma}_i, \hat{\tau}_i] = [i-1, i]$ . The  $\hat{I}_i$  are called epochs.

From the previous analysis, the CPET transform of  $\mathcal{H}$  subject to assumptions A1–A3 will result in a CPET hybrid system  $\hat{\mathcal{H}}$  subject to the following assumptions.

B1.  $\hat{P} = [\hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U] = [(\mathbf{p}^L, \delta^L), (\mathbf{p}^U, \delta^U)]$  is a nondegenerate interval vector. The vector field  $\mathcal{F}(m, \cdot)$  for each  $m \in \mathcal{M}$  is nonlinear in the continuous state variables  $\hat{\mathbf{x}}$  and the parameters  $\hat{\mathbf{p}}$  so that for each epoch  $\hat{I}_{i \in \mathcal{E}}$  the continuous state variables evolve according to the nonlinear system in (4.3), written as

$$(4.6) \quad \hat{\mathbf{x}}'(i, \hat{\mathbf{p}}, s) = \mathcal{F}(m, \hat{\mathbf{x}}, \hat{\mathbf{p}}, s) \quad \forall s \in (i-1, i].$$

Moreover, for all  $(m, i, \hat{\mathbf{p}}, \hat{x}) \in \mathcal{M} \times \mathcal{E} \times \hat{P} \times \mathbb{R}^{n_x}$ ,  $\mathcal{F}(m, \hat{\mathbf{x}}, \hat{\mathbf{p}}, \cdot)$  is piecewise continuous on epoch  $\hat{I}_i$  with a finite number of stationary simple discontinuities and is defined at any point of discontinuity.

B2. The initial conditions  $\mathcal{T}^0$  are functions so that the initial conditions for epoch  $\hat{I}_1$  are given by

$$\hat{\mathbf{x}}(1, \hat{\mathbf{p}}, 0) = \mathbf{E}(0)\mathbf{p} + \mathbf{J}(0)\boldsymbol{\delta} + \mathbf{k}(0).$$

B3. The transition functions  $\mathcal{T}(i, \cdot)$  for each  $i \in \mathcal{E} \setminus \{n_e\}$  are functions so that the initial conditions for epochs  $\hat{I}_{i \in \mathcal{E} \setminus \{1\}}$  are given by

$$(4.7) \quad \hat{\mathbf{x}}(i, \hat{\mathbf{p}}, i-1) = \mathbf{D}(i-1)\hat{\mathbf{x}}(i-1, \hat{\mathbf{p}}, i-1) + \mathbf{E}(i-1)\mathbf{p} + \mathbf{J}(i-1)\boldsymbol{\delta} + \mathbf{k}(i-1).$$

DEFINITION 4.3. *Given a value for  $\hat{\mathbf{p}}$ , the solution, or execution, of a CPET hybrid system  $\hat{\mathcal{H}}$  subject to assumptions B1–B3 is  $\hat{\mathbf{x}}(i, \hat{\mathbf{p}}, s)$ ,  $s \in \hat{I}_i$ ,  $i \in \mathcal{E}$ , where  $\hat{\mathbf{x}}(i, \hat{\mathbf{p}}, s)$  is the solution of the ODE system in B1 with initial conditions B2 if  $i = 1$  and B3 otherwise.*

The major differences between a CPET hybrid system  $\hat{\mathcal{H}}$  subject to assumptions B1–B3 and a hybrid system  $\mathcal{H}$  subject to assumptions A1–A3 are (a) the initial time for  $\hat{\mathcal{H}}$  is fixed at  $s = 0$ , and the durations of all epochs are 1, and (b) the form of the underlying differential equations for each mode. The corresponding relaxations for the implied state bounds are given by the following definition.

DEFINITION 4.4 (implied state bounds). *Define the following convex sets for all  $i \in \mathcal{E}$ . For any fixed  $\underline{s} \in \hat{I}_i$ ,*

$$\hat{X}(i, \underline{s}; \hat{P}) \equiv [\hat{\mathbf{x}}^L(\underline{s}), \hat{\mathbf{x}}^U(\underline{s})] \mid \hat{\mathbf{x}}^L(\underline{s}) \leq \hat{\mathbf{x}}(\hat{\mathbf{p}}, \underline{s}) \leq \hat{\mathbf{x}}^U(\underline{s}) \quad \forall \hat{\mathbf{p}} \in \hat{P}.$$

In addition,

$$\hat{X}(i, \hat{P}) \equiv [\hat{\mathbf{x}}^L, \hat{\mathbf{x}}^U] \mid \hat{X}(i, \underline{s}; \hat{P}) \subset [\hat{\mathbf{x}}^L, \hat{\mathbf{x}}^U] \quad \forall \underline{s} \in \hat{I}_i.$$

PROBLEM 4.5. *The transformed problem is given by*

$$\begin{aligned} & \min_{\hat{\mathbf{p}} \in \hat{P}} \hat{F}(\hat{\mathbf{p}}) \\ & \text{s.t. } \hat{\mathbf{G}}(\hat{\mathbf{p}}) \leq \mathbf{0}, \end{aligned}$$

where  $\hat{\mathbf{x}}(i, \hat{\mathbf{p}}, s)$  is given by the solution of the embedded nonlinear hybrid system in Definition 4.1 subject to assumptions B1–B3;  $\hat{F}(\hat{\mathbf{p}})$  and  $\hat{\mathbf{G}}(\hat{\mathbf{p}})$  are given by (4.4) and (4.5), respectively;  $\hat{f}_i$  and  $\hat{\mathbf{g}}_i$  are piecewise continuous mappings  $\hat{f}_i : \hat{X}(i, \hat{P}) \times \hat{P} \times \hat{I}_i \rightarrow \mathbb{R}$ , and  $\hat{\mathbf{g}}_i : \hat{X}(i, \hat{P}) \times \hat{P} \times \hat{I}_i \rightarrow \mathbb{R}^{n_e}$  for all  $i \in \mathcal{E}$ , with a finite number of stationary simple discontinuities;  $n_{\phi i}$  is the number of fixed point objectives in epoch  $\hat{I}_i$ ,  $\hat{\alpha}_{ij} \in \hat{I}_i$ ,  $\hat{\phi}_{ij}$  is a continuous mapping  $\hat{\phi}_{ij} : \hat{X}(i, \hat{\alpha}_{ij}; \hat{P}) \times \hat{P} \rightarrow \mathbb{R}$  for all  $j = 1, \dots, n_{\phi i}$ , and  $i \in \mathcal{E}$ ; and  $n_{\eta i}$  is the number of fixed point constraints in epoch  $\hat{I}_i$ ,  $\hat{\beta}_{ij} \in \hat{I}_i$ ,  $\hat{\eta}_{ij}$  is a continuous mapping  $\hat{\eta}_{ij} : \hat{X}(i, \hat{\beta}_{ij}; \hat{P}) \times \hat{P} \rightarrow \mathbb{R}^{n_e}$  for all  $j = 1, \dots, n_{\eta i}$ , and  $i \in \mathcal{E}$ . Additionally, we require that the set  $\hat{G} = \{\hat{\mathbf{p}} \in \hat{P} \mid \hat{\mathbf{G}}(\hat{\mathbf{p}}) \leq \mathbf{0}\}$  be nonempty.

LEMMA 4.6. *Consider  $\mathcal{H}$  subject to A1–A3 and  $\hat{\mathcal{H}}$  subject to B1–B3. Then, for any  $(\hat{\mathbf{p}}, s) \in \hat{P} \times \hat{I}_i$ ,  $\hat{\mathbf{x}}(i, \hat{\mathbf{p}}, s) = \mathbf{x}(i, \mathbf{p}, \boldsymbol{\delta}, t(\boldsymbol{\delta}, s))$  for all  $i \in \mathcal{E}$ , where  $t(\boldsymbol{\delta}, s)$  is given by (4.2).*

*Proof.* Consider any arbitrary  $\hat{\mathbf{p}}^* = (\mathbf{p}^*, \boldsymbol{\delta}^*) \in \hat{P}$  and the first epoch in the transformed time scale,  $\hat{I}_1$ . From (4.2),  $t(\boldsymbol{\delta}^*, \hat{\sigma}_1) = \sigma_1$ . Hence, from the initial

conditions in assumptions A2 and B2,  $\hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, \hat{\sigma}_1) = \mathbf{x}(1, \mathbf{p}^*, \boldsymbol{\delta}^*, \sigma_1)$ . Integrating (4.3), we obtain

$$\int_0^s \hat{\mathbf{x}}' dz = \int_0^s \left( \hat{\mathbf{A}}(m_1, \boldsymbol{\delta}, z) \hat{\mathbf{x}}(1, \hat{\mathbf{p}}, z) + \hat{\mathbf{B}}(m_1, \boldsymbol{\delta}, z) \mathbf{p} + \hat{\mathbf{q}}(m_1, \boldsymbol{\delta}, z) \right) v(\boldsymbol{\delta}, z) dz,$$

and with the change of variables  $t(\boldsymbol{\delta}^*, s)$  given by (4.2),

$$\int_0^s \hat{\mathbf{x}}' dz = \int_{\sigma_1}^{t(\boldsymbol{\delta}^*, s)} \left( \mathbf{A}(m_1, \boldsymbol{\delta}, w) \mathbf{x}(1, \mathbf{p}, \boldsymbol{\delta}, z) + \mathbf{B}(m_1, \boldsymbol{\delta}, w) \mathbf{p} + \mathbf{q}(m_1, \boldsymbol{\delta}, w) \right) dw$$

or

$$\hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, s) - \hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, \hat{\sigma}_1) = \mathbf{x}(1, \mathbf{p}^*, \boldsymbol{\delta}^*, t(\boldsymbol{\delta}^*, s)) - \mathbf{x}(1, \mathbf{p}^*, \boldsymbol{\delta}^*, \sigma_1).$$

Therefore, for all  $s \in \hat{I}_1$ ,  $\hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, s) = \mathbf{x}(1, \mathbf{p}^*, \boldsymbol{\delta}^*, t(\boldsymbol{\delta}^*, s))$ . At the transition to epoch 2, from (4.2),  $t(\boldsymbol{\delta}^*, \hat{\tau}_1) = \tau_1$ , thus  $\hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, \hat{\tau}_1) = \mathbf{x}(1, \mathbf{p}^*, \boldsymbol{\delta}^*, \tau_1)$ . Applying the transition functions in assumptions A3 and B3, we obtain  $\hat{\mathbf{x}}(2, \hat{\mathbf{p}}^*, \hat{\sigma}_2) = \mathbf{x}(2, \mathbf{p}^*, \boldsymbol{\delta}^*, \sigma_2)$ . Since the choice of  $\mathbf{p}^*$  was arbitrary, induction on all epochs gives  $\hat{\mathbf{x}}(i, \hat{\mathbf{p}}, s) = \mathbf{x}(i, \mathbf{p}, \boldsymbol{\delta}, t(\boldsymbol{\delta}, s))$  for all  $i \in \mathcal{E}$ , for any  $(\hat{\mathbf{p}}, s) \in \hat{P} \times \hat{I}_i$ .  $\square$

*Remark 4.7.* A solution,  $\hat{\mathbf{x}}(i, \hat{\mathbf{p}}, s)$ ,  $s \in \hat{I}_{i \in \mathcal{E}}$  will exist and be unique for all  $\hat{\mathbf{p}} \in \hat{P}$ . This follows directly from Theorem 2.5 and Lemma 4.6.

**THEOREM 4.8.** *Problem 3.1 has the solution  $(\mathbf{p}^*, \boldsymbol{\delta}^*)$  iff  $(\mathbf{p}^*, \boldsymbol{\delta}^*)$  is a solution to Problem 4.5.*

*Proof.* Consider any arbitrary  $\hat{\mathbf{p}}^* = (\mathbf{p}^*, \boldsymbol{\delta}^*) \in \hat{P}$  and any arbitrary epoch  $i \in \mathcal{E}$ . For any  $j \in \{1, \dots, n_{\phi i}\}$ , from Lemma 4.6 and (4.2), we have

$$\phi_{ij}(\hat{\mathbf{x}}(i, \hat{\mathbf{p}}^*, \hat{\alpha}_{ij}), \hat{\mathbf{p}}^*) = \phi_{ij}(\mathbf{x}(i, \mathbf{p}^*, \boldsymbol{\delta}^*, \alpha_{ij}(\boldsymbol{\delta}^*)), \mathbf{p}^*, \boldsymbol{\delta}^*).$$

Similarly,

$$\int_{\hat{\sigma}_i}^{\hat{\tau}_i} f_i(\hat{\mathbf{x}}(i, \hat{\mathbf{p}}^*, s), \hat{\mathbf{p}}^*, t(\boldsymbol{\delta}^*, s)) v(\boldsymbol{\delta}^*, s) ds = \int_{\sigma_i(\boldsymbol{\delta}^*)}^{\tau_i(\boldsymbol{\delta}^*)} f_i(\mathbf{x}(i, \mathbf{p}^*, \boldsymbol{\delta}^*, t), \mathbf{p}^*, \boldsymbol{\delta}^*, t) dt.$$

Hence, we have  $\hat{F}(\hat{\mathbf{p}}^*) = F(\mathbf{p}^*, \boldsymbol{\delta}^*)$ . Applying the same analysis to the constraints, we have  $\hat{\mathbf{G}}(\hat{\mathbf{p}}^*) = \mathbf{G}(\mathbf{p}^*, \boldsymbol{\delta}^*)$ . Since  $\hat{\mathbf{p}}^*$  was arbitrary, we have shown the equivalence of Problems 3.1 and 4.5.  $\square$

*Remark 4.9.* A consequence of the CPET transform is the destruction of the linear structure of the original hybrid system.

The following theorem presents sufficient conditions for the objective function (and analogously the constraints) to be smooth, and will be assumed to hold for the transformed problem.

**THEOREM 4.10.** *Let  $\hat{P}^o \supset \hat{P}$ ,  $\hat{X}^o(i, \hat{\alpha}_{ij}) \supset \hat{X}(i, \hat{\alpha}_{ij}; \hat{P}^o)$  and  $\hat{X}^o(i) \supset \hat{X}(i, \hat{P}^o)$  be open subsets of  $\mathbb{R}^{n_p + n_e}$ ,  $\mathbb{R}^{n_x}$ , and  $\mathbb{R}^{n_x}$ , respectively, for all  $j = 1, \dots, n_{\phi i}$ ,  $i \in \mathcal{E}$ . If the following conditions are satisfied, then the objective function  $\hat{F}$  is continuously differentiable on  $\hat{P}^o$ .*

- C1.  $\frac{\partial \phi_{ij}}{\partial \mathbf{x}}$  and  $\frac{\partial \phi_{ij}}{\partial \mathbf{p}}$  exist and are continuous on  $\hat{X}^o(i, \hat{\alpha}_{ij}) \times \hat{P}^o$  for all  $j = 1, \dots, n_{\phi i}$ ,  $i \in \mathcal{E}$ .
- C2.  $\frac{\partial f_i}{\partial \mathbf{x}}$  and  $\frac{\partial f_i}{\partial \mathbf{p}}$  exist and are piecewise continuous on  $\hat{X}^o(i) \times \hat{P}^o \times \hat{I}_i$  for all  $i \in \mathcal{E}$ , where only a finite number of stationary simple discontinuities in  $s$  are allowed.

*Proof.* Consider an arbitrary epoch  $\hat{I}_i$ . First, we show that the parametric sensitivities exist and are unique. We have a finite number of stationary discontinuities (in  $s$ ) in (4.6). Let there be  $k$  such discontinuities in  $\hat{I}_i$  found at points  $s = \zeta_l$ ,  $l = 1, \dots, k$ . Construct a sequence of  $k + 1$  subepochs  $[\xi_1, \zeta_1], \dots, [\xi_{k+1}, \zeta_{k+1}]$ , where  $\xi_1 = \hat{\sigma}_i$ ,  $\zeta_{k+1} = \hat{\tau}_i$ , and  $\xi_{l+1} = \zeta_l$ ,  $l = 1, \dots, k$ . Extend the function  $\mathcal{F}(m_i, \cdot)$  to be continuous on all subepochs,

$$\mathcal{F}(m_i, \cdot, \xi_l) \equiv \lim_{s \rightarrow \xi_l^+} \mathcal{F}(m_i, \cdot, s), \quad \mathcal{F}(m_i, \cdot, \zeta_l) \equiv \lim_{s \rightarrow \zeta_l^-} \mathcal{F}(m_i, \cdot, s),$$

for  $l = 1, \dots, k + 1$ , and impose state continuity for each transition between subepochs. A CPET hybrid system is thus defined within the epoch  $\hat{I}_i$ . Now consider an arbitrary subepoch  $[\xi_l, \zeta_l]$ . From (4.3), it is clear that the partial derivatives  $\frac{\partial \mathcal{F}(m_i, \cdot)}{\partial \mathbf{x}}$  and  $\frac{\partial \mathcal{F}(m_i, \cdot)}{\partial \mathbf{p}}$  exist and are continuous internal to this subepoch. At the transition  $\zeta_l$ , it is easy to verify that the remaining assumptions of [14, Theorem 1] are satisfied, which provides the existence and uniqueness result. Since the discontinuities are stationary, the transition times in the interior of the epoch are independent of the parameters, and the sensitivities are continuous everywhere interior to the epoch.

Consider now an arbitrary  $v \in \{1, \dots, n_p + n_e\}$ . For the point objectives, since  $\hat{P}^o$  is an open set, the parametric sensitivities exist and condition C1 holds, the summation term  $\sum_{j=1}^{n_{\phi i}} \frac{\partial \phi_{ij}}{\partial p_v}$  exists and is continuous on  $\hat{P}^o$  by the chain rule and linearity of the derivative operator. Next, consider the integral objectives. We have a finite number of stationary discontinuities (in  $s$ ) in the integrand  $f_i$ ,  $\mathcal{F}(m_i, \cdot)$ , the parametric sensitivities and the derivatives in condition C2. Let there be  $\hat{k}$  such discontinuities found at points  $s = \hat{\zeta}_{\hat{l}}$ ,  $\hat{l} = 1, \dots, \hat{k}$ . Partition the integral into the following:

$$\hat{F}_i(\hat{\mathbf{p}}) = \sum_{\hat{l}=0}^{\hat{k}} \hat{F}_{i\hat{l}}(\hat{\mathbf{p}}) = \sum_{\hat{l}=0}^{\hat{k}} \int_{\zeta_{\hat{l}}}^{\zeta_{\hat{l}+1}} f_i(\hat{\mathbf{x}}, \hat{\mathbf{p}}, s) \, ds,$$

where  $\zeta_0 = \hat{\sigma}_i$  and  $\zeta_{\hat{k}+1} = \hat{\tau}_i$ . Now, consider an arbitrary  $\hat{l} \in \{1, \dots, \hat{k}\}$ . Extend  $f_i$ ,  $\frac{\partial f_i}{\partial \mathbf{x}}$ , and  $\frac{\partial f_i}{\partial \mathbf{p}}$  to be continuous on  $\hat{X}^o(i) \times \hat{P}^o \times [\zeta_{\hat{l}}, \zeta_{\hat{l}+1}]$ , and  $\frac{\partial \hat{\mathbf{x}}(i, \cdot)}{\partial \mathbf{p}}$  and  $\hat{\mathbf{x}}(i, \cdot)$  to be continuous on  $\hat{P}^o \times [\zeta_{\hat{l}}, \zeta_{\hat{l}+1}]$ . At most, these functions are discontinuous at their endpoints in time. Removing these discontinuities does not alter the value of the integral because the endpoints comprise a set of measure zero. As above, applying the chain rule on the partial derivative  $\frac{\partial f_i}{\partial p_v}$ , we obtain continuity of said derivative on  $\hat{P}^o \times [\zeta_{\hat{l}}, \zeta_{\hat{l}+1}]$ . These continuity conditions enable us to differentiate under the integral sign [7, p. 308] to obtain

$$\frac{\partial \hat{F}_{i\hat{l}}}{\partial p_v} = \int_{\zeta_{\hat{l}}}^{\zeta_{\hat{l}+1}} \frac{\partial f_i}{\partial p_v} \, ds.$$

We can then apply [32, Proposition 2.1] to yield  $\frac{\partial \hat{F}_{i\hat{l}}}{\partial p_v}$  continuous on  $\hat{P}^o$ . Since  $\hat{l}$  was arbitrary,  $\frac{\partial \hat{F}_i}{\partial p_v}$  is continuous on  $\hat{P}^o$  as the sum of continuous functions is continuous. Since  $i$  was arbitrary,  $\frac{\partial \hat{F}}{\partial p_v}$  is continuous on  $\hat{P}^o$ . Since  $v$  was arbitrary, it follows that  $\hat{F}$  is continuously differentiable on  $\hat{P}^o$ .  $\square$



**5. Relaxation theory.** In this section, we will present the theory required for constructing convex relaxations of Problem 4.5. This theory is an extension of that developed in [30, Chapter 6] and [33] for (single-stage) nonlinear dynamic systems. The main idea behind the theorems presented below will consist of breaking down the multistage hybrid system into contiguous intervals in time, verifying that the hypotheses of the theorems in [30, Chapter 6] and [33] hold for each of these intervals, and applying the theorems sequentially for each interval via finite induction.

The ultimate goal of this section is condensed into constructing a convex relaxation for the objective function (4.4), subject to the transformed nonlinear hybrid system. The exact same theory is applied for the point and isoperimetric constraints in (4.5). The ability to construct convex relaxations for the objective function and constraints then enables a convex relaxation of the problem to be solved. Finally, it is shown that the constructed convex relaxations possess the same consistent bounding properties of the convex relaxation techniques used in their construction, so that their incorporation into a branch-and-bound framework [9, 23] leads to an infinitely convergent algorithm [17]. This implies  $\varepsilon$  global optimality within a finite number of iterations.

The steps for constructing the convex relaxation are outlined below:

1. Estimating the implied state bounds,  $\hat{X}(i, \underline{s}; \hat{P})$  in Definition 4.4.
2. Constructing convex and concave relaxations for the states.
3. Applying convex relaxation techniques on subsets of Euclidean spaces to construct the required convex relaxation.

**DEFINITION 5.1.** Let  $\hat{\mathbf{x}}(i, \hat{\mathbf{p}}, s)$  be the solution of  $\hat{\mathcal{H}}$  subject to assumptions B1–B3, and let  $\hat{x}_j(i, \hat{\mathbf{p}}, s) \in \hat{X}_j(i, \hat{\mathbf{p}}, s)$  for each  $\hat{\mathbf{p}} \in \hat{P}$ ,  $i \in \mathcal{E}$ ,  $j = 1, \dots, n_x$ , where  $\hat{X}_j(i, \hat{\mathbf{p}}, s) \subset \mathbb{R}$  is a bounding set that is known independently. For each fixed  $\underline{s} \in \hat{I}_{i \in \mathcal{E}}$ , let  $\varrho_j(i, \mathbf{r}, \underline{s}) = \inf \hat{X}_j(i, \mathbf{r}, \underline{s})$  and  $\varsigma_j(i, \mathbf{r}, \underline{s}) = \sup \hat{X}_j(i, \mathbf{r}, \underline{s})$  for each  $\mathbf{r} \in \hat{P}$ ,  $j = 1, \dots, n_x$ . Furthermore, let  $\hat{X}(i, \underline{s})$  be defined pointwise in (transformed) time for each  $i \in \mathcal{E}$  by  $\hat{X}(i, \underline{s}) = [\mathbf{z}^L, \mathbf{z}^U]$  such that

$$z_j^L = \inf_{\mathbf{r} \in \hat{P}} \varrho_j(i, \mathbf{r}, \underline{s}), \quad z_j^U = \sup_{\mathbf{r} \in \hat{P}} \varsigma_j(i, \mathbf{r}, \underline{s}) \quad \forall j = 1, \dots, n_x,$$

where  $z_j^L$  and  $z_j^U$  are in the extended real number system.

**THEOREM 5.2.** Consider  $\hat{\mathcal{H}}$  subject to assumptions B1–B3. If the following conditions are satisfied for all  $i \in \mathcal{E}$  and  $j = 1, \dots, n_x$ ,

$$\text{D1. } v_j(\hat{\sigma}_i) < \min_{\mathbf{r} \in \hat{P}} \hat{x}_j(i, \mathbf{r}, \hat{\sigma}_i),$$

$$\text{D2. } w_j(\hat{\sigma}_i) > \max_{\mathbf{r} \in \hat{P}} \hat{x}_j(i, \mathbf{r}, \hat{\sigma}_i),$$

and, additionally, for all  $\mathbf{v}(s), \mathbf{w}(s) \in H(s)$ ,  $s \in [i-1, i]$ ,

$$\text{D3. } v'_j = \underline{h}_j(m_i, \mathbf{v}, \mathbf{w}, s; \hat{P}) < \inf_{\substack{\mathbf{z} \in \hat{X}(i, s) \cap H(s), \mathbf{r} \in \hat{P} \\ z_j = v_j(s)}} \mathcal{F}_j(m_i, \mathbf{z}, \mathbf{r}, s),$$

$$\text{D4. } w'_j = \bar{h}_j(m_i, \mathbf{v}, \mathbf{w}, s; \hat{P}) > \sup_{\substack{\mathbf{z} \in \hat{X}(i, s) \cap H(s), \mathbf{r} \in \hat{P} \\ z_j = w_j(s)}} \mathcal{F}_j(m_i, \mathbf{z}, \mathbf{r}, s),$$

where  $H(s) \equiv \{\mathbf{z} \mid \mathbf{v}(s) \leq \mathbf{z} \leq \mathbf{w}(s)\}$ ; then

$$\mathbf{v}(s) < \hat{\mathbf{x}}(i, \hat{\mathbf{p}}, s) < \mathbf{w}(s) \quad \forall (\hat{\mathbf{p}}, s) \in \hat{P} \times \hat{I}_i, \quad i \in \mathcal{E}.$$

It is also assumed that the solutions, in the sense of Carathéodory, to the differential systems in  $\mathbf{v}$  and  $\mathbf{w}$  exist and are unique for all  $i \in \mathcal{E}$ .

*Proof.* From assumption B2, Theorem 4.10 (treating  $\hat{\mathbf{x}}(i, \cdot, \hat{\tau}_i)$  as the objective function) and the form of (4.7),  $\hat{\mathbf{x}}(i, \cdot, \hat{\sigma}_i)$  is continuous on  $\hat{P}$  for all  $i \in \mathcal{E}$ . Hence, the extrema in conditions D1 and D2 exist.

Consider now the first epoch  $\hat{I}_1$ . From assumption B1,  $\mathcal{F}(m_1, \cdot)$  is piecewise continuous with a finite number of stationary simple discontinuities in  $s$ . Let  $\gamma$  be the number of discontinuities occurring at  $\tilde{\tau}_k \in \hat{I}_1$ ,  $k = 1, \dots, \gamma$ . Then, the first epoch can be further subdivided into  $\gamma + 1$  contiguous subepochs, for which we have transitions at the subepoch boundaries, state continuity at each transition, and  $\mathcal{F}(m_1, \cdot)$  is continuous for each subepoch. Let the sequence of subepochs be given by  $\{\tilde{I}_k\}$ ,  $k = 1, \dots, \gamma + 1$ , where  $\tilde{I}_k = [\tilde{\sigma}_k, \tilde{\tau}_k]$ ,  $\tilde{\sigma}_1 = 0$ ,  $\tilde{\tau}_{\gamma+1} = \hat{\tau}_1$ , and  $\tilde{\sigma}_{k+1} = \tilde{\tau}_k$  for  $k = 1, \dots, \gamma$ . Consider now the first subepoch  $\tilde{I}_1$ .

The initial condition at time  $s = 0$  given by assumption B2 is clearly continuous on  $\hat{P}$ . The form of the nonlinear ODE system in the first subepoch, and the conditions D1–D4 clearly satisfy the conditions of [33, Corollary 2.6], which gives

$$(5.1) \quad \mathbf{v}(s) < \hat{\mathbf{x}}(1, \hat{\mathbf{p}}, s) < \mathbf{w}(s)$$

for all  $(\hat{\mathbf{p}}, s) \in \hat{P} \times \tilde{I}_1$ . At the transition  $\tilde{\tau}_1$ , state continuity ensures

$$(5.2) \quad \mathbf{v}(\tilde{\sigma}_2) = \mathbf{v}(\tilde{\tau}_1) < \hat{\mathbf{x}}(1, \hat{\mathbf{p}}, \tilde{\sigma}_2) < \mathbf{w}(\tilde{\tau}_1) = \mathbf{w}(\tilde{\sigma}_2) \quad \forall \hat{\mathbf{p}} \in \hat{P}.$$

From Theorem 4.10,  $\hat{\mathbf{x}}(1, \cdot, \tilde{\sigma}_2)$  is continuous on  $\hat{P}$  (simply treat  $\hat{\mathbf{x}}(1, \cdot, \tilde{\sigma}_2)$  as the objective function). The form of the nonlinear ODE system in the second subepoch, (5.2), and the conditions D3 and D4 thus satisfy the conditions of [33, Corollary 2.6], which implies that (5.1) holds for all  $(\hat{\mathbf{p}}, s) \in \hat{P} \times \tilde{I}_2$ . By induction on all subepochs, (5.1) holds for all  $(\hat{\mathbf{p}}, s) \in \hat{P} \times \hat{I}_1$ . Consider now the second epoch  $\hat{I}_2$ . From Theorem 4.10 and (4.7),  $\hat{\mathbf{x}}(2, \cdot, \hat{\sigma}_2)$  is continuous on  $\hat{P}$ . The analysis carried out for the first epoch is thus valid for the second. By induction on all epochs, we have the desired result.  $\square$

By asserting uniqueness of the solution of the bounding differential equations, the conditions of the above theorem may be relaxed to

$$\text{D1. } v_j(\hat{\sigma}_i) \leq \min_{\mathbf{r} \in \hat{P}} \hat{x}_j(\mathbf{r}, \hat{\sigma}_i),$$

$$\text{D2. } w_j(\hat{\sigma}_i) \geq \max_{\mathbf{r} \in \hat{P}} \hat{x}_j(\mathbf{r}, \hat{\sigma}_i),$$

$$\text{D3. } v'_j = \underline{h}_j(m_i, \mathbf{v}, \mathbf{w}, s; \hat{P}) \leq \inf_{\substack{\mathbf{z} \in \hat{\mathcal{X}}(i, s) \cap H(s), \mathbf{r} \in \hat{P} \\ z_j = v_j(s)}} \mathcal{F}_j(m_i, \mathbf{z}, \mathbf{r}, s),$$

$$\text{D4. } w'_j = \bar{h}_j(m_i, \mathbf{v}, \mathbf{w}, s; \hat{P}) \geq \sup_{\substack{\mathbf{z} \in \hat{\mathcal{X}}(i, s) \cap H(s), \mathbf{r} \in \hat{P} \\ z_j = w_j(s)}} \mathcal{F}_j(m_i, \mathbf{z}, \mathbf{r}, s),$$

i.e., replacing the strict inequalities with regular inequalities (see [40, Remark 12.X]). Furthermore, by asserting regular inequalities, the result of the theorem also permits

$$\mathbf{v}(s) \leq \hat{\mathbf{x}}(i, \hat{\mathbf{p}}, s) \leq \mathbf{w}(s) \quad \forall (\hat{\mathbf{p}}, s) \in \hat{P} \times \hat{I}_i, \quad i \in \mathcal{E}.$$

For the remainder of this article, we will assume that the uniqueness of the constructed bounding differential equations hold for all  $\hat{\mathbf{p}} \in \hat{P}$ ,  $s \in [0, n_e]$ , and so it is understood that reference to Theorem 5.2 also refers to the regular inequalities just described.

*Remark 5.3.* The bounding set  $\hat{\mathcal{X}}(i, \hat{\mathbf{p}}, s)$  makes it possible to tighten the implied state bounds obtained when physical insight from the problem in the form of invariants (e.g., conservation laws) and bounds is available; see [30, 33] for examples.

Theorem 5.2 enables a hybrid system of bounding differential equations to be constructed to obtain the set:

$$(5.3) \quad \hat{X}(i, \underline{s}; \hat{P}) \equiv \{\mathbf{z} \mid \mathbf{v}(\underline{s}) \leq \mathbf{z} \leq \mathbf{w}(\underline{s})\}.$$

The most difficult aspect of applying the theorem lies in obtaining the extrema in conditions D1–D4. As stated in [33], while computing the exact solution to the optimization problem would yield the tightest bounds possible from the theorem, actually solving the optimization problems at each integration step in a numerical integration would be a prohibitively expensive task. Hence, in practice, the solutions to the optimization problems are estimated by interval arithmetic [24] pointwise in time. Before we proceed, we will briefly introduce the metric topology for the set of intervals (see [24] for more details). By an *interval* we mean a compact set of real numbers  $[x^L, x^U] = \{x \mid x^L \leq x \leq x^U\}$ . As in [24], we will not distinguish between the degenerate interval  $[a, a]$  and the real number  $a$ . Define the distance  $d(X, Y) = \max(|x^L - y^L|, |x^U - y^U|)$  for the intervals  $X \equiv [x^L, x^U]$ ,  $Y \equiv [y^L, y^U]$ . The absolute value of an interval  $X \equiv [x^L, x^U]$  is given by  $|X| = \max(|x^L|, |x^U|)$ . The vector norm  $\|Z\| = \max(|Z_1|, \dots, |Z_n|)$  is used for interval vectors. An interval valued function  $F: Z \rightarrow \mathbb{IR}$ ,  $Z \subset \mathbb{IR}^n$ , is said to be *continuous* in the usual  $\varepsilon - \delta$  fashion with the metric  $d(X, Y)$ , where  $\mathbb{IR}$  is the set of all intervals. We say that an interval valued function  $F$  of the interval variables  $X_1, \dots, X_n$  is *inclusion monotonic* if  $Y_i \subseteq X_i, i = 1, \dots, n$  implies  $F(Y_1, \dots, Y_n) \subseteq F(X_1, \dots, X_n)$ . Let  $f$  be a real valued function of  $n$  real variables  $x_1, \dots, x_n$ . By an *interval extension* of  $f$ , we mean an interval valued function  $F$  of  $n$  interval variables  $X_1, \dots, X_n$  with the property  $F(x_1, \dots, x_n) = f(x_1, \dots, x_n)$  for real arguments; i.e., an interval extension of  $f$  is an interval valued function which has real values when the arguments are all real (degenerate intervals) and coincides with  $f$ .

Consider now the vector  $\mathbf{z} \in \mathbb{R}^n$ . We will introduce the following notation: for any fixed  $j \in \{1, \dots, n\}$ , let  $\mathbf{z}_{k \neq j}$  denote the vector  $\tilde{\mathbf{z}} \in \mathbb{R}^{n-1}$ , where

$$\tilde{z}_k = \begin{cases} z_k & \text{if } k < j, \\ z_{k+1} & \text{if } k \geq j. \end{cases}$$

If  $\mathbf{z} \in Z = [\mathbf{z}^L, \mathbf{z}^U]$ , then the corresponding interval vector  $Z_{k \neq j} = [\mathbf{z}_{k \neq j}^L, \mathbf{z}_{k \neq j}^U]$ . For convenience, we will also introduce the following (elementwise) maximization and minimization operations: consider the  $n$ -dimensional vectors  $\mathbf{x}$  and  $\mathbf{y}$  whose elements are in the extended real number system. Let the vector valued operation  $\text{cmin}(\mathbf{x}, \mathbf{y})$  return the  $n$ -dimensional vector  $\mathbf{z}$  whose elements are in the extended real number system where

$$z_i = \min(x_i, y_i) \quad \forall i = 1, \dots, n.$$

Similarly, let  $\text{cmax}(\mathbf{x}, \mathbf{y})$  return the  $n$ -dimensional vector  $\mathbf{z}$  in the extended real number system where

$$z_i = \max(x_i, y_i) \quad \forall i = 1, \dots, n.$$

**COROLLARY 5.4.** Consider  $\hat{\mathcal{H}}$  subject to assumptions B1–B3. Define the following interval valued functions:

$$(5.4) \quad Y(\hat{\sigma}_1) = [\mathbf{y}^L(\hat{\sigma}_1), \mathbf{y}^U(\hat{\sigma}_1)] = \mathbf{E}(0)P + \mathbf{J}(0)\Delta + \mathbf{k}(0),$$

$$(5.5) \quad \begin{aligned} Y(\hat{\sigma}_{l+1}) &= [\mathbf{y}^L(\hat{\sigma}_{l+1}), \mathbf{y}^U(\hat{\sigma}_{l+1})] \\ &= \mathbf{D}(l)[\mathbf{v}(\hat{\tau}_l), \mathbf{w}(\hat{\tau}_l)] + \mathbf{E}(l)P + \mathbf{J}(l)\Delta + \mathbf{k}(l) \quad \forall l = 1, \dots, n_e - 1, \end{aligned}$$

and let  $\Gamma_j(m_i, v_j, Z(j, i, s), \hat{P}, s) = [\gamma_j^L(m_i), \gamma_j^U(m_i)]$  and  $\Lambda_j(m_i, w_j, Z(j, i, s), \hat{P}, s) = [\lambda_j^L(m_i), \lambda_j^U(m_i)]$  be inclusion monotonic interval extensions of  $\mathcal{F}_j(m_i, \hat{x}_j, \hat{\mathbf{x}}_{k \neq j}, \hat{\mathbf{p}}, s)$   $\forall i \in \mathcal{E}, j = 1, \dots, n_x$ , where

$$Z(j, i, s) = \{\mathbf{z}_{k \neq j} \mid \text{cmax}(\mathbf{v}_{k \neq j}(s), \boldsymbol{\varphi}_{k \neq j}(i, s)) \leq \mathbf{z}_{k \neq j} \leq \text{cmin}(\mathbf{w}_{k \neq j}(s), \boldsymbol{\psi}_{k \neq j}(i, s))\},$$

and  $\hat{X}(i, s; \hat{P}) = [\boldsymbol{\varphi}(i, s), \boldsymbol{\psi}(i, s)]$  is defined in (5.3) and obtained from Theorem 5.2. Then,  $\forall j = 1, \dots, n_x, s \in [i-1, i]$ , and  $i \in \mathcal{E}$ , the system of differential equations and initial conditions

$$(5.6) \quad v_j' = \gamma_j^L(m_i, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U, s), \quad v_j(\hat{\sigma}_i) = y_j^L(\hat{\sigma}_i),$$

$$(5.7) \quad w_j' = \lambda_j^U(m_i, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U, s), \quad w_j(\hat{\sigma}_i) = y_j^U(\hat{\sigma}_i),$$

bounds the transformed hybrid system,

$$\mathbf{v}(s) \leq \hat{\mathbf{x}}(i, \hat{\mathbf{p}}, s) \leq \mathbf{w}(s) \quad \forall (\hat{\mathbf{p}}, s) \in \hat{P} \times \hat{I}_i, i \in \mathcal{E}.$$

*Proof.* The rational interval functions (5.4) and (5.5) are inclusion monotonic [24, p. 21]. Together with the inclusion monotonicity of the interval extensions of  $\mathcal{F}_j(m_i, \cdot)$ , (5.6) and (5.7) thus satisfy conditions D1–D4 of Theorem 5.2.  $\square$

It is important to note that the implied state bounds are obtained given a particular parameter set  $\hat{P}$ ; in order to ensure convergence of the branch-and-bound framework, these bounds must converge as  $\hat{P}$  becomes degenerate in the limit.

LEMMA 5.5. Let  $H(X_1, \dots, X_n) = [h^L, h^U]$  be an interval valued function, where  $X_i = [\mathbf{x}_i^L, \mathbf{x}_i^U]$  are  $n_{x_i}$ -dimensional interval vectors for all  $i = 1, \dots, n$ . Consider the following real valued functions:

$$g_1(\mathbf{x}_1^L, \dots, \mathbf{x}_n^L, \mathbf{x}_1^U, \dots, \mathbf{x}_n^U) = h^L, \quad g_2(\mathbf{x}_1^L, \dots, \mathbf{x}_n^L, \mathbf{x}_1^U, \dots, \mathbf{x}_n^U) = h^U.$$

If  $H$  is continuous on  $Y = Y_1 \times \dots \times Y_n$ , where  $Y_i = [\mathbf{y}_i^L, \mathbf{y}_i^U] \subset \mathbb{R}^{n_{x_i}}$  for all  $i = 1, \dots, n$ , then  $g_1$  and  $g_2$  are continuous on  $Y \times Y$ , and bounded by a constant  $M$  there.

*Proof.* The proof of continuity is elementary, employing the metric topology for intervals defined in [24] and the max norm. That the functions are bounded follows from continuity on the compact set  $Y \times Y$ .  $\square$

THEOREM 5.6. Let  $\{\hat{P}_k\}$  be a convergent sequence of interval vectors such that

$$(5.8) \quad \lim_{k \rightarrow \infty} \hat{P}_k = \hat{P}^* = [\hat{\mathbf{p}}^*, \hat{\mathbf{p}}^*],$$

where  $\hat{P}^* \in \hat{P}$ . Let Corollary 5.4 be used to construct (5.3). For all  $i \in \mathcal{E}$ , let the epoch  $\hat{I}_i$  be split into a finite number  $(\pi_i)$  of contiguous subepochs  $\tilde{I}_l = [\tilde{\sigma}_l, \tilde{\tau}_l]$ , where  $\tilde{\sigma}_1 = i-1$ ,  $\tilde{\tau}_{\pi_i} = i$ , and  $\tilde{\sigma}_{l+1} = \tilde{\tau}_l$  for all  $l = 1, \dots, \pi_i - 1$ . If the interval extensions  $\Gamma_j(m_i, \cdot)$  and  $\Lambda_j(m_i, \cdot)$  are continuous on  $\hat{X}_j(i, \hat{P}) \times \hat{X}_{n \neq j}(i, \hat{P}) \times \hat{P} \times [\tilde{\sigma}_l, \tilde{\tau}_l]$  for all  $i \in \mathcal{E}, j = 1, \dots, n_x$ , and  $l = 1, \dots, \pi_i$ , then

$$\lim_{k \rightarrow \infty} \hat{X}(i, \underline{s}; \hat{P}_k)_k = [\hat{\mathbf{x}}(i, \hat{\mathbf{p}}^*, \underline{s}), \hat{\mathbf{x}}(i, \hat{\mathbf{p}}^*, \underline{s})] \quad \forall \underline{s} \in \hat{I}_i, i \in \mathcal{E}.$$

*Proof.* Consider the first subepoch of the first epoch,  $\tilde{I}_1$ . By definition, interval extensions have real values when their arguments are all real (degenerate interval vectors). Hence, with the degenerate interval vector  $\hat{P}^*$  as argument, the natural

interval extension (5.4) becomes  $Y(\check{\sigma}_1) = [\hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, \hat{\sigma}_1), \hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, \hat{\sigma}_1)]$ . Thus, the initial conditions for the bounding ODE system becomes

$$\mathbf{v}(\check{\sigma}_1) = \hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, \hat{\sigma}_1) = \mathbf{w}(\check{\sigma}_1).$$

This implies that the interval vector  $Z(j, 1, s)$  defined in Corollary 5.4 is degenerate at  $s = \check{\sigma}_1$ , which implies

$$\begin{aligned} v'_j(\check{\sigma}_1) &= \gamma_j^L(m_1, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^*, \hat{\mathbf{p}}^*, \check{\sigma}_1) = \mathcal{F}_j(m_1, \hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, \check{\sigma}_1), \hat{\mathbf{p}}^*, \check{\sigma}_1), \\ w'_j(\check{\sigma}_1) &= \lambda_j^U(m_1, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^*, \hat{\mathbf{p}}^*, \check{\sigma}_1) = \mathcal{F}_j(m_1, \hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, \check{\sigma}_1), \hat{\mathbf{p}}^*, \check{\sigma}_1) \end{aligned}$$

for all  $j = 1, \dots, n_x$ . We have thus defined an initial value problem in  $\mathbf{v}(s)$  and  $\mathbf{w}(s)$ . Since the solution trajectory  $\hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, s)$  is unique (Remark 4.7), this implies that the interval vector  $Z(j, 1, s)$  is degenerate for all  $j = 1, \dots, n_x$ ,  $s \in \check{I}_1$  and equal to the value  $\hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, s)$ . Hence, (5.6) and (5.7) become

$$(5.9) \quad \mathbf{v}'(s) = \mathcal{F}(m_1, \hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, s), \hat{\mathbf{p}}^*, s) = \mathbf{w}'(s), \quad \mathbf{v}(\check{\sigma}_1) = \hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, \hat{\sigma}_1) = \mathbf{w}(\check{\sigma}_1).$$

For convenience, let  $\mathbf{z}(s) = (\mathbf{v}(s), \mathbf{w}(s))$  and  $\mathbf{y} = (\hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U)$ . The system of ODEs in (5.6) and (5.7) can then be expressed as

$$\mathbf{z}' = \mathbf{f}(\mathbf{z}, \mathbf{y}, s),$$

where a solution  $\mathbf{z}(\mathbf{y}^*, s)$  exists and is unique for  $\mathbf{y}^* = (\hat{\mathbf{p}}^*, \hat{\mathbf{p}}^*)$  (since the solution  $\hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, s)$  exists and is unique). Since the interval extensions  $\Gamma_j(m_1, \cdot)$  and  $\Lambda_j(m_1, \cdot)$  are continuous for all  $j = 1, \dots, n_x$ , an application of Lemma 5.5 (treating  $x_j$  and  $s$  as degenerate intervals) gives  $\mathbf{f}$  continuous on  $\hat{X}(1, \hat{P})^2 \times \hat{P}^2 \times \check{I}_1$  and bounded by a constant  $M$  there. With  $\mathbf{z}(\mathbf{y}^*, s)$  as the unique trajectory, we can then apply [6, Theorem 4.3] to obtain  $\mathbf{z}(s) \rightarrow (\hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, s), \hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, s))$  uniformly over  $\check{I}_1$  as  $\hat{\mathbf{p}}^L \rightarrow \hat{\mathbf{p}}^*$  and  $\hat{\mathbf{p}}^U \rightarrow \hat{\mathbf{p}}^*$  (or  $\hat{P}_k \rightarrow \hat{P}^*$ ). Consider now the transition at  $\check{\tau}_1$ . Clearly, state continuity preserves the form of (5.9),

$$\mathbf{v}' = \mathcal{F}(m_1, \hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, s), \hat{\mathbf{p}}^*, s) = \mathbf{w}', \quad \mathbf{v}(\check{\sigma}_2) = \hat{\mathbf{x}}(1, \hat{\mathbf{p}}^*, \check{\sigma}_2) = \mathbf{w}(\check{\sigma}_2).$$

We can then perform the same analysis to obtain uniform convergence of the bounds over the second subepoch. By induction on all subepochs, we obtain uniform convergence over the first epoch. Consider now the transition to the second epoch at  $\hat{\tau}_1$ . With the degenerate interval vector  $\hat{P}^*$  as argument, it is clear from the preceding analysis that the natural interval extension (5.5) becomes  $Y(\hat{\sigma}_2) = [\hat{\mathbf{x}}(2, \hat{\mathbf{p}}^*, \hat{\sigma}_2), \hat{\mathbf{x}}(2, \hat{\mathbf{p}}^*, \hat{\sigma}_2)]$ . The same analysis made for  $\check{I}_1$  in  $\hat{I}_1$  thus applies, and (5.9) becomes

$$\mathbf{v}' = \mathcal{F}(m_2, \hat{\mathbf{x}}(2, \hat{\mathbf{p}}^*, s), \hat{\mathbf{p}}^*, s) = \mathbf{w}', \quad \mathbf{v}(\hat{\sigma}_2) = \hat{\mathbf{x}}(2, \hat{\mathbf{p}}^*, \hat{\sigma}_2) = \mathbf{w}(\hat{\sigma}_2).$$

The analysis carried out for the first epoch is thus valid for the second. By induction on all epochs, we have the desired result.  $\square$

*Remark 5.7.* Note that the requirement of  $\Gamma_j(m_i, \cdot)$  and  $\Lambda_j(m_i, \cdot)$  to be inclusion monotonic and continuous for all  $j = 1, \dots, n_x$  and  $i \in \mathcal{E}$  is not a strong one. For linear time invariant hybrid systems, it is automatically satisfied since (4.3) becomes a rational function (in the sense of interval analysis [24]). For time varying hybrid systems, inclusion monotonic interval extensions of the time varying matrices in (4.3) can be constructed for most functions in computing provided that no division by an interval containing zero occurs (see, e.g., [24, Chapters 3 and 4] and [26, Chapter

1]). In addition, since the functions of interest are continuous in each subepoch, the constructed interval extensions will also be continuous (see, e.g., [2, Theorem 4 and Corollary 5]).

Next, we will show how convex and concave relaxations for the states of the transformed hybrid system can be constructed.

**DEFINITION 5.8.** *Consider the functions  $f : Z \times \hat{P} \times S \rightarrow \mathbb{R}$  and  $\mathbf{z} : S \rightarrow Z$ , where  $Z \subset \mathbb{R}^{n_x}$ ,  $\hat{P} \subset \mathbb{R}^{n_p+n_e}$ ,  $S \subset \mathbb{R}$ , and  $f(\cdot, \underline{s})$  is differentiable on some suitable open set containing  $Z \times \hat{P}$  for each  $\underline{s} \in S$ . Define the function  $\mathcal{L}_f|_{\zeta^*(s)} : Z \times \hat{P} \times S \rightarrow \mathbb{R}$  to be a linearization of  $f$  at the point  $\zeta^*(s) = (\mathbf{z}^*(s), \hat{\mathbf{p}}^*)$ , where  $(\mathbf{z}^*(s), \hat{\mathbf{p}}^*) \in Z \times \hat{P}$ , and given by the following:*

$$\begin{aligned} \mathcal{L}_f|_{\zeta^*(s)}(\mathbf{z}, \hat{\mathbf{p}}, s) = & f(\mathbf{z}^*, \hat{\mathbf{p}}^*, s) + \sum_{k=1}^{n_x} \frac{\partial f}{\partial z_k} \bigg|_{(\zeta^*(s), s)} (z_k(s) - z_k^*(s)) \\ & + \sum_{k=1}^{n_p} \frac{\partial f}{\partial \hat{p}_k} \bigg|_{(\zeta^*(s), s)} (\hat{p}_k - \hat{p}_k^*). \end{aligned}$$

**THEOREM 5.9.** *For  $i \in \mathcal{E}$  and  $j = 1, \dots, n_x$ , define the functions  $u_j(m_i, \cdot, \underline{s}) : \hat{X}(i, s; \hat{P}) \times \hat{P} \rightarrow \mathbb{R}$  and  $o_j(m_i, \cdot, \underline{s}) : \hat{X}(i, s; \hat{P}) \times \hat{P} \rightarrow \mathbb{R}$  for each fixed  $\underline{s} \in \hat{I}_i$ . Let the following conditions be satisfied for all  $i \in \mathcal{E}$ ,  $j = 1, \dots, n_x$  and each fixed  $\underline{s} \in \hat{I}_i$ :*

- E1.  $u_j(m_i, \cdot, \underline{s})$  is a convex underestimator and  $o_j(m_i, \cdot, \underline{s})$  is a concave overestimator for  $\mathcal{F}_j(m_i, \cdot, \underline{s})$  on  $\hat{X}(i, \underline{s}; \hat{P}) \times \hat{P}$ .
- E2.  $u_j(m_i, \cdot, \underline{s})$  and  $o_j(m_i, \cdot, \underline{s})$  are differentiable on some suitable open set containing  $\hat{X}(i, s; \hat{P}) \times \hat{P}$  along some reference trajectory  $\zeta^*(s) = (\mathbf{z}^*(s), \hat{\mathbf{p}}^*) \in \hat{X}(i, s; \hat{P}) \times \hat{P}$ .

Therefore the following ODE system can be constructed:

$$\begin{aligned} c'_j = h_{c,j}(m_i, \mathbf{c}, \mathbf{C}, \hat{\mathbf{p}}, s) &= \inf_{\substack{\mathbf{z} \in \mathcal{C}(\hat{\mathbf{p}}, s) \\ z_j = c_j(s)}} \mathcal{L}_{u_j(m_i, \cdot)}(\mathbf{z}, \hat{\mathbf{p}}, s) \big|_{(\zeta^*(s), s)}, \quad s \in (i-1, i], \\ C'_j = h_{C,j}(m_i, \mathbf{c}, \mathbf{C}, \hat{\mathbf{p}}, s) &= \sup_{\substack{\mathbf{z} \in \mathcal{C}(\hat{\mathbf{p}}, s) \\ z_j = C_j(s)}} \mathcal{L}_{o_j(m_i, \cdot)}(\mathbf{z}, \hat{\mathbf{p}}, s) \big|_{(\zeta^*(s), s)}, \quad s \in (i-1, i], \end{aligned}$$

with initial conditions for each epoch  $\hat{I}_i$  given by

$$(5.10) \quad \mathbf{c}(\hat{\mathbf{p}}, 0) = \mathbf{C}(\hat{\mathbf{p}}, 0) = \mathbf{E}(0)\mathbf{p} + \mathbf{J}(0)\boldsymbol{\delta} + \mathbf{k}(0),$$

$$(5.11) \quad [\mathbf{c}(\hat{\mathbf{p}}, \hat{\sigma}_{l+1}), \mathbf{C}(\hat{\mathbf{p}}, \hat{\sigma}_{l+1})] = \mathbf{D}(l)[\mathbf{c}(\hat{\mathbf{p}}, \hat{\tau}_l), \mathbf{C}(\hat{\mathbf{p}}, \hat{\tau}_l)] + \mathbf{E}(l)\mathbf{p} + \mathbf{J}(l)\boldsymbol{\delta} + \mathbf{k}(l),$$

for  $l = 1, \dots, n_e - 1$ , where  $\mathcal{C}(\hat{\mathbf{p}}, s) = \{\mathbf{z} \mid \mathbf{c}(\hat{\mathbf{p}}, s) \leq \mathbf{z} \leq \mathbf{C}(\hat{\mathbf{p}}, s)\}$ . Then, for each fixed  $\underline{s} \in \hat{I}_i$ ,  $\mathbf{c}(\cdot, \underline{s})$  is a convex underestimator and  $\mathbf{C}(\cdot, \underline{s})$  is a concave overestimator for  $\hat{\mathbf{x}}(i, \cdot, \underline{s})$  on  $\hat{P}$  for all  $i \in \mathcal{E}$ .

*Proof.* We proceed as in the proof of Theorem 5.2 by subdividing the epochs into contiguous subepochs. Consider now the first subepoch  $\tilde{I}_1$ . The initial condition given by (5.10) is clearly affine on  $\hat{P}$  and satisfies  $\mathbf{c}(\hat{\mathbf{p}}, 0) \leq \hat{\mathbf{x}}(1, \hat{\mathbf{p}}, 0) \leq \mathbf{C}(\hat{\mathbf{p}}, 0)$ . The conditions for [30, Theorem 6.16] are thus satisfied, and applying said theorem,  $\mathbf{c}(\cdot, \underline{s})$  is a convex underestimator and  $\mathbf{C}(\cdot, \underline{s})$  is a concave overestimator for  $\hat{\mathbf{x}}(1, \cdot, \underline{s})$  for

each fixed  $\underline{s} \in \hat{I}_1$ . At the transition  $\tilde{\tau}_1$ , state continuity of the hybrid system gives  $\hat{x}(1, \hat{\mathbf{p}}, \tilde{\sigma}_2) = \hat{x}(1, \hat{\mathbf{p}}, \tilde{\tau}_1)$ , which implies that

$$\mathbf{c}(\hat{\mathbf{p}}, \tilde{\sigma}_2) = \mathbf{c}(\hat{\mathbf{p}}, \tilde{\tau}_1) \leq \hat{x}(1, \hat{\mathbf{p}}, \tilde{\sigma}_2) \leq \mathbf{C}(\hat{\mathbf{p}}, \tilde{\tau}_1) = \mathbf{C}(\hat{\mathbf{p}}, \tilde{\sigma}_2) \quad \forall \hat{\mathbf{p}} \in \hat{P}.$$

From [30, Theorem 6.16], we know that  $\mathbf{c}(\cdot, \tilde{\sigma}_2)$  and  $\mathbf{C}(\cdot, \tilde{\sigma}_2)$  are affine in  $\hat{\mathbf{p}}$ . The conditions for [30, Theorem 6.16] are thus satisfied for the second subepoch. By induction on all subepochs, the desired result holds for each fixed  $\underline{s} \in \hat{I}_1$ . Consider now the second epoch  $\hat{I}_2$ . From (5.11),

$$\mathbf{c}(\hat{\mathbf{p}}, \hat{\sigma}_2) \leq \hat{\mathbf{x}}(2, \hat{\mathbf{p}}, \hat{\sigma}_2) \leq \mathbf{C}(\hat{\mathbf{p}}, \hat{\sigma}_2) \quad \forall \hat{\mathbf{p}} \in \hat{P},$$

where  $\mathbf{c}(\cdot, \hat{\sigma}_2)$  and  $\mathbf{C}(\cdot, \hat{\sigma}_2)$  are clearly affine in  $\hat{\mathbf{p}}$ . The conditions for [30, Theorem 6.16] are thus satisfied for the second epoch, and by induction on all epochs, we obtain the desired result.  $\square$

Note that the infima and suprema in Theorem 5.9 are attained at the vertices of the set  $\mathcal{C}(\hat{\mathbf{p}}, s)$  due to the properties of the linearizations, and are easily computed; see [30, Theorem 6.16]. The next theorem demonstrates the convergence properties of the convex relaxations constructed using the relaxation techniques presented in this section.

**THEOREM 5.10.** *Consider the following convex relaxation of (4.4):*

$$(5.12) \quad \hat{U}(\hat{\mathbf{p}}; \hat{P}) = \sum_{i=1}^{n_e} \left\{ \sum_{j=1}^{n_{\phi i}} \hat{\psi}_{ij} \left( \mathbf{c}(\hat{\mathbf{p}}, \hat{\alpha}_{ij}), \mathbf{C}(\hat{\mathbf{p}}, \hat{\alpha}_{ij}), \hat{\mathbf{p}}; \hat{X}(i, \hat{\alpha}_{ij}; \hat{P}), \hat{P} \right) + \int_{i-1}^i \hat{u}_i \left( \mathbf{c}, \mathbf{C}, \hat{\mathbf{p}}, s; \hat{X}(i, s; \hat{P}), \hat{P} \right) v(\delta, s) \, ds \right\},$$

where  $\hat{\psi}_{ij}$  and  $\hat{u}_i$  are constructed using any relaxation technique that possesses a consistent bounding operation [17, Definition IV.4, p. 128], the convex and concave relaxations for the state and derivatives are constructed using Theorem 5.9, and the estimation of the state bounds constructed using Corollary 5.4. If the interval vector  $\hat{P}_k$  in any partition on  $\hat{P}$  approaches degeneracy  $\hat{P}^*$ , then the lower bound on this partition  $\hat{U}(\hat{\mathbf{p}}; \hat{P}_k)$  converges pointwise to the objective function value  $\hat{F}(\hat{\mathbf{p}})$  in this same partition.

*Proof.* Choose any arbitrary partition and any fixed  $\underline{s}$  in any epoch  $\hat{I}_i$ . From Theorem 5.6, as  $\hat{P}_k \rightarrow \hat{P}^*$ , the interval vector  $\hat{X}(i, \underline{s}; \hat{P}_k)_k$  approaches the degenerate value of  $\hat{\mathbf{x}}^*(i, \hat{\mathbf{p}}^*, \underline{s})$ . To be valid, the convex and concave overestimators ( $u_j(m_i, \cdot)$  and  $o_j(m_i, \cdot)$ ) from Theorem 5.9 must themselves possess a consistent bounding operation for all  $j = 1, \dots, n_x$ . Hence, as  $\hat{X}(i, \underline{s}; \hat{P}_k)_k \times \hat{P}_k$  shrinks to degeneracy,  $u_j(m_i, \cdot, \underline{s}) \uparrow \mathcal{F}_j(m_i, \cdot, \underline{s})$  and  $o_j(m_i, \cdot, \underline{s}) \downarrow \mathcal{F}_j(m_i, \cdot, \underline{s})$  for  $j = 1, \dots, n_x$ . The right-hand sides of the equations defining  $c'_i$  and  $C'_i$  are linearizations on  $u_j(m_i, \cdot)$  and  $o_j(m_i, \cdot)$ , respectively. Since  $u_j(m_i, \cdot, \underline{s})$  and  $o_j(m_i, \cdot, \underline{s})$  are each approaching  $\mathcal{F}_j(m_i, \cdot, \underline{s})$ ,  $h_{c,j}(m_i, \cdot, \underline{s}) \uparrow \mathcal{F}_j(m_i, \cdot, \underline{s})$ , and  $h_{C,j}(m_i, \cdot, \underline{s}) \downarrow \mathcal{F}_j(m_i, \cdot, \underline{s})$  because the linearization approaches the value of the function it approximates at the point of linearization. Thus, as  $k \rightarrow \infty$ ,

$$(5.13) \quad \hat{\psi}_{ij} \left( \mathbf{c}(\hat{\mathbf{p}}, \hat{\alpha}_{ij}), \mathbf{C}(\hat{\mathbf{p}}, \hat{\alpha}_{ij}), \hat{\mathbf{p}}; \hat{X}(i, \hat{\alpha}_{ij}; \hat{P}_k), \hat{P} \right) \uparrow \phi_{ij} \left( \hat{\mathbf{x}}(i, \hat{\mathbf{p}}, \hat{\alpha}_{ij}), \hat{\mathbf{p}} \right)$$

for all  $j = 1, \dots, n_{\phi i}$  and  $\hat{u}_i(\mathbf{c}, \mathbf{C}, \hat{\mathbf{p}}, \underline{s}; \hat{X}(i, \underline{s}; \hat{P}_k), \hat{P}_k) \uparrow f_i(\hat{\mathbf{x}}, \hat{\mathbf{p}}, \underline{s})$  for all  $\underline{s} \in \hat{I}_i$ , where the convergence arises because the convex relaxations  $\hat{\psi}_{ij}$  and  $\hat{u}_i$  possess consistent

bounding operations as  $\hat{P}_k$  approaches degeneracy. Because  $\underline{s}$  is fixed arbitrarily, the convergence for the integrand is true for all  $\underline{s} \in \hat{I}_i$ . An application of the monotone convergence theorem [27, Theorem 11.28] for the integral term then gives

$$(5.14) \quad \lim_{k \rightarrow \infty} \int_{i-1}^i \hat{u}_i(\mathbf{c}, \mathbf{C}, \hat{\mathbf{p}}, s; \hat{X}(i, s; \hat{P}_k), \hat{P}_k) v(\boldsymbol{\delta}, s) \, ds = \int_{i-1}^i f_i(\hat{\mathbf{x}}, \hat{\mathbf{p}}^*, s) v(\boldsymbol{\delta}, s) \, ds.$$

Since the partition and epoch was arbitrarily chosen, (5.13) and (5.14) imply that  $\lim_{k \rightarrow \infty} \hat{U}(\hat{\mathbf{p}}; \hat{P}_k) = \hat{F}(\hat{\mathbf{p}}^*)$ .  $\square$

**6. Results and discussion.** In the first example, we walk through a procedure for bounding the solution of the transformed hybrid system using Corollary 5.4. In the second example, we present an interesting problem from chemical reaction engineering that can be posed in the form of Problem 3.1 and solved using the techniques developed in this article.

*Example 6.1.* Consider the following linear hybrid system:

$$\begin{aligned} \text{Mode 1: } & \begin{cases} \dot{x}_1 &= 0.5x_1 + x_2 + p_1, \\ \dot{x}_2 &= -x_1 + x_2 + p_1, \end{cases} \\ \text{Mode 2: } & \begin{cases} \dot{x}_1 &= x_1 + x_2 - p_2, \\ \dot{x}_2 &= -x_1 + p_2, \end{cases} \end{aligned}$$

$\mathcal{E} = \{1, 2\}$ ,  $T_\mu = 1, 2$ ,  $P \equiv [0, 1]^2$ ,  $\Delta \equiv [0, 1]^2$ , and we have state continuity as the transition functions with initial condition  $\mathbf{x}(1, \mathbf{p}, \boldsymbol{\delta}, 0) = (0, 2)$ .

Applying the CPET, we obtain the following transformed nonlinear hybrid system:

$$\begin{aligned} \text{Mode 1: } & \begin{cases} \hat{x}'_1 &= v(0.5\hat{x}_1 + \hat{x}_2 + p_1), \\ \hat{x}'_2 &= v(-\hat{x}_1 + \hat{x}_2 + p_1), \end{cases} \\ \text{Mode 2: } & \begin{cases} \hat{x}'_1 &= v(\hat{x}_1 + \hat{x}_2 - p_2), \\ \hat{x}'_2 &= v(-\hat{x}_1 + p_2), \end{cases} \end{aligned}$$

$\mathcal{E} = \{1, 2\}$ ,  $T_\mu = 1, 2$ ,  $\hat{P} \equiv [0, 1]^2 \times [0, 1]^2$ , and we have state continuity as the transition functions with initial condition  $\hat{\mathbf{x}}(1, \hat{\mathbf{p}}, 0) = (0, 2)$ .

We now apply Corollary 5.4 to obtain bounds for the transformed hybrid system. For this example, we assume that we do not have additional bounding information, and so the user defined set  $\hat{\mathcal{X}}(i, \hat{\mathbf{p}}, s)$  is set to  $\mathbb{R}^2$ . From (5.4),  $Y(0) = [\mathbf{k}(0), \mathbf{k}(0)]$ , where  $\mathbf{k}(0) = (0, 2)$  since  $\mathbf{E}(0)$  and  $\mathbf{J}(0)$  are zero matrices. Expanding the right-hand sides of the nonlinear ODEs in epoch  $\hat{I}_1$  (note that  $v = \delta_1$  for epoch  $\hat{I}_1$ ), and taking the natural interval extensions, we obtain the following forms for  $\Gamma(1, \cdot)$  and  $\Lambda(1, \cdot)$  in Corollary 5.4:

$$\begin{aligned} \Gamma_1(1, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U, s) &= [\delta_1^L, \delta_1^U] \cdot (0.5v_1(s) + [v_2(s), w_2(s)] + [p_1^L, p_1^U]), \\ \Gamma_2(1, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U, s) &= [\delta_1^L, \delta_1^U] \cdot (-[v_1(s), w_1(s)] + v_2(s) + [p_1^L, p_1^U]), \\ \Lambda_1(1, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U, s) &= [\delta_1^L, \delta_1^U] \cdot (0.5w_1(s) + [v_2(s), w_2(s)] + [p_1^L, p_1^U]), \\ \Lambda_2(1, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U, s) &= [\delta_1^L, \delta_1^U] \cdot (-[v_1(s), w_1(s)] + w_2(s) + [p_1^L, p_1^U]). \end{aligned}$$



At the transition, state continuity gives  $Y(1) = [\mathbf{v}(1), \mathbf{w}(1)]$  for (5.5) since  $\mathbf{D}(1)$  is the identity matrix,  $\mathbf{E}(1)$  and  $\mathbf{J}(1)$  are zero matrices, and  $\mathbf{k}(1)$  is a zero vector. Similarly, expanding and taking the natural interval extensions of right-hand sides of the nonlinear ODEs in epoch  $\hat{I}_2$  (and noting that  $v = \delta_2$  for epoch  $\hat{I}_2$ ), we obtain

$$\begin{aligned}\Gamma_1(2, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U, s) &= [\delta_2^L, \delta_2^U] \cdot (v_1(s) + [v_2(s), w_2(s)] - [p_2^L, p_2^U]), \\ \Gamma_2(2, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U, s) &= [\delta_2^L, \delta_2^U] \cdot (-[v_1(s), w_1(s)] + [p_2^L, p_2^U]), \\ \Lambda_1(2, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U, s) &= [\delta_2^L, \delta_2^U] \cdot (w_1(s) + [v_2(s), w_2(s)] - [p_2^L, p_2^U]), \\ \Lambda_2(2, \mathbf{v}, \mathbf{w}, \hat{\mathbf{p}}^L, \hat{\mathbf{p}}^U, s) &= [\delta_2^L, \delta_2^U] \cdot (-[v_1(s), w_1(s)] + [p_2^L, p_2^U]).\end{aligned}$$

Applying Corollary 5.4, the following ODE system bounds the transformed hybrid system,  $\mathbf{v}(s) \leq \mathbf{x}(s) \leq \mathbf{w}(s)$ :

$$\text{Epoch } \hat{I}_1: \left\{ \begin{array}{l} v_1'(s) = \min \left( \delta_1^L(0.5v_1(s) + v_2(s) + p_1^L), \delta_1^L(0.5v_1(s) + w_2(s) + p_1^U), \right. \\ \quad \left. \delta_1^U(0.5v_1(s) + v_2(s) + p_1^L), \delta_1^U(0.5v_1(s) + w_2(s) + p_1^U) \right), \\ v_2'(s) = \min \left( \delta_1^L(-w_1(s) + v_2(s) + p_1^L), \delta_1^L(-v_1(s) + v_2(s) + p_1^U), \right. \\ \quad \left. \delta_1^U(-w_1(s) + v_2(s) + p_1^L), \delta_1^U(-v_1(s) + v_2(s) + p_1^U) \right), \\ w_1'(s) = \max \left( \delta_1^L(0.5w_1(s) + v_2(s) + p_1^L), \delta_1^L(0.5w_1(s) + w_2(s) + p_1^U), \right. \\ \quad \left. \delta_1^U(0.5w_1(s) + v_2(s) + p_1^L), \delta_1^U(0.5w_1(s) + w_2(s) + p_1^U) \right), \\ w_2'(s) = \max \left( \delta_1^L(-w_1(s) + w_2(s) + p_1^L), \delta_1^L(-v_1(s) + w_2(s) + p_1^U), \right. \\ \quad \left. \delta_1^U(-w_1(s) + w_2(s) + p_1^L), \delta_1^U(-v_1(s) + w_2(s) + p_1^U) \right), \end{array} \right.$$

with initial conditions  $\mathbf{v}(0) = \mathbf{w}(0) = (0, 2)$ , and

$$\text{Epoch } \hat{I}_2: \left\{ \begin{array}{l} v_1'(s) = \min \left( \delta_2^L(v_1(s) + v_2(s) - p_2^U), \delta_2^L(v_1(s) + w_2(s) - p_2^L), \right. \\ \quad \left. \delta_2^U(v_1(s) + v_2(s) - p_2^U), \delta_2^U(v_1(s) + w_2(s) - p_2^L) \right), \\ v_2'(s) = \min \left( \delta_2^L(-w_1(s) + p_2^L), \delta_2^L(-v_1(s) + p_2^U), \right. \\ \quad \left. \delta_2^U(-w_1(s) + p_2^L), \delta_2^U(-v_1(s) + p_2^U) \right), \\ w_1'(s) = \max \left( \delta_2^L(w_1(s) + v_2(s) - p_2^U), \delta_2^L(w_1(s) + w_2(s) - p_2^L), \right. \\ \quad \left. \delta_2^U(w_1(s) + v_2(s) - p_2^U), \delta_2^U(w_1(s) + w_2(s) - p_2^L) \right), \\ w_2'(s) = \max \left( \delta_2^L(-w_1(s) + p_2^L), \delta_2^L(-v_1(s) + p_2^U), \right. \\ \quad \left. \delta_2^U(-w_1(s) + p_2^L), \delta_2^U(-v_1(s) + p_2^U) \right), \end{array} \right.$$

with initial conditions  $\mathbf{v}(1) = \mathbf{v}^*$ ,  $\mathbf{w}(1) = \mathbf{w}^*$ , where  $\mathbf{v}^*$  and  $\mathbf{w}^*$  are the final values of the bounding ODE system for the first epoch at  $s = 1$ .

This bounding ODE system can be integrated efficiently with an integrator that supports the rigorous detection of events, due to the min and max functions in the right-hand sides. The following results are obtained using the JACOBIAN Dynamic Modeling and Optimization Software [22] release 2.1A with the default options.

Figure 6.1 shows the bounding trajectories obtained for  $P \equiv [0, 1]^2$  and  $\Delta \equiv [0, 1]^2$ . To illustrate that the trajectories actually bound the transformed system, 20

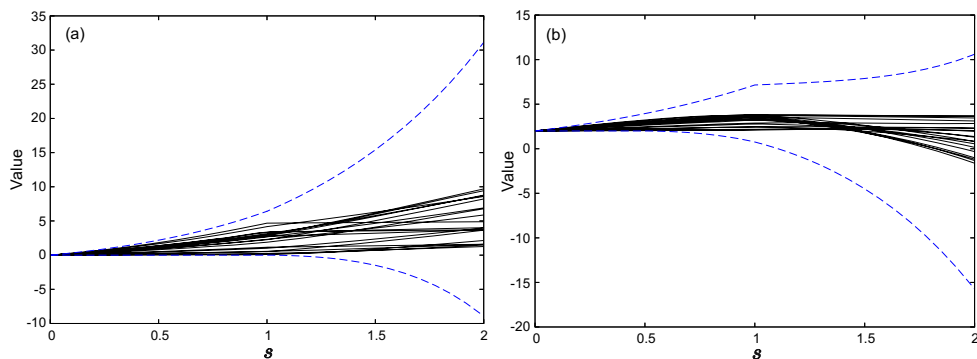


FIG. 6.1. Bounding trajectories (dashed lines) and random state trajectories (solid lines) with  $P \equiv [0, 1]^2$  and  $\Delta \equiv [0, 1]^2$  for (a)  $\hat{x}_1(s)$  and (b)  $\hat{x}_2(s)$ .

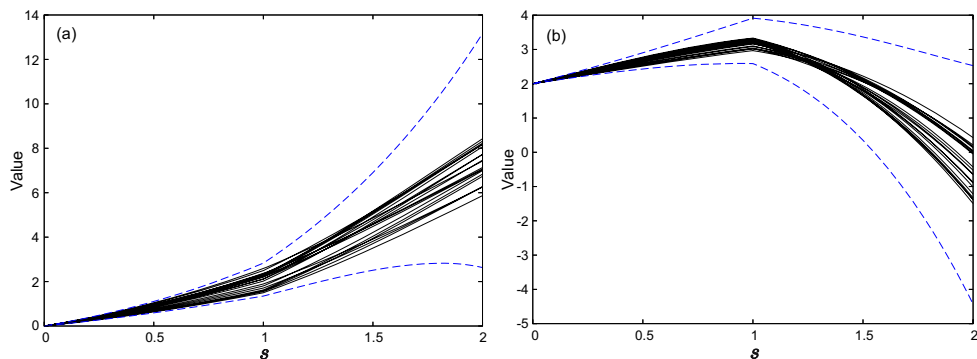


FIG. 6.2. Bounding trajectories (dashed lines) and random state trajectories (solid lines) with  $P \equiv [0, 0.25] \times [0.25, 0.5]$  and  $\Delta \equiv [0.5, 0.75] \times [0.75, 1]$  for (a)  $\hat{x}_1(s)$  and (b)  $\hat{x}_2(s)$ .

random points were generated in  $P \times \Delta$ , and the state trajectories of the transformed system were plotted alongside the bounding trajectories. It can be seen that the bounds indeed enclose the solution of the transformed system, as should be expected with the application of Corollary 5.4. Figure 6.2 shows what happens when the bounds on  $(\mathbf{p}, \delta)$  are changed to  $P \equiv [0, 0.25] \times [0.25, 0.5]$ ,  $\Delta \equiv [0.5, 0.75] \times [0.75, 1]$ . Again, 20 random points of  $(\mathbf{p}, \delta)$  were generated in  $P \times \Delta$  and plotted together with the new bounding trajectories. Besides bounding the state trajectories, it can be seen from the scales of the vertical axis that the bounding trajectories are closer together than those in Figure 6.1. Finally, Figure 6.3 illustrates the convergence of the bounding trajectories in Theorem 5.6 as  $P$  and  $\Delta$  become degenerate. Note that for the degenerate intervals  $P \equiv \Delta \equiv [0.5, 0.5]^2$  (case (f)), the bounding trajectories become the same, i.e.,  $\mathbf{v}(s) = \mathbf{w}(s) = \hat{\mathbf{x}}(i, 0.5, 0.5, s)$ .

*Example 6.2.* Consider an isothermal plug flow reactor (PFR) operating at steady state with three sections into which the catalyst can be loaded. It has been determined that catalysts 1, 2, and 3 will be loaded in that order into the reactor. The optimization decision variables are the lengths of the catalyst sections. Table 6.1 lists the associated rate constants of the following kinetic model:

$$\begin{aligned} \dot{x}_1(t) &= -(k_1 + k_2)x_1, & \dot{x}_2(t) &= k_2x_1, & \dot{x}_3(t) &= k_1x_1 - (k_3 + k_4)x_3, \\ \dot{x}_4(t) &= k_4x_3, & \dot{x}_5(t) &= k_3x_3, \end{aligned}$$

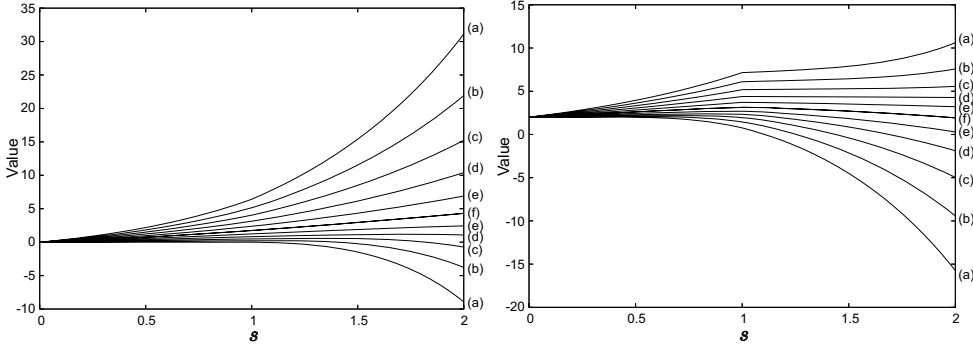


FIG. 6.3. Bounding trajectories with  $P \equiv \Delta \equiv Z^2$ , where  $Z$  is given by (a)  $[0, 1]$ , (b)  $[0.1, 0.9]$ , (c)  $[0.2, 0.8]$ , (d)  $[0.3, 0.7]$ , (e)  $[0.4, 0.6]$ , and (f)  $[0.5, 0.5]$ . The plot on the left is for  $\hat{x}_1(s)$ , while the one on the right is for  $\hat{x}_2(s)$ .

TABLE 6.1  
Rate constants for each catalyst.

Catalyst	$k_1$	$k_2$	$k_3$	$k_4$
1	2.098	1.317	0.021	0.033
2	29.53	110.2	0.295	0.079
3	182.6	2325	1.826	0.143

where  $x_i$  represents the molar concentration of component  $i$  ( $\text{mol m}^{-3}$ ), and  $k_j$  represents the rate constant of reaction  $j$  ( $\text{min}^{-1}$ ). The PFR has a uniform cross-sectional area of  $0.05 \text{ m}^2$ , and a constant volumetric flow rate of  $0.05 \text{ m}^3 \text{ min}^{-1}$ . In this example, the independent variable  $t$  is the distance,  $l$ , along the reactor. The objective function,  $F$  (\$k), is to maximize the profit from the process, which is the value of the product  $x_5$  minus the treatment costs of the byproducts  $x_2$  and  $x_4$ ,

$$\min_{\delta \in \Delta} -F = 0.01x_2(1) + 0.1x_4(1) - x_5(1).$$

After applying the CPET, the transformed problem is given by

$$\min_{\delta \in \Delta} 0.01\hat{x}_2(\delta, 3) + 0.1\hat{x}_4(\delta, 3) - \hat{x}_5(\delta, 3),$$

subject to the point constraint,

$$(6.1) \quad \delta_1 + \delta_2 + \delta_3 = 1,$$

where  $\hat{\mathbf{x}}(i, \delta, s)$  is given by the solution of the following CPET hybrid system:

$$\text{Mode } m: \left\{ \begin{array}{l} \hat{x}'_1(s) = v \left( -(k_1(i) + k_2(i))\hat{x}_1 \right), \\ \hat{x}'_2(s) = v \left( k_2(i)\hat{x}_1 \right), \\ \hat{x}'_3(s) = v \left( k_1(i)\hat{x}_1 - (k_3(i) + k_4(i))\hat{x}_3 \right), \\ \hat{x}'_4(s) = v \left( k_4(i)\hat{x}_3 \right), \\ \hat{x}'_5(s) = v \left( k_3(i)\hat{x}_3 \right) \end{array} \right\} m = 1, 2, 3,$$

$\mathcal{E} = \{1, 2, 3\}$ ,  $T_\mu = 1, 2, 3$ ,  $\hat{P} \equiv \Delta \equiv [0, 1]^3$ , and we have state continuity as the transition functions with initial condition  $\hat{\mathbf{x}}(1, \boldsymbol{\delta}, 0) = (1000, 0, 0, 0, 0)$ .

The convex relaxations for this example are constructed directly using Theorem 5.9 as the original objective function comprises an affine function of the state variables at the final time. Since the right-hand sides of the nonlinear hybrid system exhibit a bilinear structure, the convex and concave relaxations in Theorem 5.9 can be calculated from the convex envelope of a bilinear term [9]. However, since the convex envelope is composed of two intersecting hyperplanes and thus not continuously differentiable everywhere, there is no guarantee that condition E2 will be satisfied for a particular choice of a reference trajectory. Fortunately, it is clear that the condition can be relaxed to accommodate the nonsmoothness in the intersection of the two hyperplanes by constructing the linearizations using any subgradient at the point of nonsmoothness. In practice, we have implemented a heuristic that either chooses one or the other hyperplane (which are both valid convex relaxations and supply valid subgradients), where the effects of any possible chattering in the numerical integration can be mitigated; see [30, Chapter 7].

It is also possible to remove a degree of freedom,  $\delta_3$ , from the optimization problem by substituting it with  $1 - \delta_1 - \delta_2$  and eliminating the constraint (6.1) from the problem. This is attractive because it reduces the dimension of the parameter space in the branch-and-bound framework. The numerical implementation used for solving this problem is as follows: the convex relaxations constructed using Theorem 5.9 and the natural interval extensions of Corollary 5.4 were generated automatically based on an operator-overloading approach using C++; the local dynamic optimizations were performed using the code DYNO [11], which implements the control parametrization approach; and the branch-and-bound framework used was libBandB 3.2 [31]. Using a Pentium 4 2.6 GHz machine with 1 GB RAM running SuSE Linux 9.2, a reference trajectory of  $(\hat{\mathbf{x}}, \boldsymbol{\delta})_k = (\hat{\mathbf{x}}^L, \boldsymbol{\delta}^L)_k$ , a relative tolerance for libBandB of  $10^{-3}$ , relative and absolute tolerances for the numerical integrator in DYNO of  $10^{-7}$ , and an optimality tolerance for the NLP solver in DYNO of  $10^{-5}$ , an optimal solution value of 314.2 was obtained, at the point  $\boldsymbol{\delta}^* = (0.3626, 0.0196, 0.6178)$ . There was a total of 483 nodes visited in the branch-and-bound tree, with a total CPU time of 560 seconds. For comparison, if the problem only involved two sections with the mode sequence  $T_\mu = 1, 3$ , then an optimal solution value of 296.9 was obtained, at the point  $\boldsymbol{\delta}^* = (0.4181, 0.5819)$ , with a total of 17 nodes and a total CPU time of 8 seconds.

**7. Conclusion.** The global optimization problem with continuous time linear hybrid systems embedded has been considered where the embedded systems have varying time transitions. Sufficient conditions have been proposed for these problems to be smooth in the control parametrization framework. The CPET has been utilized to transform the problem into a global optimization problem with nonlinear hybrid systems embedded where the transitions are now fixed in time. A method of constructing convex relaxations for the transformed problem has been developed that is shown to be convergent within a branch-and-bound framework.

**Acknowledgments.** The authors would like to thank Dr. A. B. Singer and Dr. B. Chachuat for extremely helpful discussions.

#### REFERENCES

- [1] C. S. ADJIMAN, S. DALLWIG, C. A. FLOUDAS, AND A. NEUMAIER, *A global optimization method,  $\alpha$ BB, for general twice-differentiable constrained NLPs. I. Theoretical advances*, Comput. Chem. Eng., 22 (1998), pp. 1137–1158.

- [2] G. ALEFELD AND J. HERZBERGER, *Introduction to Interval Computations*, Academic Press, New York, 1983.
- [3] P. I. BARTON, J. R. BANGA, AND S. GALÁN, *Optimization of hybrid discrete/continuous dynamic systems*, *Comput. Chem. Eng.*, 24 (2000), pp. 2171–2182.
- [4] P. I. BARTON AND C. K. LEE, *Modeling, simulation, sensitivity analysis and optimization of hybrid systems*, *ACM Trans. Model. Comput. Simul.*, 12 (2002), pp. 256–289.
- [5] P. I. BARTON AND C. K. LEE, *Design of process operations using hybrid dynamic optimization*, *Comput. Chem. Eng.*, 28 (2004), pp. 955–969.
- [6] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [7] L. J. CORWIN AND R. H. SZCZARBA, *Multivariable Calculus*, Marcel Dekker, Inc., New York, 1982.
- [8] M. EGERSTEDT, Y. WARDI, AND F. DELMOTTE, *Optimal control of switching times in switched dynamical systems*, in *Proceedings of the 42nd IEEE Conference on Decision and Control*, Maui, HI, 2003, pp. 2138–2143.
- [9] J. E. FALK AND R. M. SOLAND, *An algorithm for separable nonconvex programming problems*, *Management Sci.*, 15 (1969), pp. 550–569.
- [10] W. F. FEEHERRY, J. E. TOLSMAN, AND P. I. BARTON, *Efficient sensitivity analysis of large-scale differential-algebraic systems*, *Appl. Numer. Math.*, 25 (1997), pp. 41–54.
- [11] M. FIKAR AND M. A. LATIFI, *User's Guide for Fortran Dynamic Optimisation Code DYNO*, Technical report, LSGC-CNRS, Nancy, France, and Slovak Technical University Bratislava, Bratislava, Slovak Republic, 2002.
- [12] R. FLETCHER AND S. LEYFFER, *Solving mixed integer nonlinear programs by outer approximation*, *Math. Programming*, 66 (1994), pp. 327–349.
- [13] S. GALÁN AND P. I. BARTON, *Dynamic optimization of hybrid systems*, *Comput. Chem. Eng.*, 22 (1998), pp. S183–S190.
- [14] S. GALÁN, W. F. FEEHERRY, AND P. I. BARTON, *Parametric sensitivity functions for hybrid discrete/continuous systems*, *Appl. Numer. Math.*, 31 (1999), pp. 17–47.
- [15] A. M. GEOFFRION, *Generalized Benders decomposition*, *J. Optim. Theory Appl.*, 10 (1972), pp. 237–260.
- [16] P. E. GILL, W. MURRAY, AND M. A. SAUNDERS, *SNOPT: An SQP algorithm for large-scale constrained optimization*, *SIAM J. Optim.*, 12 (2002), pp. 979–1006.
- [17] R. HORST AND H. TUY, *Global Optimization*, 3rd ed., Springer-Verlag, Berlin, 1996.
- [18] P. KESAVAN, R. J. ALLGOR, E. P. GATZKE, AND P. I. BARTON, *Outer approximation algorithms for separable nonconvex mixed-integer nonlinear programs*, *Math. Program.*, 100 (2004), pp. 517–535.
- [19] C. K. LEE, A. B. SINGER, AND P. I. BARTON, *Global optimization of linear hybrid systems with explicit transitions*, *Systems Control Lett.*, 51 (2004), pp. 363–375.
- [20] H. W. J. LEE, K. L. TEO, V. REHBOCK, AND L. S. JENNINGS, *Control parametrization enhancing technique for time optimal control problems*, *Dynam. Systems Appl.*, 6 (1997), pp. 243–261.
- [21] H. W. J. LEE, K. L. TEO, V. REHBOCK, AND L. S. JENNINGS, *Control parametrization enhancing technique for optimal discrete-valued control problems*, *Automatica J. IFAC*, 35 (1999), pp. 1401–1407.
- [22] NUMERICA TECHNOLOGY LLC, *JACOBIAN User Guide*, Numerica Technology, Cambridge, MA, 2005, available online at [www.numericatech.com](http://www.numericatech.com).
- [23] G. P. McCORMICK, *Computability of global solutions to factorable nonconvex programs I. Convex underestimating problems*, *Math. Programming*, 10 (1976), pp. 147–175.
- [24] R. E. MOORE, *Methods and Applications of Interval Analysis*, SIAM, Philadelphia, 1979.
- [25] K. R. MORISON AND R. W. H. SARGENT, *Optimization of multistage processes described by differential-algebraic equations*, in *Numerical Analysis, Proceedings of the 4th IIMAS Workshop*, Guanajuato, Mexico, J. P. Hennart, ed., *Lecture Notes in Math.* 1230, Springer-Verlag, Berlin, 1986, pp. 86–102.
- [26] H. RATSCHKE AND J. ROKNE, *Computer Methods for the Range of Functions*, Ellis Horwood Ltd., Chichester, UK, 1984.
- [27] W. RUDIN, *Principles of Mathematical Analysis*, 3rd ed., McGraw-Hill, New York, 1976.
- [28] H. S. RYOO AND N. V. SAHINIDIS, *A branch-and-reduce approach to global optimization*, *J. Global Optim.*, 8 (1996), pp. 107–139.
- [29] R. W. H. SARGENT AND G. R. SULLIVAN, *The development of an efficient optimal control package*, in *Proceedings of the 8th IFIP Conference on Optimization Techniques*, J. Stoer, ed., *Lecture Notes in Control and Inform. Sci.* 7, Springer-Verlag, Berlin, 1978, pp. 158–168.
- [30] A. B. SINGER, *Global Dynamic Optimization*, Ph.D. thesis, MIT, Cambridge, MA, 2004, avail-

- able online at <http://yoric.mit.edu/reports.html>.
- [31] A. B. SINGER, *LibBandB.a Version, 3.2 Manual*, Technical report, MIT, Cambridge, MA, 2004.
  - [32] A. B. SINGER AND P. I. BARTON, *Global solution of optimization problems with parameter-embedded linear dynamic systems*, J. Optim. Theory Appl., 121 (2004), pp. 613–646.
  - [33] A. B. SINGER AND P. I. BARTON, *Bounding the solutions of parameter dependent nonlinear ordinary differential equations*, SIAM J. Sci. Comput., 27 (2006), pp. 2167–2182.
  - [34] A. B. SINGER AND P. I. BARTON, *Global optimization with nonlinear ordinary differential equations*, J. Global Optim., 34 (2006), pp. 159–190.
  - [35] K. TEO, G. GOH, AND K. WONG, *A Unified Computational Approach to Optimal Control Problems*, Pitman Monographs and Surveys in Pure and Applied Mathematics 55, Wiley, New York, 1991.
  - [36] K. L. TEO, L. S. JENNINGS, H. W. J. LEE, AND V. REHBOCK, *The control parametrization enhancing transform for constrained optimal control problems*, J. Austral. Math. Soc. Ser. B, 40 (1999), pp. 314–335.
  - [37] J. E. TOLSMA AND P. I. BARTON, *DAEPACK: An open modeling environment for legacy models*, Ind. Eng. Chem. Res., 39 (2000), pp. 1826–1839.
  - [38] J. E. TOLSMA AND P. I. BARTON, *Hidden discontinuities and parametric sensitivity calculations*, SIAM J. Sci. Comput., 23 (2002), pp. 1861–1874.
  - [39] V. S. VASSILIADIS, R. W. H. SARGENT, AND C. C. PANTELIDES, *Solution of a class of multistage dynamic optimization problems. 1. Problems without path constraints*, Ind. Eng. Chem. Res., 33 (1994), pp. 2111–2122.
  - [40] W. WALTER, *Differential and Integral Inequalities*, Springer-Verlag, New York, 1970.
  - [41] X. XU AND P. J. ANTSAKLIS, *An approach for solving general switched linear quadratic optimal control problems*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 2478–2483.
  - [42] X. XU AND P. J. ANTSAKLIS, *Optimal control of switched systems based on parameterization of the switching instants*, IEEE Trans. Automat. Control, 49 (2004), pp. 2–16.

# THE MAX-PLUS FINITE ELEMENT METHOD FOR SOLVING DETERMINISTIC OPTIMAL CONTROL PROBLEMS: BASIC PROPERTIES AND CONVERGENCE ANALYSIS\*

MARIANNE AKIAN<sup>†</sup>, STÉPHANE GAUBERT<sup>†</sup>, AND ASMA LAKHOUE<sup>‡</sup>

**Abstract.** We introduce a max-plus analogue of the Petrov–Galerkin finite element method to solve finite horizon deterministic optimal control problems. The method relies on a max-plus variational formulation. We show that the error in the sup-norm can be bounded from the difference between the value function and its projections on max-plus and min-plus semimodules when the max-plus analogue of the stiffness matrix is exactly known. In general, the stiffness matrix must be approximated: this requires approximating the operation of the Lax–Oleinik semigroup on finite elements. We consider two approximations relying on the Hamiltonian. We derive a convergence result, in arbitrary dimension, showing that for a class of problems, the error estimate is of order  $\delta + \Delta x(\delta)^{-1}$  or  $\sqrt{\delta} + \Delta x(\delta)^{-1}$ , depending on the choice of the approximation, where  $\delta$  and  $\Delta x$  are, respectively, the time and space discretization steps. We compare our method with another max-plus based discretization method previously introduced by Fleming and McEneaney. We give numerical examples in dimensions 1 and 2.

**Key words.** max-plus algebra, tropical semiring, Hamilton–Jacobi equation, weak formulation, residuation, projection, idempotent semimodules, finite element method

**AMS subject classifications.** Primary, 49L20; Secondary, 65M60, 06A15, 12K10

**DOI.** 10.1137/060655286

**1. Introduction.** We consider the following optimal control problem:

$$(1a) \quad \text{maximize} \quad \int_0^T \ell(\mathbf{x}(s), \mathbf{u}(s)) ds + \phi(\mathbf{x}(T))$$

over the set of trajectories  $(\mathbf{x}(\cdot), \mathbf{u}(\cdot))$  satisfying

$$(1b) \quad \dot{\mathbf{x}}(s) = f(\mathbf{x}(s), \mathbf{u}(s)), \quad \mathbf{x}(s) \in X, \quad \mathbf{u}(s) \in U$$

for all  $0 \leq s \leq T$  and

$$(1c) \quad \mathbf{x}(0) = x.$$

Here the *state space*  $X$  is a subset of  $\mathbb{R}^n$ , the set of *control values*  $U$  is a subset of  $\mathbb{R}^m$ , the *horizon*  $T > 0$  and the *initial condition*  $x \in X$  are given, and we assume that the map  $\mathbf{u}(\cdot)$  is measurable and that the map  $\mathbf{x}(\cdot)$  is absolutely continuous. We also assume that the *instantaneous reward* or *Lagrangian*  $\ell : X \times U \rightarrow \mathbb{R}$  and the *dynamics*  $f : X \times U \rightarrow \mathbb{R}^n$  are sufficiently regular maps and that the *terminal reward*  $\phi$  is a map  $X \rightarrow \mathbb{R} \cup \{-\infty\}$ .

---

\*Received by the editors March 27, 2006; accepted for publication (in revised form) July 25, 2007; published electronically February 27, 2008. This work was supported by a fellowship of the IFC (Institut Français de Coopération), a fellowship of the AUF (Agence Universitaire de la Francophonie), grant STIC-INRIA-Universités tunisiennes I-04, and joint RFBR-CNRS grant 05-01-02807.

<http://www.siam.org/journals/sicon/47-2/65528.html>

<sup>†</sup>INRIA, Domaine de Voluceau, 78153 Le Chesnay Cédex, France (marianne.akian@inria.fr, stephane.gaubert@inria.fr).

<sup>‡</sup>INRIA, Domaine de Voluceau, 78153 Le Chesnay Cédex, France, and ENIT-LAMSIN, BP 37, 1002 Tunis Le Belvédère, Tunisie (asma.lakhoua@inria.fr, asma.lakhoua@lamsin.rnu.tn).

We are interested in the numerical computation of the *value function*  $v$  which associates with any  $(x, t) \in X \times [0, T]$  the supremum  $v(x, t)$  of  $\int_0^t \ell(\mathbf{x}(s), \mathbf{u}(s)) ds + \phi(\mathbf{x}(t))$ , under the constraints (1b), for  $0 \leq s \leq t$  and (1c). It is known that, under certain regularity assumptions,  $v$  is solution of the Hamilton–Jacobi equation

$$(2a) \quad -\frac{\partial v}{\partial t} + H\left(x, \frac{\partial v}{\partial x}\right) = 0, \quad (x, t) \in X \times (0, T],$$

with initial condition

$$(2b) \quad v(x, 0) = \phi(x), \quad x \in X,$$

where  $H(x, p) = \sup_{u \in U} \ell(x, u) + p \cdot f(x, u)$  is the *Hamiltonian* of the problem (see, for instance, [Lio82], [FS93], [Bar94]).

Several techniques have been proposed in the literature to solve this problem. We mention, for example, finite difference schemes and the method of the vanishing viscosity [CL84], the antidiffusive schemes for advection [BZ07], the finite element approach [GR85a], [GR85b] (in the case of the stopping time problem), the so-called discrete dynamic programming method or semi-Lagrangian method [CD83], [CDI84], [Fal87a], [Fal87b], [FF94], [FG99], [CFF04], and the Markov chain approximations [BD99]. Other schemes have been obtained by integration from the essentially nonoscillatory schemes for the hyperbolic conservation laws (see, for instance, [OS91]). Recently, max-plus methods have been proposed to solve first-order Hamilton–Jacobi equations [McE06], [MH98], [MH99], [FM00], [McE02], [McE03], [CM04], [McE04].

Recall that the *max-plus semiring*,  $\mathbb{R}_{\max}$ , is the set  $\mathbb{R} \cup \{-\infty\}$ , equipped with the addition  $a \oplus b = \max(a, b)$  and the multiplication  $a \otimes b = a + b$ . In what follows, let  $S^t$  denote the *evolution semigroup* of (2), or Lax–Oleinik semigroup, which associates with any map  $\phi$  the function  $v^t := v(\cdot, t)$ , where  $v$  is the value function of the optimal control problem (1). Maslov [Mas73] observed that the semigroup  $S^t$  is *max-plus linear*, meaning that for all maps  $f, g$  from  $X$  to  $\mathbb{R}_{\max}$ , and for all  $\lambda \in \mathbb{R}_{\max}$ , we have

$$S^t(f \oplus g) = S^t f \oplus S^t g,$$

$$S^t(\lambda f) = \lambda S^t f,$$

where  $f \oplus g$  denotes the map  $x \mapsto f(x) \oplus g(x)$ , and  $\lambda f$  denotes the map  $x \mapsto \lambda \otimes f(x)$ . Linear operators over max-plus-type semirings have been widely studied; see, for instance, [CG79], [MS92], [BCOQ92], [KM97], [GM01], [Fat].

In [FM00], Fleming and McEneaney introduced a max-plus-based discretization method to solve a subclass of Hamilton–Jacobi equations (with a Lagrangian  $\ell$  quadratic with respect to  $u$  and a dynamics  $f$  affine with respect to  $u$ ). They use the max-plus linearity of the semigroup  $S^t$  to approximate the value function  $v^t$  by a function  $v_h^t$  of the form

$$(3) \quad v_h^t = \sup_{1 \leq i \leq p} \{\lambda_i^t + w_i\},$$

where  $\{w_i\}_{1 \leq i \leq p}$  is a given family of functions (a max-plus “basis”) and  $\{\lambda_i^t\}_{1 \leq i \leq p}$  is a family of scalars (the “coefficients” of  $v_h^t$  on the max-plus “basis”), which must be determined. They proposed a discretization scheme in which  $\lambda^t$  is computed inductively by applying a max-plus linear operator to  $\lambda^{t-\delta}$ , where  $\delta$  is the time discretization step. Thus, their scheme can be interpreted as the dynamic programming equation of a discrete control problem.



In this paper, we introduce a max-plus analogue of the finite element method, the “MFEM,” to solve the deterministic optimal control problem (1). We still look for an approximation  $v_h^t$  of the form (3). However, to determine the “coefficients”  $\lambda_i^t$ , we use a max-plus analogue of the notion of variational formulation, which originates from the notion of generalized solution of Hamilton–Jacobi equations of Maslov and Kolokoltsov [KM88], [KM97, Section 3.2]. We choose a family  $\{z_j\}_{1 \leq j \leq q}$  of test functions and define  $v_h^t$  inductively to be the maximal function of the form (3) satisfying

$$(4) \quad \langle v_h^t \mid z_j \rangle \leq \langle S^\delta v_h^{t-\delta} \mid z_j \rangle \quad \forall 1 \leq j \leq q,$$

where  $\langle \cdot \mid \cdot \rangle$  denotes the max-plus scalar product (see section 3 for details). We show that the corresponding vector of coefficients  $\lambda^t$  can be obtained by applying to  $\lambda^{t-\delta}$  a nonlinear operator, which can be interpreted as the dynamic programming operator of a deterministic zero-sum two player game, with finite action and state spaces. Indeed,

$$\lambda_i^t = \min_{1 \leq j \leq q} \left( - (M_h)_{ji} + \max_{1 \leq k \leq p} ((K_h)_{jk} + \lambda_k^{t-\delta}) \right),$$

where the matrices  $M_h$  and  $K_h$  are max-plus analogues of the mass and stiffness matrices, respectively (see (20) and (21) for the definitions of  $M_h$  and  $K_h$  and Remark 5 for details on the game interpretation).

One interest of the MFEM is to provide, as in the case of the classical finite element method, a systematic way to compute error estimates, which can be interpreted geometrically as “projection” errors. In the classical finite element method, orthogonal projectors with respect to the energy norm must be used. In the max-plus case, projectors on semimodules must be used (note that these projectors minimize an additive analogue of Hilbert projective metric [CGQ04]).

We shall see that when the value function is nonsmooth, the space of test functions must be different from the space in which the solution is represented, so that our discretization is indeed a max-plus analogue of the Petrov–Galerkin finite element method. A convenient choice of finite elements and test functions includes quadratic functions (also considered by Fleming and McEneaney [FM00]) and norm-like functions; see section 5.

In the MFEM, we need to compute the value of the max-plus scalar product  $\langle z \mid S^\delta w \rangle$  for each finite element  $w$  and each test function  $z$ . In some special cases,  $\langle z \mid S^\delta w \rangle$  can be computed analytically. In general, we need to approximate this scalar product. Here we consider the approximation  $S^\delta w(x) = w(x) + \delta H(x, \nabla w(x))$ , for  $x \in X$ , which is also used in [MH99]. Our main result, Theorem 22, provides for the resulting discretization of the value function an error estimate of order  $\delta + \Delta x(\delta)^{-1}$ , where  $\Delta x$  is the “space discretization step,” under classical assumptions on the control problem and the additional assumption that the value function  $v^t$  is semiconvex for all  $t \in [0, T]$ . This is comparable with the order obtained in the simplest discrete dynamic programming method; see [CDI84], [Fal87a], [Fal87b], [CDF89]. To avoid solving a difficult (nonconvex) optimization problem, we propose a further approximation of the max-plus scalar product  $\langle z \mid S^\delta w \rangle$ , for which we obtain an error estimate of order  $\sqrt{\delta} + \Delta x(\delta)^{-1}$ , which is yet comparable to the order of the existing discretization methods [CDI84], [Fal87a], [Fal87b], [CDF89], [CL84].

Note that the discretization grid need not be regular: in Theorem 22,  $\Delta x$  is defined for an arbitrary grid in terms of Voronoi tessellations.

The paper is organized as follows. In section 2, we recall some basic tools and notions: residuation, semimodules, and projection. In section 3, we present the formulation of the max-plus finite element method. In section 4, we compare our method

with the method proposed by Fleming and McEneaney in [FM00]. In section 5, we state an error estimate and give the main convergence theorem. Finally, in section 6, we illustrate the method by numerical examples in dimensions 1 and 2. Preliminary results of this paper appeared in [AGL04].

**2. Preliminaries on residuation and projections over semimodules.** In this section we recall some classical residuation results (see, for example, [DJLC53], [Bir67], [BJ72], [BCOQ92]) and their application to linear maps on idempotent semimodules (see [LMS01], [CGQ04]). We also review some results of [CGQ96], [CGQ04] concerning projectors over semimodules. Other results on projectors over semimodules appeared in [Gon96], [GM01].

**2.1. Residuation, semimodules, and linear maps.** If  $(S, \leq)$  and  $(T, \leq)$  are (partially) ordered sets, we say that a map  $f : S \rightarrow T$  is *monotone* if  $s \leq s' \implies f(s) \leq f(s')$ . We say that  $f$  is *residuated* if there exists a map  $f^\sharp : T \rightarrow S$  such that

$$(5) \quad f(s) \leq t \iff s \leq f^\sharp(t).$$

The map  $f$  is residuated if and only if, for all  $t \in T$ ,  $\{s \in S \mid f(s) \leq t\}$  has a maximum element in  $S$ . Then

$$f^\sharp(t) = \max\{s \in S \mid f(s) \leq t\} \quad \forall t \in T.$$

Moreover, in that case, we have

$$(6) \quad f \circ f^\sharp \circ f = f \quad \text{and} \quad f^\sharp \circ f \circ f^\sharp = f^\sharp.$$

In what follows, we shall consider situations where  $S$  (or  $T$ ) is equipped with an idempotent monoid law  $\oplus$  (*idempotent* means that  $a \oplus a = a$ ). Then the *natural order* on  $S$  is defined by  $a \leq b \iff a \oplus b = b$ . The supremum law for the natural order, which is denoted by  $\bigvee$ , coincides with  $\oplus$ , and the infimum law for the natural order, when it exists, will be denoted by  $\bigwedge$ . We say that  $S$  is *complete* as a naturally ordered set if any subset of  $S$  has a least upper bound for the natural order. When  $S$  and  $T$  are complete, it is known that the map  $f : S \rightarrow T$  is residuated if and only if it preserves arbitrary sups [BCOQ92, Theorem 4.50].

If  $\mathcal{K}$  is an idempotent semiring, i.e., a semiring whose addition is idempotent, we say that the semiring  $\mathcal{K}$  is *complete* if it is complete as a naturally ordered set and if the left and right multiplications  $\mathcal{K} \rightarrow \mathcal{K}$ ,  $x \mapsto ax$ , and  $x \mapsto xa$  are residuated. Here and in what follows, semiring multiplication is denoted by concatenation.

The max-plus semiring,  $\mathbb{R}_{\max} = (\mathbb{R} \cup \{-\infty\}, \max, +)$ , defined in the introduction, is an idempotent semiring. It is not complete, but it can be embedded into the complete idempotent semiring  $\overline{\mathbb{R}}_{\max}$  obtained by adjoining  $+\infty$  to  $\mathbb{R}_{\max}$ , with the convention that  $-\infty$  is absorbing for the multiplication. The map  $x \mapsto -x$  from  $\overline{\mathbb{R}}$  to itself yields an isomorphism from  $\overline{\mathbb{R}}_{\max}$  to the complete idempotent semiring  $\overline{\mathbb{R}}_{\min}$ , obtained by replacing  $\max$  by  $\min$  and by exchanging the roles of  $+\infty$  and  $-\infty$  in the definition of  $\overline{\mathbb{R}}_{\max}$ .

Semimodules over semirings are defined like modules over rings, mutatis mutandis; see [LMS01], [CGQ04]. When  $\mathcal{K}$  is a complete idempotent semiring, we say that a (right)  $\mathcal{K}$ -semimodule  $\mathcal{X}$  is *complete* if it is complete as an idempotent monoid and if, for all  $u \in \mathcal{X}$  and  $\lambda \in \mathcal{K}$ , the right and left multiplications,  $R_\lambda^\mathcal{X} : \mathcal{X} \rightarrow \mathcal{X}$ ,  $v \mapsto v\lambda$  and  $L_u^\mathcal{X} : \mathcal{K} \rightarrow \mathcal{X}$ ,  $\mu \mapsto u\mu$ , are residuated (for the natural order). In a complete semimodule  $\mathcal{X}$ , we define, for all  $u, v \in \mathcal{X}$ ,

$$u \setminus v \stackrel{\text{def}}{=} (L_u^\mathcal{X})^\sharp(v) = \max\{\lambda \in \mathcal{K} \mid u\lambda \leq v\}.$$

We shall use *semimodules of functions*: when  $X$  is a set and  $\mathcal{K}$  is a complete idempotent semiring, the set of functions  $\mathcal{K}^X$  is a complete  $\mathcal{K}$ -semimodule for the componentwise addition  $(u, v) \mapsto u \oplus v$  (defined by  $(u \oplus v)(x) = u(x) \oplus v(x)$ ) and the componentwise multiplication  $(\lambda, u) \mapsto u\lambda$  (defined by  $(u\lambda)(x) = u(x)\lambda$ ). In fact, we shall need only the case where  $\mathcal{K} = \overline{\mathbb{R}}_{\max}$  or  $\overline{\mathbb{R}}_{\min}$  when applying these notions. In particular, when  $\mathcal{K} = \overline{\mathbb{R}}_{\max}$ ,  $\mathcal{K}^X$  is the set of functions from  $X$  to  $\mathbb{R} \cup \{\pm\infty\}$ , equipped with the pointwise supremum and with the action  $(\lambda, u) \mapsto \lambda + u$ , where  $(\lambda + u)(x) = \lambda + u(x)$ , for all  $x \in X$ , again with the convention that  $(-\infty) + \infty = -\infty$ .

If  $\mathcal{K}$  is an idempotent semiring, and if  $\mathcal{X}$  and  $\mathcal{Y}$  are  $\mathcal{K}$ -semimodules, we say that a map  $A : \mathcal{X} \rightarrow \mathcal{Y}$  is *linear*, or is a *linear operator*, if for all  $u, v \in \mathcal{X}$  and  $\lambda, \mu \in \mathcal{K}$ ,  $A(u\lambda \oplus v\mu) = A(u)\lambda \oplus A(v)\mu$ . Then, as in classical algebra, we use the notation  $Au$  instead of  $A(u)$ . When  $A$  is residuated and  $v \in \mathcal{Y}$ , we use the notation  $A \setminus v$  or  $A^\sharp v$  instead of  $A^\sharp(v)$ . So

$$A \setminus v = A^\sharp v = \max\{w \in \mathcal{X} \mid Aw \leq v\}.$$

We denote by  $L(\mathcal{X}, \mathcal{Y})$  the set of linear operators from  $\mathcal{X}$  to  $\mathcal{Y}$  which is a complete semimodule as soon as  $\mathcal{K}$  and  $\mathcal{Y}$  are complete. If  $\mathcal{K}$  is a complete idempotent semiring, if  $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$  are complete  $\mathcal{K}$ -semimodules, and if  $A \in L(\mathcal{X}, \mathcal{Y})$  is residuated, then the map  $L_A : L(\mathcal{Z}, \mathcal{X}) \rightarrow L(\mathcal{Z}, \mathcal{Y})$ ,  $B \mapsto A \circ B$  is residuated and we set  $A \setminus C := (L_A)^\sharp(C)$  for all  $C \in L(\mathcal{Z}, \mathcal{Y})$ .

If  $X$  and  $Y$  are two sets,  $\mathcal{K}$  is a complete idempotent semiring, and  $a \in \mathcal{K}^{X \times Y}$ , we construct the linear operator  $A$  from  $\mathcal{K}^Y$  to  $\mathcal{K}^X$  which associates with any  $u \in \mathcal{K}^Y$  the function  $Au \in \mathcal{K}^X$  such that  $Au(x) = \bigvee_{y \in Y} a(x, y)u(y)$ . We say that  $A$  is the *kernel operator* with *kernel* or *matrix*  $a$ . We shall often use the same notation  $A$  for the operator and the kernel (so  $A(x, y) = a(x, y)$ ). In particular, when  $\mathcal{K} = \overline{\mathbb{R}}_{\max}$ , we have

$$(7) \quad Au(x) = \sup_{y \in Y} (A(x, y) + u(y)).$$

As is well known (see, for instance, [BCOQ92]), the kernel operator  $A$  is residuated, and

$$(A \setminus v)(y) = \bigwedge_{x \in X} A(x, y) \setminus v(x).$$

When  $\mathcal{K} = \overline{\mathbb{R}}_{\max}$ ,  $A \setminus v$  is given by

$$(8) \quad (A \setminus v)(y) = \inf_{x \in X} (-A(x, y) + v(x)) = [-A^*(-v)](y),$$

where  $A^*$  denotes the *transposed operator*  $\mathcal{K}^X \rightarrow \mathcal{K}^Y$ , which is associated with the kernel  $A^*(y, x) = A(x, y)$ . (In (8), we use the convention that  $+\infty$  is absorbing for addition.)

**2.2. Projectors on semimodules.** Let  $\mathcal{K}$  be a complete idempotent semiring and  $\mathcal{V}$  denote a *complete subsemimodule* of a complete semimodule  $\mathcal{X}$ , i.e., a subset of  $\mathcal{X}$  that is stable by arbitrary sups and by the action of scalars. We call *canonical projector* on  $\mathcal{V}$  the map

$$(9) \quad P_{\mathcal{V}} : \mathcal{X} \rightarrow \mathcal{X}, \quad u \mapsto P_{\mathcal{V}}(u) = \max\{v \in \mathcal{V} \mid v \leq u\}.$$

Let  $W$  denote a *generating family* of a complete subsemimodule  $\mathcal{V}$ , which means that any element  $v \in \mathcal{V}$  can be written as  $v = \bigvee \{w\lambda_w \mid w \in W\}$  for some  $\lambda_w \in \mathcal{K}$ . It is known that

$$P_{\mathcal{V}}(u) = \bigvee_{w \in W} w(w \setminus u)$$

(see, for instance, [CGQ04]). If  $B : \mathcal{U} \rightarrow \mathcal{X}$  is a residuated linear operator, then when  $\mathcal{U}$  and  $\mathcal{X}$  are complete semimodules over  $\mathcal{K}$ , the image  $\text{im } B$  of  $B$  is a complete subsemimodule of  $\mathcal{X}$ , and

$$(10) \quad P_{\text{im } B} = B \circ B^{\sharp}.$$

The max-plus finite element methods relies on the notion of projection on an image, parallel to a kernel, which was introduced by Cohen, Gaubert, and Quadrat in [CGQ96]. The following theorem, of which Corollary 3 below is an immediate corollary, is a variation on the results of [CGQ96, Section 6].

**THEOREM 1** (projection on an image parallel to a kernel). *Let  $\mathcal{U}$ ,  $\mathcal{X}$ , and  $\mathcal{Y}$  be complete semimodules over  $\mathcal{K}$ . Let  $B : \mathcal{U} \rightarrow \mathcal{X}$  and  $C : \mathcal{X} \rightarrow \mathcal{Y}$  be two residuated linear operators over  $\mathcal{K}$ . Let  $\Pi_B^C = B \circ (C \circ B)^{\sharp} \circ C$ . We have  $\Pi_B^C = \Pi_B \circ \Pi^C$ , where  $\Pi_B = B \circ B^{\sharp}$  and  $\Pi^C = C^{\sharp} \circ C$ . Moreover,  $\Pi_B^C$  is a projector, meaning that  $(\Pi_B^C)^2 = \Pi_B^C$ , and for all  $x \in \mathcal{X}$*

$$\Pi_B^C(x) = \max\{y \in \text{im } B \mid Cy \leq Cx\}.$$

*Proof.* The first assertion follows from  $(C \circ B)^{\sharp} = B^{\sharp} \circ C^{\sharp}$ . For the second assertion, we have

$$\begin{aligned} (\Pi_B^C)^2 &= (B \circ (C \circ B)^{\sharp} \circ C) \circ (B \circ (C \circ B)^{\sharp} \circ C) \\ &= B \circ (C \circ B)^{\sharp} \circ C \quad (\text{using (6)}) \\ &= \Pi_B^C. \end{aligned}$$

To prove the last assertion, we use that  $\Pi_B = P_{\text{im } B}$  and (5), and we deduce that

$$\begin{aligned} \Pi_B^C(x) &= P_{\text{im } B} \circ C^{\sharp} \circ C(x) \\ &= \max\{y \in \text{im } B \mid y \leq C^{\sharp} \circ C(x)\} \\ &= \max\{y \in \text{im } B \mid Cy \leq Cx\}. \quad \square \end{aligned}$$

The results of [CGQ96] characterize the existence and uniqueness, for all  $x \in X$ , of  $y \in \text{im } B$  such that  $Cy = Cx$ . In that case,  $y = \Pi_B^C(x)$ .

When  $\mathcal{K} = \overline{\mathbb{R}}_{\max}$ , and  $C : \overline{\mathbb{R}}_{\max}^X \rightarrow \overline{\mathbb{R}}_{\max}^Y$  is a kernel operator,  $\Pi^C = C^{\sharp} \circ C$  has an interpretation similar to (10):

$$\Pi^C(v) = C^{\sharp} \circ C(v) = -P_{\text{im } C^*}(-v) = P^{-\text{im } C^*}(v),$$

where  $-\text{im } C^*$  is thought of as a  $\overline{\mathbb{R}}_{\min}$ -subsemimodule of  $\overline{\mathbb{R}}_{\min}^X$  and  $P^{\mathcal{V}}$  denotes the projector on a  $\overline{\mathbb{R}}_{\min}$ -semimodule  $\mathcal{V}$ , so that

$$P^{-\text{im } C^*}(v) = \min\{w \in -\text{im } C^* \mid w \geq v\},$$

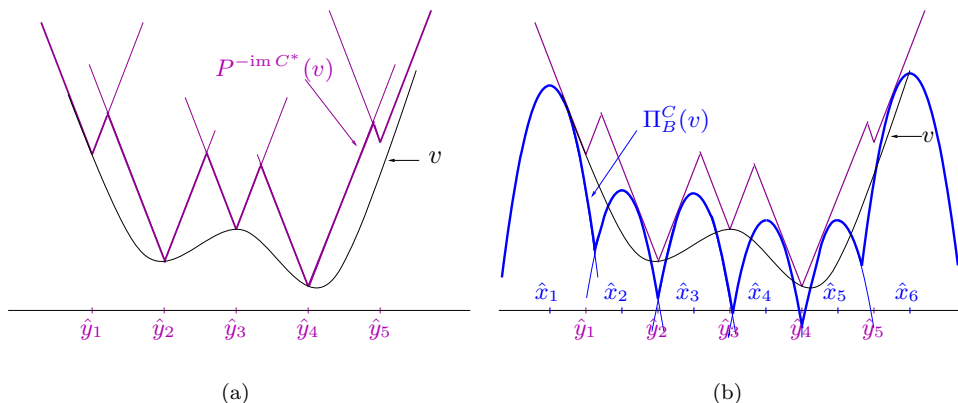


FIG. 1. Example illustrating max-plus and min-plus projectors.

where  $\leq$  denotes here the usual order on  $\overline{\mathbb{R}}^X$ . When  $B : \overline{\mathbb{R}}_{\max}^U \rightarrow \overline{\mathbb{R}}_{\max}^X$  is also a kernel operator, we have

$$\Pi_B^C = P_{\text{im } B} \circ P^{-\text{im } C^*}.$$

This factorization will be instrumental in the geometrical interpretation of the finite element algorithm.

*Example 2.* We take  $U = \{1, \dots, p\}$ ,  $X = \mathbb{R}$ , and  $Y = \{1, \dots, q\}$ . Consider the linear operators  $B : \overline{\mathbb{R}}_{\max}^U \rightarrow \overline{\mathbb{R}}_{\max}^X$  and  $C : \overline{\mathbb{R}}_{\max}^X \rightarrow \overline{\mathbb{R}}_{\max}^Y$  such that

$$B\lambda(x) = \sup_{1 \leq i \leq p} \left\{ -\frac{c}{2}(x - \hat{x}_i)^2 + \lambda_i \right\} \quad \forall \lambda \in \overline{\mathbb{R}}_{\max}^U$$

and

$$(Cf)_i = \sup_{x \in \mathbb{R}} \{-a|x - \hat{y}_i| + f(x)\} \quad \forall f \in \overline{\mathbb{R}}_{\max}^X.$$

The image of  $B$ ,  $\text{im } B$ , is the semimodule generated in the max-plus sense by the functions  $x \mapsto -\frac{c}{2}(x - \hat{x}_i)^2$  for  $i = 1, \dots, p$ . We have

$$C^\sharp \mu(x) = \inf_{1 \leq i \leq q} \{a|x - \hat{y}_i| + \mu_i\} \quad \forall \mu \in \overline{\mathbb{R}}_{\max}^Y,$$

and the image of  $C^\sharp$ , which coincides with  $-\text{im } C^*$ , is the semimodule generated in the min-plus sense by the functions  $x \mapsto a|x - \hat{y}_i|$  for  $i = 1, \dots, q$ .

In Figure 1(a), we represent a function  $v$  and its projection  $P^{-\text{im } C^*}(v)$  (in bold). In Figure 1(b), we represent (in bold) the projection  $P_{\text{im } B}(P^{-\text{im } C^*}(v)) = \Pi_B^C(v)$ .

### 3. The max-plus finite element method.

**3.1. Max-plus variational formulation.** We now describe the max-plus finite element method to solve problem (1). Let  $\mathcal{V}$  be a complete semimodule of functions from  $X$  to  $\overline{\mathbb{R}}_{\max}$ . Let  $S^t : \mathcal{V} \rightarrow \mathcal{V}$  and  $v^t$  be defined as in the introduction. Using the semigroup property  $S^{t+t'} = S^t \circ S^{t'}$ , for  $t, t' > 0$ , we get

$$(11) \quad v^{t+\delta} = S^\delta v^t, \quad t = 0, \delta, \dots, T - \delta,$$

with  $v^0 = \phi$  and  $\delta = \frac{T}{N}$  for some positive integer  $N$ . Let  $\mathcal{W} \subset \mathcal{V}$  be a complete  $\overline{\mathbb{R}}_{\max}$ -semimodule of functions from  $X$  to  $\overline{\mathbb{R}}_{\max}$  such that for all  $t \geq 0$ ,  $v^t \in \mathcal{W}$ . We choose a “dual” semimodule  $\mathcal{Z}$  of “test functions” from  $X$  to  $\overline{\mathbb{R}}_{\max}$ . Recall that the max-plus scalar product is defined by

$$\langle u | v \rangle = \sup_{x \in X} u(x) + v(x)$$

for all functions  $u, v : X \rightarrow \overline{\mathbb{R}}_{\max}$ . We replace (11) by

$$(12) \quad \langle z | v^{t+\delta} \rangle = \langle z | S^\delta v^t \rangle \quad \forall z \in \mathcal{Z}$$

for  $t = 0, \delta, \dots, T - \delta$ , with  $v^\delta, \dots, v^T \in \mathcal{W}$ . Equation (12) can be seen as the analogue of a *variational* or *weak formulation*. Kolokoltsov and Maslov used this formulation in [KM88] and [KM97, Section 3.2] to define a notion of generalized solution of Hamilton–Jacobi equations. We use it in the next section to construct an approximation algorithm for the value function, which is obtained by taking for  $\mathcal{W}$  and  $\mathcal{Z}$  finitely generated semimodules (whereas in the work of Kolokoltsov and Maslov, the semimodules  $\mathcal{W}$  and  $\mathcal{Z}$  are “infinite dimensional” spaces consisting, for instance, of continuous or convex functions).

**3.2. Ideal max-plus finite element method.** We consider a semimodule  $\mathcal{W}_h \subset \mathcal{W}$  generated by the family  $\{w_i\}_{1 \leq i \leq p}$ . We call *finite elements* the functions  $w_i$ . We approximate  $v^t$  by  $v_h^t \in \mathcal{W}_h$ , that is,

$$v^t \simeq v_h^t = \bigvee_{1 \leq i \leq p} w_i \lambda_i^t,$$

where  $\lambda_i^t \in \overline{\mathbb{R}}_{\max}$ . We also consider a semimodule  $\mathcal{Z}_h \subset \mathcal{Z}$  with generating family  $\{z_j\}_{1 \leq j \leq q}$ . The functions  $z_1, \dots, z_q$  will act as test functions. We replace (12) by

$$(13) \quad \langle z | v_h^{t+\delta} \rangle = \langle z | S^\delta v_h^t \rangle \quad \forall z \in \mathcal{Z}_h$$

for  $t = 0, \delta, \dots, T - \delta$ , with  $v_h^\delta, \dots, v_h^T \in \mathcal{W}_h$ . The function  $v_h^0$  is a given approximation of  $\phi$ . Since  $\mathcal{Z}_h$  is generated by  $z_1, \dots, z_q$ , (13) is equivalent to

$$(14) \quad \langle z_j | v_h^{t+\delta} \rangle = \langle z_j | S^\delta v_h^t \rangle \quad \forall 1 \leq j \leq q$$

for  $t = 0, \delta, \dots, T - \delta$ , with  $v_h^t \in \mathcal{W}_h$ ,  $t = 0, \delta, \dots, T$ .

Since (14) need not have a solution, we look for its maximal subsolution, i.e., the maximal solution  $v_h^{t+\delta} \in \mathcal{W}_h$  of

$$(15a) \quad \langle z_j | v_h^{t+\delta} \rangle \leq \langle z_j | S^\delta v_h^t \rangle \quad \forall 1 \leq j \leq q.$$

We also take for the approximate value function  $v_h^0$  at time 0 the maximal solution  $v_h^0 \in \mathcal{W}_h$  of

$$(15b) \quad v_h^0 \leq v^0.$$

Let us denote by  $W_h$  the max-plus linear operator from  $\overline{\mathbb{R}}_{\max}^p$  to  $\mathcal{W}$  with matrix  $W_h = \text{col}(w_i)_{1 \leq i \leq p}$  and by  $Z_h^*$  the max-plus linear operator from  $\mathcal{W}$  to  $\overline{\mathbb{R}}_{\max}^q$  whose transposed matrix is  $Z_h = \text{col}(z_j)_{1 \leq j \leq q}$ . This means that  $W_h \lambda = \bigvee_{1 \leq i \leq p} w_i \lambda_i$  for all  $\lambda = (\lambda_i)_{i=1, \dots, p} \in \overline{\mathbb{R}}_{\max}^p$ , and  $(Z_h^* v)_j = \langle z_j | v \rangle$  for all  $v \in \mathcal{W}$  and  $j = 1, \dots, q$ .

Applying Theorem 1 to  $B = W_h$  and  $C = Z_h^*$  and noting that  $\mathcal{W}_h = \text{im } W_h$ , we get the following corollary.

COROLLARY 3. *The maximal solution  $v_h^{t+\delta} \in \mathcal{W}_h$  of (15a) is given by  $v_h^{t+\delta} = S_h^\delta(v_h^t)$ , where*

$$S_h^\delta = \Pi_{W_h}^{Z_h^*} \circ S^\delta.$$

Note that  $\Pi_{W_h}^{Z_h^*} = P_{\mathcal{W}_h} \circ P^{-Z_h}$ . The following proposition provides a recursive equation verified by the vector of coordinates of  $v_h^t$ .

PROPOSITION 4. *Let  $v_h^t \in \mathcal{W}_h$  be the maximal solution of (15) for  $t = 0, \delta, \dots, T$ . Then, for every  $t = 0, \delta, \dots, T$ , there exists a maximal  $\lambda^t \in \mathbb{R}_{\max}^p$  such that  $v_h^t = W_h \lambda^t$ ,  $t = 0, \delta, \dots, T$ , which can be determined recursively from*

$$(16a) \quad \lambda^{t+\delta} = (Z_h^* W_h) \setminus (Z_h^* S^\delta W_h \lambda^t)$$

for  $t = 0, \dots, T - \delta$  with the initial condition

$$(16b) \quad \lambda^0 = W_h \setminus \phi.$$

*Proof.* Since  $v_h^t \in \mathcal{W}_h$ ,  $v_h^t = W_h \lambda^t$  for some  $\lambda^t \in \overline{\mathbb{R}}_{\max}^p$  and the maximal  $\lambda^t$  satisfying this condition is  $\lambda^t = W_h^\#(v_h^t)$  for all  $t = 0, \delta, \dots, T$ . Since  $v_h^0$  is the maximal solution of (15b), then by (9) and (10),  $v_h^0 = P_{\mathcal{W}_h}(\phi) = W_h \circ W_h^\#(\phi)$ , and hence  $\lambda^0 = W_h^\# \circ W_h \circ W_h^\#(\phi) = W_h^\#(\phi)$ . Let  $t = \delta, \dots, T$ . Using Corollary 3, Theorem 1, (6), and the property that  $(f \circ g)^\# = g^\# \circ f^\#$  for all residuated maps  $f$  and  $g$ , we get

$$\begin{aligned} \lambda^{t+\delta} &= W_h^\# \circ \Pi_{W_h}^{Z_h^*} \circ S^\delta(W_h \lambda^t) \\ &= W_h^\# \circ W_h \circ W_h^\# \circ (Z_h^*)^\# \circ Z_h^* \circ S^\delta(W_h \lambda^t) \\ &= W_h^\# \circ (Z_h^*)^\# \circ Z_h^* \circ S^\delta(W_h \lambda^t) \\ &= (Z_h^* W_h)^\#(Z_h^* S^\delta W_h \lambda^t), \end{aligned}$$

which yields (16a).  $\square$

For  $1 \leq i \leq p$  and  $1 \leq j \leq q$ , we define

$$(17) \quad (M_h)_{ji} = \langle z_j \mid w_i \rangle,$$

$$(18) \quad (K_h)_{ji} = \langle z_j \mid S^\delta w_i \rangle,$$

$$(19) \quad = \langle (S^*)^\delta z_j \mid w_i \rangle,$$

where  $S^*$  is the *transposed semigroup* of  $S$ , which is the evolution semigroup associated with the optimal control problem (1) in which the sign of the dynamics is changed. The matrices  $M_h$  and  $K_h$ , which represent, respectively, the max-plus linear operators  $Z_h^* W_h$  and  $Z_h^* S^\delta W_h$ , may be thought of as the max-plus analogues of the mass and stiffness matrices, respectively.

The ideal max-plus finite element method (Algorithm 1) is the algorithm derived from Proposition 4, assuming that the “mass” and “stiffness” matrices  $M_h$  and  $K_h$  are computed by oracles. We shall discuss in the next section the approximations of the matrices  $M_h$  and  $K_h$ , which will allow us to implement the method.

**Algorithm 1** IDEAL MAX-PLUS FINITE ELEMENT METHOD

- 
- 1: Choose the finite elements  $(w_i)_{1 \leq i \leq p}$  and  $(z_j)_{1 \leq j \leq q}$ . Choose the time discretization step  $\delta = \frac{T}{N}$ .
  - 2: Compute the matrix  $M_h$  and the matrix  $K_h$  defined in (17), (18), or (19).
  - 3: Compute  $\lambda^0 = W_h \setminus \phi$  and  $v_h^0 = W_h \lambda^0$ .
  - 4: For  $t = \delta, 2\delta, \dots, T$ , compute  $\lambda^t = M_h \setminus (K_h \lambda^{t-\delta})$  and  $v_h^t = W_h \lambda^t$ .
- 

For the convenience of the reader, we rewrite the elements of Algorithm 1 with the usual notation:

$$(20) \quad (M_h)_{ji} = \sup_{x \in X} (z_j(x) + w_i(x)),$$

$$(21) \quad \begin{aligned} (K_h)_{ji} &= \sup_{x \in X} (z_j(x) + S^\delta w_i(x)), \\ &= \sup_{x \in X} (w_i(x) + (S^*)^\delta z_j(x)). \end{aligned}$$

Equation (16a) may be written explicitly, using (7) and (8), for  $1 \leq i \leq p$ , as

$$\lambda_i^{t+\delta} = \min_{1 \leq j \leq q} \left( - (M_h)_{ji} + \max_{1 \leq k \leq p} ((K_h)_{jk} + \lambda_k^t) \right).$$

*Remark 5.* This recursion may be interpreted as the dynamic programming equation of a deterministic zero-sum two player game, with finite action and state spaces. Here the state space of the game is the finite set  $\{1, \dots, p\}$  (to each finite element corresponds a state of the game). To each test function corresponds one possible action  $j \in \{1, \dots, q\}$  of the first player, and to each finite element corresponds one possible action  $k \in \{1, \dots, p\}$  of the second player. Given these actions at the state  $i \in \{1, \dots, p\}$ , the cost of the first player, which is the reward of the second player, is  $-(M_h)_{ji} + (K_h)_{jk}$ .

Finally, we have, for all  $x \in X$  and  $t = 0, \delta, \dots, T - \delta$ ,

$$v_h^{t+\delta} = \sup_{1 \leq i \leq p} (w_i(x) + \lambda_i^{t+\delta}).$$

*Remark 6.* Since  $v_h^t \in \mathcal{W}_h$  for all  $t = 0, \dots, T$ , the dynamics of  $v_h^t$  can be written as a function of the matrices  $M_h$  and  $K_h$ :

$$(22) \quad v_h^{t+\delta} = W_h \circ M_h^\# \circ K_h \circ W_h^\#(v_h^t).$$

**3.3. Effective max-plus finite element method.** In the ideal max-plus finite element method, we assume that the matrices  $M_h$  and  $K_h$  are exactly known. We shall see in section 5 that for natural choices of finite elements and test functions, computing every entry of the matrix  $M_h$  is equivalent to solving a maximization problem in which the objective function is concave and the feasible set is convex. This problem can be approached by standard optimization methods. When the domain  $X$  has a “simple” shape, for instance when  $X$  is a hypercube, the entries of the matrix  $M_h$  can even be computed analytically. Hence, the assumption that  $M_h$  is accurately known is not a restrictive one. The same is not true for  $K_h$ . Indeed, evaluating every scalar product  $\langle z \mid S^\delta w \rangle$  leads to a new optimal control problem since

$$\langle z \mid S^\delta w \rangle = \max z(\mathbf{x}(0)) + \int_0^\delta \ell(\mathbf{x}(s), \mathbf{u}(s)) ds + w(\mathbf{x}(\delta)),$$



where the maximum is taken over the set of trajectories  $(\mathbf{x}(\cdot), \mathbf{u}(\cdot))$  satisfying (1b). This problem is simpler to approximate than problem (1), because the horizon  $\delta$  is small, and the functions  $z$  and  $w$  have a regularizing effect.

We call “the effective max-plus finite element method” the method obtained by replacing in Algorithm 1 the matrix  $K_h$  by an approximation.

We first discuss the approximation of  $S^\delta w$  for every finite element  $w$ . The Hamilton–Jacobi equation (2a) suggests approximating  $S^\delta w$  by the function  $[S^\delta w]_H$  such that

$$(23) \quad [S^\delta w]_H(x) = w(x) + \delta H(x, \nabla w(x)) \quad \forall x \in X.$$

Let  $[S^\delta W_h]_H$  denote the max-plus linear operator from  $\mathbb{R}_{\max}^p$  to  $\mathcal{W}$  with matrix  $[S^\delta W_h]_H = \text{col}([S^\delta w_i]_H)_{1 \leq i \leq p}$ , which means that

$$[S^\delta W_h]_H \lambda = \bigvee_{1 \leq i \leq p} [S^\delta w_i]_H \lambda_i$$

for all  $\lambda = (\lambda_i)_{1 \leq i \leq p} \in \mathbb{R}_{\max}^p$ . The above approximation of  $S^\delta w$  yields an approximation of the matrix  $K_h$  by the matrix  $K_{H,h} := Z_h^* [S^\delta W_h]_H$ , whose entries are given, for  $1 \leq i \leq p$  and  $1 \leq j \leq q$ , by

$$(24) \quad (K_{H,h})_{ji} = \sup_{x \in X} \left( z_j(x) + w_i(x) + \delta H(x, \nabla w_i(x)) \right).$$

Let  $A_{ji}$  denote the set where the optimum of the function  $x \mapsto z_j(x) + w_i(x)$  is attained. Computing  $(K_{H,h})_{ji}$  in (24) requires solving an optimization problem, which is nothing but a perturbation of the optimization problem associated with the computation of  $(M_h)_{ji}$ . We may exploit this observation by approximating  $K_{H,h}$  by the matrix  $\tilde{K}_{H,h}$  with entries

$$(25) \quad \begin{aligned} (\tilde{K}_{H,h})_{ji} &= \sup_{x \in A_{ji}} \left( z_j(x) + w_i(x) \right) + \sup_{x \in A_{ji}} H(x, \nabla w_i(x)) \\ &= \langle z_j \mid w_i \rangle + \delta \sup_{x \in A_{ji}} H(x, \nabla w_i(x)) \end{aligned}$$

for  $1 \leq i \leq p$  and  $1 \leq j \leq q$ . When  $A_{ji}$  has only one element, (25) yields a convenient approximation of  $K_h$ .

Of course,  $w_i$  must be differentiable for the approximation (23) to make sense. When  $w_i$  is nondifferentiable, but  $z_j$  is differentiable, the dual formula (19) suggests approximating  $(K_h)_{ji}$  by

$$\sup_{x \in X} \left( z_j(x) + \delta H(x, -\nabla z_j(x)) + w_i(x) \right).$$

We may also use the dual formula of (25), where  $\nabla w_i(x)$  is replaced by  $-\nabla z_j(x)$ .

We shall see in section 5 that the approximation (24) yields an error of order  $O(\delta^2)$  and that the further approximation (25) yields a more important error of order  $O(\delta^{\frac{3}{2}})$ .

**4. Comparison with the method of Fleming and McEneaney.** Fleming and McEneaney proposed a max-plus based method [FM00], which also uses a space  $\mathcal{W}_h$  generated by finite elements,  $w_1, \dots, w_p$ , together with the linear formulation (11).

Their method approaches the value function at time  $t$ ,  $v^t$ , by  $W_h\mu^t$ , where  $W_h = \text{col}(w_i)_{1 \leq i \leq p}$  as above, and  $\mu^t$  is defined inductively by

$$(26a) \quad \mu^0 = W_h \setminus \phi,$$

$$(26b) \quad \mu^{t+\delta} = (W_h \setminus (S^\delta W_h)) \mu^t$$

for  $t = 0, \delta, \dots, T - \delta$ . This can be compared with the limit case of our finite element method, in which the space of test functions  $\mathcal{Z}_h$  is the set of all functions. This limit case corresponds to replacing  $Z_h^*$  by the identity operator in (16a), so that

$$(27) \quad \lambda^{t+\delta} = W_h \setminus (S^\delta W_h \lambda^t).$$

**PROPOSITION 7.** *Let  $(\mu^t)$  be the sequence of vectors defined by the algorithm of Fleming and McEneaney, (26); let  $(\lambda^t)$  be the sequence of vectors defined by the max-plus finite element method in the limit case (27); and let  $v^t$  denote the value function at time  $t$ . Then*

$$(28) \quad W_h \mu^t \leq W_h \lambda^t \leq v^t \quad \text{for } t = 0, \delta, \dots, T.$$

*Proof.* We first prove that  $W_h \lambda^t \leq v^t$  for  $t = 0, \delta, \dots, T$ . This can be proved by induction. For  $t = 0$ , we have  $W_h \lambda^0 \leq v^0$  by (15b). We assume that  $W_h \lambda^t \leq v^t$ . Using (27), we have

$$\begin{aligned} W_h \lambda^{t+\delta} &= W_h W_h^\dagger S^\delta (W_h \lambda^t) \\ &= \Pi_{W_h} (S^\delta (W_h \lambda^t)). \end{aligned}$$

Using the monotonicity of the semigroup  $S^\delta$ , we obtain

$$\begin{aligned} W_h \lambda^{t+\delta} &\leq \Pi_{W_h} (S^\delta v^t) \\ &\leq S^\delta v^t \\ &= v^{t+\delta}. \end{aligned}$$

The second inequality is also proved by induction. For  $t = 0$ , we have  $\mu^0 = \lambda^0 = W_h \setminus \Phi$ . Suppose that  $\mu^t \leq \lambda^t$ . By definition of  $W_h \setminus (S^\delta W_h)$ , we have

$$W_h (W_h \setminus S^\delta W_h) \leq S^\delta W_h,$$

and hence

$$\begin{aligned} W_h \mu^{t+\delta} &= W_h (W_h \setminus S^\delta W_h) \mu^t \\ &\leq (S^\delta W_h) \mu^t \\ &\leq S^\delta W_h \lambda^t. \end{aligned}$$

Since

$$\begin{aligned} \lambda^{t+\delta} &= W_h \setminus (S^\delta W_h \lambda^t) \\ &= \max\{\lambda \in \overline{\mathbb{R}}_{\max}^p \mid W_h \lambda \leq S^\delta W_h \lambda^t\}, \end{aligned}$$

we get that  $\mu^{t+\delta} \leq \lambda^{t+\delta}$ . Then  $\mu^t \leq \lambda^t$  for  $t = 0, \delta, \dots, T$ . Since  $W_h$  is monotone, we deduce (28).  $\square$

An approximation of (26b) using formulae of the same type as (23) is also discussed in [MH99].

## 5. Error analysis.

**5.1. General error estimates.** In what follows we denote by  $\|v\|_\infty = \sup_{i \in I} |v(i)| \in \mathbb{R} \cup \{+\infty\}$  the sup-norm of any function  $v : I \rightarrow \mathbb{R}$ . We also use the same notation  $\|v\|_\infty = \sup_{i \in I} |v_i|$  for a vector  $v = (v_i)_{i \in I}$ . For any two sets  $I$  and  $J$ , a map  $\Phi : \mathbb{R}^I \rightarrow \mathbb{R}^J$  is said to be monotone and homogeneous if it is monotone for the natural order and if for all  $u \in \mathbb{R}^I$  and  $\lambda \in \mathbb{R}$ ,  $\Phi(u + \lambda) = \Phi(u) + \lambda$  with  $(u + \lambda)(i) = u(i) + \lambda$ . Monotone homogeneous maps are nonexpansive for the sup-norm:  $\|\Phi(u) - \Phi(v)\|_\infty \leq \|u - v\|_\infty$ ; see [CT80]. In particular, max-plus or min-plus linear operators are nonexpansive for the sup-norm. This property will be frequently used in what follows. In order to simplify notation, we denote  $\bar{\tau}_\delta = \{0, \delta, \dots, T\}$ ,  $\tau_\delta^+ = \bar{\tau}_\delta \setminus \{0\}$ , and  $\tau_\delta^- = \bar{\tau}_\delta \setminus \{T\}$ .

*Remark 8.* To establish the main result of the paper (Theorem 22 below), we shall need only to take the norm of finite valued functions. However, we wish to emphasize that all the computations that follow are valid for functions with values in  $\bar{\mathbb{R}}$  if one replaces every occurrence of a term of the form  $\|u - v\|_\infty$  by  $d_\infty(u, v) = \inf\{\lambda \geq 0 \mid -\lambda + v \leq u \leq \lambda + v\}$ . Observe that  $d_\infty(u, v)$  is a semidistance and that  $d_\infty(u, v) = \|u - v\|_\infty$  if  $u - v$  takes finite values. Observe also that if a map  $\Phi : \bar{\mathbb{R}}^I \rightarrow \bar{\mathbb{R}}^J$  is monotone and homogeneous,  $d_\infty(\Phi(u), \Phi(v)) \leq d_\infty(u, v)$  for all  $u, v \in \bar{\mathbb{R}}^I$ .

The following lemma shows that the error of the ideal max-plus finite element method is controlled by the projection errors  $\|\Pi_{W_h}^{Z_h^*}(v^t) - v^t\|_\infty$ . This lemma may be thought of as an analogue of Cea's lemma in the classical analysis of the errors of the finite element method. Projectors over semimodules in the MFEM correspond to orthogonal projectors in the classical finite element method.

**LEMMA 9.** *For  $t \in \bar{\tau}_\delta$ , let  $v^t$  be the value function at time  $t$  and  $v_h^t$  be its approximation given by the ideal max-plus finite element method. We have*

$$(29) \quad \|v_h^T - v^T\|_\infty \leq \|\Pi_{W_h}(v^0) - v^0\|_\infty + \sum_{t \in \tau_\delta^+} \|\Pi_{W_h}^{Z_h^*}(v^t) - v^t\|_\infty.$$

*Proof.* For all  $t \in \tau_\delta^-$ , we have

$$\begin{aligned} \|v_h^{t+\delta} - v^{t+\delta}\|_\infty &\leq \|v_h^{t+\delta} - S_h^\delta(v^t)\|_\infty + \|S_h^\delta(v^t) - v^{t+\delta}\|_\infty \\ &\leq \|S_h^\delta(v_h^t) - S_h^\delta(v^t)\|_\infty + \|\Pi_{W_h}^{Z_h^*} \circ S_h^\delta(v^t) - v^{t+\delta}\|_\infty. \end{aligned}$$

Since  $S_h^\delta$  is a nonexpansive operator, we deduce that

$$\|v_h^{t+\delta} - v^{t+\delta}\|_\infty \leq \|v_h^t - v^t\|_\infty + \|\Pi_{W_h}^{Z_h^*}(v^{t+\delta}) - v^{t+\delta}\|_\infty.$$

The result is obtained by induction on  $t$ , using the fact that  $v_h^0 = P_{W_h}(v^0) = \Pi_{W_h}(v^0)$ .  $\square$

To obtain an error estimate, we need to bound  $\|\Pi_{W_h}^{Z_h^*}(v^t) - v^t\|_\infty$  for all  $t \in \tau_\delta^+$ . Since  $\Pi_{W_h}^{Z_h^*} = \Pi_{W_h} \circ \Pi^{Z_h^*}$ , we have

$$\begin{aligned} \|\Pi_{W_h}^{Z_h^*}(v^t) - v^t\|_\infty &= \|\Pi_{W_h} \circ \Pi^{Z_h^*}(v^t) - v^t\|_\infty \\ &\leq \|\Pi_{W_h} \circ \Pi^{Z_h^*}(v^t) - \Pi_{W_h}(v^t)\|_\infty + \|\Pi_{W_h}(v^t) - v^t\|_\infty, \end{aligned}$$

and since  $\Pi_{W_h}$  is a nonexpansive operator, we get

$$(30) \quad \|\Pi_{W_h}^{Z_h^*}(v^t) - v^t\|_\infty \leq \|\Pi^{Z_h^*}(v^t) - v^t\|_\infty + \|\Pi_{W_h}(v^t) - v^t\|_\infty.$$

Using this inequality together with Lemma 9, we deduce the following corollary.

**COROLLARY 10.** *For  $t \in \bar{\tau}_\delta$ , let  $v^t$  be the value function at time  $t$  and  $v_h^t$  be its approximation given by the ideal max-plus finite element method. We have*

$$\|v_h^T - v^T\|_\infty \leq \left(1 + \frac{T}{\delta}\right) \left(\sup_{t \in \bar{\tau}_\delta} (\|\Pi^{Z_h^*}(v^t) - v^t\|_\infty + \|\Pi_{W_h}(v^t) - v^t\|_\infty)\right).$$

The following general lemma shows that the error of the effective finite element method is controlled by the projection errors and the errors resulting from the approximation of the matrix  $K_h$  by a matrix  $\tilde{K}_h$ .

**LEMMA 11.** *For  $t \in \bar{\tau}_\delta$ , let  $v^t$  be the value function at time  $t$  and  $v_h^t$  be its approximation given by the effective max-plus finite element method, where  $K_h$  is approximated by  $\tilde{K}_h$ . We have*

$$\begin{aligned} \|v_h^T - v^T\|_\infty &\leq \left(1 + \frac{T}{\delta}\right) \left(\sup_{t \in \bar{\tau}_\delta} (\|\Pi^{Z_h^*}(v^t) - v^t\|_\infty + \|\Pi_{W_h}(v^t) - v^t\|_\infty) \right. \\ &\quad \left. + \|\tilde{K}_h - K_h\|_\infty\right). \end{aligned}$$

*Proof.* Since  $v_h^t$  is computed with the approximation  $\tilde{K}_h$  of  $K_h$ , we have  $v_h^t = W_h \lambda^t$ ,  $t \in \bar{\tau}_\delta$ , with

$$\lambda^{t+\delta} = M_h^\# \circ (\tilde{K}_h \lambda^t) = W_h^\# \circ (Z_h^*)^\# \circ (\tilde{K}_h \lambda^t).$$

We have

$$\begin{aligned} \|v_h^{t+\delta} - v^{t+\delta}\|_\infty &\leq \|v_h^{t+\delta} - S_h^\delta v_h^t\|_\infty + \|S_h^\delta v_h^t - S_h^\delta v^t\|_\infty + \|S_h^\delta v^t - v^{t+\delta}\|_\infty \\ &\leq \|\Pi_{W_h} \circ (Z_h^*)^\# \circ (\tilde{K}_h \lambda^t) - \Pi_{W_h} \circ (Z_h^*)^\# \circ Z_h^* \circ S_h^\delta W_h \lambda^t\|_\infty \\ &\quad + \|v_h^t - v^t\|_\infty + \|\Pi_{W_h}^{Z_h^*}(v^{t+\delta}) - v^{t+\delta}\|_\infty \\ &\leq \|\tilde{K}_h \lambda^t - K_h \lambda^t\|_\infty + \|v_h^t - v^t\|_\infty + \|\Pi_{W_h}^{Z_h^*}(v^{t+\delta}) - v^{t+\delta}\|_\infty \\ &\leq \max_{\substack{1 \leq j \leq q \\ 1 \leq i \leq p}} |(\tilde{K}_h)_{ji} - (K_h)_{ji}| + \|v_h^t - v^t\|_\infty + \|\Pi_{W_h}^{Z_h^*}(v^{t+\delta}) - v^{t+\delta}\|_\infty. \end{aligned}$$

We deduce that

$$\|v_h^T - v^T\|_\infty \leq \|\Pi_{W_h}(v^0) - v^0\|_\infty + \sum_{t \in \tau_\delta^+} \left(\|\Pi_{W_h}^{Z_h^*}(v^t) - v^t\|_\infty + \|\tilde{K}_h - K_h\|_\infty\right),$$

and so

$$\begin{aligned} \|v_h^T - v^T\|_\infty &\leq \left(1 + \frac{T}{\delta}\right) \left(\sup_{t \in \bar{\tau}_\delta} (\|\Pi^{Z_h^*}(v^t) - v^t\|_\infty + \|\Pi_{W_h}(v^t) - v^t\|_\infty) \right. \\ &\quad \left. + \|\tilde{K}_h - K_h\|_\infty\right). \quad \square \end{aligned}$$

COROLLARY 12. For  $t \in \bar{\tau}_\delta$ , let  $v^t$  be the value function at time  $t$  and  $v_h^t$  be its approximation given by the effective max-plus finite element method, implemented with the approximation  $K_{H,h}$  of  $K_h$ , given by (24). We have

$$\begin{aligned} \|v_h^T - v^T\|_\infty &\leq \left(1 + \frac{T}{\delta}\right) \left( \sup_{t \in \bar{\tau}_\delta} (\|\Pi^{Z_h^*}(v^t) - v^t\|_\infty + \|\Pi_{W_h}(v^t) - v^t\|_\infty) \right. \\ &\quad \left. + \max_{1 \leq i \leq p} \|[S^\delta w_i]_H - S^\delta w_i\|_\infty \right). \end{aligned}$$

*Proof.* Using the same technique as in the precedent lemma and using that  $K_{H,h} = Z_h^*[S^\delta W_h]_H$  and  $K_h = Z_h^*S^\delta W_h$ , we have

$$\begin{aligned} \|K_{H,h} - K_h\|_\infty &\leq \|[S^\delta W_h]_H - S^\delta W_h\|_\infty \\ (31) \qquad \qquad \qquad &= \max_{1 \leq i \leq p} \|[S^\delta w_i]_H - S^\delta w_i\|_\infty, \end{aligned}$$

which ends the proof.  $\square$

COROLLARY 13. For  $t \in \bar{\tau}_\delta$ , let  $v^t$  be the value function at time  $t$  and  $v_h^t$  be its approximation given by the effective max-plus finite element method, implemented with the approximation  $\tilde{K}_{H,h}$  of  $K_h$ , given by (25). We have

$$\begin{aligned} \|v_h^T - v^T\|_\infty &\leq \left(1 + \frac{T}{\delta}\right) \left( \sup_{t \in \bar{\tau}_\delta} (\|\Pi^{Z_h^*}v^t - v^t\|_\infty + \|\Pi_{W_h}v^t - v^t\|_\infty) \right. \\ &\quad \left. + \max_{1 \leq i \leq p} \|[S^\delta w_i]_H - S^\delta w_i\|_\infty + \|\tilde{K}_{H,h} - K_{H,h}\|_\infty \right). \end{aligned}$$

*Proof.* We use Lemma 11, together with (31) and

$$\|\tilde{K}_{H,h} - K_h\|_\infty \leq \|\tilde{K}_{H,h} - K_{H,h}\|_\infty + \|K_{H,h} - K_h\|_\infty. \quad \square$$

**5.2. Projection errors.** In this section, we estimate the projection errors resulting from different choices of finite elements. Recall that a function  $f$  is *c-semiconvex* if  $f(x) + \frac{c}{2}\|x\|_2^2$ , where  $\|\cdot\|_2$  is the standard euclidean norm of  $\mathbb{R}^n$ , is convex. A function  $f$  is *c-semiconcave* if  $-f$  is *c-semiconvex*. Spaces of semiconvex functions were intensively used in the max-plus based approximation method of Fleming and McEneaney [FM00]; see also [MH98], [MH99], [McE02], [McE03], [McE04], [Fal87a], [Fal87b], [CDI84], [CDF89].

We shall use the following finite elements.

DEFINITION 14 ( $P_1$  finite elements). We call the  $P_1$  finite element or Lipschitz finite element centered at point  $\hat{x} \in X$ , with constant  $a > 0$ , the function  $w(x) = -a\|x - \hat{x}\|_1$ , where  $\|x\|_1 = \sum_{i=1}^n |x_i|$  is the  $l^1$ -norm of  $\mathbb{R}^n$ .

The family of Lipschitz finite elements of constant  $a$  generates, in the max-plus sense, the semimodule of Lipschitz continuous functions from  $X$  to  $\bar{\mathbb{R}}$  of Lipschitz constant  $a$  with respect to  $\|\cdot\|_1$ .

DEFINITION 15 ( $P_2$  finite elements). We call the  $P_2$  finite element or quadratic finite element centered at point  $\hat{x} \in X$ , with Hessian  $c > 0$ , the function  $w(x) = -\frac{c}{2}\|x - \hat{x}\|_2^2$ .

When  $X = \mathbb{R}^n$ , the family of quadratic finite elements with Hessian  $c$  generates, in the max-plus sense, the semimodule of l.s.c. *c*-semiconvex functions with values in  $\bar{\mathbb{R}}$ .

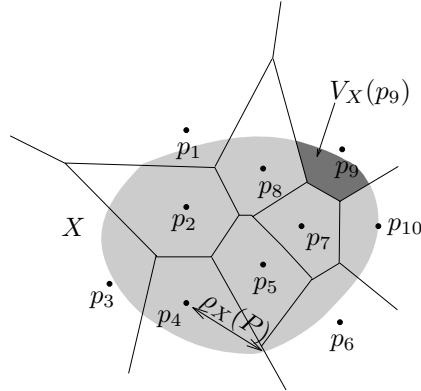


FIG. 2. Voronoi tessellation.

*Notation.* Let  $Y$  be a subset of  $\mathbb{R}^n$  and  $f$  be a function from  $Y$  to  $\overline{\mathbb{R}}$ . We will denote by  $\text{Conv}Y$  the convex hull of  $Y$ ,  $\text{ri}Y$  the relative interior of  $Y$ ,  $\text{dom}f$  the effective domain of  $f$ , and  $\partial f(x)$  the subdifferential of  $f$  at  $x \in \text{dom}f$ .

When  $C$  is a nonempty convex subset of  $\mathbb{R}^n$  and  $c > 0$ , a function is said to be  $c$ -strongly convex on  $C$  if and only if  $f - \frac{1}{2}c\|\cdot\|_2^2$  is convex on  $C$ . A function  $f$  is  $c$ -strongly concave on  $C$  if  $-f$  is  $c$ -strongly convex on  $C$ .

Let  $P$  be a finite subset of  $\mathbb{R}^n$ . The Voronoi cell of a point  $p \in P$  is defined by

$$V(p) = \{x \in \mathbb{R}^n \mid \|x - p\|_2 \leq \|x - q\|_2 \forall q \in P\}.$$

The family  $\{V(p)\}_{p \in P}$  constitutes a subdivision of  $\mathbb{R}^n$ , which is called a Voronoi tessellation (see [SU00] for an introduction to Voronoi tessellations). We define the restriction of  $V(p)$  to  $X$  to be

$$V_X(p) = V(p) \cap X.$$

We define  $\rho_X(P)$  to be the maximal radius of the restriction to  $X$  of the Voronoi cells of the points of  $P$ :

$$\rho_X(P) := \sup_{p \in P} \sup_{x \in V_X(p)} \|x - p\|_2.$$

Observe that

$$\rho_X(P) := \sup_{x \in X} \inf_{p \in P} \|x - p\|_2.$$

The previous definitions are illustrated in Figure 2. The set  $X$  is in light gray,  $P = \{p_1, \dots, p_{10}\}$ ,  $V_X(p_9)$  is in dark gray, and  $\rho_X(P)$  is indicated by a bidirectional arrow.

The next two lemmas bound the projection error in term of the radius of Voronoi cells.

**LEMMA 16** (primal projection error). *Let  $X$  be a compact convex subset of  $\mathbb{R}^n$ . Let  $v : X \rightarrow \mathbb{R}$  be a  $c$ -semiconvex and Lipschitz continuous function with Lipschitz constant  $L_v$  with respect to the euclidean norm. Let  $v_c(x) = v(x) + \frac{c}{2}\|x\|_2^2$ . Let  $\hat{X} = X + B_2(0, \frac{L_v}{c})$ , let  $\hat{X}_h$  be a finite subset of  $\mathbb{R}^n$ , and let  $\mathcal{W}_h$  denote the complete subsemimodule of  $\overline{\mathbb{R}}_{\max}^X$  generated by the family  $(w_{\hat{x}_h})_{\hat{x}_h \in \hat{X}_h}$ , where  $w_{\hat{x}_h}(x) = -\frac{c}{2}\|x - \hat{x}_h\|_2^2$ . Then*

$$\|v - P_{\mathcal{W}_h} v\|_\infty \leq c \text{diam } X \rho_{\hat{X}}(\hat{X}_h).$$

*Proof.* Let  $\mathcal{W}$  denote the complete subsemimodule of  $\overline{\mathbb{R}}_{\max}^X$  generated by the family  $(w_{\hat{x}})_{\hat{x} \in \hat{X}}$ . We will first prove that for all  $x \in X$ ,  $P_{\mathcal{W}}v(x) = v(x)$ . It is obvious that for all  $x \in X$ ,  $v(x) \geq P_{\mathcal{W}}v(x)$ . Using that  $P_{\mathcal{W}} = W \circ W^\sharp$ , with  $W = \text{col}(w_{\hat{x}})_{\hat{x} \in \hat{X}}$ , we obtain

$$\begin{aligned} P_{\mathcal{W}}v(x) &= \sup_{\hat{x} \in \hat{X}} \left( -\frac{c}{2}\|x - \hat{x}\|_2^2 + \inf_{y \in X} \left( \frac{c}{2}\|y - \hat{x}\|_2^2 + v(y) \right) \right) \\ &= \sup_{\hat{x} \in \hat{X}} \left( -\frac{c}{2}\|x\|_2^2 + c\hat{x} \cdot x - \sup_{y \in X} (c\hat{x} \cdot y - v_c(y)) \right) \\ &= -\frac{c}{2}\|x\|_2^2 + \sup_{\hat{x} \in \hat{X}} (c\hat{x} \cdot x - v_c^*(c\hat{x})), \end{aligned}$$

where  $v_c^*$  denotes the Fenchel transform of  $v_c$ . Since  $v_c$  is l.s.c., convex, and proper, we have for all  $x \in X$

$$(32) \quad v_c(x) = v_c^{**}(x) = \sup_{\theta \in \mathbb{R}^n} (\theta \cdot x - v_c^*(\theta)).$$

Using Theorem 23.4 of [Roc70], for all  $x \in \text{ri}(\text{dom}v_c)$ , the subdifferential of  $v_c$  at  $x$ ,  $\partial v_c(x) = \{\theta \in \mathbb{R}^n \mid v_c(y) - v_c(x) \geq \theta \cdot (y - x) \forall y \in X\}$ , is nonempty. Then  $\theta \in \partial v_c(x)$  if and only if  $v_c^*(\theta) = \theta \cdot x - v_c(x)$  and, consequently, the supremum of (32) is attained for all elements  $\theta$  of  $\partial v_c(x)$ .

Set  $q(x) = \frac{c}{2}\|x\|_2^2$ . Using the fact that  $q(y) - q(x) = q'(x) \cdot (y - x) + O(\|y - x\|_2^2)$  and that  $v$  is Lipschitz continuous with Lipschitz constant  $L_v$ , we obtain  $\partial v_c(x) \subset B_2(cx, L_v)$  for all  $x \in \text{ri}X$ . Therefore, for all  $x \in \text{ri}X$ ,

$$(33) \quad v_c(x) = \sup_{\hat{x} \in \hat{X}} (c\hat{x} \cdot x - v_c^*(c\hat{x})).$$

By continuity in the members of (33), we have the equality for all  $x \in X$ , and so

$$\begin{aligned} P_{\mathcal{W}}v(x) &= -\frac{c}{2}\|x\|_2^2 + \sup_{\hat{x} \in \hat{X}} (c\hat{x} \cdot x - v_c^*(c\hat{x})) \\ &= -\frac{c}{2}\|x\|_2^2 + v_c(x) \\ &= v(x) \end{aligned}$$

for all  $x \in X$ .

Now fix  $x \in X$ . For  $\hat{x} \in \hat{X}$ , we set  $\varphi(\hat{x}) = c\hat{x} \cdot x - v_c^*(c\hat{x})$ . Since  $P_{\mathcal{W}_h}v \leq P_{\mathcal{W}}v \leq v$ , we have for all  $x \in X$

$$\begin{aligned} 0 \leq v(x) - P_{\mathcal{W}_h}v(x) &= P_{\mathcal{W}}v(x) - P_{\mathcal{W}_h}v(x) \\ &= \sup_{\hat{x} \in \hat{X}} \varphi(\hat{x}) - \sup_{\hat{x}_h \in \hat{X}_h} \varphi(\hat{x}_h) \\ &= \sup_{\hat{x} \in \hat{X}} \inf_{\hat{x}_h \in \hat{X}_h} \varphi(\hat{x}) - \varphi(\hat{x}_h). \end{aligned}$$

We have  $\partial(-\varphi)(\hat{x}) = -cx + c\partial v_c^*(c\hat{x})$ . Since  $\partial v_c^* \subset X$ , we have  $\partial(-\varphi)(\hat{x}) \subset c(X - x) \subset B_2(0, c \text{diam} X)$ . Hence,  $\varphi$  is Lipschitz continuous with Lipschitz constant

$L_\varphi = c \operatorname{diam} X$ . Then for all  $x \in X$

$$\begin{aligned} v(x) - P_{\mathcal{W}_h} v(x) &\leq \sup_{\hat{x} \in \hat{X}} \inf_{\hat{x}_h \in \hat{X}_h} L_\varphi \|\hat{x} - \hat{x}_h\|_2 \\ &= c \operatorname{diam} X \rho_{\hat{X}}(\hat{X}_h). \quad \square \end{aligned}$$

LEMMA 17 (dual projection error). *Let  $X$  be a bounded subset of  $\mathbb{R}^n$  and  $X_h$  a finite subset of  $\mathbb{R}^n$ . Let  $v : X \rightarrow \mathbb{R}$  be a given Lipschitz continuous function with Lipschitz constant  $L_v$  with respect to the euclidean norm. Let  $\mathcal{Z}_h$  denote the complete semimodule of  $\overline{\mathbb{R}}_{\max}^X$  generated by the  $P_1$  finite elements  $(z_{x_h})_{x_h \in X_h}$  centered at the points of  $X_h$  with constant  $a \geq L_v$ . Then*

$$\|v - P^{-\mathcal{Z}_h} v\|_\infty \leq n(a + L_v) \rho_X(X_h).$$

*Proof.* It is clear that  $P^{-\mathcal{Z}_h} v \geq v$ , and using that  $P^{-\mathcal{Z}_h} = (Z^*)^\# \circ Z^*$ , with  $Z = \operatorname{col}(z_{x_h})_{x_h \in X_h}$ , we obtain

$$P^{-\mathcal{Z}_h} v(x) - v(x) = \inf_{x_h \in X_h} \left( a \|x - x_h\|_1 + \sup_{y \in X} (-a \|y - x_h\|_1 + v(y) - v(x)) \right)$$

for all  $x \in X$ . Since  $v$  is  $L_v$ -Lipschitz continuous, we have

$$\begin{aligned} P^{-\mathcal{Z}_h} v(x) - v(x) &\leq \inf_{x_h \in X_h} \left( a \|x - x_h\|_1 + \sup_{y \in X} (-a \|y - x_h\|_1 + L_v \|y - x\|_2) \right) \\ &\leq \inf_{x_h \in X_h} \left( a \|x - x_h\|_1 + \sup_{y \in X} (-a \|y - x_h\|_1 + L_v \|y - x\|_1) \right) \\ &\leq \inf_{x_h \in X_h} \left( a \|x - x_h\|_1 + \sup_{y \in X} (-a \|y - x_h\|_1 + L_v \|y - x_h\|_1 \right. \\ &\quad \left. + L_v \|x - x_h\|_1) \right) \\ &= \inf_{x_h \in X_h} \left( (a + L_v) \|x - x_h\|_1 + \sup_{y \in X} (L_v - a) \|y - x_h\|_1 \right). \end{aligned}$$

Since  $a \geq L_v$ , we deduce that

$$P^{-\mathcal{Z}_h} v(x) - v(x) \leq (a + L_v) \sup_{x \in X} \inf_{x_h \in X_h} \|x - x_h\|_1 \leq n(a + L_v) \rho_X(X_h). \quad \square$$

**5.3. The approximation errors.** To state an error estimate, we make the following standard assumptions (see [Bar94], for instance):

- (H1)  $f : X \times U \rightarrow \mathbb{R}^n$  is bounded and Lipschitz continuous with respect to  $x$ , meaning that there exist  $L_f > 0$  and  $M_f > 0$  such that

$$\begin{aligned} \|f(x, u) - f(y, u)\|_2 &\leq L_f \|x - y\|_2 & \forall x, y \in X, u \in U, \\ \|f(x, u)\|_2 &\leq M_f & \forall x \in X, u \in U. \end{aligned}$$

- (H2)  $\ell : X \times U \rightarrow \mathbb{R}$  is bounded and Lipschitz continuous with respect to  $x$ , meaning that there exist  $L_\ell > 0$  and  $M_\ell > 0$  such that

$$\begin{aligned} |\ell(x, u) - \ell(y, u)| &\leq L_\ell \|x - y\|_2 & \forall x, y \in X, u \in U, \\ |\ell(x, u)| &\leq M_\ell & \forall x \in X, u \in U. \end{aligned}$$



We shall also need the further assumptions:

- (H3) The domain  $X$  is invariant by the dynamics: for all  $\mathbf{u} : [0, T] \rightarrow U$  and for all  $x \in X$ , the solution  $\mathbf{x}_{\mathbf{u},x}$  of  $\dot{\mathbf{x}}_{\mathbf{u},x}(s) = f(\mathbf{x}_{\mathbf{u},x}(s), u(s))$ ,  $s \geq 0$ , and  $\mathbf{x}_{\mathbf{u},x}(0) = x$  satisfies  $\mathbf{x}_{\mathbf{u},x}(s) \in X$  for all  $s \geq 0$ .
- (H4) The domain  $X$  is invariant by the discretized dynamics in time  $\delta > 0$ : for all  $u \in U$  and for all  $x \in X$ ,  $x + \delta f(x, u) \in X$ .

In the main results, the domain will be also assumed to be convex. Then assumption (H4) implies (H3).

### 5.3.1. Approximation of $S^\delta w$ .

LEMMA 18. *Let  $X$  be a convex subset of  $\mathbb{R}^n$ . We make assumptions (H1), (H2), (H3), and (H4). Let  $w : x \rightarrow \mathbb{R}$  be such that  $w$  is  $C^1$  on a neighborhood of  $X$ , Lipschitz continuous with Lipschitz constant  $L_w$  with respect to the euclidean norm,  $c_1$ -semiconvex, and  $c_2$ -semiconcave. Then there exists  $K_1 > 0$  such that  $\|[S^\delta w]_H - S^\delta w\|_\infty \leq K_1 \delta^2$ , for  $\delta > 0$ , where  $[S^\delta w]_H$  is given by (23).*

*Proof.* We first show that there exists  $K_1 > 0$  such that

$$[S^\delta w]_H(x) - S^\delta w(x) \geq -K_1 \delta^2 \quad \forall x \in X.$$

For all  $x \in X$  and  $u \in U$ , define  $\mathbf{x}_{u,x}$  to be the trajectory such that  $\dot{\mathbf{x}}_{u,x}(s) = f(\mathbf{x}_{u,x}(s), u)$ ,  $s \geq 0$ , and  $\mathbf{x}_{u,x}(0) = x$ . In other words, we apply a constant control  $u$ . From (H3),  $\mathbf{x}_{u,x}(s) \in X$ ,  $s \geq 0$ , for all  $u \in U$  and  $x \in X$ . Hence

$$(S^\delta w)(x) \geq \sup \left\{ \int_0^\delta \ell(\mathbf{x}_{u,x}(s), u) ds + w(\mathbf{x}_{u,x}(\delta)) \mid u \in U \right\}.$$

Since  $\ell$  is Lipschitz continuous and  $f$  is bounded, we have

$$\begin{aligned} \left| \int_0^\delta [\ell(\mathbf{x}_{u,x}(s), u) - \ell(x, u)] ds \right| &\leq L_\ell \int_0^\delta \|\mathbf{x}_{u,x}(s) - x\|_2 ds \\ &\leq L_\ell \int_0^\delta M_f s ds; \end{aligned}$$

then

$$(34) \quad \left| \int_0^\delta [\ell(\mathbf{x}_{u,x}(s), u) - \ell(x, u)] ds \right| \leq \frac{1}{2} L_\ell M_f \delta^2.$$

Therefore,

$$(S^\delta w)(x) \geq -\frac{1}{2} L_\ell M_f \delta^2 + \sup \{ \delta \ell(x, u) + w(\mathbf{x}_{u,x}(\delta)) \mid u \in U \}.$$

Since  $w$  is Lipschitz continuous,  $X$  is invariant by the discretized dynamics in time  $\delta$ , and  $f$  is bounded and Lipschitz continuous, we have

$$\begin{aligned}
 \left| w(\mathbf{x}_{u,x}(\delta)) - w(x + \delta f(x, u)) \right| &\leq L_w \|\mathbf{x}_{u,x}(\delta) - x - \delta f(x, u)\|_2 \\
 &\leq L_w \int_0^\delta \|f(\mathbf{x}_{u,x}(s), u) - f(x, u)\|_2 ds \\
 &\leq L_w \int_0^\delta L_f \|\mathbf{x}_{u,x}(s) - x\|_2 ds \\
 &\leq L_w L_f \int_0^\delta M_f s ds,
 \end{aligned}$$

and so

$$(35) \quad \left| w(\mathbf{x}_{u,x}(\delta)) - w(x + \delta f(x, u)) \right| \leq \frac{1}{2} L_w L_f M_f \delta^2.$$

Moreover, since  $w$  is  $c_1$ -semiconvex, we have

$$(36) \quad w(x + \delta f(x, u)) \geq w(x) + \delta \nabla w(x) \cdot f(x, u) - \frac{c_1}{2} M_f^2 \delta^2.$$

We deduce from (34), (35), and (36) that

$$\begin{aligned}
 (S^\delta w)(x) &\geq -(L_\ell M_f + L_w L_f M_f + c_1 M_f^2) \frac{\delta^2}{2} + w(x) \\
 &\quad + \sup_{u \in U} \{ \delta \ell(x, u) + \delta \nabla w(x) \cdot f(x, u) \} \\
 &\geq -(L_\ell M_f + L_w L_f M_f + c_1 M_f^2) \frac{\delta^2}{2} + w(x) + \delta H(x, \nabla w(x)).
 \end{aligned}$$

This ends the first part of the proof.

We now prove an opposite inequality. For all  $x \in X$  and for all measurable functions  $\mathbf{u} : [0, \delta] \rightarrow U$ , define  $\mathbf{x}_{\mathbf{u},x}$  to be the trajectory such that  $\dot{\mathbf{x}}_{\mathbf{u},x}(s) = f(\mathbf{x}_{\mathbf{u},x}(s), \mathbf{u}(s))$  and  $\mathbf{x}_{\mathbf{u},x}(0) = x$ . From (H3),  $\mathbf{x}_{\mathbf{u},x}(s) \in X$ ,  $s \geq 0$ , for all  $\mathbf{u} : [0, \delta] \rightarrow U$  and  $x \in X$ . Since  $\ell(x, u) \leq H(x, p) - p \cdot f(x, u)$ , for all  $p \in \mathbb{R}^n$ ,  $x \in X$ , and  $u \in U$ , we deduce that

$$\begin{aligned}
 (S^\delta w)(x) &\leq \sup \left\{ \int_0^\delta H(\mathbf{x}_{\mathbf{u},x}(s), \nabla w(x)) ds + w(\mathbf{x}_{\mathbf{u},x}(\delta)) \right. \\
 &\quad \left. - \nabla w(x) \cdot \int_0^\delta f(\mathbf{x}_{\mathbf{u},x}(s), \mathbf{u}(s)) ds \mid \mathbf{u} : [0, \delta] \rightarrow U \right\} \\
 &= \sup \left\{ \int_0^\delta H(\mathbf{x}_{\mathbf{u},x}(s), \nabla w(x)) ds \right. \\
 &\quad \left. + w(\mathbf{x}(\delta)) - \nabla w(x) \cdot (\mathbf{x}_{\mathbf{u},x}(\delta) - x) \mid \mathbf{u} : [0, \delta] \rightarrow U \right\}.
 \end{aligned}$$

Using the fact that  $\ell$  and  $f$  are Lipschitz continuous with respect to  $x$ , we have for all  $x, x' \in X$ ,  $p \in \mathbb{R}^n$

$$\left| H(x, p) - H(x', p) \right| \leq (L_\ell + L_f \|p\|_2) \|x - x'\|_2;$$

therefore

$$\begin{aligned} (S^\delta w)(x) &\leq \sup \left\{ (L_\ell + L_f L_w) \int_0^\delta \|\mathbf{x}_{\mathbf{u},x}(s) - x\|_2 ds + \delta H(x, \nabla w(x)) \right. \\ &\quad \left. + w(\mathbf{x}_{\mathbf{u},x}(\delta)) - \nabla w(x) \cdot (\mathbf{x}_{\mathbf{u},x}(\delta) - x) \mid \mathbf{u} : [0, \delta] \rightarrow U \right\} \\ &\leq (L_\ell + L_f L_w) M_f \frac{\delta^2}{2} + \delta H(x, \nabla w(x)) \\ &\quad + \sup \left\{ w(\mathbf{x}_{\mathbf{u},x}(\delta)) - \nabla w(x) \cdot (\mathbf{x}_{\mathbf{u},x}(\delta) - x) \mid \mathbf{u} : [0, \delta] \rightarrow U \right\}. \end{aligned}$$

Since  $w$  is  $c_2$ -semiconcave, we have

$$w(\mathbf{x}_{\mathbf{u},x}(\delta)) \leq w(x) + \nabla w(x) \cdot (\mathbf{x}_{\mathbf{u},x}(\delta) - x) + \frac{c_2}{2} M_f^2 \delta^2.$$

We obtain

$$(S^\delta w)(x) \leq (L_\ell + L_f L_w + c_2 M_f) M_f \frac{\delta^2}{2} + w(x) + \delta H(x, \nabla w(x)).$$

To end the proof, we take  $K_1 = \frac{1}{2} (L_\ell M_f + L_f L_w M_f + \max(c_1, c_2) M_f^2)$ .  $\square$

### 5.3.2. Approximation of the matrix $K_h$ by the matrix $\tilde{K}_H$ .

LEMMA 19. Let  $X$  be a compact subset of  $\mathbb{R}^n$ . We consider an u.s.c. function  $\varphi : X \rightarrow \mathbb{R}$  and a Lipschitz continuous function  $\psi : X \rightarrow \mathbb{R}$  with Lipschitz constant  $L_\psi$  with respect to a norm  $\|\cdot\|$ . For  $\varepsilon \geq 0$ , we define

$$(37a) \quad F_\varepsilon = \left\{ x \in X \mid \varphi(x) \geq \sup_{x' \in X} \varphi(x') - \varepsilon \right\},$$

$$(37b) \quad g(\varepsilon) = \sup_{x \in F_\varepsilon} d(x, F_0),$$

where  $d(x, F_0) = \inf_{y \in F_0} \|y - x\|$ . We have

$$\left| \sup_{x \in X} (\varphi(x) + \delta \psi(x)) - \left[ \sup_{x \in X} \varphi(x) + \delta \sup_{x \in \arg \max \varphi} \psi(x) \right] \right| \leq L_\psi \delta g(\delta M),$$

where  $M = \sup_{x \in X} \psi(x) - \inf_{x \in X} \psi(x)$ .

*Proof.* Since  $\varphi$  is u.s.c. and  $X$  is compact,  $F_0 = \arg \max \varphi$  and

$$(38) \quad \sup_{x \in X} (\varphi(x) + \delta \psi(x)) \geq \sup_{x \in X} \varphi(x) + \delta \sup_{x \in F_0} \psi(x).$$

For  $\varepsilon > 0$ , we have

$$\sup_{x \in X} (\varphi(x) + \delta \psi(x)) = \max \left[ \sup_{x \in F_\varepsilon} (\varphi(x) + \delta \psi(x)), \sup_{x \in X \setminus F_\varepsilon} (\varphi(x) + \delta \psi(x)) \right].$$

Let  $\varepsilon = \delta(\sup_{x \in X} \psi(x) - \inf_{x \in X} \psi(x)) = M\delta$  (which is finite since  $\psi$  is continuous and  $X$  is compact). We have

$$\begin{aligned} \sup_{x \in X \setminus F_\varepsilon} (\varphi(x) + \delta\psi(x)) &\leq -\varepsilon + \sup_{x \in X} \varphi(x) + \delta \sup_{x \in X} \psi(x) \\ &= \sup_{x \in F_\varepsilon} \varphi(x) + \delta \inf_{x \in X} \psi(x) \\ &\leq \sup_{x \in F_\varepsilon} [\varphi(x) + \delta\psi(x)]. \end{aligned}$$

Therefore,

$$\begin{aligned} \sup_{x \in X} (\varphi(x) + \delta\psi(x)) &= \sup_{x \in F_\varepsilon} (\varphi(x) + \delta\psi(x)) \\ (39) \qquad \qquad \qquad &\leq \sup_{x \in X} \varphi(x) + \delta \sup_{x \in F_\varepsilon} \psi(x). \end{aligned}$$

We deduce from (38) and (39) that

$$0 \leq \sup_{x \in X} (\varphi(x) + \delta\psi(x)) - \left[ \sup_{x \in X} \varphi(x) + \delta \sup_{x \in F_0} \psi(x) \right] \leq \delta \left[ \sup_{x \in F_\varepsilon} \psi(x) - \sup_{x \in F_0} \psi(x) \right].$$

Since  $\psi$  is Lipschitz continuous, we have

$$\begin{aligned} \sup_{x \in F_\varepsilon} \psi(x) - \sup_{x \in F_0} \psi(x) &= \sup_{x \in F_\varepsilon} \inf_{y \in F_0} (\psi(x) - \psi(y)) \\ &\leq \sup_{x \in F_\varepsilon} \inf_{y \in F_0} L_\psi \|x - y\| \\ &= L_\psi \sup_{x \in F_\varepsilon} d(x, F_0) \\ &= L_\psi g(\varepsilon). \quad \square \end{aligned}$$

**COROLLARY 20.** *Let  $X$  be a compact convex subset of  $\mathbb{R}^n$ . We consider an u.s.c. and strongly concave function  $\varphi : X \rightarrow \mathbb{R}$  with modulus  $c > 0$  and a Lipschitz continuous function  $\psi : X \rightarrow \mathbb{R}$  with Lipschitz constant  $L_\psi$  with respect to the euclidean norm. Then the maximum of  $\varphi$  on  $X$  is attained at a unique point  $x_0 \in X$ , i.e.,  $\arg \max_X \varphi = \{x_0\}$  and*

$$\left| \sup_{x \in X} (\varphi(x) + \delta\psi(x)) - (\varphi(x_0) + \delta\psi(x_0)) \right| \leq L_\psi \delta \sqrt{\frac{2\delta M}{c}},$$

where  $M = \sup_{x \in X} \psi(x) - \inf_{x \in X} \psi(x)$ .

*Proof.* Define  $\Phi(x) = \varphi(x_0) - \varphi(x)$  for  $x \in X$  and  $\Phi(x) = +\infty$  elsewhere. We have  $\Phi(x) \geq 0$  for all  $x \in \mathbb{R}^n$  and  $\Phi(x_0) = 0$ . Since  $\Phi$  is l.s.c. and convex on  $\mathbb{R}^n$ , then  $0 \in \partial\Phi(x_0)$ . Moreover,  $\Phi$  is strongly convex with modulus  $c$ . Then, using Theorem 6.1.2 of [HUL93, Chapter VI], we have for all  $x, x' \in X$

$$\Phi(x) \geq \Phi(x') + \langle s \mid x - x' \rangle + \frac{c}{2} \|x - x'\|_2^2 \quad \forall s \in \partial\Phi(x').$$

Taking  $x' = x_0$  and  $s = 0$  we obtain for all  $x \in X$

$$\Phi(x) \geq \frac{c}{2} \|x - x_0\|_2^2,$$

which implies that

$$\varphi(x) \leq \varphi(x_0) - \frac{c}{2} \|x - x_0\|_2^2 \quad \forall x \in X.$$

Using the notation of Lemma 19, we get easily (see also Proposition 4.32 of [BS00]) for all  $x \in F_\varepsilon$ ,  $d(x, F_0) \leq \sqrt{\frac{2\varepsilon}{c}}$ , where  $\varepsilon = \delta(\sup_{x \in X} \psi(x) - \inf_{x \in X} \psi(x))$ .  $\square$

*Remark 21.* To have an error estimate of the approximation of the matrix  $K_{H,h}$  by the matrix  $\tilde{K}_{H,h}$ , we apply Corollary 20 in the case where

$$\varphi(x) = w_i(x) + z_j(x) \quad \text{and} \quad \psi(x) = H(x, \nabla w_i(x))$$

for a suitable choice of the finite elements  $w_i$  and test functions  $z_j$ . Using assumptions (H1) and (H2), we have that, for all  $x \in X$ ,  $|\psi(x)| \leq M_f \|\nabla w\|_\infty + M_\ell$ , where  $\|\nabla w\|_\infty = \|\nabla w\|_2$  and  $\nabla w = (\nabla w_i)_{1 \leq i \leq p}$ . We deduce that

$$\sup \psi - \inf \psi \leq 2(M_f \|\nabla w\|_\infty + M_\ell).$$

Moreover,  $H(\cdot, p)$  and  $H(x, \cdot)$  are Lipschitz continuous with Lipschitz constants  $L_f \|p\|_2 + L_\ell$  and  $M_f$ , respectively. Hence,  $\psi$  is Lipschitz continuous with Lipschitz constant

$$L_\psi = L_f \|\nabla w\|_\infty + L_\ell + M_f \|D^2 w_i\|_\infty.$$

**5.4. Final estimation of the error of the MFEM.** We now state our main convergence result, which holds for quadratic finite elements and Lipschitz test functions.

**THEOREM 22.** *Let  $X$  be a compact convex subset of  $\mathbb{R}^n$  with nonempty interior and  $\hat{X} = X + B_2(0, \frac{L}{c})$ , where  $L > 0$ ,  $c > 0$ . Choose any finite sets of discretization points  $X_h \subset \mathbb{R}^n$  and  $\hat{X}_h \subset \mathbb{R}^n$ . Let*

$$\Delta x = \max(\rho_X(X_h), \rho_{\hat{X}}(\hat{X}_h)).$$

*We make assumptions (H1), (H2), (H3), and (H4) and assume that the value function at time  $t$ ,  $v^t$ , is  $c$ -semiconvex and Lipschitz continuous with constant  $L$  with respect to the euclidean norm for all  $t \geq 0$ . Let us choose quadratic finite elements  $w_{\hat{x}_h}$  of Hessian  $c$ , centered at the points  $\hat{x}_h$  of  $\hat{X}_h$ . Let us choose, as test functions, the Lipschitz finite elements  $z_{x_h}$  with constant  $a \geq L$ , centered at the points  $x_h$  of  $X_h$ . For  $t = 0, \delta, \dots, T$ , let  $v_h^t$  be the approximation of  $v^t$  given by the max-plus finite element method implemented with the approximation  $K_{H,h}$  of  $K_h$  given by (24). Then there exists a constant  $C_1 > 0$  such that*

$$\|v_h^T - v^T\|_\infty \leq C_1 \left( \delta + \frac{\Delta x}{\delta} \right).$$

*When the approximation  $K_{H,h}$  is replaced by  $\tilde{K}_{H,h}$ , given by (25), this inequality becomes*

$$\|v_h^T - v^T\|_\infty \leq C_2 \left( \sqrt{\delta} + \frac{\Delta x}{\delta} \right)$$

*for some constant  $C_2 > 0$ .*

*Proof.* Let  $\mathcal{W}_h$  and  $\mathcal{Z}_h$  denote the complete semimodules of  $\overline{\mathbb{R}}_{\max}^X$  generated by the families  $(w_{\hat{x}_h})_{\hat{x}_h \in \hat{X}_h}$  and  $(z_{x_h})_{x_h \in X_h}$ , respectively. We index the elements of  $\hat{X}_h$  and  $X_h$  by  $\hat{x}_{h_1}, \dots, \hat{x}_{h_p}$  and  $x_{h_1}, \dots, x_{h_q}$ , respectively. Using Corollary 12, we have

$$\begin{aligned} \|v_h^T - v^T\|_\infty &\leq \left(1 + \frac{T}{\delta}\right) \left( \sup_{t \in \bar{\tau}_\delta} (\|P^{-\mathcal{Z}_h}(v^t) - v^t\|_\infty + \|P_{\mathcal{W}_h}(v^t) - v^t\|_\infty) \right. \\ &\quad \left. + \max_{1 \leq i \leq p} \|[S^\delta w_i]_H - S^\delta w_i\|_\infty \right). \end{aligned}$$

To estimate the projection error  $\|P_{\mathcal{W}_h}(v^t) - v^t\|_\infty$ , we apply Lemma 16. We obtain, for  $t \in \bar{\tau}_\delta$ ,  $\|P_{\mathcal{W}_h}(v^t) - v^t\|_\infty \leq c \operatorname{diam} X \Delta x$ . Applying Lemma 17 we obtain, for  $t \in \bar{\tau}_\delta$ ,  $\|P^{-\mathcal{Z}_h}(v^t) - v^t\|_\infty \leq n(a + L)\Delta x$ . Finally, using Lemma 18, we get

$$\|v_h^T - v^T\|_\infty \leq C_1 \left( \delta + \frac{\Delta x}{\delta} \right),$$

where

$$C_1 > (T + 1) \max \left( c \operatorname{diam} X + n(a + L), \frac{M_f}{2} \left( L_\ell + cL_f \left( \operatorname{diam} X + \frac{L}{c} \right) + cM_f \right) \right).$$

To prove the second inequality, we use Corollary 13 together with Remark 21. Using the notation of Corollary 20 and Remark 21, we have  $\sup \psi - \inf \psi \leq 2(M_\ell + M_f c(\operatorname{diam} X + \frac{L}{c}))$  and  $L_\psi = L_\ell + cM_f + L_f c(\operatorname{diam} X + \frac{L}{c})$ . Since  $\varphi = w_i + z_j$  is  $c$ -strongly concave, we deduce that

$$\begin{aligned} |(\tilde{K}_{H,h})_{ji} - (K_{H,h})_{ji}| \\ \leq 2 \left( L_\ell + cM_f + L_f c \left( \operatorname{diam} X + \frac{L}{c} \right) \right) \sqrt{\frac{M_\ell}{c} + M_f \left( \operatorname{diam} X + \frac{L}{c} \right)} \delta \sqrt{\delta} \end{aligned}$$

for  $i = 1, \dots, p$  and  $j = 1, \dots, q$ . Hence, there exists  $C_2 > 0$  such that

$$\|v_h^T - v^T\|_\infty \leq C_2 \left( \sqrt{\delta} + \frac{\Delta x}{\delta} \right)$$

when  $\delta$  is small enough.  $\square$

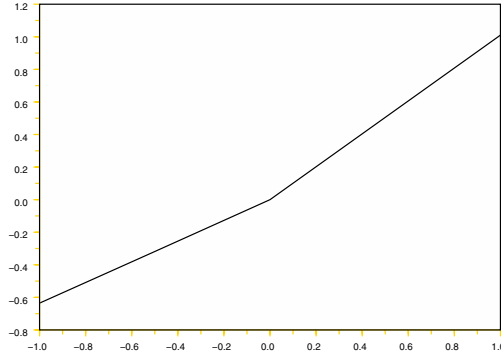
A variant of this theorem, with a stronger assumption, was proved in [Lak03].

*Remark 23.* When  $X_h$  is a rectangular grid of step  $h > 0$ , meaning that  $X_h$  is the intersection of  $(\mathbb{Z}h)^n$  with a Cartesian product of bounded intervals, we have

$$\rho_X(X_h) \leq \sqrt{n}h.$$

Hence, when  $X_h$  and  $\hat{X}_h$  are both rectangular grids of step  $h$ , we have  $\Delta x \leq \sqrt{n}h = O(h)$  in Theorem 22.

**6. Numerical results.** This section presents the results of numerical experiments with the MFEM described in section 3. We consider optimal control problems in dimensions 1 and 2 whose value functions are known or can be computed by solving the Riccati equation (in the case of linear quadratic problems).

FIG. 3. *Max-plus approximation (Example 24).*

**6.1. Implementation.** We implemented the MFEM using the max-plus toolbox of Scilab [Plu98] (in dimension 1) and specific programs written in C (in dimension 2). We used the approximation  $\tilde{K}_{H,h}$  of the matrix  $K_h$ . The matrix  $M_h$  can always be computed analytically. In all the examples below, the Hamiltonian  $H$ , and thus the stiffness matrix  $\tilde{K}_{H,h}$ , have been computed analytically. We avoided storing the (full) matrices  $M_h$  and  $\tilde{K}_{H,h}$  when the number of discretization points was large.

**6.2. Examples in dimension 1.** The next two examples are inspired by those proposed by M. Falcone in [BCD97].

*Example 24.* We consider the case where  $T = 1$ ,  $\phi \equiv 0$ ,  $X = [-1, 1]$ ,  $U = [0, 1]$ ,  $\ell(x, u) = x$ , and  $f(x, u) = -xu$ . Assumptions (H1) and (H2) are satisfied. The optimal choice is to take  $u^* = 0$  whenever  $x > 0$  and to move on the right with maximum speed ( $u^* = 1$ ) whenever  $x \leq 0$ . For all  $t \in [0, T]$ , the value function is

$$v(x, t) = \begin{cases} xt & \text{if } x > 0, \\ x(1 - e^{-t}) & \text{otherwise.} \end{cases}$$

We choose quadratic finite elements  $w_i$  of Hessian  $c$  centered at the points of the regular grid  $(\mathbb{Z}\Delta x) \cap [-2, 2]$  and Lipschitz finite elements  $z_j$  with constant  $a \geq 1$  centered at the points of the regular grid  $(\mathbb{Z}\Delta x) \cap X$ . We represent in Figure 3 the solution given by our algorithm in the case where  $\delta = 0.01$ ,  $\Delta x = 0.005$ ,  $a = 1.5$ , and  $c = 1$ . We obtain an  $L_\infty$ -error of order  $10^{-2}$ .

*Example 25.* We consider the case where  $T = 1$ ,  $\Phi \equiv 0$ ,  $X = [-1, 1]$ ,  $U = [-1, 1]$ ,  $\ell(x, u) = -3(1 - |x|)$ , and  $f(x, u) = u(1 - |x|)$ . It is clear that  $\ell$  and  $f$  are bounded and Lipschitz continuous functions. The optimal choice is to take  $u^* = -1$  whenever  $x > 0$  and  $u^* = 1$  whenever  $x < 0$ . Therefore, all the trajectories lie in  $X$ . For all  $t \in [0, T]$ , the value function is

$$v(x, t) = -3(1 - |x|)(1 - e^{-t}).$$

We choose quadratic finite elements  $w_i$  of Hessian  $c$  and Lipschitz finite elements  $z_j$  with constant  $a$ . We represent in Figure 4 the solution given by our algorithm in the case where  $\delta = 0.02$ ,  $\Delta x = 0.01$ ,  $a = 2$ , and  $c = 8$ . We obtain an  $L_\infty$ -error of order  $7.66 \cdot 10^{-3}$ .

*Example 26* (linear quadratic problem). We consider the case where  $U = \mathbb{R}$ ,  $X = \mathbb{R}$ ,

$$\ell(x, u) = -\frac{1}{2}(x^2 + u^2), \quad f(x, u) = u, \quad \text{and } \phi \equiv 0.$$

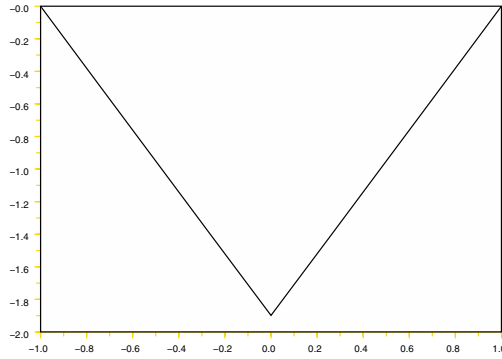


FIG. 4. *Max-plus approximation (Example 25).*

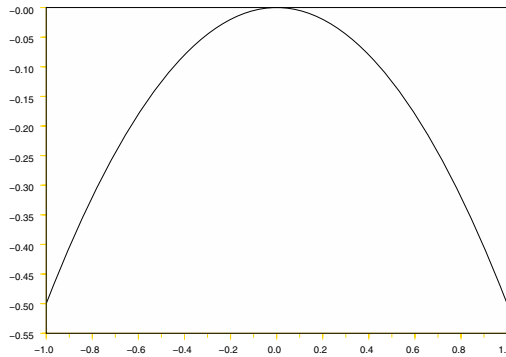


FIG. 5. *Max-plus approximation of a linear quadratic control problem (Example 26).*

The Hamiltonian is  $H(x, p) = -\frac{x^2}{2} + \frac{p^2}{2}$ . This problem can be solved analytically. For  $x \in X$ , the value function at time  $t$  is

$$v(x, t) = -\frac{1}{2} \tanh(t) x^2.$$

The domain  $X$  is unbounded, and  $\ell$  and  $f$  are unbounded and locally Lipschitz continuous. We will restrict  $X$  to the set  $[-5; 5]$  so that  $\ell$  and  $f$  satisfy assumptions (H1) and (H2).

We choose quadratic finite elements  $w_i$  and  $z_j$  of Hessian  $c = 1$ , centered at the points of the regular grid  $(\mathbb{Z}\Delta x) \cap [-6, 6]$ . We represent in Figure 5 the solution given by our algorithm in the interval  $[-1; 1]$  in the case where  $T = 5$ ,  $\delta = 0.5$ ,  $\Delta x = 0.05$ , and  $L = 1$ . We obtain an  $L_\infty$ -error of  $4.54 \cdot 10^{-5}$ .

*Example 27* (distance problem). We consider the case where  $T = 1$ ,  $\phi \equiv 0$ ,  $X = [-1, 1]$ ,  $U = [-1, 1]$ ,

$$\ell(x, u) = \begin{cases} -1 & \text{if } x \in (-1, 1), \\ 0 & \text{if } x \in \{-1, 1\}, \end{cases} \quad \text{and} \quad f(x, u) = \begin{cases} u & \text{if } x \in (-1, 1), \\ 0 & \text{if } x \in \{-1, 1\}. \end{cases}$$

Putting  $\ell = 0$  and  $f = 0$  on  $\partial X$  keeps the trajectories in the domain  $X$ , but we lose the Lipschitz continuity of  $\ell$  and  $f$ . For  $x \in X$ , the value function at time  $t$  of this



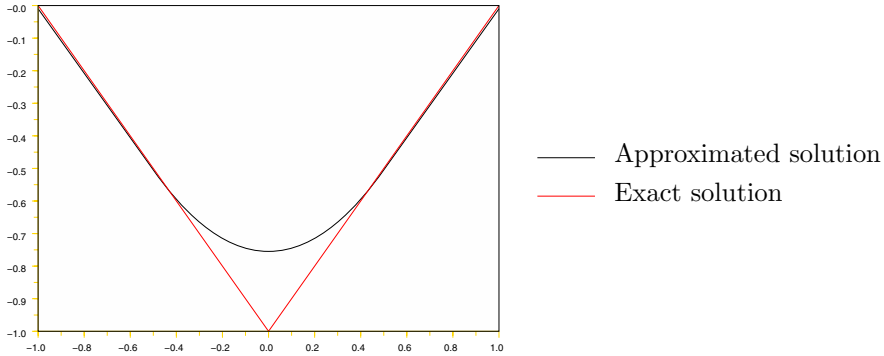


FIG. 6. A bad choice of test functions for the distance problem (Example 27).

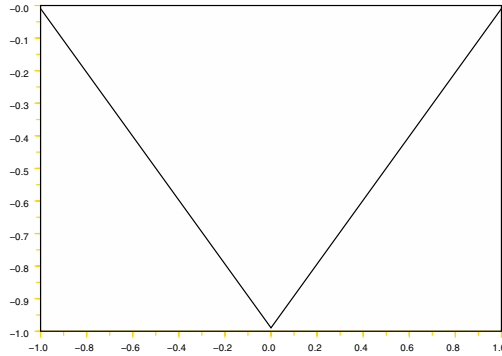


FIG. 7. A good choice of test functions for the distance problem (Example 27).

problem is

$$v(x, t) = \max(-t, |x| - 1).$$

Consider first quadratic finite elements  $w_i$  and  $z_j$  of Hessian  $c$ , centered at the points of the regular grid  $(\mathbb{Z}\Delta x) \cap (X + B_\infty(0, \frac{L}{c}))$ . In Figure 6, we represent the solution given by our algorithm in the case where  $\delta = 0.02$ ,  $\Delta x = 0.01$ ,  $c = 2$ , and  $L = 1$ . Since  $\Pi^{Z_h^*}$  is a projector on a subsemimodule of the  $\bar{\mathbb{R}}_{\min}$ -semimodule of  $c$ -semiconcave functions, and since the solution is not  $c$ -semiconcave for any  $c$ , the error of projection  $\|\Pi^{Z_h^*}(v^t) - v^t\|_\infty$  does not converge to zero when  $\Delta x$  goes to zero, which explains the magnitude of the error.

To solve this problem, it suffices to replace the test functions  $z_j$  by the Lipschitz finite elements with constant  $a \geq 1$ , centered at the points of the regular grid  $(\mathbb{Z}\Delta x) \cap [-1, 1]$ . This is illustrated in Figure 7 in the case where  $\delta = 0.02$ ,  $\Delta x = 0.01$ ,  $c = 2$ , and  $a = 1.1$ . We obtain an  $L_\infty$ -error of  $1.05 \cdot 10^{-2}$ .

### 6.3. Examples in dimension 2.

*Example 28* (linear quadratic problem in dimension 2). We consider the case where  $U = \mathbb{R}^2$ ,  $X = \mathbb{R}^2$ ,  $\phi \equiv 0$ ,

$$\ell(x, u) = -\frac{x_1^2 + x_2^2}{2} - \frac{u_1^2 + u_2^2}{2}, \quad \text{and} \quad f(x, u) = u.$$

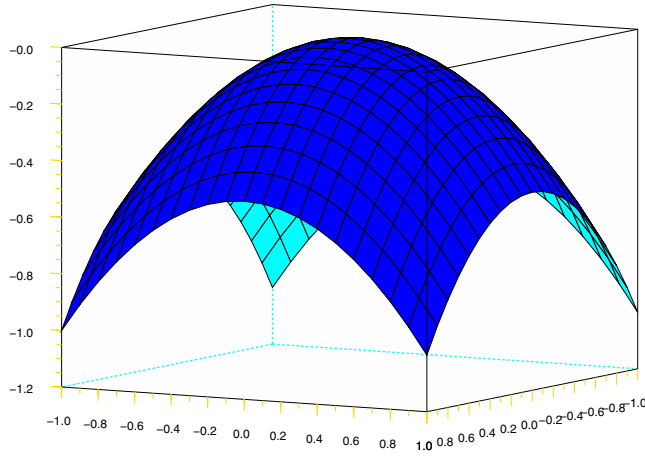


FIG. 8. *Max-plus approximation of a linear quadratic control problem (Example 28).*

For  $x \in X$ , the value functions at time  $t$  is

$$v(x, t) = -\frac{1}{2} \tanh(t)(x_1^2 + x_2^2).$$

As in Example 26, the domain  $X$  is unbounded; therefore  $\ell$  and  $f$  do not satisfy assumptions (H1) and (H2). We will restrict the domain to the set  $[-5; 5]^2$ . We choose quadratic finite elements  $w_i$  and  $z_j$  of Hessian  $c$  centered at the points of the regular grid  $((\mathbb{Z}\Delta x) \cap [-6, 6])^2$ . We represent in Figure 8 the solution given by our algorithm in the case where  $T = 5$ ,  $\delta = 0.5$ ,  $\Delta x = 0.1$ , and  $c = 1$ . The  $L_\infty$ -error is  $9 \cdot 10^{-5}$ .

*Example 29* (distance problem in dimension 2). We consider the case where  $T = 1$ ,  $\phi \equiv 0$ ,  $X = [-1, 1]^2$ ,  $U = [-1, 1]^2$ ,

$$\ell(x, u) = \begin{cases} -1 & \text{if } x \in \text{int}X, \\ 0 & \text{if } x \in \partial X, \end{cases}$$

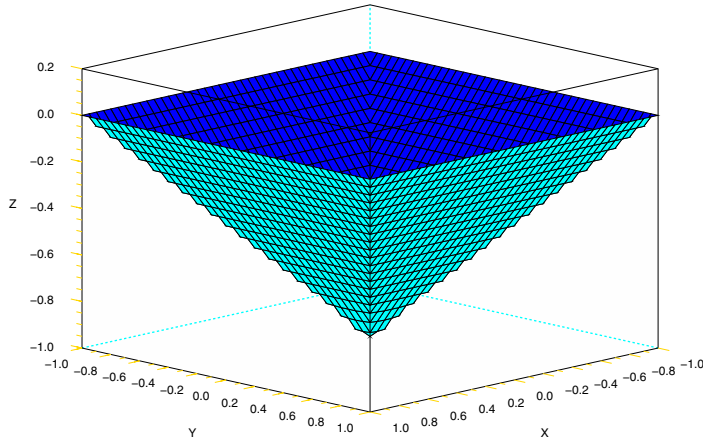
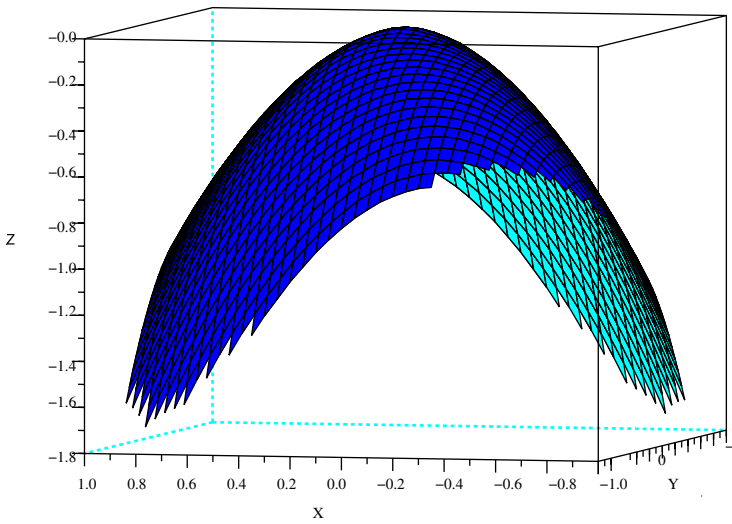
and

$$f(x, u) = 7 \begin{cases} u & \text{if } x \in \text{int}X, \\ 0 & \text{if } x \in \partial X. \end{cases}$$

For  $x \in X$ , the value function at time  $t$  is

$$v(x, t) = \max(-t, \max(|x_1|, |x_2|) - 1).$$

We choose quadratic finite elements  $w_i$  of Hessian  $c$  centered at the points of the regular grid  $((\mathbb{Z}\Delta x) \cap [-3, 3])^2$  and Lipschitz finite elements  $z_j$  with constant  $a$  centered at the points of the regular grid  $((\mathbb{Z}\Delta x) \cap [-1, 1])^2$ . We represent in Figure 9 the solution given by our algorithm in the case where  $T = 1$ ,  $\delta = 0.05$ ,  $\Delta x = 0.025$ ,  $a = 3$ , and  $c = 1$ . The  $L_\infty$ -error is of order 0.05.

FIG. 9. *Max-plus approximation of the distance problem (Example 29).*FIG. 10. *Max-plus approximation of the rotating problem (Example 30).*

*Example 30* (rotating problem). We consider here the Mayer problem where  $T = 1$ ,  $X = B_2(0, 1)$ ,  $U = \{0\}$ ,  $\phi(x) = -\frac{1}{2}x_1^2 - \frac{3}{2}x_2^2$ ,  $\ell(x, u) = 0$ , and  $f(x, u) = (-x_2, x_1)$ . For  $x \in X$ , the value function at time  $t$  is

$$v(x, t) = -\frac{1}{2}(-x_2 \sin(t) + x_1 \cos(t))^2 - \frac{3}{2}(x_2 \cos(t) + x_1 \sin(t))^2.$$

We choose quadratic finite elements  $w_i$  and  $z_j$  of Hessians  $c_w$  and  $c_z$ , respectively, centered at the points of the regular grid  $((\mathbb{Z}\Delta x) \cap [-2, 2])^2$ . We represent in Figure 10 the solution given by our algorithm in the case where  $\delta = \Delta x = 0.05$ ,  $c_w = 4$ , and  $c_z = 3$ . The  $L_\infty$ -error is 0.046.

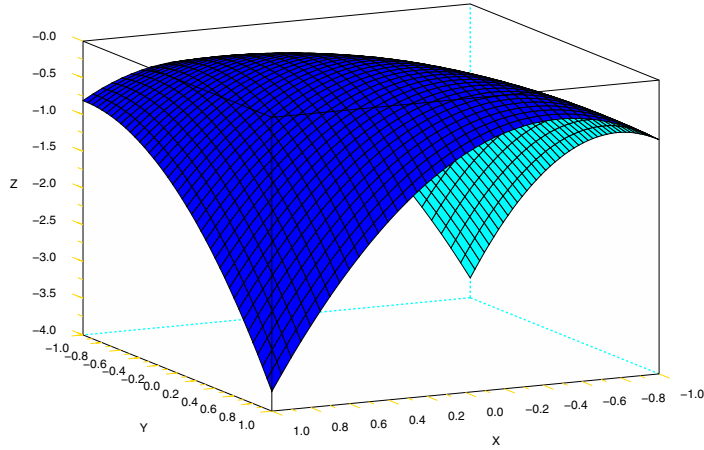


FIG. 11. Max-plus approximation of the solution of the control problem of Example 31.

*Example 31.* We consider the case where  $U = \mathbb{R}$ ,  $X = \mathbb{R}^2$ ,  $\phi(x) = -x_1^2 - 2x_2^2$ ,

$$\ell(x, u) = -x_1^2 - \frac{u^2}{2}, \quad \text{and} \quad f(x, u) = (x_2, u)^T.$$

We choose quadratic finite elements  $w_i$  and  $z_j$  of Hessian  $c_w$  and  $c_z$ , respectively, centered at the points of the grids  $((\mathbb{Z}\Delta x) \cap [-2, 2])^2$  and  $((\mathbb{Z}\Delta x) \cap [-11, 11])^2$ , respectively. We represent in Figure 11 the solution given by our algorithm in the case where  $T = 1$ ,  $\delta = 0.05$ ,  $\Delta x = 0.025$ ,  $c_w = 10$ , and  $c_z = 1$ . The  $L_\infty$ -error is 0.11. (We compared the max-plus approximation with the solution of the problem given by the Riccati equation.)

**6.4. Conclusion.** We have tested our method on examples that fulfill the assumptions of Theorem 22 (see Examples 24, 25, 30) but also on problems that do not fulfill these assumptions. The method is efficient even in the second case. The only difficulty comes from the full character of the matrices  $M_h$  and  $K_h$ , which limits the number of discretization points. To treat higher dimensional examples, we need higher-order approximations (when the value function is regular enough). This is the object of a subsequent work.

**Acknowledgment.** We thank Henda El Fekih for advice and suggestions throughout the development of the present work.

#### REFERENCES

- [AGL04] M. AKIAN, S. GAUBERT, AND A. LAKHOUA, *A max-plus finite element method for solving finite horizon deterministic optimal control problems*, in Proceedings of the Sixteenth International Symposium on Mathematical Theory of Networks and Systems (MTNS'04), Leuven, Belgium, 2004; also available online from <http://arxiv.org/abs/math/0404184> (2004).
- [Bar94] G. BARLES, *Solutions de viscosité des équations de Hamilton-Jacobi*, Springer-Verlag, Paris, 1994.
- [BCD97] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, MA, 1997.
- [BCOQ92] F. BACCELLI, G. COHEN, G. J. OLSDER, AND J.-P. QUADRAT, *Synchronization and Linearity: An Algebra for Discrete Events Systems*, John Wiley and Sons, New York, 1992.

- [BD99] M. BOUÉ AND P. DUPUIS, *Markov chain approximations for deterministic control problems with affine dynamics and quadratic cost in the control*, SIAM J. Numer. Anal., 36 (1999), pp. 667–695.
- [Bir67] G. BIRKHOFF, *Lattice Theory*, Amer. Math. Soc. Colloq. Publ. 25, AMS, Providence, RI, 1967.
- [BJ72] T. S. BLYTH AND M. F. JANOWITZ, *Residuation Theory*, Internat. Ser. Monogr. Pure Appl. Math. 102, Pergamon Press, Oxford, UK, 1972.
- [BS00] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer Ser. Oper. Res., Springer-Verlag, New York, 2000.
- [BZ07] O. BOKANOWSKI AND H. ZIDANI, *Anti-dissipative schemes for advection and application to Hamilton-Jacobi-Bellman equations*, J. Sci. Comput., 30 (2007), pp. 1–33.
- [CD83] I. CAPUZZO DOLCETTA, *On a discrete approximation of the Hamilton-Jacobi equation of dynamic programming*, Appl. Math. Optim., 10 (1983), pp. 367–377.
- [CDF89] I. CAPUZZO-DOLCETTA AND M. FALCONE, *Discrete dynamic programming and viscosity solutions of the Bellman equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (1989), pp. 161–183.
- [CDI84] I. CAPUZZO-DOLCETTA AND H. ISHII, *Approximate solutions of the Bellman equation of deterministic control theory*, Appl. Math. Optim., 11 (1984), pp. 161–181.
- [CFF04] E. CARLINI, M. FALCONE, AND R. FERRETTI, *An efficient algorithm for Hamilton-Jacobi equations in high dimension*, Comput. Vis. Sci., 7 (2004), pp. 15–29.
- [CG79] R. CUNINGHAME-GREEN, *Minimax Algebra*, Lecture Notes in Econom. and Math. Systems 166, Springer-Verlag, Berlin, New York, 1979.
- [CGQ96] G. COHEN, S. GAUBERT, AND J.-P. QUADRAT, *Kernels, images and projections in dioids*, in Proceedings of the International Workshop on Discrete Event Systems (WODES’96), IEE, Edinburgh, UK, 1996.
- [CGQ04] G. COHEN, S. GAUBERT, AND J.-P. QUADRAT, *Duality and separation theorem in idempotent semimodules*, Linear Algebra Appl., 379 (2004), pp. 395–422.
- [CL84] M. G. CRANDALL AND P.-L. LIONS, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.
- [CM04] G. COLLINS AND W. MCENEANEY, *Min-plus eigenvector methods for nonlinear  $H_\infty$  problems with active control*, in Optimal Control, Stabilization and Nonsmooth Analysis, Lecture Notes in Control and Inform. Sci. 301, Springer-Verlag, Berlin, 2004, pp. 101–120.
- [CT80] M. G. CRANDALL AND L. TARTAR, *Some relations between non expansive and order preserving maps*, Proc. Amer. Math. Soc., 78 (1980), pp. 385–390.
- [DJLC53] M. DUBREIL-JACOTIN, L. LESIEUR, AND R. CROISOT, *Théorie des treillis des structures algébriques ordonnées et des treillis géométriques*, Gauthier-Villars, Paris, 1953.
- [Fal87a] M. FALCONE, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 1–13.
- [Fal87b] M. FALCONE, *Corrigenda: A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 213–214.
- [Fat] A. FATHI, *The Weak KAM Theorem in Lagrangian Dynamics*, Cambridge University Press, Cambridge, UK, to appear.
- [FF94] M. FALCONE AND R. FERRETTI, *Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations*, Numer. Math., 67 (1994), pp. 315–344.
- [FG99] M. FALCONE AND T. GIORGI, *An approximation scheme for evolutive Hamilton-Jacobi equations*, in Stochastic Analysis, Control, Optimization and Applications, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 1999, pp. 289–303.
- [FM00] W. H. FLEMING AND W. M. MCENEANEY, *A max-plus-based algorithm for a Hamilton-Jacobi-Bellman equation of nonlinear filtering*, SIAM J. Control Optim., 38 (2000), pp. 683–710.
- [FS93] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
- [GM01] M. GONDRAAN AND M. MINOUX, *Graphes, Dioïdes et semi-anneaux*, TEC & DOC, Paris, 2001.
- [Gon96] M. GONDRAAN, *Analyse MINPLUS*, C.R. Acad. Sci. Paris Sér. I Math., 323 (1996), pp. 371–375.
- [GR85a] R. GONZALEZ AND E. ROFMAN, *On deterministic control problems: An approximation procedure for the optimal cost I. The stationary problem*, SIAM J. Control Optim., 23 (1985), pp. 242–266.
- [GR85b] R. GONZALEZ AND E. ROFMAN, *On deterministic control problems: An approximation procedure for the optimal cost II. The nonstationary case*, SIAM J. Control Optim., 23 (1985), pp. 267–285.

- [HUL93] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms I*, Springer-Verlag, Berlin, 1993.
- [KM88] V. N. KOLOKOLTSOV AND V. P. MASLOV, *The Cauchy problem for the homogeneous Bellman equation*, Soviet Math. Dokl., 36 (1988), pp. 326–330.
- [KM97] V. N. KOLOKOLTSOV AND V. P. MASLOV, *Idempotent Analysis and Its Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.
- [Lak03] A. LAKHOVA, *Résolution numérique de problèmes de commande optimale déterministe et algèbre max-plus*, Rapport de DEA, Université Paris VI, Paris, France, 2003.
- [Lio82] P.-L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, London, 1982.
- [LMS01] G. L. LITVINOV, V. P. MASLOV, AND G. B. SHPIZ, *Idempotent functional analysis: An algebraic approach*, Math. Notes, 69 (2001), pp. 696–729.
- [Mas73] V. MASLOV, *Méthodes Operatorielles*, Mir, Moscow, 1973 (French transl. 1987).
- [McE02] W. M. MCENEANEY, *Error analysis of a max-plus algorithm for a first-order HJB equation*, in Stochastic Theory and Control (Lawrence, KS, 2001), Lecture Notes in Control and Inform. Sci. 280, Springer-Verlag, Berlin, 2002, pp. 335–351.
- [McE03] W. M. MCENEANEY, *Max-plus eigenvector representations for solution of nonlinear  $H_\infty$  problems: Basic concepts*, IEEE Trans. Automat. Control, 48 (2003), pp. 1150–1163.
- [McE04] W. M. MCENEANEY, *Max-plus eigenvector methods for nonlinear  $H_\infty$  problems: Error analysis*, SIAM J. Control Optim., 43 (2004), pp. 379–412.
- [McE06] W. M. MCENEANEY, *Max-Plus Methods for Nonlinear Control and Estimation*, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 2006.
- [MH98] W. M. MCENEANEY AND M. HORTON, *Max-plus eigenvector representations for nonlinear  $H_\infty$  value functions*, in Proceedings of the 37th IEEE Conference on Decision and Control (CDC'98), 1998, pp. 3506–3511.
- [MH99] W. M. MCENEANEY AND M. HORTON, *Computation of max-plus eigenvector representations for nonlinear  $H_\infty$  value functions*, in Proceedings of the American Control Conference, 1999, pp. 1400–1404.
- [MS92] V. P. MASLOV AND S. SAMBORSKIĬ, EDS., *Idempotent Analysis*, Adv. Soviet. Math. 13, AMS, Providence, RI, 1992.
- [OS91] S. OSHER AND C.-W. SHU, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.
- [Plu98] M. PLUS, *Documentation of the Max-Plus Toolbox of Scilab*, 1998. Available from <ftp://ftp.inria.fr/INRIA/Scilab/contrib/MAXPLUS/>.
- [Roc70] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Math. Ser. 28, Princeton University Press, Princeton, NJ, 1970.
- [SU00] J.-R. SACK AND J. URRUTIA, *Handbook of Computational Geometry*, North-Holland, Amsterdam, 2000.

## NONUNIFORM SMALL-GAIN THEOREMS FOR SYSTEMS WITH UNSTABLE INVARIANT SETS\*

IVAN TYUKIN<sup>†</sup>, ERIK STEUR<sup>‡</sup>, HENK NIJMEIJER<sup>‡</sup>, AND CEES VAN LEEUWEN<sup>§</sup>

**Abstract.** We consider the problem of asymptotic convergence to invariant sets in interconnected nonlinear dynamical systems. Standard approaches often require that the invariant sets be uniformly attracting, e.g., stable in the Lyapunov sense. This, however, is neither a necessary requirement nor is always useful. Systems may, for instance, be inherently unstable (e.g., intermittent, itinerant, meta-stable) or the problem statement may include requirements that cannot be satisfied with stable solutions. This is often the case in general optimization problems and in nonlinear parameter identification or adaptation. Conventional techniques for these cases either rely on detailed knowledge of the system's vector-fields or require boundedness of its states. The presently proposed method relies only on estimates of the input-output maps and steady-state characteristics. The method requires the possibility of representing the system as an interconnection of a stable and contracting part with an unstable and exploratory part. We illustrate with examples how the method can be applied to problems of analyzing the asymptotic behavior of locally unstable systems as well as to problems of parameter identification and adaptation in the presence of nonlinear parametrizations. The relation of our results to conventional small-gain theorems is discussed.

**Key words.** nonuniform convergence, weakly attracting sets, small-gain theorems, input-output stability

**AMS subject classifications.** 40A99, 34D05, 34D45, 93D25, 93B03, 93B30

**DOI.** 10.1137/060672546

**1. Notation.** Throughout the paper we use the following notational conventions. The symbol  $\mathbb{R}$  denotes the field of real numbers; symbol  $\mathbb{R}_+$  stands for the following subset of  $\mathbb{R}$ :  $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$ ; and  $\mathbb{N}$  and  $\mathbb{Z}$  denote the set of natural numbers and its extension to the negative domain, respectively.

Let  $\Omega$  be a set; by the symbol  $\mathcal{S}\{\Omega\}$  we denote the set of all subsets of  $\Omega$ . The symbol  $\mathcal{C}^k$  denotes the space of functions that are at least  $k$  times differentiable;  $\mathcal{K}$  denotes the class of all strictly increasing functions  $\kappa : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  such that  $\kappa(0) = 0$ . If, in addition,  $\lim_{s \rightarrow \infty} \kappa(s) = \infty$ , we say that  $\kappa \in \mathcal{K}_\infty$ . Further,  $\mathcal{K}_e$  (or  $\mathcal{K}_{e,\infty}$ ) denotes the class of functions of which the restriction to the interval  $[0, \infty)$  belongs to  $\mathcal{K}$  (or  $\mathcal{K}_\infty$ ). The symbol  $\mathcal{KL}$  denotes the class of functions  $\beta : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$  such that  $\beta(\cdot, s) \in \mathcal{K}$  and  $\beta(r, \cdot)$  is monotonically decreasing for each  $s, r \in \mathbb{R}_+$ .

Let  $\mathbf{x} \in \mathbb{R}^n$ , and  $\mathbf{x}$  can be partitioned into two vectors  $\mathbf{x}_1 \in \mathbb{R}^q$ ,  $\mathbf{x}_1 = (x_{11}, \dots, x_{1q})^T$ ,  $\mathbf{x}_2 \in \mathbb{R}^p$ ,  $\mathbf{x}_2 = (x_{21}, \dots, x_{2p})^T$  with  $q + p = n$ ; then  $\oplus$  denotes their concatenation:  $\mathbf{x} = \mathbf{x}_1 \oplus \mathbf{x}_2$ .

---

\*Received by the editors October 12, 2006; accepted for publication (in revised form) October 30, 2007; published electronically February 27, 2008.

<http://www.siam.org/journals/sicon/47-2/67254.html>

<sup>†</sup>Corresponding author. Laboratory for Perceptual Dynamics, RIKEN (Institute for Physical and Chemical Research) Brain Science Institute, 2-1, Hirosawa, Wako-shi, Saitama, 351-0198, Japan (tyukinivan@brain.riken.jp). Present position: Department of Mathematics, University of Leicester, University Road, Leicester, LE1 7RH, UK (I.Tyukin@le.ac.uk).

<sup>‡</sup>Department of Mechanical Engineering, Dynamics and Control, Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands (e.steur@tue.nl, h.nijmeijer@tue.nl).

<sup>§</sup>Laboratory for Perceptual Dynamics, RIKEN (Institute for Physical and Chemical Research) Brain Science Institute, 2-1, Hirosawa, Wako-shi, Saitama, 351-0198, Japan (ceesvl@brain.riken.jp).

The symbol  $\|\mathbf{x}\|$  denotes the Euclidian norm in  $\mathbf{x} \in \mathbb{R}^n$ . By  $L_\infty^n[t_0, T]$  we denote the space of all functions  $\mathbf{f} : \mathbb{R}_+ \rightarrow \mathbb{R}^n$  such that  $\|\mathbf{f}\|_{\infty, [t_0, T]} = \sup\{\|\mathbf{f}(t)\|, t \in [t_0, T]\} < \infty$ , and  $\|\mathbf{f}\|_{\infty, [t_0, T]}$  stands for the  $L_\infty^n[t_0, T]$  norm of  $\mathbf{f}(t)$ . Let  $\mathcal{A}$  be a set in  $\mathbb{R}^n$  and  $\|\cdot\|$  be the usual Euclidian norm in  $\mathbb{R}^n$ . By the symbol  $\|\cdot\|_{\mathcal{A}}$  we denote the following induced norm:

$$\|\mathbf{x}\|_{\mathcal{A}} = \inf_{\mathbf{q} \in \mathcal{A}} \{\|\mathbf{x} - \mathbf{q}\|\}.$$

Let  $\Delta \in \mathbb{R}_+$ ; then the notation  $\|\mathbf{x}\|_{\mathcal{A}_\Delta}$  stands for the following equality:

$$\|\mathbf{x}\|_{\mathcal{A}_\Delta} = \begin{cases} \|\mathbf{x}\|_{\mathcal{A}} - \Delta, & \|\mathbf{x}\|_{\mathcal{A}} > \Delta, \\ 0, & \|\mathbf{x}\|_{\mathcal{A}} \leq \Delta. \end{cases}$$

The symbol  $\|\cdot\|_{\mathcal{A}_\infty, [t_0, t]}$  is defined as follows:

$$\|\mathbf{x}(\tau)\|_{\mathcal{A}_\infty, [t_0, t]} = \sup_{\tau \in [t_0, t]} \|\mathbf{x}(\tau)\|_{\mathcal{A}}.$$

**2. Introduction.** In many fields of science, such as systems and control theory, physics, chemistry, and biology, it is of fundamental importance to analyze the asymptotic behavior of dynamical systems. Most of these analyses are based around the concept of Lyapunov stability [15], [33], [32], i.e., continuity of the flow  $\mathbf{x}(t, \mathbf{x}_0) : \mathbb{R}_+ \times \mathbb{R}^n \rightarrow L_\infty^n[t_0, \infty]$  with respect to  $\mathbf{x}_0$  [18], in combination with the standard notion of an *attracting set* [9], defined as follows.

DEFINITION 1. A set  $\mathcal{A}$  is an *attracting set* iff it is

- (i) *closed, invariant, and*
- (ii) *for some neighborhood  $\mathcal{V}$  of  $\mathcal{A}$  and for all  $\mathbf{x}_0 \in \mathcal{V}$  the following conditions hold:*

$$(1) \quad \mathbf{x}(t, \mathbf{x}_0) \in \mathcal{V} \quad \forall t \geq 0;$$

$$(2) \quad \lim_{t \rightarrow \infty} \|\mathbf{x}(t, \mathbf{x}_0)\|_{\mathcal{A}} = 0.$$

Condition (1) in Definition 1 stipulates the existence of a trapping region  $\mathcal{V}$  which is a neighborhood of  $\mathcal{A}$ . Condition (2) assures convergence to  $\mathcal{A}$ . Due to condition (1), convergence to  $\mathcal{A}$  is uniform with respect to  $\mathbf{x}_0$  in the neighborhood of  $\mathcal{A}$ ; i.e., every trajectory which starts in  $\mathcal{V}$  remains in  $\mathcal{V}$  for  $t \geq 0$  and converges to  $\mathcal{A}$  at  $t \rightarrow \infty$ .

Although the conventional concepts of attracting set and Lyapunov stability are powerful in tandem in various applications, some problems cannot be solved within this framework. Condition (1), for example, could be violated in systems with intermittent, itinerant, or meta-stable dynamics. In general the condition does not hold when the system dynamics, loosely speaking, is exploring rather than contracting. Such systems appear naturally in the context of global optimization. For instance, in [22] finding the global minimum of a differentiable cost function  $Q : \mathbb{R}^n \rightarrow \mathbb{R}_+$  in a bounded subset  $\Omega_x \subset \mathbb{R}^n$  is achieved by splitting the search procedure into a locally attracting gradient  $\mathcal{S}_a$  and a wandering part  $\mathcal{S}_w$ :

$$(3) \quad \begin{aligned} \mathcal{S}_a : \dot{\mathbf{x}} &= -\mu_x \frac{\partial Q(\mathbf{x})}{\partial \mathbf{x}} + \mu_t T(t), \quad \mu_x, \mu_t \in \mathbb{R}_+, \\ \mathcal{S}_w : T(t) &= h\{t, \mathbf{x}(t)\}, \quad h : \mathbb{R}_+ \times L_\infty^n[t_0, t] \rightarrow L_\infty^n[t_0, t]. \end{aligned}$$

The trace function,  $T(t)$ , in (3) is supposed to cover (i.e., be dense in) the whole searching domain  $\Omega_x$ . Even though the results in [22] are purely simulation studies,



they illustrate the superior performance of algorithms (3) in a variety of benchmark problems compared to standard local minimizers and classical methods of global optimization. Abandoning Lyapunov stability is likewise advantageous in problems of identification and adaptation in the presence of general nonlinear parametrization [28], in maneuvering and path searching [26], and in decision making in intelligent systems [30], [31]. Systems with attracting, yet unstable, invariant sets are relevant for modeling complex behavior in biological and physical systems [2]. Last but not least, Lyapunov-unstable attracting sets are relevant in problems of synchronization [5], [19], [27].<sup>1</sup>

Even when it is appropriate to consider a system as stable, we may be limited in our success in meeting the requirement to identify a proper Lyapunov function. This is the case, for instance, when the system's dynamics is only partially known. Trading stability requirements for the sake of convergence might be a possible remedy. Known results in this direction can be found in [11], [21].<sup>2</sup>

In all the cases that are problematic under condition (1) of Definition 1, condition (2)—convergence of  $\mathbf{x}(t, \mathbf{x}_0)$  to an invariant set  $\mathcal{A}$ —is still a requirement that has to be met. In order to treat these cases analytically we shall, first of all, move from the standard concept of attracting sets in Definition 1 to one that does not assume that the basin of attraction is necessarily a neighborhood of the invariant set  $\mathcal{A}$ . In other words we shall allow convergence which is not uniform in initial conditions. This requirement is captured by the concept of weak, or Milnor, attraction [17], defined as follows.

DEFINITION 2. *A set  $\mathcal{A}$  is weakly attracting, or Milnor attracting, iff*

- (i) *it is closed, invariant, and*
- (ii) *for some set  $\mathcal{V}$  (not necessarily a neighborhood of  $\mathcal{A}$ ) with strictly positive measure and for all  $\mathbf{x}_0 \in \mathcal{V}$  limiting relation (2) holds.*

Conventional methods such as La Salle's invariance principle [14] or center manifold theory [7] can, in principle, address the issue of convergence to weak equilibria. They do so, however, at the expense of requiring detailed knowledge of the vector-fields of the ordinary differential equations of the model. When such information is not available the system can be thought of as a mere interconnection of input-output maps. Small-gain theorems [34], [12] are usually efficient in this case. These results, however, apply only under the assumption of stability of each component in the interconnection.

In the present study we aim to find a proper balance between the generality of input-output approaches [34], [12] in the analysis of convergence and the specificity of the fundamental notions of limit sets and invariance that play a central role in [14], [7]. The object of our study is a class of systems that can be decomposed into an attracting, or stable, component  $\mathcal{S}_a$  and an exploratory, generally unstable, part  $\mathcal{S}_w$ . Typical systems of this class are nonlinear systems in cascaded form

$$(4) \quad \begin{aligned} \mathcal{S}_a : \quad \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}, \mathbf{z}), \\ \mathcal{S}_w : \quad \dot{\mathbf{z}} &= \mathbf{q}(\mathbf{z}, \mathbf{x}), \end{aligned}$$

where the zero solution of the  $\mathbf{x}$ -subsystem is asymptotically stable in the absence of

<sup>1</sup>See also [20], where the striking difference between stable and “almost stable” synchronization in terms of the coupling strengths for a pair of the Lorenz oscillators is demonstrated analytically.

<sup>2</sup>In Examples 1 and 2 in section 6, we demonstrate how explorative dynamics can solve the problem of simultaneous state and parameter observation for a system which cannot be transformed into a canonical adaptive observer form [3].

input  $\mathbf{z}$ , and the state of the  $\mathbf{z}$ -subsystem consists of functions of the type  $\int_{t_0}^t \|\mathbf{x}(\tau)\| d\tau$ . Even when both subsystems in (4) are stable and the  $\mathbf{x}$ -subsystem does not depend on state  $\mathbf{z}$ , the cascade can still be unstable [1]. We show, however, that for unstable interconnections (4), under certain conditions that involve only input-to-state properties of  $\mathcal{S}_a$  and  $\mathcal{S}_w$ , there is a set  $\mathcal{V}$  in the system state space such that trajectories starting in  $\mathcal{V}$  remain bounded. The result is formally stated in Theorem 3. In the case when an additional measure of invariance is defined for  $\mathcal{S}_a$  (in our case a steady-state characteristic), a weak, Milnor attracting set emerges. Its location is completely determined by the zeros of the steady-state response of system  $\mathcal{S}_a$ .

We demonstrate how this basic result can be used in problems of design and analysis of control systems and identification/adaptation algorithms. In particular, we present an adaptive observer of state and parameter values for uncertain systems which cannot be transformed into a canonic adaptive observer form [3]. In Examples 1 and 2 in section 6 we present an application of this result to the problem of reconstructing a dynamic model of neuronal cell activity.

The paper is organized as follows. In section 3 we formally state the problem and provide specific assumptions for the class of systems under consideration. Section 4 contains the main results of our present study. In section 5 we provide several corollaries of the main result that apply to specific problems. Section 6 contains examples, and section 7 concludes the paper. Proofs of all lemmas, theorems, and corollaries are provided in the appendix.

**3. Problem formulation.** Consider a system that can be decomposed into two interconnected subsystems,  $\mathcal{S}_a$  and  $\mathcal{S}_w$ :

$$(5) \quad \begin{aligned} \mathcal{S}_a &: (u_a, \mathbf{x}_0) \mapsto \mathbf{x}(t), \\ \mathcal{S}_w &: (u_w, \mathbf{z}_0) \mapsto \mathbf{z}(t), \end{aligned}$$

where  $u_a \in \mathcal{U}_a \subseteq L_\infty[t_0, \infty]$ ,  $u_w \in \mathcal{U}_w \subseteq L_\infty[t_0, \infty]$  are the spaces of inputs to  $\mathcal{S}_a$  and  $\mathcal{S}_w$ , respectively,  $\mathbf{x}_0 \in \mathbb{R}^n$ ,  $\mathbf{z}_0 \in \mathbb{R}^m$  represent initial conditions, and  $\mathbf{x}(t) \in \mathcal{X} \subseteq L_\infty^n[t_0, \infty]$ ,  $\mathbf{z}(t) \in \mathcal{Z} \subseteq L_\infty^m[t_0, \infty]$  are the system states.

System  $\mathcal{S}_a$  represents the contracting dynamics. More precisely, we require that  $\mathcal{S}_a$  is input-to-state stable<sup>3</sup> [23] with respect to a compact set  $\mathcal{A}$ .

ASSUMPTION 1 (contracting dynamics).

$$(6) \quad \mathcal{S}_a : \|\mathbf{x}(t)\|_{\mathcal{A}} \leq \beta(\|\mathbf{x}(t_0)\|_{\mathcal{A}}, t - t_0) + c\|u_a(t)\|_{\infty, [t_0, t]} \quad \forall t_0 \in \mathbb{R}_+, t \geq t_0,$$

where the function  $\beta(\cdot, \cdot) \in \mathcal{KL}$ , and  $c > 0$  is some positive constant.

The function  $\beta(\cdot, \cdot)$  in (6) specifies the contraction property of the unperturbed dynamics of  $\mathcal{S}_a$ . In other words it models the rate with which the system forgets its initial conditions  $\mathbf{x}_0$ , if left unperturbed. Propagation of the input to output is estimated in terms of a continuous mapping,  $c\|u_a(t)\|_{\infty, [t_0, t]}$ , which, in our case, is chosen for simplicity to be linear. Notice that this mapping should not necessarily be contracting. In what follows we will assume that the function  $\beta(\cdot, \cdot)$  and constant  $c$  are known or can be estimated a priori.

For systems  $\mathcal{S}_a$ , of which a model is given by a system of ordinary differential equations

$$(7) \quad \dot{\mathbf{x}} = \mathbf{f}_x(\mathbf{x}, u_a), \quad \mathbf{f}_x(\cdot, \cdot) \in \mathcal{C}^1,$$

<sup>3</sup>In general, as will be demonstrated with examples, our analysis can be carried out for (integral) input-to-output/input-to-state stable systems as well.

Assumption 1 is equivalent, for instance, to the combination of the following properties:<sup>4</sup>

1. Let  $u_a(t) \equiv 0$  for all  $t$ ; then the set  $\mathcal{A}$  is Lyapunov stable and globally attracting for (7).
2. For all  $u_a \in \mathcal{U}_a$  and  $\mathbf{x}_0 \in \mathbb{R}^n$  there exists a nondecreasing function  $\kappa : \mathbb{R}_+ \rightarrow \mathbb{R}_+ : \kappa(0) = 0$  such that

$$\inf_{t \in [0, \infty)} \|\mathbf{x}(t)\|_{\mathcal{A}} \leq \kappa(\|u_a(t)\|_{\infty, [t_0, \infty)}).$$

The system  $\mathcal{S}_w$  stands for the searching or wandering dynamics. We will consider  $\mathcal{S}_w$  subject to the following conditions.

ASSUMPTION 2 (wandering dynamics). *The system  $\mathcal{S}_w$  is forward-complete:*

$$u_w(t) \in \mathcal{U}_w \Rightarrow \mathbf{z}(t) \in \mathcal{Z} \quad \forall t \geq t_0, \quad t_0 \in \mathbb{R}_+,$$

and there exists an “output” function  $h : \mathbb{R}^m \rightarrow \mathbb{R}$ , and two “bounding” functions  $\gamma_0 \in \mathcal{K}_{\infty, e}$ ,  $\gamma \in \mathcal{K}_{\infty, e}$  such that the following integral inequality holds:

$$(8) \quad \mathcal{S}_w : \int_{t_0}^t \gamma_1(u_w(\tau)) d\tau \leq h(\mathbf{z}(t_0)) - h(\mathbf{z}(t)) \leq \int_{t_0}^t \gamma_0(u_w(\tau)) d\tau$$

$$\forall t \geq t_0, \quad t_0 \in \mathbb{R}_+.$$

In the case when system  $\mathcal{S}_w$  is specified in terms of vector-fields

$$(9) \quad \dot{\mathbf{z}} = \mathbf{f}_z(\mathbf{z}, u_w), \quad \mathbf{f}_z(\cdot, \cdot) \in \mathcal{C}^1,$$

Assumption 2 can be viewed, for example, as postulating the existence of a function  $h : \mathbb{R}^m \rightarrow \mathbb{R}_+$  of which the evolution in time is a mere integration of the input  $u_w(t)$ . In general, for  $u_w : u_w(t) \geq 0$  for all  $t \in \mathbb{R}_+$ , inequality (8) implies *monotonicity* of function  $h(\mathbf{z}(t))$  in  $t$ . Regarding the function  $\gamma_0(\cdot)$  in (8), we assume that for any  $M \in \mathbb{R}_+$  there exists a function  $\gamma_{0,1} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  and a *nondecreasing* function  $\gamma_{0,2} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  such that

$$(10) \quad \gamma_0(a \cdot b) \leq \gamma_{0,1}(a) \cdot \gamma_{0,2}(b) \quad \forall a, b \in [0, M].$$

Requirement (10) is a technical assumption which will be used in the formulation and proof of the main results of the paper. Yet, it is not too restrictive; it holds, for instance, for a wide class of locally Lipschitz functions  $\gamma_0(\cdot) : \gamma_0(a \cdot b) \leq L_0(M) \cdot (a \cdot b)$ ,  $L_0(M) \in \mathbb{R}_+$ . Another example for which the assumption holds is the class of polynomial functions  $\gamma_0(\cdot) : \gamma_0(a \cdot b) = (a \cdot b)^p = a^p \cdot b^p$ ,  $p > 0$ . No further restrictions will be imposed a priori on  $\mathcal{S}_a, \mathcal{S}_w$ .

Now consider the interconnection of (6), (8) with coupling  $u_a(t) = h(\mathbf{z}(t))$  and  $u_s(t) = \|\mathbf{x}(t)\|_{\mathcal{A}}$ . Equations for the combined system can be written as

$$(11) \quad \begin{aligned} \|\mathbf{x}(t)\|_{\mathcal{A}} &\leq \beta(\|\mathbf{x}(t_0)\|_{\mathcal{A}}, t - t_0) + c\|h(\mathbf{z}(t))\|_{\infty, [t_0, t]}, \\ \int_{t_0}^t \gamma_1(\|\mathbf{x}(\tau)\|_{\mathcal{A}}) d\tau &\leq h(\mathbf{z}(t_0)) - h(\mathbf{z}(t)) \leq \int_{t_0}^t \gamma_0(\|\mathbf{x}(\tau)\|_{\mathcal{A}}) d\tau. \end{aligned}$$

<sup>4</sup>For a comprehensive characterization of the input-to-state stability and detailed mathematical arguments we refer to the paper by Sontag and Wang [24].

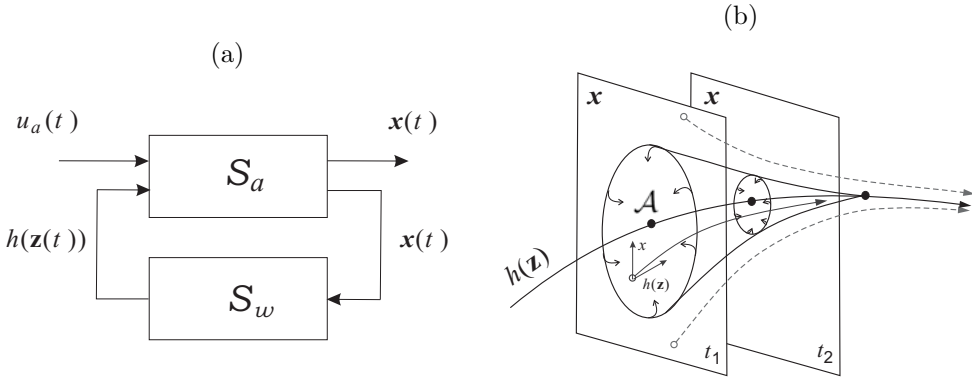


FIG. 1. The class of interconnected systems  $S_a$  and  $S_w$  (diagram (a)). System  $S_a$ , the “contracting system,” has an attracting invariant set  $A$  in its state space. System  $S_w$  does not necessarily have an attracting set. This system represents the “wandering” dynamics. A typical example of such behavior is the dynamics of the flow in a neighborhood of a saddle point in three-dimensional space (diagram (b)).

A diagram illustrating the general structure of the entire system (11) is given in Figure 1.

Equations (11) capture the relevant interplay between contracting,  $S_a$ , and wandering,  $S_w$ , dynamics inherent in a variety of searching strategies in the realm of optimization, (3), and interconnections, (4), in general systems theory. In addition, this kind of interconnection describes the behavior of systems which undergo transcritical or saddle-node bifurcations. Consider, for instance, the following system:

$$(12) \quad \begin{aligned} \dot{x}_1 &= -x_1 + x_2, \\ \dot{x}_2 &= \varepsilon + \gamma x_1^2, \gamma > 0, \end{aligned}$$

where the parameter  $\varepsilon$  varies from negative to positive values. At  $\varepsilon = 0$  stable and unstable equilibria collide, leading to the cascade satisfying (11). An alternative bifurcation scenario could be represented by the system

$$(13) \quad \begin{aligned} \dot{x}_1 &= -x_1 + x_2, \\ \dot{x}_2 &= \varepsilon + \gamma x_2^2, \gamma > 0. \end{aligned}$$

In this case, however, the dynamics of the variable  $x_2$  is *independent* of  $x_1$ , and analysis of the asymptotic behavior of (13) reduces to the analysis of each equation separately. Thus systems such as (13) are easier to deal with than (12). This constitutes an additional motivation for the present approach.

When analyzing the asymptotic behavior of interconnection (11) we will address the following question: Is there a set (a weak trapping set in the system state space) such that the trajectories which start in this set are bounded? It is natural to expect that the existence of such a set depends on the specific functions  $\gamma_0(\cdot)$ ,  $\gamma_1(\cdot)$  in (11), on properties of  $\beta(\cdot, \cdot)$ , and on values of  $c$ . In the case when such a set exists and could be defined, the next questions are, therefore, where will the trajectories converge and how can these domains be characterized?

**4. Main results.** In this section we provide a formal statement of the main results of our present study. In section 4.1, we formulate conditions ensuring that

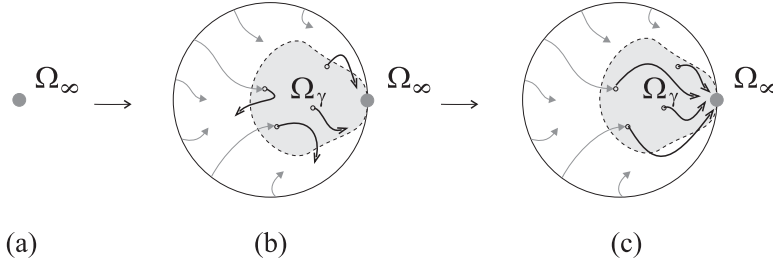


FIG. 2. Emergence of a weak (Milnor) attracting set  $\Omega_\infty$ . Panel (a) depicts the target invariant set  $\Omega_\infty$  as a filled circle. First (see Theorem 3), we investigate whether a domain  $\Omega_\gamma \subset \mathbb{R}^n \times \mathbb{R}^m$  exists such that  $\|\mathbf{x}(t)\|_{\mathcal{A}}, h(\mathbf{z}(t))$  are bounded for all  $\mathbf{x}_0 \oplus \mathbf{z}_0 \in \Omega_\gamma$ . In the text we refer to this set as a weak trapping region or simply a trapping region. The trapping region is shown as a grey domain in panel (b). In principle, the system's states can eventually leave the domain  $\Omega_\gamma$ . They must, however, satisfy (14), ensuring boundedness of  $\|\mathbf{x}(t)\|_{\mathcal{A}}, h(\mathbf{z}(t))$ . As a result they will dwell within the region shown as a circle in panel (b). Notice that neither this domain nor the previous need be neighborhoods of  $\Omega_\infty$ . Second (see Lemmas 6 and 7 and Corollary 8), we provide conditions which lead to the emergence of a weak attracting set in the trapping region  $\Omega_\gamma$ . This is illustrated in panel (c).

there exists a point  $\mathbf{x}_0 \oplus \mathbf{z}_0$  such that the  $\omega$ -limit set of  $\mathbf{x}_0 \oplus \mathbf{z}_0$ <sup>5</sup> is bounded in the following sense:

$$(14) \quad \|\omega_{\mathbf{x}}(\mathbf{x}_0 \oplus \mathbf{z}_0)\|_{\mathcal{A}} < \infty, \quad |h(\omega_{\mathbf{z}}(\mathbf{x}_0 \oplus \mathbf{z}_0))| < \infty.$$

These conditions and a specification of the set  $\Omega_\gamma$  of points  $\mathbf{x}' \oplus \mathbf{z}'$  for which the  $\omega$ -limit set satisfies property (14) are provided in Theorem 3.

In order to verify whether an attracting set exists in  $\omega(\Omega_\gamma)$  that is a subset of  $\omega(\Omega_\gamma)$  we use an additional characterization of the contracting system  $\mathcal{S}_a$ . In particular, we introduce the intuitively clear notion of the input-to-state *steady-state characteristics*<sup>6</sup> of a system. It is possible to show that in the case when system  $\mathcal{S}_a$  has a steady-state characteristic, there exists an attracting set in  $\omega(\Omega_\gamma)$ , and this set is uniquely defined by the zeros of the steady-state characteristics of  $\mathcal{S}_a$ . A diagram illustrating the steps of our analysis is provided in Figure 2, along with the sequence of conditions leading to the emergence of the attracting set in (11).

**4.1. Emergence of the trapping region. Small-gain conditions.** Before we formulate the main results of this section let us first comment briefly on the machinery of our analysis. First, we introduce three sequences

$$\begin{aligned} \mathcal{S} &= \{\sigma_i\}_{i=0}^\infty, \quad \sigma_i \in \mathbb{R}_+, \\ \Xi &= \{\xi_i\}_{i=0}^\infty, \quad \xi_i \in \mathbb{R}_+, \\ \mathcal{T} &= \{\tau_i\}_{i=0}^\infty, \quad \tau_i \in \mathbb{R}_+. \end{aligned}$$

The first sequence,  $\mathcal{S}$ , partitions the interval  $[0, h(\mathbf{z}_0)]$ ,  $h(\mathbf{z}_0) > 0$ , into the union of shrinking subintervals  $H_i$ :

$$(15) \quad [0, h(\mathbf{z}_0)] = \cup_{i=0}^\infty H_i, \quad H_i = [\sigma_{i+1}h(\mathbf{z}_0), \sigma_i h(\mathbf{z}_0)].$$

<sup>5</sup>Recall that in our current notation a point  $\mathbf{p} \in \mathbb{R}^{m+n}$  is an  $\omega$ -limit point of  $\mathbf{x}' \oplus \mathbf{z}'$  if there exists a sequence  $\{t_i\}$ ,  $i = 1, 2, \dots$ , such that  $\lim_{i \rightarrow \infty} t_i = \infty$  and  $\lim_{t_i \rightarrow \infty} \mathbf{x}(t_i, \mathbf{x}' \oplus \mathbf{z}') \oplus \mathbf{z}(t_i, \mathbf{x}' \oplus \mathbf{z}') = \mathbf{p}$ , where  $\mathbf{x}(t, \mathbf{x}' \oplus \mathbf{z}') \oplus \mathbf{z}(t, \mathbf{x}' \oplus \mathbf{z}')$  denotes the flow of interconnection (11). A set of all  $\omega$ -limit points of  $\mathbf{x}' \oplus \mathbf{z}'$  is an  $\omega$ -limit set of  $\mathbf{x}' \oplus \mathbf{z}'$ .

<sup>6</sup>A more precise definition of the steady-state characteristics is given in section 4.2.

For the sake of clarity, let us define this property formally in the form of Property 1 as follows.

PROPERTY 1 (partition of  $\mathbf{z}_0$ ). *The sequence  $\mathcal{S}$  is strictly monotone and converging:*

$$(16) \quad \{\sigma_n\}_{n=0}^\infty : \lim_{n \rightarrow \infty} \sigma_n = 0, \sigma_0 = 1.$$

Sequences  $\Xi$  and  $\mathcal{T}$  will specify the desired rates  $\xi_i \in \Xi$  of the contracting dynamics (6) in terms of function  $\beta(\cdot, \cdot)$  and  $\tau_i \in \mathcal{T}$ . Let us, therefore, impose the following constraint on the choice of  $\Xi, \mathcal{T}$ .

PROPERTY 2 (rate of contraction, part 1). *Sequences  $\Xi$  and  $\mathcal{T}$  are such that for the given function  $\beta(\cdot, \cdot) \in \mathcal{KL}$  in (6) the following inequality holds:*

$$(17) \quad \beta(\cdot, T) \leq \xi_i \beta(\cdot, 0) \quad \forall T \geq \tau_i.$$

Property 2 states that for the given, yet arbitrary, factor  $\xi_i$  and time instant  $t_0$ , time  $\tau_i$  is needed for the state  $\mathbf{x}$  in order to reach the domain:

$$\|\mathbf{x}\|_{\mathcal{A}} \leq \xi_i \beta(\|\mathbf{x}(t_0)\|_{\mathcal{A}}, 0).$$

In order to specify the desired convergence rates  $\xi_i$ , it will be necessary to define another measure in addition to (17). This is a measure of the propagation of initial conditions  $\mathbf{x}_0$  and input  $h(\mathbf{z}_0)$  to the state  $\mathbf{x}(t)$  of the contracting dynamics (6) when the system travels in  $h(\mathbf{z}(t)) \in [0, h(\mathbf{z}_0)]$ . For this reason we introduce two systems of functions,  $\Phi$  and  $\Upsilon$ :

$$(18) \quad \Phi : \begin{aligned} \phi_j(s) &= \phi_{j-1} \circ \rho_{\phi,j}(\xi_{i-j} \cdot \beta(s, 0)), \quad j = 1, \dots, i, \\ \phi_0(s) &= \beta(s, 0), \end{aligned}$$

$$(19) \quad \Upsilon : \begin{aligned} v_j(s) &= \phi_{j-1} \circ \rho_{v,j}(s), \quad j = 1, \dots, i, \\ v_0(s) &= \beta(s, 0), \end{aligned}$$

where the functions  $\rho_{\phi,j}, \rho_{v,j} \in \mathcal{K}$  satisfy the following inequality:

$$(20) \quad \phi_{j-1}(a + b) \leq \phi_{j-1} \circ \rho_{\phi,j}(a) + \phi_{j-1} \circ \rho_{v,j}(b).$$

Notice that in the case when  $\beta(\cdot, 0) \in \mathcal{K}_\infty$ , the functions  $\rho_{\phi,j}(\cdot), \rho_{v,j}(\cdot)$  will always exist [12]. The properties of sequence  $\Xi$  which ensure the desired propagation rate of the influence of initial condition  $\mathbf{x}_0$  and input  $h(\mathbf{z}_0)$  to the state  $\mathbf{x}(t)$  are specified in Property 3.

PROPERTY 3 (rate of contraction, part 2). *The sequences*

$$\sigma_n^{-1} \cdot \phi_n(\|\mathbf{x}_0\|_{\mathcal{A}}), \quad \sigma_n^{-1} \cdot \left( \sum_{i=0}^n v_i(c|h(\mathbf{z}_0)|\sigma_{n-i}) \right), \quad n = 0, \dots, \infty,$$

*are bounded from above; e.g., there exist functions  $B_1(\|\mathbf{x}_0\|), B_2(|h(\mathbf{z}_0)|, c)$  such that*

$$(21) \quad \sigma_n^{-1} \cdot \phi_n(\|\mathbf{x}_0\|_{\mathcal{A}}) \leq B_1(\|\mathbf{x}_0\|_{\mathcal{A}}),$$

$$(22) \quad \sigma_n^{-1} \cdot \left( \sum_{i=0}^n v_i(c|h(\mathbf{z}_0)|\sigma_{n-i}) \right) \leq B_2(|h(\mathbf{z}_0)|, c)$$

*for all  $n = 0, 1, \dots, \infty$ .*

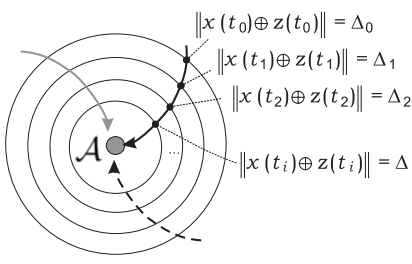
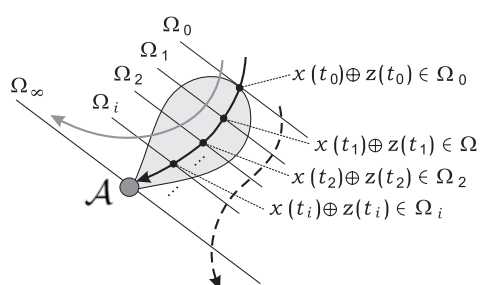
Standard	Proposed
1) Domain of attraction is a neighborhood  2) Implies Lyapunov stability  	1) Domain of attraction is a set of positive measure (not necessarily a neighborhood)  2) Allows us to analyze convergence in Lyapunov-unstable systems  
Given: a sequence of diverging time instances $t_i$	Given: a sequence of sets $\Omega_i$ whose distance $\Delta_i$ to $\mathcal{A}$ is converging to zero
Prove: convergence of norms $\ \mathbf{x}(t_i) \oplus \mathbf{z}(t_i)\  = \Delta_i$ to zero	Prove: divergence of $\{t_i\}$ , where $t_i : \mathbf{x}(t_i) \oplus \mathbf{z}(t_i) \in \Omega_i$

FIG. 3. Key differences between the conventional concept of convergence (left panel) and the concept of weak, nonuniform convergence (right panel). In the uniform case, trajectories which start in a neighborhood of  $\mathcal{A}$  remain in a neighborhood of  $\mathcal{A}$  (solid and dashed lines). In the nonuniform case, only a fraction of the initial conditions in a neighborhood of  $\mathcal{A}$  will produce trajectories which remain in a neighborhood of  $\mathcal{A}$  (solid black line). In the most general case a necessary condition for this to happen is that the sequence  $\{t_i\}$  diverges. In our current problem statement divergence of  $\{t_i\}$  implies boundedness of  $\|\mathbf{x}(t)\|_{\mathcal{A}}$ . To show state boundedness and convergence of  $\mathbf{x}(t)$  to  $\mathcal{A}$ , additional information on the system dynamics will be required.

For a large class of functions  $\beta(s, 0)$ , for instance those that are Lipschitz in  $s$ , these conditions reduce to more transparent ones which can always be satisfied by an appropriate choice of sequences  $\Xi$  and  $\mathcal{S}$ . This case is considered in detail as a corollary of our main results in section 4.3.

The main differences between the standard and the presently proposed approaches for the analysis of asymptotic behavior of dynamical systems are illustrated in Figure 3. In order to prove the emergence of the trapping region we consider the following collection of volumes induced by the sequence  $\mathcal{S}_i$  and the corresponding partition (15) of the interval  $[0, h(\mathbf{z}_0)]$ :

$$(23) \quad \Omega_i = \{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z} \mid h(\mathbf{z}(t)) \in H_i\}.$$

For the given initial conditions  $\mathbf{x}_0 \in \mathcal{X}$ ,  $\mathbf{z}_0 \in \mathcal{Z}$  two alternative possibilities exist. First, there exists an  $i$  such that the trajectory  $\mathbf{x}(t, \mathbf{x}_0) \oplus \mathbf{z}(t, \mathbf{z}_0)$  enters  $\Omega_i$  and stays there forever. Hence for  $t \rightarrow \infty$  the state will converge into

$$(24) \quad \Omega_a = \{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z} \mid \|\mathbf{x}\|_{\mathcal{A}} \leq c \cdot h(\mathbf{z}_0), \mathbf{z} : h(\mathbf{z}) \in [0, h(\mathbf{z}_0)]\}.$$

The second alternative is that for each  $i = 0, 1, \dots$  the trajectory  $\mathbf{x}(t, \mathbf{x}_0) \oplus \mathbf{z}(t, \mathbf{z}_0)$  enters  $\Omega_i$  and leaves some time later. Let  $t_i$  be the time instances when it hits

the hypersurfaces  $h(\mathbf{z}(t)) = h(\mathbf{z}_0)\sigma_i$ . Then the state of the coupled system stays in  $\cup_{i=0}^{\infty} \Omega_i$  only if the sequence  $\{t_i\}_{i=0}^{\infty}$  diverges. Theorem 3 provides sufficient conditions specifying the latter case in terms of the properties of sequences  $\mathcal{S}$ ,  $\Xi$ ,  $\mathcal{T}$  and function  $\gamma_0(\cdot)$  in (11). For a large class of interconnections (11) it is possible to formulate these conditions in terms of the input-output properties of systems  $\mathcal{S}_a$  and  $\mathcal{S}_w$  explicitly, i.e., in terms of functions  $\beta(\cdot, \cdot)$  and  $\gamma_0(\cdot)$  and the values of  $c$ . The results are presented as immediate corollaries of Theorem 3 in sections 4.3 and 5.1.

**THEOREM 3** (nonuniform small-gain theorem). *Let systems  $\mathcal{S}_a$ ,  $\mathcal{S}_w$  be given and satisfy Assumptions 1, 2. Consider their interconnection (11) and suppose there exist sequences  $\mathcal{S}$ ,  $\Xi$ , and  $\mathcal{T}$  satisfying Properties 1–3. In addition, suppose that the following conditions hold:*

1. *There exists a positive number  $\Delta_0 > 0$  such that*

$$(25) \quad \frac{1}{\tau_i} \frac{(\sigma_i - \sigma_{i+1})}{\gamma_{0,1}(\sigma_i)} \geq \Delta_0 \quad \forall i = 0, 1, \dots, \infty.$$

2. *The set  $\Omega_\gamma$  of all points  $\mathbf{x}_0, \mathbf{z}_0$  satisfying the inequality*

$$(26) \quad \gamma_{0,2}(B_1(\|\mathbf{x}_0\|_{\mathcal{A}}) + B_2(|h(\mathbf{z}_0)|, c) + c|h(\mathbf{z}_0)|) \leq h(\mathbf{z}_0)\Delta_0$$

*is not empty.*

3. *Partial sums of elements from  $\mathcal{T}$  diverge:*

$$(27) \quad \sum_{i=0}^{\infty} \tau_i = \infty.$$

*Then for all  $\mathbf{x}_0, \mathbf{z}_0 \in \Omega_\gamma$  the state  $\mathbf{x}(t, \mathbf{z}_0) \oplus \mathbf{z}(t, \mathbf{z}_0)$  of system (11) converges into the set specified by (24):*

$$\Omega_a = \{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z} \mid \|\mathbf{x}\|_{\mathcal{A}} \leq c \cdot h(\mathbf{z}_0), \mathbf{z} : h(\mathbf{z}) \in [0, h(\mathbf{z}_0)]\}.$$

The proofs of Theorem 3 and subsequent results are provided in the appendix.

The major difference between the conditions of Theorem 3 and those of conventional small-gain theorems [34], [12] is that the latter involve only input-output or input-state mappings. Formulating conditions for state boundedness of the interconnection in terms of input-output or input-state mappings is possible in the traditional case because the interconnected systems are assumed to be input-to-state stable. Hence their internal dynamics can be neglected. In our case, however, the dynamics of  $\mathcal{S}_w$  is generally unstable in the Lyapunov sense. Hence, in order to ensure boundedness of  $\mathbf{x}(t, \mathbf{x}_0)$  and  $h(\mathbf{z}(t, \mathbf{z}_0))$ , the rate/degree of stability of  $\mathcal{S}_a$  should be taken into account. Roughly speaking, system  $\mathcal{S}_a$  should ensure a sufficiently high degree of contraction in  $\mathbf{x}_0$  while the input-output response of  $\mathcal{S}_w$  should be sufficiently small. The rate of contraction in  $\mathbf{x}_0$  of  $\mathcal{S}_a$ , according to (6), is specified in terms of the function  $\beta(\cdot, \cdot)$ . Properties of this function that are relevant for convergence are explicitly accounted for in Property 3 and (27). The domain of admissible initial conditions, and actually the small-gain condition (input-state-output properties of  $\mathcal{S}_w$  and  $\mathcal{S}_a$ ), are defined by (25), (26), respectively. Notice also that  $\Omega_\gamma$  is not necessarily a neighborhood of  $\Omega_a$ ; thus the convergence ensured by Theorem 3 is allowed to be nonuniform in  $\mathbf{x}_0, \mathbf{z}_0$ .



**4.2. Characterization of the attracting set.** Even for interconnections of Lyapunov-stable systems, small-gain conditions usually are effective merely for establishing boundedness of states or outputs. Yet, even in the setting of Theorem 3 it is still possible to derive estimates (such as, for instance, (24)) of the domains to which the state will converge. These estimates, however, are often too conservative. If a more precise characterization of these domains is required, additional information on the dynamics of systems  $\mathcal{S}_a$  and  $\mathcal{S}_w$  will be needed. The question, therefore, is how detailed this information should be. It appears that some additional knowledge of the steady-state characteristics of system  $\mathcal{S}_a$  is sufficient to improve the estimates (24) substantially.

Let us formally introduce the notion of steady-state characteristic as follows.

**DEFINITION 4.** *We say that system (6) has steady-state characteristic  $\chi : \mathbb{R} \rightarrow \mathcal{S}\{\mathbb{R}_+\}$  with respect to the norm  $\|\mathbf{x}\|_{\mathcal{A}}$  iff for each constant  $\bar{u}_a$  the following holds:*

$$(28) \quad \forall u_a(t) \in \mathcal{U}_a : \lim_{t \rightarrow \infty} u_a(t) = \bar{u}_a \Rightarrow \lim_{t \rightarrow \infty} \|\mathbf{x}(t)\|_{\mathcal{A}} \in \chi(\bar{u}_a).$$

The key property captured by Definition 4 is that there exists a limit of  $\|\mathbf{x}(t)\|_{\mathcal{A}}$  as  $t \rightarrow \infty$ , provided that the limit for  $u_a(t)$ ,  $t \rightarrow \infty$ , is defined and constant. Notice that the mapping  $\chi$  is set-valued. This means that for each  $\bar{u}_a$  there is a set  $\chi(\bar{u}_a) \subset \mathbb{R}_+$  such that  $\|\mathbf{x}(t)\|_{\mathcal{A}}$  converges to an element of  $\chi(\bar{u}_a)$  as  $t \rightarrow \infty$ . Therefore, our definition allows a fairly large amount of uncertainty for  $\mathcal{S}_a$ . It will be of essential importance, however, that such a characterization exists for the system  $\mathcal{S}_a$ .

Clearly, not every system obeys a steady-state characteristic  $\chi(\cdot)$  of Definition 4. There are relatively simple systems of which the state does not converge even in the “norm” sense for constant converging inputs (condition (28)). In mechanics, physics, and biology such systems encompass the large class of nonlinear oscillators which can be excited by constant inputs. In order to take such systems into consideration, we introduce a weaker notion, that of a steady-state characteristic *on average*, defined as follows.

**DEFINITION 5.** *We say that system (6) has steady-state characteristic on average  $\chi_T : \mathbb{R} \rightarrow \mathcal{S}\{\mathbb{R}_+\}$  with respect to the norm  $\|\mathbf{x}\|_{\mathcal{A}}$  iff for each constant  $\bar{u}_a$  and some  $T > 0$  the following holds:*

$$(29) \quad \forall u_a(t) \in \mathcal{U}_a : \lim_{t \rightarrow \infty} u_a(t) = \bar{u}_a \Rightarrow \lim_{t \rightarrow \infty} \int_t^{t+T} \|\mathbf{x}(\tau)\|_{\mathcal{A}} d\tau \in \chi_T(\bar{u}_a).$$

Steady-state characterizations of system  $\mathcal{S}_a$  allow us to further specify the asymptotic behavior of interconnection (11). These results are summarized in Lemmas 6 and 7 below.

**LEMMA 6.** *Let system (11) be given and  $h(\mathbf{z}(t, \mathbf{z}_0))$  be bounded for some  $\mathbf{x}_0, \mathbf{z}_0$ . Let, furthermore, system (6) have steady-state characteristic  $\chi(\cdot) : \mathbb{R} \rightarrow \mathcal{S}\{\mathbb{R}_+\}$ . Then the following limiting relations hold:<sup>7</sup>*

$$(30) \quad \lim_{t \rightarrow \infty} \|\mathbf{x}(t, \mathbf{x}_0)\|_{\mathcal{A}} = 0, \quad \lim_{t \rightarrow \infty} h(\mathbf{z}(t, \mathbf{z}_0)) \in \chi^{-1}(0).$$

As follows from Lemma 6, in the case when the steady-state characteristic of  $\mathcal{S}_a$  is defined, the asymptotic behavior of interconnection (11) is characterized by the zeros

<sup>7</sup>The symbol  $\chi^{-1}(0)$  in (30) denotes the set  $\chi^{-1}(0) = \bigcup_{\bar{u}_a \in \mathbb{R}_+} \bar{u}_a : \chi(\bar{u}_a) \ni 0$ .

of the steady-state mapping  $\chi(\cdot)$ . For the steady-state characteristics on average a slightly modified conclusion can be derived.

LEMMA 7. *Let system (11) be given,  $h(\mathbf{z}(t, \mathbf{z}_0))$  be bounded for some  $\mathbf{x}_0, \mathbf{z}_0$ ,  $h(\mathbf{z}(t, \mathbf{z}_0)) \in [0, h(\mathbf{z}_0)]$ , and system (6) have steady-state characteristic  $\chi_T(\cdot) : \mathbb{R} \rightarrow \mathcal{S}\{\mathbb{R}_+\}$  on average. Furthermore, let there exist a positive constant  $\bar{\gamma}$  such that the function  $\gamma_1(\cdot)$  in (8) satisfies the following constraint:*

$$(31) \quad \gamma_1(s) \geq \bar{\gamma} \cdot s \quad \forall s \in [0, \bar{s}], \bar{s} \in \mathbb{R}_+ : \bar{s} > c \cdot h(\mathbf{z}_0).$$

*In addition, suppose that  $\chi_T(\cdot)$  has no zeros in the positive domain, i.e.,  $0 \notin \chi_T(\bar{u}_a)$  for all  $\bar{u}_a > 0$ . Then*

$$(32) \quad \lim_{t \rightarrow \infty} \|\mathbf{x}(t, \mathbf{x}_0)\|_{\mathcal{A}} = 0, \quad \lim_{t \rightarrow \infty} h(\mathbf{z}(t, \mathbf{z}_0)) = 0.$$

An immediate outcome of Lemmas 6 and 7 is that in the case when the conditions of Theorem 3 are satisfied and system (6) has steady-state characteristic  $\chi(\cdot)$  or  $\chi_T(\cdot)$ , the domain of convergence  $\Omega_a$  becomes

$$(33) \quad \Omega_a = \{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z} \mid \|\mathbf{x}\|_{\mathcal{A}} = 0, \mathbf{z} : h(\mathbf{z}) \in [0, h(\mathbf{z}_0)]\}.$$

It is possible, however, to improve estimate (33) further under additional hypotheses on system  $\mathcal{S}_a$  and  $\mathcal{S}_w$  dynamics. This result is formulated in the corollary below.

COROLLARY 8. *Let system (11) be given and satisfy the assumptions of Theorem 3. Let, in addition,*

(C1) *the flow  $\mathbf{x}(t, \mathbf{x}_0) \oplus \mathbf{z}(t, \mathbf{z}_0)$  be generated by a system of autonomous differential equations with a locally Lipschitz right-hand side;*

(C2) *subsystem  $\mathcal{S}_w$  be practically integral-input-to-state stable:*

$$(34) \quad \|\mathbf{z}(\tau)\|_{\infty, [t_0, t]} \leq C_z + \int_0^t \gamma_1(u_w(\tau)) d\tau,$$

*and let function  $h(\cdot) \in \mathcal{C}^0$  in (8);*

(C3) *system  $\mathcal{S}_a$  have steady-state characteristic  $\chi(\cdot)$ .*

*Then for all  $\mathbf{x}_0, \mathbf{z}_0 \in \Omega_\gamma$  the state of the interconnection converges to the set*

$$(35) \quad \Omega_a = \{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z} \mid \|\mathbf{x}\|_{\mathcal{A}} = 0, h(\mathbf{z}) \in \chi^{-1}(0)\}.$$

As follows from Corollary 8, zeros of the steady-state characteristic of system  $\mathcal{S}_a$  actually “control” the domains to which the state of interconnection (11) might potentially converge. This is illustrated in Figure 4. Notice also that in the case when condition C3 in Corollary 8 is replaced with the alternative,

(C3)' *system  $\mathcal{S}_a$  has a steady-state characteristic on average  $\chi_T(\cdot)$ ,*

*and then it is possible to show that the state converges to*

$$(36) \quad \Omega_a = \{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z} \mid \|\mathbf{x}\|_{\mathcal{A}} = 0, h(\mathbf{z}) = 0\}.$$

The proof follows straightforwardly from the proof of Corollary 8 and is therefore omitted.

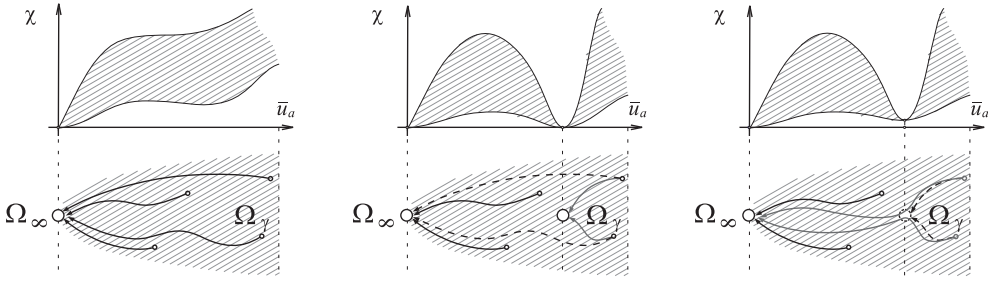


FIG. 4. Control of the attracting set by means of the system's steady-state characteristics.

**4.3. Systems with contracting dynamics separable in space-time.** In the previous sections we have presented convergence tests and estimates of the trapping region, and also characterized the attracting sets of interconnection (11) under assumptions of uniform asymptotic stability of  $\mathcal{S}_a$  and input-output properties (8), (34) of system  $\mathcal{S}_w$ . The conditions are given for rather general functions  $\beta(\cdot, \cdot) \in \mathcal{KL}$  in (6) and  $\gamma_0(\cdot)$ ,  $\gamma_1(\cdot)$  in (8). It appears, however, that these conditions can be substantially simplified if additional properties of  $\beta(\cdot, \cdot)$  and  $\gamma_0(\cdot)$  are available. This information is, in particular, the separability of function  $\beta(\cdot, \cdot)$  or, equivalently, the possibility of factorization:

$$(37) \quad \beta(\|\mathbf{x}\|_{\mathcal{A}}, t) \leq \beta_x(\|\mathbf{x}\|_{\mathcal{A}}) \cdot \beta_t(t),$$

where  $\beta_x(\cdot) \in \mathcal{K}$  and  $\beta_t(\cdot) \in \mathcal{C}^0$  is strictly decreasing<sup>8</sup> with

$$(38) \quad \lim_{t \rightarrow \infty} \beta_t(t) = 0.$$

In principle, as shown in [8], factorization (37) is achievable for a large class of uniformly asymptotically stable systems under an appropriate coordinate transformation. An immediate consequence of factorization (37) is that the elements of sequence  $\Xi$  in Property 2 are independent of  $\|\mathbf{x}(t_i)\|_{\mathcal{A}}$ . As a result, verification of Properties 2, 3 becomes easier. The most interesting case, however, occurs when the function  $\beta_x(\cdot)$  in the factorization (37) is Lipschitz. For this class of functions the conditions of Theorem 3 reduce to a single and easily verifiable inequality. Let us consider this case in detail.

Without loss of generality, we assume that the state  $\mathbf{x}(t)$  of system  $\mathcal{S}_a$  satisfies the equation

$$(39) \quad \|\mathbf{x}(t)\|_{\mathcal{A}} \leq \|\mathbf{x}(t_0)\|_{\mathcal{A}} \cdot \beta_t(t - t_0) + c \cdot \|h(\mathbf{z}(\tau, \mathbf{z}_0))\|_{\infty, [t_0, t]},$$

where  $\beta_t(0)$  is greater than or equal to one. Given that  $\beta_t(t)$  is strictly decreasing, the mapping  $\beta_t : [0, \infty) \mapsto [0, \beta_t(0)]$  is injective. Moreover  $\beta_t(t)$  is continuous, and then it is surjective and, therefore, bijective. In other words there is a (continuous) mapping  $\beta_t^{-1} : [0, \beta_t(0)] \mapsto \mathbb{R}_+$ :

$$(40) \quad \beta_t^{-1} \circ \beta_t(t) = t \quad \forall t > 0.$$

Conditions for emergence of the trapping region for interconnection (11) with dynamics of system  $\mathcal{S}_a$  governed by (39) are summarized below:

<sup>8</sup>If  $\beta_t(\cdot)$  is not strictly monotone, it can always be majorized by a strictly decreasing function.

COROLLARY 9. *Let the interconnection (11) be given, system  $\mathcal{S}_a$  satisfy (39), and function  $\gamma_0(\cdot)$  in (8) be Lipschitz:*

$$(41) \quad |\gamma_0(s)| \leq D_{\gamma,0} \cdot |s|.$$

*Domain*

$$(42) \quad \Omega_\gamma : D_{\gamma,0} \leq \left( \beta_t^{-1} \left( \frac{d}{\kappa} \right) \right)^{-1} \frac{\kappa - 1}{\kappa} \times \frac{h(\mathbf{z}_0)}{\beta_t(0) \|\mathbf{x}_0\|_{\mathcal{A}} + \beta_t(0) \cdot c \cdot |h(\mathbf{z}_0)| \left( 1 + \frac{\kappa}{1-d} \right) + c|h(\mathbf{z}_0)|}$$

is not empty for some  $d < 1$ ,  $\kappa > 1$ . Then for all initial conditions  $\mathbf{x}_0, \mathbf{z}_0 \in \Omega_\gamma$  the state  $\mathbf{x}(t, \mathbf{x}_0) \oplus \mathbf{z}(t, \mathbf{z}_0)$  of interconnection (11) converges into the set  $\Omega_a$  specified by (24). If, in addition, conditions (C1)–(C3) of Corollary 8 hold, then the domain of convergence is given by (33).

A practically important consequence of this corollary concerns systems  $\mathcal{S}_a$  which are exponentially stable:

$$(43) \quad \|\mathbf{x}(t)\|_{\mathcal{A}} \leq \|\mathbf{x}(t_0)\|_{\mathcal{A}} D_\beta \exp(-\lambda t) + c \cdot \|h(\mathbf{z}(t, \mathbf{z}_0))\|_{\infty, [t_0, t]}, \lambda > 0, D_\beta \geq 1.$$

In this case the domain (42) of initial conditions ensuring convergence into  $\Omega_a$  is defined as

$$D_{\gamma,0} \leq \max_{\kappa > 1, d \in (0,1)} -\lambda \left( \ln \frac{d}{\kappa} \right)^{-1} \frac{\kappa - 1}{\kappa} \times \frac{h(\mathbf{z}_0)}{D_\beta \|\mathbf{x}_0\|_{\mathcal{A}} + D_\beta \cdot c \cdot |h(\mathbf{z}_0)| \left( 1 + \frac{\kappa}{1-d} \right) + c|h(\mathbf{z}_0)|}.$$

**5. Discussion.** In this section we discuss some practically relevant outcomes of the results of Theorem 3 and Corollaries 8, 9 and their potential applications to problems of analysis of asymptotic behavior in nonlinear dynamic systems.

First, in section 5.1 we specify conditions for existence of a trapping region of nonzero volume in  $\mathbb{R}^n \oplus \mathbb{R}^m$  in terms of the parameters of system (11) without invoking dependence on  $\mathbf{x}(t_0), \mathbf{z}(t_0)$ , as was done in Theorem 3. The resulting criterion has a form similar to the standard small-gain conditions [34]. The differences and similarities between this new result and standard small-gain theorems are illustrated with an example.

Second, in section 5.2 we demonstrate how the results of our present contribution can be applied to address the problem of output nonlinear identification for systems which cannot be transformed into a canonic observer form and/or with nonlinear parametrization.

**5.1. Relation to conventional small-gain theorems.** Conditions specifying state boundedness formulated in Theorem 3 and Corollaries 8, 9 depend explicitly on initial conditions  $\mathbf{x}(t_0), \mathbf{z}(t_0)$ . Such dependence is inevitable when the convergence is allowed to be nonuniform. But if the mere existence of a trapping region is asked for, dependence on initial conditions may be removed from the statements of the results. The next corollary presents such modified conditions.

COROLLARY 10. *Consider interconnection (11), where the system  $\mathcal{S}_a$  satisfies inequality (39) and the function  $\gamma_0(\cdot)$  obeys (41). Then there exists a set  $\Omega_\gamma$  of initial*

conditions corresponding to the trajectories converging to  $\Omega_a$  if the following condition is satisfied:

$$(44) \quad D_{\gamma,0} \cdot c \cdot \mathcal{G} < 1,$$

where

$$\mathcal{G} = \beta_t^{-1} \left( \frac{d}{\kappa} \right) \frac{k}{k-1} \left( \beta_t(0) \left( 1 + \frac{\kappa}{1-d} \right) + 1 \right)$$

for some  $d \in (0, 1)$ ,  $\kappa \in (1, \infty)$ . In particular,  $\Omega_\gamma$  contains the following domain:

$$\|\mathbf{x}(t_0)\|_{\mathcal{A}} \leq \frac{h(\mathbf{z}(t_0))}{\beta_t(0)} \left[ \frac{1}{D_{\gamma,0}} \left( \beta_t^{-1} \left( \frac{d}{\kappa} \right) \right)^{-1} \frac{k-1}{k} - c \left( \beta_t(0) \left( 1 + \frac{\kappa}{1-d} \right) + 1 \right) \right].$$

In the case when the function  $h(\mathbf{z})$  in (11) is continuous, the volume of the set  $\Omega_\gamma$  is nonzero in  $\mathbb{R}^n \oplus \mathbb{R}^m$ .

Notice that in the case when the dynamics of the contracting subsystem  $\mathcal{S}_a$  is exponentially stable, i.e., it satisfies inequality (43), the term  $\mathcal{G}$  in condition (44) reduces to

$$(45) \quad \mathcal{G} = \frac{1}{\lambda} \cdot \ln \left( \frac{\kappa}{d} \right) \frac{k}{k-1} \left( D_\beta \left( 1 + \frac{\kappa}{1-d} \right) + 1 \right).$$

For  $D_\beta = 1$  the minimal value of  $\mathcal{G}$  in (45) can be estimated as

$$(46) \quad \mathcal{G}^* = \frac{1}{\lambda} \cdot \min_{d \in (0,1), \kappa \in (1,\infty)} \ln \left( \frac{\kappa}{d} \right) \frac{k}{k-1} \left( 2 + \frac{\kappa}{1-d} \right) \approx \frac{15.6886}{\lambda} < \frac{16}{\lambda},$$

which leads to an even more simple formulation of (45):

$$D_{\gamma,0} \cdot \frac{c}{\lambda} \leq \frac{1}{16}.$$

Corollary 10 provides an explicit and easy-to-check condition for existence of a trapping region in the state space of a class of Lyapunov unstable systems. In addition, it allows us to specify explicitly points  $\mathbf{x}(t_0)$ ,  $\mathbf{z}(t_0)$  which belong to the emergent trapping region. Notice also that the existence condition, inequality (44), has the flavor of conventional small-gain constraints. Yet, it is substantially different from these classical results. This is because the input-output gain for the wandering subsystem,  $\mathcal{S}_w$ , may not be finite or need not even be defined.

To elucidate these differences as well as the similarities between conditions of conventional small-gain theorems and those formulated in Corollary 10 we provide an example. Consider the following systems:

$$(47a) \quad \begin{cases} \dot{x}_1 = -\lambda_1 x_1 + c_1 x_2, \\ \dot{x}_2 = -\lambda_2 x_2 - c_2 |x_1|, \end{cases}$$

$$(47b) \quad \begin{cases} \dot{x}_1 = -\lambda_1 x_1 + c_1 x_2, \\ \dot{x}_2 = -c_2 |x_1|. \end{cases}$$

System (47a) can be viewed as an interconnection of two input-to-state stable systems,  $x_1$  and  $x_2$ , with input-output  $L_\infty$ -gains  $c_1/\lambda_1$  and  $c_2/\lambda_2$ , respectively. Therefore, in order to prove state boundedness of (47a) we can, in principle, invoke the conventional small-gain theorem. The small-gain condition in this case is as follows:

$$(48a) \quad \frac{c_1}{\lambda_1} \cdot \frac{c_2}{\lambda_2} < 1.$$

The theorem, however, does not apply to system (47b) because the input-output gain of its second subsystem,  $x_2$ , is infinite. Yet, by invoking Corollary 10 it is still possible to show existence of a weak attracting set in the state space of system (47b) and specify its basin of attraction. As follows from Corollary 10, condition

$$(48b) \quad \frac{c_1}{\lambda_1} \cdot \frac{c_2}{\lambda_1} < \frac{1}{16}$$

ensures existence of the trapping region, and the trapping region itself is given by

$$|x_1(t_0)| \leq \left[ \frac{1}{c_2} \lambda_1 \left( \ln \frac{\kappa}{d} \right)^{-1} \frac{k-1}{k} - \frac{c_1}{\lambda_1} \left( 2 + \frac{\kappa}{1-d} \right) \right] x_2(t_0).$$

**5.2. Output nonlinear identification problem.** In the literature on adaptive control, observation, and identification a few classes of systems are referred to as *canonic forms* because they guarantee existence of a solution to the problem and because a large variety of physical models can be transformed into this class. Among these, perhaps the most widely known is the *adaptive observer canonical form* [3]. Necessary and sufficient conditions for transformation of the original system into this canonical form can be found, for example, in [16]. These conditions, however, include restrictive requirements of linearization of uncertainty-independent dynamics by output injection, and they also require linear parametrization of the uncertainty. Alternative approaches [4] heavily rely on knowledge of the proper Lyapunov function for the uncertainty-independent part and still assume linear parametrization.

We now demonstrate how these restrictions can be lifted by application of our result to the problem of state and parameter observation. Let us consider systems which can be transformed by means of static or dynamic feedback<sup>9</sup> into the following form:

$$(49) \quad \dot{\mathbf{x}} = \mathbf{f}_0(\mathbf{x}, t) + \mathbf{f}(\boldsymbol{\xi}(t), \boldsymbol{\theta}) - \mathbf{f}(\boldsymbol{\xi}(t), \hat{\boldsymbol{\theta}}) + \boldsymbol{\varepsilon}(t),$$

where

$$\boldsymbol{\varepsilon}(t) \in L_\infty^m[t_0, \infty], \quad \|\boldsymbol{\varepsilon}(\tau)\|_{\infty, [t_0, t]} \leq \Delta_\varepsilon$$

is an external perturbation with known  $\Delta_\varepsilon$ , and  $\mathbf{x} \in \mathbb{R}^n$ . The function  $\boldsymbol{\xi} : \mathbb{R}_+ \rightarrow \mathbb{R}^\xi$  is a function of time, which possibly includes available measurements of the state, and  $\boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \in \Omega_\theta \subset \mathbb{R}^d$  are the unknown and estimated parameters of the function  $\mathbf{f}(\cdot)$ , respectively, and the set  $\Omega_\theta$  is bounded. We assume that uniformly in  $\boldsymbol{\xi}$ , the function  $\mathbf{f}(\boldsymbol{\xi}(t), \boldsymbol{\theta})$  is locally bounded in  $\boldsymbol{\theta}$ :

$$\|\mathbf{f}(\boldsymbol{\xi}(t), \boldsymbol{\theta}) - \mathbf{f}(\boldsymbol{\xi}(t), \hat{\boldsymbol{\theta}})\| \leq D_f \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| + \Delta_f$$

<sup>9</sup>Notice that conventional observers in control theory could be viewed as dynamic feedbacks.

and the values of  $D_f \in \mathbb{R}_+$ ,  $\Delta_f$  are available. The function  $\mathbf{f}_0(\cdot)$  in (49) is assumed to satisfy the following condition.

ASSUMPTION 3. *The system*

$$(50) \quad \dot{\mathbf{x}} = \mathbf{f}_0(\mathbf{x}, t) + \mathbf{u}(t)$$

is forward-complete. Furthermore, for all  $\mathbf{u}(t)$  such that

$$\|\mathbf{u}(t)\|_{\infty, [t_0, t]} \leq \Delta_u + \|\mathbf{u}_0(\tau)\|_{\infty, [t_0, t]}, \quad \Delta_u \in \mathbb{R}_+,$$

there exists a bounded set  $\mathcal{A}$ ,  $c > 0$  and a function  $\Delta : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  satisfying the following inequality:

$$\|\mathbf{x}(t)\|_{\mathcal{A}_{\Delta(\Delta_u)}} \leq \beta(t - t_0) \|\mathbf{x}(t_0)\|_{\mathcal{A}_{\Delta(\Delta_u)}} + c \|\mathbf{u}_0(\tau)\|_{\infty, [t_0, t]},$$

where  $\beta(\cdot) : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ ,  $\lim_{t \rightarrow \infty} \beta(t) = 0$  is a strictly decreasing function.

Consider the following auxiliary system:

$$(51) \quad \dot{\lambda} = S(\lambda), \quad \lambda(t_0) = \lambda_0 \in \Omega_\lambda \subset \mathbb{R}^\lambda,$$

where  $\Omega_\lambda \subset \mathbb{R}^\lambda$  is a compact set,  $\lambda(t, \lambda_0) \in \Omega_\lambda$  for all  $t \geq t_0$ , and  $S(\lambda)$  is locally Lipschitz. Furthermore, suppose that the following assumption holds for system (51).

ASSUMPTION 4. *System (51) is Poisson stable in  $\Omega_\lambda$ , that is,*

$$\forall \lambda' \in \Omega_\lambda, \quad t' \in \mathbb{R}_+ \Rightarrow \exists t'' > t : \|\lambda(t'', \lambda') - \lambda'\| \leq \epsilon,$$

where  $\epsilon$  is an arbitrary small positive constant. Moreover, the trajectory  $\lambda(t, \lambda_0)$  is dense in  $\Omega_\lambda$ :

$$\forall \lambda' \in \Omega_\lambda, \quad \epsilon \in \mathbb{R}_{>0} \Rightarrow \exists t \in \mathbb{R}_+ : \|\lambda' - \lambda(t, \lambda_0)\| < \epsilon.$$

Now we are ready to formulate the following statement.

COROLLARY 11. *Consider system (49) and suppose that the following conditions hold:*

- (C4) *the vector-field  $\mathbf{f}_0(\mathbf{x}, t)$  in (49) satisfies Assumption 3;*
- (C5) *there exists a (known) system (51) satisfying Assumption 4;*
- (C6) *there exists a locally Lipschitz  $\eta : \mathbb{R}^\lambda \rightarrow \mathbb{R}^d$ :*

$$\|\eta(\lambda') - \eta(\lambda'')\| \leq D_\eta \|\lambda' - \lambda''\|$$

such that the set  $\eta(\Omega_\lambda)$  is dense in  $\Omega_\theta$ ;

- (C7) *system (49) has a steady-state characteristic with respect to the norm*

$$\|\cdot\|_{\mathcal{A}_{\Delta(M)}}, \quad M = 2\Delta_f + \Delta_\varepsilon + \delta,$$

and input  $\hat{\theta}$ , where  $\delta$  is some positive (arbitrarily small) constant.

Consider the following interconnection of (49), (51):

$$(52) \quad \begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}_0(\mathbf{x}, t) + \mathbf{f}(\xi(t), \theta) - \mathbf{f}(\xi(t), \hat{\theta}) + \varepsilon(t), \\ \dot{\hat{\theta}} &= \eta(\lambda), \\ \dot{\lambda} &= \gamma \|\mathbf{x}(t)\|_{\mathcal{A}_{\Delta(M)}} S(\lambda), \end{aligned}$$

where  $\gamma > 0$  satisfies the following inequality:

$$(53) \quad \gamma \leq \left( \beta_t^{-1} \left( \frac{d}{\kappa} \right) \right)^{-1} \frac{\kappa - 1}{\kappa} \frac{1}{D_\lambda \left( \beta_t(0) \left( 1 + \frac{\kappa}{1-d} \right) + 1 \right)},$$

$$D_\lambda = c \cdot D_f \cdot D_\eta \cdot \max_{\lambda \in \Omega_\lambda} \|S(\lambda)\|$$

for some  $d \in (0, 1)$ ,  $\kappa \in (1, \infty)$ . Then, for  $\lambda(t_0) = \lambda_0$ , some  $\theta' \in \Omega_\theta$ , and all  $\mathbf{x}(t_0) = \mathbf{x}_0 \in \mathbb{R}^n$ , the following holds:

$$(54) \quad \lim_{t \rightarrow \infty} \|\mathbf{x}(t)\|_{\mathcal{A}_{\Delta(M)}} = 0, \quad \lim_{t \rightarrow \infty} \hat{\theta}(t) = \theta' \in \Omega_\theta.$$

Notice that, as has been pointed out in the previous section, in the case when the dynamics of (50) is exponentially stable with a rate of convergence equal to  $\rho$  and  $\beta(0) = D_\beta$ , condition (53) will have the following form:

$$\gamma \leq -\rho \left( \ln \frac{d}{\kappa} \right)^{-1} \frac{\kappa - 1}{\kappa} \frac{1}{D_\lambda \left( D_\beta \left( 1 + \frac{\kappa}{1-d} \right) + 1 \right)}.$$

According to Corollary 11, for the rather general class of systems (49) it is possible to design an estimator  $\hat{\theta}(t)$  which guarantees not only that the “error” vector  $\mathbf{x}(t)$  reaches a neighborhood of the origin, but also that the estimates  $\hat{\theta}(t)$  converge to some  $\theta'$  in  $\Omega_\theta$ . Both these facts, together with additional nonlinear persistent excitation conditions [6], [29]

$$\begin{aligned} & \exists T > 0, \rho \in \mathcal{K} : \forall \mathcal{T} = [t, t + T], t \in \mathbb{R}_+ \\ & \Rightarrow \exists \tau \in \mathcal{T} : |\mathbf{f}(\xi(\tau), \theta) - \mathbf{f}(\xi(\tau), \theta')| \geq \rho(\|\theta - \theta'\|), \end{aligned}$$

in principle allow us to estimate the domain of convergence for  $\hat{\theta}(t)$ .

Concluding this section we mention that statements of Theorem 3 and Corollaries 8–11 constitute additional theoretical tools for the analysis of asymptotic behavior of systems in cascaded form. In particular they are complementary to the results of [1], where *asymptotic stability* of systems of the following type:

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}), \\ \dot{\mathbf{z}} &= \mathbf{q}(\mathbf{x}, \mathbf{z}), \quad \mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad \mathbf{q} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^m \end{aligned}$$

was considered under the assumption that the  $\mathbf{x}$ -subsystem is globally asymptotically stable and the  $\mathbf{z}$ -subsystem is integral input-to-state stable. In contrast to this, our results apply to establishing *asymptotic convergence* for systems with the following structure:

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}, \mathbf{z}), \\ \dot{\mathbf{z}} &= \mathbf{q}(\mathbf{x}, \mathbf{z}), \quad \mathbf{f} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n, \end{aligned}$$

where the  $\mathbf{x}$ -subsystem is input-to-state stable, and the  $\mathbf{z}$ -subsystem could be practically integral input-to-state stable (see Corollary 8), although in general no stability assumptions are imposed on it.



**6. Examples.** In this section we provide two examples of parameter identification in nonlinearly parametrized systems that cannot be transformed into the canonical adaptive observer form.

The first example is merely an academical illustration of Corollary 11, where only one parameter is unknown and the system itself is a first-order differential equation. The second example illustrates a possible application of our results to the problem of identifying the dynamics in living cells.

*Example 1.* Consider the following system:

$$(55) \quad \dot{x} = -kx + \sin(x\theta + \theta) + u, \quad k > 0, \quad \theta \in [-a, a],$$

where  $\theta$  is an unknown parameter and  $u$  is the control input. Without loss of generality we let  $a = 1, k = 1$ . The problem is to estimate the parameter  $\theta$  from measurements of  $x$  and steer the system to the origin. Clearly, the choice  $u = -\sin(x\hat{\theta} + \hat{\theta})$  transforms (55) into

$$(56) \quad \dot{x} = -kx + \sin(x\theta + \theta) - \sin(x\hat{\theta} + \hat{\theta}),$$

which satisfies Assumption 3. Moreover, the system

$$\begin{aligned} \dot{\lambda}_1 &= \lambda_1, \\ \dot{\lambda}_2 &= -\lambda_2, \quad \lambda_1^2(t_0) + \lambda_2^2(t_0) = 1 \end{aligned}$$

with mapping  $\eta = (1, 0)^T \lambda$  satisfies Assumption 4 and therefore

$$(57) \quad \begin{aligned} \dot{\lambda}_1 &= \gamma|x|\lambda_1, \\ \dot{\lambda}_2 &= -\gamma|x|\lambda_2, \quad \lambda_1^2(t_0) + \lambda_2^2(t_0) = 1 \end{aligned}$$

would be a candidate for the control and parameter estimation algorithm. According to Corollary 11, the goal will be reached if the parameter  $\gamma$  in (57) obeys the following constraint:

$$\gamma \leq -\rho \left( \ln \frac{d}{\kappa} \right)^{-1} \frac{\kappa - 1}{\kappa} \frac{1}{D_\lambda \left( D_\beta \left( 1 + \frac{\kappa}{1-d} \right) + 1 \right)}, \quad \rho = k = 1, \quad D_\beta = 1, \quad D_\lambda = 1$$

for some  $d \in (0, 1), \kappa \in (1, \infty)$ . Hence, choosing, for example,  $d = 0.5, \kappa = 2$  we obtain that choice

$$0 < \gamma < -\ln \left( \frac{0.5}{2} \right)^{-1} \frac{1}{2} \cdot \frac{1}{6} = 0.0601$$

suffices to ensure that

$$\lim_{t \rightarrow \infty} x(t) = 0, \quad \lim_{t \rightarrow \infty} \hat{\theta}(t) = \theta.$$

We simulated system (56), (57) with  $\theta = 0.3, \gamma = 0.05$  and initial conditions  $x(t_0)$  randomly distributed in the interval  $[-1, 1]$ . Results of the simulation are illustrated in Figure 5, where the phase plots of system (56), (57) as well as the trajectories of  $\hat{\theta}(t)$  are given.

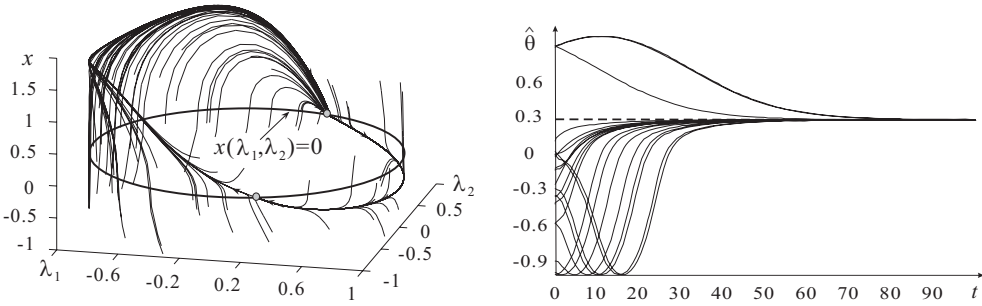


FIG. 5. Trajectories of system (56), (57) (left panel) and the family of estimates  $\hat{\theta}(t)$  of parameter  $\theta$  as functions of time  $t$  (right panel).

*Example 2.* Consider the problem of modeling electrical activity in biological cells from the input-output data in current clamp experiments. The simplest mathematical model, which captures a fairly large variety of phenomena such as periodic bursting in response to constant stimulation, is the classical Hindmarsh and Rose model neuron without adaptation currents [10]:

$$(58) \quad \begin{aligned} \dot{x}_1 &= -ax_1^3 + bx_1^2 + x_2 + \alpha u, \\ \dot{x}_2 &= c - \beta x_2 - dx_1^2, \end{aligned}$$

where variable  $x_1$  is the membrane potential,  $x_2$  stands for the ionic currents in the cell,  $u$  is the input current, and  $a, b, c, d, \alpha, \beta \in \mathbb{R}$  are parameters. While the parameters of the first equation can, in principle, be identified experimentally by blocking the ionic channels in the cells and measuring the membrane conductance, identification of parameters  $\beta, d$  is a difficult problem, as information about ionic currents  $x_2$  is rarely available.

Conventional techniques [3] cannot be applied directly to this problem as the model (58) is not in canonical adaptive observer form. Let us illustrate how our results can be used to derive the unknown parameters of (58) such that the reconstructed model fits the observed data. Assume, first, that parameters  $a, b, c, \alpha$  in the first equation of (58) are known, whereas parameters  $\beta, d$  in the second equation are unknown. This corresponds to the realistic case, where the time constant of current  $x_2$  and coupling between  $x_1$  and  $x_2$  are uncertain. In our example we assumed that

$$\beta \in \Omega_\beta = [0.3, 0.7], \quad d \in \Omega_d = [2, 3], \quad a = 1, \quad b = 3, \quad \alpha = 0.7, \quad c = 0.5.$$

As a candidate for the observer we select the following system:

$$(59) \quad \dot{\hat{x}} = \rho(x_1 - \hat{x}) - ax_1^3 + bx_1^2 + \alpha u + f(\hat{\beta}, \hat{d}, t), \quad \rho \in \mathbb{R}_{>0},$$

where  $\hat{\beta}, \hat{d}$  are parameters to be adjusted and the function  $f(\hat{\beta}, \hat{d}, t)$  is specified as

$$f(\hat{\beta}, \hat{d}, t) = \int_{t_0}^t e^{-\hat{\beta}(t-\tau)} (\hat{d}x_1^2(\tau) + c) d\tau.$$

Then the dynamics of  $\tilde{x}(t) = x(t) - \hat{x}(t)$  satisfies the following differential equation:

$$\dot{\tilde{x}} = -\rho\tilde{x} + f(\beta, d, t) - f(\hat{\beta}, \hat{d}, t).$$

The function  $f(\beta, d, t)$  satisfies the following inequality:

$$\begin{aligned} |f(\beta, d, t) - f(\hat{\beta}, \hat{d}, t)| &\leq |f(\beta, d, t) - f(\hat{\beta}, d, t)| + |f(\hat{\beta}, d, t) - f(\hat{\beta}, \hat{d}, t)| \\ &\leq D_{f,\beta}|\beta - \hat{\beta}| + D_{f,d}|d - \hat{d}| + \epsilon(t), \end{aligned}$$

where  $\epsilon(t)$  is an exponentially decaying term, and

$$(60) \quad D_{f,\beta} = \max_{\hat{\beta}, \beta \in \Omega_\beta, d \in \Omega_d} \left\{ \frac{1}{\beta \hat{\beta}} (d \|x_1(\tau)\|_{\infty, [t_0, \infty]} + c) \right\}, \quad D_{f,d} = \max_{\hat{\beta} \in \Omega_\beta} \left\{ \frac{1}{\hat{\beta}} \|x_1(\tau)\|_{\infty, [t_0, \infty]} \right\}.$$

Furthermore, Assumption 3 is satisfied for system

$$(61) \quad \dot{\tilde{x}} = -\rho \tilde{x} + v(t),$$

with

$$\Delta(\Delta_u) = \frac{\Delta_u}{\rho}.$$

In particular, for all  $v(t) : \|v(\tau)\|_{\infty, [t_0, t]} \leq \Delta_u + \|v_0(\tau)\|_{\infty, [t_0, t]}$  the following inequality holds:

$$(62) \quad \|\tilde{x}(t)\|_{\Delta(\Delta_u)} \leq e^{-\rho(t-t_0)} \|\tilde{x}(t_0)\|_{\Delta(\Delta_u)} + \frac{1}{\rho} \|v_0(\tau)\|_{\infty, [t_0, t]}.$$

To see this consider the general solution of (61):

$$\tilde{x}(t) = e^{-\rho(t-t_0)} \tilde{x}(t_0) + e^{-\rho t} \int_{t_0}^t e^{\rho \tau} v(\tau) d\tau$$

and derive an estimate of  $|\tilde{x}(t)|$ . This estimate has the following form:

$$\begin{aligned} |\tilde{x}(t)| &\leq e^{-\rho(t-t_0)} |\tilde{x}(t_0)| + \frac{1}{\rho} (1 - e^{-\rho(t-t_0)}) \|v(\tau)\|_{\infty, [t_0, t]} \\ &\leq e^{-\rho(t-t_0)} \left( |\tilde{x}(t_0)| - \frac{1}{\rho} \Delta_u \right) + \frac{1}{\rho} (\|v_0(\tau)\|_{\infty, [t_0, t]} + \Delta_u) \\ &\leq e^{-\rho(t-t_0)} \|\tilde{x}(t_0)\|_{\Delta(\Delta_u)} + \frac{1}{\rho} (\|v_0(\tau)\|_{\infty, [t_0, t]} + \Delta_u). \end{aligned}$$

Hence

$$|\tilde{x}(t)| - \frac{1}{\rho} \Delta_u \leq e^{-\rho(t-t_0)} \|\tilde{x}(t_0)\|_{\Delta(\Delta_u)} + \frac{1}{\rho} \|v_0(\tau)\|_{\infty, [t_0, t]},$$

which automatically implies (62).

Let us define subsystem (51). Consider the following system of differential equations:

$$(63) \quad \begin{aligned} \dot{\lambda}_1 &= \lambda_2, \\ \dot{\lambda}_2 &= -\omega_1^2 \lambda_1, \\ \dot{\lambda}_3 &= \lambda_4, \\ \dot{\lambda}_4 &= -\omega_2^2 \lambda_3, \quad \lambda_0 = (1, 0, 1, 0)^T, \end{aligned}$$

where  $\Omega_\lambda$  is the  $\omega$ -limit set of the point  $\lambda_0$ , and  $\omega_1, \omega_2 \in \mathbb{R}$ . System (63), therefore, satisfies Assumption 4. Given that domains  $\Omega_\beta, \Omega_d$  are known, select

(64)

$$\boldsymbol{\eta} : \mathbb{R}^n \rightarrow \mathbb{R}^2, \quad \boldsymbol{\eta} = (\eta_1(\boldsymbol{\lambda}), \eta_2(\boldsymbol{\lambda})),$$

$$\hat{\beta} = \eta_1(\boldsymbol{\lambda}) = \frac{1}{2} \left( \frac{2 \arcsin(\lambda_1)}{\pi} + 1 \right) \cdot 0.4 + 0.3, \quad \hat{d} = \eta_2(\boldsymbol{\lambda}) = \frac{1}{2} \left( \frac{2 \arcsin(\lambda_3)}{\pi} + 1 \right) + 2.$$

Choosing

$$\frac{\omega_1}{\omega_2} = \pi$$

we ensure that  $\boldsymbol{\eta}(\Omega_\lambda)$  is dense in  $\Omega_\beta \times \Omega_d$ . Given that  $\hat{\beta}, \hat{d}$  are bounded and  $\hat{\beta} \geq 0.3$ , the values of  $D_{f,\beta}, D_{f,d}$  are bounded since the signal  $x_1(t)$  is always bounded for any  $t \geq t_0$  for the given range of parameters. Hence, according to Corollary 11, interconnection of (59), (64), and

$$\begin{aligned} \dot{\lambda}_1 &= \gamma \|\tilde{x}(t)\|_{\Delta(\delta)} \cdot \lambda_2, \\ \dot{\lambda}_2 &= -\gamma \|\tilde{x}(t)\|_{\Delta(\delta)} \cdot \omega_1^2 \lambda_1, \\ \dot{\lambda}_3 &= \gamma \|\tilde{x}(t)\|_{\Delta(\delta)} \cdot \lambda_4, \\ \dot{\lambda}_4 &= -\gamma \|\tilde{x}(t)\|_{\Delta(\delta)} \cdot \omega_2^2 \lambda_3, \quad \boldsymbol{\lambda}_0 = (1, 0, 1, 0)^T, \end{aligned}$$

with arbitrary small  $\delta > 0$  and properly chosen  $\gamma > 0$ , ensures that

$$\lim_{t \rightarrow \infty} \|\tilde{x}(t)\|_{\Delta(\delta)} = 0, \quad \lim_{t \rightarrow \infty} \hat{\beta}(t) = \beta' \in \Omega_\beta, \quad \lim_{t \rightarrow \infty} \hat{d}(t) = d' \in \Omega_d.$$

This in turn implies a successful fit of the model to the observations.

We simulated the system with  $\rho = 10$  and  $\gamma = 3 \cdot 10^{-4}$  for  $\beta = 0.5, d = 2.5$ . The results of the simulations are provided in Figure 6. It can be seen from this figure that the reconstruction is successful and that the parameters converge into a small neighborhood of the actual values. Further details explaining how this technique can be applied to model the dynamics of the evoked membrane potentials in real neural cells from input-output measurements in vitro are discussed in [25].

**7. Conclusion.** We proposed tools for the analysis of asymptotic behavior of a class of dynamical systems. In particular, we consider an interconnection of an input-to-state stable system with an unstable or integral input-to-state dynamics. Our results allow us to address a variety of problems in which convergence may not be uniform with respect to initial conditions. It is necessary to notice that the proposed method does not require complete knowledge of the dynamical systems in question. Only qualitative information such as, for instance, characterization of input-to-state stability is necessary for application of our results. We demonstrated how our analysis can be used in the problems of synthesis and design—in particular for problems of nonlinear regulation and parameter identification of nonlinear parametrized systems. The examples show the relevance of our approach in those domains where application of the standard techniques is either not possible or too complicated.

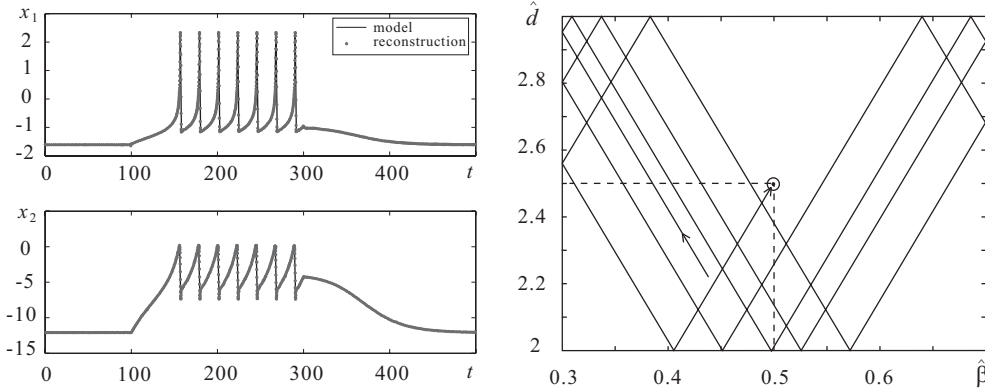


FIG. 6. Left panel: trajectories  $x_1(t)$ ,  $x_2(t)$  of system (58) plotted for the nominal values of parameters  $\beta = 0.5$ ,  $d = 2.5$  (model), and for the values  $\beta = \hat{\beta}(t_0 + T)$ ,  $d = \hat{d}(t_0 + T)$ , where  $T$  is the total simulation time (reconstruction). Input  $u(t)$  is a rectangular impulse with amplitude 0.7 starting at  $t = 100$  and ending at  $t = 300$ . Right panel: searching dynamics in the bounded parameter space (a segment of the trajectory  $\hat{\beta}(t)$ ,  $\hat{d}(t)$  towards the end of the simulation).

## Appendix. Proofs of Theorem 3, lemmas, and corollaries.

**A.1. Proof of Theorem 3.** Let the conditions of the theorem be satisfied for given  $t_0 \in \mathbb{R}_+$ :  $\mathbf{x}(t_0) = \mathbf{x}_0$ ,  $\mathbf{z}(t_0) = \mathbf{z}_0$ . Notice that in this case  $h(\mathbf{z}_0) \geq 0$ ; otherwise requirement (26) will be violated. Consider the sequence (23) of volumes  $\Omega_i$  induced by  $\mathcal{S}$ :

$$\Omega_i = \{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z} \mid h(\mathbf{z}(t)) \in H_i\}.$$

To prove the theorem we show that  $0 \leq h(\mathbf{z}(t)) \leq h(\mathbf{z}_0)$  for all  $t \geq t_0$ . For the given partition (23) we consider two alternatives.

First, in the degenerative case, the state  $\mathbf{x}(t) \oplus \mathbf{z}(t)$  enters some  $\Omega_j$ ,  $j \geq 0$ , and stays there afterward, which automatically guarantees that  $0 \leq |h(\mathbf{z})| \leq h(\mathbf{z}_0)$ . Then, according to (6) the trajectory  $\mathbf{x}(t)$  satisfies the following inequality:

$$(65) \quad \|\mathbf{x}(t)\|_{\mathcal{A}} \leq \beta(\|\mathbf{x}_0\|_{\mathcal{A}}, t - t_0) + c\|h(\mathbf{z}(t))\|_{\infty, [t_0, t]} \leq \beta(\|\mathbf{x}_0\|_{\mathcal{A}}, t - t_0) + c|h(\mathbf{z}_0)|.$$

Taking into account that  $\beta(\cdot, \cdot) \in \mathcal{KL}$  we can conclude that (65) implies that

$$(66) \quad \limsup_{t \rightarrow \infty} \|\mathbf{x}(t)\|_{\mathcal{A}} \leq c|h(\mathbf{z}_0)|.$$

Therefore the statements of the theorem hold.

Let us consider the second alternative, where the state  $\mathbf{x}(t) \oplus \mathbf{z}(t)$  enters each  $\Omega_j$  and leaves later. Given that  $h(\mathbf{z}(t))$  is monotone and nonincreasing in  $t$ , this implies that there exists an ordered sequence of time instants  $t_j$ :

$$(67) \quad t_0 < t_1 < t_2 \cdots t_j < t_{j+1} \cdots$$

such that

$$(68) \quad h(\mathbf{z}(t_i)) = \sigma_i h(\mathbf{z}_0).$$

Hence in order to prove the theorem we must show that the sequence  $\{t_i\}_{i=0}^{\infty}$  does not converge. In other words, the boundary  $\sigma_{\infty} h(\mathbf{z}_0) = 0$  will not be reached in finite time.

In order to do this let us estimate the upper bounds for the following differences:

$$T_i = t_{i+1} - t_i.$$

Taking into account inequality (8) and the fact that  $\gamma_0(\cdot) \in \mathcal{K}_e$ , we can derive that

$$(69) \quad h(\mathbf{z}(t_i)) - h(\mathbf{z}(t_{i+1})) \leq T_i \max_{\tau \in [t_i, t_{i+1}]} \gamma_0(\|\mathbf{x}(\tau)\|_{\mathcal{A}}) \leq T_i \gamma_0(\|\mathbf{x}(\tau)\|_{\mathcal{A}_{\infty}, [t_i, t_{i+1}]}).$$

According to the definition of  $t_i$  in (68) and noticing that the sequence  $\mathcal{S}$  is strictly decreasing, we have

$$h(\mathbf{z}(t_i)) - h(\mathbf{z}(t_{i+1})) = (\sigma_i - \sigma_{i+1})h(\mathbf{z}_0) > 0.$$

Hence  $h(\mathbf{z}_0) > 0$  implies that  $\gamma_0(\|\mathbf{x}(\tau)\|_{\mathcal{A}_{\infty}, [t_i, t_{i+1}]}) > 0$  and, therefore, (69) results in the following estimate of  $T_i$ :

$$(70) \quad T_i \geq \frac{h(\mathbf{z}(t_i)) - h(\mathbf{z}(t_{i+1}))}{\gamma_0(\|\mathbf{x}(\tau)\|_{\mathcal{A}_{\infty}, [t_i, t_{i+1}]})} = \frac{h(\mathbf{z}_0)(\sigma_i - \sigma_{i+1})}{\gamma_0(\|\mathbf{x}(\tau)\|_{\mathcal{A}_{\infty}, [t_i, t_{i+1}]})}.$$

Taking into account that  $h(\mathbf{z}(t))$  is nonincreasing over  $[t_i, t_{i+1}]$  and using (6) we can bound the norm  $\|\mathbf{x}(\tau)\|_{\mathcal{A}_{\infty}, [t_i, t_{i+1}]}$  as follows:

$$(71) \quad \begin{aligned} \|\mathbf{x}(\tau)\|_{\mathcal{A}_{\infty}, [t_i, t_{i+1}]} &\leq \beta(\|\mathbf{x}(t_i)\|_{\mathcal{A}}, 0) + c\|h(\mathbf{z}(\tau))\|_{\infty, [t_i, t_{i+1}]} \\ &\leq \beta(\|\mathbf{x}(t_i)\|_{\mathcal{A}}, 0) + c \cdot \sigma_i h(\mathbf{z}_0). \end{aligned}$$

Hence, combining (70) and (71), we obtain that

$$T_i \geq \frac{h(\mathbf{z}_0)(\sigma_i - \sigma_{i+1})}{\gamma_0(\sigma_i(\sigma_i^{-1}\beta(\|\mathbf{x}(t_i)\|_{\mathcal{A}}, 0) + c \cdot h(\mathbf{z}_0)))}.$$

Then, using property (10) of function  $\gamma_0$  we can derive that

$$(72) \quad T_i \geq \frac{h(\mathbf{z}_0)(\sigma_i - \sigma_{i+1})}{\gamma_{0,1}(\sigma_i)} \frac{1}{\gamma_{0,2}(\sigma_i^{-1}\beta(\|\mathbf{x}(t_i)\|_{\mathcal{A}}, 0) + c \cdot h(\mathbf{z}_0)))}.$$

Taking into account condition (27) of the theorem, the theorem will be proved if we ensure that

$$(73) \quad T_i \geq \tau_i$$

for all  $i = 0, 1, 2, \dots, \infty$ . We prove this claim by induction with respect to the index  $i = 0, 1, \dots, \infty$ . We start with  $i = 0$ , and then show that for all  $i > 0$  the following implication holds:

$$(74) \quad T_i \geq \tau_i \Rightarrow T_{i+1} \geq \tau_{i+1}.$$

Let us prove that (73) holds for  $i = 0$ . For this purpose consider the term  $(\sigma_i - \sigma_{i+1})/\gamma_{0,1}(\sigma_i)$ . As follows immediately from condition (25) of the theorem, we have that

$$(75) \quad \frac{\sigma_i - \sigma_{i+1}}{\gamma_{0,1}(\sigma_i)} \geq \tau_i \Delta_0 \quad \forall i \geq 0.$$

In particular

$$\frac{\sigma_0 - \sigma_1}{\gamma_{0,1}(\sigma_0)} \geq \tau_0 \Delta_0.$$

Therefore, inequality (72) reduces to

$$(76) \quad T_0 \geq \tau_0 \Delta_0 \frac{h(\mathbf{z}_0)}{\gamma_{0,2}(\sigma_0^{-1} \beta(\|\mathbf{x}(t_0)\|_{\mathcal{A}}, 0) + c \cdot h(\mathbf{z}_0))}.$$

Moreover, taking into account Property 3 and (18), (19), we can derive the following estimate:

$$\begin{aligned} \sigma_0^{-1} \beta(\|\mathbf{x}(t_0)\|_{\mathcal{A}}, 0) &\leq \sigma_0^{-1} \phi_0(\|\mathbf{x}(t_0)\|_{\mathcal{A}}) + \sigma_0^{-1} v_0(c \cdot |h(\mathbf{z}_0)| \sigma_0) \\ &\leq B_1(\|\mathbf{x}_0\|_{\mathcal{A}}) + B_2(|h(\mathbf{z}_0)|, c). \end{aligned}$$

According to the theorem conditions,  $\mathbf{x}_0$  and  $\mathbf{z}_0$  satisfy inequality (26). This in turn implies that

$$(77) \quad \begin{aligned} &\gamma_{0,2}(\sigma_0^{-1} \beta(\|\mathbf{x}(t_0)\|_{\mathcal{A}}, 0) + c \cdot h(\mathbf{z}_0)) \\ &\leq \gamma_{0,2}(B_1(\|\mathbf{x}_0\|_{\mathcal{A}}) + B_2(|h(\mathbf{z}_0)|, c) + c \cdot h(\mathbf{z}_0)) \leq \Delta_0 \cdot h(\mathbf{z}_0). \end{aligned}$$

Combining (76) and (77), we obtain the desired inequality

$$T_0 \geq \tau_0 \Delta_0 \frac{h(\mathbf{z}_0)}{\gamma_{0,2}(\sigma_0^{-1} \beta(\|\mathbf{x}(t_0)\|_{\mathcal{A}}, 0) + c \cdot h(\mathbf{z}_0))} \geq \tau_0 \frac{\Delta_0 h(\mathbf{z}_0)}{\Delta_0 h(\mathbf{z}_0)} = \tau_0.$$

Thus the basis of induction is proved.

Let us assume that (73) holds for all  $i = 0, \dots, n$ ,  $n \geq 0$ . We shall prove now that implication (74) holds for  $i = n + 1$ . Consider the term  $\beta(\|\mathbf{x}(t_{n+1})\|_{\mathcal{A}}, 0)$ :

$$\begin{aligned} \beta(\|\mathbf{x}(t_{n+1})\|_{\mathcal{A}}, 0) &\leq \beta(\beta(\|\mathbf{x}(t_n)\|_{\mathcal{A}}, T_n) + c \|h(\mathbf{z}(\tau))\|_{\infty, [t_n, t_{n+1}]}, 0) \\ &\leq \beta(\beta(\|\mathbf{x}(t_n)\|_{\mathcal{A}}, T_n) + c \cdot \sigma_n \cdot h(\mathbf{z}_0), 0). \end{aligned}$$

Taking into account Property 2 (specifically, inequality (17)) and (18)–(20) we can derive that

$$(78) \quad \begin{aligned} \beta(\|\mathbf{x}(t_{n+1})\|_{\mathcal{A}}, 0) &\leq \beta(\xi_n \cdot \beta(\|\mathbf{x}(t_n)\|_{\mathcal{A}}, 0) + c \cdot \sigma_n \cdot h(\mathbf{z}_0), 0) \\ &\leq \phi_1(\|\mathbf{x}(t_n)\|_{\mathcal{A}}) + v_1(c \cdot |h(\mathbf{z}_0)| \cdot \sigma_n). \end{aligned}$$

Notice that, according to the inductive hypothesis ( $T_i \geq \tau_i$ ), the following holds:

$$(79) \quad \|\mathbf{x}(t_{i+1})\|_{\mathcal{A}} \leq \beta(\|\mathbf{x}(t_i)\|_{\mathcal{A}}, T_i) + c \cdot \sigma_i \cdot h(\mathbf{z}_0) \leq \xi_i \beta(\|\mathbf{x}(t_i)\|_{\mathcal{A}}, 0) + c \cdot \sigma_i \cdot h(\mathbf{z}_0)$$

for all  $i = 0, \dots, n$ . Then (78), (79), and (18)–(20) imply that

$$(80) \quad \begin{aligned} &\beta(\|\mathbf{x}(t_{n+1})\|_{\mathcal{A}}, 0) \leq \phi_1(\xi_n \beta(\|\mathbf{x}(t_{n-1})\|_{\mathcal{A}}, 0) + c \cdot \sigma_{n-1} \cdot h(\mathbf{z}_0)) \\ &\quad + v_1(c \cdot |h(\mathbf{z}_0)| \cdot \sigma_n) \leq \phi_2(\|\mathbf{x}(t_{n-1})\|_{\mathcal{A}}) + v_2(c \cdot |h(\mathbf{z}_0)| \cdot \sigma_{n-1}) \\ &\quad + v_1(c \cdot |h(\mathbf{z}_0)| \cdot \sigma_n) \leq \phi_{n+1}(\|\mathbf{x}_0\|_{\mathcal{A}}) + \sum_{i=1}^{n+1} v_i(c \cdot |h(\mathbf{z}_0)| \sigma_{n+1-i}) \\ &\leq \phi_{n+1}(\|\mathbf{x}_0\|_{\mathcal{A}}) + \sum_{i=0}^{n+1} v_i(c \cdot |h(\mathbf{z}_0)| \sigma_{n+1-i}). \end{aligned}$$

According to Property 3, the term

$$\sigma_{n+1}^{-1} \left( \phi_{n+1}(\|\mathbf{x}_0\|_{\mathcal{A}}) + \sum_{i=0}^{n+1} v_i(c \cdot |h(\mathbf{z}_0)|\sigma_{n+1-i}) \right)$$

is bounded from above by the sum

$$B_1(\|\mathbf{x}_0\|_{\mathcal{A}}) + B_2(|h(\mathbf{z}_0)|, c).$$

Therefore, monotonicity of  $\gamma_{0,2}$ , estimate (80), and inequality (26) lead to the following inequality:

$$\begin{aligned} \gamma_{0,2}(\sigma_{n+1}^{-1}\beta(\|\mathbf{x}(t_{n+1})\|_{\mathcal{A}}, 0) + c \cdot h(\mathbf{z}_0)) &\leq \gamma_{0,2}(B_1(\|\mathbf{x}_0\|_{\mathcal{A}}) + B_2(|h(\mathbf{z}_0)|, c) + c \cdot h(\mathbf{z}_0)) \\ &\leq h(\mathbf{z}_0)\Delta_0. \end{aligned}$$

Hence, according to (72) and (75) we have

$$\begin{aligned} T_{n+1} &\geq \frac{(\sigma_{n+1} - \sigma_{n+2})}{\gamma_{0,1}(\sigma_{n+1})} \frac{h(\mathbf{z}_0)}{\gamma_{0,2}(\sigma_{n+1}^{-1}\beta(\|\mathbf{x}(t_{n+1})\|_{\mathcal{A}}, 0) + c \cdot h(\mathbf{z}_0))} \\ &\geq \tau_{n+1} \frac{\Delta_0 h(\mathbf{z}_0)}{\Delta_0 h(\mathbf{z}_0)} = \tau_{n+1}. \end{aligned}$$

Thus implication (74) is proved. This implies that  $h(\mathbf{z}(t)) \in [0, h(\mathbf{z}_0)]$  for all  $t \geq t_0$  and, consequently, that (66) holds.  $\square$

**A.2. Proof of Lemma 6.** As follows from the assumptions,  $h(\mathbf{z}(t, \mathbf{z}_0))$  is bounded. Assume it belongs to the interval  $[a, h(\mathbf{z}_0)]$ ,  $a \leq h(\mathbf{z}_0)$ . Taking into account that  $h(\mathbf{z}(t, \mathbf{z}_0))$  is bounded and monotone in  $t$  (every subsequence of which is again monotone) and applying the Bolzano–Weierstrass theorem we can conclude that  $h(\mathbf{z}(t, \mathbf{z}_0))$  converges in  $[a, h(\mathbf{z}_0)]$ . In particular, there exists  $\bar{h} \in [a, h(\mathbf{z}_0)]$  such that

$$(81) \quad \lim_{t \rightarrow \infty} h(\mathbf{z}(t, \mathbf{z}_0)) = \bar{h}.$$

Therefore, as follows from (8) we can conclude that

$$\begin{aligned} (82) \quad 0 &\leq \lim_{t \rightarrow \infty} \int_{t_0}^t \gamma_1(\|\mathbf{x}(\tau, \mathbf{x}_0)\|_{\mathcal{A}}) d\tau \leq \lim_{t \rightarrow \infty} (h(\mathbf{z}_0) - h(\mathbf{z}(t, \mathbf{z}_0))) \\ &= h(\mathbf{z}_0) - \bar{h} \leq h(\mathbf{z}_0) - a < \infty. \end{aligned}$$

According to the lemma assumptions, system  $\mathcal{S}_a$  has steady-state characteristics. This means that there exists a constant  $\bar{x} \in \mathbb{R}_+$  such that

$$(83) \quad \lim_{t \rightarrow \infty} \|\mathbf{x}(t, \mathbf{x}_0)\|_{\mathcal{A}} = \bar{x}.$$

Suppose that  $\bar{x} > 0$ . Then it follows from (83) that there exists time instant  $t_1$ ,  $t_0 \leq t_1 < \infty$  and some constant  $0 < \delta < \bar{x}$  such that

$$\|\mathbf{x}(t)\|_{\mathcal{A}} \geq \delta \quad \forall t \geq t_1.$$

Hence using (82) and noticing that  $\gamma_1 \in \mathcal{K}_e$  we obtain

$$\infty > h(\mathbf{z}_0) - \bar{h} \geq \lim_{t \rightarrow \infty} \int_{t_0}^t \gamma_1(\|\mathbf{x}(\tau, \mathbf{x}_0)\|_{\mathcal{A}}) d\tau \geq \lim_{t \rightarrow \infty} \int_{t_1}^t \gamma_1(\delta) d\tau = \infty.$$



Thus we obtained a contradiction. Hence  $\bar{x} = 0$  and, consequently,

$$\lim_{t \rightarrow \infty} \|\mathbf{x}(t)\|_{\mathcal{A}} = 0.$$

Then, according to the notion of steady-state characteristic in Definition 4, this is possible only if  $\bar{h} \in \chi^{-1}(0)$ .  $\square$

**A.3. Proof of Lemma 7.** Analogously to the proof of Lemma 6 we notice that (82) holds. This, however, implies that for any constant and positive  $T$ , the limit

$$\lim_{t \rightarrow \infty} \int_t^{t+T} \gamma_1(\|\mathbf{x}(\tau)\|_{\mathcal{A}}) d\tau$$

exists and equals zero. Furthermore,  $h(\mathbf{z}(t, \mathbf{z}_0)) \in [0, h(\mathbf{z}_0)]$  for all  $t \geq t_0$ . Hence there exists a time instant  $t'$  such that

$$\|\mathbf{x}(t)\|_{\mathcal{A}} \leq c \cdot h(\mathbf{z}_0) + \varepsilon \quad \forall t \geq t',$$

where  $\varepsilon > 0$  is arbitrarily small. Then taking into account (31) we can conclude that

$$(84) \quad \lim_{t \rightarrow \infty} \int_t^{t+T} \gamma_1(\|\mathbf{x}(\tau)\|_{\mathcal{A}}) d\tau \geq \bar{\gamma} \int_t^{t+T} \|\mathbf{x}(\tau)\|_{\mathcal{A}} d\tau = 0.$$

Given that (81) holds, system (6) has the steady-state characteristic on average, and given that  $\chi_T(\cdot)$  has no zeros in the positive domain, limiting relation (84) is possible only if  $\bar{h} = 0$ . Then, according to (6),  $\lim_{t \rightarrow \infty} \|\mathbf{x}(t)\|_{\mathcal{A}} = 0$ .  $\square$

**A.4. Proof of Corollary 8.** As follows from Theorem 3, state  $\mathbf{x}(t, \mathbf{x}_0) \oplus \mathbf{z}(t, \mathbf{z}_0)$  converges to the set  $\Omega_a$  specified by (24). Hence  $h(\mathbf{z}(t, \mathbf{z}_0))$  is bounded. Then, according to (8), estimate (82) holds. This, in combination with condition (34), implies that  $\mathbf{z}(t, \mathbf{z}_0)$  is bounded. In other words,

$$\mathbf{x}(t, \mathbf{x}_0) \oplus \mathbf{z}(t, \mathbf{z}_0) \in \Omega' \quad \forall t \geq t_0,$$

where  $\Omega'$  is a bounded subset in  $\mathbb{R}^n \times \mathbb{R}^m$ . Applying the Bolzano–Weierstrass theorem we can conclude that for every point  $\mathbf{x}_0 \oplus \mathbf{z}_0 \in \Omega_\gamma$  there is an  $\omega$ -limit set  $\omega(\mathbf{x}_0 \oplus \mathbf{z}_0) \subseteq \Omega'$  (nonempty).

As follows from (C3) and Lemma 6, the following holds:

$$\lim_{t \rightarrow \infty} h(\mathbf{z}(t, \mathbf{z}_0)) \in \chi^{-1}(0).$$

Therefore, given that  $h(\cdot) \in \mathcal{C}^0$ , we can obtain that

$$\lim_{t_i \rightarrow \infty} (\mathbf{z}(t_i, \mathbf{z}_0)) = h \left( \lim_{t_i \rightarrow \infty} \mathbf{z}(t_i, \mathbf{z}_0) \right) = h(\omega_z(\mathbf{x}_0 \oplus \mathbf{z}_0)) \in \chi^{-1}(0).$$

In other words,

$$\omega_z(\mathbf{x}_0 \oplus \mathbf{z}_0) \subseteq \Omega_h = \{\mathbf{x} \in \mathbb{R}^n, \mathbf{z} \in \mathbb{R}^m \mid h(\mathbf{z}) \in \chi^{-1}(0)\}.$$

Moreover,

$$\omega_x(\mathbf{x}_0 \oplus \mathbf{z}_0) \subseteq \Omega_a = \{\mathbf{x} \in \mathbb{R}^n, \mathbf{z} \in \mathbb{R}^m \mid \|\mathbf{x}\|_{\mathcal{A}} = 0\}.$$

According to assumption (C1), the flow  $\mathbf{x}(t, \mathbf{x}_0) \oplus \mathbf{z}(t, \mathbf{z}_0)$  is generated by a system of autonomous differential equations with a locally Lipschitz right-hand side. Then, as follows from [13, Lemma 4.1, page 127],

$$\lim_{t \rightarrow \infty} \text{dist}(\mathbf{x}(t, \mathbf{x}_0) \oplus \mathbf{z}(t, \mathbf{z}_0), \omega(\mathbf{x}_0 \oplus \mathbf{z}_0)) = 0.$$

Noticing that

$$\text{dist}(\mathbf{x}(t, \mathbf{x}_0) \oplus \mathbf{z}(t, \mathbf{z}_0), \omega(\mathbf{x}_0 \oplus \mathbf{z}_0)) \geq \text{dist}(\mathbf{x}(t, \mathbf{x}_0), \Omega_a) + \text{dist}(\mathbf{z}(t, \mathbf{z}_0), \Omega_h),$$

we can finally obtain that

$$\lim_{t \rightarrow \infty} \text{dist}(\mathbf{x}(t, \mathbf{x}_0), \Omega_a) = 0, \quad \lim_{t \rightarrow \infty} \text{dist}(\mathbf{z}(t, \mathbf{z}_0), \Omega_h) = 0. \quad \square$$

**A.5. Proof of Corollary 9.** As follows from Theorem 3, the corollary will be proved if Properties 1–3 are satisfied and also if (25), (26), and (27) hold. In order to satisfy Property 1 we select the following sequence  $\mathcal{S}$ :

$$(85) \quad \mathcal{S} = \{\sigma_i\}_{i=0}^{\infty}, \quad \sigma_i = \frac{1}{\kappa^i}, \quad \kappa \in \mathbb{R}_+, \quad \kappa > 1.$$

Let us choose sequences  $\mathcal{T}$  and  $\Xi$  as follows:

$$(86) \quad \mathcal{T} = \{\tau_i\}_{i=0}^{\infty}, \quad \tau_i = \tau^*,$$

$$(87) \quad \Xi = \{\xi_i\}_{i=0}^{\infty}, \quad \xi_i = \xi^*,$$

where  $\tau^*, \xi^*$  are positive constants yet to be defined. Notice that choosing  $\mathcal{T}$  as in (86) automatically fulfills condition (27) of Theorem 3. On the other hand, taking into account (17), (39), and that  $\beta_t(t)$  is monotonically decreasing in  $t$ , this choice defines a constant  $\xi^*$  as follows:

$$(88) \quad \beta_t(\tau^*) \leq \xi^* \beta_t(0) < \beta_t(0), \quad 0 \leq \xi^* < 1.$$

Given that the inverse  $\beta_t^{-1}$  exists, (40), this choice is always possible. In particular, (88) will be satisfied for the following values of  $\tau^*$ :

$$(89) \quad \tau^* \geq \beta_t^{-1}(\xi^* \beta_t(0)).$$

Let us now find the values for  $\tau^*$  and  $\xi^*$  such that Property 3 is also satisfied. For this purpose consider systems of functions  $\Phi, \Upsilon$  specified by (18), (19). Notice that function  $\beta(s, 0)$  in (18), (19) is linear for system (39),

$$\beta(s, 0) = s \cdot \beta_t(0),$$

and therefore the functions  $\rho_{\phi, j}(\cdot), \rho_{v, j}$  are identity maps. Hence  $\Phi, \Upsilon$  reduce to the following:

$$(90) \quad \Phi : \begin{aligned} \phi_j(s) &= \phi_{j-1} \cdot \xi^* \cdot \beta(s, 0) = \xi^* \cdot \beta_t(0) \cdot \phi_{j-1}(s), \quad j = 1, \dots, i, \\ \phi_0(s) &= \beta_t(0) \cdot s, \end{aligned}$$

$$(91) \quad \Upsilon : \begin{aligned} v_j(s) &= \phi_{j-1}(s), \quad j = 1, \dots, i, \\ v_0(s) &= \beta_t(0) \cdot s. \end{aligned}$$

Taking into account (85), (90), and (91), let us explicitly formulate requirements (21), (22) in Property 3. These conditions are equivalent to the boundedness of the following functions:

$$(92) \quad \|\mathbf{x}(t_0)\|_{\mathcal{A}} \cdot \beta_t(0) \cdot \kappa^n (\xi^* \cdot \beta_t(0))^n;$$

$$(93) \quad \begin{aligned} & \kappa^n \left( \beta_t(0) \frac{c|h(\mathbf{z}_0)|}{\kappa^n} + \frac{\beta_t(0)c|h(\mathbf{z}_0)|}{\kappa^{n-1}} + \beta_t(0) \sum_{i=2}^n c|h(\mathbf{z}_0)| \frac{1}{\kappa^{n-i}} (\xi^* \cdot \beta_t(0))^{i-1} \right) \\ &= \beta_t(0)c|h(\mathbf{z}_0)| + \beta_t(0)c|h(\mathbf{z}_0)|\kappa \left( 1 + \sum_{i=2}^n \kappa^{i-1} (\xi^* \cdot \beta_t(0))^{i-1} \right). \end{aligned}$$

Boundedness of the functions  $B_1(\|\mathbf{x}_0\|_{\mathcal{A}})$  and  $B_2(|h(\mathbf{z}_0)|, c)$  is ensured if  $\xi^*$  satisfies the inequality

$$(94) \quad \xi^* \leq \frac{d}{\kappa \cdot \beta_t(0)}$$

for some  $0 \leq d < 1$ . Notice that  $\kappa > 1$ ,  $\beta_t(0) \geq 1$  imply that  $\xi^* \leq 1$ , and therefore constant  $\tau^*$  satisfying (89) will always be defined. Hence according to (92) and (93), the functions  $B_1(\|\mathbf{x}_0\|_{\mathcal{A}})$  and  $B_2(|h(\mathbf{z}_0)|, c)$  satisfying Property 3 can be chosen as

$$(95) \quad B_1(\|\mathbf{x}_0\|_{\mathcal{A}}) = \beta_t(0) \|\mathbf{x}_0\|_{\mathcal{A}}; \quad B_2(|h(\mathbf{z}_0)|, c) = \beta_t(0) \cdot c \cdot |h(\mathbf{z}_0)| \left( 1 + \frac{\kappa}{1-d} \right).$$

In order to apply Theorem 3 we have to check the remaining conditions (25) and (26). This requires the possibility of factorization (10) for the function  $\gamma_0(\cdot)$ . According to assumption (41) of the corollary the function  $\gamma_0(\cdot)$  is Lipschitz:

$$|\gamma_0(s)| \leq D_{\gamma,0} \cdot |s|.$$

This allows us to choose functions  $\gamma_{0,1}(\cdot)$  and  $\gamma_{0,2}(\cdot)$  as follows:

$$(96) \quad \gamma_{0,1}(s) = s, \quad \gamma_{0,2}(s) = D_{\gamma,0} \cdot s.$$

Condition (25), therefore, is equivalent to solvability of the following inequality:

$$(97) \quad \left( \frac{1}{\kappa^i} - \frac{1}{\kappa^{i+1}} \right) \frac{\kappa^i}{\tau^*} \geq \Delta_0.$$

Taking into account inequalities (89) and (94), we can derive that solvability of

$$(98) \quad \Delta_0 = \left( \beta_t^{-1} \left( \frac{d}{\kappa} \right) \right)^{-1} \frac{\kappa - 1}{\kappa}$$

implies existence of  $\Delta_0 > 0$  satisfying (97) and, consequently, condition (25) of Theorem 3. Given that  $d < 1$ ,  $\kappa > 1$ , and  $\beta_t(0) \geq 1$ , a positive solution to (98) is always defined. Hence the proof will be complete and the claim is nonvacuous if the domain

$$(99) \quad \begin{aligned} & D_{\gamma,0} \leq \left( \beta_t^{-1} \left( \frac{d}{\kappa} \right) \right)^{-1} \frac{\kappa - 1}{\kappa} \\ & \times \frac{h(\mathbf{z}_0)}{\beta_t(0) \|\mathbf{x}_0\|_{\mathcal{A}} + \beta_t(0) \cdot c \cdot |h(\mathbf{z}_0)| \left( 1 + \frac{\kappa}{1-d} \right) + c|h(\mathbf{z}_0)|} \end{aligned}$$

is not empty.  $\square$

**A.6. Proof of Corollary 10.** It follows from Corollary 9 that the state of the interconnection converges into  $\Omega_a$  for all initial conditions  $\mathbf{x}_0, \mathbf{z}_0$  satisfying (99). In other words, the following inequality should hold:

$$(100) \quad \begin{aligned} & D_{\gamma,0} \left( \beta_t(0) \|\mathbf{x}_0\|_{\mathcal{A}} + \beta_t(0) \cdot c \cdot |h(\mathbf{z}_0)| \left( 1 + \frac{\kappa}{1-d} \right) + c|h(\mathbf{z}_0)| \right) \\ & \leq \left( \beta_t^{-1} \left( \frac{d}{\kappa} \right) \right)^{-1} \frac{\kappa-1}{\kappa} \cdot h(\mathbf{z}_0). \end{aligned}$$

Hence assuming that  $h(\mathbf{z}_0) > 0$ , we can rewrite (100) in the following way:

$$(101) \quad \begin{aligned} & D_{\gamma,0} \cdot \beta_t(0) \|\mathbf{x}_0\|_{\mathcal{A}} \\ & \leq \left( \left( \beta_t^{-1} \left( \frac{d}{\kappa} \right) \right)^{-1} \frac{\kappa-1}{\kappa} - D_{\gamma,0} \cdot c \left( \beta_t(0) \cdot \left( 1 + \frac{\kappa}{1-d} \right) + 1 \right) \right) h(\mathbf{z}_0). \end{aligned}$$

Solutions to (101) exist, however, if the inequality

$$\left( \beta_t^{-1} \left( \frac{d}{\kappa} \right) \right)^{-1} \frac{\kappa-1}{\kappa} \geq D_{\gamma,0} \cdot c \left( \beta_t(0) \cdot \left( 1 + \frac{\kappa}{1-d} \right) + 1 \right)$$

or, equivalently,

$$(102) \quad D_{\gamma,0} \cdot c \cdot \left( \beta_t(0) \cdot \left( 1 + \frac{\kappa}{1-d} \right) + 1 \right) \cdot \beta_t^{-1} \left( \frac{d}{\kappa} \right) \frac{\kappa}{\kappa-1} < 1$$

is satisfied. The estimate of the trapping region follows from (101).

Let us finally show that continuity of  $h(\mathbf{z})$  implies that the volume of  $\Omega_\gamma$  is nonzero in  $\mathbb{R}^n \oplus \mathbb{R}^m$ . For the sake of compactness we rewrite inequality (101) in the following form:

$$(103) \quad \|\mathbf{x}_0\|_{\mathcal{A}} \leq C_\gamma h(\mathbf{z}_0),$$

where  $C_\gamma$  is a constant depending on  $d, \kappa, \beta_t(0)$ , and  $D_{\gamma,0}$ . Given that (102) holds we can conclude that  $C_\gamma > 0$ . According to (103), domain  $\Omega_\gamma$  contains the following set:

$$\{\mathbf{x}_0 \in \mathbb{R}^n, \mathbf{z}_0 \in \mathbb{R}^m \mid h(\mathbf{z}_0) > D_z \in \mathbb{R}_+, \|\mathbf{x}_0\|_{\mathcal{A}} \leq C_\gamma D_z\}.$$

Consider the following domain:  $\Omega_{\mathbf{x},\gamma} = \{\mathbf{x}_0 \in \mathbb{R}^n \mid \|\mathbf{x}_0\|_{\mathcal{A}} \leq C_\gamma D_z\}$ . Clearly, it contains a point  $\mathbf{x}_{0,1} \in \mathbb{R}^n : \|\mathbf{x}_{0,1}\|_{\mathcal{A}} = \frac{C_\gamma D_z}{2}$ . For the point  $\mathbf{x}_{0,1}$  and for all  $\boldsymbol{\varepsilon}_1 \in \mathbb{R}^n : \|\boldsymbol{\varepsilon}_1\| \leq \frac{C_\gamma D_z}{4}$  we have that  $\|\mathbf{x}_{0,1} + \boldsymbol{\varepsilon}_1\|_{\mathcal{A}} = \inf_{\mathbf{q} \in \mathcal{A}} \|\mathbf{x}_{0,1} + \boldsymbol{\varepsilon}_1 - \mathbf{q}\| \leq \inf_{\mathbf{q} \in \mathcal{A}} \{\|\mathbf{x}_{0,1} - \mathbf{q}\| + \|\boldsymbol{\varepsilon}_1\|\} \leq \frac{3C_\gamma D_z}{4}$ . On the other hand,  $\|\mathbf{x}_{0,1} + \boldsymbol{\varepsilon}_1\|_{\mathcal{A}} = \inf_{\mathbf{q} \in \mathcal{A}} \|\mathbf{x}_{0,1} + \boldsymbol{\varepsilon}_1 - \mathbf{q}\| \geq \inf_{\mathbf{q} \in \mathcal{A}} \{\|\mathbf{x}_{0,1} - \mathbf{q}\| - \|\boldsymbol{\varepsilon}_1\|\} \geq \frac{C_\gamma D_z}{4}$ . This implies that there exists a set of points  $\mathbf{x}_{0,2} = \mathbf{x}_{0,1} + \boldsymbol{\varepsilon}_1 \in \mathbb{R}^n : \|\mathbf{x}_{0,1} - \mathbf{x}_{0,2}\| \leq \frac{C_\gamma D_z}{4}, \mathbf{x}_{0,2} \notin \mathcal{A}, \|\mathbf{x}_{0,2}\|_{\mathcal{A}} \leq C_\gamma D_z$ .

Consider now the following domain:  $\Omega_{\mathbf{z},\gamma} = \{\mathbf{z}_0 \in \mathbb{R}^m \mid h(\mathbf{z}_0) > D_z\}$ . Let us pick  $\mathbf{z}_{0,1} \in \Omega_{\mathbf{z},\gamma} : h(\mathbf{z}_{0,1}) = 2D_z$ . Because  $h(\cdot)$  is continuous we have that

$$\forall \varepsilon > 0, \exists \delta > 0 : \|\mathbf{z}_{0,1} - \mathbf{z}_{0,2}\| < \delta \Rightarrow |h(\mathbf{z}_{0,1}) - h(\mathbf{z}_{0,2})| < \varepsilon.$$

Let  $\varepsilon = D_z$ ; then  $-D_z < h(\mathbf{z}_{0,1}) - h(\mathbf{z}_{0,2}) < D_z$  and therefore  $h(\mathbf{z}_{0,2}) > D_z$ . Hence there exists a set of points  $\mathbf{z}_{0,2} \in \mathbb{R}^m : \|\mathbf{z}_{0,1} - \mathbf{z}_{0,2}\| < \delta, \mathbf{z}_{0,2} \in \Omega_{\mathbf{z},\gamma}$ .

Consider the following set:

$$\Omega_{\mathbf{xz},\gamma} = \left\{ \mathbf{x}' \in \mathbb{R}^n, \mathbf{z}' \in \mathbb{R}^m \mid \|\mathbf{x}_{0,1} - \mathbf{x}'\|^2 + \|\mathbf{z}_{0,1} - \mathbf{z}'\|^2 \leq r^2, r = \min \left\{ \delta, \frac{C_\gamma D_z}{4} \right\} \right\}.$$

For all  $\mathbf{x}_0, \mathbf{z}_0 \in \Omega_{\mathbf{xz},\gamma}$  we have that  $\mathbf{x}_0 \in \Omega_{\mathbf{x},\gamma}$ ,  $\mathbf{z}_0 \in \Omega_{\mathbf{z},\gamma}$ . Hence inequality (103) holds, and  $\mathbf{x}_0 \oplus \mathbf{z}_0 \in \Omega_\gamma$ . The volume of the set  $\Omega_{\mathbf{xz},\gamma}$  is defined by the volume of the interior of a sphere in  $\mathbb{R}^{n+m}$  with nonzero radius. Thus the volume of  $\Omega_\gamma \supset \Omega_{\mathbf{xz},\gamma}$  is also nonzero.  $\square$

**A.7. Proof of Corollary 11.** Let  $\lambda(\tau, \lambda_0)$  be a solution of system (51). Consider it as a function of variable  $\tau$ . Let us pick some monotone, strictly increasing function  $\sigma$  such that the following holds:

$$\tau = \sigma(t), \sigma : \mathbb{R}_+ \rightarrow \mathbb{R}_+.$$

Given that  $\eta(\Omega_\lambda)$  is dense in  $\Omega_\theta$ , for any  $\theta \in \Omega_\theta$  there always exists a vector  $\lambda_\theta \in \Omega_\lambda$  such that  $\eta(\lambda_\theta) = \theta + \epsilon_\theta$ , where  $\|\epsilon_\theta\|$  is arbitrarily small. Furthermore,  $\lambda(\tau)$  is dense in  $\Omega_\lambda$ , and hence there is a point  $\lambda^* = \lambda(\tau^*, \lambda_0)$  which is arbitrarily close to  $\lambda_\theta$ . Consider the following difference:

$$\mathbf{f}(\xi(t), \theta) - \mathbf{f}(\xi(t), \hat{\theta}) = \mathbf{f}(\xi(t), \theta) - \mathbf{f}(\xi(t), \eta(\lambda^*)) + \mathbf{f}(\xi, \eta(\lambda^*)) - \mathbf{f}(\xi, \eta(\lambda(\sigma(t)))).$$

The function  $\mathbf{f}(\cdot)$  is locally bounded and  $\eta(\cdot)$  is Lipschitz, and then

$$\|\mathbf{f}(\xi, \theta) - \mathbf{f}(\xi, \eta(\lambda^*))\| \leq D_f \|\epsilon_\theta\| + \Delta_f = \Delta_\theta + \Delta_f,$$

where  $\Delta_\theta$  is arbitrarily small. Hence

$$(104) \quad \begin{aligned} \|\mathbf{f}(\xi, \eta(\lambda^*)) - \mathbf{f}(\xi, \eta(\lambda(\sigma(t))))\| &\leq D_f \|\eta(\lambda^*) - \eta(\lambda(\sigma(t)))\| + \Delta_f + \Delta_\theta \\ &\leq D_f \cdot D_\eta \|\lambda^* - \lambda(\sigma(t))\| + \Delta_f + \Delta_\theta. \end{aligned}$$

Noticing that  $\lambda^* = \lambda(\tau^*, \lambda_0) = \lambda(\sigma(\tau^*), \lambda_0)$  and taking into account the Poisson stability of (51), we can always choose  $\lambda^*(\sigma^*, \lambda_0)$  such that  $\sigma^* > \sigma(t_0) = \tau_0$  for any  $\tau_0 \in \mathbb{R}_+$ . Hence, according to (104) the following estimate holds:

$$(105) \quad \begin{aligned} \|\mathbf{f}(\xi, \eta(\lambda^*)) - \mathbf{f}(\xi, \eta(\lambda(\sigma(t))))\| &\leq D_f \cdot D_\eta \left\| \int_{\sigma(t)}^{\sigma^*} S(\lambda(\sigma(\tau))) d\tau \right\| + \Delta_f + \Delta_\theta \\ &\leq D_f \cdot D_\eta \cdot \max_{\lambda \in \Omega_\lambda} \|S(\lambda)\| |\sigma^* - \sigma(t)| = \mathcal{D} \cdot |\sigma^* - \sigma(t)| + \Delta_f + \Delta_\theta, \\ \mathcal{D} &= D_f \cdot D_\eta \cdot \max_{\lambda \in \Omega_\lambda} \|S(\lambda)\|. \end{aligned}$$

Denoting  $\mathbf{u}(t) = \mathbf{f}(\xi(t), \theta) - \mathbf{f}(\xi(t), \hat{\theta}) + \varepsilon(t)$  we can now conclude that

$$(106) \quad \begin{aligned} \|\mathbf{u}(t)\| &\leq \Delta_\epsilon + \Delta_f + \|\mathbf{f}(\xi(t), \theta) - \mathbf{f}(\xi(t), \eta(\lambda^*))\| + \mathcal{D} \cdot |\sigma^* - \sigma(t)| \\ &\leq \Delta_\epsilon + 2\Delta_f + \Delta_\theta + D_f \|\theta - \eta(\lambda^*)\| + \mathcal{D} \cdot |\sigma^* - \sigma(t)|. \end{aligned}$$

Notice that due to the denseness of  $\lambda(t, \lambda_0)$  in  $\Omega_\lambda$  it is always possible to choose  $\lambda^*$  such that

$$D_f \|\theta - \eta(\lambda^*)\| = D_f \|\eta(\lambda_\theta) - \eta(\lambda^*)\| \leq D_f D_\eta \|\lambda_\theta - \eta(\lambda^*)\| \leq \Delta_\lambda.$$

Hence, according to (106), we have

$$\|\mathbf{u}(t)\|_{\infty, [t_0, t]} \leq 2\Delta_f + \Delta_\varepsilon + \delta + \mathcal{D} \cdot \|\sigma^* - \sigma(t)\|_{\infty, [t_0, t]},$$

where the term  $\delta > \Delta_\theta + \Delta_\lambda$  can be made arbitrarily small.

Therefore Assumption 3 implies that the following inequality holds:

$$(107) \quad \|\mathbf{x}(t)\|_{\mathcal{A}_{\Delta(M)}} \leq \beta(t - t_0) \|\mathbf{x}(t_0)\|_{\mathcal{A}_{\Delta(M)}} + c \cdot \mathcal{D} \cdot \|\sigma^* - \sigma(t)\|_{\infty, [t_0, t]}.$$

Let us now define  $\sigma(t)$  as follows:

$$(108) \quad \sigma(t) = \int_{t_0}^t \gamma \|\psi(\mathbf{x}(\tau))\|_{\mathcal{A}_{\Delta(M)}} d\tau.$$

Moreover, let us introduce the following notation:

$$h(t) = \sigma^* - \sigma(t) = \sigma^* - \int_{t_0}^t \gamma \|\psi(\mathbf{x}(\tau))\|_{\mathcal{A}_{\Delta(M)}} d\tau;$$

then for all  $t', t \geq t_0$ ,  $t \geq t'$  we have that

$$h(t') - h(t) = \int_{t'}^t \gamma \|\psi(\mathbf{x}(\tau))\|_{\mathcal{A}_{\Delta(M)}} d\tau.$$

Taking into account (104), (105), equality

$$\frac{\partial \boldsymbol{\lambda}(\sigma(t), \boldsymbol{\lambda}_0)}{dt} = \frac{\partial \sigma(t)}{dt} S(\boldsymbol{\lambda}(\sigma(t), \boldsymbol{\lambda}_0)) = \gamma \|\psi(\mathbf{x}(\tau))\|_{\mathcal{A}_{\Delta(M)}} S(\boldsymbol{\lambda}(\sigma(t), \boldsymbol{\lambda}_0)),$$

and (107), and denoting  $D_\lambda = c\mathcal{D}$ , we can conclude that the following holds along the trajectories of (52):

$$(109) \quad \begin{aligned} \|\mathbf{x}(t)\|_{\mathcal{A}_{\Delta(M)}} &\leq \beta(t - t_0) \|\mathbf{x}(t_0)\|_{\mathcal{A}_{\Delta(M)}} + D_\lambda \|h(t)\|_{\infty, [t_0, t]}, \\ h(t_0) - h(t) &= \int_{t_0}^t \gamma \|\psi(\mathbf{x}(\tau))\|_{\mathcal{A}_{\Delta(M)}} d\tau. \end{aligned}$$

Hence, according to Corollary 8, the limit relation (54) holds for all  $|h(t_0)|$ ,  $\|\mathbf{x}(t_0)\|_{\mathcal{A}_{\Delta(M)}}$  which belong to the domain

$$\begin{aligned} \Omega_\gamma : \gamma &\leq \left( \beta_t^{-1} \left( \frac{d}{\kappa} \right) \right)^{-1} \frac{\kappa - 1}{\kappa} \\ &\times \frac{h(t_0)}{\beta_t(0) \|\mathbf{x}(t_0)\|_{\mathcal{A}_{\Delta+\delta}} + \beta_t(0) \cdot D_\lambda \cdot |h(t_0)| \left( 1 + \frac{\kappa}{1-d} \right) + D_\lambda |h(t_0)|} \end{aligned}$$

for some  $d < 1$ ,  $\kappa > 1$ . Notice, however, that  $\|\mathbf{x}(t)\|_{\mathcal{A}_{\Delta+\delta}}$  is always bounded, as  $\mathbf{f}(\cdot)$  is Lipschitz in  $\theta$  and both  $\boldsymbol{\theta}$  and  $\hat{\boldsymbol{\theta}}$  are bounded ( $\boldsymbol{\eta}(\cdot)$  is Lipschitz and  $\boldsymbol{\lambda}(t, \boldsymbol{\lambda}_0)$  is bounded according to assumptions of the corollary). Moreover, due to the Poisson stability of (51) it is always possible to choose a point  $\boldsymbol{\lambda}^*$  such that  $h(t_0) = \sigma^*$  is arbitrarily large. Hence the choice of  $\gamma$  in (109), as (53) suffices to ensure that  $h(t)$  is bounded. Moreover, it follows that  $h(t)$  converges to a limit as  $t \rightarrow \infty$ . This implies

that  $\gamma \int_{t_0}^t \|\mathbf{x}(\tau)\|_{\mathcal{A}_{\Delta(M)}}$  also converges as  $t \rightarrow \infty$  and, consequently,  $\boldsymbol{\lambda}(t, \boldsymbol{\lambda}_0)$  converges to some  $\boldsymbol{\lambda}' \in \Omega_{\lambda}$ . Hence the following holds:

$$\lim_{t \rightarrow \infty} \hat{\boldsymbol{\theta}}(t) = \boldsymbol{\theta}'$$

for some  $\boldsymbol{\theta}' \in \Omega_{\theta}$ . According to the corollary conditions, system (50) has steady-state characteristics with respect to  $\hat{\boldsymbol{\theta}}$ . Then, in the same way as in the proof of Lemma 6, we can show that (54) holds.  $\square$

**Acknowledgments.** The authors are thankful to Peter Jurica and Tatiana Tyukina for their enthusiastic help and comments during the preparation of this manuscript.

#### REFERENCES

- [1] M. ARCAK, D. ANGELI, AND E. SONTAG, *A unifying integral ISS framework for stability of nonlinear cascades*, SIAM J. Control Optim., 40 (2002), pp. 1888–1904.
- [2] P. ASHWIN AND M. TIMME, *When instability makes sense*, Nature, 436 (2005), pp. 36–37.
- [3] G. BASTIN AND M. GEVERS, *Stable adaptive observers for nonlinear time-varying systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 650–658.
- [4] G. BESANCON, *Remarks on nonlinear adaptive observer design*, Systems Control Lett., 41 (2000), pp. 271–280.
- [5] G.-I. BISCHI, L. STEFANINI, AND L. GARDINI, *Synchronization, intermittency, and critical curves in a duopoly game*, Math. Comput. Simulation, 44 (1998), pp. 559–585.
- [6] C. CAO, A. M. ANNASWAMY, AND A. KOJIC, *Parameter convergence in nonlinearly parametrized systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 397–411.
- [7] J. CARR, *Applications of Centre Manifold Theory*, Springer-Verlag, New York, 1981.
- [8] L. GRUNE, E. SONTAG, AND F. R. WIRTH, *Asymptotic stability equals exponential stability, and ISS equals finite energy gain—if you twist your eyes*, Systems Control Lett., 38 (1999), pp. 127–134.
- [9] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer, New York, 2002.
- [10] J. L. HINDMARSH AND R. M. ROSE, *A model of the nerve impulse using two first-order differential equations*, Nature, 269 (1982), pp. 162–164.
- [11] A. ILCHMAN, *Universal adaptive stabilization of nonlinear systems*, Dynam. Control, (1997), pp. 199–213.
- [12] Z.-P. JIANG, A. R. TEEL, AND L. PRALY, *Small-gain theorem for ISS systems and applications*, Math. Control Signals Systems, (1994), pp. 95–120.
- [13] H. KHALIL, *Nonlinear Systems*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 2002.
- [14] J. P. LA SALLE, *Stability theory and invariance principles*, in Dynamical Systems, An International Symposium, J. K. Hale, L. Cesari, and J. P. La Salle, eds., Vol. 1, Academic Press, New York, 1976, pp. 211–222.
- [15] A. M. LYAPUNOV, *The general problem of the stability of motion*, Int. J. Control, Lyapunov Centenary Issue, 55 (1992), pp. 531–773.
- [16] R. MARINO, *Adaptive observers for single output nonlinear systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 1054–1058.
- [17] J. MILNOR, *On the concept of attractor*, Commun. Math. Phys., 99 (1985), pp. 177–195.
- [18] I. MIROSHNIK, V. NIKIFOROV, AND A. FRADKOV, *Nonlinear and Adaptive Control of Complex Systems*, Kluwer, Dordrecht, 1999.
- [19] E. OTT AND J. C. SOMMERER, *Blowout bifurcations: The occurrence of riddled basins and on-off intermittency*, Phys. Lett. A, 188 (1994), pp. 39–47.
- [20] A. Y. POGROMSKY, G. SANTOBONI, AND H. NIJMEIJER, *An ultimate bound on the trajectories of the Lorenz system and its applications*, Nonlinearity, 16 (2003), pp. 1597–1605.
- [21] J.-B. POMET, *Remarks on sufficient information for adaptive nonlinear regulation*, in Proceedings of the 31st Annual IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 1992, pp. 1737–1741.
- [22] Y. SHANG AND B. W. WAH, *Global optimization for neural network training*, Computer, 29 (1996), pp. 45–54.
- [23] E. SONTAG, *Further facts about input to state stabilization*, IEEE Trans. Automat. Control, 35 (1990), pp. 473–476.

- [24] E. SONTAG AND Y. WANG, *New characterizations of input-to-state stability*, IEEE Trans. Automat. Control, 41 (1996), pp. 1283–1294.
- [25] E. STEUR, I. YU. TYUKIN, H. NIJMEIJER, AND C. VAN LEEUWEN, *Reconstructing dynamics of spiking neurons from input-output measurements in vitro*, in Proceedings of the 3rd Annual IEEE Physics and Control Conference, Potsdam, Germany, 2007. Available online at <http://lib.physcon.ru/>
- [26] Y. SUEMITSU AND S. NARA, *A solution for two-dimensional mazes with use of chaotic dynamics in a recurrent neural network model*, Neural Computation, 16 (2004), pp. 1943–1957.
- [27] M. TIMME, F. WOLF, AND T. GEISEL, *Prevalence of unstable attractors in networks of pulse-coupled oscillators*, Phys. Rev. Lett., 89 (2002), 154105.
- [28] I. YU. TYUKIN AND C. VAN LEEUWEN, *Adaptation and nonlinear parametrization: Nonlinear dynamics prospective*, in Proceedings of the 16th Annual IFAC World Congress, Prague, Czech Republic, 2005. Available online at [http://arxiv.org/PS\\_cache/math/pdf/0412/0412444v1.pdf](http://arxiv.org/PS_cache/math/pdf/0412/0412444v1.pdf)
- [29] I. YU. TYUKIN, D. V. PROKHOROV, AND C. VAN LEEUWEN, *Adaptation and parameter estimation in systems with unstable target dynamics and nonlinear parametrization*, IEEE Trans. Automat. Control, 52 (2007), pp. 1543–1559.
- [30] C. VAN LEEUWEN AND A. RAFFONE, *Coupled nonlinear maps as models of perceptual pattern and memory trace dynamics*, Cognitive Processing, 2 (2001), pp. 67–111.
- [31] C. VAN LEEUWEN, S. VERVER, AND M. BRINKERS, *Visual illusions, solid/outline-invariance, and non-stationary activity patterns*, Connection Science, 12 (2000), pp. 279–297.
- [32] V. I. VOROTNIKOV, *Partial Stability and Control*, Birkhäuser Boston, Boston, MA, 1998.
- [33] T. YOSHIZAWA, *Stability and boundedness of systems*, Arch. Rational Mech. Anal., 6 (1960), pp. 409–421.
- [34] G. ZAMES, *On the input-output stability of time-varying nonlinear feedback systems. part I: Conditions derived using concepts of loop gain, conicity, and passivity*, IEEE Trans. Automat. Control, 11 (1966), pp. 228–238.



## A SET-VALUED EKELAND'S VARIATIONAL PRINCIPLE IN VECTOR OPTIMIZATION\*

C. GUTIÉRREZ<sup>†</sup>, B. JIMÉNEZ<sup>‡</sup>, AND V. NOVO<sup>‡</sup>

**Abstract.** This paper deals with Ekeland's variational principle for vector optimization problems. By using a set-valued metric, a set-valued perturbed map, and a cone-boundedness concept based on scalarization, we introduce an original approach to extending the well-known scalar Ekeland's principle to vector-valued maps. As a consequence of this approach, we obtain an Ekeland's variational principle that does not depend on any approximate efficiency notion. This result is related to other Ekeland's principles proved in the literature, and the finite-dimensional case is developed via an  $\varepsilon$ -efficiency notion that we introduced in [*Math. Methods Oper. Res.*, 64 (2006), pp. 165–185; *SIAM J. Optim.*, 17 (2006), pp. 688–710].

**Key words.** Ekeland's variational principle,  $\varepsilon$ -efficiency, set-valued metric, normal based cone, cone-boundedness

**AMS subject classifications.** 49J53, 90C48, 65K10

**DOI.** 10.1137/060672868

**1. Introduction.** Ekeland's variational principle (EVP) [6] has been used extensively in subjects such as fixed point theorems, geometry of Banach spaces, mathematical programming, control theory, etc. (See [3, p. 183] and [7] for other applications.) So, EVP is one of the most popular tools in nonlinear analysis and optimization theory.

Motivated by this wide usefulness, during the last two decades different authors have been interested in obtaining EVP for vector-valued maps (see, for instance, [2, 5, 9, 10, 15, 16, 17, 19, 21, 23, 25] and the recent survey by Chen, Huang, and Yang in [3, Chapter 4]).

Broadly speaking, one can find in the literature four different approaches to obtaining EVP for vector-valued maps: by using a scalarization process [5, 19, 23]; by transfinite induction arguments such as Zorn's lemma, the axiom of choice, the Hausdorff maximality principle, etc. [2, 21, 25]; by existence theorems of critical points in dynamic systems such as the Dancs–Hegedus–Medvegyev theorem [15]; or by means of particular classes of cones such as nuclear, well-based, or Daniell cones [9, 10, 16, 17].

A lot of EVPs proved in the literature for vector-valued maps have the same statements as the classical EVP for a scalar functional, since they give necessary conditions for  $\varepsilon$ -efficient (approximate) solutions of vector optimization problems via efficient (exact) solutions of vector-valued perturbed maps. Let us remark that several  $\varepsilon$ -efficiency concepts have been used to this end (see [11, 12] for a unified approach on the  $\varepsilon$ -efficiency notion).

---

\*Received by the editors October 20, 2006; accepted for publication (in revised form) November 7, 2007; published electronically February 27, 2008. This research was partially supported by the Spanish Ministry of Education and Science under projects MTM2006-02629 and Ingenio Mathematica (i-MATH) CSD2006-00032 (Consolider-Ingenio 2010), and by the Consejería de Educación de la Junta de Castilla y León (Spain), project VA027B06.

<http://www.siam.org/journals/sicon/47-2/67286.html>

<sup>†</sup>Departamento de Matemática Aplicada, E.T.S.I. Informática, Universidad de Valladolid, Edificio de Tecnologías de la Información y las Telecomunicaciones, Campus Miguel Delibes, s/n, 47011 Valladolid, Spain (cesargv@mat.uva.es).

<sup>‡</sup>Departamento de Matemática Aplicada, E.T.S.I. Industriales, Universidad Nacional de Educación a Distancia, c/ Juan del Rosal, 12, Ciudad Universitaria, 28040 Madrid, Spain (bjimenez@ind.uned.es, vnov@ind.uned.es).

The EVP obtained in this work is stronger than these previous EVPs from two points of view. First, because a set-valued perturbed map is considered, and second, because, in a certain sense, it does not depend on any  $\varepsilon$ -efficiency concept. Both improvements are attained by considering a set-valued metric notion, which is new to our knowledge, and a cone-boundedness concept based on scalarization.

In proving our results, we follow the fourth previous approach. Specifically, we consider normal and based cones. Let us observe that this is a broad class of cones. Moreover, the normalness and basedness of the order cone are mild hypotheses in the framework of the vector optimization problems when the final space is a Hausdorff locally convex space.

This paper is structured as follows. In section 2, the vector optimization problem and several notations used in what follows are fixed. Moreover, some definitions and technical results are recalled. In section 3, two general set-valued EVPs for vector-valued maps are proved. In section 4, we show several conditions in order to check the assumptions of these set-valued EVPs. In particular, the cone-boundedness concept based on scalarization is related to other well-known cone-boundedness notions, and some specific conditions for the finite-dimensional case are given. In section 5, we relate the EVP attained in section 3 to similar results in the literature. Let us observe that the finite-dimensional case is studied via a new  $\varepsilon$ -efficiency notion. Finally, in section 6 various conclusions are presented which summarize this work.

**2. Preliminaries.** Let  $(X, d)$  be a nontrivial complete metric space and let  $Y$  be a Hausdorff locally convex space. As usual, the topological dual space of  $Y$  is denoted by  $Y^*$ , and for all  $y^* \in Y^*$  and  $y \in Y$  the value  $y^*(y)$  is written as  $\langle y, y^* \rangle$ .

In this work, the following vector optimization problem is considered:

$$(1) \quad \text{Min}\{f(x) : x \in S\},$$

where  $f : X \rightarrow Y$  and  $S$  is a nonempty closed subset of  $X$ . The partial order in  $Y$  is given via a convex cone  $D$  by the relation

$$y_1, y_2 \in Y, \quad y_1 \leq y_2 \iff y_1 - y_2 \in -D.$$

We assume that  $D$  is pointed ( $D \cap (-D) = \{0\}$ ) and nontrivial ( $D \neq \{0\}$ ). Let us recall that  $x_0 \in S$  is an efficient solution of (1) if there is not a feasible point  $x \in S$  such that  $f(x) \leq f(x_0)$ ,  $f(x) \neq f(x_0)$ .

The positive (resp., strict positive) polar cone of  $D$  is denoted by  $D^+$  (resp.,  $D^{+s}$ ). We denote the nonnegative orthant of  $\mathbb{R}^p$  by  $\mathbb{R}_+^p$  and  $\mathbb{R}_+ := [0, \infty)$ . Moreover, the interior, the closure, the convex hull, the affine hull, and the conical hull of a set  $A \subset Y$  are denoted by  $\text{int}(A)$ ,  $\text{cl}(A)$  ( $\text{cl}^w(A)$  if the weak topology is considered),  $\text{conv}(A)$ ,  $\text{aff}(A)$ , and  $\text{cone}(A)$ , respectively. We say that  $D$  is a solid cone if  $\text{int}(D) \neq \emptyset$ .

If  $A, B \subset Y$  and  $\alpha \in \mathbb{R}$ , the sets  $A + B$  and  $\alpha A$  are defined as follows:

$$\begin{aligned} A + B &= \{z \in Y : z = x + y, x \in A, y \in B\}, \\ \alpha A &= \{z \in Y : z = \alpha x, x \in A\}. \end{aligned}$$

For a scalar optimization problem, there exist several equivalent versions of the EVP (see, for instance, [1]), but the first and most popular statement is based on approximate solutions of the problem (see [6, Theorem 1.1]). In the literature, several EVPs for vector optimization problems have been proposed following this approach and this fact has motivated different EVPs for vector-valued maps, since the  $\varepsilon$ -efficiency notion is not unique.

In this paper, the  $(C, \varepsilon)$ -efficiency concept due to Gutiérrez, Jiménez, and Novo [11, 12] is considered, because it extends and unifies several  $\varepsilon$ -efficiency notions (see [11, 12] for more details).

DEFINITION 2.1. A nonempty set  $C \subset Y$  is coradial if  $\cup_{\beta \geq 1} \beta C = C$ .

Let us refer to [24] and the references therein for further details on this class of sets.

DEFINITION 2.2. Let  $C \subset D \setminus \{0\}$  be a coradial set and  $\varepsilon > 0$ . A point  $x_0 \in S$  is a  $(C, \varepsilon)$ -efficient solution of problem (1) if  $(f(S) - f(x_0)) \cap (-\varepsilon C) = \emptyset$ .

In [11, 12] the reader can find several  $\varepsilon$ -efficiency concepts given by different coradial sets. The set of  $(C, \varepsilon)$ -efficient solutions of (1) is denoted by  $AE(C, \varepsilon)$ . It is easy to prove that if  $C$  is a coradial set, then  $\varepsilon_2 C \subset \varepsilon_1 C \ \forall \varepsilon_1, \varepsilon_2 > 0, \varepsilon_1 < \varepsilon_2$ . Therefore,

$$(2) \quad AE(C, \varepsilon_1) \subset AE(C, \varepsilon_2) \quad \forall \varepsilon_1, \varepsilon_2 > 0, \quad \varepsilon_1 < \varepsilon_2.$$

In obtaining an EVP for vector optimization problems, different conditions on the order cone  $D$  are usually required (see section 5). In this paper we assume that  $D$  is  $w$ -normal, and a nontrivial based convex cone  $K \subset D$  is considered.

DEFINITION 2.3. A convex cone  $D \subset Y$  is normal if for all nets  $(x_i), (y_i) \subset Y$  such that  $0 \leq x_i \leq y_i$  for all  $i$  and  $(y_i) \rightarrow 0$ , then one has  $(x_i) \rightarrow 0$ .

A convex cone  $D \subset Y$  is called  $w$ -normal if it verifies Definition 2.3 when the weak topology is considered (and so  $\rightarrow$  is changed by the weak convergence  $\xrightarrow{w}$ ).

DEFINITION 2.4. A convex cone  $K \subset Y$  is based (well-based) if there exists a (bounded) convex set  $B \subset K$ , called base of  $K$ , such that  $0 \notin \text{cl}(B)$  and  $\mathbb{R}_+ B = K$ .

The following results are well known. The reader can find more details about normal and based cones in [9].

PROPOSITION 2.5 (see [9, Proposition 2.2.9]). Let  $D \subset Y$  be a convex cone. Then  $D$  is  $w$ -normal if and only if  $Y^* = D^+ - D^+$ .

PROPOSITION 2.6 (see [9, Theorem 2.2.12]). A convex cone  $K \subset Y$  is based if and only if  $K^{+s} \neq \emptyset$ .

**3. A set-valued EVP for vector optimization problems.** In this section, a set-valued EVP for (1) is obtained, which extends several given in the literature (see section 5). To attain this goal, a set-valued  $D$ -metric concept is introduced.

DEFINITION 3.1. A set-valued map  $F : X \times X \rightarrow 2^D$  is called a set-valued  $D$ -metric (sv- $D$ -metric) if it satisfies the following conditions:

- (a)  $F(x, y) \neq \emptyset$  and  $F(x, x) = \{0\} \ \forall x, y \in X$ , and  $0 \notin F(x, y) \ \forall x \neq y$ .
- (b)  $F(x, y) = F(y, x) \ \forall x, y \in X$ .
- (c)  $F(x, y) + F(y, z) \subset F(x, z) + D \ \forall x, y, z \in X$ .

Example 3.2.

- (a) Let  $r : X \times X \rightarrow D$  be a vector-valued  $D$ -metric (see, for instance, [21, p. 673]). The set-valued map  $F(x, y) = \{r(x, y)\} \ \forall x, y \in X$  is an sv- $D$ -metric, and so Definition 3.1 extends to set-valued maps the vector concept of a  $D$ -metric.
- (b) Let  $H \subset D \setminus \{0\}$  be a nonempty  $D$ -convex set (i.e.,  $H + D$  is a convex set). It follows that  $F(x, y) = d(x, y)H$  is an sv- $D$ -metric.
- (c) If  $F$  is an sv- $D$ -metric, then the set-valued map  $G : X \times X \rightarrow 2^D$  defined by

$$G(x, y) = \begin{cases} F(x, y) + D & \text{if } x \neq y, \\ \{0\} & \text{if } x = y \end{cases}$$

is an sv- $D$ -metric.

Let us denote  $K_F := \text{cone}(\text{conv}(\cup\{F(x, y) : \forall x, y \in X\}))$  and

$$(3) \quad D_F := (K_F \setminus \{0\} + D) \cup \{0\}.$$

Let us observe that  $D_F$  is a convex cone.

Next, a cone-boundedness notion is introduced, which will be required in what follows. In Proposition 4.6 the reader can find some relations between different cone-boundedness concepts.

**DEFINITION 3.3.** *Let  $K \subset Y$  be a convex cone. A set  $M \subset Y$  is said to be  $K$ -bounded by scalarizations ( $K$ -bounded) if*

$$\inf\{\langle y, \xi \rangle : y \in M\} > -\infty \quad \forall \xi \in K^+.$$

In order to prove the announced set-valued EVP, several assumptions on problem (1), a constant  $\gamma > 0$ , an sv- $D$ -metric  $F$ , and a point  $x_0 \in S$  will be considered:

- (A1)  $D$  is  $w$ -normal and  $D_F$  is based.
- (B1)  $D$  is  $w$ -normal and based.
- (A2)  $0 \notin \text{cl}^w(\cup_{d(x,y) \geq \delta} F(x, y)) \forall \delta > 0$ .
- (A3) Let us denote

$$A_x^{\gamma F} := \{z \in X : (f(z) + \gamma F(z, x) - f(x)) \cap (-D) \neq \emptyset\} \quad \forall x \in S.$$

For each  $x \in S$  and  $(z_n) \subset A_x^{\gamma F}$ ,  $(z_n) \rightarrow z$  such that  $f(z_n) \leq f(z_m) \forall n > m$ , it follows that  $z \in A_x^{\gamma F}$ .

- (A4) The set  $(f(S) - f(x_0)) \cap (-D_F)$  is  $D$ -bounded.
- (B4) The set  $(f(S) - f(x_0)) \cap (-D)$  is  $D$ -bounded.

*Remark 3.4.*

- (a) If  $D$  is  $w$ -normal (in particular, when (A1) or (B1) is satisfied), then  $\text{cl}(D)$  is pointed. Indeed, according to [9, Corollary 2.1.23],  $\text{cl}^w(D)$  is pointed, and as  $\text{cl}(D)$  is a convex closed set, one has  $\text{cl}(D) = \text{cl}^w(D)$ .
- (b) It is clear that (B1) implies (A1). However, assumption (A1) is weaker than (B1). Indeed, suppose that  $D$  is closed and take a Banach space  $X$ ,  $d_0 \in D \setminus \{0\}$ , and the sv- $D$ -metric  $F(x, y) = \{\|x - y\|d_0\} \forall x, y \in X$ . Then we have that

$$D_F \setminus \{0\} = \{\alpha d_0 + d : \forall \alpha > 0, \forall d \in D\}.$$

As  $\{-d_0\} \cap D = \emptyset$ , by the separation theorem, there exists  $\lambda \in D^+ \setminus \{0\}$  such that  $\lambda(d_0) > 0$ . It follows that

$$\lambda(\alpha d_0 + d) = \alpha \lambda(d_0) + \lambda(d) \geq \alpha \lambda(d_0) > 0 \quad \forall \alpha > 0, \forall d \in D$$

and so  $D_F^{+s} \neq \emptyset$ , since  $\lambda \in D_F^{+s}$ . Consequently, if  $D$  is  $w$ -normal and  $D^{+s} = \emptyset$ , then (A1) holds and (B1) is not satisfied. Next, we show two examples in which  $D$  is closed and  $w$ -normal and  $D^{+s} = \emptyset$ .

- (i) Let  $Y = B([a, b])$  be the Banach space of all bounded real functions on the compact interval  $[a, b]$  endowed with the supremum norm, and consider the cone  $D = \{y \in Y : y(t) \geq 0 \forall t \in [a, b]\}$  (see [22, p. 27]).
- (ii) Let  $Y = \ell^2(\mathbb{R})$  be the Hilbert space of all square summable functions  $y : \mathbb{R} \rightarrow \mathbb{R}$  endowed with the norm associated to the usual scalar product  $\langle y, z \rangle = \sum_{t \in \mathbb{R}} y(t)z(t) \forall y, z \in Y$ , and consider the cone  $D = \{y \in Y : y(t) \geq 0 \forall t \in \mathbb{R}\}$ . Let us check that  $D^{+s} = \emptyset$ . If  $v \in D^{+s}$ , then  $\langle v, y \rangle > 0$

$\forall y \in D \setminus \{0\}$ . In particular, for each  $\alpha \in \mathbb{R}$  define  $y_\alpha : \mathbb{R} \rightarrow \mathbb{R}$  by  $y_\alpha(\beta) = 1$  if  $\beta = \alpha$  and  $y_\alpha(\beta) = 0$  if  $\beta \neq \alpha$ . It is clear that  $y_\alpha \in D \ \forall \alpha \in \mathbb{R}$ , and  $\langle v, y_\alpha \rangle = v(\alpha) > 0 \ \forall \alpha \in \mathbb{R}$ , but this is a contradiction because if  $v \in \ell^2(\mathbb{R})$ , then  $\{\alpha \in \mathbb{R} : v(\alpha) \neq 0\}$  is countable.

- (c) The convex cones  $D$  and  $D_F$  are very close: one has  $\text{cl}(D) = \text{cl}(D_F)$  and so  $D^+ = D_F^+$ . As a consequence, by Proposition 2.5,  $D$  is  $w$ -normal if and only if  $D_F$  is  $w$ -normal, and a set  $M$  is  $D$ -bounded if and only if  $M$  is  $D_F$ -bounded. Indeed, as  $X$  is nontrivial, there exists  $q \in D \setminus \{0\}$  such that

$$(0, \infty)q + D \subset D_F.$$

Therefore,  $D \subset \text{cl}(D_F)$  and so  $\text{cl}(D) = \text{cl}(D_F)$ .

- (d) It is obvious that (B4) implies (A4). However, hypothesis (A4) is weaker than (B4). Indeed, consider  $X = Y = \mathbb{R}^2$ ,  $D = \mathbb{R}_+^2$ ,  $S = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 x_2 = 0\}$ ,  $f(x) = x \ \forall x \in \mathbb{R}^2$ ,  $F(x, u) = \{\|x - u\|(1, 1)\} \ \forall x, u \in \mathbb{R}^2$ , and  $x_0 = (0, 0)$ . We have that  $(f(S) - f(x_0)) \cap (-\mathbb{R}_+^2)$  is not  $\mathbb{R}_+^2$ -bounded. However,  $D_F = \text{int}(\mathbb{R}_+^2) \cup \{0\}$  and so  $(f(S) - f(x_0)) \cap (-D_F) = \{0\}$ , which is  $\mathbb{R}_+^2$ -bounded.

Next we give some useful conditions to check if  $D_F$  is based. The proof of Lemma 3.5 is immediate and so it is omitted. Lemma 3.6(b)–(d) extends Lemma 4.20 in [3].

LEMMA 3.5. *Let  $D_F$  be the convex cone given by (3). Then*

- (a)  $D_F^{+s} = K_F^{+s} \cap D^+$ ,  
 (b)  $D_F$  is based if and only if  $K_F^{+s} \cap D^+ \neq \emptyset$ .

LEMMA 3.6. *Let  $K \subset D$  be a nonempty convex cone. The following statements all hold.*

- (a) *If there exists a  $D$ -convex set  $B_K$  such that  $\mathbb{R}_+ B_K = K$  and  $0 \notin \text{cl}(B_K + D)$ , then  $K^{+s} \cap D^+ \neq \emptyset$ .*  
 (b) *If  $D$  is closed and there exists a  $D$ -convex compact set  $B_K$  such that  $\mathbb{R}_+ B_K = K$  and  $0 \notin B_K$ , then  $K^{+s} \cap D^+ \neq \emptyset$ .*  
 (c) *If  $D$  is normal and there exists a  $D$ -convex set  $B_K$  such that  $\mathbb{R}_+ B_K = K$  and  $0 \notin \text{cl}(B_K)$ , then  $K^{+s} \cap D^+ \neq \emptyset$ .*  
 (d) *If  $D$  is solid and  $K \setminus \{0\} \subset \text{int}(D)$ , then  $D^+ \setminus \{0\} \subset K^{+s}$ .*

*Proof.* Part (a). It is clear that

$$(\mathbb{R}_+(B_K + D))^{+s} = (\mathbb{R}_+ B_K)^{+s} \cap D^+ = K^{+s} \cap D^+.$$

Then, the conclusion follows since  $\mathbb{R}_+(B_K + D)$  is based.

Part (b). The set  $B_K + D$  is closed because it is the sum of a compact set and a closed set, and  $B_K + D \subset D \setminus \{0\} + D \subset D \setminus \{0\}$ . Therefore, the result follows from part (a).

Part (c). Since  $D$  is normal and  $B_K \subset D$  we deduce that  $0 \in \text{cl}(B_K)$  whenever  $0 \in \text{cl}(B_K + D)$ . Therefore,  $0 \notin \text{cl}(B_K + D)$  and the result follows from part (a).

Part (d) is well known (see, for instance, [3, Lemma 4.20(xiii)]).  $\square$

The next lemma is useful to prove Theorem 3.8. Consider  $x \in S$ ,  $\gamma > 0$ , an sv-D-metric  $F$ , and

$$(4) \quad S_x := \{z \in S \setminus \{x\} : (f(z) + \gamma F(z, x) - f(x)) \cap (-D) \neq \emptyset\}.$$

LEMMA 3.7. *The following properties are satisfied:*

- (a)  $S_z \subset S_x \ \forall z \in S_x$ .

- (b) Let  $x_0 \in S$  and suppose that (A4) holds. Then  $\xi \circ f$  is bounded from below on  $S_x \forall x \in S_{x_0} \cup \{x_0\}$  and  $\forall \xi \in D^+$ , and for each  $x \in S$  such that  $S_x \neq \emptyset$  it follows that

$$\langle f(x), \lambda \rangle > \inf_{z \in S_x} \{\langle f(z), \lambda \rangle\} > -\infty \quad \forall \lambda \in D_F^{+s}.$$

*Proof.* Part (a). If  $z \in S_x$  and  $u \in S_z$ , there exists  $d \in F(u, z)$  such that  $f(u) \leq f(z) - \gamma d$ . As  $z \in S_x$  there exists  $q \in F(z, x)$  such that  $f(z) \leq f(x) - \gamma q$ . By adding these relations we see that

$$(5) \quad f(u) \leq f(x) - \gamma(d + q).$$

From Definition 3.1 it follows that  $d + q \in F(u, x) + D$  and by (5) we derive that  $(f(u) + \gamma F(u, x) - f(x)) \cap (-D) \neq \emptyset$ . If  $u = x$ , then  $q = 0$ , which is a contradiction since  $q \in F(z, x)$  and  $z \neq x$ . Then  $u \neq x$  and  $u \in S_x$ .

Part (b). Let  $x \in S_{x_0} \cup \{x_0\}$ . By part (a) it is clear that

$$\inf\{\langle f(z), \xi \rangle : z \in S_x\} \geq \inf\{\langle f(z), \xi \rangle : z \in S_{x_0}\} \quad \forall \xi \in D^+.$$

Moreover, from the definitions of the sets  $S_{x_0}$  and  $D_F$  we see that  $f(S_{x_0}) - f(x_0) \subset -D_F$ . Then, by (A4) we have that

$$\begin{aligned} \inf\{\langle f(z), \xi \rangle : z \in S_{x_0}\} &= \langle f(x_0), \xi \rangle + \inf\{\langle f(z) - f(x_0), \xi \rangle : z \in S_{x_0}\} \\ &= \langle f(x_0), \xi \rangle + \inf\{\langle y, \xi \rangle : y \in (f(S_{x_0}) - f(x_0)) \cap (-D_F)\} \\ &\geq \langle f(x_0), \xi \rangle + \inf\{\langle y, \xi \rangle : y \in (f(S) - f(x_0)) \cap (-D_F)\} > -\infty \quad \forall \xi \in D^+ \end{aligned}$$

and  $\xi \circ f$  is bounded from below on  $S_x \forall \xi \in D^+$ .

For the second part of (b), let  $u \in S_x$ . Then  $u \neq x$  and there exist  $q \in F(u, x)$ ,  $d \in D$  such that  $f(u) + \gamma q - f(x) = -d$ . As  $F$  is an sv- $D$ -metric we see that  $q \neq 0$  and so  $\gamma q + d \in D_F \setminus \{0\}$ . Therefore,

$$\langle f(x), \lambda \rangle > \langle f(u), \lambda \rangle \geq \inf_{z \in S_x} \{\langle f(z), \lambda \rangle\} > -\infty \quad \forall \lambda \in D_F^{+s},$$

where the last inequality is a consequence of the first part of (b) and Lemma 3.5(a).  $\square$

**THEOREM 3.8.** Consider  $\gamma > 0$ , an sv- $D$ -metric  $F$ , and a point  $x_0 \in S$  satisfying assumptions (A1)–(A4). Then, there exists  $\hat{x} \in S$  such that

- (a)  $(f(\hat{x}) + \gamma F(\hat{x}, x_0) - f(x_0)) \cap (-D) \neq \emptyset$ ,  
 (b)  $(f(\hat{x}) - f(x) - \gamma F(x, \hat{x})) \cap D = \emptyset \quad \forall x \in S \setminus \{\hat{x}\}$ .

*Proof.* By Proposition 2.6 we see that  $D_F^{+s} \neq \emptyset$  and then a functional  $\lambda \in D_F^{+s}$  can be fixed. For this functional, a sequence  $(x_n) \subset S$  is attained via an iterative process.

Suppose that  $x_1, x_2, \dots, x_n$  are defined in such a way that  $x_1 = x_0$  and for each  $i = 2, 3, \dots, n$ ,  $x_i \in S_{x_{i-1}}$ , where the sets  $S_{x_i}$  are given by (4). The point  $x_{n+1}$  is defined as follows. If  $S_{x_n} = \emptyset$ , then  $x_{n+1} := x_n$  and so  $(x_n)$  is stationary.

If  $S_{x_n} \neq \emptyset$ , then by Lemma 3.7(b) we deduce that

$$(6) \quad \langle f(x_n), \lambda \rangle > \inf_{x \in S_{x_n}} \{\langle f(x), \lambda \rangle\} > -\infty.$$

Then, in view of (6) the point  $x_{n+1} \in S_{x_n}$  can be chosen such that

$$(7) \quad \langle f(x_{n+1}), \lambda \rangle - \inf_{x \in S_{x_n}} \{\langle f(x), \lambda \rangle\} \leq (1/2) \left( \langle f(x_n), \lambda \rangle - \inf_{x \in S_{x_n}} \{\langle f(x), \lambda \rangle\} \right).$$

Let us prove that  $(x_n)$  is a Cauchy sequence. Indeed, suppose that there exists  $\delta > 0$  such that  $\forall k \in \mathbb{N}$  there exists  $n_k \in \mathbb{N}$ ,  $n_k > k$ , verifying  $d(x_{n_k}, x_k) \geq \delta$ . From Lemma 3.7(a) we see that  $S_{x_{n_k-1}} \subset S_{x_k}$ , and as  $x_{n_k} \in S_{x_{n_k-1}}$  we have that there exists  $d_{n_k} \in F(x_{n_k}, x_k)$  such that  $f(x_{n_k}) \leq f(x_k) - \gamma d_{n_k}$ . By Lemma 3.7(a), (b) we deduce that the real-valued sequence  $(\langle f(x_k), \xi \rangle)$  is lower bounded  $\forall \xi \in D^+$  because  $x_k \in S_{x_0}$ . Moreover,  $f(x_{k+1}) \in f(x_k) - D_F \forall k$  and so  $(\langle f(x_k), \xi \rangle)$  is nonincreasing. Thus,  $(\langle f(x_k), \xi \rangle)$  has a limit and it follows that

$$0 \leq \gamma \langle d_{n_k}, \xi \rangle \leq \langle f(x_k), \xi \rangle - \langle f(x_{n_k}), \xi \rangle \rightarrow 0 \quad \forall \xi \in D^+.$$

By Proposition 2.5 we deduce that  $(d_{n_k}) \xrightarrow{w} 0$ , which is a contradiction to (A2).

Therefore,  $(x_n)$  is a Cauchy sequence and so there exists a point  $\hat{x} \in S$  such that  $(x_n) \rightarrow \hat{x} \in S$ . Let us prove that  $\hat{x}$  verifies properties (a) and (b).

As  $x_m \in S_{x_n} \forall m \geq n+1$  and  $f(x_k) \leq f(x_m) \forall k > m \geq n+1$ , it follows that  $\hat{x} \in S_{x_n} \cup \{x_n\} \forall n$ , since assumption (A3) is satisfied. In particular,  $\hat{x} \in S_{x_0} \cup \{x_0\}$  and so

$$(f(\hat{x}) + \gamma F(\hat{x}, x_0) - f(x_0)) \cap (-D) \neq \emptyset.$$

Next, let us suppose, contrary to part (b), that there exists  $x \in S_{\hat{x}}$ , i.e.,  $f(x) = f(\hat{x}) - \gamma d - d'$  for some  $x \in S \setminus \{\hat{x}\}$ ,  $d \in F(x, \hat{x})$ , and  $d' \in D$ . As  $\gamma d + d' \in K_F \setminus \{0\} + D = D_F \setminus \{0\}$  and  $\lambda \in D_F^{+s}$  we deduce that

$$(8) \quad \langle f(x), \lambda \rangle < \langle f(\hat{x}), \lambda \rangle.$$

As  $x \in S_{\hat{x}}$  and  $\hat{x} \in S_{x_n} \cup \{x_n\} \forall n$  by Lemma 3.7(a) we deduce that  $x \in S_{x_n} \forall n \geq 1$ . Thus, from (7) we see that

$$2\langle f(x_{n+1}), \lambda \rangle - \langle f(x_n), \lambda \rangle \leq \inf_{z \in S_{x_n}} \{\langle f(z), \lambda \rangle\} \leq \langle f(x), \lambda \rangle \quad \forall n$$

and  $\lim \langle f(x_n), \lambda \rangle \leq \langle f(x), \lambda \rangle$ . As  $\hat{x} \in S_{x_n} \cup \{x_n\}$ , we have that  $\langle f(\hat{x}), \lambda \rangle \leq \langle f(x_n), \lambda \rangle \forall n$  and it follows that  $\langle f(\hat{x}), \lambda \rangle \leq \lim \langle f(x_n), \lambda \rangle$ . Therefore,  $\langle f(\hat{x}), \lambda \rangle \leq \langle f(x), \lambda \rangle$ , contrary to (8), and the proof is finished.  $\square$

*Remark 3.9.* Let us observe from Remark 3.4(b), (d) that Theorem 3.8 is also true if assumptions (A1) and (A4) are changed to (B1) and (B4), respectively.

*Remark 3.10.* Let us observe that Theorem 3.8(b) can be rewritten by saying that  $(\hat{x}, f(\hat{x}))$  is a minimizer of the following set-valued optimization problem (see [18, Definition 14.2] for more details):

$$\text{Min}\{f(x) + \gamma F(x, \hat{x}) : x \in S\}.$$

Moreover,  $(\hat{x}, f(\hat{x}))$  is a strong minimizer, since it is a minimizer and  $f(x) + \gamma d \neq f(\hat{x}) \forall x \in S \setminus \{\hat{x}\}$ ,  $\forall d \in F(x, \hat{x})$ . Therefore, as usual in the context of EVP, part (b) of Theorem 3.8 consists of perturbing the objective function  $f$  by a map  $g := \gamma F(\cdot, \hat{x})$  in such a way that  $f + g$  attains a strong minimum on  $S$  at  $\hat{x}$ . The advantage of Theorem 3.8(b) as opposed to similar principles in the literature is that the perturbation  $g$  is a set-valued map (see Example 5.2).

Next, motivated by the results obtained in [1] for scalar optimization problems, we present a different version of Theorem 3.8 based on  $(C, \varepsilon)$ -efficient solutions. Previously, a simple relation between  $D$ -bounded sets and  $(C, \varepsilon)$ -efficient points is established, from which one can easily deduce if a feasible point is an approximate efficient

solution of (1). In this sense, let us observe that if (1) is a scalar optimization problem ( $Y = \mathbb{R}$ ), then all feasible points are approximate solutions if and only if the problem is lower bounded, and the problem is lower bounded if and only if there exists some approximate solution. However,  $(C, \varepsilon)$ -efficient and not  $(C, \varepsilon)$ -efficient feasible points are possible in the same vector (nonscalar) optimization problem, or it is even possible to find a  $(C, \varepsilon)$ -efficient solution  $x_0$  whose section  $(f(x_0) - D) \cap f(S)$  is not bounded.

LEMMA 3.11. *Let us consider a coradial set  $C \subset D \setminus \{0\}$ .*

- (a) *Suppose that  $K \subset Y$  is a  $w$ -normal convex cone such that  $C \subset K$  and let  $x_0 \in S$ . If  $0 \notin \text{cl}^w(C)$  and  $(f(S) - f(x_0)) \cap (-K)$  is  $K$ -bounded, then there exists  $\varepsilon > 0$  such that  $x_0 \in \text{AE}(C, \varepsilon)$ .*
- (b) *Let  $K \subset Y$  be a convex cone and suppose that there exists  $\varepsilon > 0$  such that  $x_0 \in \text{AE}(C, \varepsilon)$ . Then  $(f(S) - f(x_0)) \cap (-K)$  is  $K$ -bounded if and only if  $(f(S) - f(x_0)) \cap (-\varepsilon(K \setminus C))$  is  $K$ -bounded.*

*Proof.* Part (a). Let us suppose that  $(f(S) - f(x_0)) \cap (-K)$  is  $K$ -bounded and  $x_0 \notin \text{AE}(C, \varepsilon) \forall \varepsilon > 0$ . Then, for each  $n \in \mathbb{N}$  there exists  $c_n \in C$  such that  $-nc_n \in (f(S) - f(x_0)) \cap (-K)$ . By the  $K$ -boundedness of this set we deduce for each  $\xi \in K^+$  that there exists  $m_\xi \in \mathbb{R}$  such that

$$m_\xi \leq -n \langle c_n, \xi \rangle \quad \forall n.$$

Therefore  $(\langle c_n, \xi \rangle) \rightarrow 0 \forall \xi \in K^+$ . By Proposition 2.5 it follows that  $(c_n) \xrightarrow{w} 0$  and so  $0 \in \text{cl}^w(C)$ , which is a contradiction.

Part (b). It is clear that

$$K = (K \setminus (\varepsilon C)) \cup (K \cap \varepsilon C) = (\varepsilon(K \setminus C)) \cup (K \cap \varepsilon C).$$

As  $x_0 \in \text{AE}(C, \varepsilon)$ , we have that

$$(f(S) - f(x_0)) \cap (-K) = (f(S) - f(x_0)) \cap (-\varepsilon(K \setminus C)),$$

and the result follows.  $\square$

The assumption  $0 \notin \text{cl}^w(C)$  is important in order to consider appropriate approximate efficiency concepts. Let us observe that in the other case, it is possible to obtain minimizing sequences  $(x_n)$  such that  $f(x_{n+1}) - f(x_n) \in -\varepsilon C \forall n$ ; i.e., there could be feasible sequences that tend to efficient points and whose elements are not  $(C, \varepsilon)$ -efficient solutions for a fixed precision  $\varepsilon > 0$ .

COROLLARY 3.12. *Let  $C \subset D \setminus \{0\}$  be a coradial set and suppose that  $-(D \setminus C)$  is  $D$ -bounded. If there exists  $\varepsilon > 0$  such that  $x_0 \in \text{AE}(C, \varepsilon)$ , then  $(f(S) - f(x_0)) \cap (-D)$  is  $D$ -bounded. In particular, the assumption on the boundedness of  $-(D \setminus C)$  is satisfied if there exists  $B \subset D$  such that  $-B$  is  $D$ -bounded,  $D = \mathbb{R}_+ B$ , and  $C = [1, \infty)B$ .*

*Proof.* It is clear that  $(f(S) - f(x_0)) \cap (-\varepsilon(D \setminus C))$  is  $D$ -bounded, since

$$(f(S) - f(x_0)) \cap (-\varepsilon(D \setminus C)) \subset -\varepsilon(D \setminus C)$$

and  $-(D \setminus C)$  is  $D$ -bounded. Then, by Lemma 3.11(b) with  $K = D$  we deduce that  $(f(S) - f(x_0)) \cap (-D)$  is  $D$ -bounded.

Next, let us suppose that  $D = \mathbb{R}_+ B$  and  $C = [1, \infty)B$ , where  $-B$  is  $D$ -bounded. It is obvious that

$$-(D \setminus C) \subset \bigcup_{0 \leq \beta < 1} (-\beta B),$$

and the result follows since  $-B$  is  $D$ -bounded.  $\square$



In some vector optimization problems, the order cone  $D$  is well-based (see Definition 2.4). In this case, if  $B$  is a base of  $D$ , by defining  $C = B + D$  we obtain that a feasible point  $x_0$  is a  $(C, \varepsilon)$ -efficient solution for some  $\varepsilon > 0$  if and only if the section  $(f(S) - f(x_0)) \cap (-D)$  is  $D$ -bounded. This fact is established in the following corollary. Unfortunately, the well-based property is not verified for several usual order cones (see [4, section 3] for more details).

**COROLLARY 3.13.** *Suppose that  $D$  is well-based with respect to the weak topology through the base  $B$ . Consider  $C = B + D$  and  $x_0 \in S$ . Then, there exists  $\varepsilon > 0$  such that  $x_0 \in \text{AE}(C, \varepsilon)$  if and only if the section  $(f(S) - f(x_0)) \cap (-D)$  is  $D$ -bounded.*

*Proof.* It is well known that  $D$  is  $w$ -normal when  $D$  is well-based (see, for example, [9, Proposition 2.2.15]). Moreover,  $C$  is a coradial set and  $0 \notin \text{cl}^w(C)$  since  $C = B + D$ ,  $D$  is  $w$ -normal, and  $0 \notin \text{cl}^w(B)$ . Therefore, by Lemma 3.11(a) with  $K = D$ , if  $(f(S) - f(x_0)) \cap (-D)$  is  $D$ -bounded, then there exists  $\varepsilon > 0$  such that  $x_0 \in \text{AE}(C, \varepsilon)$ .

Reciprocally, as  $B$  is bounded with respect to the weak topology, it follows that  $-B$  is  $D$ -bounded. Moreover, from the convexity of  $B$  and the relation  $D = \mathbb{R}_+ B$  we have that  $C = [1, \infty)B$ . Thus, by Corollary 3.12 we deduce that the section  $(f(S) - f(x_0)) \cap (-D)$  is  $D$ -bounded if there exists  $\varepsilon > 0$  such that  $x_0 \in \text{AE}(C, \varepsilon)$ .  $\square$

Let us consider the following assumptions on problem (1), an  $sv$ - $D$ -metric  $F$ , a coradial set  $C \subset D \setminus \{0\}$ , and a feasible point  $x_0 \in S$ :

(A5)  $\bigcup_{x, y \in X} F(x, y) \subset \text{cone}(C)$  and  $C + D = C$ .

(A6) There exists  $\varepsilon > 0$  such that  $x_0 \in \text{AE}(C, \varepsilon)$  and  $(f(S) - f(x_0)) \cap (-\varepsilon(D_F \setminus C))$  is  $D$ -bounded.

(B6) There exists  $\varepsilon > 0$  such that  $x_0 \in \text{AE}(C, \varepsilon)$  and  $(f(S) - f(x_0)) \cap (-\varepsilon(D \setminus C))$  is  $D$ -bounded.

**THEOREM 3.14.** *If assumptions (A1)–(A3), (A5), and (A6) are satisfied, then there exists  $x_\varepsilon \in S$  such that*

(a)  $(f(x_\varepsilon) + \gamma F(x_\varepsilon, x_0) - f(x_0)) \cap (-D) \neq \emptyset$  (and so  $f(x_\varepsilon) \leq f(x_0)$ ),

(b)  $F(x_\varepsilon, x_0) \cap (\text{cone}(C) \setminus ((\varepsilon/\gamma)C)) \neq \emptyset$ ,

(c)  $(f(x_\varepsilon) - f(x) - \gamma F(x, x_\varepsilon)) \cap D = \emptyset \ \forall x \in S \setminus \{x_\varepsilon\}$ .

*Proof.* By Remark 3.4(c) and Lemma 3.11(b) with  $D_F$  instead of  $K$ , it follows that assumption (A4) is also satisfied. Then, by applying Theorem 3.8 one obtains  $x_\varepsilon \in S$ , satisfying parts (a) and (c).

Part (b) is trivial if  $x_\varepsilon = x_0$ . If  $x_\varepsilon \neq x_0$ , then by part (a) we deduce that there exist  $d \in F(x_\varepsilon, x_0)$  and  $d' \in D$  such that  $f(x_\varepsilon) - f(x_0) = -\gamma d - d'$ . Since  $x_0 \in \text{AE}(C, \varepsilon)$  we have that  $\gamma d + d' \notin \varepsilon C$ . If  $d \in (\varepsilon/\gamma)C$ , then  $\gamma d + d' \in \varepsilon C + D = \varepsilon C$ , a contradiction. Hence  $d \in \text{cone}(C) \setminus ((\varepsilon/\gamma)C)$ , that is, (b) holds.  $\square$

**Remark 3.15.** Let us notice that Theorem 3.14 is also true if assumptions (A1) and (A6) are changed to (B1) and (B6), respectively.

In the proof of Theorem 3.14 it was deduced that the hypotheses of Theorem 3.8 hold when the hypotheses of Theorem 3.14 are satisfied, and so the former hypotheses are weaker than the latter. However, by Lemma 3.11 with  $D_F$  instead of  $K$  and Remark 3.4(c), both collections of hypotheses are equivalent when statement (A5) is satisfied,  $0 \notin \text{cl}^w(C)$ , and  $C \subset D_F$ . This fact motivates the following definition.

**DEFINITION 3.16.** *Let  $D$  be a nontrivial pointed convex cone, let  $C \subset D \setminus \{0\}$  be a coradial set such that  $0 \notin \text{cl}^w(C)$ , and let  $F$  be an  $sv$ - $D$ -metric. We say that  $D$ ,  $C$ , and  $F$  are compatible if statement (A5) is satisfied and  $C \subset D_F$ .*

Theorem 3.14 has been proved by using Theorem 3.8. Reciprocally, it is obvious that Theorem 3.8 can be proved via Theorem 3.14 when  $D$ ,  $C$ , and  $F$  are compatible.

This fact is shown in the following proposition by saying that both theorems are equivalent.

**PROPOSITION 3.17.** *Assume that  $D$ ,  $C$ , and  $F$  are compatible. Then Theorems 3.8 and 3.14 are equivalent.*

*Remark 3.18.* In Proposition 3.17, we have extended to vector optimization problems an equivalence between two versions of the original (scalar) EVP (see [1, pp. 345–346] for more details). Indeed, let us suppose that (1) is a scalar optimization problem. In this problem,  $D = \mathbb{R}_+$  and the usual concept of approximate solution is obtained by considering  $C = [1, \infty)$ . By taking  $F(x, y) = \{d(x, y)\}$  it is clear that  $D$ ,  $C$ , and  $F$  are compatible, and so Theorems 3.8 and 3.14 are equivalent. Let us observe that (A1) and (A2) are satisfied, and if  $f$  is lower semicontinuous, then (A3) is verified too. Finally,

$$(A4) \iff (A6) \iff \inf_{x \in S} \{f(x)\} > -\infty.$$

**4. On the assumptions of the set-valued EVP.** In this section we show some conditions from which one can check if a vector optimization problem satisfies the hypotheses of Theorems 3.8 and 3.14.

The reader can deduce from Remark 3.18 that hypothesis (A3) is a kind of cone lower semicontinuity of the map  $f$ . Next, we give a condition from which one can prove that hypothesis.

**DEFINITION 4.1** (see [9, Definition 2.5.7(iv)]). *Let  $G : X \rightarrow 2^Y$  be a set-valued map.  $G$  is said to be (sequentially) compact at  $x \in X$  if for every sequence  $(x_n, y_n) \subset X \times Y$  such that  $y_n \in G(x_n) \forall n$  and  $(x_n) \rightarrow x$ , there exists a subsequence  $(y_{n_k})$  converging to some  $y \in G(x)$ .  $G$  is said to be compact if it is compact at every  $x \in X$ .*

*Remark 4.2.* Let  $g : X \rightarrow \mathbb{R}$  be a continuous functional and let  $H \subset Y$  be a sequentially compact set. It is easy to check that the set-valued map  $G : X \rightarrow 2^Y$  defined by  $G(x) = g(x)H \forall x \in X$  is compact.

**DEFINITION 4.3** (see [21, p. 674]). *A vector-valued map  $f : X \rightarrow Y$  is said to be sequentially submonotone with respect to  $D$  (submonotone) if for every  $x \in X$  and for each sequence  $(x_n)$  such that  $(x_n) \rightarrow x$  and  $f(x_n) \leq f(x_m) \forall n > m$ , it follows that  $f(x) \leq f(x_n) \forall n$ .*

Let us observe that a  $D$ -lower semicontinuous vector-valued map  $f : X \rightarrow Y$  (i.e., a vector-valued map  $f$  such that the sets  $\{x \in X : f(x) \leq y\}$  are closed  $\forall y \in Y$ ) is submonotone.

**THEOREM 4.4.** *If  $D$  is closed, the set-valued map  $F(\cdot, x)$  is compact  $\forall x \in X$ , and the map  $f$  is submonotone, then (A3) is satisfied.*

*Proof.* Let us consider  $x \in S$  and a sequence  $(z_n) \subset A_x^{\gamma F}$  such that  $(z_n) \rightarrow z$  and  $f(z_m) \leq f(z_n) \forall m > n$ . Then there exists a sequence  $(d_n)$  such that  $d_n \in F(z_n, x)$  and  $f(z_n) \in -\gamma d_n + f(x) - D \forall n$ . As  $f$  is submonotone, we deduce that

$$f(z) \leq f(z_n) \leq -\gamma d_n + f(x) \quad \forall n.$$

Since the set-valued map  $F(\cdot, x)$  is compact, we see that there exist a point  $d \in Y$  and a subsequence  $(d_{n_k})$  such that  $(d_{n_k}) \rightarrow d \in F(z, x)$ . Therefore, taking the limit when  $n_k \rightarrow \infty$ , it follows that

$$f(z) + \gamma d \in f(x) - D,$$

since  $D$  is closed. Thus  $z \in A_x^{\gamma F}$  and the proof is completed.  $\square$

Next, some simple relations between different  $D$ -boundedness concepts are introduced, which show that the notion stated in Definition 3.3 is weaker than others more usual in the literature.

DEFINITION 4.5. *Let  $M$  be a subset of  $Y$ .*

- (a) (See [20, Definition 3.1, p. 13].)  *$M$  is called topologically  $D$ -bounded if for each neighborhood  $V \subset Y$  of zero there exists  $\alpha > 0$  such that  $M \subset \alpha V + D$ . In what follows we say that  $M$  is strong (resp., weak)  $D$ -bounded if  $M$  is topologically  $D$ -bounded with respect to the strong (resp., weak) topology on  $Y$ .*
- (b) (See [9, p. 15].)  *$M$  is called lower order bounded if there exists  $y \in Y$  such that  $M \subset y + D$ .*

PROPOSITION 4.6. *Consider a set  $M \subset Y$ . Then*

- (a) *if  $M$  is lower order bounded, then  $M$  is strong  $D$ -bounded.*
- (b) *if  $M$  is strong  $D$ -bounded, then  $M$  is weak  $D$ -bounded.*
- (c) *if  $M$  is weak  $D$ -bounded, then  $M$  is  $D$ -bounded by scalarization.*
- (d) *assuming that  $M \subset -D$  and  $D$  is  $w$ -normal, if  $M$  is  $D$ -bounded by scalarization, then  $M$  is weak  $D$ -bounded.*

*Proof.* Part (a) is clear because every neighborhood of zero in  $Y$  is absorbing. Part (b) is obvious.

Part (c). Let us suppose that  $M$  is weak  $D$ -bounded. For each  $\xi \in D^+$ , the set

$$V_\xi = \{y \in Y : |\langle y, \xi \rangle| < 1\}$$

is a weak neighborhood of zero. As  $M$  is weak  $D$ -bounded, there exists  $\alpha_\xi > 0$  such that  $M \subset \alpha_\xi V_\xi + D$ . Thus,

$$\begin{aligned} \inf\{\langle y, \xi \rangle : y \in M\} &\geq \inf\{\langle \alpha_\xi z + d, \xi \rangle : z \in V_\xi, d \in D\} \\ &= \alpha_\xi \inf\{\langle z, \xi \rangle : z \in V_\xi\} \geq -\alpha_\xi, \end{aligned}$$

and so  $M$  is  $D$ -bounded by scalarization.

Part (d). To obtain a contradiction, consider  $\xi \in Y^*$  and suppose that

$$M \not\subset \{y \in Y : |\langle y, \xi \rangle| < r\} + D \quad \forall r > 0.$$

Then there exists  $(y_n) \subset M$  such that  $|\langle y_n, \xi \rangle| \geq n \quad \forall n$ . Since  $D$  is  $w$ -normal, by Proposition 2.5 we deduce that  $\xi = \xi_1 - \xi_2$  for some  $\xi_1, \xi_2 \in D^+$ . Thus,  $|\langle y_n, \xi_1 - \xi_2 \rangle| \geq n \quad \forall n$ , and we can suppose (by taking a subsequence and by changing  $\xi_1$  to  $\xi_2$  and  $\xi_2$  to  $\xi_1$  if necessary) that

$$\langle y_n, \xi_1 \rangle - \langle y_n, \xi_2 \rangle \geq n \quad \forall n.$$

Then, as  $M \subset -D$  we deduce that

$$\langle y_n, \xi_2 \rangle \leq \langle y_n, \xi_1 \rangle - n \leq -n \quad \forall n,$$

and so

$$\inf\{\langle y, \xi_2 \rangle : y \in M\} = -\infty,$$

which is a contradiction because  $M$  is  $D$ -bounded by scalarization. Therefore,  $M$  is weak  $D$ -bounded and the proof is finished.  $\square$

To finish this section, we focus on finite-dimensional vector optimization problems. In this framework, an  $sv$ - $D$ -metric is proposed from which hypotheses (A1)–(A3) are satisfied if the objective function  $f$  is submonotone.

THEOREM 4.7. Assume that  $Y$  is finite-dimensional.

- (a) If  $\text{cl}(D)$  is pointed, then  $D$  is normal,  $D^{+s} \neq \emptyset$ , and  $D_F^{+s} \neq \emptyset \forall \text{sv-}D\text{-metric } F$ .
- (b) (A1) is satisfied if and only if  $\text{cl}(D)$  is pointed.
- (c) Assume that  $D$  is closed. Then  $D$  is well-based through each set

$$B_\xi = \{d \in D : \langle d, \xi \rangle = 1\} \quad \forall \xi \in D^{+s}.$$

- (d) If there exists a bounded set  $B \subset D$  such that  $D = \mathbb{R}_+ B$  and  $0 \notin \text{cl}(\text{conv}(B))$ , then  $\text{cl}(D)$  is pointed.

*Proof.* Part (a). By [9, Corollary 2.2.11 and Example 1.1.2] we see that  $D$  is normal and  $D^{+s} \neq \emptyset$ . Then  $D_F^{+s} \neq \emptyset$  since  $D_F \subset D \forall \text{sv-}D\text{-metric } F$ .

Part (b) is a consequence of part (a) and Remark 3.4(a).

Part (c). By part (a) we see that  $D^{+s} \neq \emptyset$ . For each  $\xi \in D^{+s}$  it is obvious that  $B_\xi$  is convex,  $0 \notin \text{cl}(B_\xi)$ , and  $D = \mathbb{R}_+ B_\xi$ . Then the proof is completed if we prove that  $B_\xi$  is bounded. Suppose that  $B_\xi$  is not bounded for some  $\xi \in D^{+s}$ ; then there exists  $(b_n) \subset B_\xi$  with  $\|b_n\| \rightarrow \infty$ . Taking a subsequence if necessary, we have  $\|b_n\|^{-1} b_n \rightarrow y \in \text{cl}(D) \setminus \{0\} = D \setminus \{0\}$ . Since  $\|b_n\|^{-1} = \xi(\|b_n\|^{-1} b_n)$  we get the contradiction  $\xi(y) = 0$ .

Part (d). Set  $B_1 = \text{cl}(\text{conv}(B))$ . Then  $B_1$  is a compact convex set with  $0 \notin B_1$ , and the convex cone  $D_1 = \mathbb{R}_+ B_1$  is closed (see, for instance, [14, Proposition 1.4.7, p. 102]) and based. Thus, it is pointed (see [9, Example 1.1.2]). As  $\text{cl}(D) \subset D_1$ , it follows that  $\text{cl}(D)$  is pointed.  $\square$

Let  $\mathcal{D}$  be the family of nonempty finite sets  $\Delta \subset D^{+s}$ , let  $\varphi_\Delta : Y \rightarrow \mathbb{R}$  be the functional defined by  $\varphi_\Delta(y) = \min_{\xi \in \Delta} \{\langle y, \xi \rangle\} \forall y \in Y$ , and denote

$$B_\Delta := \{y \in D : \varphi_\Delta(y) = 1\}.$$

The following facts are clear:  $\varphi_\Delta$  is positively homogeneous and concave,  $\mathbb{R}_+ B_\Delta = D = \{y \in D : \varphi_\Delta(y) \geq 0\}$ , and

$$(9) \quad B_\Delta + D = \{y \in D : \varphi_\Delta(y) \geq 1\}.$$

Moreover,  $B_\Delta \subset \cup_{\xi \in \Delta} B_\xi$ , and by Theorem 4.7(c) it is clear that  $B_\Delta$  is compact when  $\dim Y < \infty$  and  $D$  is closed.

COROLLARY 4.8. Assume that  $Y$  is finite-dimensional. If  $D$  is closed, then the family  $\mathcal{D}$  is not empty and for each  $\Delta \in \mathcal{D}$ , the map  $F_{B_\Delta}(x, y) = d(x, y)B_\Delta$  is an sv- $D$ -metric such that hypotheses (A1)–(A2) are satisfied. If, in addition,  $f$  is submonotone, then (A3) is also satisfied.

*Proof.* By Theorem 4.7(a) we have that  $D^{+s} \neq \emptyset$ . Therefore the family  $\mathcal{D}$  is not empty. Moreover, by Example 3.2(b) the map  $F_{B_\Delta}$  is an sv- $D$ -metric  $\forall \Delta \in \mathcal{D}$ , since  $0 \notin B_\Delta$  and the sets  $B_\Delta$  are  $D$ -convex, taking into account (9) and that  $\varphi_\Delta$  is concave.

Condition (A1) is satisfied by Theorem 4.7(b). Fix  $\Delta \in \mathcal{D}$  and  $\delta > 0$ ; then

$$(10) \quad \cup \{F_{B_\Delta}(x, y) : d(x, y) \geq \delta\} \subset [\delta, \infty)B_\Delta = \{y \in D : \varphi_\Delta(y) \geq \delta\},$$

where the last equality is true because  $\varphi_\Delta$  is positively homogeneous. Since  $\varphi_\Delta$  is continuous, the last set in (10) is closed, and therefore (A2) holds.

From Remark 4.2 we see that the set-valued map  $F_{B_\Delta}(\cdot, z)$  is compact  $\forall z \in X$ . Thus, by Theorem 4.4 we deduce that hypothesis (A3) is satisfied, and the proof is finished.  $\square$

### 5. Relations between several EVPs for vector optimization problems.

We begin this section by giving a set-valued EVP for finite-dimensional vector optimization problems. In order to deal with  $(C, \varepsilon)$ -efficient solutions of this problem, we consider the following coradiant set  $C_\Delta$  introduced in [11, Remark 4.6]:

$$C_\Delta = \{d \in D : \varphi_\Delta(d) \geq 1\},$$

where  $\Delta \in \mathcal{D}$ .

**PROPOSITION 5.1.** *Consider problem (1) with the following data:  $Y = \mathbb{R}^p$ ,  $D$  closed,  $f$  submonotone,  $\gamma, \varepsilon > 0$ , and  $x_0 \in \text{AE}(C_\Delta, \varepsilon)$ . Then there exists a point  $x_\varepsilon \in S$  such that*

- (a)  $f(x_\varepsilon) \leq f(x_0)$ ,
- (b)  $d(x_\varepsilon, x_0) < \varepsilon/\gamma$ ,
- (c)  $\forall d \in D$  such that  $\varphi_\Delta(d) = 1$  it follows that

$$(11) \quad f(x_\varepsilon) \notin f(x) + \gamma d(x, x_\varepsilon)d + D \quad \forall x \in S \setminus \{x_\varepsilon\}.$$

*Proof.* From (9) we see that  $C_\Delta = B_\Delta + D$ . It is clear that  $\text{cone}(C_\Delta) = D$ , and as  $X$  is not a singleton we have that  $D = D_{F_\Delta}$ . So  $D$ ,  $C_\Delta$ , and  $F_\Delta$  are compatible. Moreover,  $B_\Delta$  is bounded since it is compact,  $D = \mathbb{R}_+ B_\Delta$ , and  $C_\Delta = [1, \infty) B_\Delta$  by (10). Then, the result follows by Corollary 4.8, Corollary 3.12, and Theorem 3.14.  $\square$

In particular, if  $D = \mathbb{R}_+^p$  (i.e., (1) is a Pareto problem),  $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p) \in \mathbb{R}_+^p \setminus \{0\}$ ,  $\varepsilon = \sum_{i=1}^p \varepsilon_i$ ,  $\gamma = \sqrt{\varepsilon}$ ,  $\Delta = \{(1, 1, \dots, 1)\}$ , and statement (11) is evaluated at  $d = (1/\varepsilon)(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)$ , then Proposition 5.1 collapses to the first EVP proved in the literature for a vector optimization problem ([19, Proposition 4.2]).

In the literature, the reader can find several EVPs for vector optimization problems proved via different notions of approximate efficiency (see, for instance, [2, 3, 5, 8, 9, 10, 15, 16, 17, 19, 21, 23, 25]). These EVPs are “vector-valued” because they relate approximate solutions of (1) with efficient solutions of vector optimization problems. In other words, they use vector-valued perturbation functions. Specifically, the perturbation function is as follows (see [2, Corollary 2.1], [3, Corollary 4.17, Theorem 4.32], [5, Theorem 3.3], [8, Theorem 8], [9, Proposition 3.10.3], [10, Corollary 2, Corollary 9], [15, Theorem 4], [16, Theorem 11], [17, Theorem 10], [19, Proposition 4.2], [23, Corollary 4.1], and [25, section 4]):

$$f_q(x) := f(x) + \gamma d(x, x_\varepsilon)q \quad \forall x \in X,$$

where  $q \in D \setminus \{0\}$ ,  $\gamma > 0$ , and  $x_\varepsilon$  is a feasible point. Let us observe that some EVPs (see, for instance, [2, Theorem 2.1], [3, Theorem 4.15], [9, Corollary 3.10.19], [10, Corollary 12], and [21, Theorem 6.1]) consider the following more general perturbation function:

$$f_q(x) := f(x) + \gamma r(x, x_\varepsilon) \quad \forall x \in X,$$

where  $r : X \times X \rightarrow D$  is a vector-valued  $D$ -metric,  $\gamma > 0$ , and  $x_\varepsilon$  is a feasible point. However, in practice, this vector-valued  $D$ -metric is defined by  $r(x, y) = d(x, y)q \quad \forall x, y \in X$ , where  $q \in D \setminus \{0\}$  (see, for instance, [2, Lemma 2.1], [3, Lemma 4.16], [9, p. 209], and [10, p. 919]) and so the previous perturbation function is obtained.

In our approach, the perturbation function is set-valued, and this feature is important because in (1), the objective can be improved by using any vector  $d \in D \setminus \{0\}$ .

Next, we give an example which shows that the EVP is stronger if the perturbation function is set-valued.

*Example 5.2.* Let us consider problem (1) with the following data:  $X = Y = \mathbb{R}^2$ ,  $f(x) = x \ \forall x \in \mathbb{R}^2$ ,  $D = \mathbb{R}_+^2$ ,  $S_1 = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 \geq 0, x_2 = 0\}$ ,  $S_2 = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 0, x_2 \geq 0\}$ ,  $S = S_1 \cup S_2$ ,  $\xi = (1, 1)$ , and  $\text{AE}(B_\xi + \mathbb{R}_+^2, 1/2)$  as an approximate efficiency set.

In order to identify  $(B_\xi + \mathbb{R}_+^2, 1/2)$ -efficient solutions via the EVP and  $\gamma = 1/2$ , one can use two approaches. The first is to apply the usual vector-valued version of this principle (see, for instance, Theorem 5.4 following). Indeed, it is obvious that

$$(12) \quad \text{AE}(B_\xi + \mathbb{R}_+^2, 1/2) = \bigcap_{q \in B_\xi} \text{AE}(q + \mathbb{R}_+^2, 1/2).$$

Then, from statements (a)–(c) of Theorem 5.4, the following necessary conditions are derived:

$$(13) \quad \{(0, x_2) : 1/2 < x_2\} \cap \text{AE}((0, 1) + \mathbb{R}_+^2, 1/2) = \emptyset,$$

$$(14) \quad \{(x_1, 0) : 1/2 < x_1\} \cap \text{AE}((1, 0) + \mathbb{R}_+^2, 1/2) = \emptyset,$$

and no information is obtained regarding any vector  $q \in \text{int}(\mathbb{R}_+^2)$  because in this case statements (a)–(c) are fulfilled  $\forall x \in S$ . In summary, from (12)–(14) (and so, by considering an sv- $D$ -metric!) it follows that

$$(15) \quad \{(x_1, x_2) \in S : x_1 + x_2 > 1/2\} \cap \text{AE}(B_\xi + \mathbb{R}_+^2, 1/2) = \emptyset.$$

The reader can observe that if the EVP is applied by considering a single vector  $q \in \mathbb{R}_+^2$ , then the information on the set  $\text{AE}(B_\xi + \mathbb{R}_+^2, 1/2)$  is incomplete (or even trivial, if  $q \in \text{int}(\mathbb{R}_+^2)$ ).

Of course, one can deduce (15) directly by means of Theorem 3.14 and the sv- $D$ -metric

$$F(x, z) := \|x - z\|_{B_\xi} \quad \forall x, z \in \mathbb{R}^2.$$

Another consequence of considering a vector-valued perturbation functional is that various EVPs in the literature, based on different  $\varepsilon$ -efficiency concepts, are particular cases of Theorem 5.4. This result is well known and has been proved under different hypotheses on the objective function and the order cone (see Remark 5.5). Here, by using an approach introduced in [23], we show that the result works for any pointed convex order cone  $D$ .

The following lemma can be proved in a way similar to [9, Theorem 2.3.1].

LEMMA 5.3. *Consider  $q \in D \setminus \{0\}$ , the functional  $\varphi_q : Y \rightarrow \mathbb{R} \cup \{\pm\infty\}$  defined by*

$$(16) \quad \varphi_q(y) := \begin{cases} \inf\{t \in \mathbb{R} : y \in tq - D\} & \text{if } y \in \mathbb{R}q - D, \\ \infty & \text{if } y \notin \mathbb{R}q - D, \end{cases}$$

and the set  $\text{dom}(\varphi_q) = \{y \in Y : |\varphi_q(y)| < \infty\}$ . Then

(a) *for each  $y \in Y$ ,  $y \in \text{dom}(\varphi_q)$  if and only if  $y + \mathbb{R}q \subset \text{dom}(\varphi_q)$  and*

$$(17) \quad \varphi_q(y + \alpha q) = \varphi_q(y) + \alpha \quad \forall y \in Y, \quad \forall \alpha \in \mathbb{R};$$

(b)  *$\{y \in Y : \varphi_q(y) < 0\} = (-\infty, 0)q - D$ ;*

(c)  $\varphi_q$  is  $D$ -monotone, i.e.,

$$\forall y_1, y_2 \in Y, \quad y_1 \in y_2 - D \Rightarrow \varphi_q(y_1) \leq \varphi_q(y_2).$$

Let us consider the following assumption on problem (1),  $q \in D \setminus \{0\}$ , and a feasible point  $x_0 \in S$ :

(B3) The sets  $\{x \in X : f(x) \leq f(x_0) + rq\}$  are closed  $\forall r \in \mathbb{R}$ .

THEOREM 5.4. *Let  $q \in D \setminus \{0\}$ ,  $\gamma > 0$ ,  $\varepsilon > 0$ ,  $C_q := q + D \setminus \{0\}$ , and  $x_0 \in \text{AE}(C_q, \varepsilon)$ . Assume that (B3) is satisfied. Then there exists  $x_\varepsilon \in S$  such that*

- (a)  $f(x_\varepsilon) \leq f(x_0)$  and  $f(x_\varepsilon) \neq f(x_0)$  if  $x_\varepsilon \neq x_0$ ,
- (b)  $d(x_\varepsilon, x_0) \leq \gamma$ ,
- (c)  $f(x) + (\varepsilon/\gamma)d(x, x_\varepsilon)q \notin f(x_\varepsilon) - D \setminus \{0\} \quad \forall x \in S \setminus \{x_\varepsilon\}$ .

*Proof.* Let us define a functional  $g_{q,\varepsilon} : S \subset X \rightarrow \mathbb{R} \cup \{\pm\infty\}$  as follows:

$$g_{q,\varepsilon}(x) := \varphi_q(f(x) - f(x_0) + \varepsilon q) \quad \forall x \in S.$$

From Definition 2.2 is clear that  $(f(S) - f(x_0)) \cap (-\varepsilon C_q) = \emptyset$  and so

$$(f(S) - f(x_0) + \varepsilon q) \cap (-D \setminus \{0\}) = \emptyset,$$

since  $\varepsilon C_q = \varepsilon q + D \setminus \{0\}$ . Then, from Lemma 5.3(b) we see that

$$(18) \quad g_{q,\varepsilon}(x) = \varphi_q(f(x) - f(x_0) + \varepsilon q) \geq 0 \quad \forall x \in S.$$

Moreover, it is easy to check that  $\varphi_q(0) = 0$ . Thus, by (17) it follows that

$$g_{q,\varepsilon}(x_0) = \varphi_q(0 + \varepsilon q) = \varphi_q(0) + \varepsilon = \varepsilon,$$

and from (18) we have that

$$g_{q,\varepsilon}(x_0) - \varepsilon \leq g_{q,\varepsilon}(x) \quad \forall x \in S.$$

Therefore, the functional  $g_{q,\varepsilon}$  is lower bounded and proper. Moreover, by reasoning similar to the proof of [5, Lemma 3.1] we deduce that  $g_{q,\varepsilon}$  is lower semicontinuous. Then, by applying the scalar EVP (see, for example, [6, Theorem 1.1 and statement (1.18)]) we deduce that there exists a point  $x_\varepsilon \in S$  such that

- (i)  $g_{q,\varepsilon}(x_\varepsilon) \leq g_{q,\varepsilon}(x_0) - (\varepsilon/\gamma)d(x_\varepsilon, x_0)$ ,
- (ii)  $d(x_\varepsilon, x_0) \leq \gamma$ ,
- (iii)  $g_{q,\varepsilon}(x_\varepsilon) < g_{q,\varepsilon}(x) + (\varepsilon/\gamma)d(x, x_\varepsilon) \quad \forall x \in S \setminus \{x_\varepsilon\}$ ,

and so part (b) is verified. Part (a) is obvious if  $x_\varepsilon = x_0$ . From (i) and (17) we have that

$$\varphi_q(f(x_\varepsilon) - f(x_0)) + (\varepsilon/\gamma)d(x_\varepsilon, x_0) \leq 0,$$

and if  $x_\varepsilon \neq x_0$ , it follows that

$$\varphi_q(f(x_\varepsilon) - f(x_0)) < 0.$$

Then, by Lemma 5.3(b) we see that  $f(x_\varepsilon) \in f(x_0) - D \setminus \{0\}$  and so part (a) is completed.

In order to prove part (c), let us suppose that there exists  $x \in S \setminus \{x_\varepsilon\}$  such that  $f(x) + (\varepsilon/\gamma)d(x, x_\varepsilon)q \in f(x_\varepsilon) - D \setminus \{0\}$ . Then

$$f(x) - f(x_0) + \varepsilon q + (\varepsilon/\gamma)d(x, x_\varepsilon)q \in f(x_\varepsilon) - f(x_0) + \varepsilon q - D \setminus \{0\},$$

and by the  $D$ -monotonicity of  $\varphi_q$  and property (17) we obtain that

$$\begin{aligned} g_{q,\varepsilon}(x) + (\varepsilon/\gamma)d(x, x_\varepsilon) &= \varphi_q(f(x) - f(x_0) + \varepsilon q + (\varepsilon/\gamma)d(x, x_\varepsilon)q) \\ &\leq \varphi_q(f(x_\varepsilon) - f(x_0) + \varepsilon q) = g_{q,\varepsilon}(x_\varepsilon), \end{aligned}$$

which is contrary to (iii), and part (c) is verified.  $\square$

Property (B3) is called the lower semicontinuity of  $f$  at  $x_0$  along  $q$  (or in the direction  $q$ ) and has been used by some authors in obtaining EVPs for vector-valued maps (see, for example, [9, Corollary 3.10.14], [10, Corollary 9], [23, Corollary 4.1]).

*Remark 5.5.*

- (a) In the literature, Theorem 5.4 has been proved by assuming additional hypotheses such as  $D$  solid (see [3, Theorem 4.32], [9, Proposition 3.10.3], [23, Corollary 4.1]),  $D$  normal (see [15, Theorem 4]),  $D$  closed (see [9, Proposition 3.10.3], [10, Corollary 9], [17, Theorem 10]), or  $f(S)$  lower order bounded (see [3, Theorem 4.32], [8, Theorem 8], [10, Corollary 2], [16, Theorem 11], [17, Theorem 10], [19, Proposition 4.2], [23, Corollary 4.1]).
- (b) The conclusions of Theorem 5.4 could be obtained from Theorem 3.14 via the sv- $D$ -metric  $F(x, y) = d(x, y)q$ , but by considering stronger hypotheses. In particular, let us observe that Theorem 5.4 does not include any boundedness condition. However, Theorem 5.4 is weaker than Theorem 3.14 because it deals with a specific  $\varepsilon$ -efficiency notion and the perturbed map is vector valued.

As was pointed out, one drawback of considering EVPs with vector-valued perturbation functionals for vector optimization problems is that the conclusions are equal whatever the  $(C, \varepsilon)$ -efficiency notion considered in the hypotheses. For example, parts (7)–(9) of [2, Corollary 2.1], parts (vi)–(viii) of [3, Corollary 4.17], parts (a), (c), and (d) of [5, Theorem 3.3], and, by considering the vector-valued metric  $r(u, v) = d(u, v)q$ , part (viii) of [2, Theorem 2.1], part (viii) of [3, Theorem 4.13], and parts (ii) and (iii) of [21, Theorem 6.1] are consequences of our Theorem 5.4. This fact is shown in the following proposition and has been suggested in [10, p. 919]. Previously, two concepts of approximate efficiency due to Németh [21] and Dentcheva and Helbig [5] are recalled.

DEFINITION 5.6.

- (a) (See Németh [21].) Let  $H \subset D \setminus \{0\}$  and  $\varepsilon > 0$ . A point  $x_0 \in S$  is said to be an  $\varepsilon$ -efficient solution of (1) in the sense of Németh (with respect to  $H$ ) if  $(f(x_0) - \varepsilon H - D) \cap f(S) = \emptyset$ .
- (b) (See Dentcheva and Helbig [5].) Let  $q \in D \setminus \{0\}$ ,  $\varepsilon > 0$ , and consider a strictly  $D$ -monotone functional  $h : Y \rightarrow \mathbb{R}$ , i.e., such that  $h(y_1) < h(y_2) \forall y_1, y_2 \in Y, y_1 \in y_2 - D \setminus \{0\}$ . A point  $x_0 \in S$  is said to be an  $\varepsilon$ -efficient solution of (1) in the sense of Dentcheva and Helbig (with respect to  $h$  and  $q$ ) if  $h(f(x_0)) \leq h(f(x) + \varepsilon q) \forall x \in S$  such that  $f(x) \in f(x_0) - D \setminus \{0\}$ .

PROPOSITION 5.7. Consider  $q \in D \setminus \{0\}$ ,  $\gamma > 0$ ,  $\varepsilon > 0$ ,  $x_0 \in S$  and assume that (B3) is satisfied. Suppose that  $x_0$  is an  $\varepsilon$ -efficient solution of (1) in the sense of Dentcheva and Helbig with respect to  $h$  and  $q$  or in the sense of Németh with respect to  $H$ , where  $q \in H$ . Then, parts (a)–(c) of Theorem 5.4 are verified.

*Proof.* First, let us suppose that  $x_0$  is an  $\varepsilon$ -efficient solution in the sense of Németh with respect to  $H$  and  $q \in H$ . Then it is clear that  $x_0 \in \text{AE}(C_q, \varepsilon)$ , and so parts (a)–(c) of Theorem 5.4 hold since the hypotheses of this theorem are satisfied.

If  $x_0$  is an  $\varepsilon$ -efficient solution in the sense of Dentcheva and Helbig with respect to  $h$  and  $q$  and we see that  $x_0 \in \text{AE}(C_q, \varepsilon)$ , then the result follows from Theorem 5.4. Let us suppose that  $x_0 \notin \text{AE}(C_q, \varepsilon)$ . Then there exists  $x \in S$  and  $d \in D \setminus \{0\}$  such that  $f(x) = f(x_0) - \varepsilon q - d$ , and as  $h$  is strictly  $D$ -monotone, it follows that

$$h(f(x_0)) = h(f(x) + \varepsilon q + d) > h(f(x) + \varepsilon q),$$

which is a contradiction since  $f(x) \in f(x_0) - D \setminus \{0\}$  and  $x_0$  is an  $\varepsilon$ -efficient solution in the sense of Dentcheva and Helbig with respect to  $h$  and  $q$ .  $\square$



*Remark 5.8.* Let us observe from Proposition 5.7 that if hypothesis (B3) holds for all  $x_0 \in S$ , then parts (a)–(c) of Theorem 5.4 are verified for each  $(C, \varepsilon)$ -efficient solution if  $\text{AE}(C, \varepsilon) \subset \text{AE}(C_q, \varepsilon)$ . For example, let us consider the  $(C_h, \varepsilon)$ -efficiency notion given by the set

$$C_h = \{d \in D : h(d) > 1\},$$

where  $h \in D^+$  and  $h(q) = 1$  (see [11, Example 3.4] for more details). If  $h \in D^{+s}$ , then it is easy to check that  $\text{AE}(C_h, \varepsilon) \subset \text{AE}(C_q, \varepsilon)$ , and so we deduce that parts (a)–(c) of Theorem 5.4 are true if (B3) is verified and  $x_0 \in \text{AE}(C_h, \varepsilon)$ .

In Theorems 5.11 and 5.15 two set-valued EVPs for  $\varepsilon$ -efficient solutions in the senses of Németh and Dentcheva and Helbig are obtained, which improve the vector-valued versions proved in [21, Theorem 6.1] and [5, Theorem 3.3]. To this end, Theorem 3.14 is applied.

In what follows, a nonempty  $D$ -convex set  $H \subset D \setminus \{0\}$  and the sv- $D$ -metric  $F_H(x, y) := d(x, y)H \forall x, y \in X$  are considered. Let us denote  $C_H := H + D$ .

LEMMA 5.9. *The following statements are satisfied:*

- (a)  $K_{F_H} = (\bigcup_{\alpha > 0} \alpha \text{conv}(H)) \cup \{0\}$ .
- (b)  $D_{F_H} = \text{cone}(C_H)$ .
- (c)  $x_0 \in \text{AE}(C_H, \varepsilon)$  if and only if  $x_0$  is an  $\varepsilon$ -efficient solution of (1) in the sense of Németh with respect to  $H$ .
- (d) If  $0 \notin \text{cl}^w(H)$ , then  $F_H$  satisfies (A2).
- (e)  $F_H$  and  $C_H$  verify (A5).

*Proof.* Part (a). As  $X$  is nontrivial, it follows that

$$(\bigcup_{\alpha > 0} \alpha \text{conv}(H)) \cup \{0\} \subset \text{cone}(\text{conv}(\bigcup_{x \neq y} d(x, y)H)).$$

The reciprocal inclusion is obvious since  $\bigcup_{\alpha \geq 0} \alpha \text{conv}(H)$  is a convex cone, and so part (a) is completed.

Part (b). As  $H$  is  $D$ -convex, we have that

$$H + D \subset \text{conv}(H) + D \subset \text{conv}(H + D) + D = H + D.$$

Therefore,

$$\begin{aligned} D_{F_H} &= ((\bigcup_{\alpha > 0} \alpha \text{conv}(H)) + D) \cup \{0\} \\ &= \bigcup_{\alpha \geq 0} \alpha (\text{conv}(H) + D) = \bigcup_{\alpha \geq 0} \alpha (H + D) = \text{cone}(C_H). \end{aligned}$$

Part (c) is immediate since  $\varepsilon C_H = \varepsilon H + D$ .

Part (d). Fix  $\delta > 0$ . Then  $\bigcup \{d(x, y)H : d(x, y) \geq \delta\} \subset [\delta, \infty)H$ . If  $0 \in \text{cl}^w([\delta, \infty)H)$ , then,  $\alpha_i z_i \xrightarrow{w} 0$  for some nets  $(\alpha_i) \subset [\delta, \infty)$  and  $(z_i) \subset H$ . We can assume that  $\alpha_i \rightarrow \alpha \in [\delta, \infty]$ . It follows that  $z_i = \alpha_i^{-1}(\alpha_i z_i) \xrightarrow{w} 0$ , which is a contradiction because  $0 \notin \text{cl}^w(H)$ .

Part (e) is trivial, and so the proof is finished.  $\square$

*Remark 5.10.* If  $D$  is based, then from Lemma 5.9(b) we deduce that there exists a convex set  $B$  such that  $F(x, y) := d(x, y)B$  is an sv- $D$ -metric and  $D_{F_B} = D$ .

THEOREM 5.11. *Suppose that  $D$  is  $w$ -normal,  $\text{cone}(C_H)$  is based,  $0 \notin \text{cl}^w(H)$ , and assumption (A3) is verified by considering the sv- $D$ -metric  $F_H$ . Let  $x_0$  be an  $\varepsilon$ -efficient solution of (1) in the sense of Németh with respect to  $H$  and assume that the set  $(f(S) - f(x_0)) \cap (-\varepsilon(\text{cone}(C_H) \setminus C_H))$  is  $D$ -bounded. Then there exists  $x_\varepsilon \in S$*

such that

- (a)  $(f(x_\varepsilon) + \gamma d(x_\varepsilon, x_0)H - f(x_0)) \cap (-D) \neq \emptyset$ ,
- (b)  $d(x_\varepsilon, x_0)H \cap (\varepsilon/\gamma)(\text{cone}(C_H) \setminus C_H) \neq \emptyset$ ,
- (c)  $(f(x_\varepsilon) - f(x) - \gamma d(x, x_\varepsilon)H) \cap D = \emptyset \ \forall x \in S \setminus \{x_\varepsilon\}$ .

*Proof.* By Lemma 5.9(b) we see that  $D_{F_H} = \text{cone}(C_H)$ , and from the hypotheses we have that assumptions (A1) and (A3) hold. Moreover, by parts (c), (d), and (e) of Lemma 5.9 we deduce that assumptions (A6), (A2), and (A5) are verified, respectively. Then, the result follows from Theorem 3.14.  $\square$

Let us observe that the assumptions of Theorem 5.11 describe properties about the cone  $D_F \subset D$  and the boundedness and cone lower semicontinuity of problem (1). In the next theorem we show a particular case in which these properties are verified.

**THEOREM 5.12.** *Let us consider that  $D$  is well-based and has a compact base  $B$  such that  $0 \notin \text{aff}(B)$ . Suppose that the objective map  $f$  is submonotone with respect to  $D$ . Consider that  $x_0$  is an  $\varepsilon$ -efficient solution of (1) in the sense of Németh with respect to  $B$ . Then there exists  $x_\varepsilon \in S$  such that*

- (a)  $(f(x_\varepsilon) + \gamma d(x_\varepsilon, x_0)B - f(x_0)) \cap (-D) \neq \emptyset$ ,
- (b)  $d(x_\varepsilon, x_0) < \varepsilon/\gamma$ ,
- (c)  $(f(x_\varepsilon) - f(x) - \gamma d(x, x_\varepsilon)B) \cap D = \emptyset \ \forall x \in S \setminus \{x_\varepsilon\}$ .

*Proof.* Let us define  $H := B$ . Then  $\text{cone}(C_H) = D$ , and so  $\text{cone}(C_H)$  is based and  $0 \notin B = \text{cl}^w(H)$ . Moreover,  $D$  is  $w$ -normal (see [9, Proposition 2.2.15]) and closed, since  $D$  has a compact base. Therefore, by Remark 4.2 and Theorem 4.4 it follows that assumption (A3) is satisfied by considering the  $sv$ - $D$ -metric  $F_H$ .

By Corollary 3.13, Lemma 3.11, and Remark 5.10, we deduce that the set

$$(f(S) - f(x_0)) \cap (-\varepsilon(\text{cone}(C_H) \setminus C_H))$$

is  $D$ -bounded. Then parts (a)–(c) of Theorem 5.11 hold and the result is obtained if we prove that  $d(x_\varepsilon, x_0) < \varepsilon/\gamma$ . Indeed, from the convexity of  $B$  it is clear that  $C_H = B + D = [1, \infty)B$ ,  $D = [0, \infty)B$ , and so  $\text{cone}(C_H) \setminus C_H = [0, 1)B$ . Thus,

$$d(x_\varepsilon, x_0)H \cap (\varepsilon/\gamma)(\text{cone}(C_H) \setminus C_H) \neq \emptyset \iff d(x_\varepsilon, x_0)B \cap [0, \varepsilon/\gamma)B \neq \emptyset$$

and this last statement is satisfied if and only if  $d(x_\varepsilon, x_0) < \varepsilon/\gamma$ , since the condition  $0 \notin \text{aff}(B)$  implies that the representation of a vector of  $D$  as an element of the set  $\mathbb{R}_+ B$  is unique (see [9, Theorem 2.1.15]).  $\square$

To prove a set-valued EVP for approximate solutions of (1) in the sense of Dentcheva and Helbig, the following lemma is necessary. In what follows, we consider a vector  $q \in D \setminus \{0\}$  and a continuous positively homogeneous  $D$ -monotone functional  $h : Y \rightarrow \mathbb{R}$  such that

$$(19) \quad h(y_1 - y_2) \leq 0 \Rightarrow h(y_1) \leq h(y_2) \quad \forall y_1, y_2 \in Y.$$

For instance, any functional  $\xi \in D^+ \setminus \{0\}$  satisfies these properties. In general, if  $\xi_i \in D^+ \setminus \{0\} \ \forall i = 1, 2, \dots, m$ , then the functional

$$g(y) = \max_{1 \leq i \leq m} \{\langle y, \xi_i \rangle\}$$

verifies those properties too. Property (19) is satisfied for any subadditive functional.

Let us define  $C_{q,h} := \{d \in D : h(-q + d) > 0\}$ ,  $[h = 0] := \{d \in D : h(d) = 0\}$  and  $D_h := (D \setminus [h = 0]) \cup \{0\}$ .

**LEMMA 5.13.** *Let  $\varepsilon > 0$ . The following statements are true:*

- (a)  $C_{q,h}$  is a coradiant set and if  $[h = 0] \neq D$ , then  $C_{q,h}$  is nonempty.
- (b)  $\varepsilon C_{q,h} = \{d \in D : h(-\varepsilon q + d) > 0\}$ .
- (c) If  $x_0 \in \text{AE}(C_{q,h}, \varepsilon)$  and  $h$  is strictly  $D$ -monotone, then  $x_0$  is an  $\varepsilon$ -efficient solution of (1) in the sense of Dentcheva and Helbig.
- (d)  $\text{cone}(C_{q,h}) = D_h$  and  $D_h$  is convex.

*Proof.* Parts (a) and (b) follow easily since  $h$  is  $D$ -monotone and positively homogeneous.

Part (c). Let us suppose that  $x_0 \in \text{AE}(C_{q,h}, \varepsilon)$  and consider a point  $x \in S$  such that  $f(x) \leq f(x_0)$ . Then  $f(x_0) - f(x) \notin \varepsilon C_{q,h}$  and by property (19) we have that  $h(f(x_0)) \leq h(f(x) + \varepsilon q)$ ; i.e.,  $x_0$  is an  $\varepsilon$ -efficient solution of (1) in the sense of Dentcheva and Helbig.

Part (d). Let  $d \in \text{cone}(C_{q,h})$ ,  $d \neq 0$ . Then there exists  $\alpha > 0$  such that  $h(-\alpha q + d) > 0$ , and as  $h$  is  $D$ -monotone we have that  $h(d) > 0$ . Therefore,  $\text{cone}(C_{q,h}) \subset (D \setminus [h = 0]) \cup \{0\}$ .

Reciprocally, if  $d \in D \setminus [h = 0]$ , then  $h(d) > 0$  and so, by the continuity of  $h$ ,  $h(d - \alpha q) > 0$  for some  $\alpha > 0$ . Hence  $d \in \text{cone}(C_{q,h})$ . Moreover, it is easy to check that  $C_{q,h} + D \subset C_{q,h}$ , and then  $\alpha_1 d_1 + \alpha_2 d_2 = \alpha_1(d_1 + \alpha_1^{-1} \alpha_2 d_2) \in \text{cone}(C_{q,h})$ , since  $d_1 + \alpha_1^{-1} \alpha_2 d_2 \in C_{q,h}$  for  $d_1, d_2 \in C_{q,h}$ , and  $\alpha_1, \alpha_2 > 0$ . Therefore,  $D_h$  is convex and the proof is finished.  $\square$

*Remark 5.14.* The reader can deduce from the previous proof that  $h \in D^{+s}$  and  $h(q) = 1$  imply that  $x_0 \in \text{AE}(C_{q,h}, \varepsilon)$  if and only

$$h(f(x_0)) \leq h(f(x)) + \varepsilon \quad \forall x \in S, f(x) \leq f(x_0).$$

This notion is due to Helbig (see [13]) and was introduced by considering (not necessarily linear)  $D$ -monotone functionals.

The following result is a direct consequence of Theorem 3.14 and so its proof is omitted.

**THEOREM 5.15.** *Suppose that  $D$  is  $w$ -normal and  $D_h$  is nontrivial and based through a set  $B$ . Define  $F_B(x, y) = d(x, y)B$  and suppose that assumption (A3) is verified by the  $sv$ - $D$ -metric  $F_B$ . Consider that  $x_0 \in \text{AE}(C_{q,h}, \varepsilon)$  and assume that the set  $(f(S) - f(x_0)) \cap (-\varepsilon(D_h \setminus C_{q,h}))$  is  $D$ -bounded. Then there exists  $x_\varepsilon \in S$  such that*

- (a)  $(f(x_\varepsilon) + \gamma d(x_\varepsilon, x_0)B - f(x_0)) \cap (-D) \neq \emptyset$ ,
- (b)  $d(x_\varepsilon, x_0)B \cap (\varepsilon/\gamma)(D_h \setminus C_{q,h}) \neq \emptyset$ ,
- (c)  $(f(x_\varepsilon) - f(x) - \gamma d(x, x_\varepsilon)B) \cap D = \emptyset \quad \forall x \in S \setminus \{x_\varepsilon\}$ .

**COROLLARY 5.16.** *Consider that  $D$  is  $w$ -normal and based. For  $\xi_0 \in D^{+s}$  define  $B = \{d \in D : \langle d, \xi_0 \rangle = 1\}$  and  $F_B(x, y) = d(x, y)B$ . Suppose that assumption (A3) is verified by considering the  $sv$ - $D$ -metric  $F_B$ . Let  $q \in B$  and  $x_0 \in \text{AE}(C_{q,g}, \varepsilon)$ , where  $g = \max\{\xi_0, \xi_1, \xi_2, \dots, \xi_m\}$  and  $\xi_i \in D^+ \quad \forall i = 1, 2, \dots, m$ ,  $m \geq 0$ , and assume that the set  $(f(S) - f(x_0)) \cap (-\varepsilon(D \setminus C_{q,g}))$  is  $D$ -bounded. Then there exists  $x_\varepsilon \in S$  such that*

- (a)  $(f(x_\varepsilon) + \gamma d(x_\varepsilon, x_0)B - f(x_0)) \cap (-D) \neq \emptyset$ ;
- (b)  $d(x_\varepsilon, x_0) \leq \varepsilon/\gamma$ , and if  $m > 0$ , then there exists  $b \in B$  such that

$$d(x_\varepsilon, x_0)\langle b, \xi_i \rangle \leq (\varepsilon/\gamma)\langle q, \xi_i \rangle \quad \forall i = 1, 2, \dots, m;$$

- (c)  $(f(x_\varepsilon) - f(x) - \gamma d(x, x_\varepsilon)B) \cap D = \emptyset \quad \forall x \in S \setminus \{x_\varepsilon\}$ .

*Proof.* It is easy to check that  $[g = 0] = \{0\}$ . Then, by applying Theorem 5.15 to  $h = g$  we deduce that there exists  $x_\varepsilon \in S$ , verifying parts (a) and (c) of this corollary and the relation

$$d(x_\varepsilon, x_0)B \cap (\varepsilon/\gamma)(D \setminus C_{q,g}) \neq \emptyset.$$

Therefore, there is  $b \in B$  such that  $d(x_\varepsilon, x_0)b \notin (\varepsilon/\gamma)C_{q,g}$ , i.e.,

$$g(-(\varepsilon/\gamma)q + d(x_\varepsilon, x_0)b) \leq 0.$$

Thus,

$$\begin{aligned} \langle -(\varepsilon/\gamma)q + d(x_\varepsilon, x_0)b, \xi_0 \rangle &\leq 0, \\ \langle -(\varepsilon/\gamma)q + d(x_\varepsilon, x_0)b, \xi_i \rangle &\leq 0 \quad \forall i = 1, 2, \dots, m, \end{aligned}$$

and the result follows since  $b, q \in B$ .  $\square$

**6. Conclusions.** In this work, an original approach is introduced to extending the well-known Ekeland's variational principle to vector optimization problems. This new approach is based on considering a concept of a set-valued  $D$ -metric, a notion of cone-boundedness, and a set-valued perturbed map. The notions of an sv- $D$ -metric (which is new to our knowledge) and a cone-bounded set by scalarizations are fundamental tools in developing this approach.

In section 5 we have given several results and examples in order to show that this new approach is stronger than the usual vector-valued EVP proved in the literature. In particular let us point out that in the context of vector optimization, where the preferences of the decision-maker are given by a cone, the perturbed map can use various directions, and in this sense the principle is stronger when a set of directions is considered.

The EVP obtained in Theorem 3.8 does not depend on any  $\varepsilon$ -efficient concept. However, the EVP proved in Theorem 3.14 could be applied to several  $\varepsilon$ -efficient notions, since it has been obtained for  $(C, \varepsilon)$ -efficient solutions. In section 5 these results have been applied to different contexts such as the finite-dimensional case or the consideration of various  $\varepsilon$ -efficiency concepts.

**Acknowledgments.** The authors are grateful to the anonymous referees for their helpful comments and suggestions.

## REFERENCES

- [1] T. Q. BAO AND P. Q. KHANH, *Are several recent generalizations of Ekeland's variational principle more general than the original principle?*, Acta Math. Vietnam., 28 (2003), pp. 345–350.
- [2] G. Y. CHEN AND X. X. HUANG, *A unified approach to the existing three types of variational principles for vector valued functions*, Math. Methods Oper. Res., 48 (1998), pp. 349–357.
- [3] G. Y. CHEN, X. X. HUANG, AND X. YANG, *Vector Optimization. Set-Valued and Variational Analysis*, Lecture Notes in Econom. Math. Systems 541, Springer-Verlag, Berlin, 2005.
- [4] A. DANIILIDIS, *Arrow-Barankin-Blackwell theorems and related results in cone duality: A survey*, in Optimization (Namur, 1998), Lecture Notes in Econom. Math. Systems 481, Springer-Verlag, Berlin, 2000, pp. 119–131.
- [5] D. DENTCHEVA AND S. HELBIG, *On variational principles, level sets, well-posedness, and  $\varepsilon$ -solutions in vector optimization*, J. Optim. Theory Appl., 89 (1996), pp. 325–349.
- [6] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.
- [7] I. EKELAND, *Nonconvex minimization problems*, Bull. Amer. Math. Soc., 1 (1979), pp. 443–474.
- [8] C. FINET, *Variational principles in partially ordered Banach spaces*, J. Nonlinear Convex Anal., 2 (2001), pp. 167–174.
- [9] A. GÖPFERT, H. RIAHI, C. TAMMER, AND C. ZĂLINESCU, *Variational Methods in Partially Ordered Spaces*, Springer-Verlag, New York, 2003.
- [10] A. GÖPFERT, C. TAMMER, AND C. ZĂLINESCU, *On the vectorial Ekeland's variational principle and minimal points in product spaces*, Nonlinear Anal., 39 (2000), pp. 909–922.
- [11] C. GUTIÉRREZ, B. JIMÉNEZ, AND V. NOVO, *On approximate efficiency in multiobjective programming*, Math. Methods Oper. Res., 64 (2006), pp. 165–185.

- [12] C. GUTIÉRREZ, B. JIMÉNEZ, AND V. NOVO, *A unified approach and optimality conditions for approximate solutions of vector optimization problems*, SIAM J. Optim., 17 (2006), pp. 688–710.
- [13] S. HELBIG, *On a new concept for  $\varepsilon$ -efficiency*, talk presented at “Optimization Days 1992,” Montreal, 1992.
- [14] J.-B. HIRIART-URRUTY AND C. LEMARECHAL, *Convex Analysis and Minimization Algorithms I*, Springer-Verlag, Berlin, 1996.
- [15] G. ISAC, *The Ekeland's principle and the Pareto  $\varepsilon$ -efficiency*, in Multi-Objective Programming and Goal Programming: Theories and Applications, Lecture Notes in Econom. Math. Systems 432, Springer-Verlag, Berlin, 1996, pp. 148–163.
- [16] G. ISAC, *Nuclear cones in product spaces, Pareto efficiency and Ekeland-type variational principles in locally convex spaces*, Optimization, 53 (2004), pp. 253–268.
- [17] G. ISAC AND C. TAMMER, *Nuclear and full nuclear cones in product spaces: Pareto efficiency and an Ekeland type variational principle*, Positivity, 9 (2005), pp. 511–539.
- [18] J. JAHN, *Vector Optimization. Theory, Applications, and Extensions*, Springer-Verlag, Berlin, 2004.
- [19] P. LORIDAN,  *$\varepsilon$ -solutions in vector minimization problems*, J. Optim. Theory Appl., 43 (1984), pp. 265–276.
- [20] D. T. LUC, *Theory of Vector Optimization*, Lecture Notes in Econom. Math. Systems 319, Springer-Verlag, Berlin, 1989.
- [21] A. B. NÉMETH, *A nonconvex vector minimization problem*, Nonlinear Anal., 10 (1986), pp. 669–678.
- [22] A. PERESSINI, *Ordered Topological Vector Spaces*, Harper & Row Publishers, New York, 1967.
- [23] C. TAMMER, *A generalization of Ekeland's variational principle*, Optimization, 25 (1992), pp. 129–141.
- [24] A. ZAFFARONI, *Superlinear separation for radiant and coradiant sets*, Optimization, 56 (2007), pp. 267–285.
- [25] J. ZHU, C. K. ZHONG, AND Y. J. CHO, *Generalized variational principle and vector optimization*, J. Optim. Theory Appl., 106 (2000), pp. 201–217.

## JUMP DIFFUSION OVER FEATURE SPACE FOR OBJECT RECOGNITION\*

SÉBASTIEN GADAT†

**Abstract.** We present a dynamical model for a population of tests in pattern recognition. Taking a preprocessed initialization of a feature set, we apply a stochastic algorithm based on an efficiency criterion and a Gaussian noise to recursively build and improve the feature space. This algorithm simulates a Markov chain which estimates a probability distribution  $\mathbb{P}$  on the set of features. The features are structured as binary trees and we show that such random forests are a good way to represent the evolution of the feature set. We then obtain properties on the dynamic of the features space before applying this algorithm to practical examples such as face detection and microarray analysis. Lastly, we identify the weak limit of our process as a jump-diffusion process defined using the Skorokhod map over simplices.

**Key words.** Markov processes, jump-diffusion algorithms, stochastic approximation, Skorokhod map, feature selection, pattern recognition

**AMS subject classifications.** 60J75, 60H10, 62L20, 93E35, 62H30

**DOI.** 10.1137/060656759

**1. Introduction.** In this paper, we study a learning algorithm designed for the construction of features in pattern recognition tasks. This algorithm is constructed as the stochastic approximation of a constrained jump-diffusion process, for which we provide an asymptotic analysis.

The algorithm originates from the following issue. A pattern recognition problem corresponds to the classification of *input data* into two or more classes. To solve this, an algorithm, called a *classifier*, is used to design a function which associates a class prediction to an observation of the input variables. There exists several types of competing approaches for building classifiers. Our goal is not to build a new one, but to optimize and improve the prediction by feeding the algorithm with the “best” input variables. Poorly informative variables indeed act like noise in a dataset and reduce the quality of learning algorithms, and fewer variables generally is a guarantee for robustness and reduced generalization ability. Also, a good understanding of the features which have more impact in the classification is critical in some subjects such as biology or text categorization: In microarray analysis, for example, it is important to identify the genes which express a pathology, and in spam detection, one can expect that the presence of some special chain of words enables better detection of nondesirable spam for some classical algorithms such as support vector machines (SVMs), classification trees (CART), or random forests, for instance.

Denote by  $\mathcal{F}_0$  the initial set of variables; in the machine learning community, these are also called *features* and this is the word we will use in this paper. In several recent interesting applications,  $\mathcal{F}_0$  is a large set, which contains hundreds, maybe thousands, of elements. Given that what we want to consider are not only a few useful elements of  $\mathcal{F}_0$ , but also useful combinations of them, we face an overwhelming space of possible explanatory variables that we need to explore in the selection process. Our goal will

---

\*Received by the editors April 8, 2006; accepted for publication (in revised form) November 7, 2007; published electronically February 27, 2008.

<http://www.siam.org/journals/sicon/47-2/65675.html>

†Institut Mathématiques de Toulouse, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse, France (sebastien.gadat@math.ups-tlse.fr).

be to provide a suboptimal stochastic approach to recursively explore and build new composed features space.

For simplicity, the only combinations we consider in this paper are products of variables. If we denote  $\mathcal{P}(\mathcal{F}_0)$  parts of  $\mathcal{F}_0$ , we estimate, from a training set of samples, a subset  $\mathcal{F}$  of  $\mathcal{P}(\mathcal{F}_0)$  of “useful” variables, those which are the most important for the classification task. This set will be estimated as a jump process which will be denoted  $(\mathcal{F}_t)_{t \geq 0}$ .<sup>1</sup>

This jump process will in fact be driven by an auxiliary process, denoted  $\mathbb{P}_t$ , such that, at all times  $t$ ,  $\mathbb{P}_t$  is a probability measure supported by  $\mathcal{F}_t$ . We will define the pair  $(\mathbb{P}_t, \mathcal{F}_t)$  as a jump-diffusion process, designed to maximize the efficiency of the variables belonging to  $\mathcal{F}_t$ . The practical implementation will be a stochastic approximation of this process. The primary goal of this paper is to provide a convergence study of both algorithms, the diffusion, and its approximation.

Since there are important motivations and applications from feature extraction, finding a universal alphabet of features has intrigued researchers in computer vision, and the construction of feature sets has become an active research domain. Direct methods, based on principal (or discriminant) components analysis (PCA) or independent components analysis (ICA) [24], can be used for reduction of dimension, but are not able to create new variables by composition and do not help us to easily understand the selection. Methods based on hierarchically structured variables have also been developed: Amit and Geman [2] and Fleuret and Geman [16] build recursive sets of binary decision trees using coarse to fine procedures. These recursive algorithms combine statistical and geometric properties to assemble discriminative sequential testing and reach very low rates of error in many image classification problems. But in most cases for these algorithms, the amount of features constructed is not limited and can regularly increase if the learning procedure is not stopped [25] and conclusions about optimization results are not inferred. Our approach to the feature space structure will be largely inspired from this sequential testing method, based on statistical correlation [16], entropy [20], or mutual information [15].

Finally, methods based on the optimization of margin of support vector machines have been recently proposed to make recursive feature elimination (RFE [31], [10]). These methods use exact expressions of margin separation of SVM and optimize weights on features to keep only those with high influence on the margin formula. This yielded interesting results on several classification tasks, such as pedestrian detection and cancer morphology classification, with a quantity of features. However, all these methods perform only backward selections from an initial fixed set of features, while adding new features obtained from composition of initial ones could improve efficiency of classification.

Building a set of features derived from an initial set, which contains a reduced number of variables, and complex combinations of variables, is at this point a largely open issue. Our objective will be to handle this problem using simultaneously upward and backward stochastic strategies. Such evolutionary algorithms are commonly used in the framework of regression, adding and removing variables with respect to any information criterion (AIC (Akaike information criterion), MSE (mean square error), etc.). We show here how one can think about similar ideas for the framework of pattern classification without using logistic regression, which may be considered somewhat artificial. Moreover, contrary to most variable selection procedures for linear analysis, we provide a theoretical background for our stochastic exploration of

---

<sup>1</sup>The construction will in fact be slightly more complex, involving trees instead of subsets.

features subsets dedicated to an optimization criterion. Lastly, our method can be used with any classification algorithm. This is an important point since for the commonly investigated classification problems, there does not exist a best classifier among all methods developed by statisticians.

Our paper will be organized as follows. In the next section, we give a precise description of our framework and introduce notation. The third section is devoted to the theoretical model of our jump-diffusion Markov process. Then, section 4 gives exact rules to enable features space to evolve over time. These rules use a Metropolis–Hastings evolution based on an energy  $\mathcal{E}$  to be minimized over time. Section 5 gives dynamic properties of the model previously defined, whereas section 6 provides a statistical implementation and approximation method to simulate the jump-diffusion process of section 3. Finally, we conclude our work with experiments on synthetic data and real classification problems (face detection and leukemia classification) before giving future developments and applications to other situations in pattern recognition. Lastly, note that we choose to formalize our work in a continuous setting (Markov processes) rather than in a discrete form (Markov chains). One motivation will be to provide an understanding of the limit behavior of our exploration/extraction algorithm. Continuous setup will make it easier to precisely describe the dynamic of our constrained optimization method (section 5.1), while the formalism of the martingale problem and generator for Markov processes will be very powerful in identifying the asymptotic measure of our algorithm (Theorems 6.3 and 7.2). In fact, one can also describe our algorithm in a discrete setting (it is, moreover, the way it is numerically implemented) but the identification of the asymptotic behavior requires a time continuous approach with the use of the Skorokhod map. We thus choose to directly present the algorithm in the continuous framework to avoid some additional notation and repetitions.

## 2. Notation and settings.

**2.1. Classes and features.** We address the following pattern recognition problem. Given a large integer  $d$  which will denote the initial number of features, an input signal  $I \in \mathbb{R}^d$  must be classified into a fixed number of classes denoted  $\mathcal{C} = \{C_1, \dots, C_N\}$ . Each input  $I$  is described by its coordinates  $(X^1(I), \dots, X^d(I))$ .  $\mathcal{F}_0$  is the set of initial coordinates maps:

$$\mathcal{F}_0 = \{X^1, \dots, X^d\}.$$

In our experiments, the  $X^j$  will be the projection to the  $j$ th component, it can be binary (values in  $\{0, 1\}$ ) or ternary (values in  $\{-1, 0, 1\}$ ) for the image processing problem of section 8, or more generally, real-valued coordinates can also be considered (see the microarray analysis experiments of section 8).

A classification algorithm is a function which assigns a class  $C_i$  of the finite set  $\mathcal{C}$  to an observed signal  $I$ . This function is estimated on the basis of a training set, which is a finite family of correctly labeled signals. However, for obvious dimensional complexity, the algorithm assumes a specific parametric form for the classification function: it could be, for instance, CART, SVMs, linear discriminant analysis, nearest neighbor, etc. In the two-class problem, the simplest classification rule is based on linear separation: Compute the sum  $\beta_0 + \sum_{j=1}^d \beta_j X^j$ , and decide for the first class if it is negative and for the second otherwise. The parameters  $(\beta_0, \beta_j, j \in \{1 \dots p\})$  are estimated so that this rule is as consistent as possible with the training data. Various definitions of the consistency criterion, variants on the functional form of the decision



rule and of the optimization algorithms, yield a very large family of classifiers, as provided by the literature. We will use in our applications an SVM with a linear kernel because of the generalization ability of this algorithm. Note that the previous linear separation rule assumes that *all* the features are used *as monomials*. Our goal in this context is twofold:

- Selection: Use less than the total family of features, which can be very large ( $d > 1000$ , for instance).
- Composition: Use more complex expressions than monomials by combining the features, and thus define one way to combine them.

This last point implies heuristic or stochastic exploration of the several compositions we can produce starting from  $\mathcal{F}_0$ : simplest ones are  $X^j X^k, (j, k) \in \{1 \dots d\}$ , and  $X^j X^k X^l, (j, k) \in \{1 \dots d\}$ . One can see the exponential growth of the size of possible composition space, and our algorithm proposes a stochastic approach of this exploration step.

*Example 2.1.* Consider the following synthetic example that will be used first in the experiments section. We deal with 3 classes of signals described by 100 ternary features. We thus have  $\mathcal{F}_0 = \{X^1, \dots, X^{100}\}$ . One can imagine that these 3 classes behave differently on several subset of features  $\mathcal{G}_1, \mathcal{G}_2$ , and  $\mathcal{G}_3$  (which may overlap or not) and follow exactly the same distribution on variables in  $\mathcal{F}_0 \setminus \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$ . This is the case for most signal processing situations, where some variables act as independent noise whatever the class of the signal is, although different statistic distributions are located on some other special set of variables for each corresponding class ( $\mathcal{G}_1$  for  $C_1$ ,  $\mathcal{G}_2$  for  $C_2$ , and  $\mathcal{G}_3$  for  $C_3$ ).

We are interested in the problem of detecting interactions of features encoded in all  $\mathcal{G}_i$ , filtering out noisy features in  $\mathcal{F}_0 \setminus \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$ , and forming new compositional variables corresponding to each subset  $\mathcal{G}_i$ . We will provide more details on practical examples in section 8.

**2.2. Composition of features.** We introduce notation regarding the composition of features. Individual features from the original set will be denoted  $\mathcal{F}_0$ , while the set of features obtained at time  $t$  will be naturally named  $\mathcal{F}_t$ . In the definition of the jump diffusion, there will be many advantages in ensuring that the jumps are *reversible*. To obtain such a property, it will be necessary (see section 4.2) for each element of  $\mathcal{F}_t$  to remember how it has been constructed. For this reason, we introduce trees on the set of features as follows.

To an elementary feature  $X^j$  in  $\mathcal{F}_0$  we associate the elementary tree (and keep the same notation “ $X^j$ ”):

$$(1) \quad X^j := \begin{array}{c} X^j \\ \wedge \\ \emptyset \quad \emptyset \end{array} .$$

A tree feature  $\mathcal{A}$  is a binary tree such that each node contains a composition of elementary features, and terminal nodes (leaves) are elementary features of  $\mathcal{F}_0$ . Moreover, each nonterminal node in  $\mathcal{A}$  must be the concatenation (union) of its descendants so that one can easily infer how any tree has been formed. The root of  $\mathcal{A}$ , denoted  $r(\mathcal{A})$ , is the main node associated to the tree. Tree features  $\mathcal{A}, \mathcal{B}$  are aggregated with the construction rule “ $::$ ”

$$(2) \quad \mathcal{A} :: \mathcal{B} = \begin{array}{c} r(\mathcal{A}) \cup r(\mathcal{B}) \\ \wedge \\ \mathcal{A} \quad \mathcal{B} \end{array} .$$

Note that we do not take into account any order or repetition of elementary features taken in  $r(\mathcal{A}) \cup r(\mathcal{B})$ .

*Example 2.2.* For instance, in the operation

$$\begin{array}{c} X^1 X^2 \\ \swarrow \quad \searrow \\ X^1 \quad X^2 \\ \underbrace{\hspace{1.5cm}}_{Al} \end{array} \quad :: \quad \begin{array}{c} X^1 X^3 \\ \swarrow \quad \searrow \\ X^1 \quad X^3 \\ \underbrace{\hspace{1.5cm}}_{Ar} \end{array} = \begin{array}{c} X^1 X^2 X^3 \\ \swarrow \quad \searrow \\ \begin{array}{c} X^1 X^2 \\ \swarrow \quad \searrow \\ X^1 \quad X^2 \end{array} \quad \begin{array}{c} X^1 X^3 \\ \swarrow \quad \searrow \\ X^1 \quad X^3 \end{array} \\ \underbrace{\hspace{3cm}}_{Bl} \end{array} := A,$$

we can reform left and right sons ( $Ar$  and  $Al$ ) from  $A$  by cutting  $A$ 's main node. It is manifest here that without this tree structure of features, the same composition will be

$$\underbrace{X^1 X^2}_{Bl} :: \underbrace{X^1 X^3}_{Bl} = X^1 X^2 X^3 := B$$

but we cannot directly obtain from  $B$  the way it has been formed since some sons could be  $\{(X^1 X^2); (X^1 X^3)\}$  or  $\{(X^2 X^3); (X^2 X^1)\}$ .

To restrict the number of notations, we will keep again the notation  $\mathcal{F}_0$  for the set of elementary trees over the initial set of variables created by operation (1). Similarly,  $\mathcal{F}_t$  will be the set of features handled at time  $t$  by our algorithm. We will denote by  $\mathbf{F}^\sharp$  the set of all trees over  $\mathcal{F}_0$  defined by (1) using (2). Technically,  $\mathcal{F}_t$  will be a jump process on  $\mathbf{F}^\sharp$  (we will call them forests), and  $\mathbb{P}_t$  will be a jump diffusion process with values in the set of probability distributions on  $\mathbf{F}^\sharp$ , which will be supported by a subset of the process  $\mathcal{F}_t$ .

We use the map  $\mathcal{A} \rightarrow r(\mathcal{A})$  only for the computation of trees over input signals since each value of any tree  $\mathcal{A}$  will naturally be defined on any signal  $I$  by

$$r(\mathcal{A})(I) = X^{i_1}(I) \times \cdots \times X^{i_p}(I)$$

if  $r$  is written as  $r(\mathcal{A}) = X^{i_1} \dots X^{i_p}$ .

**2.3. Base classification algorithm  $\mathbb{A}$ .** In this paper, we consider a classification algorithm, denoted  $\mathbb{A}$ , as a “black box” with the following functionalities. We assume that  $\mathbb{A}$  can be conditioned by any subset  $\omega \subset \mathbf{F}^\sharp$  of *active variables*. In training mode,  $\mathbb{A}$  uses a database to build an optimal classifier  $\mathbb{A}_\omega : \mathcal{I} \rightarrow \mathcal{C}$ , such that  $\mathbb{A}_\omega(I)$  depends only on variables  $\omega(I)$ . The test mode simply consists in the instantiation of  $\mathbb{A}_\omega$  on a given signal of the test set.

We work with a randomized version of  $\mathbb{A}$ , in which the randomization is on the set of variables. This randomization of features spaces has been introduced by Amit and Geman [1] and Breiman [8] who build accurate random classifiers with very low dependence to outliers and noise. In the training phase, this works as follows: First, extract a collection  $\{\omega^{(1)}, \dots, \omega^{(N)}\}$  of subsets of  $\mathbf{F}^\sharp$ , and build the classifiers  $\mathbb{A}_{\omega^{(1)}}, \dots, \mathbb{A}_{\omega^{(N)}}$ . Then, derive the classification in the test phase using a majority rule within these  $N$  classifiers. This final algorithm will be denoted  $\bar{\mathbb{A}} = \bar{\mathbb{A}}(\omega^{(1)}, \dots, \omega^{(N)})$ . In test mode, it is run with fixed  $\omega^{(i)}$ 's, which have been obtained in the learning phase.

In addition to being an auxiliary process that we use for variable selection, the probability  $\mathbb{P}_t$  will also be used for sampling the  $\omega^{(k)}$  in the construction of randomized algorithms. Note that the present paper focuses on the way to construct an automatic process creating the random subsets of  $\mathbf{F}^\sharp$  and not designing the classification algorithm  $\mathbb{A}$ , for which we use standard procedures.

We will construct a process  $(\mathcal{F}_t, \mathbb{P}_t)$ , where  $\mathcal{F}_t$  is a jump process over forests (subsets of  $\mathbf{F}^\sharp$ ) and between jumps, and  $\mathbb{P}_t$  is a diffusion process, constrained to the set of probabilities on  $\mathcal{F}_t$ , designed to optimize the performance of the classification algorithm. We start by describing the diffusion process. We will then consider the transition probabilities at jump times, both for  $\mathcal{F}_t$  and  $\mathbb{P}_t$ .

### 3. Constrained diffusion.

*Important notation.* From now on, we will denote with capital letters the Markov process  $(\mathcal{F}_t)$  among the forests;  $(\mathbb{P}_t)$  will denote the Markov process among the probabilities although  $F$  (and  $F_1, F_2, \dots$ ) or  $P$  ( $P_1, P_2, \dots$ ) will be some possible realizations of these two processes. This distinction will be important to the understanding of settings of the next sections.

*Description of the dynamic.* Between jumping times, the probability will essentially evolve according to the diffusion

$$(3) \quad d\mathbb{P}_t = -\nabla \mathcal{E}_{err}(\mathbb{P}_t)dt + \sigma dW_t,$$

where  $\mathcal{E}_{err}(P)$  is a cost function measuring the quality of the classifier using variables sampled from  $P$ ; this will be precisely defined in the following paragraph. Such a process classically stabilizes around probabilities  $P$  with low cost  $\mathcal{E}_{err}$ .

This process must, however, be modified in order to ensure that  $\mathbb{P}_t$  is a probability supported by  $\mathcal{F}_t$ . If  $\mathcal{F}$  is a subset of  $\mathbf{F}^\sharp$ , we denote by  $\mathcal{H}_{\mathcal{F}}$  the hyperplane in  $\mathbb{R}^{\mathcal{F}}$  of equation  $\sum_{\delta \in \mathcal{F}} P(\delta) = 1$ . Let  $\pi_{\mathcal{F}}$  be the affine orthogonal projection onto  $\mathcal{H}_{\mathcal{F}}$  (which is  $\pi_{\mathcal{F}}(U) = U - \sum_{\delta \in \mathcal{F}} U(\delta)/|\mathcal{F}|$ ). We denote  $\nabla^{\mathcal{F}} \mathcal{E}_{err}(P) = \pi_{\mathcal{F}} \nabla \mathcal{E}_{err}(P)$ . We can restrict (3) to  $\mathcal{H}_{\mathcal{F}}$  by replacing  $\nabla$  by  $\nabla^{\mathcal{F}}$  and using a Brownian motion on  $\mathcal{H}_{\mathcal{F}}$ , or equivalently, using

$$(4) \quad d\mathbb{P}_t = -\nabla \mathcal{E}_{err}(\mathbb{P}_t)dt + \Sigma^{\mathcal{F}} dW_t,$$

where  $W$  is a Brownian motion on  $\mathbb{R}^{\mathbf{F}^\sharp}$  and  $\Sigma^{\mathcal{F}} = \sigma \pi_{\mathcal{F}}$ .

Denoting  $\mathcal{S}_{\mathcal{F}}$  for the set of all such probability distributions on  $\mathcal{F}$ , we need to modify (4) to ensure that  $\mathbb{P}_t$  belongs to  $\mathcal{S}_{\mathcal{F}_t}$  at all times. This is done using a constrained diffusion process, which is here a reflected diffusion process:

$$d\mathbb{P}_t = -\nabla \mathcal{E}_{err}(\mathbb{P}_t)dt + \Sigma^{\mathcal{F}_t} dW_t + dZ_t,$$

where  $Z_t$  acts as a correction to ensure that the positivity constraints are satisfied at all times. This means that  $d|Z_t|$  is positive only when  $\mathbb{P}_t$  hits  $\partial \mathcal{S}_{\mathcal{F}}$ .

**3.1. Cost function.** We now define two costs functions for our system forest  $F$  +probability  $P$ . The first function  $\mathcal{E}_{err}(P)$  measures the average performance of the classifier based on random feature selection according to  $P$ . The second measures a structural cost of the set of features  $F$  and does not depend on  $P$ . These two functions enable us to form the global cost for the pairwise process  $(\mathcal{F}_t, \mathbb{P}_t)$ .

**3.1.1. Measuring the mean performance of  $\mathbb{A}$ : The energy  $\mathcal{E}_{err}$ .** Consider a set of trees  $F \subset \mathbf{F}^\sharp$  and a probability distribution  $P$  on  $\mathbf{F}^\sharp$  supported by  $F$ . The algorithm  $\mathbb{A}$  provides a different classifier  $\mathbb{A}_\omega$  for each choice of a subset  $\omega$  of  $k$  features  $\omega = (\omega_1, \dots, \omega_k) \subset F$ . We let  $\eta$  be the classification error,  $\eta(\omega) = \mathbf{P}(\mathbb{A}_\omega(I) \neq \mathcal{C}(I))$ , which will be estimated by

$$g(\omega) = \hat{\mathbf{P}}(\mathbb{A}_\omega(I) \neq \mathcal{C}(I)),$$

where  $\hat{\mathbf{P}}$  is the empirical probability on the training set. As we want to use a small number of features, we fix an integer  $k$ ; the distribution  $P^{\otimes k}$  corresponds to  $k$  independent trials with replacement with respect to the distribution  $P$ . We define the cost function  $\mathcal{E}_{err}$  by

$$\mathcal{E}_{err}(P) = \mathbb{E}_{P^{\otimes k}} g(\omega) = \sum_{\omega \in F^k} g(\omega) P^{\otimes k}(\omega) = \sum_{\omega \in F^k} g(\omega) P(\omega_1) \dots P(\omega_k).$$

One can thus remark that minimizing  $\mathcal{E}_{err}$  according to the control parameter  $P$  will drive us to a distribution with important weights on useful features for the classification (low error rate induced by  $\mathbb{A}$ ).

**3.1.2. Global cost function on  $(F, P)$ : The energy  $\mathcal{E}$ .** We now describe the global cost function, denoted by  $\mathcal{E}$ . It will take the form

$$\mathcal{E}(F, P) = \mathcal{E}_{err}(P) + \mathcal{E}_{struct}(F),$$

where  $\mathcal{E}_{struct}$  is a structural energy on the forest. More precisely,

$$(5) \quad \mathcal{E}_{struct}(F) = \underbrace{\sum_{\mathcal{A} \in F} |\mathcal{A}|}_{\mathcal{E}_s^1(F)} - \underbrace{\sum_{\mathcal{A} \in F} \hat{I}(\mathcal{A}.g, \mathcal{A}.d)}_{\mathcal{E}_s^2(F)}$$

and  $\hat{I}(\mathcal{A}.g, \mathcal{A}.d)$  is the empirical mutual information function between the left and right subtrees of  $\mathcal{A}$ . The first term  $\mathcal{E}_s^1$  limits the size of the forest and comes from the minimum description length principle of information theory [28]. The last term  $\mathcal{E}_s^2$  is of a compositional nature and favors the concatenation of correlated trees (or trees with high mutual information) [16]. Our goal is now to minimize  $\mathcal{E}$  over the space  $\mathbf{F}^\# \times \mathbf{S}_{\mathbf{F}^\#}$ , which has a discrete component and a continuous one.

**4. Jumps.** We first introduce the notion of *weak reversibility* of a jump process since this property will have critical importance in the stochastic dynamic search  $(\mathcal{F}_t, \mathbb{P}_t)$ .

**4.1. General rule.** The time differences between jumps are assumed to be mutually independent, and independent from the rest of the process. Jumps occur as a Poisson process (interjump times are independent and identically distributed (i.i.d.) exponential). Coupled with the constrained diffusion process, this allows the inference algorithm to visit  $\mathbf{F}^\# \times \mathbf{S}_{\mathbf{F}^\#}$ . This accomodates the discrete nature of the problem. At jump times, the transitions  $\mathcal{F}_t \rightarrow \mathcal{F}_{t+dt}$  will correspond to deletion, addition, or combination of elements of  $\mathcal{F}_t$ . Each of these rules will be designed using an accept/reject scheme (Hastings) as follows. We handle here the complete cost function,  $\mathcal{E}(\mathcal{F}_t, \mathbb{P}_t)$  defined by (5). Below, we review some general notions on the Metropolis–Hastings method (this section may be skipped).

**4.1.1. Generality on the Metropolis–Hastings algorithm.** The situation is as follows: Let  $\Omega$  be a measurable set with a measure  $m$  and let  $\mu$  be a measure on  $\Omega$  with density (also denoted  $\mu$ ) w.r.t.  $m$ . The Metropolis–Hastings transitions follow a two-step rule:

- From state  $x \in \Omega$ , first propose a state  $y$  with probability  $Q_0(x, dy)$ ;
- then, accept the transition with a probability which is adjusted so that  $\mu$  is invariant.

We assume the following property: For all  $x \in \Omega$ , there exists a measure  $\rho_x$  such that

- (A1)  $Q_0(x, \cdot)$  has a density  $q(x, \cdot)$  w.r.t.  $\rho_x$ .
- (A2)  $q(x, y) > 0 \Leftrightarrow q(y, x) > 0$ .
- (A3) the measure  $\rho_x(dy) \otimes m(dx)$  is symmetrical: For any function  $f$  on  $\Omega^2$ ,

$$\int_{\Omega} f(x, y) \rho_x(dy) m(dx) = \int_{\Omega} f(y, x) \rho_x(dy) m(dx).$$

The transition  $Q$  is then defined by

$$(6) \quad \begin{aligned} Q(x, dy) = & \min \left( \frac{\mu(y)q(y, x)}{\mu(x)q(x, y)}, 1 \right) Q_0(x, dy) \\ & + \left( 1 - \int_{\Omega} \min \left( \frac{\mu(z)q(z, x)}{\mu(x)q(x, z)}, 1 \right) Q_0(x, dz) \right) \mathbb{1}_x(dy). \end{aligned}$$

The distribution of two consecutive states is then  $Q(x, dy) \otimes m(dx)$ , and to ensure the reversibility we need to verify that it is symmetrical. But

$$\begin{aligned} Q(x, dy) \otimes m(dx) = & \min(\mu(y)q(y, x), \mu(x)q(x, y)) \rho_x(dy) \otimes m(dx) \\ & + \left( 1 - \int_{\Omega} \min(\mu(z)q(z, x), \mu(x)q(x, z)) \rho_x(dz) \right) \mathbb{1}_x(dy) \otimes m(dx). \end{aligned}$$

The second line takes the form  $g(x) \mathbb{1}_x(dy) m(dx)$  and is obviously symmetric, although the first one is symmetric thanks to our assumption on  $\rho_x$ . Consequently, we need to give a transitions rule satisfying the previous assumptions (A1), (A2), and (A3) for our framework on weighted forests.

*Remark 4.1* (necessity of weak reversibility). It is important here to underline why the building process of  $(\mathcal{F}_t)$  must be weakly reversible (assumptions (A1), (A2), and (A3)). We can present at least two reasons for this imperative condition:

- First, note that our exploration process of  $\mathbf{F}^\sharp$  has a stochastic nature and may be mistaken for some iteration because of the Metropolis–Hastings acceptance strategy. We thus need to cancel the decision taken at this step (assumption (A2)), and weak reversibility guarantees this possibility in only one reverse jump.
- Furthermore, the Metropolis–Hastings acceptance rate computation (6) involves the ratio  $q(x, y)/q(y, x)$  because of assumptions (A1), (A3) applied to  $q(x, \cdot)$  and  $q(y, \cdot)$ . Obviously, if the features are not structured as a tree, one cannot compute this ratio since we do not have from any set of variables  $x$  the unique pair of its antecedents.

**4.1.2. Metropolis–Hastings transitions on weighted forests.** We denote by  $m_F$  the Lebesgue measure on  $\mathcal{S}_F$  and consider  $m$  as the global measure on  $\mathcal{P}(\mathbf{F}^\sharp) \times \mathcal{S}_{\mathbf{F}^\sharp}$  defined by

$$m = \sum_{F \subset \mathbf{F}^\sharp} \mathbb{1}_F \otimes m_F,$$

which means that

$$\int f(F, P) dm(F, P) = \sum_{F \subset \mathbf{F}^\sharp} \int_{\mathcal{S}_F} f(F, P) dm_F(P).$$

Here,  $\mathbb{1}_F$  is the Dirac measure at a forest  $F$ . Consider any forest  $F_1$  and an element  $P_1$  of  $\mathcal{S}_{F_1}$ . The transitions are defined as follows: Choose a new forest  $F_2 \in V_{F_1}$ , where  $V_{F_1}$  is the set of forests which are reachable in one jump, and then choose an element of  $\mathcal{S}_{F_2}$  according to a probability which depends on  $F_1$ ,  $F_2$ , and  $P_1$ . Assume that this probability has a positive density w.r.t. some measure denoted  $\psi_{F_1, F_2}(P_1, \cdot)$  on  $\mathcal{S}_{F_2}$ . This implies that the measures, w.r.t. which the densities of the transitions are computed, are

$$\rho_{F_1, P_1}(F_2, \cdot) = \mathbb{1}_{V_{F_1}}(F_2) \psi_{F_1, F_2}(P_1, \cdot),$$

where  $\rho_{F_1, P_1} = \rho_x$  is the measure defined in the former paragraph. Therefore, we need to construct  $\rho$ ,  $\psi$ , and a neighborhood  $V$  in order to satisfy assumptions (A1)–(A3). We design in the next section transitions satisfying (A1) and (A2). Next, we will show that the symmetry requirement is true. Since in the framework of a weighted forest we have

$$m_{F_1}(dP_1) \rho_{F_1, F_2}(P_1, dP_2) = \mathbb{1}_{V_{F_1}}(F_2) \psi_{F_1, F_2}(P_1, \cdot) m_{F_1}(dP_1),$$

it will be sufficient to establish

$$m_{F_1}(dP_1) \psi_{F_1, F_2}(P_1, dP_2) = m_{F_2}(dP_2) \psi_{F_2, F_1}(P_2, dP_1).$$

**4.2. Transitions between forests.** We first construct a set  $\mathcal{T}$  of compositional rules before showing the *weak reversibility* (assumptions (A1), (A2), and (A3)) of our system. This set of transitions does not seem standard and is different from what is done in genetic algorithms. However, to satisfy the *weak reversibility* needed by the Metropolis sampling scheme, this set of transitions  $\mathcal{T}$  will be necessary.

**DEFINITION 4.2** (transition rules  $\mathcal{T}$ ).  *$\mathcal{T}$  is the set of applications from  $\mathcal{P}(\mathbf{F}^\sharp) \times \mathcal{S}_{\mathbf{F}^\sharp}$  to  $\mathcal{P}(\mathbf{F}^\sharp) \times \mathcal{S}_{\mathbf{F}^\sharp}$  formed by buddings, cuttings, suppressions, or rebirths. By  $(F, P) \in \mathcal{P}(\mathbf{F}^\sharp) \times \mathcal{S}_{\mathbf{F}^\sharp}$  we enumerate the states which are reachable in one jump from  $(F, P)$ . For convenience of notation,  $U$  will denote the set of active variables in  $F$  with their associated weights in  $P$ . The quantities  $p_b, p_c, p_s$ , and  $p_r$  will represent the nonnegative probability at each jump time of choosing budding, cutting, suppression, or rebirth. We first enumerate the **budding transitions**:*

Transition	Symbol	Antecedents	Changes in $U$	Probability
Budding without suppression	$\mathcal{B}$	$(\mathcal{A}_1, p_1); (\mathcal{A}_2, p_2)$	Add $(\mathcal{A}_1 :: \mathcal{A}_2, p)$ Change the weights: $(\mathcal{A}_1, p_1 - p + x)$ $(\mathcal{A}_2, p_2 - x)$ where $p \sim \mathcal{U}_{[0; p_1 + p_2]}$ $x \sim \mathcal{U}_{[p - p_1; p_2]}$	$p_b/4$
Budding with left suppression	$\mathcal{B}_l$	$(\mathcal{A}_1, p_1); (\mathcal{A}_2, p_2)$	Add $(\mathcal{A}_1 :: \mathcal{A}_2, p_1)$ Leave $(\mathcal{A}_2, p_2)$ Remove $(\mathcal{A}_1, p_1)$	$p_b/4$
Budding with right suppression	$\mathcal{B}_r$	$(\mathcal{A}_1, p_1); (\mathcal{A}_2, p_2)$	Add $(\mathcal{A}_1 :: \mathcal{A}_2, p_2)$ Leave $(\mathcal{A}_1, p_1)$ Remove $(\mathcal{A}_2, p_2)$	$p_b/4$
Budding with both suppressions	$\mathcal{B}_{lr}$	$(\mathcal{A}_1, p_1); (\mathcal{A}_2, p_2)$	Add $(\mathcal{A}_1 :: \mathcal{A}_2, p_1 + p_2)$ Remove $(\mathcal{A}_1, p_1)$ Remove $(\mathcal{A}_2, p_2)$	$p_b/4$

We present next the **cut transitions**:

Transition	Notation	Antecedents	Changes in $U$	Probability
Cut without creation	$\mathcal{C}$	$(\mathcal{A}_1 :: \mathcal{A}_2, p)$ $(\mathcal{A}_1, p_1); (\mathcal{A}_2, p_2)$	Remove $(\mathcal{A}_1 :: \mathcal{A}_2, p)$ Change the weights: $(\mathcal{A}_1, p_1 + p - x)$ $(\mathcal{A}_2, p_2 + x)$ where $x \sim \mathcal{U}_{[-p_2; p_1 + p]}$	$p_c/4$
Cut with left creation	$\mathcal{C}_l$	$(\mathcal{A}_1 :: \mathcal{A}_2, p)$ $(\mathcal{A}_2, p_2)$	Remove $(\mathcal{A}_1 :: \mathcal{A}_2, p)$ Add $(\mathcal{A}_1, p)$ Leave $(\mathcal{A}_2, p_2)$	$p_c/4$
Cut with right creation	$\mathcal{C}_r$	$(\mathcal{A}_1 :: \mathcal{A}_2, p)$ $(\mathcal{A}_1, p_1)$	Remove $(\mathcal{A}_1 :: \mathcal{A}_2, p)$ Add $(\mathcal{A}_2, p)$ Leave $(\mathcal{A}_1, p_1)$	$p_c/4$
Cut with both creation	$\mathcal{C}_{lr}$	$(\mathcal{A}_1 :: \mathcal{A}_2, p)$	Remove $(\mathcal{A}_1 :: \mathcal{A}_2, p)$ Add $(\mathcal{A}_1, x)$ Add $(\mathcal{A}_2, p - x)$ where $x \sim \mathcal{U}_{[0; p]}$	$p_c/4$

Lastly, we have the **suppression and rebirth transitions**:

Transition	Notation	Antecedents	Changes in $U$	Probability
Suppression	$\mathcal{S}$	$(\mathcal{A}, p)$	Remove $(\mathcal{A}, p)$ Change the weights $\forall (\mathcal{B}, q) \in U \Rightarrow (\mathcal{B}, q/(1-p))$	$p_s$
Rebirth	$\mathcal{S}$	$\mathcal{A} \in \mathcal{F}_0 \setminus F$	Add $(\mathcal{A}, x)$ Change the weights $\forall (\mathcal{B}, q) \Rightarrow (\mathcal{B}, q(1-x))$	$p_r$

With these former transition rules, it is now possible to establish the *weak reversibility conditions*.

PROPOSITION 4.3 (weak reversibility of  $\mathcal{T}$ ). *Assumptions (A1), (A2), and (A3) are true under the dynamic of  $(\mathcal{F}_t, \mathbb{P}_t)$  induced by  $\mathcal{T}$ .*

*Proof.* Take a forest  $F$  in  $\mathcal{P}(\mathcal{F}^\#)$  and define  $V_F$  as the set of reachable forests using one (and only one) transition of  $\mathcal{T}$ . We first remark that if we enumerate all transitions between two forests, we have for any couples of forests  $(F_1, F_2)$ :

$$F_2 \in V_{F_1} \iff F_1 \in V_{F_2}.$$

Roughly speaking, if one tree is created, cut, or deleted using  $\mathcal{T}$ , it is instantaneously possible to flashback and cancel this transition using another rule in  $\mathcal{T}$ . This point is also true if we study weights over forests. For instance, the inverse of budding without suppression is a cut without creation, and if

$$\{(\mathcal{A}_1, p_1); (\mathcal{A}_2, p_2)\} \mapsto \{(\mathcal{A}_1, q_1 = p_1 - p + x); (\mathcal{A}_2, q_2 = p_2 - x); (\mathcal{A}_1 :: \mathcal{A}_2, q_3 = p)\},$$

one can see easily that  $q_1$  takes all values in  $[0; p_1 + p_2]$  and  $q_2$  in  $[0; p_1 + p_2]$  with, in addition,  $q_1 + q_2 + q_3 = p_1 + p_2$ . Consequently, cut without creation from such  $\{(\mathcal{A}_1, q_1); (\mathcal{A}_2, q_2); (\mathcal{A}_1 :: \mathcal{A}_2, q_3)\}$  can reach the initial state. We can verify that this

point is true for all transitions given in the three former arrays while enumerating all possible transitions.

If we then denote by  $\rho_{F,P}$  the uniform measure among reachable sets from  $(F, P)$  using  $\mathcal{T}$ , and by  $q((F, P), \cdot)$  the density of proposition law  $Q_0$  defined in section 4.1.1, we naturally obtain that (A1) and (A2) are true.

We now study assumption (A3). Denote first  $(F_1, P_1)$  as a weighted forest and  $(F_2, P_2)$  as reachable from  $(F_1, P_1)$  using  $\mathcal{T}$ . We must compare  $m_{F_1}(dP_1)\psi_{F_1, F_2}(P_1, dP_2)$  with  $m_{F_2}(dP_2)\psi_{F_2, F_1}(P_2, dP_1)$ . We must then number all transitions of  $\mathcal{T}$  and verify the symmetrical relation. This point is more or less complicated according to the relation considered. For instance, take again the case of budding without creation (remember that  $m_{F_1}$  is the Lebesgue measure defined on the simplex  $S_{F_1}$ ), and denote by  $N$  the length of vector  $P_1 = (p_1, p_2, \dots, p_N)$ . Without loss of generality, we can suppose that we choose to bud trees 1 and 2 so that other weighted trees of  $F_1$  remain unchanged. The length of  $P_2$  is consequently  $N + 1$ , and we have thus  $P_2 = (q_1, q_2, \dots, q_N, q_{N+1})$ .

Hence, to one side we have

$$m_{F_1}(dP_1)\psi_{F_1, F_2}(P_1, dP_2) = \prod_{i=1}^N m_{F_1}(dp_i) \otimes \prod_{i=3}^N \mathbb{1}_{p_i}(q_i) \otimes \mathcal{U}_{X^{p_1, p_2}}(q_1, q_2, q_{N+1}),$$

and  $\psi_{F_1, F_2}(P_1, \cdot)$  is a Dirac for all coordinates in  $P_1$  which are not modified by the bud and

$$X^{p_1, p_2} = \{(a, b, c) \in \mathbb{R}_+^3 \mid a + b + c = p_1 + p_2\}.$$

On the other hand, we can equally write the transition measure

$$m_{F_2}(dP_2)\psi_{F_2, F_1}(P_2, dP_1) = \prod_{i=1}^{N+1} m_{F_2}(dq_i) \otimes \prod_{i=3}^N \mathbb{1}_{q_i}(p_i) \otimes \mathcal{U}_{X^{q_1, q_2, q_{N+1}}}(p_1, p_2).$$

The symmetrical claim is satisfied since, for all measurable functions  $f$  on  $\mathcal{S}_{F_1} \times \mathcal{S}_{F_2}$ ,

$$\begin{aligned} & \langle m_{F_1}(dP_1)\psi_{F_1, F_2}(P_1, dP_2); f \rangle \\ &= \iint_{\mathcal{S}_{F_1} \times \mathcal{S}_{F_2}} f(p_1, \dots, p_N, q_1, \dots, q_{N+1}) m_{F_1}(dP_1)\psi_{F_1, F_2}(P_1, dP_2) \\ &= \int \prod_{i=1}^N dp_i \int_0^{p_1+p_2} dp \int_{p-p_1}^{p_2} dx f(P_1, p_1-p+x, p_2-x, p_3, \dots, p_N, p) \\ &= \int \prod_{i=1}^{N+1} dq_i \int_{-q_2}^{q_1+q_{N+1}} dx \int_{q_1+q_{N+1}-x}^{q_2+x} f(P_2, q_1+p-x, q_2+x, q_3, \dots, q_N) \\ &= \langle m_{F_2}(dP_2)\psi_{F_2, F_1}(P_2, dP_1); f \rangle. \end{aligned}$$

A similar change of variables can be done for all other types of transitions of  $\mathcal{T}$ , and we can conclude that the symmetrical assumption (A3) is also true.  $\square$

**4.3. Decision steps of the Markovian dynamic of jumps.** Taking a jump time  $t_j$  and any state of our process  $(\mathcal{F}_{t_j}, \mathbb{P}_{t_j})$ , we use rules taken from  $\mathcal{T}$  to modify  $\mathcal{F}_{t_j}$  and  $\mathbb{P}_{t_j}$  to  $\mathcal{F}_{t_j+dt}$  and  $\mathbb{P}_{t_j+dt}$ . There are exactly three steps for the choice of which transition of  $\mathcal{T}$  is applied.



- Step 1.* We first choose which kind of transition is proposed in  $\mathcal{T}$  (bud, cut, suppression, or rebirth) according to the probability distribution specified in the last columns of arrays of section 4.2.
- Step 2.* When the rule is chosen, select the trees to which the rule is applied. One can make this decision regardless of whether it is dependent on  $\mathbb{P}_{t_j}$ . The simpler method is to choose uniformly among all trees in  $\mathcal{F}_{t_j}$  or in  $\mathcal{F}_0 \setminus \mathcal{F}_{t_j}$ .
- Step 3.* Accept (or not) the transition according to a differential energy criterion,

$$(7) \quad Q((\mathcal{F}_{t_j}, \mathbb{P}_{t_j}); (F, P)) = \min \left( 1, e^{\mathcal{E}(\mathcal{F}_{t_j}, \mathbb{P}_{t_j}) - \mathcal{E}(F, P)} \times R \right),$$

where

$$R = \frac{Q_0((F, P); (\mathcal{F}_{t_j}, \mathbb{P}_{t_j}))q((F, P); (\mathcal{F}_{t_j}, \mathbb{P}_{t_j}))}{Q_0((\mathcal{F}_{t_j}, \mathbb{P}_{t_j}); (F, P))q((F, P); (\mathcal{F}_{t_j}, \mathbb{P}_{t_j}))}.$$

The computation of the first step is easy with a discrete probability distribution on the rules constituting  $\mathcal{T}$ . At Step 2, the choice of which trees to apply the rule can depend on the distribution  $\mathbb{P}_{t_j}$ . The main idea is to favor trees with high probability for *budding* and low probability for *cuts*. Trees selected for *rebirth* are chosen uniformly in the feature space  $\mathcal{F}_0 \setminus \mathcal{F}_{t_j}$ . More details can be found in [17].

**5. Existence of the jump-diffusion process.** From the beginning of this section, special attention will be dedicated to the indexing of our random processes. They will be described first (up to and including section 6) using a continuous setting  $(\mathcal{F}_t, \mathbb{P}_t)$ , which looks somewhat artificial since in section 7 the algorithm works in a discrete framework with  $(\mathcal{F}_n, \mathbb{P}_n)$ . Moreover, the description of the continuous setup will be much more complicated than the discretized one mainly owing to the projection term in the set of probability distributions.

Actually, the asymptotic behavior of the Markov chain  $(\mathcal{F}_n, \mathbb{P}_n)$  will be presented following a classical scheme of compactness/identification. The compactness is studied in section 7, although the identification of the stationary measure in section 7 will critically use uniqueness of the stationary measure for the continuous process. Thus, the heavy use of the Skorokhod map is highly motivated by this asymptotic study since the identification of the stationary measure is easily deduced from the Markov generator of the process. Since we will need this continuous approach for this last identification, we directly describe the learning process in a continuous setting. Lastly, the weak limit of our discrete Markov chain will be the continuous reflected jump diffusion, and the description of this last process will need the Skorokhod map.

But the Skorokhod map can be skipped to intuitively make the understanding easier in this section, and one can replace the continuous processes by the discretized ones using a simple convex projection to keep  $\mathbb{P}_n$  in a set of probability measures.

**5.1. Existence of the reflected diffusion between jump times.** We work in this section with fixed  $\mathcal{F}_t = F$  of size  $S$  and discuss the existence of a Markovian reflected diffusion process which drives the evolution in the absence of jumps:

$$(8) \quad d\mathbb{P}_t = - \underbrace{\nabla}_{(=\nabla^F)} \mathcal{E}_{err}(\mathbb{P}_t) dt + \sigma dW_t + dZ_t.$$

The construction of solutions of (8) relies on the Skorokhod map  $\Gamma$  associated to  $\mathcal{S}_F$  and a set of unit vectors  $d_c(x)$  for all  $x$  on the boundary  $\partial\mathcal{S}_F$ . This map associates

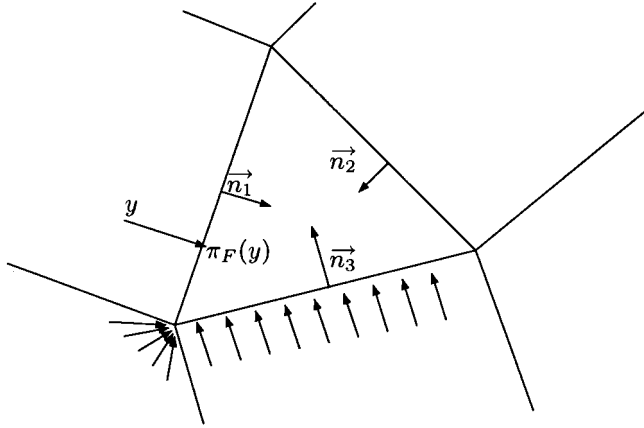


FIG. 1. Directions of reflection vectors for  $x$  in  $\partial\mathcal{S}_F$ .

to any càdlàg trajectory a constrained càdlàg trajectory that satisfies some boundary conditions based on  $d_c(\cdot)$ . We refer to [12] and [13] for further precise technical definitions on this construction. For the sake of completeness, we provide only the constraint vectors we used. The important fact is thus that  $\Gamma$  will exist and define a Lipschitz function on càdlàg trajectories.

DEFINITION 5.1 (directions of constraints  $d_c(\cdot)$ ). We call  $\vec{n}_i$  the unit vectors belonging to the hyperplane supporting  $\mathcal{S}_F$  that normally enter the  $i$ th face of the simplex. The directions of constraints are given by

$$\forall x \in \partial\mathcal{S}_F, \quad d_c(x) = \left\{ \gamma = \sum_{i \mid x_i=0} \alpha_i \vec{n}_i \mid \alpha_i \geq 0, \|\gamma\| = 1 \right\}.$$

These directions of reflection on  $\partial\mathcal{S}_F$  can be expressed easily in a different way as follows.

PROPOSITION 5.2 (directions of constraints  $d_c(\cdot)$ ). For any point  $x$  in  $\partial\mathcal{S}_F$ , vectors  $d_c(x)$  coincide exactly with the sets of unit vectors:

$$d_c(x) = \{ \vec{\gamma} \text{ with } \|\vec{\gamma}\|_2 = 1 \mid \exists y \in \mathcal{H}_F \quad y - \pi_F(y) = \alpha \gamma, \quad \alpha \leq 0, x = \pi_F(y) \},$$

where  $\pi_F$  is the natural convex projection on the simplex  $\mathcal{S}_F$ .

Figure 1 summarizes this natural property. One can remark that directions  $d_c(x)$  are strongly connected to convex projections on  $\mathcal{S}_F$ : they correspond exactly to the unitary vectors that can be used to project any exterior point to  $\mathcal{S}_F$ . In stochastic approximation algorithms, it is the usual way of introducing convex constraints. This yields a set of possible callback vectors shown in Figure 1.

The Skorokhod map allows us to formalize the reflected diffusion (8) as a system of integral equations:

$$\begin{cases} X_t = \mathbb{P}_0 - \int_0^t \nabla \mathcal{E}_{err}(\mathbb{P}_s) ds + \sigma dW(s), \\ \mathbb{P}_t = \Gamma(X)_t. \end{cases}$$

This system is equivalent to the stochastic differential equation (8) [3, 12]. Strong (and obviously weak) existence and uniqueness of such an integral system is standard using a fixed point method [30], Lipschitz regularity of  $\Gamma$ , and Lipschitz continuity of the drift  $\nabla \mathcal{E}_{err}$ . Indeed, for  $\omega \in F^k$  and  $\delta \in F$ , denote by  $C(\omega, \delta)$  the number of occurrences of  $\delta$  in  $\omega$ :

$$C(\omega, \delta) = |\{i \in \{1, \dots, k\} \mid \omega_i = \delta\}|.$$

Since the drift term is polynomial in variables  $P(\delta)$ , it is obviously Lipschitz continuous. Its exact expression is for any  $P \in \mathcal{S}_F$ ; then

$$(9) \quad \forall \delta \in F, \quad \nabla_P \mathcal{E}_{err}(\delta) = \sum_{\omega \in F^k} \frac{C(\omega, \delta) P^{\otimes k}(\omega)}{P(\delta)} g(\omega).$$

We can thus infer the following result.

**THEOREM 5.3** (existence and uniqueness of (8)). *Let  $(\Omega, \mathcal{T}, Q)$  be a probability space with an increasing filtration  $\mathcal{T}_t$ , let  $W_t$  be standard Brownian motion on  $\mathbb{R}^{|\mathcal{F}|}$ , and let  $\mathbb{P}$  be a random variable  $\mathcal{T}_0$ -measurable. Then there exists a unique pair  $(\mathbb{P}_t, Z_t)$   $\mathcal{T}_t$ -measurable satisfying (8) with*

1.

$$\forall T > 0, \quad |Z_T| < +\infty \quad \mathcal{T}_T\text{-a.s.}$$

2.

$$\forall t \geq 0, \quad |Z|_t = \int_0^t \mathbb{1}_{\mathbb{P}_s \in \partial \mathcal{S}_{\mathcal{F}}} d|Z|_s.$$

3.

$$\forall t \geq 0, \quad dZ_t \in d_c(\mathbb{P}_t).$$

*Proof.* See [30, Chapter 5].  $\square$

**5.2. Existence of the complete process.** Since the jump time is a Poisson process independent of the rest, the previous result, combined with the Markov transitions at jump times, trivially implies the existence and uniqueness of the complete jump-diffusion process. An example of the evolution of such a stochastic process is summarized in Figure 2 using a sequence of four different simplices and jumping times. Each simplex corresponds to a features space while the a.s. continuous trajectory points to the evolution of our extraction method  $\mathbb{P}_s$ . We represent here one reflection on  $\mathcal{S}_{\mathcal{F}_{t_{s_2}}}$  and several jumps between several (i.e., 4) simplices. Note that if it is possible to jump from  $\mathcal{S}_{\mathcal{F}_{t_{s_2}}}$  to  $\mathcal{S}_{\mathcal{F}_{t_{s_3}}}$ , it is equally possible to jump from  $\mathcal{S}_{\mathcal{F}_{t_{s_3}}}$  to  $\mathcal{S}_{\mathcal{F}_{t_{s_2}}}$  (weak reversibility).

We will denote by  $\Phi$  the stationary solution of the stochastic differential equation of the reflected jump diffusion based on reflected diffusion on each simplex and jumps between subspaces of features. This solution is defined as follows.

**DEFINITION 5.4** (stationary solution  $\Phi$ ). *Let  $(\Omega, \mathcal{T}, Q)$  be a probability space with an increasing filtration  $\mathcal{T}_t$ . Let  $(W_t)_{t \geq 0}$  be a standard Brownian motion on  $\mathbb{R}^{|\mathcal{F}|}$  and  $(N_t)_{t \geq 0}$  be a Poisson jump process, both adapted to filtration  $\mathcal{T}$ . Suppose likewise that  $W$  and  $N$  are independent. We call  $\Phi = (\mathbb{P}, \mathcal{F})$  the stationary solution of the*

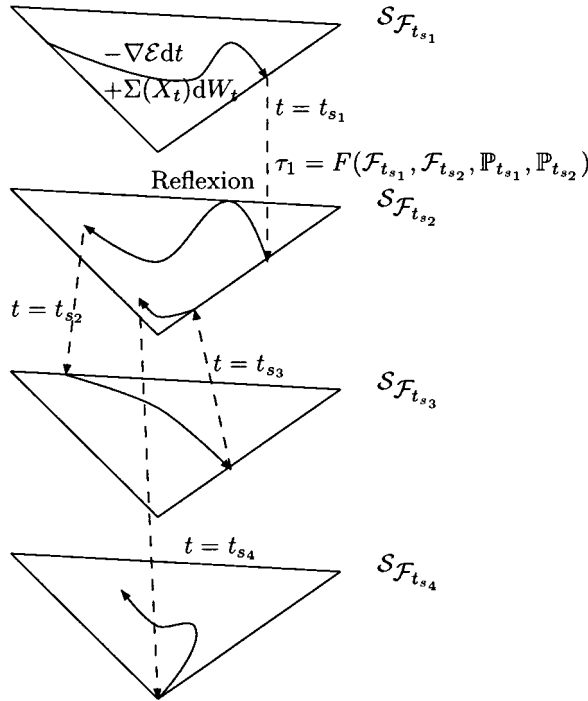


FIG. 2. General form of the stochastic jump-diffusion process.

stochastic differential equation with jumps:

$$d\left(\begin{matrix} \mathbb{P}_t \\ \mathcal{F}_t \end{matrix}\right) = - \begin{pmatrix} \nabla^{\mathcal{F}_t} \mathcal{E}(\mathbb{P}_t) dt + \Sigma^{\mathcal{F}_t} dW_t + dZ_t \\ 0 \end{pmatrix} + \int_{\mathcal{F} \subset \mathbf{F}^\sharp, \mathbb{P} \in \mathcal{S}_{\mathcal{F}}} Q\left[\begin{pmatrix} \mathcal{F}_t \\ \mathbb{P}_t \end{pmatrix}; \begin{pmatrix} \mathcal{F} \\ \mathbb{P} \end{pmatrix}\right] N\left(d\left(\begin{matrix} \mathbb{P} \\ \mathcal{F} \end{matrix}\right); dt\right).$$

**6. Dynamical properties of the algorithm.** In this section, we briefly summarize the dynamical properties of the unique solution of the reflected jump-diffusion process. Our goal is to prove that the process is positive recurrent with a unique stationary measure, given by the the density

$$(10) \quad \mu(F, P) = \frac{e^{-\mathcal{E}(F, P)}}{Z},$$

with respect to the measure on  $\mathcal{P}_{\mathbf{F}^\sharp} \times \mathcal{S}_{\mathbf{F}^\sharp}$ ,

$$m = \sum_{F \subset \mathbf{F}^\sharp} \mathbb{1}_F \otimes m_F.$$

We first give the expression of the infinitesimal generator of the process, then establish that  $(\mathbb{P}_s, X_s)_{s \geq 0}$  is positive recurrent and prove that its stationary measure is the Gibbs field  $\mu$  associated to  $\mathcal{E}$ .

**6.1. Infinitesimal generator of  $(\mathcal{F}_s, \mathbb{P}_s)_{s \geq 0}$ .** Our Markov process is a combination of a reflected diffusion process and a jump process. A generic function on  $\mathcal{P}(\mathbf{F}^\sharp) \times \mathcal{S}_{\mathbf{F}^\sharp}$  can be decomposed as

$$f(F, P) = \sum_{F' \subset \mathbf{F}^\sharp} \mathbb{1}_{F'}(F) f_{F'}(P).$$

The generator  $A$  of this process can be decomposed into a diffusion part and a jump part, yielding  $Af = A^d f + A^j f$ , with

$$A^d f(P, F) = -\langle \nabla_P^F \mathcal{E}_{err} | \nabla_P^F f_F \rangle + \frac{1}{2} \Delta^F f_F(x)$$

and

$$A^j f(P, F) = \int_{\mathcal{P}(\mathbf{F}^\sharp) \times \mathcal{S}_{\mathbf{F}^\sharp}} [f_{F'}(P') - f_F(P)] Q[(F, P), (F', P')] dm(F', P'),$$

where  $Q$  is the transition probability at jump times.

**6.2. Positive recurrence.** The main result of this section uses the positive definite nature of  $\Sigma_F$  on  $\mathcal{H}_F$  and a result of [23, Theorem 1, section 7] ensuring a positive recurrence of the reflected process (without any jump). For any reachable simplex  $\mathcal{S}_F$ , the unique process solution of

$$d\mathbb{P}_t = -\nabla \mathcal{E}_\epsilon(\mathbb{P}_t) dt + \sigma dW_t + dZ_t$$

satisfies the following for all compact sets  $S \subset \mathcal{S}_{\mathcal{F}}$  of nonnegative Lebesgue measure  $\lambda(S) > 0$  (if  $P_p$  is the probability of one event for which initialization of our process is taken at point  $p$ ):

$$(11) \quad \inf_{p \in \mathcal{S}_{\mathcal{F}}} P_p [\tau_S \leq 1] > 0,$$

where

$$\tau_S = \inf \{t / \mathbb{P}_t \in S\}.$$

Equation (11) means that starting at any point  $p$  of simplex  $\mathcal{S}_F$ , one can reach  $S$  in less time than with a probability strictly positive. This implies in particular (see [4, Theorem 2.8]) the positive recurrence of (8) without a jump, and the existence of a unique invariant measure. The extension of the results to the jump-diffusion process now requires only the following fact. Denote

$$p_{F,T}(S) = \inf_{p \in \mathcal{S}_F} P_p [\tau_S \leq T]$$

for  $S \subset \mathcal{S}_F$ , where  $\tau_S$  is the hitting time of  $S$ . We have the following result.

COROLLARY 6.1.

$$p_{F,T}(S) > 0,$$

and the general reflected process with jumps is positive recurrent.

*Proof.* The jumps have been designed so that there exists an integer  $N$  such that for any  $F$  and  $F'$ , and for any  $P \in \mathcal{S}_F$ , the transition  $(F, P) \rightarrow F'$  in  $N$  steps has a probability strictly larger than some positive constant,  $\eta$ . Since the probability of making  $N$  jumps before  $T$ , and no other jump after, is strictly positive, the result is a direct consequence of the positive recurrence of the process without a jump.  $\square$

**6.3. Invariant measure of the process.** Properties of the invariant measure can be inferred from the positive recurrence of  $(\mathbb{P}_t, \mathcal{F}_t)_{t \geq 0}$ . First, note that for any initialization  $(\mathbb{P}_0, \mathcal{F}_0)$  of the process, the family of occupation measures  $(\mu_t)_{t \geq 0}$ , defined by

$$\mu_t(A) = \frac{1}{t} \int_0^t P_{(\mathbb{P}_0, \mathcal{F}_0)}[(\mathbb{P}_s, \mathcal{F}_s) \in A] ds,$$

is tight and any weak limit is an invariant measure of  $(\mathbb{P}_t, X_t)_{t \geq 0}$  since the process is Feller–Markov. Uniqueness is derived from the nondegeneracy of the diffusion of the process into each simplex and the weak reversibility between each simplex of  $(\mathbb{P}_t, \mathcal{F}_t)$ . Identification of this measure from the characterization of [14] is used, for example, in [29].

We use here the well-posedness of the associated martingale problem. Define first the core of this generator as

$$\mathbb{D} = \left\{ f = \sum_{F \subset \mathcal{F}^\#} \mathbb{1}_{\mathcal{S}_F}(P) f_F(P) \mid \forall F \subset \mathcal{F}^\# \quad \forall P \in \partial \mathcal{S}_F \quad \nabla f_F(P) = 0 \right\}.$$

We start noticing that for any function in  $\mathbb{D}$ , the mean effect of generator  $A$  with distribution  $\mu$  given by (10) is null.

PROPOSITION 6.2. *Assume  $f$  is an element of  $\mathbb{D}$ ; then we have*

$$\int A f d\mu = 0.$$

*Proof.* This result is proved by integration by parts, using the Neumann conditions on each simplex  $\mathcal{S}_F$ , where  $F \subset \mathbf{F}^\#$ , the Ostrogradski formula, and the stability equation on the transition acceptance threshold (7). Similar arguments can be found in [29].  $\square$

We are now able to prove the next theorem.

THEOREM 6.3. *The Gibbs field  $\mu$  given by (10) is the unique invariant measure of the global reflected jump-diffusion process, and the martingale problem associated to  $A$  on  $\mathbb{D}$  is well-posed.*

*Proof.* We first apply Echeverria’s theorem (see [14, Theorem 9.17, Chapter 9]) to show that  $\mu$  is stationary. Denote by  $E$  the compact set  $\{(F, P) \mid F \subset \mathbf{F}^\#, P \in \mathcal{S}_F\}$ ; note first that  $\mathbb{D}$  is dense in  $\mathcal{C}(E)$  by the Uryshon lemma applied in each simplex  $\mathcal{S}_F$ . Now,  $A$  satisfies the positive maximum principle on  $\mathbb{D}$  ( $A$  is a classical jump-diffusion generator) and the measure  $\mu$  satisfies

$$\forall f \in \mathbb{D}, \quad \int_E A f d\mu = 0.$$

Consequently  $\mu$  is stationary for  $A$ . Since  $A$  satisfies the maximum principle,  $A$  is dissipative on  $\mathcal{C}(E)$  and  $E$  is separable. Denote then by  $\nu$  a measure on  $E$ ; we can apply the result of [14, Theorem 4.1, Chapter 4] to conclude that uniqueness holds for the martingale problem  $(A, \nu)$  and every solution of the martingale problem is Markov. The martingale problem is well-posed on  $\mathcal{C}(E)$ , every solution of the martingale problem is a weak solution of the stochastic differential equation of jump diffusion, and  $\mu$  is the unique stationary distribution of  $(\mathcal{F}_t, \mathbb{P}_t)$ .  $\square$

**7. Stochastic approximations.** We now address the computational part of the algorithm, which is not trivial because the drift term involves a sum over an untractable number of terms. Fortunately, this sum can be interpreted as an expectation, which allows us to replace it by a stochastic approximation of Robbins–Monro type. Before passing to the drift term, we first address the time discretization issues.

**7.1. Time discretization.** To solve (8), we use a time discretization scheme with a discretization step  $\alpha$ ,

$$\forall n \in \mathbb{N}, \quad \mathbb{P}_{n+1} = \mathbb{P}_n - \alpha \nabla^{\mathcal{F}} \mathcal{E}_{err}(\mathbb{P}_n) + \sqrt{\alpha} \sqrt{\sigma} d\xi_n + dz_n,$$

where  $d\xi_n$  is a centered normal  $|\mathcal{F}|$  dimensional vector and  $dz_n$  is the smaller vector that is added to make  $\mathbb{P}_{n+1} \in \mathcal{S}_{\mathcal{F}}$ . In other words,

$$\forall n \in \mathbb{N}, \quad \mathbb{P}_{n+1} = \pi_{\mathcal{F}} \left( \mathbb{P}_n - \alpha \nabla^{\mathcal{F}} \mathcal{E}_{err}(\mathbb{P}_n) + \sqrt{\alpha} \sqrt{\sigma} d\xi_n \right).$$

However, the computational issue comes from the gradient of  $\mathcal{E}_{err}$ , given in (9), which requires a sum over all  $\omega$  in  $\mathcal{F}^p$ . This is an untractable sum, since  $|\mathcal{F}|$  is typically thousands and  $p$  hundreds. However, it can be replaced by the stochastic approximation defined in the next section.

**7.2. Stochastic differential equation method for approximation.** Stochastic approximation can be seen as noisy discretizations of stochastic differential equations ([26]). They are generally expressed under the form

$$(12) \quad X_{n+1} = X_n + \alpha_n F(X_n, \zeta_{n+1}) + \sqrt{\alpha_n} \sqrt{\sigma} \xi_n + \alpha_n z_n + \alpha_n^2 T_n,$$

where  $X_n$  is the current state of the process,  $\zeta_{n+1}$  a random perturbation,  $\xi_n$  a random perturbation of known distribution,  $z_n$  a random variable designed to ensure the constraints, and  $T_n$  a secondary error term. If the distribution of  $\zeta_{n+1}$  depends only on the current value of  $X_n$ , then one defines an average drift  $X \mapsto G(X)$  by

$$G(X) = \mathbb{E}[F(X, \zeta)|X],$$

and (12) can be shown to evolve similarly to the stochastic differential equation:  $dX_t = G(X)dt + \sqrt{\sigma}dw_t + dz_t$ , in the sense that the trajectories coincide when  $(\epsilon_n)_{n \in \mathbb{N}}$  goes to 0 (a more precise statement is given below).

To implement our reflected diffusion equations (8) in this framework, we need to design a random variable  $d_n$  (identified as  $F(X_n, \zeta_n)$  in (12)) such that

$$(13) \quad \mathbb{E}[d_n] = -\nabla^{\mathcal{F}} \mathcal{E}_{err}(\mathbb{P}_n) = -\Pi_{\overrightarrow{\mathcal{H}_{\mathcal{F}}}}[\nabla \mathcal{E}_{err}(\mathbb{P}_n)],$$

where  $\Pi_{\overrightarrow{\mathcal{H}_{\mathcal{F}}}}$  is the vectorial projection on the hyperplane supporting  $\mathcal{S}_{\mathcal{F}}$ . We will then define

$$\mathbb{P}_{n+1} = \mathbb{P}_n - \alpha_n d_n + \sqrt{\alpha_n} \sqrt{\sigma} \xi_n + dz_n = \pi_{\mathcal{F}} \left( \mathbb{P}_n - \alpha_n d_n + \sqrt{\alpha_n} \sqrt{\sigma} \xi_n \right).$$

From (9), we obtain

$$\nabla \mathcal{E}_{err}(\mathbb{P})(\delta) = \mathbb{E}_{\mathbb{P}^{\otimes k}} \left[ \frac{C(\omega, \delta)g(\omega)}{\mathbb{P}(\delta)} \right].$$

Using the linearity of the projection  $\Pi_{\overrightarrow{\mathcal{H}_{\mathcal{F}}}}$ , we get

$$\Pi_{\overrightarrow{\mathcal{H}_{\mathcal{F}}}}(\nabla \mathcal{E}(\mathbb{P}))(\delta) = \mathbb{E}_{\mathbb{P}^{\otimes k}} \left[ \Pi_{\overrightarrow{\mathcal{H}_{\mathcal{F}}}} \left( \frac{C(\omega, \cdot)g(\omega)}{\mathbb{P}(\cdot)} \right) (\delta) \right].$$

Consequently, following (13), it is now natural to define the approximation term of the reflected diffusion (8) by

$$d_n(\delta) = \Pi_{\mathcal{H}_{\mathcal{F}}} \left( \frac{C(\omega_n, \cdot)}{\mathbb{P}_n(\cdot)} \right) (\delta),$$

where the set of  $k$  features  $\omega_n$  is a random variable extracted from  $\mathcal{F}$  with law  $\mathbb{P}_n^{\otimes k}$ .

This results in the following numerical simulation scheme:

1. Step 0: Initialization: Set  $\mathbb{P}_0 = \mathcal{U}_{\mathcal{F}}$ .
2. Step  $n$ : Draw a sample  $\omega_n$  in  $\mathcal{F}^k$  with respect to  $\mathbb{P}_n^{\otimes k}$ .
3. Step  $n$ : Compute  $g(\omega_n)$ .
4. Step  $n$ : Update  $\mathbb{P}_{n+1}$  with

$$(14) \quad \begin{aligned} \mathbb{P}_{n+1} &= \pi_{\mathcal{F}} \left( \mathbb{P}_n - \alpha_n \left[ \frac{C(\omega_n, \cdot)}{\mathbb{P}_n} \right] + \sqrt{\alpha_n} \sqrt{\sigma} d\xi_n \right) \\ &= \mathbb{P}_n - \alpha_n \left[ \frac{C(\omega_n, \cdot)}{\mathbb{P}_n} \right] + \sqrt{\alpha_n} \sqrt{\sigma} d\xi_n + dz_n, \end{aligned}$$

where  $-\alpha_n C(\omega_n, \cdot)/\mathbb{P}_n$  is the approximated value of  $-\nabla \mathcal{E}_{err}(\mathbb{P}_n)$  and  $d\xi_n$  is a centered normal  $|\mathcal{F}|$  dimensional vector.

To simulate the stochastic approximation of the jump-diffusion algorithm, (14) must be combined with transitions of  $(\mathcal{F}, \mathcal{P})$  at jump times. This results in the following new complete scheme:

1. Step 0: Initialization: Set  $\mathbb{P}_0 = \mathcal{U}_{\mathcal{F}_0}$ . Sample the first jumping time  $t^1$  with an exponential distribution, set  $t = 0$ , and set  $n = 0$ .
2. Step  $j$  ( $j \geq 1$ ): While  $t < t^j$ , run the previous discretization scheme (for the reflected diffusion),  $t$  being iteratively computed by  $t = \alpha_0 + \dots + \alpha_n$ .
3. When  $t > t_j$ : Update  $\mathcal{F}_t$  and  $\mathbb{P}_t$  according to the Markov transition rules.
4. Compute the next jump time with  $t^{j+1}$  by adding an exponential variable to  $t^j$  and return to 2.

**7.3. Weak convergence of the numerical scheme.** In the following paragraphs, we will define  $(\mathbb{P}^n(t)_{t \geq 0})_{n \in \mathbb{N}}$  as a sequence of continuous processes that interpolates the behavior of the discrete sequence of  $(\mathbb{P}_n)_{n \in \mathbb{N}}$ .

**7.3.1. Interpolated approximations.** Following classic notation of [26], we set up the time parameter  $\tau_n$  as

$$\tau_n = \sum_{i \leq n} \alpha_i,$$

and set up the map  $m$  permitting the association of continuous time and discrete iteration as

$$m(t) = \sup_{\tau_n \leq t} \{n \in \mathbb{N}\}.$$

Given that the  $j$ th jump occurs at time  $\nu_j$ , we construct its values according to the distribution  $Q((\mathcal{F}_{\nu_j-}, \mathbb{P}_{\nu_j-}), \cdot)$  to obtain  $(\mathcal{F}_{\nu_j}, \mathbb{P}_{\nu_j})$ . It is thus possible to define the discrete jump term in the discrete case as

$$q_j = \mathbb{P}_{m(\nu_j)+1} - \mathbb{P}_{m(\nu_j)},$$



which corresponds to the term we add to compute the jump from  $\nu_j-$  to  $\nu_j$ .

We now define the sequence of right continuous interpolation processes  $(\mathbb{P}^n(t))_{t \geq 0}$  initialized at  $\mathbb{P}_n$ .

**DEFINITION 7.1** (processes  $(\mathbb{P}^n(\cdot), Y^n(\cdot), W^n(\cdot), Z^n(\cdot))_{n \in \mathbb{N}}$ ). *We define the processes  $(\mathbb{P}^n, Y^n, W^n, Z^n)$  valued in  $\mathbb{R}^{\mathbf{F}^\sharp}$  by*

$$\forall n \in \mathbb{N}, \quad \forall t \in \mathbb{R}_+, \quad Y^n(t) = \sum_{i=n}^{m(\tau_n+t)} \alpha_i y_i,$$

where the term  $y_i$  satisfies

$$\forall \delta \in \mathcal{F}, \quad y_i(\delta) = -\frac{C(\omega_i, \delta)g(\omega_i)}{\mathbb{P}_i(\delta)} \text{ if } \mathbb{P}_i(\delta) \neq 0 \text{ and } y_i(\delta) = 0 \text{ if } \mathbb{P}_i(\delta) = 0.$$

Likewise, we define

$$W^n(t) = \sum_{i=n}^{m(\tau_n+t)} \sqrt{\alpha_i} d\xi_i,$$

where  $d\xi_i$  is considered as an element of  $\mathbb{R}^{\mathbf{F}^\sharp}$ ,

$$Z^n(t) = \sum_{i=n}^{m(\tau_n+t)} dz_i.$$

Finally,

$$\mathbb{P}^n(t) = \mathbb{P}_n + Y^n(t) + W^n(t) + Z^n(t) + \sum_{\tau_n \leq \nu_j \leq \tau_n+t} q_j.$$

With these definitions, it is obvious that  $\mathbb{P}^n$  is a process on  $\mathcal{S}_{\mathbf{F}^\sharp}$ , which is right continuous with left limits (in the space  $\mathcal{D}$  of càdlàg trajectories). To get theoretical convergence results on these sequence of processes, we will now classically choose  $(\alpha_n)$  such that  $\sum \alpha_i = \infty$  and  $\sum \alpha_i^2 < \infty$  (see [5], [26], for instance).

**7.3.2. Convergence of  $(\mathbb{P}^n, Y^n, W^n, Z^n)_{n \in \mathbb{N}}$ .** We will show that the family of processes  $(\mathbb{P}^n(\cdot), Y^n(\cdot), W^n(\cdot), Z^n(\cdot))_{n \in \mathbb{N}}$  is weakly compact in the space  $\mathcal{D}$ . The associated topology on this space is derived from the Skorokhod distance [6], [26] and we consider weak convergence of trajectories of  $\mathcal{D}([0; \infty[)$ .

**THEOREM 7.2.** *The processes  $(\mathbb{P}^n, Z^n)_{n \in \mathbb{N}}$ , which are stepwise constant, weakly converge toward the unique invariant solution of the stochastic differential equation without jumps and  $(\mathbb{P}_n, \mathcal{F}_n)_{n \in \mathbb{N}}$  converges toward the stationary measure  $\mu$ .*

The proof of Theorem 7.2 includes three steps: First, prove the tightness of the family  $(\mathbb{P}^n, Y^n, W^n, Z^n)_{n \in \mathbb{N}}$ , then identify the unique possible weak limit, and finally show the convergence toward the stationary measure  $\mu$ .

**7.3.3. Tightness.** To show tightness, we use the following criterion.

**THEOREM 7.3** (see [26], [6]). *Let  $X^n$  be a sequence in  $\mathcal{D}$ ;  $(X^n)_{n \in \mathbb{N}}$  is tight iff*

1. *for any time  $T$  and  $\epsilon > 0$ , there exist an integer  $n_0$  and a real  $K$  satisfying*

$$(15) \quad \forall n \geq n_0, \quad P \left[ \sup_{t \leq T} |X^n(t)| \geq K \right] \leq \epsilon.$$

2.

$$(16) \quad \forall \epsilon > 0, \quad \lim_{\delta \rightarrow 0} \limsup_n P[w'_{X^n}(\delta) \geq \epsilon] = 0.$$

We establish successively (15) and (16) for our family of processes  $(\mathbb{P}^n, Y^n, W^n, Z^n)_{n \in \mathbb{N}}$ . The next proposition shows that (15) is true for  $(\mathbb{P}^n, Y^n, W^n, Z^n)_{n \in \mathbb{N}}$  and consequently guarantees the tightness of the family  $(\mathbb{P}^n, Y^n, W^n, Z^n)_{n \in \mathbb{N}}$ .

PROPOSITION 7.4. *The sequence of processes  $(\mathbb{P}^n, Y^n, W^n, Z^n)_{n \in \mathbb{N}}$  satisfies (15).*

*Proof.* The result is obvious for  $\mathbb{P}^n$ , since it is compactly supported. To get the result for  $(Y^n)_{n \in \mathbb{N}}$ , we define the sequence  $\tilde{y}_n$  as

$$\tilde{y}_n = y_n - \underbrace{\mathbb{E}_{\mathbb{P}^n}[y_n]}_{=h_n}.$$

Fix any real time  $T$  and a real number  $\epsilon > 0$ . We define the sequence of processes

$$\tilde{Y}^n(t) = \sum_{i=n}^{m(\tau_n+t)} \alpha_i \tilde{y}_i.$$

Since  $\mathbb{E}_{\mathbb{P}^n}[y_n]$  is bounded by  $M$ , the first tightness criterion is true for processes  $H^n$ :

$$H^n(t) = \sum_{i=n}^{m(\tau_n+t)} \alpha_i h_i,$$

and we study the sequence of  $(\tilde{Y}^n)_{n \in \mathbb{N}}$ . Now the sum  $M_p^n$  given by

$$M_p^n = \sum_{i=n}^{n+p} \alpha_i \tilde{y}_i$$

is a martingale for the filtration generated by  $\mathbb{F}_p^n = \sigma(\mathbb{P}_i, \xi_i, w_{i-1}, i \leq n+p)$ . We can use Doob's inequality to show that

$$P\left(\sup_{q \leq p} |M_q^n| > K\right) \leq \frac{1}{K} \mathbb{E}(|M_p^n|).$$

Now,

$$\mathbb{E}(|M_p^n|) \leq \sum_{i=n}^p \alpha_i \mathbb{E}[|\tilde{y}_i|] \leq \sup_i \mathbb{E}[|\tilde{y}_i|] \sum_{i=n}^p \alpha_i.$$

Finally,  $\mathbb{E}[|\tilde{y}_i|] = \mathbb{E}(\mathbb{E}[|\tilde{y}_i| | \mathbb{F}_i^n])$ , and  $\mathbb{E}[|\tilde{y}_i| | \mathbb{F}_i^n]$  is bounded by  $2M$ . We have  $\sum_{i=n}^p \alpha_i \leq T$  and we can deduce from these upper-bounds that

$$\lim_{K \rightarrow \infty} P\left(\sup_{q \leq p} |M_q^n| > K\right) = 0.$$

The fact that

$$\lim_{K \rightarrow \infty} \sup_{n \in \mathbb{N}} P\left[\sup_{t \leq T} |W^n(t)| \geq K\right] = 0$$

is standard and can be found in [26], [17]. Finally, since  $Z^n = \mathbb{P}^n - Y^n - W^n$ ,  $(Z^n)_{n \in \mathbb{N}}$  obviously satisfies (15).  $\square$

We must now establish condition (16) to achieve tightness of  $(\mathbb{P}^n, Y^n, W^n, Z^n)_{n \in \mathbb{N}}$ .

**PROPOSITION 7.5** (condition (16)). *Each of the processes  $(\mathbb{P}^n, Y^n, W^n, Z^n)_{n \in \mathbb{N}}$  satisfies (16).*

*Proof.* We first establish (16) for  $(Y^n)_{n \in \mathbb{N}}$ . Note that

$$\mathbb{E}[Y^n(t+s) - Y^n(t)] = \sum_{i=m(\tau_n+t)}^{m(\tau_n+t+s)} \alpha_i \mathbb{E}[\mathbb{E}[y_i | \mathbb{F}_0^i]].$$

Then, use the fact that the expectations of  $y_k$  are bounded by  $M$  to obtain

$$\mathbb{E}[|Y^n(t+s) - Y^n(t)|] \leq Ms.$$

We thus conclude that (16) is true for  $(Y^n)_{n \in \mathbb{N}}$  using the Markov inequality. The argument is standard to get a similar result for  $(W^n)_{n \in \mathbb{N}}$  by Doob's inequality (see [26]). The jump component involved by terms  $q_j$  defines also a sequence of processes:

$$J^n(t) = \sum_{\tau_n \leq \nu_j \leq \tau_n+t} q_j.$$

Inequality (16) for  $(J^n)_{n \in \mathbb{N}}$  is here clearly satisfied since jumps occur exponentially as each term  $q_j$  is bounded. Consequently, we have

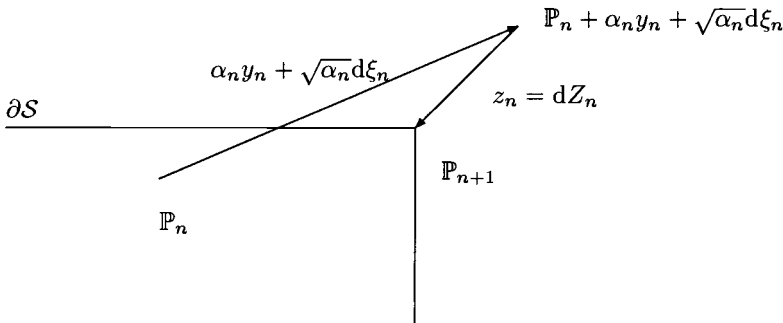
$$\limsup_n P[w'_{J^n}(\delta) \geq \epsilon] = o(\delta).$$

To deal with the processes  $(Z^n)_{n \in \mathbb{N}}$ , it is important to note that

$$|Z^n(t+s) - Z^n(t)| \leq C \sum_{i=m(\tau_n+t)}^{m(\tau_n+t+s)} |\alpha_i y_i + \sqrt{\alpha_i} d\xi_i|,$$

since

$$(17) \quad |z_i| \leq C|\alpha_i y_i + \sqrt{\alpha_i} d\xi_i|.$$



Using inequality (16) for  $(Y^n)_{n \in \mathbb{N}}$  and  $(W^n)_{n \in \mathbb{N}}$ , and (17), we obtain the second tightness inequality needed on the processes  $(Z^n)_{n \in \mathbb{N}}$ . The conclusion is immediate for  $(\mathbb{P}^n)_{n \in \mathbb{N}}$ .  $\square$

**7.3.4. Proof of Theorem 7.2.** We end the proof of Theorem 7.2 using compactness of the trajectories. Note first that if  $(\mathbb{P}^n, Y^n, W^n, Z^n)$  is weakly convergent toward  $(\mathbb{P}, Y, W, Z)$ , then  $(\mathbb{P}, Y, W, Z)$  is a solution of the reflected jump diffusion  $\Phi$  initialized to the weak limit of  $(\mathbb{P}^n(0), Y^n(0), W^n(0), Z^n(0))$  using the same argument of [26, Theorem 2.3].

While replacing  $(\mathbb{P}^n, \mathcal{F}^n)$  by  $\mathbb{P}^n$  and taking any sequence extracted from  $(\mathbb{P}^n)_{n \in \mathbb{N}}$ , we note  $(N_k)_{k \in \mathbb{N}}$  this extraction procedure and show that  $(\mathbb{P}^{N_k})_{k \in \mathbb{N}}$  is weakly convergent to the unique invariant measure  $\mu$ . Denote by  $\nu_\infty$  the weak limit of  $(\mathbb{P}^{N_k}(0))_{k \in \mathbb{N}}$ ; it is then sufficient to show that for any measurable function  $\phi$ ,

$$\mathbb{E}_{\nu_\infty} \phi = \mathbb{E}_\mu \phi.$$

Denote by  $P_\nu^t$  the law of our process at time  $t$  initialized by measure  $\nu$ , since  $\mu$  is the unique stationary measure we have for any compact set of measures  $K$ :

$$(18) \quad \forall \nu \in K \quad \forall \epsilon > 0, \quad \exists T > 0 \quad \forall t \geq T, \quad \left| \int \phi(y) dP_\nu^t - \int \phi(y) d\mu(y) \right| \leq \epsilon.$$

Taking  $\epsilon$  strictly positive and applying (18) to the family of measures  $K$  formed by the law of  $(\mathbb{P}^{N_k})_{k \in \mathbb{N}}$ , which is tight and thus compact, we find  $T$  such that

$$\forall t \geq T, \quad \left| \int \phi(y) dP_\nu^t - \int \phi(y) d\mu(y) \right| \leq \epsilon.$$

Now, if  $\nu'_\infty$  is the weak limit of the sequence of processes  $(\mathbb{P}^{N_k}(\cdot - T))_{k \in \mathbb{N}}$ , which is also the weak limit of  $(\mathbb{P}(\tau_{N_k} - T))_{k \in \mathbb{N}}$ , we have

$$\begin{aligned} \left| \int \phi(y) d\nu_\infty(y) - \int \phi(y) d\mu(y) \right| &\leq \left| \int \phi(y) d\nu_\infty(y) - \mathbb{E}[\phi(\mathbb{P}(\tau_{N_k}))] \right| \\ &\quad + \left| \mathbb{E}[\phi(\mathbb{P}(\tau_{N_k}))] - \int \phi(y) dP_{\nu'_\infty}^T(y) \right| \\ &\quad + \left| \int \phi(y) dP_{\nu'_\infty}^T(y) - \int \phi(y) d\mu(y) \right|. \end{aligned}$$

Making  $N_k \mapsto \infty$ , then  $\tau_{N_k} \mapsto \infty$ , and under our hypotheses on  $T$ ,  $\nu_\infty$ , and  $\nu'_\infty$ , we obtain

$$\left| \int \phi(y) d\nu_\infty(y) - \int \phi(y) d\mu(y) \right| \leq \epsilon.$$

Finally, we conclude that  $\nu_\infty = \mu$  and this fact ensures that  $(\mathbb{P}^n)_{n \in \mathbb{N}}$  and  $(\mathbb{P}^n(0))_{n \in \mathbb{N}}$  weakly converge toward  $\mu$ .

**8. Experiments.** We present here three experiments. The first one is a synthetic mixture model, and we compare our result with standard algorithms. The other databases are real problems on image processing and microarray data. In these last two cases, we use Fisher rule selection, random forest selection (see [8]), forward/backward selection, and OFW (optimal feature weighting) (see [18]) to draw comparisons with our method. In each of these cases, the number of selected features is computed using an internal cross-validation step.

### 8.1. Synthetic data.

**8.1.1. Description of the database.** We first test our algorithm on a simple synthetic example. We consider  $f = 100$  ternary variables ( $|\mathcal{F}| = 100$ ) and three classes (similar results can be obtained with more classes and variables). We let  $I \in \{-1; 0; 1\}^{100}$  and let  $\mathcal{G}$  be a subset of  $\mathcal{F}$ . We define the probability distribution  $\mu(\cdot; \mathcal{G})$  on  $\mathcal{I}$  to be the one for which all  $X^j$  in  $\mathcal{G}$  are independent,  $X^j(I)$  follows a uniform distribution on  $\{-1; 0; 1\}$  if  $X^j \notin \mathcal{G}$ , and  $X^j(I) = 1$  if  $X^j \in \mathcal{G}$ . We model each class by a mixture of such a distribution, including a small proportion of noise. More precisely, for a class  $C_i$ ,  $i = 1, 2, 3$ , we define

$$\mu_i(I) = \frac{q}{3} (\mu(I; \mathcal{G}_i^1) + \mu(I; \mathcal{G}_i^2) + \mu(I; \mathcal{G}_i^3)) + (1 - q)\mu(I; \emptyset),$$

with  $q = 0.9$  and

$$\begin{aligned} \mathcal{G}_1^1 &= \{X^1; X^3; X^5; X^7\}, & \mathcal{G}_1^2 &= \{X^1; X^5\}, & \mathcal{G}_1^3 &= \{X^3; X^7\}, \\ \mathcal{G}_2^1 &= \{X^2; X^4; X^6; X^8\}, & \mathcal{G}_2^2 &= \{X^2; X^4\}, & \mathcal{G}_2^3 &= \{X^6; X^8\}, \\ \mathcal{G}_3^1 &= \{X^1; X^4; X^8; X^9\}, & \mathcal{G}_3^2 &= \{X^1; X^8\}, & \mathcal{G}_3^3 &= \{X^4; X^9\}. \end{aligned}$$

We sample with this mixture model enough data to obtain well-conditioned statistical problems. We expect our learning algorithm to put large weights on features that compose the sets  $\mathcal{G}_i^j$  and to filter out the other noisy ones. The algorithm  $\mathbb{A}$  we use in this case is a  $p$  nearest neighbor classification algorithm, with distance given by

$$d(I_1, I_2) = \sum_j \mathbb{1}_{X^j(I_1) \neq X^j(I_2)}.$$

This synthetic example is interesting because it makes it possible to compute the exact gradient of  $\mathcal{E}$  for small values of  $M$  and  $k = |\omega|$ . See [17] and [18] for more details on this experiment when the set of features is fixed.

### 8.1.2. Results and comparisons with existing methods.

*OFW and jump algorithm.* We compare first the reflected diffusion (OFW of [18]) with our jump algorithm. Performances obtained with the jump algorithm are better than the ones without any jump as shown in Figure 3. The trees constructed by our algorithm are deeper than elementary ones since the mean depth achieved by our algorithm is 3. We compute the mean occupation measure of each tree in the process  $\mathcal{F}_t$  as

$$\mu_t(\mathcal{A}) = \frac{1}{t} \int_0^t \mathbb{1}_{\mathcal{A} \in \mathcal{F}_{ts}} ds.$$

We can then infer from this measure the importance of a tree while looking at the real numbers  $\mu_t(\mathcal{A})$ . We rank the nodes of these trees by decreasing importance of  $\mu_t(\mathcal{A})$  and we give the main roots detected by our algorithm below:

$$\{X^2; X^4\}, \{X^1; X^5\}, \{X^4; X^9\}, \{X^1; X^8\}, \{X^6; X^8\}, \{X^3; X^7\}, X^1, X^4, X^8.$$

It is important to remark that the nodes selected by our jump-diffusion algorithm are very similar to the sources  $\mathcal{G}_i^j$ , while the favored nodes are those which are *reusable features*.

One can, however, consider using standard feature selection techniques such as anova coupled with the logistic regression method or the more recent random forests feature selection.

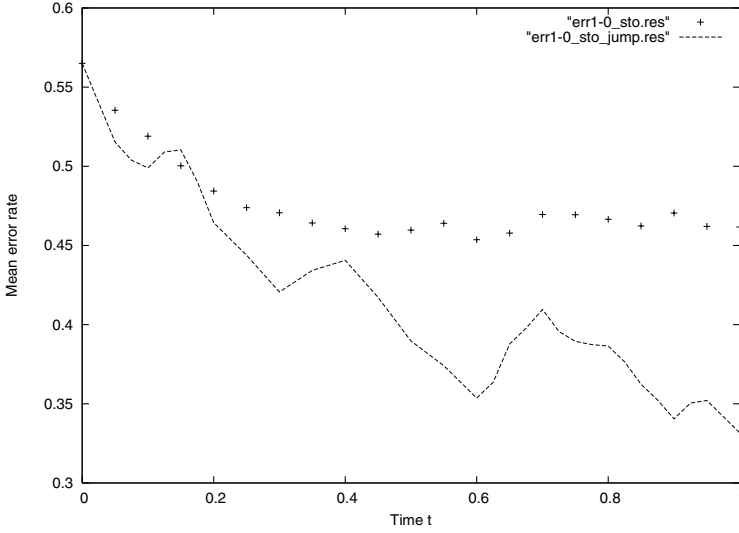


FIG. 3. Evolution of the mean error rate for the reflected diffusion (crosses) and the reflected jump diffusion (dashed line).

*Backward selection, random forest, and Fisher selection.* We run first a ternary (three classes) logistic regression coupled with a backward selection based on the anova criterion. In this particular case, the selected features cannot be composed and interactions with features are completely missed. We logically obtain the subset of selected variables  $\{X^1; X^2; X^3; X^4; X^5; X^6; X^7; X^8; X^9\}$  with small  $p$ -value. This result is not surprising and is coherent with the selected features of the OFW [18] (diffusion algorithm without the jump process). Note also that the selected features here are the same while running the random forest selection method and we are convinced that in this simple example, several other classical criteria, such as PLS (partial least square method), AIC, etc., achieve the same result. Lastly, we remark that we do not run a forward selection method with the logistic regression because of the high number of features, which make this greedy method numerically costly.

*Learning composition: The forward/backward selection.* Next, we use the classical forward/backward selection method combined with logistic regression [21] since other feature selection methods, such as random forest, do not provide any composed features. In this very simple example (there are “only” 100 variables although typical real applications will use thousands of variables), the computational time to run this forward/backward selection is much more important (it takes several hours to stabilize the model). In addition to each singleton  $X^i$ , we obtain all subsets  $\mathcal{G}_i^j$  given in the description of the way we construct our synthetic example.

*Comparison.* To conclude this section on the synthetic data, we observe that many other feature selection algorithms achieve the identification of useful variables.

Only one of them (the forward/backward method coupled with detection of interactions) can also compose features. This method has an important numerical cost. If this method succeeds in locating the interactions between features, it does not provide a selection as small as our method does.

Moreover, the main drawback in this framework is that the forward/backward criterion can be performed only with a sufficiently large database (we need to have

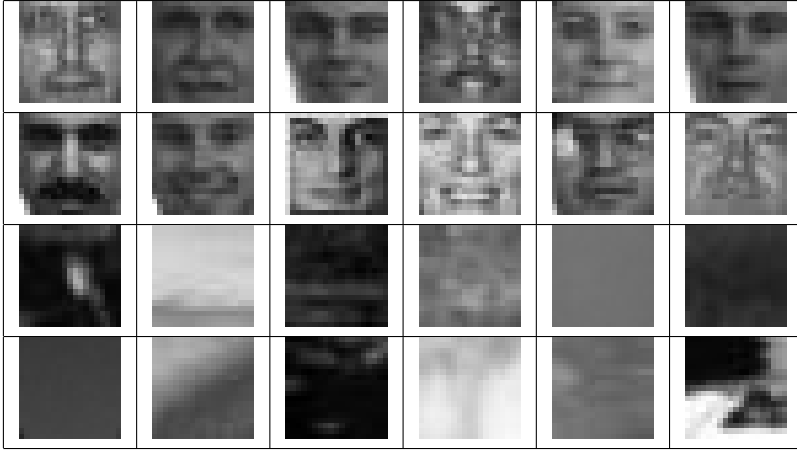


FIG. 4. Sample of images taken from [27] database.

more observations than initial number of variables). This point is very annoying in some real applications such as microarray analysis, where standard situations are described with thousands of variables for less than 100 observations.

Lastly, the forward/backward algorithm is dedicated only to very special classification algorithms such as logistic regression, although we can apply our approach to every classification algorithm  $\mathbb{A}$ . It is an important point too since there does not exist a universal classifier that beats all other algorithms. It can thus be helpful to run our meta-jump algorithm to the more appropriate  $\mathbb{A}$  regarding the database which is studied.

## 8.2. Face recognition.

**8.2.1. Description of the database.** We use in this section the face database from [27], which contains  $19 \times 19$  grayscale images. The elementary features in  $\mathcal{F}_0$  are simply edge detectors constructed by Amit and Geman in [2]. The initial number of elementary features in  $\mathcal{F}_0$  is nearly 2000. The number of observations in this database is 7000 in the training set and 23,000 in the test set. Figure 4 presents some examples of images taken in this database.

### 8.2.2. Results and comparisons.

*OFW and jump algorithm.* Efficiency of the reflected diffusion (OFW algorithm) is already described in [18]. In this paper, our approach permits largely improved error rates on the same datasets and we can easily give an interpretation of features constructed by our jump-diffusion process. To illustrate these advantages, we can plot first in Figures 5 and 6 the evolution of the number of trees selected by our algorithm with time  $t$ .

The decreasing of the number of trees is consequently important since starting with almost 2000 features, we reduce the amount of variables to below 800. Even if this number seems to be strictly decreasing in Figure 5, this is not the case if we “zoom” the evolution of the cardinal  $t \mapsto |\mathcal{F}_t|$ , as shown in Figure 6.

Moreover, by using a linear SVM and a voting procedure with the subsets  $\omega^{(i)}$  extracted with the process  $\mathbb{P}_t$ , we obtain a null false positive rate (images taken from the font class are perfectly classified) and the global misclassification rate is improved

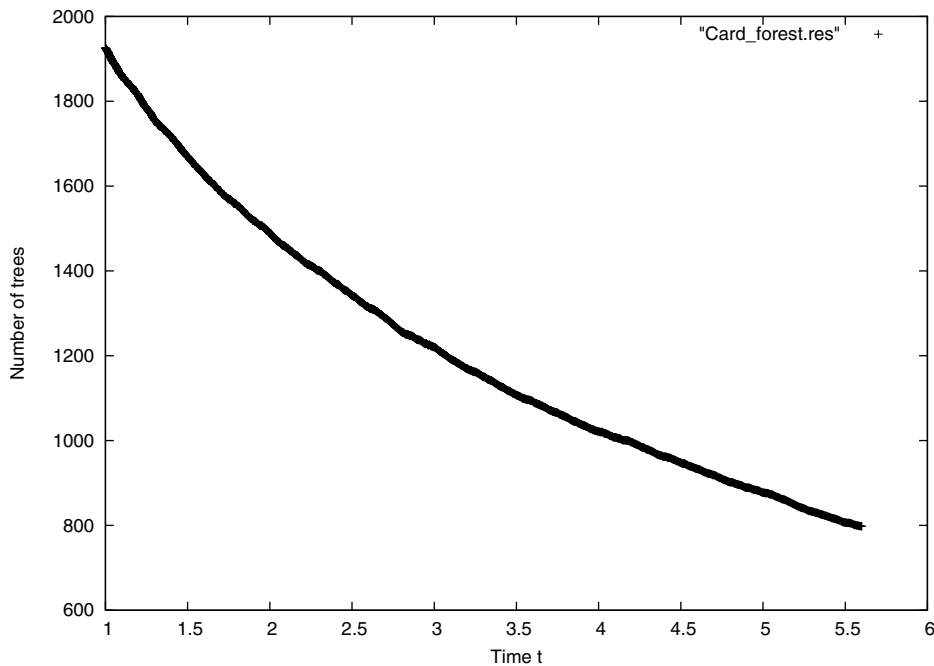


FIG. 5. Evolution of the number of trees in the forest  $\mathcal{F}_t$  with time for the face recognition problem.

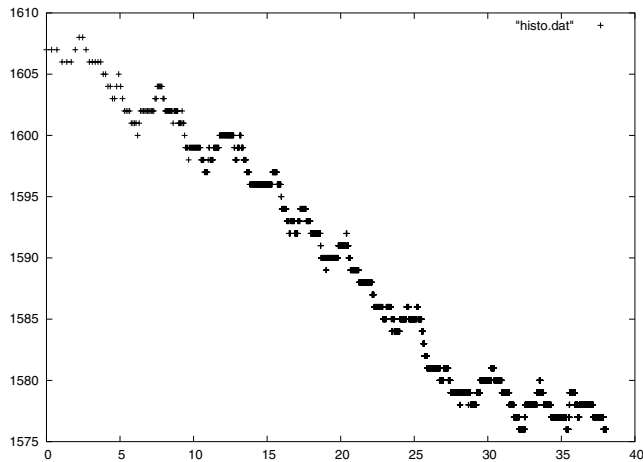


FIG. 6. Microscopic evolution of the number of trees in the forest  $\mathcal{F}_t$  for the faces experiment.

since we get 1.2% misclassified test samples using only 650 features (3.5% error rate with the OFW approach without jump and the same amount of features).

*Comparison with random forest, Fisher selection, and forward/backward selection.* Without features composition, the random forest classifier provides more than 1000 useful features and gives a general misclassification rate of 1%; note also that this rate has been achieved using 1000 trees in the random forest. The Fisher selection method combined with a linear SVM algorithm yields an error rate of 4% with a selection



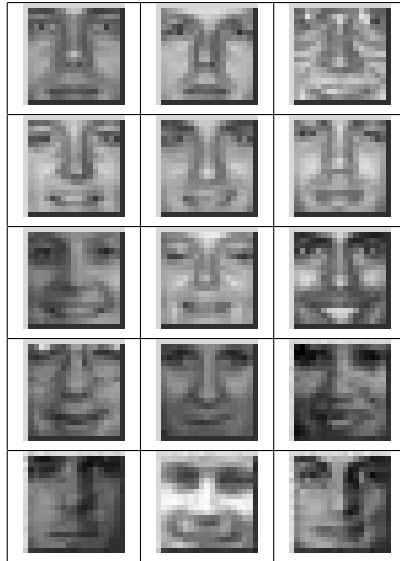


FIG. 7. Representation of the main aggregation of edge detectors selected by our process.

of more than 1000 features. Lastly, the logistic regression algorithm combined with forward/backward selection and composition performs poorly (more than 1000 useful features and 12% of misclassified signals). It seems here that the classifier  $\mathbb{A}$  used by the forward/backward (logistic regression) is not adequate for this database. This illustrates the fact that it is important to have a meta-algorithm to select features in order to apply any classifier which seems adapted to the problem. It is not the case with the forward/backward selection method.

In this case, the better global misclassification rate is obtained using the random forest selection method. Our method obtains good results too since only 1.2% of signals are misclassified. Lastly, the selection obtained using our jump algorithm is much more compact than the one obtained by random forest.

*Selected features.* We show in Figure 7 the main composition of edges selected by our process of jump diffusion. The important fact is that complex features (as well as elementary ones) are constructed and used by our algorithm. It is this point that permits us to obtain the perfect false positive rate since these complex compositions of features filter out the background images.

### 8.3. Leukemia microarray classification.

*Description of the database.* Finally, we benchmark our selection of features on the standard leukemia cancer dataset available online from the NCI.<sup>2</sup> Data are preprocessed and transformed into a collection of 3859 genes of 72 leukemia samples. They are divided into 47 samples of Acute Lymphoblastic Leukemia (ALL) and 25 samples of Acute Myeloblastic Leukemia (AML). As we cannot provide a simple meaning of concatenation of real variables, we only permit suppression ( $\mathcal{S}$ ) and rebirth ( $\mathcal{R}$ ) of some genes in  $\mathcal{F}_t$ . As this database does not contain any train or test sets, we estimate the misclassification rate using a tenfold cross-validation method. The cross-validation method is a good way to estimate performances of our algorithm [7].

<sup>2</sup>National Cancer Institute, <http://www.cancer.gov>.

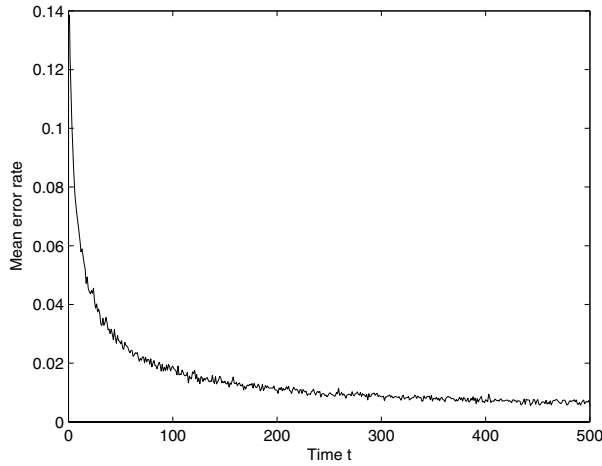


FIG. 8. Mean error rate on the training set of the ALL-AML database with time  $t$ .

TABLE 1  
Rate of misclassified samples using several number of genes selected by our algorithm.

Number of genes	OFW	Reflected jump diffusion	Random forest	F-test
4-9	6.9%	4.8%	4%	5.5%
10-19	5.5%	4.3%	3.8%	4.2%
20-24	4.1%	2.1%	4.5%	4.1%
25-45	3.5%	0.9%	4.5%	4.8%

*Comparison with random forest, Fisher selection, and forward/backward selection.* After a learning procedure, we then ranked genes by a decreasing importance criterion based on the probability distribution  $\mathbb{P}_t$ . For the jump-diffusion method, we do not run a tenfold cross validation because of the time of computation needed by this method, and we then employ a more simple three-fold cross validation.

We use for  $\mathbb{A}$  a linear SVM classifier, and  $k = 100$  genes are extracted at each step. The evolution of the error rate all along our learning algorithm is shown in Figure 8. We obtain in [9] interesting results on classification rates on this database applying other algorithms such as CART to the OFW meta-algorithm.

We present in Table 1 results obtained using our jump process. We cannot run here a logistic regression coupled with a forward/backward algorithm because of the small number of signals in the database. The several selection methods used highlight the good performance of our jump algorithm, comparing it to standard methods such as Fisher tests.

Our results improve those referred to in [22], and the genes selected by our algorithm are consistent with some of the genes selected in other works (such as Zyxin in [11]). However, our selected features are nearly similar to those reported in [19]. One can again note the improvement using the jump process (second and third columns of Table 1).

Figure 9 represents the evolution of the number of genes selected at time  $t$ . In this case, we note again the good dimensionality reduction that permits our algorithm.

Finally, we can extract from the set of variables the names of genes most selected by our algorithm. We do not obtain exactly the same results for the 10 most important

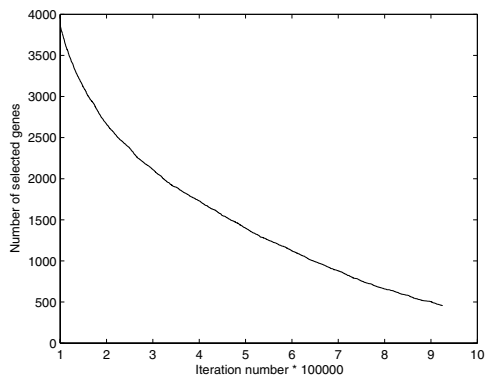


FIG. 9. Evolution of the number of trees in the forest  $\mathcal{F}_t$  with iteration number  $n$ .

OFW Algorithm	Reflected jump diffusion
CTSD Cathepsin D	CTSD Cathepsin D
MPO Myeloperoxidase	MPO Myeloperoxidase
MB-1 gene	MB-1 gene
Catalase (EC 1.11.1.6)	Catalase (EC 1.11.1.6)
PROTEASOME IOTA CHAIN	Kazal-type serine proteinase
Zyxin	PROTEASOME IOTA CHAIN
Terminal transferase mRNA	VIL2 Villin 2
Kazal-type serine proteinase	PRG1 Proteoglycan 1
CCND3 Cyclin D3	CD37 CD37 antigen
CD37 CD37 antigen	HLA CLASS I HISTO. ANTIGEN

FIG. 10. Genes most selected by our algorithm.

genes listed in Figure 10 whether we use the OFW or the jump-diffusion approach.

**9. Conclusion.** From a theoretical point of view, we provide in this paper a mathematical algorithm to select variables in a large amount of features dealing with the general untractable problems using full data. This is not the case of filter methods (forward/backward, for instance) that use a heuristic strategy to compose features, and these methods are not useful in some situations. Our approach is based on a jump-diffusion stochastic differential equation, where jumps are transitions between spaces of features. We have seen that the structure of trees is convenient to deal with Markov processes since this enables us to identify the dynamical structure of our method. This method is highly motivated by real problems and we have shown (Theorem 7.2) the “optimality” of our algorithm since it converges toward the unique Gibbs field measure inferred from an energy  $\mathcal{E}$ .

From a practical point of view, we have reached interesting results in real data such as face recognition and microarray analysis, even if we do not perform any composition rule with this last database. We have obtained similar results as other standard methods on the synthetic example and have clearly overcome the forward/backward algorithm in the face recognition problem, which is the only other method known to permit features composition. On this last point, one can consider two hypotheses. Either the selected features are not so good with the forward/backward strategy

(it would be surprising) or (and it is the more likely) the classifier used after this selection is powerless compared to the SVM used with our method. This stresses the fact that our approach is usable with any classification algorithm: One can use for  $\mathbb{A}$  SVMs, linear discriminant analysis, random forests, etc., and it is well known that at the moment there does not exist one algorithm which performs best on all pattern recognition problems.

In a forthcoming paper, we will present several computational results on this algorithm applied to several databases described by thousands of variables. Numerically, it would be interesting to use our composition strategy with real variables (instead of binary or ternary ones) since we have not used it on the leukemia database, for instance.

Similarly, it would be useful to interpret the composition of real variables as a process to learn a kernel for the SVM. We believe that using a Rademacher penalty term in energy  $\mathcal{E}$  will improve the generalization ability of the algorithm and could permit us to obtain Oracle's inequality. Another improvement can be made using a simulated annealing strategy to fix the selected features to a deterministic version in the end of the algorithm.

**Acknowledgments.** This paper contains research performed during my thesis and I am glad to thank my Ph.D. advisor Laurent Younes for the numerous and helpful discussions we had on this occasion.

#### REFERENCES

- [1] Y. AMIT AND D. GEMAN, *Shape quantization and recognition with randomized trees*, Neural Computation, 9 (1997), pp. 1545–1588.
- [2] Y. AMIT AND D. GEMAN, *A computational model for visual selection*, Neural Computation, 11 (1999), pp. 1691–1715.
- [3] R. F. ANDERSON AND S. OREY, *Small random perturbation of dynamical systems with reflecting boundary*, Nagoya Math. J., 60 (1976), pp. 189–216.
- [4] R. ATAR, A. BUDHIRAJA, AND P. DUPUIS, *Correction note: On positive recurrence of constrained diffusion processes*, Ann. Probab., 29 (2001), p. 1404.
- [5] A. BENVENISTE, M. MÉTIVIER, AND P. PRIOURET, *Algorithmes adaptatifs et approximations stochastiques*, Théorie et applications à l'identification, au traitement du signal et à la reconnaissance des formes, Masson, Paris, 1987. English translation available as *Adaptive Algorithms and Stochastic Approximations*, Appl. Math. 22, Springer-Verlag, Berlin, 1990.
- [6] P. BILLINGSLEY, *Convergence of Probability Measures*, 2nd ed., Wiley Ser. Probab. Statist. Probab. Statist., John Wiley, New York, 1999.
- [7] U. M. BRAGA-NETO AND E. R. DOUGHERTY, *Is cross-validation valid for small-sample microarray classification?*, Bioinformatics, (2003), pp. 1061–1069.
- [8] L. BREIMAN, *Arcing classifiers*, Ann. Statist., 26 (1998), pp. 801–849.
- [9] K.-A. LÊ CAO, O. GONÇALVES, P. BESSE, AND S. GADAT, *Selection of biologically relevant genes with a wrapper stochastic algorithm*, Statistical Applications in Genetics and Molecular Biology, 6 (2007), article 29.
- [10] O. CHAPPELLE, V. VAPNIK, O. BOUSQUET, AND S. MUKHERJEE, *Choosing multiple parameters for support vector machines*, Machine Learning, 46 (2002), pp. 131–159.
- [11] K. DEB AND R. REDDY, *Classification of Two-Class Cancer Data Reliably Using Evolutionary Algorithms*, KanGAL report 2003001, Kanpur Genetics Algorithms Laboratory, Kanpur, India, 2003. <http://www.iitk.ac.in/kangal/papers/k2003001.pdf>
- [12] P. DUPUIS AND H. ISHII, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics and Stochastics Rep., 35 (1991), pp. 31–62.
- [13] P. DUPUIS AND K. RAMANAN, *Convex duality and the Skorokhod problem*. I, II, Probab. Theory Related Fields, 115 (1999), pp. 153–195; 197–236.
- [14] S. ETHIER AND T. KURTZ, *Markov Processes*, John Wiley, New York, 1986.
- [15] F. FLEURET, *Fast binary feature selection with conditional mutual information*, J. Mach. Learn. Res., (2004), pp. 1531–1555.

- [16] F. FLEURET AND D. GEMAN, *Coarse-to-fine face detection*, International Journal of Computer Vision, 41 (2001), pp. 85–107.
- [17] S. GADAT, *Apprentissage d'un vocabulaire symbolique pour la détection d'objets dans une image*, Thèse de l'École Normale Supérieure de Cachan, 2004.
- [18] S. GADAT AND L. YOUNES, *A stochastic algorithm for feature selection in pattern recognition*, J. Mach. Learn. Res., 8 (2007), pp. 509–547.
- [19] D. GEMAN, C. D'AVIGNON, D. NAIMAN, AND R. WINSLOW, *Classifying gene expression profiles from pairwise MRNA comparisons*, Statistical Applications in Genetics and Molecular Biology, 3 (2004), article 19.
- [20] D. GEMAN AND B. JEDYNAK, *Model-based classification trees*, IEEE Trans. Inform. Theory, 47 (2001), pp. 1075–1082.
- [21] A. S. GOLDBERGER AND D. B. JOCHEMS, *Note on stepwise least squares*, J. Amer. Statist. Assoc., 56 (1961), pp. 105–110.
- [22] T. R. GOLUB, D. K. SLONIM, P. TAMAZYO, C. HUARD, M. GAASENBEEK, J. P. MESIROV, H. COLLIER, M. L. LOH, J. R. DOWNING, M. A. CALIGIURI, C. D. BLOOMFIELD, AND E. S. LANDER, *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring*, Science, 286 (1999), pp. 531–537.
- [23] J. M. HARRISON AND R. J. WILLIAMS, *Brownian models of feedforward queueing networks: Quasireversibility and product form solutions*, Ann. Appl. Probab., 2 (1992), pp. 263–293.
- [24] C. JUTTEN AND J. HÉRAULT, *Blind separation of sources, part 1: An adaptative algorithm based on neuromimetic architecture*, Signal Process., 24 (1991), pp. 1–10.
- [25] S. KREMPP, D. GEMAN, AND Y. AMIT, *Sequential Learning of Reusable Parts for Object Detection*, Technical report, Center for Imaging Science, Johns Hopkins University, Baltimore, MD, 2002. [http://cis.jhu.edu/publications/papers\\_in\\_database/GEMAN/seqlearning.pdf](http://cis.jhu.edu/publications/papers_in_database/GEMAN/seqlearning.pdf)
- [26] H. J. KUSHNER, AND G. G. YIN, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Appl. Math. (New York), 35, Stochastic Modelling and Applied Probability, Springer-Verlag, New York, 2003.
- [27] MIT, *CBCL Face Database*, Center for Biological and Computation Learning, MIT, Cambridge, MA, 2000. <http://cbcl.mit.edu/software-datasets/FaceData2.html>
- [28] J. RISSANEN, *A universal prior for integers and estimation by minimum description length*, Ann. Statist., 11 (1983), pp. 416–431.
- [29] A. SRIVASTAVA, M. I. MILLER, AND U. GRENANDER, *Ergodic algorithms on special Euclidean groups for ATR*, in Systems and Control in the Twenty-First Century (St. Louis, MO, 1996), Progr. Systems Control Theory 22, Birkhäuser Boston, Boston, MA, 1997, pp. 327–350.
- [30] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Grundlehren der Math. Wiss. [Fundamental Principles of Math. Sci.] 233, Springer-Verlag, Berlin, 1979.
- [31] J. WESTON, S. MUKHERJEE, O. CHAPPELLE, M. PONTIL, T. POGGIO, AND V. VAPNIK, *Feature selection for SVMs*, in Proceedings of the Neural Information Processing Systems Conference (NIPS 2000), MIT Press, Cambridge, MA, pp. 668–674.

## STABILITY ANALYSIS OF SWITCHED TIME DELAY SYSTEMS\*

PENG YAN<sup>†</sup> AND HITAY ÖZBAY<sup>‡</sup>

**Abstract.** This paper addresses the asymptotic stability of switched time delay systems with heterogeneous time invariant time delays. Piecewise Lyapunov–Razumikhin functions are introduced for the switching candidate systems to investigate the stability in the presence of an infinite number of switchings. We provide sufficient conditions in terms of the minimum dwell time to guarantee asymptotic stability under the assumptions that each switching candidate is delay-independently or delay-dependently stable. Conservatism analysis is also provided by comparing with the dwell time conditions for switched delay-free systems. Finally, a numerical example is given to validate the results.

**Key words.** asymptotic stability, switched systems, time delay, dwell time

**AMS subject classifications.** 93D05, 93D20, 93C05, 93C23

**DOI.** 10.1137/060668262

**1. Introduction.** Switching control offers a new look into the design of complex control systems (e.g., nonlinear systems, parameter varying systems, and uncertain systems) [1, 8, 9, 19, 21, 27]. Unlike the conventional adaptive control techniques that rely on continuous tuning, the switching control method updates the controller parameters in a discrete fashion based on the switching logic. The resulting closed-loop systems have hybrid behaviors (e.g., continuous dynamics, discrete time dynamics, and jump phenomena). One of the most challenging issues in the area of hybrid systems is the stability analysis in the presence of control switching. We refer to [9] for a general review on switching control methods.

In particular, we are interested in the stability analysis of switched time delay systems. In fact, time delay systems are ubiquitous in chemical processes, aerodynamics, and communication networks [3, 14]. To further complicate the situation, the time delays are usually time varying and uncertain [24, 26]. It has been shown that robust  $\mathcal{H}^\infty$  controllers can be designed for such infinite-dimensional plants, where robustness can be guaranteed within some uncertainty bounds [4]. In order to incorporate a larger operating range or better robustness, controller switching can be introduced, which results in switched closed-loop systems with time delays. For delay-free systems, stability analysis and design methodology have been investigated recently in the framework of hybrid dynamical systems [1, 2, 8, 11, 19, 21, 25]. In particular, [21] provided sufficient conditions on the stability of the switching control systems based on Filippov solutions to discontinuous differential equations and Lyapunov functionals; [19] proposed a dwell-time-based switching control, where a sufficiently large dwell time can guarantee system stability. A more flexible result was obtained in [10], where the average dwell time was introduced for switching control. In [25] the results of [10] were extended to linear parameter varying (LPV) systems. LaSalle’s invariance

---

\*Received by the editors August 24, 2006; accepted for publication (in revised form) October 28, 2007; published electronically February 29, 2008. A brief version of this paper was presented at the *IFAC* World Congress, 2005. This work was supported in part by the European Commission under contract MIRC-CT-2004-006666 and by TÜBİTAK under grant EEEAG-105E156.

<http://www.siam.org/journals/sicon/47-2/66826.html>

<sup>†</sup>Seagate Technology, 1280 Disc Drive, Shakopee, MN 55379 (Peng.Yan@seagate.com).

<sup>‡</sup>Department of Electrical & Electronics Engineering, Bilkent University, Ankara 06800, Turkey (hitay@bilkent.edu.tr).

principle was extended to a class of switched linear systems for stability analysis [8]. Despite the variety and significance of the many results on hybrid system stability, stability of switched time delay systems hasn't been adequately addressed due to the general difficulty of infinite-dimensional systems [7].

Two important approaches in the stability analysis of time delay systems are the (1) Lyapunov–Krasovskii method and (2) Lyapunov–Razumikhin method [6, 20]. Various sufficient conditions with respect to the stability of time delay systems have been given using Riccati-type inequalities or linear matrix inequalities (LMIs) [3, 12, 14, 24]. Meanwhile, stability analysis in the presence of switching has been discussed in some recent works [16, 18, 22]. In [18] stability and stabilizability were discussed for discrete time switched time delay systems; [16] considered a similar stability problem in a continuous time domain. Note that [18] and [16] produce *trajectory-dependent* results without taking admissible switching signals into consideration.

The main contribution of this paper is a collection of results on the *trajectory-independent* stability of continuous time switched time delay systems using piecewise Lyapunov–Razumikhin functions. The dwell time of the switching signals is constructively given, which guarantees asymptotic stability for the delay-independent case and the delay-dependent case, respectively. Note that the asymptotic stability of finite-dimensional linear systems indicates exponential stability, while this is not the case for infinite-dimensional systems [7, 15]. This poses the key challenge in the analysis of switched time delay systems, where we do not assume exponential convergence of the switching candidates, as opposed to most of the results in the literature [8, 10, 17, 19].

The paper is organized as follows. The problem is defined in section 2. In section 3, the main results on the stability of switched time delay systems are presented in terms of the dwell time of the switching signals. Conservatism analysis is provided by comparing with the dwell-time conditions for switching delay-free systems in section 4. The results are illustrated with a numerical example in section 5, followed by concluding remarks in section 6.

**2. Problem definition.** For convenience, we would like to employ the following notation. The general retarded functional differential equations (RFDEs) with time delay  $r$  can be described as

$$(2.1) \quad \dot{x}(t) = f(t, x_t)$$

with initial condition  $\phi(\cdot) \in C([-r, 0], \mathbb{R}^n)$ , where  $x_t$  denotes the state defined by  $x_t(\theta) = x(t + \theta)$ ,  $-r \leq \theta \leq 0$ . We use  $\|\cdot\|$  to denote the Euclidean norm of a vector in  $\mathbb{R}^n$ , and  $\|f\|_{[t-r, t]}$  for the  $\infty$ -norm of  $f$ , i.e.,

$$\|f\|_{[t-r, t]} := \sup_{t-r \leq \theta \leq t} \|f(\theta)\|,$$

where  $f$  is an element of the Banach space  $C([t-r, t], \mathbb{R}^n)$ .

Consider the following switched time delay systems:

$$(2.2) \quad \Sigma_t : \begin{cases} \dot{x}(t) = A_{q(t)}x(t) + \bar{A}_{q(t)}x(t - \tau_{q(t)}), & t \geq 0, \\ x_0(\theta) = \phi(\theta) \quad \forall \theta \in [-\tau_{max}, 0], \end{cases}$$

where  $x(t) \in \mathbb{R}^n$  and  $q(t)$  is a piecewise switching signal taking values on the set  $\mathcal{F} := \{1, 2, \dots, l\}$ , i.e.,  $q(t) = k_j$ ,  $k_j \in \mathcal{F} \quad \forall t \in [t_j, t_{j+1})$ , where  $t_j$ ,  $j \in \mathbb{Z}^+ \cup \{0\}$ , is the  $j$ th switching time instant. It is clear that the trajectory of  $\Sigma_t$  in any arbitrary

switching interval  $t \in [t_j, t_{j+1})$  obeys

$$(2.3) \quad \Sigma_{k_j} : \begin{cases} \dot{x}(t) = A_{k_j}x(t) + \bar{A}_{k_j}x(t - \tau_{k_j}), & t \in [t_j, t_{j+1}), \\ x_{t_j}(\theta) = \phi_j(\theta) \quad \forall \theta \in [-\tau_{k_j}, 0], \end{cases}$$

where  $\phi_j(\theta)$  is defined as

$$(2.4) \quad \phi_j(\theta) = \begin{cases} x(t_j + \theta), & -\tau_{k_j} \leq \theta < 0, \\ \lim_{h \rightarrow 0^-} x(t_j + h), & \theta = 0. \end{cases}$$

We introduce the triplet  $\Sigma_i := (A_i, \bar{A}_i, \tau_i) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n} \times \mathbb{R}^+$  to describe the  $i$ th candidate system of (2.2). Thus  $\forall t \geq 0$ , we have  $\Sigma_t \in \mathcal{A} := \{\Sigma_i : i \in \mathcal{F}\}$ , where  $\mathcal{A}$  is the family of candidate systems of (2.2). In (2.2),  $\phi(\cdot) : [-\tau_{max}, 0] \rightarrow \mathbb{R}^n$  is a continuous and bounded vector-valued function, where  $\tau_{max} = \max_{i \in \mathcal{F}} \{\tau_i\}$  is the maximal time delay of the candidate systems in  $\mathcal{A}$ .

Similar to [8], we say that the switched time delay system  $\Sigma_t$  described by (2.2) is *stable* if there exists a function  $\bar{\alpha}$  of class  $\mathcal{K}^1$  such that

$$(2.5) \quad \|x(t)\| \leq \bar{\alpha}(\|x\|_{[t_0 - \tau_{max}, t_0]}) \quad \forall t \geq t_0 \geq 0$$

along the trajectory of (2.2). Furthermore,  $\Sigma_t$  is *asymptotically stable* when  $\Sigma_t$  is stable and  $\lim_{t \rightarrow +\infty} x(t) = 0$ .

LEMMA 2.1 (see [3, 14]). *Suppose for a given triplet  $\Sigma_i \in \mathcal{A}$ ,  $i \in \mathcal{F}$ , there exists symmetric and positive-definite  $P_i \in \mathbb{R}^{n \times n}$ , such that the following LMI with respect to  $P_i$  is satisfied for some  $p_i > 1$  and  $\alpha_i > 0$ :*

$$(2.6) \quad \begin{bmatrix} P_i A_i + A_i^T P_i + p_i \alpha_i P_i & P_i \bar{A}_i \\ \bar{A}_i^T P_i & -\alpha_i P_i \end{bmatrix} < 0.$$

*Then  $\Sigma_i$  is asymptotically stable independent of delay.*

If all candidate systems of (2.2),  $\Sigma_i \in \mathcal{A}$ , are delay-independently asymptotically stable satisfying (2.6), we denote  $\mathcal{A}$  by  $\tilde{\mathcal{A}}$ .

LEMMA 2.2 (see [3, 14]). *Suppose for a given triplet  $\Sigma_i \in \mathcal{A}$ ,  $i \in \mathcal{F}$ , there exists symmetric and positive-definite  $P_i \in \mathbb{R}^{n \times n}$ , and a scalar  $p_i > 1$ , such that*

$$(2.7) \quad \begin{bmatrix} \tau_i^{-1} \Omega_i & P_i \bar{A}_i M_i \\ M_i^T \bar{A}_i^T P_i & -R_i \end{bmatrix} < 0,$$

where

$$\begin{aligned} \Omega_i &= (A_i + \bar{A}_i)^T P_i + P_i (A_i + \bar{A}_i) + \tau_i p_i (\alpha_i + \beta_i) P_i, \\ M_i &= [A_i \quad \bar{A}_i], \\ R_i &= \text{diag}(\alpha_i P_i, \beta_i P_i), \end{aligned}$$

*and  $\alpha_i > 0$ ,  $\beta_i > 0$  are scalars. Then  $\Sigma_i$  is asymptotically stable dependent on delay.*

Similarly we denote  $\mathcal{A}$  by  $\tilde{\mathcal{A}}_d$  if all candidate systems of (2.2) are delay-dependently asymptotically stable satisfying (2.7).

<sup>1</sup>A continuous function  $\bar{\alpha}(\cdot) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  is a class  $\mathcal{K}$  function if it is strictly increasing and  $\bar{\alpha}(0) = 0$ .



In what follows, we will establish sufficient conditions to guarantee stability of switched system (2.2) for the delay-independent case and the delay-dependent case. Therefore, we will assume that  $\mathcal{A} = \tilde{\mathcal{A}}$  and  $\mathcal{A} = \tilde{\mathcal{A}}_d$ , respectively, in the corresponding sections in this paper. An important method in stability analysis of switched systems is based on the construction of the common Lyapunov function (CLF), which allows for arbitrary switching. However, this method is too conservative from the perspective of controller design because it is usually difficult to find the CLF for all the candidate systems, particularly for time delay systems whose stability criteria are only sufficient in most of the circumstances. A recent paper [28] explored the CLF method for switched time delay systems with three very strong assumptions: (i) each candidate system has the same time delay  $\tau$ ; (ii) each candidate is assumed to be delay-independently stable; (iii) the  $A$ -matrix is always symmetric and the  $\bar{A}$ -matrix is always in the form of  $\delta I$ . In the present paper, we consider an alternative method using piecewise Lyapunov–Razumikhin functions for a general class of systems (2.2) and obtain stability conditions in terms of the dwell time of the switching signal. This method can be used for the case with delay-independent criterion (2.6) and the case with delay-dependent criterion (2.7).

**3. Main results on dwell-time-based switching.** For a given positive constant  $\tau_D$ , the switching signal set based on the dwell time  $\tau_D$  is denoted by  $S[\tau_D]$ , where for any switching signal  $q(t) \in S[\tau_D]$ , the distance between any consecutive discontinuities of  $q(t)$ ,  $t_{j+1} - t_j$ ,  $j \in \mathbb{Z}^+ \cup \{0\}$ , is larger than  $\tau_D$  [10, 19]. A sufficient condition on the minimum dwell time to guarantee the stable switching will be given using piecewise Lyapunov–Razumikhin functions. Note that the dwell-time-based switching is trajectory independent [8].

Before presenting the main result of this paper, we recall the following lemma [7] for general RFDEs (2.1).

LEMMA 3.1 (see [7]). *Suppose  $u, v, w, p : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  are continuous, nondecreasing functions,  $u(0) = v(0) = 0$ ,  $u(s), v(s), w(s), p(s)$  positive for  $s > 0$ ,  $p(s) > s$ , and  $v(s)$  strictly increasing. If there is a continuous function  $V : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$  such that*

$$(3.1) \quad u(\|x(t)\|) \leq V(t, x) \leq v(\|x(t)\|), \quad t \in \mathbb{R}, x \in \mathbb{R}^n,$$

and

$$(3.2) \quad \dot{V}(t, x(t)) \leq -w(\|x(t)\|)$$

if

$$(3.3) \quad V(t + \theta, x(t + \theta)) < p(V(t, x(t))) \quad \forall \theta \in [-r, 0],$$

then the solution  $x = 0$  of the RFDE is uniformly asymptotically stable.

A particular case of (2.1) is a linear time delay system  $\Sigma_i$ ,  $i \in \mathcal{F}$ , where we can construct the corresponding Lyapunov–Razumikhin function in the quadratic form

$$(3.4) \quad V_i(t, x) = x^T(t) P_i x(t), \quad P_i = P_i^T > 0.$$

Apparently  $V_i$  can be bounded by

$$(3.5) \quad u_i(\|x(t)\|) \leq V_i(t, x) \leq v_i(\|x(t)\|) \quad \forall x \in \mathbb{R}^n,$$

where

$$(3.6) \quad u_i(s) := \kappa_i s^2, \quad v_i(s) := \bar{\kappa}_i s^2,$$

in which  $\kappa_i := \sigma_{\min}[P_i] > 0$  denotes the smallest singular value of  $P_i$  and  $\bar{\kappa}_i := \sigma_{\max}[P_i] > 0$  the largest singular value of  $P_i$ .

PROPOSITION 3.2. *For each time delay system  $\Sigma_i$  with Lyapunov–Razumikhin function defined by (3.4), assume that (3.2) and (3.3) are satisfied for some  $w_i(s)$ . Then we have*

$$(3.7) \quad |x|_{[t_m - \tau_i, t_m]} \leq \left( \frac{\bar{\kappa}_i}{\kappa_i} \right)^{1/2} |x|_{[t_n - \tau_i, t_n]} \quad \forall t_m \geq t_n \geq 0.$$

*Proof.* Define

$$(3.8) \quad \bar{V}_i(t, x) := \sup_{-\tau_i \leq \theta \leq 0} V_i(t + \theta, x(t + \theta))$$

for  $t \geq 0$ . We have

$$(3.9) \quad \kappa_i(|x|_{[t - \tau_i, t]})^2 \leq \bar{V}_i(t, x) \leq \bar{\kappa}_i(|x|_{[t - \tau_i, t]})^2, \quad t \geq 0.$$

The definition of  $\bar{V}_i(t, x)$  implies  $\exists \theta_0 \in [-\tau_i, 0]$ , such that  $\bar{V}_i(t, x) = V(t + \theta_0, x(t + \theta_0))$ . Introduce the upper right-hand derivative of  $\bar{V}_i(t, x)$  as

$$\dot{\bar{V}}_i^+ = \limsup_{h \rightarrow 0^+} \frac{1}{h} [\bar{V}_i(t + h, x(t + h)) - \bar{V}_i(t, x(t))].$$

We have the following:

- (i) If  $\theta_0 = 0$ , i.e.,  $V_i(t + \theta, x(t + \theta)) \leq V_i(t, x(t)) < p(V_i(t, x(t)))$ , we have  $\dot{\bar{V}}_i(t, x) < 0$  by (3.2). Therefore  $\dot{\bar{V}}_i^+ \leq 0$ .
- (ii) If  $-\tau_i < \theta_0 < 0$ , we have  $\bar{V}_i(t + h, x(t + h)) = \bar{V}_i(t, x)$  for  $h > 0$  sufficiently small, which results in  $\dot{\bar{V}}_i^+ = 0$ .
- (iii) If  $\theta_0 = -\tau_i$ , the continuity of  $V_i(t, x)$  implies  $\dot{\bar{V}}_i^+ \leq 0$ .

The above analysis shows that

$$(3.10) \quad \bar{V}_i(t_m) \leq \bar{V}_i(t_n) \quad \forall t_m \geq t_n \geq 0.$$

Recalling (3.9), we have

$$(3.11) \quad \kappa_i(|x|_{[t_m - \tau_i, t_m]})^2 \leq \bar{V}_i(t_m) \leq \bar{V}_i(t_n) \leq \bar{\kappa}_i(|x|_{[t_n - \tau_i, t_n]})^2$$

for any  $t_m \geq t_n \geq 0$ . This implies (3.7) and proves the result.  $\square$

Suppose all of the conditions of Lemma 3.1 are satisfied for general RFDEs (2.1). We also have the following result.

LEMMA 3.3 (see [7]). *Suppose  $|\phi|_{[t_0 - r, t_0]} \leq \bar{\delta}_1$ ,  $\bar{\delta}_1 > 0$ , and  $\bar{\delta}_2 > 0$  such that  $v(\bar{\delta}_1) = u(\bar{\delta}_2)$ . For all  $\eta$  satisfying  $0 < \eta \leq \bar{\delta}_2$ , we have*

$$(3.12) \quad V(t, x) \leq u(\eta) \quad \forall t \geq t_0 + T.$$

Here

$$(3.13) \quad T = \frac{Nv(\bar{\delta}_1)}{\gamma}$$

is defined by  $\gamma = \inf_{v^{-1}(u(\eta)) \leq s \leq \bar{\delta}_2} w(s)$  and  $N = \lceil (v(\bar{\delta}_1) - u(\eta))/a \rceil$ , where  $\lceil \cdot \rceil$  is the ceiling integer function and  $a > 0$  satisfies  $p(s) - s > a$  for  $u(\eta) \leq s \leq v(\bar{\delta}_1)$ .

**3.1. The case with delay-independent criterion.** Consider the switched time delay systems  $\Sigma_t$  defined by (2.2) and assume each candidate system  $\Sigma_i$ ,  $i \in \mathcal{F}$ , delay-independently asymptotically stable satisfying (2.6) (i.e.,  $\mathcal{A} = \tilde{\mathcal{A}}$ ). A sufficient condition on the minimum dwell time to guarantee the asymptotic stability can be derived using multiple piecewise Lyapunov–Razumikhin functions. In order to state the main result we make some preliminary definitions.

For the switched delay systems (2.2), first assume  $\tau_D > \tau_{max}$ . Consider an arbitrary switching interval  $[t_j, t_{j+1})$  of the piecewise switching signal  $q(t) \in S[\tau_D]$ , where  $q(t) = k_j$ ,  $k_j \in \mathcal{F} \forall t \in [t_j, t_{j+1})$  and  $t_j$  is the  $j$ th switching time instant for  $j \in \mathbb{Z}^+ \cup \{0\}$  and  $t_0 = 0$ . The state variable  $x_j(t)$  defined on this interval obeys (2.3). For the convenience of using “sup”, we define  $x_j(t_{j+1}) = \lim_{h \rightarrow 0^-} x_j(t_{j+1} + h) = x_{j+1}(t_{j+1})$  based on the fact that  $x(t)$  is continuous for  $t \geq 0$ . Therefore  $x_j(t)$  is now defined on a compact set  $[t_j, t_{j+1}]$ . Recall (2.4); the initial condition  $\phi_j(t)$  of  $\Sigma_{k_j}$  is  $\phi_j(t) = x(t) = x_{j-1}(t)$ ,  $t \in [t_j - \tau_{k_j}, t_j]$  for  $j \in \mathbb{Z}^+$ , which is true because  $\tau_D > \tau_{max}$ .

Construct the Lyapunov–Razumikhin function

$$(3.14) \quad V_{k_j}(x_j, t) = x_j^T(t) P_{k_j} x_j(t), \quad t \in [t_j, t_{j+1}],$$

for (2.3). Then we have

$$(3.15) \quad \kappa_{k_j} \|x_j(t)\|^2 \leq V_{k_j}(t, x_j) \leq \bar{\kappa}_{k_j} \|x_j(t)\|^2 \quad \forall x_j \in \mathbb{R}^n.$$

A straightforward calculation gives the time derivative of  $V_{k_j}(t, x_j(t))$  along the trajectory of (2.3),

$$(3.16) \quad \dot{V}_{k_j}(t, x_j) = x_j^T (A_{k_j}^T P_{k_j} + P_{k_j} A_{k_j}) x_j + 2x_j^T(t) P_{k_j} \bar{A}_{k_j} x_j(t - \tau_{k_j}),$$

where

$$\begin{aligned} & 2x_j^T(t) P_{k_j} \bar{A}_{k_j} x_j(t - \tau_{k_j}) \\ & \leq \alpha_{k_j} x_j^T(t - \tau_{k_j}) P_{k_j} x_j(t - \tau_{k_j}) \\ & \quad + \alpha_{k_j}^{-1} x_j^T(t) P_{k_j} \bar{A}_{k_j} P_{k_j}^{-1} \bar{A}_{k_j}^T P_{k_j} x_j(t) \quad \forall \alpha_{k_j} > 0. \end{aligned}$$

Applying the Razumikhin condition with  $p(s) = p_{k_j} s$ ,  $p_{k_j} > 1$ , we obtain

$$(3.17) \quad x_j^T(t - \tau_{k_j}) P_{k_j} x_j(t - \tau_{k_j}) \leq p_{k_j} x_j^T(t) P_{k_j} x_j(t)$$

for

$$V_{k_j}(t + \theta, x_j(t + \theta)) < p_{k_j} V_{k_j}(t, x_j(t)) \quad \forall \theta \in [-\tau_{k_j}, 0].$$

Let

$$(3.18) \quad S_{k_j} := -(A_{k_j}^T P_{k_j} + P_{k_j} A_{k_j} + p_{k_j} \alpha_{k_j} P_{k_j} + \alpha_{k_j}^{-1} P_{k_j} \bar{A}_{k_j} P_{k_j}^{-1} \bar{A}_{k_j}^T P_{k_j}).$$

We have

$$(3.19) \quad \dot{V}_{k_j}(t, x_j) \leq -x_j^T(t) S_{k_j} x_j(t).$$

Because  $\Sigma_t \in \tilde{\mathcal{A}}$ , we have  $S_{k_j} > 0$  from Lemma 2.1. Furthermore we can select  $w(s) = w_{k_j} s^2$  in Lemma 3.1, such that (3.2) is satisfied, where  $w_{k_j} := \sigma_{min}[S_{k_j}] > 0$ .

Define

$$(3.20) \quad \lambda := \max_{i \in \mathcal{F}} \frac{\bar{\kappa}_i}{\kappa_i}$$

and

$$(3.21) \quad \mu := \max_{i \in \mathcal{F}} \frac{\bar{\kappa}_i}{w_i}.$$

Now we are ready to state the main result.

**THEOREM 3.4.** *Let the dwell time be defined by  $\tau_D := T^* + \tau_{max}$ , where*

$$(3.22) \quad T^* := \lambda\mu \left\lfloor \frac{\lambda - 1}{\bar{p} - 1} + 1 \right\rfloor,$$

with  $\bar{p} := \min_{i \in \mathcal{F}} \{p_i\} > 1$ , and  $\lfloor \cdot \rfloor$  being the floor integer function. Then the system (2.2) with  $\Sigma_t \in \bar{\mathcal{A}}$  is asymptotically stable for any switching rule  $q(t) \in S[\tau_D]$ .

*Proof.* First we claim that  $\forall \tau > \tau_D$ , there exist  $0 < \beta < 1$  and  $0 < \alpha < 1$ , such that  $\tau \geq \bar{T} + \tau_{max}$ , where

$$(3.23) \quad \bar{T} := \frac{\lambda\mu}{\alpha^2} \left\lfloor \frac{\lambda - \alpha^2}{\alpha^2\beta(\bar{p} - 1)} \right\rfloor.$$

For a given  $\tau$ , to find such  $\alpha$  and  $\beta$  define  $\tilde{T} + \tau_{max} := \tau > \tau_D = T^* + \tau_{max}$ , and consider the two cases below.

- (1) If  $\lfloor (\lambda - 1)/(\bar{p} - 1) \rfloor =: k < (\lambda - 1)/(\bar{p} - 1) < k + 1$ , then we can find  $\Delta_1 > 0$  and  $\Delta_2 > 0$  small enough, such that

$$\left\lfloor \frac{\lambda - \alpha_1^2}{\alpha_1^2\beta(\bar{p} - 1)} \right\rfloor = \left\lfloor \frac{\lambda - 1}{\bar{p} - 1} \right\rfloor = k + 1 = \left\lfloor \frac{\lambda - 1}{\bar{p} - 1} + 1 \right\rfloor$$

with  $\alpha_1 = (1 + \Delta_1)^{-\frac{1}{2}} < 1$  and  $\beta = (1 + \Delta_2)^{-\frac{1}{2}} < 1$ . Let  $\tilde{T} = T^* + \epsilon$ ,  $\epsilon > 0$ . It is easy to check that

$$(3.24) \quad \frac{\lambda\mu}{\alpha_2^2} \left\lfloor \frac{\lambda - \alpha_1^2}{\alpha_1^2\beta(\bar{p} - 1)} \right\rfloor = \frac{\lambda\mu}{\alpha_2^2} (k + 1) \leq (k + 1)\lambda\mu + \epsilon = \tilde{T},$$

where  $0 < \alpha_2 = (1 + \Delta_3)^{-\frac{1}{2}} < 1$  with  $0 < \Delta_3 \leq \frac{\epsilon}{(k+1)\lambda\mu}$ . Now choosing  $0 < \alpha = \max\{\alpha_1, \alpha_2\} < 1$ , we have  $\bar{T} \leq \tilde{T}$ , which is straightforward from (3.23) and (3.24).

- (2) If  $(\lambda - 1)/(\bar{p} - 1) = k > 0$  is an integer, then we can similarly find  $0 < \alpha_1 < 1$  and  $0 < \beta < 1$  such that

$$\left\lfloor \frac{\lambda - \alpha_1^2}{\alpha_1^2\beta(\bar{p} - 1)} \right\rfloor = \left\lfloor \frac{\lambda - 1}{\bar{p} - 1} + 1 \right\rfloor = k + 1 = \left\lfloor \frac{\lambda - 1}{\bar{p} - 1} + 1 \right\rfloor.$$

In the same fashion as (1), we can constructively have  $0 < \alpha < 1$  and  $0 < \beta < 1$  such that  $\bar{T} \leq \tilde{T}$ .

This proves the first claim.

The second claim we make is that  $\|x_j(t)\| \leq \alpha\delta_j$  for any  $t \geq t_j + \bar{T}$ ,  $t \in [t_j, t_{j+1}]$ , where we assume  $|\phi_j(t)|_{[t_j - \tau_{k_j}, t_j]} \leq \delta_j$ . To show this fact, we can choose  $\delta_1 = \delta_j$ ,

$\bar{\delta}_2 = \bar{\delta}_1 \sqrt{\bar{\kappa}_{k_j}/\kappa_{k_j}} \geq \bar{\delta}_1$ , and select  $\eta = \alpha \bar{\delta}_1$  in Lemma 3.3. It is straightforward that  $0 < \eta < \bar{\delta}_1 \leq \bar{\delta}_2$ . Recalling (3.12) and (3.13), we have

$$(3.25) \quad V_{k_j}(t, x_j) \leq \kappa_{k_j} \eta^2 \quad \text{for } t \geq t_j + T,$$

where

$$(3.26) \quad \begin{aligned} T &= \frac{Nv(\bar{\delta}_1)}{\gamma} \frac{[(v(\bar{\delta}_1) - u(\eta))/a]v(\bar{\delta}_1)}{\inf_{v^{-1}(u(\eta)) \leq s \leq \bar{\delta}_2} w(s)} \\ &= \frac{\bar{\kappa}_{k_j}^2 [(v(\bar{\delta}_1) - u(\eta))/a]}{\alpha^2 w_{k_j} \kappa_{k_j}}. \end{aligned}$$

Combining (3.15) and (3.25) yields

$$(3.27) \quad \|x_j(t)\| \leq \alpha \delta_j \quad \text{for } t \geq t_j + T.$$

Now choosing  $a = \beta(p_{k_j} - 1)\kappa_{k_j}\eta^2$ , we have

$$(3.28) \quad T = \frac{\bar{\kappa}_{k_j}^2 \left[ \frac{(\bar{\kappa}_{k_j}/\kappa_{k_j}) - \alpha^2}{\alpha^2 \beta(p_{k_j} - 1)} \right]}{\alpha^2 w_{k_j} \kappa_{k_j}} \leq \bar{T}.$$

Therefore from (3.27) and (3.28) we have

$$(3.29) \quad |x_j|_{[t_j + \bar{T}, t_{j+1}]} \leq \alpha \delta_j,$$

as claimed.

Now recall that  $t_{j+1} - t_j > \tau_D$ . Therefore  $t_{j+1} - t_j \geq \bar{T} + \tau_{max} \geq \bar{T} + \tau_{k_{j+1}}$ . Also notice that  $\phi_{j+1}(t) = x_j(t)$ ,  $t \in [t_{j+1} - \tau_{k_{j+1}}, t_{j+1}]$ . We have

$$(3.30) \quad \begin{aligned} |\phi_{j+1}|_{[t_{j+1} - \tau_{k_{j+1}}, t_{j+1}]} &= |x_j|_{[t_{j+1} - \tau_{k_{j+1}}, t_{j+1}]} \\ &\leq |x_j|_{[t_j + \bar{T}, t_{j+1}]} \leq \alpha \delta_j := \delta_{j+1} \end{aligned}$$

and  $\delta_0$  is defined as  $\delta_0 := |\phi|_{[-\tau_{max}, 0]} \geq |\phi|_{[-\tau_{k_0}, 0]}$ . Therefore we obtain a convergent sequence  $\{\delta_i\}$ ,  $i = 0, 1, 2, \dots$ , where  $\delta_i = \alpha^i \delta_0$ .

Meanwhile, (3.7) implies

$$(3.31) \quad |x_j|_{[t - \tau_{k_j}, t]} \leq \sqrt{\frac{\bar{\kappa}_{k_j}}{\kappa_{k_j}}} |x_j|_{[t_j - \tau_{k_j}, t_j]} \quad \forall t \in [t_j, t_{j+1}].$$

Hence

$$(3.32) \quad \begin{aligned} &\sup_{t \in [t_j, t_{j+1}]} \|x_j(t)\| \\ &\leq \sup_{t \in [t_j, t_{j+1}]} |x_j|_{[t - \tau_{k_j}, t]} \leq \sqrt{\lambda} |x_j|_{[t_j - \tau_{k_j}, t_j]} \\ &\leq \sqrt{\lambda} \delta_j = \alpha^j \sqrt{\lambda} \delta_0, \end{aligned}$$

which implies the asymptotic stability of the switched time delay system  $\Sigma_t$  with the switching signal  $q(t) \in S_{[\tau_D]}$ .  $\square$

**3.2. The case with delay-dependent criterion.** In a similar fashion, we can investigate the stability of the switched time delay system  $\Sigma_t$  of (2.2) under the assumption that  $\Sigma_t \in \bar{\mathcal{A}}_d$ . Hence each candidate system  $\Sigma_i$ ,  $i \in \mathcal{F}$ , is delay-dependently asymptotically stable satisfying (2.7). We assume  $\tau_D^d > 2\tau_{max}$  in this scenario. Similar to the proof of Theorem 3.4, we consider an arbitrary switching interval  $[t_j, t_{j+1})$  of the piecewise switching signal  $q(t) \in S[\tau_D^d]$ , where the state variable  $x_j(t)$  defined on this interval obeys (2.3). The first order model transformation [7] of (2.3) results in

$$(3.33) \quad \begin{aligned} \dot{x}_j(t) = & (A_{k_j} + \bar{A}_{k_j})x_j(t) \\ & - \bar{A}_{k_j} \int_{-\tau_{k_j}}^0 [A_{k_j}x_j(t+\theta) + \bar{A}_{k_j}x(t+\theta-\tau_{k_j})]d\theta, \end{aligned}$$

where the initial condition  $\psi_j(t)$  is defined as  $\psi_j(t) = x_{j-1}(t)$ ,  $t \in [t_j - 2\tau_{k_j}, t_j]$ , for  $j \in \mathbb{Z}^+$ , and  $\psi_0(t)$  defined by

$$\psi_0(t) = \begin{cases} \phi(t), & t \in [-\tau_{max}, 0], \\ \phi(-\tau_{max}), & t \in [-2\tau_{max}, -\tau_{max}). \end{cases}$$

By using the Lyapunov–Razumikhin function (3.14), we obtain the time derivative of  $V_{k_j}(t, x_j(t))$  along the trajectory of (3.33),

$$\begin{aligned} \dot{V}_{k_j}(t, x_j) = & x_j^T(t)[P_{k_j}(A_{k_j} + \bar{A}_{k_j}) + (A_{k_j} + \bar{A}_{k_j})^T P_{k_j}]x_j(t) \\ & - \int_{-\tau_{k_j}}^0 [2x_j^T(t)P_{k_j}\bar{A}_{k_j}(A_{k_j}x_j(t+\theta) + \bar{A}_{k_j}x_j(t+\theta-\tau_{k_j}))]d\theta. \end{aligned}$$

Assume  $V_{k_j}(t+\theta, x_j(t+\theta)) < p(V_{k_j}(t, x_j(t))) \quad \forall \theta \in [-2\tau_{k_j}, 0]$ , where  $p(s) = p_{k_j}s$ ,  $p_{k_j} > 1$ . We have [3, 14]

$$(3.34) \quad \dot{V}_{k_j}(t, x_j) \leq -x_j^T(t)S_{k_j}^d x_j(t),$$

where

$$(3.35) \quad \begin{aligned} S_{k_j}^d := & - \left\{ P_{k_j}(A_{k_j} + \bar{A}_{k_j}) + (A_{k_j} + \bar{A}_{k_j})^T P_{k_j} \right. \\ & + \tau_{k_j} \left[ \alpha_{k_j}^{-1} P_{k_j} \bar{A}_{k_j} A_{k_j} P_{k_j}^{-1} \bar{A}_{k_j}^T A_{k_j}^T P_{k_j} \right. \\ & \left. \left. + \beta_i^{-1} P_{k_j} (\bar{A}_{k_j})^2 P_{k_j}^{-1} (\bar{A}_{k_j}^T)^2 P_{k_j} + p_{k_j}(\alpha_{k_j} + \beta_{k_j}) P_{k_j} \right] \right\}. \end{aligned}$$

Because  $\Sigma_t \in \tilde{\mathcal{A}}_d$ , we have  $S_{k_j}^d > 0$  from Lemma 2.2. Therefore we can select  $w(s) = w_{k_j}^d s^2$  in Lemma 3.1, such that (3.2) holds, where  $w_{k_j}^d := \sigma_{min}[S_{k_j}^d] > 0$ .

**THEOREM 3.5.** *Let the dwell time be  $\tau_D^d := T_d^* + 2\tau_{max}$ , where*

$$(3.36) \quad T_d^* := \lambda \mu_d \left[ \frac{\lambda - 1}{\bar{p} - 1} + 1 \right],$$

with

$$(3.37) \quad \mu_d := \max_{i \in \mathcal{F}} \frac{\bar{\kappa}_i}{w_i^d}$$

and the other parameters are the same as those defined in Theorem 3.4. Then system (2.2) with  $\Sigma_t \in \tilde{\mathcal{A}}_d$  is asymptotically stable for any switching rule  $q(t) \in S[\tau_D^d]$ .

*Proof.* We can apply arguments similar to those used in the proof of Theorem 3.4 to obtain the following inequality:

$$(3.38) \quad \sup_{t \in [t_j, t_{j+1}]} \|x_j(t)\| \leq \sqrt{\lambda} \delta_j^d,$$

where  $|\psi_j(t)|_{[t_j - 2\tau_{k_j}, t_j]} \leq \delta_j^d$  and  $\delta_{j+1}^d = \alpha \delta_j^d$ . Note that  $\delta_0^d$  can be selected as

$$\delta_0^d := |\psi|_{[-2\tau_{max}, 0]} = |\phi|_{[-\tau_{max}, 0]} = \delta_0.$$

It is clear that  $|\psi|_{[-2\tau_{k_0}, 0]} \leq \delta_0^d$ , which further implies  $\delta_j^d = \delta_j$ ,  $j \in \mathbb{Z}^+ \cup \{0\}$ . The upper bound of the state variable  $x(t)$  of the switched time delay systems  $\Sigma_t$  is bounded by a decreasing sequence  $\{\delta_i\}$ ,  $i = 0, 1, 2, \dots$ , converging to zero, which implies the asymptotic stability and proves this theorem.  $\square$

The dwell-time-based stability analysis proposed in this paper is general in the sense that it can be used for other stability results based on Razumikhin theorems as long as the correspondingly Lyapunov functions are in quadratic forms. Particularly, Theorem 3.5 can be extended easily to the case where  $\Sigma_t$  has time-varying time delays and parameter uncertainties, which has important applications such as TCP (transmission control protocol) congestion control of computer networks [13, 26].

*Remark 3.6.* Note that the Lyapunov–Krasovskii method has been used to analyze the stability of time delay systems, with which some less conservative stability conditions have been provided [5]. However, it is difficult to employ piecewise Lyapunov–Krasovskii functionals for dwell-time-based analysis similar to Theorems 3.4 and 3.5. Recall the general form of Lyapunov–Krasovskii functional  $V(t, x_t)$  [20] for delay system (2.1), such that

$$u(\|x(t)\|) \leq V(t, x_t) \leq v(|x|_{[t-\tau, t]}), \quad t \in \mathbb{R}, \quad x \in \mathbb{R}^n,$$

and

$$\dot{V}(t, x(t)) \leq -w(\|x(t)\|).$$

The upper bound of  $V(t, x_t)$  is dependent on the  $\infty$ -norm of the trajectory, while other bounds on  $V(t, x_t)$  and  $\dot{V}(t, x_t)$  are on the Euclidean norm of the trajectory, which poses the technical challenge of estimating the trajectory bound and decaying rate for the switched delay systems (2.2).

**4. Conservatism analysis.** The dwell-time-based stability results had been obtained for switched linear systems free of delays [10, 19]. It is interesting to compare the conservatism of the results presented in this paper with those for delay-free systems.

In fact, one extreme case of the switched system  $\Sigma_t$  is  $\tau_i = 0$  and  $\bar{A}_i = 0$  for  $i \in \mathcal{A}$ , which corresponds to the delay-free scenario. For each candidate system  $\dot{x} = A_i x$ , a sufficient and necessary condition to guarantee asymptotic stability is  $\exists P_i = P_i^T > 0$ , such that  $Q_i := -(A_i^T P_i + P_i A_i) > 0$ . Correspondingly a dwell-time-based stability for such a switched delay-free system is  $q(t) \in S[\tilde{\tau}_D]$ , where

$$(4.1) \quad \tilde{\tau}_D = \tilde{\mu} \ln \lambda,$$

where  $\lambda$  is defined by (3.20) and

$$(4.2) \quad \tilde{\mu} := \max_{i \in \mathcal{F}} \frac{\bar{\kappa}_i}{\tilde{w}_i},$$

where  $\tilde{w}_i := \sigma_{\min}[Q_i] > 0$ .

On the other hand, in our case for  $\tau_i = 0$  and  $\bar{A}_i = 0$ , we observe that

$$(4.3) \quad \lim_{\alpha_i \rightarrow 0^+} S_i = \lim_{\alpha_i, \beta_i \rightarrow 0^+} S_i^d = Q_i, \quad i \in \mathcal{F},$$

from (3.18) and (3.35), which indicates  $\mu = \mu_d = \tilde{\mu}$  by (3.21), (3.37), and (4.2). Accordingly we can select  $p_i > 1$ ,  $i \in \mathcal{F}$ , sufficiently large such that  $\lfloor \frac{\lambda-1}{p-1} + 1 \rfloor = 1$  in (3.22) and (3.36) and obtain

$$(4.4) \quad \tau_D = T^* = \lambda\mu = \lambda\mu_d = T_d^* = \tau_D^d.$$

Therefore

$$(4.5) \quad \tau_D = \tau_D^d = \lambda\tilde{\mu} > \tilde{\mu} \ln \lambda = \tilde{\tau}_D.$$

The dwell times derived for switched time delay systems are proportional to  $\lambda$ , in contrast to the logarithm of  $\lambda$  for switched delay-free systems. This gap is due to the fact that asymptotic stability for linear delay-free systems implies exponential stability. However, for time delay systems, the sufficient stability conditions based on the Lyapunov–Razumikhin theorem do not guarantee exponential stability. As a matter of fact, the exponential estimates for time delay systems require additional assumptions besides asymptotic stability [15].

It should be noted that stability conditions for switched time delay systems are also considered in [22, 23], where the authors give a sufficient condition to guarantee *uniform* stability (see Theorem 6.1 of [22] for notation and details):  $\Gamma e^{L(\Lambda+h)} \leq 1$ . Apparently, this condition does not hold for the switched system (2.2) because in our case  $\Gamma = 1$ , and hence

$$\Gamma e^{L(\Lambda+h)} = e^{L(\Lambda+h)} > 1 \quad \forall \Lambda > 0, L > 0, h > 0.$$

**5. Numerical example.** In this section, we use an illustrative example to demonstrate the results in section 3.

*Example.* Consider the following switched time delay system with 2 candidates:

$$(5.1) \quad \Sigma_t : \begin{cases} \dot{x} = A_{q(t)}x(t) + \bar{A}_{q(t)}x(t - \tau_{q(t)}), & t \geq 0, \\ x(t) = \phi(t) & \forall t \in [-\tau_{\max}, 0], \\ q(t) \in \{1, 2\}, \end{cases}$$

where the switching candidate systems  $\Sigma_1 := (A_1, \bar{A}_1, \tau_1)$  and  $\Sigma_2 := (A_2, \bar{A}_2, \tau_2)$  are determined by

$$A_1 = \begin{bmatrix} -2 & 0 \\ 0 & -0.9 \end{bmatrix}, \quad \bar{A}_1 = \begin{bmatrix} -1 & 0 \\ -0.5 & -1 \end{bmatrix};$$

$$A_2 = \begin{bmatrix} -1 & 0.5 \\ 0 & -1 \end{bmatrix}, \quad \bar{A}_2 = \begin{bmatrix} -1 & 0 \\ 0.1 & -1 \end{bmatrix};$$



TABLE 5.1

Parameters calculated with respect to switched time delay system  $\Sigma_t$  and switched delay-free system  $\bar{\Sigma}_t$ .

Parameters	$\Sigma_t$ (with delay)	$\bar{\Sigma}_t$ (delay free)
$\lambda$	1.7224	1.7224
$\mu_d$	1.5216	N/A
$\bar{\mu}_d$	N/A	0.7056
$\bar{p}$	1.4	N/A
$T_d^*$	5.3147	N/A
Dwell time	$\tau_D^d = 6.5147$	$\tilde{\tau}_D = 0.3836$

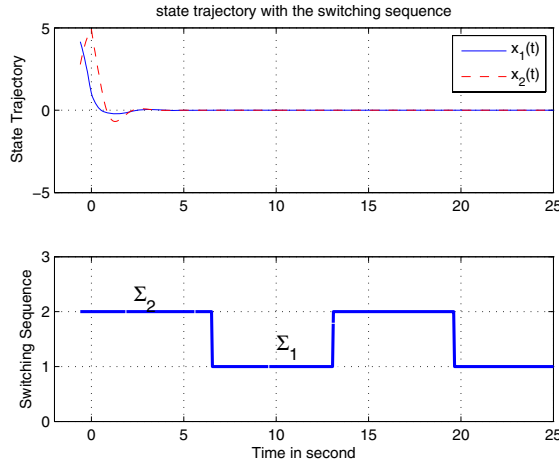


FIG. 5.1. The state trajectory of  $\Sigma_t$  in the presence of switching.

and  $\tau_1 = 0.3, \tau_2 = 0.6$ . The initial condition of (5.1) is chosen as

$$\phi(t) = \begin{bmatrix} 5 \cos\left(\frac{\pi}{2.4}t + \frac{\pi}{6}\right) \\ 5 \sin\left(\frac{\pi}{2.4}t + \frac{\pi}{6}\right) \end{bmatrix} \quad \forall t \in [-0.6, 0].$$

It is clear that  $\Sigma_1$  and  $\Sigma_2$  are delay-dependently stable, which can be verified by Lemma 2.2. Applying Theorem 3.5 gives the dwell time  $\tau_D^d = 6.52$ , which guarantees the asymptotic stability of the switched time delay system (5.1). For the purpose of comparison, we also calculate the dwell time  $\tilde{\tau}_D$  of the delay-free system  $\bar{\Sigma}_t : \dot{x} = A_{q(t)}x(t), q(t) \in \{1, 2\}$ . The results are shown in Table 5.1.

The switched time delay system  $\Sigma_t$  described by (5.1) is simulated in MATLAB, where we start with  $\Sigma_2$  and perform switching every  $\tau_D^d$  seconds. The state trajectory is depicted in Figure 5.1, where we clearly see the asymptotic convergence in the presence of switching. Also, we provide the phase portrait in Figure 5.2 with respect to  $x_1(t)$  and  $x_2(t)$ , which better illustrates the switching and the stability of the switched system.

It is also interesting to investigate the relation between time delays of (2.2) and the corresponding dwell time  $\tau_D^d$ . For this purpose, we took  $\tau_1 = \tau_2 = \tau$  in (5.1) with  $\tau$  varying from 0.1 to 0.7. The results are shown in Table 5.2. We should also indicate that the free parameters  $\alpha_i, \beta_i$ , and  $p_i$  can be further optimized to reduce the values of  $\tau_D^d$  given in the table. However, this is an open problem deserving a

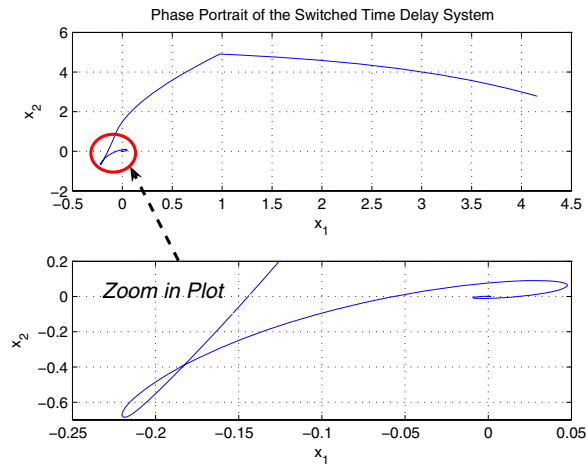


FIG. 5.2. Phase portrait of the switched time delay system  $\Sigma_t$ .

TABLE 5.2  
Dwell time values versus time delays of  $\Sigma_t$ .

$\tau$	0.1	0.2	0.3	0.4	0.5	0.6	0.7
$\tau_D^d$	0.93	1.49	3.36	4.83	9.14	106.23	950.58

separate study. Nevertheless, the results given in the table suggest an exponentially increasing behavior of  $\tau_D^d$  with the delay. Similar behavior is observed for the  $\mathcal{H}^\infty$  optimal cost in weighted sensitivity minimization for systems with delays [4].

**6. Concluding remarks.** We provided stability analysis for switched linear systems with time delays, where each candidate system is assumed to be delay-independently or delay-dependently asymptotically stable. We showed the existence of a dwell time of the switching signal, such that the switched time delay system is asymptotically stable independent of the trajectory. The dwell time values for both scenarios are constructively given. The results are compared with the dwell-time conditions for switched delay-free systems. Optimization of the minimum dwell times that we have derived, in terms of the free parameters appearing in the LMI conditions, is an interesting open problem. An interesting extension of this work is to investigate stability and controller synthesis for switched interval time delay systems, which will potentially offer a hybrid control method for large time delay systems and time varying delay systems.

REFERENCES

[1] C. BETT AND M. LEMMON, *Bounded amplitude performance of switched LPV systems with applications to hybrid systems*, Automatica, 35 (1999), pp. 491–503.  
[2] C. DE PERSIS, R. DE SANTIS, AND S. MORSE, *Supervisory control with state-dependent dwell-time logic and constraints*, Automatica, 40 (2004), pp. 269–275.  
[3] L. DUGARD AND E. I. VERRIEST, EDS., *Stability and Control of Time-Delay Systems*, Springer-Verlag, London, New York, 1998.  
[4] C. FOIAS, H. ÖZBAY, AND A. TANNENBAUM, *Robust Control of Infinite Dimensional Systems: Frequency Domain Methods*, Lecture Notes in Control and Inform. Sci. 209, Springer-Verlag, London, 1996.

- [5] K. GU AND S.-I. NICULESCU, *Survey on recent results in the stability and control of time-delay systems*, ASME J. Dynamic Systems, Measurement, and Control, 125 (2003), pp. 158–165.
- [6] K. GU, V. L. KHARITONOV, AND J. CHEN, *Stability and Robust Stability of Time-Delay Systems*, Birkhäuser, Boston, 2003.
- [7] J. HALE AND S. VERDUYN LUNEL, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.
- [8] J. HESPAÑHA, *Uniform stability of switched linear systems: Extension of LaSalle's invariance principle*, IEEE Trans. Automat. Control, 49 (2004), pp. 470–482.
- [9] J. HESPAÑHA, D. LIBERZON, AND S. MORSE, *Overcoming the limitations of adaptive control by means of logic-based switching*, Systems Control Lett., 49 (2003), pp. 49–65.
- [10] J. HESPAÑHA AND S. MORSE, *Stability of switched systems with average dwell-time*, in Proceedings of the 38th Annual Conference on Decision and Control, Phoenix, AZ, 1999, pp. 2655–2660.
- [11] J. HOCHCERMAN-FROMMER, S. KULKARNI, AND P. RAMADGE, *Controller switching based on output prediction errors*, IEEE Trans. Automat. Control, 43 (1998), pp. 596–607.
- [12] C. KAO AND B. LINCOLN, *Simple stability criteria for systems with time-varying delays*, Automatica, 40 (2004), pp. 1429–1434.
- [13] F. KELLY, *Mathematical modelling of the Internet*, in Mathematics Unlimited—2001 and Beyond, B. Engquist and W. Schmid, eds., Springer-Verlag, Berlin, 2001, pp. 685–702.
- [14] V. L. KHARITONOV, *Robust stability analysis of time delay systems: A survey*, Ann. Rev. Control, 23 (1999), pp. 185–196.
- [15] V. L. KHARITONOV AND D. HINRICHSSEN, *Exponential estimates for time delay systems*, Systems Control Lett., 53 (2004), pp. 395–405.
- [16] V. KULKARNI, M. JUN, AND J. HESPAÑHA, *Piecewise quadratic Lyapunov functions for piecewise affine time delay systems*, in Proceedings of the American Control Conference, Boston, MA, 2004, pp. 3885–3889.
- [17] D. LIBERZON AND S. MORSE, *Basic problems in stability and design of switched systems*, IEEE Control Systems Mag., 19 (1999), pp. 59–70.
- [18] V. MONTAGNER, V. LEITE, S. TARBOURIECH, AND P. PERES, *Stability and stabilizability of discrete-time switched linear systems with state delay*, in Proceedings of the American Control Conference, Portland, OR, 2005, pp. 3806–3811.
- [19] S. MORSE, *Supervisory control of families of linear set-point controllers, part 1: Exact matching*, IEEE Trans. Automat. Control, 41 (1996), pp. 1413–1431.
- [20] S.-I. NICULESCU, *Delay Effects on Stability: A Robust Control Approach*, Lecture Notes in Control and Inform. Sci. 269, Springer-Verlag, Heidelberg, 2001.
- [21] E. SKAFIDAS, R. EVANS, A. SAVKIN, AND I. PETERSON, *Stability results for switched controller systems*, Automatica, 35 (1999), pp. 553–564.
- [22] Y. SUN, A. MICHEL, AND G. ZHAI, *Stability of discontinuous retarded functional differential equations with applications to delay systems*, in Proceedings of the American Control Conference, Denver, CO, 2003, pp. 3387–3392.
- [23] Y. SUN, A. MICHEL, AND G. ZHAI, *Stability of discontinuous retarded functional differential equations with applications*, IEEE Trans. Automat. Control, 50 (2005), pp. 1090–1105.
- [24] F. WU AND K. GRIGORIADIS, *LPV systems with parameter-varying time delays: Analysis and control*, Automatica, 37 (2001), pp. 221–229.
- [25] P. YAN AND H. ÖZBAY, *On switching  $\mathcal{H}^\infty$  controllers for a class of linear parameter varying systems*, Systems Control Lett., 56 (2007), pp. 504–511.
- [26] P. YAN AND H. ÖZBAY, *Robust controller design for AQM and  $\mathcal{H}^\infty$  performance analysis*, in Advances in Communication Control Networks, S. Tarbouriech, C. Abdallah, and J. Chiasson, eds., Lecture Notes in Control and Inform. Sci. 308, Springer-Verlag, New York, 2005, pp. 49–64.
- [27] D. YUE AND Q. HAN, *Delay-dependent exponential stability of stochastic systems with time-varying delay, nonlinearity, and Markovian switching*, IEEE Trans. Automat. Control, 50 (2005), pp. 217–222.
- [28] G. ZHAI, Y. SUN, X. CHEN, AND A. MICHEL, *Stability and  $\mathcal{L}^2$  gain analysis for switched symmetric systems with time delay*, in Proceedings of the American Control Conference, Denver, CO, 2003, pp. 2682–2687.

## OUTPUT FEEDBACK $H_\infty$ CONTROL OF CONTINUOUS-TIME INFINITE MARKOVIAN JUMP LINEAR SYSTEMS VIA LMI METHODS\*

MARCOS G. TODOROV<sup>†</sup> AND MARCELO D. FRAGOSO<sup>†</sup>

**Abstract.** The output feedback  $H_\infty$  control is addressed for a class of continuous-time Markov jump linear systems with the Markov process taking values in an infinite countable set  $\mathcal{S}$ . We consider that only an output and the jump parameters are available to the controller. Via a certain bounded real lemma, together with some extensions of Schur complements and of the projection lemma, a theorem which characterizes whether there exists a full-order solution to the disturbance attenuation problem is devised in terms of two different linear matrix inequality (LMI) feasibility problems. This result connects a certain projection approach to an LMI problem which is more amenable to computer solution, and hence for design. We conclude the paper with some design algorithms for the construction of such controllers and an illustrative example.

**Key words.** continuous-time linear systems, Markov jump parameter, bounded real lemma, linear matrix inequalities, output feedback,  $H_\infty$  control

**AMS subject classifications.** 93C05, 93E03, 93B36, 60J27, 60J75

**DOI.** 10.1137/060675162

**1. Introduction.** It is a well-known fact that in order to treat adequately problems related to a wide class of dynamical systems, we do need to characterize adequately the uncertainties in the mathematical description of these systems, which can be, for instance, of an environmental and/or modeling nature. These uncertainties have many sources: noise in communications systems, atmospheric fluctuation, volatility in the economic scenario, *failures* (abrupt change in the system structure), parametric uncertainty, etc. In this paper we shall be interested in a class of linear dynamical systems which are subjected to uncertainty (change) in their structures as a consequence of abrupt phenomena. The uncertainties are characterized in the model via a Markov process. These systems are known in the literature as Markov jump linear systems (MJLS) and constitute an important class of hybrid systems.

MJLS provide a powerful framework for modeling and analyzing systems which are subject to abrupt changes in their structure [6]. They are seen as a real alternative to reaching failure tolerance behavior in control systems, and consequently are highly relevant for control theory; i.e., MJLS belong to what is known in the specialized literature as safety-critical and high-integrity systems (e.g., aircraft, chemical plants, nuclear power station, robotic manipulator systems, large scale flexible structures for space stations such as antenna, solar arrays, etc.).

This class of systems (MJLS) has been the subject of intensive research over the last few decades and the associated literature is now fairly extensive. The last 25 years have seen significant development in the nature and scope of this subject. This

---

\*Received by the editors November 16, 2006; accepted for publication (in revised form) November 14, 2007; published electronically February 29, 2008. This work was supported in part by the Research Council of the State of Rio de Janeiro-FAPERJ under grant 151.899/05, by the Brazilian National Research Council-CNPq under grants 141363/2007-0, 470527/2007-2, and 302587/2004-7, by the Research Council of the State of São Paulo-FAPESP under grant 03/06736-7, and by FAPERJ under grant E-26/100.579/2007.

<http://www.siam.org/journals/sicon/47-2/67516.html>

<sup>†</sup>National Laboratory for Scientific Computing—LNCC/CNPq, Av. Getúlio Vargas 333, Petrópolis, Rio de Janeiro, CEP 25651-070, Brazil (mtodorov@lncc.br, frag@lncc.br).

period has witnessed exciting development in creating new directions of study and understanding a great deal of subtleties in this field (see, e.g., [6] and [12] for an up-to-date discussion on this subject). While the study of MJLS is mature in many ways, a great variety of interesting problems of this field are still wide open. This is the case, for instance, with the output feedback  $H_\infty$  control problem for the class of continuous-time infinite MJLS (including many aspects of the finite case). *Infinite*, or *finite*, here refers to the state space of the Markov process.

In this paper we address the output feedback  $H_\infty$  control for the case in which the state space of the Markov chain is *infinite countable*. We consider that only an output and the jump parameters are available to the controller. A theorem which characterizes whether there exists a full-order solution to the disturbance attenuation problem is devised in terms of two different linear matrix inequality (LMI) feasibility problems. Our result establishes the connection between a certain projection approach and an LMI problem which is more suitable for design. A certain infinite Markov jump bounded real lemma (JBRL) is employed for the first time, illustrating the importance of this recent result (see [21]). In addition, we provide some *design algorithms* leading to a series of tools that, among other things, may be readily applied to the finite setting. This includes an LMI algorithm, which allows one to check, by two different ways, whether there exists a solution for the disturbance attenuation (DA) problem (Algorithm 4.10); a two-step design method (Algorithm 5.3) which provides explicit formulas for the construction of a suboptimal controller; and an alternative three-step design method (Algorithm 5.4). The main difference between Algorithms 5.3 and 5.4 is that, in the latter, one can first check if a smaller (projected) LMI problem is feasible by means of Algorithm 4.10. This amounts to verifying whether or not the DA problem has a solution, before proceeding further. Finally, in order to provide a rather complete design framework and hence bring our results to a more applicable basis, the  $H_\infty$  optimization problem is also treated from this viewpoint. A simple bisectional algorithm (Algorithm 5.5) is employed, in a parallel to [2] and [6, Algorithm 8.9], yielding a tractable way of computing the optimal  $H_\infty$  performance that may be achieved with our results.

An important issue here is that some fundamental tools such as Schur complements and the projection lemma cannot be directly applied in the *infinity* setting. Since, to the best of the authors' knowledge, this is the first time the LMI approach is employed for the infinite case, we had to solve this problem by devising extended versions of these results. We notice that, as the breadth of applicability of these tools is not restricted to the jump setting, they represent potentially powerful tools for handling a variety of LMI problems (for instance, in time-varying systems [9, 20]). Furthermore, it seems that these are new results not previously found in the literature (for a discussion on some differences in the treatment of the infinite case, readers are referred, e.g., to [11], [10], and [12]).

When reduced to the *finite case*, a closely related paper is [7]. One important aspect which distinguishes our work from [7] is that the starting point here is a novel bounded real lemma, recently devised in [21], which allows us to ignore the bounds of sufficiency conditions as used in [7]. The bounded real lemma enables us to stray into the more general realm of the  $H_\infty$  problem, providing not only *necessary* and sufficient conditions at each step, but also allowing us to decouple the existence problem from the design. We should perhaps mention here that no *regularity assumptions* are made, as, for example, in [8] and [17].

It is perhaps worth noting here that, besides being interesting in their own right,

matrices with complex coefficients are also interesting from an application point of view. As pointed out in [16] and [1], in several engineering applications, such as communication application of signals systems, whirling shafts, and vibrational systems, complex coefficients come into play. In [16] some systems with complex coefficients that naturally arise in mechanics and signal processing are presented, motivating the study of equations with complex coefficients and illustrating more general situations as in, for instance, satellite and cosmic vehicles control. We refer the reader to these papers and the references therein for further examples and results on systems with complex coefficients.

Finally, it is too early to predict the full extent to which the theory developed here, for the case of a countably infinite state space of the Markov chain, will be applied. However, a few situations where model (3.2) seems to be naturally applied are mentioned in the introduction of [10] (see also [5]).

The paper is organized as follows. In section 2 we provide the bare essentials of notation and some auxiliary results. Section 3 introduces the basic model together with the bounded real lemma, which will be important in section 4, where the  $H_\infty$  problem is dealt with. Some tools for the design of *full-order*  $H_\infty$  compensators are given in section 5, together with a nominal example. The paper is concluded in section 6 with a highlight of the main contributions. Some proofs and auxiliary results are presented in the appendix.

**2. Notation and auxiliary results.** Let  $\|\cdot\|$  denote the euclidean norm in the complex  $n$ -space  $\mathbb{C}^n$ . We define  $\mathbb{M}(\mathbb{C}^m, \mathbb{C}^n)$  as the Banach space of all complex matrices  $M \in \mathbb{C}^{n \times m}$ , equipped with the standard induced matrix norm, also denoted by  $\|\cdot\|$ . Let us also define the infinite-dimensional Banach space  $\mathbb{H}_{\text{sup}}^{m,n}$  of all matrices of the form  $H = (H_1, H_2, \dots)$ , where  $H_i \in \mathbb{M}(\mathbb{C}^m, \mathbb{C}^n)$  for every  $i \in \mathcal{S} := \{1, 2, \dots\}$ , such that  $\|H\|_{\text{sup}} := \sup_{i \in \mathcal{S}} \|H_i\| < \infty$ . We also write  $\mathbb{H}_{\text{sup}}^n$  in place of  $\mathbb{H}_{\text{sup}}^{n,n}$  and define  $\mathbb{H}_{\text{sup}}^{n*}$  as the subset of  $\mathbb{H}_{\text{sup}}^n$  whose elements  $H = (H_1, H_2, \dots)$  exhibit the additional property that  $H_i = H_i^*$  for all  $i \in \mathcal{S}$  ( $H = H^*$  for short), with  $*$  denoting the conjugate transpose (we denote plain transposition by  $'$ ). Next, we define  $\tilde{\mathbb{H}}_{\text{sup}}^{n+}$  as the set composed by all *uniformly positive* matrices  $H \gg 0$ , i.e., such that  $H = (H_1, H_2, \dots) \in \mathbb{H}_{\text{sup}}^{n*}$  and  $H_i \geq \varepsilon I_n$  for all  $i \in \mathcal{S}$  and some  $\varepsilon > 0$  independent of  $i$  (here  $I_n$  stands for the  $n \times n$  identity matrix). Accordingly, we say that  $L \in \tilde{\mathbb{H}}_{\text{sup}}^{n-}$  (or is *uniformly negative*,  $L \ll 0$ ) whenever  $-L \gg 0$ . For short, we write that such  $H_i \gg 0$  and  $L_i \ll 0$  for all  $i \in \mathcal{S}$ . Finally, given  $R = (R_1, R_2, \dots) \in \mathbb{H}_{\text{sup}}^{n*}$  and  $S = (S_1, S_2, \dots) \in \mathbb{H}_{\text{sup}}^{n,m}$ , we shall write that  $R_i \gg 0$  over  $\mathcal{N}(S_i)$  whenever there exist  $\varepsilon > 0$  such that  $R_i \geq \varepsilon I_n$  over  $\mathcal{N}(S_i)$  for all  $i \in \mathcal{S}$ , where  $\mathcal{N}(\cdot)$  stands for the null space associated to a given complex matrix. (Accordingly,  $\mathcal{R}(\cdot)$  is the range of complex matrices. We refer to Proposition 2.3, in which an extension of  $\mathcal{N}(\cdot)$  and  $\mathcal{R}(\cdot)$  to the infinite case is stated.)

For  $H \in \mathbb{H}_{\text{sup}}^{p,m}$  and  $L \in \mathbb{H}_{\text{sup}}^{n,p}$  we have, in a natural way, that  $\|HL\|_{\text{sup}} \leq \|H\|_{\text{sup}} \|L\|_{\text{sup}}$  and thus  $HL := (H_1 L_1, H_2 L_2, \dots) \in \mathbb{H}_{\text{sup}}^{n,m}$ . Moreover, given  $F \in \mathbb{H}_{\text{sup}}^{\ell,m}$  we have that  $[H \ F] := ([H_1 \ F_1], [H_2 \ F_2], \dots) \in \mathbb{H}_{\text{sup}}^{(p+\ell),m}$  (the analogous holding for vertical or diagonal block concatenation). We denote by  $0_{\ell \times m}$  the zero matrix in either  $\mathbb{C}^{\ell,m}$  or  $\mathbb{H}_{\text{sup}}^{m,\ell}$ , the same holding for the identity matrices  $I_\ell \in \mathbb{C}^{\ell \times \ell}$ ,  $I_\ell \in \tilde{\mathbb{H}}_{\text{sup}}^{\ell+}$ . Whenever the size of any of those matrices has no importance or may be easily deduced by the context, it will be omitted. In addition, we define  $\text{Her}(H) := H + H^*$ ,  $\mathfrak{C}(H, L) := L^* H L$ , and sometimes represent off-diagonal blocks of a given self-adjoint matrix (that is, a matrix in a set such as  $\mathbb{H}_{\text{sup}}^{n*}$ ) by  $*$ , while entries with absolutely

no importance are denoted by  $\star$ . Furthermore, the Kronecker product of complex matrices is denoted by  $\otimes$ , in the usual way (see [4]).

With respect to the infinite countable set  $\mathcal{S}$  consider two infinite sequences  $\mathbf{m} = (m_1, m_2, \dots)$ ,  $\mathbf{n} = (n_1, n_2, \dots)$ , where  $(m_i, n_i) \in \{1, 2, \dots, M\}^2$  for some finite integer  $M$  and every  $i \in \mathcal{S}$ . Then define the infinite-dimensional Banach space  $\mathbb{H}_{\text{sup}}^{\mathbf{m}, \mathbf{n}}$  of all *bounded matrices*, which are objects of the form  $H = (H_1, H_2, \dots)$ , such that  $H_i \in \mathbb{M}(\mathbb{C}^{m_i}, \mathbb{C}^{n_i})$  for every  $i \in \mathcal{S}$ , where  $\|H\|_{\text{sup}} := \sup_{i \in \mathcal{S}} \|H_i\| < \infty$ . Let us also define the set  $\tilde{\mathbb{H}}_{\text{sup}}^{n+} \subset \mathbb{H}_{\text{sup}}^{\mathbf{n}, \mathbf{n}}$  analogously as before, where  $H \gg 0$  once more stands for uniform positivity,  $H \in \tilde{\mathbb{H}}_{\text{sup}}^{n+}$ . In particular, whenever  $\mathbf{m} = (m, m, \dots)$  and  $\mathbf{n} = (n, n, \dots)$  we simply have that  $\mathbb{H}_{\text{sup}}^{\mathbf{m}, \mathbf{n}} \equiv \mathbb{H}_{\text{sup}}^{m, n} \equiv \mathbb{H}_{\text{sup}}^{\mathbf{m}, n} \equiv \mathbb{H}_{\text{sup}}^{m, n}$ .

Concerning the random objects, fix a complete probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  carrying a right-continuous filtration  $\mathcal{F}_t \subset \mathcal{F}$  on  $t \in \mathbb{R}_+ := [0, \infty)$ . In addition, let  $E(\cdot)$  denote the usual mathematical expectation and define  $L_2^n$  as the space of all second order random variables  $(\Omega, \mathcal{F}) \mapsto \mathbb{C}^n$ . We also define the Lebesgue space  $L_2^n(\mathbb{R}_+)$  of all stochastic processes  $y = \{(y(t), \mathcal{F}_t); t \in \mathbb{R}_+, y(\cdot) \in \mathbb{C}^n\}$  such that  $\|y\|_{\mathbb{R}_+} := (\int_0^\infty E[\|y(t)\|^2] dt)^{1/2}$  is finite.

**2.1. Some auxiliary results.** This subsection contains some independent tools which will be necessary later on, including *extended versions* of Schur complements and of the projection lemma. The reader is referred to the appendix for the proofs.

The following proposition states that if uniform positive definiteness of the identity element in  $\mathbb{H}_{\text{sup}}^{n+}$  is preserved under application of a congruence transformation, then, in particular, both this entire set and its negative counterpart are invariant by such operation. This easy-to-check condition will be employed throughout the paper.

**PROPOSITION 2.1.** *Suppose  $Q = (Q_1, Q_2, \dots) \in \mathbb{H}_{\text{sup}}^{m, n}$  is such that  $Q^*Q \gg 0$ . Then the following statements are true:*

- (i)  $X \in \tilde{\mathbb{H}}_{\text{sup}}^{n+}$  implies  $Q^*XQ \gg 0$ ;
- (ii)  $Y \in \tilde{\mathbb{H}}_{\text{sup}}^{n-}$  implies  $Q^*YQ \ll 0$ .

*Proof.* See the appendix.  $\square$

The following result extends the well-known result on Schur complements to our setting.

**THEOREM 2.2** (uniform Schur complements). *Given  $U = (U_1, U_2, \dots) \in \mathbb{H}_{\text{sup}}^{p, p}$ ,  $V = (V_1, V_2, \dots) \in \mathbb{H}_{\text{sup}}^{q, p}$ , and  $W = (W_1, W_2, \dots) \in \mathbb{H}_{\text{sup}}^{q, q}$ , the following are equivalent:*

- (i)  $\begin{bmatrix} U & V \\ V^* & W \end{bmatrix} \gg 0$ ;
- (ii)  $U \gg 0$  and  $W - V^*U^{-1}V \gg 0$ ;
- (iii)  $W \gg 0$  and  $U - VW^{-1}V^* \gg 0$ .

Moreover, the same holds replacing  $\gg$  by  $\ll$ .

*Proof.* See the appendix.  $\square$

The following proposition introduces the notion of range and null spaces to the infinite case, besides defining what we will refer to later as an “orthonormal basis.”

**PROPOSITION 2.3.** *Suppose  $\Psi = (\Psi_1, \Psi_2, \dots) \in \mathbb{H}_{\text{sup}}^{p, \ell}$  is such that  $1 \leq p_i := \dim(\mathcal{N}(\Psi_i))$  for any  $i \in \mathcal{S}$ . Then, defining  $\mathbf{p} = (p_1, p_2, \dots)$ , there exists  $\Phi = (\Phi_1, \Phi_2, \dots) \in \mathbb{H}_{\text{sup}}^{\mathbf{p}, \mathbf{p}}$  such that, for all  $i \in \mathcal{S}$ ,*

- (i)  $\mathcal{N}(\Phi_i) = \{0\}$  and  $\mathcal{R}(\Phi_i) = \mathcal{N}(\Psi_i)$ ;
- (ii)  $\Phi_i^* \Phi_i = I_{p_i}$ .

Moreover, we use the notation  $\mathcal{N}(\Psi) = \mathcal{R}(\Phi)$  and say that such  $\Phi$  is an orthonormal basis for these spaces.

*Proof.* See the appendix.  $\square$

The following result is an immediate consequence of the two previous results.

COROLLARY 2.4. Suppose  $\Psi = (\Psi_1, \Psi_2, \dots) \in \mathbb{H}_{\text{sup}}^{p,\ell}$  is such that  $\dim(\mathcal{N}(\Psi_i)) \geq 1$  for any  $i \in \mathcal{S}$ . Then the following are equivalent for any  $U = (U_1, U_2, \dots) \in \mathbb{H}_{\text{sup}}^{p,*}$ ,  $V = (V_1, V_2, \dots) \in \mathbb{H}_{\text{sup}}^{q,p}$ , and  $W = (W_1, W_2, \dots) \in \mathbb{H}_{\text{sup}}^{q,*}$ :

- (i)  $\begin{bmatrix} U & V \\ V^* & W \end{bmatrix} \gg 0$  over  $\mathcal{N}([\Psi \ 0])$ ;
- (ii)  $U - VW^{-1}V^* \gg 0$  over  $\mathcal{N}(\Psi)$  and  $W \gg 0$ .

*Proof.* See the appendix.  $\square$

The following lemma is an extension from the one depicted in [15] and is of major importance in what follows.

LEMMA 2.5 (uniform projection lemma). Assume  $N = (N_1, N_2, \dots) \in \mathbb{H}_{\text{sup}}^{p,q}$ ,  $M = (M_1, M_2, \dots) \in \mathbb{H}_{\text{sup}}^{p,r}$ , and  $H = (H_1, H_2, \dots) \in \mathbb{H}_{\text{sup}}^{p,*}$ . Then the LMI

$$(2.1) \quad H + N^* X^* M + M^* X N \gg 0$$

has a solution  $X \in \mathbb{H}_{\text{sup}}^{q,r}$  if and only if  $H$  is uniformly positive over  $\mathcal{N}(N) \cup \mathcal{N}(M)$ .

*Proof.* See the appendix.  $\square$

**3. Model setting and preliminaries.** Consider in  $(\Omega, \mathcal{F}, \mathbb{P})$  a homogeneous Markov process  $\theta = \{(\theta_t, \mathcal{F}_t), t \in \mathbb{R}_+\}$ , with right-continuous sample paths and countably infinite state space  $\mathcal{S} = \{1, 2, \dots\}$ , such that

$$(3.1) \quad \mathbb{P}(\theta_{t+dt} = j | \theta_t = i) = \begin{cases} \lambda_{ij} dt + o(dt), & i \neq j, \\ 1 + \lambda_{ii} dt + o(dt), & i = j, \end{cases}$$

where  $0 \leq \lambda_{ij}$  for  $i \neq j$ , and  $0 \leq \lambda_i := -\lambda_{ii} = \sum_{j \in \mathcal{S} \setminus \{i\}} \lambda_{ij} \leq \varrho$  for some  $\varrho < \infty$  and all  $i \in \mathcal{S}$ . We assume that  $\theta_0 : (\Omega, \mathcal{F}) \rightarrow \mathcal{S}$  is a random variable with distribution  $\nu_0$ , and that  $\theta_t$  is available for every  $t \in \mathbb{R}_+$  (this last assumption reflects the fact that the control laws we shall seek are  $\theta_t$ -dependent).

In order to introduce the bounded real lemma (JBRL), it suffices to consider now a simple version of the MJLS which will be stated in the next section (see (4.1)).

$$(3.2) \quad \Sigma : \begin{cases} \dot{\hat{x}}(t) = \hat{A}_{\theta_t} \hat{x}(t) + \hat{B}_{\theta_t} v(t), & \hat{x}(0) = \hat{x}_0, \\ z(t) = \hat{C}_{\theta_t} \hat{x}(t) + \hat{D}_{\theta_t} v(t), & t \in \mathbb{R}_+, \end{cases}$$

for  $\hat{x}_0 \in L_2^{\hat{n}}$ , where  $\hat{A} = (\hat{A}_1, \hat{A}_2, \dots) \in \mathbb{H}_{\text{sup}}^{\hat{n}}$  in the same way as  $\hat{B} \in \mathbb{H}_{\text{sup}}^{n_v, \hat{n}}$ ,  $\hat{C} \in \mathbb{H}_{\text{sup}}^{\hat{n}, n_z}$ , and  $\hat{D} \in \mathbb{H}_{\text{sup}}^{n_v, n_z}$ . In what follows we shall point out some fundamental facts regarding this model, including a modified version of the JBRL [21].

Denoting by  $\hat{x}(\cdot, \hat{x}_0, \theta_0, v)$  the *state response* of system  $\Sigma$  when subjected to arbitrary initial conditions  $(\hat{x}_0, \theta_0)$  and input  $v \in L_2^{n_v}(\mathbb{R}_+)$ , we begin with the following definition.

DEFINITION 3.1. For an initial condition  $\hat{x}(0) = 0$  and arbitrary  $\theta_0$  we define the zero-state response of (3.2) as  $\hat{x}_{zs}(\cdot) = \hat{x}(\cdot, 0, \theta_0, v)$ . On the other hand, for arbitrary initial conditions but an identically zero input, we have the zero-input response,  $\hat{x}_{zi}(\cdot) = \hat{x}(\cdot, \hat{x}_0, \theta_0, 0)$ .

Preserving the terminology, often used in the literature for MJLS, in this paper we deal exclusively with stability in the following internal sense.

DEFINITION 3.2. System (3.2) is said to be stochastically stable (SS) if, for any initial condition  $\hat{x}_0 \in L_2^{\hat{n}}$  and initial distribution  $\nu_0$ , we have that  $\|\hat{x}_{zi}\|_{\mathbb{R}_+} < \infty$ .

Concerning SS for this class of systems, an extensive series of results may be found in the recent paper [12]. In particular, it has been proved that SS of (3.2) implies  $z \in L_2^{n_z}(\mathbb{R}_+)$  for any  $v \in L_2^{n_v}(\mathbb{R}_+)$ . That is, SS leads to some kind of *external*  $L_2$  input-output stability for this system (see also [21, Remark 2]).



If  $\Sigma$  is SS, we may define, in the spirit of  $H_\infty$  theory, the following *perturbation operator*  $\mathbb{L} : L_2^{n_v}(\mathbb{R}_+) \rightarrow L_2^{n_z}(\mathbb{R}_+)$ :

$$(3.3) \quad \mathbb{L}v(t) = \hat{C}_{\theta_t} \hat{x}_{zs}(t) + \hat{D}_{\theta_t} v(t),$$

which describes how input disturbances affect the output of system  $\Sigma$ , in such a way that  $z(\cdot) = \mathbb{L}v(\cdot)$  whenever  $\hat{x}_0 = 0$ . The worst-case effect of such disturbances is measured by the induced norm of  $\mathbb{L}$  from  $L_2^{n_v}(\mathbb{R}_+)$  into  $L_2^{n_z}(\mathbb{R}_+)$ :

$$(3.4) \quad \|\mathbb{L}\| = \sup_{v \in L_2^{n_v}(\mathbb{R}_+), \|v\|_{\mathbb{R}_+} \neq 0} \frac{\|\mathbb{L}v\|_{\mathbb{R}_+}}{\|v\|_{\mathbb{R}_+}}.$$

We conclude this section by presenting one important tool which we shall be employing in the remainder of this paper. The so-called JBRL is the starting point towards a characterization of  $H_\infty$  controllers for the class of MJLS under consideration, when it comes to an LMI approach.

**LEMMA 3.3 (JBRL).** *System (3.2) is SS with  $\|\mathbb{L}\| < \gamma$  if and only if there exist  $P = (P_1, P_2, \dots) \in \tilde{\mathbb{H}}_{\sup}^{\hat{n}-}$  such that, for all  $i \in \mathcal{S}$ ,*

$$(3.5) \quad \begin{bmatrix} \hat{A}_i^* P_i + P_i \hat{A}_i + \sum_{j \in \mathcal{S}} \lambda_{ij} P_j & P_i \hat{B}_i & \hat{C}_i^* \\ \hat{B}_i^* P_i & \gamma^2 I_{n_v} & \hat{D}_i^* \\ \hat{C}_i & \hat{D}_i & I_{n_z} \end{bmatrix} \gg 0.$$

*Proof.* The proof follows directly from [21], in view of Theorem 2.2.  $\square$

**4. Disturbance attenuation.** In this section we shall establish the disturbance attenuation (DA) problem in the jump setting and apply the JBRL in order to characterize the existence of solutions in terms of LMIs.

**4.1. Problem setting.** Bearing in mind the definition of  $\{\theta_t\}$ , consider the MJLS

$$(4.1) \quad \Sigma_u : \begin{cases} \dot{x}(t) = A_{\theta_t} x(t) + B_{\theta_t} v(t) + G_{\theta_t} u(t), \\ z(t) = C_{\theta_t} x(t) + D_{\theta_t} v(t) + H_{\theta_t} u(t), \\ y(t) = \Gamma_{\theta_t} x(t) + L_{\theta_t} v(t) \end{cases}$$

for  $t \in \mathbb{R}_+$  and  $x(0) = x_0 \in L_2^n$ , where  $A = (A_1, A_2, \dots) \in \mathbb{H}_{\sup}^n$  in the same way as  $B \in \mathbb{H}_{\sup}^{n_v, n}$ ,  $G \in \mathbb{H}_{\sup}^{n_u, n}$ ,  $C \in \mathbb{H}_{\sup}^{n_u, n_z}$ ,  $D \in \mathbb{H}_{\sup}^{n_v, n_z}$ ,  $H \in \mathbb{H}_{\sup}^{n_u, n_z}$ ,  $\Gamma \in \mathbb{H}_{\sup}^{n_u, n_y}$ , and  $L \in \mathbb{H}_{\sup}^{n_v, n_y}$ .

We refer to  $x(t)$  as the state variable for a given  $t \in \mathbb{R}_+$  and to  $(x(t), \theta_t)$  as an augmented state variable. Just as before, we assume  $v \in L_2^{n_v}(\mathbb{R}_+)$  is any finite energy stochastic disturbance acting on the system; the specific structure of the  $n_u$ -dimensional control input  $u$  will be defined soon.

We call  $z$  and  $y$  the *error* and *measurement* outputs, respectively. The former represents some adversely disturbed output, which should be relatively insensitive to input disturbances. In addition, the processes  $x$  and  $z$  may only be observed through the measurement  $y$  (which without any loss of generality is not directly fed by  $u$ ), which makes this model rather general for both practical and theoretical reasons. Sometimes we refer to the system (4.1) simply as  $\Sigma_u$ .

To meet the ends of the  $H_\infty$  control problem, it is natural to ask the following question about  $\Sigma_u$ : *How does one characterize a class of (suboptimal) controllers  $\mathcal{K}$*

which guarantee not only stability, but also that the worse-case effect caused by the disturbance ( $v$ ) on the error output ( $z$ ) is smaller than some prescribed attenuation level? (In what follows it will be meaningful to measure such effect through (3.4).)

In our approach, we consider the class of dynamic compensators that map the augmented measurement process  $(y, \theta)$  into control policies  $u$  according to the following ( $k$ -dimensional) model:

$$(4.2) \quad \mathcal{K} : \begin{bmatrix} \dot{x}_{\mathcal{K}}(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} \mathcal{K}_{\theta_t}^{11} & \mathcal{K}_{\theta_t}^{12} \\ \mathcal{K}_{\theta_t}^{21} & \mathcal{K}_{\theta_t}^{22} \end{bmatrix} \begin{bmatrix} x_{\mathcal{K}}(t) \\ y(t) \end{bmatrix}, \quad x_{\mathcal{K}}(0) = 0.$$

The above compensator is completely defined through the matrix  $\mathcal{K} = (\mathcal{K}_1, \mathcal{K}_2, \dots) \in \mathbb{H}_{\text{sup}}^{(k+n_y), (k+n_u)}$ , where  $\mathcal{K}_i = [\mathcal{K}_i^{\bullet\bullet}]$ ,  $i \in \mathcal{S}$ . For this reason, the same symbol is used to denote both the system and the matrix in question without confusion.

It is possible to incorporate both systems,  $\Sigma_u$  and  $\mathcal{K}$ , into a *closed-loop* system  $\Sigma_{\mathcal{K}}$ , with the augmented state variable  $(\hat{x}(t), \theta_t) = (x(t), x_{\mathcal{K}}(t), \theta_t)$  for any  $t \in \mathbb{R}_+$ . Defining  $\hat{n} = n + k$  and  $\hat{x}_0 = (x_0, 0)$  we have that the state and output equations for this  $\hat{n}$ -dimensional system may be written as an instance of (3.2):

$$(4.3) \quad \Sigma_{\mathcal{K}} : \begin{cases} \dot{\hat{x}}(t) = \hat{A}_{\theta_t} \hat{x}(t) + \hat{B}_{\theta_t} v(t), & \hat{x}(0) = \hat{x}_0, \\ z(t) = \hat{C}_{\theta_t} \hat{x}(t) + \hat{D}_{\theta_t} v(t), & t \in \mathbb{R}_+, \end{cases}$$

where  $\hat{A} = (\hat{A}_1, \hat{A}_2, \dots) \in \mathbb{H}_{\text{sup}}^{\hat{n}}$ ,  $\hat{B} = (\hat{B}_1, \hat{B}_2, \dots) \in \mathbb{H}_{\text{sup}}^{n_v, \hat{n}}$ ,  $\hat{C} = (\hat{C}_1, \hat{C}_2, \dots) \in \mathbb{H}_{\text{sup}}^{\hat{n}, n_z}$ ,  $\hat{D} = (\hat{D}_1, \hat{D}_2, \dots) \in \mathbb{H}_{\text{sup}}^{n_v, n_z}$ , and, for  $i \in \mathcal{S}$ ,

$$\begin{aligned} \hat{A}_i &= \begin{bmatrix} A_i + G_i \mathcal{K}_i^{22} \Gamma_i & G_i \mathcal{K}_i^{21} \\ \mathcal{K}_i^{12} \Gamma_i & \mathcal{K}_i^{11} \end{bmatrix}, & \hat{B}_i &= \begin{bmatrix} B_i + G_i \mathcal{K}_i^{22} L_i \\ \mathcal{K}_i^{12} L_i \end{bmatrix}, \\ \hat{C}_i &= [C_i + H_i \mathcal{K}_i^{22} \Gamma_i \quad H_i \mathcal{K}_i^{21}], & \hat{D}_i &= D_i + H_i \mathcal{K}_i^{22} L_i, \end{aligned}$$

or, equivalently,

$$\begin{aligned} \hat{A}_i &= A_i^0 + \hat{G}_i \mathcal{K}_i \hat{\Gamma}_i, & \hat{B}_i &= B_i^0 + \hat{G}_i \mathcal{K}_i \hat{L}_i, \\ \hat{C}_i &= C_i^0 + \hat{H}_i \mathcal{K}_i \hat{\Gamma}_i, & \hat{D}_i &= D_i^0 + \hat{H}_i \mathcal{K}_i \hat{L}_i, \end{aligned}$$

with

$$\begin{bmatrix} A_i^0 & B_i^0 & \hat{G}_i \\ C_i^0 & D_i^0 & \hat{H}_i \\ \hat{\Gamma}_i & \hat{L}_i & \star \end{bmatrix} = \left[ \begin{array}{cc|cc|cc} A_i & 0 & B_i & 0 & G_i & \\ 0 & 0_k & 0_{k \times n_v} & I_k & 0 & \\ \hline C_i & 0_{n_z \times k} & D_i & 0 & H_i & \\ 0 & I_k & 0 & \star & \star & \\ \hline \Gamma_i & 0 & L_i & \star & \star & \end{array} \right],$$

where  $\star$  denotes an entry which has no importance in the rest of this paper (see section 2). For the sake of simplicity we also introduce the following definition.

**DEFINITION 4.1.** A compensator  $\mathcal{K} \in \mathbb{H}_{\text{sup}}^{(k+n_y), (k+n_u)}$  such as (4.2) is said to be  $H_{\infty}$  of level  $\gamma$  whenever the closed-loop system  $\Sigma_{\mathcal{K}}$  is SS and  $\|\mathbb{L}\| < \gamma$ , in accordance with (3.3)–(3.4).

**4.2. Characterization results (general case).** In this subsection we shall derive conditions which relate the existence of  $H_{\infty}$  controllers of a given level  $\gamma$  to the solvability of LMIs. The main results here will be further explored in the next section, where the full-order case is dealt with.

The first result we prove shows that a given compensator  $\mathcal{K}$  guarantees SS of system  $\Sigma_{\mathcal{K}}$  with a desired DA level  $\gamma$  if and only if a specific LMI feasibility problem in the variable  $P$  possesses an adequate solution. It should be noted that *any*  $H_\infty$  controller must satisfy this condition in order to solve the DA problem at hand.

PROPOSITION 4.2. *Given a compensator  $\mathcal{K}$  and a desired disturbance attenuation level  $\gamma > 0$  to be achieved, the following are equivalent:*

- (i) *System  $\Sigma_{\mathcal{K}}$  is SS with  $\|\mathbb{L}\| < \gamma$ .*
- (ii) *There exists  $P = (P_1, P_2, \dots) \in \tilde{\mathbb{H}}_{\sup}^{\hat{n}-}$  such that*

$$(4.4) \quad \mathcal{M}^0 + (\mathcal{H}\mathcal{P})^* \mathcal{K} \mathcal{J} + \mathcal{J}^* \mathcal{K}^* (\mathcal{H}\mathcal{P}) \gg 0,$$

where  $\mathcal{M}^0 = (\mathcal{M}_1^0, \mathcal{M}_2^0, \dots) \in \mathbb{H}_{\sup}^{(\hat{n}+n_v+n_z)^*}$ , with

$$\mathcal{M}_i^0 = \begin{bmatrix} P_i A_i^0 + (A_i^0)^* P_i + \sum_{j \in \mathcal{S}} \lambda_{ij} P_j & P_i B_i^0 & (C_i^0)^* \\ (B_i^0)^* P_i & \gamma^2 I_{n_v} & D_i^* \\ C_i^0 & D_i & I_{n_z} \end{bmatrix}, \quad i \in \mathcal{S},$$

and  $\mathcal{P} \in \mathbb{H}_{\sup}^{(\hat{n}+n_v+n_z)^*}$ ,  $\mathcal{H} \in \mathbb{H}_{\sup}^{(\hat{n}+n_v+n_z), (k+n_u)}$ , and  $\mathcal{J} \in \mathbb{H}_{\sup}^{(\hat{n}+n_v+n_z), (k+n_y)}$  given by

$$\mathcal{P} = \begin{bmatrix} P & 0 & 0 \\ 0 & I_{n_v} & 0 \\ 0 & 0 & I_{n_z} \end{bmatrix}, \quad \mathcal{H} = \begin{bmatrix} \hat{G}^* & 0_{(k+n_u) \times n_v} & \hat{H}^* \end{bmatrix},$$

$$\mathcal{J} = \begin{bmatrix} \hat{\Gamma} & \hat{L} & 0_{(k+n_y) \times n_z} \end{bmatrix}.$$

*Proof.* The proof relies on straightforward manipulations of (3.5), which is equivalent to condition (i) above (Lemma 3.3). It turns out that (i) holds if and only if, for some  $P = (P_1, P_2, \dots) \in \tilde{\mathbb{H}}_{\sup}^{\hat{n}-}$  and every  $i \in \mathcal{S}$ ,

$$\mathcal{M}_i^0 + \text{Her} \begin{bmatrix} P_i \hat{G}_i \mathcal{K}_i \hat{\Gamma}_i & P_i \hat{G}_i \mathcal{K}_i \hat{L}_i & 0_{\hat{n} \times n_z} \\ 0_{n_v \times \hat{n}} & 0_{n_v} & 0_{n_v \times n_z} \\ \hat{H}_i \mathcal{K}_i \hat{\Gamma}_i & \hat{H}_i \mathcal{K}_i \hat{L}_i & 0_{n_z} \end{bmatrix} \gg 0,$$

which corresponds to (ii),  $\mathcal{M}^0 + \text{Her}\{(\mathcal{H}\mathcal{P})^* \mathcal{K} \mathcal{J}\} \gg 0$ .  $\square$

The following proposition states that the existence of any compensator *at all* which is  $H_\infty$  of level  $\gamma$  depends on the feasibility of two specific sets of matrix inequalities.

PROPOSITION 4.3. *There exists a compensator  $\mathcal{K}$  which is  $H_\infty$  of level  $\gamma$  if and only if the following conditions are satisfied for some  $P = (P_1, P_2, \dots) \in \tilde{\mathbb{H}}_{\sup}^{\hat{n}-}$ :*

- (i)  $\mathcal{M}^0 \gg 0$  over  $\mathcal{N}(\mathcal{J})$ ;
- (ii)  $\mathcal{P}^{-1} \mathcal{M}^0 \mathcal{P}^{-1} \gg 0$  over  $\mathcal{N}(\mathcal{H})$ .

*Proof.* (Necessity.) Suppose that such  $\mathcal{K}$  exists. Then, from Proposition 4.2, there must exist  $P \in \tilde{\mathbb{H}}_{\sup}^{\hat{n}-}$  such that (4.4) is satisfied. From Lemma 2.5, condition (i) is just a consequence of this fact, along with the following, for some  $\mu > 0$ :

$$(4.5) \quad y_i^* \mathcal{M}_i^0 y_i \geq \mu y_i^* y_i \text{ whenever } \mathcal{H}_i \mathcal{P}_i y_i = 0.$$

Now notice that the linear map  $\mathcal{N}(\mathcal{H}_i \mathcal{P}_i) \ni y_i \mapsto \mathcal{P}_i y_i = x_i \in \mathcal{N}(\mathcal{H}_i)$  is a bijection between  $\mathcal{N}(\mathcal{H}_i \mathcal{P}_i)$  and  $\mathcal{N}(\mathcal{H}_i)$ . Thus, the above condition holds if and only if, for all  $i \in \mathcal{S}$ ,

$$(4.6) \quad (\mathcal{P}_i^{-1} x_i)^* \mathcal{M}_i^0 (\mathcal{P}_i^{-1} x_i) \geq \mu (\mathcal{P}_i^{-1} x_i)^* (\mathcal{P}_i^{-1} x_i)$$

$$= \mu x_i^* \mathcal{P}_i^{-2} x_i$$

holds true for every such  $x_i$ . But since  $P \in \tilde{\mathbb{H}}_{\text{sup}}^{\hat{n}-}$ , it follows that  $P^2 \in \tilde{\mathbb{H}}_{\text{sup}}^{\hat{n}+}$ ; thus, there must exist positive bounds  $\varepsilon_1, \varepsilon_2$  such that

$$(4.7) \quad \varepsilon_1 I \leq P^2 \leq \varepsilon_2 I \quad \Leftrightarrow \quad \varepsilon_1^{-1} I \geq P^{-2} \geq \varepsilon_2^{-1} I,$$

that is,  $P^{-2} \in \tilde{\mathbb{H}}_{\text{sup}}^{\hat{n}+}$  also. Then we have that the congruence transformation  $\mathfrak{C}(\cdot, \mathcal{P}^{-1})$  (see section 2) satisfies the conditions described in Proposition 2.1, from which (4.6) implies that, for some  $\varepsilon > 0$  sufficiently small and every  $x_i \in \mathcal{N}(\mathcal{H}_i)$ ,

$$(\mathcal{P}_i^{-1} x_i)^* \mathcal{M}_i^0 (\mathcal{P}_i^{-1} x_i) \geq \mu x_i^* \mathcal{P}_i^{-2} x_i > \varepsilon x_i^* x_i,$$

which corresponds to (ii), since the existence of a norm bound for  $\mathfrak{C}(\mathcal{M}^0, \mathcal{P}^{-1})$  is ensured by Proposition 2.1.

(Sufficiency.) Since condition (ii) is equivalent to (4.5) holding for all  $i \in \mathcal{S}$ , it follows that the existence of such  $P \in \tilde{\mathbb{H}}_{\text{sup}}^{\hat{n}-}$  guarantees (again from Lemma 2.5) that there is  $\mathcal{K}$  such that (4.4) is satisfied, which from Proposition 4.2 implies that this controller meets the desired condition.  $\square$

It should be noted that, even though the first condition of the last proposition corresponds to an LMI feasibility problem, the same doesn't hold for the second one. Besides, it will be shown that the restrictions on  $\mathcal{N}(\mathcal{J})$  and  $\mathcal{N}(\mathcal{H})$  can be expressed under simpler forms. With the aid of Corollary 2.4 we shall restate the last result in a more suitable way in what follows.

PROPOSITION 4.4. For  $X = (X_1, X_2, \dots)$ ,  $Y = (Y_1, Y_2, \dots) \in \mathbb{H}_{\text{sup}}^{n*}$ ,  $P_2 = (P_{21}, P_{22}, \dots)$ ,  $S_2 = (S_{21}, S_{22}, \dots) \in \mathbb{H}_{\text{sup}}^{k,n}$ ,  $P_3 = (P_{31}, P_{32}, \dots) \in \mathbb{H}_{\text{sup}}^k$ , and  $S_3 = (S_{31}, S_{32}, \dots) \in \mathbb{H}_{\text{sup}}^k$ , let

$$P_i := \begin{bmatrix} X_i & P_{2i} \\ P_{2i}^* & P_{3i} \end{bmatrix}, \quad S_i := P_i^{-1} = \begin{bmatrix} Y_i & S_{2i} \\ S_{2i}^* & S_{3i} \end{bmatrix}, \quad i \in \mathcal{S}.$$

Then

(i)  $\mathcal{M}^0 \gg 0$  over  $\mathcal{N}(\mathcal{J})$  if and only if

$$(4.8) \quad \begin{bmatrix} A_i^* X_i + X_i A_i + \sum_{j \in \mathcal{S}} \lambda_{ij} X_j & X_i B_i & C_i^* \\ B_i^* X_i & \gamma^2 I_{n_v} & D_i^* \\ C_i & D_i & I_{n_z} \end{bmatrix} \gg 0$$

over  $\mathcal{N} \left[ \begin{array}{ccc} \Gamma_i & L_i & 0_{n_y \times n_z} \end{array} \right];$

(ii)  $\mathcal{P}^{-1} \mathcal{M}^0 \mathcal{P}^{-1} \gg 0$  over  $\mathcal{N}(\mathcal{H})$  if and only if

$$(4.9) \quad \begin{bmatrix} Y_i A_i^* + A_i Y_i + \lambda_{ii} Y_i + \sum_{j \neq i} \lambda_{ij} (S_i S_j^{-1} S_i)_{11} & Y_i C_i^* & B_i \\ C_i Y_i & I_{n_z} & D_i \\ B_i^* & D_i^* & \gamma^2 I_{n_v} \end{bmatrix} \gg 0$$

over  $\mathcal{N} \left[ \begin{array}{ccc} G_i^* & H_i^* & 0_{n_u \times n_v} \end{array} \right],$

where  $(S_i S_j^{-1} S_i)_{11} = Y_i X_j Y_i + Y_i P_{2j} S_{2i}^* + S_{2i} P_{2j}^* Y_i + S_{2i} P_{3j} S_{2i}^*$ .

*Proof.* First of all, notice that  $\mathcal{N}(\mathcal{J}_i)$  may be rewritten as

$$\mathcal{N} \left[ \begin{array}{ccc} \hat{\Gamma}_i & \hat{L}_i & 0_{(k+n_y) \times n_z} \end{array} \right] = \mathcal{N} \left[ \begin{array}{cccc} 0 & I_k & 0 & 0 \\ \Gamma_i & 0 & L_i & 0_{n_y \times n_z} \end{array} \right] = \mathcal{R} \left[ \begin{array}{cc} \Phi_{1i} & 0 \\ 0 & 0_k \\ \Phi_{2i} & 0 \\ 0 & I_{n_z} \end{array} \right]$$

whenever  $\mathcal{N} \begin{bmatrix} \Gamma_i & L_i \end{bmatrix} = \mathcal{R} \begin{bmatrix} \Phi_{1i} \\ \Phi_{2i} \end{bmatrix}$ . Thus, noticing that  $(P_i A_i^0)_{11} = X_i A_i$ , we have that the first condition of Proposition 4.3 is equivalent to

$$\begin{aligned} 0 &\ll \mathfrak{C} \left( \begin{bmatrix} A_i^* X_i + X_i A_i + \sum_{j \in \mathcal{S}} \lambda_{ij} X_j & \star & X_i B_i & C_i^* \\ \star & \star & \star & \star \\ B_i^* X_i & \star & \gamma^2 I_{n_v} & D_i^* \\ C_i & \star & D_i & I_{n_z} \end{bmatrix}, \begin{bmatrix} \Phi_{1i} & 0 \\ 0 & 0 \\ \Phi_{2i} & 0 \\ 0 & I \end{bmatrix} \right) \\ &= \mathfrak{C} \left( \begin{bmatrix} A_i^* X_i + X_i A_i + \sum_{j \in \mathcal{S}} \lambda_{ij} X_j & X_i B_i & C_i^* \\ B_i^* X_i & \gamma^2 I_{n_v} & D_i^* \\ C_i & D_i & I_{n_z} \end{bmatrix}, \begin{bmatrix} \Phi_{1i} & 0 \\ \Phi_{2i} & 0 \\ 0 & I_{n_z} \end{bmatrix} \right), \end{aligned}$$

from which (i) follows. With respect to the second condition of Proposition 4.3 we have that

$$\begin{aligned} &\mathcal{P}_i^{-1} \mathcal{M}_i^0 \mathcal{P}_i^{-1} \\ &= \mathfrak{C} \left( \begin{bmatrix} (A_i^0)^* P_i + P_i A_i^0 + \sum_{j \in \mathcal{S}} \lambda_{ij} P_j & \star & \star \\ (B_i^0)^* P_i & \gamma^2 I_{n_v} & \star \\ C_i^0 & D_i & I_{n_z} \end{bmatrix}, \begin{bmatrix} P_i^{-1} & 0 & 0 \\ 0 & I_{n_v} & 0 \\ 0 & 0 & I_{n_z} \end{bmatrix} \right) \\ &= \begin{bmatrix} S_i (A_i^0)^* + A_i^0 S_i + \sum_{j \in \mathcal{S}} \lambda_{ij} S_i S_j^{-1} S_i & B_i^0 & S_i (C_i^0)^* \\ (B_i^0)^* & \gamma^2 I_{n_v} & D_i^* \\ C_i^0 S_i & D_i & I_{n_z} \end{bmatrix}, \end{aligned}$$

which clearly corresponds to (ii). Moreover, we have that  $\mathcal{N}(\mathcal{H}_i)$  may be rewritten as

$$\mathcal{N}(\mathcal{H}_i) = \mathcal{N} \left( \begin{bmatrix} \hat{G}_i^* & 0_{(k+n_u) \times n_v} & \hat{H}_i^* \end{bmatrix} \right) = \mathcal{R} \begin{bmatrix} \hat{\Psi}_{1i} & 0 \\ 0 & I_{n_v} \\ \hat{\Psi}_{2i} & 0 \end{bmatrix}$$

whenever  $\mathcal{R} \begin{bmatrix} \hat{\Psi}_{1i} \\ \hat{\Psi}_{2i} \end{bmatrix} = \mathcal{N} \begin{bmatrix} \hat{G}_i^* & \hat{H}_i^* \end{bmatrix}$ . We rewrite condition (ii) as

$$\begin{aligned} 0 &\ll \mathfrak{C} \left( \begin{bmatrix} S_i (A_i^0)^* + A_i^0 S_i + \sum_{j \in \mathcal{S}} \lambda_{ij} S_i S_j^{-1} S_i & \star & \star \\ (B_i^0)^* & \gamma^2 I_{n_v} & \star \\ C_i^0 S_i & D_i & I_{n_z} \end{bmatrix}, \begin{bmatrix} \hat{\Psi}_{1i} & 0 \\ 0 & I \\ \hat{\Psi}_{2i} & 0 \end{bmatrix} \right) \\ &= \mathfrak{C} \left( \begin{bmatrix} S_i (A_i^0)^* + A_i^0 S_i + \sum_{j \in \mathcal{S}} \lambda_{ij} S_i S_j^{-1} S_i & \star & \star \\ C_i^0 S_i & I_{n_z} & \star \\ (B_i^0)^* & D_i^* & \gamma^2 I_{n_v} \end{bmatrix}, \begin{bmatrix} \hat{\Psi}_{1i} & 0 \\ \hat{\Psi}_{2i} & 0 \\ 0 & I_{n_v} \end{bmatrix} \right). \end{aligned}$$

But

$$\mathcal{R} \begin{bmatrix} \hat{\Psi}_{1i} \\ \hat{\Psi}_{2i} \end{bmatrix} = \mathcal{N} \begin{bmatrix} \hat{G}_i^* & \hat{H}_i^* \end{bmatrix} = \mathcal{N} \begin{bmatrix} 0 & I_k & 0 \\ G_i^* & 0 & H_i^* \end{bmatrix} = \mathcal{R} \begin{bmatrix} \Psi_{1i} & 0 \\ 0 & 0_k \\ \Psi_{2i} & 0 \end{bmatrix}$$

whenever  $\mathcal{N} \begin{bmatrix} G_i^* & H_i^* \end{bmatrix} = \mathcal{R} \begin{bmatrix} \Psi_{1i} \\ \Psi_{2i} \end{bmatrix}$ . By substituting into the previous expression we get the equivalence to

$$\begin{aligned} 0 &\ll \mathfrak{C} \left( \begin{bmatrix} Y_i A_i^* + A_i Y_i + \sum_{j \in \mathcal{S}} \lambda_{ij} (S_i S_j^{-1} S_i)_{11} & \star & Y_i C_i^* & B_i \\ \star & \star & \star & \star \\ C_i Y_i & \star & I_{n_z} & D_i \\ B_i^* & \star & D_i^* & \gamma^2 I_{n_v} \end{bmatrix}, \begin{bmatrix} \Psi_{1i} & 0 \\ 0 & 0_{k \times n_v} \\ \Psi_{2i} & 0 \\ 0 & I_{n_v} \end{bmatrix} \right) \\ &= \mathfrak{C} \left( \begin{bmatrix} Y_i A_i^* + A_i Y_i + \sum_{j \neq i} \lambda_{ij} (S_i S_j^{-1} S_i)_{11} & \star & \star \\ C_i Y_i & I_{n_z} & \star \\ B_i^* & D_i^* & \gamma^2 I_{n_v} \end{bmatrix}, \begin{bmatrix} \Psi_{1i} & 0 \\ \Psi_{2i} & 0 \\ 0 & I_{n_v} \end{bmatrix} \right), \end{aligned}$$

and (ii) follows directly.  $\square$

*Remark 4.5.* In the single-mode case (when  $\mathcal{S} = \{1\}$ ) the above result reconciles with the LMI characterization results stated, for instance, in [14], [15], and [18].

Although the last result applies to the general-order case (when  $k$  is arbitrary), it has two major drawbacks:

- (i) Relation (4.9) depends on every entry of  $S$  (and  $P$ , consequently) through the term  $S_i S_j^{-1} S_i$ .
- (ii) The equality  $S = P^{-1}$  leads to the coupling condition  $X = (Y - S_2 S_3^{-1} S_2^*)^{-1}$ , which is nonlinear and conservative in excess.

In an effort to overcome such drawbacks, we consider in the next subsection the full-order case. The main idea is that, by restricting ourselves to a specific class of Lyapunov functions, a fairly complete LMI characterization may be derived along the lines of Proposition 4.4.

**4.3. Characterization results (full-order case).** This subsection deals with the so-called *full-order case*, in which  $k = n$  (that is, we consider controllers of the same order as the to-be-controlled system). The main result (Theorem 4.8) states an equivalent condition to the existence of  $H_\infty$  controllers of such type in terms of two distinct LMI problems (see also Algorithm 4.10).

The basic idea here is to restrict ourselves to the class of quadratic Lyapunov functions parametrized by

$$(4.10) \quad P_i = \begin{bmatrix} X_i & Y_i^{-1} - X_i \\ Y_i^{-1} - X_i & X_i - Y_i^{-1} \end{bmatrix}, \quad i \in \mathcal{S}.$$

Notice that, in this case,  $S_i := P_i^{-1} = \begin{bmatrix} Y_i & Y_i \\ Y_i & \star \end{bmatrix}$  for every such  $i$ , and hence

$$(4.11) \quad S_i S_j^{-1} S_i = \begin{bmatrix} Y_i Y_j^{-1} Y_i & \star \\ \star & \star \end{bmatrix}.$$

In what follows we shall investigate what conditions  $X, Y \in \mathbb{H}_{\text{sup}}^{n*}$  must satisfy so that a quadratic Lyapunov function may be defined with the aid of (4.10). First, we derive a sufficient condition in terms of LMIs. This result, in conjunction with an auxiliary lemma, will be germane to the proof of Theorem 4.8, which shows that the sufficient condition is also necessary.

**THEOREM 4.6.** *There exists a full-order  $H_\infty$  compensator  $\mathcal{K}$  of level  $\gamma$  whenever there exist  $X, Y \in \mathbb{H}_{\text{sup}}^{n*}$  such that the following set of LMIs is satisfied for every  $i \in \mathcal{S}$ :*

$$(4.12a) \quad \begin{bmatrix} A_i^* X_i + X_i A_i + \sum_{j \in \mathcal{S}} \lambda_{ij} X_j & X_i B_i & C_i^* \\ B_i^* X_i & \gamma^2 I_{n_v} & D_i^* \\ C_i & D_i & I_{n_z} \end{bmatrix} \gg 0$$

over  $\mathcal{N} \begin{bmatrix} \Gamma_i & L_i & 0_{n_y \times n_z} \end{bmatrix},$

$$(4.12b) \quad \begin{bmatrix} Y_i A_i^* + A_i Y_i + \lambda_{ii} Y_i & Y_i C_i^* & B_i & \underline{\lambda}'_i \otimes Y_i \\ C_i Y_i & I_{n_z} & D_i & 0 \\ B_i^* & D_i^* & \gamma^2 I_{n_v} & 0 \\ \underline{\lambda}_i \otimes Y_i & 0 & 0 & \mathbb{D}_i(Y) \end{bmatrix} \gg 0$$

over  $\mathcal{N} \begin{bmatrix} G_i^* & H_i^* & 0_{n_u \times n_v} & 0_{n_u \times \infty} \end{bmatrix},$

and

$$(4.12c) \quad \begin{bmatrix} Y_i & I \\ I & X_i \end{bmatrix} \ll 0,$$

where  $\underline{\lambda}_i = (\sqrt{\lambda_{i1}}, \dots, \sqrt{\lambda_{i(i-1)}}, \sqrt{\lambda_{i(i+1)}}, \dots)$ ,  $\mathbb{D}_i(Y) := -\text{diag}(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots)$  for any such  $i$ , and  $0_{\bullet \times \infty}$  is a zero matrix with an infinite number of columns.

*Proof.* The idea behind this proof is to further improve the previous result (Proposition 4.4) by choosing  $P$  as in (4.10). Notice that (4.12c) implies  $X - Y^{-1} \ll 0$  and

$$(4.13) \quad X - (Y^{-1} - X)(X - Y^{-1})^{-1}(Y^{-1} - X) = Y^{-1} \ll 0,$$

which guarantees that this particular choice satisfies  $P \ll 0$ .

Since (4.8) depends only on the first diagonal block of  $P$  this condition remains unchanged, (4.12a). Hence, it remains only to prove that (4.12b) yields (4.9).

Let  $\Phi = (\Phi_1, \Phi_2, \dots)$  be an orthonormal basis for  $\mathcal{N} \begin{bmatrix} G_i^* & H_i^* & 0_{n_u \times n_v} \end{bmatrix}$  and, for any such  $i \in \mathcal{S}$ , define

$$\left[ \begin{array}{c|c} \frac{U_i}{V_i^*} & \frac{V_i}{W_i} \end{array} \right] := \left[ \begin{array}{c|c} \Phi_i^* \begin{pmatrix} \text{Her}(A_i Y_i) + \lambda_{ii} Y_i & Y_i C_i^* & B_i \\ C_i Y_i & I_{n_z} & D_i \\ B_i^* & D_i^* & \gamma^2 I_{n_v} \end{pmatrix} \Phi_i & \Phi_i^* \begin{pmatrix} \underline{\lambda}'_i \otimes Y_i \\ 0 \\ 0 \end{pmatrix} \\ \hline \begin{pmatrix} \underline{\lambda}_i \otimes Y_i & 0 & 0 \end{pmatrix} \Phi_i & \mathbb{D}_i(Y) \end{array} \right].$$

Then, bearing in mind (A.5) and (A.9), it is not difficult to see that application of the transformation  $\mathfrak{C}\{\cdot, [-W_i^{-1} V_i^*]\}$ ,  $i \in \mathcal{S}$ , to both sides of (4.12b) yields

$$(4.14) \quad \left[ \begin{array}{c|c} Y_i A_i^* + A_i Y_i + \lambda_{ii} Y_i + \sum_{j \neq i} \lambda_{ij} Y_i Y_j^{-1} Y_i & Y_i C_i^* & B_i \\ C_i Y_i & I_{n_z} & D_i \\ B_i^* & D_i^* & \gamma^2 I_{n_v} \end{array} \right] \gg 0$$

over  $\mathcal{N} \begin{bmatrix} G_i^* & H_i^* & 0_{n_u \times n_v} \end{bmatrix}$ .

In fact, assume the left-hand side (LHS) of (4.12b) is  $\varepsilon$ -positive; then, in a similar way to (A.5), we have

$$(4.15) \quad U_i - V_i W_i^{-1} V_i^* = \mathfrak{C} \left\{ \begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix}, \begin{bmatrix} I \\ -W_i^{-1} V_i^* \end{bmatrix} \right\} \geq \varepsilon I$$

for every  $i \in \mathcal{S}$ . Finally, we have from (A.9) that

$$\|U_i - V_i W_i^{-1} V_i^*\| \leq (1 + \|V_i W_i^{-1}\|)^2 \left\| \begin{bmatrix} U & V \\ V^* & W \end{bmatrix} \right\|_{\sup},$$

in which, for every such  $i \in \mathcal{S}$ ,

$$(4.16) \quad \begin{aligned} \|V_i W_i^{-1}\| &\leq \|\Phi_i\| \|(\underline{\lambda}'_i \otimes Y_i) \mathbb{D}_i(Y^{-1})\| \leq \|\underline{\lambda}_i\| \|Y\|_{\sup} \|Y^{-1}\|_{\sup} \\ &= \left( \sum_{j \in \mathcal{S} \setminus \{i\}} \lambda_{ij} \right)^{1/2} \|Y\|_{\sup} \|Y^{-1}\|_{\sup} \\ &\leq \sqrt{\varrho} \|Y\|_{\sup} \|Y^{-1}\|_{\sup} \end{aligned}$$

with  $\varrho$  as defined in the beginning of section 3. This concludes the proof, since (4.14) is just a restatement of (4.9) with  $P$  given by (4.10).  $\square$

Before presenting our main result, we state the following lemma. It is proven that the feasibility of a specific LMI problem is necessary for the existence of full-order controllers, in the spirit of [20, Theorem 4].

LEMMA 4.7. *Suppose there exists a full-order  $H_\infty$  compensator  $\mathcal{K}$  of level  $\gamma$ . Then there exist  $X = (X_1, X_2, \dots)$ ,  $Y = (Y_1, Y_2, \dots) \in \mathbb{H}_{\sup}^{n_y, n}$ ,  $J = (J_1, J_2, \dots) \in \mathbb{H}_{\sup}^{n_y, n}$ ,  $F = (F_1, F_2, \dots) \in \mathbb{H}_{\sup}^{n, n_u}$ , and  $U = (U_1, U_2, \dots) \in \mathbb{H}_{\sup}^{n_y, n_u}$  such that the following LMIs are satisfied for every  $i \in \mathcal{S}$ :*

$$(4.17a) \quad \begin{bmatrix} \text{Her}(X_i A_i + J_i \Gamma_i) + \sum_{j \in \mathcal{S}} \lambda_{ij} X_j & * & * \\ (X_i B_i + J_i L_i)^* & \gamma^2 I_{n_v} & * \\ C_i + H_i U_i \Gamma_i & D_i + H_i U_i L_i & I_{n_z} \end{bmatrix} \gg 0,$$

$$(4.17b) \quad \begin{bmatrix} \text{Her}(A_i Y_i + G_i F_i) + \lambda_{ii} Y_i & * & * & * \\ (B_i + G_i U_i L_i)^* & \gamma^2 I_{n_v} & * & * \\ C_i Y_i + H_i F_i & D_i + H_i U_i L_i & I_{n_z} & * \\ \underline{\lambda}_i \otimes Y_i & 0 & 0 & \mathbb{D}_i(Y) \end{bmatrix} \gg 0,$$

$$(4.17c) \quad \begin{bmatrix} Y_i & I \\ I & X_i \end{bmatrix} \ll 0.$$

*Proof.* From the hypothesis there must exist  $P = (P_1, P_2, \dots) \ll 0$  satisfying (3.5). For later use, let us write such  $P$  and  $S := P^{-1}$  under the following compatible form:

$$(4.18) \quad P = \begin{bmatrix} X & P_2 \\ P_2^* & P_3 \end{bmatrix}, \quad S = \begin{bmatrix} Y & S_2 \\ S_2^* & S_3 \end{bmatrix}.$$

Introducing now  $\mathcal{I} = \begin{bmatrix} 0_n & I_n \end{bmatrix}$ , let us make explicit the affine dependence of  $\hat{A}$  on  $\mathcal{K}^{11}$ :

$$(4.19) \quad \hat{A}_i = \begin{bmatrix} A_i + G_i \mathcal{K}_i^{22} \Gamma_i & G_i \mathcal{K}_i^{21} \\ \mathcal{K}_i^{12} \Gamma_i & 0_n \end{bmatrix} + \begin{bmatrix} 0_n & 0 \\ 0 & \mathcal{K}_i^{11} \end{bmatrix} =: \underline{\hat{A}}_i + \mathcal{I}' \mathcal{K}_i^{11} \mathcal{I}.$$

Next, define  $\mathcal{J} = \begin{bmatrix} I_n & 0_n \end{bmatrix}'$  and notice that  $\mathcal{N}(\text{diag}(\mathcal{J}, I_{n_v}, I_{n_z})) = \{0\}$ , so uniform definiteness is preserved under application of  $\mathfrak{C}(\cdot, \text{diag}(\mathcal{J}, I_{n_v}, I_{n_z}))$  on (3.5). Moreover, since  $\mathcal{I}\mathcal{J} = 0$  the hypothesis yields

$$(4.20) \quad \begin{bmatrix} \mathcal{J}'(\underline{\hat{A}}_i^* P_i + P_i \underline{\hat{A}}_i + \sum_{j \in \mathcal{S}} \lambda_{ij} P_j) \mathcal{J} & * & * \\ \hat{B}_i^* P_i \mathcal{J} & \gamma^2 I_{n_v} & * \\ \hat{C}_i \mathcal{J} & \hat{D}_i & I_{n_z} \end{bmatrix} \gg 0.$$

By defining  $U := \mathcal{K}^{22}$ ,  $J := XGU + P_2 \mathcal{K}^{12}$  and performing the indicated calculations, the equivalence between (4.20) and (4.17a) follows immediately.

Proceeding further, we have that the transformation  $\mathfrak{C}(\cdot, \text{diag}(S_i, I_{n_v}, I_{n_z})) \equiv \mathfrak{C}(\cdot, \mathcal{P}^{-1})$ , which also preserves uniform definiteness (see the proof of Proposition 4.3), leaves (3.5) as

$$(4.21) \quad \begin{bmatrix} S_i \hat{A}_i^* + \hat{A}_i S_i + \sum_{j \in \mathcal{S}} \lambda_{ij} S_i S_j^{-1} S_i & * & * \\ \hat{B}_i^* & \gamma^2 I_{n_v} & * \\ \hat{C}_i S_i & \hat{D}_i & I_{n_z} \end{bmatrix} \gg 0.$$



In the spirit of Theorem 2.2, let us now define

$$(4.22) \quad \mathcal{U}_i - \mathcal{V}_i \mathcal{W}_i^{-1} \mathcal{V}_i^* := \begin{bmatrix} S_i \hat{A}_i^* + \hat{A}_i S_i + \lambda_{ii} S_i & * & * \\ \hat{B}_i^* & \gamma^2 I_{n_v} & * \\ \hat{C}_i S_i & \hat{D}_i & I_{n_z} \end{bmatrix} - \begin{bmatrix} \underline{\lambda}_i' \otimes S_i \\ 0 \\ 0 \end{bmatrix} \mathbb{D}_i(S)^{-1} \begin{bmatrix} \underline{\lambda}_i' \otimes S_i \\ 0 \\ 0 \end{bmatrix}^*$$

for each  $i \in \mathcal{S}$ . Then (4.21) implies that

$$\begin{bmatrix} \mathcal{U}_i & \mathcal{V}_i \\ \mathcal{V}_i^* & \mathcal{W}_i \end{bmatrix} = \begin{bmatrix} I & \mathcal{V}_i \mathcal{W}_i^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \mathcal{U}_i - \mathcal{V}_i \mathcal{W}_i^{-1} \mathcal{V}_i^* & 0 \\ 0 & \mathcal{W}_i \end{bmatrix} \begin{bmatrix} I & 0 \\ \mathcal{W}_i^{-1} \mathcal{V}_i^* & I \end{bmatrix} \gg 0$$

on  $\mathcal{S}$ , which may be rewritten more explicitly as

$$(4.23) \quad \begin{bmatrix} S_i \hat{A}_i^* + \hat{A}_i S_i + \lambda_{ii} S_i & * & * & * \\ \hat{B}_i^* & \gamma^2 I_{n_v} & * & * \\ \hat{C}_i S_i & \hat{D}_i & I_{n_z} & * \\ \underline{\lambda}_i \otimes S_i & 0 & 0 & \mathbb{D}_i(S) \end{bmatrix} \gg 0.$$

To see why this is true, we rely on two facts. First, that  $\mathcal{X}_i := \mathcal{W}_i^{-1} \mathcal{V}_i^*$  is such that

$$(4.24) \quad \sup \sigma \{ \mathcal{X}_i^* \mathcal{X}_i \} = \|\mathcal{V}_i \mathcal{W}_i^{-1}\|^2 \leq (\sqrt{\varrho} \|P\|_{\sup} \|S\|_{\sup})^2 < \infty$$

for all  $i \in \mathcal{S}$ ; this can be verified directly from (4.22), in the same vein as (4.16) (here  $\sigma\{\cdot\}$  denotes the spectrum of an infinite matrix, and  $\varrho$  is as defined in the beginning of section 3). Hence, by performing a minor modification to (A.4) and bearing in mind (A.2)–(A.3), it follows that a uniform lower bound must be attained on (4.23). Finally, the other fact to be observed is that, in the spirit of (A.8), the estimate (4.24) should guarantee the existence of a norm bound to the LHS of (4.23) over all  $i \in \mathcal{S}$ .

At this point, we already know that (4.23) is necessary for the existence of compensators. But since  $\mathcal{N}(\mathcal{J}) = \{0\}$  it follows that uniform positivity is not affected by  $\mathfrak{C}\{\cdot, \text{diag}(\mathcal{J}, I_{n_v}, I_{n_z}, \text{diag}(\mathcal{J}, \dots))\}$ . Thus, defining  $F := UTY + \mathcal{K}^{21} S_2^*$  we have (4.23) reduced to (4.17b) by such transformation.

To complete the proof, notice that  $P_2$  may be assumed invertible without loss of generality: given any admissible  $P \ll 0$ , just perturb it to  $\tilde{P} \ll 0$  in such a way that  $\tilde{P}_2$  is invertible and (3.5) continues to hold true. By a classical argument, there always exists such  $\tilde{P} \in \tilde{\mathbb{H}}_{\sup}^{n-}$  arbitrarily close to  $P$ . In light of this, the congruence transformation associated to  $\begin{bmatrix} Y & I \\ -P_3^{-1} P_2^* Y & 0 \end{bmatrix}$  leaves  $P_i \ll 0$  as (4.17c).  $\square$

The following main result of this section unifies the results obtained so far by giving equivalent conditions to the existence of full-order  $H_\infty$  compensators.

**THEOREM 4.8** (full-order characterization). *The following statements are equivalent:*

- (i) *There exists a full-order  $H_\infty$  compensator  $\mathcal{K}$  of level  $\gamma$ .*
- (ii) *There exist suitable  $X, Y$  such that (4.12) is satisfied for every  $i \in \mathcal{S}$ .*
- (iii) *There exist suitable  $X, Y, J, F$ , and  $U$  such that (4.17) is satisfied for every  $i \in \mathcal{S}$ .*

Moreover, given any  $X, Y$  satisfying (ii) there always exist suitable  $J, F$ , and  $U$  such that (iii) is satisfied.

*Proof.* First assume (ii) holds; then Theorem 4.6 guarantees that (i) is satisfied. Next, notice that (i) yields (iii) according to Lemma 4.7 (the fact that such  $X$  and  $Y$  satisfying (iii) may be chosen by (ii) will be proved last). Thus it remains only to prove (iii)  $\Rightarrow$  (ii).

Assume (iii) is true. Since relations (4.17c) and (4.12c) are the same, it remains only to prove that such  $X, Y$  satisfy (4.12a) and (4.12b), respectively. We have that (4.17a) may be written

$$(4.25) \quad \begin{bmatrix} X_i A_i + A_i^* X_i + \sum_{j \in \mathcal{S}} \lambda_{ij} X_j & * & * \\ B_i^* X_i & \gamma^2 I & * \\ C_i & D_i & I \end{bmatrix} + \text{Her} \left( \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & H_i^* \end{bmatrix}^* \begin{bmatrix} J_i \\ 0 \\ U_i \end{bmatrix} \begin{bmatrix} \Gamma_i & L_i & 0 \end{bmatrix} \right) \gg 0.$$

Moreover, the transformation

$$\mathfrak{C} \left( \cdot, \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \right)$$

leaves (4.17b) equivalent to

$$(4.26) \quad \begin{bmatrix} \text{Her}(A_i Y_i) + \sum_{j \in \mathcal{S}} \lambda_{ij} Y_i Y_j^{-1} Y_i & * & * \\ C_i Y_i & I & * \\ B_i^* & D_i^* & \gamma^2 I \end{bmatrix} + \text{Her} \left( \begin{bmatrix} G_i \\ H_i \\ 0 \end{bmatrix} \begin{bmatrix} F_i & 0 & U_i \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & L_i \end{bmatrix} \right) \gg 0$$

(this is proved just as (4.12b)  $\Rightarrow$  (4.14) in Theorem 4.6). Further on, we have that

$$(4.27) \quad \mathcal{N} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & H_i^* \end{bmatrix} = \mathcal{R} \begin{bmatrix} 0 \\ 0 \\ \Psi_{H_i^*} \end{bmatrix}, \quad \mathcal{N} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & L_i \end{bmatrix} = \mathcal{R} \begin{bmatrix} 0 \\ 0 \\ \Psi_{L_i} \end{bmatrix}$$

for some adequate  $\Psi_{H_i^*}, \Psi_{L_i}$ . These facts, along with the uniform projection lemma (Lemma 2.5), lead to the fulfillment of statement (ii).

Finally, we prove that any  $X$  and  $Y$  satisfying (ii) may be fed into (iii) and suitable  $J, F$ , and  $U$  will always exist. In fact, just assume (ii) holds. Then (i) follows, and Lemma 4.7 guarantees that  $U = \mathcal{K}_{22}$ ,  $J = XGU + (Y^{-1} - X)\mathcal{K}^{12}$ , and  $F = UTY + \mathcal{K}^{21}Y$  are such that (4.17a)–(4.17c) are satisfied, completing the proof.  $\square$

*Remark 4.9.* In case  $D$  and  $\mathcal{K}^{22} = U$  are identically zero, all data are real, and the state space of the underlying Markov chain is finite, the LMIs (4.17a)–(4.17c) coincide with those pointed out in Theorem 4.2 of [7]. In addition, notice that here it is stated that such conditions are not only sufficient but also necessary for the existence of  $\mathcal{K}$ , with the class of Lyapunov functions taken into consideration.

Bearing in mind the above characterization result we present the following algorithm as a solution to the *existence* problem. It is noteworthy that each of the design

procedures we present in the next subsection (Algorithms 5.3, 5.4, and 5.5) depend on it to some extent.

**ALGORITHM 4.10** (existence of compensators). *The existence of some  $H_\infty$  compensator of given level  $\gamma > 0$  is guaranteed by solving either one of the following convex feasibility problems:*

**e<sub>1</sub>**: Find  $X = (X_1, X_2, \dots)$ ,  $Y = (Y_1, Y_2, \dots) \in \mathbb{H}_{\sup}^{n*}$ ,  $J = (J_1, J_2, \dots) \in \mathbb{H}_{\sup}^{n_y, n}$ ,  $F = (F_1, F_2, \dots) \in \mathbb{H}_{\sup}^{n, n_u}$ , and  $U = (U_1, U_2, \dots) \in \mathbb{H}_{\sup}^{n_y, n_u}$  such that (4.17) is satisfied for every  $i \in \mathcal{S}$ .

**e<sub>2</sub>**: Find  $X = (X_1, X_2, \dots)$ ,  $Y = (Y_1, Y_2, \dots) \in \mathbb{H}_{\sup}^{n*}$  such that (4.12) is satisfied for every  $i \in \mathcal{S}$ .

On the other hand, whenever it may be proved that either of these problems doesn't have a solution, then such a compensator does not exist at all.

Finally, we would like to point out that the above algorithm is of immediate practical interest when it comes to the finite case, in the sense that it can be efficiently implemented by widely available convex programming software (see, for instance, [3], [19], and the references therein). The following remark closes this subsection with one last consideration in this direction.

**Remark 4.11.** Some evident differences between procedures **e<sub>1</sub>** and **e<sub>2</sub>** are the larger number of variables in the former and the projection-like restrictions in the latter. It would be interesting to determine under which conditions each of these aspects is more critical when aiming for practical implementation.

**5. Design.** In this section we present some tools for the design of full-order  $H_\infty$  compensators. Both the suboptimal and optimal synthesis problems are discussed.

The main theoretical result of this section, whose *statement* has been inspired in [7, Theorem 4.2] (see also [14]), provides the aforementioned formulas for construction of full-order  $H_\infty$  compensators, as follows.

**THEOREM 5.1.** *Suppose that suitable  $X$ ,  $Y$ ,  $J$ ,  $F$ , and  $U$  satisfying the conditions of Theorem 4.8 may be found. Then the following full-order compensator guarantees that SS of the closed-loop system  $\Sigma_K$  is achieved along with a DA level  $\gamma$ :*

$$(5.1) \quad K^{12} = (Y^{-1} - X)^{-1} (J - XGU),$$

$$(5.2) \quad K^{21} = (F - UTY)Y^{-1},$$

$$(5.3) \quad K^{22} = U,$$

and, for every  $i \in \mathcal{S}$ ,

$$(5.4) \quad \begin{aligned} K_i^{11} = & -(Y_i^{-1} - X_i)^{-1} \left\{ X_i(A_i Y_i + G_i F_i) + (J_i - X_i G_i U_i) \Gamma_i Y_i + \tilde{A}_i^* \right. \\ & + \sum_{j \in \mathcal{S}} \lambda_{ij} Y_j^{-1} Y_i - \tilde{C}_i^* (C_i Y_i + H_i F_i) - [X_i B_i + J_i L_i - \tilde{C}_i^* \tilde{D}_i] \\ & \left. \times (\gamma^2 I - \tilde{D}_i^* \tilde{D}_i)^{-1} [\tilde{B}_i - (C_i Y_i + H_i F_i)^* \tilde{D}_i] \right\} Y_i^{-1}, \end{aligned}$$

where

$$(5.5) \quad \begin{bmatrix} \tilde{A}_i & \tilde{B}_i \\ \tilde{C}_i & \tilde{D}_i \end{bmatrix} = \begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix} + \begin{bmatrix} G_i \\ H_i \end{bmatrix} U_i \begin{bmatrix} \Gamma_i & L_i \end{bmatrix}.$$

*Proof.* First of all, notice that such  $X$  and  $Y$  (which are known by hypothesis) satisfy the conditions of Theorem 4.6; moreover, it has been proved there that  $P = (P_1, P_2, \dots)$  given by

$$(5.6) \quad P_i = \begin{bmatrix} X_i & Y_i^{-1} - X_i \\ Y_i^{-1} - X_i & X_i - Y_i^{-1} \end{bmatrix}, \quad i \in \mathcal{S},$$

ensures the existence of some  $H_\infty$  compensator  $\mathcal{K}$  of level  $\gamma$ , and hence it remains only to explicitly present one such compensator. Also notice that, in this case,  $S_i = P_i^{-1} = \begin{bmatrix} Y_i & Y_i \\ Y_i & \star \end{bmatrix}$  for every such  $i$ .

From Theorem 4.8 we have that such  $X, Y, J, F$ , and  $U$  (all of which are known) correspond to a solution of (4.17). Now, from Lemma 4.7 we have that such  $J, F$ , and  $U$  are related to the controller matrices (and to the Lyapunov variable  $P$ ) as

$$(5.7) \quad J = XGU + P_2\mathcal{K}^{12}, \quad F = U\Gamma Y + \mathcal{K}^{21}S_2^*, \quad U = \mathcal{K}^{22},$$

which, in accordance with (4.18) and (5.6), yields (5.1), (5.2), and (5.3).

To complete the proof we have to present some  $\mathcal{K}^{11}$  in order to solve the LMI problem (3.5); from Theorem 2.2, however, this is equivalent to  $0 \ll N^\gamma(P) = (N_1^\gamma(P), N_2^\gamma(P), \dots)$ , where

$$(5.8) \quad \begin{aligned} N_i^\gamma(P) := & \hat{A}_i^* P_i + P_i \hat{A}_i + \sum_{j \in \mathcal{S}} \lambda_{ij} P_j - \hat{C}_i^* \hat{C}_i \\ & - (P_i \hat{B}_i - \hat{C}_i^* \tilde{D}_i)(\gamma^2 I - \tilde{D}_i^* \tilde{D}_i)^{-1} (\hat{B}_i^* P_i - \tilde{D}_i^* \hat{C}_i) \end{aligned}$$

for every such  $i$ , since (4.17a) guarantees that  $\gamma^2 I - \tilde{D}_i^* \tilde{D}_i \gg 0$  (bearing in mind  $\tilde{D}_i := D_i + H_i U_i L_i$ ). Now define  $\mathfrak{R} = (\mathfrak{R}_1, \mathfrak{R}_2, \dots)$  by

$$(5.9) \quad \mathfrak{R}_i = \begin{bmatrix} \mathfrak{R}_i^{11} & \mathfrak{R}_i^{21*} \\ \mathfrak{R}_i^{21} & \mathfrak{R}_i^{22} \end{bmatrix} = \begin{bmatrix} Y_i & I \\ Y_i & 0 \end{bmatrix}^* N_i^\gamma(P) \begin{bmatrix} Y_i & I \\ Y_i & 0 \end{bmatrix}$$

for  $i \in \mathcal{S}$ , with each block dimension according to  $\hat{A}_i$ . Since the above congruence transformation satisfies the conditions of Proposition 2.1, we have that the uniform positivity of  $\mathfrak{R}_i$  on  $i \in \mathcal{S}$  is actually equivalent to that of  $N^\gamma(P)$ .

A sketch of the rest of the proof is as follows. First, we shall prove that  $\mathfrak{R}^{11} \gg 0$  and  $\mathfrak{R}^{22} \gg 0$  regardless of what choice is made on  $\mathcal{K}^{11}$ . Next, we claim that it is possible to choose  $\mathcal{K}^{11}$  in such a way that the off-diagonal element  $\mathfrak{R}^{21}$  is zero, which in fact will correspond to (5.4) for all  $i \in \mathcal{S}$ . Notice, however, that  $\mathfrak{R}^{21} = 0$  is only a sufficient but not a necessary condition (because different controllers could be sought at this point).

Moving further, let us define  $\mathcal{Y}_i = \begin{bmatrix} Y_i & I \\ Y_i & 0 \end{bmatrix}$ ,  $\Xi_i = X_i(A_i Y_i + G_i F_i) + (J_i - X_i G_i U_i) \Gamma_i Y_i + (Y_i^{-1} - X_i) \mathcal{K}_i^{11} Y_i$  and calculate each entry of  $\mathfrak{R}_i$  separately. We have

$$\mathcal{Y}_i^* P_i \hat{A}_i \mathcal{Y}_i = \begin{bmatrix} A_i Y_i + G_i F_i & A_i + G_i U_i \Gamma_i \\ \Xi_i & X_i A_i + J_i \Gamma_i \end{bmatrix} \quad \text{and} \quad \mathcal{Y}_i^* P_j \mathcal{Y}_i = \begin{bmatrix} Y_i Y_j^{-1} Y_i & Y_i Y_j^{-1} \\ Y_j^{-1} Y_i & X_j \end{bmatrix}.$$

Moreover, we have that

$$\mathcal{Y}_i^* \hat{C}_i^* \hat{C}_i \mathcal{Y}_i = \begin{bmatrix} (C_i Y_i + H_i F_i)^* (C_i Y_i + H_i F_i) & (C_i Y_i + H_i F_i)^* (C_i + H_i U_i \Gamma_i) \\ * & (C_i + H_i U_i \Gamma_i)^* (C_i + H_i U_i \Gamma_i) \end{bmatrix},$$

because  $\{C_i + H_i(U_i\Gamma_i + \mathcal{K}_i^{21})\}Y_i = C_iY_i + H_iF_i$ . Finally,

$$\mathcal{Y}_i^* P_i \hat{B}_i = \begin{bmatrix} B_i + G_i U_i L_i \\ X_i B_i + J_i L_i \end{bmatrix} \quad \text{and} \quad \mathcal{Y}_i^* \hat{C}_i^* \tilde{D}_i = \begin{bmatrix} (C_i Y_i + H_i F_i)^* \\ (C_i + H_i U_i \Gamma_i)^* \end{bmatrix} (D_i + H_i U_i L_i).$$

Now substitute the above results back into the definition of  $\mathfrak{R}$  to obtain that

$$\begin{aligned} \mathfrak{R}_i^{\ell m} &= \text{Her}(\mathcal{Y}_i^* P_i \hat{A}_i \mathcal{Y}_i)_{\ell m} + \sum_{j \in \mathcal{S}} \lambda_{ij} (\mathcal{Y}_i^* P_j \mathcal{Y}_i)_{\ell m} - (\mathcal{Y}_i^* \hat{C}_i^* \hat{C}_i \mathcal{Y}_i)_{\ell m} \\ &\quad - \mathcal{Y}_i^* (P_i \hat{B}_i - \hat{C}_i^* \tilde{D}_i)_{\ell \bullet} (\gamma^2 I - \tilde{D}_i^* \tilde{D}_i)^{-1} (\hat{B}_i^* P_i - \tilde{D}_i^* \hat{C}_i)_{\bullet m}, \\ &\quad \text{where } (\ell, m) \in \{1, 2\} \times \{1, 2\}, \\ \Rightarrow \mathfrak{R}_i^{11} &= A_i Y_i + Y_i A_i^* + \sum_{j \in \mathcal{S}} \lambda_{ij} Y_i Y_j^{-1} Y_i + G_i F_i + F_i^* G_i^* - (C_i Y_i + H_i F_i)^* (C_i Y_i + H_i F_i) \\ &\quad - (B_i + G_i U_i L_i - (C_i Y_i + H_i F_i)^* (D_i + H_i U_i L_i)) (\gamma^2 I - \tilde{D}_i^* \tilde{D}_i)^{-1} \\ &\quad \times (B_i + G_i U_i L_i - (C_i Y_i + H_i F_i)^* (D_i + H_i U_i L_i))^*, \\ \mathfrak{R}_i^{22} &= X_i A_i + A_i^* X_i + \sum_{j \in \mathcal{S}} \lambda_{ij} X_j + J_i \Gamma_i + \Gamma_i^* J_i^* - (C_i + H_i U_i \Gamma_i)^* (C_i + H_i U_i \Gamma_i) \\ &\quad - (X_i B_i + J_i L_i - (C_i + H_i U_i \Gamma_i)^* \tilde{D}_i) (\gamma^2 I - \tilde{D}_i^* \tilde{D}_i)^{-1} \\ &\quad \times (X_i B_i + J_i L_i - (C_i + H_i U_i \Gamma_i)^* \tilde{D}_i)^*, \end{aligned}$$

and

$$\begin{aligned} \mathfrak{R}_i^{21} &= X_i (A_i Y_i + G_i F_i) + (J_i - X_i G_i U_i) \Gamma_i Y_i + (Y_i^{-1} - X_i) \mathcal{K}_i^{11} Y_i \\ &\quad + (A_i + G_i U_i \Gamma_i)^* + \sum_{j \in \mathcal{S}} \lambda_{ij} Y_j^{-1} Y_i - (C_i + H_i U_i \Gamma_i)^* (C_i Y_i + H_i F_i) \\ &\quad - \{X_i B_i + J_i L_i - (C_i + H_i U_i \Gamma_i)^* \tilde{D}_i\} (\gamma^2 I - \tilde{D}_i^* \tilde{D}_i)^{-1} \\ &\quad \times \{B_i + G_i U_i L_i - (C_i Y_i + H_i F_i)^* \tilde{D}_i\}^*. \end{aligned}$$

Just as mentioned before, we have from a repeated application of uniform Schur complements (Theorem 2.2, aiming for dimension reduction) that (4.17a) and (4.17b) are equivalent to  $\mathfrak{R}^{11} \gg 0$  and  $\mathfrak{R}^{22} \gg 0$ , respectively, and regardless of what choice is made with respect to  $\mathcal{K}^{11}$ .

Finally, we have that the unique solution  $\mathcal{K}^{11}$  to the algebraic equation  $\mathfrak{R}_i^{21} = 0$ ,  $i \in \mathcal{S}$ , is given by (5.4), concluding the proof.  $\square$

*Remark 5.2.* Suppose  $D$  and  $\mathcal{K}^{22} = U$  are identically zero. Then (5.4) reduces to

$$\begin{aligned} \mathcal{K}_i^{11} &= -(Y_i^{-1} - X_i)^{-1} \left\{ X_i (A_i Y_i + G_i F_i) + J_i \Gamma_i Y_i + A_i^* + \sum_{j \in \mathcal{S}} \lambda_{ij} Y_j^{-1} Y_i \right. \\ (5.10) \quad &\quad \left. - C_i^* (C_i Y_i + H_i F_i) - \gamma^{-2} (X_i B_i + J_i L_i) B_i^* \right\} Y_i^{-1}, \end{aligned}$$

which, together with (5.1) and (5.2), coincides with the result from Theorem 4.2 in [7], in case all data are real and the set  $\mathcal{S}$  is finite.

**5.1. Some algorithms.** In what follows we shall present some design procedures in order to put our results on a more practical basis. It should be noted that this whole

design framework provides a collection of tools which may be readily implemented on convex programming software, at least in the finite case.

The next algorithm provides one possible way of computing a full-order controller such as the one presented in Theorem 5.1.

**ALGORITHM 5.3** (two-step design procedure). *An  $H_\infty$  compensator of given level  $\gamma > 0$  may be constructed according to Theorem 5.1 by the following steps:*

**d<sub>1</sub>:** *Solve the existence problem by means of  $\mathbf{e}_1$ ;*

$\hookrightarrow$  *If such a solution can't be found, then **stop**.*

**d<sub>2</sub>:** *Bearing in mind  $X, Y, J, F$ , and  $U$  from the previous step, build a compensator by means of relations (5.1)–(5.4).*

Looking back at Theorem 4.8 it is possible to propose the following alternative to Algorithm 5.3. The main advantage here is that the existence of solutions depends on the feasibility of a problem of relatively smaller dimension.

**ALGORITHM 5.4** (three-step design procedure). *An  $H_\infty$  compensator of given level  $\gamma > 0$  may be constructed according to Theorem 5.1 by the following steps:*

**D<sub>1</sub>:** *Solve the existence problem by means of  $\mathbf{e}_2$ ;*

$\hookrightarrow$  *If such a solution can't be found, then **stop**.*

**D<sub>2</sub>:** *Bearing in mind such  $X$  and  $Y$  from the above step, find  $J = (J_1, J_2, \dots) \in \mathbb{H}_{\text{sup}}^{n_y, n}$ ,  $F = (F_1, F_2, \dots) \in \mathbb{H}_{\text{sup}}^{n, n_u}$ , and  $U = (U_1, U_2, \dots) \in \mathbb{H}_{\text{sup}}^{n_y, n_u}$  such that (4.17a) and (4.17b) are satisfied (from Theorem 4.8 we have that there always exists a solution to this problem).*

**D<sub>3</sub>:** *With  $X, Y, J, F$ , and  $U$  obtained from the previous steps, build a compensator by means of relations (5.1)–(5.4).*

We now consider the  $H_\infty$  optimization problem—that of computing the smallest DA level possible. The next algorithm provides, by means of a bisectional procedure (see [2] or [6, Algorithm 8.9], for instance), two different ways to compute the optimal DA level that may be achieved by a full-order controller such as  $\mathcal{K}$ , by solving each one of the following semidefinite programs [22]:

$$\begin{array}{ll} \mathbf{SDP}_1 : & \text{minimize } \gamma \\ & \text{subject to } (4.12), i \in \mathcal{S}, \quad \gamma > 0, \end{array}$$

and

$$\begin{array}{ll} \mathbf{SDP}_2 : & \text{minimize } \gamma \\ & \text{subject to } (4.17), i \in \mathcal{S}, \quad \gamma > 0. \end{array}$$

**ALGORITHM 5.5** ( $H_\infty$  optimization). *The smallest DA level  $\gamma = \gamma_*$  that can be achieved by a controller such as (4.2) in the full-order case may be computed, with arbitrary precision  $\varepsilon > 0$ , by the following steps (where  $\iota \in \{1, 2\}$ ):*

**B<sub>1</sub><sup>sec</sup>:** *Let  $\gamma_{\min} = 0$  and choose  $\gamma_{\max} > 2\varepsilon$  so large that  $\mathbf{e}_\iota$  is feasible for  $\gamma = \gamma_{\max}$ .*

**B<sub>2</sub><sup>sec</sup>:** *Let  $\gamma \leftarrow (\gamma_{\min} + \gamma_{\max})/2$  and solve the existence problem by means of  $\mathbf{e}_\iota$ ;*

$\hookrightarrow$  *If a solution to  $\mathbf{e}_\iota$  can be found, then let  $\gamma_{\max} \leftarrow \gamma$ ;*

$\hookrightarrow$  *otherwise, let  $\gamma_{\min} \leftarrow \gamma$ .*

**B<sub>3</sub><sup>sec</sup>:** *Repeat **B<sub>2</sub><sup>sec</sup>** until  $(\gamma_{\max} - \gamma_{\min})/2 < \varepsilon$ .*

**B<sub>4</sub><sup>sec</sup>:** *Return  $\gamma \approx \gamma_*$ .*

An explicit implementation of the above design procedures is presented in what follows. It is worth noting that the example under consideration does not satisfy the simplifying assumptions of [7] and, by consequence, cannot be tackled by current methods in the literature.

**5.1.1. A nominal example.** Let  $\mathcal{S} = \{1, 2\}$ ,  $n = 1$ , and consider system (4.1) in the form

$$\begin{bmatrix} A_1 & B_1 & G_1 \\ C_1 & D_1 & H_1 \\ \Gamma_1 & L_1 & \star \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & -1 & \star \end{bmatrix}, \quad \begin{bmatrix} A_2 & B_2 & G_2 \\ C_2 & D_2 & H_2 \\ \Gamma_2 & L_2 & \star \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & 1 \\ 1 & 1 & \star \end{bmatrix},$$

with Markov switching governed by

$$\begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} -1 & +1 \\ +1 & -1 \end{bmatrix}.$$

Notice that the results of [7] do not apply here, since  $D_2 \neq 0$ . In what follows we shall design a controller which ensures stochastic stability of the closed-loop system together with a prescribed DA level  $\gamma$ .

Let  $\gamma = 5$ . By employing Algorithm 4.10, we obtain that a feasible solution to (4.12)–(4.17) is given by

$i$	$X_i$	$Y_i$	$J_i$	$F_i$	$U_i$
1	-4.1633	-0.5227	19.9958	1.5509	-0.0727
2	-5.0761	-0.2459	19.4919	-0.7369	-1.0000

either by performing  $\mathbf{d}_1$  (in Algorithm 5.3) or  $\mathbf{D}_1$ – $\mathbf{D}_2$  (Algorithm 5.4). A suboptimal controller is then given by

$$\begin{bmatrix} \mathcal{K}_1^{11} & \mathcal{K}_1^{12} \\ \mathcal{K}_1^{21} & \mathcal{K}_1^{22} \end{bmatrix} = \begin{bmatrix} -9.3202 & 8.7519 \\ -2.8944 & -0.0727 \end{bmatrix}, \quad \begin{bmatrix} \mathcal{K}_2^{11} & \mathcal{K}_2^{12} \\ \mathcal{K}_2^{21} & \mathcal{K}_2^{22} \end{bmatrix} = \begin{bmatrix} -28.4562 & 24.3391 \\ 3.9967 & -1.0000 \end{bmatrix}.$$

Finally, by running Algorithm 5.5 with initial condition  $\gamma_{\max} = 5$ , we obtain, after 22 iterations and with precision  $\varepsilon = 10^{-6}$ , the optimal  $H_\infty$  performance  $\gamma_* \approx 3.42735$ .

One final remark goes as follows. For the example under consideration, the obtained results for the optimal case ( $\gamma$  close to  $\gamma_*$ ) are such that  $XY \approx 1$ , which gives rise to unbounded controller entries in (5.1) and (5.4). Although this is a drawback of the presented method, we remark that the same kind of problem arises in the LTI case (see [13] or [15, section 9.4], for example).

**6. Conclusions.** In this paper, the output feedback  $H_\infty$  control has been addressed for a class of continuous-time Markov jump linear systems with the Markov process taking values in an infinite countable set  $\mathcal{S}$ . We have obtained the following results:

- A theorem which characterizes whether there exists a full-order solution to the disturbance attenuation problem in terms of two distinct sets of LMIs (Theorem 4.8). This result connects a certain projection approach to an LMI problem which is more suitable for design.
- Extensions of Schur complements and of the projection lemma to a wider context. We remark here that one is faced with the same *uniformity* problems if dealing with, say, time-variant systems (as, e.g., in [9] or [20]), considering the case where the system parameters are time functions with uniform bounds.
- The JBRL was employed for the first time and valuable results could be obtained, thus illustrating the importance of this recent result and accomplishing an important step in the development of an  $H_\infty$ -like theory for the class of systems considered.
- An LMI algorithm (Algorithm 4.10), which allows one to check whether there exists a solution for the DA problem.

- A two-step design method (Algorithm 5.3) which provides explicit formulas for the construction of a controller.
- An alternative three-step design method, Algorithm 5.4. The main issue here is that one can first check if a smaller (projected) LMI problem is feasible, which amounts to verifying whether the DA problem has a solution or not. This partial solution is then fed into the two-step procedure.
- A bisectional procedure for  $H_\infty$  optimization of infinite Markov jump systems (Algorithm 5.5).
- A nominal example, for the finite case, which illustrates how the obtained results may be employed in a situation where the hypotheses of [7] are not satisfied.

**Appendix.** We present now some of the proofs omitted in the core of the text.

*Proof of Proposition 2.1.* Let us first prove (i). From the hypothesis, there must exist  $\eta > 0$  such that  $Q_i^* Q_i \geq \eta I$  for every  $i \in \mathcal{S}$ . Additionally, since  $X \gg 0$  there must exist  $\varepsilon_0 > 0$  such that  $X_i \geq \varepsilon_0 I_n$  for every such  $i$ . Thus

$$Q_i^* X_i Q_i \geq \varepsilon_0 Q_i^* Q_i \geq (\varepsilon_0 \eta) I = \varepsilon I$$

for some suitable  $\varepsilon > 0$ , and

$$\|Q_i^* X_i Q_i\| \leq \|Q_i\|^2 \|X_i\| \leq \|Q\|_{\sup}^2 \|X\|_{\sup},$$

from which the result follows. To prove (ii), just replace  $X$  by  $-X$ .  $\square$

*Remark A.1.* Two important issues in the above proof are the need to ensure uniform definiteness on the relations and the norm-bound invariance. It highlights two important features which come up when the state space of the Markov chain is assumed to be infinite vis-à-vis the finite case.

Before we proceed to the proof of Theorem 2.2, let us state the following auxiliary result.

**PROPOSITION A.2.** *Given  $U = (U_1, U_2, \dots) \in \mathbb{H}_{\sup}^{p*}$ ,  $V = (V_1, V_2, \dots) \in \mathbb{H}_{\sup}^{q,p}$ , and  $W = (W_1, W_2, \dots) \in \mathbb{H}_{\sup}^{q*}$ , the following are equivalent:*

- (i)  $\begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix} \geq \varepsilon I$  for every  $i \in \mathcal{S}$  and some  $\varepsilon > 0$ ;
- (ii)  $U_i \geq \mu I$  and  $W_i - V_i^* U_i^{-1} V_i \geq \mu I$  for every  $i \in \mathcal{S}$  and some  $\mu > 0$ ;
- (iii)  $W_i \geq \nu I$  and  $U_i - V_i W_i^{-1} V_i^* \geq \nu I$  for every  $i \in \mathcal{S}$  and some  $\nu > 0$ .

*Proof.* Assume (iii) holds. Thus, defining  $X = -W^{-1} V^* \in \mathbb{H}_{\sup}^{p,q}$  we have that, for every such  $i$ ,

$$(A.1) \quad \nu I \leq \begin{bmatrix} U_i - V_i W_i^{-1} V_i^* & 0 \\ 0 & W_i \end{bmatrix} = \begin{bmatrix} I & -X_i^* \\ 0 & I \end{bmatrix} \begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix} \begin{bmatrix} I & 0 \\ -X_i & I \end{bmatrix}.$$

Notice that  $\begin{bmatrix} -I & 0 \\ -X_i & I \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ X_i & I \end{bmatrix}$ ; then the above expression is equivalent to

$$(A.2) \quad \begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix} \geq \nu \begin{bmatrix} I & X_i^* \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ X_i & I \end{bmatrix} = \nu \begin{bmatrix} I + X_i^* X_i & X_i^* \\ X_i & I \end{bmatrix}.$$

But notice that, for some  $\alpha \in (0, 1)$  and every  $i \in \mathcal{S}$ ,

$$(A.3) \quad \begin{aligned} & \begin{bmatrix} I + X_i^* X_i & X_i^* \\ X_i & I \end{bmatrix} - \alpha \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \\ &= \mathfrak{C} \left\{ \begin{bmatrix} (1-\alpha)I - \frac{\alpha}{1-\alpha} X_i^* X_i & 0 \\ 0 & (1-\alpha)I \end{bmatrix}, \begin{bmatrix} I & 0 \\ \frac{1}{1-\alpha} X_i & I \end{bmatrix} \right\} \\ &> 0. \end{aligned}$$



This is due to the fact that  $\mathcal{N} \begin{bmatrix} I & 0 \\ \frac{1}{1-\alpha} X_i & I \end{bmatrix} \equiv \{0\}$ , together with  $(1-\alpha)I > 0$  and

$$\begin{aligned} (1-\alpha)I - \frac{\alpha}{1-\alpha} X_i^* X_i &\geq (1-\alpha)I - \frac{\alpha}{1-\alpha} \lambda_{\max}(X_i^* X_i)I = (1-\alpha)I - \frac{\alpha}{1-\alpha} \|X_i\|^2 I \\ &\geq I - \alpha \left\{ 1 + \frac{1}{1-\alpha} \|X\|_{\sup}^2 \right\} I \\ (A.4) \quad &> 0, \end{aligned}$$

bearing in mind that  $\|X\|_{\sup} \leq \|W^{-1}\|_{\sup} \|V\|_{\sup} < \infty$ , and  $\alpha \in (0, 1)$  may be as small as desired. Hence, (i) follows straight from (A.2)–(A.3) by choosing  $\varepsilon < \alpha\nu$ .

Now assume (i) holds. Then it is immediate that, for every  $i \in \mathcal{S}$ ,

$$\begin{aligned} U_i - V_i W_i^{-1} V_i^* &= \begin{bmatrix} I & -X_i^* \end{bmatrix} \begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix} \begin{bmatrix} I \\ -X_i \end{bmatrix} \\ &\geq \varepsilon \begin{bmatrix} I & -X_i^* \end{bmatrix} \begin{bmatrix} I \\ -X_i \end{bmatrix} = \varepsilon(I + X_i^* X_i) \\ (A.5) \quad &\geq \varepsilon I \end{aligned}$$

and

$$(A.6) \quad W_i = \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix} \begin{bmatrix} 0 \\ I \end{bmatrix} \geq \varepsilon \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} 0 \\ I \end{bmatrix} = \varepsilon I,$$

so that (iii) should hold with  $\nu \equiv \varepsilon$ .

Finally, to prove the equivalence between (i) and (ii), one has to notice only that

$$(A.7) \quad \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix} \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} = \begin{bmatrix} W_i & V_i^* \\ V_i & U_i \end{bmatrix}$$

and that Proposition 2.1 guarantees that the indicated transformation preserves uniform definiteness. Hence, the result follows immediately if we replace  $(U, V, V^*, W)$  by  $(W, V^*, V, U)$  in the first part of the proof.  $\square$

*Proof of Theorem 2.2.* Suppose (iii) holds. From Proposition A.2 we have the positivity in (i) guaranteed, and it remains only to prove that boundedness is preserved. From (A.1) we have that

$$\begin{aligned} \left\| \begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix} \right\| &= \left\| \begin{bmatrix} I & X_i^* \\ 0 & I \end{bmatrix} \begin{bmatrix} U_i - V_i W_i^{-1} V_i^* & 0 \\ 0 & W_i \end{bmatrix} \begin{bmatrix} I & 0 \\ X_i & I \end{bmatrix} \right\| \\ &\leq \left\| \begin{bmatrix} I & 0 \\ X_i & I \end{bmatrix} \right\|^2 \left\| \begin{bmatrix} U_i - V_i W_i^{-1} V_i^* & 0 \\ 0 & W_i \end{bmatrix} \right\| \\ &\leq \left\{ 1 + \left\| \begin{bmatrix} 0 & 0 \\ X_i & 0 \end{bmatrix} \right\| \right\}^2 (\|U_i - V_i W_i^{-1} V_i^*\| + \|W_i\|) \\ (A.8) \quad &\leq (1 + \|X\|_{\sup})^2 (\|U - VW^{-1}V^*\|_{\sup} + \|W\|_{\sup}) \end{aligned}$$

for every  $i \in \mathcal{S}$ , which immediately yields (i).

Assuming now (i), relation (A.1) guarantees that, for every  $i \in \mathcal{S}$ ,

$$\begin{aligned} \|U_i - V_i W_i^{-1} V_i^*\| &= \left\| \begin{bmatrix} I & -X_i^* \end{bmatrix} \begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix} \begin{bmatrix} I \\ -X_i \end{bmatrix} \right\| \\ (A.9) \quad &\leq \left\| \begin{bmatrix} I & -X_i^* \end{bmatrix} \right\|^2 \left\| \begin{bmatrix} U & V \\ V^* & W \end{bmatrix} \right\|_{\sup} \end{aligned}$$

and

$$(A.10) \quad \|W_i\| = \left\| \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix} \begin{bmatrix} 0 \\ I \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} 0 & I \end{bmatrix} \right\|^2 \left\| \begin{bmatrix} U & V \\ V^* & W \end{bmatrix} \right\|_{\sup},$$

which, similarly to (A.8), implies (iii).

Equivalence between (i) and (ii) is proven with the aid of (A.7), as we did in the proof of Proposition A.2. Finally, we have that the negative case benefits from the above proof, in the sense that some given  $Q \ll 0$  if and only if  $-Q \gg 0$ .  $\square$

The following proposition further extends Theorem 2.2 to the wider context of *bounded matrices* (see the beginning of section 2); this natural extension will be employed in the proof of the uniform projection lemma (Lemma 2.5) below.

**PROPOSITION A.3.** *Let  $\mathbf{p} = (p_1, p_2, \dots)$  and  $\mathbf{q} = (q_1, q_2, \dots)$  be such that  $(p_i, q_i) \in \{1, 2, \dots, M\}^2$  for some finite integer  $M$  and every  $i \in \mathcal{S}$ . Then the following statements are equivalent for all  $U = (U_1, U_2, \dots) \in \mathbb{H}_{\sup}^{\mathbf{p}+}$ ,  $V = (V_1, V_2, \dots) \in \mathbb{H}_{\sup}^{\mathbf{q}, \mathbf{p}}$  and  $W = (W_1, W_2, \dots) \in \mathbb{H}_{\sup}^{\mathbf{q}+}$ :*

- (i)  $\begin{bmatrix} U & V \\ V^* & W \end{bmatrix} \gg 0$ ;
- (ii)  $U \gg 0$  and  $W - V^*U^{-1}V \gg 0$ ;
- (iii)  $W \gg 0$  and  $U - VW^{-1}V^* \gg 0$ .

*Proof.* The proof is straightforward. Just notice that Proposition 2.1, as well as Proposition A.2 and Theorem 2.2, may be extended to the case where  $m, n, p$ , and  $q$  are replaced by bounded and arbitrarily chosen sequences of integers  $\mathbf{m}, \mathbf{n}, \mathbf{p}$ , and  $\mathbf{q}$ , respectively, by quite the same proofs.  $\square$

*Proof of Proposition 2.3.* Just notice that, for any  $i \in \mathcal{S}$ , we have that  $p_i \leq \min\{\ell, p\}$ . Then define  $\Phi_i = [\phi_{i,1} \ \dots \ \phi_{i,p_i}] \in \mathbb{C}^{p \times p_i}$ , where the set  $\{\phi_{i,j}\}_{j=1}^{p_i} \subset \mathbb{C}^p$  forms an orthonormal basis for  $\mathcal{N}(\Psi_i)$ , and the proof follows immediately.  $\square$

*Proof of Corollary 2.4.* From Theorem 2.2 we have that the boundedness of the LHS of each of these expressions is equivalent, so it remains only to prove that uniform definiteness is preserved.

Define  $\Phi$  to be an orthonormal basis for  $\mathcal{N}(\Psi)$  (see Proposition 2.3). Then (ii) is equivalent to the existence of some  $\varepsilon > 0$  such that  $W_i > \varepsilon I$  for all  $i \in \mathcal{S}$  and

$$(A.11) \quad 0 < \Phi_i^*(U_i - \varepsilon I - V_i W_i^{-1} V_i^*) \Phi_i = \Phi_i^*(U_i - V_i W_i^{-1} V_i^*) \Phi_i - \varepsilon I,$$

which, due to Theorem 2.2, is equivalent to the inequality

$$\begin{aligned} & \begin{bmatrix} \Phi_i^* & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} U_i & V_i \\ V_i^* & W_i \end{bmatrix} \begin{bmatrix} \Phi_i & 0 \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} \Phi_i^* U_i \Phi_i & \Phi_i^* V_i \\ V_i^* \Phi_i & W_i \end{bmatrix} \\ &> \varepsilon \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} = \varepsilon \begin{bmatrix} \Phi_i^* \Phi_i & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} \Phi_i^* & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \varepsilon I & 0 \\ 0 & \varepsilon I \end{bmatrix} \begin{bmatrix} \Phi_i & 0 \\ 0 & I \end{bmatrix}, \end{aligned}$$

which is equivalent to (i), since  $\mathcal{R} \begin{bmatrix} \Phi_i & 0 \\ 0 & I \end{bmatrix} = \mathcal{N} \begin{bmatrix} \Psi_i & 0 \end{bmatrix}$  for all  $i \in \mathcal{S}$ :

$$\begin{bmatrix} \Phi_i^* & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} U_i - \varepsilon I & V_i \\ V_i^* & W_i - \varepsilon I \end{bmatrix} \begin{bmatrix} \Phi_i & 0 \\ 0 & I \end{bmatrix} > 0. \quad \square$$

*Proof of Lemma 2.5.* The proof of necessity is straightforward. Just assume that such  $X$  exists and notice that (2.1) must hold, in particular, on the indicated set.

The proof of sufficiency goes as follows. For  $\iota \in \{1, 2, 3, 4\}$  and  $i \in \mathcal{S}$  let  $W_{\iota i}$  of adequate dimensions be an orthonormal basis for  $\mathbb{W}_{\iota i}$ , where

$$\begin{aligned}\mathbb{W}_{1i} &= \mathcal{N}(M_i) \cap \mathcal{N}(N_i), & \mathbb{W}_{3i} &= \mathcal{N}(M_i)^\perp \cap \mathcal{N}(N_i), \\ \mathbb{W}_{2i} &= \mathcal{N}(M_i) \cap \mathcal{N}(N_i)^\perp, & \mathbb{W}_{4i} &= \mathcal{N}(M_i)^\perp \cap \mathcal{N}(N_i)^\perp.\end{aligned}$$

Thus, defining  $T = (T_1, T_2, \dots) \in \mathbb{H}_{\text{sup}}^p$  by  $T_i = [W_{1i} \ W_{2i} \ W_{3i} \ W_{4i}]$ ,  $i \in \mathcal{S}$ , we have that  $T^*T = I_p \gg 0$ . Then from Proposition 2.1 it follows that (2.1) is equivalent to

$$(A.12) \quad T^*HT + (NT)^*X^*(MT) + (MT)^*X(NT) \gg 0.$$

Now, similarly to [15, Lemma 3.1], we partition each component of (A.12) just as  $T_i$ . This yields equivalence of (2.1) to

$$(A.13) \quad \begin{bmatrix} \Phi_1 & \Phi_2 & \Phi_3 & \Phi_4 \\ \Phi_2^* & \Phi_5 & \Phi_6 + \mathcal{X}_{11}^* & \Phi_7 + \mathcal{X}_{21}^* \\ \Phi_3^* & \Phi_6^* + \mathcal{X}_{11} & \Phi_8 & \Phi_9 + \mathcal{X}_{12} \\ \Phi_4^* & \Phi_7^* + \mathcal{X}_{21} & \Phi_9^* + \mathcal{X}_{12}^* & \Phi_{10} + \mathcal{X}_{22} + \mathcal{X}_{22}^* \end{bmatrix} \gg 0,$$

where  $[\Phi_\bullet] := T^*HT$  and

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \mathcal{X}_{11} & 0 & \mathcal{X}_{12} \\ 0 & \mathcal{X}_{21} & 0 & \mathcal{X}_{22} \end{bmatrix} := \begin{bmatrix} 0 \\ 0 \\ Q_1^* \\ Q_2^* \end{bmatrix} X \begin{bmatrix} 0 & R_1 & 0 & R_2 \end{bmatrix} = (MT)^*X(NT),$$

in which  $Q = [Q_1 \ Q_2]$  and  $R = [R_1 \ R_2]$  are defined by the full column rank matrices  $[Q_{1i} \ Q_{2i}]$  and  $[R_{1i} \ R_{2i}]$  for which  $M_i T_i = [0 \ 0 \ Q_{1i} \ Q_{2i}]$  and  $N_i T_i = [0 \ R_{1i} \ 0 \ R_{2i}]$ . Then from Proposition A.3 we have that (A.12) is equivalent to the two following relations being satisfied for some quadruple  $\mathcal{X} = (\mathcal{X}_{11}, \mathcal{X}_{12}, \mathcal{X}_{21}, \mathcal{X}_{22})$  of bounded matrices:

$$(A.14) \quad \Upsilon(\mathcal{X}_{11}) := \begin{bmatrix} \Phi_1 & \Phi_2 & \Phi_3 \\ \Phi_2^* & \Phi_5 & \Phi_6 + \mathcal{X}_{11}^* \\ \Phi_3^* & \Phi_6^* + \mathcal{X}_{11} & \Phi_8 \end{bmatrix} \gg 0,$$

$$(A.15) \quad \Phi_{10} + \mathcal{X}_{22} + \mathcal{X}_{22}^* - \begin{bmatrix} \Phi_4 \\ \Phi_7 + \mathcal{X}_{21}^* \\ \Phi_9 + \mathcal{X}_{12} \end{bmatrix}^* \Upsilon(\mathcal{X}_{11})^{-1} \begin{bmatrix} \Phi_4 \\ \Phi_7 + \mathcal{X}_{21}^* \\ \Phi_9 + \mathcal{X}_{12} \end{bmatrix} \gg 0.$$

Finally, after some calculations analogous to [15, Lemma 3.1], we have that (A.14) holds if and only if

$$(A.16) \quad \begin{bmatrix} \Phi_1 & \Phi_2 \\ \Phi_2^* & \Phi_5 \end{bmatrix} \gg 0 \quad \text{together with} \quad \begin{bmatrix} \Phi_1 & \Phi_3 \\ \Phi_3^* & \Phi_8 \end{bmatrix} \gg 0,$$

which from the definition of  $\Phi$  yields immediately the desired result.  $\square$

**Acknowledgment.** The authors would like to express their gratitude to the referees for their suggestions and comments, which have certainly improved the paper.

## REFERENCES

- [1] N. K. BOSE AND Y. Q. SHI, *A simple general proof of the Kharitonov's generalized stability criterion*, IEEE Trans. Circuits and Systems, 34 (1987), pp. 1233–1237.
- [2] S. BOYD, V. BALAKRISHNAN, AND P. KABAMBA, *A bisection method for computing the  $H_\infty$  norm of a transfer matrix and related problems*, Math. Control Signals Systems, 2 (1989), pp. 207–219.
- [3] S. BOYD, L. EL GHAOU, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.
- [4] J. W. BREWER, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits and Systems, 25 (1978), pp. 772–781.
- [5] O. L. V. COSTA AND M. D. FRAGOSO, *Discrete-time LQ-optimal control problems for infinite Markov jump parameter systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 2076–2088.
- [6] O. L. V. COSTA, M. D. FRAGOSO, AND R. P. MARQUES, *Discrete-Time Markov Jump Linear Systems*, Probab. Appl. (N.Y.), Springer-Verlag, New York, 2005.
- [7] D. P. DE FARIAS, J. C. GEROMEL, J. B. R. DO VAL, AND O. L. V. COSTA, *Output feedback control of Markov jump linear systems in continuous-time*, IEEE Trans. Automat. Control, 45 (2000), pp. 944–949.
- [8] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State-space solutions to standard  $H_2$  and  $H_\infty$  control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.
- [9] V. DRAGAN AND T. MOROZAN, *Stability and robust stabilization to linear stochastic systems described by differential equations with Markovian jumping and multiplicative noise*, Stochastic Anal. Appl., 20 (2002), pp. 33–92.
- [10] M. D. FRAGOSO AND J. BACZYNSKI, *Optimal control for continuous-time linear quadratic problems with infinite Markov jump parameters*, SIAM J. Control Optim., 40 (2001), pp. 270–297.
- [11] M. D. FRAGOSO AND J. BACZYNSKI, *Lyapunov coupled equations for continuous-time infinite Markov jump linear systems*, J. Math. Anal. Appl., 274 (2002), pp. 319–355.
- [12] M. D. FRAGOSO AND O. L. V. COSTA, *A unified approach for stochastic and mean square stability of continuous-time linear systems with Markovian jumping parameters and additive disturbances*, SIAM J. Control Optim., 44 (2005), pp. 1165–1191.
- [13] P. GAHINET, *Reliable computation of  $H_\infty$  central controllers near the optimum*, in Proceedings of the 1992 American Control Conference, Chicago, 1992, pp. 738–742.
- [14] P. GAHINET, *Explicit controller formulas for LMI-based  $H_\infty$  synthesis*, Automatica, 32 (1996), pp. 1007–1014.
- [15] P. GAHINET AND P. APKARIAN, *A linear matrix inequality approach to  $H_\infty$  control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.
- [16] D. HERION, J. JEŽEK, AND M. ŠEBEK, *Discrete-time symmetric polynomial equations with complex coefficients*, Kybernetika, 38 (2002), pp. 113–139.
- [17] D. HINRICHSSEN AND A. J. PRITCHARD, *Stochastic  $H^\infty$* , SIAM J. Control Optim., 36 (1998), pp. 1504–1538.
- [18] T. IWASAKI AND R. E. SKELTON, *All controllers for the general  $H_\infty$  control problem: LMI existence conditions and state space formulas*, Automatica, 30 (1994), pp. 1307–1317.
- [19] M. A. RAMI AND L. E. GHAOU, *LMI optimization for nonstandard Riccati equations arising in stochastic control*, IEEE Trans. Automat. Control, 41 (1996), pp. 1666–1671.
- [20] C. W. SCHERER, *Mixed  $H_2/H_\infty$  control*, in Trends in Control: A European Perspective, A. Isidori, ed., Springer-Verlag, Berlin, 1995, pp. 173–216.
- [21] M. G. TODOROV AND M. D. FRAGOSO, *Infinite Markov jump bounded real lemma*, Systems Control Lett., 57 (2008), pp. 64–70.
- [22] L. VANDENBERGHE AND S. BOYD, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

## ON STABILITY AND ROBUST STABILITY OF POSITIVE LINEAR VOLTERRA EQUATIONS\*

PHAM HUU ANH NGOC<sup>†</sup>, TOSHIKI NAITO<sup>†</sup>, JONG SON SHIN<sup>†</sup>, AND  
SATORU MURAKAMI<sup>‡</sup>

**Abstract.** We first introduce the notion of positive linear Volterra integrodifferential equations. Then we give some characterizations of these positive equations. An explicit criterion and a Perron–Frobenius-type theorem for positive linear Volterra integrodifferential equations are given. Then we offer a new criterion for uniformly asymptotic stability of positive equations. Finally, we study stability radii of positive linear Volterra integrodifferential equations. It is proved that complex, real, and positive stability radii of positive linear Volterra equations under structured perturbations (or affine perturbations) coincide and can be computed by explicit formulae. To the best of our knowledge, most of the results of this paper are new.

**Key words.** linear Volterra equation, positive system, uniformly asymptotic stability, stability radius

**AMS subject classifications.** 34K20, 93D09

**DOI.** 10.1137/070679740

**1. Introduction.** Generally speaking, a dynamical system is called *positive* if for any nonnegative initial condition, the corresponding solution of the system is also nonnegative. In particular, a dynamical system with state space  $\mathbb{R}^n$  is positive if any trajectory of the system starting at an initial state in the positive orthant  $\mathbb{R}_+^n$  remains forever in  $\mathbb{R}_+^n$ . Positive dynamical systems play an important role in the modeling of dynamical phenomena whose variables are restricted to be nonnegative. This model class is used in many areas such as economics, population dynamics, and ecology; see [2], [25]. They are often encountered in applications, for example, networks of reservoirs, industrial processes involving chemical reactors, heat exchangers, distillation columns, storage systems, hierarchical systems, compartmental systems used for modeling transport and accumulation phenomena of substances, etc. The mathematical theory of positive systems is based on the theory of nonnegative matrices founded by Perron and Frobenius. As references we mention [2], [5].

Problems of positive systems have attracted a lot of attention from researchers for a long time; see, e.g., [5], [6], [8], [9], [10], [14], [15], [17], [18], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [43], [44], [45], [46]. In the literature, there are some criteria for familiar positive linear systems such as positive linear invariant-time differential (difference) systems, and positive linear time-delay systems of retarded type. For example, it is well known that a linear time-delay system of the form  $\dot{x}(t) = A_0x(t) + A_1x(t-h)$ ,  $t \geq 0$ , is positive if and only if  $A_0$  is a Metzler matrix and  $A_1$  is a nonnegative matrix, and a linear discrete time system of the form  $x(k+1) = A_0x(k) + A_1x(k-h)$ ,  $k \in \mathbb{N}$ ,  $k \geq h$ , is positive if and only if  $A_0, A_1$  are nonnegative matrices; see, e.g., [33], [34], [45].

\*Received by the editors January 10, 2007; accepted for publication (in revised form) November 15, 2007; published electronically February 29, 2008.

<http://www.siam.org/journals/sicon/47-2/67974.html>

<sup>†</sup>Department of Mathematics, The University of Electro-Communications, Chofu, Tokyo, 182-8585, Japan (phanhngoc@yahoo.com, naito@e-one.uec.ac.jp, shinjs@jcom.home.ne.jp).

<sup>‡</sup>Department of Applied Mathematics, Okayama University of Science, Ridai, Okayama, Okaya 700 Japan (murakami@youhei.xmath.ous.ac.jp).

Recently, we developed a theory of positive systems for some new classes of linear systems such as positive linear functional differential equations [37], [35], [32], positive linear functional difference equations [38], and positive linear Volterra integral equations [29]. More precisely, we introduced various notions of positive systems for these classes of systems. Then we offered explicit criteria for them in terms of positivity of system matrices. Furthermore, we gave some extensions of the classical Perron–Frobenius theorems which are important tools for analyzing stability and robust stability of these positive systems. Finally, we obtained new criteria for asymptotic stability of positive systems. For example, in the recent paper [37], we showed that a linear functional differential equation of the form

$$(1) \quad \dot{x}(t) = Ax(t) + \int_{-h}^0 d[\eta(\theta)]x(t+\theta), \quad x(t) \in \mathbb{R}^n, \quad t \geq 0,$$

is positive (meaning that its solution semigroup is positive) if and only if  $A$  is a Metzler matrix and  $\eta$  is an increasing matrix function. Then such a positive equation is exponentially stable if and only if the spectral abscissa of the matrix  $A + \eta(0)$  is strictly less than zero. Moreover, stability radius problems of positive linear functional differential equations (1) under multiperturbations or affine perturbations have been studied in [35], where the explicit formulae for the stability radii are given.

In the present paper, we first introduce the notion of positive linear Volterra integrodifferential equations. Then we give an explicit criterion for equations of this class. Finally, we study stability and robust stability of positive equations. It is important to note that Volterra equations are studied extensively in many various areas such as control theory, optimization, probability and statistics, economics, etc. In particular, problems of stability and robust stability of Volterra equations have been studied quite some time; see, e.g., [3], [4], [11], [19], [20], [26], [27], [47], [48]. However, to the best of our knowledge, aspects of positivity of problems of Volterra equations have not been exploited yet in the literature and the main purpose of this paper is to fill this gap. This paper is motivated by a series of our works on the problems of stability and robust stability of positive linear systems; see, e.g., [15], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [45], [46].

The organization of the paper is as follows. In the next section, we summarize some notation and preliminary results which will be used in what follows. In section 3, we give an explicit criterion and a Perron–Frobenius theorem for positive linear Volterra integrodifferential equations. In the next section, we offer a new criterion for uniformly asymptotic stability of positive equations. In section 5, we study stability radii of positive linear Volterra integrodifferential equations. It is proved that complex, real, and positive stability radii of positive linear Volterra equations under structured perturbations (or affine perturbations) coincide and can be computed by explicit formulae. Finally, in section 6, we briefly summarize the obtained results and give a suggestion for further research of positive Volterra differential equations with delays and positive Volterra–Stieltjes differential equations.

**2. Preliminaries.** In this section we shall define some notation and recall some well-known results which will be used in subsequent sections. Let  $\mathbb{K} = \mathbb{C}$  or  $\mathbb{R}$ , where  $\mathbb{C}$  and  $\mathbb{R}$  denote the sets of all complex and all real numbers, respectively. Let us denote  $\Re z$  the real part of  $z \in \mathbb{C}$ . For an integer  $l, q \geq 1$ ,  $\mathbb{K}^l$  denotes the  $l$ -dimensional vector space over  $\mathbb{K}$ ,  $(\mathbb{K}^l)^*$  is its dual, and  $\mathbb{K}^{l \times q}$  stands for the set of all  $l \times q$  matrices with entries in  $\mathbb{K}$ . Inequalities between real matrices or vectors will be understood componentwise; i.e., for two real matrices  $A = (a_{ij})$  and  $B = (b_{ij})$

in  $\mathbb{R}^{l \times q}$ , we write  $A \geq B$  if and only if  $a_{ij} \geq b_{ij}$  for  $i = 1, \dots, l$ ,  $j = 1, \dots, q$ . In particular, if  $a_{ij} > b_{ij}$  for  $i = 1, \dots, l$ ,  $j = 1, \dots, q$ , then we write  $A \gg B$  instead of  $A \geq B$ . We denote by  $\mathbb{R}_+^{l \times q}$  the set of all nonnegative matrices  $A \geq 0$ . A similar notation is adopted for vectors. For  $x \in \mathbb{K}^n$  and  $P \in \mathbb{K}^{l \times q}$  we define  $|x| = (|x_i|)$  and  $|P| = (|p_{ij}|)$ . For any matrix  $A \in \mathbb{K}^{n \times n}$  the *spectral radius* and *spectral abscissa* of  $A$  are denoted by  $\rho(A) = \max\{|\lambda|; \lambda \in \sigma(A)\}$  and  $\mu(A) = \max\{\Re \lambda; \lambda \in \sigma(A)\}$ , where  $\sigma(A) := \{s \in \mathbb{C}; \det(sI_n - A) = 0\}$  is the spectrum of  $A$ . A matrix  $A \in \mathbb{R}^{n \times n}$  is called a *Metzler matrix* if all off-diagonal elements of  $A$  are nonnegative or, equivalently,  $tI_n + A \geq 0$  for some  $t \geq 0$ . It is clear that any  $A \in \mathbb{R}_+^{n \times n}$  is a Metzler matrix and, moreover,  $\rho(A) = \mu(A)$ .

A norm  $\|\cdot\|$  on  $\mathbb{K}^n$  is said to be *monotonic* if  $\|x\| \leq \|y\|$  whenever  $|x| \leq |y|$ ,  $x, y \in \mathbb{K}^n$ . Every  $p$ -norm on  $\mathbb{K}^n$ ,  $1 \leq p \leq \infty$ , is monotonic. Throughout the paper, if not otherwise stated, the norm of a matrix  $P \in \mathbb{K}^{l \times q}$  is understood as its operator norm associated with a given pair of monotonic vector norms on  $\mathbb{K}^l$  and  $\mathbb{K}^q$ , that is,  $\|P\| = \max\{\|Py\|; \|y\| = 1\}$ . We note that the operator norm is in general not monotonic norm on  $\mathbb{K}^{l \times q}$  even if  $\mathbb{K}^l, \mathbb{K}^q$  are provided with monotonic norms. However, such monotonicity holds for nonnegative matrices. Moreover, we have (see, e.g., [44])

$$(2) \quad P \in \mathbb{K}^{l \times q}, Q \in \mathbb{R}_+^{l \times q}, |P| \leq Q \quad \Rightarrow \quad \|P\| \leq \| |P| \| \leq \|Q\|.$$

The following theorem summarizes some existing results on properties of Metzler matrices which will be used in what follows (see, e.g., [44]).

**THEOREM 2.1.** *Suppose that  $A \in \mathbb{R}^{n \times n}$  is a Metzler matrix. Then*

- (i) (Perron–Frobenius)  $\mu(A)$  is an eigenvalue of  $A$  and there exists a nonnegative eigenvector  $x \geq 0$ ,  $x \neq 0$  such that  $Ax = \mu(A)x$ ;
- (ii) given  $\alpha \in \mathbb{R}$ , there exists a nonzero vector  $x \geq 0$  such that  $Ax \geq \alpha x$  if and only if  $\mu(A) \geq \alpha$ ;
- (iii)  $(tI_n - A)^{-1}$  exists and is nonnegative if and only if  $t > \mu(A)$ ;
- (iv) given  $B \in \mathbb{R}_+^{n \times n}$ ,  $C \in \mathbb{C}^{n \times n}$ , then

$$(3) \quad |C| \leq B \quad \Rightarrow \quad \mu(A + C) \leq \mu(A + B).$$

For  $\phi \in C([\alpha, \beta], \mathbb{R}^{m \times n})$ , the notation  $\phi \geq 0$  means that  $\phi(\theta) \geq 0$  for every  $\theta \in [\alpha, \beta]$ . To make the presentation self-contained we present here some basic facts on vector functions of bounded variation and relative ones.

A matrix function  $\eta(\cdot) : [\alpha, \beta] \rightarrow \mathbb{R}^{l \times q}$  is called an increasing matrix function if

$$\eta(\theta_2) \geq \eta(\theta_1) \quad \text{for } \alpha \leq \theta_1 \leq \theta_2 \leq \beta.$$

A matrix function  $\eta(\cdot) : [\alpha, \beta] \rightarrow \mathbb{K}^{m \times n}$  is said to be of bounded variation if

$$(4) \quad \text{Var}(\eta; \alpha, \beta) := \sup_{P[\alpha, \beta]} \sum_k \|\eta(\theta_k) - \eta(\theta_{k-1})\| < +\infty,$$

where the supremum is taken over the set of all finite partitions of the interval  $[\alpha, \beta]$ . The set  $BV([\alpha, \beta], \mathbb{K}^{m \times n})$  of all matrix functions  $\eta(\cdot)$  of bounded variation on  $[\alpha, \beta]$  satisfying  $\eta(\alpha) = 0$  is a Banach space endowed with the norm  $\|\eta\| = \text{Var}(\eta; \alpha, \beta)$ . Since all matrix norms on  $\mathbb{K}^{m \times n}$  are equivalent, it follows that the matrix function  $\eta(\cdot) = (\eta_{ij}(\cdot)) \in \mathbb{K}^{m \times n}$  is of bounded variation if and only if each  $\eta_{ij}(\cdot)$  is of bounded variation. Moreover, it is easy to show that if  $\mathbb{K}^{m \times n}$  is provided with the  $\infty$ -norm, then

$$(5) \quad \max_{1 \leq i \leq m} \sum_{j=1}^n \text{Var}(\eta_{ij}; \alpha, \beta) \leq \|\eta\| = \text{Var}(\eta; \alpha, \beta) \leq \sum_{i=1}^m \sum_{j=1}^n \text{Var}(\eta_{ij}; \alpha, \beta).$$

Given  $\eta(\cdot) \in BV([\alpha, \beta], \mathbb{K}^{m \times n})$ , then for any continuous functions  $\gamma \in C([\alpha, \beta], \mathbb{K})$  and  $\phi \in C([\alpha, \beta], \mathbb{K}^n)$ , the integrals

$$\int_{\alpha}^{\beta} \gamma(\theta) d[\eta(\theta)] \quad \text{and} \quad \int_{\alpha}^{\beta} d[\eta(\theta)] \phi(\theta)$$

exist and are defined, respectively, as the limits of  $S_1(P) := \sum_{k=1}^p \gamma(\zeta_k)(\eta(\theta_k) - \eta(\theta_{k-1}))$  and  $S_2(P) := \sum_{k=1}^p (\eta(\theta_k) - \eta(\theta_{k-1}))\phi(\zeta_k)$  as  $d(P) := \max_k |\theta_k - \theta_{k-1}| \rightarrow 0$ , where  $P = \{\theta_1 = \alpha \leq \theta_2 \leq \dots \leq \theta_p = \beta\}$  is any finite partition of the interval  $[\alpha, \beta]$  and  $\zeta_k \in [\theta_{k-1}, \theta_k]$ . It is immediate from the definition that

$$(6) \quad \left\| \int_{\alpha}^{\beta} \gamma(\theta) d[\eta(\theta)] \right\| \leq \max_{\theta \in [\alpha, \beta]} |\gamma(\theta)| \|\eta\|,$$

$$\left\| \int_{\alpha}^{\beta} d[\eta(\theta)] \phi(\theta) \right\| \leq \max_{\theta \in [\alpha, \beta]} \|\phi(\theta)\| \|\eta\|.$$

Let  $\mathbb{K}^n$  be endowed with a vector norm  $\|\cdot\|$  and let  $C([\alpha, \beta], \mathbb{K}^n)$  be the Banach space of all continuous functions on  $[\alpha, \beta]$  with values in  $\mathbb{K}^n$  normed by the maximum norm  $\|\phi\| = \max_{\theta \in [\alpha, \beta]} \|\phi(\theta)\|$ . Let  $L : C([\alpha, \beta], \mathbb{K}^n) \rightarrow \mathbb{K}^n$  be a linear bounded operator. Then, by the Riesz representation theorem, there exists a unique matrix function  $\eta = (\eta_{ij}(\cdot)) \in BV([\alpha, \beta], \mathbb{K}^{n \times n})$  which is *continuous from the left* (c.f.l.) on  $(\alpha, \beta)$  such that

$$(7) \quad L\phi = \int_{\alpha}^{\beta} d[\eta(\theta)] \phi(\theta) \quad \forall \phi \in C([\alpha, \beta], \mathbb{K}^n).$$

For any vector norm on  $\mathbb{K}^n$ , we have by (6),  $\|L\| \leq \|\eta\|$ . Moreover, if  $\mathbb{K}^n$  is provided with the  $\infty$ -norm so that  $\forall x \in \mathbb{K}^n$  and  $\forall \theta \in [\alpha, \beta]$

$$\|x\| = \max_{1 \leq i \leq n} |x_i| \quad \text{and} \quad \|\eta(\theta)\| = \max_{1 \leq i \leq n} \sum_{j=1}^n |\eta_{ij}(\theta)|,$$

then it can be shown immediately that  $\|L\| = \|\eta\|$ . Let  $X$  be a subspace of  $C([\alpha, \beta], \mathbb{R}^n)$ . Then the operator  $L$  is called positive on  $X$  if  $L\phi \geq 0$  for every  $\phi \in X$ ,  $\phi \geq 0$ .

In subsequent sections the following subspace of  $BV([\alpha, \beta], \mathbb{K}^{m \times n})$  will be used:  $NBV([\alpha, \beta], \mathbb{K}^{m \times n}) := \{\eta \in BV([\alpha, \beta], \mathbb{K}^{m \times n}) : \eta \text{ is c.f.l. on } [\alpha, \beta]\}$ . It is clear that  $NBV([\alpha, \beta], \mathbb{K}^{m \times n})$  is closed in  $BV([\alpha, \beta], \mathbb{K}^{m \times n})$  and thus it is a Banach space with the norm  $\|\delta\| = \text{Var}(\delta; \alpha, \beta)$ .

To end this section, we give below a list of notation:

$\mathbb{R}$	$(-\infty, +\infty)$ ;
$\mathbb{C}$	the complex plane;
$\mathbb{K}$	either $\mathbb{R}$ or $\mathbb{C}$ ;
$\mathbb{N}$	the set of all natural numbers;
$\mathbb{R}^n$	the set of $n$ -dimensional column vectors with real entries;
$\mathbb{C}^n$	the set of $n$ -dimensional column vectors with complex entries;
$\mathbb{R}_+^n$	the set of $n$ -dimensional column vectors with nonnegative entries;
$\mathbb{R}^{m \times n}$	the set of $m \times n$ -dimensional matrices with real entries;
$\mathbb{C}^{m \times n}$	the set of $m \times n$ -dimensional matrices with complex entries;
$\mathbb{R}_+^{m \times n}$	the set of $m \times n$ -dimensional matrices with nonnegative entries;
$\Re z$	the real part of $z \in \mathbb{C}$ ;
$A, B, C$ , etc.	matrices;



$I_n$	the identity of $\mathbb{C}^{n \times n}$ ;
$x, y, z, \text{ etc.}$	column vectors;
$\mu(A)$	the spectral abscissa of $A \in \mathbb{C}^{n \times n}$ ;
$A \geq B$	$A, B \in \mathbb{R}^{m \times n}$ and $A - B \in \mathbb{R}_+^{m \times n}$ for some $m, n \in \mathbb{N}$ ;
$C([0, +\infty), \mathbb{K}^{m \times n})$	the vector space of all continuous functions on $[0, +\infty)$ with values in $\mathbb{K}^{m \times n}$ ;
$C([\alpha, \beta], \mathbb{K}^{m \times n})$	the Banach space of all continuous functions on $[\alpha, \beta]$ with values in $\mathbb{K}^{m \times n}$ , endowed with the maximum norm;
$BV([\alpha, \beta], \mathbb{K}^{m \times n})$	the Banach space of all matrix functions $\eta(\cdot)$ of bounded variation on $[\alpha, \beta]$ with values in $\mathbb{K}^{m \times n}$ satisfying $\eta(\alpha) = 0$ , endowed with the norm $\ \eta\  = \text{Var}(\eta; \alpha, \beta)$ ;
$NBV([\alpha, \beta], \mathbb{K}^{m \times n})$	functions in $BV([\alpha, \beta], \mathbb{K}^{m \times n})$ that are c.f.l. on $[\alpha, \beta]$ ;
$L^1([0, +\infty), \mathbb{K}^{m \times n})$	the Banach space of $L^1$ -integrable matrix functions on $[0, +\infty)$ , with values in $\mathbb{K}^{m \times n}$ and endowed with the $L^1$ -norm;
$\mu(A, B(\cdot))$	the spectral abscissa of a Volterra differential equation of convolution type defined by $A \in \mathbb{C}^{n \times n}$ and $B(\cdot) \in C([0, +\infty), \mathbb{C}^{n \times n})$ .

### 3. Positive linear Volterra integrodifferential equations.

**3.1. Explicit criterion for positive linear Volterra integrodifferential equations.** Consider a linear Volterra integrodifferential equation of convolution type

$$(8) \quad \dot{x}(t) = Ax(t) + \int_0^t B(t-s)x(s)ds,$$

where  $A \in \mathbb{R}^{n \times n}$  is a given matrix and  $B : [0, +\infty) \rightarrow \mathbb{R}^{n \times n}$  is a given continuous matrix function.

**DEFINITION 3.1.** Let  $\sigma \geq 0$  and  $\phi \in C([0, \sigma], \mathbb{R}^n)$  be given. A function  $x : [0, +\infty) \rightarrow \mathbb{R}^n$  such that

$$(9) \quad x(t) = \phi(t), \quad t \in [0, \sigma],$$

and fulfilling (8) for every  $t \geq \sigma$  is called a solution of (8) with the initial condition (9).

It is well known that for every  $\sigma \geq 0$  and  $\phi \in C([0, \sigma], \mathbb{R}^n)$ , (8) has a unique solution satisfying the initial condition (9); see, e.g., [4, p. 177]. We denote it by  $x(t, \sigma, \phi), t \geq 0$ .

**DEFINITION 3.2.** We say that (8) is positive if for every  $\sigma \geq 0$  and every  $\phi \in C([0, \sigma], \mathbb{R}^n)$ , with  $\phi \geq 0$ , the corresponding solution  $x(t, \sigma, \phi)$  satisfies  $x(t, \sigma, \phi) \geq 0 \forall t \geq \sigma$ .

**Remark 3.3.** The notion of positive linear Volterra equations given in Definition 3.2 is similar to that of positive linear functional differential equations of the form (1); see, e.g., [35], [37]. However, it is worth noting that in general a Volterra equation of the form (8) cannot convert into a linear functional differential equation of the form (1).

We believe that positive linear Volterra equations are interesting objects not only in mathematics but also in other sciences such as economics, physics, and biology. To prove a criterion for positive linear Volterra equations, we need the following technical lemmas.

**LEMMA 3.4.** Let  $T > 0$  and  $C_0([0, T], \mathbb{R}^n) := \{\phi \in C([0, T], \mathbb{R}^n) : \phi(T) = 0\}$ . Suppose that the linear operator  $L$  is defined by

$$L : C_0([0, T], \mathbb{R}^n) \rightarrow \mathbb{R}^n, \quad \phi \mapsto L\phi = \int_0^T d[\eta(\theta)]\phi(\theta),$$

where  $\eta \in NBV([0, T], \mathbb{R}^{n \times n})$  is given. Then  $L$  is a positive operator if and only if  $\eta$  is an increasing matrix function.

*Proof.* Let  $\eta$  be an increasing matrix function. Then by the definition of Riemann–Stieltjes integrals, we have

$$L\phi = \lim_{d(P) \rightarrow 0} \sum_{k=1}^p (\eta(\theta_k) - \eta(\theta_{k-1}))\phi(\zeta_k) \geq 0$$

for every  $\phi \in C_0([0, T], \mathbb{R}^n)$ ,  $\phi \geq 0$ . This means that  $L$  is positive.

Conversely, assume that  $L$  is positive on  $C_0([0, T], \mathbb{R}^n)$ . Let  $\eta(\cdot) = (\eta_{ij}(\cdot))$ . We show that  $\eta_{ij}(\cdot) \in NBV([0, T], \mathbb{R})$  is an increasing scalar function for every  $i, j \in \{1, 2, \dots, n\}$ . Since  $L$  is positive, it is easy to see that the operator

$$L_{ij} : C_0([0, T], \mathbb{R}) \rightarrow \mathbb{R}, \quad \phi \mapsto L_{ij}\phi := \int_0^T \phi(\theta) d[\eta_{ij}(\theta)]$$

is also positive for every  $i, j \in \{1, 2, \dots, n\}$ . Fix  $\theta_1, \theta_2 \in (0, T)$ ,  $\theta_1 < \theta_2$  and  $k \in \mathbb{N}$ ,  $k > \max\{\frac{1}{\theta_1}, \frac{1}{\theta_2}, \frac{1}{\theta_2 - \theta_1}\}$ , and consider the continuous function  $\phi_k$  defined by

$$(10) \quad \phi_k(\theta) := \begin{cases} 0 & \text{if } \theta \in [0, \theta_1 - \frac{1}{k}], \\ k\theta + 1 - k\theta_1 & \text{if } \theta \in (\theta_1 - \frac{1}{k}, \theta_1], \\ 1 & \text{if } \theta \in (\theta_1, \theta_2 - \frac{1}{k}], \\ -k\theta + k\theta_2 & \text{if } \theta \in (\theta_2 - \frac{1}{k}, \theta_2], \\ 0 & \text{if } \theta \in (\theta_2, T]. \end{cases}$$

Since  $\phi_k$  is a continuous on  $[0, T]$ , it follows from a standard property of Riemann–Stieltjes integrals that

$$\int_0^T \phi_k(\theta) d[\eta_{ij}(\theta)] = \left( \int_0^{\theta_1 - \frac{1}{k}} + \int_{\theta_1 - \frac{1}{k}}^{\theta_1} + \int_{\theta_1}^{\theta_2 - \frac{1}{k}} + \int_{\theta_2 - \frac{1}{k}}^{\theta_2} + \int_{\theta_2}^T \right) \phi_k(\theta) d[\eta_{ij}(\theta)];$$

see, e.g., [41]. This gives

$$\int_{\theta_1 - \frac{1}{k}}^{\theta_1} \phi_k(\theta) d[\eta_{ij}(\theta)] + \eta_{ij}\left(\theta_2 - \frac{1}{k}\right) - \eta_{ij}(\theta_1) + \int_{\theta_2 - \frac{1}{k}}^{\theta_2} \phi_k(\theta) d[\eta_{ij}(\theta)] \geq 0$$

for every  $k \in \mathbb{N}$  large enough. Taking into account that  $\eta_{ij}$  is c.f.l. at  $\theta_1, \theta_2$  and letting  $k \rightarrow +\infty$ , we have  $\eta_{ij}(\theta_2) \geq \eta_{ij}(\theta_1)$  for every  $\theta_1, \theta_2 \in (0, T)$ ,  $\theta_2 \geq \theta_1$ . In the case of  $\theta_1 = 0 < \theta_2 < T$ , in a similar way we also get  $\eta_{ij}(\theta_2) \geq \eta_{ij}(\theta_1)$ . Finally, since  $\eta_{ij}$  is c.f.l. at  $T$ , we have  $\eta_{ij}(T) \geq \eta_{ij}(\theta)$  for  $\theta \in [0, T]$ . This completes our proof.  $\square$

Let  $h : [0, +\infty) \rightarrow \mathbb{R}$ . Then the Laplace transform of  $h$  is formally defined to be

$$\hat{h}(s) := \int_0^{+\infty} e^{-st} h(t) dt.$$

If  $\beta \in \mathbb{R}$  and  $\int_0^{+\infty} e^{-\beta t} |h(t)| dt < +\infty$ , then  $\hat{h}(s)$  exists for  $s \in \mathbb{C}$ ,  $\Re s \geq \beta$ . Furthermore,  $\hat{h}(s)$  is an analytic function in the domain  $\{s \in \mathbb{C} : \Re s > \beta\}$ . If  $D(t) = (d_{ij}(t))$  is a matrix function, then we define

$$\hat{D} := (\hat{d}_{ij}).$$

In the rest of this paper, we always assume that

$$(11) \quad B(\cdot) \in L^1([0, +\infty), \mathbb{R}^{n \times n}).$$

Let  $Z(t)$  be the matrix whose columns are solutions of (8) with  $Z(0) = I_n$ . Then  $Z(t)$  satisfies the resolvent equation

$$(12) \quad \frac{d}{dt}Z(t) = AZ(t) + \int_0^t B(t-s)Z(s)ds, \quad Z(0) = I_n.$$

LEMMA 3.5. *Suppose that for every  $x_0 \in \mathbb{R}_+^n$ , the corresponding solution  $x(t, 0, x_0)$ ,  $t \geq 0$ , satisfies  $x(t, 0, x_0) \geq 0 \forall t \geq 0$ . Then  $A \in \mathbb{R}^{n \times n}$  is a Metzler matrix and  $(sI_n - A - \hat{B}(s))^{-1} \geq 0$  for  $s \in \mathbb{R}$  large enough.*

*Proof.* By the assumption,  $Z(t) \geq 0 \forall t \geq 0$ . It is well known that  $Z(\cdot)$  is of exponential order; see [3, p. 29]. Therefore, taking the Laplace transform of both sides of (12), we get

$$[sI_n - A - \hat{B}(s)]\hat{Z}(s) = Z(0) = I_n$$

for  $s \in \mathbb{R}$  large enough. From  $Z(t) \geq 0 \forall t \geq 0$ , it follows that  $\hat{Z}(s) = (sI_n - A - \hat{B}(s))^{-1} \geq 0$  for  $s \in \mathbb{R}$  large enough. It remains only to show that  $A$  is a Metzler matrix. Let  $A = (a_{ij})$  and assume to the contrary that  $a_{i_0 j_0} < 0$  for some  $i_0 \neq j_0$ . Since  $B(t) \in L^1([0, +\infty), \mathbb{R}^{n \times n})$ , it is easy to see that  $\hat{B}(s) \rightarrow 0$  as  $s \rightarrow +\infty$ . This implies that the spectral radius of the matrix  $s^{-1}(A + \hat{B}(s))$  is strictly less than 1 for  $s > 0$  large enough. Therefore, we can represent

$$\begin{aligned} (sI_n - A - \hat{B}(s))^{-1} &= s^{-1} \left( I_n - s^{-1}(A + \hat{B}(s)) \right)^{-1} \\ &= s^{-1}I_n + s^{-2}(A + \hat{B}(s)) + \sum_{k=2}^{+\infty} s^{-(k+1)}(A + \hat{B}(s))^k \end{aligned}$$

for  $s > 0$  large enough. We thus get

$$(13) \quad sI_n + (A + \hat{B}(s)) + \sum_{k=2}^{+\infty} s^{-(k-1)}(A + \hat{B}(s))^k \geq 0$$

for  $s > 0$  large enough. It is important to note that

$$\lim_{s \rightarrow +\infty} \sum_{k=2}^{+\infty} s^{-(k-1)}(A + \hat{B}(s))^k = 0.$$

Then from (13) it follows that the entry  $\tilde{b}_{i_0 j_0}$  of the matrix on the left-hand side of (13) is negative for  $s > 0$  large enough. This is a contradiction. Hence,  $A$  must be a Metzler matrix. This completes our proof.  $\square$

LEMMA 3.6. (see [3, p. 24]) *Let  $0 < \alpha \leq +\infty$  and suppose that  $f : [0, \alpha) \rightarrow \mathbb{R}^n$  is continuous and that  $D(t, s)$  is an  $n \times n$  matrix of functions continuous for  $0 \leq s \leq t < \alpha$ . If  $0 < T < \alpha$ , then there exists a unique solution  $y(t)$  of the integral equation*

$$(14) \quad y(t) = f(t) + \int_0^t D(t, s)y(s)ds$$

on  $[0, T]$ . Moreover,  $y(t)$  is the limit in  $C([0, T], \mathbb{R}^n)$  of the sequence of Picard's successive approximations  $(y_n(t))_n$ , given by

$$(15) \quad y_1(t) = f(t); \quad y_{n+1}(t) = f(t) + \int_0^t D(t, s)y_n(s)ds, \quad t \in [0, T], n \geq 1.$$

We are now in the position to prove the first main result of this paper.

**THEOREM 3.7.** *Equation (8) is positive if and only if  $A \in \mathbb{R}^{n \times n}$  is a Metzler matrix and  $B(t) \geq 0$  for every  $t \geq 0$ .*

*Proof.* Assume that  $A \in \mathbb{R}^{n \times n}$  is a Metzler matrix and  $B(t) \geq 0$  for every  $t \geq 0$ . Let  $\sigma \geq 0$  and  $\phi \in C([0, \sigma], \mathbb{R}^n)$ ,  $\phi \geq 0$ . Denote  $x(t) = x(t, \sigma, \phi)$ ,  $t \geq 0$ . Then  $x(t)$  satisfies

$$x(t) = e^{A(t-\sigma)}\phi(\sigma) + \int_{\sigma}^t e^{A(t-s)}g(s)ds, \quad t \geq \sigma,$$

where  $g(s) := \int_0^s B(s-\tau)x(\tau)d\tau$ ,  $s \in [\sigma, t]$ . This is equivalent to

$$x(t+\sigma) = e^{At}\phi(\sigma) + \int_0^t e^{A(t-s)}g(s+\sigma)ds, \quad t \geq 0.$$

Therefore,

$$x(t+\sigma) = e^{At}\phi(\sigma) + \int_0^t e^{A(t-s)} \left( \int_0^{\sigma} B(s+\sigma-\tau)\phi(\tau)d\tau + \int_{\sigma}^{\sigma+s} B(s+\sigma-\tau)x(\tau)d\tau \right) ds, \quad t \geq 0.$$

This gives

$$(16) \quad x(t+\sigma) = f(t) + \int_0^t \left( e^{A(t-s)} \int_0^s B(s-\tau)x(\tau+\sigma)d\tau \right) ds, \quad t \geq 0,$$

where

$$f(t) = e^{At}\phi(\sigma) + \int_0^t \left( e^{A(t-s)} \int_0^{\sigma} B(s+\sigma-\tau)\phi(\tau)d\tau \right) ds, \quad t \geq 0.$$

Set  $y(t) = x(t+\sigma)$ ,  $t \geq 0$ . Then (16) can be rewritten in the form

$$y(t) = f(t) + \int_0^t \left( e^{A(t-s)} \int_0^s B(s-\tau)y(\tau)d\tau \right) ds, \quad t \geq 0.$$

Interchanging the order of integration in the last term, we get

$$(17) \quad y(t) = f(t) + \int_0^t \left( \int_s^t e^{A(t-\tau)}B(\tau-s)d\tau \right) y(s)ds, \quad t \geq 0.$$

Since  $A \in \mathbb{R}^{n \times n}$  is a Metzler matrix, it follows that  $A + \alpha I_n \geq 0$  for some  $\alpha > 0$ . This implies that  $e^{(A+\alpha I_n)t} = e^{\alpha t}e^{At} \geq 0 \quad \forall t \geq 0$ . Thus,  $e^{At} \geq 0 \quad \forall t \geq 0$ . From  $e^{At}$ ,  $B(t) \geq 0$  for every  $t \geq 0$ , it follows that  $f(t) \geq 0 \quad \forall t \geq 0$  and  $D(t, s) := \int_s^t e^{A(t-\tau)}B(\tau-s)d\tau \geq 0$ ,  $0 \leq s \leq t \leq T$ . Furthermore, the functions  $y_n(t)$ ,  $n \in \mathbb{N}$ , defined by (15) in Lemma 3.6 are also nonnegative for every  $t \geq 0$ . Applying Lemma 3.6 to (17), we conclude that the solution  $y(t)$ ,  $t \in [0, T]$ , of integral equation (17) is nonnegative on  $[0, T]$ . Since  $T > 0$  is arbitrary, it implies that  $y(t) \geq 0 \quad \forall t \geq 0$ . Thus,  $x(t) \geq 0 \quad \forall t \geq 0$ .

Conversely, assume that (8) is positive. By Lemma 3.5, it remains only to show that  $B(t) \geq 0$  for every  $t \geq 0$ . Fix  $T > 0$  and let  $\phi \in C_0([0, T], \mathbb{R}^n)$ ,  $\phi \geq 0$ . Since (8) is positive, it follows that

$$\frac{d}{dt}x(t)\Big|_{t=T} = \lim_{t \rightarrow T+0} \frac{x(t) - x(T)}{t - T} = \lim_{t \rightarrow T+0} \frac{x(t)}{t - T} \geq 0.$$

This implies that

$$\frac{d}{dt}x(t)\Big|_{t=T} = Ax(T) + \int_0^T B(T-s)x(s)ds = \int_0^T B(T-s)\phi(s)ds \geq 0$$

for every  $\phi \in C_0([0, T], \mathbb{R}^n)$ ,  $\phi \geq 0$ . Thus, the linear operator defined by

$$L : C_0([0, T], \mathbb{R}^n) \rightarrow \mathbb{R}^n, \quad \phi \mapsto L\phi := \int_0^T B(T-s)\phi(s)ds$$

is a positive operator. Applying Lemma 3.4 to the positive operator  $L$ , we conclude that the function

$$\eta(t) = \int_0^t B(T-s)ds, \quad t \in [0, T],$$

is an increasing matrix function. This gives  $B(t) \geq 0$  for every  $t \in [0, T]$ . Since  $T > 0$  is arbitrary, it follows that  $B(t) \geq 0$  for every  $t \geq 0$ . This completes our proof.  $\square$

**3.2. A Perron–Frobenius theorem for positive linear Volterra integro-differential equations.** It is well known that Perron–Frobenius-type theorems are principal tools for analysis of stability and robust stability of positive linear time-invariant systems. To our knowledge, there is a large number of extensions of classical Perron–Frobenius theorems; see; e.g., [1], [6], [21], [31], [32], [36], [42], and the references therein.

Recall that a linear time-invariant differential system of the form  $\dot{x}(t) = Ax(t)$ ,  $t \geq 0$ , is positive if and only if the system matrix  $A \in \mathbb{R}^{n \times n}$  is a Metzler matrix. Therefore, the classical Perron–Frobenius theorem (Theorem 2.1) can be seen as the Perron–Frobenius theorem for the class of these positive systems. From this dynamic point of view, we recently presented some extensions of the classical Perron–Frobenius theorem such as the Perron–Frobenius theorem for positive linear higher order difference equations [36], the Perron–Frobenius theorem for positive linear time-delay systems [31], and the Perron–Frobenius theorem for positive linear functional differential equations [32].

In this subsection, we give a new Perron–Frobenius theorem for positive linear Volterra integrodifferential equations of the form (8). Let us denote

$$(18) \quad H(s) := (sI_n - A - \hat{B}(s)) = \left( sI_n - A - \int_0^{+\infty} e^{-st} B(t)dt \right)$$

for appropriate  $s \in \mathbb{C}$ . Let us define

$$(19) \quad \mu(A, B(\cdot)) := \sup \left\{ \Re s : \int_0^{+\infty} e^{-\Re st} \|B(t)\|dt < +\infty, \det H(s) = 0 \right\},$$

where  $\mu(A, B(\cdot)) := -\infty$  if  $\det H(s) \neq 0 \forall s \in \mathbb{C}$ ,  $\int_0^{+\infty} e^{-\Re st} \|B(t)\|dt < +\infty$ . Then  $\mu(A, B(\cdot))$  is called *spectral abscissa* of the Volterra equation (8).

**THEOREM 3.8.** *Suppose that the Volterra equation (8) is positive. Let  $\beta := \inf\{\gamma \in \mathbb{R} : \int_0^{+\infty} e^{-\gamma t} \|B(t)\| dt < +\infty\}$  and  $\alpha \in (\beta, +\infty)$  be given. Denote by  $\mu_0 := \mu(A, B(\cdot))$ . If  $\mu_0 > -\infty$ , then*

- (i) *(Perron–Frobenius theorem for positive linear Volterra integrodifferential equations)  $\mu_0$  is a root of the characteristic equation, that is,  $\det H(\mu_0) = 0$ ; moreover, there exists a nonzero vector  $x \geq 0$  such that*

$$\left(A + \int_0^{+\infty} e^{-\mu_0 t} B(t) dt\right)x = \mu_0 x;$$

- (ii) *there exists a nonzero vector  $x \geq 0$  such that  $(A + \int_0^{+\infty} e^{-\alpha t} B(t) dt)x \geq \alpha x$  if and only if  $\mu_0 \geq \alpha$ ;*

- (iii) *the matrix  $H(\alpha)^{-1} = (\alpha I_n - A - \int_0^{+\infty} e^{-\alpha t} B(t) dt)^{-1}$  exists and is nonnegative if and only if  $\alpha > \mu_0$ .*

*Proof.* We first show that  $\mu_0$  is finite. Since (8) is positive, it follows that  $A$  is a Metzler matrix and  $B(t) \in \mathbb{R}_+^{n \times n} \forall t \geq 0$ . Assume that  $\beta > -\infty$  and  $\det H(s) = 0$  for some  $s \in \mathbb{C}, \Re s \geq \beta + 1$ . Then, by Theorem 2.1(iv),  $\Re s \leq \mu(A + \int_0^{+\infty} e^{-st} B(t) dt) \leq \mu(A + \int_0^{+\infty} e^{-\Re s t} B(t) dt) \leq \mu(A + \int_0^{+\infty} e^{-(\beta+1)t} B(t) dt)$ . It follows that

$$\mu_0 \leq \max \left\{ \beta + 1, \mu \left( A + \int_0^{+\infty} e^{-(\beta+1)t} B(t) dt \right) \right\}.$$

If  $\beta = -\infty$ , then by a similar argument, we get  $\mu_0 \leq \mu(A + \int_0^{+\infty} B(t) dt)$ . Next, we prove that

$$(20) \quad \mu_0 \leq \mu \left( A + \int_0^{+\infty} e^{-\mu_0 t} B(t) dt \right).$$

In fact, by the assumption  $\mu_0 > -\infty$ , there exists  $s_0 \in \mathbb{C}$  such that  $\det H(s_0) = 0$ ,  $\beta \leq \Re s_0 \leq \mu_0$ . If  $\Re s_0 = \mu_0$ , then, using Theorem 2.1(iv) again, we get (20). If  $\beta \leq \Re s_0 < \mu_0$ , then there exists a sequence  $(s_k)_k$  such that  $\det H(s_k) = 0$ ,  $\beta < \Re s_k < \mu_0 \forall k \in \mathbb{N}$  and  $\Re s_k \rightarrow \mu_0$  as  $k \rightarrow \infty$ . Then by Theorem 2.1(iv), we get

$$(21) \quad \Re s_k \leq \mu \left( A + \int_0^{+\infty} e^{-\Re s_k t} B(t) dt \right).$$

Letting  $k \rightarrow \infty$  in (21), we also get (20). We now consider the continuous real function

$$(22) \quad f(\theta) := \theta - \mu \left( A + \int_0^{+\infty} e^{-\theta t} B(t) dt \right),$$

where  $\theta \in [\beta, +\infty)$  if  $\int_0^{+\infty} e^{-\beta t} \|B(t)\| dt < +\infty$ ; otherwise  $\theta \in (\beta, +\infty)$ . Then  $f(\mu_0) \leq 0$  by (20). Assume that  $f(\mu_0) < 0$ . Since, clearly,  $\lim_{\theta \rightarrow +\infty} f(\theta) = +\infty$ , we derive that  $f(\theta_0) = 0$  for some  $\theta_0 > \mu_0$ . This gives  $\theta_0 = \mu(A + \int_0^{+\infty} e^{-\theta_0 t} B(t) dt)$ . By Theorem 2.1(i), it implies that  $\det(\theta_0 I_n - A - \int_0^{+\infty} e^{-\theta_0 t} B(t) dt) = 0$ . However, this conflicts with the definition of  $\mu_0$ . Thus  $f(\mu_0) = 0$ , or, equivalently,  $\mu_0 = \mu(A + \int_0^{+\infty} e^{-\mu_0 t} B(t) dt)$ . Then Theorem 3.8(i) now follows from Theorem 2.1(i).

Moreover, by Theorem 2.1(iv), we get

$$\mu \left( A + \int_0^{+\infty} e^{-\theta_2 t} B(t) dt \right) \leq \mu \left( A + \int_0^{+\infty} e^{-\theta_1 t} B(t) dt \right), \quad \beta < \theta_1 \leq \theta_2.$$

Therefore,  $f$  is strictly increasing on  $(\beta, +\infty)$ . Moreover,  $f(\mu_0) = 0$ . Now, it is easy to see that Theorem 3.8(ii), (iii) follow from Theorem 2.1(ii), (iii), respectively.  $\square$

**4. An explicit criterion for uniformly asymptotic stability of positive linear Volterra integrodifferential equations.** In this section, we offer a new and novel criterion for uniformly asymptotic stability of positive linear Volterra integrodifferential equations of the form (8). First, we quote here the various notions of stability of zero solution of (8) which are taken from the standard books on theory of Volterra equations; see [3], [4].

DEFINITION 4.1. *The zero solution of (8) is (Lyapunov) stable if, for each  $\epsilon > 0$  and each  $t_0 \geq 0$ , there exists  $\delta > 0$  such that*

$$\phi \in C([0, t_0], \mathbb{R}^n), \quad \|\phi\| < \delta,$$

*implies that  $\|x(t, t_0, \phi)\| < \epsilon \quad \forall t \geq t_0$ .*

DEFINITION 4.2. *The zero solution of (8) is uniformly stable if, for each  $\epsilon > 0$ , there exists  $\delta > 0$  such that*

$$t_0 \geq 0, \phi \in C([0, t_0], \mathbb{R}^n), \quad \|\phi\| < \delta,$$

*implies that  $\|x(t, t_0, \phi)\| < \epsilon \quad \forall t \geq t_0$ .*

DEFINITION 4.3. *The zero solution of (8) is asymptotically stable (AS) if it is stable and if, for each  $t_0 \geq 0$ , there exists  $\delta > 0$  such that  $\phi \in C([0, t_0], \mathbb{R}^n)$ ,  $\|\phi\| < \delta$ , implies that  $x(t, t_0, \phi) \rightarrow 0$  as  $t \rightarrow +\infty$ .*

DEFINITION 4.4. *The zero solution of (8) is uniformly asymptotically stable (UAS) if it is uniformly stable and if there exists  $\delta > 0$  such that, for each  $\epsilon > 0$ , there is a  $T > 0$  such that*

$$t_0 \geq 0, \phi \in C([0, t_0], \mathbb{R}^n), \quad \|\phi\| < \delta,$$

*implies that  $\|x(t, t_0, \phi)\| < \epsilon \quad \forall t \geq t_0 + T$ .*

If the zero solution of (8) is AS (UAS), then we say that (8) is AS (UAS), respectively. Recall that we continue to assume that (11) holds true, that is,  $B(\cdot) \in L^1([0, \infty), \mathbb{R}^{n \times n})$ . Let us denote by  $\sigma(A, B(\cdot))$  the set of all root of the characteristic equation of (8). That is,  $\sigma(A, B(\cdot)) := \{s \in \mathbb{C} : \det H(s) = 0\}$ . Denote  $\mathbb{C}^- := \{s \in \mathbb{C} : \Re s < 0\}$ .

THEOREM 4.5 (see [7]). *Let assumption (11) hold true. Then (8) is UAS if and only if  $\sigma(A, B(\cdot)) \subset \mathbb{C}^-$ .*

The following theorem provides an explicit criterion for uniformly asymptotic stability of positive linear Volterra equations (8).

THEOREM 4.6. *Suppose that (8) is positive. Then (8) is UAS if and only if  $\mu(A + \int_0^{+\infty} B(t)dt) < 0$ .*

*Proof.* Assume on the contrary that (8) is UAS but  $\mu(A + \int_0^{+\infty} B(t)dt) \geq 0$ . Since  $B(\cdot)$  is integrable on  $[0, \infty)$ , it follows that the real function given by (22) is continuous on  $[0, +\infty)$ . By our assumption,  $f(0) \leq 0$ . Since, clearly,  $\lim_{\theta \rightarrow +\infty} f(\theta) = +\infty$ , we have  $f(\theta_0) = 0$  for some  $\theta_0 \geq 0$ . This gives  $\theta_0 = \mu(A + \int_0^{+\infty} e^{-\theta_0 t} B(t)dt)$ . Since  $A + \int_0^{+\infty} e^{-\theta_0 t} B(t)dt$  is a Metzler matrix, it follows from Theorem 2.1(i) that  $\det(\theta_0 I_n - A - \int_0^{+\infty} e^{-\theta_0 t} B(t)dt) = 0$ ,  $\theta_0 \geq 0$ . However, by Theorem 4.5, this conflicts with the fact that (8) is UAS.

Conversely, let  $\mu(A + \int_0^{+\infty} B(t)dt) < 0$ . Assume on the contrary that there exists  $s \in \mathbb{C}$ ,  $\Re s \geq 0$  such that  $\det H(s) = 0$ . This implies that  $\Re s \leq \mu(A + \int_0^{+\infty} e^{-st} B(t)dt)$ . By  $\Re s \geq 0$ , we have

$$\left| \int_0^{+\infty} e^{-st} B(t)dt \right| \leq \int_0^{+\infty} e^{-\Re s t} B(t)dt \leq \int_0^{+\infty} B(t)dt.$$

Then it follows from Theorem 2.1(iv) that

$$0 \leq \mu \left( A + \int_0^{+\infty} e^{-st} B(t) dt \right) \leq \mu \left( A + \int_0^{+\infty} e^{-\Re st} B(t) dt \right) \leq \mu \left( A + \int_0^{+\infty} B(t) dt \right).$$

We reach a contradiction. This completes our proof.  $\square$

*Remark 4.7.* We encountered a scalar version of Theorem 4.6 in some standard books on Volterra equations (see, e.g., [3], [23]), where it was proven by using a Lyapunov function. However, to the best of our knowledge, there is not such a result for  $k$ -dimensional systems with  $k \geq 2$ . As a consequence of the above theorem, we have the following.

**THEOREM 4.8.** *Suppose that (8) is positive and  $b \in \mathbb{R}^n, b \gg 0$ . Then the following statements are equivalent:*

- (i) *equation (8) is AS;*
- (ii) *equation (8) is UAS;*
- (iii)  $\mu(A + \int_0^{+\infty} B(t) dt) < 0$ ;
- (iv) *if  $x(t)$  is a solution of (8) on  $[0, +\infty)$ , then  $x \in L^1([0, +\infty), \mathbb{R}^n)$ ;*
- (v) *there exists a nonzero vector  $p \in \mathbb{R}_+^n$  such that*

$$(23) \quad \left( A + \int_0^{+\infty} B(t) dt \right) p = -b.$$

Moreover, under one of the above conditions, each solution  $y(t)$  of

$$(24) \quad \dot{y}(t) = Ay(t) + \int_0^t B(t-s)y(s)ds + b$$

satisfies  $\lim_{t \rightarrow +\infty} y(t) = p$ .

*Proof.* (i)  $\Leftrightarrow$  (ii) was proved in [7]. By Theorem 4.6, we have (ii)  $\Leftrightarrow$  (iii). Moreover, (ii)  $\Leftrightarrow$  (iv) was found in [26]. We now show that (i)  $\Leftrightarrow$  (v). In fact, the implication (i)  $\Rightarrow$  (v) follows from Theorem 2.1(iii). Conversely, assume on the contrary that (23) holds true but we have  $\mu_0 := \mu(A + \int_0^{+\infty} B(t) dt) \geq 0$ . By Theorem 2.1(i), we get

$$\left( A + \int_0^{+\infty} B(t) dt \right)^T p_0 = \mu_0 p_0$$

for some nonzero vector  $p_0 \in \mathbb{R}_+^n$ . Then it follows from (23) that

$$\mu_0 p^T p_0 + b^T p_0 = 0.$$

However, this conflicts with  $p, p_0 \geq 0$ ,  $p, p_0 \neq 0$ ,  $b \gg 0$ , and  $\mu_0 \geq 0$ . We now assume that (8) is UAS. Then each solution  $y$  of (24) satisfying  $y(0) = y_0$  is given by

$$y(t) = Z(t)y_0 + \int_0^t Z(s)b ds, \quad t \geq 0.$$

It follows that  $\lim_{t \rightarrow +\infty} y(t) = \left( \int_0^{+\infty} Z(t) dt \right) b = \hat{Z}(0)b = \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} b = p$ . This completes our proof.  $\square$

We illustrate the obtained result by a simple example.

*Example 4.9.* Consider a linear positive Volterra integrodifferential equation in  $\mathbb{R}^2$  given by

$$(25) \quad \dot{x}(t) = Ax(t) + \int_0^t B(t-\tau)x(\tau)d\tau, \quad x(t) \in \mathbb{R}^2, t \geq 0,$$



where

$$(26) \quad A = \begin{pmatrix} a & 1 \\ 0 & a \end{pmatrix}; \quad B(t) = \begin{pmatrix} e^{-pt} & 0 \\ \frac{1}{(t+1)^2} & e^{-pt} \end{pmatrix}, \quad t \geq 0,$$

and  $a, p \in \mathbb{R}, p > 0$  are parameters. From  $p > 0$ , it follows that the condition (11) is satisfied. By Theorem 4.6, (25) is UAS if and only if

$$\mu \left( A + \int_0^{+\infty} B(t) dt \right) = \mu \left( \begin{bmatrix} a + 1/p & 1 \\ 1 & a + 1/p \end{bmatrix} \right) = a + 1 + 1/p < 0.$$

Taking into account that  $p > 0$ , it is easy to see that this is equivalent to  $a < -1$ ,  $p > -1/(a+1)$ .

**5. Stability radius of positive linear Volterra integrodifferential equations.** Motivated by many applications in control engineering, researchers have focused much attention on problems of robust stability of dynamical systems over the last twenty years. In the study of these problems, the notion of the stability radius was proved to be an effective tool. By definition, the stability radius of an AS linear differential system  $\dot{x}(t) = Ax(t), t \geq 0$ , is the *maximal*  $r > 0$  for which all systems of the form

$$\dot{x}(t) = (A + D\Delta E)x(t), \quad \|\Delta\| < r,$$

are AS. Here,  $\Delta$  is an unknown disturbance matrix, and  $D$  and  $E$  are given matrices defining the structure of the perturbations. Depending upon whether complex or real disturbances  $\Delta$  are considered, this maximal  $r$  is called a complex or real stability radius, respectively. The basic problem in the study of robustness of stability of the system is to characterize and compute these radii in terms of given matrices  $A, D, E$ . It is important to note that these two stability radii are in general distinct. The analysis and computation of the complex stability radius for systems under structured perturbations was first shown in [12] and extended later in many subsequent papers (see [13] for a survey up till 1990), while the computation of the real stability radius, being a much more difficult problem, has been solved only quite recently; see, e.g., [40]. The situation is much simpler for the class of positive systems. It had been shown in [14], [43] that if  $A$  is a Metzler matrix (i.e.,  $\dot{x}(t) = Ax(t), t \geq 0$ , is a positive system) and  $D, E$  are nonnegative matrices, then the complex and the real stability radii coincide and can be computed directly by a simple formula. These results have been extended recently to many various classes of positive systems such as positive linear time-delay differential systems [34], [30], [45], positive linear discrete time-delay systems [15], [33], and positive linear functional differential equations [35], [46].

Although there have been many works dedicated to studying the stability radius problems of linear dynamical systems, the problem of computing stability radii of linear Volterra integrodifferential equations has not yet been studied in the literature. In this section, we deal with the stability radii of positive linear Volterra integrodifferential equation (8) under structured perturbations and affine perturbations.

**5.1. Stability radius of positive linear Volterra integrodifferential equation under structured perturbations.** We now assume that (8) is UAS and consider the following perturbed equations of the form

$$(27) \quad \dot{x}(t) = (A + D_0\Delta E)x(t) + \int_0^t (B(s) + D_1\delta(s)E)x(t-s)ds,$$

where  $D_i \in \mathbb{R}^{n \times l_i}$  ( $i \in I := \{0, 1\}$ ),  $E \in \mathbb{R}^{q \times n}$  are given matrices determining the *structure* of perturbations, and  $\Delta \in \mathbb{K}^{l_0 \times q}$ ,  $\delta(\cdot) \in L^1([0, +\infty), \mathbb{K}^{l_1 \times q}) \cap C([0, +\infty), \mathbb{K}^{l_1 \times q})$ , ( $\mathbb{K} = \mathbb{R}, \mathbb{C}$ ) are unknown disturbances. In this case, we say that  $A, B(\cdot)$  are subjected to *structured perturbations* of the form

$$(28) \quad A \rightsquigarrow A + D_0 \Delta E; \quad B(\cdot) \rightsquigarrow B(\cdot) + D_1 \delta(\cdot) E.$$

We shall measure the size of each perturbation  $(\Delta, \delta(\cdot)) \in \mathbb{K}^{l_0 \times q} \times L^1([0, +\infty), \mathbb{K}^{l_1 \times q})$  by the norm

$$(29) \quad \|(\Delta, \delta(\cdot))\| := \|\Delta\| + \int_0^{+\infty} \|\delta(t)\| dt.$$

The main problem here is to find the maximal  $r > 0$  for which the perturbed equations (27) remain UAS whenever  $\|(\Delta, \delta(\cdot))\| < r$ .

Let  $\sigma(A + D_0 \Delta E, B(\cdot) + D_1 \delta(\cdot) E)$  be the set of all roots of the characteristic equation of a perturbed equation (27). Recall that, by Theorem 4.5, the perturbed equation (27) is UAS if and only if  $\sigma(A + D_0 \Delta E, B(\cdot) + D_1 \delta(\cdot) E) \subset \mathbb{C}^-$ . Denote

$$\mathcal{D}_{\mathbb{C}} := \{(\Delta, \delta(\cdot)) : \Delta \in \mathbb{C}^{l_0 \times q}, \delta(\cdot) \in L^1([0, +\infty), \mathbb{C}^{l_1 \times q}) \cap C([0, +\infty), \mathbb{C}^{l_1 \times q})\},$$

$$\mathcal{D}_{\mathbb{R}} := \{(\Delta, \delta(\cdot)) : \Delta \in \mathbb{R}^{l_0 \times q}, \delta(\cdot) \in L^1([0, +\infty), \mathbb{R}^{l_1 \times q}) \cap C([0, +\infty), \mathbb{R}^{l_1 \times q})\},$$

$$\mathcal{D}_+ := \{(\Delta, \delta(\cdot)) : \Delta \in \mathbb{R}_+^{l_0 \times q}, \delta(\cdot) \in L^1([0, +\infty), \mathbb{R}_+^{l_1 \times q}) \cap C([0, +\infty), \mathbb{R}_+^{l_1 \times q})\}.$$

Then  $\mathcal{D}_{\mathbb{C}}, \mathcal{D}_{\mathbb{R}}, \mathcal{D}_+$  are called, respectively, the *class of complex, real, and nonnegative perturbations*. In what follows, we always define  $\inf \emptyset = +\infty, 0^{-1} = +\infty$ . To study robustness of stability of the linear Volterra equation (8), we introduce the following.

**DEFINITION 5.1.** *Let (8) be UAS. Then the complex, real, and positive stability radii of the equation under structured perturbations (28) are defined, respectively, by*

$$r_{\mathbb{C}} = \inf\{\|(\Delta, \delta(\cdot))\| : (\Delta, \delta(\cdot)) \in \mathcal{D}_{\mathbb{C}}, \sigma(A + D_0 \Delta E, B(\cdot) + D_1 \delta(\cdot) E) \not\subset \mathbb{C}^-\},$$

$$r_{\mathbb{R}} = \inf\{\|(\Delta, \delta(\cdot))\| : (\Delta, \delta(\cdot)) \in \mathcal{D}_{\mathbb{R}}, \sigma(A + D_0 \Delta E, B(\cdot) + D_1 \delta(\cdot) E) \not\subset \mathbb{C}^-\},$$

$$r_+ = \inf\{\|(\Delta, \delta(\cdot))\| : (\Delta, \delta(\cdot)) \in \mathcal{D}_+, \sigma(A + D_0 \Delta E, B(\cdot) + D_1 \delta(\cdot) E) \not\subset \mathbb{C}^-\}.$$

From the above definition, it is easy to see that

$$(30) \quad 0 < r_{\mathbb{C}} \leq r_{\mathbb{R}} \leq r_+ \leq +\infty.$$

To get characterizations of the stability radii for the class of positive equations, we need the following technical lemmas.

**LEMMA 5.2.** *Suppose that (8) is positive and UAS and  $D \in \mathbb{R}_+^{n \times l}, E \in \mathbb{R}_+^{q \times n}$ . Then*

$$(31) \quad \left| \left( sI_n - A - \int_0^{+\infty} e^{-st} B(t) dt \right)^{-1} x \right| \leq \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} |x|, \quad x \in \mathbb{C}^n,$$

for every  $s \in \mathbb{C}, \Re s \geq 0$ . Moreover, we have

$$(32) \quad \max_{s \in \mathbb{C}, \Re s \geq 0} \left\| E \left( sI_n - A - \int_0^{+\infty} e^{-st} B(t) dt \right)^{-1} D \right\| = \left\| E \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} D \right\|.$$

*Proof.* (i) Since (8) is positive, it follows that  $A$  is a Metzler matrix and  $B(t) \geq 0 \forall t \geq 0$ . For every  $s \in \mathbb{C}, \Re s \geq 0$ , by Theorem 2.1(iv), we get

$$\mu \left( A + \int_0^{+\infty} e^{-st} B(t) dt \right) \leq \mu \left( A + \int_0^{+\infty} e^{-\Re s t} B(t) dt \right) \leq \mu \left( A + \int_0^{+\infty} B(t) dt \right).$$

On the other hand, by Theorem 4.6, we have  $\mu(A + \int_0^{+\infty} B(t) dt) < 0$ . Therefore,  $\mu(A + \int_0^{+\infty} e^{-st} B(t) dt) < 0$  for every  $s \in \mathbb{C}, \Re s \geq 0$ . For a fixed  $s \in \mathbb{C}, \Re s \geq 0$ , we can represent

$$(33) \quad \left( sI_n - \left( A + \int_0^{+\infty} e^{-st} B(t) dt \right) \right)^{-1} x = \int_0^{+\infty} e^{-s\theta} e^{\theta(A + \int_0^{+\infty} e^{-st} B(t) dt)} x d\theta, \quad x \in \mathbb{C}^n,$$

for every  $s \in \mathbb{C}, \Re s \geq 0$ ; see [28, p. 8] or [39]. Since  $A$  is a Metzler matrix, there exists a real number  $\alpha_0 > 0$  such that  $(A + \alpha_0 I_n) \geq 0$ . From  $(A + \alpha_0 I_n)$  and  $B(t) \geq 0 \forall t \geq 0$ , it follows that

$$e^{\alpha_0 \theta} |e^{\theta(A + \int_0^{+\infty} e^{-st} B(t) dt)}| = |e^{\alpha_0 \theta} e^{\theta(A + \int_0^{+\infty} e^{-st} B(t) dt)}| = |e^{\alpha_0 \theta I_n} e^{\theta(A + \int_0^{+\infty} e^{-st} B(t) dt)}| \\ = |e^{\theta((A + \alpha_0 I_n) + \int_0^{+\infty} e^{-st} B(t) dt)}| \leq e^{\theta((\alpha_0 I_n + A) + \int_0^{+\infty} B(t) dt)} = e^{\alpha_0 \theta} e^{\theta(A + \int_0^{+\infty} B(t) dt)}, \quad \theta \geq 0.$$

This implies that

$$(34) \quad |e^{\theta(A + \int_0^{+\infty} e^{-st} B(t) dt)}| \leq e^{\theta(A + \int_0^{+\infty} B(t) dt)}, \quad \theta \geq 0, \quad s \in \mathbb{C}, \Re s \geq 0.$$

Taking (33), (34) into account, we get

$$\left| \left( sI_n - A - \int_0^{+\infty} e^{-st} B(t) dt \right)^{-1} x \right| \leq \int_0^{+\infty} e^{\theta(A + \int_0^{+\infty} B(t) dt)} d\theta |x| \\ = \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} |x|$$

for every  $s \in \mathbb{C}, \Re s \geq 0$ . Furthermore, since  $D, E$  are the nonnegative matrices, it follows that

$$\left| E \left( sI_n - A - \int_0^{+\infty} e^{-st} B(t) dt \right)^{-1} D x \right| \leq E \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} D |x|, \quad x \in \mathbb{C}^n,$$

for every  $s \in \mathbb{C}, \Re s \geq 0$ . By the monotonicity property of the vector norm and the definition of the operator norm, we get

$$\left\| E \left( sI_n - A - \int_0^{+\infty} e^{-st} B(t) dt \right)^{-1} D \right\| \leq \left\| E \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} D \right\|$$

for every  $s \in \mathbb{C}, \Re s \geq 0$ . This completes our proof.  $\square$

LEMMA 5.3. Suppose that (8) is positive and UAS and  $D_i \in \mathbb{R}_+^{n \times l_i}$ ,  $i \in I$ ,  $E \in \mathbb{R}_+^{q \times n}$ . If  $\max_{i \in I} \|E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_i\| \neq 0$ , then there exists a nonnegative perturbation  $(\Delta_0, \delta_1(\cdot)) \in \mathcal{D}_+$  such that

$$(35) \quad \|(\Delta_0, \delta_1(\cdot))\| = \frac{1}{\max_{i \in I} \|E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_i\|}$$

and

$$(36) \quad 0 \in \sigma(A + D_0 \Delta_0 E, B(\cdot) + D_1 \delta_1(\cdot) E).$$

*Proof.* Since (8) is UAS, it follows from Theorem 4.6 that  $\mu(A + \int_0^{+\infty} B(t)dt) < 0$ . By Theorem 2.1(iii), we get  $E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_i \in \mathbb{R}_+^{q \times l_i}$ ,  $i \in I$ . Assume that

$$\max_{i \in I} \left\{ \left\| E \left( -A - \int_0^{+\infty} B(t)dt \right)^{-1} D_i \right\| \right\} = \left\| E \left( -A - \int_0^{+\infty} B(t)dt \right)^{-1} D_{i_0} \right\|, \quad i_0 \in I.$$

Then  $\|E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0}\| = \max_{u \in \mathbb{R}_+^{l_{i_0}}, \|u\|=1} \|E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0} u\|$ ; see, e.g., [16]. Therefore, we can choose  $u_0 \in \mathbb{R}_+^{l_{i_0}}$  such that  $\|u_0\| = 1$  and  $\|E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0} u_0\| = \|E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0}\|$ . By a theorem of Krein and Rutman [22], from  $E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0} u_0 \geq 0$ , there exists a positive linear form  $y^* \in (\mathbb{C}^q)^*$  of dual norm  $\|y^*\| = 1$  such that  $y^* E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0} u_0 = \|E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0} u_0\|$ . Define  $\Delta := \|E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0}\|^{-1} u_0 y^* \in \mathbb{R}_+^{l_{i_0} \times q}$ . It is easy to see that  $\|\Delta\| = \|E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0}\|^{-1}$ . Set  $x_0 := (-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0} u_0$ . This implies that  $\Delta E x_0 = u_0$ . Therefore,  $x_0 \neq 0$  and  $x_0 = (-A - \int_0^{+\infty} B(t)dt)^{-1} D_{i_0} \Delta E x_0$ . It follows that

$$\begin{aligned} \left( (A + D_{i_0} \Delta E) + \int_0^{+\infty} B(t)dt \right) x_0 &= \left( A + \int_0^{+\infty} (B(t) + D_{i_0} \delta(t) E) dt \right) x_0 = 0, \\ x_0 &\in \mathbb{R}^n, x_0 \neq 0, \end{aligned}$$

where  $\delta(t) := e^{-t} \Delta \in L^1([0, +\infty), \mathbb{R}_+^{l_{i_0} \times q}) \cap C([0, +\infty), \mathbb{R}_+^{l_{i_0} \times q})$ . We consider two separate cases as follows:

- If  $i_0 = 0$ , then we set  $\Delta_0 = \Delta$  and  $\delta_1(\cdot) = 0$ . Then it is easy to see that  $(\Delta_0, \delta_1(\cdot)) \in \mathcal{D}_+$  satisfies (35)–(36).
- If  $i_0 = 1$ , then we set  $\Delta_0 = 0$  and  $\delta_1(\cdot) = \delta(\cdot)$ . Then  $(\Delta_0, \delta_1(\cdot)) \in \mathcal{D}_+$  satisfies (35)–(36).

This completes our proof.  $\square$

The following theorem shows that for the class of positive equations, the complex, real, and positive stability radii coincide and can be computed by an explicit formula.

THEOREM 5.4. Suppose that (8) is positive and UAS and  $D_i \in \mathbb{R}_+^{n \times l_i}$  ( $i \in I$ ),  $E \in \mathbb{R}_+^{q \times n}$ . Then

$$r_{\mathbb{C}} = r_{\mathbb{R}} = r_+ = \frac{1}{\max_{i \in I} \|E(-A - \int_0^{+\infty} B(t)dt)^{-1} D_i\|}.$$

*Proof.* Suppose that  $r_{\mathbb{C}} < +\infty$ , as otherwise there is nothing to show. Let  $(\Delta, \delta(\cdot)) \in \mathcal{D}_{\mathbb{C}}$  be a destabilizing complex disturbance. That is,  $\sigma(A + D_0 \Delta E, B(\cdot) +$

$D_1\delta(\cdot)E) \not\subset \mathbb{C}^-$ . Then there exist  $s \in \mathbb{C}$ ,  $\Re s \geq 0$ , and a nonzero vector  $x \in \mathbb{C}^n$  such that

$$\left( A + D_0\Delta E + \int_0^{+\infty} e^{-st}(B(t) + D_1\delta(t)E)dt \right) x = sx.$$

Since (8) is UAS, it follows that

$$\left( sI_n - A - \int_0^{+\infty} e^{-st}B(t)dt \right)^{-1} \left( D_0\Delta E + D_1 \int_0^{+\infty} e^{-st}\delta(t)dt E \right) x = x.$$

Therefore,  $Ex \neq 0$  and we thus get

$$E \left( sI_n - A - \int_0^{+\infty} e^{-st}B(t)dt \right)^{-1} \left( D_0\Delta E + D_1 \int_0^{+\infty} e^{-st}\delta(t)dt E \right) x = Ex.$$

This implies that

$$\begin{aligned} & \left\| E \left( sI_n - A - \int_0^{+\infty} e^{-st}B(t)dt \right)^{-1} D_0 \right\| \|\Delta\| \|Ex\| \\ & + \left\| E \left( sI_n - A - \int_0^{+\infty} e^{-st}B(t)dt \right)^{-1} D_1 \right\| \left\| \int_0^{+\infty} e^{-st}\delta(t)dt \right\| \|Ex\| \geq \|Ex\|. \end{aligned}$$

Hence,

$$\max_{i \in I} \left\{ \left\| E \left( sI_n - A - \int_0^{+\infty} e^{-st}B(t)dt \right)^{-1} D_i \right\| \right\} \left( \|\Delta\| + \int_0^{+\infty} \|\delta(t)\|dt \right) \geq 1.$$

Using Lemma 3.6, we get

$$\max_{i \in I} \left\{ \left\| E \left( -A - \int_0^{+\infty} B(t)dt \right)^{-1} D_i \right\| \right\} \|(\Delta, \delta(\cdot))\| \geq 1.$$

This is equivalent to

$$\|(\Delta, \delta(\cdot))\| \geq \frac{1}{\max_{i \in I} \left\{ \left\| E \left( -A - \int_0^{+\infty} B(t)dt \right)^{-1} D_i \right\| \right\}}.$$

Since this inequality holds true for an arbitrary destabilizing complex perturbation, we conclude that

$$r_{\mathbb{C}} \geq \frac{1}{\max_{i \in I} \left\{ \left\| E \left( -A - \int_0^{+\infty} B(t)dt \right)^{-1} D_i \right\| \right\}}.$$

Taking into account the inequalities (30), it remains only to show that

$$r_+ \leq \frac{1}{\max_{i \in I} \left\{ \left\| E \left( -A - \int_0^{+\infty} B(t)dt \right)^{-1} D_i \right\| \right\}}.$$

However, this is immediate from Lemma 5.3. This completes our proof.  $\square$

*Example 5.5.* We revisit the linear positive Volterra integrodifferential equation in  $\mathbb{R}^2$  defined by (25)–(26) in Example 4.9. Assume that the matrix  $A$  and the matrix function  $B(\cdot)$  are subjected to the structured perturbations of the form (28), where

$$D_1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{and} \quad E = I_2.$$

By an easy computation, we get

$$\begin{aligned} E \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} D_1 &= \begin{pmatrix} \frac{-2b}{b^2-1} \\ \frac{2}{b^2-1} \end{pmatrix}, \\ E \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} D_2 &= \begin{pmatrix} \frac{1}{b^2-1} \\ \frac{-b}{b^2-1} \end{pmatrix}, \end{aligned}$$

where  $b := a + 1/p < -1$ . Therefore, if  $\mathbb{R}^2$  is endowed, respectively, with the 1-norm, 2-norm, and  $\infty$ -norm, then by Theorem 5.4, we have

$$r_{\mathbb{C}} = r_{\mathbb{R}} = r_+ = \frac{-1-b}{2}, \quad \frac{(b^2-1)\sqrt{(b^2+1)}}{2(b^2+1)}, \quad \frac{b^2-1}{-2b},$$

respectively.

**5.2. Stability radius of positive linear Volterra integrodifferential equations under affine perturbations.** We now deal with the problem of computing the stability radius of positive linear Volterra equation (8) under affine perturbations. To do this, we assume that (8) is UAS and the system matrices  $A, B(\cdot)$  are subjected to *affine perturbation* of the form

$$(37) \quad A \rightsquigarrow A + \sum_{i=1}^N \alpha_i A_i, \quad B(\cdot) \rightsquigarrow B(\cdot) + \sum_{i=1}^N \beta_i B_i(\cdot),$$

where  $A_i \in \mathbb{R}^{n \times n}$ ,  $B_i(\cdot) \in L^1([0, +\infty), \mathbb{R}^{n \times n}) \cap C([0, +\infty), \mathbb{R}^{n \times n})$ ,  $i \in \underline{N} := \{1, 2, \dots, N\}$  are given and  $\alpha_i, \beta_i \in \mathbb{K}$  ( $\mathbb{K} = \mathbb{R}, \mathbb{C}$ ),  $i \in \underline{N}$ , are unknown scalars.

We assume that  $A$  is a Metzler matrix and  $B(t) \geq 0 \ \forall t \geq 0$ , under which the linear Volterra equation (8) is positive. In view of positivity of the equation under consideration, we assume that  $A_i \in \mathbb{R}_+^{n \times n}$  and  $B_i(t) \geq 0 \ \forall t \geq 0, \forall i \in \underline{N}$ . We define the complex and the real stability radii of the linear Volterra equation (8) under affine parameter perturbations (37) by setting, for  $\mathbb{K} = \mathbb{C}$  and, respectively,  $\mathbb{K} = \mathbb{R}$ ,

$$\begin{aligned} r_{\mathbb{K}}^a &= \inf \left\{ \max \left( \max_{i \in \underline{N}} |\alpha_i|; \max_{i \in \underline{N}} |\beta_i| \right) : \right. \\ &\quad \left. \alpha_i, \beta_i \in \mathbb{K}, \sigma \left( A + \sum_{i=1}^N \alpha_i A_i, B(\cdot) + \sum_{i=1}^N \beta_i B_i(\cdot) \right) \not\subset \mathbb{C}^- \right\}. \end{aligned}$$

Similarly, the positive stability radius  $r_+^a$  is obtained by restricting, in the above definition, the disturbances  $(\alpha, \beta) := ((\alpha_i)_{i \in \underline{N}}, (\beta_j)_{j \in \underline{N}})$  to be nonnegative.

It is clear that

$$(38) \quad 0 < r_{\mathbb{C}}^a \leq r_{\mathbb{R}}^a \leq r_+^a.$$

The following theorem proves that the equalities in (38) hold true and provides computable formulae for the stability radii  $r_{\mathbb{K}}^a, r_+^a$  ( $\mathbb{K} = \mathbb{C}, \mathbb{R}$ ).

**THEOREM 5.6.** *Suppose that the linear Volterra equation (8) is UAS and  $A, B(\cdot)$  are subjected to multi-affine perturbations of the form (37). If the stability radii of the equation are given by (38), then*

$$(39) \quad r_{\mathbb{C}}^a = r_{\mathbb{R}}^a = r_+^a = \frac{1}{\mu[(-A - \int_0^{+\infty} B(t)dt)^{-1}(\sum_{i=1}^N A_i + \sum_{i=1}^N \int_0^{+\infty} B_i(t)dt)]}.$$

*Proof.* We first prove that

$$r_+^a = \frac{1}{\mu[(-A - \int_0^{+\infty} B(t)dt)^{-1}(\sum_{i=1}^N A_i + \sum_{i=1}^N \int_0^{+\infty} B_i(t)dt)]}.$$

Let  $(\alpha, \beta) = ((\alpha_i)_{i \in \underline{N}}, (\beta_i)_{i \in \underline{N}})$  be an arbitrary nonnegative destabilizing perturbation, that is,

$$\sigma\left(A + \sum_{i=1}^N \alpha_i A_i, B(\cdot) + \sum_{i=1}^N \beta_i B_i(\cdot)\right) \not\subset \mathbb{C}^-.$$

Then there exist a complex number  $s, \Re s \geq 0$  and a nonzero vector  $x \in \mathbb{C}^n$  such that

$$\left(\left(A + \sum_{i=1}^N \alpha_i A_i\right) + \int_0^{+\infty} e^{-st} \left(B(t) + \sum_{i=1}^N \beta_i B_i(t)\right) dt\right) x = sx.$$

Since (8) is UAS, this implies that

$$\left(sI_n - A - \int_0^{+\infty} e^{-st} B(t)dt\right)^{-1} \left(\sum_{i=1}^N \alpha_i A_i + \sum_{i=1}^N \beta_i \int_0^{+\infty} e^{-st} B_i(t)dt\right) x = x.$$

Using (31), we obtain the estimates

$$\begin{aligned} |x| &= \left| \left(sI_n - A - \int_0^{+\infty} e^{-st} B(t)dt\right)^{-1} \left(\sum_{i=1}^N \alpha_i A_i + \sum_{i=1}^N \beta_i \int_0^{+\infty} e^{-st} B_i(t)dt\right) x \right| \\ &\leq \left(-A - \int_0^{+\infty} B(t)dt\right)^{-1} \left| \left(\sum_{i=1}^N \alpha_i A_i + \sum_{i=1}^N \beta_i \int_0^{+\infty} e^{-st} B_i(t)dt\right) x \right| \\ &\leq \left(-A - \int_0^{+\infty} B(t)dt\right)^{-1} \left(\sum_{i=1}^N \alpha_i A_i + \sum_{i=1}^N \beta_i \int_0^{+\infty} B_i(t)dt\right) |x| \\ &\leq \gamma \left[ \left(-A - \int_0^{+\infty} B(t)dt\right)^{-1} \left(\sum_{i=1}^N A_i + \sum_{i=1}^N \int_0^{+\infty} B_i(t)dt\right) \right] |x|, \end{aligned}$$

where  $\gamma := \max\{\max_{i \in \underline{N}} \alpha_i, \max_{i \in \underline{N}} \beta_i\}$ . Since  $B := [(-A - \int_0^{+\infty} B(t)dt)^{-1}(\sum_{i=1}^N A_i + \sum_{i=1}^N \int_0^{+\infty} B_i(t)dt)]$  is a nonnegative matrix, it follows from Theorem 2.1(ii) that

$$\mu(B) \geq \frac{1}{\gamma} > 0.$$

Hence,

$$\gamma \geq \frac{1}{\mu(B)}.$$

Since this holds for arbitrary destabilizing nonnegative perturbation  $(\alpha, \beta)$ , we conclude that

$$r_+^a \geq \frac{1}{\mu(B)}.$$

We shall prove that the converse inequality holds true. In fact, by Theorem 2.1(i) (Perron–Frobenius), there exists a nonzero vector  $y \in \mathbb{R}_+^n$  such that  $By = \mu(B)y$ . This implies that

$$\left( \left( A_0 + \sum_{i=1}^N \frac{1}{\mu(B)} A_i \right) + \int_0^{+\infty} \left( B(t) + \sum_{i=1}^N \frac{1}{\mu(B)} B_i(t) \right) dt \right) y = 0.$$

This means that the nonnegative perturbation  $(\alpha^*, \beta^*)$  defined by  $\alpha_i^* = 1/\mu(B)$ ,  $\beta_i^* = 1/\mu(B)$ ,  $i \in \underline{N}$ , is destabilizing. By the definition of  $r_+^a$ , we have

$$r_+^a \leq \frac{1}{\mu(B)}.$$

Thus, we obtain

$$r_+^a = \frac{1}{\mu(B)} = \frac{1}{\mu \left[ \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} \left( \sum_{i=1}^N A_i + \sum_{i=1}^N \int_0^{+\infty} B_i(t) dt \right) \right]}.$$

We are now ready to show that  $r_{\mathbb{C}}^a = r_{\mathbb{R}}^a = r_+^a$ . Suppose  $(\alpha, \beta) = ((\alpha_i)_{i \in \underline{N}}, (\beta_j)_{j \in \underline{N}})$  is an arbitrary complex destabilizing perturbation. By an argument similar to the above, we get

$$(40) \quad \left[ \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} \left( \sum_{i=1}^N |\alpha_i| A_i + \sum_{i=1}^N |\beta_i| \int_0^{+\infty} B_i(t) dt \right) \right] |x_0| \geq |x_0|$$

for some  $x_0 \in \mathbb{C}^n$ ,  $x_0 \neq 0$ . By Theorem 2.1(ii),

$$\mu \left[ \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} \left( \sum_{i=1}^N |\alpha_i| A_i + \sum_{i=1}^N |\beta_i| \int_0^{+\infty} B_i(t) dt \right) \right] \geq 1.$$

Since  $C := \left( -A - \int_0^{+\infty} B(t) dt \right)^{-1} \left( \sum_{i=1}^N |\alpha_i| A_i + \sum_{i=1}^N |\beta_i| \int_0^{+\infty} B_i(t) dt \right)$  is nonnegative, using Theorem 2.1(i) again, we have  $Cx_1 = \mu(C)x_1$  for some nonzero vector  $x_1 \in \mathbb{R}_+^n$ . This gives

$$\left( \left( A_0 + \sum_{i=1}^N \frac{|\alpha_i|}{\mu(C)} A_i \right) + \int_0^{+\infty} \left( B(t) + \sum_{i=1}^N \frac{|\beta_i|}{\mu(C)} B_i(t) \right) dt \right) x_1 = 0,$$

which means that

$$(|\alpha|, |\beta|) := \left( \left( \frac{|\alpha_i|}{\mu(C)} \right)_{i \in \underline{N}}, \left( \frac{|\beta_i|}{\mu(C)} \right)_{i \in \underline{N}} \right)$$



is a nonnegative destabilizing perturbation. Hence, it follows from the definition of  $r_+^a$  that

$$\max \left( \max_{i \in \underline{N}} \left( \frac{|\alpha_i|}{\mu(C)} \right), \max_{i \in \underline{N}} \left( \frac{|\beta_i|}{\mu(C)} \right) \right) \geq r_+^a,$$

or, equivalently,  $\max(\max_{i \in \underline{N}} |\alpha_i|, \max_{i \in \underline{N}} |\beta_i|) \geq \mu(C)r_+^a \geq r_+^a$ , which implies that  $r_{\mathbb{C}}^a \geq r_+^a$ . Combined with the inequalities  $r_{\mathbb{C}}^a \leq r_{\mathbb{R}}^a \leq r_+^a$ , this implies that  $r_{\mathbb{C}}^a = r_{\mathbb{R}}^a = r_+^a$ . In addition, from the above arguments, we observe that  $r_{\mathbb{C}}^a = r_{\mathbb{R}}^a = r_+^a = +\infty$  if and only if  $\mu(B) = 0$ . This completes our proof.  $\square$

**6. Concluding remarks.** A new class of positive linear systems (namely, positive linear Volterra integrodifferential equations) is introduced and an explicit criterion for equations of this class is presented. A new Perron–Frobenius theorem for positive linear Volterra integrodifferential equations is given. Furthermore, explicit criteria for uniformly asymptotic stability of positive equations are provided. Finally, two explicit formulae for the stability radii of positive equations under structured perturbations and affine perturbations are given.

The proposed approach can be extended to study positive linear Volterra integrodifferential equations with delay and positive linear Volterra–Stieltjes differential equations. These works will be done in the future.

#### REFERENCES

- [1] D. AEYELS AND P. DE LEENHEER, *Extension of the Perron–Frobenius theorem to homogeneous systems*, SIAM J. Control Optim., 41 (2002), pp. 563–582.
- [2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in Mathematical Sciences*, Academic Press, New York, 1979.
- [3] T. A. BURTON, *Volterra Integral and Differential Equations*, Math. Sci. Eng. 167, Academic Press, New York, 1983.
- [4] T. A. BURTON, *Stability and Periodic Solutions of Ordinary and Functional Differential Equations*, Math. Sci. Eng. 178, Academic Press, New York, 1985.
- [5] L. FARINA AND S. RINALDI, *Positive Linear Systems: Theory and Applications*, John Wiley, New York, 2000.
- [6] S. GAUBERT AND J. GUNAWARDENA, *The Perron–Frobenius theorem for homogeneous, monotone functions*, Trans. Amer. Math. Soc., 356 (2004), pp. 4931–4950.
- [7] G. S. GROSSMAN AND R. K. MILLER, *Nonlinear Volterra integro-differential systems with  $L^1$ -kernels*, J. Differential Equations, 13 (1973), pp. 551–566.
- [8] W. M. HADDAD AND V. CHELLABOINA, *Stability and dissipativity theory for nonnegative and compartmental dynamical systems with time delay*, in Advances in Time-Delay Systems, Lect. Notes Comput. Sci. Eng. 38, Springer, Berlin, 2004, pp. 421–435.
- [9] W. M. HADDAD AND V. CHELLABOINA, *Stability theory for nonnegative and compartmental dynamical systems with delay*, Systems Control Lett., 51 (2004), pp. 355–361.
- [10] W. M. HADDAD AND V. CHELLABOINA, *Stability and dissipativity theory for nonnegative dynamical systems: A unified analysis framework for biological and physiological systems*, Nonlinear Anal. Real World Appl., 6 (2005), pp. 35–65.
- [11] Y. HINO AND S. MURAKAMI, *Stability properties of linear Volterra integro-differential equations in a Banach space*, Funkcial. Ekvac., 48 (2005), pp. 367–392.
- [12] D. HINRICHSSEN AND A. J. PRITCHARD, *Stability radius for structured perturbations and the algebraic Riccati equation*, Systems Control Lett., 8 (1986), pp. 105–113.
- [13] D. HINRICHSSEN AND A. J. PRITCHARD, *Real and complex stability radii: A survey*, in Control of Uncertain Systems, Progr. Systems Control Theory 6, D. Hinrichsen and B. Mårtensson, eds., Birkhäuser, Basel, 1990, pp. 119–162.
- [14] D. HINRICHSSEN AND N. K. SON,  *$\mu$ -analysis and robust stability of positive linear systems*, Appl. Math. Comput. Sci., 8 (1998), pp. 253–268.
- [15] D. HINRICHSSEN, N. K. SON, AND P. H. A. NGOC, *Stability radii of positive higher order difference systems*, Systems Control Lett., 49 (2003), pp. 377–388.
- [16] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1993.
- [17] G. JAMES AND V. RUMCHEV, *Stability of positive linear discrete-time systems*, Systems Sci., 30 (2004), pp. 51–67.

- [18] G. JAMES, S. P. KOSTOVA, AND V. G. RUMCHEV, *Pole-assignment for a class of positive linear systems*, Internat. J. Systems Sci., 32 (2001), pp. 1377–1388.
- [19] G. S. JORDAN AND R. L. WHEELER, *Structure of resolvents of Volterra integral and integro-differential systems*, SIAM J. Math. Anal., 11 (1980), pp. 119–132.
- [20] G. S. JORDAN, O. J. STAFFANS, AND R. L. WHEELER, *Local analyticity in weighted  $L^1$ -spaces and applications to stability problems for Volterra equations*, Trans. Amer. Math. Soc., 274 (1982), pp. 749–782.
- [21] F. E. KLOEDEN AND A. M. RUBINOV, *A generalization of Perron–Frobenius theorem*, Nonlinear Anal., 41 (2000), pp. 97–115.
- [22] M. J. KREIN AND M. A. RUTMAN, *Linear operators leaving invariant a cone in a Banach space*, Amer. Math. Soc. Transl., 26 (1950), pp. 199–325.
- [23] V. LAKSHMIKANTHAM AND M. R. M. RAO, *Theory of Integro-differential Equations*, Stability Control Theory Methods Appl. 1, Gordon and Breach, Lausanne, 1995.
- [24] P. DE LEENHEER AND D. P. AYEYLS, *Stabilization of positive systems with first integrals*, Automatica, 38 (2002), pp. 1583–1589.
- [25] D. G. LUENBERGER, *Introduction to Dynamic Systems, Theory, Models and Applications*, John Wiley, New York, 1979.
- [26] R. K. MILLER, *Asymptotic stability properties of Volterra integro-differential systems*, J. Differential Equations, 10 (1971), pp. 485–506.
- [27] R. K. MILLER, *Structure of solutions of unstable linear Volterra integro-differential equations*, J. Differential Equations, 15 (1974), pp. 129–157.
- [28] R. NAGEL, ED., *One-Parameter Semigroups of Positive Operators*, Springer, Berlin, 1986.
- [29] T. NAITO, J. S. SHIN, S. MURAKAMI, AND P. H. A. NGOC, *Characterizations of linear Volterra integral equations with nonnegative kernels*, J. Math. Anal. Appl., 335 (2007), pp. 298–313.
- [30] P. H. A. NGOC, *Strong stability radii of positive linear time-delay systems*, Internat. J. Robust Nonlinear Control, 15 (2005), pp. 459–472.
- [31] P. H. A. NGOC, *A Perron–Frobenius for a class of positive quasi-polynomial matrices*, Appl. Math. Lett., 19 (2006), pp. 747–751.
- [32] P. H. A. NGOC AND B. S. LEE, *A characterization of spectral abscissa and Perron–Frobenius theorem of positive linear functional differential equations*, IMA J. Math. Control Inform., 23 (2006), pp. 259–268.
- [33] P. H. A. NGOC AND N. K. SON, *Stability radii of positive linear difference equations under affine parameter perturbations*, Appl. Math. Comput., 134 (2003), pp. 577–594.
- [34] P. H. A. NGOC AND N. K. SON, *Stability radii of linear systems under multi-perturbations*, Numer. Funct. Anal. Optim., 25 (2004), pp. 221–238.
- [35] P. H. A. NGOC AND N. K. SON, *Stability radii of positive linear functional differential equations under multi-perturbations*, SIAM J. Control Optim., 43 (2005), pp. 2278–2295.
- [36] P. H. A. NGOC, B. S. LEE, AND N. K. SON, *Perron–Frobenius theorem for positive polynomial matrices*, Vietnam J. Math., 32 (2004), pp. 475–481.
- [37] P. H. A. NGOC, T. NAITO, AND J. S. SHIN, *Characterizations of positive linear functional differential equations*, Funkcial. Ekvac., 50 (2007), pp. 1–17.
- [38] P. H. A. NGOC, T. NAITO, AND J. S. SHIN, *On stability of a class of positive linear functional difference equations*, Math. Control Signals Systems, 19 (2007), pp. 361–382.
- [39] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, Berlin, 1983.
- [40] L. QIU, B. BERNHARDSSON, A. RANTZER, E. J. DAVISON, P. M. YOUNG, AND J. C. DOYLE, *A formula for computation of the real stability radius*, Automatica, 31 (1995), pp. 879–890.
- [41] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1987.
- [42] S. M. RUMP, *Theorems of Perron–Frobenius type for matrices without sign restrictions*, Linear Algebra Appl., 266 (1997), pp. 1–42.
- [43] B. SHAFAI, J. CHEN, AND M. KOTHANDARAMAN, *Explicit formulas for stability radii of nonnegative and Metzlerian matrices*, IEEE Trans. Automat. Control, 42 (1997), pp. 265–269.
- [44] N. K. SON AND D. HINRICHSSEN, *Robust stability of positive continuous time systems*, Numer. Funct. Anal. Optim., 17 (1996), pp. 649–659.
- [45] N. K. SON AND P. H. A. NGOC, *Robust stability of positive linear time delay systems under affine parameter perturbations*, Acta Math. Vietnam., 24 (1999), pp. 353–372.
- [46] N. K. SON AND P. H. A. NGOC, *Robust stability of linear functional differential equations*, Adv. Stud. Contemp. Math., 3 (2001), pp. 43–59.
- [47] B. ZHANG, *Asymptotic stability criteria and integrability properties of the resolvent of Volterra and functional equations*, Funkcial. Ekvac., 40 (1997), pp. 335–351.
- [48] B. ZHANG, *Necessary and sufficient conditions for stability in Volterra equations of nonconvolution type*, Dynam. Systems Appl., 14 (2005), pp. 525–549.

## PERTURBATION METHODS IN STABILITY AND NORM ANALYSIS OF SPATIALLY PERIODIC SYSTEMS\*

MAKAN FARDAD<sup>†</sup> AND BASSAM BAMIEH<sup>‡</sup>

**Abstract.** We consider systems governed by partial differential equations with spatially periodic coefficients over unbounded domains. These spatially periodic systems are considered as perturbations of spatially invariant ones, and we develop perturbation methods to study their stability and  $\mathcal{H}^2$  system norm. The operator Lyapunov equations characterizing the  $\mathcal{H}^2$  norm are studied by using a special frequency representation, and formulas are given for the perturbation expansion of their solution. The structure of these equations allows for a recursive method of solving for the expansion terms. Our analysis provides conditions that capture possible resonances between the periodic coefficients and the spatially invariant part of the system. These conditions can be regarded as useful guidelines when spatially periodic coefficients are to be designed to increase or decrease the  $\mathcal{H}^2$  norm of a spatially distributed system. The developed perturbation framework also gives simple conditions for checking whether a spatially periodic operator generates a holomorphic  $C_0$  semigroup and thus satisfies the spectrum-determined growth condition.

**Key words.** PDE with periodic coefficients, perturbation analysis,  $\mathcal{H}^2$  norm, sectorial operator, spectrum-determined growth condition

**AMS subject classifications.** 35B27, 35B34, 35B35, 93C20, 93C73, 93D20, 93D25

**DOI.** 10.1137/06065012X

**1. Introduction.** The terms distributed-parameter and infinite-dimensional are used to describe those systems in which the state belongs to an infinite-dimensional vector space [1]. Such systems include, but are not limited to, time-delay (retarded) and spatially distributed systems [2]. The latter includes systems in which the dynamics are governed by partial differential equations (PDEs) and it is a subclass of these systems that will be the subject of this study. More specifically, we will analyze certain properties of spatially distributed systems in which the underlying PDEs have spatially periodic coefficients. We refer to such systems as *spatially periodic*. Spatially periodic systems have many real life applications, for example, in boundary layer and channel flow problems with corrugated walls and in nonlinear optics.

Our motivation for this work is to study the effect of spatially periodic coefficients on stability and system norms of spatially distributed systems. This can be thought of as using the periodic coefficients as static feedback controls for spatially distributed systems. For example, in flow control problems where corrugated wall geometries or spatially periodic body forces are introduced, the PDEs that describe the resulting flow dynamics have periodic coefficients that are related to either the wall shape or the spatially distributed body force. An important objective is to “design” such wall shapes or body forces to obtain certain stability or instability properties of the resulting dynamics. There are currently no systematic methods for achieving this.

An analogy can be made between the present work and the use of time-periodic coefficients in ordinary differential equations (ODEs). It is known that the introduction

---

\*Received by the editors January 17, 2006; accepted for publication (in revised form) August 17, 2007; published electronically March 5, 2008. This work was partially supported by AFOSR grant FA9550-04-1-0207.

<http://www.siam.org/journals/sicon/47-2/65012.html>

<sup>†</sup>Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 (makan@umn.edu).

<sup>‡</sup>Department of Mechanical Engineering, University of California, Santa Barbara, CA 93106 (bamieh@engineering.ucsb.edu).

of time-periodic coefficients in ODEs with constant coefficients can change the stability properties of the linear time invariant (LTI) system described by the original ODE. A useful picture is to think of an ODE with periodic coefficients as an LTI system modified by time-periodic (memoryless) feedback. It is known that certain unstable LTI systems can be stabilized by being put in feedback with periodic gains of properly designed amplitudes and frequencies [3]. This can be roughly considered as an example of “vibrational control” [4]. On the other hand, certain stable or neutrally stable LTI systems can be destabilized by periodic feedback gains. This is sometimes referred to as “parametric resonance” in the dynamical systems literature [3]. In the above scenarios, the stabilization/destabilization process depends in subtle ways on “resonances” between the natural modes of the LTI subsystem and the frequency and amplitude of the periodic feedback. Although Floquet analysis can be used to ascertain stability of the resulting periodic systems, it is cumbersome to use for *designing* the periodic coefficients. This requires an exhaustive search over frequencies and amplitudes of the periodic coefficients. Alternatively, simple resonance conditions can be derived by using a perturbation analysis [3], which in turn can be used for a preliminary selection of the coefficient’s frequency. In this way, perturbation analysis serves as a useful design tool.

In related recent work [5] we developed computational tools to study stability and system norms for spatially periodic systems. However, for problems where the spatially periodic coefficients are to be designed, using these tools involves a computationally expensive search over spatial frequencies and amplitudes of the coefficients. Therefore, our aim in the present work is to develop a *perturbation analysis* that can be used to derive resonance conditions and provide a useful design tool in a similar manner to the case of ODEs discussed earlier. These resonance conditions can then identify candidate spatial frequencies to be used for the periodic coefficients. The exact behavior with respect to amplitudes can then be ascertained by using the full analysis of [5]. In this manner we reduce the dimension of the search space required for design problems.

Another challenging problem is checking the stability of a spatially periodic or, in general, any infinite-dimensional system. It is well known that, for a finite-dimensional LTI system, the spectrum of the infinitesimal generator (i.e., the  $A$ -matrix) being contained in the open left half of the complex plane is equivalent to exponential stability. In this sense the spectrum of the infinitesimal generator determines stability. Therefore it is said that the system satisfies the *spectrum-determined growth condition* (SDGC). But the SDGC may not hold for some infinite-dimensional LTI systems; indeed the evolution can grow exponentially even though the infinitesimal generator (i.e., the  $A$ -operator) has its spectrum inside the left half of the complex plane [6, 7, 8]. In the present work we use perturbation analysis to find simple conditions under which the spatially periodic system satisfies the SDGC and is exponentially stable.

Our presentation is organized as follows. Section 2 outlines the main results of the paper. Section 3 briefly reviews the frequency representation of spatially periodic operators. Section 4 introduces the problem setup. Section 5 discusses the analytic perturbation of the  $\mathcal{H}^2$  norm, and section 6 provides related illustrative examples. Section 7 studies conditions under which a spatially periodic system is exponentially stable. Many proofs and technical details have been relegated to the appendix to improve readability.

**Notation.** We use  $k \in \mathbb{R}$  to characterize the spatial-frequency variable, also known as the *wave number*.  $\Sigma(T)$  is the spectrum of the operator  $T$ ,  $\Sigma_p(T)$  its point spectrum, and  $\rho(T)$  its resolvent set.  $\mathbb{C}^-$  denotes all complex numbers with real part

less than zero, and  $j := \sqrt{-1}$ . “\*” denotes the complex-conjugate transpose and also the adjoint of a linear operator.  $\overline{\mathfrak{S}}$  is the closure of the set  $\mathfrak{S} \subset \mathbb{C}$ . The function  $\hat{u}(t, k)$  denotes the Fourier transform (in the spatial variable  $x$ ) of the spatiotemporal function  $u(t, x)$ . Similarly, the operator  $\hat{A}$  is the Fourier domain representation of the spatial operator  $A$ . We will use the same symbol for a linear operator and its kernel representation. Where there is no chance of confusion, we use the same notation for a spatially invariant operator and its Fourier symbol.

**Terminology.** Throughout the paper, we use the terms *spatial “operators”* and *spatial “systems.”* By the former we mean a *purely spatial* operator with no temporal dynamics (i.e., a memoryless operator that acts on a spatial function and yields a spatial function), whereas by the latter we refer to a *spatiotemporal* system (a system with an internal state which evolves on some spatial domain; i.e., for every time  $t$  the state is a function on a spatial domain). By a *scalar system*, we mean that the Euclidean dimension of the state is equal to one. When spatially periodic feedback operators are small in norm, we will use the phrases *periodic feedback* and *periodic perturbation* interchangeably. Finally, in the case of doubly infinite matrices we will use the phrases biinfinite matrix and (biinfinite or lifted) operator interchangeably.

**2. Main results.** We consider systems described by linear, time-invariant, integro partial differential equations defined on an unbounded one-dimensional domain. We use a standard state-space representation of the form

$$\begin{aligned} [\partial_t \psi](t, x) &= [A \psi](t, x) + [B u](t, x), \\ (1) \quad y(t, x) &= [C \psi](t, x), \end{aligned}$$

where  $t \in [0, \infty)$  and  $x \in \mathbb{R}$ ,  $\psi$ ,  $u$ ,  $y$  are spatiotemporal functions, and  $A$ ,  $B$ ,  $C$  are spatial integrodifferential operators with coefficients that are periodic functions with a common period  $X$ . We refer to (1) as a *spatially periodic system*.

It is often physically meaningful to regard the spatially periodic operators as additive or multiplicative perturbations of spatially invariant ones (and by spatially invariant we mean integrodifferential operators with constant coefficients). For example, the infinitesimal generator  $A$  in (1) can often be decomposed as  $A = A^\circ + \epsilon E$ , where  $A^\circ$  is a spatially invariant operator and  $E$  is an operator that includes multiplication by periodic functions. In some control application the operator  $E$  is something to be “designed.” Therefore it is desirable to have easily verifiable conditions for stability and norms of such systems. This would then allow for the selection of the spatial period and amplitude of periodic functions in  $E$  to achieve the desired behavior. The perturbation analysis we present, though limited to small values of  $\epsilon$ , provides useful guidelines for selecting candidate “periods” for  $E$ .

Our results are derived by using a special frequency representation. We show that the spatial periodicity of the operators  $A$ ,  $B$ , and  $C$  implies that (1) can be rewritten as

$$\begin{aligned} [\partial_t \psi_\theta](t) &= [\mathcal{A}_\theta \psi_\theta](t) + [\mathcal{B}_\theta u_\theta](t), \\ (2) \quad y_\theta(t) &= [\mathcal{C}_\theta \psi_\theta](t), \end{aligned}$$

where  $\theta \in [0, 2\pi/X)$ . For every value of  $\theta$ ,  $\psi_\theta$ ,  $u_\theta$ ,  $y_\theta$  are *biinfinite vectors*, and  $\mathcal{A}_\theta$ ,  $\mathcal{B}_\theta$ ,  $\mathcal{C}_\theta$  are *biinfinite matrices*. The systems (2) and (1) are related through a unitary transformation, and in particular quadratic forms are preserved by this transformation. Consequently, the stability and quadratic norm properties of (2) and (1) are equivalent. With this transformation the analysis of the original system (1) is reduced to that

of the family of systems (2) that are *decoupled* in the parameter  $\theta$ . In particular, perturbation analysis for (2) is easier and less technical than for the original system (1).

To make for easier reading, both in this section and in the body of the paper, we first present the results on perturbation analysis of the  $\mathcal{H}^2$  norm and then deal with the issue of stability.

The first set of results concerns  $\mathcal{H}^2$  norm analysis. For a large class of infinite-dimensional systems, computing the  $\mathcal{H}^2$  norm involves solving an *operator* algebraic Lyapunov equation

$$AP + PA^* = -BB^*.$$

In general this is a difficult task that must be done by using appropriate discretization techniques. However, if  $A$  and  $B$  are spatially periodic operators, then so is the solution  $P$ . Then the special frequency representation described above allows this operator Lyapunov equation to be rewritten as a  $\theta$ -decoupled family of Lyapunov equations

$$(3) \quad \mathcal{A}_\theta \mathcal{P}_\theta + \mathcal{P}_\theta \mathcal{A}_\theta^* = -\mathcal{B}_\theta \mathcal{B}_\theta^*,$$

where  $\mathcal{A}_\theta$ ,  $\mathcal{B}_\theta$ , and  $\mathcal{P}_\theta$  are the *biinfinite matrix* representations of  $A$ ,  $B$ , and  $P$ . Once  $\mathcal{P}_\theta$  is found, the  $\mathcal{H}^2$  norm of the system can be computed from [5]

$$\frac{1}{2\pi} \int_0^\Omega \text{trace}[\mathcal{C}_\theta \mathcal{P}_\theta \mathcal{C}_\theta^*] d\theta, \quad \Omega = 2\pi/X.$$

Solving (3) is still a difficult problem in general, since it involves biinfinite matrices. We use perturbation analysis as follows: The infinitesimal generator  $\mathcal{A}_\theta$  is expressed as  $\mathcal{A}_\theta = \mathcal{A}_\theta^0 + \epsilon \mathcal{E}_\theta$ , where  $\mathcal{A}_\theta^0$  and  $\mathcal{E}_\theta$  correspond to the spatially invariant and spatially periodic components of  $\mathcal{A}_\theta$ , respectively. It follows that the solution  $\mathcal{P}_\theta$  is analytic in  $\epsilon$ , and the terms of its power series expansion  $\mathcal{P}_\theta^{(i)}$  satisfy a sequence of forward coupled Lyapunov equations. Furthermore, the terms  $\mathcal{P}_\theta^{(i)}$  are banded matrices with the number of bands increasing with the index  $i$ . These Lyapunov equations can then be solved recursively for  $i = 0, 1, 2, \dots$ . Formulas for these Lyapunov equations are derived in section 5. In some examples that we present in section 6, these formulas reveal simple “resonance” conditions for stabilization or destabilization of PDEs by using spatially periodic feedback.

The second set of results concerns the problem of stability. As mentioned in the introduction, when the infinitesimal generator  $A$  is an infinite-dimensional operator it is possible that its spectrum  $\Sigma(A)$  lies inside  $\mathbb{C}^-$  and yet  $\|e^{At}\|$  grows exponentially [6, 7, 8]. In such cases it is said that the *spectrum-determined growth condition* is not satisfied [8]. Yet there exists a wide range of infinite-dimensional systems for which the spectrum-determined growth condition *is* satisfied. These include, but are not limited to, systems for which the infinitesimal generator is *sectorial* (also known as an operator which generates a *holomorphic* or *analytic* semigroup) [9, 10, 11] or the infinitesimal generator is a *Riesz-spectral* operator [2]. In this paper we focus on sectorial operators.

Therefore, to establish exponential stability of a system, one possibility would be to show simultaneously that

- (i) the operator  $A$  is sectorial (and thus its spectrum determines stability) and
- (ii) the spectrum  $\Sigma(A)$  lies in  $\mathbb{C}^-$ .

The problem is that proving an infinite-dimensional operator is sectorial and finding its spectrum can in general be extremely difficult.

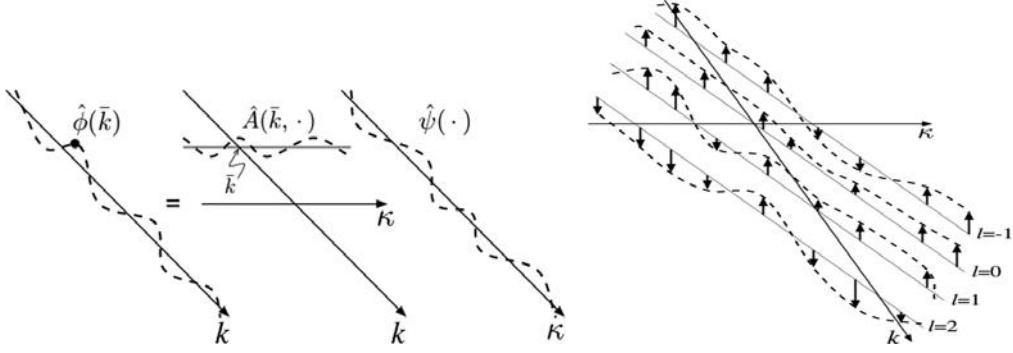


FIG. 1. Left: Pictorial representation of  $\hat{\phi}(\bar{k}) = \int_{\mathbb{R}} \hat{A}(\bar{k}, \kappa) \hat{\psi}(\kappa) d\kappa$ . Right: The frequency kernel  $\hat{A}$  of a spatially periodic operator  $A$ .

Once again we use perturbation methods to show (i) and (ii). We consider  $A$  to have the form  $A = A^\circ + \epsilon E$ , where  $A^\circ$  is a spatially invariant operator,  $E$  is a spatially periodic operator, and  $\epsilon$  is a small complex scalar. By using the biinfinite matrix representation, we first find conditions on the spatially invariant operator  $A^\circ$  such that (i) and (ii) are satisfied. We then show that (i) and (ii) will *remain* satisfied if  $\epsilon$  is small enough and if the spatially periodic operator  $E$  is “weaker” than  $A^\circ$  (in the sense that  $E$  is *relatively bounded* with respect to  $A^\circ$ ). The advantage of this approach is that (i) and (ii) are much easier to check for a spatially invariant operator than they are for a spatially periodic one.

**3. Frequency representation of periodic operators.** In this section we briefly discuss the frequency domain representation of spatially periodic operators. We then show how this representation can be used to convert a spatially periodic system into a family of matrix-valued LTI systems. For a detailed account the reader is referred to [5, 12].

Let  $\hat{\psi}(k)$  and  $\hat{\phi}(k)$  denote the Fourier transforms of two spatial functions  $\psi(x)$  and  $\phi(x)$ , respectively. If  $\psi$  and  $\phi$  are related by a linear operator  $A$  which admits a kernel representation, then  $\hat{\psi}$  and  $\hat{\phi}$  too are related by a linear operator  $\hat{A}$  which admits a kernel representation, and we have

$$(4) \quad \phi(x) = \int_{\mathbb{R}} A(x, \chi) \psi(\chi) d\chi \quad \xleftrightarrow{\mathcal{F}_x} \quad \hat{\phi}(k) = \int_{\mathbb{R}} \hat{A}(k, \kappa) \hat{\psi}(\kappa) d\kappa,$$

where  $A(\cdot, \cdot)$  and  $\hat{A}(\cdot, \cdot)$  are the *kernel functions* corresponding to the operators  $A$  and  $\hat{A}$ , respectively, and  $\mathcal{F}_x$  denotes the Fourier transform. Figure 1 (left) shows a cartoon way of picturing the equation  $\hat{\phi} = \hat{A} \hat{\psi}$ .

A linear operator is called *spatially periodic* with period  $X$  if its kernel has the property

$$A(x + Xm, \chi + Xm) = A(x, \chi) \quad \text{for all } x, \chi \in \mathbb{R}, m \in \mathbb{Z}.$$

**PROPOSITION 1.** *Let  $A$  be a spatially periodic operator with spatial period  $X = 2\pi/\Omega$ . Then its Fourier transform  $\hat{A}$  has the kernel representation*

$$(5) \quad \hat{A}(k, \kappa) = \sum_{l \in \mathbb{Z}} \hat{A}_l(k) \delta(k - \kappa - \Omega l).$$

*Proof.* See [12, Appendix to Chapter 2].  $\square$

Thus the kernel function corresponding to  $\hat{A}$  is composed of parallel and equally spaced “impulse sheets” which can be visualized in Figure 1 (right). References [5, 12] further describe how the particular structure (5) of  $\hat{A}$  allows the right-hand equation in (4) to be given a biinfinite<sup>1</sup> matrix representation

$$(6) \quad \begin{bmatrix} \vdots \\ \hat{\phi}(\theta - \Omega) \\ \hat{\phi}(\theta) \\ \hat{\phi}(\theta + \Omega) \\ \vdots \end{bmatrix} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \ddots \\ \cdots & \hat{A}_0(\theta - \Omega) & \hat{A}_{-1}(\theta - \Omega) & \hat{A}_{-2}(\theta - \Omega) & \cdots \\ \cdots & \hat{A}_1(\theta) & \hat{A}_0(\theta) & \hat{A}_{-1}(\theta) & \cdots \\ \cdots & \hat{A}_2(\theta + \Omega) & \hat{A}_1(\theta + \Omega) & \hat{A}_0(\theta + \Omega) & \cdots \\ \ddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \vdots \\ \hat{\psi}(\theta - \Omega) \\ \hat{\psi}(\theta) \\ \hat{\psi}(\theta + \Omega) \\ \vdots \end{bmatrix}$$

for which we adopt the notation

$$\phi_\theta = \mathcal{A}_\theta \psi_\theta, \quad \theta \in [0, \Omega).$$

In other words, a general spatially periodic operator  $A$  can be described by a family of biinfinite matrices  $\mathcal{A}_\theta$  parameterized by a variable  $\theta$ .

*Remark 1.* We emphasize that the functions  $\hat{A}_l(\cdot)$ ,  $l \in \mathbb{Z}$ , in (5) and (6) can be matrix-valued. Thus, in general,  $\mathcal{A}_\theta$  has a “block” structure. But throughout this paper and for the sake of brevity we choose not to explicitly refer to this block structure, even though all of our results are derived for matrix-valued  $\hat{A}_l(k)$ . Similarly, we do not refer explicitly to the Euclidean dimension of the vectors  $\hat{\psi}(k)$  and  $\hat{\phi}(k)$  for a given  $k \in \mathbb{R}$ .

*Spatially invariant* [13] and *spatially periodic pure multiplication* operators constitute special subclasses of spatially periodic operators. In the framework established above,  $\mathcal{A}_\theta$  is *diagonal* for spatially invariant operators and *Toeplitz* for periodic pure multiplication operators.

*Example 1.* The operators  $A = \partial_x$  and  $F(x) = \cos(\Omega x)$  have the following biinfinite representations

$$\mathcal{A}_\theta = \text{diag}\{j(\theta + \Omega n)\}_{n \in \mathbb{Z}}, \quad \mathcal{F} = \text{toep}\left\{\dots, 0, \frac{1}{2}, \boxed{0}, \frac{1}{2}, 0, \dots\right\},$$

where the boxed element corresponds to the main diagonal of the biinfinite Toeplitz operator  $\mathcal{F}$ . Notice that since  $A$  is spatially invariant it is fully described by its *Fourier symbol*  $\hat{A}_0(k) = jk$ ,  $k \in \mathbb{R}$ , and it is the samples of  $\hat{A}_0(\cdot)$  at  $k = \theta + \Omega n$ ,  $n \in \mathbb{Z}$ , that make up the diagonal of  $\mathcal{A}_\theta$  for a given  $\theta$ . We have dropped the  $\theta$  subscript in  $\mathcal{F}$  because it is independent of this variable.

*Remark 2.* It is possible to define a unitary operator  $\mathcal{M}_\theta$ ,  $\theta \in [0, \Omega)$ , such that  $\psi_\theta = \mathcal{M}_\theta \hat{\psi}$ ,  $\phi_\theta = \mathcal{M}_\theta \hat{\phi}$ , and thus  $\mathcal{A}_\theta = \mathcal{M}_\theta \hat{A} \mathcal{M}_\theta^*$  [12].  $\mathcal{M}_\theta$  is equivalent to a *frequency domain lifting* operation [14, 15, 16] which breaks an  $L^2(\mathbb{R})$  function into a family of  $\ell^2$  vectors. By the unitary property of the lifting operator it follows that

$$(7) \quad \int_0^\Omega \|\mathcal{A}_\theta\|_{\text{HS}}^2 d\theta = \int_0^\Omega \text{trace}[\mathcal{A}_\theta \mathcal{A}_\theta^*] d\theta = \sum_{l \in \mathbb{Z}} \int_{\mathbb{R}} \text{trace}[\hat{A}_l(k) \hat{A}_l^*(k)] dk,$$

with  $\|T\|_{\text{HS}}^2 := \text{trace}[T T^*]$  being the square of the Hilbert–Schmidt norm<sup>2</sup> of  $T$ .

<sup>1</sup>Sometimes referred to as doubly infinite.

<sup>2</sup>The Hilbert–Schmidt norm of an operator is a generalization of the Frobenius norm of finite-dimensional matrices  $\|A\|_{\text{F}}^2 = \sum_{m,n} |a_{mn}|^2 = \text{trace}[A A^*]$ .



Finally, given a spatially periodic system in state-space form (1) with spatially periodic operators  $A$ ,  $B$ , and  $C$ , one can replace each of these operators with its biinfinite matrix representation to obtain the family of LTI systems (2).

*Example 2.* Consider the spatially periodic heat equation on the real line

$$(8) \quad \begin{aligned} \partial_t \psi(t, x) &= (\partial_x^2 - \alpha \cos(\Omega x)) \psi(t, x) + u(t, x), \\ y(t, x) &= \psi(t, x), \end{aligned}$$

with real  $\alpha \neq 0$  and  $\Omega > 0$ .<sup>3</sup> Clearly  $A = \partial_x^2 + \alpha \cos(\Omega x)$  with domain

$$\mathcal{D} = \left\{ \phi \in L^2(\mathbb{R}) \mid \phi, \frac{d\phi}{dx} \text{ absolutely continuous, } \frac{d^2\phi}{dx^2} \in L^2(\mathbb{R}) \right\},$$

and  $B$  and  $C$  are equal to the identity operator on  $L^2(\mathbb{R})$ . Rewriting the system in its lifted representation, we have

$$(9) \quad \begin{aligned} \partial_t \psi_\theta(t) &= \mathcal{A}_\theta \psi_\theta(t) + u_\theta(t), \\ y_\theta(t) &= \psi_\theta(t), \end{aligned}$$

where  $\mathcal{B}_\theta$  and  $\mathcal{C}_\theta$  are equal to the identity operator on  $\ell^2$  and  $\mathcal{A}_\theta$  can be found from Example 1

$$\mathcal{A}_\theta = - \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \ddots \\ \cdots & (\theta + \Omega(n-1))^2 & \alpha/2 & 0 & \cdots \\ \cdots & \alpha/2 & (\theta + \Omega n)^2 & \alpha/2 & \cdots \\ \cdots & 0 & \alpha/2 & (\theta + \Omega(n+1))^2 & \cdots \\ \ddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Notice that (9) is now fully decoupled in the variable  $\theta$ . In other words, (8) is equivalent to the family of state-space representations (9) parameterized by  $\theta \in [0, \Omega)$ .

**4. Problem setup.** In this section we set the stage for the analysis that will follow in the proceeding sections. We first describe the class of spatially periodic LTI systems that we will be considering in this paper. We then review some facts and definitions regarding the spectrum of the infinitesimal generators of such systems, their exponential stability, and conditions under which the spectrum determines exponential stability. Finally, we introduce the notion of the  $\mathcal{H}^2$  norm for exponentially stable spatially periodic systems and give a procedure for computing this norm.

Let us consider a distributed system of the form

$$(10) \quad \begin{aligned} \partial_t \psi(t, x) &= A \psi(t, x) + B u(t, x) \\ &= (A^\circ + B^\circ \epsilon F C^\circ) \psi(t, x) + B u(t, x), \\ y(t, x) &= C \psi(t, x), \end{aligned}$$

where  $t \in [0, \infty)$  and  $x \in \mathbb{R}$  with the following assumptions. The (possibly unbounded) operators  $A^\circ$ ,  $B^\circ$ , and  $C^\circ$  are spatially invariant, and the bounded operators  $B$  and  $C$  are spatially periodic with period  $X = 2\pi/\Omega$ .  $F(x) = 2L \cos(\Omega x)$ , with  $L$  a constant

<sup>3</sup>By  $\partial_t \psi(t, x)$  and  $\partial_x^2 \psi(t, x)$  we mean the spatiotemporal functions  $\partial_t \psi$  and  $\partial_x^2 \psi$ , respectively, evaluated at the point  $(t, x)$ .

matrix, and  $\epsilon$  is a complex scalar.  $A^\circ$ ,  $B^\circ$ ,  $C^\circ$ , and  $E := B^\circ F C^\circ$  are all defined on a common dense domain  $\mathcal{D} \subset L^2(\mathbb{R})$ .  $A = A^\circ + \epsilon E$  is closed and generates a *strongly continuous* semigroup (also known as a  $C_0$  semigroup)  $e^{At}$  [2]. We will refer to  $A$  as the *infinitesimal generator* of the system. The functions  $u$ ,  $y$ , and  $\psi$  are, respectively, the spatiotemporal input, output, and state of the system and belong to  $L^2(\mathbb{R})$  for all  $t$ .

**Comment on notation.** To avoid clutter, we henceforth drop the “ $\wedge$ ” from the representation of frequency domain functions. For example, we use  $A_0(\cdot)$  [instead of  $\hat{A}_0(\cdot)$ ] to represent the Fourier symbol of the spatially invariant operator  $A^\circ$ .

As shown in [12, 5, 17], system (10) can be represented in the (spatial) frequency domain by the family of systems

$$\begin{aligned} \partial_t \psi_\theta(t) &= (\mathcal{A}_\theta^\circ + \epsilon \mathcal{B}_\theta^\circ \mathcal{F} \mathcal{C}_\theta^\circ) \psi_\theta(t) + \mathcal{B}_\theta u_\theta(t) \\ (11) \quad &= (\mathcal{A}_\theta^\circ + \epsilon \mathcal{E}_\theta) \psi_\theta(t) + \mathcal{B}_\theta u_\theta(t), \\ y_\theta(t) &= \mathcal{C}_\theta \psi_\theta(t), \end{aligned}$$

parameterized by  $\theta \in [0, \Omega)$ . The vectors  $u_\theta$ ,  $y_\theta$ , and  $\psi_\theta$  belong to  $\ell^2$  for all  $t$ .  $\mathcal{B}_\theta$  and  $\mathcal{C}_\theta$  have no particular structure and can be any bounded operators on  $\ell^2$ .  $\mathcal{F}$  has the form given in Example 1 with  $\frac{1}{2}$  replaced by the matrix  $L$ .  $\mathcal{A}_\theta^\circ$ ,  $\mathcal{B}_\theta^\circ$ , and  $\mathcal{C}_\theta^\circ$  are (possibly unbounded) block diagonal operators on  $\ell^2$ ,

$$\begin{aligned} \mathcal{A}_\theta^\circ &= \begin{bmatrix} \ddots & & \\ & A_0(\theta_n) & \\ & & \ddots \end{bmatrix}, \quad \mathcal{B}_\theta^\circ = \begin{bmatrix} \ddots & & \\ & B^\circ(\theta_n) & \\ & & \ddots \end{bmatrix}, \quad \mathcal{C}_\theta^\circ = \begin{bmatrix} \ddots & & \\ & C^\circ(\theta_n) & \\ & & \ddots \end{bmatrix}, \\ (12) \quad \mathcal{E}_\theta &:= \mathcal{B}_\theta^\circ \mathcal{F} \mathcal{C}_\theta^\circ = \begin{bmatrix} \ddots & & \ddots & & \\ & \ddots & & 0 & A_{-1}(\theta_n) \\ & & A_1(\theta_{n+1}) & 0 & \ddots \\ & & & \ddots & \ddots \end{bmatrix}, \end{aligned}$$

with  $\theta_n := \theta + \Omega n$ ,  $n \in \mathbb{Z}$ , and<sup>4</sup>

$$(13) \quad A_1(\cdot) := B^\circ(\cdot) L C^\circ(\cdot - \Omega), \quad A_{-1}(\cdot) := B^\circ(\cdot) L C^\circ(\cdot + \Omega).$$

$A_0(\cdot)$ ,  $B^\circ(\cdot)$ , and  $C^\circ(\cdot)$  denote the Fourier symbols of the spatially invariant operators  $A^\circ$ ,  $B^\circ$ , and  $C^\circ$ , respectively.

In sections 7 and 5 we will establish conditions for exponential stability of system (10) and compute its  $\mathcal{H}^2$  norm. We will see that the biinfinite representation (11)–(12) will play a key role in simplifying the perturbation analysis of system (10).

*Remark 3.* Taking  $F(x)$  to be a pure cosine is not restrictive. The results obtained here can be easily extended to problems where  $F(x)$  is any periodic function with absolutely convergent Fourier series coefficients.

*Remark 4.* The system (10) can be considered as the linear fractional transformation [18] of a spatially periodic system  $G^\circ$  with *spatially invariant* infinitesimal

<sup>4</sup>We emphasize that the notational convention used in the elements of  $\mathcal{E}_\theta$  is the same as that used in (6); the  $n$ th row of  $\mathcal{E}_\theta$  is  $\{\dots, 0, A_1(\theta_n), 0, A_{-1}(\theta_n), 0, \dots\}$ .

generator  $A^\circ$ ,

$$G^\circ = \left[ \begin{array}{c|cc} A^\circ & B & B^\circ \\ \hline C & 0 & 0 \\ \hline C^\circ & 0 & 0 \end{array} \right],$$

and the (memoryless and bounded) spatially periodic pure multiplication operator  $\epsilon F(x) = \epsilon 2L \cos(\Omega x)$ .

**Sectorial operators and exponential stability.** We introduce the class of *sectorial* operators. These operators generate *holomorphic* (also known as *analytic*) semigroups  $e^{At}$ . Sectorial operators have the important property that their spectrum determines the decay or growth rate of their semigroup; i.e., they satisfy the spectrum-determined growth condition.

Suppose  $A$  is densely defined,  $\rho(A)$  contains a sector of the complex plane  $|\arg(z - \alpha)| \leq \frac{\pi}{2} + \gamma$ ,  $\gamma > 0$ ,  $\alpha \in \mathbb{R}$ , and there exists some  $M > 0$  such that

$$(14) \quad \|(zI - A)^{-1}\| \leq \frac{M}{|z - \alpha|} \quad \text{for } |\arg(z - \alpha)| \leq \frac{\pi}{2} + \gamma.$$

Then  $A$  generates a holomorphic semigroup, and we write  $A \in \mathcal{H}(\gamma, \alpha, M)$  [11, 9]. We say that  $A$  is *sectorial* if  $A$  belongs to some  $\mathcal{H}(\gamma, \alpha, M)$ .

A semigroup is called exponentially stable if there exist positive constants  $M$  and  $\varrho$  such that [2]

$$\|e^{At}\| \leq M e^{-\varrho t} \quad \text{for } t \geq 0.$$

The following theorem constitutes the reason for our interest in sectorial operators and forms the foundation of our analysis in section 7.

**THEOREM 2.** *Assume that  $A$  is sectorial. Then  $A$  generates an exponentially stable  $C_0$  semigroup if and only if  $\Sigma(A) \subset \mathbb{C}^-$ .*

*Proof.* If  $A$  is sectorial, it defines a holomorphic  $C_0$  semigroup and  $e^{At}$  is differentiable for  $t > 0$  [10, 19]. Then [8] shows that this is sufficient for the spectrum-determined growth condition to hold. Since  $\Sigma(A) \subset \mathbb{C}^-$  and  $\Sigma(A)$  belongs to a left sector, it follows that  $\Sigma(A)$  is bounded away from the imaginary axis. Let  $\omega_\sigma = \sup_{z \in \Sigma(A)} \operatorname{Re}(z)$ . Then  $\omega_\sigma < 0$ , and  $A$  generates an exponentially stable  $C_0$  semigroup. Clearly  $\Sigma(A) \subset \mathbb{C}^-$  is also a necessary condition for exponential stability, and the proof is complete.  $\square$

**Spectrum of spatially periodic operators.** We show how the spectrum of  $A$  relates to the spectrum of the  $\theta$ -parameterized family of operators  $\mathcal{A}_\theta$ . We will use this in section 7 to find the spectrum of  $A$ , as needed in Theorem 2 to establish exponential stability.

Since the spatially periodic operator  $A$  and the family of biinfinite matrices  $\mathcal{A}_\theta$ ,  $\theta \in [0, \Omega)$ , are related via a unitary transformation, it follows that [5]

$$(15) \quad \Sigma(A) = \overline{\bigcup_{\theta \in [0, \Omega)} \Sigma(\mathcal{A}_\theta)}.$$

In the case where  $A$  is spatially invariant (and thus  $\mathcal{A}_\theta = \operatorname{diag}\{\dots, A_0(\theta_n), \dots\}$ ), (15) further simplifies to

$$(16) \quad \Sigma(A) = \overline{\bigcup_{k \in \mathbb{R}} \Sigma_p(A_0(k))}.$$

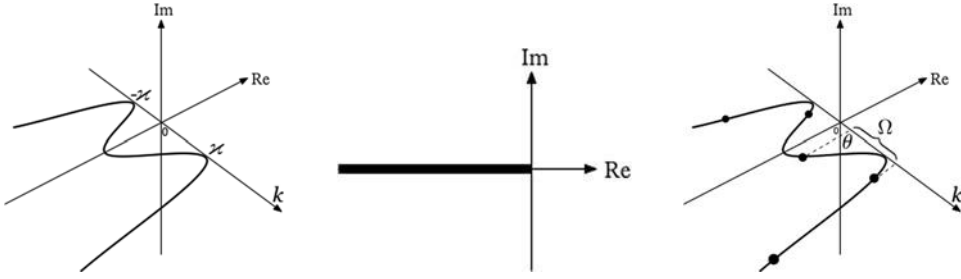


FIG. 2. Left: The symbol  $A_0(\cdot)$  of Example 3 in “complex-plane  $\times$  spatial-frequency” space. Center: The spectrum  $\Sigma(A)$  in the complex plane. Right: For spatially invariant  $A$ , the (diagonal) elements of  $A_\theta$  are samples of the Fourier symbol  $A_0(\cdot)$ .

Example 3. Let  $A = -(\partial_x^2 + \varkappa^2)^2$  with domain

$$(17) \quad \mathcal{D} = \left\{ \phi \in L^2(\mathbb{R}) \mid \phi, \frac{d\phi}{dx}, \frac{d^2\phi}{dx^2}, \frac{d^3\phi}{dx^3} \text{ absolutely continuous, } \frac{d^4\phi}{dx^4} \in L^2(\mathbb{R}) \right\}.$$

Integration by parts shows that  $A$  is a self-adjoint operator and thus closed. The function  $A_0(k) = -(k^2 - \varkappa^2)^2$  is the Fourier symbol of  $A$ ; see Figure 2 (left). Since  $A_0(\cdot)$  is scalar, we have  $\Sigma_p(A_0(k)) = A_0(k)$  for every  $k$ . It is easy to see that  $A_0(\cdot)$  takes every real negative value, and thus from (16)  $A$  has continuous spectrum  $\Sigma(A) = (-\infty, 0]$ ; see Figure 2 (center).

Remark 5. When  $A$  is spatially invariant, a helpful way to think about  $\Sigma(A)$  in terms of the symbol  $A_0(\cdot)$  of  $A$  is suggested by the previous example. First plot  $\Sigma_p(A_0(\cdot))$  in the “complex-plane  $\times$  spatial-frequency” space, as in Figure 2 (left). Then the orthogonal projection onto the complex plane of this plot would give  $\Sigma(A)$ . This can be considered as a geometric interpretation of (16). In Example 3, since  $A_0(\cdot)$  is real scalar and takes all negative values, this projection yields the negative real axis. But, in general, if  $A_0(\cdot) \in \mathbb{C}^{q \times q}$ , this projection would lead to  $q$  curves in the complex plane. Notice also that in this setting  $\Sigma(A_\theta)$  is the projection onto the complex plane of samples of  $\Sigma_p(A_0(\cdot))$  taken at  $k = \theta + \Omega n$ ,  $n \in \mathbb{Z}$ , in the complex-plane  $\times$  spatial-frequency space. As  $\theta$  varies in  $[0, \Omega)$  these projections trace out  $\Sigma(A)$  in the complex plane. This can be considered as a geometric interpretation of (15). Figure 2 (right) shows the said samples in the complex-plane  $\times$  spatial-frequency space for a scalar  $A$ .

**$\mathcal{H}^2$  norm of spatially periodic systems.** We define the  $\mathcal{H}^2$  norm of an exponentially stable spatially periodic system  $G$  as

$$\|G\|_{\mathcal{H}^2}^2 = \frac{1}{2\pi} \int_0^\Omega \int_0^\infty \text{trace}[\mathcal{G}_\theta(t) \mathcal{G}_\theta^*(t)] dt d\theta,$$

where  $\mathcal{G}_\theta(t) = C_\theta e^{A_\theta t} B_\theta$  is the impulse response of the system (11). The proof of the following theorem can be found in [5, 12].

THEOREM 3. Consider the exponentially stable spatially periodic LTI system  $G$ , with spatial period  $X = 2\pi/\Omega$  and state-space realization (10)–(11). Then

$$\|G\|_{\mathcal{H}^2}^2 = \frac{1}{2\pi} \int_0^\Omega \text{trace}[C_\theta \mathcal{P}_\theta C_\theta^*] d\theta = \frac{1}{2\pi} \int_0^\Omega \text{trace}[B_\theta^* \mathcal{Q}_\theta B_\theta] d\theta,$$

where  $\mathcal{P}_\theta$  and  $\mathcal{Q}_\theta$  are the solutions of the  $\theta$ -parameterized algebraic Lyapunov equations

$$\mathcal{A}_\theta \mathcal{P}_\theta + \mathcal{P}_\theta \mathcal{A}_\theta^* = -\mathcal{B}_\theta \mathcal{B}_\theta^*, \quad \mathcal{A}_\theta^* \mathcal{Q}_\theta + \mathcal{Q}_\theta \mathcal{A}_\theta = -\mathcal{C}_\theta^* \mathcal{C}_\theta.$$

**5. Perturbation analysis of the  $\mathcal{H}^2$  norm.** The difficulty in calculating the  $\mathcal{H}^2$  norm using Theorem 3 is that, unless  $\mathcal{A}_\theta$ ,  $\mathcal{B}_\theta$ , and  $\mathcal{C}_\theta$  are diagonal, the operators  $\mathcal{P}_\theta$  and  $\mathcal{Q}_\theta$  are “full,” meaning that they possess *all* of their (infinite number of) subdiagonals. This makes the computation of the  $\mathcal{H}^2$  norm numerically expensive. Namely, one has to solve an *infinite-dimensional* algebraic Lyapunov equation to find the operator  $\mathcal{P}_\theta$  (or  $\mathcal{Q}_\theta$ ) for every value of  $\theta \in [0, \Omega)$ . In this section we will see how one can use analytic perturbation techniques to compute the  $\mathcal{H}^2$  norm in a simple and numerically efficient way and without having to explicitly find the full  $\mathcal{P}_\theta$  and  $\mathcal{Q}_\theta$  operators. Such a perturbation analysis is very useful in predicting general trends and extracting valuable information about the  $\mathcal{H}^2$  norm, as needed in the design of periodic feedback.

Consider the general setup of (10), where we take  $\epsilon$  to be a small *real* scalar. We assume that  $A = A^\circ + \epsilon E$  defines an exponentially stable  $C_0$  semigroup on  $L^2(\mathbb{R})$  with finite  $\mathcal{H}^2$  norm for small enough  $\epsilon$  and that  $B$  and  $C$  are spatially invariant operators. We are interested in the changes in the  $\mathcal{H}^2$  norm of this system for small magnitudes of  $\epsilon$  and different values of the frequency  $\Omega$ .

Let

$$\mathcal{P}_\theta(\epsilon) = \mathcal{P}_\theta^{(0)} + \epsilon \mathcal{P}_\theta^{(1)} + \epsilon^2 \mathcal{P}_\theta^{(2)} + \dots,$$

with  $\mathcal{P}_\theta^*(\epsilon) = \mathcal{P}_\theta(\epsilon)$ . This implies that  $\mathcal{P}_\theta^{(m)*} = \mathcal{P}_\theta^{(m)}$  for all  $m = 0, 1, 2, \dots$ ; i.e.,  $\mathcal{P}_\theta^{(m)}$  are self-adjoint operators for all  $m$ . The proof of convergence of the above series is relegated to the appendix. Our aim is to find  $\mathcal{P}_\theta^{(m)}$  by solving the Lyapunov equation

$$(18) \quad \mathcal{A}_\theta(\epsilon) \mathcal{P}_\theta(\epsilon) + \mathcal{P}_\theta(\epsilon) \mathcal{A}_\theta^*(\epsilon) = -\mathcal{B}_\theta \mathcal{B}_\theta^*$$

$\Downarrow$

$$(19) \quad (\mathcal{A}_\theta^\circ + \epsilon \mathcal{E}_\theta) (\mathcal{P}_\theta^{(0)} + \epsilon \mathcal{P}_\theta^{(1)} + \epsilon^2 \mathcal{P}_\theta^{(2)} + \dots) + (\mathcal{P}_\theta^{(0)} + \epsilon \mathcal{P}_\theta^{(1)} + \epsilon^2 \mathcal{P}_\theta^{(2)} + \dots) (\mathcal{A}_\theta^\circ + \epsilon \mathcal{E}_\theta)^* = -\mathcal{B}_\theta \mathcal{B}_\theta^*$$

and to compute the  $\mathcal{H}^2$  norm of the system by using Theorem 3 and

$$\|G\|_{\mathcal{H}^2}^2 = \frac{1}{2\pi} \int_0^\Omega \text{trace}[\mathcal{C}_\theta \mathcal{P}_\theta(\epsilon) \mathcal{C}_\theta^*] d\theta.$$

By equating equal powers of  $\epsilon$  on both sides of (19), we have

$$(20) \quad \mathcal{A}_\theta^\circ \mathcal{P}_\theta^{(0)} + \mathcal{P}_\theta^{(0)} \mathcal{A}_\theta^{\circ*} = -\mathcal{B}_\theta \mathcal{B}_\theta^*,$$

$$(21) \quad \mathcal{A}_\theta^\circ \mathcal{P}_\theta^{(1)} + \mathcal{P}_\theta^{(1)} \mathcal{A}_\theta^{\circ*} = -(\mathcal{E}_\theta \mathcal{P}_\theta^{(0)} + \mathcal{P}_\theta^{(0)} \mathcal{E}_\theta^*),$$

$$(22) \quad \mathcal{A}_\theta^\circ \mathcal{P}_\theta^{(2)} + \mathcal{P}_\theta^{(2)} \mathcal{A}_\theta^{\circ*} = -(\mathcal{E}_\theta \mathcal{P}_\theta^{(1)} + \mathcal{P}_\theta^{(1)} \mathcal{E}_\theta^*),$$

$\vdots$

The existence of a unique solution to each of these equations is guaranteed by the exponential stability of the unperturbed system. Furthermore, in (20) since the operators  $\mathcal{A}_\theta^\circ$  and  $\mathcal{B}_\theta \mathcal{B}_\theta^*$  are diagonal, so is  $\mathcal{P}_\theta^{(0)}$ . In (21) the right-hand side operator has

the structure of being nonzero only on the first upper and lower subdiagonals, and hence  $\mathcal{P}_\theta^{(1)}$  inherits the same structure since  $\mathcal{A}_\theta^o$  is diagonal. In the same manner, the operator  $\mathcal{P}_\theta^{(2)}$  is nonzero only on the main diagonal and on the second upper and lower subdiagonals. This type of argument can be applied to all other  $\mathcal{P}_\theta^{(m)}$ ,  $m = 3, 4, \dots$ , and we have

$$\mathcal{P}_\theta^{(0)} = \begin{bmatrix} \ddots & & & \\ & P_0(\theta_n) & & \\ & & \ddots & \end{bmatrix}, \quad \mathcal{P}_\theta^{(1)} = \begin{bmatrix} \ddots & \ddots & & \\ \ddots & 0 & P_1^*(\theta_n) & \\ & P_1(\theta_n) & 0 & \ddots \\ & & \ddots & \ddots \end{bmatrix},$$

$$\mathcal{P}_\theta^{(2)} = \begin{bmatrix} \ddots & \ddots & \ddots & & \\ \ddots & \ddots & 0 & P_2^*(\theta_{n+1}) & \\ \ddots & 0 & Q_0(\theta_n) & 0 & \ddots \\ & P_2(\theta_{n+1}) & 0 & \ddots & \ddots \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \quad \dots,$$

where we have used the self-adjointness of these operators in writing the elements of their upper subdiagonals.

*Remark 6.* Note that not only is  $\mathcal{P}_\theta^{(m)}$  not a full operator, it has at most  $m$  nonzero upper and lower subdiagonals. Furthermore, all  $\mathcal{P}_\theta^{(m)}$  for odd  $m$  have zero diagonal and are thus trace-free operators.

*Remark 7.* Although  $\mathcal{A}_\theta = \mathcal{A}_\theta^o + \epsilon \mathcal{E}_\theta$  has only one nonzero subdiagonal, the operator  $\mathcal{P}_\theta(\epsilon) = \mathcal{P}_\theta^{(0)} + \epsilon \mathcal{P}_\theta^{(1)} + \dots$  possesses all of its subdiagonals. This is precisely the reason why direct calculation of  $\mathcal{P}_\theta$  in Theorem 3 is computationally difficult.

From (20)–(22) the operators  $\mathcal{P}_\theta^{(0)}$ ,  $\mathcal{P}_\theta^{(1)}$ , and  $\mathcal{P}_\theta^{(2)}$  are found by equating, element by element, the biinfinite matrices on both sides of these equations. For example, (20) leads to

$$A_0(\theta + \Omega n) P_0(\theta + \Omega n) + P_0(\theta + \Omega n) A_0^*(\theta + \Omega n) = -B(\theta + \Omega n) B^*(\theta + \Omega n)$$

for every  $n \in \mathbb{Z}$  and  $\theta \in [0, \Omega)$ . But as  $n$  takes all integer values and  $\theta$  changes in  $[0, \Omega)$ , the variable  $k = \theta + \Omega n$  takes all real values, and one can rewrite the above equation as

$$A_0(k) P_0(k) + P_0(k) A_0^*(k) = -B(k) B^*(k),$$

with  $k \in \mathbb{R}$ . By applying the same procedure to (21)–(22) one arrives at

$$(23) \quad A_0(k) P_0(k) + P_0(k) A_0^*(k) = -B(k) B^*(k),$$

$$(24) \quad A_0(k) P_1(k) + P_1(k) A_0^*(k - \Omega) = -(A_1(k) P_0(k - \Omega) + P_0(k) A_{-1}^*(k - \Omega)),$$

$$(25) \quad \begin{aligned} A_0(k) Q_0(k) + Q_0(k) A_0^*(k) &= -(A_1(k) P_1^*(k) + P_1(k) A_1^*(k) \\ &\quad + A_{-1}(k) P_1(k + \Omega) + P_1^*(k + \Omega) A_{-1}^*(k)), \end{aligned}$$

and so on for all nonzero diagonals of  $\mathcal{P}_\theta^{(m)}$ ,  $m = 3, 4, \dots$ . The existence of a unique solution to each of these equations is guaranteed by the exponential stability of the unperturbed system.

From the above equations one first finds  $P_0(\cdot)$  from (23), then  $P_1(\cdot)$  from (24), and so on. In other words, computing the subdiagonals of  $\mathcal{P}_\theta$  becomes *decoupled in one direction*. This decoupling would not have been possible had we not employed a perturbation approach and had attempted to solve (18) directly.

Returning to the calculation of the  $\mathcal{H}^2$  norm, let us first separate the diagonal part of  $\mathcal{P}_\theta^{(2)}$  by rewriting it as  $\mathcal{P}_\theta^{(2)} = \overline{\mathcal{P}}_\theta^{(2)} + \widetilde{\mathcal{P}}_\theta^{(2)}$ , where

$$\overline{\mathcal{P}}_\theta^{(2)} := \begin{bmatrix} \ddots & & \\ & Q_0(\theta_n) & \\ & & \ddots \end{bmatrix}$$

and  $\widetilde{\mathcal{P}}_\theta^{(2)}$  contains the rest of  $\mathcal{P}_\theta^{(2)}$ . Clearly  $\text{trace}[\mathcal{C}_\theta \widetilde{\mathcal{P}}_\theta^{(2)} \mathcal{C}_\theta^*] = 0$ . Also recall that

$$(26) \quad \text{trace}[\mathcal{C}_\theta \mathcal{P}_\theta^{(2m+1)} \mathcal{C}_\theta^*] = 0, \quad m = 0, 1, 2, \dots$$

Therefore

$$\begin{aligned} \|G\|_{\mathcal{H}^2}^2 &= \frac{1}{2\pi} \int_0^\Omega \text{trace}[\mathcal{C}_\theta \mathcal{P}_\theta(\epsilon) \mathcal{C}_\theta^*] d\theta \\ &= \frac{1}{2\pi} \int_0^\Omega \text{trace}[\mathcal{C}_\theta \mathcal{P}_\theta^{(0)} \mathcal{C}_\theta^* + \epsilon^2 \mathcal{C}_\theta \mathcal{P}_\theta^{(2)} \mathcal{C}_\theta^*] d\theta + O(\epsilon^4) \\ &= \frac{1}{2\pi} \int_0^\Omega \text{trace}[\mathcal{C}_\theta \mathcal{P}_\theta^{(0)} \mathcal{C}_\theta^* + \epsilon^2 \mathcal{C}_\theta \overline{\mathcal{P}}_\theta^{(2)} \mathcal{C}_\theta^*] d\theta + O(\epsilon^4), \end{aligned}$$

where the absence of odd powers of  $\epsilon$  is due to (26) and the last equation follows from the fact that  $\text{trace}[\mathcal{C}_\theta \widetilde{\mathcal{P}}_\theta^{(2)} \mathcal{C}_\theta^*] = 0$ . By using the unitary property of the lifting transform we have

$$\begin{aligned} \int_0^\Omega \text{trace}[\mathcal{C}_\theta \mathcal{P}_\theta^{(0)} \mathcal{C}_\theta^*] d\theta &= \int_{-\infty}^\infty \text{trace}[C(k) P_0(k) C^*(k)] dk, \\ \int_0^\Omega \text{trace}[\mathcal{C}_\theta \overline{\mathcal{P}}_\theta^{(2)} \mathcal{C}_\theta^*] d\theta &= \int_{-\infty}^\infty \text{trace}[C(k) Q_0(k) C^*(k)] dk, \end{aligned}$$

and we arrive at the final result

$$(27) \quad \|G\|_{\mathcal{H}^2}^2 = \frac{1}{2\pi} \int_{-\infty}^\infty \text{trace}[C(k) P_0(k) C^*(k) + \epsilon^2 C(k) Q_0(k) C^*(k)] dk + O(\epsilon^4).$$

We have thus proved the following theorem, which is the main result of this section.

**THEOREM 4.** *Consider the exponentially stable spatially periodic LTI system  $G$  with finite  $\mathcal{H}^2$  norm, spatial period  $X = 2\pi/\Omega$ , and state-space realization (10). Then for small values of  $|\epsilon|$  the  $\mathcal{H}^2$  norm of the system (10) can be computed from (27), where  $P_0(\cdot)$  and  $Q_0(\cdot)$  are solutions of the family of finite-dimensional Lyapunov and Sylvester equations described by (23)–(25).*

The described procedure can be continued to find higher-order terms in the perturbation series of the  $\mathcal{H}^2$  norm. The interested can refer to [20] for details.

**6. Examples.** As an application of the perturbation results of the previous section, we first investigate the occurrence of “parametric resonance” for a class of spatially periodic systems. Parametric resonance occurs when a specific frequency  $\Omega_{\text{res}}$

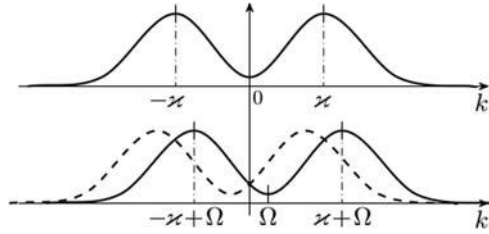


FIG. 3. *Top: Plot of  $P_0(\cdot)$ . Bottom: Plot of  $P_0(\cdot - \Omega)$  and  $P_0(\cdot + \Omega)$  (dashed line).*

of the periodic perturbation resonates with some “natural frequency”  $\varkappa$  of the unperturbed system, leading to a local (in  $\Omega$ ) change in system behavior [3]. In the systems we consider in this section this change in behavior is captured by the value of the  $\mathcal{H}^2$  norm.

*Example 4.* We consider the Swift–Hohenberg equation, which is of interest in hydrodynamics [21, 22, 23] and nonlinear optics [24, 25] as well as in other branches of physics [23]. The linearization of the Swift–Hohenberg equation around its time-independent spatially periodic solution leads to a PDE with spatially periodic coefficients of the form [26]

$$\begin{aligned} \partial_t \psi &= -(\partial_x^2 + \varkappa^2)^2 \psi - c\psi + f\psi + u, \\ y &= \psi, \end{aligned} \quad (28)$$

with  $0 \neq \varkappa \in \mathbb{R}$ ,  $c > 0$ , and  $f(x) = f(x + 2\pi/\Omega)$ . We assume here that  $f(x) = \epsilon \cos(\Omega x)$ , with  $\epsilon \in \mathbb{R}$  small. By comparing (28) and (10) we have

$$A_0(k) = -(k^2 - \varkappa^2)^2 - c, \quad B^o(k) = 1, \quad C^o(k) = 1, \quad B(k) = 1, \quad C(k) = 1, \quad L = \frac{1}{2}.$$

For this system the functions  $P_0(k)$  and  $Q_0(k)$  of the previous section simplify to<sup>5</sup>

$$\begin{aligned} (29) \quad P_0(k) &= \frac{-1}{2A_0(k)}, \\ Q_0(k) &= \frac{1}{(A_0(k))^2} \left( \frac{-1}{2A_0(k - \Omega)} + \frac{-1}{2A_0(k + \Omega)} \right) \\ (30) \quad &= 4(P_0(k))^2 (P_0(k - \Omega) + P_0(k + \Omega)), \end{aligned}$$

and our aim is to find the  $\mathcal{H}^2$  norm

$$(31) \quad \|G\|_{\mathcal{H}^2}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} (P_0(k) + \epsilon^2 Q_0(k)) dk + O(\epsilon^4)$$

for different values of the parameter  $\Omega > 0$ . More specifically, we are interested in the values of  $\Omega$  for which the  $\mathcal{H}^2$  norm is maximized.

From (29) we have  $P_0(k) = (1/2)/[(k^2 - \varkappa^2)^2 + c]$ . The first plot of Figure 3 shows  $P_0(\cdot)$ , while the second shows  $P_0(\cdot - \Omega)$  and  $P_0(\cdot + \Omega)$  [dashed line] for a given value of  $\Omega \neq 0$ . As  $\Omega$  is increased,  $P_0(\cdot - \Omega)$  slides to the right and  $P_0(\cdot + \Omega)$  to the left. From (30) it is clear that, to find  $Q_0(\cdot)$  for a given  $\Omega$ , one would sum the two functions in the second plot and multiply the result by the square of the first

<sup>5</sup>To find  $Q_0(k)$  one needs to first find  $P_1(k)$ , but we have omitted the details for brevity.



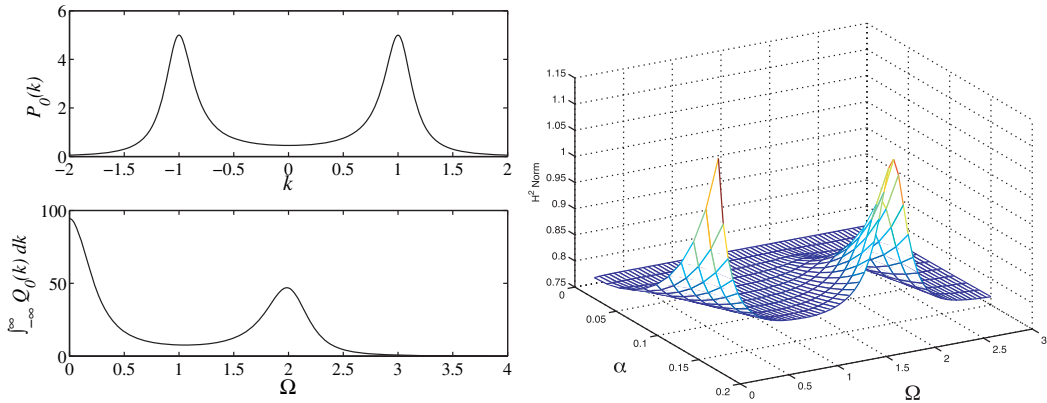


FIG. 4. Left: Plots of Example 4 for  $\kappa = 1$  and  $c = 0.1$ . Notice that the first graph is plotted against  $k$  and the second against  $\Omega$ . Right: The plot of the  $\mathcal{H}^2$  norm of the same example but calculated by taking large truncations of the  $\mathcal{A}_\theta$ ,  $\mathcal{B}_\theta$ , and  $\mathcal{C}_\theta$  matrices and using Theorem 3.

plot. The interesting question now is for what value(s) of  $\Omega \in (0, \infty)$  the  $\mathcal{H}^2$  norm in (31) would be maximized.

One can easily see that as  $\Omega \rightarrow 0$  the peaks of  $P_0(\cdot - \Omega)$  and  $P_0(\cdot + \Omega)$  merge toward those of  $(P_0(\cdot))^2$ . Thus  $\int_{-\infty}^{\infty} Q_0(k) dk$  grows, and hence  $\|G\|_{\mathcal{H}^2}^2$  grows.<sup>6</sup> This is not surprising; as  $\Omega \rightarrow 0$  the perturbation is tending toward a constant function  $F(x) = \cos(\Omega x) \rightarrow 1$ . This results in shifting the whole spectrum of  $A^0$  toward the right half of the complex plane, thus increasing the  $\mathcal{H}^2$  norm of the perturbed system.

But we are more interested in frequencies  $\Omega \gg 0$  that lead to a local (in  $\Omega$ ) increase in the  $\mathcal{H}^2$  norm. Notice that a different alignment of the peaks can also occur, which leads to another local maximum of the  $\mathcal{H}^2$  norm as a function of  $\Omega$ . This happens when the peak of  $P_0(\cdot - \Omega)$  at  $k = -\kappa + \Omega$  becomes aligned with the peak of  $(P_0(\cdot))^2$  at  $k = \kappa$  and, simultaneously, the peak of  $P_0(\cdot + \Omega)$  at  $k = \kappa - \Omega$  becomes aligned with the peak of  $(P_0(\cdot))^2$  at  $k = -\kappa$ . Clearly this occurs when

$$-\kappa + \Omega_{\text{res}} = \kappa \implies \Omega_{\text{res}} = 2\kappa.$$

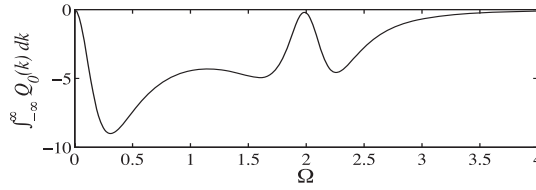
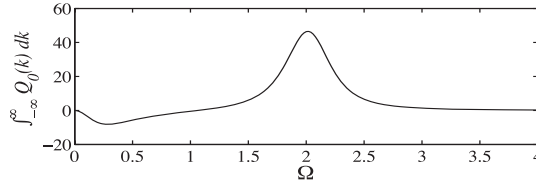
This result agrees exactly with that obtained in [27], where in the analysis of the same problem it is shown that the part of the spectrum of  $A$  closest to the imaginary axis “resonates” with perturbations whose frequency satisfies the relation  $\Omega = 2\kappa$ .

Consider (28) with  $\kappa = 1$  and  $c = 0.1$ . For this system  $\int_{-\infty}^{\infty} P_0(k) dk \approx 4.74$ . Figure 4 (left) shows the graphs of  $P_0(\cdot)$  plotted against  $k$  and  $\int_{-\infty}^{\infty} Q_0(k) dk$  plotted against  $\Omega$ . The peak at  $\Omega = 2$  in the lower plot verifies the relation  $\Omega_{\text{res}} = 2\kappa$  obtained previously.

Figure 4 (right) shows the  $\mathcal{H}^2$  norm of this system computed by taking large enough truncations [5] of the  $\mathcal{A}_\theta$ ,  $\mathcal{B}_\theta$ , and  $\mathcal{C}_\theta$  matrices and then applying Theorem 3. The figure shows that the trends were indeed correctly predicted by the perturbation analysis and the peaks at  $\Omega = 0, 2$  correspond to the peaks of  $\int_{-\infty}^{\infty} Q_0(k) dk$ .

Now consider (28) with  $\kappa = 1$  and  $c = 0.1$  but with  $\epsilon$  replaced by  $\epsilon j$ . This would correspond to  $L = j/2$  in the notation of (10). Obviously the unperturbed system

<sup>6</sup>Remember that  $P_0(k)$  is independent of  $\Omega$ , and thus  $\int_{-\infty}^{\infty} P_0(k) dk$  remains constant for different  $\Omega$ .

FIG. 5. The plot of Example 4 for  $\varkappa = 1$ ,  $c = 0.1$ , and a purely imaginary perturbation.FIG. 6. Plot of Example 5 for  $\varkappa = 1$  and  $c = 0.1$ .

remains the same as before, and hence  $\int_{-\infty}^{\infty} P_0(k) dk \approx 4.74$ . Figure 5 shows the graph of  $\int_{-\infty}^{\infty} Q_0(k) dk$  which demonstrates that for this system the purely imaginary perturbation reduces the  $\mathcal{H}^2$  norm at all frequencies. The physical interpretation of such a perturbation is investigated in [17].

*Example 5.* We consider a slightly different version of the Swift–Hohenberg equation in the previous example [5]

$$\begin{aligned} \partial_t \psi &= -(\partial_x^2 + \varkappa^2) \psi - c \psi + \epsilon \cos(\Omega x) \partial_x \psi + u, \\ y &= \psi. \end{aligned} \quad (32)$$

By comparing (32) and (10) we have

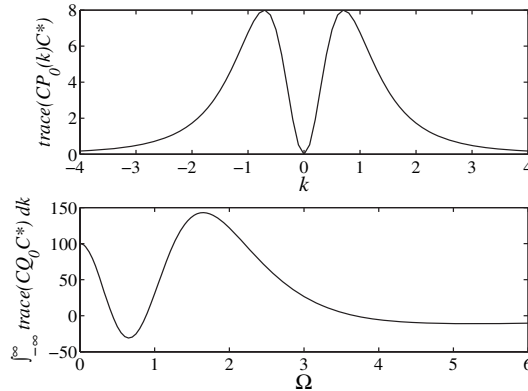
$$(33) \quad A_0(k) = -(k^2 - \varkappa^2)^2 - c, \quad B^o(k) = 1, \quad C^o(k) = jk, \quad B(k) = 1, \quad C(k) = 1, \quad L = \frac{1}{2}.$$

The difference between this system and (28) is that here  $C^o$  is a spatial derivative. The plot of Figure 6 demonstrates  $\int_{-\infty}^{\infty} Q_0(k) dk$  for  $\varkappa = 1$  and  $c = 0.1$ . Notice that the peak at  $\Omega_{\text{res}} = 2$  remains the same as in Figure 4, but we now have a decrease at small frequencies. This is due to the derivative operator  $C^o = \partial_x$ . These results are in agreement with the (nonperturbation) calculations for the same system in Example B, section VII of [5]. Our perturbation methods correctly predict the increase at  $\Omega = 2$  and the decrease around  $\Omega \approx 0.4$  of the  $\mathcal{H}^2$  norm.

*Example 6.* The system in this example is inspired by boundary layer and channel flow problems, where the introduction of corrugated walls or periodic body forces influences drag reduction or enhancement in such geometries. The following PDE has an analogous structure to the linearized Navier–Stokes equations in these geometries. Consider

$$A_0(k) = \begin{bmatrix} -\frac{1}{R}(k^2 + c) & 0 \\ jk & -\frac{1}{R}(k^2 + c) \end{bmatrix},$$

$$B^o(k) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C^o(k) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B(k) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad C(k) = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad L = \frac{1}{2} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$


 FIG. 7. Plots of Example 6 for  $R = 6$  and  $c = 1$ .

Numerical calculations for  $R = 6$ ,  $c = 1$  give  $\int_{-\infty}^{\infty} \text{trace}[C(k)P_0(k)C^*(k)] dk \approx 20.72$ . Figure 7 shows that the  $\mathcal{H}^2$  norm can be decreased by the application of periodic perturbations with frequency  $\Omega \approx 0.7$ . It is interesting that, if one uses the locations of the peaks in the first plot of Figure 7 to find  $\varkappa = 0.75$ , then from the peak at  $\Omega_{\text{res}} = 1.6$  in the second plot it seems that the relationship  $\Omega_{\text{res}} \approx 2\varkappa = 1.5$  (see Example 4) still holds with an acceptable error even for this matrix-valued system.

**7. Stability and the spectrum-determined growth condition.** In the literature on semigroups there exist examples in which  $\Sigma(A)$  lies entirely inside  $\mathbb{C}^-$  but  $\|e^{At}\|$  does not decay exponentially; see [6] and more recently [7]. In such cases it is said that the semigroup does not satisfy the *spectrum-determined growth condition* [8]. The determining factor in the examples presented in [6, 7] is the accumulation of the eigenvalues of  $\mathcal{A}_\theta$  around  $\pm j\infty - c$ ,  $c > 0$ , in the form of Jordan blocks of ever-increasing size (i.e., as the eigenvalues tend to  $\pm j\infty - c$ , their algebraic multiplicity increases while their geometric multiplicity stays finite). But such cases are ruled out when one deals with semigroups generated by *sectorial* operators.

Our ultimate aim in this section is to verify exponential stability. From Theorem 2, in order to prove exponential stability of a holomorphic  $C_0$  semigroup generated by a sectorial operator  $A$ , it is necessary and sufficient to verify that  $\Sigma(A) \subset \mathbb{C}^-$ . Hence in the first part of this section we give conditions under which the infinitesimal generator  $A$  in (10) is a sectorial operator. In the second part we find conditions which guarantee  $\Sigma(A) \subset \mathbb{C}^-$ .

Once again, the setup is that of (10) where  $\epsilon$  is a small complex scalar. In addition, assume that  $A_0(k) \in \mathbb{C}^{q \times q}$  is diagonalizable for every  $k \in \mathbb{R}$ .

**Conditions for sectorial infinitesimal generator.** To find conditions under which an infinitesimal generator  $A$  is sectorial, we have to verify (14), i.e., verify whether  $\|(zI - A)^{-1}\| \leq M/|z - \alpha|$  for all  $z$  belonging to some sector of  $\mathbb{C}$ . This involves finding the inverse of the operator  $zI - A$  and then calculating its norm. Such a computation can in general be very difficult. On the other hand, finding  $\|(zI - A^\circ)^{-1}\|$  is easy because of the spatial invariance of  $A^\circ$ . Indeed, from the norm-preserving property of the Fourier transform, it follows that  $\|(zI - A^\circ)^{-1}\| = \sup_{k \in \mathbb{R}} \|(zI - A_0(k))^{-1}\|$ .

Thus to establish conditions for  $A$  to be sectorial we again use perturbation theory. We first find conditions under which  $A^\circ$  is sectorial. We then show that  $A = A^\circ + \epsilon E$

remains sectorial if  $E$  is “weaker” than  $A^\circ$  in a certain sense we will describe and if  $\epsilon$  is small enough.

In the next theorem we present a condition for a spatially invariant operator  $A^\circ$  with Fourier symbol  $A_0(\cdot)$  to be sectorial.

**THEOREM 5.** *Let  $A_0(k)$  be diagonalizable for every  $k \in \mathbb{R}$ , and let  $U(k)$  be the transformation that diagonalizes  $A_0(k)$ , i.e.,  $A_0(k) = U(k) \Lambda(k) U^{-1}(k)$ , with  $\Lambda(k)$  diagonal. Let  $\kappa(k) := \|U(k)\| \|U^{-1}(k)\|$  denote the condition number of  $U(k)$ . If  $\sup_{k \in \mathbb{R}} \kappa(k) < \infty$ , and for every  $k \in \mathbb{R}$  the resolvent set  $\rho(A_0(k))$  contains a sector of the complex plane  $|\arg(z - \alpha)| < \frac{\pi}{2} + \gamma$ , with  $\gamma > 0$  and  $\alpha \in \mathbb{R}$  both independent of  $k$ , then  $A^\circ$  is sectorial.*

*Proof.* See the appendix.  $\square$

This theorem has a particularly simple interpretation when  $A_0(\cdot)$  is scalar. In this case  $\kappa(k) = 1$  for all  $k \in \mathbb{R}$ . Since  $A_0(\cdot)$  traces a curve in the complex plane, by Theorem 5 if this curve stays outside some sector  $|\arg(z - \alpha)| \leq \frac{\pi}{2} + \gamma$ ,  $\gamma > 0$ , of the complex plane, then  $A^\circ$  is sectorial.

The following theorem is the main result of this section and uses the notion of *relative boundedness* of one unbounded operator with respect to another unbounded operator [9] to prove that  $A = A^\circ + \epsilon E$  is sectorial.

**THEOREM 6.** *Let  $A^\circ$  with domain  $\mathcal{D}$  be a closed operator, and  $A^\circ \in \mathcal{H}(\gamma, \alpha, M)$ . Let  $E = B^\circ F C^\circ$  with domain  $\mathcal{D}' \supset \mathcal{D}$  be relatively bounded with respect to  $A^\circ$  such that*

$$(34) \quad \|E\psi\| \leq a \|\psi\| + b \|A^\circ \psi\|, \quad \psi \in \mathcal{D},$$

with  $0 \leq a < \infty$  and  $0 \leq b|\epsilon| < 1/(1 + M)$ . Then  $A = A^\circ + \epsilon E$  is sectorial, closed, and generates a  $C_0$  semigroup.

*Proof.* From (34) we have

$$\|\epsilon E \psi\| \leq a |\epsilon| \|\psi\| + b |\epsilon| \|A^\circ \psi\|.$$

Then from [11, Theorem 4.5.7] it follows that  $A = A^\circ + \epsilon E$  is sectorial for all  $\epsilon$  such that  $0 \leq b|\epsilon| < 1/(1 + M)$ . Since  $M > 0$  we have  $b|\epsilon| < 1$ , and [9, Theorem IV.1.1] gives that  $A$  is closed. Finally, it follows from [19, p. 100] that  $A$  is the generator of a  $C_0$  semigroup.  $\square$

This theorem says that if  $A^\circ$  is sectorial and closed, then so is  $A = A^\circ + \epsilon E$  if  $E$  is weaker than  $A^\circ$  in the sense of (34) and if  $|\epsilon|$  is small enough. Notice that at this point condition (34) cannot be reduced to a condition in terms of Fourier symbols as in Theorem 5 (i.e., a condition that can be checked pointwise in  $k$ ). This is because  $E$  is not spatially invariant. But once the exact form of the operators  $B^\circ$  and  $C^\circ$  is known, (34) can be simplified to a condition on the Fourier symbols of  $A^\circ$ ,  $B^\circ$ , and  $C^\circ$ . Let us clarify this statement with the aid of an example.

*Example 7.* Consider the spatially periodic system

$$\begin{aligned} \partial_t \psi &= -(\partial_x^2 + \varkappa^2)^2 \psi - c \psi + \epsilon \partial_x \cos(\Omega x) \psi + u, \\ y &= \psi, \end{aligned}$$

where  $\psi \in \mathcal{D}$ , and  $\mathcal{D}$  is defined as in (17). It is easy to see that  $A^\circ = -(\partial_x^2 + \varkappa^2)^2 - c$  is sectorial by Theorem 5 and closed by Example 3, and  $E = \partial_x \cos(\Omega x)$ . By formal differentiation we have

$$E \psi = \partial_x \cos(\Omega x) \psi = -\Omega \sin(\Omega x) \psi + \cos(\Omega x) \partial_x \psi.$$

By using the triangle inequality and  $\|\sin(\Omega x)\| = \|\cos(\Omega x)\| = 1$ , we have

$$(35) \quad \|E\psi\| \leq |\Omega| \|\psi\| + \|\partial_x \psi\|.$$

Thus we have effectively commuted out the bounded spatially periodic component of  $E$  and are left with only spatially invariant operators on the right of (35). Hence after applying a Fourier transformation to the right side of (35), a sufficient condition for (34) to hold is that

$$|\Omega| + |k| \leq a + b|(k^2 - \varkappa^2)^2 + c|, \quad k \in \mathbb{R},$$

which can be shown to be satisfied for large enough  $a > 0$  and  $b > 0$ . By using Theorem 6 we get that  $A$  is sectorial, closed, and the generator of a  $C_0$  semigroup.

*Remark 8.* The above example makes clear the notion of  $E$  being “weaker” than  $A^\circ$  that we mentioned at the beginning of this subsection. If in Example 7 we had  $B^\circ = \partial_x^\nu$ ,  $C^\circ = \partial_x^\mu$ , and  $\nu + \mu = 5$ , then  $E$  would contain a 5th-order derivative, whereas the highest order of  $\partial_x$  in  $A^\circ$  is 4. This would mean that (34) cannot be satisfied for any choice of  $a$  and  $b$ .

**Conditions for infinitesimal generator with spectrum in  $\mathbb{C}^-$ .** The final step in establishing exponential stability is to check whether  $\Sigma(A) \subset \mathbb{C}^-$ . Since this is, in general, a difficult problem, we proceed as follows. We consider the (block) diagonal operators  $\mathcal{A}_\theta^\circ$  and extend Geršgorin-type methods to find bounds on the location of  $\Sigma(\mathcal{A}_\theta)$ ,  $\mathcal{A}_\theta = \mathcal{A}_\theta^\circ + \epsilon \mathcal{E}_\theta$ . We then use this to find conditions on  $\epsilon \mathcal{E}_\theta$  that yield  $\Sigma(\mathcal{A}_\theta) \subset \mathbb{C}^-$  for all  $\theta \in [0, \Omega]$ .

In locating the spectrum of a finite-dimensional matrix  $T \in \mathbb{C}^{q \times q}$ , the theory of Geršgorin circles [28] provides us with a region of the complex plane that contains the eigenvalues of  $T$ . This region is composed of the union of  $q$  disks, the centers of which are the diagonal elements of  $T$ , and their radii depend on the magnitude of the off-diagonal elements [28]. This theory has also been extended to the case of finite-dimensional *block* matrices (i.e., matrices whose elements are themselves matrices) in [29]. We will further extend this theory to include biinfinite block matrices  $\mathcal{A}_\theta$ .

Take  $\mathfrak{B}_k$  to be the set of complex numbers  $z$  that satisfy

$$(36) \quad \sigma_{\min}(zI - A_0(k)) \leq |\epsilon| (\|A_{-1}(k)\| + \|A_1(k)\|),$$

where  $\sigma_{\min}(zI - A_0(k))$  denotes the smallest singular value of the matrix  $zI - A_0(k)$ .

LEMMA 7. *For every  $\theta$  we have  $\Sigma(\mathcal{A}_\theta) \subseteq \mathfrak{S}_\theta$ , where*

$$\mathfrak{S}_\theta = \overline{\bigcup_{n \in \mathbb{Z}} \mathfrak{B}_{\theta_n}}.$$

*Proof.* See the appendix.  $\square$

*Example 8.* Let us consider again the spatially periodic system in Example 5,

$$\begin{aligned} \partial_t \psi &= -(\partial_x^2 + \varkappa^2)^2 \psi - c\psi + \epsilon \cos(\Omega x) \partial_x \psi + u, \\ y &= \psi. \end{aligned}$$

From (33) and (13) we have  $A_1(k) = \frac{i}{2}(k - \Omega)$ ,  $A_{-1}(k) = \frac{i}{2}(k + \Omega)$ , and thus  $\|A_{-1}(k)\| + \|A_1(k)\| = \frac{1}{2}(|k - \Omega| + |k + \Omega|)$ . Hence (36) leads to

$$\sigma_{\min}(zI - A_0(k)) = |zI - A_0(k)| \leq \frac{|\epsilon|}{2} (|k - \Omega| + |k + \Omega|) = \begin{cases} \Omega |\epsilon| & |k| \leq \Omega, \\ |k| |\epsilon| & |k| \geq \Omega, \end{cases}$$

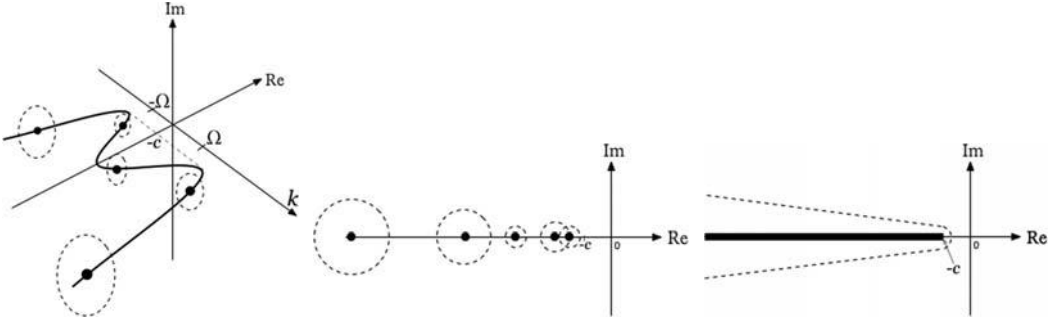


FIG. 8. Left: The  $\mathfrak{B}_{\theta_n}$  regions viewed in the complex-plane  $\times$  spatial-frequency space (the disks are parallel to the complex plane). Center:  $\Sigma(\mathcal{A}_\theta)$  is contained inside the union of the regions  $\mathfrak{B}_{\theta_n}$ . Right: The bold line shows  $\Sigma(A^\circ)$ , and the dotted region contains  $\Sigma(A)$ ,  $A = A^\circ + \epsilon E$ .

which means that the set  $\mathfrak{S}_\theta$  is composed of the union of disks with centers at  $A_0(\theta_n)$  and radii  $\frac{|\epsilon|}{2}(|\theta_n - \Omega| + |\theta_n + \Omega|)$ . This is nothing but an extension of the classical Geršgorin disks to biinfinite matrices. Figure 8 (left and center) show  $\mathfrak{S}_\theta$  in the complex-plane  $\times$  spatial-frequency space and in  $\mathbb{C}$ , respectively.<sup>7</sup>

*Remark 9.* The set

$$\begin{aligned}
 \Sigma_\varepsilon(T) &:= \{z \in \mathbb{C} \mid \sigma_{\min}(zI - T) \leq \varepsilon\} \\
 &= \{z \in \mathbb{C} \mid \|(zI - T)\varphi\| \leq \varepsilon \text{ for some } \|\varphi\| = 1\} \\
 (37) \quad &= \{z \in \mathbb{C} \mid z \in \Sigma_p(T + Z) \text{ for some } \|Z\| \leq \varepsilon\}
 \end{aligned}$$

is called the  $\varepsilon$ -pseudospectrum of the matrix  $T$  [30, 31]. Clearly  $\Sigma_{\varepsilon'}(T) \subseteq \Sigma_\varepsilon(T)$  if  $\varepsilon' \leq \varepsilon$ , and  $\Sigma_\varepsilon(T) = \Sigma_p(T)$  for  $\varepsilon = 0$ . The pseudospectrum is composed of small sets around the eigenvalues of  $T$ . For instance, if  $T$  has simple eigenvalues then for small enough values of  $\varepsilon$  the pseudospectrum consists of disjoint compact and convex neighborhoods of each eigenvalue [32]. By comparing (37) and the definition of  $\mathfrak{B}_k$  in (36), it is clear that  $\mathfrak{B}_k = \Sigma_\varepsilon(A_0(k))$ , with  $\varepsilon = |\epsilon|(\|A_{-1}(k)\| + \|A_1(k)\|)$ . Thus for every  $k \in \mathbb{R}$ , the inequality (36) defines a closed region of  $\mathbb{C}$  that includes the eigenvalues of  $A_0(k)$ .

We now employ Lemma 7 to determine whether  $\Sigma(A)$  resides completely inside  $\mathbb{C}^-$ , as needed to assess system stability. Take  $\mathfrak{D}_\varepsilon$  to be the closed disk of radius  $\varepsilon$  and center at the origin and  $\mathfrak{B}_k$  to be the region described by (36). Define the sum of sets by  $\mathfrak{U}_1 + \mathfrak{U}_2 = \{z \mid z = z_1 + z_2, z_1 \in \mathfrak{U}_1, z_2 \in \mathfrak{U}_2\}$ . For every  $k \in \mathbb{R}$  let  $\lambda_{\max}(k)$  represent the eigenvalue of  $A_0(k)$  with the maximum real part, and let  $\kappa(k)$  be defined as in Theorem 5.

**THEOREM 8.** For a given  $k$  the set  $\mathfrak{B}_k$  is contained inside  $\Sigma_p(A_0(k)) + \mathfrak{D}_{r(k)}$ , with

$$r(k) = |\epsilon|(\|A_{-1}(k)\| + \|A_1(k)\|)\kappa(k).$$

In particular, if  $\Sigma(A^\circ) \subset \mathbb{C}^-$  and

$$(38) \quad r(k) < |\operatorname{Re}(\lambda_{\max}(k))| + \beta$$

for every  $k \in \mathbb{R}$  and some  $\beta < 0$  independent of  $k$ , then  $\Sigma(A) \subset \mathbb{C}^-$ .

<sup>7</sup>We point out that Figure 8 (left) is technically incorrect; once the spatially invariant system is perturbed by a spatially periodic perturbation, it is no longer spatially invariant and thus cannot be represented by a Fourier symbol. Hence its spectrum can no longer be demonstrated in the complex-plane  $\times$  spatial-frequency space. Figure 8 (center) demonstrates the correct representation of the Geršgorin disks for  $\mathcal{A}_\theta$ .

*Proof.* See the appendix.  $\square$

*Example 9.* Consider the system of Example 8.  $\kappa(k) = 1$  since  $A_0(k)$  is scalar,  $|\operatorname{Re}(\lambda_{\max}(k))| = |(k^2 - \varkappa^2)^2 + c|$ , and

$$\|A_{-1}(k)\| + \|A_1(k)\| = \frac{1}{2} (|k - \Omega| + |k + \Omega|).$$

Thus condition (38) becomes

$$\frac{|\epsilon|}{2} (|k - \Omega| + |k + \Omega|) < |(k^2 - \varkappa^2)^2 + c| + \beta.$$

If this condition is satisfied for some  $\beta < 0$ , the dotted region in Figure 8 (right) will remain inside  $\mathbb{C}^-$ , and thus  $\Sigma(A) \subset \mathbb{C}^-$ .

In summary, to assess exponential stability we first find sufficient conditions on the infinitesimal generator  $A$  such that it belongs to the class of sectorial operators, for which the spectrum-determined growth condition holds. We then find sufficient conditions for  $A$  to have  $\mathbb{C}^-$  spectrum. We do this via an extension of Geršgorin circles to biinfinite (block) matrices.

**8. Conclusions and future work.** We use perturbation analysis to find a computationally efficient way of revealing trends in the  $\mathcal{H}^2$  norm of spatially periodic systems. We show that for certain classes of systems the periodicity can be chosen so as to increase the  $\mathcal{H}^2$  norm and induce parametric resonance. An application of this would be in fluid mixing problems. It is also shown that the  $\mathcal{H}^2$  norm can be made to decrease for an appropriate choice of the frequency of the perturbation. This would be the desired scenario, for example, in the design of the body of an aircraft. We demonstrate that for certain scalar systems the value of the spatial period that achieves the desired increase or decrease of the  $\mathcal{H}^2$  norm can be characterized exactly based on the description of the nominal system.

We also study the problem of verifying the exponential stability of a spatially periodic system. We do this by (i) finding conditions under which its infinitesimal generator is a sectorial operator (i.e., generates a holomorphic  $C_0$  semigroup) and thus satisfies the spectrum-determined growth condition and (ii) deriving conditions which guarantee that the infinitesimal generator has its spectrum contained inside the open left half of the complex plane.

The methods presented here can also be used in systems with many spatial directions. For example, consider the PDE

$$\psi_t = \psi_{yy} + \psi_{xx} + c\psi + \epsilon \cos(\Omega x)\psi,$$

with  $y \in [-1, 1]$  and  $x \in R$ . To put this system into the developed framework one would only have to perform a discrete approximation of the operator  $\partial_y^2$  with the appropriate boundary conditions. Furthermore, the techniques developed in this paper can be applied to spatially periodic systems defined on a torus with minor changes.

Future research in this direction would include an exact (analytical) characterization of the frequencies for which the  $\mathcal{H}^2$  norm is most increased or decreased for the general case of matrix-valued  $A_0(\cdot)$ . The perturbation methods presented here could also be generalized to biinfinite Sylvester equations which arise frequently in fluids problems.

### Appendix.

**Convergence of the perturbation series.** For the series expansion

$$\mathcal{P}_\theta(\epsilon) = \mathcal{P}_\theta^{(0)} + \epsilon \mathcal{P}_\theta^{(1)} + \epsilon^2 \mathcal{P}_\theta^{(2)} + \dots$$

to be valid, we must show that all elements of the biinfinite matrix  $\mathcal{P}_\theta(\epsilon)$  converge for  $\epsilon$  contained in some small enough neighborhood of the origin. Let us assume that  $B^\circ$  and  $C^\circ$  are bounded operators and (without loss of generality) that  $\sup_{k \in \mathbb{R}} \|B(k)\| = 1$ . Assume that  $\|e^{A_0(k)t}\| \leq M_k e^{\varrho_k t}$ , and define

$$M := \sup_{k \in \mathbb{R}} M_k < \infty, \quad -\alpha_0 := \sup_{k \in \mathbb{R}} \varrho_k < 0,$$

$$\alpha_1 := \max\{\sup_{k \in \mathbb{R}} \|A_1(k)\|, \sup_{k \in \mathbb{R}} \|A_{-1}(k)\|\}.$$

Notice that the finiteness of  $M$  and the negativity of  $-\alpha_0$  follow from the exponential stability of the unperturbed system. Now from (23) we have

$$P_0(k) = \int_0^\infty e^{A_0(k)t} B(k) B^*(k) e^{A_0^*(k)t} dt,$$

and therefore

$$\sup_{k \in \mathbb{R}} \|P_0(k)\| \leq \frac{M^2}{2\alpha_0} =: \mu.$$

Similarly, from (24) we have

$$P_1(k) = \int_0^\infty e^{A_0(k)t} (A_1(k) P_0(k - \Omega) + P_0(k) A_{-1}^*(k - \Omega)) e^{A_0^*(k - \Omega)t} dt,$$

and therefore

$$\sup_{k \in \mathbb{R}} \|P_1(k)\| \leq \frac{M^2}{2\alpha_0} (2\alpha_1 \mu) \leq \mu (4\alpha_1 \mu) = 4\alpha_1 \mu^2.$$

From (25) we have

$$Q_0(k) = \int_0^\infty e^{A_0(k)t} (A_1(k) P_1^*(k) + \dots + P_1^*(k + \Omega) A_{-1}^*(k)) e^{A_0^*(k)t} dt,$$

and therefore

$$\sup_{k \in \mathbb{R}} \|Q_0(k)\| \leq \frac{M^2}{2\alpha_0} (4\alpha_1 (4\alpha_1 \mu^2)) \leq \mu (4\alpha_1)^2 \mu^2 = (4\alpha_1)^2 \mu^3.$$

In fact it is possible to show that any element of  $\mathcal{P}_\theta^{(m)}$  is bounded by

$$(4\alpha_1)^m \mu^{(m+1)}.$$

Thus for all  $|\epsilon| < 4\alpha_1 \mu = 2M^2 \alpha_1 / \alpha_0$  the series expansion of  $\mathcal{P}_\theta(\epsilon)$  converges.

*Proof of Theorem 5.* Recall that  $A^\circ$  is a sectorial operator if  $\rho(A^\circ)$  contains a (right) sector of the complex plane  $|\arg(z - \alpha)| \leq \frac{\pi}{2} + \gamma$ ,  $\gamma > 0$ ,  $\alpha \in \mathbb{R}$ , and there exists some  $M > 0$  such that

$$|z - \alpha| \|(zI - A^\circ)^{-1}\| \leq M \quad \text{for} \quad |\arg(z - \alpha)| \leq \frac{\pi}{2} + \gamma.$$



Since  $A_0(k) \in \mathbb{C}^{q \times q}$  is diagonalizable for every  $k$ , there exists a matrix  $U(k)$  such that  $A_0(k) = U(k) \Lambda(k) U^{-1}(k)$ , with  $\Lambda(k)$  a diagonal matrix. Let  $\lambda_i(k)$ ,  $i = 1, \dots, q$ , denote the diagonal elements of  $\Lambda(k)$ , which are also the eigenvalues of  $A_0(k)$ . Then

$$\begin{aligned} |z - \alpha| \|(zI - A^\circ)^{-1}\| &\leq \sup_{k \in \mathbb{R}} \left( |z - \alpha| \|(zI - A_0(k))^{-1}\| \right) \\ &\leq \sup_{k \in \mathbb{R}} \left( |z - \alpha| \|U(k)\| \|U^{-1}(k)\| \|(zI - \Lambda(k))^{-1}\| \right) \\ &= \sup_{k \in \mathbb{R}} \left( \kappa(k) \frac{|z - \alpha|}{\text{dist}[z, \Sigma_p(A_0(k))]} \right) \\ &\leq \kappa_{\max} \sup_{k \in \mathbb{R}} \left( \frac{|z - \alpha|}{\text{dist}[z, \Sigma_p(A_0(k))]} \right), \end{aligned}$$

where  $\kappa_{\max} := \sup_{k \in \mathbb{R}} \kappa(k)$ .

Set  $M = (M' + 1) \kappa_{\max}$ , with  $M' > 0$ . Consider for a given  $k$  the region of the complex plane defined by

$$\kappa_{\max} \frac{|z - \alpha|}{\text{dist}[z, \Sigma_p(A_0(k))]} \geq M.$$

This region (which contains the eigenvalues  $\lambda_i(k)$ ) is contained inside the union of the disks

$$\kappa_{\max} \frac{|z - \alpha|}{|z - \lambda_i(k)|} \geq M, \quad i = 1, \dots, q,$$

which are themselves contained inside the larger disks

$$(A1) \quad |z - \lambda_i(k)| \leq \frac{|\lambda_i(k) - \alpha|}{M'}, \quad i = 1, \dots, q.$$

Notice that (A1) describes disks whose radii increase like  $|\lambda_i(k) - \alpha|/M'$ ,  $M' > 0$ , as their centers  $\lambda_i(k)$  grow distant from the point  $\alpha$ . A sufficient condition for these disks to belong to some open (left) sector of the complex plane  $|\arg(z - \alpha)| > \frac{\pi}{2} + \gamma$ ,  $\gamma > 0$ , for all  $k \in \mathbb{R}$  and large enough  $M'$  is that  $\Sigma_p(A_0(k))$ ,  $k \in \mathbb{R}$ , reside inside some open (left) sector of the complex plane  $|\arg(z - \alpha)| > \frac{\pi}{2} + \gamma'$ ,  $\gamma' > \gamma$ .

Finally, if the conditions of the previous paragraph are satisfied then for  $z \in \mathbb{C}$  that belong to the sector  $|\arg(z - \alpha)| \leq \frac{\pi}{2} + \gamma$  we have  $z \in \rho(A_0(k))$  and

$$\kappa_{\max} \sup_{k \in \mathbb{R}} \left( \frac{|z - \alpha|}{\text{dist}[z, \Sigma_p(A_0(k))]} \right) \leq M.$$

Thus  $|z - \alpha| \|(zI - A^\circ)^{-1}\| \leq M$ , and  $A^\circ$  is sectorial.  $\square$

*Proof of Lemma 7.* We use  $\Pi_N T \Pi_N$  to denote the  $(2N + 1) \times (2N + 1)$  [block] truncation of an operator  $T$  on  $\ell^2$ , where  $\Pi_N$  is the projection defined by

$$\text{diag} \left\{ \dots, 0, \underbrace{I, \dots, I}_{2N+1 \text{ times}}, \dots, I, 0, \dots \right\},$$

center  
↓

and  $I$  is the identity matrix. Notice that  $\Pi_N T \Pi_N$  is still an operator on  $\ell^2$ ; it is made from the biinfinite matrix  $T$  by replacing all entries outside the center  $(2N+1) \times (2N+1)$

block with zeros. We now form the finite-dimensional matrix  $\Pi_N \mathcal{A}_\theta \Pi_N|_{\Pi_N \ell^2}$  by restricting  $\Pi_N \mathcal{A}_\theta \Pi_N$  to the finite-dimensional space  $\Pi_N \ell^2$ . Clearly  $\Pi_N \mathcal{A}_\theta \Pi_N|_{\Pi_N \ell^2}$  has pure point spectrum. Hence, using a generalized form of the Gershgorin circle theorem [29] for finite-dimensional (block) matrices, we conclude that

$$\Sigma(\Pi_N \mathcal{A}_\theta \Pi_N|_{\Pi_N \ell^2}) \subset \bigcup_{|n| \leq N} \mathfrak{B}_{\theta_n} \subseteq \overline{\bigcup_{n \in \mathbb{Z}} \mathfrak{B}_{\theta_n}},$$

where  $\mathfrak{B}_{\theta_n}$  are regions of  $\mathbb{C}$  defined by (36). Since this holds for all  $N \geq 0$ , we have  $\Sigma(\mathcal{A}_\theta) \subseteq \mathfrak{S}_\theta$ .  $\square$

*Proof of Theorem 8.* If  $U(k)$  diagonalizes  $A_0(k)$ ,  $A_0(k) = U(k) \Lambda(k) U^{-1}(k)$ , and  $\kappa(k) = \|U(k)\| \|U^{-1}(k)\|$  denotes the condition number of  $U(k)$ , then from [33] the pseudospectrum of  $A_0(k)$  satisfies

$$(A2) \quad \Sigma_p(A_0(k)) + \mathfrak{D}_\varepsilon \subseteq \Sigma_\varepsilon(A_0(k)) \subseteq \Sigma_p(A_0(k)) + \mathfrak{D}_{\varepsilon \kappa(k)}$$

for all  $\varepsilon \geq 0$ . Thus the first statement of the theorem follows immediately from (A2) and the fact that  $\mathfrak{B}_k = \Sigma_\varepsilon(A_0(k))$ , with  $\varepsilon = |\epsilon| (\|A_{-1}(k)\| + \|A_1(k)\|)$  [see Remark 9].

To prove the second statement, let  $\mathbb{C}_\beta^-$  denote all complex numbers with a real part less than  $\beta \in \mathbb{R}$ . It follows from  $\Sigma(A^\circ) \subset \mathbb{C}^-$  that  $\Sigma(\mathcal{A}_\theta^\circ) \subset \mathbb{C}^-$  for every  $\theta$ . If (38) holds, then

$$\mathfrak{B}_{\theta_n} \subseteq \Sigma_p(A_0(\theta_n)) + \mathfrak{D}_{r(\theta_n)} \subset \mathbb{C}_\beta^-$$

for every  $n \in \mathbb{Z}$ , and from Lemma 7 we have  $\Sigma(\mathcal{A}_\theta) \subseteq \mathfrak{S}_\theta = \overline{\bigcup_{n \in \mathbb{Z}} \mathfrak{B}_{\theta_n}} \subset \mathbb{C}_{\beta'}^-$  for some  $\beta < \beta' < 0$  and every  $\theta$ . Thus  $\Sigma(A) \subset \mathbb{C}^-$ .

**Acknowledgments.** The first author thanks Prof. Mihai Putinar and Prof. Farhad Jafari for many valuable discussions and suggestions.

#### REFERENCES

- [1] E. W. KAMEN, *Stabilization of linear spatially-distributed continuous-time and discrete-time systems*, in Multidimensional Systems Theory, N. K. Bose, ed., Reidel, Boston, 1985.
- [2] R. F. CURTAIN AND H. J. ZWART, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.
- [3] V. ARNOLD, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1989.
- [4] S. M. MEERKOV, *Principle of vibrational control: Theory and applications*, IEEE Trans. Automat. Control, 25 (1980), pp. 755–762.
- [5] M. FARDAD, M. R. JOVANOVIĆ, AND B. BAMIEH, *Frequency analysis and norms of distributed spatially periodic systems*, IEEE Trans. Automat. Control, to appear.
- [6] J. ZABCZYK, *A note on  $C_0$ -semigroups*, Bull. Acad. Polon. Sci., 23 (1975), pp. 895–898.
- [7] M. RENARDY, *On the linear stability of hyperbolic PDEs and viscoelastic flows*, Z. Angew. Math. Phys., 45 (1994), pp. 854–865.
- [8] Z. LUO, B. GUO, AND O. MORGUL, *Stability and Stabilization of Infinite Dimensional Systems with Applications*, Springer-Verlag, New York, 1999.
- [9] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1995.
- [10] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, American Mathematical Society, Providence, RI, 1957.
- [11] M. MIKLAVČIČ, *Applied Functional Analysis and Partial Differential Equations*, World Scientific, River Edge, NJ, 1998.
- [12] M. FARDAD, *The Analysis of Distributed Spatially Periodic Systems*, Ph.D. thesis, University of California, Santa Barbara, 2006.
- [13] B. BAMIEH, F. PAGANINI, AND M. A. DAHLEH, *Distributed control of spatially invariant systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 1091–1107.

- [14] G. DULLERUD, *Control of Uncertain Sampled-Data Systems*, Birkhauser, Cambridge, MA, 1996.
- [15] M. ARAKI, Y. ITO, AND T. HAGIWARA, *Frequency response of sampled-data systems*, *Automatica*, 32 (1996), pp. 483–497.
- [16] E. MOELLERSTEDT, *Dynamic Analysis of Harmonics in Electrical Systems*, Ph.D. thesis, Department of Automatic Control, Lund Institute of Technology, 2000.
- [17] M. FARDAD AND B. BAMIEH, *A perturbation approach to the  $H^2$  analysis of spatially periodic systems*, in Proceedings of the 2005 American Control Conference, IEEE, Piscataway, NJ, 2005, pp. 4838–4843.
- [18] K. ZHOU, J. DOYLE, AND K. GLOVER, *Robust and Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1996.
- [19] K. ENGEL AND R. NAGEL, *One-Parameter Semigroups for Linear Evolution Equations*, Springer-Verlag, New York, 2000.
- [20] R. MOARREF AND M. R. JOVANOVIĆ, *Transition control using an array of streamwise vortices*, in Proceedings of the 45th IEEE Conference on Decision and Control, 2006, pp. 107–112.
- [21] J. SWIFT AND P. S. HOHENBERG, *Hydrodynamic fluctuations at the convective instability*, *Phys. Rev. A*, 15 (1977), p. 319.
- [22] I. S. ARANSON, K. A. GORSHKOV, A. S. LOMOV, AND M. I. RABINOVICH, *Stable particle-like solutions of multidimensional nonlinear fields*, *Phys. D*, 43 (1990), p. 435.
- [23] M. CROSS AND P. C. HOHENBERG, *Pattern formation outside of equilibrium*, *Rev. Modern Phys.*, 65 (1993), p. 851.
- [24] M. TLIDI, M. GEORGIOU, AND P. MANDEL, *Transverse patterns in nascent optical bistability*, *Phys. Rev. A*, 48 (1993), p. 4605.
- [25] S. LONGHI AND A. GERACI, *Swift-Hohenberg equation for optical parametric oscillators*, *Phys. Rev. A*, 54 (1996), p. 4581.
- [26] J. BURKE AND E. KNOBLOCH, *Localized states in the generalized Swift-Hohenberg equation*, *Phys. Rev. E*, 73 (2006), p. 056211.
- [27] M. FARDAD AND B. BAMIEH, *A perturbation analysis of parametric resonance and periodic control in spatially distributed systems*, in Proceedings of the 43rd IEEE Conference on Decision and Control, 2004, pp. 3786–3791.
- [28] R. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, London, 1985.
- [29] D. G. FEINGOLD AND R. S. VARGA, *Block diagonally dominant matrices and generalizations of the Gerschgorin circle theorem*, *Pacific J. Math.*, 12 (1962), pp. 1241–1250.
- [30] L. N. TREFETHEN, *Pseudospectra of matrices*, in Numerical Analysis 1991, Research Notes in Mathematics Series 260, Pitman, Boston, 1992, pp. 234–266.
- [31] L. N. TREFETHEN AND M. EMBREE, *Spectra and Pseudospectra*, Princeton University Press, Princeton, NJ, 2005.
- [32] J. V. BURKE, A. S. LEWIS, AND M. L. OVERTON, *Optimization and pseudospectra, with applications to robust stability*, *SIAM J. Matrix Anal. Appl.*, 25 (2003), pp. 80–104.
- [33] S. C. REDDY, P. J. SCHMID, AND D. S. HENNINGSON, *Pseudospectra of the Orr-Sommerfeld operator*, *SIAM J. Appl. Math.*, 53 (1993), pp. 15–47.

# A VISCOSITY SOLUTION APPROACH TO THE INFINITE-DIMENSIONAL HJB EQUATION RELATED TO A BOUNDARY CONTROL PROBLEM IN A TRANSPORT EQUATION\*

GIORGIO FABBRI†

**Abstract.** The paper concerns the infinite-dimensional Hamilton–Jacobi–Bellman equation related to an optimal control problem regulated by a linear transport equation with boundary control. A suitable viscosity solution approach is needed in view of the presence of the unbounded control-related term in the state equation in the Hilbert setting. An existence-and-uniqueness result is obtained.

**Key words.** Hamilton–Jacobi–Bellman equation, viscosity solution, boundary control

**AMS subject classifications.** 49J20, 49L25

**DOI.** 10.1137/050638813

**1. Introduction.** We study the Hamilton–Jacobi–Bellman (HJB) equation related to the infinite-dimensional formulation of an optimal control problem whose state equation is a PDE of transport type.

More precisely we consider the PDE

$$(1) \quad \begin{cases} \frac{\partial}{\partial s}x(s, r) + \beta \frac{\partial}{\partial r}x(s, r) = -\mu x(s, r) + \alpha(s, r), & (s, r) \in (0, +\infty) \times (0, \bar{s}), \\ x(s, 0) = a(s) & \text{if } s > 0, \\ x(0, r) = x^0(r) & \text{if } r \in [0, \bar{s}], \end{cases}$$

where  $\bar{s}, \beta$  are positive constants,  $\mu \in \mathbb{R}$ , the initial data  $x^0$  is in  $L^2(0, \bar{s})$ , and we consider two controls: A boundary control  $a$  is in  $L^2_{loc}([0, +\infty); \mathbb{R})$  and a distributed control  $\alpha \in L^2_{loc}([0, +\infty) \times [0, \bar{s}]; \mathbb{R})$ .<sup>1</sup>

By using the approach and the references described in section 2, the above equation can be written as an ordinary differential equation in the Hilbert space  $\mathcal{H} = L^2(0, \bar{s})$  as follows:

$$(2) \quad \begin{cases} \frac{d}{ds}x(s) = Ax(s) - \mu x(s) + \alpha(s) + \beta \delta_0 a(s), \\ x(0) = x^0, \end{cases}$$

where  $A$  is the generator of a suitable  $C_0$  semigroup and  $\delta_0$  is the Dirac delta in 0. Such an unbounded contribution in the Hilbert formulation comes from the presence in the PDE of a boundary control (see [8]). Besides we consider the problem of minimizing the cost functional

$$(3) \quad J(x, \alpha(\cdot), a(\cdot)) = \int_0^\infty e^{-\rho s} L(x(s), \alpha(s), a(s)) ds,$$

where  $\rho > 0$  and  $L$  is globally bounded and satisfies some Lipschitz-type condition, as better described in section 2. The HJB equation related to the control problem with

\*Received by the editors August 24, 2005; accepted for publication (in revised form) October 16, 2007; published electronically March 5, 2008.

<http://www.siam.org/journals/sicon/47-2/63881.html>

†Dipartimento di Scienze Economiche ed Aziendali, Università LUISS - Guido Carli, Roma, Italy (gfabbri@luiss.it).

<sup>1</sup>We write “ $-\mu x$ ” instead of “ $\mu x$ ” because it is the standard way to write the equation in the economic literature, where  $-\mu$  has the meaning of a depreciation factor (and only the case  $\mu \geq 0$  is used). Here we consider a generic  $\mu \in \mathbb{R}$ .

state equation (2) and target functional (3) is

$$(4) \quad \rho u(x) - \langle \nabla u(x), Ax \rangle - \langle \nabla u(x), -\mu x \rangle_{L^2(0, \bar{s})} \\ - \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( \langle \beta \delta_0(\nabla u(x)), a \rangle_{\mathbb{R}} + \langle \nabla u(x), \alpha \rangle_{L^2(0, \bar{s})} + L(x, \alpha, a) \right) = 0.$$

The sets  $\Gamma$  and  $\Sigma$  will be introduced in section 2; they are suitable subsets, respectively, of  $\mathbb{R}$  and  $\mathcal{H}$ . If we define the value function of the control problem as

$$V(x) \stackrel{\text{def}}{=} \inf_{(\alpha(\cdot), a(\cdot)) \in \mathcal{E} \times \mathcal{A}} J(x, \alpha(\cdot), a(\cdot)),$$

we wish to prove that  $V(\cdot)$  is the unique solution, in a suitable sense, of the HJB equation.

We use the viscosity approach. Our main problem is to write a suitable definition of the viscosity solution, so that an existence and uniqueness theorem can be derived for such a solution. The main difficulties we encounter, with respect to the existing literature, are in dealing with the boundary term and the nonanalyticity of the semigroup. We substantially follow the original idea of Crandall and Lions [14], [15]—with some changes, as the reader will rate in Definitions 2.14 and 2.15—of writing test functions as the sum of a “good part” as it is a regular function with differential in  $D(A^*)$  and a “bad part” represented by some radial function. The main problems arise from the evaluation of the boundary term on the radial part.

In order to write a working definition in our case, some further requirements are needed, such as a  $C^2$  regularity of the test functions, the presence of a “remainder term” in the definition of a sub/supersolution, and the  $B$ -Lipschitz continuity (see Definition 2.10) of the solution. This last feature guarantees that the maxima and the minima in the definition of a sub/supersolution remain in  $D(A^*)$  (see Proposition 4.1). Some other comments on the definition of solution (Definitions 2.14 and 2.15) need some technical details and can be found in Remark 2.17.

The technique used cannot be easily extended to treat a general nonlinear problem because we use the explicit form of the PDE that we give in (6). A nontrivial generalization would be also that of replacing the constant  $\mu$  with a function  $\mu(r)$  in  $L^\infty(0, \bar{s})$  (see Remark 4.9 for details). Nevertheless, the problem remains challenging.

**A brief summary of the literature.** Hamilton–Jacobi equations in infinite dimensions, especially when arising from optimal control problems in Hilbert spaces, were first studied by Barbu and Da Prato [3], [4] with a strong solutions approach. The viscosity method, introduced in the study of finite-dimensional HJ equations in [13], was generalized by the same authors in a series of works: The most important for our approach are [14] and [15]. Moreover new variants of the notion of a viscosity solution for HJB equations in Hilbert spaces were given in [28], [34], [35], [33], [16].

The study of viscosity solutions for HJB equations associated to boundary control in PDE is more recent. In this research field there is not an organic and complete theory but some works that adapt the ideas and the techniques of viscosity solutions to problems for particular PDEs, by exploiting their peculiarities (as we do in this work for the problem regulated by a transport equation). For the first order HJB equations see [9], [12] (see also [10], [11]), where some classes of parabolic equations are treated, and [24], where the authors study the HJB equation related to a two-dimensional Navier–Stokes equation (see also [32]). It must be noted that all of these works treat the case of analytic  $A$ .

A HJB equation like (4) was treated, but only in the case of a convex objective functional, with a strong solutions approach by adapting Barbu and Da Prato's technique of convex regularization in [20], [19].

**A motivating economic problem.** Transport equations are used to model a large variety of phenomena. They are used, for example, in age-structured population models (see, for instance, [26], [2], [27]), in population economics [23], epidemiologic studies, socioeconomic science, and transport phenomena in physics.

Problems such as (1) can be used to describe, in economics, capital accumulation processes where an heterogeneous capital is involved, and this is the reason why the study of the infinite-dimensional control problem is of growing interest in the economic fields. For instance, in the vintage capital models  $x(t, s)$  may be regarded as the stock of capital goods differentiated with respect to the time  $t$  and the age  $s$ . Heterogeneous capital, in both the finite- and infinite-dimensional approaches, is used to study depreciation and obsolescence of physical capital, geographical difference in growth, innovation, and R&D.

Regarding problems modeled by a transport equation where an infinite-dimensional setting is used, we cite the following papers: [5], [7] on optimal technology adoption in a vintage capital context (in the case of a quadratic cost functional), [25] on capital accumulation, [6] on optimal advertising, and [19], [21] on the case of a general objective convex functional with a strong solutions approach. See also [22].

Moreover, we mention that the infinite-dimensional approach may apply to problems such as issuance of public debt (see [1] for a description of the problem). In that problem a stochastic setting and simple state-control constraints appear, but hopefully the present work can be a first step in this direction.

**Plan of the paper.** The work is organized as follows: In the first section we recall some results on the state equation, we introduce some preliminary remarks on the main operators involved in the problem, we explain some notations, we define the HJB equation, and we give the definition of solution. The second section regards some properties of the value function (in particular, some regularity properties) that will be used in the third section to prove that it is the unique (viscosity) solution of the HJB equation.

## 2. Notation and preliminary results.

**2.1. State equation.** In this subsection we will see some properties of the state equation: We write it in three different (and equivalent) forms that point out different properties of the solution. We will use all three forms in the following proofs.

We consider the PDE on  $[0, +\infty) \times [0, \bar{s}]$  given by

$$(5) \quad \begin{cases} \frac{\partial}{\partial s} x(s, r) + \beta \frac{\partial}{\partial r} x(s, r) = -\mu x(s, r) + \alpha(s, r), & (s, r) \in (0, +\infty) \times (0, \bar{s}), \\ x(s, 0) = a(s) & \text{if } s > 0, \\ x(0, r) = x^0(r) & \text{if } r \in [0, \bar{s}]. \end{cases}$$

Given an initial datum  $x^0 \in L^2((0, \bar{s}); \mathbb{R})$  (from now on simply  $L^2(0, \bar{s})$ ), a boundary control  $a(\cdot) \in L^2_{loc}([0, +\infty); \mathbb{R})$ , and a distributed control  $\alpha(\cdot) \in L^2_{loc}([0, +\infty) \times [0, \bar{s}]; \mathbb{R})$ , (5) has a unique solution in  $L^2_{loc}([0, +\infty) \times [0, \bar{s}]; \mathbb{R})$  given by

$$(6) \quad x(s, r) = \begin{cases} e^{-\mu s} x^0(r - \beta s) + \int_0^s e^{-\mu \tau} \alpha(s - \tau, r - \beta \tau) d\tau, & r \in [\beta s, \bar{s}], \\ e^{-\frac{\mu}{\beta} r} a(s - r/\beta) + \int_0^{r/\beta} e^{-\mu \tau} \alpha(s - \tau, r - \beta \tau) d\tau, & r \in [0, \beta s]. \end{cases}$$

In the following  $x(s, r)$  is the function defined in (6).

We can rewrite such an equation in a suitable Hilbert space setting. We take the Hilbert space  $\mathcal{H} \stackrel{\text{def}}{=} L^2(0, \bar{s})$  and the  $C_0$  semigroup  $S(t)$  given by

$$S(s)f[r] \stackrel{\text{def}}{=} \begin{cases} f(r - \beta s) & \text{for } r \in [\beta s, \bar{s}], \\ 0 & \text{for } r \in [0, \beta s]. \end{cases}$$

The generator of  $S(s)$  is the operator  $A$  given by

$$\begin{cases} D(A) = \{f \in H^1[0, \bar{s}] : f(0) = 0\}, \\ A(f)[r] = -\beta \frac{d}{dr} f(r) \end{cases}$$

(see [7] for a proof in the case where  $\beta = 1$ ; the proof in our case can be obtained by simply taking  $s' = \beta s$ ). In the following we will use the notation  $e^{sA}$  instead of  $S(s)$ .

*Remark 2.1.* To avoid confusion if  $x \in L^2(0, \bar{s})$  we will use  $[\cdot]$  to denote the pointwise evaluation, so  $x[r]$  is the value of  $x$  in  $r \in [0, \bar{s}]$ . On the other hand,  $x(s)$  will denote the evolution of the solution of the state equation (in the Hilbert space) at time  $s$  (as in (7)). That is,  $x(s)$  is an element of  $\mathcal{H}$ , while  $x[r]$  is a real number.

We want to write an infinite-dimensional formulation of (5), but in  $L^2(0, \bar{s})$  it should appear like

$$(7) \quad \begin{cases} \frac{d}{ds} x(s) = Ax(s) - \mu x(s) + \alpha(s) + \beta \delta_0 a(s), \\ x(0) = x^0, \end{cases}$$

where  $\alpha(s) \in L^2(0, \bar{s})$  is the function  $r \mapsto \alpha(s, r)$ . Such an expression does not make sense in  $L^2(0, \bar{s})$  for the presence of the unbounded term  $\beta \delta_0 a(s)$ . We can anyway apply formally the variation of constants method to (7) and obtain a mild form of (7) that is continuous from  $[0, +\infty)$  to  $L^2(0, \bar{s})$ . This is what we do in the next definition.

**DEFINITION 2.2.** Given  $x^0 \in L^2(0, \bar{s})$ ,  $a(\cdot) \in L^2_{loc}([0, +\infty); \mathbb{R})$ , and  $\alpha(\cdot) \in L^2_{loc}([0, +\infty); L^2(0, \bar{s}))$ , the function in  $C([0, +\infty); L^2(0, \bar{s}))$  given by

$$(8) \quad x(s) = e^{-\mu s} e^{sA} x^0 - A \int_0^s e^{-\mu(s-\tau)} e^{(s-\tau)A} (a(\tau)\nu) d\tau + \int_0^s e^{-\mu(s-\tau)} e^{(s-\tau)A} \alpha(\tau) d\tau,$$

where

$$\begin{aligned} \nu: [0, \bar{s}] &\rightarrow \mathbb{R}, \\ \nu: r &\mapsto e^{-\frac{\mu}{\beta} r}, \end{aligned}$$

is called the mild solution of (7).

*Remark 2.3.* We could include the term  $-\mu x$  in the generator of the semigroup  $A$  taking a  $\tilde{A} = A - \mu \mathbb{1}$  as done in [7]. The problem of this approach is that often we will use, in the estimates, the dissipativity of the generator, and  $\tilde{A}$  is dissipative only if  $\mu \geq 0$ .

**PROPOSITION 2.4.** Given  $x(s)$  the function from  $\mathbb{R}^+$  to  $L^2(0, \bar{s})$  given by (8) and  $x(s, r)$  the function from  $\mathbb{R}^+ \times [0, \bar{s}]$  to  $\mathbb{R}$  given by (6), we have  $x(s)[r] = x(s, r)$ .

*Proof.* See [7].  $\square$

Eventually we observe that (7) can be rewritten in a precise way in a larger space in which  $\beta \delta_0$  belongs. To this extent, we consider the adjoint operator  $A^*$ , whose explicit expression is given by

$$\begin{cases} D(A^*) \stackrel{\text{def}}{=} \{f \in H^1(0, \bar{s}) : f(\bar{s}) = 0\}, \\ A^*(f)[r] = \beta \frac{d}{dr} f(r). \end{cases}$$

We endow, in all of the paper,  $D(A^*)$  with the graph norm and the related Hilbert structure. We consider the inclusion

$$i: D(A^*) \hookrightarrow L^2(0, \bar{s})$$

and its continuous adjoint

$$i^*: L^2(0, \bar{s}) \rightarrow D(A^*)',$$

where we have identified  $L^2$  with its dual.

We can extend  $A$  to a generator of a  $C_0$  semigroup on  $D(A^*)'$  (the domain of the extension will contain  $L^2(0, \bar{s})$ ), and we observe that Dirac's measure  $\delta_0 \in D(A^*)'$  (see [20, Proposition 4.5, page 60] for details).

**PROPOSITION 2.5.** *Given  $T > 0$ ,  $x^0 \in L^2(0, \bar{s})$ ,  $a(\cdot) \in L^2(0, T)$ ,  $\alpha(\cdot) \in L^2([0, T]; L^2(0, \bar{s}))$ , (8) is the unique solution of*

$$(9) \quad \begin{cases} \frac{d}{ds} i^* x(s) = Ax(s) - \mu x(s) + \alpha(s) + \beta \delta_0 a(s), \\ x(0) = x^0 \end{cases}$$

in  $W^{1,2}(0, T; D(A^*)') \cap C(0, T, \mathcal{H})$ . Moreover, if  $a(\cdot) \in W^{1,2}(0, T)$ , then such a solution will belong to  $C^1(0, T; D(A^*)') \cap C(0, T; \mathcal{H})$ .

*Proof.* See [8, Chapter 3.2] (in particular, Theorem 3.1, page 173).  $\square$

**2.2. The definition of the operator  $B$ .** In this subsection we give the definition of the operator  $B$  that will have a fundamental role.<sup>2</sup>

Note that  $A^*$  and  $A$  are negative operators. We take  $\phi \in D(A^*)$ , so that  $\phi(\bar{s}) = 0$ , and then

$$\langle A^* \phi, \phi \rangle = \int_0^{\bar{s}} \beta \phi'(r) \phi(r) dr = \frac{-\beta \phi(0)^2}{2}$$

and for  $\phi \in D(A)$  (so that  $\phi(0) = 0$ )

$$\langle A \phi, \phi \rangle = \int_0^{\bar{s}} -\beta \phi'(r) \phi(r) dr = \frac{-\beta \phi(\bar{s})^2}{2}.$$

Therefore, given a  $\lambda > 0$ , the operators  $(A - \lambda I)$  and  $(A^* - \lambda I)$  are strongly negative:  $\langle (A - \lambda I)x, x \rangle \leq -\lambda |x|_{\mathcal{H}}^2$  for all  $x \in D(A)$  and  $\langle (A^* - \lambda I)x, x \rangle \leq -\lambda |x|_{\mathcal{H}}^2$  for all  $x \in D(A^*)$ .

We can also directly prove that

$$(A - \lambda I)^{-1}: \mathcal{H} \rightarrow D(A)$$

is a continuous negative linear operator whose explicit expression is given by

$$(A - \lambda I)^{-1}(\phi)[r] = \frac{1}{\beta} \left( -e^{-\frac{\lambda}{\beta} r} \int_0^r e^{\frac{\lambda}{\beta} \tau} \phi(\tau) d\tau \right).$$

The continuity can be proved directly with not difficult estimates, and the negativity can be proved directly by using an integration by part argument.

<sup>2</sup>We could use an abstract approach, noting that  $A$  and  $A^*$  are both generators of  $C_0$  semigroups of contractions, and then both are negative (see [17, page 424]), and the set  $\{\lambda \in \mathbb{C} : \operatorname{Re}(\lambda) > 0\}$  is in the resolvent of both  $A$  and  $A^*$  (Hille–Yosida theorem; see [29, page 53]). Here we can also follow a more direct approach that allows us to find the explicit form of the operator.



In the same way we can prove that

$$(A^* - \lambda I)^{-1}: \mathcal{H} \rightarrow D(A^*)$$

is a continuous and negative linear operator and that its explicit expression is given by

$$(A^* - \lambda I)^{-1}(\phi)[r] = \frac{1}{\beta} \left( -e^{\frac{\lambda}{\beta}r} \int_r^{\bar{s}} e^{-\frac{\lambda}{\beta}\tau} \phi(\tau) d\tau \right).$$

Eventually we can define  $B \stackrel{\text{def}}{=} (A^* - \lambda I)^{-1}(A - \lambda I)^{-1} = ((A - \lambda I)^{-1})^*(A - \lambda I)^{-1}$  that is continuous, positive, and self-adjoint.<sup>3</sup> Moreover

$$(A^* - \lambda I)B = (A - \lambda I)^{-1} \leq 0,$$

and so

$$A^*B = (A - \lambda I)^{-1} + \lambda B \leq \lambda B;$$

then  $A^*B$  is continuous, and if we choose  $\lambda < 1$ , we have

$$(10) \quad A^*B \leq B.$$

Thus  $B$  satisfies all requirements of the so-called “weak case” of [14].

*Remark 2.6.* Note that  $B^{1/2}$  is a particular case of the operator that Renardy found in more generality in [31], and so  $B^{1/2}: \mathcal{H} \rightarrow D(A^*)$  continuously and in particular  $\mathcal{R}(B^{1/2}) \subseteq D(A^*)$ .

*Notation 2.7.* For every  $x \in \mathcal{H}$  we will indicate with  $|x|_B$  the  $B$ -norm that is  $\sqrt{\langle Bx, x \rangle_{\mathcal{H}}}$ . We will write  $\mathcal{H}_B$  for the completion of  $\mathcal{H}$  with respect to the  $B$ -norm.

*Remark 2.8.* Thanks to the definition of  $A^*$ , the graph norm on  $D(A^*)$  is equivalent to the  $H^1(0, \bar{s})$ -norm. In particular  $D(A^*)$  is the completion of

$$K = \{f|_{[0, \bar{s}]} : f \in C_c^\infty(\mathbb{R}) \text{ with } \text{supp}(f) \subseteq (-\infty, \bar{s})\}$$

with respect to the  $H^1(0, \bar{s})$ -norm. So, since  $H^1(0, \bar{s}) \hookrightarrow C([0, \bar{s}]; \mathbb{R})$ , we can apply  $\beta\delta_0$  on the elements of  $D(A^*)$  and  $\delta_0 \in D(A^*)'$ .

*Notation 2.9.* The notation  $\langle x, y \rangle_H$  will indicate the inner product in the Hilbert space  $H$  (for example,  $H = \mathcal{H} \equiv L^2(0, \bar{s})$  or  $H = \mathbb{R}$  or  $D(A^*) \dots$ ). Otherwise, if  $Z$  is a Banach space (possibly a Hilbert space) and  $Z'$  its dual, the notation  $\langle x, y \rangle_{Z \times Z'}$  will indicate the duality pairing. Eventually  $\langle x, y \rangle \equiv \langle x, y \rangle_{L^2(0, \bar{s})}$ .

**2.3. The control problem and the HJB equation.** In this subsection we describe the optimal control problem, state the hypotheses, define the HJB equation of the system, and give a suitable definition of solution of the HJB equation.

We consider the optimal control problem governed by the state equation

$$(11) \quad \begin{cases} \frac{d}{ds} i^* x(s) = Ax(s) - \mu x(s) + \alpha(s) + \beta\delta_0 a(s), \\ x(0) = x \end{cases}$$

that has a unique solution in the sense described in section 2.1. Given two compact subsets  $\Gamma$  and  $\Lambda$  of  $\mathbb{R}$ , we consider the set of admissible boundary controls given by

$$\mathcal{A} \stackrel{\text{def}}{=} \{a: [0, +\infty) \rightarrow \Gamma \subseteq \mathbb{R} : a(\cdot) \text{ is measurable}\}.$$

<sup>3</sup>See [36, Proposition 2, page 273] for a proof of the equality  $(A^* - \lambda I)^{-1} = ((A - \lambda I)^{-1})^*$ .

Moreover we call

$$\Sigma \stackrel{def}{=} \{\gamma: [0, \bar{s}] \rightarrow \Lambda \subseteq \mathbb{R} : \gamma(\cdot) \text{ is measurable}\}.$$

In view of the compactness of  $\Lambda$  we have  $\Sigma \subseteq L^2(0, \bar{s})$ . We define the set of admissible distributed controls as

$$\mathcal{E} \stackrel{def}{=} \{\alpha: [0, +\infty) \rightarrow \Sigma \subseteq L^2(0, \bar{s}) : \alpha(\cdot) \text{ is measurable}\}.$$

In view of the compactness of  $\Gamma$  and  $\Lambda$ , we have  $\mathcal{A} \subseteq L^2_{loc}([0, +\infty); \mathbb{R})$  and  $\mathcal{E} \subseteq L^2_{loc}([0, +\infty) \times [0, \bar{s}]; \mathbb{R})$ . We call  $\|\Gamma\| \stackrel{def}{=} \sup_{a \in \Gamma}(|a|)$ ,  $\|\Lambda\| \stackrel{def}{=} \sup_{b \in \Lambda}(|b|)$ , and  $\|\Sigma\| \stackrel{def}{=} \sup_{\alpha \in \Sigma}(\|\alpha\|_{\mathcal{H}})$  (they are bounded thanks to the boundedness of  $\Gamma$  and  $\Lambda$ ).

We call *admissible control* a couple  $(\alpha(\cdot), a(\cdot)) \in \mathcal{E} \times \mathcal{A}$ . The cost functional will be of the form

$$J(x, \alpha(\cdot), a(\cdot)) = \int_0^\infty e^{-\rho s} L(x(s), \alpha(s), a(s)) ds,$$

where  $L$  is uniformly continuous and satisfies the following conditions: There exists a  $C_L \geq 0$  with

$$((L1)) \quad |L(x, \alpha, a) - L(y, \alpha, a)| \leq C_L \langle B(x - y), (x - y) \rangle_{\mathcal{H} \times \mathcal{H}} \forall (\alpha, a) \in \Sigma \times \Gamma,$$

$$((L2)) \quad |L| \leq C_L < +\infty.$$

We define formally the HJB equation of the system as

$$(HJB) \quad \rho u(x) - \langle \nabla u(x), Ax \rangle - \langle \nabla u(x), -\mu x \rangle - H(x, \nabla u(x)) = 0,$$

where  $H$  is the Hamiltonian of the system and is defined as

$$\begin{cases} H: \mathcal{H} \times D(A^*) \rightarrow \mathbb{R}, \\ H(x, p) \stackrel{def}{=} \inf_{(\alpha, a) \in \Sigma \times \Gamma} (\langle \beta \delta_0(p), a \rangle_{\mathbb{R}} + \langle p, \alpha \rangle_{\mathcal{H}} + L(x, \alpha, a)). \end{cases}$$

Before introducing a suitable definition of the (viscosity) solution of the HJB equation, we give some preliminary definitions.

DEFINITION 2.10. A function  $v \in C(\mathcal{H})$  is *Lipschitz with respect to the B-norm*, or *B-Lipschitz*, if there exists a constant  $C$  such that

$$|v(x) - v(y)| \leq C|(x - y)|_B \stackrel{def}{=} C|B^{1/2}(x - y)|_{\mathcal{H}}$$

for every choice of  $x$  and  $y$  in  $\mathcal{H}$ . In the same way we can give the definition of a *locally B-Lipschitz function*.

DEFINITION 2.11. A function  $v \in C(\mathcal{H})$  is said to be *B-continuous* at a point  $x \in \mathcal{H}$  if for every  $x_n \in \mathcal{H}$ , with  $x_n \rightarrow x$  and  $|B(x_n - x)| \rightarrow 0$ , it holds that  $v(x_n) \rightarrow v(x)$ . In the same way we can define the *B-upper/lower semicontinuity*.

DEFINITION 2.12. We say that a function  $\phi$  such that  $\phi \in C^1(\mathcal{H})$  and  $\phi$  is *B-lower semicontinuous* is a *test function of type 1*, and we write  $\phi \in \text{test1}$ , if  $\nabla \phi(x) \in D(A^*)$  for all  $x \in \mathcal{H}$  and  $A^* \nabla \phi: \mathcal{H} \rightarrow \mathcal{H}$  is continuous.

DEFINITION 2.13. We say that  $g \in C^2(\mathcal{H})$  is a *test function of type 2*, and we write  $g \in \text{test2}$ , if  $g(x) = g_0(|x|)$  for some nondecreasing function  $g_0: \mathbb{R}^+ \rightarrow \mathbb{R}$ .

DEFINITION 2.14. A function  $u \in C(\mathcal{H})$  bounded and Lipschitz with respect to the  $B$ -norm, is a subsolution of the HJB equation (or simply a “subsolution”) if for every  $\phi \in \text{test1}$ ,  $g \in \text{test2}$ , and local maximum point  $x$  of  $u - (\phi + g)$  we have

$$(12) \quad \begin{aligned} & \rho u(x) - \langle A^* \nabla \phi(x), x \rangle - \langle \nabla \phi(x) + \nabla g(x), -\mu x \rangle \\ & - \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( \langle \beta \delta_0(\nabla \phi(x), a) \rangle_{\mathbb{R}} + \langle \nabla \phi(x) + \nabla g(x), \alpha \rangle_{\mathcal{H}} + L(x, \alpha, a) \right) \\ & \leq \frac{g'_0(|x|)}{|x|} \beta \frac{\|\Gamma\|^2}{2}. \end{aligned}$$

DEFINITION 2.15. A function  $v \in C(\mathcal{H})$  bounded and Lipschitz with respect to the  $B$ -norm is a supersolution of the HJB equation (or simply a “supersolution”) if for every  $\phi \in \text{test1}$ ,  $g \in \text{test2}$ , and local minimum point  $x$  of  $v + (\phi + g)$  we have

$$(13) \quad \begin{aligned} & \rho v(x) + \langle A^* \nabla \phi(x), x \rangle + \langle \nabla \phi(x) + \nabla g(x), -\mu x \rangle \\ & - \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( -\langle \beta \delta_0(\nabla \phi(x), a) \rangle_{\mathbb{R}} - \langle \nabla \phi(x) + \nabla g(x), \alpha \rangle_{\mathcal{H}} + L(x, \alpha, a) \right) \\ & \geq -\frac{g'_0(|x|)}{|x|} \beta \frac{\|\Gamma\|^2}{2}. \end{aligned}$$

DEFINITION 2.16. A function  $v \in C(\mathcal{H})$  bounded and Lipschitz with respect to the  $B$ -norm is a solution of the HJB equation if it is at the same time a supersolution and a subsolution.

Remark 2.17. In the definition of viscosity solution we have used two kinds of test functions: those in *test1* and those in *test2*, which, as usual in the literature, play a different role. In view of their properties the functions of the first set (*test1*) represent the “good part.” More difficult is to deal with the functions of the set *test2*, which have the role of localizing the problem. A difficulty of our case is the following: The trajectory is not Lipschitz with respect to the  $\mathcal{H}$ -norm, and so, given a function  $g \in \text{test2}$ , the term

$$(14) \quad \frac{g(x(s)) - g(x)}{s}$$

(where  $x(s)$  is a trajectory starting from  $x$ ) cannot be treated with standard arguments. The idea then is to consider only a  $B$ -Lipschitz solution so that the maxima/minima considered in Definitions 2.14 and 2.15 are in  $D(A^*)$ . If the starting point  $x$  is in  $D(A^*)$ , there are some advantages in the estimate of (14), but some problems remain: In such a case we will prove in Proposition 4.5 that (if  $\alpha(\cdot)$  is continuous in 0)

$$\left| \frac{g(x(s)) - g(x)}{s} - \langle \nabla g(x), -\mu x + \alpha(0) \rangle \right| \leq \frac{g'_0(|x|)}{|x|} \beta \frac{\|\Gamma\|^2}{2} + O(s),$$

where the rest  $O(s) \xrightarrow{s \rightarrow 0} 0$  and does not depend on the control. So the most challenging case is the one described in the definition.

**3. The value function and its properties.** The value function is, as usual, the candidate unique solution of the HJB equation. In this section we define the value function  $V(\cdot)$  of the problem, and then we verify that it has the regularity properties

required to be a solution. Namely, we will check that  $V(\cdot)$  is  $B$ -Lipschitz (Proposition 3.4). To obtain such a result we prove an approximation result (Proposition 3.1) and then a suitable estimate for the solution of the state equation (Proposition 3.3).

The value function of our problem is defined as

$$V(x) \stackrel{\text{def}}{=} \inf_{(\alpha(\cdot), a(\cdot)) \in \mathcal{E} \times \mathcal{A}} J(x, \alpha(\cdot), a(\cdot)).$$

We consider the functions

$$\begin{cases} \eta_n: [0, \bar{s}] \rightarrow \mathbb{R}, \\ \eta_n(r) \stackrel{\text{def}}{=} [2n - 2n^2 r]^+ \end{cases}$$

(where  $[\cdot]^+$  is the positive part). We then define

$$\begin{cases} \mathcal{C}_n^*: \mathbb{R} \rightarrow \mathcal{H}, \\ \mathcal{C}_n^*: \gamma \mapsto \gamma \eta_n. \end{cases}$$

Such functions are linear and continuous, and their adjoints are

$$(15) \quad \begin{cases} \mathcal{C}_n: \mathcal{H} \rightarrow \mathbb{R}, \\ \mathcal{C}_n: x \mapsto \langle x, \eta_n \rangle. \end{cases}$$

The functions  $\mathcal{C}_n$  approximate the delta measure in some sense. The approximating state equations we consider are

$$(16) \quad \begin{cases} \frac{d}{ds} x_n(s) = Ax_n(s) - \mu x_n(s) + \alpha(s) + \beta \mathcal{C}_n^* a(s), \\ x_n(0) = x. \end{cases}$$

In the following proofs we will use (8) together with the mild form of the approximating state equations (that can be found in [30, page 105, equation (2.3)]):

$$(17) \quad x_n(s) = e^{-\mu s} e^{sA} x + \int_0^s e^{-(s-\tau)\mu} e^{(s-\tau)A} \alpha(\tau) d\tau + \int_0^s e^{-(s-\tau)\mu} e^{(s-\tau)A} \beta \mathcal{C}_n^* a(\tau) d\tau.$$

PROPOSITION 3.1. For  $T > 0$  and  $(\alpha(\cdot), a(\cdot)) \in \mathcal{E} \times \mathcal{A}$

$$\lim_{n \rightarrow \infty} \sup_{s \in [0, T]} |x_n(s) - x(s)|_{\mathcal{H}} = 0.$$

*Proof.* By using (8) and (17) we find

$$(18) \quad |x(s) - x_n(s)| = \left| -A \int_0^s e^{-(s-\tau)\mu} e^{(s-\tau)A} (a(\tau) \nu) d\tau - \int_0^s e^{-(s-\tau)\mu} e^{(s-\tau)A} \beta \mathcal{C}_n^* (a(\tau)) d\tau \right|.$$

To estimate such an expression we will use the explicit expression of the two terms (as a two-variable function). We simplify the notation by using the extension  $\tilde{a}(\cdot)$  of  $a(\cdot)$  on all  $\mathbb{R}$  given by

$$\tilde{a}(s) = \begin{cases} 0 & \text{if } s < 0, \\ a(s) & \text{if } s \geq 0. \end{cases}$$

So

$$\begin{aligned}
 y(s, r) &\stackrel{def}{=} \left( -A \int_0^s e^{-(s-\tau)\mu} e^{(s-\tau)A} (a(\tau)\nu) d\tau \right) [r] = e^{-\frac{\mu}{\beta}r} \tilde{a}(s - r/\beta), \\
 (19) \quad y_n(s, r) &\stackrel{def}{=} \left( \int_0^s e^{-(s-\tau)\mu} e^{(s-\tau)A} \beta \mathcal{C}_n^*(a(\tau)) d\tau \right) [r] \\
 &= \int_0^{r \wedge (1/n)} e^{-\frac{\mu}{\beta}(r-\theta)} [2n - 2n^2\theta]^+ \tilde{a} \left( \frac{\theta - r}{\beta} + s \right) d\theta.
 \end{aligned}$$

Now for all  $s \in [0, T]$  we have

$$\begin{aligned}
 (20) \quad &|y(s, \cdot) - y_n(s, \cdot)|_{\mathcal{H}}^2 \\
 &\leq \left( \int_{1/n}^{\bar{s}} \left| e^{-\frac{\mu}{\beta}r} \tilde{a}(s - r/\beta) - \int_0^{1/n} e^{-\frac{\mu}{\beta}(r-\theta)} [2n - 2n^2\theta]^+ \tilde{a} \left( \frac{\theta - r}{\beta} + s \right) d\theta \right|^2 dr \right) \\
 &\quad + \left( \int_0^{1/n} \left| e^{-\frac{\mu}{\beta}r} \tilde{a}(s - r/\beta) - \int_0^r e^{-\frac{\mu}{\beta}(r-\theta)} [2n - 2n^2\theta]^+ \tilde{a} \left( \frac{\theta - r}{\beta} + s \right) d\theta \right|^2 dr \right) \\
 &\quad (\text{for } \bar{s} \leq T)
 \end{aligned}$$

$$\begin{aligned}
 (21) \quad &\leq \left( e^{|\mu|s} \int_0^T \left| e^{-\mu(\frac{r}{\beta}-s)} \tilde{a}(s - r/\beta) \right. \right. \\
 &\quad \left. \left. - \int_0^{1/n} e^{-\mu(\frac{r-\theta}{\beta}-s)} [2n - 2n^2\theta]^+ \tilde{a} \left( s + \frac{\theta - r}{\beta} \right) d\theta \right|^2 dr \right) + \left( \frac{1}{n} e^{|\mu|/\beta T} 2\|\Gamma\| \right).
 \end{aligned}$$

Such an estimate does not depend on  $s$ ; the integral term goes to zero because it is the convolution of a function in  $L^2(0, T)$  with an approximate unit, and the second goes to zero for  $n \rightarrow \infty$ .  $\square$

**PROPOSITION 3.2.** *Let  $\phi \in C^1(\mathcal{H})$  be such that  $\nabla\phi: \mathcal{H} \rightarrow D(A^*)$  is continuous. Then, for an admissible control  $(\alpha(\cdot), a(\cdot))$ , if we call  $x(\cdot)$  the trajectory starting from  $x$  and subject to the control  $(\alpha(\cdot), a(\cdot))$ , we have that, for every  $s > 0$ ,*

$$\begin{aligned}
 (22) \quad \phi(x(s)) &= \phi(x) + \int_0^s [\langle A^* \nabla\phi(x(\tau)), x(\tau) \rangle + \langle \beta\delta_0(\nabla\phi(x(\tau))), a(\tau) \rangle_{\mathbb{R}} \\
 &\quad + \langle \nabla\phi(x(\tau)), \alpha(\tau) \rangle + \langle \nabla\phi(x(\tau)), -\mu x(\tau) \rangle] d\tau.
 \end{aligned}$$

*Proof.* In the approximating state equation (16) the unbounded term  $\beta\delta_0$  does not appear ( $\beta\mathcal{C}_n^*$  are continuous), and then (see [29, Proposition 5.5, page 67]) for every  $\phi(\cdot) \in C^1(\mathcal{H})$  such that  $A^*\nabla\phi(\cdot) \in C(\mathcal{H})$  we have

$$\begin{aligned}
 (23) \quad \phi(x_n(s)) &= \phi(x) + \int_0^s [\langle A^* \nabla\phi(x_n(\tau)), x_n(\tau) \rangle + \langle \nabla\phi(x_n(\tau)), \beta\mathcal{C}_n^* a(\tau) \rangle \\
 &\quad + \langle \nabla\phi(x_n(\tau)), \alpha(\tau) \rangle + \langle \nabla\phi(x_n(\tau)), -\mu x_n(\tau) \rangle] d\tau \\
 &= \phi(x) + \int_0^s [\langle A^* \nabla\phi(x_n(\tau)), x_n(\tau) \rangle + \langle \beta\mathcal{C}_n \nabla\phi(x_n(\tau)), a(\tau) \rangle \\
 &\quad + \langle \nabla\phi(x_n(\tau)), \alpha(\tau) \rangle + \langle \nabla\phi(x_n(\tau)), -\mu x_n(\tau) \rangle] d\tau,
 \end{aligned}$$

where we passed to the adjoint in view of the continuity of the operator  $\mathcal{C}_n^*$  (see (15) for the explicit form of  $\mathcal{C}_n$ ).

Now we prove that every integral term of (23) converges to the corresponding term of (22). This fact, together with the pointwise convergence of  $\phi(x_n(s))$  to  $\phi(x(s))$  (due to Proposition 3.1), will imply the claim.

First we note that, in view of Proposition 3.1 and of the continuity of  $x$ ,  $x_n(\tau)$  is bounded uniformly in  $n$  and  $\tau \in [0, s]$ , and, in view of the continuity of  $\nabla\phi$ ,  $\nabla\phi(x_n(\tau))$  is bounded uniformly in  $n$  and  $\tau \in [0, s]$ . So we can apply the Lebesgue theorem (the pointwise convergence is given by Proposition 3.1 and  $|\alpha(\tau)| \leq \|\Sigma\|$ ), and we derive that

$$(24) \quad \int_0^s [\langle \nabla\phi(x_n(\tau)), \alpha(\tau) \rangle + \langle \nabla\phi(x_n(\tau)), -\mu x_n(\tau) \rangle] d\tau \\ \xrightarrow{n \rightarrow \infty} \int_0^s [\langle \nabla\phi(x(\tau)), \alpha(\tau) \rangle + \langle \nabla\phi(x(\tau)), -\mu x(\tau) \rangle] d\tau.$$

Next we observe that, in view of the continuity of  $A^*\nabla\phi$  and of Proposition 3.1, the term  $A^*\nabla\phi(x_n(\tau))$  is bounded uniformly in  $n$  and  $\tau \in [0, s]$ , so the same is true for

$$|A^*\nabla\phi(x_n(\tau)) - A^*\nabla\phi(x(\tau))|.$$

Therefore we can use the Lebesgue theorem (the pointwise convergence is given by Proposition 3.1) to conclude that

$$\int_0^s \langle A^*\nabla\phi(x_n(\tau)), x_n(\tau) \rangle d\tau \rightarrow \int_0^s \langle A^*\nabla\phi(x(\tau)), x(\tau) \rangle d\tau.$$

We now have to prove that

$$(25) \quad \int_0^s \langle \beta \mathcal{C}_n \nabla\phi(x_n(\tau)), a(\tau) \rangle d\tau \rightarrow \int_0^s \langle \beta \delta_0(\nabla\phi(x(\tau))), a(\tau) \rangle_{\mathbb{R}} d\tau.$$

We first note that  $\mathcal{C}_n \xrightarrow{n \rightarrow \infty} \delta_0$  in  $H^{-1}(0, \bar{s})$  and then in  $D(A^*)'$ . Indeed given  $z \in H^1(0, \bar{s})$  we have

$$(26) \quad |(\mathcal{C}_n - \delta_0)z| = \left| \int_0^{\bar{s}} z[\tau] \eta_n[\tau] d\tau - z[0] \right| \\ = \left| \int_0^{1/n} \left( z[0] + \int_0^\tau \partial_\omega z[r] dr \right) \eta_n[\tau] d\tau - z[0] \right|$$

( $\partial_\omega z$  is the weak derivative of  $z$ ), and integrating by parts

$$(27) \quad = \left| \left( z[0] + \int_0^{1/n} \partial_\omega z[r] dr \right) \left( \int_0^{1/n} \eta_n[r] dr \right) \right. \\ \left. - \int_0^{1/n} \partial_\omega z[\tau] \int_0^\tau \eta_n[r] dr d\tau - z[0] \right|$$

and then, writing  $\eta_n$  in explicit form (note that  $\int_0^{1/n} \eta_n[r] dr = 1$ ),

$$\leq \left| \int_0^{\bar{s}} \chi_{[0, 1/n]}[\tau] |\partial_\omega z[\tau]| d\tau \right| \leq \frac{1}{\sqrt{n}} \|z\|_{H^1(0, \bar{s})}.$$

In summary, by Proposition 3.1,  $x_n(\cdot) \xrightarrow{n \rightarrow \infty} x(\cdot)$  in  $C([0, T]; \mathcal{H})$ , then (by hypothesis on  $\phi$ )  $\nabla \phi(x_n(\cdot)) \xrightarrow{n \rightarrow \infty} \nabla \phi(x(\cdot))$  in  $C([0, T]; D(A^*))$ , and then, by the last estimate,  $\beta \mathcal{C}_n(\nabla \phi(x_n(\cdot))) \xrightarrow{n \rightarrow \infty} \beta \delta_0(\nabla \phi(x(\cdot)))$  in  $C([0, T]; \mathbb{R})$ . Then (25) follows by Cauchy–Schwartz inequality.  $\square$

PROPOSITION 3.3. *Given  $T > 0$  and a control  $(\alpha(\cdot), a(\cdot)) \in \mathcal{E} \times \mathcal{A}$ , there exists  $c_T$  such that for every  $x, y \in \mathcal{H}$*

$$\sup_{s \in [0, T]} |x_x(s) - x_y(s)|_B^2 \leq c_T |x - y|_B^2,$$

where  $x_y(\cdot)$  is the solution of

$$\begin{cases} \frac{d}{ds} i^* x(s) = Ax(s) + \alpha(s) - \mu x(s) + \beta \delta_0 a(s), \\ x(0) = y \end{cases}$$

and  $x_x(\cdot)$  the solution with initial data  $x$ .

*Proof.* We use Proposition 3.2 with  $\phi(x) = \langle Bx, x \rangle$  so that  $\nabla \phi(x) = 2Bx$ . Moreover  $x_x(\cdot) - x_y(\cdot)$  satisfies the equation

$$\begin{cases} \frac{d}{ds} i^* (x_x(s) - x_y(s)) = A(x_x(s) - x_y(s)) - \mu(x_x(s) - x_y(s)), \\ (x_x - x_y)(0) = x - y \end{cases}$$

(the one of Proposition 3.2 with control identically 0), and then by (10)

$$\begin{aligned} (28) \quad |x_x(s) - x_y(s)|_B^2 &= |x - y|_B^2 + 2 \int_0^s \langle A^* B(x_x(r) - x_y(r)), (x_x(r) - x_y(r)) \rangle \\ &\quad - \mu \langle B(x_x(s) - x_y(s)), x_x(s) - x_y(s) \rangle dr \\ &\leq |x - y|_B^2 + 2(1 + |\mu|) \int_0^s \langle B(x_x(r) - x_y(r)), (x_x(r) - x_y(r)) \rangle dr. \end{aligned}$$

Finally we can use Gronwall's lemma to obtain the claim.  $\square$

PROPOSITION 3.4. *The value function  $V$  is Lipschitz with respect to the  $B$ -norm.*

*Proof.* Assume  $V(y) > V(x)$ ,  $\varepsilon > 0$ , and  $(\alpha(\cdot), a(\cdot)) \in \mathcal{E} \times \mathcal{A}$  an  $\varepsilon$ -optimal control for  $x$ . We have

$$|V(y) - V(x)| - \varepsilon \leq \int_0^\infty e^{-\rho t} |L(x_y(s), \alpha(s), a(s)) - L(x_x(s), \alpha(s), a(s))| ds.$$

If we look, the explicit form of  $x_x(\cdot)$  and  $x_y(\cdot)$  as two-variable functions depends on the initial data only for  $s \in [0, \frac{\bar{s}}{\beta}]$  and afterwards only on the control. Indeed, for  $s > \frac{\bar{s}}{\beta}$ ,  $x_x(s) = x_y(s)$ , and the previous integral is equal to

$$\int_0^{\bar{s}/\beta} e^{-\rho t} |L(x_y(s), \alpha(s), a(s)) - L(x_x(s), \alpha(s), a(s))| ds$$

by (L1) and Proposition 3.3

$$\leq \int_0^{\bar{s}/\beta} e^{-\rho t} C_L |x_y(s) - x_x(s)|_B ds \leq \frac{\bar{s}}{\beta} c_{\bar{s}} C_L |x - y|_B.$$

By letting  $\varepsilon \rightarrow 0$  we have the thesis.  $\square$

**4. Existence and uniqueness of a solution.** In this section we will prove that the value function is a viscosity solution of the HJB equation (Theorem 4.7) and that the HJB equation admits at most one solution (Theorem 4.8).

We remind the reader that we use  $\mathcal{H}_B$  to denote the completion of  $\mathcal{H}$  in the  $B$ -norm. This notation will be used in the next propositions.

**PROPOSITION 4.1.** *Let  $u \in C(\mathcal{H})$  be a locally  $B$ -Lipschitz function. Let  $\psi \in C^1(\mathcal{H})$ , and let  $x$  be a local maximum (or a local minimum) of  $u - \psi$ . Then  $\nabla\psi(x) \in \mathcal{R}(B^{1/2}) \subseteq D(A^*)$ .*

*Proof.* We give the proof only when  $x$  is a local maximum, as the case of the minima is similar.

We take  $\omega \in \mathcal{H}$ , with  $|\omega| = 1$  and  $h \in (0, 1)$ . Then for every  $h$  small enough

$$\frac{(u(x - h\omega) - \psi(x - h\omega))}{h} \leq \frac{u(x) - \psi(x)}{h},$$

so that

$$\frac{\psi(x) - \psi(x - h\omega)}{h} \leq C|\omega|_B,$$

and by passing to the limit we have  $\langle \nabla\psi(x), \omega \rangle \leq C|\omega|_B$ . Likewise

$$\frac{(u(x + h\omega) - \psi(x + h\omega))}{h} \leq \frac{u(x) - \psi(x)}{h},$$

so that

$$\frac{\psi(x) - \psi(x + h\omega)}{h} \leq C|\omega|_B,$$

and by passing to the limit we have  $-\langle \nabla\psi(x), \omega \rangle \leq C|\omega|_B$ .

The two inequalities together then give

$$|\langle \nabla\psi(x), \omega \rangle| \leq C|\omega|_B$$

for all  $\omega \in \mathcal{H}$ . Then we can consider the linear extension to  $\mathcal{H}_B$  of the continuous linear functional  $\omega \mapsto \langle \nabla\psi(x), \omega \rangle$  that we denote with  $\Phi_x$ . By the Riesz representation theorem, we can find  $z_x \in \mathcal{H}_B$  such that

$$\Phi_x(\omega) = \langle z_x, \omega \rangle_{\mathcal{H}_B} \quad \forall \omega \in \mathcal{H}_B$$

and note afterwards that

$$\begin{aligned} (29) \quad \langle z_x, \omega \rangle_{\mathcal{H}_B} &= \left\langle B^{1/2}(z_x), B^{1/2}(\omega) \right\rangle_{\mathcal{H}} \\ &= \left\langle B^{1/2}(B^{1/2}(z_x)), \omega \right\rangle_{(\mathcal{H}_B)' \times (\mathcal{H}_B)} = \left\langle B^{1/2}(m_x), \omega \right\rangle_{(\mathcal{H}_B)' \times (\mathcal{H}_B)}, \end{aligned}$$

where  $m_x \stackrel{\text{def}}{=} (B^{1/2}(z_x)) \in \mathcal{H}$ . Now for  $\omega \in \mathcal{H}$

$$\left\langle B^{1/2}(m_x), \omega \right\rangle_{(\mathcal{H}_B)' \times (\mathcal{H}_B)} = \left\langle B^{1/2}(m_x), \omega \right\rangle_{\mathcal{H}},$$

and therefore  $\nabla\psi(x) = B^{1/2}(m_x) \in \mathcal{R}(B^{1/2}) \subseteq D(A^*)$ , where the last inclusion follows from Remark 2.6.  $\square$



**4.1. Existence.** We start by proving a lemma and two propositions that will be used to prove the existence theorem (Theorem 4.7). We will use the notation introduced in Remark 2.1 on “ $x(s)$ ” and “ $x[r]$ .” Moreover we will continue to use the symbol  $\delta_0$  in the text so that  $x[0] = \delta_0 x$  if  $x \in D(A^*)$ .

LEMMA 4.2. *Let  $x$  be a function of  $H^1(0, \bar{s})$ , and then*

$$(30) \quad (i) \lim_{s \rightarrow 0^+} \left( \int_s^{\bar{s}} \frac{(x[r] - x[r-s])^2}{s} dr \right) = 0,$$

$$(31) \quad (ii) \lim_{s \rightarrow 0^+} \left( \int_s^{\bar{s}-s} \frac{(x[r+s] - x[r])}{s} x[r] dr \right) = \frac{x^2[\bar{s}] - x^2[0]}{2}.$$

*Proof.* Part (i): We have

$$\int_s^{\bar{s}} \frac{(x[r] - x[r-s])^2}{s} dr = \int_0^{\bar{s}} \psi_s[r] dr,$$

where  $\psi_s: [0, \bar{s}] \rightarrow \mathbb{R}$  is defined in the following way:

$$\psi_s[r] = \begin{cases} 0 & \text{if } r \in [0, s), \\ \frac{(x[r] - x[r-s])^2}{s} & \text{if } r \in [s, \bar{s}]. \end{cases}$$

We prove the thesis by means of the Lebesgue theorem. First we show that  $\psi_s$  converges a.e. to zero. For  $r > s$ :

$$\psi_s[r] \leq \frac{\left| \int_{r-s}^r \partial_\omega x[\tau] d\tau \right|}{s} |x[r] - x[r-s]|,$$

where  $\partial_\omega x$  is the weak derivative of  $x$ . Now almost every  $r$  is a Lebesgue point, which implies that

$$\frac{\left| \int_{r-s}^r \partial_\omega x(\tau) d\tau \right|}{s} \xrightarrow{s \rightarrow 0^+} |\partial_\omega x[r]| \quad \text{a.e. in } r \in (0, \bar{s}),$$

while the term  $|x[r] - x[r-s]|$  goes uniformly to 0.

In order to dominate the convergence we note that by Morrey's theorem ([18, Theorem 4, page 266]) every  $x \in H^1(0, \bar{s})$  is 1/2-Holder continuous, and then there exists a positive constant  $C$  such that for every  $s \in (0, \bar{s})$  and every  $r \in [s, \bar{s}]$  we have

$$\frac{|x[r] - x[r-s]|}{\sqrt{s}} \leq C.$$

Then

$$|\psi_s[r]| \leq \frac{|x[r] - x[r-s]|^2}{s} \leq C^2,$$

and the proof of part (i) is complete.

Now we prove part (ii):

$$\begin{aligned}
 (32) \quad I(s) &\stackrel{\text{def}}{=} \int_s^{\bar{s}-s} \frac{(x[r+s] - x[r])}{s} x[r] dr \\
 &= \int_s^{\bar{s}-s} \frac{(x[r+s]x[r])}{s} dr - \int_0^{\bar{s}-2s} \frac{(x[r+s]x[r+s])}{s} dr \\
 &= - \int_s^{\bar{s}-2s} \frac{(x[r+s] - x[r])}{s} x[r+s] dr + \int_{\bar{s}-2s}^{\bar{s}-s} \frac{(x[r+s]x[r])}{s} dr \\
 &\quad + \int_0^s - \frac{(x[r+s])^2}{s} dr \stackrel{\text{def}}{=} -I_1(s) + I_2(s) + I_3(s).
 \end{aligned}$$

By the continuity of  $x$  we see that

$$I_2(s) \xrightarrow{s \rightarrow 0^+} x^2[\bar{s}]$$

and

$$I_3(s) \xrightarrow{s \rightarrow 0^+} -x^2[0].$$

Moreover, by using arguments similar to those in (i), we find that

$$\begin{aligned}
 (33) \quad \lim_{s \rightarrow 0^+} (I(s) - I_1(s)) &= \lim_{s \rightarrow 0^+} \int_s^{\bar{s}-2s} - \frac{(x[r+s] - x[r])^2}{s} dr \\
 &\quad + \lim_{s \rightarrow 0^+} \int_{\bar{s}-2s}^{\bar{s}-s} \frac{(x[r+s] - x[r])}{s} x[r] dr = 0,
 \end{aligned}$$

so, since  $I(s) + I_1(s) = I_2(s) + I_3(s)$ , the limit  $\lim_{s \rightarrow 0^+} I(s)$  exists if and only if there exists the limit  $\lim_{s \rightarrow 0^+} \frac{I_1(s) + I(s)}{2}$ , and in such a case they have the same value. As

$$\frac{I_1(s) + I(s)}{2} = \frac{I_2(s) + I_3(s)}{2} \xrightarrow{s \rightarrow 0^+} \frac{x^2[\bar{s}] - x^2[0]}{2},$$

then

$$\lim_{s \rightarrow 0^+} \left( \int_s^{\bar{s}-s} \frac{(x[r+s] - x[r])}{s} x[r] dr \right) = \frac{x^2[\bar{s}] - x^2[0]}{2}. \quad \square$$

*Notation 4.3.* We will call from now on  $O(s)$  a generic function  $O(\cdot): [0, +\infty) \rightarrow [0, +\infty)$  such that  $O(s) \xrightarrow{s \rightarrow 0^+} 0$  and  $O(0) = 0$ . In what follows such notation will be used to express in particular the estimates that do not depend on the control.

**LEMMA 4.4.** *Given  $x \in D(A^*)$ , there exists an  $O(s)$ , independent of the control, such that, for every control  $(\alpha(\cdot), a(\cdot)) \in \mathcal{E} \times \mathcal{A}$ , we have that*

$$|x(s) - x| \leq O(s)$$

(where we called  $x(s)$  the trajectory that starts from  $x$  and subject to the control  $(\alpha(\cdot), a(\cdot))$ ).

*Proof.* We consider  $s \in (0, 1]$ . By means of (6) we have

$$\begin{aligned}
 (34) \quad \|x(s) - x\|_{\mathcal{H}}^2 &= \int_{\beta s}^{\bar{s}} \left| e^{-\mu s} x[r - \beta s] + \int_0^s e^{-\mu \tau} \alpha(s - \tau, r - \beta \tau) d\tau - x[r] \right|^2 dr \\
 &+ \int_0^{\beta s} \left| e^{-\frac{\mu}{\beta} r} a(s - r/\beta) + \int_0^{r/\beta} e^{-\mu \tau} \alpha(s - \tau, r - \beta \tau) d\tau - x[r] \right|^2 dr \\
 &\leq 2 \int_{\beta s}^{\bar{s}} |e^{-\mu s} x[r - \beta s] - x[r]|^2 dr + 2 \int_{\beta s}^{\bar{s}} \left| \int_0^s e^{|\mu|} \|\Lambda\| d\tau \right|^2 dr \\
 &\quad + \int_0^{\beta s} \left| e^{|\mu|/\beta} \|\Gamma\| + \int_0^{r/\beta} e^{|\mu|} \|\Lambda\| d\tau + |x|_{L^\infty(0, \bar{s})} \right|^2 dr,
 \end{aligned}$$

where we have used that  $x \in D(A^*) \subseteq W^{1,2}(0, \bar{s}) \subseteq L^\infty(0, \bar{s})$

$$\begin{aligned}
 (35) \quad &\leq 2 \int_0^{\bar{s}} |e^{-\mu s} x[(r - \beta s) \wedge 0] - x[r]|^2 dr + 2s^2 \bar{s} \left( e^{|\mu|} \|\Lambda\| \right)^2 \\
 &\quad + s\beta \left( e^{|\mu|/\beta} \|\Gamma\| + |x|_{L^\infty} + se^{|\mu|} \|\Lambda\| \right)^2.
 \end{aligned}$$

Observe that in this estimate the control  $(\alpha(\cdot), a(\cdot))$  does not appear. The second and the third terms go to zero for  $s \rightarrow 0$ , while for the first we can use the Lebesgue theorem by observing that

$$|e^{-\mu s} x[(r - \beta s) \wedge 0] - x[r]| \leq e^{|\mu|} |x|_{L^\infty} + |x|_{L^\infty} \quad \forall (s, r) \in (0, 1] \times [0, \bar{s}]$$

and that  $|e^{-\mu s} x[(r - \beta s) \wedge 0] - x[r]| \xrightarrow{s \rightarrow 0} 0$  pointwise.  $\square$

**PROPOSITION 4.5.** *Given  $x \in D(A^*)$  and  $g \in \text{test2}$ , there exists an  $O(s)$ , independent of the control, such that, for every control  $(\alpha(\cdot), a(\cdot)) \in \mathcal{E} \times \mathcal{A}$ , with  $a(\cdot)$  continuous, we have that*

$$\left| \frac{g(x(s)) - g(x)}{s} - \frac{\int_0^s \langle \nabla g(x), \alpha(r) \rangle}{s} - \langle \nabla g(x), -\mu x \rangle \right| \leq \frac{g'_0(|x|)}{|x|} \beta \frac{\|\Gamma\|^2}{2} + O(s)$$

(where we called  $x(s)$  the trajectory that starts from  $x$  and subject to the control  $(\alpha(\cdot), a(\cdot))$ ).

*Proof.* First we observe

$$\begin{aligned}
 (36) \quad &\frac{g(x(s)) - g(x)}{s} - \langle \nabla g(x), -\mu x \rangle - \frac{\int_0^s \langle \nabla g(x), \alpha(r) \rangle}{s} \\
 &= \frac{g(x(s)) - g(y(s)) + g(y(s)) - g(x)}{s} - \langle \nabla g(x), -\mu x \rangle - \frac{\int_0^s \langle \nabla g(x), \alpha(r) \rangle}{s},
 \end{aligned}$$

where  $y(\cdot)$  is the mild solution of

$$(37) \quad \begin{cases} \dot{y}(s) = Ay(s) + \beta \delta_0 a(s), \\ y(0) = x \end{cases}$$

corresponding to (11) with  $\mu = 0$  and  $\alpha(\cdot) = 0$ . The difference  $x(s) - y(s)$  can be expressed in the mild form as

$$x(s) - y(s) = \int_0^s e^{(s-\tau)A}(\alpha(\tau) - \mu x(\tau))d\tau.$$

Now we come back to (36), and we have

$$(38) \quad \left| \frac{g(x(s)) - g(x)}{s} - \frac{\int_0^s \langle \nabla g(x), \alpha(r) \rangle dr}{s} - \langle \nabla g(x), -\mu x \rangle \right| \\ \leq \left| \frac{g(x(s)) - g(y(s))}{s} - \frac{\int_0^s \langle \nabla g(x), \alpha(r) \rangle dr}{s} - \langle \nabla g(x), -\mu x \rangle \right| + \left| \frac{g(y(s)) - g(x)}{s} \right|.$$

In order to estimate the first addendum we use the Taylor expansion as follows:

$$(39) \quad \frac{g(x(s)) - g(y(s))}{s} = \left\langle \nabla g(y(s)), \frac{x(s) - y(s)}{s} \right\rangle \\ + \left\langle \nabla g(\xi(s)) - \nabla g(y(s)), \frac{x(s) - y(s)}{s} \right\rangle,$$

where  $\xi(s)$  is a point between  $x(s)$  and  $y(s)$

$$(40) \quad = \left\langle \nabla g(y(s)), \frac{\int_0^s e^{(s-\tau)A}(\alpha(\tau) - \mu x(\tau))d\tau}{s} \right\rangle \\ + \left\langle \nabla g(\xi(s)) - \nabla g(y(s)), \frac{\int_0^s e^{(s-\tau)A}(\alpha(\tau) - \mu x(\tau))d\tau}{s} \right\rangle.$$

We know by Lemma 4.4 that  $x(s) \xrightarrow{s \rightarrow 0} x$  and  $y(s) \xrightarrow{s \rightarrow 0} x$  uniformly in the control  $(\alpha(\cdot), a(\cdot))$ , and then  $\nabla g(y(s)) \xrightarrow{s \rightarrow 0} \nabla g(x)$  uniformly in the control and  $|\nabla g(y(s)) - \nabla g(\xi(s))| \xrightarrow{s \rightarrow 0} 0$  uniformly in the control. Moreover, since in addition the control is bounded and  $x(s) \xrightarrow{s \rightarrow 0} x$  uniformly in the control, we infer that the term

$$\left| \frac{\int_0^s e^{(s-\tau)A}(\alpha(\tau) - \mu x(\tau))d\tau}{s} \right|_{\mathcal{H}}$$

is bounded uniformly in the control  $\alpha(\cdot)$  and in  $s$ . From this we conclude that the second term in (40) goes to zero uniformly in  $(\alpha(\cdot), a(\cdot))$  and that

$$(41) \quad \left| \frac{g(x(s)) - g(y(s))}{s} - \frac{\int_0^s \langle \nabla g(x), \alpha(r) \rangle dr}{s} - \langle \nabla g(x), -\mu x \rangle \right| \leq O(s)$$

for some function  $O(s)$  independent of the control.

Now we estimate the second term of (38).

We first note that

$$\nabla g(x) = g'_0(|x|) \frac{x}{|x|}$$

and

$$D^2g(x) = g_0''(|x|)\frac{x}{|x|} \otimes \frac{x}{|x|} + g_0'(|x|)\left(\frac{\mathbf{I}}{|x|} - \frac{x \otimes x}{|x|^3}\right).$$

We consider the Taylor's expansion of  $g$  at  $x$ :

$$\begin{aligned} (42) \quad \frac{g(y(s)) - g(x)}{s} &= \frac{\langle \nabla g(x), y(s) - x \rangle}{s} + \frac{1}{2} \frac{(y(s) - x)^T (D^2g(x))(y(s) - x)}{s} \\ &\quad + \frac{o(|y(s) - x|^2)}{s} \\ &= \frac{g_0'(|x|)}{|x|} \left( \left\langle x, \frac{y(s) - x}{s} \right\rangle + \frac{1}{2} \frac{\langle y(s) - x, y(s) - x \rangle}{s} \right) \\ &\quad + \frac{1}{2} \left( \frac{g_0''(|x|)}{|x|^2} - \frac{g_0'(|x|)}{|x|^3} \right) \frac{\langle x, y(s) - x \rangle^2}{s} + \frac{o(|y(s) - x|^2)}{s} \\ &\stackrel{def}{=} P1 + P2 + P3. \end{aligned}$$

First we prove that  $P2$  and  $P3$  go to zero uniformly in  $(\alpha(\cdot), a(\cdot))$ , and then we estimate  $P1$ . We proceed in three steps.

*Step 1.* There exists a constant  $C$  such that for every admissible control  $(\alpha(\cdot), a(\cdot)) \in \mathcal{E} \times \mathcal{A}$  with  $a(\cdot)$  continuous, and every  $s \in (0, 1]^4$

$$\left| \frac{\langle x, y(s) - x \rangle}{s} \right| \leq C.$$

We observe first that the explicit solution of  $y(s)[r]$  can be found by taking  $\mu = 0$  and  $\alpha = 0$  in (6). We have

$$y(s, r) = \begin{cases} x[r - \beta s], & r \in [\beta s, \bar{s}], \\ a(s - r/\beta), & r \in [0, \beta s), \end{cases}$$

so

$$\begin{aligned} (43) \quad \frac{\langle x, y(s) - x \rangle}{s} &= \int_{\beta s}^{\bar{s}} x[r] \frac{(x[r - \beta s] - x[r])}{s} dr + \frac{\int_0^{\beta s} x[r](a(s - r/\beta) - x[r]) dr}{s} \\ &= \int_{\beta s}^{\bar{s} - \beta s} x[r] \frac{(x[r + \beta s] - x[r])}{s} dr + \frac{\int_{\bar{s} - \beta s}^{\bar{s}} -x^2[r] dr}{s} + \frac{\int_0^{\beta s} x[r]x[r + \beta s] dr}{s} \\ &\quad + \frac{\int_0^{\beta s} x[r]a(s - r/\beta) dr}{s} - \frac{\int_0^{\beta s} x^2[r] dr}{s}. \end{aligned}$$

When  $s \rightarrow 0$ , the third and the fifth terms have opposite limits, while the second goes to zero as  $x$  is continuous and  $x(\bar{s}) = 0$ . The first term goes to  $-\frac{\beta}{2}x^2[0] = \langle A^*x, x \rangle$  in view of Lemma 4.2. The control appears only in the fourth term that we estimate as follows:

$$\left| \frac{\int_0^{\beta s} x[r]a(s - r/\beta) dr}{s} \right| \leq \frac{\int_0^{\beta s} |x[r]| \|\Gamma\| dr}{s} \leq \beta \max_{r \in [0, \bar{s}]} |x[r]| \|\Gamma\|.$$

<sup>4</sup>In the expression of  $y(\cdot)$  the distributed control  $\alpha(\cdot)$  does not appear, so we will speak from now on only of the boundary control  $a(\cdot)$

*Step 2.* There exists a constant  $C$  such that for every admissible control  $a(\cdot) \in \mathcal{A}$ , with  $a(\cdot)$  continuous, and every  $s \in (0, 1]$

$$\frac{|y(s) - x|^2}{s} \leq C.$$

Indeed

$$(44) \quad \left| \frac{\langle y(s) - x, y(s) - x \rangle}{s} \right| = \left| \int_{\beta s}^{\bar{s}} \frac{(x[r - \beta s] - x[r])^2}{s} dr \right| + \left| \frac{\int_0^{\beta s} (a(s - r/\beta) - x[r])^2 dr}{s} \right|,$$

since  $x \in D(A^*) \subseteq H^1(0, \bar{s})$  and Lemma 4.2 holds; then the first term goes to zero. Moreover the second term is less than or equal to

$$(45) \quad \frac{\int_0^{\beta s} \|\Gamma\|^2 dr}{s} + \frac{\int_0^{\beta s} 2|x[r]||\|\Gamma\| dr}{s} + \frac{\int_0^{\beta s} |x[r]|^2 dr}{s} \leq C.$$

This completes Step 2.

From Step 2 it follows that

$$\frac{o(|y(s) - x|^2)}{s} = \frac{o(|y(s) - x|^2)}{|y(s) - x|^2} \frac{|y(s) - x|^2}{s} \xrightarrow{s \rightarrow 0^+} 0$$

uniformly in  $a(\cdot)$ . Thus  $|P3| \xrightarrow{s \rightarrow 0} 0$  uniformly in  $a(\cdot)$ . Moreover

$$\frac{\langle x, y(s) - x \rangle^2}{s} \leq \frac{|\langle x, y(s) - x \rangle|}{s} |x| |y(s) - x|,$$

and then, from Step 1 and Lemma 4.4,  $|P2| \xrightarrow{s \rightarrow 0} 0$  uniformly in  $a(\cdot)$ .

*Step 3. (Conclusion.)* We now estimate  $P1$ . We can write a more explicit form of  $P1$  as in the proofs of Steps 1 and 2 ((43), (44), and (45)), and by using the same arguments we can see that there exists a function  $O(s)$  (depending only on  $x$  and independent of the control) such that for every continuous control  $a(\cdot)$

$$(46) \quad P1 = \frac{g'_0(|x|)}{|x|} \left( \langle A^* x, x \rangle + \frac{\int_0^{\beta s} x[s] a(s - r/\beta) dr}{s} + \frac{1}{2} \frac{\int_0^{\beta s} (a(s - r/\beta))^2 dr}{s} \right. \\ \left. + \frac{1}{2} \frac{\int_0^{\beta s} x^2[r] dr}{s} + \frac{1}{2} \frac{\int_0^{\beta s} -2x[r] a(s - r/\beta) dr}{s} \right) + O(s).$$

The fourth term above does not depend on the control and converges to  $\beta \frac{x[0]^2}{2}$  that is the opposite of the first term. The second and the fifth terms are opposite. Then we have that

$$P1 = O(s) + \frac{g'_0(|x|)}{|x|} \left( \frac{1}{2} \frac{\int_0^{\beta s} (a(s - r/\beta))^2 dr}{s} \right) \leq O(s) + \frac{1}{2} \frac{g'_0(|x|)}{|x|} \beta \|\Gamma\|^2.$$

Now, by using the estimates on  $P1$ ,  $P2$ , and  $P3$ , we see that

$$\left| \frac{g(y(s)) - g(x)}{s} \right| \leq O(s) + \frac{1}{2} \frac{g'_0(|x|)}{|x|} \beta \|\Gamma\|^2.$$

By using this fact and (41) in (38), we have proved the proposition.  $\square$

**PROPOSITION 4.6.** *If  $x \in D(A^*)$  and  $\phi \in \text{test1}$ , then there exists an  $O(s)$ , independent of the control, such that, for every control  $(\alpha(\cdot), a(\cdot)) \in \mathcal{E} \times \mathcal{A}$ , with  $a(\cdot)$  continuous, we have that*

$$(47) \quad \left| \frac{\phi(x(s)) - \phi(x)}{s} - \frac{\int_0^s \langle \nabla \phi(x), \alpha(r) \rangle dr}{s} - \langle \nabla \phi(x), -\mu x \rangle - \langle A^* \nabla \phi(x), x \rangle - \frac{\int_0^s \langle \beta \delta_0(\nabla \phi(x)), a(r) \rangle_{\mathbb{R}} dr}{s} \right| \leq O(s)$$

(where we called  $x(s)$  the trajectory that starts from  $x$  and subject to the control  $(\alpha(\cdot), a(\cdot))$ ).

*Proof.* We proceed as in the proof of Proposition 4.5 by observing that

$$\frac{\phi(x(s)) - \phi(x)}{s} = \frac{\phi(x(s)) - \phi(y(s))}{s} + \frac{\phi(y(s)) - \phi(x)}{s},$$

where  $y(\cdot)$  is the solution of (37). It is possible to prove, by using exactly the same arguments in the proof of Proposition 4.5, that

$$\left| \frac{\phi(x(s)) - \phi(y(s))}{s} - \langle \nabla \phi(x), -\mu x \rangle - \frac{\int_0^s \langle \nabla \phi(x), \alpha(r) \rangle dr}{s} \right| \leq O(s),$$

where  $O(s)$  does not depend on the control. What is left to show is

$$\left| \frac{\phi(y(s)) - \phi(x)}{s} - \langle A^* \nabla \phi(x), x \rangle - \frac{\int_0^s \beta \langle \delta_0 \nabla \phi(x), a(r) \rangle_{\mathbb{R}} dr}{s} \right| \leq O(s),$$

where  $O(s)$  does not depend on the control.

We write

$$(48) \quad \frac{\phi(y(s)) - \phi(x)}{s} = I_0 + I_1 \stackrel{\text{def}}{=} \left\langle \nabla \phi(x), \frac{y(s) - x}{s} \right\rangle + \left\langle \nabla \phi(\xi(s)) - \nabla \phi(x), \frac{y(s) - x}{s} \right\rangle,$$

where  $\xi(s)$  is a point between  $x$  and  $y(s)$ . In view of Lemma 4.4,  $|y(s) - x| \xrightarrow{s \rightarrow 0} 0$  uniformly in the control, so that  $|\xi(s) - x| \xrightarrow{s \rightarrow 0} 0$  uniformly in  $a(\cdot)$ . By hypothesis  $\nabla \phi: \mathcal{H} \rightarrow D(A^*)$  is continuous so that

$$(49) \quad |\nabla \phi(\xi(s)) - \nabla \phi(x)|_{D(A^*)} \xrightarrow{s \rightarrow 0} 0$$

uniformly in  $a(\cdot)$ .

If we read (37) in  $D(A^*)'$  it appears as an equation of the form

$$\begin{cases} \dot{u}(t) = \bar{A}u(t) + f(t), \\ u(0) = x, \end{cases}$$

where  $f(t)$  is a bounded measurable function, with  $|f(t)|_{D(A^*)'} \leq \beta|\delta_0|_{D(A^*)'}\|\Gamma\|$ , and  $\bar{A}$  is an extension of  $A$  that generates a  $C_0$ -semigroup on  $D(A^*)'$ . So<sup>5</sup> we can choose a constant  $C$  that depends on  $x$  such that, for all admissible continuous control  $a(\cdot)$  and all  $s \in (0, 1]$ ,

$$(50) \quad \frac{|y(s) - x|_{D(A^*)'}}{s} \leq C.$$

Thus from (49) and (50), we can say that  $|I_1| \xrightarrow{s \rightarrow 0} 0$  uniformly in  $a(\cdot)$ . Therefore

$$\left| \frac{\phi(y(s)) - \phi(x)}{s} - \frac{\langle \nabla \phi(x), y(s) - x \rangle}{s} \right| \xrightarrow{s \rightarrow 0} 0$$

uniformly in  $a(\cdot)$ . We now write

$$\begin{aligned} (51) \quad \frac{\langle \nabla \phi(x), y(s) - x \rangle}{s} &= \int_{\beta s}^{\bar{s}} \nabla \phi(x)[r] \frac{(x[r - \beta s] - x[r])}{s} dr \\ &\quad + \frac{\int_0^{\beta s} \nabla \phi(x)[r] (a(s - r/\beta) - x[r]) dr}{s} \\ &= \int_{\beta s}^{\bar{s} - \beta s} x[r] \frac{\nabla \phi(x)[r + \beta s] - \nabla \phi(x)[r]}{s} dr + \int_{\bar{s} - \beta s}^{\bar{s}} \frac{(-\nabla \phi(x)[r] x[r])}{s} dr \\ &\quad + \frac{\int_0^{\beta s} (\nabla \phi(x)[r + \beta s] x[r]) dr}{s} + \frac{\int_0^{\beta s} \nabla \phi(x)[r] a(s - r/\beta) dr}{s} \\ &\quad + \frac{\int_0^{\beta s} -\nabla \phi(x)[r] x[r] dr}{s}. \end{aligned}$$

The third and the fifth terms, which do not depend on the control, have opposite limits, the second goes to zero because  $\nabla \phi(x)$  and  $x$  are in  $D(A^*)$ , and then  $x[\bar{s}] = 0 = \nabla \phi(x)[\bar{s}]$ . The first term tends to  $\langle A^* \nabla \phi(x), x \rangle$ . Finally we observe that the only term that depends on the control is the fourth and

$$\left| \frac{\int_0^{\beta s} \nabla \phi(x)[r] a(s - r/\beta) dr}{s} - \beta \frac{\int_0^s \nabla \phi(x)[0] a(s - r') dr'}{s} \right| \xrightarrow{s \rightarrow 0} 0$$

uniformly in  $a(\cdot)$  and, since  $\phi(x)[0]$  is a constant,

$$\beta \frac{\int_0^s \nabla \phi(x)[0] a(s - r) dr}{s} = \frac{\int_0^s \langle \beta \delta_0 \nabla \phi(x), a(r) \rangle_{\mathbb{R}} dr}{s}.$$

This complete the proof.  $\square$

We can now prove that the value function is a solution of the HJB equation.

**THEOREM 4.7.** *The value function  $V$  is bounded and  $B$ -Lipschitz, and it is a solution of the HJB equation.*

<sup>5</sup>In view of the fact that  $x$  is in  $\mathcal{H} \subseteq D(\bar{A}) \subseteq D(A^*)'$ ; see [20] for a proof.



*Proof.* The boundedness of  $V$  follows from the boundedness of  $L$  (assumption (L2)). The  $B$ -Lipschitz property is the result of Proposition 3.4. It remains to verify that  $V$  is a solution of the HJB equation.

**$V$  is a subsolution:** Let  $x$  be a local maximum of  $V - (\phi + g)$  for  $\phi \in \text{test1}$  and  $g \in \text{test2}$ . Due to Proposition 4.1 we know that  $\nabla(\phi + g)(x) \in D(A^*)$ . Moreover we know that  $\nabla\phi(x) \in D(A^*)$  for the definition of the set  $\text{test1}$ . So  $\nabla g(x) = g'_0(|x|) \frac{x}{|x|} \in D(A^*)$ , which implies that  $x \in D(A^*)$ . We can assume that  $V(x) - (\phi + g)(x) = 0$ . We consider the constant control  $(\alpha(\cdot), a(\cdot)) \equiv (\alpha, a) \in \Sigma \times \Gamma$  and  $x(s)$  the trajectory starting from  $x$  and subject to  $(\alpha, a)$ . Then for  $s$  small enough

$$V(x(s)) - (\phi + g)(x(s)) \leq V(x) - (\phi + g)(x),$$

and thanks to Bellman's principle of optimality we know that

$$V(x) \leq e^{-\rho s} V(x(s)) + \int_0^s e^{-\rho r} L(x(r), \alpha, a) dr,$$

so that

$$(52) \quad \frac{1 - e^{-\rho s}}{s} V(x(s)) - \frac{\phi(x(s)) - \phi(x)}{s} - \frac{g(x(s)) - g(x)}{s} - \frac{\int_0^s e^{-\rho r} L(x(r), \alpha, a) dr}{s} \leq 0.$$

By using Propositions 4.5 and 4.6 and letting  $s \rightarrow 0$ , we obtain

$$(53) \quad \begin{aligned} & \rho V(x) - \langle \nabla\phi(x), -\mu x \rangle - \langle \nabla g(x), -\mu x \rangle \\ & - \left( \langle A^* \nabla\phi(x), x \rangle + \langle \beta \delta_0(\nabla\phi(x)), a \rangle_{\mathbb{R}} + \langle \nabla\phi(x), \alpha \rangle + \langle \nabla g(x), \alpha \rangle + L(x, \alpha, a) \right) \\ & \leq \frac{g'_0(|x|)}{|x|} \beta \frac{\|\Gamma\|^2}{2}. \end{aligned}$$

By taking the  $\inf_{(\alpha, a) \in \Sigma \times \Gamma}$  we obtain the subsolution inequality.

**$V$  is a supersolution:** Let  $\phi \in \text{test1}$  and  $g \in \text{test2}$  and  $x$  be a minimum for  $V + (\phi + g)$  and such that  $V + (\phi + g)(x) = 0$ . Then as observed above  $x \in D(A^*)$ . For some  $\varepsilon > 0$  let  $(\alpha_\varepsilon(\cdot), a_\varepsilon(\cdot))$  be an  $\varepsilon^2$ -optimal strategy. With no loss of generality we can assume  $a_\varepsilon(\cdot)$  continuous. We call  $x(s)$  the trajectory starting from  $x$  and subject to  $(\alpha_\varepsilon(\cdot), a_\varepsilon(\cdot))$ . Now for  $s$  small enough

$$V(x(s)) + (\phi + g)(x(s)) \geq V(x) + (\phi + g)(x),$$

and due to the  $\varepsilon^2$ -optimality and Bellman's principle we know that

$$V(x) + \varepsilon^2 \geq e^{-\rho s} V(x(s)) + \int_0^s e^{-\rho r} L(x(r), \alpha_\varepsilon(r), a_\varepsilon(r)) dr.$$

Then for  $s = \varepsilon$  we have

$$(54) \quad \frac{1 - e^{-\rho \varepsilon}}{\varepsilon} V(x(\varepsilon)) + \frac{\phi(x(\varepsilon)) - \phi(x)}{\varepsilon} + \frac{g(x(\varepsilon)) - g(x)}{\varepsilon} - \frac{\int_0^\varepsilon e^{-\rho r} L(x(r), \alpha_\varepsilon(r), a_\varepsilon(r)) dr}{\varepsilon} + \frac{\varepsilon^2}{\varepsilon} \geq 0.$$

In view of Propositions 4.5 and 4.6 we can choose, independently of the control  $(\alpha_\varepsilon(\cdot), a_\varepsilon(\cdot))$ , an  $O(s)$  such that:

$$(55) \quad \rho V(x) + \langle A^* \nabla \phi(x), x \rangle + \langle \nabla \phi(x) + \nabla g(x), -\mu x \rangle \\ - \left( \frac{\int_0^\varepsilon \langle -\beta \delta_0(\nabla \phi(x), a_\varepsilon(r)) \rangle_{\mathbb{R}} + e^{-\rho r} L(x, \alpha_\varepsilon(r), a_\varepsilon(r)) dr}{\varepsilon} \right. \\ \left. - \frac{\int_0^\varepsilon \langle \nabla \phi(x) + \nabla g(x), \alpha_\varepsilon(r) \rangle dr}{\varepsilon} \right) \geq O(\varepsilon) - \frac{g'_0(|x|)}{|x|} \beta \frac{\|\Gamma\|^2}{2}.$$

Next we take the infimum as  $\alpha$  and  $a$  varying in  $\Sigma \times \Gamma$  inside the integral and let  $\varepsilon \rightarrow 0$  to obtain that

$$(56) \quad \rho V(x) + \langle A^* \nabla \phi(x), x \rangle + \langle \nabla \phi(x) + \nabla g(x), -\mu x \rangle \\ - \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( -\langle \beta \delta_0(\nabla \phi(x)), a \rangle_{\mathbb{R}} + L(x, \alpha, a) - \langle \nabla \phi(x) + \nabla g(x), \alpha \rangle \right) \\ \geq -\frac{g'_0(|x|)}{|x|} \beta \frac{\|\Gamma\|^2}{2}$$

(we observe again that the fact that  $O(s) \xrightarrow{\varepsilon \rightarrow 0} 0$  uniformly in the control is essential). Therefore  $V$  is a solution of the HJB equation.  $\square$

**4.2. Uniqueness.** Now we prove a uniqueness result in the case  $\mu \neq 0$ . The proof of the case  $\mu = 0$  is similar with minor changes.

**THEOREM 4.8.** *Given a supersolution  $v$  of the HJB equation and a subsolution  $u$  we have*

$$u(x) \leq v(x) \text{ for every } x \in \mathcal{H}.$$

*In particular there exists at most one solution of the HJB equation.*

*Proof.* We proceed by contradiction. Assume that  $u$  is a subsolution of the HJB equation and  $v$  a supersolution, and suppose that there exists  $\tilde{x} \in \mathcal{H}$  and  $\gamma > 0$  such that

$$(u(\tilde{x}) - v(\tilde{x})) > \frac{3\gamma}{\rho} > 0.$$

We choose  $\gamma < 1$ . Given  $\vartheta > 0$  small enough we have

$$(57) \quad u(\tilde{x}) - v(\tilde{x}) - \vartheta |\tilde{x}|^2 > \frac{2\gamma}{\rho} > 0.$$

We consider  $\varepsilon > 0$  and  $\psi: \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$  given by

$$\psi(x, y) \stackrel{def}{=} u(x) - v(y) - \frac{1}{2\varepsilon} |B^{1/2}(x - y)|^2 - \frac{\vartheta}{2} |x|^2 - \frac{\vartheta}{2} |y|^2.$$

Thanks to the boundedness of  $u$  and  $v$ , chosen  $\vartheta > 0$ , there exist  $R_\vartheta > 0$  such that

$$(58) \quad \psi(0, 0) \geq \left( \sup_{(|x| \geq R_\vartheta) \text{ or } (|y| \geq R_\vartheta)} (\psi(x, y)) \right) + 1.$$

We set

$$S = \{(x, y) \in \mathcal{H} \times \mathcal{H} : |x| \leq R_\vartheta \text{ and } |y| \leq R_\vartheta\}.$$

If we choose  $R_\vartheta$  big enough  $\tilde{x} \in S$ . By standard techniques (see [29, page 252]), given  $\sigma > 0$ , we can find  $p$  and  $q$  in  $\mathcal{H}$ , with  $|p| < \sigma$  and  $|q| < \sigma$ , and such that

$$(x, y) \mapsto \psi(x, y) - \langle Bp, x \rangle - \langle Bq, y \rangle$$

attains a maximum  $(\bar{x}, \bar{y})$  in  $\bar{S}$ . If we choose  $\sigma$  small enough (for example, such that  $\sigma \|B\| R_\vartheta < \frac{1}{4} \frac{\gamma}{\rho}$ ), we know from (58) that such a maximum is in the interior of  $S$  and, thanks to (57), that

$$\psi(\bar{x}, \bar{y}) - \langle Bp, \bar{x} \rangle - \langle Bq, \bar{y} \rangle > \frac{3\gamma}{2\rho}.$$

Moreover

$$(59) \quad \psi(\bar{x}, \bar{y}) > \frac{\gamma}{\rho} \quad \text{and so} \quad u(\bar{x}) - v(\bar{y}) > \frac{\gamma}{\rho}.$$

In order to find an estimate in contradiction with (59) we prove some preliminary estimates.

**Estimates 1 (on  $\varepsilon$ ):** We observe that

$$\begin{cases} M: (0, 1] \rightarrow \mathbb{R}, \\ M: \varepsilon \mapsto \sup_{(x, y) \in \mathcal{H} \times \mathcal{H}} \left( u(x) - v(y) - \frac{1}{2\varepsilon} |B^{1/2}(x - y)|^2 \right) \end{cases}$$

is nonincreasing and bounded and so it admits a limit for  $\varepsilon \rightarrow 0^+$ . So there exists a  $\bar{\varepsilon} > 0$  such that for every  $\varepsilon_1, \varepsilon_2 \in (0, \bar{\varepsilon}]$  we have that

$$(60) \quad |M(\varepsilon_1) - M(\varepsilon_2)| < \left( \frac{\gamma}{16(1 + |\mu|)} \right)^2.$$

We choose now  $\varepsilon$  that will be fixed in the proof:

$$(61) \quad \varepsilon := \min \left\{ \bar{\varepsilon}, \frac{1}{32C_L^2} \right\}$$

( $C_L$  is the constant introduced in hypotheses (L1) and (L2)). This concludes the estimates on  $\varepsilon$ .

Now we state and prove a claim that we will use in the estimate on  $\sigma$  and  $\vartheta$ .

*Claim.* If  $\tilde{x} \in \mathcal{H}$  and  $\tilde{y} \in \mathcal{H}$  satisfy

$$(62) \quad u(\tilde{x}) - v(\tilde{y}) - \frac{1}{2\varepsilon} |B^{1/2}(\tilde{x} - \tilde{y})|^2 \geq M(\varepsilon) - \left( \frac{\gamma}{16(1 + |\mu|)} \right)^2,$$

then

$$(63) \quad \frac{1}{\varepsilon} |B^{1/2}(\tilde{x} - \tilde{y})|^2 \leq \frac{1}{32} \left( \frac{\gamma}{(1 + |\mu|)} \right)^2.$$

*Proof of the claim.* (We follow the idea used in Lemma 3.2 of [16])

$$\begin{aligned}
 (64) \quad M(\varepsilon/2) &\geq u(\tilde{x}) - y(\tilde{y}) - \frac{1}{4\varepsilon} \left| B^{1/2}(\tilde{x} - \tilde{y}) \right|^2 \\
 &= u(\tilde{x}) - y(\tilde{y}) - \frac{1}{2\varepsilon} \left| B^{1/2}(\tilde{x} - \tilde{y}) \right|^2 + \frac{1}{4\varepsilon} \left| B^{1/2}(\tilde{x} - \tilde{y}) \right|^2 \\
 &\geq M(\varepsilon) - \left( \frac{\gamma}{16(1+|\mu|)} \right)^2 + \frac{1}{4\varepsilon} \left| B^{1/2}(\tilde{x} - \tilde{y}) \right|^2,
 \end{aligned}$$

so that

$$\begin{aligned}
 (65) \quad \frac{1}{4\varepsilon} \left| B^{1/2}(\tilde{x} - \tilde{y}) \right|^2 &\leq M(\varepsilon/2) - M(\varepsilon) + \left( \frac{\gamma}{16(1+|\mu|)} \right)^2 \\
 &\leq \left( \frac{\gamma}{16(1+|\mu|)} \right)^2 + \left( \frac{\gamma}{16(1+|\mu|)} \right)^2 = 2 \left( \frac{\gamma}{16(1+|\mu|)} \right)^2,
 \end{aligned}$$

where the inequality  $M(\varepsilon/2) - M(\varepsilon) < \left( \frac{\gamma}{16(1+|\mu|)} \right)^2$  follows from the definition of  $\varepsilon$  (61) that implies  $\varepsilon \leq \bar{\varepsilon}$  and then (60). The claim follows.

Note that from (61) we have

$$\frac{1}{\sqrt{\varepsilon}} \geq 4\sqrt{2}C_L,$$

and then, if  $\tilde{x}, \tilde{y}$  satisfy the hypothesis (62) of the claim, we have

$$(66) \quad C_L |\tilde{x} - \tilde{y}|_B \leq \frac{\gamma}{32(1+|\mu|)}.$$

**Estimates 2 (on  $\sigma$ ):** We have already imposed  $\sigma < \frac{\gamma/\rho}{4\|B\|R_\vartheta}$  we set here and in the following

$$(67) \quad \sigma = \min \left\{ \frac{\gamma}{8\rho\|B\|R_\vartheta}, \vartheta, \frac{\vartheta}{R_\vartheta} \right\},$$

so that

$$(68) \quad \sigma \xrightarrow{\vartheta \rightarrow 0} 0$$

and

$$(69) \quad \sigma R_\vartheta \xrightarrow{\vartheta \rightarrow 0} 0.$$

We recall that we have already fixed  $\varepsilon$  in (61). From the choice of  $\sigma$  (67) follows that

$$(70) \quad |\langle Bp, \bar{x} \rangle| \leq \|B\|\sigma R_\vartheta \xrightarrow{\vartheta \rightarrow 0} 0, \quad |\langle Bq, \bar{y} \rangle| \leq \|B\|\sigma R_\vartheta \xrightarrow{\vartheta \rightarrow 0} 0.$$

Moreover, in view of the continuity of the linear operator  $A^*B: \mathcal{H} \rightarrow \mathcal{H}$  with norm  $\|A^*B\|$ , we have

$$(71) \quad |\langle A^*Bp, \bar{x} \rangle| \leq \|A^*B\|\sigma R_\vartheta \xrightarrow{\vartheta \rightarrow 0} 0, \quad |\langle A^*Bq, \bar{y} \rangle| \leq \|B\|\sigma R_\vartheta \xrightarrow{\vartheta \rightarrow 0} 0.$$

This concludes the estimates on  $\sigma$ .

**Estimates 3 (on  $\vartheta$ ):** One can prove that with fixed  $\varepsilon$  we have

$$(72) \quad \vartheta |\bar{x}|^2 \xrightarrow{\vartheta \rightarrow 0} 0, \quad \vartheta |\bar{y}|^2 \xrightarrow{\vartheta \rightarrow 0} 0$$

(it is a quite standard fact; see, for example, [16]). So

$$(73) \quad \lim_{\vartheta \rightarrow 0} (\psi(\bar{x}, \bar{y}) - \langle Bp, \bar{x} \rangle - \langle Bq, \bar{y} \rangle) \\ = \sup_{(x,y) \in \mathcal{H} \times \mathcal{H}} \left( x(x) - v(y) - \frac{1}{2\varepsilon} \left| B^{1/2}(x-y) \right|^2 \right) > 2\frac{\gamma}{\rho}$$

(where the last inequality follows from (57)). In (67) we chose  $\sigma$  as a function of  $\vartheta$ , and now we fix  $\vartheta$ . We begin by taking

$$\vartheta < \frac{\gamma}{64\beta\|\Gamma\|^2}$$

so that

$$(74) \quad \beta\vartheta\|\Gamma\|^2 < \frac{\gamma}{64}.$$

We know from (70) and (71) that, if we choose  $\vartheta$  small enough, we have

$$(75) \quad |\mu| |\langle Bp, \bar{x} \rangle| < \frac{\gamma}{16}, \quad |\mu| |\langle Bq, \bar{y} \rangle| < \frac{\gamma}{16}, \\ |\langle A^*Bp, \bar{x} \rangle| < \frac{\gamma}{16}, \quad |\langle A^*Bq, \bar{y} \rangle| < \frac{\gamma}{16}.$$

From (72) we know that, if we choose  $\vartheta$  small enough, we have

$$(76) \quad |\mu|\vartheta |\bar{x}|^2 < \frac{\gamma}{32}, \quad |\mu|\vartheta |\bar{y}|^2 < \frac{\gamma}{32}.$$

Moreover (72) implies also that

$$\vartheta |\bar{x}| \xrightarrow{\vartheta \rightarrow 0} 0, \quad \vartheta |\bar{y}| \xrightarrow{\vartheta \rightarrow 0} 0,$$

and then, if we choose  $\vartheta$  small enough, we have

$$(77) \quad \vartheta \|\Sigma\| (|\bar{x}| + |\bar{y}|) < \frac{\gamma}{32}.$$

Moreover in view of (73) we know that, if we choose  $\vartheta$  small enough, then  $\bar{x}$  and  $\bar{y}$  satisfy the hypothesis (62) of the claim, and then, from (63), we have

$$(78) \quad \frac{1}{\varepsilon} \left| B^{1/2}(\bar{x} - \bar{y}) \right|^2 \leq \frac{1}{32} \left( \frac{\gamma}{(1+|\mu|)} \right)^2 \leq \frac{1}{32} \frac{\gamma}{(1+|\mu|)} \leq \frac{\gamma}{32}$$

(we took  $0 < \gamma < 1$  and then  $\gamma^2 < \gamma$ ). From (63) in the same way we obtain

$$(79) \quad \frac{|\mu|}{\varepsilon} \left| B^{1/2}(\bar{x} - \bar{y}) \right|^2 \leq \frac{\gamma}{32}$$

and, from (66),

$$(80) \quad C_L \left| B^{1/2}(\bar{x} - \bar{y}) \right| \leq \frac{\gamma}{32}.$$

Eventually, if we choose  $\vartheta$  small enough in (68), we have

$$(81) \quad 2\|B\|\|p\|\|\Sigma\| \leq 2\|B\|\|\sigma\|\|\Sigma\| \leq \frac{\gamma}{64}$$

and

$$(82) \quad 2\sigma\beta\|\delta_0 \circ B\|\|\Gamma\| \leq \frac{\gamma}{32},$$

where  $\|\delta_0 \circ B\|$  is the norm of the linear continuous functional  $\delta_0 \circ B: \mathcal{H} \rightarrow \mathbb{R}$ .

Finally we fix  $\vartheta > 0$  small enough to satisfy (75), (76), (77), (78), (79) (80), (81), and (82).

Now we proceed with the proof of the theorem. The map

$$x \mapsto u(x) - \left( \frac{1}{2\varepsilon} |B^{1/2}(x - \bar{y})|^2 + \frac{\vartheta}{2} |x|^2 + \langle Bp, x \rangle \right)$$

attains a maximum at  $\bar{x}$ , and

$$y \mapsto v(y) + \left( \frac{1}{2\varepsilon} |B^{1/2}(\bar{x} - y)|^2 + \frac{\vartheta}{2} |y|^2 + \langle Bq, y \rangle \right)$$

attains a minimum at  $\bar{y}$ .

Note that, due to Proposition 4.1,  $\bar{x}$  and  $\bar{y}$  are in  $D(A^*)$ . We can now use the definition of sub- and supersolution to obtain

$$(83) \quad \begin{aligned} & \rho u(\bar{x}) - \frac{1}{\varepsilon} \langle A^* B(\bar{x} - \bar{y}), \bar{x} \rangle - \frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), -\mu \bar{x} \rangle - \langle A^* Bp, \bar{x} \rangle \\ & - \langle Bp, -\mu \bar{x} \rangle - \vartheta \langle \bar{x}, -\mu \bar{x} \rangle - \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( \frac{1}{\varepsilon} \langle \beta \delta_0(B(\bar{x} - \bar{y})), a \rangle_{\mathbb{R}} + \langle \beta \delta_0(Bp), a \rangle_{\mathbb{R}} \right. \\ & \left. + \vartheta \langle \bar{x}, \alpha \rangle + \frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), \alpha \rangle + \langle Bp, \alpha \rangle + L(\bar{x}, \alpha, a) \right) \leq \frac{\vartheta \beta \|\Gamma\|^2}{2} \end{aligned}$$

and

$$(84) \quad \begin{aligned} & \rho v(\bar{y}) - \frac{1}{\varepsilon} \langle A^* B(\bar{x} - \bar{y}), \bar{y} \rangle - \frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), -\mu \bar{y} \rangle + \langle A^* Bq, \bar{y} \rangle \\ & + \langle Bq, -\mu \bar{y} \rangle + \vartheta \langle \bar{y}, -\mu \bar{y} \rangle - \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( \frac{1}{\varepsilon} \langle \beta \delta_0(B(\bar{x} - \bar{y})), a \rangle_{\mathbb{R}} - \langle \beta \delta_0(Bq), a \rangle_{\mathbb{R}} \right. \\ & \left. - \vartheta \langle \bar{y}, \alpha \rangle + \frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), \alpha \rangle - \langle Bq, \alpha \rangle + L(\bar{y}, \alpha, a) \right) \geq -\frac{\vartheta \beta \|\Gamma\|^2}{2}. \end{aligned}$$

The two estimates together give

$$\begin{aligned}
 (85) \quad & \rho u(\bar{x}) - \rho v(\bar{y}) - \frac{1}{\varepsilon} \langle A^* B(\bar{x} - \bar{y}), (\bar{x} - \bar{y}) \rangle \\
 & - \frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), -\mu(\bar{x} - \bar{y}) \rangle - \langle A^* Bp, \bar{x} \rangle - \langle A^* Bq, \bar{y} \rangle \\
 & - \langle Bp, -\mu\bar{x} \rangle - \langle Bq, -\mu\bar{y} \rangle - \vartheta \langle \bar{x}, -\mu\bar{x} \rangle - \vartheta \langle \bar{y}, -\mu\bar{y} \rangle \\
 & - \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( \frac{1}{\varepsilon} \langle \beta \delta_0(B(\bar{x} - \bar{y})), a \rangle_{\mathbb{R}} + \langle \beta \delta_0(Bp), a \rangle_{\mathbb{R}} \right. \\
 & \left. + \vartheta \langle \bar{x}, \alpha \rangle + \frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), \alpha \rangle + \langle Bp, \alpha \rangle + L(\bar{x}, \alpha, a) \right) \\
 & + \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( \frac{1}{\varepsilon} \langle \beta \delta_0(B(\bar{x} - \bar{y})), a \rangle_{\mathbb{R}} - \langle \beta \delta_0(Bq), a \rangle_{\mathbb{R}} \right. \\
 & \left. - \vartheta \langle \bar{y}, \alpha \rangle + \frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), \alpha \rangle - \langle Bq, \alpha \rangle + L(\bar{y}, \alpha, a) \right) \leq \beta \vartheta \|\Gamma\|^2.
 \end{aligned}$$

We now note that from (10)  $A^* B \leq B$ , and then we have

$$(86) \quad -\frac{1}{\varepsilon} \langle A^* B(\bar{x} - \bar{y}), (\bar{x} - \bar{y}) \rangle \geq -\frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), (\bar{x} - \bar{y}) \rangle = -\frac{1}{\varepsilon} |\bar{x} - \bar{y}|_B^2.$$

Moreover we observe that

$$\begin{aligned}
 (87) \quad & - \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( \frac{1}{\varepsilon} \langle \beta \delta_0(B(\bar{x} - \bar{y})), a \rangle_{\mathbb{R}} + \langle \beta \delta_0(Bp), a \rangle_{\mathbb{R}} \right. \\
 & \left. + \vartheta \langle \bar{x}, \alpha \rangle + \frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), \alpha \rangle + \langle Bp, \alpha \rangle + L(\bar{x}, \alpha, a) \right) \\
 & + \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( \frac{1}{\varepsilon} \langle \beta \delta_0(B(\bar{x} - \bar{y})), a \rangle_{\mathbb{R}} - \langle \beta \delta_0(Bq), a \rangle_{\mathbb{R}} \right. \\
 & \left. - \vartheta \langle \bar{y}, \alpha \rangle + \frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), \alpha \rangle - \langle Bq, \alpha \rangle + L(\bar{y}, \alpha, a) \right) \\
 & \geq \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( - \langle \beta \delta_0(Bp), a \rangle_{\mathbb{R}} - \langle \beta \delta_0(Bq), a \rangle_{\mathbb{R}} + L(\bar{y}, \alpha, a) - L(\bar{x}, \alpha, a) \right. \\
 & \quad \left. - \vartheta \langle \bar{y}, \alpha \rangle - \vartheta \langle \bar{x}, \alpha \rangle - \langle Bq, \alpha \rangle - \langle Bp, \alpha \rangle \right) \\
 & \geq \inf_{(\alpha, a) \in \Sigma \times \Gamma} \left( L(\bar{y}, \alpha, a) - L(\bar{x}, \alpha, a) \right) \\
 & \quad - \sup_{(\alpha, a) \in \Sigma \times \Gamma} \left( \langle \beta \delta_0(Bp), a \rangle_{\mathbb{R}} + \langle \beta \delta_0(Bq), a \rangle_{\mathbb{R}} \right) \\
 & - \sup_{(\alpha, a) \in \Sigma \times \Gamma} \left( \vartheta \langle \bar{y}, \alpha \rangle + \vartheta \langle \bar{x}, \alpha \rangle \right) - \sup_{(\alpha, a) \in \Sigma \times \Gamma} \left( \langle Bq, \alpha \rangle + \langle Bp, \alpha \rangle \right) \\
 & \geq -C_L |\bar{x} - \bar{y}|_B - 2\sigma\beta \|\delta_0 \circ B\| \|\Gamma\| - \|\Sigma\| \vartheta(|\bar{x}| + |\bar{y}|) - 2\|B\| \sigma \|\Sigma\|.
 \end{aligned}$$

Thus by using (86) and (87) in (85) we derive

$$\begin{aligned}
 (88) \quad & \rho(u(\bar{x}) - v(\bar{y})) - \frac{1}{\varepsilon} |\bar{x} - \bar{y}|_B^2 \\
 & - \frac{\mu}{\varepsilon} \langle B(\bar{x} - \bar{y}), -(\bar{x} - \bar{y}) \rangle - \langle A^* Bp, \bar{x} \rangle - \langle A^* Bq, \bar{y} \rangle \\
 & - \langle Bp, -\mu\bar{x} \rangle - \langle Bq, -\mu\bar{y} \rangle - \vartheta \langle \bar{x}, -\mu\bar{x} \rangle - \vartheta \langle \bar{y}, -\mu\bar{y} \rangle \\
 & - C_L |\bar{x} - \bar{y}|_B - 2\sigma\beta \|\delta_0 \circ B\| \|\Gamma\| - \|\Sigma\| \vartheta(|\bar{x}| + |\bar{y}|) - 2\|B\| \sigma \|\Sigma\| - \beta\vartheta \|\Gamma\|^2 \leq 0,
 \end{aligned}$$

and from the preceding and from (78), (79), (75), (76), (80), (82), (77), (81), and (74) we obtain

$$(89) \quad \rho(u(\bar{x}) - v(\bar{y})) - 2\left(\frac{\gamma}{32}\right) - 4\left(\frac{\gamma}{16}\right) - 2\left(\frac{\gamma}{32}\right) - \frac{\gamma}{32} - \frac{\gamma}{32} - \frac{\gamma}{32} - \frac{\gamma}{64} - \frac{\gamma}{64} \leq 0;$$

that is,

$$(90) \quad \rho(u(\bar{x}) - v(\bar{y})) - \frac{1}{2}\gamma \leq 0.$$

Now we recall that from (59) we have  $\rho(u(\bar{x}) - v(\bar{y})) > \gamma$ , and then we obtain from (90)

$$\frac{1}{2}\gamma = \gamma - \frac{1}{2}\gamma < \rho(u(\bar{x}) - v(\bar{y})) - \frac{1}{2}\gamma \leq 0,$$

which yields a contradiction because  $\gamma > 0$ . The theorem is proved.  $\square$

**Remark 4.9.** Now we can better explain the remark in the introduction saying that it is difficult to treat the case of a nonconstant coefficient. We can estimate the term  $\frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), -\mu(\bar{x} - \bar{y}) \rangle$  because we use the term  $\frac{1}{\varepsilon} |\bar{x} - \bar{y}|_B^2$  to penalize the doubling of the variables with respect to the  $B$ -norm. If  $\mu$  is a function of  $r$ , such a term is replaced by  $\frac{1}{\varepsilon} \langle B(\bar{x} - \bar{y}), -\mu(\cdot)(\bar{x} - \bar{y}) \rangle$ , where  $-\mu(\cdot)(\bar{x} - \bar{y})$  is the pointwise product of the  $L^\infty(0, \bar{s})$  function  $\mu(\cdot)$  and the  $L^2(0, \bar{s})$  function  $(\bar{x} - \bar{y})$ , which cannot be estimated by using similar arguments.

**Acknowledgments.** The author thanks Prof. Andrzej Świąch for his hospitality, his great kindness, many useful suggestions, and stimulating conversations. Thanks to Silvia Faggian for the invaluable advice.

#### REFERENCES

- [1] M. ADAMO, A. L. AMADORI, M. BERNASCHI, C. LA CHIOMA, A. MARIGO, B. PICCOLI, S. SBARAGLIA, A. UBOLDI, D. VERGNI, P. FABBRI, D. IACOVONI, F. NATALE, S. SCALERA, L. SPILOTRO, AND A. VALLETTA, *Optimal strategies for the issuances of public debt securities*, Int. J. Theor. Appl. Finance, 7 (2004), pp. 805–822.
- [2] S. ANİTA, *Analysis and Control of Age-Dependent Population Dynamics*, Math. Model. Theory Appl. 11, Kluwer Academic Publishers, Dordrecht, 2000.
- [3] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Res. Notes Math. 86, Pitman (Advanced Publishing Program), Boston, MA, 1983.
- [4] V. BARBU, G. DA PRATO, AND C. POPA, *Existence and uniqueness of the dynamic programming equation in Hilbert space*, Nonlinear Anal., 7 (1983), pp. 283–299.
- [5] E. BARUCCI AND F. GOZZI, *Optimal investment in a vintage capital model*, Res. Econ., (1998), pp. 159–188.
- [6] E. BARUCCI AND F. GOZZI, *Optimal advertising with a continuum of goods*, Ann. Oper. Res., 88 (1999), pp. 15–29.



- [7] E. BARUCCI AND F. GOZZI, *Technology adoption and accumulation in a vintage capital model*, J. Econ., 74 (2001), pp. 1–30.
- [8] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite-Dimensional Systems*. Vol. 1, Systems & Control: Foundations & Applications, Birkhäuser Boston Inc., Boston, MA, 1992.
- [9] P. CANNARSA, F. GOZZI, AND H. M. SONER, *A dynamic programming approach to nonlinear boundary control problems of parabolic type*, J. Funct. Anal., 117 (1993), pp. 25–61.
- [10] P. CANNARSA AND M. E. TESSITORE, *Cauchy problem for Hamilton-Jacobi equations and Dirichlet boundary control problems of parabolic type*, in Control of Partial Differential Equations and Applications (Laredo, 1994), Lecture Notes in Pure and Appl. Math. 174, Dekker, New York, 1996, pp. 31–42.
- [11] P. CANNARSA AND M. E. TESSITORE, *Dynamic programming equation for a class of nonlinear boundary control problems of parabolic type*, Control Cybernet., 25 (1996), pp. 483–495.
- [12] P. CANNARSA AND M. E. TESSITORE, *Infinite-dimensional Hamilton-Jacobi equations and Dirichlet boundary control problems of parabolic type*, SIAM J. Control Optim., 34 (1996), pp. 1831–1847.
- [13] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
- [14] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations in infinite dimensions. IV. Hamiltonians with unbounded linear terms*, J. Funct. Anal., 90 (1990), pp. 237–283.
- [15] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations in infinite dimensions. V. Unbounded linear terms and B-continuous solutions*, J. Funct. Anal., 97 (1991), pp. 417–465.
- [16] M. G. CRANDALL AND P. L. LIONS, *Hamilton-Jacobi equations in infinite dimensions. VI. Nonlinear A and Tataru's method refined*, in Evolution Equations, Control Theory, and Biomathematics (Han sur Lesse, 1991), Lecture Notes in Pure and Appl. Math. 155, Dekker, New York, 1994, pp. 51–89.
- [17] G. DA PRATO AND J. ZABCZYK, *Stochastic Equations in Infinite Dimensions*, Encyclopedia of Mathematics and Its Applications 44, Cambridge University Press, London, 1992, p. 454.
- [18] L. C. EVANS, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998.
- [19] S. FAGGIAN, *Applications of dynamic programming to economic problems with vintage capital*, Dyn. Contin. Discrete Impuls. Syst., to appear.
- [20] S. FAGGIAN, *First Order Hamilton-Jacobi Equations in Hilbert Spaces, Boundary Optimal Control and Applications to Economics*, Ph.D. thesis, Università degli studi di Pisa, 2001/2002.
- [21] S. FAGGIAN, *Regular solutions of first-order Hamilton-Jacobi equations for boundary control problems and applications to economics*, Appl. Math. Optim., 51 (2005), pp. 123–162.
- [22] S. FAGGIAN AND F. GOZZI, *On the dynamic programming approach for optimal control problems of PDE's with age structure*, Math. Popul. Stud., 11 (2004), pp. 233–270.
- [23] G. FEICHTINGER, A. PRSKAWETZ, AND V. M. VELIOV, *Age-structured optimal control in population economics*, Theor. Popul. Biol., 65 (2004), pp. 373–387.
- [24] F. GOZZI, S. S. SRITHARAN, AND A. ŚWIĘCH, *Viscosity solutions of dynamic-programming equations for the optimal control of the two-dimensional Navier-Stokes equations*, Arch. Ration. Mech. Anal., 163 (2002), pp. 295–327.
- [25] R. HARTL, P. KORT, V. VELIOV, AND G. FEICHTINGER, *Capital Accumulation under Technological Progress and Learning: A Vintage Capital Approach*, mimeo, 2003, Institute of management, University of Vienna.
- [26] M. IANNELLI, *Mathematical Theory of Age-Structured Population Dynamics*, Applied Mathematics Monographs, Comitato Nazionale per le scienze matematiche. C.N.R. 7, Giardini, Pisa, 1995.
- [27] M. IANNELLI, M. MARTCHEVA, AND F. A. MILNER, *Gender-Structured Population Modeling*, Frontiers Appl. Math. 31, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.
- [28] H. ISHII, *Viscosity solutions of nonlinear second-order partial differential equations in Hilbert spaces*, Comm. Partial Differential Equations, 18 (1993), pp. 601–650.
- [29] X. J. LI AND J. M. YONG, *Optimal control theory for infinite-dimensional systems*, in Systems & Control: Foundations & Applications, Birkhäuser Boston Inc., Boston, MA, 1995.
- [30] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983, p. 279.
- [31] M. RENARDY, *Polar decomposition of positive operators and a problem of Crandall and Lions*, Appl. Anal., 57 (1995), pp. 383–385.

- [32] K. SHIMANO, *A class of Hamilton-Jacobi equations with unbounded coefficients in Hilbert spaces*, Appl. Math. Optim., 45 (2002), pp. 75–98.
- [33] D. TATARU, *Viscosity solutions for the dynamic programming equations*, Appl. Math. Optim., 25 (1992), pp. 109–126.
- [34] D. TATARU, *Viscosity solutions of Hamilton-Jacobi equations with unbounded nonlinear terms*, J. Math. Anal. Appl., 163 (1992), pp. 345–392.
- [35] D. TATARU, *Viscosity solutions for Hamilton-Jacobi equations with unbounded nonlinear term: A simplified approach*, J. Differential Equations, 111 (1994), pp. 123–146.
- [36] K. YOSIDA, *Functional Analysis*, Classics Math., Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.

## STABILITY AND ERGODICITY OF PIECEWISE DETERMINISTIC MARKOV PROCESSES\*

O. L. V. COSTA<sup>†</sup> AND F. DUFOUR<sup>‡</sup>

**Abstract.** The main goal of this paper is to establish some equivalence results on stability, recurrence, and ergodicity between a piecewise deterministic Markov process (PDMP)  $\{X(t)\}$  and an embedded discrete-time Markov chain  $\{\Theta_n\}$  generated by a Markov kernel  $G$  that can be explicitly characterized in terms of the three local characteristics of the PDMP, leading to tractable criterion results. First we establish some important results characterizing  $\{\Theta_n\}$  as a sampling of the PDMP  $\{X(t)\}$  and deriving a connection between the probability of the first return time to a set for the discrete-time Markov chains generated by  $G$  and the resolvent kernel  $R$  of the PDMP. From these results we obtain equivalence results regarding irreducibility, existence of  $\sigma$ -finite invariant measures, and (positive) recurrence and (positive) Harris recurrence between  $\{X(t)\}$  and  $\{\Theta_n\}$ , generalizing the results of [F. Dufour and O. L. V. Costa, *SIAM J. Control Optim.*, 37 (1999), pp. 1483–1502] in several directions. Sufficient conditions in terms of a modified Foster–Lyapunov criterion are also presented to ensure positive Harris recurrence and ergodicity of the PDMP. We illustrate the use of these conditions by showing the ergodicity of a capacity expansion model.

**Key words.** piecewise deterministic Markov process, recurrence, ergodicity

**AMS subject classifications.** 60J25, 60J10, 93E15

**DOI.** 10.1137/060670109

**1. Introduction.** Piecewise deterministic Markov processes (PDMPs) were introduced in the literature by Davis [6] as a general class of stochastic models. PDMPs are a family of Markov processes involving deterministic motion punctuated by random jumps. The motion of the PDMP  $\{X(t)\}$  depends on three local characteristics, namely, the flow  $\Phi$ , the jump rate  $\lambda$ , and the transition measure  $Q$ , which specifies the postjump location. Starting from  $x$  the motion of the process follows the flow  $\Phi(x, t)$  until the first jump time  $T_1$ , which occurs either spontaneously in a Poisson-like fashion with rate  $\lambda$  or when the flow  $\Phi(x, t)$  hits the boundary of the state space. In either case the location of the process at the jump time  $T_1$  is selected by the transition measure  $Q(\Phi(x, T_1), \cdot)$  and the motion restarts from this new point as before. A suitable choice of the state space and the local characteristics  $\Phi$ ,  $\lambda$ , and  $Q$  provide stochastic models covering a great number of problems in operations research [6].

Over the past decades a great deal of attention has been given to the stability properties and related ergodic theory of Markov processes. One of the main approaches to dealing with these problems is to show that the recurrence properties of the Markov process under consideration are related to the recurrence properties of an associated discrete-time Markov chain obtained from a sampling of the original process, so that the well-known discrete-time Markov chains results could be used (see, for example, the books [12, 17, 18] and the references therein).

---

\*Received by the editors September 18, 2006; accepted for publication (in revised form) November 7, 2007; published electronically March 5, 2008.

<http://www.siam.org/journals/sicon/47-2/67010.html>

<sup>†</sup>Departamento de Engenharia de Telecomunicações e Controle, Escola Politécnica da Universidade de São Paulo, CEP 05508-900, São Paulo, S.P., Brazil (oswaldo@lac.usp.br). This author received financial support from CNPq (Brazilian National Research Council) through grant 304866/03-2, FAPESP (Research Council of the State of São Paulo) through grant 03/06736-7, PRONEX through grant 015/98, and IM-AGIMB.

<sup>‡</sup>Université Bordeaux 1, Institut de Mathématiques de Bordeaux and INRIA Futurs Team: CQFD, 351 cours de la Libération, 33405 Talence Cedex, France (dufour@math.u-bordeaux1.fr).

In the continuous-time context, Azéma, Duflo, and Revuz [1, 2] showed that a general Markov process and its associated resolvent admit the same recurrence properties. It was proved by Tuominen and Tweedie [19] that the recurrence structure of a Markov process  $\{X(t)\}$  with transition semigroup  $\{P^t\}$  and the Markov chain with kernel  $K_F = \int P^t F(dt)$ , where  $F$  is a distribution on  $[0, \infty)$ , are essentially equivalent, provided that a continuity assumption on  $\{P^t\}$  is satisfied, an assumption later suppressed in a fundamental paper by Meyn and Tweedie [11]. It must be pointed out that these results are related to the sampling of a continuous-time process  $\{X(t)\}$ , sampled at random times defined by an independent undelayed renewal process. This idea of randomized sampling was generalized to state-dependent sampling to provide some more powerful state-dependent drift criteria in order to ensure stability of the original Markov process. Within this context, Malyšev and Men'sikov [9] derived a modified Foster–Lyapunov criterion to establish recurrence properties for discrete-time Markov chains with countable state space. Meyn and Tweedie [15] generalized this work to discrete-time Markov chains with a general state space and furthermore obtained state-dependent drift conditions to get geometric ergodic properties. The generalization to continuous-time models was established by Dai and Meyn [5] in the context of general state space Markovian queuing models. In particular, they provided sufficient conditions for the existence of bounds on the long-run average moments and rates of convergence of the  $p$ th moments to their steady-state values. Another paper related to this subject is [4].

The main goal of this paper is to establish equivalence results on stability, recurrence, and ergodicity between a PDMP and a discrete-time Markov chain generated by a kernel  $G$  (see (2.2)–(2.4) for its definition) that can be explicitly characterized in terms of the three local characteristics of the PDMP leading to tractable criterion results. From a practical point of view, it should be noted that the results developed in [2, 11, 19] would be hard to apply for PDMPs because the transition semigroup of the PDMP as well as its associated resolvent kernel cannot be explicitly calculated from its local characteristics, which is not the case regarding the kernel  $G$ . As shown in Theorem 3.1 below,  $G$  generates a Markov chain that corresponds to a state-dependent sampling of the PDMP  $\{X(t)\}$ , providing an interesting parallel between our work in the continuous-time context and the results obtained in [15] in the discrete-time setting. However, it must be stressed that [15] provides general sufficient conditions to ensure that stability of the sampled chain implies stability of the Markov process, but not the converse. One of the main goals of our paper is to show the converse for PDMPs and, in fact, that the PDMP and the discrete-time Markov chain generated by this tractable kernel  $G$  have an equivalent recurrence structure. This is an important feature that distinguishes our work from [15]. We show that the following equivalence results hold:

- (i) The PDMP  $\{X(t)\}$  is irreducible if and only if the Markov chain  $\{\Theta_n\}$  associated to  $G$  is irreducible; see Proposition 4.1.
- (ii) There is a one-to-one correspondence between the set of invariant measures for the PDMP  $\{X(t)\}$  and for the Markov chain  $\{\Theta_n\}$  associated to  $G$ ; see Theorem 4.2.
- (iii) The PDMP  $\{X(t)\}$  is recurrent if and only if the Markov chain  $\{\Theta_n\}$  associated to  $G$  is recurrent; see Theorem 4.4.
- (iv) The PDMP is Harris recurrent if and only if the Markov chain associated to  $G$  is Harris recurrent; see Theorem 4.6.
- (v) The PDMP is positive recurrent (respectively, positive Harris recurrent) if and

only if the Markov chain associated to  $G$  is recurrent (respectively, Harris recurrent) with invariant measure satisfying a boundedness condition; see Corollary 4.5 (respectively, Corollary 4.7).

It is interesting to note that (v) also highlights some differences between our approach and some general results in the literature [2, 11, 15, 19]. Indeed, as shown in [11, Theorem 3.1], the Markov chain generated by  $K_F$  is positive Harris recurrent if and only if the process  $\{X(t)\}$  is positive Harris recurrent, while in our case the PDMP is positive Harris recurrent if and only if the Markov chain associated to  $G$  is Harris recurrent with its unique invariant measure satisfying a boundedness condition, which is far less demanding than positive Harris recurrence.

It should be pointed out that the proof of the Harris recurrence equivalence (item (iv)) requires two important results. One of them, Theorem 3.3, establishes a connection between the probability of the first return time to a set considering the Markov kernels generated by  $G$  and the resolvent kernel  $R$ . From this result we obtain the first part of the equivalence; if the process  $\{X(t)\}$  is Harris recurrent, then so is the Markov chain  $\{\Theta_n\}$ . The other result is presented in Theorem 3.1, which provides the sample path characterization of the Markov chain generated by  $G$  through the PDMP  $\{X(t)\}$ . From this result we get the second part of the equivalence; if the Markov chain  $\{\Theta_n\}$  is Harris recurrent, then so is the PDMP  $\{X(t)\}$ .

After we obtain the recurrence structure for PDMPs in terms of  $G$ , a natural question that would arise is what could be said about tractable ergodicity conditions for the PDMPs. As shown in [13, Theorem 6.1], if  $\{X(t)\}$  is positive Harris recurrent, then it is ergodic if and only if some skeleton chain is irreducible. Again, the problem is that the transition semigroup and consequently the skeleton chain for  $\{X(t)\}$  cannot be explicitly calculated from the local characteristics of the PDMP. We provide a tractable and equivalent condition to ensure that a skeleton chain is irreducible.

We conclude the paper by presenting sufficient conditions in a modified Foster–Lyapunov criterion form, written in terms of  $G$ , to ensure the following for the PDMP: the existence of an invariant probability measure, positive Harris recurrence, and ergodicity.

The paper is organized as follows. In section 2 we present some basic definitions related to the motion of a PDMP, introduce the Markov kernel  $G$ , and recall some classical definitions related to Markov processes both in the discrete-time and continuous-time contexts. Some preliminary results are derived in section 3 that will be important to obtain the equivalence properties for the stability of the PDMPs and the Markov kernel  $G$ . In section 4, it will be established that the stability and ergodic properties are equivalent for the PDMPs and the kernel  $G$ . In section 5 we establish some sufficient conditions to ensure various concepts of stability of the PDMP through a Foster–Lyapunov criterion for the kernel  $G$ . Section 6 illustrates the use of these conditions by showing the ergodicity of a capacity expansion model. The last two sections (7 and 8) are devoted to the proofs of Theorems 3.1 and 3.3.

**2. Definition of the PDMP and the Markov kernel  $G$ .** In this section we first present some standard notation and some basic definitions related to the motion of a PDMP  $\{X(t)\}$ . For further details the reader is referred to [6]. Afterwards we introduce the Markov kernel  $G$ , which we will use for characterizing the recurrence and the Harris recurrence structure of the PDMP  $\{X(t)\}$ . At the end of this section, we recall some classical definitions related to Markov processes both in the discrete-time and continuous-time contexts. For a complete exposition on the subject, the reader is referred to the works of Meyn and Tweedie [12, 10, 14, 13]. We follow closely the

notation in Meyn and Tweedie [12].

Let  $\mathbb{R}_+$  be the set of nonnegative real numbers. The set of natural numbers is denoted by  $\mathbb{N}$ , and  $\mathbb{N}^* \doteq \mathbb{N} - \{0\}$ . For any metric space  $H$ , the Borel  $\sigma$ -field of  $H$  is denoted by  $\mathcal{B}(H)$ . The indicator of a set  $A$  is denoted by  $\mathbf{1}_A$  ( $\mathbf{1}_A(x) = 1$  if  $x \in A$ ,  $\mathbf{1}_A(x) = 0$  if  $x \notin A$ ). Let  $E$  and  $F$  be two metric spaces. A kernel  $K$  on  $E \times \mathcal{B}(F)$  is a map  $K : E \times \mathcal{B}(F) \longrightarrow \mathbb{R}_+ \cup \{+\infty\}$  such that for  $x \in E$ ,  $K(x, \cdot)$  is a nonnegative  $\sigma$ -finite measure on  $(F, \mathcal{B}(F))$  and for any  $A \in \mathcal{B}(F)$ ,  $K(\cdot, A)$  is a measurable function on  $E$ . The kernel  $I_A$  on  $(E, \mathcal{B}(E))$  is defined for any set  $A \in \mathcal{B}(E)$  by  $I_A(x, B) = \mathbf{1}_{A \cap B}(x)$ . For two kernels  $K_1$  and  $K_2$  defined on  $E \times (\mathcal{B}(E) \otimes \mathcal{B}(\mathbb{R}_+))$ , the convolution  $K_1 * K_2$  is again a kernel on  $E \times (\mathcal{B}(E) \otimes \mathcal{B}(\mathbb{R}_+))$  defined as

$$K_1 * K_2(x, A \times \Gamma) \doteq \int_E \int_{\mathbb{R}_+} K_1(x, dy \times dt) K_2(y, A \times (\Gamma - t)),$$

where  $A \in \mathcal{B}(E)$ ,  $\Gamma \in \mathcal{B}(\mathbb{R}_+)$ , and  $\Gamma - t = \{u - t : u \in \Gamma\}$ . The notation  $K^{*n}$  represents the  $n$ -fold convolution  $K * \cdots * K$ . For a positive real-valued measurable function  $f$  defined on  $E \times \mathbb{R}_+$  the convolution  $K_1 * f$  is defined for  $x \in E$ ,  $t \in \mathbb{R}_+$  as

$$K_1 * f(x, t) \doteq \int_E \int_0^t f(y, t - s) K_1(x, dy \times ds).$$

For a substochastic kernel  $D$  we define  $U^D$  as  $U^D \doteq \sum_{k=1}^{\infty} D^k$ .

We denote by  $\gamma$  the Lebesgue measure on  $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ .

We present next the definition of the motion of a PDMP. Let  $E^0$  be an open subset of  $\mathbb{R}^n$  and let  $\partial E^0$  be its boundary. A PDMP is determined by its local characteristics  $(\mathfrak{X}, \lambda, Q)$  where the following hold:

- $\mathfrak{X}$  is a Lipschitz continuous vector field.  $\mathfrak{X} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ , which determines a flow  $\Phi(x, t)$  such that  $\frac{\partial}{\partial t} \Phi(x, t) = \mathfrak{X}(\Phi(x, t))$  and  $\Phi(x, 0) = x$  for all  $x \in \mathbb{R}^n$ .

Define

$$\Gamma^+ \doteq \{x \in \partial E^0 : x = \Phi(y, t) \text{ for some } y \in E^0, t > 0, \text{ and } \Phi(y, s) \in E^0 \forall s \in [0, t]\},$$

and

$$\Gamma^- \doteq \{x \in \partial E^0 : x = \Phi(y, -t) \text{ for some } y \in E^0, t > 0, \text{ and } \Phi(y, -s) \in E^0 \forall s \in [0, t]\}.$$

$\Gamma^+ \subset \partial E^0$  represents the boundary points at which the flow exits from  $E^0$ .  $\Gamma^- \subset \partial E^0$  is characterized by the fact that the flow starting from a point in  $\Gamma^-$  will not leave  $E^0$  immediately. Therefore it is natural to define the state space for the PDMP by  $E \doteq E^0 \cup \Gamma^- - \Gamma^- \cap \Gamma^+$ . For all  $x$  in  $E$ , let us denote  $t_*(x) \doteq \inf\{t > 0 : \Phi(x, t) \in \partial E^0\}$ , with the convention  $\inf \emptyset = \infty$ .

- The jump rate  $\lambda : E \rightarrow \mathbb{R}_+$  is assumed to be a measurable function satisfying  $(\forall x \in E) (\exists \varepsilon > 0)$  such that  $\int_0^\varepsilon \lambda(\Phi(x, s)) ds < \infty$ .

- $Q : E \cup \Gamma^+ \times \mathcal{B}(E) \rightarrow [0, 1]$  is a transition measure satisfying the following property:  $(\forall x \in E \cup \Gamma^+) Q(x, E - \{x\}) = 1$ .

From these characteristics, it can be shown [6, pp. 62–66] that there exists a filtered probability space  $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \{P_x\}_{x \in E})$  such that the motion of the process  $\{X(t)\}$  starting from a point  $x \in E$  may be constructed as follows. Take a random variable  $T_1$  such that

$$P_x(T_1 > t) \doteq \begin{cases} e^{-\Lambda(x, t)} & \text{for } t < t_*(x), \\ 0 & \text{for } t \geq t_*(x), \end{cases}$$

where for  $x \in E$  and  $t \in [0, t_*(x)[$

$$(2.1) \quad \Lambda(x, t) \doteq \int_0^t \lambda(\Phi(x, s)) ds.$$

If  $T_1$  generated according to the above probability is equal to infinity, then for  $t \in \mathbb{R}_+$ ,  $X(t) = \Phi(x, t)$ . Otherwise select independently an  $E$ -valued random variable (labeled  $X_1$ ) having distribution  $Q(\Phi(x, T_1), \cdot)$ . The trajectory of  $\{X(t)\}$  starting at  $x$ , for  $t \leq T_1$ , is given by

$$X(t) \doteq \begin{cases} \Phi(x, t) & \text{for } t < T_1, \\ X_1 & \text{for } t = T_1. \end{cases}$$

Starting from  $X(T_1) = X_1$ , we now select the next interjump time  $T_2 - T_1$  and post-jump location  $X(T_2) = X_2$  in a similar way.

This gives a strong Markov process  $\{X(t)\}$  with jump times  $\{T_k\}_{k \in \mathbb{N}}$  (where  $T_0 \doteq 0$ ). The transition semigroup of the process  $\{X(t)\}$  is denoted by  $\{P^t\}_{t \in \mathbb{R}_+}$ . We denote by  $\{\mathcal{F}_t^X\}_{t \in \mathbb{R}_+}$  the filtration generated by the process  $\{X(t)\}$ .

It is assumed throughout that for all  $(t, x) \in \mathbb{R}_+ \times E$ ,  $E_x[\sum_k \mathbf{1}_{\{T_k \leq t\}}] < \infty$ , implying in particular that  $T_k \rightarrow \infty$  as  $k \rightarrow \infty$ . This is a standard assumption; see, for example, (24.4) or (24.8) in [6].

Now let us introduce the substochastic kernels  $H$  and  $J$  and the Markov kernel  $G$ :

$$(2.2) \quad H(x, A) \doteq \int_0^{t_*(x)} e^{-\{s+\Lambda(x,s)\}} \mathbf{1}_A(\Phi(x, s)) ds,$$

$$(2.3) \quad \begin{aligned} J(x, A) &\doteq \int_0^{t_*(x)} \lambda(\Phi(x, s)) e^{-\{s+\Lambda(x,s)\}} Q(\Phi(x, s), A) ds \\ &\quad + e^{-\{t_*(x)+\Lambda(x, t_*(x))\}} Q(\Phi(x, t_*(x)), A), \end{aligned}$$

$$(2.4) \quad G(x, A) \doteq J(x, A) + H(x, A).$$

In [8], it was shown that  $G$  as defined in (2.4) is a Markov kernel.

The resolvent kernel associated to the process  $\{X(t)\}_{t \in \mathbb{R}_+}$  is denoted by

$$(2.5) \quad R(x, A) \doteq \int_0^\infty P^t(x, A) e^{-t} dt.$$

As shown in [8],  $R$  can be written in terms of  $H$  and  $J$  as follows:

$$(2.6) \quad R = \sum_{j=0}^\infty J^j H.$$

Let  $\{\Theta_n\}$  (respectively,  $\{\Upsilon_n\}$ ) be the Markov chain associated to the Markov kernel  $G$  (respectively,  $R$ ). In Theorem 3.1 below it will be shown how the Markov chain  $\{\Theta_n\}$  can be generated from the sample paths of the PDMP  $\{X(t)\}$ .

In what follows we will present some definitions considering a discrete-time Markov chain  $\{\chi_n\}$  with Markov kernel  $S$  that could be either  $\{\Theta_n\}$  (with  $S = G$ ) or  $\{\Upsilon_n\}$  (with  $S = R$ ). The first return time of a set  $A \in \mathcal{B}(E)$  for the PDMP  $\{X(t)\}$  and for the Markov chain  $\{\chi_n\}$  are defined, respectively, as follows:

$$\tau_A^X \doteq \inf\{t > 0 : X(t) \in A\}, \quad \tau_A^S \doteq \inf\{n \geq 1 : \chi_n \in A\}.$$

Associated to these first return times are the return time probability of a set  $A \in \mathcal{B}(E)$  for the PDMP  $\{X(t)\}$  and for the Markov chain  $\{\chi_n\}$ , given, respectively, by

$$L^X(x, A) \doteq P_x(\tau_A^X < \infty), \quad L^S(x, A) \doteq P_x(\tau_A^S < \infty).$$

The number of visits to a set  $A$  is defined for the PMDP  $\{X(t)\}$  and for the Markov chain  $\{\chi_n\}$ , respectively, as

$$\eta_A^X \doteq \int_0^\infty \mathbf{1}_A(X(t))dt, \quad \eta_A^S \doteq \sum_{n=1}^\infty \mathbf{1}_A(\chi_n).$$

If  $F$  is a probability distribution on  $\mathbb{R}_+$  (respectively,  $b$  is a probability on  $\mathbb{N}^*$ ), then the stochastic kernel  $K_F^X$  (respectively,  $K_b^S$ ) associated to  $\{X(t)\}$  (respectively,  $\{\chi_n\}$ ) is defined on  $E \times \mathcal{B}(E)$  by

$$(2.7) \quad (\forall x \in E) \quad (\forall A \in \mathcal{B}(E)) \quad K_F^X(x, A) \doteq \int_0^\infty P^t(x, A)F(dt),$$

$$(2.8) \quad (\forall x \in E) \quad (\forall A \in \mathcal{B}(E)) \quad K_b^S(x, A) \doteq \sum_{k=0}^\infty b(k)S^k(x, A).$$

A set  $C \in \mathcal{B}(E)$  is called a petite set for  $\{X(t)\}$  (respectively,  $\{\chi_n\}$ ) if there exist a probability distribution  $F$  on  $\mathbb{R}_+$  (respectively, a probability  $b$  on  $\mathbb{N}^*$ ) and a non-trivial measure  $\nu$  on  $(E, \mathcal{B}(E))$  such that  $(\forall A \in \mathcal{B}(E)) \quad (\forall x \in C) \quad K_F^X(x, A) \geq \nu(A)$  (respectively,  $(\forall A \in \mathcal{B}(E)) \quad (\forall x \in C) \quad K_b^S(x, A) \geq \nu(A)$ ).

A positive measure  $\mu$  (respectively,  $\pi$ ) is called an invariant for the PMDP  $\{X(t)\}$  (respectively, for the Markov chain  $\{\chi_n\}$ ) if it is a  $\sigma$ -finite measure satisfying  $\mu = \mu P^t$  for all  $t \geq 0$  (respectively,  $\pi = \pi S$ ). The PDMP  $\{X(t)\}$  is said to be ergodic if it has an invariant probability measure  $\mu$  such that

$$(\forall x \in E) \quad \lim_{t \rightarrow \infty} \|P^t(x, \cdot) - \mu(\cdot)\| = 0,$$

where  $\|\cdot\|$  denotes the total variation norm.

The following definitions apply for both the continuous-time as well as the discrete-time processes, and therefore we suppress the superscript  $X$  or  $S$ . A Markov process is called  $\varphi$ -irreducible (and  $\varphi$  is an irreducibility measure) if for some  $\sigma$ -finite measure  $\varphi$  we have that  $E_x(\eta_A) > 0$  for all  $x \in E$  whenever  $\varphi(A) > 0$ . A set  $A \in \mathcal{B}(E)$  is said to be full if  $\varphi(A^c) = 0$ . An irreducibility measure  $\psi$  is called maximal irreducible if for any other  $\varphi$  irreducibility measure we have that  $\psi \gg \varphi$ . A Markov process is called recurrent if for some  $\sigma$ -finite measure  $\varphi$  we have that  $E_x(\eta_A) = \infty$  for all  $x \in E$  whenever  $\varphi(A) > 0$ , and is called Harris recurrent if  $E_x(\eta_A) = \infty$  is replaced by  $P_x(\eta_A = \infty) = 1$ . If the Markov process is Harris recurrent, then there exists a unique (up to constant multiples) invariant measure. The Markov process is said to be positive Harris recurrent if it is Harris recurrent and the invariant measure is finite.

**3. Preliminary results.** In this section we present some preliminary results that will be very important in characterizing the recurrence and Harris recurrence structure of the PDMP  $\{X(t)\}$ . First, in Theorem 3.1 the Markov chain  $\{\Theta_n\}$  generated by the kernel  $G$  is shown to be related to the sample path of the PDMP. Since the proof of Theorem 3.1 is rather long and technical, it is presented in an independent section near the end of the paper (see section 7). It is interesting to note that  $\{\Theta_n\}$  corresponds to a sampling of the continuous-time process  $\{X(t)\}$  at random



times that depends on a combination of a sequence of independent and identically distributed exponential times with the sequence  $\{T_k\}$  of jump times of the PDMP  $\{X(t)\}$ . Moreover it must be pointed out that the Markov kernel  $G$  does not correspond to a generalized resolvent, as studied in the fundamental paper of Meyn and Tweedie [11]. An easy consequence of Theorem 3.1 presented in Corollary 3.2 is that if the first return time of the Markov chain  $\{\Theta_n\}$  to a set  $A$  is finite, then the first return time of the PDMP  $\{X(t)\}$  to the same set  $A$  is finite. Consequently, it will be easy to deduce from this result that if the Markov chain  $\{\Theta_n\}$  is Harris recurrent, then so is the process  $\{X(t)\}$ . The last two theorems of this section show that

- the probability of the first return time of  $\{\Theta_n\}$  to a set  $A$  to be finite is bounded below by the probability of the first return time of  $\{\Upsilon_n\}$  to the same set  $A$  to be finite (see Theorem 3.3);
- the average number of visits of  $\{\Theta_n\}$  to a set  $A$  is bounded below by the average number of visits of the Markov chain  $\{\Upsilon_n\}$  (generated by the resolvent) to the same set  $A$  (see Theorem 3.4).

Theorem 3.4 will be used in the next section to show that if the process  $\{X(t)\}$  is recurrent, then so is the Markov chain  $\{\Theta_n\}$ . An important consequence of Theorem 3.3 is that if the process  $\{X(t)\}$  is Harris recurrent, then so is the Markov chain  $\{\Theta_n\}$ . Theorem 3.3 is surprising and far from trivial to show. The last section of the paper (see section 8) is devoted to its proof.

We have the following result, which shows how the Markov chain  $\{\Theta_n\}$  could be generated from the sample path realizations of  $\{X(t)\}$ .

**THEOREM 3.1.** *On the probability space  $(\Omega, \mathcal{F}_t^X, \mathcal{F}, P_x)$  let  $\{s_n\}_{n \geq 0}$  be a sequence of independent and identically distributed  $\mathbb{R}_+$ -valued random variables with exponential distribution with parameter equal to 1 such that  $\forall_{t \geq 0} \mathcal{F}_t^X$  and  $\sigma\{s_k : k \geq 0\}$  are independent. Let the sequence of stopping times  $\{\tau_n\}_{n \in \mathbb{N}}$  be defined as follows:  $\tau_0 = 0$  and*

$$(3.1) \quad \tau_{n+1} \doteq \sum_{k=0}^n \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} \left[ (\tau_n + s_{n+1}) \wedge T_{k+1} \right].$$

*Then  $\{X(\tau_n)\}$  is a Markov chain with transition probability given by  $G$ .*

The proof of this result is presented in section 7.

**Remark 3.1.** Roughly speaking the sequence  $\{\tau_n\}_{n \in \mathbb{N}}$  is defined as

$$\begin{aligned} \tau_1 &= \mathbf{1}_{\{T_0 \leq \tau_0 < T_1\}} \left[ (\tau_0 + s_1) \wedge T_1 \right] \\ &= \begin{cases} s_1 & \text{if } s_1 < T_1, \\ T_1 & \text{otherwise,} \end{cases} \end{aligned}$$

$$\begin{aligned} \tau_2 &= \mathbf{1}_{\{T_0 \leq \tau_1 < T_1\}} \left[ (\tau_1 + s_2) \wedge T_1 \right] + \mathbf{1}_{\{T_1 \leq \tau_1 < T_2\}} \left[ (\tau_1 + s_2) \wedge T_2 \right] \\ &= \begin{cases} \tau_1 + s_2 & \text{if } \tau_1 < T_1 \text{ (equivalently } s_1 < T_1), \text{ and } \tau_1 + s_2 < T_1, \\ T_1 & \text{if } \tau_1 < T_1 \text{ (equivalently } s_1 < T_1), \text{ and } \tau_1 + s_2 \geq T_1, \\ T_1 + s_2 & \text{if } \tau_1 = T_1 \text{ (equivalently } s_1 \geq T_1), \text{ and } T_1 + s_2 < T_2, \\ T_2 & \text{otherwise,} \end{cases} \end{aligned}$$

and so on.

Without loss of generality, it can be considered that  $\Theta_n = X(\tau_n)$ , since  $\{\Theta_n\}$  was defined in section 2 as a Markov chain generated by  $G$ , and, from the previous theorem, that  $\{\Theta_n\}$  and  $\{X(\tau_n)\}$  have the same probability distribution. An important corollary of the previous theorem is the following inclusion for the first return times of the Markov chain  $\{\Theta_n\}$  and the process  $\{X(t)\}$ .

COROLLARY 3.2. *For any set  $A \in \mathcal{B}(E)$ ,*

$$(3.2) \quad \{\tau_A^G < \infty\} \subset \{\tau_A^X < \infty\}.$$

*Proof.* This is a straightforward consequence from the fact that we can consider  $\Theta_n = X(\tau_n)$ , as shown in Theorem 3.1.  $\square$

We have the following important theorem establishing a link between the probability of the first return time to a set for the Markov chains  $\{\Theta_n\}$  and  $\{\Upsilon_n\}$ .

THEOREM 3.3. *For every  $x \in E$  and for  $A \in \mathcal{B}(E)$ ,*

$$(3.3) \quad L^G(x, A) \geq L^R(x, A).$$

The proof of this result is presented in section 8.

Combining (3.2) and (3.3) we have, for every  $x \in E$  and  $A \in \mathcal{B}(E)$ , the following important inequalities:

$$L^R(x, A) \leq L^G(x, A) \leq L^X(x, A).$$

We conclude this section with the following theorem, providing a link between the average numbers of visits for the Markov chains generated by the kernel  $G$  and  $R$ .

THEOREM 3.4. *For every  $x \in E$  and for  $A \in \mathcal{B}(E)$ ,*

$$(3.4) \quad U^G(x, A) \geq U^R(x, A).$$

*Proof.* From Lemma 4.1 in [17], we have  $G^n = (J+H)^n = J^n + \sum_{j=1}^n J^{j-1} H G^{n-j}$  for any  $n \in \mathbb{N}^*$ . Consequently, it follows that

$$\begin{aligned} U^G &= U^J + \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} J^{j-1} H G^{n-j} \mathbf{1}_{\{j \leq n\}} \\ &= U^J + R + R U^G, \end{aligned}$$

where the last equality follows from (2.6). Hence, by iteration on  $n$  in the previous equation, we get that for all  $n \geq 1$ ,

$$U^G = U^J + \sum_{j=1}^{n-1} R^j U^J + \sum_{j=1}^n R^j + R^n U^G \geq \sum_{j=1}^n R^j.$$

Taking the limit as  $n$  goes to  $\infty$ , we get the result.  $\square$

**4. Characterization of the recurrence and Harris recurrence structures of the PDMP in terms of the Markov kernel  $G$ .** The aim of this section is to characterize the (Harris) recurrence properties between the PMDP  $\{X(t)\}$  and the Markov chain  $\{\Theta_n\}$  generated by the kernel  $G$ . First, it is proved in Proposition 4.1 that  $\{X(t)\}$  is irreducible if and only if  $\{\Theta_n\}$  is irreducible. Then a generalization of Theorem 3.5 in [8] is presented in Theorem 4.2 giving a one-to-one correspondence between the invariant (positive and  $\sigma$ -finite) measures for the PDMP  $\{X(t)\}$  and the

Markov chain  $\{\Theta_n\}$ . Using the preliminary results derived in the previous section, it is shown in Theorems 4.4 and 4.6 that the PDMP  $\{X(t)\}$  is recurrent (respectively, Harris recurrent) if and only if the Markov chain  $\{\Theta_n\}$  is recurrent (respectively, Harris recurrent). One would expect a natural generalization of such equivalence results for positivity between the processes  $\{X(t)\}$  and  $\{\Theta_n\}$ . In fact, this result does not hold. Indeed, it is shown in Corollary 4.5 that the positive recurrence of the process  $\{X(t)\}$  is equivalent to a weaker form of stability for the Markov chain  $\{\Theta_n\}$ . Namely,  $\{X(t)\}$  is positive recurrent if and only if  $\{\Theta_n\}$  is recurrent and its unique invariant measure  $\pi$  satisfies the condition given by  $\pi H(E) < \infty$ , which is far less demanding than positive recurrence for  $\{\Theta_n\}$ . A similar result will be proved for the positive Harris recurrence of  $\{X(t)\}$  (see Corollary 4.7).

We have the following proposition characterizing the irreducibility of the process  $\{X(t)\}$  and the Markov chain  $\{\Theta_n\}$ .

**PROPOSITION 4.1.** *The PDMP  $\{X(t)\}$  is irreducible if and only if the Markov chain  $\{\Theta_n\}$  is irreducible.*

*Proof.* Suppose that the Markov chain  $\{\Theta_n\}$  is  $\varphi$ -irreducible. From Proposition 4.2.1 in [12, p. 87], whenever  $\varphi(A) > 0$  for any  $A \in \mathcal{B}(E)$  we have that  $L^G(x, A) > 0$  for all  $x \in E$ . From (3.2) we have that  $L^X(x, A) > 0$  for all  $x \in E$  whenever  $\varphi(A) > 0$  for  $A \in \mathcal{B}(E)$ . By using Proposition 2.1 in [13], it follows that the PDMP  $\{X(t)\}$  is  $\mu$ -irreducible with  $\mu = \varphi R$ , where we recall that  $R$  is the resolvent defined in (2.5).

Now suppose that  $\{X(t)\}$  is  $\Psi$ -irreducible. Then for  $A \in \mathcal{B}(E)$  with  $\Psi(A) > 0$ , we have, for all  $x \in E$ ,  $E_x[\eta_A^X] > 0$ . Since  $E_x[\eta_A^X] = U^R(x, A)$  (see, for instance, [19]), it implies that  $L^R(x, A) > 0$ . From Theorem 3.3, we get the result.  $\square$

Recall that by definition an invariant measure is always  $\sigma$ -finite and positive. The next result shows that there exists a one-to-one correspondence between the set of invariant measures for the PDMP  $\{X(t)\}$  and the set of invariant measures for the Markov chain  $\{\Theta_n\}$  generated by  $G$ . It extends Theorem 3.5 in [8], which was restricted to the set of invariant probability measures for the PDMPs.

**THEOREM 4.2.** (i) *If  $\mu$  is an invariant measure for  $\{X(t)\}$ , then  $\mu \sum_{j=0}^{\infty} J^j$  is invariant for  $\{\Theta_n\}$  and  $\mu \sum_{j=0}^{\infty} J^j H = \mu$ .*

(ii) *If  $\pi$  is an invariant measure for  $\{\Theta_n\}$ , then  $\pi H$  is invariant for  $\{X(t)\}$  and  $\pi H \sum_{j=0}^{\infty} J^j = \pi$ .*

*Proof.* Let us show (i). Let  $\mu$  be an invariant measure for  $\{X(t)\}$  and set  $\pi = \sum_{j=0}^{\infty} \mu J^j$ . Let us show that  $\pi$  is  $\sigma$ -finite. Since  $\mu$  is  $\sigma$ -finite, there exists a partition  $\{A_i\}$  of  $E$  such that  $\mu(A_i) < \infty$ . Define  $C_n = \bigcup_{i=1}^n A_i$  and  $B_{n,m} = \{y \in E : H(y, C_n) > \frac{1}{m}\}$  for  $m \in \mathbb{N}^*$ . Notice now that for every  $x \in E$ , we have that  $0 < H(x, E) < 1$ , and so  $\bigcup_{n,m} B_{n,m} = E$ . From (2.6) we have that  $\mu = \mu R = \sum_{j=0}^{\infty} \mu J^j H = \pi H$ , so that

$$\infty > \mu(C_n) = \int_E H(y, C_n) \pi(dy) \geq \int_{B_{n,m}} H(y, C_n) \pi(dy) \geq \frac{1}{m} \pi(B_{n,m}),$$

showing that  $\pi$  is  $\sigma$ -finite. Finally notice from (2.4) and (2.6) that

$$\pi G = \pi J + \pi H = \sum_{j=1}^{\infty} \mu J^j + \mu R = \sum_{j=1}^{\infty} \mu J^j + \mu = \sum_{j=0}^{\infty} \mu J^j = \pi,$$

showing that  $\pi$  is invariant for  $\{\Theta_n\}$ . Moreover,

$$\mu \sum_{j=0}^{\infty} J^j H = \mu R = \mu,$$

completing the proof of (i).

Let us now show (ii). Let  $\pi$  be an invariant measure for  $\{\Theta_n\}$ . For any  $n \in \mathbb{N}^*$ , we have

$$\pi \sum_{j=1}^n J^j H + \pi H = \pi G \sum_{j=0}^n J^j H = \pi \sum_{j=1}^n J^j H + \pi J^{n+1} H + \pi H \sum_{j=0}^n J^j H.$$

In order to cancel out the identical term  $\pi \sum_{j=1}^n J^j H$  on both sides of the previous equation, one must first check that all the measures under consideration are  $\sigma$ -finite. Since  $\pi = \pi G = \pi H + \pi J$ , it can be shown easily by induction that  $\pi J^j H \leq \pi$  and so  $\pi J^j H$  is  $\sigma$ -finite for all  $j \in \mathbb{N}$ . Consequently, for all  $j \in \mathbb{N}^*$ , the measures  $\pi \sum_{j=1}^n J^j H$ ,  $\pi H$ ,  $\pi J^{n+1} H$ , and  $\pi H \sum_{j=0}^n J^j H$  are  $\sigma$ -finite, implying that

$$(4.1) \quad \pi H = \pi J^{n+1} H + \pi H \sum_{j=0}^n J^j H.$$

Moreover, it can be shown that  $J^n(x, A) = E_x[e^{-T_n} \mathbf{1}_A[X(T_n)]]$  for all  $n \in \mathbb{N}^*$ ,  $x \in E$ , and  $A \in \mathcal{B}(E)$ . By using the dominated convergence theorem and the fact that  $\lim_{n \rightarrow \infty} T_n = +\infty$ , it follows that for all  $A \in \mathcal{B}(E)$

$$(4.2) \quad \lim_{n \rightarrow \infty} \pi J^n(A) = 0.$$

Combining (4.1) and (4.2), we have that  $\mu = \pi H = \pi H \sum_{j=0}^{\infty} J^j H = \mu R$ , and from Lemma 1 in [1] it follows that  $\mu$  is an invariant measure for  $\{X(t)\}$ . Now we have that

$$\pi H \sum_{j=0}^n J^j + \pi J^{n+1} + \pi \sum_{j=1}^n J^j = \pi + \pi \sum_{j=1}^n J^j.$$

It follows that  $\pi H \sum_{j=0}^n J^j + \pi J^{n+1} = \pi$  by using the same arguments as above. Equation (4.2) yields that  $\lim_{n \rightarrow \infty} \pi H \sum_{j=0}^n J^j = \pi$ , showing (ii).  $\square$

*Remark 4.1.* A straightforward consequence of Theorem 4.2 is the following result: There exists a *finite* invariant measure for  $\{X(t)\}$  if and only if there exists an invariant measure  $\pi$  for  $\{\Theta_n\}$  satisfying  $\pi H(E) < \infty$ . Note that this result was already proved in Theorem 3.5 in [8].

The next two results show that if the PDMP is recurrent, then so is the Markov chain generated by  $G$ , and vice versa.

**PROPOSITION 4.3.** *If  $H \in \mathcal{B}(E)$  is recurrent for the process  $\{X(t)\}$ , then  $H$  is recurrent for the Markov chain  $\{\Theta_n\}$ .*

*Proof.* If  $H \in \mathcal{B}(E)$  is recurrent for  $\{X(t)\}$ , then there exists a measure  $\nu$  on  $(H, \mathcal{B}(H))$  such that for all  $A \in \mathcal{B}(H)$  with  $\nu(A) > 0$ ,  $E_x[\eta_A^X] = U^R(x, A) = \infty$  for every  $x \in H$ . From Theorem 3.4, it follows that  $U^G(x, A) = \infty$ , showing the result.  $\square$

**THEOREM 4.4.** *The PDMP  $\{X(t)\}$  is recurrent if and only if the Markov chain  $\{\Theta_n\}$  associated to  $G$  is recurrent.*

*Proof.* Suppose that  $\{X(t)\}$  is transient and  $\{\Theta_n\}$  is recurrent. Let  $\varphi$  be a maximal irreducibility measure for  $\{\Theta_n\}$ . Then from Proposition 9.0.1 in [12],  $E = H \cup T$ , where  $T \in \mathcal{B}(E)$  is  $\varphi$ -null and transient for  $\{\Theta_n\}$ ,  $H \in \mathcal{B}(E)$  is nonempty and absorbing for  $\{\Theta_n\}$ , and every subset of  $H$  in  $\mathcal{B}(E)^+ \doteq \{A \in \mathcal{B}(E) : \varphi(A) > 0\}$

is Harris recurrent. On the other hand, since  $\{X(t)\}$  is transient, it follows that  $E = \cup_{i=1}^{\infty} E_i$  with  $E_x[\eta_{E_i}^X] \leq M_i < \infty$  for every  $x \in E$ . Define  $A_n \doteq \cup_{i=1}^n E_i \cap H$ . Consider  $k \in \mathbb{N}^*$  such that  $\varphi(A_k) > 0$ . For every  $x \in E$ , we have

$$E_x[\eta_{A_k}^X] \leq \sum_{i=1}^k E_x[\eta_{E_i}^X] \leq M,$$

with  $M \doteq \sum_{i=1}^k M_i$ . However, since  $A_k$  is a subset of  $H$  and belongs to  $\mathcal{B}(E)^+$ , the set  $A_k$  is Harris recurrent for  $\{\Theta_n\}$ . Consequently, by using (3.2), we have that for every  $x \in A_k$ ,  $1 = L^G(x, A) \leq L^X(x, A)$ , contradicting the fact that  $E_x[\eta_{A_k}^X] \leq M$ . Therefore, if  $\{\Theta_n\}$  is recurrent, then  $\{X(t)\}$  is recurrent.

The converse follows from Proposition 4.3, giving the result.  $\square$

The next corollary emphasizes a split with the previous equivalence results. Indeed it is shown that the process is positive recurrent if and only if  $\{\Theta_n\}$  satisfies a weaker condition: recurrence and a technical condition for its unique ( $\sigma$ -finite) invariant measure.

**COROLLARY 4.5.** *The PDMP  $\{X(t)\}$  is positive recurrent if and only if the Markov chain  $\{\Theta_n\}$  associated to  $G$  is recurrent with invariant measure  $\pi$  satisfying  $\pi H(E) < \infty$ .*

*Proof.* The result easily follows from Remark 4.1 and Theorem 4.4.  $\square$

We prove now that the Harris recurrent properties are equivalent for  $\{X(t)\}$  and  $\{\Theta_n\}$ .

**THEOREM 4.6.** *The PDMP  $\{X(t)\}$  is Harris recurrent if and only if the Markov chain  $\{\Theta_n\}$  is Harris recurrent.*

*Proof.* Suppose that the Markov chain  $\{\Theta_n\}$  is Harris recurrent. Denote by  $\Psi^G$  a maximal irreducibility measure for the Markov chain  $\{\Theta_n\}$ . Then for any set  $A \in \mathcal{B}(E)$  satisfying  $\Psi^G(A) > 0$  it follows from Corollary 3.2 that  $1 = L^G(x, A) \leq L^X(x, A)$  for all  $x \in E$ . Therefore,  $\{X(t)\}$  is Harris recurrent by using Theorem 1.1 in [11].

Now assume that the PDMP  $\{X(t)\}$  is Harris recurrent. From the equivalence results in [19], if the PDMP  $\{X(t)\}$  is Harris recurrent, then the Markov chain  $\{\Upsilon_n\}$  associated to the resolvent  $R$  is Harris recurrent. Moreover, by using Proposition 4.1  $\{\Theta_n\}$  is irreducible. Let us denote by  $\Psi^G$  (respectively,  $\Psi^R$ ) a maximal irreducible measure for  $\{\Theta_n\}$  (respectively,  $\{\Upsilon_n\}$ ). According to the definition of Harris recurrence (see [12, p. 200]), we want to show that if  $\Psi^G(A) > 0$ , then  $P_x(\cap_{j=1}^{\infty} \cup_{n=j}^{\infty} \{\Theta_n \in A\}) = 1$  for all  $x \in E$ . From (ii) and (iii) of Proposition 5.5.5 in [12], it follows that there exists an increasing sequence of petite sets  $\{C_k\}_{k \in \mathbb{N}}$  for  $\{\Theta_n\}$  such that  $E = \cup_{k \in \mathbb{N}} C_k$  with  $\Psi^G(C_k) > 0$  and  $\Psi^R(C_k) > 0$  for all  $k \in \mathbb{N}$ . Since  $\{\Upsilon_n\}$  is Harris recurrent we have for all  $k \in \mathbb{N}$  that  $L^R(x, C_k) = 1$  for all  $x \in C_k$ . From Theorem 3.3 we have that  $L^G(x, C_k) = 1$  for all  $x \in C_k$ , and from Proposition 9.1.1 in [12], it follows that  $C_k$  is Harris recurrent for  $\{\Theta_n\}$ . The remainder of the proof now follows the same steps as the end of the proof of Theorem 9.1.4 in [12], and it will be presented for the sake of completeness. From Lemma 5.5.1 in [12], we have that for all  $A \in \mathcal{B}(E)$  with  $\Psi^G(A) > 0$  there exists  $\delta > 0$  such that  $\inf_{x \in C_k} L^G(x, A) > \delta$ . However,  $C_k$  is Harris recurrent for  $\{\Theta_n\}$ , and from Theorem 9.1.3(i) in [12], we have that for all  $x \in C_k$ ,  $P_x(\cap_{j=1}^{\infty} \cup_{n=j}^{\infty} \{\Theta_n \in A\}) = 1$ . The result follows after recalling that  $E = \cup_{k \in \mathbb{N}} C_k$ .  $\square$

**Remark 4.2.** In the previous proof note that if  $A$  is a set such that  $\psi^G(A) > 0$ , then it does not necessarily imply that  $\psi^R(A) > 0$ . That is why we needed to proceed

through the tool of petite sets.

As for the positive recurrence property, the following result points out the split with the previous theorem by showing that the positive Harris recurrence is equivalent to a weaker form of stability for the chain  $\{\Theta_n\}$ .

**COROLLARY 4.7.** *The PDMP  $\{X(t)\}$  is positive Harris recurrent if and only if the Markov chain  $\{\Theta_n\}$  associated to  $G$  is Harris recurrent with invariant measure  $\pi$  satisfying  $\pi H(E) < \infty$ .*

*Proof.* Combining Remark 4.1 and Theorem 4.6, we obtain the result.  $\square$

The final result of this section gives a condition for the PDMP to be ergodic. As shown in [13, Theorem 6.1], if  $\{X(t)\}$  is positive Harris recurrent, then it is ergodic if and only if some skeleton chain is irreducible. Thus from Corollary 4.7 it remains only to obtain a tractable and equivalent condition to show that some skeleton chain is irreducible. Let  $N$  be the kernel defined on  $E \times \mathcal{B}(E) \otimes \mathcal{B}(\mathbb{R}_+)$  by

$$(4.3) \quad \begin{aligned} N(x, A \times \Gamma) &= \int_0^{t_*(x)} \mathbf{1}_\Gamma(s) \lambda(\Phi(x, s)) e^{-\Lambda(x, s)} Q(\Phi(x, s), A) ds \\ &+ \mathbf{1}_\Gamma(t_*(x)) e^{-\Lambda(x, t_*(x))} Q(\Phi(x, t_*(x)), A), \end{aligned}$$

where  $A \in \mathcal{B}(E)$ ,  $\Gamma \in \mathcal{B}(\mathbb{R}_+)$ . Define  $M : E \times \mathcal{B}(E) \times \mathbb{R}_+ \rightarrow \mathbb{R}$  by

$$(4.4) \quad M^t(x, A) \doteq \mathbf{1}_{\{t < t_*(x)\}} \mathbf{1}_A(\Phi(x, t)) e^{-\Lambda(x, t)}$$

for  $A \in \mathcal{B}(E)$ ,  $t \in \mathbb{R}_+$ , and  $x \in E$ . For  $x \in E$  and  $k \in \mathbb{N}$ , define the measure  $\alpha^k(x, \cdot)$  on  $(E \times \mathbb{R}_+, \mathcal{B}(E) \otimes \mathcal{B}(\mathbb{R}_+))$  by

$$\alpha^k(x, B) \doteq \int_B N^{*k} * M^t(x, dy) \gamma(dt)$$

for  $B \in \mathcal{B}(E) \otimes \mathcal{B}(\mathbb{R}_+)$ .

**THEOREM 4.8.** *Suppose that the Markov chain  $\{\Theta_n\}$  is positive Harris recurrent and denote by  $\Psi^G$  a maximal irreducibility measure for  $\{\Theta_n\}$ . The following propositions are equivalent:*

- (i) *There exist a set  $C \in \mathcal{B}(E)$  with  $\Psi^G H(C) > 0$  such that for all  $x \in C$  there exists  $k \in \mathbb{N}$  for which the measure  $\alpha^k(x, \cdot)$  is nonsingular with respect to the measure  $\Psi^G H \otimes \gamma$ .*
- (ii) *The PDMP  $\{X(t)\}$  is ergodic.*

*Proof.* From Lemma (27.3) in [6], it is easy to obtain that for all  $x \in E$  and for  $A \in \mathcal{B}(E)$ ,

$$(4.5) \quad P^t(x, A) = \sum_{k=0}^{\infty} N^{*k} * M^t(x, A).$$

Consequently, for any  $B \in \mathcal{B}(E) \otimes \mathcal{B}(\mathbb{R}_+)$ , we have

$$(4.6) \quad \int_B P^t(x, dy) \gamma(dt) = \sum_{k=0}^{\infty} \alpha^k(x, B).$$

From Corollary 4.7, the PDMP  $\{X(t)\}$  is positive Harris recurrent. Denote by  $\mu$  the invariant probability measure for  $\{X(t)\}$  and by  $\Psi^X$  a maximal irreducibility measure for  $\{X(t)\}$ . From Proposition 4.1 there exists a positive  $\sigma$ -finite measure

(labeled  $\pi$ ) invariant for  $G$  satisfying  $\pi H = \mu$ . By hypothesis, we also have that  $G$  is irreducible. From Theorem 10.4.9 in [12] we obtain that  $\mu \sim \Psi^X$  and  $\pi \sim \Psi^G$ , and thus  $\Psi^G H \sim \Psi^X$ .

According to Proposition 6.2 of Niemi and Nummelin [16] and by using (4.6), a skeleton chain is irreducible if and only if there exists a set  $C$  satisfying  $\Psi^X(C) > 0$  and such that for all  $x \in C$  there exists an integer  $k$  for which the measure  $\alpha^k(x, \cdot)$  is nonsingular with respect to the measure  $\Psi^X \otimes \gamma$ . However, from Theorem 6.1 in [13] and by using the fact that  $\Psi^G H \sim \Psi^X$ , we have that the PDMP  $\{X(t)\}$  is ergodic if and only if item (i) is satisfied, showing the result.  $\square$

**5. Sufficient conditions for stability and ergodicity of PDMPs.** Based on the results derived in the previous section we now present some sufficient conditions for the existence of an invariant probability measure, positive Harris recurrence, and ergodicity for the PDMP  $\{X(t)\}$ . These various stability conditions are inspired by the results obtained in [12, 15] and are based on modified Foster–Lyapunov criteria through the Markov kernel  $G$ . As mentioned before, this will lead to tractable criteria since the kernel  $G$  can be explicitly characterized in terms of the three local characteristics of the PDMP. It must be pointed out that the results derived in [12, 15] are not directly applicable to our case mainly because we need to obtain a criterion ensuring the existence of  $\sigma$ -finite invariant measure for  $G$  satisfying  $\pi H(E) < \infty$ .

We present next a Foster–Lyapunov criterion, based on the stochastic kernel  $G$ , for obtaining an invariant probability measure  $\mu$  for the PDMP  $\{X(t)\}$  satisfying  $\int_E f(x)\mu(dx) < \infty$ , for any measurable function  $f \geq 1$ . This result generalizes Corollary 4.5 in [8] by relaxing the hypotheses on the Lyapunov function  $V$  and the test set  $D$ , and allowing us to consider the measurable function  $f \geq 1$ .

**PROPOSITION 5.1.** *Let  $f : E \rightarrow [1, \infty)$  be a measurable function. Suppose that the Markov kernel  $G$  is recurrent and that the following Foster–Lyapunov criterion is satisfied:*

$$(5.1) \quad (\forall x \in E) \quad GV(x) \leq V(x) - Hf(x) + bI_D(x),$$

where  $D$  is a petite set for the Markov chain associated to  $G$ , and  $V : E \rightarrow [0, \infty]$  is a measurable function (with  $V(x_0) < +\infty$  for at least one  $x_0 \in E$ ). Then there exists a unique invariant probability measure  $\mu$  for the PDMP  $\{X(t)\}$  and moreover  $\int_E f(x)\mu(dx) < \infty$ .

*Proof.* Let  $\Psi^G$  be a maximal irreducibility measure for  $G$ . From Proposition 5.5.5 in [12], there exists a sequence of petite sets, say  $\{C_n\}$ , such that  $C_n \uparrow E$ . Define  $B_n = C_n \cap \{x \in E : V(x) \leq n\}$ . Notice that the set  $A = \{x \in E : V(x) < +\infty\}$  is absorbing (indeed, from (5.1), if  $x \in A$ , then  $GV(x) < \infty$  and thus  $G(x, A) = 1$ ), and thus from Proposition 4.2.3 in [12], it is full. It follows that there exists  $k \in \mathbb{N}$  such that  $\Psi^G(B_k) > 0$ . Clearly,  $B_k$  is a petite set for  $G$ , and  $V$  is bounded in  $B_k$ . Now applying Theorem 14.2.3 in [12] we obtain that  $\sup_{x \in B_k} E_x \left[ \sum_{j=0}^{\tau_{B_k}-1} Hf(\Theta_j) \right] < +\infty$ . By using the same arguments as in Theorem 4.4 of [8], it follows that there exists a unique invariant measure  $\pi$  of  $G$  which, moreover, satisfies  $\int_E f(x)\pi H(dx) < \infty$ . Consequently, it follows from Theorem 4.2 and Remark 4.1 that the measure  $\mu$  defined by  $\pi H$  is finite, invariant for the PDMP, and satisfies  $\int_E f(x)\mu(dx) < \infty$ . We get the uniqueness by using Proposition 4.1 since  $G$  is irreducible, showing the result.  $\square$

Next we present a theorem that requires only  $\varphi$ -irreducibility of the Markov kernel  $G$ , replacing the recurrence condition by another Lyapunov criterion. Moreover this result shows that the PDMP  $\{X(t)\}$  is positive Harris recurrent.

THEOREM 5.2. *Let  $f : E \rightarrow [1, \infty)$  be a measurable function. Suppose that the Markov kernel  $G$  is irreducible and satisfies the following Lyapunov criterion:*

$$(5.2) \quad (\forall x \in C^c) \quad GW(x) \leq W(x),$$

$$(5.3) \quad (\forall x \in E) \quad GV(x) \leq V(x) - Hf(x) + bI_D(x),$$

where  $C$  and  $D$  are petite sets for the Markov chain associated to  $G$ ,  $V : E \rightarrow [0, \infty]$  is a measurable function (with  $V(x_0) < +\infty$  for at least one  $x_0 \in E$ ), and  $W$  is a function unbounded off petite sets. Then the PDMP  $\{X(t)\}$  is positive Harris recurrent and its unique invariant probability measure  $\mu$  satisfies  $\int_E f(x)\mu(dx) < \infty$ .

*Proof.* From Theorem 9.1.8 in [12], inequality (5.2) implies that  $G$  is Harris recurrent, showing that the PDMP is Harris recurrent from Theorem 4.6. From Proposition 5.1 we get the result.  $\square$

Next we present a state-dependent criterion based on Theorem 2.2 of [15]. For this consider a random variable  $\varsigma_x$  on  $\mathbb{N}^*$  for each  $x \in E$ , independent of  $\{\Theta_k\}$ , and with distribution  $a_x(k)$ ,  $k \in \mathbb{N}^*$ . Recall that  $K_{a_x}^G(x, A) = \sum_{k=1}^{\infty} a_x(k)G^k(x, A)$  for any  $A \in \mathcal{B}(E)$ . Suppose that  $\sum_{i>k} a_x(i) \leq B_x a_x(k)$ . We have the following theorem.

THEOREM 5.3. *Let  $f : E \rightarrow [1, \infty)$  be a measurable function. Suppose that the Markov kernel  $G$  is  $\varphi$ -irreducible and satisfies the following Lyapunov criterion:*

$$(5.4) \quad (\forall x \in C^c) \quad K_{a_x}^G W(x) \leq W(x),$$

$$(5.5) \quad (\forall x \in E) \quad K_{a_x}^G V(x) \leq V(x) - Hf(x) + bI_D(x),$$

where  $C$  and  $D$  are petite sets for the Markov chain associated to  $G$ ,  $V : E \rightarrow [0, \infty]$  is a measurable function bounded on  $D$ , and  $W$  is a nonnegative function unbounded off petite sets. Then the PDMP  $\{X(t)\}$  is positive Harris recurrent and its unique invariant probability measure  $\mu$  satisfies  $\int_E f(x)\mu(dx) < \infty$ .

*Proof.* From Theorem 2.2(i) in [15], inequality (5.4) implies that  $G$  is Harris recurrent. From (5.5) and the proof of Theorem 2.2(i) in [15, p. 158], we have that for some nonnegative function  $W(x)$ , with  $W(x) < \infty$  for all  $x \in E$ , that  $GW(x) \leq W(x) - Hf(x) + bI_D(x)$  for all  $x \in E$ . From Theorem 5.2 we get the desired result.  $\square$

Finally we present a sufficient tractable condition for the PDMP to be ergodic.

THEOREM 5.4. *Suppose that the hypotheses of Theorem 5.2 and item (i) of Theorem 4.8 are satisfied. Then the PDMP  $\{X(t)\}$  is ergodic.*

*Proof.* The result follows by combining Theorems 4.8 and 5.2.  $\square$

REMARK 5.1. The results have been developed considering the stochastic kernel  $G$  defined on the measurable state space  $(E, \mathcal{B}(E))$ . It may be convenient in some applications to consider the extended state space  $E \cup \Gamma^+$  instead of  $E$ , with the stochastic kernel  $G$  extended to include the points in  $\Gamma^+$ . Let us define the kernel  $\overline{G}$  on  $(E \cup \Gamma^+, \mathcal{B}(E \cup \Gamma^+))$  (corresponding to an extension of  $G$ ) by  $(\forall x \in E \cup \Gamma^+) (\forall A \in \mathcal{B}(E \cup \Gamma^+))$

$$\overline{G}(x, A) \doteq \mathbf{1}_E(x)G(x, A \cap E) + \mathbf{1}_{\Gamma^+}(x)Q(x, A \cap E).$$

It is easy to see that  $\overline{G}$  is a stochastic kernel on  $(E \cup \Gamma^+, \mathcal{B}(E \cup \Gamma^+))$ . Moreover, it can be easily shown that  $\overline{G}$  has the following important properties:

- (a)  $\varphi$  is an irreducibility measure for  $G$  if and only if  $\varphi$  is an irreducibility measure for  $\overline{G}$ .
- (b)  $\pi$  is an invariant measure for  $G$  if and only if  $\pi$  is an invariant measure for  $\overline{G}$ .



(c)  $G$  is Harris recurrent if and only if  $\bar{G}$  is Harris recurrent. Consequently, the result presented in Theorem 5.4 remains valid if the Markov kernel  $G$  is replaced by the Markov kernel  $\bar{G}$ .

**6. Example.** This example is based on the capacity expansion model, analyzed in [7], [6, Example (34.45)], and [3, section 7]. The existence of an invariant probability measure for a generalized version of this model was studied in [8]. In the present work this example is revisited and the result of [8] is strengthened: Proposition 6.1 gives sufficient conditions to ensure that this general capacity expansion model is ergodic.

Capacity expansions are general processes of adding facilities to meet a rising demand. The demand for some utility is modeled as a random point process; i.e., it increases by one unit at random times. This demand is met by consecutive construction of expansion projects. Each project meets  $K_i$  units of demand when completed, where  $i$  corresponds the present level of excess demand. We assume that if there is an excess demand of at least  $p$  units, then the construction of a new project is started at a rate  $r_i$  per unit of time and is completed after a lead time of  $\mathcal{L}_i$  units of time. If the excess demand is less than  $p$ , then no construction takes place. New demand occurs with rate  $\lambda_i(u)$ , where  $u$  is the time spent by the current project. This problem can be modeled as a PDMP  $\{x_t\}$  with state space

$$(6.1) \quad E \doteq \{ \{p - K_p, \dots, p - 1\} \times \{0\} \} \cup \bigcup_{n=p}^{\infty} \{n, [0, \mathcal{L}_n[ \},$$

where  $p \in \mathbb{N}$  and  $(K_i)_{i \geq p}$  is a sequence of integers, and  $(\mathcal{L}_i)_{i \geq p}$  is an increasing sequence of strictly positive real numbers.  $\mathbb{N}_p$  denotes the set of integers greater than or equal to  $p$ . The local characteristics are given by

$$(6.2) \quad \lambda((i, u)) \doteq \lambda_i(u),$$

$$(6.3) \quad \Phi((i, u), t) = \begin{cases} (i, 0) & \text{for } i \in \{p - K_p, \dots, p - 1\}, \\ (i, u + r_i t) & \text{for } i \geq p, \end{cases}$$

$$(6.4) \quad Q((i, u), \{(i + 1, u)\}) = 1 \quad \text{and} \quad Q((i, \mathcal{L}_i), \{(i - K_i, 0)\}) = 1.$$

Consequently, we have  $\Gamma^+ = \{(n, \mathcal{L}_n) : n \geq p\}$ . Let us introduce the following assumptions:

- (H1) There exists  $r > 0$  such that for all  $i$ ,  $r_i \geq r$ .
- (H2) For  $i \geq p$ ,  $i - K_i \geq p - K_p > 0$  and  $\lim_{i \rightarrow \infty} i - K_i = \infty$ .
- (H3) There exists an integer  $k_0$  such that for all  $i \geq k_0$ ,  $\frac{K_i}{\mathcal{L}_i} \leq \frac{K_{i+1}}{\mathcal{L}_{i+1}}$ .
- (H4) For  $i \geq p$ ,  $\lambda_i(u)$  is a continuous real-valued function on  $[0, \mathcal{L}_i]$ .
- (H5) For  $i \in \{p - K_p, \dots, p - 1\}$ ,  $\lambda_i > 0$ .
- (H6)  $\limsup_{i \rightarrow \infty} \frac{\mathcal{L}_i}{K_i r_i} \max_{u \in [0, \mathcal{L}_i]} \lambda_i(u) < 1$ .

We have the following result.

**PROPOSITION 6.1.** *If (H1)–(H6) hold, then the PDMP  $\{X(t)\}$  is ergodic.*

*Proof.* Following the proof of Proposition 5.1 in [8], we have that the Markov kernel  $\bar{G}$  defined on  $(E \cup \Gamma^+, \mathcal{B}(E \cup \Gamma^+))$  is weak Feller and for the real-valued function defined on  $E \cup \Gamma^+$  by

$$V((i, u)) = i - \frac{K_i}{\mathcal{L}_i} u,$$

there exist  $(\varepsilon, b) \in \mathbb{R}_+^2$  and an integer  $k_0$  such that

$$\bar{G}V(i, u) - V(i, u) \leq -\varepsilon H((i, u), E) + b \mathbf{1}_{C_{k_0}}((i, u)),$$

where  $C_{k_0} \doteq \{\{p - K_p, \dots, p - 1\} \times \{0\}\} \cup \bigcup_{n=p}^{k_0} \{n, [0, \mathcal{L}_n]\}$ .

Using the hypotheses, it can be shown that  $\delta_{\{(p-1,0)\}}$  is an irreducibility measure for  $\{X(t)\}$ . The support of  $\delta_{\{(p-1,0)\}}$  clearly has a nonempty interior. It follows from item (ii) of Theorem 3.4 in [13] that all compact subsets of  $E \cup \Gamma^+$  are petite sets. Therefore,  $C_{k_0}$  is a petite set and the function  $V$  is unbounded off petite sets. Consequently, the hypotheses of Theorem 5.2 are satisfied.

Now let us show that item (i) of Theorem 4.8 is satisfied for the Markov kernel  $\overline{G}$ . Define  $C \doteq \{(p-1, 0)\}$ . Clearly,  $\delta_{\{(p-1,0)\}}(C) > 0$  and

$$(\forall x \in C) (\forall A \in \mathcal{B}(E) \otimes \mathcal{B}(\mathbb{R}_+)) \quad \alpha^0(x, A) = \int_A \delta_{\{(p-1,0)\}}(dy) e^{-\lambda_{p-1}(0)s} \gamma(ds),$$

since  $t_*(p-1, 0) = +\infty$  and  $\Phi((p-1, 0), s) = (p-1, 0)$ . Using assumption (H5), we obtain that

$$(6.5) \quad (\forall x \in C) \quad \alpha^0(x, \cdot) \sim \delta_{\{(p-1,0)\}} \otimes \gamma(\cdot).$$

Consequently, for all  $x \in C$ ,  $\alpha^0(x, \cdot)$  is nonsingular with respect to the measure  $\delta_{\{(p-1,0)\}} \otimes \gamma$ . However, if  $\Psi^{\overline{G}}$  denotes an irreducibility measure for  $\overline{G}$ , then  $\Psi^{\overline{G}}H$  is a maximal irreducibility for  $\{X(t)\}$  and so  $\delta_{\{(p-1,0)\}} \ll \Psi^{\overline{G}}H$ . Consequently, item (i) of Theorem 4.8 is satisfied for the Markov kernel  $\overline{G}$ . Combining Remark 5.1 and Theorem 5.4, we obtain the result.  $\square$

**7. Proof of Theorem 3.1.** In this section we present the proof of Theorem 3.1. First, Lemma 7.1 gives another expression for the Markov kernel  $G$ . The following results (see Lemmas 7.2 and 7.4 and Proposition 7.3) will give some relations between the sequence of stopping times  $\{\tau_n\}$  and the jump times of the PDMP  $\{T_n\}$  that will be used in what follows. The filtration generated by the PDMP stopped at times  $\{\tau_n\}$  and at times  $\{T_n\}$  are studied in Lemma 7.5. Finally, the proof of Theorem 3.1 follows by combining Proposition 7.6 and the previous results.

LEMMA 7.1. Assume that  $\bigvee_{t \geq 0} \mathcal{F}_t^x$  and  $\sigma\{s_k : k \geq 0\}$  are independent. Then

$$G(x, A) = E_x \left[ \mathbf{1}_A [X(T_1 \wedge s_0)] \right].$$

*Proof.* From the definition of  $G$  it follows by a straightforward calculation that  $G(x, A) = \int_0^{+\infty} E_x [\mathbf{1}_A [X(T_1 \wedge t)]] e^{-t} dt$ , showing the result.  $\square$

LEMMA 7.2. For any  $x \in E$  and for all  $n \in \mathbb{N}$ , we have

$$(7.1) \quad P_x(\tau_n \leq T_n) = 1,$$

$$(7.2) \quad P_x(\tau_n < \tau_{n+1}) = 1,$$

$$(7.3) \quad P_x \left( \sum_{j=n}^{\infty} \mathbf{1}_{\{\tau_j = T_n\}} = 1 \right) = 1.$$

For all  $n \in \mathbb{N}^*$ ,  $k < n$ ,  $j \in \{k, k+1, \dots, n-1\}$ , and  $x \in E$ ,

$$(7.4) \quad \tau_n = T_k + \sum_{i=j+1}^n s_i,$$

$P_x$ -a.s. on  $\{T_k \leq \tau_n < T_{k+1}\} \cap \{\tau_j = T_k\}$ .

*Proof.* These results can be shown easily by induction on the definition of  $\tau_n$ . The proof is therefore omitted.  $\square$

PROPOSITION 7.3. For all  $n \in \mathbb{N}_*$ ,  $k < n$ ,

$$(7.5) \quad \tau_n \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} = \sum_{j=k}^{n-1} \mathbf{1}_{\{\tau_j = T_k\}} \left[ T_k + \sum_{i=j+1}^n s_i \right] \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} + \mathbf{1}_{\{T_k = \tau_n\}} T_k.$$

*Proof.* From (7.3), we have

$$(7.6) \quad \tau_n \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} = \tau_n \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} \left[ \sum_{j=k}^n \mathbf{1}_{\{\tau_j = T_k\}} + \sum_{j=n+1}^{\infty} \mathbf{1}_{\{\tau_j = T_k\}} \right],$$

and from (7.2), we get  $\{T_k \leq \tau_n < T_{k+1}\} \cap \{\tau_j = T_k\} = \emptyset$  for  $j \geq n+1$ . Consequently, the last term in (7.6) cancels out, and, by using (7.4), the result follows.  $\square$

It is important to point out that in the next lemma the function  $H_{k,p}(\tau_{p-1}, T_1, \dots, T_k, s_1, \dots, s_p)$  does not depend on  $T_{k+1}$ .

LEMMA 7.4. For all  $(k, p, n) \in \mathbb{N}_*^3$  such that  $p \leq n$  and  $k < n$ , there exists a measurable function  $H_{k,p}$  such that

$$\tau_p \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} = H_{k,p}(T_1, \dots, T_k, s_1, \dots, s_p) \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}}.$$

*Proof.* For  $p \leq n$  and  $k < n$  and from the definition of  $\tau_p$ , we have

$$(7.7) \quad \tau_p \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} = \sum_{j=0}^{p-1} \mathbf{1}_{\{T_j \leq \tau_{p-1} < T_{j+1}\}} \left[ (\tau_{p-1} + s_p) \wedge T_{j+1} \right] \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}}.$$

However, if  $k < p-1$ , then  $\{T_j \leq \tau_{p-1} < T_{j+1}\} \cap \{T_k \leq \tau_n < T_{k+1}\} = \emptyset$  for  $j \in \{k+1, \dots, p-1\}$  by using (7.2). Consequently, it follows from (7.7) that

$$\begin{aligned} \tau_p \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} &= \sum_{j=0}^{k \wedge (p-1)} \mathbf{1}_{\{T_j \leq \tau_{p-1} < T_{j+1}\}} \left[ (\tau_{p-1} + s_p) \wedge T_{j+1} \right] \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} \\ &= \sum_{j=0}^{(k-1) \wedge (p-2)} \mathbf{1}_{\{T_j \leq \tau_{p-1} < T_{j+1}\}} \left[ (\tau_{p-1} + s_p) \wedge T_{j+1} \right] \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} \\ &\quad + \mathbf{1}_{\{p-1 < k\}} \mathbf{1}_{\{T_{p-1} \leq \tau_{p-1} < T_p\}} \left[ (\tau_{p-1} + s_p) \wedge T_p \right] \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} \\ (7.8) \quad &+ \mathbf{1}_{\{p-1 \geq k\}} \mathbf{1}_{\{T_k \leq \tau_{p-1} < T_{k+1}\}} \left[ (\tau_{p-1} + s_p) \wedge T_{k+1} \right] \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}}. \end{aligned}$$

However, by using (7.1) we have  $\{T_{p-1} \leq \tau_{p-1} < T_p\} = \{T_{p-1} = \tau_{p-1}\}$ , and  $\{T_k \leq \tau_{p-1} < T_{k+1}\} \cap \{T_k \leq \tau_n < T_{k+1}\} = \{T_k \leq \tau_{p-1}\} \cap \{T_k \leq \tau_n < T_{k+1}\}$ . On the set  $\{T_k \leq \tau_{p-1} < T_{k+1}\} \cap \{T_k \leq \tau_n < T_{k+1}\}$ , we get  $(\tau_{p-1} + s_p) \wedge T_{k+1} = \tau_{p-1} + s_p$ . Taking these remarks into consideration, (7.8) becomes

$$\begin{aligned} \tau_p \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} &= \left\{ \sum_{j=0}^{(k-1) \wedge (p-2)} \mathbf{1}_{\{T_j \leq \tau_{p-1} < T_{j+1}\}} \left[ (\tau_{p-1} + s_p) \wedge T_{j+1} \right] \right. \\ &\quad + \mathbf{1}_{\{p-1 < k\}} \mathbf{1}_{\{T_{p-1} = \tau_{p-1}\}} \left[ (\tau_{p-1} + s_p) \wedge T_p \right] \\ &\quad \left. + \mathbf{1}_{\{p-1 \geq k\}} \mathbf{1}_{\{T_k \leq \tau_{p-1}\}} (\tau_{p-1} + s_p) \right\} \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}}. \end{aligned}$$

Consequently, there exists a measurable function  $H_{p,k}$  such that

$$\tau_p \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} = H_{k,p}(\tau_{p-1}, T_1, \dots, T_k, s_1, \dots, s_p) \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}}.$$

Now repeating the same arguments for  $\tau_{p-1} \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}}$ , the result follows by induction.  $\square$

LEMMA 7.5. Define  $\mathcal{G}_n = \sigma\{X(t \wedge \tau_n) : t \in \mathbb{R}_+\} \vee \sigma\{\tau_j : j \leq n\}$ , and  $\mathcal{S}_n = \sigma\{s_j : j \leq n\}$ . For  $n \in \mathbb{N}_*$  and  $k < n$ , we have that

$$\left[ \mathcal{G}_n \cap \{T_k \leq \tau_n < T_{k+1}\} \right] \subset \left[ \left( \mathcal{S}_n \vee \sigma\{X(t \wedge T_k) : t \in \mathbb{R}_+\} \right) \cap \{T_k \leq \tau_n < T_{k+1}\} \right].$$

*Proof.* For all  $t \in \mathbb{R}_+$ , it follows from Lemma 7.4 that for all  $p \leq n$ ,

$$\{\tau_p \leq t\} \cap \{T_k \leq \tau_n < T_{k+1}\} \in \left( \mathcal{S}_n \vee \sigma\{X(t \wedge T_k) : t \in \mathbb{R}_+\} \right) \cap \{T_k \leq \tau_n < T_{k+1}\}.$$

Now, from Proposition 7.3, we have on the set  $\{t \geq T_k\} \cap \{T_k \leq \tau_n < T_{k+1}\}$

$$X(t \wedge \tau_n) = \sum_{j=k}^{n-1} \mathbf{1}_{\{\tau_j = T_k\}} \Phi \left( X(T_k), (t - T_k) \wedge \sum_{i=j+1}^n s_i \right) + \mathbf{1}_{\{\tau_n = T_k\}} \Phi(X(T_k), 0).$$

Therefore, for any  $B \in \mathcal{B}(\mathbb{R})$  and for all  $t \in \mathbb{R}_+$ , we have

$$\begin{aligned} \{X(t \wedge \tau_n) \in B\} \cap \{T_k \leq \tau_n < T_{k+1}\} &= \left( \left[ \{t < T_k\} \cap \{X(t \wedge T_k) \in B\} \right] \right. \\ &\quad \cup \left[ \bigcup_{j=k}^{n-1} \left[ \{\tau_j = T_k\} \cap \left\{ \Phi \left( X(T_k), (t - T_k) \wedge \sum_{i=j+1}^n s_i \right) \in B \right\} \cap \{T_k \leq t\} \right] \right] \\ &\quad \left. \cup \left[ \{\tau_n = T_k\} \cap \{X(T_k) \in B\} \cap \{T_k \leq t\} \right] \right) \cap \{T_k \leq \tau_n < T_{k+1}\}. \end{aligned}$$

Using again Lemma 7.4, it follows that

$$\{X(t \wedge \tau_n) \in B\} \cap \{T_k \leq \tau_n < T_{k+1}\} \in \left( \mathcal{S}_n \vee \sigma\{X(t \wedge T_k) : t \geq 0\} \right) \cap \{T_k \leq \tau_n < T_{k+1}\},$$

showing the result.  $\square$

PROPOSITION 7.6. Assume that  $\forall_{t \geq 0} \mathcal{F}_t^x$  and  $\sigma\{s_k : k \geq 0\}$  are independent. For  $n \in \mathbb{N}_*$  and  $k \leq n$ , we have that

$$E_x \left[ \mathbf{1}_{\{t + \tau_n < T_{k+1}\}} | \mathcal{G}_n \right] \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} = F(t, \Theta_n) \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}},$$

where  $F(t, x) = \mathbf{1}_{\{t < t_*(x)\}} e^{-\int_0^t \lambda(\Phi(x, s)) ds}$ .

*Proof.* Denote  $S_{k+1} = T_{k+1} - T_k$ . Consider first the case where  $n \in \mathbb{N}_*$  and  $k = n$ . From (7.1) and by Theorem 25.5 in [6] and its proof, it follows that

$$E_x \left[ \mathbf{1}_{\{t + \tau_n < T_{k+1}\}} | \mathcal{G}_n \right] \mathbf{1}_{\{T_k \leq \tau_k < T_{k+1}\}} = F(t, X(\tau_k)) \mathbf{1}_{\{T_k \leq \tau_k < T_{k+1}\}}.$$

Now consider the case where  $n \in \mathbb{N}_*$  and  $k < n$ . From Proposition 7.3, we have

$$(7.9) \quad \begin{aligned} E_x \left[ \mathbf{1}_{\{t+\tau_n < T_{k+1}\}} | \mathcal{G}_n \right] \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} &= \left[ \sum_{j=k}^{n-1} E_x \left[ \mathbf{1}_{\{\tau_j = T_k\}} \mathbf{1}_{\{S_{k+1} > \sum_{i=j+1}^n s_i + t\}} | \mathcal{G}_n \right] \right. \\ &\quad \left. + E_x \left[ \mathbf{1}_{\{\tau_n = T_k\}} \mathbf{1}_{\{S_{k+1} > t\}} | \mathcal{G}_n \right] \right] \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}}. \end{aligned}$$

From Lemma 7.5, we obtain that for any  $A \in \mathcal{G}_n$ , there exists  $B \in \mathcal{S}_n \vee \sigma\{X(t \wedge T_k) : t \in \mathbb{R}_+\}$  such that

$$A \cap \{T_k \leq \tau_n < T_{k+1}\} = B \cap \{T_k \leq \tau_n < T_{k+1}\}.$$

By denoting  $\mathcal{H}_k = \sigma\{s_j : j \geq 0\} \vee \sigma\{X(t \wedge T_k) : t \in \mathbb{R}_+\}$ , we have for  $j \leq n-1$

$$\begin{aligned} \int_A \mathbf{1}_{\{S_{k+1} > \sum_{i=j+1}^n s_i + t\}} \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} \mathbf{1}_{\{\tau_j = T_k\}} dP \\ = \int_B \mathbf{1}_{\{\tau_j = T_k\}} E_x \left[ \mathbf{1}_{\{S_{k+1} > \sum_{i=j+1}^n s_i + t\}} | \mathcal{H}_k \right] dP. \end{aligned}$$

However, since  $\vee_{t \geq 0} \mathcal{F}_t^X$  and  $\sigma\{s_k : k \geq 0\}$  are independent, we have that

$$E_x \left[ \mathbf{1}_{\{S_{k+1} > \sum_{i=j+1}^n s_i + t\}} | \mathcal{H}_{n,k} \right] = F \left( \sum_{i=j+1}^n s_i + t, X(T_k) \right).$$

Therefore, by using the semigroup property of  $\Phi$ , we obtain

$$\begin{aligned} \int_A \mathbf{1}_{\{S_{k+1} > \sum_{i=j+1}^n s_i + t\}} \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} \mathbf{1}_{\{\tau_j = T_k\}} dP \\ = \int_A \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} \mathbf{1}_{\{\tau_j = T_k\}} F(t, X(\tau_n)) dP. \end{aligned}$$

Consequently, it follows that for  $j \leq n-1$

$$(7.10) \quad E_x \left[ \mathbf{1}_{\{S_{k+1} > \sum_{i=j+1}^n s_i + t\}} | \mathcal{G}_n \right] = F(t, X(\tau_n))$$

on the set  $\{\tau_j = T_k\} \cap \{T_k \leq \tau_n < T_{k+1}\}$ . Similarly, we have that

$$(7.11) \quad \mathbf{1}_{\{\tau_n = T_k\}} E_x \left[ \mathbf{1}_{\{S_{k+1} > t\}} | \mathcal{G}_n \right] = \mathbf{1}_{\{\tau_n = T_k\}} F(t, X(\tau_n)).$$

Combining (7.9), (7.10), and (7.11), we have that

$$E_x \left[ \mathbf{1}_{\{t+\tau_n < T_{k+1}\}} | \mathcal{G}_n \right] \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} = \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} F(t, X(\tau_n)) \sum_{j=k}^n \mathbf{1}_{\{\tau_j = T_k\}},$$

showing the result.  $\square$

Using the previous results, we now present the proof of Theorem 3.1.

*Proof.* Consider  $n \in \mathbb{N}_*$  and denote  $\sigma\{X(\tau_k) : k \leq n\}$  by  $\mathcal{F}_n^\Theta$ . From the definition of  $\tau_n$  (see (3.1)), we have

$$E_x \left[ \mathbf{1}_A(\Theta_{n+1}) | \mathcal{F}_n^\Theta \right] = E_x \left[ \sum_{k=0}^n \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} \mathbf{1}_A[X((\tau_n + s_{n+1}) \wedge T_{k+1})] | \mathcal{F}_n^\Theta \right].$$

However, by using the facts that  $\vee_{t \geq 0} \mathcal{F}_t^X$  and  $\sigma\{s_k : k \geq 0\}$  are independent and that  $\{s_n\}_{n \geq 0}$  is a sequence of independent and identically distributed  $\mathbb{R}_+$ -valued random variables with exponential distribution, we obtain

$$\begin{aligned} E_x \left[ \mathbf{1}_A(\Theta_{n+1}) | \mathcal{F}_n^\Theta \right] &= \int_0^{+\infty} E_x \left[ \sum_{k=0}^n \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} \mathbf{1}_A[X((\tau_n + t) \wedge T_{k+1})] | \mathcal{F}_n^\Theta \right] e^{-t} dt \\ (7.12) \quad &= \int_0^{+\infty} E_x \left[ \sum_{k=0}^n \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} E_x \left[ \mathbf{1}_A[X((\tau_n + t) \wedge T_{k+1})] | \mathcal{G}_n \right] | \mathcal{F}_n^\Theta \right] e^{-t} dt, \end{aligned}$$

where for the last equality we have used  $\{T_k \leq \tau_n < T_{k+1}\} \in \mathcal{G}_n$  and  $\mathcal{F}_n^\Theta \subset \mathcal{G}_n$ . For  $k \leq n$ , by using Proposition 7.6 we have that on the set  $\{T_k \leq \tau_n < T_{k+1}\}$

$$(7.13) \quad E_x \left[ \mathbf{1}_A[X((\tau_n + t) \wedge T_{k+1})] | \mathcal{G}_n \right] = E_{X(\tau_n)} \left[ \mathbf{1}_A[X(T_1 \wedge t)] \right].$$

Combining (7.12) and (7.13), it follows that

$$\begin{aligned} E_x \left[ \mathbf{1}_A(\Theta_{n+1}) | \mathcal{F}_n^\Theta \right] &= \int_0^{+\infty} E_x \left[ \sum_{k=0}^n \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} E_{X(\tau_n)} \left[ \mathbf{1}_A[X(T_1 \wedge t)] \right] | \mathcal{F}_n^\Theta \right] e^{-t} dt \\ &= \int_0^{+\infty} E_{X(\tau_n)} \left[ \mathbf{1}_A[X(T_1 \wedge t)] \right] E_x \left[ \sum_{k=0}^n \mathbf{1}_{\{T_k \leq \tau_n < T_{k+1}\}} | \mathcal{F}_n^\Theta \right] e^{-t} dt \\ &= G(X(\tau_n), A), \end{aligned}$$

showing the result.  $\square$

**8. Proof of Theorem 3.3.** In this section we present the proof of Theorem 3.3. Propositions 8.1 and 8.2 show how we can represent  $L^R$  as an infinite sum of substochastic kernels. Proposition 8.3 presents iterative equations for  $L^R$  and  $L^G$ , which, combined with the previous propositions and some limit convergence proved in Proposition 8.4, yield the proof of Theorem 3.3.

**PROPOSITION 8.1.** *Let  $f$  be an  $\mathbb{R}_+$ -valued measurable function defined on  $E$ . For  $B \in \mathcal{B}(E)$  and  $n \in \mathbb{N}$ ,*

$$(8.1) \quad (HI_B)^{n+1} f(x) = E_x \left[ \int_0^{T_1} v_{n,B}(x, s) e^{-s} f(X(s)) \mathbf{1}_B(X(s)) ds \right],$$

$$(8.2) \quad (HI_B)^n Jf(x) = E_x \left[ v_{n,B}(x, T_1) e^{-T_1} f(X(T_1)) \right],$$

where

$$(8.3) \quad v_{n,B}(x, s) = \frac{1}{n!} \left[ \int_0^s \mathbf{1}_B(\Phi(x, u)) du \right]^n.$$

*Proof.* Notice that for  $s \geq t$  we have that

$$v_{n,B}(\Phi(x, t), s - t) = \frac{1}{n!} \left[ \int_0^{s-t} \mathbf{1}_B(\Phi(x, u + t)) du \right]^n = \frac{1}{n!} \left[ \int_t^s \mathbf{1}_B(\Phi(x, u)) du \right]^n$$

and thus

$$(8.4) \quad \frac{dv_{n,B}(\Phi(x, t), s - t)}{dt} = -v_{n-1,B}(\Phi(x, t), s - t) \mathbf{1}_B(\Phi(x, t)).$$

Let us first show (8.1) by induction on  $n$ . For  $n = 1$  the result follows from the fact that

$$(8.5) \quad H(x, A) = E_x \left[ \int_0^{T_1} e^{-s} \mathbf{1}_A(X(s)) ds \right].$$

Suppose it holds for  $n$ . Note that for any  $\mathbb{R}_+$ -valued measurable function  $h$  defined on  $\mathbb{R}$ , we have

$$(8.6) \quad E_x \left[ \int_0^{T_1} h(s) ds \right] = \int_0^{t_*(x)} h(s) e^{-\Lambda(x,s)} ds.$$

Combining (8.5) and (8.6), we have

$$(8.7) \quad \begin{aligned} (HI_B)^{n+1} f(x) &= HI_B (HI_B)^n f(x) \\ &= \int_0^{t_*(x)} \mathbf{1}_B(\Phi(x, s)) (HI_B)^n f(\Phi(x, s)) e^{-\{s+\Lambda(x,s)\}} ds. \end{aligned}$$

However, from the induction hypothesis and (8.6), we have

$$(HI_B)^n f(x) = \int_0^{t_*(x)} v_{n-1,B}(x, u) \mathbf{1}_B(\Phi(x, u)) f(\Phi(x, u)) e^{-\{u+\Lambda(x,u)\}} du.$$

Using the semigroup property of  $\Phi$ , it follows that

$$(8.8) \quad \begin{aligned} (HI_B)^n f(\Phi(x, s)) &= e^{\{s+\Lambda(x,s)\}} \int_s^{t_*(x)} v_{n-1,B}(\Phi(x, s), u-s) \mathbf{1}_B(\Phi(x, u)) \\ &\quad \times f(\Phi(x, u)) e^{-\{u+\Lambda(x,u)\}} du. \end{aligned}$$

Combining (8.7) and (8.8), we obtain that

$$\begin{aligned} (HI_B)^{n+1} f(x) &= \int_0^{t_*(x)} \mathbf{1}_B(\Phi(x, u)) f(\Phi(x, u)) e^{-\{u+\Lambda(x,u)\}} \\ &\quad \times \int_0^{t_*(x)} v_{n-1,B}(\Phi(x, s), u-s) \mathbf{1}_B(\Phi(x, s)) \mathbf{1}_{[0,u]}(s) ds du. \end{aligned}$$

Now from (8.4), we have

$$\int_0^u v_{n-1,B}(\Phi(x, s), u-s) \mathbf{1}_B(\Phi(x, s)) ds = v_{n,B}(x, u),$$

and so

$$(HI_B)^{n+1} f(x) = \int_0^{t_*(x)} v_{n,B}(x, u) \mathbf{1}_B(\Phi(x, u)) f(\Phi(x, u)) e^{-\{u+\Lambda(x,u)\}} du.$$

Finally, using (8.6), we obtain

$$(HI_B)^{n+1} f(x) = E_x \left[ \int_0^{T_1} v_{n,B}(x, s) e^{-s} f(X(s)) \mathbf{1}_B(X(s)) ds \right],$$

showing the first part of the result.

The second part of the result can be obtained by using the same arguments and the fact that

$$\begin{aligned} E_x[g(T_1)f(X(T_1))] &= \int_0^{t_*(x)} \lambda(\Phi(x, s))e^{-\Lambda(x, s)}g(s)Qf(\Phi(x, s))ds \\ &\quad + e^{-\Lambda(x, t_*(x))}g(t_*(x))Qf(\Phi(x, t_*(x))) \end{aligned}$$

for any  $\mathbb{R}_+$ -valued measurable function  $f$  (respectively,  $g$ ) defined on  $\mathbb{R}$  (respectively,  $E$ ). Indeed, for  $n = 0$  the result follows from the fact that

$$J(x, A) = E_x[e^{-T_1}\mathbf{1}_A(X(T_1))].$$

Now, suppose it holds for  $n$ . Then

$$(HI_B)^{n+1}Jf(x) = \int_0^{t_*(x)} \mathbf{1}_B(\Phi(x, s))(HI_B)^nJf(\Phi(x, s))e^{-\{s+\Lambda(x, s)\}}ds$$

and

$$\begin{aligned} (HI_B)^nJf(\Phi(x, t)) &= \left[ e^{-\Lambda(x, t_*(x))}Qf(\Phi(x, t_*(x)))v_{n,B}(\Phi(x, t), t_*(x) - t) \right. \\ &\quad \left. + \int_t^{t_*(x)} \lambda(\Phi(x, s))e^{-\Lambda(x, s)}Qf(\Phi(x, s))v_{n,B}(\Phi(x, t), s - t)ds \right] e^{\{t+\Lambda(x, t)\}}. \end{aligned}$$

Combining the two previous equations and (8.4) gives

$$\begin{aligned} (HI_B)^{n+1}Jf(x) &= \int_0^{t_*(x)} \lambda(\Phi(x, s))e^{-\{s+\Lambda(x, s)\}}Qf(\Phi(x, s))v_{n+1,B}(x, s)ds \\ &\quad + e^{-\{t_*(x)+\Lambda(x, t_*(x))\}}Qf(\Phi(x, t_*(x)))v_{n+1,B}(x, t_*(x)). \end{aligned}$$

From (8.9) we have  $(HI_B)^{n+1}Jf(x) = E_x[v_{n,B}(x, T_1)e^{-T_1}f(X(T_1))]$ , completing the proof.  $\square$

PROPOSITION 8.2. *For any  $A \in \mathcal{B}(E)$  define the kernel  $\mathbb{H}_A$  on  $(E, \mathcal{B}(E))$  as*

$$\mathbb{H}_A = \sum_{n=0}^{\infty} (HI_{A^c})^n.$$

Then

$$(8.9) \quad L^R(x, A) = \sum_{n=0}^{\infty} (\mathbb{H}_AJ)^n \mathbb{H}_AH\mathbf{1}_A(x).$$

*Proof.* From Proposition 8.1 we have that for any  $\mathbb{R}_+$ -valued measurable function  $g$  defined on  $E$

$$\begin{aligned} \mathbb{H}_Ag(x) &= g(x) + E_x \left[ \int_0^{T_1} \sum_{n=0}^{\infty} v_{n,A^c}(x, s)e^{-s}g(X(s))\mathbf{1}_{A^c}(X(s))ds \right] \\ &= g(x) + E_x \left[ \int_0^{T_1} e^{-s+\int_0^s \mathbf{1}_{A^c}(\Phi(x, u))du} g(\Phi(x, s))\mathbf{1}_{A^c}(\Phi(x, s))ds \right] \\ &= g(x) + \int_0^{t_*(x)} e^{-\int_0^s \mathbf{1}_A(\Phi(x, u))du} g(\Phi(x, s))\mathbf{1}_{A^c}(\Phi(x, s))e^{-\Lambda(x, s)}ds. \end{aligned}$$



However,  $H\mathbf{1}_A(\Phi(x, s)) = \int_t^{t_*(x)} \mathbf{1}_A(\Phi(x, s))e^{-[s+\Lambda(x, s)]}ds e^{t+\Lambda(x, t)}$ . Consequently,

$$\begin{aligned}
 \mathbb{H}_A H\mathbf{1}_A(x) &= \int_0^{t_*(x)} \mathbf{1}_A(\Phi(x, s))e^{-[s+\Lambda(x, s)]} \int_0^s e^{t-\int_0^t \mathbf{1}_A(\Phi(x, u))du} \mathbf{1}_{A^c}(\Phi(x, t))dt ds \\
 &\quad + H\mathbf{1}_A(x) \\
 &= \int_0^{t_*(x)} \mathbf{1}_A(\Phi(x, s))e^{-[s+\Lambda(x, s)]} \left[ e^{-\int_0^s \mathbf{1}_{A^c}(\Phi(x, u))du} - 1 \right] ds + H\mathbf{1}_A(x) \\
 &= \int_0^{t_*(x)} \mathbf{1}_A(\Phi(x, s))e^{-\Lambda(x, s)} e^{-\int_0^s \mathbf{1}_A(\Phi(x, u))du} ds \\
 (8.10) \quad &= E_x \left[ \int_0^{T_1} e^{-\int_0^t \mathbf{1}_A(X(u))du} \mathbf{1}_A(X(t))dt \right].
 \end{aligned}$$

Using Proposition 8.1, we have that for any  $\mathbb{R}_+$ -valued measurable function  $f$  defined on  $E$

$$\begin{aligned}
 \mathbb{H}_A Jf(x) &= E_x \left[ \sum_{n=0}^{\infty} v_{n, A^c}(x, T_1) e^{-T_1} f(X(T_1)) \right] \\
 &= E_x \left[ e^{-T_1 + \int_0^{T_1} \mathbf{1}_{A^c}(\Phi(x, u))du} f(X(T_1)) \right] \\
 (8.11) \quad &= E_x \left[ e^{-\int_0^{T_1} \mathbf{1}_A(\Phi(x, u))du} f(X(T_1)) \right].
 \end{aligned}$$

From (8.10), (8.11), and the Markov property of the process  $\{X(t)\}$ , we get by induction

$$(8.12) \quad (\mathbb{H}_A J)^n \mathbb{H}_A H\mathbf{1}_A(x) = E_x \left[ \int_{T_n}^{T_{n+1}} e^{-\int_0^t \mathbf{1}_A(X(u))du} \mathbf{1}_A(X(t))dt \right].$$

Taking the sum over  $n$  in (8.12) and recalling that a.s.  $T_n \rightarrow \infty$  as  $n \rightarrow \infty$ , it follows that

$$\begin{aligned}
 \sum_{n=0}^{\infty} (\mathbb{H}_A J)^n \mathbb{H}_A H\mathbf{1}_A(x) &= E_x \left[ \int_0^{\infty} e^{-\int_0^t \mathbf{1}_A(X(u))du} \mathbf{1}_A(X(t))dt \right] \\
 &= 1 - E_x \left[ e^{-\int_0^{\infty} \mathbf{1}_A(X(u))du} \right] \\
 &= 1 - E_x \left[ e^{-\eta_A^x} \right] = L^R(x, A),
 \end{aligned}$$

where the last equality follows from, for instance, Theorem 2.3(ii) of [11].  $\square$

PROPOSITION 8.3. *We have that*

$$(8.13) \quad L^R(x, A) = \mathbb{H}_A H(x, A) + \mathbb{H}_A J L^R(x, A),$$

$$(8.14) \quad L^G(x, A) = \mathbb{H}_A G(x, A) + \mathbb{H}_A J I_{A^c} L^G(x, A).$$

*Proof.* Note that  $L^R(x, A) = \sum_{n=1}^{\infty} (R I_{A^c})^{(n-1)} R(x, A)$ . Since  $R = \sum_{j=0}^{\infty} J^j H$  we have that  $R = H + JR$ . Therefore, from [17, Lemma 4.1], it follows that for any  $n \in \mathbb{N}^*$

$$\begin{aligned}
 (R I_{A^c})^n &= (H I_{A^c} + J R I_{A^c})^n \\
 &= (H I_{A^c})^n + \sum_{j=1}^n (H I_{A^c})^{j-1} J (R I_{A^c}) (R I_{A^c})^{n-j} \\
 &= (H I_{A^c})^n + \sum_{j=1}^n (H I_{A^c})^{j-1} J (R I_{A^c})^{n+1-j}.
 \end{aligned}$$

Taking the sum over  $n$  and using the fact that  $R = H + JR$ , we get that

$$\begin{aligned} L^R(x, A) &= \sum_{n=0}^{\infty} (HI_{A^c})^n R(x, A) + \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} (HI_{A^c})^{j-1} J(RI_{A^c})^{n+1-j} R(x, A) I_{\{j \leq n\}} \\ &= \mathbb{H}_A R(x, A) + \sum_{j=1}^{\infty} (HI_{A^c})^{j-1} J \sum_{n=j}^{\infty} (RI_{A^c})^{n+1-j} R(x, A) \\ &= \mathbb{H}_A H(x, A) + \mathbb{H}_A JR(x, A) + \mathbb{H}_A J \sum_{n=1}^{\infty} (RI_{A^c})^n R(x, A), \end{aligned}$$

giving (8.13). Similarly for  $G$ , we have  $L^G(x, A) = \sum_{n=1}^{\infty} (GI_{A^c})^{(n-1)} G(x, A)$  and so

$$\begin{aligned} L^G(x, A) &= \sum_{n=0}^{\infty} (HI_{A^c})^n G(x, A) + \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} (HI_{A^c})^{j-1} JI_{A^c} (GI_{A^c})^{n-j} G(x, A) I_{\{j \leq n\}} \\ &= \mathbb{H}_A G(x, A) + \sum_{j=1}^{\infty} (HI_{A^c})^{j-1} JI_{A^c} \sum_{n=j}^{\infty} (GI_{A^c})^{n-j} GI_{A^c}(x), \end{aligned}$$

showing the last part of the result.  $\square$

PROPOSITION 8.4. *For every  $x \in E$  and for  $A \in \mathcal{B}(E)$ , we have that*

$$\lim_{n \rightarrow \infty} (\mathbb{H}_A J)^n L^R(x, A) = 0$$

and the limit

$$\lim_{n \rightarrow \infty} (\mathbb{H}_A JI_{A^c})^n L^G(x, A)$$

exists.

*Proof.* From (8.13) we have that for every  $x \in E$  and all  $n \in \mathbb{N}^*$ ,

$$(8.15) \quad \sum_{k=0}^n (\mathbb{H}_A J)^k \mathbb{H}_A H \mathbf{1}_A(x) + (\mathbb{H}_A J)^{n+1} G^R(x, A) = L^R(x, A).$$

Taking the limit as  $n \rightarrow \infty$  in (8.15), the first part of the result follows from (8.9).

Similarly, from (8.14) we have that every  $x \in E$  and all  $n \in \mathbb{N}^*$

$$(8.16) \quad \sum_{k=0}^n (\mathbb{H}_A JI_{A^c})^k \mathbb{H}_A G \mathbf{1}_A(x) + (\mathbb{H}_A JI_{A^c})^{n+1} L^G(x, A) = L^G(x, A).$$

Since  $\sum_{k=0}^n (\mathbb{H}_A JI_{A^c})^k \mathbb{H}_A G \mathbf{1}_A(x) \geq 0$ , and  $(\mathbb{H}_A JI_{A^c})^{n+1} L^G(x, A) \geq 0$  for all  $x \in E$  and all  $n \in \mathbb{N}^*$ , it follows that  $\lim_{n \rightarrow \infty} \sum_{k=0}^n (\mathbb{H}_A JI_{A^c})^k \mathbb{H}_A G \mathbf{1}_A(x)$  exists in  $\mathbb{R}_+$ , implying the existence of  $\lim_{n \rightarrow \infty} (\mathbb{H}_A JI_{A^c})^n L^G(x, A)$ , completing the proof.  $\square$

Using the previous results, we now present the proof of Theorem 3.3.

*Proof.* From (8.13) and (8.14), it follows that for all  $x \in E$ ,

$$\begin{aligned} L^G(x, A) - L^R(x, A) &= \mathbb{H}_A (G(x, A) - H(x, A)) + \mathbb{H}_A JI_{A^c} U^G(x, A) \\ &\quad - \mathbb{H}_A J (I_{A^c} + I_A) L^R(x, A) \\ &= \mathbb{H}_A J(x, A) - \mathbb{H}_A JI_{A^c} L^R(x, A) \\ &\quad + \mathbb{H}_A JI_{A^c} [L^G(x, A) - L^R(x, A)] \\ &= \mathbb{H}_A JI_{A^c} [\mathbf{1}_E(x) - L^R(x, A)] \\ &\quad + \mathbb{H}_A JI_{A^c} [L^G(x, A) - L^R(x, A)]. \end{aligned}$$

Consequently, for all  $x \in E$  and all  $n \in \mathbb{N}^*$ ,

$$\begin{aligned} L^G(x, A) - L^R(x, A) &= \sum_{k=0}^n [\mathbb{H}_A J I_{A^c}]^k \mathbb{H}_A J I_A [\mathbf{1}_E(x) - L^R(x, A)] \\ &\quad + [\mathbb{H}_A J I_{A^c}]^{n+1} [L^G(x, A) - L^R(x, A)]. \end{aligned}$$

From Proposition 8.4, and using the fact that  $\lim_{n \rightarrow \infty} (\mathbb{H}_A J \mathbf{1}_{A^c})^n L^G(x, A)$  is non-negative, it follows that

$$L^G(x, A) - L^R(x, A) \geq \sum_{k=0}^{\infty} [\mathbb{H}_A J I_{A^c}]^k \mathbb{H}_A J I_A [\mathbf{1}_E(x) - L^R(x, A)].$$

Since for all  $x \in E$ ,  $A \in \mathcal{B}(E)$ ,  $L^R(x, A) \leq \mathbf{1}_E(x)$ , the result follows.  $\square$

**Acknowledgment.** The authors would like to express their gratitude to the associate editor and referees for their suggestions and helpful comments.

#### REFERENCES

- [1] J. AZÉMA, M. DUFLO, AND D. REVUZ, *Mesure invariante sur les classes récurrentes des processus de Markov*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 8 (1967), pp. 157–181.
- [2] J. AZÉMA, M. DUFLO, AND D. REVUZ, *Propriétés relatives des processus de Markov récurrents*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 13 (1969), pp. 286–314.
- [3] O. L. V. COSTA, *Stationary distributions for piecewise-deterministic Markov processes*, J. Appl. Probab., 27 (1990), pp. 60–73.
- [4] J. G. DAI, *On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models*, Ann. Appl. Probab., 5 (1995), pp. 49–77.
- [5] J. G. DAI AND S. P. MEYN, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Trans. Automat. Control, 40 (1995), pp. 1889–1904.
- [6] M. H. A. DAVIS, *Markov Models and Optimization*, Chapman and Hall, London, 1993.
- [7] M. H. A. DAVIS, M. A. H. DEMPSTER, S. P. SETHI, AND D. VERMES, *Optimal capacity expansion under uncertainty*, Adv. in Appl. Probab., 19 (1987), pp. 156–176.
- [8] F. DUFOUR AND O. L. V. COSTA, *Stability of piecewise-deterministic Markov processes*, SIAM J. Control Optim., 37 (1999), pp. 1483–1502.
- [9] V. A. MALÝŠEV AND M. V. MEN'SIKOV, *Ergodicity, continuity and analyticity of countable Markov chains*, Trans. Moscow Math. Soc., 1 (1982), pp. 1–48.
- [10] S. P. MEYN AND R. L. TWEEDIE, *Stability of Markovian processes I: Criteria for discrete-time chains*, Adv. in Appl. Probab., 24 (1992), pp. 542–574.
- [11] S. P. MEYN AND R. L. TWEEDIE, *Generalized resolvents and Harris recurrence of Markov processes*, in Doeblin and Modern Probability (Blaubeuren, 1991), Contemp. Math. 149, AMS, Providence, RI, 1993, pp. 227–250.
- [12] S. P. MEYN AND R. L. TWEEDIE, *Markov Chains and Stochastic Stability*, Springer-Verlag, Berlin, 1993.
- [13] S. P. MEYN AND R. L. TWEEDIE, *Stability of Markovian processes II: Continuous-time processes and sampled chains*, Adv. in Appl. Probab., 25 (1993), pp. 487–517.
- [14] S. P. MEYN AND R. L. TWEEDIE, *Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes*, Adv. in Appl. Probab., 25 (1993), pp. 518–548.
- [15] S. P. MEYN AND R. L. TWEEDIE, *State-dependent criteria for convergence of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 149–168.
- [16] S. NIEMI AND E. NUMMELIN, *On nonsingular renewal kernels with application to a semigroup of transition kernels*, Stochastic Process. Appl., 22 (1986), pp. 177–202.
- [17] E. NUMMELIN, *General Irreducible Markov Chains and Nonnegative Operators*, Cambridge University Press, Cambridge, UK, 1984.
- [18] D. REVUZ, *Markov Chains*, 2nd ed., North–Holland Math. Library 11, North–Holland, Amsterdam, 1984.
- [19] P. TUOMINEN AND R. TWEEDIE, *The recurrence structure of general Markov processes*, Proc. London Math. Soc., 39 (1978), pp. 554–576.

## SINGULAR TRAJECTORIES OF CONTROL-AFFINE SYSTEMS\*

YACINE CHITOUR<sup>†</sup>, FRÉDÉRIC JEAN<sup>‡</sup>, AND EMMANUEL TRÉLAT<sup>§</sup>

**Abstract.** When applying methods of optimal control to motion planning or stabilization problems, we see that some theoretical or numerical difficulties may arise, due to the presence of specific trajectories, namely, minimizing singular trajectories of the underlying optimal control problem. In this article, we provide characterizations for singular trajectories of control-affine systems. We prove that, under generic assumptions, such trajectories share nice properties, related to computational aspects; more precisely, we show that, for a generic system—with respect to the Whitney topology—all nontrivial singular trajectories are of minimal order and of corank one. These results, established both for driftless and for control-affine systems, extend results of [Y. Chitour, F. Jean, and E. Trélat, *Comptes Rendus Math.*, 337 (2003), pp. 49–52 (in French); Y. Chitour, F. Jean, and E. Trélat, *J. Differential Geom.*, 73 (2006), pp. 45–73]. As a consequence, for generic control-affine systems (with or without drift) defined by more than two vector fields, and for a fixed cost, there do not exist minimizing singular trajectories. Besides, we prove that, given a control-affine system satisfying the Lie algebra rank condition (LARC), singular trajectories are strictly abnormal, generically with respect to the cost. We then show how these results can be used to derive regularity results for the value function and in the theory of Hamilton–Jacobi equations, which in turn have applications for stabilization and motion planning, from both theoretical and implementational points of view.

**Key words.** singular trajectory, control-affine system, genericity, optimal control

**AMS subject classifications.** 93B99, 37C20, 49J15

**DOI.** 10.1137/060663003

**1. Introduction.** When addressing standard issues of control theory such as motion planning and stabilization, one may adopt an approach based on optimal control, e.g., Hamilton–Jacobi type methods and shooting algorithms. One is then immediately facing intrinsic difficulties due to the possible presence of singular trajectories. It is therefore important to characterize these trajectories by studying, in particular, their existence, their optimality status, and the related computational aspects. In this paper, we provide solutions to the aforementioned difficulties for control-affine systems, under generic assumptions, and then investigate consequences in optimal control and its applications.

Let  $M$  be a smooth (i.e.,  $C^\infty$ ) manifold of dimension  $n$ . Consider the control-affine system

$$(\Sigma) \quad \dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x),$$

where  $x \in M$ ,  $m$  is a positive integer,  $(f_0, \dots, f_m)$  is an  $(m+1)$ -tuple of smooth vector fields on  $M$ , and the control  $u = (u_1, \dots, u_m)$  takes values in an open subset  $\Omega$  of  $\mathbb{R}^m$ . For  $x_0 \in M$  and  $T > 0$ , a control  $u \in L^\infty([0, T], \Omega)$  is said to be *admissible* if the trajectory  $x(\cdot, x_0, u)$  of  $(\Sigma)$  associated to  $u$  and starting at  $x_0$  is well defined on

\*Received by the editors June 15, 2006; accepted for publication (in revised form) May 22, 2007; published electronically March 19, 2008.

<http://www.siam.org/journals/sicon/47-2/66300.html>

<sup>†</sup>Labo. des Signaux et Systèmes, Université Paris-Sud, CNRS, Supélec, 91192 Gif-sur-Yvette cedex, France (Yacine.Chitour@lss.supelec.fr).

<sup>‡</sup>ENSTA, UMA, 32 bld Victor, 75739 Paris, France (Frederic.Jean@ensta.fr).

<sup>§</sup>Université Paris-Sud, Math., UMR 8628, Bat. 425, 91405 Orsay cedex, France (Emmanuel.Trelat@math.u-psud.fr).

$[0, T]$ . On the set  $\mathcal{U}_{x_0, T}$  of admissible controls, define the *end-point mapping* by

$$E_{x_0, T}(u) := x(T, x_0, u).$$

It is classical that  $\mathcal{U}_{x_0, T}$  is an open subset of  $L^\infty([0, T], \Omega)$  and that  $E_{x_0, T} : \mathcal{U}_{x_0, T} \rightarrow M$  is a smooth map.

DEFINITION 1.1. A control  $u \in \mathcal{U}_{x_0, T}$  is said to be *singular* if  $u$  is a critical point of the end-point mapping  $E_{x_0, T}$ ; i.e., its differential at  $u$ ,  $DE_{x_0, T}(u)$ , is not surjective. A trajectory  $x(\cdot, x_0, u)$  is said to be *singular* if  $u$  is singular and of corank one if the codimension in the tangent space of the range of  $E_{x_0, T}(u)$  is equal to one.

In other words, a control  $u \in \mathcal{U}_{x_0, T}$  is singular if the linearized system along the trajectory  $x(\cdot, x_0, u)$  is not controllable on  $[0, T]$ . Singular trajectories appear as singularities in the set of trajectories of  $(\Sigma)$  joining two given points, and hence, they play a crucial role in variational problems associated to  $(\Sigma)$  and in optimal control, as described next.

Let  $x_0$  and  $x_1$  be two points of  $M$ , and let  $T > 0$ . Consider the following optimal control problem: From among all the trajectories of  $(\Sigma)$  steering  $x_0$  to  $x_1$ , determine a trajectory minimizing the *cost*

$$(1.1) \quad C_{U, \alpha, g}(T, u) = \int_0^T \left( \frac{1}{2} u(t)^T U(x(t)) u(t) + \alpha(x(t))^T u(t) + g(t, x(t)) \right) dt,$$

where  $\alpha = (\alpha_1, \dots, \alpha_m) \in C^\infty(M, \mathbb{R}^m)$ ,  $g \in C^\infty(\mathbb{R} \times M)$ , and  $U$  takes values in the set of symmetric positive definite  $m \times m$  matrices.

According to the Pontryagin maximum principle (see [21]), for every optimal trajectory  $x(\cdot) := x(\cdot, x_0, u)$ , there exists a nonzero pair  $(\lambda(\cdot), \lambda^0)$ , where  $\lambda^0$  is a nonpositive real number and  $\lambda(\cdot)$  is an absolutely continuous function on  $[0, T]$  (called *adjoint vector*) with  $\lambda(t) \in T_{x(t)}^* M$  such that, almost everywhere on  $[0, T]$ ,

$$(1.2) \quad \begin{aligned} \dot{x}(t) &= \frac{\partial H}{\partial \lambda}(t, x(t), \lambda(t), \lambda^0, u(t)), \\ \dot{\lambda}(t) &= -\frac{\partial H}{\partial x}(t, x(t), \lambda(t), \lambda^0, u(t)), \\ \frac{\partial H}{\partial u}(t, x(t), \lambda(t), \lambda^0, u(t)) &= 0, \end{aligned}$$

where

$$H(t, x, \lambda, \lambda^0, u) := \sum_{i=1}^m u_i \langle \lambda, f_i(x) \rangle + \lambda^0 \left( \frac{1}{2} u^T U(x) u + \alpha(x)^T u + g(t, x) \right)$$

is the *Hamiltonian* of the system. An *extremal* is a 4-tuple  $(x(\cdot), \lambda(\cdot), \lambda^0, u(\cdot))$  solution of the system of equations (1.2). The extremal is said to be *normal* if  $\lambda^0 \neq 0$  and *abnormal* if  $\lambda^0 = 0$ .

The relevance of singular trajectories in optimal control lies in the fact that they are exactly the projections of abnormal extremals. Note that a singular trajectory may be the projection of several abnormal extremals, and also of a normal extremal. A singular trajectory is said to be *strictly abnormal* if it is not the projection of a normal extremal. Notice that a singular trajectory is of corank one if and only if it admits a unique (up to scalar normalization) abnormal extremal lift; it is strictly abnormal and of corank one if and only if it admits a unique extremal lift which is abnormal.

For a normal extremal, it is standard to adopt the normalization  $\lambda^0 = -1$  and to derive the control  $u$  as the feedback function of  $(x, \lambda)$ ,

$$(1.3) \quad u(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_m(t) \end{pmatrix} = U(x(t))^{-1} \begin{pmatrix} h_1(t) - \alpha_1(x(t)) \\ \vdots \\ h_m(t) - \alpha_m(x(t)) \end{pmatrix},$$

for every  $t \in [0, T]$ , where  $h_i(t) := \langle \lambda(t), f_i(x(t)) \rangle$ , for  $i = 1, \dots, m$ . In particular, normal extremals are smooth on  $[0, T]$ .

For abnormal extremals, the situation is much more involved, since system (1.2) does not provide directly an expression for abnormal controls. Abnormal extremals may be nonsmooth, and it is not always possible to determine an explicit expression for singular controls. Indeed, it follows from (1.2) that

$$(1.4) \quad h_i(\cdot) \equiv 0 \text{ on } [0, T], \quad i = 1, \dots, m,$$

along every abnormal extremal. At that point, in order to compute the singular control, one usually differentiates iteratively (1.4) with respect to  $t$  until the control appears explicitly (in an affine way). To recover the control, an invertibility property is then required, which may not hold in general.

In this paper, we prove that, in a generic context, such an invertibility property is obtained with a minimal number of differentiations (cf. Theorem 2.6). This is the concept of *minimal order*, defined in Definition 2.5. Here, genericity means that the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  belongs to an open and dense subset of the set of vector fields equipped with the Whitney topology. The corank one property is also proved to hold generically. We obtain similar results in the driftless case for generic  $m$ -tuples  $(f_1, \dots, f_m)$  (cf. Theorem 2.17).

Note that the latter result can be directly derived from [14] under the additional assumption that the  $m$ -tuples  $(f_1, \dots, f_m)$  are everywhere linearly independent. Such a geometric assumption is not adapted for control applications, e.g., whenever the state space is a product of manifolds involving a sphere of even dimension. One of the main novelties of this paper consists in dropping that assumption. As pointed out in [12], this raises serious technical difficulties, which furthermore cannot be treated by the methods of [14].

In a preliminary step for deriving the above theorems, we establish two results of independent interest, asserting that any trajectory of a generic control-affine system satisfies  $\dot{x} = 0$  almost everywhere on the set where the vector fields are linearly dependent (cf. Theorems 2.1 and 2.13).

When considering optimal control problems, we see that minimizing singular trajectories may exist, and play a major role, since they are not dependent on the specific minimization problem. The issue of such minimizing trajectories was already well known in the classical theory of calculus of variations (see, for instance, [9, 32]) and proved to be a major focus, during the 1940s, when the whole domain eventually developed into optimal control theory (cf. [10]). The optimality status of singular trajectories was chiefly investigated in [11, 30] in relation to control-affine systems with  $m = 1$ , in [1, 18, 19, 30] regarding driftless systems with  $m = 2$ , and in [2, 27] for general nonlinear control systems.

In this paper, we prove that, for generic systems with  $m \geq 2$  (and  $m \geq 3$  in the driftless case) and for a fixed cost  $C_{U,\alpha,g}$ , there does not exist minimizing singular trajectories (cf. Corollaries 2.9 and 2.20). We also prove that, given a fixed system

( $\Sigma$ ), singular trajectories are strictly abnormal, generically with respect to the cost (1.1) (cf. Propositions 2.12 and 2.22). We then show how the above mentioned results can be used to derive regularity results for the value function and in the theory of Hamilton–Jacobi equations, which in turn have applications for stabilization and motion planning.

This paper is organized as follows. Section 2 is devoted to the statement of the main results, first in the control-affine case, and second in the driftless case. The consequences are detailed in section 3, and proofs are provided in section 4.

**2. Statement of the main results.** Let  $M$  be a smooth,  $n$ -dimensional manifold. Throughout the paper,  $VF(M)$  denotes the set of smooth vector fields on  $M$ , endowed with the  $C^\infty$  Whitney topology.

**2.1. Trajectories of control-affine systems.** Let  $T$  be a positive real number. Consider the control-affine system

$$(2.1) \quad \dot{x}(t) = f_0(x(t)) + \sum_{i=1}^m u_i(t) f_i(x(t)),$$

where  $(f_0, \dots, f_m)$  is an  $(m+1)$ -tuple of smooth vector fields on  $M$ , and the set of admissible controls  $u = (u_1, \dots, u_m)$  is an open subset of  $L^\infty([0, T], \Omega)$ .

For every trajectory  $x(\cdot) := x(\cdot, x_0, u)$  of (2.1), define  $I_{\text{dep}}(x(\cdot))$  as the closed subset of  $[0, T]$ ,

$$(2.2) \quad I_{\text{dep}}(x(\cdot)) := \{t \in [0, T] \mid \text{rank}\{f_0(x(t)), \dots, f_m(x(t))\} < m+1\}.$$

Note that, on the open subset of  $\mathbb{R}^n$ , where  $\text{rank}\{f_0, \dots, f_m\} = m+1$ , there is a one-to-one correspondence between trajectories and controls. In contrast, on  $I_{\text{dep}}(x(\cdot))$ , there is no uniqueness of the control associated to  $x(\cdot)$ ; in particular,  $x(\cdot)$  may be associated to both singular and nonsingular controls. This fact emphasizes the following result, which describes, in a generic context, trajectories on the subset of  $\mathbb{R}^n$ , where  $\text{rank}\{f_0, \dots, f_m\} < m+1$ .

**THEOREM 2.1.** *Let  $m < n$  be a nonnegative integer. There exists an open and dense subset  $O_{m+1}$  of  $VF(M)^{m+1}$  so that, if the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  belongs to  $O_{m+1}$ , then every trajectory  $x(\cdot)$  of the associated control-affine system  $\dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$  verifies*

$$(2.3) \quad \dot{x}(t) = 0, \text{ for almost every } t \in I_{\text{dep}}(x(\cdot)).$$

*In addition, for every integer  $N$ , the set  $O_{m+1}$  can be chosen so that its complement has codimension greater than  $N$ .*

**Remark 2.2.** In light of the previous result, one can choose the admissible control  $u$  on  $I_{\text{dep}}(x(\cdot))$  such that, for every  $t \in I_{\text{dep}}(x(\cdot))$ ,  $u(t)$  consists of any  $m$ -tuple  $(\alpha_1, \dots, \alpha_m)$  so that

$$f_0(x(t)) + \sum_{i=1}^m \alpha_i f_i(x(t)) = 0.$$

In particular, on any subinterval of  $I_{\text{dep}}(x(\cdot))$ , the trajectory  $x(\cdot)$  is constant, and the control can be chosen constant as well.

**Remark 2.3.** A trajectory  $x(\cdot)$  is said to be *trivial* if it reduces to a point; otherwise it is said to be *nontrivial*. It is clear that, if  $I_{\text{dep}}(x(\cdot)) \neq [0, T]$ , then  $\dot{x}(t) \neq 0$  for  $t \notin I_{\text{dep}}(x(\cdot))$  and  $x(\cdot)$  is nontrivial.

Let  $x(\cdot)$  be a trajectory of a control-affine system associated to an  $(m+1)$ -tuple of  $O_{m+1}$ . As a consequence of Theorem 2.1,  $x(\cdot)$  is trivial if and only if  $I_{\text{dep}}(x(\cdot)) = [0, T]$ .

**2.2. Singular trajectories.** Recall that a singular trajectory  $x(\cdot)$  is the projection of an abnormal extremal  $(x(\cdot), \lambda(\cdot))$ . For  $t \in [0, T]$  and  $i, j \in \{0, \dots, m\}$ , we define

$$h_i(t) := \langle \lambda(t), f_i(x(t)) \rangle, \quad h_{ij}(t) := \langle \lambda(t), [f_i, f_j](x(t)) \rangle.$$

Along an abnormal extremal, we have for every  $t \in [0, T]$ ,

$$(2.4) \quad h_0(t) = \text{constant}, \quad h_i(t) = 0, \quad i = 1, \dots, m.$$

Differentiating (2.4), one gets, almost everywhere on  $[0, T]$ ,

$$(2.5) \quad h_{i0}(t) + \sum_{j=1}^m h_{ij}(t) u_j(t) = 0, \quad i \in \{0, \dots, m\}.$$

DEFINITION 2.4. *Along an abnormal extremal  $(x(\cdot), \lambda(\cdot), u(\cdot))$  of the system (2.1), the Goh matrix  $G(t)$  at time  $t \in [0, T]$  is the  $m \times m$  skew-symmetric matrix given by*

$$(2.6) \quad G(t) := (h_{ij}(t))_{1 \leq i, j \leq m}.$$

Since  $G(t)$  is skew-symmetric, rank  $G(t)$  is even, and (2.5) is rewritten as, almost everywhere on  $[0, T]$ ,

$$(2.7) \quad G(t)u(t) = b(t),$$

with  $b(t) := -(h_{i0}(t))_{1 \leq i \leq m}$ .

Note that, if  $G(t)$  is invertible, then  $u(t)$  is uniquely determined by (2.7). This only occurs for  $m$  even.

If  $m$  is odd,  $G(t)$  is never invertible. However, a similar construction is derived as follows. Define

$$(2.8) \quad \overline{G}(t) := (h_{ij}(t))_{0 \leq i, j \leq m}.$$

Since  $\overline{G}(t)$  is skew-symmetric, the determinant of  $\overline{G}(t)$  is the square of a polynomial  $\overline{P}(t)$  in the  $h_{ij}(t)$  with degree  $(m+1)/2$ , called the *Pfaffian* of  $\overline{G}(t)$  (see [6]). From (2.5),  $\overline{G}(t)$  is not invertible, and thus, along the extremal,  $\overline{P}(t) = 0$ . After differentiation, one gets, almost everywhere on  $[0, T]$ ,

$$(2.9) \quad \{\overline{P}, h_0\}(t) + \sum_{i=1}^m u_i(t) \{\overline{P}, h_i\}(t) = 0.$$

Define the  $(m+1) \times m$  matrix  $\tilde{G}(t)$  as  $G(t)$  augmented with the row  $(\{\overline{P}, h_j\}(t))_{1 \leq j \leq m}$ , and the  $(m+1)$ -dimensional vector  $\tilde{b}(t)$  as  $b(t)$  augmented with the coefficient  $-\{\overline{P}, h_0\}(t)$ . Then, from (2.7) and (2.9), there holds, almost everywhere on  $[0, T]$ ,

$$(2.10) \quad \tilde{G}(t)u(t) = \tilde{b}(t).$$

If  $\tilde{G}(t)$  is of rank  $m$ , then  $u(t)$  is uniquely determined by (2.10).

These facts, combined with Remark 2.2, motivate the following definition.

DEFINITION 2.5. *If  $m$  is even (resp., odd), a singular trajectory  $x(\cdot)$  is said to be of minimal order if*



- (i)  $\dot{x}(t) = 0$  for almost every  $t \in I_{\text{dep}}(x(\cdot))$ ;
- (ii) it admits an abnormal extremal lift such that, for almost every  $t \in [0, T] \setminus I_{\text{dep}}$ ,  
 $\text{rank } G(t) = m$  if  $m$  is even, resp.,  $\text{rank } \tilde{G}(t) = m$  if  $m$  is odd.

On the opposite side, for arbitrary  $m$ , a singular trajectory is said to be a *Goh trajectory* if it admits an abnormal extremal lift along which the Goh matrix is identically equal to 0.

**THEOREM 2.6.** *Let  $m < n$  be a positive integer. There exists an open and dense subset  $O_{m+1}$  of  $VF(M)^{m+1}$  so that, if the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  belongs to  $O_{m+1}$ , then every nontrivial singular trajectory of the associated control-affine system  $\dot{x}(t) = f_0(x(t)) + \sum_{i=1}^m u_i(t)f_i(x(t))$  is of minimal order and of corank one. In addition, for every integer  $N$ , the set  $O_{m+1}$  can be chosen so that its complement has codimension greater than  $N$ .*

**COROLLARY 2.7.** *With the notation of Theorem 2.6 and if  $m \geq 2$ , there exists an open and dense subset  $O_{m+1}$  of  $VF(M)^{m+1}$  so that every control-affine system defined with an  $(m+1)$ -tuple of  $O_{m+1}$  does not admit nontrivial Goh singular trajectories.*

**2.3. Minimizing singular trajectories.** We keep here the notation of the previous sections. Consider the control-affine system

$$(2.11) \quad \dot{x}(t) = f_0(x(t)) + \sum_{i=1}^m u_i(t)f_i(x(t)),$$

and the quadratic cost given by

$$(2.12) \quad C_{U,g}(T, u) = \frac{1}{2} \int_0^T \left( u(t)^T U(x(t)) u(t) + g(t, x(t)) \right) dt,$$

where  $U \in \mathcal{S}_m^+(M)$  and  $g \in C^\infty(\mathbb{R} \times M)$ . Here,  $\mathcal{S}_m^+(M)$  denotes the set of smooth mappings  $x \mapsto U(x)$  on  $M$ , taking values in the set  $\mathcal{S}_m^+$  of  $m \times m$  real positive definite matrices.

For  $x_0 \in M$  and  $T > 0$ , define the optimal control problem

$$(2.13) \quad \inf \{ C_{U,g}(T, u) \mid E_{x_0, T}(u) = x \}.$$

We next state two sets of genericity results, which depend, resp., on whether the cost or the control system is fixed.

### 2.3.1. Genericity with respect to the control system, with a fixed cost.

**PROPOSITION 2.8.** *Fix  $U \in \mathcal{S}_m^+(M)$  and  $g \in C^\infty(\mathbb{R} \times M)$ . There exists an open and dense subset  $O_{m+1}$  of  $VF(M)^{m+1}$  such that every nontrivial singular trajectory of a control-affine system defined by an  $(m+1)$ -tuple of  $O_{m+1}$  is strictly abnormal for the optimal control problem (2.13).*

Corollary 2.7, together with Proposition 2.8, yields the next corollary.

**COROLLARY 2.9.** *Fix  $U \in \mathcal{S}_m^+(M)$  and  $g \in C^\infty(\mathbb{R} \times M)$ . Let  $m \geq 2$  be an integer. There exists an open and dense subset  $O_{m+1}$  of  $VF(M)^{m+1}$  so that the optimal control problem (2.13) defined with an  $(m+1)$ -tuple of  $O_{m+1}$  does not admit nontrivial minimizing singular trajectories.*

**Remark 2.10.** In both previous results, the set  $O_{m+1}$  can be chosen so that its complement has an arbitrary codimension.

### 2.3.2. Genericity with respect to the cost, with a fixed control system.

We endow  $\mathcal{S}_m^+(M)$  with the Whitney topology. An  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  of

$VF(M)^{m+1}$  is said to verify the *Lie algebra rank condition* (LARC) if the Lie algebra generated by  $f_0, \dots, f_m$  is of dimension  $n$  at every point of  $M$ .

**PROPOSITION 2.11.** *Fix  $(f_0, \dots, f_m) \in VF(M)^{m+1}$  so that the LARC is satisfied and the zero control  $u \equiv 0$  is not singular. Let  $g \in C^\infty(\mathbb{R} \times M)$ . Then, there exists an open and dense subset  $\mathcal{A}_m$  of  $\mathcal{S}_m^+(M)$  such that every nontrivial singular trajectory of the control-affine system associated to the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  is strictly abnormal for the optimal control problem (2.13) defined with  $U \in \mathcal{A}_m$  and  $g$ .*

Assuming that the zero control  $u \equiv 0$  is not singular is a necessary hypothesis. Indeed, the fact that a control  $u$  is singular is a property of the sole  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  and is independent of the cost. On the other hand, every trajectory  $x := x(\cdot, x_0, 0)$  associated to the zero control is always the projection of the normal extremal  $(x(\cdot), 0, -1, 0)$  of any optimal control problem (2.13). As a consequence, if the zero control is singular, such a trajectory  $x(\cdot, x_0, 0)$  cannot be strictly abnormal.

In order to handle the case of a singular zero control, it is therefore necessary to consider more general quadratic costs such as

$$(2.14) \quad C_{U,\alpha,g}(T, u) = \int_0^T \left( \frac{1}{2} u(t)^T U(x(t)) u(t) + \alpha(x(t))^T u(t) + g(t, x(t)) \right) dt,$$

where  $U \in \mathcal{S}_m^+(M)$ ,  $\alpha \in C^\infty(M, \mathbb{R}^m)$  and  $g \in C^\infty(\mathbb{R} \times M)$ .

**PROPOSITION 2.12.** *Fix  $(f_0, \dots, f_m) \in VF(M)^{m+1}$  satisfying the LARC and  $g \in C^\infty(\mathbb{R} \times M)$ . Then, there exists an open and dense subset  $\mathcal{B}_m$  of  $\mathcal{S}_m^+(M) \times C^\infty(M, \mathbb{R}^m)$  such that every nontrivial singular trajectory of the control-affine system associated to the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  is strictly abnormal for the optimal control problem (2.11)–(2.14) defined with  $(U, \alpha) \in \mathcal{B}_m$  and  $g$ .*

**2.4. Driftless control-affine systems.** Let  $T$  be a positive real number. Consider the driftless control-affine system

$$(2.15) \quad \dot{x}(t) = \sum_{i=1}^m u_i(t) f_i(x(t)),$$

where  $(f_1, \dots, f_m)$  is an  $m$ -tuple of smooth vector fields on  $M$ , and the set of admissible controls  $u = (u_1, \dots, u_m)$  is an open subset of  $L^\infty([0, T], \Omega)$ .

For every trajectory  $x(\cdot) := x(\cdot, x_0, u)$  of (2.1), define  $I_{\text{dep}}(x(\cdot))$  as the closed subset of  $[0, T]$ ,

$$I_{\text{dep}}(x(\cdot)) := \{t \in [0, T] \mid \text{rank}\{f_1(x(t)), \dots, f_m(x(t))\} < m\}.$$

**THEOREM 2.13.** *Let  $m \leq n$  be a positive integer. There exists an open and dense subset  $O_m$  of  $VF(M)^m$  so that, if the  $m$ -tuple  $(f_1, \dots, f_m)$  belongs to  $O_m$ , then every trajectory  $x(\cdot)$  of the associated driftless control-affine system  $\dot{x} = \sum_{i=1}^m u_i f_i(x)$  verifies*

$$\dot{x}(t) = 0 \text{ for almost every } t \in I_{\text{dep}}(x(\cdot)).$$

*In addition, for every integer  $N$ , the set  $O_m$  can be chosen so that its complement has codimension greater than  $N$ .*

**Remark 2.14.** As a consequence, one can simply choose the admissible control  $u$  on  $I_{\text{dep}}(x(\cdot))$  such that, for every  $t \in I_{\text{dep}}(x(\cdot))$ ,  $u(t) = 0$ . This choice induces a one-to-one correspondence between trajectories and controls.

**2.4.1. Singular trajectories.** Let  $x(\cdot)$  be a singular trajectory; it is the projection of an abnormal extremal  $(x(\cdot), \lambda(\cdot))$ . Similarly to the previous section, we define, for  $t \in [0, T]$  and  $i, j \in \{1, \dots, m\}$ ,

$$h_i(t) := \langle \lambda(t), f_i(x(t)) \rangle, \quad h_{ij}(t) := \langle \lambda(t), [f_i, f_j](x(t)) \rangle.$$

For every  $t \in [0, T]$ ,

$$(2.16) \quad h_i(t) = 0, \quad i = 1, \dots, m.$$

Differentiating (2.16), one gets, almost everywhere on  $[0, T]$ ,

$$(2.17) \quad \sum_{j=1}^m h_{ij}(t) u_j(t) = 0, \quad i \in \{1, \dots, m\}.$$

DEFINITION 2.15. *Along an abnormal extremal  $(x(\cdot), \lambda(\cdot), u(\cdot))$  of the system (2.1), the Goh matrix  $G(t)$  at time  $t \in [0, T]$  is the  $m \times m$  skew-symmetric matrix given by*

$$(2.18) \quad G(t) := (h_{ij}(t))_{1 \leq i, j \leq m}.$$

Since  $G(t)$  is skew-symmetric,  $\text{rank } G(t)$  is even, and (2.17) is rewritten, almost everywhere on  $[0, T]$ , as

$$(2.19) \quad G(t)u(t) = 0.$$

Note that, if  $\text{rank } G(t) = m - 1$ , one can deduce from (2.19) an expression for  $u(t)$ , up to time reparameterization. This only occurs for  $m$  odd.

If  $m$  is even,  $\text{rank } G(t)$  is always smaller than  $m - 1$ . However, a similar construction is derived as follows. The determinant of  $G(t)$  is the square of the Pfaffian  $P(t)$ , and, along the extremal,  $P(t) \equiv 0$ . After differentiation, one gets, almost everywhere on  $[0, T]$ ,

$$(2.20) \quad \sum_{i=1}^m u_i(t) \{P, h_i\}(t) = 0.$$

Define the  $(m+1) \times m$  matrix  $\tilde{G}(t)$  as  $G(t)$  augmented with the row  $(\{P, h_j\}(t))_{1 \leq j \leq m}$ . Then, from (2.19) and (2.20), there holds, almost everywhere on  $[0, T]$ ,

$$(2.21) \quad \tilde{G}(t)u(t) = 0.$$

If  $\tilde{G}(t)$  is of rank  $m - 1$ , one can deduce from (2.21) an expression for  $u(t)$ , up to time reparameterization.

DEFINITION 2.16. *If  $m$  is odd (resp., even), a singular trajectory  $x(\cdot)$  is said to be of minimal order if*

- (i)  $\dot{x}(t) = 0$  for almost every  $t \in I_{\text{dep}}(x(\cdot))$ ;
- (ii) it admits an abnormal extremal lift such that, for almost every  $t \in [0, T] \setminus I_{\text{dep}}$ ,  $\text{rank } G(t) = m - 1$  if  $m$  is odd, resp.,  $\text{rank } \tilde{G}(t) = m - 1$  if  $m$  is even.

On the opposite side, for arbitrary  $m$ , a singular trajectory is said to be a *Goh trajectory* if it admits an abnormal extremal lift along which the Goh matrix is identically equal to 0.

**THEOREM 2.17.** *Let  $m$  be an integer such that  $2 \leq m \leq n$ . There exists an open and dense subset  $O_m$  of  $VF(M)^m$  so that, if the  $m$ -tuple  $(f_1, \dots, f_m)$  belongs to  $O_m$ , then every nontrivial singular trajectory of the associated driftless control-affine system  $\dot{x}(t) = \sum_{i=1}^m u_i(t)f_i(x(t))$  is of minimal order and of corank one. In addition, for every integer  $N$ , the set  $O_m$  can be chosen so that its complement has codimension greater than  $N$ .*

**COROLLARY 2.18.** *With the notation of Theorem 2.17 and if  $m \geq 3$ , there exists an open and dense subset  $O_m$  of  $VF(M)^m$  so that every driftless control-affine system defined with an  $m$ -tuple of  $O_m$  does not admit nontrivial Goh singular trajectories.*

**2.4.2. Minimizing singular trajectories.** Consider the optimal control problem associated to the driftless control-affine system

$$(2.22) \quad \dot{x}(t) = \sum_{i=1}^m u_i(t)f_i(x(t)),$$

with the quadratic cost given by

$$(2.23) \quad C_{U,g}(T, u) = \frac{1}{2} \int_0^T \left( u(t)^T U(x(t)) u(t) + g(t, x(t)) \right) dt,$$

where  $U \in \mathcal{S}_m^+(M)$  and  $g \in C^\infty(\mathbb{R} \times M)$ .

For  $x_0 \in M$  and  $T > 0$ , define the optimal control problem

$$(2.24) \quad \inf \{ C_{U,g}(T, u) \mid E_{x_0, T}(u) = x \}.$$

We next state genericity results with respect to the control system, with a fixed cost.

**PROPOSITION 2.19.** *Fix  $U \in \mathcal{S}_m^+(M)$  and  $g \in C^\infty(\mathbb{R} \times M)$ . There exists an open and dense subset  $O_m$  of  $VF(M)^m$  such that every nontrivial singular trajectory of a driftless control-affine system defined by an  $m$ -tuple of  $O_m$  is strictly abnormal for the optimal control problem (2.24).*

Corollary 2.18, together with Proposition 2.19, yields the next corollary.

**COROLLARY 2.20.** *Fix  $U \in \mathcal{S}_m^+(M)$  and  $g \in C^\infty(\mathbb{R} \times M)$ . Let  $m \geq 3$  be an integer. There exists an open and dense subset  $O_m$  of  $VF(M)^m$  so that the optimal control problem (2.24) defined with an  $m$ -tuple of  $O_m$  does not admit nontrivial minimizing singular trajectories.*

**Remark 2.21.** In both previous results, the set  $O_m$  can be chosen so that its complement has an arbitrary codimension.

We also have a genericity result with respect to the cost, with a fixed control system.

**PROPOSITION 2.22.** *Fix  $(f_1, \dots, f_m) \in VF(M)^m$  so that the LARC is satisfied. Let  $g \in C^\infty(\mathbb{R} \times M)$ . Then, there exists an open and dense subset  $\mathcal{A}_m$  of  $\mathcal{S}_m^+(M)$  such that every nontrivial singular trajectory of the driftless control-affine system associated to the  $m$ -tuple  $(f_1, \dots, f_m)$  is strictly abnormal for the optimal control problem (2.24) defined with  $U \in \mathcal{A}_m$  and  $g$ .*

**Remark 2.23.** In the driftless case, the control  $u \equiv 0$  is always singular but corresponds to a trivial trajectory. Therefore, in opposition to the control-affine case, it is not necessary to add the linear term  $\alpha(x)^T u$  in the cost.

### 3. Consequences.

**3.1. Regularity of the value function.** Consider the optimal control problem (2.13), associated to the control-affine system (2.11) and the cost (2.12). The value function is defined by

$$(3.1) \quad S_{x_0,T}(x) := \inf\{C_{U,g}(T, u) \mid E_{x_0,T}(u) = x\}$$

for every  $x \in \mathbb{R}^n$  (with, as usual,  $\inf \emptyset := -\infty$ ). We assume in what follows that all data are analytic.

The regularity of  $S_{x_0,T}$  is closely related to the existence of nontrivial minimizing singular trajectories starting from  $x_0$ . It is proved in [29] that, in the absence of minimizing singular trajectories, the value function is continuous and subanalytic (see, e.g., [16] for a definition of a subanalytic function). For driftless control-affine systems and  $g \equiv 0$ , the value function coincides with the square of a sub-Riemannian distance (see [7] for an introduction to sub-Riemannian geometry). In particular, in this case, the value function is always continuous, but the trivial trajectory  $x(\cdot) \equiv x_0$  is always minimizing and singular. Moreover, if there is no nontrivial minimizing singular trajectories, then the value function is subanalytic outside  $x_0$  (see [3, 4]). This situation holds for generic distributions of rank greater than or equal to three (see [5, 14]).

The results of section 2.3 have the following consequence on the regularity of  $S_{x_0,T}$ .

**COROLLARY 3.1.** *With the notation of Corollary 2.9, and if in addition the functions  $g$ ,  $U$  and the vector fields of the  $(m+1)$ -tuple in  $O_{m+1}$  are analytic, then the associated value function  $S_{x_0,T}$  is continuous and subanalytic on its domain of definition.*

**Remark 3.2.** If there exists a nontrivial minimizing singular trajectory, the value function may fail to be subanalytic or even continuous. For example, consider the control-affine system in  $\mathbb{R}^2$  given by

$$\dot{x}(t) = 1 + y(t)^2, \quad \dot{y}(t) = u(t),$$

and the cost  $C(T, u) = \int_0^T u(t)^2 dt$ . The trajectory  $(x(t) = t, y(t) = 0)$ , associated to the control  $u = 0$ , is a nontrivial minimizing singular trajectory, and the value function  $S_{(0,0),T}$  has the asymptotic expansion, near the point  $(T, 0)$ ,

$$S_{(0,0),T}(x, y) = \frac{1}{4} \frac{y^4}{x - T} + \frac{y^4}{x - T} \exp\left(-\frac{y^2}{x - T}\right) + o\left(\frac{y^4}{x - T} \exp\left(-\frac{y^2}{x - T}\right)\right)$$

(see [29] for details). Hence, it is neither continuous nor subanalytic at the point  $(T, 0)$ .

In the driftless control-affine case, by using the results of section 2.4.2, we derive the following similar consequence.

**COROLLARY 3.3.** *With the notation of Corollary 2.20, and if in addition the functions  $g$ ,  $U$  and the vector fields of the  $m$ -tuple in  $O_m$  are analytic, then the associated value function  $S_{x_0,T}$  is subanalytic outside  $x_0$ .*

### 3.2. Regularity of viscosity solutions of Hamilton–Jacobi equations.

Assume that the assumptions of the previous subsection hold. It is standard (see [15, 17]) that the value function  $v(t, x) = S_{x_0,t}(x)$  is a viscosity solution of the

Hamilton–Jacobi equation

$$(3.2) \quad \frac{\partial v}{\partial t} + \mathcal{H}\left(x, \frac{\partial v}{\partial x}\right) = g(t, x),$$

where  $\mathcal{H}(x, p) = \langle p, f_0(x) \rangle + \frac{1}{2} \sum_{i,j=1}^m (U^{-1}(x))_{ij} \langle p, f_i(x) \rangle \langle p, f_j(x) \rangle$ .

Conversely, the viscosity solution of (3.2) with analytic Dirichlet-type conditions is subanalytic, as soon as the corresponding optimal control problem does not admit minimizing singular trajectories (see [31]). Using the results of the previous sections, this situation holds generically if  $m \geq 2$  (and, similarly for driftless control-affine systems, if  $m \geq 3$ ).

As a consequence, the analytic singular set  $\text{Sing}(v)$  of the viscosity solution  $v$ , i.e., the subset of  $\mathbb{R}^n$  where  $v$  is not analytic, is a (subanalytic) stratified manifold of codimension greater than or equal to one (see [28] for more details on the subject). Since  $\text{Sing}(v)$  is also the locus where characteristic curves intersect, the above mentioned property turns out to be instrumental for the global convergence of numerical schemes for (3.2) (see [15]). Indeed, the analytic singular set must be as “nice” as possible in order to integrate energy functions on the set of characteristic curves.

**3.3. Applications to stabilization and motion planning.** For a driftless control-affine system verifying the LARC, there exist general stabilizing strategies stemming from dynamic programming. As usual, the stabilizing feedback is computed using the gradient of the value function  $S$  for a suitable optimal control problem. Of course this is only possible outside the singular set  $\text{Sing}(S)$ , and one must devise another construction for the feedback on  $\text{Sing}(S)$ . Let us mention two such strategies, the first one providing a hybrid feedback (see [22]) and the second one a smooth repulsive stabilizing (SRS) feedback (see [23, 24]). Both strategies crucially rely on the fact that  $\text{Sing}(S)$  is a stratified manifold of codimension greater than or equal to one.

As seen before, the latter fact holds generically in the analytic category for  $m \geq 3$ .

On the other hand, the absence of minimizing singular trajectories is the basic requirement for the convergence of usual algorithms in optimal control (such as direct or indirect methods; see, e.g., [8, 20]). We have proved that this situation holds generically for control-affine systems if  $m \geq 2$  and for driftless control-affine systems if  $m \geq 3$ .

As a final application, consider a driftless control-affine system verifying the LARC. According to Proposition 2.22, it is possible to choose a (generic) cost function  $C_{U,g}$  such that all singular trajectories are strictly abnormal. Combining that fact with [25, Theorem 1.1], we deduce that there exists a dense subset  $N$  of  $\mathbb{R}^n$  such that every point of  $N$  is reached by a unique minimizing trajectory, which is, moreover, nonsingular. As a consequence, a shooting method with a target in  $N$  will converge. That fact may be used for solving (at least approximately) motion planning problems.

## 4. Proofs of the results.

**4.1. Proofs of Theorems 2.1 and 2.13.** Every trajectory of the control-affine system  $\dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$  is also a trajectory of the driftless control system  $\dot{x} = \sum_{i=0}^m u_i f_i(x)$ , with  $u_0 \equiv 1$ . Therefore, Theorem 2.1 follows from Theorem 2.13, whose proof is provided next.

Let  $x(\cdot) = x(\cdot, x_0, u)$  be a trajectory of the driftless control system  $\dot{x} = \sum_{i=1}^m u_i f_i(x)$ , with  $2 \leq m \leq n$ . Consider the set  $I_{\text{dep}}(x(\cdot))$  defined by (2.2). We argue by contraposition and assume that  $I_{\text{dep}}(x(\cdot))$  contains a subset  $I$  of positive

measure such that  $\dot{x}(t) \neq 0$  for  $t \in I$ . Since Lebesgue points of  $u$  are of full Lebesgue measure, we assume that  $u$  is continuous on  $I$ .

Up to considering a subset of  $I$ , and relabeling the  $f_i$ 's, we assume that, for every  $t \in I$ , that

- (i) there exists  $1 \leq k < m$  such that

$$\text{rank}\{f_1(x(t)), \dots, f_m(x(t))\} = k;$$

- (ii)  $f_1(x(t)), \dots, f_k(x(t))$  are linearly independent, and thus, there exist real numbers  $\alpha_i^j(t)$ ,  $i = 1, \dots, k$ ,  $j = k+1, \dots, m$ , such that

$$f_j(x(t)) = \sum_{i=1}^k \alpha_i^j(t) f_i(x(t)), \quad j = k+1, \dots, m.$$

Therefore,  $\dot{x}(t) = \sum_{i=1}^k \delta_i(t) f_i(x(t))$ , where  $\delta_i(t) := u_i(t) + \sum_{j=k+1}^m \alpha_i^j(t) u_j(t)$ ;

- (iii)  $\delta_1(t) \neq 0$ .

*Remark 4.1.* Up to reducing  $I$ , we furthermore assume that  $I$  is contained in an open interval  $\mathcal{I}$  on which  $\text{rank}\{f_1(x(t)), \dots, f_k(x(t))\} = k$ .

Set  $\text{ad}^0 g(h) = h$ , where  $g, h \in \text{VF}(M)$ , and  $\text{ad}^k g(h) = [g, \text{ad}^{k-1} g(h)]$  for  $k \geq 1$ . The length of the iterated Lie bracket  $[f_{i_1}, [f_{i_2}, [\dots, f_{i_k}]\dots]]$  of  $f_1, \dots, f_m$  is the integer  $k$ .

**PROPOSITION 4.2.** *Let  $N$  be a positive integer. There exists a subset  $J_N \subset I$  of positive measure such that, for every  $t \in J_N$  and every  $\ell \in \{1, \dots, N\}$ ,*

$$(4.1) \quad \delta_1(t)^{\ell-1} \text{ad}^{\ell-1} f_1(f_m)(x(t)) = h_t^\ell(x(t)) + R_t^\ell(x(t)),$$

where

- $h_t^\ell(x(t)) \in \text{Span}\{f_1(x(t)), \dots, f_k(x(t))\}$ ;
- $R_t^\ell$  is a linear combination of iterated Lie brackets of  $f_1, \dots, f_m$ , of length smaller than  $\ell-1$ , and of iterated Lie brackets of  $f_1, \dots, f_k$ , of length smaller than or equal to  $\ell$ .

*Proof.* For  $t \in I$ , let  $F_t \in \text{VF}(M)$  be the vector field defined by

$$F_t(x) := \sum_{i=1}^k \delta_i(t) f_i(x).$$

Notice that  $\dot{x}(t) = F_t(x(t))$  for  $t \in I$ . For the argument of Proposition 4.2, we need the following lemma.

**LEMMA 4.3.** *Consider a set  $J \subset I$  of positive measure and  $h \in \text{VF}(M)$  so that  $h(x(t)) \in \text{Span}\{f_1(x(t)), \dots, f_k(x(t))\}$  on  $J$ ; i.e., for every  $t \in J$ , there exist real numbers  $\beta_i(t)$ ,  $i = 1, \dots, k$ , such that*

$$(4.2) \quad h(x(t)) = \sum_{i=1}^k \beta_i(t) f_i(x(t)).$$

For  $t \in J$ , define  $g_t \in \text{VF}(M)$  by

$$g_t(x) := h(x) - \sum_{i=1}^k \beta_i(t) f_i(x).$$

Then, there exists a set  $J' \subset J$  of positive measure such that

$$(4.3) \quad [F_t, g_t](x(t)) \in \text{Span}\{f_1(x(t)), \dots, f_k(x(t))\} \quad \text{on } J'.$$

*Proof of Lemma 4.3.* Using Remark 4.1, we see that there exist  $e_j \in VF(M)$ ,  $k+1 \leq j \leq n$ , so that, for every  $t \in \mathcal{I}$ , the vectors  $f_1(x(t)), \dots, f_k(x(t)), e_{k+1}(x(t)), \dots, e_n(x(t))$  span  $T_{x(t)}M$ . Thus, there exist  $n$  smooth functions  $b_i$ ,  $1 \leq i \leq n$ , defined on  $M$ , such that

$$h(x) = \sum_{i=1}^k b_i(x) f_i(x) + \sum_{i=k+1}^n b_i(x) e_i(x),$$

for  $x$  in an open neighborhood of  $x(\mathcal{I})$ . For  $i = 1, \dots, n$ , define  $\beta_i(t) := b_i(x(t))$  for  $t \in \mathcal{I}$  (this notation is consistent with (4.2)). The  $\beta_i$ 's are absolutely continuous on  $\mathcal{I}$  and differentiable everywhere on  $J$ . For  $i = k+1, \dots, n$ , there holds  $\beta_i \equiv 0$  on  $J$  and therefore, it follows that  $\dot{\beta}_i \equiv 0$  on a subset  $J' \subset J$  of full measure (cf. [26, Lemma p. 177]).

For  $t \in J$ , using that  $g_t(x(t)) = 0$  and  $F_t(x(t)) = \dot{x}(t)$ , it holds that

$$\begin{aligned} [F_t, g_t](x(t)) &= dg_t \circ F_t(x(t)) \\ &= \sum_{i=1}^k (db_i(x(t)) \cdot \dot{x}(t)) f_i(x(t)) + \sum_{i=k+1}^n (db_i(x(t)) \cdot \dot{x}(t)) e_i(x(t)) \\ &= \sum_{i=1}^k \dot{\beta}_i(t) f_i(x(t)) + \sum_{i=k+1}^n \dot{\beta}_i(t) e_i(x(t)). \end{aligned}$$

On  $J'$ , the second sum of the right-hand side of the last equation vanishes, and the lemma follows.  $\square$

Applying Lemma 4.3 to  $h = f_m$  and  $J = I$ , we get

$$[F_t, g_t^1](x(t)) \in \text{Span}\{f_1(x(t)), \dots, f_k(x(t))\} \quad \text{on } J_1,$$

where  $J_1 \subset I$  and  $g_t^1 := f_m - \sum_{i=1}^k \alpha_i^m(t) f_i$ .

Set  $h_t^1 = [F_t, g_t^1]$ . We next iterate the above procedure for  $1 \leq \ell \leq N$ . Assume that the vector fields  $h_t^\ell$ ,  $g_t^\ell$  and the set  $J_\ell$  of positive measure are defined such that  $h_t^\ell(x(t)) \in \text{Span}\{f_1(x(t)), \dots, f_k(x(t))\}$  on  $J_\ell$ . For every  $t \in J_\ell$ , let  $\beta_i^\ell(t)$ ,  $i = 1, \dots, k$ , be the real numbers such that

$$h_t^\ell(x(t)) = \sum_{i=1}^k \beta_i^\ell(t) f_i(x(t)),$$

and define  $g_t^{\ell+1} \in VF(M)$  by  $g_t^{\ell+1} := h_t^\ell - \sum_{i=1}^k \beta_i^\ell(t) f_i$ . Set  $h_t^{\ell+1} := [F_t, g_t^{\ell+1}]$ . Applying Lemma 4.3, there exists a subset  $J_{\ell+1} \subset J_\ell$  of positive measure such that  $h_t^{\ell+1}(x(t)) \in \text{Span}\{f_1(x(t)), \dots, f_k(x(t))\}$  on  $J_{\ell+1}$ .

For  $t \in J_N$ , and for  $\ell = 1, \dots, N$ , we express  $h_t^\ell(x(t))$  using iterated Lie brackets of  $f_1, \dots, f_m$ , and an easy induction yields (4.1).  $\square$

Combining Proposition 4.2 with routine transversality arguments (see, for instance, [12] and [14]), it follows that the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  belongs to a closed subset of  $VF(M)^{m+1}$  of codimension greater than or equal to  $N$ . Theorem 2.13 follows.

*Remark 4.4.* The fact that  $f_1(x(t)) \neq 0$  is essential in order to derive, from (4.1), an infinite number of independent relations, and then to apply the above mentioned transversality arguments.



**4.2. Proof of Theorem 2.6.** The minimal order and corank one properties are proved separately in the following lemmas.

LEMMA 4.5. *There exists an open and dense subset  $O_{m+1}^1$  of  $VF(M)^{m+1}$  so that, if the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  belongs to  $O_{m+1}^1$ , then every singular trajectory of the associated control-affine system  $\dot{x}(t) = f_0(x(t)) + \sum_{i=1}^m u_i(t)f_i(x(t))$  is of minimal order. In addition, for every integer  $N$ , the set  $O_{m+1}^1$  can be chosen so that its complement has codimension greater than  $N$ .*

LEMMA 4.6. *There exists an open and dense subset  $O_{m+1}$  of  $O_{m+1}^1$  so that, if the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  belongs to  $O_{m+1}$ , then every nontrivial singular trajectory of the associated control-affine system  $\dot{x}(t) = f_0(x(t)) + \sum_{i=1}^m u_i(t)f_i(x(t))$  is of corank one. In addition, for every integer  $N$ , the set  $O_{m+1}$  can be chosen so that its complement has codimension greater than  $N$ .*

The conclusion of Theorem 2.6 follows.

**4.2.1. Proof of Lemma 4.5.** From Theorem 2.1, there exists an open and dense subset  $O_{m+1}^{11}$  of  $VF(M)^{m+1}$  such that, if  $(f_0, \dots, f_m) \in O_{m+1}^{11}$ , then every trajectory  $x(\cdot)$  of  $\dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$  verifies item (i) of Definition 2.5.

It is therefore enough to show the existence of an open and dense subset  $O_{m+1}^{12}$  of  $VF(M)^{m+1}$  such that, if  $(f_0, \dots, f_m) \in O_{m+1}^{12}$ , then every singular trajectory  $x(\cdot)$  of  $\dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$  verifies item (ii) of Definition 2.5. Then, by choosing  $O_{m+1}^1 := O_{m+1}^{11} \cap O_{m+1}^{12}$ , the conclusion of Lemma 4.5 follows.

Consider a singular trajectory  $x(\cdot) := x(\cdot, x_0, u)$  of  $\dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$ , admitting an abnormal extremal  $(x(\cdot), \lambda(\cdot))$ . Assume that there exists  $J \subset [0, T] \setminus I_{\text{dep}}(x(\cdot))$  of positive measure such that  $G(t)$  is not of rank  $m$  if  $m$  is even, resp.,  $\tilde{G}(t)$  is not of rank  $m$  if  $m$  is odd. We will show that the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  belongs to a subset of arbitrary codimension in  $VF(M)^{m+1}$  whose complement contains an open and dense subset.

Note that, on  $[0, T] \setminus I_{\text{dep}}(x(\cdot))$ , the vector fields  $f_0(x(t)), \dots, f_m(x(t))$  are linearly independent. The remaining part of the argument consists of reformulating the problem in order to follow the chain of arguments in the proof of [14, Theorem 2.4] concerning the case of everywhere linearly independent vector fields. For that purpose, we distinguish the cases  $m$  even and  $m$  odd.

Assume first that  $m$  is even. As in (2.8), define, for  $t \in J$ ,  $\overline{G}(t) := (h_{ij}(t))_{0 \leq i, j \leq m}$ . From (2.7), we have, for  $t \in J$ ,

$$\overline{G}(t) = \begin{pmatrix} 0 & (G(t)u(t))^T \\ -G(t)u(t) & G(t) \end{pmatrix}.$$

Since the ranks of both  $\overline{G}(t)$  and  $G(t)$  are even, they must be equal, for  $t \in J$ , and hence, the rank of  $\overline{G}(t)$  is smaller than  $m$  on  $J$ . This is exactly the starting point of the proof of [14, Lemma 3.8]. The machinery of [14] then applies and we deduce that the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  belongs to a subset of arbitrary codimension in  $VF(M)^{m+1}$  whose complement contains an open and dense subset  $O_{m+1}^2$  of  $VF(M)^{m+1}$ .

Assume next that  $m$  is odd. Define the  $(m+2) \times (m+1)$  matrix  $\widehat{G}(t)$  as  $\overline{G}(t)$  augmented in the last row with  $(\{P, h_j\}(t))_{0 \leq j \leq m}$ .

LEMMA 4.7. *With the notation above,  $\text{rank } \widehat{G}(t) \leq \text{rank } \tilde{G}(t) + 1$ .*

*Proof.* It amounts to showing that  $\xi \in \ker \tilde{G}(t)$  implies  $(0, \xi) \in \ker \widehat{G}(t)$ . This follows from the fact that if  $\tilde{G}(t)\xi = 0$ , then  $G(t)\xi = 0$ , and thus  $\xi$  is orthogonal to the range of  $G(t)$  since  $G(t)$  is skew-symmetric.  $\square$

Using Lemma 4.7, we see that the rank of  $\widehat{G}(t)$  is less than  $m + 1$  on  $J$ . This is exactly the starting point of the proof of [14, Lemma 3.9]. The machinery of [14] then applies and we deduce that the  $(m + 1)$ -tuple  $(f_0, \dots, f_m)$  belongs to a subset of arbitrary codimension in  $VF(M)^{m+1}$  whose complement contains an open and dense subset  $O_{m+1}^{12}$  of  $VF(M)^{m+1}$ .

**4.2.2. Proof of Lemma 4.6.** We argue by contraposition. Consider a nontrivial singular trajectory  $x(\cdot) := x(\cdot, x_0, u)$  of  $\dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$ , with  $(f_0, \dots, f_m) \in O_{m+1}^1$ . Assume  $x(\cdot)$  admits two abnormal extremal lifts  $(x(\cdot), \lambda^{[1]}(\cdot))$  and  $(x(\cdot), \lambda^{[2]}(\cdot))$  such that, for some  $t_0 \in [0, T]$ ,  $\lambda^{[1]}(t_0)$  and  $\lambda^{[2]}(t_0)$  are linearly independent. By linearity,  $\lambda^{[1]}(\cdot)$  and  $\lambda^{[2]}(\cdot)$  are linearly independent everywhere on  $[0, T]$ . Since  $x(\cdot)$  is nontrivial, it follows from Remark 2.3 that there exists a nonempty subinterval  $J$  of  $[0, T] \setminus I_{\text{dep}}(x(\cdot))$ . We are now in a position to exactly follow the arguments of [14] corresponding to the corank one property, i.e., [14, Lemma 4.4].

**4.3. Proof of Theorem 2.17.** We start with the proof of the statement dealing with the minimal order property.

From Theorem 2.13, there exists an open and dense subset  $O_m^1$  of  $VF(M)^m$  such that, if  $(f_1, \dots, f_m) \in O_m^1$ , then every trajectory  $x(\cdot)$  of  $\dot{x} = \sum_{i=1}^m u_i f_i(x)$  verifies item (i) of Definition 2.16.

It is therefore enough to show the existence of an open and dense subset  $O_m^2$  of  $VF(M)^m$  such that, if  $(f_1, \dots, f_m) \in O_m^2$ , then every singular trajectory  $x(\cdot)$  of  $\dot{x} = \sum_{i=1}^m u_i f_i(x)$  verifies item (ii) of Definition 2.16. Then, by choosing  $O_m := O_m^1 \cap O_m^2$ , the statement dealing with the minimal order property in Theorem 2.17 follows.

Consider a singular trajectory  $x(\cdot) := x(\cdot, x_0, u)$  of  $\dot{x} = \sum_{i=1}^m u_i f_i(x)$  admitting an abnormal extremal  $(x(\cdot), \lambda(\cdot))$ . Assume that there exists  $J \subset [0, T] \setminus I_{\text{dep}}(x(\cdot))$  of positive measure such that  $G(t)$  is not of rank  $m - 1$  if  $m$  is odd, resp.,  $\widetilde{G}(t)$  is not of rank  $m - 1$  if  $m$  is even. Following exactly the proofs of Lemmas 3.8 and 3.9 in [14], the  $m$ -tuple  $(f_1, \dots, f_m)$  belongs to a subset of arbitrary codimension in  $VF(M)^m$  whose complement contains an open and dense subset.

We proceed similarly for an argument of the statement dealing with the corank one property.

**4.4. Proofs of Propositions 2.8 and 2.19.** We only treat the control-affine case, as the argument for the driftless control-affine case is identical. We argue by contraposition. Consider a nontrivial singular trajectory  $x(\cdot) := x(\cdot, x_0, u)$  of  $\dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$ , with  $(f_0, \dots, f_m) \in VF(M)^{m+1}$ . Assume that  $x(\cdot)$  admits on the one part a normal extremal lift  $(x(\cdot), \lambda^{[n]}(\cdot))$  and on the other part an abnormal extremal lift  $(x(\cdot), \lambda^{[a]}(\cdot))$ .

Let us introduce some notation. For  $k \in \mathbb{N}$ , let  $L = l_1 \cdots l_k$  be a multi-index of  $\{0, \dots, m\}$ . The length of  $L$  is  $|L| = k$  and  $f_L$  is the vector field defined by

$$f_L := [[\cdots [f_{l_1}, f_{l_2}], \cdots], f_{l_k}].$$

A multi-index  $L = jl \cdots l$  with  $k$  consecutive occurrences of the index  $l$  is denoted as  $L = j l^k$ .

For every multi-index  $L$  of  $\{0, \dots, m\}$  and  $t \in [0, T]$ , set

$$h_L^{[n]}(t) = \langle \lambda^{[n]}(t), f_L(x(t)) \rangle \text{ and } h_L^{[a]}(t) = \langle \lambda^{[a]}(t), f_L(x(t)) \rangle.$$

After time differentiation, we have on  $[0, T]$ ,

$$(4.4) \quad \frac{d}{dt} h_L^{[n]}(t) = \sum_{l=1}^m u_l(t) h_{Ll}^{[n]}(t),$$

$$(4.5) \quad \frac{d}{dt} h_L^{[a]}(t) = \sum_{l=1}^m u_l(t) h_{Ll}^{[a]}(t).$$

Recall that, according to the Pontryagin maximum principle, there holds

$$(4.6) \quad u(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_m(t) \end{pmatrix} = U(x(t))^{-1} \begin{pmatrix} h_1^{[n]}(t) \\ \vdots \\ h_m^{[n]}(t) \end{pmatrix},$$

and, for every  $t \in [0, T]$ ,

$$(4.7) \quad h_0^{[a]}(t) = \text{constant}, \quad h_i^{[a]}(t) = 0,$$

for every  $l \in \{1, \dots, m\}$ , and  $t \in [0, T]$ . Since the trajectory  $x(\cdot)$  is nontrivial, there exists an open interval  $J \subset [0, T]$  and  $i \in \{0, \dots, m\}$  such that  $u_i(\cdot) f_i(x(\cdot))$  is never vanishing (with the convention  $u_0 \equiv 1$ ). Fix  $j \in \{0, \dots, m\} \setminus \{i\}$ . Differentiating  $s$  times (with  $s \geq 1$ ) the relation  $h_j^{[a]}(t) = \text{constant}$  with respect to  $t \in J$ , one gets, by using (4.4), (4.5), and (4.6), that

$$(4.8) \quad 0 = \frac{d^s}{dt^s} h_j^{[a]}(t) = (u_i(t))^s h_{ji^s}^{[a]}(t) + R_s(t),$$

where  $R_s(t)$  is polynomial in  $h_L^{[n]}(t)$  and  $h_K^{[a]}(t)$ ,  $|L| \leq s$ ,  $|K| \leq s+1$ , with  $K$  different from  $ji^s$  and  $iji^{s-1}$ . Fix  $t \in J$ . Since  $u_i(t) \neq 0$  and  $f_i(x(t)) \neq 0$ , we are in a position to apply routine transversality arguments. It follows that the  $(m+1)$ -tuple  $(f_0, \dots, f_m)$  belongs to a closed subset of  $VF(M)^{m+1}$  of arbitrary codimension. Proposition 2.8 follows.

**4.5. Proofs of Propositions 2.11, 2.12, and 2.22.** We first prove Proposition 2.12 and argue by contraposition. Consider a nontrivial singular trajectory  $x(\cdot) := x(\cdot, x_0, u)$  of  $\dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$ . Assume that  $x(\cdot)$  admits on the one part a normal extremal lift  $(x(\cdot), \lambda^{[n]}(\cdot))$  and on the other part an abnormal extremal lift  $(x(\cdot), \lambda^{[a]}(\cdot))$ .

From the Pontryagin maximum principle, there holds, for  $l = 1, \dots, m$ ,

$$u_l(t) = \sum_{p=1}^m Q^{lp}(x(t)) \beta_p(x(t)), \quad \beta_p(x(t)) := h_p^{[n]}(t) - \alpha_p(x(t)),$$

where the  $Q^{lp}(x)$  and the  $\alpha_p(x)$  are, resp., the coefficients of  $U^{-1}(x)$  and of  $\alpha(x)$ . Note that the  $u_l$ 's are smooth functions of the time.

Since the trajectory  $x(\cdot)$  is nontrivial, there exists an open interval  $J \subset [0, T]$  such that  $\dot{x}$  is never vanishing on  $J$  and one of the two following cases holds.

*Case 1.*  $u \equiv 0$  on  $J$ .

In that case,  $\dot{x}(t) = f_0(x(t))$  for  $t \in J$ , and  $f_0(x(\cdot))$  is never vanishing on  $J$ . Moreover, for  $p = 1, \dots, m$ ,  $\beta_p \equiv 0$  on  $J$ , i.e.,  $\alpha_p(x(t)) = h_p^{[n]}(t)$  for  $t \in J$ . By differentiating the latter relation with respect to the time, we deduce that, for all  $N \geq 0$ ,  $t \in J$ , and  $p = 1, \dots, m$ ,

$$L_{f_0}^N \alpha_p(x(t)) = L_{f_0}^N h_p^{[n]}(x(t)),$$

where  $L_{f_0}$  denotes the Lie derivative with respect to the vector field  $f_0$ . Applying routine transversality arguments, it follows that  $\alpha$  belongs to a closed subset of  $C^\infty(M, \mathbb{R}^m)$  of arbitrary codimension.

*Case 2.*  $u$  is never vanishing on  $J$ .

Using (2.4) and the LARC, there exist a multi-index  $L$ , an index  $j_0 \in \{0, \dots, m\}$ , and a subinterval of  $J$  (still denoted  $J$ ), such that

$$h_L^{[a]}(t) = \text{constant and } h_{L_{j_0}}^{[a]}(t) \neq 0$$

for every  $t \in J$ . Differentiating  $h_L^{[a]}$  on  $J$ , one gets

$$\begin{aligned} 0 &= \frac{d}{dt} h_L^{[a]}(t) = h_{L_0}^{[a]}(t) + \sum_{l=1}^m u_l(t) h_{L_l}^{[a]}(t) \\ (4.9) \quad &= h_{L_0}^{[a]}(t) + \sum_{1 \leq l \leq p \leq m} c_{lp}(t) Q^{lp}(x(t)), \end{aligned}$$

where  $c_{ll}(t) := \beta_l(t) h_{L_l}^{[a]}(t)$ , and  $c_{lp}(t) := \beta_p(t) h_{L_l}^{[a]}(t) + \beta_l(t) h_{L_p}^{[a]}(t)$  if  $l < p$ .

LEMMA 4.8. *Up to reducing the interval  $J$ , there exist indices  $j$  and  $l$  in  $\{1, \dots, m\}$  such that  $c_{lj}(t)$  or  $c_{jl}(t)$  is never vanishing on  $J$ .*

*Proof.* If  $j_0 = 0$ , then  $h_{L_0}^{[a]}(t) \neq 0$ , and it follows from (4.9) that there exist  $l, j \in \{1, \dots, m\}$  such that  $c_{lj}(t) \neq 0$ . Otherwise, take  $j := j_0$ . In that case, one of the  $\beta_p$ 's does not vanish on  $J$  since  $u$  is not zero. First, assume that  $\beta_j(t)$  is not identically equal to zero on  $J$ ; then, up to reducing  $J$ ,  $c_{jj}(t)$  is never vanishing on  $J$ . Otherwise, there exists  $l \neq j$  such that, up to reducing  $J$ ,  $\beta_l$  is never vanishing on  $J$  and thus similarly for  $c_{lj}$  (or  $c_{jl}$ ).  $\square$

For  $t \in J$ , let  $F_t \in VF(M)$  be the vector field defined by

$$F_t(x) := f_0(x) + \sum_{i=1}^m u_i(t) f_i(x).$$

Notice that  $F_t(x(t)) = \dot{x}(t) \neq 0$ . For all  $N \geq 0$  and  $t \in J$ , we get, by taking the  $(N+1)$ th time derivative of  $h_L^{[a]}$  on  $J$ ,

$$0 = \frac{d^{N+1}}{dt^{N+1}} h_L^{[a]}(t) = c_{lj}(t) L_{F_t}^N Q^{jl}(x(t)) + R_N(t),$$

where  $R_N(t)$  is a linear combination of  $L_{F_t}^s Q^{pi}(x(t))$  with  $s \leq N$ ,  $p \leq i$  in  $\{1, \dots, m\}$ , and  $s < N$  if  $(p, i) = (j, l)$ , and of  $L_{f_r}^s Q^{pi}(x(t))$  with  $s < N$ ,  $p \leq i$  in  $\{1, \dots, m\}$ , and  $r \in \{0, \dots, m\}$ . Applying routine transversality arguments, it follows that  $(U, \alpha)$  belongs to a closed subset of  $\mathcal{S}_m^+(M) \times C^\infty(M, \mathbb{R}^m)$  of arbitrary codimension. Proposition 2.12 is proved.

To show Propositions 2.11 and 2.22, we simply note that the argument of Case 2 with  $\alpha = 0$  applies with suitable modifications.

## REFERENCES

- [1] A. AGRACHEV AND A. SARYCHEV, *Strong minimality of abnormal geodesics for 2-distributions*, J. Dynam. Control Systems, 1 (1995), pp. 139–176.
- [2] A. AGRACHEV AND A. SARYCHEV, *On abnormal extremals for Lagrange variational problems*, J. Math. Systems Estim. Control, 8 (1998), pp. 87–118.
- [3] A. AGRACHEV, *Compactness for sub-Riemannian length minimizers and subanalyticity*, Rend. Semin. Mat. Univ. Politec. Torino, 56 (1998), pp. 1–12.
- [4] A. AGRACHEV AND A. SARYCHEV, *Sub-Riemannian metrics: Minimality of abnormal geodesics versus subanalyticity*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 377–403.
- [5] A. AGRACHEV AND J.-P. GAUTHIER, *On subanalyticity of Carnot-Carathéodory distances*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 18 (2001), pp. 359–382.
- [6] E. ARTIN, *Geometric Algebra*, John Wiley and Sons, New York, 1988.
- [7] A. BELLAÏCHE, *The tangent space in sub-Riemannian geometry*, in Sub-Riemannian Geometry, Progr. Math. 114, Birkhäuser, Basel, 1996, pp. 1–78.
- [8] J. T. BETTS, *Practical Methods for Optimal Control Using Nonlinear Programming*, Adv. Des. Control 3, SIAM, Philadelphia, 2001.
- [9] G. A. BLISS, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, IL, 1946.
- [10] B. BONNARD AND M. CHYBA, *The Role of Singular Trajectories in Control Theory*, Math. Appl. 40, Springer-Verlag, Berlin, New York, 2003.
- [11] B. BONNARD AND I. KUPKA, *Théorie des singularités de l'application entrée/sortie et optimalité des trajectoires singulières dans le problème du temps minimal*, Forum Math. 5 (1993), pp. 111–159.
- [12] B. BONNARD AND I. KUPKA, *Generic properties of singular trajectories*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 167–186.
- [13] Y. CHITOUR, F. JEAN, AND E. TRÉLAT, *Propriétés génériques des trajectoires singulières*, Comptes Rendus Math., 337 (2003), pp. 49–52.
- [14] Y. CHITOUR, F. JEAN, AND E. TRÉLAT, *Genericity results for singular curves*, J. Differential Geom., 73 (2006), pp. 45–73.
- [15] L. C. EVANS, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.
- [16] H. HIRONAKA, *Subanalytic sets*, in Number Theory, Algebraic Geometry and Commutative Algebra, in Honor of Y. Akizuki, Kinokuniya, Tokyo, 1973, pp. 453–493.
- [17] P.-L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.
- [18] W. S. LIU AND H. J. SUSSMANN, *Shortest paths for sub-Riemannian metrics on rank-two distributions*, Mem. Amer. Math. Soc., 118 (1995), no. 564.
- [19] R. MONTGOMERY, *Abnormal minimizers*, SIAM J. Control Optim., 32 (1994), pp. 1605–1620.
- [20] H. J. PESCH, *A practical guide to the solution of real-life optimal control problems. Parametric optimization*, Control Cybernet., 23 (1994), pp. 7–60.
- [21] L. PONTRYAGIN, V. BOLTYANSKII, R. GAMKRELIDZE, AND E. MISCHENKO, *The Mathematical Theory of Optimal Processes*, Wiley Interscience, New York, 1962.
- [22] C. PRIEUR AND E. TRÉLAT, *Quasi-optimal robust stabilization of control systems*, SIAM J. Control Optim., 45 (2006), pp. 1875–1897.
- [23] L. RIFFORD, *On the existence of local smooth repulsive stabilizing feedbacks in dimension three*, J. Differential Equations, 226 (2006), pp. 429–500.
- [24] L. RIFFORD, *The stabilization problem: AGAS and SRS feedbacks*, in Optimal Control, Stabilization, and Nonsmooth Analysis, Lecture Notes in Control and Inform. Sci. 301, Springer-Verlag, Heidelberg, 2004, pp. 173–184.
- [25] L. RIFFORD AND E. TRÉLAT, *Morse-Sard type results in sub-Riemannian geometry*, Math. Ann., 332 (2005), pp. 145–159.
- [26] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
- [27] A. SARYCHEV, *The index of the second variation of a control system*, Math. USSR Sb., 41 (1982), pp. 383–401.
- [28] M. TAMM, *Subanalytic sets in the calculus of variation*, Acta Math., 146 (1981), pp. 167–199.
- [29] E. TRÉLAT, *Some properties of the value function and its level sets for affine control systems with quadratic cost*, J. Dynam. Control Systems, 6 (2000), pp. 511–541.
- [30] E. TRÉLAT, *Asymptotics of accessibility sets along an abnormal trajectory*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 387–414 (electronic).
- [31] E. TRÉLAT, *Global subanalytic solutions of Hamilton-Jacobi type equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 23 (2006), pp. 363–387.
- [32] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, Chelsea, New York, 1980.

## TIME-VARYING REGULAR BILINEAR SYSTEMS\*

HAMID BOUNITH<sup>†</sup> AND ABDELALI IDRISSE<sup>‡</sup>

**Abstract.** The theory of infinite-dimensional systems introduced and developed by Salamon [*Trans. Amer. Math. Soc.*, 300 (1987), pp. 383–431] and Weiss [*SIAM J. Control Optim.*, 27 (1989), pp. 527–545] has been recently applied for various systems; see, e.g., the Jacob–Partington survey [*Current Trends in Operator Theory and its Applications*, Birkhäuser, Boston, 2004]. In the same spirit, Schnaubelt, in his recent work [*SIAM J. Control Optim.*, 41 (2002), pp. 1141–1165], has extended their approach to the time-varying linear systems. In this paper, we follow this spirit and prove that one can extend the results recently obtained for infinite-dimensional time-invariant bilinear systems to the time-varying setting. In particular, we show that the absolute regularity and detectability assumptions ensure the existence of an observer for this kind of systems.

**Key words.** evolution family, well-posed bilinear systems, Lebesgue extension, representation, feedback bilinear systems, detectability, observer

**AMS subject classifications.** 34K35, 34K17, 47D06, 93C23

**DOI.** 10.1137/050632245

**1. Introduction.** In this paper, we deal with the large class of infinite-dimensional time-varying bilinear systems which allow some degree of unboundedness in control and observation. These systems are of the following form:

$$(1.1) \quad \begin{cases} \dot{x}(t) &= A(t)x(t) + u(t)B(t)x(t), \quad t \geq s \geq 0, \\ y(t) &= C(t)x(t), \quad x(s) = x_0, \end{cases}$$

where  $A(t)$ ,  $t \geq 0$ , generate an evolution family  $\mathbb{T}(\cdot, \cdot)$  on  $X$ . The inputs  $u(\cdot)$  are  $L^2_{loc}$ -integrable scalar functions. The Banach spaces  $X$  and  $Y$  are state space and output space, respectively. The linear operators  $B(t)$  are bounded from  $X$  to its extension  $\bar{X}_t$ , which is a Banach space, and  $C(t)$  are densely defined linear operators from  $\underline{X}_t$  to  $Y$ , with  $\underline{X}_t \subset X \subset \bar{X}_t$ . Such operators, called “unbounded,” arise from the modeling of point or boundary control and sensing; see, e.g., [6]. The bilinear systems can be considered as an intermediate subclass between the linear and nonlinear systems. Their relation to several phenomena, especially in engineering areas (see, e.g., [7, 26]) justify their importance. The system (1.1) is a generalized time-varying version of the systems considered recently in [6].

If we consider the particular case  $\underline{X}_t = X = \bar{X}_t$ , then the state  $x$  and the output  $y$  of the system (1.1) can be interpreted by the following functional equations:

$$(1.2) \quad x(t) = \mathbb{T}(t, s)x_0 + \Phi_{t,s}(u, x) \quad \text{for } t \geq s,$$

$$(1.3) \quad y = \Psi_s x_0 + \mathbb{F}_s(u, x) \quad \text{on } [0, t],$$

where  $\Phi_{t,s}(u, x) := \int_s^t \mathbb{T}(t, \tau)u(\tau)B(\tau)x(\tau)d\tau$  (the bilinear input-state operator),  $(\Psi_s x_0)(t) := C(t)\mathbb{T}(t, s)x_0$  (the linear state-output operator), and  $\mathbb{F}_s(u, x)(t) := C(t)$

\*Received by the editors May 24, 2005; accepted for publication (in revised form) October 5, 2007; published electronically March 19, 2008.

<http://www.siam.org/journals/sicon/47-3/63224.html>

<sup>†</sup>Department of Mathematics, Laboratory of Applied Mathematics and Applications (LAMA), Faculty of Sciences, Ibn Zohr University, P.O. Box 8106, 80000 Agadir, Morocco (bounith@yahoo.fr).

<sup>‡</sup>Department of Mathematics, Faculty of Sciences Semlalia, Cadi Ayyad University, P.O. Box 2390, 40000 Marrakech, Morocco (idrissi@ucam.ac.ma).

$\Phi_{t,s}(u, x)$  (the bilinear input-output operator). In our setting, where the unboundedness of control and observation operators is considered, we just conserve, in an abstract formulation, the algebraic relations interconnecting the operators  $\mathbb{T}$ ,  $\Phi$ ,  $\Psi$ , and  $\mathbb{F}$ . These also interpret the different relations between the three components of the time-varying bilinear systems: input, state, and output. This approach was first introduced by Salamon [29] and developed by Weiss [36] for time-invariant linear systems. Later, many studies of time-varying linear systems have proposed different generalizations of this approach; see, e.g., [19, 21, 22, 25, 31].

In the present work we combine the notion of well-posed time-varying linear systems due to Schnaubelt [31] and that of well-posed time-invariant bilinear systems (i.e.,  $A(t) = A, B(t) = B$  and  $C(t) = C$ ), which we have introduced in [6]. The main motivation of our framework is the study of the observer design problem for the well-posed time-varying bilinear system constituted by the quadruple  $\Sigma := (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  (see section 4).

By setting up our abstract framework, we give some topological properties of the bilinear input-state operator  $\Phi$  and the bilinear input-output operator  $\mathbb{F}$  (see Lemmas 3.3 and 3.4). This will be used to give a representation of the abstract system  $\Sigma$  in terms of some control and observation operators. More precisely, we prove that the bilinear input-state operator  $\Phi$  is still “approximately” represented by some bounded operators  $B_n(\cdot)$  (see Proposition 3.9), and under regularity assumptions, which assume a certain smoothness of  $\mathbb{F}_s$  at initial time  $s$  (see Definition 3.10), we also prove that  $\mathbb{F}_s$  can be represented via some unbounded observation operator  $C(\cdot)$  (see Theorem 3.12). This later is extracted by means of the Lebesgue points of the functions  $\Psi_s x_0$  (the output of  $\Sigma$  with  $u = 0$  and initial state  $x_0$ ). This technique was used by Weiss [35] for time-invariant linear systems and later by Schnaubelt [31] for the time-varying situation. We have to mention that in the time-invariant setting we can give, by means of the extrapolation theory, a representation of the bilinear input-state operators  $\Phi$  as convolutions of the semigroups with some unbounded control operators  $B$  (see [20, Cor. 3.11]). The extrapolation theory developed for semigroups (see, e.g., [15, sect. VI.9]) shows that the approximation leads to the existence of the unbounded admissible control operator  $B$ , and if  $X$  is reflexive, the operator  $B$  has to be bounded (see [20, Cor. 4.8]). This theory, however, has no counterpart for evolution families. But the approximation procedure still holds, in our setting, for the convolutions associated with  $\mathbb{T}$  and  $B_n(\cdot)$ . A similar result has been also established by Schnaubelt for time-varying linear systems [31, Prop. 3.5].

In Theorem 3.14 we show that the state trajectory  $x(\cdot)$  of the well-posed time-varying bilinear system  $\Sigma$  exists and is unique by solving the equation (1.2). Further, for any input  $u \in L^2(s, \infty)$  the associated state is given by  $x(t) = \mathbb{T}_u(t, s)x_0$ , where the evolution family  $\mathbb{T}_u$  satisfies the “variation of constants formula”

$$\mathbb{T}_u(t, s)x_0 = \mathbb{T}(t, s)x_0 + \Phi_{t,s}(u, \mathbb{T}_u(\cdot, s)x_0)$$

for  $t \geq s$  and  $x_0 \in X$  (see Theorem 3.14). On the other hand, the regularity gives sense to the output function  $y$  of  $\Sigma$  given by (1.3), since  $x(t)$  fits in  $D(C(t))$  for a.e.  $t \geq s$  (see Theorem 3.15). We show also that the unbounded observation operators  $C(\cdot)$  issued from  $\Sigma$  are still admissible for the evolution family  $\mathbb{T}_u$  whence  $u \in L^2(s, \infty) \cup L^\infty(s, \infty)$ ; i.e., the linear state-output operator  $\Psi_u$  associated with  $C(\cdot)$  and  $\mathbb{T}_u$  can map any initial state  $x_0 \in X$  to an  $L^2_{loc}$ -function which coincides with the output of  $\Sigma$  (see Proposition 3.17).

The results thus obtained in section 3 will serve us to study the observer design problem for the well-posed time-varying bilinear systems  $\Sigma$ . In contrast to the linear systems, the existence and the type of the observer depend on the injected inputs and this even for time-invariant bilinear systems. In the finite-dimensional situation we can cite, e.g., the works by Funahashi [16] and Hara and Furuta [18], who have been interested in the inputs which make the system unobservable called “bad” inputs. Bornard, Couenne, and Celle [2] and Celle et al. [8] were interested in the observer design exited by a class of inputs called “regularly persistent” inputs. In the infinite-dimensional case there are the works by Gauthier, Xu, and Bounabat [17] and Bounit and Hammouri [5], who have adapted the notion of regular persistent input to the infinite-dimensional skew-adjoint bilinear systems. In [4] Bounit and Hammouri have also proposed an exponential observer for bilinear systems when the semigroup is contractive on Hilbert space  $X$ , and the control and observation operators are supposed bounded. Recently, we have given some generalization of the previous results to the regular time-invariant bilinear systems on Banach spaces [6]. We have proved the existence of a Luenberger-like observer for such systems without restrictive conditions on the semigroup [6, Thm. 5.7].

In this paper, we investigate the problem of observer design for the well-posed time-varying bilinear systems  $\Sigma$ . We establish, in Theorem 4.5, the existence of the so-called “feedback” bilinear systems  $\Sigma^K := (\mathbb{T}^K, \Phi^K, \Psi^K, \mathbb{F}^K)$  which describe exactly the estimation error and inherit the absolute regularity. In Proposition 4.7, we prove that the system  $\Sigma^K$  can be represented by the same control and observation operators issued from the observed system  $\Sigma$ . The main result in section 4 (Theorem 4.9) gives the existence of an exponential Luenberger-like observer for the systems  $\Sigma$ . This extends the results of [4, 6] to the present setting. Our condition on the pair  $(\mathbb{T}, \Psi)$ —detectability (see Definition 4.8)—is somewhat stronger than that introduced by Schnaubelt [31, Def. 5.8], but it coincides with that considered by Rebarber [28, Def. 1.5] for time-invariant linear systems.

The article is organized as follows: In section 2 we review the notions of well-posed and regular time-varying linear systems and collect necessary tools for our study. In section 3 we introduce and develop the notions of well-posed and regular time-varying bilinear systems. Section 4 is devoted to designing an observer for absolutely regular time-varying bilinear systems. We give a time-varying version of the results already obtained for this kind of system. Finally, in section 5 we present an example of a system with boundary control and unbounded observation which can be written as an absolutely regular time-varying bilinear systems. We show also the existence of exponential observers for this system.

**2. Well-posed and regular time-varying linear systems: Review.** In this section we recall the necessary background on time-varying linear systems which allow some unboundedness in controls and observations. These systems were initially introduced and studied by Curtain and Pritchard [10], Hinrichsen and Pritchard [19], Jacob [22, 23], and Jacob, Dragan, and Pritchard [25]. Here we recall some notions and collect recent results due to Schnaubelt [31]. The proofs will be omitted and will be referenced.

We begin, first, by fixing some notations which will be used throughout the paper. For a complex Banach space  $E$  and  $p \in [1, +\infty]$  we set  $E^p := L^p([0, \infty), E)$ , and for  $s \geq 0$  we denote by  $E_{loc}^{p,s}$  and  $E_{loc}^p := E_{loc}^{p,0}$  the spaces  $L_{loc}^p([s, \infty); E)$  and  $L_{loc}^p([0, \infty); E)$ , respectively, endowed with their standard Fréchet topology.  $L^\infty([0, \infty),$



$\mathcal{L}_s(E, F)$  is the space of (essentially) bounded strongly measurable operator-valued functions. For  $f \in E_{loc}^{p,s}$  we denote by  $f|_J$  the restriction of  $f$  to the interval  $J$ . We say that  $t \in \mathbb{R}_+$  is a  $p$ -Lebesgue point of  $f \in E_{loc}^p$  if

$$\lim_{\tau \searrow 0} \frac{1}{\tau} \int_t^{t+\tau} \|f(t) - f(\sigma)\|^p d\sigma = 0.$$

It is known (see, e.g., [14, sect. I.1.8], [37, Lem. 6.1]) that almost every  $t \in \mathbb{R}_+$  is a Lebesgue point of  $f$ . We use  $\chi_I$  to denote the characteristic function of the interval  $I$ . For a vector  $z \in E$  we shall use the tensor notation  $\chi_{\mathbb{R}_+} \otimes z$  to denote the function defined by  $(\chi_{\mathbb{R}_+} \otimes z)(t) = z$ . For Banach spaces  $E$ ,  $F$ , and  $G$ ,  $\mathcal{L}(E, F)$  denotes the space of bounded linear operators from  $E$  into  $F$  ( $\mathcal{L}(E) := \mathcal{L}(E, E)$ ), and  $\mathcal{BL}(E \times F, G)$  is the space of bounded bilinear operators from  $E \times F$  to  $G$ .

DEFINITION 2.1. Let  $X$  be a Banach space and  $\Delta := \{(t, s), 0 \leq s \leq t\}$ . An evolution family is a set  $\mathbb{T} := (\mathbb{T}(t, s))_{(t,s) \in \Delta}$  of bounded linear operators on  $X$  such that

- (i)  $\mathbb{T}(s, s) = I$ ;
- (ii)  $\mathbb{T}(t, s) = \mathbb{T}(t, r)\mathbb{T}(r, s)$  for all  $0 \leq s \leq r \leq t$ ;
- (iii)  $(t, s) \mapsto \mathbb{T}(t, s)$  is strongly continuous on  $X$ .

We define the exponential growth bound of evolution family  $\mathbb{T}$  by

$$\omega_0(\mathbb{T}) := \inf\{\omega \in \mathbb{R} : \exists M_\omega \geq 1 \text{ with } \|\mathbb{T}(t, \tau)\| \leq M_\omega e^{\omega(t-\tau)} \text{ for } (t, \tau) \in \Delta\}.$$

In contrast to strongly continuous semigroups, the evolution family may not be exponentially bounded (i.e.  $\omega_0(\mathbb{T}) = +\infty$ ). For this reason and in order to obtain some estimations, we need to deal with the exponentially bounded evolution family. If  $\omega_0(\mathbb{T}) < 0$ , we say that the evolution family  $\mathbb{T}$  is *exponentially stable*. For more details on evolution families and the associated time-varying Cauchy problems, we refer to [9], [15, sect. VI.9], and [30].

We now recall some properties of well-posed and regular time-varying linear systems as introduced by Schnaubelt [31]. Those systems are necessary for studying the problem of observer design for the introduced time-varying bilinear systems; see section 4.

DEFINITION 2.2. Let  $U, X$ , and  $Y$  be Banach spaces. A well-posed time-varying linear system on  $(X, U, Y)$  for state space  $X$ , control space  $U$ , and output space  $Y$  is the quadruple  $\Gamma := (\mathbb{T}, \mathbb{K}, \Psi, \mathbb{L})$ , where

- (i)  $\mathbb{T} := (\mathbb{T}(t, s))_{(t,s) \in \Delta}$  is an evolution family on  $X$ ;
- (ii)  $\mathbb{K} := (\mathbb{K}_{t,s})_{(t,s) \in \Delta}$  is a family of operators in  $\mathcal{L}(U_{loc}^{2,s}, X)$  such that

$$(2.1) \quad \mathbb{K}_{t,s}u = \mathbb{K}_{t,r}(u|_{[r,\infty)}) + \mathbb{T}(t, r)\mathbb{K}_{r,s}(u), \quad s \leq r \leq t, \text{ and}$$

$$(2.2) \quad \|\mathbb{K}_{t,s}(u)\|_X \leq \beta \|u\|_{L^2([s,t], U)}, \quad s \leq t \leq s + t_0,$$

for  $u \in U_{loc}^{2,s}$  and some/all  $t_0$  and  $\beta = \beta(t_0) > 0$ . The pair  $(\mathbb{T}, \mathbb{K})$  is called a time-varying linear control system on  $(X, U)$ ;

- (iii)  $\Psi := (\Psi_s)_{s \geq 0}$  is a family of operators in  $\mathcal{L}(X, Y_{loc}^{2,s})$  such that

$$(2.3) \quad \Psi_s x = \Psi_t \mathbb{T}(t, s)x, \quad \text{on } [t, \infty), \text{ and}$$

$$(2.4) \quad \int_s^{s+t_0} \|(\Psi_s x)(t)\|^2 dt \leq \delta^2 \|x\|^2,$$

for  $(t, s) \in \Delta$ ,  $x \in X$ , and some/all  $t_0$  and constant  $\delta = \delta(t_0) > 0$ . The pair  $(\mathbb{T}, \Psi)$  is called a time-varying linear observation system on  $(X, Y)$ ;

(iv)  $\mathbb{L} := (\mathbb{L}_s)_{s \geq 0}$  is a family of operators in  $\mathcal{L}(U_{loc}^{2,s}, Y_{loc}^{2,s})$  such that

$$(2.5) \quad \mathbb{L}_s u = \Psi_t \mathbb{K}_{t,s} u + \mathbb{L}_t(u|_{[t,\infty)}) \quad \text{on } [t, \infty), \text{ and}$$

$$(2.6) \quad \|\mathbb{L}_s u\|_{L^2([s, s+t_0], Y)} \leq \kappa \|u\|_{L^2([s, s+t_0], U)}$$

for  $u \in U_{loc}^{2,s}$ ,  $(t, s) \in \Delta$ , and some/all  $t_0$  and constant  $\kappa = \kappa(t_0) > 0$ . The operators  $\mathbb{L}_s$ ,  $s \geq 0$ , are called input-output operators.

The above definition can be extended to the case of  $p \in [1, \infty)$  instead of  $p = 2$ . As for time-invariant linear observation systems, it was shown [31] that time-varying observation systems are typically given by some unbounded operators which we define as follows.

DEFINITION 2.3. Let  $\mathbb{T}$  be an evolution family on  $X$  and  $C(s) : X \supset D(C(s)) \rightarrow Y$ ,  $s \geq 0$ , be densely defined linear operators. We say that  $C(\cdot)$  are  $\mathbb{T}$ -admissible observation operators if

$$(i) \quad \mathbb{T}(\cdot, s)x \in \mathcal{D}_s(C(\cdot)),$$

$$(ii) \quad \int_s^{s+t_0} \|C(t)\mathbb{T}(t, s)x\|^2 dt \leq \delta^2 \|x\|^2$$

for  $s \geq 0$ ,  $x \in D(C(s))$ , and some constants  $\delta, t_0 > 0$ .

Here  $\mathcal{D}_s(C(\cdot)) := \{f \in X_{loc}^{2,s} : f(t) \in D(C(t)) \text{ for a.e. } t \geq s, C(\cdot)f(\cdot) \in Y_{loc}^{2,s}\}$ .

By the above admissibility, the mapping

$$(2.7) \quad \begin{aligned} \Psi_s : D(C(s)) &\longrightarrow Y_{loc}^{2,s}, \quad s \geq 0, \\ x &\longmapsto \Psi_s x := C(\cdot)\mathbb{T}(\cdot, s)x \end{aligned}$$

has a unique extension (again noted by  $\Psi_s$ ) to linear bounded operators from  $X$  to  $Y_{loc}^{2,s}$ , and the pair  $(\mathbb{T}, \Psi)$  is a time-varying linear observation system represented by  $C(\cdot)$ . Conversely, it was shown [31] that a time-varying observation system  $(\mathbb{T}, \Psi)$  can be represented by admissible observation operators. This extends the time-invariant case given in [35]. In fact, we define the operators  $C(\cdot)$  as follows:

$$(2.8) \quad D(C(s)) = \underline{X}_s := \left\{ x \in X : \exists C(s)x := Y - \lim_{\tau \searrow 0} \frac{1}{\tau} \int_s^{s+\tau} (\Psi_s x)(\sigma) d\sigma \right\},$$

equipped with the norm

$$\|x\|_{\underline{X}_s} := \|x\| + \sup_{0 < \tau \leq 1} \left\| \frac{1}{\tau} \int_s^{s+\tau} (\Psi_s x)(\sigma) d\sigma \right\|_Y$$

for  $x \in \underline{X}_s$ . Then  $(\underline{X}_s, \|\cdot\|_s)$ ,  $s \geq 0$ , are Banach spaces,  $\underline{X}_s \subset X$ , and  $C(s)$  is linear continuous. It was shown (see [31, Thm. 2.7]) that the operators  $C(\cdot)$  are admissible and  $(\Psi_s x)(t) = C(t)\mathbb{T}(t, s)x$  for all Lebesgue points  $t$  of  $\Psi_s x$ .

A more important subclass of well-posed linear systems is those which are regular. This concept was first introduced by Weiss [37] and generalized latter by Schnaubelt [31] to the time-varying systems.

DEFINITION 2.4. Let  $\Gamma = (\mathbb{T}, K, \Psi, \mathbb{L})$  be a time-varying linear system. Then  $\Gamma$  is called a regular linear system (with feedthrough  $D = 0$ ) if for  $v \in U, t \geq 0$  the following limit exists in  $Y$ :

$$(2.9) \quad \lim_{\tau \searrow 0} \frac{1}{\tau} \int_t^{t+\tau} \mathbb{L}_t(\chi_{\mathbb{R}_+} \otimes v)(\sigma) d\sigma = 0,$$

and absolutely regular if

$$\lim_{\tau \searrow 0} \frac{1}{\tau} \int_t^{t+\tau} \|\mathbb{L}_t(\chi_{\mathbb{R}_+} \otimes v)(\sigma)\|^2 d\sigma = 0.$$

One can remark that  $\Gamma = (\mathbb{T}, K, \Psi, \mathbb{L})$  is regular if the step response  $\mathbb{L}_t(\chi_{\mathbb{R}_+} \otimes v)$  has a Lebesgue point at  $t$  equal to zero for all  $t \geq 0$ .

**3. Well-posed and regular time-varying bilinear systems.** In this section, we follow what we have presented in the previous section. We give the time-varying analogue of the well-posed bilinear systems introduced in [6].

**DEFINITION 3.1.** *Let  $X$  and  $Y$  be Banach spaces. A well-posed time-varying bilinear system on  $(X, \mathbb{C}, Y)$ , for state space  $X$ , control space  $\mathbb{C}$ , and output space  $Y$ , is the quadruple  $\Sigma := (\mathbb{T}, \Phi, \Psi, \mathbb{F})$ , where*

- (i)  $\mathbb{T}$  and  $\Psi$  satisfy (i) and (iii) of Definition 2.2, respectively;
- (ii)  $\Phi := (\Phi_{t,s})_{(t,s) \in \Delta}$  is a family of operators in  $\mathcal{BL}(\mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s}, X)$  satisfying

$$(3.1) \quad \Phi_{t,s}(u, x) = \Phi_{t,r}(u|_{[r,\infty)}, x|_{[r,\infty)}) + \mathbb{T}(t, r)\Phi_{r,s}(u, x), \quad s \leq r \leq t, \text{ and}$$

$$(3.2) \quad \|\Phi_{t,s}(u, x)\|_X \leq \beta \|u\|_{L^2[s,t]} \|x\|_{L^2([s,t], X)}, \quad s \leq t \leq s + t_0,$$

for  $u \in \mathbb{C}_{loc}^{2,s}$ ,  $x \in X_{loc}^{2,s}$ , and constants  $t_0, \beta = \beta(t_0) > 0$ . Then the pair  $(\mathbb{T}, \Phi)$  is called a time-varying bilinear control system on  $X$ ;

- (iii)  $\mathbb{F} := (\mathbb{F}_s)_{s \geq 0}$  is a family of operators in  $\mathcal{BL}(\mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s}, Y_{loc}^{2,s})$  such that

$$(3.3) \quad \mathbb{F}_s(u, x) = \Psi_t \Phi_{t,s}(u, x) + \mathbb{F}_t(u|_{[t,\infty)}, x|_{[t,\infty)}) \quad \text{on } [t, \infty) \text{ and}$$

$$(3.4) \quad \|\mathbb{F}_s(u, x)\|_{L^2([s, s+t_0], Y)} \leq \kappa \|u\|_{L^2[s, s+t_0]} \|x\|_{L^2([s, s+t_0], X)}$$

for  $u \in \mathbb{C}_{loc}^{2,s}$ ,  $x \in X_{loc}^{2,s}$ ,  $(t, s) \in \Delta$ , and constants  $t_0, \kappa = \kappa(t_0) > 0$ .

**Remark 3.2.**

- (i) The property (3.3) shows that the operator  $\mathbb{F}_s$  is causal, i.e.,  $\mathbb{F}_s(u, x)|_{[t,\infty)} = \mathbb{F}_t(u|_{[t,\infty)}, x|_{[t,\infty)})$ . Thus, one can define the restriction

$$\mathbb{F}_{t,s} := \mathbb{F}_s : L^2[s, t] \times L^2([s, t], X) \rightarrow L^2([s, t], Y).$$

- (ii) If  $(\mathbb{T}, \Phi)$  is a well-posed time-varying bilinear control system on  $X$ , then  $(\mathbb{T}, \Phi(\chi_{\mathbb{R}_+}, \cdot))$  is a well-posed time-varying linear control system on  $(X, U := X)$ . Conversely, let  $X = U$ , and assume that the pair  $(\mathbb{T}, \varphi)$  is a linear control system for  $p = 1$ ; then  $(\mathbb{T}, \Phi)$  is a bilinear control system on  $(X, U)$  for all  $p \in (1, \infty)$  where  $\Phi_{t,s}(u, x) := \varphi_{t,s}(u \cdot x)$ . Moreover, if  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  is a well-posed time-varying bilinear system on  $(X, \mathbb{C}, Y)$ , then the associated quadruple  $\Gamma_\Sigma = (\mathbb{T}, \Phi(\chi_{\mathbb{R}_+}, \cdot), \Psi, \mathbb{F}(\chi_{\mathbb{R}_+}, \cdot))$  is a time-varying linear system on  $(X, X, Y)$ .

We start with some lemmas which will be used in the subsequent developments.

**LEMMA 3.3.** *Let  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  be a well-posed time-varying bilinear system on  $(X, \mathbb{C}, Y)$  with  $\omega_0(\mathbb{T}) < \infty$ . Then*

$$(3.5) \quad \|\Phi_{t,s}(u, x)\|_X \leq d(\omega, t - s) \|u\|_{L^2[s,t]} \|x\|_{L^2([s,t], X)},$$

$$(3.6) \quad \|\Phi_{\cdot,s}(u, x)\|_{L^2([s,t], X)} \leq d(\omega, t - s) \|u\|_{L^2[s,t]} \|x\|_{L^2([s,t], X)},$$

$$(3.7) \quad \|\mathbb{F}_s(u, x)\|_{L^2([s, s+t_1], Y)} \leq d(\omega, t_1) \|u\|_{L^2[s, s+t_1]} \|x\|_{L^2([s, s+t_1], X)}$$

for  $t_1 \geq 0$ ,  $t \geq s \geq 0$ ,  $u \in \mathbb{C}_{loc}^{2,s}$ ,  $x \in X_{loc}^{2,s}$ , where  $\omega > \omega_0(\mathbb{T})$ ,

$$d(\omega, t) := \begin{cases} d_0(w)e^{\omega t} & \text{if } \omega > 0, \\ d_0(w) \left(1 + \frac{t}{t_0}\right)^{\frac{1}{2}} & \text{if } \omega = 0, \\ d_0(\omega)e^{\omega t_0} & \text{if } \omega < 0, \end{cases}$$

and  $d_0, t_0$  are positive constants.

*Proof.* Let  $t_0$  be the constant given in Definition 3.1(ii). For  $s \leq t \leq s + t_0$ , the assertion (3.5) is clear. Now let  $t > s + t_0$  and  $n := [\frac{t-s}{t_0}]$  (the integer part of  $\frac{t-s}{t_0}$ ), and set  $s_k := s + kt_0$  for  $k \in \mathbb{N} := \{0, 1, \dots\}$ . Then  $t \in I_n := [s_n, s_{n+1}]$ . Let  $u_k$  and  $x_k$  be the restriction of  $u$  and  $x$  to  $I_k \cap [s, t]$  for  $k = 0, \dots, n$ , respectively. By using (3.1) and proceeding by induction we obtain

$$\Phi_{t,s}(u, x) = \Phi_{t,s_n}(u_n, x_n) + \sum_{k=1}^n \mathbb{T}(t, s_k) \Phi_{s_k, s_{k-1}}(u_{k-1}, x_{k-1}).$$

That means

$$\begin{aligned} \|\Phi_{t,s}(u, x)\| &\leq \beta(t_0) \|u_n\|_{L^2[s,t]} \|x_n\|_{L^2([s,t], X)} \\ &\quad + M\beta(t_0) \sum_{k=1}^n e^{\omega(t-s_k)} \|u_{k-1}\|_{L^2[s,t]} \|x_{k-1}\|_{L^2([s,t], X)} \\ &\leq M\beta(t_0) e^{\omega^- t_0} \left( \sum_{k=0}^{\infty} e^{\omega t_0(n-k)} \|u_k\|_{L^2[s,t]} \|x_k\|_{L^2([s,t], X)} \right) \\ &= M\beta(t_0) e^{\omega^- t_0} \left( \sum_{k=0}^{\infty} e_{(n-k)} \|u_k\|_{L^2[s,t]} \|x_k\|_{L^2([s,t], X)} \right), \end{aligned}$$

where  $\omega^- := \min\{\omega, 0\}$  and  $e_k := e^{\omega t_0 k}$  if  $k = 0, \dots, n$  and  $e_k := 0$  elsewhere. Thus, Young's and Hölder's inequalities yield estimation (3.5). Estimation (3.6) is straightforward.

The assertion (3.7) is clear for  $t_1 \leq t_0$ . Assume, now that  $t_1 > t_0$ , as in (i) that there is  $n \in \mathbb{N}^* := \mathbb{N} - \{0\}$  such that  $t_1 \in I_n := [s_n, s_{n+1}]$ . By using (3.1) and (3.3) we obtain

$$\begin{aligned} \mathbb{F}_s(u, x) &= \Psi_{s_k} \Phi_{s_k, s}(u, x) + \mathbb{F}_{s_k}(u_k, x_k) \quad \text{on } I_k \\ &= \mathbb{F}_{s_k}(u_k, x_k) + \sum_{j=1}^k \Psi_{s_k} \mathbb{T}(s_k, s_j) \Phi_{s_j, s_{j-1}}(u_{j-1}, x_{j-1}). \end{aligned}$$

Hence

$$\begin{aligned} \|\mathbb{F}_s(u, x)\|_{L^2(I_k, Y)} &\leq \kappa(t_0) \|u_k\|_{L^2(I_k)} \|x_k\|_{L^2(I_k, X)} \\ &\quad + M\kappa(t_0) \delta(t_0) \sum_{j=1}^k e^{\omega t_0(k-j)} \|u_{j-1}\|_{L^2(I_{j-1})} \|x_{j-1}\|_{L^2(I_{j-1}, X)} \\ &\leq c(t_0) e^{\omega^- t_0} \sum_{j=1}^k e^{\omega t_0(k-j)} \|u_{j-1}\|_{L^2(I_{j-1})} \|x_{j-1}\|_{L^2(I_{j-1}, X)}, \end{aligned}$$

where  $c(t_0) := \max\{\kappa(t_0)M\kappa(t_0)\delta(t_0)\}$ . Therefore it follows that

$$\begin{aligned} \|\mathbb{F}_s(u, x)\|_{L^2([s, s+t_1], Y)} &\leq \left( \sum_{k=0}^n \|\mathbb{F}_s(u, x)\|_{L^2(I_k, Y)}^2 \right)^{1/2} \\ &\leq c(t_0)e^{\omega^- t_0} \left( \sum_{k=0}^n \left( \sum_{j=0}^k e^{\omega(k-j)} \|u_j\|_{L^2(I_j)} \|x_j\|_{L^2(I_j, X)} \right)^2 \right)^{\frac{1}{2}} \\ &\leq c(t_0)e^{\omega^- t_0} \left( \sum_{k=0}^{\infty} \left( \sum_{j=0}^k e^{\omega(k-j)} \|u_j\|_{L^2(I_j)} \|x_j\|_{L^2(I_j, X)} \right)^2 \right)^{\frac{1}{2}}. \end{aligned}$$

By applying once again Young's and Hölder's inequalities, we then obtain

$$\|\mathbb{F}_s(u, x)\|_{L^2([s, s+t_1], Y)} \leq c(t_0)e^{\omega^- t_0} \left( \sum_{k=0}^n e^{2\omega t_0 k} \right)^{\frac{1}{2}} \|u\|_{L^2[s, s+t_1]} \|x\|_{L^2([s, s+t_1], X)}.$$

Thus the asserted estimate follows.  $\square$

LEMMA 3.4. *Let  $(\mathbb{T}, \Phi)$  be a time-varying bilinear control system on  $X$  with  $\omega_0(\mathbb{T}) < \infty$ . If  $u \in \mathbb{C}_{loc}^{2,s}$  and  $x \in X_{loc}^{2,s}$ , then the following properties hold:*

- (i)  $t \mapsto \Phi_{t,s}(u, x) \in X$  is continuous from the right for  $t \geq s$ ;
- (ii)  $s \mapsto \Phi_{t,s}(u, x) \in X$  is continuous for  $s \in [0, t]$  (locally uniformly in  $t$ );
- (iii)  $(t, s) \mapsto \Phi_{t,s}(u, x)$  is measurable in  $X$ .

*Proof.* In view of (3.1), we get

$$\begin{aligned} \Phi_{t',s}(u, x) - \Phi_{t,s}(u, x) &= \Phi_{t',t}(u, x) + (\mathbb{T}(t', t) - I)\Phi_{t,s}(u, x) \text{ and} \\ \Phi_{t,s}(u, x) - \Phi_{t,s'}(u, x) &= \mathbb{T}(t, s')\Phi_{s',s}(u, x) \end{aligned}$$

for all  $t' \geq t \geq s' \geq s \geq 0$ . Thus, we can estimate

$$\begin{aligned} \|\Phi_{t',s}(u, x) - \Phi_{t,s}(u, x)\| &\leq \kappa \|u\|_{L^2[t, t']} \|x\|_{L^2([t, t'], X)} + \|(\mathbb{T}(t', t) - I)\Phi_{t,s}(u, x)\| \text{ and} \\ \|\Phi_{t,s}(u, x) - \Phi_{t,s'}(u, x)\| &\leq M\kappa e^{|\omega|(t-s)} \|u\|_{L^2[s, s']} \|x\|_{L^2([s, s'], X)}. \end{aligned}$$

Thus the lemma is proved.  $\square$

DEFINITION 3.5. *Let  $(\mathbb{T}, \Phi)$  be a time-varying bilinear control system on  $X$ ,  $(\mathbb{T}, \Psi)$  be a time-varying linear observation system on  $(X, Y)$ , and  $C(\cdot)$  be the observation operator associated with  $(\mathbb{T}, \Psi)$  given by (2.8). We say that the bilinear triple  $(\mathbb{T}, \Phi, \Psi)$  is admissible on  $(X, \mathbb{C}, Y)$  if for all  $s \geq 0$ ,  $u \in \mathbb{C}_{loc}^{2,s}$ , and  $x \in X_{loc}^{2,s}$*

- (i)  $\Phi_{\cdot,s}(u, x) \in \mathcal{D}_s(C(\cdot))$  and

$$(ii) \|C(\cdot)\Phi_{\cdot,s}(u, x)\|_{L^2([s, s+t_0], Y)} \leq \kappa \|u\|_{L^2[s, s+t_0]} \|x\|_{L^2([s, s+t_0], X)}$$

for some constants  $\kappa, t_0 > 0$ .

PROPOSITION 3.6. *Let  $(\mathbb{T}, \Phi, \Psi)$  be an admissible triple on  $(X, \mathbb{C}, Y)$ . Define  $\mathbb{F}_s(u, x) := C(\cdot)\Phi_{\cdot,s}(u, x)$ . Then  $(\mathbb{T}, \Phi, \Psi, \mathbb{F})$  is a well-posed time-varying bilinear system on  $(X, \mathbb{C}, Y)$ .*

*Proof.* We have only to verify (3.3) for all  $u \in \mathbb{C}_{loc}^{2,s}$ ,  $x \in X_{loc}^{2,s}$ , and  $(t, s) \in \Delta$ . By definition, we have

$$\begin{aligned}\mathbb{F}_s(u, x)(\tau) &= C(\tau)\Phi_{\tau,s}(u, x) \\ &= C(\tau)(\Phi_{\tau,t}(u|_{[t,\infty)}, x|_{[t,\infty)}) + \mathbb{T}(\tau, t)\Phi_{t,s}(u, x)) \\ &= C(\tau)\Phi_{\tau,t}(u|_{[t,\infty)}, x|_{[t,\infty)}) + C(\tau)\mathbb{T}(\tau, t)\Phi_{t,s}(u, x) \\ &= \mathbb{F}_t(u|_{[t,\infty)}, x|_{[t,\infty)})(\tau) + (\Psi_t\Phi_{t,s}(u, x))(\tau)\end{aligned}$$

for a.e.  $\tau \geq t$ .  $\square$

Let  $\overline{X}_t$ ,  $t \geq 0$ , be Banach spaces in which  $X$  is densely and continuously embedded. We assume that  $\mathbb{T}(t, s)$  has a locally uniformly bounded extension  $\overline{\mathbb{T}}(t, s) : \overline{X}_s \rightarrow \overline{X}_t$  verifying (i) and (ii) of Definition 2.1. We assume that  $\overline{\mathbb{T}}(t, \cdot)$  is strongly continuous. The following definition of admissible control operators is a simple extension of that given in [20] for the time-invariant case. For  $(t, s) \in \Delta$  and  $f \in L_{loc}^1([s, \infty), X)$  we define

$$(\mathbb{V}_s^{\overline{\mathbb{T}}}f)(t) := \int_s^t \overline{\mathbb{T}}(t, \tau)f(\tau) d\tau.$$

DEFINITION 3.7. We say that  $B(t) \in \mathcal{L}(X, \overline{X}_t)$ ,  $t \geq 0$ , is  $\mathbb{T}$ -admissible control operators if the following hold:

- (i) The function  $\overline{\mathbb{T}}(t, \cdot)u(\cdot)B(\cdot)x(\cdot)$  is integrable in  $\overline{X}_t$  and

$$(3.8) \quad [\mathbb{V}_s^{\overline{\mathbb{T}}}(uB(\cdot))x](t) := \int_s^t \overline{\mathbb{T}}(t, \tau)u(\tau)B(\tau)x(\tau) d\tau \in X.$$

- (ii) There are constants  $t_0, \beta > 0$  such that

$$(3.9) \quad \|[\mathbb{V}_s^{\overline{\mathbb{T}}}(uB(\cdot))x](t)\|_X \leq \beta \|u\|_{L^2(s,t)} \|x\|_{L^2([s,t], X)}$$

for all  $0 \leq s \leq t \leq s + t_0$ ,  $u \in L^2[s, t]$ , and  $x \in L^2([s, t], X)$ .

Remark 3.8.

- (i) By means of the admissible control operators  $B(\cdot)$  we can define a time-varying bilinear control system  $(\mathbb{T}, \Phi)$  on  $X$  by setting

$$(3.10) \quad \Phi_{t,s}(u, x) := [\mathbb{V}_s^{\overline{\mathbb{T}}}(uB(\cdot))x](t), \quad (t, s) \in \Delta.$$

Let us mention that every time-invariant bilinear control system on  $X$  is given by a  $\mathbb{T}$ -admissible control operator due to [20, Thm. 3.9], where  $\overline{X}_t$  coincides with the extrapolation space  $X_{-1}$  associated with  $\mathbb{T}$ . However, one can extend this result only to the time-dependent setting in the “approximate” sense because we have not an extrapolation theory for evolution families.

- (ii) If  $B(s)$  and  $C(s)$ ,  $s \geq 0$ , are  $\mathbb{T}$ -admissible control and observation operators, then the triple  $(\mathbb{T}, B(\cdot), C(\cdot))$  is called admissible if  $(\mathbb{T}, \Phi, \Psi)$  is admissible, where  $\Phi$  and  $\Psi$  are given by (3.10) and (2.7), respectively. Moreover, an admissible triple  $(\mathbb{T}, B(\cdot), C(\cdot))$  on  $(X, \mathbb{C}, Y)$  gives rise to a well-posed time-varying bilinear system  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$ , on  $(X, \mathbb{C}, Y)$ , where  $\mathbb{F}_s = C(\cdot)\mathbb{V}_s^{\overline{\mathbb{T}}}(uB(\cdot))x$  (via Proposition 3.6).

Let  $(\mathbb{T}, \Phi)$  be a time-varying bilinear control system on  $X$ . Let  $(u, x) \in \mathbb{C}_{loc}^2 \times X_{loc}^2$ ,  $z \in X$ , and  $n \in \mathbb{N}^*$ . We define

$$(3.11) \quad B_n(u, x)(t) := \begin{cases} n\Phi_{t, t-\frac{1}{n}}(u, x) & \text{if } t - \frac{1}{n} \geq 0, \\ n\Phi_{t, 0}(u, x) & \text{if } t - \frac{1}{n} < 0, \end{cases} \quad \text{and}$$

$$(3.12) \quad B_n(t)z := B_n(\chi_{\mathbb{R}_+}, \chi_{\mathbb{R}_+} \otimes z)(t) \quad \text{for } t \geq 0.$$

Due to Lemma 3.4,  $B_n(u, x) \in X_{loc}^1$  and  $B_n(t) \in \mathcal{L}(X)$ . The following result shows that the input operator  $\Phi_{t,s}$  can be approximated by the convolutions:

$$(3.13) \quad \Phi_{t,s}^n(u, x) := [\mathbb{V}_s^{\mathbb{T}} B_n(u, x)](t), \quad (t, s) \in \Delta.$$

**PROPOSITION 3.9.** *Let  $(\mathbb{T}, \Phi)$  be a time-varying bilinear control system on  $X$ ,  $0 \leq t - s \leq t_0$ ,  $z \in X$ ,  $u \in \mathbb{C}_{loc}^2$ , and  $x \in X_{loc}^2$ . Then the following hold:*

(i)  $\Phi_{t,s}^n(u, x) \rightarrow \Phi_{t,s}(u, x)$  in  $X$  as  $n \rightarrow \infty$ , and

$$\|\Phi_{t,s}^n(u, x)\|_X \leq \beta \|u\|_{L^2[s,t]} \|x\|_{L^2([s,t], X)} \quad (\text{with } \beta > 0).$$

(ii)  $(t, s) \mapsto \Phi_{t,s}(u, x)$  and  $B_n(\cdot)z$  are continuous in  $X$ .

(iii)  $[\mathbb{V}_s^{\mathbb{T}}(uB_n(\cdot)x)](t) \rightarrow \Phi_{t,s}(u, x)$  as  $n \rightarrow \infty$  if  $u \in W_{loc}^{1,2}(\mathbb{R})$  and  $x \in W_{loc}^{1,2}(\mathbb{R}, X)$ .

(iv)  $\|[\mathbb{V}_s^{\mathbb{T}}(uB_n(\cdot)x)](t)\|_X \leq \beta \|u \cdot x\|_{L^2([s,t], X)}$  if  $u \cdot x \in L^2([s,t], X)$ .

The limits in (i) and (iv) are taken in  $X$  and locally uniform in  $(t, s) \in \Delta$ .

*Proof.* The assertions (i)–(ii) can be obtained by proceeding as in the proof of [31, Prop. 3.5]. We now concentrate on the claims in (iii) and (iv). To prove (iii) we proceed by steps.

(iii) *Step 1.* We first show that  $\Xi_k(u, x) := B_k(u, x) - uB_k(\cdot)x \rightarrow 0$  as  $k \rightarrow \infty$  in  $X_{loc}^1$ . Take  $u \in W_{loc}^{1,2}(\mathbb{R})$  and  $x \in W_{loc}^{1,2}(\mathbb{R}, X)$ . Then, via (3.2), we have

$$\begin{aligned} & \|\Xi_k(u, x)\|_{L^1([s, s+t_0], X)} \\ & \leq k \int_s^{s+t_0} \|\Phi_{\tau, \tau-\frac{1}{k}}(u, x) - u(\tau)\Phi_{\tau, \tau-\frac{1}{k}}(\chi_{\mathbb{R}_+}, \chi_{\mathbb{R}_+} \otimes x(\tau))\| d\tau \\ & \leq k \int_s^{s+t_0} (\|\Phi_{\tau, \tau-\frac{1}{k}}(u - \chi_{\mathbb{R}_+} \otimes u(\tau), x)\| + \|\Phi_{\tau, \tau-\frac{1}{k}}(\chi_{\mathbb{R}_+} \otimes u(\tau), x - \chi_{\mathbb{R}_+} \otimes x(\tau))\|) d\tau \\ & \leq \beta k \int_s^{s+t_0} \left( \int_{\tau-\frac{1}{k}}^{\tau} |u(\sigma) - u(\tau)|^2 d\sigma \right)^{1/2} \left( \int_{\tau-\frac{1}{k}}^{\tau} \|x(\sigma)\|^2 d\sigma \right)^{1/2} d\tau \\ & \quad + \beta k^{1/2} \int_s^{s+t_0} |u(\tau)| \left( \int_{\tau-\frac{1}{k}}^{\tau} \|x(\sigma) - x(\tau)\|^2 d\sigma \right)^{1/2} d\tau. \end{aligned}$$

This can be also decomposed as

$$\begin{aligned} & \|\Xi_k(u, x)\|_{L^1([s, s+t_0], X)} \\ & \leq \beta k \int_s^{s+t_0} \left( \int_{\tau-\frac{1}{k}}^{\tau} |u(\sigma) - u(\tau)|^2 d\sigma \right)^{1/2} \left( \int_{\tau-\frac{1}{k}}^{\tau} \|x(\sigma) - x(\tau)\|^2 d\sigma \right)^{1/2} d\tau \\ & \quad + \beta k^{1/2} \int_s^{s+t_0} \|x(\tau)\| \left( \int_{\tau-\frac{1}{k}}^{\tau} |u(\sigma) - u(\tau)|^2 d\sigma \right)^{1/2} d\tau \\ & \quad + \beta k^{1/2} \int_s^{s+t_0} |u(\tau)| \left( \int_{\tau-\frac{1}{k}}^{\tau} \|x(\sigma) - x(\tau)\|^2 d\sigma \right)^{1/2} d\tau. \end{aligned}$$

Therefore, one can write

$$\begin{aligned} \|\Xi_k(u, x)\|_{L^1([s, s+t_0], X)} &\leq \beta k \int_s^{s+t_0} |u_k(\tau)| \|x_k(\tau)\| d\tau + c_u \beta k^{1/2} \int_s^{s+t_0} \|x_k(\tau)\| d\tau \\ &\quad + c_x \beta k^{1/2} \int_s^{s+t_0} |u_k(\tau)| d\tau, \end{aligned}$$

where  $f_k(\tau) := (\int_{\tau-\frac{1}{k}}^{\tau} \|f(\sigma) - f(\tau)\|^2 d\sigma)^{1/2}$  and  $c_f := \sup_{s \leq \sigma \leq s+t_0} \|f(\sigma)\|_E$  for  $f \in W_{loc}^{1,2}(\mathbb{R}, E)$ .

*Step 2.* We have to show that

$$(3.14) \quad \lim_{k \rightarrow \infty} k \int_s^{s+t_0} \|f_k(\tau)\|^2 d\tau = 0$$

for all  $f \in W_{loc}^{1,2}(\mathbb{R}, E)$ . In fact, we have

$$\begin{aligned} \int_s^{s+t_0} \|f_k(\tau)\|^2 d\tau &= \int_s^{s+t_0} \int_{\tau-\frac{1}{k}}^{\tau} \|f(\sigma) - f(\tau)\|^2 d\sigma d\tau \\ &= \int_s^{s+t_0} \int_{\tau-\frac{1}{k}}^{\tau} \left\| \int_{\sigma}^{\tau} f'(\mu) d\mu \right\|^2 d\sigma d\tau \\ &\leq \int_s^{s+t_0} \int_{\tau-\frac{1}{k}}^{\tau} \left( \int_{\sigma}^{\tau} \|f'(\mu)\| d\mu \right)^2 d\sigma d\tau \\ &\leq \frac{1}{k} \int_s^{s+t_0} \left( \int_{\tau-\frac{1}{k}}^{\tau} \|f'(\mu)\| d\mu \right)^2 d\tau. \end{aligned}$$

By using Hölder's inequality and interchanging integrals we obtain

$$\begin{aligned} \int_s^{s+t_0} \|f_k(\tau)\|^2 d\tau &\leq \frac{1}{k^2} \int_s^{s+t_0} \int_0^{\frac{1}{k}} \|f'(\tau - \mu)\|^2 d\mu d\tau \\ &\leq \frac{1}{k^2} \int_0^{\frac{1}{k}} \int_s^{s+t_0} \|f'(\tau - \mu)\|^2 d\tau d\mu. \end{aligned}$$

It follows that

$$k \int_s^{s+t_0} \|f_k(\tau)\|^2 d\tau \leq \frac{1}{k} \int_0^{\frac{1}{k}} \int_s^{s+t_0} \|f'(\tau - \mu)\|^2 d\tau d\mu.$$

Therefore by the fact that

$$k \int_s^{s+t_0} |u_k(\tau)| \|x_k(\tau)\| d\tau \leq \frac{k}{2} \int_s^{s+t_0} |u_k(\tau)|^2 d\tau + \frac{k}{2} \int_s^{s+t_0} \|x_k(\tau)\|^2 d\tau$$

we obtain the limit (3.14).



(iv) Consider now  $u \in L^2([s, t])$ ,  $x \in L^2([s, t], X)$ ,  $t > s \geq 0$ , and extend  $u$  and  $x$  by zero to  $\mathbb{R}$ . By making a straightforward computation based on (3.1) we get

$$\begin{aligned} (\mathbb{V}_s^T u B_n(\cdot)x)(t) &= n \int_{s-\frac{1}{n}}^t \left( \Phi_{t,\tau} \left( \chi_{\mathbb{R}_+} \otimes u \left( \tau + \frac{1}{n} \right), \chi_{\mathbb{R}_+} \otimes x \left( \tau + \frac{1}{n} \right) \right) \right. \\ &\quad \left. - \Phi_{t,\tau}(\chi_{\mathbb{R}_+} \otimes u(\tau), \chi_{\mathbb{R}_+} \otimes x(\tau)) \right) d\tau \\ &= \lim_{k \rightarrow \infty} n \int_{s-\frac{1}{n}}^t \int_{\tau}^t \mathbb{T}(t, \mu) B_k(\mu) \left( u \left( \tau + \frac{1}{n} \right) \cdot x \left( \tau + \frac{1}{n} \right) \right. \\ &\quad \left. - u(\tau) \cdot x(\tau) \right) d\mu d\tau \\ &= \lim_{k \rightarrow \infty} n \int_{s-\frac{1}{n}}^t \mathbb{T}(t, \mu) B_k(\mu) \int_{s-\frac{1}{n}}^{\mu} \left( u \left( \tau + \frac{1}{n} \right) \cdot x \left( \tau + \frac{1}{n} \right) \right. \\ &\quad \left. - u(\tau) \cdot x(\tau) \right) d\mu d\tau. \end{aligned}$$

Observe that  $B_k(\cdot)$  are exactly the control operators associated with time-varying linear control system  $(\mathbb{T}, \Phi(\chi_{\mathbb{R}_+}, \cdot))$  on  $(X, X)$ ; see Remark 3.2. It follows from [31, Prop. 3.15(iii)] that

$$(\mathbb{V}_s^T u B_n(\cdot)x)(t) = \Phi_{t, s-\frac{1}{n}}(\chi_{\mathbb{R}_+}, (u \cdot x)^{(n)}),$$

with

$$(u \cdot x)^{(n)}(\sigma) = n \int_{\sigma}^{\sigma + \frac{1}{n}} (u \cdot x)(\tau) d\tau$$

satisfying the estimate

$$\|(u \cdot x)^{(n)}\|_{L^2([s, t], X)} \leq \|u \cdot x\|_{L^2([s, t], X)}.$$

Finally, by means of the latter and (2.2) we then obtain

$$\|(\mathbb{V}_s^T u B_n(\cdot)x)(t)\| \leq \beta(t_0 + 1) \|u \cdot x\|_{L^2([s, t], X)}. \quad \square$$

As for the time-invariant setting, to show that the converse of Proposition 3.6 holds we need the following concepts of regularity.

**DEFINITION 3.10.** *A well-posed time-varying bilinear system  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  is called regular (with feedthrough  $D = 0$ ) if the following limit exists in  $Y$ :*

$$(3.15) \quad \lim_{\tau \searrow 0} \frac{1}{\tau} \int_t^{t+\tau} [\mathbb{F}_t(\chi_{\mathbb{R}_+}, \chi_{\mathbb{R}_+} \otimes x)](\sigma) d\sigma = 0,$$

and absolutely regular if

$$(3.16) \quad \lim_{\tau \searrow 0} \frac{1}{\tau} \int_t^{t+\tau} \|\mathbb{F}_t(\chi_{\mathbb{R}_+}, \chi_{\mathbb{R}_+} \otimes x)(\sigma)\|_Y^2 d\sigma = 0$$

for all  $t \geq 0$  and  $x \in X$ .

**Remark 3.11.** One can remark that the (absolute) regularity of time-varying bilinear system  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  is equivalent to that of time-varying linear system  $\Gamma_{\Sigma} :=$

$(\mathbb{T}, \Phi(\chi_{\mathbb{R}_+}, \cdot), \Psi, \mathbb{F}(\chi_{\mathbb{R}_+}, \cdot))$ ; see Remark 3.2. Trivially, the triple  $(\mathbb{T}, B(\cdot), C(\cdot) := I)$  gives rise to an absolutely regular well-posed time-varying bilinear control system where  $B(\cdot)$  is an admissible control operator for  $\mathbb{T}$ ; see Remark 3.8.

We say that  $(\mathbb{T}, \Phi, \Psi)$  is an admissible (absolutely) regular triple if the associated well-posed time-varying bilinear system (see Proposition 3.6) is (absolutely) regular. For the time-invariant case, the regularity of  $\Sigma$  has been characterized by that of  $\Gamma_\Sigma$  (cf. [6, Thm. 4.8]).

Now we show a time-varying version of the representation theorem (see [6, Thm. 4.11]).

**THEOREM 3.12.** *Let  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  be a regular time-varying bilinear system and  $C(s)$  be given by (2.8). Then  $\Phi_{\cdot, s}(u, x) \in \mathcal{D}_s(C(\cdot))$  and  $\mathbb{F}_s(u, x) = C(\cdot)\Phi_{\cdot, s}(u, x)$  for all  $s \geq 0, u \in \mathbb{C}_{loc}^{2, s}$ , and  $x \in X_{loc}^{2, s}$ .*

*Proof.* Set  $\epsilon_f(\sigma, t) := f(\sigma) - f(t)$ ,  $\sigma \geq t$ . Let  $(u, x) \in \mathbb{C}_{loc}^{2, s} \times X_{loc}^{2, s}$  and  $t \geq s$ , where  $t \in P_u \cap P_x \cap P_{\mathbb{F}_s(u, x)}$  such that the regularity condition (3.15) holds at  $t$ . Here  $P_f$  is the set of 2-Lebesgue points of  $f$ ; see section 2. These intersections are not void, since its complement in  $[s, \infty)$  is negligible. Then, by invoking (3.3), we obtain

$$\mathbb{F}_s(u, x) = \Psi_t \Phi_{t, s}(u, x) + \mathbb{F}_t(\epsilon_u(\cdot, t) + \chi_{\mathbb{R}_+} \otimes u(t), \epsilon_x(\cdot, t) + \chi_{\mathbb{R}_+} \otimes x(t)) \quad \text{on } [t, \infty). \quad (3.17)$$

Set  $I(\tau, t) := \frac{1}{\tau} \left\| \int_t^{t+\tau} \mathbb{F}_t(\epsilon_u(\cdot, t) + \chi_{\mathbb{R}_+} \otimes u(t), \epsilon_x(\cdot, t) + \chi_{\mathbb{R}_+} \otimes x(t))(\sigma) d\sigma \right\|$ . We get

$$\begin{aligned} I(\tau, t) &\leq \frac{1}{\tau} \left\| \int_t^{t+\tau} \mathbb{F}_t(\epsilon_u(\cdot, t), \epsilon_x(\cdot, t) + \chi_{\mathbb{R}_+} \otimes x(t))(\sigma) d\sigma \right\| \\ &\quad + \frac{1}{\tau} \left\| \int_t^{t+\tau} \mathbb{F}_t(\chi_{\mathbb{R}_+} \otimes u(t), \epsilon_x(\cdot, t))(\sigma) d\sigma \right\| \\ &\quad + \frac{1}{\tau} \left\| \int_t^{t+\tau} \mathbb{F}_t(\chi_{\mathbb{R}_+} \otimes u(t), \chi_{\mathbb{R}_+} \otimes x(t))(\sigma) d\sigma \right\| \\ &=: I_1(\tau, t) + I_2(\tau, t) + I_3(\tau, t). \end{aligned}$$

By regularity of  $\Sigma$ , it follows that  $I_3(\tau, t)$  goes to zero when  $\tau \rightarrow 0$ . By means of Hölder's inequality and (3.4) we obtain

$$\begin{aligned} I_1^2(\tau, t) &\leq \frac{1}{\tau^2} \left( \int_t^{t+\tau} \|\mathbb{F}_t(\epsilon_u(\cdot, t), x)(\sigma)\| d\sigma \right)^2 \\ &\leq \frac{\kappa^2}{\tau} \int_t^{t+\tau} \|u(\sigma) - u(t)\|^2 d\sigma \int_t^{t+\tau} \|x(\sigma)\|^2 d\sigma. \end{aligned}$$

Thus,  $I_1(\tau, t)$  goes to zero when  $\tau \rightarrow 0$ . Similarly, we can also show that  $I_2(\tau, t)$  goes to zero when  $\tau \rightarrow 0$ . Therefore we conclude, via (3.17) and (2.8), that  $\Phi_{\cdot, s}(u, x) \in \mathcal{D}_s(C(\cdot))$  and  $\mathbb{F}_s(u, x)(\cdot) = C(\cdot)\Phi_{\cdot, s}(u, x)$  for a.e.  $t \geq s$ .  $\square$

**Remark 3.13.** By adopting a technique from the proof of [31, Prop. 3.12], it is easy to show that for an absolutely regular time-varying bilinear system

$$C(\cdot)\Phi_{\cdot, s}^n(u, x) \rightarrow \mathbb{F}_s(u, x) \quad \text{in } Y_{loc}^{2, s} \quad \text{as } n \rightarrow \infty$$

for all  $u \in \mathbb{C}_{loc}^{2, s}$ ,  $x \in X_{loc}^{2, s}$ , and  $s \geq 0$ .

Now let  $(\mathbb{T}, \Phi)$  be a time-varying bilinear control system on  $X$ , and consider the equation

$$(3.18) \quad x(t) = \mathbb{T}(t, s)x_0 + \Phi_{t, s}(u, x), \quad (t, s) \in \Delta,$$

for a given  $u \in \mathbb{C}_{loc}^{2,s}$ ,  $x_0 \in X$ , and  $s \geq 0$ . So we are looking for the function (called the state trajectory of  $(\mathbb{T}, \Phi)$ )  $x(\cdot) \in \mathcal{C}([s, \infty), X)$  solution of (3.18). In particular, if  $\Phi_{\cdot,s}(u, z) = \mathbb{V}_s^\mathbb{T} u B(\cdot) z$  for admissible control operators  $B(\cdot)$ , then  $x(\cdot)$  is a solution of the following variation of constants formula:

$$(3.19) \quad x(t) = \mathbb{T}(t, s)x_0 + \int_s^t \mathbb{T}(t, \tau)u(\tau)B(\tau)x(\tau)d\tau, \quad (t, s) \in \Delta.$$

Equation (3.19) can also be related to the absolutely regular time-varying linear system  $(\mathbb{T}, \phi, \Psi, \mathbb{L})$  on  $(X, X, X)$  where

$$\phi_{t,s} x := [\mathbb{V}_s^\mathbb{T}(B(\cdot)x)](t), \quad \Psi_s x_0 := \mathbb{T}(\cdot, s)x_0 \text{ (i.e., } C(\cdot) = I) \quad \text{and} \quad \mathbb{L}_s x := \mathbb{V}_s^\mathbb{T} B(\cdot)x$$

for  $x \in X_{loc}^{2,s}$ ,  $x_0 \in X$ , and  $(t, s) \in \Delta$ . Thus, if the input  $u \in \mathbb{C}^\infty$ , then one can apply [31, Thm. 4.4, Rem. 4.6(a)], with admissible feedback operators  $\Delta(\cdot) := u(\cdot)I$  (see [31, Def. 4.1] or Remark 4.4(ii) for definition), to show the existence of a solution to (3.19). Yet, for inputs  $u \in \mathbb{C}_{loc}^{2,s}$  one cannot apply this result since, in this case, the feedback operators  $\Delta(\cdot)$  do not fit in the setting of [31, Thm. 4.4]. Here we show that (3.18) admits a unique solution and is given by an evolution family on  $X$ .

**THEOREM 3.14.** *Let  $(\mathbb{T}, \Phi)$  be a time-varying bilinear control system on  $X$  with  $\omega_0(\mathbb{T}) < \infty$ . Then for every  $u \in \mathbb{C}_{loc}^{2,s}$  and  $x_0 \in X$  there exists an evolution family  $\mathbb{T}_u := (\mathbb{T}_u(t, s))_{t \geq s}$  on  $X$  such that  $x(\cdot) := \mathbb{T}_u(\cdot, s)x_0$  is the unique solution of the functional equation*

$$(3.20) \quad x(t) = \mathbb{T}(t, s)x_0 + \Phi_{t,s}(u, x), \quad (t, s) \in \Delta.$$

*Proof: Existence.* Define a sequence  $\mathbb{T}_u^n(t, s)$  for  $(t, s) \in \Delta$  by

$$\mathbb{T}_u^0(t, s) := \mathbb{T}(t, s) \quad \text{and} \quad \mathbb{T}_u^{n+1}(t, s)x_0 := \Phi_{t,s}(u, \mathbb{T}_u^n(\cdot, s)x_0).$$

The operators  $\mathbb{T}_u^n(t, s) \in \mathcal{L}(X)$ . By using Proposition 3.9 and Lemma 3.3, we obtain by induction

$$(3.21) \quad \|\mathbb{T}_u^n(t, s)\| \leq c_1 \left( \frac{(c_2 \|u\|_{L^2[s,t]}(t-s))^n}{n!} \right)^{\frac{1}{2}}$$

for some positive constants  $c_1$  and  $c_2$ . It follows from (3.21) that  $\sum_{n=0}^\infty \mathbb{T}_u^n(t, s)$  converges with respect to the  $\mathcal{L}(X)$ -norm and uniformly on any compact subset of  $\Delta$ . The operators  $\mathbb{T}_u(t, s) := \sum_{n=0}^\infty \mathbb{T}_u^n(t, s) \in \mathcal{L}(X)$  are strongly continuous in  $X$ , since  $(t, s) \mapsto \mathbb{T}_u^n(t, s)x_0 \in X$  is continuous.

Now let us show that  $\mathbb{T}_u(\cdot, s)x_0$  satisfies (3.18). In fact, we can write

$$(3.22) \quad \begin{aligned} \sum_{k=0}^n \mathbb{T}_u^k(t, s)x_0 &= \mathbb{T}(t, s)x_0 + \sum_{k=1}^n \mathbb{T}_u^k(t, s)x_0 \\ &= \mathbb{T}(t, s)x_0 + \sum_{k=0}^{n-1} \Phi_{t,s}(u, \mathbb{T}_u^k(\cdot, s)x_0) \\ &= \mathbb{T}(t, s)x_0 + \Phi_{t,s} \left( u, \sum_{k=0}^{n-1} \mathbb{T}_u^k(\cdot, s)x_0 \right). \end{aligned}$$

Since  $S_u^n(\cdot, s) := \sum_{k=0}^n \mathbb{T}_u^k(\cdot, s)$  and  $\mathbb{T}_u(\cdot, s)$  are strongly continuous on  $X$ , it follows that  $S_u^n(\cdot, s)x_0 \rightarrow \mathbb{T}_u(\cdot, s)x_0$  in  $X_{loc}^{2,s}$  as  $n \rightarrow \infty$ . By (3.22) and Lemma 3.3 we can show that  $\mathbb{T}_u(t, s)x_0$  verifies (3.18). Clearly property (i) of Definition 2.1 is satisfied by  $\mathbb{T}_u$ . To prove the evolution property (ii), let  $T \geq t \geq r \geq s \geq 0$ . By invoking (3.18) and (3.1) we get

$$\begin{aligned} & \mathbb{T}_u(t, r)\mathbb{T}_u(r, s)x_0 \\ &= \mathbb{T}(t, r)\mathbb{T}_u(r, s)x_0 + \Phi_{t,r}(u, \mathbb{T}_u(\cdot, r)\mathbb{T}_u(r, s)x_0) \\ &= \mathbb{T}(t, r)\mathbb{T}(r, s)x_0 + \mathbb{T}(t, r)\Phi_{r,s}(u, \mathbb{T}_u(\cdot, s)x_0) + \Phi_{t,r}(u, \mathbb{T}_u(\cdot, r)\mathbb{T}_u(r, s)x_0) \\ &= \mathbb{T}(t, s)x_0 + \mathbb{T}(t, r)\Phi_{r,s}(u, \mathbb{T}_u(\cdot, s)x_0) + \Phi_{t,r}(u, \mathbb{T}_u(\cdot, s)x_0) \\ &\quad + \Phi_{t,r}(u, \mathbb{T}_u(\cdot, r)\mathbb{T}_u(r, s)x_0 - \mathbb{T}_u(\cdot, s)x_0) \\ &= \mathbb{T}(t, s)x_0 + \Phi_{t,s}(u, \mathbb{T}_u(\cdot, s)x_0) + \Phi_{t,r}(u, \mathbb{T}_u(\cdot, r)\mathbb{T}_u(r, s)x_0 - \mathbb{T}_u(\cdot, s)x_0) \\ &= \mathbb{T}_u(t, s)x_0 + \Phi_{t,r}(u, \mathbb{T}_u(\cdot, r)\mathbb{T}_u(r, s)x_0 - \mathbb{T}_u(\cdot, s)x_0), \end{aligned}$$

which implies that

$$(3.23) \quad \mathbb{T}_u(t, r)\mathbb{T}_u(r, s)x_0 - \mathbb{T}_u(t, s)x_0 = \Phi_{t,r}(u, \mathbb{T}_u(\cdot, r)\mathbb{T}_u(r, s)x_0 - \mathbb{T}_u(\cdot, s)x_0).$$

Set  $\varphi_r(\tau) := \mathbb{T}_u(\tau, r)\mathbb{T}_u(r, s)x_0 - \mathbb{T}_u(\tau, s)x_0$  for  $\tau \in [r, t]$ . Then (3.23) and (3.2) give

$$\begin{aligned} \|\varphi_r(\tau)\|^2 &= \|\Phi_{t,r}(u, \varphi_r)\|^2 \\ &\leq c(u, T) \int_r^t \|\varphi_r(\tau)\|^2 d\tau. \end{aligned}$$

By Gronwall's lemma we obtain  $\varphi_r(\tau) = 0$  on  $[r, t]$  and  $\mathbb{T}_u(t, r)\mathbb{T}_u(r, s)x_0 = \mathbb{T}_u(t, s)x_0$ . Since  $T > 0$  and  $x_0 \in X$  are arbitrariness, we have (ii) for  $\mathbb{T}_u$ .

*Uniqueness.* Suppose that  $\tilde{x}(\cdot) \in \mathcal{C}([s, \infty), X)$  is another solution of (3.18), and set  $\xi(t) := x(t) - \tilde{x}(t)$ ; then we have  $\xi(t) = \Phi_{t,s}(u, \xi)$ . By using Lemma 3.3, we obtain

$$\|\xi(t)\|^2 \leq \text{const} \|u\|_{L^2[s,t]}^2 \int_s^t \|\xi(\tau)\|^2 d\tau.$$

Again, from Gronwall's lemma we deduce that  $\xi(t) = 0$  for all  $t \geq s$ .  $\square$

The following result shows that the regularity allows an explicit expression of the output function  $y(\cdot)$  in terms of the observation operators  $C(\cdot)$  issued from the system  $(\mathbb{T}, \Psi)$ . This extends the representation theorem obtained for time-invariant bilinear systems; see [6, Thm. 11].

**THEOREM 3.15.** *Let  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  be a well-posed time-varying bilinear system on  $(X, \mathbb{C}, Y)$  and  $\omega_0(\mathbb{T}) < \infty$ . Let  $C(\cdot)$  be the observation operator associated with  $(\mathbb{T}, \Psi)$  given by (2.8). Then, for any  $s \geq 0$ ,  $x_0 \in X$ , and  $u \in \mathbb{C}_{loc}^{2,s}$ , there is a unique solution  $(x_\Sigma, y_\Sigma) \in \mathcal{C}([s, \infty), X) \times Y_{loc}^{2,s}$  of the functional equations*

$$(3.24) \quad \begin{cases} x_\Sigma(t) = \mathbb{T}(t, s)x_0 + \Phi_{t,s}(u, x_\Sigma), & t \geq s, \\ y_\Sigma = \Psi_s x_0 + \mathbb{F}_s(u, x_\Sigma). \end{cases}$$

If, furthermore,  $\Sigma$  is regular, then  $x_\Sigma(\cdot) \in \mathcal{D}_s(C(\cdot))$ , and the output  $y_\Sigma$  satisfies

$$(3.25) \quad y_\Sigma(t) = C(t)x_\Sigma(t) \quad \text{for a.e. } t \geq s.$$

*Proof.* The existence and the uniqueness of the  $x_\Sigma(\cdot)$  solution of the first equation in (3.24) have been proved in Theorem 3.14 and that  $y_\Sigma \in Y_{loc}^{2,s}$  is trivial. Now assume

that  $\Sigma$  is regular, and let us show that  $x_\Sigma(\cdot) \in \mathcal{D}_s(C(\cdot))$  and  $y_\Sigma$  satisfies (3.25). By [31, Thm. 2.7], we have

$$(3.26) \quad \mathbb{T}(\cdot, s)x_0 \in \mathcal{D}_s(C(\cdot)) \quad \text{and} \quad (\Psi_s x_0)(\cdot) = C(\cdot)\mathbb{T}(\cdot, s)x_0.$$

By using (3.26) and applying Theorem 3.12 we obtain that  $x_\Sigma(\cdot) \in \mathcal{D}_s(C(\cdot))$  and

$$\begin{aligned} y_\Sigma(t) &= C(t)\mathbb{T}(t, s)x_0 + C(t)\Phi_{t,s}(u, x_\Sigma) \\ &= C(t)x_\Sigma(t) \end{aligned}$$

for a.e.  $t \geq s$ .  $\square$

*Remark 3.16.* If we consider a regular time-varying bilinear system  $\Sigma$ , Theorems 3.14–3.15 give us some additional information: The evolution family  $\mathbb{T}_u$  is such that  $\mathbb{T}_u(\cdot, s)x_0 \in \mathcal{D}_s(C(\cdot))$  for all  $u \in \mathbb{C}_{loc}^{2,s}$ ,  $x_0 \in X$ , and  $s \geq 0$ . Then we ask under what condition  $C(\cdot)$  is still an admissible observation operator for the evolution  $\mathbb{T}_u$  and that the observation system  $(\mathbb{T}_u, \Psi_u)$  is also represented by  $C(\cdot)$ , where  $\Psi_u$  is given by  $(\Psi_u)_s x_0 := C(\cdot)\mathbb{T}_u(\cdot, s)x_0$ .

The following proposition provides a sufficient condition for the invariance of admissibility of observation for a perturbed evolution family in terms of input  $u$ .

**PROPOSITION 3.17.** *Let  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  be a regular time-varying bilinear system on  $(X, \mathbb{C}, Y)$  and  $\omega_0(\mathbb{T}) < \infty$ . Let  $C(\cdot)$  be the observation operator associated with  $(\mathbb{T}, \Psi)$  given by (2.8). If  $u \in \mathbb{C}^2 \cup \mathbb{C}^\infty$ , then the pair  $(\mathbb{T}_u, \Psi_u)$  ( $\Psi_u$  is defined in Remark 3.16) is a time-varying observation system which is represented by  $C(\cdot)$ .*

*Proof.* By taking into account Theorems 3.14 and 3.15, we have only to verify (ii) in Definition 2.3. So, let  $x_0 \in X$  and  $t_0, s \geq 0$ ; then in view of the result above we have

$$\begin{aligned} \int_s^{s+t_0} \|[(\Psi_u)_s x_0](t)\|^2 dt &= \int_s^{s+t_0} \|y_\Sigma(t)\|^2 dt \\ &= \int_s^{s+t_0} \|(\Psi_s x_0)(t) + (\mathbb{F}_s(u, \mathbb{T}_u(\cdot, s)x_0))(t)\|^2 dt \\ &\leq 2 \int_s^{s+t_0} \left( \|(\Psi_s x_0)(t)\|^2 + \|(\mathbb{F}_s(u, \mathbb{T}_u(\cdot, s)x_0))(t)\|^2 \right) dt \\ &\leq 2 \left( \gamma^2 \|x_0\|^2 + \kappa^2 \|u\|_{L^2[s, s+t_0]}^2 \int_s^{s+t_0} \|\mathbb{T}_u(t, s)x_0\|^2 dt \right). \end{aligned}$$

By making a straightforward computation based on the fact that  $u \in \mathbb{C}^2 \cup \mathbb{C}^\infty$  and (3.21), we can show that the evolution family  $\mathbb{T}_u$  is exponentially bounded. Therefore, we obtain

$$\int_s^{s+t_0} \|[(\Psi_u)_s x_0](t)\|^2 dt \leq 2 \left( \gamma^2 + \kappa^2 M_u \frac{(e^{2\omega_u t_0} - 1)}{2\omega_u} \|u\|_{\mathbb{C}^2}^2 \right) \|x\|^2$$

for some positive constants  $M_u$  and  $\omega_u$ .

It remains to show that  $\Psi_u$  is represented by  $C(\cdot)$ . Let  $s \geq 0, t > 0$ , and  $x \in X$ . Then

$$(3.27) \quad \frac{1}{t} \int_s^{s+t} [(\Psi_u)_s x](\tau) d\tau - \frac{1}{t} \int_s^{s+t} (\Psi_s x)(\tau) d\tau = \frac{1}{t} \int_s^{s+t} [\mathbb{F}_s(u, \mathbb{T}_u(\cdot, s)x)](\tau) d\tau.$$

By using (3.7) and Hölder's inequality we obtain

$$\left\| \frac{1}{t} \int_s^{s+t} [\mathbb{F}_s(u, \mathbb{T}_u(\cdot, s)x)](\tau) d\tau \right\| \leq d(\omega, t) \|u\|_{L^2[s, s+t]} \|\mathbb{T}_u(\cdot, s)x\|_{L^2([s, s+t], X)}. \quad (3.28)$$

As the right-hand side of (3.28) tends to zero as  $t \rightarrow 0$ , identity (3.27) shows that

$$\lim_{t \rightarrow 0} \frac{1}{t} \int_s^{s+t} [(\Psi_u)_s x](\tau) d\tau = \lim_{t \rightarrow 0} \frac{1}{t} \int_s^{s+t} (\Psi_s x)(\tau) d\tau,$$

which ends the proof.  $\square$

Proposition 3.17 means that a regular time-varying (or -invariant) bilinear system  $\Sigma$  has the same state and output as the linear time-varying observation system  $(\mathbb{T}_u, \Psi_u)$  whenever the input  $u \in \mathbb{C}^2 \cup \mathbb{C}^\infty$ . Consequently, an input  $u \in \mathbb{C}^2 \cup \mathbb{C}^\infty$  is universal on  $[s, t]$  for a well-posed time-varying bilinear system  $\Sigma$  if the associated time-varying linear observation system  $(\mathbb{T}_u, \Psi_u)$  is observable on  $[s, t]$  (see the comment after Definition 4.1 about these notions). Based on this, the author in [3] gives some topological properties for observers input extending the results given in [2, 8] and in [4, 5, 17] in infinite dimensions to absolutely regular time-varying bilinear systems.

**4. Observer design.** Motivated by the results obtained for time-invariant bilinear systems [6], this section deals with the problem of synthesis of observers for more general systems, namely, for the absolutely regular time-varying bilinear systems introduced in section 3. Thus, the problem of approximating the state of an absolutely regular time-varying bilinear system is solved by a Luenberger-like observer whenever the system is detectable. Contrary to the linear case, this observer has the property that the dynamics of the error estimation depend on the inputs being applied.

**DEFINITION 4.1.** Let  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  be a well-posed time-varying bilinear system. An observer for  $\Sigma$  is given by a system  $(\tilde{\mathcal{O}})$  (not necessarily bilinear) controlled by the inputs and the outputs of  $\Sigma$ :

$$(\tilde{\mathcal{O}}) \quad \begin{cases} \zeta(t) &= \Omega(\zeta, u, y_\Sigma), & \zeta(s) = \zeta_0 \in X, \\ \hat{x}(t) &= \Xi(\zeta(t)) \end{cases}$$

such that the output  $\hat{x}$  is an approximation to the state  $x_\Sigma$  for any initial states of  $\Sigma$  and  $(\tilde{\mathcal{O}})$ . That means

$$(4.1) \quad \lim_{t \rightarrow +\infty} \|\hat{x}(t, \zeta_0, u, y_\Sigma) - x_\Sigma(t, x_0, u)\| = 0 \text{ for any } \zeta_0, x_0 \in X.$$

Let  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  be a well-posed time-varying bilinear system on  $(X, \mathbb{C}, Y)$ . A pair of points  $x_0$  and  $\bar{x}_0$  ( $x_0 \neq \bar{x}_0$ ) is *indistinguishable* if for every input  $u \in L^2[s, t]$  the associated outputs  $y_\Sigma$  and  $\bar{y}_\Sigma$  are identically equal on  $[s, t]$ . The system  $\Sigma$  with an input  $u \in \mathbb{C}_{loc}^{2,s}$  is called *observable* on  $[s, t]$  if every pair of points  $x_0$  and  $\bar{x}_0$  is distinguishable on  $[s, t]$ . In this case we say that  $u$  is a universal input for  $\Sigma$  on  $[s, t]$ .

If the time-varying linear observation system  $(\mathbb{T}, \Psi)$  is observable on  $[s, t]$  (i.e.,  $\Psi_{t,s} := \chi_{[s,t]} \Psi_s$  is injective), then automatically the null input is universal for  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  on  $[s, t]$ . It is worth mentioning that, when a well-posed time-varying (or -invariant) linear system is observable, any input distinguishes any two distinct states, while for a bilinear (generally nonlinear) system this is no longer true. Even if the system is observable, some inputs cannot distinguish between states. Furthermore,

these “bad” inputs give rise to singularities in the observation design. The study of universal inputs has been initiated by Sontag [32] for the discrete case. For finite-dimensional bilinear systems, universal inputs do exist and are generic in the analytic case [34]. Now we give a definition of a Luenberger-like observer for the well-posed time-varying bilinear systems.

DEFINITION 4.2. *Let  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  be a well-posed time-varying bilinear system, let  $x_\Sigma$  and  $y_\Sigma$  be the corresponding state trajectory and the output given by (3.24), respectively, let  $u \in \mathbb{C}_{loc}^{2,s}$ , and let  $\mathbb{K} := (\mathbb{K}_{t,s})_{(t,s) \in \Delta} \in \mathcal{L}(Y_{loc}^{2,s}, X)$ . The system*

$$(\mathcal{O}) \quad \begin{cases} \zeta(t) &= \mathbb{T}(t,s)\zeta_0 + \Phi_{t,s}(u, \zeta) + \mathbb{K}_{t,s}(C(\cdot)\zeta(\cdot) - y_\Sigma), & t \geq s, \quad \zeta_0 \in X, \\ \hat{x}(t) &= \zeta(t), & t \geq s, \end{cases}$$

*is called an exponential Luenberger-like observer for  $\Sigma$  if  $\zeta(\cdot) \in \mathcal{C}([s, \infty), X) \cap \mathcal{D}_s(C(\cdot))$  and*

$$\|\hat{x}(t) - x_\Sigma(t)\| \leq d \|\zeta_0 - x_0\| e^{-\delta t}, \quad t \geq s,$$

*for any  $\zeta_0, x_0 \in X$  with constants  $d, \delta > 0$ .*

As in the finite-dimensional case, the observer system is obtained as an output injection of the original system where the difference  $\hat{y} - y_\Sigma$  (with  $\hat{y} := C(\cdot)\zeta(\cdot)$ ) is another input used to correct the estimated state  $\hat{x}(\cdot)$ .

Notice that, by taking the difference between the equations  $(\mathcal{O})$  and (3.24), we formally obtain the system which describes the estimation error

$$(4.2) \quad \varepsilon_{\Sigma^K}(t) = \mathbb{T}(t,s)\varepsilon_0 + \Phi_{t,s}(u, \varepsilon_{\Sigma^K}) + \mathbb{K}_{t,s}(\hat{y} - y_\Sigma)$$

for a.e.  $t \geq s$ , where  $\varepsilon_0 := \zeta_0 - x_0$ .

Contrary to the observer system  $(\mathcal{O})$ , we can show that the system (4.2) augmented by the output  $h_{\Sigma^K} := \hat{y} - y_\Sigma$  generates a well-posed time-varying bilinear system  $\Sigma^K$  in the sense of Theorem 3.15; i.e., the solution  $\varepsilon_{\Sigma^K}(\cdot)$  and the output function  $h_{\Sigma^K}$  associated with  $\Sigma^K$ , as in (3.24), are given by

$$\begin{cases} \varepsilon_{\Sigma^K}(t) &= \mathbb{T}^K(t,s)\varepsilon_0 + \Phi_{t,s}^K(u, \varepsilon_{\Sigma^K}), \\ h_{\Sigma^K} &= \Psi_s^K \varepsilon_0 + \mathbb{F}_s^K(u, \varepsilon_{\Sigma^K}) \end{cases}$$

for  $t \geq s$ .

DEFINITION 4.3. *Let  $(\mathbb{T}, \Psi)$  be a time-varying observation system. The family  $\mathbb{K} = (\mathbb{K}_{t,s})_{(t,s) \in \Delta} \in \mathcal{L}(Y_{loc}^{2,s}, X)$  is called admissible for  $(\mathbb{T}, \Psi)$  if the following hold:*

- (i)  *$(\mathbb{T}, \mathbb{K}, \Psi)$  is an admissible regular linear triple, and*
- (ii)  *$I$  is an admissible feedback for the time-varying linear system  $(\mathbb{T}, \mathbb{K}, \Psi, \mathbb{L}^K)$  where the notions in (i) and (ii) are defined and commented in the following remark.*

Remark 4.4.

- (i) As for the bilinear case (see Remark 3.11), condition (i) means that the system  $(\mathbb{T}, \mathbb{K}, \Psi, \mathbb{L}^K)$ , where  $\mathbb{L}_s^K y := C(\cdot)\mathbb{K}_{\cdot,s}y$ ,  $y \in Y_{loc}^{2,s}$ , is a time-varying regular linear system (see Definition 2.4), with  $C(\cdot)$  is defined by (2.8).
- (ii) Condition (ii) of Definition 4.3 means that there exists  $t_0 > 0$  such that the operators  $I - \mathbb{L}_{s+t_0,s}^K$ ,  $s \geq 0$ , are invertible on  $L^2([s, s+t_0], Y)$  and have uniformly bounded inverses. This is equivalent to the invertibility of  $I - \mathbb{L}_{s+t_1,s}^K$  on  $L^2([s, s+t_1], Y)$  for all  $t_1 > 0$  [31, Lem. 4.2]. A time-invariant version of this result was given in the paper of Staffans and Weiss [33] about the flow inversion of abstract linear systems.

Now we state the following result on the existence of the so-called feedback bilinear system without any regularity assumption on the open-loop system.

**THEOREM 4.5.** *Let  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  be a well-posed time-varying bilinear system. If a family  $\mathbb{K} = (\mathbb{K}_{t,s})_{(t,s) \in \Delta}$  is admissible for  $(\mathbb{T}, \Psi)$ , then the quadruple  $\Sigma^K := (\mathbb{T}^K, \Phi^K, \Psi^K, \mathbb{F}^K)$  defined by*

$$(4.3) \quad \begin{aligned} \mathbb{F}_s^K &:= (I - \mathbb{L}_s^K)^{-1} \mathbb{F}_s, & \Psi_s^K &:= (I - \mathbb{L}_s^K)^{-1} \Psi_s, \\ \Phi_{t,s}^K &:= \Phi_{t,s} + \mathbb{K}_{t,s} \mathbb{F}_s^K, & \mathbb{T}^K(t, s) &:= \mathbb{T}(t, s) + \mathbb{K}_{t,s} \Psi_s^K \end{aligned}$$

for all  $(t, s) \in \Delta$  is a well-posed time-varying bilinear system, where  $\mathbb{L}_s^K$  is the linear input-output map defined in Remark 4.4(i). Moreover, if  $\Sigma$  is absolutely regular, then  $\Sigma^K$  is so as well.

*Proof.* Let us prove that the four operators in (4.3) satisfy the conditions of Definition 3.1. Thus, the two upper identities in (4.3) give

$$(4.4) \quad \mathbb{F}_s^K(u, x) = \mathbb{F}_s(u, x) + \mathbb{L}_s^K \mathbb{F}_s^K(u, x) \quad \text{and} \quad \Psi_s^K x_0 = \Psi_s x_0 + \mathbb{L}_s^K \Psi_s^K x_0$$

for all  $u \in \mathbb{C}_{loc}^{2,s}$ ,  $x \in X_{loc}^{2,s}$ ,  $x_0 \in X$ , and  $s \geq 0$ . By combining the first identity in (4.4) with the properties (2.5) and (3.3), we obtain

$$(4.5) \quad \begin{aligned} \mathbb{F}_s^K(u, x) &= \mathbb{F}_t(u, x) + \Psi_t \Phi_{t,s}(u, x) + \mathbb{L}_t^K \mathbb{F}_s^K(u, x) + \Psi_t \mathbb{K}_{t,s} \mathbb{F}_s^K(u, x) \\ &= \mathbb{F}_t(u, x) + \Psi_t (\Phi_{t,s}(u, x) + \mathbb{K}_{t,s} \mathbb{F}_s^K(u, x)) + \mathbb{L}_t^K \mathbb{F}_s^K(u, x) \\ &= \mathbb{F}_t(u, x) + \Psi_t \Phi_{t,s}^K + \mathbb{L}_t^K \mathbb{F}_s^K(u, x) \end{aligned}$$

on  $[t, \infty)$ , which implies that

$$(I - \mathbb{L}_t^K) \mathbb{F}_s^K(u, x) = \mathbb{F}_t(u, x) + \Psi_t \Phi_{t,s}^K \quad \text{on } [t, \infty).$$

Thus, the definitions of  $\mathbb{F}_t^K$  and  $\Psi_t^K$  in (4.3) give the equation required in (3.3) for  $\mathbb{F}_s^K$ .

Now, by using the second identity in (4.4) and applying (2.5) and (2.3), we obtain

$$(4.6) \quad \begin{aligned} \Psi_s^K x_0 &= \Psi_t \mathbb{T}(t, s) x_0 + \mathbb{L}_t^K \Psi_s^K x_0 + \Psi_t \mathbb{K}_{t,s} \Psi_s^K x_0 \\ &= \Psi_t (\mathbb{T}(t, s) + \mathbb{K}_{t,s} \Psi_s^K) x_0 + \mathbb{L}_t^K \Psi_s^K x_0 \end{aligned}$$

on  $[t, \infty)$  for  $(t, s) \in \Delta$  and  $x_0 \in X$ .

Therefore, the equation required in (2.3) for  $\Psi_s^K$  follows from the definitions of  $\mathbb{T}^K(t, s)$  and  $\Psi_t^K$  in (4.3).

By combining now (2.1) and (3.1) we obtain

$$\begin{aligned} \Phi_{t,s}^K(u, x) &= \Phi_{t,s}(u, x) + \mathbb{K}_{t,s} \mathbb{F}_s^K(u, x) \\ &= \Phi_{t,l}(u, x) + \mathbb{T}(t, l) \Phi_{l,s}(u, x) + \mathbb{K}_{t,l} \mathbb{F}_s^K(u, x) + \mathbb{T}(t, l) \mathbb{K}_{l,s} \mathbb{F}_s^K(u, x) \end{aligned}$$

for  $t \geq l \geq s$ . By using (3.3) for  $\mathbb{F}_s^K$  and definitions of  $\Phi^K$  and  $\mathbb{T}^K$  in (4.3), it follows that

$$(4.7) \quad \begin{aligned} \Phi_{t,s}^K(u, x) &= \Phi_{t,l}(u, x) + \mathbb{T}(t, l) \Phi_{l,s}(u, x) + \mathbb{K}_{t,l} (\mathbb{F}_l^K(u, x) \\ &\quad + \Psi_l^K \Phi_{l,s}^K(u, x)) + \mathbb{T}(t, l) \mathbb{K}_{l,s} \mathbb{F}_s^K(u, x) \\ &= \Phi_{t,l}^K(u, x) + \mathbb{T}(t, l) (\Phi_{l,s}(u, x) + \mathbb{K}_{l,s} \mathbb{F}_s^K(u, x)) + \mathbb{K}_{t,l} \Psi_l^K \Phi_{l,s}^K(u, x) \\ &= \Phi_{t,l}^K(u, x) + (\mathbb{T}(t, l) + \mathbb{K}_{t,l} \Psi_l^K) \Phi_{l,s}^K(u, x) \\ &= \Phi_{t,l}^K(u, x) + \mathbb{T}^K(t, l) \Phi_{l,s}^K(u, x) \end{aligned}$$



for  $t \geq l \geq s$ , which proves (3.1) for  $\Phi_{t,s}^K$ . Since  $(I - \mathbb{L}_s^K)^{-1}$  is uniformly bounded on  $L^2([s, s+t], Y)$  and  $\Sigma$  is a well-posed time-varying bilinear system, the estimations (2.4), (3.2), and (3.4) hold for  $\Psi_s^K$ ,  $\Phi_{t,s}^K$ , and  $\mathbb{F}_s^K$ , respectively. It remains to show that  $\mathbb{T}^K$  is an evolution family on  $X$ . Due to the fact that  $\mathbb{K}_{t,t} = 0$  and  $\mathbb{T}(t, t) = I$ , we obtain  $\mathbb{T}^K(t, t) = I$ . By applying (2.1) and (2.3) for  $(\mathbb{T}, \mathbb{K})$  and  $(\mathbb{T}^K, \Psi^K)$ , respectively, we get

$$\begin{aligned} \mathbb{T}^K(t, s)x_0 &= \mathbb{T}(t, s)x_0 + \mathbb{K}_{t,s}\Psi_s^K x_0 \\ &= \mathbb{T}(t, r)\mathbb{T}(r, s)x_0 + \mathbb{K}_{t,r}\Psi_r^K x_0 + \mathbb{T}(t, r)\mathbb{K}_{r,s}\Psi_s^K x_0 \\ &= \mathbb{T}(t, r)(\mathbb{T}(r, s) + \mathbb{K}_{r,s}\Psi_s^K)x_0 + \mathbb{K}_{t,r}\Psi_r^K x_0 \\ &= \mathbb{T}(t, r)\mathbb{T}^K(r, s)x_0 + \mathbb{K}_{t,r}\Psi_r^K \mathbb{T}^K(r, s)x_0 \\ &= \mathbb{T}^K(t, r)\mathbb{T}^K(r, s)x_0 \end{aligned}$$

for all  $(t, s) \in \Delta$  and  $x_0 \in X$ . The strong continuity of  $\mathbb{T}^K$  is obtained as in [31, Thm. 4.4].

Finally, the definition of  $\mathbb{F}^K$  in (4.3) yields

$$\int_t^{t+\tau} \|\mathbb{F}_t^K(\chi_{\mathbb{R}_+}, \chi_{\mathbb{R}_+} \otimes x)(\sigma)\|^2 d\sigma \leq \text{const} \int_t^{t+\tau} \|\mathbb{F}_t(\chi_{\mathbb{R}_+}, \chi_{\mathbb{R}_+} \otimes x)(\sigma)\|^2 d\sigma.$$

Thus  $\Sigma^K$  is a well-posed bilinear system, and if  $\Sigma$  is absolutely regular, then  $\Sigma^K$  is so as well.  $\square$

*Remark 4.6.* In the time-invariant setting (see [6, sect. 5]) we have proved that the regularity is even conserved by the systems  $\Sigma^K$ . This was established by means of the transfer functions introduced for well-posed time-invariant bilinear systems  $\Sigma^K$  and its relation with that of  $\Sigma$ ; see [6, Rem. 6(ii)]. But the proof is essentially based on the Laplace transform of the input-output maps and the use of [6, Prop. 4.9], which has no version in our setting.

We now show more representation of  $\Sigma^K$ . We begin by giving the expressions of  $\Psi^K$ ,  $\mathbb{F}^K$ , and the observation operators  $C^K(\cdot)$  associated with  $(\mathbb{T}^K, \Psi^K)$ . This requires the initial observation operators associated with  $(\mathbb{T}, \Psi)$ . We show that, if  $\Phi$  is represented by an admissible control operators,  $\Phi^K$  has a representation by the same control operators. This is the time-varying analogue of the result given in [6, Prop. 5.5].

**PROPOSITION 4.7.** *Let  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  be an absolutely regular time-varying bilinear system and  $\omega_0(\mathbb{T}) < \infty$ . Let  $C(\cdot)$  be the observation operator associated with  $(\mathbb{T}, \Psi)$  and  $\mathbb{K}$  is admissible for  $(\mathbb{T}, \Psi)$ , and let  $\Sigma^K$  be the absolutely regular time-varying bilinear system given in Theorem 4.5. Then the following hold:*

- (i)  $\mathbb{T}^K(\cdot, s)x_0$  and  $\Phi_{\cdot,s}^K(u, x)$  belong to  $\mathcal{D}_s(C(\cdot))$ , and we have

$$(4.8) \quad \Psi_s^K x_0 = C(\cdot)\mathbb{T}^K(\cdot, s)x_0, \quad \mathbb{F}_s^K(u, x) = C(\cdot)\Phi_{\cdot,s}^K(u, x)$$

for all  $x_0 \in X, u \in \mathbb{C}_{loc}^{2,s}$ , and  $x \in X_{loc}^{2,s}$ .

- (ii) For any  $z_0 \in X, u \in \mathbb{C}_{loc}^{2,s}$ , the unique solutions  $(z_{\Sigma^K}, h_{\Sigma^K}) \in \mathcal{C}([s, \infty), X) \times Y_{loc}^{2,s}$  of the functional equations

$$(4.9) \quad \begin{cases} z_{\Sigma^K}(t) = \mathbb{T}^K(t, s)z_0 + \Phi_{t,s}^K(u, z_{\Sigma^K}), \\ h_{\Sigma^K} = \Psi_s^K z_0 + \mathbb{F}_s^K(u, z_{\Sigma^K}) \end{cases}$$

are such that  $z_{\Sigma^K} \in \mathcal{D}_s(C(\cdot)) \cap \mathcal{D}_s(C^K(\cdot))$  and  $h_{\Sigma^K}(t) = C(t)z_{\Sigma^K}(t)$  for a.e.  $t \geq s$ .

- (iii) Let  $\overline{X}$  be a Banach space in which  $X$  is dense and continuously embedded. Assume that  $\Phi_{t,s}$  is represented by admissible control operators  $B(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(X, \overline{X}))$  (see Remark 3.8(i)) and  $\mathbb{T}^K(t, s)$  has a locally uniformly bounded extension  $\overline{\mathbb{T}}^K(t, s) : \overline{X} \rightarrow \overline{X}$ . Thus,  $\overline{\mathbb{T}}^K$  satisfies (i) and (ii) of Definition 2.1. Then

$$(4.10) \quad \Phi_{\cdot,s}^K(u, x) = \mathbb{V}_s^{\overline{\mathbb{T}}^K}(uB(\cdot)x) \quad \text{and} \quad \mathbb{F}_s^K(u, x) = C(\cdot)\mathbb{V}_s^{\overline{\mathbb{T}}^K}(uB(\cdot)x)$$

for all  $(u, x) \in \mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s}$ .

*Proof.* (i) From identity (4.4) we have, for all  $x_0 \in X$ ,

$$\begin{aligned} \Psi_s^K x_0 &= \Psi_s x_0 + \mathbb{L}_s^K \Psi_s^K x_0 \\ &= C(\cdot)\mathbb{T}(\cdot, s)x_0 + C(\cdot)\mathbb{K}_{\cdot,s}\Psi_s^K x_0 && \text{(by Remark 4.4(i))} \\ &= C(\cdot)(\mathbb{T}(\cdot, s) + \mathbb{K}_{\cdot,s}\Psi_s^K)x_0 \\ &= C(\cdot)\mathbb{T}^K(\cdot, s)x_0 && \text{(by (4.3)),} \end{aligned}$$

and this shows that  $\mathbb{T}^K(\cdot, s)x_0 \in \mathcal{D}_s(C(\cdot))$ .

By using again (4.4) and Remark 4.4(i) and applying Theorem 3.12, we obtain

$$\begin{aligned} \mathbb{F}_s^K(u, x) &= \mathbb{F}_s(u, x) + \mathbb{L}_s^K \mathbb{F}_s^K(u, x) \\ &= C(\cdot)\Phi_{\cdot,s}(u, x) + C(\cdot)\mathbb{K}_{\cdot,s}\mathbb{F}_s^K(u, x) \\ &= C(\cdot)(\Phi_{\cdot,s}(u, x) + \mathbb{K}_{\cdot,s}\mathbb{F}_s^K(u, x)) \\ &= C(\cdot)\Phi_{\cdot,s}^K(u, x) && \text{(by (4.3))} \end{aligned}$$

for all  $u \in \mathbb{C}_{loc}^{2,s}$  and  $x \in X_{loc}^{2,s}$ . Therefore  $\Phi_{\cdot,s}^K(u, x) \in \mathcal{D}_s(C(\cdot))$ .

(ii) As  $\omega_0(\mathbb{T}) < \infty$  we deduce from [31, Thm. 4.4] that  $\omega_0(\mathbb{T}^K) < \infty$ . Then, according to Theorems 4.5 and 3.15, the functional equations (4.9) have a unique solution  $(z_{\Sigma^K}, h_{\Sigma^K})$  such that  $z_{\Sigma^K} \in \mathcal{D}_s(C^K(\cdot))$  and  $h_{\Sigma^K}(t) = C^K(t)z_{\Sigma^K}(t)$  for a.e.  $t \geq s$ . From (i) it follows that  $z_{\Sigma^K}(\cdot) \in \mathcal{D}_s(C(\cdot))$ , and, for a.e.  $t \geq s$ , we obtain

$$\begin{aligned} h_{\Sigma^K}(t) &= (\Psi_s^K z_0)(t) + \mathbb{F}_s^K(u, z_{\Sigma^K})(t) \\ &= C(t)\mathbb{T}^K(t, s)z_0 + C(t)\Phi_{t,s}^K(u, z_{\Sigma^K}) \\ &= C(t)z_{\Sigma^K}(t). \end{aligned}$$

(iii) By [31, Prop. 2.11] we can see that  $\mathbb{V}_s^{\mathbb{T}} B_k(u, x) \in \mathcal{D}_s(C(\cdot))$ , and by proceeding as in the proof of [31, Form. (4.14)] we can obtain

$$(4.11) \quad \mathbb{V}_s^{\mathbb{T}^K} B_k(u, x) = \mathbb{V}_s^{\mathbb{T}} B_k(u, x) + \mathbb{K}_{\cdot,s}(I - \mathbb{L}_{s+t_1,s}^K)^{-1} C(\cdot) \mathbb{V}_s^{\mathbb{T}} B_k(u, x)$$

for all  $u \in \mathbb{C}_{loc}^{2,s}$ ,  $x \in X_{loc}^{2,s}$ , and  $t_1$  large enough. By using Proposition 3.9 and Remark 3.13 and taking the limit in (4.11) on  $\mathcal{C}([s, s+t_1], X)$ , we obtain

$$(4.12) \quad \begin{aligned} \Phi_{t,s}^K(u, x) &= \lim_{k \rightarrow \infty} [\mathbb{V}_s^{\mathbb{T}^K} B_k(u, x)](t) \\ &= \Phi_{t,s}(u, x) + \mathbb{K}_{t,s}(I - \mathbb{L}_{s+t_1,s}^K)^{-1} \mathbb{F}_s(u, x). \end{aligned}$$

Hence, it follows from Remark 4.4(i) that  $\Phi_{\cdot,s}^K(u, x) \in \mathcal{D}_s(C(\cdot))$ . Our aim now is to show that for all  $(u, x) \in \mathcal{A}^{2,s} := \{(u, x) \in \mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s} \mid u.x \in X_{loc}^{2,s}\}$

$$(4.13) \quad B_k(u, x) \rightarrow uB(\cdot)x \quad \text{as } k \rightarrow \infty \quad \text{in } \overline{X}_{loc}^{2,s},$$

where the operator  $B_k$  is defined in (3.11). Let  $(u, x) \in \mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s}$ , and set  $D_k(u, x) := B_k(u, x) - uB(\cdot)x$ . Then  $D_k(u, x) \in \overline{X}_{loc}^{2,s}$  if  $(u, x) \in \mathcal{A}^{2,s}$ . By means of Hölder's inequality and Fubini's theorem we estimate

$$\begin{aligned} \|D_k(u, x)\|_{L^2([0,T], \overline{X})} &= k^2 \int_0^T \left\| \int_{\tau-\frac{1}{k}}^\tau \overline{\mathbb{T}}(\tau, \mu) u(\mu) B(\mu) x(\mu) d\mu - u(\tau) B(\tau) x(\tau) \right\|^2 d\tau \\ &\leq k \int_0^{\frac{1}{k}} \int_0^T \|\overline{\mathbb{T}}(\tau + \mu, \tau) u(\tau) B(\tau) x(\tau) - u(\tau + \mu) B(\tau + \mu) x(\tau + \mu)\|^2 d\tau d\mu, \end{aligned}$$

which gives (4.13). By combining now (4.12) and (4.13) we obtain

$$(4.14) \quad \mathbb{V}_s^{\mathbb{T}^K} B_k(u, x)(t) = \mathbb{V}_s^{\overline{\mathbb{T}}^K} B_k(u, x)(t) \rightarrow (\mathbb{V}_s^{\overline{\mathbb{T}}^K} uB(\cdot)x)(t) \text{ as } k \rightarrow \infty.$$

Thus  $\Phi_{\cdot, s}^K(u, x) = \mathbb{V}_s^{\overline{\mathbb{T}}^K} (uB(\cdot)x)$  for all  $(u, x) \in \mathcal{A}^{2,s}$ .

Let  $\mathcal{E} := \mathcal{C}_c([s, \infty), \mathbb{C}) \times \mathcal{C}_c([s, \infty), X)$ , where  $\mathcal{C}_c([s, \infty), X)$  is the space of continuous functions with compact support in  $[s, \infty)$ . Since the space  $\mathcal{E}$  is dense in  $\mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s}$  and is included in  $\mathcal{A}^{2,s}$ , we deduce that this latter is also dense in  $\mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s}$ . Thus, by the continuity of both sides of (4.14) we obtain the first equality in (4.10). The second equality in (4.10) is immediate from the first equality.  $\square$

We now turn back to the problem of the observer. This leads us to introduce the following concept of detectability.

**DEFINITION 4.8.** *Let  $(\mathbb{T}, \Psi)$  be a time-varying observation system. We say that  $(\mathbb{T}, \Psi)$  is admissibly detectable if there exists a family  $\mathbb{K}$  such that:*

- (i)  $\mathbb{K}$  is admissible for  $(\mathbb{T}, \Psi)$  (see Definition 4.3);
- (ii) the evolution family  $\mathbb{T}^K$  on  $X$  defined in (4.3) is such that  $\omega_0(\mathbb{T}^K) < 0$ .

For the time-invariant Hilbert space setting, Rebarber [28, Def. 1.5] has introduced a somewhat general version of detectability. However, in the time-varying Banach space setting, Schnaubelt [31, Def. 5.8] has proposed a more general definition of detectability with no regularity assumption on the system. But one can remark that if the pair  $(\mathbb{T}, \Psi)$  is detectable in the sense of Definition 4.8, then, by Theorem 4.5, the closed-loop evolution family

$$\mathbb{T}^K(t, s) = \mathbb{T}(t, s) + \mathbb{K}_{t,s}(I - \mathbb{L}_s^K)^{-1} \Psi_s, \quad (t, s) \in \Delta,$$

is exponentially stable on  $X$ . By letting  $\phi_{t,s}^K := \mathbb{K}_{t,s}(I - \mathbb{L}_s^K)^{-1}$ , it is easy to see that the pair  $(\mathbb{T}^K, \phi^K)$  is a linear control system on  $X$  which implies that  $(\mathbb{T}, \Psi)$  is detectable in the sense of [31, Def. 5.8]. It is not known if Schnaubelt's definition is genuinely more general; i.e., there exists a pair  $(\mathbb{T}, \Psi)$  which is detectable in the sense of [31, Def. 5.8] but it is not in the sense of Definition 4.8.

The problem to design an observer for any input holds even in the finite-dimensional case; see, e.g., [16, 18]. The following result gives sufficient conditions for the existence of a Luenberger-like observer for an absolutely regular time-varying bilinear system.

**THEOREM 4.9.** *Let  $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$  be an absolutely regular time-varying bilinear system and  $\omega_0(\mathbb{T}) < \infty$ . If  $(\mathbb{T}, \Psi)$  is admissibly detectable by  $\mathbb{K}$ , then the system  $(\mathcal{O})$  is an exponential Luenberger-like observer for  $\Sigma$  working for some small inputs.*

*Proof.* First, for  $\varepsilon_0 \in X$  and  $s \geq 0$ , we are looking for a function  $\varepsilon_{\Sigma^K}(\cdot) \in \mathcal{D}_s(C(\cdot)) \cap \mathcal{D}_s(C^K(\cdot))$  satisfying (4.2). So let  $\Sigma^K$  be the absolutely regular time-varying bilinear systems associated with  $\Sigma$  as defined in Theorem 4.5 and consider

the system

$$(4.15) \quad \begin{cases} \varepsilon_{\Sigma^K}(t) &= \mathbb{T}^K(t, s)\varepsilon_0 + \Phi_{t,s}^K(u, \varepsilon_{\Sigma^K}), \\ h_{\Sigma^K} &= \Psi_s^K \varepsilon_0 + \mathbb{F}_s^K(u, \varepsilon_{\Sigma^K}) \end{cases}$$

for  $u \in \mathbb{C}_{loc}^{2,s}$ . According to Theorem 3.15, the solution  $\varepsilon_{\Sigma^K}(\cdot)$  of  $\Sigma^K$  associated with  $u$  exists in  $\mathcal{C}([s, \infty), X)$ , and, via Proposition 4.7(i), it fits in  $\mathcal{D}_s(C(\cdot)) \cap \mathcal{D}_s(C^K(\cdot))$ . Therefore, we have  $h_{\Sigma^K}(t) = C^K(t)\varepsilon_{\Sigma^K}(t) = C(t)\varepsilon_{\Sigma^K}(t)$  for a.e.  $t \geq s$ . Again by Theorem 3.15, the solution  $x_{\Sigma}(\cdot)$  of  $\Sigma$  associated with  $u$  exists in  $\mathcal{C}([s, \infty), X) \cap \mathcal{D}_s(C^K(\cdot))$ . Thus,  $\zeta(\cdot) := x_{\Sigma}(\cdot) + \varepsilon_{\Sigma^K}(\cdot) \in \mathcal{C}([s, \infty), X) \cap \mathcal{D}_s(C(\cdot))$ , and the output function is then given by

$$(4.16) \quad \begin{aligned} h_{\Sigma^K}(t) &= C^K(t)\varepsilon_{\Sigma^K}(t) \\ &= C(t)(\zeta(t) - x_{\Sigma}(t)) \\ &= \hat{y}(t) - y_{\Sigma}(t) \end{aligned}$$

for a.e.  $t \geq s$ . On the other hand, by using the first equation in (4.15) and (4.3), we can write

$$\begin{aligned} \varepsilon_{\Sigma^K}(t) &= \mathbb{T}(t, s)\varepsilon_0 + \Phi_{t,s}(u, \varepsilon_{\Sigma^K}) + \mathbb{K}_{t,s}(\Psi_s^K \varepsilon_0 + \mathbb{F}_s^K(u, \varepsilon_{\Sigma^K})) \\ &= \mathbb{T}(t, s)\varepsilon_0 + \Phi_{t,s}(u, \varepsilon_{\Sigma^K}) + \mathbb{K}_{t,s}h_{\Sigma^K}. \end{aligned}$$

By combining this with (4.16), we deduce that  $\varepsilon_{\Sigma^K}(\cdot)$  is the unique solution of (4.2) and that  $\zeta(\cdot)$  is the unique solution of the observer  $(\mathcal{O})$ . Now, by detectability and Lemma 3.3,  $\delta := \sup_{(t,s) \in \Delta} \|\Phi_{t,s}^K\|_{\mathcal{BL}(L^2(s,t) \times L^2(s,t;X), X)}$  is finite, and, via (4.15), we have

$$\|\varepsilon_{\Sigma^K}(t)\| \leq M e^{-\theta(t-s)} \|\varepsilon_0\| + \delta \|u\|_{\mathbb{C}^{2,s}} \left( \int_s^t \|\varepsilon_{\Sigma^K}(\sigma)\|^2 d\sigma \right)^{\frac{1}{2}},$$

where  $\omega := \omega_0(\mathbb{T}^K) < 0$ ,  $\theta \in ]0, -\omega[$ , and  $M \geq 1$ . By taking the square of both terms of the above inequality and using Gronwall's lemma, it then follows that

$$\|\varepsilon_{\Sigma^K}(t)\| \leq M\sqrt{2} e^{(\delta^2 \|u\|_{\mathbb{C}^{2,s}}^2 - \epsilon)(t-s)} \|\varepsilon_0\|.$$

Thus, the system  $(\mathcal{O})$  is an exponential observer for  $\Sigma$  for all inputs satisfying  $\|u\|_{\mathbb{C}^{2,s}} < \varrho$ , with  $\varrho := \frac{\sqrt{-\omega}}{\delta}$ .  $\square$

**5. Application.** We consider the following partial differential equation with mixed boundary conditions:

$$(5.1) \quad \begin{cases} \partial_t^1 \phi(t, \xi) = \alpha(t) \partial_{\xi}^2 \phi(t, \xi) & \text{for } t \geq s, \xi \in [0, 1], \\ \phi(t, 0) = 0, \quad \partial_{\xi}^1 \phi(t, 1) + u(t) \int_0^1 \pi(\xi) \phi(t, \xi) d\xi = 0 & \text{for } t \geq s, \\ \phi(s, \xi) = \varphi(\xi) & \text{for } \xi \in [0, 1], \end{cases}$$

where  $\partial_{\xi}^i := \frac{\partial^i}{\partial \xi^i}$ ,  $i = 1, 2$ ,  $\varphi \in X := L^1(0, 1)$ , the input function  $u \in L_{loc}^2(\mathbb{R}_+)$ , and the coefficient  $\alpha \in \mathcal{C}(\mathbb{R}_+)$  is a strictly positive function such that  $0 < \alpha_0 \leq \alpha(t) \leq \alpha_m$  for all  $t \geq 0$ . For simplicity, we assume that the boundary coefficient  $\pi \in L^\infty(0, 1)$ . Let  $A := \partial_{\xi}^2$  with  $D(A) := \{f \in W^{2,1}[0, 1], f(0) = f'(1) = 0\}$ . It is known that  $(A, D(A))$  generates an analytic semigroup  $(S(t))_{t \geq 0}$  on the space  $X$ . For the

reformulation of our system in bilinear form we define the space  $X_{-1}$  as the completion of  $X$  with respect to the new norm  $\|x\|_{-1} := \|(\lambda I - A)^{-1}x\|$  for some  $\lambda$  fixed in  $\rho(A)$ . This space is independent of the choice of  $\lambda$  and is called the *extrapolation space* of  $X$  with respect to  $A$ . The semigroup  $(S(t))_{t \geq 0}$  can be extended to a  $C_0$ -semigroup  $(S_{-1}(t))_{t \geq 0}$  on  $X_{-1}$ , with the generator denoted by  $(A_{-1}, D(A_{-1}))$ . Notice that  $A_{-1}$  is an extension of  $A$  to  $X$ , with  $D(A_{-1}) = X$ , and the norm of  $X$  is equivalent to the graph norm of  $A_{-1}$ . On the other hand, let  $X_a$  (respectively,  $X_{b-1}$ ) be the domain of the fractional power  $(\lambda - A)^a$  (respectively,  $(\lambda - A_{-1})^b$ ), endowed with the graph norm, with  $a, b \in [0, 1]$ . Then it is known that

$$(5.2) \quad S_{-1}(t) \in \mathcal{L}(X_{b-1}, X_a) \quad \text{and}$$

$$(5.3) \quad \|S(t)(\lambda - A)^b x\|_{X_a} \leq h e^{\omega t} \max\{1, t^{-(a+b)}\} \|x\|$$

for all  $t \geq 0$  and  $x \in X$ , where  $h = h(\lambda) > 0$ ,  $\omega \geq 0$ , and  $a, b \in [0, 1]$ . For more details we can see, e.g., [15, sect. II.5] and [1].

Here we suppose that our system (5.1) is coupled with the output function

$$(5.4) \quad y(t) = \int_0^1 \xi^\theta \phi(t, \xi) d\xi$$

for  $t \geq s$ , where  $\theta \in ]-1, 0[$ . We will show that (5.1) and (5.4) can be reformulated as a well-posed time-varying bilinear system  $\Sigma$ .

Let  $t \geq 0$  and  $\lambda > 0$ . We first remark that the solution of the elliptic problem

$$(5.5) \quad (\lambda - \alpha(t)\partial_\xi^2)\beta_{\lambda,t} = 0, \quad \text{with} \quad \beta_{\lambda,t}(0) = 0 \quad \text{and} \quad \partial_\xi^1 \beta_{\lambda,t}(1) = 1,$$

exists, and it is explicitly given by

$$(5.6) \quad \beta_{\lambda,t}(\xi) = \sqrt{\frac{\alpha(t)}{\lambda}} \frac{\sinh(\sqrt{\frac{\lambda}{\alpha(t)}} \xi)}{\cosh(\sqrt{\frac{\lambda}{\alpha(t)}})}.$$

In what follows, we fix  $\lambda > 0$ , and we introduce the time-varying operators

$$A(t) := \alpha(t)A_{-1}, \quad \text{with} \quad D(A(t)) := D(A_{-1}) = X, \quad t \geq 0,$$

and

$$B_\lambda(t) := (\lambda - A(t))(\beta_{\lambda,t} \otimes L),$$

where

$$L(f) := \int_0^1 \pi(\xi) f(\xi) d\xi \quad \text{and} \quad (\beta_{\lambda,t} \otimes L)f := L(f)\beta_{\lambda,t}, \quad f \in X.$$

Thus, we obtain the following result, which shows that the system (5.1) is equivalent to some time-varying bilinear system.

**PROPOSITION 5.1.** *If  $z$  is a solution of (5.1), i.e.,  $z \in \mathcal{C}^1([s, \infty), X) \cap \mathcal{C}([s, \infty), \mathcal{C}^2[0, 1])$  and verifies (5.1), then  $x(t) := z(t, \cdot)$  is a solution of the bilinear system*

$$(5.7) \quad \dot{x}(t) = A(t)x(t) + u(t)B_\lambda(t)x(t), \quad x(s) = \varphi, \quad t \geq s.$$

Conversely, if  $x \in \mathcal{C}^1([s, \infty), X) \cap \mathcal{C}([s, \infty), \mathcal{C}^2[0, 1])$  is a solution of (5.7) and verifies (5.8)

$$x(t) - u(t)(\beta_{\lambda,t} \otimes L)x(t) \in D(A)$$

for all  $t \geq s$ , then  $z(t, \cdot) := x(t)$  is a solution of (5.1).

*Proof.* Let  $z$  be a solution of (5.1). Then, from the boundary conditions, we deduce that  $x(t) := z(t, \cdot)$  verifies the condition (5.8). By invoking (5.5) together with (5.8), we get

$$\begin{aligned} \dot{x}(t) &= \lambda x(t) - \left( \lambda - \alpha(t) \frac{\partial^2}{\partial \xi^2} \right) x(t) \\ &= \lambda x(t) - (\lambda - \alpha(t)A)(x(t) + u(t)(\beta_{\lambda,t} \otimes L)x(t)) \\ &= \lambda x(t) - (\lambda - A(t))(x(t) + u(t)(\beta_{\lambda,t} \otimes L)x(t)) \\ &= A(t)x(t) + u(t)B_\lambda(t)x(t). \end{aligned}$$

It is straightforward to see that the converse also holds.  $\square$

We know that the evolution family generated by  $\alpha(\cdot)A$  is given by  $\mathbb{T}(t, s) := S(\int_s^t \alpha(\tau) d\tau)$ ,  $(t, s) \in \Delta$ . This family can be extended on the extrapolated space  $X_{-1}$  to the evolution family  $\overline{\mathbb{T}}(t, s) := S_{-1}(\int_s^t \alpha(\tau) d\tau)$  which is generated by  $A(\cdot)$ . Let us define the operator

$$\Phi_{t,s}(u, x) := \int_s^t \overline{\mathbb{T}}(t, \sigma) u(\sigma) B_\lambda(\sigma) x(\sigma) d\sigma$$

for  $(u, x) \in \mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s}$  and  $(t, s) \in \Delta$ . We show that the pair  $(\mathbb{T}, \Phi)$  is a time-varying bilinear control system on  $X$ . We proceed by steps.

*Step 1.* We show that the operator  $\Phi_{t,s} \in \mathcal{BL}(\mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s}, X)$ : For each  $(s, t) \in \Delta$  and  $(u, x) \in \mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s}$  we define the functions

$$\gamma(\sigma) := \int_\sigma^t \alpha(\tau) d\tau \quad \text{and} \quad f(\sigma) := u(\sigma)(\lambda - A(t))L(x(\sigma))\beta_{\lambda,\sigma}, \quad \sigma \in [s, t].$$

The function  $\gamma \in C^1([s, t])$  and is strictly decreasing and hence bijective. Thus, by changing variables, one can write

$$\begin{aligned} \Phi_{t,s}(u, x) &= \int_s^t S_{-1}(\gamma(\sigma)) f(\sigma) d\sigma \\ &= \int_0^{\gamma(s)} S_{-1}(r) \frac{f(\gamma^{-1}(r))}{\alpha(\gamma^{-1}(r))} dr \\ (5.9) \quad &= \int_0^{\gamma(s)} S_{-1}(\gamma(s) - r) g(\gamma(s) - r) dr, \end{aligned}$$

where  $g(\cdot) := \frac{f(\gamma^{-1}(\cdot))}{\alpha(\gamma^{-1}(\cdot))}$ .

We now introduce the Favard space associated with a semigroup  $(V(t))_{t \geq 0}$  (with generator  $G$ ) on some Banach space  $E$  which is defined by

$$F_G := \left\{ x \in E : \sup_{t > 0} \frac{1}{t} \|e^{-\omega t} V(t)x - x\| < \infty \right\},$$

with  $\omega > \omega_0$  fixed ( $\omega_0$  is the growth bound of the semigroup  $(V(t))_{t \geq 0}$ ). The space  $F_G$  endowed with the norm  $\|x\|_{F_G} := \sup_{t > 0} \frac{1}{t} \|e^{-\omega t} V(t)x - x\|$  is a Banach space;

see, e.g., [15, sect. II.5] for further properties. We have to note that in our case  $F_{\alpha(t)A} = F_A$  for all  $t \geq 0$ .

PROPOSITION 5.2. *We have the following properties:*

- (i) *The function  $\beta_{\lambda,t} \in F_A$  for each  $t \geq 0$ .*
- (ii) *The function  $t \mapsto \beta_{\lambda,t}$  is continuous from  $\mathbb{R}_+$  to  $F_A$ .*
- (iii) *The function  $g$  defined in (5.9) is locally integrable with values in  $F_{A_{-1}}$ .*

*Proof.* (i) In fact, a simple computation gives

$$(5.10) \quad \|\beta_{\lambda,t}\|_X \leq \frac{\alpha_m}{\lambda}$$

for all  $\lambda > 0$ , and then by applying [13, Thm. 16] we obtain our claim.

(ii) The estimate (5.10) implies that the space  $X_m := \{f \in W^{1,2}(0,1), f(0) = 0\}$ , endowed with the  $W^{1,2}$ -norm, is continuously embedded in the Favard space  $F_A$ ; see [11, Prop. 4.1]. Hence, via (i) and (5.5), we can write

$$\begin{aligned} \|\beta_{\lambda,t} - \beta_{\lambda,t_0}\|_{F_A} &\leq \text{const } \|\beta_{\lambda,t} - \beta_{\lambda,t_0}\|_{X_m} \\ &\leq \text{const } (1 + \lambda) \|\beta_{\lambda,t} - \beta_{\lambda,t_0}\|_X. \end{aligned}$$

Thus, it is enough to prove the continuity of  $\beta_{\lambda,\cdot}$  with respect to the topology of  $X$ . For this we can simply use (5.6) and the fact that the function  $\cosh$  is locally Lipschitz.

(iii) Remark that  $\lambda_0 - A_{-1} \in \mathcal{L}(F_A, F_{A_{-1}})$  for all  $\lambda_0 \in \rho(A) = \rho(A_{-1})$ ; see, e.g., [15, sect. II.5]. Then the function  $g$  takes values in  $F_{A_{-1}}$ . To obtain our aim we show that the function  $(\lambda - \alpha(\cdot)A_{-1})\beta_{\lambda,\cdot}$  is continuous from  $\mathbb{R}_+$  to  $F_{A_{-1}}$ . In fact, let  $t, t' \geq 0$  and  $\lambda_0 \in \rho(A)$ . Then by the estimate

$$\begin{aligned} \|(\lambda - \alpha(t)A_{-1})\beta_{\lambda,t} - (\lambda - \alpha(t')A_{-1})\beta_{\lambda,t'}\|_{F_{A_{-1}}} &\leq \alpha(t) \left| \frac{\lambda}{\alpha(t)} - \frac{\lambda}{\alpha(t')} \right| \|\beta_{\lambda,t}\|_{F_A} \\ &\quad + \left\{ \left| \frac{\lambda}{\alpha(t')} - \lambda_0 \right| + \|\lambda_0 - A_{-1}\|_{\mathcal{L}(F_A, F_{A_{-1}})} \right\} \|\alpha(t)\beta_{\lambda,t} - \alpha(t')\beta_{\lambda,t'}\|_{F_A} \end{aligned}$$

together with the continuity of the function  $\alpha$  and (ii), we obtain (iii).

To conclude this step, we apply the Nagel–Sinestrari result (see [27, Prop. 3.1]) and use (iii) to obtain that the range of  $\Phi_{t,s}$  is contained in  $X$ . Therefore, by using the fact that  $X (= D(A_{-1}))$  is continuously injected in  $X_{-1}$  and the open mapping theorem for bilinear operators, we can deduce that  $\Phi_{t,s} \in \mathcal{BL}(\mathbb{C}_{loc}^{2,s} \times X_{loc}^{2,s}, X)$ .  $\square$

*Step 2.* We prove the estimate (3.2): By applying again [27, Prop. 3.1] and using (5.9) we obtain

$$(5.11) \quad \|\Phi_{t,s}(u, x)\| \leq N e^{\omega\gamma(s)} \|g\|_{L^1(0, \gamma(s); F_{A_{-1}})},$$

with constants  $\omega, N \geq 0$  independent of  $s$  and  $t$ . Now, by using successively Hölder's

inequality and changing variables we can estimate

$$\begin{aligned}
 \|g\|_{L^1(0,\gamma(s);F_{A_{-1}})} &\leq \left( \int_0^{\gamma(s)} \frac{1}{\alpha(\gamma^{-1}(r))} |u(\gamma^{-1}(r))|^2 dr \right)^{\frac{1}{2}} \\
 &\quad \times \left( \int_0^{\gamma(s)} (|Lx(\gamma^{-1}(r))| \|(\lambda - \alpha(\gamma^{-1}(r))A_{-1})\beta_{\lambda,\gamma^{-1}(r)}\|_{F_{A_{-1}}})^2 dr \right)^{\frac{1}{2}} \\
 &\leq \frac{\alpha_m^2}{\alpha_0} \|\pi\|_{L^\infty} \left( \int_s^t |u(r)|^2 dr \right)^{\frac{1}{2}} \left( \int_s^t \left( \|x(r)\| \left\| \left( \frac{\lambda}{\alpha(r)} - A_{-1} \right) \beta_{\lambda,r} \right\|_{F_{A_{-1}}} \right)^2 dr \right)^{\frac{1}{2}}.
 \end{aligned}
 \tag{5.12}$$

On the other hand, we have the following continuous injection  $X_m \hookrightarrow F_{A_{-1}}$  ( $X_m$  is defined in (ii)). Thus, by using the fact that  $\|\beta_{\lambda,r}\|_{X_m} \leq \text{const } (1 + \lambda)\|\beta_{\lambda,r}\|_X$  and (5.10), we obtain

$$\begin{aligned}
 \left\| \left( \frac{\lambda}{\alpha(r)} - A_{-1} \right) \beta_{\lambda,r} \right\|_{F_{A_{-1}}} &\leq \left\| \left( \frac{\lambda}{\alpha(r)} - \lambda_0 \right) \beta_{\lambda,r} \right\|_{F_{A_{-1}}} + \|(\lambda_0 - A_{-1})\beta_{\lambda,r}\|_{F_{A_{-1}}} \\
 &\leq \text{const } \alpha_m \frac{1 + \lambda}{\lambda} \left( \frac{\lambda}{\alpha_0} + \lambda_0 + 1 \right).
 \end{aligned}
 \tag{5.13}$$

Therefore, the estimate (3.2) can be deduced from (5.11) and (5.13) with the constant

$$\beta(t_0) = \text{const } \left\{ \frac{\alpha_m^3}{\alpha_0} N \|\pi\|_{L^\infty} \frac{1 + \lambda}{\lambda} \left( \frac{\lambda}{\alpha_0} + \lambda_0 + 1 \right) \right\} e^{\omega \alpha_m t_0}.$$

Finally, the property of composition (3.1) is immediate, and then the pair  $(\mathbb{T}, \Phi)$  is a time-varying bilinear control system on  $X$ .

Our observation operator  $C_\theta f := \int_0^1 \xi^\theta f(\xi) d\xi$  is unbounded with maximal domain  $D(C_\theta) = L^r(0, 1)$ ,  $r > \frac{1}{1+\theta}$ . But one can prove by using the change of variables in (5.9) and [12, Lem. 6.5] that  $C_\theta$  is a  $\mathbb{T}$ -admissible observation operator. Remark also that the operator  $C_\theta$  is bounded on  $X_a$ , where  $a \in ]0, \frac{1}{2}[$ . Hence, according to (5.2), the output map  $\Psi$  is given by  $\Psi_s(x) = C_\theta \mathbb{T}(\cdot, s)x$  for  $x \in X$ .

Since the Favard space  $F_A$  is continuously embedded in  $X_a$  when  $a \in [0, 1[$  and due to Proposition 5.2, we have that  $\bar{\mathbb{T}}(t, s)B_\lambda(t) \in \mathcal{L}(X, X_a)$ , which is continuous w.r.t.  $t \geq s$ . On the other hand, by using (5.3) we can easily deduce that

$$\|\bar{\mathbb{T}}(t, s)B_\lambda(t)\|_{\mathcal{L}(X, X_a)} \leq \hbar_1 e^{\omega(t-s)}(t-s)^{(b-a-1)},
 \tag{5.14}$$

where  $a \in ]0, 1[$ ,  $b \in ]0, \frac{3}{4}[$ , and  $\hbar_1 = \hbar_1(L, a, b, \alpha) > 0$ . This proves that  $\Phi_{t,s}(u, x) \in X_a$ , and if  $a \in ]0, \frac{1}{2}[$  and  $b \in ]0, \frac{3}{4}[$ , then we estimate

$$|C_\theta \Phi_{t,s}(u, x)| \leq \hbar_2 \int_s^t e^{\omega(\tau-s)}(\tau-s)^{(b-a-1)} u(\tau) x(\tau) d\tau,$$

where  $\hbar_2 := \hbar_1 \|C_\theta\|_{\mathcal{L}(X_a, \mathbb{C})}$ . This implies that

$$\|C_\theta \Phi_{t,s}(u, x)\|_{L^2(s, s+t_0)}^2 \leq \hbar_2 \int_s^{s+t_0} \left( \int_s^t e^{\omega(\tau-s)}(\tau-s)^{(b-a-1)} u(\tau) x(\tau) d\tau \right)^2 dt.$$



By applying successively Young and Hölder inequalities and assuming that  $b-a > 1/2$ , we have

$$(5.15) \quad \|C_\theta \Phi_{\cdot,s}(u, x)\|_{L^2(s, s+t_0)} \leq \hbar_3 e^{\omega t_0} t_0^{(b-a-\frac{1}{2})} \|u\|_{L^2(s, s+t_0)} \|x\|_{L^2([s, s+t_0], X)},$$

where  $\hbar_3 = \hbar_3(a, b, \hbar_2) > 0$ .

Thus we have proved that, if  $a \in ]0, \frac{1}{2}[$ ,  $b \in ]0, \frac{3}{4}[$ , and  $b-a > 1/2$ , the triple  $(\mathbb{T}, \Phi, \Psi)$  is admissible on  $(L^1(0,1), \mathbb{C}, \mathbb{C})$ , which means that the system  $\Sigma := (\mathbb{T}, \Phi, \Psi, \mathbb{F})$ , where  $\mathbb{F}_s(u, x) := C_\theta \Phi_{\cdot,s}(u, x)$ , is a well-posed time-varying bilinear system on  $(L^1(0,1), \mathbb{C}, \mathbb{C})$ ; see Proposition 3.6. Under the same condition we can also deduce from (5.15) that our system is absolutely regular.

Now it remains to prove that the pair  $(\mathbb{T}, \Psi)$  is detectable. Since the spectral bound  $s(A) = -\pi^2/4$  and  $S(\cdot)$  is an analytic semigroup, then via the spectral mapping theorem the semigroup  $S(\cdot)$  is exponentially stable. Hence the evolution family  $\mathbb{T}$  is so as well. Therefore  $(\mathbb{T}, \Psi)$  is detectable by the operator  $\mathbb{K} = 0$ . However, one can see that this pair is still detectable by nontrivial, but not “too greater,” operators. In fact, let  $k \in X_{loc}^2$ , and set

$$(5.16) \quad \mathbb{K}_{t,s}(y) := \int_s^t \mathbb{T}(t, \sigma) y(\sigma) k(\sigma) d\sigma,$$

where  $y \in \mathbb{C}_{loc}^{2,s}$  and  $t \geq s \geq 0$ .

**PROPOSITION 5.3.** *There is a constant  $k_0 > 0$  such that the pair  $(\mathbb{T}, \Psi)$  is admissibly detectable by  $\mathbb{K}$  for all functions  $k \in X^\infty$ , with  $\|k\|_{X^\infty} < k_0$ .*

*Proof.* The time-varying linear observation system  $(\mathbb{T}, \Psi)$  is represented by  $C_\theta$ ; then by [31, Prop. 2.11],  $\mathbb{K}_{t,s}y \in D(C_\theta)$  for all  $t \geq s$ , and the time-varying linear system  $\Gamma := (\mathbb{T}, \mathbb{K}, \Psi, \mathbb{L}^K)$ , where  $\mathbb{L}_s^K := C_\theta \mathbb{K}_{\cdot,s}$ , is well-posed. The system  $\Gamma$  is also regular since  $k \in X^\infty$ . Similarly, the time-varying linear system  $\Gamma^I := (\mathbb{T}, \mathbb{K}^I, \Psi, \mathbb{L}^I)$ , where  $\mathbb{K}_{t,s}^I f := (\mathbb{V}_s^\mathbb{T} f)(t)$  for  $f \in X_{loc}^{2,s}$ , and  $\mathbb{L}_s^I := C_\theta \mathbb{K}_{\cdot,s}^I$ , is well-posed and regular. On the other hand,  $I$  is an admissible feedback for  $\Gamma$ , since the constant  $\kappa(t_0) = \kappa_0 t_0^{1/2}$  (see (2.6)) tends to 0 when  $t_0$  tends to 0. Similarly, the operators  $k(\cdot) \in L^\infty([0, \infty), \mathcal{L}_s(\mathbb{C}, X))$  are admissible feedback for  $\Gamma^I$ . This means that there exists  $t_0 > 0$  such that the operators  $I - \mathbb{L}_{s+t_0,s}^I k(\cdot)$ ,  $s \geq 0$ , are invertible on  $L^2[s, s+t_0]$  and have uniformly bounded inverses. Therefore, according to [31, Thm. 5.6], there exists a constant  $k_0 > 0$  such that for all  $k \in X^\infty$ , with  $\|k\|_{X^\infty} < k_0$ , the feedback evolution family  $\mathbb{T}^K$ , associated with  $\Gamma^I$  and the feedback  $k$ , is exponentially stable. Finally, it is easy to see that the feedback evolution family  $\mathbb{T}^K$  coincides with that of the feedback bilinear system  $\Sigma^K$ .  $\square$

From this result and Theorem 4.9 we can deduce that, with  $\mathbb{K}$  given in (5.16) and under conditions in Proposition 5.3, the system  $(\mathcal{O})$  is an exponential Luenberger-like observer for  $\Sigma$  with some small inputs.

## REFERENCES

- [1] P. ACQUISTAPASE, *Evolution operators and strong solutions of abstract linear parabolic equations*, Differential Integral Equations, 1 (1988), pp. 433–457.
- [2] G. BORNARD, N. COUENNE, AND F. CELLE, *Regularly persistent observers for bilinear systems*, in New Trends in Nonlinear Control Theory, Lecture Notes in Control and Inform. Sci. 122, Springer-Verlag, Berlin, 1989, manuscript.
- [3] H. BOUNIT, *Topological Properties for Observers Inputs for Regular Time-Varying Bilinear Systems*, in preparation.
- [4] H. BOUNIT AND H. HAMMOURI, *Observers for infinite-dimensional bilinear systems*, Eur. J. Control, 3 (1997), pp. 325–339.

- [5] H. BOUNIT AND H. HAMMOURI, *Observer design for distributed parameter dissipative bilinear systems*, Appl. Math. and Comp. Sci., 8 (1997), pp. 381–402.
- [6] H. BOUNIT AND A. IDRISSE, *Regular bilinear systems*, IMA J. Math. Control Inform., 22 (2005), pp. 26–57.
- [7] C. BRUNI, G. DIPILLO, AND K. KOCH, *Bilinear systems. An appealing class of nearly linear systems in theory and applications*, IEEE Trans. Automat. Control, 19 (1974), pp. 334–348.
- [8] F. CELLE, J. P. GAUTHIER, D. KAZAKOS, AND G. SALLET, *Synthesis of nonlinear observers: A harmonic-analysis approach*, Math. Systems Theory, 22 (1989), pp. 291–322.
- [9] C. CHICONE AND Y. LATUSHKIN, *Evolution Semigroups in Dynamical Systems and Differential Equations*, American Mathematical Society, Providence, RI, 1999.
- [10] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear System Theory*, Lecture Notes in Control and Inform. Sci. 8, Springer-Verlag, Berlin, 1978.
- [11] W. DESCH, J. MILOTA, AND W. SCHAPPACHER, *Least square control problems in nonreflexive spaces*, Semigroup Forum, 62 (2001), pp. 337–357.
- [12] W. DESCH, E. FASANGOVA, J. MILOTA, AND W. SCHAPPACHER, *Riccati operators in non-reflexive spaces*, Differential Integral Equations, 15 (2002), pp. 1493–1510.
- [13] W. DESCH AND W. SCHAPPACHER, *Generation results for perturbed semigroups*, in Semigroup Theory and Applications, Lecture Notes in Pure and Appl. Math., 116, P. Clément, S. Invernizzi, E. Mitidieri, and I. Vrabie, eds., Marcel Dekker, New York, 1989, pp. 127–152.
- [14] J. DIESTEL AND J. J. UHL, *Vector Measures*, Math. Surveys 15, American Mathematical Society, Providence, RI, 1977.
- [15] K. ENGEL AND R. NAGEL, *One-Parameter Semigroups for Linear Evolution Equations*, Springer-Verlag, Berlin, 2000.
- [16] Y. FUNAHASHI, *Stable state estimator for bilinear systems*, Internat. J. Control, 29 (1979), pp. 181–188.
- [17] J. P. GAUTHIER, C. Z. XU, AND A. BOUNABAT, *Observer for infinite dimensional dissipative bilinear systems*, J. Math. Syst. Estim. Contr., 5 (1995), pp. 1–20.
- [18] S. HARA AND K. FURUTA, *Minimal order state observers for bilinear systems*, Internat. J. Control, 24 (1976), pp. 705–718.
- [19] D. HINRICHSSEN AND A. J. PRITCHARD, *Robust stability of linear evolution operators on Banach spaces*, SIAM J. Control Optim., 32 (1994), pp. 1503–1541.
- [20] A. IDRISSE, *On the unboundedness of control operators for bilinear systems*, Quaest. Math., 26 (2003), pp. 105–123.
- [21] A. IDRISSE AND A. RHANDI, *Admissibility of time-varying observations for time-varying systems*, J. Comput. Anal. Appl., 6 (2004), pp. 229–242.
- [22] B. JACOB, *Time-Varying Infinite Dimensional State-Space Systems*, Ph.D. thesis, Bremen, 1995.
- [23] B. JACOB, *Optimal control of time-varying well-posed linear systems on finite time horizon*, in Mathematical Theory of Networks and Systems (Proceedings of the MTNS-98), A. Beghi, I. Finesso, and G. Picci, eds., Il Poligrafo, Padova, 1998, pp. 483–486.
- [24] B. JACOB AND J. R. PARTINGTON, *Admissibility of control and observation operators for semigroups: A survey*, in Current Trends in Operator Theory and its Applications, Proceedings of IWOTA 2002, Operator Theory: Advances and Applications, Vol. 149, J. A. Ball, J. W. Helton, M. Klaus, and L. Rodman, eds., Birkhäuser, Boston, 2004, pp. 199–221.
- [25] B. JACOB, V. DRAGAN, AND A. J. PRITCHARD, *Robust stability of infinite dimensional time-varying systems with respect to non-linear perturbation*, Integral Equations Operator Theory, 22 (1995), pp. 440–462.
- [26] R. R. MOHLER, *Bilinear Control Processes*, Academic Press, New York, 1973.
- [27] R. NAGEL AND E. SINISTRARI, *Inhomogeneous Volterra integrodifferential equations for Hille-Yosida operators*, Lect. Notes Pure Appl. Math., 150 (1994), pp. 51–70.
- [28] R. REBARBER, *Conditions for the equivalence of internal and external stability for distributed parameter systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 994–998.
- [29] D. SALAMON, *Infinite-dimensional linear systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
- [30] R. SCHNAUBELT, *Well-Posedness and Asymptotic Behaviour of Time-Varying Linear Evolution Equations*, preprint Reports of the Institut Analysis, Halle-Wittenberg University, Germany, Report 11, 2001.
- [31] R. SCHNAUBELT, *Feedbacks for nonautonomous regular linear systems*, SIAM J. Control Optim., 41 (2002), pp. 1141–1165.
- [32] E. SONTAG, *On the observability of polynomial systems, I: Finite-time problems*, SIAM J. Control Optim., 17 (1979), pp. 139–151.
- [33] O. J. STAFFANS AND G. WEISS, *Transfer functions of regular linear systems. Part III: Inversions and duality*, Integral Equations Operator Theory, 49 (2004), pp. 517–558.

- [34] H. SUSSMANN, *Single input observability of continuous time systems*, Math. Systems Theory, 12 (1979), pp. 371–393.
- [35] G. WEISS, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.
- [36] G. WEISS, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.
- [37] G. WEISS, *Transfer functions of regular linear systems. Part I: Characterization of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

## OPTIMAL TRANSPORTATION PROBLEM BY STOCHASTIC OPTIMAL CONTROL\*

TOSHIO MIKAMI<sup>†</sup> AND MICHÈLE THIEULLEN<sup>‡</sup>

**Abstract.** We address an optimal mass transportation problem by means of optimal stochastic control. We consider a stochastic control problem which is a natural extension of the Monge–Kantorovich problem. Using a vanishing viscosity argument we provide a probabilistic proof of two fundamental results in mass transportation: the Kantorovich duality and the graph property for the support of an optimal measure for the Monge–Kantorovich problem. Our key tool is a stochastic duality result involving solutions of the Hamilton–Jacobi–Bellman PDE.

**Key words.** optimal mass transportation theory, Monge–Kantorovich problem, Monge problem, duality, stochastic control, Hamilton–Jacobi–Bellman PDE, value function, vanishing viscosity, semiconvex functions

**AMS subject classifications.** 60J25, 60J60, 60G99, 93E20, 49J20, 70H20

**DOI.** 10.1137/050631264

**1. Introduction.** Our goal in the present paper is to show that stochastic optimal control theory can be used efficiently to study deterministic optimal mass transportation problems. Let us recall that optimal transportation theory consists of the following two minimization problems, where  $P_0$  and  $P_1$  are given Borel probability measures on  $\mathbf{R}^d$  and the cost function  $c : \mathbf{R}^d \times \mathbf{R}^d \rightarrow \mathbf{R}^+ \cup \{+\infty\}$  is measurable. In this paper the cost function has the form

$$(1.1) \quad c(x, y) = L(y - x)$$

with  $L(u) : \mathbf{R}^d \rightarrow [0, \infty)$  convex in  $u$ . In the Monge problem the object of study is

$$(1.2) \quad T_M(P_0, P_1) := \inf \left\{ \int_{\mathbf{R}^d} L(g(x) - x) P_0(dx) \right\},$$

and the infimum is taken over all measurable maps  $g : \mathbf{R}^d \mapsto \mathbf{R}^d$  such that the image of  $P_0$  by  $g$  is  $P_1$ . In the Monge–Kantorovich problem (MKP), one considers

$$(1.3) \quad T_{MK}(P_0, P_1) := \inf \left\{ \int_{\mathbf{R}^d \times \mathbf{R}^d} L(y - x) \mu(dxdy) \right\}$$

on the set of probability measures  $\mu$  on  $\mathbf{R}^d \times \mathbf{R}^d$  with marginals  $P_0$  and  $P_1$  (namely, such that  $\mu(A \times \mathbf{R}^d) = P_0(A)$  and  $\mu(\mathbf{R}^d \times B) = P_1(B)$ ). The resolution of (1.2) is a difficult problem. Kantorovich introduced the relaxed version (1.3) as a step to solve (1.2). It is easy to check that

$$(1.4) \quad T_{MK}(P_0, P_1) \leq T_M(P_0, P_1).$$

---

\*Received by the editors May 11, 2005; accepted for publication (in revised form) November 27, 2007; published electronically March 19, 2008.

<http://www.siam.org/journals/sicon/47-3/63126.html>

<sup>†</sup>Department of Mathematics, Hokkaido University, Sapporo 060-0810, Japan (mikami@math.sci.hokudai.ac.jp). This author's research was partially supported by the Grant-in-Aid for Scientific research 15340047, 15340051, 16654031, JSPS.

<sup>‡</sup>Corresponding author. Laboratoire de Probabilités et Modèles Aléatoires, Boite 188, 4, Place Jussieu, Université Paris VI, 75252 Paris cedex 05, France (mth@ccr.jussieu.fr). This author's research was supported by the Grant-in-Aid for Scientific research 15340051, 16654031, JSPS.

Indeed, any measurable mapping  $g : \mathbf{R}^d \mapsto \mathbf{R}^d$  such that the image measure of  $P_0$  by  $g$  is  $P_1$  satisfies

$$(1.5) \quad \int_{\mathbf{R}^d} L(g(x) - x) P_0(dx) = \int_{\mathbf{R}^d \times \mathbf{R}^d} L(y - x) \mu_g(dxdy),$$

where  $\mu_g$  is the image measure of  $P_0$  by the mapping

$$(1.6) \quad \mathbf{R}^d \mapsto \mathbf{R}^d \times \mathbf{R}^d,$$

$$(1.7) \quad x \mapsto (x, g(x)).$$

Inequality (1.4) follows since  $\mu_g$  is a probability measure on  $\mathbf{R}^d \times \mathbf{R}^d$  with marginals  $P_0$  and  $P_1$ . Moreover, an optimal measure for (1.3) always exists (cf. [14]). If any such measure is supported by the graph of a measurable map, we say that the *graph property* holds; that is, if for any  $\mu^*$  optimal for (1.3), there exists a set  $\Gamma$  satisfying  $\mu^*(\Gamma) = 1$  and

$$(1.8) \quad \Gamma = \{(x, \theta(x)); x \in \mathbf{R}^d\}$$

for some measurable mapping  $\theta$ . If the graph property holds, it provides a solution to Monge problem (1.2). Indeed, in this case  $T_{MK}(P_0, P_1) = \int_{\mathbf{R}^d \times \mathbf{R}^d} L(y - x) \mu^*(dxdy) = \int_{\mathbf{R}^d} L(\theta(x) - x) P_0(dx)$ . Using (1.4), we see that the mapping  $\theta$  minimizes Monge problem (1.2). In order to check whether the graph property is satisfied, *Kantorovich duality* for (1.3) plays a fundamental role. It was first proved by Kantorovich (cf. [7]) when the cost function is a distance and later generalized by Kellerer (cf. [8]). It runs as follows:

$$(1.9) \quad T_{MK}(P_0, P_1) = \sup \left\{ \int_{\mathbf{R}^d} \psi(y) P_1(dy) - \int_{\mathbf{R}^d} \varphi(x) P_0(dx) \right\},$$

where the supremum is taken over all pairs  $(\varphi, \psi) \in L^1(P_0) \times L^1(P_1)$  satisfying  $\psi(y) - \varphi(x) \leq L(y - x)$ . To go from Kantorovich duality to the graph property, two types of arguments have been used: differentiability properties of convex functions for the quadratic cost (cf. [1]) and geometrical properties of cyclically monotone sets for general costs (cf. [6]).

In the present paper we show that Kantorovich duality and the graph property can be proved by stochastic optimal control combined with a vanishing viscosity argument. It is not clear a priori that stochastic optimal control theory is well suited to studying problems such as  $T_{MK}$  or  $T_M$ , where the initial and the final distributions are both imposed. However, for the case when the cost is  $L(u) = |u|^2$ , one of us (cf. [10]) addressed (1.2) directly without using (1.3) and gave a probabilistic proof of existence and uniqueness of a solution to (1.2). The proof in [10] relies on h-path processes and cyclically monotone sets. In the present paper, on the contrary, we focus on (1.3) and on duality arguments. We rely on a stochastic duality result which we proved in [11]. The basis of this result is the correspondence between solutions of the Hamilton–Jacobi–Bellman (HJB) partial differential equation (PDE) and value functions of stochastic control. We do not use cyclically monotone sets. By *vanishing viscosity* we prove Kantorovich duality and we recover the graph property. Thus the present paper together with [11] provides a global treatment of these two building blocks of optimal transportation theory by stochastic optimal control. Let us mention that here  $L$  is more general than  $|u|^2$  and our method greatly simplifies the arguments

of [10]. Classically (cf. [14]) the graph property is proved for  $L$  satisfying a cone-type condition which is easy to check only for radial  $L$ . We prove the graph property without any additional cone condition when  $L$  is not necessarily radial such that  $L(u) \sim |u|^2$  at infinity.

In our approach by vanishing viscosity there are still open questions left. One question is the convergence of the optimal process for the stochastic control problem (see section 2 below) to the optimal trajectory for (1.2). It is known that when  $L(u) = |u|^2$ , each optimal process is an h-path process, which is a rather explicit property. Using this information, it was proved in [10] that these optimal h-path processes converge to the deterministic optimal trajectory of (1.2) when their diffusion part tends to zero. To prove an analogous convergence when  $L(u) \sim |u|^2$  at infinity, we may be willing to use the following result obtained in [11] which can be compared to the h-path process property: when  $L(u) \sim |u|^2$  at infinity, the optimal process of the stochastic control problem solves a forward-backward system (cf. [11] Theorem 2.2).

The paper is organized as follows. In section 2 we review the stochastic duality theorem that we have proved in [11]. Sections 3 and 4 present two applications of this stochastic duality combined with a vanishing viscosity argument: Kantorovich duality in section 3 as well as the graph property in section 4 are proved using this method.

**2. A stochastic duality result.** We will be working under the following assumptions:  $L(u) : \mathbf{R}^d \rightarrow [0, \infty)$  is convex in  $u$ ,

(A.1) for some  $\delta > 1$ ,

$$\liminf_{|u| \rightarrow \infty} \frac{L(u)}{|u|^\delta} > 0.$$

(A.2) (i)  $L \in C^3(\mathbf{R}^d)$ ,

(ii)  $D_u^2 L(u)$  is positive definite for all  $u \in \mathbf{R}^d$ .

We denote by  $H$  the Legendre transform of  $L$ :

$$(2.1) \quad H(z) := \sup_{u \in \mathbf{R}^d} \{ \langle z, u \rangle - L(u) \},$$

for  $z \in \mathbf{R}^d$ ;  $\nabla := (\partial/\partial x_i)_{i=1}^d$  and  $\langle \cdot, \cdot \rangle$  denotes the inner product in  $\mathbf{R}^d$ .

**2.1. The stochastic control problem.** We consider the following stochastic optimization problem. For  $\epsilon > 0$ , let

$$(2.2) \quad V_\epsilon(P_0, P_1) := \inf \left\{ E \left[ \int_0^1 L(\beta_X(t, X)) dt \right] \mid \forall X \in \mathcal{A}^\epsilon \text{ such that } PX_0^{-1} = P_0, PX_1^{-1} = P_1 \right\},$$

where  $\mathcal{A}^\epsilon$  is the set of all  $\mathbf{R}^d$ -valued, continuous semimartingales  $\{X(t)\}_{0 \leq t \leq 1}$  on a probability space  $(\Omega, \mathbf{B}, P)$  such that there exists a Borel measurable  $\beta_X : [0, 1] \times C([0, 1]) \mapsto \mathbf{R}^d$  for which

(i)  $\omega \mapsto \beta_X(t, X(\omega))$  is  $\mathcal{B}_t(C)_+$ -measurable for all  $t \in [0, 1]$ , where  $\mathcal{B}_t(C)$  denotes the Borel  $\sigma$ -field of  $C([0, t])$ ;

(ii)  $\{X(t) - X(0) - \int_0^t \beta_X(s, X) ds := \sqrt{\epsilon} W_X(t)\}_{0 \leq t \leq 1}$ , where  $W_X$  is a  $\sigma[X(s) : 0 \leq s \leq t]$ -Brownian motion.

Results about existence and uniqueness of a minimizer for  $V_\epsilon$  are gathered in the following statement.

**THEOREM 2.1.** *Let  $\epsilon > 0$ . Let us assume that  $V_\epsilon(P_0, P_1) < +\infty$  and that assumptions (A.1) and (A.2) hold. Then*

- (i)  $V_\epsilon(P_0, P_1)$  admits a minimizer.
- (ii) if assumption (A.1) holds with  $\delta = 2$ ,  $V_\epsilon(P_0, P_1)$  admits a Markovian minimizer.
- (iii) if  $L$  is strictly convex and assumption (A.1) holds with  $\delta = 2$ , then  $V_\epsilon(P_0, P_1)$  admits a unique minimizer (which is Markovian from (ii)).

Actually statements (ii) and (iii) will be of no use in the present paper. They were important in [11] in order to characterize the minimizer of (2.2) as the solution of a forward-backward system which consists of the coupling of a usual stochastic differential equation (SDE) with a backward one (we refer the reader to [2] for a study of such systems).

**2.2. Stochastic duality.** We now recall (Theorem 2.3 below) the stochastic duality result we obtained in [11]. In order to set the framework, we first quote a fundamental result of optimal stochastic control theory.

In the same way as  $\mathcal{A}^\epsilon$ , we define the set of semimartingales  $\mathcal{A}_t^\epsilon$  in  $C([t, 1])$  and we notice that (A.2)(ii) implies the strict convexity of  $u \mapsto L(u)$ . Moreover, the HJB equation with diffusion coefficient (or viscosity)  $\epsilon$  is the following PDE with given terminal value  $\varphi(1, \cdot) = f(\cdot)$ :

$$(2.3) \quad \frac{\partial \varphi(t, x)}{\partial t} + \frac{\epsilon}{2} \Delta \varphi(t, x) + H(\nabla \varphi(t, x)) = 0 \quad ((t, x) \in (0, 1) \times \mathbf{R}^d),$$

where  $\Delta := \sum_{i=1}^d \partial^2 / \partial x_i^2$  and  $\nabla := (\frac{\partial}{\partial x_i}; 1 \leq i \leq d)$ .

**THEOREM 2.2** (cf. [5]). *Suppose that (A.1) and (A.2) hold. Then for any  $f \in C_b^\infty(\mathbf{R}^d)$ , the HJB equation (2.3) with  $\varphi(1, \cdot) = f$  has a unique solution  $\varphi \in C^{1,2}([0, 1] \times \mathbf{R}^d) \cap C_b^{0,1}([0, 1] \times \mathbf{R}^d)$ , which can be written as follows (as a value function):*

$$(2.4) \quad \varphi(t, x) = \sup_{X \in \mathcal{A}_t^\epsilon} \left\{ E[f(X(1)) | X(t) = x] - E \left[ \int_t^1 L(\beta_X(s, X)) ds \middle| X(t) = x \right] \right\},$$

and for the minimizer  $X \in \mathcal{A}_t^\epsilon$ , the following holds:

$$\beta_X(s, X) = D_x H(\nabla \varphi(s, X(s))).$$

In other words, this theorem establishes a one-to-one correspondence between classical solutions of (2.3) and value functions of stochastic control problems with smooth terminal cost. Actually it is a duality result since the supremum in (2.4) involves  $L$ , while (2.3) involves its Legendre transform  $H$ .

In [11] we proved the following duality theorem for the minimization problem (2.2).

**THEOREM 2.3** (stochastic duality). *Let  $\epsilon > 0$  be fixed and  $V_\epsilon(P_0, P_1)$  be as defined in (2.2). Let us assume that (A.1), (A.2) are satisfied and*

$$(2.5) \quad V_\epsilon(P_0, P_1) < +\infty.$$

Then, the following identity holds:

$$(2.6) \quad V_\epsilon(P_0, P_1) = \nu_\epsilon(P_0, P_1)$$

with  $\nu_\epsilon(P_0, P_1)$  defined by

$$(2.7) \quad \nu_\epsilon(P_0, P_1) := \sup \left\{ \int_{\mathbf{R}^d} \varphi(1, y) P_1(dy) - \int_{\mathbf{R}^d} \varphi(0, x) P_0(dx) \right\},$$

where the supremum is taken over all classical solutions  $\varphi$ , to HJB equation (2.3) for which  $\varphi(1, \cdot) \in C_b^\infty(\mathbf{R}^d)$ .

*Remark 2.1.* Actually a stronger version of Theorem 2.3 is proved in [11]: (2.6) holds true without assuming that  $V_\epsilon(P_0, P_1) < +\infty$ ; moreover, this identity still holds when the supremum in  $\nu_\epsilon$  is taken over all bounded, uniformly Lipschitz continuous viscosity solutions  $\varphi$  of (2.3).

Before proceeding further we check that the right-hand side of (2.6) is finite when  $V_\epsilon(P_0, P_1) < +\infty$  and give an outline of the proof of this theorem. For the detailed proof we refer the reader to [11].

Let us first notice that Theorem 2.2 recalled above ensures, in particular, that given  $f \in C_b^\infty(\mathbf{R}^d)$ , a classical solution of HJB PDE (2.3) exists with  $f$  as a terminal value and belongs to  $C^{1,2}([0, 1] \times \mathbf{R}^d) \cap C_b^{0,1}([0, 1] \times \mathbf{R}^d)$ . For more details we refer the reader to [5, p. 206, Theorem 11.1 and p. 210, Remark 11.2]. Therefore the set on which the supremum in (2.7) is taken is not empty.

Let us now assume that  $V_\epsilon(P_0, P_1) < +\infty$ . For all  $X \in \mathcal{A}^\epsilon$  and  $\varphi$  solution of (2.3) satisfying  $\varphi(1, \cdot) \in C_b^\infty(\mathbf{R}^d)$ , the constraints on the marginals of  $X$  combined with the Ito formula imply the following identities:

$$(2.8) \quad \int_{\mathbf{R}^d} \varphi(1, y) P_1(dy) - \int_{\mathbf{R}^d} \varphi(0, x) P_0(dx) = E(\varphi(1, X_1) - \varphi(0, X_0))$$

$$(2.9) \quad = E \int_0^1 \left( \frac{\partial \varphi(s, X_s)}{\partial t} + \frac{\epsilon}{2} \Delta \varphi(s, X_s) + \beta_X(s, X) \nabla \varphi(s, X_s) \right) ds.$$

Since  $\varphi$  solves (2.3) and  $H$  is the Legendre transform of  $L$ , it follows that

$$(2.10) \quad \int_{\mathbf{R}^d} \varphi(1, y) P_1(dy) - \int_{\mathbf{R}^d} \varphi(0, x) P_0(dx) \leq E \int_0^1 L(\beta_X(s, X)) ds.$$

Hence

$$(2.11) \quad \int_{\mathbf{R}^d} \varphi(1, y) P_1(dy) - \int_{\mathbf{R}^d} \varphi(0, x) P_0(dx) \leq V_\epsilon(P_0, P_1)$$

since  $\varphi$  and  $X$  have been chosen independently.

The scheme of the proof of Theorem 2.3 proceeds as follows. We show first that the function  $Q \mapsto V_\epsilon(P_0, Q)$  is lower semicontinuous and convex on the set of probability measures on  $\mathbf{R}^d$ . Therefore it coincides with its double dual, in particular at point  $P_1$ ; namely,

$$(2.12) \quad V_\epsilon(P_0, P_1) = \sup_{f \in C_b(\mathbf{R}^d)} \left\{ \int_{\mathbf{R}^d} f(x) P_1(dx) - V_\epsilon(P_0, \cdot)^*(f) \right\},$$



where for  $f \in C_b(\mathbf{R}^d)$ ,

$$(2.13) \quad V_\epsilon(P_0, \cdot)^*(f) := \sup_{Q \in \mathcal{M}_1(\mathbf{R}^d)} \left\{ \int_{\mathbf{R}^d} f(x) Q(dx) - V_\epsilon(P_0, Q) \right\}.$$

In this identity,  $Q$  plays the role of a terminal law and is arbitrary. Hence we are back in the framework of classical stochastic control; in particular, we can use Theorem 2.2. For all of the details, see [11].

**3. Kantorovich duality by vanishing viscosity.** We start with a precise statement of Kantorovich duality mentioned in the introduction. For the sake of coherence we keep a cost function of the form  $c(x, y) = L(y - x)$ , but this result holds true for general costs (cf. [12] or [14]).

**THEOREM 3.1** (cf. [7], [8]). *Let  $L : \mathbf{R}^d \rightarrow \mathbf{R}^+ \cup \{+\infty\}$  be a lower semicontinuous function. Let  $P_0$  and  $P_1$  be given Borel probability measures on  $\mathbf{R}^d$ . Let us keep the notation  $T_{MK}(P_0, P_1)$  for the MKP as in (1.3) and define*

$$(3.1) \quad \mathcal{T}(P_0, P_1) = \sup \left\{ \int_{\mathbf{R}^d} \psi(y) P_1(dy) - \int_{\mathbf{R}^d} \varphi(x) P_0(dx) \right\},$$

where the supremum is taken over all pairs  $(\varphi, \psi) \in L^1(P_0) \times L^1(P_1)$  satisfying  $\psi(y) - \varphi(x) \leq L(y - x)$ . Then

$$(3.2) \quad T_{MK}(P_0, P_1) = \mathcal{T}(P_0, P_1).$$

We now apply our stochastic duality (Theorem 2.3) in order to prove *Kantorovich duality* with the help of a vanishing viscosity argument. The first part of the following statement is our key tool to go from  $\epsilon > 0$  to  $\epsilon = 0$ . Remember that  $\mathcal{T}(P_0, P_1)$  is defined by (3.1).

**THEOREM 3.2.** *Let us assume that  $T_{MK}(P_0, P_1) < +\infty$  and that assumptions (A.1)–(A.2) hold. Let us recall that  $V_\epsilon$  (resp.,  $\nu_\epsilon$ ) has been defined in (2.2) (resp., (2.7)). We denote by  $g_\epsilon \star P_1$  the convolution of  $P_1$  with the Gaussian kernel  $g_\epsilon(x) = (2\pi\epsilon)^{-\frac{d}{2}} \exp(-\frac{|x|^2}{2\epsilon})$ . Then*

(1.i) *for all  $\epsilon > 0$ ,*

$$(3.3) \quad \nu_\epsilon(P_0, g_\epsilon \star P_1) \leq \mathcal{T}(P_0, P_1).$$

(1.ii) *Moreover,*

$$(3.4) \quad T_{MK}(P_0, P_1) \leq \liminf_{\epsilon \rightarrow 0} V_\epsilon(P_0, g_\epsilon \star P_1).$$

(2) *As a consequence we recover the Kantorovich duality,*

$$(3.5) \quad T_{MK}(P_0, P_1) = \mathcal{T}(P_0, P_1).$$

*Proof of Theorem 3.2.* The only thing to prove is  $T_{MK}(P_0, P_1) \leq \mathcal{T}(P_0, P_1)$ . Indeed, the converse inequality is easy, as we now check. Let  $(u, v) \in L^1(P_0) \times L^1(P_1)$  be such that  $v(y) - u(x) \leq L(y - x)$  and  $\mu$  with marginals  $P_0$  and  $P_1$ . Then

$$\begin{aligned} \int_{\mathbf{R}^d} v(y) P_1(dy) - \int_{\mathbf{R}^d} u(x) P_0(dx) &= \int_{\mathbf{R}^d \times \mathbf{R}^d} (v(y) - u(x)) \mu(dxdy) \\ &\leq \int_{\mathbf{R}^d \times \mathbf{R}^d} L(y - x) \mu(dxdy), \end{aligned}$$

which yields

$$(3.6) \quad \mathcal{T}(P_0, P_1) \leq T_{MK}(P_0, P_1).$$

Let us now prove the converse.

(1.i) Take  $\epsilon > 0$  and let  $\varphi(t, x)$  denote a solution to the HJB PDE (2.3) with  $\varphi(1, \cdot) \in C_b^\infty(\mathbf{R}^d)$ , which implies that  $\varphi \in C^{1,2}([0, 1] \times \mathbf{R}^d) \cap C_b^{0,1}([0, 1] \times \mathbf{R}^d)$ . Let us define

$$(3.7) \quad u_\epsilon(x) := \varphi(0, x),$$

$$(3.8) \quad v_\epsilon(y) := E(\varphi(1, y + \sqrt{\epsilon}W_1)).$$

The pair  $(u_\epsilon, v_\epsilon)$  belongs to  $L^1(P_0) \times L^1(P_1)$  and satisfies

$$\begin{aligned} & \int_{\mathbf{R}^d} \varphi(1, y) g_\epsilon \star P_1(dy) - \int_{\mathbf{R}^d} \varphi(0, x) P_0(dx) \\ &= \int_{\mathbf{R}^d} v_\epsilon(y) P_1(dy) - \int_{\mathbf{R}^d} u_\epsilon(x) P_0(dx). \end{aligned}$$

Moreover, by definition

$$(3.9) \quad v_\epsilon(y) - u_\epsilon(x) = E(\varphi(1, X_1^{x,y}) - \varphi(0, X_0^{x,y})),$$

where  $X_t^{x,y} := x + t(y - x) + \sqrt{t}W_t$ . Using the Ito formula and the fact that  $\varphi$  solves (2.3), we obtain

$$(3.10) \quad E(\varphi(1, X_1^{x,y}) - \varphi(0, X_0^{x,y})) = E \int_0^1 (\langle y - x, \nabla \varphi \rangle - H(\nabla \varphi))(s, X_s^{x,y}) ds,$$

which implies  $v_\epsilon(y) - u_\epsilon(x) \leq L(y - x)$ . Inequality (3.3) follows.

(1.ii) Let us first notice that

$$(3.11) \quad 0 \leq \liminf_{\epsilon \rightarrow 0} V_\epsilon(P_0, g_\epsilon \star P_1)$$

since by definition  $V_\epsilon$  is positive. Moreover,

$$(3.12) \quad \liminf_{\epsilon \rightarrow 0} V_\epsilon(P_0, g_\epsilon \star P_1) < +\infty.$$

Indeed, (3.6) and (1.i) imply  $V_\epsilon(P_0, g_\epsilon \star P_1) \leq T_{MK}(P_0, P_1) < +\infty$ , and stochastic duality applied to the pair  $(P_0, g_\epsilon \star P_1)$  yields (3.12). Let us now consider a sequence  $(\epsilon_n)$  which converges to 0 such that  $V_{\epsilon_n}(P_0, g_{\epsilon_n} \star P_1)$  converges to  $\liminf_{\epsilon \rightarrow 0} V_\epsilon(P_0, g_\epsilon \star P_1)$ . Let us denote by  $X^n$  a minimizer of  $V_{\epsilon_n}(P_0, g_{\epsilon_n} \star P_1)$  (cf. Theorem 2.1). For each  $n$ ,  $X^n \in \mathcal{A}^{\epsilon_n}$ . In particular, with the notation of (2.2),

$$(3.13) \quad \lim_{n \rightarrow +\infty} E \int_0^1 L(\beta_{X^n}(s, X^n)) ds = \liminf_{\epsilon \rightarrow 0} V_\epsilon(P_0, g_\epsilon \star P_1).$$

The superlinearity of  $L$  (namely,  $L(u) \geq |u|^\delta$  with  $\delta > 1$ ) ensures that the sequence of semimartingales  $(X^n)$  is tight and any converging subsequence converges to an absolutely continuous process (cf. [15]). Let  $X_t = X_0 + \int_0^t b_X(s) ds$  be the limit of a converging subsequence. From the convexity property of  $L$ ,

$$(3.14) \quad E \int_0^1 L(b_X(s)) ds \geq E(L(X_1 - X_0)).$$

The law of  $X_1$  is equal to  $P_1$  since it is the limit in distribution of a subsequence of  $g_\epsilon \star P_1$ . Using Fatou's lemma we obtain (3.4).

(2) By combining inequalities (3.3), (3.4), and (3.6) with stochastic duality applied to the pair  $(P_0, g_\epsilon \star P_1)$ , we recover *Kantorovich duality*.  $\square$

#### 4. Graph property by vanishing viscosity.

**4.1. Graph property.** We sketch the argument briefly, for the quadratic cost. For a complete exposition we refer the reader to [14].

So, let us assume for a short while that  $L(y - x) = \frac{1}{2}|y - x|^2$ . In this case,

$$(4.1) \quad T_{MK}(P_0, P_1) := \inf \left\{ \int_{\mathbf{R}^d \times \mathbf{R}^d} \frac{1}{2} |y - x|^2 \mu(dx dy) \right\}$$

on the set of probability measures  $\mu$  on  $\mathbf{R}^d \times \mathbf{R}^d$  with marginals  $P_0$  and  $P_1$  satisfying  $\int_{\mathbf{R}^d} |y|^2 P_1(dy) < +\infty$  (resp.,  $\int_{\mathbf{R}^d} |x|^2 P_0(dx) < +\infty$ ). *Kantorovich duality* (3.2) takes the form

$$(4.2) \quad T_{MK}(P_0, P_1) = \sup \left\{ \int_{\mathbf{R}^d} \psi(y) P_1(dy) - \int_{\mathbf{R}^d} \varphi(x) P_0(dx) \right\},$$

where the supremum is taken over all pairs  $(\varphi, \psi) \in L^1(P_0) \times L^1(P_1)$  satisfying  $\psi(y) - \varphi(x) \leq \frac{1}{2}|y - x|^2$ . Using the identity  $\int_{\mathbf{R}^d \times \mathbf{R}^d} \frac{1}{2} |y|^2 \mu(dx dy) = \int_{\mathbf{R}^d} \frac{1}{2} |y|^2 P_1(dy)$  (resp.,  $\int_{\mathbf{R}^d \times \mathbf{R}^d} \frac{1}{2} |x|^2 \mu(dx dy) = \int_{\mathbf{R}^d} \frac{1}{2} |x|^2 P_0(dx)$ ), and setting  $u(x) := \varphi(x) + \frac{1}{2}|x|^2$  (resp.,  $v(y) := \frac{1}{2}|y|^2 - \psi(y)$ ), identity (4.2) can be rewritten as

$$(4.3) \quad \sup \int_{\mathbf{R}^d \times \mathbf{R}^d} \langle x, y \rangle \mu(dx dy) = \inf \int_{\mathbf{R}^d} v(y) P_1(dy) + \int_{\mathbf{R}^d} u(x) P_0(dx),$$

where the supremum on the left-hand side is taken over all probabilities with marginals  $P_0$  and  $P_1$  with finite second order moments, and the infimum on the right-hand side is over pairs  $(u, v) \in L^1(P_0) \times L^1(P_1)$  satisfying

$$(4.4) \quad \langle x, y \rangle \leq u(x) + v(y) \quad \forall (x, y) \in \mathbf{R}^d \times \mathbf{R}^d.$$

This simple remark has an important consequence (cf. [14]): On the right-hand side of (4.3) it is sufficient to consider pairs  $(u, v)$  such that  $u$  is convex and  $v$  is the Legendre transform of  $u$ . Now let  $\mu^*$  be optimal for (1.3) and  $(u^*, v^*)$  be optimal for the right-hand side of (4.3) (cf. [14] for the respective existence of these optima). Then

$$(4.5) \quad \langle x, y \rangle = u^*(x) + v^*(y) \quad \text{for } \mu^*\text{-a.a. } (x, y).$$

Let us assume, moreover, that  $P_0$  is absolutely continuous w.r.t. Lebesgue measure. Then, differentiability properties of convex functions imply that  $u^*$  is differentiable  $P_0$ -a.s. on  $\mathbf{R}^d$ . Let us consider  $x_0 \in \mathbf{R}^d$  such that  $(x_0, y_0) \in \text{Supp} \mu^*$  for some  $y_0$ . Let  $u^*$  be differentiable at  $x_0$ . Comparison of identity (4.4), written for  $(u^*, v^*)$  valid for all  $(x, y_0), x \in \mathbf{R}^d$  on the one hand and (4.5) at  $(x_0, y_0)$  on the other hand, implies that  $y_0 = \nabla u^*(x_0)$ . We conclude that  $\mu^*$  is indeed supported on a graph, which is the graph of  $\nabla u^*$ : the *graph property* holds.

We now come back to general costs and recall precise statements on the *graph property* obtained, respectively, by Brenier and Benamou (quadratic cost) and Gangbo and McCann (more general costs). The proof of Gangbo and McCann relies on cyclically monotone sets. These authors assume that  $L$  has the following property: For  $(p, \theta, r) \in \mathbf{R}^d \times ]0, \pi[ \times ]0, +\infty[$ , when the norm of  $p$  is large enough, there exists a  $z \in \mathbf{R}^d$  such that the restriction of  $L$  to the set

$$(4.6) \quad K(p, z, \theta, r) := \left\{ x \in \mathbf{R}^d; \quad |x - p| |z| \cos \left( \frac{\theta}{2} \right) \leq \langle z, x - p \rangle \leq r |z| \right\}$$

attains its maximum at  $p$ . Let us notice that  $K(p, z, \theta, r)$  is a truncated cone (with vertex  $p$ , angle  $\frac{1}{2}\theta$ , direction  $z$ ); for this reason we call this condition the cone condition. A drawback of the cone condition is that it can be easily checked only for radial functions  $L$ .

**THEOREM 4.1** (cf. [1], [6]). *Let us assume that  $T_{MK}(P_0, P_1) < +\infty$  and  $P_0$  is absolutely continuous w.r.t. the Lebesgue measure on  $\mathbf{R}^d$ .*

(1) *Let  $L(u) = \frac{1}{2}|u|^2$ . There exists a unique  $\mu^*$  minimizing (1.3). The support of  $\mu^*$  is the graph of  $\nabla u$ , where  $u$  is convex.*

(2) *Let  $L$  be superlinear ( $\lim_{|u| \rightarrow \infty} \frac{L(u)}{|u|} = +\infty$ ) and strictly convex, satisfying the cone condition. Let  $H$  denote the Legendre transform of  $L$ . Then there exists a unique  $\mu^*$  minimizing (1.3). There exists  $\phi$ ,  $L$ -concave, such that the support of  $\mu^*$  is the graph of the mapping*

$$(4.7) \quad g(x) = x + \nabla H(-\nabla \phi(x)).$$

Let us recall that a function  $\gamma : \mathbf{R}^d \rightarrow \mathbf{R} \cup \{-\infty\}$  is  $L$ -concave if there exists  $\beta : \mathbf{R}^d \rightarrow \mathbf{R} \cup \{-\infty\}$  with  $\beta \not\equiv -\infty$  such that

$$(4.8) \quad \forall x \in \mathbf{R}^d \quad \gamma(x) = \inf_{y \in \mathbf{R}^d} (L(y - x) - \beta(y)).$$

**4.2. Vanishing viscosity method.** In this section we apply stochastic duality to recover the fact that an optimal measure for (1.3) is supported on a graph when  $P_0$  is absolutely continuous w.r.t. the Lebesgue measure on  $\mathbf{R}^d$ . More precisely, stochastic duality will enable us to reach a weak form of the situation just described for the quadratic cost, where we had at the same time identities (4.4) and (4.5). However, this weak form will turn out to be sufficient to conclude. In what follows we denote by  $\mu^*$  an optimal measure for (1.3). The following statement exhibits a set  $S$  which supports  $\mu^*$ .

**THEOREM 4.2.** *Let us assume that (A.1)–(A.2) hold true. Let  $\mu^*$  be optimal for the MKP (1.3). There exists a sequence  $(\epsilon_n, \varphi_{\epsilon_n})$  such that  $\epsilon_n \rightarrow 0$ ,*

$$(4.9) \quad \frac{\partial \varphi_{\epsilon_n}(t, x)}{\partial t} + \frac{\epsilon_n}{2} \Delta \varphi_{\epsilon_n}(t, x) + H(\nabla \varphi_{\epsilon_n}(t, x)) = 0,$$

as well as  $\varphi_{\epsilon_n}(1, \cdot) \in C_b^\infty(\mathbf{R}^d)$  for all  $n$ , and  $\mu^*(S) = 1$ , where

$$(4.10) \quad S := \left\{ (x, y); \lim_{n \rightarrow +\infty} E(\varphi_{\epsilon_n}(1, y + \sqrt{\epsilon_n} W_1)) - \varphi_{\epsilon_n}(0, x) = L(y - x) \right\}.$$

*Proof.* For each  $\epsilon > 0$ , by the definition of  $\nu_\epsilon(P_0, g_\epsilon \star P_1)$ , we can choose  $\varphi_\epsilon(t, x)$  such that

$$(4.11) \quad \frac{\partial \varphi_\epsilon(t, x)}{\partial t} + \frac{\epsilon}{2} \Delta \varphi_\epsilon(t, x) + H(\nabla \varphi_\epsilon(t, x)) = 0,$$

and choose  $\varphi_\epsilon(1, \cdot) \in C_b^\infty(\mathbf{R}^d)$  as well as

$$(4.12) \quad \nu_\epsilon(P_0, g_\epsilon \star P_1) - \epsilon \leq \int_{\mathbf{R}^d} \varphi_\epsilon(1, y) g_\epsilon \star P_1(dy) - \int_{\mathbf{R}^d} \varphi_\epsilon(0, x) P_0(dx).$$

Since  $\mu^*$  has marginals  $P_0$  and  $P_1$ , the right-hand side of this inequality can be written as

$$(4.13) \quad \int_{\mathbf{R}^d \times \mathbf{R}^d} (E(\varphi_\epsilon(1, y + \sqrt{\epsilon} W_1)) - \varphi_\epsilon(0, x)) \mu^*(dxdy).$$

Let us subtract  $T_{MK}(P_0, P_1) = \int_{\mathbf{R}^d \times \mathbf{R}^d} L(y-x) \mu^*(dxdy)$  from both sides of (4.12). We know that  $L(y-x) - (E(\varphi_\epsilon(1, y + \sqrt{\epsilon}W_1)) - \varphi_\epsilon(0, x))$  always remains nonnegative and  $\lim_{\epsilon \rightarrow 0} \nu_\epsilon(P_0, g_\epsilon \star P_1) = T_{MK}(P_0, P_1)$  (cf. Theorem 3.2 and its proof). Therefore

$$(4.14) \quad E(\varphi_\epsilon(1, y + \sqrt{\epsilon}W_1)) - \varphi_\epsilon(0, x) - L(y-x)$$

converges to 0 in  $L^1(\mu^*)$  when  $\epsilon$  goes to 0, which implies  $\mu^*$ -a.s. convergence for a subsequence  $(\epsilon_n)$ .  $\square$

At this stage, let us assume that both sequences of functions  $(\varphi_{\epsilon_n}(0, \cdot))$  and  $(E(\varphi_{\epsilon_n}(1, \cdot + \sqrt{\epsilon_n}W_1)))$  admit limits when  $n \rightarrow +\infty$  which we denote, respectively, by  $u(\cdot)$  and  $v(\cdot)$ . The pair  $(u, v)$  then satisfies

$$(4.15) \quad v(y) - u(x) \leq L(y-x) \quad \forall (x, y) \in \mathbf{R}^d \times \mathbf{R}^d$$

with equality on  $S$ . If we know that  $u(\cdot)$  is differentiable at any interior point  $(x_0, y_0)$  of  $S$ , we can conclude, as we did for the quadratic cost, that the following holds:

$$(4.16) \quad \nabla u(x_0) = \nabla L(y_0 - x_0)$$

and consequently  $y_0 = x_0 + \nabla H(\nabla u(x_0))$ . The *graph property* will hold if this argument is applicable at any  $(x_0, y_0)$  in  $S$ . Unfortunately neither separate convergence nor differentiability holds true in general; we also do not know whether interior points exist. Moreover, in the quadratic example, differentiability of  $u(\cdot)$  was a consequence of its convexity; actually existence of partial derivatives  $\frac{\partial u}{\partial x_i}$  would have been sufficient to conclude. In what follows we will approach the ideal quadratic situation by taking advantage of semiconvexity properties of value functions under relevant assumptions.

**DEFINITION 4.1.** *Let  $\Phi$  be a function defined on a convex subset of  $\mathbf{R}^d$  with values in  $\mathbf{R} \cup \{+\infty\}$ . The function  $\Phi$  is semiconvex with constant  $C$  if there exists  $C > 0$  such that  $x \mapsto \Phi(x) + C\frac{|x|^2}{2}$  is convex.*

**PROPOSITION 4.1** (cf. [5, p. 229]). *Let  $G$  be a compact subset of  $\mathbf{R}^d$  and  $\Phi$  a semiconvex function on  $G$ . Let us assume that  $x_0$  maximizes  $\Phi$  on  $G$  and belongs to the interior of  $G$ . Then  $\Phi$  is differentiable at  $x_0$  with  $D\Phi(x_0) = 0$ .*

**THEOREM 4.3.** *Let us assume that  $L$  satisfies*

$$(4.17) \quad \exists \quad C > 0, \quad D_u^2 L \leq C.$$

*Let  $\varphi(t, x)$  be a value function given by (2.4). Then,  $\varphi(0, \cdot)$  is semiconvex with constant  $C$ .*

The proofs of Theorem 4.3 and Proposition 4.1 are given in the appendix. For other sufficient conditions which guarantee that the value function is semiconvex we refer the reader to [5]. Definition 4.1 is equivalent to the requirement

$$(4.18) \quad \forall (x, z) \quad \Phi(x+z) + \Phi(x-z) - 2\Phi(x) \geq -C|z|^2$$

**Note.** From now on we will be working under assumption (4.17). We do *not* require  $L$  to satisfy the cone condition.

**DEFINITION 4.2.** *Let  $(\epsilon_n)$  denote a sequence given by Theorem 4.2. For  $a \in \mathbf{R}^d$ ,*

$$(4.19) \quad \psi_a(x) := \limsup_{n \rightarrow +\infty} (\varphi_{\epsilon_n}(0, x) - \varphi_{\epsilon_n}(0, a)).$$

**PROPOSITION 4.2.** *Under assumption (4.17), the set  $D_a := \{x \in \mathbf{R}^d; \psi_a(x) < +\infty\}$  is a convex set independent of  $a \in \pi_1(S) := \{x \in \mathbf{R}^d; \exists y \in \mathbf{R}^d, (x, y) \in S\}$ . Moreover,  $\psi_a$  is semiconvex on  $D_a$ .*

PROPOSITION 4.3. *Let us denote by  $(\mathbf{e}_i, 1 \leq i \leq n)$  the canonical basis of  $\mathbf{R}^d$  and assume that  $L \in C^1(\mathbf{R}^d)$ . Let  $a$  belong to  $\pi_1(S)$  and  $(a, b) \in S$ . If for some  $i \in \{1, \dots, d\}$  there exist sequences  $(h_n^{(i)})$  and  $(y_n^{(i)})$  such that  $h_n^{(i)} \rightarrow 0$ ,  $y_n^{(i)} \rightarrow b$ , and for all  $n$ ,  $(a + h_n^{(i)} \mathbf{e}_i, y_n^{(i)}) \in S$ , then  $\lim_{n \rightarrow +\infty} \frac{1}{h_n^{(i)}} \psi_a(a + h_n^{(i)} \mathbf{e}_i)$  exists and coincides with  $\partial_i L(b - a)$ .*

*Proof of Proposition 4.2.* Since  $(a, b) \in S$  for all  $(u, c) \in \mathbf{R}^d \times \mathbf{R}^d$ ,

$$(4.20) \quad \psi_a(u) \geq L(b - a) - L(b - u),$$

$$(4.21) \quad \psi_a(u) \geq \psi_c(u) + L(b - a) - L(b - c).$$

Indeed, since  $\varphi_{\epsilon_n}(0, c) - E(\varphi_{\epsilon_n}(1, b + \sqrt{\epsilon_n} W_1)) \geq -L(b - c)$ , the following holds:

$$\begin{aligned} \varphi_{\epsilon_n}(0, u) - \varphi_{\epsilon_n}(0, a) \\ \geq \varphi_{\epsilon_n}(0, u) - \varphi_{\epsilon_n}(0, c) + E(\varphi_{\epsilon_n}(1, b + \sqrt{\epsilon_n} W_1)) - \varphi_{\epsilon_n}(0, a) - L(b - c). \end{aligned}$$

To obtain (4.21) it remains to let  $n$  go to  $+\infty$  and apply Theorem 4.2. Inequality (4.20) follows when  $u$  equals  $c$ . Semiconvexity of  $\psi_a$  on its domain  $D_a$  follows from Theorem 4.3 and the fact that if  $(\Phi_n)$  is a sequence of semiconvex functions with the same constant  $C$ , then  $\limsup \Phi_n$  is itself semiconvex with this same constant. Therefore the set  $D_a$  is convex since it coincides with the domain of the convex function  $\psi_a + \frac{C}{2} |\cdot|^2$ . Moreover, let  $a$  and  $a'$  in  $\pi_1(S)$ . By applying (4.21) twice, to  $(a, a')$  and to  $(a', a)$ , we conclude that  $D_a = D_{a'}$ .  $\square$

*Proof of Proposition 4.3.* Take  $i \in \{1, \dots, d\}$ ,  $a \in \pi_1(S)$  and  $b$  such that  $(a, b) \in S$ . Inequality (4.21) implies

$$(4.22) \quad L(b - a) - L(b - (a + h_n^{(i)} \mathbf{e}_i)) \leq \psi_a(a + h_n^{(i)} \mathbf{e}_i) \leq L(y_n^{(i)} - a) - L(y_n^{(i)} - (a + h_n^{(i)} \mathbf{e}_i)).$$

The desired statement follows since  $L$  is  $C^1$  by letting  $n \rightarrow +\infty$ .  $\square$

For  $a \in \pi_1(S)$  and  $b$  such that  $(a, b)$  belongs to  $S$ , we see that the function  $x \mapsto L(b - a) - \psi_a(x)$  plays the same role as  $x \mapsto v(b) - u(x)$  in (4.15):  $L(b - a) - \psi_a(x) \leq L(b - x)$  on  $\mathbf{R}^d$  with equality when  $x = a$ . We cannot apply Proposition 4.1 since we do not know whether  $\pi_1(S)$  has interior points. However, suppose that the assumptions of Proposition 4.3 are satisfied for all  $i \in \{1, \dots, d\}$ . Then we can set  $\nabla \psi_a$  to be the vector

$$(4.23) \quad \nabla \psi_a := \left( \lim_{n \rightarrow +\infty} \frac{1}{h_n^{(i)}} \psi_a(a + h_n^{(i)} \mathbf{e}_i); \quad i = 1 \leq d \right).$$

Hence  $b$  is uniquely given by  $a + \nabla H(\nabla \Psi_a)$ .

Let us now assume that  $P_0$  is absolutely continuous w.r.t. the Lebesgue measure on  $\mathbf{R}^d$ . We prove below that the set of points  $a \in \pi_1(S)$ , where  $\nabla \psi_a(a)$  does not exist, has Lebesgue measure 0. It suffices to show that the set  $\Pi := \{a \in \pi_1(S); \partial_1 \psi_a(a) \text{ does not exist}\}$  has Lebesgue measure 0. Let us first make the following remark: For  $\alpha \in \pi_1(S)$  and  $\beta$  such that  $(\alpha, \beta) \in S$ , consider

$$\begin{aligned} U_n^+(\alpha, \beta) &= \{(x, y) \in \mathbf{R}^d \times \mathbf{R}^d; x = \alpha + h \mathbf{e}_1, h > 0, |h|^2 + |y - \beta|^2 < n^{-2}\}, \\ U_n^-(\alpha, \beta) &= \{(x, y) \in \mathbf{R}^d \times \mathbf{R}^d; x = \alpha + h \mathbf{e}_1, h < 0, |h|^2 + |y - \beta|^2 < n^{-2}\}; \end{aligned}$$

if  $U_n^+(\alpha, \beta) \cap S \neq \emptyset$  or  $U_n^-(\alpha, \beta) \cap S \neq \emptyset$  for a sequence of integers  $n$  going to  $+\infty$ , then  $\partial_1 \psi_\alpha(\alpha)$  exists as a consequence of Proposition 4.3. Take now  $a \in \Pi$  and  $(a, b) \in S$ .

There exists  $N = N(a, b) \geq 1$  such that  $U_N^+(a, b) \cap S = U_N^-(a, b) \cap S = \emptyset$ . Therefore the set  $\sigma := \{x \in \mathbf{R}; a + x\mathbf{e}_1 \in \mathbf{\Pi}\}$  is at most countable and thus has Lebesgue measure 0. We have just proved the following theorem.

**THEOREM 4.4.** *Let  $P_0(dx) \ll dx$  and  $T_{MK}(P_0, P_1) < +\infty$ . Under assumption (4.17), the graph property holds.*

### 5. Appendix.

*Proof of Theorem 4.3.* Let  $(X_t^x; t \in [0, 1])$  in  $\mathcal{A}_\epsilon$  be optimal for (2.4); namely,  $X_t^x = x + \int_0^t \nabla H(\nabla \varphi(s, X_s^x)) ds + \sqrt{\epsilon} W_t$  and

$$(5.1) \quad \varphi(0, x) = E \left[ f(X_1^x) - \int_0^1 L(\nabla H(\nabla \varphi(s, X_s^x))) ds \right]$$

with  $\varphi(1, \cdot) = f(\cdot)$ . For  $z \in \mathbf{R}^d$  let us set  $X_t^1 := X_t^x + (1-t)z$  and  $X_t^2 := X_t^x - (1-t)z$ ; these processes both belong to  $\mathcal{A}_\epsilon$  and satisfy  $X_0^1 = x + z$ ,  $X_0^2 = x - z$ , and  $X_1^1 = X_1^2 = X_1^x$ . Let  $\beta_t^x := \nabla H(\nabla \varphi(t, X_t^x))$ . From the definition of  $\varphi$  in (2.4) it follows that

$$(5.2) \quad \begin{aligned} & \varphi(0, x+z) + \varphi(0, x-z) - 2\varphi(0, x) \\ & \geq E \int_0^1 (2L(\beta_t^x) - L(\beta_t^x + z) - L(\beta_t^x - z)) dt. \end{aligned}$$

The conclusion follows from assumption (4.17).  $\square$

*Proof of Proposition 4.1.* Let  $\mathbf{B}$  be an open ball centered at  $x_0$  included in  $G$ . Such a ball exists since  $x_0$  is an interior point of  $G$ . The function  $x \mapsto \Psi(x) := \Phi(x) + C \frac{|x-x_0|^2}{2}$  is convex on  $\mathbf{B}$  for some constant  $C > 0$ . Therefore there exists a vector  $b \in \mathbf{R}^d$  such that, for all  $x \in \mathbf{B}$ ,  $\Psi(x) \geq \Psi(x_0) + \langle b, x - x_0 \rangle$ . Moreover, since  $x_0$  maximizes  $\Phi$  on  $\mathbf{B}$ ,

$$(5.3) \quad \langle b, x - x_0 \rangle \leq C \frac{|x - x_0|^2}{2} \quad \forall x \in \mathbf{B}.$$

For  $\epsilon > 0$  small enough, the point  $x = x_0 + \epsilon b$  belongs to  $\mathbf{B}$ . We conclude that  $b = 0$  since it must satisfy  $\epsilon|b|^2 \leq \frac{C}{2} \epsilon^2 |b|^2$  for all  $\epsilon$  small enough.  $\square$

**Acknowledgments.** We thank two anonymous referees for their comments and suggestions which helped us to improve the first version of this paper. This work was done during the visit of the second author (M. Thieullen) to the University of Hokkaido. She would like to thank this university for its hospitality.

### REFERENCES

- [1] Y. BRENIER AND J. D. BENAMOU, *A numerical method for the optimal mass transport problem and related problems*, in Monge Ampère Equation: Applications to Geometry and Optimization, Proceedings of the NSF-CBMS Conference (Deerfield Beach, FL, 1997), L. A. Caffarelli and M. Milman, eds., Contemp. Math. 226, AMS, Providence, RI, 1999, pp. 1–11.
- [2] F. DELARUE, *On the existence and uniqueness of solutions to FBSDEs in a nondegenerate case*, Stochastic Process. Appl., 99 (2002), pp. 209–286.
- [3] L. C. EVANS, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.
- [4] L. C. EVANS, *Partial differential equations and Monge–Kantorovich mass transfer*, in Current Developments in Mathematics (Cambridge, MA, 1997), S. T. Yau, ed., Int. Press, Boston, MA, 1999, pp. 65–126.
- [5] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1993.

- [6] W. GANGBO AND R. J. MCCANN, *The geometry of optimal transportation*, Acta Math., 177 (1996), pp. 113–161.
- [7] L. V. KANTOROVICH, *On the translocation of masses*, C. R. (Dokl.) Acad. Sci. URSS, 37 (1942), pp. 199–201; reprinted in J. Math. Sci., 133 (2006), pp. 1381–1382.
- [8] H. G. KELLERER, *Duality theorem for marginal problems*, Z. Wahrsch. Verw. Gebiete, 67 (1984), pp. 399–432.
- [9] T. MIKAMI, *Optimal control for absolutely continuous stochastic processes and the mass transportation problem*, Electron. Comm. Probab., 7 (2002), pp. 199–213.
- [10] T. MIKAMI, *Monge’s problem with a quadratic cost by the zero noise limit of  $h$ -path processes*, Probab. Theory Related Fields, 129 (2004), pp. 245–260.
- [11] T. MIKAMI AND M. THIEULLEN, *Duality theorem for the stochastic optimal control problem*, Stochastic Process Appl., 116 (2006), pp. 1815–1835.
- [12] S. T. RACHEV AND L. RÜSCHENDORF, *Mass Transportation Problems, Vol. I: Theory, Vol. II: Application*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1998.
- [13] L. RÜSCHENDORF AND W. THOMSEN, *Note on the Schrödinger equation and  $I$ -projections*, Statist. Probab. Lett., 17 (1993), pp. 369–375.
- [14] C. VILLANI, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.
- [15] W. A. ZHENG, *Tightness results for laws of diffusion processes application to stochastic mechanics*, Ann. Inst. H. Poincaré Probab. Statist., 21 (1985), pp. 103–124.



## CONTROLLABILITY OF LINEAR DISCRETE SYSTEMS WITH CONSTANT COEFFICIENTS AND PURE DELAY\*

JOSEF DIBLÍK<sup>†</sup>, DENYS YA. KHUSAINOV<sup>‡</sup>, AND M. RŮŽIČKOVÁ<sup>§</sup>

**Abstract.** The purpose of this contribution is to develop a controllability method for linear discrete systems with constant coefficients and with pure delay. To do this, a representation of solutions with the aid of a discrete matrix delayed exponential is used. Such an approach leads to new conditions of controllability. Except for a criterion of relative controllability, a control function is constructed as well.

**Key words.** discrete system, delay, discrete matrix delayed exponential, relative controllability, control function

**AMS subject classifications.** 39A10, 39A11, 93B05

**DOI.** 10.1137/070689085

**1. Introduction.** In control theory, one of the best solved problems is the problem of controllability of linear time-independent differential systems with one input,

$$(1.1) \quad \dot{x}(t) = Ax(t) + bu(t),$$

where  $x: R_+ := [0, \infty) \rightarrow R^n$  is continuously differentiable,  $A$  is an  $n \times n$  constant matrix,  $b$  is a constant vector, and input  $u: R_+ \rightarrow R$  (see, e.g., [9, 11]). As usual, system (1.1) is called controllable if, for arbitrary finite initial state  $x(0) = x_0$  and terminal state  $x(t_1) = x_1$  with finite  $t_1 > 0$ , there exists a control  $u^*(t)$  such that the system

$$\dot{x}(t) = Ax(t) + bu^*(t)$$

has a solution  $x = x^*(t)$  satisfying  $x^*(0) = x_0$  and  $x^*(t_1) = x_1$ .

A much more complicated situation occurs if we deal with differential equations with an aftereffect having, e.g., the form

$$\begin{aligned} \dot{x}(t) &= Bx(t - \tau) + bu(t), \quad t > 0, \\ x(t) &= \varphi(t), \quad -\tau \leq t \leq 0, \quad \tau > 0, \end{aligned}$$

where  $x: [-\tau, \infty) \rightarrow R^n$  is continuously differentiable on  $(0, \infty)$ ,  $B$  is an  $n \times n$  constant matrix,  $b$  is a constant vector, and input  $u: R_+ \rightarrow R$ ,  $\tau = \text{const}$  and  $\varphi: [-\tau, 0] \rightarrow R^n$ . We refer to [6], where various classes of control for delayed differential systems are considered and different variants of controllability are developed.

---

\*Received by the editors April 23, 2007; accepted for publication (in revised form) November 27, 2007; published electronically March 19, 2008. The preliminary version of this paper was written at Oberwolfach, Germany during the Research in Pairs (RiP) program.

<http://www.siam.org/journals/sicon/47-3/68908.html>

<sup>†</sup>Brno University of Technology, 602 00 Brno, Czech Republic (diblik@feec.vutbr.cz). This author was supported by grant 201/07/0145 of the Czech Grant Agency (Prague) and by the Council of Czech Government grant MSM 00216 30503.

<sup>‡</sup>Department of Complex Systems Modelling, Kiev University, 01 033 Kiev, Ukraine (khusainov@unicyb.kiev.ua).

<sup>§</sup>Department of Applied Mathematics, Žilina University, 010 26 Žilina, Slovak Republic (miroslava.ruzickova@fpv.uniza.sk). This author was supported by grant 1/3238/06 of the Grant Agency of Slovak Republic (VEGA).

We use the following notation: For integers  $s, q$ ,  $s \leq q$  we define  $Z_s^q := \{s, s+1, \dots, q\}$ , where either  $s = -\infty$  or  $q = \infty$  is admitted, too. Throughout this paper, using notation  $Z_s^q$  (perhaps with another pair of integers), we always assume  $s \leq q$ .

In the present paper we consider discrete controlled systems with pure delay,

$$(1.2) \quad \Delta x(k) = Bx(k-m) + bu(k),$$

where  $m \geq 1$  is a fixed integer,  $k \in Z_0^\infty$ ,  $B$  is a constant  $n \times n$  matrix,  $\Delta x(k) = x(k+1) - x(k)$ ,  $x: Z_{-m}^\infty \rightarrow R^n$  is an unknown solution,  $b \in R^n$  is a given nonzero vector, and  $u: Z_0^\infty \rightarrow R$  is the input scalar function. Following the standard terminology (used, e.g., in [1, 2]) we refer to (1.2) as a delayed discrete system if  $m \geq 1$  and as a nondelayed discrete system if  $m = 0$ . Discrete systems containing only one delay are often called *systems with pure delay*.

Together with (1.2) we consider an initial (Cauchy) problem

$$(1.3) \quad x(k) = \varphi(k)$$

with a given initial function  $\varphi: Z_{-m}^0 \rightarrow R^n$ . The *existence* and *uniqueness* of the solution of the initial problem (1.2), (1.3) on  $Z_{-m}^\infty$  are obvious. We recall that a *solution*  $x: Z_{-m}^\infty \rightarrow R^n$  of initial Cauchy problem (1.2), (1.3) is defined as an *infinite sequence*

$$\{x(-m) = \varphi(-m), x(-m+1) = \varphi(-m+1), \dots, x(0) = \varphi(0), x(1), x(2), \dots, x(k), \dots\}$$

such that, for any  $k \in Z_0^\infty$ , equality (1.2) holds.

Throughout the paper we adopt the customary notation for the sum and the product:  $\sum_{i=k+s}^k \circ(i) = 0$  and  $\prod_{i=k+s}^k \circ(i) = 1$ , where  $k$  is an integer,  $s$  is a positive integer, and “ $\circ$ ” denotes the function considered independently of whether it is defined for indicated arguments or not. For  $\varphi: Z_s^q \rightarrow R^n$  we define  $\|\varphi\|_{sq} := \max_{j=s, \dots, q} \|\varphi(j)\|$  with  $\|\varphi(j)\| := \max_{i=1, \dots, n} |\varphi_i(j)|$ .

**1.1. Description of the problem considered.** Problems of control in discrete systems are considered, e.g., in [5, 7, 10, 12]. In [10] attention is paid to discrete delayed systems as well. Unlike the given recommendation to transform delayed systems into systems without delay but with greater dimensionality, we perform direct investigation using a new technique.

We will consider controllability on finite intervals. In this paper we deal with the so-called relatively controllable discrete systems (1.2) within the meaning of the following definition.

**DEFINITION 1.1.** *System (1.2) is relatively controllable if for any initial function  $\varphi: Z_{-m}^0 \rightarrow R^n$ , any finite terminal state  $x = x^* \in R^n$ , and any finite terminal point  $k_1$  greater than or equal to a fixed integer  $k^* \in Z_1^\infty$  there exists a discrete function  $u^*: Z_0^{k_1-1} \rightarrow R$  such that the system (1.2) with the input  $u = u^*$ , i.e., the system*

$$\Delta x(k) = Bx(k-m) + bu^*(k),$$

*has a solution  $x^*: Z_{-m}^{k_1} \rightarrow R^n$  such that  $x^*(k_1) = x^*$  and  $x^*(k) = \varphi(k)$  if  $k \in Z_{-m}^0$ .*

In this paper we will derive not only conditions for relative controllability of system (1.2), but also an explicit form of a possible control function  $u^*$ .

**2. Preliminaries.** In investigating relative controllability we use an auxiliary result, which is an analogy of the well-known basic lemma of variational calculus.

LEMMA 2.1. *Let a function  $\Phi: Z_s^q \rightarrow R$  be given. Then the equality*

$$(2.1) \quad \sum_{k=s}^q \Phi(k) \eta(k) = 0$$

*with an arbitrary function  $\eta: Z_s^q \rightarrow R$  holds if and only if  $\Phi(k) = 0$  on  $Z_s^q$ .*

The proof of the lemma is obvious and we omit it.

The known proof of controllability for linear differential systems with one input (Kalman's criterion) is based on two important results. The first is the formula for integral representation of a solution of a Cauchy problem for nonhomogeneous system (1.1),

$$x(t) = e^{At}x_0 + b \int_0^t e^{A(t-s)}u(s)ds,$$

where  $\exp(At)$  is the matrix exponential defined as

$$e^{At} = I + \frac{1}{1!}At + \frac{1}{2!}(At)^2 + \cdots + \frac{1}{k!}(At)^k + \cdots$$

with a unit  $n \times n$  matrix  $I$ . The second is the Cayley–Hamilton theorem, which states that every power  $A^i$ ,  $i = n, n+1, \dots$ , of a given  $n \times n$  constant matrix  $A$  can be expressed as a linear combination of a finite number of matrices  $I, A, A^2, \dots, A^{n-1}$ . Concerning these facts (which serve as motivation in our research) we refer the reader to, e.g., [8].

In this paper we will use what is called a discrete matrix delayed exponential  $\exp_m(Bk)$ . Its definition was given in [3, 4]. We reproduce this definition. Let a constant  $n \times n$  matrix  $B$  be given. Then

$$e_m^{Bk} := \begin{cases} \Theta & \text{if } k \in Z_{-\infty}^{-m-1}, \\ I & \text{if } k \in Z_{-m}^0, \\ I + B \cdot \binom{k}{1} & \text{if } k \in Z_1^{m+1}, \\ I + B \cdot \binom{k}{1} + B^2 \cdot \binom{k-m}{2} & \text{if } k \in Z_{(m+1)+1}^{2(m+1)}, \\ I + B \cdot \binom{k}{1} + B^2 \cdot \binom{k-m}{2} + B^3 \cdot \binom{k-2m}{3} & \text{if } k \in Z_{2(m+1)+1}^{3(m+1)}, \\ \dots & \\ I + B \cdot \binom{k}{1} + B^2 \cdot \binom{k-m}{2} + \dots + B^\ell \cdot \binom{k-(\ell-1)m}{\ell} & \\ \text{if } k \in Z_{(\ell-1)(m+1)+1}^{\ell(m+1)}, \quad \ell = 0, 1, 2, \dots, \end{cases}$$

where  $\Theta$  is an  $n \times n$  null matrix.

The following result is proved in [4].

LEMMA 2.2. *Let  $B$  be a constant  $n \times n$  matrix. Then*

$$(2.2) \quad \Delta e_m^{Bk} = B e_m^{B(k-m)}$$

*for  $k \in Z_{-m}^\infty$ .*

The discrete matrix delayed exponential  $\exp_m(Bk)$  of the  $n \times n$  constant matrix  $B$  can be successfully used for representing the solutions of discrete systems. In [4] a formula expressing the solution of an initial Cauchy problem with the aid of a discrete matrix exponential was given. In the case of the problem (1.2), (1.3) it takes the form

$$(2.3) \quad x(k) = e_m^{Bk} \varphi(-m) + \sum_{j=-m+1}^0 e_m^{B(k-m-j)} \Delta \varphi(j-1) \\ + \sum_{j=1}^k e_m^{B(k-m-j)} bu(j-1),$$

where  $k \in Z_1^\infty$ .

**DEFINITION 2.3.** Let a positive number  $r$  be given. We define a class of bounded discrete functions  $\Omega_r(s, q)$  as

$$\Omega_r(s, q) := \{\omega: Z_s^q \rightarrow R, \|\omega\|_{sq} \leq r\}.$$

**DEFINITION 2.4.** We define a domain of reachability  $Q_\varphi \subseteq R^n$  as a set of all points  $x(k_1)$ , where  $x$  is a solution of the system (1.2) corresponding to fixed initial data (1.3) and to an arbitrary control  $u \in \Omega_r(0, k_1 - 1)$ .

**3. Main results.** We obtain conditions for relative controllability and we find the control function  $u = u^*(k)$ ,  $k \in Z_0^{k_1-1}$ , for the problem

$$(3.1) \quad \Delta x(k) = Bx(k-m) + bu(k), \quad k \in Z_0^{k_1-1},$$

$$(3.2) \quad x(k) = \varphi(k), \quad k \in Z_{-m}^0,$$

$$(3.3) \quad x(k_1) = x^*.$$

We define an auxiliary  $n \times n$  matrix

$$S := (b, Bb, B^2b, \dots, B^{n-1}b)$$

and the vector

$$(3.4) \quad \xi := x^* - e_m^{Bk_1} \varphi(-m) - \sum_{j=-m+1}^0 e_m^{B(k_1-m-j)} \Delta \varphi(j-1).$$

Moreover, we define a number

$$(3.5) \quad k^* := (n-1)(m+1) + 1.$$

### 3.1. Criterion of relative controllability of the problem (3.1)–(3.3).

**THEOREM 3.1.** Problem (3.1)–(3.3) is relatively controllable if and only if assumptions

$$(3.6) \quad \text{rank } S = n$$

and

$$(3.7) \quad k_1 \geq k^*$$

hold simultaneously.

*Proof.* (a) Let the system (3.1) be relatively controllable. Then, in accordance with Definition 1.1 for any arbitrary initial function  $\varphi: Z_{-m}^0 \rightarrow R^n$ , arbitrary finite terminal state  $x = x^* \in R^n$ , and arbitrary finite terminal point  $k_1$  greater than or equal to a fixed integer  $k^* \in Z_1^\infty$ , there exists a discrete function  $u^*: Z_0^{k_1-1} \rightarrow R$  such that the system (3.1) has a solution  $x^*: Z_{-m}^{k_1} \rightarrow R^n$  satisfying assumption (3.3). We show that, in this case, (3.6) and (3.7) with  $k^*$  defined by (3.5) hold. We use the formula (2.3) to represent the solution  $x^*$  of the problem (3.1)–(3.3). At the moment  $k = k_1$  we have

$$(3.8) \quad x^* = x^*(k_1) = e_m^{Bk_1} \varphi(-m) + \sum_{j=-m+1}^0 e_m^{B(k_1-m-j)} \Delta \varphi(j-1) + \sum_{j=1}^{k_1} e_m^{B(k_1-m-j)} bu^*(j-1).$$

We rewrite (3.8) (using (3.4)) as

$$(3.9) \quad \sum_{j=1}^{k_1} e_m^{B(k_1-m-j)} bu^*(j-1) = \xi.$$

The discrete matrix delayed exponential  $\exp_m(Bk)$  contains the same maximal degree  $p(k)$  of the matrix  $B$  for every set of values

$$k = (\ell - 1)(m + 1) + 1, (\ell - 1)(m + 1) + 2, \dots, \ell(m + 1)$$

with fixed  $\ell = 0, 1, 2, \dots$ . It is easy to see (the function  $[\cdot]$  is the greatest integer function) that

$$p(k) = \begin{cases} \left[ \frac{k-1}{m+1} \right] + 1 & \text{if } k = 1, 2, \dots, \\ 0 & \text{if } k < 1. \end{cases}$$

Consequently, the definition of a discrete matrix delayed exponential can be shortened to

$$e_m^{Bk} = \begin{cases} 0 & \text{if } k \in Z_{-\infty}^{-m}, \\ I + \sum_{i=1}^{p(k)} B^i \cdot \binom{k - (i-1)m}{i} & \text{if } k \in Z_{-m}^{\infty}. \end{cases}$$

Since  $k_1 \geq 1$ , the relation (3.9) becomes

$$(3.10) \quad \sum_{j=1}^{k_1} \left[ I + \sum_{i=1}^{p(k_1-m-j)} B^i \cdot \binom{k_1 - m - j - (i-1)m}{i} \right] bu^*(j-1) = \xi.$$

The left-hand side is a linear combination of  $q + 1$  vectors

$$(3.11) \quad b, Bb, \dots, B^q b$$

with

$$(3.12) \quad q = p(k_1 - m - 1) = \left[ \frac{k_1 - m - 1 - 1}{m + 1} \right] + 1 = \left[ \frac{k_1 - 1}{m + 1} \right].$$

In other words, (3.10) can be written in the form

$$(3.13) \quad C_1 \cdot b + C_2 \cdot Bb + \cdots + C_{q+1} \cdot B^q b = \xi$$

with “constants”  $C_1, C_2, \dots, C_{q+1}$  (being functions of  $u^*(0), u^*(1), \dots, u^*(k_1 - 1)$ ). We conclude that, for  $q + 1 < n$ , system (3.13) is, in general, overdetermined and its solution (for arbitrary  $\xi$ ) does not always exist. Therefore, a necessary condition for the solvability of (3.13) is the inequality  $q + 1 \geq n$ . Then, the solution of (3.13) always exists. Finally, we have that (see (3.12))

$$q = \left\lceil \frac{k_1 - 1}{m + 1} \right\rceil \geq n - 1$$

or

$$k_1 \geq (n - 1)(m + 1) + 1 = k^*.$$

The validity of inequality (3.7) is now obvious. By the Cayley–Hamilton theorem, any matrix  $B^i$  with  $i \geq n$  can be expressed as a linear combination of matrices

$$I, B, B^2, \dots, B^{n-1}.$$

If  $q \geq n$ , then the linear combination (3.13) of vectors of system (3.11) can be reduced to a linear combination of vectors of the following system:

$$b, Bb, \dots, B^{n-1}b.$$

Thus, the above-mentioned system (3.13) of  $n$  linear equations is solvable for any arbitrary right-hand side  $\xi$  if and only if  $\det S \neq 0$ , i.e.,  $\text{rank } S = n$ . Consequently, condition (3.6) holds.

(b) Let (3.6), (3.7) be valid. We show that the system (3.1) is relatively controllable. First, we prove that the dimension of the domain of reachability equals  $n$ . Suppose the contrary, i.e.,  $\dim Q_\varphi < n$ . Then there exists a nontrivial constant vector  $y = (y_1, y_2, \dots, y_n)^T \in R^n$  such that, for any control  $u \in \Omega_r(0, k_1 - 1)$  and the corresponding solution  $x = x(k; u, \varphi)$  of the problem (3.1), (3.2), we have

$$(3.14) \quad y^T x(k_1) = 0.$$

We choose the zero initial function  $\varphi(k) = 0$ ,  $k \in Z_{-m}^0$ . Then, with the aid of formula (2.3), we get

$$(3.15) \quad x(k_1) = \sum_{j=1}^{k_1} e_m^{B(k_1-m-j)} b u(j-1),$$

and (3.14) becomes

$$(3.16) \quad y^T \sum_{j=1}^{k_1} e_m^{B(k_1-m-j)} b u(j-1) = 0.$$

Since (3.16) holds for arbitrary control  $u \in \Omega_r(0, k_1 - 1)$  then, by Lemma 2.1,

$$(3.17) \quad y^T e_m^{B(k_1-m-j)} b = 0, \quad j \in Z_0^{k_1}.$$

Now we apply (2.2) to (3.17). We obtain

$$\Delta \left[ y^T e_m^{B(k_1-m-j)} b \right] = y^T \Delta \left[ e_m^{B(k_1-m-j)} \right] b = y^T B e_m^{B(k_1-2m-j)} b = 0$$

or

$$(3.18) \quad y^T B e_m^{B(k_1-2m-j)} b = 0, \quad j \in Z_0^{k_1}.$$

We repeatedly apply the operator  $\Delta$  to (3.18) for a total  $(n-2)$  times. Then, in accordance with (2.2),

$$(3.19) \quad y^T B^2 e_m^{B(k_1-3m-j)} b = 0, \quad j \in Z_0^{k_1},$$

$$(3.20) \quad y^T B^3 e_m^{B(k_1-4m-j)} b = 0, \quad j \in Z_0^{k_1},$$

...

$$(3.21) \quad y^T B^{n-1} e_m^{B(k_1-nm-j)} b = 0, \quad j \in Z_0^{k_1}.$$

We put  $j = k_1$  in (3.17),  $j = k_1 - m$  in (3.18) and, step by step,

$$j = k_1 - 2m, j = k_1 - 3m, \dots, j = k_1 - (n-1)m$$

in (3.19), (3.20), and (3.21). Since, due to (3.7), inequality  $k_1 \geq (n-1)(m+1) + 1$  holds, we get  $k_1 - (n-1)m \geq n \geq 1$ , and all choices of indices  $j$  are admissible. Since  $e_m^{B(-m)} \equiv I$ , system (3.17)–(3.21) (with respect to  $b$ ) reduces to

$$(3.22) \quad y^T b = 0, y^T B b = 0, \dots, y^T B^{n-1} b = 0.$$

We use the definition of the matrix  $S$  and rewrite (3.22) as

$$(3.23) \quad y^T S = 0.$$

The homogeneous system (3.23) has a nontrivial solution if and only if its matrix is singular. This means  $\det S = 0$ . We get a contradiction with (3.6). Consequently, the inequality  $\dim Q_\varphi < n$  does not hold and the dimension of the domain of reachability equals  $n$ .

Since the domain of reachability  $Q_\varphi$  contains the point  $-x(k_1)$  corresponding to a control  $-u \in \Omega_r(0, k_1 - 1)$  together with a point  $x(k_1)$  corresponding to a control  $u \in \Omega_r(0, k_1 - 1)$  (in view of the choice of a zero initial function and in view of representation (3.15)), we conclude that  $Q_\varphi$  is symmetric. Moreover, the domain of reachability contains the segment connecting the point  $x(k_1)$  with the point  $-x(k_1)$ . Consequently, due to the linearity of the problem considered, it contains a ball of a radius  $\delta$ :

$$U_\delta := \{x \in R^n, \|x\| < \delta\}$$

with positive  $\delta$ . Obviously, if  $r \rightarrow \infty$  in  $\Omega_r(0, k_1 - 1)$ , then, due to the finiteness of the interval  $Z_0^{k_1-1}$ , we conclude that  $\delta \rightarrow \infty$  in the definition of  $U_\delta$ . Therefore, the domain  $Q_\varphi$  coincides with the whole space  $R^n$ . Simultaneously it says that, for every point  $x^* \in R^n$ , there exists a control  $u = u^*$  solving the problem (3.1)–(3.3). This conclusion remains valid even in the case of the nonzero initial function  $\varphi$ . Indeed, a simple transformation  $x(k) = x_\varphi(k) + z(k)$ , where  $x_\varphi(k)$  is a solution of the homogeneous problem

$$\begin{aligned} \Delta x_\varphi(k) &= Bx_\varphi(k-m), \quad k \in Z_0^{k_1-1}, \\ x_\varphi(k) &= \varphi(k), \quad k \in Z_{-m}^0, \end{aligned}$$

leads to the same problem with respect to  $z$  with a zero initial function.  $\square$

**3.2. A control function for the problem (3.1)–(3.3).** In this section we construct a control function  $u = u^*$  for the problem (3.1)–(3.3). To do this, we need some auxiliary results.

DEFINITION 3.2. We say that the pair  $(B, b)$  is controlled if  $\text{rank } S = n$ .

LEMMA 3.3. Let the pair  $(B, b)$  be controlled and  $k_1 \geq k^*$ . Then the elements  $(e_m^{Bk}b)_i$ ,  $i = 1, 2, \dots, n$ , of the vector  $e_m^{Bk}b$  are linearly independent on  $Z_{-m}^{k^*-1}$ , i.e., no nontrivial constant vector  $l = (l_1, l_2, \dots, l_n)^T$  exists such that

$$(3.24) \quad l^T e_m^{Bk}b = 0 \quad \text{for } k \in Z_{-m}^{k^*-1}.$$

*Proof.* Suppose that, on the contrary, there exists a nontrivial vector  $l$  such that (3.24) holds. Itemizing (3.24) in accordance with the definition of the discrete matrix exponential  $\exp_m(Bk)$  and noting that  $k^* - 1 = (n - 1)(m + 1)$ , we get

$$(3.25) \quad 0 = l^T e_m^{Bk}b = \begin{cases} l^T b & \text{if } k \in Z_{-m}^0, \\ l^T \left[ I + B \cdot \binom{k}{1} \right] b & \text{if } k \in Z_1^{m+1}, \\ l^T \left[ I + B \cdot \binom{k}{1} + B^2 \cdot \binom{k-m}{2} \right] b & \text{if } k \in Z_{(m+1)+1}^{2(m+1)}, \\ \dots \\ l^T \left[ I + B \cdot \binom{k}{1} + \dots + B^{n-1} \cdot \binom{k-(n-2)m}{n-1} \right] b \\ \text{if } k \in Z_{(n-2)(m+1)+1}^{(n-1)(m+1)}, \text{ where } (n-1)(m+1) = k^* - 1. \end{cases}$$

From (3.25) we get (due to formula (2.2))

$$(3.26) \quad \begin{aligned} 0 &= \Delta [l^T e_m^{Bk}b] = l^T \Delta [e_m^{Bk}] b = l^T B e_m^{B(k-m)} b, \\ &= \begin{cases} 0 & \text{if } k \in Z_{-m}^1, \\ l^T B b & \text{if } k \in Z_0^m, \\ l^T B \left[ I + B \cdot \binom{k}{1} \right] b & \text{if } k \in Z_{m+1}^{2m+1}, \\ \dots \\ l^T \left[ I + B \cdot \binom{k}{1} + \dots + B^{n-2} \cdot \binom{k-(n-3)m}{n-2} \right] b \\ \text{if } k \in Z_{(n-2)(m+1)}^{(n-1)(m+1)-1}, \text{ where } (n-1)(m+1) - 1 = k^* - 2. \end{cases} \end{aligned}$$

We continue to compute further differences up to those of order  $(n - 1)$ . Finally, we get from (3.26) (due to formula (2.2))

$$(3.27) \quad \begin{aligned} 0 &= \Delta^{n-1} [l^T e_m^{Bk}b] = l^T B^{n-1} e_m^{B(k-(n-1)m)} b \\ &= \begin{cases} 0 & \text{if } k \in Z_{-m}^{(n-2)m-1}, \\ l^T B^{n-1} b = 0 & \text{if } k \in Z_{(n-2)m}^{(n-1)m}, \text{ where } (n-1)m = k^* - n. \end{cases} \end{aligned}$$

Now we select the equation  $l^T b = 0$  from (3.25), the equation  $l^T B b = 0$  from (3.26), and, proceeding similarly up to the last formula (3.27), we finally select the equality  $l^T B^{n-1} b = 0$ . We get the system of equations

$$l^T S = 0.$$



This is a system of homogeneous equations and its nontrivial solution can exist only in the case of  $\det S = 0$ . This is a contradiction with our supposition, and the components of the vector  $e_m^{Bk}b$  in (3.24) are linearly independent on the interval  $Z_{-m}^{k^*-1}$ .  $\square$

LEMMA 3.4. *Let  $k_1 \geq k^*$ . Then the matrix*

$$G = \sum_{j=1}^{k_1} e_m^{B(k_1-m-j)} b b^T \left( e_m^{B(k_1-m-j)} \right)^T$$

*is nonsingular.*

*Proof.* Since the components  $(e_m^{Bk}b)_i$ ,  $i = 1, 2, \dots, n$ , of the vector  $e_m^{Bk}b$  are linearly independent on interval  $Z_{-m}^{k^*-1}$  by Lemma 3.3, we have

$$\left( e_m^{B(k_1-m-j)} b \right)^T l = \sum_{i=1}^n l_i \left( e_m^{B(k_1-m-j)} b \right)_i \neq 0$$

for arbitrary nontrivial vectors  $l = (l_1, l_2, \dots, l_n)^T$  and for  $j = k_1 - n(m+1) + 1, \dots, k_1$  (note that  $k_1 - m - j \in Z_{-m}^{k^*-1}$ ), or

$$\sum_{j=1}^{k_1} \left[ \left( e_m^{B(k_1-m-j)} b \right)^T l \right]^2 > 0$$

since  $k_1 \geq n$ . Let us transform the expression on the left-hand side of the last inequality. We have

$$\begin{aligned} \sum_{j=1}^{k_1} \left[ \left( e_m^{B(k_1-m-j)} b \right)^T l \right]^2 &= \sum_{j=1}^{k_1} \left[ \left( e_m^{B(k_1-m-j)} b \right)^T l \right] \left[ \left( e_m^{B(k_1-m-j)} b \right)^T l \right] \\ &= \sum_{j=1}^{k_1} \left[ l^T e_m^{B(k_1-m-j)} b \right] \left[ b^T \left( e_m^{B(k_1-m-j)} \right)^T l \right] \\ &= l^T \left[ \sum_{j=1}^{k_1} e_m^{B(k_1-m-j)} b b^T \left( e_m^{B(k_1-m-j)} \right)^T \right] l \\ &= l^T G l. \end{aligned}$$

Thus,  $l^T G l > 0$  and therefore  $\det G \neq 0$ .  $\square$

Using the assertions given by Lemmas 3.3 and 3.4, it is easy to prove the final result.

THEOREM 3.5. *Let the conditions of relative controllability (3.6), (3.7) be valid. Then a control function  $u = u^*$  for the problem (3.1)–(3.3) can be expressed in the form*

$$(3.28) \quad u^*(k) = b^T \left( e_m^{B(k_1-m-k-1)} \right)^T G^{-1} \xi,$$

where the vector  $\xi$  is defined by formula (3.4) and  $k \in Z_0^{k_1-1}$ .

*Proof.* Since the control function  $u^*(j)$ ,  $j \in Z_0^{k_1-1}$ , should satisfy (3.8), it is necessary to prove that the system

$$(3.29) \quad \sum_{j=1}^{k_1} e_m^{B(k_1-m-j)} b u(j-1) = \xi$$

has a solution  $u(j-1) = u^*(j-1)$ ,  $j \in Z_1^{k_1}$ . Let us try to find the control function in the form of a linear combination

$$(3.30) \quad u(j-1) = \left( e_m^{B(k_1-m-j)} b \right)^T D,$$

where  $D = (D_1, D_2, \dots, D_n)^T$  is an unknown vector. We apply representation (3.30) to (3.29). Then we get a linear homogeneous system with respect to  $D_1, D_2, \dots, D_n$ :

$$\left[ \sum_{j=1}^{k_1} e_m^{B(k_1-m-j)} b b^T \left( e_m^{B(k_1-m-j)} \right)^T \right] D = \xi$$

or, in a matrix form,

$$GD = \xi.$$

By Lemma 3.4, the matrix  $G$  is nonsingular and it is clear that  $D = G^{-1}\xi$ . From (3.30) we get

$$u(j-1) = \left( e_m^{B(k_1-m-j)} b \right)^T G^{-1}\xi,$$

which is equivalent to (3.28).  $\square$

**Acknowledgments.** The authors would like to thank the referees for their helpful suggestions incorporated into this paper.

#### REFERENCES

- [1] J. BAŠTINEC AND J. DIBLÍK, *One case of appearance of positive solutions of delayed discrete equations*, Appl. Math., 48 (2003), pp. 429–436.
- [2] J. BAŠTINEC AND J. DIBLÍK, *Subdominant positive solutions of the discrete equation  $\Delta u(k+n) = -p(k)u(k)$* , Abstr. Appl. Anal., no. 6, 2004, pp. 461–470.
- [3] J. DIBLÍK AND D. YA. KHUSAINOV, *Representation of solutions of discrete delayed system  $x(k+1) = Ax(k) + Bx(k-m) + f(k)$  with commutative matrices*, J. Math. Anal. Appl., 318 (2006), pp. 63–76.
- [4] J. DIBLÍK AND D. YA. KHUSAINOV, *Representation of solutions of linear discrete systems with constant coefficients and pure delay*, Adv. Difference Equ., 2006, article 80825.
- [5] S. N. ELAYDI, *An Introduction to Difference Equations*, 3rd ed., Springer, New York, 2005.
- [6] R. GABASOV AND F. M. KIRILLOVA, *The Qualitative Theory of Optimal Processes*, Marcel Dekker, New York, Basel, 1976.
- [7] I. V. GAJSHUN, *Discrete Time Systems*, Institut Matematiki NAN Belarusi, Minsk, 2001 (in Russian).
- [8] F. P. GANTMACHER, *The Theory of Matrices, Vol. I*, AMS Chelsea Publishing, Providence, RI, 2002.
- [9] R. E. KALMAN, P. L. FALB, AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
- [10] J. KLAMKA, *Controllability of dynamical systems—a survey*, Arch. Control Sci., 2 (1993), pp. 283–310.
- [11] N. N. KRASOVSKII, *Theory of Control of Motion: Linear Systems*, Izdat. “Nauka,” Moscow, 1968 (in Russian).
- [12] K. N. MURTY, L. V. FAUSETT, AND Y. WU, *Fundamental theory of control of general first-order matrix difference systems*, Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal., 13 (2006), pp. 301–310.

# A PRIORI ERROR ESTIMATES FOR SPACE-TIME FINITE ELEMENT DISCRETIZATION OF PARABOLIC OPTIMAL CONTROL PROBLEMS PART I: PROBLEMS WITHOUT CONTROL CONSTRAINTS\*

DOMINIK MEIDNER<sup>†</sup> AND BORIS VEXLER<sup>‡</sup>

**Abstract.** In this paper we develop a priori error analysis for Galerkin finite element discretizations of optimal control problems governed by linear parabolic equations. The space discretization of the state variable is done using usual conforming finite elements, whereas the time discretization is based on discontinuous Galerkin methods. For different types of control discretizations we provide error estimates of optimal order with respect to both space and time discretization parameters. The paper is divided into two parts. In the first part we develop some stability and error estimates for space-time discretization of the state equation and provide error estimates for optimal control problems without control constraints. In the second part of the paper, the techniques and results of the first part are used to develop a priori error analysis for optimal control problems with pointwise inequality constraints on the control variable.

**Key words.** optimal control, parabolic equations, error estimates, finite elements

**AMS subject classifications.** 49N10, 49M25, 65M15, 65M60

**DOI.** 10.1137/070694016

**1. Introduction.** In this paper we develop a priori error analysis for space-time finite element discretizations of parabolic optimization problems. We consider the following linear-quadratic optimal control problem for the state variable  $u$  and the control variable  $q$ :

$$(1.1a) \quad \text{Minimize } J(q, u) = \frac{1}{2} \int_0^T \int_{\Omega} (u(t, x) - \hat{u}(t, x))^2 dx dt + \frac{\alpha}{2} \int_0^T \int_{\Omega} q(t, x)^2 dx dt$$

subject to

$$(1.1b) \quad \begin{aligned} \partial_t u - \Delta u &= f + q && \text{in } (0, T) \times \Omega, \\ u(0) &= u_0 && \text{in } \Omega, \end{aligned}$$

combined with either homogeneous Dirichlet or homogeneous Neumann boundary conditions on  $(0, T) \times \partial\Omega$ . A precise formulation of this problem including a functional analytic setting is given in the next section.

While the a priori error analysis for finite element discretization of optimal control problems governed by elliptic equations is discussed in many publications (see, e.g., [12, 15, 1, 16, 22, 4]), there are only a few published results on this topic for parabolic problems; see [20, 28, 17, 19, 24].

---

\*Received by the editors June 8, 2007; accepted for publication (in revised form) November 27, 2007; published electronically March 19, 2008. The first author was supported by the German research Foundation DFG through the International Research Training Group 710 “Complex Processes: Modeling, Simulation, and Optimization.” The second author was partially supported by the Austrian Science Fund FWF project P18971-N18 “Numerical analysis and discretization strategies for optimal control problems with singularities.”

<http://www.siam.org/journals/sicon/47-3/69401.html>

<sup>†</sup>Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, INF 294, 69120 Heidelberg, Germany (dominik.meidner@iwr.uni-heidelberg.de).

<sup>‡</sup>Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenberger Straße 69, 4040 Linz, Austria (boris.vexler@oeaw.ac.at).

In this paper, we will use discontinuous finite element methods for time discretization of the state equation (1.1b), as proposed, e.g., in [7, 10]. The spatial discretization will be based on usual  $H^1$ -conforming finite elements. In [2] this type of discretization is shown to allow for a natural translation of the optimality conditions from the continuous to the discrete level. This gives rise to exact computation of the derivatives required in the optimization algorithms on the discrete level. In [21] a posteriori error estimates for this type of discretization are derived and an adaptive algorithm is developed.

Throughout, we will use a general discretization parameter  $\sigma$  consisting of three discretization parameters  $\sigma = (k, h, d)$ , where  $k$  corresponds to the time discretization of the state variable,  $h$  to the space discretization of the state variable, and  $d$  to the discretization of the control variable  $q$ , respectively. The space and time discretizations of the control variable may differ from the discretizations of the state. Therefore, the discretization parameter  $d$  consists of the discretization parameters  $k_d$  and  $h_d$  for the time and space discretizations of the control variable. In this paper we will derive a priori error estimates of optimal order with respect to all discretization parameters, where the influences of different parts of the discretization are clearly separated. Moreover, the temporal and spatial regularity properties of the solution to the continuous problem (1.1) are separated as well.

For the discretization error between the solution of the continuous optimization problem  $(\bar{q}, \bar{u})$  and the solution of the discretized problem  $(\bar{q}_\sigma, \bar{u}_\sigma)$ , we will prove error estimates of the following structure:

$$(1.2) \quad \|\bar{q} - \bar{q}_\sigma\|_{L^2((0,T) \times \Omega)} \leq C_1(\bar{u}, \bar{z}) k^{r+1} + C_2(\bar{u}, \bar{z}) h^{s+1} + C_3(\bar{q}) k_d^{r_d+1} + C_4(\bar{q}) h_d^{s_d+1},$$

where  $r, r_d$  are the highest degrees of the polynomials in the time discretization of the state and the control variable, respectively, and  $s, s_d$  are the highest degree of the polynomials in the space discretization of the control and the state variable. The constants  $C_1(\bar{u}, \bar{z})$  and  $C_2(\bar{u}, \bar{z})$  depend on the temporal and the spatial regularity of the optimal state  $\bar{u}$  and the corresponding adjoint state  $\bar{z}$ , respectively; cf. Theorem 6.1. The temporal and spatial regularity of the optimal control  $\bar{q}$  determines the constants  $C_3(\bar{q})$  and  $C_4(\bar{q})$ , respectively.

In [19] a similar result is proved for the case  $r = 0, s = 1$ , and under the assumption  $k \approx h^2$ . We would like to emphasize that the discretization parameters  $k, h, k_d, h_d$  in estimate (1.2) can be chosen independently of each other.

The purpose of this paper is twofold. The first goal is to derive a priori error estimates for optimal control problem (1.1) of the above structure. The second goal is to provide techniques which will be used in the second part of the paper for derivation of a priori error estimates for problems involving pointwise inequality constraints on the control variable.

The paper is organized as follows. In the next section we recall the function analytic setting and optimality conditions for the optimal control problem under consideration. In section 3 the space-time finite element discretization is presented. Based on stability estimates developed in section 4, we provide a priori error analysis for the state equation in section 5. The main result on the error analysis for the considered optimal control problem is given in section 6. In this section error estimates for the error in the control, state, and adjoint variables are developed. In the last section we present a numerical example illustrating our results.

**2. Optimization.** In this section we briefly discuss the precise formulation of the optimization problem under consideration. Furthermore, we recall theoretical results

on existence, uniqueness, and regularity of optimal solutions as well as optimality conditions.

To set up a weak formulation of the state equation (1.1b), we introduce the following notation: For a convex polygonal domain  $\Omega \subset \mathbb{R}^n$ ,  $n \in \{2, 3\}$ , we denote  $V$  to be either  $H^1(\Omega)$  or  $H_0^1(\Omega)$  depending on the prescribed type of boundary conditions (homogeneous Neumann or homogeneous Dirichlet). Together with  $H = L^2(\Omega)$ , the Hilbert space  $V$  and its dual  $V^*$  build a Gelfand triple  $V \hookrightarrow H \hookrightarrow V^*$ . Here and in what follows, we employ the usual notion for Lebesgue and Sobolev spaces.

For a time interval  $I = (0, T)$  we introduce the state space

$$X := \{v \mid v \in L^2(I, V) \text{ and } \partial_t v \in L^2(I, V^*)\}$$

and the control space

$$Q = L^2(I, L^2(\Omega)).$$

In addition, we use the following notation for the inner products and norms on  $L^2(\Omega)$  and  $L^2(I, L^2(\Omega))$ :

$$\begin{aligned} (v, w) &:= (v, w)_{L^2(\Omega)}, & (v, w)_I &:= (v, w)_{L^2(I, L^2(\Omega))}, \\ \|v\| &:= \|v\|_{L^2(\Omega)}, & \|v\|_I &:= \|v\|_{L^2(I, L^2(\Omega))}. \end{aligned}$$

In this setting, a standard weak formulation of the state equation (1.1b) for given control  $q \in Q$ ,  $f \in L^2(I, H)$ , and  $u_0 \in V$  reads as follows: Find a state  $u \in X$  satisfying

$$\begin{aligned} (2.1) \quad & (\partial_t u, \varphi)_I + (\nabla u, \nabla \varphi)_I = (f + q, \varphi)_I \quad \forall \varphi \in X, \\ & u(0) = u_0. \end{aligned}$$

For simplicity of notation, we skip here and throughout the paper the dependence of the solution variable on  $x$  and  $t$ .

For this formulation of the state equation, we recall the following result on existence and regularity.

**PROPOSITION 2.1.** *For fixed control  $q \in Q$ ,  $f \in L^2(I, H)$ , and  $u_0 \in V$  there exists a unique solution  $u \in X$  of problem (2.1). Moreover, the solution exhibits the improved regularity*

$$u \in L^2(I, H^2(\Omega) \cap V) \cap H^1(I, L^2(\Omega)) \hookrightarrow C(\bar{I}, V).$$

*It holds the stability estimate*

$$\|\partial_t u\|_I + \|\nabla^2 u\|_I \leq C\{\|f + q\|_I + \|\nabla u_0\|\}.$$

*Proof.* The proof of existence and uniqueness is given, e.g., in [18] and [29]. The improved regularity relies on the fact that  $\Omega$  is polygonal and convex and is proved, e.g., in [11]. The embedding of  $L^2(I, H^2(\Omega) \cap V) \cap H^1(I, L^2(\Omega))$  into  $C(\bar{I}, V)$  can be found, for instance, in [6].  $\square$

The weak formulation of the optimal control problem (1.1) is given as follows:

$$(2.2) \quad \text{Minimize } J(q, u) := \frac{1}{2}\|u - \hat{u}\|_I^2 + \frac{\alpha}{2}\|q\|_I^2 \text{ subject to (2.1) and } (q, u) \in Q \times X,$$

where  $\hat{u} \in L^2(I, H)$  is a given desired state and  $\alpha > 0$  is the regularization parameter.

PROPOSITION 2.2. *For given  $f, \hat{u} \in L^2(I, H)$ ,  $u_0 \in V$ , and  $\alpha > 0$ , the optimal control problem (2.2) admits a unique solution  $(\bar{q}, \bar{u}) \in Q \times X$ . The optimal control  $\bar{q}$  possesses the regularity*

$$\bar{q} \in L^2(I, H^2(\Omega)) \cap H^1(I, L^2(\Omega)).$$

*Proof.* For existence and uniqueness we refer to [18]. First order necessary optimality conditions and Proposition 2.1 imply the stated regularity of the optimal control.  $\square$

The existence result for the state equation in Proposition 2.1 ensures the existence of a control-to-state mapping  $q \mapsto u = u(q)$  defined through (2.1). By means of this mapping we introduce the reduced cost functional  $j: Q \rightarrow \mathbb{R}$ :

$$j(q) := J(q, u(q)).$$

The optimal control problem (2.2) can then be equivalently reformulated as follows:

$$(2.3) \quad \text{Minimize } j(q) \text{ subject to } q \in Q.$$

The first order necessary optimality condition for (2.3) reads as

$$(2.4) \quad j'(\bar{q})(\delta q) = 0 \quad \forall \delta q \in Q.$$

Due to the linear-quadratic structure of the optimal control problem this condition is also sufficient for optimality.

Utilizing the adjoint state equation for  $z = z(q) \in X$  given by

$$(2.5) \quad \begin{aligned} -(\varphi, \partial_t z)_I + (\nabla \varphi, \nabla z)_I &= (\varphi, u(q) - \hat{u})_I \quad \forall \varphi \in X, \\ z(T) &= 0, \end{aligned}$$

the first derivative of the reduced cost functional can be expressed as

$$(2.6) \quad j'(q)(\delta q) = (\alpha q + z(q), \delta q)_I.$$

**3. Discretization.** In this section we describe the space-time finite element discretization of the optimal control problem (2.2).

**3.1. Semidiscretization in time.** At first, we present the semidiscretization in time of the state equation by discontinuous Galerkin methods. We consider a partitioning of the time interval  $\bar{I} = [0, T]$  as

$$(3.1) \quad \bar{I} = \{0\} \cup I_1 \cup I_2 \cup \cdots \cup I_M$$

with subintervals  $I_m = (t_{m-1}, t_m]$  of size  $k_m$  and time points

$$0 = t_0 < t_1 < \cdots < t_{M-1} < t_M = T.$$

We define the discretization parameter  $k$  as a piecewise constant function by setting  $k|_{I_m} = k_m$  for  $m = 1, 2, \dots, M$ . Moreover, we denote by  $k$  the maximal size of the time steps, i.e.,  $k = \max k_m$ .

The semidiscrete trial and test space is given as

$$X_k^r = \left\{ v_k \in L^2(I, V) \mid v_k|_{I_m} \in \mathcal{P}_r(I_m, V), \ m = 1, 2, \dots, M \right\}.$$

Here,  $\mathcal{P}_r(I_m, V)$  denotes the space of polynomials up to order  $r$  defined on  $I_m$  with values in  $V$ . On  $X_k^r$  we use the notation

$$(v, w)_{I_m} := (v, w)_{L^2(I_m, L^2(\Omega))} \quad \text{and} \quad \|v\|_{I_m} := \|v\|_{L^2(I_m, L^2(\Omega))}.$$

To define the discontinuous Galerkin (dG( $r$ )) approximation using the space  $X_k^r$  we employ the following definitions for functions  $v_k \in X_k^r$ :

$$v_{k,m}^+ := \lim_{t \rightarrow 0^+} v_k(t_m + t), \quad v_{k,m}^- := \lim_{t \rightarrow 0^+} v_k(t_m - t) = v_k(t_m), \quad [v_k]_m := v_{k,m}^+ - v_{k,m}^-$$

and define the bilinear form  $B(\cdot, \cdot)$  for  $u_k, \varphi \in X_k^r$  by

$$(3.2) \quad B(u_k, \varphi) := \sum_{m=1}^M (\partial_t u_k, \varphi)_{I_m} + (\nabla u_k, \nabla \varphi)_I + \sum_{m=2}^M ([u_k]_{m-1}, \varphi_{m-1}^+) + (u_{k,0}^+, \varphi_0^+).$$

Then, the dG( $r$ ) semidiscretization of the state equation (2.1) for a given control  $q \in Q$  reads as follows: Find a state  $u_k = u_k(q) \in X_k^r$  such that

$$(3.3) \quad B(u_k, \varphi) = (f + q, \varphi)_I + (u_0, \varphi_0^+) \quad \forall \varphi \in X_k^r.$$

The existence and uniqueness of solutions to (3.3) can be shown by using Fourier analysis; see [27] for details.

*Remark 3.1.* Using a density argument, it is possible to show that the exact solution  $u = u(q) \in X$  also satisfies the identity

$$B(u, \varphi) = (f + q, \varphi)_I + (u_0, \varphi_0^+) \quad \forall \varphi \in X_k^r.$$

Thus, we have here the property of Galerkin orthogonality

$$B(u - u_k, \varphi) = 0 \quad \forall \varphi \in X_k^r,$$

although the dG( $r$ ) semidiscretization is a nonconforming Galerkin method ( $X_k^r \not\subset X$ ).

The semidiscrete optimization problem for the dG( $r$ ) time discretization has the following form:

$$(3.4) \quad \text{Minimize } J(q_k, u_k) \text{ subject to (3.3) and } (q_k, u_k) \in Q \times X_k^r.$$

**PROPOSITION 3.2.** *The semidiscrete optimal control problem (3.4) admits for  $\alpha > 0$  a unique solution  $(\bar{q}_k, \bar{u}_k) \in Q \times X_k^r$ .*

*Proof.* The proof is done by translating standard arguments from the proof in the continuous case and by employing the continuity of the mapping  $q \mapsto u_k(q)$  provided by the stability estimates derived in the next section (cf. Theorem 4.3).  $\square$

Note that the optimal control  $\bar{q}_k$  is searched for in the continuous space  $Q$  and the subscript  $k$  indicates the usage of the semidiscretized state equation.

Similar to the continuous case, we introduce the semidiscrete reduced cost functional  $j_k: Q \rightarrow \mathbb{R}$ :

$$j_k(q) := J(q, u_k(q))$$

and reformulate the semidiscrete optimal control problem (3.4) as follows:

$$\text{Minimize } j_k(q_k) \text{ subject to } q_k \in Q.$$

The first order necessary optimality condition reads as

$$(3.5) \quad j'_k(\bar{q}_k)(\delta q) = 0 \quad \forall \delta q \in Q,$$

and the derivative of  $j_k$  can be expressed as

$$(3.6) \quad j'_k(q)(\delta q) = (\alpha q + z_k(q), \delta q)_I.$$

Here,  $z_k = z_k(q) \in X_k^r$  denotes the solution of the semidiscrete adjoint equation

$$(3.7) \quad B(\varphi, z_k) = (\varphi, u_k(q) - \hat{u})_I \quad \forall \varphi \in X_k^r.$$

Note that by using integration by parts in time, the bilinear form  $B(\cdot, \cdot)$  defined by (3.2) can equivalently be expressed as

$$(3.8) \quad B(\varphi, z_k) = - \sum_{m=1}^M (\varphi, \partial_t z_k)_{I_m} + (\nabla \varphi, \nabla z_k)_I - \sum_{m=1}^{M-1} (\varphi_m^-, [z_k]_m) + (\varphi_M^-, z_{k,M}^-).$$

**3.2. Discretization in space.** To define the finite element discretization in space, we consider two or three dimensional shape-regular meshes; see, e.g., [5]. A mesh consists of quadrilateral or hexahedral cells  $K$ , which constitute a non-overlapping cover of the computational domain  $\Omega$ . The corresponding mesh is denoted by  $\mathcal{T}_h = \{K\}$ , where we define the discretization parameter  $h$  as a cellwise constant function by setting  $h|_K = h_K$  with the diameter  $h_K$  of the cell  $K$ . We use the symbol  $h$  also for the maximal cell size, i.e.,  $h = \max h_K$ .

On the mesh  $\mathcal{T}_h$  we construct a conforming finite element space  $V_h \subset V$  in a standard way:

$$V_h^s = \{v \in V \mid v|_K \in \mathcal{Q}_s(K) \text{ for } K \in \mathcal{T}_h\}.$$

Here,  $\mathcal{Q}_s(K)$  consists of shape functions obtained via (bi-/tri-)linear transformations of polynomials in  $\widehat{\mathcal{Q}}_s(\widehat{K})$  defined on the reference cell  $\widehat{K} = (0, 1)^n$ , where

$$\widehat{\mathcal{Q}}_s(\widehat{K}) = \text{span} \left\{ \prod_{j=1}^n x_j^{\alpha_j} \mid \alpha_j \in \mathbb{N}_0, \alpha_j \leq s \right\}.$$

*Remark 3.3.* The definition of  $V_h^s$  can be extended to the case of triangular meshes in the obvious way.

To obtain the fully discretized versions of the time discretized state equation (3.3), we utilize the space-time finite element space

$$X_{k,h}^{r,s} = \left\{ v_{kh} \in L^2(I, V_h^s) \mid v_{kh}|_{I_m} \in \mathcal{P}_r(I_m, V_h^s) \right\} \subset X_k^r.$$

*Remark 3.4.* Here, the spatial mesh and, therefore, also the space  $V_h^s$  is fixed for all time intervals. We refer to [25] for a discussion of treatment of different meshes  $\mathcal{T}_h^m$  for each of the subintervals  $I_m$ .

The so-called cG( $s$ )dG( $r$ ) discretization of the state equation for given control  $q \in Q$  has the following form: Find a state  $u_{kh} = u_{kh}(q) \in X_{k,h}^{r,s}$  such that

$$(3.9) \quad B(u_{kh}, \varphi) = (f + q, \varphi)_I + (u_0, \varphi_0^+) \quad \forall \varphi \in X_{k,h}^{r,s}.$$



*Remark 3.5.* The notation  $\text{cG}(s)\text{dG}(r)$  is taken from [7] and describes a method with conforming (continuous) discretization in space of order  $s$  and discontinuous discretization in time of order  $r$ .

Then, the corresponding optimal control problem is given as follows:

$$(3.10) \quad \text{Minimize } J(q_{kh}, u_{kh}) \text{ subject to (3.9) and } (q_{kh}, u_{kh}) \in Q \times X_{k,h}^{r,s},$$

and by means of the discrete reduced cost functional  $j_{kh}: Q \rightarrow \mathbb{R}$ ,

$$j_{kh}(q) := J(q, u_{kh}(q)),$$

it can be reformulated as follows:

$$\text{Minimize } j_{kh}(q_{kh}) \text{ subject to } q_{kh} \in Q.$$

The uniquely determined optimal solution of (3.10) is denoted by  $(\bar{q}_{kh}, \bar{u}_{kh}) \in Q \times X_{k,h}^{r,s}$ .

The optimal control  $\bar{q}_{kh} \in Q$  fulfills the first order optimality condition

$$(3.11) \quad j'_{kh}(\bar{q}_{kh})(\delta q) = 0 \quad \forall \delta q \in Q,$$

where  $j'_{kh}(q)(\delta q)$  is given by

$$(3.12) \quad j'_{kh}(q)(\delta q) = (\alpha q + z_{kh}(q), \delta q)_I$$

with the discrete adjoint solution  $z_{kh} = z_{kh}(q) \in X_{k,h}^{r,s}$  of

$$(3.13) \quad B(\varphi, z_{kh}) = (\varphi, u_{kh}(q) - \hat{u})_I \quad \forall \varphi \in X_{k,h}^{r,s}.$$

**3.3. Discretization of the controls.** To obtain the fully discrete optimal control problem we restrict the control space  $Q$  to a finite dimensional subspace  $Q_d \subset Q$ . The optimal control problem on this level of discretization is given as follows:

$$(3.14) \quad \text{Minimize } J(q_\sigma, u_\sigma) \text{ subject to (3.9) and } (q_\sigma, u_\sigma) \in Q_d \times X_{k,h}^{r,s}.$$

The unique optimal solution of (3.14) is denoted by  $(\bar{q}_\sigma, \bar{u}_\sigma) \in Q_d \times X_{k,h}^{r,s}$ , where the subscript  $\sigma$  collects the discretization parameters  $k$ ,  $h$ , and  $d$ . The optimality condition is given using the discrete reduced cost functional  $j_{kh}$  introduced in the previous section by

$$(3.15) \quad j'_{kh}(\bar{q}_\sigma)(\delta q) = 0 \quad \forall \delta q \in Q_d.$$

Most of our results presented below hold true independently of the choice of the control discretization; see Theorem 6.1. However, we present here some possibilities for construction of the discrete control space  $Q_d$ , which will play a role in the discussion of the error in the state and adjoint variables (see section 6.2) and which will be employed for the numerical example in section 7.

For the construction of  $Q_d$  it is possible to use spatial and temporal meshes, which are different from those employed for the discretization of the state variable. However, for simplicity of notation we will use the same time-partitioning (3.1). Using a spatial mesh  $\mathcal{T}_{h_d}$  we consider two corresponding finite element spaces:

$$V_{h_d}^{s_d} = \{v \in C(\bar{\Omega}) \mid v|_K \in \mathcal{Q}_{s_d}(K) \text{ for } K \in \mathcal{T}_{h_d}\}$$

and

$$\tilde{V}_{h_d}^{s_d} = \{v \in L^2(\Omega) \mid v|_K \in \mathcal{Q}_{s_d}(K) \text{ for } K \in \mathcal{T}_{h_d}\}.$$

The space  $V_{h_d}^{s_d}$  consists of continuous cellwise polynomial functions of order  $s_d$ , whereas the functions in the space  $\tilde{V}_{h_d}^{s_d}$  are discontinuous. Using these spaces we define two possibilities for the choice of  $Q_d$ .

The first possibility is similar to the construction of the state space  $X_{k,h}^{r,s}$ , consisting of functions which are continuous in space and discontinuous in time, and results in the following definition:

$$Q_d = \left\{ v_{kh} \in L^2(I, V_{h_d}^{s_d}) \mid v_{kh}|_{I_m} \in \mathcal{P}_{r_d}(I_m, V_{h_d}^{s_d}) \right\}.$$

We will refer to this control discretization as  $\text{cG}(s_d)\text{dG}(r_d)$ . If the control mesh  $\mathcal{T}_{h_d}$  coincides with the state mesh  $\mathcal{T}_h$  and one chooses the same order of polynomials ( $r = r_d, s = s_d$ ), then the state space  $X_{k,h}^{r,s}$  coincides with the control space  $Q_d$  in case of homogeneous Neumann boundary conditions and is a subspace of it, i.e.,  $X_{k,h}^{r,s} \subset Q_d$  in the presence of homogeneous Dirichlet boundary conditions. In this case one can show (cf. the discussion in section 6) that  $\bar{q}_{kh} = \bar{q}_\sigma$ . This means that a complete discretization of the optimal control problem is achieved already after discretization of the state equation; cf. [16].

For the second possibility we employ the space  $\tilde{V}_{h_d}^{s_d}$  of discontinuous cellwise polynomials and obtain the following definition:

$$Q_d = \left\{ v_{kh} \in L^2(I, \tilde{V}_{h_d}^{s_d}) \mid v_{kh}|_{I_m} \in \mathcal{P}_{r_d}(I_m, \tilde{V}_{h_d}^{s_d}) \right\}.$$

We will refer to this control discretization as  $\text{dG}(s_d)\text{dG}(r_d)$ . The special choice  $s_d = 0$  leads to cellwise constant discretization in space.

**4. Stability estimates for the state and adjoint equations.** The first step in proving the desired a priori estimate is to prove stability estimates for the solution of the semidiscrete (3.3) and the fully discretized (3.9) state equation. Throughout this section we discuss the uncontrolled situation and set therefore  $q = 0$ .

In the following theorem we provide a stability estimate for semidiscretization in time, which has a structure similar to the estimate on the continuous level given in Proposition 2.1. A comparable estimate is shown in [8, 9] for the case  $f = 0$ .

**THEOREM 4.1.** *For the solution  $u_k \in X_k^r$  of the  $\text{dG}(r)$  semidiscretized state equation (3.3) with right-hand side  $f \in L^2(I, H)$ , initial condition  $u_0 \in V$ , and  $q = 0$ , the stability estimate*

$$\sum_{m=1}^M \|\partial_t u_k\|_{I_m}^2 + \|\Delta u_k\|_I^2 + \sum_{m=1}^M k_m^{-1} \|[u_k]_{m-1}\|^2 \leq C \{ \|f\|_I^2 + \|\nabla u_0\|^2 \}$$

*holds. The constant  $C$  depends only on the polynomial degree  $r$  and the domain  $\Omega$ . The jump term  $[u_k]_0$  at  $t = 0$  is defined as  $u_{k,0}^+ - u_0$ .*

*Proof.* By means of the definition  $[u_k]_0 = u_{k,0}^+ - u_0$ , the solution  $u_k \in X_k^r$  of (3.3) fulfills for all  $\varphi \in \mathcal{P}_r(I_m, V)$  the following system of equations:

$$(4.1) \quad (\partial_t u_k, \varphi)_{I_m} + (\nabla u_k, \nabla \varphi)_{I_m} + ([u_k]_{m-1}, \varphi_{m-1}^+) = (f, \varphi)_{I_m}, \quad m = 1, 2, \dots, M.$$

The proof of the desired estimate consist of three steps—one for each term of the left-hand side of (4.1). The steps are based on consecutively testing with  $\varphi = -\Delta u_k$ ,  $\varphi = (t - t_{m-1})\partial_t u_k$ , and  $\varphi = [u_k]_{m-1}$ .

*Step* (i). At first, we want to choose  $\varphi = -\Delta u_k$ . For applying integration by parts in space to (4.1), it is necessary to prove  $\Delta u_k|_{I_m} \in \mathcal{P}_r(I_m, H)$ . This assertion follows immediately from applying elliptic regularity theory (cf. [11]) to the transformed time stepping equation

$$(\nabla u_k, \nabla \varphi)_{I_m} = (f - \partial_t u_k, \varphi)_{I_m} - ([u_k]_{m-1}, \varphi_{m-1}^+).$$

The fact that  $u_k|_{I_m}$  is polynomial in time with values in  $V \subseteq H$  implies that the right-hand side is in  $H$  for almost all  $t \in I_m$ . Thus,  $\Delta u_k|_{I_m}$  is also in  $H$  for almost all  $t \in I_m$ , and since  $u_k|_{I_m}$  is polynomial with respect to time, this yields  $\Delta u_k|_{I_m} \in \mathcal{P}_r(I_m, H)$ .

Consequently, it is feasible to integrate (4.1) by parts in space to obtain the formulation

$$(4.2) \quad (\partial_t u_k, \varphi)_{I_m} - (\Delta u_k, \varphi)_{I_m} + ([u_k]_{m-1}, \varphi_{m-1}^+) = (f, \varphi)_{I_m}, \quad m = 1, 2, \dots, M.$$

The arising boundary terms vanish for both homogeneous Neumann and homogeneous Dirichlet boundary conditions.

Since there are no spatial derivatives on the test function  $\varphi$  anymore, formulation (4.2) holds not only for all  $\varphi \in \mathcal{P}_r(I_m, V)$  but by the density of  $V$  in  $H$  also for all  $\varphi \in \mathcal{P}_r(I_m, H)$ . Hence, we may choose  $\varphi = -\Delta u_k$  as a test function and get, by applying integration by parts in space a second time,

$$(\partial_t \nabla u_k, \nabla u_k)_{I_m} + (\Delta u_k, \Delta u_k)_{I_m} + ([\nabla u_k]_{m-1}, \nabla u_{k,m-1}^+) = (f, -\Delta u_k)_{I_m}.$$

Again, the arising boundary terms vanish due to the prescribed homogeneous boundary conditions of Neumann or Dirichlet type.

By means of the identities

$$(4.3a) \quad (\partial_t v, v)_{I_m} = \frac{1}{2} \|v_m^-\|^2 - \frac{1}{2} \|v_{m-1}^+\|^2,$$

$$(4.3b) \quad ([v]_{m-1}, v_{m-1}^+) = \frac{1}{2} \|v_{m-1}^+\|^2 + \frac{1}{2} \|[v]_{m-1}\|^2 - \frac{1}{2} \|v_{m-1}^-\|^2,$$

we achieve

$$\frac{1}{2} \|\nabla u_{k,m}^-\|^2 + \frac{1}{2} \|[\nabla u_k]_{m-1}\|^2 - \frac{1}{2} \|\nabla u_{k,m-1}^-\|^2 + \|\Delta u_k\|_{I_m}^2 = (f, -\Delta u_k)_{I_m}.$$

Summation of the equations for  $m = 1, 2, \dots, M$  leads to

$$\frac{1}{2} \|\nabla u_{k,M}^-\|^2 + \frac{1}{2} \sum_{m=1}^M \|[\nabla u_k]_{m-1}\|^2 + \|\Delta u_k\|_I^2 = (f, -\Delta u_k)_I + \frac{1}{2} \|\nabla u_0\|^2.$$

Using Young's inequality for the right-hand side, we obtain the first intermediary result

$$(4.4) \quad \|\Delta u_k\|_I^2 \leq \|f\|_I^2 + \|\nabla u_0\|^2.$$

Step (ii). To bound the time derivative  $\partial_t u_k$ , we will use the inverse estimate

$$(4.5) \quad \|v_k\|_{I_m}^2 \leq C k_m^{-1} \int_{I_m} (t - t_{m-1}) \|v_k\|^2 dt,$$

which holds true for all  $v_k \in \mathcal{P}_r(I_m, V)$ . To obtain this estimate one transforms both sides of the inequality into the reference time interval  $[0, 1]$ , uses equivalence of norms for finite dimensional spaces, and transforms the inequality back into the real interval  $I_m$ .

We choose  $\varphi = (t - t_{m-1}) \partial_t u_k$  and obtain, utilizing the fact that  $\varphi_{m-1}^+ = 0$ ,

$$\begin{aligned} \int_{I_m} (t - t_{m-1}) \|\partial_t u_k\|^2 dt &= \int_{I_m} (t - t_{m-1}) (f + \Delta u_k, \partial_t u_k) dt \\ &\leq \left( \int_{I_m} (t - t_{m-1}) \|f + \Delta u_k\|^2 dt \right)^{\frac{1}{2}} \left( \int_{I_m} (t - t_{m-1}) \|\partial_t u_k\|^2 dt \right)^{\frac{1}{2}}. \end{aligned}$$

The inverse estimate (4.5) yields, by means of Hölder's inequality,

$$\|\partial_t u_k\|_{I_m}^2 \leq C k_m^{-1} \int_{I_m} (t - t_{m-1}) \|f + \Delta u_k\|^2 dt \leq C \{ \|f\|_{I_m}^2 + \|\Delta u_k\|_{I_m}^2 \}.$$

Then, (4.4) implies the second intermediary result

$$(4.6) \quad \sum_{m=1}^M \|\partial_t u_k\|_{I_m}^2 \leq C \{ \|f\|_I^2 + \|\nabla u_0\|^2 \}.$$

Step (iii). It remains to estimate the jump terms. To this end, we choose  $\varphi = [u_k]_{m-1}$  constant in time on  $I_m$  and obtain

$$\begin{aligned} \|[u_k]_{m-1}\|^2 &= (f + \Delta u_k - \partial_t u_k, [u_k]_{m-1})_{I_m} \\ &\leq \frac{k_m}{2} \|f + \Delta u_k - \partial_t u_k\|_{I_m}^2 + \frac{1}{2k_m} \|[u_k]_{m-1}\|_{I_m}^2. \end{aligned}$$

Since  $[u_k]_{m-1}$  is constant in time, we have  $\|[u_k]_{m-1}\|_{I_m}^2 = k_m \|[u_k]_{m-1}\|^2$ . This implies

$$k_m^{-1} \|[u_k]_{m-1}\|^2 \leq \|f + \Delta u_k - \partial_t u_k\|_{I_m}^2.$$

The results (4.4) and (4.6) yield the remaining estimate

$$\sum_{m=1}^M k_m^{-1} \|[u_k]_{m-1}\|^2 \leq C \{ \|f\|_I^2 + \|\nabla u_0\|^2 \}. \quad \square$$

The result of the previous theorem will also be applied to dual (adjoint) equations. Let  $g \in L^2(I, H)$  be a given right-hand side and  $z_T \in V$  a given terminal condition; then the corresponding semidiscretized dual equation is given by

$$(4.7) \quad B(\varphi, z_k) = (\varphi, g)_I + (\varphi_M^-, z_T) \quad \forall \varphi \in X_k^r.$$

Note that the semidiscrete adjoint solution defined in (3.7) can be obtained by setting  $g = u_k(q) - \hat{u}$  and  $z_T = 0$ .

**COROLLARY 4.2.** *For the solution  $z_k \in X_k^r$  of the semidiscrete dual equation (4.7) with right-hand side  $g \in L^2(I, H)$  and terminal condition  $z_T \in V$ , the estimate from Theorem 4.1 reads as*

$$\sum_{m=1}^M \|\partial_t z_k\|_{I_m}^2 + \|\Delta z_k\|_I^2 + \sum_{m=1}^M k_m^{-1} \|[z_k]_m\|^2 \leq C \{\|g\|_I^2 + \|\nabla z_T\|^2\}.$$

Here, the jump term  $[z_k]_M$  at  $t = T$  is defined as  $z_T - z_{k,M}^-$ .

*Proof.* Let  $z_k \in X_k^r$  be the solution of (4.7). Then formula (3.8) implies that it also fulfills for all  $\varphi \in \mathcal{P}_r(I_m, V)$  the following system of equations:

$$-(\varphi, \partial_t z_k)_{I_m} + (\nabla \varphi, \nabla z_k)_{I_m} - (\varphi_m^-, [z_k]_m) = (g, \varphi)_{I_m}, \quad m = 1, 2, \dots, M.$$

Based on this representation, all steps of the proof of Theorem 4.1 can be repeated similarly to obtain the stated result.  $\square$

For proving a priori estimates for the control problem (2.2), we will additionally need stability estimates for the  $L^2(I, H)$ -norm of the solution  $\|u_k\|_I$  and of its gradient  $\|\nabla u_k\|_I$ , which are given in the following theorem.

**THEOREM 4.3.** *For the solution  $u_k \in X_k^r$  of the dG(r) semidiscretized state equation (3.3) with right-hand side  $f \in L^2(I, H)$ , initial condition  $u_0 \in V$ , and  $q = 0$ , the stability estimate*

$$\|u_k\|_I^2 + \|\nabla u_k\|_I^2 \leq C \{\|f\|_I^2 + \|\nabla u_0\|^2 + \|u_0\|^2\}$$

holds true with a constant  $C$  that depends only on the polynomial degree  $r$ , the domain  $\Omega$ , and the final time  $T$ .

*Remark 4.4.* In the case of homogeneous Dirichlet boundary conditions, the estimate can be proved by means of Poincaré's inequality with a constant independent of  $T$ .

*Proof.* The proof is done using a duality argument: Let  $\tilde{z} \in X$  be the solution of

$$-(\varphi, \partial_t \tilde{z})_I + (\nabla \varphi, \nabla \tilde{z})_I = (\varphi, u_k)_I \quad \forall \varphi \in X$$

together with the terminal condition  $\tilde{z}(T) = \tilde{z}_T = 0$ . Thus, due to Remark 3.1, which applies similarly to dual or adjoint equations,  $\tilde{z}$  also fulfills

$$B(\varphi, \tilde{z}) = (\varphi, u_k)_I \quad \forall \varphi \in X_k^r.$$

By means of this equality, we write

$$\begin{aligned} \|u_k\|_I^2 &= B(u_k, \tilde{z}) \\ &= \sum_{m=1}^M (\partial_t u_k, \tilde{z})_{I_m} + (\nabla u_k, \nabla \tilde{z})_I + \sum_{m=2}^M ([u_k]_{m-1}, \tilde{z}(t_{m-1})) + (u_{k,0}^+, \tilde{z}(0)). \end{aligned}$$

Using the setting  $[u_k]_0 = u_{k,0}^+ - u_0$  we obtain

$$\|u_k\|_I^2 = \sum_{m=1}^M (\partial_t u_k, \tilde{z})_{I_m} + (\nabla u_k, \nabla \tilde{z})_I + \sum_{m=1}^M ([u_k]_{m-1}, \tilde{z}(t_{m-1})) + (u_0, \tilde{z}(0)),$$

from which we get, with integration by parts in space and Hölder's inequality,

$$\begin{aligned} \|u_k\|_I^2 &\leq \left( \sum_{m=1}^M \|\partial_t u_k\|_{I_m}^2 \right)^{\frac{1}{2}} \|\tilde{z}\|_I + \|\Delta u_k\|_I \|\tilde{z}\|_I \\ &\quad + \left( \sum_{m=1}^M k_m^{-1} \|[u_k]_{m-1}\|^2 \right)^{\frac{1}{2}} \left( \sum_{m=1}^M k_m \|\tilde{z}(t_{m-1})\|^2 \right)^{\frac{1}{2}} + \|u_0\| \|\tilde{z}(0)\|. \end{aligned}$$

The stability estimate for the continuous solution  $\tilde{z} \in X$

$$\max_{t \in \bar{I}} \|\tilde{z}(t)\| \leq C \|u_k\|_I,$$

which makes use of the continuity of the mapping  $u_k \mapsto \tilde{z}$  (cf. [18]) and the continuous embedding of  $X$  into  $C(\bar{I}, H)$ , implies

$$\begin{aligned} \|u_k\|_I &\leq C\sqrt{T} \left( \sum_{m=1}^M \|\partial_t u_k\|_{I_m}^2 \right)^{\frac{1}{2}} + C\sqrt{T} \|\Delta u_k\|_I \\ &\quad + C\sqrt{T} \left( \sum_{m=2}^M k_m^{-1} \|[u_k]_{m-1}\|^2 \right)^{\frac{1}{2}} + C\|u_0\|, \end{aligned}$$

from which the desired estimate for  $\|u_k\|_I^2$  follows by application of Theorem 4.1.

To prove the estimate for  $\|\nabla u_k\|_I^2$ , we proceed similarly to the proof of Theorem 4.1 and test (4.1) with  $\varphi = u_k$ . We obtain for  $m = 1, 2, \dots, M$

$$(\partial_t u_k, u_k)_{I_m} + (\nabla u_k, \nabla u_k)_{I_m} + ([u_k]_{m-1}, u_{k,m-1}^+) = (f, u_k)_{I_m}.$$

The identities (4.3) lead to

$$\frac{1}{2} \|u_{k,m}^-\|^2 + \frac{1}{2} \|[u_k]_{m-1}\|^2 - \frac{1}{2} \|u_{k,m-1}^-\|^2 + \|\nabla u_k\|_{I_m}^2 = (f, u_k)_{I_m}.$$

After summing up these equations for  $m = 1, 2, \dots, M$  and by application of Young's inequality, we have

$$\|\nabla u_k\|_I^2 \leq \frac{1}{2} \{ \|f\|_I^2 + \|u_k\|_I^2 + \|u_0\|^2 \}.$$

Insertion of the already proved estimate for  $\|u_k\|_I^2$  completes the proof.  $\square$

**COROLLARY 4.5.** *For the solution  $z_k \in X_k^r$  of the semidiscrete dual equation (4.7) with right-hand side  $g \in L^2(I, H)$  and terminal condition  $z_T \in V$ , the estimate from Theorem 4.3 reads as*

$$\|z_k\|_I^2 + \|\nabla z_k\|_I^2 \leq C \{ \|g\|_I^2 + \|\nabla z_T\|^2 + \|z_T\|^2 \}.$$

*Proof.* The proof is done similarly to the proof of Theorem 4.3.  $\square$

All the estimates proved in this section hold true also for the fully discrete cG(s)dG(r) solutions  $u_{kh}$ ,  $z_{kh} \in X_{k,h}^{r,s}$  almost without any changes. Only two differences have to be regarded: We have to replace the continuous Laplacian  $\Delta$  by a discrete analogue  $\Delta_h: V_h^s \rightarrow V_h^s$  defined by

$$(\Delta_h u, \varphi) = -(\nabla u, \nabla \varphi) \quad \forall \varphi \in V_h^s,$$

and the jump terms  $[u_{kh}]_0$  and  $[z_{kh}]_M$  are given here by means of the spatial  $L^2$ -projection  $\Pi_h: V \rightarrow V_h^s$  as

$$[u_{kh}]_0 = u_{kh,0}^+ - \Pi_h u_0 \quad \text{and} \quad [z_{kh}]_M = \Pi_h z_T - z_{kh,M}^-.$$

Here,  $z_{kh} \in X_{k,h}^{r,s}$  is the solution of the fully discretized dual equation with given right-hand side  $g \in L^2(I, H)$  and terminal condition  $z_T \in V$  given by

$$(4.8) \quad B(\varphi, z_{kh}) = (\varphi, g)_I + (\varphi_M^-, z_T) \quad \forall \varphi \in X_{k,h}^{r,s}.$$

For convenience of the reader, we state here the estimates for the fully discrete solutions.

**THEOREM 4.6.** *For the solution  $u_{kh} \in X_{k,h}^{r,s}$  of the discrete state equation (3.9) with right-hand side  $f \in L^2(I, H)$ , initial condition  $u_0 \in V$ , and  $q = 0$ , the stability estimate*

$$\sum_{m=1}^M \|\partial_t u_{kh}\|_{I_m}^2 + \|\Delta_h u_{kh}\|_I^2 + \sum_{m=1}^M k_m^{-1} \|[u_{kh}]_{m-1}\|^2 \leq C \{ \|f\|_I^2 + \|\nabla \Pi_h u_0\|^2 \}$$

*holds. The constant  $C$  depends only on the polynomial degree  $r$  and the domain  $\Omega$ . The jump term  $[u_{kh}]_0$  at  $t = 0$  is defined as  $u_{kh,0}^+ - \Pi_h u_0$ . Furthermore, the estimate*

$$\|u_{kh}\|_I^2 + \|\nabla u_{kh}\|_I^2 \leq C \{ \|f\|_I^2 + \|\nabla \Pi_h u_0\|^2 + \|\Pi_h u_0\|^2 \}$$

*holds true with a constant  $C$  that depends only on the polynomial degree  $r$ , the domain  $\Omega$ , and the final time  $T$ .*

**COROLLARY 4.7.** *For the solution  $z_{kh} \in X_{k,h}^{r,s}$  of the discrete dual equation (4.8) with right-hand side  $g \in L^2(I, H)$  and terminal condition  $z_T \in V$ , the estimates from Theorem 4.6 read as*

$$\sum_{m=1}^M \|\partial_t z_{kh}\|_{I_m}^2 + \|\Delta_h z_{kh}\|_I^2 + \sum_{m=1}^M k_m^{-1} \|[z_{kh}]_m\|^2 \leq C \{ \|g\|_I^2 + \|\nabla \Pi_h z_T\|^2 \}$$

*and*

$$\|z_{kh}\|_I^2 + \|\nabla z_{kh}\|_I^2 \leq C \{ \|g\|_I^2 + \|\nabla \Pi_h z_T\|^2 + \|\Pi_h z_T\|^2 \}.$$

*Here, the jump term  $[z_{kh}]_M$  at  $t = T$  is defined as  $\Pi_h z_T - z_{kh,M}^-$ .*

**5. Analysis of the discretization error for the state equation.** The goal of this section is to prove an a priori error estimate for the discretization error of the (uncontrolled) state equation. Due to the choice of the control space  $Q = L^2(I, L^2(\Omega))$ , we will need error estimates for the error in the state (and adjoint) variable with respect to the norm of  $L^2(I, L^2(\Omega))$ ; cf. the discussion in section 6. Similar error estimates with respect to the  $L^\infty(I, L^2(\Omega))$ -norm can be found in [8, 9], and with respect to the  $L^2(I, H^1(\Omega))$ -norm in [13].

Let  $u \in X$  be the solution of the state equation (2.1) for  $q = 0$ ,  $u_k \in X_k^r$  be the solution of the corresponding semidiscretized equation (3.3), and  $u_{kh} \in X_{k,h}^{r,s}$  be the solution of the fully discretized state equation (3.9). To separate the influences of the space and time discretizations, we split the total discretization error  $e := u - u_{kh}$  into its temporal part  $e_k := u - u_k$  and its spatial part  $e_h := u_k - u_{kh}$ . The temporal

discretization error will be estimated in the following subsection, and the spatial discretization error is treated in section 5.2.

Throughout this section we will assume that the solutions  $u \in X$  and  $u_k \in X_k^r$  possess the regularity  $\partial_t^{r+1}u \in L^2(I, L^2(\Omega))$  and  $\nabla^{s+1}u_k \in L^2(I, L^2(\Omega))$ . Note that Proposition 2.1 and Theorem 4.1 ensure this assumption for  $s = 1$  and  $r = 0$  for convex polygonal domains  $\Omega$ . Better regularity results ( $r > 0$ ,  $s > 1$ ) usually require stronger assumptions on the domain  $\Omega$  and additional compatibility relations.

**5.1. Analysis of the temporal discretization error.** In this section, we will prove the following error estimate for the temporal discretization error  $e_k$ .

**THEOREM 5.1.** *For the error  $e_k := u - u_k$  between the continuous solution  $u \in X$  of (2.1) and the  $dG(r)$  semidiscretized solution  $u_k \in X_k^r$  of (3.3), we have the error estimate*

$$\|e_k\|_I \leq Ck^{r+1}\|\partial_t^{r+1}u\|_I,$$

where the constant  $C$  is independent of the size of the time steps  $k$ .

For clarity of presentation, we divide the proof of this theorem into several steps, which are discussed in the following lemmas.

Before doing so, we define a semidiscrete projection  $\pi_k: C(\bar{I}, V) \rightarrow X_k^r$  for  $m = 1, 2, \dots, M$  by  $\pi_k u|_{I_m} \in \mathcal{P}_r(I_m, V)$  and

$$(5.1a) \quad (\pi_k u - u, \varphi)_{I_m} = 0 \quad \forall \varphi \in \mathcal{P}_{r-1}(I_m, V) \quad \text{for } r > 0,$$

$$(5.1b) \quad \pi_k u(t_m^-) = u(t_m^-).$$

In the case  $r = 0$ ,  $\pi_k u$  is defined solely by condition (5.1b). The projection  $\pi_k$  is well-defined by these conditions; see, for instance, [27] or [26]. By Proposition 2.1 the solution  $u$  of (2.1) belongs to  $C(\bar{I}, V)$ , and therefore  $\pi_k$  is applicable to the state  $u$ .

To shorten the notation in the following analysis, we introduce the abbreviations

$$\eta_k := u - \pi_k u \quad \text{and} \quad \xi_k := \pi_k u - u_k$$

and split the error  $e_k$  as

$$e_k = \eta_k + \xi_k.$$

**LEMMA 5.2.** *For the projection error  $\eta_k$  defined above, the identity*

$$B(\eta_k, \varphi) = (\nabla \eta_k, \nabla \varphi)_I$$

holds for all  $\varphi \in X_k^r$ .

*Proof.* By means of (3.8), we have

$$B(\eta_k, \varphi) = - \sum_{m=1}^M (\eta_k, \partial_t \varphi)_{I_m} + (\nabla \eta_k, \nabla \varphi)_I - \sum_{m=1}^{M-1} (\eta_{k,m}^-, [\varphi]_m) + (\eta_{k,M}^-, \varphi_{k,M}^-).$$

The term  $(\eta_k, \partial_t \varphi)_{I_m}$  vanishes due to (5.1a), and  $\eta_{k,m}^- = 0$  for all  $m$  due to (5.1b). This completes the proof.  $\square$

**LEMMA 5.3.** *The temporal discretization error  $e_k = u - u_k$  is bounded by the projection error  $\eta_k$  with respect to the  $L^2(I, L^2(\Omega))$ -norm, that is,*

$$\|e_k\|_I \leq C\|\eta_k\|_I.$$



*Proof.* We define  $\tilde{z}_k \in X_k^r$  to be the solution of

$$B(\varphi, \tilde{z}_k) = (\varphi, e_k)_I \quad \forall \varphi \in X_k^r.$$

Thus, we obtain by Galerkin orthogonality the relation  $B(\xi_k + \eta_k, \tilde{z}_k) = 0$  (cf. Remark 3.1) which implies

$$\|e_k\|_I^2 = (\xi_k, e_k)_I + (\eta_k, e_k)_I = B(\xi_k, \tilde{z}_k) + (\eta_k, e_k)_I = -B(\eta_k, \tilde{z}_k) + (\eta_k, e_k)_I.$$

Using Lemma 5.2 and integration by parts in space, and the stability estimate from Corollary 4.2, it follows that

$$-B(\eta_k, \tilde{z}_k) = -(\nabla \eta_k, \nabla \tilde{z}_k)_I = (\eta_k, \Delta \tilde{z}_k)_I \leq \|\eta_k\|_I \|\Delta \tilde{z}_k\|_I \leq C \|\eta_k\|_I \|e_k\|_I.$$

Note that the arising boundary terms vanish for both homogeneous Neumann and homogeneous Dirichlet boundary conditions. This leads, by means of Cauchy's inequality, to the desired assertion.  $\square$

LEMMA 5.4. *For the projection error  $\eta_k = u - \pi_k u$  the following estimate holds:*

$$\|\eta_k\|_{I_m} \leq C k_m^{r+1} \|\partial_t^{r+1} u\|_{I_m}.$$

*Proof.* Similarly to [27], the proof is done by standard arguments utilizing the Bramble–Hilbert lemma from [3].  $\square$

After these preparations, we are able to give the proof of Theorem 5.1.

*Proof of Theorem 5.1.* From the Lemmas 5.3 and 5.4 we directly obtain

$$\|e_k\|_I^2 \leq C \|\eta_k\|_I^2 = C \sum_{m=1}^M \|\eta_k\|_{I_m}^2 \leq C \sum_{m=1}^M k_m^{2r+2} \|\partial_t^{r+1} u\|_{I_m}^2 \leq C k^{2r+2} \|\partial_t^{r+1} u\|_I^2,$$

which implies the stated result.  $\square$

**5.2. Analysis of the spatial discretization error.** In this section we give a proof of the following result.

THEOREM 5.5. *For the error  $e_h := u_k - u_{kh}$  between the  $dG(r)$  semidiscretized solution  $u_k \in X_k^r$  of (3.3) and the fully  $cG(s)dG(r)$  discretized solution  $u_{kh} \in X_{k,h}^{r,s}$  of (3.9), we have the error estimate*

$$\|e_h\|_I \leq C h^{s+1} \|\nabla^{s+1} u_k\|_I,$$

where the constant  $C$  is independent of the mesh size  $h$  and the size of the time steps  $k$ .

Similar to the previous subsection, the proof is divided into several steps which are collected in the following lemmas.

We define the projection  $\pi_h: X_k^r \rightarrow X_{k,h}^{r,s}$  by means of the spatial  $L^2$ -projection  $\Pi_h: V \rightarrow V_h^s$  pointwise in time as

$$(\pi_h u_k)(t) = \Pi_h u_k(t).$$

For the solutions of the semidiscrete and fully discretized state equations  $u_k \in X_k^r$  and  $u_{kh} \in X_{k,h}^{r,s}$ , and for  $\tilde{z}_k \in X_k^r$  being the solution of the dual equation (4.7) with right-hand side  $g = e_h$  and terminal condition  $\tilde{z}_T = 0$ , we use the abbreviations

$$\eta_h := u_k - \pi_h u_k, \quad \xi_h := \pi_h u_k - u_{kh}, \quad \text{and} \quad \eta_h^* := \tilde{z}_k - \pi_h \tilde{z}_k,$$

and split the error  $e_h$  as

$$e_h = \eta_h + \xi_h.$$

LEMMA 5.6. *For the projection errors  $\eta_h$  and  $\eta_h^*$  defined above, the identities*

$$B(\eta_h, \varphi) = (\nabla \eta_h, \nabla \varphi)_I \quad \text{and} \quad B(\varphi, \eta_h^*) = (\nabla \varphi, \nabla \eta_h^*)_I$$

hold for all  $\varphi \in X_{k,h}^{r,s}$ .

*Proof.* As in the proof of Lemma 5.2 we obtain

$$\begin{aligned} B(\eta_h, \varphi) &= - \sum_{m=1}^M (\eta_h, \partial_t \varphi)_{I_m} + (\nabla \eta_h, \nabla \varphi)_I - \sum_{m=1}^{M-1} (\eta_{h,m}^-, [\varphi]_m) + (\eta_{h,M}^-, \varphi_M^-) \\ &= (\nabla \eta_h, \nabla \varphi)_I \end{aligned}$$

by means of the definition of  $\pi_h$ . The assertion for  $B(\varphi, \eta_h^*)$  follows directly when employing representation (3.2) instead of (3.8).  $\square$

LEMMA 5.7. *For the error  $\xi_h$  and the projection error  $\eta_h$ , the estimate*

$$\|\nabla \xi_h\|_I \leq \|\nabla \eta_h\|_I$$

holds.

*Proof.* As in [13], we have for all  $v \in X_k^r$  by (3.2) and (3.8)

$$\begin{aligned} B(v, v) &= \sum_{m=1}^M (\partial_t v, v)_{I_m} + (\nabla v, \nabla v)_I + \sum_{m=1}^{M-1} ([v]_m, v_m^+) + (v_0^+, v_0^+), \\ B(v, v) &= - \sum_{m=1}^M (v, \partial_t v)_{I_m} + (\nabla v, \nabla v)_I + \sum_{m=1}^{M-1} (-v_m^-, [v]_m) + (v_M^-, v_M^-). \end{aligned}$$

We arrive at

$$B(v, v) \geq (\nabla v, \nabla v)_I$$

by adding these two identities. Utilizing the Galerkin orthogonality of the space discretization, we obtain

$$\|\nabla \xi_h\|_I^2 = (\nabla \xi_h, \nabla \xi_h)_I \leq B(\xi_h, \xi_h) = -B(\eta_h, \xi_h) = -(\nabla \eta_h, \nabla \xi_h)_I \leq \|\nabla \eta_h\|_I \|\nabla \xi_h\|_I.$$

Division by  $\|\nabla \xi_h\|_I$  leads to the asserted result.  $\square$

LEMMA 5.8. *For the projection errors  $\eta_h$  and  $\eta_h^*$  we have the intermediary result*

$$B(\eta_h, \eta_h^*) \leq \|\nabla \eta_h\|_I \|\nabla \eta_h^*\|_I + C \|\eta_h\|_I \|e_h\|_I.$$

*Proof.* Since  $\pi_h \tilde{z}_k \in X_{k,h}^{r,s}$ , it holds by (3.8) and the definition of  $\pi_h$  that

$$B(\eta_h, \eta_h^*) = - \sum_{m=1}^M (\eta_h, \partial_t \tilde{z}_k)_{I_m} + (\nabla \eta_h, \nabla \eta_h^*)_I - \sum_{m=1}^{M-1} (\eta_{h,m}^-, [\tilde{z}_k]_m) + (\eta_{h,M}^-, \tilde{z}_{k,M}^-).$$

Using  $\tilde{z}_T = 0$ , we subtract the term  $(\eta_{h,M}^-, \tilde{z}_T)$  and obtain by means of the definition  $[\tilde{z}_k]_M = \tilde{z}_T - \tilde{z}_{k,M}^-$

$$(5.2) \quad B(\eta_h, \eta_h^*) = - \sum_{m=1}^M (\eta_h, \partial_t \tilde{z}_k)_{I_m} + (\nabla \eta_h, \nabla \eta_h^*)_I - \sum_{m=1}^M (\eta_{h,m}^-, [\tilde{z}_k]_m).$$

Now, we separately treat the three terms on the right-hand side above: For the term containing spatial derivatives, we have immediately

$$(5.3) \quad (\nabla \eta_h, \nabla \eta_h^*)_I \leq \|\nabla \eta_h\|_I \|\nabla \eta_h^*\|_I.$$

For the term containing the time derivatives, we achieve by Cauchy's inequality and with the stability estimate from Corollary 4.2

$$(5.4) \quad - \sum_{m=1}^M (\eta_h, \partial_t \tilde{z}_k)_{I_m} \leq \|\eta_h\|_I \left( \sum_{m=1}^M \|\partial_t \tilde{z}_k\|_{I_m}^2 \right)^{\frac{1}{2}} \leq C \|\eta_h\|_I \|e_h\|_I.$$

For the jump terms, we obtain again by Cauchy's inequality

$$- \sum_{m=1}^M (\eta_{h,m}^-, [\tilde{z}_k]_m) \leq \left( \sum_{m=1}^M k_m \|\eta_{h,m}^-\|^2 \right)^{\frac{1}{2}} \left( \sum_{m=1}^M k_m^{-1} \|[\tilde{z}_k]_m\|^2 \right)^{\frac{1}{2}}.$$

Utilizing the inverse estimate (cf. [8])

$$k_m \|\eta_{h,m}^-\|^2 \leq C \|\eta_h\|_{I_m}^2,$$

which holds true for polynomials in time, and the stability estimate from Corollary 4.2, we finally obtain

$$(5.5) \quad - \sum_{m=1}^M (\eta_{h,m}^-, [\tilde{z}_k]_m) \leq C \|\eta_h\|_I \|e_h\|_I.$$

We complete the proof by inserting the three estimates (5.3), (5.4), and (5.5) into (5.2).  $\square$

We are now prepared to give the proof of Theorem 5.5.

*Proof of Theorem 5.5.* The solution  $\tilde{z}_k \in X_k^r$  defined above satisfies

$$B(\varphi, \tilde{z}_k) = (\varphi, e_h)_I \quad \forall \varphi \in X_k^r.$$

Due to Galerkin orthogonality, which is applicable for  $\pi_h \tilde{z}_k \in X_{k,h}^{r,s}$ , the identity

$$\|e_h\|_I^2 = B(e_h, \tilde{z}_k) = B(e_h, \tilde{z}_k - \pi_h \tilde{z}_k) = B(\xi_h, \eta_h^*) + B(\eta_h, \eta_h^*)$$

is fulfilled. For the first term we obtain, using Lemma 5.6 and Lemma 5.7,

$$B(\xi_h, \eta_h^*) = (\nabla \xi_h, \nabla \eta_h^*)_I \leq \|\nabla \xi_h\|_I \|\nabla \eta_h^*\|_I \leq \|\nabla \eta_h\|_I \|\nabla \eta_h^*\|_I.$$

This yields, together with Lemma 5.8,

$$(5.6) \quad \|e_h\|_I^2 \leq 2 \|\nabla \eta_h\|_I \|\nabla \eta_h^*\|_I + C \|\eta_h\|_I \|e_h\|_I.$$

Due to the definition of  $\pi_h$ , well-known a priori estimates for the spatial  $L^2$ -projection  $\Pi_h$  can be employed to directly obtain estimates for  $\eta_h$  and  $\eta_h^*$ . We have

$$\|\eta_h\|_I \leq Ch^{s+1} \|\nabla^{s+1} u_k\|_I, \quad \|\nabla \eta_h\|_I \leq Ch^s \|\nabla^{s+1} u_k\|_I, \quad \|\nabla \eta_h^*\|_I \leq Ch \|\nabla^2 \tilde{z}_k\|_I.$$

These estimates applied to (5.6) lead to

$$\|e_h\|_I^2 \leq Ch^{s+1} \|\nabla^{s+1} u_k\|_I \{ \|\nabla^2 \tilde{z}_k\|_I + \|e_h\|_I \}.$$

Due to the fact that the domain  $\Omega$  is polygonal and convex, elliptic regularity theory yields

$$\|\nabla^2 z_k\|_I \leq C \|\Delta z_k\|_I,$$

and we obtain the stated result by the stability estimate from Corollary 4.2.  $\square$

**6. Error analysis for the optimal control problem.** In this section, we prove the main results of this article, namely, an estimate of the error between the solution  $(\bar{q}, \bar{u})$  of the continuous optimal control problem (2.2) and the solution  $(\bar{q}_\sigma, \bar{u}_\sigma)$  of the discretized problem (3.14).

Throughout this section, we will indicate the dependence of the state and the adjoint state on the specific control  $q \in Q$  by the notation introduced in section 2 and section 3, that is,  $u(q)$ ,  $z(q)$  on the continuous level,  $u_k(q)$ ,  $z_k(q)$  on the semidiscrete and  $u_{kh}(q)$ ,  $z_{kh}(q)$  on the discrete level.

**6.1. Error in the control variable.** In this section we analyze the error with respect to the control variable and prove the following result.

**THEOREM 6.1.** *The error between the solution  $\bar{q} \in Q$  of the continuous optimization problem (2.2) and the solution  $\bar{q}_\sigma \in Q_d$  of the discrete optimization problem (3.14) can be estimated as*

$$\begin{aligned} \|\bar{q} - \bar{q}_\sigma\|_I &\leq \frac{C}{\alpha} k^{r+1} \{ \|\partial_t^{r+1} u(\bar{q})\|_I + \|\partial_t^{r+1} z(\bar{q})\|_I \} \\ &\quad + \frac{C}{\alpha} h^{s+1} \{ \|\nabla^{s+1} u_k(\bar{q})\|_I + \|\nabla^{s+1} z_k(\bar{q})\|_I \} + \left( 2 + \frac{C}{\alpha} \right) \inf_{p_d \in Q_d} \|\hat{q} - p_d\|_I, \end{aligned}$$

where  $\hat{q} \in Q$  can be chosen either as the continuous solution  $\bar{q}$  or as the solution  $\bar{q}_{kh}$  of the purely state discretized problem (3.10). The constants  $C$  are independent of the mesh size  $h$ , the size of the time steps  $k$ , and the choice of the discrete control space  $Q_d \subset Q$ .

We first discuss the infimum term appearing on the right-hand side of the error estimate above. Thereby, we make use of the two possible formulations of this term for  $\hat{q} = \bar{q}$  or  $\hat{q} = \bar{q}_{kh}$ : From the optimality conditions (3.11) for the optimal control problem (3.10) obtained after the discretization of the state equation in space and time, we get

$$(\bar{q}_{kh}, \delta q)_I = \frac{1}{\alpha} (z_{kh}(\bar{q}_{kh}), \delta q)_I \quad \forall \delta q \in Q,$$

and therefore  $\bar{q}_{kh} = \frac{1}{\alpha} z_{kh}(\bar{q}_{kh}) \in X_{k,h}^{r,s} \subset Q$ . Thus, if  $Q_d$  is chosen such that  $Q_d \supset X_{k,h}^{r,s}$ , the term

$$\inf_{p_d \in Q_d} \|\bar{q}_{kh} - p_d\|_I$$

vanishes. In this case, the solution  $\bar{q}_\sigma$  of the fully discretized optimal control problem (3.14) coincides with the solution  $\bar{q}_{kh}$ ; cf. [16]. Consequently, it is reasonable to discretize the control at most as fine as the adjoint state. The same conclusion can be drawn by inspection of the a posteriori error estimates developed in [21].

If the discrete control space  $Q_d$  does not fulfill the condition  $Q_d \supset X_{k,h}^{r,s}$ , it is desirable to choose  $\hat{q} = \bar{q}$  in the above theorem to obtain an estimate for the infimum term. For both choices of the space  $Q_d$  described in section 3.3 we obtain the following estimate using interpolation theory:

$$\inf_{p_d \in Q_d} \|\bar{q} - p_d\|_I \leq Ck^{r_d+1} \|\partial_t^{r_d+1} \bar{q}\|_I + Ch_d^{s_d+1} \|\nabla^{s_d+1} \bar{q}\|_I.$$

Here,  $h_d$  is the discretization parameter corresponding to the spatial mesh employed for the control discretization.

The proof of Theorem 6.1 makes use of the assertions of the following lemmas and will be given at the end of this section.

LEMMA 6.2. *Let  $q \in Q$  be a given control. The error between the continuous state  $u = u(q) \in X$  determined by (2.1) and the discrete state  $u_{kh} = u_{kh}(q) \in X_{k,h}^{r,s}$  determined by (3.9) can be estimated as*

$$\|u(q) - u_{kh}(q)\|_I \leq Ck^{r+1} \|\partial_t^{r+1} u(q)\|_I + Ch^{s+1} \|\nabla^{s+1} u_k(q)\|_I.$$

For the error between the continuous adjoint state  $z = z(q) \in X$  determined by (2.5) and the discrete adjoint state  $z_{kh} = z_{kh}(q) \in X_{k,h}^{r,s}$  determined by (3.13), the following estimate holds:

$$\begin{aligned} \|z(q) - z_{kh}(q)\|_I &\leq Ck^{r+1} \{ \|\partial_t^{r+1} u(q)\|_I + \|\partial_t^{r+1} z(q)\|_I \} \\ &\quad + Ch^{s+1} \{ \|\nabla^{s+1} u_k(q)\|_I + \|\nabla^{s+1} z_k(q)\|_I \}. \end{aligned}$$

*Proof.* The estimate for the error in terms of the state variable is immediately obtained by Theorems 5.1 and 5.5 since for  $q \in Q$  the right-hand side  $f + q$  of the state equation (2.1) is in  $L^2(I, H)$  and thus fulfills the assumptions of Proposition 2.1.

For estimating the error in  $z$ , we introduce additionally the solutions  $\tilde{z}_k \in X_k^r$  and  $\tilde{z}_{kh} \in X_{k,h}^{r,s}$  which solve

$$\begin{aligned} B(\varphi, \tilde{z}_k) &= (\varphi, u(q) - \hat{u})_I \quad \forall \varphi \in X_k^r \quad \text{and} \\ B(\varphi, \tilde{z}_{kh}) &= (\varphi, u_k(q) - \hat{u})_I \quad \forall \varphi \in X_{k,h}^{r,s}. \end{aligned}$$

Since the adjoint solution  $z(q) \in X$  is determined by (2.5), we may apply Theorem 5.1 to obtain

$$\|z(q) - \tilde{z}_k\|_I \leq Ck^{r+1} \|\partial_t^{r+1} z(q)\|_I.$$

Correspondingly, due to the definition of the semidiscrete adjoint solution  $z_k(q) \in X_k^r$  by (3.7), Theorem 5.5 yields the estimate

$$\|z_k(q) - \tilde{z}_{kh}\|_I \leq Ch^{s+1} \|\nabla^{s+1} z_k(q)\|_I.$$

Using (3.7) for  $z_k(q)$ , we obtain that the difference  $\tilde{z}_k - z_k(q)$  solves

$$B(\varphi, \tilde{z}_k - z_k(q)) = (\varphi, u(q) - u_k(q))_I \quad \forall \varphi \in X_k^r.$$

Then, the stability estimate from Corollary 4.2 yields

$$\|\tilde{z}_k - z_k(q)\|_I \leq C\|u(q) - u_k(q)\|_I.$$

Similarly, using (3.13) for  $z_{kh}(q)$ , we obtain for the difference  $\tilde{z}_{kh} - z_{kh}(q)$  the identity

$$B(\varphi, \tilde{z}_{kh} - z_{kh}(q)) = (\varphi, u_k(q) - u_{kh}(q))_I \quad \forall \varphi \in X_{k,h}^{r,s},$$

and the stability estimate from Corollary 4.7 implies

$$\|\tilde{z}_{kh} - z_{kh}(q)\|_I \leq C\|u_k(q) - u_{kh}(q)\|_I.$$

Finally, the triangle inequality and the error estimates from Theorems 5.1 and 5.5 for the error in the state variable lead to the proposed result.  $\square$

LEMMA 6.3. *For given controls  $q, r \in Q$ , the difference between the derivatives of the continuous reduced functional  $j$  and the discrete reduced functional  $j_{kh}$  can be estimated by*

$$|j'(q)(r) - j'_{kh}(q)(r)| \leq \|z(q) - z_{kh}(q)\|_I \|r\|_I.$$

*Proof.* The representations (2.6) and (3.12) for  $j'$  and  $j'_{kh}$ , respectively, imply directly the assertion

$$|j'(q)(r) - j'_{kh}(q)(r)| = |(z(q) - z_{kh}(q), r)_I| \leq \|z(q) - z_{kh}(q)\|_I \|r\|_I. \quad \square$$

LEMMA 6.4. *The derivatives of the discrete reduced functional  $j_{kh}$  are Lipschitz continuous on  $Q$ . That is, for arbitrary  $p, q, r \in Q$ , the estimate*

$$|j'_{kh}(q)(r) - j'_{kh}(p)(r)| \leq (C + \alpha)\|q - p\|_I \|r\|_I$$

*holds true.*

*Proof.* By means of (3.12), we have

$$\begin{aligned} |j'_{kh}(q)(r) - j'_{kh}(p)(r)| &\leq \alpha|(q - p, r)_I| + |(z_{kh}(q) - z_{kh}(p), r)_I| \\ &\leq \alpha\|q - p\|_I \|r\|_I + \|z_{kh}(q) - z_{kh}(p)\|_I \|r\|_I. \end{aligned}$$

Since  $z_{kh}(q) - z_{kh}(p)$  solves

$$B(\varphi, z_{kh}(q) - z_{kh}(p)) = (\varphi, u_{kh}(q) - u_{kh}(p))_I \quad \forall \varphi \in X_{k,h}^{r,s},$$

and  $u_{kh}(q) - u_{kh}(p)$  satisfies

$$B(u_{kh}(q) - u_{kh}(p), \varphi) = (q - p, \varphi)_I \quad \forall \varphi \in X_{k,h}^{r,s},$$

the stability estimates for  $z_{kh}$  from Corollary 4.7 and for  $u_{kh}$  from Theorem 4.6 yield

$$\|z_{kh}(q) - z_{kh}(p)\|_I \leq C\|u_{kh}(q) - u_{kh}(p)\|_I \leq C\|q - p\|_I,$$

which implies the desired result.  $\square$

With the aid of these preliminary results, we now prove Theorem 6.1.

*Proof of Theorem 6.1.* To obtain the asserted result, we split the error to be estimated in two different ways:

$$(6.1) \quad \|\bar{q} - \bar{q}_\sigma\|_I \leq \|\bar{q} - p_d\|_I + \|p_d - \bar{q}_\sigma\|_I,$$

$$(6.2) \quad \|\bar{q} - \bar{q}_\sigma\|_I \leq \|\bar{q} - \bar{q}_{kh}\|_I + \|\bar{q}_{kh} - p_d\|_I + \|p_d - \bar{q}_\sigma\|_I.$$

Here,  $p_d$  is an arbitrary element of  $Q_d$  and  $\bar{q}$ ,  $\bar{q}_{kh}$ , and  $\bar{q}_\sigma$  are the optimal solutions on the different levels of discretization.

Due to the linear-quadratic structure of the optimal control problem under consideration, we have for all  $p, r \in Q$ ,

$$j''_{kh}(p)(r, r) \geq \alpha \|r\|_I^2,$$

and  $j''_{kh}(p)$  does not depend on  $p$ . This implies, for arbitrary  $p, p_d \in Q_d$ ,

$$\alpha \|p_d - \bar{q}_\sigma\|_I^2 \leq j''_{kh}(p)(p_d - \bar{q}_\sigma, p_d - \bar{q}_\sigma) = j'_{kh}(p_d)(p_d - \bar{q}_\sigma) - j'_{kh}(\bar{q}_\sigma)(p_d - \bar{q}_\sigma).$$

Since  $\bar{q}$ ,  $\bar{q}_{kh}$ , and  $\bar{q}_\sigma$  are the optimal solutions of the continuous, semidiscrete, and discrete optimization problems, we have by (2.4), (3.11), and (3.15),

$$j'_{kh}(\bar{q}_\sigma)(p_d - \bar{q}_\sigma) = j'_{kh}(\bar{q}_{kh})(p_d - \bar{q}_\sigma) = j'(\bar{q})(p_d - \bar{q}_\sigma) = 0.$$

Using these identities, we obtain for the separation (6.1) (which we use to prove the theorem in the case  $\hat{q} = \bar{q}$ ) the estimate

$$\begin{aligned} \alpha \|p_d - \bar{q}_\sigma\|_I^2 &\leq j'_{kh}(p_d)(p_d - \bar{q}_\sigma) - j'(\bar{q})(p_d - \bar{q}_\sigma) \\ &= j'_{kh}(p_d)(p_d - \bar{q}_\sigma) - j'_{kh}(\bar{q})(p_d - \bar{q}_\sigma) + j'_{kh}(\bar{q})(p_d - \bar{q}_\sigma) - j'(\bar{q})(p_d - \bar{q}_\sigma). \end{aligned}$$

By means of Lemmas 6.3 and 6.4, we achieve

$$\alpha \|p_d - \bar{q}_\sigma\|_I^2 \leq (C + \alpha) \|p_d - \bar{q}\|_I \|p_d - \bar{q}_\sigma\|_I + \|z(\bar{q}) - z_{kh}(\bar{q})\|_I \|p_d - \bar{q}_\sigma\|_I.$$

Using (6.1) we get the estimate

$$(6.3) \quad \|\bar{q} - \bar{q}_\sigma\|_I \leq \frac{1}{\alpha} \|z(\bar{q}) - z_{kh}(\bar{q})\|_I + \left(2 + \frac{C}{\alpha}\right) \|\bar{q} - p_d\|_I.$$

To use separation (6.2) for proving the theorem in the case  $\hat{q} = \bar{q}_{kh}$ , we estimate alternatively by means of Lemma 6.4,

$$\alpha \|p_d - \bar{q}_\sigma\|_I^2 \leq j'_{kh}(p_d)(p_d - \bar{q}_\sigma) - j'_{kh}(\bar{q}_{kh})(p_d - \bar{q}_\sigma) \leq (C + \alpha) \|p_d - \bar{q}_{kh}\|_I \|p_d - \bar{q}_\sigma\|_I.$$

In the same manner as before, we can estimate  $\|\bar{q} - \bar{q}_{kh}\|_I$  by Lemma 6.3 as

$$\begin{aligned} \alpha \|\bar{q} - \bar{q}_{kh}\|_I^2 &\leq j''_{kh}(p)(\bar{q} - \bar{q}_{kh}, \bar{q} - \bar{q}_{kh}) \\ &= j'_{kh}(\bar{q})(\bar{q} - \bar{q}_{kh}) - j'_{kh}(\bar{q}_{kh})(\bar{q} - \bar{q}_{kh}) \\ &= j'_{kh}(\bar{q})(\bar{q} - \bar{q}_{kh}) - j'(\bar{q})(\bar{q} - \bar{q}_{kh}) \\ &\leq \|z(\bar{q}) - z_{kh}(\bar{q})\|_I \|\bar{q} - \bar{q}_{kh}\|_I. \end{aligned}$$

Then, the two latter estimates imply

$$(6.4) \quad \|\bar{q} - \bar{q}_\sigma\|_I \leq \frac{1}{\alpha} \|z(\bar{q}) - z_{kh}(\bar{q})\|_I + \left(2 + \frac{C}{\alpha}\right) \|\bar{q}_{kh} - p_d\|_I.$$

Finally, the inequalities (6.3) and (6.4) prove the assertion by means of the estimate for  $\|z(\bar{q}) - z_{kh}(\bar{q})\|_I$  from Lemma 6.2.  $\square$

To concretize the result of Theorem 6.1, we consider the following choice of discretizations: The state space is discretized by the cG(1)dG(0) method, that is, we consider the case when  $r = 0$  and  $s = 1$ . Using for simplicity the same triangulation of the spatial domain ( $h_d = h$ ) and the same distribution of the time steps ( $k_d = k$ ) as for the discretization of the state, we discuss the following two possibilities for the

control discretization (cf. section 3.3):

1. *cG(1)dG(0) discretization, i.e., cellwise (bi-/tri-)linear in space and piecewise constant in time:* In this case the infimum term in the estimate in Theorem 6.1 vanishes, since  $Q_d \supset X_{k,h}^{r,s}$ ; see the above discussion. Thus, Theorem 6.1 implies that the discretization error is of order

$$\|\bar{q} - \bar{q}_\sigma\|_I = \mathcal{O}(k + h^2).$$

2. *dG(0)dG(0) discretization, i.e., cellwise constant in space and piecewise constant in time:* In this case the infimum term of the error estimation from Theorem 6.1 has to be taken into account, leading to the discretization error of order

$$\|\bar{q} - \bar{q}_\sigma\|_I = \mathcal{O}(k + h).$$

Note that the regularity of the optimal solutions required for these estimates is ensured by Propositions 2.1 and 2.2 for the continuous solutions  $q$ ,  $u$ , and  $z$ , and by Theorem 4.1 and Corollary 4.2 for the time-discrete solutions  $u_k$  and  $z_k$ . A numerical validation of these estimates will be given in section 7.

**6.2. Error in the state and in the adjoint variable.** In this subsection we prove error estimates for the state and adjoint state variables. That is, we consider the discretization errors

$$\|\bar{u} - \bar{u}_\sigma\|_I = \|u(\bar{q}) - u_{kh}(\bar{q}_\sigma)\|_I \quad \text{and} \quad \|\bar{z} - \bar{z}_\sigma\|_I = \|z(\bar{q}) - z_{kh}(\bar{q}_\sigma)\|_I.$$

By means of the stability estimates derived in section 4, one simply obtains the following result.

**THEOREM 6.5.** *Let  $(\bar{q}, \bar{u})$  be the solution of the continuous optimal control problem (2.2) and  $\bar{z} = z(\bar{q})$  be the corresponding adjoint state. Let, moreover,  $(\bar{q}_\sigma, \bar{u}_\sigma)$  be the solution of the discrete optimal control problem (3.14) with the corresponding discrete adjoint state  $\bar{z}_\sigma = z_{kh}(\bar{q}_\sigma)$ . Then, the following estimates hold:*

- (i)  $\|\bar{u} - \bar{u}_\sigma\|_I \leq \|u(\bar{q}) - u_{kh}(\bar{q})\|_I + C\|\bar{q} - \bar{q}_\sigma\|_I,$
- (ii)  $\|\bar{z} - \bar{z}_\sigma\|_I \leq \|z(\bar{q}) - z_{kh}(\bar{q})\|_I + C\|\bar{q} - \bar{q}_\sigma\|_I.$

*Proof.* Using the fact that  $\bar{u} = u(\bar{q})$  and  $\bar{u}_\sigma = u_{kh}(\bar{q}_\sigma)$ , we have

$$(6.5) \quad \|\bar{u} - \bar{u}_\sigma\|_I \leq \|u(\bar{q}) - u_{kh}(\bar{q})\|_I + \|u_{kh}(\bar{q}) - u_{kh}(\bar{q}_\sigma)\|_I.$$

By means of the stability result from Theorem 4.6, we obtain

$$\|u_{kh}(\bar{q}) - u_{kh}(\bar{q}_\sigma)\|_I \leq C\|\bar{q} - \bar{q}_\sigma\|_I.$$

This proves the first assertion. The second assertion follows in the same way utilizing the stability of the adjoint state given by Corollary 4.7.  $\square$

Employing the discretization of the control by cG(1)dG(0), the above theorem leads to the optimal order of convergence using Lemma 6.2 and Theorem 6.1. That is, we have

$$\begin{aligned} \|u(\bar{q}) - u_{kh}(\bar{q})\|_I &= \mathcal{O}(k + h^2), & \|z(\bar{q}) - z_{kh}(\bar{q})\|_I &= \mathcal{O}(k + h^2), \\ \|\bar{q} - \bar{q}_\sigma\|_I &= \mathcal{O}(k + h^2), \end{aligned}$$



and thus

$$\|\bar{u} - \bar{u}_\sigma\|_I = \mathcal{O}(k + h^2) \quad \text{and} \quad \|\bar{z} - \bar{z}_\sigma\|_I = \mathcal{O}(k + h^2).$$

However, in the case of dG(0)dG(0) discretization, this theorem does not lead to the optimal order of convergence: In this case, we have indeed as before,

$$\|u(\bar{q}) - u_{kh}(\bar{q})\|_I = \mathcal{O}(k + h^2) \quad \text{and} \quad \|z(\bar{q}) - z_{kh}(\bar{q})\|_I = \mathcal{O}(k + h^2)$$

since the discretization of the state space is unaffected by the discretization of the controls, but we have only

$$\|\bar{q} - \bar{q}_\sigma\|_I = \mathcal{O}(k + h)$$

due to the first order discretization of the control space. This would lead to  $\mathcal{O}(k + h)$  for the state and the adjoint variable.

Utilizing a more detailed analysis, we can prove also in this case the optimal order of convergence  $\mathcal{O}(k + h^2)$  for the errors  $\|\bar{u} - \bar{u}_\sigma\|_I$  and  $\|\bar{z} - \bar{z}_\sigma\|_I$ .

For both choices of the space  $Q_d$  described in section 3.3 the following results hold.

**THEOREM 6.6.** *Let  $(\bar{q}, \bar{u})$  be the solution of the continuous optimal control problem (2.2) and  $\bar{z} = z(\bar{q})$  be the corresponding adjoint state. Let, moreover,  $(\bar{q}_\sigma, \bar{u}_\sigma)$  be the solution of the discrete optimal control problem (3.14) with the corresponding discrete adjoint state  $\bar{z}_\sigma = z_{kh}(\bar{q}_\sigma)$ . In addition we assume  $r = r_d$ , i.e., the same discretization of the state and control variable in time. Then, the following estimates hold:*

$$\begin{aligned} \text{(i)} \quad & \|\bar{u} - \bar{u}_\sigma\|_I \leq \|u(\bar{q}) - u_{kh}(\bar{q})\|_I + Ch_d \left(1 + \frac{1}{\alpha}\right) \|\bar{q} - \pi_d \bar{q}\|_I + \frac{C}{\alpha} \|z(\bar{q}) - z_{kh}(\bar{q})\|_I, \\ \text{(ii)} \quad & \|\bar{z} - \bar{z}_\sigma\|_I \leq Ch_d \left(1 + \frac{1}{\alpha}\right) \|\bar{q} - \pi_d \bar{q}\|_I + C \left(1 + \frac{1}{\alpha}\right) \|z(\bar{q}) - z_{kh}(\bar{q})\|_I, \end{aligned}$$

where  $\pi_d: Q \rightarrow Q_d$  is the space-time  $L^2$ -projection on  $Q_d$ .

*Proof.* For proving (i) we split the error  $\|\bar{u} - \bar{u}_\sigma\|_I$  as follows:

$$(6.6) \quad \|\bar{u} - \bar{u}_\sigma\|_I \leq \|u(\bar{q}) - u_{kh}(\bar{q})\|_I + \|u_{kh}(\bar{q}) - u_{kh}(\pi_d \bar{q})\|_I + \|u_{kh}(\pi_d \bar{q}) - u_{kh}(\bar{q}_\sigma)\|_I.$$

The second term on the right-hand side of (6.6) is estimated using the following duality argument: Let  $\tilde{z}_{kh} \in X_{k,h}^{r,s}$  be the solution of

$$B(\varphi, \tilde{z}_{kh}) = (\varphi, u_{kh}(\bar{q}) - u_{kh}(\pi_d \bar{q}))_I \quad \forall \varphi \in X_{k,h}^{r,s}.$$

By means of the discrete state equation (3.9) for  $u_{kh}(\bar{q})$  and  $u_{kh}(\pi_d \bar{q})$ , we obtain

$$\|u_{kh}(\bar{q}) - u_{kh}(\pi_d \bar{q})\|_I^2 = B(u_{kh}(\bar{q}) - u_{kh}(\pi_d \bar{q}), \tilde{z}_{kh}) = (\bar{q} - \pi_d \bar{q}, \tilde{z}_{kh})_I.$$

Since  $\pi_d$  is the  $L^2$ -projection, we have

$$(6.7) \quad \|u_{kh}(\bar{q}) - u_{kh}(\pi_d \bar{q})\|_I^2 = (\bar{q} - \pi_d \bar{q}, \tilde{z}_{kh} - \pi_d \tilde{z}_{kh})_I \leq \|\bar{q} - \pi_d \bar{q}\|_I \|\tilde{z}_{kh} - \pi_d \tilde{z}_{kh}\|_I.$$

Using the fact that  $r = r_d$  and that therefore the same time discretization is employed for the control and state variable, the space-time  $L^2$ -projection  $\pi_d$  applied to  $\tilde{z}_{kh}$  can be expressed as spatial  $L^2$ -projection  $\Pi_{h_d} \tilde{z}_{kh}$ .

Applying an interpolation estimate and the stability estimate from Corollary 4.7 we obtain

$$\|\tilde{z}_{kh} - \pi_d \tilde{z}_{kh}\|_I = \|\tilde{z}_{kh} - \Pi_{h_d} \tilde{z}_{kh}\|_I \leq Ch_d \|\nabla \tilde{z}_{kh}\|_I \leq Ch_d \|u_{kh}(\bar{q}) - u_{kh}(\pi_d \bar{q})\|_I.$$

Plugging this estimate into (6.7) yields

$$(6.8) \quad \|u_{kh}(\bar{q}) - u_{kh}(\pi_d \bar{q})\|_I \leq Ch_d \|\bar{q} - \pi_d \bar{q}\|_I.$$

For the third term in (6.6) we obtain, using Theorem 4.6,

$$\|u_{kh}(\pi_d \bar{q}) - u_{kh}(\bar{q}_\sigma)\|_I \leq C \|\pi_d \bar{q} - \bar{q}_\sigma\|_I.$$

For estimating the term  $\|\pi_d \bar{q} - \bar{q}_\sigma\|_I$  we proceed as in the proof of Theorem 6.1 for the term  $\|p_d - \bar{q}_\sigma\|_I$ :

$$\alpha \|\pi_d q - q_\sigma\|_I^2 \leq j'_{kh}(\pi_d q)(\pi_d q - q_\sigma) - j'(q)(\pi_d q - q_\sigma).$$

Using representation (3.12) of  $j'_{kh}$  and (2.6) of  $j'$  we have

$$\alpha \|\pi_d \bar{q} - \bar{q}_\sigma\|_I^2 \leq \alpha (\pi_d \bar{q} - \bar{q}, \pi_d \bar{q} - \bar{q}_\sigma)_I + (z_{kh}(\pi_d \bar{q}) - z(\bar{q}), \pi_d \bar{q} - \bar{q}_\sigma)_I.$$

Since  $\pi_d \bar{q} - \bar{q}_\sigma \in Q_d$ , the term  $(\pi_d \bar{q} - \bar{q}, \pi_d \bar{q} - \bar{q}_\sigma)_I$  vanishes, and due to Corollary 4.7, we end up with

$$\begin{aligned} \alpha \|\pi_d \bar{q} - \bar{q}_\sigma\|_I &\leq \|z_{kh}(\pi_d \bar{q}) - z(\bar{q})\|_I \\ &\leq \|z_{kh}(\pi_d \bar{q}) - z_{kh}(\bar{q})\|_I + \|z_{kh}(\bar{q}) - z(\bar{q})\|_I \\ &\leq C \|u_{kh}(\pi_d \bar{q}) - u_{kh}(\bar{q})\|_I + \|z_{kh}(\bar{q}) - z(\bar{q})\|_I, \end{aligned}$$

which implies, by using (6.8), the estimate

$$\begin{aligned} (6.9) \quad \|u_{kh}(\pi_d \bar{q}) - u_{kh}(\bar{q}_\sigma)\|_I &\leq \frac{C}{\alpha} \|u_{kh}(\pi_d \bar{q}) - u_{kh}(\bar{q})\|_I + \frac{C}{\alpha} \|z_{kh}(\bar{q}) - z(\bar{q})\|_I \\ &\leq \frac{C}{\alpha} h_d \|\bar{q} - \pi_d \bar{q}\|_I + \frac{C}{\alpha} \|z_{kh}(\bar{q}) - z(\bar{q})\|_I. \end{aligned}$$

Plugging (6.8) and (6.9) into (6.6) we complete the proof of (i). The assertion (ii) follows using (6.8), (6.9), and the following estimate exploiting the stability result from Corollary 4.7:

$$\begin{aligned} \|\bar{z} - \bar{z}_\sigma\|_I &\leq \|z(\bar{q}) - z_{kh}(\bar{q})\|_I + \|z_{kh}(\bar{q}) - z_{kh}(\pi_d \bar{q})\|_I + \|z_{kh}(\pi_d \bar{q}) - z_{kh}(\bar{q}_\sigma)\|_I \\ &\leq \|z(\bar{q}) - z_{kh}(\bar{q})\|_I + C \|u_{kh}(\bar{q}) - u_{kh}(\pi_d \bar{q})\|_I + C \|u_{kh}(\pi_d \bar{q}) - u_{kh}(\bar{q}_\sigma)\|_I. \quad \square \end{aligned}$$

For the case of dG(0)dG(0) discretization of the control space with  $h_d = h$  and  $k_d = k$  this theorem leads to the improved (optimal) order of convergence

$$\|\bar{u} - \bar{u}_\sigma\|_I = \mathcal{O}(k + h^2) \quad \text{and} \quad \|\bar{z} - \bar{z}_\sigma\|_I = \mathcal{O}(k + h^2).$$

**7. Numerical results.** In this section, we are going to validate the a priori error estimates for the error in the control, state, and adjoint state numerically. To this end, we consider the following concretization of the model problem (2.2) with known analytical exact solution on  $\Omega \times I = (0, 1)^2 \times (0, 0.1)$  and homogeneous Dirichlet boundary conditions. The right-hand side  $f$ , the desired state  $\hat{u}$ , and the initial condition  $u_0$  are given in terms of the eigenfunctions

$$w_a(t, x_1, x_2) := \exp(a\pi^2 t) \sin(\pi x_1) \sin(\pi x_2), \quad a \in \mathbb{R},$$

of the operator  $\pm \partial_t - \Delta$  as

$$f(t, x_1, x_2) := -\pi^4 w_a(T, x_1, x_2),$$

$$\hat{u}(t, x_1, x_2) := \frac{a^2 - 5}{2 + a} \pi^2 w_a(t, x_1, x_2) + 2\pi^2 w_a(T, x_1, x_2),$$

$$u_0(x_1, x_2) := \frac{-1}{2 + a} \pi^2 w_a(0, x_1, x_2).$$

For this choice of data and with the regularization parameter  $\alpha$  chosen as  $\alpha = \pi^{-4}$ , the optimal solution triple  $(\bar{q}, \bar{u}, \bar{z})$  of the optimal control problem (2.2) is given by

$$\bar{q}(t, x_1, x_2) := -\pi^4 \{w_a(t, x_1, x_2) - w_a(T, x_1, x_2)\},$$

$$\bar{u}(t, x_1, x_2) := \frac{-1}{2 + a} \pi^2 w_a(t, x_1, x_2),$$

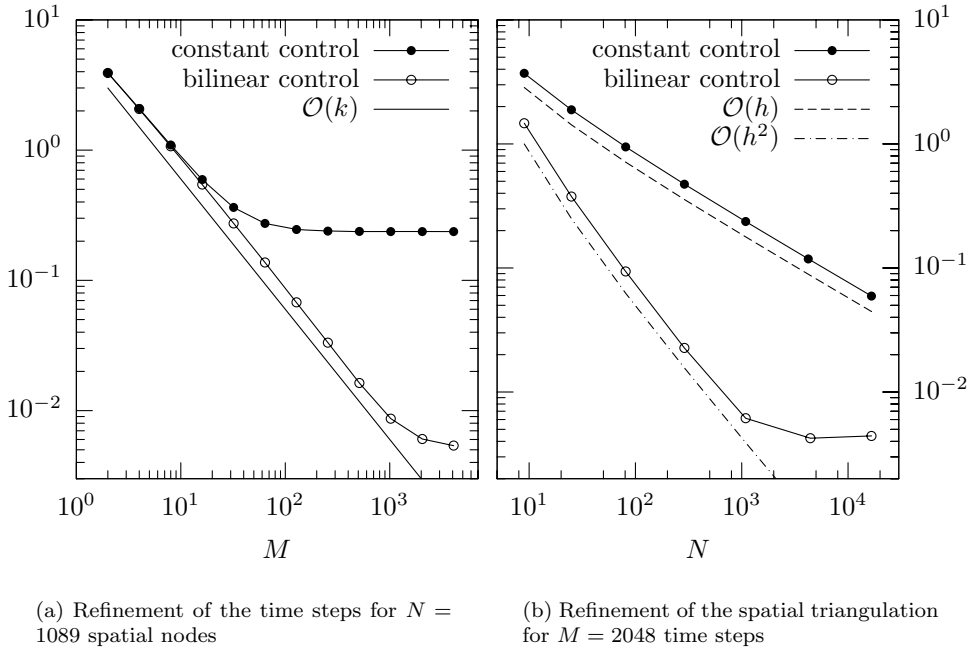
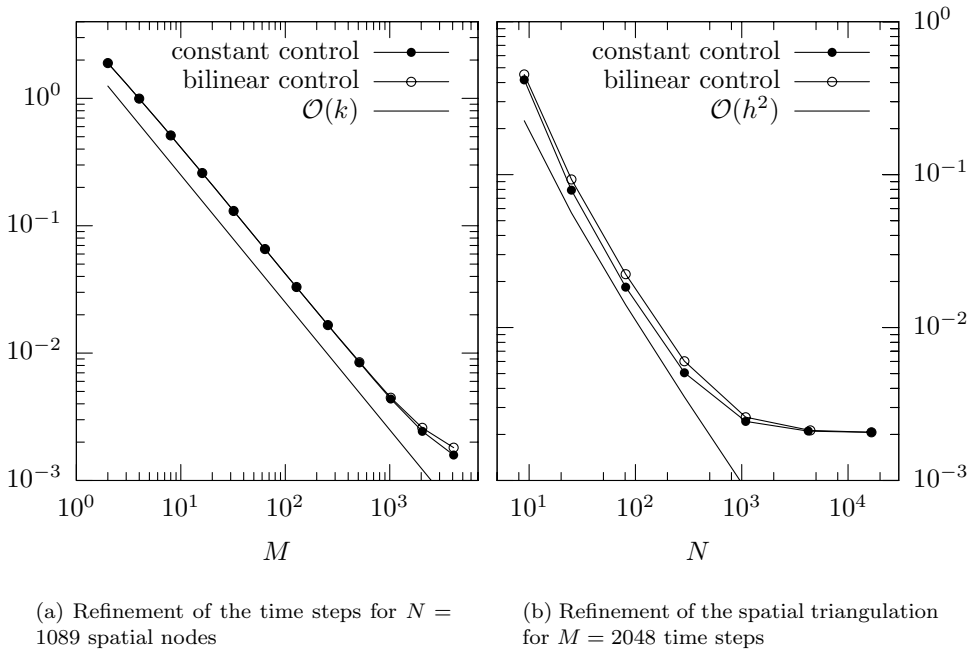
$$\bar{z}(t, x_1, x_2) := w_a(t, x_1, x_2) - w_a(T, x_1, x_2).$$

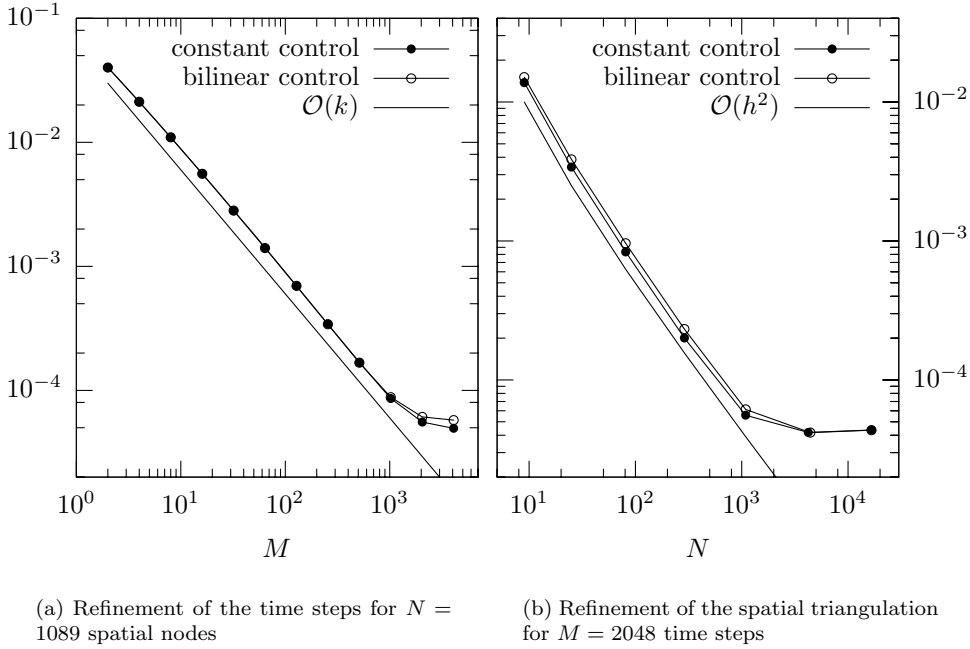
We are going to validate the estimates developed in the previous section by separating the discretization errors. That is, we consider at first the behavior of the error for a sequence of discretizations with decreasing size of the time steps and a fixed spatial triangulation with  $N = 1089$  nodes. Second, we examine the behavior of the error under refinement of the spatial triangulation for  $M = 2048$  time steps.

The state discretization is chosen as cG(1)dG(0), i.e.,  $r = 0$ ,  $s = 1$ . For the control discretization we use the same temporal and spatial meshes as for the state variable and present the result for two choices of the discrete control space  $Q_d$ : cG(1)dG(0) and dG(0)dG(0). For the following computations, we choose the free parameter  $a$  to be  $-\sqrt{5}$ . For this choice the right-hand side  $f$  and the desired state  $\hat{u}$  do not depend on time which avoids side effects introduced by numerical quadrature.

The optimal control problems are solved by the optimization library RoDoBo [23] and the finite element toolkit GASCOIGNE [14] using a conjugate gradient method applied to the reduced problem (3.14).

Figure 7.1(a) depicts the development of the error under refinement of the temporal step size  $k$ . Up to the spatial discretization error it exhibits the proven convergence order  $\mathcal{O}(k)$  for both kinds of spatial discretization of the control space. For piecewise constant control (dG(0)dG(0) discretization), the discretization error is already reached at 128 time steps, whereas in the case of bilinear control (cG(1)dG(0) discretization), the number of time steps could be increased up to  $M = 4096$  until reaching the spatial accuracy.

FIG. 7.1. Discretization error  $\|\bar{q} - \bar{q}_\sigma\|_I$ .FIG. 7.2. Discretization error  $\|\bar{u} - \bar{u}_\sigma\|_I$ .

FIG. 7.3. Discretization error  $\|\bar{z} - \bar{z}_\sigma\|_I$ .

In Figure 7.1(b) the development of the error in the control variable under spatial refinement is shown. The expected order  $\mathcal{O}(h)$  for piecewise constant control (dG(0)dG(0) discretization) and  $\mathcal{O}(h^2)$  for bilinear control (cG(1)dG(0) discretization) is observed.

Figures 7.2 and 7.3 show the errors in the state and the adjoint variables,  $\|\bar{u} - \bar{u}_\sigma\|_I$  and  $\|\bar{z} - \bar{z}_\sigma\|_I$ , for separate refinement of the time and space discretization. Thereby, we observe convergence of order  $\mathcal{O}(k + h^2)$  regardless of the type of spatial discretization used for the controls. This is consistent with the results proven in the previous section.

## REFERENCES

- [1] N. ARADA, E. CASAS, AND F. TRÖLTZSCH, *Error estimates for a semilinear elliptic optimal control problem*, Comput. Optim. Appl., 23 (2002), pp. 201–229.
- [2] R. BECKER, D. MEIDNER, AND B. VEXLER, *Efficient numerical solution of parabolic optimization problems by finite element methods*, Optim. Methods Softw., 22 (2007), pp. 813–833.
- [3] J. H. BRAMBLE AND S. R. HILBERT, *Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolation*, SIAM J. Numer. Anal., 7 (1970), pp. 112–124.
- [4] E. CASAS, M. MATEOS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl., 31 (2005), pp. 193–220.
- [5] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.
- [6] R. DAUTRAY AND J.-L. LIONS, *Mathematical Analysis and Numerical Methods for Science and Technology: Evolution Problems I*, Vol. 5, Springer-Verlag, Berlin, 1992.
- [7] K. ERIKSSON, D. ESTEP, P. HANSBO, AND C. JOHNSON, *Computational Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

- [8] K. ERIKSSON AND C. JOHNSON, *Adaptive finite element methods for parabolic problems I: A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.
- [9] K. ERIKSSON AND C. JOHNSON, *Adaptive finite element methods for parabolic problems II: Optimal error estimates in  $L_\infty L_2$  and  $L_\infty L_\infty$* , SIAM J. Numer. Anal., 32 (1995), pp. 706–740.
- [10] K. ERIKSSON, C. JOHNSON, AND V. THOMÉE, *Time discretization of parabolic problems by the discontinuous Galerkin method*, M2AN Math. Model. Numer. Anal., 19 (1985), pp. 611–643.
- [11] L. C. EVANS, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, 2002.
- [12] R. FALK, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.
- [13] M. FEISTAUER AND K. ŠVADLENKA, *Space-Time Discontinuous Galerkin Method for Solving Nonstationary Convection-Diffusion-Reaction Problems*, Preprint MATH-knm-2005/2, Charles University, Prague, 2005.
- [14] GASCOIGNE: *The Finite Element Toolkit*, <http://www.gascoigne.uni-hd.de/>.
- [15] T. GEVECI, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, M2AN Math. Model. Numer. Anal., 13 (1979), pp. 313–328.
- [16] M. HINZE, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–61.
- [17] I. LASIECKA AND K. MALANOWSKI, *On discrete-time Ritz-Galerkin approximation of control constrained optimal control problems for parabolic systems*, Control Cybern., 7 (1978), pp. 21–36.
- [18] J.-L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Grundlehren Math. Wiss. 170, Springer, Berlin, 1971.
- [19] K. MALANOWSKI, *Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal control problems*, Appl. Math. Optim., 8 (1981), pp. 69–95.
- [20] R. S. MCNIGHT AND W. E. BOSARGE, JR., *The Ritz-Galerkin procedure for parabolic control problems*, SIAM J. Control Optim., 11 (1973), pp. 510–524.
- [21] D. MEIDNER AND B. VEXLER, *Adaptive space-time finite element methods for parabolic optimization problems*, SIAM J. Control Optim., 46 (2007), pp. 116–142.
- [22] C. MEYER AND A. RÖSCH, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim., 43 (2004), pp. 970–985.
- [23] *RoDoBo: A C++ Library for Optimization with Stationary and Nonstationary PDEs*, <http://rodoobo.uni-hd.de/>.
- [24] A. RÖSCH, *Error estimates for parabolic optimal control problems with control constraints*, Z. Anal. Anwendungen, 23 (2004), pp. 353–376.
- [25] M. SCHMICH AND B. VEXLER, *Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations*, SIAM J. Sci. Comput., 30 (2008), pp. 369–393.
- [26] D. SCHÖTZAU, *hp-DGFEM for Parabolic Evolution Problems*, Ph.D. thesis, Swiss Federal Institute of Technology, Zürich, 1999.
- [27] V. THOMÉE, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer Ser. Comput. Math. 25, Springer-Verlag, Berlin, 2006.
- [28] R. WINTHER, *Error estimates for a Galerkin approximation of a parabolic control problem*, Ann. Math. Pura Appl. (4), 117 (1978), pp. 173–206.
- [29] J. WLOKA, *Partielle Differentialgleichungen: Sobolevräume und Randwertaufgaben*, B. G. Teubner, Stuttgart, 1982.

## THE “PRINCESS AND MONSTER” GAME ON AN INTERVAL\*

STEVE ALPERN<sup>†</sup>, ROBERT FOKKINK<sup>‡</sup>, ROY LINDELAUF<sup>‡</sup>, AND  
GEERT-JAN OLSDER<sup>‡</sup>

**Abstract.** A minimizing searcher  $S$  and a maximizing hider  $H$  move at unit speed on a closed interval until the first (*capture*, or *payoff*) time  $T = \min\{t : S(t) = H(t)\}$  that they meet. This zero-sum *princess and monster game* or less colorfully *search game with mobile hider* was proposed by Rufus Isaacs for general networks  $Q$ . While the existence and finiteness of the value  $V = V(Q)$  has been established for such games, only the circle network has been solved (value and optimal mixed strategies). It seems that the interval network  $Q = [-1, 1]$  had not been studied because it was assumed to be trivial, with value  $3/2$  and “obvious” searcher mixed strategy going equiprobably from one end to the other. We establish that this game is in fact nontrivial by showing that  $V < 3/2$ . Using a combination of continuous and discrete mixed strategies for both players, we show that  $15/11 \leq V \leq 13/9$ . The full solution of this very simple game is still open and appears difficult, though many properties of the optimal strategies are derived here.

**Key words.** differential game, search game, zero-sum game, interval

**AMS subject classifications.** 91A24, 90B40

**DOI.** 10.1137/060672054

**1. Introduction.** In the final chapter of his classic book *Differential Games* [15], Rufus Isaacs introduced search games with mobile hiders, which he also called princess and monster games (see also [7], Example 1.4). A searcher (monster) and a hider (princess) move about a space  $Q$ , which we take to be a network (and later specialize to an interval). The searcher chooses as his pure strategy a path  $S = S(t)$  of known speed, which we take to be 1. He says, “We permit the princess full freedom of locomotion,” which we take to be any continuous path  $H = H(t)$ . (We will establish for the interval that she need never go faster than the searcher.) The payoff for this zero-sum game  $\Gamma(Q)$  is the capture time

$$(1) \quad T = T(S, H) = \min\{t : S(t) = H(t)\}.$$

Taking the topology of uniform convergence on compact subsets, the payoff function  $T$  is upper semicontinuous and the searcher mixed strategy space is compact Hausdorff. Consequently the minimax theorem of Alpern and Gal [4] can be used, as shown in Appendix A of [6] or [11] to establish the existence of the value  $V(Q)$ , an optimal mixed searcher strategy, and an  $\varepsilon$ -optimal hider mixed strategy. Recall that a strategy is  $\varepsilon$ -optimal if the expected payoff is at least  $V - \varepsilon$  against any strategy of the opponent. Upper bounds on  $V = V(Q)$  in terms of the structure of  $Q$  (and hence the finiteness of  $V$ ) are derived in [3]. For general networks  $Q$ , it is sometimes advantageous for the searcher to wait for a while at a node, a so-called *ambush strategy*, and these games are known to be difficult—none have been solved. So the only networks that might appear

---

\*Received by the editors October 11, 2006; accepted for publication (in revised form) August 9, 2007; published electronically March 21, 2008. This research was carried out while Steve Alpern visited Delft University on NWO visitor grant B61-590.

<http://www.siam.org/journals/sicon/47-3/67205.html>

<sup>†</sup>London School of Economics, Houghton Street, London WC2A 2AE, UK (alpern@lse.ac.uk).

<sup>‡</sup>Delft Institute of Applied Mathematics, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands (r.j.fokkink@tudelft.nl, r.h.lindelauf@student.tudelft.nl, g.j.olsder@tudelft.nl).

possible to solve are those with no nodes (of degree greater than 2)—namely, the circle and the interval. The game  $\Gamma(Q)$  when  $Q$  is a circle was indeed a problem suggested by Isaacs [15, Example 12.4.2] and was solved a long time ago [20], [1]. The solution, for both players, is the *cohatu* strategy: start randomly (uniform distribution); flip a coin; if head (tails) go to antipodal point half way around circle clockwise (anticlockwise), at unit speed. No one seems to have considered this problem for the other network without nodes, the interval. It seems to have generally been believed that the game on the interval was trivial. The searcher should simply start at a random end and go directly to the other end. (Against this, the hider waits at 0 until time  $1 - \epsilon$ , then goes equiprobably to either end.)

It should be noted that this “simple” search strategy, of starting at a random end and going to the other, is indeed optimal in the related *search game with immobile hider*, also introduced by Isaacs [15, section 12.3]. Indeed, for trees [9] and trees with Eulerian networks attached [2], an optimal search strategy is to traverse a Chinese postman path (minimal covering path) equiprobably in either direction [10, 12]. Of course this is obvious for the interval, where the hider can hide uniformly or equiprobably at the ends (or many other optimal strategies—the full class has not been determined). The value is 1 for this game on an interval of length 2. A related problem on the interval, also with two mobile agents (they both have unit speed), was analyzed by Howard [14] and by Chester and Tutuncu [8]. In this rendezvous version of the problem, both players wish to *minimize* the meeting time  $T$ . Related problems have been analyzed in [13, 16, 18].

The paper is organized as follows. In section 2 we establish some elementary lemmas which restrict the strategies we need to consider in the rest of the paper. In section 3 we show that the game  $\Gamma(I)$  is not trivial by using some finitely supported mixed strategies to obtain estimates on its value  $V$ . In section 4 we obtain the bound  $V \leq 13/9$  by using a continuous mixed searcher strategy. In section 5 we obtain the bound  $V \geq 15/11$  by using a continuous mixed hider strategy.

**2. Properties of optimal strategies.** In this section we present results which restrict the strategies (pure and mixed) that we will need to consider in the remainder of the paper. We begin by noting that the pure strategy space for the hider is the space  $\mathcal{H}$  consisting of all continuous paths  $H : [0, \infty) \rightarrow I = [-1, 1]$ . For the searcher the pure strategy space  $\mathcal{S}$  consists of all paths in  $\mathcal{H}$  with Lipschitz constant 1, that is

$$(2) \quad \mathcal{S} = \{S : [0, \infty) \rightarrow I ; |S(t) - S(t')| \leq |t - t'| \quad \forall t, t' \geq 0\}.$$

If a searcher chooses paths that do not cover the entire interval, then hiding at some end gives a payoff that is infinite. This is absurd, so we may assume that searcher paths are onto. We show that once the searcher reaches an end, he should go directly to the other end, but if the hider reaches an end, he should stay there. The hider, while unrestricted in speed, need never go faster than speed 1 (the searcher’s maximum speed). We show that both players can optimally use mixed strategies which respect the symmetry of the interval, that is, the reflection  $\phi(x) = -x$ , and that in an optimal hider mixed strategy, the pure strategies do not intersect. Finally, we give some properties of optimal response searcher strategies.

We say that a pure strategy  $S$  is *end-reflecting* if whenever  $S(t_0) = \pm 1$ , we have  $|S(t) - S(t_0)| = t - t_0$  for  $t_0 \leq t \leq t_0 + 1$ . We say that a pure strategy  $H$  is *end-absorbing* if  $H(t_0) = \pm 1$  implies  $H(t) = H(t_0)$  for  $t \geq t_0$ . We say that a mixed strategy (for either player) is *symmetric* if it is invariant under the reflection  $\phi(x) = -x$ .



The first three lemmas concern pure strategies which we may ignore in our subsequent analysis because they are dominated.

**PROPOSITION 1.** *Every pure searcher strategy  $S \in \mathcal{S}$  is dominated by one which is end-reflecting.*

*Proof.* Suppose  $S$  is not end-reflecting and reaches, say,  $+1$  at first time  $t_0$ . Define  $S^*$  as  $S$  up to time  $t_0$  and then equal to  $1+t_0-t$ . Consider any  $H \in \mathcal{H}$ . If  $T(S, H) \leq t_0$ , then  $T(S^*, H) = T(S, H)$ . If  $T(S, H) = t_1 > t_0$ , then  $H(t_1) - S^*(t_1) \geq 0$ , and since  $H$  is continuous and  $H(t_0) - S^*(t_0) = H(t_0) - 1 \leq 0$  the intermediate value theorem implies that  $T(S^*, H) = t_2$  for some  $t_2$  with  $t_0 \leq t_2 \leq t_1$ . Thus in all cases the end-reflecting strategy  $S^*$  satisfies  $T(S^*, H) \leq T(S, H)$ .  $\square$

**PROPOSITION 2.** *Every pure hider strategy  $H$  is dominated by one which is end-absorbing.*

*Proof.* Assume that  $H$  is not end-absorbing and arrives at, say,  $+1$  at first time  $t_0$ . Let  $H^*$  be the end-absorbing strategy that agrees with  $H$  up to  $t_0$  and then stays at  $+1$ . Consider any  $S \in \mathcal{S}$  and assume, as we may, that  $T(S, H^*) = t_1 > t_0$ . Then  $S(t_1) = 1 \geq H(t_1)$  and  $S(t_0) < 1 = H(t_0)$  so the searcher meets the hider  $H$  between  $t_0$  and  $t_1$  by the intermediate value theorem. It follows that  $T(S, H) \leq T(S, H^*)$  for any pure searcher strategy.  $\square$

We say that a continuously differentiable function is *smooth*. By reasons that will become clear below, we may restrict the pure hider strategies to any subset that is dense in  $\mathcal{S}$ . In particular, we may consider smooth paths only, without changing the value of the game.

**PROPOSITION 3.** *Every smooth hider strategy in  $\mathcal{H}$  is dominated by one in  $\mathcal{S}$ , that is, one with speed bounded by 1.*

*Proof.* Essentially, the idea is that if  $H$  is any smooth hider and if  $H^*$  is a hider that follows  $H$  but has speed bounded by 1, then  $H^*$  cannot be caught from behind since the searcher has the same speed limit. Let  $t_0 = \inf\{t \in [0, \infty): |H'(t)| > 1\}$  and assume, as we may, that  $t_0$  is finite. Define  $H^*(t) = H(t)$  for  $t \leq t_0$ . For  $t > t_0$  the hider  $H^*$  continues to move at speed 1. Since the interval is bounded,  $H^*$  meets  $H$  at some time  $\tau > t_0$ . Now let  $t_1 = \inf\{t \in [\tau, \infty): |H'(t)| > 1\}$  and repeat the construction inductively.

Suppose that a searcher  $S$  finds  $H$  at a time  $T = T(S, H)$  when the hiders  $H$  and  $H^*$  are in different locations. So suppose that  $t_0 < T < \tau$ . By symmetry, we may assume as well that  $H'(t_0) = +1$ . Then  $H^*$  has velocity  $+1$  and  $H^*(t) < H(t)$  for all  $t \in (t_0, t_1)$ . Since  $S(T) = H(T) > H^*(T)$  and since the searcher moves with bounded speed,  $S(t) > H^*(t)$  for all  $t \in (t_0, T)$ . This implies that  $T(S, H^*) > T(S, H)$ .  $\square$

From now on, we shall only consider hider paths of speed  $\leq 1$ . We say that a hider or a searcher runs at time  $t$  if  $|H'(t)| = 1$  or  $|S'(t)| = 1$ , respectively. We now consider mixed strategies, i.e., probability measures on the Borel  $\sigma$ -algebra of  $\mathcal{S}$ .

**PROPOSITION 4.** *There is an optimal searcher mixed strategy and (for any  $\varepsilon$ ) an  $\varepsilon$ -optimal hider mixed strategy, which are invariant under the reflection  $\phi(x) = -x$ .*

*Proof.* The proof is just a special case of Theorem 3 of Alpern and Asic [3], where the existence of such strategies invariant under the isometry group (distance-preserving homeomorphisms) of a network  $Q$  is established. In the case of  $Q = I$ , this group consists just of the identity and  $\phi$ .  $\square$

The capture time  $T(S, H)$  is upper semicontinuous as a function on the pure strategies  $H \in \mathcal{H}$  and  $S \in \mathcal{S}$ . This implies that for any  $\varepsilon$  there exists a  $\varepsilon$ -optimal finite mixed strategy and in principle it is possible to determine the value of the game by considering finite mixed strategies only.

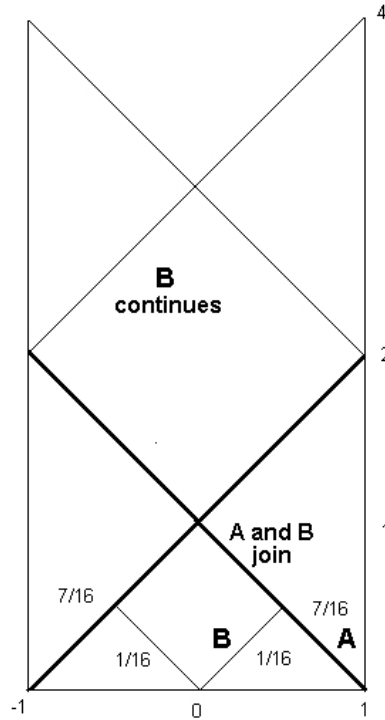


FIG. 1. The searcher strategy for  $V < 47/32$  in a space–time diagram.

**PROPOSITION 5.** Suppose the hider is using a mixed strategy concentrated on a finite number of pure strategies  $H_j \in \mathcal{S}$ ,  $j = 1, \dots, J$ . Then the optimal response by the searcher is concentrated on a finite number of strategies as well. In particular, the searcher picks a random permutation  $\sigma$  of  $1, \dots, J$  and runs towards  $H_{\sigma(j+1)}$  as soon as he has met  $H_{\sigma(j)}$ .

*Proof.* Let  $S$  be any optimal response to the hider mixed strategy. The searcher  $S$  meets the  $H_j$  in a certain order and this defines a permutation  $\sigma$  of  $1, \dots, J$ . Now it is obviously optimal for the searcher to run toward  $H_{\sigma(j+1)}$  after he has met  $H_{\sigma(j)}$ .  $\square$

This result is a simplified version of a similar observation for rendezvous on the line due to Alpern and Gal [5] (repeated as Theorem 16.10 of [6]). It follows that we can restrict our attention to pure strategies in which the searcher runs all the time. In a space–time diagram  $[-1, 1] \times [0, \infty)$ , such search paths are depicted as broken lines of slope  $\pm 1$  with finitely many turning points; see Figure 1.

**DEFINITION 6.** A pair of pure hider strategies  $H_1, H_2$  is called noncrossing if (possibly after reordering) we have  $H_1(t) \leq H_2(t)$  for all  $t \geq 0$ . If the inequality holds strictly, we say that they are nonintersecting.

For pure hider strategies  $H_1$  and  $H_2$ , define new pure strategies  $H_1 \wedge H_2(t) = \min\{H_1(t), H_2(t)\}$  and  $H_1 \vee H_2(t) = \max\{H_1(t), H_2(t)\}$ . Obviously,  $H_1 \wedge H_2$  and  $H_1 \vee H_2$  are noncrossing.

**PROPOSITION 7.** The hider strategy that mixes two pure strategies  $H_1, H_2$  with equal probability is dominated by the noncrossing hider strategy that mixes  $H_1 \wedge H_2, H_1 \vee H_2$  with equal probability. Consequently, any finite mixed hider strategy may be assumed to consist of noncrossing pure strategies.

*Proof.* Note that at for all  $t$  the sets  $\{H_1(t), H_2(t)\}$  and  $\{H_1 \wedge H_2(t), H_1 \vee H_2(t)\}$  are the same. So if  $S$  catches the first of the two original hiders  $H_1, H_2$ , then at the same time he catches the first of the noncrossing hiders  $H_1 \wedge H_2, H_1 \vee H_2$ . Denote this time by  $t_1$ . By renumbering indices or reflecting the interval we may assume that  $S(t_1) = H_1(t_1)$  and that  $H_1(t_1) \leq H_2(t_1)$ . Now under these assumptions,  $H_2$  and  $H_1 \vee H_2$  are in the same location at time  $t_1$  and the searcher is to their left. Since  $H_1 \vee H_2(t) \geq H_2(t)$  for all  $t$ , the searcher cannot catch  $H_1 \vee H_2$  before he catches  $H_2$ .  $\square$

It follows from Proposition 4 that there exist noncrossing  $\varepsilon$ -optimal hider strategies that are symmetric. Any finite collection of noncrossing paths can be approximated arbitrarily closely by a collection of nonintersecting paths. So, there exist  $\varepsilon$ -optimal mixed hider strategies that are finite, symmetric and nonintersecting.

**PROPOSITION 8.** *Any pure strategy  $H$  in a nonintersecting symmetric hider strategy is contained in half of the interval, that is,  $H(t) \in [-1, 0]$  or  $H(t) \in [0, 1]$  for all  $t$ .*

*Proof.* If the pure strategy  $H$  is used in a symmetric mixed strategy, then so is  $-H$ . If  $H(t) = 0$ , then  $H$  and  $-H$  intersects  $-H(t)$ , but the mixed strategy is nonintersecting. So either  $H(t) \neq 0$  for all  $t$  or  $H$  is immobile and remains in 0. In both cases,  $H$  is contained in half of the interval.  $\square$

**3. The interval game is not trivial.** The interval game has some fairly obvious strategies for each player that appear like they might be optimal. If any of these were indeed optimal, we would consider the game to be trivial. The purpose of this section is to show that none of these strategies are in fact optimal. For the searcher, the obvious strategy is to start at a random end and run to the other end; this gives an estimate  $V \leq 3/2$  and the bound can be obtained if the hider waits at the center until time  $1 - \varepsilon$  and then runs to a random end. For the hider, the two stationary strategies may be considered that are optimal in the immobile version of the game: one of these is to wait at a random end; the other is to wait randomly along the interval. Both guarantee  $V \geq 1$ , where 1 is the value of the immobile hider game. We show that the game is not trivial by exhibiting fairly simple strategies establishing that

$$1 < \frac{97}{75} < V < \frac{47}{32} < \frac{3}{2}.$$

Suppose that the hider starts out with the strategy of hiding at an end point  $E(t) = 1$  and symmetrically  $-E(t) = -1$ . According to Lemma 5 the optimal response of the searcher is to start at an end point and run to the other end  $A(t) = 1 - t$ , or symmetrically  $-A(t) = -1 + t$ . The strategies  $A$  and  $-A$  are what Howard [14] called the *sweepers* in his rendezvous version. The expected meeting time for this optimal response is 1 and this is a lower bound on  $V$ . Similarly, suppose that the searcher adopts the sweeper strategies  $A$  and  $-A$ . Against any hider path  $H$  either  $V(A, H) \leq 1$  or  $V(-A, H) \leq 1$ . So if the searcher adopts the sweeping strategy, then the payoff is  $\leq 3/2$  against any mixed hider strategy and this puts an upper bound on  $V$ . The game would be trivial if  $V = 1$  or  $V = 3/2$ , but it is not. In this section we show that  $1 < V < 3/2$ . It is  $\varepsilon$ -optimal against the sweeping strategy to loiter around 0 until time  $1 - \varepsilon$  and then run to one of the end points. The searcher can ambush such loiterers by adding a search path that patrols the center. More specifically, strategy  $B$  starts at 0; runs to the left; joins sweeper  $-A$  from the time  $(1/2)$  when he meets him, until reaching  $-1$ ; then (by Lemma 1) goes to  $-1$ . Strategies  $\pm A$  are each used with probability  $7/16$ , while  $\pm B$  are each used with probability  $1/16$ . The searcher strategies  $\pm A, \pm B$  are drawn in a space-time diagram in Figure 1.

We show that this mixed strategy ensures a meeting time less than  $3/2$ , though admittedly not much less.

THEOREM 9.  $V \leq \frac{47}{32} = 1.4688$ .

*Proof.* Consider the mixed strategy  $s$  in which the searcher uses  $\pm A$  each with probability  $7/16$  and  $\pm B$  each with probability  $1/16$ . Let  $H$  be *any* hiding strategy. Let  $P(t)$  denote the probability that  $T \leq t$ . There are two cases: (1)  $|H(1/2)| \leq 1/2$ , and (2)  $|H(1/2)| > 1/2$ .

- (i) In this case  $P(1/2) \geq 1/16$  (because  $B$  or  $-B$  has been met) and  $P(1) \geq 8/16$  (because also  $A$  or  $-A$  has been met). Furthermore  $P(2) = 1$ . Thus

$$T \leq \frac{1}{16} \cdot \frac{1}{2} + \frac{7}{16} \cdot 1 + \frac{1}{2} \cdot 2 = \frac{47}{32}.$$

- (ii) By symmetry we may assume  $H(1/2) > 1/2$ . Then  $P(1/2) \geq 7/16$  ( $A$  has been met),  $P(2) \geq 15/16$  ( $A, -A, B$  have been met), and  $P(4) = 1$  (all met). Hence

$$(3) \quad T \leq \frac{7}{16} \cdot \frac{1}{2} + \frac{8}{16} \cdot 2 + \frac{1}{16} \cdot 4 = \frac{47}{32}. \quad \square$$

We now consider the lower bound on  $V$ . In the search game with an immobile hider, the hider has two particular mixed strategies that guarantee him an expected capture time of half the length of the interval: (1) Hide equiprobably at the ends (end-hiding is optimal on trees for such games [9], and symmetry implies the equiprobability), or (2) hide uniformly (this is optimal for all networks with Eulerian paths [2]). Mobility usually helps the hider, for example, when  $Q$  is the circle of circumference  $c$  an immobile hider can be found (by a random tour) in mean time  $c/2$ , while a mobile hider can be found with best play on both ends in time  $3c/4$  [1]. The intuitive explanation is usually that when the hider is immobile, the searcher does not have to search again any part of  $Q$  already searched (and hence can employ a minimal, Chinese postman, search path)—but a *mobile* hider might not be met by such a path. However this explanation does not apply to the interval (though to all other networks), since a Chinese postman path on an interval will indeed find a mobile hider—it has search number 1 in the sense of Parsons [17].

Mobility helps the hider, since  $V > 1$  if the hider is mobile. To prove that this is true, we select hider strategies by considering the searcher strategies  $A$  and  $B$ . The optimal hider strategy against  $A, B$  is to loiter around  $\pm \frac{1}{2}$  and just before time  $\frac{1}{2}$  run either to the middle and back, or to the end. These two possible paths  $G, H$  and their symmetric counterparts are depicted in thick lines in the left-hand diagram in Figure 2 (the paths do not cross the center). We combine these with the two other hider strategies that we considered above:  $E$ , hiding at an end, or  $F$ , loitering in the center  $F(t) = \max\{0, t + \varepsilon - 1\}$ . Then we get a mixed strategy in which the hider uses  $\{E, F, G, H\}$  and their symmetric counterparts. According to Lemma 5 the searcher adopts strategies that start at  $0, \pm \frac{1}{2} \pm 1$  and then run between hider paths. If the searcher starts out from an end, then by Proposition 1 it is optimal to adopt strategy  $A$ . If he starts out in  $0$ , then the searcher runs to  $\pm \frac{1}{2}$  at which point he may turn, “strategy  $B$ ,” or continue to an end and run back, “strategy  $M$ ” in Figure 2. Starting from  $\pm \frac{1}{2}$  the searcher either runs to the near end, “strategy  $C$ ,” or to the remote end, “strategy  $D$ .” So the optimal optimal response to a mixed hider strategy on  $G$  and  $H$  is a mixed searcher strategy on  $\{A, B, C, D, M\}$  and their symmetric counterparts.

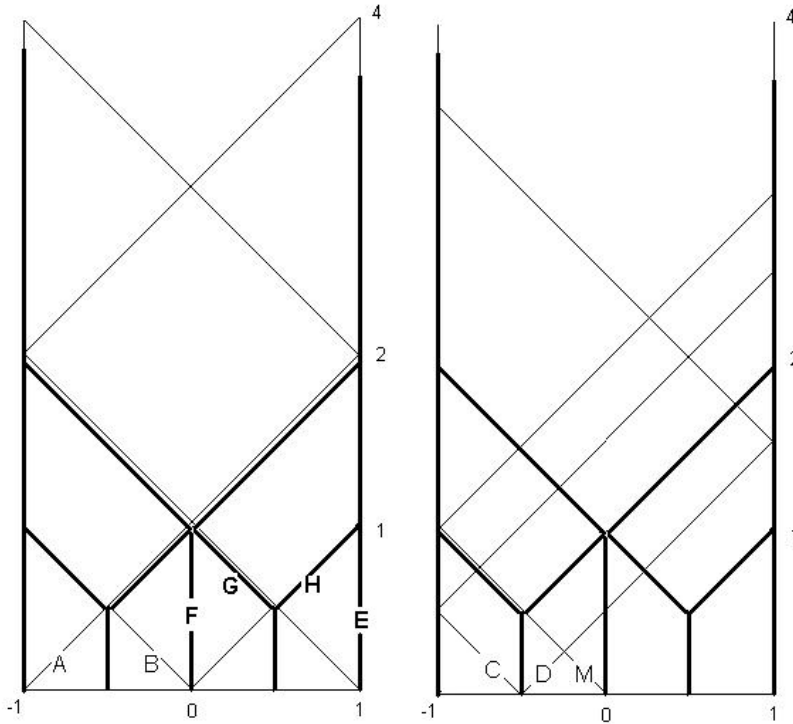


FIG. 2. *Hide paths are bold. Left: the pure paths in an optimal mixed hide strategy against A, B in a space-time diagram are E, F, G, H. Right: the searcher's response to {E, F, G, H}. Apart from A and B the only relevant paths, up to symmetry, are C, D, M.*

Ignoring  $\varepsilon$  these strategies give the zero-sum game matrix:

$$(4) \quad \begin{array}{c} A \\ B \\ C \\ D \\ M \end{array} \begin{array}{c} E \quad F \quad G \quad H \\ \left[ \begin{array}{cccc} 1 & \frac{3}{2} & \frac{3}{2} & \frac{5}{4} \\ 3 & 0 & \frac{5}{4} & 3 \\ \frac{3}{2} & \frac{15}{8} & \frac{5}{4} & \frac{5}{4} \\ \frac{5}{2} & \frac{1}{2} & \frac{3}{8} & \frac{3}{4} \\ 2 & 0 & \frac{7}{4} & 2 \end{array} \right] \end{array}$$

The value of this matrix game  $\frac{97}{75}$  puts a lower bound on  $V$ .

THEOREM 10.  $V \geq \frac{97}{75} = 1.29\bar{3}$ .

In principle, it is possible to compute the value of the game in the following iterative manner. Give finitely many pure strategies to the hider; according to Proposition 5 the searcher needs finitely many pure strategies to optimize his response to the hider. From the space-time diagram it is obvious that the hider can now optimize his response to the searcher by a finite number of additional pure strategies, etc. In this way we get an ever-increasing set of pure strategies for the hider and the searcher, but the convergence is slow and the increase is exponential. The convergence to the value of the game appears to be slow and not effectively computable.

**4. A searcher strategy with a continuous initial distribution.** To improve on the upper bound  $V < 47/32$  of Theorem 9, we extend the mixed searcher strategy which uses  $\{A, B\}$ . We replace the pure strategy  $B$  by a continuous mixed strategy  $s_\Phi$ . In this strategy  $s_\Phi$  the searcher picks a point  $x$  according to a continuous distribution function on the interval  $\Phi(x)$  and runs to the right until he meets the sweeper  $A$ , then joins the sweeper until he reaches  $-1$  and runs back to the other end. The symmetric strategy  $-s_\Phi$  starts according to  $\Phi(-x)$  and the searcher runs to the left until he meets the sweeper  $-A$ , etc.

LEMMA 11. *Suppose that the searcher uses the mixed strategy  $s_\Phi$ . Let  $H$  be a pure hider strategy and let  $y = y(H)$  be the first time that the hider meets a sweeper. Then the searcher finds the hider before time  $y$  if and only if he starts in  $(H(y) - y, H(0)]$  and runs to the right, or if he starts in  $[H(0), H(y) + y)$  and runs to the left.*

*Proof.* By Proposition 8 we may assume that  $H \geq 0$ , so  $H$  meets the right sweeper  $A$  first and  $H(y) = 1 - y$ . We consider only the case that the searcher  $S$  initially runs to the right until he meets  $A$ . If  $S$  runs to the left the argument is the same. Suppose that  $S$  starts in  $(1 - 2y, H(0)]$ , so  $S$  meets  $A$  at time  $t = \frac{1-S(0)}{2} < y$ . Then  $S(0) < H(0)$  and  $S(t) > H(t)$  since  $H(t) < A(t)$  for  $t < y$ . Therefore  $S$  finds  $H$  in between time 0 and time  $t$ . Now suppose that  $S$  does not start in the interval  $(1 - 2y, H(0)]$ , so either  $S(0) > H(0)$  or  $S(0) < 1 - 2y$ . In the first case,  $S$  runs toward  $A$  and finds the hider at time  $y$ . In the second case, the searcher cannot meet  $A$  at time  $y$ , while the hider can. So the paths of  $S$  and  $H$  cannot cross before time  $y$ .  $\square$

LEMMA 12. *Let  $f = \Phi'$  be the probability density. Searchers that start in  $(H(y) - y, H(0)]$  and run to the right catch the hider with expected time*

$$(5) \quad \int_0^y t f(H(t) - t) (1 - H'(t)) dt.$$

*Searchers that start in  $[H(0), H(y) + y)$  and run to the left catch the hider with expected time*

$$(6) \quad \int_0^y t f(-H(t) - t) (1 + H'(t)) dt.$$

*Proof.* Consider a small time interval  $[t, t + \Delta t]$  when the hider moves from  $H(t)$  to  $H(t) + \Delta H$ . Searchers that meet the hider in that time interval and that have started out from the right have started in  $[H(t) - t + \Delta H - \Delta t, H(t) - t]$ . Here we use that  $|\Delta H| \leq \Delta t$  since we may assume that a hider moves at no greater speed than 1. The probability of a hider starting out in that interval is  $f(H(t) - t)(\Delta H - \Delta t)$  up to first order. The time of capture is  $t$  up to first order. By taking the limit of  $\Delta t \rightarrow 0$  we obtain the integral in (5). By symmetry we obtain (6).  $\square$

If a hider has chosen  $x = H(0)$  and  $y$ , then he maximizes the expected time of capture, which comes down to maximizing

$$(7) \quad \int_0^y t [f(H(t) - t) (1 - H'(t)) + f(-H(t) - t) (1 + H'(t))] dt.$$

This integral can be simplified by partial integration, which gives the sum of a constant  $-y\Phi(1 - 2y)$  and the integral in (8).

LEMMA 13. *Let  $y$  be the first time that the hider meets a sweeper. The optimal hider path from  $H(0)$  to  $H(y)$  maximizes*

$$(8) \quad \int_0^y \Phi(-t + H(t)) + \Phi(-t - H(t)) dt.$$

This is a variational problem. Its Euler–Lagrange equation is  $f(-t + H(t)) = f(-t - H(t))$ , where as before  $f$  is the probability density. A stationary value is  $H(t) = 0$  and this makes sense, since  $s_\Phi$  is designed against loitering hiders and  $H(t) = 0$  is a minimum. If  $\Phi$  is the uniform distribution, then the Euler–Lagrange equation is satisfied by any path and the integral does not depend on  $H$ : it is equal to  $y(1 - y/2)$ .

The integral in (8) represents the payoff against the searchers that start in  $(H(0) - y, H(0) + y)$  and run toward  $H(0)$ . Once the hider has met a sweeper, he should run to the end since if the searcher uses  $s_\Phi$  he joins the sweeper. So we can determine the payoff  $V(s_\Phi, H)$ . Denote  $x = H(0)$ ; then this payoff is

$$(9) \quad 1 - \Phi(-x) + 2\Phi(1 - 2y) + \frac{y}{2}(1 - \Phi(x)) - \frac{y}{2}\Phi(1 - 2y) + \frac{1}{2} \int_0^y \Phi(-t + H(t)) + \Phi(-t - H(t)) dt.$$

This equation is derived as follows. A searcher starts out left from  $x$  and runs to the left with probability  $(1 - \Phi(-x))/2$ . Such a searcher catches  $H$  at time 2. This gives the first term. A searcher starts out left from  $1 - 2y$  and runs to the right with probability  $\Phi(1 - 2y)/2$ . Such a searcher catches  $H$  at time 4 and this gives the second term. A searcher starts out right from  $x$  and runs right with probability  $(1 - \Phi(x))/2$ . Such a searcher catches  $H$  at time  $T$ . This gives the third term. The fourth term  $-y\Phi(1 - 2y)$  turns up in the partial integration and the final term is the variational integral.

We consider a mixed strategy  $\sigma$  for the searcher, as follows. Use the sweeper strategy  $\pm A$  with probability  $p$  and use the continuous mixed strategy  $s_\Phi$  with probability  $1 - p$ . If  $\Phi$  is the uniform distribution then (9) is equal to

$$(10) \quad \frac{10 + 2x - (7 + x)y + y^2}{4},$$

which is maximal at  $x = 1$  for any  $0 \leq y \leq 1$  and decreasing in  $y$ . If the searcher takes  $p = 7/9$ , then  $V(\sigma, H) = 13/9 - y/18 + y^2/18$  which is maximal at  $y = 0$  and  $y = 1$ . So against  $\sigma$ , the hider optimizes in either of two ways: stick to an end, or run from an end to the middle and back. The payoff is  $13/9$ , which puts an upper bound on the value of the game that is sharper than the bound  $47/32$  that we found in the previous section. We summarize this in a theorem.

**THEOREM 14.** *If the searcher uses the mixed strategy  $\sigma$ , then the optimal response of the hider gives a matrix game with value  $13/9$ . In particular  $V \leq 13/9$ .*

Now the obvious way to try to improve on this bound is by varying the distribution  $\Phi$ . Our computer experiments indicated that the bound of  $13/9$  can be improved only marginally in this way. To prove that there is only room for marginal improvement, we consider a specific noncrossing hider strategy:  $H_x$  starts in  $H(0) = x \geq 0$  and runs toward the sweeper  $A$ ; turning just  $\varepsilon$  in front of the sweeper; then  $H_x$  runs back toward the middle but turns once again at time  $y$ , before reaching the middle, to run to the end. So, ignoring  $\varepsilon$  we can describe this path by  $H_x(t) = x + t$  if  $t \leq (1 - x)/2$  and  $H_x(t) = 1 - t$  if  $(1 - x)/2 \leq t \leq y$  for  $x \geq 0$ . The variational integral (8) over this path is

$$(11) \quad \frac{1}{2} \left( \Phi(x)(1 - x) + \int_0^{-x} \Phi(t) dt + \int_{1-2y}^x \Phi(t) dt \right)$$

Suppose that  $y = 1$ . Then the payoff  $V(H_x, A)$  against the sweeper strategy is  $3/2$ . Against the continuous strategy it is

$$(12) \quad V(H_x, s_\Phi) = \frac{3}{2} - \Phi(-x) - \Phi(x) \frac{(1+x)}{4} + \frac{1}{4} \int_{-1}^{-x} \Phi(t) dt + \frac{1}{4} \int_{-1}^x \Phi(t) dt.$$

Suppose that the searcher uses a mixed strategy  $\sigma_\Phi$  on  $\{A, s_\Phi\}$ . Suppose that the hider uses a mixed strategy  $\gamma$  on  $\{E, H_1, H_{\frac{1}{2}}, H_0\}$ , where  $E$  is the end point strategy. The integrals in (12) can be bounded from below by finite sums such as  $\int_{-1}^0 \Phi(t) dt \geq \frac{1}{2}\Phi(-\frac{1}{2}) + \frac{1}{2}\Phi(0)$ . Bounding the integrals in this way we get a  $4 \times 2$  zero-sum matrix game, the value of which is a lower bound on  $V(\gamma, \sigma_\Phi)$ :

$$\begin{bmatrix} 1 & \frac{3}{2} & \frac{3}{2} & \frac{3}{2} \\ 3 & 1 + \frac{1}{8}\Phi(-\frac{1}{2}) + \frac{1}{8}\Phi(0) + \frac{1}{8}\Phi(\frac{1}{2}) & \frac{3}{2} + \frac{1}{4}\Phi(-\frac{1}{2}) - \frac{5}{4}\Phi(0) & \frac{3}{2} - \frac{7}{8}\Phi(-\frac{1}{2}) + \frac{1}{8}\Phi(0) - \frac{3}{8}\Phi(\frac{1}{2}) \end{bmatrix}.$$

By linear programming we find that the value of the matrix game is  $\frac{337}{237} = 1.4219\dots$ , which is only marginally smaller than  $13/9$ .

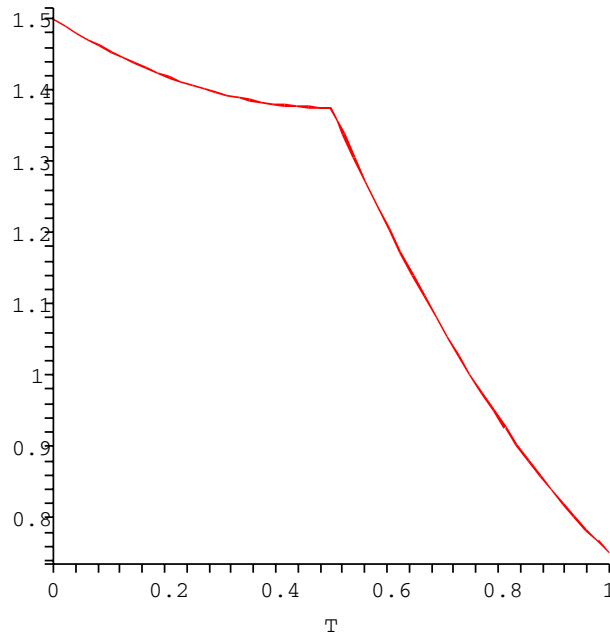
**THEOREM 15.** *If the searcher uses a mixed strategy  $\sigma_\Phi$  on  $\{A, s_\Phi\}$  for any distribution  $\Phi$ , then the hider can respond by a strategy  $\gamma_\Phi$  such that  $V(\sigma_\Phi, \gamma) > 1.421$ .*

**5. A hider strategy with a continuous initial distribution.** To improve on the lower bound  $V > 97/75$  of Theorem 10, we extend the mixed hider strategy which uses  $\{E, F, G\}$ . We replace the loitering strategies  $F$  and  $G$  by a continuous mixed strategy  $h_\Phi$ . In this strategy the hider picks a point  $x \in [\varepsilon, 1 - \varepsilon]$  according to a distribution function  $\Phi(x)$  and he waits at  $x$  until the sweeper  $A$  is  $\varepsilon$  close. Then the hider runs to  $\varepsilon$ , turns, and runs back to 1. In the symmetric strategy  $-h_\Phi$  the hider picks a point in  $[-1 + \varepsilon, -\varepsilon]$ .

It follows from Proposition 5 that the searcher  $S$  either starts at an end, in which case the  $S$  is a sweeper, or starts in  $[-1 + \varepsilon, -\varepsilon] \cup [\varepsilon, 1 - \varepsilon]$ . Let  $y$  be the first time that  $S$  gets  $\varepsilon$  close to a sweeper. By symmetry we may assume that this sweeper is  $A$ . Since  $A$  runs all the time  $S(t) < A(t)$  for  $t < y$ . We conclude that  $S$  approaches  $A$  from the left, so up to time  $y$  he catches hidens that have remained immobile. Clearly, the searcher should maximize the interval  $[S(0), S(y)]$  to catch as many immobile hidens as possible. So  $S$  starts in  $1 - 2T - \varepsilon$  and runs to  $1 - T - \varepsilon$ . At time  $y$  the searcher catches all the loitering hidens that have started out from  $x > S(0)$  and are running  $\varepsilon$  in front of  $A$ . The searcher now effectively has two possibilities, and we leave it to the reader to check this: either he turns and runs to  $-1$  and back to  $+1$ , let's call this  $S_1$ , or he continues and runs to  $+1$  and then back to  $-1$ , let's call this  $S_2$ . Against the end point strategy the payoffs are  $V(S_1, E) = 3$  and  $V(S_2, E) = 1 + 2T$ . Ignoring  $\varepsilon$  the payoffs of these two strategies against  $h_\Phi$  are

$$\begin{aligned} V(S_1, h_\Phi) &= (1 - \Phi(2y - 1)) + \frac{1}{2} \int_0^{2y-1} f(t) dt \\ &\quad + \frac{y(1 - \Phi(1 - y)) + \Phi(1 - 2y) + \int_{1-2y}^{1-y} f(t) dt}{2}, \\ V(S_2, h_\Phi) &= (1 + y)(1 - \Phi(2y - 1)) + \frac{1}{2} \int_0^{2y-1} f(t) dt \\ &\quad + \frac{y(1 - \Phi(1 - y)) + (1 + y)\Phi(1 - 2y) + \int_{1-2y}^{1-y} f(t) dt}{2}, \end{aligned}$$



FIG. 3.  $V(S_1, h_\Phi)$  as a function of  $y$ .

where, as before,  $f$  denotes the density of  $\Phi$ . To see why this is true, note that the first two terms in  $V(S_1, h_\Phi)$  concern loitering hidiers that start out from  $x < 0$ : the term first represents the hidiers that start from  $x < 0$  and that are found at the end point  $-1$ ; the second term represents the hidiers that start from  $1 - 2y < x < 0$ , which are caught before time  $y$ . The third term represents hidiers that start from  $x > 0$ . In the same way we obtain  $V(S_2, h_\Phi)$ . Both payoffs are functions of  $y$ .

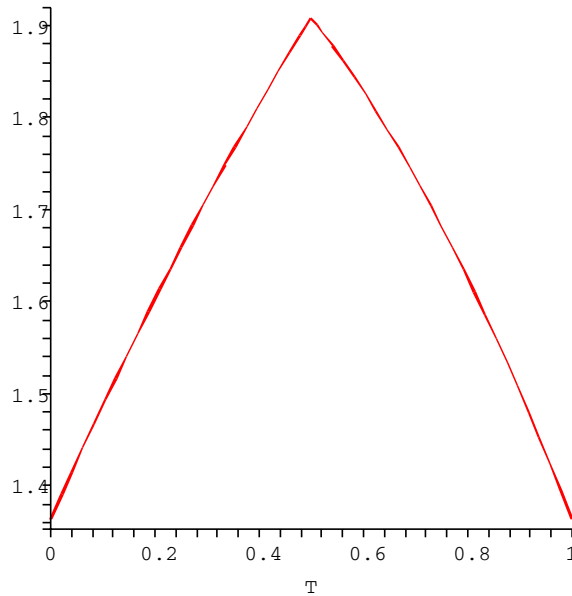
We simplify the analysis and consider only the case that  $\Phi$  is the uniform distribution. Let  $\gamma$  be the mixed strategy in which the hider uses  $\{E, h_\Phi\}$  with  $\Phi$  equal to the uniform distribution.

Since  $V(S_1, E)$  does not depend on  $T$  it is optimal for the searcher to choose  $y$  such that  $V(S_1, h_\Phi)$  is minimal.

**THEOREM 16.** *If the hider uses the mixed strategy  $\gamma$ , then the optimal response of the searcher gives a matrix game with value  $15/11$ . In particular  $V \geq 15/11$ .*

*Proof.* Since  $V(S_1, E)$  does not depend on  $y$ , the searcher should choose  $y$  such that  $V(S_1, h_\Phi)$  is minimal. As is shown in Figure 3, the minimum is at  $y = 1$ . So the searcher  $S_1$  runs  $\varepsilon$  ahead of the sweeper and the payoff is  $\frac{3}{4}$ . If the searcher only uses a mixed strategy on  $\{A, S_1\}$ , then we get a  $2 \times 2$  matrix game  $\begin{bmatrix} 1 & 3/2 \\ 3 & 3/4 \end{bmatrix}$  which has value  $15/11$ . In this game it is optimal for the hider to choose the end point strategy  $E$  with probability  $\frac{3}{11}$  and the loitering strategy  $h_\Phi$  with probability  $\frac{8}{11}$ . It turns out that the searcher cannot decrease the value of the game by including the strategy  $S_2$ :  $\frac{3}{11}V(S_2, E) + \frac{8}{11}V(S_2, h_\Phi) \geq \frac{15}{11}$  as illustrated in Figure 4. So if the hider uses the mixed strategy  $\gamma$  in which he chooses  $E$  with probability  $\frac{3}{11}$  then the searcher cannot do better than expected capture time  $15/11$ .  $\square$

**6. Conclusions.** This paper introduces the apparently easy problem of how best to search for a mobile hider who is restricted to a known interval. The existence of

FIG. 4.  $\frac{3}{11}V(S_2, h_\Phi) + \frac{8}{11}V(S_2, h_\Phi)$  as a function of  $y$ .

a Value for this game, and of  $\varepsilon$ -optimal strategies for the searcher and the hider, follows from a minimax theorem by Alpern and Gal. However, the determination of the value  $V$ , much less optimal strategies, seems difficult. The problem has resisted our attempts to solve it, but we have made significant progress in that direction. We have established many properties of optimal searcher and hider paths, that is, those that can be used in optimal mixed strategies. We have established bounds  $15/11 \leq V \leq 13/9$  on the value of the game by developing a variational theory that can be used to evaluate certain mixed strategies which start according to a continuous distribution on the interval. We present this problem, the princess and monster game on an interval as a challenge to the zero-sum game community. We conjecture that its value is 1.4.

## REFERENCES

- [1] S. ALPERN, *The search game with mobile hider on the circle*, in Proceedings of the Conference on Differential Games and Control Theory (Kingston, RI, July 1973), Differential Games and Control Theory, E. O. Roxin, P. T. Liu, and R. L. Sternberg, eds., Marcel Dekker, New York, 1974, pp. 181–200.
- [2] S. ALPERN, *Hide-and-seek games on trees to which Eulerian networks are attached*, Networks, 2008, to appear.
- [3] S. ALPERN AND M. ASIC, *The search value of a network*, Networks, 15 (1985), pp. 229–238.
- [4] S. ALPERN AND S. GAL, *A mixed strategy minimax theorem without compactness*, SIAM J. Control Optim., 26 (1988), pp. 1357–1361.
- [5] S. ALPERN AND S. GAL, *Rendezvous on the line with distinguishable players*, SIAM J. Control Optim., 33 (1998), pp. 1270–1276.
- [6] S. ALPERN AND S. GAL, *The Theory of Search Games and Rendezvous*, Kluwer International Series in Operations Research and Management Sciences, Kluwer, Boston, 55 (2003), p. 319.
- [7] T. BASAR AND G. J. OLSDER, *Dynamic Noncooperative Game Theory*, 2nd ed., Academic Press, New York, 1995.
- [8] E. CHESTER AND R. J. TUTUNCU, *Rendezvous search on the labelled line*, Oper. Res., 52 (2004), pp. 330–334.

- [9] A. DAGAN AND S. GAL, *Searching a network from an arbitrary starting point*, Networks, 2008, to appear.
- [10] J. EDMONDS AND E. L. JOHNSON, *Matching Euler tours and the Chinese postman problem*, Math. Program., 5 (1973), pp. 88–124.
- [11] S. GAL, *Search Games*, Academic Press, New York, 1980.
- [12] S. GAL, *On the optimality of a simple strategy for searching graphs*, Int. J. Game Theory, 29 (2000), pp. 533–542.
- [13] A. Y. GARNAEV, *Search Games and Other applications of game theory*, Springer Verlag, Berlin, 2000.
- [14] J. V. HOWARD, *Rendezvous search on the interval and the circle*, Oper. Res., 47 (1999), pp. 550–557.
- [15] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.
- [16] K. KIKUTA, *A search game with travelling cost on a tree*, J. Oper. Res. Soc. Japan, 38 (1995), pp. 70–88.
- [17] T. D. PARSONS, *The search number of a connected graph*, in Proceedings of the 9th Annual Southwestern Conference on Combinatorics, Graph Theory, and Computing, Boca Raton, FL, 1978.
- [18] J. H. REIJNIERSE AND J. A. M. POTTERS, *Search games with immobile hider*, Int. J. Game Theory, 21 (1993), pp. 385–394.
- [19] W. H. RUCKLE, *Geometric Games and Their Applications*, Pitman, Boston, 1983.
- [20] M. I. ZELIKEN, *On a differential game with incomplete information*, Soviet Math. Dokl. 13 (1972), pp. 228–231.

## A DIRECT SOLUTION METHOD FOR STOCHASTIC IMPULSE CONTROL PROBLEMS OF ONE-DIMENSIONAL DIFFUSIONS\*

MASAHIKO EGAMI<sup>†</sup>

**Abstract.** We consider stochastic impulse control problems where the process is driven by one-dimensional diffusions. Impulse control problems are widely applied to financial engineering and decision-making problems including the dividend payout problem, portfolio optimization with transaction costs, and inventory control. We shall show a new mathematical characterization of the value function in the continuation region as a linear function in a certain transformed space. The merits of our approach are as follows: (1) One does not have to guess optimal strategies or verify the optimality via a verification lemma, (2) the method of finding the solution (based on the new characterization of the value function) is simple and direct, and thereby (3) one can handle a broader class of reward and cost functions than the conventional methods that use quasi-variational inequalities.

**Key words.** stochastic impulse control, diffusions, optimal stopping, concavity

**AMS subject classifications.** Primary: 49N25; Secondary: 60G40

**DOI.** 10.1137/060669905

**1. Introduction.** This paper proposes a general solution method of stochastic impulse control problems for one-dimensional diffusion processes. Stochastic impulse control problems have attracted a growing interest of many researchers for the past two decades. Under a typical setting, the controller faces some underlying process and reward/cost structure. There exist continuous and instantaneous components of reward/cost functions. By exercising impulse controls, the controller moves the underlying process from one point to another. At the same time, the controller receives rewards associated with the instantaneous shifts of the process. Then the controller's objective is to maximize the total discounted expected net income.

The mathematical framework to these types of problems is in Bensoussan and Lions [4]. Impulse control has been studied widely in inventory control (Harrison, Sellke, and Taylor [15]), the exchange rate problem (Jeanblanc-Piqué [17], Mundaca and Øksendal [23], Cadenillas and Zapatero [7]), dividend payout problems (for example, Jeanblanc-Piqué and Shiryaev [18]), and portfolio optimization with transaction costs (Korn [19], Morton and Pliska [22]). It is worth mentioning that Korn [20] surveys the applications in mathematical finance. Also see Chancelier, Øksendal, and Sulem [10] for a combination of optimal stopping and impulse control problems. In many economic and financial applications where the controlled process is described as an Itô diffusion, the solution to the problem demands a thorough study of a related Hamilton–Jacobi–Bellman equation and quasi-variational inequalities. The method based on quasi-variational inequalities is the following: One guesses the form of (a) the continuation region and the intervention region, (b) the associated optimal policy, and (c) the value function. Then one specifies the value function by using appropriate boundary conditions and verifies optimality of the candidate policy. Both steps are

---

\*Received by the editors September 14, 2006; accepted for publication (in revised form) September 26, 2007; published electronically March 21, 2008.

<http://www.siam.org/journals/sicon/47-3/66990.html>

<sup>†</sup>Department of Mathematics, University of Michigan, Ann Arbor, MI 48189-1043 (egami@umich.edu), and Graduate School of Economics, Kyoto University, Kyoto, 606-8501, Japan (egami@econ.kyoto-u.ac.jp).

often very difficult, and success depends heavily on the form of the controlled process, reward, and cost functions.

Alternatively, an impulse control problem can be viewed as a sequence of optimal stopping problems. The connection between impulse control and optimal stopping has been investigated by Davis [12] and Øksendal and Sulem [25] among others. In this setting, the value functions of a sequence of optimal stopping problems converge to the value function of the impulse control problem under suitable conditions. In this regard, we utilize the results of Dynkin [14] (see, e.g., Theorem 16.4) about the functional concavity characterization of  $\alpha$ -excessive mappings and Dayanik and Karatzas [13] that give a general characterization of optimal stopping times of one-dimensional diffusions. We use these results to identify a new and useful characterization of the solution of the original impulse control problem. At the end, we get rid of the sequence of optimal stopping problems altogether and directly find the value function: The new characterization allows us to propose a new direct solution method for impulse control problems. Other works similar to our approach include Alvarez [1], Alvarez and Lempa [3], and Alvarez and Virtanen [2].

Similar recursive formulations for one-dimensional diffusions appear in multiple stopping problems in the context of swing options; see, for example, Carmona and Dayanik [8]. See also Carmona and Touzi [9] and Dahlgren and Korn [11] for the valuation of swing options. It is worth pointing out the differences of this article from Carmona and Dayanik [8], where, given a reward function, the controller can exercise their options  $n$  times under the condition that the controller has to wait at least certain units of time between two exercising times. By using a recursive formulation, we are able to obtain the result that the value function has to be linear in the continuation region of the transformed space. This linear characterization enables us to present the three-step optimization procedure (described in section 3.3) that does not require us to use the recursive stopping scheme, while Carmona and Dayanik [8] do not have this kind of geometric specification due to the nature of their problem; one needs to solve multiple stopping problems.

In the next section, we briefly go over the solution method for optimal stopping problems of one-dimensional diffusions. We describe the impulse control problem and its solution in section 3. Examples are presented in section 4. Finally, extensions and concluding remarks are in section 5.

**2. Summary of the key results of optimal stopping.** Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a complete probability space with a standard Brownian motion  $W = \{W_t; t \geq 0\}$ , and consider the diffusion process  $X^0$  with state space  $\mathcal{I} \subseteq \mathbb{R}$  and dynamics

$$(2.1) \quad dX_t^0 = \mu(X_t^0)dt + \sigma(X_t^0)dW_t$$

for some Borel functions  $\mu : \mathcal{I} \rightarrow \mathbb{R}$  and  $\sigma : \mathcal{I} \rightarrow (0, \infty)$ . We emphasize here that  $X^0$  is an uncontrolled process. We assume that  $\mathcal{I}$  is an interval with end points  $-\infty \leq c < d \leq +\infty$  and that  $X^0$  is regular in  $(c, d)$ ; in other words,  $X^0$  reaches  $y$  with positive probability starting at  $x$  for every  $x$  and  $y$  in  $(c, d)$ . We shall denote by  $\mathbb{F} = \{\mathcal{F}_t\}$  the natural filtration generated by  $X^0$  and assume that the usual condition is satisfied.

Let  $\alpha \geq 0$  be a real constant and  $h(\cdot)$  a Borel function such that  $\mathbb{E}^x[e^{-\alpha\tau}h(X_\tau^0)]$  is well-defined for every  $\mathbb{F}$ -stopping time  $\tau$  and  $x \in \mathcal{I}$ . We first consider

$$(2.2) \quad V(x) \triangleq \sup_{\tau \in \mathcal{S}} \mathbb{E}^x[e^{-\alpha\tau}h(X_\tau^0)], \quad x \in \mathcal{I},$$

the value function of the optimal stopping problem with the reward function  $h(\cdot) : \mathcal{I} \rightarrow \mathbb{R}$  where the supremum is taken over the class  $\mathcal{S}$  of all  $\mathbb{F}$ -stopping times.

Let  $z \in \mathcal{I}$  be a fixed point of the state space and  $\tau_z$  be the first hitting time of  $z \in \mathcal{I}$  by  $X^0$ . Let us denote the infinitesimal generator of  $X^0$  by  $\mathcal{A}$  and consider the ODE  $(\mathcal{A} - \alpha)v(x) = 0$ . This equation has two fundamental solutions:  $\psi(\cdot)$  and  $\varphi(\cdot)$ . We set  $\psi(\cdot)$  to be the increasing and  $\varphi(\cdot)$  to be the decreasing solution. They are linearly independent positive solutions and uniquely determined up to multiplication. It is well known that

$$\mathbb{E}^x[e^{-\alpha\tau_z}] = \begin{cases} \frac{\psi(x)}{\psi(z)}, & x \leq z, \\ \frac{\varphi(x)}{\varphi(z)}, & x \geq z. \end{cases}$$

For the complete characterization of  $\psi(\cdot)$  and  $\varphi(\cdot)$  corresponding to various types of boundary behavior, refer to Itô and McKean [16]. Let us now define

$$(2.3) \quad F(x) \triangleq \frac{\psi(x)}{\varphi(x)}, \quad x \in \mathcal{I}.$$

Then  $F(\cdot)$  is continuous and strictly increasing. Next, following Dynkin [14, pp. 238], we define concavity of a function with respect  $F$  as follows: A real-valued function  $u$  is called  $F$ -concave on  $\mathcal{I}$  if, for every  $x \in [l, r] \subseteq \mathcal{I}$ ,

$$u(x) \geq u(l) \frac{F(r) - F(x)}{F(r) - F(l)} + u(r) \frac{F(x) - F(l)}{F(r) - F(l)}.$$

We will use the notion of  $F$ -concavity to provide a new characterization of the value function of stochastic impulse control problems. Before doing that, the first step is to present the following results of optimal stopping problems, the proofs of which we refer to Dayanik and Karatzas [13].

**PROPOSITION 2.1.** *The value function  $V(\cdot)$  of (2.2) is the smallest nonnegative majorant of  $h(\cdot)$  such that  $V(\cdot)/\varphi(\cdot)$  is  $F$ -concave on  $\mathcal{I}$ .*

**PROPOSITION 2.2.** *Let  $W(\cdot)$  be the smallest nonnegative concave majorant of  $H \triangleq (h/\varphi) \circ F^{-1}$  on  $[F(c), F(d)]$ , where  $F^{-1}(\cdot)$  is the inverse of the strictly increasing function  $F(\cdot)$  in (2.3). Then  $V(x) = \varphi(x)W(F(x))$  for every  $x \in \mathcal{I}$ .*

**PROPOSITION 2.3.** *Define*

$$(2.4) \quad S \triangleq \{x \in [c, d] : V(x) = h(x)\} \quad \text{and} \quad \tau^* \triangleq \inf\{t \geq 0 : X_t^0 \in S\}.$$

*If  $h(\cdot)$  is continuous on  $\mathcal{I}$ , then  $\tau^*$  is an optimal stopping rule.*

**3. Impulse control problem and its solution.** Suppose that, at any time  $t \in \mathbb{R}_+$  and any state  $x \in \mathcal{I}$ , we can intervene and give the system an impulse  $\xi \in \mathbb{R}$ . Once the system gets intervened, the point moves from  $x$  to  $y \in \mathbb{R}$ . An impulse control for the system is a double sequence:

$$\nu = (T_1, T_2, \dots, T_i, \dots; \xi_1, \xi_2, \dots, \xi_i, \dots),$$

where  $0 \leq T_1 < T_2 < \dots$  are an increasing sequence of  $\mathbb{F}$ -stopping times and  $\xi_1, \xi_2, \dots$  are  $\mathcal{F}_{T_i}$ -measurable random variables representing impulses exercised at the corresponding intervention times  $T_i$ , with  $\xi_i \in Z$  for all  $i$ , where  $Z \in \mathbb{R}$  is a given

set of admissible impulse amounts. The controlled process is, in general, described as follows:

$$(3.1) \quad dX_t = \mu(X_t)dt + \sigma(X_t)dW_t, \quad T_{i-1} \leq t < T_i,$$

$$(3.2) \quad X_{T_i} = \Gamma(X_{T_i-}, \xi_i)$$

with some mapping  $\Gamma : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ .

Let 0 be the absorbing state, without loss of generality, and  $\tau_0 \triangleq \inf\{t : X_t = 0\}$  the ruin time. With the absorbing state at 0, it is natural to consider a set of problems where  $Z \in \mathbb{R}_+$  (i.e.,  $\xi_i = x_i - y_i > 0$  for all  $i$ ) and  $X_{T_i} = X_{T_i-} - \xi_i$ . (We shall comment on cases where interventions are allowed in both positive and negative directions in section 5.)

With each pair  $(T_i, \xi_i)$ , we associate the intervention reward

$$(3.3) \quad K(X_{T_i-}, X_{T_i}),$$

where  $K(x, y) : \mathcal{I} \times \mathbb{R} \rightarrow \mathbb{R}$  is a given function continuous in the first and second arguments. Our result below does not depend on the specification of  $K(\cdot)$ . We assume that, for any point  $x \in \mathcal{I}$ ,

$$(3.4) \quad K(x, x) < 0$$

due to the existence of fixed costs. We consider the following performance measure with  $\nu \in \mathcal{V}$ , a collection of admissible strategies:

$$(3.5) \quad J^\nu(x) = \mathbb{E}^x \left[ \int_0^{\tau_0} e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau_0} P + \sum_{T_i < \tau_0} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right],$$

where  $P \in \mathbb{R}_-$  is a constant penalty<sup>1</sup> at the ruin time and  $f : \mathbb{R} \rightarrow \mathbb{R}$  is a continuous function, satisfying:

$$(3.6) \quad \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} |f(X_s)| ds \right] < \infty.$$

We also assume the standard polynomial growth condition on  $f(\cdot)$  and  $K(x, y) = K(z)$  by setting  $z := x - y$ : That is, there exist constants  $C_i$  and  $m_i$  ( $i = 1, 2$ ) such that, for all  $x, z \in \mathbb{R}$ ,

$$(3.7) \quad |f(x)| \leq C_1(1 + |x|^{m_1}) \quad \text{and} \quad |K(z)| \leq C_2(1 + |z|^{m_2}).$$

Our goal is to find the optimal strategy  $\nu^*(T_i, \xi_i)_{i \geq 0}$  and the corresponding value function

$$(3.8) \quad v(x) \triangleq \sup_{\nu \in \mathcal{V}} J^\nu(x) = J^{\nu^*}(x).$$

Let us briefly go over our plan. In section 3.1, we introduce a recursive optimal stopping scheme that eventually solves the original impulse control problem as in Øksendal and Sulem [25]. In section 3.2, we consider a special case where the mapping

<sup>1</sup>Equation (3.5) can be rewritten as  $J^\nu(x) = P + \mathbb{E}^x[\int_0^{\tau_0} e^{-\alpha s} (f(X_s) - \alpha P) ds + \sum_{T_i < \tau_0} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i})]$ . Hence we could define  $\tilde{f}(x) \triangleq f(x) - \alpha P$  to get rid of  $P$ . We, however, maintain the original formulation to make the penalty term explicit.

$x \rightarrow \frac{K}{\varphi}(x) : \mathcal{I} \rightarrow \mathbb{R}$  is  $F$ -concave. We show, under this assumption, that the optimal intervention times  $T_i$  are characterized as exit times from an interval, say,  $(0, b^*)$ , for every  $i$ . Then we characterize the value function for impulse control problems and present a solution method based on the characterization of the intervention times and value function. In section 3.4, we consider the general case where the  $F$ -concavity assumption above does not hold.

**3.1. A sequence of optimal stopping problems.** In this subsection, we consider a recursive optimal stopping with a view to characterizing intervention times for the impulse control problems. Here we assume that no absorbing boundary exists. As we will see in the next subsection, the existence of an absorbing state is easily incorporated. Hence by using the same  $v(x)$ , we consider the problem

$$(3.9) \quad v(x) = \sup_{\nu \in \mathcal{V}} \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} f(X_s) ds + \sum_i e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right].$$

Let us also define the set  $S_n$  and the objective function  $v_n$  as follows:

$$S_n \triangleq \{\nu \in S; \nu = (T_1, T_2, \dots, T_{n+1}; \xi_1, \xi_2, \dots, \xi_n); T_{n+1} = +\infty\}$$

and

$$(3.10) \quad v_n(x) \triangleq \sup_{\nu \in S_n} \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} f(X_s) ds + \sum_{T_i} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right].$$

In other words, we are allowed to make at most  $n$  interventions. For this recursive approach, see, for example, Davis [12] and Øksendal and Sulem [25]. We use the following notation for convenience:

$$(3.11) \quad g(x) \triangleq \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} f(X_s^0) ds \right].$$

Let  $\mathcal{H}$  denote the space of all Borel functions. Define the two operators  $\mathcal{M} : \mathcal{H} \rightarrow \mathcal{H}$  and  $\mathcal{L} : \mathcal{H} \rightarrow \mathcal{H}$  as follows:

$$(3.12) \quad \mathcal{M}u(x) \triangleq \sup_{y \in \mathbb{R}} [K(x, y) - (g(x) - g(y)) + u(y)]$$

and

$$(3.13) \quad \mathcal{L}u(x) \triangleq \sup_{\tau \in S} \mathbb{E}^x [e^{-\alpha \tau} \mathcal{M}u(X_{\tau-})]$$

for  $u \in \mathcal{H}$ . From the definition of the two operators,  $a_1(x) \leq a_2(x)$  for  $x \in \mathbb{R}$ ,  $a_1(\cdot), a_2(\cdot) \in \mathcal{H}$  implies  $\mathcal{M}a_1(x) \leq \mathcal{M}a_2(x)$  and  $\mathcal{L}a_1(x) \leq \mathcal{L}a_2(x)$  for all  $x \in \mathbb{R}$ .

On the other hand, consider the following recursive formula:

$$(3.14) \quad \begin{cases} w_0(x) = g(x), \\ w_n(x) = \sup_{\tau \in S, \xi \in \mathbb{R}_+} \mathbb{E}^x \left[ \int_0^\tau e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau} (K(X_{\tau-}, X_\tau) + w_{n-1}(X_\tau)) \right] \end{cases}$$

for  $n = 1, 2, 3, \dots$ . It should be noted that, for each  $n$ , this is an optimal stopping problem over  $\tau$ . The second line of (3.14) is equivalent to

$$(3.15) \quad w_{n+1}(x) - g(x) = \sup_{\tau \in S, \xi \in \mathbb{R}_+} \mathbb{E}^x [e^{-\alpha \tau} (K(X_{\tau-}, X_\tau) - g(X_{\tau-}) + w_n(X_\tau))]$$



by applying the strong Markov property (with (3.6)) at time  $\tau$  to the integral term. Note that, on  $\{\omega : 0 \leq t \leq \tau(\omega)-\}$ , we have  $X(\omega) = X^0(\omega)$  almost surely. In fact, this derivation is explained in detail in subsection 3.2. By defining

$$\phi \triangleq w - g,$$

and adding and subtracting  $g(X_\tau)$  on the right-hand side of (3.15), it becomes

$$\phi_{n+1}(x) = \sup_{\tau \in \mathcal{S}} \mathbb{E}^x[e^{-\alpha\tau} \mathcal{M}\phi_n(X_{\tau-})].$$

Then it can be further simplified, by using the operator  $\mathcal{L}$  defined in (3.13):

$$(3.16) \quad \phi_{n+1}(x) = \mathcal{L}\phi_n(x).$$

Let us start this recursive scheme with  $w_0(x) = g(x)$  (i.e., no interventions are allowed, equivalently  $\phi_0(x) = 0$ ) and define recursively  $\phi_n(x) \triangleq w_n(x) - g(x) = \mathcal{L}(w_{n-1}(x) - g(x)) = \mathcal{L}\phi_{n-1}$ . Clearly,

$$\begin{aligned} \phi_1(x) &= \mathcal{L}\phi_0(x) = \sup_{\tau \in \mathcal{S}} \mathbb{E}^x[e^{-\alpha\tau} (\mathcal{M}(w_0(X_{\tau-}) - g(X_{\tau-})))] \\ &= \sup_{\tau \in \mathcal{S}, \xi \in \mathbb{R}_+} \mathbb{E}^x[e^{-\alpha\tau} \{K(X_{\tau-}, X_\tau) - g(X_{\tau-}) + g(X_\tau)\}]. \end{aligned}$$

On the other hand, by looking at the first (and the sole) intervention time,

$$\begin{aligned} v_1(x) - g(x) &= \sup_{\nu \in \mathcal{S}_1} \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} f(X_s) ds + e^{-\alpha\tau} K(X_{\tau-}, X_\tau) \right] \\ &\quad - \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} f(X_s^0) ds \right] \\ &= \sup_{\tau \in \mathcal{S}, \xi \in \mathbb{R}_+} \mathbb{E}^x \left[ \int_0^\tau e^{-\alpha s} f(X_s) ds + e^{-\alpha\tau} \left( \mathbb{E}^{X_\tau} \left[ \int_0^\infty e^{-\alpha s} f(X_s) ds \right] \right. \right. \\ &\quad \left. \left. + K(X_{\tau-}, X_\tau) \right) - \int_0^\tau e^{-\alpha s} f(X_s^0) ds - e^{-\alpha\tau} g(X_{\tau-}) \right] \\ &= \sup_{\tau \in \mathcal{S}, \xi \in \mathbb{R}_+} \mathbb{E}^x[e^{-\alpha\tau} \{K(X_{\tau-}, X_\tau) - g(X_{\tau-}) + g(X_\tau)\}]. \end{aligned}$$

The last equation is due to the fact that only one intervention is allowed. Hence we have  $w_1(x) = v_1(x)$ . By the definition of the recursive scheme,  $w_n$  is an increasing sequence (i.e.,  $w_1(x) \leq w_2(x) \leq \dots$  for all  $x \in \mathbb{R}$ ). In fact, we shall prove that  $w_n = v_n$  for all  $n$  in Lemma 3.2. Before that, we need the following lemma to relate this recursive scheme with the method described in section 2.

LEMMA 3.1. *The mapping  $x \rightarrow \frac{\mathcal{L}\phi(x)}{\varphi(x)} : \mathcal{I} \rightarrow \mathbb{R}$  is  $F$ -concave.*

*Proof.* We shall fix some  $x \in (l, r) \subseteq \mathcal{I}$ . Since  $\mathcal{M}\phi(\cdot)$  is bounded there, for a given  $\varepsilon > 0$ , there are admissible  $\varepsilon$ -optimal intervention pairs  $(\sigma_\varepsilon^l, \xi_\varepsilon^l)$  and  $(\sigma_\varepsilon^r, \xi_\varepsilon^r)$  such that

$$\mathbb{E}^l[e^{-\alpha\sigma_\varepsilon^l} \mathcal{M}\phi(X_{\sigma_\varepsilon^l})] > \mathcal{L}\phi(l) - \varepsilon \quad \text{and} \quad \mathbb{E}^r[e^{-\alpha\sigma_\varepsilon^r} \mathcal{M}\phi(X_{\sigma_\varepsilon^r})] > \mathcal{L}\phi(r) - \varepsilon.$$

Define another stopping time  $\sigma_\varepsilon^{lr} \in \mathcal{S}$ , with

$$\sigma_\varepsilon^{lr} \triangleq \begin{cases} \tau^l + \sigma_\varepsilon^l \circ \theta_{\tau^l} & \text{if } \tau^l < \tau^r, \\ \tau^r + \sigma_\varepsilon^r \circ \theta_{\tau^r} & \text{if } \tau^l > \tau^r. \end{cases}$$

Putting it all together, with the strong Markov property of  $X$ , which is  $X^0$  a.s. until the stopping time  $\tau$ ,

$$\begin{aligned}\mathcal{L}\phi(x) &\geq \mathbb{E}^x[e^{-\alpha\sigma_\varepsilon^{lr}} \mathcal{M}\phi(X_{\sigma_\varepsilon^{lr}})] \\ &> (\mathcal{L}\phi(l) - \varepsilon)\mathbb{E}^x[e^{-\alpha\tau^l} 1_{\{\tau^l < \tau^r\}}] + (\mathcal{L}\phi(r) - \varepsilon)\mathbb{E}^x[e^{-\alpha\tau^r} 1_{\{\tau^l > \tau^r\}}] \\ &\geq \frac{\mathcal{L}\phi(l)}{\varphi(l)}\varphi(x)\frac{F(r) - F(x)}{F(r) - F(l)} + \frac{\mathcal{L}\phi(r)}{\varphi(r)}\varphi(x)\frac{F(x) - F(l)}{F(r) - F(l)} - \varepsilon.\end{aligned}$$

Since  $\varepsilon$  is arbitrary, we have an  $F$ -concavity.  $\square$

This lemma guarantees that we can use Propositions 2.1–2.3 to identify the value function and an optimal stopping rule for each of the recursive optimal stopping problems (3.14).

**3.2. Characterization of the intervention times and the value function:  $F$ -concave reward case.** Based on the results in the previous subsection, we first consider a special case where the mapping  $x \rightarrow \frac{\bar{K}}{\varphi}(x) : \mathcal{I} \rightarrow \mathbb{R}$  is  $F$ -concave. Let us define, for notational convenience,

$$(3.17) \quad \bar{K}(x, y) \triangleq K(x, y) - (g(x) - g(y)).$$

Further, we prove the following properties of the recursive optimization scheme.

LEMMA 3.2. *If we define  $w_n$  by (3.14) (with  $w_0 = g$ ) and  $v_n$  by (3.10), then*

$$w_n(x) = v_n(x) \quad \text{for each } n \quad \text{and} \quad v(x) = \lim_{n \rightarrow \infty} w_n(x).$$

Moreover,  $w(x) \triangleq \lim_{n \rightarrow +\infty} w_n(x)$  is the smallest solution majorizing  $g$  of the functional equation  $w - g = \mathcal{L}(w - g)$ .

*Proof.* The proof is given in the appendix.  $\square$

Hence if we solve the optimal stopping problem

$$(3.18) \quad \phi_{n+1}(x) = \sup_{\tau \in \mathcal{S}} \mathbb{E}^x[e^{-\alpha\tau} \mathcal{M}\phi_n(X_{\tau-})]$$

recursively for each  $n$ , then we obtain  $\phi(x) \triangleq \lim_{n \rightarrow \infty} \phi_n(x) = \lim_{n \rightarrow \infty} v_n(x) - g(x) = v(x) - g(x)$ . By summarizing the above argument, we have the following proposition.

PROPOSITION 3.3. *The value function  $v(x)$  for (3.9) is given by the smallest solution majorizing  $g$  of the functional equation  $v - g = \mathcal{L}(v - g)$ , and  $\frac{v-g}{\varphi} (= \frac{\phi}{\varphi})$  is always  $F$ -concave.*

*Proof.* The first statement comes from Lemma 3.2. By the recursive method that we described above, we are solving a series of optimal stopping problems for each  $\phi_n$ . Hence Lemma 3.1 and Proposition 2.1 give the second statement.  $\square$

Note that the functional equation  $v - g = \mathcal{L}(v - g)$  is in essence the same as (2.8) in Alvarez [1], where the optimal boundary is found by ordinary optimization techniques in a case where the size of the control is exogenously given. Similar ideas are also in Alvarez and Virtanen [2] and Alvarez and Lempa [3] that extend Alvarez [1] by finding optimal size of control as well under some practically reasonable assumptions on the reward and cost function. These papers identify the function (along with optimal control) that satisfies this relationship, by using the idea of  $\alpha$ -excessive mappings

and proving optimality with the verification of a weaker version of quasi-variational inequalities under the assumptions on the reward function. Now we consider different paths to reach the solution and develop a new method.

The argument in the previous subsection is modified to incorporate the existence of the ruin state. Instead of (3.10) and (3.14), we define, respectively,

$$\begin{aligned} v_n(x) &\triangleq \sup_{\nu \in S_n} \mathbb{E}^x \left[ \int_0^{\tau_0} e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau_0} P + \sum_{T_i < \tau_0} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right], \\ w_{n+1}(x) &\triangleq \sup_{\tau \in \mathcal{S}, \xi} \mathbb{E}^x \left[ \int_0^{\tau_0 \wedge \tau} e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau_0} P 1_{\{\tau_0 < \tau\}} \right. \\ &\quad \left. + e^{-\alpha \tau} \{K(X_{\tau-}, X_{\tau}) + w_n(X_{\tau})\} 1_{\{\tau < \tau_0\}} \right], \end{aligned}$$

with

$$w_0(x) = \mathbb{E}^x \left[ \int_0^{\infty} e^{-\alpha s} f(X_s^0) 1_{\{s < \tau_0\}} ds + e^{-\alpha \tau_0} P \right] \triangleq g_0(x).$$

Then by defining the operator  $\mathcal{L} : \mathcal{H} \rightarrow \mathcal{H}$  instead of (3.13),

$$(3.19) \quad \mathcal{L}u(x) \triangleq \sup_{\tau \in \mathcal{S}} \mathbb{E}^x [e^{-\alpha \tau} \mathcal{M}u(X_{\tau-}) 1_{\{\tau < \tau_0\}} + e^{-\alpha \tau_0} (P - g(0)) 1_{\{\tau_0 < \tau\}}],$$

we have the same recursion formula as in (3.16). We can obtain the same results as in Lemmas 3.1 and 3.2. Proposition 3.3 also holds with one change that the value function is given by the smallest solution majorizing  $g_0$  of the functional equation  $v - g = \mathcal{L}(v - g)$ , where  $\mathcal{L} : \mathcal{H} \rightarrow \mathcal{H}$  is given by (3.19). Now we consider the characterization of the intervention times.

**PROPOSITION 3.4.** *If the mapping  $x \rightarrow \frac{K}{\varphi}(x) : \mathbb{R}_+ \rightarrow \mathbb{R}$  is  $F$ -concave and 0 is an absorbing state, then the optimal intervention times  $T_i^*$  are given, for some  $b^* \in \mathbb{R}_+$ , by*

$$T_i^* = \inf\{t > T_{i-1}^*; X_t \notin (0, b^*), \quad i = 1, 2, \dots\}.$$

*Proof.* Our proof is constructive, describing the procedure of recursive optimization steps. For any  $n \geq 1$ , in view of Lemma 3.1,  $\phi_n(x)/\varphi(x)$  is the smallest  $F$ -concave majorant of  $\mathcal{L}\phi_{n-1}(x)/\varphi(x)$ . We claim that this majorant (that passes  $(F(0), \frac{P-g(0)}{\varphi(0)})$  in the transformed space) always exists. Indeed, since we consider the case of  $\xi_i > 0$ , i.e.,  $x > y$  for  $K(x, y)$  and

$$\begin{aligned} \mathcal{M}\phi_0(x) &= \sup_{y \in \mathbb{R}_+} [K(x, y) - (g(x) - g(y)) + \phi_0(y)] \\ &= \sup_{y \in \mathbb{R}_+} [K(x, y) - (g(x) - g(y)) + (g_0(y) - g(y))], \end{aligned}$$

we should check whether the concave majorant exists, namely, that

$$(3.20) \quad \lim_{x \downarrow 0} (K(x, y) - g(x) + g_0(y)) < P - g(0)$$

holds when  $y \downarrow 0$ . Note that  $\lim_{y \downarrow 0} g_0(y) = P$  and  $g(x) \rightarrow g(0)$  as  $x \rightarrow 0$  due to the continuity of  $f$ . Hence (3.20) holds in the neighborhood of  $y = 0$  because of (3.4). In the subsequent iterations, we consider

$$\mathcal{M}\phi_1(x) = \sup_{y \in \mathbb{R}_+} [K(x, y) - (g(x) - g(y)) + \phi_1(y)].$$

We should check if the expression inside the supremum operator becomes less than  $P - g(0)$  as  $x \downarrow 0$  and  $y \downarrow 0$ . Since  $\lim_{y \downarrow 0} \phi_1(y) = \phi_1(0) = P - g(0)$  by the concavity (hence continuity) of  $\phi_1$  and since  $\lim_{x \downarrow 0} g(x) = \lim_{y \downarrow 0} g(y)$ , we have, in the neighborhood of  $y = 0$ , that

$$\lim_{x \downarrow 0} K(x, 0) + P - g(0) < P - g(0)$$

holds. Hence the concave majorant always exists also in the subsequent iterations.

Now the  $F$ -concavity of  $\phi_n$  is obviously maintained for all  $n$ . The limit function of the increasing sequence of functions  $\phi(x) = \lim_{n \rightarrow \infty} \phi_n(x)$  exists and is also  $F$ -concave. Accordingly,  $\bar{K}(x, y)/\varphi(x) + \phi(y)/\varphi(x)$  is  $F$ -concave for all  $y$ . Hence  $\phi(x)/\varphi(x)$  and  $(\bar{K}(x, y) + \phi(y))/\varphi(x)$  meet once and only once in the transformed space. Recall that the value function satisfies  $\phi = \mathcal{L}\phi$ . This implies that the continuous region is in the form of  $(0, b^*)$  for some  $b^* \in \mathbb{R}_+$ , which completes the proof.  $\square$

By using the above characterization of intervention times, we next want to characterize the value function and reduce the impulse control problem (3.8) to some optimal stopping problem. Moreover, we shall present a method that does not have to go through the iteration scheme. Let us first simplify  $J^\nu$ :

$$(3.21) \quad J^\nu(x) = \mathbb{E}^x \left[ \int_0^{\tau_0} e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau_0} P + \sum_{T_i < \tau_0} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right].$$

This is just a reproduction of (3.5). Let us split the right-hand side of (3.21) into pieces and use the strong Markov property of the uncontrolled process  $X^0$  at the first intervention time  $T_1$  (together with the shift operator  $\theta(\cdot)$ ) with each of them. Note that, on  $\{\omega : 0 \leq t \leq T_1(\omega)-\}$ , we have  $X(\omega) = X^0(\omega)$  almost surely. The first term becomes

$$\begin{aligned} \mathbb{E}^x \left[ \int_0^{\tau_0} e^{-\alpha s} f(X_s) ds \right] &= \mathbb{E}^x \left[ 1_{\{T_1 < \tau_0\}} \left\{ \int_0^{T_1} e^{-\alpha s} f(X_s) ds \right. \right. \\ &\quad \left. \left. + e^{-\alpha T_1} \mathbb{E}^{X_{T_1}} \int_0^\infty e^{-\alpha s} f(X_s) 1_{\{s < \tau_0\}} ds \right\} \right] \\ &\quad + \mathbb{E}^x \left[ 1_{\{T_1 > \tau_0\}} \int_0^{\tau_0} f(X_s) ds \right] \end{aligned}$$

since  $\int_{T_1-}^{T_1} e^{-\alpha s} f(X_s) ds = 0$ . The second and third terms become

$$\begin{aligned} \mathbb{E}^x [e^{-\alpha \tau_0} P] &= \mathbb{E}^x \left[ 1_{\{T_1 < \tau_0\}} e^{-\alpha T_1} \mathbb{E}^x [e^{-\alpha(\tau_0 - T_1)} P | \mathcal{F}_{T_1}] \right] + \mathbb{E}^x [1_{\{T_1 > \tau_0\}} e^{-\alpha \tau_0} P] \\ &= \mathbb{E}^x \left[ 1_{\{T_1 < \tau_0\}} e^{-\alpha T_1} \mathbb{E}^x [e^{-\alpha(\tau_0 \circ \theta(T_1))} P | \mathcal{F}_{T_1}] \right] + \mathbb{E}^x [1_{\{T_1 > \tau_0\}} e^{-\alpha \tau_0} P] \\ &= \mathbb{E}^x [1_{\{T_1 < \tau_0\}} e^{-\alpha T_1} \mathbb{E}^{X_{T_1}} (e^{-\alpha \tau_0} P)] + \mathbb{E}^x [1_{\{T_1 > \tau_0\}} e^{-\alpha \tau_0} P] \end{aligned}$$

and

$$\begin{aligned}
& \mathbb{E}^x \left[ \sum_{T_i < \tau_0} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right] \\
&= \mathbb{E}^x \left[ 1_{\{T_1 < \tau_0\}} \left\{ e^{-\alpha T_1} K(X_{T_1-}, X_{T_1}) + e^{-\alpha T_1} \sum_{i=2} e^{-\alpha(T_i - T_1)} K(X_{T_i-}, X_{T_i}) 1_{\{T_i < \tau_0\}} \right\} \right] \\
&= \mathbb{E}^x \left[ 1_{\{T_1 < \tau_0\}} \left\{ e^{-\alpha T_1} K(X_{T_1-}, X_{T_1}) \right. \right. \\
&\quad \left. \left. + e^{-\alpha T_1} \mathbb{E}^x \left[ \sum_{T_i < \tau_0} e^{-\alpha(T_i \circ \theta(T_1))} K(X_{S_i-}, X_{S_i}) \mid \mathcal{F}_{T_1} \right] \right\} \right] \\
&= \mathbb{E}^x \left[ 1_{\{T_1 < \tau_0\}} e^{-\alpha T_1} \left\{ K(X_{T_1-}, X_{T_1}) + \mathbb{E}^{X_{T_1}} \sum_{i=1} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) 1_{\{T_i < \tau_0\}} \right\} \right],
\end{aligned}$$

where  $S_i \triangleq T_1 + T_i \circ \theta(T_1)$  and the index  $i$  runs from 1 for the sum in the second equality. By combining the three terms and rearranging, we have

$$\begin{aligned}
(3.22) \quad J^\nu(x) &= \mathbb{E}^x \left[ 1_{\{T_1 < \tau_0\}} \left\{ \int_0^{T_1} e^{-\alpha s} f(X_s) ds + e^{-\alpha T_1} K(X_{T_1-}, X_{T_1}) + e^{-\alpha T_1} J^\nu(X_{T_1}) \right\} \right] \\
&\quad + \mathbb{E}^x \left[ 1_{\{T_1 > \tau_0\}} \left\{ \int_0^{\tau_0} e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau_0} P \right\} \right].
\end{aligned}$$

For any  $\mathbb{F}$  stopping time  $\tau$ , the strong Markov property of  $X^0$ , with our assumption (3.6), gives us

$$\mathbb{E}^x \left[ \int_0^\tau e^{-\alpha s} f(X_s^0) ds \right] = g(x) - \mathbb{E}^x \left[ e^{-\alpha \tau} g(X_\tau^0) \right],$$

where  $g(\cdot)$  is defined as in (3.11). We apply this result to (3.22) by reading  $\tau = T_1$  and  $\tau = \tau_0$  to derive

$$\begin{aligned}
(3.23) \quad J^\nu(x) &= \mathbb{E}^x \left[ 1_{\{T_1 < \tau_0\}} e^{-\alpha T_1} \{ K(X_{T_1-}, X_{T_1}) - g(X_{T_1}^0) + J^\nu(X_{T_1}) \} \right] \\
&\quad + \mathbb{E}^x \left[ 1_{\{T_1 > \tau_0\}} e^{-\alpha \tau_0} \{ P - g(X_{\tau_0}) \} \right] + g(x).
\end{aligned}$$

By noting that  $g(X_{T_1}^0) = g(X_{T_1-})$ , adding and subtracting  $g(X_{T_1})$ , and further defining

$$u(x) \triangleq J^\nu(x) - g(x),$$

(3.23) finally becomes

$$\begin{aligned}
(3.24) \quad u(x) &= \mathbb{E}^x \left[ 1_{\{T_1 < \tau_0\}} e^{-\alpha T_1} \{ K(X_{T_1-}, X_{T_1}) + u(X_{T_1}) - g(X_{T_1-}) + g(X_{T_1}) \} \right] \\
&\quad + \mathbb{E}^x \left[ 1_{\{T_1 > \tau_0\}} e^{-\alpha \tau_0} \{ P - g(X_{\tau_0}) \} \right],
\end{aligned}$$

and we consider the maximization of this  $u(\cdot)$  function and add back  $g(x)$  since  $\sup u(x) = \sup \{ J^\nu(x) - g(x) \} = \sup J^\nu(x) - g(x)$ . Note that this simplification leading to (3.24) does not depend on the  $F$ -concavity assumption.

Since we have confirmed that optimal intervention times are exit times of the process from an interval, let us use a simpler notation:  $X_{T_i-} = b$  and  $X_{T_i} = a$  for all  $i = 0, 1, 2, \dots$ . We can denote  $T_i- = \tau_b \triangleq \inf\{t > T_{i-1}; X_t \geq b\}$ . By observing (3.24),

$$\begin{aligned} u(b) &= u(X_{T_1-}) = K(X_{T_1-}, X_{T_1}) + g(X_{T_1}) - g(X_{T_1-}) + u(X_{T_1}) \\ (3.25) \quad &= K(b, a) + g(a) - g(b) + u(a) = \bar{K}(b, a) + u(a), \\ u(0) &= u(X_{\tau_0}) = P - g(X_{\tau_0}) = P - g(0), \end{aligned}$$

and we have

$$(3.26) \quad u(x) = \begin{cases} u_0(x), & x \in [0, b), \\ \bar{K}(x, a) + u_0(a), & x \in [b, \infty), \end{cases}$$

where

$$u_0(x) \triangleq \mathbb{E}^x[1_{\{\tau_b < \tau_0\}} e^{-\alpha \tau_b} u(b)] + \mathbb{E}^x[1_{\{\tau_b > \tau_0\}} e^{-\alpha \tau_0} u(0)].$$

The second equation of (3.26) is obtained from (3.24) by noticing that, on  $x \in [b, \infty)$ ,  $\mathbb{P}^x(T_1 < \tau_0) = 1$ . Indeed, in this case, we immediately jump to  $a$ , so that  $X_{T_1-} = x$  and  $X_{T_1} = a$ . Since  $a \in (0, b)$ ,  $u(a) = u_0(a)$ . Now let us note that we have the following representations in (3.24):

$$\mathbb{E}^x[e^{-\alpha \tau_r} 1_{\{\tau_r < \tau_l\}}] = \frac{\psi(l)\varphi(x) - \psi(x)\varphi(l)}{\psi(l)\varphi(r) - \psi(r)\varphi(l)}, \quad x \in [l, r],$$

where  $\tau_l \triangleq \inf\{t > 0; X_t = l\}$  and  $\tau_r \triangleq \inf\{t > 0; X_t = r\}$ ;  $\varphi(\cdot)$  and  $\psi(\cdot)$  are defined in the previous section. Finally, with  $F(\cdot)$  being defined as in (2.3), we have a characterization of  $u(x)$  in the continuation region:

$$(3.27) \quad \frac{u(x)}{\varphi(x)} = \frac{u(b)(F(x) - F(0))}{\varphi(b)(F(b) - F(0))} + \frac{u(0)(F(b) - F(x))}{\varphi(0)(F(b) - F(0))}, \quad x \in [0, b].$$

Define the transformation

$$(3.28) \quad W \triangleq \frac{u}{\varphi} \circ F^{-1},$$

and (3.27) becomes, for any  $a > 0$  and  $b > 0$ ,

$$(3.29) \quad W(F(x)) = W(F(b)) \frac{F(x) - F(0)}{F(b) - F(0)} + W(F(0)) \frac{F(b) - F(x)}{F(b) - F(0)}, \quad x \in [0, b],$$

which represents a *linear function* that passes a fixed point  $(F(0), W(F(0)))$ .

To discuss how to find the optimal pair  $(a^*, b^*)$ , we write  $u(x)$  as  $u_{a,b}(x)$  to emphasize the dependence on  $a, b$ , then on  $x \in [0, b]$ :

$$\begin{aligned} (3.30) \quad \sup_{a \in \mathbb{R}_+, b \in \mathbb{R}_+} u_{a,b}(x) &= \sup_{a \in \mathbb{R}_+} \sup_{b \in \mathbb{R}_+} \{ \mathbb{E}^x[1_{\{\tau_b < \tau_0\}} e^{-\alpha \tau_b} (\bar{K}(b, a) \\ &\quad + u_{a,b}(a))] + \mathbb{E}^x[1_{\{\tau_b > \tau_0\}} e^{-\alpha \tau_0} u_{a,b}(0)] \}. \end{aligned}$$

This can be considered as a two-stage optimization problem. First, let  $a$  be fixed. For each  $a$ , the inner maximization of (3.30) becomes

$$(3.31) \quad V_a(x) \triangleq \sup_{\tau_b \in \mathcal{S}} \{ \mathbb{E}^x [1_{\{\tau_b < \tau_0\}} e^{-\alpha \tau_b} (\bar{K}(b, a) + V_a(a))] + \mathbb{E}^x [1_{\{\tau_b > \tau_0\}} e^{-\alpha \tau_0} (P - g(0))] \},$$

and, among  $a$ 's, choose an optimal  $a$  in the sense that  $\tilde{v}(x) \triangleq \sup_a V_a(x)$  for any  $x$ . It should be pointed out that  $V_a(x)$  may take negative values if  $P - g(0)$  does. Now we discuss a solution method of the first stage optimization (3.31). For this purpose, we need a technical lemma.

LEMMA 3.5. *If we define*

$$G(x, \gamma) \triangleq \sup_{\tau \in \mathcal{S}} \mathbb{E}^x [e^{-\alpha \tau} (h(X_\tau^0) + \gamma)], \quad x \in \mathbb{R}, \gamma \in \mathbb{R},$$

for some Borel function  $h : \mathbb{R} \rightarrow \mathbb{R}$  and with condition (3.6), then, for  $\gamma_1 > \gamma_2 \geq 0$ ,

$$(3.32) \quad G(x, \gamma_1) - G(x, \gamma_2) \leq \gamma_1 - \gamma_2$$

for any  $x$ .

*Proof.* The left-hand side of (3.32) is well-defined due to (3.6). It is clear that  $G(x, \gamma)$  is convex in  $\gamma$  for any  $x$ . Then  $D_\gamma^+ G(x, \gamma_0) \triangleq \lim_{\gamma \downarrow \gamma_0} \frac{G(x, \gamma_0) - G(x, \gamma)}{\gamma_0 - \gamma}$  exists at every  $\gamma_0 \in \mathbb{R}$ , and

$$(3.33) \quad \frac{G(x, \gamma_1) - G(x, \gamma_2)}{\gamma_1 - \gamma_2} \leq D_\gamma^+ G(x, \gamma_1).$$

Consider the bound of  $G(x, \gamma)$  for  $x$  fixed:

$$G(x, \gamma) \leq \sup_{\tau \in \mathcal{S}} \mathbb{E}^x [e^{-\alpha \tau} |h(X_\tau^0)|] + |\gamma| \sup_{\tau \in \mathcal{S}} \mathbb{E}^x [e^{-\alpha \tau}].$$

The first term on the right-hand side is constant in  $\gamma$ ; the second term is linear in  $\gamma$ , and  $\mathbb{E}^x [e^{-\alpha \tau}] \leq 1$  for any  $\tau \in \mathcal{S}$ . Due to the convexity of  $G(x, \gamma)$  in  $\gamma$ , for the above inequality to hold,  $D_\gamma^+ G(x, \gamma) \leq 1$  for all  $\gamma \in \mathbb{R}$ . On account of (3.33), we have (3.32).  $\square$

Coming back to (3.31), we need some care because the value function  $V_a(x)$  contains its value at  $a$ , i.e.,  $V_a(a)$  in the definitive equation. Let us consider a family of optimal stopping problem parameterized by  $\gamma \in \mathbb{R}$ :

$$(3.34) \quad \begin{aligned} V_a^\gamma(x) &\triangleq \sup_{\tau \in \mathcal{S}} \{ \mathbb{E}^x [1_{\{\tau < \tau_0\}} e^{-\alpha \tau} (\bar{K}(X_\tau, a) + \gamma)] + \mathbb{E}^x [1_{\{\tau > \tau_0\}} e^{-\alpha \tau_0} (P - g(0))] \} \\ &= \sup_{\tau \in \mathcal{S}} \mathbb{E}^x [e^{-\alpha \tau} r^\gamma(X_\tau, a)], \end{aligned}$$

where

$$r^\gamma(x, a) = \begin{cases} P - g(0), & x = 0, \\ \bar{K}(x, a) + \gamma, & x > 0. \end{cases}$$

Obviously, this parameterized problem can be solved by using Propositions 2.1–2.3. Now we link this parameterized optimal stopping problem to (3.31).

LEMMA 3.6. *For  $a > 0$  given, if there exists a solution to (3.34), then there always exists unique  $\gamma$  such that  $\gamma = V_a^\gamma(a)$  holds, provided that (3.4) holds.*

*Proof.* Without loss of generality, we need only to consider the case where

$$(3.35) \quad \sup_{x \in \mathbb{R}_+} \bar{K}(x, a) > 0$$

for some  $a > 0$ . Indeed, suppose that there is no such  $a$ , and let us consider a sequence of optimal stopping scheme. In each iteration, the value function for the optimal stopping problem takes negative values, so that  $\phi_n(\cdot) < 0$  for all  $n$ . Then in the next iteration,  $\bar{K}(x, y)$  function will be shifted downwards, leading to  $\phi_{n+1}(\cdot) < 0$ . Hence the “no interventions” strategy is trivially optimal.

In (3.34), since  $\gamma$  is some constant parameter, we benefit from Proposition 2.1 and claim that  $V_a^\gamma(x)$  is characterized as the smallest  $F$ -concave majorant of  $r^\gamma(\cdot, a)$  that passes  $(F(0), \frac{P-g(0)}{\varphi(0)})$ . In terms of the notation of Proposition 2.3, if we define  $W_a^\gamma(\cdot)$  such that

$$V_a^\gamma(x) = \varphi(x)W_a^\gamma(F(x)),$$

then  $W_a^\gamma(\cdot)$  passes through the fixed point  $A = (F(0), W_a^\gamma(F(0)))$  and is the smallest concave majorant of  $H^\gamma(\cdot, a) \triangleq \frac{r^\gamma(F^{-1}(\cdot), a)}{\varphi(F^{-1}(\cdot))} = \frac{\bar{K}(F^{-1}(\cdot), a)}{\varphi(F^{-1}(\cdot))} + \frac{\gamma}{\varphi(F^{-1}(\cdot))}$ .

Now fix  $a$ . Our approach here is that, by starting with  $\gamma = 0$ , we move  $\gamma$  upwards, evaluate  $V_a^\gamma(a)$ , and try to find  $\gamma$  such that  $\gamma = V_a^\gamma(a)$ . Due to (3.35), we have  $W_a^0(F(a)) > 0$ . By the monotonicity of  $F$ , it is equivalent to saying that  $V_a^0(a) > 0 = \gamma$ . As  $\gamma$  increases,  $W_a^\gamma(F(a))$  increases monotonically: See the right-hand side of (3.34).

Lemma 3.5 implies that, for  $\gamma_1 > \gamma_2 \geq 0$ ,

$$(3.36) \quad V_a^{\gamma_1}(x) - V_a^{\gamma_2}(x) \leq \gamma_1 - \gamma_2$$

for any  $x \in \mathbb{R}_+$ . Note that  $W_a^\gamma(F(a)) \geq H^\gamma(F(a), a)$ . However, since  $V_a^\gamma$  has less than the linear growth in  $\gamma$  as demonstrated by (3.36), there is a certain  $\gamma'$  large enough such that  $W_a^\gamma(F(a)) = H^\gamma(F(a), a)$  for  $\gamma \geq \gamma'$ . This implies that

$$\begin{aligned} \varphi(a)W_a^{\gamma'}(F(a)) &= \varphi(a)H^{\gamma'}(F(a), a) \\ \Leftrightarrow V_a^{\gamma'}(a) &= \bar{K}(a, a) + \gamma' < \gamma', \end{aligned}$$

where the inequality is due to the assumption (3.4). For this  $\gamma'$ , we have  $V_a^{\gamma'}(a) < \gamma'$ .

The monotonicity and continuity of  $W^\gamma(F(a))$  (due to the convexity of  $V_a^\gamma(\cdot)$ ) with respect to  $\gamma$ , together with (3.36), implies that, for any  $a$ , there exists one and only one  $\gamma$  such that  $V_a^\gamma(a) = \gamma$ .  $\square$

**3.3. Methodology to find  $v(x)$  and  $(a^*, b^*)$ .** By using (3.27), namely, the characterization of  $u_{a,b}$ , we describe an optimization procedure based on Propositions 2.2 and 2.3.

1. Fix  $a > 0$ . Consider the function

$$(3.37) \quad R(\cdot, a) \triangleq \frac{\bar{K}(F^{-1}(\cdot), a)}{\varphi(F^{-1}(\cdot))}.$$

Define  $W_a(\cdot)$  such that  $V_a(x) = \varphi(x)W_a(F(x))$  and, by the linear characterization (3.29), it is a straight line with a slope, say,  $\beta(a)$ , and passes through  $(F(0), W_a(F(0))) = (F(0), \frac{P-g(0)}{\varphi(0)})$ . We can write the linear majorant, in general,

$$(3.38) \quad W_a(y) = \beta(a)y + \delta.$$



2. *First stage optimization:* For each slope  $\beta(a)$ , we can calculate the value of  $W_a(F(a))$ , but we have to find the  $W_a(\cdot)$  function such that, at some point  $F(b(a))$ , we have

$$(3.39) \quad W_a(F(b)) = R(F(b), a) + W_a(F(a)) \frac{\varphi(a)}{\varphi(b)},$$

where we write  $b(a) \equiv b$  for notational simplicity. This requirement is equivalent to finding  $\gamma$  in (3.34) in Lemma 3.6 such that

$$\frac{\gamma}{\varphi(a)} = W_a^\gamma(F(a)).$$

Let us denote the right-hand side of (3.39) by

$$(3.40) \quad \Phi(y, a) \triangleq R(y, a) + W_a(F(a)) \frac{\varphi(a)}{\varphi(F^{-1}(y))}, \quad y \in (F(c), F(d)).$$

By Proposition 3.4,  $(0, b(a))$  is the continuation region. If  $R$  is a differentiable function with respect to the first argument, we can find the optimal point  $b(a)$  analytically. In fact, it is to find a point  $b(a)$  such that the linear majorant and the shifted function  $\Phi(y, a)$  have a tangency point. This is equivalent to calculating the smallest linear majorant of  $\Phi(y, a)$  (due to Proposition 3.3 and the linear characterization (3.29)). Explicitly, we solve

$$(3.41) \quad \left( \frac{\bar{K}(b, a)}{\varphi(b)} \right)' - \frac{\varphi'(b)\varphi(a)}{\varphi(b)^2} \delta = \beta(a) \left( F'(b) + \frac{\varphi'(b)\varphi(a)}{\varphi(b)^2} F(a) \right)$$

for  $b(a)$ , where  $\beta(a)$  is

$$(3.42) \quad \beta(a) = \frac{\varphi(b)R(F(b), a) - \delta(\varphi(b) - \varphi(a))}{F(b)\varphi(b) - F(a)\varphi(a)}.$$

For the absorbing boundary case, these equations can be easily modified. Let us denote that  $D \triangleq W_a(F(0)) = (P - g(0))/\varphi(0)$ . Then (3.41) and (3.42) become, by substituting

$$(3.43) \quad \delta = D - \beta(a)F(0)$$

in (3.41),

$$(3.44) \quad \left( \frac{\bar{K}(b, a)}{\varphi(b)} \right)' - \frac{\varphi'(b)\varphi(a)}{\varphi(b)^2} D = \beta(a) \left( F'(b) + \frac{\varphi'(b)\varphi(a)}{\varphi(b)^2} (F(a) - F(0)) \right)$$

and

$$(3.45) \quad \beta(a) = \frac{\varphi(b)R(F(b), a) - D(\varphi(b) - \varphi(a))}{(F(b) - F(0))\varphi(b) - (F(a) - F(0))\varphi(a)},$$

respectively.

3. *Second stage optimization:* To summarize up to this point, we set  $a$  and find  $b(a)$  and in turn  $\beta(a)$ . Now, let  $a$  vary, and choose, among  $\beta(a)$ , the largest slope  $\beta^* \triangleq \max_{a \in \mathbb{R}_+} \beta(a)$ , if it exists, and also the corresponding  $a^*$  and

$b(a^*)$ . Due to the characterization of the value function with (3.30), these  $a^*$  and  $b^* \triangleq b(a^*)$  must be the solution to (3.8).

If  $\bar{K}(x, y)$  is a differentiable function with respect to both the first and second arguments, then we can find  $a^*$  analytically. If, for any  $a \in \mathbb{R}_+$ ,  $\Phi(y, a)$  is strictly concave in  $y$  at  $F(b)$ ,  $a^*$  must satisfy

$$(3.46) \quad \left( \frac{\partial \bar{K}(b, a)}{\partial a} + \delta \varphi'(a) \right) (\psi(b) - \psi(a)) \\ = - (\bar{K}(b, a) - \delta(\varphi(b) - \varphi(a))) (F'(a)\varphi(a) + F(a)\varphi'(a)).$$

Therefore, in this case, our nonlinear optimization procedure is just to solve (3.41) and (3.46) with (3.42), simultaneously. We postpone the derivation of (3.46) to the appendix.

*Remark 3.1.* With respect to the third point of the proposed method above, we should check if there exists a concave majorant as  $a \downarrow 0$ . Namely, we consider whether

$$\lim_{a \downarrow 0} (K(x, a) - (g(x) - g(a)) + u(a)) < P - g(0)$$

holds in the neighborhood of  $a = 0$ . Since  $\lim_{x \downarrow 0} g(x) = \lim_{a \downarrow 0} g(a)$  and  $\lim_{a \rightarrow 0} u(a) = u(0) = P - g(0)$  by the continuity of  $u$ , the last inequality holds due to (3.4).

### 3.4. Characterization of the intervention times and the value function:

**General case.** Let us move on to a general case where the mapping  $x \rightarrow \frac{\bar{K}}{\varphi}(x) : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is not necessarily  $F$ -concave. First, we extend Proposition 3.4 to characterize optimal intervention times.

**PROPOSITION 3.7.** *The value function  $v(x)$  for (3.8) is given by the smallest solution majorizing  $g$  of  $v - g = \mathcal{L}(v - g)$ , and optimal intervention times  $T_i^*$  are given by exit times from an interval if and only if, for all  $y \in \mathbb{R}_+$ ,*

$$(3.47) \quad x \rightarrow \bar{K}(x, y) \text{ is continuous and } q \triangleq \limsup_{x \rightarrow \infty} D^- \left( \frac{\bar{K}}{\varphi} \circ F^{-1} \right) (x) \text{ is finite,}$$

where  $D^- f(x_0) \triangleq \limsup_{x \uparrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$ .

*Proof.* For any given  $a \in \mathbb{R}_+$ , if we can find the smallest linear majorant of  $\frac{\bar{K}(F^{-1}(\cdot), a) + \gamma}{\varphi(F^{-1}(\cdot))}$  for an arbitrary  $\gamma \in \mathbb{R}_+$ , we can find  $\gamma = \varphi(a)W_a(F(a))$  by Lemma 3.6. Due to the constancy of  $\gamma$ , it suffices to show that condition (3.47) is necessary and sufficient for the existence of concave majorant of  $\frac{\bar{K}}{\varphi} \circ F^{-1}$  on  $F(\mathcal{I})$ . The sufficiency is immediate. For the necessity, we assume that  $q = +\infty$ . We can take a sequence of points  $\{x^k\} \subset \mathbb{R}$  such that  $x^k \rightarrow \infty$  and  $D^-(\frac{\bar{K}}{\varphi} \circ F^{-1})(x^k) \rightarrow \infty$  as  $k \rightarrow \infty$ . If necessary, by taking a subsequence, we can make this sequence  $\{x^k\}$  monotone. Consider the smallest concave majorant of  $\frac{\bar{K}}{\varphi} \circ F^{-1}$  on  $[F(0), F(x^k)]$ . Call it  $v^k(x)$ . It is clear that  $v^k(x)$  is monotone increasing in  $k$  for all  $x \in [F(0), F(x^k)]$ . As  $k \rightarrow \infty$ ,  $x^k \rightarrow \infty$  and  $v(x) \geq v^k(x)$ . We thus have  $v(x) = \lim_{k \rightarrow \infty} v^k(x) = \infty$  for all  $x \in \mathbb{R}_+$ . There is no optimal intervention policy.  $\square$

Suppose that the  $F$ -concavity of the reward function is violated, so that the intervention point may be multiple. For the rest of this subsection, as is practically the case (see our examples in section 4), we study the case where  $\Phi(y, a)$ , with  $y = F(x)$ , is increasing to infinity and becomes eventually concave and then derive linear characterization as in section 3.2.

Let us consider a strategy where we have two intervention points  $b_1$  and  $b_2$  being arbitrarily chosen such that  $0 < b_1 < b_2$ . We want to characterize function  $J^\nu(x)$  as in (3.5) again. Recall that there are no controls in a way that the process is pulled up to avoid ruin. In other words,  $\mathbb{P}^x[\tau_0 < \infty] = 1$ . Assume, for the moment, that we always apply control at these boundaries  $b_1$  and  $b_2$  and then, once applied, the process moves to  $a_1 < b_1$  and  $a_2 < b_2$ , respectively.

If we start with a point  $x \in [0, b_1]$ , the problem is equivalent to the case we considered already, since the process cannot go beyond the level  $b_1$ . Hence by following (3.26), we have for  $x \in [0, b_1]$

$$J_1^\nu(x) \triangleq \mathbb{E}^x \left[ \int_0^{\tau_0} e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau_0} P + \sum_{T_i < \tau_0} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right]$$

and

$$u_1(x) = \mathbb{E}^x[1_{\{\tau_{b_1} < \tau_0\}} e^{-\alpha \tau_{b_1}} u_1(b_1)] + \mathbb{E}^x[1_{\{\tau_{b_1} > \tau_0\}} e^{-\alpha \tau_0} u_1(0)], \quad x \in [0, b_1],$$

by defining  $u_1(x) \triangleq J_1^\nu(x) - g(x)$ . If we start with a point  $x \in [b_1, b_2]$ , there are two strategies available:

- (A) Let  $X_t$  move along. (It hits either  $b_1$  or  $b_2$  first.)
- (B) Apply the control immediately ( $t = 0$ ) by moving the process from  $x$  to  $a_1$  (the postcontrol point that corresponds to  $b_1$ ), and let the process start at  $a_1$ . (Recall that we do not let  $X$  enter into  $(b_1, \infty)$  after moving to  $a_1$ .)

Consider strategy (A) first. Let us define

$$J_2^\nu(x) \triangleq \mathbb{E}^x \left[ \int_0^{\tau_{b_1}} e^{-\alpha s} f(X_s) ds + \sum_{T_i < \tau_{b_1}} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right] \quad x \in [b_1, b_2].$$

By using the strong Markov property at the first intervention, we can reduce  $J_2^\nu$  to a simpler form. For any  $(a_1, b_1)$  and  $(a_2, b_2)$ , we have

$$\begin{aligned} J_2^\nu(x) &= \mathbb{E}^x[1_{\{\tau_{b_1} < \tau_{b_2}\}} e^{-\alpha \tau_{b_1}} K(X_{\tau_{b_1}-}, X_{\tau_{b_1}}) - g(X_{\tau_{b_1}}) + J_1^\nu(X_{\tau_{b_1}}) + g(X_{\tau_{b_1}}) - g(X_{\tau_{b_1}-})] \\ &+ \mathbb{E}^x[1_{\{\tau_{b_1} > \tau_{b_2}\}} e^{-\alpha \tau_{b_2}} K(X_{\tau_{b_2}-}, X_{\tau_{b_2}}) - g(X_{\tau_{b_2}}) + J_2^\nu(X_{\tau_{b_2}}) + g(X_{\tau_{b_2}}) - g(X_{\tau_{b_2}-})] + g(x). \end{aligned}$$

We shall use  $u_1(x) = J_1^\nu(x) - g(x)$  in the first term. Now let us define  $u_2(x) \triangleq J_2^\nu(x) - g(x)$ . Then the last equation becomes

$$\begin{aligned} u_2(x) &= \mathbb{E}^x[1_{\{\tau_{b_1} < \tau_{b_2}\}} e^{-\alpha \tau_{b_1}} K(X_{\tau_{b_1}-}, X_{\tau_{b_1}}) + u_1(X_{\tau_{b_1}}) + g(X_{\tau_{b_1}}) - g(X_{\tau_{b_1}-})] \\ &+ \mathbb{E}^x[1_{\{\tau_{b_1} > \tau_{b_2}\}} e^{-\alpha \tau_{b_2}} K(X_{\tau_{b_2}-}, X_{\tau_{b_2}}) + u_2(X_{\tau_{b_2}}) + g(X_{\tau_{b_2}}) - g(X_{\tau_{b_2}-})] \\ &= \mathbb{E}^x[1_{\{\tau_{b_1} < \tau_{b_2}\}} e^{-\alpha \tau_{b_1}} (\bar{K}(b_1, a_1) + u_1(a_1))] \\ (3.48) \quad &+ \mathbb{E}^x[1_{\{\tau_{b_1} > \tau_{b_2}\}} e^{-\alpha \tau_{b_2}} (\bar{K}(b_2, a_2) + u_2(a_2))] \end{aligned}$$

on  $x \in [b_1, b_2]$ . By identifying  $\bar{K}(b_2, a_2) + u_2(a_2) = u_2(b_2)$  and  $u_2(b_1) = \bar{K}(b_1, a_1) + u_1(a_1) = u_1(b_1)$  (the latter shows that  $u_1(x)$  and  $u_2(x)$  are connected at  $x = b_1$ ),

$$(3.49) \quad u_2(x) = \frac{\varphi(x)}{\varphi(b_1)} \frac{F(b_2) - F(x)}{F(b_2) - F(b_1)} u_2(b_1) + \frac{\varphi(x)}{\varphi(b_2)} \frac{F(x) - F(b_1)}{F(b_2) - F(b_1)} u_2(b_2), \quad x \in [b_1, b_2].$$

To summarize this result, if we define  $W_i(\cdot) \triangleq \frac{u_i}{\varphi} \circ F^{-1}(\cdot)$  for  $i = 1, 2$  on  $F(\mathcal{I})$ , this is again a linear function for each  $i$ . Hence by defining

$$W_A(F(x)) \triangleq \begin{cases} W_1(F(x)) = W_1(F(0)) \frac{F(b_1) - F(x)}{F(b_1) - F(0)} + W_1(F(b_1)) \frac{F(x) - F(0)}{F(b_1) - F(0)}, & x \in [0, b_1], \\ W_2(F(x)) = W_2(F(b_1)) \frac{F(b_2) - F(x)}{F(b_2) - F(b_1)} + W_2(F(b_2)) \frac{F(x) - F(b_1)}{F(b_2) - F(b_1)}, & x \in [b_1, b_2], \end{cases}$$

we have a piecewise linear function on  $F(\mathcal{I})$ .

Next consider strategy (B), whose value function is

$$(3.50) \quad W_B(F(x)) \triangleq \begin{cases} W_1(F(x)), & 0 \leq x \leq b_1, \\ \overline{W}_1(F(x)) \triangleq \frac{\varphi(a_1)}{\varphi(x)} W_1(F(a_1)) + R(F(x), a_1), & b_1 < x. \end{cases}$$

LEMMA 3.8. (A) is better than (B) only if

$$\beta_1 \triangleq \frac{W(F(b_1)) - W(F(0))}{F(b_1) - F(0)} < \frac{W(F(b_2)) - W(F(b_1))}{F(b_2) - F(b_1)} \triangleq \beta_2.$$

*Proof.* Since the value function of strategy (B) is (3.50), choosing (A) over (B) is equivalent to

$$\overline{W}_1(F(x)) < W_2(F(x)) \quad \text{on } x > b_1.$$

If  $W_1(F(x))$  majorizes  $\overline{W}_1(F(x))$  on  $x \in [0, \infty)$ , then this problem reduces to the  $F$ -concavity case discussed in the previous subsection. Hence we consider the case where there exists some  $x \in [b_1, \infty)$  such that

$$W_1(F(x)) < \overline{W}_1(F(x)).$$

Now suppose that we have  $\beta_1 \geq \beta_2$ . Then it is clear that we cannot have  $W_2(F(x)) > \overline{W}_1(F(x))$  on  $x \in [b_1, \infty)$ .  $\square$

There are two cases to consider:

- (1) If  $W_2(F(x))$  majorizes  $\overline{W}_1(F(x))$  on  $x \in [b_1, \infty)$ , then we adopt the point  $b_2$  as an intervention point. In this case,  $\beta_2 > \beta_1$  holds. However, this implies that if we connect  $A \triangleq (F(0), W_1(F(0)))$  and  $C \triangleq (F(b_2), W_2(F(b_2)))$ , then this line segment  $AC$  is above the line segment connecting, piece by piece, points  $A$ ,  $B \triangleq (F(b_1), W_1(F(b_1)))$  and  $C$ . We can show that there exists a point  $b' \geq b_2$  such that its corresponding linear majorant  $W'(F(x))$  satisfies  $W'(F(x)) > W_1(F(x))$  on  $x \in [0, b_1]$  and  $W'(F(x)) > W_2(F(x))$  on  $[b_1, b_2]$ . The proof of the existence of a postintervention point  $a'$  corresponding to this point  $b'$  follows in a similar manner to Lemma 3.6.
- (2) If  $W_2(F(x))$  does not majorize  $\overline{W}_1(F(x))$ , we can find another point  $\bar{b}$ , instead of  $b_1$ , such that the linear (not piecewise linear) function  $W(F(x))$  corresponding to  $\bar{b}$  majorizes  $R(F(x), \bar{a}) + W(F(\bar{a})) \frac{\varphi(\bar{a})}{\varphi(x)}$  on  $x \in \mathbb{R}_+$  by Proposition 3.7.

In either case, the value function in the transformed space should be a linear function that attains the largest slope among all of the possible linear majorants. This argument holds true for any  $b_1$  and  $b_2$  with  $b_1 < b_2$ . We can continue this argument inductively to the case of  $n$  intervention points  $(b_1, \dots, b_n)$ . We here summarize our argument up to this point as a main proposition.

PROPOSITION 3.9. Suppose that (3.47) holds and the optimal continuation region is connected. The value function corresponding to (3.5) of the impulse problem described in (3.3)–(3.8) is written as

$$(3.51) \quad v(x) = \begin{cases} v_0(x) \triangleq \varphi(x)W^*(F(x)) + g(x), & 0 \leq x \leq b^*, \\ v_0(a^*) + K(x, a^*), & b^* \leq x, \end{cases}$$

where  $W^*(\cdot)$  is the line segment connecting  $(F(0), W^*(F(0)))$  and  $(F(b^*), W^*(F(b^*)))$  and satisfies the following:

1.  $W^*(F(\cdot))$  is the smallest linear majorant of  $W^*(F(a^*))\frac{\varphi(a^*)}{\varphi(\cdot)} + R(F(\cdot), a^*)$  and meets with  $W^*(F(a^*)) + R(F(\cdot), a^*)\frac{\varphi(a^*)}{\varphi(\cdot)}$  at point  $F(b^*)$  and passes  $(F(0), \frac{P-g(0)}{\varphi(0)})$ . If  $R$  is differentiable,  $(a^*, b^*)$  satisfy (3.41).
2. The slope of  $W^*(\cdot)$ , denoted as  $\beta^*$ , is the largest slope among  $\beta(a)$ 's of all of the possible linear majorants  $W_a(\cdot)$ .

Moreover, if the mapping  $x \rightarrow \frac{\bar{K}}{\varphi}(x) : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is  $F$ -concave, then the optimal continuation region  $(0, b^*)$  is uniquely determined.

Note that, at  $x = 0$ ,

$$v(0) = \varphi(0)W^*(F(0)) + g(0) = \varphi(0)\frac{P-g(0)}{\varphi(0)} + g(0) = P$$

as expected.

Remark 3.2. If the  $F$ -concavity of  $\bar{K}/\varphi$  is violated, there are at least two possible cases (and a combination of them) where we have multiple continuation regions with a linear value function in the continuation region of the transformed space.

1. For some  $a_i^*$ , with  $i = 1, 2, \dots$ , we have the common  $\beta^*$ . This is the case which we shall show in the next example. In this case, the continuation region is  $C = \{(0, b_1^*), (b_1^*, b_2^*), (b_2^*, b_3^*) \dots\}$ , where  $b_i^*$  corresponds to  $a_i^*$  for each  $i$ , and the intervention region is  $\Gamma = \{\{b_1^*\}, \{b_2^*\}, \{b_3^*\} \dots\}$ . Each time the process hits one of the points  $\{b_i^*\}$ , the control pulls the process back to the corresponding  $a_i^*$ .
2. Another case is that, for the unique optimal  $a^*$ , there exist nonunique  $b_1^*$  and  $b_2^*$ . In this case, the continuation region is  $C = \{(0, b_1^*), (b_1^*, b_2^*)\}$ , and the stopping region is  $\Gamma = \{\{b_1^*\}, [b_2^*, \infty)\}$ . If the process hits  $b_1^*$  or  $b_2^*$ , then the control pulls the process back to  $a^*$  in either situation. It makes sense to continue in the region  $(b_1^*, b_2^*)$  because there is a positive probability that one can extract  $\bar{K}(b_2^*, a^*) (> \bar{K}(b_1^*, a^*))$  within a finite time.

**3.5. No absorbing boundary case.** Next, we extend our argument to a problem without the absorbing boundary. Hence the process can move along in the state space in an infinite amount of time. The problem becomes

$$(3.52) \quad J^\nu(x) = \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} f(X_s) ds + \sum_{i=1} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right].$$

We can characterize intervention times as exit times from a certain boundary and simplify the performance measure (3.52):

$$\begin{aligned} J^\nu(x) &= \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} f(X_s) ds + \sum_{i=1} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right] \\ &= \mathbb{E}^x [e^{-\alpha T_1} \{K(X_{T_1-}, X_{T_1}) - g(X_{T_1-}) + J^\nu(X_{T_1})\}] + g(x). \end{aligned}$$

The second equation is easily obtained in the same way as in the previous section by noting that  $\mathbb{P}^x(T_1 < \infty) = 1$ . The last term does not depend on controls, so we define  $u(x) \triangleq J^\nu(x) - g(x)$ :

$$u(x) = \mathbb{E}^x[e^{-\alpha T_1}\{K(X_{T_1-}, X_{T_1}) - g(X_{T_1-}) + g(X_{T_1}) + u(X_{T_1})\}].$$

Again, we consider the  $F$ -concave case with the notation  $T_i- = \tau_b$  for all  $i$ , and we have

$$u(x) = \mathbb{E}^x[e^{-\alpha \tau_b}(K(b, a) - g(b) + g(a) + u(a))] = \mathbb{E}^x[e^{-\alpha \tau_b}(\bar{K}(b, a) + u(a))].$$

By defining  $W = (u/\varphi) \circ F^{-1}$ , we have

$$W(F(x)) = W(F(c))\frac{F(b) - F(x)}{F(b) - F(c)} + W(F(b))\frac{F(x) - F(c)}{F(b) - F(c)}, \quad x \in (c, b].$$

We should note that  $F(c) \triangleq F(c+) = \psi(c+)/\varphi(c+) = 0$  and

$$W(F(c)) = l_c \triangleq \limsup_{x \downarrow c} \frac{\bar{K}(x, a)^+}{\varphi(x)}$$

for any  $a \in (c, d]$ . For a more detailed mathematical meaning of this value  $l_c$ , we refer the reader to Dayanik and Karatzas [13]. We can effectively consider  $(F(c), l_c)$  as the absorbing boundary.

**4. Examples.** In this section, we work out some examples from financial engineering problems. As described in section 3.3, the main task to find solutions now reduces to analyzing the reward function in the transformed space and finding the smallest linear majorant. Let us recall that, for given  $a \in \mathbb{R}$ , the shifted reward function in the transformed space is (3.40), that is,

$$(4.1) \quad \Phi(y, a) = R(y, a) + W_a(F(a))\frac{\varphi(a)}{\varphi(F^{-1}(y))}, \quad y \in (F(c), F(d)),$$

in which  $R(\cdot, a)$  is given by (3.37) and  $W_a(\cdot)$  is given by (3.38). For the purpose of analyzing  $\Phi(\cdot, a)$ , we recall some useful observations: If  $h(\cdot)$  is twice-differentiable at  $x \in \mathcal{I}$  and  $y \triangleq F(x)$ , then  $H'(y) = m(x)$  and  $H''(y) = m'(x)/F'(x)$ , with

$$(4.2) \quad m(x) = \frac{1}{F'(x)} \left( \frac{h}{\varphi} \right)'(x), \quad \text{and} \quad H''(y)(\mathcal{A} - \alpha)h(x) \geq 0, \quad y = F(x),$$

with strict inequality if  $H''(y) \neq 0$ . These identities are of practical use in identifying the concavities of  $H(\cdot)$  when it is hard to calculate its derivatives explicitly. Before the transformation defined by (3.28), (4.1) is of the form  $\bar{K}(x, a) + u(a)$ . Hence, for a fixed  $a$ , we read  $h(x) = \bar{K}(x, a)$  for the first term of (4.1) and  $h(x) = \text{constant}$  for the second term of (4.1) to apply (4.2).

*Remark 4.1.* It is worth examining  $h(x) = k$ , where  $k \in \mathbb{R}_+$  is a constant. The first equation of (4.2) is  $m(x) = -\frac{k\varphi'(x)}{F'(x)\varphi(x)^2} > 0$  since  $\varphi'(x) < 0$  and  $F'(x) > 0$  for all  $x \in \mathcal{I}$ . As to the second inequality of (4.2),  $(\mathcal{A} - \alpha)k = -\alpha k < 0$ . These facts imply that the second term of (4.1) is always increasing and concave in the transformed space for an  $a$  that makes  $W_a(F(a)) > 0$ .

*Example 4.1.* Øksendal [24] considers the following problem:

$$(4.3) \quad J_o^\nu(x) = \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} X_s^2 ds + \sum_i^\infty e^{-\alpha T_i} (c + \lambda \xi_i) \right],$$

where  $X_t^0 = B_t$  is a standard Brownian motion and  $c > 0$  and  $\lambda \geq 0$  are constants. The Brownian motion represents the exchange rate of some currency, and each impulse represents an interventions taken by the central bank in order to keep the exchange rate in a given target zone. Here we are allowed only to give the system impulses  $\zeta$  with values in  $(0, +\infty)$ . By reducing a level from  $b$  to  $a$  (i.e.,  $b > a$ ) through interventions, one can save continuously incurred cost (which is high if the process is at a high level). The problem is to minimize the expected total discounted cost  $v_o(x) = \inf_\nu J_o^\nu(x)$ . We want to solve its sup version and change the sign afterwards (i.e.,  $v_o(x) = -v(x)$ ):

$$v(x) = \sup_\nu \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} (-X_s^2) ds - \sum_i^\infty e^{-\alpha T_i} (c + \lambda \xi_i) \right].$$

*Data:* The continuous cost rate  $f(x) = -x^2$ , and the intervention cost is  $K(x, y) = -c - \lambda(x - y)$  in our terminology. By solving the equation  $(\mathcal{A} - \alpha)v(x) = \frac{1}{2}v''(x) - \alpha v(x) = 0$ , we find  $\psi(x) = e^{x\sqrt{2\alpha}}$  and  $\varphi(x) = e^{-x\sqrt{2\alpha}}$ . Hence  $F(x) = e^{2x\sqrt{2\alpha}}$  and  $F^{-1}(x) = \frac{\log x}{2\sqrt{2\alpha}}$ .  $g(x)$  can be calculated by Fubini's theorem:

$$g(x) = \mathbb{E}^x \left[ \int_0^\infty -e^{-\alpha s} (x + B_s)^2 ds \right] = - \left( \frac{x^2}{\alpha} + \frac{1}{\alpha^2} \right).$$

Note that, when  $b > a$ ,  $g(a) - g(b) > 0$  is the source of cost savings. Hence  $\bar{K}(x, a) = -c - \lambda(x - a) + \frac{x^2 - a^2}{\alpha}$ .

*Analysis of the reward function:* Let us fix  $a > 0$  and consider  $h(x) \triangleq \bar{K}(x, a) = -c - \lambda(x - a) + \frac{x^2 - a^2}{\alpha}$  and  $H(y) \triangleq (h/\varphi)(F^{-1}(y))$ ,  $y > 0$ . By the first equation in (4.2), the sign of  $(\frac{h}{\varphi})'(x)$  will lead us to conclude that  $H(F(x))$  is increasing to infinity from a certain point, say,  $x = p$  on  $(p, \infty)$ , and so is  $H(F(x))$ . Also, by direct calculation,  $H'(+\infty) = 0$ , from which we can assert that the value function is finite by Proposition 3.7.

If we set  $p(x) \triangleq -x^2 + a^2 + \lambda\alpha(x - a) + \alpha c + 1/\alpha$ , then  $(\mathcal{A} - \alpha)h(x) = p(x)$  for every  $x > 0$ . This quadratic function  $p(x)$  possibly has one or two positive roots. Let  $k$  be the largest one. Since  $\lim_{x \rightarrow \infty} p(x) = -\infty$ , by the second inequality in (4.2),  $H(\cdot)$  is concave on  $(F(k), +\infty)$ . Hence  $H(y, a)$  is increasing (to infinity) and concave on  $y \in (F(k), \infty)$ . Since it is obvious that there exists  $a$  such that  $W(F(a)) > 0$ , due to Remark 4.1, the second term of (4.1) is increasing and concave on  $\mathbb{R}_+$ .

Since the cost function in the transformed space is increasing and concave from a certain point on, there is a linear majorant that touches the cost function once and only once. We can conclude that, for any  $a > 0$  and the parameter set, we have a connected continuation region in the form of  $(0, b^*)$ .

*Solution:* For a fixed  $a$ , let us define  $W_a(\cdot)$  such that  $V_a(x) = \varphi(x)W_a(F(x))$  and  $r(x, a) = -c$  if  $x < a$  and  $r(x, a) = h(x) = -c - \lambda(x - a) + \frac{x^2}{\alpha} - \frac{a^2}{\alpha}$  if  $x \geq a$ . Then we have, for any  $a > 0$ ,

$$(4.4) \quad l_{-\infty} = \limsup_{x \downarrow -\infty} \frac{r(x, a)^+}{\varphi(x)} = 0.$$

Recall that the left boundary  $-\infty$  is natural for a Brownian motion. Hence  $W_a(y)$  that passes the origin of the transformed space is the straight-line majorant of  $R(\cdot, a) + W_a(F(a))/\varphi(F^{-1}(\cdot))$ , where  $R(\cdot, a)$  is defined in (3.37):

$$R(y, a) = \begin{cases} -c\sqrt{y}, & 0 \leq y \leq F(a), \\ H(y, a) = \sqrt{y} \left( -c - \frac{\lambda}{2\sqrt{2\alpha}} \log y + \lambda a + \frac{(\log y)^2}{8\alpha^2} - \frac{a^2}{\alpha} \right), & y > F(a). \end{cases}$$

We can represent  $W_a$  as  $W(y) = \beta y$ . Since  $R(x, a)$  is differentiable with respect to  $x$  on  $x \geq a$ , we can use (3.41) to find  $b(a)$  and corresponding  $\beta(a)$ . Then by varying  $a$ , one can find the optimal  $(a^*, b^*, \beta^*)$ . Going back to the original space, on  $x \in (-\infty, b^*]$

$$\tilde{v}(x) \triangleq \sup u(x) = \varphi(x)W^*(F(x)) = \varphi(x)(\beta^*)F(x) = \beta^*e^{x\sqrt{2\alpha}}.$$

To get  $v(x) = \sup_{\nu} J^{\nu}(x)$ , we add back  $g(x)$ :

$$v(x) = \tilde{v}(x) + g(x) = \beta^*e^{x\sqrt{2\alpha}} - \left( \frac{x^2}{\alpha} + \frac{1}{\alpha^2} \right).$$

Finally, flip the sign, and obtain the optimal cost function

$$v_o(x) = \begin{cases} \hat{v}_o(x) \triangleq \left( \frac{x^2}{\alpha} + \frac{1}{\alpha^2} \right) - \beta^*e^{x\sqrt{2\alpha}}, & 0 \leq x \leq b^*, \\ \hat{v}_o(a^*) + c + \lambda(x - a^*), & b^* \leq x, \end{cases}$$

which coincides with the solution given by Øksendal [24]. Figure 1 displays the solution with parameters  $(c, \lambda, \alpha) = (150, 50, 0.2)$ .

*Example 4.2.* This example is a dividend payout problem where the underlying process follows an Ornstein–Uhlenbeck process. This problem was originally studied by Cadenillas, Sarkar, and Zapatero [6]. Suppose that  $X^0$  has the dynamics

$$dX_t^0 = \delta(m - X_t)dt + \sigma dW_t, \quad t \geq 0,$$

where  $\delta > 0$ ,  $\sigma > 0$ , and  $m \in \mathbb{R}$ . Only positive impulse is allowed in this problem. We consider the impulse control problem

$$v(x) \triangleq \sup_{\nu \in S} \mathbb{E}^x \left[ \sum_{T_i < \tau_0}^{\infty} e^{-\alpha T_i} (-K + k\xi_i^{\gamma}) \right],$$

with some positive constants  $K, k$  and the risk-aversion parameter  $\gamma \in (0, 1]$ .

*Data:*  $x = 0$  is an absorbing boundary. Since  $\xi \in \mathbb{R}_+$ , we have

$$\bar{K}(x, y) = k(x - y)^{\gamma} - K, \quad x > y > 0.$$

Since  $f(x) = 0$  for all  $x \in \mathbb{R}$ , we have  $g(x) = 0$ . The functions  $\psi(\cdot)$  and  $\varphi(\cdot)$  are positive, increasing, and decreasing solutions of the differential equation  $(\mathcal{A} - \alpha)v(x) = (1/2)\sigma^2 v''(x) + \delta(m - x)v'(x) - \alpha v(x) = 0$ . We denote, by  $\tilde{\psi}(\cdot)$  and  $\tilde{\varphi}(\cdot)$ , the functions of the fundamental solutions for the auxiliary process  $Z_t \triangleq (X_t - m)/\sigma, t \geq 0$ , which satisfies  $dZ_t = -\delta Z_t dt + dW_t$ . For every  $x \in \mathbb{R}$ ,

$$\tilde{\psi}(x) = e^{\delta x^2/2} \mathcal{D}_{-\alpha/\delta}(-x\sqrt{2\delta}), \quad \tilde{\varphi}(x) = e^{\delta x^2/2} \mathcal{D}_{-\alpha/\delta}(x\sqrt{2\delta}),$$



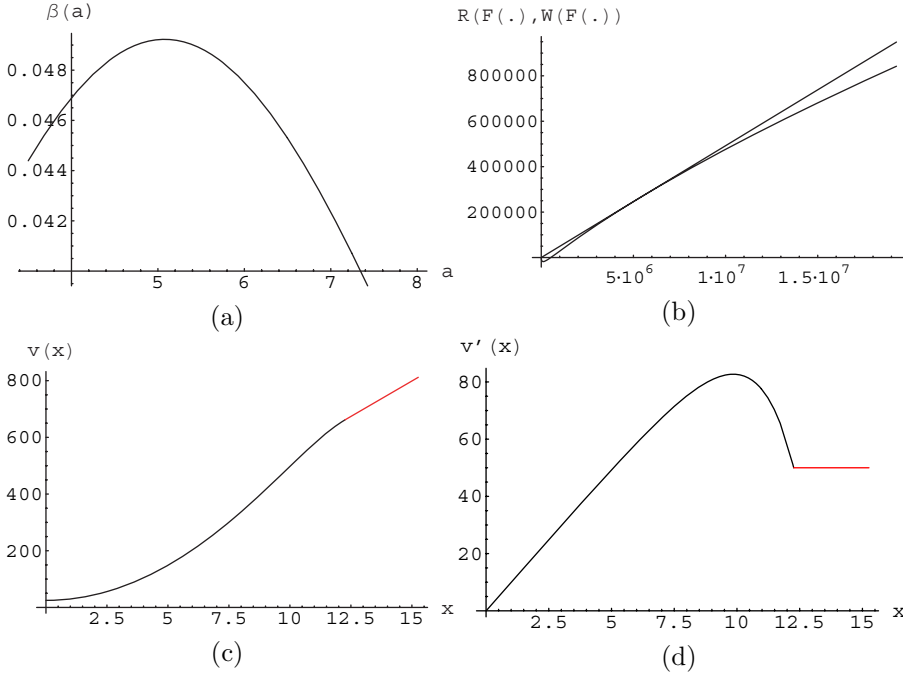


FIG. 1. (a) The plot of  $\beta(a)$  against  $a$ , the former being maximized at  $a^* = 5.077$  with  $\beta^* = 0.0492$ . (b) The functions  $R(F(\cdot), a^*)$  shifted by the amount  $W_{a^*}(F(a^*)) \frac{\varphi(a)}{\varphi(x)}$  (lower curve) and the majorant  $W_{a^*}(F(\cdot))$  (upper curve) corresponding to  $a^*$ , giving us  $b^* = 12.261$ . (c) The cost function  $v_o(x)$ . (d) The derivative of  $v_o(x)$ , showing that the smooth-fit principle holds at  $b^*$ .

$\psi(x) = \tilde{\psi}((x-m)/\sigma)$ , and  $\varphi(x) = \tilde{\varphi}((x-m)/\sigma)$ , where  $\mathcal{D}_\nu(\cdot)$  is the parabolic cylinder function (see [5, Appendices 1.24 and 2.9] and [8, Section 6.3]). By using the relation

$$(4.5) \quad \mathcal{D}_\nu(z) = 2^{-\nu/2} e^{-z^2/4} \mathcal{H}_\nu(z/\sqrt{2}), \quad z \in \mathbb{R},$$

in terms of the Hermite function  $\mathcal{H}_\nu$  of degree  $\nu$  and its integral representation,

$$(4.6) \quad \mathcal{H}_\nu(z) = \frac{1}{\Gamma(-\nu)} \int_0^\infty e^{-t^2 - 2tz} t^{-\nu-1} dt, \quad \operatorname{Re}(\nu) < 0$$

(see, for example, [21, pp. 284, 290]).

*Analysis of the reward function:* Let us consider the function

$$h(x) \triangleq k(x-a)^\gamma - K, \quad x > 0, \gamma \in (0, 1].$$

Since the function  $h(\cdot)$  is positive when  $x$  is large enough and it is increasing to infinity on the whole real line, so is the function  $H(y) = (h/\varphi) \circ F^{-1}(y)$ ,  $y \in (0, \infty)$ .

Since it is obvious that there exists  $a$  such that  $W(F(a)) > 0$ , due to Remark 4.1, the second term of (4.1) is increasing and concave on  $(F(0), \infty)$ . Let us concentrate on the first term and define the function

$$p(x) \triangleq \frac{1}{2} \sigma^2 k \gamma (\gamma - 1) x^{\gamma-2} + m \delta k \gamma x^{\gamma-1} - k(\delta \gamma + \alpha) x^\gamma + \alpha K,$$

which satisfies  $(\mathcal{A} - \alpha)h(x) = p(x)$ . By using (4.2),  $H''(y)$  and  $p(F^{-1}(y))$  have the same sign at every  $y$  where  $h$  is twice-differentiable. Hence we study the (positive)

roots of  $p(x) = 0$ . We have to divide two cases: (1)  $\gamma = 1$  and (2)  $\gamma < 1$ . In either case, it can be shown that  $H'(+\infty) < \infty$  by using (4.5) and (4.6) and the identity  $\mathcal{H}'_\nu(z) = 2\nu\mathcal{H}_{\nu-1}(z)$ ,  $z \in \mathbb{R}$ . Therefore, the finiteness of the value function is proved.

- (1)  $\gamma = 1$ :  $h(\cdot)$  reduces to a linear function, and  $p(x) = 0$  always has one positive root, say,  $p > 0$ . The  $H(\cdot)$  function is convex on  $[0, F(p))$  and concave on  $(F(p), +\infty)$ . Hence we have a connected continuation region  $(0, b^*)$ .
- (2)  $\gamma < 1$ : We observe that  $\lim_{x \downarrow 0} p(x) = -\infty$ ,  $\lim_{x \uparrow +\infty} p(x) = -\infty$ ,  $\lim_{x \downarrow 0} p'(x) = +\infty$ , and  $\lim_{x \uparrow +\infty} p'(x) = 0-$ . A direct analysis of  $p'(x)$  shows that there is only one stationary point in  $(0, \infty)$ , and the number of the roots of  $p(x) = 0$  is either 0, 1, or 2. Hence, in the first two cases,  $H(\cdot)$  is concave on  $[0, \infty)$  and the continuation region is connected. In the last case there are two roots, say,  $0 < p_1 < p_2$ . The  $H(\cdot)$  function is then concave on  $[0, F(p_1)) \cup (F(p_2), +\infty)$  and is convex on  $(F(p_1), F(p_2))$ . Since  $H(\cdot)$  increases and is concave on  $y \in (F(p_2), \infty)$ , we can conclude that the continuation region is connected in this case as well.

*Solution:* Let us move on to finding an optimal continuation region. Unlike the previous example, it is not easy (at least analytically) to find  $F^{-1}(y)$  explicitly. But it is not necessary. We can solve (3.44) for  $b(a)$  with  $D = 0$ :

$$\left( \frac{\bar{K}(b, a)}{\varphi(b)} \right)' = \frac{\bar{K}(b, a)}{\varphi(b)(F(b) - F(0)) - \varphi(a)(F(a) - F(0))} \left( F'(b) + \frac{\varphi'(b)\varphi(a)}{\varphi(b)^2}(F(a) - F(0)) \right).$$

As in the previous examples,  $W_a(\cdot)$  is a straight line passing  $(F(0), 0)$  in the form of  $W_a(y) = \beta(y - F(0))$ . The value function  $v(x)$  in  $x \in (0, b^*)$  is

$$\begin{aligned} \hat{v}(x) &= \varphi(x)W(F(x)) = \beta(F(x) - F(0))\varphi(x) \\ &= \beta^*(\psi(x) - F(0)\varphi(x)) = \beta^*e^{\frac{\delta}{2}\frac{(x-m)^2}{\sigma^2}} \left\{ \mathcal{D}_{-\alpha/\delta} \left( -\left( \frac{x-m}{\sigma} \right) \sqrt{2\delta} \right) \right. \\ &\quad \left. - F(0)\mathcal{D}_{-\alpha/\delta} \left( \left( \frac{x-m}{\sigma} \right) \sqrt{2\delta} \right) \right\}. \end{aligned}$$

Therefore, the solution to the problem is

$$v(x) = \begin{cases} \hat{v}(x), & 0 \leq x \leq b^*, \\ \hat{v}(a^*) + k(x - a^*)^\gamma - K, & b^* \leq x. \end{cases}$$

This solves the problem. See Figure 2(b) for the value function in the case of parameters  $\delta = 0.1$ ,  $m = 0.9$ ,  $\sigma = 0.35$ ,  $\alpha = 0.105$  for the diffusions. As for the reward/cost function parameters,  $k = 0.7$ ,  $K = 0.1$ , and  $\gamma = 0.75$ . The solution is  $(a^*, b^*, \beta) = (0.2192, 0.6220, 0.5749)$ .

*Example 4.3.* We show a simple example where we have multiple continuation regions, the first case of Remark 3.2. Let the uncontrolled process be a standard Brownian motion  $B_t$ , and let  $\alpha = 0$ ,  $f = 0$ , and

$$K(x, y) = -c(\sin x - \sin y) - \delta,$$

with  $c \in \mathbb{R}_+$  and  $\delta \in \mathbb{R}_+$  being some constant parameters. We want to solve

$$v(x) = \sup_{\nu \in \mathcal{S}} \mathbb{E}^x \left[ \sum_{T_i < \tau_0} (\xi_i - \delta) \right].$$

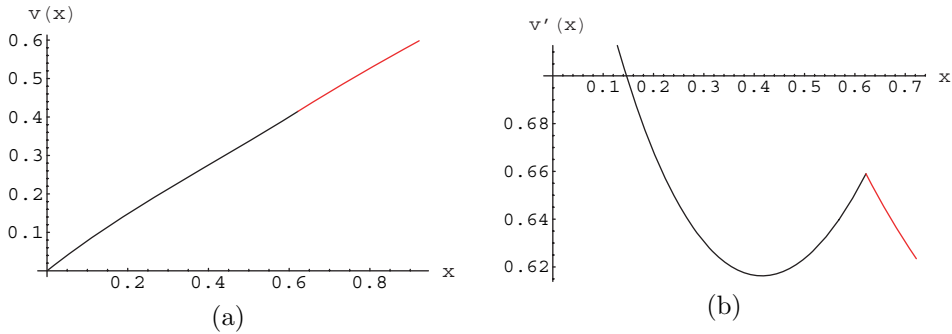


FIG. 2. (a) The value function for the Cadenillas–Sarkar–Zapatero [6] problem and (b) its derivative.

In this case  $F(x) = x$ , and let us define

$$R(x, a) = r(x, a) = \begin{cases} 0, & x = 0, \\ -c(\sin x - \sin a) - \delta, & x > 0. \end{cases}$$

By solving (3.41) with some parameter  $(c, \delta) = (10, 0.35)$ , we find that  $a_k^* = 2.75 + 4k\pi$  and  $b_k^* = 3.52 + 4k\pi$ , with  $k = 0, 1, 2, \dots$ . For all of these pairs,  $\beta^*$  has a common value of 9.30. Hence all of these pairs are optimal. This implies that if the initial state  $x \in (b_k^*, b_{k+1}^*)$ , then we let the process move until it reaches  $b_k^*$  or  $b_{k+1}^*$ . If it reaches  $b_k^*$  first, then an intervention is made to  $a_k^*$ . Now we are in the interval  $(b_{k-1}^*, b_k^*)$ . We continue until the process is absorbed at  $x = 0$ .

**5. Conclusions.** Before we conclude this article, we shall mention an immediate extension to two boundary impulse control problems:

(5.1)

$$J^\nu(x) = \mathbb{E}^x \left[ \int_0^\infty e^{-\alpha s} f(X_s) ds + \sum_{i=1} e^{-\alpha T_i} C_1(X_{T_i-}, X_{T_i}) + \sum_{j=1} e^{-\alpha S_j} C_2(X_{S_j-}, X_{S_j}) \right]$$

and

$$(5.2) \quad v(x) = \sup_{\nu} J^\nu(x) = J^{\nu^*}(x)$$

for all  $x \in \mathbb{R}$ , where

$$\nu = (T_1, T_2, \dots; \zeta_1, \zeta_2, \dots; S_1, S_2, \dots; \eta_1, \eta_2, \dots),$$

with  $\zeta_i > 0$  corresponding to interventions at the upper boundary at intervention time  $T_i$  and  $\eta_j < 0$  at the lower boundary at intervention time  $S_j$ .

Examples of this type include the storage model analyzed by Harrison, Sellke, and Taylor [15] and the foreign exchange rate model studied by Jeanblanc-Picqué [17]. The former problem, for example, is that a controller continuously monitors the inventory so that the inventory level will not fall below the zero level. He is allowed to make interventions by increasing and decreasing the inventory by paying costs associated with interventions. In this case, the process remains within some band(s). In other words, the optimal intervention times are characterized as exit times from an interval in the form of  $(p^*, b^*)$  for  $0 \leq p^* < b^*$ . See Korn [20] for a survey.

Under suitable assumptions, we can develop an argument similar to the previous sections. Among others, the intervention times can be characterized as exit times from an interval  $(p^*, b^*)$ . We can also simplify the performance measure:

$$J^\nu(x) = \mathbb{E}^x[1_{\{T_1 < S_1\}} e^{-\alpha T_1} \{C_1(X_{T_1-}, X_{T_1}) - g(X_{T_1-}) + J^\nu(X_{T_1})\}] \\ + \mathbb{E}^x[1_{\{T_1 > S_1\}} e^{-\alpha S_1} \{C_2(X_{S_1-}, X_{S_1}) - g(X_{S_1-}) + J^\nu(X_{S_1})\}] + g(x),$$

where  $g(x) = \mathbb{E}^x \int_0^\infty e^{-\alpha s} f(X_s^0) ds$  as usual. Again, the last term does not depend on controls, we define  $u(x)$  as  $u(x) = J^\nu(x) - g(x)$ ,

$$(5.3) \quad u(x) = \mathbb{E}^x[1_{\{\tau_b < \tau_p\}} e^{-\alpha \tau_b} u(b)] + \mathbb{E}^x[1_{\{\tau_b > \tau_p\}} e^{-\alpha \tau_p} u(p)], \quad x \in [p, b],$$

where  $T_1 = \tau_b$  and  $S_1 = \tau_p$ , and it follows that

$$(5.4) \quad \frac{u(x)}{\varphi(x)} = \frac{u(b)(F(x) - F(p))}{\varphi(b)(F(b) - F(p))} + \frac{u(p)(F(b) - F(x))}{\varphi(p)(F(b) - F(p))}, \quad x \in [p, b].$$

Hence if we define  $W \triangleq \frac{u}{\varphi} \circ F^{-1}$ , we have linear characterization again in the transformed space:

$$(5.5) \quad W(F(x)) = W(F(b)) \frac{F(x) - F(p)}{F(b) - F(p)} + W(F(p)) \frac{F(b) - F(x)}{F(b) - F(p)}, \quad x \in [p, b],$$

and the solution to the problem is described as

$$u(x) = \begin{cases} \bar{C}_2(x, q) + u_0(q), & x \leq p, \\ u_0(x) \triangleq \mathbb{E}^x[1_{\{\tau_b < \tau_p\}} e^{-\alpha \tau_b} u(b)] + \mathbb{E}^x[1_{\{\tau_b > \tau_p\}} e^{-\alpha \tau_p} u(p)], & p \leq x \leq b, \\ \bar{C}_1(x, a) + u_0(a), & b \leq x, \end{cases}$$

where  $\bar{C}_i(x, y) = C_i(x, y) - g(x) + g(y)$  for  $i = 1$  and  $2$ .

We have studied impulse control problems. The intervention times are characterized as exit times of the process from a finite union of disjoint intervals on the real line. A sufficient condition is given for the connectedness of the continuation region. The value function is shown to be linear in the continuation region of the transformed space, and a direct calculation method is described. This method can handle impulse control problems with nonsmooth reward and cost functions. The finiteness of the value function is shown to be equivalent to the existence of a concave majorant of the suitable transformed reward function. The latter is easier to check by using elementary geometric arguments.

The new characterization of the value function and optimal strategies can be extended to other optimization problems, such as optimal switching and combined problems of optimal stopping and impulse control. If an optimal strategy exists in the class of exit times, then the problem can be reduced to a sequence of optimal stopping problems, and an effective characterization of the value function is possible.

## Appendix A.

**A.1. Proof of Lemma 3.2.** To make the proof more intuitive, we will work with (3.14) rather than with (3.15) where the integration part is converted to  $g$  functions. For this purpose, it is convenient to define the following two operators  $\mathcal{M}_o : \mathcal{H} \rightarrow \mathcal{H}$  and  $\mathcal{L}_o : \mathcal{H} \rightarrow \mathcal{H}$ :

$$(A.1) \quad \mathcal{M}_o u(x) = \sup_{y \in \mathbb{R}} [K(x, y) + u(y)]$$

and

$$(A.2) \quad \mathcal{L}_o u(x) = \sup_{\tau \in \mathcal{S}} \mathbb{E}^x \left[ \int_0^\tau e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau} \mathcal{M}_o u(X_{\tau-}) \right].$$

Hence we can proceed with the arguments developed by Davis [12]. In terms of the two operators just defined, (3.14) becomes

$$(A.3) \quad w_{n+1}(x) = \sup_{\tau \in \mathcal{S}} \mathbb{E}^x \left[ \int_0^\tau e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau} \mathcal{M}_o w_n(X_{\tau-}) \right]$$

$$(A.4) \quad = \mathcal{L}_o w_n(x).$$

(1)  $w_n = v_n$  for all  $n$ : Note that  $F$ -concavity of the function  $\bar{K}/\varphi$  guarantees that the value  $u(y)$  in (3.12) is a finite number as is shown in Proposition 3.4. Hence  $f$  and  $\mathcal{M}w_n$  satisfy the condition (polynomial growth) of Theorem 7.2 in Øksendal and Sulem [25], which shows that  $w_n = v_n$ .

(2)  $v(x) = \lim_{n \rightarrow \infty} w_n(x)$ : Since  $w_n$  is monotone increasing, the limit  $w(x) = \lim_{n \rightarrow \infty} w_n(x)$  exists. Since  $S_n \subset S$ ,  $w_n(x) \leq v(x)$ . Hence  $w(x) \leq v(x)$ . To show the reverse inequality, we define  $S^*$  be a set of interventions such that

$$S^* = \{\nu \in \mathcal{S} : J^\nu(x) < \infty \text{ for all } x \in \mathbb{R}\}.$$

Let us assume that  $\underline{v(x)} < +\infty$  and consider strategy  $\nu^* \in S^*$  such that

$$(A.5) \quad J^{\nu^*}(x) \geq v(x) - \epsilon$$

for some  $\epsilon > 0$  and another strategy  $\nu_n$  that coincides with  $\nu^*$  up to and including time  $T_n$  and then takes no further interventions.

$$J^{\nu^*}(x) - J^{\nu_n}(x) = \mathbb{E}^x \left[ \int_{T_n}^\infty e^{-\alpha s} (f(X_s) - f(X_s^0)) ds + \sum_{i \geq n+1} e^{-\alpha T_i} K(X_{T_i-}, X_{T_i}) \right],$$

which implies that

$$|J^{\nu^*}(x) - J^{\nu_n}(x)| \leq \mathbb{E}^x \left[ \int_{T_n}^\infty e^{-\alpha s} (|f(X_s)| + |f(X_s^0)|) ds + \sum_{i \geq n+1} e^{-\alpha T_i} |K(X_{T_i-}, X_{T_i})| \right].$$

As  $n \rightarrow +\infty$ ,  $T_n \rightarrow +\infty$ , and the first term of the right-hand side can be arbitrarily small due to (3.6); so is the second term by the finiteness of  $v(x)$  with (3.7). Hence it is shown that  $|J^{\nu^*}(x) - J^{\nu_n}(x)| < \epsilon$  for  $n$  large enough, which implies with (A.5) that

$$\liminf_{n \rightarrow +\infty} J^{\nu_n}(x) \geq J^{\nu^*}(x) - \epsilon \geq v(x) - 2\epsilon$$

so that  $v(x) \leq \lim_{n \rightarrow +\infty} v_n(x)$  since  $\epsilon$  is arbitrary. Now we have established that  $v(x) = \lim_{n \rightarrow \infty} w_n(x)$  when  $v(x) < +\infty$ . Next, consider the case of  $v(x) = +\infty$ . Then by the recursive method described in section 3.1, we see that  $v_1(\overline{x}) = \overline{w_1(x)} = \infty$ . By the first statement of this lemma, we can conclude that  $v_n(x) = w_n(x) = \infty$  for all  $n \in \mathbb{N}$ , obtaining  $v(x) = \lim_{n \rightarrow \infty} w_n(x)$ . This completes the proof of the second statement.

(3)  $w = \mathcal{L}_o w$  : Since  $w_n \uparrow w$ , we have the following chain of equalities:

$$\begin{aligned}\mathcal{M}_o w(x) &= \sup_{y \in \mathbb{R}} [K(x, y) + w(y)] = \sup_{y \in \mathbb{R}} \sup_{n \in \mathbb{N}} [K(x, y) + w_n(y)] \\ &= \sup_{n \in \mathbb{N}} \sup_{y \in \mathbb{R}} [K(x, y) + w_n(y)] = \sup_{n \in \mathbb{N}} \mathcal{M}_o w_n(x).\end{aligned}$$

In view of this, if we take the limit on both sides of (A.3) as  $n \rightarrow \infty$ , by the monotone convergence theorem,

$$w(x) = \sup_{\tau \in \mathcal{S}} \mathbb{E}^x \left[ \int_0^\tau e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau} \mathcal{M}_o w(X_{\tau-}) \right].$$

This shows that  $w = \mathcal{L}_o w$ . Suppose that  $w'(x)$  satisfies  $w' = \mathcal{L}_o w'$  and majorizes  $g(x) = v_0(x)$ . Then  $w' = \mathcal{L}_o w' \geq \mathcal{L}_o v_0 = w_1$ . If we assume that  $w' \geq v_n$ , then

$$w' = \mathcal{L}_o w' \geq \mathcal{L}_o v_n = v_{n+1} = w_{n+1}.$$

This shows that, by the induction argument, we have  $w' \geq w_n$  for all  $n$ , leading to  $w' \geq \lim_{n \rightarrow \infty} w_n = w$ . Thus it shows that  $w$  is the smallest solution majorizing  $g$  of the functional equation  $w - g = \mathcal{L}(w - g)$ . This completes the third statement of the lemma.

**A.2. Derivation of (3.46).** The first order condition of the optimality with respect to  $a$  is

$$(A.6) \quad d\beta(a)/da = d\beta/db \times db/da = 0.$$

For a fixed  $a$ , from (3.41) by viewing  $\beta$  as a function of state  $x = F^{-1}(y)$ , that is,  $\beta = \beta(F^{-1}(y))$ , it is clear that  $d\beta(F^{-1}(y))/dy = \frac{1}{F'(x)} d\beta/dx$ . Since  $F'(x) > 0$  for all  $x \in \mathcal{I}$ , we have  $d\beta(F^{-1}(y))/dy = 0$  if and only if  $d\beta/dx = 0$ . But  $d\beta(F^{-1}(y))/dy = 0$  at  $x = F^{-1}(y) = b$  implies the following: At  $x = b(a)$ , where the shifted function  $\Phi(y, a)$  becomes tangent to the linear function  $W_a(F(x)) = \beta(a)F(x) + \delta$ , the second derivative of the shifted function vanishes. But it is clearly impossible if  $\Phi(y, a)$  is strictly concave at  $F(b)$ . Hence in view of (A.6), we can claim that  $d\beta(a)/da = 0$  only if  $db/da = 0$ , provided that  $\Phi(y, a)$  is strictly concave at  $y = F(b)$ . Now we differentiate (3.42) with respect to  $a$  with noting  $db/da = 0$ , and we obtain (3.46). In particular, for the case of absorbing boundary, (3.46) is of the form:

$$\begin{aligned}& \left( \frac{\partial \bar{K}(b, a)}{\partial a} + D\varphi'(a) \right) ((F(b) - F(a))\varphi(b) - (F(a) - F(0))\varphi(a)) \\ &= -(\bar{K}(b, a) - D(\varphi(b) - \varphi(a))) (F'(a)\varphi(a) + (F(a) - F(0))\varphi'(a)),\end{aligned}$$

which is to be solved simultaneously with (3.44) and (3.45).

**Acknowledgments.** The author thanks Savas Dayanik for valuable comments. He is also grateful to Erhan Bayraktar and the participants at the INFORMS 2004 Annual Meeting in Denver, CO, at the Civitas Foundation Finance Seminar in Princeton, NJ, and at various university seminars. The author also thanks the two anonymous referees for their comments that improved the manuscript.

## REFERENCES

- [1] L. H. R. ALVAREZ, *A class of solvable impulse control problems*, Appl. Math. Optim., 49 (2004), pp. 265–295.
- [2] L. H. R. ALVAREZ AND J. A. VIRTANEN, *A class of solvable stochastic dividend optimization problems: On the general impact of flexibility on valuation*, Econom. Theory, 28 (2006), pp. 373–398.
- [3] L. H. R. ALVAREZ AND J. LEMPA, *On the optimal stochastic impulse control of linear diffusions*, SIAM J. Control Optim., 47 (2008), pp. 703–732.
- [4] A. BENSOUSSAN AND J. L. LIONS, *Impulse Control and Quasi-Variational Inequalities*, Gauthier-Villars, Paris, 1984.
- [5] A. N. BORODIN AND P. SALMINEN, *Handbook of Brownian Motion - Facts and Formulae*, 2nd ed., Birkhäuser, Basel, 2002.
- [6] A. CADENILLAS, S. SARKAR, AND F. ZAPATERO, *Optimal dividend policy with mean-reverting cash reservoir*, Math. Finance, 17 (2007), pp. 81–109.
- [7] A. CADENILLAS AND F. ZAPATERO, *Classical and impulse stochastic control of the exchange rate using interest rates and reserves*, Math. Finance, 10 (2000), pp. 141–156.
- [8] R. CARMONA AND S. DAYANIK, *Optimal multiple-stopping of linear diffusions*, Math. Oper. Res., to appear.
- [9] R. CARMONA AND N. TOUZI, *Optimal multiple stopping and valuations of swing options*, Math. Finance, to appear.
- [10] J. PH. CHANCELIER, B. ØKSENDAL, AND A. SULEM, *Combined stochastic control and optimal stopping, and application to numerical approximation of combined stochastic and impulse control*, Proc. Steklov Inst. Math., 237 (2002), pp. 140–163.
- [11] M. DAHLGREN AND R. KORN, *The swing option on the stock market*, Int. J. Appl. Theor. Finance, 8 (2005), pp. 123–139.
- [12] M. H. A. DAVIS, *Markov Models and Optimization*, Chapman and Hall, London, 1992.
- [13] S. DAYANIK AND I. KARATZAS, *On the optimal stopping problem for one-dimensional diffusions*, Stochastic Process. Appl., 107 (2003), pp. 173–212.
- [14] E. DYKIN, *Markov Processes*, Vol. II, Springer-Verlag, Berlin, 1965.
- [15] J. M. HARRISON, T. M. SELKE, AND A. J. TAYLOR, *Impulse control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 454–466.
- [16] K. ITÔ AND H. P. MCKEAN, *Diffusions Processes and Their Sample Paths*, Springer-Verlag, Berlin, 1974.
- [17] M. JEANBLANC-PICQUÉ, *Impulse control method and exchange rate*, Math. Finance, 3 (1993), pp. 161–177.
- [18] M. JEANBLANC-PICQUÉ AND A. N. SHIRYAEV, *Optimization of the flow of dividends*, Russian Math. Surveys, 50 (1995), pp. 257–277.
- [19] R. KORN, *Portfolio optimisation with strictly positive transaction costs*, Finance Stoch., 2 (1998), pp. 84–114.
- [20] R. KORN, *Some applications of impulse control in mathematical finance*, Math. Methods Oper. Res., 50 (1999), pp. 493–528.
- [21] N. N. LEBEDEV, *Special Functions and Their Applications*, revised ed., R. A. Silverman, ed., Dover, New York, 1972 (in English).
- [22] A. J. MORTON AND S. R. PLISKA, *Optimal portfolio management with fixed transaction costs*, Math. Finance, 5 (1995), pp. 337–356.
- [23] G. MUNDACA AND B. ØKSENDAL, *Optimal stochastic intervention control with application to the exchange rate*, J. Math. Econom., 29 (1998), pp. 223–241.
- [24] B. ØKSENDAL, *Stochastic control problems where small intervention costs have big effects*, Appl. Math. Optim., 40 (1999), pp. 355–375.
- [25] B. ØKSENDAL AND A. SULEM, *Applied Stochastic Control of Jump Diffusions*, Springer-Verlag, Berlin, 2005.

## LOCAL CONTROLLABILITY OF A ONE-DIMENSIONAL BEAM EQUATION\*

KARINE BEAUCHARD†

**Abstract.** We prove that the beam equation with clamped ends is locally controllable in a  $H^{5+\epsilon} \times H^{3+\epsilon}((0, 1), \mathbb{R})$ -neighborhood of a particular trajectory of the free system, with  $\epsilon > 0$  and with control functions in  $H_0^1((0, T), \mathbb{R})$ . Ball, Marsden, and Slemrod already proved that this equation is not controllable in  $H_0^2 \times L^2((0, 1), \mathbb{R})$  with control functions in  $L_{\text{loc}}^r(\mathbb{R}, \mathbb{R})$ ,  $r > 1$ . This article justifies that their negative result is due to a choice of functional spaces which does not allow controllability. Our proof uses moment theory and the Nash–Moser theorem.

**Key words.** control of partial differential equations, bilinear control problem, beam equation, Nash–Moser theorem, trigonometric moment problem

**AMS subject classifications.** 35B37, 93C10, 93C20, 35Q72, 42A70

**DOI.** 10.1137/050642034

### 1. Introduction.

**1.1. Main result.** We consider the beam equation with clamped ends

$$(\Sigma) \begin{cases} u_{tt} + u_{xxxx} + p(t)u_{xx} = 0, (t, x) \in \mathbb{R}_+ \times (0, 1), \\ u = u_x = 0 \text{ at } x = 0, 1. \end{cases}$$

It is a nonlinear control system where

- the state is the couple  $(u, u_t)$  and
- the control is the real-valued function  $t \mapsto p(t)$ .

We introduce the operator  $A$  defined by

$$(1.1) \quad D(A) := H^4 \cap H_0^2((0, 1), \mathbb{R}), \quad Av := \frac{d^4 v}{dx^4}.$$

Let  $(\lambda_n)_{n \in \mathbb{N}^*} \subset \mathbb{R}_+^*$  be the increasing sequence of eigenvalues of  $A$  and  $(\varphi_n)_{n \in \mathbb{N}^*}$  associated orthonormalized eigenvectors. Then, for every  $n \in \mathbb{N}^*$ , the functions

$$\varphi_n(x) \cos(\sqrt{\lambda_n} t) \text{ and } \varphi_n(x) \sin(\sqrt{\lambda_n} t)$$

are solutions of  $(\Sigma)$ , with  $p \equiv 0$ . For  $s > 0$ , we introduce the space

$$(1.2) \quad H_{(0)}^s((0, 1), \mathbb{R}) := D(A^{s/4}),$$

equipped with the norm

$$(1.3) \quad \|\varphi\|_{H_{(0)}^s((0, 1), \mathbb{C})} := \left( \sum_{k=1}^{\infty} |\lambda_k^{s/4} \langle \varphi, \varphi_k \rangle|^2 \right)^{1/2}.$$

\*Received by the editors October 6, 2005; accepted for publication (in revised form) October 2, 2007; published electronically March 21, 2008.

<http://www.siam.org/journals/sicon/47-3/64203.html>

†CMLA, ENS Cachan, CNRS, Universud, 61 avenue du président Wilson, F-94230 Cachan, France (Karine.Beauchard@cmla.ens-cachan.fr).



In particular, we have

$$(1.4) \quad \begin{aligned} H_{(0)}^s((0, 1), \mathbb{R}) &= \{\varphi \in H^s((0, 1), \mathbb{R}); \varphi = \varphi' = 0 \text{ at } x = 0, 1\} \text{ for } s \in \{2, 3, 4\}, \\ H_{(0)}^5((0, 1), \mathbb{R}) &= \{\varphi \in H^5((0, 1), \mathbb{R}); \varphi = \varphi' = \varphi^{(4)} = 0 \text{ at } x = 0, 1\}, \\ H_{(0)}^s((0, 1), \mathbb{R}) &= \{\varphi \in H^s((0, 1), \mathbb{R}); \varphi = \varphi' = \varphi^{(4)} = \varphi^{(5)} = 0 \text{ at } x = 0, 1\} \\ &\quad \text{for } s = 6, 7, 8. \end{aligned}$$

The main result of this article is the following one.

**THEOREM 1.** *Let  $T := 8/\pi$ ,  $\epsilon > 0$ , and*

$$(1.5) \quad u^{ref}(t, x) := \varphi_2(x) \sin(\sqrt{\lambda_2}t) + \varphi_3(x) \sin(\sqrt{\lambda_3}t).$$

*There exists a neighborhood  $V_0$  of  $(u^{ref}(0), \dot{u}^{ref}(0))$  and a neighborhood  $V_T$  of  $(u^{ref}(T), \dot{u}^{ref}(T))$  in  $H_{(0)}^{5+\epsilon} \times H_{(0)}^{3+\epsilon}((0, 1), \mathbb{R})$  such that, for every  $(u_0, \dot{u}_0) \in V_0$ , for every  $(u_T, \dot{u}_T) \in V_T$ , there exists  $p \in H_0^1((0, T), \mathbb{R})$  such that the solution of  $(\Sigma)$  with  $(u(0), \dot{u}(0)) = (u_0, \dot{u}_0)$  and control  $p$  satisfies  $(u(T), \dot{u}(T)) = (u_T, \dot{u}_T)$ .*

**1.2. A previous noncontrollability result.** In [1], Ball, Marsden, and Slemrod discuss the controllability of infinite dimensional bilinear control systems of the form

$$(1.6) \quad \dot{w}(t) = \mathcal{A}w(t) + p(t)\mathcal{B}w(t).$$

Thanks to the Baire lemma, they prove the following noncontrollability result.

**THEOREM 2.** *Let  $X$  be a Banach space with  $\dim(X) = +\infty$ . Let  $\mathcal{A}$  generate a  $C^0$ -semigroup of bounded linear operators on  $X$  and  $\mathcal{B} : X \rightarrow X$  be a bounded linear operator. Let  $w_0 \in X$  be fixed, and let  $w(t; p, w_0)$  denote the unique solution of (1.6) for  $p \in L_{loc}^1([0, +\infty), \mathbb{R})$ . The set of states accessible from  $w_0$  defined by*

$$S(w_0) := \{w(t; p, w_0); t \geq 0, p \in L_{loc}^r([0, \infty), \mathbb{R}), r > 1\}$$

*is contained in a countable union of compact subsets of  $X$  and, in particular, has a dense complement.*

We can write  $(\Sigma)$  in the first order form (1.6), with

$$(1.7) \quad w := \begin{pmatrix} u \\ u_t \end{pmatrix}, \quad \mathcal{A} := \begin{pmatrix} 0 & I \\ -A & 0 \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} 0 & 0 \\ -\frac{d^2}{dx^2} & 0 \end{pmatrix}.$$

Let  $X := H_0^2((0, 1), \mathbb{R}) \times L^2((0, 1), \mathbb{R})$ , with inner product

$$\langle (u_1, u_2), (v_1, v_2) \rangle_X := \int_0^1 \left( A^{1/2} u_1 A^{1/2} v_1 + u_2 v_2 \right) dx.$$

As noticed in [1], Theorem 2 shows that, for every  $(u_0, \dot{u}_0) \in X$ , the set of  $(u, u_t)$  in  $X$  accessible from  $(u_0, \dot{u}_0)$  with controls in  $L_{loc}^r([0, \infty), \mathbb{R})$ ,  $r > 1$ , has a dense complement in  $X$ . In particular,  $(\Sigma)$  is not controllable in  $H_0^2((0, 1), \mathbb{R}) \times L^2((0, 1), \mathbb{R})$ , with control functions  $p$  in  $L_{loc}^2([0, +\infty), \mathbb{R})$ .

However, this theorem does not give any obstruction for having controllability in other spaces. For example, Theorem 2 does not apply with

$$\tilde{X} := H_{(0)}^s \times H_{(0)}^{s-2}((0, 1), \mathbb{R}) \text{ for } s \in \mathbb{N}^*, s \geq 3$$

instead of  $X$ . Indeed,

- $\mathcal{A}$  generates a  $C^0$ -semigroup of bounded linear operators of  $\tilde{X}$  (we refer to Remark 1 in section 2 for a justification of this statement),
- for  $w = (w_1, w_2)^t \in \tilde{X}$ ,  $\mathcal{B}w = (0, -w_2'')^t$  belongs to  $H_{(0)}^s \times H^{s-2}((0, 1), \mathbb{R})$ ,
- but in general  $w_2''$  does not vanish at  $x = 0, 1$ , so  $\mathcal{B}w$  does not belong to  $\tilde{X}$  (see (1.4)), and thus  $\mathcal{B}$  does not map  $\tilde{X}$  into  $\tilde{X}$ .

Notice that Theorem 2 does not apply either with

$$\overline{X} := H_{(0)}^3 \times H^1((0, 1), \mathbb{R})$$

(which is a space such that  $\mathcal{B}$  maps  $\overline{X}$  into  $\overline{X}$ ) instead of  $X$ , because  $\mathcal{A}$  does not generate a  $C^0$  semigroup of bounded operators of  $\overline{X}$  (we refer to Proposition 8 for a proof of this statement).

In this article, we prove a local controllability result in  $H_{(0)}^{5+\epsilon} \times H_{(0)}^{3+\epsilon}((0, 1), \mathbb{R})$ , with  $\epsilon > 0$  and with control functions  $p$  in  $H_{loc}^1(\mathbb{R}_+, \mathbb{R})$ . Thus, the noncontrollability result proved by Ball, Marsden, and Slemrod relies on the fact that the choice of functional spaces does not allow controllability. In order to state affirmative controllability results, one must

- either control  $(u, u_t)$  in  $H_0^2 \times L^2((0, 1), \mathbb{R})$  but with a control functions set larger than  $\cup_{r>0} L_{loc}^r(\mathbb{R}_+, \mathbb{R})$ , for example,  $H_{loc}^{-1}(\mathbb{R}_+, \mathbb{R})$ ,
- or control  $(u, u_t)$  using the control functions set  $L_{loc}^2(\mathbb{R}_+, \mathbb{R})$  but in a smaller space than  $H_0^2 \times L^2((0, 1), \mathbb{R})$ , for example,  $H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R})$ .

**1.3. Sketch of the proof.** The result of Theorem 1 is a local controllability result, in time  $T = 8/\pi$ , around the trajectory  $(u^{ref}, \dot{u}^{ref}, p \equiv 0)$  for the nonlinear control system  $(\Sigma)$ . It is equivalent to a local surjectivity property for the map

$$(1.8) \quad \Phi_T : (u_0, \dot{u}_0, p) \mapsto (u_0, \dot{u}_0, u(T), u_t(T)),$$

where  $T = 8/\pi$  and  $u$  is a solution of  $(\Sigma)$  with initial condition  $u(0) = u_0$ ,  $u_t(0) = \dot{u}_0$ . For the proof of such results, there exists a classical approach which consists of

- first proving the global controllability of the linearized system around the trajectory considered (which corresponds to the surjectivity of  $d\Phi_T(u^{ref}(0), u_t^{ref}(0), 0)$ ) and
- then applying the inverse mapping theorem to the nonlinear map  $\Phi_T$ , which gives a local surjectivity property for  $\Phi_T$  around  $(u_0^{ref}, \dot{u}_0^{ref}, 0)$ .

The general strategy adopted in this article is this one, but we will see that the classical inverse mapping theorem is not sufficient to conclude. Thus, we use a more elaborate version of the inverse mapping theorem, namely, the Nash–Moser implicit functions theorem. This strategy has already been used for Schrödinger equations in [3], [4], [2].

First, we consider the linearized system around the trajectory  $(u^{ref}, \dot{u}^{ref}, p \equiv 0)$  defined by (1.5), which is

$$(\Sigma_l^{ref}) \begin{cases} U_{tt} + U_{xxxx} + P(t)u_{xx}^{ref} = 0, (t, x) \in (0, \infty) \times \mathbb{R}, \\ U = U_x = 0 \text{ at } x = 0, 1. \end{cases}$$

It is a linear control system where

- the state is the couple  $(U, U_t)$  and
- the control is the real-valued function  $t \mapsto P(t)$ .

We prove the following controllability result for this system.

THEOREM 3. (1) For every  $T > \pi/\sqrt{\lambda_2}$ , for every  $(U_T, \dot{U}_T) \in H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R})$ , there exists  $P \in L^2((0, T), \mathbb{R})$  such that the solution of  $(\Sigma_l^{ref})$  with  $(U(0), \dot{U}(0)) = (0, 0)$  and control  $P$  satisfies  $(U(T), \dot{U}(T)) = (U_T, \dot{U}_T)$ .

(2) For every  $T > 0$ , there exists  $(U_T, \dot{U}_T) \in H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R})$ , (resp.,  $(U_T, \dot{U}_T) \in H_0^2 \times L^2((0, 1), \mathbb{R})$ ) such that, for every  $P \in H^1((0, T), \mathbb{R})$  (resp.,  $P \in L^2((0, T), \mathbb{R})$ ), the solution of  $(\Sigma_l^{ref})$  with  $(U(0), \dot{U}(0)) = (0, 0)$  and control  $P$  satisfies  $(U(T), \dot{U}(T)) \neq (U_T, \dot{U}_T)$ .

Therefore, the linear system  $(\Sigma_l^{ref})$  is controllable in  $H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R})$  with control functions in  $L^2((0, T), \mathbb{R})$ , but it is neither controllable in  $H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R})$  with control functions in  $H^1((0, T), \mathbb{R})$  nor controllable in  $H_0^2 \times L^2((0, 1), \mathbb{R})$  with control functions in  $L^2((0, T), \mathbb{R})$ . As we will see precisely in section 4, statement (2) of Theorem 3 shows that the local controllability around  $(u^{ref}, \dot{u}^{ref}, 0)$  cannot be obtained by applying the classical inverse mapping theorem. Roughly speaking, the controls built for the control of the linearized system are not smooth enough for the application of the inverse mapping theorem.

Thus, in order to prove Theorem 1, we use a Nash–Moser theorem. This theorem is an elaborate version of the inverse mapping theorem. It gives the local surjectivity of a nonlinear map  $\Theta$  around a point  $x^{ref}$  thanks to essentially 3 assumptions:

- (1) There exist decreasing sequences of spaces  $(E_a)_{a \geq 0}$ ,  $(F_b)_{b \geq 0}$  such that
  - $\Theta : E_a \rightarrow F_a$  is  $C^2$  for every  $a$ ,
  - there exist smoothing linear operators  $(S_\theta)_{\theta > 0}$  (resp.,  $(\tilde{S}_\theta)_{\theta > 0}$ ) defined on the spaces  $E_a$  (resp.,  $F_b$ ) with  $S_\theta \rightarrow Id$  (resp.,  $\tilde{S}_\theta \rightarrow Id$ ) when  $\theta \rightarrow +\infty$ ,
  - the norms of the operators  $Id - S_\theta, S_\theta : E_a \rightarrow E_A$  for  $a \neq A$  (resp.,  $Id - \tilde{S}_\theta, \tilde{S}_\theta : F_b \rightarrow F_B$  for  $b \neq B$ ) are bounded by an explicit expression in terms of  $\theta, a, A$  (resp.,  $\theta, b, B$ )
- (2) one knows a particular explicit bound on the second differential  $d^2\Theta(x)$  for  $x$  in a neighborhood of  $x^{ref}$ ;
- (3) for every  $x$  in a neighborhood of  $x^{ref}$ , the differential  $d\Theta(x)$  has a right inverse  $d\Theta(x)^{-1} : F_{b_1} \rightarrow E_{a_1}$ , where  $a_1 < b_1$ , that satisfies particular explicit estimates, called “tame estimates.”

We refer to section 5 to see more precisely what the explicit expressions mentioned above look like (the bounds on the smoothing operators mentioned in (1) are given in (5.1), (5.2), (5.3), (5.4); the bound of the second differential mentioned in (2) is given in (5.6); the tame estimates mentioned in (3) are given in (5.7), (5.8)).

The main differences between the inverse mapping theorem and the Nash–Moser theorem are the following:

- The Nash–Moser theorem needs a weaker surjectivity property on  $d\Theta(x^{ref})$ . Indeed, the inverse mapping theorem needs the surjectivity of the map  $d\Theta(x^{ref}) : E_{a_1} \rightarrow F_{a_1}$  (with the same index  $a_1$  in both sides; i.e., the nonlinear map  $\Theta$  has to be  $C^1$  between  $E_{a_1}$  and  $F_{a_1}$ ), whereas the Nash–Moser theorem needs the existence of a right inverse  $d\Theta(x^{ref})^{-1}$  defined on a space  $F_{b_1}$  strictly included in  $F_{a_1}$ , with values in  $E_{a_1}$ . Thus  $d\Theta(x^{ref})$  does not need to be surjective from  $E_{a_1}$  to  $F_{a_1}$ .
- The Nash–Moser theorem needs a surjectivity property for all of the differentials  $d\Theta(x)$  for  $x$  in a neighborhood of  $x^{ref}$ , whereas the inverse mapping theorem requires the surjectivity of the differential  $d\Theta(x)$  only at the point  $x = x^{ref}$ . This assumption is often the most difficult to check in the applications of this theorem.

The rest of the paper is organized as follows.

In section 2, we give the definition of solutions for the nonlinear system  $(\Sigma)$ , and we recall classical results about the existence, the uniqueness, and the regularity of these solutions. We also prove bounds on those solutions that will be used many times in this article. Finally, we define spaces  $E_a$  and  $F_b$  such that the map  $\Phi_T$  defined by (1.8) is a  $C^1$  map from  $E_a$  to  $F_a$  for every  $a$ .

Section 3 is devoted to the proof of Theorem 3. In subsection 3.1, we state and prove some preliminary results about the eigenvalues and eigenvectors of the operator  $A$  defined by (1.1). In subsection 3.2, we prove Theorem 3.

In section 4, we explain in detail why Theorem 1 cannot be deduced from the first statement of Theorem 3 by applying the classical inverse mapping theorem to the map  $\Phi_T$ .

In section 5, we state and prove a Nash–Moser implicit function theorem inspired by [20]. The following sections are dedicated to the verification of each assumption of this theorem, i.e., points (1), (2), and (3) presented above.

In section 6, we present a construction of smoothing operators  $S_\theta$ , on the spaces  $E_a$  and  $F_b$  (defined in section 2), with the desired explicit bounds in terms of  $\theta$ ,  $a$ , and  $A$  for the linear map  $S_\theta : E_a \rightarrow E_A$ ,  $a < A$  (i.e., point (1)).

In section 7, we prove a bound on the second differential  $d^2\Phi_T$  (i.e., point (2)).

In section 8, we prove the existence of a right inverse for  $d\Phi_T(u_0^{ref}, \dot{u}_0^{ref}, 0)$  with tame estimates (i.e., the case  $x = x^{ref}$  in point (3)).

In section 9, we prove the existence of a right inverse for  $d\Phi_T(u_0, \dot{u}_0, p)$  with tame estimates, for every  $(u_0, \dot{u}_0, p)$  in a small neighborhood of  $(u_0^{ref}, \dot{u}_0^{ref}, 0)$  (i.e., point (3)). This part of the proof is the most technical one.

Finally, in section 10, we give some remarks, conjectures, and prospects.

In this work, we use the same letter  $C$  to denote different constants. The value of  $C$  can change from one expression to another.

**2. Regularity and bound for the solutions of the nonlinear system.** This section is dedicated to the statement of existence, uniqueness, regularity results, and bounds for the solutions of the Cauchy problem

$$(CY) : \begin{cases} u_{tt} + u_{xxxx} + p(t)u_{xx} + f(t) = 0, & x \in (0, 1), t \in \mathbb{R}_+, \\ u = u_x = 0 \text{ at } x = 0, 1, \\ u(0, x) = u_0(x), \dot{u}(0, x) = \dot{u}_0(x). \end{cases}$$

These bounds, presented in subsection 2.1, will be used many times in this article. Then, in subsection 2.2, we deduce the spaces  $E_a$  and  $F_b$  between which the map  $\Phi_T$  defined by (1.8) is of class  $C^1$ .

All of these results are classical, but we give proofs for the sake of completeness. The reading of the proofs in this section is not necessary for the understanding of the next sections of this article.

**2.1. Existence, uniqueness, regularity, and bounds.** We introduce the first order Cauchy problem

$$(CY) : \begin{cases} \frac{dw}{dt} = -\mathcal{A}w - p(t)\mathcal{B}w + F(t), \\ w(0) = w_0, \end{cases}$$

where  $\mathcal{A}$  and  $\mathcal{B}$  are linear operators with domains

$$D(\mathcal{A}) := H_{(0)}^4 \times H_0^2((0, 1), \mathbb{R}), \quad D(\mathcal{B}) := H_0^2 \times L^2((0, 1), \mathbb{R})$$

defined by

$$\mathcal{A} \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} := \begin{pmatrix} -w^2 \\ w^1_{xxxx} \end{pmatrix}, \quad \mathcal{B} \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} := \begin{pmatrix} 0 \\ w^1_{xx} \end{pmatrix},$$

and  $F : (0, T) \rightarrow H_0^2 \times L^2((0, 1), \mathbb{R})$ .

The Cauchy problem (CY) is equivalent to the Cauchy problem (CY) with  $w = (u, u_t)$ ,  $w_0 = (u_0, \dot{u}_0)$ , and  $F = (0, -f)$ . Thus, in this section, we work only on (CY).

The operator  $\mathcal{A}$  generates a  $C^0$ -group of isometries of  $H_0^2 \times L^2((0, 1), \mathbb{R})$  with the following explicit expression

$$(2.1) \quad e^{-t\mathcal{A}} \begin{pmatrix} u_0 \\ \dot{u}_0 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{\infty} \left( \langle u_0, \varphi_k \rangle \cos(\sqrt{\lambda_k}t) + \frac{1}{\sqrt{\lambda_k}} \langle \dot{u}_0, \varphi_k \rangle \sin(\sqrt{\lambda_k}t) \right) \varphi_k \\ \sum_{k=1}^{\infty} \left( -\sqrt{\lambda_k} \langle u_0, \varphi_k \rangle \sin(\sqrt{\lambda_k}t) + \langle \dot{u}_0, \varphi_k \rangle \cos(\sqrt{\lambda_k}t) \right) \varphi_k \end{pmatrix}$$

for every  $(u_0, \dot{u}_0) \in H_0^2 \times L^2((0, 1), \mathbb{R})$ , where  $\langle \cdot, \cdot \rangle$  denotes the usual  $L^2((0, 1), \mathbb{R})$ -scalar product.

*Remark 1.* With this explicit expression of the  $C^0$ -group generated by  $\mathcal{A}$ , it is clear that  $e^{-\mathcal{A}t}$  is an isometry of  $H_{(0)}^{s+2} \times H_{(0)}^s((0, 1), \mathbb{R})$  for every  $s \in \mathbb{N}^*$  (with the definition (1.2) and (1.3)) as claimed in section 1.2.

**PROPOSITION 1.** *Let  $T > 0$ ,  $p \in L^1((0, T), \mathbb{R})$ ,  $w_0 \in H_0^2 \times L^2((0, 1), \mathbb{R})$ , and  $F \in L^1((0, T), H_0^2 \times L^2((0, 1), \mathbb{R}))$ . There exists a unique weak solution of (CY), i.e., a function  $w \in C^0([0, T], H_0^2 \times L^2((0, 1), \mathbb{R}))$ , such that the following equality holds in  $H_0^2 \times L^2((0, 1), \mathbb{R})$  for every  $t \in [0, T]$ :*

$$(2.2) \quad w(t) = e^{-t\mathcal{A}}w_0 + \int_0^t e^{-(t-s)\mathcal{A}}[-p(s)\mathcal{B}w(s) + F(s)]ds.$$

Moreover, it satisfies

$$(2.3) \quad \|w\|_{C^0([0, T], H_0^2 \times L^2)} \leq \left( \|w_0\|_{H_0^2 \times L^2} + \|F\|_{L^1((0, T), H_0^2 \times L^2)} \right) e^{\|p\|_{L^1((0, T), \mathbb{R})}}.$$

*Proof of Proposition 1.* The existence and uniqueness result comes from a fixed point argument on the map  $\Theta : C^0([0, T], H_0^2 \times L^2) \rightarrow C^0([0, T], H_0^2 \times L^2)$ ,  $\Theta(\xi) = w$ , where  $w$  is defined by

$$w(t) = e^{-\mathcal{A}t}w_0 + \int_0^t e^{-\mathcal{A}(t-s)}[-p(s)\mathcal{B}\xi(s) + F(s)]ds.$$

When  $\|p\|_{L^1((0, T), \mathbb{R})}$  is small enough,  $\Theta$  is a contraction of  $C^0([0, T], H_0^2 \times L^2)$ , and thus it has a unique fixed point  $w \in C^0([0, T], H_0^2 \times L^2)$  that satisfies (2.2). If  $\|p\|_{L^1((0, T), \mathbb{R})}$  is not small enough, one may use  $0 = T_0 < T_1 < \dots < T_n = T$  where, for  $i = 0, \dots, n-1$ ,  $\|p\|_{L^1((T_i, T_{i+1}), \mathbb{R})}$  is small enough so that the previous result holds on  $[T_i, T_{i+1}]$  for  $i = 0, \dots, n-1$ . Then we glue the solutions defined on  $[T_0, T_1], [T_1, T_2], \dots, [T_{n-1}, T_n]$ . We deduce from the equality (2.2) that

$$\|w(t)\|_{H_0^2 \times L^2} \leq \|w_0\|_{H_0^2 \times L^2} + \int_0^t [p(s)\|w(s)\|_{H_0^2 \times L^2} + \|F(s)\|_{H_0^2 \times L^2}]ds,$$

and thus

$$\|w(t)\|_{H_0^2 \times L^2} \leq \|w_0\|_{H_0^2 \times L^2} + \|F\|_{L^1((0, T), H_0^2 \times L^2)} + \int_0^t p(s)\|w(s)\|_{H_0^2 \times L^2}ds,$$

and Gronwall's Lemma gives (2.3).  $\square$

PROPOSITION 2. Let  $T > 0$ ,  $p \in W^{1,1}((0, T), \mathbb{R})$ ,  $w_0 \in H_{(0)}^4 \times H_0^2((0, 1), \mathbb{R})$ , and  $F \in W^{1,1}((0, T), H_0^2 \times L^2((0, 1), \mathbb{R}))$ . The solution  $w$  of (2.2) belongs to  $C^1([0, T], H_0^2 \times L^2)$  and to  $C^0([0, T], H_{(0)}^4 \times H_0^2)$ . Moreover, for every  $r > 0$ , there exists  $C(r) > 0$  such that, when  $\|p\|_{W^{1,1}} \leq r$ , the quantities

$$(2.4) \quad \|w\|_{C^0([0, T], H_{(0)}^4 \times H_0^2)} \text{ and } \|w\|_{C^1([0, T], H_0^2 \times L^2)}$$

are bounded by

$$(2.5) \quad C(r)[\|w_0\|_{H_{(0)}^4 \times H_0^2} + \|F\|_{W^{1,1}((0, T), H_0^2 \times L^2)}].$$

*Proof of Proposition 2.* Under the assumptions of Proposition 2, one can prove, by using the equality (2.2), that  $w \in C^1([0, T], H_0^2 \times L^2)$  and that

$$(2.6) \quad \frac{dw}{dt}(t) = -\mathcal{A}w(t) - p(t)\mathcal{B}w(t) + F(t) \text{ in } H_0^2 \times L^2((0, 1), \mathbb{R}) \quad \forall t \in [0, T].$$

We also have

$$\begin{aligned} \frac{dw}{dt}(t) &= e^{-\mathcal{A}t}[-\mathcal{A}w_0 - p(0)\mathcal{B}w_0 + F(0)] \\ &\quad + \int_0^t e^{-\mathcal{A}(t-s)} \left[ -p(s)\mathcal{B} \frac{dw}{dt}(s) - \dot{p}(s)\mathcal{B}w(s) + \dot{F}(s) \right] ds. \end{aligned}$$

By applying Proposition 1 to  $\frac{dw}{dt}$ , we get  $\frac{dw}{dt} \in C^0([0, T], H_0^2 \times L^2)$  and

$$\begin{aligned} \left\| \frac{dw}{dt} \right\|_{C^0([0, T], H_0^2 \times L^2)} &\leq C \left[ \|w_0\|_{H_{(0)}^4 \times H_0^2} + |p(0)| \|w_0\|_{H_0^2 \times L^2} + |F(0)|_{H_0^2 \times L^2} \right. \\ &\quad \left. + \|\dot{p}\|_{L^1} \|w\|_{C^0([0, T], H_0^2 \times L^2)} + \|\dot{F}\|_{L^1((0, T), H_0^2 \times L^2)} \right] e^{\|p\|_{L^1}}. \end{aligned}$$

Therefore, by using (2.3), we get a universal constant  $C_1 > 0$  such that

$$\begin{aligned} (2.7) \quad \left\| \frac{dw}{dt} \right\|_{C^0([0, T], H_0^2 \times L^2)} &\leq C_1 [\|w_0\|_{H_{(0)}^4 \times H_0^2} + \|p\|_{W^{1,1}} \|w_0\|_{H_0^2 \times L^2} + |F(0)|_{H_0^2 \times L^2} \\ &\quad + \|F\|_{W^{1,1}((0, T), H_0^2 \times L^2)} + \|p\|_{W^{1,1}} \|F\|_{L^1((0, T), H_0^2 \times L^2)}] e^{2\|p\|_{L^1}}. \end{aligned}$$

Then by using (2.6), we get  $w \in C^0([0, T], H_{(0)}^4 \times H_0^2((0, 1), \mathbb{R}))$  and

$$\begin{aligned} \|\mathcal{A}w\|_{C^0([0, T], H_0^2 \times L^2)} &= \left\| \frac{dw}{dt} + p\mathcal{B}w - F \right\|_{C^0([0, T], H_0^2 \times L^2)} \\ &\leq \left\| \frac{dw}{dt} \right\|_{C^0([0, T], H_0^2 \times L^2)} + \|p\|_{W^{1,1}} \|w\|_{C^0([0, T], H_0^2 \times L^2)} \\ &\quad + \|F\|_{C^0([0, T], H_0^2 \times L^2)}. \end{aligned}$$

Thanks to (2.7) and (2.3), the quantities in (2.4) are bounded by

$$\begin{aligned} C'(r)[\|w_0\|_{H_{(0)}^4 \times H_0^2} + \|p\|_{W^{1,1}} \|w_0\|_{H_0^2 \times L^2} + \|F\|_{W^{1,1}((0, T), H_0^2 \times L^2)} \\ + \|p\|_{W^{1,1}} \|F\|_{L^1((0, T), H_0^2 \times L^2)}] \end{aligned}$$

when  $\|p\|_{L^1((0, T), \mathbb{R})} \leq r$ , where  $C'(r) > 0$ . Finally, (2.5) is a direct consequence of the previous inequality.  $\square$

PROPOSITION 3. Let  $T > 0$ ,  $p \in W^{2,1}((0, T), \mathbb{R})$ , with  $p(0) = p(T) = 0$ ,  $w_0 \in H_{(0)}^6 \times H_{(0)}^4((0, 1), \mathbb{R})$ , and  $F \in W^{2,1}((0, T), H_0^2 \times L^2) \cap C^0([0, T], H^4 \times H^2)$ , with  $F(0), F(T) \in H_{(0)}^4 \times H_{(0)}^2((0, 1), \mathbb{R})$ . The solution  $w$  of (2.2) belongs to  $C^2([0, T], H_0^2 \times L^2)$ ,  $C^1([0, T], H_{(0)}^4 \times H_{(0)}^2)$ ,  $C^0([0, T], H^6 \times H^4)$ , and  $w(T) \in H_{(0)}^6 \times H_{(0)}^4$ . Moreover, for every  $r > 0$ , there exists  $C(r) > 0$  such that, when  $\|p\|_{W^{1,1}} \leq r$ , the quantities

$$\|w\|_{C^0([0, T], H^6 \times H^4)}, \|w\|_{C^1([0, T], H^4 \times H^2)}, \text{ and } \|w\|_{C^2([0, T], H_0^2 \times L^2)}$$

are bounded by

$$(2.8) \quad C(r)[\|w_0\|_{H_{(0)}^6 \times H_{(0)}^4} + \|p\|_{W^{2,1}}\|w_0\|_{H_0^2 \times L^2} + \|F\|_{W^{2,1}((0, T), H_0^2 \times L^2)} + \|p\|_{W^{2,1}}\|F\|_{L^1((0, T), H_0^2 \times L^2)} + \|F\|_{C^0([0, T], H^4 \times H^2)}].$$

*Proof of Proposition 3.* By applying Proposition 2 to  $\frac{dw}{dt}$ , we get  $\frac{dw}{dt} \in C^0([0, T], H_{(0)}^4 \times H_0^2) \cap C^1([0, T], H_0^2 \times L^2)$ . Notice that the assumptions  $p(0) = 0$  and  $F(0) \in H_{(0)}^4 \times H_{(0)}^2((0, 1), \mathbb{R})$  are useful to ensure that the initial condition

$$\frac{dw}{dt}(0) = -\mathcal{A}w_0 - p(0)\mathcal{B}w_0 + F(0)$$

belongs to  $H_{(0)}^4 \times H_0^2$ . Indeed, when  $w = (w_1, w_2) \in H_{(0)}^6 \times H_{(0)}^4((0, 1), \mathbb{R})$ , then  $w_1'' \in H^2((0, 1), \mathbb{R})$ , but in general  $w_1''$  does not vanish at  $x = 0, 1$ , and thus  $\mathcal{B}w$  does not belong to  $H_{(0)}^4 \times H_0^2((0, 1), \mathbb{R})$ . Proposition 2 also gives the existence of a constant  $C(r) > 0$  such that, when  $\|p\|_{W^{1,1}((0, T), \mathbb{R})} \leq r$ , the quantities

$$(2.9) \quad \left\| \frac{dw}{dt} \right\|_{C^0([0, T], H_{(0)}^4 \times H_0^2)} \text{ and } \left\| \frac{dw}{dt} \right\|_{C^1([0, T], H_0^2 \times L^2)}$$

are bounded by

$$\begin{aligned} & C \left[ \left\| \frac{dw}{dt}(0) \right\|_{H_{(0)}^4 \times H_0^2} + \|\dot{p}w - \dot{F}\|_{W^{1,1}((0, T), H_0^2 \times L^2)} \right] \\ & \leq C \left[ \|w_0\|_{H_{(0)}^6 \times H_{(0)}^4} + \|F(0)\|_{H_{(0)}^4 \times H_0^2} + \|p\|_{W^{2,1}}\|w\|_{C^0([0, T], H_0^2 \times L^2)} \right. \\ & \quad \left. + \|p\|_{W^{1,1}} \left\| \frac{dw}{dt} \right\|_{C^0([0, T], H_0^2 \times L^2)} + \|F\|_{W^{2,1}((0, T), H_0^2 \times L^2)} \right]. \end{aligned}$$

By using (2.3) and (2.5), we get the following bounds for the quantities written in (2.9):

$$(2.10) \quad C \left[ \|w_0\|_{H^6 \times H^4} + \|p\|_{W^{2,1}}\|w_0\|_{H_0^2 \times L^2} + \|F(0)\|_{H_{(0)}^4 \times H_0^2} + \|F\|_{W^{2,1}((0, T), H_0^2 \times L^2)} + \|p\|_{W^{2,1}}\|F\|_{L^1((0, T), H_0^2 \times L^2)} \right].$$

Now, by using (2.6), we deduce that  $\mathcal{A}w \in C^0([0, T], H^4 \times H^2)$  and  $\mathcal{A}w(T) \in H_{(0)}^4 \times H_0^2$  because  $p(T) = 0$  and  $F(T) \in H_{(0)}^4 \times H_0^2$ . Moreover,

$$\begin{aligned} \|\mathcal{A}w\|_{C^0([0, T], H^4 \times H^2)} & \leq \left\| \frac{dw}{dt} \right\|_{C^0([0, T], H_{(0)}^4 \times H_0^2)} + \|p\|_{W^{1,1}}\|w\|_{C^0([0, T], H_{(0)}^4 \times H_0^2)} \\ & \quad + \|F\|_{C^0([0, T], H^4 \times H^2)}. \end{aligned}$$

Thanks to the previous inequality, (2.10), and (2.5), we get (2.8).  $\square$

PROPOSITION 4. Let  $T > 0$ ,  $p \in W^{3,1}((0, T), \mathbb{R})$ , with  $p(0) = p(T) = \dot{p}(0) = \dot{p}(T) = 0$ ,  $w_0 \in H_{(0)}^8 \times H_{(0)}^6((0, 1), \mathbb{R})$ , and  $F \in W^{3,1}((0, T), H_0^2 \times L^2) \cap C^1([0, T], H^4 \times H^2) \cap C^0([0, T], H^6 \times H^4)$ , with  $F(0), F(T) \in H_{(0)}^6 \times H_{(0)}^4((0, 1), \mathbb{R})$ ,  $\dot{F}(0), \dot{F}(T) \in H_{(0)}^4 \times H_{(0)}^2((0, 1), \mathbb{R})$ . The solution  $w$  of (2.2) belongs to  $C^3([0, T], H_0^2 \times L^2)$ ,  $C^2([0, T], H_{(0)}^4 \times H_{(0)}^2)$ ,  $C^1([0, T], H^6 \times H^4)$ ,  $C^0([0, T], H^8 \times H^6)$ , and  $w(T) \in H_{(0)}^8 \times H_{(0)}^6$ . Moreover, for every  $r > 0$ , there exists  $C(r) > 0$  such that, when  $\|p\|_{W^{1,1}} \leq r$ , the quantities

$$\|w\|_{C^0([0, T], H^8 \times H^6)}, \|w\|_{C^1([0, T], H^6 \times H^4)}, \|w\|_{C^2([0, T], H^4 \times H^2)}, \text{ and } \|w\|_{C^3([0, T], H_0^2 \times L^2)}$$

are bounded by

$$(2.11) \quad C(r) \left[ \|w_0\|_{H^8 \times H^6} + \|p\|_{W^{2,1}} \|w_0\|_{H_{(0)}^4 \times H^2} + \|p\|_{W^{3,1}} \|w_0\|_{H_0^2 \times L^2} \right. \\ \left. + \|F\|_{W^{3,1}((0, T), H_0^2 \times L^2)} + \|p\|_{W^{2,1}} \|F\|_{W^{1,1}((0, T), H_0^2 \times L^2)} \right. \\ \left. + \|p\|_{W^{3,1}} \|F\|_{L^1((0, T), H_0^2 \times L^2)} \right. \\ \left. \|F\|_{C^0([0, T], H^6 \times H^4)} + \|p\|_{W^{2,1}} \|F\|_{C^0((0, T), H_0^2 \times L^2)} + \|F\|_{C^1([0, T], H^4 \times H^2)} \right].$$

*Proof of Proposition 4.* We apply Proposition 3 to  $\frac{dw}{dt}$ .  $\square$

**2.2. Spaces between which  $\Phi_T$  is  $C^1$ .** For  $T > 0$  fixed, we introduce the spaces

$$(2.12) \quad \begin{aligned} E_2 &:= H_0^2((0, 1), \mathbb{R}) \times L^2((0, 1), \mathbb{R}) \times L^2((0, T), \mathbb{R}), \\ E_4 &:= H_{(0)}^4((0, 1), \mathbb{R}) \times H_0^2((0, 1), \mathbb{R}) \times H^1((0, T), \mathbb{R}), \\ E_6 &:= H_{(0)}^6((0, 1), \mathbb{R}) \times H_{(0)}^4((0, 1), \mathbb{R}) \times H^2 \cap H_0^1((0, T), \mathbb{R}), \\ E_8 &:= H_{(0)}^8((0, 1), \mathbb{R}) \times H_{(0)}^6((0, 1), \mathbb{R}) \times H^3 \cap H_0^2((0, T), \mathbb{R}), \end{aligned}$$

and, for  $s > 0$ ,

$$(2.13) \quad F_s := H_{(0)}^s((0, 1), \mathbb{R}) \times H_{(0)}^{s-2}((0, 1), \mathbb{R}) \times H_{(0)}^s((0, 1), \mathbb{R}) \times H_{(0)}^{s-2}((0, 1), \mathbb{R}),$$

where  $H_{(0)}^s((0, 1), \mathbb{R})$  is defined by (1.2). Notice that the spaces  $E_a$  depend on  $T$  and should be called  $E_{a,T}$ . However, since no confusion is possible, we omit the subscript  $T$  in order to simplify the notations.

PROPOSITION 5. For every  $T > 0$  and for every  $a \in \{2, 4, 6, 8\}$ , the map  $\Phi_T$  defined by (1.8) is  $C^1$  from  $E_a$  to  $F_a$ , and, for every  $(u_0, \dot{u}_0, p) \in E_a$ ,  $d\Phi_T(u_0, \dot{u}_0, p) \cdot (U_0, \dot{U}_0, P) = (U_0, \dot{U}_0, U(T), U_t(T))$ , where  $U$  is the weak solution of

$$(2.14) \quad \begin{cases} U_{tt} + U_{xxxx} + p(t)U_{xx} + P(t)u_{xx} = 0, x \in (0, 1), t \in [0, T], \\ U = U_x = 0 \text{ at } x = 0, 1, \\ U(0, x) = U_0(x), \\ U_t(0, x) = \dot{U}_0(x), \end{cases}$$

and  $u$  is the weak solution of

$$(2.15) \quad \begin{cases} u_{tt} + u_{xxxx} + p(t)u_{xx} = 0, x \in (0, 1), t \in [0, T], \\ u = u_x = 0 \text{ at } x = 0, 1, \\ u(0, x) = u_0(x), \\ u_t(0, x) = \dot{u}_0(x). \end{cases}$$

*Proof of Proposition 5.* By using Propositions 1, 2, and 3 we see that  $\Phi_T : E_a \rightarrow F_a$  is continuous for  $a = 2, 4, 6$ . Let us prove that  $\Phi_T : E_2 \rightarrow F_2$  is  $C^1$  (the cases  $a =$



4, 6, 8 can be treated in the same way). Let  $(u_0, \dot{u}_0, p), (U_0, \dot{U}_0, P) \in E_2$ ,  $u$  be the weak solutions of (2.15),  $U$  be the weak solutions of (2.14), and  $\tilde{u}$  be the weak solution of

$$\begin{cases} \tilde{u}_{tt} + \tilde{u}_{xxxx} + (p + P)(t)\tilde{u}_{xx} = 0, x \in (0, 1), t \in [0, T], \\ \tilde{u} = \tilde{u}_x = 0 \text{ at } x = 0, 1, \\ \tilde{u}(0, x) = (u_0 + U_0)(x), \\ \tilde{u}_t(0, x) = (\dot{u}_0 + \dot{U}_0)(x). \end{cases}$$

Then  $\Delta := \tilde{u} - u - U$  is the weak solution of

$$\begin{cases} \Delta_{tt} + \Delta_{xxxx} + p(t)\Delta_{xx} + P(t)(\tilde{u} - u)_{xx} = 0, x \in (0, 1), t \in [0, T], \\ \Delta = \Delta_x = 0 \text{ at } x = 0, 1, \\ \Delta(0, x) = 0, \\ \Delta_t(0, x) = 0. \end{cases}$$

Thus, Proposition 1 gives

$$(2.16) \quad \begin{aligned} \|\Delta(T)\|_{H_0^2} + \|\Delta_t(T)\|_{L^2} &\leq \|P(t)(\tilde{u} - u)_{xx}\|_{L^1((0,T),L^2)} e^{\|p\|_{L^1}} \\ &\leq C\|P\|_{L^1}\|\tilde{u} - u\|_{C^0([0,T],H^2)}. \end{aligned}$$

Moreover,  $\tilde{u} - u$  is the weak solution of

$$\begin{cases} (\tilde{u} - u)_{tt} + (\tilde{u} - u)_{xxxx} + p(t)(\tilde{u} - u)_{xx} + P(t)\tilde{u}_{xx} = 0, x \in (0, 1), t \in [0, T], \\ \tilde{u} - u = (\tilde{u} - u)_x = 0 \text{ at } x = 0, 1, \\ (\tilde{u} - u)(0, x) = U_0(x), \\ (\tilde{u} - u)_t(0, x) = \dot{U}_0(x). \end{cases}$$

Thus, Proposition 1 gives

$$(2.17) \quad \|\tilde{u} - u\|_{C^0([0,T],H^2)} \leq \left( \|(U_0, \dot{U}_0)\|_{H^2 \times L^2} + \|P\|_{L^1} \|\tilde{u}_{xx}\|_{C^0([0,T],L^2)} \right) e^{\|p\|_{L^1}}.$$

Again, thanks to Proposition 1, we have

$$(2.18) \quad \|\tilde{u}\|_{C^0([0,T],H^2)} \leq \|(u_0 + U_0, \dot{u}_0 + \dot{U}_0)\|_{H^2 \times L^2} e^{\|p+P\|_{L^1}}.$$

Therefore, by using (2.16), (2.17), and (2.18), we get

$$\|\Delta(T)\|_{H_0^2} + \|\Delta_t(T)\|_{L^2} = o\left(\|(U_0, \dot{U}_0, P)\|_{E_2}\right)$$

when  $\|(U_0, \dot{U}_0, P)\|_{E_2} \rightarrow 0$ . This proves that  $\Phi_T$  is differentiable at  $(u_0, \dot{u}_0, p)$  and that

$$d\Phi_T(u_0, \dot{u}_0, p).(U_0, \dot{U}_0, P) = (U_0, \dot{U}_0, U(T), U_t(T)).$$

The continuity of the map

$$\begin{aligned} E_2 &\rightarrow \mathcal{L}(E_2, F_2) \\ (u_0, \dot{u}_0, p) &\mapsto d\Phi_T(u_0, \dot{u}_0, p) \end{aligned}$$

is a consequence of the estimate (2.3).  $\square$

### 3. Controllability of the linearized system around $(u^{ref}, \dot{u}^{ref}, p \equiv 0)$ .

The aim of this section is the proof of Theorem 3. First, in subsection 3.1, we prove some preliminary results, mainly about the eigenvalues and the eigenvectors of the operator  $A$  defined by (1.1). The reading of the proofs in this subsection is not necessary for the understanding of the next sections. Then, in subsection 3.2, we prove Theorem 3.

### 3.1. Preliminaries.

PROPOSITION 6. *The eigenvalues  $(\lambda_n)_{n \in \mathbb{N}^*}$  of the operator  $A$  are the numbers*

$$(3.1) \quad \lambda_n = \nu_n^4,$$

where  $(\nu_n)_{n \in \mathbb{N}^*}$  is the increasing sequence of positive solutions of the equation

$$(3.2) \quad \cos(\nu_n) \cosh(\nu_n) = 1.$$

We have, for every  $k \in \mathbb{N}^*$ ,

$$(3.3) \quad \nu_{2k-1} \in (2k\pi - \pi/2, 2k\pi - \pi/4), \quad \nu_{2k} \in (2k\pi + \pi/4, 2k\pi + \pi/2).$$

We have, for every  $n \in \mathbb{N}^*$ ,

$$(3.4) \quad \nu_n = \frac{\pi}{2}(2n+1) - (-1)^n x_n, \text{ where } 0 < x_n < \frac{\pi}{2 \cosh(\nu_n)},$$

$$(3.5) \quad \left| \sqrt{\lambda_n} - \frac{\pi^2}{4} K_n \right| < \frac{\pi^2}{16},$$

where, for every  $n \in \mathbb{N}^*$ ,

$$(3.6) \quad K_n := (2n+1)^2.$$

For every  $n \in \mathbb{N}^*$ , the function

$$(3.7) \quad v_n := \xi_n(\cos(\nu_n x) - \cosh(\nu_n x)) + \zeta_n(\sin(\nu_n x) - \sinh(\nu_n x)),$$

where

$$(3.8) \quad \xi_n := \sin(\nu_n) - \sinh(\nu_n), \quad \zeta_n := -\cos(\nu_n) + \cosh(\nu_n),$$

is an eigenvector of  $A$  associated to the eigenvalue  $\lambda_n$  and satisfies

$$(3.9) \quad v_n(1-x) = (-1)^n v_n(x).$$

Moreover, there exists a constant  $C \in \mathbb{R}^*$  such that, when  $n \rightarrow +\infty$ ,

$$(3.10) \quad \|v_n\|_{L^2((0,1),\mathbb{R})} \sim C e^{\nu_n}.$$

*Proof of Proposition 6.* The relation (3.2) comes from the condition  $v = v' = 0$  on  $x = 0, 1$  imposed on any eigenvector  $v$  of the operator  $A$ . The intermediate values theorem gives (3.3). Equation (3.2) provides, for every  $n \in \mathbb{N}^*$ ,

$$(3.11) \quad \sin(x_n) \cosh(\nu_n) = 1,$$

which gives (3.4) thanks to the convexity inequality

$$(3.12) \quad x \leq \frac{\pi}{2} \sin(x) \quad \forall x \in [0, \pi/2].$$

For every  $n \in \mathbb{N}^*$ , we have

$$\sqrt{\lambda_n} - \frac{\pi^2}{4} K_n = \pi(2n+1)x_n - (-1)^n x_n^2.$$

Thus, we just need to justify that, for every  $n \in \mathbb{N}^*$ ,

$$\frac{2(2n+1)}{\cosh \nu_n} + \frac{1}{\cosh(\nu_n)^2} < \frac{1}{4}.$$

This can be proved for  $n = 1$  by using  $\nu_1 > 3\pi/2$  and for  $n \geq 2$  by using  $\nu_n \geq n\pi$ . The property (3.9) comes from the explicit expression (3.7) together with (3.2) and the relation  $\sin(\nu_n) = (-1)^n \sqrt{1 - \cos(\nu_n)^2}$ , which is a consequence of (3.3). By using (3.7) and a change of variable, we get

$$\int_0^1 v_n(x)^2 dx = \frac{\xi_n^2}{\nu_n} I_1(\nu_n) + \frac{\zeta_n^2}{\nu_n} I_2(\nu_n) + \frac{2\xi_n \zeta_n}{\nu_n} I_3(\nu_n),$$

where

$$\begin{aligned} I_1(\nu) &:= \int_0^\nu [\cos(y) - \cosh(y)]^2 dy \\ &= \frac{\sinh(2\nu)}{4} + \nu - \sin(\nu) \cosh(\nu) - \cos(\nu) \sinh(\nu) + \frac{\sin(2\nu)}{4}, \\ I_2(\nu) &:= \int_0^\nu [\sin(y) - \sinh(y)]^2 dy \\ &= \frac{\sinh(2\nu)}{4} - \sin(\nu) \cosh(\nu) + \cos(\nu) \sinh(\nu) - \frac{\sin(2\nu)}{4}, \\ I_3(\nu) &:= \int_0^\nu [\cos(y) - \cosh(y)][\sin(y) - \sinh(y)] dy \\ &= \frac{\cosh(2\nu)}{4} - \sin(\nu) \sinh(\nu) - \frac{\cos(2\nu)}{4}. \end{aligned}$$

By using the following behaviors, when  $n \rightarrow +\infty$ ,

$$(3.13) \quad \begin{aligned} \cos(\nu_n) &= O(e^{-\nu_n}), & \sin(\nu_n) &= (-1)^n + O(e^{-2\nu_n}), \\ \sin(2\nu_n) &= O(e^{-\nu_n}), & \cos(2\nu_n) &= -1 + O(e^{-2\nu_n}), \end{aligned}$$

which are consequences of (3.4), we get

$$\int_0^1 v_n(x)^2 dx \sim C e^{2\nu_n}. \quad \square$$

The orthonormal basis  $(\varphi_n)_{n \in \mathbb{N}^*}$  of  $L^2((0,1), \mathbb{R})$  made of eigenvectors of  $A$  has been introduced in section 1.1. Up to a change of sign, one may assume that

$$(3.14) \quad \varphi_n = \frac{v_n}{\|v_n\|_{L^2((0,1), \mathbb{R})}}.$$

This equality will be assumed in the remainder of this article.

**PROPOSITION 7.** *Let  $m \in \mathbb{N}^*$ . For every  $n \in \mathbb{N}^*$ , we have*

$$(3.15) \quad \langle \varphi_m'', \varphi_n \rangle \neq 0 \text{ iff } m \text{ and } n \text{ have the same parity.}$$

Moreover, there exists a constant  $C_m$  such that, when  $n$  tends to  $+\infty$  with the same parity as  $m$ ,

$$(3.16) \quad \langle \varphi_m'', \varphi_n \rangle \sim \frac{C_m}{n}.$$

*Proof of Proposition 7.* When  $m$  and  $n$  have different parities, the equality  $\langle \varphi_m'', \varphi_n \rangle = 0$  comes from (3.14) and (3.9). We have

$$\langle \varphi_m'', \varphi_m \rangle = - \int_0^1 |\varphi_m'(x)|^2 dx < 0.$$

Let  $n \in \mathbb{N}^*$  with the same parity as  $m$  and different from  $m$ . Thanks to integrations by parts and (3.9), we get

$$(\lambda_n - \lambda_m) \langle v_m'', v_n \rangle = 2(v_m''' v_n'' - v_m'' v_n''')(0).$$

The explicit expression (3.7) leads to

$$(3.17) \quad (\lambda_n - \lambda_m) \langle v_m'', v_n \rangle = 8\nu_m^2 \nu_n^2 (-\xi_m \nu_n \zeta_n + \zeta_m \nu_m \xi_n).$$

The function  $f : \mathbb{R}_+ \rightarrow \mathbb{R}$  defined by

$$f(x) := \frac{\sinh(x) - \sin(x)}{x(\cosh(x) - \cos(x))}$$

decreases on  $\mathbb{R}_+$ . If  $n < m$  (resp.,  $n > m$ ), the inequality  $f(\nu_n) > f(\nu_m)$  (resp.,  $f(\nu_n) < f(\nu_m)$ ) provides  $-\xi_m \nu_n \zeta_n + \zeta_m \nu_m \xi_n < 0$  (resp.,  $> 0$ ). Thus  $\langle v_m'', v_n \rangle \neq 0$ . Thanks to (3.8), we get

$$(3.18) \quad \langle v_m'', v_n \rangle \sim -\frac{4\nu_m^2 \xi_m}{\nu_n} e^{\nu_n} \text{ when } n \rightarrow +\infty, \text{ with the same parity as } m,$$

which together with (3.10) gives the asymptotic behavior (3.16).  $\square$

At this step, one can justify the following property, claimed in section 1.2.

**PROPOSITION 8.** *The operator  $\mathcal{A}$  defined by (1.7) and (1.1) does not generate a  $C^0$ -semigroup of bounded operators of  $\bar{X} := H_{(0)}^3 \times H^1((0, 1), \mathbb{R})$ .*

*Proof of Proposition 8.* We argue by contradiction. Let us assume that  $\mathcal{A}$  generates a  $C^0$ -semigroup of bounded operators of  $\bar{X}$ . Then there exists  $m > 0$  such that, for every  $t \in [0, 1]$ , for every  $(u_0, \dot{u}_0) \in H_{(0)}^3 \times H^1((0, 1), \mathbb{R})$ ,

$$e^{-t\mathcal{A}} \begin{pmatrix} u_0 \\ \dot{u}_0 \end{pmatrix} \in H_{(0)}^3 \times H^1((0, 1), \mathbb{R})$$

and

$$(3.19) \quad \left\| e^{-t\mathcal{A}} \begin{pmatrix} u_0 \\ \dot{u}_0 \end{pmatrix} \right\|_{H_{(0)}^3 \times H^1} \leq m \|(u_0, \dot{u}_0)\|_{H_{(0)}^3 \times H^1}.$$

Let us consider  $(u_0, \dot{u}_0) \in H_{(0)}^3 \times H^1((0, 1), \mathbb{R})$  defined by  $u_0 := 0$  and  $\dot{u}_0 := \varphi_1''$ . By using (2.1), we get

$$e^{-t\mathcal{A}} \begin{pmatrix} u_0 \\ \dot{u}_0 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{\infty} \frac{1}{\sqrt{\lambda_k}} \langle \varphi_1'', \varphi_k \rangle \sin(\sqrt{\lambda_k} t) \varphi_k \\ \sum_{k=1}^{\infty} \langle \varphi_1'', \varphi_k \rangle \cos(\sqrt{\lambda_k} t) \varphi_k \end{pmatrix}.$$

Thanks to (3.19), the  $H_{(0)}^3((0, 1), \mathbb{R})$ -norm of the first component of the right-hand side is bounded by  $m\|\varphi_1''\|_{H^1}$  for every  $t \in [0, 1]$ ; i.e., (see (1.3) for the definition of the  $H_{(0)}^3((0, 1), \mathbb{R})$ -norm)

$$(3.20) \quad \sum_{k=1}^{\infty} |\lambda_k^{1/4} \langle \varphi_1'', \varphi_k \rangle \sin(\sqrt{\lambda_k} t)|^2 \leq m^2 \|\varphi_1''\|_{H^1}^2 \quad \forall t \in [0, 1].$$

Thanks to (3.1), (3.3), and Proposition 7, there exists an odd integer  $N_0 \in \mathbb{N}^*$  and  $C_0 > 0$  such that

$$(3.21) \quad |\lambda_k^{1/4} \langle \varphi_1'', \varphi_k \rangle|^2 \geq C_0 \quad \forall k \geq N_0, k \text{ odd}.$$

Thus, (3.20) and (3.21) imply

$$(3.22) \quad C_0 \sum_{k=N_0, k \text{ odd}}^{\infty} |\sin(\sqrt{\lambda_k} t)|^2 \leq m^2 \|\varphi_1''\|_{H^1}^2 \quad \forall t \in [0, 1].$$

Let  $N \in \mathbb{N}^*$  be such that  $N \geq N_0$  and  $\pi/(2\sqrt{\lambda_N}) \leq 1$ . Let  $t_N := \pi/(2\sqrt{\lambda_N})$ . For every  $k \in \mathbb{N}^*$ , odd with  $N_0 \leq k \leq N$ , we have  $\sqrt{\lambda_k} t_N \in (0, \pi/2)$ , and thus  $\sin(\sqrt{\lambda_k} t_N) \geq (2/\pi)\sqrt{\lambda_k} t_N$ . We also have  $t_N \in [0, 1]$ , so we deduce from (3.22) that

$$\frac{1}{2} C_0 \left( \frac{2}{\pi} \right)^2 \sum_{k=N_0}^N \lambda_k t_N^2 \leq m^2 \|\varphi_1''\|_{H^1}^2.$$

Thanks to (3.1), (3.3), and the definition of  $t_N$ , there exists  $C_1 > 0$  such that

$$\sum_{k=N_0}^N \lambda_k t_N^2 \geq \frac{C_1}{N^4} \sum_{k=N_0}^N k^4 \geq \frac{C_1}{N^4} \int_{N_0}^N x^4 dx \geq \frac{C_1}{5N^4} (N^5 - N_0^5).$$

We have proved that, for every  $N \in \mathbb{N}^*$  such that  $N \geq N_0$  and  $\pi/(2\sqrt{\lambda_N}) \leq 1$ , the following inequality holds:

$$\frac{1}{2} C_0 \left( \frac{2}{\pi} \right)^2 \frac{C_1}{5N^4} (N^5 - N_0^5) \leq m^2 \|\varphi_1''\|_{H^1}^2.$$

We get a contradiction by considering large enough  $N$ .  $\square$

LEMMA 1. *The frequencies*

$$0, 2\sqrt{\lambda_2}, 2\sqrt{\lambda_3}, \sqrt{\lambda_3} \pm \sqrt{\lambda_1}, \sqrt{\lambda_4} \pm \sqrt{\lambda_2}, \sqrt{\lambda_{2k}} \pm \sqrt{\lambda_2}, \sqrt{\lambda_{2k-1}} \pm \sqrt{\lambda_3}$$

for  $k \in \mathbb{N}^*, k \geq 3$ ,

are all different.

*Proof of Lemma 1.* First, we claim that the nonnegative integers

$$0, 2K_2, 2K_3, K_3 \pm K_1, K_4 \pm K_2, K_{2k-1} \pm K_3, K_{2k} \pm K_2 \text{ for } k \in \mathbb{N}^*, k \geq 3,$$

where  $K_n$  is defined by (3.6), are all different. First, for every  $k \geq 5$ , we have

$$K_{2k-1} + K_3 < K_{2k} - K_2 < K_{2k} + K_2 < K_{2k+1} - K_3.$$

Indeed, for  $k \geq 5$ , we have

$$\begin{aligned} (K_{2k} - K_2) - (K_{2k-1} + K_3) &= (4k+1)^2 - (4k-1)^2 - 25 - 49 \geq 16k - 74 \geq 6, \\ (K_{2k+1} - K_3) - (K_{2k} + K_2) &= (4k+3)^2 - (4k+1)^2 - 25 - 49 \geq 16k - 66 \geq 14. \end{aligned}$$

Moreover, we have

$$\begin{aligned} 2K_2 &= 50, & 2K_3 &= 98, & K_3 - K_1 &= 40, & K_3 + K_1 &= 58, \\ K_4 - K_2 &= 56, & K_4 + K_2 &= 106, & K_5 - K_3 &= 72, & K_5 + K_3 &= 170, \\ K_6 - K_2 &= 144, & K_6 + K_2 &= 194, & K_7 - K_3 &= 176, & K_7 + K_3 &= 274, \\ K_8 - K_2 &= 264, & K_8 + K_2 &= 314, & K_9 - K_3 &= 312, & & \end{aligned}$$

and thus the claim is proved. We deduce Lemma 1 from the previous result and the inequality (3.5).  $\square$

### 3.2. Proof of Theorem 3.

*Proof of statement (1) of Theorem 3.* Since the system  $(\Sigma_l^{ref})$  is linear and  $e^{-T\mathcal{A}}$  is a bounded operator of  $H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R})$ , it is sufficient to prove Theorem 3 with  $U_0 = \dot{U}_0 = 0$ .

Let  $T > 0$  and  $U$  be a solution of  $(\Sigma_l^{ref})$  for some  $P \in L^2((0, T), \mathbb{R})$ , with  $U(0) = \dot{U}(0) = 0$ . Then  $(U, U_t) \in C^0([0, T], H_0^2 \times L^2((0, 1), \mathbb{R}))$ . The family  $(\varphi_k)_{k \in \mathbb{N}^*}$  is an orthonormal basis of  $L^2((0, 1), \mathbb{R})$ , and thus, for every  $t \in [0, T]$ ,

$$U(t) = \sum_{k=1}^{\infty} x_k(t) \varphi_k, \text{ where } x_k(t) := \int_0^1 U(t, x) \varphi_k(x) dx.$$

By recalling that  $u^{ref}$  is given by (1.5), the partial differential equation satisfied by  $U$  provides, for every  $k \in \mathbb{N}^*$ , the following explicit expression:

$$x_k(t) = -\frac{1}{\sqrt{\lambda_k}} \int_0^t P(\tau) [b_k \sin(\sqrt{\lambda_2} \tau) + c_k \sin(\sqrt{\lambda_3} \tau)] \sin[\sqrt{\lambda_k}(t - \tau)] d\tau,$$

where, for every  $k \in \mathbb{N}^*$ ,

$$b_k := \langle \varphi_2'', \varphi_k \rangle, \quad c_k := \langle \varphi_3'', \varphi_k \rangle.$$

The equality  $(U(T), \dot{U}(T)) = (U_T, \dot{U}_T)$  is equivalent to

$$(3.23) \quad \begin{aligned} & \int_0^T P(t) [b_k \sin(\sqrt{\lambda_2} t) + c_k \sin(\sqrt{\lambda_3} t)] e^{-i\sqrt{\lambda_k} t} dt \\ &= -e^{-i\sqrt{\lambda_k} T} \left( \langle \dot{U}_T, \varphi_k \rangle + i\sqrt{\lambda_k} \langle U_T, \varphi_k \rangle \right) \quad \forall k \in \mathbb{N}^*. \end{aligned}$$

Thanks to Proposition 7, the equality (3.23) is satisfied if  $P$  solves the following moment problem:

$$(3.24) \quad \begin{cases} \int_0^T P(t) \sin(\sqrt{\lambda_3} t) e^{-i\sqrt{\lambda_{2k-1}} t} dt = d_{2k-1}(U_T, \dot{U}_T) \quad \forall k \in \mathbb{N}^*, \\ \int_0^T P(t) \sin(\sqrt{\lambda_2} t) e^{-i\sqrt{\lambda_{2k}} t} dt = d_{2k}(U_T, \dot{U}_T) \quad \forall k \in \mathbb{N}^*, \end{cases}$$

where, for every  $k \in \mathbb{N}^*$ ,

$$\begin{aligned} d_{2k-1}(U_T, \dot{U}_T) &:= -\frac{e^{-i\sqrt{\lambda_{2k-1}} T}}{c_{2k-1}} \left( \langle \dot{U}_T, \varphi_{2k-1} \rangle + i\sqrt{\lambda_{2k-1}} \langle U_T, \varphi_{2k-1} \rangle \right), \\ d_{2k}(U_T, \dot{U}_T) &:= -\frac{e^{-i\sqrt{\lambda_{2k}} T}}{b_{2k}} \left( \langle \dot{U}_T, \varphi_{2k} \rangle + i\sqrt{\lambda_{2k}} \langle U_T, \varphi_{2k} \rangle \right). \end{aligned}$$

The moment problem (3.24) is satisfied, in particular, when

$$(3.25) \quad \begin{cases} \int_0^T P(t) dt = 0, \\ \int_0^T P(t) e^{-i2\sqrt{\lambda_3} t} dt = -2id_3, \\ \int_0^T P(t) e^{i(-\sqrt{\lambda_{2k-1}} \pm \sqrt{\lambda_3}) t} dt = \pm id_{2k-1}(U_T, \dot{U}_T) \quad \forall k \in \mathbb{N}^*, k \neq 2, \\ \int_0^T P(t) e^{-i2\sqrt{\lambda_2} t} dt = -2id_2, \\ \int_0^T P(t) e^{i(-\sqrt{\lambda_{2k}} \pm \sqrt{\lambda_2}) t} dt = \pm id_{2k}(U_T, \dot{U}_T) \quad \forall k \in \mathbb{N}^*, k \neq 1. \end{cases}$$

Let  $(\omega_n)_{n \in \mathbb{N}}$  be the nondecreasing sequence of the frequencies appearing in the previous moment problem, written in the following way:

$$\int_0^T P(t) e^{i\omega_n t} dt = \delta_n \quad \forall n \in \mathbb{N}.$$

For large enough indexes, the successive terms of the sequence  $(\omega_n)_{n \in \mathbb{N}}$  are

$$\sqrt{\lambda_{2k-1}} - \sqrt{\lambda_3} < \sqrt{\lambda_{2k-1}} + \sqrt{\lambda_3} < \sqrt{\lambda_{2k}} - \sqrt{\lambda_2} < \sqrt{\lambda_{2k}} + \sqrt{\lambda_2} < \sqrt{\lambda_{2k+1}} - \sqrt{\lambda_3} < \dots.$$

The gap between the first and the second terms (resp., the third and the fourth terms) is  $2\sqrt{\lambda_3}$  (resp.,  $2\sqrt{\lambda_2}$ ). The gap between the second and the third terms (resp., the fourth and the fifth terms) tends to  $+\infty$  when  $k \rightarrow +\infty$ ; indeed, by using (3.1) and (3.3), we have

$$\begin{aligned} (\sqrt{\lambda_{2k}} - \sqrt{\lambda_2}) - (\sqrt{\lambda_{2k-1}} + \sqrt{\lambda_3}) &= \nu_{2k}^2 - \nu_{2k-1}^2 - \nu_2^2 - \nu_3^2 \\ &= (\nu_{2k} - \nu_{2k-1})(\nu_{2k} + \nu_{2k-1}) - \nu_2^2 - \nu_3^2 \\ &\geq \frac{\pi}{2} \left( 4k\pi - \frac{\pi}{4} \right) - \nu_2^2 - \nu_3^2, \\ (\sqrt{\lambda_{2k+1}} - \sqrt{\lambda_3}) - (\sqrt{\lambda_{2k}} + \sqrt{\lambda_2}) &= \nu_{2k+1}^2 - \nu_{2k}^2 - \nu_2^2 - \nu_3^2 \\ &= (\nu_{2k+1} - \nu_{2k})(\nu_{2k+1} + \nu_{2k}) - \nu_2^2 - \nu_3^2 \\ &\geq \pi \left( 4k\pi + \frac{7\pi}{4} \right) - \nu_2^2 - \nu_3^4. \end{aligned}$$

Thus,

$$\liminf_{n \rightarrow +\infty} (\omega_{n+1} - \omega_n) = 2\sqrt{\lambda_2}.$$

The frequencies appearing in (3.25) are all different (see Lemma 1), and thus the moment problem (3.25) has a solution  $P \in L^2((0, T), \mathbb{R})$  when  $T > \pi/\sqrt{\lambda_2}$  and  $(d_n(U_T, \dot{U}_T))_{n \in \mathbb{N}^*}$  belongs to  $l^2(\mathbb{N}^*, \mathbb{C})$  (see [22, Chap. 1.2]). Thanks to (3.16), the assumption that  $(U_T, \dot{U}_T) \in H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R})$  guarantees that  $(d_n(U_T, \dot{U}_T))_{n \in \mathbb{N}^*}$  belongs to  $l^2(\mathbb{N}^*, \mathbb{C})$ .  $\square$

*Proof of statement (2) of Theorem 3.* We assume that the system  $(\Sigma_l^{ref})$  is controllable in  $H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R})$  with control functions in  $H^1((0, T), \mathbb{R})$ . Thanks to the equivalence between the controllability of  $(\Sigma_l^{ref})$  and the solvability of the moment problem (3.24), we deduce that, for every  $d = (d_k)_{k \in \mathbb{N}^*} \in l^2(\mathbb{N}^*, \mathbb{C})$ , there exists  $P \in H^1((0, T), \mathbb{R})$  such that

$$(3.26) \quad \int_0^T P(t) \sin(\sqrt{\lambda_3} t) e^{-i\sqrt{\lambda_{2k-1}} t} dt = d_k \quad \forall k \in \mathbb{N}^*.$$

However, thanks to (3.1) and (3.3), for any  $P \in H^1((0, T), \mathbb{R})$ , an integration by parts shows that

$$\left| \int_0^T P(t) \sin(\sqrt{\lambda_3} t) e^{-i\sqrt{\lambda_{2k-1}} t} dt \right| \leq \frac{C}{k^2} \|P\|_{H^1((0, T), \mathbb{R})}.$$

Thus, for every  $d = (d_k)_{k \in \mathbb{N}^*} \in l^2(\mathbb{N}^*, \mathbb{C})$ , there exists  $C > 0$  such that

$$|d_k| \leq \frac{C}{k^2}.$$

We get a contradiction by considering, for example,  $d_k = 1/k$ . Therefore, the system  $(\Sigma_l^{ref})$  is not controllable in  $H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R})$  with control functions in  $H^1((0, T), \mathbb{R})$ .

We assume that the system  $(\Sigma_l^{ref})$  is controllable in  $H_0^2 \times L^2((0, 1), \mathbb{R})$  with control functions in  $L^2((0, T), \mathbb{R})$ . Then, for every  $d = (d_k)_{k \in \mathbb{N}^*} \in h^{-1}(\mathbb{N}^*, \mathbb{C})$ , i.e.,

$$\sum_{k=1}^{\infty} \left| \frac{1}{k} d_k \right|^2 < \infty,$$

there exists  $P \in L^2((0, T), \mathbb{R})$  such that (3.26) holds. However, for any  $P \in L^2((0, T), \mathbb{R})$ , the Cauchy–Schwarz inequality shows that

$$\left| \int_0^T P(t) \sin(\sqrt{\lambda_3} t) e^{-i\sqrt{\lambda_{2k-1}} t} dt \right| \leq \sqrt{T} \|P\|_{L^2((0, T), \mathbb{R})}.$$

Thus, for every  $d \in h^{-1}(\mathbb{N}^*, \mathbb{C})$ , there exists  $C > 0$  such that  $|d_k| \leq C$  for every  $k \in \mathbb{N}^*$ . We get a contradiction by considering, for example,  $d_k = k^{1/4}$ . Therefore, the system  $(\Sigma_l^{ref})$  is not controllable in  $H_0^2 \times L^2((0, 1), \mathbb{R})$  with control functions in  $L^2((0, T), \mathbb{R})$ .  $\square$

**4. Why the inverse mapping theorem does not apply.** As we have seen in Proposition 5, for every  $T > 0$ , the map  $\Phi_T$  defined by (1.8) is  $C^1$  between the following spaces:

$$H_0^2((0, 1), \mathbb{R}) \times L^2((0, 1), \mathbb{R}) \times L^2((0, T), \mathbb{R}) \rightarrow H_0^2 \times L^2 \times H_0^2 \times L^2((0, 1), \mathbb{R}).$$

In the same way we proved that  $\Phi_T$  is  $C^1$  from  $H_{(0)}^4((0, 1), \mathbb{R}) \times H_{(0)}^2((0, 1), \mathbb{R}) \times H^1((0, T), \mathbb{R})$  to  $H_{(0)}^4 \times H_{(0)}^2 \times H_{(0)}^4 \times H_{(0)}^2((0, 1), \mathbb{R})$ , one can prove that  $\Phi_T$  is  $C^1$  between the following spaces:

$$H_{(0)}^3((0, 1), \mathbb{R}) \times H_0^1((0, 1), \mathbb{R}) \times H^1((0, T), \mathbb{R}) \rightarrow H_{(0)}^3 \times H_0^1 \times H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R}).$$

Thus, in order to apply the inverse mapping theorem on the map  $\Phi_T$ , one needs to prove the surjectivity of  $d\Phi_T(u_0^{ref}, \dot{u}_0^{ref}, 0)$

- from  $H_0^2((0, 1), \mathbb{R}) \times L^2((0, 1), \mathbb{R}) \times L^2((0, T), \mathbb{R})$  to  $H_0^2 \times L^2 \times H_0^2 \times L^2((0, 1), \mathbb{R})$
- or from  $H_{(0)}^3((0, 1), \mathbb{R}) \times H_0^1((0, 1), \mathbb{R}) \times H^1((0, T), \mathbb{R})$  to  $H_{(0)}^3 \times H_0^1 \times H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R})$ .

It means that one needs to prove the controllability of the linear system  $(\Sigma_l^{ref})$

- in  $H_0^2 \times L^2$  with controls in  $L^2$
- or in  $H_{(0)}^3 \times H_0^1$  with controls in  $H^1$ .

However, as seen in statement (2) of Theorem 3, this is impossible. Thus the inverse mapping theorem cannot be applied with these spaces.

Let us emphasize that the inverse mapping theorem could be applied if we could prove that the map  $\Phi_T$  is well-defined and  $C^1$  between the following spaces:

$$H_{(0)}^3((0, 1), \mathbb{R}) \times H_0^1((0, 1), \mathbb{R}) \times L^2((0, T), \mathbb{R}) \rightarrow H_{(0)}^3 \times H_0^1 \times H_{(0)}^3 \times H_0^1((0, 1), \mathbb{R}).$$

With further developments, one can prove that the map  $\Phi_T$  is well-defined between these spaces but it is not  $C^1$ . Thus, we have the same pathology as in [8], [3], [4], [2]. The strategy developed in this article to solve this pathology is different from the one of [8] and similar to the one of [3], [4], [2].



**5. The Nash–Moser theorem used.** In order to get local controllability for the nonlinear system  $(\Sigma)$  around  $(u^{ref}, \dot{u}^{ref}, p \equiv 0)$ , we use a Nash–Moser theorem given by Hörmander in [20]. In this section, we recall the context and the statement of this theorem. We repeat the proof in order to justify that, in our situation, we need only a finite number of bounds on the right inverse of the differential map.

We consider a family of Hilbert spaces  $(E_a)_{a \in [2,8]}$  with continuous injections  $E_b \rightarrow E_a$  of norm  $\leq 1$  when  $b \geq a$ . We suppose that we have linear operators  $S_\theta : E_2 \rightarrow E_8$  for  $\theta \geq 1$ . We also assume that there exists a constant  $K > 0$  such that, for every  $a, b \in [2, 8]$  and for every  $u \in E_a$ ,

$$(5.1) \quad \|S_\theta u\|_b \leq K \|u\|_a \text{ when } b \leq a,$$

$$(5.2) \quad \|S_\theta u\|_b \leq K \theta^{b-a} \|u\|_a \text{ when } b > a,$$

$$(5.3) \quad \|u - S_\theta u\|_b \leq K \theta^{b-a} \|u\|_a \text{ when } b < a,$$

$$(5.4) \quad \left\| \frac{d}{d\theta} S_\theta u \right\|_b \leq K \theta^{b-a-1} \|u\|_a.$$

We fix a sequence  $(\theta_j)_{j \in \mathbb{N}}$  of the form  $\theta_j := (j+1)^\delta$  where  $0 < \delta$ , and we set, for every  $j \in \mathbb{N}$ ,  $\Delta_j := \theta_{j+1} - \theta_j$ . For every  $u \in E_a$ , we have a decomposition

$$u = \sum_{j=0}^{\infty} \Delta_j R_j u$$

with convergence in  $E_b$  when  $b < a$ , where

$$R_j u := \frac{1}{\Delta_j} (S_{\theta_{j+1}} - S_{\theta_j}) u \text{ if } j > 0 \text{ and } R_0 u := \frac{1}{\Delta_0} S_{\theta_1} u.$$

Moreover there exists a constant  $K'$  such that, for every  $b \in [2, 8]$ ,

$$\|R_j u\|_b \leq K' \theta_j^{b-a-1} \|u\|_a.$$

From (5.2) and (5.3), we get the logarithmic convexity of the norms: There exists a constant  $c \geq 1$  such that, for every  $a, b \in [2, 8]$  with  $a < b$ ,  $\lambda \in [0, 1]$ , and  $u \in E_b$ ,

$$(5.5) \quad \|u\|_{\lambda a + (1-\lambda)b} \leq c \|u\|_a^\lambda \|u\|_b^{1-\lambda}.$$

We refer to [20] for the proof of the two previous properties.

We have another family  $(F_a)_{a \in [2,8]}$  with the same properties as above, and we use the same notations for the smoothing operators. Moreover, we assume that the injection  $F_b \rightarrow F_a$  is compact when  $b > a$ .

**THEOREM 4.** *Let  $\beta$  be a real number such that*

$$5 < \beta < 6.$$

*Let  $V$  be a  $E_4$ -neighborhood of zero and  $\Phi$  a map from  $V$  to  $F_4$  which is twice differentiable and satisfies*

$$(5.6) \quad \|\Phi''(u; v, w)\|_6 \leq C \sum (1 + \|u\|_m) \|v\|_{m'} \|w\|_{m''},$$

where the sum is taken over the following values:

$m$	$m'$	$m''$
6	2	2
4	6	2
4	2	6

We assume that  $\Phi : E_4 \rightarrow F_4$  is continuous for every  $a \in [2, 8]$ . We assume that, for every  $v \in V \cap E_8$ ,  $\Phi'(v)$  has a right inverse  $\psi(v) : F_7 \rightarrow E_6$ , that  $(v, g) \mapsto \psi(v, g)$  is continuous from  $(V \cap E_6) \times F_7 \rightarrow E_6$ , and that there exists a constant  $C$  such that, for every  $(v, g) \in (V \cap E_6) \times F_7$ ,

$$(5.7) \quad \|\psi(v)g\|_2 \leq C\|g\|_3,$$

$$(5.8) \quad \|\psi(v)g\|_6 \leq C[\|g\|_7 + \|v\|_8\|g\|_3].$$

Then, for every  $f \in F_\beta$  with a sufficiently small norm, there exists  $u \in E_4$  such that  $\Phi(u) = \Phi(0) + f$ .

*Remark 2.* The differences between the previous statement and Hörmander's statement are the following ones:

- First, here we do not use the weak spaces  $E'_a$  defined by Hörmander, which simplifies the statement a little bit;
- then, here, the number of tame estimates to be proved is finite, which is more practical for the application of the theorem.

*Proof of Theorem 4.* Let  $g \in F_\beta$ . We have a decomposition

$$(5.9) \quad g = \sum_{j=0}^{\infty} \Delta_j g_j, \text{ with } \|g_j\|_b \leq K' \|g\|_\beta \theta_j^{b-\beta-1} \quad \forall b \in [2, 8],$$

where  $g_j := R_j g$  for every  $j \in \mathbb{N}$ . We claim that, when  $\|g\|_\beta$  is small enough, we can define a sequence  $(u_j)_{j \in \mathbb{N}}$  with  $u_0 = 0$  and the recursive formula

$$u_{j+1} := u_j + \Delta_j \dot{u}_j, \quad \dot{u}_j := \psi(v_j)g_j, \quad v_j := S_{\theta_j} u_j.$$

We also claim that there exist constants  $C_1, C_2, C_3, C_4$  such that, for every  $j \in \mathbb{N}^*$ ,

$$(5.10) \quad \|\dot{u}_j\|_a \leq C_1 \|g\|_\beta \theta_j^{a-\beta} \quad \forall a \in \{2, 4, 6\},$$

$$(5.11) \quad \|u_j\|_4 \leq C_2 \|g\|_\beta \text{ and } \|u_j\|_6 \leq C_2 \|g\|_\beta \theta_j^{7-\beta},$$

$$(5.12) \quad \|v_j\|_4 \leq C_3 \|g\|_\beta, \quad \|v_j\|_a \leq C_3 \|g\|_\beta \theta_j^{a-\beta+1} \quad \forall a \in \{6, 8\},$$

$$(5.13) \quad \|u_j - v_j\|_a \leq C_4 \|g\|_\beta \theta_j^{a-\beta+1} \quad \forall a \in \{2, 4, 6\}.$$

More precisely, we prove by induction on  $k \in \mathbb{N}$  the following property:

$$\begin{aligned} \mathcal{P}_k : u_j \text{ is well-defined for } j = 0, \dots, k+1, \\ (5.10) \text{ is satisfied for } j = 0, \dots, k, \\ (5.11), (5.12), \text{ and } (5.13) \text{ are satisfied for } j = 0, \dots, k+1. \end{aligned}$$

We introduce  $r > 0$  such that, for every  $u \in E_4$ ,  $\|u\|_\alpha < r$  implies  $u \in V$ .

Property  $\mathcal{P}_0$  is obvious. Let  $k \in \mathbb{N}^*$ . We assume that property  $\mathcal{P}_{k-1}$  is satisfied. Let us prove  $\mathcal{P}_k$ .

The vector  $u_{k+1}$  is well-defined if  $v_k \in V$ . By using (5.1) and (5.10), we get

$$\|v_k\|_4 \leq KC_1 \|g\|_\beta \sum_{j=0}^{k-1} \Delta_j \theta_j^{4-\beta}.$$

Since  $\beta < 5$ , the sum

$$S := \sum_{j=0}^{\infty} \Delta_j \theta_j^{4-\beta}$$

is convergent. Therefore, when

$$\|g\|_\beta \leq \frac{r}{C_1 K S},$$

$v_k \in V$  and  $u_{k+1}$  is well-defined.

Let us prove (5.10) for  $j = k$ . By using (5.7) and (5.9), we get

$$\|\dot{u}_k\|_2 \leq CK' \|g\|_\beta \theta_k^{2-\beta}.$$

By using (5.8), (5.12), and (5.9), we get

$$\|\dot{u}_k\|_6 \leq CK' \|g\|_\beta [\theta_k^{6-\beta} + C_3 \|g\|_\beta \theta_k^{9-\beta} \theta_k^{2-\beta}] \leq 2CK' \|g\|_\beta \theta_k^{6-\beta},$$

when  $\|g\|_\beta < 1/C_3$ . The convexity of the norm (5.5) provides

$$\|\dot{u}_k\|_4 \leq c\sqrt{2}CK' \|g\|_\beta \theta_k^{4-\beta}.$$

Therefore, we have (5.10) for  $j = k$ , when  $\|g\|_\beta < 1/C_3$  for  $C_1 = \max\{2, c\sqrt{2}\}CK'$ .

Let us prove (5.11) for  $j = k + 1$ . As noticed at the beginning of the proof by induction, we have

$$\|u_{k+1}\|_4 \leq C_1 \|g\|_\beta \sum_{j=0}^k \Delta_j \theta_j^{4-\beta} \leq C_1 \|g\|_\beta S.$$

Thanks to (5.10), we have

$$\|u_{k+1}\|_6 \leq C_1 \|g\|_\beta \sum_{j=0}^k \Delta_j \theta_j^{6-\beta} \leq C_1 \|g\|_\beta \frac{\theta_{k+1}^{7-\beta}}{7-\beta}.$$

Therefore, we have (5.11) for  $j = k + 1$ , with

$$C_2 := C_1 \max \left\{ S, \frac{1}{7-\beta} \right\}.$$

We get (5.12) for  $j = k + 1$  thanks to (5.1) and (5.2), with  $C_3 := KC_2$ . We get (5.13) for  $j = k + 1$  thanks to (5.11) and (5.12) for  $a = 6$  and thanks to (5.3) and (5.11) for  $a \in \{2, 4\}$ , with  $C_4 := \max\{C_2 + C_3; KC_2\}$ .

Inequality (5.10) proves that  $(u_k)$  converges in  $E_4$  toward

$$u := \sum_{j=0}^{\infty} \Delta_j \dot{u}_j.$$

The continuity of the map  $\Phi : E_4 \rightarrow F_4$  implies that  $\Phi(u_k)$  converges to  $\Phi(u)$  in  $F_4$ .

Let us study the limit of the sequence  $(\Phi(u_k))_{k \in \mathbb{N}}$  in a different way. We have

$$\Phi(u_{j+1}) - \Phi(u_j) = \Delta_j(e'_j + e''_j + g_j),$$

where

$$\begin{aligned} e'_j &:= \frac{1}{\Delta_j} (\Phi(u_j + \Delta_j \dot{u}_j) - \Phi(u_j) - \Phi'(u_j) \Delta_j \dot{u}_j) \\ &= \Delta_j \int_0^1 (1-t) \Phi''(u_j + t \Delta_j \dot{u}_j; \dot{u}_j, \dot{u}_j) dt, \\ e''_j &:= (\Phi'(u_j) - \Phi'(v_j)) \dot{u}_j = \int_0^1 \Phi''(v_j + t(u_j - v_j); u_j - v_j, \dot{u}_j) dt. \end{aligned}$$

Thanks to (5.6), we have

$$\begin{aligned} \|e'_j\|_6 &\leq C \sum (1 + \|u_j\|_m + \Delta_j \|\dot{u}_j\|_m) \|\dot{u}_j\|_{m'} \|\dot{u}_j\|_{m''} \\ &\leq C[(1 + (C_1 + C_2) \|g\|_\beta \theta_j^{7-\beta}) C_1^2 \|g\|_\beta^2 \theta_j^{4-2\beta} \\ &\quad + 2(1 + (C_1 + C_2) \|g\|_\beta) C_1^2 \|g\|_\beta^2 \theta_j^{8-2\beta}] \\ &\leq C \|g\|_\beta^2 \theta_j^{8-2\beta}, \\ \|e''_j\|_6 &\leq C \sum (1 + \|v_j\|_m + \|u_j - v_j\|_m) \|u_j - v_j\|_{m'} \|\dot{u}_j\|_{m''} \\ &\leq C[(1 + (C_3 + C_4) \|g\|_\beta \theta_j^{7-\beta}) C_1 C_4 \|g\|_\beta^2 \theta_j^{5-2\beta} \\ &\quad + 2(1 + (C_3 + C_4) \|g\|_\beta) C_1 C_4 \|g\|_\beta^2 \theta_j^{9-2\beta}] \\ &\leq C \|g\|_\beta^2 \theta_j^{9-2\beta}. \end{aligned}$$

Since  $9 - 2\beta < -1$ , then  $\sum \Delta_j(e'_j + e''_j)$  converges in  $F_6$  and

$$\left\| \sum_{j=0}^{\infty} \Delta_j(e'_j + e''_j) \right\|_6 \leq C \|g\|_\beta^2.$$

The uniqueness of the limit of the sequence  $(\Phi(u_k))_{k \in \mathbb{N}}$  gives the following equality in  $F_4$ :

$$\Phi(u) = g + \mathcal{T}(g),$$

where  $\mathcal{T}(g) \in F_6$  and

$$\|\mathcal{T}(g)\|_6 \leq C \|g\|_\beta^2.$$

We conclude by applying the Leray–Schauder fixed point theorem.  $\square$

*Remark 3.* The proof can also be done thanks to the Banach fixed point theorem, instead of the Leray–Schauder fixed point theorem, provided one adds new assumptions (see [3, Appendix C]). In this situation, one does not need any longer the compactness of the injections  $F_b \rightarrow F_a$  for  $b > a$ .

**6. Smoothing operators.** In this section, we build smoothing operators on the spaces  $E_a$  and  $F_b$  defined by (2.12) and (2.13). First, we smooth the functions in  $H_{(0)}^a((0, 1), \mathbb{R})$  for every integer  $a \in \{2, \dots, 8\}$ . Let  $s \in C^\infty(\mathbb{R}, \mathbb{R})$  be such that

$$s = 1 \text{ on } [0, 1], 0 \leq s \leq 1, s = 0 \text{ on } [2, +\infty).$$

We define

$$S_\theta u := \sum_{k=1}^{\infty} s\left(\frac{k}{\theta}\right) \langle u, \varphi_k \rangle \varphi_k.$$

The proof of the following proposition is easy.

**PROPOSITION 9.** *There exists a constant  $K > 0$  such that, for every integer  $a \in \{2, \dots, 8\}$ , for every  $u \in H_{(0)}^a((0, 1), \mathbb{R})$  and for every  $\theta \geq 1$ , we have*

$$\begin{aligned} \|S_\theta u\|_{H^b((0,1),\mathbb{R})} &\leq K \|u\|_{H^a((0,1),\mathbb{R})} \text{ for every } b \in \{2, \dots, a\}, \\ \|S_\theta u\|_{H^b((0,1),\mathbb{R})} &\leq K \theta^{b-a} \|u\|_{H^a((0,1),\mathbb{R})} \text{ for every } b \in \{a+1, \dots, 8\}, \\ \|u - S_\theta u\|_{H^b((0,1),\mathbb{R})} &\leq K \theta^{b-a} \|u\|_{H^a((0,1),\mathbb{R})} \text{ for every } b \in \{2, \dots, a-1\}, \\ \left\| \frac{d}{d\theta} S_\theta u \right\|_{H^b((0,1),\mathbb{R})} &\leq K \theta^{b-a-1} \|u\|_{H^a((0,1),\mathbb{R})} \text{ for every } b \in \{2, \dots, 8\}. \end{aligned}$$

Suitable smoothing operators for the control function  $p$  can be built with convolution products and truncations with  $C^\infty$ -function having compact support as in [3, sect. 3.3.2]. This construction is inspired by [19].

**7. Bound on  $\Phi_T''$ .** The aim of this section is the proof of inequality (5.6) on the map  $\Phi_T$  defined by (1.8). More precisely, we prove the following proposition.

**PROPOSITION 10.** *Let  $T > 0$ . The map  $\Phi_T : E_6 \mapsto F_6$  is twice differentiable, and, for every  $(u_0, \dot{u}_0, p), (\lambda_0, \dot{\lambda}_0, \rho), (\mu_0, \dot{\mu}_0, \theta) \in E_6$ ,*

$$\Phi_T''(u_0, \dot{u}_0, p) \cdot ((\lambda_0, \dot{\lambda}_0, \rho), (\mu_0, \dot{\mu}_0, \theta)) = (0, 0, h(T), h_t(T)),$$

where

$$\begin{cases} u_{tt} + u_{xxxx} + p(t)u_{xx} = 0, \\ u = u_x = 0 \text{ at } x = 0, 1, \\ u(0) = u_0, \dot{u}(0) = \dot{u}_0, \end{cases} \quad \begin{cases} \lambda_{tt} + \lambda_{xxxx} + p(t)\lambda_{xx} + \rho(t)u_{xx} = 0, \\ \lambda = \lambda_x = 0 \text{ at } x = 0, 1, \\ \lambda(0) = \lambda_0, \dot{\lambda}(0) = \dot{\lambda}_0, \end{cases} \\ \begin{cases} \mu_{tt} + \mu_{xxxx} + p(t)\mu_{xx} + \theta(t)u_{xx} = 0, \\ \mu = \mu_x = 0 \text{ at } x = 0, 1, \\ \mu(0) = \mu_0, \dot{\mu}(0) = \dot{\mu}_0, \end{cases} \\ \begin{cases} h_{tt} + h_{xxxx} + p(t)h_{xx} + \rho(t)\mu_{xx} + \theta(t)\lambda_{xx} = 0, \\ h = h_x = 0 \text{ at } x = 0, 1, \\ h(0) = 0, \dot{h}(0) = 0. \end{cases} \end{cases}$$

For every  $r > 0$ , there exists a constant  $C(r) > 0$  such that, for every  $(u_0, \dot{u}_0, p), (\lambda_0, \dot{\lambda}_0, \rho), (\mu_0, \dot{\mu}_0, \theta) \in E_6$ , with  $\|(u_0, \dot{u}_0, p)\|_4 \leq r$ ,

$$(7.1) \quad \begin{aligned} &\|\Phi_T''(u_0, \dot{u}_0, p) \cdot ((\lambda_0, \dot{\lambda}_0, \rho), (\mu_0, \dot{\mu}_0, \theta))\|_6 \\ &\leq C(r) \sum (1 + \|(u_0, \dot{u}_0, p)\|_m) \|(\lambda_0, \dot{\lambda}_0, \rho)\|_{m'} \|(\mu_0, \dot{\mu}_0, \theta)\|_{m''} \end{aligned}$$

where the sum is taken over the following values:

$m$	$m'$	$m''$
6	2	2
4	6	2
4	2	6

*Proof of Proposition 10.* The regularity  $C^2$  of the map  $\Phi_T$  can be proved thanks to Propositions 1, 2, and 3 in a very similar way as for the regularity  $C^1$  in Proposition 5. Here we only justify (7.1). By applying Proposition 3 (which is possible because  $\|p\|_{W^{1,1}((0,T),\mathbb{R})} \leq \sqrt{T}\|p\|_{H^1((0,T),\mathbb{R})} \leq \sqrt{T}r$ ) we get  $C = C(r) > 0$  such that

$$(7.2) \quad \begin{aligned} & \| (h, h_t) \|_{C^0([0,T], H^6 \times H^4)} \\ & \leq C \left[ \|f\|_{W^{2,1}((0,T), L^2)} + \|p\|_{W^{2,1}} \|f\|_{L^1((0,T), L^2)} + \|f\|_{C^0([0,T], H^2)} \right], \end{aligned}$$

where  $f(t) := \rho(t)\mu_{xx}(t) + \theta(t)\lambda_{xx}(t)$ . Let us define  $f_1(t) := \rho(t)\mu_{xx}(t)$  and  $f_2(t) := \theta(t)\lambda_{xx}(t)$ . We work only on  $f_1$ , and we deduce the same results for  $f_2$  just by exchanging  $(\mu, \rho)$  and  $(\lambda, \theta)$ . By using Propositions 1, 2, and 3, we get constants  $C = C(r)$  such that, when  $\|(u_0, \dot{u}_0, p)\|_{E_4} \leq r$ ,

$$\begin{aligned} \|f_1\|_{L^1((0,T), L^2)} & \leq C \|\rho\|_{L^1} \|\mu\|_{C^0((0,T), H^2)}, \\ \|f_1\|_{W^{2,1}((0,T), L^2)} & \leq C [\|\rho\|_{H^2} \|\mu\|_{C^0([0,T], H^2)} + \|\rho\|_{H^1} \|\mu\|_{C^1([0,T], H^2)} \\ & \quad + \|\rho\|_{L^2} \|\mu\|_{C^2([0,T], H^2)}], \\ \|f_1\|_{C^0([0,T], H^2)} & \leq C \|\rho\|_{H^1} \|\mu\|_{C^0([0,T], H^4)}, \end{aligned}$$

where

$$\begin{aligned} \|\mu\|_{C^0([0,T], H^2)} & \leq C [\|(\mu_0, \dot{\mu}_0)\|_{H_0^2 \times L^2} + \|\theta u_{xx}\|_{L^1((0,T), L^2)}] \\ & \leq C [\|(\mu_0, \dot{\mu}_0)\|_{H_0^2 \times L^2} + \|\theta\|_{L^2} \|(u_0, \dot{u}_0)\|_{H_0^2 \times L^2}] \\ & \leq C \|(\mu_0, \dot{\mu}_0, \theta)\|_{E_2}, \\ \|\mu\|_{C^0([0,T], H^4)}, \|\mu\|_{C^1([0,T], H^2)} & \leq C [\|(\mu_0, \dot{\mu}_0)\|_{H_{(0)}^4 \times H_0^2} + \|\theta u_{xx}\|_{W^{1,1}((0,T), L^2)}] \\ & \leq C [\|(\mu_0, \dot{\mu}_0)\|_{H_{(0)}^4 \times H_0^2} + \|\theta\|_{H^1} \|(u_0, \dot{u}_0)\|_{H_0^2 \times L^2} \\ & \quad + \|\theta\|_{L^1} \|(u_0, \dot{u}_0)\|_{H_{(0)}^4 \times H_0^2}] \\ & \leq C \|(\mu_0, \dot{\mu}_0, \theta)\|_{E_4}, \\ \|\mu\|_{C^2([0,T], H^2)} & \leq C [\|(\mu_0, \dot{\mu}_0)\|_{H_{(0)}^6 \times H_{(0)}^4} + \|p\|_{W^{2,1}} \|(\mu_0, \dot{\mu}_0)\|_{H_0^2 \times L^2} \\ & \quad + \|\theta u_{xx}\|_{W^{2,1}((0,T), L^2)} + \|p\|_{W^{2,1}} \|\theta u_{xx}\|_{L^1((0,T), L^2)} \\ & \quad + \|\theta u_{xx}\|_{C^0([0,T], H^2)}]. \end{aligned}$$

We have

$$\begin{aligned} \|\theta u_{xx}\|_{W^{2,1}((0,T), L^2)} & \leq C [\|\theta\|_{H^2} \|u\|_{C^0([0,T], H^2)} + \|\theta\|_{H^1} \|u\|_{C^1([0,T], H^2)} \\ & \quad + \|\theta\|_{L^2} \|u\|_{C^2([0,T], H^2)}] \\ & \leq C \{ \|\theta\|_{H^2} \|(u_0, \dot{u}_0)\|_{H_0^2 \times L^2} + \|\theta\|_{H^1} \|(u_0, \dot{u}_0)\|_{H_{(0)}^4 \times H_0^2} \\ & \quad + \|\theta\|_{L^2} [\|(u_0, \dot{u}_0)\|_{H_{(0)}^6 \times H_{(0)}^4} + \|p\|_{W^{2,1}} \|(u_0, \dot{u}_0)\|_{H_0^2 \times L^2}] \} \\ & \leq C [\|\theta\|_{H^2} + \|\theta\|_{L^2} \|(u_0, \dot{u}_0, p)\|_{E_6}], \\ \|\theta u_{xx}\|_{L^1((0,T), L^2)} & \leq C \|\theta\|_{L^2} \|(u_0, \dot{u}_0)\|_{H_0^2 \times L^2} \\ & \leq C \|\theta\|_{L^2}, \\ \|\theta u_{xx}\|_{C^0([0,T], H^2)} & \leq C \|\theta\|_{H^1} \|(u_0, \dot{u}_0)\|_{H_{(0)}^4 \times H_0^2} \\ & \leq C \|\theta\|_{H^1}, \end{aligned}$$

and thus

$$\|\mu\|_{C^2([0,T], H^2)} \leq C [\|(\mu_0, \dot{\mu}_0, \theta)\|_{E_6} + \|(u_0, \dot{u}_0, p)\|_{E_6} \|(\mu_0, \dot{\mu}_0, \theta)\|_{E_2}].$$

We deduce from the previous computations that

$$\begin{aligned}
 \|f_1\|_{L^1((0,T),L^2)} &\leq C\|\rho\|_{L^2}\|(\mu_0, \dot{\mu}_0, \theta)\|_{E_2}, \\
 \|f_1\|_{W^{2,1}((0,T),L^2)} &\leq C\{\|\rho\|_{H^2}\|(\mu_0, \dot{\mu}_0, \theta)\|_{E_2} + \|\rho\|_{H^1}\|(\mu_0, \dot{\mu}_0, \theta)\|_{E_4} \\
 (7.3) \quad &\quad + \|\rho\|_{L^2}\|(\mu_0, \dot{\mu}_0, \theta)\|_{E_6} \\
 &\quad + \|\rho\|_{L^2}\|(u_0, \dot{u}_0, p)\|_{E_6}\|(\mu_0, \dot{\mu}_0, \theta)\|_{E_2}\}, \\
 \|f_1\|_{C^0([0,T],H^2)} &\leq C\|\rho\|_{H^1}\|(\mu_0, \dot{\mu}_0, \theta)\|_{E_4}.
 \end{aligned}$$

The inequalities (7.2) and (7.3) give the conclusion.  $\square$

**8. Controllability of the linearized system around  $(u^{ref}, \dot{u}^{ref}, p \equiv 0)$  with Nash–Moser bounds.** In all of this section,  $T := 8/\pi$ . We recall that the spaces  $E_a$  and  $F_b$  are defined by (2.12) and (2.13). The goal of this section is the proof of the following result.

**PROPOSITION 11.** *There exists  $C > 0$  such that the map  $d\Phi_T(u_0^{ref}, \dot{u}_0^{ref}, 0)$  has a right inverse*

$$d\Phi_T(u_0^{ref}, \dot{u}_0^{ref}, 0)^{-1} : F_7 \rightarrow E_6$$

and, for every  $(U_0, \dot{U}_0, U_T, \dot{U}_T) \in F_7$ ,

$$\begin{aligned}
 \|d\Phi_T(u_0^{ref}, \dot{u}_0^{ref}, 0)^{-1} \cdot (U_0, \dot{U}_0, U_T, \dot{U}_T)\|_{E_2} &\leq C\|(U_0, \dot{U}_0, U_T, \dot{U}_T)\|_{F_3}, \\
 \|d\Phi_T(u_0^{ref}, \dot{u}_0^{ref}, 0)^{-1} \cdot (U_0, \dot{U}_0, U_T, \dot{U}_T)\|_{E_6} &\leq C\|(U_0, \dot{U}_0, U_T, \dot{U}_T)\|_{F_7}.
 \end{aligned}$$

We introduce the linear map  $M$ , defined on  $L^2((0,T), \mathbb{R})$  by  $M(P) = (M(P)_k)_{k \in \mathbb{N}^*}$ ,

$$\begin{aligned}
 M(P)_{2k-1} &:= \frac{1}{T} \int_0^T P(t) \sin(\sqrt{\lambda_3}t) e^{-i\sqrt{\lambda_{2k-1}}t} dt \quad \forall k \in \mathbb{N}^*, \\
 M(P)_{2k} &:= \frac{1}{T} \int_0^T P(t) \sin(\sqrt{\lambda_2}t) e^{-i\sqrt{\lambda_{2k}}t} dt \quad \forall k \in \mathbb{N}^*.
 \end{aligned}$$

For  $s > 0$ , we introduce the space

$$h^s(\mathbb{N}^*, \mathbb{C}) := \left\{ d = (d_n)_{n \in \mathbb{N}^*}; \sum_{n=1}^{\infty} |n^s d_n|^2 < +\infty \right\}.$$

Thanks to the equivalence between the controllability of the linearized system  $(\Sigma_l^{ref})$  and the solvability of the moment problem (3.24) (which is explained in the proof of statement (1) of Theorem 3, in section 3.2), Proposition 11 is a consequence of the following result.

**PROPOSITION 12.** *There exists  $C > 0$  such that the map  $M$  has a right inverse*

$$M^{-1} : h^4(\mathbb{N}^*, \mathbb{C}) \rightarrow H_0^2((0,T), \mathbb{R})$$

and, for every  $d \in h^4(\mathbb{N}^*, \mathbb{C})$ ,

$$\|M^{-1}(d)\|_{L^2((0,T), \mathbb{R})} \leq C\|d\|_{l^2(\mathbb{N}^*, \mathbb{C})},$$

$$\|M^{-1}(d)\|_{H_0^2((0,T), \mathbb{R})} \leq C\|d\|_{h^4(\mathbb{N}^*, \mathbb{C})}.$$

Proposition 11 is a consequence of Proposition 12, but they are not equivalent. Indeed, Proposition 12 provides a right inverse  $d\Phi_T(u_0, \dot{u}_0, 0)^{-1}$  defined on  $F_7$  with values in  $H_{(0)}^7((0, 1), \mathbb{R}) \times H_{(0)}^5((0, 1), \mathbb{R}) \times H_0^2((0, T), \mathbb{R})$  which is strictly smaller than  $E_6$  (see (2.12) for a definition of  $E_6$ ).

We will prove Proposition 12 by using an auxiliary linear map  $\widetilde{M}$ , which is easier to deal with than  $M$  and close enough to  $M$  so that the surjectivity of  $\widetilde{M}$  implies the surjectivity of  $M$ . Let us introduce the map  $\widetilde{M} : L^2((0, T), \mathbb{R}) \rightarrow l^2(\mathbb{N}^*, \mathbb{C})$  defined by

$$(8.1) \quad \begin{aligned} \widetilde{M}(P)_{2k-1} &:= \frac{1}{T} \int_0^T P(t) \sin\left(\frac{\pi^2}{4} K_3 t\right) e^{-i \frac{\pi^2}{4} K_{2k-1} t} dt \quad \forall k \in \mathbb{N}^*, \\ \widetilde{M}(P)_{2k} &:= \frac{1}{T} \int_0^T P(t) \sin\left(\frac{\pi^2}{4} K_2 t\right) e^{-i \frac{\pi^2}{4} K_{2k} t} dt \quad \forall k \in \mathbb{N}^*, \end{aligned}$$

where  $(K_n)_{n \in \mathbb{N}^*}$  is defined by (3.6). The linear map  $\widetilde{M}$  maps  $L^2((0, T), \mathbb{R})$  into  $l^2(\mathbb{N}^*, \mathbb{C})$ ,  $H_0^1((0, T), \mathbb{R})$  into  $h^2(\mathbb{N}^*, \mathbb{C})$ , and  $H_0^2((0, T), \mathbb{R})$  into  $h^4(\mathbb{N}^*, \mathbb{C})$ . Indeed, each term of the sequence is the sum of two Fourier coefficients of  $P$ . Note that  $T = 8/\pi$  is chosen so that, for every  $n \in \mathbb{N}^*$ ,  $e^{i \frac{\pi^2}{4} K_n t}$  is  $T$ -periodic. Thus, the previous statement is a consequence of Bessel Parseval equality and integrations by parts.

For technical reasons, the space  $h^s(\mathbb{N}^*, \mathbb{C})$  is equipped with the unusual norm

$$\|d\|_{h^s(\mathbb{N}^*, \mathbb{C})} := \left( \sum_{n=1}^{\infty} |K_n^{s/2} d_n|^2 \right)^{1/2}.$$

On the spaces  $L^2((0, T), \mathbb{R})$ ,  $H_0^1((0, T), \mathbb{R})$ , and  $H_0^2((0, T), \mathbb{R})$  we use

$$\begin{aligned} \|f\|_{L^2((0, T), \mathbb{R})} &:= \left( \frac{1}{T} \int_0^T |f(t)|^2 dt \right)^{1/2}, \\ \|f\|_{H_0^1((0, T), \mathbb{R})} &:= \|f'\|_{L^2((0, T), \mathbb{R})}, \quad \|f\|_{H_0^2((0, T), \mathbb{R})} := \|f''\|_{L^2((0, T), \mathbb{R})}. \end{aligned}$$

More precisely, we prove Proposition 12 by applying the next proposition to  $\mathcal{M} = M$  and  $\widetilde{\mathcal{M}} = \widetilde{M}$ .

**PROPOSITION 13.** *Let  $\mathcal{M}$  and  $\widetilde{\mathcal{M}}$  be continuous linear maps from  $L^2((0, T), \mathbb{R})$  to  $l^2(\mathbb{N}^*, \mathbb{C})$ , from  $H_0^1((0, T), \mathbb{R})$  to  $h^2(\mathbb{N}^*, \mathbb{C})$ , and from  $H_0^2((0, T), \mathbb{R})$  to  $h^4(\mathbb{N}^*, \mathbb{C})$ . We assume that there exist positive constants  $C_0, C_1, C_2, C$  such that  $\widetilde{\mathcal{M}}$  has a right inverse*

$$\widetilde{\mathcal{M}}^{-1} : h^4(\mathbb{N}^*, \mathbb{C}) \rightarrow H_0^2((0, T), \mathbb{R})$$

which satisfies, for every  $d \in h^4(\mathbb{N}^*, \mathbb{C})$ ,

$$\begin{aligned} \|\widetilde{\mathcal{M}}^{-1}(d)\|_{L^2((0, T), \mathbb{R})} &\leq C_0 \|d\|_{l^2(\mathbb{N}^*, \mathbb{C})}, \\ \|\widetilde{\mathcal{M}}^{-1}(d)\|_{H_0^1((0, T), \mathbb{R})} &\leq C_1 \|d\|_{h^2(\mathbb{N}^*, \mathbb{C})} + C \|d\|_{l^2(\mathbb{N}^*, \mathbb{C})}, \\ \|\widetilde{\mathcal{M}}^{-1}(d)\|_{H_0^2((0, T), \mathbb{R})} &\leq C_2 \|d\|_{h^4(\mathbb{N}^*, \mathbb{C})} + C \|d\|_{h^2(\mathbb{N}^*, \mathbb{C})}. \end{aligned}$$

We also assume that there exist positive constants  $C_0, C_1, C_2, C_3, C_4, C_5$  such that, for



every  $P \in H_0^2((0, T), \mathbb{R})$ ,

$$\|(\widetilde{\mathcal{M}} - \mathcal{M})(P)\|_{l^2(\mathbb{N}^*, \mathbb{C})} \leq \mathcal{C}_0 \|P\|_{L^2((0, T), \mathbb{R})},$$

$$\|(\widetilde{\mathcal{M}} - \mathcal{M})(P)\|_{h^2(\mathbb{N}^*, \mathbb{C})} \leq \mathcal{C}_1 \|P\|_{H_0^1((0, T), \mathbb{R})} + \mathcal{C}_3 \|P\|_{L^2((0, T), \mathbb{R})},$$

$$\|(\widetilde{\mathcal{M}} - \mathcal{M})(P)\|_{h^4(\mathbb{N}^*, \mathbb{C})} \leq \mathcal{C}_2 \|P\|_{H_0^2((0, T), \mathbb{R})} + \mathcal{C}_4 \|P\|_{H_0^1((0, T), \mathbb{R})} + \mathcal{C}_5 \|P\|_{L^2((0, T), \mathbb{R})}.$$

We assume that  $C_0\mathcal{C}_0$ ,  $C_1\mathcal{C}_1$ , and  $C_2\mathcal{C}_2$  are  $< 1$ . Then there exists a constant  $m > 0$  such that  $\mathcal{M}$  has a right inverse

$$\mathcal{M}^{-1} : h^4(\mathbb{N}^*, \mathbb{C}) \rightarrow H_0^2((0, T), \mathbb{R})$$

which satisfies, for every  $d \in h^4(\mathbb{N}^*, \mathbb{C})$ ,

$$\begin{aligned} \|\mathcal{M}^{-1}(d)\|_{L^2((0, T), \mathbb{R})} &\leq m \|d\|_{l^2(\mathbb{N}^*, \mathbb{C})}, \\ \|\mathcal{M}^{-1}(d)\|_{H_0^1((0, T), \mathbb{R})} &\leq m \|d\|_{h^2(\mathbb{N}^*, \mathbb{C})}, \\ \|\mathcal{M}^{-1}(d)\|_{H_0^2((0, T), \mathbb{R})} &\leq m \|d\|_{h^4(\mathbb{N}^*, \mathbb{C})}. \end{aligned}$$

*Remark 4.* One of the interests of this proposition, when we apply it, is that the constants  $C$  and  $\mathcal{C}_j$  for  $j = 3, 4, 5$  can be large. In the application of this proposition to the maps  $M$  and  $\widetilde{M}$ , we will put all of the possible terms in the factor with  $C$  or  $\mathcal{C}_j$  for  $j = 3, 4, 5$  in order to reduce  $\mathcal{C}_1$  and  $\mathcal{C}_2$  and ensure  $C_1\mathcal{C}_1 < 1$  and  $C_2\mathcal{C}_2 < 1$ .

*Proof of Proposition 13.* We introduce  $\Delta_i := C_i\mathcal{C}_i$  for  $i = 0, 1, 2$ . Let  $d \in h^4(\mathbb{N}^*, \mathbb{C})$ . We define by induction the sequence  $(w_n)_{n \in \mathbb{N}}$  in  $H_0^2((0, T), \mathbb{R})$  by

$$\begin{cases} w_0 := \widetilde{\mathcal{M}}^{-1}(d), \\ w_{n+1} := \widetilde{\mathcal{M}}^{-1}[(\widetilde{\mathcal{M}} - \mathcal{M})(w_n)]. \end{cases}$$

Then we have, for every  $n \in \mathbb{N}^*$ ,

$$\|w_n\|_{L^2((0, T), \mathbb{R})} \leq C_0 \Delta_0^n \|d\|_{l^2(\mathbb{N}^*, \mathbb{C})},$$

$$\|w_n\|_{H_0^1((0, T), \mathbb{R})} \leq C_1 \Delta_1^n \|d\|_{h^2(\mathbb{N}^*, \mathbb{C})} + x_n \|d\|_{l^2(\mathbb{N}^*, \mathbb{C})},$$

$$\|w_n\|_{H_0^2((0, T), \mathbb{R})} \leq C_2 \Delta_2^n \|d\|_{h^4(\mathbb{N}^*, \mathbb{C})} + y_n \|d\|_{h^2(\mathbb{N}^*, \mathbb{C})} + z_n \|d\|_{l^2(\mathbb{N}^*, \mathbb{C})},$$

where  $x_0 = y_0 = C$ ,  $z_0 = 0$ , and, for every  $n \in \mathbb{N}^*$ ,

$$x_{n+1} = \Delta_1 x_n + XC_0 \Delta_0^n, \quad y_{n+1} = \Delta_2 y_n + YC_1 \Delta_1^n, \quad z_{n+1} = \Delta_2 z_n + Yx_n + ZC_0 \Delta_0^n,$$

$$X := CC_0 + C_1\mathcal{C}_3, \quad Y := C_2\mathcal{C}_4 + CC_1, \quad Z := C_2\mathcal{C}_5 + C\mathcal{C}_3.$$

Under the assumption that  $\Delta_i < 1$  for  $i = 0, 1, 2$ , the sums  $\sum x_n$ ,  $\sum y_n$ , and  $\sum z_n$  converge and

$$\begin{aligned} (1 - \Delta_1) \sum_{n=0}^{\infty} x_n &= C + \frac{XC_0}{1 - \Delta_0}, & (1 - \Delta_2) \sum_{n=0}^{\infty} y_n &= C + \frac{YC_1}{1 - \Delta_1}, \\ (1 - \Delta_2) \sum_{n=0}^{\infty} z_n &= Y \sum_{n=0}^{\infty} x_n + \frac{ZC_0}{1 - \Delta_0}. \end{aligned}$$

Thus  $\sum w_n$  converges in  $H_0^2((0, T), \mathbb{R})$  to a function  $w$  which satisfies

$$\begin{aligned} \|w\|_{L^2((0, T), \mathbb{R})} &\leq \frac{C_0}{1 - \Delta_0} \|d\|_{l^2}, \\ \|w\|_{H_0^1((0, T), \mathbb{R})} &\leq \frac{C_1}{1 - \Delta_1} \|d\|_{h^2} + \frac{1}{1 - \Delta_1} \left( C + \frac{XC_0}{1 - \Delta_0} \right), \\ \|w\|_{H_0^2((0, T), \mathbb{R})} &\leq \frac{C_2}{1 - \Delta_2} \|d\|_{h^4} + \frac{1}{1 - \Delta_1} \left( C + \frac{YC_1}{1 - \Delta_1} \right) \|d\|_{h^2} \\ &\quad + \frac{1}{1 - \Delta_2} \left[ \frac{Y}{1 - \Delta_1} \left( C + \frac{XC_0}{1 - \Delta_0} \right) + \frac{ZC_0}{1 - \Delta_0} \right] \|d\|_{l^2}. \end{aligned}$$

For every  $n \in \mathbb{N}^*$ , we have

$$\mathcal{M} \left( \sum_{k=0}^n w_k \right) = d + (\mathcal{M} - \widetilde{\mathcal{M}})(w_n),$$

where  $(\mathcal{M} - \widetilde{\mathcal{M}})(w_n) \rightarrow 0$  in  $l^2$  because  $w_n \rightarrow 0$  in  $L^2$  and  $(\mathcal{M} - \widetilde{\mathcal{M}})$  is continuous from  $L^2$  to  $l^2$ , and thus  $\mathcal{M}(w) = d$ .  $\square$

In the next proposition, we check the first assumption of Proposition 13 with  $\widetilde{\mathcal{M}} = \widetilde{M}$ .

**PROPOSITION 14.** *The linear map  $\widetilde{M}$  defined by (8.1) has a right inverse  $\widetilde{M}^{-1} : h^4(\mathbb{N}^*, \mathbb{C}) \rightarrow H_0^2((0, T), \mathbb{R})$  such that, for every  $d \in h^4(\mathbb{N}^*, \mathbb{C})$ ,*

$$(8.2) \quad \|\widetilde{M}^{-1}(d)\|_{L^2((0, T), \mathbb{R})} \leq C_0 \|d\|_{l^2(\mathbb{N}^*, \mathbb{C})},$$

$$(8.3) \quad \|\widetilde{M}^{-1}(d)\|_{H_0^1((0, T), \mathbb{R})} \leq C_1 \|d\|_{h^2(\mathbb{N}^*, \mathbb{C})} + C \|d\|_{l^2(\mathbb{N}^*, \mathbb{C})},$$

$$(8.4) \quad \|\widetilde{M}^{-1}(d)\|_{H_0^2((0, T), \mathbb{R})} \leq C_2 \|d\|_{h^4(\mathbb{N}^*, \mathbb{C})} + C \|d\|_{h^2(\mathbb{N}^*, \mathbb{C})},$$

where  $C_0 := 2\sqrt{3}$ ,  $C_1 := \frac{\sqrt{3}}{2\sqrt{2}}\pi^2$ ,  $C_2 := \frac{\sqrt{3}}{8\sqrt{2}}\pi^4$ , and  $C$  is a positive constant.

*Proof of Proposition 14.* Let  $d \in h^4(\mathbb{N}^*, \mathbb{C})$ . For a function  $P \in H_0^2((0, T), \mathbb{R})$ , the equality  $\widetilde{M}(P) = d$  is satisfied in particular when

$$(8.5) \quad \begin{cases} \frac{1}{T} \int_0^T P(t) dt = 0, \\ \frac{1}{T} \int_0^T P(t) e^{-i\frac{\pi^2}{4} 2K_3 t} dt = -2id_3, \\ \frac{1}{T} \int_0^T P(t) e^{-i\frac{\pi^2}{4} (K_{2k-1} - K_3) t} dt = id_{2k-1} \quad \forall k \in \mathbb{N}^* \text{ such that } k \neq 2, \\ \frac{1}{T} \int_0^T P(t) e^{-i\frac{\pi^2}{4} (K_{2k-1} + K_3) t} dt = -id_{2k-1} \quad \forall k \in \mathbb{N}^* \text{ such that } k \neq 2, \\ \frac{1}{T} \int_0^T P(t) e^{-i\frac{\pi^2}{4} 2K_2 t} dt = -2id_2, \\ \frac{1}{T} \int_0^T P(t) e^{-i\frac{\pi^2}{4} (K_{2k} - K_2) t} dt = id_{2k} \quad \forall k \in \mathbb{N}^* \text{ such that } k \neq 2, \\ \frac{1}{T} \int_0^T P(t) e^{-i\frac{\pi^2}{4} (K_{2k} + K_2) t} dt = -id_{2k} \quad \forall k \in \mathbb{N}^* \text{ such that } k \neq 2. \end{cases}$$

Let us consider the candidate

$$\begin{aligned} P(t) := 2\Re &\left[ -2id_3 e^{i\frac{\pi^2}{4} 2K_3 t} - 2id_2 e^{i\frac{\pi^2}{4} 2K_2 t} \right. \\ &+ \sum_{k=1, k \neq 2}^{\infty} id_{2k-1} \left( e^{i\frac{\pi^2}{4} (K_{2k-1} - K_3) t} - e^{i\frac{\pi^2}{4} (K_{2k-1} + K_3) t} \right) \\ &\left. + \sum_{k=2}^{\infty} id_{2k} \left( e^{i\frac{\pi^2}{4} (K_{2k} - K_2) t} - e^{i\frac{\pi^2}{4} (K_{2k} + K_2) t} \right) \right] \frac{1}{2} (1 - e^{i\frac{\pi^2}{4} t}) (1 - e^{-i\frac{\pi^2}{4} t}). \end{aligned}$$

Then  $P \in H_0^2((0, T), \mathbb{R})$  because  $d \in h^4(\mathbb{N}^*, \mathbb{C})$  and  $t \mapsto (1 - e^{i\pi^2 t/4})(1 - e^{-i\pi^2 t/4})$  belongs to  $H_0^2((0, T), \mathbb{R})$ . The positive integers

$$2K_3, 2K_3 \pm 1, 2K_2, 2K_2 \pm 1, K_3 \pm K_1, K_3 \pm K_1 \pm 1, K_4 \pm K_2, K_4 \pm K_2 \pm 1, \\ K_{2k-1} \pm K_3, K_{2k-1} \pm K_3 \pm 1, K_{2k} \pm K_2, K_{2k} \pm K_2 \pm 1, \text{ for } k \in \mathbb{N} \text{ such that } k \geq 3,$$

are all different except  $K_3 + K_1 - 1 = K_4 - K_2 + 1$  (which concerns the coefficients  $d_1$  and  $d_4$  in  $P$ ) and  $K_8 + K_2 - 1 = K_9 - K_3 + 1$  (which concerns the coefficients  $d_8$  and  $d_9$  in  $P$ ). This statement can be proved in the same way as the claim in the proof of Lemma 1. Thanks to the orthogonality of the different terms in  $P$ , the function  $P$  solves the moment problem (8.5). We have

$$\begin{aligned} & \frac{2id_3 e^{i\frac{\pi^2}{4} 2K_3 t} (1 - e^{i\frac{\pi^2}{4} t})(1 - e^{-i\frac{\pi^2}{4} t})}{2} \\ &= 2id_3 \left[ -\frac{1}{2} e^{i\frac{\pi^2}{4} (2K_3+1)t} + e^{i\frac{\pi^2}{4} 2K_3 t} - \frac{1}{2} e^{i\frac{\pi^2}{4} (2K_3-1)t} \right], \end{aligned}$$

where all of the terms are orthogonal, and thus

$$\begin{aligned} \left\| 2\Re[2id_3 e^{i\frac{\pi^2}{4} 2K_3 t}] \frac{(1 - e^{i\frac{\pi^2}{4} t})(1 - e^{-i\frac{\pi^2}{4} t})}{2} \right\|_{L^2((0, T), \mathbb{R})}^2 &= 4|d_3|^2 \left( \frac{1}{4} + 1 + \frac{1}{4} \right) * 2 \\ &= 12|d_3|^2. \end{aligned}$$

For the same reason, we have

$$\left\| 2\Re[2id_2 e^{i\frac{\pi^2}{4} 2K_2 t}] \frac{(1 - e^{i\frac{\pi^2}{4} t})(1 - e^{-i\frac{\pi^2}{4} t})}{2} \right\|_{L^2((0, T), \mathbb{R})}^2 = 12|d_2|^2.$$

We have

$$\begin{aligned} & id_{2k} \left( e^{i\frac{\pi^2}{4} (K_{2k}-K_2)t} - e^{i\frac{\pi^2}{4} (K_{2k}+K_2)t} \right) \frac{(1 - e^{i\frac{\pi^2}{4} t})(1 - e^{-i\frac{\pi^2}{4} t})}{2} \\ &= id_{2k} \left[ -\frac{1}{2} e^{i\frac{\pi^2}{4} (K_{2k}-K_2-1)t} + e^{i\frac{\pi^2}{4} (K_{2k}-K_2)t} - \frac{1}{2} e^{i\frac{\pi^2}{4} (K_{2k}-K_2+1)t} \right] \\ &\quad - id_{2k} \left[ -\frac{1}{2} e^{i\frac{\pi^2}{4} (K_{2k}+K_2-1)t} + e^{i\frac{\pi^2}{4} (K_{2k}+K_2)t} - \frac{1}{2} e^{i\frac{\pi^2}{4} (K_{2k}+K_2+1)t} \right], \end{aligned}$$

where the terms are orthogonal, and thus

$$\begin{aligned} & \left\| 2\Re \left[ id_{2k} \left( e^{i\frac{\pi^2}{4} (K_{2k}-K_2)t} - e^{i\frac{\pi^2}{4} (K_{2k}+K_2)t} \right) \right] \frac{(1 - e^{i\frac{\pi^2}{4} t})(1 - e^{-i\frac{\pi^2}{4} t})}{2} \right\|_{L^2((0, T), \mathbb{R})}^2 \\ &= |d_{2k}|^2 \left( \frac{1}{4} + 1 + \frac{1}{4} \right) * 4 = 6|d_{2k}|^2. \end{aligned}$$

For the same reason, we have

$$\begin{aligned} & \left\| 2\Re \left[ id_{2k-1} \left( e^{i\frac{\pi^2}{4} (K_{2k-1}-K_3)t} - e^{i\frac{\pi^2}{4} (K_{2k-1}+K_3)t} \right) \right] \frac{(1 - e^{i\frac{\pi^2}{4} t})(1 - e^{-i\frac{\pi^2}{4} t})}{2} \right\|_{L^2((0, T), \mathbb{R})}^2 \\ &= 6|d_{2k-1}|^2. \end{aligned}$$

Since  $K_8 + K_2 - 1 = K_9 - K_3 + 1$ , we have

$$\begin{aligned}
 & \left[ id_8 \left( e^{i\frac{\pi^2}{4}(K_8-K_2)t} - e^{i\frac{\pi^2}{4}(K_8+K_2)t} \right) \right. \\
 & \quad \left. + id_9 \left( e^{i\frac{\pi^2}{4}(K_9-K_3)t} - e^{i\frac{\pi^2}{4}(K_9+K_3)t} \right) \right] \frac{(1 - e^{i\frac{\pi^2}{4}t})(1 - e^{-i\frac{\pi^2}{4}t})}{2} \\
 &= id_8 \left[ -\frac{1}{2} e^{i\frac{\pi^2}{4}(K_8-K_2-1)t} + e^{i\frac{\pi^2}{4}(K_8-K_2)t} - \frac{1}{2} e^{i\frac{\pi^2}{4}(K_8-K_2+1)t} \right] \\
 & \quad - id_8 \left[ e^{i\frac{\pi^2}{4}(K_8+K_2)t} - \frac{1}{2} e^{i\frac{\pi^2}{4}(K_8+K_2+1)t} \right] \\
 & \quad + id_9 \left[ -\frac{1}{2} e^{i\frac{\pi^2}{4}(K_9-K_3-1)t} + e^{i\frac{\pi^2}{4}(K_9-K_3)t} \right] \\
 & \quad - id_9 \left[ -\frac{1}{2} e^{i\frac{\pi^2}{4}(K_9+K_3-1)t} + e^{i\frac{\pi^2}{4}(K_9+K_3)t} - \frac{1}{2} e^{i\frac{\pi^2}{4}(K_9+K_3+1)t} \right] \\
 & \quad + i(d_8 - d_9) \frac{1}{2} e^{i\frac{\pi^2}{4}(K_8+K_2-1)t},
 \end{aligned}$$

where all of the terms are orthogonal, and thus

$$\begin{aligned}
 & \left\| 2\Re \left[ id_8 \left( e^{i\frac{\pi^2}{4}(K_8-K_2)t} - e^{i\frac{\pi^2}{4}(K_8+K_2)t} \right) \right. \right. \\
 & \quad \left. \left. + id_9 \left( e^{i\frac{\pi^2}{4}(K_9-K_3)t} - e^{i\frac{\pi^2}{4}(K_9+K_3)t} \right) \right] \frac{(1 - e^{i\frac{\pi^2}{4}t})(1 - e^{-i\frac{\pi^2}{4}t})}{2} \right\|_{L^2((0,T),\mathbb{R})}^2 \\
 &= [|d_8|^2 + |d_9|^2] * \frac{11}{4} * 2 + |d_8 - d_9|^2 * \frac{1}{4} * 2 \\
 &\leq \frac{13}{2} [|d_8|^2 + |d_9|^2].
 \end{aligned}$$

For the same reasons, we have

$$\begin{aligned}
 & \left\| 2\Re \left[ id_4 \left( e^{i\frac{\pi^2}{4}(K_4-K_2)t} - e^{i\frac{\pi^2}{4}(K_4+K_2)t} \right) \right. \right. \\
 & \quad \left. \left. + id_1 \left( e^{i\frac{\pi^2}{4}(K_1-K_3)t} - e^{i\frac{\pi^2}{4}(K_1+K_3)t} \right) \right] \frac{(1 - e^{i\frac{\pi^2}{4}t})(1 - e^{-i\frac{\pi^2}{4}t})}{2} \right\|_{L^2((0,T),\mathbb{R})}^2 \\
 &\leq \frac{13}{2} [|d_1|^2 + |d_4|^2].
 \end{aligned}$$

Thus, we have

$$\|P\|_{L^2}^2 \leq 12 \sum_{n \in \{1,2,3,4,8,9\}} |d_n|^2 + 6 \sum_{n \in \mathbb{N}^*, n \notin \{1,2,3,4,8,9\}} |d_n|^2$$

which justifies (8.2) with  $C_0 = 2\sqrt{3}$ . For the computation of  $\|P'\|_{L^2}^2$ , we use the same kind of arguments together with

$$\begin{aligned}
 & (K_n - K_j)^2 + (K_n + K_j)^2 \\
 & \quad + \frac{(K_n - K_j - 1)^2 + (K_n + K_j + 1)^2 + (K_n - K_j + 1)^2 + (K_n + K_j - 1)^2}{4} \\
 &= 3 \left( K_n^2 + K_j^2 + \frac{1}{3} \right).
 \end{aligned}$$

This gives (8.3) with  $C_1 = \pi^2 \sqrt{3}/(2\sqrt{2})$ . For the computation of  $\|P''\|_{L^2}^2$ , we also use the same kind of arguments, together with

$$\begin{aligned} & (K_n - K_j)^4 + (K_n + K_j)^4 \\ & + \frac{(K_n - K_j - 1)^4 + (K_n + K_j + 1)^4 + (K_n - K_j + 1)^4 + (K_n + K_j - 1)^4}{4} \\ & = 3K_n^4 + 6K_n^2(3K_j^2 + 1) + 3K_j^4 + 6K_j^2 + 1. \end{aligned}$$

This gives (8.4) with  $C_2 = \pi^4 \sqrt{3}/(8\sqrt{2})$ .  $\square$

In the next proposition, we check the second assumption of Proposition 13 with  $\widetilde{\mathcal{M}} = \widetilde{M}$  and  $\mathcal{M} = M$ .

PROPOSITION 15. *For every  $P \in H_0^2((0, T), \mathbb{R})$ , we have*

$$\|(\widetilde{M} - M)(P)\|_{l^2(\mathbb{N}^*, \mathbb{C})} \leq \mathcal{C}_0 \|P\|_{L^2((0, T), \mathbb{R})},$$

$$\|(\widetilde{M} - M)(P)\|_{h^2(\mathbb{N}^*, \mathbb{C})} \leq \mathcal{C}_1 \|P\|_{H_0^1((0, T), \mathbb{R})} + \mathcal{C}_3 \|P\|_{L^2((0, T), \mathbb{R})},$$

$$\|(\widetilde{M} - M)(P)\|_{h^4(\mathbb{N}^*, \mathbb{C})} \leq \mathcal{C}_2 \|P\|_{H_0^2((0, T), \mathbb{R})} + \mathcal{C}_4 \|P\|_{H_0^1((0, T), \mathbb{R})} + \mathcal{C}_5 \|P\|_{L^2((0, T), \mathbb{R})},$$

where

$$\begin{aligned} \mathcal{C}_0 &:= \sqrt{S_0} + \frac{T}{\sqrt{2}} \left( \left| \sqrt{\lambda_3} - \frac{\pi^2}{4} K_3 \right|^2 + \left| \sqrt{\lambda_2} - \frac{\pi^2}{2} K_2 \right|^2 \right)^{1/2}, \quad \mathcal{C}_2 := \frac{4}{\pi^2} \mathcal{C}_1, \\ \mathcal{C}_1 &:= \frac{2\sqrt{2}}{\pi^2} T \left( \left| \sqrt{\lambda_3} - \frac{\pi^2}{4} K_3 \right| + \left| \sqrt{\lambda_2} - \frac{\pi^2}{2} K_2 \right| \right), \\ S_0 &:= 2 \sum_{n=1}^{\infty} \left( 1 - \operatorname{sinc} \left[ T \left( \sqrt{\lambda_n} - \frac{\pi^2}{4} K_n \right) \right] \right), \end{aligned}$$

$\operatorname{sinc}(x) := \sin(x)/x$ , and  $\mathcal{C}_j$  is a positive constant for  $j = 3, 4, 5$ .

*Proof of Proposition 15.* By using decompositions of the form

$$\begin{aligned} (M - \widetilde{M})(P)_{2k-1} &= \frac{1}{T} \int_0^T P(t) \sin(\sqrt{\lambda_3} t) (e^{-i\sqrt{\lambda_{2k-1}} t} - e^{-i\frac{\pi^2}{4} K_{2k-1} t}) dt \\ (8.6) \quad &+ \frac{1}{T} \int_0^T P(t) \left[ \sin(\sqrt{\lambda_3} t) - \sin\left(\frac{\pi^2}{4} K_3 t\right) \right] e^{-i\frac{\pi^2}{4} K_{2k-1} t} dt, \end{aligned}$$

the Cauchy–Schwarz inequality, and Bessel Parseval inequality, we get

$$\begin{aligned} \|(M - \widetilde{M})(P)\|_{l^2(\mathbb{N}^*, \mathbb{C})} &\leq \left( \frac{1}{2} \left\| P(t) \left[ \sin(\sqrt{\lambda_3} t) - \sin\left(\frac{\pi^2}{4} K_3 t\right) \right] \right\|_{L^2}^2 \right. \\ &\quad \left. + \frac{1}{2} \left\| P(t) \left[ \sin(\sqrt{\lambda_2} t) - \sin\left(\frac{\pi^2}{4} K_2 t\right) \right] \right\|_{L^2}^2 \right)^{1/2} \\ &\quad + \|P\|_{L^2} \left( \sum_{n=1}^{\infty} \frac{1}{T} \int_0^T |e^{-i\sqrt{\lambda_n} t} - e^{-i\frac{\pi^2}{4} K_n t}|^2 dt \right)^{1/2} \end{aligned}$$

(the factor  $\frac{1}{2}$  comes from the fact that we sum only positive frequencies of a real-valued function), which gives the first bound of the proposition. By using decompositions of the form (8.6), the triangular inequality, the Cauchy–Schwarz inequality, and integrations by parts, we get

$$\begin{aligned} \|(M - \widetilde{M})(P)\|_{h^2} &\leq \|P\|_{L^2} \sqrt{S_1} + \frac{2\sqrt{2}}{\pi^2} \left( \left\| \frac{d}{dt} \left\{ P(t) \left[ \sin(\sqrt{\lambda_3}t) - \sin\left(\frac{\pi^2}{4}K_3t\right) \right] \right\} \right\|_{L^2} \right. \\ &\quad \left. + \left\| \frac{d}{dt} \left\{ P(t) \left[ \sin(\sqrt{\lambda_2}t) - \sin\left(\frac{\pi^2}{4}K_2t\right) \right] \right\} \right\|_{L^2} \right), \end{aligned}$$

where

$$S_1 := \sum_{n=1}^{\infty} K_n^2 \frac{1}{T} \int_0^T \left| e^{-i\sqrt{\lambda_n}t} - e^{-i\frac{\pi^2}{4}K_nt} \right|^2 dt.$$

In the same way, we get

$$\begin{aligned} \|(M - \widetilde{M})(P)\|_{h^4} &\leq \|P\|_{L^2} \sqrt{S_2} + \frac{8\sqrt{2}}{\pi^4} \left( \left\| \frac{d^2}{dt^2} \left\{ P(t) \left[ \sin(\sqrt{\lambda_3}t) - \sin\left(\frac{\pi^2}{4}K_3t\right) \right] \right\} \right\|_{L^2} \right. \\ &\quad \left. + \left\| \frac{d^2}{dt^2} \left\{ P(t) \left[ \sin(\sqrt{\lambda_2}t) - \sin\left(\frac{\pi^2}{4}K_2t\right) \right] \right\} \right\|_{L^2} \right), \end{aligned}$$

where

$$S_2 := \sum_{n=1}^{\infty} K_n^4 \frac{1}{T} \int_0^T \left| e^{-i\sqrt{\lambda_n}t} - e^{-i\frac{\pi^2}{4}K_nt} \right|^2 dt. \quad \square$$

Finally, in the next proposition, we check the last assumption of Proposition 13 with  $\widetilde{\mathcal{M}} = \widetilde{M}$  and  $\mathcal{M} = M$ .

**PROPOSITION 16.** *The constants  $C_0, C_1, C_2, \mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2$  defined in Propositions 14 and 15 satisfy  $C_1\mathcal{C}_1 = C_2\mathcal{C}_2$ ,  $C_0\mathcal{C}_0 < 1$ , and  $C_1\mathcal{C}_1 < 1$ .*

*Proof of Proposition 16.* First, let us give a bound on the last term in  $\mathcal{C}_0$ . We have

$$\begin{aligned} \cos\left(\frac{5\pi}{2}\right) \cosh\left(\frac{5\pi}{2}\right) &= 0 < 1, \\ \cos\left(\frac{5\pi}{2} - \frac{1}{1000}\right) \cosh\left(\frac{5\pi}{2} - \frac{1}{1000}\right) &= 1.286\dots > 1, \end{aligned}$$

and thus, thanks to the intermediate values theorem,  $x_2 \in (0, 1/1000)$ . We have

$$\sqrt{\lambda_2} - \frac{\pi^2}{4}K_2 = \left(\frac{5\pi}{2} - x_2\right)^2 - \frac{\pi^2}{4}K_2 = -5\pi x_2 + x_2^2.$$

The two terms of the right-hand side have different signs and  $x_2 < 5\pi$ , and thus

$$(8.7) \quad \left(\sqrt{\lambda_2} - \frac{\pi^2}{4}K_2\right)^2 < (5\pi x_2)^2 < 25\pi^2 10^{-6}.$$

In the same way, we deduce from

$$\cos\left(\frac{7\pi}{2}\right)\cosh\left(\frac{7\pi}{2}\right) = 0 < 1$$

and  $\cos\left(\frac{7\pi}{2} + \frac{1}{10000}\right)\cosh\left(\frac{7\pi}{2} + \frac{1}{10000}\right) = 2.98\dots > 1$

that  $x_3 \in (0, 1/10000)$ . Moreover,

$$\sqrt{\lambda_3} - \frac{\pi^2}{4}K_3 = \left(\frac{7\pi}{2} + x_3\right)^2 - \frac{\pi^2}{4}K_3 = 7\pi x_3 + x_3^2,$$

and thus

$$\left(\sqrt{\lambda_3} - \frac{\pi^2}{4}K_3\right)^2 < (7\pi 10^{-4} + 10^{-8})^2 < \pi^2 10^{-6}.$$

Therefore

$$(8.8) \quad \frac{T}{\sqrt{2}} \left[ \left(\sqrt{\lambda_2} - \frac{\pi^2}{4}K_2\right)^2 + \left(\sqrt{\lambda_3} - \frac{\pi^2}{4}K_3\right)^2 \right]^{1/2} < \frac{1}{\sqrt{2}} \frac{8}{\pi} (26\pi^2 10^{-6})^{1/2} = 8\sqrt{13} * 10^{-3}.$$

Now let us give a bound on  $S_0$ . We have

$$(8.9) \quad S_0 \leq 2 \sum_{n=1}^{\infty} \frac{1}{6} \left[ T \left( \sqrt{\lambda_n} - \frac{\pi^2}{4}K_n \right) \right]^2 \leq \frac{T^2}{3} \sum_{n=1}^{\infty} \left( \sqrt{\lambda_n} - \frac{\pi^2}{4}K_n \right)^2.$$

First, we study the cases where  $n$  is odd. We have

$$\nu_{2k-1} = 2k\pi - \frac{\pi}{2} + x_{2k-1} = (4k-1)\frac{\pi}{2} + x_{2k-1},$$

and thus

$$\left(\sqrt{\lambda_{2k-1}} - \frac{\pi^2}{4}K_{2k-1}\right)^2 = (\pi(4k-1)x_{2k-1} + x_{2k-1}^2)^2.$$

For every  $k \geq 2$ , we have  $x_{2k-1} \leq \pi(4k-1)$ , and thus, by using (3.3) and (3.4), we get

$$\begin{aligned} \left(\sqrt{\lambda_{2k-1}} - \frac{\pi^2}{4}K_{2k-1}\right)^2 &\leq [2\pi(4k-1)x_{2k-1}]^2 \\ &\leq 4 \left[ \pi(4k-1) \frac{\pi}{2 \cosh(\nu_{2k-1})} \right]^2 \\ &\leq \pi^2 \frac{[(4k-1)\pi]^2}{\cosh(2k\pi - \pi/2)^2} \\ &\leq \frac{\pi^2}{a} \frac{1}{\cosh(2k\pi - \pi/2)} \\ &\leq \frac{2\pi^2}{a} e^{-2k\pi + \frac{\pi}{2}}, \end{aligned}$$

where

$$a := \frac{1}{2(7\pi)^2} \left( e^{\frac{7\pi}{2}} - 1 - \frac{7\pi}{2} \right).$$

Indeed, for  $k \geq 2$ , we have

$$\begin{aligned} \cosh \left( 2k\pi - \frac{\pi}{2} \right) &= \cosh \left( (4k-1) \frac{\pi}{2} \right) \\ &\geq \frac{1}{2} \sum_{n=2}^{\infty} \frac{1}{n!} \left( (4k-1) \frac{\pi}{2} \right)^2 \left( \frac{7\pi}{2} \right)^{n-2} \\ &\geq \frac{1}{2} \left( \frac{2}{7\pi} \right)^2 \left( e^{\frac{7\pi}{2}} - 1 - \frac{7\pi}{2} \right) \left( (4k-1) \frac{\pi}{2} \right)^2 \\ &\geq a[(4k-1)\pi]^2. \end{aligned}$$

Thus,

$$\frac{T^2}{3} \sum_{n=1, n \text{ odd}}^{\infty} \left( \sqrt{\lambda_n} - \frac{\pi^2}{4} K_n \right)^2 \leq \frac{T^2}{3} \left\{ \left( \sqrt{\lambda_1} - \frac{\pi^2}{4} K_1 \right)^2 + \frac{2\pi^2}{a} \int_1^{\infty} e^{-2\pi x + \frac{\pi}{2}} dx \right\},$$

We have

$$\cos \left( \frac{3\pi}{2} + \frac{1}{55} \right) \cosh \left( \frac{3\pi}{2} + \frac{1}{55} \right) = 1.030 \dots > 1,$$

and thus  $x_1 \in (0, 1/55)$ . Therefore

$$(8.10) \quad \frac{T^2}{3} \sum_{n=1, n \text{ odd}}^{\infty} \left( \sqrt{\lambda_n} - \frac{\pi^2}{4} K_n \right)^2 \leq \frac{1}{3} \left( \frac{8}{\pi} \right)^2 \left\{ \left( \frac{3\pi}{55} + \left( \frac{1}{55} \right)^2 \right)^2 + \frac{\pi}{a} e^{-3\pi/2} \right\}$$

Now we deal with even integers in the sum  $S_0$ . We have

$$\nu_{2k} = 2k\pi + \frac{\pi}{2} - x_{2k} = (4k+1) \frac{\pi}{2} - x_{2k},$$

and thus, for  $k \geq 2$ , by using (3.4), we get

$$\begin{aligned} \left( \sqrt{\lambda_{2k}} - \frac{\pi^2}{4} K_{2k} \right)^2 &= (\pi(4k+1)x_{2k} - x_{2k}^2)^2 \\ &\leq (\pi(4k+1)x_{2k})^2 \\ &\leq \frac{\pi^2}{4} \frac{[(4k+1)\pi]^2}{\cosh(2k\pi + \pi/4)^2} \\ &\leq \frac{\pi^2}{4b} \frac{1}{\cosh(2k\pi + \pi/4)} \\ &\leq \frac{\pi^2}{2b} e^{-2k\pi - \pi/4}, \end{aligned}$$

where

$$b := \frac{2}{(18\pi)^2} \left( e^{17\pi/4} - 1 - \frac{17\pi}{4} \right).$$



Indeed, for  $k \geq 2$ , we have

$$\begin{aligned} \cosh \left[ 2k\pi + \frac{\pi}{4} \right] \cosh \left[ \left( 4k + \frac{1}{2} \right) \frac{\pi}{2} \right] \\ \geq \frac{1}{2} \sum_{n=2}^{\infty} \frac{1}{n!} \left[ \left( 4k + \frac{1}{2} \right) \frac{\pi}{2} \right]^2 \left( \frac{17\pi}{4} \right)^{n-2} \\ \geq \frac{1}{2} \left( \frac{4}{17\pi} \right)^2 \left\{ e^{\frac{17\pi}{4}} - 1 - \frac{17\pi}{4} \right\} \left( \frac{4k + \frac{1}{2}}{4k + 1} \right)^2 \left( (4k + 1) \frac{\pi}{2} \right)^2 \\ \geq \frac{1}{2} \left( \frac{4}{17\pi} \right)^2 \left\{ e^{\frac{17\pi}{4}} - 1 - \frac{17\pi}{4} \right\} \left( \frac{17}{18} \right)^2 \left( (4k + 1) \frac{\pi}{2} \right)^2 \\ \geq b[(4k + 1)\pi]^2. \end{aligned}$$

Thus,

$$(8.11) \quad \frac{T^2}{3} \sum_{n=2, n \text{ even}}^{\infty} \left( \sqrt{\lambda_n} - \frac{\pi^2}{4} K_n \right)^2 \leq \frac{T^2}{3} \left\{ \left( \sqrt{\lambda_2} - \frac{\pi^2}{4} K_2 \right)^2 + \frac{\pi^2}{2b} \int_1^{\infty} e^{-2\pi x - \frac{\pi}{2}} dx \right\}.$$

Thanks to (8.11), we get

$$(8.12) \quad \frac{T^2}{3} \sum_{n=2, n \text{ even}}^{\infty} \left( \sqrt{\lambda_n} - \frac{\pi^2}{4} K_n \right)^2 \leq \frac{1}{3} \left( \frac{8}{\pi} \right)^2 \left\{ 25\pi^2 10^{-6} + \frac{\pi}{4b} e^{-5\pi/2} \right\}.$$

Thanks to the explicit expression of  $C_0$  and  $\mathcal{C}_0$  and the inequalities (8.8), (8.9), (8.10), and (8.12), we get

$$C_0 \mathcal{C}_0 = 0.9847 \dots < 1.$$

Finally, we have

$$\begin{aligned} C_2 \mathcal{C}_2 = C_1 \mathcal{C}_1 &= \frac{\sqrt{3}}{2\sqrt{2}} \pi^2 \frac{2\sqrt{2}}{\pi^2} \frac{8}{\pi} (5\pi x_2 + 7\pi x_3 + x_3^2) \\ &\leq \sqrt{3} \frac{8}{\pi} \left( \frac{5\pi}{1000} + \frac{14\pi}{10000} \right) = 0.0886 \dots < 1. \end{aligned}$$

In this proof, there are two numerical values. They were computed thanks to the software Maple, with a precision that guarantees the validity of the first decimals.  $\square$

**9. Controllability of the linearized system around  $(u, u_t, p)$ .** In all of this section,  $T := 8/\pi$ . The goal of this section is the proof of the following result, which is the only assumption of Theorem 4 which is missing, for its application to the map  $\Phi_T$ , defined by (1.8).

PROPOSITION 17. *There exist  $\delta^* > 0$ ,  $C > 0$  such that*

- *for every  $(u_0, \dot{u}_0, p) \in E_8 \cap V$ , where*

$$V := \{(u_0, \dot{u}_0, p) \in E_4; \|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_4} < \delta^*\},$$

*the map  $d\Phi_T(u_0, \dot{u}_0, p)$  has a right inverse*

$$d\Phi_T(u_0, \dot{u}_0, p)^{-1} : F_7 \rightarrow E_6$$

such that, for every  $(U_0, \dot{U}_0, U_T, \dot{U}_T) \in F_7$ ,

$$\begin{aligned} \|d\Phi_T(u_0, \dot{u}_0, p)^{-1} \cdot (U_0, \dot{U}_0, U_T, \dot{U}_T)\|_{E_2} &\leq C \|(U_0, \dot{U}_0, U_T, \dot{U}_T)\|_{E_3}, \\ \|d\Phi_T(u_0, \dot{u}_0, p)^{-1} \cdot (U_0, \dot{U}_0, U_T, \dot{U}_T)\|_{E_6} &\leq C [\|(U_0, \dot{U}_0, U_T, \dot{U}_T)\|_{E_7} \\ &\quad + \|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_8} \|(U_0, \dot{U}_0, U_T, \dot{U}_T)\|_{E_3}] \end{aligned}$$

• and the map

$$\begin{array}{ccc} E_8 \cap V & \times & F_7 \\ ((u_0, \dot{u}_0, p) & , & (U_0, \dot{U}_0, U_T, \dot{U}_T)) \end{array} \begin{array}{c} \rightarrow \\ \mapsto \end{array} \begin{array}{c} E_6 \\ d\Phi_T(u_0, \dot{u}_0, p)^{-1} \cdot (U_0, \dot{U}_0, U_T, \dot{U}_T) \end{array}$$

is continuous.

*Remark 5.* A natural strategy for proving Proposition 17 consists in deducing it from Proposition 11 together with an argument of close linear maps. Indeed, if we prove that, for  $(u_0, \dot{u}_0, p) \in E_8 \cap V$ ,  $d\Phi_T(u_0, \dot{u}_0, p)$  is a continuous linear map from  $E_6$  to  $F_7$  such that

$$(9.1) \quad \left\| d\Phi_T(u_0, \dot{u}_0, p) - d\Phi_T(u_0^{ref}, \dot{u}_0^{ref}, 0) \right\|_{\mathcal{L}(E_6, F_7)} < 1,$$

then we know that  $d\Phi_T(u_0, \dot{u}_0, p)$  has a right inverse  $d\Phi_T(u_0, \dot{u}_0, p)^{-1} : F_7 \rightarrow E_6$ . This strategy is more or less the one used in this section. However, we chose to use the closeness between moment problems (associated to the controllability of the linearized systems) instead of the closeness between  $d\Phi_T(u_0, \dot{u}_0, p)$  and  $d\Phi_T(u_0^{ref}, \dot{u}_0^{ref}, 0)$ . Note that the results of section 2 are not sufficient to prove (9.1). Indeed, Proposition 3 ensures only that  $d\Phi_T(u_0, \dot{u}_0, p)$  is a continuous linear map from  $E_6$  to  $F_6$  (and not  $F_7$ ) and provides bounds only for

$$(9.2) \quad \left\| d\Phi_T(u_0, \dot{u}_0, p) - d\Phi_T(u_0^{ref}, \dot{u}_0^{ref}, 0) \right\|_{\mathcal{L}(E_6, F_6)}.$$

Therefore, in order to get (9.1), we need to work more.

In subsection 9.1, we detail the strategy of the proof. We explain that the proof of Proposition 17 is the consequence of two other results stated in Proposition 20s and 21. The proof of these propositions needs preliminary work done in subsection 9.2. Finally, in subsection 9.3, we prove Proposition 20, and, in subsection 9.4, we prove Proposition 21.

**9.1. Strategy.** In order to prove Proposition 17, we will transform the controllability of the linearized system around a trajectory  $(u, u_t, p)$  into the solvability of a “generalized moment problem.” Then we will use the closeness between the generalized moment problem associated to the linearized system around  $(u, u_t, p)$  and the moment problem associated to the linearized system around  $(u^{ref}, u_t^{ref}, 0)$  to deduce the solvability of the first one from the solvability of the second one (which was proved in the previous section).

First, let us write the controllability of the linearized system around a trajectory  $(u, u_t, p)$  into a generalized moment problem. In order to do that, we need some new notations. For  $\gamma \in \mathbb{R}$ , we introduce the operator  $A_\gamma$  defined by

$$D(A_\gamma) := H^4 \cap H_0^2((0, 1), \mathbb{C}), \quad A_\gamma u := u_{xxxx} + \gamma u_{xx}.$$

It is a symmetric unbounded operator on  $L^2((0, 1), \mathbb{R})$ . Let  $(\lambda_{k, \gamma})_{k \in \mathbb{N}^*}$  be the non-decreasing sequence of the eigenvalues of  $A_\gamma$  and  $(\varphi_{k, \gamma})_{k \in \mathbb{N}^*}$  associated eigenvectors,

which form an orthonormal basis of  $L^2((0, 1), \mathbb{R})$ . The maps  $\gamma \mapsto \lambda_{k,\gamma}$  and  $\gamma \mapsto \varphi_{k,\gamma}$  are analytic, which gives a sense to the notations

$$\left. \frac{d\varphi_{k,\gamma}}{d\gamma} \right|_{\gamma_1} \quad \text{and} \quad \left. \frac{d}{d\gamma} \left[ \frac{1}{\sqrt{\lambda_{k,\gamma}}} \varphi_{k,\gamma} \right] \right|_{\gamma_1},$$

which are, respectively, the derivative of the map  $\gamma \mapsto \varphi_{k,\gamma}$  considered at the point  $\gamma = \gamma_1$  and the derivative of the map  $\gamma \mapsto \varphi_{k,\gamma}/\sqrt{\lambda_{k,\gamma}}$  considered at the point  $\gamma = \gamma_1$ .

Now we transform the controllability of the linearized system around a trajectory  $(u, u_t, p)$  into a generalized moment problem. Let  $(u_0, \dot{u}_0, p) \in E_8$  and  $u \in C^0([0, T], H^8((0, 1), \mathbb{R}))$  be the solution of

$$\begin{cases} u_{tt} + u_{xxxx} + p(t)u_{xx} = 0, x \in (0, 1), t \in (0, T), \\ u = u_x = 0 \text{ at } x = 0, 1, \\ u(0) = u_0, u_t(0) = \dot{u}_0. \end{cases}$$

Let  $(U_0, \dot{U}_0, P) \in E_6$ . We have

$$d\Phi_T(u_0, \dot{u}_0, P) \cdot (U_0, \dot{U}_0, P) = (U_0, \dot{U}_0, U(T), U_t(T)),$$

where  $U \in C^0([0, T], H^6((0, 1), \mathbb{R}))$  is the solution of

$$\begin{cases} U_{tt} + U_{xxxx} + p(t)U_{xx} + P(t)u_{xx} = 0, x \in (0, 1), t \in (0, T), \\ U = U_x = 0 \text{ at } x = 0, 1, \\ U(0) = U_0, U_t(0) = \dot{U}_0. \end{cases}$$

We have, for every  $t \in [0, T]$ ,

$$(9.3) \quad \begin{pmatrix} U(t) \\ U_t(t) \end{pmatrix} = \sum_{k=1}^{\infty} 2\Re(x_k(t)X_{k,p(t)}),$$

where, for every  $k \in \mathbb{N}^*$ ,

$$(9.4) \quad x_k(t) := \frac{1}{2} \int_0^1 \left[ U(t, x)\varphi_{k,p(t)}(x) - \dot{U}(t, x) \frac{1}{i\sqrt{\lambda_{k,p(t)}}} \varphi_{k,p(t)}(x) \right] dx,$$

$$(9.5) \quad X_{k,\gamma} := \begin{pmatrix} \varphi_{k,\gamma} \\ -i\sqrt{\lambda_{k,\gamma}}\varphi_{k,\gamma} \end{pmatrix}.$$

By using the partial differential equation solved by  $U$ , we get

$$\begin{aligned} \dot{x}_k &= \frac{1}{2} \int_0^1 \left( U_t \varphi_{k,p} + [U_{xxxx} + p(t)U_{xx} + P(t)u_{xx}] \frac{\varphi_{k,p}}{i\sqrt{\lambda_{k,p}}} \right) dx \\ &\quad + \frac{\dot{p}}{2} \int_0^1 \left( U \left[ \frac{d\varphi_{k,\gamma}}{d\gamma} \right]_{p(t)} - U_t \frac{d}{d\gamma} \left[ \frac{\varphi_{k,\gamma}}{i\sqrt{\lambda_{k,\gamma}}} \right]_{p(t)} \right) dx \\ &= -i\sqrt{\lambda_{k,p(t)}}x_k(t) - \frac{i}{2}P(t) \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle \\ &\quad + \frac{1}{2}\dot{p}(t) \left( \left\langle U(t), \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{p(t)} + i \left\langle \dot{U}(t), \frac{d}{d\gamma} \left[ \frac{\varphi_{k,\gamma}}{\sqrt{\lambda_{k,\gamma}}} \right]_{p(t)} \right\rangle \right). \end{aligned}$$

This first order ordinary differential equation can be solved explicitly. Then the equality

$$d\Phi_T(u_0, \dot{u}_0, p) \cdot (U_0, \dot{U}_0, P) = (U_0, \dot{U}_0, U_T, \dot{U}_T)$$

is equivalent to  $(U(T), \dot{U}(T)) = (U_T, \dot{U}_T)$ , which is equivalent to the generalized moment problem

$$\Upsilon_{(u_0, \dot{u}_0, p)}(P) = d(U_0, \dot{U}_0, U_T, \dot{U}_T),$$

where  $\Upsilon_{(u_0, \dot{u}_0, p)}(P) := (\Upsilon_{(u_0, \dot{u}_0, p)}(P)_k)_{k \in \mathbb{N}^*}$ ,  $d(U_0, \dot{U}_0, U_T, \dot{U}_T) := (d(U_0, \dot{U}_0, U_T, \dot{U}_T)_k)_{k \in \mathbb{N}^*}$ , and, for every  $k \in \mathbb{N}^*$ ,

(9.6)

$$\begin{aligned} \Upsilon_{(u_0, \dot{u}_0, p)}(P)_k &:= \int_0^T \left\{ \frac{-i}{2} P(t) \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle + \frac{1}{2} \dot{p}(t) \left\langle U(t), \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{p(t)} \right. \\ &\quad \left. + \frac{i}{2} \dot{p}(t) \left\langle \dot{U}(t), \frac{d}{d\gamma} \left[ \frac{\varphi_{k,\gamma}}{\sqrt{\lambda_{k,\gamma}}} \right]_{p(t)} \right\rangle \right\} e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt, \\ d(U_0, \dot{U}_0, U_T, \dot{U}_T)_k &:= \frac{1}{2} e^{i \int_0^T \sqrt{\lambda_{k,p(s)}} ds} \left[ \langle U_T, \varphi_k \rangle - \frac{1}{i\sqrt{\lambda_k}} \langle \dot{U}_T, \varphi_k \rangle \right] \\ &\quad - \frac{1}{2} \left[ \langle U_0, \varphi_k \rangle - \frac{1}{i\sqrt{\lambda_k}} \langle \dot{U}_0, \varphi_k \rangle \right]. \end{aligned}$$

Notice that the right-hand side  $d(U_0, \dot{U}_0, U_T, \dot{U}_T)$  belongs to  $h^7(\mathbb{N}^*, \mathbb{C})$  when  $(U_0, \dot{U}_0, U_T, \dot{U}_T) \in F_7$ . Thus, Proposition 17 is equivalent to the following proposition.

PROPOSITION 18. *There exist  $\delta^* > 0$ ,  $C > 0$  such that*

- *for every  $(u_0, \dot{u}_0, p) \in E_8 \cap V$ , where*

$$V := \{(u_0, \dot{u}_0, p) \in E_4; \|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_4} < \delta^*\},$$

*the map  $\Upsilon_{(u_0, \dot{u}_0, p)}$  has a right inverse*

$$\Upsilon_{(u_0, \dot{u}_0, p)}^{-1} : h^7(\mathbb{N}^*, \mathbb{C}) \rightarrow H^2 \cap H_0^1((0, T), \mathbb{R})$$

*such that, for every  $d \in h^7(\mathbb{N}^*, \mathbb{C})$ ,*

$$\|\Upsilon_{(u_0, \dot{u}_0, p)}^{-1} \cdot d\|_{L^2((0, T), \mathbb{R})} \leq C \|d\|_{h^3},$$

$$\|\Upsilon_{(u_0, \dot{u}_0, p)}^{-1} \cdot d\|_{H^2} \leq C [\|d\|_{h^7} + \|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_8} \|d\|_{h^3}],$$

- *the map*

$$\begin{array}{ccc} E_8 \cap V & \times & h^7(\mathbb{N}^*, \mathbb{C}) \rightarrow H^2 \cap H_0^1((0, T), \mathbb{R}) \\ ((u_0, \dot{u}_0, p) & , & d) \mapsto \Upsilon_{(u_0, \dot{u}_0, p)}^{-1} \cdot d \end{array}$$

*is continuous.*

We will get Proposition 18 by applying the following proposition with  $\widetilde{\mathcal{M}}$  replaced by  $\Upsilon_{(u_0^{ref}, \dot{u}_0^{ref}, 0)}$ ,  $\mathcal{M}$  replaced by  $\Upsilon_{(u_0, \dot{u}_0, p)}$ ,  $\Delta_4$  replaced by  $C\|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_4}$ , and  $\Delta_8$  replaced by  $C\|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_8}$ .

PROPOSITION 19. Let  $\widetilde{\mathcal{M}}$  be a continuous linear map  $L^2((0, T), \mathbb{R}) \rightarrow h^3(\mathbb{N}^*, \mathbb{C})$  and  $H_0^2((0, T), \mathbb{R}) \rightarrow h^7(\mathbb{N}^*, \mathbb{C})$  that has a continuous right inverse  $\widetilde{\mathcal{M}}^{-1} : h^7(\mathbb{N}^*, \mathbb{C}) \rightarrow H_0^2((0, T), \mathbb{R})$  satisfying, for every  $d \in h^7(\mathbb{N}^*, \mathbb{C})$ ,

$$\begin{aligned}\|\widetilde{\mathcal{M}}^{-1}(d)\|_{L^2} &\leq C_0 \|d\|_{h^3}, \\ \|\widetilde{\mathcal{M}}^{-1}(d)\|_{H_0^2} &\leq C_0 \|d\|_{h^7},\end{aligned}$$

with some positive constant  $C_0$ .

(1) Every linear map  $\mathcal{M}$  for which there exists  $\Delta_4, \Delta_8 > 0$ , with  $C_0 \Delta_4 < 1$  such that, for every  $P \in H_0^2((0, T), \mathbb{R})$ ,

$$(9.7) \quad \begin{aligned}\|(\widetilde{\mathcal{M}} - \mathcal{M})(P)\|_{h^3} &\leq \Delta_4 \|P\|_{L^2}, \\ \|(\widetilde{\mathcal{M}} - \mathcal{M})(P)\|_{h^7} &\leq \Delta_4 \|P\|_{H_0^2} + \Delta_8 \|P\|_{L^2},\end{aligned}$$

has a right inverse  $\mathcal{M}^{-1} : h^7(\mathbb{N}^*, \mathbb{C}) \rightarrow H_0^2((0, T), \mathbb{R})$  such that, for every  $d \in h^7(\mathbb{N}^*, \mathbb{C})$ ,

$$(9.8) \quad \begin{aligned}\|\mathcal{M}^{-1}(d)\|_{L^2} &\leq \frac{C_0}{1 - C_0 \Delta_4} \|d\|_{h^3}, \\ \|\mathcal{M}^{-1}(d)\|_{H_0^2} &\leq \frac{C_0}{1 - C_0 \Delta_4} \|d\|_{h^7} + \left( \frac{C_0}{1 - C_0 \Delta_4} + \frac{C_0 \Delta_8}{(1 - C_0 \Delta_4)^2} \right) \|d\|_{h^3}.\end{aligned}$$

(2) Let  $(\mathcal{M}_\epsilon)_{\epsilon>0}$ ,  $\mathcal{M}$  be linear maps that satisfy the assumptions of statement (1) with constants  $\Delta_4^\epsilon, \Delta_8^\epsilon, \Delta_4, \Delta_8$ . Let  $(d_\epsilon)_{\epsilon>0}, d \in h^7(\mathbb{N}^*, \mathbb{C})$ . We assume the following:

- (a)  $\mathcal{M}_\epsilon \rightarrow \mathcal{M}$  weakly in  $\mathcal{L}(H_0^2((0, T), \mathbb{R}), h^7(\mathbb{N}^*, \mathbb{C}))$ , i.e., for every  $P \in H_0^2((0, T), \mathbb{R})$ ,  $\mathcal{M}_\epsilon(P) \rightarrow \mathcal{M}(P)$  in  $h^7(\mathbb{N}^*, \mathbb{C})$  when  $\epsilon \rightarrow 0$ ;
- (b)  $d_\epsilon \rightarrow d$  in  $h^7(\mathbb{N}^*, \mathbb{C})$  when  $\epsilon \rightarrow 0$ ;
- (c) there exists  $\Delta_4^*, \Delta_8^* > 0$  such that  $C_0 \Delta_4^* < 1$ , and for every  $\epsilon > 0$ ,  $\Delta_4^\epsilon \leq \Delta_4^*$  and  $\Delta_8^\epsilon \leq \Delta_8^*$ .

Then  $\mathcal{M}_\epsilon^{-1}(d_\epsilon) \rightarrow \mathcal{M}^{-1}(d)$  in  $H_0^2((0, T), \mathbb{R})$ .

*Proof of Proposition 19.* First, we prove statement (1). Let  $d \in h^7(\mathbb{N}^*, \mathbb{C})$ . We define a sequence  $(P_n)_{n \in \mathbb{N}^*} \subset H_0^2((0, T), \mathbb{R})$  by

$$\begin{cases} P_0 := \widetilde{\mathcal{M}}^{-1}(P_0), \\ P_{n+1} := \widetilde{\mathcal{M}}^{-1}[(\widetilde{\mathcal{M}} - \mathcal{M})(P_n)] \quad \forall n \in \mathbb{N}. \end{cases}$$

Then, for every  $n \in \mathbb{N}$ , we have

$$(9.9) \quad \mathcal{M} \left( \sum_{k=0}^n P_k \right) = d + (\mathcal{M} - \widetilde{\mathcal{M}})(P_n).$$

Thanks to (9.7), we have, for every  $n \in \mathbb{N}$ ,

$$(9.10) \quad \begin{aligned}\|P_n\|_{L^2} &\leq C_0 (C_0 \Delta_4)^n \|d\|_{h^3}, \\ \|P_n\|_{H_0^2} &\leq C_0 (C_0 \Delta_4)^n \|d\|_{h^7} + y_n \|d\|_{h^3},\end{aligned}$$

where  $(y_n)_{n \in \mathbb{N}} \subset \mathbb{R}$  is defined by

$$(9.11) \quad \begin{cases} y_0 = C_0, \\ y_{n+1} = C_0 \Delta_4 y_n + C_0 \Delta_8 (C_0 \Delta_4)^{n+1} \quad \forall n \in \mathbb{N}. \end{cases}$$

Since  $C_0\Delta_4 < 1$ , then  $\sum P_n$  converges in  $H_0^2((0, T), \mathbb{R})$  to  $P := \sum_{n=0}^{\infty} P_n$ . By using (9.9), the convergence of  $P_n$  to zero in  $H_0^2((0, T), \mathbb{R})$ , and the continuity of  $\mathcal{M} - \widetilde{\mathcal{M}} : H_0^2 \rightarrow h^7$ , we get  $\mathcal{M}(P) = d$ . Moreover, we have

$$\begin{aligned} \|P\|_{L^2} &\leq \sum_{n=0}^{\infty} \|P_n\|_{L^2} \leq \sum_{n=0}^{\infty} C_0(C_0\Delta_4)^n \|d\|_{h^3} \leq \frac{C_0}{1 - C_0\Delta_4} \|d\|_{h^3}, \\ \|P\|_{H_0^2} &\leq \sum_{n=0}^{\infty} \|P_n\|_{H_0^2} \leq \frac{C_0}{1 - C_0\Delta_4} \|d\|_{h^7} + \left( \sum_{n=0}^{\infty} y_n \right) \|d\|_{h^3}. \end{aligned}$$

By using (9.11), we get

$$\left( \sum_{n=0}^{\infty} y_n \right) - C_0 = C_0\Delta_4 \left( \sum_{n=0}^{\infty} y_n \right) + \frac{C_0^2\Delta_4\Delta_8}{1 - C_0\Delta_4},$$

which gives (9.8).

Now we prove statement (2). Let  $P^\epsilon := \mathcal{M}_\epsilon^{-1}(d_\epsilon)$ ,  $P := \mathcal{M}^{-1}(d) \in H_0^2((0, T), \mathbb{R})$  built with the previous construction, i.e.,  $P^\epsilon = \sum_{n=0}^{\infty} P_n^\epsilon$  and  $P = \sum_{n=0}^{\infty} P_n$ . We want to prove that  $P^\epsilon \rightarrow P$  in  $H_0^2((0, T), \mathbb{R})$  when  $\epsilon \rightarrow 0$ .

First we prove by induction on  $n \in \mathbb{N}$  that, for every  $n \in \mathbb{N}$ ,

$$\mathbf{H}(n) : P_n^\epsilon \rightarrow P_n \text{ in } H_0^2((0, T), \mathbb{R}) \text{ when } \epsilon \rightarrow 0.$$

It is clear that  $P_0^\epsilon := \widetilde{\mathcal{M}}^{-1}(d_\epsilon)$  converges to  $P_0 = \widetilde{\mathcal{M}}^{-1}(d)$  in  $H_0^2((0, T), \mathbb{R})$ . Indeed,  $\widetilde{\mathcal{M}}^{-1} : h^7 \rightarrow H_0^2$  is continuous and  $d_\epsilon \rightarrow d$  in  $h^7$ . This proves  $\mathbf{H}(0)$ .

Let  $n \in \mathbb{N}$ . We assume  $\mathbf{H}(n)$ . Let us recall that

$$P_{n+1}^\epsilon = \widetilde{\mathcal{M}}^{-1}[(\widetilde{\mathcal{M}} - \mathcal{M}_\epsilon)(P_n^\epsilon)].$$

Since

$$\widetilde{\mathcal{M}} - \mathcal{M}_\epsilon \rightarrow \widetilde{\mathcal{M}} - \mathcal{M} \text{ weakly in } \mathcal{L}(H_0^2, h^7) \text{ when } \epsilon \rightarrow 0,$$

and  $P_n^\epsilon \rightarrow P_n$  strongly in  $H_0^2$ , then

$$(\widetilde{\mathcal{M}} - \mathcal{M}_\epsilon)(P_n^\epsilon) \rightarrow (\widetilde{\mathcal{M}} - \mathcal{M})(P_n) \text{ in } h^7 \text{ when } \epsilon \rightarrow 0.$$

Thanks to the continuity of  $\widetilde{\mathcal{M}}^{-1} : h^7 \rightarrow H_0^2$ , we deduce that  $\mathbf{H}(n+1)$  holds. This ends the proof by induction.

Now, let us prove statement (2) thanks to the dominated convergence theorem. For every  $\epsilon > 0$ , for every  $n \in \mathbb{N}^*$ , we know that

$$\|P_n^\epsilon\|_{H_0^2} \leq C_0(C_0\Delta_4^\epsilon)^n \|d\|_{h^7} + y_n^\epsilon \|d\|_{h^3},$$

where  $(y_n^\epsilon)_{\epsilon>0} \subset \mathbb{R}$  is defined by

$$\begin{cases} y_0^\epsilon = C_0, \\ y_{n+1}^\epsilon = C_0\Delta_4^\epsilon y_n^\epsilon + C_0\Delta_8^\epsilon (C_0\Delta_4^\epsilon)^n \quad \forall n \in \mathbb{N}. \end{cases}$$

Let  $(y_n^*)_{\epsilon>0} \subset \mathbb{R}$  be defined by

$$\begin{cases} y_0^* = C_0, \\ y_{n+1}^* = C_0\Delta_4^* y_n^* + C_0\Delta_8^* (C_0\Delta_4^*)^n. \end{cases}$$

Since  $C_0\Delta_4^* < 1$ , then  $(y_n^*)_{n \in \mathbb{N}} \in l^1(\mathbb{N}, \mathbb{C})$ . Moreover, for every  $\epsilon > 0$ , for every  $n \in \mathbb{N}$ ,  $y_n^\epsilon \leq y_n^*$ . Therefore, we have

- $P_n^\epsilon \rightarrow P_n$  in  $H_0^2((0, T), \mathbb{R})$  when  $\epsilon \rightarrow 0$ ,
- and for every  $\epsilon > 0$  and for every  $n \in \mathbb{N}$ ,

$$\|P_n^\epsilon\|_{H_0^2} \leq C_0(C_0\Delta_4^*)^n \|d\|_{h^7} + C_0\Delta_8^* y_n^* \|d\|_{h^3}.$$

The right-hand side of the previous inequality defines a sequence in  $l^1(\mathbb{N}, \mathbb{C})$ , and thus the dominated convergence theorem allows one to conclude that

$$P^\epsilon = \sum_{n=0}^{\infty} P_n^\epsilon \rightarrow \sum_{n=0}^{\infty} P_n = P \text{ in } H_0^2((0, T), \mathbb{R}) \text{ when } \epsilon \rightarrow 0.$$

This ends the proof of Proposition 19.  $\square$

Now let us explain how we apply Proposition 19 with  $\widetilde{\mathcal{M}} = \Upsilon_{(u_0^{ref}, \dot{u}_0^{ref}, 0)}$ ,  $\mathcal{M} = \Upsilon_{(u_0, \dot{u}_0, p)}$ ,  $\mathcal{M}_\epsilon = \Upsilon_{(u_0^\epsilon, \dot{u}_0^\epsilon, p^\epsilon)}$ , where  $(u_0^\epsilon, \dot{u}_0^\epsilon, p^\epsilon) \rightarrow (u_0, \dot{u}_0, p)$  in  $E_8$ . First, note that the first assumption of Proposition 19, with  $\widetilde{\mathcal{M}} = \Upsilon_{(u_0^{ref}, \dot{u}_0^{ref}, 0)}$ , holds thanks to Proposition 12. In order to prove the estimate (9.7), we use the following decomposition:

$$\begin{aligned} & (\Upsilon_{(u_0, \dot{u}_0, p)} - \Upsilon_{(u^{ref}(0), \dot{u}^{ref}(0), 0)})(P) \\ &= \frac{-i}{2} \Upsilon_{(u_0, \dot{u}_0, p)}^1(P) + \frac{1}{2} \Upsilon_{(u_0, \dot{u}_0, p)}^2(P) + \frac{i}{2} \Upsilon_{(u_0, \dot{u}_0, p)}^3(P), \end{aligned}$$

where, for every  $k \in \mathbb{N}^*$ ,

$$\begin{aligned} \Upsilon_{(u_0, \dot{u}_0, p)}^1(P)_k &:= \int_0^T P(t) \left( \left\langle u_{xx}(t), \frac{\varphi_{k,p}(t)}{\sqrt{\lambda_{k,p}(t)}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p}(s)} ds} - \left\langle u_{xx}^{ref}(t), \frac{\varphi_k}{\sqrt{\lambda_k}} \right\rangle e^{i \sqrt{\lambda_k} t} \right) dt, \\ \Upsilon_{(u_0, \dot{u}_0, p)}^2(P)_k &:= \int_0^T \dot{p}(t) \left\langle U(t), \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{p(t)} e^{i \int_0^t \sqrt{\lambda_{k,p}(s)} ds} dt, \\ \Upsilon_{(u_0, \dot{u}_0, p)}^3(P)_k &:= \int_0^T \dot{p}(t) \left\langle \dot{U}(t), \frac{d}{d\gamma} \left[ \frac{\varphi_{k,\gamma}}{\sqrt{\lambda_{k,\gamma}}} \right]_{p(t)} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p}(s)} ds} dt. \end{aligned}$$

In sections 9.3 and 9.4, we prove the following proposition that allows us to deduce Proposition 18 from Proposition 19.

**PROPOSITION 20.** *There exist  $\delta^* > 0$ ,  $C > 0$  such that, for every  $j \in \{1, 2, 3\}$ , for every  $(u_0, \dot{u}_0, p) \in E_8$ , and for every  $P \in H_0^2((0, T), \mathbb{R})$ ,*

$$\|\Upsilon_{(u_0, \dot{u}_0, p)}^j(P)\|_{h^3(\mathbb{N}^*, \mathbb{C})} \leq C \|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_4} \|P\|_{L^2},$$

$$\begin{aligned} & \|\Upsilon_{(u_0, \dot{u}_0, p)}^j(P)\|_{h^7(\mathbb{N}^*, \mathbb{C})} \\ & \leq C [\|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_8} \|P\|_{L^2} + \|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_4} \|P\|_{H_0^2}]. \end{aligned}$$

**PROPOSITION 21.** *Let  $((u_0^\epsilon, \dot{u}_0^\epsilon, p^\epsilon))_{\epsilon>0}$ ,  $(u_0, \dot{u}_0, p)$  in  $E_8$  such that  $(u_0^\epsilon, \dot{u}_0^\epsilon, p^\epsilon) \rightarrow (u_0, \dot{u}_0, p)$  in  $E_8$  when  $\epsilon \rightarrow 0$ . Then  $\Upsilon_{(u_0^\epsilon, \dot{u}_0^\epsilon, p^\epsilon)} \rightarrow \Upsilon_{(u_0, \dot{u}_0, p)}$  weakly in  $\mathcal{L}(H_0^2, h^7)$ , when  $\epsilon \rightarrow 0$ .*

The proofs of Propositions 20 and 21 need preliminary work that is developed in the next subsection.

**9.2. Preliminaries.** In this subsection, we prove technical results which are useful for the proof of Propositions 20 and 21.

**9.2.1. Technical results about  $\lambda_{k,\gamma}$  and  $\varphi_{k,\gamma}$ .** In this section, we state some useful results on the eigenvalues  $(\lambda_{k,\gamma})_{k \in \mathbb{N}^*}$  and eigenfunctions  $(\varphi_{k,\gamma})_{k \in \mathbb{N}^*}$  introduced in section 9.1. When  $\gamma = 0$ , we write  $\lambda_k$  and  $\varphi_k$  instead of  $\lambda_{k,0}$  and  $\varphi_{k,0}$ . In this case, explicit expressions and asymptotic behaviors are given in Proposition 6. It is well known that  $\varphi_{k,\gamma}$  and  $\lambda_{k,\gamma}$  are analytic functions of the parameter  $\gamma$ :

$$(9.12) \quad \begin{aligned} \varphi_{k,\gamma} &= \varphi_k + \gamma \varphi_k^{(1)} + \gamma^2 \varphi_k^{(2)} + \gamma^3 \varphi_k^{(3)} + \cdots, \\ \lambda_{k,\gamma} &= \lambda_k + \gamma \lambda_k^{(1)} + \gamma^2 \lambda_k^{(2)} + \gamma^3 \lambda_k^{(3)} + \cdots. \end{aligned}$$

LEMMA 2. *There exists  $c \in (0, 1)$  such that, for every  $j, k \in \mathbb{N}^*$  with the same parity*

$$(9.13) \quad (1 - c)\pi|k - j| \leq |\nu_k - \nu_j| \leq (1 + c)\pi|k - j|,$$

$$(9.14) \quad (1 - c) \left( \frac{\pi}{2} \right)^4 |(2k + 1)^4 - (2j + 1)^4| \leq |\lambda_k - \lambda_j| \leq (1 + c) \left( \frac{\pi^2}{2} \right)^4 |(2k + 1)^4 - (2j + 1)^4|.$$

*Proof of Lemma 2.* When  $k, j \in \mathbb{N}^*$ ,  $k \neq j$ , and  $k, j$  have the same parity, thanks to (3.3), we have

$$\frac{\pi}{2}|k - j| \leq |\nu_k - \nu_j| \leq 2\pi|k - j|$$

and

$$\begin{aligned} & \left| (\lambda_k - \lambda_j) - \left( \frac{\pi}{2} \right)^4 [(2k + 1)^4 - (2j + 1)^4] \right| \\ & \leq \max \left\{ \left| \lambda_k - \left( \frac{\pi}{2} (2k + 1) \right)^4 \right|, \left| \lambda_j - \left( \frac{\pi}{2} (2j + 1) \right)^4 \right| \right\}. \end{aligned}$$

Moreover, by using (3.4), we get

$$\begin{aligned} \left| \lambda_k - \left( \frac{\pi}{2} (2k + 1) \right)^4 \right| & \leq 15 \left( \frac{\pi}{2} (2k + 1) \right)^3 x_k \text{ because } 0 < x_k < 1 \\ & \leq \left( \frac{\pi}{2} \right)^4 m \frac{15}{2320} \frac{(2k + 1)^3}{\cosh(k\pi)}, \end{aligned}$$

where  $m := \inf\{|(2l + 1)^4 - (2i + 1)^4|; l \neq i \in \mathbb{N}^* \text{ with the same parity}\} = 2320$ . Thus,

$$\begin{aligned} \left| \lambda_k - \left( \frac{\pi}{2} (2k + 1) \right)^4 \right| & \leq \left( \frac{\pi}{2} \right)^4 m c_1 \\ & \leq \left( \frac{\pi}{2} \right)^4 |(2k + 1)^4 - (2j + 1)^4| c_1, \end{aligned}$$

where

$$c_1 := \sup \left\{ \frac{15}{2320} \frac{(2k + 1)^3}{\cosh(\pi)}; k \in \mathbb{N}^* \right\} = \frac{81}{464 \cosh(\pi)} < 1,$$

which ends the proof.  $\square$



PROPOSITION 22. *For every  $k \in \mathbb{N}^*$ , we have*

$$(9.15) \quad \frac{d^4 \varphi_k^{(1)}}{dx^4} + \frac{d^2 \varphi_k}{dx^2} = \lambda_k \varphi_k^{(1)} + \lambda_k^{(1)} \varphi_k,$$

$$(9.16) \quad \lambda_k^{(1)} = -\|\varphi_k'\|_{L^2((0,1),\mathbb{R})}^2,$$

$$(9.17) \quad \varphi_k^{(1)} = \sum_{j \in \mathbb{N}^*, j \neq k, P(j)=P(k)} x_{k,j} \varphi_j, \text{ where } x_{k,j} := \frac{\langle \varphi_k'', \varphi_j \rangle}{(\lambda_k - \lambda_j)},$$

where the sum is taken over all of the integers  $j$  different from  $k$ , with the same parity as  $k$ . There exists a constant  $C > 0$  such that, for every  $k \in \mathbb{N}^*$ ,

$$(9.18) \quad |\lambda_k^{(1)}| \leq Ck^2,$$

$$(9.19) \quad \|\varphi_k^{(1)}\|_{L^2((0,1),\mathbb{R})} \leq \frac{C}{k}.$$

*Proof of Proposition 22.* Equation (9.15) corresponds to the term of first order with respect to  $\gamma$ , in the equality  $A_\gamma \varphi_{k,\gamma} = \lambda_{k,\gamma} \varphi_{k,\gamma}$  developed thanks to (9.12). Notice that the equality

$$\|\varphi_{k,\gamma}\|_{L^2((0,1),\mathbb{R})} = 1$$

implies  $\langle \varphi_k^{(1)}, \varphi_k \rangle = 0$ . Then we get (9.16) by taking the  $L^2$ -scalar product of (9.15) with the vector  $\varphi_k$ . The expression (9.17) comes from (9.15), and the parity of the functions  $\varphi_j$  justify that half of the components vanish.

Thanks to (9.16), the convexity of the  $H^s$ -norms, the behavior (3.3), and the equalities

$$\|\varphi_k\|_{L^2((0,1),\mathbb{R})} = 1, \quad \left\| \frac{d^4 \varphi_k}{dx^4} \right\|_{L^2((0,1),\mathbb{R})} = \lambda_k,$$

we get (9.18). By using (9.14) and  $\|\varphi_k''\|_{L^2} = \sqrt{\lambda_k} \leq Ck^2$ , we get

$$\begin{aligned} \|\varphi_k^{(1)}\|_{L^2((0,1),\mathbb{R})} &= \left[ \sum_{j \in \mathbb{N}^*, j \neq k, P(j)=P(k)} \left( \frac{\langle \varphi_k'', \varphi_j \rangle}{\lambda_k - \lambda_j} \right)^2 \right]^{1/2} \\ &\leq Ck^2 \left[ \sum_{j \in \mathbb{N}^*, j \neq k, P(j)=P(k)} \frac{1}{[(2k+1)^4 - (2j+1)^4]^2} \right]^{1/2}. \end{aligned}$$

Thanks to the explicit expression for  $x > 0$ ,  $x \neq K$ ,

$$\begin{aligned} \int \frac{dx}{(x^4 - K^4)^2} &= \frac{3}{16K^7} \ln \left( \frac{x+K}{|x-K|} \right) \\ &\quad - \frac{1}{16K^6} \frac{2x}{x^2 - K^2} + \frac{3}{8K^7} \arctan \left( \frac{x}{K} \right) + \frac{1}{8K^6} \frac{x}{x^2 + K^2} \end{aligned}$$

we get

$$(9.20) \quad \sum_{j \in \mathbb{N}^*, j \neq k, P(j)=P(k)} \frac{1}{[(2k+1)^4 - (2j+1)^4]^2} \leq \frac{C}{k^6},$$

which gives the conclusion.  $\square$

COROLLARY 1. *There exist  $\gamma^* > 0$  and  $C > 0$  such that, for every  $\gamma_1 \in (-\gamma^*, \gamma^*)$ , for every  $k \in \mathbb{N}^*$ , we have*

$$(9.21) \quad \|\varphi_{k,\gamma_1} - \varphi_k\|_{H^s((0,1),\mathbb{R})} \leq C|\gamma_1|k^{s-1} \text{ for every integer } s \in [0, 4],$$

$$(9.22) \quad |\lambda_{k,\gamma_1} - \lambda_k| \leq C|\gamma_1|k^2,$$

$$(9.23) \quad |\sqrt{\lambda_{k,\gamma_1}} - \sqrt{\lambda_k}| \leq C|\gamma_1|.$$

*Proof of Corollary 1.* The inequalities (9.22) and (9.23) are consequences of (9.18). The inequality (9.21) for  $s = 0$  is a consequence of (9.19). By using the equation

$$\frac{d^4}{dx^4} [\varphi_{k,\gamma_1} - \varphi_k] + \gamma_1 \varphi_{k,\gamma_1}'' = \lambda_{k,\gamma_1} (\varphi_{k,\gamma_1} - \varphi_k) + (\lambda_{k,\gamma_1} - \lambda_k) \varphi_k,$$

we get (9.21) for  $s = 4$ . Indeed, we have (9.22) and

$$\|\varphi_{k,\gamma_1}''\|_{L^2((0,1),\mathbb{R})} = \sqrt{\lambda_{k,\gamma_1}} \leq Ck^2.$$

Then (9.21) for  $s = 2, 3$  comes from the logarithmic convexity of the  $H^s$ -norm.  $\square$

PROPOSITION 23. *For every  $k \in \mathbb{N}^*$ , we have*

$$(9.24) \quad \frac{d^4 \varphi_k^{(2)}}{dx^4} + \frac{d^2 \varphi_k^{(1)}}{dx^2} = \lambda_k \varphi_k^{(2)} + \lambda_k^{(1)} \varphi_k^{(1)} + \lambda_k^{(2)} \varphi_k,$$

$$(9.25) \quad \lambda_k^{(2)} = \left\langle \frac{d^2 \varphi_k^{(1)}}{dx^2}, \varphi_k \right\rangle,$$

$$(9.26) \quad \varphi_k^{(2)} = -\frac{1}{2} \|\varphi_k^{(1)}\|_{L^2}^2 \varphi_k + \sum_{j \in \mathbb{N}^*, j \neq k, P(j)=P(k)} \frac{\langle \frac{d^2 \varphi_k^{(1)}}{dx^2}, \varphi_j \rangle - \lambda_k^{(1)} x_{k,j}}{(\lambda_k - \lambda_j)} \varphi_j.$$

*There exists a constant  $C > 0$  such that, for every  $k \in \mathbb{N}^*$ ,*

$$(9.27) \quad \|\varphi_k^{(2)}\|_{L^2((0,1),\mathbb{R})} \leq \frac{C}{k^2},$$

$$(9.28) \quad |\lambda_k^{(2)}| \leq Ck.$$

*Proof of Proposition 23.* The proof of this proposition is very similar to the one of Proposition 22; thus, here, we justify only the bound (9.27). By using (9.14), we get

$$\|\varphi_k^{(2)}\|_{L^2} \leq \frac{1}{2} \|\varphi_k^{(1)}\|_{L^2}^2 + C \left( \left\| \frac{d^2 \varphi_k^{(1)}}{dx^2} \right\|_{L^2} + |\lambda_k^{(1)}| \|\varphi_k^{(1)}\|_{L^2} \right) \left[ \sum_{j=1, j \neq k, P(j)=P(k)}^{\infty} \frac{1}{[(2k+1)^4 - (2j+1)^4]^2} \right]^{1/2}.$$

Equation (9.24) gives

$$\left\| \frac{d^4 \varphi_k^{(1)}}{dx^4} \right\|_{L^2} \leq Ck^3,$$

which, with (9.19) and the convexity of the norms, leads to

$$\left\| \frac{d^2 \varphi_k^{(1)}}{dx^2} \right\|_{L^2} \leq Ck.$$

We conclude by using also (9.18), (9.19), and (9.20).  $\square$

The vectors  $\varphi_k$  and the real numbers  $\lambda_{k,\gamma}$  are analytic functions of the parameter  $\gamma$ , so we can consider their derivatives with respect to  $\gamma$ . We introduce the notations

$$\left. \frac{d^j \varphi_{k,\gamma}}{d\gamma^j} \right]_{\gamma_1}$$

for the  $j$ th derivative of the function  $\gamma \mapsto \varphi_{k,\gamma}$  evaluated at the point  $\gamma = \gamma_1$  and

$$\lambda'_{k,\gamma_1}, \lambda''_{k,\gamma_1}$$

for the first and the second derivative, respectively, of the function  $\gamma \mapsto \lambda_{k,\gamma}$  evaluated at the point  $\gamma = \gamma_1$ .

**COROLLARY 2.** *There exist  $\gamma^* > 0$ ,  $C > 0$ , and  $l \in \mathbb{R}$  such that, for every  $\gamma_1 \in (-\gamma^*, \gamma^*)$ , for every  $k \in \mathbb{N}^*$ , we have*

$$(9.29) \quad \left\| \left[ \frac{d\varphi_{k,\gamma}}{d\gamma} \right]_{\gamma_1} - \frac{d\varphi_{k,\gamma}}{d\gamma} \right\|_{H^s((0,1),\mathbb{R})} \leq C|\gamma_1|k^{s-2} \text{ for every integer } s \in [0,4],$$

$$(9.30) \quad \left\| \left[ \frac{d\varphi_{k,\gamma}}{d\gamma} \right]_{\gamma_1} \right\|_{H^s((0,1),\mathbb{R})} \leq Ck^{s-1},$$

$$(9.31) \quad \|\varphi_{k,\gamma_1} - \varphi_k - \gamma_1 \varphi_k^{(1)}\|_{L^2((0,1),\mathbb{R})} \leq C \frac{|\gamma_1|^2}{k^2},$$

$$(9.32) \quad |\lambda'_{k,\gamma_1} - \lambda'_k| \leq C|\gamma_1|k \text{ and } |\lambda'_{k,\gamma_1}| \leq Ck^2,$$

$$(9.33) \quad \left| \sqrt{\lambda_{k,\gamma_1}} - \sqrt{\lambda_k} - \gamma_1 l \right| \leq C \frac{|\gamma_1|}{k}.$$

*Proof of Corollary 2.* The proof of Corollary 2 is similar to the one of Corollary 1; thus we justify only (9.33). We deduce from (9.28) that there exist  $C > 0$  and  $\gamma^* > 0$  such that, for every  $\gamma \in (-\gamma^*, \gamma^*)$ , for every  $k \in \mathbb{N}^*$ ,

$$\left| \sqrt{\lambda_{k,\gamma}} - \sqrt{\lambda_k} - \gamma \frac{\lambda'_k}{2\sqrt{\lambda_k}} \right| \leq C \frac{\gamma^2}{k}.$$

Thus, we need only to prove the existence of constants  $C > 0$  and  $l \in \mathbb{R}$  such that, for every  $k \in \mathbb{N}^*$ ,

$$(9.34) \quad \left| \frac{\lambda'_k}{2\sqrt{\lambda_k}} - l \right| \leq \frac{C}{k}.$$

By using (9.16), (3.7), and (3.14), we get

$$\frac{\lambda'_k}{\sqrt{\lambda_k}} = -\frac{1}{\nu_k \|v_k\|_{L^2}^2} (\xi_k^2 I_1(\nu_k) + \zeta_k I_2(\nu_k) + 2\xi_k \zeta_k I_3(\nu_k)),$$

where

$$\begin{aligned}
 I_1(x) &:= \int_0^x (\sin(y) + \sinh(y))^2 dy \\
 &= \frac{\sinh(2x)}{4} + \sin(x) \cosh(x) - \cos(x) \sinh(x) - \frac{\sin(2x)}{4}, \\
 I_2(x) &:= \int_0^x (\cos(y) - \cosh(y))^2 dy \\
 &= \frac{\sinh(2x)}{4} - \cos(x) \sinh(x) - \sin(x) \cosh(x) + x + \frac{\sin(2x)}{4}, \\
 I_3(x) &:= \int_0^x (\sin(y) + \sinh(y))(-\cos(y) + \cosh(y)) dy \\
 &= \frac{\cosh(2x)}{4} - \cos(x) \cosh(x) + \frac{1}{4} + \frac{\cos(x)^2}{2}.
 \end{aligned}$$

We get (9.34) thanks to (3.8) and the asymptotic behaviors (3.13).  $\square$

In the same way as we proved the two previous propositions, we can get the next one.

PROPOSITION 24. *There exists  $C > 0$  such that, for every  $k \in \mathbb{N}^*$ ,*

$$\|\varphi_k^{(3)}\|_{L^2} \leq \frac{C}{k^2}.$$

Thus, there exist  $\gamma^* > 0$  and  $C > 0$  such that, for every  $\gamma \in (-\gamma^*, \gamma^*)$ , for every  $k \in \mathbb{N}^*$ , we have

$$(9.35) \quad \left\| \left[ \frac{d^2 \varphi_{k,\gamma}}{d\gamma^2} \right]_{\gamma_1} \right\|_{L^2} \leq \frac{C}{k^2}.$$

**9.2.2. Technical results about sequences with the same form as  $\Upsilon_{(u_0, \dot{u}_0, p)}(P)$ .** In this section, we prove bounds for the  $h^1$ - or  $h^3$ -norm of sequences  $S = (S_k)_{k \in \mathbb{N}^*}$  that have a form similar to the one of  $\Upsilon_{(u_0, \dot{u}_0, p)}(P)$ . For example,

$$S_k := \int_0^T w(t) \left\langle f(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds},$$

where  $p, w : [0, T] \rightarrow \mathbb{R}$  and  $f : [0, T] \rightarrow L^2((0, 1), \mathbb{R})$ .

LEMMA 3. *Let  $\gamma^*$  be as in Corollaries 1 and 2. There exists  $C > 0$  such that, for every  $\gamma_1 \in (-\gamma^*, \gamma^*)$ , for every  $f \in L^2((0, 1), \mathbb{R})$ ,*

$$(9.36) \quad \sum_{k=1}^{\infty} \left| k \left\langle f, \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{\gamma_1} \right|^2 \leq C \|f\|_{L^2((0,1),\mathbb{R})}^2,$$

$$(9.37) \quad \sum_{k=1}^{\infty} \left| k \langle f, \varphi_{k,\gamma_1} - \varphi_k \rangle \right|^2 \leq C \gamma_1^2 \|f\|_{L^2((0,1),\mathbb{R})}^2.$$

*Proof of Lemma 3.* Note that, in order to get (9.36), it is sufficient to prove it with  $\gamma_1 = 0$ . Indeed, by using Corollary 2, we have, for  $\gamma_1 \in (-\gamma^*, \gamma^*)$  and  $f \in L^2((0, 1), \mathbb{R})$ ,

$$\left| k \left\langle f, \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{\gamma_1} - \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_0 \right| \leq \|f\|_{L^2((0,1),\mathbb{R})} \frac{C|\gamma_1|}{k}.$$

For  $f \in L^2((0, 1), \mathbb{R})$ , we have

$$\sum_{k=1}^{\infty} \left| k \left\langle f, \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_0 \right|^2 = \sum_{k=1}^{\infty} \left| \sum_{j=1, j \neq k, P(j)=P(k)}^{\infty} a_{k,j} \langle f, \varphi_j \rangle \right|^2,$$

where the second sum is taken over all  $j \in \mathbb{N}^*$  having the same parity as  $k$  such that  $j \neq k$ ,  $a_{k,j} := kx_{k,j}$ ,  $x_{k,j}$  is defined in (9.17) when  $P(k) = P(j)$ , and  $a_{k,j} = 0$  when  $P(k) \neq P(j)$ .

If we prove that there exists  $C > 0$  such that

$$(9.38) \quad \forall j \in \mathbb{N}^* \quad \sum_{k \in \mathbb{N}^*, k \neq j, P(k)=P(j)} |a_{k,j}| \leq C \text{ and } \forall k \in \mathbb{N}^* \quad \sum_{j \in \mathbb{N}^*, j \neq k, P(j)=P(k)} |a_{k,j}| \leq C,$$

then Cauchy–Schwarz inequality provides

$$\forall (x_j)_{j \in \mathbb{N}^*} \in l^2(\mathbb{N}^*, \mathbb{C}), \quad \sum_{k=1}^{\infty} \left| \sum_{j=1}^{\infty} a_{j,k} x_j \right| \leq C^2 \sum_{j=1}^{\infty} |x_j|^2,$$

which gives the conclusion. Let us prove (9.38).

By using (3.14), (3.10), (3.17), and (9.14), we get

$$|a_{k,j}| \leq C \frac{k^3 j^2 \max\{k, j\}}{[(2k+1)^4 - (2j+1)^4]^2} \text{ when } P(k) \neq P(j).$$

Thus, we have, for every  $k \in \mathbb{N}^*$ ,

$$\begin{aligned} \sum_{j=1}^{\infty} |a_{k,j}| &\leq Ck^4 \sum_{j < k, P(j)=P(k)}^{\infty} \frac{j^2}{[(2k+1)^4 - (2j+1)^4]^2} \\ &\quad + Ck^3 \sum_{j > k, P(j)=P(k)}^{\infty} \frac{j^3}{[(2j+1)^4 - (2k+1)^4]^2} \\ &\leq Ck^4 \int_3^{2k} \frac{x^2}{[(2k+1)^4 - x^4]^2} dx + Ck^3 \int_{2k+2}^{\infty} \frac{x^3}{[(2k+1)^4 - x^4]^2} dx \leq C. \end{aligned}$$

For every  $j \in \mathbb{N}^*$ , we have

$$\begin{aligned} \sum_{k=1}^{\infty} |a_{k,j}| &\leq Cj^3 \sum_{k < j, P(j)=P(k)}^{\infty} \frac{k^3}{[(2j+1)^4 - (2k+1)^4]^2} \\ &\quad + Cj^2 \sum_{k > j, P(j)=P(k)}^{\infty} \frac{k^4}{[(2j+1)^4 - (2k+1)^4]^2} \\ &\leq Cj^3 \int_3^{2j} \frac{x^3}{[(2j+1)^4 - x^4]^2} dx + Cj^2 \int_{2j+2}^{\infty} \frac{x^4}{[(2j+1)^4 - x^4]^2} dx \leq C. \end{aligned}$$

Now let us prove (9.37). We use the decomposition

$$\varphi_{k,\gamma_1} - \varphi_k = \left( \varphi_{k,\gamma_1} - \varphi_k - \gamma_1 \frac{d\varphi_{k,\gamma}}{d\gamma} \right)_0 + \gamma_1 \frac{d\varphi_{k,\gamma}}{d\gamma} \Big|_0.$$

For the first term, we use (9.31), and, for the second one, we apply (9.36).  $\square$

LEMMA 4. Let  $\gamma^*$  be as in Corollaries 1 and 2. There exists  $C > 0$  such that, for every  $\gamma_1 \in (-\gamma^*, \gamma^*)$ , for every  $f \in H_0^2((0, 1), \mathbb{R})$ ,

$$(9.39) \quad \sum_{k=1}^{\infty} \left| k^3 \left\langle f, \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{\gamma_1} \right|^2 \leq C \|f\|_{H_0^2((0,1),\mathbb{R})}^2.$$

*Proof of Lemma 4.* First, we prove Lemma 4 for  $\gamma_1 = 0$  with the same argument as for the previous lemma. For  $f \in H_0^2((0, 1), \mathbb{R})$ , thanks to integrations by parts, we get

$$\sum_{k=1}^{\infty} \left| k^3 \left\langle f, \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_0 \right|^2 = \sum_{k=1}^{\infty} \left| \sum_{j=1}^{\infty} b_{k,j} \left\langle f'', \frac{\varphi_j''}{\sqrt{\lambda_j}} \right\rangle \right|^2,$$

where

$$b_{k,j} := \frac{k^3}{\sqrt{\lambda_j}} x_{k,j} \text{ when } P(k) = P(j), k \neq j,$$

$b_{k,j} = 0$  in the other cases, and  $x_{k,j}$  is defined in (9.17). Note that  $(\varphi_j''/\sqrt{\lambda_j})_{j \in \mathbb{N}^*}$  is an orthonormal family of  $L^2((0, 1), \mathbb{R})$ , and thus

$$\sum_{j=1}^{\infty} \left| \left\langle f'', \frac{\varphi_j''}{\sqrt{\lambda_j}} \right\rangle \right|^2 \leq \|f\|_{H_0^2((0,1),\mathbb{R})}^2.$$

We have

$$|b_{k,j}| \leq C \frac{k^5 \max\{k, j\}}{[(2k+1)^4 - (2j+1)^4]^2} \text{ when } P(k) \neq P(j).$$

We have, for every  $k \in \mathbb{N}^*$ ,

$$\begin{aligned} \sum_{j=1}^{\infty} |b_{k,j}| &\leq Ck^6 \sum_{j < k, P(j)=P(k)}^{\infty} \frac{1}{[(2k+1)^4 - (2j+1)^4]^2} \\ &\quad + Ck^5 \sum_{j > k, P(j)=P(k)}^{\infty} \frac{j}{[(2j+1)^4 - (2k+1)^4]^2} \\ &\leq Ck^6 \int_3^{2k} \frac{1}{[(2k+1)^4 - x^4]^2} dx + Ck^5 \int_{2k+2}^{\infty} \frac{x}{[(2k+1)^4 - x^4]^2} dx \leq C. \end{aligned}$$

For every  $j \in \mathbb{N}^*$ , we have

$$\begin{aligned} \sum_{k=1}^{\infty} |b_{k,j}| &\leq Cj \sum_{k < j, P(j)=P(k)}^{\infty} \frac{k^5}{[(2j+1)^4 - (2k+1)^4]^2} \\ &\quad + C \sum_{k > j, P(j)=P(k)}^{\infty} \frac{k^6}{[(2j+1)^4 - (2k+1)^4]^2} \\ &\leq Cj \int_3^{2j} \frac{x^5}{[(2j+1)^4 - x^4]^2} dx + C \int_{2j+2}^{\infty} \frac{x^6}{[(2j+1)^4 - x^4]^2} dx \leq C. \end{aligned}$$

This gives the conclusion for  $\gamma_1 = 0$ .

Now let us prove that, for  $\gamma_1 \in (-\gamma^*, \gamma^*)$  and  $f \in H_0^2((0, 1), \mathbb{C})$ , we have

$$(9.40) \quad \left| \left\langle f, \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{\gamma_1} - \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_0 \right| \leq \frac{C|\gamma_1|}{k^4} \|f\|_{H_0^2((0,1),\mathbb{R})},$$

which gives the conclusion. Thanks to the equation

$$A_{\gamma_1} \frac{d\varphi_{k,\gamma}}{d\gamma} \Big|_{\gamma_1} + \varphi_{k,\gamma_1}'' = \lambda_{k,\gamma_1} \frac{d\varphi_{k,\gamma}}{d\gamma} \Big|_{\gamma_1} + \lambda_{k,\gamma_1}' \varphi_{k,\gamma_1},$$

we have

$$\begin{aligned} \left\langle f, \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{\gamma_1} - \frac{d\varphi_{k,\gamma}}{d\gamma} \Big|_0 \right\rangle &= \frac{1}{\lambda_{k,\gamma_1}} \left\langle f'', \left( \frac{d\varphi_{k,\gamma}}{d\gamma} \right)_{\gamma_1} - \frac{d\varphi_{k,\gamma}}{d\gamma} \Big|_0 \right\rangle'' \\ &\quad + \frac{\gamma_1}{\lambda_{k,\gamma_1}} \left\langle f'', \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{\gamma_1} \\ &\quad + \frac{1}{\lambda_{k,\gamma_1}} \langle f'', \varphi_{k,\gamma_1} - \varphi_k \rangle - \frac{\lambda_{k,\gamma_1}'}{\lambda_{k,\gamma_1}} \langle f, \varphi_{k,\gamma_1} - \varphi_k \rangle \\ &\quad + \left( \frac{1}{\lambda_{k,\gamma_1}} - \frac{1}{\lambda_k} \right) \left( \left\langle f'', \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_0 \right)'' + \langle f'', \varphi_k \rangle \\ &\quad - \left( \frac{\lambda_{k,\gamma_1}'}{\lambda_{k,\gamma_1}} - \frac{\lambda_k'}{\lambda_k} \right) \langle f, \varphi_k \rangle. \end{aligned}$$

We get (9.40) by using (9.29), (9.30), (9.21), (9.22), and (9.32).  $\square$

**PROPOSITION 25.** *There exists a constant  $C > 0$  such that, for every  $w \in L^2((0, T), \mathbb{R})$  and for every  $f \in C^0([0, T], H^1((0, 1), \mathbb{R}))$ , the  $h^1(\mathbb{N}^*, \mathbb{C})$ -norm of the sequence  $(S_k)_{k \in \mathbb{N}^*}$  defined by*

$$S_k := \int_0^T w(t) \langle f(t), \varphi_k \rangle e^{-i\sqrt{\lambda_k}t} dt$$

*is bounded by*

$$C \|w\|_{L^2((0,T),\mathbb{R})} \|f\|_{C^0((0,T),H^1((0,T),\mathbb{R}))}.$$

*Proof of Proposition 25.* We introduce the function  $g \in C^0([0, T], H_0^1((0, T), \mathbb{R}))$  defined by

$$g(t, x) := f(t, x) - f(t, 0)(1 - x) - f(t, 1)x.$$

We have

$$\begin{aligned} (9.41) \quad S_k &= \int_0^T w(t) \langle g(t), \varphi_k \rangle e^{-i\sqrt{\lambda_k}t} dt \\ &\quad + \langle (1 - x), \varphi_k \rangle \int_0^T w(t) f(t, 0) e^{-i\sqrt{\lambda_k}t} + \langle x, \varphi_k \rangle \int_0^T w(t) f(t, 1) e^{-i\sqrt{\lambda_k}t}. \end{aligned}$$

Thanks to Cauchy–Schwarz inequality in  $L^2((0, T), \mathbb{R})$ , we get the following bound for the  $h^1$ -norm of the first term of the right-hand side of (9.41):

$$\|w\|_{L^2} \left( \int_0^T \sum_{k=1}^{\infty} |k \langle g(t), \varphi_k \rangle|^2 dt \right)^{1/2} \leq C \|w\|_{L^2} \|g\|_{C^0([0, T], H_0^1((0, 1), \mathbb{R}))}.$$

Thanks to (3.7), (3.8), and (3.10) we get

$$(-1)^k \langle (1-x), \varphi_k \rangle = \langle x, \varphi_k \rangle \sim \frac{C}{\nu_k},$$

which gives the conclusion.  $\square$

**PROPOSITION 26.** *Let  $\gamma^*$  be as in Corollaries 1 and 2. There exists a constant  $C > 0$  such that, for every  $p \in H^1((0, T), \mathbb{R})$ , with  $\|p\|_{L^\infty((0, T), \mathbb{R})} < \gamma^*$ , for every  $w \in L^2((0, T), \mathbb{R})$ , and for every  $f \in C^0([0, T], H^1((0, 1), \mathbb{R}))$ , the  $h^3(\mathbb{N}^*, \mathbb{C})$ -norm of the sequences  $(S_k)_{k \in \mathbb{N}^*}$ ,  $(\tilde{S}_k)_{k \in \mathbb{N}^*}$  defined by*

$$S_k := \int_0^T w(t) \left\langle f(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{-i\sqrt{\lambda_k}t} dt,$$

$$\tilde{S}_k := \int_0^T w(t) \left\langle f(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{-i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt$$

are bounded by

$$C \|w\|_{L^2((0, T), \mathbb{R})} \|f\|_{C^0([0, T], H^1((0, T), \mathbb{R}))},$$

and, moreover,

$$\|S - \tilde{S}\|_{h^3} \leq C \|p\|_{L^\infty} \|w\|_{L^2} \|f\|_{C^0([0, T], L^2)}.$$

*Proof of Proposition 26.* Let us consider the decomposition  $S_k = S^1 + S^2 + S^3$ , where, for every  $k \in \mathbb{N}^*$ ,

$$(9.42) \quad \begin{aligned} S_k^1 &:= \int_0^T w(t) \left( \frac{1}{\sqrt{\lambda_{k,p(t)}}} - \frac{1}{\sqrt{\lambda_k}} \right) \langle f(t), \varphi_{k,p(t)} \rangle e^{i\sqrt{\lambda_k}t} dt, \\ S_k^2 &:= \frac{1}{\sqrt{\lambda_k}} \int_0^T w(t) \langle f(t), \varphi_{k,p(t)} - \varphi_k \rangle e^{i\sqrt{\lambda_k}t} dt, \\ S_k^3 &:= \frac{1}{\sqrt{\lambda_k}} \int_0^T w(t) \langle f(t), \varphi_k \rangle e^{i\sqrt{\lambda_k}t} dt. \end{aligned}$$

Thanks to (9.23) we have

$$\left| \frac{1}{\sqrt{\lambda_{k,p(t)}}} - \frac{1}{\sqrt{\lambda_k}} \right| \leq \frac{C|p(t)|}{k^4}.$$

Thus, thanks to the Cauchy–Schwarz inequality, the  $h^3(\mathbb{N}^*, \mathbb{C})$ -norm of  $S^1$  is bounded by

$$C \|w\|_{L^2} \|p\|_{L^\infty} \|f\|_{C^0([0, T], L^2)} \leq C \|w\|_{L^2} \|f\|_{C^0([0, T], L^2)}.$$



Thanks to Lemma 3, the  $h^3(\mathbb{N}^*, \mathbb{C})$ -norm of  $S^2$  is bounded by

$$C\|w\|_{L^2}\|p\|_{L^\infty}\|f\|_{C^0([0,T],L^2)} \leq C\|w\|_{L^2}\|f\|_{C^0([0,T],L^2)}.$$

We conclude the study of  $S$  by applying Proposition 25 to  $S^3$ .

By using (see Corollary 1)

$$(9.43) \quad \sqrt{\lambda_{k,p(s)}} = \sqrt{\lambda_k} + p(s)l + \epsilon_k(s), \text{ where } l \in \mathbb{R} \text{ and } |\epsilon_k(s)| \leq C \frac{\|p\|_{L^\infty}}{k},$$

we get  $\tilde{S} - S = \delta S^1 + \delta S^2$ , where, for every  $k \in \mathbb{N}^*$ ,

$$(9.44) \quad \begin{aligned} \delta S_k^1 &:= \int_0^T w(t) (e^{il \int_0^t p(s) ds} - 1) \left\langle f(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i\sqrt{\lambda_k}t} dt \\ \delta S_k^2 &:= \int_0^T w(t) e^{il \int_0^t p(s) ds} \left\langle f(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i\sqrt{\lambda_k}t} [e^{i \int_0^t \epsilon_k(s) ds} - 1] dt. \end{aligned}$$

The first part of Proposition 26 gives the following bound for the  $h^3(\mathbb{N}^*, \mathbb{C})$ -norm of  $\delta S^1$ :

$$C\|w\|_{L^2}\|p\|_{L^\infty}\|f\|_{C^0([0,T],H^1)}.$$

Thanks to Cauchy-Schwarz inequality and the bound on  $\epsilon_k$  given in (9.43), we get the following bound for the  $h^3(\mathbb{N}^*, \mathbb{C})$ -norm of  $\delta S^2$ :

$$C\|w\|_{L^2}\|p\|_{L^\infty}\|f\|_{C^0([0,T],L^2)}. \quad \square$$

**9.3. Proof of Proposition 20.** In all of this section,  $(u_0, \dot{u}_0, p) \in E_8$  is fixed, and we use the notations

$$(9.45) \quad \begin{aligned} \delta_4 &:= \|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_4}, \\ \delta_6 &:= \|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_6}, \\ \delta_8 &:= \|(u_0 - u_0^{ref}, \dot{u}_0 - \dot{u}_0^{ref}, p)\|_{E_8}. \end{aligned}$$

We assume that  $\delta_4 \in [0, 1]$  and that  $\delta_4$  is small enough so that  $\|p\|_{L^\infty((0,T),\mathbb{R})} < \gamma^*$ , where  $\gamma^*$  is given in Corollaries 1 and 2. We write  $\Upsilon^j$  instead of  $\Upsilon_{(u_0, \dot{u}_0, p)}^j(P)$ .

*Proof of Proposition 20 for  $j = 1$ .* For the study of  $\Upsilon^1$  in  $h^3(\mathbb{N}^*, \mathbb{C})$ , we consider the decomposition

$$(9.46) \quad \begin{aligned} \Upsilon^1(P)_k &= \int_0^T P(t) \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle [e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} - e^{i\sqrt{\lambda_k}t}] dt \\ &+ \int_0^T P(t) \left( \frac{1}{\sqrt{\lambda_{k,p(t)}}} - \frac{1}{\sqrt{\lambda_k}} \right) \langle u_{xx}(t), \varphi_{k,p(t)} \rangle e^{i\sqrt{\lambda_k}t} dt \\ &+ \frac{1}{\sqrt{\lambda_k}} \int_0^T P(t) \langle u_{xx}(t), \varphi_{k,p(t)} - \varphi_k \rangle e^{i\sqrt{\lambda_k}t} dt \\ &+ \int_0^T P(t) \left\langle (u - u^{ref})_{xx}(t), \frac{\varphi_k}{\sqrt{\lambda_k}} \right\rangle e^{i\lambda_k t} dt. \end{aligned}$$

Thanks to Proposition 26, we have the following bound for the  $h^3(\mathbb{N}^*, \mathbb{C})$ -norm of the first term of the right-hand side of (9.46):

$$C\|P\|_{L^2}\|p\|_{L^\infty}\|u\|_{C^0([0,T],H^2)} \leq C\delta_4(1+\delta_4)\|P\|_{L^2} \leq C\delta_4\|P\|_{L^2}.$$

Thanks to (9.23) and Proposition 1, the  $h^3(\mathbb{N}^*, \mathbb{C})$ -norm of the second term of the right-hand side of (9.46) is bounded by

$$C\|P\|_{L^2}\|p\|_{L^\infty}\|u\|_{C^0([0,T],H^2)} \leq C\delta_4(1+\delta_4)\|P\|_{L^2} \leq C\delta_4\|P\|_{L^2}.$$

Thanks to Lemma 3 and Proposition 1, we have the following bound for the  $h^3(\mathbb{N}^*, \mathbb{C})$ -norm of the third term of the right-hand side of (9.46):

$$C\|P\|_{L^2}\|p\|_{L^\infty}\|u\|_{C^0([0,T],H^2)} \leq C\delta_4(1+\delta_4)\|P\|_{L^2} \leq C\delta_4\|P\|_{L^2}.$$

Thanks to Propositions 25 and 2, we have the following bound for the  $h^3(\mathbb{N}^*, \mathbb{C})$ -norm of the last term of the right-hand side of (9.46):

$$C\|P\|_{L^2}\|(u - u^{ref})_{xx}\|_{C^0([0,T],H^1)} \leq C\|P\|_{L^2}\delta_4.$$

In conclusion, we have proved that

$$\|\Upsilon^1\|_{h^3} \leq C\delta_4\|P\|_{L^2}.$$

In order to study  $\Upsilon^1$  in  $h^7$ , first we study it in  $h^5(\mathbb{N}^*, \mathbb{C})$ . By using an integration by parts, one gets

$$\begin{aligned} (9.47) \quad -\Upsilon_k^1 = & \frac{1}{i\sqrt{\lambda_k}} \int_0^T \dot{P}(t) \left( \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} - \langle u_{xx}^{ref}(t), \varphi_k \rangle e^{i\lambda_k t} \right) dt \\ & + \int_0^T \left( \frac{1}{i\sqrt{\lambda_{k,p(t)}}} - \frac{1}{i\sqrt{\lambda_k}} \right) \dot{P}(t) \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ & + \frac{1}{i\sqrt{\lambda_k}} \int_0^T P(t) \left( \left\langle \dot{u}_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} - \left\langle \dot{u}_{xx}^{ref}(t), \frac{\varphi_k}{\sqrt{\lambda_k}} \right\rangle e^{i\lambda_k t} \right) dt \\ & + \int_0^T \left( \frac{1}{i\sqrt{\lambda_{k,p(t)}}} - \frac{1}{i\sqrt{\lambda_k}} \right) P(t) \left\langle \dot{u}_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ & + i \int_0^T P(t) \dot{p}(t) \frac{\lambda'_{k,p(t)}}{\lambda_{k,p(t)}^2} \langle u_{xx}(t), \varphi_{k,p(t)} \rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ & + \int_0^T P(t) \dot{p}(t) \frac{1}{i\lambda_{k,p(t)}} \left\langle u_{xx}(t), \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{p(t)} e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt. \end{aligned}$$

Thanks to the study of  $\Upsilon^1$  in  $h^3(\mathbb{N}^*, \mathbb{C})$ , we have the following bound for the  $h^5(\mathbb{N}^*, \mathbb{C})$ -norm of the first (resp., third) term of the right-hand side of (9.47):

$$C\|P\|_{H_0^1}\delta_4 \text{ (resp., } C\|P\|_{L^2}\delta_6).$$

Thanks to (9.23), we get the following bound for the  $h^5(\mathbb{N}^*, \mathbb{C})$ -norm of the second (resp., fourth) term of the right-hand side of (9.47):

$$C\|P\|_{H_0^1}\delta_4 \text{ (resp., } C\|P\|_{L^2}\delta_4).$$

Thanks to (9.32), we get the following bound for the  $h^5(\mathbb{N}^*, \mathbb{C})$ -norm of the fifth term of the right-hand side of (9.47):

$$C\|P\|_{L^2}\|p\|_{H^2}\|u\|_{C^0([0,T],H^2)} \leq C\|P\|_{L^2}\delta_6.$$

Thanks to Lemma 3, we get the following bound for the  $h^5(\mathbb{N}^*, \mathbb{C})$ -norm of the last term of the right-hand side of (9.47):

$$C\|P\|_{L^2}\delta_4.$$

In conclusion, we have proved that

$$\|\Upsilon^1\|_{h^5(\mathbb{N}^*, \mathbb{C})} \leq C[\|P\|_{L^2}\delta_6 + \|P\|_{H_0^1}\delta_6].$$

Finally, let us study  $\Upsilon^1$  in  $h^7(\mathbb{N}^*, \mathbb{C})$ . Thanks to the study of  $\Upsilon^1$  in  $h^5(\mathbb{N}^*, \mathbb{C})$ , we have the following bound for the  $h^7(\mathbb{N}^*, \mathbb{C})$ -norm of the first (resp., third) term of the right-hand side of (9.47):

$$C[\|P\|_{H_0^1}\delta_6 + \|P\|_{H_0^2}\delta_4] \text{ (resp., } C[\|P\|_{L^2}\delta_8 + \|P\|_{H_0^1}\delta_6]).$$

The study of the second and fourth terms of the right-hand side of (9.47) can be done in the same way. Let us work on the first one. We use (9.43). We have (9.48)

$$\begin{aligned} & \int_0^T \frac{lp(t) + \epsilon_k(t)}{\sqrt{\lambda_{k,p(t)}}\lambda_k} \dot{P}(t) \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt = \\ & \frac{l}{\lambda_k} \int_0^T \dot{P}(t)p(t) \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ & + \frac{l}{\sqrt{\lambda_k}} \int_0^T \left( \frac{1}{\sqrt{\lambda_{k,p(t)}}} - \frac{1}{\sqrt{\lambda_k}} \right) \dot{P}(t)p(t) \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ & + \int_0^T \frac{\epsilon_k(s)}{\sqrt{\lambda_k}\sqrt{\lambda_{k,p(t)}}} \dot{P}(t) \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt. \end{aligned}$$

Thanks to Proposition (26), we get the following bound for the  $h^7(\mathbb{N}^*, \mathbb{C})$ -norm of the first term of the right-hand side of (9.48):

$$C\|P\|_{H_0^1}\|p\|_{L^\infty}\|u\|_{C^0([0,T],H^3)} \leq C\delta_4(1 + \delta_4)\|P\|_{H_0^1} \leq C\delta_4\|P\|_{H_0^1}.$$

By using (9.23), we get the following bound for the  $h^7(\mathbb{N}^*, \mathbb{C})$ -norm of the second term of the right-hand side of (9.48):

$$C\|P\|_{H_0^1}\|p\|_{L^\infty}\|u\|_{C^0([0,T],H^2)} \leq C\delta_4\|P\|_{H_0^1}.$$

Thanks to the asymptotic behavior of  $\epsilon_k$ , we get the same bound for the  $h^7(\mathbb{N}^*, \mathbb{C})$ -norm of the last term of the right-hand side of (9.48).

Finally, we get the following bound for the  $h^7(\mathbb{N}^*, \mathbb{C})$ -norm of the second (resp., fourth) term of the right-hand side of (9.47):

$$C\|P\|_{H_0^1}\delta_4 \text{ (resp., } C\|P\|_{L^2}\delta_6).$$

For the study of the fifth term of the right-hand side of (9.47), we consider the decomposition

$$\begin{aligned} & \int_0^T P(t) \dot{p}(t) \frac{\lambda'_{k,p(t)}}{\lambda_{k,p(t)}^2} \langle u_{xx}(t), \varphi_{k,p(t)} \rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ &= \frac{\lambda'_k}{\lambda_k^{3/2}} \int_0^T P(t) \dot{p}(t) \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ &+ \int_0^T P(t) \dot{p}(t) \left( \frac{\lambda'_{k,p(t)}}{\lambda_{k,p(t)}^{3/2}} - \frac{\lambda'_k}{\lambda_k^{3/2}} \right) \left\langle u_{xx}(t), \frac{\varphi_{k,p(t)}}{\sqrt{\lambda_{k,p(t)}}} \right\rangle e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt. \end{aligned}$$

We study the first term of this decomposition with Proposition 26 and (9.32) and the second one with (9.32); then we get the following bound for the  $h^7(\mathbb{N}^*, \mathbb{C})$ -norm of the fifth term of the right-hand side of (9.47):

$$C \|P\|_{L^2} \delta_6.$$

For the last term of the right-hand side of (9.47), we cannot apply Lemma 4 because  $u_{xx} \notin H_0^2((0, 1), \mathbb{R})$ ; thus, we perform another integration by parts:

$$\begin{aligned} & -i \int_0^T P(t) \dot{p}(t) \frac{1}{\lambda_{k,p(t)}} \left\langle u_{xx}(t), \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{p(t)} e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ &= \int_0^T [\dot{P}(t) \dot{p}(t) + P(t) \ddot{p}(t)] \frac{1}{(\lambda_{k,p(t)})^{3/2}} \left\langle u_{xx}(t), \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{p(t)} e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ (9.49) \quad &+ \int_0^T P(t) \dot{p}(t) \frac{1}{(\lambda_{k,p(t)})^{3/2}} \left\langle \dot{u}_{xx}(t), \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{p(t)} e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ &- \int_0^T P(t) \dot{p}(t)^2 \frac{3\lambda'_{k,p(t)}}{2(\lambda_{k,p(t)})^{5/2}} \left\langle u_{xx}(t), \frac{d\varphi_{k,\gamma}}{d\gamma} \right\rangle_{p(t)} e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt \\ &+ \int_0^T P(t) \dot{p}(t)^2 \frac{1}{(\lambda_{k,p(t)})^{3/2}} \left\langle u_{xx}(t), \frac{d^2\varphi_{k,\gamma}}{d\gamma^2} \right\rangle_{p(t)} e^{i \int_0^t \sqrt{\lambda_{k,p(s)}} ds} dt. \end{aligned}$$

We study the first two terms of the right-hand side of (9.49) thanks to Lemma 3 and the last two thanks to (9.32) and (9.35). In conclusion, we have proved that

$$\|\Upsilon^1\|_{h^7(\mathbb{N}^*, \mathbb{C})} \leq C[\|P\|_{L^2} \delta_8 + \|P\|_{H_0^1} \delta_6 + \|P\|_{H_0^2} \delta_4],$$

and the logarithmic convexity of the norms justifies that

$$\|\Upsilon^1\|_{h^7(\mathbb{N}^*, \mathbb{C})} \leq C[\|P\|_{L^2} \delta_8 + \|P\|_{H_0^2} \delta_4]. \quad \square$$

*Proof of Proposition 20 for  $j = 2, 3$ .* Lemma 4 gives the bound in  $h^3(\mathbb{N}^*, \mathbb{C})$  for  $\Upsilon^2$ . As in the previous proof, one can study  $\Upsilon^2$  in  $h^5$  thanks to an integration by parts. Then we deduce the bound in  $h^7(\mathbb{N}^*, \mathbb{C})$ . The study of  $\Upsilon^3$  can be done in the same way.  $\square$

**9.4. Proof of Proposition 21.** The goal of this section is to sketch the proof of Proposition 21. Let  $((u_0^\epsilon, \dot{u}_0^\epsilon, p^\epsilon))_{\epsilon>0}$ ,  $(u_0, \dot{u}_0, p) \in E_8$  such that  $(u_0^\epsilon, \dot{u}_0^\epsilon, p^\epsilon) \rightarrow (u_0, \dot{u}_0, p)$  in  $E_8$  and  $P \in H_0^2((0, T), \mathbb{R})$ . By doing again for

$$\Upsilon_{(u_0^\epsilon, \dot{u}_0^\epsilon, p^\epsilon)}(P) - \Upsilon_{(u_0, \dot{u}_0, p)}(P)$$

the same analysis we did for

$$\Upsilon_{(u_0, \dot{u}_0, p)}(P) - \Upsilon_{(u_0^{ref}, \dot{u}_0^{ref}, 0)}(P)$$

in the previous subsection, one can prove that

$$\Upsilon_{(u_0^\epsilon, \dot{u}_0^\epsilon, p^\epsilon)}(P) \rightarrow \Upsilon_{(u_0, \dot{u}_0, p)}(P) \text{ in } h^7(\mathbb{N}^*, \mathbb{C}) \text{ when } \epsilon \rightarrow 0.$$

**10. Remarks, conjectures, prospects.** In Theorem 1, the regularity assumption  $H_{(0)}^{5+\epsilon} \times H_{(0)}^{3+\epsilon}((0, 1), \mathbb{R})$ , with  $\epsilon > 0$ , is technical and related to the use of the Nash–Moser theorem. We conjecture that  $(\Sigma)$  is controllable

- in  $H_{(0)}^3 \times H_{(0)}^1((0, 1), \mathbb{R})$  with control functions in  $L_{loc}^2(\mathbb{R}, \mathbb{R})$ ,
- in  $H_{(0)}^5 \times H_{(0)}^3((0, 1), \mathbb{R})$  with control functions in  $H_{loc}^1(\mathbb{R}, \mathbb{R})$ ,
- in  $H_{(0)}^7 \times H_{(0)}^5((0, 1), \mathbb{R})$  with control functions in  $H_{loc}^2(\mathbb{R}, \mathbb{R})$ , etc.,

because it is the case for the linearized system studied in section 3.

Theorem 1 provides the local controllability in time  $T := 8/\pi$ . This choice ( $T := 8/\pi$ ) is also technical. The existence of a minimal time for the controllability of this system is an open problem.

Most probably, the proof presented in this article also works for the proof of the local controllability of the same system around

$$\varphi_k(x) \left\{ \begin{array}{c} \cos(\sqrt{\lambda_k}t) \\ \text{or} \\ \sin(\sqrt{\lambda_k}t) \end{array} \right\} + \varphi_j(x) \left\{ \begin{array}{c} \cos(\sqrt{\lambda_j}t) \\ \text{or} \\ \sin(\sqrt{\lambda_j}t) \end{array} \right\}$$

when  $k$  and  $j$  are integers with different parities. The same argument should also be adaptable to the beam equation with follower loads

$$\left\{ \begin{array}{l} u_{tt} + u_{xxxx} + p(t)u_{xx} = 0, (t, x) \in \mathbb{R}_+ \times (0, 1), \\ u = u_x = 0 \text{ at } x = 0, \\ u_{xx} = u_{xxx} = 0 \text{ at } x = 1, \end{array} \right.$$

which is also proved to be not controllable in  $H^2 \times L^2((0, 1), \mathbb{R})$  with  $L_{loc}^r(\mathbb{R}, \mathbb{R})$ -control functions, in [1].

Since the coefficients  $\langle \varphi_1'', \varphi_k \rangle$  vanish when  $k$  is even, the linearized system of  $(\Sigma)$  around the trajectory  $(u(t, x) = \varphi_1(x) \sin(\sqrt{\lambda_1}t), p \equiv 0)$  is not controllable. In order to prove the local controllability of  $(\Sigma)$  around this trajectory, the return method would probably work, as in [3]. This method was introduced by Coron in [5] in order to solve a stabilization problem. It has been used in order to get controllability results for partial differential equations by Coron in [8], [6], [7], by Coron and Fursikov in [9], by Fursikov and Imanuvilov in [11], by Glass in [12], [14], [16], [13], [15], [17], [18], and by Horsin in [21]; see also the book [10] by Coron.

The strategy developed in [4] could be used in order to prove steady-state controllability results of the type: For every  $k, l \in \mathbb{N}^*$ , there exists  $T > 0$  and  $p \in H_0^1((0, T), \mathbb{R})$  such that the solution of  $(\Sigma)$  with  $u(0) = \varphi_k$  and control  $p$  satisfies  $u(T) = \varphi_l$ .

**Acknowledgments.** The author thanks J.-M. Coron for fruitful discussions and advice on this work, E. Zuazua for having attracted her attention to this controllability problem, and O. Glass for interesting remarks.

## REFERENCES

- [1] J. M. BALL, J. E. MARSDEN, AND M. SLEMROD, *Controllability for distributed bilinear systems*, SIAM J. Control Optim., 20 (1982), pp. 575–597.
- [2] K. BEAUCHARD, *Controllability of a quantum particle in a 1D variable domain*, ESAIM Control Optim. Calc. Var., 14 (2008), pp. 105–147.
- [3] K. BEAUCHARD, *Local controllability of a 1D Schrödinger equation*, J. Math. Pures Appl., 9 (2005), pp. 851–956.
- [4] K. BEAUCHARD AND J.-M. CORON, *Controllability of a quantum particle in a moving potential well*, J. Funct. Anal., 232 (2006), pp. 328–389.
- [5] J.-M. CORON, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.
- [6] J.-M. CORON, *Contrôlabilité exacte frontière de l'équation d'Euler des fluides parfaits incompressibles bidimensionnels*, C. R. Acad. Sci. Paris, 317 (1993), pp. 271–276.
- [7] J.-M. CORON, *On the controllability of the 2-D incompressible perfect fluids*, J. Math. Pures Appl., 75 (1996), pp. 155–188.
- [8] J.-M. CORON, *Local controllability of a 1-D tank containing a fluid modeled by the shallow water equations*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 513–554.
- [9] J.-M. CORON AND A. FURSIKOV, *Global exact controllability of the 2D Navier-Stokes equations on a manifold without boundary*, Russian J. Math. Phys., 4 (1996), pp. 429–448.
- [10] J.-M. CORON, *Control and nonlinearity*, Math. Surveys Monogr., 136 (2007), 427 pp.
- [11] A. V. FURSIKOV AND O. YU. EMANUILOV, *Exact controllability of the Navier-Stokes and Boussinesq equations*, Uspekhi Mat. Nauk, 54 (1999), pp. 93–146 (in Russian); Russian Math. Surveys, 54 (1999), pp. 565–618 (in English).
- [12] O. GLASS, *Exact boundary controllability of 3D-Euler equation*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 1–44.
- [13] O. GLASS, *An addendum to a J.-M. Coron theorem concerning the controllability of the Euler system for 2D incompressible inviscid fluids*, J. Math. Pures Appl., 80 (2001), pp. 845–877.
- [14] O. GLASS, *On the controllability of the Vlasov Poisson system*, J. Differential Equations, 195 (2003), pp. 332–379.
- [15] O. GLASS, *Asymptotic stabilizability by stationary feedback of the two-dimensional Euler equation: The multiconnected case*, SIAM J. Control Optim., 44 (2005), pp. 1105–1147.
- [16] O. GLASS, *Controllability and Asymptotic Stabilization of the Camassa-Holm Equation*, preprint, 2007.
- [17] O. GLASS, *On the controllability of the 1-D isentropic Euler equation*, J. Eur. Math. Soc. (JEMS), 9 (2007), pp. 427–486.
- [18] O. GLASS AND S. GUERRERO, *On the uniform controllability of the Burgers equation*, SIAM J. Control Optim., 46 (2007), pp. 1211–1238.
- [19] M. GROMOV, *Partial Differential Relations*, Springer-Verlag, Berlin, 1986.
- [20] L. HÖRMANDER, *On the Nash-Moser implicit function theorem*, Ann. Acad. Sci. Fenn. Ser. A I Math., 10 (1985), pp. 255–259.
- [21] T. HORSIN, *On the controllability of the Burgers equation*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 83–95.
- [22] W. KRABS, *On Moment Theory and Controllability of One-Dimensional Vibrating Systems and Heating Processes*, Springer-Verlag, Berlin, 1992.

## STOCHASTIC CONTROL WITH IMPERFECT MODELS\*

ARNAB BASU<sup>†</sup> AND VIVEK S. BORKAR<sup>‡</sup>

**Abstract.** We consider the problem of worst case performance estimation for a stochastic dynamic model in the presence of model uncertainty. This is cast as a nonclassical controlled diffusion problem. An infinite dimensional linear programming formulation is given and its dual is derived. The dual is successively approximated on a bounded domain by a semi-infinite and a finite linear program. This uses function approximation based on a reproducing kernel Hilbert space. Error analysis for the approximation is provided along with an estimate of the sample complexity.

**Key words.** worst case performance, controlled diffusion, approximation of linear programs, reproducing kernel Hilbert spaces, sample complexity, credit risk

**AMS subject classifications.** Primary, 93E20; Secondary, 93E25, 90C48

**DOI.** 10.1137/060667049

**1. Introduction.** An important problem in finance as well as several other areas is that of estimating risk (or more generally, an appropriate performance index) under model uncertainty. We consider this problem in the context of an underlying state process that is a continuous time diffusion. We take the “worst case” approach, that is, we estimate the minimal performance ( $\approx$  negative of maximal risk in risk estimation) over the allowed class of models.

We identify the following two basic forms of uncertainties (see section 2):

- The first is the *modelling uncertainty* wherein the drift and diffusion coefficients of the diffusion are not exactly known. We model this by introducing a hypothetical control process. This changes it to a controlled diffusion model. We then minimize the performance over the allowed control processes. With finance applications in mind, we consider this problem with additional constraints, thereby making it a “constrained” control problem [9].
- The second form of uncertainty has to do with *unmodelled dynamics*. Here we assume that there may be certain state variables about whose dynamics something is known, but these are not observed. On the other hand, certain other state variables are observed and are modelled *separately* by Markov diffusions, but there is uncertainty about their dependence structure. We call this scenario “uncertainty in dynamics.” This formulation is explicitly motivated by problems arising in credit risk (see, e.g., [22, Chap. 9]), where stochastic dynamic models for two or more processes are separately available, but their dependence structure is unknown.

Combining both of these, we cast the problem as an abstract linear program over appropriately defined “occupation measures” (see section 3). This is an infinite

---

\*Received by the editors August 8, 2006; accepted for publication (in revised form) November 21, 2007; published electronically March 21, 2008.

<http://www.siam.org/journals/sicon/47-3/66704.html>

<sup>†</sup>Indian Institute of Management, Bannerghatta Road, Bangalore 560076, India (arnab.basu@iimb.ernet.in). This work was done when this author was with the School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India. This author’s research was supported by an Infosys research fellowship.

<sup>‡</sup>School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India (borkar@tifr.res.in). This author’s research was supported by grant III.5(157)/99-ET from the Department of Science and Technology, Government of India.

dimensional linear program over a space of functions with an unbounded domain. Motivated by computational considerations, we restrict these functions to a bounded domain. The resulting approximate linear program is then successively approximated further first by a semi-infinite linear program and then by a finite linear program using a specific function approximation scheme based on reproducing kernel Hilbert spaces (RKHS). Error estimates for the approximation (see section 4) provide a rigorous justification for its use. Section 5 of this paper deals with the calculation of the sample complexity required to achieve a specified tolerance on the approximation error for RKHS functions (as described above) defined on a compact domain with a smooth boundary. Section 6 sketches an extension to Markov-modulated diffusions, which capture the regime-switching phenomena.

This work is purely theoretical. In a related work [3], we also present some preliminary computational results for the simplest case of modelling uncertainty alone. To put this work in context, we briefly mention some relevant literature that it builds upon. The constrained nature of the above problem and the unmodelled dynamics make the problem inconvenient for a dynamic programming based approach. This suggests a linear programming formulation. The linear programming approach to discrete time or state stochastic control is by now classical [21] and its extensions to continuous state space have been extensively developed in [8], [15], [16], and [17]. Similar developments are available in continuous time as well (see [5], [13], [28]).

Another important antecedent to our work is the work on the celebrated Monge–Kantorovich optimal mass transportation problem [2]. Here the objective is to maximize or minimize the expected value of a function of two random variables over all possible joint distributions thereof, subject to the constraint that they have prescribed marginals. Our formulation can be viewed as a dynamic version of this with the additional provision for constraints and unmodelled components, as will become clear later.

As these are perforce infinite dimensional problems, considerable attention has been devoted in recent times to approximation issues (see [14], [18], [19], [24]). Our approach is that of [18], [19], except that we specialize to a specific form of function approximation: We approximate functions by finite linear combinations of translates of a  $C^\infty$  Mercer kernel. This procedure is justified by the theory of RKHS associated with such kernels. These functions are attractive on many grounds. First and foremost, the “best” approximation turns out to be simply a linear combination of the *same* kernel function centered at the sampled points. This offers tremendous computational advantages. This fact has driven the extensive use of RKHS in signal processing. Further, a rich theory recently developed in statistical learning theory literature (wherein the applications of RKHS are motivated by similar considerations) allows us to obtain precise error estimates (see [10], [25], [26], [27], [30]).

To summarize, the main contributions of this article are the following:

- A novel and comprehensive formulation of the risk/performance estimation problem under uncertainty;
- conversion thereof to an infinite dimensional linear program that subsumes classical linear programming formulations of stochastic control;
- a rigorous approximation theory for the same.

We qualify the last of the above by observing that our error analysis for the approximation yields only “order of magnitude” estimates of the kind prevalent in theoretical computer science, not the tight bounds that numerical analysts love. Nevertheless, we feel that this is an important first step, as such bounds did not seem to be around



until now. Furthermore, our approach charts out a step-by-step procedure for error analysis, many features of which are bound to have a much broader applicability.

**2. Problem formulation.** Our underlying state process is a  $d$ -dimensional diffusion process  $x(\cdot) = [x_1(\cdot), \dots, x_d(\cdot)]^T$  described by the stochastic differential equation (SDE)

$$(1) \quad dx(t) = m^*(x(t))dt + \sigma^*(x(t))dW(t), \quad x(0) = x_0, \quad t \geq 0,$$

where  $W(\cdot)$  is a standard Brownian motion in  $\mathcal{R}^d$  and the drift vector  $m^*(\cdot) : \mathcal{R}^d \rightarrow \mathcal{R}^d$  and the diffusion matrix  $\sigma^*(\cdot) : \mathcal{R}^d \rightarrow \mathcal{R}^{d \times d}$  are assumed to be Lipschitz. This ensures the well-posedness of (1), but can be relaxed in special cases. (Alternatively, one may invoke the “genericity” arguments of [20].) The law of  $x_0$  is fixed, say,  $\phi_0$ . The two forms of uncertainty we consider are captured by the following two basic formulations.

**Problem 1: Model uncertainty.** Here we suppose that  $m^*, \sigma^*$  are not known exactly, but only up to a certain approximation. The latter is captured by the conditions

$$(2) \quad m^*(x) \in m(x, U), \quad (\sigma^*(\sigma^*)^T)(x) \in (\sigma\sigma^T)(x, U) \quad \forall x,$$

where

- $U$  is a prescribed compact metric space;
- $m(\cdot, \cdot) = [m_1(\cdot, \cdot), \dots, m_d(\cdot, \cdot)]^T : \mathcal{R}^d \times U \rightarrow \mathcal{R}^d$  is continuous and Lipschitz in the first argument uniformly w.r.t. the second, and, in addition,  $m(x, U)$  is convex for each  $x$ ;
- $\sigma(\cdot, \cdot) = [[\sigma_{ij}(\cdot, \cdot)]]_{1 \leq i, j \leq d} : \mathcal{R}^d \times U \rightarrow \mathcal{R}^{d \times d}$  is continuous and Lipschitz in the first argument uniformly w.r.t. the second, and, in addition,  $(\sigma\sigma^T)(x, U)$  is convex for each  $x$ . Also, the *nondegeneracy* condition holds, i.e., there exists a constant  $c > 0$  such that for all  $(x, u) \in \mathcal{R}^d \times U$  and  $\xi \in \mathcal{R}^d$ ,

$$(3) \quad \xi^T \sigma \sigma^T(x, u) \xi \geq c|\xi|^2.$$

Familiar “measurable selection” arguments from stochastic control (see, e.g., [6, Chaps. I–II]) then allow us to replace (1) by

$$(4) \quad dx(t) = m(x(t), u'(t))dt + \sigma(x(t), u''(t))dW(t),$$

where  $u'(t), u''(t)$  are  $U$ -valued processes of the form  $v'(x(t)), v''(x(t))$ ,  $t \geq 0$ , for some measurable  $v', v'' : \mathcal{R}^d \rightarrow U$ . One can be a bit more ambitious and allow for “model drift,” thus permitting  $u'(t), u''(t)$  of the form  $v'(x(t), t), v''(x(t), t)$  for measurable  $v', v'' : \mathcal{R}^d \times \mathcal{R}^+ \rightarrow U$ . This in fact will also be more convenient, as explicit time dependence is hard to suppress in our formulation. The final liberty we take is to permit more general arbitrary nonanticipative  $u'(\cdot), u''(\cdot)$ , i.e., satisfying that  $u'(\cdot), u''(\cdot)$  have measurable paths and that for any  $t \geq 0$ ,  $W(t + \cdot) - W(t)$  is independent of  $x_0, u'(s), u''(s), W(s), s \leq t$ . But this relaxation is no relaxation at all, as we shall soon observe. We shall refer to  $[u'(\cdot), u''(\cdot)]$  of the form  $[v'(x(\cdot), \cdot), v''(x(\cdot), \cdot)]$  as *Markov controls*.

Fix the time horizon  $T > 0$  and let  $Z \stackrel{\text{def}}{=} [0, T] \times \mathcal{R}^d$ . Let  $u(t) = [u'(t), u''(t)]$ , viewed as a  $U^2$ -valued process. Further, by abuse of terminology, we shall replace  $U$  by  $U^2$ , using the former symbol to mean the latter, and write  $m(x, u), \sigma(x, u)$  for  $m(x, u'), \sigma(x, u'')$ , respectively, with  $u = [u', u'']$ . For a Polish space  $S$ , let  $\mathcal{P}(S)$  denote the Polish space of probability measures on  $S$  with the Prohorov topology (see [7, Chap. 2]). Associated with (4) is the *occupation measure*  $\nu \in \mathcal{P}(Z \times U)$  defined by

$$(5) \quad \int_{Z \times U} f d\nu \stackrel{\text{def}}{=} \frac{1}{T} E \left[ \int_0^T f(t, x(t), u(t)) dt \right], \quad f \in C_b(Z \times U).$$

Associated with  $x(\cdot)$  is the controlled extended generator  $\mathcal{L} : \mathcal{D}(\mathcal{L}) \stackrel{\text{def}}{=} \{f \in C_b^{1,2}([0, T] \times \mathcal{R}^d) : f(T, \cdot) \equiv 0\} \rightarrow C_b(Z \times U)$  given by

$$\mathcal{L}g(t, x, u) \stackrel{\text{def}}{=} \frac{\partial g}{\partial t}(t, x) + \langle m(x, u), \nabla g(t, x) \rangle + \frac{1}{2} \text{tr}(\sigma(x, u) \sigma^T(x, u) \nabla^2 g(t, x)).$$

Let  $\mathcal{M}$  denote the set of occupation measures.

LEMMA 2.1.  $\mathcal{M}$  is convex compact and is characterized by

$$\mathcal{M} = \left\{ \nu \in \mathcal{P}(Z \times U) : T \int_{Z \times U} \mathcal{L}g d\nu = - \int_{\mathcal{R}^d} g(0, \cdot) d\phi_0 \quad \forall g \in \mathcal{D}(\mathcal{L}) \right\}.$$

Furthermore, this set does not change if we restrict  $u(\cdot)$  to the class of Markov controls.

This is proved in [5]. The second half of the lemma justifies our relaxation to nonanticipative controls without loss of generality. Our goal will be to estimate a performance metric of the form

$$(6) \quad \int_{Z \times U} k d\nu \equiv \frac{1}{T} E \left[ \int_0^T k(t, x(t), u(t)) dt \right]$$

for a prescribed  $k(\cdot) \in C_b(Z \times U)$ , for  $x(\cdot)$  described by (1). Without loss of generality, we can assume  $k(\cdot) > 0$ . In view of the uncertainty about the coefficients in (1), we take a “worst case” viewpoint and estimate the minimum of this functional over all possible solutions to (4), i.e., over  $\mathcal{M}$ . As  $\mathcal{M}$  is characterized by linear constraints, this is an infinite dimensional linear program.

We qualify this formulation further by including additional constraints

$$(7) \quad \int_{Z \times U} k_i d\nu = c_i, \quad 1 \leq i \leq m.$$

Here  $k_i(\cdot) \in C_b(Z \times U)$  for all  $i$ , and the  $c_i$ ’s are prescribed values. Again, without loss of generality, we can assume  $k_i(\cdot) > 0, 1 \leq i \leq m$ . (One can also consider *inequalities* in (7). The analysis will be similar.) Note that  $k_i(\cdot) > 0$  implies that  $c_i > 0$ . We also assume

$$(\dagger) \quad c \stackrel{\text{def}}{=} [c_1, \dots, c_m] \text{ is an interior point of the set } \left\{ \left[ \int k_1 d\mu, \dots, \int k_m d\mu \right] : \mu \in \mathcal{M} \right\} \subset \mathcal{R}^m.$$

This is the *constrained control* problem studied in [9] and has found many applications. The inclusion of (7) complicates the dynamic programming approach to such problems, but makes only a small alteration to the linear programming approach.

Note that if it were a cost or risk rather than a (positive) performance metric, one would consider the maximum thereof. This can be reduced to the above framework by taking negative risk as the performance metric.

**Problem 2: Unmodelled dynamics.** Here we assume that we can *separately* observe for some integer  $l \geq 1$  only the following:

$$(8) \quad \tilde{x}_j(\cdot) \stackrel{\text{def}}{=} \left[ x_{\sum_{k=1}^{j-1} \tilde{d}_{k+1}}(\cdot), \dots, x_{\sum_{k=1}^j \tilde{d}_k}(\cdot) \right], \quad 1 \leq j \leq l, \quad \tilde{d}_j \geq 1, \quad \sum_{k=1}^l \tilde{d}_k \leq d.$$

Let  $\tilde{\nu}_j \in \mathcal{P}(\tilde{Z}_j)$  for  $\tilde{Z}_j \stackrel{\text{def}}{=} [0, T] \times \mathcal{R}^{\tilde{d}_j}$  denote the corresponding occupation measures defined by

$$(9) \quad \int_{\tilde{Z}_j} f d\tilde{\nu}_j \stackrel{\text{def}}{=} \frac{1}{T} \tilde{E}_j \left[ \int_0^T f(t, \tilde{x}_j(t)) dt \right], \quad f \in C_b(\tilde{Z}_j), \quad 1 \leq j \leq l.$$

We posit that observing samples of  $\tilde{x}_j(\cdot)$  allows us to estimate  $\tilde{\nu}_j$ . On the other hand, Lemma 2.1 allows us to mimic  $\tilde{\nu}_j$  by a  $\tilde{d}_j$ -dimensional diffusion  $\hat{x}_j(\cdot)$  given by (say)

$$(10) \quad d\hat{x}_j(t) = \tilde{m}_j(t, \hat{x}_j(t))dt + \tilde{\sigma}_j(t, \hat{x}_j(t))d\tilde{W}_j(t), \quad 1 \leq j \leq l.$$

Here  $\tilde{W}_j(\cdot)$  is a  $\tilde{d}_j$ -dimensional standard Brownian motion and the law  $\tilde{\phi}_{0j}$  of  $\hat{x}_j(0)$  is the image of  $\phi_0$  under the projection  $\mathcal{R}^d \rightarrow \mathcal{R}^{\tilde{d}_j}$ .

We view (10) as the *model fitted to the observed data*  $\tilde{x}_j(\cdot)$ . The justification for this is twofold. First, as already observed, Lemma 2.1 does justify a model such as (10) as long as we are interested in only the occupation measures. Second, fitting such a model is precisely what one does in much of time series analysis.

Finally, the unobserved components  $[x_{\sum_{k=1}^l \tilde{d}_{k+1}}(\cdot), \dots, x_d(\cdot)]$  then correspond to *unmodelled dynamics*, a notion familiar from robust control literature.

*Remark 1.* It is important to note that there is a compatibility issue here. Not every  $\tilde{\nu}_j$  as above will be compatible with the feasible  $\nu$ 's in the background. We assume that the feasible set  $\mathcal{M}'$  defined below is nonempty. This is reasonable in view of our underlying premise that the  $\tilde{\nu}_j$ 's have been arrived at empirically and therefore are perforce compatible with the underlying dynamics.

If  $l = 2$  and  $\tilde{d}_1 + \tilde{d}_2 = d$ , then we have models for two component processes whose dependence structure is unknown. This is the typical situation in credit risk problems, which is generalized above. It is also the exact dynamic counterpart of the Monge–Kantorovich optimal mass transportation problem.

We shall assume that for all  $j = 1, \dots, l$ ,  $\tilde{m}_j, \tilde{\sigma}_j$  are continuous and Lipschitz in the space variable uniformly w.r.t. the time variable. This is not free, i.e., it does not follow from our assumptions until now, but is an additional assumption. It is not unreasonable in view of the fact that one usually fits “nice” models (linear, composition of sigmoids, etc.) to data. Its main purpose here is to simplify the well-posedness issues, which now become nonissues. More generally, one must allow for measurable coefficients, in which case one may need to invoke genericity arguments for well-posedness facilitated by [20]. We avoid the additional complications implicit therein by taking this simple and not entirely unreasonable option. Also, we assume the *nondegeneracy* condition similar to (3) for the  $\tilde{\sigma}_j$ 's.

Let  $\tilde{\mathcal{L}}_j, 1 \leq j \leq l$ , denote the extended generator of  $\hat{x}_j(\cdot)$ ,

$$\tilde{\mathcal{L}}_j f(t, x) \stackrel{\text{def}}{=} \frac{\partial f}{\partial t}(t, x) + \langle \tilde{m}_j(t, x), \nabla f(t, x) \rangle + \frac{1}{2} \text{tr}(\tilde{\sigma}_j(t, x) \tilde{\sigma}_j^T(t, x) \nabla^2 f(t, x))$$

for all  $f \in C_b^{1,2}([0, T] \times \mathcal{R}^{\tilde{d}_j})$ . Let  $\mathcal{D}(\tilde{\mathcal{L}}_j) \stackrel{\text{def}}{=} \{f \in C_b^{1,2}([0, T] \times \mathcal{R}^{\tilde{d}_j}) : f(T, \cdot) \equiv 0\}$ . Then  $\tilde{\nu}_j$  is characterized by

$$T \int \tilde{\mathcal{L}}_j f d\tilde{\nu}_j = - \int f(0, \cdot) d\tilde{\phi}_{0j} \quad \forall f \in \mathcal{D}(\tilde{\mathcal{L}}_j).$$

This is the same as

$$(11) \quad \begin{aligned} & T \int \tilde{\mathcal{L}}_j f \nu(dtdx^j \times \mathcal{R}^{d-\tilde{d}_j} \times U) \\ &= - \int f(0, \cdot) \phi_0(dx^j \times \mathcal{R}^{d-\tilde{d}_j}) \quad \forall f \in \mathcal{D}(\tilde{\mathcal{L}}_j), \end{aligned}$$

where we partition  $x \in \mathcal{R}^d \approx \mathcal{R}^{\tilde{d}_1} \times \cdots \times \mathcal{R}^{\tilde{d}_j} \times \cdots \times \mathcal{R}^{\tilde{d}_l} \times \mathcal{R}^{d-\sum_{j=1}^l \tilde{d}_j}$  as  $x = [x^1 : \cdots : x^j : \cdots : x^l : y]$ , with  $x^j \in \mathcal{R}^{\tilde{d}_j}, y \in \mathcal{R}^{d-\sum_{j=1}^l \tilde{d}_j}$ . The goal then is to minimize  $\int_{Z \times U} k d\nu$  over a convex compact set  $\mathcal{M}'$  defined as

$$\mathcal{M}' \stackrel{\text{def}}{=} \{\nu \in \mathcal{M} : (11) \text{ holds for } 1 \leq j \leq l\}.$$

In the remainder of this paper, we consider a combined problem that has both features. That is, our problem is as follows:

$$\text{Minimize } \int_{Z \times U} k d\nu \text{ over } \mathcal{M}^* \stackrel{\text{def}}{=} \{\nu \in \mathcal{M}' : (7) \text{ holds}\}.$$

We shall assume that the *feasibility* condition holds; i.e., the *feasible* set of the above minimization problem is a nonempty subset of  $\mathcal{M}'$  (see Remark 1 above). It is clear that  $\mathcal{M}^*$  is closed compact and  $\int k d\nu$  is a continuous linear functional over  $\mathcal{M}^*$ . Thus by Weierstrass' theorem, we get the following.

LEMMA 2.2. *There exists a  $\nu^* \in \mathcal{M}^*$  such that*

$$\int_{Z \times U} k d\nu^* = \min_{\nu \in \mathcal{M}^*} \int_{Z \times U} k d\nu.$$

We conclude this section with a remark on notation. Superscript “\*” will stand for the formal adjoint of an operator throughout and  $\mathcal{D}(A)$  for the domain of an operator  $A$ .  $\langle \cdot, \cdot \rangle_{X,Y}$  denotes the standard pairing (bilinear form) between two topological vector spaces  $X, Y$  in duality.

**3. Dual linear programs.** We now consider the dual linear program associated with the above problem and its approximation. The advantages of working with a dual are that, first of all, it provides an approximate lower bound. In the kind of applications we have in mind, this is more useful than an approximate upper bound that the primal might provide. Furthermore, when additional constraints as in (7) are present, it converts a “min-max” problem into a pure maximization problem, because the Lagrange multipliers associated with these constraints are not being dualized.

We introduce the following notation:  $X \stackrel{\text{def}}{=} \mathcal{P}(Z \times U)$ ,  $Y \stackrel{\text{def}}{=} C_b(Z \times U)$ ,  $\Gamma \stackrel{\text{def}}{=} \mathcal{D}(\mathcal{L})^*$ ,  $\tilde{\Gamma}_j \stackrel{\text{def}}{=} \mathcal{D}(\tilde{\mathcal{L}}_j)^*$ ,  $1 \leq j \leq l$ . Given  $k(\cdot) \in C_b(Z \times U)$ , our primal problem can be stated as follows: Find

$$(12) \quad \inf_{\nu \in X} \langle k, \nu \rangle : F\nu = \delta_0 \times \phi_0, \tilde{F}_j \nu = \delta_0 \times \tilde{\phi}_{0j}, F_c \nu = -c, 1 \leq j \leq l,$$

where

- $F : X \rightarrow \Gamma$  is defined by  $F\nu = \mu$ , where

$$\langle g, \mu \rangle_{\mathcal{D}(\mathcal{L}), \Gamma} = -T \int_{Z \times U} \mathcal{L} g d\nu \quad \forall g \in \mathcal{D}(\mathcal{L}).$$

- $\tilde{F}_j : X \rightarrow \tilde{\Gamma}_j$  is defined by  $\tilde{F}_j \nu = \mu_j$ , where

$$\langle \tilde{g}_j, \mu_j \rangle_{\mathcal{D}(\tilde{\mathcal{L}}_j), \tilde{\Gamma}_j} = -T \int_{Z \times U} \tilde{\mathcal{L}}_j \tilde{g}_j d\nu \quad \forall \tilde{g}_j \in \mathcal{D}(\tilde{\mathcal{L}}_j), 1 \leq j \leq l.$$

- $c \stackrel{\text{def}}{=} [c_1, \dots, c_m]^T \in \mathcal{R}^m$  and  $F_c : X \rightarrow \mathcal{R}^m$  is defined by  $F_c \nu = r$ , where

$$r = [-\int k_1 d\nu, \dots, -\int k_m d\nu].$$

Following the developments of Chapter 3 of [2], the dual program is to find

$$(13) \quad \sup_{g \in \mathcal{D}(\mathcal{L}), \tilde{g}_j \in \mathcal{D}(\tilde{\mathcal{L}}_j) \forall j, \lambda \in \mathcal{R}^m} \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \sum_{i=1}^m \lambda_i c_i$$

$$: \frac{k(x, u) + \sum_{i=1}^m \lambda_i k_i(x, u)}{T} + \mathcal{L}g(x, u)$$

$$+ \sum_{j=1}^l \tilde{\mathcal{L}}_j \tilde{g}_j(\tilde{x}_j) \geq 0, (x, u) \in Z \times U$$

for  $\tilde{x}_j$  as defined in (8). In view of the comments that follow Remark 1 above, this may be viewed as the dynamic version of the Kantorovich dual in mass transportation theory (see, e.g., [29]). Hence we dub it the *generalized Kantorovich dual*.

**THEOREM 3.1.** *There exists strong duality between (12) and (13).*

*Proof.* It is easy to verify that the set

$$\{(F\nu, \tilde{F}_1\nu, \dots, \tilde{F}_l\nu, F_c\nu, \langle k, \nu \rangle_{Y,X}) : \nu \in X\}$$

is closed. Then the result follows from [2, Theorem 3.22, p. 61] and the discussion following it.  $\square$

Let  $\mathcal{D} \stackrel{\text{def}}{=} C_b^{1,2}(Z) \times C_b^{1,2}(\tilde{Z}_1) \times \dots \times C_b^{1,2}(\tilde{Z}_l)$  equipped with the product Sobolev norm defined by

$$\|(g, \tilde{g}_1, \dots, \tilde{g}_l)\|_{C_b^{1,2}} \stackrel{\text{def}}{=} \max\left(\|g\|_{C_b^{1,2}}, \max_{1 \leq j \leq l} \|\tilde{g}_j\|_{C_b^{1,2}}\right),$$

where  $\|f\|_{C_b^{1,2}} \stackrel{\text{def}}{=} (\max_{0 \leq |\alpha| \leq 2} \|D_x^\alpha f\|_\infty) \vee \|D_t f\|_\infty$ . (Here  $D_t$  denotes the time derivative and  $D_x^\alpha$  the space derivative of multi-index  $\alpha$ .) Let

$$BLR \stackrel{\text{def}}{=} \{(g, \tilde{g}_1, \dots, \tilde{g}_l) \in \mathcal{D} : \|(g, \tilde{g}_1, \dots, \tilde{g}_l)\|_{C_b^{1,2}} \leq R\}.$$

We have the following trivial consequence of Lemma 2.2 and Theorem 3.1, stated here for later use.

(\*) There exists a *large enough*  $R^* < \infty$ ,<sup>1</sup> and a *sufficiently large* compact subset  $\Lambda \subset \mathcal{R}^m$ , such that the supremum in (13) is attained at some  $(g^*, \tilde{g}_1^*, \dots, \tilde{g}_l^*, \lambda^*) \in BLR^* \times \Lambda$ , with  $g^* \in \mathcal{D}(\mathcal{L})$  and  $\tilde{g}_j^* \in \mathcal{D}(\tilde{\mathcal{L}}_j)$ ,  $1 \leq j \leq l$ .

#### 4. Approximation of dual programs.

**4.1. Bounded domain approximation.** We shall now describe the approximation scheme for (13) in several steps. Let  $\Omega_M \stackrel{\text{def}}{=} [-M, M]^d$ . As a first step of approximation, we can take  $Z_M \stackrel{\text{def}}{=} [0, T] \times \Omega_M$  for some  $M \gg 0$ ,  $T > 0$ . We assume that  $\text{support}(\phi_0) \subset \Omega_M$ . (Otherwise, there is an additional error term, which

<sup>1</sup>See also Lemma 4.5.

can be easily handled.) Let  $\partial Z_M \stackrel{\text{def}}{=} \{T\} \times \Omega_M \cup [0, T] \times \partial\Omega_M$ . Let  $\tilde{Z}_{Mj}$  denote the projection of  $Z_M$  on  $[0, T] \times \mathcal{R}^{\tilde{d}_j}$ ,  $1 \leq j \leq l$ . We define  $\partial\tilde{Z}_{Mj}$  similarly. Let  $\tau_M \stackrel{\text{def}}{=} \inf\{t \geq 0 : X(t) \notin \Omega_M\}$ . We shall now, in a manner analogous to the previous section, define  $\mathcal{M}_M, \mathcal{M}'_M, \mathcal{M}^*_M, \mathcal{D}_M(\mathcal{L}), \mathcal{D}_M(\tilde{\mathcal{L}}_j), \mathcal{D}_M$ , etc. Define the occupation measure  $\nu_M$  by

$$\int f d\nu_M \stackrel{\text{def}}{=} \frac{1}{T} E \left[ \int_0^{T \wedge \tau_M} f(t, x(t), u(t)) dt \right], \quad f \in C_b(Z_M \times U).$$

Let  $\mathcal{D}_M(\mathcal{L}) \stackrel{\text{def}}{=} \{f \in C_b^{1,2}(Z_M) : f|_{\partial Z_M} \equiv 0\}$ . Then the set of occupation measures  $\mathcal{M}_M$  is equivalently defined as

$$\mathcal{M}_M = \left\{ \nu_M \in \mathcal{P}(Z_M \times U) : T \int \mathcal{L} g d\nu_M = - \int g(0, \cdot) d\phi_0 \quad \forall g \in \mathcal{D}_M(\mathcal{L}) \right\}.$$

Our additional constraints (7) become

$$(14) \quad \int_{Z_M \times U} k_i d\nu_M = c_i, \quad 1 \leq i \leq m.$$

Once again, we assume the following counterpart of (†), which is reasonable for large  $M$ , as the  $\nu_M$ 's are then close to the corresponding  $\nu$ 's:

$$(††) \quad c \stackrel{\text{def}}{=} [c_1, \dots, c_m] \text{ is an interior point of the set } \left\{ \left[ \int k_1 d\mu, \dots, \int k_m d\mu \right] : \mu \in \mathcal{M}_M \right\} \subset \mathcal{R}^m.$$

For (9), (10), let  $\tilde{\nu}_{Mj} \in \mathcal{P}(\tilde{Z}_{Mj})$  denote the corresponding occupation measures defined by

$$(15) \quad \int f d\tilde{\nu}_{Mj} \stackrel{\text{def}}{=} \frac{1}{T} \tilde{E}_j \left[ \int_0^{\tau_M \wedge T} f(t, \tilde{x}_j(t)) dt \right], \quad f \in C_b(\tilde{Z}_{Mj}), \quad 1 \leq j \leq l.$$

Let  $\mathcal{D}_M(\tilde{\mathcal{L}}_j) \stackrel{\text{def}}{=} \{f \in C_b^{1,2}(\tilde{Z}_{Mj}) : f|_{\partial\tilde{Z}_{Mj}} \equiv 0\}$ . Then  $\tilde{\nu}_{Mj}$  is characterized by

$$T \int \tilde{\mathcal{L}}_j f d\tilde{\nu}_{Mj} = - \int f(0, \cdot) d\tilde{\phi}_{0j} \quad \forall f \in \mathcal{D}_M(\tilde{\mathcal{L}}_j).$$

This is the same as, for all  $f \in \mathcal{D}_M(\tilde{\mathcal{L}}_j)$ ,

$$(16) \quad T \int \tilde{\mathcal{L}}_j f \nu_M (dt dx^j \times \Omega_M^{d-\tilde{d}_j} \times U) = - \int f(0, \cdot) \phi_0 (dx^j \times \Omega_M^{d-\tilde{d}_j}),$$

where we partition  $x \in \Omega_M \approx \Omega_M^{\tilde{d}_1} \times \dots \times \Omega_M^{\tilde{d}_l} \times \Omega_M^{d-\sum_{j=1}^l \tilde{d}_j}$  as  $x = [x^1 : \dots : x^j : \dots : x^l : y]$  with  $x^j \in \Omega_M^{\tilde{d}_j}$ ,  $y \in \Omega_M^{d-\sum_{j=1}^l \tilde{d}_j}$  and  $\Omega_M^{\tilde{d}_j}$  denotes the  $\tilde{d}_j$ -dimensional image of  $\Omega_M$  under the projection to the appropriate factor space. The goal then is to minimize  $\int_{Z_M \times U} k d\nu$  over a convex compact set  $\mathcal{M}'_M$  defined as

$$\mathcal{M}'_M \stackrel{\text{def}}{=} \{\nu_M \in \mathcal{M}_M : (16) \text{ holds for } 1 \leq j \leq l\}.$$

As before, we consider a combined problem that has both constraints (14) and unmodelled dynamics. Then the bounded domain approximation to the original problem (6) is as follows:

Minimize over  $\mathcal{M}_M^* \stackrel{\text{def}}{=} \{\nu_M \in \mathcal{M}'_M : (14) \text{ holds}\}$  the functional

$$(17) \quad \int_{Z_M \times U} k d\nu_M \equiv \frac{1}{T} E \left[ \int_0^{\tau_M \wedge T} k(t, x(t), u(t)) dt \right].$$

As before, define  $\mathcal{D}_M \stackrel{\text{def}}{=} C_b^{1,2}(Z_M) \times C_b^{1,2}(\tilde{Z}_{M1}) \times \cdots \times C_b^{1,2}(\tilde{Z}_{Ml})$  equipped with the appropriate Sobolev norm (see the discussion between Theorem 3.1 and (\*)). The corresponding dual becomes the following: For all  $(x, u) \in Z_M \times U$ ,

$$(18) \quad \begin{aligned} \sup_{g \in \mathcal{D}_M(\mathcal{L}), \tilde{g}_j \in \mathcal{D}_M(\tilde{\mathcal{L}}_j) \forall j, \lambda \in \mathcal{R}^m} & \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \sum_{i=1}^m \lambda_i c_i \\ & : \frac{k(x, u) + \sum_{i=1}^m \lambda_i k_i(x, u)}{T} + \mathcal{L}g(x, u) \\ & + \sum_{j=1}^l \tilde{\mathcal{L}}_j \tilde{g}_j(\tilde{x}_j) \geq 0 \end{aligned}$$

for  $\tilde{x}_j$  defined as in (8). Results analogous to Lemma 2.2, Theorem 3.1, and (\*) also hold. Let us define  $\varepsilon_j(M), 1 \leq j \leq m+1$ , to be

$$\varepsilon_j(M) \stackrel{\text{def}}{=} \left| \frac{1}{T} E \left[ \int_0^T k_j(t, x(t), u(t)) dt \right] - \frac{1}{T} E \left[ \int_0^{\tau_M \wedge T} k_j(t, x(t), u(t)) dt \right] \right|,$$

where  $k_{m+1}(\cdot) \stackrel{\text{def}}{=} k(\cdot)$ . We define the *unconstrained domain approximation error* as

$$(19) \quad \varepsilon(M) \stackrel{\text{def}}{=} \max_{1 \leq j \leq m+1} \varepsilon_j(M).$$

The following lemma shows that (19) goes to zero as  $\Omega_M \uparrow \mathcal{R}^m$  and in the process, gives an upper bound on  $\varepsilon(M)$ .

LEMMA 4.1.  $\varepsilon(M) \xrightarrow{M \uparrow \infty} 0$ .

*Proof.*

$$\begin{aligned} \varepsilon_j(M) &= \left| \frac{1}{T} E \left[ \int_0^T k_j(t, x(t), u(t)) I\{t \geq \tau_M\} dt \right] \right| \\ &\leq \frac{1}{T} E \left[ \int_0^T (k_j(t, x(t), u(t)))^2 dt \right]^{1/2} E \left[ \int_0^T I\{t \geq \tau_M\} dt \right]^{1/2} \\ &= \bar{C}_j \left[ \int_0^T P(T \geq t \geq \tau_M) dt \right]^{1/2} \\ &\leq \bar{C}_j \sqrt{TP} (T \geq \tau_M)^{1/2} \\ &= \bar{C}_j \sqrt{TP} \left( \sup_{t \in [0, T]} \|x(t)\|_\infty \geq M \right)^{1/2} \end{aligned}$$

$$\begin{aligned} &\leq \bar{C}_j \sqrt{T} \frac{E[\sup_{t \in [0, T]} \|x(t)\|_\infty^2]^{1/2}}{M} \\ &\leq \frac{\bar{C}_j \sqrt{T} \tilde{C} (1 + E[\|x(0)\|^2])^{1/2}}{M}, \end{aligned}$$

where  $\bar{C}_j$  is a constant bound on  $k_j(\cdot)$ , and  $\tilde{C}$  is a constant depending on  $T$  and the Lipschitz constant of the drift/diffusion terms in (1). The last inequality is an easy consequence of Gronwall's inequality. Taking  $\bar{C} = \max_{1 \leq j \leq m+1} \bar{C}_j$  we get

$$\varepsilon(M) \leq \frac{\bar{C} \sqrt{T} \tilde{C} (1 + E[\|x(0)\|^2])^{1/2}}{M}.$$

The result follows.  $\square$

Let  $\beta_o$  and  $\beta_d$  denote the optimal values for (13) and (18), respectively. Next we find a bound on the *constrained domain approximation error*  $\zeta(M) \stackrel{\text{def}}{=} |\beta_o - \beta_d|$ . Using (\*), we can write

$$\beta_o = \min_{\nu \in \mathcal{M}^*} \langle \nu, k \rangle,$$

where, as before,  $\mathcal{M}^* = \{\nu \in \mathcal{M}' : \int_{Z \times U} k_j d\nu = c_j, 1 \leq j \leq m\}$ , and

$$\beta_d = \min_{\nu_M \in \mathcal{M}_M^*} \langle \nu_M, k \rangle,$$

where  $\mathcal{M}_M^* = \{\nu_M \in \mathcal{M}'_M : \int_{Z_M \times U} k_j d\nu_M = c_j, 1 \leq j \leq m\}$ .

Let us define a map  $F : \mathcal{M}' \rightarrow \mathcal{R}^{m+1}$  as follows:

$$\begin{aligned} \mathcal{M}' \ni \nu &\longmapsto F\nu = \left[ \int_{Z \times U} k_1 d\nu, \dots, \int_{Z \times U} k_m d\nu, \int_{Z \times U} k d\nu \right] \\ &\equiv [y_1(\nu), \dots, y_m(\nu), y_{m+1}(\nu)] \in \mathcal{R}^{m+1}. \end{aligned}$$

Map  $F'$  is defined over  $\mathcal{M}'_M$  in an analogous manner. Note that  $F, F'$  are bounded linear maps and map the compact convex sets  $\mathcal{M}', \mathcal{M}'_M$ , respectively, to compact convex sets in  $\mathcal{R}^{m+1}$ .

The following lemma is the main result of this subsection.

LEMMA 4.2.  $\zeta(M) \leq C_d \varepsilon(M) \xrightarrow{M \uparrow \infty} 0$  for some constant  $C_d = C_d(M) > 0$  which remains uniformly bounded as  $M \uparrow \infty$ .

*Proof.* Using the definition of  $F, F'$ , we can write

$$\beta_o = \min_{[c_1, \dots, c_m, y_{m+1}(\nu)] \in F(\mathcal{M}')} y_{m+1}(\nu)$$

and

$$\beta_d = \min_{[c_1, \dots, c_m, y_{m+1}(\nu_M)] \in F'(\mathcal{M}'_M)} y_{m+1}(\nu_M).$$

Note that by ( $\dagger\dagger$ ), the hypersurfaces  $y_j(\nu) = c_j, 1 \leq j \leq m$ , generate a  $2^m$ -partition of the compact convex set  $F(\mathcal{M}')$ . Then we must have that  $\beta_o$  = the *minimum* of  $y_{m+1}(\nu)$  over *at least one* of these  $2^m$ -partitions. Without loss of generality, let us assume that this is so for the orthant  $Y_1 \stackrel{\text{def}}{=} \{[y_1(\nu), \dots, y_m(\nu), y_{m+1}(\nu)] \in F(\mathcal{M}') :$



$y_j(\nu) \leq c_j, 1 \leq j \leq m\}$ . Hence for  $\mathcal{M}_1^* \stackrel{\text{def}}{=} F^{-1}(Y_1) = \{\nu \in \mathcal{M}' : \int_{Z \times U} k_j d\nu \leq c_j, 1 \leq j \leq m\}$ , we have

$$(20) \quad \beta_o = \min_{\nu \in \mathcal{M}_1^*} \langle \nu, k \rangle.$$

Using analogous arguments, we suppose that

$$\beta_d = \min_{\nu_M \in \mathcal{M}_{M,1}^*} \langle \nu_M, k \rangle$$

for  $\mathcal{M}_{M,1}^* \stackrel{\text{def}}{=} \{\nu_M \in \mathcal{M}'_M : \int_{Z_M \times U} k_j d\nu_M \leq c_j, 1 \leq j \leq m\}$ . (We have chosen the orthant corresponding to (20) above. This is purely for notational convenience; in reality it could be a different orthant.) Using the Lagrange multiplier formulation we can write

$$\begin{aligned} \beta_o &= \max_{\lambda \geq 0} \min_{\nu \in \mathcal{M}'} \left\{ \langle \nu, k \rangle + \sum_{i=1}^m \int_{Z \times U} \lambda_i k_i d\nu - \langle \lambda, c \rangle \right\}, \\ \beta_d &= \max_{\lambda_M \geq 0} \min_{\nu_M \in \mathcal{M}'_M} \left\{ \langle \nu_M, k \rangle + \sum_{i=1}^m \int_{Z_M \times U} \lambda_{Mi} k_i d\nu_M - \langle \lambda_M, c \rangle \right\}. \end{aligned}$$

Let  $\lambda^*, \lambda_M^* \geq 0$  be the Lagrange multipliers (i.e., the *maximizing* values of  $\lambda, \lambda_M$ , respectively, on the r.h.s.) for the two cases, respectively. By  $(\dagger)$ , there exists an *interior* point  $\bar{\nu}$  of  $\mathcal{M}_1^*$ , i.e.,  $\int_{Z \times U} k_i d\bar{\nu} < c_i, 1 \leq i \leq m$ . Let  $\nu^* = \operatorname{argmin}_{\nu \in \mathcal{M}_1^*} \langle \nu, k \rangle$ . Similarly, by  $(\dagger\dagger)$ , we have an *interior* point  $\bar{\nu}_M \in \mathcal{M}_{M,1}^*$ . Let  $\nu_M^* = \operatorname{argmin}_{\nu_M \in \mathcal{M}_{M,1}^*} \langle \nu_M, k \rangle$ . Suppose  $\beta_o \geq \beta_d$ . Then we have

$$\begin{aligned} |\beta_o - \beta_d| &= \max_{\lambda \geq 0} \min_{\nu \in \mathcal{M}'} \left\{ \langle \nu, k \rangle + \sum_{i=1}^m \int_{Z \times U} \lambda_i k_i d\nu - \langle \lambda, c \rangle \right\} \\ &\quad - \max_{\lambda_M \geq 0} \min_{\nu_M \in \mathcal{M}'_M} \left\{ \langle \nu_M, k \rangle + \sum_{i=1}^m \int_{Z_M \times U} \lambda_{Mi} k_i d\nu_M - \langle \lambda_M, c \rangle \right\} \\ &\leq \min_{\nu \in \mathcal{M}'} \left\{ \langle \nu, k \rangle + \sum_{i=1}^m \int_{Z \times U} \lambda_i^* k_i d\nu - \langle \lambda^*, c \rangle \right\} \\ &\quad - \min_{\nu_M \in \mathcal{M}'_M} \left\{ \langle \nu_M, k \rangle + \sum_{i=1}^m \int_{Z_M \times U} \lambda_i^* k_i d\nu_M - \langle \lambda^*, c \rangle \right\} \\ &\leq \max_{\mathcal{L}(x(\cdot), u(\cdot))} \left| \left( \langle \nu, k \rangle - \langle \nu_M, k \rangle \right) + \sum_{i=1}^m \lambda_i^* (\langle \nu, k_i \rangle - \langle \nu_M, k_i \rangle) \right| \\ &\leq (1 + \|\lambda^*\|_1) \varepsilon(M), \end{aligned}$$

where  $\mathcal{L}(x(\cdot), u(\cdot))$  denotes the joint law of  $(x(\cdot), u(\cdot))$ . (Note that each pair  $(\nu, \nu_M)$  corresponds to one such law.) The  $\beta_d \geq \beta_o$  case is exactly symmetric with  $\lambda_M^*$  in place of  $\lambda^*$ . Hence we have

$$|\beta_o - \beta_d| \leq (1 + (\|\lambda^*\|_1 \vee \|\lambda_M^*\|_1)) \varepsilon(M).$$

Now it follows from Exercise 5.3.1, p. 516 of [4], that

$$\begin{aligned} \|\lambda^*\|_1 &\leq \frac{\langle \bar{\nu}, k \rangle - \langle \nu^*, k \rangle}{\min_{1 \leq i \leq m} (c_i - \int_{Z \times U} k_i d\bar{\nu})}, \\ \|\lambda_M^*\|_1 &\leq \frac{\langle \bar{\nu}_M, k \rangle - \langle \nu_M^*, k \rangle}{\min_{1 \leq i \leq m} (c_i - \int_{Z_M \times U} k_i d\bar{\nu}_M)}. \end{aligned}$$

Since  $k(\cdot) > 0$ , we have  $\langle \nu^*, k \rangle > 0$ ,  $\langle \nu_M^*, k \rangle > 0$ , and hence

$$|\beta_o - \beta_d| \leq (1 + \|k\|_\infty A(c, k_1, \dots, k_m))\varepsilon(M),$$

where

$$A(c, k_1, \dots, k_m) \stackrel{\text{def}}{=} \frac{1}{\min_{1 \leq i \leq m} (c_i - \int_{Z \times U} k_i d\bar{\nu}) \wedge \min_{1 \leq i \leq m} (c_i - \int_{Z_M \times U} k_i d\bar{\nu}_M)}.$$

Also,

$$A(c, k_1, \dots, k_m) \rightarrow \frac{1}{\min_{1 \leq i \leq m} (c_i - \|k_i\|_\infty)} < \infty$$

as  $M \uparrow \infty$ . The result follows.  $\square$

In anticipation of the finite dimensional approximation to be introduced in the next subsection, we will have to specify boundary conditions to be *not exactly zero*, but bounded by a *given small constant*  $\eta > 0$ . This is because the approximation of an RKHS function vanishing on the boundary by one in a finite dimensional subspace thereof will be small but not necessarily zero on the boundary. This converts the problem (18) to the following: For all  $(x, u) \in Z_M \times U$ ,

$$(21) \quad \sup_{(g, \tilde{g}_1, \dots, \tilde{g}_l) \in \mathcal{D}_M, \lambda \in \mathcal{R}^m} \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \sum_{i=1}^m \lambda_i c_i \\ : \frac{k(x, u) + \sum_{i=1}^m \lambda_i k_i(x, u)}{T} + \mathcal{L}g(x, u) \\ + \sum_{j=1}^l \tilde{\mathcal{L}}_j \tilde{g}_j(\tilde{x}_j) \geq 0; \\ |g|_{\partial Z_M} \leq \eta; \quad |\tilde{g}_j|_{\partial \tilde{Z}_{Mj}} \leq \eta \quad \forall j$$

with  $\tilde{x}_j$  as in (8).

Let  $\beta_x$  denote the optimal value for (21). We define the error function  $\hat{\varepsilon}(\eta)$  due to this further approximation as  $\hat{\varepsilon}(\eta) \stackrel{\text{def}}{=} |\beta_d - \beta_x|$ . Define a family of functions

$$F_\eta \stackrel{\text{def}}{=} \{h(\cdot) : Z_M \rightarrow \mathcal{R} \text{ continuous, } |h(x)| \leq \eta \quad \forall x \in \partial Z_M\}.$$

LEMMA 4.3.  $\hat{\varepsilon}(\eta) \leq \eta(l+1)$ .

*Proof.* Let

$$f(x, u) \stackrel{\text{def}}{=} \frac{k(x, u) + \sum_{i=1}^m \lambda_i k_i(x, u)}{T},$$

$$G \stackrel{\text{def}}{=} \left\{ (g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in \mathcal{D}_M \times \mathcal{R}^m : f(x, u) + \mathcal{L}g(x, u) + \sum_{j=1}^l \tilde{\mathcal{L}}_j \tilde{g}_j(\tilde{x}_j) \geq 0 \right\},$$

and

$$G' \stackrel{\text{def}}{=} \{ (g, \tilde{g}_1, \dots, \tilde{g}_l) \in \mathcal{D}_M : \mathcal{L}g(x, u) = 0; \quad \tilde{\mathcal{L}}_j \tilde{g}_j(\tilde{x}_j) = 0 \quad \forall j; \quad g|_{\partial Z_M} = h|_{\partial Z_M}, \\ \tilde{g}_j|_{\partial \tilde{Z}_{Mj}} = h_j|_{\partial \tilde{Z}_{Mj}} \quad \forall j, \quad h, h_j \in F_\eta \}.$$

Then,

$$\begin{aligned}\beta_d &= \sup_{(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in G} \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \\ &: \quad g|_{\partial Z_M} \equiv 0, \quad \tilde{g}_j|_{\partial \tilde{Z}_{Mj}} \equiv 0, \\ \beta_x &= \sup_{(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in G} \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \\ &: \quad |g|_{\partial Z_M} \leq \eta, \quad |\tilde{g}_j|_{\partial \tilde{Z}_{Mj}} \leq \eta.\end{aligned}$$

Note that  $\beta_x$  can be written as

$$\begin{aligned}\beta_x &= \sup \int (g^1(0, \cdot) + g^2(0, \cdot)) d\phi_0 + \sum_{j=1}^l \int (\tilde{g}_j^1(0, \cdot) + \tilde{g}_j^2(0, \cdot)) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \\ &: \quad (g^1, \tilde{g}_1^1, \dots, \tilde{g}_l^1, \lambda) \in G, \quad g^1|_{\partial Z_M} \equiv 0; \quad \tilde{g}_j^1|_{\partial \tilde{Z}_{Mj}} \equiv 0, \\ &\quad (g^2, \tilde{g}_1^2, \dots, \tilde{g}_l^2) \in G' .\end{aligned}$$

But then  $g_j^2$  is of the form  $\tilde{g}_j^2(\tilde{x}_j) = E_{\tilde{x}_j}[h_j(\hat{X}_j(\tau_M \wedge T))]$ , where  $\hat{X}_j(\cdot)$  is a  $\tilde{d}_j$ -dimensional diffusion given by (10). In particular,  $|g_j^2(\cdot)| \leq \eta$ . Similar results hold for  $g^2(\cdot)$ . So, we get

$$\begin{aligned}|\beta_d - \beta_x| &\leq \sup_{(g^2, \tilde{g}_1^2, \dots, \tilde{g}_l^2) \in G'} \left| \int g^2(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}_j^2(0, \cdot) d\tilde{\phi}_{0j} \right| \\ &= \sup_{h(\cdot), h_1(\cdot), \dots, h_l(\cdot)} \left| \int E_x[h(X(\tau_M \wedge T))] d\phi_0(x) \right. \\ &\quad \left. + \sum_{j=1}^l \int E_{\tilde{x}_j}[h_j(\hat{X}_j(\tau_M \wedge T))] d\tilde{\phi}_{0j}(\tilde{x}_j) \right| \\ &\leq \eta(l+1).\end{aligned}$$

The claim follows.  $\square$

Henceforth we drop the subscript  $M$  in the notation and assume that all definitions are for the bounded domain  $Z_M$  under consideration.

**4.2. Finite dimensional approximation.** In this subsection, we give approximation error bounds for our problem (21) using finite dimensional functions with a finite number of constraints.

Let  $K(x, y)$  denote a  $C^\infty$  Mercer (i.e., continuous, symmetric, and positive definite) kernel on  $Z \times Z$ . Hence so will be its restriction  $\tilde{K}(x, y)$  to  $\tilde{Z}_j \times \tilde{Z}_j$ . Associated with this is a unique *RKHS*  $\mathcal{H}_K$  of functions on  $Z$  and  $\tilde{\mathcal{H}}_{Kj}$  of functions on  $\tilde{Z}_j$ , characterized by ([10, p. 35]):

1. For all  $x \in Z$ ,  $K_x \stackrel{\text{def}}{=} K(x, \cdot) \in \mathcal{H}_K$ ;
2.  $\text{span}\{K_x : x \in Z\}$  is dense in  $\mathcal{H}_K$ ;
3. For all  $f \in \mathcal{H}_K$ ,  $f(x) = \langle K_x, f \rangle$ ,

and likewise for  $\tilde{\mathcal{H}}_{Kj}$ ,  $1 \leq j \leq l$ . The kernel  $K$  defines (see [10, Proposition 1, p. 33]) the Hilbert–Schmidt integral operator  $L_K$  by

$$L_K(f)(z) = \int_Z K(z, t) f(t) d\nu(t), \quad z \in Z, f \in L^2_\nu(Z),$$

with  $\vartheta$  = the Lebesgue measure on  $Z$ . Let  $\{\phi_k\}_{k \geq 1}$  be the eigenvectors corresponding to the eigenvalues  $\{\lambda_k\}_{k \geq 1}$  of  $L_K$ . By [10, Theorem 4, p. 37], the inner product on  $\mathcal{H}_K$  can be defined by

$$\langle f, g \rangle_K = \sum_{k=1}^{\infty} \frac{f_k g_k}{\lambda_k}.$$

Here  $f = \sum_{k=1}^{\infty} f_k \phi_k$ ,  $g = \sum_{k=1}^{\infty} g_k \phi_k$ . The corresponding norm is defined by  $\|f\|_K \stackrel{\text{def}}{=} (\sum_{k=1}^{\infty} \frac{f_k^2}{\lambda_k})^{\frac{1}{2}}$ . Analogous statements hold for  $\tilde{Z}_j, \tilde{\mathcal{H}}_{Kj}$ .

From an approximation point of view, the attractive features of  $\mathcal{H}_K$  are the following:

- They are spanned by the translates of a single function  $K_x$ , and the best least squares estimate of any  $f$  sampled at a discrete subset  $\tilde{Z}$  of  $Z$ , with a regularity penalty, turns out to be a linear combination of the  $K_z, z \in \tilde{Z}$ . The latter is a version of the Shannon sampling theorem (see [25], [26], [27]).
- For any  $n > d + 1$ ,  $\mathcal{H}_K$  is densely embedded in  $H^{\frac{n}{2}}(Z)$  and the embedding is bounded (see the proof of [10, Theorem D, p. 41]). By the Rellich–Kondrachov theorem ([1, Theorem 6.3, part II, p. 168]),  $H^{m+2}(Z)$  is densely embedded in  $C_b^{1,2}(Z)$  and the embedding is compact whenever  $m > \frac{d+1}{2}$ . Thus  $\mathcal{H}_K$  is compactly and densely embedded in  $C_b^{1,2}(Z)$  if we choose  $n = 2m + 4 > d + 5$ . Similar statements apply to  $\tilde{Z}_j, \tilde{d}_j, \tilde{m}_j, \tilde{\mathcal{H}}_{Kj}$  in place of  $Z, d, m, \mathcal{H}_K$ . Thus  $\mathcal{H} \stackrel{\text{def}}{=} \mathcal{H}_K \times \tilde{\mathcal{H}}_{K1} \times \cdots \times \tilde{\mathcal{H}}_{Kl}$  is compactly and densely embedded in  $\mathcal{D}$ . This makes it a convenient space for approximation (*hypothesis space* in the language of statistical learning theory; see [10]).

For future reference, let  $\mathcal{I}$  denote this injection. Let us define the product Sobolev space  $\mathcal{W}$  as follows:

$$(22) \quad \mathcal{W} \stackrel{\text{def}}{=} \left\{ (u, \tilde{u}_1, \dots, \tilde{u}_l) \in H^{m+2}(Z) \times H^{\tilde{m}_1+2}(\tilde{Z}_1) \times \cdots \times H^{\tilde{m}_l+2}(\tilde{Z}_l) \right. \\ \left. : m > \frac{d+1}{2}, \tilde{m}_j > \frac{\tilde{d}_j+1}{2}, 1 \leq j \leq l \right\}.$$

Let  $\bar{Z} \subset Z, \bar{\tilde{Z}}_j \subset \tilde{Z}_j, 1 \leq j \leq l$ , be finite. Define a Hilbert space  $\mathcal{H}_{K,\bar{Z}}$  of functions  $f : Z \rightarrow \mathcal{R}$  as the linear span of  $K(z_i, \cdot), z_i \in \bar{Z}$ , with inner product

$$\left\langle \sum_i a_i K_{z_i}, \sum_j b_j K_{z_j} \right\rangle_{K,\bar{Z}} \stackrel{\text{def}}{=} \sum_{i,j} a_i b_j K(z_i, z_j).$$

This is a finite dimensional Hilbert space of dimension  $|\bar{Z}|$ . The  $|\bar{\tilde{Z}}_j|$ -dimensional Hilbert spaces  $\tilde{\mathcal{H}}_{Kj,\bar{Z}}, 1 \leq j \leq l$ , of functions  $\tilde{f}_j : \tilde{Z}_j \rightarrow \mathcal{R}$  are defined analogously. These norms are compatible with the norms of  $\mathcal{H}_K, \tilde{\mathcal{H}}_{Kj}$ , respectively, relativized to  $\mathcal{H}_{K,\bar{Z}}, \tilde{\mathcal{H}}_{Kj,\bar{Z}}$ , viewed as subspaces thereof.

For  $R > 0$ ,  $B_R \stackrel{\text{def}}{=} \text{the closed ball of radius } R \text{ in } \mathcal{H}, \text{ centered at the origin.}$  By the properties of  $\mathcal{H}$  noted above,  $\overline{\mathcal{I}(B_R)}$  is compact in  $\mathcal{D}$  and  $\cup_{R>0} \overline{\mathcal{I}(B_R)}$  is dense in  $\mathcal{D}$ . For  $R^*$  as in (\*), we shall fix any  $R > 0$  such that  $R \geq \|\mathcal{I}\|^{-1} R^*$ . (See also Lemma 4.5 below for a specific choice.) Thus we get  $BL_{R^*} \subseteq \overline{\mathcal{I}(B_R)}$ . Let  $A^\epsilon$  for any set  $A \subset \mathcal{D}$  denote its  $\epsilon$ -thickening, i.e., the set of all points which are at most  $\epsilon$  away

from  $A$  in the  $\|\cdot\|_{C_b^{1,2}}$  norm. We shall approximate  $\overline{\mathcal{I}(B_R)}$  by its intersection with  $\bar{\mathcal{H}} \stackrel{\text{def}}{=} \mathcal{H}_{K,Z} \times \tilde{\mathcal{H}}_{K1,Z} \times \cdots \times \tilde{\mathcal{H}}_{Kl,Z}$ . The quality of this approximation is described by Theorem 5.1 in section 5. Let the *sampling approximation error*  $\Xi$  be defined as

$$(23) \quad \Xi \stackrel{\text{def}}{=} \sup_{(g, \tilde{g}_1, \dots, \tilde{g}_l) \in \overline{\mathcal{I}(B_R)}} \inf_{(\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l) \in \bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}} \|(g, \tilde{g}_1, \dots, \tilde{g}_l) - (\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l)\|_{C_b^{1,2}}.$$

Using  $R$  fixed as above, we now consider a semi-infinite approximation to (21). This is semi-infinite because, while it has finitely many variables, it still has infinitely many constraints. Specifically, this will be

$$(24) \quad \begin{aligned} & \sup \left\{ \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle : (g, \tilde{g}_1, \dots, \tilde{g}_l) \in \bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}, \right. \\ & \quad \lambda \in \Lambda, \text{ and, } \frac{k(x, u) + \sum_{i=1}^m \lambda_i k_i(x, u)}{T} + \mathcal{L}g(x, u) + \sum_{j=1}^l \tilde{\mathcal{L}}_j \tilde{g}_j(\tilde{x}_j) \geq 0 \\ & \quad \left. \forall (x, u) \in Z \times U, |g|_{\partial Z} \leq \eta, |\tilde{g}_j|_{\partial \tilde{Z}_j} \leq \eta \right\} \end{aligned}$$

with  $\tilde{x}_j$  is defined as in (8). Let  $\beta_s$  denote the optimal value for (24).

Let us now prove two important lemmas. The first is a technical one which is a slight generalization of the result in Exercise 5.3.1, p. 516 of [4]. Let  $G$  be a convex subset of  $\mathcal{D} \times \mathcal{R}^m$ . Let  $\mathcal{F}$  be a concave mapping from  $G$  into  $\mathcal{R}$ . Assume that there exists a  $(\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda}) \in G$  such that  $\mathcal{F}((\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda})) > 0$ . Let  $f$  be a real-valued convex functional on  $G$ . Let

$$\beta^* \stackrel{\text{def}}{=} \operatorname{argmax}_G f(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) : \mathcal{F}(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \geq 0.$$

LEMMA 4.4. *There exists an element  $\nu^* \in C_b(Z)^*$  with  $\nu^* \geq 0$  such that*

$$\|\nu^*\| \leq \frac{\beta^* - f(\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda})}{\min_{x \in Z} \mathcal{F}((\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda}))(x)}.$$

*Proof.* It follows from [23, Theorem 1, p. 217] that there exists  $\nu^* \in C_b(Z)^*$  such that  $\nu^* \geq 0$  and

$$-\beta^* = \inf_{(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in G} [-f(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) + \langle -\mathcal{F}((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)), \nu^* \rangle].$$

Then

$$-\beta^* + f(\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda}) \leq \langle -\mathcal{F}((\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda})), \nu^* \rangle.$$

Since  $\min_{x \in Z} \mathcal{F}((\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda}))(x) \leq \mathcal{F}((\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda}))(x)$ , we have

$$\begin{aligned} \beta^* - f(\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda}) & \geq \left( \min_{x \in Z} \mathcal{F}((\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda}))(x) \right) \int_Z d\nu^* \\ & = \|\nu^*\| \left( \min_{x \in Z} \mathcal{F}((\bar{g}, \bar{\tilde{g}}_1, \dots, \bar{\tilde{g}}_l, \bar{\lambda}))(x) \right). \end{aligned}$$

The last equality holds because  $\nu^* \geq 0$ . The claim follows.  $\square$

Now define  $\mathcal{F} : \mathcal{H} \times \mathcal{R}^m \rightarrow C_b(Z)$  as

$$\begin{aligned} \mathcal{F}((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda))(x) \stackrel{\text{def}}{=} \inf_{u \in U} \left( \frac{k(x, u) + \sum_{i=1}^m \lambda_i k_i(x, u)}{T} + \mathcal{L}g(x, u) \right) \\ + \sum_{j=1}^l \tilde{\mathcal{L}}_j \tilde{g}_j(\tilde{x}_j) \quad \forall x \in Z \end{aligned}$$

with  $\tilde{x}_j$  defined as in (8). Note that this is a concave functional. Furthermore,

$$\begin{aligned} \|\mathcal{L}\| &\leq 1 + \sup_{(x,u) \in \Omega_M \times U} |m(x, u)| + \frac{1}{2} \sup_{(x,u) \in \Omega_M \times U} |\text{tr}(\sigma(x, u)\sigma^T(x, u))| \text{ and} \\ \|\tilde{\mathcal{L}}_j\| &\leq 1 + \sup_{(t,x) \in Z} |\tilde{m}_j(t, x)| + \frac{1}{2} \sup_{(t,x) \in Z} |\text{tr}(\tilde{\sigma}_j(t, x)\tilde{\sigma}_j^T(t, x))|, \quad 1 \leq j \leq l. \end{aligned}$$

LEMMA 4.5. *Let  $\Xi = \epsilon$ . Given any  $\xi, \eta > 0$ , for sufficiently small  $\epsilon > 0$ , there exists a solution  $(v, \tilde{v}_1, \dots, \tilde{v}_l, \lambda) \in (\mathcal{H} \cap \mathcal{I}(B_R)) \times \Lambda$  to the inequality*

$$\begin{aligned} \mathcal{F}((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) &\geq \xi(l+1) - \epsilon(\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|) > 0, \\ |g|_{\partial Z} &\leq \eta, \\ |\tilde{g}_j|_{\partial \tilde{Z}_j} &\leq \eta, \quad 1 \leq j \leq l, \end{aligned}$$

for large enough  $R^*$ , and  $R$  satisfying  $R \geq \|\mathcal{I}\|^{-1} R^*$ .

*Proof.* Consider the HJB equation

$$\begin{aligned} \inf_{u \in U} \left( \frac{k(t, x, u) + \sum_{i=1}^m \lambda_i k_i(t, x, u)}{T} + \mathcal{L}g(t, x, u) - \xi \right) &= 0, \quad (t, x) \in Z, \\ g|_{\partial Z} &= \frac{\eta}{2}, \end{aligned}$$

for a prescribed  $\xi > 0$ . Since *nondegeneracy* (3) holds, by [12, Theorem 4.2, p. 169], it has a solution  $g(\cdot) \in C_b^{1,2}(Z)$ . By [12, Theorem 3.1, p. 163],

$$g(t, x) = \inf_{u(\cdot)} E_{t,x} \left[ \int_t^{\tau_M \wedge T} \left( \frac{k(s, x(s), u(s)) + \sum_{i=1}^m \lambda_i k_i(s, x(s), u(s))}{T} - \xi \right) dt + \frac{\eta}{2} \right], \quad (25)$$

where  $x(\cdot)$  is a  $d$ -dimensional diffusion given by (4). Similarly, for  $1 \leq j \leq l$ , we get that

$$\begin{aligned} \tilde{\mathcal{L}}_j \tilde{g}_j(t, \tilde{x}_j) &= \xi, \quad (t, \tilde{x}_j) \in \tilde{Z}_j, \\ \tilde{g}_j|_{\partial \tilde{Z}_j} &= \frac{\eta}{2} \end{aligned}$$

has a solution  $\tilde{g}_j(\cdot) \in C_b^{1,2}(\tilde{Z}_j)$  given by

$$\tilde{g}_j(t, \tilde{x}_j) = E_{t, \tilde{x}_j} \left[ - \int_t^{\tau_M \wedge T} \xi dt + \frac{\eta}{2} \right]. \quad (26)$$

Then, adding these equations, we get that

$$\begin{aligned} \inf_{u \in U} \left( \frac{k(t, x, u) + \sum_{i=1}^m \lambda_i k_i(t, x, u)}{T} + \mathcal{L}g(t, x, u) \right) \\ + \sum_{j=1}^l \tilde{\mathcal{L}}_j \tilde{g}_j(t, \tilde{x}_j) = \xi(l+1) > 0, \\ g|_{\partial Z} = \frac{\eta}{2}, \\ \tilde{g}_j|_{\partial \tilde{Z}_j} = \frac{\eta}{2} \end{aligned}$$

has a solution  $(g, \tilde{g}_1, \dots, \tilde{g}_l) \in \mathcal{D}$  such that  $(g, \tilde{g}_1, \dots, \tilde{g}_l) \in B_{\frac{R^*}{2}}$  for  $R^*$  chosen *large enough*. Choose  $\epsilon > 0$  such that

$$(27) \quad \epsilon < \min \left( \frac{\xi(l+1)}{\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|}, \frac{\eta}{2} \right).$$

Since  $\bar{\mathcal{H}} \cap B_{\frac{R^*}{2}}$  is  $\epsilon$ -dense in  $B_{\frac{R^*}{2}}$ , we can find a  $(v, \tilde{v}_1, \dots, \tilde{v}_l) \in \bar{\mathcal{H}} \cap B_{R^*}$  such that

$$\begin{aligned} \inf_{u \in U} \left( \frac{k(t, x, u) + \sum_{i=1}^m \lambda_i k_i(t, x, u)}{T} + \mathcal{L}v(t, x, u) \right) + \sum_{j=1}^l \tilde{\mathcal{L}}_j \tilde{v}_j(t, \tilde{x}_j) \\ \geq \xi(l+1) - \epsilon(\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|) > 0, \\ |v|_{\partial Z} \leq \frac{\eta}{2} + \epsilon \leq \eta, \\ \forall j, \quad |\tilde{v}_j|_{\partial \tilde{Z}_j} \leq \frac{\eta}{2} + \epsilon \leq \eta. \end{aligned}$$

The claim follows.  $\square$

We now state the first main result of this section.

**THEOREM 4.6.** *If  $\Xi = \epsilon > 0$  and (27) holds, then  $|\beta_o - \beta_s| \leq C_1(\epsilon + \eta + \frac{1}{M})$  for some  $\eta > 0$  chosen small enough and a constant  $C_1 = C_1(M) > 0$  independent of  $\epsilon, \eta$  which remains uniformly bounded as  $M \uparrow \infty$ .*

*Proof.* Define the concave functional  $\mathcal{F}_\epsilon : \mathcal{H} \times \mathcal{R}^m \rightarrow C_b(Z)$  as

$$\mathcal{F}_\epsilon((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) \stackrel{\text{def}}{=} \mathcal{F}((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) + \epsilon \left( \|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\| \right).$$

Now define

$$\begin{aligned} G_1 \stackrel{\text{def}}{=} \{ (g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in (\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}) \times \Lambda : \mathcal{F}((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) \geq 0; \\ |g|_{\partial Z} \leq \eta; |\tilde{g}_j|_{\partial \tilde{Z}_j} \leq \eta, 1 \leq j \leq l \}, \end{aligned}$$

$$\begin{aligned} G_2 \stackrel{\text{def}}{=} \{ (g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in \overline{\mathcal{I}(B_R)} \times \Lambda : \mathcal{F}((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) \geq 0; \\ |g|_{\partial Z} \leq \eta; |\tilde{g}_j|_{\partial \tilde{Z}_j} \leq \eta, 1 \leq j \leq l \}, \end{aligned}$$

$$\begin{aligned} G_3 \stackrel{\text{def}}{=} \{ (g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in (\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}) \times \Lambda : \mathcal{F}_\epsilon((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) \geq 0; \\ |g|_{\partial Z} \leq \eta; |\tilde{g}_j|_{\partial \tilde{Z}_j} \leq \eta, 1 \leq j \leq l \}, \end{aligned}$$

and

$$G'_3 \stackrel{\text{def}}{=} \{(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in (\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}) \times \Lambda : \mathcal{F}_\epsilon((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) \geq 0; \\ |g|_{\partial Z} \leq \eta + \epsilon; |\tilde{g}_j|_{\partial \tilde{Z}_j} \leq \eta + \epsilon, 1 \leq j \leq l\}.$$

Note that these sets are convex and  $G_1, G_3, G'_3$  compact. Let

$$G_1^\epsilon \stackrel{\text{def}}{=} \{(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in (\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)})^\epsilon \times \Lambda : \mathcal{F}((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) \geq 0; \\ |g|_{\partial Z} \leq \eta; |\tilde{g}_j|_{\partial \tilde{Z}_j} \leq \eta, 1 \leq j \leq l\}$$

and

$$G_3^\epsilon \stackrel{\text{def}}{=} \{(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in (\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)})^\epsilon \times \Lambda : \mathcal{F}_\epsilon((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) \geq 0; \\ |g|_{\partial Z} \leq \eta + \epsilon; |\tilde{g}_j|_{\partial \tilde{Z}_j} \leq \eta + \epsilon, 1 \leq j \leq l\}.$$

Note that these sets are also convex and relatively compact in the  $C_b^{0,1} \times \mathcal{R}^m$  topology. It follows from Lemma 4.5 that all these sets are nonempty. As discussed before, using (\*) and Lemma 4.5, choosing  $R \geq \|\mathcal{I}\|^{-1}R^*$ , and the fact that  $\mathcal{H}$  is dense in  $\mathcal{D}$ , we see that

$$(28) \quad \beta_x = \sup_{G_2} \left( \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \right).$$

Let  $\beta_s, \beta', \beta'', \beta_\epsilon^1, \beta'_\epsilon$ , respectively, denote the supremum of the quantity in parentheses on the right over  $G_1, G_3, G'_3, G_1^\epsilon, G_3^\epsilon$ . Note that  $G_1 \subset G_2 \subset G_1^\epsilon \subset G_3^\epsilon$ . Hence,  $\beta_s \leq \beta_x \leq \beta_\epsilon^1 \leq \beta'_\epsilon$ . Therefore

$$(29) \quad |\beta_x - \beta_s| \leq |\beta_\epsilon^1 - \beta''| + |\beta'' - \beta'| + |\beta' - \beta_s|.$$

Of these,  $\beta_s \leq \beta' \leq \beta''$  as  $G_1 \subset G_3 \subset G'_3$ . An argument similar to the proof of Lemma 4.3 allows us to conclude that

$$|\beta'' - \beta'| \leq \epsilon(l+1).$$

Also,

$$\beta' - \beta_s = |\beta' - \beta_s| \leq (|\nu| \vee |\nu^\epsilon|)(\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|)\epsilon$$

by [23, Theorem 1, p. 222], where  $\nu, \nu^\epsilon \geq 0$  are the *minimizing* Lagrange multipliers (measures) for the convex programs

$$\sup_{G_1} \left( \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \right)$$

and

$$\sup_{G_3} \left( \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \right),$$



respectively. It follows from Lemma 4.4 that

$$\|\nu\| \leq \frac{\beta_s - \bar{\beta}}{\min_{x \in Z} \mathcal{F}((\bar{g}, \bar{g}_1, \dots, \bar{g}_l, \bar{\lambda}))(x)},$$

where  $\bar{\beta} = \int \bar{g} d\phi_0 + \sum_{j=1}^l \int \bar{g}_j d\tilde{\phi}_{0j} - \langle \bar{\lambda}, c \rangle$  for some  $(\bar{g}, \bar{g}_1, \dots, \bar{g}_l, \bar{\lambda}) \in G_1$  such that  $\mathcal{F}(\bar{g}, \bar{g}_1, \dots, \bar{g}_l, \bar{\lambda}) \geq \xi(l+1) - \epsilon(\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|) > 0$ . That such a  $(\bar{g}, \bar{g}_1, \dots, \bar{g}_l, \bar{\lambda})$  indeed exists is guaranteed by Lemma 4.5. Hence, using the stochastic representation of Lemma 4.5, we have, by the  $\epsilon$ -density of  $\mathcal{H} \cap \mathcal{I}(B_R)$  in  $\mathcal{I}(B_R)$  that

$$\bar{g}(t, x) \geq \inf_{u(\cdot)} E_{t,x} \left[ \int_t^{\tau_M \wedge T} \left( \frac{k(s, x(s), u(s)) + \sum_{i=1}^m \bar{\lambda}_i k_i(s, x(s), u(s))}{T} - \xi \right) dt + \frac{\eta}{2} \right] - \epsilon$$

and

$$\bar{g}_j(t, \tilde{x}_j) \geq E_{t, \tilde{x}_j} \left[ - \int_t^{\tau_M \wedge T} \xi dt + \frac{\eta}{2} \right] - \epsilon.$$

Hence we have

$$\bar{\beta} \geq C_{\bar{\lambda}} - \xi T(l+1) - \epsilon(l+1) + \frac{\eta}{2}(l+1),$$

where

$$C_{\bar{\lambda}} \stackrel{\text{def}}{=} \int \left( \inf_{u(\cdot)} E_{t,x} \left[ \int_0^{\tau_M \wedge T} \left( \frac{k(s, x(s), u(s)) + \sum_{i=1}^m \bar{\lambda}_i k_i(s, x(s), u(s))}{T} \right) dt \right] \right) d\phi_0 - \langle \bar{\lambda}, c \rangle.$$

Since  $\bar{\lambda}$  is fixed, and the  $k(\cdot)$  and the  $k_i(\cdot)$ 's are bounded functions,  $C_{\bar{\lambda}}$  is independent of  $\xi$  and we have  $\frac{C_{\bar{\lambda}}}{\xi} \rightarrow 0$  as  $\xi \uparrow \infty$ . Also,

$$\begin{aligned} \|\nu\| &\leq \frac{\beta_s - \bar{\beta}}{\min_{x \in Z} \mathcal{F}((\bar{g}, \bar{g}_1, \dots, \bar{g}_l, \bar{\lambda}))(x)} \\ &\leq \frac{\beta_s + \xi T(l+1) + \epsilon(l+1) - \frac{\eta}{2}(l+1) - C_{\bar{\lambda}}}{\xi(l+1) - \epsilon(\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|)} \\ &\leq T + \frac{\beta_s + \epsilon(l+1) + \frac{\eta}{2}(l+1) + \epsilon T(\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|) - C_{\bar{\lambda}}}{\xi(l+1) - \epsilon(\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|)}. \end{aligned} \tag{30}$$

Since  $\xi$  can be arbitrarily large, we let  $\xi \uparrow \infty$  (note that  $\epsilon$  remains fixed by (27) as  $\eta$  is fixed) to obtain

$$\|\nu\| \leq T.$$

Similarly, we have

$$\|\nu^\epsilon\| \leq \frac{\beta' - \bar{\beta}_\epsilon}{\min_{x \in Z} \mathcal{F}_\epsilon((\bar{g}_\epsilon, \bar{g}_{\epsilon 1}, \dots, \bar{g}_{\epsilon l}, \bar{\lambda}_\epsilon))(x)},$$

where  $\bar{\beta}_\epsilon = \int \bar{g}_\epsilon d\phi_0 + \sum_{j=1}^l \int \bar{g}_{\epsilon j} d\tilde{\phi}_{0j} - \langle \bar{\lambda}_\epsilon, c \rangle$  for some  $(\bar{g}_\epsilon, \bar{g}_{\epsilon 1}, \dots, \bar{g}_{\epsilon l}, \bar{\lambda}_\epsilon) \in G_3$  such that  $\mathcal{F}_\epsilon(\bar{g}_\epsilon, \bar{g}_{\epsilon 1}, \dots, \bar{g}_{\epsilon l}, \bar{\lambda}_\epsilon) \geq \xi(l+1) - \epsilon(\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|) > 0$ . We argue as above to deduce  $\|\nu^\epsilon\| \leq T$ . Thus we have

$$(31) \quad |\beta' - \beta_s| \leq \epsilon T \left( \|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\| \right).$$

On the other hand, let  $(g^*, \tilde{g}_1^*, \dots, \tilde{g}_l^*, \lambda^*) \in G_1^\epsilon$  be such that the value of  $(\int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle)$  at this point is within  $\epsilon$  of its supremum over  $G_1^\epsilon$ . Suppose  $\beta_\epsilon^1 \geq \beta''$ . Then, for any  $(g^\epsilon, \tilde{g}_1^\epsilon, \dots, \tilde{g}_l^\epsilon, \lambda^*) \in G'_3$ , we have

$$\begin{aligned} 0 \leq \beta_\epsilon^1 - \beta'' &= \sup_{G_1^\epsilon} \left( \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \right) \\ &\quad - \sup_{G'_3} \left( \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \right) \\ &\leq \int g^*(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}_j^*(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda^*, c \rangle + \epsilon \\ &\quad - \int g^\epsilon(0, \cdot) d\phi_0 - \sum_{j=1}^l \int \tilde{g}_j^\epsilon(0, \cdot) d\tilde{\phi}_{0j} + \langle \lambda^*, c \rangle \\ &= \int (g^*(0, \cdot) - g^\epsilon(0, \cdot)) d\phi_0 + \sum_{j=1}^l \int (\tilde{g}_j^*(0, \cdot) - \tilde{g}_j^\epsilon(0, \cdot)) d\tilde{\phi}_{0j} + \epsilon. \end{aligned}$$

Now since  $(g^*, \tilde{g}_1^*, \dots, \tilde{g}_l^*, \lambda^*) \in G_1^\epsilon$ , this  $(g^\epsilon, \tilde{g}_1^\epsilon, \dots, \tilde{g}_l^\epsilon, \lambda^*) \in G'_3$  can be chosen such that

$$\|(g^*, \tilde{g}_1^*, \dots, \tilde{g}_l^*) - (g^\epsilon, \tilde{g}_1^\epsilon, \dots, \tilde{g}_l^\epsilon)\|_{C_b^{1,2}} \leq \epsilon.$$

Thus we get

$$\beta_\epsilon^1 - \beta'' \leq \|g^* - g^\epsilon\|_\infty + \sum_{j=1}^l \|\tilde{g}_j^* - \tilde{g}_j^\epsilon\|_\infty + \epsilon \leq 2\epsilon.$$

Now consider the case  $\beta'' \geq \beta_\epsilon^1$ . Then, since  $G'_3 \subset G_3^\epsilon$  we have  $\beta'_\epsilon \geq \beta''$ . Hence

$$0 \leq \beta'' - \beta_\epsilon^1 \leq \beta'_\epsilon - \beta_\epsilon^1 = |\beta'_\epsilon - \beta_\epsilon^1|.$$

Arguments similar to the proofs of Lemmas 4.3 and 4.5 and the analysis exactly as in the derivation of (31) allow us to conclude that

$$|\beta'_\epsilon - \beta_\epsilon^1| \leq \epsilon \left( T(\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|) + (l+1) \right).$$

So we have

$$|\beta'_\epsilon - \beta''| \leq \epsilon \left( (T(\|\mathcal{L}\| + \|\tilde{\mathcal{L}}_1\| + \dots + \|\tilde{\mathcal{L}}_l\|) + (l+1)) \vee 2 \right).$$

Also, by Lemmas 4.1 and 4.2, we have

$$|\beta_o - \beta_d| \leq C_d(M) \frac{\bar{C}\sqrt{T}\tilde{C}(1 + E[\|x(0)\|^2])^{1/2}}{M},$$

and by Lemma 4.3, we have

$$|\beta_d - \beta_x| \leq \eta(l+1).$$

From (29), we get

$$|\beta_o - \beta_s| \leq |\beta_o - \beta_d| + |\beta_d - \beta_x| + |\beta_\epsilon^1 - \beta''| + |\beta'' - \beta'| + |\beta' - \beta_s|.$$

Combining both the inequalities above, we get the result.  $\square$

We next prove the second main result of this section. We consider a *finite* linear program that approximates the semi-infinite linear program (24). For this purpose, let  $\bar{Z}_0 \subset Z$  be a finite set such that

$$\Delta \stackrel{\text{def}}{=} \max_{y \in Z} \min_{z \in \bar{Z}_0} \|y - z\| = \delta$$

for a prescribed  $\delta > 0$ . For example,  $\bar{Z}_0$  could be a regular grid of width  $\Delta = \delta$ . By arguments similar to those used to claim a compact embedding of  $\mathcal{H}$  in  $\mathcal{D}$ , one can also claim compact embedding thereof in  $C_b^{1,3}(Z) \times C_b^{1,3}(\bar{Z}_1) \times \cdots \times C_b^{1,3}(\bar{Z}_l)$ , whence there is a common Lipschitz constant  $L$  for  $(f, \tilde{f}_1, \dots, \tilde{f}_l) \in B_R$ . Define  $\mathcal{F}_\delta : \mathcal{H} \times \mathcal{R}^m \rightarrow C_b(Z)$  as

$$\mathcal{F}_\delta((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) \stackrel{\text{def}}{=} \mathcal{F}((g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda)) + L\delta.$$

Note that this is a concave functional. Now define convex compact sets  $G_4, G_5$ , and  $G'_5$  as follows:

$$G_4 \stackrel{\text{def}}{=} \{(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in (\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}) \times \Lambda : \mathcal{F}(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \geq 0, x \in \bar{Z}_0, \\ |g|_{\bar{Z}_0 \cap \partial Z} \leq \eta; |\tilde{g}_j|_{Z_0 \cap \partial \bar{Z}_j} \leq \eta, 1 \leq j \leq l\},$$

$$G_5 \stackrel{\text{def}}{=} \{(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in (\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}) \times \Lambda : \mathcal{F}_\delta(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \geq 0, x \in Z, \\ |g|_{\partial Z} \leq \eta + L\delta; |\tilde{g}_j|_{\partial \bar{Z}_j} \leq \eta + L\delta, 1 \leq j \leq l\},$$

$$G'_5 \stackrel{\text{def}}{=} \{(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \in (\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}) \times \Lambda : \mathcal{F}_\delta(g, \tilde{g}_1, \dots, \tilde{g}_l, \lambda) \geq 0, x \in Z, \\ |g|_{\partial Z} \leq \eta; |\tilde{g}_j|_{\partial \bar{Z}_j} \leq \eta, 1 \leq j \leq l\}.$$

These are nonempty by Lemma 4.5. Let

$$\beta_f \stackrel{\text{def}}{=} \sup_{G_4} \left( \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \right), \\ \beta_\delta \stackrel{\text{def}}{=} \sup_{G_5} \left( \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \right), \\ \beta'_\delta \stackrel{\text{def}}{=} \sup_{G'_5} \left( \int (0, \cdot) d\phi_0 + \sum_{j=1}^l \int g(0, \cdot) d\phi_0 + \langle \lambda, c \rangle \right).$$

THEOREM 4.7. *Given  $\Xi = \epsilon > 0, \eta > 0$ , as in Theorem 4.6, and grid size  $\Delta = \delta > 0$ , it follows that  $|\beta_o - \beta_f| \leq C_2(\epsilon + \eta + \frac{1}{M} + \delta)$  for some constant  $C_2 = C_2(M) > 0$  independent of  $\epsilon, \eta, \delta$  which remains uniformly bounded as  $M \uparrow \infty$ .*

*Proof.* Note that  $G_1 \subset G_4 \subset G_5$ . Hence,  $\beta_s \leq \beta_f \leq \beta_\delta$ . Therefore,

$$\begin{aligned} |\beta_o - \beta_f| &\leq |\beta_o - \beta_s| + |\beta_s - \beta_f|, \\ |\beta_s - \beta_f| &\leq |\beta_s - \beta_\delta| \leq |\beta_s - \beta'_\delta| + |\beta'_\delta - \beta_\delta|. \end{aligned}$$

An argument similar to the proof of Lemma 4.3 allows us to conclude that

$$|\beta'_\delta - \beta_\delta| \leq L\delta(l+1).$$

Thus the claim follows by arguments similar to those used to prove Theorem 4.6, with  $C_2 \stackrel{\text{def}}{=} L((\|\nu\| \vee \|\nu^\delta\|) + (l+1)) \vee C_1(M)$ , where  $\nu^\delta \geq 0$  is the *minimizing* Lagrange multiplier (measure) for the convex program

$$\sup_{G_5^c} \left( \int g(0, \cdot) d\phi_0 + \sum_{j=1}^l \int \tilde{g}(0, \cdot) d\tilde{\phi}_{0j} - \langle \lambda, c \rangle \right),$$

and  $\nu, C_1(M)$  are as in Theorem 4.6. Note also that by Lemmas 4.4 and 4.5, we have, as in (30),

$$\|\nu^\delta\| \leq T.$$

Thus the claim follows.  $\square$

**5. Estimation of sample complexity.** In this section we calculate the sample complexity for RKHS functions, i.e., the minimum number of sample points (and hence minimum number of basis functions) required to approximate the cost function to a given degree of accuracy. We shall assume the functions to be defined over a set  $Z$  which is now a compact subset of  $[0, T] \times \Omega_M$  with  $C^\infty$  boundary. Hence, by definition, the  $\tilde{Z}_j$ 's are also compact with a  $C^\infty$  boundary. This additional regularity of the boundary is needed for technical reasons. Refer to section 4 for definitions.

Note that the map  $\mathcal{H} \xrightarrow{\mathcal{I}} \mathcal{D}$  can be factored into

$$\mathcal{H} \xrightarrow{\mathcal{J}} \mathcal{W} \xhookrightarrow{J} \mathcal{D},$$

where  $\mathcal{I} = J\mathcal{J}$ . Given a normed space  $S$  with norm  $\|\cdot\|_S$ , let  $B_1 \stackrel{\text{def}}{=} \{x \in S : \|x\|_S \leq 1\}$ . Also, given Banach spaces  $E$  and  $F$  and a linear map  $T : E \rightarrow F$ , let  $e_k(T) \stackrel{\text{def}}{=} e_k(T(B_1))$  for  $k \geq 1$ , where the *entropy number*  $e_k(S)$  is defined as

$$(32) \quad e_k(S) \stackrel{\text{def}}{=} \inf \left\{ \epsilon > 0 : S \subset \bigcup_{j=1}^{2^k-1} (b_j + \epsilon B_1), b_1, \dots, b_{2^k-1} \in S \right\}.$$

Thus under the product topology in  $\mathcal{W}$  and  $\mathcal{D}$ , by taking  $s_1 = m + 2(\tilde{m}_j + 2)$ ,  $p_1 = s_2 = 2, p_2 = \infty$  in a very general “theorem” of [11, p. 105], we get

$$(33) \quad e_k(J) \leq \max \left( C(Z, m) \left( \frac{1}{k} \right)^{m/(d+1)}, \max_{1 \leq j \leq l} \left( C(\tilde{Z}_j, \tilde{m}_j) \left( \frac{1}{k} \right)^{\tilde{m}_j/(\tilde{d}_j+1)} \right) \right)$$

for  $k \geq 1$ , where  $m, d$  and  $\tilde{m}_j, \tilde{d}_j$  satisfy the conditions in definition (22) of  $\mathcal{W}$  above and  $C$  is a constant depending on the  $(Z, m)$  or the  $(\tilde{Z}_j, \tilde{m}_j)$ 's, but not on  $k$ .

Given some  $\eta > 0$  and a bounded set  $A$  in some normed space  $S$  as above, the *covering number*  $\mathcal{N}(A, \eta)$  is the minimum number of  $\eta$ -balls required to cover the set  $A$ , i.e.,

$$(34) \quad \mathcal{N}(A, \eta) \stackrel{\text{def}}{=} \inf \left\{ N > 0 : A \subset \bigcup_{j=1}^N (b_j + \eta B_1), b_1, \dots, b_N \in A \right\},$$

where  $B_1$  is a unit ball in  $S$  as defined above.

Given  $\epsilon > 0$ , the sample complexity is defined as  $d^*(\epsilon) \stackrel{\text{def}}{=}$  the *minimum*  $|\bar{Z}| + \sum_{j=1}^l |\tilde{Z}_j|$  such that the *sampling approximation error*  $\Xi$  satisfies  $\Xi \leq \epsilon$ . We look for a good estimate  $d(\epsilon)$  of  $d^*(\epsilon)$ . First, we state and prove the following theorem.

**THEOREM 5.1.** *Let  $C' \stackrel{\text{def}}{=} \|\mathcal{J}\|$ ,  $C'' \stackrel{\text{def}}{=} \|\mathcal{I}\|$ , and*

$$r \stackrel{\text{def}}{=} \frac{1}{\min(m/(d+1), \min_{1 \leq j \leq l} \tilde{m}_j/(\tilde{d}_j+1))}.$$

*Then*

$$\text{A.} \quad \ln(\mathcal{N}(\overline{\mathcal{I}(B_R)}, 2\epsilon)) < 1 + \left( \frac{RC' \max(C(Z, m), \max_{1 \leq j \leq l} C(\tilde{Z}_j, \tilde{m}_j))}{2\epsilon} \right)^r.$$

$$\text{B.} \quad d^*(\epsilon) \geq \frac{\ln(\mathcal{N}(\overline{\mathcal{I}(B_R)}, 2\epsilon))}{\ln(\frac{4RC''}{\epsilon})}.$$

*Proof.* A. By (33), we get

$$\begin{aligned} e_k(\mathcal{I}) &= e_k(J\mathcal{J}) \leq e_k(J)\|\mathcal{J}\| \\ &\leq C' \max \left( C(Z, m) \left( \frac{1}{k} \right)^{m/(d+1)}, \max_{1 \leq j \leq l} \left( C(\tilde{Z}_j, \tilde{m}_j) \left( \frac{1}{k} \right)^{\tilde{m}_j/(\tilde{d}_j+1)} \right) \right) \\ (35) \quad &\leq C' \max \left( C(Z, m), \max_{1 \leq j \leq l} C(\tilde{Z}_j, \tilde{m}_j) \right) \left( \frac{1}{k} \right)^{1/r} \end{aligned}$$

for  $k \geq 1$ . By [10, Lemma 4, p. 16] we have

$$(36) \quad \mathcal{N}(\overline{\mathcal{I}(B_R)}, 2\epsilon) \leq 2^k - 1 \iff e_k(\overline{\mathcal{I}(B_R)}) = Re_k(\mathcal{I}) \leq 2\epsilon.$$

Let

$$k = \left\lceil \left( \frac{RC' \max(C(Z, m), \max_{1 \leq j \leq l} C(\tilde{Z}_j, \tilde{m}_j))}{2\epsilon} \right)^r \right\rceil.$$

Then from (35) and (36) we have (see [10, Proposition 6, p. 16]),

$$\begin{aligned} \ln(\mathcal{N}(\overline{\mathcal{I}(B_R)}, 2\epsilon)) &= \ln \left( \mathcal{N} \left( \overline{\mathcal{I}(B_1)}, \frac{2\epsilon}{R} \right) \right) \\ &< k < 1 + \left( \frac{RC' \max(C(Z, m), \max_{1 \leq j \leq l} C(\tilde{Z}_j, \tilde{m}_j))}{2\epsilon} \right)^r. \end{aligned} \quad (37)$$

B. Since  $\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}$  is a finite dimensional Banach space, we have by [10, Proposition 5, p. 15] that

$$(38) \quad \ln \mathcal{N}(\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}, \epsilon) \leq d^*(\epsilon) \ln \left( \frac{4R^*}{\epsilon} \right) \leq d^*(\epsilon) \ln \left( \frac{4RC'''}{\epsilon} \right),$$

where the last inequality follows from the fact  $\|\cdot\|_{C_b^{1,2}} \leq C'' \|\cdot\|_{\mathcal{H}}$ . Also since  $\Xi \leq \epsilon$ , i.e., every point in  $\overline{\mathcal{I}(B_R)}$  is at most  $\epsilon$ -distance away from some point in  $\bar{\mathcal{H}} \cap \overline{\mathcal{I}(B_R)}$ , we have

$$(39) \quad \mathcal{N}(\overline{\mathcal{I}(B_R)}, 2\epsilon) \leq \mathcal{N}(\overline{\mathcal{I}(B_R)} \cap \bar{\mathcal{H}}, \epsilon).$$

Assuming  $0 < \epsilon < \frac{R}{2}$  since  $\epsilon$  is sufficiently small, we see, combining (38) and (39), that

$$(40) \quad d^*(\epsilon) \geq \frac{\ln(\mathcal{N}(\overline{\mathcal{I}(B_R)}, 2\epsilon))}{\ln(\frac{4RC'''}{\epsilon})}.$$

Hence we have the claim.  $\square$

*Remark 2.* Note that (37) and (40) together suggest that  $d(\epsilon)$  given by

$$d(\epsilon) = \frac{1 + \left( \frac{RC' \max(C(Z, m), \max_{1 \leq j \leq l} C(\tilde{Z}_j, \tilde{m}_j))}{2\epsilon} \right)^{1/\min(m/(d+1), \min_{1 \leq j \leq l} \tilde{m}_j/(\tilde{d}_j+1))}}{\ln(\frac{4RC'''}{\epsilon})}$$

is a reasonable estimate for  $d^*(\epsilon)$ .

*Remark 3.* The above bound on the sample complexity shows that by choosing  $m > \frac{d+1}{2}$ ,  $\tilde{m}_j > \frac{\tilde{d}_j+1}{2}$  large enough we can make the exponent almost a constant independent of the dimension of the underlying state space. On the other hand, the dependence of the constant  $C$  on  $d, m$  (respectively,  $(\tilde{d}_j, \tilde{m}_j)$ 's) may be exponential as is the case for approximation of  $C^0$  functions on  $[0, 1]^d$  by Gaussian kernels (see [30, Example 4, p. 23]).

**6. Markov-modulated diffusions.** Now we briefly sketch the extension of the foregoing to Markov-modulated diffusions, a popular model when regime-switching phenomena are present. Let  $\alpha(\cdot)$  be a continuous time stationary Markov chain taking values in a finite state space  $\mathcal{S} = \{1, \dots, N\}$ . Let  $W(\cdot)$  be a standard Brownian motion in  $\mathcal{R}^d$  independent of  $\alpha(\cdot)$ . The Markov chain has a generator  $Q^* = [[q_{ij}^*]]_{1 \leq i, j \leq N} \in \mathcal{R}^{N \times N}$ , i.e., its transition probabilities are given by

$$(41) \quad P(\alpha(t + \delta t) = j | \alpha(t) = i) = q_{ij}^* \delta t + o(\delta t), \quad i \neq j, \quad i, j \in \mathcal{S}.$$

We shall assume that our underlying state process is a  $d$ -dimensional process  $x(\cdot) = [x_1(\cdot), \dots, x_d(\cdot)]^T$  described by the SDE

$$(42) \quad dx(t) = m^*(x(t), \alpha(t))dt + \sigma^*(x(t), \alpha(t))dW(t), \quad x(0) = x_0, \quad \alpha(0) = i_0, \quad t \geq 0,$$

where the drift vector  $m^*(\cdot) : \mathcal{R}^d \times \mathcal{S} \rightarrow \mathcal{R}^d$  and the diffusion matrix  $\sigma^*(\cdot) : \mathcal{R}^d \times \mathcal{S} \rightarrow \mathcal{R}^{d \times d}$  are assumed to be Lipschitz. (This ensures the well-posedness of (42) and can be relaxed in special cases.) The law of  $(x_0, i_0)$  is fixed; say,  $\phi_0$ . See section 2 for definitions. The two forms of uncertainty we considered in section 2 are captured by the following two basic formulations.

**Problem 1: Model uncertainty.** We suppose that  $m^*, \sigma^*$ , and  $Q^* = [[q_{ij}^*]]$  are not known exactly, but only approximately. The latter is captured by the conditions

$$(43) \quad m^*(x, i) \in m(x, i, U), \quad (\sigma^*(\sigma^*)^T)(x, i) \in (\sigma\sigma^T)(x, i, U), \quad Q^* \in Q(U) \quad \forall x, i,$$

where

- $U$  is a prescribed compact metric space;
- $m(\cdot, i, \cdot) = [m_1(\cdot, i, \cdot), \dots, m_d(\cdot, i, \cdot)]^T : \mathcal{R}^d \times \mathcal{S} \times U \rightarrow \mathcal{R}^d$  is continuous and Lipschitz in  $x \in \mathcal{R}^d$  uniformly w.r.t.  $u \in U$ , and, in addition,  $m(x, i, U)$  is convex for each  $x$ , for each  $i \in \mathcal{S}$ ;
- $\sigma(\cdot, i, \cdot) = [[\sigma_{ij}(\cdot, i, \cdot)]]_{1 \leq i, j \leq d} : \mathcal{R}^d \times \mathcal{S} \times U \rightarrow \mathcal{R}^{d \times d}$  is continuous and Lipschitz in  $x \in \mathcal{R}^d$  uniformly w.r.t.  $u \in U$ , and, in addition,  $(\sigma\sigma^T)(x, i, U)$  is convex for each  $x$ , for each  $i \in \mathcal{S}$  and satisfies the *nondegeneracy* condition as in (3);
- $Q(\cdot) = [[q_{ij}(\cdot)]]_{1 \leq i, j \leq N} : U \rightarrow \mathcal{R}^{N \times N}$  is bounded and convex.

Now using familiar *measurable selection* arguments as in section 2, we can rewrite the SDE (42) as

$$(44) \quad dx(t) = m(x(t), \alpha(t), u(t))dt + \sigma(x(t), \alpha(t), u(t))dW(t), \quad t \geq 0,$$

where the Markov chain has the controlled extended generator

$$Q(\cdot) = [[q_{ij}(\cdot)]]_{1 \leq i, j \leq N} : U \rightarrow \mathcal{R}^{N \times N}$$

and transition probabilities

$$(45) \quad P(\alpha(t + \delta t) = j | \alpha(t) = i) = q_{ij}(u(t))\delta t + o(\delta t), \quad i \neq j, \quad i, j \in \mathcal{S}.$$

Associated with  $x(\cdot)$  is the controlled extended generator  $\mathcal{A}$ :

$$\begin{aligned} \mathcal{A}g(t, x, i, u) &\stackrel{\text{def}}{=} \frac{\partial g}{\partial t}(t, x, i) + \langle m(x, i, u), \nabla g(t, x, i) \rangle \\ &\quad + \frac{1}{2} \text{tr}(\sigma(x, i, u)\sigma^T(x, i, u)\nabla^2 g(t, x, i)) \\ &\quad + \sum_{j=1}^N q_{ij}(u)g(t, x, j) \end{aligned}$$

for all  $f \in \mathcal{D}(\mathcal{A}) \stackrel{\text{def}}{=} \bigcup_{i \in \mathcal{S}} \mathcal{D}_i(\mathcal{A})$ , where  $\mathcal{D}_i(\mathcal{A}) \stackrel{\text{def}}{=} \{f(\cdot, \cdot, i) \in C_b^{1,2}(Z) : f(T, \cdot, i) \equiv 0\}$ .

**Problem 2: Unmodelled dynamics.** As in section 2, we *separately* observe for a given integer  $l \geq 1$  only  $\tilde{x}_j(\cdot)$ ,  $j = 1, \dots, l$  (see definition (8)) and mimic it by a  $\tilde{d}_j$ -dimensional process given by (say)

$$(46) \quad d\hat{x}_j(t) = \tilde{m}_j(t, \hat{x}_j(t), \tilde{\alpha}_j(t))dt + \tilde{\sigma}_j(t, \hat{x}_j(t), \tilde{\alpha}_j(t))d\tilde{W}_j(t), \quad 1 \leq j \leq l,$$

where, as before,  $\tilde{\alpha}_j(\cdot)$  is the *correspondingly observed*  $\mathcal{S}$ -valued Markov chain with generator  $\tilde{Q}_j(\cdot) = [[\tilde{q}_{kl}^j(\cdot)]]_{1 \leq k, l \leq N} : [0, T] \rightarrow \mathcal{R}^{N \times N}$  and transition probabilities

$$(47) \quad \tilde{P}_j(\tilde{\alpha}_j(t + \delta t) = l | \tilde{\alpha}_j(t) = k) = \tilde{q}_{kl}^j(t)\delta t + o(\delta t), \quad k \neq l, \quad k, l \in \mathcal{S}, \quad 1 \leq j \leq l.$$

We view (46) and (47) as the *model fitted to the observed data*  $(\tilde{x}_j(t), \tilde{\alpha}_j(t))$ . Define

the extended generator  $\tilde{\mathcal{A}}_j$  of  $\hat{x}_j(\cdot)$  by

$$\begin{aligned}\tilde{\mathcal{A}}_j f(t, x, i) &\stackrel{\text{def}}{=} \frac{\partial f}{\partial t}(t, x, i) + \langle \tilde{m}_j(t, x, i), \nabla f(t, x, i) \rangle \\ &\quad + \frac{1}{2} \text{tr}(\tilde{\sigma}_j(t, x, i) \tilde{\sigma}_j^T(t, x, i) \nabla^2 f(t, x, i)) \\ &\quad + \sum_{k=1}^N \tilde{q}_{ik}^j(t) f(t, x, k)\end{aligned}$$

for all  $f \in \mathcal{D}(\tilde{\mathcal{A}}_j) \stackrel{\text{def}}{=} \bigcup_{i \in \mathcal{S}} \mathcal{D}_i(\tilde{\mathcal{A}}_j)$ , where  $\mathcal{D}_i(\tilde{\mathcal{A}}_j) \stackrel{\text{def}}{=} \{f(\cdot, \cdot, i) \in C_b^{1,2}(\tilde{Z}_j) : f(T, \cdot, i) \equiv 0\}$ .

Now define *occupation measures* as in section 2 (see definitions (5) and (9)). Replacing  $\mathcal{L}$  by  $\mathcal{A}$  and the  $\tilde{\mathcal{L}}_j$ 's by  $\tilde{\mathcal{A}}_j$ 's and using the general results of [5], we get the characterizations of the corresponding  $\mathcal{M}$ ,  $\mathcal{M}'$ , and  $\mathcal{M}^*$ . Then the minimization problem becomes

$$\text{minimize } \int k d\nu \text{ over } \mathcal{M}^*.$$

The rest of the analysis is exactly similar and the details are omitted.

In conclusion, we note some potential extensions that hold promise:

1. extensions to SDEs driven by both a Brownian motion and a point process;
2. an additional layer of approximation replacing the differential operators by their finite difference approximations;
3. families of approximating functions other than those from an RKHS.

**Acknowledgment.** The first author thanks Prof. K. T. Joseph for helpful discussions.

#### REFERENCES

- [1] R. A. ADAMS AND J. J. F. FOURNIER, *Sobolev Spaces*, 2nd ed., Academic Press, New York, 2003.
- [2] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces*, John Wiley, Chichester, 1987.
- [3] A. BASU, V. S. BORKAR, AND D. GARG, *A linear programming approach to risk estimation*, in Proceedings of 1st Conference on Advances in Control and Optimization of Dynamical Systems, Bangalore, 2007. Available online at <http://www.aero.iisc.ernet.in/acods2007/ppt/acods2007.pdf>.
- [4] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.
- [5] A. BHATT AND V. S. BORKAR, *Occupation measures for controlled Markov processes: Characterization and optimality*, Ann. Probab., 24 (1996), pp. 1531–1562.
- [6] V. S. BORKAR, *Optimal Control of Diffusion Processes*, Pitman Res. Notes in Math. 203, Longman Scientific and Technical, Harlow, UK, 1989.
- [7] V. S. BORKAR, *Probability Theory: An Advanced Course*, Springer-Verlag, New York, 1995.
- [8] V. S. BORKAR, *Convex analytic methods in Markov decision processes*, in Handbook of Markov Decision Processes: Methods and Applications, E. A. Feinberg and A. Shwartz, eds., Kluwer Academic Publishers, Boston, MA, 2002, pp. 347–375.
- [9] V. S. BORKAR AND M. K. GHOSH, *Controlled diffusions with constraints*, J. Math. Anal. Appl., 152 (1990), pp. 88–108.
- [10] F. CUCKER AND S. SMALE, *On the mathematical foundations of learning*, Bull. AMS, 39 (2002), pp. 1–49.
- [11] D. E. EDMUNDS AND H. TRIEBEL, *Function Spaces, Entropy Numbers, and Differential Operators*, Cambridge University Press, Cambridge, UK, 1996.
- [12] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.



- [13] V. GAITSGORY AND S. ROSSOMAKHINE, *Linear programming approach to deterministic long run average problems of optimal control*, SIAM J. Control Optim., 44 (2006), pp. 2006–2037.
- [14] K. HELMES AND R. H. STOCKBRIDGE, *Numerical comparison of control and verification of optimality for stochastic control problems*, J. Optim. Theory Appl., 106 (2000), pp. 107–127.
- [15] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.
- [16] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.
- [17] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Approximation schemes for infinite linear programs*, SIAM J. Optim., 8 (1998), pp. 973–988.
- [18] O. HERNÁNDEZ-LERMA AND J.-B. LASSERRE, *Linear programming approximations for Markov control processes in metric spaces*, Acta Appl. Math., 51 (1998), pp. 123–139.
- [19] O. HERNÁNDEZ-LERMA AND J.-B. LASSERRE, *The linear programming approach*, in Handbook of Markov Decision Processes: Methods and Applications, E. A. Feinberg and A. Shwartz, eds., Kluwer Academic Publishers, Boston, MA, 2002, pp. 377–407.
- [20] A. J. HEUNIS, *On the prevalence of stochastic differential equations with unique strong solutions*, Ann. Probab., 14 (1986), pp. 653–662.
- [21] L. C. M. KALLENBERG, *Linear Programming and Finite Markovian Control Problems*, Math. Centrum Tracts 148, Mathematisch Centrum, Amsterdam, 1983.
- [22] D. LANDO, *Credit Risk Modeling*, Princeton University Press, Princeton, NJ, 2005.
- [23] D. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [24] M. S. MENDIONDO AND R. H. STOCKBRIDGE, *Approximation of infinite-dimensional linear programming problems which arise in stochastic control*, SIAM J. Control Optim., 36 (1998), pp. 1448–1472.
- [25] S. SMALE AND D.-X. ZHOU, *Estimating the approximation error in learning theory*, Anal. Appl., 1 (2003), pp. 1–25.
- [26] S. SMALE AND D.-X. ZHOU, *Shannon sampling and function reconstruction from point values*, Bull. AMS, 41 (2004), pp. 279–305.
- [27] S. SMALE AND D.-X. ZHOU, *Shannon sampling II: Connections to learning theory*, Appl. Comput. Harmon. Anal., 19 (2005), pp. 285–302.
- [28] R. H. STOCKBRIDGE, *Time-average control of Martingale problems: A linear programming formulation*, Ann. Probab., 18 (1990), pp. 206–217.
- [29] C. VILLANI, *Topics in Optimal Transportation*, AMS, Providence, RI, 2003.
- [30] D.-X. ZHOU, *The covering number in learning theory*, J. Complexity, 18 (2002), pp. 737–769.

# A PRIORI ERROR ESTIMATES FOR SPACE-TIME FINITE ELEMENT DISCRETIZATION OF PARABOLIC OPTIMAL CONTROL PROBLEMS PART II: PROBLEMS WITH CONTROL CONSTRAINTS\*

DOMINIK MEIDNER<sup>†</sup> AND BORIS VEXLER<sup>‡</sup>

**Abstract.** This paper is the second part of our work on a priori error analysis for finite element discretizations of parabolic optimal control problems. In the first part [*SIAM J. Control Optim.*, 47 (2008), pp. 1150–1177] problems without control constraints were considered. In this paper we derive a priori error estimates for space-time finite element discretizations of parabolic optimal control problems with pointwise inequality constraints on the control variable. The space discretization of the state variable is done using usual conforming finite elements, whereas the time discretization is based on discontinuous Galerkin methods. For the treatment of the control discretization we discuss different approaches, extending techniques known from the elliptic case.

**Key words.** optimal control, parabolic equations, error estimates, finite elements, pointwise inequality constraints

**AMS subject classifications.** 49N10, 49M25, 65M15, 65M60

**DOI.** 10.1137/070694028

**1. Introduction.** In this paper we develop a priori error analysis for space-time finite element discretizations of parabolic optimization problems. We consider the following linear-quadratic optimal control problem for the state variable  $u$  and the control variable  $q$  involving pointwise control constraints:

$$(1.1a) \quad \text{Minimize } J(q, u) = \frac{1}{2} \int_0^T \int_{\Omega} (u(t, x) - \hat{u}(t, x))^2 dx dt + \frac{\alpha}{2} \int_0^T \int_{\Omega} q(t, x)^2 dx dt$$

subject to

$$(1.1b) \quad \begin{aligned} \partial_t u - \Delta u &= f + q && \text{in } (0, T) \times \Omega, \\ u(0) &= u_0 && \text{in } \Omega \end{aligned}$$

and subject to

$$(1.1c) \quad q_a \leq q(t, x) \leq q_b \quad \text{a.e. in } (0, T) \times \Omega,$$

combined with either homogeneous Dirichlet or homogeneous Neumann boundary conditions on  $(0, T) \times \partial\Omega$ . A precise formulation of this problem including a functional analytic setting is given in the next section.

---

\*Received by the editors June 8, 2007; accepted for publication (in revised form) November 27, 2007; published electronically March 21, 2008. The first author was supported by the German research foundation DFG through the International Research Training Group 710 “Complex Processes: Modeling, Simulation, and Optimization.” The second author was partially supported by the Austrian Science Fund FWF project P18971-N18 “Numerical analysis and discretization strategies for optimal control problems with singularities.”

<http://www.siam.org/journals/sicon/47-3/69402.html>

<sup>†</sup>Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, INF 294, 69120 Heidelberg, Germany (dominik.meidner@iwr.uni-heidelberg.de).

<sup>‡</sup>Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenberger Straße 69, 4040 Linz, Austria (boris.vexler@oeaw.ac.at).

Although the a priori error analysis for finite element discretization of optimal control problems governed by elliptic equations is discussed in many publications (see, e.g., [10, 12, 1, 13, 21, 7]), there are only a few published results on this topic for parabolic problems; see [19, 28, 16, 18, 23].

In the first part of our work on a priori error analysis of parabolic optimal control problems [20], we developed a priori error estimates for problems without control constraints. The consideration of control constraints (1.1c) leads to many additional difficulties. In the absence of inequality constraints the regularity of the optimal solution  $(\bar{q}, \bar{u})$  of (1.1a)–(1.1b) is restricted only by the regularity of the domain  $\Omega$ ; by the regularity of the data  $f, u_0, \hat{u}$ ; and possibly by some compatibility conditions. Therefore, in this case it is reasonable to assume high regularity of  $(\bar{q}, \bar{u})$  leading to a corresponding order of convergence of the finite element discretization; see the discussion in [20].

The presence of control constraints (1.1c) leads to a stronger restriction of the regularity of the optimal solution, which is often reflected in a reduction of the order of convergence of the finite element discretization. For a discussion of the regularity of solutions to parabolic optimal control problems with control constraints, we refer, e.g., to [15].

In order to describe the claims and challenges of a priori error analysis for finite element discretization of (1.1), we first recall some corresponding results in the elliptic case. Using a finite element discretization with discretization parameter  $h$ , one can define a discretized optimal control problem with the discrete solution  $(\bar{q}_h, \bar{u}_h)$ .

Many authors made an effort to analyze the behavior of  $\|\bar{q} - \bar{q}_h\|_{L^2(\Omega)}$  with respect to  $h$ : In the first papers concerning approximation of elliptic optimal control problems (see [10, 12]), the convergence order  $\mathcal{O}(h)$  was established using a cellwise constant discretization of the control variable; see also [8, 1, 7]. For finite element discretization of the control variable by (bi-/tri-)linear  $H^1$ -conforming elements, the convergence order  $\mathcal{O}(h^{\frac{3}{2}})$  can be shown; see, e.g., [5, 24, 2]. Recently, two approaches achieving  $\mathcal{O}(h^2)$ -convergence for the error in the control variable have been established; see [13, 21]. In [13] a *variational approach* is proposed, where no explicit discretization of the control variable is used. The discrete control variable is obtained by the projection of the discretized adjoint state on the set of admissible controls. In [21] a cellwise constant discretization is utilized, and a *postprocessing step* is used to obtain the desired accuracy. The latter technique is extended to optimal control of the Stokes equations in [25].

For discretization of parabolic problems such as (1.1), the state variable has to be discretized with respect to space and time leading to two discretization parameters  $h, k$ ; see section 3 for a detailed description. The solution of the discretized optimal control problem is denoted by  $(\bar{q}_\sigma, \bar{u}_\sigma)$ , where  $\sigma = (k, h, d)$  is a general discretization parameter and  $d$  denotes an abstract discretization parameter for the control space; cf. [20].

The main purpose of this paper is to analyze the behavior of  $\|\bar{q} - \bar{q}_\sigma\|_{L^2(0,T;L^2(\Omega))}$  with respect to all involved discretization parameters. Our aim is to discuss the following four approaches for the discretization of the control variable, which extend some techniques known from the elliptic case:

1. Discretization using cellwise constant ansatz functions with respect to space and time. In this case we obtain, similar to [16, 18], the order of convergence  $\mathcal{O}(h + k)$ : The result is obtained under weaker regularity assumptions than

in [16, 18]. Moreover, we separate the influences of the spatial and temporal regularity on the discretization error; see Corollary 5.3.

2. Discretization using cellwise (bi-/tri-)linear,  $H^1$ -conforming finite elements in space, and piecewise constant functions in time: For this type of discretization we obtain the improved order of convergence  $\mathcal{O}(k + h^{\frac{3}{2} - \frac{1}{p}})$ ; see Corollary 5.8. Here,  $p$  depends on the regularity of the adjoint solution. In two space dimensions we show the assertion for any  $p < \infty$ , whereas in three space dimensions the result is proved for  $p \leq 6$ . Under an additional regularity assumption, one can choose  $p = \infty$  leading to  $\mathcal{O}(k + h^{\frac{3}{2}})$ . Again the influences of spatial and temporal regularity as well as of spatial and temporal discretization are clearly separated.
3. The discretization following the variational approach from [13], where no explicit discretization of the control variable is used: In this case we obtain an optimal result  $\mathcal{O}(k + h^2)$ ; see Corollary 5.11. The usage of this approach requires a nonstandard implementation and more involved stopping criteria for optimization algorithms, since the control variable does not lie in any finite element space associated with the given mesh. However, there are no additional difficulties caused by the time discretization.
4. The postprocessing strategy extending the technique from [21] to parabolic problems: In this case we use the cellwise constant ansatz functions with respect to space and time. For the discrete solution  $(\bar{q}_\sigma, \bar{u}_\sigma)$ , a postprocessing step based on a projection formula is proposed leading to an approximation  $\tilde{q}_\sigma$  with order of convergence  $\|\bar{q} - \tilde{q}_\sigma\|_{L^2(0,T;L^2(\Omega))} = \mathcal{O}(k + h^{2 - \frac{1}{p}})$ ; see Corollary 5.17. Here,  $p$  can be chosen as discussed for the cellwise linear discretization. Under an additional regularity assumption, one can also choose  $p = \infty$  leading to  $\mathcal{O}(k + h^2)$ .

The paper is organized as follows. In the next section, we present a functional analytic setting for the optimal control problem (1.1) and discuss optimality conditions and the regularity of optimal solutions. Section 3 is devoted to the discretization of the considered optimal control problem. Therein, we address the temporal and spatial discretizations of the state equation by Galerkin finite element methods. Moreover, we give a detailed presentation of the four possibilities for discretizing the control variable introduced above. In section 4 we provide basic results on stability and approximation quality proved in the first part of this article [20]. In section 5 we develop our main results on a priori error analysis for the four mentioned types of control discretizations. Finally, we illustrate our theoretical results by numerical experiments.

**2. Optimization.** In this section we briefly discuss the precise formulation of the optimization problem under consideration. Furthermore, we recall theoretical results on existence, uniqueness, and regularity of optimal solutions as well as optimality conditions.

To set up a weak formulation of the state equation (1.1b), we introduce the following notation: For a convex polygonal domain  $\Omega \subset \mathbb{R}^n$ ,  $n = 2, 3$ , we denote  $V$  to be either  $H^1(\Omega)$  or  $H_0^1(\Omega)$  depending on the prescribed type of boundary conditions (homogeneous Neumann or homogeneous Dirichlet). Together with  $H = L^2(\Omega)$ , the Hilbert space  $V$  and its dual  $V^*$  build a Gelfand triple  $V \hookrightarrow H \hookrightarrow V^*$ . Here and in what follows, we employ the usual notion for Lebesgue and Sobolev spaces.

For a time interval  $I = (0, T)$  we introduce the state space

$$X := \{v | v \in L^2(I, V) \text{ and } \partial_t v \in L^2(I, V^*)\}$$

and the control space

$$Q = L^2(I, L^2(\Omega)).$$

In addition, we use the following notation for the inner products and norms on  $L^2(\Omega)$  and  $L^2(I, L^2(\Omega))$ :

$$\begin{aligned} (v, w) &:= (v, w)_{L^2(\Omega)}, & (v, w)_I &:= (v, w)_{L^2(I, L^2(\Omega))}, \\ \|v\| &:= \|v\|_{L^2(\Omega)}, & \|v\|_I &:= \|v\|_{L^2(I, L^2(\Omega))}. \end{aligned}$$

In this setting, a standard weak formulation of the state equation (1.1b) for given control  $q \in Q$ ,  $f \in L^2(I, H)$ , and  $u_0 \in V$  reads as follows: Find a state  $u \in X$  satisfying

$$\begin{aligned} (2.1) \quad & (\partial_t u, \varphi)_I + (\nabla u, \nabla \varphi)_I = (f + q, \varphi)_I \quad \forall \varphi \in X, \\ & u(0) = u_0. \end{aligned}$$

As in Proposition 2.1 in [20] the following result on existence and regularity holds.

**PROPOSITION 2.1.** *For fixed control  $q \in Q$ ,  $f \in L^2(I, H)$ , and  $u_0 \in V$  there exists a unique solution  $u \in X$  of problem (2.1). Moreover, the solution exhibits the improved regularity*

$$u \in L^2(I, H^2(\Omega) \cap V) \cap H^1(I, L^2(\Omega)) \hookrightarrow C(\bar{I}, V).$$

Furthermore, the stability estimate

$$\|\partial_t u\|_I + \|\nabla^2 u\|_I \leq C\{\|f + q\|_I + \|\nabla u_0\|\}$$

holds.

To formulate the optimal control problem we introduce the admissible set  $Q_{\text{ad}}$ , collecting the inequality constraints (1.1c) as

$$Q_{\text{ad}} := \{q \in Q \mid q_a \leq q(t, x) \leq q_b \quad \text{a.e. in } I \times \Omega\},$$

where the bounds  $q_a, q_b \in \mathbb{R}$  fulfill  $q_a < q_b$ .

The weak formulation of the optimal control problem (1.1) is given as follows:

$$(2.2) \quad \text{Minimize } J(q, u) := \frac{1}{2}\|u - \hat{u}\|_I^2 + \frac{\alpha}{2}\|q\|_I^2 \text{ subject to (2.1) and } (q, u) \in Q_{\text{ad}} \times X,$$

where  $\hat{u} \in L^2(I, H)$  is a given desired state and  $\alpha > 0$  is the regularization parameter.

**PROPOSITION 2.2.** *For given  $f, \hat{u} \in L^2(I, H)$ ,  $u_0 \in V$ , and  $\alpha > 0$  the optimal control problem (2.2) admits a unique solution  $(\bar{q}, \bar{u}) \in Q_{\text{ad}} \times X$ .*

For the standard proof we refer, e.g., to [17].

The existence result for the state equation in Proposition 2.1 ensures the existence of a control-to-state mapping  $q \mapsto u = u(q)$  defined through (2.1). By means of this mapping we introduce the reduced cost functional  $j: Q \rightarrow \mathbb{R}$ :

$$j(q) := J(q, u(q)).$$

The optimal control problem (2.2) can then be equivalently reformulated as follows:

$$(2.3) \quad \text{Minimize } j(q) \text{ subject to } q \in Q_{\text{ad}}.$$

The first order necessary optimality condition for (2.3) reads as

$$(2.4) \quad j'(\bar{q})(\delta q - \bar{q}) \geq 0 \quad \forall \delta q \in Q_{\text{ad}}.$$

Due to the linear-quadratic structure of the optimal control problem, this condition is also sufficient for optimality.

Utilizing the adjoint state equation for  $z = z(q) \in X$  given by

$$(2.5) \quad \begin{aligned} -(\varphi, \partial_t z)_I + (\nabla \varphi, \nabla z)_I &= (\varphi, u(q) - \hat{u})_I \quad \forall \varphi \in X, \\ z(T) &= 0, \end{aligned}$$

the first derivative of the reduced cost functional can be expressed as

$$(2.6) \quad j'(q)(\delta q) = (\alpha q + z(q), \delta q)_I.$$

The second derivative  $j''(q)(\cdot, \cdot)$  is independent of  $q$  and positive definite, i.e.,

$$(2.7) \quad j''(q)(p, p) \geq \alpha \|p\|_I^2 \quad \forall p \in Q.$$

Using a pointwise projection on the admissible set  $Q_{\text{ad}}$ ,

$$(2.8) \quad P_{Q_{\text{ad}}} : Q \rightarrow Q_{\text{ad}}, \quad P_{Q_{\text{ad}}}(r)(t, x) = \max(q_a, \min(q_b, r(t, x))),$$

the optimality condition (2.4) can be expressed as

$$(2.9) \quad \bar{q} = P_{Q_{\text{ad}}} \left( -\frac{1}{\alpha} z(\bar{q}) \right).$$

It is well known that the projection  $P_{Q_{\text{ad}}}$  possesses the following property:

$$(2.10) \quad \|\nabla(P_{Q_{\text{ad}}}(v))(t)\|_{L^p(\Omega)} \leq \|\nabla v(t)\|_{L^p(\Omega)} \quad \forall v \in L^2(I, W^{1,p}(\Omega)) \text{ and for a.a. } t \in I.$$

Employing formulation (2.9) of the optimality condition, we obtain the following regularity result.

**PROPOSITION 2.3.** *Let  $(\bar{q}, \bar{u})$  be the solution of the optimization problem (2.2) and  $\bar{z} = z(\bar{q})$  be the corresponding adjoint state. Then there holds*

$$\bar{u}, \bar{z} \in L^2(I, H^2(\Omega)) \cap H^1(I, L^2(\Omega)),$$

$$\bar{q} \in L^2(I, W^{1,p}(\Omega)) \cap H^1(I, L^2(\Omega)) \cap L^\infty(I \times \Omega)$$

for any  $p < \infty$  when  $n = 2$  and  $p \leq 6$  when  $n = 3$ .

*Proof.* The regularity of  $\bar{u}, \bar{z}$  follows directly from Proposition 2.1. The embedding  $H^2(\Omega) \hookrightarrow W^{1,p}(\Omega)$  and property (2.10) imply the desired result for  $\bar{q}$ .  $\square$

**3. Discretization.** In this section we describe the space-time finite element discretization of optimal control problem (2.2).

**3.1. Semidiscretization in time.** At first, we present the semidiscretization in time of the state equation by discontinuous Galerkin methods following along the lines of the first part of this article [20]. We consider a partitioning of the time interval  $\bar{I} = [0, T]$  as

$$(3.1) \quad \bar{I} = \{0\} \cup I_1 \cup I_2 \cup \cdots \cup I_M$$

with subintervals  $I_m = (t_{m-1}, t_m]$  of size  $k_m$  and time points

$$0 = t_0 < t_1 < \cdots < t_{M-1} < t_M = T.$$

We define the discretization parameter  $k$  as a piecewise constant function by setting  $k|_{I_m} = k_m$  for  $m = 1, 2, \dots, M$ . Moreover, we denote by  $k$  the maximal size of the time steps, i.e.,  $k = \max k_m$ .

The semidiscrete trial and test space is given as

$$X_k^r = \left\{ v_k \in L^2(I, V) \mid v_k|_{I_m} \in \mathcal{P}_r(I_m, V), \ m = 1, 2, \dots, M \right\}.$$

Here,  $\mathcal{P}_r(I_m, V)$  denotes the space of polynomials up to order  $r$  defined on  $I_m$  with values in  $V$ . On  $X_k^r$  we use the notation

$$(v, w)_{I_m} := (v, w)_{L^2(I_m, L^2(\Omega))} \quad \text{and} \quad \|v\|_{I_m} := \|v\|_{L^2(I_m, L^2(\Omega))}.$$

To define the discontinuous Galerkin approximation (dG( $r$ )) using the space  $X_k^r$ , we employ the following definition for functions  $v_k \in X_k^r$ :

$$v_{k,m}^+ := \lim_{t \rightarrow 0^+} v_k(t_m + t), \quad v_{k,m}^- := \lim_{t \rightarrow 0^+} v_k(t_m - t) = v_k(t_m), \quad [v_k]_m := v_{k,m}^+ - v_{k,m}^-$$

and define the bilinear form  $B(\cdot, \cdot)$  for  $u_k, \varphi \in X_k^r$  by

$$(3.2) \quad B(u_k, \varphi) := \sum_{m=1}^M (\partial_t u_k, \varphi)_{I_m} + (\nabla u_k, \nabla \varphi)_I + \sum_{m=2}^M ([u_k]_{m-1}, \varphi_{m-1}^+) + (u_{k,0}^+, \varphi_0^+).$$

Then, the dG( $r$ ) semidiscretization of the state equation (2.1) for given control  $q \in Q$  reads as follows: Find a state  $u_k = u_k(q) \in X_k^r$  such that

$$(3.3) \quad B(u_k, \varphi) = (f + q, \varphi)_I + (u_0, \varphi_0^+) \quad \forall \varphi \in X_k^r.$$

The existence and uniqueness of solutions to (3.3) can be shown by using Fourier analysis; see [27] for details.

*Remark 3.1.* Using a density argument, it is possible to show that the exact solution  $u = u(q) \in X$  satisfies the identity

$$B(u, \varphi) = (f + q, \varphi)_I + (u_0, \varphi_0^+) \quad \forall \varphi \in X_k^r.$$

Thus, we have here the property of Galerkin orthogonality

$$B(u - u_k, \varphi) = 0 \quad \forall \varphi \in X_k^r,$$

although the dG( $r$ ) semidiscretization is a nonconforming Galerkin method ( $X_k^r \not\subset X$ ).

Throughout the paper we restrict ourselves to the case  $r = 0$ . The resulting dG(0) scheme is a variant of the implicit Euler method. In this case the semidiscrete state equation (3.3) can be explicitly rewritten as the following time-stepping scheme, using the fact that  $u_k$  is piecewise constant in time. We use the notation  $U_m = u_k|_{I_m} \in V$  and obtain

$$(U_1, \psi) + k_1(\nabla U_1, \nabla \psi) = (f + q, \psi)_{I_m} + (u_0, \psi) \quad \forall \psi \in V,$$

$$(U_m, \psi) + k_m(\nabla U_m, \nabla \psi) = (f + q, \psi)_{I_m} + (U_{m-1}, \psi) \quad \forall \psi \in V, \ m = 2, 3, \dots, M.$$

The semidiscrete optimization problem for the dG(0) time discretization has the following form:

$$(3.4) \quad \text{Minimize } J(q_k, u_k) \text{ subject to (3.3) and } (q_k, u_k) \in Q_{\text{ad}} \times X_k^0.$$

As in Proposition 3.2 in [20] the following result holds.

**PROPOSITION 3.2.** *For  $\alpha > 0$ , the semidiscrete optimal control problem (3.4) admits a unique solution  $(\bar{q}_k, \bar{u}_k) \in Q_{\text{ad}} \times X_k^0$ .*

Note that the optimal control  $\bar{q}_k$  is searched for in the subset  $Q_{\text{ad}}$  of the continuous space  $Q$ , and the subscript  $k$  indicates the usage of the semidiscretized state equation.

Similarly to the continuous case, we introduce the semidiscrete reduced cost functional  $j_k: Q \rightarrow \mathbb{R}$ :

$$j_k(q) := J(q, u_k(q))$$

and reformulate the semidiscrete optimal control problem (3.4) as follows:

$$\text{Minimize } j_k(q_k) \text{ subject to } q_k \in Q_{\text{ad}}.$$

The first order necessary optimality condition reads as

$$(3.5) \quad j'_k(\bar{q}_k)(\delta q - \bar{q}_k) \geq 0 \quad \forall \delta q \in Q_{\text{ad}},$$

and the derivative of  $j_k$  can be expressed as

$$(3.6) \quad j'_k(q)(\delta q) = (\alpha q + z_k(q), \delta q)_I.$$

Here,  $z_k = z_k(q) \in X_k^0$  denotes the solution of the semidiscrete adjoint equation

$$(3.7) \quad B(\varphi, z_k) = (\varphi, u_k(q) - \hat{u})_I \quad \forall \varphi \in X_k^0.$$

As on the continuous level, the second derivative  $j''_k(q)$  is independent of  $q$  and positive definite, i.e.,

$$(3.8) \quad j''_k(q)(p, p) \geq \alpha \|p\|_I^2 \quad \forall p \in Q.$$

Similarly to (2.9), the optimality condition (3.5) can be rewritten as

$$(3.9) \quad \bar{q}_k = P_{Q_{\text{ad}}} \left( -\frac{1}{\alpha} z_k(\bar{q}_k) \right).$$

This projection formula implies particularly that the optimal solution  $\bar{q}_k$  is piecewise constant in time. We will make use of this fact in section 5.

**3.2. Discretization in space.** To define the finite element discretization in space, we consider two or three dimensional shape-regular meshes; see, e.g., [9]. A mesh consists of quadrilateral or hexahedral cells  $K$ , which constitute a non-overlapping cover of the computational domain  $\Omega$ . The corresponding mesh is denoted by  $\mathcal{T}_h = \{K\}$ , where we define the discretization parameter  $h$  as a cellwise constant function by setting  $h|_K = h_K$  with the diameter  $h_K$  of the cell  $K$ . We use the symbol  $h$  also for the maximal cell size, i.e.,  $h = \max h_K$ .

On the mesh  $\mathcal{T}_h$  we construct a conform finite element space  $V_h \subset V$  in a standard way:

$$V_h^s = \{v \in V | v|_K \in \mathcal{Q}_s(K) \text{ for } K \in \mathcal{T}_h\}.$$



Here,  $\mathcal{Q}_s(K)$  consists of shape functions obtained via (bi-/tri-)linear transformations of polynomials in  $\widehat{\mathcal{Q}}_s(\widehat{K})$  defined on the reference cell  $\widehat{K} = (0, 1)^n$ ; cf. section 3.2 in [20].

To obtain the fully discretized versions of the time discretized state equation (3.3), we utilize the space-time finite element space

$$X_{k,h}^{r,s} = \left\{ v_{kh} \in L^2(I, V_h^s) \mid v_{kh}|_{I_m} \in \mathcal{P}_r(I_m, V_h^s) \right\} \subset X_k^r.$$

*Remark 3.3.* Here, the spatial mesh, and therefore also the space  $V_h^s$ , is fixed for all time intervals. We refer to [26] for a discussion of treatment of different meshes  $\mathcal{T}_h^m$  for each of the subintervals  $I_m$ .

The so-called cG( $s$ )dG( $r$ ) discretization of the state equation for given control  $q \in Q$  has the following form: Find a state  $u_{kh} = u_{kh}(q) \in X_{k,h}^{r,s}$  such that

$$(3.10) \quad B(u_{kh}, \varphi) = (f + q, \varphi)_I + (u_0, \varphi_0^+) \quad \forall \varphi \in X_{k,h}^{r,s}.$$

Throughout this paper we will restrict ourselves to the consideration of (bi-/tri-)linear elements, i.e., we set  $s = 1$  and consider the cG(1)dG(0) scheme.

Then, the corresponding optimal control problem is given as follows:

$$(3.11) \quad \text{Minimize } J(q_{kh}, u_{kh}) \text{ subject to (3.10) and } (q_{kh}, u_{kh}) \in Q_{\text{ad}} \times X_{k,h}^{0,1},$$

and by means of the discrete reduced cost functional  $j_{kh}: Q \rightarrow \mathbb{R}$ ,

$$j_{kh}(q) := J(q, u_{kh}(q)),$$

it can be reformulated as follows:

$$\text{Minimize } j_{kh}(q_{kh}) \text{ subject to } q_{kh} \in Q_{\text{ad}}.$$

The uniquely determined optimal solution of (3.11) is denoted by  $(\bar{q}_{kh}, \bar{u}_{kh}) \in Q_{\text{ad}} \times X_{k,h}^{0,1}$ .

The optimal control  $\bar{q}_{kh} \in Q_{\text{ad}}$  fulfills the first order optimality condition

$$(3.12) \quad j'_{kh}(\bar{q}_{kh})(\delta q - \bar{q}_{kh}) \geq 0 \quad \forall \delta q \in Q_{\text{ad}},$$

where  $j'_{kh}(q)(\delta q)$  is given by

$$(3.13) \quad j'_{kh}(q)(\delta q) = (\alpha q + z_{kh}(q), \delta q)_I$$

with the discrete adjoint solution  $z_{kh} = z_{kh}(q) \in X_{k,h}^{0,1}$  of

$$(3.14) \quad B(\varphi, z_{kh}) = (\varphi, u_{kh}(q) - \hat{u})_I \quad \forall \varphi \in X_{k,h}^{0,1}.$$

For the second derivative of  $j_{kh}$  we have, as before,

$$(3.15) \quad j''_{kh}(q)(p, p) \geq \alpha \|p\|_I^2 \quad \forall p \in Q.$$

**3.3. Discretization of the controls.** In this subsection, we describe four different approaches for the discretization of the control variable. Choosing a subspace  $Q_d \subset Q$ , we introduce the corresponding admissible set

$$Q_{d,\text{ad}} = Q_d \cap Q_{\text{ad}}.$$

Note that in what follows, the space  $Q_d$  will be either finite dimensional or the whole space  $Q$ . The optimal control problem on this level of discretization is given as follows:

$$(3.16) \quad \text{Minimize } J(q_\sigma, u_\sigma) \text{ subject to (3.10) and } (q_\sigma, u_\sigma) \in Q_{d,\text{ad}} \times X_{k,h}^{0,1}.$$

The unique optimal solution of (3.16) is denoted by  $(\bar{q}_\sigma, \bar{u}_\sigma) \in Q_{d,\text{ad}} \times X_{k,h}^{0,1}$ , where the subscript  $\sigma$  collects the discretization parameters  $k$ ,  $h$ , and  $d$ . The optimality condition is given using the discrete reduced cost functional  $j_{kh}$  introduced before:

$$(3.17) \quad j'_{kh}(\bar{q}_\sigma)(\delta q - \bar{q}_\sigma) \geq 0 \quad \forall \delta q \in Q_{d,\text{ad}}.$$

**3.3.1. Cellwise constant discretization.** The first possibility for the control discretization is to use cellwise constant functions. Employing the same time partitioning and the same spatial mesh as for the discretization of the state variable, we set

$$Q_d = \left\{ q \in Q \mid q|_{I_m \times K} \in \mathcal{P}_0(I_m \times K), \ m = 1, 2, \dots, M, \ K \in \mathcal{T}_h \right\}.$$

The discretization error for this type of discretization will be analyzed in section 5.1.

**3.3.2. Cellwise linear discretization.** Another possibility for the discretization of the control variable is to choose the same control discretization as for the state variable, i.e., piecewise constant in time and cellwise (bi-/tri-)linear in space. Using a spatial space

$$Q_h = \left\{ v \in C(\bar{\Omega}) \mid v|_K \in \mathcal{Q}_1(K) \text{ for } K \in \mathcal{T}_h \right\},$$

we set

$$Q_d = \left\{ q \in Q \mid q|_{I_m} \in \mathcal{P}_0(I_m, Q_h) \right\}.$$

The state space  $X_{k,h}^{0,1}$  coincides with the control space  $Q_d$  in the case of homogeneous Neumann boundary conditions and is a subspace of it, i.e.,  $Q_d \supset X_{k,h}^{0,1}$  in the presence of homogeneous Dirichlet boundary conditions.

The discretization error for this type of discretization will be analyzed in section 5.2.

**3.3.3. Variational approach.** Extending the discretization approach presented in [13], we can choose  $Q_d = Q$ . In this case the optimization problems (3.11) and (3.16) coincide, and therefore  $\bar{q}_\sigma = \bar{q}_{kh} \in Q_{\text{ad}}$ .

We use the fact that the optimality condition (3.12) can be rewritten employing the projection (2.8) as

$$\bar{q}_{kh} = P_{Q_{\text{ad}}} \left( -\frac{1}{\alpha} z_{kh}(\bar{q}_{kh}) \right),$$

and we obtain that  $\bar{q}_{kh}$  is a piecewise constant function in time. However,  $\bar{q}_{kh}$  is in general not a finite element function corresponding to the spatial mesh  $\mathcal{T}_h$ . This fact requires more care for the construction of algorithms for computation of  $\bar{q}_{kh}$ ; see [13] for details.

The discretization error for this type of discretization will be analyzed in section 5.3.

**3.3.4. Postprocessing strategy.** The strategy described in this section extends the approach from [21] to parabolic problems. For the discretization of the control space we employ the same choice as in section 3.3.1, i.e., cellwise constant discretization. After the computation of the corresponding solution  $\bar{q}_\sigma$ , a better approximation  $\tilde{q}_\sigma$  is constructed by a postprocessing, making use of the projection operator (2.8):

$$(3.18) \quad \tilde{q}_\sigma = P_{Q_{\text{ad}}} \left( -\frac{1}{\alpha} z_{kh}(\bar{q}_\sigma) \right).$$

Note that, similar to the solution obtained by the variational approach in section 3.3.3, the solution  $\tilde{q}_\sigma$  is piecewise constant in time and is generally not a finite element function in space with respect to the spatial mesh  $\mathcal{T}_h$ . This solution can be simply evaluated pointwise; however, the corresponding error analysis requires an additional assumption on the structure of active sets; see the discussion in section 5.4.

**4. Auxiliary results.** In this section we recall some results provided in the first part of this article [20], which will be used in what follows.

The first proposition provides a stability result for the purely time discretized state and adjoint solutions. It follows from Theorems 4.1 and 4.3 and Corollaries 4.2 and 4.5 of [20] as well as from elliptic regularity.

**PROPOSITION 4.1.** *For  $q \in Q$  let the solutions  $u_k(q) \in X_k^0$  and  $z_k(q) \in X_k^0$  be given by the semidiscrete state equation (3.3) and adjoint equation (3.7), respectively. Then it holds that*

$$\begin{aligned} \|\nabla^2 u_k(q)\|_I + \|\nabla u_k(q)\|_I + \|u_k(q)\|_I &\leq C\{\|f + q\|_I + \|\nabla u_0\| + \|u_0\|\}, \\ \|\nabla^2 z_k(q)\|_I + \|\nabla z_k(q)\|_I + \|z_k(q)\|_I &\leq C\|u_k(q) - \hat{u}\|_I. \end{aligned}$$

A similar result holds for the fully discretized solutions of the state and adjoint equations; cf. Theorem 4.6 and Corollary 4.7 in [20].

**PROPOSITION 4.2.** *For  $q \in Q$  let the solutions  $u_{kh}(q) \in X_{k,h}^{0,1}$  and  $z_{kh}(q) \in X_{k,h}^{0,1}$  be given by the discrete state equation (3.10) and adjoint equation (3.14), respectively. Then it holds that*

$$\begin{aligned} \|\nabla u_{kh}(q)\|_I + \|u_{kh}(q)\|_I &\leq C\{\|f + q\|_I + \|\nabla \Pi_h u_0\| + \|\Pi_h u_0\|\}, \\ \|\nabla z_{kh}(q)\|_I + \|z_{kh}(q)\|_I &\leq C\|u_{kh}(q) - \hat{u}\|_I, \end{aligned}$$

where  $\Pi_h: V \rightarrow V_h$  denotes the spatial  $L^2$ -projection.

In the following two propositions, we recall a priori estimates for the errors due to temporal and spatial discretizations of the state and adjoint variables. The assertions are proved in [20] by means of Theorems 5.1 and 5.5 as well as by Lemma 6.2 presented therein.

**PROPOSITION 4.3.** *For  $q \in Q$  let the solutions  $u(q) \in X$  and  $z(q) \in X$  be given by the state equation (2.1) and adjoint equation (2.5), respectively. Moreover, let  $u_k(q) \in X_k^0$  and  $z_k(q) \in X_k^0$  be determined as solutions of the semidiscrete state equation (3.3) and adjoint equation (3.7). Then the following error estimates hold:*

$$\begin{aligned} \|u(q) - u_k(q)\|_I &\leq Ck\|\partial_t u(q)\|_I, \\ \|z(q) - z_k(q)\|_I &\leq Ck\{\|\partial_t u(q)\|_I + \|\partial_t z(q)\|_I\}. \end{aligned}$$

**PROPOSITION 4.4.** *For  $q \in Q$  let the solutions  $u_k(q) \in X_k^0$  and  $z_k(q) \in X_k^0$  be given by the semidiscrete state equation (3.3) and adjoint equation (3.7), respectively.*

Moreover, let  $u_{kh}(q) \in X_{k,h}^{0,1}$  and  $z_{kh}(q) \in X_{k,h}^{0,1}$  be determined as solutions of the discrete state equation (3.10) and adjoint equation (3.14). Then the following error estimates hold:

$$\|u_k(q) - u_{kh}(q)\|_I \leq Ch^2 \|\nabla^2 u_k(q)\|_I,$$

$$\|z_k(q) - z_{kh}(q)\|_I \leq Ch^2 \{ \|\nabla^2 u_k(q)\|_I + \|\nabla^2 z_k(q)\|_I \}.$$

Proposition 4.2 provides a stability result for the discrete adjoint solution with respect to the norm of  $L^2(I, H^1(\Omega))$ . For later use we additionally prove a corresponding result with respect to the norm of  $L^2(I, L^\infty(\Omega))$ .

LEMMA 4.5. For  $q \in Q$  let the solutions  $u_{kh}(q) \in X_{k,h}^{0,1}$  and  $z_{kh}(q) \in X_{k,h}^{0,1}$  be given by the discrete state equation (3.10) and adjoint equation (3.14), respectively. Then it holds that

$$\|z_{kh}(q)\|_{L^2(I, L^\infty(\Omega))} \leq C \|u_{kh}(q) - \hat{u}\|_I.$$

*Proof.* We define an additional adjoint solution  $\tilde{z}_k \in X_k^0$  as solution of

$$B(\varphi, \tilde{z}_k) = (\varphi, u_{kh}(q) - \hat{u})_I \quad \forall \varphi \in X_k^0.$$

Since  $\tilde{z}_k$  and  $z_{kh}(q)$  are given by means of the same right-hand side  $u_{kh}(q) - \hat{u}$ , it is possible to apply standard a priori error estimates to the discretization error  $z_{kh}(q) - \tilde{z}_k$  similar to Proposition 4.4.

By inserting the solution  $\tilde{z}_k$  and utilizing the embedding  $L^2(I, H^2(\Omega)) \hookrightarrow L^2(I, L^\infty(\Omega))$ , we get

$$\begin{aligned} \|z_{kh}(q)\|_{L^2(I, L^\infty(\Omega))} &\leq \|z_{kh}(q) - \tilde{z}_k\|_{L^2(I, L^\infty(\Omega))} + \|\tilde{z}_k\|_{L^2(I, L^\infty(\Omega))} \\ &\leq \|z_{kh}(q) - \tilde{z}_k\|_{L^2(I, L^\infty(\Omega))} + C \|\tilde{z}_k\|_{L^2(I, H^2(\Omega))}. \end{aligned}$$

For the first term we obtain, by inserting a spatial interpolation  $i_h \tilde{z}_k \in X_{k,h}^{0,1}$ ,

$$(4.1) \quad \|z_{kh}(q) - \tilde{z}_k\|_{L^2(I, L^\infty(\Omega))} \leq \|z_{kh}(q) - i_h \tilde{z}_k\|_{L^2(I, L^\infty(\Omega))} + \|i_h \tilde{z}_k - \tilde{z}_k\|_{L^2(I, L^\infty(\Omega))}.$$

For the first term on the right-hand side of (4.1) we proceed by means of an inverse estimate between  $L^\infty(\Omega)$  and  $L^2(\Omega)$  for discrete functions, an estimate for the error due to space discretization (cf. Theorem 5.1 of [20]), and an estimate for the spatial interpolation error as

$$\begin{aligned} \|z_{kh}(q) - i_h \tilde{z}_k\|_{L^2(I, L^\infty(\Omega))}^2 &= \sum_{m=1}^M k_m \|z_{kh}(q)(t_m) - i_h \tilde{z}_k(t_m)\|_{L^\infty(\Omega)}^2 \\ &\leq Ch^{-n} \sum_{m=1}^M k_m \|z_{kh}(q)(t_m) - i_h \tilde{z}_k(t_m)\|^2 \\ &\leq Ch^{-n} \{ \|z_{kh}(q) - \tilde{z}_k\|_I^2 + \|\tilde{z}_k - i_h \tilde{z}_k\|_I^2 \} \\ &\leq Ch^{4-n} \|\nabla^2 \tilde{z}_k\|_I^2. \end{aligned}$$

By standard interpolation estimates, we have for the second term on the right-hand

side of (4.1),

$$\begin{aligned}
 \|i_h \tilde{z}_k - \tilde{z}_k\|_{L^2(I, L^\infty(\Omega))}^2 &= \sum_{m=1}^M k_m \|i_h \tilde{z}_k(t_m) - \tilde{z}_k(t_m)\|_{L^\infty(\Omega)}^2 \\
 &\leq Ch^{4-n} \sum_{m=1}^M k_m \|\nabla^2 \tilde{z}_k(t_m)\|^2 \\
 &= Ch^{4-n} \|\nabla^2 \tilde{z}_k\|_I^2.
 \end{aligned}$$

We complete the proof by collecting all estimates and application of the stability result from Proposition 4.1:

$$\|z_{kh}(q)\|_{L^2(I, L^\infty(\Omega))} \leq Ch^{4-n} \|\nabla^2 \tilde{z}_k\|_I + C \|\tilde{z}_k\|_{L^2(I, H^2(\Omega))} \leq C \|u_{kh}(q) - \hat{u}\|_I. \quad \square$$

**5. Error estimates.** In this section we provide a priori error estimates for the different discretization approaches described in section 3. We start with an assertion of the error between the solution  $\bar{q}$  of the continuous problem (2.2) and the solution  $\bar{q}_k$  of the semidiscretized problem (3.4).

**THEOREM 5.1.** *Let  $\bar{q} \in Q_{ad}$  be the solution of optimization problem (2.2) and  $\bar{q}_k$  be the solution of the semidiscretized problem (3.4). Then the following estimate holds:*

$$\|\bar{q} - \bar{q}_k\|_I \leq \frac{1}{\alpha} \|z(\bar{q}) - z_k(\bar{q})\|_I.$$

*Proof.* Using the optimality conditions (2.4) and (3.5), we obtain the relation

$$-j'_k(\bar{q}_k)(\bar{q} - \bar{q}_k) \leq 0 \leq -j'(\bar{q})(\bar{q} - \bar{q}_k).$$

From (3.8) we have with any  $p \in Q$ :

$$\begin{aligned}
 \alpha \|\bar{q} - \bar{q}_k\|_I^2 &\leq j''_k(p)(\bar{q} - \bar{q}_k, \bar{q} - \bar{q}_k) \\
 &= j'_k(\bar{q})(\bar{q} - \bar{q}_k) - j'_k(\bar{q}_k)(\bar{q} - \bar{q}_k) \\
 &\leq j'_k(\bar{q})(\bar{q} - \bar{q}_k) - j'(\bar{q})(\bar{q} - \bar{q}_k).
 \end{aligned}$$

By means of the representations (2.6) and (3.6) of  $j'$  and  $j'_k$ , respectively, we obtain

$$\alpha \|\bar{q} - \bar{q}_k\|_I^2 \leq (z(\bar{q}) - z_k(\bar{q}), \bar{q} - \bar{q}_k)_I.$$

The desired assertion follows by Cauchy's inequality.  $\square$

**5.1. Cellwise constant discretization.** In this section we are going to prove an estimate for the error  $\|\bar{q} - \bar{q}_\sigma\|_I$  when the control is discretized by cellwise constant polynomials in space and time; see section 3.3.1.

For doing so, we will extend the techniques presented in [8] to the case of parabolic optimal control problems. This demands the introduction of the solution  $\bar{q}_d$  of the following purely control discretized problem:

$$(5.1) \quad \text{Minimize } j(q_d) \text{ subject to } q_d \in Q_{d,ad}.$$

The uniquely determined solution  $\bar{q}_d$  fulfills the optimality condition

$$(5.2) \quad j'(\bar{q}_d)(\delta q - \bar{q}_d) \geq 0 \quad \forall \delta q \in Q_{d,\text{ad}}.$$

To formulate the main result of this section, we introduce the  $L^2$ -projection  $\pi_d: Q \rightarrow Q_d$  and note that, due to the cellwise constant discretization, the following property holds true:

$$\pi_d Q_{\text{ad}} \subset Q_{d,\text{ad}}.$$

**THEOREM 5.2.** *Let  $\bar{q} \in Q_{\text{ad}}$  be the solution of the optimal control problem (2.2), and let  $\bar{q}_\sigma \in Q_{d,\text{ad}}$  be the solution of the discretized problem (3.16), where the cellwise constant discretization for the control variable is employed. Moreover, let  $\bar{q}_d \in Q_{d,\text{ad}}$  be the solution of the purely control discretized problem (5.1). Then the following estimate holds:*

$$\|\bar{q} - \bar{q}_\sigma\|_I \leq \|\bar{q} - \pi_d \bar{q}\|_I + \frac{1}{\alpha} \|z(\bar{q}_d) - \pi_d z(\bar{q}_d)\|_I + \frac{1}{\alpha} \|z(\bar{q}_d) - z_{kh}(\bar{q}_d)\|_I.$$

*Proof.* We split the error

$$\|\bar{q} - \bar{q}_\sigma\|_I \leq \|\bar{q} - \bar{q}_d\|_I + \|\bar{q}_d - \bar{q}_\sigma\|_I$$

and estimate both terms on the right-hand side separately. For treating the first term, we use the fact that  $\pi_d \bar{q} \in Q_{d,\text{ad}}$  and obtain from the optimality conditions (2.4) and (5.2) the inequalities

$$j'(\bar{q})(\bar{q} - \bar{q}_d) \leq 0 \quad \text{and} \quad -j'(\bar{q}_d)(\pi_d \bar{q} - \bar{q}_d) \leq 0.$$

Using (2.7) we proceed with any  $p \in Q$ :

$$\begin{aligned} \alpha \|\bar{q} - \bar{q}_d\|_I^2 &\leq j''(p)(\bar{q} - \bar{q}_d, \bar{q} - \bar{q}_d) \\ &= j'(\bar{q})(\bar{q} - \bar{q}_d) - j'(\bar{q}_d)(\bar{q} - \bar{q}_d) \\ &= j'(\bar{q})(\bar{q} - \bar{q}_d) - j'(\bar{q}_d)(\bar{q} - \pi_d \bar{q}) - j'(\bar{q}_d)(\pi_d \bar{q} - \bar{q}_d) \\ &\leq -j'(\bar{q}_d)(\bar{q} - \pi_d \bar{q}). \end{aligned}$$

By means of the representation of the derivative  $j'$  from (2.6) and the properties of  $\pi_d$ , we have

$$\begin{aligned} \alpha \|\bar{q} - \bar{q}_d\|_I^2 &\leq -j'(\bar{q}_d)(\bar{q} - \pi_d \bar{q}) \\ &= -(\alpha \bar{q}_d + z(\bar{q}_d), \bar{q} - \pi_d \bar{q})_I \\ &= (\pi_d z(\bar{q}_d) - z(\bar{q}_d), \bar{q} - \pi_d \bar{q})_I, \end{aligned}$$

and by Young's inequality we obtain the intermediary result

$$(5.3) \quad \|\bar{q} - \bar{q}_d\|_I^2 \leq \|\bar{q} - \pi_d \bar{q}\|_I^2 + \frac{1}{4\alpha^2} \|z(\bar{q}_d) - \pi_d z(\bar{q}_d)\|_I^2.$$

In order to estimate the term  $\|\bar{q}_d - \bar{q}_\sigma\|_I$  we exploit the optimality conditions (5.2) and (3.17) leading to the following relation:

$$-j'_{kh}(\bar{q}_\sigma)(\bar{q}_d - \bar{q}_\sigma) \leq 0 \leq -j'(\bar{q}_d)(\bar{q}_d - \bar{q}_\sigma).$$

Using (3.15) and the representations (2.6) for  $j'$  and (3.13) for  $j'_{kh}$ , respectively, we obtain

$$\begin{aligned} \alpha \|\bar{q}_d - \bar{q}_\sigma\|_I^2 &\leq j''_{kh}(p)(\bar{q}_d - \bar{q}_\sigma, \bar{q}_d - \bar{q}_\sigma) \\ &= j'_{kh}(\bar{q}_d)(\bar{q}_d - \bar{q}_\sigma) - j'_{kh}(\bar{q}_\sigma)(\bar{q}_d - \bar{q}_\sigma) \\ &\leq j'_{kh}(\bar{q}_d)(\bar{q}_d - \bar{q}_\sigma) - j'(\bar{q}_d)(\bar{q}_d - \bar{q}_\sigma) \\ &\leq \|z(\bar{q}_d) - z_{kh}(\bar{q}_d)\|_I \|\bar{q}_d - \bar{q}_\sigma\|_I. \end{aligned}$$

Thus, we achieve

$$(5.4) \quad \|\bar{q}_d - \bar{q}_\sigma\|_I \leq \frac{1}{\alpha} \|z(\bar{q}_d) - z_{kh}(\bar{q}_d)\|_I.$$

Collecting estimates (5.3) and (5.4), we complete the proof.  $\square$

This theorem directly implies the following result.

**COROLLARY 5.3.** *Under the conditions of Theorem 5.2, the following estimate holds:*

$$\begin{aligned} \|\bar{q} - \bar{q}_\sigma\|_I &\leq \frac{C}{\alpha} k \{ \|\partial_t \bar{q}\|_I + \|\partial_t u(\bar{q}_d)\|_I + \|\partial_t z(\bar{q}_d)\|_I \} \\ &+ \frac{C}{\alpha} h \{ \|\nabla \bar{q}\|_I + \|\nabla z(\bar{q}_d)\|_I + h (\|\nabla^2 u_k(\bar{q}_d)\|_I + \|\nabla^2 z_k(\bar{q}_d)\|_I) \} = \mathcal{O}(k + h). \end{aligned}$$

*Proof.* The assertion follows from Theorem 5.2 by interpolation estimates and Propositions 4.3 and 4.4. Due to the fact that  $\bar{q}, \bar{q}_d \in Q_{\text{ad}}$ , we obtain, using the stability estimates from Proposition 4.1, that all norms involved in this estimate are bounded by a constant independent of all discretization parameters.  $\square$

**5.2. Cellwise linear discretization.** This section is devoted to the error analysis for the discretization of the control variable by piecewise constants in time and cellwise (bi-/tri-)linear functions in space as described in section 3.3.2. To this end we split the error

$$\|\bar{q} - \bar{q}_\sigma\|_I \leq \|\bar{q} - \bar{q}_k\|_I + \|\bar{q}_k - \bar{q}_\sigma\|_I$$

and use the result of Theorem 5.1 for the first part. For treating the error  $\|\bar{q}_k - \bar{q}_\sigma\|_I$  we adapt the technique described in [4] and [6] to parabolic problems.

The analysis in this section is based on an assumption on the structure of the active sets. For each time interval  $I_m$  we group the cells  $K$  of the mesh  $\mathcal{T}_h$  depending on the value of  $\bar{q}_k$  on  $K$  into three sets  $\mathcal{T}_h = \mathcal{T}_{h,m}^1 \cup \mathcal{T}_{h,m}^2 \cup \mathcal{T}_{h,m}^3$  with  $\mathcal{T}_{h,m}^i \cap \mathcal{T}_{h,m}^j = \emptyset$  for  $i \neq j$ . The sets are chosen as follows:

$$\begin{aligned} \mathcal{T}_{h,m}^1 &:= \{K \in \mathcal{T}_h \mid \bar{q}_k(t_m, x) = q_a \text{ or } \bar{q}_k(t_m, x) = q_b \quad \forall x \in K\}, \\ \mathcal{T}_{h,m}^2 &:= \{K \in \mathcal{T}_h \mid q_a < \bar{q}_k(t_m, x) < q_b \quad \forall x \in K\}, \\ \mathcal{T}_{h,m}^3 &:= \mathcal{T}_h \setminus (\mathcal{T}_{h,m}^1 \cup \mathcal{T}_{h,m}^2). \end{aligned}$$

Hence, the set  $\mathcal{T}_{h,m}^3$  consists of the cells which contain the free boundary between the active and the inactive sets for the time interval  $I_m$ .

*Assumption 1.* We assume that there exists a positive constant  $C$  independent of  $k$ ,  $h$ , and  $m$  such that

$$\sum_{K \in \mathcal{T}_{h,m}^3} |K| \leq Ch$$

separately for all  $m = 1, 2, \dots, M$ .

*Remark 5.4.* A similar assumption is used in [21, 25, 2]. This assumption is valid if the boundary of the level sets

$$\{x \in \Omega | \bar{q}_k(t_m, x) = q_a\} \quad \text{and} \quad \{x \in \Omega | \bar{q}_k(t_m, x) = q_b\}$$

consists of a finite number of rectifiable curves.

We consider the usual nodal interpolation operator  $I_d$  which maps into the space of cellwise (bi-/tri-)linear functions  $Q_h$ . It is defined for functions  $g \in C(\Omega)$  by pointwise setting

$$(5.5) \quad I_d g(x_i) = g(x_i) \quad \text{for each node } x_i \text{ of } \mathcal{T}_h.$$

The operator  $I_d$  will also be applied to time-dependent functions  $g$  by the setting  $(I_d g)(t) = I_d g(t)$ .

In the following theorem we provide an assertion on the error  $\|\bar{q}_k - \bar{q}_\sigma\|_I$ .

**THEOREM 5.5.** *Let  $\bar{q}_k \in Q_{ad}$  be the solution of the semidiscretized optimal control problem (3.4) and  $\bar{q}_\sigma \in Q_{d,ad}$  be the solution of the discrete problem (3.16), where the cellwise (bi-/tri-)linear discretization for the control variable is employed. Then the following estimate holds:*

$$\begin{aligned} \|\bar{q}_k - \bar{q}_\sigma\|_I &\leq C \left(1 + \frac{1}{\alpha}\right) \|I_d \bar{q}_k - \bar{q}_k\|_I \\ &\quad + \frac{C}{\alpha} \|z_k(\bar{q}_k) - z_{kh}(\bar{q}_k)\|_I + \frac{C}{\sqrt{\alpha}} (j'_k(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_k))^{\frac{1}{2}}. \end{aligned}$$

*Proof.* We split

$$(5.6) \quad \|\bar{q}_k - \bar{q}_\sigma\|_I \leq \|\bar{q}_k - I_d \bar{q}_k\|_I + \|I_d \bar{q}_k - \bar{q}_\sigma\|_I$$

and estimate the term  $\|I_d \bar{q}_k - \bar{q}_\sigma\|_I$ . Due to the optimality conditions (3.17) and (3.5), and since  $I_d \bar{q}_k \in Q_{d,ad}$ , we have

$$-j'_{kh}(\bar{q}_\sigma)(I_d \bar{q}_k - \bar{q}_\sigma) \leq 0 \leq -j'_k(\bar{q}_k)(\bar{q}_k - \bar{q}_\sigma),$$

and due to (3.15) we obtain for any  $p \in Q$ ,

$$\begin{aligned} \alpha \|I_d \bar{q}_k - \bar{q}_\sigma\|_I^2 &\leq j''_{kh}(p)(I_d \bar{q}_k - \bar{q}_\sigma, I_d \bar{q}_k - \bar{q}_\sigma) \\ &\leq j'_{kh}(I_d \bar{q}_k)(I_d \bar{q}_k - \bar{q}_\sigma) - j'_{kh}(\bar{q}_\sigma)(I_d \bar{q}_k - \bar{q}_\sigma) \\ &\leq j'_{kh}(I_d \bar{q}_k)(I_d \bar{q}_k - \bar{q}_\sigma) - j'_k(\bar{q}_k)(\bar{q}_k - \bar{q}_\sigma) \\ (5.7) \quad &= j'_{kh}(I_d \bar{q}_k)(I_d \bar{q}_k - \bar{q}_\sigma) - j'_{kh}(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_\sigma) \\ &\quad + j'_{kh}(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_\sigma) - j'_k(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_\sigma) \\ &\quad + j'_k(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_k). \end{aligned}$$



The representations (3.6) of  $j'_k$  and (3.13) of  $j'_{kh}$  yield, by means of Proposition 4.2, that for any  $p, q, r \in Q$ ,

$$|j'_{kh}(p)(r) - j'_{kh}(q)(r)| \leq \{\alpha \|p - q\|_I + \|z_{kh}(p) - z_{kh}(q)\|_I\} \|r\|_I \leq (C + \alpha) \|p - q\|_I \|r\|_I$$

and

$$|j'_{kh}(q)(r) - j'_k(q)(r)| \leq \|z_k(q) - z_{kh}(q)\|_I \|r\|_I.$$

Applying these inequalities to the right-hand side of (5.7) leads to

$$\begin{aligned} \alpha \|I_d \bar{q}_k - \bar{q}_\sigma\|_I^2 &\leq (C + \alpha) \|I_d \bar{q}_k - \bar{q}_k\|_I \|I_d \bar{q}_k - \bar{q}_\sigma\|_I \\ &\quad + \|z_k(\bar{q}_k) - z_{kh}(\bar{q}_k)\|_I \|I_d \bar{q}_k - \bar{q}_\sigma\|_I + j'_k(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_k). \end{aligned}$$

With Young's inequality, we obtain

$$\begin{aligned} \|I_d \bar{q}_k - \bar{q}_\sigma\|_I^2 &\leq C \left(1 + \frac{1}{\alpha^2}\right) \|I_d \bar{q}_k - \bar{q}_k\|_I^2 \\ &\quad + \frac{C}{\alpha^2} \|z_k(\bar{q}_k) - z_{kh}(\bar{q}_k)\|_I^2 + \frac{C}{\alpha} j'_k(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_k). \end{aligned}$$

Inserting this estimate into (5.6) completes the proof.  $\square$

In the following two lemmas we provide estimates for the terms  $j'_k(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_k)$  and  $\|I_d \bar{q}_k - \bar{q}_k\|_I$  appearing on the right-hand side of the assertion of Theorem 5.5.

**LEMMA 5.6.** *Let  $\bar{q}_k \in Q_{ad}$  be the solution of the semidiscretized optimization problem (3.4) and  $I_d \bar{q}_k$  be the interpolation constructed by (5.5). Then, if Assumption 1 is fulfilled, the following estimate holds for  $n < p \leq \infty$ , provided  $z_k(\bar{q}_k) \in L^2(I, W^{1,p}(\Omega))$ :*

$$|j'_k(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_k)| \leq \frac{C}{\alpha} h^{3-\frac{2}{p}} \|\nabla z_k(\bar{q}_k)\|_{L^2(I, L^p(\Omega))}^2.$$

*Proof.* Using representation (3.6) of  $j'_k$  we have

$$\begin{aligned} j'_k(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_k) &= (\alpha \bar{q}_k + z_k(\bar{q}_k), I_d \bar{q}_k - \bar{q}_k)_I \\ (5.8) \quad &= \sum_{m=1}^M \int_{I_m} (\alpha \bar{q}_k(t) + z_k(\bar{q}_k)(t), I_d \bar{q}_k(t) - \bar{q}_k(t)) dt \\ &= \sum_{m=1}^M k_m (\alpha \bar{q}_k(t_m) + z_k(\bar{q}_k)(t_m), I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)). \end{aligned}$$

With the abbreviation  $d_m := \alpha \bar{q}_k(t_m) + z_k(\bar{q}_k)(t_m)$  we obtain

$$\begin{aligned} (5.9) \quad (d_m, I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)) &= \sum_{K \in \mathcal{T}_h} (d_m, I_d \bar{q}_k(t_m) - \bar{q}_k(t_m))_{L^2(K)} \\ &= \sum_{K \in \mathcal{T}_{h,m}^3} (d_m, I_d \bar{q}_k(t_m) - \bar{q}_k(t_m))_{L^2(K)}, \end{aligned}$$

since it holds  $I_d \bar{q}_k(t_m) = \bar{q}_k(t_m)$  on  $\mathcal{T}_{h,m}^1$  by construction and  $d_m = 0$  on  $\mathcal{T}_{h,m}^2$  due to representation formula (3.9).

In every cell  $K \in \mathcal{T}_{h,m}^3$  there is a point  $x_K$  with  $d_m(x_K) = 0$ . Thus, we get

$$\begin{aligned} |(d_m, I_d \bar{q}_k(t_m) - \bar{q}_k(t_m))_{L^2(K)}| & \leq |K|^{1-\frac{2}{p}} \|d_m\|_{L^p(K)} \|I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)\|_{L^p(K)} \\ & = |K|^{1-\frac{2}{p}} \|d_m - d_m(x_K)\|_{L^p(K)} \|I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)\|_{L^p(K)} \\ & \leq Ch^2 |K|^{1-\frac{2}{p}} \|\nabla d_m\|_{L^p(K)} \|\nabla \bar{q}_k(t_m)\|_{L^p(K)}. \end{aligned}$$

Inserting this estimate into (5.9) yields, together with Assumption 1,

$$\begin{aligned} |(d_m, I_d \bar{q}_k(t_m) - \bar{q}_k(t_m))| & \leq Ch^2 \sum_{K \in \mathcal{T}_{h,m}^3} |K|^{1-\frac{2}{p}} \|\nabla d_m\|_{L^p(K)} \|\nabla \bar{q}_k(t_m)\|_{L^p(K)} \\ & \leq Ch^2 \left( \sum_{K \in \mathcal{T}_{h,m}^3} |K| \right)^{1-\frac{2}{p}} \|\nabla d_m\|_{L^p(\Omega)} \|\nabla \bar{q}_k(t_m)\|_{L^p(\Omega)} \\ & \leq Ch^{3-\frac{2}{p}} \|\nabla d_m\|_{L^p(\Omega)} \|\nabla \bar{q}_k(t_m)\|_{L^p(\Omega)}. \end{aligned}$$

Then, the estimate

$$\|\nabla d_m\|_{L^p(\Omega)} \leq \alpha \|\nabla q_k(\bar{q}_k)(t_m)\|_{L^p(\Omega)} + \|\nabla z_k(\bar{q}_k)(t_m)\|_{L^p(\Omega)},$$

representation formula (3.9), and property (2.10) imply

$$|(d_m, I_d \bar{q}_k(t_m) - \bar{q}_k(t_m))| \leq \frac{C}{\alpha} h^{3-\frac{2}{p}} \|\nabla z_k(\bar{q}_k)(t_m)\|_{L^p(\Omega)}^2.$$

Hence, by inserting this last estimate into (5.8) we obtain the proposed assertion

$$\begin{aligned} |j'(\bar{q}_k)(I_d \bar{q}_k - \bar{q}_k)| & \leq \frac{C}{\alpha} h^{3-\frac{2}{p}} \sum_{m=1}^M k_m \|\nabla z_k(\bar{q}_k)(t_m)\|_{L^p(\Omega)}^2 \\ & = \frac{C}{\alpha} h^{3-\frac{2}{p}} \|\nabla z_k(\bar{q}_k)\|_{L^2(I, L^p(\Omega))}^2. \quad \square \end{aligned}$$

**LEMMA 5.7.** *Let  $\bar{q}_k \in Q_{ad}$  be the solution of the semidiscretized optimization problem (3.4) and  $I_d \bar{q}_k$  be the interpolation constructed by (5.5). Then, if Assumption 1 is fulfilled, the following estimate holds for  $n < p \leq \infty$ , provided  $z_k(\bar{q}_k) \in L^2(I, W^{1,p}(\Omega))$ :*

$$\|I_d \bar{q}_k - \bar{q}_k\|_I \leq \frac{C}{\alpha} \{h^2 \|\nabla^2 z_k(\bar{q}_k)\|_I + h^{\frac{3}{2}-\frac{1}{p}} \|\nabla z_k(\bar{q}_k)\|_{L^2(I, L^p(\Omega))}\}.$$

*Proof.* Since  $\bar{q}_k$  is piecewise constant in time we write

$$(5.10) \quad \|I_d \bar{q}_k - \bar{q}_k\|_I^2 = \sum_{m=1}^M \int_{I_m} \|I_d \bar{q}_k(t) - \bar{q}_k(t)\|^2 dt = \sum_{m=1}^M k_m \|I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)\|^2.$$

For each  $m = 1, 2, \dots, M$ , we split

$$\begin{aligned}
 \|I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)\|^2 &= \sum_{K \in \mathcal{T}_h} \|I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)\|_{L^2(K)}^2 \\
 (5.11) \qquad &= \sum_{K \in \mathcal{T}_{h,m}^2} \|I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)\|_{L^2(K)}^2 \\
 &\quad + \sum_{K \in \mathcal{T}_{h,m}^3} \|I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)\|_{L^2(K)}^2.
 \end{aligned}$$

Here, the sum over  $K \in \mathcal{T}_{h,m}^1$  vanishes since on  $\mathcal{T}_{h,m}^1$  it holds that  $I_d \bar{q}_k = \bar{q}_k$ .

The first term on the right-hand side of (5.11) can be estimated as

$$\begin{aligned}
 \sum_{K \in \mathcal{T}_{h,m}^2} \|I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)\|_{L^2(K)}^2 &\leq Ch^4 \sum_{K \in \mathcal{T}_{h,m}^2} \|\nabla^2 \bar{q}_k(t_m)\|_{L^2(K)}^2 \\
 &\leq \frac{C}{\alpha^2} h^4 \|\nabla^2 z_k(\bar{q}_k)(t_m)\|^2,
 \end{aligned}$$

since  $\bar{q}_k(t_m) = -\frac{1}{\alpha} z_k(\bar{q}_k)(t_m)$  on all cells  $K \in \mathcal{T}_{h,m}^2$ . For the second term on the right-hand side of (5.11) we proceed by means of representation formula (3.9), property (2.10), and Assumption 1:

$$\begin{aligned}
 \sum_{K \in \mathcal{T}_{h,m}^3} \|I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)\|_{L^2(K)}^2 &\leq \sum_{K \in \mathcal{T}_{h,m}^3} |K|^{1-\frac{2}{p}} \|I_d \bar{q}_k(t_m) - \bar{q}_k(t_m)\|_{L^p(K)}^2 \\
 &\leq Ch^2 \sum_{K \in \mathcal{T}_{h,m}^3} |K|^{1-\frac{2}{p}} \|\nabla \bar{q}_k(t_m)\|_{L^p(K)}^2 \\
 &\leq Ch^2 \left( \sum_{K \in \mathcal{T}_{h,m}^3} |K| \right)^{1-\frac{2}{p}} \|\nabla \bar{q}_k(t_m)\|_{L^p(\Omega)}^2 \\
 &\leq \frac{C}{\alpha^2} h^{3-\frac{2}{p}} \|\nabla z_k(\bar{q}_k)(t_m)\|_{L^p(\Omega)}^2.
 \end{aligned}$$

Inserting the last two estimates into (5.11) and plugging (5.11) into (5.10) implies the stated result.  $\square$

**COROLLARY 5.8.** *Under the conditions of Theorem 5.5 and Lemmas 5.6 and 5.7, the following estimate holds:*

$$\begin{aligned}
 \|\bar{q} - \bar{q}_\sigma\|_I &\leq \frac{C}{\alpha} k \{ \|\partial_t u(\bar{q})\|_I + \|\partial_t z(\bar{q})\|_I \} + \frac{C}{\alpha} \left(1 + \frac{1}{\alpha}\right) \{ h^2 \|\nabla^2 u_k(\bar{q}_k)\|_I \\
 &\quad + h^2 \|\nabla^2 z_k(\bar{q}_k)\|_I + h^{\frac{3}{2}-\frac{1}{p}} \|\nabla z_k(\bar{q}_k)\|_{L^2(I, L^p(\Omega))} \} = \mathcal{O}(k + h^{\frac{3}{2}-\frac{1}{p}}).
 \end{aligned}$$

*Proof.* The result follows directly from Theorems 5.1 and 5.5, Lemmas 5.6 and 5.7, and Proposition 4.4.  $\square$

In what follows we discuss the result from Corollary 5.8 in more details. This result holds under the assumption that  $z_k(\bar{q}_k) \in L^2(I, W^{1,p}(\Omega))$ . From the stability result in Proposition 4.1 and the fact that  $\bar{q}_k \in Q_{\text{ad}}$ , we know that

$$\|z_k(\bar{q}_k)\|_{L^2(I, H^2(\Omega))} \leq C.$$

By a Sobolev embedding theorem we have  $H^2(\Omega) \hookrightarrow W^{1,p}(\Omega)$  for all  $p < \infty$  in two space dimensions and for  $p \leq 6$  in three dimensions. This implies the order of convergence  $\mathcal{O}(k + h^{\frac{3}{2} - \frac{1}{p}})$  for all  $2 < p < \infty$  in two dimensions and  $\mathcal{O}(k + h^{\frac{4}{3}})$  in three dimensions, respectively. If in addition  $\|z_k(\bar{q}_k)\|_{L^2(I, W^{1,\infty}(\Omega))}$  is bounded, then we have in both cases the order of convergence  $\mathcal{O}(k + h^{\frac{3}{2}})$ .

*Remark 5.9.* The above result relies on Assumption 1. This assumption is valid in the majority of practical cases; cf. Remark 5.4. In the absence of this assumption a weaker result for the behavior of the spatial error can be shown, i.e.,

$$\lim_{h \rightarrow 0} \frac{1}{h} \|\bar{q}_k - \bar{q}_\sigma\|_I = 0.$$

The proofs in this section can simply be adapted to this situation. For the corresponding result for elliptic optimal control problems, we refer to [4].

**5.3. Variational approach.** In this subsection we prove an estimate for the error  $\|\bar{q} - \bar{q}_\sigma\|_I$  in the case of no control discretization; see section 3.3.3. In this case we choose  $Q_d = Q$ , and thus  $Q_{d,\text{ad}} = Q_{\text{ad}}$ . This implies  $\bar{q}_\sigma = \bar{q}_{kh}$ .

**THEOREM 5.10.** *Let  $\bar{q} \in Q_{\text{ad}}$  be the solution of optimization problem (2.2) and  $\bar{q}_{kh} \in Q_{\text{ad}}$  be the solution of the discretized problem (3.11). Then the following estimate holds:*

$$\|\bar{q} - \bar{q}_{kh}\|_I \leq \frac{1}{\alpha} \|z(\bar{q}) - z_{kh}(\bar{q})\|_I.$$

*Proof.* The proof is similar to the proof of Theorem 5.1. The optimality conditions (2.4) and (3.12) lead to

$$-j'_{kh}(\bar{q}_{kh})(\bar{q} - \bar{q}_{kh}) \leq 0 \leq -j'(\bar{q})(\bar{q} - \bar{q}_{kh}).$$

Using (3.15) we have with any  $p \in Q$ ,

$$\begin{aligned} \alpha \|\bar{q} - \bar{q}_{kh}\|_I^2 &\leq j''_{kh}(p)(\bar{q} - \bar{q}_{kh}, \bar{q} - \bar{q}_{kh}) \\ &= j'_{kh}(\bar{q})(\bar{q} - \bar{q}_{kh}) - j'_{kh}(\bar{q}_{kh})(\bar{q} - \bar{q}_{kh}) \\ &\leq j'_{kh}(\bar{q})(\bar{q} - \bar{q}_{kh}) - j'(\bar{q})(\bar{q} - \bar{q}_{kh}) \\ &= (z(\bar{q}) - z_{kh}(\bar{q}), \bar{q} - \bar{q}_{kh})_I. \end{aligned}$$

The desired assertion follows by Cauchy's inequality.  $\square$

This approach provides the optimal order of convergence stated in the following corollary.

**COROLLARY 5.11.** *Let the conditions of Theorem 5.10 be fulfilled. Then there holds*

$$\begin{aligned} \|\bar{q} - \bar{q}_{kh}\|_I &\leq \frac{C}{\alpha} k \{ \|\partial_t u(\bar{q})\|_I + \|\partial_t z(\bar{q})\|_I \} \\ &\quad + \frac{C}{\alpha} h^2 \{ \|\nabla^2 u_k(\bar{q})\|_I + \|\nabla^2 z_k(\bar{q})\|_I \} = \mathcal{O}(k + h^2). \end{aligned}$$

*Proof.* The proof follows directly from Theorem 5.10 and Propositions 4.3 and 4.4.  $\square$

**5.4. Postprocessing strategy.** In this section, we extend the postprocessing techniques initially proposed in [21] to the parabolic case. As described in section 3.3.4 we discretize the control by piecewise constants in time and space. To improve the quality of the approximation, we additionally employ the postprocessing step (3.18).

In what follows we will use the operator  $R_d$  defined for functions  $g \in C(\bar{\Omega})$  cellwise by

$$R_d g|_K = g(S_K), \quad K \in \mathcal{T}_h,$$

where  $S_K$  denotes the barycenter of the cell  $K$ . This operator allows for the following interpolation estimates.

LEMMA 5.12. *Let  $K \in \mathcal{T}_h$  be a given cell. Then we have that*

- *for  $g \in H^2(K)$ ,*

$$\left| \int_K (g(x) - (R_d g)(x)) dx \right| \leq Ch^2 |K|^{\frac{1}{2}} \|\nabla^2 g\|_{L^2(K)};$$

- *for  $g \in W^{1,p}(K)$  with  $n < p \leq \infty$ ,*

$$\|g - R_d g\|_{L^p(K)} \leq Ch \|\nabla g\|_{L^p(K)}.$$

*Proof.* The proof is done by standard arguments using the Bramble–Hilbert lemma; see [21] for details.  $\square$

The operator  $R_d$  will also be used for time-dependent functions  $g$  by setting  $(R_d g)(t) = R_d g(t)$ . There holds the following lemma.

LEMMA 5.13. *For a function  $g_k \in X_k^0 \cap L^2(I, H^2(\Omega))$  and a cellwise constant function  $p_d \in Q_d$ , the estimate*

$$(p_d, g_k - R_d g_k)_I \leq Ch^2 \|p_d\|_I \|\nabla^2 g_k\|_I$$

*holds.*

*Proof.* Using Lemma 5.12 we obtain

$$\begin{aligned} (p_d, g_k - R_d g_k)_I &= \sum_{m=1}^M \int_{I_m} (p_d(t), g_k(t) - R_d g_k(t)) dt \\ &= \sum_{m=1}^M k_m (p_d(t_m), g_k(t_m) - R_d g_k(t_m)) \\ &= \sum_{m=1}^M k_m \sum_{K \in \mathcal{T}_h} p_d(t_m, S_K) \int_K (g_k(t_m, x) - (R_d g_k)(t_m, x)) dx \\ &\leq Ch^2 \sum_{m=1}^M k_m \sum_{K \in \mathcal{T}_h} |p_d(t_m, S_K)| |K|^{\frac{1}{2}} \|\nabla^2 g_k(t_m)\|_{L^2(K)}. \end{aligned}$$

We complete the proof by Cauchy's inequality.  $\square$

LEMMA 5.14. *Let  $\bar{q}_k \in Q_{ad}$  be the solution of the semidiscrete optimization problem (3.4) and  $\bar{q}_\sigma \in Q_{d,ad}$  be the solution of the discrete problem (3.16), where the cellwise constant control discretization is employed. Then the following relation holds:*

$$(\alpha R_d \bar{q}_k + R_d z_k(\bar{q}_k), \bar{q}_\sigma - R_d \bar{q}_k)_I \geq 0.$$

*Proof.* From the optimality condition (3.5) for  $\bar{q}_k$ , we obtain

$$(\alpha \bar{q}_k(t_m, x) + z_k(\bar{q}_k)(t_m, x)) \cdot (\delta q(t_m, x) - \bar{q}_k(t_m, x)) \geq 0$$

for any  $\delta q \in Q_{d,\text{ad}}$  pointwise a.e. in  $\Omega$  and for  $m = 1, 2, \dots, M$ . For an arbitrary cell  $K \in \mathcal{T}_h$  we apply this formula for  $x = S_K$  and  $\delta q = \bar{q}_\sigma$ :

$$(\alpha \bar{q}_k(t_m, S_K) + z_k(\bar{q}_k)(t_m, S_K)) \cdot (\bar{q}_\sigma(t_m, S_K) - \bar{q}_k(t_m, S_K)) \geq 0.$$

This can be done because of the spatial continuity of  $z_k(\bar{q}_k)$ ,  $\bar{q}_k$ , and  $\bar{q}_\sigma$ . Due to the definition of  $R_d$ , this is equivalent to

$$(\alpha R_d \bar{q}_k(t_m, S_K) + R_d z_k(\bar{q}_k)(t_m, S_K)) \cdot (\bar{q}_\sigma(t_m, S_K) - R_d \bar{q}_k(t_m, S_K)) \geq 0.$$

Then, integration over  $K$  and  $I_m$ , summation over all  $K \in \mathcal{T}_h$ , and  $m = 1, 2, \dots, M$  lead to the proposed relation.  $\square$

LEMMA 5.15. *Let  $\bar{q}_k \in Q_{\text{ad}}$  be the solution of the semidiscrete optimization problem (3.4) and let  $\psi_{kh} \in X_{k,h}^{0,1}$ . Moreover, let Assumption 1 be fulfilled and  $n < p \leq \infty$ . Then, it holds that*

$$\begin{aligned} (\psi_{kh}, \bar{q}_k - R_d \bar{q}_k)_I &\leq \frac{C}{\alpha} h^2 \{ \|\nabla \psi_{kh}\|_I \|\nabla z_k(\bar{q}_k)\|_I + \|\psi_{kh}\|_{L^2(I, L^\infty(\Omega))} \|\nabla^2 z_k(\bar{q}_k)\|_I \} \\ &\quad + \frac{C}{\alpha} h^{2-\frac{1}{p}} \|\psi_{kh}\|_{L^2(I, L^\infty(\Omega))} \|\nabla z_k(\bar{q}_k)\|_{L^2(I, L^p(\Omega))}, \end{aligned}$$

provided that  $z_k(\bar{q}_k) \in L^2(I, W^{1,p}(\Omega))$ .

*Proof.* By means of the  $L^2$ -projection  $\pi_d: Q \rightarrow Q_d$ , we split

$$(\psi_{kh}, \bar{q}_k - R_d \bar{q}_k)_I = (\psi_{kh}, \bar{q}_k - \pi_d \bar{q}_k)_I + (\psi_{kh}, \pi_d \bar{q}_k - R_d \bar{q}_k)_I.$$

Using the optimality condition (3.9) and property (2.10) of the projection operator  $P_{Q_{\text{ad}}}$ , we have for the first term

$$\begin{aligned} (\psi_{kh}, \bar{q}_k - \pi_d \bar{q}_k)_I &= (\psi_{kh} - \pi_d \psi_{kh}, \bar{q}_k - \pi_d \bar{q}_k)_I \leq C h^2 \|\nabla \psi_{kh}\|_I \|\nabla \bar{q}_k\|_I \\ (5.12) \quad &\leq \frac{C}{\alpha} h^2 \|\nabla \psi_{kh}\|_I \|\nabla z_k(\bar{q}_k)\|_I. \end{aligned}$$

For the second term we obtain

$$\begin{aligned} (\psi_{kh}, \pi_d \bar{q}_k - R_d \bar{q}_k)_I &= \sum_{m=1}^M \int_{I_m} (\psi_{kh}(t), \pi_d \bar{q}_k(t) - R_d \bar{q}_k(t)) dt \\ &= \sum_{m=1}^M k_m (\psi_{kh}(t_m), \pi_d \bar{q}_k(t_m) - R_d \bar{q}_k(t_m)). \end{aligned}$$

Utilizing the fact that  $\pi_d \bar{q}_k(t_m)$  as well as  $R_d \bar{q}_k(t_m)$  are constant on each cell  $K$ , we

proceed with

$$\begin{aligned}
 (5.13) \quad & (\psi_{kh}(t_m), \pi_d \bar{q}_k(t_m) - R_d \bar{q}_k(t_m)) \\
 &= \sum_{K \in \mathcal{T}_h} \int_K \psi_{kh}(t_m, x) (\pi_d \bar{q}_k(t_m, x) - (R_d \bar{q}_k)(t_m, x)) dx \\
 &= \sum_{K \in \mathcal{T}_h} \frac{1}{|K|} \int_K \psi_{kh}(t_m, x) dx \int_K (\pi_d \bar{q}_k(t_m, x) - (R_d \bar{q}_k)(t_m, x)) dx \\
 &\leq \|\psi_{kh}(t_m)\|_{L^\infty(\Omega)} \sum_{K \in \mathcal{T}_h} \left| \int_K (\bar{q}_k(t_m, x) - (R_d \bar{q}_k)(t_m, x)) dx \right|.
 \end{aligned}$$

As in section 5.2, we split the last sum using the separation  $\mathcal{T}_h = \mathcal{T}_{h,m}^1 \cup \mathcal{T}_{h,m}^2 \cup \mathcal{T}_{h,m}^3$  for  $m = 1, 2, \dots, M$ . For the sum over  $\mathcal{T}_{h,m}^1 \cup \mathcal{T}_{h,m}^2$  we obtain by means of Lemma 5.12 and the fact that  $\bar{q}_k(t_m)$  equals either  $q_a$ ,  $q_b$ , or  $-\frac{1}{\alpha} z_k(\bar{q}_k)(t_m)$ :

$$\begin{aligned}
 (5.14) \quad & \sum_{K \in \mathcal{T}_{h,m}^1 \cup \mathcal{T}_{h,m}^2} \left| \int_K (\bar{q}_k(t_m, x) - R_d \bar{q}_k(t_m, x)) dx \right| \\
 &\leq Ch^2 \sum_{K \in \mathcal{T}_{h,m}^1 \cup \mathcal{T}_{h,m}^2} |K|^{\frac{1}{2}} \|\nabla^2 \bar{q}_k(t_m)\|_{L^2(K)} \\
 &\leq \frac{C}{\alpha} h^2 \|\nabla^2 z_k(\bar{q}_k)(t_m)\|.
 \end{aligned}$$

For the part of the sum over  $\mathcal{T}_{h,m}^3$ , the estimate of Lemma 5.12, Assumption 1, the optimality condition (3.9), and property (2.10) lead to

$$\begin{aligned}
 (5.15) \quad & \sum_{K \in \mathcal{T}_{h,m}^3} \left| \int_K (\bar{q}_k(t_m, x) - R_d \bar{q}_k(t_m, x)) dx \right| \leq \sum_{K \in \mathcal{T}_{h,m}^3} |K|^{1-\frac{1}{p}} \|\bar{q}_k(t_m) - R_d \bar{q}_k(t_m)\|_{L^p(K)} \\
 &\leq Ch \sum_{K \in \mathcal{T}_{h,m}^3} |K|^{1-\frac{1}{p}} \|\nabla \bar{q}_k(t_m)\|_{L^p(K)} \\
 &\leq \frac{C}{\alpha} h^{2-\frac{1}{p}} \|\nabla z_k(\bar{q}_k(t_m))\|_{L^p(\Omega)}.
 \end{aligned}$$

Inserting (5.14) and (5.15) into (5.13) and collecting the estimates (5.12) and (5.13) completes the proof.  $\square$

The following theorem provides a supercloseness result on the difference  $R_d \bar{q}_k - \bar{q}_\sigma$ .

**THEOREM 5.16.** *Let  $\bar{q}_k \in Q_{ad}$  be the solution of the semidiscretized optimization problem (3.4) and  $\bar{q}_\sigma \in Q_{d,ad}$  be the solution of the discrete problem (3.16), where the cellwise constant discretization for the control variable is employed. Moreover, let*

Assumption 1 be fulfilled and  $n < p \leq \infty$ . Then, it holds that

$$\begin{aligned} \|R_d \bar{q}_k - \bar{q}_\sigma\|_I &\leq \frac{C}{\alpha} h^2 \left\{ \|\nabla^2 u_k(\bar{q}_k)\|_I + \frac{1}{\alpha} \|\nabla z_k(\bar{q}_k)\|_I + \left(1 + \frac{1}{\alpha}\right) \|\nabla^2 z_k(\bar{q}_k)\|_I \right\} \\ &\quad + \frac{C}{\alpha^2} h^{2-\frac{1}{p}} \|\nabla z_k(\bar{q}_k)\|_{L^2(I, L^p(\Omega))}, \end{aligned}$$

provided that  $z_k(\bar{q}_k) \in L^2(I, W^{1,p}(\Omega))$ .

*Proof.* As before, we proceed with an arbitrary  $p \in Q$ ,

$$\begin{aligned} \alpha \|R_d \bar{q}_k - \bar{q}_\sigma\|_I^2 &\leq j''_{kh}(p)(R_d \bar{q}_k - \bar{q}_\sigma, R_d \bar{q}_k - \bar{q}_\sigma) \\ &= j'_{kh}(R_d \bar{q}_k)(R_d \bar{q}_k - \bar{q}_\sigma) - j'_{kh}(\bar{q}_\sigma)(R_d \bar{q}_k - \bar{q}_\sigma). \end{aligned}$$

By means of the inequality

$$-j'_{kh}(\bar{q}_\sigma)(R_d \bar{q}_k - \bar{q}_\sigma) \leq 0 \leq -(\alpha R_d \bar{q}_k + R_d z_k(\bar{q}_k), R_d \bar{q}_k - \bar{q}_\sigma)_I,$$

which is implied by the optimality of  $\bar{q}_\sigma$  and Lemma 5.14, and by means of the explicit representation of  $j'_{kh}$  from (3.13), we obtain

$$\begin{aligned} \alpha \|R_d \bar{q}_k - \bar{q}_\sigma\|_I^2 &\leq (z_{kh}(R_d \bar{q}_k) - R_d z_k(\bar{q}_k), R_d \bar{q}_k - \bar{q}_\sigma)_I \\ (5.16) \quad &\leq (z_{kh}(R_d \bar{q}_k) - z_k(\bar{q}_k), R_d \bar{q}_k - \bar{q}_\sigma)_I \\ &\quad + (z_k(\bar{q}_k) - R_d z_k(\bar{q}_k), R_d \bar{q}_k - \bar{q}_\sigma)_I. \end{aligned}$$

For the first term on the right-hand side of (5.16), we have by Cauchy's inequality,

$$(z_{kh}(R_d \bar{q}_k) - z_k(\bar{q}_k), R_d \bar{q}_k - \bar{q}_\sigma)_I \leq \|z_{kh}(R_d \bar{q}_k) - z_k(\bar{q}_k)\|_I \|R_d \bar{q}_k - \bar{q}_\sigma\|_I.$$

By insertion of  $z_{kh}(\bar{q}_k)$ , the term  $\|z_{kh}(R_d \bar{q}_k) - z_k(\bar{q}_k)\|_I$  is further estimated as

$$(5.17) \quad \|z_{kh}(R_d \bar{q}_k) - z_k(\bar{q}_k)\|_I \leq \|z_{kh}(R_d \bar{q}_k) - z_{kh}(\bar{q}_k)\|_I + \|z_{kh}(\bar{q}_k) - z_k(\bar{q}_k)\|_I.$$

Due to the stability estimate of the fully discrete adjoint solution (see Proposition 4.2), the first term is bounded by

$$(5.18) \quad \|z_{kh}(R_d \bar{q}_k) - z_{kh}(\bar{q}_k)\|_I \leq C \|u_{kh}(R_d \bar{q}_k) - u_{kh}(\bar{q}_k)\|_I.$$

Further, we have by means of the discrete state equation (3.10) and the discrete adjoint equation (3.14),

$$\|u_{kh}(R_d \bar{q}_k) - u_{kh}(\bar{q}_k)\|_I^2 = (z_{kh}(\bar{q}_k) - z_{kh}(R_d \bar{q}_k), \bar{q}_k - R_d \bar{q}_k)_I.$$

With  $\psi_{kh} = z_{kh}(\bar{q}_k) - z_{kh}(R_d \bar{q}_k)$  in Lemma 5.15, we have

$$\begin{aligned} \|u_{kh}(R_d \bar{q}_k) - u_{kh}(\bar{q}_k)\|_I^2 &\leq \frac{C}{\alpha} h^2 \left\{ \|\nabla(z_{kh}(\bar{q}_k) - z_{kh}(R_d \bar{q}_k))\|_I \|\nabla z_k(\bar{q}_k)\|_I \right. \\ &\quad \left. + \|z_{kh}(\bar{q}_k) - z_{kh}(R_d \bar{q}_k)\|_{L^2(I, L^\infty(\Omega))} \|\nabla^2 z_k(\bar{q}_k)\|_I \right\} \\ &\quad + \frac{C}{\alpha} h^{2-\frac{1}{p}} \|z_{kh}(\bar{q}_k) - z_{kh}(R_d \bar{q}_k)\|_{L^2(I, L^\infty(\Omega))} \|\nabla z_k(\bar{q}_k)\|_{L^2(I, L^p(\Omega))}, \end{aligned}$$



and the stability estimates from Proposition 4.2 and Lemma 4.5,

$$\begin{aligned}\|\nabla(z_{kh}(q_k) - z_{kh}(R_d q_k))\|_I &\leq C\|u_{kh}(R_d q_k) - u_{kh}(q_k)\|_I, \\ \|z_{kh}(q_k) - z_{kh}(R_d q_k)\|_{L^2(I, L^\infty(\Omega))} &\leq C\|u_{kh}(R_d q_k) - u_{kh}(q_k)\|_I,\end{aligned}$$

yield the following intermediary result:

$$\begin{aligned}\|u_{kh}(R_d \bar{q}_k) - u_{kh}(\bar{q}_k)\|_I &\leq \frac{C}{\alpha} h^2 \{ \|\nabla z_k(\bar{q}_k)\|_I + \|\nabla^2 z_k(\bar{q}_k)\|_I \} \\ &\quad + \frac{C}{\alpha} h^{2-\frac{1}{p}} \|\nabla z_k(\bar{q}_k)\|_{L^2(I, L^p(\Omega))}.\end{aligned}$$

We proceed by inserting this in (5.18) and in (5.17). Together with an estimate for the second term on the right-hand side of (5.17) from Proposition 4.4, this leads to

$$\begin{aligned}(5.19) \quad \|z_{kh}(R_d \bar{q}_k) - z_k(\bar{q}_k)\|_I &\leq C h^2 \left\{ \|\nabla^2 u_k(\bar{q}_k)\|_I + \frac{1}{\alpha} \|\nabla z_k(\bar{q}_k)\|_I \right. \\ &\quad \left. + \left(1 + \frac{1}{\alpha}\right) \|\nabla^2 z_k(\bar{q}_k)\|_I \right\} + \frac{C}{\alpha} h^{2-\frac{1}{p}} \|\nabla z_k(\bar{q}_k)\|_{L^2(I, L^p(\Omega))}.\end{aligned}$$

By applying Lemma 5.13 with  $p_d = R_d \bar{q}_k - \bar{q}_\sigma$  to the second term on the right-hand side of (5.16), we get

$$(z_k(\bar{q}_k) - R_d z_k(\bar{q}_k), R_d \bar{q}_k - \bar{q}_\sigma)_I \leq C h^2 \|R_d \bar{q}_k - \bar{q}_\sigma\|_I \|\nabla^2 z_k(\bar{q}_k)\|_I.$$

The asserted result is obtained by insertion of the last two estimates into (5.16).  $\square$

Based on this theorem, we state the main result of this section concerning the order of convergence of the error between  $\bar{q}$  and  $\tilde{q}_\sigma$ , where  $\tilde{q}_\sigma$  is defined using the postprocessing step (3.18).

**COROLLARY 5.17.** *Let the conditions of Theorem 5.16 be fulfilled. Then, there holds*

$$\begin{aligned}\|\bar{q} - \tilde{q}_\sigma\|_I &\leq \frac{C}{\alpha} \left(1 + \frac{1}{\alpha}\right) k \{ \|\partial_t u(\bar{q})\|_I + \|\partial_t z(\bar{q})\|_I \} \\ &\quad + \frac{C}{\alpha} \left(1 + \frac{1}{\alpha}\right) h^2 \left\{ \|\nabla^2 u_k(\bar{q}_k)\|_I + \frac{1}{\alpha} \|\nabla z_k(\bar{q}_k)\|_I + \left(1 + \frac{1}{\alpha}\right) \|\nabla^2 z_k(\bar{q}_k)\|_I \right\} \\ &\quad + \frac{C}{\alpha^2} \left(1 + \frac{1}{\alpha}\right) h^{2-\frac{1}{p}} \|\nabla z_k(\bar{q}_k)\|_{L^2(I, L^p(\Omega))} = \mathcal{O}(k + h^{2-\frac{1}{p}}).\end{aligned}$$

*Proof.* From the optimality condition (2.9) and the definition (3.18) of  $\tilde{q}_\sigma$  we have the representation

$$\|\bar{q} - \tilde{q}_\sigma\|_I = \left\| P_{Q_{\text{ad}}} \left( -\frac{1}{\alpha} z(\bar{q}) \right) - P_{Q_{\text{ad}}} \left( -\frac{1}{\alpha} z_{kh}(\bar{q}_\sigma) \right) \right\|_I.$$

By means of the Lipschitz continuity of  $P_{Q_{\text{ad}}}$  on  $L^2(I, L^2(\Omega))$ , this leads to

$$(5.20) \quad \alpha \|\bar{q} - \tilde{q}_\sigma\|_I \leq \|z(\bar{q}) - z_{kh}(\bar{q}_\sigma)\|_I \leq \|z(\bar{q}) - z_k(\bar{q}_k)\|_I + \|z_k(\bar{q}_k) - z_{kh}(\bar{q}_\sigma)\|_I.$$

The first term is controlled by means of Proposition 4.1, Theorem 5.1, and Proposition 4.3 as

$$\begin{aligned}
 \|z(\bar{q}) - z_k(\bar{q}_k)\|_I &\leq \|z(\bar{q}) - z_k(\bar{q})\|_I + \|z_k(\bar{q}) - z_k(\bar{q}_k)\|_I \\
 &\leq \|z(\bar{q}) - z_k(\bar{q})\|_I + C\|u_k(\bar{q}) - u_k(\bar{q}_k)\|_I \\
 &\leq \|z(\bar{q}) - z_k(\bar{q})\|_I + C\|\bar{q} - \bar{q}_k\|_I \\
 &\leq \left(1 + \frac{C}{\alpha}\right) \|z(\bar{q}) - z_k(\bar{q})\|_I \\
 &\leq C\left(1 + \frac{1}{\alpha}\right) k\{\|\partial_t u(\bar{q})\|_I + \|\partial_t z(\bar{q})\|_I\}.
 \end{aligned}$$

The second term can be estimated by means of the stability result of Proposition 4.2 as

$$\begin{aligned}
 \|z_k(\bar{q}_k) - z_{kh}(\bar{q}_\sigma)\|_I &\leq \|z_k(\bar{q}_k) - z_{kh}(R_d \bar{q}_k)\|_I + \|z_{kh}(R_d \bar{q}_k) - z_{kh}(\bar{q}_\sigma)\|_I \\
 &\leq \|z_k(\bar{q}_k) - z_{kh}(R_d \bar{q}_k)\|_I + C\|u_{kh}(R_d \bar{q}_k) - u_{kh}(\bar{q}_\sigma)\|_I \\
 &\leq \|z_k(\bar{q}_k) - z_{kh}(R_d \bar{q}_k)\|_I + C\|R_d \bar{q}_k - \bar{q}_\sigma\|_I.
 \end{aligned}$$

Inserting the two last inequalities into (5.20) and applying the estimates from (5.19) and Theorem 5.16 yield the stated assertion.  $\square$

The choice of  $p$  in Corollary 5.17 follows the description in section 5.2 requiring  $z_k(\bar{q}_k) \in L^2(I, W^{1,p}(\Omega))$ . Due to the fact that  $\|z_k(\bar{q}_k)\|_{L^2(I, H^2(\Omega))}$  is bounded independently of  $k$ , the result in Corollary 5.17 holds for any  $n < p < \infty$  in the two dimensional case, leading to the order of convergence  $\mathcal{O}(k + h^{2-\frac{1}{p}})$ . In the three dimensional case we obtain  $p = 6$  and therefore  $\mathcal{O}(k + h^{\frac{11}{6}})$ . If in addition  $\|z_k(\bar{q}_k)\|_{L^2(I, W^{1,\infty}(\Omega))}$  is bounded, then we have in both cases the order of convergence  $\mathcal{O}(k + h^2)$ .

**6. Numerical results.** In this section, we are going to validate the a priori error estimates for the error in the control, state, and adjoint state numerically. To this end, we consider the following concretion of the optimal control problem (2.2) with known exact solution on  $\Omega \times I = (0, 1)^2 \times (0, 0.1)$  and homogeneous Dirichlet boundary conditions. According to the first part of this article [20], the right-hand side  $f$ , the desired state  $\hat{u}$ , and the initial condition  $u_0$  are given in terms of the eigenfunctions

$$w_a(t, x_1, x_2) := \exp(a\pi^2 t) \sin(\pi x_1) \sin(\pi x_2), \quad a \in \mathbb{R},$$

of the operator  $\pm \partial_t - \Delta$  as

$$\begin{aligned}
 f(t, x_1, x_2) &:= -\pi^4 w_a(t, x_1, x_2) - P_{Q_{\text{ad}}}(-\pi^4 \{w_a(t, x_1, x_2) - w_a(T, x_1, x_2)\}), \\
 \hat{u}(t, x_1, x_2) &:= \frac{a^2 - 5}{2 + a} \pi^2 w_a(t, x_1, x_2) + 2\pi^2 w_a(T, x_1, x_2), \\
 u_0(x_1, x_2) &:= \frac{-1}{2 + a} \pi^2 w_a(0, x_1, x_2),
 \end{aligned}$$

with  $P_{Q_{\text{ad}}}$  given by (2.8) with  $q_a = -70$  and  $q_b = -1$ . For this choice of data and with the regularization parameter  $\alpha$  chosen as  $\alpha = \pi^{-4}$ , the optimal solution triple

$(\bar{q}, \bar{u}, \bar{z})$  of the optimal control problem (2.2) is given by

$$\bar{q}(t, x_1, x_2) := P_{Q_{\text{ad}}}(-\pi^4\{w_a(t, x_1, x_2) - w_a(T, x_1, x_2)\}),$$

$$\bar{u}(t, x_1, x_2) := \frac{-1}{2+a}\pi^2 w_a(t, x_1, x_2),$$

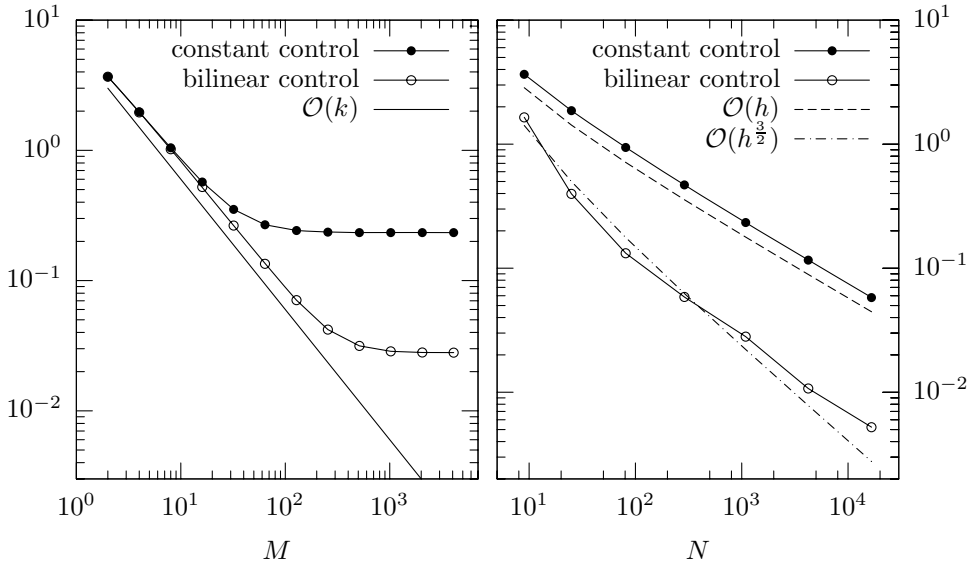
$$\bar{z}(t, x_1, x_2) := w_a(t, x_1, x_2) - w_a(T, x_1, x_2).$$

We are going to validate the estimates developed in the previous section by separating the discretization errors. That is, we consider at first the behavior of the error for a sequence of discretizations with decreasing size of the time steps and a fixed spatial triangulation with  $N = 1089$  nodes. Second, we examine the behavior of the error under refinement of the spatial triangulation for  $M = 2048$  time steps.

The state discretization is chosen as cG(1)dG(0), i.e.,  $r = 0, s = 1$ . For the control discretization we use the same temporal and spatial meshes as for the state variable and present results for two choices of the discrete control space  $Q_a$ : cG(1)dG(0) and dG(0)dG(0). For the following computations, we choose the free parameter  $a$  to be  $-\sqrt{5}$ .

The optimal control problems are solved by the optimization library RoDoBo [22] and the finite element toolkit GASCOIGNE [11] using a primal-dual active set strategy (cf. [3, 14]) in combination with a conjugate gradient method applied to the reduced problem (3.16).

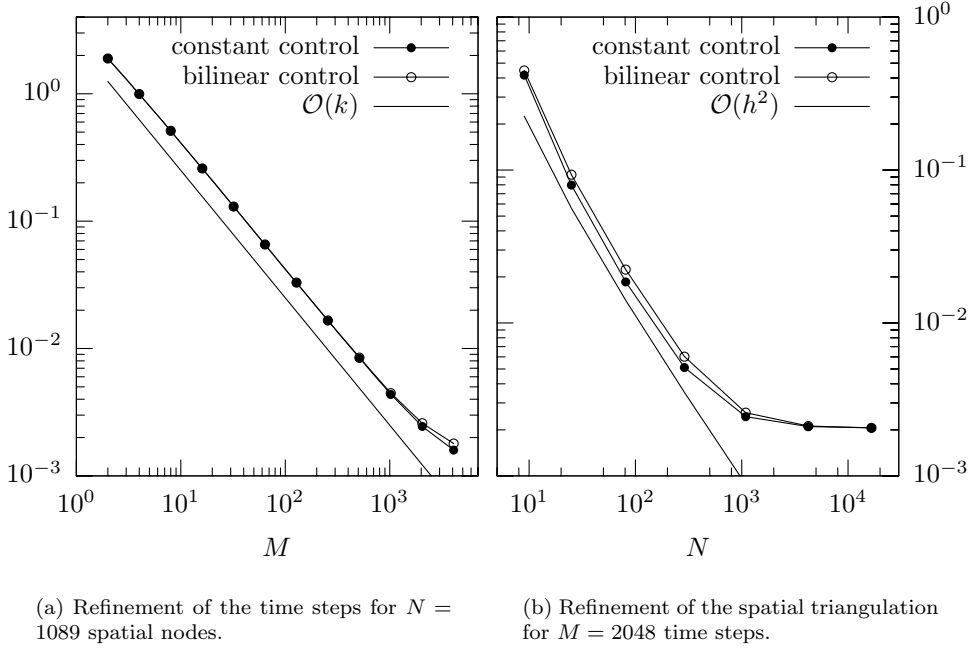
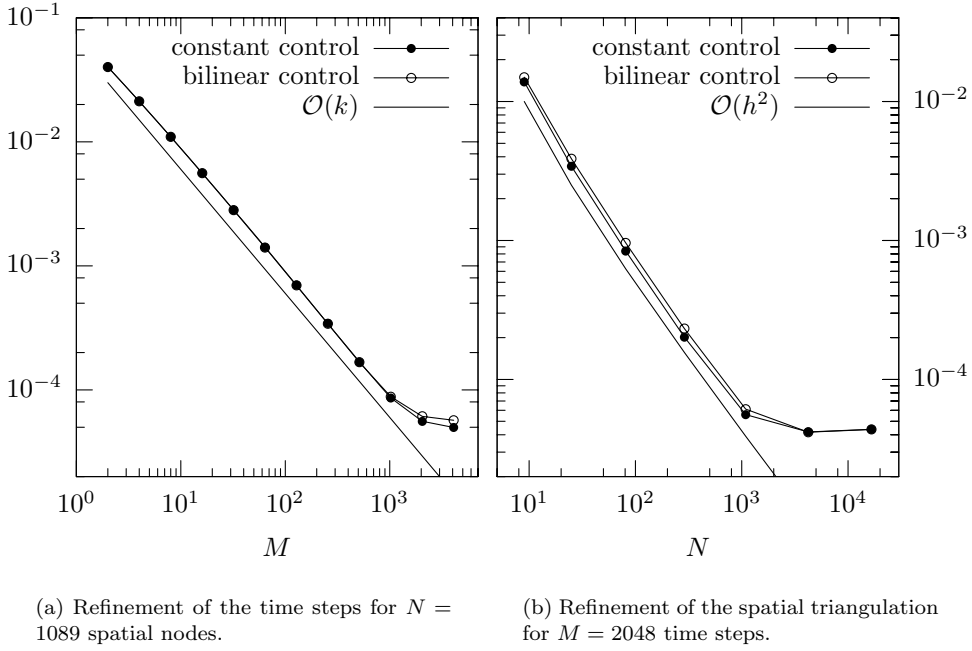
Figure 6.1(a) depicts the development of the error under refinement of the temporal step size  $k$ . Up to the spatial discretization error it exhibits the proven convergence order  $\mathcal{O}(k)$  for both kinds of spatial discretization of the control space.



(a) Refinement of the time steps for  $N = 1089$  spatial nodes.

(b) Refinement of the spatial triangulation for  $M = 2048$  time steps.

FIG. 6.1. Discretization error  $\|\bar{q} - \bar{q}_\sigma\|_I$ .

FIG. 6.2. Discretization error  $\|\bar{u} - \bar{u}_\sigma\|_I$ .FIG. 6.3. Discretization error  $\|\bar{z} - \bar{z}_\sigma\|_I$ .

For piecewise constant control (dG(0)dG(0) discretization), the discretization error is already reached at 128 time steps, whereas in the case of bilinear control (cG(1)dG(0) discretization), the number of time steps could be increased up to  $M = 1024$  until reaching the spatial accuracy. This illustrates the convergence results from sections 5.1 and 5.2 with respect to the *temporal* discretization.

In Figure 6.1(b) the development of the error in the control variable under spatial refinement is shown. The expected order  $\mathcal{O}(h)$  for piecewise constant control (dG(0)dG(0) discretization) and  $\mathcal{O}(h^{\frac{3}{2}})$  for bilinear control (cG(1)dG(0) discretization) are observed. This illustrates the convergence results from sections 5.1 and 5.2 with respect to the *spatial* discretization.

Figures 6.2 and 6.3 show the errors in the state and in the adjoint variables,  $\|\bar{u} - \bar{u}_\sigma\|_I$  and  $\|\bar{z} - \bar{z}_\sigma\|_I$ , for separate refinement of the time and space discretization. Thereby, we observe convergence of order  $\mathcal{O}(k + h^2)$  regardless of the type of spatial discretization used for the controls. This is consistent with the results proved in the previous section. Since the postprocessing strategy presented in section 5.4 relies essentially on the convergence properties of the adjoint variable, Figure 6.3 confirms the proven order of convergence of the error  $\|\bar{q} - \bar{q}_\sigma\|_I$ .

#### REFERENCES

- [1] N. ARADA, E. CASAS, AND F. TRÖLTZSCH, *Error estimates for a semilinear elliptic optimal control problem*, Comput. Optim. Appl., 23 (2002), pp. 201–229.
- [2] R. BECKER AND B. VEXLER, *Optimal control of the convection-diffusion equation using stabilized finite element methods*, Numer. Math., 106 (2007), pp. 349–367.
- [3] M. BERGOUNIOUX, K. ITO, AND K. KUNISCH, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.
- [4] E. CASAS, *Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems*, Adv. Comput. Math., 26 (2007), pp. 137–153.
- [5] E. CASAS AND M. MATEOS, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems. Continuous piecewise linear approximations*, in Systems, Control, Modeling and Optimization, IFIP Int. Fed. Inf. Process. 202, Springer, New York, 2006, pp. 91–101.
- [6] E. CASAS AND M. MATEOS, *Error estimates for the numerical approximation of Neumann control problems*, Comput. Optim. Appl., 2007, published electronically.
- [7] E. CASAS, M. MATEOS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl., 31 (2005), pp. 193–220.
- [8] E. CASAS AND F. TRÖLTZSCH, *Error estimates for linear-quadratic elliptic control problems*, in Analysis and Optimization of Differential Systems, V. Barbu, I. Lasiecka, D. Tiba, and C. Varsan, eds., Kluwer Academic Publishers, Boston, 2003, pp. 89–100.
- [9] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, Classics Appl. Math., 40, SIAM, Philadelphia, 2002.
- [10] R. FALK, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.
- [11] GASCOIGNE: *The Finite Element Toolkit*, <http://www.gascoigne.uni-hd.de/>.
- [12] T. GEVECI, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, M2AN Math. Model. Numer. Anal., 13 (1979), pp. 313–328.
- [13] M. HINZE, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–61.
- [14] K. KUNISCH AND A. RÖSCH, *Primal-dual active set strategy for a general class of constrained optimal control problems*, SIAM J. Optim., 13 (2002), pp. 321–334.
- [15] I. LASIECKA AND K. MALANOWSKI, *On regularity of solutions to convex optimal control problems with control constraints for parabolic systems*, Control Cybern., 6 (1977), pp. 57–74.
- [16] I. LASIECKA AND K. MALANOWSKI, *On discrete-time Ritz-Galerkin approximation of control constrained optimal control problems for parabolic systems*, Control Cybern., 7 (1978), pp. 21–36.

- [17] J.-L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Grundlehren Math. Wiss., 170, Springer, Berlin, 1971.
- [18] K. MALANOWSKI, *Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal-control problems*, Appl. Math. Optim., 8 (1981), pp. 69–95.
- [19] R. S. MCNIGHT AND W. E. BOSARGE, JR., *The Ritz–Galerkin procedure for parabolic control problems*, SIAM J. Control Optim., 11 (1973), pp. 510–524.
- [20] D. MEIDNER AND B. VEXLER, *A priori error estimates for space-time finite element discretization of parabolic optimal control problems. Part I: Problems without control constraints*, SIAM J. Control Optim., 47 (2008), pp. 1150–1177.
- [21] C. MEYER AND A. RÖSCH, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim., 43 (2004), pp. 970–985.
- [22] *RoDoBo: A C++ Library for Optimization with Stationary and Nonstationary PDEs*, <http://rodoobo.uni-hd.de/>.
- [23] A. RÖSCH, *Error estimates for parabolic optimal control problems with control constraints*, Z. Anal. Anwendungen, 23 (2004), pp. 353–376.
- [24] A. RÖSCH, *Error estimates for linear-quadratic control problems with control constraints*, Optim. Methods Softw., 21 (2006), pp. 121–134.
- [25] A. RÖSCH AND B. VEXLER, *Optimal control of the Stokes equations: A priori error analysis for finite element discretization with postprocessing*, SIAM J. Numer. Anal., 44 (2006), pp. 1903–1920.
- [26] M. SCHMICH AND B. VEXLER, *Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations*, SIAM J. Sci. Comput., 30 (2008), pp. 369–393.
- [27] V. THOMÉE, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer Ser. Comput. Math., 25, Springer-Verlag, Berlin, 1997.
- [28] R. WINTHER, *Error estimates for a Galerkin approximation of a parabolic control problem*, Ann. Math. Pura Appl. (4), 117 (1978), pp. 173–206.

## OPTIMAL TRANSPORTATION WITH TRAFFIC CONGESTION AND WARDROP EQUILIBRIA\*

G. CARLIER<sup>†</sup>, C. JIMENEZ<sup>†</sup>, AND F. SANTAMBROGIO<sup>‡</sup>

**Abstract.** In the classical Monge–Kantorovich problem, the transportation cost depends only on the amount of mass sent from sources to destinations and not on the paths followed by this mass. Thus, it does not allow for congestion effects. Using the notion of traffic intensity, we propose a variant, taking into account congestion. This variant is a continuous version of a well-known traffic problem on networks that is studied both in economics and in operational research. The interest of this problem is in its relations with traffic equilibria of Wardrop type. What we prove in the paper is exactly the existence and the variational characterization of equilibria in a continuous space setting.

**Key words.** optimal transportation, traffic congestion, Wardrop equilibria

**AMS subject classifications.** 90C46, 91A13

**DOI.** 10.1137/060672832

**1. Introduction.** Researchers in the field of applied traffic modeling have long emphasized the role of congestion in networks. In the early 1950s, Wardrop (see [14]) considered the situation where a large number of vehicles have to go from one location to another, connected by a finite number of different roads. Each vehicle has to choose one road (or a probability on the set of possible roads) to minimize some transportation cost which depends not only on the road chosen but also on the total flow of vehicles on this road. Some roads may be better than others, for instance, because they are shorter or wider, but they are all affected by congestion effects: the “cost” (in terms of time, say) for a vehicle of following a road depends increasingly on the total number of vehicles that choose to use it. Wardrop gave a minimal stability requirement for transportation strategies: the cost of every actually used road should be less than or equal to that which would be experienced by a single vehicle on any other road. In particular, there is an equilibrium concept (all actually used roads have the same cost, i.e., they compensate in terms of congestion for their differences given by length and other conditions) and a minimality concept as well (those roads are minimal among all the possible ones). This natural equilibrium concept has been very popular since its introduction because of its applications to networks, of course, but also due to the development of noncooperative game theory. To the best of our knowledge, the study of Wardrop equilibria has mainly been restricted to the case where admissible roads are given by a finite graph. In the present paper our main goal is to introduce Wardrop’s concepts in a continuous state setting, to prove the existence of such equilibria, and to relate it to the optimal transportation problem with congestion (Theorem 4.2).

The key point of the discrete theory about Wardrop equilibria is the fact that such an equilibrium problem may be linked to a variational one (see, for instance, [2] and the references therein). Suppose the cost for a road  $\sigma$  is given by  $\int_{\sigma} g(i)$ ,

---

\*Received by the editors October 20, 2006; accepted for publication (in revised form) December 10, 2007; published electronically March 21, 2008.

<http://www.siam.org/journals/sicon/47-3/67283.html>

<sup>†</sup>CEREMADE, UMR CNRS 7534, Université Paris IX Dauphine, Pl. de Lattre de Tassigny, 75775 Paris Cedex 16, France (carlier@ceremade.dauphine.fr, jimenez@ceremade.dauphine.fr).

<sup>‡</sup>Scuola Normale Superiore, Classe di Scienze, Piazza dei Cavalieri 7, 56126, Pisa, Italy (f.santambrogio@sns.it).

where  $i$  denotes the traffic intensity along the road (it can be nonconstant, since at branching points of different roads, the traffic splits according to Kirchhoff's law) where  $g \geq 0$  is an increasing function modeling congestion effects. We may say that  $g(i)$  is the cost per unit length for a road where the traffic intensity is  $i$ . It is known that looking for an equilibrium is equivalent to solving a minimization problem with total congestion cost  $\int_N H(i)$ , where  $N$  represents the whole network and  $H$  is another increasing function which is linked to  $g$ . The relation which is necessary to ensure that minima of the variational problem coincide with solutions of the equilibrium one is  $H' = g$ . Notice that solving this problem does not amount, in general, to finding the configuration which minimizes the total cost paid by vehicles, since this quantity is represented instead by the integral  $\int_N ig(i)$ . The two functions  $H(i)$  and  $ig(i)$  are the same up to multiplicative constants in the case of power functions, but otherwise they give rise to different optimization problems.

The unknown, both in the equilibrium and in the optimization problems, is the distribution of vehicles along the possible paths. In a continuous-setting language, this is a measure on the set of paths. The constraints on such a measure are given by the data: usually the total amount of vehicles commuting from a point  $x$  to a point  $y$  of the network is prescribed for every pair  $(x, y)$ . This corresponds to fixing a measure  $\gamma$  on the set of sources-destinations pairs. As a possible alternative, one could prescribe the total quantity of vehicles leaving  $x$  and the total quantity reaching  $y$ , and look at all possible couplings between these "boundary data." In this case, the fixed data are two measures  $\mu_0$  and  $\mu_1$  modeling the distribution of sources and destinations separately. Also the coupling or *transport plan*  $\gamma$  (i.e., a measure on the sources-destinations space whose projections on the two coordinates are  $\mu_0$  and  $\mu_1$ ) is part of the unknowns.

In the present paper, we introduce a variational problem in a continuous setting, i.e., when the data on sources and destinations are arbitrary probability measures on a domain  $\Omega \subset \mathbb{R}^2$  and the allowed paths are all possible Lipschitz curves connecting points of  $\Omega$ . In this problem, the functional is built as a total cost arising from the congested transport problem. This means that we start from a congestion function  $g$  and we look at a minimization problem involving the function  $i \mapsto H(i) = ig(i)$ . Obviously, the concept of traffic intensity associated to a probability measure on a suitable set of paths has to be carefully defined in the continuous framework, which we do in section 2.2. The definition of traffic intensity we provide is the path-dependent analogue of the well-known notion of *transport density* in Monge's problem (see Bouchitté, Buttazzo, and Seppecher [4], Bouchitté and Buttazzo [3], Caffarelli, Feldman, and McCann [7]).

After defining the cost and the constraint, we may forget about the origin of the function  $H$  and, as in the discrete case, look at the minimization problem: we are concerned with existence, finiteness of the minimum value, and optimality conditions. Interestingly, in the continuous case, this variational problem takes the form of a *path-dependent* optimal transportation problem.

We show, under some extra assumptions on  $H$ , that solutions of the variational problem exist and are characterized by two optimality conditions. One of them is nothing but the continuous counterpart of Wardrop equilibria: the quantity  $H'(i) \geq 0$  defines a metric on  $\Omega$  and the paths that are actually used must be geodesics for this metric. The other optimality condition is peculiar to the case of nonfixed couplings and is much more linked to Monge–Kantorovich optimal transport theory (which can be traced back to Monge [12]; see [1], [5], and [13] for a modern account of the theory). In fact, once a solution is found, we can call  $c(x, y)$  the minimal cost for commuting



from  $x$  to  $y$ , according to the previously mentioned metric  $H'(i)$ . It turns out that, for an optimal solution,  $\gamma$  solves the optimal transport problem between  $\mu_0$  and  $\mu_1$  with respect to the cost  $c$ . The case of a nonfixed transport plan is the one which is developed in this paper because it is the richest case from a mathematical point of view. Moreover, it allows for a direct comparison with optimal transport problems à la Monge and Kantorovich. A variant of the problem we study is the case where, instead of allowing any transport plan between  $\mu_0$  and  $\mu_1$ , we prescribe a given convex and compact subset of couplings compatible with the data. All of the results of this paper can be extended to such a variant, which includes the case of a single prescribed transport plan.

In the minimization problem we study, the functional is linked to a Monge transport cost, in the case of a nonuniform metric. It corresponds to a cost  $\int d(x, y) d\gamma(x, y)$ , where  $d$  is a distance which in this case is unknown as it depends on the traffic intensity itself. Other transportation costs, for instance, squared distances, are very important for the applications of the Monge–Kantorovich theory but are not directly linked to this equilibrium problem. Notice, however, that, from an individual point of view (i.e., the viewpoint of the equilibrium issue), minimizing the displacement cost or its square is the same.

The presentation of the model, the construction of the functional, and its links with the Monge–Kantorovich theory follow the discrete case in its generality. Yet, most of the results require some additional assumptions on the function  $H$ . In particular,  $H$  is required to behave like a power  $H(i) = i^q$  with  $1 < q < 2$ . This is needed both for technical and feasibility reasons. First, it ensures the validity of some crucial estimates giving continuity results, providing well-defined transport costs  $c(x, y)$ . Second, it turns out that if  $\gamma$  is discrete (i.e., any path is allowed but the set of sources and destinations is finite), then the condition for having a finite minimal cost is exactly  $q < 2$ . Moreover, if we come back to the congestion function  $g$ , then this condition on  $H$  corresponds to  $g$  behaving like a concave power  $i^{q-1}$ , which is very natural from an economic point of view.

## 2. Optimal transportation with congestion.

**2.1. Notation.** Given a Polish (i.e., metrizable, separable, and complete) space  $X$ , we will denote, respectively, by  $\mathcal{M}_+(X)$  and  $\mathcal{M}_+^1(X)$  the set of positive and finite Radon measures on  $X$  and the set of Radon probability measures on  $X$ . If  $X$  and  $Y$  are Polish spaces,  $\mu \in \mathcal{M}_+^1(X)$ , and  $f : X \rightarrow Y$  is a Borel map, we shall denote by  $f\#\mu$  the push forward of  $\mu$  through  $f$ , i.e., the element of  $\mathcal{M}_+^1(Y)$  defined by  $f\#\mu(B) = \mu(f^{-1}(B))$  for every Borel subset  $B$  of  $Y$ .

In what follows,  $\mathcal{L}^d$  denotes the  $d$ -dimensional Lebesgue measure. If  $\mu$  and  $\nu$  are in  $\mathcal{M}_+^1(\mathbb{R}^d)$ , then  $\frac{d\mu}{d\nu}$  denotes the Radon–Nikodym derivative of  $\mu$  with respect to  $\nu$ . We shall write  $\mu \ll \nu$  to express that  $\mu$  is absolutely continuous with respect to  $\nu$ , in which case, slightly abusing notation, we will identify  $\mu$  with the Radon–Nikodym derivative  $\frac{d\mu}{d\nu}$ .

The data of our problem are  $\Omega$  (its closure  $\overline{\Omega}$  modeling the city, say), which is some open bounded convex subset of  $\mathbb{R}^2$ , and two probability measures  $\mu_0$  and  $\mu_1$  in  $\mathcal{M}_+^1(\overline{\Omega})$  giving, respectively, the distribution of residents and services in the city  $\overline{\Omega}$ . The set of transportation plans associated to  $\mu_0$  and  $\mu_1$  will be denoted  $\Pi(\mu_0, \mu_1)$  and consists of the probability measures on  $\overline{\Omega} \times \overline{\Omega}$  having  $\mu_0$  and  $\mu_1$  as marginals:

$$(2.1) \quad \Pi(\mu_0, \mu_1) := \{\gamma \in \mathcal{M}_+^1(\overline{\Omega} \times \overline{\Omega}) : \pi_0\#\gamma = \mu_0, \pi_1\#\gamma = \mu_1\},$$

where  $(\pi_0(x, y), \pi_1(x, y)) := (x, y)$  stand for the canonical projections ( $x$  and  $y$  in  $\overline{\Omega}$ ).

Introducing congestion naturally leads us to consider spaces of paths, lengths of such paths, and sets of probability measures on sets of paths. From now on, we shall denote the following:

- $C := W^{1,\infty}([0, 1], \overline{\Omega})$ , viewed as a subset of  $C^0([0, 1], \mathbb{R}^2)$ , i.e., equipped with the uniform topology;
- $C^{x,y} := \{\sigma \in C : \sigma(0) = x, \sigma(1) = y\}$  ( $x, y$  in  $\overline{\Omega}$ );
- $l(\sigma) := \int_0^1 |\dot{\sigma}(t)| dt$ , the length of  $\sigma \in C$ ;
- for  $\sigma \in C$ ,  $\tilde{\sigma}$  denotes the constant speed reparameterization of  $\sigma$  belonging to  $C$ , and hence  $|\dot{\tilde{\sigma}}(t)| = l(\sigma) = l(\tilde{\sigma})$  for a.e.  $t \in [0, 1]$ ;
- $\tilde{C} := \{\sigma \in C : |\dot{\sigma}| \text{ is constant}\} = \{\tilde{\sigma}, \sigma \in C\}$ ;
- slightly abusing notation, we will denote  $Q \in \mathcal{M}_+^1(C)$ , whenever  $Q \in \mathcal{M}_+^1(C^0([0, 1], \mathbb{R}^2))$  and  $Q(C) = 1$ ;
- for  $Q \in \mathcal{M}_+^1(C)$ , we define  $\tilde{Q} \in \mathcal{M}_+^1(\tilde{C})$  as the push forward of  $Q$  through the map  $\sigma \mapsto \tilde{\sigma}$ ;
- for  $\varphi \in C^0(\overline{\Omega}, \mathbb{R})$  and  $\sigma \in C$ , we define

$$L_\varphi(\sigma) := \int_0^1 \varphi(\sigma(t)) |\dot{\sigma}(t)| dt = l(\sigma) \int_0^1 \varphi(\tilde{\sigma}(t)) dt;$$

- $e_0(\sigma) := \sigma(0)$ ,  $e_1(\sigma) := \sigma(1)$  for all  $\sigma \in C^0([0, 1], \mathbb{R}^2)$ .

**2.2. Traffic congestion modeling.** The classical Monge–Kantorovich optimal transportation problem for a given cost function  $c \in C^0(\overline{\Omega} \times \overline{\Omega}, \mathbb{R})$  is

$$(2.2) \quad \inf \left\{ \int_{\overline{\Omega} \times \overline{\Omega}} c(x, y) d\gamma(x, y) : \gamma \in \Pi(\mu_0, \mu_1) \right\}.$$

Note that, in the linear problem (2.2), the cost of transporting one unit of mass from  $x$  to  $y$ ,  $c(x, y)$ , is given and does not depend on the path(s) followed by the mass from  $x$  to  $y$ . In order to take into account congestion effects, we explicitly introduce probabilities over  $C^{x,y}$  as part of the optimization problem. More precisely, the overall transportation cost will depend not only on the transportation plan  $\gamma \in \Pi(\mu_0, \mu_1)$  but also on the way travelers commuting from  $x$  to  $y$  use the different possible paths  $\sigma \in C^{x,y}$ . In what follows, the way commuters from  $x$  to  $y$  are split according to the different paths will be given by a probability measure  $p^{x,y}$  on  $C^{x,y}$ . Put differently,  $p^{x,y}(\Sigma)$  is the proportion of travelers from  $x$  to  $y$  using a path  $\sigma \in \Sigma \subset C^{x,y}$ . This naturally leads to the following definition.

**DEFINITION 2.1.** A transportation strategy consists of a pair  $(\gamma, p)$  with  $\gamma \in \Pi(\mu_0, \mu_1)$  and where  $p = (p^{x,y})_{(x,y) \in \overline{\Omega} \times \overline{\Omega}}$  is a Borel family of probability measures on  $C$  (i.e.,  $(x, y) \rightarrow \int_C F(\sigma) dp^{x,y}(\sigma)$  is Borel for every bounded Borel function  $F: C \rightarrow \mathbb{R}$ ) such that  $p^{x,y}(C^{x,y}) = 1$  for  $\gamma$ -a.e.  $(x, y) \in \overline{\Omega} \times \overline{\Omega}$ .

Thanks to Lemma 2.7 proved below, for every  $\varphi \in C^0(\overline{\Omega}, \mathbb{R}_+)$ , the map  $\sigma \rightarrow L_\varphi(\sigma)$  is l.s.c., and hence Borel on  $C$  (equipped with the uniform topology). For  $Q \in \mathcal{M}_+^1(C)$  one can therefore define  $\int_C L_\varphi(\sigma) dQ(\sigma)$  and this integral is finite for every  $\varphi \in C^0(\overline{\Omega}, \mathbb{R}_+)$  whenever  $\int_C l(\sigma) dQ(\sigma) < +\infty$ , i.e., the average for the probability  $Q$  length is finite. Hence, there results, from the use of a transportation strategy  $(\gamma, p)$ ,

an overall *traffic intensity*  $I_{\gamma,p} \in \mathcal{M}_+(\overline{\Omega})$  defined by

$$\begin{aligned} \int_{\overline{\Omega}} \varphi(x) dI_{\gamma,p}(x) &:= \int_{\overline{\Omega} \times \overline{\Omega}} \left( \int_{C^{x,y}} \left( \int_0^1 \varphi(\sigma(t)) |\dot{\sigma}(t)| dt \right) dp^{x,y}(\sigma) \right) d\gamma(x, y) \\ &= \int_{\overline{\Omega} \times \overline{\Omega}} \left( \int_{C^{x,y}} L_{\varphi}(\sigma) dp^{x,y}(\sigma) \right) d\gamma(x, y) \quad \forall \varphi \in C^0(\overline{\Omega}, \mathbb{R}_+) \end{aligned} \quad (2.3)$$

and an overall probability over paths  $Q_{\gamma,p} \in \mathcal{M}_+^1(C)$  given by  $Q_{\gamma,p} = p^{x,y} \otimes \gamma$ , i.e.,

$$(2.4) \quad \int_C F(\sigma) dQ_{\gamma,p}(\sigma) = \int_{\overline{\Omega} \times \overline{\Omega}} \left( \int_{C^{x,y}} F(\sigma) dp^{x,y}(\sigma) \right) d\gamma(x, y) \quad \forall F \in C^0(C, \mathbb{R}).$$

One could consider the probability  $Q_{\gamma,p}$  as if it represented the total number of travelers that use a path  $\sigma \in \Sigma$  given the global transportation strategy  $(\gamma, p)$ .

Let us remark that if we set  $Q := Q_{\gamma,p} \in \mathcal{M}_+^1(C)$ , then  $I_{\gamma,p}$  depends only on  $Q$  and can be written as  $I_{\gamma,p} = i_Q \in \mathcal{M}_+(\overline{\Omega})$ , where  $i_Q$  is defined for every  $Q \in \mathcal{M}_+^1(C)$  by

$$(2.5) \quad \int_{\overline{\Omega}} \varphi(x) di_Q(x) = \int_C L_{\varphi}(\sigma) dQ(\sigma) \quad \forall \varphi \in C^0(\overline{\Omega}, \mathbb{R}_+).$$

Let us also remark that since  $L_{\varphi}(\sigma) = L_{\varphi}(\tilde{\sigma})$ , one has  $i_Q = i_{\tilde{Q}}$ , for all  $Q \in \mathcal{M}_+^1(C)$ . Finally, let us note that the total mass of  $i_Q$  is the average length with respect to  $Q$ :

$$(2.6) \quad i_Q(\overline{\Omega}) = \int_C l(\sigma) dQ(\sigma).$$

If the probability  $Q$  is concentrated on injective curves, one can also express the measure  $i_Q$  through  $\mathcal{H}^1$ -integrals as explained in the next remark.

*Remark 2.2.* If a curve  $\sigma$  is injective, then one could also write  $l(\sigma) = \mathcal{H}^1(\sigma([0, 1]))$  and  $L_{\varphi}(\sigma) = \int_{\sigma([0,1])} \varphi d\mathcal{H}^1$ . Moreover, if for  $\gamma$ -a.e.  $(x, y)$  the probability  $p^{x,y}$  is concentrated on the set of injective curves from  $x$  to  $y$ , one could also define the measure  $I_{\gamma,p}$  by replacing the integral with respect to  $|\dot{\sigma}(t)|dt$  in (2.3) with an integral in  $d\mathcal{H}^1$ . Notice, moreover, that for every Borel subset  $A \subset \overline{\Omega}$  one would have

$$I_{\gamma,p}(A) = \int_{\overline{\Omega} \times \overline{\Omega}} \left( \int_{C^{x,y}} \mathcal{H}^1(A \cap \sigma) dp^{x,y}(\sigma) \right) d\gamma(x, y) = \int_C \mathcal{H}^1(A \cap \sigma) dQ_{\gamma,p}(\sigma).$$

If we imagine that, for each  $\sigma \in C^{x,y}$ , the mass of travelers commuting on  $\sigma$  is uniformly distributed on  $\sigma$ , this means that  $I_{\gamma,p}(A)$  represents the cumulative traffic through the region  $A$ . The same formula remains true, under no injectivity assumption, if we replace  $\mathcal{H}^1(A \cap \sigma)$  with  $L_{\mathbf{1}_A}(\sigma)$ , and in this case the cumulative traffic takes into account the number of times a path  $\sigma$  passes through the points of  $A$ .

In what follows, it will be convenient to formulate our optimization problem in terms of  $Q = Q_{\gamma,p}$  rather than in the transportation strategy  $(\gamma, p)$ . To that end, we shall use the following.

LEMMA 2.3. *Let us define*

$$\mathcal{Q}(\mu_0, \mu_1) := \{Q_{\gamma,p} : (\gamma, p) \text{ transportation strategy}\};$$

*then one has*

$$\mathcal{Q}(\mu_0, \mu_1) = \{Q \in \mathcal{M}_+^1(C) : e_0 \# Q = \mu_0, e_1 \# Q = \mu_1\}.$$

*Proof.* If  $(\gamma, p)$  is a transportation strategy, then  $e_0\#Q_{\gamma,p} = \pi_0\#\gamma = \mu_0$  and  $e_1\#Q_{\gamma,p} = \pi_1\#\gamma = \mu_1$ . Now let  $Q \in \mathcal{M}_+^1(C)$  be such that  $e_0\#Q = \mu_0$ ,  $e_1\#Q = \mu_1$ . If we define  $\gamma := (e_0, e_1)\#Q$ , we have  $\gamma \in \Pi(\mu_0, \mu_1)$ . It then follows from the disintegration theorem (see [9]) that there exists  $p = (p^{x,y})_{(x,y) \in \bar{\Omega} \times \bar{\Omega}}$ , a Borel family of probability measures on  $C$ , such that  $p^{x,y}(C^{x,y}) = 1$  for  $\gamma$ -a.e.  $(x, y) \in \bar{\Omega} \times \bar{\Omega}$  and  $Q = p^{x,y} \otimes \gamma$ . Hence  $Q = Q_{\gamma,p}$  for a transportation strategy  $(\gamma, p)$ .  $\square$

At this point, a natural way to model traffic congestion is, for a given transportation strategy  $(\gamma, p)$ , to consider that the transportation cost per unit of mass between  $x$  and  $y$  is given by

$$(2.7) \quad c_{\gamma,p}(x, y) = \int_{C^{x,y}} L_{G_{I_{\gamma,p}}}(\sigma) dp^{x,y}(\sigma),$$

where  $G_{I_{\gamma,p}}$  is a nonnegative function which depends (in a way that will be specified later) on the traffic intensity  $I_{\gamma,p}$ . The optimal transportation with traffic congestion then takes the form (to be compared with the usual Monge–Kantorovich problem (2.2))

$$(2.8) \quad \inf \left\{ \int_{\bar{\Omega} \times \bar{\Omega}} c_{\gamma,p}(x, y) d\gamma(x, y) : (\gamma, p) \text{ transportation strategy} \right\}.$$

Setting  $Q = Q_{\gamma,p}$  and using formally (2.5), we see that the total transportation cost in (2.8) can be rewritten as

$$\int_{\bar{\Omega} \times \bar{\Omega}} c_{\gamma,p}(x, y) d\gamma(x, y) = \int_C L_{G_{i_Q}}(\sigma) dQ(\sigma) = \int_{\bar{\Omega}} G_{i_Q}(x) di_Q(x).$$

Hence using Lemma 2.3, we can reformulate (2.8) in terms of  $Q$  only:

$$(2.9) \quad \inf \left\{ \int_{\bar{\Omega}} G_{i_Q}(x) di_Q(x) : Q \in \mathcal{Q}(\mu_0, \mu_1) \right\}.$$

Note that in the definition (2.7), it is required that  $G_{I_{\gamma,p}}$  is continuous (or at least l.s.c.), whereas the form (2.9) allows for more general forms of congestion through  $i \mapsto G_i$ . From now on, we assume that  $G$  has the following local form:

$$(2.10) \quad G_i(x) = g \left( \frac{di}{d\mathcal{L}^2}(x) \right),$$

where  $\frac{di}{d\mathcal{L}^2}$  is the Radon–Nikodym derivative of  $i$  with respect to the Lebesgue measure and  $g$  is a nondecreasing function  $\mathbb{R}_+ \rightarrow \mathbb{R}_+$  such that the function  $H$  defined by  $H(z) = zg(z)$  for all  $z \in \mathbb{R}_+$  is convex and superlinear (i.e.,  $\lim_{z \rightarrow +\infty} g(z) = +\infty$ ).

The optimization problem we shall study now then reads as

$$(2.11) \quad \inf_{Q \in \mathcal{Q}(\mu_0, \mu_1)} \mathcal{H}(i_Q), \text{ where } \mathcal{H}(i) = \begin{cases} \int_{\bar{\Omega}} H(i(x)) dx & \text{if } i \ll \mathcal{L}^2, \\ +\infty & \text{otherwise.} \end{cases}$$

In what follows, we shall say that a transportation strategy  $(\gamma, p)$  is optimal if  $Q_{\gamma,p}$  solves (2.11).

*Remark 2.4.* It will be clear in what follows that the probability  $Q_{\gamma,p}$  associated to an optimal transportation strategy  $(\gamma, p)$  will be concentrated on injective curves, so that the interpretation in terms of  $\mathcal{H}^1$ -integrals (see Remark 2.2) may apply.

**2.3. Existence of minimizers.** From now on, we make the following assumptions:

- $H$  is convex and nondecreasing on  $\mathbb{R}_+$  with  $H(0) = 0$ ;
- there exists  $q > 1$ , and positive constants  $a$  and  $b$  such that  $az^q \leq H(z) \leq b(z^q + 1)$  for all  $z \in \mathbb{R}_+$ ;
- $H$  is differentiable on  $\mathbb{R}_+$ , and there exists a positive constant  $c$  such that  $0 \leq H'(z) \leq c(z^{q-1} + 1)$ , for all  $z \in \mathbb{R}_+$ ;
- the set

$$(2.12) \quad \mathcal{Q}^q(\mu_0, \mu_1) := \{Q \in \mathcal{Q}(\mu_0, \mu_1) : i_Q \in L^q\}$$

is nonempty (in the definition of  $\mathcal{Q}^q(\mu_0, \mu_1)$ , we intend, of course, both  $i_Q \ll \mathcal{L}^2$  and  $\frac{di_Q}{d\mathcal{L}^2} \in L^q$ ).

These assumptions enable us to simply rewrite (2.11) as

$$(2.13) \quad \inf_{Q \in \mathcal{Q}^q(\mu_0, \mu_1)} \int_{\Omega} H(i_Q(x)) dx.$$

*Remark 2.5.* Let us discuss the assumption that  $\mathcal{Q}^q(\mu_0, \mu_1) \neq \emptyset$  which, at first glance, may seem difficult to check. In order to have the existence of a  $Q \in \mathcal{Q}(\mu_0, \mu_1)$  such that  $i_Q \in L^q$  it is sufficient that  $\mu_0$  and  $\mu_1$  are in  $L^q$ . This result, which is not obvious, follows from the regularity results of De Pascale and Pratelli (see [10] and [11]) who proved that  $L^q$  regularity of  $\mu_0$  and  $\mu_1$  implies that for  $\gamma$  solving the Monge–Kantorovich problem (2.2) with  $c(x, y) = |x - y|$  and  $p^{x,y} = \delta_{[x,y]}$  (the Dirac mass at the segment  $[x, y]$ ) for every  $x$  and  $y$  the corresponding traffic density  $I_{\gamma,p}$  is  $L^q$ .

*Remark 2.6.* It is not necessary, however, that  $\mu_0$  and  $\mu_1$  are absolutely continuous for the assumption to be satisfied. Indeed, in the discrete case, i.e., when  $\mu_0$  and  $\mu_1$  have finite support, one can easily prove that  $\mathcal{Q}^q(\mu_0, \mu_1) \neq \emptyset$  for every  $q \in (1, 2)$ . Finally, let us consider, for instance, the case where  $\overline{\Omega} = [0, 1]^2$  and  $\mu_0$  and  $\mu_1$  are, respectively, the one-dimensional Hausdorff measures of the segments  $[(0, 0), (0, 1)]$  and  $[(1, 0), (1, 1)]$ . If we define  $\gamma := (\text{id}, \text{id} + (1, 0))\# \mu_0$  and  $p^{x,y} = \delta_{[x,y]}$ , then a straightforward computation shows that  $I_{\gamma,p}$  is uniform on  $[0, 1]^2$ .

Under the assumptions above, we are going to prove that (2.13) admits a solution. The proof of existence involves some preliminary lemmas.

**LEMMA 2.7.** *For any  $\varphi \in C^0(\overline{\Omega}, \mathbb{R}_+)$ ,  $L_\varphi$  is l.s.c. on  $C$  for the uniform topology; indeed, for any  $\sigma \in C$ , one has*

$$(2.14) \quad L_\varphi(\sigma) = \sup \left\{ \sum_{i=1}^n \left( \inf_{[t_i, t_{i+1}]} (\varphi \circ \sigma) \right) |\sigma(t_{i+1}) - \sigma(t_i)| : \right. \\ \left. ([t_i, t_{i+1}])_i \text{ is a subdivision of } [0, 1] \right\}.$$

*Proof.* For any subdivision  $([t_i, t_{i+1}])_{i=1, \dots, n}$ , we have

$$\begin{aligned} L_\varphi(\sigma) &= \sum_{i=1}^n \int_{t_i}^{t_{i+1}} \varphi(\sigma(t)) |\dot{\sigma}(t)| \, dt \\ &\geq \sum_{i=1}^n \inf_{[t_i, t_{i+1}]} (\varphi \circ \sigma) \int_{t_i}^{t_{i+1}} |\dot{\sigma}(t)| \, dt \\ &\geq \sum_{i=1}^n \inf_{[t_i, t_{i+1}]} (\varphi \circ \sigma) |\sigma(t_{i+1}) - \sigma(t_i)|. \end{aligned}$$

Taking the supremum over all such divisions, we get

$$L_\varphi(\sigma) \geq \sup \left\{ \sum_{i=1}^n \inf_{[t_i, t_{i+1}]} (\varphi \circ \sigma) |\sigma(t_{i+1}) - \sigma(t_i)| : ([t_i, t_{i+1}])_i \text{ is a subdivision of } [0, 1] \right\}.$$

Let us prove the converse inequality. Let  $\varepsilon > 0$ ; since  $\varphi \circ \sigma$  is uniformly continuous, there is a  $\delta > 0$  such that

$$\forall t, t' \in [0, 1]^2, \quad (|t - t'| \leq \delta \Rightarrow |\varphi(\sigma(t)) - \varphi(\sigma(t'))| \leq \varepsilon).$$

For any subdivision  $([t_i, t_{i+1}])_{i=1, \dots, n}$  such that  $|t_i - t_{i+1}| \leq \delta$  for all  $i$ , we have

$$\begin{aligned} L_\varphi(\sigma) &\leq \sum_{i=1}^n \left( \inf_{[t_i, t_{i+1}]} (\varphi \circ \sigma) + \varepsilon \right) \int_{t_i}^{t_{i+1}} |\dot{\sigma}(t)| \, dt \\ &= \sum_{i=1}^n \left( \inf_{[t_i, t_{i+1}]} (\varphi \circ \sigma) + \varepsilon \right) \sup \left\{ \sum_j |\sigma(\tau_j) - \sigma(\tau_{j+1})| : \right. \\ &\quad \left. ([\tau_j, \tau_{j+1}])_j \text{ is a subdivision of } [t_i, t_{i+1}] \right\} \\ &\leq \sup \left\{ \sum_i \sum_j \left( \inf_{[\tau_j, \tau_{j+1}]} (\varphi \circ \sigma) + \varepsilon \right) |\sigma(\tau_j) - \sigma(\tau_{j+1})| : \right. \\ &\quad \left. ([\tau_j, \tau_{j+1}])_j \text{ is a subdivision of } [t_i, t_{i+1}] \right\} \\ &= \sup \left\{ \sum_{i=1}^n \left( \inf_{t \in [t_i, t_{i+1}]} (\varphi \circ \sigma) + \varepsilon \right) |\sigma(t_{i+1}) - \sigma(t_i)| : \right. \\ &\quad \left. ([t_i, t_{i+1}])_i \text{ is a subdivision of } [0, 1] \right\}. \end{aligned}$$

As this last inequality is true for any  $\varepsilon > 0$ , we get (2.14). The lower semicontinuity is then obvious since, by (2.14),  $L_\varphi$  is the supremum of the family of l.s.c. functions on  $C^0([0, 1], \bar{\Omega})$ .  $\square$

**LEMMA 2.8.** *Let  $(Q_n)_n \in \mathcal{M}_+^1(C^0([0, 1], \mathbb{R}^2))^{\mathbb{N}}$  such that  $Q_n(C) = 1$  for all  $n$ , and let there exist a constant  $M > 0$  such that*

$$\sup_n \int_C l(\sigma) \, dQ_n(\sigma) \leq M.$$

Then the sequence  $(\tilde{Q}_n)_n$  is tight and admits a subsequence that converges weakly\* to a probability  $Q$  such that  $Q(C) = 1$ .

*Proof.* The tightness of  $(\tilde{Q}_n)_n$  easily follows from the inequality

$$\begin{aligned} \tilde{Q}_n \left( \{ \sigma \in \tilde{C} : |\dot{\sigma}| > K \} \right) &= Q_n \left( \{ \sigma \in C : l(\sigma) > K \} \right) \\ (2.15) \qquad \qquad \qquad &\leq \frac{1}{K} \int_C l(\sigma) dQ_n(\sigma). \end{aligned}$$

By the Prokhorov theorem, we may therefore assume, passing to a subsequence if necessary, that  $(\tilde{Q}_n)_n$  converges weakly\* to  $Q \in \mathcal{M}_+^1(C^0([0, 1], \mathbb{R}^2))$ . It remains to show that  $Q(C) = 1$ . For  $K > 0$  let us define  $C_K := \{ \sigma \in C : |\dot{\sigma}| \leq K \}$ ; then inequality (2.15) and the fact that the measures  $\tilde{Q}_n$  are concentrated on  $\tilde{C}$  yield

$$\sup_n \tilde{Q}_n \frac{C}{C_K} = \sup_n \tilde{Q}_n \frac{\tilde{C}}{C_K} \leq \frac{M}{K},$$

for every  $K > 0$ , which implies

$$\begin{aligned} 1 = \limsup_n \tilde{Q}_n(C) &\leq \limsup_n \tilde{Q}_n(C_K) + \limsup_n \tilde{Q}_n \frac{C}{C_K} \\ &\leq Q(C_K) + \frac{M}{K}. \end{aligned}$$

Letting  $K$  tend to  $\infty$ , we then get  $Q(C) = \sup_K Q(C_K) = 1$ .  $\square$

LEMMA 2.9. *Let  $(Q_n)_n$  be a sequence in  $\mathcal{M}_+^1(C)$  that converges weakly\* to some  $Q \in \mathcal{M}_+^1(C)$ . If there exists  $i \in M_+(\bar{\Omega})$  such that  $i_{Q_n}$  converges weakly\* to  $i$  in  $\mathcal{M}_+(\bar{\Omega})$ , then we have  $i_Q \leq i$ .*

*Proof.* Let  $\varphi \in C^0(\bar{\Omega}, \mathbb{R}_+)$ . We first have

$$\int_{\bar{\Omega}} \varphi di = \lim_n \int_{\bar{\Omega}} \varphi di_{Q_n} = \lim_n \int_C L_\varphi dQ_n;$$

it easily follows from Lemma 2.7 that  $Q \mapsto \int_C L_\varphi dQ$  is l.s.c. for the weak\* topology of  $\mathcal{M}_+^1(C)$ . We then have

$$\int_{\bar{\Omega}} \varphi di \geq \int_C L_\varphi dQ = \int_{\bar{\Omega}} \varphi di_Q. \quad \square$$

Now, we are in position to prove the following.

THEOREM 2.10. *The minimization problem (2.13) admits a solution.*

*Proof.* Our assumptions imply that the value of (2.13) is finite. Let  $(Q_n)_n$  be some minimizing sequence of (2.13). From the identity  $i_Q = i_{\tilde{Q}}$ , we may assume  $Q_n = \tilde{Q}_n$  for all  $n$ . We deduce from our growth condition on  $H$  that  $(i_{Q_n})_n$  is bounded in  $L^q$ . On the one hand, extracting a subsequence if necessary, we may therefore assume that  $(i_{Q_n})_n$  converges weakly in  $L^q$  to some  $i$ . On the other hand, since  $i_{Q_n}$  is bounded in  $L^q$  and hence in  $L^1$ , we have

$$\sup_n \int_C l(\sigma) dQ_n(\sigma) = \sup_n \int_{\Omega} i_{Q_n} < +\infty.$$

Moreover,  $Q_n = \tilde{Q}_n$  and we deduce from Lemma 2.8 that (up to some subsequence)  $(Q_n)_n$  weakly\* converges to some  $Q$  in  $\mathcal{M}_+^1(C)$ . Since  $Q(\mu_0, \mu_1)$  is obviously weakly\*

closed, we have  $Q \in \mathcal{Q}(\mu_0, \mu_1)$ , and Lemma 2.9 implies that  $i_Q \leq i$  (consequently, to this inequality  $i_Q$  is absolutely continuous). From the monotonicity and convexity of  $H$  we then have

$$\int_{\Omega} H(i_Q(x)) dx \leq \int_{\Omega} H(i(x)) dx \leq \liminf_n \int_{\Omega} H(i_{Q_n}(x)) dx,$$

which proves that  $Q$  solves (2.13).  $\square$

Let us remark that if  $H$  is furthermore assumed to be strictly convex, then if  $Q_1$  and  $Q_2$  solve (2.13), then  $i_{Q_1} = i_{Q_2}$  so that the optimal traffic intensity is unique (of course, this does not imply, in general, that  $Q_1 = Q_2$  or that the corresponding optimal transportation strategy is unique).

**3. Characterization of the minimizers.** In what follows, we shall denote by  $q^*$  the conjugate exponent of  $q$ , given by  $q^* = q/(q-1)$ .

**3.1. Optimality conditions.** The variational inequalities characterizing solutions of the convex problem (2.13) can be expressed as follows.

PROPOSITION 3.1.  $\bar{Q} \in \mathcal{Q}^q(\mu_0, \mu_1)$  solves (2.13) if and only if

$$(3.1) \quad \int_{\Omega} \bar{\xi} i_{\bar{Q}} = \inf \left\{ \int_{\Omega} \bar{\xi} i_Q : Q \in \mathcal{Q}^q(\mu_0, \mu_1) \right\} \text{ with } \bar{\xi} := H'(i_{\bar{Q}}) \in L^{q^*}.$$

*Proof.* Assume that  $\bar{Q}$  solves (2.13); then for every  $Q \in \mathcal{Q}^q(\mu_0, \mu_1)$ , one has

$$\begin{aligned} 0 &\leq \lim_{\varepsilon \rightarrow 0^+} \frac{1}{\varepsilon} [\mathcal{H}(i_{\bar{Q} + \varepsilon(Q - \bar{Q})}) - \mathcal{H}(i_{\bar{Q}})] = \lim_{\varepsilon \rightarrow 0^+} \frac{1}{\varepsilon} [\mathcal{H}(i_{\bar{Q}} + \varepsilon(i_Q - i_{\bar{Q}})) - \mathcal{H}(i_{\bar{Q}})] \\ &= \int_{\Omega} H'(i_{\bar{Q}})(i_Q - i_{\bar{Q}}) = \int_{\Omega} \bar{\xi}(i_Q - i_{\bar{Q}}), \end{aligned}$$

which proves (3.1). Conversely, if  $\bar{Q} \in \mathcal{Q}^q(\mu_0, \mu_1)$  satisfies (3.1), then by convexity of  $H$ , for every  $Q \in \mathcal{Q}^q(\mu_0, \mu_1)$ , one has

$$\mathcal{H}(i_Q) - \mathcal{H}(i_{\bar{Q}}) \geq \int_{\Omega} \bar{\xi}(i_Q - i_{\bar{Q}}) \geq 0. \quad \square$$

Sections 3.2 and 3.3 will be devoted to investigating the precise meaning of (3.1). Before going further, let us do some formal manipulations to give a formal interpretation of (3.1) in terms of optimal transportation strategy. Let us assume that  $\bar{Q}$  solves (2.13) and let us write  $\bar{Q} = Q_{\bar{\gamma}, \bar{p}}$  for some (optimal) transportation strategy  $(\bar{\gamma}, \bar{p})$  and define  $\bar{\xi} := H'(i_{\bar{Q}})$ ; then (3.1) formally can be rewritten as

$$\begin{aligned} \int_{\Omega} \bar{\xi} i_{\bar{Q}} &= \int_C L_{\bar{\xi}}(\sigma) d\bar{Q}(\sigma) \\ &= \int_{\bar{\Omega} \times \bar{\Omega}} \left( \int_{C^{x,y}} L_{\bar{\xi}}(\sigma) d\bar{p}^{x,y}(\sigma) \right) d\bar{\gamma}(x, y) \\ &= \inf_{(\gamma, p) \text{ transp. strategy}} \int_{\bar{\Omega} \times \bar{\Omega}} \left( \int_{C^{x,y}} L_{\bar{\xi}}(\sigma) dp^{x,y}(\sigma) \right) d\gamma(x, y) \\ &= \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \left( \inf_{p \in \mathcal{M}_+^1(C^{x,y})} \int_{C^{x,y}} L_{\bar{\xi}}(\sigma) dp(\sigma) \right) d\gamma(x, y) \\ &= \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \left( \inf_{\sigma \in C^{x,y}} L_{\bar{\xi}}(\sigma) \right) d\gamma(x, y), \end{aligned}$$



defining (again formally) the transportation cost

$$c_{\bar{\xi}}(x, y) = \inf_{\sigma \in C^{x, y}} L_{\bar{\xi}}(\sigma).$$

We then first have

$$\int_{\bar{\Omega} \times \bar{\Omega}} c_{\bar{\xi}}(x, y) d\bar{\gamma}(x, y) \leq \int_C L_{\bar{\xi}} d\bar{Q} = \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} c_{\bar{\xi}}(x, y) d\gamma(x, y)$$

so that  $\bar{\gamma}$  solves the Monge–Kantorovich problem:

$$\inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} c_{\bar{\xi}}(x, y) d\gamma(x, y).$$

Second,

$$\begin{aligned} \int_C L_{\bar{\xi}}(\sigma) d\bar{Q}(\sigma) &= \int_{\bar{\Omega} \times \bar{\Omega}} c_{\bar{\xi}}(x, y) d\bar{\gamma}(x, y) \\ &= \int_C c_{\bar{\xi}}(\sigma(0), \sigma(1)) d\bar{Q}(\sigma), \end{aligned}$$

and since  $L_{\bar{\xi}}(\sigma) \geq c_{\bar{\xi}}(\sigma(0), \sigma(1))$ , we get

$$L_{\bar{\xi}}(\sigma) = c_{\bar{\xi}}(\sigma(0), \sigma(1)) \quad \text{for } \bar{Q}\text{-a.e. } \sigma,$$

or, in an equivalent way, for  $\bar{\gamma}$ -a.e.  $(x, y)$  one has

$$L_{\bar{\xi}}(\sigma) = c_{\bar{\xi}}(x, y) \quad \text{for } \bar{p}^{x, y}\text{-a.e. } \sigma.$$

Since  $\bar{\xi}$  is only  $L^{q^*}$ ,  $L_{\bar{\xi}}$  and  $c_{\bar{\xi}}$  are not well defined and the previous arguments are purely formal. In section 3.2, we will extend the definition of  $c_{\xi}$  to the case where  $\xi$  is only  $L^{q^*}$  under the additional assumption  $q < 2$ . This will enable us to make the formal argument above rigorous and to characterize optimal transportation strategies in section 3.3.

**3.2. The transportation cost  $\bar{c}_{\xi}$  when  $\xi$  is  $L^{q^*}$ .** For a nonnegative function  $\xi \in C^0(\bar{\Omega}, \mathbb{R}_+)$  we define

$$c_{\xi}(x, y) = \inf\{L_{\xi}(\sigma) : \sigma \in C^{x, y}\}.$$

**PROPOSITION 3.2.** *Assume that  $q < 2$  and define  $\alpha := 1 - 2/q^*$ ; then there exists a nonnegative constant  $C$  such that for every  $\xi \in C^0(\bar{\Omega}, \mathbb{R}_+)$  and every  $(x_1, y_1, x_2, y_2) \in \Omega^4$ , one has*

$$(3.2) \quad |c_{\xi}(x_1, y_1) - c_{\xi}(x_2, y_2)| \leq C \|\xi\|_{L^{q^*}(\Omega)} (|x_1 - x_2|^{\alpha} + |y_1 - y_2|^{\alpha}).$$

Consequently, if  $(\xi_n)_n \in C^0(\bar{\Omega}, \mathbb{R}_+)^{\mathbb{N}}$  is bounded in  $L^{q^*}$ , then  $(c_{\xi_n})_n$  admits a subsequence that converges in  $C^0(\bar{\Omega} \times \bar{\Omega}, \mathbb{R}_+)$ .

*Proof.* Let  $\xi \in C^0(\bar{\Omega}, \mathbb{R}_+)$  and  $x, y \in \Omega^2$ . For  $k > 0$  let  $\sigma_k \in C^{x, y}$  be such that

$$\int_0^1 \xi(\sigma_k(t)) |\dot{\sigma}_k(t)| dt \leq c_{\xi}(x, y) + \frac{1}{k}.$$

Then for all  $\varepsilon > 0$  and  $z$  such that  $y + \varepsilon z \in \Omega$  and  $t_0 \in (0, 1)$  we consider the following element of  $C^{x, y + \varepsilon z}$ :

$$\sigma_{k, t_0}(t) := \begin{cases} \sigma_k\left(\frac{t}{t_0}\right) & \text{if } t \in [0, t_0], \\ y + \left(\frac{t - t_0}{1 - t_0}\right) \varepsilon z & \text{if } t \in [t_0, 1]. \end{cases}$$

We then have, for all  $k > 0$ ,

$$\begin{aligned} c_\xi(x, y + \varepsilon z) &\leq \int_0^1 \xi(\sigma_{k, t_0}(t)) |\dot{\sigma}_{k, t_0}(t)| dt \\ &= \int_0^1 \xi(\sigma_k(t)) |\dot{\sigma}_k(t)| dt + \int_0^1 \xi(y + t\varepsilon z) \varepsilon |z| dt \\ &\leq c_\xi(x, y) + \varepsilon |z| \int_0^1 \xi(y + t\varepsilon z) dt + \frac{1}{k}. \end{aligned}$$

Now we let  $k$  tend to  $\infty$  and we get

$$\frac{c_\xi(x, y + \varepsilon z) - c_\xi(x, y)}{\varepsilon} \leq |z| \int_0^1 \xi(y + t\varepsilon z) dt,$$

and by a similar argument,

$$\frac{c_\xi(x, y) - c_\xi(x, y + \varepsilon z)}{\varepsilon} \leq |z| \int_0^1 \xi(y + (1 - t)\varepsilon z) dt.$$

This implies that  $c_\xi(x, \cdot) \in W^{1, \infty}$  and

$$(3.3) \quad |\nabla_y c_\xi(x, \cdot)| \leq |\xi(\cdot)| \quad \forall x.$$

By symmetry we also have

$$(3.4) \quad |\nabla_x c_\xi(\cdot, y)| \leq |\xi(\cdot)| \quad \forall y.$$

Since  $q^* > 2$ , we deduce from (3.3), (3.4), and Morrey's theorem (see [6, Chapter IX]), that there is a constant  $C$  such that

$$\begin{aligned} |c_\xi(x, y_1) - c_\xi(x, y_2)| &\leq C \|\xi\|_{L^{q^*}} |y_1 - y_2|^\alpha \quad \forall x, y_1, y_2 \text{ in } \Omega, \\ |c_\xi(x_1, y) - c_\xi(x_2, y)| &\leq C \|\xi\|_{L^{q^*}} |x_1 - x_2|^\alpha \quad \forall x_1, x_2, y \text{ in } \Omega. \end{aligned}$$

This proves (3.2). The second claim in the proposition then follows from (3.2), the identity  $c_{\xi_n}(x, x) = 0$ , and Ascoli's theorem.  $\square$

From now on, we further assume that  $q < 2$ . For a nonnegative function  $\xi \in L^{q^*}(\Omega)$  we then define

$$\bar{c}_\xi(x, y) = \sup \{c(x, y) : c \in \mathcal{A}(\xi)\},$$

where

$$\mathcal{A}(\xi) = \left\{ \lim_n c_{\xi_n} \text{ in } C^0(\bar{\Omega} \times \bar{\Omega}) : (\xi_n)_n \in C^0(\bar{\Omega}), \xi_n \geq 0, \xi_n \rightarrow \xi \text{ in } L^{q^*} \right\}.$$

*Remark 3.3.* The definition of  $\bar{c}_\xi$  is unchanged if one replaces  $\xi_n \rightarrow \xi$  in  $L^{q^*}$  by  $\xi_n \rightharpoonup \xi$  in  $L^{q^*}$  in the definition of  $\mathcal{A}(\xi)$ . Indeed, if we do so, we obviously obtain a

function which is larger than  $\bar{c}_\xi$ . Now, let us assume that  $\xi_n \rightarrow \xi$  in  $L^q$ , and  $c_{\xi_n}$  converges to  $c$  in  $C^0(\bar{\Omega} \times \bar{\Omega})$ ; using Mazur's lemma there exists a sequence  $\eta_n$  which converges strongly to  $\xi$  and such that each  $\eta_n$  is in the convex hull of  $\{\xi_k, k \geq n\}$ . It is clear that for fixed  $x, y$ ,  $\xi \rightarrow c_\xi(x, y)$  is concave, and hence  $c(x, y) = \lim c_{\xi_n}(x, y) \leq \limsup c_{\eta_n}(x, y) \leq \bar{c}_\xi(x, y)$ .

When  $\xi$  is continuous, one has the following.

LEMMA 3.4. *If  $\xi$  is continuous and nonnegative, then  $\bar{c}_\xi = c_\xi$ .*

*Proof.* The inequality  $\bar{c}_\xi \geq c_\xi$  is obvious, as one can always choose the constant sequence  $\xi_n = \xi$  in the definition of  $\bar{c}_\xi$ . Let us show now the opposite inequality. Take  $x, y \in \Omega$ ,  $\varepsilon > 0$ , and  $\sigma \in C^{x,y}$  such that  $L_\xi(\sigma) < c_\xi(x, y) + 1/k$ . We can choose  $\sigma$  so that it is piecewise linear by density of this kind of curve and using the continuity of  $\xi$ . Let  $(S_i)_{i=0,\dots,m-1}$  be the segments which compose  $\sigma$  with  $S_i = x_i x_{i+1}$ ,  $x_0 = x$ , and  $x_m = y$ . Let us fix, moreover, a sequence  $\xi_n \rightarrow \xi$  such that  $c_{\xi_n} \rightarrow c$ . Now, we want to prove  $c \leq c_\xi$ . Fix a small number  $\delta > 0$ , and for any  $\alpha \in [0, \delta]$  let us define a curve  $\sigma^\alpha$  in the following way: let  $R$  be the clockwise  $90^\circ$  rotation in the plane; let  $x'_i(\alpha)$  and  $x''_i(\alpha)$  be the only points such that  $x_i x'_i(\alpha) = \alpha R e_i$  and  $x_{i+1} x''_i(\alpha) = \alpha R e_i$ , where  $e_i$  is the tangent unit vector to  $\sigma$  in the  $S_i$  part; define  $\sigma^\alpha$  by linking any point  $x'_i(\alpha)$  to  $x''_i(\alpha)$  by some segments  $S'_i(\alpha)$  and  $x''_i(\alpha)$  to  $x'_{i+1}(\alpha)$  by some arcs  $A_{i+1}(\alpha)$  with center  $x_{i+1}$  and radius  $\alpha$ . In this way we have  $\sigma^\alpha \in C^{x_\alpha, y_\alpha}$ , where  $x_\alpha = x'_0(\alpha)$  and  $y_\alpha = x''_m(\alpha)$ . Let  $R_i(\delta)$  be the rectangle whose vertices are the points  $x_i$ ,  $x'_i(\delta)$ ,  $x''_i(\delta)$ , and  $x_{i+1}$  and let  $B_i(\delta)$  be the circular sector centered at  $x_i$  and whose vertices are  $x''_{i-1}(\delta)$  and  $x'_i(\delta)$ .

If we compute  $\int_0^\delta L_{\xi_n}(\sigma_\alpha) d\alpha$ , it is not difficult to see that we get

$$\int_0^\delta L_{\xi_n}(\sigma_\alpha) d\alpha = \sum_{i=0}^{m-1} \left( \int_{R_i(\delta)} \xi_n d\mathcal{L}^2 \right) + \sum_{i=1}^{m-1} \left( \int_{B_i(\delta)} \xi_n d\mathcal{L}^2 \right).$$

Moreover, it holds that  $c_{\xi_n}(x^\alpha, y^\alpha) \leq L_{\xi_n}(\sigma^\alpha)$ , and hence we get

$$\int_0^\delta c_{\xi_n}(x^\alpha, y^\alpha) d\alpha \leq \sum_{i=0}^{m-1} \left( \int_{R_i(\delta)} \xi_n d\mathcal{L}^2 \right) + \sum_{i=1}^{m-1} \left( \int_{B_i(\delta)} \xi_n d\mathcal{L}^2 \right).$$

If we pass to the limit as  $n \rightarrow \infty$ , then by using the uniform convergence of  $c_{\xi_n}$  to  $c$  on the left-hand side and the  $L^q$  convergence of  $\xi_n$  to  $\xi$  on the right-hand side, we get

$$\int_0^\delta c(x^\alpha, y^\alpha) d\alpha \leq \sum_{i=0}^{m-1} \left( \int_{R_i(\delta)} \xi d\mathcal{L}^2 \right) + \sum_{i=1}^{m-1} \left( \int_{B_i(\delta)} \xi d\mathcal{L}^2 \right).$$

Then we divide by  $\delta$  and pass to the limit as  $\delta \rightarrow 0$ . Using the fact that  $c$  is continuous, we have

$$\lim_{\delta \rightarrow 0} \frac{1}{\delta} \int_0^\delta c(x^\alpha, y^\alpha) d\alpha = c(x, y).$$

On the other hand, we may notice that the areas of the sectors  $B_i(\delta)$  may be estimated by  $C\delta^2$ , and hence we have, for  $\delta \rightarrow 0$ ,

$$\frac{1}{\delta} \sum_{i=1}^{m-1} \left( \int_{B_i(\delta)} \xi d\mathcal{L}^2 \right) \leq mC|\xi|\delta \rightarrow 0.$$

On the contrary the integrals over  $R_i(\delta)$ , when divided by  $\delta$ , converge on the integrals on the segments  $S_i$ , which give exactly the integral over the curve  $\sigma$ , i.e.,  $L_\xi(\sigma)$ . We have, consequently,

$$\lim_{\delta \rightarrow 0} \frac{1}{\delta} \left( \sum_{i=0}^{m-1} \left( \int_{R_i(\delta)} \xi d\mathcal{L}^2 \right) + \sum_{i=1}^{m-1} \left( \int_{B_i(\delta)} \xi d\mathcal{L}^2 \right) \right) = L_\xi(\sigma) < c_\xi(x, y) + \frac{1}{k}.$$

This gives

$$c(x, y) < c_\xi(x, y) + \frac{1}{k},$$

and,  $k$  being arbitrary, we also get  $c \leq c_\xi$  and the thesis.  $\square$

LEMMA 3.5. *Let us assume that  $q < 2$  and let  $\xi$  be a nonnegative function belonging to  $L^{q^*}$ ; then there exists a sequence  $(\xi_n)_n \in C^0(\Omega)$ ,  $\xi_n \geq 0$ ,  $\xi_n \rightarrow \xi$  in  $L^{q^*}$ , such that  $c_{\xi_n}$  converges uniformly to  $\bar{c}_\xi$  on  $\Omega \times \Omega$ .*

*Proof.* It is easy to see that for every  $(x, y) \in \Omega^2$  there exists a sequence of nonnegative continuous functions  $(\xi_n)_n$  converging to  $\xi$  in  $L^{q^*}(\Omega)$  such that  $c_{\xi_n}$  converges in  $C^0$  and  $\bar{c}_\xi(x, y) = \lim_n c_{\xi_n}(x, y)$ . Let  $I$  be a finite set,  $(x_i, y_i) \in \Omega^2$  for all  $i \in I$ , and for every  $i$ , let  $(\xi_n^i)$  be a sequence of nonnegative continuous functions converging to  $\xi$  in  $L^{q^*}(\Omega)$  such that  $\bar{c}_\xi(x_i, y_i) = \lim_n c_{\xi_n^i}(x_i, y_i)$ . Let us set  $\xi_n := \max_{i \in I} \xi_n^i$ . We then have  $(\xi_n)_n$  converging to  $\xi$  in  $L^{q^*}(\Omega)$ , and

$$\bar{c}_\xi(x_i, y_i) \leq \liminf_n c_{\xi_n}(x_i, y_i) \leq \limsup_n c_{\xi_n}(x_i, y_i) \leq \bar{c}_\xi(x_i, y_i).$$

We thus have  $\bar{c}_\xi(x_i, y_i) = \lim_n c_{\xi_n}(x_i, y_i)$  for every  $i \in I$ . Now, let  $(x_i, y_i)_{i \in \mathbb{N}}$  be a dense sequence of points of  $\Omega^2$ . From what precedes, for every  $n$ , there exists a continuous nonnegative  $\xi_n$  such that

$$\|\xi_n - \xi\|_{L^{q^*}} \leq \frac{1}{n}, \quad |\bar{c}_\xi(x_k, y_k) - c_{\xi_n}(x_k, y_k)| \leq \frac{1}{n} \quad \forall k \leq n.$$

By the Hölder estimate of Proposition 3.2 and Ascoli's theorem, passing to a subsequence, if necessary, we may assume that  $c_{\xi_n}$  converges in  $C^0$  to some  $c$ . Since obviously  $c(x_k, y_k) = \bar{c}_\xi(x_k, y_k)$  for all  $k$ , we deduce  $c = \bar{c}_\xi$ , and the desired result follows.  $\square$

The next lemma enables us to extend  $L_\xi$  in some sense when  $\xi \geq 0$  is only  $L^{q^*}$ .

LEMMA 3.6. *Let us assume that  $q < 2$ . Let  $Q \in \mathcal{Q}^q(\mu_0, \mu_1)$ ,  $\xi$  be a nonnegative element of  $L^{q^*}$ , and  $(\xi_n)_n$  be a sequence of nonnegative continuous functions that converges to  $\xi$  in  $L^{q^*}$ . Then we have the following:*

- (i)  $(L_{\xi_n})_n$  converges strongly in  $L^1(C, Q)$  to some limit which is independent of the approximating sequence  $(\xi_n)_n$  and which will again be denoted  $L_\xi$ .
- (ii) The following equality holds:

$$(3.5) \quad \int_{\Omega} \xi(x) i_Q(x) \, dx = \int_C L_\xi(\sigma) \, dQ(\sigma).$$

- (iii) The following inequality holds for  $Q$ -a.e.  $\sigma \in C$ :

$$(3.6) \quad L_\xi(\sigma) \geq \bar{c}_\xi(\sigma(0), \sigma(1)).$$

*Proof.* For all  $n$  and  $m$  in  $\mathbb{N}$ , we have

$$\begin{aligned} \int_C |L_{\xi_n}(\sigma) - L_{\xi_m}(\sigma)| \, dQ(\sigma) &= \int_C \left| \int_0^1 (\xi_n(\sigma(t)) - \xi_m(\sigma(t))) |\dot{\sigma}(t)| \, dt \right| \, dQ(\sigma) \\ &\leq \int_{\Omega} |\xi_n(x) - \xi_m(x)| i_Q(x) \, dx \\ &\leq \|\xi_n - \xi_m\|_{L^{q^*}} \|i_Q\|_{L^q}. \end{aligned}$$

This implies that  $(L_{\xi_n})_n$  is a Cauchy sequence in  $L^1(C, Q)$ , and it is obvious from the previous inequality that its  $L^1(C, Q)$  limit does not depend on the approximating sequence  $(\xi_n)_n$ .

The proof of (ii) follows from (i):

$$\begin{aligned} \int_{\Omega} \xi(x) i_Q(x) \, dx &= \lim_n \int_{\Omega} \xi_n(x) i_Q(x) \, dx \\ &= \lim_n \int_C L_{\xi_n}(\sigma) \, dQ(\sigma) \\ &= \int_C L_{\xi}(\sigma) \, dQ(\sigma). \end{aligned}$$

To prove (iii) we choose an approximating sequence  $(\xi_n)_n$  as in Lemma 3.5 and pass to the limit in

$$L_{\xi_n}(\sigma) \geq c_{\xi_n}(\sigma(0), \sigma(1)). \quad \square$$

*Remark 3.7.* The condition  $q < 2$  may seem restrictive; however, if, for instance,  $\mu_0$  and  $\mu_1$  are discrete (and  $\mu_0 \neq \mu_1$ ), this condition is in fact sharp. Indeed, in this case, it can easily be proved that  $\mathcal{Q}^q(\mu_0, \mu_1) \neq \emptyset$  for  $q \in (1, 2)$  but  $\mathcal{Q}^2(\mu_0, \mu_1) = \emptyset$ . To see that  $\mathcal{Q}^2(\mu_0, \mu_1) = \emptyset$ , assume on the contrary that there is a  $Q \in \mathcal{Q}(\mu_0, \mu_1)$  such that  $i_Q \in L^2$ , and define the vector measure  $\vec{i}_Q$  by

$$\int_{\Omega} X(x) d\vec{i}_Q(x) = \int_C \left( \int_0^1 X(\sigma(t)) \cdot \dot{\sigma}(t) dt \right) dQ(\sigma) \quad \forall X \in C^0(\overline{\Omega}, \mathbb{R}^2).$$

It is easy to check that

$$\operatorname{div}(\vec{i}_Q) = \mu_0 - \mu_1 \quad \text{and} \quad \|\vec{i}_Q\|_{L^2} \leq \|i_Q\|_{L^2} < +\infty.$$

But, since  $\mu_0 - \mu_1 \notin H^{-1}(\Omega)$ , we get the desired contradiction (and in fact we have proved that  $\mathcal{Q}^2(\mu_0, \mu_1) = \emptyset$  as soon as  $\mu_0 - \mu_1 \notin H^{-1}(\Omega)$ ). In other words, if  $q \geq 2$ , the congestion effect is so strong that the total congested cost in (2.13) is always  $+\infty$  as soon as  $\mu_0 - \mu_1 \notin H^{-1}(\Omega)$ .

*Remark 3.8.* We have assumed throughout the paper that the ambient dimension is 2, and we have seen in this case that one can extend the definitions of  $c_{\xi}$  and  $L_{\xi}$  to the case  $\xi \in L^{q^*}$ ,  $\xi \geq 0$ , provided  $q < 2$ . In dimension  $d \geq 2$ , it is easy to see that the Hölder estimate of Proposition 3.2 still holds for  $q^* > d$ , i.e.,  $q < d/(d-1)$ . All the results of the paper in fact extend to this more general case.

**3.3. Characterization of optimal transport strategies.** In this section, our goal is to make the formal arguments of section 3.1 rigorous in order to characterize optimal transport strategies. This can be done under the additional assumption that

$H$  is strictly convex. First, we relate the optimality condition (3.1) to the Monge–Kantorovich problem with cost  $\bar{c}_{\bar{\xi}}$  as follows.

PROPOSITION 3.9. *Let us assume that  $q < 2$  and that  $H$  is strictly convex. If  $\bar{Q}$  solves (2.13) and  $\bar{\xi} := H'(i_{\bar{Q}})$ , then we have*

$$(3.7) \quad \int_{\Omega} \bar{\xi} i_{\bar{Q}} = \inf_{Q \in \mathcal{Q}^q(\mu_0, \mu_1)} \int_{\Omega} \bar{\xi} i_Q = \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\bar{\xi}}(x, y) d\gamma(x, y).$$

*Proof.* Let us recall that from Proposition 3.1, we have

$$(3.8) \quad \int_{\Omega} \bar{\xi} i_{\bar{Q}} = \inf_{Q \in \mathcal{Q}^q(\mu_0, \mu_1)} \int_{\Omega} \bar{\xi} i_Q.$$

Let  $\xi$  be a nonnegative element of  $L^{q^*}$  and let  $Q \in \mathcal{Q}^q(\mu_0, \mu_1)$ . Using Lemma 3.6 and the definition of  $\mathcal{Q}^q(\mu_0, \mu_1)$  yields

$$\begin{aligned} \int_{\Omega} \xi i_Q &= \int_C L_{\xi} dQ \geq \int_C \bar{c}_{\xi}(\sigma(0), \sigma(1)) dQ(\sigma) \\ &\geq \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\xi}(x, y) d\gamma(x, y). \end{aligned}$$

We then have, for all  $\xi \in L^{q^*}$ ,  $\xi \geq 0$ ,

$$(3.9) \quad \inf_{Q \in \mathcal{Q}^q(\mu_0, \mu_1)} \int_{\Omega} \xi i_Q \geq \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\xi}(x, y) d\gamma(x, y).$$

By a similar argument and using Lemma 3.4, we also have, for all  $\xi \in C^0(\bar{\Omega}, \mathbb{R}_+)$ ,

$$(3.10) \quad \inf_{Q \in \mathcal{Q}(\mu_0, \mu_1)} \int_{\bar{\Omega}} \xi di_Q \geq \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\xi}(x, y) d\gamma(x, y).$$

Let  $\xi \in C^0(\bar{\Omega}, \mathbb{R}_+)$  and  $\varepsilon > 0$ . For every  $x$  and  $y$  in  $\bar{\Omega}$ , there exists  $\sigma_{\varepsilon}^{x,y} \in C^{x,y}$  such that  $(x, y) \mapsto \sigma_{\varepsilon}^{x,y}$  is measurable (see, for instance, [8]), and by Lemma 3.4,

$$(3.11) \quad L_{\xi}(\sigma_{\varepsilon}^{x,y}) \leq c_{\xi}(x, y) + \varepsilon = \bar{c}_{\xi}(x, y) + \varepsilon.$$

Let  $\gamma \in \Pi(\mu_0, \mu_1)$  and let us define the element of  $\mathcal{Q}(\mu_0, \mu_1)$ :  $Q_{\varepsilon} := \delta_{\sigma_{\varepsilon}^{x,y}} \otimes \gamma$ . We then have

$$\int_{\bar{\Omega}} \xi di_{Q_{\varepsilon}} = \int_{\bar{\Omega} \times \bar{\Omega}} L_{\xi}(\sigma_{\varepsilon}^{x,y}) d\gamma(x, y) \leq \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\xi}(x, y) d\gamma(x, y) + \varepsilon.$$

Since  $\gamma$  and  $\varepsilon$  are arbitrary, using (3.10) we obtain

$$(3.12) \quad \inf_{Q \in \mathcal{Q}(\mu_0, \mu_1)} \int_{\bar{\Omega}} \xi di_Q = \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\xi}(x, y) d\gamma(x, y) \quad \forall \xi \in C^0(\bar{\Omega}, \mathbb{R}_+).$$

In what follows, for every  $\mu \in \mathcal{M}_+(\bar{\Omega})$ , we extend  $\mu$  by 0 outside  $\bar{\Omega}$ . Let  $(\rho_n)_n$  be a standard mollifying sequence. For  $n \in \mathbb{N}^*$  let us consider the regularized problem

$$(3.13) \quad \inf_{Q \in \mathcal{Q}(\mu_0, \mu_1)} \int_{\mathbb{R}^2} H(\rho_n \star i_Q).$$

The existence of a solution  $\overline{Q}_n$  of (3.13) can be obtained by similar arguments as in Theorem 2.10 (using Lemma 2.8 and the fact that the  $L^1$  norm of  $\rho_n \star i_Q$  equals the total mass of  $i_Q$ ). Proceeding as in Proposition 3.1 and defining  $j_n := \rho_n \star i_{\overline{Q}_n}$ ,  $\xi_n := H'(j_n)$ ,  $\eta_n := \rho_n \star \xi_n$ , we have

$$(3.14) \quad \int_{\mathbb{R}^2} H'(\rho_n \star i_{\overline{Q}_n})(\rho_n \star i_{\overline{Q}_n}) = \int_{\mathbb{R}^2} \xi_n j_n = \int_{\mathbb{R}^2} \eta_n d i_{\overline{Q}_n} = \inf_{Q \in \mathcal{Q}(\mu_0, \mu_1)} \int_{\mathbb{R}^2} \eta_n d i_Q.$$

With (3.12), we then get

$$(3.15) \quad \int_{\mathbb{R}^2} \eta_n d i_{\overline{Q}_n} = \int_{\mathbb{R}^2} \xi_n j_n = \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\overline{\Omega} \times \overline{\Omega}} \bar{c}_{\eta_n}(x, y) d\gamma(x, y).$$

By convexity of  $H$ , we also have

$$(3.16) \quad \int_{\mathbb{R}^2} H(j_n) \leq \int_{\mathbb{R}^2} H(\rho_n \star i_{\overline{Q}}) \leq \int_{\mathbb{R}^2} \rho_n \star H(i_{\overline{Q}}),$$

which implies that  $j_n$  is bounded in  $L^q$ . Passing to subsequences, we may therefore assume

$$(3.17) \quad j_n \rightharpoonup j \text{ in } L^q, \quad \xi_n \rightharpoonup \xi \text{ in } L^{q^*}, \quad \eta_n \rightharpoonup \xi \text{ in } L^{q^*}.$$

Since the total mass of  $i_{\overline{Q}_n}$  is the same as that of  $j_n$ , and  $j_n$  is bounded in  $L^q$  (and hence in  $L^1$ ), we get a bound on  $i_{\overline{Q}_n}(\Omega)$  and, from Lemma 2.8, we may also assume

$$(3.18) \quad \overline{Q}_n \xrightarrow{*} Q \text{ in } \mathcal{M}_+(C), \quad i_{\overline{Q}_n} \xrightarrow{*} i \text{ in } \mathcal{M}_+(\overline{\Omega}).$$

It is obvious that  $i = j$  and that Lemma 2.9 implies  $j \geq i_Q$ . With (3.16) and the monotonicity of  $H$ , we then get

$$(3.19) \quad \int_{\Omega} H(i_Q) \leq \int_{\Omega} H(j) \leq \liminf_n \int_{\mathbb{R}^2} H(j_n) \leq \int_{\Omega} H(i_{\overline{Q}}).$$

With the strict convexity of  $H$  and the optimality of  $\overline{Q}$ , this also yields

$$(3.20) \quad i_{\overline{Q}} = i_Q = j \in L^q \text{ and } \liminf_n \int_{\mathbb{R}^2} H(j_n) = \int_{\mathbb{R}^2} H(i_{\overline{Q}}).$$

Up to some subsequence (as  $\bar{\xi} = H'(i_{\overline{Q}}) \in L^{q^*}$ ), this also implies

$$H(j_n) - H(i_{\overline{Q}}) - \bar{\xi}(j_n - i_{\overline{Q}}) \rightarrow 0 \text{ a.e. and in } L^1,$$

and using the strict convexity of  $H$ , we deduce that  $j_n$  converges a.e. to  $i_{\overline{Q}}$ . This implies that  $\xi_n$  converges a.e. to  $\bar{\xi} = H'(i_{\overline{Q}})$  and that  $\xi = \bar{\xi}$ . With Fatou's lemma and (3.15), we therefore obtain

$$\begin{aligned} \int_{\Omega} \bar{\xi} i_{\overline{Q}} &= \int_{\mathbb{R}^2} H'(i_{\overline{Q}}) i_{\overline{Q}} \\ &\leq \liminf_n \int_{\mathbb{R}^2} \xi_n j_n \\ &= \liminf_n \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\overline{\Omega} \times \overline{\Omega}} \bar{c}_{\eta_n}(x, y) d\gamma(x, y). \end{aligned}$$

Using  $\eta_n \rightarrow \bar{\xi}$  and Remark 3.3, from the uniform convergence of  $c_{\eta_n}$  to a cost  $c \leq c_{\bar{\xi}}$  we get

$$\int_{\Omega} \bar{\xi} i_{\bar{Q}} \leq \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\bar{\xi}}(x, y) d\gamma(x, y).$$

Together with (3.8) and (3.9), this completes the proof.  $\square$

The characterization of optimal transport strategies then reads as follows.

**THEOREM 3.10.** *Let us assume that  $q < 2$  and that  $H$  is strictly convex. A transportation strategy  $(\bar{\gamma}, \bar{p})$  is optimal if and only if, setting  $\bar{Q} := Q_{\bar{\gamma}, \bar{p}}$  and  $\bar{\xi} := H'(i_{\bar{Q}})$ , one has that*

1.  $\bar{\gamma}$  solves the Monge–Kantorovich problem

$$(3.21) \quad \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\bar{\xi}}(x, y) d\gamma(x, y);$$

2. for  $\bar{Q}$ -a.e.  $\sigma \in C$ , one has

$$(3.22) \quad L_{\bar{\xi}}(\sigma) = \bar{c}_{\bar{\xi}}(\sigma(0), \sigma(1)).$$

*Proof.* Let us assume first that the transportation strategy  $(\bar{\gamma}, \bar{p})$  is optimal and set  $\bar{Q} := Q_{\bar{\gamma}, \bar{p}}$  and  $\bar{\xi} := H'(i_{\bar{Q}})$ . From Proposition 3.9 and Lemma 3.6, we get

$$\begin{aligned} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\bar{\xi}}(x, y) d\bar{\gamma}(x, y) &= \int_C \bar{c}_{\bar{\xi}}(\sigma(0), \sigma(1)) d\bar{Q}(\sigma) \\ &\leq \int_C L_{\bar{\xi}} d\bar{Q} = \int_{\Omega} \bar{\xi} i_{\bar{Q}} \\ &= \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\bar{\xi}}(x, y) d\gamma(x, y). \end{aligned}$$

This proves that  $\bar{\gamma}$  solves (3.21) and implies that the inequalities above are equalities. We therefore deduce (3.22) from the inequality  $\bar{c}_{\bar{\xi}}(\sigma(0), \sigma(1)) \leq L_{\bar{\xi}}(\sigma)$ .

Conversely, assume that the transportation strategy  $(\bar{\gamma}, \bar{p})$  satisfies the two conditions of the theorem. Condition (3.22) first yields

$$\int_{\Omega} \bar{\xi} i_{\bar{Q}} = \int_C L_{\bar{\xi}} d\bar{Q} = \int_C \bar{c}_{\bar{\xi}}(\sigma(0), \sigma(1)) d\bar{Q}(\sigma) = \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\bar{\xi}}(x, y) d\bar{\gamma}(x, y).$$

Second, if  $Q = Q_{\gamma, p} \in \mathcal{Q}^q(\mu_0, \mu_1)$ , one has

$$\int_{\Omega} \bar{\xi} i_Q = \int_C L_{\bar{\xi}} dQ \geq \int_C \bar{c}_{\bar{\xi}}(\sigma(0), \sigma(1)) dQ(\sigma) = \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\bar{\xi}}(x, y) d\gamma(x, y),$$

and since  $\bar{\gamma}$  solves (3.21), we finally have

$$\int_{\Omega} \bar{\xi} i_{\bar{Q}} \leq \int_{\Omega} \bar{\xi} i_Q \quad \forall Q \in \mathcal{Q}^q(\mu_0, \mu_1),$$

which, with Proposition 3.1, proves that  $(\bar{\gamma}, \bar{p})$  is optimal.  $\square$

Let us see, through an easy example, an application of Theorem 3.10.

**Example 3.11.** Suppose that  $\bar{\Omega}$  contains the two segments  $A = \{0\} \times [0, 1]$  and  $B = \{1\} \times [0, 1]$  and the square  $S = [0, 1] \times [0, 1]$ , which is their convex hull. Set



$\mu_1 = \mathcal{H}^1 \llcorner A$  and  $\mu_2 = \mathcal{H}^1 \llcorner B$  and denote by  $T : A \rightarrow C$  the map that associates to every point  $(0, x) \in A$  the curve  $T(x)$  given by  $T(x)(t) = (t, x)$ , i.e., the horizontal segment from  $A$  to  $B$  starting from  $x$ . Set  $Q = T\# \mu_1$ . It is clear that  $Q$  comes from an admissible transportation strategy linking  $\mu_1$  to  $\mu_2$ , and it is not difficult to see that the traffic intensity  $i_Q$  has constant density 1 on  $S$  and 0 elsewhere. We consider only two particular cases: we claim that  $Q$  is optimal if  $\Omega = ]0, 1[ \times ]0, 1[$ , while it is not if  $S$  is compactly contained in  $\Omega$ . Indeed, if  $\bar{\Omega} = S$ , the metric induced by  $i_Q$  is the Euclidean metric, the paths  $T(x)$  are geodesic, and the transport plan induced by  $Q$  is optimal according to this metric. On the other hand, if  $\Omega$  is larger than  $S$ , then all the segments  $T(x)$  that are very close to the upper or lower boundary of  $S$  are not geodesic according to this metric, because they could be improved by nonstraight line paths which arrive up to zone  $\Omega \setminus S$ , where  $i_Q = 0$  and the transportation is cheaper. In the former case, consequently, the sufficient optimality conditions are satisfied, while in the latter the geodesic conditions on the paths (Wardrop condition; see the next section) is not and prevents optimality.

*Remark 3.12.* Let us remark that  $\bar{\xi} = H'(i_{\bar{Q}})$  (with  $\bar{Q}$  solving (2.13)) solves the following (dual of (2.13)) problem:

$$(3.23) \quad \sup_{\xi \in L^{q^*}, \xi \geq 0} W(\xi) - \int_{\Omega} H^*(\xi(x)) dx,$$

where  $H^*$  is the Fenchel transform of  $H$ , and

$$W(\xi) := \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\xi}(x, y) d\gamma(x, y).$$

Indeed, it follows from Proposition 3.9 and  $\bar{\xi} = H'(i_{\bar{Q}})$  that

$$W(\bar{\xi}) - \int_{\Omega} H^*(\bar{\xi}(x)) dx = \int_{\Omega} (\bar{\xi} i_{\bar{Q}} - H^*(\bar{\xi})) = \int_{\Omega} H(i_{\bar{Q}}).$$

If  $\xi \in L^{q^*}$  is nonnegative, we deduce from (3.9) and Young's inequality,

$$W(\xi) \leq \int_{\Omega} \xi i_{\bar{Q}} \leq \int_{\Omega} H(i_{\bar{Q}}) + \int_{\Omega} H^*(\xi),$$

which proves that  $\bar{\xi}$  solves (3.23).

*Remark 3.13.* It is very natural to investigate numerical schemes for the primal problem (2.13). We believe that the dual problem (3.23) offers a tractable and convenient way to address this numerical issue. We are not developing this point further here and leave the numerical approximation of (2.13) and (3.23) for future research.

**4. Application to equilibria of Wardrop type.** In this final section, we relate the results of the previous sections to some concepts of equilibria of Wardrop type. Modeling congestion as in section 2.2 enables us to extend the concept of Wardrop equilibrium to a continuous setting.

Let us consider a congestion function  $g : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ , which is continuous increasing and satisfies  $az^{q-1} \leq g(z) \leq b(z^{q-1} + 1)$  for all  $z \in \mathbb{R}_+$  and some  $q \in (1, 2)$  and nonnegative constants  $a$  and  $b$ . Then for any transportation strategy  $(\gamma, p)$  such that  $I_{\gamma, p}$  (defined by (2.3)) belongs to  $L^q$ , the transportation cost function resulting from the strategy  $(\gamma, p)$  is  $\bar{c}_{\xi}$  for  $\xi := g \circ I_{\gamma, p} \in L^{q^*}$ . Roughly speaking, an equilibrium is then a transportation strategy  $(\gamma, p)$  that satisfies the Wardrop stability condition

(i.e.,  $Q_{\gamma,p}$  gives full mass to the set of geodesics for the metric  $\xi = g \circ I_{\gamma,p}$ ) and the additional requirement that  $\gamma$  is an optimal transportation plan between  $\mu_0$  and  $\mu_1$  for the cost resulting from  $(\gamma, p)$ . This leads to the following.

DEFINITION 4.1. *A transportation strategy  $(\bar{\gamma}, \bar{p})$  is said to be an equilibrium if  $I_{\bar{\gamma}, \bar{p}} \in L^q$  and, setting  $\bar{\xi} := g \circ I_{\bar{\gamma}, \bar{p}}$ , one has*

1.  $L_{\bar{\xi}}(\sigma) = \bar{c}_{\bar{\xi}}(\sigma(0), \sigma(1))$  for  $Q_{\bar{\gamma}, \bar{p}}$ -a.e.  $\sigma \in C$ ;
2.  $\bar{\gamma}$  solves the Monge–Kantorovich problem

$$\inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\bar{\Omega} \times \bar{\Omega}} \bar{c}_{\bar{\xi}}(x, y) d\gamma(x, y).$$

Only the first condition above is linked to Wardrop's original equilibrium concept. Imagine that some social planner chooses the transportation plan  $\gamma$ ; then the second equilibrium condition expresses that  $\gamma$  is optimal for the transportation cost resulting from  $\gamma$  itself and the traveler's individual behavior. Our notion of equilibrium can therefore be viewed either as a refinement of the Wardrop equilibrium or as its generalization to the case where the transportation plan is not given a priori.

Defining the function  $H_g : \mathbb{R}_+ \rightarrow \mathbb{R}$  by

$$(4.1) \quad H'_g(z) = g(z) \quad \forall z \in \mathbb{R}_+, \quad H_g(0) = 0,$$

a direct application of Theorems 2.10 and 3.10 then gives the existence of equilibria together with a variational characterization as follows.

THEOREM 4.2. *If  $Q^q(\mu_0, \mu_1) \neq \emptyset$ , there exists an equilibrium. Moreover,  $(\bar{\gamma}, \bar{p})$  is an equilibrium if and only if  $\bar{Q} := Q_{\bar{\gamma}, \bar{p}}$  solves the minimization problem*

$$(4.2) \quad \inf_{Q \in Q^q(\mu_0, \mu_1)} \int_{\Omega} H_g(i_Q(x)) dx,$$

where  $H_g$  is defined by (4.1).

REMARK 4.3. We have always assumed that the congestion function  $g$  depends only on the intensity  $z$ . It is, of course, straightforward to extend all our results to the case of a congestion function  $(x, z) \rightarrow g(x, z)$  that also depends on  $x$ . In this case, one finds equilibria by solving

$$\inf_{Q \in Q^q(\mu_0, \mu_1)} \int_{\Omega} H_g(x, i_Q(x)) dx,$$

where  $H_g$  is defined by  $H_g(x, 0) = 0$  and  $\partial_z H_g(x, z) = g(x, z)$ .

REMARK 4.4. A slightly different situation, which can be relevant in some applications, occurs when not only the marginals  $\mu_0$  and  $\mu_1$  are fixed but also when the transportation plan  $\bar{\gamma} \in \Pi(\mu_0, \mu_1)$  is fixed. In this case, one defines equilibria as the set of  $\bar{p}$ 's such that  $(\bar{\gamma}, \bar{p})$  satisfies the first condition (Wardrop) of Definition 4.1. If the set

$$Q^q(\bar{\gamma}) := \{Q \in \mathcal{M}_+^1(C) : (e_0, e_1) \# Q = \bar{\gamma}, i_Q \in L^q\}$$

is nonempty, then by slightly adapting our arguments we have existence of equilibria, and  $\bar{p}$  is an equilibrium if and only if  $\bar{Q} := Q_{\bar{\gamma}, \bar{p}}$  solves the minimization problem

$$\inf_{Q \in Q^q(\bar{\gamma})} \int_{\Omega} H_g(i_Q(x)) dx.$$

## REFERENCES

- [1] L. AMBROSIO, *Lecture notes on optimal transport problems*, in Mathematical Aspects of Evolving Interfaces, Lecture Notes in Math. 1812, Springer, Berlin, 2003, pp. 1–52.
- [2] J.-B. BAILLON AND R. COMINETTI, *Markovian traffic equilibrium*, Math. Prog., 111 (2008), pp. 33–56.
- [3] G. BOUCHITTÉ AND G. BUTTAZZO, *Characterization of optimal shapes and masses through Monge–Kantorovich equation*, J. Eur. Math. Soc., 3 (2001), pp. 139–168.
- [4] G. BOUCHITTÉ, G. BUTTAZZO, AND P. SEPPECHER, *Shape optimization solutions via Monge–Kantorovich equation*, C. R. Acad. Sci. Paris, 324 (1997), pp. 1185–1191.
- [5] Y. BRENIER, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.
- [6] H. BRÉZIS, *Analyse Fonctionnelle*, Masson, Paris, 1983.
- [7] L. CAFFARELLI, M. FELDMAN, AND R. J. MCCANN, *Constructing optimal maps for Monge’s transport problem as a limit of strictly convex costs*, J. Amer. Math. Soc., 15 (2002), pp. 1–26.
- [8] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, 1977.
- [9] C. DELLACHERIE AND P.-A. MEYER, *Probabilities and Potential*, Math. Stud. 29, North-Holland, Amsterdam, 1978.
- [10] L. DE PASCALE AND A. PRATELLI, *Regularity properties for Monge transport density and for solutions of some shape optimization problem*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 249–274.
- [11] L. DE PASCALE AND A. PRATELLI, *Sharp summability for Monge transport density via interpolation*, ESAIM Control Opt. Calc. Var., 10 (2004), pp. 549–552.
- [12] G. MONGE, *Mémoire sur la théorie des déblais et des remblais*, Hist. de l’Acad. Roy. Sci. Paris, 1781, pp. 666–704.
- [13] C. VILLANI, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.
- [14] J. G. WARDROP, *Some theoretical aspects of road traffic research*, Proc. Inst. Civ. Eng., 2 (1952), pp. 325–378.

## CONTROLLABILITY AND OBSERVABILITY OF SECOND ORDER DESCRIPTOR SYSTEMS\*

PHILIP LOSSE<sup>†</sup> AND VOLKER MEHRMANN<sup>‡</sup>

**Abstract.** We analyze controllability and observability conditions for second order descriptor systems and show how the classical conditions for first order systems can be generalized to this case. We show that performing a classical transformation to first order form may destroy some controllability and observability properties. As an example, we demonstrate that the loss of impulse controllability in constrained multibody systems is due to the representation as a first order system. To avoid this problem, we will derive a canonical form and new first order formulations that do not destroy the controllability and observability properties.

**Key words.** descriptor system, impulse controllability, impulse observability, second order system, order reduction, index reduction, complete controllability, strong controllability, complete observability, strong observability

**AMS subject classifications.** 93B05, 93B07, 93B10

**DOI.** 10.1137/060673977

**1. Introduction.** We study linear second order constant coefficient descriptor control problems of the form

$$(1.1) \quad M\ddot{x} + G\dot{x} + Kx = Bu,$$

$$(1.2) \quad y = Cx,$$

$$(1.3) \quad x(0) = x_0, \quad \dot{x}(0) = \dot{x}_0$$

with coefficients  $M, G, K \in \mathbb{R}^{n,n}$ ,  $C \in \mathbb{R}^{p,n}$ , and  $B \in \mathbb{R}^{n,m}$ . Here  $\mathbb{R}^{n,\ell}$  denotes the vector space of  $n \times \ell$  real matrices,  $x$  is the state,  $u$  is the input or control, and  $y$  is the output of the system. In particular, we study descriptor systems, where the matrix  $M$  is singular, and, despite the fact that formally  $\ddot{x}$  and  $\dot{x}$  occur in (1.1), we require that  $\ddot{x}$  only has to exist outside the kernel of  $M$  and that  $\dot{x}$  has to exist outside the kernel of  $G$ .

All of the results in this paper also carry over to the complex case, and they can also be easily extended to systems of higher than second order, but, for ease of notation and because this is the most important case in practice, we restrict ourselves to the real second order case.

In the following we denote by  $I$  or  $I_n$  the identity matrix of size  $n \times n$  and by  $A^T$  the transpose of a matrix  $A$ . We denote a matrix with orthonormal columns spanning the right null space of the matrix  $M$  by  $S_\infty(M)$  and a matrix with orthonormal columns spanning the left null space of  $M$  by  $T_\infty(M)$ . These matrices are not uniquely determined, although the corresponding spaces are. Nevertheless, for simplicity, we speak of these matrices as the corresponding spaces.

---

\*Received by the editors November 2, 2006; accepted for publication (in revised form) October 2, 2007; published electronically March 26, 2008.

<http://www.siam.org/journals/sicon/47-3/67397.html>

<sup>†</sup>Fakultät für Mathematik, TU Chemnitz, D-09107 Chemnitz, Germany (philip.losse@mathematik.tu-chemnitz.de). This author was supported by Deutsche Forschungsgemeinschaft through project BE-2174/6-1,2.

<sup>‡</sup>Institut für Mathematik, TU Berlin, Str. des 17. Juni 136, D-10623 Berlin, Germany (mehrmann@math.tu-berlin.de). This author was partially supported by Deutsche Forschungsgemeinschaft through project ME 790/16-1.

Second order descriptor systems arise in the control of constrained mechanical systems (see, e.g., [14, 19, 23, 36, 38, 39, 40]) in the control of electrical and electromechanical systems [2, 3], and in particular in heterogeneous systems, where different models are coupled together [37].

Usually, in the classical theory of ordinary differential equations and *classical state space systems* (i.e., descriptor systems where the leading coefficient is the identity), second order systems are turned into first order systems by introducing new variables for the first derivative. This gives rise to linear first order *descriptor* (or *generalized state space*) systems of the form

$$(1.4) \quad E\dot{\xi} = A\xi + B_1 u,$$

$$(1.5) \quad y = C_1 \xi,$$

$$(1.6) \quad \xi(0) = \xi_0.$$

Let us briefly recall some results for first order descriptor systems; see, e.g., [4, 9, 12, 44]. In contrast to classical state space systems, where  $E = I$ , the response of a descriptor system can consist of step functions or can be discussed only in a distributional setting [10, 18, 43], if the input function  $u$  is not sufficiently smooth. But here we are interested only in classical solutions in the sense that  $M\ddot{x}$  and  $G\dot{x}$  exist, and we explicitly want to avoid impulsive terms in the solution; thus, we do not use this formulation.

The response of system (1.4) can be described in terms of the eigenstructure of the matrix pencil  $\alpha E - \beta A$ . The pencil and the corresponding system (1.4)–(1.5) are said to be *regular* if  $\det(\alpha E - \beta A) \neq 0$  for some  $(\alpha, \beta) \in \mathbb{C}^2$ . Regular systems are *solvable* in the sense that (1.4) admits a classical solution  $\xi : \mathbb{R} \rightarrow \mathbb{R}^n$ , with  $\xi$  differentiable in the image of  $E$  for all sufficiently smooth controls  $u$  and consistent initial conditions  $\xi_0$  [9, 12, 44].

For regular pencils, *generalized eigenvalues* are the pairs  $(\alpha, \beta) \in \mathbb{C}^2 \setminus \{(0, 0)\}$  for which  $\det(\alpha E - \beta A) = 0$ . If  $\beta \neq 0$ , then the pair represents the finite eigenvalue  $\lambda = \alpha/\beta$ . If  $\beta = 0$ , then  $(\alpha, \beta)$  represents an “infinite” eigenvalue. In the following, for simplicity, we use the notation with  $\lambda$ .

The solution and many properties of the *free descriptor system* (with  $u = 0$ ) can be characterized in terms of the Weierstraß canonical form (WCF) for regular matrix pencils.

**THEOREM 1.1** (see [16]). *If  $\lambda E - A$  is a regular pencil, then there exist nonsingular matrices  $X = \begin{bmatrix} X_r & X_\infty \end{bmatrix} \in \mathbb{R}^{n,n}$  and  $Y = \begin{bmatrix} Y_r & Y_\infty \end{bmatrix} \in \mathbb{R}^{n,n}$  for which*

$$(1.7) \quad Y^T E X = \begin{bmatrix} Y_r^T \\ Y_\infty^T \end{bmatrix} E \begin{bmatrix} X_r & X_\infty \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}$$

and

$$(1.8) \quad Y^T A X = \begin{bmatrix} Y_r^T \\ Y_\infty^T \end{bmatrix} A \begin{bmatrix} X_r & X_\infty \end{bmatrix} = \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix},$$

where  $J$  is a matrix in real Jordan canonical form whose eigenvalues are the finite eigenvalues of the pencil and  $N$  is a nilpotent matrix, also in Jordan form.  $J$  and  $N$  are unique up to permutation of Jordan blocks.

Usually, the index of nilpotency  $\nu$  of the nilpotent matrix  $N$  in (1.7) is called the *differentiation index* or *index* of the system, and if  $E$  is nonsingular, then the pencil is said to be of index zero. In recent years the theory of descriptor systems has

been extended to rectangular, time-varying, and even nonlinear systems, and different index concepts, in particular the *strangeness index*, have been introduced; see [30] for a recent textbook. The strangeness index generalizes the index of a linear descriptor system to over- and underdetermined linear and nonlinear systems, and it uses a slightly different counting; i.e., systems of the form (1.4) with an index of at most one have a strangeness index of zero and are called *strangeness-free*. For all other systems where the differentiation index is defined, it is the strangeness index plus 1. Since we restrict ourselves to square systems, we will use only the differentiation index  $\nu$  and call it the *index of the system*.

For first order systems (1.4) the index describes the degree of differentiability of the input function that is needed to achieve a continuous solution with the property that  $Ex$  is continuously differentiable. In analogy we define the *index of the second order system* (1.1) to be the degree of differentiability of the input function that is needed to achieve a continuous solution with the property that  $Mx$  is twice and  $Gx$  is once continuously differentiable.

In the notation of (1.7)–(1.8), classical solutions of (1.4) take the form

$$\xi(t) = X_r z_1(t) + X_\infty z_2(t),$$

where

$$(1.9) \quad \begin{aligned} \dot{z}_1 &= Jz_1 + Y_r^T B_1 u, \\ N\dot{z}_2 &= z_2 + Y_\infty^T B_1 u. \end{aligned}$$

This system admits the explicit solution

$$(1.10) \quad \begin{aligned} z_1(t) &= e^{tJ} z_1(t_0) + \int_{t_0}^t e^{(t-s)J} Y_r^T B_1 u(s) ds, \\ z_2(t) &= - \sum_{i=0}^{\nu-1} \frac{d^i}{dt^i} (N^i Y_\infty^T B_1 u(t)), \end{aligned}$$

where  $\nu$  is the index of the system. Equation (1.10) shows that, for regular systems that are not of index at most one, in order to have classical, continuous solutions, the input  $u$  has to be sufficiently smooth, and, to ensure a smooth response for every continuous input  $u$ , the system must be regular and of index at most one. Under certain further requirements that we discuss below, this property may, however, be achieved by feedback. If this is the case, then the system is said to be *regularizable*.

Equation (1.10) also shows that the initial condition  $\xi_0$  is restricted. For a given input function  $u$ , the set of *consistent* initial conditions is given by

$$(1.11) \quad \mathcal{S} = \left\{ X_r z_1 + X_\infty z_2 \mid z_1 \in \mathbb{R}^r, z_2 = - \sum_{i=0}^{\nu-1} \left( \frac{d^i}{dt^i} (N^i Y_\infty^T B_1 u)(0) \right) \right\}.$$

The set  $\mathcal{R}$  of *reachable* states or *reachable set* of (1.4) from the set  $\mathcal{S}$  of consistent initial conditions is  $\mathcal{S}$  itself [44].

Coming back to second order descriptor systems and their first order representations, one should note first that there is no unique way of performing this transformation to first order; see [33] for large vector spaces of first order formulations in the context of eigenvalue problems. As a consequence, the solution space and the set of admissible controls may be different for different first order formulations.

This has recently been shown in the context of the numerical solution of higher order differential-algebraic systems [35, 41]. There, it also has been demonstrated that the classical first order formulations may even lead to false results if certain smoothness conditions are not met or if the initial conditions are not chosen properly.

Let us illustrate these difficulties with the well-known example of mechanical multibody systems.

*Example 1.2.* Consider a simplified, linearized model of a two-dimensional, three-link mobile manipulator [22]. The Lagrangian equations of motion in its linearized form are given by a linear second order system

$$\begin{aligned} M_0 \ddot{z} + G_0 \dot{z} + K_0 z &= B_0 u - H_0^T \phi, \\ H_0 z &= 0, \end{aligned}$$

where  $M_0$  represents the nonsingular mass matrix,  $G_0$  the coefficient matrix associated with damping, centrifugal, gravity, and Coriolis forces,  $K_0$  the stiffness matrix, and  $H_0$  the constraint, whereas  $\phi$  is a vector of Lagrange multipliers.

By setting  $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} z \\ \phi \end{bmatrix}$ , and adding an output equation

$$y = Cz = \begin{bmatrix} C_0 & 0 \end{bmatrix} x,$$

we obtain a descriptor system of the form (1.1)–(1.2) given by

$$\begin{aligned} \begin{bmatrix} M_0 & 0 \\ 0 & 0 \end{bmatrix} \ddot{x} + \begin{bmatrix} G_0 & 0 \\ 0 & 0 \end{bmatrix} \dot{x} + \begin{bmatrix} K_0 & H_0^T \\ H_0 & 0 \end{bmatrix} x &= \begin{bmatrix} B_0 \\ 0 \end{bmatrix} u, \\ y &= \begin{bmatrix} C_0 & 0 \end{bmatrix} x. \end{aligned}$$

If one would follow the usual approach for ordinary differential equations, then one would introduce a new state vector, often called a *descriptor vector*,

$$\xi = \begin{bmatrix} \dot{x} \\ x \end{bmatrix} = \begin{bmatrix} \dot{z} \\ \dot{\phi} \\ z \\ \phi \end{bmatrix}.$$

Under the usual assumptions that  $M_0$  is invertible and that  $H_0$  has full row rank, it is easy to check that the resulting descriptor system has blocks of size 4 in the Weierstraß form associated with the eigenvalue  $\infty$  and thus an index  $\nu = 4$ . It follows that the input functions have to be at least three times continuously differentiable to obtain a continuous solution.

As a simple example, consider the system

$$(1.12) \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \ddot{x} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \dot{x} + \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} u,$$

which has the structure of a constrained and damped mechanical system. The classical first order version yields

$$E = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}.$$

A transformation  $(\hat{E}, \hat{A}) = P(E, A)Q$ , with

$$P = \begin{bmatrix} 0 & 0 & 0 & -1 \\ -1 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & 0 & -1 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 \end{bmatrix},$$

yields the Weierstraß canonical form

$$(\hat{E}, \hat{A}) = \left( \left[ \begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array} \right], \left[ \begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array} \right] \right),$$

which consists only of one block associated with the eigenvalue  $\infty$  of size 4.

This classical approach, however, is usually not taken in practice, since on one hand it would introduce the unnecessary derivative of the Lagrange multiplier  $\phi$ , which may not be differentiable, and also this approach would require extra initial values associated with  $\dot{\phi}(t_0)$  which usually are not available. In practice, one therefore uses the knowledge about the structure of the system and introduces the reduced descriptor vector

$$\xi = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} = \begin{bmatrix} \dot{z} \\ z \\ \phi \end{bmatrix}.$$

In this way one obtains a first order descriptor system of the form

$$\begin{aligned} \begin{bmatrix} M_0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & I & 0 \end{bmatrix} \dot{\xi} &= \begin{bmatrix} -G_0 & -K_0 & -H_0^T \\ 0 & -H_0 & 0 \\ I & 0 & 0 \end{bmatrix} \xi + \begin{bmatrix} B_0 \\ 0 \\ 0 \end{bmatrix} u, \\ (1.13) \quad y &= \begin{bmatrix} 0 & C_0 & 0 \end{bmatrix} \xi, \end{aligned}$$

which has index  $\nu = 3$ .

For the simple system (1.12) the first order formulation (1.13) has

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} -1 & -1 & -1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

If we use the transformation  $(\hat{E}, \hat{A}) = P(E, A)Q$ , with

$$P = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix},$$

then we obtain the Weierstraß canonical form

$$(\hat{E}, \hat{A}) = \left( \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right),$$

which has only one block of size 3.



From this example we see that different first order formulations lead to different indexes and therefore to different differentiability requirements for the input functions  $u$  which we assume to be at least piecewise continuous functions.

But there is a second difficulty which both first order formulations in Example 1.2 share that is connected to the controllability and observability of the descriptor system and its first order formulations.

To describe this second difficulty we return again to our review of results for first order descriptor systems (1.4)–(1.5). Typically one or more of the following conditions is essential for most classical control design aims:

$$\begin{aligned}
 \mathbf{C0}: \quad & \text{rank}[\alpha E - \beta A, B_1] = n \text{ for all } (\alpha, \beta) \in \mathbb{C}^2 \setminus \{(0, 0)\}; \\
 \mathbf{C1}: \quad & \text{rank}[\lambda E - A, B_1] = n \text{ for all } \lambda \in \mathbb{C}; \\
 (1.14) \quad \mathbf{C2}: \quad & \text{rank}[E, AS_\infty(E), B_1] = n.
 \end{aligned}$$

A regular first order descriptor system is called *completely controllable* or *C-controllable* if **C0** holds [44] and *controllable in the reachable set* or  *$\mathcal{R}$ -controllable* if condition **C1** holds. The system is called *strongly controllable* or *S-controllable* if **C1** and **C2** hold [5]. C-controllability ensures that for any given initial and final states  $\xi_0, \xi_f$  there exists a piecewise continuous control  $u$  that transfers the system from  $\xi_0$  to  $\xi_f$  in finite time [44], while S-controllability ensures the same for any given initial and final states in the reachable set, i.e.,  $\xi_0, \xi_f \in \mathcal{R}$ . Systems that satisfy condition **C2** are called *controllable at infinity*, *impulse-controllable*, or *I-controllable* [11, 27, 43]. For these systems, impulsive modes that arise from a high index of  $(E, A)$  can be avoided by a suitable linear feedback; see [4, 5]. It has been shown in [12] that a first order descriptor system is C-controllable if and only if it is  $\mathcal{R}$ -controllable and  $\text{rank} \begin{bmatrix} E & B_1 \end{bmatrix} = n$ . To have S-controllability, however, the condition that  $\text{rank} \begin{bmatrix} E & B_1 \end{bmatrix} = n$  is not needed; see [4, 5].

Observability for descriptor systems is the dual of controllability. Consider the following conditions:

$$\begin{aligned}
 \mathbf{O0}: \quad & \text{rank} \begin{bmatrix} \alpha E - \beta A \\ C_1 \end{bmatrix} = n \text{ for all } (\alpha, \beta) \in \mathbb{C}^2 \setminus \{(0, 0)\}; \\
 \mathbf{O1}: \quad & \text{rank} \begin{bmatrix} \lambda E - A \\ C_1 \end{bmatrix} = n \text{ for all } \lambda \in \mathbb{C}; \\
 (1.15) \quad \mathbf{O2}: \quad & \text{rank} \begin{bmatrix} E \\ T_\infty^T(E)A \\ C_1 \end{bmatrix} = n.
 \end{aligned}$$

A regular descriptor system is called *completely observable* or *C-observable* if condition **O0** holds, *observable in the reachable set* or  *$\mathcal{R}$ -observable* if condition **O2** holds, and *strongly observable* or *S-observable* if conditions **O1** and **O2** hold. A system that satisfies condition **O2** is called *observable at infinity*, *impulse-observable*, or *I-observable*. Analogous to the controllable case, a system is C-observable if and only if it is  $\mathcal{R}$ -observable and  $\text{rank} \begin{bmatrix} E \\ C_1 \end{bmatrix} = n$ ; see [12].

Note that the conditions (1.14) are preserved under equivalence transformations of the system and under state and output feedback. Analogous properties hold for (1.15).

Classical design approaches in control require the system to be at least S-controllable and S-observable; see [12, 30, 34]. But it is well known that in many practical examples, e.g., in the context of constrained mechanical systems, the resulting system

in neither of the first order formulations as described in Example 1.2 is I-controllable and I-observable.

*Example 1.3.* Consider the first order formulation (1.13) in Example 1.2 with

$$E = \begin{bmatrix} M_0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & I & 0 \end{bmatrix}, \quad A = \begin{bmatrix} -G_0 & -K_0 & -H_0^T \\ 0 & -H_0 & 0 \\ I & 0 & 0 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} B_0 \\ 0 \\ 0 \end{bmatrix}, \quad C_1 = [0 \quad C_0 \quad 0].$$

Here we immediately see that

$$\begin{bmatrix} E & AS_\infty(E) & B_1 \end{bmatrix} = \begin{bmatrix} M_0 & 0 & 0 & -H_0^T & B_0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \end{bmatrix}$$

does not have full row rank if constraints are present, and hence the system is not I-controllable. Similarly,

$$\begin{bmatrix} E \\ T_\infty^T(E)A \\ C_1 \end{bmatrix} = \begin{bmatrix} M_0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & -H_0 & 0 \\ 0 & C_0 & 0 \end{bmatrix}$$

does not have full column rank either; i.e., the system is not I-observable. Furthermore, neither  $[E \quad B_1]$  has full row rank, nor  $\begin{bmatrix} E \\ C_1 \end{bmatrix}$  has full column rank, and hence the system is neither C-controllable nor C-observable.

It should be noted that a first order system that is regular and of index at most one is always I-controllable, since already  $\text{rank}[E, AS_\infty(E)]$  is full, which follows directly from the Weierstraß canonical form.

Since the conditions of I-controllability and I-observability are so important, it has been discussed, for the first order case in [6] for linear systems with constant coefficients and in [7, 26, 29, 31] for linear variable coefficient and nonlinear systems (see also [30]), how systems that are not I-controllable can be modified by a combination of index reduction and feedback to have this property. It has also been argued in [6] that if the system is not I-observable, then the modeling of the system should be reconsidered, since this means that the solution can be formulated only with the help of distributions, but these impulsive parts are not observed.

In view of all of these difficulties it is a natural question to ask whether the choice of the first order formulation may be the reason for the described difficulties with the I-controllability and I-observability. To analyze this question is the topic of the present paper which is organized as follows.

In section 2 we derive normal forms that allow us to check the controllability and observability conditions and the construction of adequate first order formulations. In sections 3 and 4 we then derive the controllability and observability conditions for second order systems analogous to **C0**, **C1**, **C2** and **O1**, **O2**, **O3**. We demonstrate that we can always find first order formulations which are guaranteed to be I-controllable and I-observable, so that the described difficulties can be avoided. We finish with some conclusions.

**2. Normal forms.** In this section we will discuss partial normal forms for matrix triples. The general results for matrix tuples can be found in [35].

DEFINITION 2.1. *Two second order descriptor systems of the form (1.1) with system matrices  $(M, G, K, B)$ , and  $(\hat{M}, \hat{G}, \hat{K}, \hat{B})$  are called strongly equivalent if there exist nonsingular matrices  $P \in \mathbb{R}^{n,n}$ ,  $Q \in \mathbb{R}^{n,n}$ , and  $V \in \mathbb{R}^{m,m}$  such that*

$$(2.1) \quad \hat{M} = PMQ, \quad \hat{G} = PGQ, \quad \hat{K} = PKQ, \quad \hat{B} = PBV.$$

We write  $(M, G, K, B) \sim (\hat{M}, \hat{G}, \hat{K}, \hat{B})$ .

Canonical forms under strong equivalence are known only for the case of matrix pairs, giving the Weierstraß and Kronecker canonical forms [15, 16]. For matrix triples or larger tuples, such canonical forms are not known. Condensed forms which present partial information about the invariants associated with the eigenvalue  $\infty$  and the singular chains have recently been given in [35]. We will recall and extend these results below.

Another class of equivalence transformations that is studied in matrix polynomials is unimodular transformations such as adding the  $\lambda a$  multiple of one row to another (or the same for columns) without increasing the degree of the polynomial. The analogue of these transformations in the context of descriptor systems is well studied [30] and has been discussed in the context of higher order systems in [35]. We reformulate these transformations by using the concept of differential polynomials; see, e.g., [25]. Let  $\mathbb{R}[D_i]$  be the set of *ith order differential polynomials with coefficients in  $\mathbb{R}$* , i.e.,

$$\mathbb{R}[D_i] := \left\{ a_0 + a_1 \frac{d}{dt} + a_2 \frac{d^2}{dt^2} + \dots + a_i \frac{d^i}{dt^i} \mid a_k \in \mathbb{R}, k = 0, 1, \dots, i \right\}.$$

Since we do not want to increase the order of the polynomial, we consider only the following restricted transformations.

DEFINITION 2.2. *Systems  $M\ddot{x} + G\dot{x} + Kx = Bu$  and  $\hat{M}\ddot{x} + \hat{G}\dot{x} + \hat{K}x = \hat{B}u$ , with  $M, G, K, \hat{M}, \hat{G}, \hat{K} \in \mathbb{R}^{n,n}$ ,  $B, \hat{B} \in \mathbb{R}^{n,m}$ , are called order-preserving unimodularly equivalent, or opu-equivalent, if there exists  $P \in \mathbb{R}[D_2]^{n,n}$  with a constant nonzero determinant such that*

$$P(M\ddot{x} + G\dot{x} + Kx - Bu) = \hat{M}\ddot{x} + \hat{G}\dot{x} + \hat{K}x - \hat{B}u.$$

The concept of opu-equivalence requires that the order of differentiation in  $x, u$  does not increase. In section 4 we will make use of analogous transformations which do not increase the order of differentiation in  $x$  but allow derivatives of the input function  $u$  to be introduced. To distinguish these two types of transformations we call the latter ones *state order-preserving unimodularly equivalences, or sopu-equivalences*.

Note that the set of consistent initial conditions is not altered by opu- and sopu-equivalence transformations, since the solution set is not altered.

In the following we will show that, as in the first order case, regularization and index reduction can be obtained via a combination of unimodular equivalence transformations and appropriate feedback transformations.

DEFINITION 2.3. *Systems  $M\ddot{x} + G\dot{x} + Kx = Bu$  and  $M\ddot{x} + \hat{G}\dot{x} + \hat{K}x = B\hat{u}$  are called equivalent under proportional feedback if there exists a matrix  $F_0$  of appropriate dimension such that  $\hat{K} = K + BF_0$ .*

*They are called equivalent under first order derivative feedback if there exists a matrix  $F_1$  of appropriate dimension such that  $\hat{G} = G + BF_1$ .*

After introducing the definitions, we now describe a condensed form under strong equivalence.

THEOREM 2.4. Consider the system (1.1). Then there exist nonsingular matrices  $P, Q \in \mathbb{R}^{n,n}$  such that the coefficients in the transformed system

$$(2.2) \quad \hat{M}\ddot{\hat{x}} + \hat{G}\dot{\hat{x}} + \hat{K}\hat{x} - \hat{B}u = PMQ\ddot{\hat{x}} + PGQ\dot{\hat{x}} + PKQ\hat{x} - PBu$$

that are obtained by setting  $x = Q\hat{x}$  have the following form:

$$(2.3) \quad \begin{pmatrix} \begin{bmatrix} I_{s^{(0,1,2)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{s^{(1,2)}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{s^{(0,2)}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{d^{(2)}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \\ \begin{bmatrix} 0 & 0 & * & * & 0 & 0 & * & * \\ 0 & 0 & * & * & 0 & 0 & * & * \\ 0 & 0 & * & * & 0 & 0 & * & * \\ 0 & 0 & * & * & 0 & 0 & * & * \\ 0 & 0 & 0 & * & I_{s^{(0,1)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{d^{(1)}} & 0 & 0 \\ I_{s^{(0,1,2)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{s^{(1,2)}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \\ \begin{bmatrix} 0 & * & 0 & * & 0 & * & 0 & * \\ 0 & * & 0 & * & 0 & * & 0 & * \\ 0 & * & 0 & * & 0 & * & 0 & * \\ 0 & * & 0 & * & 0 & * & 0 & * \\ 0 & * & 0 & * & 0 & * & 0 & * \\ 0 & * & 0 & * & 0 & * & 0 & * \\ 0 & * & 0 & * & 0 & * & 0 & * \\ 0 & * & 0 & * & 0 & * & 0 & * \\ 0 & 0 & 0 & 0 & 0 & 0 & I_a & 0 \\ 0 & 0 & 0 & 0 & I_{s^{(0,1)}} & 0 & 0 & 0 \\ 0 & 0 & I_{s^{(0,2)}} & 0 & 0 & 0 & 0 & 0 \\ I_{s^{(0,1,2)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \hat{B}_3 \\ \hat{B}_4 \\ \hat{B}_5 \\ \hat{B}_6 \\ \hat{B}_7 \\ \hat{B}_8 \\ \hat{B}_9 \\ \hat{B}_{10} \\ \hat{B}_{11} \\ \hat{B}_{12} \\ \hat{B}_{13} \end{bmatrix} \end{pmatrix},$$

where  $s^{(0,1,2)}, s^{(1,2)}, s^{(0,2)}, s^{(0,1)}, d^{(2)}, d^{(1)}, a$ , and  $v$  are nonnegative integers and the blocks denoted by  $*$  are not specified.

*Proof.* This result follows directly from Theorem 12 in [35] with  $f = Bu$ .  $\square$

Based on Theorem 2.4 we can then show the following result.

THEOREM 2.5. *Consider system (1.1)–(1.2). Then there exists a sequence of strong and opu-equivalence transformations such that the transformed system*

$$\hat{M}\ddot{\hat{x}} + \hat{G}\dot{\hat{x}} + \hat{K}\hat{x} = \hat{B}\hat{u}$$

*has the form*

$$\begin{aligned} & \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \ddot{\hat{x}} + \begin{bmatrix} \hat{G}_{11} & 0 & \hat{G}_{13} & 0 \\ 0 & I & 0 & 0 \\ \hat{G}_{31} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \dot{\hat{x}} + \begin{bmatrix} \hat{K}_{11} & \hat{K}_{12} & \hat{K}_{13} & 0 \\ \hat{K}_{21} & \hat{K}_{22} & \hat{K}_{23} & 0 \\ \hat{K}_{31} & \hat{K}_{32} & \hat{K}_{33} & 0 \\ \hat{K}_{41} & \hat{K}_{42} & \hat{K}_{43} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \hat{x} \\ &= \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \hat{B}_3 \\ \hat{B}_4 \\ \hat{B}_5 \\ 0 \end{bmatrix} \hat{u}, \\ & y = [\hat{C}_1 \quad \hat{C}_2 \quad \hat{C}_3 \quad \hat{C}_4] \hat{x}, \end{aligned}$$

where

$$\hat{x} = [\hat{x}_1^T \quad \hat{x}_2^T \quad \hat{x}_3^T \quad \hat{x}_4^T]^T,$$

and, furthermore,  $\hat{B}_3, \hat{B}_4$ , and  $\hat{B}_5$  have full row rank.

*Proof.* A detailed constructive proof is given in Appendix A of [32].  $\square$

THEOREM 2.6. *Consider system (1.1)–(1.2). Then there exists a sequence of strong and opu-equivalence transformations, as well as proportional feedbacks and first order derivative feedbacks such that the transformed system*

$$\hat{M}\ddot{\hat{x}} + \hat{G}\dot{\hat{x}} + \hat{K}\hat{x} = \hat{B}\hat{u}$$

*has the form*

$$\begin{aligned} & \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \ddot{\hat{x}} + \begin{bmatrix} \hat{G}_{11} & 0 & \hat{G}_{13} & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \dot{\hat{x}} + \begin{bmatrix} \hat{K}_{11} & \hat{K}_{12} & \hat{K}_{13} & 0 \\ \hat{K}_{21} & \hat{K}_{22} & \hat{K}_{23} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \hat{x} \\ &= \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \hat{B}_3 \\ 0 \end{bmatrix} \hat{u}, \\ (2.4) \quad y &= [\hat{C}_1 \quad \hat{C}_2 \quad \hat{C}_3 \quad \hat{C}_4] \hat{x}, \end{aligned}$$

where

$$\hat{x} = [\hat{x}_1^T \quad \hat{x}_2^T \quad \hat{x}_3^T \quad \hat{x}_4^T]^T,$$

and  $\hat{B}_3$  has full row rank.

*Proof.* A detailed constructive proof is given in Appendix B of [32].  $\square$

*Remark 2.7.* It is important to note that, in general, a combination of all of the described types of transformations in Theorem 2.6 is needed to achieve the condensed form (2.4). Furthermore, even though the proofs to Theorems 2.5 and 2.6 as well as that of the condensed form (2.3) are constructive (see [32, 35]), in general they cannot be implemented in a numerically reliable way. As an alternative way, for matrix pencils, staircase algorithms have been constructed that determine the structural information in the condensed forms via orthogonal transformations; see, e.g., [13, 42].

We can use the normal form (2.4) to derive a first order formulation which, as we will show later, avoids the difficulties of other first order formulations.

**COROLLARY 2.8.** *Consider system (1.1)–(1.2). Then there exists a bijective map between the solutions of (1.1) and the components  $\xi_2, \dots, \xi_5$  of the first order system  $\hat{E}\dot{\xi} = \hat{A}\xi + \hat{B}_1\hat{u}$  given by*

$$(2.5) \quad \begin{bmatrix} I & \hat{G}_{11} & 0 & \hat{G}_{13} & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \end{bmatrix} \dot{\xi} = \begin{bmatrix} 0 & -\hat{K}_{11} & -\hat{K}_{12} & -\hat{K}_{13} & 0 \\ 0 & -\hat{K}_{21} & -\hat{K}_{22} & -\hat{K}_{23} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -I \\ I & 0 & 0 & 0 & 0 \end{bmatrix} \xi + \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \hat{B}_3 \\ 0 \\ 0 \end{bmatrix} \hat{u},$$

$$(2.5) \quad y = \begin{bmatrix} 0 & \hat{C}_1 & \hat{C}_2 & \hat{C}_3 & \hat{C}_4 \end{bmatrix} \xi,$$

where  $\hat{B}_3$  has full row rank,

$$\xi = \begin{bmatrix} \xi_1^T & \xi_2^T & \xi_3^T & \xi_4^T & \xi_5^T \end{bmatrix}^T = \begin{bmatrix} \hat{x}_1^T & \hat{x}_1^T & \hat{x}_2^T & \hat{x}_3^T & \hat{x}_4^T \end{bmatrix}^T,$$

and  $\hat{x} = [\hat{x}_1^T \ \hat{x}_2^T \ \hat{x}_3^T \ \hat{x}_4^T]^T$  is a solution of (2.4).

Furthermore, the first order system (2.5) is I-controllable.

*Proof.* By solving for  $\xi_1$  in the last block row of (2.5) we obtain (2.4), which is equivalent to (1.1)–(1.2).

The I-controllability of (2.5) follows immediately from the definition, since in this case

$$[E, AS_\infty(E), B] = \left[ \begin{array}{ccccc|cc|c} I & \hat{G}_{11} & 0 & \hat{G}_{13} & 0 & -\hat{K}_{13} & 0 & \hat{B}_1 \\ 0 & 0 & I & 0 & 0 & -\hat{K}_{23} & 0 & \hat{B}_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & -I & 0 \\ 0 & I & 0 & 0 & 0 & -\hat{G}_{13} & 0 & 0 \end{array} \right]$$

has full rank.  $\square$

*Remark 2.9.* The bijectivity of the map in Corollary 2.8 follows from the fact that strong equivalence transformations form a bijection and that the linear combination of derivatives of equations that do contain the output (opu-equivalence) does not change the solution sets of classical solutions. Note that the relationship between  $u$  and  $\hat{u}$  is just a change of basis.

If the system is considered in the distributional setting, then one has to be careful with opu-equivalences, since then the impulse order may change but the smooth parts of the solution are still mapped in a bijective way. See [17, 30] for a detailed discussion of this issue.

Let us illustrate these results with some examples.

*Example 2.10.* Consider the artificial second order descriptor system (1.1) with

$$M = 0, \quad G = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad K = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Since this is in fact a first order system, we can check its I-controllability using condition **C2** with  $E = G$ ,  $A = -K$ ,  $B_1 = B$  and see that

$$\text{rank} \begin{bmatrix} E & AS_\infty(E) & B \end{bmatrix} = \text{rank} \left[ \begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & -1 \end{array} \right] = 2.$$

But if we perform the classical transformation to first order, then we obtain

$$\tilde{E} = \left[ \begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right], \quad \tilde{A} = \left[ \begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \end{array} \right], \quad \tilde{B} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

In this case

$$\begin{bmatrix} \tilde{E} & \tilde{A}S_\infty(\tilde{E}) & \tilde{B} \end{bmatrix} = \left[ \begin{array}{cccc|cc|c} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right]$$

does not have full row rank, and hence the system is not I-controllable.

In this system we can easily reduce the classical first order formulation to one which is I-controllable by carrying out an index reduction procedure on the first order formulation.

The previous example seems artificial, but a similar phenomenon arises for constrained mechanical systems.

*Example 2.11.* Consider again the system (1.12). The first order version (1.13) yields

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} -1 & -1 & -1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

which is obviously not I-controllable.

If, however, we use the construction to the normal form (2.4), then we obtain a system with

$$\hat{M} = 0, \quad \hat{G} = 0, \quad \hat{K} = K, \quad \hat{B} = B,$$

and this system is I-controllable.

We see from these examples that the choice of the first order formulation is important. While the classical first order formulation of the system may not be I-controllable, the normal form (2.4) allows one to obtain a first order formulation that is I-controllable.

*Remark 2.12.* A natural question that arises is whether we could not just formally first introduce a first order formulation and then perform index reduction via transformation to normal form, including opu-equivalence transformations. This is indeed possible if the original triple is regular, i.e.,  $\det(\lambda^2 M + \lambda G + K) \neq 0$ , since

it is then known that the length of chains associated with the eigenvalue  $\infty$  is kept invariant under the classical companion formulation [20]. This is not true any longer if  $\det(\lambda^2 M + \lambda G + K)$  vanishes identically as the following example of [8] shows. For the singular matrix polynomial

$$P(\lambda) = \begin{bmatrix} \lambda^2 + \lambda & 0 \\ 1 & 0 \end{bmatrix},$$

the right null space is  $x(\lambda) = e_2$ , which creates a chain of length 1, and the left null space is  $y(\lambda) = [-1 \quad \lambda^2 + \lambda]$ , which gives  $y_0 = -e_1$ ,  $y_1 = e_2$ , and  $y_2 = e_2$ , and thus the chain has length 3.

By considering the first companion linearization, we get

$$L(\lambda) = \lambda \left[ \begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array} \right] + \left[ \begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{array} \right].$$

The right and left null space vector polynomials are

$$x(\lambda) = \begin{bmatrix} 0 \\ \lambda \\ 0 \\ 1 \end{bmatrix}, \quad y(\lambda) = \begin{bmatrix} 1 \\ -\lambda^2 - \lambda \\ \lambda + 1 \\ 0 \end{bmatrix},$$

and clearly the right chain does not have the same length as in the original matrix polynomial.

Furthermore, performing a first order formulation first may substantially change the sensitivity of the problem, in particular when computing the index reduction; see [35] for an illustrative example.

Thus it is generally preferable to perform index reductions and reformulations on the original data which is the second order pencil.

In the next section we will show how we can use the condensed forms of this section to check the different controllability and observability conditions directly for second order systems and how to derive first order formulations that preserve these conditions.

**3. Controllability for second order systems.** For a descriptor system (1.1)–(1.3), the following definitions extend the concepts of C-controllability and C-observability to second order descriptor systems.

**DEFINITION 3.1.** Consider a system as in (1.1)–(1.3). A set  $\mathcal{R} \subseteq \mathbb{R}^n$  is called reachable from  $x_0, \dot{x}_0$  if for every  $x_f \in \mathcal{R}$  there exists a piecewise continuous input function  $u$  that transfers the system in finite time from  $x(t_0) = x_0$  to  $x_f$ .

A set  $\mathcal{R} \subseteq \mathbb{R}^n \times \mathbb{R}^n$  is called  $\mathcal{R}2$ -reachable from  $x_0, \dot{x}_0$  if for every  $x_f, \dot{x}_f \in \mathcal{R}$  there exists a piecewise continuous input function  $u$  that transfers the system in finite time from  $x(t_0) = x_0, \dot{x}(t_0) = \dot{x}_0$  to  $x_f, \dot{x}_f$ . The system is called

- (i) C-controllable if for any  $x_0$  and  $\dot{x}_0$  and any  $x_f \in \mathbb{R}^n$  there exist a time  $t_f$  and a piecewise continuous input function  $u : [t_0, t_f] \rightarrow \mathbb{R}^m$  such that  $x(t_f) = x_f$ ;
- (ii) strongly C2-controllable if for any  $x_0, \dot{x}_0$  and any  $x_f, \dot{x}_f \in \mathbb{R}^n$  there exist a time  $t_f$  and a piecewise continuous input function  $u : [t_0, t_f] \rightarrow \mathbb{R}^m$  such that  $x(t_f) = x_f, \dot{x}(t_f) = \dot{x}_f$ ;



- (iii)  $\mathcal{R}$ -controllable if any state  $x_f$  in the reachable set  $\mathcal{R}$  can be reached from any admissible  $x_0, \dot{x}_0$  in finite time;
- (iv)  $\mathcal{R}2$ -controllable if any state and derivative  $(x_f, \dot{x}_f)$  in the  $\mathcal{R}2$ -reachable set can be reached from any admissible  $x_0, \dot{x}_0$  in finite time.

By using the normal form (2.4) we get a variation of C2-controllability, which is better adapted to the problem.

**DEFINITION 3.2.** *A system in normal form (2.4) is called C2-controllable if for any  $\hat{x}(0), \dot{\hat{x}}_1(0)$  and any  $\hat{x}_f \in \mathbb{R}^n, \hat{x}_{1,f} \in \mathbb{R}^{\dim(\hat{x}_1)}$  there exists a time  $t_f$  and a piecewise continuous input function  $u : [t_0, t_f] \rightarrow \mathbb{R}^m$  such that  $\hat{x}(t_f) = \hat{x}_f, \dot{\hat{x}}(t_f) = \dot{\hat{x}}_{1,f}$ .*

We immediately see that a strongly C2-controllable second order descriptor system is also C2-controllable, that a C2-controllable second order descriptor system is also C-controllable, and that an  $\mathcal{R}2$ -controllable second order descriptor system is also  $\mathcal{R}$ -controllable.

For the analysis of controllability conditions let us first discuss the case that  $M$  is invertible; i.e., we have an implicitly defined second order ordinary differential equation. Then it is known that C-controllability is equivalent to C2-, strong C2-,  $\mathcal{R}2$ -, and  $\mathcal{R}$ -controllability, and all five are characterized by the Hautus criterion [1, 24],

$$(3.1) \quad \text{rank} \begin{bmatrix} \lambda^2 M + \lambda G + K & B \end{bmatrix} = n \text{ for all } \lambda \in \sigma(M, G, K),$$

where  $\sigma(M, G, K)$  denotes the spectrum of the matrix polynomial  $P(\lambda) = \lambda^2 M + \lambda G + K$ , i.e., the roots of  $\det P(\lambda)$ .

Strong C2-controllability and  $\mathcal{R}2$ -controllability is trivially characterized via the classical (companion) first order form.

**COROLLARY 3.3.** *Consider a second order descriptor system (1.1) and its classical (companion) first order form*

$$(3.2) \quad \begin{bmatrix} M & 0 \\ 0 & I \end{bmatrix} \dot{z} = \begin{bmatrix} -G & -K \\ I & 0 \end{bmatrix} z + \begin{bmatrix} B \\ 0 \end{bmatrix} u.$$

- (i) *System (1.1) is strongly C2-controllable if and only if (3.2) is C-controllable.*
- (ii) *System (1.1) is  $\mathcal{R}2$ -controllable if and only if (3.2) is  $\mathcal{R}$ -controllable.*

To characterize the other controllability conditions for second order descriptor systems, we make use of the condensed forms of section 2. From (2.4) we see that, for a consistent initial condition in the variables that occurs only in first order, we can prescribe only initial values and not initial derivatives. This immediately implies the following corollary.

**COROLLARY 3.4.**

- (i) *A second order system in normal form (2.4) is C2-controllable if and only if the associated first order system (2.5) is C-controllable.*
- (ii) *A second order system in normal form (2.4) is  $\mathcal{R}2$ -controllable if and only if the associated first order system (2.5) is  $\mathcal{R}$ -controllable.*

We can illustrate the difference between strong C2-controllability and C2-controllability by the following example.

**Example 3.5.** Consider the following C2-controllable second order system in normal form (2.4):

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \ddot{x} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \dot{x} + \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} u.$$

The first order formulation given in Corollary 2.8 is given by

$$\left[ \begin{array}{c|cc} 1 & 0 & 0 \\ 0 & 0 & 1 \\ \hline 0 & 1 & 0 \end{array} \right] \dot{\xi} = \left[ \begin{array}{c|cc} 0 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 1 & 0 & 0 \end{array} \right] \xi + \left[ \begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right] u.$$

This system is C-controllable, since

$$\text{rank} \left[ \begin{array}{cc|c} \alpha E - \beta A & B_1 \end{array} \right] = \left[ \begin{array}{ccc|c} \alpha & 0 & 0 & 1 \\ 0 & -\beta & \alpha & 0 \\ -\beta & \alpha & 0 & 0 \end{array} \right] = 3 \text{ for all } (\alpha, \beta) \in \mathbb{C}^2 \setminus \{(0, 0)\}.$$

But the classical first order formulation

$$\left[ \begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array} \right] \dot{z} = \left[ \begin{array}{cc|cc} 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ \hline -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{array} \right] z + \left[ \begin{array}{c} 1 \\ 0 \\ 0 \\ 0 \end{array} \right] u$$

is not C-controllable, since

$$\text{rank} \left[ \begin{array}{cc|ccc} \alpha & 0 & 0 & 0 & 1 \\ 0 & -\beta & \beta & 0 & 0 \\ \beta & 0 & \alpha & 0 & 0 \\ 0 & \beta & 0 & \alpha & 0 \end{array} \right] \neq 4 \text{ for } \beta = 0,$$

and hence the original system is not strongly C2-controllable. It is easily checked that both first order formulations are  $\mathcal{R}$ -controllable.

In order to study I-controllability we will make use of different types of feedback. DEFINITION 3.6. Consider a second order descriptor system (1.1)–(1.2).

- (i) The system is called proportionally I-controllable if there exists a state feedback  $u = \hat{u} - F_0 x$  such that the resulting system with coefficients  $(M, G, K + BF_0)$  is regular and of index at most one.
- (ii) The system is called differentially I-controllable if there exists a first order derivative feedback  $u = \hat{u} - F_1 \dot{x}$  such that the resulting system with coefficients  $(M, G + BF_1, K)$  is regular and of index at most one.
- (iii) The system is called proportionally and differentially I-controllable or just I-controllable if there exist a proportional and a first order derivative feedback  $u = \hat{u} - F_0 x - F_1 \dot{x}$  such that the resulting system given by  $(M, G + BF_1, K + BF_0)$  is regular and of index at most one.

It is straightforward to show that strong equivalence transformations preserve all types of controllability for second order descriptor systems.

The same is true for proportional and first order derivative feedback. On the other hand, opu-equivalence transformations preserve C-, C2-, and strong C2-controllability as well as  $\mathcal{R}$ - and  $\mathcal{R}2$ -controllability but may turn a system that is not I-controllable into one that is I-controllable.

Example 3.7. Consider the first order descriptor system (1.4) given by

$$\left[ \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right] \dot{\xi} + \left[ \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{array} \right] \xi = \left[ \begin{array}{c} 1 \\ 1 \\ 0 \end{array} \right] u.$$

The system is not I-controllable, since

$$\text{rank} \begin{bmatrix} E & AS_\infty(E) & B_1 \end{bmatrix} = \text{rank} \left[ \begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right] = 2,$$

but if we apply the opu-equivalence transformation of multiplying by

$$P = \begin{bmatrix} 1 & 0 & -\frac{d}{dt} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

from the left, then we obtain the I-controllable system

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \dot{\xi} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \xi = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} u.$$

As we will see below, a combination of opu-equivalence transformations and proportional and first order derivative feedbacks together always allows one to make a second order system regular and of index at most one, which then implies I-controllability. See also [30, 31] for similar results in the first order case.

In the following we derive algebraic characterizations for the different controllability conditions. We begin with systems in normal form (2.4).

**THEOREM 3.8.** *Consider a second order descriptor system (1.1) in normal form (2.4), and let (2.5) be the first order system derived from this normal form.*

- (i) *The first order system (2.5) is  $\mathcal{R}$ -controllable if the system matrices of the normal form (2.4) satisfy (3.1).*
- (ii) *System (2.5) is I-controllable.*
- (iii) *System (2.5) is C-controllable if and only if it is  $\mathcal{R}$ -controllable and the 4th row in (2.4) is void.*

*Proof.* Let  $n_1$  be the size of the component  $\xi_1$ . To see that (i) holds, we observe that (2.5) is  $\mathcal{R}$ -controllable if and only if

$$\text{rank} \begin{bmatrix} \lambda I & \lambda \hat{G}_{11} - \hat{K}_{11} & -\hat{K}_{12} & \lambda \hat{G}_{13} - \hat{K}_{13} & 0 & \hat{B}_1 \\ 0 & -\hat{K}_{21} & \lambda I - \hat{K}_{22} & -\hat{K}_{23} & 0 & \hat{B}_2 \\ 0 & 0 & 0 & 0 & 0 & \hat{B}_3 \\ 0 & 0 & 0 & 0 & -I & 0 \\ -I & -\lambda I & 0 & 0 & 0 & 0 \end{bmatrix} = n + n_1,$$

and this is the case if and only if

$$\text{rank} \begin{bmatrix} 0 & -\lambda^2 I + \lambda \hat{G}_{11} - \hat{K}_{11} & -\hat{K}_{12} & \lambda \hat{G}_{13} - \hat{K}_{13} & 0 & \hat{B}_1 \\ 0 & -\hat{K}_{21} & \lambda I - \hat{K}_{22} & -\hat{K}_{23} & 0 & \hat{B}_2 \\ 0 & 0 & 0 & 0 & 0 & \hat{B}_3 \\ 0 & 0 & 0 & 0 & -I & 0 \\ I & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = n + n_1,$$

which holds if and only if

$$\text{rank} \begin{bmatrix} -\lambda^2 I + \lambda \hat{G}_{11} - \hat{K}_{11} & -\hat{K}_{12} & \lambda \hat{G}_{13} - \hat{K}_{13} & 0 & \hat{B}_1 \\ -\hat{K}_{21} & \lambda I - \hat{K}_{22} & -\hat{K}_{23} & 0 & \hat{B}_2 \\ 0 & 0 & 0 & 0 & \hat{B}_3 \\ 0 & 0 & 0 & -I & 0 \end{bmatrix} = n.$$

By comparison with (2.4) we see that this holds if and only if  $\text{rank}[-\lambda^2 M + \lambda G - K B] = n$  for all  $\lambda \in \mathbb{C}$ , which proves the assertion.

(ii) We first carry out a strong equivalence transformation by a change of basis that eliminates  $\hat{G}_{11}$ ,  $\hat{G}_{13}$  and turn (2.5) to the form

$$\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -I & 0 & 0 & 0 \end{bmatrix} \dot{\hat{\xi}} + \begin{bmatrix} 0 & \hat{K}_{11} & \hat{K}_{12} & \hat{K}_{13} & 0 \\ 0 & \hat{K}_{21} & \hat{K}_{22} & \hat{K}_{23} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I \\ I & \hat{G}_{11} & 0 & \hat{G}_{13} & 0 \end{bmatrix} \hat{\xi} = \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \hat{B}_3 \\ 0 \\ 0 \end{bmatrix} \hat{u}.$$

This system is I-controllable if the matrix

$$\begin{bmatrix} I & 0 & 0 & 0 & 0 & \hat{K}_{13} & 0 & \hat{B}_1 \\ 0 & 0 & I & 0 & 0 & \hat{K}_{23} & 0 & \hat{B}_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & I & 0 \\ 0 & -I & 0 & 0 & 0 & \hat{G}_{13} & 0 & 0 \end{bmatrix}$$

has full row rank. But this follows since  $\hat{B}_3$  has full row rank.

(iii) Since (see, e.g., [4]) a first order descriptor system is C-controllable if it is  $\mathcal{R}$ -controllable and  $[E \ B_1]$  has full row rank, we can just check this rank condition. In the given case this matrix has the form

$$\begin{bmatrix} I & \hat{G}_{11} & 0 & \hat{G}_{13} & 0 & \hat{B}_1 \\ 0 & 0 & I & 0 & 0 & \hat{B}_2 \\ 0 & 0 & 0 & 0 & 0 & \hat{B}_3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -I & 0 & 0 & 0 & 0 \end{bmatrix},$$

and since  $\hat{B}_3$  has full row rank, this matrix has full row rank if and only if the 4th block row is void. By considering system (2.5), we see that this holds if and only if the part  $\xi_5$  is void. But  $\xi_5 = \hat{x}_4$ , and thus we have finished the proof.  $\square$

**THEOREM 3.9.** *Consider a second order descriptor system (1.1) in normal form (2.4). Then the 4th component  $\hat{x}_4$  in (2.4) is void if and only if  $\text{rank}[M \ G \ B] = n$ .*

*Proof.* From the proof of Theorem 2.6 we see that the component  $\hat{x}_4$  is void if and only if there is no rank deficit in  $[\hat{B}_9^T \ \hat{B}_{10}^T \ \hat{B}_{11}^T \ \hat{B}_{12}^T \ \hat{B}_{13}^T]^T$ , with  $\hat{B}_9, \dots, \hat{B}_{13}$  as in (2.3).

It remains to show that this is equivalent to  $\text{rank}[M \ G \ B] = n$ . Since this rank is invariant under strong equivalence transformations, it remains to show that  $\text{rank}[\hat{M} \ \hat{G} \ \hat{B}] = n$  in (2.3) if and only if  $[\hat{B}_9^T \ \hat{B}_{10}^T \ \hat{B}_{11}^T \ \hat{B}_{12}^T \ \hat{B}_{13}^T]^T$  has full row

rank. But

$$\begin{aligned} & \text{rank} \begin{bmatrix} \hat{M} & \hat{G} & \hat{B} \end{bmatrix} \\ &= \text{rank} \begin{bmatrix} I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & 0 & 0 & * & * & \hat{B}_1 \\ 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & 0 & 0 & * & * & \hat{B}_2 \\ 0 & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & 0 & 0 & * & * & \hat{B}_3 \\ 0 & 0 & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & 0 & 0 & * & * & \hat{B}_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & I & 0 & 0 & 0 & \hat{B}_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & 0 & 0 & \hat{B}_6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{10} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{13} \end{bmatrix} \\ &= \text{rank} \begin{bmatrix} I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{10} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{13} \end{bmatrix}, \end{aligned}$$

and hence the assertion follows.  $\square$

**THEOREM 3.10.** *Consider a second order descriptor system (1.1) and the corresponding classical (companion) first order form (3.2). Then (3.2) is C-controllable if and only if (1.1) is  $\mathcal{R}2$ -controllable and  $\text{rank}[M, B] = n$ .*

*Proof.* System (3.2) is C-controllable if and only if

$$\text{rank} \begin{bmatrix} \alpha M - \beta G & -\beta K & B \\ \beta I & \alpha I & 0 \end{bmatrix} = 2n.$$

Setting  $\beta = 0$  gives  $\text{rank}[M, B] = n$ , and  $\beta \neq 0$  gives the  $\mathcal{R}2$ -controllability condition.  $\square$

Obviously, for a second order descriptor system (1.1),  $\text{rank}[\lambda^2 M + \lambda G + K \quad B]$  is invariant under strong equivalence transformations, proportional and first order derivative feedback, and opu-equivalence transformations. Thus, we can combine these results with Corollaries 3.3 and 3.4.

**COROLLARY 3.11.** *A second order descriptor system of the form (1.1) is*

(i)  $\mathcal{R}2$ -controllable if and only if

$$\text{rank} \begin{bmatrix} \lambda^2 M + \lambda G + K & B \end{bmatrix} = n \quad \text{for all } \lambda \in \mathbb{C};$$

(ii)  $\mathcal{C}2$ -controllable if and only if it is  $\mathcal{R}2$ -controllable and

$$\text{rank} \begin{bmatrix} M & G & B \end{bmatrix} = n;$$

(iii) *strongly C2-controllable if and only if it is R2-controllable and*

$$\text{rank} \begin{bmatrix} M & B \end{bmatrix} = n.$$

Let us illustrate this result with an example.

*Example 3.12.* By continuing with the data of Example 2.11, we obtain

$$\text{rank} \begin{bmatrix} \lambda^2 M + \lambda G + K & B \end{bmatrix} = \text{rank} \begin{bmatrix} \lambda^2 + \lambda + 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} = 2 \text{ for all } \lambda \in \mathbb{C}.$$

Hence, the system is R2-controllable, while

$$\text{rank} \begin{bmatrix} M & G & B \end{bmatrix} = \text{rank} \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = 1,$$

and thus the system is not C2-controllable.

To characterize I-controllability we use the condensed form (2.3) which can be obtained by using only strong equivalence transformations.

**THEOREM 3.13.** *Consider a second order descriptor system (1.1) in condensed form (2.3). The system is*

- (i) *proportionally I-controllable if and only if in (2.3) the 7th and 8th block rows are void and the matrix  $[\hat{B}_{10}^T \dots \hat{B}_{13}^T]^T$  has full row rank;*
- (ii) *first order derivative I-controllable if and only if in (2.3) the 10th to 12th block rows are void and the matrix  $[\hat{B}_7^T \hat{B}_8^T \hat{B}_{13}^T]^T$  has full row rank;*
- (iii) *proportional and first order derivative I-controllable if and only if in (2.3) the matrix  $[\hat{B}_7^T \hat{B}_8^T \hat{B}_{10}^T \dots \hat{B}_{13}^T]^T$  has full row rank.*

*Proof.* From the proof of Theorem 2.6 we observe the following:

(a) If in (2.3) the 7th and 8th block rows are void, then we do not need a first order derivative feedback to make the system regular and of index at most one. If these are not void, then proportional feedback is not enough to achieve this.

(b) Similarly, if in (2.3) the 10th to 12th block rows are void, then we do not need a proportional feedback to make the system regular and of index at most one. If these are not void, then first order derivative feedback is not enough to achieve this.

(c) If in (2.3) the matrix

$$\begin{bmatrix} \hat{B}_7^T & \hat{B}_8^T & \hat{B}_{10}^T & \dots & \hat{B}_{13}^T \end{bmatrix}^T$$

has full row rank, then we do not need opu-equivalence transformations to make the system regular and of index at most one. If there is a rank deficit, then proportional and first order derivative feedback is not sufficient to make the system regular and of index at most one.

Then with (c) we obtain (iii), with (a) and (c) we obtain (i), and with (b) and (c) we get (ii).  $\square$

Theorem 3.13 shows that the condensed form (2.3) and the canonical form (2.4) allow one to check the different controllability properties for second order descriptor systems. For mathematical elegance and a simpler description it would also be nice to have a coordinate-free algebraic characterization. This is given in the following theorem.

**THEOREM 3.14.** *Consider a second order descriptor system (1.1) and its condensed form (2.3), and let  $s^{(0,1,2)}$ ,  $s^{(0,2)}$ ,  $s^{(0,1)}$ , and  $s^{(1,2)}$  be the integer quantities defined in Theorem 2.4. Then the system is*

(i) *proportionally and first order derivative I-controllable if and only if*

$$(3.3) \quad \text{rank} \begin{bmatrix} M & GS_\infty^1 & KS_\infty^2 & B \end{bmatrix} = n,$$

where the columns of the matrix  $S_\infty^1$  form a basis of  $\ker M$ , the columns of  $S_\infty^2$  form a basis of

$$\ker \begin{bmatrix} M \\ Z_1^T G \end{bmatrix} \setminus \ker \begin{bmatrix} M \\ Z_3^T G \\ Z_3^T K \end{bmatrix},$$

the columns of  $Z_1$  form a basis of  $\ker M^T$ , and those of  $Z_3$  form a basis of  $\ker \begin{bmatrix} M^T \\ G^T \end{bmatrix}$ ;

(ii) *proportionally I-controllable if and only if it satisfies (i) and  $s^{(0,1,2)} = s^{(1,2)} = 0$ ;*

(iii) *first order derivative I-controllable if and only if it satisfies (i) and  $s^{(0,1,2)} = s^{(0,2)} = s^{(0,1)} = 0$ .*

*Proof.* (i) In the condensed form (2.3) we have

$$\begin{bmatrix} M & GS_\infty^1 & KS_\infty^2 & B \end{bmatrix} = \left[ \begin{array}{cccccccc|cccc|c|c} I_{s^{(0,1,2)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & 0 & \hat{B}_1 \\ 0 & I_{s^{(1,2)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & 0 & \hat{B}_2 \\ 0 & 0 & I_{s^{(0,2)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & 0 & \hat{B}_3 \\ 0 & 0 & 0 & I_{d^{(2)}} & 0 & 0 & 0 & 0 & 0 & 0 & * & * & 0 & \hat{B}_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{s^{(0,1)}} & 0 & 0 & 0 & 0 & \hat{B}_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{d^{(1)}} & 0 & 0 & 0 & \hat{B}_6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_a & \hat{B}_9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{10} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{B}_{13} \end{array} \right].$$

Thus,  $\text{rank}[M \ GS_\infty^1 \ KS_\infty^2 \ B] = n$  if and only if the matrix  $\text{sfif}[\hat{B}_7^T \ \hat{B}_8^T \ \hat{B}_{10}^T \ \dots \ \hat{B}_{13}^T]^T$  has full row rank. Then by Theorem 3.13(iii) the system is proportionally and first order derivative I-controllable.

It remains to show that  $\text{rank} \begin{bmatrix} M & GS_\infty^1 & KS_\infty^2 & B \end{bmatrix}$  is invariant under strong equivalence. For this let

$$\tilde{M} = PMQ, \quad \tilde{G} = PGQ, \quad \tilde{K} = PKQ, \quad \tilde{B} = PBV,$$

and let  $\tilde{Z}_1, \tilde{Z}_3, \tilde{S}_\infty^1, \tilde{S}_\infty^2$  be the corresponding subspaces. Since  $\tilde{M}v = 0$  if and only if  $PMQv = 0$  if and only if  $MQv = 0$ , we obtain  $Q\tilde{S}_\infty^1 = S_\infty^1$  and analogously in Theorem  $P^T\tilde{Z}_1 = Z_1$  and  $P^T\tilde{Z}_3 = Z_3$ . Since, furthermore,

$$\begin{bmatrix} \tilde{M} \\ \tilde{Z}_1^T \tilde{G} \end{bmatrix} v = 0 \Leftrightarrow \begin{bmatrix} PMQ \\ Z_1^T P^{-1} PGQ \end{bmatrix} v = 0 \Leftrightarrow \begin{bmatrix} M \\ Z_1^T G \end{bmatrix} Qv = 0$$

and

$$\begin{bmatrix} \tilde{M} \\ \tilde{Z}_1^T \tilde{G} \\ \tilde{Z}_3^T \tilde{K} \end{bmatrix} v = 0 \Leftrightarrow \begin{bmatrix} PMQ \\ Z_1^T P^{-1} PGQ \\ Z_3^T P^{-1} PKQ \end{bmatrix} v = 0 \Leftrightarrow \begin{bmatrix} M \\ Z_1^T G \\ Z_3^T K \end{bmatrix} Qv = 0,$$

we have  $Q\tilde{S}_\infty^2 = S_\infty^2$ . Thus, altogether we have

$$\begin{aligned} \text{rank}[\tilde{M}, \tilde{G}\tilde{S}_\infty^1, \tilde{K}\tilde{S}_\infty^2, \tilde{B}] \\ = \text{rank}[PMQ, PGQQ^{-1}S_\infty^1, PKQQ^{-1}S_\infty^2, PBV] \\ = \text{rank}[M, GS_\infty^1, KS_\infty^2, B]. \end{aligned}$$

This finishes the proof of (i). Parts (ii) and (iii) then follow from Theorem 3.13.  $\square$

*Remark 3.15.* If in Theorem 3.14 we have  $M = 0$ , then  $S_\infty^1 = I$ ,  $Z_1 = I$ ,  $Z_3$  is a basis of  $\text{kernel } G^T$ , and  $S_\infty^2$  is a basis of

$$\text{kernel } G \setminus \text{kernel} \begin{bmatrix} G \\ Z_3 K \end{bmatrix}.$$

Thus,  $\text{rank}[M \quad GS_\infty^1 \quad KS_\infty^2 \quad B] = \text{rank}[G \quad KS_\infty^2 \quad B]$ . In this case, the condensed form is

$$\left( 0, \begin{bmatrix} I_{s^{(0,1)}} & 0 & 0 & 0 \\ 0 & I_{d^{(1)}} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & * & 0 & * \\ 0 & * & 0 & * \\ 0 & 0 & I_a & 0 \\ I_{s^{(0,1)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{B}_5 \\ \hat{B}_6 \\ \hat{B}_9 \\ \hat{B}_{10} \\ \hat{B}_{13} \end{bmatrix} \right),$$

and thus

$$\begin{aligned} \text{rank} \begin{bmatrix} G & KS_\infty^2 & B \end{bmatrix} &= \text{rank} \left[ \begin{array}{cccc|c} I_{s^{(0,1)}} & 0 & 0 & 0 & 0 \\ 0 & I_{d^{(1)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_a \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \begin{array}{c} \hat{B}_5 \\ \hat{B}_6 \\ \hat{B}_9 \\ \hat{B}_{10} \\ \hat{B}_{13} \end{array} \right] \\ &= \text{rank} \begin{bmatrix} G & KS_\infty(G) & B \end{bmatrix}. \end{aligned}$$

This shows that Theorem 3.14 is a direct generalization of the I-controllability results for first order systems.

*Example 3.16.* By continuing with Example 2.10, we obtain that the system is proportionally and first order derivative I-controllable if and only if  $\text{rank}[G \quad KS_\infty(G) \quad B] = n$ , which we have seen already. Since  $M = 0$  we have  $s^{(0,1,2)} = s^{(1,2)} = s^{(0,2)} = 0$ , and, thus, the system is proportionally I-controllable as well as first order derivative I-controllable.

This example also demonstrates that condition (3.3) in Theorem 3.14(i) is not equivalent to condition **C2** in (1.14), since the classical first order companion form does not satisfy **C2** but (3.3). Other examples with  $M \neq 0$  are easily constructed.

*Example 3.17.* In Example 1.2 we have

$$GS_\infty^1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad Z_1^T G = \begin{bmatrix} 0 & 0 \end{bmatrix}, \quad Z_3^T K = \begin{bmatrix} H_1 & 0 \end{bmatrix},$$

and

$$\text{kernel} \begin{bmatrix} M \\ Z_1^T G \end{bmatrix} \setminus \text{kernel} \begin{bmatrix} M \\ Z_1^T G \\ Z_3^T K \end{bmatrix} = \emptyset.$$

Then  $\text{rank}[M \quad GS_\infty^1 \quad KS_\infty^2 \quad B] = 3 < n = 5$ ; i.e., the system is not I-controllable.



We also have similar  $\mathcal{R}$ -controllability.

**THEOREM 3.18.** *Consider a second order descriptor system (1.1) and its first order formulation (2.5). Let  $\mathcal{R}$  be the reachable set of (2.5), and let*

$$E_0 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix}$$

be partitioned as  $\hat{E}$  in (2.5). Then the following are equivalent:

- (i) The system is  $\mathcal{C}$ -controllable.
- (ii) In the first order formulation (2.5) for  $\xi_2(t_0), \dots, \xi_5(t_0)$  and  $\xi_{2f}, \dots, \xi_{5f}$ , there exist  $t_f$  and an input function  $u : [t_0, t_f] \rightarrow \mathbb{R}^m$  such that  $\xi_2(t_f) = \xi_{2f}, \dots, \xi_5(t_f) = \xi_{5f}$ .
- (iii) The system is  $\mathcal{R}$ -controllable and  $\text{Im}(E_0) \subset \mathcal{R}$ .
- (iv) The system is  $\mathcal{R}$ -controllable and  $\text{rank} \begin{bmatrix} M & G & B \end{bmatrix} = n$ .

*Proof.* The equivalence of (i) and (ii) is obvious. To prove the other equivalences, consider the first order system (2.5). By carrying out a strong equivalence transformation with

$$P = \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -I \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} I & -\hat{G}_{11} & 0 & -\hat{G}_{13} & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \end{bmatrix}$$

from left and right, respectively, we obtain the system

$$\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \dot{\hat{\xi}} + \begin{bmatrix} 0 & \hat{K}_{11} & \hat{K}_{12} & \hat{K}_{13} & 0 \\ -I & \hat{G}_{11} & 0 & \hat{G}_{13} & 0 \\ 0 & \hat{K}_{21} & \hat{K}_{22} & \hat{K}_{23} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix} \hat{\xi} = \begin{bmatrix} \hat{B}_1 \\ 0 \\ \hat{B}_2 \\ \hat{B}_3 \\ 0 \end{bmatrix} \hat{u},$$

where  $\xi = Q\hat{\xi}$ . Since  $\hat{B}_3$  has full row rank, we can compress its columns and eliminate with the full-rank part upwards. This gives the system

$$\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \dot{\hat{\xi}} + \begin{bmatrix} 0 & \hat{K}_{11} & \hat{K}_{12} & \hat{K}_{13} & 0 \\ -I & \hat{G}_{11} & 0 & \hat{G}_{13} & 0 \\ 0 & \hat{K}_{21} & \hat{K}_{22} & \hat{K}_{23} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix} \hat{\xi} = \begin{bmatrix} 0 & \tilde{B}_1 \\ 0 & 0 \\ 0 & \tilde{B}_2 \\ I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

By choosing a proportional feedback  $u_1 = v_1 - \hat{\xi}_4$ ,  $u_2 = v_2$ , which does not change the  $\mathcal{R}$ -controllability or the reachable set  $\mathcal{R}$ , we obtain a closed loop system  $E\dot{\hat{\xi}} + A\hat{\xi} = Bv$  of the form

$$\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \dot{\hat{\xi}} + \begin{bmatrix} 0 & \hat{K}_{11} & \hat{K}_{12} & \hat{K}_{13} & 0 \\ -I & \hat{G}_{11} & 0 & \hat{G}_{13} & 0 \\ 0 & \hat{K}_{21} & \hat{K}_{22} & \hat{K}_{23} & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix} \hat{\xi} = \begin{bmatrix} 0 & \tilde{B}_1 \\ 0 & 0 \\ 0 & \tilde{B}_2 \\ I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

By eliminating further in the second coefficient matrix, we get

$$\left[ \begin{array}{ccc|ccc} I & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right] \dot{\hat{\xi}} + \left[ \begin{array}{ccc|cc} 0 & \hat{K}_{11} & \hat{K}_{12} & 0 & 0 \\ -I & \hat{G}_{11} & 0 & 0 & 0 \\ 0 & \hat{K}_{21} & \hat{K}_{22} & 0 & 0 \\ \hline 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{array} \right] \hat{\xi} = \left[ \begin{array}{cc} -\hat{K}_{13} & \tilde{B}_1 \\ -\hat{G}_{13} & 0 \\ -\hat{K}_{23} & \tilde{B}_2 \\ \hline I & 0 \\ 0 & 0 \end{array} \right] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

This system has the form

$$(3.4) \quad \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} \dot{\hat{\xi}} + \begin{bmatrix} A_1 & 0 \\ 0 & I \end{bmatrix} \hat{\xi} = \begin{bmatrix} \bar{B}_1 \\ \bar{B}_2 \end{bmatrix} v.$$

By following [12], we can determine the reachable set as

$$\mathcal{R} = \mathbb{R}^{n_1} \oplus \mathcal{K}(N, \bar{B}_2),$$

where  $n_1 = \text{rank}(E)$ ,  $n_2 = n - n_1$  and

$$\mathcal{K}(N, \bar{B}_2) = \text{Im}[\bar{B}_2, N\bar{B}_2, N^2\bar{B}_2, \dots, N^{n_2-1}\bar{B}_2].$$

Since  $N = 0$  we obtain

$$\mathcal{R} = \text{Im} \left[ \begin{array}{c|c} I & 0 \\ \hline 0 & \bar{B}_2 \end{array} \right] = \text{Im} \left[ \begin{array}{ccc|cc} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ \hline 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right].$$

By incorporating the change of variables in the beginning, it remains to show that

$$(3.5) \quad \text{Im} \left[ \begin{array}{cccc} 0 & 0 & 0 & 0 \\ I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{array} \right] \subset \text{Im} \left[ \begin{array}{cccc} I & -\hat{G}_{11} & 0 & -\hat{G}_{13} \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \\ 0 & 0 & 0 & 0 \end{array} \right]$$

if and only if  $\text{rank}[M \ G \ B] = n$ . But (3.5) holds if and only if the last row is void, which is by Theorem 3.9 the case if and only if  $\text{rank} \begin{bmatrix} M & G & B \end{bmatrix} = n$ .  $\square$

**THEOREM 3.19.** *The second order descriptor system (1.1) is  $\mathcal{R}$ -controllable if and only if for the corresponding first order system (3.4) the matrix*

$$\begin{bmatrix} 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} [\bar{B}_1, A_1 \bar{B}_1, A_1^2 \bar{B}_1, \dots, A_1^{n_1-1} \bar{B}_1]$$

*has full row rank.*

*Proof.* From [12] it is known that for a first order system in the form (3.4) the reachable set is  $\mathcal{R} = \mathbb{R}^{n_1} \oplus \text{Im}(\mathcal{K}(N, \bar{B}_2))$  and the reachable set from  $\xi_0 = 0$  is  $\mathcal{R}(0) = \text{Im}(\mathcal{K}(A_1, \bar{B}_1)) \oplus \text{Im}(\mathcal{K}(N, \bar{B}_2))$ . Thus, the first order system is  $\mathcal{R}$ -controllable if and only if  $\text{Im}(\mathcal{K}(A_1, \bar{B}_1)) = \mathbb{R}^{n_1}$ . The second order descriptor system has in its state only the variables  $\xi_2, \dots, \xi_4$ ; the other variables come from the transformation to first order and are not relevant. Hence the proof follows.  $\square$

We conclude this section with a summary of the obtained results. We have shown that natural extensions of the rank conditions **C0**, **C1**, **C2** allow one to characterize C-, C2-, strong C2-  $\mathcal{R}$ -,  $\mathcal{R}2$ -, and I-controllability for second order systems but that the common transformations to first order form may destroy the I-controllability. This implies two possible routes for second order descriptor systems. Either one works directly with the second order form and avoids the transformation to first order, or one performs a transformation to first order that preserves the I-controllability. The latter approach would require the computation of the normal form (2.4). If a first order formulation is desirable, then, however, it is essential to first regularize the system and to reduce the index to at most one.

**4. Observability of second order descriptor systems.** In this section we derive the corresponding observability conditions for second order descriptor systems and analyze, in particular, the duality between controllability and observability. For this we will need the subspace spanned by the right eigenvectors and principal vectors corresponding to the finite eigenvalues of  $\lambda^2 M + \lambda G + K$ ; see [21]. We call this space the *right finite eigenspace* of  $\lambda^2 M + \lambda G + K$  and denote by  $P_{r,2}$  the projection onto this space.

DEFINITION 4.1. *Consider a system as in (1.1)–(1.2). The system is called*

- (i) C-observable *if from an output  $y = 0$  for the input  $u = 0$  it already follows that the system has only the trivial solution  $x = 0$ ;*
- (ii)  $\mathcal{R}$ -observable *if from an output  $y = 0$  for the input  $u = 0$  it already follows that the solution  $x$  satisfies  $P_{r,2}x = 0$ ;*
- (iii) I-observable *if the impulsive behavior of the solution is uniquely determined by the impulsive behavior of the output  $y$  and the jump behavior of the input  $u$ .*

Remark 4.2. Since for the trivial solution also its derivative vanishes, it makes no sense to define a concept like C2-observability.

Because the transformation from (2.4) to (2.5) leaves input and output unchanged and the impulsive behavior of the newly introduced variables is uniquely determined by the impulsive behavior of the old variables, I-observability of second order systems is a direct generalization of I-observability for first order systems. Thus, it follows immediately that a system (2.4) is I-observable if and only if the corresponding first order system (2.5) is I-observable.

THEOREM 4.3. *Consider a second order descriptor system (1.1)–(1.2), in normal form (2.4), and let (2.5) be the first order system derived from this normal form. Then the system (2.4) is  $\mathcal{R}$ -observable if and only if the first order system (2.5) is  $\mathcal{R}$ -observable.*

*Proof.* Let  $\hat{P}_{r,2}$  be the projection onto the right finite eigenspace of  $\lambda^2 \hat{M} + \lambda \hat{G} + \hat{K}$ , with  $\hat{M}$ ,  $\hat{G}$ ,  $\hat{K}$  as in (2.4), and let  $\hat{P}_{r,1}$  be the projection onto the right finite eigenspace of  $\lambda \hat{E} + \hat{A}$ , with  $\hat{E}$ ,  $\hat{A}$  as in (2.5). If we choose the partitioning as in (2.4), then

$$\hat{P}_{r,2} = \begin{bmatrix} I & 0 \\ 0 & I \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

So, if (2.4) is  $\mathcal{R}$ -observable and if we set  $u = 0$  and  $y = 0$ , then it follows that  $\hat{x}_1 = 0$  and thus also  $\hat{\dot{x}}_1 = 0$ . From the fifth block row of (2.5) it then follows that

$\xi_1 = \dot{x}_1 = 0$ . Accordingly  $\xi$  has the form  $\xi = [0, \hat{x}^T]^T$ . Because

$$(4.1) \quad \hat{P}_{r,1} = \begin{bmatrix} I & 0 \\ 0 & \hat{P}_{r,2} \end{bmatrix},$$

it follows that  $\hat{P}_{r,1}\xi = 0$ , and so (2.5) is  $\mathcal{R}$ -observable. For the converse, observe that the solution  $\xi$  of (2.5) has the form  $\xi = [\xi_1]$ , where  $\hat{x}$  is the solution of (2.4). From  $\hat{P}_{r,1}\xi = 0$  and (4.1) it then follows immediately that  $\hat{P}_{r,2}\hat{x} = 0$ .  $\square$

It is again straightforward to show that strong equivalence preserves all types of observability for second order descriptor systems. The same is true for opu-equivalence transformations. Proportional or first order derivative feedback, on the other hand, may change the observability properties.

*Example 4.4.* The second order descriptor system  $M\ddot{x} + G\dot{x} + Kx = Bu$ ,  $y = Cx$ , with

$$M = \begin{bmatrix} 1 & & \\ & 0 & \\ & & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & & \\ & 1 & \\ & & 0 \end{bmatrix}, \quad K = \begin{bmatrix} 0 & & \\ & 0 & \\ & & 1 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

is clearly C-observable, because from  $u = 0$  one obtains  $x_3 = 0$  and from  $y = 0$  one gets  $x_1 = x_2 = 0$ . For the proportional feedback  $u = v + x_3$  and the closed loop system with input  $v$ , we obtain

$$\hat{M} = \begin{bmatrix} 1 & & \\ & 0 & \\ & & 0 \end{bmatrix}, \quad \hat{G} = \begin{bmatrix} 0 & & \\ & 1 & \\ & & 0 \end{bmatrix}, \quad \hat{K} = \begin{bmatrix} 0 & & \\ & 0 & \\ & & 0 \end{bmatrix},$$

$$\hat{B} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \hat{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Here we can no longer make any statement about  $x_3$ . Similar examples can be constructed by using first order derivative feedback. Analogously one can also show that  $\mathcal{R}$ -observability is not invariant.

To see the noninvariance of I-observability, consider a modification of system (1.12)

$$(4.2) \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \ddot{x} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \dot{x} + \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} x = \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \quad y = \begin{bmatrix} 1 & 0 \end{bmatrix} x.$$

By using as input the Heaviside function  $H(t)$ , which is 0 for  $t < 0$  and 1 for  $t \geq 0$ , the solution is  $x_1 = H(t)$  and  $x_2 = -H(t) - \dot{H} - \ddot{H}$ , but this impulsive solution is not observed in the output  $y = x_1$ . By choosing the proportional feedback  $u = -x_2 + x_1 + v$ , we obtain  $x_2 = v$  and  $x_1$  solves the second order differential equation  $\ddot{x}_1 + \dot{x}_1 + x_1 + x_2 = 0$ . A jump in the input  $v$  will be integrated, and hence the output cannot contain impulsive parts if the input is piecewise continuous; i.e., all potential impulsive parts of the solution (of which there are none) are observed in the output.

The noninvariance under proportional or first order derivative feedback poses a problem insofar as we cannot use Theorem 2.6 to construct a system that can be

correctly transformed to first order. For this reason we proceed in a different way and make use of Theorem 14 in [35], which implies the following result.

**THEOREM 4.5.** *Consider a second order descriptor system (1.1)–(1.2) with differentiation index  $\nu$ , and suppose that  $Bu$  is  $\nu-1$  times continuously differentiable. Then there exists a sequence of strong equivalence transformations and sopu-equivalence transformations such that the transformed system has the coefficients*

$$(\hat{M}, \hat{G}, \hat{K}, \hat{B}) = \left( \begin{bmatrix} I_{d^2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \tilde{G}_{1,1} & 0 & 0 & \tilde{G}_{1,4} \\ 0 & I_{d^1} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \tilde{K}_{1,1} & \tilde{K}_{1,2} & 0 & \tilde{K}_{1,4} \\ \tilde{K}_{2,1} & \tilde{K}_{2,2} & 0 & \tilde{K}_{2,4} \\ 0 & 0 & I_a & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \\ \hat{B}_3 \\ \hat{B}_4 \end{bmatrix} \right),$$

where  $\hat{M}, \hat{G}, \hat{K} \in \mathbb{R}^{n \times n}$  and  $\hat{B} \in \mathbb{R}[D_\mu]^{n \times m}$ .

Note here that we allow sopu-equivalences, which do not increase the differentiation order of  $x$  but may introduce derivatives of  $u$ .

*Remark 4.6.* In contrast to Theorem 2.6, the transformed system with coefficients as in Theorem 4.5 requires derivatives of  $u$ . But since we consider  $u = 0$  only to check  $\mathcal{R}$ - and  $\mathcal{C}$ -observability, this is not a problem.

Now that we have a transformation to normal form that preserves  $\mathcal{R}$ -,  $\mathcal{C}$ -, and  $\mathcal{I}$ -observability, we immediately observe that the first order duality of controllability and observability [12, 28] also holds in the second order case if the particular output  $y = Cx$  is used, since transposing and changing the roles of  $B$  and  $C^T$  can be carried out also in the specific reduction order given by (2.5). Thus we have the following immediate consequences for the dual system to (1.1) given by

$$(4.3) \quad M^T \ddot{x} + G^T \dot{x} + K^T x = C^T u.$$

**THEOREM 4.7.** *Consider a second order descriptor system (1.1)–(1.2). The system is  $\mathcal{C}$ -observable if and only if the dual system (4.3) is  $\mathcal{C}2$ -controllable.*

*Proof.* Let (1.1)–(1.2) be in normal form (2.4). The system is  $\mathcal{C}$ -observable if and only if the corresponding first order system (2.5) is  $\mathcal{C}$ -observable. This, however, is the case if and only if the dual first order system is  $\mathcal{C}$ -controllable; see, e.g., [12]. But the dual first order system is  $\mathcal{C}$ -controllable if and only if the dual second order system is  $\mathcal{C}2$ -controllable.  $\square$

The result for  $\mathcal{R}$ -observability is analogous.

**THEOREM 4.8.** *Consider a second order descriptor system (1.1)–(1.2). The system is  $\mathcal{R}$ -observable if and only if the dual system (4.3) is  $\mathcal{R}2$ -controllable.*

*Proof.* By using Theorem 4.3 the proof is analogous to that of Theorem 4.7.  $\square$

**THEOREM 4.9.** *Consider a second order descriptor system (1.1)–(1.2). The system is  $\mathcal{I}$ -observable if and only if the dual system (4.3) is proportionally and first order derivative  $\mathcal{I}$ -controllable.*

*Proof.* The proof is analogous to that of Theorem 4.7.  $\square$

For completeness we will also present coordinate-free algebraic conditions that can be immediately derived from the duality between controllability and observability.

**COROLLARY 4.10.** *A second order descriptor system (1.1)–(1.2) is*

(i)  $\mathcal{R}$ -observable if and only if

$$\text{rank} \begin{bmatrix} \lambda^2 M + \lambda G + K \\ C \end{bmatrix} = n;$$

(ii)  $\mathcal{C}$ -observable if and only if it is  $\mathcal{R}$ -observable and

$$\text{rank} \begin{bmatrix} M \\ G \\ C \end{bmatrix} = n;$$

(iii)  $\mathcal{I}$ -observable if and only if

$$\text{rank} \begin{bmatrix} M \\ T_\infty^1 G \\ T_\infty^2 K \\ C \end{bmatrix} = n,$$

where the rows of the matrix  $T_\infty^1$  form a basis of cokernel  $M$  and the rows of  $T_\infty^2$  form a basis of

$$\text{cokernel} \begin{bmatrix} M & GZ_2 \end{bmatrix} \setminus \text{cokernel} \begin{bmatrix} M & GZ_2 & KZ_5 \end{bmatrix},$$

the columns of  $Z_2$  form a basis of kernel  $M$ , and those of  $Z_5$  form a basis of kernel  $\begin{bmatrix} M \\ G \end{bmatrix}$ .

*Remark 4.11.* In the output equation (1.2) we could have also considered a term  $C_1 \dot{x}$ . If such a term is present, then we can still transform to the form (2.4) and investigate the observability. In this case, however, the duality may be lost if derivatives of  $\hat{x}_2, \dots, \hat{x}_5$  occur.

**5. Conclusion.** We have shown how to extend the analysis of controllability and observability conditions to second order descriptor systems. We have demonstrated that the straightforward idea of using a classical first order formulation and then applying the first order results does not work, because in particular  $\mathcal{I}$ -controllability and  $\mathcal{I}$ -observability are not invariant under this transformation to first order. We have derived normal forms which can be used to check the controllability and observability conditions and from which we can obtain new first order formulations which preserve  $\mathcal{I}$ -controllability and  $\mathcal{I}$ -observability.

All of the presented results can be extended to nonreal, rectangular, and also higher order descriptor systems.

**Acknowledgment.** We thank three anonymous referees for several comments and suggestions which helped to improve the paper.

#### REFERENCES

- [1] W. F. ARNOLD AND A. J. LAUB, *Controllability and observability criteria for multivariable linear second-order model*, IEEE Trans. Automat. Control, 29 (1984), pp. 163–165.
- [2] Z. BAI, D. BINDEL, J. CLARK, J. DEMMEL, K. S. J. PISTER, AND N. ZHOU, *New numerical techniques and tools in SUGAR for 3D MEMS simulation*, in Technical Proceedings of the Fourth International Conference on Modeling and Simulation of Microsystems, NSTI, Cambridge, MA, 2000, pp. 31–34.
- [3] Z. BAI, P. DE WILDE, AND R. W. FREUND, *Reduced order modeling*, in Numerical Methods in Electromagnetics, Handb. Numer. Anal. XIII, W. Schilders and E. J. W. ter Maten, eds., Elsevier, New York, 2005, pp. 825–895.

- [4] A. BUNSE-GERSTNER, R. BYERS, V. MEHRMANN, AND N. K. NICHOLS, *Feedback design for regularizing descriptor systems*, Linear Algebra Appl., 299 (1999), pp. 119–151.
- [5] A. BUNSE-GERSTNER, V. MEHRMANN, AND N. K. NICHOLS, *Regularization of descriptor systems by derivative and proportional state feedback*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 46–67.
- [6] R. BYERS, T. GEERTS, AND V. MEHRMANN, *Descriptor systems without controllability at infinity*, SIAM J. Control Optim., 35 (1997), pp. 462–479.
- [7] R. BYERS, P. KUNKEL, AND V. MEHRMANN, *Regularization of linear descriptor systems with variable coefficients*, SIAM J. Control Optim., 35 (1997), pp. 117–133.
- [8] R. BYERS, V. MEHRMANN, AND H. XU, *Staircase Forms and Trimmed Linearization for Structured Matrix Polynomials*, Linear Algebra Appl., to appear.
- [9] S. L. CAMPBELL, *Singular Systems of Differential Equations I*, Pitman, San Francisco, CA, 1980.
- [10] D. COBB, *Controllability, observability and duality in singular systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 1076–1082.
- [11] J. D. COBB, *On the solutions of linear differential equations with singular coefficients*, J. Differential Equations, 46 (1982), pp. 310–323.
- [12] L. DAI, *Singular Control Systems*, Springer-Verlag, Berlin, 1989.
- [13] J. W. DEMMEL AND B. KÅGSTRÖM, *Computing stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88 (1987), pp. 139–186.
- [14] E. EICH-SOELLNER AND C. FÜHRER, *Numerical Methods in Multibody Systems*, Teubner Verlag, Stuttgart, 1998.
- [15] F. R. GANTMACHER, *The Theory of Matrices I*, Chelsea Publishing, New York, 1959.
- [16] F. R. GANTMACHER, *The Theory of Matrices II*, Chelsea Publishing, New York, 1959.
- [17] T. GEERTS, *Invariant subspaces and invertibility properties for singular systems: The general case*, Linear Algebra Appl., 183 (1993), pp. 61–88.
- [18] T. GEERTS, *Solvability conditions, consistency, and weak consistency for linear differential-algebraic equations and time-invariant linear systems: The general case*, Linear Algebra Appl., 181 (1993), pp. 111–130.
- [19] M. GERDTS, *Local minimum principle for optimal control problems subject to index-two differential-algebraic equations*, J. Optim. Theory Appl., 130 (2006), pp. 443–462.
- [20] I. GOHBERG, M. A. KAASHOEK, AND P. LANCASTER, *General theory of regular matrix polynomials and band Toeplitz operators*, Integral Equations Operator Theory, 11 (1988), pp. 776–882.
- [21] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.
- [22] M. HOU, *A Three-Link Planar Manipulator Model*, Sicherheitstechnische Regelungs- und Meßtechnik, Bergische Universität–GH Wuppertal, Germany, 1994.
- [23] M. HOU, *Descriptor Systems: Observer and Fault Diagnosis*, Fortschr.-Ber. VDI Reihe 8, Nr. 482, VDI Verlag, Düsseldorf, Germany, 1999.
- [24] A. ILCHMANN, *Contributions to Time-Varying Linear Systems*, Verlag an der Lottbek, Hamburg, 1989.
- [25] A. ILCHMANN AND V. MEHRMANN, *A behavioural approach to time-varying linear systems. Part 1: General theory*, SIAM J. Control Optim., 44 (2005), pp. 1725–1747.
- [26] A. ILCHMANN AND V. MEHRMANN, *A behavioural approach to time-varying linear systems. Part II: Descriptor systems*, SIAM J. Control Optim., 44 (2005), pp. 1748–1765.
- [27] J. KAUTSKY, N. K. NICHOLS, AND E. K-W. CHU, *Robust pole assignment in singular control systems*, Linear Algebra Appl., 121 (1989), pp. 9–37.
- [28] H. W. KNOBLOCH AND H. KWAKERNAK, *Lineare Kontrolltheorie*, Springer-Verlag, Berlin, 1985.
- [29] P. KUNKEL AND V. MEHRMANN, *Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems*, Math. Control Signals Systems, 14 (2001), pp. 233–256.
- [30] P. KUNKEL AND V. MEHRMANN, *Differential-Algebraic Equations. Analysis and Numerical Solution*, EMS Publishing House, Zürich, 2006.
- [31] P. KUNKEL, V. MEHRMANN, AND W. RATH, *Analysis and numerical solution of control problems in descriptor form*, Math. Control Signals Systems, 14 (2001), pp. 29–61.
- [32] P. LOSSE AND V. MEHRMANN, *Algebraic Characterization of Controllability and Observability for Second Order Descriptor Systems*, preprint 2006/21, Institut für Mathematik, TU Berlin, D-10623 Berlin, FRG, 2006; also available online from url: <http://www.math.tu-berlin.de/preprints/>.
- [33] D. S. MACKEY, N. MACKEY, C. MEHL, AND V. MEHRMANN, *Vector spaces of linearizations for matrix polynomials*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 971–1004.

- [34] V. MEHRMANN, *The Autonomous Linear Quadratic Control Problem*, Springer-Verlag, Berlin, 1991.
- [35] V. MEHRMANN AND C. SHI, *Transformation of high order linear differential-algebraic systems to first order*, Numer. Algorithms, 42 (2006), pp. 281–307.
- [36] P. C. MÜLLER, P. RENTROP, W. KORTÜM, AND W. FÜHRER, *Constrained mechanical systems in descriptor form: Identification, simulation and control*, in Advanced Multibody System Dynamics, W. Schiehlen, ed., pp. 451–456, Kluwer Academic, Stuttgart, 1993.
- [37] M. OTTER, H. ELMQVIST, AND S. E. MATTSON, *Multi-domain modeling with modelica*, in CRC Handbook of Dynamic System Modeling, Paul Fishwick, ed., CRC Press, Boca Raton, FL, 2006.
- [38] P. J. RABIER AND W. C. RHEINBOLDT, *Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint*, SIAM, Philadelphia, 2000.
- [39] T. SCHMIDT AND M. HOU, *Rollringgetriebe*, Internal Report, Sicherheitstechnische Regelungs- und Meßtechnik, Bergische Universität, GH Wuppertal, Wuppertal, Germany, 1992.
- [40] R. SCHÜPPHAUS AND P. C. MÜLLER, *Control analysis and synthesis of linear mechanical descriptor systems*, in Advanced Multibody System Dynamics, W. Schiehlen, ed., Kluwer Academic, Stuttgart, 1990, pp. 463–468.
- [41] C. SHI, *Linear Differential-Algebraic Equations of Higher-Order and the Regularity or Singularity of Matrix Polynomial*, Ph.D. thesis, TU Berlin, Institut für Mathematik, Str. des 17. Juni 136, D-10623 Berlin, 2004.
- [42] P. VAN DOOREN, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–141.
- [43] G. C. VERGHESE, B. C. LÉVY, AND T. KAILATH, *A general state space for singular systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 811–831.
- [44] E. L. YIP AND R. F. SINCOVEC, *Solvability, controllability and observability of continous descriptor systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 702–707.



## A POST-TREATMENT OF THE HOMOGENIZATION METHOD FOR SHAPE OPTIMIZATION\*

O. PANTZ<sup>†</sup> AND K. TRABELSI<sup>‡</sup>

**Abstract.** We propose an alternative to the classical post-treatment of the homogenization method for shape optimization. Rather than penalize the material density once the optimal composite shape is obtained (by the homogenization method) in order to produce a workable shape close to the optimal one, we macroscopically project the microstructure of the former through an appropriate procedure that roughly consists in laying the material along the directions of lamination of the composite. We have tested our approach in the framework of compliance minimization in two-dimensional elasticity. Numerical results are provided.

**Key words.** shape optimization, homogenization, compliance, elasticity

**AMS subject classifications.** 49Q10, 74P05, 74B05, 65K10

**DOI.** 10.1137/070688900

**1. Introduction.** Shape optimization consists in finding the optimal shape (represented by an open set) that minimizes a given cost-function (a mapping from the set of open sets into  $\mathbb{R}$ ). The homogenization method in shape optimization extends the space of admissible shapes to composite shapes, that is, shapes containing microholes. This approach is motivated by theoretical as well as practical considerations and is now well established (see the books [1], [6], [8], and [10]). Theoretically, the problem does not, in general, admit solutions in a classical sense if the shape is not submitted to conditions; it is often useful to nucleate a multitude of tiny holes. Practically, optimization by the homogenization method allows the substitution of a nonstandard admissible space (the space of open sets) by a vector space (the parameters of the composite material). Solving the relaxed problem then leads to the obtention of a composite optimal shape.

However, more than an optimal shape, we are looking for a workable shape, i.e., a sufficiently smooth open set. A commonly used procedure consists in continuing the optimization, by the homogenization method, of the optimal composite solution obtained, albeit simultaneously penalizing the intermediate material densities. The main drawback of this method lies in the difficulty of controlling the level of detail with respect to the qualitative loss of optimality of the shape. Moreover, only the material density of the composite shape is explicitly used in the penalization procedure. The pattern of its microstructure is not directly exploited when the latter holds important information that can be turned into profit. In this paper, we propose an alternative method that consists in reconstituting a shape close to the optimal by straightforwardly reproducing the underlying microstructure at a macroscopic scale. A parameter allows us to control the desired degree of detail.

Let us mention that our work was prompted by another algorithm, which we have proposed in [13], that automatically combined the homogenization method with the boundary variation, and a topological criterion for nucleating holes. Indeed, we had

---

\*Received by the editors April 23, 2007; accepted for publication (in revised form) November 26, 2007; published electronically March 26, 2008.

<http://www.siam.org/journals/sicon/47-3/68890.html>

<sup>†</sup>CMAP, Ecole Polytechnique, 91128 Palaiseau Cedex, France (olivier.pantz@polytechnique.org).

<sup>‡</sup>ENST/TSI, 37-39 rue Dareau, 75014 Paris, France (karim.trabelsi@polytechnique.edu).

noticed that following the nucleation of a hole, the algorithm tended to produce areas of low density around the hole that were reminiscent of the underlying microstructure. Finally, note that our method is not a visual post-treatment intended to display the local microstructure of the optimal composite shape (see, for instance, [11]).

We first recall some classical results from the homogenization theory in elasticity. At this point, we present two important classes of composite materials: the periodic materials and the laminates. We have tested our procedure on the compliance minimization problem for an elastic structure in dimension two. It is a paradigm of the shape optimization problem, since it has the nonnegligible advantage of possessing an explicit relaxed formulation. The optimal shape can be attained in the special class of composite shapes made of laminates. This allows the numerical computation of (at least almost) optimal composite shapes. The compliance minimization problem as well as its relaxed version are presented in section 3. Lastly, in section 4, we introduce a new method that builds a sequence of shapes close to the optimal solution of the compliance minimization problem obtained through the homogenization method. This is illustrated by some numerical examples.

**2. Preliminaries. The theory of homogenization.** This section is a brief review of classical homogenization results in linear elasticity. We refer the reader to the monograph of Allaire [1] for more details on this topic (see also [6], [8], and [10]).

As stated in the introduction, the majority of problems arising in shape optimization are ill posed. The minimizing sequences do not converge to a classical shape. In effect, the nucleation of microscopic holes turns out to be profitable more often than not. The sequence of open sets thus obtained does not “converge” in the space of open sets. However, we can still define a notion of convergence for which these sequences are compact. The limits that may be attained in this fashion constitute the set of composite bodies. For simplicity, instead of considering bodies made of a mixture of material and void, we shall study composite bodies, occupying a fixed domain, made of two distinct materials. The theory of homogenization is actually better understood in this context. Numerically, void shall be approximated by a material of weak resistance, also called soft material.

**2.1. Composite materials: Homogenization in elasticity.** All the results presented in this section are classical. Their proofs can be found, for instance, in [1]. Let  $\Omega$  be an open set in  $\mathbb{R}^N$ . An elastic body occupying the domain  $\Omega$  is characterized in every point  $x$  of the domain  $\Omega$  by its Hooke law  $A(x)$ , a fourth-order tensor operating on  $N \times N$  symmetric matrices. Let  $\alpha$  and  $\beta$  be two positive real numbers. We introduce the following subsets of Hooke laws:

$$\mathcal{M}_{\alpha,\beta} := \{A \in \mathcal{M}_N^4 : A\xi : \xi \geq \alpha|\xi|^2 \text{ and } A^{-1}\xi : \xi \geq \beta|\xi|^2 \text{ for all } \xi \in \mathcal{M}_N^s\},$$

where  $\mathcal{M}_N^4$  denotes the space of fourth-order tensors operating on symmetric matrices and  $\mathcal{M}_N^s$  is the space of  $N \times N$  symmetric matrices. We assume the body  $\Omega$  to be clamped on its boundary, submitted to dead loads  $f \in L^2(\Omega; \mathbb{R}^N)$ . The elasticity system consists in determining the displacement  $u$  of the structure, i.e., the unique solution to the boundary value problem:

$$\begin{cases} -\operatorname{div}(Ae(u)) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where  $e(u)$  is the linearized metric tensor associated with the displacement  $u$ :

$$e(u) = \frac{1}{2}(\nabla u + \nabla u^T).$$

Consider a sequence of Hooke laws  $A^\varepsilon \in L^\infty(\Omega; \mathcal{M}_{\alpha,\beta})$ . The sequence  $A^\varepsilon$  is said to H-converge to  $A^* \in L^\infty(\Omega; \mathcal{M}_{\alpha,\beta})$  if and only if, for all  $f \in L^2(\Omega; \mathbb{R}^N)$ , the sequence of displacements  $u_\varepsilon$  of the boundary value problems:

$$\begin{cases} -\operatorname{div}(A^\varepsilon e(u_\varepsilon)) = f & \text{in } \Omega, \\ u_\varepsilon = 0 & \text{on } \partial\Omega, \end{cases}$$

converges in  $L^2(\Omega; \mathbb{R}^N)$  to the solution  $u$  of the boundary value problem:

$$\begin{cases} -\operatorname{div}(A^* e(u)) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

A remarkable property of H-convergence is that every sequence in  $L^\infty(\Omega; \mathcal{M}_{\alpha,\beta})$  admits a convergent subsequence.

*Remark 1.* If a sequence  $A^\varepsilon$  H-converges to  $A^*$ , the convergence result extends to elasticity problems with mixed boundary conditions.

Consider two elastic materials of Hooke laws  $A$  and  $B$ . A composite body made of these materials is a linear elastic body whose Hooke law may be obtained as an H-limit of a sequence of elements  $A^\varepsilon$  such that for all  $x \in \Omega$ ,  $A^\varepsilon(x)$  is equal to either  $A$  or  $B$ . For all mapping  $\theta \in L^\infty(\Omega; [0, 1])$ , we call  $\mathcal{G}_\theta$  the set of Hooke laws  $A^* \in L^\infty(\Omega; \mathcal{M}_{\alpha,\beta})$  derived using the materials  $A$  and  $B$  with respective local proportions  $\theta(x)$  and  $1 - \theta(x)$ . In other words,

$$\mathcal{G}_\theta := \left\{ A^* \in L^\infty(\Omega; \mathcal{M}_{\alpha,\beta}) : \exists \chi^\varepsilon \in L^\infty(\Omega; \{0, 1\}) \right. \\ \left. \text{such that } A^\varepsilon = \chi^\varepsilon A + (1 - \chi^\varepsilon)B \xrightarrow{H} A^* \text{ and } \chi^\varepsilon \xrightarrow{L^\infty-*} \theta \right\}.$$

Another noteworthy property of H-convergence is that the set of Hooke laws  $A(x)$  achieved at a point depends only on the value of  $\theta(x)$ . Hence, for all real  $\theta \in [0, 1]$ , there exists a closed subspace  $G_\theta$  of  $\mathcal{M}_{\alpha,\beta}$  satisfying

$$\mathcal{G}_\theta = \left\{ A^* \in L^\infty(\Omega; \mathcal{M}_{\alpha,\beta}) : A^*(x) \in G_{\theta(x)} \right\}.$$

In particular, this allows us to be content with the study of homogeneous composite solids in order to determine the properties of composite materials.

**2.2. Periodic composites.** A special family of composites is obtained by arranging periodically the two types of material  $A$  and  $B$ , as shown in Figure 1. Let  $Y = ]0, 1[^N$ . Let  $\chi \in L^\infty(\mathbb{R}^N; \{0, 1\})$  be a  $Y$ -periodic function, i.e.,  $\chi(x + f_i) = \chi(x)$  for all  $x \in \mathbb{R}^N$  and all vector  $f_i$  ( $i = 1, \dots, N$ ) of the canonical basis of  $\mathbb{R}^N$ . For all real  $\varepsilon > 0$ , we denote by  $A^\varepsilon$  the element of  $L^\infty(\Omega; \mathcal{M}_{\alpha,\beta})$  defined for all  $x \in \Omega$  by

$$A^\varepsilon(x) = \chi(x/\varepsilon)A + (1 - \chi(x/\varepsilon))B.$$

The sequence  $A^\varepsilon$  H-converges to a constant element  $A^* \in L^\infty(\Omega; \mathcal{M}_{\alpha,\beta})$  defined for all symmetric matrices of  $N$ th order by

$$A^* \sigma \cdot \sigma = \min_{\substack{\tau \in L^2_\#(Y; \mathcal{M}_N^s) \\ \int_Y \tau dx = \sigma \\ \operatorname{div}(\tau) = 0}} \int_Y (\chi(x)A + (1 - \chi(x))B)^{-1} \tau \cdot \tau dx,$$

where  $L^2_\#(Y; \mathcal{M}_N^s)$  is the set of  $Y$ -periodic  $L^2$  functions with values in  $\mathcal{L}_N^s$ . We call  $P_\theta$  the set of periodic composites obtained using  $A$  and  $B$  with respective proportions  $\theta$  and  $1 - \theta$ . Yet another remarkable property is the fact that  $\overline{P}_\theta = G_\theta$ ; i.e., any composite may be approximated by a periodic material.

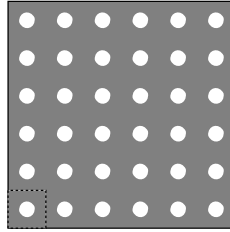


FIG. 1. *Periodic composite.*

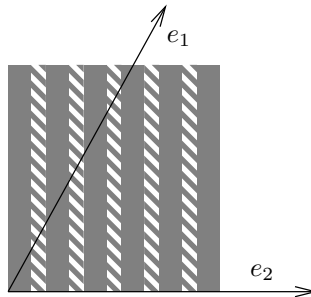


FIG. 2. *Rank-2 laminate.*

**2.3. Sequential laminates.** A rank-1 sequential laminate is obtained by successively layering the materials  $A$  and  $B$ . Higher-rank laminates are obtained recursively by repeating the procedure with the lower-rank laminate and  $A$  (or  $B$ ). Figure 2 shows a rank-2 laminate. The significant property of laminates consists in the fact that, unlike periodic composites, which necessitate solving variational problems, there are explicit formulas that enable us to compute the associated homogenized Hooke laws. In what follows, we shall consider only rank-2 laminates. Let us consider the limit case  $B = 0$ . A laminate is determined by several parameters: the density  $\theta$  of the material used, the successive directions of lamination  $e_i$ , and the proportions of lamination  $m_i$ , which represent the fraction of material with respect to the total quantity of material  $\theta$  used at each step of the manufacture of the laminate. In particular,  $\sum_i m_i = 1$ . The Hooke law  $A^*$  associated with such a composite is

$$A^{*-1} = A^{-1} + \frac{1 - \theta}{\theta} \left( \sum_{i=1}^N m_i f_A^c(e_i) \right)^{-1},$$

where for all unit vectors  $e$  of  $\mathbb{R}^2$ ,  $f_A^c(e)$  is the fourth-order tensor defined for all symmetric matrix  $\xi$  by the quadratic form

$$f_A^c(e)\xi \cdot \xi = A\xi \cdot \xi - \frac{1}{\mu} |A\xi|^2 + \frac{\mu + \lambda}{\mu(2\mu + \lambda)} ((A\xi)e \cdot e)^2.$$

For all densities  $\theta \in [0, 1]$ , we call  $L_{\theta,p}$  the set of all the Hooke laws generated by rank- $p$  laminates.

**3. Compliance optimization.** We restrict our analysis henceforth to the two-dimensional case,  $N = 2$ . Let  $\Omega$ , a connected open set of  $\mathbb{R}^2$ , be the reference configuration of a homogeneous and isotropic linearly elastic body. Assume that the boundary

of  $\Omega$  consists of three parts: the portion  $\Gamma_D$  along which the body is clamped, the portion  $\Gamma_F$  which is left free, and the last portion  $\Gamma_N$  which is submitted to surface loads  $g$ . The displacement of the structure  $u$  is the unique element,

$$u \in V := \{v \in H^1(\Omega) : u = 0 \text{ a.e. on } \Gamma_D\},$$

where for all  $v \in V$ , we have

$$\int_{\Omega} Ae(u) \cdot e(v) \, dx = \int_{\Gamma_N} g \cdot v \, dx.$$

In the above,  $A$  is the used material's Hooke law defined for all symmetric tensors  $\xi$  by

$$A\xi = 2\mu\xi + \lambda \operatorname{tr} \xi \operatorname{Id},$$

where  $\lambda$  and  $\mu$  are the Lamé coefficients of the constitutive material of the solid  $\Omega$ , and  $e(v)$  is the linearized metric tensor of  $v$ :

$$e(v) = \frac{1}{2}(\nabla v + \nabla v^T).$$

The compliance associated with the shape  $\Omega$  is defined as the work of exterior forces exerted on the solid, that is,

$$J(\Omega) = \int_{\Gamma_N} g \cdot u \, dx = \int_{\Omega} 2\mu|e(u)|^2 + \lambda(\operatorname{div} u)^2 \, dx.$$

The weaker the compliance, the more rigid the structure. Since we neglect the weight of the body, it is always advantageous to add material in order to strengthen the structure and reduce its compliance. We consider the compliance minimization problem that consists in determining  $\Omega^* \in \mathcal{U}_{ad}$  satisfying

$$(3.1) \quad J(\Omega^*) = \min_{\Omega \in \mathcal{U}_{ad}} J(\Omega),$$

where  $\mathcal{U}_{ad}$  denotes the set of open sets whose boundaries contain  $\Gamma_N$  and  $\Gamma_D$ , and whose volumes do not exceed a given maximal volume  $V$ :

$$\mathcal{U}_{ad} = \{\Omega \text{ open set in } \mathbb{R}^2 \text{ such that } \Gamma_N \cup \Gamma_D \subset \partial\Omega \text{ and } |\Omega| \leq V\}.$$

**3.1. Relaxation by the homogenization method.** Problem (3.1) is ill posed: it does not have an optimal solution. Minimizing sequences of  $J$  consist of shapes  $\Omega$  having an increasing number of holes, and do not *converge* to an element of  $\mathcal{U}_{ad}$ . In order to solve this problem, it is necessary to enlarge the space of admissible designs by allowing for composite shapes. Such a structure is determined through the local density of the used material  $\theta(x)$ , and through its effective Hooke law  $A^*(x)$  corresponding to its microstructure. In the particular case we are concerned with, an optimal solution may be obtained thanks to the particular class of composite materials that consists of the rank-2 sequential laminates. Should we impose upon the shape to be contained in a fixed box  $D$ , the relaxed minimization problem associated with the compliance minimization problem consists in solving the following minimization problem:

$$\min_{\substack{0 \leq \theta \leq 1 \\ A^* \in L_{\theta,2} \\ \int_D \theta dx \leq V}} J(\theta, A^*),$$

with

$$J(\theta, A^*) = \int_{\Gamma_N} g \cdot u \, dx,$$

where

$$u \in W := \{v \in H^1(D)^2 : v = 0 \text{ a.e. on } \Gamma_D\},$$

and the displacement of the structure satisfies for all  $v \in W$ ,

$$\int_D A^* e(u) \cdot e(v) \, dx = \int_{\Gamma_N} g \cdot v \, dx.$$

To conclude, let us recall that the above problem may be rewritten, using the dual (or complementary) energy principle, as the minimization of a functional that does not directly involve the solution of a variational problem. More precisely, we have

$$(3.2) \quad J(\theta, A^*) = \min_{\substack{\sigma \in L^2(D; \mathcal{M}_2^s) \\ \operatorname{div} \sigma = 0 \text{ in } \bar{D} \\ \sigma n = g \text{ on } \Gamma_N}} \int_D A^{*-1} \sigma \cdot \sigma \, dx.$$

The asset of such a formulation is that it enables swapping the different steps of minimization. Now, for a given  $\theta$  and  $\sigma$ , we may explicitly determine the tensor  $A^*$  of  $L_{\theta,2}$  that minimizes  $A^* \sigma \cdot \sigma$ . In addition, we infer that the directions of lamination coincide with the eigenvectors of the matrix  $\sigma$  (in particular, they are orthogonal) and that the respective proportions of lamination are

$$m_1 = \frac{|\sigma_2|}{|\sigma_1| + |\sigma_2|} \quad \text{and} \quad m_2 = \frac{|\sigma_1|}{|\sigma_1| + |\sigma_2|}.$$

Lastly, we have

$$\min_{A^* \in L_{\theta,2}} A^{*-1} \sigma \cdot \sigma = A^{-1} \sigma \cdot \sigma + \frac{1-\theta}{\theta} g^*(\sigma),$$

with

$$g^*(\sigma) = \frac{\kappa + \mu}{2\mu\kappa} (|\sigma_1| + |\sigma_2|)^2.$$

The relaxed problem (3.2) is classically solved by successive minimizations with respect to  $(\theta, A^*)$  and  $\sigma$ ; minimizing with respect to  $\sigma$  requires, at each iteration, solving a variational problem.

**4. Projection of a composite shape.** In the previous section, we have recalled how the optimal solution to the compliance minimization problem is determined. Unfortunately, the solution obtained is a composite shape that is not workable in practice. To make up for this problem, it is natural to try to build up a sequence of shapes  $\Omega^\varepsilon$  that reproduce at a macroscopic scale the underlying microstructure of the optimal composite. This sequence shall be built so that its limit behavior is that of the optimal composite shape. However, this construction is difficult to achieve because of the different scales the construction of the shape involves: the size of the structure  $L$ , and the two scales of the microstructure. Indeed, the principal microstructure (of

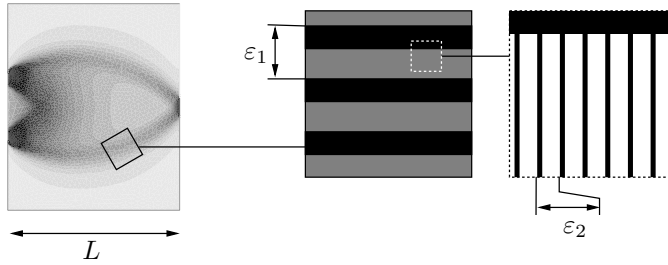
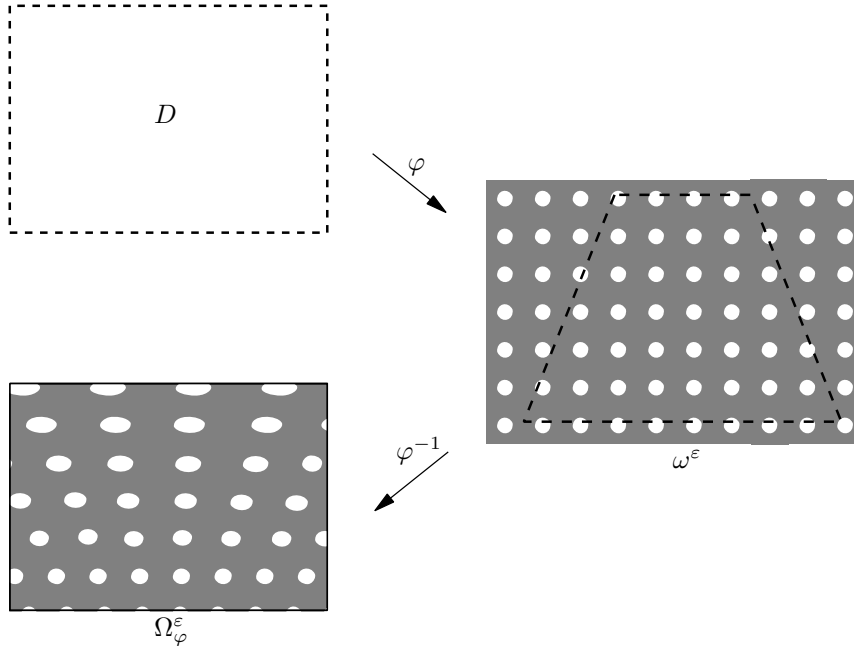


FIG. 3. *Three scales for one shape ( $L \gg \varepsilon_1 \gg \varepsilon_2$ ).*

characteristic scale  $\varepsilon_1$ ) itself contains a composite material (a micro-microstructure in a certain sense), namely, a rank-1 laminate (of characteristic scale  $\varepsilon_2$ ); see Figure 3. We may consider proceeding in two steps. First, reproduce a microstructure (at the macroscopic level) composed of a mixture of material, and a rank-1 laminate. Next, apply once again this procedure in order to end up with a noncomposite shape. Seemingly, three different scales have to be introduced. From a practical viewpoint, the smallest scale  $\varepsilon_2$  is bounded from below by the size of the smallest workable detail. The other magnitude scale  $\varepsilon_1$  is submitted to two contradictory constraints: it has to be small compared with the size of the structure  $L$ , and it has to be large with respect to the former scale  $\varepsilon_2$ , i.e., the smallest details allowed for; see Figure 3. Details must therefore be of a size order twice as small as the structure. Since the size of the smallest details, as well as the scale of the structure, are both parameters of the problem, they do not necessarily agree with this constraint. Moreover, the relationship between the Hooke law and the microstructure is not univocal. The same Hooke law may be achieved by different microstructures. Hence, the choice of reproducing the microstructure of a rank-2 laminate at a macroscopic scale proves to be somewhat arbitrary, all the more so as, even in this subclass of composite materials, different microstructures lead to the same Hooke law. For instance, changing the order of lamination ( $e_1$  with  $e_2$ , and  $m_1$  with  $m_2$ ) produces different microstructures without changing the Hooke law. These observations have induced us into opting for a slightly different approach that consists in reproducing, at a macroscopic level, a locally periodic shape whose associated constitutive law remains close (without being identical, however) to that of a rank-2 laminate. Such a shape has (locally) only one scale parameter, that is, the period.

**4.1. Construction of locally periodic elastic bodies.** As stated in section 2, in the nondegenerate case ( $A$  and  $B \in \mathcal{M}_{\alpha,\beta}$ ), every solid, whose Hooke law is that of a homogeneous composite solid pointwise (i.e., belonging to  $G_\theta$ ), may be obtained as the limit of solid bodies with Hooke laws taking their values in  $\{A, B\}$ . However, we do not have at our disposal a similar result in the general case (in particular, if  $B = 0$ ). Therefore, it is not obvious that we shall manage to build up a sequence of shapes  $\Omega^\varepsilon$  whose limit behavior converges to an optimal composite solid belonging to  $\mathcal{G}_\theta$ . Since we are not able to identify the set of composite solids that can be obtained with a unique elastic material  $A$  (combined with void), we set out to exhibit some of them. This shall enable us to project the optimal composite on an element of this subclass and accordingly build up a sequence of *real* shapes that converges to this element.

The constructions of homogeneous periodic solids and homogeneous sequential


 FIG. 4. Construction of  $\Omega_\varphi^\varepsilon$ .

laminates proposed in section 2 extend to the case  $B = 0$ . Homogeneous periodic solids are obtained as limits of the open sets:

$$\Omega^\varepsilon = D \cap \omega^\varepsilon,$$

where

$$\omega^\varepsilon = \{x \in \mathbb{R}^2 : \varepsilon^{-1}x \in \omega\},$$

and  $\omega$  is a  $Y$ -periodic open set, i.e.,

$$x \in \omega \Leftrightarrow x + f_1 \in \omega \Leftrightarrow x + f_2 \in \omega,$$

where  $(f_1, f_2)$  is the canonical basis of  $\mathbb{R}^2$ . It is easy to modify this definition in order to produce nonhomogeneous solids by allowing the open set  $\omega$  to depend on the considered point  $x$ . Let  $\omega$  be a smooth enough mapping from  $D$  into the space of  $Y$ -periodic open sets of  $\mathbb{R}^2$ . The sequence of open sets

$$\Omega^\varepsilon = \{x \in D : \varepsilon^{-1}x \in \omega(x)\}$$

converges to a composite solid whose Hooke law at every point  $x$  is that of a homogeneous periodic solid associated with  $\omega(x)$ . The set of composite solids that we may build up, in this fashion, remains limited after all. In particular, all the cells of periodicity are square (when we may likewise use rectangular cells) and identically oriented. Let  $\varphi$  be a smooth mapping of  $D$  into  $\mathbb{R}^2$ ; the sequence of open sets

$$(4.1) \quad \Omega_\varphi^\varepsilon = \{x \in D : \varepsilon^{-1}x \in \varphi^{-1}(\omega(x))\}$$



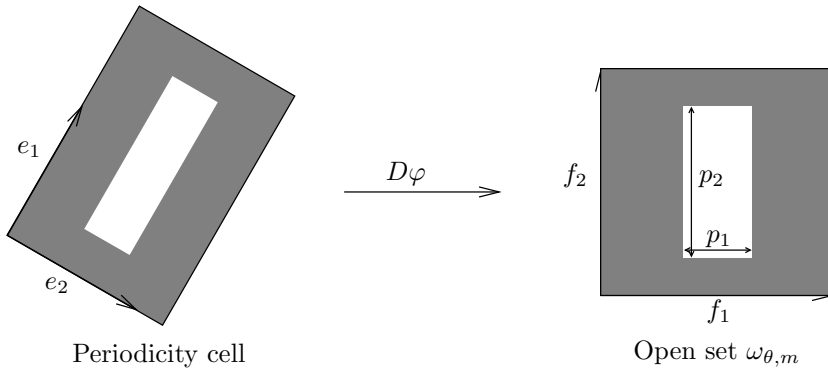


FIG. 5. Open set  $\omega_{\theta,m}$ .

converges to a composite solid whose Hooke law  $A^*$  is, at every point  $x$ , that of a periodic homogeneous body of period  $D\varphi^{-1}(x)Y$ , associated with the open set  $D\varphi^{-1}(x)\omega(x)$ ; see Figure 4. It should be mentioned that Briane [9] proposes a different construction that is also based on the deformation of a periodic network.

**4.2. An alternate composite material.** We substitute the optimal composite shape by a composite whose behavior is close and that may be obtained by the construction (4.1) described in the previous section.

At each point  $x$  of the domain  $D$ , the optimal laminate is determined by the orthogonal directions of lamination  $e_1$  and  $e_2$ , the density of the material  $\theta$ , and the proportions of lamination  $m_1$  and  $m_2$ . First of all, with all parameters  $m_1$ ,  $m_2$ , and  $\theta$ , we associate the  $Y$ -periodic open set  $\omega_{\theta,m}$  defined by

$$\omega_{\theta,m} \cap Y = \{x \in Y : 2|x_1 - 1/2| > p_1 \text{ and } 2|x_1 - 1/2| > p_2\},$$

where  $p_1$  and  $p_2$  are positive reals that satisfy

$$1 - p_1 p_2 = \theta \quad \text{and} \quad (1 - p_1)m_2 = (1 - p_2)m_1.$$

We assume that the directions of lamination  $e = (e_1, e_2)$  of the optimal shape constitute a regular field and that  $D$  is simply connected. Furthermore, we introduce the set  $\mathcal{F}_e$  of functions from  $D$  into  $\mathbb{R}^2$  defined by

$$(4.2) \quad \mathcal{F}_e = \left\{ \varphi : D \rightarrow \mathbb{R}^2 : \det(D_x \varphi) \neq 0, \right. \\ \left. D_x \varphi(e_1) \wedge f_1(x) = 0 \text{ and } D_x \varphi(e_2) \wedge f_2(x) = 0 \right\}.$$

The following lemma ensures that  $\mathcal{F}_e$  is not empty.

**LEMMA 4.1.** *Let  $D$  be a smooth connected bounded open set, and let  $e_1 \in C^1(\overline{D}; \mathbb{R}^2)$  be a vector field of unit norm defined in a neighborhood of  $\Omega$ ; then the set  $\mathcal{F}_e$  defined by (4.2) is not empty.*

This result is standard. However, it is not easy to find in the literature under the current formulation. For the reader's convenience, a proof is given in the appendix.

*Remark 2.* Should the directions of lamination reveal pointwise singularities or should  $D$  be not simply connected,  $\mathcal{F}_e$  may be empty.

For all elements  $\varphi \in \mathcal{F}_e$ , we call  $\Omega_\varphi^\varepsilon$  the sequence of open sets produced by the previous construction applied to  $\varphi$  and  $\omega = \omega_{\theta,m}$ ,

$$\Omega_\varphi^\varepsilon = \left\{ x \in D : \varepsilon^{-1}x \in \varphi^{-1}(\omega_{\theta,m}(x)) \right\}.$$

The sequence of open sets  $\Omega_\varphi^\varepsilon$  converges to a locally periodic composite whose density is that of the optimal laminate. Conditions  $D_x\varphi(e_i) \wedge f_i(x) = 0$  ( $i = 1, 2$ ) ensure that periodicity cells are oriented along the directions of lamination.

*Remark 3.* The limit behavior of the sequence of open sets  $\Omega_\varphi^\varepsilon$  is a locally periodic body whose periodicity cells are of the form  $R\omega_{\theta,m}$ , where  $R$  is a linear mapping (more precisely,  $R = D\varphi^{-1}$ ). Instead of solving the compliance minimization problem over rank-2 laminates, it is possible to perform only a partial relaxation by restricting the minimization to this class of composites. This approach was developed by Bendsøe and Kikuchi [7] for periodicity cells of the form above, where  $R$  is a rotation.

**5. Numerical applications.** It remains to determine an element  $\varphi$  of  $\mathcal{F}_e$  in order to infer the sequence of open sets  $\Omega_\varphi^\varepsilon$ . We choose to determine the element  $\varphi$  of  $\mathcal{F}_e$  that minimizes

$$I(\varphi) = \frac{1}{2} \int_D |\nabla\varphi_1 - e_1|^2 + |\nabla\varphi_2 - e_2|^2 dx.$$

The shape  $\Omega_\varphi^\varepsilon$  is then defined by

$$\Omega_\varphi^\varepsilon = \left\{ x \in D : \cos(\varphi_1(x)/\varepsilon) > \cos(\pi(1 - p_1)), \right. \\ \left. \text{and } \cos(\varphi_2(x)/\varepsilon) > \cos(\pi(1 - p_2)) \right\}.$$

The choice made is somewhat arbitrary, but it entails a linear system that is easily solved. Other choices are conceivable. The selection of the element  $\varphi$  of  $\mathcal{F}_e$  has an influence on the relative size of the nucleated holes and on their (more or less elongated) shape. We may consider selecting  $\varphi$  in order to minimize the error when substituting the optimal composite shape by the limit composite of the sequence  $\Omega_\varphi^\varepsilon$ . However, an explicit formula expressing a periodic material's Hooke law is not available, so this seems to be quite delicate to carry out.

**5.1. Orientation of eigenvectors.** The procedure described here requires the orientation of eigenvalues in a consistent fashion. Nevertheless, we can avoid such a computation. Assume that such an orientation has been established. We introduce vectors  $v_1$  and  $v_2$  defined by  $v_1 = \varphi_1 e_1$  and  $v_2 = \varphi_2 e_2$ . In this case, for  $i = 1, 2$  and  $j = 1, 2$ , we have

$$\nabla\varphi_i \cdot e_j = (\nabla v_i \cdot e_i) \cdot e_j.$$

Hence, minimizers of  $I(\varphi)$  also minimize the functional

$$I(v) = \frac{1}{2} \int_D |(\nabla v_1 \cdot e_1) \cdot e_1 - 1|^2 + |(\nabla v_2 \cdot e_2) \cdot e_2 - 1|^2 dx$$

on the set of vectors  $v = (v_1, v_2) : D \rightarrow \mathbb{R}^2 \times \mathbb{R}^2$  that satisfy

$$(5.1) \quad v_1 \wedge e_1 = 0, v_2 \wedge e_2 = 0, (\nabla v_1 \cdot e_2) \cdot e_1 = 0, \text{ and } (\nabla v_2 \cdot e_1) \cdot e_2 = 0.$$

Moreover, we have  $|\varphi_1| = \|v_1\|$  and  $|\varphi_2| = \|v_2\|$ . Now, the definition of  $\Omega_\varphi^\varepsilon$  depends only on the moduli of  $\varphi_1$  and  $\varphi_2$ . Therefore, in order to fully determine  $\Omega_\varphi^\varepsilon$ , it suffices to minimize  $I$  under the constraints (5.1), which is independent from the orientation of the eigenvectors  $e_1$  and  $e_2$ .

**5.2. Computing a minimizer of  $I$ .** Note that the computations of  $v_1$  and  $v_2$  are disconnected so that they can be carried out separately. To compute  $v_1$ , we introduce the Lagrangian

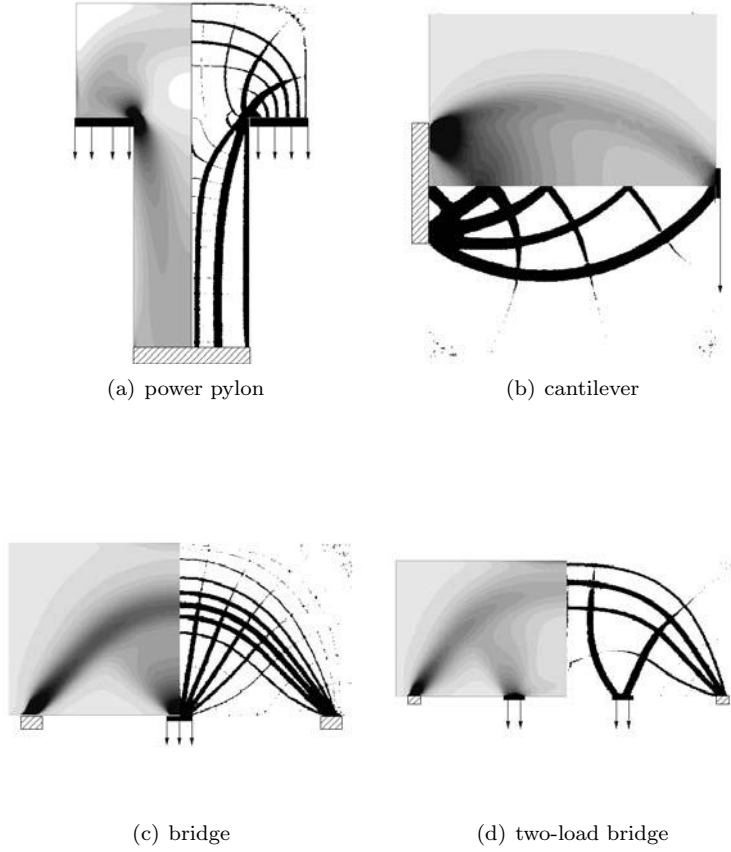
$$\begin{aligned} \mathcal{L}(v_1, p, P) = & \frac{1}{2} \int_D |(\nabla v_1 \cdot e_1) \cdot e_1 - 1|^2 + |\nabla(v_1 \cdot e_2)|^2 + |(\nabla v_1 \cdot e_2) \cdot e_1|^2 \\ & + \left| \frac{\sigma v_1 - \lambda v_1}{\lambda_1 - \lambda_2} \right|^2 + |v_1 \cdot (\nabla e_1 \cdot e_1)|^2 - |P \wedge e_1|^2 dx \\ & + \int_D (\nabla v_1 \cdot e_2) \cdot e_1 p + \frac{\sigma v_1 - \lambda_1 v_1}{\lambda_1 - \lambda_2} \cdot P dx, \end{aligned}$$

where the functions  $p$  and  $P$  are valued in  $\mathbb{R}$  and  $\mathbb{R}^2$ , respectively. Minimizers of  $I$  are saddle points of the Lagrangian. In order to determine a stationary point of the Lagrangian, we have used Lagrange finite elements  $P_1$  for all unknowns. An easy, but somewhat tedious, computation allows us to explicitly specify the different terms that appear in the Lagrangian  $\mathcal{L}$  in terms of  $v_1$  and  $\sigma$ . In particular,

$$(\nabla v_1 \cdot e_1) \cdot e_1 = \delta^{-1} \left( (\lambda_1 - \sigma_{22}) \frac{\partial v_1^1}{\partial x_1} + \sigma_{12} \left( \frac{\partial v_1^2}{\partial x_1} + \frac{\partial v_1^1}{\partial x_2} \right) + (\lambda_1 - \sigma_{11}) \frac{\partial v_1^2}{\partial x_2} \right),$$

$$\begin{aligned} \nabla(v_1 \cdot e_2) \cdot e_2 = & \delta^{-1} \left( \sigma_{12} \left( \frac{\partial v_1^1}{\partial x_1} - \frac{\partial v_1^2}{\partial x_2} \right) + (\lambda_1 - \sigma_{11}) \frac{\partial v_1^2}{\partial x_1} - (\lambda_1 - \sigma_{22}) \frac{\partial v_1^1}{\partial x_2} \right) \\ & - \delta^{-3} \left[ \left( \sigma_{12}(\lambda_1 - \sigma_{11}) \frac{\partial \sigma_{22} - \sigma_{11}}{\partial x_1} - \sigma_{12}^2 \frac{\partial \sigma_{22} - \sigma_{11}}{\partial x_2} \right. \right. \\ & \quad \left. \left. - (\sigma_{22} - \sigma_{11})(\lambda_1 - \sigma_{11}) \frac{\partial \sigma_{12}}{\partial x_1} + \sigma_{12}(\sigma_{22} - \sigma_{11}) \frac{\partial \sigma_{12}}{\partial x_2} \right) v_1^1 \right. \\ & \quad \left. - \left( \sigma_{12}^2 \frac{\partial \sigma_{22} - \sigma_{11}}{\partial x_1} - \sigma_{12}(\lambda_1 - \sigma_{22}) \frac{\partial \sigma_{22} - \sigma_{11}}{\partial x_2} \right. \right. \\ & \quad \left. \left. - (\sigma_{22} - \sigma_{11}) \sigma_{12} \frac{\partial \sigma_{12}}{\partial x_1} + (\sigma_{22} - \sigma_{11})(\lambda_1 - \sigma_{22}) \frac{\partial \sigma_{12}}{\partial x_2} \right) v_1^2 \right], \end{aligned}$$

$$\begin{aligned} \nabla(v_1 \cdot e_2) \cdot e_1 = & \delta^{-1} \left( \sigma_{12} \left( \frac{\partial v_1^2}{\partial x_1} + \frac{\partial v_1^1}{\partial x_2} \right) - (\lambda_1 - \sigma_{22}) \frac{\partial v_1^1}{\partial x_1} + (\lambda_1 - \sigma_{11}) \frac{\partial v_1^2}{\partial x_2} \right) \\ & - \delta^{-3} \left[ \left( -\sigma_{12}^2 \frac{\partial \sigma_{22} - \sigma_{11}}{\partial x_1} - \sigma_{12}(\lambda_1 - \sigma_{11}) \frac{\partial \sigma_{22} - \sigma_{11}}{\partial x_2} \right. \right. \\ & \quad \left. \left. + (\sigma_{22} - \sigma_{11}) \sigma_{12} \frac{\partial \sigma_{12}}{\partial x_1} + (\sigma_{22} - \sigma_{11})(\lambda_1 - \sigma_{11}) \frac{\partial \sigma_{12}}{\partial x_2} \right) v_1^1 \right. \\ & \quad \left. + \left( \sigma_{12}(\lambda_1 - \sigma_{22}) \frac{\partial \sigma_{22} - \sigma_{11}}{\partial x_1} + \sigma_{12}^2 \frac{\partial \sigma_{22} - \sigma_{11}}{\partial x_2} \right. \right. \\ & \quad \left. \left. - (\sigma_{22} - \sigma_{11})(\lambda_1 - \sigma_{22}) \frac{\partial \sigma_{12}}{\partial x_1} - \sigma_{12}(\sigma_{22} - \sigma_{11}) \frac{\partial \sigma_{12}}{\partial x_2} \right) v_1^2 \right], \end{aligned}$$

FIG. 6. *Optimal composite shapes and their projections.*

and

$$(\nabla v_1 \cdot e_2) \cdot e_1 = \delta^{-1} \left( \sigma_{12} \left( \frac{\partial v_1^2}{\partial x_2} - \frac{\partial v_1^1}{\partial x_1} \right) + (\lambda_1 - \sigma_{22}) \frac{\partial v_1^1}{\partial x_2} - (\lambda_1 - \sigma_{11}) \frac{\partial v_1^2}{\partial x_1} \right),$$

where

$$\lambda_1 = \frac{1}{2} (\sigma_{11} + \sigma_{22} + \delta),$$

is the greatest eigenvalue with  $\delta = \sqrt{(\sigma_{11} - \sigma_{22})^2 + 4\sigma_{12}^2}$  and  $v_1 = (v_1^1, v_1^2)$ . We have tested our approach for the compliance minimization of various structures: a power pylon, a cantilever, and several bridges. The results obtained are displayed in Figure 6. For each configuration, we show, on the one hand, the density of the optimal composite produced by the homogenization method, and, on the other, the projected shape  $\Omega_\varphi^\varepsilon$  derived by our method. Each structure is clamped on a part of its boundary (represented by a hatched block) and submitted to dead surface loads on another part of it (applied forces are represented by arrows). The weight of the structures was not taken into account. Finally, let us mention that the optimization may be naturally pursued by a geometric optimization method. To this end, a level set method (see [5], [3], [4], [2], [14], [15], [12]) seems quite appropriate. Figure 7 displays the shapes  $\Omega_\varphi^\varepsilon$

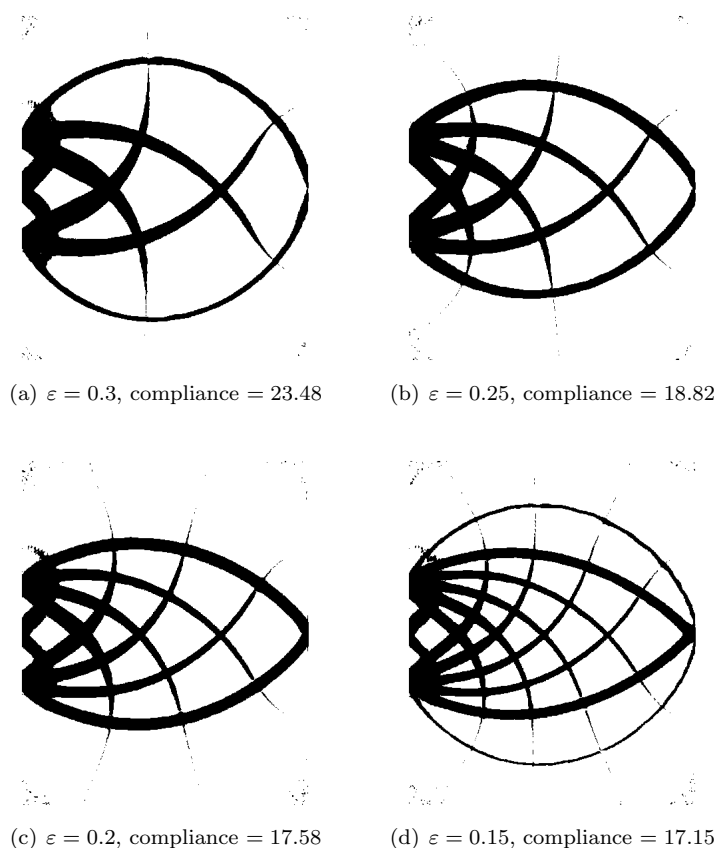


FIG. 7. Cantilever  $\Omega_\varphi^\varepsilon$  for different values of  $\varepsilon$ .

obtained for different values of  $\varepsilon$ . In the case at hand, the compliance of the optimal composite is 16.28. We notice that the compliance decreases proportionally with the parameter  $\varepsilon$ .

**5.3. Singularity of the field of eigenvectors.** The proposed method does not allow for the projection of laminates whose directions of lamination (i.e., the eigenvectors of the stress tensor  $\sigma$  of the optimal composite) show singularities. Generically, the singularities of the eigenvector field are made of a finite set of points for which  $\sigma$  is proportional to the identity. In such an instance, the eigenvector field is not orientable: it performs u-turns around singularities. Two different kinds of singularities are likely to occur according to the direction of rotation of the eigenvector field around the singularity. A singularity is said to be positive if along a small circle around it the eigenvector field has the standard trigonometric orientation; otherwise, it is said to be negative. Such a singularity typically appears in the case of a bridge with two loads as displayed in Figure 6(d). Figure 8 shows the (nonorientable) eigenvector field of the constraint  $\sigma$  of the optimal composite associated with the greatest eigenvalue  $\lambda_1$  around the singularity (located between the two loads, slightly above the platform of the bridge). We have also plotted the *level sets* orthogonal to the eigenvector fields  $e_1$  and  $e_2$ . We notice that the network thus obtained is not diffeomorphic to a square

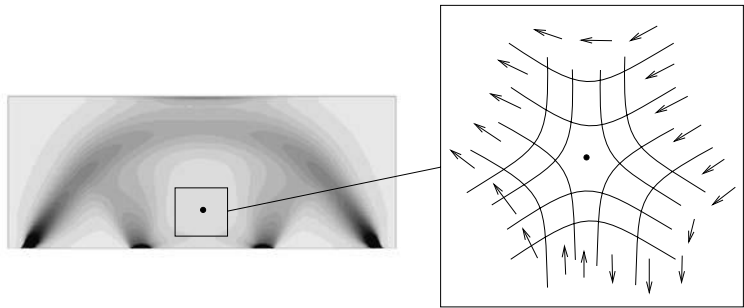


FIG. 8. *Nonorientable eigenvector field.*

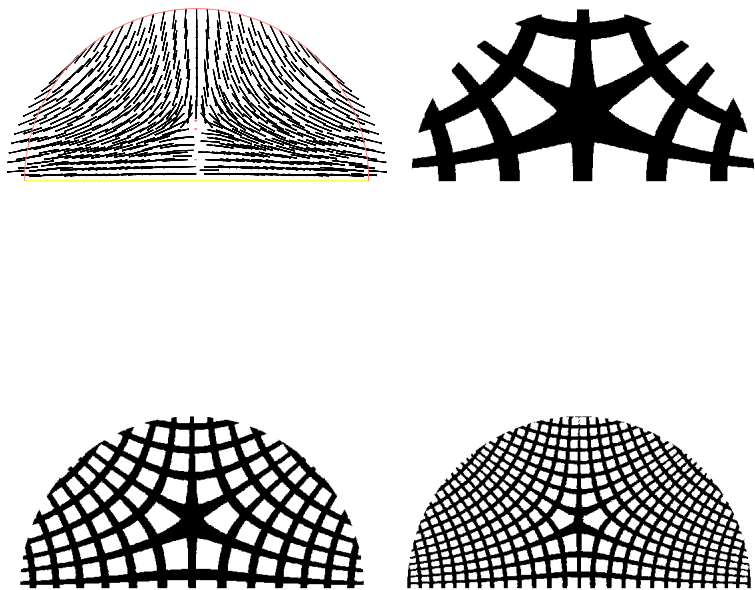
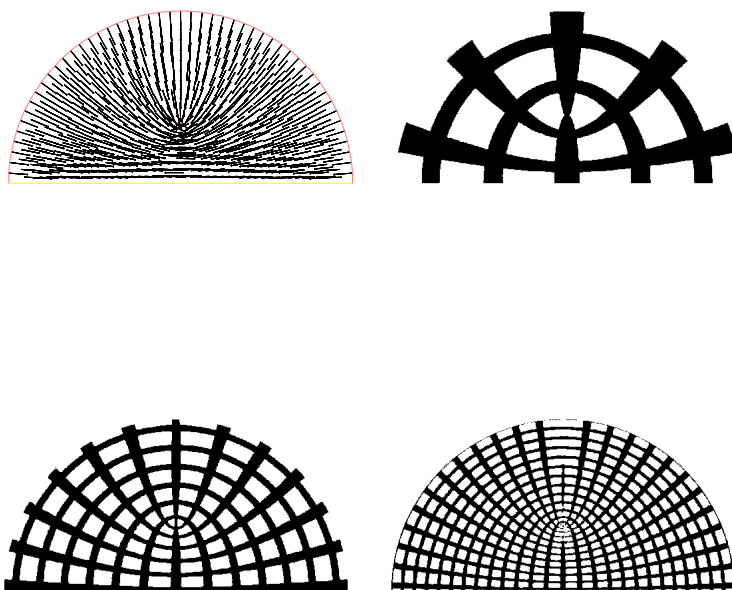


FIG. 9. *Negative singularity of the eigenvector field.*

network. It has a defect; i.e., the cell containing the singularity has five *right angles*. The trick, introduced earlier, that consists in bringing in the vectors  $v_1$  and  $v_2$  allows us to circumvent the problem in the case of a single singularity. Actually, the set of vector fields  $v_1$  of zero gradient and satisfying the constraints (5.1) is not empty, contrary to  $\mathcal{F}_e$ . Figures 9 and 10 display the eigenvector field  $e_1$  associated with, respectively, a negative and a positive singularity, as well as the projection plotted for different values of the parameter  $\varepsilon$  (with  $\theta = 1/2$  and  $m_1 = m_2 = 1/2$ ).

FIG. 10. *Positive singularity of the eigenvector field.*

In the case of several singularities, the set of vector fields  $v_1$  of zero gradient meeting the constraints (5.1) may be empty, and our algorithm is no longer adapted. Figures 11 and 12 display the eigenvector field  $e_1$  associated with, respectively, negative and positive singularities, as well as the projection plotted for several values of the parameter  $\varepsilon$  (with  $\theta = 1/2$  and  $m_1 = m_2 = 1/2$ ). The result of the projection is not satisfactory. In particular, it is different from the shape produced by symmetrizing shapes obtained for isolated singularities.

Our method cannot be applied as is if such a combination of singularities occurs. We may always consider partitioning the domain in order to isolate each singularity, apply our method to each part, then glue back the pieces. Yet, such a procedure is difficult to automate.

**6. Conclusion.** The post-treatment of the homogenization method presented here provides, in the framework of compliance minimization, quite interesting results compared to the classical penalization method. The main advantage is that it enables a sharp control of the size of the details of the final shape. Furthermore, the computational time it demands is negligible with respect to the homogenization procedure, whereas the material density penalization requires as much time as the latter. Several issues deserve to be investigated. First of all, this method should be coupled with a level set algorithm in order to sharpen the final shape. Besides, the suggested post-treatment of the homogenization method applies to the case where the directions of lamination exhibit at most one singularity. This is a strong limitation if we wanted to extend it to more general objective functions and/or complex geometries. In both

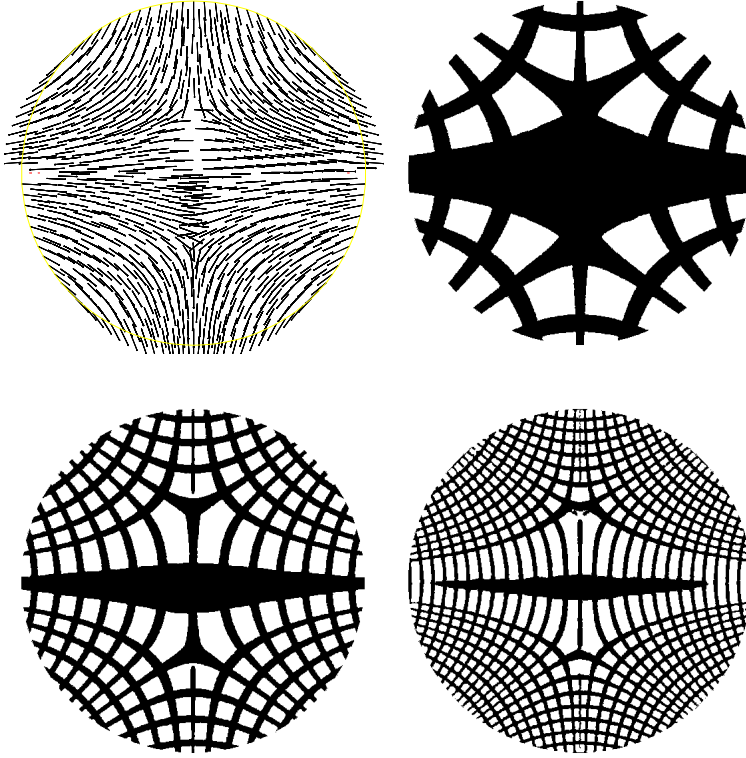


FIG. 11. *Two negative singularities of the eigenvector field.*

cases, the directions of lamination of the solutions computed by the homogenization method are likely to have more than one singularity. In the three-dimensional case, even more tricky situations may arise as lines (and not points) of singularities will have to be handled. Therefore, one needs to break free from this limitation. To conclude, this method potentially offers the prospect of an alternative to the classical topological gradient method by allowing us to nucleate a multitude of holes (or bars) in one iteration.

**Appendix A. Proof of Lemma 4.1.** We set out to exhibit an element of  $\mathcal{F}_e$ . First of all, we note that a function  $\varphi = (\varphi_1, \varphi_2)$  belongs to  $\mathcal{F}_e$  if and only if

$$(A.1) \quad \nabla \varphi_1 \neq 0, \quad \nabla \varphi_1 \cdot e_2 = 0$$

and

$$(A.2) \quad \nabla \varphi_2 \neq 0, \quad \nabla \varphi_2 \cdot e_1 = 0.$$

The conditions, to which  $\varphi_1$  and  $\varphi_2$  are submitted, are independent of one another. Therefore, it suffices to build a function  $\varphi_1$  satisfying hypotheses (A.1), since the function  $\varphi_2$  is produced by a similar procedure.

If  $\varphi_1$  satisfies (A.1), the level sets are smooth curves tangent to  $e_2$ . For all elements  $x$  in  $D$ , we denote by  $X_2(x, t)$  the solution of the following equation:

$$\begin{cases} X_2(x, 0) = x, \\ \frac{\partial X_2}{\partial t}(x, t) = e_2(X(x, t)). \end{cases}$$



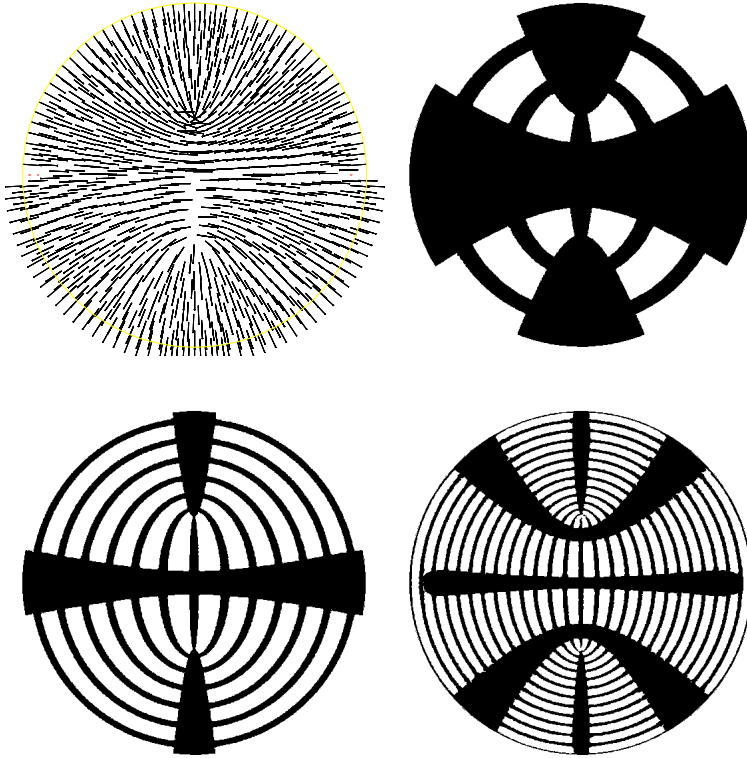


FIG. 12. Two positive singularities of the eigenvector field.

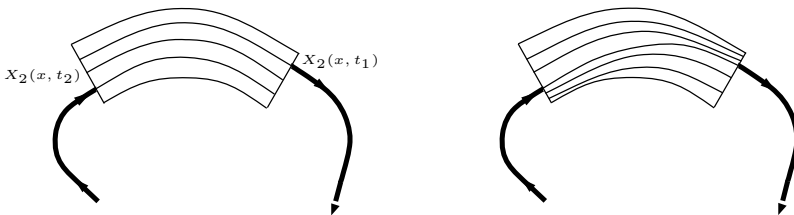


FIG. 13. Alteration of the vector fields in  $G_y$ .

Likewise, we define  $X_1(x, t)$  (by changing  $e_2$  into  $e_1$ ). By the Cauchy–Lipschitz theorem, there exists a unique maximal solution to this equation, since  $e_2$  is Lipschitzian. We call  $T_x^- \in \mathbb{R} \cup \{-\infty\}$  and  $T_x^+ \in \mathbb{R} \cup \{+\infty\}$  the lower and upper bounds of the time interval in which the maximal solution is defined, and we call  $S_x$  the set of points spanned by  $X(x, t)$ :

$$S_x = \{y = X(x, t)\}.$$

Note that there exists  $T_x > 0$  for which the mapping  $G_x$  from  $] -T_x, T_x[$  into  $\Omega$  that maps  $(t_1, t_2)$  onto  $X_2(X_1(x, t_2), t_1)$  is defined. Moreover, for  $T_x > 0$  small enough,  $G_x$  is a diffeomorphism, since  $D(G_x)(x) = (e_1, (x), e_2(x))$ . Now,  $\overline{D}$  is a compact set included in the image of functions  $G_x$ , so there exists a finite subset  $\mathcal{X}$  of elements of  $\overline{D}$  such that  $\overline{D} \subset \cup_{x \in \mathcal{X}} \text{Im}(G_x)$ . We set  $\omega = \cup_{x \in \mathcal{X}} \text{Im}(G_x)$ . Should we replace  $\Omega$  by

$\omega$ , we may assume that  $\Omega = \omega$ . We shall prove, on the one hand, that  $S_x$  cannot be a closed curve (it necessarily has endpoints), and, on the other hand, that it is of finite length, that is,  $T_x^-$  and  $T_x^+$  are finite. Assume that this is not true; every component of the preimage of  $G_y$  (for an element  $y \in \mathcal{X}$ ) by the map  $]T_x^-, T_x^+[\rightarrow \Omega, t \mapsto X_2(x, t)$  has its diameter bounded from below (by  $2 \max_{z \in \mathcal{X}} T_z$ ). Thus, if the interval  $]T_x^-, T_x^+[[$  is not bounded, there exists an element  $y \in \mathcal{X}$  such that  $X_2(x, \cdot)^{-1}(G_y)$  has an infinity of connected components. As we shall show, this is impossible. If this were the case, there would exist  $t_1$  and  $t_2$  in  $]T_x^-, T_x^+[[$  such that  $t_2 > t_1$ ,

$$X_2(x, t_1) = G_y(T_y, h_1), \quad \text{and} \quad X_2(x, t_2) = G_y(-T_y, h_2).$$

Without loss of generality, we may assume that  $t_1 = 0$ .

In addition, should we alter the field  $e_1$  as it is illustrated in Figure 13, we may assume that  $h_1 = h_2$ , and so  $S_x$  is diffeomorphic to a circle. However, all injections of the circle into  $\mathbb{R}^2$  are isotopic either to the canonical injection, or to the latter composed with a central symmetry. Therefrom, we infer the existence of a diffeomorphism  $F$  from the unit disc  $D^1$  onto the open set  $U$  included in  $D$ , and satisfying  $F(S^1) = S_x$ . Let  $f \in \mathcal{C}^0(D^1; S^1)$  be defined by  $f(x) = DF^{-1}(e_1(F(x)))/|DF^{-1}(e_1(F(x)))|$ . For all elements  $s \in S^1$ ,  $f(s)$  is nothing but  $s$  rotated by  $\pi/2$ . Now, according to Brouwer's theorem, such a field cannot be extended into a continuous field defined on the whole circle, which is exactly what  $f$  achieves. We have come to a contradiction, our initial hypothesis is accordingly false, and  $R_x$  is an open curve of finite length.

For all elements  $x$  in  $D$ , we denote by  $H_x$  the mapping defined in a neighborhood of the origin of  $\mathbb{R}^2$  by  $H_x(t, h) = X_2(X_1(x, h), t)$ . The restriction of  $H_x(t, h)$  to the axis  $h = 0$  is nothing but the injection  $t \mapsto X_2(x, t)$  of image  $S_x$ . Furthermore, the map  $H_x$  is differentiable, and

$$\frac{\partial H_x}{\partial t} = \dot{X}_2(X_1(x, h), t) = e_2(X_2(X_1(x, h)), t),$$

$$\frac{\partial H_x}{\partial h} = D_{X_1((x, h), t)} X_2 e_1(X_1(x, h)).$$

Wherefrom, we deduce, in particular, that

$$\frac{\partial}{\partial t} D_{h, t} H_x = D_{X_2(X_1(x, h), t)} e_2 D_{h, t} H_x$$

and that

$$\det(D_{(h, t)} H) = \text{Tr}(D_{X_2(X_1(x, h), t)} e_2) \det(D_{(h, t)} H).$$

However,  $D_{h=0, t=0} H_x = (e_1, e_2)$ , so that  $\det(D_{(h=0, t)} H) > 0$  for all  $t$ . Let  $\delta_x > 0$  be small enough for the endpoints of  $X_2(x, ]T_x^- + \delta_x, T_x^+ - \delta_x[)$  to belong to  $\omega \setminus \overline{D}$ . According to the local inversion theorem, for  $h_x > 0$  small enough, the map  $]T_x^- + \delta_x, T_x^+ - \delta_x[ \times ]-h_x, h_x[ \rightarrow \omega, (t, h) \mapsto H_x(t, h) = X_2(X_1(x, h), t)$  is a diffeomorphism onto its image, called  $V_x$ .

Once this tubular neighborhood is built, it is somewhat easy to build a function  $\varphi_x$  having the aforementioned properties in a neighborhood of  $S_x \cap \overline{D}$ . To do so, we first remark that  $D \setminus S_x$  consists of two distinct connected components (since  $\omega$  is simply connected). We call  $D_x^+$  the connected component of  $D \setminus S_x$  that contains  $H_x(t = 0, h_x)$ , and  $D_x^-$  the one that contains  $H_x(t = 0, -h_x)$ . We denote by  $\pi_2$

the projection of  $\mathbb{R}^2$  on the second coordinate. Let  $T$  be an infinitely differentiable increasing mapping satisfying  $T(x) = x$  in  $] -1/2, 1/2[$ , and  $T(x)$  is constant if  $|x| > 1$ . We define the mapping  $\varphi_x : \overline{D} \rightarrow \mathbb{R}$  by  $\varphi_x(y) = T(\pi_2 \circ H_x^{-1}(y)/h_x)$  for all  $y \in V_x \cap D$ ,  $\varphi(y) = T(1)$  for all  $y \in \overline{D}_x^+ \setminus V_x$ , and  $\varphi(y) = T(-1)$  for all  $y \in \overline{D}_x^- \setminus V_x$ . The map  $\varphi_x$  is  $C^1$  and satisfies  $\nabla \varphi_x \cdot e_2 = 0$ ,  $\nabla \varphi_x \cdot e_1 \geq 0$ . Finally,  $\nabla \varphi_x \cdot e_1 > 0$  in a neighborhood  $W_x$  of  $\overline{D} \cap S_x$  in  $\overline{D}$ . Now,  $\overline{D}$  is compact, which implies the existence of a finite family  $\tau$  of elements  $x \in D$  such that  $\overline{D} = \cup_{x \in \tau} W_x$ . The function  $\varphi_1 = \sum_{x \in \tau} \varphi_x$  satisfies (A.1); thus  $\mathcal{F}_e$  is not empty as claimed.

## REFERENCES

- [1] G. ALLAIRE, *Shape Optimization by the Homogenization Method*, Appl. Math. Sci. 146, Springer-Verlag, New York, 2002.
- [2] G. ALLAIRE, F. DE GOURNAY, F. JOUVE, AND A.-M. TOADER, *Structural optimization using topological and shape sensitivity via a level set method*, Control Cybernet., 34 (2005), pp. 59–80.
- [3] G. ALLAIRE, F. JOUVE, AND A.-M. TOADER, *A level-set method for shape optimization*, C. R. Math. Acad. Sci. Paris, 334 (2002), pp. 1125–1130.
- [4] G. ALLAIRE, F. JOUVE, AND A.-M. TOADER, *Structural optimization by the level-set method*, in Free Boundary Problems (Trento, 2002), Internat. Ser. Numer. Math. 147, Birkhäuser, Basel, 2004, pp. 1–15.
- [5] G. ALLAIRE, F. JOUVE, AND A.-M. TOADER, *Structural optimization using sensitivity analysis and a level-set method*, J. Comput. Phys., 194 (2004), pp. 363–393.
- [6] M. P. BENDSØE, *Optimization of Structural Topology, Shape, and Material*, Springer-Verlag, Berlin, 1995.
- [7] M. P. BENDSØE AND N. KIKUCHI, *Generating optimal topologies in structural design using a homogenization method*, Comput. Methods Appl. Mech. Engrg., 71 (1988), pp. 197–224.
- [8] M. P. BENDSØE AND O. SIGMUND, *Topology Optimization. Theory, Methods, and Applications*, Springer-Verlag, Berlin, 2003.
- [9] M. BRIANE, *Homogenization of a nonperiodic material*, J. Math. Pures Appl., 73 (1994), pp. 47–66.
- [10] A. CHERKAEV, *Variational Methods for Structural Optimization*, Appl. Math. Sci. 140, Springer-Verlag, New York, 2000.
- [11] J. HASLINGER, A. HILLEBRAND, T. KÄRKKÄINEN, AND M. MIETTINEN, *Optimization of conducting structures by using the homogenization method*, Struct. Multidiscip. Optim., 24 (2002), pp. 125–140.
- [12] S. J. OSHER AND F. SANTOSA, *Level set methods for optimization problems involving geometry and constraints. I. Frequencies of a two-density inhomogeneous drum*, J. Comput. Phys., 171 (2001), pp. 272–288.
- [13] O. PANTZ AND K. TRABELSI, *Simultaneous shape, topology, and homogenized properties optimization*, Struct. Multidiscip. Optim., 34 (2007), pp. 361–365.
- [14] J. A. SETHIAN AND A. WIEGMANN, *Structural boundary design via level set and immersed interface methods*, J. Comput. Phys., 163 (2000), pp. 489–528.
- [15] M. Y. WANG, X. WANG, AND D. GUO, *A level set method for structural topology optimization*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 227–246.

## REDUCTION FOR CONSTRAINED VARIATIONAL PROBLEMS ON 3-DIMENSIONAL NULL CURVES\*

EMILIO MUSSO<sup>†</sup> AND LORENZO NICOLODI<sup>‡</sup>

**Abstract.** We consider the optimal control problem for null curves in de Sitter 3-space defined by a functional which is linear in the curvature of the trajectory. We show how techniques based on the method of moving frames and exterior differential systems, coupled with the reduction procedure for systems with a Lie group of symmetries, lead to the integration by quadratures of the extremals. Explicit solutions are found in terms of elliptic functions and integrals.

**Key words.** null curves, invariant variational problems, extremal trajectories, optimal control systems, moving frames, Lax formulation, Marsden–Weinstein reduction

**AMS subject classifications.** 49F05, 58E10, 58A17

**DOI.** 10.1137/070686470

**1. Introduction.** Let  $M^3$  be a 3-dimensional Lorentz space form and  $\gamma \subset M^3$  a null curve parametrized by the natural (pseudo-arc) parameter  $s$  which normalizes the derivative of its tangent vector field. It is known that, in general,  $\gamma$  admits a curvature  $k_\gamma(s)$  that is a Lorentz invariant and that uniquely determines  $\gamma$  up to Lorentz transformations. We consider the variational problem on null curves defined by the Lorentz invariant functional

$$(1.1) \quad \mathcal{L}(\gamma) = \int_\gamma (m + k_\gamma) ds, \quad m \in \mathbb{R},$$

and ask how to determine the explicit form for the extremal trajectories. Motivations are provided by optimal control theory and recent work on relativistic particle models associated with action functionals of the type above (cf. [19], [18], [17], [7], and references therein).

From the Euler–Lagrange equation of the action it follows that the curvature of an extremal trajectory either is constant or is an elliptic function (possibly degenerate) of the natural parameter. In the first case, the extremals are orbits of 1-parameter subgroups of the group of Lorentz transformations and can be described in terms of elementary functions [6]. In the second case, we are led to a linear system of ODEs whose coefficients are doubly periodic functions. By the Fuchsian theory of ODEs, and in particular the results of Picard [20], the trajectories are then expressible in terms of the Weierstrass elliptic functions  $\wp$ ,  $\sigma$ , and  $\zeta$ . Alternatively, we follow a general scheme for the reduction of constrained variational problems on homogeneous spaces. We will use techniques from optimal control theory based on the method of moving frames and on Cartan’s exterior differential systems [4], [11], [8], [9], coupled with the reduction procedure for systems admitting a Lie group of symmetries extended to

---

\*Received by the editors March 27, 2007; accepted for publication (in revised form) September 28, 2007; published electronically April 4, 2008. This work was partially supported by MIUR projects *Metrische riemanniane e varietà differenziali* (E.M.) and *Proprietà geometriche delle varietà reali e complesse* (L.N.), and by the GNSAGA of INDAM.

<http://www.siam.org/journals/sicon/47-3/68647.html>

<sup>†</sup>Dipartimento di Matematica Pura ed Applicata, Università degli Studi dell’Aquila, Via Vetoio, I-67010 Coppito (L’Aquila), Italy (musso@univaq.it).

<sup>‡</sup>Dipartimento di Matematica, Università degli Studi di Parma, Viale G. P. Usberti 53/A, I-43100 Parma, Italy (lorenzo.nicolodi@unipr.it).

this setting [2]. For other applications of this general scheme of integration we refer to [10], [15], [16].

In this article, we determine the explicit form of the extremal curves when the target manifold is de Sitter 3-space. In this case, the functional (1.1) is invariant under the group  $\mathrm{SL}(2, \mathbb{C})$ , which doubly covers the identity component of the isometry group of de Sitter 3-space. The starting point of our study is the replacement of the original variational problem on null curves in de Sitter 3-space by an  $\mathrm{SL}(2, \mathbb{C})$ -invariant variational problem for integral curves of a control system on  $M \cong \mathrm{SL}(2, \mathbb{C}) \times \mathbb{R}$  defined by a suitable Pfaffian differential ideal  $(\mathcal{I}, \omega)$  with an independence condition. This is accomplished by proving the existence of a preferred  $\mathrm{SL}(2, \mathbb{C})$ -invariant frame along null curves without flex points (cf. section 2). We then follow a general construction due to Griffiths [11] and carry out a calculation to associate to the variational problem a Pfaffian differential system  $\mathcal{J}$ , the *Euler–Lagrange system*, whose integral curves are stationary for the associated functional. The Euler–Lagrange system is defined on the *momentum space*  $Y \cong \mathrm{SL}(2, \mathbb{C}) \times \mathbb{R}^3$ , which turns out to carry a contact structure, whose characteristic curves coincide with the integral curves of  $\mathcal{J}$ . As a matter of fact, in the case at hand all extremal trajectories arise as projections of integral curves of the Euler–Lagrange system. The theoretical reason for this is that all the derived systems of  $(\mathcal{I}, \omega)$  have constant rank (cf. [1]). Further, we show that the characteristic flow factors over a flow in an affine 3-dimensional subspace of  $\mathfrak{sl}(2, \mathbb{C})$  and find a Lax formulation of its defining differential equation. This implies that the momentum map induced by the Hamiltonian action of  $\mathrm{SL}(2, \mathbb{C})$  on  $Y$  is constant on solution curves of the Euler–Lagrange system, which leads to the integration by quadratures of the extremals (cf. section 4).

The paper is organized as follows. Section 2 gives the details of the construction of the canonical frame along null curves with no flex points by the method of moving frames, and defines the Pfaffian differential system of such frames. Section 3 studies the action functional (1.1), introduces the corresponding Euler–Lagrange system, and proves the constancy of the momentum map on its integral curves. Section 4 focuses on the integration procedure. It first outlines some facts from the theory of elliptic functions and then carries out the explicit integration of the extremals in terms of elliptic functions and elliptic integrals of the third kind.

## 2. Preliminaries.

**2.1. The geometry of de Sitter 3-space.** Let  $\mathrm{Herm}(2)$  be the 4-dimensional space of  $2 \times 2$  Hermitian complex matrices endowed with the Lorentz metric given by the quadratic form  $\langle X, X \rangle = -\det X$  for all  $X \in \mathrm{Herm}(2)$ . De Sitter 3-space,  $\mathbb{S}_1^3$ , can be viewed as the set of  $2 \times 2$  Hermitian matrices of determinant  $-1$ ,

$$(2.1) \quad \mathbb{S}_1^3 = \{X \in \mathrm{Herm}(2) \mid \det X = -1\},$$

with the induced metric  $g$ . The special linear group  $\mathrm{SL}(2, \mathbb{C})$  acts transitively by isometries on  $\mathbb{S}_1^3$  via the action

$$A \cdot X = AXA^*,$$

where  $A^*$  stands for the conjugate transpose of  $A$ . The stability subgroup at

$$J = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$

is the group  $\mathrm{SL}(2, \mathbb{R})$ , and  $\mathbb{S}_1^3$  may be described as a Lorentzian symmetric space

$$\mathbb{S}_1^3 \cong \mathrm{SL}(2, \mathbb{C}) / \mathrm{SL}(2, \mathbb{R}).$$

The projection

$$\pi : \mathrm{SL}(2, \mathbb{C}) \ni A \mapsto AJA^* \in \mathbb{S}_1^3$$

makes  $\mathrm{SL}(2, \mathbb{C})$  into a principal bundle with structure group  $\mathrm{SL}(2, \mathbb{R})$ .

Let  $\Omega = \alpha + i\beta$  be the Maurer–Cartan form of  $\mathrm{SL}(2, \mathbb{C})$ , where

$$(2.2) \quad \alpha = \begin{pmatrix} \alpha_1^1 & \alpha_2^1 \\ \alpha_1^2 & -\alpha_1^1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1^1 & \beta_2^1 \\ \beta_1^2 & -\beta_1^1 \end{pmatrix}.$$

Note that the matrix of 1-forms  $\beta$  is semibasic<sup>1</sup> for the projection  $\pi$ , and that the Lorentz metric  $g$  on  $\mathbb{S}_1^3$  is given by

$$g = (\beta_1^1)^2 - \beta_1^2 \beta_2^1.$$

The matrix  $\alpha$  amounts to the Levi–Civita (spinor) connection of  $g$ . The Maurer–Cartan equations of  $\mathrm{SL}(2, \mathbb{C})$ , or the structure equations, are given by

$$\begin{cases} d\alpha_1^1 = -\alpha_2^1 \wedge \alpha_1^2 + \beta_2^1 \wedge \beta_1^2, \\ d\alpha_1^2 = 2\alpha_1^1 \wedge \alpha_1^2 - 2\beta_1^1 \wedge \beta_1^2, \\ d\alpha_2^1 = -2\alpha_1^1 \wedge \alpha_2^1 + 2\beta_1^1 \wedge \beta_2^1, \\ \\ d\beta_1^1 = -\beta_2^1 \wedge \alpha_1^2 + \beta_1^2 \wedge \alpha_2^1, \\ d\beta_1^2 = 2\beta_1^1 \wedge \alpha_1^2 - 2\beta_1^2 \wedge \alpha_1^1, \\ d\beta_2^1 = -2\beta_1^1 \wedge \alpha_2^1 + 2\beta_2^1 \wedge \alpha_1^1. \end{cases}$$

**2.2. The canonical frame along a null curve.** A smooth parametrized curve

$$\gamma : I \rightarrow \mathbb{S}_1^3,$$

where  $I$  denotes any open interval of real numbers, is *null* (or *light-like*) if the velocity vector field  $\gamma'$  is null along  $\gamma$ , i.e.,  $g(\gamma'(t), \gamma'(t)) = 0$ , for each  $t \in I$ . We will assume throughout that  $\gamma$  has no flex points, i.e.,  $\gamma'(t)$  and  $\gamma''(t)$  are linearly independent, for each  $t \in I$ , where  $\gamma''$  denotes the covariant derivative of  $\gamma'$  along the curve.

A frame field along  $\gamma$  is a smooth map  $\Gamma : I \rightarrow \mathrm{SL}(2, \mathbb{C})$  such that  $\gamma = \pi \circ \Gamma$ . For any such frame, let  $\Theta = \Gamma^* \Omega$  denote the pullback of the Maurer–Cartan form of  $\mathrm{SL}(2, \mathbb{C})$  and write  $\Theta = \phi + i\theta$ . Given a frame field along  $\gamma$ , any other is given by

$$\tilde{\Gamma} = \Gamma X,$$

where  $X : I \rightarrow \mathrm{SL}(2, \mathbb{R})$  is a smooth map. If  $\tilde{\Theta} = \tilde{\Gamma}^* \Omega = \tilde{\phi} + i\tilde{\theta}$ , then

$$(2.3) \quad \tilde{\Theta} = X^{-1} \Theta X + X^{-1} dX.$$

<sup>1</sup>We recall that a differential form  $\varphi$  on the total space of a fiber bundle  $\pi : P \rightarrow B$  is said to be *semibasic* if its contraction with any vector field tangent to the fibers of  $\pi$  vanishes, or equivalently, if its value at each point  $p \in P$  is the pullback via  $\pi_p^*$  of some form at  $\pi(p) \in B$ . Some authors call such a form *horizontal*. A stronger condition is that  $\varphi$  is *basic*, meaning that it is locally the pullback via  $\pi^*$  of a form on the base  $B$ .

A frame field  $\Gamma : I \rightarrow \mathrm{SL}(2, \mathbb{C})$  along  $\gamma$  is said to be of *first order* if

$$(2.4) \quad \theta_1^1 = \theta_2^1 = 0, \quad \theta_1^2 \neq 0.$$

It is easily seen that first-order frame fields exist locally. If  $\Gamma : I \rightarrow \mathrm{SL}(2, \mathbb{C})$  is a first-order frame along  $\gamma$ , then any other is given by  $\tilde{\Gamma} = \Gamma X$ , where  $X : I \rightarrow G_1 \subset \mathrm{SL}(2, \mathbb{R})$  is a smooth map, and

$$G_1 = \left\{ \begin{pmatrix} a & 0 \\ c & a^{-1} \end{pmatrix} : a \neq 0, c \in \mathbb{R} \right\}.$$

According to (2.3), one computes

$$(2.5) \quad \tilde{\phi}_2^1 = a^2 \phi_2^1, \quad \tilde{\theta}_1^2 = \frac{1}{a^2} \theta_1^2.$$

Moreover, for first-order frames the form  $\phi_2^1$  is semibasic. If the curve  $\gamma$  has no flex points, then  $\phi_2^1 \neq 0$ . We say that the curve has *positive* or *negative spin* according to whether  $\phi_2^1$  is a positive or negative multiple of  $\theta_1^2$ .

Under our assumption, it follows from the transformation formula (2.3) that there always exist local first-order frames along  $\gamma$  such that

$$(2.6) \quad \phi_2^1 = \varepsilon \theta_1^2,$$

where  $\varepsilon = \pm 1$ , according to whether  $\gamma$  has positive or negative spin. A first-order frame field is said to be of *second order* if it satisfies (2.6) on  $I$ .

A second-order frame field along  $\gamma$  is said to be a *canonical* frame if

$$(2.7) \quad \phi_1^1 = 0.$$

Note that canonical frame fields exist on  $I$ , and that if  $\Gamma$  is a canonical frame, then any other is given by  $\pm \Gamma$ .

Summarizing, we have proved the following.

**PROPOSITION 2.1.** *Let  $\gamma : I \subset \mathbb{R} \rightarrow \mathbb{S}_1^3$  be a null curve with no flex points. Then there exists a frame along  $\gamma$ , the canonical frame,*

$$\Gamma : I \rightarrow \mathrm{SL}(2, \mathbb{C}),$$

*such that*

$$(2.8) \quad \Gamma^{-1} d\Gamma = \begin{pmatrix} 0 & \varepsilon \\ k+i & 0 \end{pmatrix} \omega,$$

*where  $\varepsilon = \pm 1$ ,  $\omega$  is a nowhere vanishing 1-form, the canonical pseudo-arc element, and  $k : I \rightarrow \mathbb{R}$  is a smooth function, the curvature of  $\gamma$ . Moreover, if  $\Gamma$  is a canonical frame field along  $\gamma$ , then any other canonical frame field is given by  $\pm \Gamma$ .*

*Remark 1.* Henceforth, we abuse the terminology and refer to the  $\mathbb{Z}_2$ -class  $[\Gamma] = \{\pm \Gamma\}$  as the canonical frame  $\Gamma$  of a null curve  $\gamma$ .

*Remark 2.* Conversely, for a smooth function  $k : I \rightarrow \mathbb{R}$ , let  $H(k) : I \rightarrow \mathfrak{sl}(2, \mathbb{C})$  be

$$(2.9) \quad H(k) = \begin{pmatrix} 0 & \varepsilon \\ k+i & 0 \end{pmatrix}.$$

Then by solving a linear system of ODEs, we see that there exists a unique (up to left multiplication)

$$\Gamma : I \rightarrow \mathrm{SL}(2, \mathbb{C})$$

such that

$$(2.10) \quad \Gamma^{-1}\Gamma' = H(k).$$

In particular,  $\gamma = \Gamma J\Gamma^* : I \rightarrow \mathbb{S}_1^3$  is a null curve without flex points and with curvature  $k$ .

*Remark 3* (null helices). The simplest examples are null helices, that is, null curves with constant curvature. Such curves are orbits of 1-parameter subgroups of  $\mathrm{SL}(2, \mathbb{C})$  (cf. Remark 9) and have been described by elementary functions in [6].

**2.3. The Pfaffian system of canonical frames.** Let  $(\mathcal{I}, \omega)$  be the Pfaffian differential system on  $M := \mathrm{SL}(2, \mathbb{C}) \times \mathbb{R}$  defined by the differential ideal  $\mathcal{I}$  generated by the linearly independent 1-forms

$$\begin{cases} \eta^1 = \beta_1^1, & \eta^2 = \beta_2^1, & \eta^3 = \alpha_1^1 - \varepsilon\omega, \\ \eta^4 = \alpha_1^1, & \eta^5 = \alpha_1^2 - k\omega, \end{cases}$$

where

$$\omega := \beta_1^2$$

gives the independence condition  $\omega \neq 0$ .

Now, let  $\gamma : I \rightarrow \mathbb{S}_1^3$  be a null curve without flex points. Then, by Proposition 2.1, the curve  $g = (\Gamma_\gamma, k_\gamma) : I \rightarrow M$ , whose components are, respectively, the canonical frame field along  $\gamma$  and the curvature of  $\gamma$ , is an integral curve of the Pfaffian system  $(\mathcal{I}, \omega)$ . Conversely, if  $g = (\Gamma, k) : I \rightarrow M$  is an integral curve of the Pfaffian system  $(\mathcal{I}, \omega)$ , then  $\gamma = \Gamma J\Gamma^* : I \rightarrow \mathbb{S}_1^3$  defines a null curve with no flex points,  $\Gamma$  is the canonical frame field along  $\gamma$ , and  $k$  is the curvature of  $\gamma$ . For this reason, null curves without flex points in  $\mathbb{S}_1^3$  can be identified with the integral curves of the Pfaffian system  $(\mathcal{I}, \omega)$ .

**DEFINITION 2.2.** *The Pfaffian differential system  $(\mathcal{I}, \omega)$  will be referred to as the canonical system.*

*Remark 4.* A smooth curve  $g = (\Gamma, k) : I \rightarrow M$  is an integral curve of the canonical system if and only if  $\Gamma : I \rightarrow \mathrm{SL}(2, \mathbb{C})$  is a solution of the linear system

$$\Gamma^{-1}(t)\Gamma'(t) = H(k(t)).$$

The function  $k$  plays the role of a control. Note that if we assign a smooth map  $k : I \rightarrow \mathbb{R}$  and a point  $A_0 \in \mathrm{SL}(2, \mathbb{C})$ , then there exists a unique integral curve  $g = (\Gamma, k)$  of the control system satisfying the initial condition  $\Gamma(t_0) = A_0$  for  $t_0 \in I$ .

Exterior differentiation and use of the Maurer–Cartan equations give, modulo the algebraic ideal generated by  $\eta^1, \dots, \eta^5$ , the *quadratic equations* of  $(\mathcal{I}, \omega)$ :

$$(2.11) \quad \begin{cases} d\omega \equiv 2(k\eta^1 + \eta^4) \wedge \omega, \\ d\eta^1 \equiv -(k\eta^2 + \eta^3) \wedge \omega, \\ d\eta^2 \equiv -2\varepsilon\eta^1 \wedge \omega, \\ d\eta^3 \equiv -2\varepsilon(k\eta^1 + 2\eta^4) \wedge \omega, \\ d\eta^4 \equiv (\eta^2 - k\eta^3 + \varepsilon\eta^5) \wedge \omega, \\ d\eta^5 \equiv -(dk + 2(1 + k^2)\eta^1) \wedge \omega. \end{cases}$$



### 3. The variational problem and the Euler–Lagrange system.

**3.1. The constrained variational problem.** Let  $\mathcal{N}$  be the space of null curves in  $\mathbb{S}_1^3$  without flex points. We consider the action functional

$$(3.1) \quad \mathcal{L}_m : \gamma \in \mathcal{N} \mapsto \int_{I_\gamma} (m + k_\gamma) \omega_\gamma, \quad m \in \mathbb{R},$$

where  $I_\gamma$  is the domain of definition of the curve,  $k_\gamma$  is its curvature, and  $\omega_\gamma$  the canonical pseudo-arc element (cf. section 2). We refer to [18], [19], [17], [7], and the references therein for a discussion on the particle model associated with this action functional.

**DEFINITION 3.1.** *A curve  $\gamma \in \mathcal{N}$  is said to be an extremal trajectory (or simply a trajectory) in  $\mathbb{S}_1^3$  if it is a critical point of the action functional  $\mathcal{L}_m$  when one considers compactly supported variations. The constant  $m$  is called the Lagrange multiplier of the trajectory.*

*Remark 5.* As usual, by a compactly supported variation of  $\gamma \in \mathcal{N}$  we mean a mapping  $V : I \times (-\epsilon, \epsilon) \rightarrow \mathbb{S}_1^3$  such that (1) for all  $u \in (-\epsilon, \epsilon)$ , the map  $\gamma_u := V(t, u) : I \rightarrow \mathbb{S}_1^3$  is a null curve without flex points; (2)  $\gamma_0 = \gamma(t)$  for all  $t \in I$ ; and (3) there exists a closed interval  $[a, b] \subset I$  such that

$$(3.2) \quad V(t, u) = \gamma(t) \quad \forall t \in I \setminus [a, b], \forall u \in (-\epsilon, \epsilon).$$

Accordingly, a curve  $\gamma \in \mathcal{N}$  is an extremal trajectory if, for every compactly supported variation  $V$ , we have that

$$\left. \frac{d}{du} \left( \int_{a_V}^{b_V} (m + k_{\gamma_u}) ds_u \right) \right|_{u=0} = 0,$$

where  $[a_V, b_V]$  is the support of the variation, i.e., the smallest closed interval for which (3.2) holds, and  $ds_u$  is the canonical pseudo-arc element of the curve  $\gamma_u$ .

In [7], the authors derive the Euler–Lagrange equation associated with (3.1) for null curves with prescribed endpoints and the same canonical frame at each end.

By the preceding discussion (cf. Proposition 2.1 and section 2.3), a curve  $\gamma \in \mathcal{N}$  is an extremal trajectory if and only if the pair  $g = (\Gamma_\gamma, k_\gamma)$  of its canonical frame field and curvature function is a critical point of the variational problem on the space  $\mathcal{V}(\mathcal{I}, \omega)$  of all integral curves of  $(\mathcal{I}, \omega)$  defined by the functional,

$$(3.3) \quad \widehat{\mathcal{L}} : g \in \mathcal{V}(\mathcal{I}, \omega) \mapsto \int_{I_g} g^*((m + k)\omega),$$

when one considers compactly supported variations through integral curves of  $(\mathcal{I}, \omega)$ .

*Remark 6.* The replacement of the original functional by the functional (3.3) is the starting point in the application of the Griffiths formalism. This approach to constrained variational problems with one independent variable provides conditions for criticality in terms of Pfaffian differential systems and is particularly well suited when one considers compactly supported variations among constrained curves. More importantly, it furnishes the appropriate setting for the explicit integration of the extremals (cf. [11], [1], [2], [12], and below).

**3.2. The Euler–Lagrange system.** Associated to the functional  $\widehat{\mathcal{L}}$  we will introduce, following Griffiths [11], the Euler–Lagrange system  $(\mathcal{J}, \omega)$  on a new manifold  $Y$ , which will be made explicit below.

For this, let  $Z \subset T^*M$  be the affine subbundle defined by

$$Z = (m + k)\omega + I \subset T^*M,$$

where  $I$  is the subbundle of  $T^*M$  associated to the differential ideal  $\mathcal{I}$ . The 1-forms  $(\eta^1, \dots, \eta^5, \omega)$  induce a global affine trivialization of  $Z$ , which may be identified with  $M \times \mathbb{R}^5$  by setting

$$M \times \mathbb{R}^5 \ni ((\Gamma, k); x_1, \dots, x_5) \mapsto \omega|_{(\Gamma, k)} + x_j \eta^j|_{(\Gamma, k)} \in Z$$

(throughout we use summation convention). Thus, the Liouville (canonical) 1-form of  $T^*M$  restricted to  $Z$  is given by

$$\mu = (m + k)\omega + x_j \eta^j.$$

Exterior differentiation and use of the quadratic equations (2.11) give

$$\begin{aligned} d\mu \equiv & dk \wedge \omega + 2(m + k)(k\eta^1 + \eta^4) \wedge \omega + dx_j \wedge \eta^j \\ & - x_1(k\eta^2 + \eta^3) \wedge \omega - 2\varepsilon x_2 \eta^1 \wedge \omega \\ & - 2\varepsilon x_3(k\eta^1 + 2\eta^4) \wedge \omega + x_4(\eta^2 - k\eta^3 + \varepsilon\eta^5) \wedge \omega \\ & - x_5(dk + 2(1 + k^2)\eta^1) \wedge \omega \quad \text{mod } \{\eta^i \wedge \eta^j\}. \end{aligned}$$

Next, we compute the Cartan system  $\mathcal{C}(d\mu) \subset T^*Z$  determined by the 2-form  $d\mu$ , i.e., the Pfaffian system generated by the 1-forms

$$\{i_\xi d\mu \mid \xi \in \mathfrak{X}(Z)\} \subset \Omega^1(Z).$$

Contracting  $d\mu$  with the vector fields of the tangent frame

$$\left( \frac{\partial}{\partial \omega}, \frac{\partial}{\partial k}, \frac{\partial}{\partial \eta^1}, \dots, \frac{\partial}{\partial \eta^5}, \frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_5} \right)$$

on  $Z$ , dual to the coframe

$$(\omega, dk, \eta^1, \dots, \eta^5, dx_1, \dots, dx_5),$$

we find the 1-forms

$$(3.4) \quad \eta^1, \dots, \eta^5,$$

$$(3.5) \quad \pi_1 = (x_5 - 1)dk,$$

$$(3.6) \quad \pi_2 = (1 - x_5)\omega,$$

$$(3.7) \quad \beta_1 = dx_1 - 2\{km + k^2 - \varepsilon x_2 - \varepsilon kx_3 - x_5(1 + k^2)\}\omega,$$

$$(3.8) \quad \beta_2 = dx_2 + (kx_1 - x_4)\omega,$$

$$(3.9) \quad \beta_3 = dx_3 + (x_1 + kx_4)\omega,$$

$$(3.10) \quad \beta_4 = dx_4 - \{2(m + k) - 4\varepsilon x_3\}\omega,$$

$$(3.11) \quad \beta_5 = dx_5 - \varepsilon x_4\omega.$$

We have proven the following.

LEMMA 3.2. *The Cartan system  $(\mathcal{C}(d\mu), \omega)$  associated to  $(I, \omega)$  is the differential ideal on  $Z \cong M \times \mathbb{R}^5$  generated by*

$$\{\eta^1, \dots, \eta^5, \pi_1, \pi_2, \beta_1, \dots, \beta_5\}$$

*and with independence condition  $\omega$ .*

DEFINITION 3.3. *The involutive prolongation of  $(\mathcal{C}(d\mu), \omega)$  on  $Z$  gives rise to a Pfaffian differential system  $(\mathcal{J}, \omega)$  on a submanifold  $Y \subset Z$ , which is called the Euler–Lagrange differential system associated to the variational problem. The submanifold  $Y$  is called the momentum space. We refer the reader to the book of Griffiths [11] for a discussion of how this system is derived and for more details on Pfaffian systems.*

LEMMA 3.4. *The momentum space  $Y$  is the 9-dimensional submanifold of  $Z$  defined by the equations*

$$x_5 = 1, \quad x_4 = 0, \quad x_3 = \frac{\varepsilon}{2}(m + k).$$

*The Euler–Lagrange system  $(\mathcal{J}, \omega)$  is the Pfaffian differential system on  $Y$  with independence condition  $\omega$  generated by the 1-forms*

$$\begin{cases} \eta^1|_Y, \dots, \eta^5|_Y, \\ \sigma_1 = dx_1 + (k^2 - mk + 2\varepsilon x_2 + 2)\omega, \\ \sigma_2 = dx_2 + kx_1\omega, \\ \sigma_3 = dk + 2\varepsilon x_1\omega. \end{cases}$$

Moreover,

$$\begin{aligned} \mu|_Y = & \frac{1}{2}(m - k)\beta_1^2 - \frac{\varepsilon}{2}k'\beta_1^1 + \frac{1}{2}\left(\frac{k''}{2} - \varepsilon k(k - m) - 2\varepsilon\right)\beta_2^1 \\ & + \frac{\varepsilon}{2}(m + k)\alpha_2^1 + \alpha_1^2. \end{aligned}$$

*Proof.* Let  $V_1(d\mu) \hookrightarrow \mathbb{P}[T(Z)] \rightarrow Z$  be the totality of 1-dimensional integral elements of  $\mathcal{C}(d\mu)$ . In view of (3.5) and (3.6), we find that

$$V_1(d\mu)|_{((\Gamma, k); x)} \neq \emptyset \iff x_5 = 1.$$

Thus, the first involutive prolongation of  $(\mathcal{C}(d\mu), \omega)$ , i.e., the image  $Z_1 \subset Z$  of  $V_1(d\mu)$  with respect to the natural projection  $V_1(d\mu) \rightarrow Z$ , is given by

$$Z_1 = \{((\Gamma, k); x) \in Z : x_5 = 1\}.$$

Next, the restriction of  $\beta_5$  to  $Z_1$  takes the form  $-\varepsilon x_4\omega$ . Thus, the second involutive prolongation  $Z_2$  is characterized by the equations

$$x_5 = 1, \quad x_4 = 0.$$

Considering then the restriction of  $\beta_4$  to  $Z_2$  yields the equations

$$x_5 = 1, \quad x_4 = 0, \quad x_3 = \frac{\varepsilon}{2}(m + k),$$

which define the third involutive prolongation  $Z_3$ . Now, the restriction  $\mathcal{C}_3(d\mu)$  to  $Z_3$  of  $\mathcal{C}(d\mu)$  is generated by the 1-forms  $\eta^1, \dots, \eta^5$  and

$$\begin{aligned} \sigma_1 &= dx_1 + (k^2 - mk + 2\varepsilon x_2 + 2)\omega, \\ \sigma_2 &= dx_2 + kx_1\omega, \\ \sigma_3 &= dk + 2\varepsilon x_1\omega. \end{aligned}$$

This implies that there exists an integral element of  $V_1(d\mu)$  over each point of  $Z_3$ , i.e.,  $V_1(d\mu)_p \neq \emptyset$ , for each  $p \in Z_3$ . Hence  $Y := Z_3$  and  $(\mathcal{J}, \omega) := (\mathcal{C}_3(d\mu), \omega)$  is the involutive prolongation of the Cartan system  $(\mathcal{C}(d\mu), \omega)$ .  $\square$

*Remark 7.* The importance of this construction is that the natural projection  $\pi_Y : Y \rightarrow M$  maps integral curves of the Euler–Lagrange system to extremals of the variational problem associated to  $(M, \mathcal{I})$ . The converse is not true in general. However, it is known to be true if all the derived systems of  $(\mathcal{I}, \omega)$  are of constant rank (cf. [1], [12]). In our case, one can easily check, using (2.11), that all the derived systems of  $(\mathcal{I}, \omega)$  have indeed constant rank, so that all the extremals do arise as projections of integral curves of the Euler–Lagrange system (see also section 3.3).

*Remark 8.* A direct calculation shows that

$$(3.12) \quad \mu|_Y \wedge (d\mu|_Y)^4 \neq 0$$

on  $Y$ , i.e., the variational problem is nondegenerate.<sup>2</sup> This implies that  $\mu|_Y$  is a contact form and that there exists a unique vector field  $\zeta \in \mathfrak{X}(Y)$ , the *characteristic vector field* of the contact structure, such that  $\mu|_Y(\zeta) = 1$  and  $i_\zeta d\mu|_Y = 0$ . In particular, the integral curves of the Euler–Lagrange system coincide with the characteristic curves of  $\zeta$ .

**3.3. The natural equation of integral curves.** Let  $\mathcal{V}(\mathcal{J}, \omega)$  be the set of integral curves of the Euler–Lagrange Pfaffian system  $(\mathcal{J}, \omega)$ . If  $y = ((\Gamma, k); x_1, x_2) : I \rightarrow Y$  is in  $\mathcal{V}(\mathcal{J}, \omega)$ , then equations

$$\eta^1 = \eta^2 = \cdots = \eta^5 = 0$$

and the independence condition  $\omega \neq 0$  tell us that  $\Gamma$  defines a canonical frame along the null curve  $\gamma = \Gamma J\Gamma^*$  and that  $k$  is the curvature of  $\gamma$ .

Next, for the smooth function  $k : I \rightarrow \mathbb{R}$ , let  $k'$ ,  $k''$ , and  $k'''$  be defined by

$$dk = k'\omega, \quad dk' = k''\omega, \quad dk'' = k'''\omega.$$

Equation  $\sigma_3 = 0$  implies

$$x_1 = -\frac{\varepsilon}{2}k'.$$

Further,  $\sigma_1 = 0$  gives

$$x_2 = \frac{1}{4}k'' - \frac{\varepsilon}{2}(k^2 - mk + 2).$$

Finally,  $\sigma_2 = 0$  yields

$$(3.13) \quad k''' - 6\varepsilon k k' + 2\varepsilon m k' = 0.$$

<sup>2</sup>A variational problem is said to be *nondegenerate* in the case when

$$\dim Y = 2m + 1 \quad \text{and} \quad \mu|_Y \wedge (d\mu|_Y)^m \neq 0.$$

Let  $V(\mathcal{J}, \omega)$  and  $V(\mathcal{C}(d\mu|_Y), \omega)$  denote the set of integral elements of the Euler–Lagrange system and of the Cartan system. For nondegenerate problems we have  $V(\mathcal{J}, \omega) = V(\mathcal{C}(d\mu|_Y), \omega)$ , whereas in general we have only inclusion  $V(\mathcal{J}, \omega) \subset V(\mathcal{C}(d\mu|_Y), \omega)$  (cf. [11, p. 84]). For a discussion on the relation between the classical Legendre transform and the construction of the Euler–Lagrange system on the momentum space, with special attention to the nondegeneracy condition, we refer the reader to [11, Chapter I, section e]).

This is the Euler–Lagrange equation of the extremals of (3.1). It has been computed, for example, in [7]. Thus, an integral curve of the Euler–Lagrange system projects to an extremal trajectory in  $\mathbb{S}_1^3$ .

Conversely, let  $\gamma : I \rightarrow \mathbb{S}_1^3$  be a null curve without flex points,  $\Gamma_\gamma$  its canonical frame, and  $k_\gamma$  its curvature. Define the lift  $y_\gamma : I \rightarrow Y$  of  $\gamma$  to the momentum space  $Y$  by

$$y_\gamma(t) = \left( (\Gamma_\gamma, k_\gamma); -\frac{\varepsilon}{2}k'_\gamma, \frac{1}{4}k''_\gamma - \frac{\varepsilon}{2}(k_\gamma^2 - mk + 2) \right).$$

Then,  $y_\gamma$  is an integral curve of the Euler–Lagrange system if and only if  $k_\gamma$  satisfies (3.13) if and only if  $\gamma$  is an extremal trajectory. Thus, the integral curves of the Euler–Lagrange system arise as lifts of trajectories in  $\mathbb{S}_1^3$ .

**3.4. The Lax formulation.** Introduce the *reduced curvature*

$$h := \frac{\varepsilon}{2} \left( k - \frac{m}{3} \right)$$

and identify  $Y \cong \mathrm{SL}(2, \mathbb{C}) \times \mathbb{R}^3$ , where  $\mathbb{R}^3$  has coordinates  $(h, h', h'')$ . Then, the Pfaffian equations defining the Euler–Lagrange system  $\mathcal{J}$  are given by

$$(3.14) \quad \begin{cases} \eta^j = 0, & (j = 1, \dots, 5), \\ dh = h'\omega, \\ dh' = h''\omega, \\ dh'' = 12hh'\omega, \end{cases}$$

where  $\omega \neq 0$  is the independence condition. Equation (3.13) becomes

$$(3.15) \quad h''' - 12hh' = 0,$$

$$(3.16) \quad H(h) = \begin{pmatrix} 0 & \varepsilon \\ 2\varepsilon h + \frac{m}{3} + i & 0 \end{pmatrix}$$

and we also have

$$\begin{aligned} \mu = & -\left(\varepsilon h - \frac{m}{3}\right)\beta_1^2 - h'\beta_1^1 + \frac{\varepsilon}{2}\left(h'' - 4h^2 + \frac{2}{3}\varepsilon mh + \frac{2}{9}m^2 - 2\right)\beta_2^1 \\ & + \left(h + \frac{2}{3}\varepsilon m\right)\alpha_2^1 + \alpha_1^2. \end{aligned}$$

Next, define the *momentum* associated with  $h$ ,  $U(h) \in \mathfrak{sl}(2, \mathbb{C})$ , by

$$(3.17) \quad \begin{pmatrix} ih' & 2i\varepsilon\left(h - \varepsilon\left(\frac{m}{3} + i\right)\right) \\ 2\left(h + \frac{2\varepsilon m}{3}\right) - i\varepsilon\left(h'' - 4h^2 + \frac{2\varepsilon mh}{3} + \frac{2m^2}{9} - 2\right) & -ih' \end{pmatrix}.$$

A direct computation shows that (3.15) is equivalent to

$$U(h)' = [U(h), H(h)].$$

The above discussion yields the following result.

**PROPOSITION 3.5.** *A map  $(A; h, h', h'') : I \subset \mathbb{R} \rightarrow Y$  is an integral curve of the Euler–Lagrange system  $(\mathcal{J}, \omega)$  if and only if*

$$(3.18) \quad \begin{cases} A^{-1}A' = H(h), \\ U(h)' = [U(h), H(h)]. \end{cases}$$

As a consequence, we have the following.

**COROLLARY 3.6.** *The momentum map*

$$\Phi : Y \rightarrow \mathfrak{sl}(2, \mathbb{C}), \quad (A; h, h', h'') \mapsto AU(h)A^{-1}$$

*is constant on integral curves of the Euler–Lagrange system.*

*Remark 9.* The momentum space  $Y$  may be identified with  $\mathrm{SL}(2, \mathbb{C}) \times \mathfrak{a}$ , where  $\mathfrak{a} = \mathrm{span}\{U(h)\}$  is an affine subspace of  $\mathfrak{sl}(2, \mathbb{C})$ . The group  $\mathrm{SL}(2, \mathbb{C})$  acts on  $(Y, \mu)$  by

$$g \cdot (A; U(h)) = (gA; U(h)), \quad \text{for each } g \in \mathrm{SL}(2, \mathbb{C}), U(h) \in \mathfrak{a},$$

in a Hamiltonian way. Using the isomorphism of  $\mathfrak{sl}(2, \mathbb{C})$  with its dual Lie algebra induced by the Killing form, one sees that the momentum map associated with this action is given by  $\Phi$ . Moreover, if  $y = (A(t), U(h)(t))$  is an integral curve of the characteristic vector field  $\zeta$ , then  $U(h)(t)$  is an integral curve of the vector field

$$X_\zeta : U(h) \mapsto [U(h), H(h)]$$

and  $\zeta$  can be written

$$\zeta|_y = H(h)|_A + X_\zeta(U(h)),$$

for all  $y = (A, U(h)) \in Y$ . If  $\mathfrak{a}_s$  denotes the singular set of  $X_\zeta$ , then the integral curves through  $(A, U(h)) \in \mathrm{SL}(2, \mathbb{C}) \times \mathfrak{a}_s$  are orbits of the 1-parameter subgroups generated by  $H(h)$ . By (3.17), these project to curves with constant curvature (null helices). Next, consider  $\Phi : \mathrm{SL}(2, \mathbb{C}) \times \mathfrak{a}_r \rightarrow \mathfrak{sl}(2, \mathbb{C})$ , where  $\mathfrak{a}_r$  denotes the complement of  $\mathfrak{a}_s$  in  $\mathfrak{a}$ . For each regular value  $\ell \in \mathfrak{sl}(2, \mathbb{C})$  of  $\Phi$ , the isotropy subgroup at  $\ell$ ,  $\mathrm{SL}(2, \mathbb{C})_\ell$ , is abelian and  $\dim \mathrm{SL}(2, \mathbb{C})_\ell = \mathrm{rank} \mathrm{SL}(2, \mathbb{C})_\ell = 2$ . The reduced space  $Y_\ell = \Phi^{-1}(\ell)/\mathrm{SL}(2, \mathbb{C})_\ell$  is then 1-dimensional. This implies that an integral curve  $y$  with momentum  $\ell$  (i.e.,  $\Phi \circ y = \ell$ ) can be found by quadratures. Any other integral curve with momentum  $\ell$  is given by  $b \cdot y$  for some  $b \in \mathrm{SL}(2, \mathbb{C})_\ell$ .

Note that when the action of the symmetry group on the momentum space is co-isotropic (as in the present case), the equation governing the flow of  $X_\zeta$  can always be written in Lax form. See, for instance, [10].

#### 4. Integration of the trajectories.

**4.1. Preparatory material.** From (3.15), it follows that the reduced curvature  $h$  satisfies

$$(4.1) \quad (h')^2 = 4h^3 - g_2h - g_3$$

for real constants  $g_2$  and  $g_3$ . Hence  $h$  is expressed by the real values of either a Weierstrass  $\wp$ -function with invariants  $g_2, g_3$ , or one of its degenerate forms.

We call a solution to (4.1) a *potential* with analytic invariants  $g_2, g_3$ . Two potentials are considered equivalent if they differ by a reparametrization of the form  $s \mapsto s + c$ , where  $c$  is a constant.<sup>3</sup> For real  $g_2$  and  $g_3$ , let  $\Delta(g_2, g_3) = 27g_3^2 - g_2^3$  be the discriminant of the cubic polynomial

$$P(t; g_2, g_3) = 4t^3 - g_2t - g_3.$$

<sup>3</sup>When invariants  $g_2$  and  $g_3$  are given, such that  $27g_3^2 \neq g_2^3$ , the general solution of the differential equation  $(\frac{dy}{dz})^2 = 4y^3 - g_2y - g_3$  can be written in the form  $\wp(z + \alpha; g_2, g_3)$ , where  $\alpha$  is a constant of integration.

The study of the real values of the Weierstrass  $\wp$ -function with real invariants  $g_2, g_3$  (and its degenerate forms) leads to primitive half-periods  $\omega_1, \omega_3$  such that (see for instance [14])

- $\Delta(g_2, g_3) < 0$ :  $\omega_1 > 0, \omega_3 = i\nu\omega_1, \nu > 0$ .
- $\Delta(g_2, g_3) > 0$ :  $\omega_1 > 0, \omega_3 = \frac{1}{2}(1 + i\nu)\omega_1, \nu > 0$ .
- $\Delta(g_2, g_3) = 0$  and  $g_3 > 0$ :  $\omega_1 > 0, \omega_3 = +i\infty$ .
- $\Delta(g_2, g_3) = 0$  and  $g_3 < 0$ :  $\omega_1 = +\infty, -i\omega_3 > 0$ .
- $g_2 = g_3 = 0$ :  $\omega_1 = +\infty, \omega_3 = +i\infty$ .

Accordingly, denoting by  $\mathcal{D}(g_2, g_3)$  the fundamental period-parallelogram spanned by  $2\omega_1$  and  $2\omega_3$ , the only possible cases for the potential function  $h : I \rightarrow \mathbb{R}$  are

- $\Delta < 0$ :  $h(s) = \wp(s; g_2, g_3), I = (0, 2\omega_1)$ .
- $\Delta < 0$ :  $h(s) = \wp_3(s; g_2, g_3) = \wp(s + \omega_3; g_2, g_3), I = \mathbb{R}$ .
- $\Delta > 0$ :  $h(s) = \wp(s; g_2, g_3), I = (0, 2\omega_1)$ .
- $\Delta = 0, g_3 = -8a^3 > 0$ :

$$h(s) = -3a \tan^2(\sqrt{-3a}s) - 2a, \quad I = \left(-\frac{\pi}{\sqrt{-12a}}, \frac{\pi}{\sqrt{-12a}}\right).$$

- $\Delta = 0, g_3 = -8a^3 < 0$ :

$$h(s) = 3a \tanh^2(\sqrt{3a}s) - 2a, \quad I = \mathbb{R}.$$

- $g_2 = g_3 = 0$ :  $h(s) = s^{-2}, I = (-\infty, 0)$ , or  $I = (0, +\infty)$ .

Let  $h$  be a Weierstrass potential with real invariants  $g_2, g_3$ , and let  $U(h)$  be the corresponding momentum as given by (3.17). Then

$$\begin{aligned} \det U(h) &= \left(\frac{4}{27}m^3 - 4m - \frac{m}{3}g_2 - \varepsilon g_3\right) + i\varepsilon \left(\frac{4}{3}m^2 - g_2 - 4\right) \\ &= P\left(\varepsilon \left(\frac{m}{3} + i\right); g_2, g_3\right). \end{aligned}$$

Let

$$\nu(m, h) := \sqrt{P\left(\varepsilon \left(\frac{m}{3} + i\right); g_2, g_3\right)},$$

chosen once for all. Then  $\pm\nu(m, h)$  are the eigenvalues of the momentum  $U(h)$ .

Next, define

$$(4.2) \quad \phi(m, h) := \begin{cases} \int \frac{\nu(m, h)}{h - \varepsilon \left(\frac{m}{3} + i\right)} ds, & \nu(m, h) \neq 0, \\ \int \frac{1}{h - \varepsilon \left(\frac{m}{3} + i\right)} ds, & \nu(m, h) = 0. \end{cases}$$

These are elliptic integrals of the third kind. Let  $w(m, h)$  be the unique point in the period-parallelogram  $\mathcal{D}(g_2, g_3)$  such that

$$h(w) = \varepsilon \left(\frac{m}{3} + i\right) \quad \text{and} \quad h'(w) = \nu(m, h).$$

Denote by  $\sigma_h$  and  $\zeta_h$ , respectively, the sigma and zeta Weierstrassian functions corresponding to the potential  $h$ , i.e., the unique analytic odd functions whose meromorphic

extensions satisfy  $\zeta'_h = -h$  and  $\sigma'_h/\sigma_h = \zeta_h$ . Under the above assumptions, we now compute the elliptic integrals (4.2). Three cases are considered.

*Case I.*  $\nu(m, h) \neq 0$ . In this case,

$$\phi(m, h) = \int \frac{h'(w)}{h(s) - h(w)} ds = \log \frac{\sigma_h(s - w)}{\sigma_h(s + w)} + 2s\zeta_h(w) + \text{const.}$$

*Case II.*  $\nu(m, h) = 0$  and  $g_2^2 + g_3^2 \neq 0$ . In this case,  $h(w) = \varepsilon \left( \frac{m}{3} + i \right)$  is a root of the cubic polynomial  $P$ , say  $e_3$ . If  $e_1, e_2$  denote the other two roots, we have

$$\begin{aligned} \phi(m, h) &= \int \frac{ds}{h(s) - e_3} = \int \frac{h(s + w) - e_3}{(e_3 - e_1)(e_3 - e_2)} ds \\ &= \frac{1}{\frac{g_2}{4} - 3 \left( \frac{m}{3} + i \right)^2} \left\{ \zeta_h(s + w) + \varepsilon \left( \frac{m}{3} + i \right) s \right\} + \text{const.} \end{aligned}$$

*Case III.*  $\nu(m, h) = 0$  and  $g_2 = g_3 = 0$ . In this case,

$$\phi(m, h) = \frac{1}{3}s^3 + \text{const.}$$

**4.2. Explicit integration.** We are now in a position to explicitly integrate the extremal trajectories. This amounts to integrating by quadratures the reduced system associated to the Hamiltonian action of  $\text{SL}(2, \mathbb{C})$  on  $Y$  (cf. Remark 9). The key to explicit integration is the conservation of the momentum map along integral curves of the Euler–Lagrange system.

**THEOREM 4.1.** *Let  $\gamma : I \rightarrow \mathbb{S}_1^3$  be an extremal trajectory with Lagrange multiplier  $m$  and reduced curvature  $h$  with real invariants  $g_2, g_3$ . Let  $U_\gamma(h)$  be the momentum of  $h$  given by (3.17), and assume that  $\gamma$  is parametrized by the canonical parameter  $s$ , i.e.,  $\omega = ds$ . According to whether  $\det U_\gamma(h)$  is zero or different from zero, we distinguish two cases.*

*Case I.* If  $\det U(h) \neq 0$ , then the canonical frame field  $\Gamma : I \rightarrow \text{SL}(2, \mathbb{C})$  along  $\gamma$  is given by

$$\Gamma(s) = A \cdot M(s),$$

where  $A \in \text{SL}(2, \mathbb{C})$  and  $M(s)$  takes the form

$$\frac{1}{\sqrt{-4i\varepsilon\nu}} \begin{pmatrix} e^{\phi(m, h)} & 0 \\ 0 & e^{-\phi(m, h)} \end{pmatrix} \begin{pmatrix} \frac{ih' + \nu}{\sqrt{h - \varepsilon \left( \frac{m}{3} + i \right)}} & 2i\varepsilon\sqrt{h - \varepsilon \left( \frac{m}{3} + i \right)} \\ \frac{-ih' + \nu}{\sqrt{h - \varepsilon \left( \frac{m}{3} + i \right)}} & -2i\varepsilon\sqrt{h - \varepsilon \left( \frac{m}{3} + i \right)} \end{pmatrix}$$

*Case II.* If  $\det U(h) = 0$ , then the canonical frame field  $\Gamma : I \rightarrow \text{SL}(2, \mathbb{C})$  along  $\gamma$  is given by

$$\Gamma(s) = A \cdot M(s),$$

where  $A \in \text{SL}(2, \mathbb{C})$  and  $M(s)$  takes the form

$$\frac{1}{\sqrt{-2i\varepsilon}} \begin{pmatrix} 1 & -\phi(m, h) \\ \sqrt{h - \varepsilon \left( \frac{m}{3} + i \right)} & 2i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -ih' & -2i\varepsilon\sqrt{h - \varepsilon \left( \frac{m}{3} + i \right)} \\ \sqrt{h - \varepsilon \left( \frac{m}{3} + i \right)} & 1 \end{pmatrix}$$



*Proof of Case I.* Let  $\Gamma = (C_1, C_2) : I \rightarrow \mathrm{SL}(2, \mathbb{C})$  be a canonical frame along  $\gamma$  and  $U_\gamma(h)$  be the momentum of  $\gamma$  given by (3.17). Consider the eigenvalues  $\pm\nu(m, h)$  of  $U_\gamma(h)$  and denote by  $\mathbf{L}_\pm$  the corresponding eigenspaces. From the definition of  $U_\gamma(h)$ , it follows that

$$\begin{aligned} L_+ &= -2i\varepsilon \left( h - \varepsilon \left( \frac{m}{3} + i \right) \right) C_1 + (ih' - \nu(m, h)) C_2 : I \rightarrow \mathbf{L}_+, \\ L_- &= -2i\varepsilon \left( h - \varepsilon \left( \frac{m}{3} + i \right) \right) C_1 + (ih' + \nu(m, h)) C_2 : I \rightarrow \mathbf{L}_- \end{aligned}$$

are eigenvectors of  $U_\gamma(h)$  corresponding to  $\nu(m, h)$  and  $-\nu(m, h)$ , respectively. Thus, we must have

$$L'_+ = \rho_1 L_+, \quad L'_- = \rho_2 L_-$$

for analytic functions  $\rho_1, \rho_2$ . Using the Maurer–Cartan equation  $\Gamma' = \Gamma H(m, h)$ , we compute

$$L'_+ = \frac{h' + \nu(m, h)}{2(h - \varepsilon(\frac{m}{3} + i))} L_+, \quad L'_- = \frac{h' - \nu(m, h)}{2(h - \varepsilon(\frac{m}{3} + i))} L_-.$$

We thus see that the two vectors

$$\begin{aligned} \Lambda_1 &:= \exp \left( - \int \frac{h' + \nu(m, h)}{2(h - \varepsilon(\frac{m}{3} + i))} ds \right) L_+, \\ \Lambda_2 &:= \exp \left( - \int \frac{h' - \nu(m, h)}{2(h - \varepsilon(\frac{m}{3} + i))} ds \right) L_- \end{aligned}$$

are constant along  $\gamma$ . By (4.2), they become

$$\Lambda_1 = \frac{\exp(-\phi(m, h))}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} L_+, \quad \Lambda_2 = \frac{\exp(\phi(m, h))}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} L_-.$$

Hence

$$\Gamma \cdot R(m, h) \cdot S(m, h) = \Lambda = (\Lambda_1, \Lambda_2),$$

where

$$R(m, h) = \begin{pmatrix} -2i\varepsilon(h - \varepsilon(\frac{m}{3} + i)) & -2i\varepsilon(h - \varepsilon(\frac{m}{3} + i)) \\ ih' - \nu(m, h) & ih' + \nu(m, h) \end{pmatrix}$$

and

$$S(m, h) = \begin{pmatrix} \frac{1}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} & 0 \\ 0 & \frac{1}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} \end{pmatrix} \begin{pmatrix} \exp(-\phi(m, h)) & 0 \\ 0 & \exp(\phi(m, h)) \end{pmatrix}.$$

From this, we obtain

$$\Gamma \begin{pmatrix} -2i\varepsilon\sqrt{h - \varepsilon(\frac{m}{3} + i)} & -2i\varepsilon\sqrt{h - \varepsilon(\frac{m}{3} + i)} \\ \frac{ih' - \nu}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} & \frac{ih' + \nu}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} \end{pmatrix} \begin{pmatrix} e^{-\phi(m, h)} & 0 \\ 0 & e^{\phi(m, h)} \end{pmatrix} = \Lambda,$$

and hence

$$\Gamma = \tilde{\Lambda} \begin{pmatrix} e^{\phi(m,h)} & 0 \\ 0 & e^{-\phi(m,h)} \end{pmatrix} \begin{pmatrix} \frac{ih' + \nu}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} & 2i\varepsilon\sqrt{h - \varepsilon(\frac{m}{3} + i)} \\ \frac{-ih' + \nu}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} & -2i\varepsilon\sqrt{h - \varepsilon(\frac{m}{3} + i)} \end{pmatrix}. \quad \square$$

*Proof of Case II.* Again, let  $\Gamma = (C_1, C_2) : I \rightarrow \mathrm{SL}(2, \mathbb{C})$  be a canonical frame along  $\gamma$  and  $U_\gamma(h)$  be the momentum of  $\gamma$ . If  $\nu(m, h) = 0$ , then

$$L_1 = -2i\varepsilon \left( h - \varepsilon \left( \frac{m}{3} + i \right) \right) C_1 + ih' C_2$$

belongs to the kernel of  $U_\gamma(h)$ , and proceeding as in Case I, we see that the vector

$$(4.3) \quad \Lambda_1 = \frac{1}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} L_1$$

is a first integral. In order to find another first integral, we look for analytic functions  $f$  and  $g$  such that

$$(4.4) \quad \Lambda_2 := gC_2 + fL_1$$

is a constant vector. Differentiating and using the Maurer–Cartan equation  $\Gamma' = \Gamma H(m, h)$ , we obtain

$$g'C_2 + g\varepsilon C_1 + f'L_1 = 0,$$

from which we compute

$$g = \frac{1}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}}, \quad f = \frac{1}{2i} \int \frac{ds}{h - \varepsilon(\frac{m}{3} + i)} = \frac{1}{2i} \phi(m, h).$$

Now, from (4.3) and (4.4), we obtain

$$\Gamma \begin{pmatrix} -2i\varepsilon\sqrt{h - \varepsilon(\frac{m}{3} + i)} & 0 \\ \frac{ih'}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2i}\phi(m, h) \\ 0 & \frac{1}{\sqrt{h - \varepsilon(\frac{m}{3} + i)}} \end{pmatrix} = \Lambda = (\Lambda_1, \Lambda_2),$$

and hence the required result.  $\square$

**Acknowledgments.** The authors would like to thank the anonymous referees for their useful comments and suggestions.

#### REFERENCES

- [1] R. L. BRYANT, *On notions of equivalence of variational problems with one independent variable*, Contemp. Math., 68 (1987), pp. 65–76.
- [2] R. L. BRYANT AND P. A. GRIFFITHS, *Reduction for constrained variational problems and  $\int \frac{k^2}{2}$* , Amer. J. Math., 108 (1986), pp. 525–570.
- [3] E. CARTAN, *Sur un problème du Calcul des variations en Géométrie projective plane*, in *Oeuvres Complètes*, Partie III, Vol. 2, Gauthier–Villars, Paris, 1955, pp. 1105–1119.

- [4] E. CARTAN, *Leçons sur les invariants intégraux*, Hermann, Paris, 1922.
- [5] M. CASTAGNINO, *Sulle formule di Frenet-Serret per le curve nulle di una  $V_4$  riemanniana a metrica iperbolica normale*, Rend. Mat. e Appl. (5), 23 (1964), pp. 438–461.
- [6] A. FERRÁNDEZ, A. GIMÉNEZ, AND P. LUCAS, *Null helices in Lorentzian space forms*, Internat. J. Modern Phys. A, 16 (2001), pp. 4845–4863.
- [7] A. FERRÁNDEZ, A. GIMÉNEZ, AND P. LUCAS, *Geometrical particle models on 3D null curves*, Phys. Lett. B, 543 (2002), pp. 311–317; also available online at <http://www.arxiv.org/abs/hep-th/0205284>.
- [8] R. G. GARDNER, *Differential geometric methods interfacing control theory*, in Differential Geometric Control Theory (Houghton, MI, 1982), Progr. Math. 27, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhäuser Boston, Inc., Boston, MA, 1983, pp. 117–180.
- [9] R. B. GARDNER, *The Method of Equivalence and Its Applications*, CBMS-NSF Reg. Conf. Ser. Appl. Math. 58, SIAM, Philadelphia, 1989.
- [10] J. D. E. GRANT AND E. MUSSO, *Coisotropic variational problems*, J. Geom. Phys., 50 (2004), pp. 303–338; also available online at <http://www.arxiv.org/abs/math.DG/0307216>.
- [11] P. A. GRIFFITHS, *Exterior Differential Systems and the Calculus of Variations*, Progr. Math. 25, Birkhäuser Boston, Inc., Boston, MA, 1982.
- [12] L. HSU, *Calculus of variations via the Griffiths formalism*, J. Differential Geom., 36 (1992), pp. 551–589.
- [13] J. LANGER AND D. SINGER, *Liouville integrability of geometric variational problems*, Comment. Math. Helv., 69 (1994), pp. 272–280.
- [14] D. F. LAW DEN, *Elliptic Functions and Applications*, Appl. Math. Sci. 80, Springer-Verlag, New York, 1989.
- [15] E. MUSSO AND L. NICOLodi, *Reduction for the projective arclength functional*, Forum Math., 17 (2005), pp. 569–590.
- [16] E. MUSSO AND L. NICOLodi, *Closed trajectories of a particle model on null curves in anti-de Sitter 3-space*, Classical Quantum Gravity, 24 (2007), pp. 5401–5411; also available online at <http://www.arxiv.org/abs/math.DG/0709.2017>.
- [17] A. NERSESSIAN, R. MANVELYAN, AND H. J. W. MÜLLER-KIRSTEN, *Particle with torsion on 3d null-curves*, Nuclear Phys. B, 88 (2000), pp. 381–384; also available online at <http://www.arxiv.org/abs/hep-th/9912061>.
- [18] A. NERSESSIAN AND E. RAMOS, *Massive spinning particles and the geometry of null curves*, Phys. Lett. B, 445 (1998), pp. 123–128; also available online at <http://www.arxiv.org/abs/hep-th/9807143>.
- [19] M. S. PLYUSHCHAY, *The model of the relativistic particle with torsion*, Nuclear Phys. B, 362 (1991), pp. 54–72.
- [20] E. PICARD, *Sur les équations différentielles linéaires à coefficients doublement périodiques*, J. Reine Angew. Math., 90 (1881), pp. 281–302.

# ON THE BOUNDEDNESS OF THE SPECTRAL FACTORIZATION MAPPING ON DECOMPOSING BANACH ALGEBRAS\*

HOLGER BOCHE<sup>†</sup> AND VOLKER POHL<sup>‡</sup>

**Abstract.** In [SIAM J. Control Optim., 40 (2001), pp. 88–106], Jacob and Partington studied the continuity and boundedness of the spectral factorization mapping on decomposing Banach algebras. Many function spaces considered in systems theory are decomposing Banach algebras. The most well known example is the Wiener algebra, the space of all absolutely convergent Fourier series. Jacob and Partington showed in the above paper that the spectral factorization is locally Lipschitz continuous on all decomposing algebras, but unbounded on the most important examples of decomposing algebras. Our paper gives an extension of this result and shows that the spectral factorization mapping is unbounded on every decomposing algebra.

**Key words.** spectral factorization, boundedness of the factorization, decomposing Banach algebras

**AMS subject classifications.** 47A68, 46J10, 46J15

**DOI.** 10.1137/070680485

**1. Introduction.** Let  $f$  be a real valued function given on the unit circle. Then the spectral factorization of  $f$  is the process by which a function  $f_+$  is determined such that  $f$  can be written as  $f = f_+ f_+^*$ , in which the spectral factor  $f_+$  is an outer function and  $f_+^*$  its para-Hermitian conjugate. This operation arises in many different applications. In communications and signal processing, it is an important tool for filter design [6, 9, 11, 23] and spectral estimation [5, 17]. In control theory, it is closely related to  $H_\infty$  and quadratic optimal control [10, 20]. Moreover, it plays a prominent roll in the theory of stochastic processes [8, 24].

The question of whether the spectral factorization mapping  $\mathfrak{S} : f \mapsto f_+$  is continuous and bounded is of fundamental importance in practical applications since these properties are linked to the robustness of the operation with respect to disturbances of the given data and to the stability of the resulting spectral factor (see, e.g., [3, 13]). Whether the spectral factorization mapping is continuous or bounded depends crucially on the normed space on which one is working. Consequently, there exists several results [1, 2, 16] on the continuity of  $\mathfrak{S}$  on different function spaces. In [7], Clancy and Gohberg introduced a general class of Banach algebras of continuous functions, so-called decomposing Banach algebras, on which the spectral factorization exists. In [18] a slightly more restricted definition of a decomposing Banach algebra was given and used to study the problem of best approximation of analytic function. In fact, many of the most important Banach algebras used in system theory are decomposing. Examples include the Wiener algebra, the set of Hölder continuous functions, and the space of functions of vanishing mean oscillation (VMO).

---

\*Received by the editors January 19, 2007; accepted for publication (in revised form) December 17, 2007; published electronically April 4, 2008. This work was partly supported by the German Research Foundation (DFG) under grants BO 1734/11–2 and PO 1347/1–1.

<http://www.siam.org/journals/sicon/47-3/68048.html>

<sup>†</sup>Heinrich-Hertz Chair for Mobile Communications, Technical University Berlin, Einsteinufer 25, 10587 Berlin, Germany (holger.boche@mk.tu-berlin.de).

<sup>‡</sup>Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel (pohl@ee.technion.ac.il).

The present paper was motivated by the work of Jacob and Partington [15], where the continuity and boundedness of the spectral factorization mapping on decomposing Banach algebras was investigated in depth. It was shown that the spectral factorization is locally Lipschitz continuous on all decomposing Banach algebras, but since the spectral factorization mapping is nonlinear, this continuity does not imply the boundedness of  $\mathfrak{S}$ . However, in [15] a necessary condition for the boundedness of  $\mathfrak{S}$  on decomposing Banach algebras was derived. Based on this condition, it was shown that the spectral factorization is unbounded on a large number of concrete Banach algebras. In fact, [15] contains no example of a decomposing Banach algebra on which the spectral factorization is bounded. This observation may suggest that there exists no decomposing Banach algebra at all on which the spectral factorization is bounded. This conjecture will be proved in the present paper. It will be shown that the condition for the boundedness of the spectral factorization mapping given in [15] is contradictory to the axioms of a decomposing Banach algebra.

The outline of the paper is as follows. Sections 2 and 3 briefly review the concept of decomposing Banach algebras and of the spectral factorization, respectively. Section 4 gives an extension of the boundedness condition in [15] and derives some consequences regarding Fourier series. Based on these consequences, we prove in section 5 that there exists no decomposing Banach algebra on which the spectral factorization is bounded. The paper closes with some discussions in section 6 and a conclusion in section 7.

**2. Decomposing Banach algebras.** This section introduces notation and briefly recalls the concept of decomposing Banach algebras. Moreover, it states some known results which are used frequently in this paper and gives some examples of decomposing algebras.

Throughout this paper,  $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$  and  $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$  denote the unit disc and the unit circle in the complex plane, respectively. The imaginary unit is denoted by  $i = \sqrt{-1}$ , and  $\bar{z}$  is the conjugate complex of  $z$ . The set  $\mathcal{C}(\mathbb{T})$  of all continuous functions on  $\mathbb{T}$  is a Banach space under the usual supremum norm  $\|f\|_\infty = \sup_{\zeta \in \mathbb{T}} |f(\zeta)|$ . The Banach space of  $p$ -integrable functions ( $1 < p < \infty$ ) on  $\mathbb{T}$  with the usual norm  $\|f\|_p := (\frac{1}{2\pi} \int_{\mathbb{T}} |f(\zeta)|^p d\zeta)^{1/p}$  is denoted by  $L^p(\mathbb{T})$ . Moreover,  $L^\infty(\mathbb{T})$  denotes the set of all measurable and essentially bounded functions on  $\mathbb{T}$ . For any function  $f \in L^1(\mathbb{T})$ , the *Fourier coefficients* are defined by

$$(2.1) \quad \hat{f}(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{i\tau}) e^{-ik\tau} d\tau, \quad k = 0, \pm 1, \pm 2, \dots$$

Let  $\mathcal{A} \subset L^1(\mathbb{T})$  be a commutative Banach algebra with unity and with norm  $\|\cdot\|_{\mathcal{A}}$ . We define the two subspaces  $\mathcal{A}_+$  and  $\mathcal{A}_-$  of  $\mathcal{A}$  by

$$\mathcal{A}_+ := \{f \in \mathcal{A} : \hat{f}(k) = 0 \text{ for all } k < 0\},$$

$$\mathcal{A}_- := \{f \in \mathcal{A} : \hat{f}(k) = 0 \text{ for all } k \geq 0\}.$$

In particular, let  $\mathcal{A} = L^p(\mathbb{T})$ ,  $1 < p \leq \infty$ ; then  $[L^p(\mathbb{T})]_+$  coincides with the *Hardy space*  $H^p(\mathbb{D})$  of functions in  $L^p(\mathbb{T})$  which are analytic in  $\mathbb{D}$ . The mapping

$$(2.2) \quad \sum_{k=-\infty}^{\infty} \hat{f}(k) z^k \mapsto \sum_{k=0}^{\infty} \hat{f}(k) z^k,$$

which assigns to every  $f \in \mathcal{A}$  with Fourier coefficients (2.1) a corresponding function in  $\mathcal{A}_+$ , will be denoted by  $\mathfrak{P}_+$  and is called the *Riesz projection*. Inserting the Fourier

coefficients (2.1) into (2.2) yields a closed form integral representation of the Riesz projection  $\mathfrak{P}_+ : \mathcal{A} \rightarrow \mathcal{A}_+$ :

$$(\mathfrak{P}_+ f)(z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{i\tau}) \frac{e^{i\tau}}{e^{i\tau} - z} d\tau.$$

A linear functional  $h : \mathcal{A} \rightarrow \mathbb{C}$  is called a *homomorphism* (a multiplicative linear functional) on  $\mathcal{A}$  if  $h(\phi\psi) = h(\phi)h(\psi)$  for all  $\phi, \psi \in \mathcal{A}$ . The set of all homomorphisms on  $\mathcal{A}$  will be denoted by  $\Gamma(\mathcal{A})$ . Moreover,  $\mathcal{G}(\mathcal{A})$  will stand for the set of all invertible elements of  $\mathcal{A}$ . Thus  $\mathcal{G}(\mathcal{A})$  is the set of all  $f \in \mathcal{A}$  for which there exists a function  $f^{-1} \in \mathcal{A}$  such that  $f \cdot f^{-1} = \mathbf{1}$ , where  $\mathbf{1}$  denotes the unity of  $\mathcal{A}$ . Accordingly,  $\mathcal{G}(\mathcal{A}_+)$  denotes the set of all invertible elements in  $\mathcal{A}_+$ . Let  $f \in \mathcal{A}$  and define  $(\exp f)(\zeta) := \sum_{k=0}^{\infty} \frac{1}{k!} f(\zeta)^k$ . By the properties of a Banach algebra, it is clear that  $\|\exp f\|_{\mathcal{A}} \leq \exp(\|f\|_{\mathcal{A}})$ ; from what follows, that  $\exp f \in \mathcal{A}$ . Moreover, it holds that  $\exp(f) \in \mathcal{G}(\mathcal{A})$ . Another important subset of  $\mathcal{A}$  that will be needed frequently is the set

$$\mathcal{A}_{\text{pos}} = \{f \in \mathcal{A} : f(\zeta) > 0 \text{ for all } \zeta \in \mathbb{T}\}$$

of all strictly positive functions in  $\mathcal{A}$ .

Next, we recall the notion of decomposing Banach algebras and review some of their most important properties. In general, a decomposing Banach algebra is a Banach algebra  $\mathcal{A}$  with an identity which can be written as a direct sum  $\mathcal{A} = \mathcal{A}_+ \oplus \mathcal{A}_-$  of two closed subalgebras  $\mathcal{A}_+$  and  $\mathcal{A}_-$  [7]. However, as in [15], we are interested in decomposing Banach algebras  $\mathcal{B}$ , which are subspaces of  $L^2(\mathbb{T})$ . Then, a decomposing Banach algebra is defined by the following four axioms [15].

DEFINITION 2.1. *A commutative Banach algebra  $\mathcal{B} \subset L^2(\mathbb{T})$  is called a decomposing Banach algebra if the following hold:*

- (A1) *If  $f \in \mathcal{B}$ , then  $\bar{f} \in \mathcal{B}$  and  $\mathfrak{P}_+ f \in \mathcal{B}$ .*
- (A2)  *$\mathcal{B}$  is a Banach algebra with respect to pointwise multiplication.*
- (A3) *The set of all trigonometric polynomials is dense in  $\mathcal{B}$ .*
- (A4) *Every multiplicative functional on  $\mathcal{B}$  coincides with a functional  $f \mapsto f(\zeta)$  defined as the value  $f(\zeta)$  at some point  $\zeta \in \mathbb{T}$ .*

Note that this definition coincides with the definition given by Peller and Khrushchev in [18]. The following results recall some essential properties of decomposing Banach algebras.

PROPOSITION 2.2. *Let  $\mathcal{B}$  be a decomposing Banach algebra. Every  $f \in \mathcal{B}$  with  $f(\zeta) \neq 0$  for all  $\zeta \in \mathbb{T}$  belongs to  $\mathcal{G}(\mathcal{B})$ .*

*Proof.* Let  $\zeta \in \mathbb{T}$  be arbitrary; then (A2) implies that the functional  $h : \mathcal{B} \rightarrow \mathbb{C}$  defined by  $h(\phi) := \phi(\zeta)$  is a homomorphism on  $\mathcal{B}$ , and (A4) shows that all possible homomorphisms on  $\mathcal{B}$  are obtained in this way. Therefore,  $\phi(\zeta) \neq 0$  for all  $\zeta \in \mathbb{T}$  implies  $h(\phi) \neq 0$  for all  $h \in \Gamma(\mathcal{B})$ , and by the Beurling–Gelfand theorem,  $\phi \in \mathcal{B}$  is invertible in  $\mathcal{B}$  if and only if  $h(\phi) \neq 0$  for all  $h \in \Gamma(\mathcal{B})$ ; see, e.g., [19, section 11.5].  $\square$

PROPOSITION 2.3. *Every decomposing Banach algebra  $\mathcal{B}$  is continuously embedded in  $\mathcal{C}(\mathbb{T})$  with*

$$\|f\|_{\infty} \leq \|f\|_{\mathcal{B}} \quad \text{for all } f \in \mathcal{B}.$$

*Proof.* By (A2) and (A4), for every  $\zeta \in \mathbb{T}$  there exists an  $h \in \Gamma(\mathcal{B})$  such that  $h(\phi) = \phi(\zeta)$  for every  $\phi \in \mathcal{B}$ . Since each complex homomorphism of  $\mathcal{B}$  is bounded by 1 (see, e.g., [19, section 10.7]), one has  $|\phi(\zeta)| = |h(\phi)| \leq \|\phi\|_{\mathcal{B}}$ , which implies  $\|\phi\|_{\infty} \leq \|\phi\|_{\mathcal{B}}$ . By (A3), for every  $\epsilon > 0$  there exists a trigonometric polynomial  $\psi$  such that  $\|\phi - \psi\|_{\infty} \leq \|\phi - \psi\|_{\mathcal{B}} < \epsilon$ , which proves that  $\phi \in \mathcal{C}(\mathbb{T})$ .  $\square$

The next lemma, taken from [15], will be used frequently in what follows.

LEMMA 2.4. *Let  $\mathcal{B}$  be a decomposing Banach algebra. Then there exists a constant  $m_1 > 0$  such that*

$$\|\bar{f}\|_{\mathcal{B}} \leq m_1 \|f\|_{\mathcal{B}} \quad \text{for all } f \in \mathcal{B}.$$

From now on, the symbol  $\mathcal{B}$  always stands for a decomposing Banach algebra. The unity in  $\mathcal{B}$  will be denoted by  $\mathbf{1}$  is given by  $\mathbf{1}(\zeta) = 1$  for all  $\zeta \in \mathbb{T}$ .

We complete this section with some examples of decomposing and nondecomposing Banach algebras. All these examples can be found in [15] or [18] along with their corresponding proofs.

Example 2.5 (Wiener algebra). The Wiener algebra  $\mathcal{W}$  is the set of all functions of the form

$$f(e^{i\omega}) = \sum_{k=-\infty}^{\infty} a_k e^{ik\omega},$$

which are bounded with respect to the norm

$$\|f\|_{\mathcal{W}} = \sum_{k=-\infty}^{\infty} |a_k|.$$

The Wiener algebra  $\mathcal{W}$  is a decomposing Banach algebra.

Example 2.6 (Hölder continuous functions). For  $0 < \alpha < 1$ , denote by  $\Lambda_{\alpha}$  the set of all functions on  $\mathbb{T}$  for which the norm

$$\|f\|_{\alpha} = \|f\|_{\infty} + \sup_{\zeta, z \in \mathbb{T}, \zeta \neq z} \frac{|f(\zeta) - f(z)|}{|\zeta - z|^{\alpha}}$$

is bounded. The closure of all trigonometric polynomials under the norm  $\|\cdot\|_{\alpha}$  is denoted by  $\lambda_{\alpha}$ , and it can be shown that  $\lambda_{\alpha}$  is a decomposing Banach algebra.

Example 2.7 (nondecomposing Banach algebras). The set of continuous functions  $\mathcal{C}(\mathbb{T})$  on the unit circle  $\mathbb{T}$  is a Banach algebra under pointwise multiplication and satisfies also the axioms (A3) and (A4) of a decomposing Banach algebra. However, it does not satisfy the second part of property (A1) since there exist functions  $f \in \mathcal{C}(\mathbb{T})$  for which the Riesz projection  $\mathfrak{P}_+ f$  is not continuous. This follows, e.g., from the counterexample given in [21], or it can be seen from the last paragraph of the proof of Theorem 5.3. The same holds also for  $L^{\infty}(\mathbb{T})$ . It does not satisfy the second part of (A1) since there exists no bounded projection  $L^{\infty}(\mathbb{T})$  to  $H^{\infty}(\mathbb{D})$  [14, Chapter 9].

**3. Spectral factorization.** This sections briefly reviews the notation and definitions of the spectral factorization, as far as it is needed in this paper.

DEFINITION 3.1. *Let  $\mathcal{B}$  be a Banach algebra of functions on the unit circle  $\mathbb{T}$ . An element  $f \in \mathcal{B}$  possesses a spectral factorization if there exists an invertible element  $f_+ \in \mathcal{G}(\mathcal{B}_+)$  such that*

$$(3.1) \quad f(e^{i\omega}) = f_+(e^{i\omega}) \overline{f_+(e^{i\omega})} \quad \text{for all } \omega \in [-\pi, \pi].$$

The function  $f_+$  is called a spectral factor of  $f$ , and every  $f \in \mathcal{B}$  which possesses a spectral factorization is called a spectral density.

Which elements of a decomposing Banach algebra  $\mathcal{B}$  are spectral densities? The answer is given by the following proposition, which is taken from [15].

PROPOSITION 3.2. *Let  $\mathcal{B}$  be a decomposing Banach algebra. An element  $f \in \mathcal{B}$  is a spectral density if and only if  $f \in \mathcal{B}_{\text{pos}}$ . One spectral factor is given by*

$$(3.2) \quad f_+(z) = (\mathfrak{S}f)(z) = \exp \left( \frac{1}{4\pi} \int_{-\pi}^{\pi} \log f(e^{i\tau}) \frac{e^{i\tau} + z}{e^{i\tau} - z} d\tau \right), \quad z \in \mathbb{D}.$$

*This spectral factor is unique up to a constant of modulus 1.*

Remark 3.3. The above proposition holds also for some nondecomposing Banach algebras. It holds, e.g., for  $L^\infty(\mathbb{T})$ , but it does not hold for  $\mathcal{C}(\mathbb{T})$  because, as is well known [14, 21], there exist functions  $f \in \mathcal{C}(\mathbb{T})$  with  $f(\zeta) > 0$  for all  $\zeta \in \mathbb{T}$  but for which the spectral factor (3.2) is not continuous.

Remark 3.4. Note that the spectral factorization mapping (3.2) is well defined already for all functions  $f \in \mathcal{B}$  which satisfy the so-called *Paley–Wiener condition*,  $\int_{\mathbb{T}} \log f(\zeta) d\zeta > -\infty$ . However, if  $f \in \mathcal{B}$  had a zero on  $\mathbb{T}$ , then by (3.1) the function  $\mathfrak{S}f$  also would have a zero on  $\mathbb{T}$ , and because of property (A4) of a decomposing Banach algebra,  $\mathfrak{S}f$  would not be invertible in  $\mathcal{B}_+$ . By Definition 3.1 of the spectral factorization, this implies that  $\mathfrak{S}f$  is no spectral factor of  $f$  in this case.

Note that the spectral factorization mapping (3.2) can also be written in terms of the Riesz projection  $\mathfrak{P}_+$ :

$$(3.3) \quad \begin{aligned} f_+(z) &= (\mathfrak{S}f)(z) = \exp \left[ (\mathfrak{P}_+[\log f])(z) - \frac{1}{2}(\mathfrak{P}_+[\log f])(0) \right] \\ &= A(f) \exp [(\mathfrak{P}_+[\log f])(z)] \end{aligned}$$

with the constant  $A(f)$ , which is determined by the zeroth Fourier coefficient of  $\log f$ . Note also that the spectral factorization mapping is nonlinear. Therefore, for every constant  $\alpha > 0$ , it holds in particular that  $\mathfrak{S}(\alpha f) = \mathfrak{S}(\alpha) \mathfrak{S}(f) = \sqrt{\alpha} \mathfrak{S}(f)$ .

**4. Basic results.** This paper investigates the boundedness of the spectral factorization mapping on decomposing Banach algebras. The boundedness of the mapping guarantees that the norm of the spectral factor is small whenever the norm of the spectral density is small.

DEFINITION 4.1. *Let  $\mathcal{B}$  be a decomposing Banach algebra. Then the spectral factorization mapping  $\mathfrak{S} : \mathcal{B} \rightarrow \mathcal{B}_+$  is said to be bounded if there exists a constant  $C < \infty$  such that for every  $f \in \mathcal{B}_{\text{pos}}$  with  $\|f\|_{\mathcal{B}} \leq 1$ , it always holds that*

$$(4.1) \quad \|f_+\|_{\mathcal{B}} = \|\mathfrak{S}f\|_{\mathcal{B}} \leq C.$$

In [15] a simple equivalent condition for the boundedness of the spectral factorization on a decomposing Banach algebra was given as follows.

PROPOSITION 4.2. *The spectral factorization mapping is bounded on a decomposing algebra  $\mathcal{B}$  if and only if there exists a constant  $m_2 > 0$  such that*

$$(4.2) \quad m_2 \|h\|_{\mathcal{B}}^2 \leq \|h \bar{h}\|_{\mathcal{B}} \quad \text{for all } h \in \mathcal{G}(\mathcal{B}_+).$$

Thus, in order for the spectral factorization to be bounded on the decomposing Banach algebra  $\mathcal{B}$ , condition (4.2) has to be satisfied for all functions  $h \in \mathcal{G}(\mathcal{B}_+)$ . It was annotated in [15] that condition (4.2) cannot be satisfied for all elements of the decomposing algebra  $\mathcal{B}$  because this would imply that  $\mathcal{B}$  is isomorphic to  $\mathcal{C}(\mathbb{T})$ , which is not a decomposing Banach algebra. However, our paper will show that there exists no decomposing algebra on which the spectral factorization mapping is



bounded. This will be done in several steps: In the following subsection, we show that the boundedness of the spectral factorization implies that condition (4.2) holds for all  $f \in \mathcal{B}_+$ , and in section 4.2 we derive consequences of this extension with respect to Fourier series in decomposing Banach algebras in which the spectral factorization is bounded. These results are used in section 5 to show that condition (4.2) holds even for all  $f \in \mathcal{B}$ . Then, the above cited argument from [15] implies immediately that there exists no decomposing Banach algebra on which the spectral factorization is bounded.

**4.1. Extension to  $\mathcal{B}_+$ .** The result of this subsection shows that condition (4.2) for the boundedness of the spectral factorization on a decomposing Banach algebra can be extended to every  $f \in \mathcal{B}_+$ .

LEMMA 4.3. *The spectral factorization is bounded on a decomposing Banach algebra  $\mathcal{B}$  if and only if there exists a constant  $m_3 > 0$  such that*

$$(4.3) \quad m_3 \|f\|_{\mathcal{B}}^2 \leq \|f \bar{f}\|_{\mathcal{B}} \quad \text{for all } f \in \mathcal{B}_+ .$$

Remark 4.4. Lemmas 4.3 and 2.4 together give the relation

$$m_3 \|f\|_{\mathcal{B}}^2 \leq \|f \bar{f}\|_{\mathcal{B}} \leq m_1 \|f\|_{\mathcal{B}}^2 \quad \text{for all } f \in \mathcal{B}_+$$

in every decomposing Banach algebra  $\mathcal{B}$  on which the spectral factorization is bounded.

*Proof.* Since  $\mathcal{G}(\mathcal{B}_+) \subset \mathcal{B}_+$ , Proposition 4.2 implies that (4.3) is sufficient for the boundedness of the spectral factorization. The necessity of (4.3) is shown by contradiction: Assume that the spectral factorization is bounded but that there exists no constant  $m_3 > 0$  such that (4.3) holds. Then, for every  $\epsilon > 0$  there exists a  $g \in \mathcal{B}_+$  with  $\|g\|_{\mathcal{B}} = 1$  and such that  $\|g \bar{g}\|_{\mathcal{B}} \leq \epsilon^2$ . Moreover, it holds obviously that  $\|g \bar{g}\|_{\infty} = \sup_{\zeta \in \mathbb{T}} |g(\zeta)|^2 = \|g\|_{\infty}^2$ . These facts and Proposition 2.3 together give

$$(4.4) \quad \|g\|_{\infty}^2 = \|g \bar{g}\|_{\infty} \leq \|g \bar{g}\|_{\mathcal{B}} \leq \epsilon^2 .$$

Next, we consider the function  $f := g + \|g\|_{\infty} + \epsilon$ . By this definition, it is clear that  $|f(z)| \geq \epsilon$  for all  $z \in \mathbb{D}$ , i.e.,  $f \in \mathcal{G}(\mathcal{B}_+)$ . Using the triangle inequality and (4.4), one gets a lower bound and an upper bound for the norm of this function,

$$(4.5) \quad 1 - 2\epsilon \leq \|f\|_{\mathcal{B}} \leq 1 + 2\epsilon .$$

Therewith, one obtains that

$$\begin{aligned} \|f \bar{f}\|_{\mathcal{B}} &\leq \|g \bar{g}\|_{\mathcal{B}} + (\|g\|_{\infty} + \epsilon) \|g + \bar{g}\|_{\mathcal{B}} + (\|g\|_{\infty} + \epsilon)^2 \|1\|_{\mathcal{B}} \\ &\leq \epsilon^2 + 2\epsilon(1 + m_1) + 4\epsilon^2 \|1\|_{\mathcal{B}} = C_1 \epsilon^2 + C_2 \epsilon \end{aligned}$$

with the constants  $C_1 = 1 + 4\|1\|_{\mathcal{B}}$  and  $C_2 = 2(1 + m_1)$  and in which inequality (4.4) and Lemma 2.4 were used to obtain the second line. On the other hand, since it was assumed that the spectral factorization is bounded, and since  $f \in \mathcal{G}(\mathcal{B}_+)$ , one can apply Proposition 4.2. Together with (4.5) this gives the lower bound

$$\|f \bar{f}\|_{\mathcal{B}} \geq m_2 \|f\|_{\mathcal{B}}^2 \geq m_2 (1 - 2\epsilon)^2$$

with  $m_2 > 0$ . Combining the lower bound and the upper bound for  $\|f \bar{f}\|_{\mathcal{B}}$ , one obtains that

$$0 < m_2 \leq \epsilon \frac{C_1 \epsilon + C_2}{(1 - 2\epsilon)^2} .$$

The right-hand side of this expression tends to zero as  $\epsilon \rightarrow 0$ . Therefore, this last inequality gives a contradiction for sufficiently small  $\epsilon$ . This shows that if the spectral factorization is bounded on  $\mathcal{B}$ , then there exists a constant  $m_3 > 0$  such that (4.3) holds.  $\square$

This extension of the statement of Proposition 4.2 from  $\mathcal{G}(\mathcal{B}_+)$  onto the whole  $\mathcal{B}_+$  has some important consequences for decomposing Banach algebras on which the spectral factorization is bounded. Some of these consequences are discussed in the following subsection.

**4.2. Fourier series.** Recall [12, 14] that an *inner function* is a function  $\varphi \in H^\infty(\mathbb{D})$  such that  $|\varphi(\zeta)| = 1$  for almost all  $\zeta \in \mathbb{T}$ . Assume now that  $\varphi \in \mathcal{B}$  is an arbitrary inner function. Then it is clear that  $\varphi \in \mathcal{B}_+$ , and Lemma 4.3 implies that  $m_3 \|\varphi\|_{\mathcal{B}}^2 \leq \|\varphi \bar{\varphi}\|_{\mathcal{B}} = \|\mathbf{1}\|_{\mathcal{B}}$ . This shows that in every decomposing Banach algebra  $\mathcal{B}$  on which the spectral factorization is bounded, one has a universal upper bound on the norm for every inner function in  $\mathcal{B}$ ,

$$\|\varphi\|_{\mathcal{B}} \leq \sqrt{\frac{\|\mathbf{1}\|_{\mathcal{B}}}{m_3}} =: C_3,$$

which depends only on the algebra  $\mathcal{B}$ .

Next, we show that in a decomposing Banach algebra  $\mathcal{B}$  on which the spectral factorization is bounded, the Fourier series of every  $f \in \mathcal{B}$  converges uniformly. To this end, define for  $n = 0, 1, 2, \dots$  the function  $s_n(z) = z^n$ . Since the trigonometric polynomials are dense in every decomposing algebra,  $s_n \in \mathcal{B}_+$  for all  $n \in \mathbb{N}$  and every  $s_n$  is obviously an inner function. Moreover, the functions  $s_{-n} := \overline{s_n}$ ,  $n \in \mathbb{N}$ , belong to  $\mathcal{B}$ , and because of Lemmas 2.4 and 4.3, one obtains a uniform upper bound

$$(4.6) \quad \|s_n\|_{\mathcal{B}} \leq m_1 C_3 \quad \text{for all } n \in \mathbb{Z}$$

for the norms of these functions.

Let  $f \in \mathcal{B} \subset L^2(\mathbb{T})$  with Fourier series representation  $f(\zeta) = \sum_{k=-\infty}^{\infty} \hat{f}(k) \zeta^k$ . For every  $n \in \mathbb{N}$ , we consider the operators  $\mathfrak{A}_n : \mathcal{B} \rightarrow \mathcal{B}$  and  $\mathfrak{A}_{-n} : \mathcal{B} \rightarrow \mathcal{B}$  defined by

$$(4.7) \quad \begin{aligned} (\mathfrak{A}_n f)(\zeta) &= \sum_{k=-\infty}^{n-1} \hat{f}(k) \zeta^k = f(\zeta) - s_n(\zeta) [\mathfrak{P}_+(\overline{s_n} f)](\zeta), \\ (\mathfrak{A}_{-n} f)(\zeta) &= \sum_{k=-n}^{\infty} \hat{f}(k) \zeta^k = \overline{s_n}(\zeta) [\mathfrak{P}_+(s_n f)](\zeta). \end{aligned}$$

As  $n$  tends to infinity, both  $\mathfrak{A}_n f$  and  $\mathfrak{A}_{-n} f$  should converge to  $f$ . That this really happens is the statement of the following proposition.

**PROPOSITION 4.5.** *Let  $\mathcal{B}$  be a decomposing Banach algebra on which the spectral factorization is bounded. Then there exist constants  $m_4$  and  $m_5$  such that*

$$(4.8) \quad \|\mathfrak{A}_n f\|_{\mathcal{B}} \leq m_4 \|f\|_{\mathcal{B}} \quad \text{and} \quad \|\mathfrak{A}_{-n} f\|_{\mathcal{B}} \leq m_5 \|f\|_{\mathcal{B}}$$

for all  $f \in \mathcal{B}$  and all  $n \in \mathbb{N}$ . Moreover, for all  $f \in \mathcal{B}$  it holds that

$$(4.9) \quad \lim_{n \rightarrow \infty} \|f - \mathfrak{A}_n f\|_{\mathcal{B}} = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \|f - \mathfrak{A}_{-n} f\|_{\mathcal{B}} = 0.$$

*Proof.* Since  $\mathcal{B}$  is decomposing, the Riesz projection is bounded; i.e., there exists a constant  $C_4$  such that  $\|\mathfrak{P}_+ f\|_{\mathcal{B}} \leq C_4 \|f\|_{\mathcal{B}}$  for all  $f \in \mathcal{B}$ . The uniform boundedness of  $\mathfrak{A}_n$  follows immediately from its definition (4.7) and from the properties of the Banach algebra  $\mathcal{B}$  and the shift functions  $s_n$ :

$$\|\mathfrak{A}_n f\|_{\mathcal{B}} \leq \|f\|_{\mathcal{B}} + C_4 \|s_n\|_{\mathcal{B}} \|\overline{s_n}\|_{\mathcal{B}} \|f\|_{\mathcal{B}} \leq (1 + C_4 m_1^2 C_3^2) \|f\|_{\mathcal{B}}.$$

It is clear that (4.9) holds for all polynomials in  $\mathcal{B}$ , and since the polynomials are dense in  $\mathcal{B}$ , and since the operator sequence  $\mathfrak{A}_n$  is uniformly bounded, (4.9) follows. The statements for  $\mathfrak{A}_{-n}$  are proved in exactly the same way.  $\square$

**5. Boundedness of the spectral factorization.** In this section we finally show that there exists no decomposing Banach algebra on which the spectral factorization is bounded. First, we extend the statement of Lemma 4.3 to the whole decomposing Banach algebra  $\mathcal{B}$ .

LEMMA 5.1. *Let  $\mathcal{B}$  be a decomposing Banach algebra and assume that there exists a constant  $m_3$  such that (4.3) holds. Then there exists a constant  $m_6$  such that*

$$(5.1) \quad m_6 \|f\|_{\mathcal{B}}^2 \leq \|f \bar{f}\|_{\mathcal{B}} \quad \text{for all } f \in \mathcal{B}.$$

*Proof.* Let  $f(\zeta) = \sum_{k=-\infty}^{\infty} \hat{f}(k) \zeta^k$  with  $\zeta \in \mathbb{T}$  be an arbitrary element of  $\mathcal{B}$ , and let  $s_n(\zeta) = \zeta^n$ . Note that the Fourier series of  $f$  (and all the following Fourier series in this proof) converge in the  $\mathcal{B}$ -norm by Proposition 4.5. Applying the upper bound (4.6) for the shift function  $s_n$ , one obtains for an arbitrary  $n \in \mathbb{N}$  that

$$(5.2) \quad \|f\|_{\mathcal{B}} = \|\overline{s_n} s_n f\|_{\mathcal{B}} \leq \|s_{-n}\|_{\mathcal{B}} \|s_n f\|_{\mathcal{B}} \leq m_1 C_3 \|s_n f\|_{\mathcal{B}}.$$

It is easily checked that

$$(5.3) \quad s_n f - \mathfrak{P}_+(s_n f) = s_n \mathfrak{R}_n f \quad \text{with} \quad (\mathfrak{R}_n f)(\zeta) = \sum_{k=-\infty}^{-n-1} \hat{f}(k) \zeta^k.$$

For the operator  $\mathfrak{R}_n$  it holds that  $\|\mathfrak{R}_n f\|_{\mathcal{B}} \rightarrow 0$  as  $n \rightarrow \infty$  for every  $f \in \mathcal{B}$ . This follows from Proposition 4.5 after (5.3) is multiplied by  $\overline{s_n}$ . Therefore, one gets from (5.3) that for every  $\epsilon > 0$  there exists an  $n_0 > 0$  such that

$$\|s_n f\|_{\mathcal{B}} \leq \|\mathfrak{P}_+(s_n f)\|_{\mathcal{B}} + \epsilon \quad \text{for all } n \geq n_0.$$

Together with (5.2) this shows that

$$(5.4) \quad \|f\|_{\mathcal{B}} - m_1 C_3 \epsilon \leq m_1 C_3 \|\mathfrak{P}_+(s_n f)\|_{\mathcal{B}} \quad \text{for all } n \geq n_0.$$

It is clear that  $\mathfrak{P}_+(s_n f) \in \mathcal{B}_+$ . Therefore, by the requirements of Lemma 5.1, inequality (4.3) holds such that

$$(5.5) \quad m_3 \|\mathfrak{P}_+(s_n f)\|_{\mathcal{B}}^2 \leq \|\mathfrak{P}_+(s_n f) \overline{\mathfrak{P}_+(s_n f)}\|_{\mathcal{B}} \\ = \|\overline{s_n} \mathfrak{P}_+(s_n f) \cdot s_n \overline{\mathfrak{P}_+(s_n f)}\|_{\mathcal{B}} = \|g_n \overline{g_n}\|_{\mathcal{B}}$$

wherein  $g_n = \mathfrak{A}_{-n} f = \overline{s_n} \mathfrak{P}_+(s_n f)$ . By Proposition 4.5, it holds therefore that  $\lim_{n \rightarrow \infty} \|f - g_n\|_{\mathcal{B}} = 0$ .

Finally, we consider the expression  $f\bar{f} - g_n\overline{g_n}$ . By the triangle inequality, one obtains

$$\begin{aligned}\|f\bar{f} - g_n\overline{g_n}\|_{\mathcal{B}} &\leq \|\bar{f}\|_{\mathcal{B}} \|f - g_n\|_{\mathcal{B}} + \|g_n\|_{\mathcal{B}} \|\overline{f - g_n}\|_{\mathcal{B}} \\ &\leq m_1(1 + m_5) \|f\|_{\mathcal{B}} \|f - g_n\|_{\mathcal{B}},\end{aligned}$$

where the second line follows from Lemma 2.4 and Proposition 4.5. Since the right-hand side of the last inequality converges to zero as  $n$  tends to infinity, so does the left-hand side. Therefore, for every  $\epsilon > 0$  there exists an  $n_1 \geq n_0 > 0$  such that

$$(5.6) \quad \|g_n\overline{g_n}\|_{\mathcal{B}} \leq \|f\bar{f}\|_{\mathcal{B}} + \epsilon \quad \text{for all } n \geq n_1.$$

Now, we can put together all the previous steps. This gives

$$\begin{aligned}(\|f\|_{\mathcal{B}} - m_1 C_3 \epsilon)^2 &\stackrel{(5.4)}{\leq} m_1^2 C_3^2 \|\mathfrak{P}_+(s_n f)\|_{\mathcal{B}}^2 \\ &\stackrel{(5.5)}{\leq} \frac{m_1^2 C_3^2}{m_3} \|g_n\overline{g_n}\|_{\mathcal{B}} \stackrel{(5.6)}{\leq} \frac{m_1^2 C_3^2}{m_3} (\|f\bar{f}\|_{\mathcal{B}} + \epsilon)\end{aligned}$$

for all  $n \geq n_1$ . Since  $\epsilon$  was chosen arbitrarily, one obtains

$$\frac{m_3}{m_1^2 C_3^2} \|f\|_{\mathcal{B}}^2 \leq \|f\bar{f}\|_{\mathcal{B}}$$

for all  $f \in \mathcal{B}$ . This is the statement of the theorem with  $m_6 = m_3/(m_1 C_3)^2$ .  $\square$

Lemmas 4.3 and 5.1 show that if the spectral factorization is bounded on a decomposing Banach algebra  $\mathcal{B}$ , then (5.1) holds for all elements in  $\mathcal{B}$ . However, as already mentioned in [15], this implies that  $\mathcal{B}$  is isomorphic to  $\mathcal{C}(\mathbb{T})$ , which is no decomposing Banach algebra. For the sake of completeness, we give a separate proof of this reasoning in the following.

LEMMA 5.2. *Let  $\mathcal{B}$  be a decomposing Banach algebra and assume that there exists a constant  $0 < m_6 < \infty$  such that (5.1) holds for all  $f \in \mathcal{B}$ . Then it holds that*

$$(5.7) \quad \|f\|_{\infty} \leq \|f\|_{\mathcal{B}} \leq \frac{2}{m_6} \|f\|_{\infty} \quad \text{for all } f \in \mathcal{B}.$$

*Proof.* The lower bound in (5.7) is a consequence of Proposition 2.3. To prove the upper bound, first let  $h \in \mathcal{B}$  be a real valued function. Then because of assumption (5.1), it holds that  $m_6 \|h\|_{\mathcal{B}}^2 \leq \|h\bar{h}\|_{\mathcal{B}} = \|h^2\|_{\mathcal{B}}$ . Moreover, since  $h^2$  belongs again to  $\mathcal{B}$ , (5.1) can be applied to  $h^2$ , which gives  $m_6 \|h^2\|_{\mathcal{B}}^2 \leq \|h^4\|_{\mathcal{B}}$ . Together with the previous inequality, one has  $m_6 m_6^2 \|h\|_{\mathcal{B}}^4 \leq m_6 \|h^2\|_{\mathcal{B}}^2 \leq \|h^4\|_{\mathcal{B}}$ . Applying this upper bound repeatedly, one obtains

$$\|h\|_{\mathcal{B}} \leq C_5 \|h\|_{\infty}^{1/2^n} \quad \text{with} \quad C_5 = \left( \prod_{k=0}^{n-1} (m_6)^{2^k} \right)^{-1/2^n} = (m_6)^{-\frac{2^n-1}{2^n}}.$$

By the spectral radius formula, it holds  $\|h\|_{\infty} = \lim_{n \rightarrow \infty} \|h^n\|_{\mathcal{B}}^{1/n}$ . For  $n \rightarrow \infty$  this gives in our case that

$$(5.8) \quad \|h\|_{\mathcal{B}} \leq \frac{1}{m_6} \|h\|_{\infty}$$

for every real valued function  $h \in \mathcal{B}$ . Now let  $f = f_1 + i f_2$  be a complex function in  $\mathcal{B}$  with real functions  $f_1, f_2 \in \mathcal{B}$ . Then it follows from (5.8) and Proposition 2.3 that  $\|f\|_{\mathcal{B}} \leq \frac{1}{m_6} (\|f_1\|_{\infty} + \|f_2\|_{\infty}) \leq \frac{2}{m_6} \|f\|_{\infty}$ , which is the upper bound in (5.7).  $\square$

With these preparations, we prove the following main result of this paper.

**THEOREM 5.3.** *There exists no decomposing Banach algebra  $\mathcal{B}$  on which the spectral factorization mapping is bounded.*

*Proof.* The theorem is proved by contradiction. Assume that the spectral factorization is bounded on the decomposing Banach algebra  $\mathcal{B}$ ; then (5.7) holds. Now, let  $f \in \mathcal{C}(\mathbb{T})$  be an arbitrary continuous function, and let  $p_n$ ,  $n = 1, 2, \dots$ , be a sequence of polynomials which converges to  $f$  in  $\mathcal{C}(\mathbb{T})$ :  $\lim_{n \rightarrow \infty} \|f - p_n\|_\infty = 0$ . This shows, together with (5.7), that  $p_n$  is a Cauchy sequence in  $\mathcal{B}$ , and since  $\mathcal{B}$  is complete, there exists a  $g \in \mathcal{B}$  such that  $\lim_{n \rightarrow \infty} \|g - p_n\|_{\mathcal{B}} = 0$ . Because of (5.7) this implies also that  $\lim_{n \rightarrow \infty} \|g - p_n\|_\infty = 0$ . By the triangle inequality, one has  $\|f - g\|_\infty \leq \|f - p_n\|_\infty + \|p_n - g\|_\infty$ . Both terms on the right-hand side converge to zero as  $n$  tends to infinity. This shows that  $f = g$  and therefore  $f \in \mathcal{B}$ .

Finally, let again  $f \in \mathcal{C}(\mathbb{T})$ , and let  $\mathfrak{A}_n$  be the approximation operator defined in (4.7). Then by (5.7) and Proposition 4.5, it holds that

$$(5.9) \quad \|\mathfrak{A}_n f\|_\infty \leq \|\mathfrak{A}_n f\|_{\mathcal{B}} \leq m_4 \|f\|_{\mathcal{B}} \leq \frac{2m_4}{m_6} \|f\|_\infty,$$

where the constants  $m_4, m_6$  are independent of  $n$ . However, this inequality does not hold for all continuous functions  $f \in \mathcal{C}(\mathbb{T})$ . As an example, consider for an arbitrary but fixed  $n \in \mathbb{N}$  the function  $f_n(e^{i\omega}) = \left(\frac{1}{\pi} \sum_{k=1}^n \frac{\sin(k\omega)}{k}\right) e^{in\omega}$ . Clearly,  $f_n \in \mathcal{C}(\mathbb{T})$  with  $\|f_n\|_\infty \leq 1$ , and it holds that

$$(\mathfrak{A}_n f_n)(e^{i\omega}) = \frac{1}{2\pi i} \sum_{k=0}^{n-1} \frac{e^{ik\omega}}{n-k}.$$

It follows that  $|(\mathfrak{A}_n f_n)(1)| = \frac{1}{2\pi} \sum_{k=1}^n \frac{1}{k}$ , which shows that  $\|\mathfrak{A}_n f_n\|_\infty \geq \frac{1}{2\pi} \log n$ . Since the right-hand side of (5.9) is a constant, (5.9) gives a contradiction for the function  $f_n$  and for a sufficiently large  $n$ . This shows that (5.7) cannot hold. Consequently, the spectral factorization mapping has to be unbounded in  $\mathcal{B}$ .  $\square$

**6. Discussions.** In a sense, the most vital axiom of a decomposing Banach algebra is the second part of the axiom (A1), namely, that  $f \in \mathcal{B}$  always implies that  $\mathfrak{P}_+ f \in \mathcal{B}$ , which is equivalent to the boundedness of the Riesz projection  $\mathfrak{P}_+$  on  $\mathcal{B}$ . Let us consider *Banach algebras of continuous functions*. These are Banach algebras  $\mathcal{A}$  that are closed subspaces of  $\mathcal{C}(\mathbb{T})$  and that satisfy the axioms (A2)–(A4) and the first part of (A1) of a decomposing Banach algebra. Then the boundedness of the Riesz projection  $\mathfrak{P}_+$  on  $\mathcal{A}$  will determine whether  $\mathcal{A}$  is a decomposing algebra or not.

In the following, let  $\mathcal{A}$  be a Banach algebra of continuous functions. Then Theorem 5.3 implies that the spectral factorization mapping  $\mathfrak{S}$  is unbounded on  $\mathcal{A}$  if the Riesz projection is bounded on  $\mathcal{A}$ . Next, we give a different condition for the unboundedness of the spectral factorization. To this end, we consider Banach algebras with the following property.

**DEFINITION 6.1** (log-property). *Let  $\mathcal{A}$  be a Banach algebra of continuous functions on  $\mathbb{T}$ . We say that  $\mathcal{A}$  possesses the log-property if every  $f \in \mathcal{A}_+$  has the property that  $|f(z)| > 0$  for all  $z \in \mathbb{D}$  implies  $\log f \in \mathcal{A}_+$ .*

**Remark 6.2.** The Wiener algebra (Example 2.5) and the algebra  $\lambda_\alpha$  of Hölder continuous functions (Example 2.6) are examples of algebras having the log-property.

With this property, one gets the following alternative condition for the unboundedness of the spectral factorization on Banach algebras of continuous functions.

**COROLLARY 6.3.** *If a Banach algebra of continuous functions  $\mathcal{A}$  possesses the log-property, then the spectral factorization is unbounded on  $\mathcal{A}$ .*

*Proof.* Since  $\mathcal{A} \subset \mathcal{C}(\mathbb{T}) \subset L^\infty(\mathbb{T})$ , every  $f \in \mathcal{A}_{\text{pos}}$  possesses a spectral factorization with a spectral factor  $f_+ \in H^\infty(\mathbb{D})$  (cf. Remark 3.3). Moreover, since  $f$  is continuous and  $\mathbb{T}$  is closed, there exists a  $\zeta_0 \in \mathbb{T}$  such that  $\inf_{\zeta \in \mathbb{T}} |f(\zeta)| = f(\zeta_0) =: c > 0$ . In contradiction to the statement of the corollary, we assume that the spectral factorization is bounded on  $\mathcal{A}$ . Then  $f_+ \in \mathcal{A}_+$ , and (3.1) shows that  $|f_+(\zeta)| \geq \sqrt{c} > 0$  for all  $\zeta \in \mathbb{D}$ . Since  $\mathcal{A}$  possesses the log-property,  $\log f_+ \in \mathcal{A}$ . Because of (3.3) it holds that  $\log f_+(z) = (\mathfrak{P}_+[\log f])(z) - \frac{1}{2} \alpha_0$ , in which the constant  $\alpha_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log f(e^{i\tau}) d\tau$  is just the zeroth Fourier coefficient of  $\log f$ . Since  $\log f_+ \in \mathcal{A}$ , this shows that the Riesz projection of  $\log f$  belongs to  $\mathcal{A}$  as well. Thus, under the assumption that  $\mathcal{A}$  has the log-property and that the spectral factorization is bounded on  $\mathcal{A}$ , we have that

$$(6.1) \quad \mathfrak{P}_+[\log f] \in \mathcal{A} \quad \text{for all } f \in \mathcal{A}_{\text{pos}}.$$

Let  $g \in \mathcal{A}$  be an arbitrary real valued function in  $\mathcal{A}$ . Then  $f := \exp g \in \mathcal{A}_{\text{pos}}$ , and it follows from (6.1) that  $\mathfrak{P}_+g \in \mathcal{A}$  whenever  $g \in \mathcal{A}$ . For complex functions  $g = g_1 + i g_2 \in \mathcal{A}$  with real functions  $g_1, g_2 \in \mathcal{A}$  one obtains  $\mathfrak{P}_+g = \mathfrak{P}_+g_1 + i \mathfrak{P}_+g_2 \in \mathcal{A}$ . This implies that  $\mathcal{A}$  is a decomposing Banach algebra, and by Theorem 5.3 it follows that the spectral factorization mapping  $\mathfrak{S}$  is unbounded on  $\mathcal{A}$ . This conflicts with the assumption that  $\mathfrak{S}$  is bounded and therefore proves the corollary.  $\square$

By (3.3), the spectral factorization mapping  $\mathfrak{S}$  is a concatenation of the logarithm of the Riesz projection and of the exponential function:  $\mathfrak{S} = \exp \circ \mathfrak{P}_+ \circ \log$ . If one considers Banach algebras of continuous functions  $\mathcal{A}$ , Theorem 5.3 shows that  $\mathfrak{S}$  will be unbounded if the Riesz projector  $\mathfrak{P}_+$  is bounded, and Corollary 6.3 shows that  $\mathfrak{S}$  will be unbounded if  $\mathcal{A}$  has the log-property. It is interesting to compare this behavior with the Banach algebra  $L^\infty(\mathbb{T})$ . On this algebra, the spectral factorization is bounded, but the Riesz projection is unbounded and  $L^\infty(\mathbb{T})$  does not possess the log-property.

It was shown that the spectral factorization is unbounded on every decomposing Banach algebra  $\mathcal{B}$ . However, it is possible to characterize subsets  $\mathcal{M}$  of  $\mathcal{B}$  such that the spectral factor is bounded for every  $f \in \mathcal{M}$ . These subsets  $\mathcal{M}$  are characterized by the minimum value of the spectral densities in  $\mathcal{M}$ : For a fixed constant  $c > 0$  and a certain decomposing Banach algebra  $\mathcal{B}$ , we consider the set  $\mathcal{M}(c, \mathcal{B}) := \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq 1, f(\zeta) \geq c \text{ for all } \zeta \in \mathbb{T}\}$  and the number

$$C(c, \mathcal{B}) = \sup_{f \in \mathcal{M}(c, \mathcal{B})} \|\log f\|_{\mathcal{B}}.$$

Assume that for a certain constant  $c > 0$  it holds that  $C(c, \mathcal{B}) < \infty$ ; then the spectral factorization is bounded on  $\mathcal{M}(c, \mathcal{B})$ .

**PROPOSITION 6.4.** *Let  $\mathcal{B}$  be a decomposing Banach algebra, and let  $c > 0$  such that the constant  $C(c, \mathcal{B}) < \infty$ ; then for all  $f \in \mathcal{M}(c, \mathcal{B})$  it holds that*

$$(6.2) \quad \|f_+\|_{\mathcal{B}} = \|\mathfrak{S}f\|_{\mathcal{B}} \leq \exp \left( \frac{1}{2} C(c, \mathcal{B}) \|\mathfrak{P}_+\|_{\mathcal{B} \rightarrow \mathcal{B}} \right).$$

*Proof.* For every  $f \in \mathcal{M}(c, \mathcal{B})$  holds

$$\left\| \frac{1}{2} \mathfrak{P}_+(\log f) \right\|_{\mathcal{B}} \leq \frac{1}{2} \|\mathfrak{P}_+\|_{\mathcal{B} \rightarrow \mathcal{B}} \|\log f\|_{\mathcal{B}} \leq \frac{1}{2} \|\mathfrak{P}_+\|_{\mathcal{B} \rightarrow \mathcal{B}} C(c, \mathcal{B}).$$

By the definition of the spectral factorization mapping (3.2), one obtains therewith

$$\|f_+\|_{\mathcal{B}} \leq \sum_{k=0}^{\infty} \frac{1}{k!} \left( \frac{1}{2} \right)^k \|\mathfrak{P}_+(\log f)\|_{\mathcal{B}}^k \leq \exp \left( \frac{1}{2} \|\mathfrak{P}_+\|_{\mathcal{B} \rightarrow \mathcal{B}} C(c, \mathcal{B}) \right)$$

which is the statement of the proposition.  $\square$

For which constants  $c$  the value  $C(c, \mathcal{B})$  remains bounded depends strongly on the algebra  $\mathcal{B}$ . This is illustrated by two examples.

*Example 6.5.* For the Wiener algebra  $\mathcal{W}$ , it follows from Wiener's classical result [22] that  $C(c, \mathcal{W}) \leq \log(1/[2c - 1])$  for every  $c > 1/2$ , and it can be shown [4] that  $C(c, \mathcal{W}) = +\infty$  for all  $c \leq 1/4$ . It is not quite clear whether  $C(c, \mathcal{W})$  is bounded for  $1/4 < c \leq 1/2$  or not.

*Example 6.6.* For the decomposing Banach algebra  $\lambda_\alpha$  of all Hölder continuous function on  $\mathbb{T}$ , it is easy to verify [3] that for every arbitrary constants  $c > 0$ , the value  $C(c, \lambda_\alpha)$  is upper bounded by  $|\log c| < \infty$ .

**7. Conclusions.** This paper presented a supplement to a result of Jacob and Partington [15] on the boundedness of the spectral factorization mapping on decomposing Banach algebras. It was shown that there exists no decomposing Banach algebra on which the spectral factorization is bounded. Together with the continuity result in [15], one has therefore that the spectral factorization is continuous but unbounded on every decomposing Banach algebra.

The boundedness of the spectral factorization mapping would allow us to bound the norm of the spectral factors in terms of the norm of the given spectrum. This possibility is of some importance in certain applications. As a practical example we mention the approximation of the spectral factors by polynomials. To this end, consider the Banach algebra  $\lambda_\alpha$  of Hölder continuous functions (see Example 2.6), and let  $\phi \in \lambda_\alpha$  be an arbitrary spectrum which should be approximated by a polynomial  $\phi_N$  of degree  $N$ . It is well known (see, e.g., [25, section III.13]) that the approximation error is upper bounded by

$$\|\phi - \phi_N\|_\infty \leq \frac{C}{N^\alpha} \|\phi\|_\alpha$$

with a fixed constant  $C$ . Thus, for all spectral densities with, say,  $\|\phi\|_\alpha \leq 1$ , a sufficient approximation degree  $N_0$  can be found such that the approximation error remain always below a desired threshold. However, if one wants to approximate the corresponding spectral factor  $\phi_+$ , no such minimal approximation degree  $N_0$  for all  $\phi$  with  $\|\phi\|_\alpha \leq 1$  can be found since  $\|\phi_+ - \phi_N\|_\infty \leq \frac{C}{N^\alpha} \|\phi_+\|_\alpha$ , and because of the unboundedness of the spectral factorization mapping, the right-hand side cannot be upper bounded by the norm of  $\phi$ . Thus, in order to determine a sufficient approximation degree for the approximation polynomial of  $\phi_+$ , one needs to determine the spectral factor and its norm, in general. If, on the other hand, the spectral factorization mapping would be bounded, it would be possible to find for every  $\epsilon > 0$  a degree  $N_0$  such that there always exists a polynomial  $\phi_{N_0}$  such that  $\|\phi_+ - \phi_{N_0}\|_\infty < \epsilon$  for all spectra with  $\|\phi\|_\alpha < 1$ , without the need to control the norm of the spectral factors.

## REFERENCES

- [1] B. D. O. ANDERSON, *Continuity of the spectral factorization operation*, Math. Appl. Comput., 4 (1985), pp. 139–156.
- [2] S. BARCLAY, *Continuity of the spectral factorization mapping*, J. London Math. Soc. (2), 70 (2004), pp. 763–779.
- [3] H. BOCHE AND V. POHL, *Structural properties of the Wiener filter—stability, smoothness properties, and FIR approximation behavior*, IEEE Trans. Inform. Theory, 52 (2006), pp. 4272–4282.
- [4] H. BOCHE AND V. POHL, *The stability and continuity behavior of the spectral factorization in the Wiener algebra with applications in Wiener filtering*, IEEE Trans. Circuits and Systems, submitted.

- [5] CH. I. BYRNES, T. T. GEORGIU, AND A. LINDQUIST, *A new approach to spectral estimation: A tunable high-resolution spectral estimator*, IEEE Trans. Signal Process., 48 (2000), pp. 3189–3205.
- [6] J. M. CIOFFI, G. P. DUDEVOIR, M. V. EYUBOGLU, AND G. D. FORNEY, *MMSE decision-feedback equalizers and coding*, IEEE Trans. Commun., 43 (1995), pp. 2582–2594.
- [7] K. F. CLANCEY AND I. GOHBERG, *Factorization of Matrix Functions and Singular Integral Operators*, Birkhäuser Verlag, Basel, 1981.
- [8] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.
- [9] A. T. ERDOGAN, B. HASSIBI, AND T. KAILATH, *MIMO decision feedback equalization from an  $H^\infty$  perspective*, IEEE Trans. Signal Process., 52 (2004), pp. 734–745.
- [10] B. A. FRANCIS, *A Course in  $H_\infty$  Control Theory*, Springer-Verlag, New York, 1987.
- [11] L. FRANKS, *Signal Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
- [12] J. B. GARNETT, *Bounded Analytic Functions*, Academic Press, New York, 1981.
- [13] M. GREEN AND M. C. SMITH, *Continuity properties of LQG optimal controllers*, Systems Control Lett., 26 (1995), pp. 33–39.
- [14] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
- [15] B. JACOB AND J. R. PARTINGTON, *On the boundedness and continuity of the spectral factorization mapping*, SIAM J. Control Optim., 40 (2001), pp. 88–106.
- [16] B. JACOB, J. WINKIN, AND H. ZWART, *Continuity of the spectral factorization on a vertical strip*, Systems Control Lett., 37 (1999), pp. 183–192.
- [17] T. KAILATH, A. H. SAYED, AND B. HASSIBI, *Linear Estimation*, Prentice-Hall, Upper Saddle River, NJ, 2000.
- [18] V. V. PELLER AND S. V. KHRUSHCHEV, *Hankel operators, best approximations, and stationary Gaussian processes*, Russian Math. Surveys, 37 (1982), pp. 61–144.
- [19] W. RUDIN, *Functional Analysis*, 2nd ed., McGraw-Hill, Boston, 1991.
- [20] O. J. STAFFANS, *Quadratic optimal control of stable systems through spectral factorization*, Math. Control Signals Systems, 8 (1995), pp. 167–197.
- [21] S. TREIL, *A counterexample on continuous coprime factors*, IEEE Trans. Automat. Control, 39 (1994), pp. 1262–1263.
- [22] N. WIENER, *Tauberian theorems*, Ann. of Math. (2), 33 (1932), pp. 1–100.
- [23] N. WIENER, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series. With Engineering Applications*, The MIT Press, Cambridge, MA; John Wiley, New York, 1949.
- [24] N. WIENER AND P. MASANI, *The prediction theory of multivariate stochastic processes*, Acta Math., 98 (1957), pp. 111–150.
- [25] A. ZYGMUND, *Trigonometric Series*, 3rd ed., Cambridge University Press, Cambridge, UK, 2002.



# RELAXATION OF A CONTROL PROBLEM IN THE COEFFICIENTS WITH A FUNCTIONAL OF QUADRATIC GROWTH IN THE GRADIENT\*

JUAN CASADO-DÍAZ<sup>†</sup>, JULIO COUCE-CALVO<sup>†</sup>, AND JOSÉ D. MARTÍN-GÓMEZ<sup>†</sup>

**Abstract.** We study an optimal design problem consisting in mixing two anisotropic (electric or thermal) materials in order to minimize a functional depending on the gradient of the state. It is known that this type of problem has no solution in general, and then it is necessary to introduce a relaxed formulation. Here we prove that this relaxation is obtained by using composite materials, is constructed by homogenization, and takes a particular extension of the cost functional to these new materials. We obtain an integral representation of this relaxed cost functional. Besides, we show that our results contain some previous results obtained by other authors for isotropic materials.

**Key words.** control in the coefficients, elliptic PDE, optimal design

**AMS subject classification.** 49K20

**DOI.** 10.1137/070685890

**1. Introduction.** We consider a control problem for a linear elliptic partial differential equation where the control variable is the diffusion matrix (control problem in the coefficients). This type of problem appears in optimal design. Recall that the thermic or electric properties of a material are given by the corresponding diffusion matrix, and so choosing an optimal matrix diffusion is equivalent to choosing an optimal material. To simplify the exposition we consider a two-phase optimization problem; i.e., we assume that the materials are constructed by mixing two fixed materials (nonisotropic in general) represented by their diffusion matrices  $A$  and  $B$ , which we take to be symmetric and elliptic.

As a model problem we consider the following one: For a bounded open set  $\Omega \subset \mathbf{R}^N$ , we look for a measurable set  $\omega \subset \Omega$  such that for a given source term  $f \in L^2(\Omega)$  (or more generally in  $H^{-1}(\Omega)$ ) the solution  $u$  of

$$\begin{cases} -\operatorname{div}(A\chi_\omega + B\chi_{\Omega \setminus \omega})\nabla u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

minimizes a given functional  $J$  on the Sobolev space  $H_0^1(\Omega)$ . We also assume that the measure of  $\omega$  is less than or equal to  $\kappa|\Omega|$ , with  $0 < \kappa < 1$ ; i.e., we dispose only of a limited quantity of material  $A$ .

It is well known (see, e.g., [20], [21]) that, in general, this problem has no solution (some existence results can be obtained in particular situations [12]), and so it is necessary to relax the problem. By denoting, for  $p \in [0, 1]$ ,  $\mathcal{K}(A, B, p)$  as the set of matrices constructed via homogenization, mixing  $A$  and  $B$  with respective proportions  $p$  and  $1-p$ , and assuming  $J$  sequentially continuous for the weak topology of  $H_0^1(\Omega)$ , it is well known (see, e.g., [1], [11], [16], [17], [22], [24], [26], [27], [29]) that the relaxation

---

\*Received by the editors March 21, 2007; accepted for publication (in revised form) December 3, 2007; published electronically May 7, 2008. This work was supported by projects MTM2005-04914 of the Ministerio de Educación y Ciencia of Spain and FQM309 of the Junta de Andalucía.

<http://www.siam.org/journals/sicon/47-3/68589.html>

<sup>†</sup>Dpto. de Ecuaciones Diferenciales y Análisis Numérico, Facultad de Matemáticas, C. Tarfia s/n, 41012 Sevilla, Spain (jcasadod@us.es, couce@us.es, jdmartin@us.es).

of the model control problem is obtained by replacing the set of controls

$$(1.1) \quad \{A\chi_\omega + B\chi_{\Omega \setminus \omega} : \omega \subset \Omega \text{ measurable}, |\omega| \leq \kappa|\Omega|\}$$

by the larger set

$$(1.2) \quad \left\{ (M, \theta) \text{ measurable} : \theta \in [0, 1], M \in \mathcal{K}(A, B, \theta) \text{ a.e. in } \Omega, \int_\Omega \theta dx \leq \kappa|\Omega| \right\}.$$

In the present paper we are interested in functionals depending on the gradient of the state function, and so they are not sequentially continuous in general on the weak topology of  $H_0^1(\Omega)$ . For simplicity, we restrict ourselves to the functional

$$(1.3) \quad J(u) = \int_\Omega F(\nabla u) dx + G(u), \quad u \in H_0^1(\Omega),$$

where  $F$  is a Hölder-continuous function in  $\mathbf{R}^N$  with a quadratic growth and  $G$  is sequentially continuous for the weak topology of  $H_0^1(\Omega)$ . Thus, the control problem considered in this paper is given by

$$(1.4) \quad \begin{cases} \inf \left\{ \int_\Omega F(\nabla u) dx + G(u) \right\}, \\ -\operatorname{div} (A\chi_\omega + B\chi_{\Omega \setminus \omega}) \nabla u = f \text{ in } \Omega, \\ u \in H_0^1(\Omega), \quad \omega \subset \Omega \text{ measurable}, |\omega| \leq \kappa|\Omega|. \end{cases}$$

Some control problems related to (1.4) have also been considered by other authors. In this way, for the case

$$(1.5) \quad A = \alpha I, \quad B = \beta I, \quad J(u) = \int_\Omega |\nabla(u - v)|^2 dx,$$

it is proved in [30] that there exists a dense subset of  $v \in H^1(\Omega)$  such that, by taking as control variable the functions of  $L^\infty(\Omega)$  valued in  $[0, 1]$ , instead of the characteristic functions of measurable subsets of  $\Omega$ , problem (1.4) has a unique solution (the fact that  $A = \alpha I$ ,  $B = \beta I$  is not relevant in the reasoning used in [30]).

Related to this result, we mention that, for  $N = 2$ ,  $A = \alpha I$ ,  $B = \beta I$ , and  $F(\xi) = G(|\xi|)$ , convex in  $\xi$ , and growth not necessarily quadratic, it is proved in [25] that a relaxation of (1.4) can be obtained by just replacing the characteristic functions by functions valued in  $[0, 1]$ .

On the other hand, a relaxation of problem (1.4) when  $A, B$ , and  $J$  are given by (1.5) is obtained in [3], [10], [18]. We also refer to [15], [31], where, from a partial relaxation, it is realized a numerical study of the problem.

Some relaxation and numerical results for (1.4) have also been obtained in [2] when  $A$  and  $B$  are not necessarily scalar matrices but  $B - A$  is small.

Other relaxation problems for anisotropic materials (in diffusion and elasticity) have been considered in [12], [13], [14], where the functional is sequentially continuous with respect to the weak topology of  $H_0^1(\Omega)$ , but constraints appear on the gradients of the state functions.

In the present paper, for general  $A, B$ , and  $F$ , we show that the relaxation of (1.4) is given by replacing (as when  $J$  was sequentially continuous in the weak topology of  $H_0^1(\Omega)$ ) the set of controls (1.1) by (1.2) and the functional  $J$  by

$$\int_\Omega H(\nabla u, M \nabla u, \theta) dx + G(u),$$

where  $H : \{(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1] : \eta \in \mathcal{K}(A, B, p)\xi\} \rightarrow \mathbf{R}$  is defined by (4.6).

We prove that the function  $H$  is continuous, satisfies the growth condition

$$|H(\xi, \eta, p)| \leq C(1 + |\xi|^2),$$

and has some convexity properties given in Proposition 4.6.

We also obtain a characterization of the set  $\mathcal{K}(A, B, p)\xi$  for every  $\xi \in \mathbf{R}^N$ , and we explicitly calculate  $H$  on

$$\{(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1] : \eta \in \partial\mathcal{K}(A, B, p)\xi\}$$

(here  $\partial$  denotes the boundary with respect to the affine hull). This can be useful to the study of optimality conditions for the relaxed problem. We refer to [1], [5], [6], [16], [24] for the study of optimality conditions for control problems in the coefficients.

Finally, we obtain an explicit expression of  $H$  in its whole domain for the case where

$$F(\xi) = sA\xi \cdot \xi,$$

where  $s \in \mathbf{R}$  and  $s(A - B)$  is nonnegative. This example contains in particular (see Corollary 6.2) the case where  $A$ ,  $B$ , and  $J$  are given by (1.5).

Although, to simplify the exposition, we have assumed that  $J$  is given by (1.3), our techniques apply for a more general functional given by

$$\int_{\omega} F_1(x, u, \nabla u) dx + \int_{\Omega \setminus \omega} F_2(x, u, \nabla u) dx + G(u),$$

with  $G$  as above and  $F_1, F_2$  satisfying similar conditions to  $F$ . It is also possible to consider more than two materials (but then, we do not have an explicit characterization of the domain of  $H$ ) and the more realistic case where the set of materials is invariable by rotations.

**2. Notation.** The space of linear applications from  $\mathbf{R}^N$  into  $\mathbf{R}^N$ , which we assume to be identified with the space of matrices of dimension  $N \times N$ , is denoted by  $\mathcal{M}_N$ . The subspace of  $\mathcal{M}_N$  corresponding to the symmetric matrices is denoted by  $\mathcal{M}_N^s$ .

The kernel and the range of  $M \in \mathcal{M}_N$  are, respectively, denoted by  $\text{Ker}(M)$  and  $\text{Ran}(M)$ .

For  $M \in \mathcal{M}_N^s$ , not necessarily invertible, we define  $M^\dagger : \text{Ran}(M) \rightarrow \text{Ran}(M)$  the pseudoinverse of  $M$ , i.e., the inverse of the restriction of  $M$  to its range.

The unitary cube in  $\mathbf{R}^N$ ,  $(0, 1)^N$  is denoted by  $Y$ .

For two sets  $Z$  and  $Z'$ , we denote by  $Z \triangle Z'$  its symmetric difference; i.e.,  $Z \triangle Z' = (Z \setminus Z') \cup (Z' \setminus Z)$ .

We use the subindex  $\sharp$  to mean  $Y$ -periodicity. For example, the space of functions in the Sobolev space  $H_{loc}^1(\mathbf{R}^N)$  which are  $Y$ -periodic is denoted by  $H_\sharp^1(Y)$ . Indeed, in the present paper all of the functions defined on  $Y$  are assumed to be extended to  $\mathbf{R}^N$  by  $Y$ -periodicity.

Throughout the paper we denote by  $A$  and  $B$  two fixed positive symmetric matrices.

We define  $\Xi : \text{Ran}(A - B) \rightarrow A^{-1}\text{Ran}(A - B) (= B^{-1}\text{Ran}(A - B))$  by

$$(2.1) \quad \Xi\zeta = \nu \Leftrightarrow \begin{cases} (B - A)\nu = \zeta \\ A\nu \in \text{Ran}(A - B) \end{cases} \Leftrightarrow \begin{cases} (B - A)\nu = \zeta \\ B\nu \in \text{Ran}(A - B) \end{cases}$$

for every  $\zeta \in \text{Ran}(A - B)$  or, equivalently,

$$(2.2) \quad \Xi = (rA + sB)^{-1}(B - A) \left( (B - A)(rA + sB)^{-1}(B - A) \right)^\dagger$$

for every  $r, s \geq 0$ ,  $r + s > 0$ . Note that if  $B - A$  is invertible, then  $\Xi = (B - A)^{-1}$ .

For  $p \in [0, 1]$ , we define  $\Lambda_p$  as the arithmetic mean of  $A$  and  $B$  with respective proportions  $p$  and  $1 - p$ ; i.e.,

$$\Lambda_p = pA + (1 - p)B.$$

We denote by  $\Omega \subset \mathbf{R}^N$  a fixed bounded open set, smooth enough for Meyer's theorem [19] to be satisfied and such that there exist  $\hat{\Omega}$  open bounded with  $\bar{\Omega} \subset \hat{\Omega}$  and a linear continuous prolongation operator from  $H^1(\Omega)$  into  $H_0^1(\hat{\Omega})$ .

For a sequence  $M_n \in L^\infty(\Omega; \mathcal{M}_N)$ , uniformly elliptic and bounded, and  $M \in L^\infty(\Omega; \mathcal{M}_N)$ , we write  $M_n \xrightarrow{H} M$  to mean that  $M_n$  converges to  $M$  in the sense of the  $H$ -convergence [22], [27]. Indeed, as we usually deal with symmetric matrices,  $H$ -convergence is equivalent to the  $G$ -convergence of Spagnolo [26].

For  $p \in [0, 1]$  we denote by  $\mathcal{K}(A, B, p)$  the set of materials constructed via homogenization mixing the materials corresponding to the diffusion matrices  $A$  and  $B$ , with respective proportions  $p$  and  $1 - p$ ; i.e.,

$$(2.3) \quad \mathcal{K}(A, B, p) = \left\{ M \in \mathcal{M}_N^s : \exists \omega_n \subset \mathbf{R}^N \text{ measurable, } \chi_{\omega_n} \xrightarrow{*} p \text{ in } L^\infty(\mathbf{R}^N), A\chi_{\omega_n} + B(1 - \chi_{\omega_n}) \xrightarrow{H} M \right\}.$$

Clearly,  $\mathcal{K}(A, B, 1) = \{A\}$ ,  $\mathcal{K}(A, B, 0) = \{B\}$ .

For  $\xi \in \mathbf{R}^N$ ,  $p \in [0, 1]$  we write

$$\mathcal{K}(A, B, p)\xi = \{ \eta \in \mathbf{R}^N : \exists M \in \mathcal{K}(A, B, p), \text{ with } \eta = M\xi \}.$$

**3. Preliminary results.** To our knowledge, an explicit characterization of the set  $\mathcal{K}(A, B, p)$  is known only for isotropic materials (see, e.g., [17], [29]). Fortunately, for the purpose of the present paper, we need only to know the set  $\mathcal{K}(A, B, p)\xi$  for every  $\xi \in \mathbf{R}^N$ . A characterization of this set is obtained in the present section.

We recall the following result due to Dal Maso and Kohn [9] (see also [1], [11]), which shows that the set  $\mathcal{K}(A, B, p)$  can be obtained via periodic homogenization.

**THEOREM 3.1.** *For  $p \in [0, 1]$ , the set of matrices  $M$  for which there exists  $Z \subset Y$ , with  $|Z| = p$ , such that for every  $\xi \in \mathbf{R}^N$ ,*

$$(3.1) \quad M\xi = \int_Y (A\chi_Z + B\chi_{Y \setminus Z})(\xi + \nabla w) dy,$$

with  $w$  the solution of

$$(3.2) \quad \begin{cases} w \in H_\#^1(Y), & \int_Y w dy = 0, \\ -\text{div}((A\chi_Z + B\chi_{Y \setminus Z})(\xi + \nabla w)) = 0 & \text{in } \mathbf{R}^N, \end{cases}$$

is dense in  $\mathcal{K}(A, B, p)$ .

The following result gives some properties of the solution  $w$  of (3.2).

LEMMA 3.2. For  $p \in (0, 1)$ , we consider  $\xi, \eta \in \mathbf{R}^N$  and  $Z \subset Y$  measurable, with  $|Z| = p$ , such that the solution  $w$  of (3.2) satisfies

$$(3.3) \quad \int_Y (A\chi_Z + B\chi_{Y \setminus Z})(\xi + \nabla w) dy = \eta.$$

Then the following equalities hold:

$$(3.4) \quad \int_Z A \nabla w \cdot \nabla w dy + \int_{Y \setminus Z} B \nabla w \cdot \nabla w dy = (\Lambda_p \xi - \eta) \cdot \xi,$$

$$(3.5) \quad (B - A) \int_Z \nabla w dy = -(B - A) \int_{Y \setminus Z} \nabla w dy = \Lambda_p \xi - \eta,$$

$$(3.6) \quad \begin{cases} \int_Z A \left( \nabla w - \frac{1}{p} \int_Z \nabla w dz \right) \cdot \left( \nabla w - \frac{1}{p} \int_Z \nabla w dz \right) dy \\ + \int_{Y \setminus Z} B \left( \nabla w - \frac{1}{1-p} \int_{Y \setminus Z} \nabla w dz \right) \cdot \left( \nabla w - \frac{1}{1-p} \int_{Y \setminus Z} \nabla w dz \right) dy \\ = (\Lambda_p \xi - \eta) \cdot \xi - \left( \frac{A}{p} + \frac{B}{1-p} \right) \left( \int_Z \nabla w dy \right) \cdot \left( \int_Z \nabla w dy \right). \end{cases}$$

*Proof.* By using  $w$  as a test function in (3.2), and taking into account (3.3), we easily get (3.4).

Since  $w$  is periodic, we have

$$\int_Z \nabla w dy + \int_{Y \setminus Z} \nabla w dy = 0.$$

On the other hand, by (3.3) we obtain

$$A \int_Z \nabla w dy + B \int_{Y \setminus Z} \nabla w dy = \eta - \Lambda_p \xi.$$

From these equalities we conclude (3.5).

To prove (3.6), it is enough to develop the left-hand side and then use (3.4).  $\square$

As a consequence of Lemma 3.2, we have the following.

PROPOSITION 3.3. If  $N = 1$ ,  $p \in [0, 1]$ , we have

$$(3.7) \quad \mathcal{K}(A, B, p) = \left\{ \left( \frac{p}{A} + \frac{1-p}{B} \right)^{-1} \right\}.$$

If  $N \geq 2$ ,  $p \in (0, 1)$ ,  $\xi \in \mathbf{R}^N$ , then, by denoting by  $E(\xi, p)$  the ellipsoid

$$E(\xi, p) = \left\{ \nu \in \mathbf{R}^N : \left( \frac{A}{p} + \frac{B}{1-p} \right) \nu \cdot \nu \leq (B - A) \nu \cdot \xi \right\},$$

we have

$$(3.8) \quad \mathcal{K}(A, B, p)\xi = \Lambda_p \xi + (A - B)E(\xi, p).$$

*Proof.* The case  $N = 1$  is well known. Indeed, it can be easily obtained by using the fact that for every  $Z \subset Y$ , with  $|Z| = p$ , the solution  $w$  of (3.2) satisfies

$$(3.9) \quad \xi + \frac{dw}{dx} = \begin{cases} \frac{\eta}{A} & \text{a.e. in } Z, \\ \frac{\eta}{B} & \text{a.e. in } Y \setminus Z, \end{cases}$$

with

$$(3.10) \quad \eta = \left( \frac{p}{A} + \frac{1-p}{B} \right)^{-1} \xi.$$

Assume now that  $N \geq 2$ ,  $p \in (0, 1)$ . For  $Z \subset Y$ , with  $|Z| = p$ , and the  $w \in H^1_{\sharp}(Y)$  solution of (3.2), the left-hand side of (3.6) is nonnegative, and so equality (3.5) proves that the vector

$$\nu = \int_Z \nabla w \, dy$$

is in  $E(\xi, p)$  and satisfies  $\eta = \Lambda_p \xi + (A - B)\nu$ . Since by Theorem 3.1 the set of  $\eta$  constructed in this way is dense in  $\mathcal{K}(A, B, p)\xi$ , we then deduce the inclusion

$$\mathcal{K}(A, B, p)\xi \subset \Lambda_p \xi + (A - B)E(\xi, p).$$

Reciprocally, let us now prove that every  $\nu \in E(\xi, p)$  satisfies that  $\Lambda_p \xi + (A - B)\nu$  belongs to  $\mathcal{K}(A, B, p)\xi$ .

If  $\nu$  belongs to  $\partial E(\xi, p) \setminus \{0\}$ , this can be easily shown by using a lamination of  $A$  and  $B$  with respective proportions  $p$  and  $1 - p$  in the direction of  $\nu$ . If  $\nu = 0$ , we consider a lamination as above but now in an orthogonal direction to  $(B - A)\xi$ .

If  $\nu$  belongs to the interior of  $E(\xi, p)$ , then, for  $\lambda > 1$  such that  $\lambda\nu$  belongs to  $\partial E(\xi, p)$ , we take two matrices  $M_1, M_2 \in \mathcal{K}(A, B, p)$  such that  $M_1\xi = \Lambda_p \xi$ ,  $M_2\xi = \Lambda_p \xi + \lambda(A - B)\nu$ . A lamination of  $M_1$  and  $M_2$  with respective proportions  $1 - 1/\lambda$  and  $1/\lambda$  in an orthogonal direction to  $(M_2 - M_1)\xi$  provides a matrix  $M \in \mathcal{K}(A, B, p)$  such that  $\Lambda_p \xi + (A - B)\nu = M\xi$ .  $\square$

**COROLLARY 3.4.** For  $N \geq 2$ ,  $p \in (0, 1)$ , and  $\xi \in \mathbf{R}^N$ , we have

$$(3.11) \quad \mathcal{K}(A, B, p)\xi = \left\{ \eta \in \mathbf{R}^N : \begin{aligned} &\Lambda_p \xi - \eta \in \text{Ran}(A - B), \\ &\left( (A - B) \left( \frac{A}{p} + \frac{B}{1-p} \right)^{-1} (A - B) \right)^{\dagger} (\Lambda_p \xi - \eta) \cdot (\Lambda_p \xi - \eta) \leq \xi \cdot (\Lambda_p \xi - \eta) \end{aligned} \right\}.$$

*Proof.* By Proposition 3.3, we have that  $\eta$  belongs to  $\mathcal{K}(A, B, p)\xi$  if and only if there exists  $\nu \in \mathbf{R}^N$ , such that  $\Lambda_p \xi - \eta = (B - A)\nu$  and

$$\left( \frac{A}{p} + \frac{B}{1-p} \right) \nu \cdot \nu \leq (\Lambda_p \xi - \eta) \cdot \xi$$

or, equivalently, if and only if  $\Lambda_p \xi - \eta \in \text{Ran}(A - B)$  and

$$(3.12) \quad \min \left\{ \left( \frac{A}{p} + \frac{B}{1-p} \right) \nu \cdot \nu : \Lambda_p \xi - \eta = (B - A)\nu \right\} \leq (\Lambda_p \xi - \eta) \cdot \xi.$$

The minimum of this problem is attained in  $\nu = \Xi(\Lambda_p \xi - \eta)$ , with  $\Xi$  defined by (2.1). By using the fact that by (2.2)

$$\Xi = \left( \frac{A}{p} + \frac{B}{1-p} \right)^{-1} (B - A) \left( (B - A) \left( \frac{A}{p} + \frac{B}{1-p} \right)^{-1} (B - A) \right)^{\dagger},$$

we obtain (3.11).  $\square$

As a consequence of this result, we have the following.

**COROLLARY 3.5.** *For  $N \geq 2$ ,  $p \in (0, 1)$ ,  $\xi \in \mathbf{R}^N$ , the set  $\mathcal{K}(A, B, p)\xi$  reduces to  $\{\Lambda_p \xi\}$  if and only if  $\xi \in \text{Ker}(A - B)$ .*

*Proof.* By Corollary 3.4, we have  $\mathcal{K}(A, B, p)\xi = \{\Lambda_p \xi\}$ , if and only if for every  $\zeta \in \text{Ran}(A - B)$ ,  $\zeta \neq 0$ , one has

$$\left( (A - B) \left( \frac{A}{p} + \frac{B}{1-p} \right)^{-1} (A - B) \right)^{\dagger} \zeta \cdot \zeta > \xi \cdot \zeta,$$

but this is equivalent to  $\xi \in \text{Ran}(A - B)^{\perp} = \text{Ker}(A - B)$ .  $\square$

**4. Formulation of the problem and main results.** Let us consider a function  $F : \mathbf{R}^N \rightarrow \mathbf{R}$  such that there exist  $L > 0$ ,  $\varrho \in (0, 1]$  which satisfy

$$(4.1) \quad |F(\xi) - F(\xi')| \leq L(1 + |\xi| + |\xi'|)^{2-\varrho} |\xi - \xi'|^{\varrho} \quad \forall \xi, \xi' \in \mathbf{R}^N.$$

Without loss of generality, we can also assume that

$$(4.2) \quad F(0) = 0.$$

These properties imply that  $F$  satisfies

$$(4.3) \quad |F(\xi)| \leq L(1 + |\xi|)^2 \quad \forall \xi \in \mathbf{R}^N.$$

For the open set  $\Omega$  and the matrices  $A, B$  given in section 2, our aim here is to obtain a relaxation of the problem

$$(4.4) \quad \begin{cases} \inf \left\{ \int_{\Omega} F(\nabla u) dx + G(u) \right\}, \\ -\text{div}(A\chi_{\omega} + B\chi_{\Omega \setminus \omega})\nabla u = f \text{ in } \Omega, \\ u \in H_0^1(\Omega), \quad \omega \subset \Omega \text{ measurable, } |\omega| \leq \kappa|\Omega|, \end{cases}$$

where  $G$  is a sequentially continuous functional in the weak topology of  $H_0^1(\Omega)$ ,  $f \in H^{-1}(\Omega)$ , and  $\kappa \in (0, 1)$ . For this purpose, given  $\delta > 0$ , we define  $H_{\delta} : \mathbf{R}^N \times \mathbf{R}^N \times [0, 1] \rightarrow \mathbf{R} \cup \{+\infty\}$  by

$$(4.5) \quad \begin{cases} H_{\delta}(\xi, \eta, p) = \inf \int_Y F(\xi + \nabla w) dy, \\ -\text{div}(A\chi_Z + B\chi_{Y \setminus Z})(\xi + \nabla w) = 0 \text{ in } \mathbf{R}^N, \quad w \in H_{\sharp}^1(Y), \\ \left| \int_Y (A\chi_Z + B\chi_{Y \setminus Z})(\xi + \nabla w) dy - \eta \right| < \delta, \\ Z \subset Y \text{ measurable, } |Z| = p, \end{cases}$$

for every  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1]$ . In the above expression, the infimum over the empty set is defined as  $+\infty$ .

By using the fact that  $H_\delta$  is decreasing with respect to  $\delta$ , we define  $H : \mathbf{R}^N \times \mathbf{R}^N \times [0, 1] \rightarrow \mathbf{R} \cup \{+\infty\}$  by

$$(4.6) \quad H(\xi, \eta, p) = \lim_{\delta \rightarrow 0} H_\delta(\xi, \eta, p) \quad \forall (\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1].$$

*Remark 4.1.* Definition (4.6) of  $H$  implies that for every  $(\xi, \eta, p) \in D(H)$  there exists a sequence of measurable sets  $Z_n \subset Y$ , with  $|Z_n| = p$ , such that by defining  $S_n \in L^\infty(\Omega; \mathcal{M}_n^s)$  by  $S_n = A\chi_{Z_n} + B\chi_{Y \setminus Z_n}$ , and taking the  $w_n \in H_\#^1(Y)$  solution of

$$-\operatorname{div} S_n(\xi + \nabla w_n) = 0 \quad \text{in } \mathbf{R}^N$$

and

$$\eta_n = \int_Y S_n(\xi + \nabla w_n) dy,$$

we have

$$(4.7) \quad \eta = \lim_{n \rightarrow \infty} \eta_n,$$

$$(4.8) \quad H(\xi, \eta, p) = \lim_{n \rightarrow \infty} \int_Y F(\xi + \nabla w_n) dy.$$

For  $N = 1$ , the following proposition gives an explicit expression of  $H$ .

**PROPOSITION 4.1.** *If  $N = 1$ , the function  $H$  is given by*

$$(4.9) \quad H(\xi, \eta, p) = \begin{cases} pF\left(\frac{\eta}{A}\right) + (1-p)F\left(\frac{\eta}{B}\right) & \text{if } \eta = \left(\frac{p}{A} + \frac{1-p}{B}\right)^{-1} \xi, \\ +\infty & \text{in another case.} \end{cases}$$

For  $N \geq 2$  we do not have an explicit expression for  $H$ , but we can show the following result.

**THEOREM 4.2.** *If  $N \geq 2$ , the function  $H$  satisfies the following properties.*

*The domain of  $H$  is given by*

$$(4.10) \quad \begin{aligned} \operatorname{Dom}(H) &\stackrel{\text{def}}{=} \{(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1] : H(\xi, \eta, p) < +\infty\} \\ &= \{(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1] : \eta \in \mathcal{K}(A, B, p)\xi\}. \end{aligned}$$

*The function  $H$  is lower semicontinuous in  $\operatorname{Dom}(H)$ , and, for*

$$\alpha = \min\{\text{eigenvalues of } A \text{ and } B\}, \quad \beta = \max\{\text{eigenvalues of } A \text{ and } B\},$$

*it satisfies*

$$(4.11) \quad |H(\xi, \eta, p)| \leq L \left(\frac{\beta}{\alpha} |\xi|\right)^q \left(1 + \frac{\beta}{\alpha} |\xi|\right)^{2-q} \quad \forall (\xi, \eta, p) \in \operatorname{Dom}(H).$$

*Moreover, we have*

$$(4.12) \quad H(\xi, A\xi, 1) = H(\xi, B\xi, 0) = F(\xi) \quad \forall \xi \in \mathbf{R}^N,$$



$$(4.13) \quad \begin{cases} H(\xi, \eta, p) = pF\left(\xi + \Xi\left(\frac{\Lambda_p \xi - \eta}{p}\right)\right) + (1-p)F\left(\xi - \Xi\left(\frac{\Lambda_p \xi - \eta}{1-p}\right)\right) \\ \forall (\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times (0, 1) \text{ with } \Lambda_p \xi - \eta \in \text{Ran}(A - B) \text{ and} \\ \left(\left((A - B)\left(\frac{A}{p} + \frac{B}{1-p}\right)^{-1}(A - B)\right)^\dagger (\Lambda_p \xi - \eta) \cdot (\Lambda_p \xi - \eta) = \xi \cdot (\Lambda_p \xi - \eta)\right). \end{cases}$$

If  $F$  is convex, then

$$(4.14) \quad \begin{aligned} F(\xi) &\leq \min \left\{ pF\left(\xi + \frac{\nu}{p}\right) + (1-p)F\left(\xi - \frac{\nu}{1-p}\right) : (B - A)\nu = \Lambda_p \xi - \eta, \nu \in E(\xi, p) \right\} \\ &\leq H(\xi, \eta, p) \quad \forall (\xi, \eta, p) \in \text{Dom}(H), p \in (0, 1). \end{aligned}$$

*Remark 4.2.* A consequence of Theorem 4.2 is that if  $N \geq 2$ , then  $H(\xi, \Lambda_p \xi, p) = F(\xi)$  for every  $\xi \in \mathbf{R}^N$  and every  $p \in [0, 1]$ .

*Remark 4.3.* If  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times (0, 1)$  is such that  $\Lambda_p \xi - \eta \in \text{Ran}(A - B)$  and

$$(4.15) \quad \left(\left((A - B)\left(\frac{A}{p} + \frac{B}{1-p}\right)^{-1}(A - B)\right)^\dagger (\Lambda_p \xi - \eta) \cdot (\Lambda_p \xi - \eta) = \xi \cdot (\Lambda_p \xi - \eta)\right),$$

then the set of  $\nu \in E(\xi, p)$  such that  $(B - A)\nu = \Lambda_p \xi - \eta$  reduces to  $\nu = \Xi(\Lambda_p \xi - \eta)$ . Therefore, (4.13) shows that the second inequality in (4.14) is an equality for such  $(\xi, \eta, p)$ .

By using the function  $H$ , we obtain the following theorem.

**THEOREM 4.3.** *For every  $G : H_0^1(\Omega) \rightarrow \mathbf{R}$  sequentially continuous in the weak topology of  $H_0^1(\Omega)$ , every  $\kappa \in (0, 1)$ , and every  $f \in H^{-1}(\Omega)$ , a relaxation of problem (4.4) is given by*

$$(4.16) \quad \begin{cases} \min \left\{ \int_{\Omega} H(\nabla u, M\nabla u, \theta) dx + G(u) \right\}, \\ -\text{div } M\nabla u = f \text{ in } \Omega, \\ u \in H_0^1(\Omega), \\ \theta \in L^\infty(\Omega), 0 \leq \theta \leq 1 \text{ a.e. in } \Omega, \int_{\Omega} \theta dx \leq \kappa|\Omega|, \\ M \text{ measurable, } M(x) \in \mathcal{K}(A, B, \theta(x)) \text{ for a.e. } x \in \Omega, \end{cases}$$

with  $H$  given by (4.6).

*Remark 4.4.* Problem (4.16) can also be written as

$$(4.17) \quad \begin{cases} \min \left\{ \int_{\Omega} H(\nabla u, \sigma, \theta) dx + G(u) \right\}, \\ -\text{div } \sigma = f \text{ in } \Omega, \\ u \in H_0^1(\Omega), \\ \theta \in L^\infty(\Omega), 0 \leq \theta \leq 1 \text{ a.e. in } \Omega, \int_{\Omega} \theta dx \leq \kappa|\Omega|, \\ \sigma \in \mathcal{K}(A, B, \theta)\nabla u \text{ a.e. in } \Omega \text{ measurable.} \end{cases}$$

Theorem 4.3 is a consequence of Theorem 4.5 below, which is interesting by itself. We need the following definition.

DEFINITION 4.4. *We say that  $(u_n, \sigma_n, \theta_n) \in H^1(\Omega) \times L^2(\Omega)^N \times L^\infty(\Omega)$   $\mathcal{T}$ -converges to  $(u, \sigma, \theta) \in H^1(\Omega) \times L^2(\Omega)^N \times L^\infty(\Omega)$  if and only if*

$$\begin{aligned} u_n &\rightharpoonup u \text{ in } H^1(\Omega), & |\nabla u_n|^2 &\text{ equi-integrable,} \\ \sigma_n &\rightharpoonup \sigma \text{ in } L^2(\Omega)^N, & \operatorname{div} \sigma_n &\rightarrow \operatorname{div} \sigma \text{ in } H^{-1}(\Omega)^N, \\ \theta_n &\overset{*}{\rightharpoonup} \theta \text{ in } L^\infty(\Omega). \end{aligned}$$

Remark 4.5. In the applications, we are interested in sequences  $(u_n, \sigma_n, \theta_n)$  such that there exists a sequence of uniformly elliptic and bounded matrix functions  $M_n$ , which satisfies  $M_n \nabla u_n = \sigma_n$ . Then we recall that, thanks to Meyer's regularity theorem [19], the weak convergence of  $u_n$  in  $H^1(\Omega)$  and the strong convergence of  $\sigma_n$  in  $L^2(\Omega)^N$  imply the equi-integrability of  $|\nabla u_n|^2$  at least for  $\Omega$  smooth and  $u_n$  satisfying "good" boundary conditions (if not, we always hold the equi-integrability on compact subsets of  $\Omega$ ).

THEOREM 4.5. *The lower semicontinuous envelope with respect to the  $\mathcal{T}$ -convergence of the functional  $\mathcal{F} : H^1(\Omega) \times L^2(\Omega)^N \times L^\infty(\Omega) \rightarrow \mathbf{R} \cup \{+\infty\}$  defined by*

$$(4.18) \quad \mathcal{F}(u, \sigma, \theta) = \begin{cases} \int_{\Omega} F(\nabla u) dx & \text{if } \theta = \chi_{\omega}, \omega \subset \Omega \text{ measurable, } \sigma = \Lambda_{\theta} \nabla u, \\ +\infty & \text{in another case} \end{cases}$$

is given by

$$(4.19) \quad \overline{\mathcal{F}}(u, \sigma, \theta) = \begin{cases} \int_{\Omega} H(\nabla u, \sigma, \theta) dx & \text{if } 0 \leq \theta \leq 1, \sigma \in \mathcal{K}(A, B, \theta) \nabla u, \text{ a.e. in } \Omega, \\ +\infty & \text{in another case.} \end{cases}$$

Remark 4.6. Analogously to the proof of Theorem 4.3, we can use Theorem 4.5 to obtain the relaxation of some other related control problems. For example (assuming smoothness enough to have the equi-integrability of  $|\nabla u_n|^2$ , with  $u_n$  the state functions corresponding to a minimizing sequence), we can consider different boundary conditions for the state equation and some other restrictions. In this way, we can apply Theorem 4.5 to obtain a relaxation of the control problem defining  $H_{\delta}$  (see (4.5)). This permits us to prove that the function  $H$  given by (4.6) satisfies

$$(4.20) \quad \left\{ \begin{aligned} H(\xi, \eta, p) &= \inf \int_Y H(\xi + \nabla w, M(\xi + \nabla w), \theta) dy \\ &\quad \theta \in L^{\infty}_{\#}(Y), \quad 0 \leq \theta \leq 1 \text{ a.e. in } \mathbf{R}^N, \quad \int_Y \theta dy = p, \\ &\quad M \in \mathcal{K}(A, B, \theta) \text{ a.e. in } \mathbf{R}^N, \quad w \in H^1_{\#}(Y), \\ &\quad -\operatorname{div} M(\xi + \nabla w) = 0 \text{ in } \mathbf{R}^N, \\ &\quad \int_Y M(\xi + \nabla w) dy = \eta. \end{aligned} \right.$$

*Remark 4.7.* In Step 3 in the proof of Theorem 4.5, for  $(u, \sigma, \theta) \in \text{Dom}(\overline{\mathcal{F}})$  given, we show how to construct  $\omega_n \subset \Omega$  and  $u_n \in H^1(\Omega_n)$  such that the sequence  $(u_n, (A\chi_{\omega_n} + B\chi_{\omega_n})\nabla u_n, \chi_{\omega_n}) \in \text{Dom}(\mathcal{F})$  satisfies

$$\bar{\mathcal{F}}(u, \sigma, \theta) = \lim_{n \rightarrow \infty} \mathcal{F}(u_n, (A\chi_{\omega_n} + B\chi_{\omega_n})\nabla u_n, \chi_{\omega_n}).$$

By applying this procedure to a solution  $(u, \sigma, \theta)$  of problem (4.17), this gives a way to construct a minimizing sequence for problem (4.4). Unfortunately, to apply this procedure it is necessary, for  $(\xi, \eta, p) \in \text{Dom}(H)$ , to know how to construct  $Z_n \subset Y$  in the conditions of Remark 4.1. We do not know how to make this, in general. In particular, we do not know if this can be carried out by using laminations, as it happens in some particular cases where the function  $H$  can be explicitly calculated (see, e.g., [3], [10], and Remark 6.2 in the present paper). When  $\eta \in \partial\mathcal{K}(A, B, p)\xi$  (here  $\partial$  denotes the boundary with respect to the affine hull), the set  $Z_n$  is obtained by using a simple lamination in the direction of  $\nu = \Xi(\Lambda_p\xi - \eta)$ , if  $\eta \neq \Lambda_p\xi$ , or in an orthogonal direction to  $(B - A)\xi$ , if  $\eta = \Lambda_p\xi$ . In this sense, we remark that if  $H$  is derivable (which we do not know if it is true) and  $\theta, M, u$  is a solution of (4.16), then, by introducing the adjoint state  $q$  as the solution of

$$\begin{cases} -\text{div } M\nabla q = -\text{div}(\nabla_\xi H(\nabla u, M\nabla u, \theta) + M\nabla_\eta H(\nabla u, M\nabla u, \theta)) & \text{in } \Omega, \\ q \in H_0^1(\Omega), \end{cases}$$

the optimality conditions for problem (4.16) show (see, e.g., [1], [5], [16], [24] for related results) that a.e. on the set  $\{x \in \Omega : \nabla q(x) \neq \nabla_\eta H(\nabla u(x), M(x)\nabla u(x), \theta(x))\}$ , one has that  $M\nabla u \in \partial\mathcal{K}(A, B, \theta)\nabla u$ .

*Remark 4.8.* In order to solve numerically problem (4.16) the main difficulty is, as in the previous remark, that we have only an explicit expression of  $H$  on the points  $(\xi, \eta, p)$  such that  $\eta \in \partial\mathcal{K}(A, B, p)\xi$  (see (4.13)). But, as we observed above, if  $\theta, M, u$  is a solution of (4.16),  $H$  is sufficiently smooth, and  $\nabla q \neq \nabla_\eta H(\nabla u, M\nabla u, \theta)$  a.e. in  $\Omega$ , then  $M\nabla u \in \partial\mathcal{K}(A, B, \theta)\nabla u$  a.e. in  $\Omega$ . Moreover, in this case  $M$  is obtained by just one lamination. By taking into account these remarks, one can consider a numerical method consisting, for example, of taking a triangulation of  $\Omega$  and then searching the state function  $u$  piecewise affine, the proportion  $\theta$ , and the matrix  $M$  piecewise constants, with  $M$  corresponding to a lamination in each triangle (so the choice of  $M$  in each triangle is reduced to the choice of the corresponding lamination vector). This provides a numerical method similar to the one used in [15] and [31] for the case where  $F(\xi) = |\xi|^2$ ,  $A = \alpha I$ ,  $B = \beta I$ ,  $\alpha, \beta > 0$ .

By using the fact that by Theorem 4.5 the functional  $\bar{\mathcal{F}}$  is lower semicontinuous for the  $\mathcal{T}$ -convergence, we can deduce some convexity properties for  $H$ . The result is essentially a consequence of the compensated compactness theory of Murat [23] and Tartar [28].

**PROPOSITION 4.6.** *The function  $H$  defined by (4.6) satisfies the following convexity properties:*

(i) *If  $N = 1$ , then*

$$(4.21) \quad H(\lambda\xi_1 + (1 - \lambda)\xi_2, \eta, \lambda p_1 + (1 - \lambda)p_2) \leq \lambda H(\xi_1, \eta, p_1) + (1 - \lambda)H(\xi_2, \eta, p_2)$$

*for every  $\xi_1, \xi_2, \eta \in \mathbf{R}$  and every  $p_1, p_2, \lambda \in [0, 1]$ .*

(ii) *If  $N \geq 2$ , then*

$$(4.22) \quad \begin{aligned} & H(\lambda\xi_1 + (1 - \lambda)\xi_2, \lambda\eta_1 + (1 - \lambda)\eta_2, \lambda p_1 + (1 - \lambda)p_2) \\ & \leq \lambda H(\xi_1, \eta_1, p_1) + (1 - \lambda)H(\xi_2, \eta_2, p_2) \end{aligned}$$

for every  $\xi_1, \xi_2, \eta_1, \eta_2 \in \mathbf{R}^N$ , with  $(\xi_2 - \xi_1) \cdot (\eta_2 - \eta_1) = 0$ , and every  $p_1, p_2, \lambda \in [0, 1]$ .

As a consequence of this result we will prove the following proposition which improves Theorem 4.2.

**PROPOSITION 4.7.** *The function  $H$  is continuous on its domain.*

*Remark 4.9.* In Theorem 4.2, we gave a lower bound for  $H$  by assuming  $F$  convex. An analogous proof shows that for  $F$  concave we have

$$H(\xi, \eta, p) \leq \max \left\{ pF \left( \xi + \frac{\nu}{p} \right) + (1-p)F \left( \xi - \frac{\nu}{1-p} \right) : (B-A)\nu = \Lambda_p \xi - \eta, \nu \in E(\xi, p) \right\}$$

for every  $(\xi, \eta, p) \in \text{Dom}(H)$ . Indeed, since  $H$  is defined by a minimum, it is not difficult to obtain upper bounds for  $H$ . In this way, by using (4.13), (4.22), and  $H$  lower semicontinuous for the  $\mathcal{T}$ -convergence, we can use the reasoning at the end of the proof of Proposition 3.3 to show that for every  $F$  satisfying (4.1) and (4.2) (not necessarily concave), every  $(\xi, \eta, p) \in \text{Dom}(H)$ ,  $p \in (0, 1)$ , and every  $\nu \in E(\xi, p)$ , with  $(B-A)\nu = \Lambda_p \xi - \eta$ , we have

$$H(\xi, \eta, p) \leq \left( 1 - \frac{1}{\lambda} \right) F(\xi) + \frac{1}{\lambda} \left( pF \left( \xi + \frac{\lambda\nu}{p} \right) + (1-p)F \left( \xi - \frac{\lambda\nu}{1-p} \right) \right),$$

where

$$\lambda = \frac{(\Lambda_p \xi - \eta) \cdot \xi}{\left( \frac{A}{p} + \frac{B}{1-p} \right) \nu \cdot \nu} \in [0, 1].$$

In particular, for  $F$  concave, this proves that

$$H(\xi, \eta, p) \leq \min \left\{ pF \left( \xi + \frac{\nu}{p} \right) + (1-p)F \left( \xi - \frac{\nu}{1-p} \right) : (B-A)\nu = \Lambda_p \xi - \eta, \nu \in E(\xi, p) \right\}.$$

**5. Proofs of the results of section 4.** Throughout this section, for a measurable set  $Z \subset Y$  and  $\xi \in \mathbf{R}^N$ , we usually associate a matrix function  $S$ , defined by

$$(5.1) \quad S = A\chi_Z + B\chi_{Y \setminus Z},$$

and a function  $w$  solution of (3.2).

*Proof of Proposition 4.1.* The result is a simple consequence of Remark 4.1 and the fact that the solution  $w$  of (3.2) with  $|Z| = p \in [0, 1]$  satisfies (3.9), with  $\eta$  given by (3.10).  $\square$

In order to prove Theorem 4.2, we first obtain some bounds for  $\nabla w$ , with the  $w$  solution of (3.2). This will be done in Lemmas 5.1 and 5.2 below.

**LEMMA 5.1.** *For every  $Z \subset Y$  measurable and  $\xi \in \mathbf{R}^N$ , the function  $w$  satisfies*

$$(5.2) \quad \|\xi + \nabla w\|_{L^2(Y)^N} \leq \frac{\beta}{\alpha} |\xi|.$$

Moreover, there exist  $r > 2$  and  $C > 0$ , which depend only on  $\beta/\alpha$  and  $N$ , such that  $w \in W_{\sharp}^{1,r}(Y)$  and

$$(5.3) \quad \|\xi + \nabla w\|_{L^r(Y)^N} \leq C|\xi|.$$

*Proof.* The proof of (5.2) easily follows by using  $w$  as a test function in (3.2). Estimate (5.3) is a consequence of Meyer's regularity theorem [19] and (5.2).  $\square$

LEMMA 5.2. *There exist  $C > 0$  and  $\rho \in (0, 1)$  (which depend only on  $\beta/\alpha$  and  $N$ ) such that, for every  $\xi, \xi' \in \mathbf{R}^N$  and every  $Z, Z' \subset Y$  measurable, the corresponding functions  $S, S', w, w'$  satisfy*

$$(5.4) \quad \|\xi + \nabla w - \xi' - \nabla w'\|_{L^2(Y)^N} \leq C(|\xi - \xi'| + |\xi'| |Z \triangle Z'|^\rho),$$

$$(5.5) \quad \left| \int_Y S(\xi + \nabla w) dy - \int_Y S'(\xi' + \nabla w') dy \right| \leq C(|\xi - \xi'| + |\xi'| |Z \triangle Z'|^\rho).$$

*Proof.* By taking  $w - w'$  as a test function in the difference of the equations satisfied by  $w$  and  $w'$  and adding and subtracting convenient terms, we get

$$\begin{aligned} & \int_Y S(\xi + \nabla w - \xi' - \nabla w') \cdot (\xi + \nabla w - \xi' - \nabla w') dx \\ &= \int_Y (S' - S)(\xi' + \nabla w') \cdot (\xi + \nabla w - \xi' - \nabla w') dy \\ &+ \int_Y S(\xi + \nabla w - \xi' - \nabla w') \cdot (\xi - \xi') dy - \int_Y (S' - S)(\xi' + \nabla w') \cdot (\xi - \xi') dy. \end{aligned}$$

By using the ellipticity of  $S$ , Young's inequality, and  $S = S'$  in  $Y \setminus (Z \triangle Z')$ , we obtain

$$(5.6) \quad \int_Y |\xi + \nabla w - \xi' - \nabla w'|^2 dy \leq C \left( \int_{Z \triangle Z'} |\xi' + \nabla w'|^2 dy + |\xi - \xi'|^2 \right),$$

where  $C$  depends only on  $\beta/\alpha$ . The first term on the right-hand side of this inequality can be estimated by (5.3), which gives

$$(5.7) \quad \int_{Z \triangle Z'} |\xi' + \nabla w'|^2 dy \leq \left( \int_Y |\xi' + \nabla w'|^r dy \right)^{\frac{2}{r}} |Z \triangle Z'|^{1 - \frac{2}{r}} \leq C |\xi'|^2 |Z \triangle Z'|^{1 - \frac{2}{r}}.$$

By substituting (5.7) into (5.6) we get (5.4).

In order to prove (5.5) we now use

$$\begin{aligned} & \left| \int_Y S(\xi + \nabla w) dy - \int_Y S'(\xi' + \nabla w') dy \right| \\ & \leq \left| \int_Y S(\xi + \nabla w - \xi' - \nabla w') dy \right| + \left| \int_Y (S - S')(\xi' + \nabla w') dy \right| \\ & \leq C \left( \int_Y |\xi + \nabla w - \xi' - \nabla w'| dy + \int_{Z \triangle Z'} |\xi' + \nabla w'| dy \right). \end{aligned}$$

By using the Cauchy-Schwarz inequality, (5.4), and (5.7), we obtain (5.5).  $\square$

Let us now use Lemma 5.2 to study some semicontinuity properties for  $H_\delta$ .

LEMMA 5.3. *There exist  $C > 0$  (depending on  $L, \beta/\alpha, \varrho$ , and  $N$ ) and  $\rho \in (0, 1)$  (depending on  $\beta/\alpha$  and  $N$ ) such that, for every  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1]$  and every  $\delta, \varepsilon > 0$ , there exists  $\tau \in [0, \delta)$ , with*

$$(5.8) \quad \begin{aligned} & H_{\tau + \lambda' + |\eta - \eta'|}(\xi', \eta', p') \leq H_\delta(\xi, \eta, p) + (1 + |\xi| + |\xi'|)^{2-e} (\lambda')^e + \varepsilon \\ & \forall (\xi', \eta', p') \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1], \end{aligned}$$

where we have denoted

$$\lambda' = C(|\xi - \xi'| + |\xi'| |p - p'|^\rho).$$

*Proof.* To prove (5.8), we consider  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1]$  and  $\delta, \varepsilon > 0$ . If  $(\xi, \eta, p)$  does not belong to the domain of  $H_\delta$ , then (5.8) is trivial. In another case, by definition (4.5) of  $H_\delta$  there exists  $Z \subset Y$ , with  $|Z| = p$ , such that, by taking  $S$  and  $w$  as the corresponding functions associated to  $Z$  and  $\xi$  and defining  $\tau$  by

$$\tau = \left| \int_Y S(\xi + \nabla w) dy - \eta \right|,$$

we have  $\tau < \delta$  and

$$H_\delta(\xi, \eta, p) > \int_Y F(\xi + \nabla w) dy - \varepsilon.$$

Now we consider  $(\xi', \eta', p') \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1]$ . Then (by adding or subtracting a measurable set of  $Y$  to  $Z$ ) we construct a set  $Z'$ , with  $|Z'| = p'$  and  $|Z \triangle Z'| = |p - p'|$ . We take  $w'$  and  $S'$  as the corresponding functions associated to  $Z'$  and  $\xi'$ . From Lemma 5.2, there exist  $C > 0$  and  $\rho \in (0, 1)$ , which depend only on  $\beta/\alpha$  and  $N$ , such that (5.4) and (5.5) hold. By using then (4.1), (5.2), and (5.4), we have (for another constant  $C$ )

$$\begin{aligned} & \left| \int_Y (F(\xi + \nabla w) - F(\xi' + \nabla w')) dy \right| \\ & \leq L \int_Y (1 + |\xi + \nabla w| + |\xi' + \nabla w'|)^{2-e} |\xi + \nabla w - \xi' - \nabla w'|^e dy \\ & \leq L (1 + \|\xi + \nabla w\|_{L^2(Y)} + \|\xi' + \nabla w'\|_{L^2(Y)})^{2-e} \|\xi + \nabla w - \xi' - \nabla w'\|_{L^2(Y)}^e \\ & \leq C(1 + |\xi| + |\xi'|)^{2-e} (|\xi - \xi'| + |\xi'| |p - p'|^\rho)^e \leq (1 + |\xi| + |\xi'|)^{2-e} (\lambda')^e. \end{aligned}$$

On the other hand, by (5.5) and the definition of  $\tau$ , we have

$$\begin{aligned} & \left| \int_Y S'(\xi' + \nabla w') dy - \eta' \right| \leq \left| \int_Y S'(\xi' + \nabla w') dy - \int_Y S(\xi + \nabla w) dy \right| \\ & + \left| \int_Y S(\xi + \nabla w) dy - \eta \right| + |\eta - \eta'| \leq \lambda' + \tau + |\eta - \eta'|. \end{aligned}$$

Then, by definition (4.5) of  $H_\delta$ , we get

$$\begin{aligned} H_{\tau+\lambda'+|\eta-\eta'|}(\xi', \eta', p') & \leq \int_Y F(\xi' + \nabla w') dy \\ & \leq \int_Y F(\xi + \nabla w) dy + \left| \int_Y (F(\xi + \nabla w) - F(\xi' + \nabla w')) dy \right| \\ & \leq H_\delta(\xi, \eta, p) + \varepsilon + (1 + |\xi| + |\xi'|)^{2-e} (\lambda')^e. \quad \square \end{aligned}$$

*Remark 5.1.* Since  $H_\delta(\xi, \eta, p)$  is decreasing in  $\delta$  inequality (5.8) implies that

$$\begin{aligned} H_{\delta+\lambda'+|\eta-\eta'|}(\xi', \eta', p') & \leq H_\delta(\xi, \eta, p) + (1 + |\xi| + |\xi'|)^{2-e} (\lambda')^e + \varepsilon \\ \forall (\xi', \eta', p') & \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1] \end{aligned}$$

for every  $\varepsilon > 0$ . So by taking  $\varepsilon$  converging to zero we get

$$(5.9) \quad \begin{aligned} H_{\delta+\lambda'+|\eta-\eta'|}(\xi', \eta', p') &\leq H_\delta(\xi, \eta, p) + (1 + |\xi| + |\xi'|)^{2-e}(\lambda')^e \\ \forall (\xi', \eta', p') &\in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1]. \end{aligned}$$

We are now in position to prove that  $H_\delta$  satisfies the following properties.

**PROPOSITION 5.4.** *For every  $\delta > 0$ ,  $H_\delta$  is upper semicontinuous in  $\mathbf{R}^N \times \mathbf{R}^N \times [0, 1]$  and satisfies*

$$(5.10) \quad |H_\delta(\xi, \eta, p)| \leq L \left( \frac{\beta}{\alpha} |\xi| \right)^e \left( 1 + \frac{\beta}{\alpha} |\xi| \right)^{2-e} \quad \forall (\xi, \eta, p) \in \text{Dom}(H_\delta).$$

Moreover, the following lower semicontinuity result holds:

$$(5.11) \quad H_{\delta+s}(\xi, \eta, p) \leq \liminf_{n \rightarrow \infty} H_\delta(\xi_n, \eta_n, p_n) \quad \forall \delta, s > 0$$

for every  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1]$  and for every sequence  $(\xi_n, \eta_n, p_n) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1]$  which converges to  $(\xi, \eta, p)$ .

*Proof.* The proof of (5.10) immediately follows from definition (4.5) of  $H_\delta$ , by taking into account that (5.2), (4.2), and (4.1) imply that for every  $\xi \in \mathbf{R}^N$  and every  $Z \subset Y$  measurable the solution  $w$  of (3.2) satisfies

$$(5.12) \quad \left| \int_Y F(\xi + \nabla w) dy \right| \leq L \left( \frac{\beta}{\alpha} |\xi| \right)^e \left( 1 + \frac{\beta}{\alpha} |\xi| \right)^{2-e}.$$

To prove the upper semicontinuity of  $H$  we consider  $(\xi, \eta, p)$  and  $(\xi_n, \eta_n, p_n)$  as above. By (5.8), for every  $\delta, \varepsilon > 0$ , there exists  $\tau \in [0, \delta)$  (which does not depend on  $n$ ) such that

$$H_{\tau+\lambda_n+|\eta-\eta_n|}(\xi_n, \eta_n, p_n) \leq H_\delta(\xi, \eta, p) + (1 + |\xi| + |\xi_n|)^{2-e} \lambda_n^e + \varepsilon \quad \forall n \in \mathbf{N},$$

with  $\lambda_n = C(|\xi - \xi_n| + |\xi_n| |p - p_n|^\rho)$ . So, since for  $n$  large enough  $\tau + \lambda_n + |\eta - \eta_n| < \delta$  and  $H_\mu$  is decreasing in  $\mu$ , we have

$$H_\delta(\xi_n, \eta_n, p_n) \leq H_\delta(\xi, \eta, p) + (1 + |\xi| + |\xi_n|)^{2-e} \lambda_n^e + \varepsilon.$$

By taking the limsup in  $n$  and then letting  $\varepsilon$  decrease to zero, we deduce the upper semicontinuity of  $H_\delta$ .

In order to prove (5.11), we take  $(\xi, \eta, p)$ ,  $(\xi_n, \eta_n, p_n)$  as above. By (5.9) we have

$$H_{\delta+\lambda_n+|\eta-\eta_n|}(\xi, \eta, p) \leq H_\delta(\xi_n, \eta_n, p_n) + (1 + |\xi| + |\xi'_n|)^{2-e} \lambda_n^e,$$

with  $\lambda_n = C(|\xi - \xi_n| + |\xi| |p - p_n|^\rho)$ . So, by using as above the fact that  $H_\mu$  is decreasing in  $\mu$ , we have for every  $s > 0$  and  $n$  large enough

$$H_{\delta+s}(\xi, \eta, p) \leq H_\delta(\xi_n, \eta_n, p_n) + (1 + |\xi| + |\xi'_n|)^{2-e} \lambda_n^e.$$

By taking the liminf in  $n$  we deduce (5.11).  $\square$

Thanks to the previous results, we can now prove Theorem 4.2.

*Proof of Theorem 4.2.* By definition (4.6) of  $H$  and (5.12), we have that  $(\xi, \eta, p) \in \text{Dom}(H)$  if and only if for every  $\delta > 0$  there exists  $Z \subset Y$  measurable, with  $|Z| = p$ , such that the solution  $w$  of (3.2) satisfies

$$\left| \int_Y S(\nabla w + \xi) dy - \eta \right| < \delta.$$

By Theorem 3.1 this holds if and only if  $\eta$  belongs to  $\mathcal{K}(A, B, p)\xi$ . This proves (4.10).

In order to prove the lower semicontinuity of  $H$ , we consider  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1]$  and  $(\xi_n, \eta_n, p_n) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1]$ , which converges to  $(\xi, \eta, p)$ . By (5.11) with  $s = \delta$  and  $H_\delta(\xi_n, \eta_n, p_n) \leq H(\xi_n, \eta_n, p_n)$ , we have

$$H_{2\delta}(\xi, \eta, p) \leq \liminf_{n \rightarrow \infty} H(\xi_n, \eta_n, p_n) \quad \forall \delta > 0.$$

By taking the limit when  $\delta$  tends to zero we conclude that

$$H(\xi, \eta, p) \leq \liminf_{n \rightarrow \infty} H(\xi_n, \eta_n, p_n)$$

and then the lower semicontinuity of  $H$ .

Inequality (4.11) immediately follows from (5.10).

To show (4.12) it is enough to use Remark 4.1 and the fact that, if  $Z \subset Y$  has measure 0 or 1, the solution  $w$  of (3.2) is zero.

To prove (4.13), we consider  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times (0, 1)$  such that  $\Lambda_p \xi - \eta \in \text{Ran}(A - B)$  and

$$(5.13) \quad \left( (A - B) \left( \frac{A}{p} + \frac{B}{1-p} \right)^{-1} (A - B) \right)^\dagger (\Lambda_p \xi - \eta) \cdot (\Lambda_p \xi - \eta) = \xi \cdot (\Lambda_p \xi - \eta).$$

Then we consider  $Z_n, S_n, w_n, \eta_n$  as in Remark 4.1, and we define

$$\nu_n = \int_{Z_n} \nabla w_n dy = - \int_{Y \setminus Z_n} \nabla w_n dy.$$

By (3.5) and (3.6),  $\nu_n$  satisfies that  $(B - A)\nu_n = \Lambda_p \xi - \eta_n$  and

$$(5.14) \quad \begin{cases} \int_{Z_n} A \left( \nabla w_n - \frac{\nu_n}{p} \right) \cdot \left( \nabla w_n - \frac{\nu_n}{p} \right) dy \\ + \int_{Y \setminus Z_n} B \left( \nabla w_n + \frac{\nu_n}{1-p} \right) \cdot \left( \nabla w_n + \frac{\nu_n}{1-p} \right) dy \\ = (\Lambda_p \xi - \eta_n) \cdot \xi - \left( \frac{A}{p} + \frac{B}{1-p} \right) \nu_n \cdot \nu_n. \end{cases}$$

This implies in particular (use the fact that the left-hand side of (5.14) is nonnegative) that  $\nu_n$  is bounded, and so, up to a subsequence, we have that  $\nu_n$  converges to some  $\hat{\nu} \in \mathbf{R}^N$  such that  $(B - A)\hat{\nu} = \Lambda_p \xi - \eta$  and

$$(5.15) \quad \left( \frac{A}{p} + \frac{B}{1-p} \right) \hat{\nu} \cdot \hat{\nu} \leq (\Lambda_p \xi - \eta) \cdot \xi.$$



By using the fact that (5.13) can also be written as

$$(\Lambda_p \xi - \eta) \cdot \xi = \min \left\{ \left( \frac{A}{p} + \frac{B}{1-p} \right) \nu \cdot \nu : \Lambda_p \xi - \eta = (B - A)\nu \right\},$$

we then deduce that  $\hat{\nu}$  gives the minimum above, and so  $\hat{\nu} = \Xi(\Lambda_p \xi - \eta)$ . In particular, this implies that (5.15) is an equality, and then by passing to the limit in (5.14) we have

$$\lim_{n \rightarrow 0} \left( \int_{Z_n} \left| \nabla w_n - \frac{\nu_n}{p} \right|^2 dz + \int_{Y \setminus Z_n} \left| \nabla w_n + \frac{\nu_n}{1-p} \right|^2 dz \right) = 0,$$

which, joining to (4.1), allows us to calculate the limit which appears in the right-hand side of (4.8) and then to conclude (4.13).

To finish the proof of Theorem 4.2, let us now prove (4.14). For  $(\xi, \eta, p) \in \text{Dom}(H)$ ,  $p \in (0, 1)$ , we take  $Z_n$  and  $w_n$  as in Remark 4.1. By Lemma 3.2, we can assume that

$$\int_{Z_n} \nabla w_n dy \rightarrow \hat{\nu} \in E(\xi, p), \quad (B - A)\hat{\nu} = \Lambda_p \xi - \eta.$$

Jensen's inequality,  $w_n$ -periodic, and  $|Z_n| = p$  give

$$\begin{aligned} \int_Y F(\xi + \nabla w_n) dy &= \int_{Z_n} F(\xi + \nabla w_n) dy + \int_{Y \setminus Z_n} F(\xi + \nabla w_n) dy \\ &\geq pF\left(\xi + \frac{1}{p} \int_{Z_n} \nabla w_n dy\right) + (1-p)F\left(\xi - \frac{1}{1-p} \int_{Z_n} \nabla w_n dy\right). \end{aligned}$$

By taking the limit in this inequality we deduce that

$$H(\xi, \eta, p) = \lim_{n \rightarrow \infty} \int_Y F(\xi + \nabla w_n) dy \geq pF\left(\xi + \frac{\hat{\nu}}{p}\right) + (1-p)F\left(\xi - \frac{\hat{\nu}}{1-p}\right),$$

and then (4.14).  $\square$

To prove Theorem 4.5 we need the following corrector result. We use some ideas which appear in the proof of Theorem 3.1 given in [1] and [11].

LEMMA 5.5. *We consider a bounded open set  $\Omega \subset \mathbf{R}^N$ , a sequence of matrices  $M_n \in L^\infty(\Omega; \mathcal{M}_N)$ , uniformly bounded and elliptic, and a sequence  $u_n \in H^1(\Omega)$ . We assume that  $u_n$  converges weakly in  $H^1(\Omega)$  to a function  $u$  and that  $-\text{div } M_n \nabla u_n$  is compact in  $H^{-1}(\Omega)$ . For  $\varepsilon > 0$  small enough, we take*

$$(5.16) \quad \Omega_\varepsilon = \{x \in \Omega : \text{dis}(x, \partial\Omega) > \varepsilon\},$$

and, for  $h \in (0, \frac{\varepsilon}{\sqrt{N}})$ , we define  $w_n^h \in L^2(\Omega_\varepsilon; H_\#^1(Y))$  as the unique solution of

$$(5.17) \quad \begin{cases} w_n^h(x, \cdot) \in H_\#^1(Y), & \int_Y w_n^h(x, y) dy = 0, \\ -\text{div}_y M_n(x + hy)(\nabla u(x) + \nabla_y w_n^h(x, y)) = 0 & \text{in } \mathbf{R}^N, \text{ a.e. } x \in \Omega_\varepsilon. \end{cases}$$

Then we have

$$(5.18) \quad \lim_{h \rightarrow 0} \limsup_{n \rightarrow \infty} \|\nabla u_n(x + hy) - \nabla u - \nabla_y w_n^h\|_{L^2(\Omega_\varepsilon \times Y)^N} = 0.$$

*Proof.* By extracting a subsequence if necessary, we can assume that  $M_n$   $H$ -converges to a matrix-valued function  $M$ . Then, by (5.17), we get

$$(5.19) \quad \nabla u(x) \cdot y + w_n^h(x, y) \rightharpoonup \nabla u(x) \cdot y + w^h(x, y) \text{ in } H^1(Y), \text{ a.e. } x \in \Omega_\varepsilon,$$

when  $n$  tends to infinity, with  $w^h \in L^2(\Omega_\varepsilon; H_\#^1(Y))$  the unique solution of the problem

$$(5.20) \quad \begin{cases} w^h(x, \cdot) \in H_\#^1(Y), & \int_Y w^h(x, y) dy = 0, \\ -\operatorname{div}_y (M(x + hy)(\nabla u(x) + \nabla_y w^h(x, y))) = 0 & \text{in } \mathbf{R}^N, \text{ a.e. } x \in \Omega_\varepsilon. \end{cases}$$

By using in (5.20) the fact that  $M(x + hy)$  converges to  $M$  strongly in  $L^q(\Omega_\varepsilon \times Y; \mathcal{M}_N)$ ,  $1 \leq q < +\infty$ , and  $*$ -weakly in  $L^\infty(\Omega_\varepsilon \times Y; \mathcal{M}_N)$ , when  $h$  tends to zero, it is easy to prove that

$$(5.21) \quad w^h \rightarrow 0 \text{ in } L^2(\Omega_\varepsilon; H_\#^1(Y)).$$

On the other hand, the strong convergence in  $H^{-1}(\Omega)$  of  $-\operatorname{div} M_n \nabla u_n$  implies that

$$-\operatorname{div}_y [M_n(x + hy) \nabla u_n(x + hy)] \rightarrow -\operatorname{div}_y [M(x + hy) \nabla u(x + hy)] \text{ in } H^{-1}(Y)$$

for every  $x \in \Omega_\varepsilon$ , when  $n$  tends to infinity. Thanks to (5.17), we can then apply the div-curl lemma (see, e.g., [23], [28]) to deduce that

$$\begin{aligned} & M_n(x + hy) (\nabla u_n(x + hy) - \nabla u - \nabla_y w_n^h) \cdot (\nabla u_n(x + hy) - \nabla u - \nabla_y w_n^h) \\ & \rightharpoonup M(x + hy) (\nabla u(x + hy) - \nabla u - \nabla_y w^h) \cdot (\nabla u(x + hy) - \nabla u - \nabla_y w^h) \end{aligned}$$

in the sense of the distributions in  $Y$ , for a.e.  $x \in \Omega_\varepsilon$ , when  $n$  tends to infinity. By Meyer's theorem this convergence holds in  $L^1(Y)$  weakly, for a.e.  $x \in \Omega_\varepsilon$ , and thus we have

$$(5.22) \quad \begin{aligned} & \int_Y M_n(x + hy) (\nabla u_n(x + hy) - \nabla u - \nabla_y w_n^h) \cdot (\nabla u_n(x + hy) - \nabla u - \nabla_y w_n^h) dy \\ & \rightarrow \int_Y M(x + hy) (\nabla u(x + hy) - \nabla u - \nabla_y w^h) \cdot (\nabla u(x + hy) - \nabla u - \nabla_y w^h) dy \end{aligned}$$

for a.e.  $x \in \Omega_\varepsilon$ . Moreover, by (5.17) there exists  $C > 0$  such that  $\|\nabla u + \nabla_y w_n^h\|_{L^2(Y)^N} \leq C|\nabla u|$  a.e. in  $\Omega_\varepsilon$ , and so, for another constant  $C$ , we have

$$\begin{aligned} & \int_Y M_n(x + hy) (\nabla u_n(x + hy) - \nabla u - \nabla_y w_n^h) \cdot (\nabla u_n(x + hy) - \nabla u - \nabla_y w_n^h) dy \\ & \leq \frac{C}{h^N} \int_\Omega |\nabla u_n|^2 dx + C|\nabla u|^2, \text{ a.e. in } \Omega_\varepsilon, \end{aligned}$$

which together with (5.22) allows us to apply the Lebesgue dominated convergence theorem to deduce that

$$\begin{aligned} & \int_{\Omega_\varepsilon} \int_Y M_n(x + hy) (\nabla u_n(x + hy) - \nabla u - \nabla_y w_n^h) \cdot (\nabla u_n(x + hy) - \nabla u - \nabla_y w_n^h) dy dx \\ & \rightarrow \int_{\Omega_\varepsilon} \int_Y M(x + hy) (\nabla u(x + hy) - \nabla u - \nabla_y w^h) \cdot (\nabla u(x + hy) - \nabla u - \nabla_y w^h) dy dx, \end{aligned}$$

when  $n$  tends to infinity, for every  $h \in (0, \varepsilon/\sqrt{N})$ . By (5.21), the right-hand side of this equality tends to zero when  $h$  tends to zero, and then, thanks to the ellipticity of  $M$ , we get (5.18).  $\square$

*Proof of Theorem 4.5.* In the proof we will separate the cases  $N = 1$  and  $N \geq 2$ . We make this distinction because for  $N \geq 2$  we will use a convexity property of the set  $K(A, B, p)\xi$  which does not hold for  $N = 1$  (see Step 3 in the proof). On the other hand, we think it is interesting to show that the one-dimensional case follows by using elementary arguments.

The proof of the theorem will be divided in three steps.

In Step 1 we will consider the case  $N = 1$ , while steps 2 and 3 are devoted to  $N \geq 2$ .

In Step 2 we will prove the inequality

$$(5.23) \quad \begin{cases} \liminf_{n \rightarrow \infty} \mathcal{F}(u_n, \sigma_n, \theta_n) \geq \overline{\mathcal{F}}(u, \sigma, \theta) & \forall (u, \sigma, \theta) \in H^1(\Omega) \times L^2(\Omega)^N \times L^\infty(\Omega) \\ \forall (u_n, \sigma_n, \theta_n) \in H^1(\Omega) \times L^2(\Omega)^N \times L^\infty(\Omega), (u_n, \sigma_n, \theta_n) \xrightarrow{\mathcal{T}} (u, \sigma, \theta). \end{cases}$$

The proof of (5.23) will follow from Lemma 5.5, which provides an approximation of  $\nabla u_n$  in the strong topology of  $H^1(\omega)$  ( $\omega \subset \subset \Omega$ ) by using periodic homogenization.

In Step 3 we prove that for every  $(u, \sigma, \theta) \in H^1(\Omega) \times L^2(\Omega)^N \times L^\infty(\Omega)$  there exists  $(u_n, \sigma_n, \theta_n) \in H^1(\Omega) \times L^2(\Omega)^N \times L^\infty(\Omega)$  such that

$$(5.24) \quad (u_n, \sigma_n, \theta_n) \xrightarrow{\mathcal{T}} (u, \sigma, \theta), \quad \limsup_{n \rightarrow \infty} \mathcal{F}(u_n, \sigma_n, \theta_n) \leq \overline{\mathcal{F}}(u, \sigma, \theta)$$

which joined to (5.23) will give the proof of Theorem 4.5 in the case  $N \geq 2$ . The main idea to prove the existence of  $(u_n, \sigma_n, \theta_n)$  satisfying (5.24) will be to use an approximation by finite elements of  $(u, \sigma, \theta)$  which reduces the problem to the case where there exists a triangulation  $\tau$  such that  $\nabla u$ ,  $\sigma$ , and  $\theta$  are constant in each element of  $\tau$ .

*Step 1.* Let us first prove the result for  $N = 1$ .

Consider  $(u, \sigma, \theta) \in H^1(\Omega) \times L^2(\Omega) \times L^\infty(\Omega)$  and a sequence  $(u_n, \sigma_n, \theta_n) \in H^1(\Omega) \times L^2(\Omega) \times L^\infty(\Omega)$  which  $\mathcal{T}$ -converges to  $(u, \sigma, \theta)$ . Let us prove that

$$(5.25) \quad \overline{\mathcal{F}}(u, \sigma, \theta) \leq \liminf_{n \rightarrow \infty} \mathcal{F}(u_n, \sigma_n, \theta_n).$$

By definition (4.18) of  $\mathcal{F}$ , it is enough to consider the case where there exists a sequence of measurable sets  $\omega_n \subset \Omega$  such that

$$\theta_n = \chi_{\omega_n}, \quad \sigma_n = (A\chi_{\omega_n} + B\chi_{\Omega \setminus \omega_n}) \frac{du_n}{dx} \quad \text{a.e. in } \Omega,$$

and thus

$$(5.26) \quad \frac{du_n}{dx} = \frac{\sigma_n}{A} \chi_{\omega_n} + \frac{\sigma_n}{B} \chi_{\Omega \setminus \omega_n} \quad \text{a.e. in } \Omega.$$

Since  $\sigma_n$  converges weakly to  $\sigma$  in  $L^2(\Omega)$  and  $\frac{d\sigma_n}{dx}$  converges strongly to  $\frac{d\sigma}{dx}$  in  $H^{-1}(\Omega)$ , we have that  $\sigma_n$  converges strongly to  $\sigma$  in  $L^2(\Omega)$ . So by (5.26) we get

$$\frac{du}{dx} = \left( \frac{\theta}{A} + \frac{1-\theta}{B} \right) \sigma \quad \text{a.e. in } \Omega$$

and

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathcal{F}(u_n, \sigma_n, \theta_n) &= \lim_{n \rightarrow \infty} \int_{\Omega} F\left(\frac{du_n}{dx}\right) dx \\ &= \lim_{n \rightarrow \infty} \int_{\Omega} \left( F\left(\frac{\sigma_n}{A}\right) \chi_{\omega_n} + F\left(\frac{\sigma_n}{B}\right) \chi_{\Omega \setminus \omega_n} \right) dx = \int_{\Omega} \left( F\left(\frac{\sigma}{A}\right) \theta + F\left(\frac{\sigma}{B}\right) (1 - \theta) \right) dx. \end{aligned}$$

By (4.9), we have then proved that  $(u, \sigma, \theta)$  belongs to  $\text{Dom}(H)$  a.e. in  $\Omega$  and

$$\overline{\mathcal{F}}(u, \sigma, \theta) = \int_{\Omega} H\left(\frac{du}{dx}, \sigma, \theta\right) dx = \lim_{n \rightarrow \infty} \mathcal{F}(u_n, \sigma_n, \theta_n).$$

To finish the proof of Step 1, we need to prove that for every  $(u, \sigma, \theta) \in H^1(\Omega) \times L^2(\Omega) \times L^\infty(\Omega)$  there exists  $(u_n, \sigma_n, \theta_n) \in H^1(\Omega) \times L^2(\Omega) \times L^\infty(\Omega)$ , which  $\mathcal{T}$ -converges to  $(u, \sigma, \theta)$  and satisfies

$$(5.27) \quad \limsup_{n \rightarrow \infty} \mathcal{F}(u_n, \sigma_n, \theta_n) \leq \overline{\mathcal{F}}(u, \sigma, \theta).$$

Clearly, it is enough to consider the case where  $(u, \sigma, \theta) \in \text{Dom}(H)$  a.e. in  $\Omega$ , but then  $\theta \in [0, 1]$  a.e. in  $\Omega$ , and so there exists  $\omega_n \subset \Omega$  such that  $\chi_{\omega_n}$  converges weakly-\* in  $L^\infty(\Omega)$  to  $\theta$ . By taking

$$\theta_n = \chi_{\omega_n}, \quad \sigma_n = \sigma, \quad \frac{du_n}{dx} = \frac{\sigma}{A} \chi_{\omega_n} + \frac{\sigma}{B} \chi_{\Omega \setminus \omega_n}$$

and reasoning as above, we deduce (5.27).

In the remainder of the proof we always assume that  $N \geq 2$ .

*Step 2.* Let us prove (5.23).

We can assume that

$$\liminf_{n \rightarrow \infty} \mathcal{F}(u_n, \sigma_n, \theta_n) < +\infty,$$

and thus, by extracting a subsequence if necessary, there exists a sequence of measurable sets  $\omega_n \subset \Omega$  such that  $\theta_n = \chi_{\omega_n}$  and  $\sigma_n = M_n \nabla u_n$ , with  $M_n = A \chi_{\omega_n} + B \chi_{\Omega \setminus \omega_n}$ . By using the compactness theorem for the  $H$ -convergence (see, e.g., [1], [11], [22], [26], [27]), we can also assume that there exists an elliptic matrix  $M \in L^\infty(\Omega; \mathcal{M}_N^s)$  such that  $M_n$   $H$ -converges to  $M$ . Since  $\text{div } \sigma_n$  converges strongly to  $\text{div } \sigma$  in  $H^{-1}(\Omega)$ , and  $M_n \nabla u_n = \sigma_n$ , we have

$$(5.28) \quad \sigma = M \nabla u \quad \text{a.e. in } \Omega.$$

In order to prove (5.23), we consider  $\varepsilon > 0$ . By defining  $\Omega_\varepsilon$  by (5.16) and taking  $h \in (0, \varepsilon/\sqrt{N})$  in such a way that  $\Omega_\varepsilon \subset \Omega - hy$  for every  $y \in Y$ , we have

$$\begin{aligned} \int_{\Omega} F(\nabla u_n) dx &= \int_{\Omega - hy} F(\nabla u_n(x + hy)) dx \\ (5.29) \quad &= \int_{\Omega_\varepsilon} F(\nabla u_n(x + hy)) dx + \int_{(\Omega - hy) \setminus \Omega_\varepsilon} F(\nabla u_n(x + hy)) dx \\ &= \int_{\Omega_\varepsilon} F(\nabla u_n(x + hy)) dx + \int_{\Omega \setminus (\Omega_\varepsilon + hy)} F(\nabla u_n) dx, \end{aligned}$$

and so, by integrating in  $y \in Y$ , we have

$$(5.30) \quad \int_{\Omega} F(\nabla u_n) dx = \int_Y \int_{\Omega_{\varepsilon}} F(\nabla u_n(x+hy)) dx dy + \int_Y \int_{\Omega \setminus (\Omega_{\varepsilon} + hy)} F(\nabla u_n) dx dy.$$

By using the fact that  $\Omega \setminus (\Omega_{\varepsilon} + hy) \subset \Omega \setminus \Omega_{2\varepsilon}$  for every  $y \in Y$ , (4.3), and the fact that  $|\nabla u_n|^2$  is equi-integrable (which follows from the definition of  $\mathcal{T}$ -convergence), we have

$$(5.31) \quad \begin{aligned} & \lim_{\varepsilon \rightarrow 0} \limsup_{n \rightarrow \infty} \left| \int_Y \int_{\Omega \setminus (\Omega_{\varepsilon} + hy)} F(\nabla u_n) dx dy \right| \\ & \leq L \lim_{\varepsilon \rightarrow 0} \limsup_{n \rightarrow \infty} \int_{\Omega \setminus \Omega_{2\varepsilon}} (1 + |\nabla u_n|)^2 dx = 0. \end{aligned}$$

To estimate the first term of (5.30), we use the decomposition

$$(5.32) \quad \begin{aligned} \int_Y \int_{\Omega_{\varepsilon}} F(\nabla u_n(x+hy)) dx dy &= \int_Y \int_{\Omega_{\varepsilon}} (F(\nabla u_n(x+hy)) - F(\nabla u + \nabla_y w_n^h)) dx dy \\ &\quad + \int_{\Omega_{\varepsilon}} \int_Y F(\nabla u + \nabla_y w_n^h) dy dx, \end{aligned}$$

with  $w_n^h$  given by (5.17). Thanks to (4.1) and (5.18), we have

$$(5.33) \quad \lim_{h \rightarrow 0} \limsup_{n \rightarrow \infty} \int_Y \int_{\Omega_{\varepsilon}} |F(\nabla u_n(x+hy)) - F(\nabla u + \nabla_y w_n^h)| dx dy = 0 \quad \forall \varepsilon > 0.$$

To estimate the second term on the right-hand side of (5.32), we denote for a.e.  $x \in \Omega_{\varepsilon}$  and a.e.  $y \in Y$

$$Z_n^h(x) = Y \cap \frac{\omega_n - x}{h}, \quad \theta_n^h(x) = |Z_n^h(x)| = \int_Y \theta_n(x+hy) dy,$$

$$M_n(x+hy) = A\chi_{\omega_n}(x+hy) + B\chi_{\Omega \setminus \omega_n}(x+hy) = A\chi_{Z_n^h(x)}(y) + B\chi_{Y \setminus Z_n^h(x)}(y),$$

$$\sigma_n^h(x) = \int_Y M_n(x+hy)(\nabla u + \nabla_y w_n^h) dy.$$

Then, by definition (5.17) of  $w_n^h$  and definition (4.6) of  $H$ , we obtain

$$(5.34) \quad \int_Y F(\nabla u + \nabla_y w_n^h) dy \geq H(\nabla u, \sigma_n^h, \theta_n^h), \quad \text{a.e. in } \Omega_{\varepsilon}.$$

By the  $H$ -convergence of  $M_n$  to  $M$  and the convergence in  $L^{\infty}(\Omega)$  weak-\* of  $\theta_n$  to  $\theta$ , we have that  $\sigma_n^h$  and  $\theta_n^h$ , respectively, converge a.e. in  $\Omega_{\varepsilon}$  to  $\sigma^h \in L^2(\Omega_{\varepsilon})^N$  and  $\theta^h \in L^{\infty}(\Omega_{\varepsilon})$ , defined by

$$\sigma^h(x) = \int_Y M(x+hy)(\nabla u + \nabla_y w^h) dy, \quad \theta^h(x) = \int_Y \theta(x+hy) dy, \quad \text{a.e. } x \in \Omega_{\varepsilon}.$$

From (4.11) and  $\|\nabla u + \nabla_y w_n^h\|_{L^2(Y)} \leq C|\nabla u|$  a.e. in  $\Omega$ , we also have

$$|H(\nabla u, \sigma_n^h, \theta_n^h)| \leq C(1 + |\nabla u|)^2 \quad \text{a.e. in } \Omega_{\varepsilon}.$$

Then, by (5.34), the lower semicontinuity of  $H$ , and Fatou's lemma we deduce that

$$\begin{aligned} \liminf_{n \rightarrow \infty} \int_{\Omega_\varepsilon} \int_Y F(\nabla u + \nabla_y w_n^h) dy dx &\geq \liminf_{n \rightarrow \infty} \int_{\Omega_\varepsilon} H(\nabla u, \sigma_n^h, \theta_n^h) dx \\ &\geq \int_{\Omega_\varepsilon} H(\nabla u, \sigma^h, \theta^h) dx \end{aligned}$$

for every  $h \in (0, \varepsilon/\sqrt{N})$ . By now using the fact that, for  $h$  tending to zero,  $\sigma^h$  converges strongly to  $\sigma$  in  $L^2(\Omega_\varepsilon)^N$  and  $\theta^h$  converges to  $\theta$  strongly in  $L^r(\Omega_\varepsilon)$ ,  $1 \leq r < +\infty$ , weak-\* in  $L^\infty(\Omega_\varepsilon)$ , we can use again the lower semicontinuity of  $H$  and Fatou's lemma to obtain

$$(5.35) \quad \liminf_{h \rightarrow 0} \liminf_{n \rightarrow \infty} \int_{\Omega_\varepsilon} \int_Y F(\nabla u + \nabla_y w_n^h) dy dx \geq \int_{\Omega_\varepsilon} H(\nabla u, \sigma, \theta) dx.$$

By (5.30), (5.31), (5.32), (5.33), and (5.35), by passing to the limit in (5.29), first in  $n$ , then in  $h$ , and then in  $\varepsilon$ , we have

$$\liminf_{n \rightarrow \infty} \int_{\Omega} F(\nabla u_n) dx \geq \int_{\Omega} H(\nabla u, \sigma, \theta) dx.$$

This proves (5.23).

*Step 3.* Let us prove that for every  $(u, \sigma, \theta) \in H^1(\Omega) \times L^2(\Omega)^N \times L^\infty(\Omega)$  there exists  $(u_n, \sigma_n, \theta_n) \in H^1(\Omega) \times L^2(\Omega)^N \times L^\infty(\Omega)$ , which satisfies (5.24).

It is enough to consider the case where  $0 \leq \theta \leq 1$  and  $\sigma \in \mathcal{K}(A, B, \theta) \nabla u$  a.e. in  $\Omega$ .

We consider an open cube  $Q$ , with  $\bar{\Omega} \subset Q$ , and a prolongation of  $u$ , still denoted by  $u$ , in  $H_0^1(Q)$ . This prolongation exists thanks to  $\Omega$  being smooth. We also extend  $\theta$  and  $\sigma$  to the whole  $Q$  by  $\theta = 1$ ,  $\sigma = A \nabla u$  a.e. in  $Q \setminus \Omega$ . The multiapplication  $x \in \Omega \mapsto \{\tilde{M} \in \mathcal{K}(A, B, \theta(x)) : \tilde{M} \nabla u(x) = \sigma(x)\}$  is closed and measurable. The measurability follows by using the fact that for every closed set  $\mathcal{C} \subset \mathcal{M}_N^s$  we have

$$\{x \in \Omega : \exists \tilde{M} \in \mathcal{C} \cap \mathcal{K}(A, B, \theta(x)), \text{ with } \tilde{M} \nabla u(x) = \sigma(x)\} = \bigcap_{m \in \mathbf{N}} (\nabla u, \sigma, \theta)^{-1}(K_m),$$

where

$$K_m = \left\{ (\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times [0, 1] : \exists \tilde{M} \in \mathcal{K}(A, B, p) \cap \mathcal{C} \text{ with } |\tilde{M}\xi - \eta| \leq \frac{1}{m} \right\}$$

is closed for every  $m \in \mathbf{N}$ . Thus (see, e.g., [7]) there exists  $M \in L^\infty(\Omega; \mathcal{M}_N^s)$  such that  $M \in \mathcal{K}(A, B, \theta)$  and  $M \nabla u = \sigma$  a.e. in  $Q$ .

For a regular sequence of triangulations  $\tau_n = \{T_n^k\}_{1 \leq k \leq k_n}$  of  $Q$  by  $N$ -simplex, whose diameter tends to zero when  $n$  tends to infinity (see, e.g., [8]), we consider the space of finite elements

$$V_n = \{v_n \in C_0^0(\bar{Q}) : v_n \text{ is affine in } T_n^k \quad \forall k \in \{1, \dots, k_n\}\}.$$

Then we define  $\hat{u}_n$  as the solution of

$$\begin{cases} \hat{u}_n \in V_n, \\ \int_Q M \nabla \hat{u}_n \nabla v = \int_Q \sigma \nabla v dx \quad \forall v \in V_n, \end{cases}$$

and we take

$$\hat{\theta}_n = \sum_{k=1}^{k_n} \left( \frac{1}{|T_n^k|} \int_{T_n^k} \theta \, dx \right) \chi_{T_n^k}, \quad \hat{\sigma}_n = \sum_{k=1}^{k_n} \left( \frac{1}{|T_n^k|} \int_{T_n^k} M \nabla \hat{u}_n \, dx \right) \chi_{T_n^k}.$$

The sequence  $\hat{\theta}_n$  converges to  $\theta$  in  $L^r(Q)$ ,  $1 \leq r < +\infty$ , and in  $L^\infty(Q)$  weak-\*. Since  $\tau_n$  is regular, and  $\sigma = M \nabla u$  a.e. in  $Q$ , we also have that  $\hat{u}_n$  converges strongly in  $H_0^1(Q)$  to  $u$  (see, e.g., [8]). Thus,  $M \nabla \hat{u}_n$  and then  $\hat{\sigma}_n$  converge strongly to  $\sigma$  in  $L^2(Q)^N$ . By Egorov's theorem, there exist a subsequence of  $n$ , still denoted by  $n$ , and a sequence  $Q_n$  of closed subsets of  $Q$  such that

$$(5.36) \quad |Q - Q_n| < \frac{1}{n}, \quad C(|\nabla \hat{u}_n - \nabla u| + |\nabla u| |\hat{\theta}_n - \theta|^\rho) + |\hat{\sigma}_n - \sigma| < \frac{1}{n} \text{ in } Q_n,$$

with  $C > 0$  and  $\rho \in (0, 1)$  given by Lemma 5.3.

Since  $\nabla \hat{u}_n$  is constant in every  $N$ -simplex  $T_n^k$  of  $\tau_n$  and, for every  $\xi \in \mathbf{R}^N$ , the set  $\mathcal{K}(A, B, p)\xi$  is closed in  $\mathbf{R}^N$  and satisfies the following convexity property (this property does not hold for  $N = 1$ , and so it is the reason to prove the cases  $N = 1$ ,  $N \geq 2$  separately):

$$\lambda \mathcal{K}(A, B, p_1)\xi + (1 - \lambda) \mathcal{K}(A, B, p_2)\xi \subset \mathcal{K}(A, B, \lambda p_1 + (1 - \lambda)p_2)\xi \quad \forall \lambda, p_1, p_2 \in [0, 1],$$

we have

$$\hat{\sigma}_n = \left( \frac{1}{|T_n^k|} \int_{T_n^k} M \, dx \right) \nabla \hat{u}_n \in \mathcal{K} \left( A, B, \frac{1}{|T_n^k|} \int_{T_n^k} \theta \, dx \right) \nabla \hat{u}_n$$

in  $T_n^k$ , for every  $k \in \{1, \dots, k_n\}$ , i.e.,  $\hat{\sigma}_n \in \mathcal{K}(A, B, \hat{\theta}_n) \nabla \hat{u}_n$  a.e. in  $Q$ , or equivalently

$$(\nabla \hat{u}_n, \hat{\sigma}_n, \hat{\theta}_n) \in \text{Dom}(H) \text{ a.e. in } Q.$$

We consider an  $N$ -simplex  $T_n^k$ , with  $k \in \{1, \dots, k_n\}$ , and we denote by  $\xi_n^k = \nabla \hat{u}_n$ ,  $\eta_n^k = \hat{\sigma}_n$ , and  $p_n^k = \hat{\theta}_n$  the constant values of  $\nabla \hat{u}_n$ ,  $\hat{\sigma}_n$ , and  $\hat{\theta}_n$ , respectively, in  $T_n^k$ . From definition (4.5) of  $H_{\frac{2}{n}}(\xi_n^k, \eta_n^k, p_n^k)$ , there exists  $Z_n^k \subset Y$ , with  $|Z_n^k| = p_n^k$ , such that, by defining  $S_n^k \in L_{\#}^\infty(Y; \mathcal{M}_N^s)$  by  $S_n^k = A \chi_{Z_n^k} + B \chi_{Y \setminus Z_n^k}$  a.e. in  $Y$ , and taking  $w_n^k$  the solution of

$$\begin{cases} w_n^k \in H_{\#}^1(Y), & \int_Y w_n^k \, dy = 0, \\ -\text{div } S_n^k(\xi_n^k + \nabla w_n^k) = 0 & \text{in } \mathbf{R}^N, \end{cases}$$

we have

$$(5.37) \quad \left| \int_Y S_n^k(\xi_n^k + \nabla w_n^k) \, dy - \eta_n^k \right| < \frac{2}{n}$$

and

$$(5.38) \quad \int_Y F(\xi_n^k + \nabla w_n^k) \, dy < H_{\frac{2}{n}}(\xi_n^k, \eta_n^k, p_n^k) + \frac{1}{n}.$$

We take

$$(5.39) \quad \check{\sigma}_n = \sum_{k=1}^{k_n} \left( \int_Y S_n^k(\xi_n^k + \nabla w_n^k) \, dy \right) \chi_{T_n^k},$$

and by (5.37), we observe that

$$(5.40) \quad \int_Q |\check{\sigma}_n - \hat{\sigma}_n|^2 dx < \frac{4}{n^2} |Q|.$$

For  $n \in \mathbf{N}$ , and  $\varepsilon > 0$ , we define

$$\omega_{n,\varepsilon} = \bigcup_{k=1}^{k_n} \left[ T_n^k \cap \left( \bigcup_{l \in \mathbf{Z}^N} (\varepsilon l + \varepsilon Z_n^k) \right) \right],$$

$$M_{n,\varepsilon} = A\chi_{\omega_{n,\varepsilon}} + B(1 - \chi_{\omega_{n,\varepsilon}}),$$

and  $u_{n,\varepsilon} \in H_0^1(Q)$  as the solution of

$$(5.41) \quad -\operatorname{div} M_{n,\varepsilon} \nabla u_{n,\varepsilon} = -\operatorname{div} \check{\sigma}_n \quad \text{in } Q.$$

Since  $|Z_n^k| = p_n^k$ , the sequence  $\omega_{n,\varepsilon}$  satisfies

$$(5.42) \quad \chi_{\omega_{n,\varepsilon}} \xrightarrow{*} \hat{\theta}_n \quad \text{in } L^\infty(\Omega), \quad \text{when } \varepsilon \rightarrow 0.$$

By using the fact that  $\nabla \hat{u}_n = \xi_n^k$  in each  $T_n^k$  and the definition (5.39) of  $\check{\sigma}_n$ , we deduce by periodic homogenization (see, e.g., [1], [4], [11]) that

$$(5.43) \quad u_{n,\varepsilon} \rightharpoonup \hat{u}_n \quad \text{in } H_0^1(Q), \quad \text{when } \varepsilon \rightarrow 0,$$

$$(5.44) \quad M_{n,\varepsilon} \nabla u_{n,\varepsilon} \rightharpoonup \check{\sigma}_n \quad \text{in } L^2(Q)^N, \quad \text{when } \varepsilon \rightarrow 0,$$

$$(5.45) \quad \nabla u_{n,\varepsilon} - \sum_{k=1}^{k_n} \left( \xi_n^k + \nabla w_n^k \left( \frac{x}{\varepsilon} \right) \right) \chi_{T_n^k} \rightarrow 0 \quad \text{in } L^2(Q)^N, \quad \text{when } \varepsilon \rightarrow 0.$$

By (5.45), (4.1),  $\nabla w_n^k$ -periodic, and (5.38), we then have

$$(5.46) \quad \lim_{\varepsilon \rightarrow 0} \int_\Omega F(\nabla u_{n,\varepsilon}) dx = \sum_{k=1}^{k_n} \lim_{\varepsilon \rightarrow 0} \int_{T_n^k \cap \Omega} F \left( \xi_n^k + \nabla w_n^k \left( \frac{x}{\varepsilon} \right) \right) dx$$

$$= \sum_{k=1}^{k_n} \int_{T_n^k \cap \Omega} \int_Y F(\xi_n^k + \nabla w_n^k) dy dx < \int_\Omega H_{\frac{2}{n}}(\nabla \hat{u}_n, \hat{\sigma}_n, \hat{\theta}_n) dx + \frac{|\Omega|}{n}.$$

We consider a dense countable subset  $\{h_j\}$  of  $L^1(\Omega)$  and a dense countable subset  $\{g_j\}$  of  $L^2(\Omega)^N$ . By using (5.42), (5.43), (5.46), for every  $n \in \mathbf{N}$ , we choose  $\varepsilon_n > 0$  such that

$$(5.47) \quad \left| \int_\Omega (\chi_{\omega_{n,\varepsilon_n}} - \hat{\theta}_n) h_j dx \right| < \frac{1}{n} \quad \forall j \in \{1, \dots, n\},$$

$$(5.48) \quad \left| \int_\Omega (M_{n,\varepsilon_n} \nabla u_{n,\varepsilon_n} - \check{\sigma}_n) g_j dx \right| < \frac{1}{n} \quad \forall j \in \{1, \dots, n\},$$

$$(5.49) \quad \int_\Omega |u_{n,\varepsilon_n} - \hat{u}_n|^2 dx < \frac{1}{n},$$



$$(5.50) \quad \int_{\Omega} F(\nabla u_{n,\varepsilon_n}) dx < \int_{\Omega} H_{\frac{2}{n}}(\nabla \hat{u}_n, \hat{\sigma}_n, \hat{\theta}_n) dx + \frac{|\Omega|}{n}.$$

Then we define

$$u_n = u_{n,\varepsilon_n}, \quad \sigma_n = M_{n,\varepsilon_n} \nabla u_{n,\varepsilon_n}, \quad \theta_n = \chi_{\omega_{n,\varepsilon_n}}.$$

Thanks to (5.40) and  $\hat{\sigma}_n$  converging strongly to  $\sigma$  in  $L^2(\Omega)^N$ , we have that  $\check{\sigma}_n$  converges strongly to  $\sigma$  in  $L^2(\Omega)^N$ . So, by (5.41),  $u_n$  is bounded in  $H^1(\Omega)$ , which joined to (5.49) and  $\hat{u}_n$  converging strongly to  $u$  in  $H^1(\Omega)$  implies that  $u_n$  converges weakly to  $u$  in  $H^1(\Omega)$ . Equation (5.41) and the strong convergence of  $\check{\sigma}_n$  (see Remark 4.5) also give that  $|\nabla u_n|^2$  is equi-integrable and that  $-\operatorname{div} \sigma_n = -\operatorname{div} \check{\sigma}_n$  converges strongly to  $-\operatorname{div} \sigma$  in  $H^{-1}(\Omega)$ . By (5.48), we also have that  $\sigma_n$  converges weakly to  $\sigma$  in  $L^2(\Omega)^N$ . Finally (5.47) and the convergence of  $\hat{\theta}_n$  to  $\theta$  in  $L^\infty(Q)$  weak-\* proves that  $\theta_n$  converges to  $\theta$  in  $L^\infty(\Omega)$  weak-\*. Hence,  $(u_n, \sigma_n, \theta_n)$   $\mathcal{T}$ -converges to  $(u, \sigma, \theta)$ .

By (5.50) and the definition (4.18) of  $\mathcal{F}$ , we obtain

$$(5.51) \quad \mathcal{F}(u_n, \sigma_n, \theta_n) = \int_{\Omega} F(\nabla u_n) dx < \int_{\Omega} H_{\frac{2}{n}}(\nabla \hat{u}_n, \hat{\sigma}_n, \hat{\theta}_n) dx + \frac{|\Omega|}{n}.$$

On the other hand, by (5.36),  $H_\delta(\xi, \eta, p)$  decreasing in  $\delta$ , and (5.9), we have

$$\begin{aligned} H_{\frac{2}{n}}(\nabla \hat{u}_n, \hat{\sigma}_n, \hat{\theta}_n) &\leq H_{\frac{1}{n} + \lambda_n(x) + |\hat{\sigma}_n - \sigma|}(\nabla \hat{u}_n, \hat{\sigma}_n, \hat{\theta}_n) \\ &\leq H_{\frac{1}{n}}(\nabla u, \sigma, \theta) + (1 + |\nabla u| + |\nabla \hat{u}_n|)^{2-e} (\lambda_n)^e \\ &\leq H_{\frac{1}{n}}(\nabla u, \sigma, \theta) + (1 + |\nabla u| + |\nabla \hat{u}_n|)^{2-e} \frac{1}{n^e}, \end{aligned}$$

a.e. in  $Q_n$ , with

$$\lambda_n = C(|\nabla(\hat{u}_n - u)| + |\nabla u| |\hat{\theta}_n - \theta|^\rho).$$

By then using (5.10), we get

$$\begin{aligned} (5.52) \quad &\int_{\Omega} H_{\frac{2}{n}}(\nabla \hat{u}_n, \hat{\sigma}_n, \hat{\theta}_n) dx \\ &\leq \int_{Q_n \cap \Omega} H_{\frac{1}{n}}(\nabla u, \sigma, \theta) dx + \frac{1}{n^e} \int_{Q_n \cap \Omega} (1 + |\nabla u| + |\nabla \hat{u}_n|)^{2-e} dx \\ &\quad + L \int_{\Omega \setminus Q_n} \left( \frac{\beta}{\alpha} |\nabla \hat{u}_n| \right)^e \left( 1 + \frac{\beta}{\alpha} |\nabla \hat{u}_n| \right)^{2-e} dx. \end{aligned}$$

By the definition of  $H$  and (5.10) we can apply the Lebesgue dominated convergence theorem to deduce that

$$H_{\frac{1}{n}}(\nabla u, \sigma, \theta) \rightarrow H(\nabla u, \sigma, \theta) \quad \text{in } L^1(\Omega).$$

By using also the fact that  $\nabla \hat{u}_n$  converges strongly in  $L^2(\Omega)^N$  and that  $|\Omega \setminus Q_n|$  tends to zero, we then deduce by (5.51) and (5.52) that

$$\limsup_{n \rightarrow \infty} \mathcal{F}(u_n, \sigma_n, \theta_n) \leq \int_{\Omega} H(\nabla u, \sigma, \theta) dx = \overline{\mathcal{F}}(u, \sigma, \theta).$$

This proves (5.24).  $\square$

*Proof of Theorem 4.3.* We denote by  $I$  the infimum of problem (4.4) and by  $J$  the infimum of problem (4.16).

*Step 1.* Let us first prove that  $I$  is bigger than or equal to  $J$ . For this purpose it is enough to observe that thanks to (4.12)

$$\int_{\Omega} F(\nabla u) dx = \int_{\Omega} H(\nabla u, M\nabla u, \theta) dx,$$

when  $\theta = \chi_{\omega}$  with  $\omega \subset \Omega$  measurable and  $M = A\chi_{\omega} + B\chi_{\Omega \setminus \omega}$ . So in (4.4) we are minimizing the same functional as that in (4.16) but in a smaller set. This proves  $I \geq J$ .

*Step 2.* Let us now use the direct method of the calculus of variations to prove that problem (4.16) has a minimum. We consider  $\theta_n \in L^{\infty}(\Omega)$ , with  $0 \leq \theta_n \leq 1$  a.e. in  $\Omega$ ,

$$\int_{\Omega} \theta_n dx \leq \kappa |\Omega|,$$

and  $M_n \in \mathcal{K}(A, B, \theta)$  a.e. in  $\Omega$  such that the solution  $u_n$  of

$$\begin{cases} -\operatorname{div} M_n \nabla u_n = f & \text{in } \Omega, \\ u_n \in H_0^1(\Omega) \end{cases}$$

satisfies

$$\exists \lim_{n \rightarrow \infty} \left( \int_{\Omega} H(\nabla u_n, M_n \nabla u_n, \theta_n) dx + G(u_n) \right) = J.$$

Thanks to  $\theta_n$  being bounded in  $L^{\infty}(\Omega)$  and the compactness of the  $H$ -convergence, by extracting a subsequence if necessary, we can assume that there exist  $\theta \in L^{\infty}(\Omega)$  and  $M \in L^{\infty}(\Omega; \mathcal{M}_N^s)$  such that  $\theta_n$  converges weak-\* in  $L^{\infty}(\Omega)$  to  $\theta$  and  $M_n$   $H$ -converges to  $M$ . Therefore,  $(u_n, M_n \nabla u_n, \theta_n)$   $\mathcal{T}$ -converges to  $(u, M \nabla u, \theta)$  (the equi-integrability of  $|\nabla u_n|^2$  is an easy consequence of Meyer's theorem [19]), where

$$0 \leq \theta \leq 1 \quad \text{a.e. in } \Omega, \quad \int_{\Omega} \theta dx \leq \kappa |\Omega|, \quad M \in \mathcal{K}(A, B, \theta) \quad \text{a.e. in } \Omega,$$

$$\begin{cases} -\operatorname{div} M \nabla u = f & \text{in } \Omega, \\ u \in H_0^1(\Omega). \end{cases}$$

Since  $\overline{\mathcal{F}}$  is lower semicontinuous for the  $\mathcal{T}$ -convergence and  $G$  is sequentially continuous for the weak convergence in  $H_0^1(\Omega)$ , we have

$$J = \lim_{n \rightarrow \infty} \left( \int_{\Omega} H(\nabla u_n, M_n \nabla u_n, \theta_n) dx + G(u_n) \right) \geq \int_{\Omega} H(\nabla u, M \nabla u, \theta) dx + G(u).$$

Thus  $u, M, \theta$  is a solution of (4.16), and hence  $J$  is a minimum.

*Step 3.* To finish the proof of Theorem 4.3, let us now prove that, for every solution  $u, M, \theta$  of (4.16), there exists a sequence  $\omega_n$  of measurable subsets of  $\Omega$ , with  $|\omega_n| \leq \kappa |\Omega|$  such that the solution  $u_n$  of

$$(5.53) \quad \begin{cases} -\operatorname{div} (A\chi_{\omega_n} + B\chi_{\Omega \setminus \omega_n}) \nabla u_n = f & \text{in } \Omega, \\ u_n \in H_0^1(\Omega) \end{cases}$$

satisfies

$$(5.54) \quad \lim_{n \rightarrow \infty} \left( \int_{\Omega} F(\nabla u_n) dx + G(u_n) \right) = J,$$

and it is such that  $(u_n, (A\chi_{\omega_n} + B\chi_{\Omega \setminus \omega_n})\nabla u_n, \chi_{\omega_n})$   $\mathcal{T}$ -converges to  $(u, M\nabla u, \theta)$ .

By Theorem 4.5, we know that there exist  $\tilde{\omega}_n \subset \Omega$  measurable and  $\tilde{u}_n \in H_0^1(\Omega)$  such that, for  $\tilde{M}_n = (A\chi_{\tilde{\omega}_n} + B\chi_{\Omega \setminus \tilde{\omega}_n})$ , the sequence  $(\tilde{u}_n, \tilde{M}_n \nabla \tilde{u}_n, \chi_{\tilde{\omega}_n})$   $\mathcal{T}$ -converges to  $(u, M\nabla u, \theta)$  and

$$(5.55) \quad \lim_{n \rightarrow \infty} \int_{\Omega} F(\nabla \tilde{u}_n) dx = \int_{\Omega} H(\nabla u, M\nabla u, \theta) dx.$$

From the compactness of the  $H$ -convergence, we can also assume that there exists  $\hat{M}$  such that  $\tilde{M}_n$   $H$ -converges to  $\hat{M}$ . Then  $\hat{M} \in \mathcal{K}(A, B, \theta)$  a.e. in  $\Omega$  and, thanks to the definition of  $\mathcal{T}$ -convergence,  $\hat{M}\nabla u = M\nabla u$  a.e. in  $\Omega$ . The weak- $*$  convergence of  $\chi_{\tilde{\omega}_n}$  to  $\theta$  also implies that

$$\lim_{n \rightarrow \infty} |\tilde{\omega}_n| = \int_{\Omega} \theta dx \leq \kappa |\Omega|.$$

Now, for every  $n \in \mathbb{N}$ , we consider  $\omega_n \subset \tilde{\omega}_n$  measurable such that

$$|\tilde{\omega}_n \setminus \omega_n| = \max \{ |\tilde{\omega}_n| - \kappa |\Omega|, 0 \} \rightarrow 0,$$

and we define  $u_n$  as the solution of (5.53).

By taking into account

$$-\operatorname{div} \tilde{M}_n \nabla u_n = f - \operatorname{div} (\tilde{M}_n - M_n) \nabla u_n \quad \text{in } \Omega,$$

with  $M_n = A\chi_{\omega_n} + B\chi_{\Omega \setminus \omega_n}$ , the equi-integrability of  $|\nabla u_n|^2$ , and the fact that

$$(5.56) \quad \lim_{n \rightarrow \infty} \left| \{x \in \Omega : \tilde{M}_n(x) - M_n(x) \neq 0\} \right| = \lim_{n \rightarrow \infty} |\tilde{\omega}_n \setminus \omega_n| = 0,$$

we have that  $(\tilde{M}_n - M_n)\nabla u_n$  tends to zero in  $L^2(\Omega)^N$  strongly. Therefore  $-\operatorname{div} \tilde{M}_n \nabla u_n$  tends to  $f$  strongly in  $H^{-1}(\Omega)$ . By the definition of  $\mathcal{T}$ -convergence and  $-\operatorname{div} M\nabla u = f$  in  $\Omega$ , we also have that  $-\operatorname{div} \tilde{M}_n \nabla \tilde{u}_n$  converges strongly to  $f$  in  $H^{-1}(\Omega)$ . The div-curl lemma then gives that  $\tilde{M}_n \nabla(u_n - \tilde{u}_n) \cdot \nabla(u_n - \tilde{u}_n)$  converges to zero in the sense of the distributions. The equi-integrability of  $|\nabla u_n|^2$  and  $|\nabla \tilde{u}_n|^2$  implies that this convergence holds in fact in  $L^1(\Omega)$  weakly, and so

$$\lim_{n \rightarrow \infty} \int_{\Omega} \tilde{M}_n \nabla(u_n - \tilde{u}_n) \cdot \nabla(u_n - \tilde{u}_n) dx = 0,$$

which by the ellipticity of  $\tilde{M}_n$  implies that  $\nabla(u_n - \tilde{u}_n)$  converges strongly to zero in  $L^2(\Omega)^N$ . Thus,  $(u_n, M_n \nabla u_n, \chi_{\omega_n})$   $\mathcal{T}$ -converges to  $(u, M\nabla u, \theta)$ , and, by (4.1), the sequential continuity of  $G$  with respect to the weak topology in  $H_0^1(\Omega)$ , and (5.55), we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \left( \int_{\Omega} F(\nabla u_n) dx + G(u_n) \right) &= \lim_{n \rightarrow \infty} \left( \int_{\Omega} F(\nabla \tilde{u}_n) dx + G(u_n) \right) \\ &= \int_{\Omega} H(\nabla u, M\nabla u, \theta) dx + G(u) = J. \quad \square \end{aligned}$$

*Proof of Proposition 4.6.* We consider  $\lambda, p_1, p_2 \in [0, 1]$ ,  $\xi_1, \xi_2, \eta_1, \eta_2 \in \mathbf{R}^N$  such that  $\eta_1 = \eta_2$  for  $N = 1$ , and  $(\xi_1 - \xi_2) \cdot (\eta_1 - \eta_2) = 0$  for  $N > 1$ .

We take  $\chi = \sum_{k \in \mathbf{Z}} \chi_{(k, k+\lambda)} \in L^\infty_\#(0, 1)$ .

If  $\xi_1 \neq \xi_2$ , we define  $u_n \in W^{1,\infty}_{loc}(\mathbf{R}^N)$ ,  $\sigma_n \in L^\infty(\mathbf{R}^N)^N$ , and  $\theta_n \in L^\infty(\mathbf{R}^N)$  by

$$u_n(x) = \xi_2 \cdot x + \frac{1}{n} \int_0^{n(\xi_1 - \xi_2) \cdot x} \chi(s) ds, \quad \sigma_n(x) = \eta_2 + (\eta_1 - \eta_2) \chi(n(\xi_1 - \xi_2) \cdot x),$$

$$\theta_n(x) = p_2 + (p_1 - p_2) \chi(n(\xi_1 - \xi_2) \cdot x), \quad \text{a.e. } x \in \mathbf{R}^N,$$

respectively, and we observe that, by defining

$$u(x) = (\lambda \xi_1 + (1 - \lambda) \xi_2) \cdot x, \quad \sigma(x) = \lambda \eta_1 + (1 - \lambda) \eta_2,$$

$$\theta(x) = \lambda p_1 + (1 - \lambda) p_2, \quad \text{a.e. } x \in \mathbf{R}^N,$$

we have that, for every smooth bounded open set  $\omega \subset \mathbf{R}^N$ ,  $u_n$  converges to  $u$  in  $W^{1,\infty}(\omega)$  weak-\*,  $\sigma_n$  converges to  $\sigma$  in  $L^\infty(\omega)^N$  weak-\*, and  $\theta_n$  converges to  $\theta$  in  $L^\infty(\omega)$  weak-\* and  $\operatorname{div} \sigma_n = 0$  in  $\omega$ . In particular,  $(u_n, \sigma_n, \theta_n)$   $\mathcal{T}$ -converges to  $(u, \sigma, \theta)$ . Since the functional  $\overline{\mathcal{F}}$  defined by (4.19) is lower semicontinuous in  $\omega$  for the  $\mathcal{T}$ -convergence (use Theorem 4.5, with  $\Omega$  replaced by  $\omega$ ), we get

$$\begin{aligned} & \int_\omega H(\lambda \xi_1 + (1 - \lambda) \xi_2, \lambda \eta_1 + (1 - \lambda) \eta_2, \lambda p_1 + (1 - \lambda) p_2) dx \\ &= \overline{\mathcal{F}}(u, \sigma, \theta) \leq \lim_{n \rightarrow \infty} \overline{\mathcal{F}}(u_n, \sigma_n, \theta_n) \\ &= \lim_{n \rightarrow \infty} \int_\omega \left( H(\xi_1, \eta_1, p_1) \chi(n(\xi_1 - \xi_2) \cdot x) + H(\xi_2, \eta_2, p_2) (1 - \chi(n(\xi_1 - \xi_2) \cdot x)) \right) dx \\ &= \int_\omega \left( \lambda H(\xi_1, \eta_1, p_1) + (1 - \lambda) H(\xi_2, \eta_2, p_2) \right) dx, \end{aligned}$$

which by the arbitrariness of  $\omega$  implies (4.22) (or (4.21) if  $N = 1$ ).

If  $\xi_1 = \xi_2$  (and then  $N > 1$ ), we reason analogously, by taking  $\zeta \neq 0$  orthogonal to  $\eta_1 - \eta_2$  and defining

$$u_n(x) = \xi_1 \cdot x, \quad \sigma_n(x) = \eta_2 + (\eta_1 - \eta_2) \chi(n\zeta \cdot x),$$

$$\theta_n(x) = p_2 + (p_1 - p_2) \chi(n\zeta \cdot x), \quad \text{a.e. } x \in \mathbf{R}^N. \quad \square$$

*Proof of Proposition 4.7.* We recall that convex and bounded functions are locally Lipschitz in the interior of their domain. This can be easily generalized to functions of several variables, convex in each one of these variables. Then Proposition 4.6 and (4.11) prove that  $H$  is continuous in the interior of its domain. On the other hand, we consider  $(\xi, \eta, p) \in \operatorname{Dom}(H)$ ,  $p \in (0, 1)$ . By reasoning similarly as in the proof of Theorem 4.2, it is not difficult to show that if  $p$  is close to 0 or 1, then  $H(\xi, \eta, p)$  is close to  $F(\xi)$  and that if

$$\xi \cdot (\Lambda_p \xi - \eta) - \left( (A - B) \left( \frac{A}{p} + \frac{B}{1-p} \right)^{-1} (A - B) \right)^\dagger (\Lambda_p \xi - \eta) \cdot (\Lambda_p \xi - \eta)$$

is small, then

$$\left| H(\xi, \eta, p) - pF \left( \xi + \Xi \left( \frac{\Lambda_p \xi - \eta}{p} \right) \right) - (1-p)F \left( \xi - \Xi \left( \frac{\Lambda_p \xi - \eta}{1-p} \right) \right) \right|$$

is small. This proves the continuity of  $H$  in the boundary of its domain.  $\square$

**6. Example.** In this section we give an example in dimension  $N \geq 2$  of a quadratic function  $F$  for which we obtain an explicit expression of the function  $H$  defined by (4.6) in its whole domain. This is given by the following theorem.

**THEOREM 6.1.** *We consider  $N \geq 2$ ,  $A, B \in \mathcal{M}_N^s$  definite positive,  $s \in \mathbf{R}$  such that  $s(A - B)$  is definite nonnegative, and*

$$F(\xi) = sA\xi \cdot \xi \quad \forall \xi \in \mathbf{R}^N.$$

*Then, for every  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times (0, 1)$  such that  $\eta \in \mathcal{K}(A, B, p)\xi$ , we have that the function  $H$  defined by (4.6) is given by*

$$(6.1) \quad H(\xi, \eta, p) = sA\xi \cdot \xi + s(\Lambda_p \xi - \eta) \cdot \xi + \frac{s}{1-p}(A - B)\nu \cdot \nu,$$

*where  $\nu$  is any vector in  $\mathbf{R}^N$  which satisfies  $(B - A)\nu = \Lambda_p \xi - \eta$ .*

*In the cases  $p = 0$ ,  $p = 1$ , we have*

$$H(\xi, A\xi, 1) = H(\xi, B\xi, 0) = sA\xi \cdot \xi \quad \forall \xi \in \mathbf{R}^N.$$

*Remark 6.1.* Since  $\text{Ran}(A - B) = \text{Ker}(A - B)^\perp$ , it is clear that the second member of (6.1) does not depend on the choice of  $\nu$ . By taking  $\nu = (A - B)^\dagger(\Lambda_p \xi - \eta)$ , we have

$$H(\xi, \eta, p) = sA\xi \cdot \xi + s(\Lambda_p \xi - \eta) \cdot \xi + \frac{s}{1-p}(A - B)^\dagger(\Lambda_p \xi - \eta) \cdot (\Lambda_p \xi - \eta)$$

for every  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times (0, 1)$  such that  $\eta \in \mathcal{K}(A, B, p)\xi$ .

*Proof of Theorem 6.1.* By (4.12) it is enough to prove (6.1).

For  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times (0, 1)$  such that  $\eta \in \mathcal{K}(A, B, p)\xi$ , we consider  $Z_n$ ,  $S_n$ ,  $w_n$ ,  $\eta_n$  as in Remark 4.1, and we define

$$\nu_n = \int_{Z_n} \nabla w_n dy = - \int_{Y \setminus Z_n} \nabla w_n dy.$$

By using that  $\nabla w_n$  has mean value zero and (3.4), we have

$$(6.2) \quad \begin{aligned} \int_Y sA(\xi + \nabla w_n) \cdot (\xi + \nabla w_n) dy &= sA\xi \cdot \xi + \int_Y sA \nabla w_n \cdot \nabla w_n dy \\ &= sA\xi \cdot \xi + s(\Lambda_p \xi - \eta_n) \cdot \xi + \int_{Y \setminus Z_n} s(A - B) \nabla w_n \cdot \nabla w_n dy. \end{aligned}$$

Thanks to  $s(A - B)$  nonnegative, the Cauchy-Schwarz inequality proves that

$$\begin{aligned} s(A - B)\nu_n \cdot \nu_n &= - \int_{Y \setminus Z_n} s(A - B) \nabla w_n \cdot \nu_n dy \\ &\leq \int_{Y \setminus Z_n} (s(A - B) \nabla w_n \cdot \nabla w_n)^{\frac{1}{2}} (s(A - B)\nu_n \cdot \nu_n)^{\frac{1}{2}} dy \\ &\leq \left( \int_{Y \setminus Z_n} s(A - B) \nabla w_n \cdot \nabla w_n dy \right)^{\frac{1}{2}} \left( (1 - p)s(A - B)\nu_n \cdot \nu_n \right)^{\frac{1}{2}} \end{aligned}$$

and so

$$\frac{s}{1-p}(A - B)\nu_n \cdot \nu_n \leq \int_{Y \setminus Z_n} s(A - B) \nabla w_n \cdot \nabla w_n dy,$$

which substituted in (6.2) and then passed to the limit in  $n$  proves that

$$(6.3) \quad H(\xi, \eta, p) \geq sA\xi \cdot \xi + s(\Lambda_p \xi - \eta) \cdot \xi + \frac{s}{1-p}(A-B)\nu \cdot \nu,$$

where  $\nu$  equals the limit of  $\nu_n$  (it exists for a subsequence) and satisfies  $(A-B)\nu = \eta - \Lambda_p \xi$ .

In order to prove the contrary inequality, we remark that if  $\eta \in \Lambda_p \xi + \text{Ran}(A-B)$  is such that

$$(6.4) \quad \left( (A-B) \left( \frac{A}{p} + \frac{B}{1-p} \right)^{-1} (A-B) \right)^\dagger (\Lambda_p \xi - \eta) \cdot (\Lambda_p \xi - \eta) = \xi \cdot (\Lambda_p \xi - \eta),$$

then (6.1) follows from Theorem 4.2 by taking  $\nu = \Xi(\Lambda_p \xi - \eta)$  and using

$$\left( \frac{A}{p} + \frac{B}{1-p} \right) \nu \cdot \nu = (\Lambda_p \xi - \eta) \cdot \xi.$$

Thus, it is enough to consider the case where  $\eta \in \Lambda_p \xi + \text{Ran}(A-B)$  satisfies

$$(6.5) \quad \left( (A-B) \left( \frac{A}{p} + \frac{B}{1-p} \right)^{-1} (A-B) \right)^\dagger (\Lambda_p \xi - \eta) \cdot (\Lambda_p \xi - \eta) < \xi \cdot (\Lambda_p \xi - \eta).$$

By taking as above  $\nu = \Xi(\Lambda_p \xi - \eta)$ , we have

$$\eta = \Lambda_p \xi + (A-B)\nu, \quad \left( \frac{A}{p} + \frac{B}{1-p} \right) \nu \cdot \nu < (B-A)\nu \cdot \xi.$$

The last inequality implies that there exists  $r \in (0, p)$  such that

$$\left( \frac{1}{p-r} - \frac{1}{p} \right) A\nu \cdot \nu = (B-A)\nu \cdot \xi - \left( \frac{A}{p} + \frac{B}{1-p} \right) \nu \cdot \nu.$$

For  $\hat{\nu} = \nu/(1-r)$  we then have that  $\hat{\eta} = \Lambda_{\hat{p}} \xi + (A-B)\hat{\nu}$ ,  $\hat{p} = (p-r)/(1-r)$  are such that  $\hat{\eta} \in \Lambda_{\hat{p}} \xi + \text{Ran}(A-B)$  and that (6.4) holds with  $\eta$  and  $p$ , respectively, replaced by  $\hat{\eta}$  and  $\hat{p}$ . By taking into account

$$p = r + (1-r)\hat{p}, \quad \eta = rA\xi + (1-r)\hat{\eta}$$

and using the convexity property (4.22) of  $H$ , we then obtain

$$(6.6) \quad \begin{aligned} H(\xi, \eta, p) &\leq rH(\xi, A\xi, 1) + (1-r)H(\xi, \hat{\eta}, \hat{p}) \\ &= sA\xi \cdot \xi + s(A-B)\nu \cdot \xi + \frac{s}{1-p}(A-B)\nu \cdot \nu. \end{aligned}$$

This finishes the proof of Theorem 6.1.  $\square$

*Remark 6.2.* By the proof of Theorem 6.1, we see that the sequence  $Z_n$  given by Remark 4.1 must be chosen in such a way that the value of  $\nabla w_n$  in  $Y \setminus Z_n$  is close to a constant. For  $N \geq 2$ ,  $(\xi, \eta, p) \in \mathbf{R}^N \times \mathbf{R}^N \times (0, 1)$ ,  $\eta \in \mathcal{K}(A, B, p)\xi$ ,  $\eta \neq \Lambda_p \xi$ , this can be carried on by using two laminations as follows: We take  $\nu = \Xi(\Lambda_p \xi - \eta)$  and  $r$  and  $\hat{p}$  as in the proof of Theorem 6.1, and then we construct a matrix  $M$  by a lamination of  $A$  and  $B$  in the direction of  $\nu$  with respective proportions  $\hat{p}$  and  $1-\hat{p}$  and

then a lamination of  $A$  and  $M$  in an orthogonal direction to  $(M - A)\xi$  with respective proportions  $r$  and  $1 - r$ . This is the idea that we used in the proof of inequality (6.6).

As a particular case of Theorem 6.1, we have the following result.

**COROLLARY 6.2.** *Assume that  $N \geq 2$ ,  $A = \alpha I$ ,  $B = \beta I$ , with  $0 < \alpha < \beta$ , and for  $s \in \mathbf{R}$  take  $F(\xi) = s|\xi|^2$  for every  $\xi \in \mathbf{R}^N$ . Then the function  $H$  defined by (4.6) satisfies*

$$\begin{aligned} H(\xi, \eta, p) &= s|\xi|^2 + \frac{s}{\beta}(\Lambda_p \xi - \eta) \cdot \xi + \frac{s}{\beta(\beta - \alpha)p} |\Lambda_p \xi - \eta|^2 & \text{if } s \geq 0, \\ H(\xi, \eta, p) &= s|\xi|^2 + \frac{s}{\alpha}(\Lambda_p \xi - \eta) \cdot \xi - \frac{s}{\alpha(\beta - \alpha)(1 - p)} |\Lambda_p \xi - \eta|^2 & \text{if } s \leq 0 \end{aligned}$$

for every  $\xi \in \mathbf{R}^N$ ,  $p \in (0, 1)$ ,  $\eta \in \mathcal{K}(A, B, p)\xi$ .

*Proof.* It is enough to apply Theorem 6.1 where the matrices  $A$  and  $B$  and the constant  $s$  must be replaced, respectively, by  $\beta I$ ,  $\alpha I$ , and  $s/\beta$  in the case  $s \geq 0$  and by  $\alpha I$ ,  $\beta I$ , and  $s/\alpha$  in the case  $s \leq 0$ .  $\square$

**Remark 6.3.** Corollary 6.2 and Theorem 4.3 give the relaxation of the problem

$$\begin{cases} \inf \left\{ s \int_{\Omega} |\nabla u|^2 dx \right\}, \\ -\operatorname{div}(\alpha \chi_{\omega} + \beta \chi_{\Omega \setminus \omega}) \nabla u = f \quad \text{in } \Omega, \\ u \in H_0^1(\Omega), \\ \omega \subset \Omega \text{ measurable, } |\omega| \leq \kappa |\Omega|. \end{cases}$$

In the case where  $s > 0$ , this relaxation was obtained by Bellido and Pedregal in [3] (see also [15]) for  $N = 2$  and independently by Grabovsky in [10] for arbitrary  $N$ . The method used in the present paper is different from the ones used by these authors.

## REFERENCES

- [1] G. ALLAIRE, *Shape Optimization by the Homogenization Method*, Appl. Math. Sci. 146, Springer-Verlag, New York, 2002.
- [2] G. ALLAIRE AND S. GUTIÉRREZ, *Optimal design in small amplitude homogenization*, M2AN Math. Model. Numer. Anal., 41 (2007), pp. 543–574.
- [3] J. C. BELLIDO AND P. PEDREGAL, *Explicit quasiconvexification for some cost functionals depending on derivatives of the state in optimal designing*, Discrete Contin. Dyn. Syst., 8 (2002), pp. 967–982.
- [4] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North-Holland, Amsterdam, 1978.
- [5] J. CASADO-DÍAZ, J. COUCE-CALVO, AND J. D. MARTÍN-GÓMEZ, *Optimality conditions for non-convex multistate control problems in the coefficients*, SIAM J. Control Optim., 43 (2004), pp. 216–239.
- [6] J. CASADO-DÍAZ, J. COUCE-CALVO, AND J. D. MARTÍN-GÓMEZ, *A density result for the variation of a material with respect to small inclusions*, C. R. Math. Acad. Sci. Paris, 342 (2006), pp. 353–358.
- [7] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, New York, 1977.
- [8] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [9] G. DAL MASO AND R. V. KOHN, *The Local Character of  $G$ -closure*, manuscript.
- [10] Y. GRABOVSKY, *Optimal design for two-phase conducting composites with weakly discontinuous objective functionals*, Adv. Appl. Math., 27 (2001), pp. 683–704.
- [11] V. V. JIKOV, S. M. KOZLOV, AND O. A. OLEINIK, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.

- [12] R. LIPTON, *Configurations of nonlinear materials with electric fields that minimize  $L^p$  norms*, Phys. B, 338 (2003), pp. 48–53.
- [13] R. LIPTON, *Relaxation through homogenization for optimal design problems with gradient constraint*, J. Optim. Theory Appl., 114 (2002), pp. 27–53.
- [14] R. LIPTON, *Stress constrained  $G$  closure and relaxation of structural design problems*, Quart. Appl. Math., 62 (2004), pp. 295–321.
- [15] R. LIPTON AND A. P. VELO, *Optimal design of gradient fields with applications to electrostatics*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. XIV, Stud. Math. Appl. 31, D. Cioranescu and J. L. Lions, eds., North-Holland, Amsterdam, 2002, pp. 509–532.
- [16] K. A. LURIE, *Applied Optimal Control Theory of Distributed Systems*, Plenum Press, New York, 1993.
- [17] K. A. LURIE AND A. V. CHERKAEV, *Exact estimates of the conductivity of a binary mixture of isotropic materials*, Proc. Roy. Soc. Edinburgh Sect. A, 104 (1986), pp. 21–38.
- [18] F. MAESTRE AND P. PEDREGAL, *Quasiconvexification in 3-D for a variational reformulation of an optimal design problem in conductivity*, Nonlinear Anal., 64 (2006), pp. 1962–1976.
- [19] N. G. MEYERS, *An  $L^p$ -estimate for the gradient of solutions of second order elliptic divergence equations*, Ann. Scuola Norm. Sup. Pisa (3), 17 (1963), pp. 189–206.
- [20] F. MURAT, *Un contre-exemple pour le problème du contrôle dans les coefficients*, C. R. Acad. Sci. Paris Sér. A-B, 273 (1971), pp. A708–A711.
- [21] F. MURAT, *Théorèmes de non-existence pour des problèmes de contrôle dans les coefficients*, C. R. Acad. Sci. Paris Sér. A, 274 (1972), pp. A395–A398.
- [22] F. MURAT AND L. TARTAR,  *$H$ -convergence*, in Topics in the Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, L. Cherkhev and R. V. Kohn, eds., Birkhäuser, Boston, 1998, pp. 21–43.
- [23] F. MURAT, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 5 (1978), pp. 489–507.
- [24] F. MURAT AND L. TARTAR, *Calculus of variations and homogenization*, in Topics in the Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, L. Cherkhev and R. V. Kohn, eds., Birkhäuser, Boston, 1998, pp. 139–174.
- [25] P. PEDREGAL, *Optimal design in two-dimensional conductivity for a general cost depending on the field*, Arch. Ration. Mech. Anal., 182 (2006), pp. 367–385.
- [26] S. SPAGNOLO, *Sulla convergenza di soluzioni di equazioni paraboliche ed ellittiche*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3), 22 (1968), pp. 571–597.
- [27] L. TARTAR, *Cours Peccot*, Collège de France, Paris, 1977; part of this work appears in [22].
- [28] L. TARTAR, *Compensated compactness and applications to partial differential equations, nonlinear analysis and mechanics*, in Heriot-Watt Symposium IV, Res. Notes Math. 39, R. J. Knops, ed., Pitman, San Francisco, 1979, pp. 136–212.
- [29] L. TARTAR, *Estimations fines de coefficients homogénéisés*, in Ennio de Giorgi Colloquium (Paris, 1983), Res. Notes Math. 125, P. Kree, ed., Pitman, London, 1985, pp. 168–187.
- [30] L. TARTAR, *Remarks on optimal design problems*, in Calculus of Variations, Homogenization and Continuum Mechanics, Adv. Math. Appl. Sci. 18, G. Buttazzo, G. Bouchitte, and P. Suquet, eds., World Scientific, Singapore, 1994, pp. 279–296.
- [31] P. VELO, *Optimal Design of Gradient Fields and Currents*, Ph.D. thesis, WPI, Worcester, MA, 2000; part of this work appears in [15].



# DISSIPATIVE BOUNDARY CONDITIONS FOR ONE-DIMENSIONAL NONLINEAR HYPERBOLIC SYSTEMS\*

JEAN-MICHEL CORON<sup>†</sup>, GEORGES BASTIN<sup>‡</sup>, AND BRIGITTE D'ANDRÉA-NOVEL<sup>§</sup>

**Abstract.** We give a new sufficient condition on the boundary conditions for the exponential stability of one-dimensional nonlinear hyperbolic systems on a bounded interval. Our proof relies on the construction of an explicit strict Lyapunov function. We compare our sufficient condition with other known sufficient conditions for nonlinear and linear one-dimensional hyperbolic systems.

**Key words.** nonlinear hyperbolic systems, boundary conditions, stability, Lyapunov function

**AMS subject classifications.** 35F30, 35F25, 93D20, 93D30

**DOI.** 10.1137/070706847

**1. Introduction.** We are concerned with the following one-dimensional  $n \times n$  nonlinear hyperbolic system:

$$(1.1) \quad u_t + F(u)u_x = 0, \quad x \in [0, 1], \quad t \in [0, +\infty),$$

where  $u : [0, \infty) \times [0, 1] \rightarrow \mathbb{R}^n$  and  $F : \mathbb{R}^n \rightarrow \mathcal{M}_{n,n}(\mathbb{R})$ ,  $\mathcal{M}_{n,n}(\mathbb{R})$  denoting, as usual, the set of  $n \times n$  real matrices. We consider the case where, possibly after an appropriate state transformation,  $F(0)$  is a diagonal matrix with distinct and nonzero eigenvalues:

$$(1.2) \quad \begin{aligned} F(0) &:= \text{diag} (\Lambda_1, \Lambda_2, \dots, \Lambda_n), \\ \Lambda_i &> 0 \quad \forall i \in \{1, \dots, m\}, \\ \Lambda_i &< 0 \quad \forall i \in \{m+1, \dots, n\}, \end{aligned}$$

$$(1.3) \quad \Lambda_i \neq \Lambda_j \quad \forall (i, j) \in \{1, \dots, n\}^2 \text{ such that } i \neq j.$$

In (1) and in what follows,  $\text{diag} (\Lambda_1, \Lambda_2, \dots, \Lambda_n)$  denotes the diagonal matrix whose  $i$ th element on the diagonal is  $\Lambda_i$ .

Our concern is to analyze the asymptotic behavior of the classical solutions of the system under the following boundary condition:

$$(1.4) \quad \begin{pmatrix} u_+(t, 0) \\ u_-(t, 1) \end{pmatrix} = G \begin{pmatrix} u_+(t, 1) \\ u_-(t, 0) \end{pmatrix}, \quad t \in [0, +\infty),$$

where the map  $G : \mathbb{R}^n \rightarrow \mathbb{R}^n$  vanishes at 0, while  $u_+ \in \mathbb{R}^m$ ,  $u_- \in \mathbb{R}^{n-m}$  are defined by requiring that  $u := (u_+^{\text{tr}}, u_-^{\text{tr}})^{\text{tr}}$ . The problem is to find the map  $G$  such that

\*Received by the editors October 30, 2007; accepted for publication (in revised form) January 16, 2008; published electronically May 7, 2008.

<http://www.siam.org/journals/sicon/47-3/70684.html>

<sup>†</sup>Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie and Institut Universitaire de France, B.C. 187, 4 place Jussieu, 75252 Paris Cedex 05, France (coron@ann.jussieu.fr). This author's research was partially supported by the "Agence Nationale de la Recherche" (ANR), Project C-QUID, number BLAN-3-139579.

<sup>‡</sup>Center for Systems Engineering and Applied Mechanics (CESAME), Université Catholique de Louvain, 4 Avenue G. Lemaître, 1348 Louvain-la-Neuve, Belgium (Georges.Bastin@uclouvain.be).

<sup>§</sup>Centre de Robotique, École des Mines de Paris, 60 boulevard Saint Michel, 75272 Paris Cedex 06, France (brigitte.dandrea-novel@ensmp.fr).

the boundary condition (1.4) is dissipative, i.e., implies that the equilibrium solution  $u \equiv 0$  of system (1.1) with the boundary condition (1.4) is exponentially stable.

This problem has been considered in the literature for more than 20 years. To our knowledge, the first results were published by Slemrod in [21] and Greenberg and Li in [9] for the special case of  $2 \times 2$  (i.e.,  $u \in \mathbb{R}^2$ ) systems. A generalization to  $n \times n$  systems was given by the Li school. Let us mention in particular [17] by Qin, [25] by Zhao, and [14, Theorem 1.3, page 173] by Li. All these results rely on a systematic use of direct estimates of the solutions and their derivatives along the characteristic curves. They give rise to sufficient dissipative boundary conditions which are kinds of “small gain conditions.” The weakest sufficient condition [14, Theorem 1.3, page 173] is formulated as follows:  $\rho(|G'(0)|) < 1$ , where  $\rho(A)$  denotes the spectral radius of  $A \in \mathcal{M}_{n,n}(\mathbb{R})$  and  $|A|$  denotes the matrix whose elements are the absolute values of the elements of  $A \in \mathcal{M}_{n,n}(\mathbb{R})$ .

In this paper we follow a different approach, which is based on a Lyapunov stability analysis. The special case of  $2 \times 2$  systems and  $F(u)$  diagonal has recently been treated in our previous paper [6]. In the present paper, by using a more general strict Lyapunov function (see section 4), we get a new weaker dissipative boundary condition, stated as follows:

$$\inf \{ \|\Delta G'(0)\Delta^{-1}\|; \Delta \in \mathcal{D}_{n,+} \} < 1,$$

where  $\|\cdot\|$  denotes the usual 2-norm of matrices in  $\mathcal{M}_{n,n}(\mathbb{R})$  and  $\mathcal{D}_{n,+}$  denotes the set of diagonal matrices whose elements on the diagonal are strictly positive.

Moreover, our proof is rather elementary, and the existence of a strict Lyapunov function may be useful for studying robustness issues.

Our paper is organized as follows. In section 2, after some mathematical preliminaries, a precise technical definition of our new dissipative boundary condition is followed by the statement of our exponential stability theorem. Section 3 is then devoted to a discussion of the optimality properties of our dissipative boundary condition and to a comparison of this condition with other stability criteria from the literature, namely the criterion [14, Theorem 1.3, p. 173] mentioned above and a stability criterion for *linear* hyperbolic systems due to Silkowski. The proof of our exponential stability theorem, including the Lyapunov stability analysis, is thoroughly given in section 4. The paper ends with two appendices, where some technical properties of our dissipative boundary condition are given.

## 2. A sufficient condition for exponential stability. For

$$x := (x_1, \dots, x_n)^{\text{tr}} \in \mathbb{C}^n,$$

$|x|$  denotes the usual Hermitian norm of  $x$ :

$$|x| := \sqrt{\sum_{i=1}^n |x_i|^2}.$$

For  $n \in \mathbb{N} \setminus \{0\}$  and  $m \in \mathbb{N} \setminus \{0\}$ , we denote by  $\mathcal{M}_{n,m}(\mathbb{R})$  the set of  $n \times m$  real matrices. We define, for  $K \in \mathcal{M}_{n,m}(\mathbb{R})$ ,

$$\|K\| := \max\{|Kx|; x \in \mathbb{R}^n, |x| = 1\} = \max\{|Kx|; x \in \mathbb{C}^n, |x| = 1\},$$

and, if  $n = m$ ,

$$(2.1) \quad \rho_1(K) := \inf \{ \|\Delta K \Delta^{-1}\|; \Delta \in \mathcal{D}_{n,+} \},$$

where  $\mathcal{D}_{n,+}$  denotes the set of  $n \times n$  real diagonal matrices with strictly positive diagonal elements.

For  $\varepsilon$ , let  $B_\varepsilon$  be the open ball of  $\mathbb{R}^n$  of radius  $\varepsilon$ . We assume that, for some  $\varepsilon_0 > 0$ ,  $F : B_{\varepsilon_0} \rightarrow \mathcal{M}_{n,n}(\mathbb{R})$  is of class  $C^2$  and that there exists  $m \in \{0, \dots, n\}$  and  $n$  real numbers  $\Lambda_1, \dots, \Lambda_n$  such that

$$(2.2) \quad \Lambda_i > 0 \quad \forall i \in \{1, \dots, m\} \text{ and } \Lambda_i < 0 \quad \forall i \in \{m+1, \dots, n\},$$

$$(2.3) \quad F(0) = \text{diag}(\Lambda_1, \dots, \Lambda_n),$$

$$(2.4) \quad \Lambda_i \neq \Lambda_j \quad \forall (i, j) \in \{1, \dots, n\}^2 \text{ such that } i \neq j.$$

For  $u \in \mathbb{R}^n$ ,  $u_+ \in \mathbb{R}^m$  and  $u_- \in \mathbb{R}^{n-m}$  are defined by requiring

$$u = \begin{pmatrix} u_+ \\ u_- \end{pmatrix}.$$

As mentioned in the introduction, we are mainly interested in analyzing the asymptotic convergence of the classical solutions of the following Cauchy problem:

$$(2.5) \quad u_t + F(u)u_x = 0, \quad x \in [0, 1], \quad t \in [0, +\infty),$$

$$(2.6) \quad \begin{pmatrix} u_+(t, 0) \\ u_-(t, 1) \end{pmatrix} = G \begin{pmatrix} u_+(t, 1) \\ u_-(t, 0) \end{pmatrix}, \quad t \in [0, +\infty),$$

$$(2.7) \quad u(0, x) = u^0(x), \quad x \in [0, 1].$$

Concerning  $G$ , we assume that  $G : B_{\varepsilon_0} \rightarrow \mathbb{R}^n$  is of class  $C^2$  and satisfies  $G(0) = 0$ . We define  $F_+(u) \in \mathcal{M}_{m,n}(\mathbb{R})$ ,  $F_-(u) \in \mathcal{M}_{(n-m),n}(\mathbb{R})$ ,  $G_+(u) \in \mathbb{R}^m$ , and  $G_-(u) \in \mathbb{R}^{n-m}$  by requiring

$$F(u) = \begin{pmatrix} F_+(u) \\ F_-(u) \end{pmatrix}, \quad G(u) = \begin{pmatrix} G_+(u) \\ G_-(u) \end{pmatrix}.$$

Regarding the existence of the solutions to the Cauchy problem (2.5)–(2.7), we have the following proposition.

**PROPOSITION 2.1.** *There exists  $\delta_0 > 0$  such that, for every  $u^0 \in H^2((0, 1), \mathbb{R}^n)$  satisfying*

$$|u^0|_{H^2((0,1),\mathbb{R}^n)} \leq \delta_0$$

*and the compatibility conditions*

$$(2.8) \quad \begin{pmatrix} u_+^0(0) \\ u_-^0(1) \end{pmatrix} = G \begin{pmatrix} u_+^0(1) \\ u_-^0(0) \end{pmatrix},$$

$$(2.9) \quad F_+(u^0(0))u_x^0(0) = \left[ G'_{+u_+} \begin{pmatrix} u_+^0(1) \\ u_-^0(0) \end{pmatrix} \right] F_+(u^0(1))u_x^0(1) \\ + \left[ G'_{+u_-} \begin{pmatrix} u_+^0(1) \\ u_-^0(0) \end{pmatrix} \right] F_-(u^0(0))u_x^0(0),$$

$$(2.10) \quad F_-(u^0(1))u_x^0(1) = \left[ G'_{-u_+} \begin{pmatrix} u_+^0(1) \\ u_-^0(0) \end{pmatrix} \right] F_+(u^0(1))u_x^0(1) \\ + \left[ G'_{-u_-} \begin{pmatrix} u_+^0(1) \\ u_-^0(0) \end{pmatrix} \right] F_-(u^0(0))u_x^0(0),$$

the Cauchy problem (2.5)–(2.7) has a unique maximal classical solution

$$u \in C^0([0, T], H^2((0, 1), \mathbb{R}^n))$$

with  $T \in [0, +\infty]$ . Moreover, if

$$|u(t, \cdot)|_{H^2((0, 1), \mathbb{R}^n)} \leq \delta_0 \quad \forall t \in [0, T],$$

then  $T = +\infty$ .

For a proof of this proposition, see, for instance, [12] by Kato, [13, pp. 2–3] by Lax, [16, pp. 35–43] by Majda, or [20, pp. 106–114] by Serre. Actually [12, 13, 16, 20] deal with  $\mathbb{R}$  instead of  $[0, 1]$ , but the proofs given there can be adapted to treat this new case. See also [15, pp. 96–107] by Li and Yu for the well-posedness of the Cauchy problem (2.5)–(2.7) in the framework of functions  $u$  of class  $C^1$ . Let us briefly explain how to adapt these proofs in order to get, for example, the existence of a solution  $u \in C^0([0, T], H^2((0, 1), \mathbb{R}^n))$  to the Cauchy problem (2.5)–(2.7) if  $m = n$  (just to simplify the notation), for  $T \in (0, +\infty)$  given, and for every  $u^0 \in H^2((0, 1), \mathbb{R}^n)$  satisfying the compatibility conditions (2.8)–(2.9) (when  $m = n$ , condition (2.10) disappears) and such that  $|u^0|_{H^2((0, 1), \mathbb{R}^n)}$  is small enough (the smallness depending on  $T$  in general). We first deal with the case where

$$T \in (0, \min\{\Lambda_1^{-1}, \dots, \Lambda_n^{-1}\}).$$

The basic ingredient is the following fixed point method, which is related to the one given in [15, page 97] (see also the pioneering works [12] and [13, pp. 2–3], where the authors deal with  $\mathbb{R}$  instead of  $[0, 1]$ ). For  $R > 0$  and for  $u^0 \in H^2((0, 1), \mathbb{R}^n)$  satisfying the compatibility conditions (2.8)–(2.9), let  $C_R(u^0)$  be the set of

$$u \in L^\infty((0, T), H^2((0, 1), \mathbb{R}^n)) \cap W^{1, \infty}((0, T), H^1((0, 1), \mathbb{R}^n)) \\ \cap W^{2, \infty}((0, T), L^2((0, 1), \mathbb{R}^n))$$

such that

$$|u|_{L^\infty((0, T), H^2((0, 1), \mathbb{R}^n))} \leq R, \\ |u|_{W^{1, \infty}((0, T), H^1((0, 1), \mathbb{R}^n))} \leq R, \\ |u|_{W^{2, \infty}((0, T), L^2((0, 1), \mathbb{R}^n))} \leq R, \\ u(\cdot, 1) \in H^2((0, T), \mathbb{R}^n) \text{ and } |u(\cdot, 1)|_{H^2((0, T), \mathbb{R}^n)} \leq R^2, \\ u(0, \cdot) = u^0, \\ u_t(0, \cdot) = -F(u^0)u_x^0.$$

The set  $C_R(u^0)$  is a closed subset of  $L^\infty((0, T), L^2((0, 1), \mathbb{R}^n))$  (at least if  $|u^0|_{H^2((0, 1), \mathbb{R}^n)}$  is small enough so that  $|u^0|_{C^0([0, 1], \mathbb{R}^n)} < \varepsilon_0$ ). Given  $R > 0$ , the set  $C_R(u^0)$  is not empty if  $|u^0|_{H^2((0, 1), \mathbb{R}^n)}$  is small enough. Let  $\mathcal{F} : C_R(u^0) \rightarrow L^\infty((0, T), H^2((0, 1), \mathbb{R}^n)) \cap W^{1, \infty}((0, T), H^1((0, 1), \mathbb{R}^n)) \cap W^{2, \infty}((0, T), L^2((0, 1), \mathbb{R}^n))$  be defined by  $\mathcal{F}(\tilde{u}) = u$ , where  $u$  is the solution of the linear hyperbolic Cauchy problem

$$(2.11) \quad u_t + F(\tilde{u})u_x = 0, \quad u(t, 0) = G(\tilde{u}(t, 1)), \quad t \in [0, T],$$

$$u(0, x) = u^0(x), \quad x \in [0, 1].$$

The set  $C_R(u^0)$  is a closed subset of  $L^\infty((0, T), L^2((0, 1), \mathbb{R}^n))$  (at least if  $|u^0|_{H^2((0, 1), \mathbb{R}^n)}$  is small enough so that  $|u^0|_{L^\infty((0, 1), \mathbb{R}^n)} \leq \varepsilon_0/2$ ). Moreover, given  $R > 0$ ,  $C_R(u^0)$  is not empty if  $|u^0|_{H^2((0, 1), \mathbb{R}^n)}$  is small enough. Using standard energy estimates and the finite speed of propagation inherent in (2.11), one gets the existence of  $M > 0$  and  $R_0 > 0$  such that, for every  $R \in (0, R_0]$ , there exists  $\delta > 0$  such that, for every  $u^0 \in H^2((0, 1), \mathbb{R}^n)$  such that  $|u^0|_{H^2((0, 1), \mathbb{R}^n)} \leq \delta$  and satisfying the compatibility conditions (2.8)–(2.9),

$$(2.12) \quad \mathcal{F}(C_R(u^0)) \subset C_R(u^0)$$

and

$$\begin{aligned} & |\mathcal{F}(\tilde{u}_2) - \mathcal{F}(\tilde{u}_1)|_{L^\infty((0, T), L^2((0, 1), \mathbb{R}^n))} + M|\mathcal{F}(\tilde{u}_2)(\cdot, 1) - \mathcal{F}(\tilde{u}_1)(\cdot, 1)|_{L^2((0, 1), \mathbb{R}^n)} \\ & \leq \frac{1}{2}|\tilde{u}_2 - \tilde{u}_1|_{L^\infty((0, T), L^2((0, 1), \mathbb{R}^n))} + \frac{M}{2}|\tilde{u}_2(\cdot, 1) - \tilde{u}_1(\cdot, 1)|_{L^2((0, 1), \mathbb{R}^n)} \quad \forall (\tilde{u}_1, \tilde{u}_2) \in C_R(u^0). \end{aligned}$$

This allows us to prove that  $\mathcal{F}$  has a fixed point  $u \in C_R(u^0)$ ; i.e., there exists a solution  $u \in C_R(u^0)$  to the Cauchy problem (2.5)–(2.7). In order to get the extra regularity property  $u \in C^0([0, T], H^2((0, 1), \mathbb{R}^n))$ , one can adapt [16, pp. 44–46] by noticing that, when one uses usual energy estimates to get (2.12), one also gets, for  $u := \mathcal{F}(\tilde{u})$  with  $\tilde{u} \in C_R(u^0)$ , the “hidden regularity”  $u_{xx}(\cdot, 1) \in L^2((0, T), \mathbb{R}^n)$  together with estimates on  $|u_{xx}(\cdot, 1)|_{L^2((0, T), \mathbb{R}^n)}$  which are sufficient to take care of the boundary terms which now appear when one does integrations by parts. The case of general  $T \in (0 + \infty)$  follows by applying the above result to  $[0, T_1], [T_1, 2T_1], [2T_1, 3T_1], \dots$ , with  $T_1$  given in  $(0, \min\{\Lambda_1^{-1}, \dots, \Lambda_n^{-1}\})$ . This concludes our sketch of the proof of Proposition 2.1.

We adopt the following definition of the exponential stability of the equilibrium solution  $u \equiv 0$ .

**DEFINITION 2.2.** *The equilibrium solution  $u \equiv 0$  of the nonlinear hyperbolic system (2.5)–(2.6) is exponentially stable (for the  $H^2$ -norm) if there exist  $\varepsilon > 0$ ,  $\nu > 0$ , and  $C > 0$  such that, for every  $u^0 \in H^2((0, 1), \mathbb{R}^n)$  satisfying  $|u^0|_{H^2((0, 1), \mathbb{R}^n)} \leq \varepsilon$  and the compatibility conditions (2.8)–(2.10), the classical solution  $u$  to the Cauchy problem (2.5)–(2.7) is defined on  $[0, +\infty)$  and satisfies*

$$(2.13) \quad |u(t, \cdot)|_{H^2((0, 1), \mathbb{R}^n)} \leq Ce^{-\nu t}|u_0|_{H^2((0, 1), \mathbb{R}^n)} \quad \forall t \in [0, +\infty).$$

Our main result is the following theorem.

**THEOREM 2.3.** *If  $\rho_1(G'(0)) < 1$ , then the equilibrium  $u \equiv 0$  of the quasi-linear hyperbolic system (2.5)–(2.6) is exponentially stable.*

The proof of this theorem is given in section 4.

As mentioned in the introduction, the next section is devoted to a comparison of our dissipative boundary condition (i.e.,  $\rho_1(G'(0)) < 1$ ) with other stability criteria from the literature, namely the criterion given in [14, Theorem 1.3, page 173] and a stability criterion for linear hyperbolic systems.

**3. Comparison with other stability conditions.** In this section, we first compare our condition  $\rho_1(G'(0)) < 1$  for exponential stability (Theorem 2.3) with a prior condition found by Li [14, Theorem 1.3, page 173]. In the second part of this section we shall compare our condition to conditions for the stability of *linear* hyperbolic systems.

**3.1. Comparison with the Li condition.** Let us first introduce some notation and definitions. For  $K \in \mathcal{M}_{n,m}(\mathbb{R})$ , we denote by  $K_{ij}$  the term on the  $i$ th line and  $j$ th column of the matrix  $K$  and denote by  $|K|$  the matrix in  $\mathcal{M}_{n,m}(\mathbb{R})$  defined by

$$|K|_{ij} := |K_{ij}| \quad \forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, m\}.$$

We define, for  $K \in \mathcal{M}_{n,n}(\mathbb{R})$ ,

$$R_2(K) := \text{Max} \left\{ \sum_{j=1}^n |K_{ij}|; i \in \{1, \dots, n\} \right\},$$

$$\rho_2(K) := \text{Inf} \{R_2(\Delta K \Delta^{-1}); \Delta \in \mathcal{D}_{n,+}\}.$$

Note that, by [14, Lemma 2.4, page 146],

$$(3.1) \quad \rho_2(K) = \rho(|K|),$$

where, for  $A \in \mathcal{M}_{n,n}(\mathbb{R})$ ,  $\rho(A)$  is the spectral radius of  $A$ . In the following theorem, we recall the sufficient condition for exponential stability introduced by Li.

**THEOREM 3.1** (see [14, Theorem 1.3, page 173]). *Assume that  $\rho_2(G'(0)) < 1$ ; then  $0 \in C^1([0, 1], \mathbb{R}^n)$  is locally exponentially stable in the  $C^1([0, 1])$ -norm for the hyperbolic system (2.5)–(2.6); i.e., there exist  $\varepsilon > 0$ ,  $\nu > 0$ , and  $C > 0$  such that, for every  $u^0 \in C^1([0, 1], \mathbb{R}^n)$  satisfying  $|u^0|_{C^1([0, 1], \mathbb{R}^n)} \leq \varepsilon$  and the compatibility conditions (2.8)–(2.10), the Cauchy problem (2.5)–(2.7) has a unique solution  $u$  in  $C^1([0, +\infty) \times [0, 1], \mathbb{R}^n)$ , and this solution satisfies*

$$|u(t, \cdot)|_{C^1([0, 1], \mathbb{R}^n)} \leq C e^{-\nu t} |u^0|_{C^1([0, 1], \mathbb{R}^n)} \quad \forall t \in [0, +\infty).$$

The following proposition and (3.3) show that our new sufficient condition, namely  $\rho_1(G'(0)) < 1$ , is weaker than the previous one.

**PROPOSITION 3.2.** *For every  $K \in \mathcal{M}_{n,n}(\mathbb{R})$ ,*

$$(3.2) \quad \rho_1(K) \leq \rho_2(K).$$

Let us point out that there are matrices  $K$  such that inequality (3.2) is strict. For example, for  $a > 0$ , let

$$K_a := \begin{pmatrix} a & a \\ -a & a \end{pmatrix} \in \mathcal{M}_{2,2}(\mathbb{R}).$$

Then

$$(3.3) \quad \rho_1(K_a) = \sqrt{2}a < 2a = \rho_2(K_a).$$

**Remark 3.3.** In fact, in [14, Theorem 1.3, page 173], it is assumed that  $G_+$  depends only on  $u_-$  and that  $G_-$  depends only on  $u_+$ . However, if one takes

$$K := \begin{pmatrix} 0 & K_a \\ K_a & 0 \end{pmatrix} \in \mathcal{M}_{4,4}(\mathbb{R}),$$

$n = 4$ ,  $m = 2$ , and  $G(u) := Ku$ , which are allowed by the type of boundary conditions considered in [14, Theorem 1.3, page 173], one again gets  $\rho_1(K) = \sqrt{2}a < 2a = \rho_2(K)$ .

*Proof of Proposition 3.2.* Let us first prove the following lemma.

LEMMA 3.4. *For every  $K \in \mathcal{M}_{n,n}(\mathbb{R})$ , for every  $D \in \mathcal{D}_{n,+}$ , for every  $\Delta \in \mathcal{D}_{n,+}$ , for every  $X \in \mathbb{R}^n$ , and for every  $Y \in \mathbb{R}^n$ ,*

$$(3.4) \quad Y^{\text{tr}} \Delta K \Delta^{-1} X \leq \frac{1}{2} R_2(D \Delta^{-1} K^{\text{tr}} \Delta D^{-1}) |X|^2 + \frac{1}{2} R_2(D \Delta K \Delta^{-1} D^{-1}) |Y|^2.$$

*Proof of Lemma 3.4.* Replacing, if necessary,  $K$  by  $\Delta K \Delta^{-1}$ , we may assume without loss of generality that  $\Delta$  is the identity map of  $\mathbb{R}^n$ . We write  $X := (X_1, \dots, X_n)^{\text{tr}} \in \mathbb{R}^n$ ,  $Y := (Y_1, \dots, Y_n)^{\text{tr}} \in \mathbb{R}^n$ ,  $D := \text{diag}(D_1, \dots, D_n)$ . One has

$$(3.5) \quad \begin{aligned} Y^{\text{tr}} K X &= \sum_{i=1}^n Y_i \left( \sum_{j=1}^n K_{ij} X_j \right) = \sum_{i=1}^n \sum_{j=1}^n \frac{K_{ij}}{D_i D_j} D_i Y_i D_j X_j \\ &\leq \frac{1}{2} Q_1 + \frac{1}{2} Q_2, \end{aligned}$$

with

$$Q_1 := \sum_{i=1}^n \sum_{j=1}^n \frac{|K_{ij}|}{D_i D_j} D_j^2 X_j^2 \text{ and } Q_2 := \sum_{i=1}^n \sum_{j=1}^n \frac{|K_{ij}|}{D_i D_j} D_i^2 Y_i^2.$$

Note that

$$(3.6) \quad \begin{aligned} Q_1 &= \sum_{j=1}^n \left( \sum_{i=1}^n |K_{ij}| D_i^{-1} D_j \right) X_j^2 = \sum_{j=1}^n \left( \sum_{i=1}^n |(D^{-1} K D)^{\text{tr}}|_{ji} \right) X_j^2 \\ &\leq \sum_{j=1}^n R_2((D^{-1} K D)^{\text{tr}}) X_j^2 = R_2(D K^{\text{tr}} D^{-1}) |X|^2. \end{aligned}$$

Similarly,

$$(3.7) \quad \begin{aligned} Q_2 &= \sum_{i=1}^n \left( \sum_{j=1}^n D_i |K_{ij}| D_j^{-1} \right) Y_i^2 = \sum_{i=1}^n \left( \sum_{j=1}^n |(D K D^{-1})_{ij}| \right) Y_i^2 \\ &\leq \sum_{i=1}^n R_2(D K D^{-1}) Y_i^2 = R_2(D K D^{-1}) |Y|^2. \end{aligned}$$

Inequality (3.4) follows from (3.5), (3.6), and (3.7). This concludes the proof of Lemma 3.4.  $\square$

Let us go back to the proof of Proposition 3.2. One easily sees that

$$(3.8) \quad \{(D \Delta^{-1}, D \Delta); D \in \mathcal{D}_{n,+}, \Delta \in \mathcal{D}_{n,+}\} = \mathcal{D}_{n,+} \times \mathcal{D}_{n,+}.$$

Equality (3.8) implies that

$$(3.9) \quad \begin{aligned} &\rho_2(K^{\text{tr}}) + \rho_2(K) \\ &= \inf \{ R_2(D \Delta^{-1} K^{\text{tr}} \Delta D^{-1}) + R_2(D \Delta K \Delta^{-1} D^{-1}); D \in \mathcal{D}_{n,+}, \Delta \in \mathcal{D}_{n,+} \}. \end{aligned}$$

Using (3.1), we have

$$(3.10) \quad \rho_2(K^{\text{tr}}) = \rho(|K^{\text{tr}}|) = \rho(|K|^{\text{tr}}) = \rho(|K|) = \rho_2(K),$$

which, together with (3.9), gives

$$(3.11) \quad \inf \{R_2(D\Delta^{-1}|K|^{\text{tr}}\Delta D^{-1}) + R_2(D\Delta|K|\Delta^{-1}D^{-1}); D \in \mathcal{D}_{n,+}, \Delta \in \mathcal{D}_{n,+}\} \\ = 2\rho_2(K).$$

Finally, let us note that, for every  $\Delta$  in  $\mathcal{D}_{n,+}$ ,

$$(3.12) \quad \sup \{Y^{\text{tr}}\Delta K\Delta^{-1}X; X \in \mathbb{R}^n, Y \in \mathbb{R}^n, |X| = |Y| = 1\} = \|\Delta K\Delta^{-1}\| \geq \rho_1(K).$$

Proposition 3.2 follows from (3.4), (3.11), and (3.12).  $\square$

**3.2. Comparison with stability conditions for linear hyperbolic systems.** Replacing, if necessary,  $y(t, x)$  by

$$\begin{pmatrix} y_+(t, x) \\ y_-(t, 1-x) \end{pmatrix},$$

it may be assumed, without loss of generality, that the speeds of propagation  $\Lambda_i$  are all positive. More precisely we consider the special case of linear hyperbolic systems

$$(3.13) \quad y_t + \Lambda y_x = 0, \quad y(t, 0) = Ky(t, 1),$$

where

$$(3.14) \quad \Lambda := \text{diag}(\Lambda_1, \dots, \Lambda_n), \quad \text{with } \Lambda_i > 0 \quad \forall i \in \{1, \dots, n\}.$$

In order to avoid compatibility conditions, one can deal with the case where  $y(t, \cdot) \in L^2((0, 1), \mathbb{R}^n)$  (instead of  $y(t, \cdot) \in H^2((0, 1), \mathbb{R}^n)$ , as we consider above for the nonlinear hyperbolic system (2.5)–(2.6)). It is well known that the Cauchy problem associated with (3.13) is well posed in  $L^2((0, 1), \mathbb{R}^n)$ ; that is, for every  $y^0 \in L^2((0, 1), \mathbb{R}^n)$ , there exists a unique

$$y \in C^0([0, +\infty), L^2((0, 1), \mathbb{R}^n))$$

solution of (3.13) satisfying the initial condition

$$(3.15) \quad y(0, \cdot) = y^0.$$

Of course, (3.13) has to be understood in the classical weak sense; i.e., for every  $\varphi \in C^1([0, +\infty) \times [0, 1]; \mathbb{R}^n)$  with compact support and satisfying

$$\varphi^{\text{tr}}(t, 1)\Lambda - \varphi_+^{\text{tr}}(t, 0)\Lambda K = 0 \quad \forall t \in [0, +\infty),$$

we have

$$\int_0^{+\infty} \int_0^1 (\varphi_t^{\text{tr}} + \varphi_x^{\text{tr}}\Lambda) y dx dt + \int_0^1 \varphi^{\text{tr}}(0, x) y^0(x) dx = 0.$$



See, for example, [5, section 2.1].

As usual, we say that  $0 \in L^2((0, 1), \mathbb{R}^n)$  is exponentially stable for (3.13) (for the norm of  $L^2((0, 1), \mathbb{R}^n)$ ) if there exist  $\nu > 0$  and  $C > 0$  such that, for every  $y^0 \in L^2((0, 1), \mathbb{R}^n)$ , the solution of the Cauchy problem (3.13), (3.15) satisfies

$$|y(t, \cdot)|_{L^2((0, 1), \mathbb{R}^n)} \leq C e^{-\nu t} |y^0|_{L^2((0, 1), \mathbb{R}^n)} \quad \forall t \in [0, +\infty).$$

One easily checks that (3.13) is equivalent to

$$(3.16) \quad \phi_i(t) = \sum_{j=1}^n K_{ij} \phi_j(t - r_j) \quad \forall i \in \{1, \dots, n\},$$

with

$$\phi_j(t) := y_j(t, 0), \quad r_j := \frac{1}{\Lambda_j}, \quad j \in \{1, \dots, n\}.$$

Hence (3.13) can be considered as a linear time-delay system. By a classical result on linear time-delay systems (see, e.g., [10, Theorem 3.5 page 275] by Hale and Verduyn Lunel),  $0 \in L^2((0, 1), \mathbb{R}^n)$  is exponentially stable for the system (3.13) if and only if there exists  $\delta > 0$  such that

$$(3.17) \quad \left( \det (\text{Id}_n - (\text{diag} (e^{-r_1 z}, \dots, e^{-r_n z})) K) = 0, z \in \mathbb{C} \right) \Rightarrow (\Re(z) \leq -\delta),$$

where  $\text{Id}_n$  is the identity map of  $\mathbb{R}^n$  and  $\Re(z)$  denotes the real part of the complex number  $z$ . Note that  $\rho_1(K) < 1$  implies the existence of  $\delta > 0$  such that (3.17) holds. Indeed, let us assume that  $\rho_1(K) < 1$ . Then, by (2.1), there exist  $\mu \in (0, 1)$  and  $D \in \mathcal{D}_{n,+}$  such that

$$(3.18) \quad \|DKD^{-1}\| \leq \mu.$$

Let us assume that  $z \in \mathbb{C}$  is such that

$$\det (\text{Id}_n - (\text{diag} (e^{-r_1 z}, \dots, e^{-r_n z})) K) = 0.$$

Then

$$\begin{aligned} & \det (\text{Id}_n - (\text{diag} (e^{-r_1 z}, \dots, e^{-r_n z})) DKD^{-1}) \\ &= \det (D(\text{Id}_n - (\text{diag} (e^{-r_1 z}, \dots, e^{-r_n z})) K)D^{-1}) \\ &= \det (\text{Id}_n - (\text{diag} (e^{-r_1 z}, \dots, e^{-r_n z})) K) = 0, \end{aligned}$$

which implies that

$$(3.19) \quad \|(\text{diag} (e^{-r_1 z}, \dots, e^{-r_n z})) DKD^{-1}\| \geq 1.$$

Since

$$\begin{aligned} \|(\text{diag} (e^{-r_1 z}, \dots, e^{-r_n z})) DKD^{-1}\| &\leq \|\text{diag} (e^{-r_1 z}, \dots, e^{-r_n z})\| \|DKD^{-1}\| \\ &\leq e^{-\min\{r_1 \Re(z), \dots, r_n \Re(z)\}} \|DKD^{-1}\|, \end{aligned}$$

one has, also using (3.18) and (3.19),

$$(3.20) \quad e^{-\min\{r_1 \Re(z), \dots, r_n \Re(z)\}} \mu \geq 1.$$

Inequality (3.20) implies that (3.17) holds with  $\delta := \ln(\mu)/(\max\{r_1, \dots, r_n\}) < 0$ .

The converse is false: the existence of  $\delta > 0$  such that (3.17) holds does not imply that  $\rho_1(K) < 1$ . For example, let us choose  $r_1 := 1$ ,  $r_2 := 2$ , and

$$K := \begin{pmatrix} a & a \\ a & a \end{pmatrix}, \quad a \in \mathbb{R}.$$

(This example is borrowed from [10, page 285].) It is easily seen that  $\rho_1(K) = 2|a|$ . Hence  $\rho_1(K) < 1$  is equivalent to  $a \in (-1/2, 1/2)$ . However, the existence of  $\delta > 0$  such that (3.17) holds is equivalent to  $a \in (-1, 1/2)$ .

If we want to try to apply results on the stability of the linear hyperbolic system (3.13) in order to get the stability of our nonlinear hyperbolic system (2.5)–(2.6), since  $F(u)$  depends on  $u$ , it is natural to ask for the robustness of the stability of the linear hyperbolic system (3.13) with respect to small changes on the  $\Lambda_i$ 's, i.e., on the speeds of propagation. (One can easily see that the stability is robust with respect to small changes on  $K$ .) Let us adopt the following definition.

**DEFINITION 3.5.** *The linear system (3.13) is robustly exponentially stable with respect to the speeds of propagation if there exists  $\varepsilon > 0$  such that, for every  $\tilde{\Lambda} := \text{diag}(\tilde{\Lambda}_1, \dots, \tilde{\Lambda}_n) \in \mathcal{D}_{n,+}$  such that*

$$|\tilde{\Lambda}_i - \Lambda_i| \leq \varepsilon \quad \forall i \in \{1, \dots, n\},$$

*$0 \in L^2((0, 1), \mathbb{R}^n)$  is exponentially stable for the perturbed linear hyperbolic system*

$$y_t + \tilde{\Lambda} y_x = 0, \quad y(t, 0) = K y(t, 1).$$

One has, then, the following theorem, which is due to Silkowski (see [10, Theorem 6.1, page 286]; see also [26, 11]).

**THEOREM 3.6.** *Let*

$$(3.21) \quad \rho_0(K) := \max\{\rho(\text{diag}(e^{\iota\theta_1}, \dots, e^{\iota\theta_n})K); (\theta_1, \dots, \theta_n)^{\text{tr}} \in \mathbb{R}^n\},$$

*with  $\iota := \sqrt{-1}$ . If the  $(r_1, \dots, r_n)$  are rationally independent, the linear system (3.13) is exponentially stable if and only if  $\rho_0(K) < 1$ . In particular (note that  $\rho_0(K)$  depends continuously on  $K$ ), whatever  $(r_1, \dots, r_n) \in (0, +\infty)^n$  is, the linear system (3.13) is robustly exponentially stable with respect to the speeds of propagation if and only if  $\rho_0(K) < 1$ .*

From this theorem the interest of comparing  $\rho_0(K)$  and  $\rho_1(K)$  is clear. This is done in the following proposition.

**PROPOSITION 3.7.** *For every  $n \in \mathbb{N}$  and for every  $K \in \mathcal{M}_{n,n}(\mathbb{R})$ ,*

$$(3.22) \quad \rho_0(K) \leq \rho_1(K).$$

*For every  $n \in \{1, 2, 3, 4, 5\}$  and for every  $K \in \mathcal{M}_{n,n}(\mathbb{R})$ ,*

$$(3.23) \quad \rho_0(K) = \rho_1(K).$$

*For every  $n \in \mathbb{N} \setminus \{1, 2, 3, 4, 5\}$ , there exists  $K \in \mathcal{M}_{n,n}(\mathbb{R})$  such that*

$$(3.24) \quad \rho_0(K) < \rho_1(K).$$

The proof of Proposition 3.7 is given in Appendix B.

**4. Proof of Theorem 2.3.** For the clarity of the analysis, we first deal in detail with the case where  $m = n$  and then give only the main modifications to deal with the case  $m < n$ . When  $m = n$  the boundary condition (2.6) reads

$$(4.1) \quad u(t, 0) = G(u(t, 1)), \quad t \in [0, +\infty),$$

and the compatibility conditions (2.8)–(2.10) become

$$(4.2) \quad u^0(0) = G(u^0(1)),$$

$$(4.3) \quad F(u^0(0))u_x^0(0) = G'(u^0(1))F(u^0(1))u_x^0(1).$$

Let us introduce some simplifying notation,

$$(4.4) \quad \Lambda := F(0) \in \mathcal{D}_{n,+}, \quad K := G'(0), \quad v := u_x, \quad w := v_x = u_{xx},$$

and let us denote by  $\mathcal{S}_n$  the set of  $n \times n$  real symmetric matrices and by  $\mathcal{S}_{n,+}$  the set of  $n \times n$  real symmetric positive definite matrices.

We shall repeatedly use the following lemma.

**LEMMA 4.1.** *Let  $\Lambda := \text{diag}(\Lambda_1, \dots, \Lambda_n) \in \mathcal{D}_n$  be such that (2.4) holds. Let  $\Delta \in \mathcal{D}_n$ . Then there exist a positive real number  $\eta$  and a map  $N : \{M \in \mathcal{M}_{n,n}(\mathbb{R}); \|M - \Lambda\| < \eta\} \rightarrow \mathcal{S}_n$  of class  $C^\infty$  such that*

$$N(\Lambda) = \Delta,$$

$$N(M)M - M^{\text{tr}}N(M) = 0 \quad \forall M \in \mathcal{M}_{n,n}(\mathbb{R}) \text{ such that } \|M - \Lambda\| < \eta.$$

*Proof of Lemma 4.1.* Let  $\mathcal{A}_n$  be the set of matrices  $A \in \mathcal{M}_{n,n}(\mathbb{R})$  such that  $A^{\text{tr}} = -A$ . For  $M \in \mathcal{M}_{n,n}(\mathbb{R})$ , let us consider the following linear map:

$$\begin{aligned} \mathcal{L}_M : \mathcal{S}_n &\rightarrow \mathcal{A}_n \times \mathcal{D}_n, \\ S &\mapsto (SM - M^{\text{tr}}S, \text{Diag}(S)), \end{aligned}$$

where  $\text{Diag}(S) := \text{diag}(S_{11}, \dots, S_{nn})$ . Noticing that

$$S\Lambda - \Lambda^{\text{tr}}S = (\Lambda_j - \Lambda_i)S_{ij} \quad \forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, n\}, \forall S \in \mathcal{S}_n$$

and using (2.4), it is easily checked that  $\mathcal{L}_\Lambda : \mathcal{S}_n \rightarrow \mathcal{A}_n \times \mathcal{D}_n$  is an isomorphism. Hence there exists  $\eta > 0$  such that, for every  $M \in \mathcal{M}_{n,n}(\mathbb{R})$  such that  $\|M - \Lambda\| < \eta$ ,  $\mathcal{L}_M$  is an isomorphism. It then suffices to define  $N$  by

$$N(M) = \mathcal{L}_M^{-1}(0, \Delta).$$

This concludes the proof of Lemma 4.1.  $\square$

For the stability analysis, we now introduce the Lyapunov function candidate

$$(4.5) \quad V(u, v, w) = V_1(u) + V_2(u, v) + V_3(u, v, w),$$

with

$$(4.6) \quad V_1(u) = \int_0^1 u^{\text{tr}} Q(u) u e^{-\mu x} dx,$$

$$(4.7) \quad V_2(u, v) = \int_0^1 v^{\text{tr}} R(u) v e^{-\mu x} dx,$$

$$(4.8) \quad V_3(u, v, w) = \int_0^1 w^{\text{tr}} S(u) w e^{-\mu x} dx,$$

where  $\mu > 0$ ,  $Q(u)$ ,  $R(u)$ , and  $S(u)$  are symmetric positive definite matrices which will be defined later.

*Remark 4.2.* The weight  $e^{-\mu x}$  is essential to get a strict Lyapunov function. It is similar to the one introduced in [4] to stabilize the Euler equation of incompressible fluids (see the definition of  $V$  given on page 1886 of [4]). It has also been used by Xu and Sallet in [24] for quite general linear hyperbolic systems (see also [22] by Tchoussou, Besson, and Xu).

Let us compute the time derivative  $\dot{V}_1$  of  $V_1$  along the classical  $C^1$ -solutions of system (2.5) with boundary conditions (4.1). One has

$$\begin{aligned}\dot{V}_1 &= \int_0^1 \left\{ 2u^{\text{tr}} Q(u) u_t + u^{\text{tr}} (Q(u))_t u \right\} e^{-\mu x} dx \\ &= \int_0^1 \left\{ -2u^{\text{tr}} Q(u) F(u) u_x + u^{\text{tr}} [Q'(u) u_t] u \right\} e^{-\mu x} dx,\end{aligned}$$

where  $Q'(u)$  is the linear map from  $\mathbb{R}^n$  to  $\mathcal{S}_n$  which stands for the derivative of  $Q$  at the point  $u$ . Hence

$$(4.9) \quad \dot{V}_1 = \int_0^1 \left\{ -\left(u^{\text{tr}} Q(u) F(u) u\right)_x + u^{\text{tr}} (Q(u) F(u))_x u - u^{\text{tr}} [Q'(u) F(u) v] u \right\} e^{-\mu x} dx.$$

For  $f \in C^0([0, 1], \mathbb{R}^n)$ , we denote by  $|f|_0$  its  $C^0$ -norm:  $|f|_0 := \max\{|f(x)|; x \in [0, 1]\}$ . From now on,  $V_1$  and  $\dot{V}_1$  are considered as functionals defined, respectively, by (4.6) and (4.9) on the set  $\mathcal{V}_1$  of  $u \in C^1([0, 1], \mathbb{R}^n)$  satisfying  $|u|_0 < \varepsilon_0$  and the compatibility condition

$$(4.10) \quad u_0 = G(u_1),$$

with  $u_0 := u(0)$  and  $u_1 := u(1)$ .

Since  $\rho_1(K) < 1$  by assumption, there exists  $D \in \mathcal{D}_{n,+}$  such that  $\|DKD^{-1}\| < 1$ . The matrix  $Q(u)$  is selected as the matrix  $N(F(u))$  of Lemma 4.1 with  $\Delta := D^2 \Lambda^{-1}$ .

Our estimates on  $V_1$  and  $\dot{V}_1$  are in the following lemma.

**LEMMA 4.3.** *There exists  $\mu_1 > 0$  such that, for every  $\mu \in (0, \mu_1)$ , there exist positive real constants  $\alpha_1, \beta_1, \delta_1$  such that, for every  $u \in \mathcal{V}_1$  such that  $|u|_0 \leq \delta_1$ ,*

$$(4.11) \quad \frac{1}{\beta_1} \int_0^1 |u|^2 dx \leq V_1(u) \leq \beta_1 \int_0^1 |u|^2 dx,$$

$$(4.12) \quad \dot{V}_1(u) \leq -\alpha_1 V_1(u) + \beta_1 \int_0^1 |u|^2 |u_x| dx.$$

*Proof of Lemma 4.3.* Throughout this proof,  $u$  is assumed to be in  $\mathcal{V}_1$ . From the construction of  $Q$ ,

$$(4.13) \quad Q(0)F(0) = Q(0)\Lambda = D^2 \in \mathcal{D}_{n,+},$$

and there exists  $\delta_{11} \in (0, \varepsilon_0/2)$  such that

$$(4.14) \quad Q(a) \in \mathcal{S}_{n,+} \text{ and } Q(a)F(a) \in \mathcal{S}_{n,+} \quad \forall a \in \mathbb{R}^n \text{ such that } |a| \leq \delta_{11}.$$

Clearly, from (4.14), we obtain that, for every  $\mu > 0$ , there exists  $\beta_1 > 0$  such that (4.11) holds if  $|u|_0 \leq \delta_{11}$ .

Let us now deal with the estimate (4.12) on  $\dot{V}_1 (= \dot{V}_1(u))$ . Let us decompose  $\dot{V}_1$  in the following way:

$$(4.15) \quad \dot{V}_1 = \mathcal{T}_{11} + \mathcal{T}_{12} + \mathcal{T}_{13},$$

with

$$(4.16) \quad \mathcal{T}_{11} := -\mu \int_0^1 \left( u^{\text{tr}} Q(u) F(u) u \right) e^{-\mu x} dx,$$

$$(4.17) \quad \mathcal{T}_{12} := - \int_0^1 \left( u^{\text{tr}} Q(u) F(u) u e^{-\mu x} \right)_x dx,$$

$$(4.18) \quad \mathcal{T}_{13} := \int_0^1 \left\{ u^{\text{tr}} \left( [Q'(u)v] F(u) + Q(u) [F'(u)v] - [Q'(u)F(u)v] \right) u \right\} e^{-\mu x} dx.$$

*Analysis of the first term  $\mathcal{T}_{11}$ .* By (4.11) and (4.14), for every  $\mu > 0$ , there exists a positive real constant  $\alpha_1 > 0$  such that, if  $|u|_0 \leq \delta_{11}$ ,

$$(4.19) \quad \mathcal{T}_{11} \leq -\alpha_1 V_1.$$

*Analysis of the second term  $\mathcal{T}_{12}$ .* One has

$$\begin{aligned} \mathcal{T}_{12} &= - \left[ u^{\text{tr}} Q(u) F(u) u e^{-\mu x} \right]_0^1 \\ &= - \left( u_1^{\text{tr}} Q(u_1) F(u_1) u_1 e^{-\mu} - u_0^{\text{tr}} Q(u_0) F(u_0) u_0 \right). \end{aligned}$$

Let us introduce a notation in order to deal with estimates on “higher order terms.” We denote by  $\mathcal{O}(X, Y)$ , with  $X \geq 0$  and  $Y \geq 0$ , quantities such that there exist  $C > 0$  and  $\varepsilon > 0$ , independent of  $u, v$  and  $w$ , satisfying

$$(Y \leq \varepsilon) \Rightarrow (|\mathcal{O}(X, Y)| \leq CX).$$

Using the compatibility condition (4.10), we have

$$\begin{aligned} \mathcal{T}_{12} &= - \left( u_1^{\text{tr}} Q(u_1) F(u_1) u_1 e^{-\mu} - (G(u_1))^{\text{tr}} Q(G(u_1)) F(G(u_1)) G(u_1) \right) \\ (4.20) \quad &= -u_1^{\text{tr}} \left( Q(0) \Lambda e^{-\mu} - K^{\text{tr}} Q(0) \Lambda K \right) u_1 + \mathcal{O}(|u_1|^3; |u_1|). \end{aligned}$$

For  $u_1 \in \mathbb{R}^n$ , we define  $\zeta := Du_1$ . Then, using (4.13), we have, for every  $u_1 \in \mathbb{R}^n$ ,

$$u_1^{\text{tr}} K^{\text{tr}} Q(0) \Lambda K u_1 = u_1^{\text{tr}} K^{\text{tr}} D D K u_1 = (\zeta^{\text{tr}} D^{-1} K^{\text{tr}} D) (D K D^{-1} \zeta) = |D K D^{-1} \zeta|^2.$$

Hence, using (4.13) once again, we have, for every  $u_1 \in \mathbb{R}^n$ ,

$$(4.21) \quad u_1^{\text{tr}} K^{\text{tr}} Q(0) \Lambda K u_1 \leq \|D K D^{-1}\|^2 \zeta^{\text{tr}} \zeta = \|D K D^{-1}\|^2 u_1^{\text{tr}} Q(0) \Lambda u_1.$$

From this inequality and the fact that  $\|D K D^{-1}\| < 1$ , it follows that, taking  $\mu > 0$  small enough (which is always implicitly assumed),  $Q(0) \Lambda e^{-\mu} - K^{\text{tr}} Q(0) \Lambda K$  is a

positive definite matrix. Then, using (4.20), there exists  $\delta_{12} > 0$  such that, if  $|u_1| \leq \delta_{12}$ ,

$$(4.22) \quad \mathcal{T}_{12} \leq 0.$$

*Analysis of the third term  $\mathcal{T}_{13}$ .* The integrand of  $\mathcal{T}_{13}$  is linear with respect to  $v$  and, at least, quadratic with respect to  $u$ . It follows that, increasing the value of  $\beta_1$  if necessary, there exists a real positive constant  $\delta_{13}$  such that, for  $|u|_0 \leq \delta_{13}$ ,

$$(4.23) \quad \mathcal{T}_{13} \leq \beta_1 \int_0^1 |u|^2 |v| dx.$$

Then, collecting inequalities (4.19), (4.22), and (4.23) together, if

$$|u|_0 \leq \delta_1 := \min\{\delta_{11}, \delta_{12}, \delta_{13}\},$$

we conclude that

$$(4.24) \quad \dot{V}_1 = \mathcal{T}_{11} + \mathcal{T}_{12} + \mathcal{T}_{13} \leq -\alpha_1 V_1 + \beta_1 \int_0^1 |u|^2 |v| dx.$$

This completes the proof of Lemma 4.3.  $\square$

From Lemma 4.3 it appears that it is clearly necessary to examine the dynamics of  $v = u_x$  in order to carry out the Lyapunov stability analysis. This is the reason why the Lyapunov function (4.5) is extended with terms involving  $v$ . By time differentiation of the system equations (2.5) and (4.1), it may be shown that  $v$  satisfies the dynamics

$$(4.25) \quad v_t + F(u)v_x + [F'(u)v]v = 0, \quad x \in [0, 1], \quad t \in [0, +\infty),$$

$$(4.26) \quad F(u(t, 0))v(t, 0) = G'(u(t, 1))F(u(t, 1))v(t, 1), \quad t \in [0, +\infty).$$

Let us compute the time derivative of  $V_2$  along the classical  $C^1$ -solutions of system (4.25) with boundary conditions (4.26). One has

$$(4.27) \quad \begin{aligned} \dot{V}_2 &= \int_0^1 \left\{ 2v^{\text{tr}} R(u)v_t + v^{\text{tr}} (R(u))_t v \right\} e^{-\mu x} dx \\ &= \int_0^1 \left\{ -2v^{\text{tr}} R(u)F(u)v_x - 2v^{\text{tr}} R(u)[F'(u)v]v - v^{\text{tr}} [R'(u)F(u)v]v \right\} e^{-\mu x} dx. \end{aligned}$$

From now on,  $V_2$  and  $\dot{V}_2$  are considered as functionals defined, respectively, by (4.7) and (4.27) on the set  $\mathcal{V}_2$  of  $(u, v) \in C^2([0, 1], \mathbb{R}^n) \times C^1([0, 1], \mathbb{R}^n)$  such that

$$(4.28) \quad |u|_0 < \varepsilon_0,$$

$$(4.29) \quad u_x = v,$$

$$(4.30) \quad u_0 = G(u_1),$$

$$(4.31) \quad F(u_0)v_0 = G'(u_1)F(u_1)v_1,$$

where  $u_0 := u(0)$ ,  $u_1 := u(1)$  as above, and  $v_0 := v(0)$ ,  $v_1 := v(1)$ .

The matrix  $R(u)$  is selected as the matrix  $N(F(u))$  of Lemma 4.1 now with  $\Delta := \Lambda D^2$ . Our estimates on  $V_2$  and  $\dot{V}_2$  are in the following lemma.

LEMMA 4.4. *There exists  $\mu_2 > 0$  such that, for every  $\mu \in (0, \mu_2)$ , there exist positive real constants  $\alpha_2, \beta_2, \delta_2$  such that, for every  $(u, v) \in \mathcal{V}_2$  such that  $|u|_0 \leq \delta_2$ ,*

$$(4.32) \quad \frac{1}{\beta_2} \int_0^1 |v|^2 dx \leq V_2(u, v) \leq \beta_2 \int_0^1 |v|^2 dx,$$

$$(4.33) \quad \dot{V}_2(u, v) \leq -\alpha_2(V_2(u, v) + |v_1|^2) + \beta_2 \int_0^1 |v|^3 dx.$$

*Proof of Lemma 4.4.* Throughout this proof,  $(u, v)$  is assumed to be in  $\mathcal{V}_2$ . By the construction of  $R$ , we have

$$(4.34) \quad F(0)^{-1}R(0) = \Lambda^{-1}R(0) = D^2 \in \mathcal{D}_{n,+}$$

and the existence of  $\delta_{21} \in (0, \varepsilon_0/2)$  such that

$$(4.35) \quad R(a) \in \mathcal{S}_{n,+} \text{ and } R(a)F(a) \in \mathcal{S}_{n,+} \quad \forall a \in \mathbb{R}^n \text{ such that } |a| \leq \delta_{21}.$$

Clearly, from (4.35), for every  $\mu > 0$ , there exists  $\beta_2 > 0$  such that (4.32) holds if  $|u|_0 \leq \delta_{21}$ .

Let us now deal with the estimate (4.33) on  $\dot{V}_2 (= \dot{V}_2(u, v))$ . Let us decompose  $\dot{V}_2$  in the following way:

$$(4.36) \quad \dot{V}_2 = \mathcal{T}_{21} + \mathcal{T}_{22} + \mathcal{T}_{23},$$

with

$$\mathcal{T}_{21} := -\mu \int_0^1 (v^{\text{tr}} R(u) F(u) v) e^{-\mu x} dx,$$

$$\mathcal{T}_{22} := - \int_0^1 (v^{\text{tr}} R(u) F(u) v e^{-\mu x})_x dx,$$

$$\mathcal{T}_{23} := \int_0^1 \left\{ v^{\text{tr}} \left( [(R(u)F(u))_x v] - 2R(u)[F'(u)v] - [R'(u)F(u)v] \right) v \right\} e^{-\mu x} dx.$$

*Analysis of the first term  $\mathcal{T}_{21}$ .* By (4.32) and (4.35), for every  $\mu > 0$ , there exists a positive real constant  $\alpha_{21} > 0$  such that, if  $|u|_0 \leq \delta_{21}$ ,

$$(4.37) \quad \mathcal{T}_{21} \leq -\alpha_{21} V_2.$$

*Analysis of the second term  $\mathcal{T}_{22}$ .* One has

$$\begin{aligned} \mathcal{T}_{22} &= - \left[ v^{\text{tr}} R(u) F(u) v e^{-\mu x} \right]_0^1 \\ &= - \left( v_1^{\text{tr}} R(u_1) F(u_1) v_1 e^{-\mu} - v_0^{\text{tr}} R(u_0) F(u_0) v_0 \right). \end{aligned}$$

Under the boundary condition (4.31), we have

$$\begin{aligned} \mathcal{T}_{22} &= -v_1^{\text{tr}} \left( R(u_1) F(u_1) e^{-\mu} - F(u_1)^{\text{tr}} (G'(u_1))^{\text{tr}} \right. \\ &\quad \left. (F(G(u_1))^{-1})^{\text{tr}} R(G(u_1)) G'(u_1) F(u_1) \right) v_1, \end{aligned}$$

which implies that

$$(4.38) \quad \mathcal{T}_{22} = -v_1^{\text{tr}} \left( R(0)\Lambda e^{-\mu} - \Lambda K^{\text{tr}} \Lambda^{-1} R(0) K \Lambda \right) v_1 + \mathcal{O}(|v_1|^2 |u_1|; |u_1|).$$

We define  $\zeta := Dv_1$ . Then, using (4.34), we have, for every  $v_1 \in \mathbb{R}^n$ ,

$$(4.39) \quad v_1^{\text{tr}} K^{\text{tr}} \Lambda^{-1} R(0) K v_1 = v_1^{\text{tr}} K^{\text{tr}} D D K v_1 = (\zeta^{\text{tr}} D^{-1} K^{\text{tr}} D) (D K D^{-1} \zeta) = |D K D^{-1} \zeta|^2.$$

Therefore, using (4.34) once again, we get that, for every  $v_1 \in \mathbb{R}^n$ ,

$$(4.40) \quad v_1^{\text{tr}} K^{\text{tr}} \Lambda^{-1} R(0) K v_1 \leq \|D K D^{-1}\|^2 \zeta^{\text{tr}} \zeta = \|D K D^{-1}\|^2 v_1^{\text{tr}} \Lambda^{-1} R(0) v_1.$$

From (4.40) and the fact that  $\|D K D^{-1}\| < 1$ , it follows that, choosing  $\mu > 0$  small enough,  $\Lambda^{-1} R(0) e^{-\mu} - K^{\text{tr}} \Lambda^{-1} R(0) K$  is a positive definite matrix, which, in turn, implies that the matrix  $R(0)\Lambda e^{-\mu} - \Lambda K^{\text{tr}} \Lambda^{-1} R(0) K \Lambda$  is also positive definite. Hence there exist  $\alpha_{22} > 0$  and  $\delta_{22} > 0$  such that, if  $|u_1| \leq \delta_{22}$ , we have

$$(4.41) \quad \mathcal{T}_{22} \leq -\alpha_{22} |v_1|^2.$$

*Analysis of the third term  $\mathcal{T}_{23}$ .* The integrand of  $\mathcal{T}_{23}$  is, at least, cubic with respect to  $v$ . It follows that, increasing the value of  $\beta_2$  if necessary, there exists  $\delta_{23} \in (0, \varepsilon_0/2)$  such that, for  $|u|_0 \leq \delta_{23}$ ,

$$(4.42) \quad \mathcal{T}_{23} \leq \beta_2 \int_0^1 |v|^3 dx.$$

Then, collecting inequalities (4.37), (4.41), and (4.42) together, we conclude that if  $|u|_0 \leq \delta_2 := \min\{\delta_{21}, \delta_{22}, \delta_{23}\}$  and if  $\alpha_2 := \min\{\alpha_{21}, \alpha_{22}\}$ , then

$$(4.43) \quad \dot{V}_2 = \mathcal{T}_{21} + \mathcal{T}_{22} + \mathcal{T}_{23} \leq -\alpha_2 (V_2 + |v_1|^2) + \beta_2 \int_0^1 |v|^3 dx.$$

This completes the proof of Lemma 4.4.  $\square$

Note that  $V_2$  is not sufficient to get an upper bound on  $\int_0^1 |v|^3 dx$ . For that reason, we also need to consider the dynamics of  $w$  to complete the Lyapunov stability analysis. By a further time differentiation of the system equations (4.25)–(4.26), we obtain

$$(4.44) \quad \begin{aligned} w_t + F(u)w_x + [F'(u)w]v + 2[F'(u)v]w + [F''(u)(v, v)]v &= 0, \\ x \in [0, 1], \quad t \in [0, +\infty), \end{aligned}$$

under the boundary condition

$$(4.45) \quad F(u_0)w_0 + [F'(u_0)v_0]v_0 = [H'(u_1)F(u_1)v_1]v_1 + H(u_1)F(u_1)w_1 + H(u_1)[F'(u_1)v_1]v_1,$$

with the notation  $w_0 := w(0)$ ,  $w_1 := w(1)$ , and  $H(u) := F(G(u))^{-1}G'(u)F(u)$ . Using the previous boundary conditions (4.30) and (4.31), this boundary condition (4.45) may be written in compact form as

$$(4.46) \quad w_0 = F(G(u_1))^{-1}H(u_1)F(u_1)w_1 + \mathcal{Z}(u_1, v_1),$$



where  $\mathcal{Z}$  is continuous on a neighborhood of  $0 \in \mathbb{R}^n \times \mathbb{R}^n$  and such that

$$(4.47) \quad \mathcal{Z}(u_1, v_1) = \mathcal{O}(|v_1|^2; |u_1|).$$

Let us compute the time derivative of  $V_3$  along the classical  $C^1$ -solutions of system (4.44) with boundary conditions (4.46). One has

$$(4.48) \quad \begin{aligned} \dot{V}_3 &= \int_0^1 \left\{ 2w^{\text{tr}} S(u) w_t + w^{\text{tr}} (S(u))_t w \right\} e^{-\mu x} dx \\ &= \int_0^1 \left\{ -2w^{\text{tr}} S(u) F(u) w_x - 2w^{\text{tr}} S(u) \left( [F'(u)w]v + 2[F'(u)v]w + [F''(u)(v, v)]v \right) \right. \\ &\quad \left. - w^{\text{tr}} [S'(u)F(u)v]w \right\} e^{-\mu x} dx. \end{aligned}$$

From now on,  $V_3$  and  $\dot{V}_3$  are considered as functionals defined, respectively, by (4.8) and (4.48) on the set  $\mathcal{V}_3$  of  $(u, v, w) \in C^3([0, 1], \mathbb{R}^n) \times C^2([0, 1], \mathbb{R}^n) \times C^1([0, 1], \mathbb{R}^n)$  such that

$$(4.49) \quad |u|_0 < \varepsilon_0,$$

$$(4.50) \quad u_x = v, \quad v_x = w,$$

$$(4.51) \quad u_0 = G(u_1),$$

$$(4.52) \quad F(u_0)v_0 = G'(u_1)F(u_1)v_1,$$

$$(4.53) \quad w_0 = F(G(u_1))^{-1}H(u_1)F(u_1)w_1 + \mathcal{Z}(u_1, v_1),$$

with  $u_0 := u(0)$ ,  $u_1 := u(1)$ ,  $v_0 := v(0)$ , and  $v_1 := v(1)$  as above, and  $w_0 := w(0)$  and  $w_1 := w(1)$ .

The matrix  $S(u)$  is selected as the matrix  $N(F(u))$  of Lemma 4.1 now with  $\Delta := \Lambda^2 D^2 \Lambda$ . Our estimates on  $V_3$  and  $\dot{V}_3$  are in the following lemma.

LEMMA 4.5. *There exists  $\mu_3 > 0$  such that, for every  $\mu \in (0, \mu_3)$ , there exist positive real constants  $\alpha_3, \beta_3, \delta_3$  such that, for every  $(u, v, w) \in \mathcal{V}_3$  such that  $|u|_0 + |v|_0 \leq \delta_3$ , one has*

$$(4.54) \quad \frac{1}{\beta_3} \int_0^1 |w|^2 dx \leq V_3(u, v, w) \leq \beta_3 \int_0^1 |w|^2 dx,$$

$$(4.55) \quad \dot{V}_3(u, v, w) \leq -\alpha_3 V_3(u, v, w) + \beta_3 |v_1|^4 + \beta_3 \int_0^1 (|v|^2 |w| + |w|^2 |v|) dx.$$

*Proof of Lemma 4.5.* Throughout this proof, we assume that  $(u, v, w) \in \mathcal{V}_3$ . By the construction of  $S$ , we have

$$(4.56) \quad \Lambda^{-2} S(0) \Lambda^{-1} = D^2 \in \mathcal{D}_{n,+}$$

and the existence of  $\delta_{31} \in (0, \varepsilon_0/2)$  such that

$$(4.57) \quad S(a) \in \mathcal{S}_{n,+} \text{ and } S(a)F(a) \in \mathcal{S}_{n,+} \quad \forall a \in \mathbb{R}^n \text{ such that } |u| \leq \delta_{31}.$$

Clearly, from (4.57), we obtain that, for every  $\mu > 0$ , there exists  $\beta_3 > 0$  such that (4.54) holds if  $|u|_0 \leq \delta_{31}$ .

Let us now deal with the estimate (4.55) on  $\dot{V}_3 (= \dot{V}_3(u, v, w))$ . Let us decompose  $\dot{V}_3$  in the following way:

$$(4.58) \quad \dot{V}_3 = \mathcal{T}_{31} + \mathcal{T}_{32} + \mathcal{T}_{33},$$

with

$$\begin{aligned} \mathcal{T}_{31} &:= -\mu \int_0^1 \left( w^{\text{tr}} S(u) F(u) w \right) e^{-\mu x} dx, \\ \mathcal{T}_{32} &:= - \int_0^1 \left( w^{\text{tr}} S(u) F(u) w e^{-\mu x} \right)_x dx, \\ \mathcal{T}_{33} &:= - \int_0^1 \left\{ -w^{\text{tr}} \left( [(S(u)F(u))_x v] + [S'(u)F(u)v] \right) w + 2w^{\text{tr}} S(u) [F'(u)w] v \right. \\ &\quad \left. + 4w^{\text{tr}} S(u) [F'(u)v] w + 2w^{\text{tr}} [F''(u)(v, v)] v \right\} e^{-\mu x} dx. \end{aligned}$$

*Analysis of the first term  $\mathcal{T}_{31}$ .* By (4.54) and (4.57), for every  $\mu > 0$ , there exists a positive real constant  $\alpha_3 > 0$  such that, if  $|u|_0 \leq \delta_{31}$ ,

$$(4.59) \quad \mathcal{T}_{31} \leq -\alpha_3 V_3.$$

*Analysis of the second term  $\mathcal{T}_{32}$ .*

$$\begin{aligned} \mathcal{T}_{32} &= - \left[ w^{\text{tr}} S(u) F(u) w e^{-\mu x} \right]_0^1 \\ &= - \left( w_1^{\text{tr}} S(u_1) F(u_1) w_1 e^{-\mu} - w_0^{\text{tr}} S(u_0) F(u_0) w_0 \right). \end{aligned}$$

Under the boundary conditions (4.51) and (4.53), we have, also using (4.47),

$$\begin{aligned} \mathcal{T}_{32} &= -w_1^{\text{tr}} \left( S(u_1) F(u_1) e^{-\mu} \right. \\ &\quad \left. - F(u_1)^{\text{tr}} H(u_1)^{\text{tr}} (F(G(u_1))^{-1})^{\text{tr}} S(G(u_1)) H(u_1) F(u_1) \right) w_1 \\ &\quad + \mathcal{O}(|v_1|^4 + |v_1|^2 |w_1|; |u_1|) \\ &= -w_1^{\text{tr}} \left( S(0) \Lambda e^{-\mu} - \Lambda^2 K^{\text{tr}} \Lambda^{-2} S(0) \Lambda^{-1} K \Lambda^2 \right) w_1 \\ &\quad + \mathcal{O}(|v_1|^4 + |v_1|^2 |w_1| + |w_1|^2 |u_1|; |u_1|). \end{aligned} \quad (4.60)$$

For  $w_1 \in \mathbb{R}^n$ , we define  $\zeta := Dw_1$ . Then, using (4.56), we have, for every  $w_1 \in \mathbb{R}^n$ ,

$$\begin{aligned} w_1^{\text{tr}} K^{\text{tr}} \Lambda^{-2} S(0) \Lambda^{-1} K w_1 &= w_1^{\text{tr}} K^{\text{tr}} D D K w_1 \\ &= (\zeta^{\text{tr}} D^{-1} K^{\text{tr}} D) (D K D^{-1} \zeta) = |D K D^{-1} \zeta|^2. \end{aligned}$$

Therefore, for every  $w_1 \in \mathbb{R}^n$ , we have, using (4.56) once again,

$$w_1^{\text{tr}} K^{\text{tr}} \Lambda^{-2} S(0) \Lambda^{-1} K w_1 \leq \|D K D^{-1}\|^2 \zeta^{\text{tr}} \zeta = \|D K D^{-1}\|^2 w_1^{\text{tr}} \Lambda^{-2} S(0) \Lambda^{-1} w_1.$$

From this inequality and the fact that  $\|DKD^{-1}\|^2 < 1$ , it follows that, choosing  $\mu > 0$  small enough,  $\Lambda^{-2}S(0)\Lambda^{-1}e^{-\mu} - K^{\text{tr}}\Lambda^{-2}S(0)\Lambda^{-1}K$  is a positive definite symmetric matrix, which, in turn, implies that the matrix

$$(4.61) \quad S(0)\Lambda e^{-\mu} - \Lambda^2 K^{\text{tr}} \Lambda^{-2} S(0) \Lambda^{-1} K \Lambda^2$$

is also positive definite. Moreover, for every  $\eta > 0$  and for every  $(v_1, w_1) \in \mathbb{R}^n \times \mathbb{R}^n$ ,

$$|v_1|^2 |w_1| \leq \frac{1}{4\eta} |v_1|^4 + \eta |w_1|^2.$$

Hence, taking  $\eta > 0$  small enough and also using (4.60), one gets the existence of  $\delta_{32} > 0$  and  $\beta_{32} > 0$  such that, if  $|u|_0 + |v|_0 \leq \delta_{32}$ ,

$$(4.62) \quad \mathcal{T}_{32} \leq \beta_{32} |v_1|^4.$$

*Analysis of the third term  $\mathcal{T}_{33}$ .* Note that

$$(F(u)G(u))_x = [F'(u)v]G(u) + F(u)G'(u)v.$$

It follows that there exist  $\delta_{33} > 0$  and  $\beta_{33} > 0$  such that, if  $|u|_0 + |v|_0 \leq \delta_{33}$ , then

$$(4.63) \quad \mathcal{T}_{33} \leq \beta_{33} \int_0^1 (|v|^2 |w| + |w|^2 |v|) dx.$$

Then, collecting inequalities (4.59), (4.62), and (4.63) together, we conclude that if  $|u|_0 + |v|_0 \leq \delta_3 := \min\{\delta_{31}, \delta_{32}, \delta_{33}\}$  and  $\beta_3 := \max\{\beta_{32}, \beta_{33}\}$ , then

$$(4.64) \quad \dot{V}_3 = \mathcal{T}_{31} + \mathcal{T}_{32} + \mathcal{T}_{33} \leq -\alpha_3 V_3 + \beta_3 |v_1|^4 + \beta_3 \int_0^1 (|v|^2 |w| + |w|^2 |v|) dx.$$

This completes the proof of Lemma 4.5.  $\square$

Finally, we deal with  $V$  (see (4.5)) and  $\dot{V}$ , which are now considered as functionals on the set  $\mathcal{V}$  of  $u \in C^3([0, 1], \mathbb{R}^n)$  satisfying (4.49), (4.51), (4.52), and (4.53) with  $v := u_x$  and  $w := u_{xx}$ ,  $u_0 := u(0)$ ,  $u_1 := u(1)$ ,  $v_0 := u_x(0)$ ,  $v_1 := u_x(1)$ ,  $w_0 := u_{xx}(0)$ , and  $w_1 := u_{xx}(1)$ . Of course, we “define”  $\dot{V}$  by  $\dot{V}(u) := \dot{V}_1(u) + \dot{V}_2(u, u_x) + \dot{V}_3(u, u_x, u_{xx})$ . The following lemma holds.

LEMMA 4.6. *Let  $\mu \in (0, \min\{\mu_1, \mu_2, \mu_3\})$ . There exist positive real constants  $\alpha$ ,  $\beta$ , and  $\delta$  such that, for every  $u \in \mathcal{V}$  such that  $|u|_0 + |u_x|_0 \leq \delta$ , we have*

$$(4.65) \quad \frac{1}{\beta} \int_0^1 (|u|^2 + |u_x|^2 + |u_{xx}|^2) dx \leq V(u) \leq \beta \int_0^1 (|u|^2 + |u_x|^2 + |u_{xx}|^2) dx,$$

$$(4.66) \quad \dot{V} \leq -\alpha V.$$

*Proof of Lemma 4.6.* Throughout this proof,  $u$  is assumed to be in  $\mathcal{V}$ . Let  $\bar{\delta} := \min\{\delta_1, \delta_2, \delta_3\}$ ,  $\bar{\alpha} := \min\{\alpha_1, \alpha_2, \alpha_3\}$ , and  $\bar{\beta} := \max\{\beta_1, \beta_2, \beta_3\}$ . It readily follows from (4.11), (4.32), and (4.54) that if  $|u|_0 + |u_x|_0 \leq \bar{\delta}$ , then

$$(4.67) \quad \frac{1}{\bar{\beta}} \int_0^1 (|u|^2 + |u_x|^2 + |u_{xx}|^2) dx \leq V(u) \leq \bar{\beta} \int_0^1 (|u|^2 + |u_x|^2 + |u_{xx}|^2) dx.$$

In order to check (4.66) (for  $\delta > 0$  small enough and  $\beta > 0$  large enough), let us first point out that, for every  $\eta > 0$ ,

$$\begin{aligned} \int_0^1 |u_x|^2 |u_{xx}| dx &\leq \int_0^1 \left( \frac{1}{4\eta} |u_x|^4 + \eta |u_{xx}|^2 \right) dx \\ (4.68) \qquad \qquad \qquad &\leq \frac{1}{4\eta} |u_x|_0^2 \int_0^1 |u_x|^2 dx + \eta \int_0^1 |u_{xx}|^2 dx. \end{aligned}$$

In order to get (4.66), it suffices to use (4.12), (4.33), (4.54), (4.55), (4.67), (4.68) with  $\eta := \alpha_3/(2\beta_3)^2$  and to point out that

$$\begin{aligned} \int_0^1 |u|^2 |u_x| dx &\leq |u_x|_0 \int_0^1 |u|^2 dx, \\ \int_0^1 |u_x|^3 dx &\leq |u_x|_0 \int_0^1 |u_x|^2 dx, \\ \int_0^1 |u_{xx}|^2 |u_x| dx &\leq |u_x|_0 \int_0^1 |u_{xx}|^2 dx. \end{aligned}$$

This concludes the proof of Lemma 4.6.  $\square$

Finally, let us explain how to deduce Theorem 2.3 from Proposition 2.1 and Lemma 4.6. By the Sobolev inequality (see, for instance, [3, Théorème VII, page 129]), there exists  $C > 0$  such that, for every  $u$  in the Sobolev space  $H^2((0, 1), \mathbb{R}^n)$ ,

$$(4.69) \qquad |u|_0 + |u_x|_0 \leq C_0 |u|_{H^2((0,1), \mathbb{R}^n)},$$

with

$$|u|_{H^2((0,1), \mathbb{R}^n)} := \left( \int_0^1 (|u|^2 + |u_x|^2 + |u_{xx}|^2) dx \right)^{1/2}.$$

We choose  $\mu \in (0, \min\{\mu_1, \mu_2, \mu_3\})$ . Let us point out that a simple density argument shows that (4.65) and (4.66) hold for every  $u \in H^2((0, 1), \mathbb{R}^n)$  satisfying (4.51), (4.52), and  $|u|_0 + |u_x|_0 \leq \delta$ . Let

$$(4.70) \qquad \varepsilon := \min \left\{ \frac{\delta}{2C_0\beta}, \frac{\delta_0}{\beta} \right\}.$$

Note that  $\beta \geq 1$ . Using Lemma 4.6, (4.69), and (4.70), the following implications hold for every  $u \in H^2((0, 1), \mathbb{R}^n)$  satisfying (4.51) and (4.52):

$$(4.71) \qquad (|u|_{H^2((0,1), \mathbb{R}^n)} \leq \varepsilon) \Rightarrow \left( |u|_0 + |u_x|_0 \leq \frac{\delta}{2} \text{ and } V(u) \leq \beta \varepsilon^2 \right),$$

$$(4.72) \qquad (|u|_0 + |u_x|_0 \leq \delta \text{ and } V(u) \leq \beta \varepsilon^2) \Rightarrow \left( |u|_0 + |u_x|_0 \leq \frac{\delta}{2} \text{ and } |u|_{H^2((0,1), \mathbb{R}^n)} \leq \delta_0 \right),$$

$$(4.73) \qquad (|u|_0 + |u_x|_0 \leq \delta) \Rightarrow (\dot{V}(u) \leq 0).$$

Now let  $u^0 \in H^2((0, 1), \mathbb{R}^n)$  satisfying (4.2), (4.3), and

$$|u^0|_{H^2((0,1),\mathbb{R}^n)} \leq \varepsilon.$$

Let  $u \in C^0([0, T], H^2((0, 1), \mathbb{R}^n))$  be the maximal classical solution the Cauchy problem (2.5)–(2.7). Using implications (4.71) to (4.73), one gets that

$$(4.74) \quad |u(t, \cdot)|_{H^2((0,1),\mathbb{R}^n)} \leq \delta_0 \quad \forall t \in [0, T],$$

$$(4.75) \quad |u(t, \cdot)|_0 + |u_x(t, \cdot)|_0 \leq \delta \quad \forall t \in [0, T].$$

Using Proposition 2.1 and (4.74), one gets that  $T = +\infty$ . Using Lemma 4.6 and (4.75), one gets that

$$|u(t, \cdot)|_{H^2((0,1),\mathbb{R}^n)}^2 \leq \beta V(u(t, \cdot)) \leq \beta V(u^0) e^{-\alpha t} \leq \beta^2 |u^0|_{H^2((0,1),\mathbb{R}^n)}^2 e^{-\alpha t}.$$

This concludes the proof of Theorem 2.3 when  $m = n$ .  $\square$

Let us now explain the modifications we use in order to deal with the case  $0 < m < n$  (of course, the case  $m = 0$  can be reduced to the case  $m = n$  by considering  $\tilde{u}(t, x) := u(t, 1 - x)$ ).

One first needs the following parametric version of Lemma 4.1.

**LEMMA 4.7.** *Let  $\Lambda := \text{diag}(\Lambda_1, \dots, \Lambda_n) \in \mathcal{D}_n$  be such that (2.4) holds. There exist a positive real number  $\eta$  and a map  $\mathcal{N} : \{M \times \Delta \in \mathcal{M}_{n,n}(\mathbb{R}) \times \mathcal{D}_n; \|M - \Lambda\| < \eta\} \rightarrow \mathcal{S}_n$  of class  $C^\infty$  such that*

$$\mathcal{N}(\Lambda, \Delta) = \Delta \quad \forall \Delta \in \mathcal{D}_n^\rho,$$

$$\mathcal{N}(M, \Delta)M - M^{\text{tr}}\mathcal{N}(M, \Delta) = 0 \quad \forall (M, \Delta) \in \mathcal{M}_{n,n}(\mathbb{R}) \times \mathcal{D}_n \text{ such that } \|M - \Lambda\| < \eta.$$

*Proof of Lemma 4.7.* With the notation of the proof of Lemma 4.1, it suffices to define  $\mathcal{N}(M, D)$  by  $\mathcal{N}(M, D) := \mathcal{L}_M^{-1}(0, \Delta)$ .  $\square$

The Lyapunov function  $V$  now has the following structure:

$$(4.76) \quad V(u, v, w) = V_1(u) + V_2(u, v) + V_3(u, v, w),$$

with

$$(4.77) \quad V_1(u) = \int_0^1 u^{\text{tr}} Q(x, u) u dx,$$

$$(4.78) \quad V_2(u, v) = \int_0^1 v^{\text{tr}} R(x, u) v dx,$$

$$(4.79) \quad V_3(u, v, w) = \int_0^1 w^{\text{tr}} S(x, u) w dx,$$

where  $Q(x, u)$ ,  $R(x, u)$ , and  $S(x, u)$  are symmetric positive definite matrices depending on  $x \in [0, 1]$  defined in the following way. We fix  $D \in \mathcal{D}_{n,+}$  such that  $\|DKD^{-1}\| < 1$ . Let  $\mu \in (0, +\infty)$ , which will be chosen small enough later. Let us recall that  $|\Lambda| = \text{diag}(|\Lambda_1|, \dots, |\Lambda_n|) = \text{diag}(\Lambda_1, \dots, \Lambda_m, |\Lambda_{m+1}|, \dots, |\Lambda_n|)$ .

(i) We define  $Q(x, u)$  by

$$Q(x, u) := \mathcal{N}(F(u), D^2 |\Lambda|^{-1} \text{diag}(e^{-\mu x}, \dots, e^{-\mu x}, e^{\mu x}, \dots, e^{\mu x})).$$

(ii) We define  $R(x, u)$  by

$$R(x, u) := \mathcal{N}(F(u), D^2|\Lambda|\text{diag}(e^{-\mu x}, \dots, e^{-\mu x}, e^{\mu x}, \dots, e^{\mu x})).$$

(iii) Finally, we define  $R(x, u)$  by

$$S(x, u) := \mathcal{N}(F(u), D^2|\Lambda|^3\text{diag}(e^{-\mu x}, \dots, e^{-\mu x}, e^{\mu x}, \dots, e^{\mu x})).$$

(In the above equalities and in the following, in  $\text{diag}(e^{-\mu x}, \dots, e^{-\mu x}, e^{\mu x}, \dots, e^{\mu x})$ ,  $e^{-\mu x}$  is repeated  $m$  times and  $e^{\mu x}$  is repeated  $(n - m)$  times.) In order to deal with the boundary conditions on  $u$  and  $v$ , let us define

$$(4.80) \quad a_0 := \begin{pmatrix} u_+(0) \\ u_-(1) \end{pmatrix}, \quad a_1 := \begin{pmatrix} u_+(1) \\ u_-(0) \end{pmatrix}, \quad b_0 := \begin{pmatrix} v_+(0) \\ v_-(1) \end{pmatrix}, \quad b_1 := \begin{pmatrix} v_+(1) \\ v_-(0) \end{pmatrix}.$$

The boundary condition (4.10) is now (see (2.6))

$$(4.81) \quad a_0 = G(a_1).$$

Now  $\mathcal{V}_1$  is defined as the set of  $u \in C^1([0, 1], \mathbb{R}^n)$  such that (4.81) holds and  $|u|_0 < \varepsilon_0$ .

Clearly, the estimate on  $V_1$  given in Lemma 4.3 still holds. Let us check that the estimate of this lemma on  $\dot{V}_1$  also holds.

The decomposition (4.15)–(4.18) becomes

$$\dot{V}_1 = \mathcal{T}_{11} + \mathcal{T}_{12} + \mathcal{T}_{13},$$

with

$$\mathcal{T}_{11} := \int_0^1 u^{\text{tr}} Q_x(x, u) F(u) u dx,$$

$$\mathcal{T}_{12} := - \int_0^1 (u^{\text{tr}} Q(x, u) F(u) u)_x dx,$$

$$\mathcal{T}_{13} := \int_0^1 u^{\text{tr}} \left( [Q'_u(x, u)v] F(u) + Q(x, u)[F'(u)v] - [Q'_u(x, u)F(u)v] \right) u dx.$$

Noticing that

$$Q_x(x, 0) = -\mu D^2 \Lambda^{-1} \text{diag}(e^{-\mu x}, \dots, e^{-\mu x}, e^{\mu x}, \dots, e^{\mu x}),$$

the term  $\mathcal{T}_{11}$  can be treated as above. Similarly the term  $\mathcal{T}_{13}$  can also be treated as above. Concerning  $\mathcal{T}_{12}$ , one has

$$(4.82) \quad \begin{aligned} \mathcal{T}_{12} &= -u_1 Q(1, u_1) F(u_1) u_1 + u_0 Q(0, u_0) F(u_0) u_0 \\ &= -u_1^{\text{tr}} D^2 |\Lambda|^{-1} \Lambda u_1 + u_0^{\text{tr}} D^2 |\Lambda|^{-1} \Lambda u_0 + \mathcal{O}(|u_1|^3; |u_1|) + \mathcal{O}(\mu |u_1|^2; \mu). \end{aligned}$$

Let

$$\begin{aligned} K_{++} &\in \mathcal{M}_{m,m}(\mathbb{R}), & K_{+-} &\in \mathcal{M}_{m,(n-m)}(\mathbb{R}), \\ K_{-+} &\in \mathcal{M}_{(n-m),n}(\mathbb{R}), & K_{--} &\in \mathcal{M}_{(n-m),(n-m)}(\mathbb{R}) \end{aligned}$$

be such that

$$K = \begin{pmatrix} K_{++} & K_{+-} \\ K_{-+} & K_{--} \end{pmatrix}.$$

Using (4.80) and (4.81), one has

$$(4.83) \quad u_0 = \begin{pmatrix} K_{++} & K_{+-} \\ 0 & \text{Id}_{n-m} \end{pmatrix} a_1 + \mathcal{O}(|a_1|^2; |a_1|), \quad u_1 = \begin{pmatrix} \text{Id}_m & 0 \\ K_{-+} & K_{--} \end{pmatrix} a_1 + \mathcal{O}(|a_1|^2; |a_1|).$$

Using (4.82) and (4.83), straightforward computations lead to

$$(4.84) \quad \mathcal{T}_{12} = -a_1^{\text{tr}}(D^2 - K^{\text{tr}}D^2K)a_1 + \mathcal{O}(|a_1|^3; |a_1|) + \mathcal{O}(\mu|a_1|^2; \mu).$$

However,  $\|DKD^{-1}\| < 1$  implies (and is in fact equivalent to) the property “the symmetric matrix  ${}^{\text{tr}}(D^2 - K^{\text{tr}}D^2K)$  is positive definite,” which, together with (4.84), implies again the existence of  $\delta_{12} > 0$  such that (4.22) holds if  $|a_1| \leq \delta_{12}$ . Hence Lemma 4.3 still holds.

Similarly it can be checked that Lemmas 4.4 and 4.5 also hold, except that in (4.33) and (4.55),  $|v_1|$  has to be replaced by  $|b_1|$  (and the definitions of  $\mathcal{V}_2$  and  $\mathcal{V}_3$  have to be modified in order to deal with the new compatibility conditions). The proof of Theorem 2.3 is then completed as in the case  $m = n$ .

*Remark 4.8.* One can give a lower bound on the exponential decay in Theorem 2.3. Indeed, it follows from our proof of this theorem that, if  $\rho_1(G'(0)) < 1$ , for every  $\nu \in (0, -\min\{|\lambda_1|, \dots, |\lambda_n|\} \ln(\rho_1(G'(0))))$ , there exist  $\varepsilon > 0$  and  $C > 0$  such that, for every  $u^0 \in H^2((0, 1), \mathbb{R}^n)$  satisfying  $|u^0|_{H^2((0, 1), \mathbb{R}^n)} \leq \varepsilon$  and the compatibility conditions (2.8)–(2.10), the classical solution  $u$  to the Cauchy problem (2.5)–(2.7) is defined on  $[0, +\infty)$  and satisfies (2.13).

**5. Conclusion and final remarks.** We have presented a new sufficient condition on the boundary conditions for the exponential stability of one-dimensional nonlinear hyperbolic systems on a bounded interval. Our analysis relies on the construction of an explicit strict Lyapunov function. Moreover, we have compared our sufficient condition with other known sufficient conditions for nonlinear and linear systems. We conclude the paper with two additional comments.

1. The Lyapunov stability analysis presented in this paper can be extended to nonlinear hyperbolic systems of the form

$$(5.1) \quad u_t + F(u)u_x = h(u),$$

i.e., systems having a nonzero right-hand side  $h(u)$  with the map  $h : \mathbb{R}^n \rightarrow \mathbb{R}^n$  of class  $C^2$  vanishing at zero ( $h(0) = 0$ ). Our main theorem (Theorem 2.3) can be extended, in a straightforward way, to system (5.1) with boundary conditions (1.4), provided  $\|h'(0)\|$  is sufficiently small.

2. For the sake of simplicity, we have assumed throughout the paper that the diagonal matrix  $F(0)$  has *distinct* nonzero diagonal entries. It turns out that this assumption may be slightly relaxed when the matrix  $F(u)$  is block-diagonal. Indeed, in such a case, it is sufficient to assume that the  $\Lambda_i$  values are different in each block, but different blocks may share identical  $\Lambda_i$  values. This situation typically occurs when the system  $u_t + F(u)u_x = 0$  is a model

for a network of interconnected  $2 \times 2$  hyperbolic systems. Typical examples are hydraulic networks modeled by Saint Venant equations [7], road networks modeled by Aw–Rascle equations [1, 8], or pipeline networks modeled by isentropic Euler equations [2].

**Appendix A. Some properties of the function  $\rho_1$ .** In this appendix we give some properties which are useful for estimating and computing  $\rho_1$ . Some of these properties are used to prove Proposition 3.7.

PROPOSITION A.1. *Let  $l \in \{1, \dots, n-1\}$ . Let  $K_1 \in \mathcal{M}_{l,l}(\mathbb{R})$ ,  $K_2 \in \mathcal{M}_{l,n-l}(\mathbb{R})$ ,  $K_3 \in \mathcal{M}_{n-l,l}(\mathbb{R})$ ,  $K_4 \in \mathcal{M}_{n-l,n-l}(\mathbb{R})$  and let  $K \in \mathcal{M}_{n,n}(\mathbb{R})$  be defined by*

$$K := \begin{pmatrix} K_1 & K_2 \\ K_3 & K_4 \end{pmatrix}.$$

Then

$$(A.1) \quad \rho_1(K) \geq \max\{\rho_1(K_1), \rho_1(K_4)\}.$$

Moreover, if  $K_2 = 0$  or  $K_3 = 0$ , then

$$(A.2) \quad \rho_1(K) = \max\{\rho_1(K_1), \rho_1(K_4)\}.$$

*Proof of Proposition A.1.* Let  $D \in \mathcal{D}_{n,+}$ . Let  $D_1 \in \mathcal{D}_{l,+}$  and  $D_2 \in \mathcal{D}_{n-l,+}$  be such that

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}.$$

Let

$$M := DKD^{-1}.$$

We have

$$M^{\text{tr}}M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix},$$

with

$$M_{11} := D_1^{-1}K_1^{\text{tr}}D_1^2K_1D_1^{-1} + D_1^{-1}K_3^{\text{tr}}D_2^2K_3D_1^{-1},$$

$$M_{12} := D_1^{-1}K_1^{\text{tr}}D_1^2K_2D_2^{-1} + D_1^{-1}K_3^{\text{tr}}D_2^2K_4D_2^{-1},$$

$$M_{21} := D_2^{-1}K_2^{\text{tr}}D_1^2K_1D_1^{-1} + D_2^{-1}K_4^{\text{tr}}D_2^2K_3D_1^{-1},$$

$$M_{22} := D_2^{-1}K_2^{\text{tr}}D_1^2K_2D_2^{-1} + D_2^{-1}K_4^{\text{tr}}D_2^2K_4D_2^{-1}.$$

For  $X \in \mathbb{R}^l$ , let  $\tilde{X} \in \mathbb{R}^n$  be defined by

$$\tilde{X} := \begin{pmatrix} X \\ 0 \end{pmatrix}.$$



Note that  $|\tilde{X}| = |X|$  and that

$$\begin{aligned}\tilde{X}^{\text{tr}} M^{\text{tr}} M \tilde{X} &= X^{\text{tr}} D_1^{-1} K_1^{\text{tr}} D_1^2 K_1 D_1^{-1} X + X^{\text{tr}} D_1^{-1} K_3^{\text{tr}} D_2^2 K_3 D_1^{-1} X \\ &\geq X^{\text{tr}} D_1^{-1} K_1^{\text{tr}} D_1^2 K_1 D_1^{-1} X.\end{aligned}$$

Hence

$$\begin{aligned}\max\{Z^{\text{tr}} M^{\text{tr}} Z; Z \in \mathbb{R}^n, |Z| = 1\} &\geq \max\{X^{\text{tr}} D_1^{-1} K_1^{\text{tr}} D_1^2 K_1 D_1^{-1} X; X \in \mathbb{R}^l, |X| = 1\} \\ &\geq \rho_1(K_1)^2,\end{aligned}$$

which implies that  $\rho_1(K_1) \leq \rho_1(K)$ . Similarly  $\rho_1(K_4) \leq \rho_1(K)$ . This proves (A.1).

Let us now prove (A.2). We deal only with the case  $K_3 = 0$  (the case  $K_2 = 0$  being similar). Let  $\eta > 0$ . Let  $D_1 \in \mathcal{D}_{l,+}$  and  $D_2 \in \mathcal{D}_{n-l,+}$  be such that

$$(A.3) \quad \|D_1 K_1 D_1^{-1}\| \leq \rho_1(K_1) + \eta, \quad \|D_2 K_4 D_2^{-1}\| \leq \rho_1(K_4) + \eta.$$

Let  $\varepsilon > 0$  and

$$D := \begin{pmatrix} \varepsilon D_1 & 0 \\ 0 & D_2 \end{pmatrix} \in \mathcal{D}_{n,+}, \quad M := D K D^{-1} \in \mathcal{M}_{n,n}(\mathbb{R}).$$

Let  $Z \in \mathbb{R}^n$  and let  $X \in \mathbb{R}^l$  and  $Y \in \mathbb{R}^{n-l}$  be such that

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix}.$$

We have

$$\begin{aligned}Z^{\text{tr}} M^{\text{tr}} M Z &= X^{\text{tr}} D_1^{-1} K_1^{\text{tr}} D_1^2 K_1 D_1^{-1} X + 2\varepsilon X^{\text{tr}} D_1^{-1} K_1^{\text{tr}} D_1^2 K_2 D_2^{-1} Y \\ &\quad + \varepsilon^2 Y^{\text{tr}} D_2^{-1} K_2^{\text{tr}} D_1^2 K_2 D_2^{-1} Y + Y^{\text{tr}} D_2^{-1} K_4^{\text{tr}} D_2^2 K_4 D_2^{-1} Y.\end{aligned}$$

Hence there exists a constant  $C > 0$  independent of  $Z$  and  $\varepsilon > 0$  such that

$$(A.4) \quad Z^{\text{tr}} M^{\text{tr}} M Z \leq (\|D_1 K_1 D_1^{-1}\| |X|)^2 + (\|D_2 K_4 D_2^{-1}\| |Y|)^2 + C\varepsilon |Z|^2.$$

From (2.1), (A.3), and (A.4), we obtain that

$$(A.5) \quad \rho_1(K)^2 \leq \max\{(\rho_1(K_1) + \eta)^2, (\rho_1(K_4) + \eta)^2\} + C\varepsilon.$$

Letting  $\varepsilon \rightarrow 0$  and  $\eta \rightarrow 0$  in (A.5), one gets that  $\rho_1(K)^2 \leq \max\{\rho_1(K_1)^2, \rho_1(K_4)^2\}$ . This concludes the proof of Proposition A.1.  $\square$

**PROPOSITION A.2.** *The map  $\rho_1 : \mathcal{M}_{n,n}(\mathbb{R}) \rightarrow [0, +\infty)$  is continuous.*

*Proof of Proposition A.2.* We proceed by induction on  $n$ . For  $n = 1$  the function  $\rho_1$  satisfies  $\rho(k) = |k|$  for every  $k \in \mathbb{R} = \mathcal{M}_{1,1}(\mathbb{R})$  and is therefore continuous. We now assume that  $\rho_1$  is continuous on  $\mathcal{M}_{p,p}(\mathbb{R})$  for every  $p \in \{1, \dots, n-1\}$  and prove that  $\rho_1$  is continuous on  $\mathcal{M}_{n,n}(\mathbb{R})$ . Since, for every  $D \in \mathcal{D}_{n,+}$ , the function  $K \in \mathcal{M}_{n,n}(\mathbb{R}) \mapsto \|K\| \in \mathbb{R}$  is continuous, it readily follows from (2.1) that  $\rho_1$  is upper semicontinuous on  $\mathcal{M}_{n,n}(\mathbb{R})$ . It remains only to check that  $\rho_1$  is lower semicontinuous.

We argue by contradiction: let  $K \in \mathcal{D}_{n,n}(\mathbb{R})$  and let  $(K_k)_{k \in \mathbb{N}}$  be a sequence of elements of  $\mathcal{M}_{n,n}(\mathbb{R})$  such that

$$(A.6) \quad K_k \rightarrow K \text{ as } k \rightarrow +\infty,$$

$$(A.7) \quad \lim_{k \rightarrow +\infty} \rho_1(K_k) < \rho_1(K).$$

Let  $(D_k)_{k \in \mathbb{N}}$  be a sequence of elements of  $\mathcal{D}_{n,+}$  such that

$$(A.8) \quad \|D_k K_k D_k^{-1}\| \leq \rho_1(K_k) + k^{-1} \quad \forall k \in \mathbb{N} \setminus \{0\}.$$

Note that, denoting by  $(e_1, \dots, e_n)$  the canonical basis of  $\mathbb{R}^n$ ,

$$|A_{ij}| = |e_i^{\text{tr}} A e_j| \leq \|A\| \quad \forall A \in \mathcal{M}_{n,n}(\mathbb{R}), \forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, n\}.$$

Hence, if we denote by  $K_{ijk}$  the term on the  $i$ th line and  $j$ th column of the matrix  $K_k$ ,

$$(A.9) \quad |K_{ijk}| \frac{d_{ik}}{d_{jk}} \leq \|D_k K_k D_k^{-1}\| \quad \forall (i, j) \in \{1, \dots, n\}^2, \forall k \in \mathbb{N},$$

where  $(d_{ik})_{i \in \{1, \dots, n\}}$  is defined by  $D_k = \text{diag}(d_{1k}, \dots, d_{nk})$ . After suitable reorderings (note that  $\rho_1(\Sigma A \Sigma^{-1}) = \rho_1(A)$  for every  $A \in \mathcal{M}_{n,n}(\mathbb{R})$  and for every permutation matrix  $\Sigma$ ) and extracting subsequences if necessary, we may assume without loss of generality that

$$(A.10) \quad d_{1k} \leq d_{2k} \leq \dots \leq d_{(n-1)k} \leq d_{nk} \quad \forall k \in \mathbb{N}.$$

A simple scaling argument also shows that we may assume without loss of generality that

$$(A.11) \quad d_{1k} = 1 \quad \forall k \in \mathbb{N}.$$

Extracting subsequences if necessary, there exist  $l \in \{1, \dots, n\}$ ,  $(d_1, \dots, d_l) \in [1, +\infty)^l$  such that

$$(A.12) \quad d_{ik} \rightarrow d_i \text{ as } k \rightarrow +\infty \quad \forall i \in \{1, \dots, l\},$$

$$(A.13) \quad d_{ik} \rightarrow +\infty \text{ as } k \rightarrow +\infty \quad \forall i \in \{l+1, \dots, n\}.$$

We first treat the case where  $l = n$ . Let  $D := \text{diag}(d_1, \dots, d_n) \in \mathcal{D}_{n,+}$ . From (A.12), we have

$$(A.14) \quad D_k \rightarrow D \text{ as } k \rightarrow +\infty.$$

From (2.1), we have

$$(A.15) \quad \rho_1(K) \leq \|DKD^{-1}\|,$$

which, together with (A.8) and (A.14), implies that

$$\liminf_{k \rightarrow +\infty} \rho_1(K_k) \geq \rho_1(K),$$

in contradiction with (A.7).

It remains to deal with the case where  $l < n$ . Let us denote  $K_{ij}$  the term on the  $i$ th line and  $j$ th column of the matrix  $K$ . From (A.6), (A.7), (A.8), (A.9), (A.12), and (A.13), one gets that

$$(A.16) \quad K_{ij} = 0 \quad \forall (i, j) \in \{l+1, \dots, n\} \times \{1, \dots, l\}.$$

Let  $K^1 \in \mathcal{M}_{l,l}(\mathbb{R})$ ,  $K^2 \in \mathcal{M}_{l,n-l}(\mathbb{R})$ ,  $K^4 \in \mathcal{M}_{n-l,n-l}(\mathbb{R})$  be such that

$$K = \begin{pmatrix} K^1 & K^2 \\ 0 & K^4 \end{pmatrix}.$$

Similarly, for  $k \in \mathbb{N}$ , let  $K_k^1 \in \mathcal{M}_{l,l}(\mathbb{R})$ ,  $K_k^2 \in \mathcal{M}_{l,n-l}(\mathbb{R})$ ,  $K_k^3 \in \mathcal{M}_{n-l,l}(\mathbb{R})$ ,  $K_k^4 \in \mathcal{M}_{n-l,n-l}(\mathbb{R})$  be defined by

$$K := \begin{pmatrix} K_k^1 & K_k^2 \\ K_k^3 & K_k^4 \end{pmatrix}.$$

From (A.2), we have

$$(A.17) \quad \rho_1(K) = \max\{\rho_1(K^1), \rho_1(K^4)\}.$$

From (A.1), we have

$$(A.18) \quad \rho_1(K_k) \geq \max\{\rho_1(K_k^1), \rho_1(K_k^4)\} \quad \forall k \in \mathbb{N}.$$

From our induction hypothesis (the continuity of  $\rho_1$  on  $\mathcal{M}_{p,p}(\mathbb{R})$  for every  $p \in \{1, \dots, n-1\}$ ) and (A.6), we get that

$$\lim_{k \rightarrow +\infty} \rho_1(K_k^1) = \rho_1(K^1), \quad \lim_{k \rightarrow +\infty} \rho_1(K_k^4) = \rho_1(K^4),$$

which, together with (A.17) and (A.18), again leads to a contradiction with (A.7). This concludes the proof of Proposition A.2.  $\square$

Our next proposition shows a case where the value of  $\rho_1(K)$  may be given directly. (For a converse of this proposition, see Proposition B.1.)

**PROPOSITION A.3.** *Let  $l \in \{1, \dots, n\}$ . Let  $(A_j)_{j \in \{1, \dots, l\}}$  and  $(B_j)_{j \in \{1, \dots, l\}}$  be two sequences of vectors in  $\mathbb{R}^n$  such that*

$$(A.19) \quad A_j^{\text{tr}} A_k = B_j^{\text{tr}} B_k \quad \forall (j, k) \in \{1, \dots, l\}^2,$$

$$(A.20) \quad \sum_{j=1}^l A_{ij}^2 = \sum_{j=1}^l B_{ij}^2 \quad \forall i \in \{1, \dots, n\},$$

where  $A_{ij}$  (resp.,  $B_{ij}$ ) is the element on the  $i$ th line of the vector  $A_j$  (resp.,  $B_j$ ). We assume that the  $l$  vectors  $A_1, \dots, A_l$  are linearly independent. Let  $R \geq 0$  and let  $K \in \mathcal{M}_{n,n}(\mathbb{R})$  be such that

$$(A.21) \quad KA_j = RB_j \quad \forall j \in \{1, \dots, l\},$$

$$(A.22) \quad |KX| \leq R|X| \quad \forall X \in \mathbb{R}^n \text{ such that } X^{\text{tr}} A_j = 0 \quad \forall j \in \{1, l\}.$$

Then  $\rho_1(K) = R$ .

*Proof of Proposition A.3.* It readily follows from the assumptions of this proposition that  $\|K\| = R$ . Hence it remains only to check that

$$(A.23) \quad \|DKD^{-1}\| \geq R \quad \forall D \in \mathcal{D}_{n,+}.$$

Let  $D := \text{diag}(D_1, \dots, D_n) \in \mathcal{D}_{n,+}$ . For  $j \in \{1, \dots, l\}$ , let us define

$$E_j := (E_{1j}, \dots, E_{nj})^{\text{tr}} \in \mathbb{R}^n \setminus \{0\}, \quad F_j := (F_{1j}, \dots, F_{nj})^{\text{tr}} \in \mathbb{R}^n \setminus \{0\}$$

by

$$E_j := DA_j, \quad F_j := DB_j.$$

We have, for every  $j \in \{1, \dots, l\}$ ,

$$(A.24) \quad DKD^{-1}E_j = RF_j,$$

$$(A.25) \quad E_{ij} = D_i A_{ij}, \quad F_{ij} = D_i B_{ij} \quad \forall i \in \{1, \dots, n\}.$$

Using (A.20), (A.24), and (A.25), we get

$$\begin{aligned} \sum_{j=1}^l |DKD^{-1}E_j|^2 &= R^2 \sum_{j=1}^l \left( \sum_{i=1}^n F_{ij}^2 \right) \\ &= R^2 \sum_{i=1}^n D_i^2 \left( \sum_{j=1}^l B_{ij}^2 \right) \\ &= R^2 \sum_{i=1}^n D_i^2 \left( \sum_{j=1}^l A_{ij}^2 \right) \\ &= R^2 \sum_{j=1}^l |E_j|^2. \end{aligned}$$

In particular, there exists  $p \in \{1, \dots, l\}$ , such that

$$|DKD^{-1}E_p|^2 \geq R^2 |E_p|^2,$$

which, together with the fact that  $E_p \neq 0$ , implies that  $\|DKD^{-1}\| \geq R$ . This concludes the proof of Proposition A.3.  $\square$

**Appendix B. Proof of Proposition 3.7.** Inequality (3.22) is obvious: indeed, for every  $(\theta_1, \dots, \theta_n)^{\text{tr}} \in \mathbb{R}^n$  and for every  $D \in \mathcal{D}_{n,+}$ ,

$$\begin{aligned} \rho(\text{diag}(e^{\iota\theta_1}, \dots, e^{\iota\theta_n})K) &= \rho(D \text{diag}(e^{\iota\theta_1}, \dots, e^{\iota\theta_n})KD^{-1}) \\ &= \rho(\text{diag}(e^{\iota\theta_1}, \dots, e^{\iota\theta_n})DKD^{-1}) \\ &\leq \|\text{diag}(e^{\iota\theta_1}, \dots, e^{\iota\theta_n})DKD^{-1}\| \\ &\leq \|\text{diag}(e^{\iota\theta_1}, \dots, e^{\iota\theta_n})\| \|DKD^{-1}\| = \|DKD^{-1}\|. \end{aligned}$$

The proof of (3.23) for every  $n \in \{1, 2, 3, 4, 5\}$  is more complicated and relies on various independent propositions. The first proposition provides the converse (up to the  $D$ ) to Proposition A.3 for generic  $K \in \mathcal{M}_{n,n}(\mathbb{R})$ .

PROPOSITION B.1. *Let  $K \in \mathcal{M}_{n,n}(\mathbb{R})$  be such that, for every  $M > 0$ , there exists  $\delta > 0$  such that*

$$(B.1) \quad \left( D := (D_1, \dots, D_n) \in \mathcal{D}_{n,+}, \sum_{i=1}^n D_i = 1, \min\{D_1, \dots, D_n\} < \delta \right) \\ \Rightarrow (\|DKD^{-1}\| > M).$$

(It is easily checked that this property holds, for example, if  $K_{ij} \neq 0$ , for every  $(i, j) \in \{1, \dots, n\}^2$  such that  $i \neq j$ , which is a generic property.) Then there exist  $D \in \mathcal{D}_{n,+}$ , an integer  $l \in \{1, \dots, n\}$ ,  $l$  vectors  $A_j \in \mathbb{R}^n$ ,  $j \in \{1, \dots, l\}$ , and  $l$  vectors  $B_j \in \mathbb{R}^n$ ,  $j \in \{1, \dots, l\}$ , such that (A.19) and (A.20) hold and

$$(B.2) \quad \text{the vectors } A_j \in \mathbb{R}^n, j \in \{1, \dots, l\}, \text{ are linearly independent,}$$

$$(B.3) \quad DKD^{-1}A_j = \rho_1(K)B_j \quad \forall j \in \{1, \dots, l\},$$

$$(B.4) \quad |DKD^{-1}X| \leq \rho_1(K)|X| \quad \forall X \in \mathbb{R}^n.$$

Remark B.2. Proposition B.1 is false if assumption (B.1) is removed. Indeed, let us take  $n = 2$  and

$$K = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then  $\rho_1(K) = 0$ , and it is easily seen that the conclusion of Proposition B.1 does not hold.

*Proof of Proposition B.1.* From (B.1), one gets the existence of  $\tilde{D} \in \mathcal{D}_{n,+}$  such that

$$(B.5) \quad \|\tilde{D}K\tilde{D}^{-1}\| = \rho_1(K).$$

Replacing  $K$  by  $\tilde{D}K\tilde{D}^{-1}$ , we may assume without loss of generality that  $\tilde{D}$  is the identity map  $\text{Id}_n$  of  $\mathbb{R}^n$ . Then

$$(B.6) \quad \|K\| = \rho_1(K).$$

Clearly, (B.1) implies that  $K \neq 0$ , and therefore, by (B.6),

$$(B.7) \quad \rho_1(K) \neq 0.$$

(In fact, if  $K = 0$ , the conclusion of Proposition B.1 obviously holds.) Note that (B.6) implies (B.4) with  $D := \text{Id}_n$ . Let  $p \in \{1, \dots, n\}$  be the dimension of the kernel of  $K^{\text{tr}}K - \rho_1(K)^2 \text{Id}_n$  and let  $(X_1, \dots, X_p)$  be an orthonormal basis of this kernel. For  $j \in \{1, \dots, p\}$ , let  $Y_j := KX_j$ . One has

$$(B.8) \quad |Y_j|^2 = X_j^{\text{tr}}K^{\text{tr}}KX_j = \rho_1(K)^2|X_j|^2 \quad \forall j \in \{1, \dots, p\},$$

$$(B.9) \quad Y_k^{\text{tr}}Y_j = X_k^{\text{tr}}K^{\text{tr}}KX_j = \rho_1(K)^2X_k^{\text{tr}}X_j = 0 \quad \forall (k, j) \in \{1, \dots, p\}^2 \text{ such that } k \neq j.$$

For  $i \in \{1, \dots, n\}$  and  $j \in \{1, \dots, p\}$ , let us denote by  $X_{ij}$  (resp.,  $Y_{ij}$ ) the  $i$ th component of  $X_j$  (resp.,  $Y_j$ ). For  $j \in \{1, \dots, p\}$ , let us denote by  $E_j$  the element of  $\mathbb{R}^n$  whose  $i$ th component is

$$(B.10) \quad E_{ij} := Y_{ij}^2 - X_{ij}^2.$$

Let us assume, for the moment, that

$$(B.11) \quad \forall \tau \in \mathbb{R}^n, \text{ there exists } j \in \{1, \dots, p\} \text{ such that } \tau^{\text{tr}} E_j \geq 0.$$

Applying the separation principle for convex sets to  $\{0\}$  and the convex hull of the vectors  $E_j$ ,  $j \in \{1, \dots, p\}$  (see, e.g., [19, Theorem 3.4(b), page 58]), it follows from (B.11) that  $0 \in \mathbb{R}^n$  is in the convex hull of the vectors  $E_1, \dots, E_p$ : there exist  $p$  nonnegative real numbers  $t_1, \dots, t_p$  such that

$$\sum_{j=1}^p t_j = 1, \quad \sum_{j=1}^p t_j E_j = 0.$$

Let  $l \in \{1, \dots, p\}$  be the number of the  $t_i$ 's which are not equal to 0. Reordering the  $X_i$ 's if necessary, we may assume that

$$t_j > 0 \quad \forall j \in \{1, \dots, l\}, \quad t_j = 0 \quad \forall j \in \{l+1, \dots, p\}.$$

For  $j \in \{1, \dots, l\}$ , we define  $A_j \in \mathbb{R}^n$  and  $B_j \in \mathbb{R}^n$  by

$$(B.12) \quad A_j := \sqrt{t_j} X_j, \quad B_j := \sqrt{t_j} Y_j.$$

Then it is easily checked that the vectors  $A_1, \dots, A_l$  are linearly independent, that (A.19) and (A.20) hold (one even has  $A_k^{\text{tr}} A_j = B_k^{\text{tr}} B_j = 0$  for every  $(k, j) \in \{1, \dots, l\}^2$  such that  $k \neq j$ ), and that (B.3) holds with  $D := \text{Id}_n$ .

It remains only to prove (B.11). Let  $\tau := (\tau_1, \dots, \tau_n)^{\text{tr}} \in \mathbb{R}^n$ . For  $s \in \mathbb{R}$ , let

$$D(s) := \text{diag}(1 + s\tau_1, \dots, 1 + s\tau_n) \in \mathcal{D}_n.$$

For  $s$  small enough,  $D(s) \in \mathcal{D}_{n,+}$ , and therefore, by (B.6),

$$(B.13) \quad \|D(s)KD(s)^{-1}\|^2 \geq \|K\|^2 = \|D(0)KD(0)^{-1}\|^2.$$

Let us estimate the left-hand side of (B.13). By a classical theorem due to Rellich (see, e.g., [18, Theorem XII.3, page 4]) on perturbations of the spectrum of self-adjoint operators, there exist  $\varepsilon > 0$ ,  $p$  real functions  $\lambda_1, \dots, \lambda_p$  of class  $C^1$  from  $(-\varepsilon, \varepsilon)$  into  $\mathbb{R}$ , and  $p$  maps  $x_1, \dots, x_p$  of class  $C^1$  from  $(-\varepsilon, \varepsilon)$  into  $\mathbb{R}^n$  such that

$$(B.14) \quad \lambda_j(0) = \rho_1(K)^2, \quad x_j(0) = X_j \quad \forall j \in \{1, \dots, p\},$$

$$(B.15) \quad D(s)^{-1}K^{\text{tr}}D(s)^2KD(s)^{-1}x_j(s) = \lambda_j(s)x_j(s) \quad \forall s \in (-\varepsilon, \varepsilon), \forall j \in \{1, \dots, p\},$$

$$(B.16) \quad x_j(s)^{\text{tr}}x_j(s) = 1 \quad \forall s \in (-\varepsilon, \varepsilon), \forall j \in \{1, \dots, p\},$$

$$(B.17) \quad x_j(s)^{\text{tr}}x_k(s) = 0 \quad \forall s \in (-\varepsilon, \varepsilon), \forall (j, k) \in \{1, \dots, p\}^p \text{ such that } k \neq j,$$

$$(B.18) \quad \|D(s)KD(s)^{-1}\|^2 = \max\{\lambda_1(s), \dots, \lambda_p(s)\} \quad \forall s \in (-\varepsilon, \varepsilon).$$

Differentiating (B.15) with respect to  $s$  and using (B.10), (B.14), (B.16), and (B.17), one gets

$$(B.19) \quad \lambda'_j(0) = 2\rho_1(K)^2 \tau^{\text{tr}} E_j \quad \forall j \in \{1, \dots, p\}.$$

Property (B.11) follows from (B.7), (B.13), (B.18), and (B.19). This concludes the proof of Proposition B.1.  $\square$

The number  $l$  appearing in Proposition B.1 turns out to be important to compare  $\rho_0$  and  $\rho_1$ : we have the following proposition.

**PROPOSITION B.3.** *Let  $K \in \mathcal{M}_{n,n}(\mathbb{R})$ ,  $D \in \mathcal{D}_{n,+}$ ,  $l \in \{1, \dots, n\}$ ,  $l$  vectors  $A_j \in \mathbb{R}^n$ ,  $j \in \{1, \dots, l\}$ , and  $l$  vectors  $B_j \in \mathbb{R}^n$ ,  $j \in \{1, \dots, l\}$ , be such that (A.20), (B.2), (B.3), and (B.4) hold. If  $l = 1$ , there exist  $X \in \mathbb{R}^n$  and  $\Upsilon := \text{diag}(\Upsilon_1, \dots, \Upsilon_n) \in \mathcal{D}_n$  such that*

$$(B.20) \quad |X| \neq 0,$$

$$(B.21) \quad \Upsilon_i \in \{1, -1\} \quad \forall i \in \{1, \dots, n\},$$

$$(B.22) \quad KX = \rho_1(K)\Upsilon X.$$

If  $l = 2$ , there exist  $X \in \mathbb{C}^n$  and  $(\Upsilon_1, \dots, \Upsilon_n) \in \mathbb{C}^n$  such that

$$(B.23) \quad |X| \neq 0,$$

$$(B.24) \quad |\Upsilon_i| = 1 \quad \forall i \in \{1, \dots, n\},$$

$$(B.25) \quad KX = \rho_1(K)\text{diag}(\Upsilon_1, \dots, \Upsilon_n)X.$$

In both cases ( $l = 1$  or  $l = 2$ ), one has (3.23).

*Proof of Proposition B.3.* Let us first consider the case  $l = 1$ . Let  $i \in \{1, \dots, n\}$ . From (A.20), one has  $|A_{i1}| = B_{i1}$ , and therefore there exists  $\Upsilon_i \in \{-1, 1\}$  such that  $B_{i1} = \varepsilon_i A_{i1}$ . From (B.3), one gets (B.22) if one defines  $X$  by  $X := D^{-1}A_1$ . Let us check that (3.23) holds. Let  $(\theta_1, \dots, \theta_n)^{\text{tr}} \in \mathbb{R}^n$  be defined by

$$\theta_i = 0 \text{ if } \Upsilon_i = 1, \quad \theta_i = -\pi \text{ if } \Upsilon_i = -1.$$

Then (B.22) implies that

$$(B.26) \quad \text{diag}(e^{\iota\theta_1}, \dots, e^{\iota\theta_n})KX = \rho_1(K)X.$$

From (3.21), (B.20), and (B.26), we get that

$$(B.27) \quad \rho_0(K) \geq \rho_1(K),$$

which, together with (3.22), gives (3.23).

Let us now turn to the case  $l = 2$ . Let  $i \in \{1, \dots, n\}$ . From (A.20), one has

$$|A_{i1} + \iota A_{i2}| = |B_{i1} + \iota B_{i2}|,$$

and therefore there exists  $\Upsilon_i \in \mathbb{C}$  such that  $|\Upsilon_i| = 1$  and  $B_{i1} + \iota B_{i2} = \Upsilon_i(A_{i1} + \iota A_{i2})$ . From (B.3), one gets (B.22) if one defines  $X$  by  $X := D^{-1}(A_1 + \iota A_2)$ . Finally, the proof of (3.23) is the same as in the case  $l = 1$ . This concludes the proof of Proposition B.3.  $\square$

The next proposition deals with the case  $n = l$ .

PROPOSITION B.4. Let  $K \in \mathcal{M}_{n,n}(\mathbb{R})$ ,  $D \in \mathcal{D}_{n,+}$ ,  $l \in \{1, \dots, n\}$ ,  $l$  vectors  $A_j \in \mathbb{R}^n$ ,  $j \in \{1, \dots, l\}$ , and  $l$  vectors  $B_j \in \mathbb{R}^n$ ,  $j \in \{1, \dots, l\}$ , be such that (A.20), (B.2), (B.3), and (B.4) hold. If  $l = n$ , there exist  $X \in \mathbb{C}^n$  satisfying (B.23) and  $\theta \in \mathbb{R}$  such that

$$(B.28) \quad KX = e^{-i\theta} \rho_1(K)X$$

and (3.23) again holds.

*Proof of Proposition B.4.* If  $\rho_1(K) = 0$ , then  $K = 0$  and the conclusion of Proposition B.4 holds. If  $\rho_1(K) > 0$ , it follows from (A.19), (B.2), and (B.3) and the assumption  $l = n$  that  $\rho_1(K)^{-1}KDKD^{-1}$  is an isometry. Hence there exist  $Y \in \mathbb{C}^n \setminus \{0\}$  and  $\theta \in \mathbb{R}$  such that  $\rho_1(K)^{-1}KDKD^{-1}Y = e^{-i\theta}Y$ , which implies (B.28) if  $X := D^{-1}Y$ . Finally, (3.23) again follows from (B.23) and (B.28). This concludes the proof of Proposition B.4.  $\square$

Note that  $\rho_0$  is continuous. Hence, from Proposition A.2, Proposition B.1, Proposition B.3, and Proposition B.4, in order to get (3.23) (for every  $n \in \{1, \dots, 5\}$ ) of Proposition 3.7, it remains to address, with the notation of the conclusion of Proposition B.1, the cases  $(l, n) = (3, 4)$ ,  $(l, n) = (3, 5)$ , and  $(l, n) = (4, 5)$ . This is done in the following proposition.

PROPOSITION B.5. Let  $K \in \mathcal{M}_{n,n}(\mathbb{R})$ ,  $D \in \mathcal{D}_{n,+}$ ,  $l \in \{1, \dots, n\}$ ,  $l$  vectors  $A_j \in \mathbb{R}^n$ ,  $j \in \{1, \dots, l\}$ , and  $l$  vectors  $B_j \in \mathbb{R}^n$ ,  $j \in \{1, \dots, l\}$ , be such that (A.20), (B.2), (B.3), and (B.4) hold. If  $(l, n) \in \{(3, 4), (3, 5), (4, 5)\}$ , there exist  $X \in \mathbb{C}^n$  and  $(Y_1, \dots, Y_n) \in \mathbb{C}^n$  such that (B.23), (B.24), and (B.25) hold. In particular, one has (3.23).

*Proof of Proposition B.5.* The fact that (3.23) is implied by the assumptions of Proposition B.5, (B.23), (B.24), and (B.25) has already been pointed out in the proof of Proposition B.3. The case  $(l, n) = (3, 4)$  follows from the case  $(l, n) = (3, 5)$  by replacing  $K \in \mathcal{M}_{4,4}(\mathbb{R})$  by the matrix

$$\tilde{K} := \begin{pmatrix} K & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence we may assume that  $n = 5$ . Taking  $X := D^{-1}(Y_1A_1 + Y_2A_2 + \dots + Y_lA_l)$  it suffices to prove the existence of  $Y := (Y_1, Y_2, \dots, Y_l)^{\text{tr}} \in \mathbb{C}^l \setminus \{0\}$  such that

$$(B.29) \quad |Y_1B_{i1} + Y_2B_{i2} + \dots + Y_lB_{il}|^2 - |Y_1A_{i1} + Y_2A_{i2} + \dots + Y_lA_{il}|^2 = 0 \quad \forall i \in \{1, 2, 3, 4, 5\}.$$

Let us recall that, for  $p \in \mathbb{N} \setminus \{0\}$ ,  $\mathcal{S}_p$  denotes the set of elements  $Q \in \mathcal{M}_{p,p}$  such that  $Q^{\text{tr}} = Q$ . For  $i \in \{1, 2, 3, 4, 5\}$ , there exists a unique  $Q_i \in \mathcal{S}_l$  such that, for every  $Y := (Y_1, Y_2, \dots, Y_l)^{\text{tr}} \in \mathbb{C}^l$ ,

$$Y^{\text{tr}}Q_i\bar{Y}^{\text{tr}} = |Y_1B_{i1} + Y_2B_{i2} + \dots + Y_lB_{il}|^2 - |Y_1A_{i1} + Y_2A_{i2} + \dots + Y_lA_{il}|^2,$$

with  $\bar{Y} := (\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_l)$  ( $\bar{z}$  denoting the complex conjugate of  $z \in \mathbb{C}$ ). Then (B.29) is equivalent to

$$(B.30) \quad Y^{\text{tr}}Q_i\bar{Y}^{\text{tr}} = 0 \quad \forall i \in \{1, 2, 3, 4, 5\}.$$

For a matrix  $M \in \mathcal{M}_{p,p}(\mathbb{C})$ , let us denote by  $\text{tr}(M)$  its trace. Using (A.20) we have that

$$\text{tr}(Q_i) = 0 \quad \forall i \in \{1, 2, 3, 4, 5\}.$$



Using (A.19), one gets that

$$Y^{\text{tr}} Q_1 \bar{Y}^{\text{tr}} + Y^{\text{tr}} Q_2 \bar{Y} + Y^{\text{tr}} Q_3 \bar{Y} + Y^{\text{tr}} Q_4 \bar{Y} + Y^{\text{tr}} Q_5 \bar{Y} = 0 \quad \forall Y \in \mathbb{C}^l.$$

Hence (B.30) is equivalent to

$$Y^{\text{tr}} Q_i \bar{Y}^{\text{tr}} = 0 \quad \forall i \in \{1, 2, 3, 4\}.$$

Therefore Proposition B.5 is a consequence of the following proposition due to Voisin [23].

PROPOSITION B.6. *Let  $l \in \{3, 4\}$ . Let  $Q_1, Q_2, Q_3$ , and  $Q_4$  be four elements of  $\mathcal{S}_l$  such that*

$$(B.31) \quad \text{tr}(Q_i) = 0 \quad \forall i \in \{1, 2, 3, 4\}.$$

*Then there exists  $Y \in \mathbb{C}^l \setminus \{0\}$  such that*

$$(B.32) \quad Y^{\text{tr}} Q_i \bar{Y} = 0 \quad \forall i \in \{1, 2, 3, 4\}.$$

*Proof of Proposition B.6.* We reproduce the proof of [23]. For  $l \in \mathbb{N}$ , let  $\overline{\mathcal{S}_{l,+}}$  be the set of semidefinite positive  $S \in \mathcal{S}_l$ . The first step is the following lemma.

LEMMA B.7. *Let  $l, p$ , and  $n$  be three positive integers. Let  $Q_i, i \in \{1, \dots, n\}$ , be  $n$  elements of  $\mathcal{S}_l$ . Assume that*

$$(B.33) \quad \text{tr}(Q_i) = 0 \quad \forall i \in \{1, \dots, n\},$$

$$(B.34) \quad n < \frac{(p+1)(p+2)}{2} - 1.$$

*Then there exists  $S \in \overline{\mathcal{S}_{l,+}} \setminus \{0\}$  such that*

$$(B.35) \quad \text{the rank of } S \text{ is less than or equal to } p,$$

$$(B.36) \quad \text{tr}(SQ_i) = 0 \quad \forall i \in \{1, \dots, n\}.$$

*Proof of Proposition B.7.* Let

$$C := \{S \in \overline{\mathcal{S}_{l,+}}; \text{tr}(S) = l, \text{tr}(SQ_i) = 0 \quad \forall i \in \{1, \dots, n\}\}.$$

The set  $C$  is a closed convex bounded subset of  $\mathcal{M}_{l,l}(\mathbb{R})$ . By (B.33),  $\text{Id}_l \in C$ , and therefore  $C$  is not empty. Hence, by the Krein–Milman theorem (see, e.g., [19, Theorem 3.21, page 70]), the convex set  $C$  has at least an extreme point. Let  $S$  be an extreme point of  $C$ . Then  $S \in \overline{\mathcal{S}_{l,+}} \setminus \{0\}$  and satisfies (B.36). It remains only to check that (B.35) holds. Let  $k$  be the rank of  $S$ . There exist an orthonormal matrix  $O \in \mathcal{M}_{l,l}(\mathbb{R})$  and a definite positive matrix  $S_0 \in \mathcal{S}_k$  such that

$$(B.37) \quad S = O^{\text{tr}} \begin{pmatrix} S_0 & 0 \\ 0 & 0 \end{pmatrix} O.$$

Let

$$(B.38) \quad \Pi := \left\{ O^{\text{tr}} \begin{pmatrix} S' & 0 \\ 0 & 0 \end{pmatrix} O; S' \in \mathcal{S}_k, \text{tr}(S') = 0 \right\} \subset \mathcal{S}_l.$$

Let us assume that

$$(B.39) \quad n < \frac{k(k+1)}{2} - 1.$$

Since  $\Pi$  is a vector subspace of  $\mathcal{S}_l$  of dimension  $(k(k+1)/2) - 1$ , (B.39) implies that there exists  $S_0 \in \Pi \setminus \{0\}$  such that

$$(B.40) \quad \text{tr} (S_0 Q_i) = 0 \quad \forall i \in \{1, \dots, n\}.$$

Then, for  $\tau \in \mathbb{R}$  with  $|\tau|$  small enough,  $S + \tau S_0$  is in  $C$ , which contradicts the fact that  $S$  is an extreme point of  $C$ . Hence (B.39) does not hold, which, together with (B.34), implies that  $k \leq p$ . This concludes the proof of Lemma B.7.  $\square$

Let us go back to the proof of Proposition B.6. We apply Lemma B.7 with  $n = 4$  and  $p = 2$  (then (B.34) holds). We get the existence of  $S \in \overline{\mathcal{S}_{l,+}} \setminus \{0\}$  satisfying

$$(B.41) \quad \text{the rank of } S \text{ is less than or equal to } 2,$$

$$(B.42) \quad \text{tr} (S Q_i) = 0 \quad \forall i \in \{1, \dots, 4\}.$$

Let  $\lambda_1 > 0$ ,  $\lambda_2 \geq 0$ , and 0 be the eigenvalues of  $S$ . Let

$$S_0 = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \overline{\mathcal{S}_{l,+}}.$$

There exists an orthonormal matrix  $O$  such that

$$(B.43) \quad S = O^{\text{tr}} S_0 O.$$

Let  $Z := (\sqrt{\lambda_1}, \iota\sqrt{\lambda_2}, 0) \in \mathbb{C}^l \setminus \{0\}$  and  $Y := O^{\text{tr}} Z \in \mathbb{C}^l \setminus \{0\}$ . Then, using (B.42) and (B.43), one gets that, for every  $i \in \{1, \dots, 4\}$ ,

$$\begin{aligned} 2Y^{\text{tr}} Q_i \bar{Y} &= \text{tr} ((Y \bar{Y}^{\text{tr}} + \bar{Y} Y)^{\text{tr}} Q_i) = \text{tr} (O^{\text{tr}} (Z \bar{Z}^{\text{tr}} + \bar{Z} Z^{\text{tr}}) O Q_i) \\ &= 2\text{tr} (O^{\text{tr}} S_0^{\text{tr}} O Q_i) = \text{tr} (S Q_i) = 0, \end{aligned}$$

which concludes the proof of Proposition B.6 and therefore the proof of Proposition B.5.  $\square$

Finally, in order to end the proof of Proposition 3.7, it remains only to check that, for  $n = 6$  and therefore for every  $n \geq 6$ , there exists  $K \in \mathcal{M}_{n,n}(\mathbb{R})$  such that  $l = 3$  and (3.24) hold. This is done in the following example.

*Example B.8.* Let  $(u_1, v_1, w_1)^{\text{tr}} \in \mathbb{R}^3$ ,  $(u_2, v_2, w_2)^{\text{tr}} \in \mathbb{R}^3$ ,  $(x_1, y_1, z_1)^{\text{tr}} \in \mathbb{R}^3$ , and  $(x_2, y_2, z_2)^{\text{tr}} \in \mathbb{R}^3$ . We define  $A_1 \in \mathbb{R}^6$ ,  $A_2 \in \mathbb{R}^6$ ,  $A_3 \in \mathbb{R}^6$ ,  $B_1 \in \mathbb{R}^6$ ,  $B_2 \in \mathbb{R}^6$ , and

$B_3 \in \mathbb{R}^6$  by

$$A_1 := \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ u_1 \\ u_2 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ v_1 \\ v_2 \end{pmatrix}, \quad A_3 := \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ w_1 \\ w_2 \end{pmatrix},$$

$$B_1 := \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1/\sqrt{2} \\ x_1 \\ x_2 \end{pmatrix}, \quad B_2 := \begin{pmatrix} 1 \\ 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \\ y_1 \\ y_2 \end{pmatrix}, \quad B_3 := \begin{pmatrix} 0 \\ 1 \\ 1/\sqrt{2} \\ 0 \\ z_1 \\ z_2 \end{pmatrix}.$$

One easily checks that (A.20) holds if (and only if)

$$(B.44) \quad u_1^2 + v_1^2 + w_1^2 = x_1^2 + y_1^2 + z_1^2,$$

$$(B.45) \quad u_2^2 + v_2^2 + w_2^2 = x_2^2 + y_2^2 + z_2^2.$$

Similarly (A.19) holds if (and only if)

$$(B.46) \quad \frac{3}{2} + u_1^2 + u_2^2 - x_1^2 - x_2^2 = 0,$$

$$(B.47) \quad -1 + v_1^2 + v_2^2 - y_1^2 - y_2^2 = 0,$$

$$(B.48) \quad -\frac{1}{2} + w_1^2 + w_2^2 - z_1^2 - z_2^2 = 0,$$

$$(B.49) \quad \frac{1}{2} + u_1 v_1 + u_2 v_2 - x_1 y_1 - x_2 y_2 = 0,$$

$$(B.50) \quad u_1 w_1 + u_2 w_2 - x_1 z_1 - x_2 z_2 = 0,$$

$$(B.51) \quad -\frac{1}{2} + v_1 w_1 + v_2 w_2 - y_1 z_1 - y_2 z_2 = 0.$$

Note that (B.44), (B.46), (B.47), and (B.48) imply (B.45).

We take  $l := 3$  and  $R := 1$ . We define  $K \in \mathcal{M}_{6,6}(\mathbb{R})$  by requiring (A.21) and

$$KX = 0 \quad \forall X \in \mathbb{R}^6 \text{ such that } X^{\text{tr}} A_1 = X^{\text{tr}} A_2 = X^{\text{tr}} A_3 = 0.$$

From Proposition A.3 we get that if (B.44) and (B.46) to (B.51) hold, then

$$\rho_1(K) = 1.$$

Let us assume, for the moment, that (B.46) to (B.51) hold. If (3.24) does not hold, we have  $\rho_0(K) = \rho_1(K) = 1$ , and therefore there exist  $X \in \mathbb{C}^6$  and  $(\Upsilon_1, \dots, \Upsilon_6)^{\text{tr}} \in \mathbb{C}^n$  such that (B.23), (B.24), and (B.25) hold. Clearly,

$$(B.52) \quad |KX| = |X|.$$

Since

$$|K(Y + Z)| = |Y| \quad \forall Y \in \mathbb{C}A_1 + \mathbb{C}A_2 + \mathbb{C}A_3,$$

$$\forall Z \in \mathbb{C}^n \text{ such that } Z^{\text{tr}}A_1 = Z^{\text{tr}}A_2 = Z^{\text{tr}}A_3 = 0,$$

it follows from (B.52) that  $X \in \mathbb{C}A_1 + \mathbb{C}A_2 + \mathbb{C}A_3$ . Hence, there exist  $\xi_1 \in \mathbb{C}$ ,  $\xi_2 \in \mathbb{C}$ , and  $\xi_3 \in \mathbb{C}$  such that

$$(B.53) \quad X = \xi_1 A_1 + \xi_2 A_2 + \xi_3 A_3.$$

Using (B.25), one gets  $(KX)_1 = \Upsilon_1 X_1$  and  $(KX)_2 = \Upsilon_1 X_2$ , which, together with (A.21) and (B.24), imply that

$$(B.54) \quad |\xi_1| = |\xi_2| = |\xi_3|.$$

Using (B.23) and (B.54) one sees that, without loss of generality, we may assume that

$$\xi_1 = 1, \quad |\xi_2| = |\xi_3| = 1.$$

Hence there exist  $\theta_2 \in \mathbb{R}$  and  $\theta_3 \in \mathbb{R}$  such that

$$(B.55) \quad \xi_2 = e^{i\theta_2}, \quad \xi_3 = e^{i\theta_3}.$$

Now using  $|(KX)_3| = |X_3|$ , one gets

$$|\xi_2 + \xi_3| = \sqrt{2},$$

which, together with (B.55), is equivalent to

$$\cos(\theta_3 - \theta_2) = 0;$$

i.e., there exists  $\varepsilon_3 \in \{1, -1\}$  such that

$$(B.56) \quad \xi_3 = \varepsilon_3 \iota \xi_2.$$

Proceeding similarly with the fourth of  $KX$ , one gets the existence of  $\varepsilon_2 \in \{1, -1\}$  such that

$$(B.57) \quad \xi_2 = \varepsilon_2 \iota.$$

Then  $|(KX)_5| = |X_5|$  and  $|(KX)_6| = |X_6|$  are equivalent to

$$(B.58) \quad (u_1 + \varepsilon_1 w_1)^2 + v_1^2 = (x_1 + \varepsilon_1 z_1)^2 + y_1^2,$$

$$(B.59) \quad (u_2 + \varepsilon_1 w_2)^2 + v_2^2 = (x_2 + \varepsilon_1 z_2)^2 + y_2^2$$

with

$$\varepsilon_1 := -\varepsilon_2 \varepsilon_3 \in \{1, -1\}.$$

Let

$$(B.60) \quad \begin{aligned} F : \quad & \mathbb{R}^{12} & \rightarrow & \mathbb{R}^7, \\ P := (u_1, v_1, w_1, x_1, y_1, z_1, u_2, v_2, w_2, x_2, y_2, z_2)^{\text{tr}} & \mapsto & F(P) \end{aligned}$$

be defined by

$$F(P) := \begin{pmatrix} \frac{3}{2} + u_1^2 + u_2^2 - x_1^2 - x_2^2 \\ -1 + v_1^2 + v_2^2 - y_1^2 - y_2^2 \\ -\frac{1}{2} + w_1^2 + w_2^2 - z_1^2 - z_2^2 \\ \frac{1}{2} + u_1 v_1 + u_2 v_2 - x_1 y_1 - x_2 y_2 \\ u_1 w_1 + u_2 w_2 - x_1 z_1 - x_2 z_2 \\ -\frac{1}{2} + v_1 w_1 + v_2 w_2 - y_1 z_1 - y_2 z_2 \\ u_1^2 + v_1^2 + w_1^2 - x_1^2 - y_1^2 - z_1^2 \end{pmatrix}.$$

Let  $\Sigma$  be the subset of  $\mathbb{R}^{12}$  defined by

$$\Sigma := \{P \in \mathbb{R}^{12}; F(P) = 0 \text{ and the rank of } F'(P) \text{ is } 7\}.$$

Let

$$\tilde{P} := \left(0, 1, 0, 1, 0, 0, -\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{3}{4}, \frac{1}{2}, -\frac{1}{4}\right)^{\text{tr}} \in \mathbb{R}^{12}.$$

One easily checks that  $F(\tilde{P}) = 0$ . Straightforward computations give

$$(B.61) \quad F'(\tilde{P}) = \begin{pmatrix} 0 & 0 & 0 & -2 & 0 & 0 & -\frac{1}{2} & 0 & 0 & -\frac{3}{2} & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{2} & 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 & -1 & 0 & \frac{1}{2} & -\frac{1}{4} & 0 & -\frac{1}{2} & -\frac{3}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & \frac{3}{4} & 0 & -\frac{1}{4} & \frac{1}{4} & 0 & -\frac{3}{4} \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \frac{3}{4} & \frac{1}{2} & 0 & \frac{1}{4} & -\frac{1}{2} \\ 0 & 2 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

In particular, the rank of  $F'(\tilde{P})$  is 7. Hence  $\tilde{P}$  is in  $\Sigma$ , and the set  $\Sigma$  is not empty and is a submanifold of  $\mathbb{R}^{12}$  of dimension  $12 - 7 = 5$ . The tangent space to this manifold at  $\tilde{P}$  is  $\text{Ker } F'(\tilde{P})$ . Let  $G_+$  be the map

$$\begin{aligned} G_+ : \quad & \mathbb{R}^{12} \quad \rightarrow \quad \mathbb{R}^2, \\ P := (u_1, v_1, w_1, x_1, y_1, z_1, u_2, v_2, w_2, x_2, y_2, z_2)^{\text{tr}} \quad & \mapsto \quad G_+(P) \end{aligned}$$

defined by

$$G_+(P) := \begin{pmatrix} (u_1 + w_1)^2 + v_1^2 - (x_1 + z_1)^2 - y_1^2 \\ (u_2 + w_2)^2 + v_2^2 - (x_2 + z_2)^2 - y_2^2 \end{pmatrix}.$$

Similarly, let  $G_-$  be the map

$$\begin{aligned} G_- : \quad \mathbb{R}^{12} &\rightarrow \mathbb{R}^2, \\ P := (u_1, v_1, w_1, x_1, y_1, z_1, u_2, v_2, w_2, x_2, y_2, z_2)^{\text{tr}} &\mapsto G_-(P) \end{aligned}$$

defined by

$$G_-(P) := \begin{pmatrix} (u_1 - w_1)^2 + v_1^2 - (x_1 - z_1)^2 - y_1^2 \\ (u_2 - w_2)^2 + v_2^2 - (x_2 - z_2)^2 - y_2^2 \end{pmatrix}.$$

Let  $S_+ \subset \mathbb{R}^{12}$  and  $S_- \subset \mathbb{R}^{12}$  be defined by

$$S_+ := \{P \in \mathbb{R}^{12}; G_+(P) = 0\},$$

$$S_- := \{P \in \mathbb{R}^{12}; G_-(P) = 0\}.$$

It suffices to check that

$$(B.62) \quad \Sigma \text{ is not a subset of } S_- \cup S_+.$$

Note that  $\tilde{P} \in S_- \cap S_+$  and

$$(B.63) \quad G'_-(\tilde{P}) = \begin{pmatrix} 0 & 2 & 0 & -2 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & 1 & 2 & -2 & -1 & 2 \end{pmatrix},$$

$$(B.64) \quad G'_+(\tilde{P}) = \begin{pmatrix} 0 & 2 & 0 & -2 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \end{pmatrix}.$$

In particular the rank of  $G'_-(\tilde{P})$  and the rank of  $G'_+(\tilde{P})$  are both equal to 2. Hence, if  $r > 0$  is small enough, the set  $\{P \in S_-; |P - \tilde{P}| < r\}$  and the set  $\{P \in S_+; |P - \tilde{P}| < r\}$  are submanifolds of  $\mathbb{R}^{12}$  of dimension  $12 - 2 = 10$  whose tangent spaces at  $\tilde{P}$  are  $\text{Ker } G'_-(\tilde{P})$  and  $\text{Ker } G'_+(\tilde{P})$ , respectively. Therefore (B.62) holds if

$$(B.65) \quad \text{Ker } F'(\tilde{P}) \text{ is not a subset of } \text{Ker } G'_-(\tilde{P}) \cup \text{Ker } G'_+(\tilde{P}).$$

Property (B.65) follows from (B.61), (B.63), and (B.64). This concludes the proof of Proposition 3.7.  $\square$

## REFERENCES

- [1] A. AW AND M. RASCLE, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.
- [2] M. K. BANDA, M. HERTY, AND A. KLAR, *Gas flow in pipeline networks*, Netw. Heterog. Media, 1 (2006), pp. 41–56.
- [3] H. BREZIS, *Analyse fonctionnelle Théorie et applications*, Collection Mathématiques Appliquées pour la Maîtrise, Masson, Paris, 1983.
- [4] J.-M. CORON, *On the null asymptotic stabilization of the two-dimensional incompressible Euler equations in a simply connected domain*, SIAM J. Control Optim., 37 (1999), pp. 1874–1896.
- [5] J.-M. CORON, *Control and Nonlinearity*, Math. Surveys Monogr. 136, AMS, Providence, RI, 2007.

- [6] J.-M. CORON, B. D'ANDRÉA-NOVEL, AND G. BASTIN, *A strict Lyapunov function for boundary control of hyperbolic systems of conservation laws*, IEEE Trans. Automat. Control, 52 (2007), pp. 2–11.
- [7] J. DE HALLEUX, C. PRIEUR, J.-M. CORON, B. D'ANDRÉA-NOVEL, AND G. BASTIN, *Boundary feedback control in networks of open channels*, Automatica J. IFAC, 39 (2003), pp. 1365–1376.
- [8] M. GARAVELLO AND B. PICCOLI, *Traffic flow on a road network using the Aw-Rascle model*, Comm. Partial Differential Equations, 31 (2006), pp. 243–275.
- [9] J. M. GREENBERG AND T. T. LI, *The effect of boundary damping for the quasilinear wave equation*, J. Differential Equations, 52 (1984), pp. 66–75.
- [10] J. K. HALE AND S. M. VERDUYN LUNEL, *Introduction to Functional-Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.
- [11] J. K. HALE AND S. M. VERDUYN LUNEL, *Strong stabilization of neutral functional differential equations*, IMA J. Math. Control Inform., 19 (2002), pp. 5–23.
- [12] T. KATO, *The Cauchy problem for quasi-linear symmetric hyperbolic systems*, Arch. Rational Mech. Anal., 58 (1975), pp. 181–205.
- [13] P. D. LAX, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS-NSF Regional Conf. Ser. Appl. Math. 11, SIAM, Philadelphia, 1973.
- [14] T. T. LI, *Global Classical Solutions for Quasilinear Hyperbolic Systems*, RAM Res. Appl. Math. 32, Masson, Paris, 1994.
- [15] T. T. LI AND W. C. YU, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Duke Univ. Math. Ser. V, Duke University Mathematics Department, Durham, NC, 1985.
- [16] A. MAJDA, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Appl. Math. Sci. 53, Springer-Verlag, New York, 1984.
- [17] T. H. QIN, *Global smooth solutions of dissipative boundary value problems for first order quasilinear hyperbolic systems*, Chinese Ann. Math. Ser. B, 6 (1985), pp. 289–298; a Chinese summary appears in Chinese Ann. Math. Ser. A, 6 (1985), p. 514.
- [18] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics. IV. Analysis of Operators*, Academic Press, New York, 1978.
- [19] W. RUDIN, *Functional Analysis*, McGraw-Hill Series in Higher Mathematics, McGraw-Hill, New York, 1973.
- [20] D. SERRE, *Systèmes de Lois de Conservation. I., Hyperbolicité, Entropies, Ondes de Choc*, Fondations, Diderot Editeur, Paris, 1996.
- [21] M. SLEMROD, *Boundary feedback stabilization for a quasilinear wave equation*, in Control Theory for Distributed Parameter Systems and Applications, Lecture Notes in Control and Inform. Sci. 54, Springer-Verlag, Berlin, 1983, pp. 221–237.
- [22] A. TCHOUSO, T. BESSON, AND C.-Z. XU, *Exponential stability of distributed parameter systems governed by symmetric hyperbolic partial differential equations using Lyapunov's second method*, ESAIM Control Optim. Calc. Var., to appear.
- [23] C. VOISIN, *private communication*, 2007.
- [24] C.-Z. XU AND G. SALLET, *Exponential stability and transfer functions of processes governed by symmetric hyperbolic systems*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 421–442.
- [25] Y. C. ZHAO, *The boundary value problem for systems of first-order quasilinear hyperbolic equations*, Chinese Ann. Math. Ser. A, 7 (1986), pp. 629–643; an English summary appears in Chinese Ann. Math. Ser. B, 8 (1987), pp. 127–128.
- [26] L. A. ŽIVOTOVSKIĀ, *Absolute stability of the solutions of differential equations with several lags*, Trudy Sem. Teor. Differencial. Uravneniĭ s Otklon. Argumentom Univ. Družby Narodov Patrisa Lumumby, 7 (1969), pp. 82–91.

## A HYBRID EXTRAGRADIENT-VISCOSITY METHOD FOR MONOTONE OPERATORS AND FIXED POINT PROBLEMS\*

PAUL-EMILE MAINGÉ†

**Abstract.** This paper deals with an iterative method, in a real Hilbert space, for approximating a common element of the set of fixed points of a demicontractive operator (possibly quasi-nonexpansive or strictly pseudocontractive) and the set of solutions of a variational inequality problem for a monotone, Lipschitz continuous mapping. The considered algorithm can be regarded as a combination of a variation of the hybrid steepest descent method and the so-called extragradient method. Under classical conditions, we prove the strong convergence of the sequences of iterates given by the considered scheme.

**Key words.** fixed point problem, variational inequality, monotone mapping, hybrid steepest descent, extragradient method

**AMS subject classifications.** 90C25, 49M45, 65C25

**DOI.** 10.1137/060675319

**1. Introduction.** Variational inequalities as well as fixed point problems are well known to be very useful and efficient tools in mathematics. They provide a unified framework for studying many problems arising in engineering sciences, structural analysis, and other fields (see, e.g., [14, 15, 46, 47, 51, 53]). A closely related subject of current interest is the problem of finding common elements of the set of fixed points of operators and the set of solutions of variational inequalities [35, 48, 54]. The motivation for this subject is mainly due to its possible applications to mathematical modeling of concrete complex problems. Indeed, a classical strategy to construct such mathematical models consists in introducing constraints which can be expressed as subproblems of a more general problem. In some cases, these constraints can be given by variational inequalities [37], by fixed point problems [46, 47], or by problems of different types [1]. The purpose of our work is to propose a strongly convergent iterative method for computing a common element of the set of fixed points of a wide class of operators and the set of solutions of a variational inequality problem for a non-strictly monotone mapping. It is worthwhile recalling that strongly convergent algorithms are of fundamental importance for solving problems in infinite dimensional spaces (see, for instance, [3]).

Throughout this paper,  $\mathcal{H}$  is a real Hilbert space with an inner product  $\langle \cdot, \cdot \rangle$  and its induced norm  $\| \cdot \|$ . Let us denote by  $\Omega$  the set of solutions of the following variational inequality problem, which we abbreviate as  $VIP(A, C)$  (see [14, 15, 24, 43]):

$$(1.1) \quad \text{find } u \in C \quad \text{such that} \quad \langle v - u, Au \rangle \geq 0 \quad \forall v \in C,$$

where  $C$  is a nonempty closed convex set in  $\mathcal{H}$  and  $A : \mathcal{H} \rightarrow \mathcal{H}$  is a monotone mapping on  $C$ , i.e.,

$$\langle Ax - Ay, x - y \rangle \geq 0 \quad \forall x, y \in C.$$

---

\*Received by the editors November 18, 2006; accepted for publication (in revised form) January 20, 2008; published electronically May 14, 2008.

<http://www.siam.org/journals/sicon/47-3/67531.html>

†GRIMAAG, Université des Antilles-Guyane, Département Scientifique Interfacultaire, Campus de Schoelcher, 97230 Cedex, Martinique (F.W.I.) (Paul-Emile.Mainge@martinique.univ-ag.fr).



Also consider an operator  $T : \mathcal{H} \rightarrow \mathcal{H}$  such that  $\text{Fix}(T) \cap \Omega \neq \emptyset$ , where  $\text{Fix}(T)$  is the fixed point set of  $T$ , namely the set of solutions of the following problem:

$$(1.2) \quad \text{find } u \in \mathcal{H} \quad \text{such that} \quad u = Tu.$$

In the present work, we are interested in approximating a common solution of (1.1) and (1.2), which can be obviously written as

$$(1.3) \quad \text{find } u \in \Omega \cap \text{Fix}(T).$$

There is an extensive literature on numerical approaches to solving either variational inequalities (e.g., projection methods [5, 15, 22, 23, 31, 41, 42], merit functions [45], outer approximation methods [9]), or fixed point problems for different classes of mappings (e.g., outer approximation methods [3], viscosity approximation methods [2, 7, 8, 17, 20, 25, 28, 34, 38, 39, 49], relaxed or Mann-type algorithms [10, 26, 30], inertial-type algorithms [29]). A natural way we can attempt to construct a method for solving (1.3) is to combine solution techniques for fixed point problems and those for variational inequalities. Such combinations would probably give rise to different kinds of methods of more or less practical interest regarding convergence properties and requirements on the operators. In the present work, we will restrict ourselves to the study of an iterative process which involves projection methods, relaxation, and viscosity approximations.

An outline of this paper is as follows. In section 2, an overview of some related methods is presented. In section 3, we propose a new extragradient-viscosity algorithm for solving (1.3). Section 4 is devoted to our main convergence result and its proof. In section 5, we apply the main convergence theorem to some specific examples.

**2. Some existing algorithms.** Before we proceed with other related methods, we give some reminders regarding projection-type methods, relaxed (or Mann-type) iterations, and viscosity-type approximations.

On the one hand, projection-type methods related to (1.1) were suggested by its equivalent fixed point formulation:

$$(2.1) \quad \text{find } u \in C \quad \text{such that} \quad u = P_C(u - \lambda Au),$$

where  $\lambda \in (0, +\infty)$  and  $P_C : \mathcal{H} \rightarrow C$  is the metric projection from  $\mathcal{H}$  onto  $C$ , characterized for all  $x \in \mathcal{H}$  by  $P_C(x) \in C$  and  $|P_C(x) - x| = \min_{y \in C} |x - y|$  (also see Remark 4.1). In particular, the simplest of these projection methods was given for  $\mathcal{H} = \mathbb{R}^n$  (finite dimensional case) by the following algorithm ([5, 41]):

$$(2.2) \quad \begin{cases} x_0 \in \mathbb{R}^n, \\ x_{n+1} = P_C(x_n - \lambda Ax_n) \quad \forall n \geq 0, \end{cases}$$

where  $\lambda$  is a judiciously chosen positive stepsize. However, for convergence, this last algorithm requires the restrictive assumption that  $A$  be either strongly monotone and Lipschitz continuous on  $C$ , or co-coercive on  $C$ . Let us recall that  $A$  is called *strongly monotone* on  $C$  if there exists  $\theta > 0$  such that  $\langle Ax - Ay, x - y \rangle \geq \theta |x - y|^2$  for all  $x, y \in C$ ; *Lipschitz continuous* on  $C$  if there exists  $k > 0$  such that  $|Ax - Ay| \leq k|x - y|$  for all  $x, y \in C$ ; *co-coercive* (also called *inverse strongly monotone*) on  $C$  if there exists  $\theta > 0$  such that  $\langle Ax - Ay, x - y \rangle \geq \theta |Ax - Ay|^2$  for all  $x, y \in C$ .

To overcome the drawbacks of (2.2), Korpelevich [23] introduced the following so-called extragradient method (also see [22, 31, 42] for extensions):

$$(2.3) \quad \begin{cases} x_0 \in \mathbb{R}^n, \\ \bar{x}_n = P_C(x_n - \lambda A x_n), \\ x_{n+1} = P_C(x_n - \lambda A \bar{x}_n) \quad \forall n \geq 0, \end{cases}$$

where  $A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is monotone,  $k$ -Lipschitz continuous on  $C$  (i.e.,  $|Ax - Ay| \leq k|x - y|$  for all  $x, y \in C$ ) and  $\lambda \in (0, 1/k)$ . He proved that the two sequences  $(x_n)$  and  $(\bar{x}_n)$  converge to the same point in the solution set  $\Omega$ .

On the other hand, a classical method for computing fixed points of an operator  $T : \mathcal{H} \rightarrow \mathcal{H}$  was given by Mann's iteration [30]:

$$(2.4) \quad x_{n+1} = \alpha_n x_n + (1 - \alpha_n) T x_n,$$

where  $(\alpha_n) \subset (0, 1)$ . It has been applied successfully to various concrete problems in the finite dimensional setting [10]. However, in infinite dimensional spaces, the iterative method (2.4) does not converge strongly in general, even when  $T$  belongs to a relatively small class of nonexpansive mappings (i.e.,  $|Tx - Ty| \leq |x - y|$  for all  $x, y \in \mathcal{H}$ ), except for very restrictive conditions on the operator  $T$  and the space  $\mathcal{H}$ .

As an alternative to Mann's iteration, viscosity approximations were proposed to select a particular fixed point of a nonexpansive operator  $T : \mathcal{H} \rightarrow \mathcal{H}$ , by the following process [2, 7, 8, 17, 25, 49]:

$$(2.5) \quad \begin{cases} x_0 \in \mathcal{H}, \\ x_{n+1} = \alpha_n w + (1 - \alpha_n) T x_n \quad \forall n \geq 0, \end{cases}$$

where  $w$  is any element in  $\mathcal{H}$  and where  $(\alpha_n) \subset (0, 1)$  is a slow decreasing sequence, in the sense that  $\alpha_n \rightarrow 0$  and  $\sum_{n \geq 0} \alpha_n = \infty$ . The iteration (2.5) was introduced first by Halpern [17] in the case when  $w = 0$ . It is well known that, under some additional conditions on  $(\alpha_n)$ , the sequence given by (2.5) converges strongly to  $P_{\text{Fix}(T)}(w)$ , where  $P_{\text{Fix}(T)}$  is the metric projection from  $\mathcal{H}$  onto  $\text{Fix}(T)$ . Later on, similar convergence results were established by Moudafi [34] in the case when the constant element  $w$  is replaced by a more general contraction (also see [20, 28]).

Now, let us recall some existing methods for solving (1.3), such as the following.

(1) A typical algorithm when  $T$  is nonexpansive (i.e.,  $|Tx - Ty| \leq |x - y|$  for all  $x, y \in \mathcal{H}$ ) and  $A$  is  $(1/k)$ -co-coercive (i.e.,  $\langle Ax - Ay, x - y \rangle \geq (1/k)|Ax - Ay|^2$  for all  $x, y \in \mathcal{H}$ ) was discussed by Takahashi and Toyoda [48]:

$$(2.6) \quad \begin{cases} x_0 \in C, \\ x_{n+1} = \alpha_n x_n + (1 - \alpha_n) T P_C(x_n - \lambda_n A x_n) \quad \forall n \geq 0, \end{cases}$$

where  $(\alpha_n) \subset (0, 1)$  and  $(\lambda_n) \subset (0, 2/k)$ . It combines Mann's iteration [30] and the basic projection method. They proved that the sequence  $(x_n)$  generated by (2.6) converges weakly to some element in  $\text{Fix}(T) \cap \Omega$ .

(2) Another iterative process based on Korpelevich's extragradient method [23] and Mann's iteration [30] was proposed later by Nadezhkina and Takahashi [35] when  $T$  is nonexpansive and  $A$  is monotone and  $k$ -Lipschitz continuous (i.e.,  $|Ax - Ay| \leq k|x - y|$  for all  $x, y \in \mathcal{H}$ ):

$$(2.7) \quad \begin{cases} x_0 \in \mathcal{H}, \\ y_n = P_C(x_n - \lambda_n A x_n), \\ x_{n+1} = \alpha_n x_n + (1 - \alpha_n) T P_C(x_n - \lambda_n A y_n) \quad \forall n \geq 0, \end{cases}$$

where  $(\lambda_n) \subset [a, b]$  for some  $a, b \in (0, 1/k)$  and  $(\alpha_n) \subset [c, d]$  for some  $c, d \in (0, 1)$ . They showed that the two sequences  $(x_n)$  and  $(y_n)$  given by (2.7) converge weakly to the same point in  $\text{Fix}(T) \cap \Omega$ . Observe that the assumptions required for convergence are weaker for (2.7) than for (2.6), since co-coercivity implies Lipschitz continuity.

(3) Recently, a variation of Nadezhkina and Takahashi's process [35] was suggested by Zeng and Yao [54]. This last algorithm combines Korpelevich's extragradient method [23] and the Halpern-type iteration [2, 25, 49]. The main result obtained can be summarized as follows.

**THEOREM 2.1** (see [54, Theorem 3.1]). *Let  $C$  be a nonempty closed convex subset of a real Hilbert space  $\mathcal{H}$ . Let  $A : C \rightarrow \mathcal{H}$  be a monotone,  $k$ -Lipschitz continuous mapping and  $T : C \rightarrow C$  be a nonexpansive mapping such that  $\text{Fix}(T) \cap \Omega \neq \emptyset$ . Let  $(x_n)$  and  $(y_n)$  be the sequences generated by*

$$(2.8) \quad \begin{cases} x_0 \in \mathcal{H}, \\ y_n = P_C(x_n - \lambda_n A x_n), \\ x_{n+1} = \alpha_n x_0 + (1 - \alpha_n) T P_C(x_n - \lambda_n A y_n) \quad \forall n \geq 0, \end{cases}$$

where  $(\lambda_n)$  and  $(\alpha_n)$  satisfy the conditions

- (a)  $(k\lambda_n) \subset [a, b]$  for some  $a, b \in (0, 1)$ ;
- (b)  $(\alpha_n) \subset (0, 1)$ ,  $\sum_{n \geq 0} \alpha_n = \infty$ ,  $\alpha_n \rightarrow 0$ .

Then the sequences  $(x_n)$  and  $(y_n)$  converge strongly to the same point  $P_{\text{Fix}(T) \cap \Omega} x_0$  provided  $|x_{n+1} - x_n| \rightarrow 0$ .

*Remark 2.1.* The original condition (a) of Theorem 3.1 in [54] is “ $(k\lambda_n) \subset (0, 1 - \delta)$  for some  $\delta \in (0, 1)$ ,” which allows  $\lambda_n \rightarrow 0$ , but it seems there is something wrong; nevertheless, a correct condition is given in Theorem 2.1.

**3. A new extragradient-viscosity algorithm.** It is noteworthy that the convergence results given by Takahashi and Toyoda [48] for (2.6) do not include important classes of mappings, even in the finite dimensional setting. Consider, for instance, the case of a linear complementary problem, with an associated matrix which is positively semidefinite but not positively definite. It is then easily checked that the corresponding mapping  $A$  is monotone and Lipschitz continuous but not co-coercive. The method (2.7) of Nadezhkina and Takahashi [35] is then applicable to a wider class of problems; however, its efficiency is valid only in finite dimensional spaces, because only weak convergence is obtained. Furthermore, in all the previously mentioned examples, the mapping  $T$  is assumed to be nonexpansive, while certain operators arising for instance in subgradient-projection techniques are not nonexpansive but quasi-nonexpansive (see Definition 3.2 and also, e.g., [3, 29, 52]).

In this paper, we propose an alternative method (algorithm (3.6)) for solving (1.3) in the more general case when  $A$  is monotone and Lipschitz continuous while  $T$  is demicontractive and demiclosed. Regarding this framework,  $\Omega$  is a closed convex set (see [6]), as is  $\text{Fix}(T)$  (see Remark 4.2); hence so is  $\Omega \cap \text{Fix}(T)$ . A strong convergence theorem is also established (Theorem 4.1). For convergence, our method does not require such a restrictive condition as  $|x_{n+1} - x_n| \rightarrow 0$ , which is needed by Zeng and Yao [54]; additionally it covers a much more general class of problems.

Let us recall some definitions of common use in optimization theory.

**DEFINITION 3.1** (see [16]). *A mapping  $T : \mathcal{H} \rightarrow \mathcal{H}$  is called demiclosed if, for any sequence  $(z_k) \subset \mathcal{H}$  and  $z \in \mathcal{H}$ , the following holds:*

$$z_k \rightarrow z \quad \text{weakly}, \quad (I - T)(z_k) \rightarrow 0 \quad \text{strongly} \quad \Rightarrow \quad z \in \text{Fix}(T).$$

DEFINITION 3.2 (see [32, 33, 36]). A mapping  $T : \mathcal{H} \rightarrow \mathcal{H}$  is called (i) quasi-nonexpansive if  $|Tx - q| \leq |x - q|$  for all  $(x, q) \in \mathcal{H} \times \text{Fix}(T)$ ; (ii)  $\rho$ -strictly pseudocontractive if  $|Tx - Ty|^2 \leq |x - y|^2 + \rho|x - y - (Tx - Ty)|^2$  for all  $(x, y) \in \mathcal{H} \times \mathcal{H}$  (for some  $\rho \in [0, 1)$ ); (iii) demicontractive if there exists  $\beta \in [0, 1)$  such that

$$(3.1) \quad |Tx - q|^2 \leq |x - q|^2 + \beta|x - Tx|^2 \quad \forall (x, q) \in \mathcal{H} \times \text{Fix}(T).$$

Remark 3.1. It is easily checked (thanks to (4.3)) that (3.1) can be equivalently rewritten as

$$(3.2) \quad \langle x - Tx, x - q \rangle \geq (1/2)(1 - \beta)|x - Tx|^2 \quad \forall (x, q) \in \mathcal{H} \times \text{Fix}(T).$$

A mapping satisfying (3.1) or (3.2) will be referred to as  $\beta$ -demicontractive. Let us observe that the set of demicontractive operators contains fundamental classes of mappings (nonexpansive, quasi-nonexpansive, and strictly pseudocontractive maps with fixed points), commonly found in various areas of applied mathematics [19, 32, 33]. More precisely, the class of 0-demicontractive mappings contains all the quasi-nonexpansive operators, while this latter set of mappings contains that of nonexpansive operators with fixed points. In addition, it is easily checked that all the nonexpansive operators are  $\rho$ -strictly pseudocontractive, while this latter type of operators (with fixed points) is  $\rho$ -demicontractive. Moreover, the strictly pseudocontractive maps as well as many quasi-nonexpansive operators (including subgradient-projection operators [3, 29]) are well known to be demiclosed.

To select a particular solution of (1.3), we will focus our attention on solving the variational inequality problem  $VIP(\mathcal{F}, \Omega \cap \text{Fix}(T))$ :

$$(3.3) \quad \text{find } x_* \in \text{Fix}(T) \cap \Omega \quad \text{s.t.} \quad \langle v - x_*, \mathcal{F}(x_*) \rangle \geq 0 \quad \forall v \in \text{Fix}(T) \cap \Omega,$$

where  $T$  is assumed to be demicontractive and  $\mathcal{F} : C \rightarrow \mathcal{H}$  satisfies the following two conditions:

- (LC)  $\mathcal{F}$  is  $L$ -Lipschitz continuous (for some  $L \geq 0$ ),  
i.e.,  $|\mathcal{F}(x) - \mathcal{F}(y)| \leq L|x - y| \quad \forall x, y \in C$ ;
- (SM)  $\mathcal{F}$  is  $\eta$ -strongly monotone (for some  $\eta > 0$ ),  
i.e.,  $\langle \mathcal{F}(x) - \mathcal{F}(y), x - y \rangle \geq \eta|x - y|^2 \quad \forall x, y \in C$ .

It is routine to see that the existence and the uniqueness of a solution of (3.3) are ensured by the conditions (LC) and (SM) and by the fact that  $\Omega \cap \text{Fix}(T)$  is a nonempty closed and convex set (thanks to the considered hypothesis).

The algorithm, which we propose in the sequel, is based on the extragradient method and some variant of the following so-called *hybrid steepest descent method* initiated in [51, 52] (also see [50]):

$$(3.4) \quad \begin{cases} x_0 \in \mathcal{H}, \\ x_{n+1} = Tx_n - \alpha_{n+1}\mathcal{F}(Tx_n) \quad \forall n \geq 0, \end{cases}$$

where  $(\alpha_n) \subset (0, 1)$  is a slow decreasing sequence. This latter process was suggested by Yamada [51] as an extension of viscosity approximation methods for solving the variational inequality  $VIP(\mathcal{F}, \text{Fix}(T))$  in the case when  $\mathcal{F}$  is strongly monotone and Lipschitz continuous. Strong convergence theorems were obtained for (3.4) when  $T$  belongs to a subclass of the quasi-nonexpansive maps (so-called *quasi-shrinking operators*). Note that, by letting  $z_n = Tx_n$  ( $(x_n)$  being the sequence given by (3.4)), we immediately obtain

$$(3.5) \quad z_{n+1} = T(z_n - \alpha_{n+1}\mathcal{F}(z_n)).$$

Furthermore, when  $T$  is a projection operator and  $\mathcal{F}$  is the gradient of a real-valued function, (3.5) reduces to the gradient projection method. This latter algorithm is nothing but an extension of the steepest descent method from unconstrained to constrained minimization (see, e.g., [4] for a brief review on this topic).

For solving (3.3), we investigate the asymptotic behavior of the sequence  $(x_n)$  generated, from an arbitrary  $x_0$  in  $\mathcal{H}$ , by the following algorithm:

$$(3.6) \quad \begin{cases} x_0 \in \mathcal{H}, \\ y_n = P_C(x_n - \lambda_n A x_n), \\ t_n = P_C(x_n - \lambda_n A y_n), \\ x_{n+1} := [(1-w)I + wT]v_n, \quad v_n := t_n - \alpha_n \mathcal{F}(t_n), \end{cases}$$

where  $I : \mathcal{H} \rightarrow \mathcal{H}$  is the identity mapping and the parameters are such that  $(\alpha_n) \subset [0, 1)$ ,  $(\lambda_n) \subset (0, \infty)$ , and  $w \in (0, 1)$ . In the special case when  $A = 0$  and  $C = \mathcal{H}$  (hence  $\Omega = \mathcal{H}$ ), (3.6) reduces to a relaxed version of (3.5). In the same frame, (3.3) reduces to  $VIP(\mathcal{F}, \text{Fix}(T))$ , which is exactly the problem considered by Yamada and Ogura [51, 52] in the less general case of certain quasi-nonexpansive mappings. In the case when  $T = I$  and  $\alpha_n = 0$ , (3.6) becomes Korpelevich's extragradient method [23]. Clearly, (3.6) can be regarded as a combination of the extragradient method and a relaxed variant of the hybrid steepest descent method. Note that the relaxation process induced by the mapping  $(1-w)I + wT$  in (3.6) was mainly suggested by the work of Suzuki [44] (see also [19, 30, 51]). It is also interesting to see that the limit attained by the iterates of the method (2.8) is nothing but the solution of (3.3) in the particular case when  $\mathcal{F} := I - x_0$ .

Under classical assumptions on the operators and the parameters, we prove that the sequences  $(x_n)$ ,  $(y_n)$ , and  $(t_n)$  generated by the scheme (3.6) converge strongly to the unique solution of (3.3). Then by the algorithm (3.6), we provide an efficient selecting-type method for solving the initial mixed problem (1.3) for a new broad class of maps. Moreover, the techniques used are simple and different from the usual ones.

**4. Main convergence result.** In this section, we establish our main convergence result given by the following theorem.

**THEOREM 4.1.** *Let  $\beta \in [0, 1)$ . Suppose  $A : \mathcal{H} \rightarrow \mathcal{H}$  is monotone on  $C$  and  $k$ -Lipschitz continuous on  $\mathcal{H}$ . Suppose  $T : \mathcal{H} \rightarrow \mathcal{H}$  is  $\beta$ -demicomtractive, demiclosed with  $\text{Fix}(T) \cap \Omega \neq \emptyset$ . Let  $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$  satisfy (LC) and (SM) and assume that the following conditions hold:*

- (H1)  $w \in (0, \frac{1-\beta}{2}]$ ;
- (H2)  $(\alpha_n) \subset [0, 1)$ ,  $\alpha_n \rightarrow 0$ ;
- (H3)  $(k\lambda_n) \subset [\delta_1, \delta_2]$  (where  $0 < \delta_1 \leq \delta_2 < 1$ );
- (SP)  $\sum_{n \geq 0} \alpha_n = \infty$ .

*Then the sequences  $(x_n)$ ,  $(y_n)$ , and  $(t_n)$  generated by (3.6) converge strongly to  $x_*$ , the unique solution of (3.3).*

Before proving Theorem 4.1, we need a sequence of preliminaries. In particular, we state some crucial inequalities resulting from (3.6). We also prove the boundedness of the sequences generated by (3.6), and we give sufficient conditions so that the weak cluster points of these sequences lies in the solution set  $\Omega \cap \text{Fix}(T)$ .

**4.1. Estimates on the numerical method.** In what follows, we first establish useful inequalities related to the extragradient part and the fixed point part of the method (3.6). To begin with, we recall some classical results related to  $P_C$ , the metric projection from  $\mathcal{H}$  onto  $C$  (a nonempty convex and closed subset of  $\mathcal{H}$ ).

*Remark 4.1.*  $P_C : \mathcal{H} \rightarrow C$  is a nonexpansive operator defined for all  $x \in \mathcal{H}$  by  $P_C x := \operatorname{argmin}_{z \in C} |z - x|$ . Furthermore,  $P_C$  is alternatively characterized by the following inequality (see, for instance, [47])

$$(4.1) \quad \langle x - P_C x, P_C x - y \rangle \geq 0 \quad \forall (x, y) \in \mathcal{H} \times C,$$

or equivalently

$$(4.2) \quad |x - y|^2 \geq |x - P_C x|^2 + |y - P_C x|^2 \quad \forall (x, y) \in \mathcal{H} \times C.$$

The equivalence of (4.1) and (4.2) can be deduced from the obvious equality:

$$(4.3) \quad \langle a, b \rangle = -(1/2)|a - b|^2 + (1/2)|a|^2 + (1/2)|b|^2.$$

The next lemma is concerned with estimates regarding the extragradient part of the method. It is obtained with the same techniques as in [54], and its proof is given for the sake of completeness.

**LEMMA 4.2.** *Suppose  $A : \mathcal{H} \rightarrow \mathcal{H}$  is  $k$ -Lipschitz continuous on  $\mathcal{H}$  and let  $(t_n)$ ,  $(y_n)$ , and  $(x_n)$  be sequences in  $\mathcal{H}$  such that*

$$(4.4) \quad y_n = P_C(x_n - \lambda_n A x_n) \quad \text{and} \quad t_n = P_C(x_n - \lambda_n A y_n).$$

*Then the following inequality holds:*

$$(4.5) \quad |y_n - t_n| \leq k\lambda_n |x_n - y_n|.$$

*If, in addition,  $\Omega \neq \emptyset$  and  $A$  is monotone on  $C$ , then the following holds:*

$$(4.6) \quad |t_n - u|^2 \leq |x_n - u|^2 - (1 - k^2 \lambda_n^2) |x_n - y_n|^2,$$

*where  $u$  is any element in  $\Omega$ .*

*Proof.* From (4.4), we have  $|y_n - t_n| = |P_C(x_n - \lambda_n A x_n) - P_C(x_n - \lambda_n A y_n)|$ ; since  $P_C$  is nonexpansive and  $A$  is assumed to be  $k$ -Lipschitz continuous, we get  $|y_n - t_n| \leq \lambda_n |A x_n - A y_n| \leq k\lambda_n |x_n - y_n|$ , that is, (4.5). Let us prove (4.6). Given  $u \in \Omega$  (hence  $u \in C$ ) and using (4.4), we have  $|t_n - u| = |u - P_C(x_n - \lambda_n A y_n)|$ , which by (4.2) yields  $|t_n - u|^2 \leq |x_n - u - \lambda_n A y_n|^2 - |x_n - t_n - \lambda_n A y_n|^2$ , namely

$$(4.7) \quad |t_n - u|^2 \leq |x_n - u|^2 - |x_n - t_n|^2 + 2\lambda_n \langle A y_n, u - t_n \rangle.$$

The last term in the right-hand side of (4.7) can be rewritten as

$$(4.8) \quad \langle A y_n, u - t_n \rangle = \langle A y_n - A u, u - y_n \rangle + \langle A u, u - y_n \rangle + \langle A y_n, y_n - t_n \rangle.$$

It is easily seen that the first and second terms in the right-hand side of (4.8) are negative, because of the monotonicity of  $A$  on  $C$ , the fact that  $u$  is a solution of  $VIP(A, C)$ , and since  $y_n \in C$ . Clearly, we then deduce  $\langle A y_n, u - t_n \rangle \leq \langle A y_n, y_n - t_n \rangle$ , which by (4.7) amounts to  $|t_n - u|^2 \leq |x_n - u|^2 - |x_n - t_n|^2 + 2\lambda_n \langle A y_n, y_n - t_n \rangle$ , namely

$$(4.9) \quad |t_n - u|^2 \leq |x_n - u|^2 - |x_n - y_n|^2 - |y_n - t_n|^2 + 2\langle x_n - \lambda_n A y_n - y_n, t_n - y_n \rangle.$$

Furthermore, setting  $u_n := x_n - \lambda_n A x_n$  (hence  $y_n := P_C(u_n)$ ), we immediately have

$$\langle x_n - \lambda_n A y_n - y_n, t_n - y_n \rangle = \langle u_n - P_C(u_n), t_n - P_C(u_n) \rangle + \lambda_n \langle A x_n - A y_n, t_n - y_n \rangle.$$

In light of (4.1), we observe that the first term in the right-hand side of this last equality is negative, which entails  $\langle x_n - \lambda_n A y_n - y_n, t_n - y_n \rangle \leq \lambda_n \langle A x_n - A y_n, t_n - y_n \rangle$ . Hence, using the Lipschitz continuity of  $A$ , we obtain

$$\langle x_n - \lambda_n A y_n - y_n, t_n - y_n \rangle \leq (k\lambda_n |x_n - y_n|) \times |t_n - y_n|,$$

which by Young's inequality leads to

$$\langle x_n - \lambda_n A y_n - y_n, t_n - y_n \rangle \leq (1/2)(k\lambda_n)^2 |x_n - y_n|^2 + (1/2)|t_n - y_n|^2.$$

By joining this last inequality to (4.9), we get the desired result.  $\square$

In the following remark, we put out some fundamental properties induced by the relaxation process occurring in the fixed point part of the method.

*Remark 4.2.* Let  $T$  be a  $\beta$ -demicontractive self-mapping on  $\mathcal{H}$  with  $\text{Fix}(T) \neq \emptyset$  and set  $T_w := (1 - w)I + wT$  for  $w \in (0, 1]$ . Then  $T_w$  is quasi-nonexpansive if  $w \in [0, 1 - \beta]$ . Indeed, for any arbitrary element  $(x, q) \in \mathcal{H} \times \text{Fix}(T)$ , we have

$$|T_w x - q|^2 = |x - q|^2 - 2w \langle x - q, x - Tx \rangle + w^2 |Tx - x|^2,$$

which by (3.2) yields

$$(4.10) \quad |T_w x - q|^2 \leq |x - q|^2 - w(1 - \beta - w)|Tx - x|^2.$$

Furthermore, it is routine to see that  $\text{Fix}(T) = \text{Fix}(T_w)$  if  $w \neq 0$ . As a consequence,  $T_w$  is quasi-nonexpansive for  $w \in [0, 1 - \beta]$  and  $\text{Fix}(T)$  is then a closed convex subset of  $\mathcal{H}$  (see [52, Proposition 1], [21, Corollary 1]).

Now, we state our main inequality resulting from the combination of the extra-gradient part and the fixed point part of the method.

**LEMMA 4.3.** *Let  $\beta \in [0, 1)$ . Suppose  $T : \mathcal{H} \rightarrow \mathcal{H}$  is  $\beta$ -demicontractive with  $\text{Fix}(T) \cap \Omega \neq \emptyset$ , and let a mapping  $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$  satisfy (LC) and (SM). Suppose  $A : \mathcal{H} \rightarrow \mathcal{H}$  is monotone on  $C$ ,  $k$ -Lipschitz continuous on  $\mathcal{H}$ . Assume in addition that the following condition holds:*

$$(H1) \quad w \in (0, \frac{1-\beta}{2}].$$

*Then the sequence  $(x_n)$ ,  $(t_n)$ , and  $(y_n)$  given by (3.6) satisfies, for all  $n \geq 0$ ,*

$$(4.11) \quad \begin{aligned} & |x_{n+1} - q|^2 - |x_n - q|^2 \\ & + |x_{n+1} - t_n|^2 + (1 - \lambda_n^2 k^2) |x_n - y_n|^2 \leq -2\alpha_n \langle x_{n+1} - q, \mathcal{F}(t_n) \rangle, \end{aligned}$$

where  $q$  is any element in  $\text{Fix}(T) \cap \Omega$ .

*Proof.* Let  $q \in \text{Fix}(T) \cap \Omega$ . From (3.6) and (4.10), we get

$$(4.12) \quad |x_{n+1} - q|^2 \leq |v_n - q|^2 - w(1 - \beta - w)|v_n - Tv_n|^2,$$

while by (3.6) we have  $Tv_n - v_n = (1/w)(x_{n+1} - v_n)$ . Consequently, setting  $\rho := (1/w)(1 - \beta - w)$ , we obtain

$$(4.13) \quad |x_{n+1} - q|^2 \leq |v_n - q|^2 - \rho |x_{n+1} - v_n|^2.$$

Hence, for  $w \in (0, \frac{1-\beta}{2}]$  (so that  $\rho \geq 1$ ) and recalling that  $v_n = t_n - \alpha_n \mathcal{F}(t_n)$ , we deduce

$$(4.14) \quad \begin{aligned} & |x_{n+1} - q|^2 \leq |v_n - q|^2 - |x_{n+1} - v_n|^2 \\ & = |(t_n - q) - \alpha_n \mathcal{F}(t_n)|^2 - |(t_n - x_{n+1}) - \alpha_n \mathcal{F}(t_n)|^2 \\ & = |t_n - q|^2 - 2\alpha_n \langle x_{n+1} - q, \mathcal{F}(t_n) \rangle - |x_{n+1} - t_n|^2, \end{aligned}$$

which by Lemma 4.2 entails the desired result.  $\square$

The underlying idea for proving the strong convergence of the method (3.6) relies on the estimate (4.11), but we also need to check that the iterates of (3.6) satisfy some other classical properties.

**4.2. Boundedness of the iterates.** In the next lemma, we establish the boundedness of the sequences generated by the method (3.6). This property will be needed, for instance, to ensure the existence of weak cluster points for these sequences.

**LEMMA 4.4.** *Let  $\beta \in [0, 1]$ . Suppose  $T : \mathcal{H} \rightarrow \mathcal{H}$  is  $\beta$ -demicontractive with  $\Omega \cap \text{Fix}(T) \neq \emptyset$ , and let  $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$  satisfy (LC) and (SM). Suppose  $A : \mathcal{H} \rightarrow \mathcal{H}$  is monotone on  $C$  and  $k$ -Lipschitz continuous on  $\mathcal{H}$ . Assume, in addition,  $w \in (0, 1 - \beta]$ ,  $(k\lambda_n) \subset [0, 1]$ , and  $(\alpha_n) \subset [0, \delta)$  (for some small enough  $\delta > 0$ ). Then the sequences  $(x_n)$ ,  $(t_n)$ , and  $(y_n)$  generated by (3.6) are bounded.*

*Proof.* Without loss of generality, we assume  $0 < \eta < L$  (where  $\eta$  and  $L$  are constants given by conditions (SM) and (LC)). Fix  $\mu > 0$  and let  $x, y \in \mathcal{H}$ . From (SM) and (LC), we obtain

$$\begin{aligned} |(\mu\mathcal{F} - I)(x) - (\mu\mathcal{F} - I)(y)|^2 &= \mu^2 |\mathcal{F}(x) - \mathcal{F}(y)|^2 - 2\mu \langle x - y, \mathcal{F}(x) - \mathcal{F}(y) \rangle + |x - y|^2 \\ &\leq \mu^2 L^2 |x - y|^2 - 2\mu\eta |x - y|^2 + |x - y|^2, \end{aligned}$$

so that

$$(4.15) \quad |(\mu\mathcal{F} - I)(x) - (\mu\mathcal{F} - I)(y)| \leq (\sqrt{1 - 2\mu\eta + \mu^2 L^2}) |x - y|.$$

Furthermore, taking  $q \in \text{Fix}(T) \cap \Omega$  and recalling that  $v_{n+1} = t_{n+1} - \alpha_{n+1}\mathcal{F}(t_{n+1})$ , we have

$$\begin{aligned} |v_{n+1} - (q - \alpha_{n+1}\mathcal{F}(q))| &= |(t_{n+1} - \alpha_{n+1}\mathcal{F}(t_{n+1})) - (q - \alpha_{n+1}\mathcal{F}(q))| \\ &= \left| \left(1 - \frac{\alpha_{n+1}}{\mu}\right) (t_{n+1} - q) - \frac{\alpha_{n+1}}{\mu} [(\mu\mathcal{F} - I)(t_{n+1}) - (\mu\mathcal{F} - I)(q)] \right| \\ &\leq \left(1 - \frac{\alpha_{n+1}}{\mu}\right) |t_{n+1} - q| + \frac{\alpha_{n+1}}{\mu} |(\mu\mathcal{F} - I)(t_{n+1}) - (\mu\mathcal{F} - I)(q)|, \end{aligned}$$

provided that  $(\alpha_n) \subset [0, \mu)$ , which by (4.15) yields

$$(4.16) \quad |v_{n+1} - (q - \alpha_{n+1}\mathcal{F}(q))| \leq \left(1 - \frac{\alpha_{n+1}}{\mu}\nu\right) |t_{n+1} - q|,$$

where  $\nu := 1 - \sqrt{1 - 2\mu\eta + \mu^2 L^2}$ . Clearly, we have  $\nu \in (0, 1)$  provided that  $\mu \in (0, \mu_0)$ , where  $\mu_0$  is some small enough positive value. Using (4.6) and observing that  $T_w := (1 - w)I + wT$  is quasi-nonexpansive for  $w \in (0, 1 - \beta]$  (see Remark 4.2), we additionally have

$$(4.17) \quad |t_{n+1} - q| \leq |x_{n+1} - q| = |T_w v_n - q| \leq |v_n - q|,$$

provided that  $(k\lambda_n) \subset [0, 1]$ . Combining (4.16) and (4.17), we then get

$$|v_{n+1} - (q - \alpha_{n+1}\mathcal{F}(q))| \leq \left(1 - \frac{\alpha_{n+1}}{\mu}\nu\right) |v_n - q|.$$



As a consequence, we deduce

$$\begin{aligned} |v_{n+1} - q| &\leq |v_{n+1} - (q - \alpha_{n+1}\mathcal{F}(q))| + |(q - \alpha_{n+1}\mathcal{F}(q)) - q| \\ &\leq \left(1 - \frac{\alpha_{n+1}\nu}{\mu}\right) |v_n - q| + \alpha_{n+1}|\mathcal{F}(q)| \\ &= \left(1 - \frac{\alpha_{n+1}\nu}{\mu}\right) |v_n - q| + \left(\frac{\alpha_{n+1}\nu}{\mu}\right) \left(\frac{\mu|\mathcal{F}(q)|}{\nu}\right), \end{aligned}$$

hence, for all  $n \geq 0$ , we get  $\max\{|v_{n+1} - q|, \frac{\mu|\mathcal{F}(q)|}{\nu}\} \leq \max\{|v_n - q|, \frac{\mu|\mathcal{F}(q)|}{\nu}\}$ , so that  $|v_n - q| \leq \max\{|v_0 - q|, \frac{\mu|\mathcal{F}(q)|}{\nu}\}$ . Thus  $(v_n)$  is bounded, which by (4.17) leads to the boundedness of  $(x_n)$  and  $(t_n)$ . That of  $(y_n)$  is then deduced from the formula  $y_n = P_C(x_n - \lambda_n A x_n)$  (because  $A$  and  $P_C$  are Lipschitz continuous and  $(\lambda_n)$  is assumed to be bounded).  $\square$

**4.3. Preliminary convergence analysis.** In what follows, we are interested in locating the set of weak cluster points of the sequence  $(t_n)$  generated by (3.6). To this end, we recall some results of fundamental importance.

*Remark 4.3.* Let  $N_C : C \rightarrow 2^{\mathcal{H}}$  be the normal cone to  $C$ , defined for all  $x \in C$  by  $N_C x := \{w \in \mathcal{H}; \langle x - z, w \rangle \geq 0 \ \forall z \in C\}$  and let  $B : \mathcal{H} \rightarrow 2^{\mathcal{H}}$  be the mapping defined for any  $x \in \mathcal{H}$  by

$$(4.18) \quad Bx := \begin{cases} Ax + N_C x & \text{if } x \in C, \\ \emptyset & \text{otherwise,} \end{cases}$$

where  $A : C \rightarrow \mathcal{H}$  is assumed to be monotone and Lipschitz continuous. Then  $B$  is maximal monotone (see, for instance, [6, 40, 54]); namely its graph defined by  $G(B) = \{(x, y) \in \mathcal{H} \times \mathcal{H}; y \in Bx\}$  is not properly contained in the graph of any other monotone mapping. In other words, for all  $(x, y) \in \mathcal{H} \times \mathcal{H}$ , there holds

$$\langle x - v, y - w \rangle \geq 0 \quad \forall (v, w) \in G(B) \quad \Rightarrow \quad y \in Bx.$$

Furthermore, as a classical result we have

$$0 \in Bu \quad \Leftrightarrow \quad u \in \Omega.$$

Consequently, we immediately deduce the following key property:

$$(4.19) \quad \langle u - v, -w \rangle \geq 0 \quad \forall (v, w) \in G(B) \quad \Rightarrow \quad u \in \Omega.$$

Throughout the next lemma, we give sufficient conditions ensuring that the weak cluster points of  $(t_n)$  belong to the set  $\Omega$ . This lemma was given implicitly in [54], but its proof needs to be detailed for the sake of completeness.

**LEMMA 4.5.** *Suppose  $A : \mathcal{H} \rightarrow \mathcal{H}$  is monotone and Lipschitz continuous on  $C$ . Let  $(\lambda_n) \subset [\delta, \infty)$  (for some  $\delta > 0$ ) and let  $(t_n)$ ,  $(y_n)$ , and  $(x_n)$  be sequences in  $\mathcal{H}$  such that  $(y_n) \subset C$  and  $t_n = P_C(x_n - \lambda_n A y_n)$ . Assume in addition that there exists a subsequence  $(t_{n_k})$  of  $(t_n)$  such that*

- (i)  $(t_{n_k})$  converges weakly to some  $u$  in  $C$ ;
- (ii)  $|x_{n_k} - y_{n_k}| \rightarrow 0$  and  $|t_{n_k} - y_{n_k}| \rightarrow 0$ .

*Then  $u$  belongs to the set  $\Omega$ .*

*Proof.* Let  $B$  be the mapping defined in (4.18). According to Remark 4.3,  $B$  is maximal monotone, because  $A$  is assumed to be monotone and Lipschitz continuous

on  $C$ , so that the key property (4.19) holds. To reach the conclusion of the lemma, we then prove that the limit  $u$  in condition (i) satisfies this latter property. Given  $(v, w)$  in  $G(B)$  (the graph of  $B$ ), we easily have  $w - Av \in N_C v$ , so that  $\langle v - z, w - Av \rangle \geq 0$  for all  $z \in C$ . In particular, since  $(t_n) \subset C$ , we get

$$(4.20) \quad \langle v - t_{n_k}, w \rangle \geq \langle v - t_{n_k}, Av \rangle.$$

From (4.1), we additionally have  $\langle x_{n_k} - \lambda_{n_k} Ay_{n_k} - t_{n_k}, t_{n_k} - v \rangle \geq 0$ , because  $t_n := P_C(x_n - \lambda_n Ay_n)$  and  $v \in C$ , which by (4.20) yields

$$\begin{aligned} \langle v - t_{n_k}, w \rangle &\geq \langle v - t_{n_k}, Av \rangle - \frac{1}{\lambda_{n_k}} \langle t_{n_k} - v, x_{n_k} - \lambda_{n_k} Ay_{n_k} - t_{n_k} \rangle \\ &= \langle v - t_{n_k}, Av - At_{n_k} \rangle + \langle v - t_{n_k}, At_{n_k} - Ay_{n_k} \rangle - \left\langle v - t_{n_k}, \frac{t_{n_k} - x_{n_k}}{\lambda_{n_k}} \right\rangle; \end{aligned}$$

hence, by the monotonicity of  $A$ , we obtain

$$(4.21) \quad \langle v - t_{n_k}, w \rangle \geq \langle v - t_{n_k}, At_{n_k} - Ay_{n_k} \rangle - \left\langle v - t_{n_k}, \frac{t_{n_k} - x_{n_k}}{\lambda_{n_k}} \right\rangle.$$

Clearly,  $(t_{n_k})$  is a bounded sequence, because it is assumed to be weakly convergent (by condition (i)). Then, in light of condition (ii) (hence  $|At_{n_k} - Ay_{n_k}| \rightarrow 0$  by the Lipschitz continuity of  $A$ ), recalling that  $t_{n_k}$  converges weakly to  $u$ , and passing to the limit in (4.21), we obtain  $\langle v - u, w \rangle \geq 0$  (since  $\lambda_n \geq \delta > 0$ ). Therefore, by the key property (4.19) we conclude that  $u \in \Omega$ .  $\square$

At once, we give sufficient conditions so that any weak cluster-point of the sequence  $(t_n)$  given by (3.6) lies in the solution set  $\text{Fix}(T) \cap \Omega$ .

LEMMA 4.6. *Suppose  $A : \mathcal{H} \rightarrow \mathcal{H}$  is monotone on  $C$  and  $k$ -Lipschitz continuous on  $\mathcal{H}$ ,  $T : \mathcal{H} \rightarrow \mathcal{H}$  is demiclosed with  $\text{Fix}(T) \cap \Omega \neq \emptyset$  and let  $\mathcal{F} : \mathcal{H} \rightarrow \mathcal{H}$  satisfy (LC). Suppose also the following conditions hold:*

(C1)  $(\lambda_n) \subset [\delta, \infty)$  (for some  $\delta > 0$ );

(H2)  $(\alpha_n) \subset [0, 1)$ ,  $\alpha_n \rightarrow 0$ .

*Let  $(t_n)$ ,  $(y_n)$ ,  $(x_n)$  be the sequences generated by (3.6) and assume further that there exists a subsequence  $(t_{n_k})$  of  $(t_n)$  such that*

(i)  $|t_{n_k} - x_{n_k+1}| \rightarrow 0$ ; (ii)  $|x_{n_k} - y_{n_k}| \rightarrow 0$ ; (iii)  $|t_{n_k} - y_{n_k}| \rightarrow 0$ .

*Then any weak cluster-point of  $(t_{n_k})$  belongs to  $\text{Fix}(T) \cap \Omega$ .*

*Proof.* Let  $u \in \mathcal{H}$  be a weak cluster-point of  $(t_{n_k})$ ; that is, there exists a bounded subsequence of  $(t_{n_k})$  (labeled  $(t_{m_k})$ ) which converges weakly to  $u$ . By (i), (ii), and (iii), we also have  $|x_{m_k+1} - t_{m_k}| \rightarrow 0$  and  $|x_{m_k} - y_{m_k}| \rightarrow 0$  and  $|t_{m_k} - y_{m_k}| \rightarrow 0$ . If in addition  $\alpha_n \rightarrow 0$ , we easily deduce that  $v_{m_k} := t_{m_k} - \alpha_{m_k} \mathcal{F}(t_{m_k})$  converges weakly to  $u$  (because  $\mathcal{F}(t_{m_k})$  is bounded thanks to (LC); hence  $\alpha_{m_k} |\mathcal{F}(t_{m_k})| \rightarrow 0$ ), which by (3.6) entails

$$|Tv_{m_k} - v_{m_k}| = (1/w)|x_{m_k+1} - v_{m_k}| = (1/w)|(x_{m_k+1} - t_{m_k}) + \alpha_{m_k} \mathcal{F}(t_{m_k})| \rightarrow 0.$$

As  $T$  is assumed to be demiclosed, we then obtain  $u \in \text{Fix}(T)$ . Furthermore, recalling that  $|x_{m_k} - y_{m_k}| \rightarrow 0$ ,  $|t_{m_k} - y_{m_k}| \rightarrow 0$ ,  $\lambda_n \geq \delta > 0$  and applying Lemma 4.5, we get  $u \in \Omega$ . Consequently, the set of weak cluster-points of  $(t_{n_k})$  is included in  $\Omega \cap \text{Fix}(T)$ , which ends the proof.  $\square$

Now we are in position to prove our main convergence result.

**4.4. Proof of Theorem 4.1.** To prove the strong convergence of the method, we set  $\Gamma_n := |x_n - x_*|^2$  and we focus our analysis on the fact that the real sequence  $(\Gamma_n)$  is either monotonous at infinity (Case 1) or not (Case 2):

• *Case 1.* There exists  $n_0$  such that the sequence  $(\Gamma_n)_{n \geq n_0}$  is either nonincreasing or nondecreasing.

• *Case 2.* For any  $n_0$ , there exist integers  $p \geq n_0$  and  $q \geq n_0$  such that  $\Gamma_p \leq \Gamma_{p+1}$  and  $\Gamma_{q+1} \leq \Gamma_q$ .

More precisely, regarding the situation when  $(\Gamma_n)$  is monotonous at infinity (Case 1) and bounded (hence convergent), we prove that its only possible limit is zero. Concerning the alternative situation (Case 2), we exhibit some subsequence  $(\Gamma_{\phi(n)})$  of  $(\Gamma_n)$  such that  $\Gamma_n \leq \Gamma_{\phi(n)}$  for all  $n \geq m_0$  (where  $m_0$  is some large enough integer) and  $\Gamma_{\phi(n)} \rightarrow 0$  (as  $n \rightarrow \infty$ ).

*Remark 4.4.* In Case 2 above, we know that, for any integer  $n_0$ , there exists another integer  $q$  such that  $q \geq n_0$  and  $\Gamma_q \leq \Gamma_{q+1}$ . Let  $n_0$  be such that  $\Gamma_{n_0} \leq \Gamma_{n_0+1}$  and introduce the set of integers defined for all  $n \geq n_0$  by

$$(4.22) \quad J_n = \{k \in \mathbb{N}; \quad n_0 \leq k \leq n, \quad \Gamma_k \leq \Gamma_{k+1}\}.$$

It is obviously seen that  $J_n \neq \emptyset$ ,  $J_n \subset J_{n+1}$ ,  $1 \leq \text{Card}(J_n) \leq n - n_0 + 1$  and  $\text{Card}(J_n) \rightarrow +\infty$  as  $n \rightarrow \infty$ , where  $\text{Card}(J_n)$  denotes the cardinal of  $J_n$ . Also consider the sequence  $(\tau(n)) \subset \mathbb{N}$  defined for all  $n \geq n_0$  by

$$(4.23) \quad \tau(n) := \max(J_n).$$

Clearly,  $(\tau(n))_{n \geq n_0}$  is a nondecreasing sequence (since  $J_n \subset J_{n+1}$ ) such that  $\tau(n) \rightarrow +\infty$  as  $n \rightarrow +\infty$  and  $\Gamma_{\tau(n)} \leq \Gamma_{\tau(n)+1}$  (for  $n \geq n_0$ ). Furthermore, for all  $n \geq n_0$ , we have either  $\tau(n) = n$  or  $\tau(n) < n$ . The former case yields  $\Gamma_n = \Gamma_{\tau(n)} \leq \Gamma_{\tau(n)+1}$ , while the latter case amounts to  $\Gamma_{\tau(n)+1} > \Gamma_{\tau(n)+2} > \dots > \Gamma_{n-1} > \Gamma_n$  (i.e.,  $\Gamma_j > \Gamma_{j+1}$  for  $\tau(n) + 1 \leq j \leq n - 1$ ), so that  $\Gamma_n \leq \Gamma_{\tau(n)+1}$ . As a consequence, we obtain

$$(4.24) \quad 0 \leq \Gamma_n \leq \Gamma_{\tau(n)+1} \quad \forall n \geq n_0.$$

Then, to prove the strong convergence of the method, it is sufficient to prove that  $\Gamma_{\tau(n)+1} \rightarrow 0$  as  $n \rightarrow \infty$ .

To begin with, the boundedness of  $(x_n)$ ,  $(y_n)$ , and  $(t_n)$  is a straightforward consequence of Lemma 4.4. It is then immediate that there exists a constant  $C \geq 0$  verifying  $|\langle x_{n+1} - x_*, \mathcal{F}(t_n) \rangle| \leq C$  for all  $n \geq 0$ , which by Lemma 4.3 and condition (H3) entails

$$(4.25) \quad \Gamma_{n+1}^2 - \Gamma_n^2 + |x_{n+1} - t_n|^2 + (1 - \delta_2^2)|x_n - y_n|^2 \leq 2C\alpha_n,$$

where  $\delta_2$  is some constant in  $(0, 1)$ . The rest of the proof can be divided into the following two cases.

*Case 1.* Assume  $(\Gamma_n)$  is monotonous at infinity. In other words, for  $n_0$  large enough,  $(\Gamma_n)_{n \geq n_0}$  is either nondecreasing or nonincreasing; hence  $(\Gamma_n)$  is convergent (being also bounded). Set  $\lambda = \lim_{n \rightarrow \infty} |x_n - x_*|$ ; we assume  $\lambda > 0$  and we show that this latter hypothesis is impossible. Clearly, we have  $\Gamma_{n+1}^2 - \Gamma_n^2 \rightarrow 0$  (since the sequence  $(\Gamma_n)$  is convergent), which by (4.25) yields

$$(4.26) \quad |x_{n+1} - t_n| \rightarrow 0 \quad \text{and} \quad |x_n - y_n| \rightarrow 0$$

(since  $\alpha_n \rightarrow 0$  by (H2)); hence by (4.5) we get  $|t_n - y_n| \rightarrow 0$ . Using condition (SM), we additionally have  $\langle t_n - x_*, \mathcal{F}(t_n) - \mathcal{F}(x_*) \rangle \geq \eta |t_n - x_*|^2$ , namely

$$(4.27) \quad \langle x_{n+1} - x_*, \mathcal{F}(t_n) \rangle \geq \eta |t_n - x_*|^2 + \langle t_n - x_*, \mathcal{F}(x_*) \rangle + \langle x_{n+1} - t_n, \mathcal{F}(t_n) \rangle.$$

Let us estimate separately each term in the right-hand side of the above inequality. Since  $(t_n)$  is a bounded sequence, so is the quantity  $\langle t_n - u_*, \mathcal{F}(u_*) \rangle$ . It is then immediate that there exists a subsequence  $(t_{m_k})$  of  $(t_n)$  such that

$$(4.28) \quad \liminf_{n \rightarrow \infty} \langle t_n - u_*, \mathcal{F}(u_*) \rangle = \lim_{k \rightarrow \infty} \langle t_{m_k} - u_*, \mathcal{F}(u_*) \rangle.$$

Observe also that there exists another subsequence of  $(t_{m_k})$  (again denoted  $(t_{m_k})$ ) which converges weakly to some element  $v$  in  $\mathcal{H}$ . Applying Lemma 4.6, we then have  $v \in \text{Fix}(T) \cap \Omega$ . Consequently, by the weak convergence of  $(t_{m_k})$  and recalling that  $u_*$  is solution of (3.3), we get  $\liminf_{k \rightarrow \infty} \langle t_{m_k} - u_*, \mathcal{F}(u_*) \rangle = \langle v - u_*, \mathcal{F}(u_*) \rangle \geq 0$ , which by (4.28) amounts to

$$(4.29) \quad \liminf_{n \rightarrow \infty} \langle t_n - x_*, \mathcal{F}(x_*) \rangle \geq 0.$$

Moreover, observing that  $|x_n - t_n| \rightarrow 0$  (since  $|x_n - y_n| \rightarrow 0$  and  $|t_n - y_n| \rightarrow 0$ ) and recalling that  $\lim_{n \rightarrow \infty} |x_n - x_*| = \lambda$ , we immediately have

$$(4.30) \quad \lim_{n \rightarrow \infty} |t_n - x_*| = \lambda.$$

In addition, as  $(t_n)$  is a bounded sequence,  $|x_{n+1} - t_n| \rightarrow 0$ , and using the Lipschitz continuity of  $\mathcal{F}$  (given by (LC)), we get

$$(4.31) \quad \lim_{n \rightarrow \infty} \langle x_{n+1} - t_n, \mathcal{F}(t_n) \rangle = 0.$$

Therefore, by (4.26)–(4.31), we obtain  $\liminf_{n \rightarrow \infty} \langle x_{n+1} - x_*, \mathcal{F}(t_n) \rangle \geq \eta \lambda^2$ . Thus, for  $\epsilon = (1/2)\eta \lambda^2$  (hence  $\epsilon > 0$ ), there exists some integer  $n_1$  such that, for  $n \geq n_1$ , there holds  $\langle x_{n+1} - x_*, \mathcal{F}(t_n) \rangle \geq \eta \lambda^2 - \epsilon = \frac{1}{2}\eta \lambda^2$ . Then, for  $n \geq n_1$  and taking into account Lemma 4.3, we deduce  $|x_{n+1} - x_*|^2 - |x_n - x_*|^2 \leq -\alpha_n(\eta \lambda^2)$ , which leads to  $|x_{n+1} - x_*|^2 - |x_{n_1} - x_*|^2 \leq -\lambda^2 \eta \sum_{k=n_1}^n \alpha_k$ . As  $\sum_k \alpha_k = \infty$  (by condition (SP)), we observe that this last inequality amounts to  $\lim_{n \rightarrow \infty} |x_{n+1} - x_*|^2 = -\infty$  if  $\lambda > 0$ , which is absurd. As a straightforward consequence, we deduce  $\lambda = 0$ ; namely,  $(x_n)$  converges strongly to  $x_*$  and so do  $(y_n)$  and  $(t_n)$  (since  $|x_n - y_n| \rightarrow 0$  and  $|t_n - y_n| \rightarrow 0$ ).

*Case 2.* Assume  $(\Gamma_n)$  is not monotonous at infinity and let  $(\tau(n))_{n \geq n_0}$  be the sequence introduced in (4.23). As  $\Gamma_{\tau(n)} \leq \Gamma_{\tau(n)+1}$  (for  $n \geq n_0$ ) and referring to (4.25), we get  $|x_{\tau(n)+1} - t_{\tau(n)}|^2 + (1 - \delta_2^2)|x_{\tau(n)} - y_{\tau(n)}|^2 \leq 2C\alpha_{\tau(n)} \rightarrow 0$ , so that

$$(4.32) \quad |x_{\tau(n)+1} - t_{\tau(n)}| \rightarrow 0 \quad \text{and} \quad |x_{\tau(n)} - y_{\tau(n)}| \rightarrow 0;$$

hence by (4.5) we obtain

$$(4.33) \quad |t_{\tau(n)} - y_{\tau(n)}| \rightarrow 0.$$

Now we can prove that the sequence  $(t_{\tau(n)})$  converges strongly to  $x_*$  as  $n \rightarrow \infty$ . Let  $(t_{\tau(n_k)})$  be any subsequence of  $(t_{\tau(n)})$  which converges weakly to some  $q$  in  $\mathcal{H}$  (as  $k \rightarrow \infty$ ). Clearly, the existence of such a subsequence is ensured by the boundedness

of  $(t_{\tau(n)})$ . It is also observed that  $|x_{\tau(n_k)+1} - t_{\tau(n_k)}| \rightarrow 0$ ,  $|x_{\tau(n_k)} - y_{\tau(n_k)}| \rightarrow 0$ , and  $|t_{\tau(n_k)} - y_{\tau(n_k)}| \rightarrow 0$  as  $k \rightarrow \infty$  (thanks to (4.32) and (4.33)), which by Lemma 4.6 amounts to  $q \in \text{Fix}(T) \cap \Omega$ . Furthermore, using Lemma 4.3 and given any  $j \geq 0$ , we have  $\langle x_{j+1} - x_*, \mathcal{F}(t_j) \rangle > 0 \Rightarrow \Gamma_{j+1} < \Gamma_j$ . As a consequence, since  $\Gamma_{\tau(n)} \leq \Gamma_{\tau(n)+1}$ , we get

$$(4.34) \quad \langle x_{\tau(n)+1} - x_*, \mathcal{F}(t_{\tau(n)}) \rangle \leq 0 \quad \forall n \geq n_0.$$

For  $n \geq n_0$  and using (SM), we additionally have

$$\begin{aligned} \eta |t_{\tau(n)} - x_*|^2 &\leq \langle t_{\tau(n)} - x_*, \mathcal{F}(t_{\tau(n)}) - \mathcal{F}(x_*) \rangle \\ &= \langle x_{\tau(n)+1} - x_*, \mathcal{F}(t_{\tau(n)}) \rangle + \langle t_{\tau(n)} - x_{\tau(n)+1}, \mathcal{F}(t_{\tau(n)}) \rangle - \langle t_{\tau(n)} - x_*, \mathcal{F}(x_*) \rangle, \end{aligned}$$

which by (4.34) entails

$$(4.35) \quad |t_{\tau(n)} - x_*|^2 \leq (1/\eta) [\langle t_{\tau(n)} - x_{\tau(n)+1}, \mathcal{F}(t_{\tau(n)}) \rangle - \langle t_{\tau(n)} - x_*, \mathcal{F}(x_*) \rangle].$$

It is also immediate that  $\lim_{k \rightarrow \infty} \langle t_{\tau(n_k)} - x_*, \mathcal{F}(x_*) \rangle = \langle q - x_*, \mathcal{F}(x_*) \rangle$  and that  $\lim_{k \rightarrow \infty} \langle t_{\tau(n_k)} - x_{\tau(n_k)+1}, \mathcal{F}(t_{\tau(n_k)}) \rangle = 0$ , which by (4.35) and (3.3) amounts to

$$\limsup_{k \rightarrow +\infty} |t_{\tau(n_k)} - x_*|^2 \leq -(1/\eta) \langle q - x_*, \mathcal{F}(x_*) \rangle \leq 0;$$

hence  $\lim_{k \rightarrow +\infty} |t_{\tau(n_k)} - x_*| = 0$ . It is then routine to see that the whole sequence  $(t_{\tau(n)})$  converges strongly to  $x_*$ . Therefore, by (4.32), (4.33), and the uniqueness of  $x_*$ , we deduce that  $|x_{\tau(n)} - x_*| \rightarrow 0$  and  $|x_{\tau(n)+1} - x_{\tau(n)}| \rightarrow 0$ . It is then immediate that  $\lim_{n \rightarrow \infty} \Gamma_{\tau(n)+1} = 0$ , which by (4.24) leads to  $\lim_{n \rightarrow \infty} \Gamma_n = 0$ ; that is,  $(x_n)$  converges strongly to  $x_*$ . In light of (4.25) and recalling that  $\Gamma_n \rightarrow 0$ , we also obtain  $|x_n - y_n| \rightarrow 0$  (hence  $|y_n - x_*| \rightarrow 0$ ), which by (4.5) entails  $|t_n - y_n| \rightarrow 0$  (hence  $|t_n - x_*| \rightarrow 0$ ).  $\square$

**5. Applications.** In the same manner as in [35, 54], we derive two applications of Theorem 4.1.

**THEOREM 5.1.** *Suppose  $A : \mathcal{H} \rightarrow \mathcal{H}$  is monotone and  $k$ -Lipschitz continuous on  $\mathcal{H}$  and  $T : \mathcal{H} \rightarrow \mathcal{H}$  is nonexpansive with  $\text{Fix}(T) \cap A^{-1}(0) \neq \emptyset$ . Let  $(x_n)$ ,  $(y_n)$ , and  $(t_n)$  be the sequences generated by*

$$(5.1) \quad \begin{cases} x_0 \in \mathcal{H}, \\ y_n = x_n - \lambda_n A x_n, \\ t_n = x_n - \lambda_n A y_n, \\ x_{n+1} := [(1-w)I + wT]v_n, \quad v_n := \alpha_n u + (1 - \alpha_n)t_n, \end{cases}$$

where  $u$  is any element in  $\mathcal{H}$  and where  $(\alpha_n)$ ,  $(\lambda_n)$ , and  $w$  satisfy the following conditions:

- (H1)'  $w \in (0, \frac{1}{2}]$ ;
- (H2)  $(\alpha_n) \subset [0, 1)$ ,  $\alpha_n \rightarrow 0$ ;
- (H3)  $(k\lambda_n) \subset [\delta_1, \delta_2]$  (where  $0 < \delta_1 \leq \delta_2 < 1$ );
- (SP)  $\sum_{n \geq 0} \alpha_n = \infty$ .

Then the sequences  $(x_n)$ ,  $(y_n)$ , and  $(t_n)$  converge strongly to  $P_{\text{Fix}(T) \cap A^{-1}(0)}(u)$  (i.e., the nearest element of  $u$  in  $\text{Fix}(T) \cap A^{-1}(0)$ ).

*Proof.* This result is a straightforward consequence of Theorem 4.1 with  $C := \mathcal{H}$  (hence  $\Omega = A^{-1}(0)$  and  $P_C = I$ ),  $\mathcal{F} := I - u$  (hence  $\mathcal{F}$  is Lipschitz continuous and strongly monotone) and recalling that a nonexpansive mapping is 0-demicontractive and demiclosed.  $\square$

The second application is related to the approximation of common zeroes of monotone operators (see also [27]).

**THEOREM 5.2.** *Let  $A : \mathcal{H} \rightarrow \mathcal{H}$  be monotone and  $k$ -Lipschitz continuous mapping on  $\mathcal{H}$  and let  $J_r^D := (I + \lambda D)^{-1}$  (for  $r > 0$ ) be the resolvent of  $D : \mathcal{H} \rightarrow 2^{\mathcal{H}}$ , a maximal monotone mapping such that  $A^{-1}(0) \cap D^{-1}(0) \neq \emptyset$ . Let  $(x_n)$ ,  $(y_n)$ , and  $(t_n)$  be the sequences generated by*

$$(5.2) \quad \begin{cases} x_0 \in \mathcal{H}, \\ y_n = x_n - \lambda_n A x_n, \\ t_n = x_n - \lambda_n A y_n, \\ x_{n+1} := [(1-w)I + wJ_r^D]v_n, \quad v_n := \alpha_n u + (1 - \alpha_n)t_n, \end{cases}$$

where  $u \in \mathcal{H}$  and  $(\alpha_n)$ ,  $(\lambda_n)$ , and  $w$  satisfy the conditions (H1)', (H2), (H3), and (SP) in Theorem 5.1.

Then the sequences  $(x_n)$ ,  $(y_n)$ , and  $(t_n)$  converge strongly to  $P_{D^{-1}(0) \cap A^{-1}(0)}(u)$  (i.e., the nearest element of  $u$  in  $D^{-1}(0) \cap A^{-1}(0)$ ).

*Proof.* This result is immediately deduced from Theorem 4.1 with  $C := \mathcal{H}$  (hence  $\Omega = A^{-1}(0)$  and  $P_C = I$ ),  $\mathcal{F} := I - u$  (hence  $\mathcal{F}$  is Lipschitz continuous and strongly monotone) and recalling that  $J_r^D$  is a nonexpansive mapping (hence 0-demicontractive and demiclosed) such that  $\text{Fix}(J_r^D) = D^{-1}(0)$ .  $\square$

**6. Conclusion.** The purpose of this article is to establish the convergence in norm of some new splitting iterative methods for finding solutions of infinite dimensional problems involving both monotone variational inequalities and fixed point problems. These algorithms are based on known processes (extragradient, relaxation, viscosity), and the main theorem (Theorem 4.1) improves, in either form or requirements of operators, most of the existing known results on this subject, including some recent ones: Nadezhkina and Takahashi [35], Zeng and Yao [54]. To the best of our knowledge, there is no existing result that is an equivalent of Theorem 4.1 in terms of a continuous dynamical system. Note also that the (demicontractive) operators occurring in the considered fixed point problems are not necessarily continuous (contrary to nonexpansive mappings which are Lipschitz continuous), so that our analysis is completely different from usual analyses regarding viscosity approximations. Moreover, it is very probable that the techniques of analysis proposed in this paper may pave the way for forthcoming developments regarding more general equilibrium [12] and fixed point problems, especially in designing new algorithms including inexact or perturbed technology [11, 13, 18] as well as inertial-type extrapolation [29].

## REFERENCES

- [1] A. S. ANTIPIN, *Solution methods for variational inequalities with coupled constraints*, Comput. Math. Math. Phys., 40 (2000), pp. 1291–1307.
- [2] H. H. BAUSCHKE, *The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space*, J. Math. Anal. Appl., 202 (1996), pp. 150–159.
- [3] H. H. BAUSCHKE AND P. L. COMBETTES, *A weak-to-strong convergence principle for Fejer monotone methods in Hilbert space*, Math. Oper. Res., 26 (2001), pp. 248–264.
- [4] D. P. BERTSEKAS, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 174–184.

- [5] D. P. BERTSEKAS AND E. M. GAFNI, *Projection methods for variational inequalities with applications to the traffic assignment problem*, Math. Prog. Study, 17 (1982), pp. 139–159.
- [6] H. BREZIS, *Opérateurs maximaux monotones*, North-Holland Math. Stud. 5, North-Holland, Amsterdam, 1973.
- [7] F. E. BROWDER, *Convergence of approximants to fixed points of non-expansive maps in Banach spaces*, Arch. Ration. Mech. Anal., 24 (1967), pp. 82–90.
- [8] F. E. BROWDER AND W. V. PETRYSHYN, *Construction of fixed points on nonlinear mappings in Hilbert space*, J. Math. Anal. Appl., 20 (1967), pp. 197–228.
- [9] R. S. BURACHIK, J. O. LOPES, AND B. S. SVAITER, *An outer approximation method for the variational inequality problem*, SIAM J. Control Optim., 43 (2005), pp. 2071–2088.
- [10] C. L. BYRNE, *A unified treatment of some iterative algorithms in signal processing and image reconstruction*, Inverse Problems, 18 (2004), pp. 441–453.
- [11] Y. CENSOR, A. MOTOVA, AND A. SEGAL, *Perturbed projections and subgradients projections for the multiple-sets split feasibility problem*, J. Math. Anal. Appl., 327 (2007), pp. 1244–1256.
- [12] P. L. COMBETTES AND S. A. HIRSTOAGA, *Equilibrium programming in Hilbert space*, J. Nonlinear Convex Anal., 6 (2005) pp. 117–136.
- [13] J. ECKSTEIN AND D. P. BERTSEKAS, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Program., 55 (1992), pp. 293–318.
- [14] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, Classics Appl. Math. 28, SIAM, Philadelphia, 1999.
- [15] R. GLOWINSKI, *Numerical Methods for Variational Problems*, Springer-Verlag, New York, 1984.
- [16] K. GOEBEL AND W. A. KIRK, *Topics in Metric Fixed Point Theory*, Cambridge Stud. Adv. Math. 28, Cambridge University Press, Cambridge, UK, 1990.
- [17] B. HALPERN, *Fixed points of nonexpanding maps*, Bull. Amer. Math. Soc., 73 (1967), pp. 957–961.
- [18] B. HE, *Inexact implicit methods for monotone general variational inequalities*, Math. Program., 86 (1999), pp. 199–217.
- [19] T. L. HICKS AND J. D. KUBICEK, *On the Mann iteration process in Hilbert spaces*, J. Math. Anal. Appl., 59 (1977), pp. 498–504.
- [20] S. A. HIRSTOAGA, *Iterative selection methods for common fixed point problems*, J. Math. Anal. Appl., 324 (2006), pp. 1020–1035.
- [21] S. ITOH AND W. TAKAHASHI, *The common fixed point theory of singlevalued mappings and multivalued mappings*, Pacific J. Math., 79 (1978), pp. 493–508.
- [22] E. N. KHOBOTOV, *A modification of the extragradient method for the solution of variational inequalities and some optimization problems*, Zh. Vychisl. Mat. Mat. Fiz., 27 (1987), pp. 1462–1473.
- [23] G. M. KORPELEVICH, *The extragradient method for finding saddle points and other problems*, Matecon, 12 (1976), pp. 747–756.
- [24] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.
- [25] P. L. LIONS, *Approximation de points fixes de contractions*, C. R. Acad. Sci. Paris Sér. A, 284 (1977), pp. 1357–1359.
- [26] L. LIU, *Approximation of fixed points of a strictly pseudocontractive mapping*, Proc. Amer. Math. Soc., 125 (1997), pp. 1363–1366.
- [27] P. E. MAINGÉ, *Viscosity methods for zeroes of accretive operators*, J. Approx. Theory, 140 (2006), pp. 127–140.
- [28] P. E. MAINGÉ, *Approximation methods for common fixed points of nonexpansive mappings in Hilbert spaces*, J. Math. Anal. Appl., 325 (2007), pp. 469–479.
- [29] P. E. MAINGÉ, *Inertial iterative process for fixed points of certain quasi-nonexpansive mappings*, Set Valued Anal., 15 (2007), pp. 67–79.
- [30] W. R. MANN, *Mean value methods in iteration*, Proc. Amer. Math. Soc., 4 (1953), pp. 506–510.
- [31] P. MARCOTTE, *Applications of Khobotov's algorithm to variational and network equilibrium problems*, Inform. Systems Oper. Res., 29 (1991), pp. 258–270.
- [32] S. MARUSTER, *The solution by iteration of nonlinear equations in Hilbert spaces*, Proc. Amer. Math. Soc., 63 (1997), pp. 69–73.
- [33] C. MOORE, *Iterative Approximation of Fixed Points of Demicontractive Maps*, The Abdus Salam. Intern. Centre for Theoretical Physics, Trieste, Italy, Scientific Report, IC /98/214, November 1998.
- [34] A. MOUDAFI, *Viscosity approximations methods for fixed point problems*, J. Math. Anal. Appl., 241 (2000), pp. 46–55.
- [35] N. NADEZHKINA AND W. TAKAHASHI, *Weak convergence theorem by an extragradient method for nonexpansive mappings and monotone mappings*, J. Optim. Theory Appl., 128 (2006),

- pp. 191–201.
- [36] M. O. OSILIKE, *Implicit iteration process for common fixed points of a finite family of strictly pseudocontractive maps*, J. Math. Anal. Appl., 294 (2004), pp. 73–81.
  - [37] L. D. POPOV, *On a one-stage method for solving lexicographic variational inequalities*, Russian Math. (Izv. VUZ), 42 (1998), pp. 71–81.
  - [38] S. REICH, *Asymptotic behavior of contractions in Banach spaces*, J. Math. Anal. Appl., 44 (1973), pp. 57–70.
  - [39] S. REICH, *Strong convergence theorems for resolvents of accretive operators in Banach spaces*, J. Math. Anal. Appl., 75 (1980), pp. 287–292.
  - [40] R. T. ROCKAFELLAR, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 55–88.
  - [41] M. SIBONY, *Méthodes itératives pour les équations et inéquations aux dérivées partielles non-linéaires de type monotone*, Calcolo, 7 (1970), pp. 65–183.
  - [42] M. V. SOLODOV AND P. TSENG, *Modified projection methods for monotone variational inequalities*, SIAM J. Control Optim., 34 (1996), pp. 1814–1830.
  - [43] G. STAMPACCHIA, *Formes bilinéaires coercitives sur les ensembles convexes*, C. R. Acad. Sci. Paris, 258 (1964), pp. 4413–4416.
  - [44] T. SUZUKI, *A sufficient and necessary condition for Halpern-type strong convergence to fixed points of nonexpansive mappings*, Proc. Amer. Math. Soc., 135 (2007), pp. 99–106.
  - [45] K. TAJI AND M. FUKUSHIMA, *A new merit function and a successive quadratic programming for variational inequality problems*, SIAM J. Optim., 6 (1996), pp. 704–713.
  - [46] W. TAKAHASHI, *Convex Analysis and Approximation of Fixed Points*, Yokohama Publishers, Yokohama, Japan, 2000.
  - [47] W. TAKAHASHI, *Nonlinear Functional Analysis*, Yokohama Publishers, Yokohama, Japan, 2000.
  - [48] W. TAKAHASHI AND M. TOYODA, *Weak convergence theorems for nonexpansive mappings and monotone mappings*, J. Optim. Theory Appl., 118 (2003), pp. 417–428.
  - [49] R. WITTMAN, *Approximation of fixed points of nonexpansive mappings*, Arch. Math., 58 (1992), pp. 486–491.
  - [50] H. K. XU AND T. H. KIM, *Convergence of hybrid steepest descent methods for variational inequalities*, J. Optim. Theory Appl., 119 (2003), pp. 185–201.
  - [51] I. YAMADA, *The hybrid steepest descent method for the variational inequality over the intersection of fixed point sets of nonexpansive mappings*, in *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, D. Butnariu, Y. Censor, and S. Reich, eds., Elsevier, Amsterdam, 2001, pp. 473–504.
  - [52] I. YAMADA AND N. OGURA, *Hybrid steepest descent method for the variational inequality problem over the fixed point set of certain quasi-nonexpansive mappings*, Numer. Funct. Anal. Optim., 25 (2004), pp. 619–655.
  - [53] E. ZEIDLER, *Nonlinear Functional Analysis and Its Applications*, Springer-Verlag, New York, 1985.
  - [54] L. C. ZENG AND J. C. YAO, *Strong convergence theorem by an extragradient method for fixed point problems and variational inequality problems*, Taiwanese J. Math., 10 (2006), pp. 1293–1303.



## MIXED $H_2/H_\infty$ CONTROL VIA NONSMOOTH OPTIMIZATION\*

P. APKARIAN<sup>†</sup>, D. NOLL<sup>‡</sup>, AND A. RONDEPIERRE<sup>§</sup>

**Abstract.** We present a new approach to mixed  $H_2/H_\infty$  output feedback control synthesis. Our method uses nonsmooth mathematical programming techniques to compute locally optimal  $H_2/H_\infty$ -controllers, which may have a predefined structure. We prove global convergence of our method and present tests to validate it numerically.

**Key words.** mixed  $H_2/H_\infty$  output feedback control, multiobjective control, robustness and performance, nonsmooth optimization, trust region technique

**AMS subject classifications.** 93B36, 93B50, 90C29, 49J52, 90C26, 90C34, 49J35

**DOI.** 10.1137/070685026

**1. Introduction.** Mixed  $H_2/H_\infty$  output feedback control is a prominent example of a multiobjective design problem, where the feedback controller has to respond favorably to several performance specifications. Typically in  $H_2/H_\infty$  synthesis, the  $H_\infty$  channel is used to enhance the robustness of the design, whereas the  $H_2$  channel guarantees good performance of the system. Due to its importance in practice, mixed  $H_2/H_\infty$  control has been addressed in various ways over the years, and we briefly review the main trends.

The interest in  $H_2/H_\infty$  synthesis was originally piqued by three publications [22, 23, 27] in the late 1980s and early 1990s. The numerical methods proposed by these authors are based on coupled Riccati equations in tandem with homotopy methods, but the numerical success of these strategies remains to be established. With the rise of linear matrix inequalities (LMIs) in the later 1990s, different strategies which convexify the problem became increasingly popular. The price to pay for convexifying is either a considerable conservatism, or controllers with a large state dimension [29, 25].

In [45, 47, 48] Scherer developed LMI formations for  $H_2/H_\infty$  synthesis for full-order controllers [48] and reduced the problem to solving LMIs in tandem with nonlinear algebraic equalities [48, 45]. In this form,  $H_2/H_\infty$  problems could in principle be solved via nonlinear semidefinite programming techniques like specSDP [24, 40, 50] or Pennon [31, 32, 36], if only these techniques were suited for medium or large size problems. Alas, one of the disappointing lessons learned in recent years from investigating BMI (bilinear matrix inequality) and LMI problems is that this is just not the case. Due to the presence of Lyapunov variables, whose number grows quadratically with the system size [14, p. 20ff], BMI and LMI programs quickly lead to problem sizes where existing numerical methods fail.

---

\*Received by the editors March 12, 2007; accepted for publication (in revised form) January 28, 2008; published electronically May 14, 2008. This work was supported by research grants from Fondation d'entreprise EADS under the contract "Solving Challenging Problems in Feedback Control," from the Agence Nationale de Recherche (ANR) under contract Guidage, and from ANR under contract Controvert.

<http://www.siam.org/journals/sicon/47-3/68502.html>

<sup>†</sup>CERT-ONERA, 2, avenue Edouard Belin, 31055 Toulouse, France (Pierre.Apkarian@cert.fr).

<sup>‡</sup>Institut de Mathématiques, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France (noll@mip.ups-tlse.fr).

<sup>§</sup>Corresponding author. Institut de Mathématiques, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France (rondep@mip.ups-tlse.fr).

Following [3, 4, 5, 7, 6], we address  $H_2/H_\infty$  synthesis by employing a new strategy which avoids the use of Lyapunov variables. This leads to a nonsmooth and semi-infinite optimization program, which we solve with a spectral bundle method, inspired by the nonconvex spectral bundle method of [37, 38] and [3, 5]. Important forerunners [19, 41, 28] are based on convexity and optimize functions of the form  $\lambda_1 \circ A$  with affine  $A$ . We have developed our method further to deal with typical control applications like multidisk [6] and multifrequency band synthesis [7], design under integral quadratic constraints (IQCs) [4, 9, 8], and loop-shaping techniques [2, 1].

The structure of the paper is as follows. The problem setting is given in section 2. Computing the  $H_2$  and  $H_\infty$  norms is briefly recalled in sections 3 and 4. The algorithm and its rationale are presented in section 5. Global convergence is established in section 6. The implementation is discussed in section 7, and numerical test examples are discussed in section 8. Our terminology follows [17, 30, 52].

**2. Problem setting.** We consider a plant in state space form:

$$(2.1) \quad P : \begin{bmatrix} \dot{x} \\ z_\infty \\ z_2 \\ y \end{bmatrix} = \begin{bmatrix} A & B_\infty & B_2 & B \\ C_\infty & D_\infty & 0 & D_{\infty u} \\ C_2 & 0 & 0 & D_{2u} \\ C & D_{y\infty} & D_{y2} & 0 \end{bmatrix} \begin{bmatrix} x \\ w_\infty \\ w_2 \\ u \end{bmatrix},$$

where  $x \in \mathbb{R}^{n_x}$  is the state,  $u \in \mathbb{R}^{n_u}$  is the control,  $y \in \mathbb{R}^{n_y}$  is the output,  $w_\infty \rightarrow z_\infty$  is the  $H_\infty$  performance channel, and  $w_2 \rightarrow z_2$  is the  $H_2$  performance channel. We seek an output feedback controller:

$$(2.2) \quad K : \begin{bmatrix} \dot{x}_K \\ u \end{bmatrix} = \begin{bmatrix} A_K & B_K \\ C_K & D_K \end{bmatrix} \begin{bmatrix} x_K \\ y \end{bmatrix},$$

where  $x_K \in \mathbb{R}^{n_K}$  is the state of the controller, such that the closed-loop system, obtained by substituting (2.2) into (2.1), satisfies the following properties:

1. *Internal stability.*  $K$  stabilizes  $P$  exponentially in closed-loop.
2. *Fixed  $H_\infty$  performance.* The  $H_\infty$  performance channel has a prespecified performance level  $\|T_{w_\infty \rightarrow z_\infty}(K)\|_\infty \leq \gamma_\infty$ .
3. *Optimal  $H_2$  performance.* The  $H_2$  performance  $\|T_{w_2 \rightarrow z_2}(K)\|_2$  is minimized among all  $K$  satisfying properties 1 and 2.

We will solve the  $H_2/H_\infty$  synthesis problem by way of the following mathematical program:

$$(2.3) \quad \begin{aligned} & \text{minimize} && f(K) := \|T_{w_2 \rightarrow z_2}(K)\|_2^2 \\ & \text{subject to} && g(K) := \|T_{w_\infty \rightarrow z_\infty}(K)\|_\infty^2 \leq \gamma_\infty^2, \end{aligned}$$

where  $T_{w_2 \rightarrow z_2}(K, s)$  denotes the transfer function of the  $H_2$  closed-loop performance channel, while  $T_{w_\infty \rightarrow z_\infty}(K, s)$  stands for the  $H_\infty$  robustness channel. Notice that  $f(K)$  is a smooth function, whereas  $g(K)$  is not, being an infinite maximum of maximum eigenvalue functions. The unknown  $K$  is in the space  $\mathbb{R}^{(n_K+n_u) \times (n_K+n_y)}$ , so the dimension  $n = (n_K + n_y)(n_K + n_u)$  of (2.3) is usually small, which is particularly attractive when small or medium size controllers for large systems are sought. Notice that as a BMI or LMI problem,  $H_2/H_\infty$  synthesis (2.3) would feature  $n_x^2$  additional Lyapunov variables, which would arise through the use of the bounded real lemma. See, e.g., [46, 14].

*Remark.* Naturally, the approach chosen in (2.3) to fix the  $H_\infty$  performance and optimize  $H_2$  performance is just one among many other strategies in multiobjective optimization. One could just as well optimize the  $H_\infty$  norm subject to an  $H_2$  norm constraint, or minimize a weighted sum or even the maximum of both criteria. Other ideas have been considered, and even game theoretic approaches exist [35].

**3. The  $H_2$  norm.** In program (2.3) we minimize composite functions  $f = \|\cdot\|_2^2 \circ T_{w_2 \rightarrow z_2}$ , where  $\|\cdot\|_2$  denotes the  $H_2$  norm. Let us for brevity write  $T_2 := T_{w_2 \rightarrow z_2}$  for the  $H_2$  transfer channel in (2.1). The corresponding plant  $P^2$  is obtained by deleting the  $w_\infty$  column and the  $z_\infty$  line in  $P$ . The objective function can be written as

$$f(K) = \|T_2(K, \cdot)\|_2^2 = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \text{Tr}(T_2(K, j\omega)^H T_2(K, j\omega)) d\omega.$$

Algorithmically it is convenient to compute function values using a state space realization of  $P^2$ :

$$P^2(s) = \begin{bmatrix} 0 & D_{2u} \\ D_{y2} & 0 \end{bmatrix} + \begin{bmatrix} C_2 \\ C \end{bmatrix} (sI - A)^{-1} \begin{bmatrix} B_2 & B \end{bmatrix}.$$

Introducing the closed-loop state space data,

$$\begin{aligned} \mathcal{A}(K) &= \begin{bmatrix} A + BD_K C & BC_K \\ B_K C & A_K \end{bmatrix}, & \mathcal{B}_2(K) &= \begin{bmatrix} B_2 + BD_K D_{y2} \\ B_K D_{y2} \end{bmatrix}, \\ \mathcal{C}_2(K) &= [C_2 + D_{2u} D_K C \quad D_{2u} C_K], & \mathcal{D}_2(K) &= D_{2u} D_K D_{y2} = 0, \end{aligned}$$

we assume either  $D_{2u} = 0$  or  $D_{y2} = 0$ , or that the controller  $K$  is strictly proper, to ensure finiteness of the  $H_2$  norm. Then a realization of the closed-loop transfer function  $T_2$  is given as

$$T_2(K, s) = \mathcal{C}_2(K)(sI - \mathcal{A}(K))^{-1} \mathcal{B}_2(K)$$

and (see, e.g., [21]) the objective function  $f$  may be rewritten as

$$f(K) = \text{Tr}(\mathcal{B}_2(K)^T X(K) \mathcal{B}_2(K)) = \text{Tr}(\mathcal{C}_2(K) Y(K) \mathcal{C}_2(K)^T),$$

where  $X(K)$  and  $Y(K)$  are the solutions of two Lyapunov equations:

$$\begin{aligned} (3.1) \quad & \mathcal{A}(K)^T X(K) + X(K) \mathcal{A}(K) + \mathcal{C}_2(K)^T \mathcal{C}_2(K) = 0, \\ & \mathcal{A}(K) Y(K) + Y(K) \mathcal{A}(K)^T + \mathcal{B}_2(K) \mathcal{B}_2(K)^T = 0. \end{aligned}$$

As observed in [43, section 3], one proves differentiability of the objective  $f$  over the set  $D$  of closed-loop stabilizing controllers  $K$ . In order to write the derivative  $f'(K)dK$  in a gradient form, we introduce the gradient  $\nabla f(K)$  of  $f$  at  $K$  defined by

$$f'(K)dK = \text{Tr}[\nabla f(K)^T dK],$$

meaning that  $\nabla f(K)$  is now an element of the same matrix space as  $K$ . These results lead to the following lemma, which is an extension of [43, Theorem 3.2].

**LEMMA 3.1.** *The objective function  $f$  is differentiable on the open set  $D$  of closed-loop stabilizing gains. For  $K \in D$ , the gradient of  $f$  at  $K$  is*

$$\nabla f(K) = 2 [B^T X(K) + D_{2u}^T \mathcal{C}_2(K)] Y(K) C^T + 2 B^T X(K) \mathcal{B}_2(K) D_{y2}^T,$$

where  $X(K)$  and  $Y(K)$  solve (3.1).

**4. The  $H_\infty$  norm.** The next element required in (2.3) is the constraint function  $g = \|\cdot\|_\infty^2 \circ T_{w_\infty \rightarrow z_\infty}$ , a composite function of the  $H_\infty$  norm. To compute it we will use a frequency domain representation of the  $H_\infty$  norm. Let us for brevity write  $T_\infty := T_{w_\infty \rightarrow z_\infty}$ . The corresponding plant is  $P^\infty$ , obtained by deleting the  $w_2$  column and the  $z_2$  line in  $P$ . The constraint function  $g$  may be written as

$$g(K) = \max_{\omega \in [0, \infty]} \bar{\sigma}(T_\infty(K, j\omega))^2 = \max_{\omega \in [0, \infty]} \lambda_1(T_\infty(K, j\omega)^H T_\infty(K, j\omega)),$$

where  $\bar{\sigma}$  is the maximum singular value of a matrix, and  $\lambda_1$  is the maximum eigenvalue of a Hermitian matrix. We rewrite this as

$$g(K) = \max_{\omega \in [0, \infty]} g(K, \omega), \quad g(K, \omega) = \lambda_1(T_\infty(K, j\omega)^H T_\infty(K, j\omega)).$$

Then it is clear that  $g(K)$  is nonsmooth with two possible sources of nonsmoothness, the infinite maximum and the maximum eigenvalue function, which is convex but nonsmooth. We present two basic results, which allow us to exploit the structure of  $g$  algorithmically. The following can be found in several places; see, e.g., [12, 11].

LEMMA 4.1. *Let  $K$  be closed-loop stabilizing. Then  $g(K) = \|T_\infty(K)\|_\infty^2 < \infty$ , and the set of active frequencies at  $K$ , defined as  $\Omega(K) = \{\omega \in [0, \infty] : g(K) = g(K, \omega)\}$ , is either finite or  $\Omega(K) = [0, \infty]$ .*

The case  $\Omega(K) = [0, \infty]$  is when the closed-loop system is all-pass. It may very well arise in practice; for instance, full order ( $n_x = n_K$ ) optimal  $H_\infty$  controllers are all-pass; see [26]. A similar result holds for full-order  $H_2/H_\infty$  control; see [20]. But we never observed it in cases where the order of the controller  $n_K < n_x$  is way smaller than the order of the system.

The following result was already used in [5, 6]. It allows us to compute Clarke subgradients of the  $H_\infty$  norm and its composite function  $g$ . To represent it, we find it convenient to introduce the notation

$$\begin{bmatrix} T_\infty(K, s) & G_{12}^\infty(K, s) \\ G_{21}^\infty(K, s) & \star \end{bmatrix} = \begin{bmatrix} \mathcal{C}_\infty(K) \\ C \end{bmatrix} (sI - \mathcal{A}(K))^{-1} \begin{bmatrix} \mathcal{B}_\infty(K) & B \end{bmatrix} + \begin{bmatrix} \mathcal{D}_\infty(K) & D_{\infty u} \\ D_{y\infty} & \star \end{bmatrix},$$

where the closed-loop state space data  $(\mathcal{A}(K), \mathcal{B}_\infty(K), \mathcal{C}_\infty(K), \mathcal{D}_\infty(K))$  are given by

$$\begin{aligned} \mathcal{A}(K) &= \begin{bmatrix} A + BD_K C & BC_K \\ B_K C & A_K \end{bmatrix}, & \mathcal{B}_\infty(K) &= \begin{bmatrix} B_\infty + BD_K D_{y\infty} \\ B_K D_{y\infty} \end{bmatrix}, \\ \mathcal{C}_\infty(K) &= [C_\infty + D_{\infty u} D_K C \quad D_{\infty u} C_K], & \mathcal{D}_\infty(K) &= D_\infty + D_{\infty u} D_K D_{y\infty}. \end{aligned}$$

LEMMA 4.2 (see [5, section IV], [13, p. 304]). *Suppose  $K$  is closed-loop stabilizing and  $\Omega(K)$  is finite. Then the Clarke subdifferential of  $g$  at  $K$  is the set*

$$\partial g(K) = \left\{ \Phi_Y : Y = (Y_\omega)_{\omega \in \Omega(K)}, Y_\omega \succeq 0, \sum_{\omega \in \Omega(K)} \text{Tr}(Y_\omega) = 1, Y_\omega \in \mathbb{S}^{r_\omega} \right\},$$

where  $r_\omega$  is the multiplicity of  $\lambda_1(T_\infty(K, j\omega)^H T_\infty(K, j\omega))$ , and where

$$\Phi_Y = \sum_{\omega \in \Omega(K)} 2\text{Re}(G_{21}^\infty(K, j\omega) T_\infty(K, j\omega)^H Q_\omega Y_\omega Q_\omega^H G_{12}^\infty(K, j\omega))^T.$$

Here the columns of the  $m \times r_\omega$  matrix  $Q_\omega$  form an orthonormal basis of the eigenspace of  $T_\infty(K, j\omega)^H T_\infty(K, j\omega) \in \mathbb{S}^m$  associated with its maximum eigenvalue.

*Remark.* Notice that the result extends to the all-pass case by replacing convex combinations over a finite set  $\Omega(K)$  by Radon probability measures on  $[0, \infty]$ . This may still be exploited algorithmically, should the case of an all-pass system ever arise in practice. Since this never occurred in our tests, this line is not investigated here.

**5. Nonsmooth algorithm.** In this central section we present our main result, a nonsmooth and nonconvex optimization method for program (2.3). In subsection 5.1 we will have a look at the necessary optimality conditions for program (2.3). The algorithm is elaborated and presented in subsections 5.2–5.4. The convergence proof will follow in section 6.

As the reader will notice, our method can be applied to a larger class of programs with a structure similar to (2.3). In consequence, during what follows we aim at a certain level of generality. In particular, to comply with the more standard notation in optimization, we denote the decision variable as  $x \in \mathbb{R}^n$ , where  $n = (n_K + n_u)(n_K + n_y)$  in our previous terminology. This means vectorization of the matrix variable previously denoted  $K$ .

**5.1. Optimality conditions.** Following an idea in [42], we address program (2.3) by introducing a progress function:

$$(5.1) \quad F(y; x) = \max \left\{ f(y) - f(x) - \mu[g(x) - \gamma_\infty^2]_+; [g(y) - \gamma_\infty^2] - [g(x) - \gamma_\infty^2]_+ \right\},$$

where  $\mu > 0$  is a fixed parameter. All we need to know about  $f$  is that it is of class  $C^2$ , while  $g$  is assumed to be of the form

$$g(x) = \max_{\omega \in [0, \infty]} g(x, \omega) = \max_{\omega \in [0, \infty]} \lambda_1(G(x, \omega)),$$

with  $G : \mathbb{R}^n \times [0, \infty] \rightarrow \mathbb{S}^m$  of class  $C^2$  in the variable  $x \in \mathbb{R}^n$ , and jointly continuous in  $(x, \omega)$ . This is in accordance with our previous terminology, where  $G(x, \omega) = T_\infty(K, j\omega)^H T_\infty(K, j\omega)$  with  $x = \text{vec}(K)$ , and where  $m = n_{z_\infty}$  or  $m = n_{w_\infty}$ , and where  $n = (n_K + n_y)(n_K + n_u)$ . We have the following preparatory lemma.

**LEMMA 5.1.** (1) *If  $\bar{x} \in \mathbb{R}^n$  is a local minimum of (2.3), then  $\bar{x}$  is also a local minimum of  $F(\cdot; \bar{x})$ . In particular, this implies  $0 \in \partial_1 F(\bar{x}; \bar{x})$ .*

(2) *If  $\bar{x}$  satisfies the F. John necessary optimality conditions for program (2.3), then  $0 \in \partial_1 F(\bar{x}; \bar{x})$ .*

(3) *Conversely, suppose  $0 \in \partial_1 F(\bar{x}; \bar{x})$  for some  $\bar{x} \in \mathbb{R}^n$ . Then we have the following possibilities: Either*

- (i)  *$g(\bar{x}) > \gamma_\infty^2$ , in which case  $\bar{x}$  is a critical point of  $g$ , called a critical point of constraint violation, or*
- (ii)  *$g(\bar{x}) \leq \gamma_\infty^2$ , in which case  $\bar{x}$  satisfies the F. John necessary optimality conditions for program (2.3). In addition, there are two subcases: Either*
  - (a)  *$\bar{x}$  is a Karush–Kuhn–Tucker (KKT) point of (2.3), or*
  - (b)  *$\bar{x}$  fails to be a KKT point of (2.3). This could only happen when  $g(\bar{x}) = \gamma_\infty^2$  and at the same time  $0 \in \partial g(\bar{x})$ .*

*Proof.* (a) Let us prove statement (1). Notice that  $F(\bar{x}; \bar{x}) = 0$ . We therefore have to show  $F(x; \bar{x}) \geq 0$  for  $x$  in a neighborhood of  $\bar{x}$ . If  $x$  is feasible in (2.3), i.e.,  $g(x) \leq \gamma_\infty^2$ , then  $F(x; \bar{x}) = \max\{f(x) - f(\bar{x}); g(x) - \gamma_\infty^2\} \geq f(x) - f(\bar{x}) \geq 0$  for  $x$  in a neighborhood of  $\bar{x}$ . Here we use the fact that  $\bar{x}$ , being optimal, is feasible, so  $[g(\bar{x}) - \gamma_\infty^2]_+ = 0$ . On the other hand, when  $x$  is infeasible, we find  $F(x; \bar{x}) \geq g(x) - \gamma_\infty^2 > 0$ . This settles statement (1).

(b) To prepare the remaining statements, let us first notice that  $0 \in \partial_1 F(\bar{x}; \bar{x})$  is equivalent to the following condition: There exists  $0 \leq \bar{t} \leq 1$  such that  $0 = \bar{t}f'(\bar{x}) + (1 - \bar{t})\phi$  for some  $\phi \in \partial g(\bar{x})$ , where both branches of  $F(\bar{x}; \bar{x})$  have to be active as soon as  $0 < \bar{t} < 1$ . The latter allows one to distinguish the cases  $g(\bar{x}) > \gamma_\infty^2$  and  $g(\bar{x}) \leq \gamma_\infty^2$ .

(c) First consider the case  $g(\bar{x}) > \gamma_\infty^2$ . Here the left-hand branch of  $F(\bar{x}; \bar{x})$ , being strictly negative, cannot be active, which means  $\bar{t} = 0$ . In consequence,  $0 \in \partial g(\bar{x})$ . This is the case of a critical point of constraint violation, so it proves (i) in (3).

(d) Next consider the case  $g(\bar{x}) \leq \gamma_\infty^2$ . In order to show that  $\bar{x}$  satisfies the F. John necessary optimality conditions, it remains to check complementarity. If  $g(\bar{x}) = \gamma_\infty^2$ , there is nothing to prove, so assume  $g(\bar{x}) < \gamma_\infty^2$ . Then the right-hand branch of  $F(\bar{x}; \bar{x})$  is negative, so it cannot be active, meaning that  $(1 - \bar{t}) = 0$ . Since this is the Lagrange multiplier for the constraint, this proves the first part of statement (3)(ii).

(e) It remains to distinguish the two cases (ii)(a) and (ii)(b). Let us see in which cases an F. John critical point can fail to satisfy the KKT conditions. That concerns the case where  $\bar{t} = 0$ , and at the same time  $g(\bar{x}) \leq \gamma_\infty^2$ . But  $g(\bar{x}) < \gamma_\infty^2$  is impossible here, because the right-hand branch of  $F(\bar{x}; \bar{x})$  has to be active. Then it turns out that  $g(\bar{x}) = \gamma_\infty^2$  and  $0 \in \partial g(\bar{x})$  is the only case where KKT fails. It may be considered as the limiting case of a critical point  $\bar{x}$  of constraint violation. This settles all cases in statement (3).

(f) Finally, to prove statement (2), let  $\bar{x}$  satisfy the F. John necessary optimality conditions for (2.3). From (b) we immediately see that it also satisfies  $0 \in \partial_1 F(\bar{x}; \bar{x})$ .  $\square$

*Remark.* (1) Lemma 5.1 shows why we should search for points  $\bar{x}$  satisfying  $0 \in \partial_1 F(\bar{x}; \bar{x})$ . It also indicates that minimizing  $F$  leads to so-called phase I/phase II methods (see [42, section 2.6]). Namely, as long as iterates stay infeasible, the right-hand term in  $F$  is dominant, so reducing  $F$  reduces constraint violation. This corresponds to phase I. Once a feasible iterate has been found, phase I terminates successfully, and iterates will henceforth stay feasible. This is where phase II begins and  $f$  is optimized.

(2) Condition (i) above addresses the case where phase I fails because iterates get stuck at a limit point  $\bar{x}$  with value  $g(\bar{x}) > \gamma_\infty^2$ , which is a local minimum (a critical point) of  $g$  alone. A first-order method may get trapped at such points, and in classical mathematical programming second-order techniques are used to avoid them. Here we are working with a nonsmooth program, where second-order methods are difficult to come up with (see, however, [38], where such a method is discussed, and also [11, 39]). Fortunately, in  $H_2/H_\infty$  control, feasible iterates are usually available, so phase I can even be avoided. Notice also that case (ii)(b) may be considered the limiting case of (i).

(3) In [44] Sagastizábal and Solodov use a different progress function, referred to as an *improvement function*, which does not feature the penalty term  $\mu[g(x) - \gamma_\infty^2]_+$ . Since this term equals 0 in phase II, both criteria lead essentially to the same steps in phase II, and differences could occur only in phase I. Now observe that with the *improvement function*, every step has to be a descent step for both the objective  $f$  and the constraint  $g$ . In contrast, in our approach, when reducing constraint violation, a slight increase in  $f$  not exceeding  $\mu[g(x) - \gamma_\infty^2]_+$  is granted. This helps the algorithm in not being trapped at infeasible local minima of  $f$  alone, and is therefore a possible advantage. Naturally, the difficulty of local minima of  $g$  alone (local minima of constraint violation) remains with both criteria. We will come

back to this issue in section 7.5, where numerical results are discussed. It turns out that a sound choice of  $\mu$  is important and gives better numerical results.

**5.2. First local model.** In this section we introduce a local model for  $F$  in a neighborhood of the current iterate  $x$ . Let us first introduce an approximation of  $g$  in a neighborhood of  $x$  by linearizing the operator  $y \mapsto G(y, \omega)$  around  $x$ :

$$(5.2) \quad \begin{aligned} \tilde{g}(y; x) &= \max_{\omega \in [0, \infty]} \lambda_1 (G(x, \omega) + G'(x, \omega)(y - x)) \\ &= \max_{\omega \in [0, \infty]} \max_{Z \in \mathcal{C}} Z \bullet (G(x, \omega) + G'(x, \omega)(y - x)), \end{aligned}$$

where  $\mathcal{C} = \{Z \in \mathbb{S}^m : Z \succeq 0, \text{tr}(Z) = 1\}$ , and where the derivative  $G'(x, \omega)$  refers to the variable  $x$ . Notice that  $\tilde{g}(x; x) = g(x)$ . By Taylor's theorem we expect  $\tilde{g}(y; x)$  to be a good approximation of  $g(y)$  for  $y$  in a neighborhood of  $x$ .

We now obtain an approximation of  $F$  in a neighborhood of  $x$  by introducing the following:

$$(5.3) \quad \tilde{F}(y; x) = \max \{f'(x)(y - x) - \mu[g(x) - \gamma_\infty^2]_+; [\tilde{g}(y; x) - \gamma_\infty^2] - [g(x) - \gamma_\infty^2]_+\}.$$

Notice that  $\tilde{F}(x; x) = F(x; x)$ , and that  $\tilde{F}(y; x)$  is close to  $F(y; x)$  for  $y$  close to  $x$ . The following result renders these statements exact.

**LEMMA 5.2.** *Let  $B \subset \mathbb{R}^n$  be a bounded set. Then there exists  $L > 0$  such that for all  $x, y \in B$ ,*

$$|g(y) - \tilde{g}(y; x)| \leq L\|y - x\|^2 \quad \text{and} \quad |F(y; x) - \tilde{F}(y; x)| \leq L\|y - x\|^2.$$

*Proof.* By Weyl's theorem we have  $\lambda_m(E) \leq \lambda_1(A + E) - \lambda_1(A) \leq \lambda_1(E)$  for all matrices  $A, E \in \mathbb{S}^m$ . We apply this to  $A = G(y, \omega)$  and  $A + E = G(x, \omega) + G'(x, \omega)(y - x)$ . Then  $E = \mathcal{O}(\|y - x\|^2)$ , uniformly over  $x, y \in B$  and uniformly over  $\omega \in [0, \infty]$ , which is a compact set. Here we use the fact that the operators  $G(\cdot, \omega)$  are of class  $C^2$  in  $x$  and jointly continuous in  $(x, \omega)$ . More precisely

$$\sup_{\omega \in [0, \infty]} \sup_{z \in \text{co}(B)} \|G''(z, \omega)\| < \infty.$$

This proves  $|g(y) - \tilde{g}(y; x)| \leq L_1\|y - x\|^2$  for some  $L_1 > 0$  and all  $x, y \in B$ .

Moreover,  $f$  is of class  $C^2$ , so that by Taylor's formula there exists  $L_2 > 0$  such that  $|f(y) - f(x) - f'(x)(y - x)| \leq L_2\|y - x\|^2$  uniformly over  $x, y \in B$ . With  $L = \max\{L_1, L_2\}$  we obtain

$$\begin{aligned} |F(y; x) - \tilde{F}(y; x)| &\leq \max \{|f(y) - f(x) - f'(x)(y - x)|; |g(y) - \tilde{g}(y; x)|\} \\ &\leq L\|y - x\|^2. \quad \square \end{aligned}$$

It is convenient to represent the local model (5.2) differently. Let us introduce

$$\alpha(\omega, Z) = [Z \bullet G(x, \omega) - \gamma_\infty^2] - [g(x) - \gamma_\infty^2]_+ \in \mathbb{R}, \quad \phi(\omega, Z) = G'(x, \omega)^* Z \in \mathbb{R}^n,$$

where dependence on the point  $x$  is suppressed for convenience. Then the right-hand branch of  $\tilde{F}(y; x)$  may be written as the envelope of cutting planes,

$$[\tilde{g}(y; x) - \gamma_\infty^2] - [g(x) - \gamma_\infty^2]_+ = \sup_{\omega \in [0, \infty]} \sup_{Z \in \mathcal{C}} \alpha(\omega, Z) + \phi(\omega, Z)^T (y - x).$$

Adding the left-hand branch of  $\tilde{F}(y; x)$  by introducing

$$\alpha_0 = -\mu[g(x) - \gamma_\infty^2]_+, \quad \phi_0 = f'(x),$$

we can introduce

$$\mathcal{G} = \text{co}(\{(\alpha(\omega, Z), \phi(\omega, Z)) : \omega \in [0, \infty], Z \in \mathcal{C}\} \cup \{(\alpha_0, \phi_0)\}).$$

Then the local model  $\tilde{F}(y; x)$  may be written as

$$(5.4) \quad \tilde{F}(y; x) = \max\{\alpha + \phi^T(y - x) : (\alpha, \phi) \in \mathcal{G}\}.$$

The advantage of (5.4) over (5.3) is that elements  $(\alpha, \phi)$  of  $\mathcal{G}$  are easier to store than elements  $(\omega, Z) \in [0, \infty] \times \mathcal{C}$ . Also, as we shall see, it is more convenient to construct approximations  $\mathcal{G}_k$  of  $\mathcal{G}$ . This is addressed in the next section.

**5.3. Second local model and tangent program.** Suppose  $x$  is the current iterate of our algorithm to be designed. In order to generate trial steps away from  $x$ , we will recursively generate approximations  $\tilde{F}_k(y; x)$  of  $\tilde{F}(y; x)$ , referred to as the working models. Using (5.4), these will be of the form

$$(5.5) \quad \tilde{F}_k(y; x) = \max\{\alpha + \phi^T(y - x) : (\alpha, \phi) \in \mathcal{G}_k\},$$

where  $\mathcal{G}_k \subset \mathcal{G}$ . In particular,  $\tilde{F}_k(y; x) \leq \tilde{F}(y; x)$ , with exactness  $\tilde{F}_k(x; x) = \tilde{F}(x; x) = F(x; x) = 0$  at  $y = x$ . Moreover, our construction presented below ensures that  $\partial_1 \tilde{F}_k(x; x) \subset \partial_1 F(x; x)$  for all  $k$  and that the  $\tilde{F}_k$  get closer to  $\tilde{F}$  as  $k$  increases. In tandem with the proximity control management described in section 6, this will also ensure that the  $\tilde{F}_k$  get closer to the true  $F$ . Once the set  $\mathcal{G}_k$  is formed, a new trial step  $y^{k+1}$  is computed via the tangent program:

$$(5.6) \quad \min_{y \in \mathbb{R}^n} \tilde{F}_k(y; x) + \frac{\delta_k}{2} \|y - x\|^2.$$

Here  $\delta_k > 0$  is the so-called proximity control parameter, which is specified anew at each step. How this should be organized will be explained in section 6.

Notice that by convexity  $y^{k+1}$  is a solution of (5.6) as soon as

$$(5.7) \quad 0 \in \partial_1 \tilde{F}_k(y^{k+1}; x) + \delta_k(y^{k+1} - x).$$

The first question is, what happens if the solution of the program (5.6) is  $y^{k+1} = x$ ?

**LEMMA 5.3.** *Suppose  $y^{k+1} = x$  is the solution of the tangent program (5.6). Then  $0 \in \partial_1 F(x; x)$ .*

This is indeed clear in view of (5.7), because we get  $0 \in \partial_1 \tilde{F}_k(x; x)$ , which implies  $0 \in \partial_1 F(x; x)$  by the property  $\partial_1 \tilde{F}_k(x; x) \subset \partial_1 F(x; x)$  of a working model. The conclusion is that as soon as  $0 \notin \partial_1 F(x; x)$ , then  $0 \notin \partial_1 \tilde{F}_k(x; x)$ , and the trial step  $y^{k+1}$  will always offer something new. In particular, if  $0 \notin \partial_1 F(x; x)$ , then we know for sure that  $\tilde{F}_k(y^{k+1}; x) < \tilde{F}(x; x) = 0$ , so that there is always a progress predicted by  $\tilde{F}_k$ .

*Remark.* In light of Lemma 5.3 it may seem natural to confine the test  $0 \in \partial_1 \tilde{F}(x; x)$  (step 2 of the algorithm) to the first instance of the tangent program  $k = 1$ . Indeed, if  $0 \in \partial_1 \tilde{F}_1(x; x)$ , then the first tangent program will detect this and return  $y^2 = x$ , in which case we quit. However, notice that this does not work the other



way round. If  $0 \in \partial_1 F(x; x)$ , then the tangent program based on  $\tilde{F}_k$  may still find  $y^{k+1} \neq x$ , in which case we would not necessarily stop the inner loop. Only when  $\partial_1 \tilde{F}_k(x; x) = \partial_1 F(x; x)$  are we certain that  $y^{k+1} = x$ . In other words, if we wish to confine the test in step 2 of the algorithm to the first instance of the tangent program in step 4, we have to use the full subdifferential  $\partial_1 \tilde{F}_1(x; x) = \partial_1 F(x; x)$ . As soon as  $y^2 \neq x$ , then the inner loop is entered, and this condition is no longer required for the following  $\tilde{F}_k$ . In any case, letting  $\partial_1 \tilde{F}_k(x; x) = \partial_1 F(x; x)$  does not pose a numerical problem if  $\partial_1 F(x; x)$  is not exceedingly large.

From now on we assume  $0 \notin \partial_1 F(x, x)$ . The solution  $y^{k+1}$  of (5.6) is then predicting a decrease of the value of the progress function (5.1) at  $y^{k+1}$ . This gives  $y^{k+1}$  the option to improve over the current iterate  $x$  and become the new iterate  $x^+$ . For this to happen, we have to make sure that  $\tilde{F}_k$  is a good model of  $F$  in the neighborhood of  $x$ .

According to standard terminology, when  $y^{k+1}$  is accepted as the new iterate  $x^+$ , it is called a serious step, while trial points  $y^{k+1}$  which are rejected are called null steps. If  $y^{k+1}$  is a null step and has to be rejected, we use it to improve the model  $\mathcal{G}_{k+1}$  at the next sweep.

Let us now show in detail how to construct the sets  $\mathcal{G}_k$ . We choose them of the form

$$(5.8) \quad \mathcal{G}_k = \text{co}(\mathcal{G}_0 \cup \mathcal{G}_k^c \cup \mathcal{G}_k^*), \quad k = 1, 2, \dots,$$

where we refer to  $\mathcal{G}_0$  as the subgradient elements, to  $\mathcal{G}_k^c$  as the cutting planes, and to  $\mathcal{G}_k^*$  as the aggregate element. The first property concerns  $\mathcal{G}_0$ , which is held fixed during the iteration  $k$ .

LEMMA 5.4. *Let  $\omega_0 \in \Omega(x)$  be any of the active frequencies at  $x$ . Choose a normalized eigenvector  $e_0$  associated with the maximum eigenvalue  $g(x) = \lambda_1(G(x, \omega_0))$  of  $G(x, \omega_0)$ , and let  $Z_0 := e_0 e_0^T \in \mathcal{C}$ . If we let  $(\alpha_0, \phi_0) \in \mathcal{G}_0$  and  $(\alpha(\omega_0, Z_0), \phi(\omega_0, Z_0)) \in \mathcal{G}_0$ , and if  $\mathcal{G}_0 \subset \mathcal{G}_k$ , then we have  $\tilde{F}_k(x; x) = F(x; x) = 0$  at all times  $k$ .*

In practice it is useful to enrich the set  $\mathcal{G}_0$  so that it contains the subdifferential  $\partial_1 F(x; x)$  at  $x$ . This can be arranged in those cases where  $\Omega(x)$ , the set of active frequencies, is finite. For every  $\omega \in \Omega(x)$  let  $r_\omega \geq 1$  be the eigenvalue multiplicity of  $\lambda_1(G(x, \omega))$ . Let the  $r_\omega$  columns of  $Q_\omega$  be an orthonormal basis of the maximum eigenspace of  $G(x, \omega)$ . Then put

$$(5.9) \quad \mathcal{G}_0 = \text{co} \left( \{(\alpha(\omega, Z_\omega), \phi(\omega, Z_\omega)) : \omega \in \Omega(x), Z_\omega = Q_\omega^T Y_\omega Q_\omega, Y_\omega \in \mathbb{S}^{r_\omega}, Y_\omega \succeq 0, \text{Tr}(Y_\omega) = 1\} \cup \{(\alpha_0, \phi_0)\} \right).$$

We observe that this set is not finitely generated, but can be handled as a semidefinite programming constraint via the matrices  $Y_\omega$ . However, for our convergence proof it would be sufficient to keep just the one element required by Lemma 5.4 in  $\mathcal{G}_0$ .

Let us now look at the cutting plane sets  $\mathcal{G}_k^c$ . Here we use a recursive construction. Suppose the solution  $y^{k+1}$  of tangent program (5.6) based on the latest model  $\mathcal{G}_k$  is a null step. Then we need to improve the next model  $\mathcal{G}_{k+1}$ , and this is done by including a cutting plane in the new set  $\mathcal{G}_{k+1}^c$ , which cuts away the unsuccessful trial step  $y^{k+1}$ .

LEMMA 5.5. *Let  $y^{k+1}$  be the solution of tangent program (5.6) at stage  $k$  and suppose  $y^{k+1}$  is a null step. Suppose the right-hand branch of (5.3) is active at*

$y^{k+1}$ , and let  $\omega_{k+1} \in [0, \infty]$  and  $Z_{k+1} \in \mathcal{C}$  be one of the pairs where the maximum (5.2) is attained, that is,  $\tilde{g}(y^{k+1}; x) - \gamma_\infty^2 - [g(x) - \gamma_\infty^2]_+ = \alpha(\omega_{k+1}, Z_{k+1}) + \phi(\omega_{k+1}, Z_{k+1})^T(y^{k+1} - x)$ . If we keep  $(\alpha(\omega_{k+1}, Z_{k+1}), \phi(\omega_{k+1}, Z_{k+1})) \in \mathcal{G}_{k+1}^c$ , then

$$\tilde{F}_{k+1}(y^{k+1}; x) = \tilde{F}(y^{k+1}; x).$$

*Remark.* (1) Following standard terminology, we refer to this procedure as the cutting plane element. In fact, adding  $\omega_{k+1}$  and  $Z_{k+1}$  to the approximations at the next step  $k+1$  will cut away the unsuccessful null step  $y^{k+1}$ , paving the way for a better  $y^{k+2}$  at the next sweep.

(2) If the right-hand branch in (5.3) is not active, it suffices to have the pair  $(\alpha_0, \phi_0) \in \mathcal{G}_0$ . As we keep this in  $\mathcal{G}_0$  anyway, no action on cutting planes is required in this event; i.e., we may have  $\mathcal{G}_{k+1}^c = \emptyset$ .

In practice it will be useful to enrich the set  $\mathcal{G}_{k+1}^c$  by what we call anticipating cutting planes. Let us again consider the case of a finite set  $\Omega(x)$ . We select a finite extension  $\Omega_e(x)$  of  $\Omega(x)$  along the lines described in [5]. We let

(5.10)

$$\mathcal{G}_{k+1}^c = \text{co} \left( \{(\alpha(\omega_{k+1}, Z_{k+1}), \phi(\omega_{k+1}, Z_{k+1}))\} \cup \{(\alpha(\omega, Z_\omega), \phi(\omega, Z_\omega)) : \omega \in \Omega_e(x) \setminus \Omega(x), Z_\omega = Q_\omega^T Y_\omega Q_\omega, Y_\omega \succeq 0, \text{Tr}(Y_\omega) = 1\} \right),$$

where the columns of  $Q_\omega$  are an orthonormal basis of some invariant subspace of  $\lambda_1(G(x, \omega))$ . Notice that for  $\omega \in \Omega_e(x) \setminus \Omega(x)$ , the support planes belonging to  $(\alpha(\omega, Z_\omega), \phi(\omega, Z_\omega))$  are indeed different in nature from those retained in  $\mathcal{G}_0$ , because they will not be exact at  $y = x$ . We may have  $\alpha(\omega, Z_\omega) < 0$ , so these planes resemble cutting planes, which are exact at the null steps  $y^{k+1}$ .

Notice that convergence theory requires only  $(\alpha(\omega_{k+1}, Z_{k+1}), \phi(\omega_{k+1}, Z_{k+1})) \in \mathcal{G}_{k+1}^c$  for the element of Lemma 5.5.

*Remark.* Notice that the planes in  $\mathcal{G}_0$  are exact at  $x$ , while genuine cutting planes are exact at the null steps  $y^{k+1}$ . Anticipated cutting planes need not be exact anywhere, but we have observed that they often behave similarly to true cutting planes and can help to avoid a large number of unsuccessful null steps.

We need yet another process to improve the model  $\mathcal{G}_{k+1}$ , which in the nonsmooth terminology is referred to as aggregation, and which is needed in order to avoid storing an increasing number of cutting planes. Suppose that the solution  $y^{k+1}$  of the old tangent program (5.6) based on  $\mathcal{G}_k$  is a null step. By the optimality condition we have  $0 \in \partial_1 \tilde{F}_k(y^{k+1}; x) + \delta_k(y^{k+1} - x)$ . Using the representation (5.4) and the form (5.8), we find  $(\alpha_0, \phi_0) \in \mathcal{G}_0$ ,  $(\alpha_{k+1}, \phi_{k+1}) \in \mathcal{G}_k^c$ , and  $(\alpha_k^*, \phi_k^*) \in \mathcal{G}_k^*$  together with convex coefficients  $\tau_0 \geq 0$ ,  $\tau_{k+1} \geq 0$ ,  $\tau_k^* \geq 0$ ,  $\tau_0 + \tau_{k+1} + \tau_k^* = 1$ , such that

$$0 = \tau_0 \phi_0 + \tau_{k+1} \phi_{k+1} + \tau_k^* \phi_k^* + \delta_k(y^{k+1} - x).$$

We put  $\alpha_{k+1}^* = \tau_0 \alpha_0 + \tau_{k+1} \alpha_{k+1} + \tau_k^* \alpha_k^* \in \mathbb{R}$ ,  $\phi_{k+1}^* = \tau_0 \phi_0 + \tau_{k+1} \phi_{k+1} + \tau_k^* \phi_k^* \in \mathbb{R}^n$  and keep  $(\alpha_{k+1}^*, \phi_{k+1}^*) \in \mathcal{G}_{k+1}^*$ , calling it the aggregate element. Notice that we have  $(\alpha_{k+1}^*, \phi_{k+1}^*) \in \mathcal{G}$  by convexity. Altogether, this shows

$$(5.11) \quad 0 = \phi_{k+1}^* + \delta_k(y^{k+1} - x).$$

LEMMA 5.6. *Keeping the aggregate element  $(\alpha_{k+1}^*, \phi_{k+1}^*)$  in the new  $\mathcal{G}_{k+1}^*$  ensures that  $\tilde{F}_{k+1}(y^{k+1}; x) \geq \tilde{F}_k(y^{k+1}; x)$  and that (5.11) is satisfied.*

To conclude this section, let us outline how the tangent program based on the forms (5.4) and (5.8) is solved. Notice first that elements of  $\mathcal{G}_0 \cup \mathcal{G}_k^c$  have the same form  $(\alpha(\omega, Z_\omega), \phi(\omega, Z_\omega))$ , where  $\omega \in \Omega_e(x)$  for some finite extension of  $\Omega(x)$ , and  $Z_\omega = Q_\omega^T Y_\omega Q_\omega$  for some  $Y_\omega \succeq 0$ ,  $\text{Tr}(Y_\omega) = 1$ . To this we add the aggregate element  $(\alpha_k^*, \phi_k^*)$ , and the element  $(\alpha_0, \phi_0)$  coming from the left-hand branch of  $\tilde{F}$ . This means, after relabeling the finite set  $\Omega_e(x)$  as  $\{\omega_1, \dots, \omega_p\}$ , we can write (5.6) in the form

$$\min_{y \in \mathbb{R}^n} \max \left\{ \alpha_0 + \phi_0^T(y - x); \max_{r=1, \dots, p} \max_{Y_r \succeq 0, \text{Tr}(Y_r)=1} \alpha_r(Y_r) + \phi_r(Y_r)^T(y - x); \right. \\ \left. \alpha_{p+1} + \phi_{p+1}^T(y - x) \right\} + \frac{\delta_k}{2} \|y - x\|^2,$$

where  $\alpha_r(Y_r) = \alpha(\omega_r, Z_{\omega_r})$ , etc., and where the aggregate element  $(\alpha_k^*, \phi_k^*)$  is relabeled  $(\alpha_{p+1}, \phi_{p+1})$ . Replacing the maximum over the three branches by a maximum over the convex hull of the three does not change the value of this program. Using Fenchel duality, we may then swap the min and max operators. Then the inner minimum can be computed explicitly, which leads to the expression

$$y^{k+1} = x - \frac{1}{\delta_k} \left( \tau_0 \phi_0 + \sum_{r=1}^p \tau_r \phi_r(Y_r) + \tau_{p+1} \phi_{p+1} \right),$$

where  $(\tau, Y)$  is the dual variable. Substituting this back into the dual program, using linearity of  $\phi(Y)$  in  $Y$ , and rewriting  $\tau_r Y_r$  as a new matrix  $Y_r$  with  $\text{Tr}(Y_r) = \tau_r$  leads to the dual program

$$\begin{aligned} \text{maximize} \quad & \tau_0 \alpha_0 + \sum_{r=1}^p \alpha_r(Y_r) + \tau_{p+1} \alpha_{p+1} - \frac{1}{2\delta_k} \left\| \tau_0 \phi_0 + \sum_{r=1}^p \phi_r(Y_r) + \tau_{p+1} \phi_{p+1} \right\|^2 \\ \text{subject to} \quad & \tau_0 \geq 0, \tau_{p+1} \geq 0, Y_r \succeq 0, \text{ and } \tau_0 + \sum_{r=0}^{p+1} \text{Tr}(Y_r) + \tau_{p+1} = 1, \end{aligned}$$

which we recognize as the concave form of a semidefinite program (SDP), as soon as we write  $\phi_r(Y_r)$  in its original form  $G'(x, \omega_r)^* Q_{\omega_r}^T Y_r Q_{\omega_r}$ . The return formula becomes

$$(5.12) \quad y^{k+1} = x - \frac{1}{\delta_k} \left( \tau_0^* \phi_0 + \sum_{r=1}^p \phi_r(Y_r^*) + \tau_{p+1}^* \phi_{p+1} \right),$$

where the dual optimal solution is  $(\tau_0^*, Y_1^*, \dots, Y_p^*, \tau_{p+1}^*)$ . Notice that this SDP is usually of small size, so that solving a succession of these programs seems a satisfactory strategy.

To conclude, we consider the case of particular interest in which the eigenvalue multiplicity of all matrices involved is 1, or where we decide to keep only one eigenvector for each leading eigenvalue. If  $\lambda_1(G(x, \omega))$  has eigenvalue multiplicity  $r_\omega = 1$ , the matrices  $Q_\omega$  are just column vectors  $e_\omega$ , where  $e_\omega$  is the normalized eigenvector associated with  $\lambda_1(G(x, \omega))$  and  $Y_\omega = 1$ . Similarly, for the latest cutting plane we then have  $Q_\omega = e_\omega$  for the normalized eigenvector of  $\lambda_1(G(x, \omega) + G'(x, \omega)(y^{k+1} - x))$ . In this case the sets  $\mathcal{G}_0, \mathcal{G}_k^c$  are finite, and so  $\mathcal{G}_k$  itself is a polyhedron  $\text{co}\{(\alpha_0, \phi_0), \dots,$

$(\alpha_{p+1}, \phi_{p+1})\}$ , where  $\text{card}(\Omega_e(x)) = p$ . In this case the dual program is a convex quadratic program which can be solved very efficiently:

$$\begin{aligned} & \text{maximize} \quad \sum_{r=0}^{p+1} \tau_r \alpha_r - \frac{1}{2\delta_k} \left\| \sum_{r=0}^{p+1} \tau_r \phi_r \right\|^2 \\ & \text{subject to} \quad \tau_r \geq 0, r = 0, \dots, p+1, \text{ and } \sum_{r=0}^{p+1} \tau_r = 1, \end{aligned}$$

with dual optimal solution  $\tau^*$ , and the return formula is  $y^{k+1} = x - \frac{1}{\delta_k} \sum_{r=0}^{p+1} \tau_r^* \phi_r$ .

**5.4. The algorithm.** In this section we present the nonsmooth spectral bundle algorithm for program (2.3).

**6. Management of the proximity parameter.** In this section the convergence proof of Algorithm 1 will be given.

To begin with, let us explain the management of the proximity control parameter in steps 5 and 8. Notice that there are two control mechanisms, governed by the control parameters  $\rho_k$  and  $\tilde{\rho}_k$ . In step 5, test parameter  $\rho_k$  compares the current model  $\tilde{F}_k$  to the truth  $F$ . The ideal case would be  $\rho_k \approx 1$ , but we accept  $y^{k+1} = x^{j+1}$  much earlier, namely, if  $\rho_k \geq \gamma$ , where the reader might for instance imagine  $\gamma = \frac{1}{4}$ . Let us call  $y^{k+1}$  bad if  $\rho_k < \gamma$ . So null steps are bad, while serious steps are not bad. Imagine further that  $\Gamma = \frac{3}{4}$ ; then steps  $y^{k+1}$  with  $\rho_k > \Gamma$  are good steps. In the good case the model  $\tilde{F}_k$  seems very reliable, so we can relax proximity control a bit at the next outer step. This is arranged by memorizing  $\delta^+ = \delta_k/2$  in step 5 of the algorithm.

It is more intriguing to decide what we should do when  $\rho_k < \gamma$ , meaning that  $y^{k+1}$  is bad (a null step). Here we need the second control parameter  $\tilde{\rho}_k$  in step 8 to support our decision. Adopting the same terminology, we say that the agreement between  $\tilde{F}$  and  $\tilde{F}_k$  is bad if  $\tilde{\rho}_k < \tilde{\gamma}$ . If this is the case, we keep  $\delta_{k+1} = \delta_k$  unchanged, being reluctant to increase the  $\delta$ -parameter prematurely, and continue to rely on cutting planes and aggregation, hoping that this will drive  $\tilde{F}_k$  closer to  $\tilde{F}$  (and also to  $F$ ) and bring home the bacon in the end. On the other hand, if  $\tilde{\rho}_k \geq \tilde{\gamma}$ , then we have to accept that driving  $\tilde{F}_k$  closer to  $\tilde{F}$  alone will not do the job, simply because  $\tilde{F}$  itself is too far from the true  $F$ . Here we need to tighten proximity control by increasing  $\delta_{k+1} = 2\delta_k$  at the next sweep. This is done in step 8 and brings  $\tilde{F}$  closer to  $F$ .

*Remark.* Notice that the control parameters  $\rho_k$  and  $\tilde{\rho}_k$  in steps 5 and 8 are well defined because we enter the inner loop only when  $0 \notin \partial_1 F(x; x)$ , in which case we have  $\tilde{F}_k(y^{k+1}; x) < \tilde{F}_k(x; x) = 0$ .

**6.1. Finiteness of inner loop.** Let  $x$  be the current iterate of the outer loop. We start our convergence analysis by showing that the inner loop terminates after a finite number of updates  $k$  with a serious step  $y^{k+1} = x^+$ . This will be proved in the next three lemmas.

Recall that  $y^{k+1}$  is the solution of the tangent program (5.6) and may be obtained from the dual optimal solution by the return formula (5.12), which is of the form

$$y^{k+1} = x - \frac{1}{\delta_k} \left[ \tau_0 f'(x) + \sum_{\omega \in \Omega_e(x)} \tau_\omega G'(x, \omega)^* Z_\omega \right]$$

for a finite extension  $\Omega_e(x)$  of  $\Omega(x)$  and for certain  $Z_\omega \in \mathcal{C}$ . Since the sequence  $\delta_k$  in the inner loop is nondecreasing, we have the following lemma.

---

**Algorithm 1.** Proximity control algorithm for the  $H_2/H_\infty$  program (2.3).

---

**Parameters:**  $0 < \gamma < \tilde{\gamma} < \Gamma < 1$ .

- 1: **Initialize outer loop.** Find initial  $x^1$  such that  $f(x^1) < \infty$  and  $g(x^1) < \infty$ . Put outer loop counter  $j = 1$ .
- 2: **Outer loop.** At outer loop counter  $j$ , stop at the current iterate  $x^j$  if  $0 \in \partial_1 F(x^j; x^j)$ . Otherwise compute  $\Omega(x^j)$  and continue with inner loop.
- 3: **Initialize inner loop.** Choose approximation  $\mathcal{G}_1$  of  $\mathcal{G}$  as in (5.8), where  $\mathcal{G}_0$  contains  $(\alpha(\omega_0, Z_0), \phi(\omega_0, Z_0))$  for some fixed  $\omega_0 \in \Omega(x^j)$  and  $Z_0 = e_0 e_0^T$ , where  $e_0$  is a normalized eigenvector associated with  $\lambda_1(G(x^j, \omega_0))$ . Possibly enrich  $\mathcal{G}_0$  as in (5.9). Initialize  $\mathcal{G}_1^c = \emptyset$ ,  $\mathcal{G}_1^* = \emptyset$ , but possibly enrich using anticipated cutting planes (5.10). Initialize proximity parameter  $\delta_1 > 0$ . If memory element  $\delta^+$  for  $\delta$  is available, use it to initialize  $\delta_1$ . Put inner loop counter  $k = 1$ .
- 4: **Trial step.** At inner loop counter  $k$  for given approximation  $\mathcal{G}_k$  and proximity control parameter  $\delta_k > 0$ , solve tangent program:

$$\min_{y \in \mathbb{R}^n} \tilde{F}_k(y; x^j) + \frac{\delta_k}{2} \|y - x^j\|^2,$$

whose solution is  $y^{k+1}$ .

- 5: **Test of progress.** Check whether

$$\rho_k = \frac{F(y^{k+1}; x^j)}{\tilde{F}_k(y^{k+1}; x^j)} \geq \gamma.$$

If this is the case, accept trial step  $y^{k+1}$  as the new iterate  $x^{j+1}$  (serious step). Compute new memory element  $\delta^+$  as:

$$\delta^+ = \begin{cases} \frac{\delta_k}{2} & \text{if } \rho_k > \Gamma, \\ \delta_k & \text{otherwise.} \end{cases}$$

Increase outer loop counter  $j \rightarrow j + 1$ , and go back to step 2. If  $\rho_k < \gamma$ , continue inner loop with step 6 (null step).

- 6: **Cutting plane.** Select a frequency  $\omega_{k+1}$  where  $\tilde{g}(y^{k+1}, x^j)$  is active and pick a normalized eigenvector  $e_{k+1}$  associated with the maximum eigenvalue of  $G(x^j, \omega_{k+1}) + G'(x^j, \omega_{k+1})(y^{k+1} - x^j)$ . Put  $Z_{k+1} = e_{k+1} e_{k+1}^T$  and assure  $(\alpha(\omega_{k+1}, Z_{k+1}), \phi(\omega_{k+1}, Z_{k+1})) \in \mathcal{G}_{k+1}^c$ . Possibly enrich  $\mathcal{G}_{k+1}^c$  by anticipating cutting planes as in (5.10).
- 7: **Aggregation.** Keep aggregate pair  $(\alpha_{k+1}^*, \phi_{k+1}^*)$  as in (5.11) in  $\mathcal{G}_{k+1}^*$ .
- 8: **Proximity control.** Compute control parameter

$$\tilde{\rho}_k = \frac{\tilde{F}(y^{k+1}; x^j)}{\tilde{F}_k(y^{k+1}; x^j)}.$$

Update proximity parameter  $\delta_k$  as

$$\delta_{k+1} = \begin{cases} \delta_k & \text{if } \rho_k < \gamma \text{ and } \tilde{\rho}_k < \tilde{\gamma}, \\ 2\delta_k & \text{if } \rho_k < \gamma \text{ and } \tilde{\rho}_k \geq \tilde{\gamma}. \end{cases}$$

Increase inner loop counter  $k$  and go back to step 4.

---

LEMMA 6.1. *The solutions  $y^{k+1}$  of (5.6) satisfy*

$$(6.1) \quad \|y^{k+1}\| \leq \|x\| + \delta_1^{-1} \left( \|f'(x)\| + \max_{\omega \in [0, \infty]} \|G'(x, \omega)^*\| \right) < \infty.$$

We are now ready to prove finite termination of the inner loop. Our first step is the following.

LEMMA 6.2. *Suppose the inner loop turns forever and creates an infinite sequence  $y^{k+1}$  of null steps with  $\rho_k < \gamma$ . Then there must be an instant  $k_0$  such that the control parameter  $\tilde{\rho}_k$  satisfies  $\tilde{\rho}_k < \tilde{\gamma}$  for all  $k \geq k_0$ .*

*Proof.* Indeed, by assumption none of the trial steps  $y^{k+1}$  passes the acceptance test in step 5, so  $\rho_k < \gamma$  at all times  $k$ . Suppose now that  $\tilde{\rho}_k \geq \tilde{\gamma}$  for infinitely many times  $k$ . Then according to step 8 the proximity control parameter  $\delta_k$  is increased infinitely often, meaning  $\delta_k \rightarrow \infty$ .

Using the fact that  $y^{k+1}$  is the optimal solution of the tangent program (5.6) gives  $0 \in \partial_1 \tilde{F}_k(y^{k+1}; x) + \delta_k(y^{k+1} - x)$ . Using convexity of  $\tilde{F}_k(\cdot; x)$ , we deduce that

$$-\delta_k(y^{k+1} - x)^T(x - y^{k+1}) \leq \tilde{F}_k(x; x) - \tilde{F}_k(y^{k+1}; x).$$

Using  $\tilde{F}_k(x; x) = F(x; x) = 0$ , ensured by keeping  $(\alpha(\omega_0, Z_0), \phi(\omega_0, Z_0)) \in \mathcal{G}_0 \subset \mathcal{G}_k$  at all times (Lemma 5.4), we obtain

$$(6.2) \quad \frac{\delta_k \|y^{k+1} - x\|^2}{-\tilde{F}_k(y^{k+1}; x)} \leq 1.$$

Next, applying Lemma 5.2 to the bounded set  $B = \{y^{k+1} : k \in \mathbb{N}\} \cup \{x\}$  gives

$$(6.3) \quad \left| F(y^{k+1}; x) - \tilde{F}(y^{k+1}; x) \right| \leq L \|y^{k+1} - x\|^2$$

for some  $L > 0$  and every  $k \in \mathbb{N}$ . Now we expand the control parameters  $\rho_k$  and  $\tilde{\rho}_k$  as follows:

$$\begin{aligned} \tilde{\rho}_k &= \rho_k + \frac{F(y^{k+1}; x) - \tilde{F}(y^{k+1}; x)}{-\tilde{F}_k(y^{k+1}; x)} \\ &\leq \rho_k + \frac{L \|y^{k+1} - x\|^2}{-\tilde{F}_k(y^{k+1}; x)} \leq \rho_k + \frac{L}{\delta_k} \quad (\text{using (6.3) and then (6.2)}). \end{aligned}$$

Since  $L/\delta_k \rightarrow 0$ , we deduce  $\limsup \tilde{\rho}_k \leq \limsup \rho_k \leq \gamma < \tilde{\gamma}$ , which contradicts  $\tilde{\rho}_k > \tilde{\gamma}$  for infinitely many  $k$ .  $\square$

So far we know that if the inner loop turns forever, this implies  $\rho_k < \gamma$  and  $\tilde{\rho}_k < \tilde{\gamma}$  from some counter  $k_0$  onwards. Our next lemma shows that this cannot happen. We refer the interested reader to [18, Proposition 4.3], where essentially the same result is proved. For the sake of completeness and the coherence of notation we give our own proof below.

LEMMA 6.3. *Suppose the inner loop turns forever and produces iterates  $y^{k+1}$  with  $\rho_k < \gamma$  and  $\tilde{\rho}_k < \tilde{\gamma}$  for all  $k \geq k_0$ . Then  $0 \in \partial_1 F(x; x)$ .*

*Proof.* (1) Step 8 of the algorithm tells us that from counter  $k_0$  onwards we are in the case where the proximity parameter is no longer increased. We may therefore assume that it remains unchanged for  $k \geq k_0$ , that is,  $\delta := \delta_k$  for all  $k \geq k_0$ .

(2) For later use, let us introduce the function

$$\psi_k(y; x) = \tilde{F}_k(y; x) + \frac{\delta}{2} \|y - x\|^2.$$

As we have seen already, the necessary optimality condition for the tangent program implies

$$\delta \|y^{k+1} - x\|^2 \leq F(x; x) - \tilde{F}_k(y^{k+1}; x) = -\tilde{F}_k(y^{k+1}; x).$$

Now remember that in step 7 of the algorithm we keep the aggregate  $(\alpha_{k+1}^*, \phi_{k+1}^*) \in \mathcal{G}_{k+1}$ . Let us define the function

$$\psi_k^*(y; x) = \alpha_{k+1}^* + \phi_{k+1}^{*T}(y - x) + \frac{\delta}{2} \|y - x\|^2.$$

We claim that

$$(6.4) \quad \psi_k^*(y^{k+1}; x) = \psi_k(y^{k+1}; x) \quad \text{and} \quad \psi_k^*(y; x) \leq \psi_{k+1}(y; x).$$

Indeed, the inequality on the right is clear because  $(\alpha_{k+1}^*, \phi_{k+1}^*)$  is retained in  $\mathcal{G}_{k+1}$  and therefore contributes to the supremum building  $\psi_{k+1}$ . As for the equality on the left, observe that the aggregate subgradient  $\phi_k^*$  is the one which realizes the necessary optimality condition for tangent program (5.6) at stage  $k$ . Now  $\psi_k(\cdot; x)$  is just the objective of this program, so the function  $\psi_k^*(\cdot; x)$  must be exact at  $y^{k+1}$ .

We now prove the relationship

$$(6.5) \quad \psi_k^*(y; x) = \psi_k^*(y^{k+1}; x) + \frac{\delta}{2} \|y - y^{k+1}\|^2.$$

Indeed, notice that  $\psi_k^*$  is a quadratic function, so expanding it gives

$$\begin{aligned} \psi_k^*(y; x) &= \psi_k^*(y^{k+1}; x) + \nabla \psi_k^*(y^{k+1}; x)^T (y - y^{k+1}) \\ &\quad + \frac{1}{2} (y - y^{k+1})^T \nabla^2 \psi_k^*(y^{k+1}; x) (y - y^{k+1}). \end{aligned}$$

But  $\nabla^2 \psi_k^*(y^{k+1}; x) = \delta I$ , so in order to establish (6.5), we have but to show that  $\nabla \psi_k^*(y^{k+1}; x) = 0$ . To prove this observe that

$$\begin{aligned} \nabla \psi_k^*(y^{k+1}; x) &= \phi_{k+1}^* + \delta(y^{k+1} - x) \\ &= -\delta(y^{k+1} - x) + \delta(y^{k+1} - x) = 0 \end{aligned} \quad (\text{using (5.11)}),$$

so (6.5) is proved. Using this and the previous relations gives

$$\begin{aligned} \psi_k(y^{k+1}; x) &\leq \psi_k^*(y^{k+1}; x) + \frac{\delta}{2} \|y^{k+2} - y^{k+1}\|^2 && (\text{using (6.4) left}) \\ &= \psi_k^*(y^{k+2}; x) && (\text{using (6.5)}) \\ &\leq \psi_{k+1}(y^{k+2}; x) && (\text{using (6.4) right}) \\ &\leq \psi_{k+1}(x; x) && (y^{k+2} \text{ is minimizer of } \psi_{k+1}) \\ &= \tilde{F}_k(x; x) = F(x; x) = 0. \end{aligned}$$

This proves that the sequence  $\psi_k(y^{k+1}; x)$  is monotonically increasing and bounded above, so it converges to some limit  $\psi^* \leq F(x; x) = 0$ . Since the term  $\frac{\delta}{2} \|y^{k+2} - y^{k+1}\|^2$  is squeezed in between two terms with the same limit  $\psi^*$ , we deduce that

$$\frac{\delta}{2} \|y^{k+2} - y^{k+1}\|^2 \rightarrow 0.$$

Since the sequence  $y^{k+1}$  is bounded by Lemma 6.1, we deduce using a geometric argument that

$$(6.6) \quad \|y^{k+2} - x\|^2 - \|y^{k+1} - x\|^2 \rightarrow 0.$$

Recalling the relation  $\tilde{F}_k(y; x) = \psi_k(y; x) - \frac{\delta}{2}\|y - x\|^2$ , we finally obtain

$$(6.7) \quad \begin{aligned} & \tilde{F}_{k+1}(y^{k+2}; x) - \tilde{F}_k(y^{k+1}; x) \\ &= \psi_{k+1}(y^{k+2}; x) - \psi_k(y^{k+1}; x) - \frac{\delta}{2}\|y^{k+2} - x\|^2 + \frac{\delta}{2}\|y^{k+1} - x\|^2, \end{aligned}$$

which converges to 0 due to  $\psi_k(y^{k+1}; x) \rightarrow \psi^*$  proved above and property (6.6).

(3) Let  $(\alpha_{k+1}, \phi_{k+1})$  be the cutting plane element obtained from the null step  $y^{k+1}$  which we retain in  $\mathcal{G}_{k+1}$ . By construction this defines an affine support plane of  $\tilde{F}(\cdot; x)$  at  $y^{k+1}$ . But on the other hand the pair  $(\alpha_{k+1}, \phi_{k+1})$  also contributes to the building of the new model  $\tilde{F}_{k+1}(\cdot; x)$ ; thus the new model must be exact at  $y^{k+1}$ , because always  $\tilde{F}_{k+1} \leq \tilde{F}$ , so the value of  $\tilde{F}$  is the best  $\tilde{F}_{k+1}$  could possibly attain. In other words,  $\phi_{k+1}$  is also a subgradient of  $\tilde{F}_{k+1}(\cdot; x)$  at  $y^{k+1}$ . That means

$$\phi_{k+1}^T(y - y^{k+1}) \leq \tilde{F}_{k+1}(y; x) - \tilde{F}_{k+1}(y^{k+1}; x).$$

Using  $\tilde{F}_{k+1}(y^{k+1}; x) = \tilde{F}(y^{k+1}; x)$  we therefore have

$$(6.8) \quad \tilde{F}(y^{k+1}; x) + \phi_{k+1}^T(y - y^{k+1}) \leq \tilde{F}_{k+1}(y; x).$$

Now observe that

$$\begin{aligned} 0 & \leq \tilde{F}(y^{k+1}; x) - \tilde{F}_k(y^{k+1}; x) \\ &= \tilde{F}(y^{k+1}; x) + \phi_{k+1}^T(y^{k+2} - y^{k+1}) - \tilde{F}_k(y^{k+1}; x) - \phi_{k+1}^T(y^{k+2} - y^{k+1}) \\ &\leq \tilde{F}_{k+1}(y^{k+2}; x) - \tilde{F}_k(y^{k+1}; x) + \|\phi_{k+1}\| \|y^{k+2} - y^{k+1}\| \quad (\text{using (6.8)}), \end{aligned}$$

and this term tends to 0 because of (6.7) and the boundedness of  $\phi_{k+1}$ , and because  $y^{k+1} - y^{k+2} \rightarrow 0$ . We conclude that

$$(6.9) \quad \tilde{F}(y^{k+1}; x) - \tilde{F}_k(y^{k+1}; x) \rightarrow 0.$$

(4) We now show that  $\tilde{F}_k(y^{k+1}; x) \rightarrow F(x; x) = 0$ , and therefore by (6.9) also  $\tilde{F}(y^{k+1}; x) \rightarrow F(x; x) = 0$ . Suppose, contrary to the claim, that  $\eta := F(x; x) - \limsup \tilde{F}_k(y^{k+1}; x) > 0$ . Choose  $0 < \theta < (1 - \tilde{\gamma})\eta$ . It follows from (6.9) that there exists  $k_1 \geq k_0$  such that

$$\tilde{F}(y^{k+1}; x) - \theta \leq \tilde{F}_k(y^{k+1}; x)$$

for all  $k \geq k_1$ . Using  $\tilde{\rho}_k < \tilde{\gamma}$  for all  $k \geq k_1$  gives

$$\begin{aligned} \tilde{\gamma}(\tilde{F}_k(y^{k+1}; x) - F(x; x)) & \leq \tilde{F}(y^{k+1}; x) - F(x; x) \\ & \leq \tilde{F}_k(y^{k+1}; x) + \theta - F(x; x). \end{aligned}$$

Passing to the limit implies  $\tilde{\gamma}\eta \geq \eta - \theta$ , contradicting the choice of  $\theta$ . This proves  $\eta = 0$ , as claimed.



(5) Having shown  $\tilde{F}_k(y^{k+1}; x) \rightarrow F(x; x) = 0$ , we now argue that we must have  $y^{k+1} \rightarrow x$ . This follows from the definition of  $y^{k+1}$ , because

$$\psi_k(y^{k+1}; x) = \tilde{F}_k(y^{k+1}; x) + \frac{\delta}{2} \|y^{k+1} - x\|^2 \leq \psi_k(x; x) = F(x; x) = 0.$$

Since  $\tilde{F}_k(y^{k+1}; x) \rightarrow 0$  by part (4), we have indeed  $y^{k+1} \rightarrow x$ . To finish the proof, observe that  $0 \in \partial_1 \psi_k(y^{k+1}; x)$  implies

$$\begin{aligned} \delta(x - y^{k+1})^T(y - y^{k+1}) &\leq \tilde{F}_k(y; x) - \tilde{F}_k(y^{k+1}; x) \\ (6.10) \qquad \qquad \qquad &\leq \tilde{F}(y; x) - \tilde{F}_k(y^{k+1}; x) \end{aligned}$$

for every  $y$ . Passing to the limit gives

$$0 \leq \tilde{F}(y; x) - \tilde{F}(x; x),$$

because the left-hand side in (6.10) converges to 0 in view of  $y^{k+1} \rightarrow x$ , and since  $\tilde{F}_k(y^{k+1}; x) \rightarrow F(x; x)$  by (3) above. Since  $\partial_1 \tilde{F}(x; x) \subset \partial_1 F(x; x)$ , we are done.  $\square$

**6.2. Convergence of outer loop.** Let us consider the sequence  $(x^j)_{j \in \mathbb{N}}$  of serious steps generated by Algorithm 1. We want to show that  $0 \in \partial_1 F(\bar{x}; \bar{x})$  for every accumulation point  $\bar{x}$  of  $(x^j)_{j \in \mathbb{N}}$ . We start by proving that under reasonable hypotheses, the sequence of serious iterates of our algorithm is bounded.

LEMMA 6.4. *Suppose the following two hypotheses are satisfied:*

- (H<sub>1</sub>)  *$g$  is weakly coercive in the sense that if a sequence  $x^j$  satisfies  $\|x^j\| \rightarrow \infty$  and  $g(x^j) > \gamma_\infty^2$ , then  $g(x^j)$  is not strictly monotonically decreasing.*
- (H<sub>2</sub>)  *$f$  is weakly coercive on the level set  $\{x \in \mathbb{R}^n : g(x) \leq \gamma_\infty^2\}$  in the sense that if  $x^j$  is a sequence of feasible iterates with  $\|x^j\| \rightarrow \infty$ , then  $f(x^j)$  is not strictly monotonically decreasing.*

*Then the sequence  $x^j$  of serious iterates with starting point  $x^1$  generated by our algorithm is bounded.*

*Proof.* There are two cases to be discussed.

(a) Suppose the iterates are all infeasible  $g(x^j) > \gamma_\infty^2$ . In that case we use axiom (H<sub>1</sub>). Notice that in phase I we have  $g(x^{j+1}) - g(x^j) \leq F(x^{j+1}, x^j) < 0$ , so the sequence  $g(x^j)$  is strictly decreasing. Then  $x^j$  is bounded by axiom (H<sub>1</sub>).

(b) Suppose next that the iterates are feasible for  $j \geq j_0$ . In phase II we have  $F(x^{j+1}, x^j) = \max\{f(x^{j+1}) - f(x^j), g(x^{j+1}) - \gamma_\infty^2\} \leq 0$ , and hence  $f(x^{j+1}) < f(x^j)$  for  $j \geq j_0$ . Then by axiom (H<sub>2</sub>) the sequence  $x^j$  could not be unbounded.  $\square$

*Remark.* Notice that axiom (H<sub>2</sub>) is certainly satisfied if  $f$  is coercive in the usual sense on the feasible set, that is, if  $f(x^j) \rightarrow \infty$  for feasible iterates with  $\|x^j\| \rightarrow \infty$ . Similarly, (H<sub>1</sub>) could be replaced by the hypothesis that the set  $\{x \in \mathbb{R}^n : \gamma_\infty^2 < g(x) \leq g(x^1)\}$  is bounded.

We are now ready to prove convergence of the outer loop of Algorithm 1.

THEOREM 6.5. *Let axioms (H<sub>1</sub>) and (H<sub>2</sub>) be satisfied. Then every accumulation point  $\bar{x}$  of the sequence of serious steps  $x^j$  generated by the algorithm satisfies  $0 \in \partial_1 F(\bar{x}; \bar{x})$ . In particular,  $\bar{x}$  is either a critical point of constraint violation or an  $F$ . John critical point of the mixed  $H_2/H_\infty$  program (2.3).*

*Proof.* The second part of the statement follows from Lemma 5.1. Let us prove  $0 \in \partial_1 F(\bar{x}; \bar{x})$ .

(1) We first prove convergence  $F(x^{j+1}; x^j) \rightarrow 0$  ( $j \rightarrow \infty$ ). By construction, we know that  $F(x^{j+1}; x^j) \leq 0$  for every  $j \in \mathbb{N}$ . We now distinguish two cases.

Case 1: there exists  $j_0 \in \mathbb{N}$  such that  $g(x^{j_0}) \leq \gamma_\infty^2$ . From that index onwards we have

$$F(x^{j+1}; x^j) = \max \{ f(x^{j+1}) - f(x^j); g(x^{j+1}) - \gamma_\infty^2 \} \leq 0,$$

and hence  $f(x^{j+1}) \leq f(x^j)$  and  $g(x^j) \leq \gamma_\infty^2$ . That means the sequence  $(f(x^i))_{i \in \mathbb{N}}$  is monotone decreasing from  $j_0$  onwards. For any accumulation point  $\bar{x}$  of  $(x^j)_{j \in \mathbb{N}}$ , continuity of  $f$  shows that  $f(\bar{x})$  is an accumulation point of  $(f(x^i))_{i \in \mathbb{N}}$ , and by the monotone sequences theorem, this implies  $f(x^j) \rightarrow f(\bar{x})$ . Now for  $j \geq j_0$  we have

$$F(x^{j+1}; x^j) = \max \{ f(x^{j+1}) - f(x^j); g(x^{j+1}) - \gamma_\infty^2 \},$$

and hence  $\liminf_{j \rightarrow \infty} F(x^{j+1}; x^j) \geq \lim_{j \rightarrow \infty} f(x^{j+1}) - f(x^j) = 0$ . In tandem with  $F(x^{j+1}; x^j) \leq 0$  this clearly implies  $F(x^{j+1}; x^j) \rightarrow 0$  ( $j \rightarrow \infty$ ).

Case 2:  $g(x^j) > \gamma_\infty^2$  for all  $j \in \mathbb{N}$ . Here

$$F(x^{j+1}; x^j) = \max \{ f(x^{j+1}) - f(x^j) - \mu[g(x^j) - \gamma_\infty^2]; g(x^{j+1}) - g(x^j) \} \leq 0.$$

Hence  $(g(x^j))_{j \in \mathbb{N}}$  is monotonically decreasing. As in the first case, we prove that by continuity of  $g$ ,  $g(\bar{x})$  is an accumulation point and so a limit point of  $(g(x^j))_{j \in \mathbb{N}}$ . We deduce in the same way that  $F(x^{j+1}; x^j) \rightarrow 0$ .

(2) Suppose that at the  $j$ th stage of the outer loop the inner loop accepts a serious step at  $k = k_j$ . Then  $x^{j+1} = y^{k_j+1}$ . By the definition of  $y^{k_j+1}$  as minimizer of the tangent program (5.6), this means

$$\delta_{k_j} (x^j - x^{j+1}) \in \partial_1 \tilde{F}_{k_j}(x^{j+1}; x^j).$$

By the subgradient inequality this gives

$$\delta_{k_j} (x^j - x^{j+1})^T (x^j - x^{j+1}) \leq \tilde{F}_{k_j}(x^j; x^j) - \tilde{F}_{k_j}(x^{j+1}; x^j) = -\tilde{F}_{k_j}(x^{j+1}; x^j),$$

where  $\tilde{F}_{k_j}(x^j; x^j) = F(x^j; x^j) = 0$  by Lemma 5.4. Since  $x^{j+1} = y^{k_j+1}$  was accepted in step 4 of the algorithm, we have  $\rho_{k_j} \geq \gamma$ , i.e.,  $-\tilde{F}_{k_j}(x^{j+1}; x^j) \leq -\gamma^{-1}F(x^{j+1}; x^j)$ . Altogether

$$0 \leq \delta_{k_j} \|x^j - x^{j+1}\|^2 \leq -\gamma^{-1}F(x^{j+1}; x^j).$$

Since  $F(x^{j+1}; x^j)$  converges to 0 by part (1), we deduce  $\delta_{k_j} \|x^j - x^{j+1}\|^2 \rightarrow 0$ . We claim that this implies  $\phi_j = \delta_{k_j} (x^j - x^{j+1}) \rightarrow 0$  ( $j \rightarrow \infty$ ).

(3) Suppose on the contrary that there exists an infinite subsequence  $j \in \mathcal{N}$  of  $\mathbb{N}$  such that  $\|\phi_j\| = \delta_{k_j} \|x^j - x^{j+1}\| \geq \eta > 0$  for some  $\eta > 0$  and every  $j \in \mathcal{N}$ . Therefore

$$\delta_{k_j} \|x^j - x^{j+1}\|^2 \geq \eta \|x^j - x^{j+1}\| \geq 0$$

for  $j \in \mathcal{N}$ , which implies  $(x^j - x^{j+1})_{j \in \mathcal{N}} \rightarrow 0$ . That is possible only when  $(\delta_{k_j})_{j \in \mathcal{N}} \rightarrow \infty$ . We now argue that there exists yet another infinite subsequence  $\mathcal{N}'$  of  $\mathbb{N}$  with the property that  $\delta_{k_j} \rightarrow \infty$ , ( $j \in \mathcal{N}'$ ), and such that in addition for each  $j \in \mathcal{N}'$ , the doubling rule to increase  $\delta_k$  in step 7 of the algorithm was applied at least once before  $x^{j+1} = y^{k_j+1}$  was accepted by the inner loop. To construct  $\mathcal{N}'$ , we associate with every  $j \in \mathcal{N}$  the last outer-loop instant  $j' \leq j$  where the  $\delta$ -parameter was increased at least once while the inner loop was turning, and we let  $\mathcal{N}'$  consist of all these  $j'$ ,  $j \in \mathcal{N}$ . It could happen that  $j' = j$ , but in general we know only that

$$2\delta_{k_{j'-1}} \leq \delta_{k_{j'}} \quad \text{and} \quad \delta_{k_{j'}} \geq \delta_{k_{j'+1}} \geq \dots \geq \delta_{k_j}.$$

The latter ensures  $\delta_{k_{j'}} \rightarrow \infty$ ,  $j' \in \mathcal{N}'$ .

Let us say that for  $j \in \mathcal{N}'$ , the doubling rule was applied for the last time at  $k_j - \nu_j$  for some  $\nu_j \geq 1$ . That is, we have  $\delta_{k_j - \nu_j + 1} = 2\delta_{k_j - \nu_j}$ , while the  $\delta$ -parameter was frozen during the remaining steps before acceptance, i.e.,

$$(6.11) \quad \delta_{k_j} = \delta_{k_j - 1} = \cdots = \delta_{k_j - \nu_j + 1} = 2\delta_{k_j - \nu_j}.$$

Recall from step 7 of the algorithm that we have  $\rho_k < \gamma$  and  $\tilde{\rho}_k \geq \tilde{\gamma}$  for those  $k$ , where the step was not accepted and the doubling rule was applied. That is,

$$\rho_{k_j - \nu_j} = \frac{F(x^j; x^j) - F(y^{k_j - \nu_j + 1}; x^j)}{F(x^j; x^j) - \tilde{F}_{k_j - \nu_j}(y^{k_j - \nu_j + 1}; x^j)} = \frac{F(y^{k_j - \nu_j + 1}; x^j)}{\tilde{F}_{k_j - \nu_j}(y^{k_j - \nu_j + 1}; x^j)} < \gamma$$

and

$$\tilde{\rho}_{k_j - \nu_j} = \frac{F(x^j; x^j) - \tilde{F}(y^{k_j - \nu_j + 1}; x^j)}{F(x^j; x^j) - \tilde{F}_{k_j - \nu_j}(y^{k_j - \nu_j + 1}; x^j)} = \frac{\tilde{F}(y^{k_j - \nu_j + 1}; x^j)}{\tilde{F}_{k_j - \nu_j}(y^{k_j - \nu_j + 1}; x^j)} \geq \tilde{\gamma}.$$

By definition of  $y^{k_j - \nu_j + 1}$  and according to (6.11), we now have

$$\frac{1}{2}\delta_{k_j} (x^j - y^{k_j - \nu_j + 1}) \in \partial_1 \tilde{F}_{k_j - \nu_j}(y^{k_j - \nu_j + 1}; x^j).$$

Using  $\tilde{F}_{k_j - \nu_j}(x^j; x^j) = F(x^j; x^j) = 0$  and the subgradient inequality for  $\tilde{F}_{k_j - \nu_j}(\cdot; x^j)$  at  $y^{k_j - \nu_j + 1}$  gives

$$\begin{aligned} \frac{1}{2}\delta_{k_j} (x^j - y^{k_j - \nu_j + 1})^T (x^j - y^{k_j - \nu_j + 1}) &\leq \tilde{F}_{k_j - \nu_j}(x^j; x^j) - \tilde{F}_{k_j - \nu_j}(y^{k_j - \nu_j + 1}; x^j) \\ &\leq -\tilde{F}_{k_j - \nu_j}(y^{k_j - \nu_j + 1}; x^j). \end{aligned}$$

This could also be written as

$$(6.12) \quad \frac{\delta_{k_j} \|x^j - y^{k_j - \nu_j + 1}\|^2}{-\tilde{F}_{k_j - \nu_j}(y^{k_j - \nu_j + 1}; x^j)} \leq 2.$$

Now we know from Lemma 6.4 that the set of serious iterates  $x^j$  is bounded. In tandem with Lemma 6.1, which relates the norm of the null steps  $y^{k+1}$  to the norm of  $x^j$ , we deduce that the set  $B = \{x^j : j \in \mathbb{N}\} \cup \{y^{k+1} : k = 1, \dots, k_j, j \in \mathbb{N}\}$  is bounded. Then Lemma 5.2 provides  $L > 0$  such that

$$(6.13) \quad |F(y^{k_j - \nu_j + 1}; x^j) - \tilde{F}(y^{k_j - \nu_j + 1}; x^j)| \leq L \|y^{k_j - \nu_j + 1} - x^j\|^2$$

for all  $j \in \mathcal{N}'$ . Now expanding the expression  $\tilde{\rho}_{k_j - \nu_j}$  gives

$$\begin{aligned} \tilde{\rho}_{k_j - \nu_j} &= \rho_{k_j - \nu_j} + \frac{F(y^{k_j - \nu_j + 1}; x^j) - \tilde{F}(y^{k_j - \nu_j + 1}; x^j)}{-\tilde{F}_{k_j - \nu_j}(y^{k_j - \nu_j + 1}; x^j)} \\ &\leq \rho_{k_j - \nu_j} + \frac{L \|x^j - y^{k_j - \nu_j + 1}\|^2}{-\tilde{F}_{k_j - \nu_j}(y^{k_j - \nu_j + 1}; x^j)} \quad (\text{using (6.13)}) \\ &\leq \rho_{k_j - \nu_j} + \frac{2L}{\delta_{k_j}} \quad (\text{using (6.12)}). \end{aligned}$$

Since  $\rho_j < \gamma$  and  $L/2\delta_{k_j} \rightarrow 0$  for the infinite subsequence  $j \in \mathcal{N}'$ , we deduce  $\limsup_{j \in \mathcal{N}'} \tilde{\rho}_{k_j - \nu_j} \leq \limsup_{j \in \mathcal{N}'} \rho_{k_j - \nu_j} \leq \gamma < \tilde{\gamma}$ , contradicting  $\tilde{\rho}_j \geq \tilde{\gamma} > \gamma$  for the infinitely many  $j \in \mathcal{N}'$ . This proves that an infinite sequence  $j \in \mathcal{N}$  with  $\|\phi_j\| \geq \eta > 0$  could not exist. The conclusion is that  $(\phi_j)_{j \in \mathbb{N}} = (\delta_{k_j}(x^j - x^{j+1}))_{j \in \mathbb{N}}$  converges to 0.

(4) Let  $\bar{x}$  be an accumulation point of the sequence of serious steps  $x^j$  and pick a convergent subsequence  $x^j \rightarrow \bar{x}$ ,  $j \in \mathcal{N}$ . We have to prove  $0 \in \partial_1 F(\bar{x}; \bar{x})$ .

Since  $\phi_j = \delta_{k_j}(x^j - x^{j+1})$  is a subgradient of  $\tilde{F}_{k_j}(\cdot, x^j)$  at  $y^{k_j+1} = x^{j+1}$  we have

$$\begin{aligned} \phi_j^T h &\leq \tilde{F}_{k_j}(x^{j+1} + h; x^j) - \tilde{F}_{k_j}(x^{j+1}; x^j) \\ &\leq \tilde{F}(x^{j+1} + h; x^j) - \tilde{F}_{k_j}(x^{j+1}; x^j) \quad (\text{using } \tilde{F}_{k_j} \leq \tilde{F}) \end{aligned}$$

for every test vector  $h \in \mathbb{R}^n$ . Now we use the fact that  $y^{k_j+1} = x^{j+1}$  was accepted in step 4 of the algorithm. That means

$$-\tilde{F}_{k_j}(x^{j+1}; x^j) \leq -\gamma^{-1} F(x^{j+1}; x^j).$$

Combining these two estimates gives

$$(6.14) \quad \phi_j^T h \leq \tilde{F}(x^{j+1} + h; x^j) - \gamma^{-1} F(x^{j+1}; x^j)$$

for every test vector  $h$ . Now fix  $h' \in \mathbb{R}^n$  and choose the test vector  $h^j = x^j - x^{j+1} + h'$  for  $j \in \mathcal{N}'$ . Substituting this in (6.14) we obtain

$$(6.15) \quad \delta_{k_j} \|x^j - x^{j+1}\|^2 + \phi_j^T h' \leq \tilde{F}(x^j + h'; x^j) - \gamma^{-1} F(x^{j+1}; x^j).$$

Now observe that  $\delta_{k_j} \|x^j - x^{j+1}\|^2 \rightarrow 0$  by part (2), and  $\phi_j = \delta_{k_j}(x^j - x^{j+1}) \rightarrow 0$  by part (3). This means that the left-hand side of (6.15) converges to 0. As for the terms on the right, recall that  $F(x^{j+1}; x^j) \rightarrow 0$  by part (1) of the proof. Finally, by joint continuity of  $\tilde{F}(\cdot; \cdot)$ , the term  $\tilde{F}(x^j + h'; x^j)$  converges to  $\tilde{F}(\bar{x} + h'; \bar{x})$ . We conclude, passing to the limit  $j \in \mathcal{N}'$  in (6.15) and using  $\tilde{F}(\bar{x}; \bar{x}) = 0$ , that

$$0 \leq \tilde{F}(\bar{x} + h'; \bar{x}) = \tilde{F}(\bar{x} + h'; \bar{x}) - \tilde{F}(\bar{x}; \bar{x}).$$

As this works for every  $h' \in \mathbb{R}^n$ , we have shown  $0 \in \partial_1 \tilde{F}(\bar{x}; \bar{x})$ , and hence also  $0 \in \partial_1 F(\bar{x}; \bar{x})$ .  $\square$

**7. Implementation.** Algorithm 1 has been implemented for both structured and unstructured mixed synthesis, and we use the enriched versions of  $\mathcal{G}_0$  and  $\mathcal{G}_k^c$  to speed up convergence. Notice that in some of the examples in section 8, the controller has to be strictly proper to ensure well-posedness of the  $H_2$  norm. (Namely,  $D_K = 0$  in (2.2) when  $D_{2u}$  and  $D_{y2}$  are nonzero in the plant (2.1).) In those cases the data in (2.2) are no longer freely assigned, the parameterizations being  $K = \mathcal{K}(A_K, B_K, C_K)$  with a linear operator  $\mathcal{K}$ . More general types of parameterizations would equally well fit into our approach and are referred to as structural constraints on the controller.

**7.1. Stopping criteria.** Notice that Algorithm 1 is a first-order method, which may be slow in the neighborhood of a local solution of (2.3). As in [5], we have therefore implemented termination criteria, which avoid pointless computational efforts during the final phase, where iterates make minor progress. Our first stopping test checks criticality  $0 \in \partial_1 F(x; x)$  by computing

$$\inf\{\|h\| : h \in \partial_1 F(x; x)\} < \varepsilon_1.$$

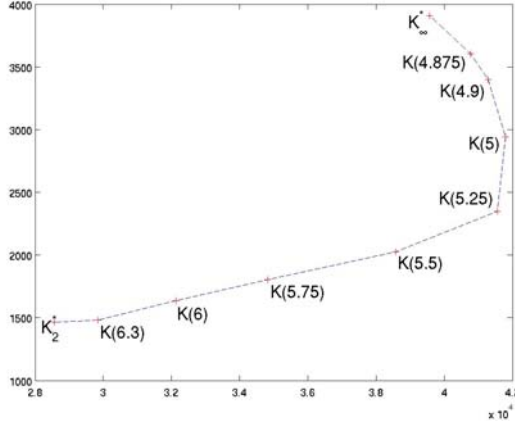


FIG. 1.  $H_2/H_\infty$  optimal static controllers  $K(\gamma_\infty) = (K_1(\gamma_\infty), K_2(\gamma_\infty)) \in \mathbb{R}^2$  for the vehicular suspension control problem.  $[\|T_\infty(K_\infty)\|_\infty, \|T_\infty(K_2)\|_\infty] \ni \gamma_\infty \mapsto K(\gamma_\infty)$  continuously transforms the  $H_\infty$  optimal gain  $K_\infty$  into the  $H_2$  optimal gain  $K_2$ .

Notice that this program is similar (but easier) than the SDP discussed in section 5.3, because the linear terms in that cast are not needed.

A second test compares the progress of the local model around the current iterate:

$$(7.1) \quad |F(x^+; x)| \leq \varepsilon_2.$$

Our third test compares the relative step length to the controller gains:

$$(7.2) \quad \|x^+ - x\| \leq \varepsilon_3(1 + \|x\|).$$

For stopping, we require that either the first or the second and third to be satisfied.

**7.2. Choice of the performance level  $\gamma_\infty$ .** In all test examples we first compute (locally) optimal  $H_2$ - and  $H_\infty$ -controllers  $K_2$  and  $K_\infty$ . It is now trivial (see, e.g., [10]) that the performance level  $\gamma_\infty$  in program (2.3) has to satisfy

$$(7.3) \quad \|T_\infty(K_\infty)\|_\infty \leq \gamma_\infty < \|T_\infty(K_2)\|_\infty.$$

Indeed, the mixed  $H_2/H_\infty$  problem (2.3) is infeasible for  $\gamma_\infty < \|T_\infty(K_\infty)\|_\infty$ , while for  $\gamma_\infty \geq \|T_\infty(K_2)\|_\infty$  the optimal  $H_2$ -controller  $K_2$  is also optimal for (2.3). Disregarding complications due to (multiple) local minima, it would make sense, in a specific case study, to consider the entire one-parameter family  $K(\gamma_\infty)$  of solutions of (2.3) as a function of the gain value  $\gamma_\infty$  over the range (7.3), as this would transform  $K_\infty$  continuously into  $K_2$  (see, e.g., Figure 1). In our tests we only compute  $K(\gamma_\infty)$  for those values  $\gamma_\infty$  which allow comparison to previous results in the literature.

Table 1 reports the problem dimensions  $n_x, n_y, n_u$  and the synthesized controller orders  $n_K$ . Columns 5 and 6 report  $\|T_\infty(K_\infty)\|_\infty$  and  $\|T_\infty(K_2)\|_\infty$ , which are the bounds in (7.3), needed to choose  $\gamma_\infty$  correctly. In column 4 we report  $\|T_2(K_2)\|_2$ , because it gives a lower bound on the optimal value  $\|T_2(K(\gamma_\infty))\|_2$  of (2.3).

Notice that in columns  $\|T_2(K_2)\|_2$  and  $\|T_\infty(K_\infty)\|_\infty$  we would expect decreasing values for a fixed example as  $n_K$  increases. However, in CM4 the orders 0 and 50 give, respectively, 9.2645e-01 and 9.3844e-01, which is not as it should be, because the order 50 controller is worse than the static controller. This phenomenon is due

TABLE 1

Problem dimensions and bounds obtained from locally optimal  $H_2$  and  $H_\infty$  synthesis for the test examples in section 8.

Problem	$(n_x, n_y, n_u)$	$n_K$	$\ T_2(K_2)\ _2$	$\ T_\infty(K_2)\ _\infty$	$\ T_\infty(K_\infty)\ _\infty$
Academic ex. [10]	(2, 1, 1)	0	$6^{\frac{1}{4}}$	$\frac{3}{\sqrt{5}}$	1
Academic ex. [49]	(3, 1, 1)	3	7.748	23.586	9.5196
Vehicular suspension [51]	(4, 2, 1)	0	32.416	6.3287	4.8602
		2	32.299	6.1828	4.8573
		4	32.267	6.3260	4.6797
Four disks [27]	(8, 1, 1)	2	0.5319	3.1658	0.31411
		4	0.4767	2.6194	0.31393
		8	0.3782	1.39	0.27537
From <i>COMPlib</i> : AC14	(40, 4, 3)	1	21.369	230.8318	104.15
		10	8.1039	100.4121	100.11
		20	7.5628	100.3566	100
BDT2	(82, 4, 4)	0	7.9389e-01	1.3167	0.67421
		10	7.8877e-01	1.1386	0.72423
		41	7.7867e-01	1.1302	0.77405
HF1	(130, 1, 2)	0	5.8193e-02	0.4611	0.44721
		10	5.8151e-02	0.4617	0.44721
		25	5.8149e-02	0.4613	0.44721
CM4	(240, 1, 2)	0	9.2645e-1	1.6546	0.81650
		50	9.3844e-1	4.2541	0.81746

to the fact that in all cases  $n_K < n_x$ , we only compute local minima of the  $H_\infty$  program, and similarly, of the  $H_2$  program. As  $n_K$  increases, more local minima appear, and it may be very difficult to improve the situation. This is obviously very unsatisfactory, and appropriate procedures to initialize at a given order  $n_K$  are currently being investigated.

**7.3. Initialization by a stabilizing controller.** In all our test examples, we use the techniques in [11] to compute a closed-loop stabilizing initial  $K^0$ , which is not necessarily feasible for (2.3). This allows us to test phase I of our method.  $K_\infty$  may always be chosen as a feasible initial iterate, so that phase I could in principle be avoided, but we prefer to use various ways to initialize Algorithm 1. In the full-order case  $n_K = n_x$ ,  $K_2$  and  $K_\infty$  are computed by algebraic Riccati equations (AREs) as routinely available in the MATLAB control toolbox. In the reduced-order case  $n_K < n_x$ , things are more complicated, and minima are in general only local. The locally optimal  $H_\infty$ -controller  $K_\infty$  is computed by the method of [5], which uses the initial closed-loop stabilizing  $K^0$  to initialize the procedure. Methods to compute  $K_2$  in the reduced-order case  $n_K < n_x$  are discussed in [43]. Since the objective function  $f(K)$  is not defined everywhere, standard software for unconstrained programming may face difficulties, and we have implemented a Polak–Rivière conjugate gradient method (with a special safeguard to stay in the set  $D$  of exponentially stabilizing controllers) to compute  $K_2$ . An alternative is of course to use Algorithm 1 with  $\gamma_\infty$  so large that  $\gamma_\infty > \|T_\infty(K_2)\|_2$  can be ensured. But this is often slow, because Algorithm 1 is a first-order method. This confirms the observation of the authors of

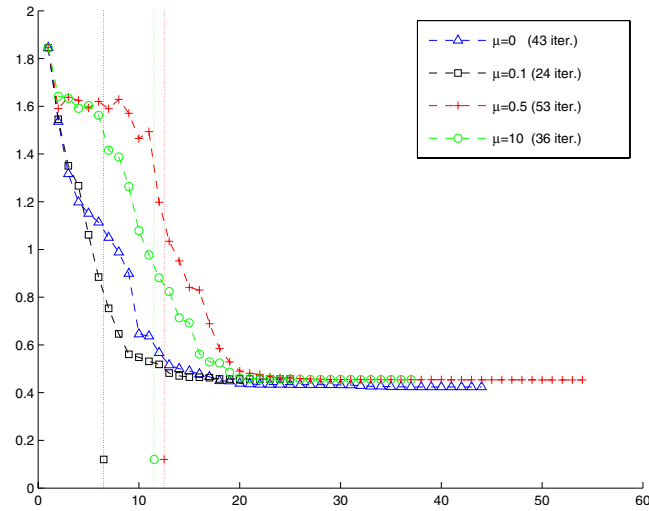


FIG. 2. Full-order mixed  $H_2/H_\infty$  synthesis for the four disks problem: the values of the  $H_2$  norm in relation to the number of serious steps are shown for four different values of the penalty parameter  $\mu$ . Here we choose  $\mu \in \{0, 0.1, 0.5, 10\}$ . Vertical lines point out the instant at which the iterates become feasible.

[43], who report slow convergence for  $H_2$  synthesis based on first-order (gradient-type) methods and recommend using second-order methods instead.

**7.4. Stability constraint.** Notice that closed-loop stability of  $K$  is not a constraint in the usual sense of mathematical programming, because the set  $D$  of closed-loop exponentially stabilizing  $K$  is an open domain. In the program (2.3), closed-loop stability  $K \in D$  is a hidden constraint, which may cause problems because the functions  $f$  and  $g$  are not defined outside  $D$ . The strategy which we adopt here is to compute an initial closed-loop stabilizing controller  $K^0 \in D$  and ignore the hidden constraint during the optimization process. Since  $f(K^0) < \infty$  and  $g(K^0) < \infty$ , our algorithm produces iterates  $K^j$  with  $f(K^j) < \infty$  and  $g(K^j) < \infty$  at all times  $j$ , and most of the time this ensures that  $K^j$  remains closed-loop stabilizing, i.e.,  $K^j \in D$ .

**7.5. Choice of  $\mu$ .** In [42, section 2.6] a similar progress function is discussed for objectives which are maxima of finite or infinite families of smooth functions, but a line search method is obtained. In both cases convergence theory works for arbitrary values of the parameter  $\mu$ , so that no immediate insight into the choice of  $\mu$  is obtained. Yet in practice the choice of  $\mu$  may influence the actual performance of the algorithm.

Figure 2 and Table 2 present the numerical results of our nonsmooth algorithm for the four disks problem presented in section 8. After computing an initial stabilizing controller  $K_0$ , the nonsmooth algorithm is run with four different values of the penalty parameter  $\mu$ , including the case  $\mu = 0$  to compare with the improvement function of [44].

As we can see in Table 2, for  $\mu = 0$  the algorithm fails to reach a feasible point. This is indeed a case where we could identify the final  $K$  of phase I where  $g(K) > \gamma_\infty^2$  as a local minimum of  $f$  alone. Recall that when  $\mu = 0$ , every descent step of the improvement function is a descent step of both  $f$  and  $g$ , and the algorithm then gets trapped as soon as it reaches a local minimum of either  $f$  or  $g$ . Choosing  $\mu > 0$

TABLE 2

*Data and numerical results of  $H_2/H_\infty$  synthesis for the four disks for four different values of  $\mu$ .*

Problem	Four disks [27]			
$(n_x, n_y, n_u)$	(8, 1, 1)			
$\gamma_\infty$	0.6			
$\mu$	0	0.1	0.5	10
Serious steps	43	24	53	36
$\ T_2(K(\gamma_\infty))\ _2^2$	0.1795	0.2087	0.2054	0.2068
$\ T_\infty(K(\gamma_\infty))\ _\infty$	0.7411	0.6000	0.6000	0.6000
Stop test	Tests (7.1) and (7.2)		Criticality	

allows a possible increase of the objective  $f$  during phase I, so that being trapped at an infeasible local minimum of  $f$  alone can be avoided.

Among the choices  $\mu > 0$  we have noticed that when  $\mu$  is not too small, the number of iterations needed to reach a feasible point decreases as  $\mu$  increases. However, choosing too large a  $\mu$ , as shown by the two last columns in Table 2, does not give the best results either, so this trend seems to be true only for a certain range. Nothing decisive can be proposed to date, but  $\mu$  of the same order of magnitude as the progress function without the penalty term so far gave the best results in practice.

**7.6. Choice of  $\Gamma$ .** The last issue we address is the choice of  $\Gamma$ , which is crucial, because step 5 is the only place in the algorithm where the proximity parameter  $\delta_k$  can be reduced. Too large a  $\Gamma$  gives few reductions of  $\delta_k$ , and since the latter is often increased during the inner loop, this bears the risk of exceedingly large  $\delta_k$ , causing the algorithm to stop.

To illustrate this observation, we have run the four disks example in section 8 for three different values  $\Gamma \in \{0.4, 0.6, 0.8\}$ ; see Table 3. The results are illustrated in Figure 3 and Table 3. We observe that the number of iterations increases with the values of  $\Gamma$ . The best numerical results were obtained for  $\Gamma = 0.6$ , and this is the value we retained for all the numerical tests. At least over a certain range one can say that the larger  $\Gamma$ , the smaller the steps accepted as serious steps  $x \rightarrow x^+$ , and the more outer iterations are needed to reach the same  $H_2$  performance.

TABLE 3

*Full-order mixed  $H_2/H_\infty$  synthesis for the four disks problem for three different values of the parameter  $\Gamma \in \{0.4, 0.6, 0.8\}$ .*

Problem	Four disks [27]		
$(n_x, n_y, n_u)$	(8, 1, 1)		
$\gamma_\infty$	0.6		
$\mu$	0.5		
$\Gamma$	0.4	0.6	0.8
Serious steps	36	53	77
$\ T_2(K(\gamma_\infty))\ _2^2$	0.2062	0.2054	0.2106
$\ T_\infty(K(\gamma_\infty))\ _\infty$	0.6000	0.6000	0.6000

**8. Numerical experiments.** In this section we test our nonsmooth algorithm on a variety of  $H_2/H_\infty$  synthesis problems from the literature.

**8.1. Two academic examples.** We first present two academic examples whose models are described in [10] and [49, Example 1]. Notice that the first one is simple



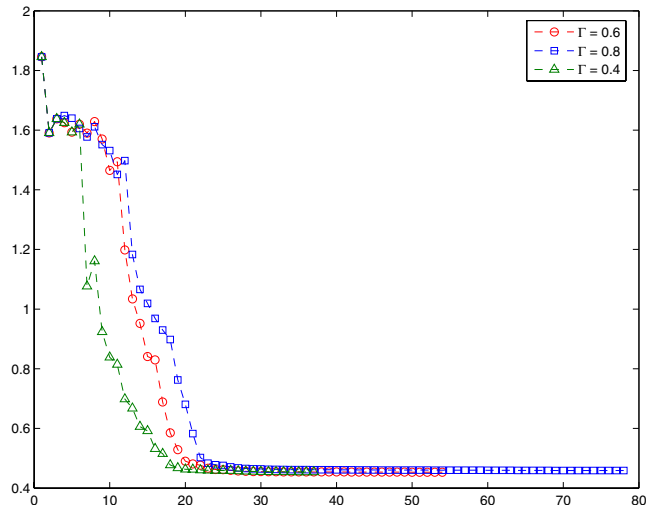


FIG. 3. Full-order mixed  $H_2/H_\infty$  synthesis for the four disks problem: the values of the  $H_2$  norm versus the number of serious steps for three different values of the parameter  $\Gamma \in \{0.4, 0.6, 0.8\}$ .

enough to allow explicit computation of static output feedback controllers  $u = Ky$  for  $H_2$ ,  $H_\infty$ , and  $H_2/H_\infty$  synthesis. The problem data are given in Table 1.

Table 4 confirms that our proximity control algorithm successfully performs the  $H_2/H_\infty$  synthesis on the two considered examples. We not only improve the results computed by the LMI approaches in [10] and [49], we even obtain the theoretical values of the  $H_2$  and  $H_\infty$  norms.

TABLE 4  
Results of  $H_2/H_\infty$  synthesis for two academic examples.

Problem	Academic ex. [10]			Academic ex. [49]	
$(n_x, n_y, n_u)$	(2, 1, 1)			(3, 1, 1)	
$\mu$	10			1	
$n_K$	0	0	1	3	
$\gamma_\infty$	2	1.2	1.2	23.6	12
Serious steps	6	8	14	83	56
$\ T_2(K(\gamma_\infty))\ _2$	1.5651	1.5735	1.5394	7.7484	10.4538
$\ T_\infty(K(\gamma_\infty))\ _\infty$	1.3416	1.2	1.2	23.591	12.0000
$K(\gamma_\infty)$	[ -0.8165 ]	[ -0.9458 ]	$K_{f_1}$	$K_{f_2}$	$K_{f_3}$
Stop test	Criticality				
(LMI) $H_2$ norm	-	1.5778	-	8.07	-
Explicit $H_2$ norm	-	1.5735	-	7.748	-

$$K_{f_1} = \begin{bmatrix} -1.437 & -0.8101 \\ 0.8141 & -0.4998 \end{bmatrix}, K_{f_2} = \begin{bmatrix} -2.5810 & 1.0823 & -0.0623 & -0.5097 \\ -0.5748 & -1.5170 & 2.1121 & 1.6238 \\ -0.1396 & -2.8266 & -2.1852 & 1.7986 \\ 0.2724 & -0.4702 & -2.6967 & 0 \end{bmatrix},$$
$$K_{f_3} = \begin{bmatrix} -1.9113 & -0.7161 & -1.8332 & -0.0065 \\ 0.6940 & -4.4787 & 1.7584 & -2.1896 \\ 0.5231 & -3.2821 & -3.0458 & 3.4518 \\ -3.2830 & 1.1238 & -2.6107 & 0 \end{bmatrix}$$

TABLE 5  
Mixed  $H_2/H_\infty$  synthesis for the vehicular suspension problem.

Problem	Vehicular suspension controller design [51]						
$(n_x, n_y, n_u)$	(4, 2, 1)						
$n_K$	0	0	2	4			
$\mu$	1	1	$10^2$	$10^2$			
$\gamma_\infty$	10	5.225	5.225	5.225			
Serious steps	502	155	496	157			
$\ T_2(K(\gamma_\infty))\ _2$	32.474	34.446	33.312	33.311			
$\ T_\infty(K(\gamma_\infty))\ _\infty$	6.2641	5.2250	5.2250	5.2236			
$K(\gamma_\infty)$	[ 37016 1473 ]	[ 41600 2393 ]	$K_{f_1}$	$K_{f_2}$			
Stop test	Criticality						
$K_{f_1} =$	0.0895	0.3310	0.7272	0.0644	$e + 03,$		
	-0.0670	-0.1540	0.4871	-0.0103			
	0.7986	0.2764	0.1618	1.7366			
$K_{f_2} =$	-0.2156	-0.5614	-0.0012	0.0006	-0.0927	0.0817	$e + 03$
	0.0728	0.1511	-0.0042	0.0020	-0.5677	-0.0378	
	0.0010	-0.0003	0.0011	0.0044	0.0004	0.0005	
	0.0044	-0.0018	-0.0006	-0.0018	0.0007	0.0005	
	0.4743	-0.2764	0.0004	0.0004	0.0922	1.7370	

**8.2. Vehicular suspension controller design.** The model of the vehicular suspension is described in [15] and [51]. We first focus on static  $H_2/H_\infty$  synthesis. The  $H_\infty$  performance level in (2.3) is chosen as  $\gamma_\infty = 5.225$  and the optimal solution we obtain is

$$K(\gamma_\infty) = [ 41600 \ 2393 ].$$

The  $H_2$  norm computed by our algorithm is  $\|T_2(K(\gamma_\infty))\|_2 = 34.446$ , compared to 35.8065 obtained in [51], which gives an improvement of 3.8%. Moreover, the  $H_\infty$  performance is  $\|T_\infty(K(\gamma_\infty))\|_\infty = 5.2250$ , compared to 5.0506 obtained in [51]. This shows that the  $H_\infty$  constraint is not active in the heuristic [51], highlighting the inevitable conservatism of the LMI approaches. In contrast, our method always attains the constraint within the numerical precision.

These results are shown in Table 5, which also gives the results of  $H_2/H_\infty$  synthesis for dynamic controllers of order  $n_K = 2, 4$ . Notice that in the first column of Table 5 by choosing the  $H_\infty$  performance level  $\gamma_\infty > \gamma_2 = \|T_\infty(K_2)\|_\infty$ , where  $\|T_\infty(K_2)\|_\infty$  is given in Table 1, the  $H_2/H_\infty$  solution is close to the solution of the  $H_2$  synthesis.

**8.3. Four disks.** The four disks model is originally described in [16] and has previously been studied to evaluate reduced-order design methods. The open loop plant is of order  $n_x = 8$  and has two stable poles.

We first focus on mixed  $H_2/H_\infty$  synthesis of full-order controllers in order to compare our nonsmooth algorithm to the original Riccati equation approach in [27]. The results are presented in Table 6. We also give results of  $H_2/H_\infty$  synthesis of reduced-order controllers in Table 7.

As can be seen in Table 6, our method gives significant improvement over the older results in [27] based on coupled Riccati equations. This highlights the reduction of conservatism of our approach compared to Riccati and LMI methods.

**8.4. COMPl<sub>e</sub>ib examples.** The models in this section are from the COMPl<sub>e</sub>ib collection [34]: aircraft model AC14, distillation tower BDT2, heat flow in a thin rod

TABLE 6

Full-order mixed  $H_2/H_\infty$  synthesis for the four disks problem ( $n_K = 8$ ): the square  $H_2$  norm is computed in order to compare our results to those in [27].

Problem	Four disks [27]				
$(n_x, n_y, n_u)$	$(8, 1, 1)$				
$\mu$	0.1	0.1	0.1	0.1	0.5
$\gamma_\infty$	1	0.9	0.8	0.7	0.52
Serious steps	35	17	29	49	39
$\ T_2(K(\gamma_\infty))\ _2^2$	0.1558	0.1612	0.1707	0.1829	0.2299
$\ T_\infty(K(\gamma_\infty))\ _\infty$	1.000	0.9000	0.8000	0.7000	0.5200
Stop test	Criticality				
Square $H_2$ norm in [27]	0.168	0.176	0.187	0.203	0.262
$H_\infty$ norm in [27]	0.855	0.797	0.732	0.661	0.511
Improvement	7.26%	8.41 %	8.72%	9.90%	12.25%

TABLE 7

Reduced-order mixed  $H_2/H_\infty$  synthesis for the four disks problem.

Problem	Four disks [27]			
$(n_x, n_y, n_u)$	$(8, 1, 1)$			
$\mu$	1			
$\gamma_\infty$	0.52			
$n_K$	2	4	6	7
Serious steps	23	18	30	47
$\ T_2(K(\gamma_\infty))\ _2$	0.2321	0.2308	0.23041	0.2304
$\ T_\infty(K(\gamma_\infty))\ _\infty$	0.52	0.52	0.52	0.52
Stop test	Criticality			

HF1, and cable mass model CM4. They are originally designed for  $H_\infty$  synthesis, so an  $H_2$  channel was added as suggested by Leibfritz [33, 34]. The same channel is used for both  $H_2$  and  $H_\infty$  performance in example AC14, while we choose  $B_2 = B_\infty$  and  $D_{y2} = 0$  for the three other models. This way the  $H_2$  norm is well-posed.

In each example, we first choose the  $H_\infty$  performance level  $\gamma_\infty$  larger than  $\|T_\infty(K_2)\|_\infty$ . In doing this we have to obtain an estimate of the optimal  $H_2$  performance  $\|T_2(K_2)\|_2$  given in Table 1. Numerical results are in Tables 8 and 9.

As an illustration, Figures 4 and 5 show the evolution of the  $H_2$  and  $H_\infty$  norms, for example, BDT2, during the first iterations.

In Figure 4 we observe phases I and II of the algorithm. As long as iterates remain infeasible, descent steps to reduce constraint violation are generated, sometimes causing the objective to increase. As soon as the feasible domain  $g(x) \leq \gamma_\infty^2$  is reached, descent of the objective  $f$  begins, and iterates stay feasible.

Figure 5 shows the frequency plot  $\omega \mapsto \lambda_1(T_\infty(K_i, j\omega)^H T_\infty(K_i, j\omega))$  of the  $H_\infty$  constraint during the first 6 iterations (serious steps)  $K_i$ ,  $i = 1, \dots, 6$ , along with the second eigenvalue  $\lambda_2$ . As can be seen, the maximum  $\|T_\infty(K_i)\|_\infty^2$  is sometimes attained at a single marked peak  $\omega$ , while other cases feature rather a flat plateau in the low frequency band. Multiple peaks appear usually at the end of the process, but cannot be ruled out at any moment, as shown by the lower right plot, which has a plateau where  $\lambda_1$  and  $\lambda_2$  are close. Stars indicate frequencies kept in the extended set  $\Omega_e(K_i)$ .

TABLE 8

Results of mixed  $H_2/H_\infty$  synthesis for test examples from COMPl<sub>e</sub>ib. Criticality is pointed out by a \* on the number of serious steps.

Problem	$(n_x, n_y, n_u)$	$n_K$	$\gamma_\infty$	Serious steps	$\ T_2(K(\gamma_\infty))\ _2$	$\ T_\infty(K(\gamma_\infty))\ _\infty$
AC14	(40, 4, 3)	1	1000	300(max.)	21.370	231.31
		10	1000	300(max.)	8.7813	101.26
		1	200	263*	21.476	200
		20	200	300(max.)	7.9879	100
BDT2	(82, 4, 4)	0	10	148*	8.0402e-01	1.0585
		10	10	543*	7.6480e-01	1.1438
		0	0.8	324*	7.9092e-01	7.9999e-01
		10	0.8	404*	7.7146e-01	0.8000
		41	0.8	115*	7.8882e-01	0.8000
HF1	(130, 1, 2)	0	10	7	5.8193e-02	4.6087e-01
		0	0.45	7*	5.8795e-02	4.4999e-01
		10	0.45	7*	5.8706e-02	4.5000e-01
		25	0.45	33*	5.8700e-02	4.4993e-01
CM4	(240, 1, 2)	0	10	5*	9.2645e-01	1.6555
		0	1	20*	9.8438e-01	1
		25	1	15*	9.5330e-01	1.000
		50	1	41*	9.4038e-01	1.000

TABLE 9

Static  $H_2/H_\infty$  output feedback controllers for examples from COMPl<sub>e</sub>ib.

Problem	$\gamma_\infty$	$K(\gamma_\infty)$
BDT2	10	$\begin{bmatrix} -0.6186 & -0.1426 & -0.5414 & 4.929 \\ 0.6357 & -0.5457 & -3.851 & 16.85 \\ -0.07527 & 0.2962 & -1.287 & 6.601 \\ 0.9223 & 0.4668 & -4.091 & 22.34 \end{bmatrix}$
	0.8	$\begin{bmatrix} -0.9207 & 0.9647 & -5.4243 & 9.8225 \\ 0.7452 & -1.3280 & -4.4241 & -0.8141 \\ -0.7119 & 2.1754 & -10.226 & 14.3827 \\ 0.0887 & 1.7433 & -13.4102 & 12.1358 \end{bmatrix}$
HF1	10	$\begin{bmatrix} -0.1002 & -1.1230 \end{bmatrix}$
	0.45	$\begin{bmatrix} -0.2521 & -1.116 \end{bmatrix}$
CM4	10	$\begin{bmatrix} -0.5448 & -1.3322 \end{bmatrix}$
	1	$\begin{bmatrix} -0.5146 & -0.8073 \end{bmatrix}$

**9. Conclusion.** We have studied and tested a nonlinear mathematical programming approach to the mixed  $H_2/H_\infty$ -controller synthesis problem. The importance of this problem was recognized in the late 1980s, but approaches based on AREs could not be brought to work satisfactorily. It is possible to characterize the optimal  $H_2/H_\infty$ -controller by way of the Q-parameterization, but as soon as the controller has to satisfy additional structural constraints, such as, for instance, reduced-order  $n_K < n_x$ , an analytic solution does not exist. In that situation convexity methods based on LMIs and AREs are no longer suitable, and finding the globally optimal solution is known to be NP-hard. As a consequence, we propose a strategy based on local optimization, which comes with a weaker certificate, but has the benefit of working in practice. The problem being nonconvex, nonsmooth, and semi-infinite, we have

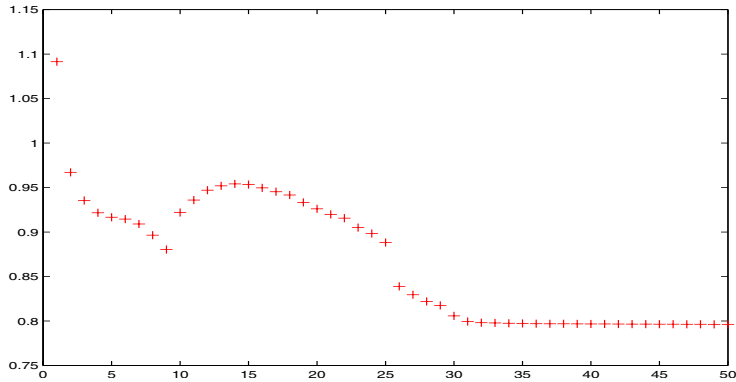


FIG. 4. Example BDT2:  $H_2$  norm during the first 50 iterations.

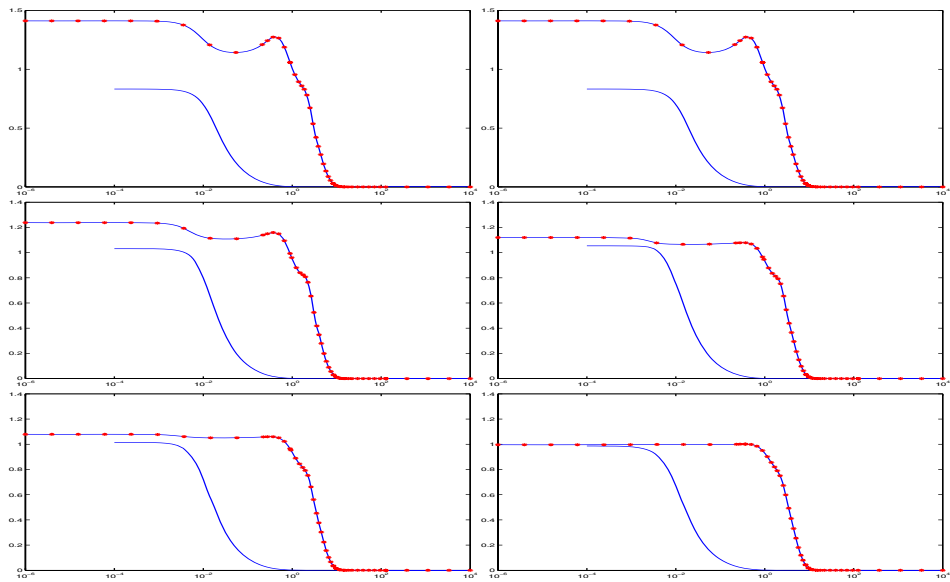


FIG. 5. Example BDT2: Largest and second largest eigenvalues versus frequency in logarithmic scale, the first 6 iterations. Observe that the second eigenvalue  $\lambda_2$  is strictly below the first one on the range  $\omega \leq 10^4$ , except for the bottom right plot, where coalescence on a low frequency band seems to occur.

developed a nonsmooth constrained programming technique suited for the  $H_2/H_\infty$  problem and other programs of a similar structure. The new method has been tested on several benchmark studies and shown to perform better than existing methods.

REFERENCES

[1] P. APKARIAN, V. BOMPART, AND D. NOLL, *Control design in the time and frequency domain using nonsmooth techniques*, Systems Control Lett., 57 (2008), pp. 271–282.  
[2] P. APKARIAN, V. BOMPART, AND D. NOLL, *Nonsmooth structured control design with application to PID loop-shaping of a process*, Internat. J. Robust Nonlinear Control, 17 (2007), pp. 1320–1342.  
[3] P. APKARIAN AND D. NOLL, *Controller design via nonsmooth multidirectional search*, SIAM J. Control Optim., 44 (2006), pp. 1923–1949.

- [4] P. APKARIAN AND D. NOLL, *IQC analysis and synthesis via nonsmooth optimization*, Systems Control Lett., 55 (2006), pp. 971–981.
- [5] P. APKARIAN AND D. NOLL, *Nonsmooth  $H_\infty$  synthesis*, IEEE Trans. Automat. Control, 51 (2006), pp. 71–86.
- [6] P. APKARIAN AND D. NOLL, *Nonsmooth optimization for multidisk  $H_\infty$  synthesis*, European J. Control, 12 (2006), pp. 229–244.
- [7] P. APKARIAN AND D. NOLL, *Nonsmooth optimization for multiband frequency domain control design*, Automatica, 43 (2007), pp. 724–731.
- [8] P. APKARIAN, D. NOLL, AND O. PROT, *A trust region spectral bundle method for nonconvex eigenvalue optimization*, SIAM J. Optim., 19 (2008), pp. 281–306.
- [9] P. APKARIAN, D. NOLL, AND O. PROT, *Nonsmooth methods for analysis and synthesis with integral quadratic constraints*, in Proceedings of the IEEE Conference on Decision and Control, New Orleans, 2007, pp. 824–829.
- [10] D. ARZELIER AND D. PEAUCELLE, *An iterative method for mixed  $H_2/H_\infty$  synthesis via static output-feedback*, in Proceedings of the IEEE Conference on Decision and Control, 2002, pp. 3464–3469.
- [11] V. BOMPART, D. NOLL, AND P. APKARIAN, *Second-order nonsmooth optimization for feedback control*, Numer. Math., 107 (2007), pp. 433–454.
- [12] S. BOYD AND V. BALAKRISHNAN, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its  $L_\infty$ -norm*, Systems Control Lett., 15 (1990), pp. 1–7.
- [13] S. BOYD AND C. BARRATT, *Linear Controller Design: Limits of Performance*, Prentice-Hall, Upper Saddle River, NJ, 1991.
- [14] S. BOYD, L. EL GHAOU, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.
- [15] J. CAMINO, D. ZAMPIERI, AND P. PERES, *Design of a vehicular suspension controller by static output feedback*, in Proceedings of the IEEE American Control Conference, 1999, pp. 3168–3172.
- [16] R. H. CANNON AND D. E. ROSENTHAL, *Experiments in control of flexible structures with non-collocated sensors and actuators*, J. Guidance Control Dynam., 7 (1984), pp. 546–553.
- [17] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Canad. Math. Soc. Ser. Monogr. Adv. Texts, John Wiley & Sons, New York, 1983.
- [18] R. CORREA AND C. LEMARÉCHAL, *Convergence of some algorithms for convex minimization*, Math. Programming, 62 (1993), pp. 261–275.
- [19] J. CULLUM, W. E. DONATH, AND P. WOLFE, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Programming Stud., 3 (1975), pp. 35–55.
- [20] S. M. DJOUADI, C. D. CHARALAMBOUS, AND D. W. REPPERGER, *On multiobjective  $H_2/H_\infty$  optimal control*, in Proceedings of the IEEE American Control Conference, Arlington, VA, 2001, pp. 4091–4096.
- [21] J. C. DOYLE, K. GLOVER, P. KHARGONEKAR, AND B. A. FRANCIS, *State-space solutions to standard  $H_2$  and  $H_\infty$  control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.
- [22] J. C. DOYLE, K. ZHOU, AND B. BODENHEIMER, *Optimal control with mixed  $H_2$  and  $H_\infty$  performance objectives*, in Proceedings of the IEEE American Control Conference, Vol. 3, 1989, pp. 2065–2070.
- [23] J. C. DOYLE, K. ZHOU, K. GLOVER, AND B. BODENHEIMER, *Mixed  $H_2$  and  $H_\infty$  performance objectives. II. Optimal control*, IEEE Trans. Automat. Control, 39 (1994), pp. 1575–1587.
- [24] B. FARES, D. NOLL, AND P. APKARIAN, *Robust control via sequential semidefinite programming*, SIAM J. Control Optim., 40 (2002), pp. 1791–1820.
- [25] J. C. GEROMEL, P. L. D. PERES, AND S. R. SOUZA, *A convex approach to the mixed  $\mathcal{H}_2/\mathcal{H}_\infty$  control problem for discrete-time uncertain systems*, SIAM J. Control Optim., 33 (1995), pp. 1816–1833.
- [26] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their  $L_\infty$ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.
- [27] W. M. HADDAD AND D. S. BERNSTEIN, *LQG control with a  $H_\infty$  performance bound: A Riccati equation approach*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 293–305.
- [28] C. HELMBERG AND F. RENDL, *A spectral bundle method for semidefinite programming*, SIAM J. Optim., 10 (2000), pp. 673–696.
- [29] H. A. HINDI, B. HASSIBI, AND S. BOYD, *Multiobjective  $H_2/H_\infty$ -optimal control via finite dimensional  $Q$ -parametrization and linear matrix inequalities*, in Proceedings of the IEEE American Control Conference, 1998, pp. 3244–3248.

- [30] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*, Grundlehren Math. Wiss. 306, Springer-Verlag, New York, 1993.
- [31] M. KOČVARA AND M. STINGL, *PENNON: A code for convex nonlinear and semidefinite programming*, Optim. Methods Softw., 18 (2003), pp. 317–333.
- [32] M. KOČVARA AND M. STINGL, *Solving nonconvex SDP problems of structural optimization with stability control*, Optim. Methods Softw., 19 (2004), pp. 595–609.
- [33] F. LEIBFRTZ, *An LMI-based algorithm for designing suboptimal static  $H_2/H_\infty$  output feedback controllers*, SIAM J. Control Optim., 39 (2001), pp. 1711–1735.
- [34] F. LEIBFRTZ, *COMPL<sub>e</sub>IB, Constraint Matrix-optimization Problem Library: A Collection of Test Examples for Nonlinear Semidefinite Programs, Control System Design and Related Problems*, Technical report, Universität Trier, 2003.
- [35] D. J. N. LIMBEER, B. D. O. ANDERSEN, AND B. HENDEL, *A Nash game approach to mixed  $H_2/H_\infty$  control*, IEEE Trans. Automat. Control, 39 (1994), pp. 69–82.
- [36] D. NOLL, *Local convergence of an augmented Lagrangian method for matrix inequality constrained programming*, Optim. Methods Softw., 22 (2007), pp. 777–802.
- [37] D. NOLL AND P. APKARIAN, *Spectral bundle methods for nonconvex maximum eigenvalue functions: First-order methods*, Math. Programming Ser. B, 104 (2005), pp. 701–727.
- [38] D. NOLL AND P. APKARIAN, *Spectral bundle methods for nonconvex maximum eigenvalue functions: Second-order methods*, Math. Programming Ser. B, 104 (2005), pp. 729–747.
- [39] D. NOLL, O. PROT, AND A. RONDEPIERRE, *A proximity control algorithm to minimize non-smooth and nonconvex functions*, Pacific J. Optim., to appear.
- [40] D. NOLL, M. TORKI, AND P. APKARIAN, *Partially augmented Lagrangian method for matrix inequality constraints*, SIAM J. Optim., 15 (2004), pp. 161–184.
- [41] F. OUSTRY, *A second-order bundle method to minimize the maximum eigenvalue function*, Math. Programming Ser. A, 89 (2000), pp. 1–33.
- [42] E. POLAK, *Optimization: Algorithms and Consistent Approximations*, Appl. Math. Sci. 124, Springer-Verlag, New York, 1997.
- [43] T. RAUTERT AND E. W. SACHS, *Computational design of optimal output feedback controllers*, SIAM J. Optim., 7 (1997), pp. 837–852.
- [44] C. SAGASTIZÁBAL AND M. SOLODOV, *An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter*, SIAM J. Optim., 16 (2005), pp. 146–169.
- [45] C. W. SCHERER, *Multiobjective  $H_2/H_\infty$  control*, IEEE Trans. Automat. Control, 40 (1995), pp. 1054–1062.
- [46] C. W. SCHERER, *Mixed  $H_2/H_\infty$  control*, in Trends in Control: A European Perspective, A. Isidori, ed., Springer-Verlag, Berlin, 1995, pp. 173–216.
- [47] C. W. SCHERER, *Lower bounds in multi-objective  $H_2/H_\infty$  problems*, in Proceedings of the 38th IEEE Conference on Decision and Control, 1999, pp. 3605–3610.
- [48] C. W. SCHERER, *An efficient solution to multi-objective control problems with LMI objectives*, Systems Control Lett., 40 (2000), pp. 43–57.
- [49] C. SCHERER, P. GAHINET, AND M. CHILALI, *Multi-objective output-feedback control via LMI optimization*, IEEE Trans. Automat. Control, 42 (1997), pp. 896–911.
- [50] J.-B. THEVENET, D. NOLL, AND P. APKARIAN, *Non linear spectral SDP method for BMI constrained problems: Applications to control design*, in Proceedings of the First International Conference on Informatics in Control, Automation and Robotics, INSTICC Press, Setúbal, Portugal, 2004, pp. 237–248.
- [51] J. YU, *A new static output feedback approach to the suboptimal mixed  $H_2/H_\infty$  problem*, Internat. J. Robust Nonlinear Control, 14 (2004), pp. 1023–1034.
- [52] K. ZHOU, J. C. DOYLE, AND K. GLOVER, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1996.

# GENERALIZED $\mathcal{S}$ -PROCEDURE FOR INEQUALITY CONDITIONS ON ONE-VECTOR-LOSSLESS SETS AND LINEAR SYSTEM ANALYSIS\*

YOSHIO EBIHARA<sup>†</sup>, KATSUTOSHI MAEDA<sup>†</sup>, AND TOMOMICHI HAGIWARA<sup>†</sup>

**Abstract.** The generalized version of the  $\mathcal{S}$ -procedure, recently introduced by Iwasaki and co-authors and Scherer independently, has proved to be very useful for robustness analysis and synthesis of control systems. In particular, this procedure provides a nonconservative way to convert specific inequality conditions on lossless sets into numerically verifiable conditions represented by linear matrix inequalities (LMIs). In this paper, we introduce a new notion, one-vector-lossless sets, and propose a generalized  $\mathcal{S}$ -procedure to reduce inequality conditions on one-vector-lossless sets into LMIs without any conservatism. By means of the proposed generalized  $\mathcal{S}$ -procedure, we can examine various properties of matrix-valued functions over some regions on the complex plane. To illustrate the usefulness, we show that full rank property analysis problems of polynomial matrices over some specific regions on the complex plane can be reduced into LMI feasibility problems. It turns out that many existing results such as Lyapunov's inequalities and LMIs for state-feedback controller synthesis readily follow from the suggested generalized  $\mathcal{S}$ -procedure.

**Key words.**  $\mathcal{S}$ -procedure, one-vector-lossless set, linear matrix inequalities

**AMS subject classifications.** 90C22, 90C25, 93B51, 93C05

**DOI.** 10.1137/050627551

**1. Introduction.** Recently, the generalized version of the  $\mathcal{S}$ -procedure has been introduced independently by Iwasaki and Hara [7, 8], Iwasaki, Meinsma, and Fu [9], Iwasaki and Shibata [10], and Scherer [14, 15, 16, 17]. Basically speaking, this procedure enables us to convert intractable semi-infinite parametrized linear matrix inequalities into numerically verifiable finite-dimensional linear matrix inequalities (LMIs). The scope of its application is wide and includes a variety of robustness analysis and synthesis problems in linear control system theory.

Among these recent papers, in [9, 8], the following inequality condition with respect to a Hermitian matrix  $\Theta$  and a subset  $\mathcal{S}$  of Hermitian matrices is discussed:

$$(1.1) \quad \zeta^* \Theta \zeta > 0 \quad \forall \zeta \in \mathcal{G}, \quad \mathcal{G} := \{\zeta \in \mathbb{C}^n : \zeta \neq 0, \quad \zeta^* S \zeta \geq 0 \quad \forall S \in \mathcal{S}\}.$$

It can be easily seen that a sufficient condition for (1.1) is given by

$$(1.2) \quad \exists S \in \mathcal{S} \text{ such that } \Theta \succ S.$$

The procedure to replace (1.1) by (1.2) is called the generalized  $\mathcal{S}$ -procedure in [9, 8]. Generally, this replacement introduces conservatism; the condition (1.2) is only sufficient for (1.1) and may not be necessary. The significance of the studies in [9, 8] lies in the fact that the generalized  $\mathcal{S}$ -procedure has been proved to be nonconservative if the set  $\mathcal{S}$  is *lossless*<sup>1</sup> [9, 8]. If the set  $\mathcal{S}$  is lossless, then the set  $\mathcal{S}$  is convex and hence

\*Received by the editors March 25, 2005; accepted for publication (in revised form) February 8, 2008; published electronically May 14, 2008. This work was supported in part by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 18760319. A conference version of this paper was presented at the 43rd IEEE Conference on Decision and Control, The Bahamas, 2004.

<http://www.siam.org/journals/sicon/47-3/62755.html>

<sup>†</sup>Department of Electrical Engineering, Kyoto University, Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto 615-8510, Japan (ebihara@kuee.kyoto-u.ac.jp, maeda@jaguar.kuee.kyoto-u.ac.jp, hagiwara@kuee.kyoto-u.ac.jp).

<sup>1</sup>In the paper [8], the terminology “rank-one separable” is used in place of “lossless.”



the LMI condition (1.2) can be verified numerically via sophisticated interior-point methods [1, 4].

When we deal with linear system analysis and synthesis problems by working with the generalized  $\mathcal{S}$ -procedure in [9, 8], the underlying idea is that inequality conditions on matrix-valued functions  $G(\lambda)$  over *curves*  $\lambda \in \Lambda$  ( $\Lambda \subset \mathbf{C}$ ) can be reformulated into a conformable form to the condition (1.1) by considering an appropriate Hermitian matrix  $\Theta$  and a lossless set  $\mathcal{S}$  [9, 8]. When dealing with a linear system, its various properties can be characterized by inequality conditions on their transfer functions in the frequency domain [1, 20, 21]. In [9, 8], those frequency domain inequalities are reformulated in the form of (1.1) so that the generalized  $\mathcal{S}$ -procedure can be applied. It follows that we can verify various properties of linear systems without introducing any conservatism by solving LMIs resulting from the generalized  $\mathcal{S}$ -procedure.

For linear system analysis and synthesis, however, we also need to verify inequality conditions on matrix-valued functions  $G(\lambda)$  over *region*  $\lambda \in \mathcal{D}$  ( $\mathcal{D} \subset \mathbf{C}$ ). For example, full rank property analysis of polynomial matrices over some specific regions on the complex plane forms an important basis for the stability analysis of linear systems [1, 6]. In view of these facts, it should be quite natural to pose the following question: Can we verify various properties of matrix-valued functions  $G(\lambda)$  over *region*  $\lambda \in \mathcal{D}$  ( $\mathcal{D} \subset \mathbf{C}$ ) by following similar lines to the generalized  $\mathcal{S}$ -procedure?

To answer this question, in this paper, we first introduce a new notion, *one-vector-lossless sets*, and provide a nonconservative generalized  $\mathcal{S}$ -procedure for inequality conditions on this set. More precisely, by taking account of the fact that the properties of lossless sets are fully used to represent *curves* on the complex plane [9, 8], we first consider to relax the requirements for the lossless sets and define one-vector-lossless sets, which enables us to represent *regions* on the complex plane. Then, we clarify under what condition the generalized  $\mathcal{S}$ -procedure for inequality conditions on this set is nonconservative. It follows that we can provide a counterpart result of [9, 8] in the case of the one-vector-lossless sets.

To illustrate the usefulness of the proposed generalized  $\mathcal{S}$ -procedure, we show that full rank property analysis problems of polynomial matrices over some regions  $\mathcal{D} \subset \mathbf{C}$  can be reduced into LMI feasibility problems. It turns out that the well-known results such as Lyapunov's inequalities for stability analysis of linear systems [1, 5] and LMIs for state-feedback controller synthesis [1, 18] follow immediately from the full rank property analysis by means of the proposed generalized  $\mathcal{S}$ -procedure. Thus, in conjunction with the results in [8, 16], the present paper reveals that most LMI results in linear system theory can be captured in a unified fashion within the framework of the generalized  $\mathcal{S}$ -procedure.

We use the following notation in this paper. For a matrix  $A$ , its transpose and complex conjugate transpose are denoted by  $A^T$  and  $A^*$ , respectively. For a matrix  $A \in \mathbf{C}^{n \times m}$  with  $\text{rank}(A) = r < n$ ,  $A^\perp \in \mathbf{C}^{(n-r) \times n}$  is a matrix such that  $A^\perp A = 0$  and  $A^\perp (A^\perp)^* \succ 0$ . The symbols  $\mathbf{H}_n$  and  $\mathbf{P}_n$  denote the sets of  $n \times n$  Hermitian matrices and positive-definite Hermitian matrices, respectively. For matrices  $\Psi$  and  $P$ , we denote by  $\Psi \otimes P$  their Kronecker product. For  $\lambda \in \mathbf{C}$  and  $\Psi \in \mathbf{H}_2$ , we define a function  $\sigma : \mathbf{C} \times \mathbf{H}_2 \rightarrow \mathbf{R}$  by

$$\sigma(\lambda, \Psi) := \begin{bmatrix} \lambda \\ 1 \end{bmatrix}^* \Psi \begin{bmatrix} \lambda \\ 1 \end{bmatrix}.$$

**2. Generalized  $\mathcal{S}$ -procedure for inequality conditions on one-vector-lossless sets.** The notion of *one-vector-lossless sets* plays an important role in this

paper. In this section, we first describe its precise definition and provide a nonconservative generalized  $\mathcal{S}$ -procedure for inequality conditions on this set.

**DEFINITION 2.1** (one-vector-lossless sets). *A subset  $\mathcal{S} \subset \mathbf{H}_n$  is said to be one-vector-lossless if it has the following properties:*

- (a)  $\mathcal{S}$  is convex.
- (b)  $S \in \mathcal{S} \Rightarrow \tau S \in \mathcal{S} \ \forall \tau > 0$ .
- (c) For each nonzero matrix  $H \in \mathbf{C}^{n \times n}$  with rank  $r$  that satisfies

$$(2.1) \quad H = H^* \succeq 0, \quad \text{trace}(SH) \geq 0 \quad \forall S \in \mathcal{S},$$

there exist vectors  $\zeta_i \in \mathbf{C}^n$  ( $i = 1, \dots, r$ ) such that  $H = \sum_{i=1}^r \zeta_i \zeta_i^*$  and the condition  $\zeta_j^* S \zeta_j \geq 0$  ( $\forall S \in \mathcal{S}$ ) holds for at least one index  $j$ .

This definition has been introduced by relaxing the requirements for the *lossless* sets given in [9]. Indeed, Definition 2.1 becomes the requirements for the lossless sets by replacing (c) by (c') given in the following:

- (c') For each nonzero matrix  $H \in \mathbf{C}^{n \times n}$  with rank  $r$  that satisfies (2.1), there exist vectors  $\zeta_i \in \mathbf{C}^n$  ( $i = 1, \dots, r$ ) such that

$$H = \sum_{i=1}^r \zeta_i \zeta_i^*, \quad \zeta_i^* S \zeta_i \geq 0 \quad \forall i, \quad \forall S \in \mathcal{S}.$$

This property is referred to as *rank-one separable* in [8]. Detailed analysis on this property can also be found in [13].

By comparing the conditions (c) and (c'), we see that the condition  $\zeta_j^* S \zeta_j \geq 0$  ( $\forall S \in \mathcal{S}$ ) is required only for one index  $j$  in the definition of the one-vector-lossless sets. Hence, it is obvious that a lossless set is one-vector-lossless.

In the case where the set  $\mathcal{S}$  is lossless, the condition (1.1) can be converted into (1.2) without introducing any conservatism. The following theorem gives a counterpart of this result in the case where the set  $\mathcal{S}$  is one-vector-lossless.

**THEOREM 2.2** (the generalized  $\mathcal{S}$ -procedure for inequality conditions on one-vector-lossless sets). *Let  $\Theta \in \mathbf{H}_n$  and a one-vector-lossless set  $\mathcal{S} \subset \mathbf{H}_n$  be given. If  $\Theta = \Theta^* \succeq 0$ , then the following statements are equivalent:*

- (i)  $\zeta^* \Theta \zeta > 0 \ \forall \zeta \in \mathcal{G}$ ,  $\mathcal{G} := \{\zeta \in \mathbf{C}^n : \zeta \neq 0, \zeta^* S \zeta \geq 0 \ \forall S \in \mathcal{S}\}$ .
- (ii) There exists  $S \in \mathcal{S}$  such that  $\Theta \succ S$ .

*Proof.* (ii)  $\Rightarrow$  (i). Suppose (ii) holds. Then, there exists  $S_0 \in \mathcal{S}$  such that  $\zeta^* (\Theta - S_0) \zeta > 0$  ( $\forall \zeta \neq 0$ ). This inequality implies that

$$\zeta^* \Theta \zeta > 0 \quad \forall \zeta \in \mathcal{G}_0, \quad \mathcal{G}_0 := \{\zeta \in \mathbf{C}^n : \zeta \neq 0, \quad \zeta^* S_0 \zeta \geq 0\}.$$

Since  $\mathcal{G} \subset \mathcal{G}_0$ , we can conclude that the condition (ii) implies (i).

(i)  $\Rightarrow$  (ii). Suppose (ii) does not hold, i.e., there is no  $S \in \mathcal{S}$  such that  $\Theta \succ S$ . Then, since  $\mathcal{S}$  is convex, it follows from the separating hyperplane theorem [11] that there exists a nonzero matrix  $H \in \mathbf{C}^{n \times n}$  such that

$$(2.2) \quad H = H^* \succeq 0, \quad \text{trace}((\Theta - S)H) \leq 0 \quad \forall S \in \mathcal{S}.$$

In view of the property (b) of the one-vector-lossless set, we see that the following conditions are necessary for the second condition in (2.2) to hold:

$$(2.3) \quad \text{trace}(\Theta H) \leq 0, \quad \text{trace}(SH) \geq 0 \quad \forall S \in \mathcal{S}.$$

Since  $\mathcal{S}$  is one-vector-lossless, it follows from the property (c) of Definition 2.1 that the second condition from (2.3) implies the existence of the vectors  $\zeta_i$  ( $i = 1, \dots, r$ ) such that  $H = \sum_{i=1}^r \zeta_i \zeta_i^*$  and  $\zeta_j^* S \zeta_j \geq 0$  ( $\forall S \in \mathcal{S}$ ) for some  $j$ , where  $r$  is the rank of  $H$ . For those vectors  $\zeta_i$ , the first condition in (2.3) implies  $\zeta_i^* \Theta \zeta_i = 0$  ( $i = 1, \dots, r$ ) due to the assumption  $\Theta = \Theta^* \succeq 0$ . These facts in particular imply that  $\zeta_j^* \Theta \zeta_j = 0$  and  $\zeta_j \in \mathcal{G}$  for at least one index  $j$ . This clearly contradicts the condition (i).  $\square$

We note that, in comparison with the case where the set  $\mathcal{S}$  is lossless [9], an additional condition  $\Theta = \Theta^* \succeq 0$  has been imposed in Theorem 2.2. This could be regarded as a price to pay for relaxing the requirements on the set  $\mathcal{S}$  from a lossless one to a one-vector-lossless one.

By means of the generalized  $\mathcal{S}$ -procedure in Theorem 2.2, we can convert the semi-infinite inequality condition (i) into the numerically verifiable LMI condition in (ii). Hence, when we deal with control system analysis and synthesis problems at hand, a crucial step is to reduce those problems into a form conformable to the condition (i). This step is not obvious in general. When exploring such reduction, it is indispensable to see concretely what sets are indeed one-vector-lossless. In the next theorem, we will show a class of one-vector-lossless sets that is relevant to control system analysis and synthesis.

**THEOREM 2.3.** *Let  $\Psi \in \mathbf{H}_2$  with  $\det(\Psi) < 0$  and  $\Gamma \in \mathbf{C}^{2n \times l}$  be given. Define a subset of Hermitian matrices by*

$$(2.4) \quad \mathcal{S} := \{\Gamma^*(\Psi \otimes P)\Gamma : P \in \mathbf{P}_n\}.$$

*Then the set  $\mathcal{S}$  is one-vector-lossless.*

*Proof.* The proof is rather technical and thus given in the appendix.  $\square$

It is meaningful to examine the property of one-vector-lossless set  $\mathcal{S}$  given by (2.4) in comparison with the lossless set  $\mathcal{S}_1$  discussed in [8], where

$$\mathcal{S}_1 := \{\Gamma^*(\Psi \otimes P)\Gamma : P \in \mathbf{H}_n\}.$$

To see a significant difference between these two sets, let us take  $\Psi = \text{diag}(-1, 1)$  and  $\Gamma = I_{2n}$  for simplicity and consider the following set that concerns the condition (i) in Theorem 2.2:

$$(2.5) \quad \mathcal{G} := \left\{ \begin{bmatrix} f_1 \\ f_0 \end{bmatrix} \in \mathbf{C}^{2n} : f_0, f_1 \in \mathbf{C}^n, \begin{bmatrix} f_1 \\ f_0 \end{bmatrix} \neq 0, \begin{bmatrix} f_1 \\ f_0 \end{bmatrix}^* S \begin{bmatrix} f_1 \\ f_0 \end{bmatrix} \geq 0 \quad \forall S \in \mathcal{S} \right\}.$$

Then, we can show that the above set defined from the one-vector-lossless set  $\mathcal{S}$  coincides with

$$(2.6) \quad \mathcal{L} := \left\{ \begin{bmatrix} f_1 \\ f_0 \end{bmatrix} \in \mathbf{C}^{2n} : f_1 = s f_0 \text{ for some } s \in \overline{\mathbf{D}} \right\},$$

where  $\overline{\mathbf{D}}$  denotes the closure of the open unit disc  $\mathbf{D}$  on the complex plane. On the other hand, if we replace the one-vector-lossless set  $\mathcal{S}$  in (2.5) by the lossless set  $\mathcal{S}_1$ , then the resulting set  $\mathcal{G}_1$  coincides with the set  $\mathcal{L}_1$  obtained by replacing  $\overline{\mathbf{D}}$  in (2.6) by  $\partial\mathbf{D}$ . These observations clearly indicate that the lossless sets are related to *curves* on the complex plane, while the one-vector-lossless sets are related to *regions* on the complex plane. This is the key observation to develop the generalized  $\mathcal{S}$ -procedure for inequality conditions on the one-vector-lossless sets. We show in the next section that full rank property analysis problems of polynomial matrices over some regions on the complex plane can be dealt with by the proposed generalized  $\mathcal{S}$ -procedure.

**3. Linear system analysis using generalized  $\mathcal{S}$ -procedure.** For given complex matrices  $M_k \in \mathbf{C}^{n \times m}$  ( $k = 0, \dots, N$ ) with  $n \geq m$ , let us consider the  $n \times m$  complex polynomial matrix represented by  $M(s) = \sum_{k=0}^N s^k M_k$ . We assume that the normal rank of  $M(s)$  is  $m$ . Following the discussions in [6, 19], we define a (finite) zero of  $M(s)$  as a complex value  $z \in \mathbf{C}$  for which the rank of  $M(s)$  drops from its normal value, i.e.,  $\text{rank}(M(z)) < m$ . In linear system analysis and synthesis, it is of great importance to determine whether the zeros of given polynomial matrix  $M(s)$  belong to a specific region  $\mathcal{D} \subset \mathbf{C}$ . This can be restated equivalently in the way that the polynomial matrix  $M(s)$  is of full-column rank for all  $s \in \mathcal{D}^c$ , where  $\mathcal{D}^c$  denotes the complement of the region  $\mathcal{D}$  in  $\mathbf{C}$ . In the subsequent discussions, we restrict our attention to the regions defined below.

**DEFINITION 3.1.** For given  $\Psi \in \mathbf{H}_2$  with  $\det(\Psi) < 0$ , we define a set  $\mathcal{D}_\Psi$  and its complement  $\mathcal{D}_\Psi^c$  by

$$(3.1) \quad \mathcal{D}_\Psi := \{\lambda \in \mathbf{C} : \sigma(\lambda, \Psi) < 0\}, \quad \mathcal{D}_\Psi^c := \{\lambda \in \mathbf{C} : \sigma(\lambda, \Psi) \geq 0\}.$$

By selecting the Hermitian matrix  $\Psi$  in (3.1) appropriately, we can obtain several important regions in linear system analysis and synthesis. In particular, by letting

$$(3.2) \quad \Psi_c := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \Psi_d := \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

we see that  $\mathcal{D}_{\Psi_c}$  and  $\mathcal{D}_{\Psi_d}$  coincide with the open left half plane  $\mathbf{C}_-$  and the open unit disc  $\mathbf{D}$ , respectively. These regions are particularly important for stability analysis of continuous- and discrete-time linear systems.

We are now in the right position to show that the full rank property analysis problems of polynomial matrices can be reduced into LMI feasibility problems by means of the proposed generalized  $\mathcal{S}$ -procedure. We note that such reduction into LMIs is also investigated in the preceding studies, and similar results to the next theorem can also be found in the literature; see, for example, [6].

**THEOREM 3.2.** Let complex matrices  $M_k \in \mathbf{C}^{n \times m}$  ( $k = 0, \dots, N$ ) with  $n \geq m$  and  $\Psi = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{12}^* & \psi_{22} \end{bmatrix} \in \mathbf{H}_2$  with  $\det(\Psi) < 0$  be given, and define  $M(s) := \sum_{k=0}^N s^k M_k$ ,  $\mathcal{M} := [M_N \cdots M_0]$ . Suppose either of the following assumptions holds:

1.  $M_N$  is of full-column rank.
2.  $\psi_{11} < 0$ .

Then, the following conditions are equivalent:

- (i) The polynomial matrix  $M(s)$  is of full-column rank for all  $s \in \mathcal{D}_\Psi^c$ .
- (ii)

$$\begin{aligned} & f^* \mathcal{M}^* \mathcal{M} f > 0 \quad \forall f \in \mathcal{L}, \\ & \mathcal{L} := \{f \in \mathbf{C}^{(N+1)m} : f \neq 0, \quad f^* S f \geq 0 \quad \forall S \in \mathcal{S}_W\}, \\ & \mathcal{S}_W := \{W^*(\Psi \otimes P)W : P \in \mathbf{P}_{Nm}\}, \\ & W := \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}, \quad W_1 := \begin{bmatrix} I_{Nm} \\ 0_{m, Nm} \end{bmatrix}^*, \quad W_2 := \begin{bmatrix} 0_{m, Nm} \\ I_{Nm} \end{bmatrix}^*. \end{aligned}$$

- (iii) There exists  $P \in \mathbf{P}_{Nm}$  such that

$$(3.3) \quad \mathcal{M}^* \mathcal{M} - W^*(\Psi \otimes P)W \succ 0.$$

If the matrices  $M_k$  ( $k = 0, \dots, N$ ) and  $\Psi$  are all real, then the equivalence still holds when we restrict  $P$  to be real.

*Proof.* The proof for the equivalence of (i) and (ii) is given in the appendix. The main step of the proof, the equivalence of (ii) and (iii), follows immediately from Theorems 2.2 and 2.3. Indeed, the set  $\mathcal{S}_W$  is one-vector-lossless by Theorem 2.3 while it is clear that  $\mathcal{M}^*\mathcal{M} \succeq 0$ . Hence the generalized  $\mathcal{S}$ -procedure in Theorem 2.2 establishes the equivalence of (ii) and (iii). Noting that the real case results can be shown by following similar arguments to [9], we complete the proof.  $\square$

From this theorem, we see that full rank property of polynomial matrices can be assessed by simply solving the LMI (3.3), provided that either assumption 1 or 2 is satisfied. From the definition of  $\mathcal{D}_\Psi^c$  in (3.1), we see that the assumption  $\psi_{11} < 0$  enforces the region  $\mathcal{D}_\Psi^c$  to be bounded. To put it another way, our assumption requires that the matrix  $M_N$  is of full-column rank if the region  $\mathcal{D}_\Psi^c$  is unbounded. When studying the full rank property of polynomial matrices over unbounded regions, it is well-known that we have to take a special care on zeros at infinity [2, 3, 6, 19], and the assumption 1 is surely a sufficient condition for the absence of the zeros at infinity. Hence, under the assumption 1, delicate problems stemming from zeros at infinity have been excluded from our discussions.

It is obvious that the result in Theorem 3.2 forms an important basis for dealing with stability related issues in linear system analysis. In particular, the (generalized) Lyapunov's inequality [1, 5] is surely a special case of (3.3). In addition, existing LMI results for  $\mathcal{D}$ -stabilizability also follow from Theorem 3.2, where a matrix pair  $A \in \mathbf{C}^{n \times n}$ ,  $B \in \mathbf{C}^{n \times m}$  is said to be  $\mathcal{D}$ -stabilizable iff there exists  $K \in \mathbf{C}^{m \times n}$  such that  $sI - (A + BK)$  is nonsingular for all  $s \in \mathcal{D}^c$ . From the Popov–Belevitch–Hautus (PBH) tests [20, 21], this condition can be restated equivalently as  $[sI - A \quad B]^*$  is of full-column rank for all  $s \in \mathcal{D}^c$ . Hence, for the region  $\mathcal{D}_\Psi$ , we can conclude from Theorem 3.2 that the pair  $(A, B)$  is  $\mathcal{D}_\Psi$ -stabilizable iff there exists  $P \in \mathbf{P}_n$  such that

$$\begin{bmatrix} I_n & -A^* \\ 0_{m,n} & B^* \end{bmatrix}^* \begin{bmatrix} I_n & -A^* \\ 0_{m,n} & B^* \end{bmatrix} - \Psi^T \otimes P \succ 0.$$

From Finsler's lemma [1], this LMI can be rewritten as

$$(3.4) \quad B^\perp \begin{bmatrix} A & I \end{bmatrix} (\Psi^T \otimes P) \begin{bmatrix} A^* \\ I \end{bmatrix} (B^\perp)^* \prec 0.$$

The condition (3.4) is known as the *elimination-of-variables* type LMI condition for state-feedback stabilizing controller synthesis with respect to the stability region  $\mathcal{D}_\Psi$  [18]. In this way, we can derive existing stability-related LMI conditions straightforwardly by means of the generalized  $\mathcal{S}$ -procedure for inequality conditions on one-vector-lossless set.

**4. Conclusion.** In this paper, we first introduced a new notion, one-vector-lossless sets, and provided a nonconservative generalized  $\mathcal{S}$ -procedure for inequality conditions on the one-vector-lossless sets. We next showed that full rank property analysis problems of polynomial matrices over some regions on the complex plane can be reduced into LMI feasibility problems by means of the proposed generalized  $\mathcal{S}$ -procedure. It turned out that many existing results such as Lyapunov's inequalities for stability analysis of linear systems and LMIs for state-feedback controller synthesis can be viewed as particular cases of this result. To summarize, in conjunction with [8, 16], this paper clarified that most LMI results in linear control system theory can be grasped within the unified framework of the generalized  $\mathcal{S}$ -procedure.

## 5. Appendix.

**5.1. Proof of Theorem 2.3.** We need the following lemma for the proof.

LEMMA 5.1. For given  $F, G \in \mathbf{C}^{n \times m}$ , and  $\Psi = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{12}^* & \psi_{22} \end{bmatrix} \in \mathbf{H}_2$  with  $\det(\Psi) < 0$ , suppose

$$(5.1) \quad \psi_{11}FF^* + \psi_{12}^*FG^* + \psi_{12}GF^* + \psi_{22}GG^* \succeq 0.$$

Then, there exists a unitary matrix  $U \in \mathbf{C}^{m \times m}$  such that

$$(5.2) \quad \psi_{11}\tilde{f}_1\tilde{f}_1^* + \psi_{12}^*\tilde{f}_1\tilde{g}_1^* + \psi_{12}\tilde{g}_1\tilde{f}_1^* + \psi_{22}\tilde{g}_1\tilde{g}_1^* \succeq 0,$$

where

$$(5.3) \quad [\tilde{f}_1 \ \cdots \ \tilde{f}_m] := FU, \quad [\tilde{g}_1 \ \cdots \ \tilde{g}_m] := GU, \quad \tilde{f}_i, \tilde{g}_i \in \mathbf{C}^n \ (i = 1, \dots, m).$$

*Proof.* We give the proof only for the case  $\psi_{11} > 0$ . Other cases can be proved similarly. If  $\psi_{11} > 0$ , then the condition (5.1) can be rewritten equivalently as

$$\left(F + \frac{\psi_{12}}{\psi_{11}}G\right) \left(F + \frac{\psi_{12}}{\psi_{11}}G\right)^* \succeq -\psi_{11}^{-2}\det(\Psi)GG^*.$$

From [12], this condition holds iff there exists a matrix  $W \in \mathbf{C}^{m \times m}$  such that

$$(5.4) \quad \sqrt{-\psi_{11}^{-2}\det(\Psi)}G = \left(F + \frac{\psi_{12}}{\psi_{11}}G\right)W, \quad \|W\| \leq 1.$$

Since  $\|W\| \leq 1$ , for each eigenvalue  $\lambda$  of  $W$  and its associated eigenvector  $\xi$ , we have  $W\xi = \lambda\xi$ ,  $|\lambda| \leq 1$ ,  $\xi^*\xi = 1$ . Taking one such  $\xi$ , we can construct a unitary matrix  $U$  of the form  $U = \begin{bmatrix} \xi & \bar{U} \end{bmatrix}$ ,  $\bar{U} \in \mathbf{C}^{m \times (m-1)}$ . Then, we see from (5.4) that the vectors  $\tilde{f}_1$  and  $\tilde{g}_1$  defined by (5.3) with this unitary matrix  $U$  satisfy

$$\sqrt{-\psi_{11}^{-2}\det(\Psi)}\tilde{g}_1 = \left(\tilde{f}_1 + \frac{\psi_{12}}{\psi_{11}}\tilde{g}_1\right)\lambda \quad (|\lambda| \leq 1).$$

This implies (5.2).  $\square$

Now we are ready to prove Theorem 2.3.

*Proof.* It is obvious that the set  $\mathcal{S}$  given in (2.4) has the properties (a) and (b) in Definition 2.1. To prove the property (c), let  $H \in \mathbf{C}^{l \times l}$  be a nonzero matrix that satisfies (2.1). In addition, we denote the full rank factorization of  $H$  by

$$(5.5) \quad H = LL^*, \quad L \in \mathbf{C}^{l \times r}, \quad r := \text{rank}(H).$$

With this  $L$ , define

$$(5.6) \quad \begin{bmatrix} F \\ G \end{bmatrix} := \Gamma L, \quad F, G \in \mathbf{C}^{n \times r}.$$

Then, the second condition in (2.1) can be rewritten as

$$\text{trace}((\psi_{11}FF^* + \psi_{12}^*FG^* + \psi_{12}GF^* + \psi_{22}GG^*)P) \geq 0 \quad \forall P \in \mathbf{P}_n.$$

It can be seen that this condition holds iff (5.1) holds. Hence, from Lemma 5.1, there exists a unitary matrix  $U \in \mathbf{C}^{r \times r}$  that satisfies (5.2) with  $\tilde{f}_i, \tilde{g}_i \in \mathbf{C}^n$  ( $i = 1, \dots, m$ ) given by (5.3). Here, note that (5.2) is equivalent to

$$(5.7) \quad \text{trace}((\psi_{11}\tilde{f}_1\tilde{f}_1^* + \psi_{12}^*\tilde{f}_1\tilde{g}_1^* + \psi_{12}\tilde{g}_1\tilde{f}_1^* + \psi_{22}\tilde{g}_1\tilde{g}_1^*)P) \geq 0 \quad \forall P \in \mathbf{P}_n.$$

By defining  $[\zeta_1 \ \cdots \ \zeta_r] := LU$ , we have  $H = \sum_{i=1}^r \zeta_i \zeta_i^*$ . On the other hand, from (5.6) and (5.3), it is apparent that  $[\tilde{f}_1^* \ \tilde{g}_1^*]^* = \Gamma \zeta_1$ . Hence, we see from (5.7) that the

condition  $\zeta_1^* \Gamma^*(\Psi \otimes P) \Gamma \zeta_1 \geq 0$  ( $\forall P \in \mathbf{P}_n$ ) holds or, equivalently,  $\zeta_1^* S \zeta_1 \geq 0$  ( $\forall S \in \mathcal{S}$ ). This clearly shows that the set  $\mathcal{S}$  satisfies the property (c) of Definition 2.1.  $\square$

**5.2. Proof of the equivalence of (i) and (ii) in Theorem 3.2.** The condition (i) can be restated equivalently as

$$\begin{aligned} & f^* \mathcal{M}^* \mathcal{M} f > 0 \quad \forall f \in \mathcal{K}, \\ \mathcal{K} := \{ & f = [f_N^* \cdots f_0^*]^* \in \mathbf{C}^{(N+1)m} : f_0 \neq 0, \\ & \exists s \in \mathcal{D}_{\Psi}^c \text{ such that } f_{k+1} = s f_k \ (k = 0, \dots, N-1) \}. \end{aligned}$$

Hence, to prove the equivalence of (i) and (ii), it suffices to show that  $\mathcal{K} = \mathcal{L}$ . To this end, we first note that  $\mathcal{K} = \mathcal{L}'$ , where

$$\begin{aligned} \mathcal{L}' := \{ & f = [f_N^* \cdots f_0^*]^* \in \mathbf{C}^{(N+1)m} : \\ & f_k \in \mathbf{C}^m \ (k = 0, \dots, N), \\ & f_0 \neq 0, \quad f^* S f \geq 0 \quad \forall S \in \mathcal{S}_W \}. \end{aligned}$$

To see this, suppose  $f = [f_N^* \cdots f_0^*]^* \in \mathcal{K}$ , and define  $f_u, f_l \in \mathbf{C}^{Nm}$  by

$$(5.8) \quad f_u := [f_N^* \cdots f_1^*]^* = W_1 f, \quad f_l := [f_{N-1}^* \cdots f_0^*]^* = W_2 f.$$

Then, from the definition of  $\mathcal{K}$ , the following inequality holds for all  $P \in \mathbf{P}_{Nm}$ :

$$\begin{aligned} f^* W^* (\Psi \otimes P) W f &= \begin{bmatrix} f_u \\ f_l \end{bmatrix}^* (\Psi \otimes P) \begin{bmatrix} f_u \\ f_l \end{bmatrix} \\ &= \begin{bmatrix} s f_l \\ f_l \end{bmatrix}^* (\Psi \otimes P) \begin{bmatrix} s f_l \\ f_l \end{bmatrix} = \sigma(s, \Psi) f_l^* P f_l \geq 0. \end{aligned}$$

This shows that  $f \in \mathcal{L}'$ , and hence  $\mathcal{K} \subset \mathcal{L}'$ . On the other hand, suppose  $f = [f_N^* \cdots f_0^*]^* \in \mathcal{L}'$ , and define  $f_u$  and  $f_l$  by (5.8). Then, from the definition of  $\mathcal{L}'$ , we have

$$\text{trace}((\psi_{11} f_u f_u^* + \psi_{12}^* f_u f_l^* + \psi_{12} f_l f_u^* + \psi_{22} f_l f_l^*) P) \geq 0 \quad \forall P \in \mathbf{P}_{Nm}.$$

It can be easily seen that the above condition implies

$$(5.9) \quad \psi_{11} f_u f_u^* + \psi_{12}^* f_u f_l^* + \psi_{12} f_l f_u^* + \psi_{22} f_l f_l^* \succeq 0.$$

Moreover, under the assumption  $f_0 \neq 0$ , we have  $f_l \neq 0$ , and hence (5.9) holds iff  $f_u = s f_l$  for some  $s \in \mathcal{D}_{\Psi}^c$  [12]. This clearly shows that  $f \in \mathcal{K}$  and hence  $\mathcal{L}' \subset \mathcal{K}$ . Thus we can conclude  $\mathcal{K} = \mathcal{L}'$ .

To complete the proof, note that  $\mathcal{L} = \mathcal{L}' \cup \mathcal{J}'$ , where

$$\begin{aligned} \mathcal{J}' := \{ & f = [f_N^* \cdots f_0^*]^* \in \mathbf{C}^{(N+1)m} : \\ & f_k \in \mathbf{C}^m \ (k = 0, \dots, N), \quad f \neq 0, \quad f_0 = 0, \quad f^* S f \geq 0 \quad \forall S \in \mathcal{S}_W \}. \end{aligned}$$

Moreover, we can show that the set  $\mathcal{J}'$  is equivalent to

$$\begin{aligned} \mathcal{J} := \{ & f = [f_N^* \cdots f_0^*]^* \in \mathbf{C}^{(N+1)m} : \\ & f_k \in \mathbf{C}^m \ (k = 0, \dots, N), \\ & f_N \neq 0, \quad f_k = 0 \ (k = 0, \dots, N-1), \quad f^* S f \geq 0 \quad \forall S \in \mathcal{S}_W \}. \end{aligned}$$

To see the equivalence of  $\mathcal{J}'$  and  $\mathcal{J}$ , let us take a vector  $f \in \mathcal{J}'$ . Furthermore, define  $f_u$  and  $f_l$  by (5.8) and suppose  $f_l \neq 0$ . Then, the vectors  $f_u$  and  $f_l$  should satisfy

(5.9), and thus  $f_u = sf_l$  holds for some  $s \in \mathcal{D}_\Psi^c$ . Since  $f_0 = 0$ , however,  $f_u = sf_l$  implies  $f_l = 0$ . This clearly contradicts the assumption  $f_l \neq 0$ . Hence, we have  $f_l = 0$  and thus  $f \in \mathcal{J}$ . This shows that  $\mathcal{J}' \subset \mathcal{J}$ . On the other hand, it is apparent that  $\mathcal{J} \subset \mathcal{J}'$ , and hence we have  $\mathcal{J}' = \mathcal{J}$ .

Summarizing the above arguments,  $\mathcal{L} = \mathcal{K} \cup \mathcal{J}$  holds. Hence, the equivalence of (i) and (ii) can be verified by showing that the condition  $f^* \mathcal{M}^* \mathcal{M} f > 0$  holds for all  $f \in \mathcal{J}$  under either assumption 1 or 2. If  $\psi_{11} < 0$ , however, it can be easily seen that the set  $\mathcal{J}$  is empty. On the other hand, if  $\psi_{11} \geq 0$ , then we have  $M_N^* M_N \succ 0$  from the assumption. This indicates that  $f^* \mathcal{M}^* \mathcal{M} f = f_N^* M_N^* M_N f_N > 0$  ( $\forall f \in \mathcal{J}$ ). Hence, we can conclude that the conditions (i) and (ii) are equivalent under either assumption 1 or 2. This completes the proof.  $\square$

## REFERENCES

- [1] S. P. BOYD, L. E. GHAOUI, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, SIAM Studies in Applied Mathematics 15, SIAM, Philadelphia, 1994.
- [2] F. M. CALLIER, *On polynomial matrix spectral factorization by symmetric extraction*, IEEE Trans. Automat. Control, 30 (1985), pp. 453–464.
- [3] P. VAN DOOREN, P. DEWILDE, AND J. VANDEWALLE, *On the determination of the Smith-McMillan form of a rational matrix from its Laurent expansion*, IEEE Trans. Circuits Systems, 26 (1979), pp. 180–189.
- [4] P. GAHINET, A. NEMIROVSKII, A. J. LAUB, AND M. CHILALI, *LMI Control Toolbox*, The MathWorks, Natick, MA, 1994.
- [5] D. HENRION AND G. MEINSMA, *Rank-one LMIs and Lyapunov's inequality*, IEEE Trans. Automat. Control, 46 (2001), pp. 1285–1288.
- [6] D. HENRION, O. BACHELIER, AND M. ŠEBEK,  *$\mathcal{D}$ -stability of polynomial matrices*, Internat. J. Control, 74 (2001), pp. 845–856.
- [7] T. IWASAKI AND S. HARA, *Well-posedness of feedback systems: Insight into exact robustness analysis and approximate computations*, IEEE Trans. Automat. Control, 43 (1998), pp. 619–630.
- [8] T. IWASAKI AND S. HARA, *Generalized KYP lemma: Unified frequency domain inequalities with design applications*, IEEE Trans. Automat. Control, 50 (2005), pp. 41–59.
- [9] T. IWASAKI, G. MEINSMA, AND M. FU, *Generalized  $\mathcal{S}$ -procedure and finite frequency KYP lemma*, Math. Probl. Eng., 6 (2000), pp. 305–320.
- [10] T. IWASAKI AND G. SHIBATA, *LPV system analysis via quadratic separator for uncertain implicit systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 1195–1208.
- [11] G. MEINSMA, Y. SHRIVASTAVA, AND M. FU, *A dual formulation of mixed  $\mu$  and on the losslessness of  $(D, G)$  scaling*, IEEE Trans. Automat. Control, 42 (1997), pp. 1032–1036.
- [12] A. RANTZER, *On the Kalman-Yakubovich-Popov lemma*, Systems Control Lett., 28 (1996), pp. 7–10.
- [13] J. F. STURM AND S. ZHANG, *On cones of nonnegative quadratic functions*, Math. Oper. Res., 28 (2003), pp. 246–267.
- [14] C. W. SCHERER, *LPV control and full block multipliers*, Automatica J. IFAC, 37 (2001), pp. 361–375.
- [15] C. W. SCHERER, *LMI Relaxations in Robust Control: How to Reduce Conservatism?*, Plenary talk at the 4th IFAC Symposium on Robust Control Design, Milan, Italy, 2003.
- [16] C. W. SCHERER, *Relaxations for robust linear matrix inequality problems with verifications for exactness*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 365–395.
- [17] C. W. SCHERER, *LMI relaxations in robust control*, Eur. J. Control, 12 (2006), pp. 3–29.
- [18] R. E. SKELTON, T. IWASAKI, AND K. GRIGORIADIS, *A Unified Algebraic Approach to Linear Control Design*, Taylor & Francis, London, 1998.
- [19] A. I. G. VARDULAKIS, *Linear Multivariable Control. Algebraic Analysis and Synthesis Methods*, John Wiley & Sons, Chichester, UK, 1991.
- [20] K. ZHOU, K. GLOVER, AND J. C. DOYLE, *Robust and Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1996.
- [21] K. ZHOU AND J. C. DOYLE, *Essentials of Robust Control*, Prentice-Hall, Englewood Cliffs, NJ, 1998.



## ON SECOND ORDER SHAPE OPTIMIZATION METHODS FOR ELECTRICAL IMPEDANCE TOMOGRAPHY\*

L. AFRAITES<sup>†</sup>, M. DAMBRINE<sup>†</sup>, AND D. KATEB<sup>†</sup>

**Abstract.** This paper is devoted to the analysis of a second order method for recovering the a priori unknown shape of an inclusion  $\omega$  inside a body  $\Omega$  from boundary measurement. This inverse problem—known as electrical impedance tomography—has many important practical applications and hence has been the focus of much attention during the past few years. However, to the best of our knowledge, no work has yet considered a second order approach for this problem. This paper aims to fill that void: We investigate the existence of second order derivative of the state  $u$  with respect to perturbations of the shape of the interface  $\partial\omega$ . Then we choose a cost function in order to recover the geometry of  $\partial\omega$  and derive the expression of the derivatives needed to implement the corresponding Newton method. We then investigate the stability of the process and explain why this inverse problem is severely ill-posed by proving the compactness of the Hessian at the global minimizer.

**Key words.** inverse conductivity problem, shape optimization, second order method

**AMS subject classifications.** 49Q10, 34A55, 49Q12

**DOI.** 10.1137/070687438

**1. Introduction.** Let  $\Omega$  be a bounded open set with smooth boundary in  $\mathbb{R}^N$ , with  $N \geq 3$ . Consider an  $L^\infty$  function  $\sigma$  such that there exists a real  $c$  with  $\sigma(x) \geq c > 0$ . Consider the elliptic equation

$$(1) \quad -\operatorname{div}(\sigma(x)\nabla u) = 0 \text{ in } \Omega,$$

with the Dirichlet boundary condition

$$(2) \quad u = f \text{ on } \partial\Omega.$$

Define the Dirichlet-to-Neumann map as

$$\Lambda_\sigma : f \mapsto \sigma(\partial_{\mathbf{n}}u)|_{\partial\Omega},$$

where  $u$  solves (1), (2) and  $\mathbf{n}$  is the outer unit normal vector to  $\partial\Omega$ . The inverse conductivity problem of Calderón is to determine  $\sigma$  from  $\Lambda_\sigma$ . Electrical impedance tomography aims to form an image of the conductivity distribution  $\sigma$  from the knowledge of  $\Lambda_\sigma$ . When  $\sigma$  is smooth enough, one can reconstruct  $\sigma$  from  $\Lambda_\sigma$  (see the works of Sylvester and Uhlmann [23], Nachman [16, 17], and Novikov [18]). When the conductivity distribution is only  $L^\infty$ , Astala and Päivärinta have recently shown in [4] that, in dimension two, the map  $\Lambda_\sigma$  determines  $\sigma \in L^\infty(\Omega)$ .

We are interested in a particular case of that problem: when a body is inserted inside a given object with a distinct conductivity, the question of determining its shape from boundary measurement arises in many fields of modern technology. In the context of the inverse problem of conductivity of Calderón, we restrict the range

---

\*Received by the editors April 5, 2007; accepted for publication (in revised form) November 29, 2007; published electronically May 16, 2008.

<http://www.siam.org/journals/sicon/47-3/68743.html>

<sup>†</sup>Laboratoire de Mathématiques Appliquées de Compiègne, Université de Technologie de Compiègne (lekhir.afraites@utc.fr, marc.dambrine@utc.fr, djalil.kateb@utc.fr).

of admissible conductivity distributions to the family of piecewise constant functions which take only two distinct values,  $\sigma_1, \sigma_2 > 0$ , which are assumed to be known. The conductivity distribution is then defined by an open subset  $\omega$  as

$$(3) \quad \sigma = \sigma_1 \chi_{\Omega \setminus \omega} + \sigma_2 \chi_{\omega}.$$

Here, the only unknown of the problem is  $\omega$ , a subdomain of  $\Omega$  with a smooth boundary  $\partial\omega$ ; its outer unit normal vector is denoted by  $\mathbf{n}$ . The notation  $\chi_{\omega}$  (respectively,  $\chi_{\Omega \setminus \omega}$ ) denotes the characteristic function of  $\omega$  (respectively,  $\Omega \setminus \omega$ ). The second main difference arises from practical considerations: it is unrealistic from the point of view of applications to know the full graph of Dirichlet-to-Neumann. Therefore, we will assume that one has access to a single point in that graph. This nondestructive testing problem is usually written from a numerical point of view as the minimization of a cost function: typically a least-squares matching criterion. Many authors have investigated the steepest descent method for this problem [13, 8, 11, 20, 3] with the methods of shape optimization since the unknown parameter is a geometrical domain.

This work is devoted to the study of second order methods for this problem, which has only been considered before for simplified models in [6, 2]. By introducing second order methods, one aims to reach two distinct objectives.

- On one hand, we provide all the needed material to design a Newton algorithm. By Newton algorithm, we mean the usual second order scheme of optimization. We will give differentiability results for the state function and for the objective that we have chosen to study in this work. Nevertheless, we point out that the discretization of a Newton method for this problem turns out to be very delicate; this is why, in the present paper, we will neither discuss this problem nor present numerical examples. Some simulations in the bidimensional case are presented in [1].
- On the other hand, we analyze rigorously the well-posedness of the optimization method. This is justified by the huge numerical literature devoted to the numerical study of this question in the field of inverse problems; the numerical experiments insist on the ill-posedness of this problem. We will explain the instability in the continuous settings in terms of shape optimization. We show that the shape Hessian is not coercive. In fact, its Riesz operator is compact. This property explains the instability of the minimization process.

**2. Statement of the results.** Let us describe the precise problem under consideration and the notation. We consider a bounded domain  $\Omega \subset \mathbb{R}^N$  with a  $\mathcal{C}^2$  boundary. It is filled with a material whose conductivity is  $\sigma_1$  and with an unknown inclusion  $\omega$  in  $\Omega$  of conductivity  $\sigma_2 \neq \sigma_1$ . We search to reconstruct the shape of  $\omega$  by measuring  $\partial\Omega$ , the input voltage, and the corresponding output current. In what follows, we fix  $d_0 > 0$  and consider inclusions  $\omega$  such that  $\omega \subset\subset \Omega_{d_0} = \{x \in \Omega, d(x, \partial\Omega) > d_0\}$ . We also assume that the boundary  $\partial\omega$  is of class  $\mathcal{C}^{4,\alpha}$ . The inverse problem arises when one has access to the normal vector derivative of the potential  $u$  that solves (1), (2) when the conductivity distribution is defined by (3). Assume that one knows

$$(4) \quad \sigma_1 \partial_{\mathbf{n}} u = g \text{ on } \partial\Omega.$$

Then the problem (1), (2), (4) is overdetermined. The electrical impedance tomography problem we consider is to recover the shape of  $\omega$  from the knowledge of the single Cauchy pair  $(f, g)$ .

In order to recover the shape of the inclusion  $\omega$ , an usual strategy is to minimize a cost function. Many choices are possible; however, it turns out that a Kohn

and Vogelius-type objective leads to a minimization problem with nicer properties than the least-squares fitting approaches (we refer to [3] for a comparison of different objectives with order one methods, and to [2] for the case of a perfectly insulated inclusion). Therefore, we study such a cost function in this work.

Let us define this criterion. Its distinctive feature is to involve two state functions,  $u_d$  and  $u_n$ : the state  $u_d$  solves (1), (2), while  $u_n$  solves (1), (4). The Kohn–Vogelius objective  $J_{KV}$  is then defined as

$$(5) \quad J_{KV}(\omega) = \int_{\Omega} \sigma |\nabla(u_d - u_n)|^2.$$

Let us sum up the results of this paper concerning the minimization of this objective. We first prove differentiability results for the state  $u_d$ . In what follows, we use the convention that a bold character denotes a vector. If  $\mathbf{h}$  denotes a deformation field, it can be written as  $\mathbf{h} = \mathbf{h}_\tau + h_n \mathbf{n}$  on  $\partial\omega$ . Note also that in the following lines,  $\mathbf{n}$  denotes the outer normal field to  $\partial\omega$  pointing into  $\Omega \setminus \bar{\omega}$ . Hence, for  $x \in \partial\omega$ , we define, when the limit exists,  $u^\pm(x)$  (respectively,  $(\partial_n u)^\pm(x)$ ) as the limit of  $u(x \pm t\mathbf{n}(x))$  (respectively,  $\langle \nabla u(x \pm t\mathbf{n}(x)), \mathbf{n}(x) \rangle$ ) when  $t > 0$  tends to 0. Note that  $\mathbf{h}_\tau$  is a vector while  $h_n$  is a scalar quantity.

The admissible deformation fields have to preserve  $\partial\Omega$  and the regularity of the boundaries: therefore the space of admissible fields is

$$\mathcal{H} = \{\mathbf{h} \in \mathcal{C}^{4,\alpha}(\mathbb{R}^d, \mathbb{R}^d), \text{Supp}(\mathbf{h}) \subset \Omega_{d_0}\}.$$

The following result concerns the first order derivative of the state functions  $u_d$  and  $u_n$ . It was derived in [8, 20, 3].

**THEOREM 1.** *Let  $\Omega$  be an open smooth subset of  $\mathbb{R}^N$  and let  $\omega$  be an element of  $\Omega_{d_0}$  with a boundary of class  $\mathcal{C}^{4,\alpha}$ . Then the state functions  $u_d$  and  $u_n$  are shape differentiable; furthermore their shape derivatives  $u'_d$  and  $u'_n$  belong to  $H^1(\Omega \setminus \bar{\omega}) \cup H^1(\omega)$  and satisfy*

$$(6) \quad \left\{ \begin{array}{l} \Delta u'_d = 0 \text{ in } \Omega \setminus \bar{\omega} \text{ and in } \omega, \\ [u'_d] = h_n \frac{[\sigma]}{\sigma_1} \partial_{\mathbf{n}} u_d^- \text{ on } \partial\omega, \\ [\sigma \partial_n u'_d] = [\sigma] \text{div}_\tau(h_n \nabla_\tau u_d) \text{ on } \partial\omega, \\ u'_d = 0 \text{ on } \partial\Omega. \end{array} \right.$$

$$(7) \quad \left\{ \begin{array}{l} \Delta u'_n = 0 \text{ in } \Omega \setminus \bar{\omega} \text{ and in } \omega, \\ [u'_n] = h_n \frac{[\sigma]}{\sigma_1} \partial_{\mathbf{n}} u_n^- \text{ on } \partial\omega, \\ [\sigma \partial_n u'_n] = [\sigma] \text{div}_\tau(h_n \nabla_\tau u_n) \text{ on } \partial\omega, \\ \partial u'_n = 0 \text{ on } \partial\Omega. \end{array} \right.$$

The first main result of this work concerns the second order derivative. It is given in the following theorem.

**THEOREM 2.** *Let  $\Omega$  be an open smooth subset of  $\mathbb{R}^N$  and let  $\omega$  be an element of  $\Omega_{d_0}$  with a  $\mathcal{C}^{4,\alpha}$  boundary. Let  $\mathbf{h}_1$  and  $\mathbf{h}_2$  be two deformation fields in  $\mathcal{H}$ . Then the*

state  $u_d$  has a second order shape derivative  $u_d'' \in H^1(\Omega \setminus \bar{\omega}) \cup H^1(\omega)$  that solves

$$(8) \quad \left\{ \begin{array}{l} \Delta u_d'' = 0 \text{ in } \Omega \setminus \bar{\omega} \text{ and in } \omega, \\ [u_d''] = (h_{1,n}h_{2,n}H - \mathbf{h}_{1\tau} \cdot (D\mathbf{n} \mathbf{h}_{2\tau})) [\partial_{\mathbf{n}} u_d] - (h_{1,n}[\partial_{\mathbf{n}}(u_d)']_2 + h_{2,n}[\partial_{\mathbf{n}}(u_d)']_1) \\ \quad + (\mathbf{h}_{1\tau} \cdot \nabla h_{2,n} + \mathbf{h}_{2\tau} \cdot \nabla h_{1,n}) [\partial_{\mathbf{n}} u_d] \text{ on } \partial\omega, \\ [\sigma \partial_{\mathbf{n}} u_d''] = \operatorname{div}_{\tau} \left( h_{2,n} [\sigma \nabla_{\tau}(u_d)']_1 + h_{1,n} [\sigma \nabla_{\tau}(u_d)']_2 + \mathbf{h}_{1\tau} \cdot (D\mathbf{n} \mathbf{h}_{2\tau}) [\sigma \nabla_{\tau} u_d] \right) \\ \quad - \operatorname{div}_{\tau} ((\mathbf{h}_{1\tau} \cdot \nabla_{\tau} h_{2,n} + \nabla_{\tau} h_{1,n} \cdot \mathbf{h}_{2\tau}) [\sigma \nabla_{\tau} u_d]) \\ \quad - \operatorname{div}_{\tau} (h_{2,n} h_{1,n} (2D\mathbf{n} - HI) [\sigma \nabla_{\tau} u_d]) \text{ on } \partial\omega, \\ u_d'' = 0 \text{ on } \partial\Omega. \end{array} \right.$$

Here,  $(u_d)'_i$  denotes the first order derivative of  $u$  in the direction of  $h_i$  as given in (6),  $D\mathbf{n}$  stands for the second fundamental form of the manifold  $\partial\omega$ , and  $H$  stands for the mean curvature of  $\partial\omega$ . Note that  $H$  is then the sum of the main curvatures and not the scaled version (divided by  $n - 1$ ) in dimension  $n$ .

The twin result concerning  $u_n$  is an easy adaption of Theorem 2. Once the differentiability of the state function has been established, one can consider the objectives. In [3], we have shown the first order result.

**THEOREM 3.** *Let  $\Omega$  be an open smooth subset of  $\mathbb{R}^N$  and let  $\omega$  be an element of  $\Omega_{d_0}$  with a  $C^{4,\alpha}$  boundary. Let  $\mathbf{h}_1$  and  $\mathbf{h}_2$  be two deformation fields in  $\mathcal{H}$ . The Kohn–Vogelius objective is differentiable with respect to the shape, and its derivative in the direction of a deformation field  $\mathbf{h}$  is given by*

$$(9) \quad DJ_{KV}(\omega)\mathbf{h} = [\sigma] \int_{\partial\omega} \left[ \frac{\sigma_1}{\sigma_2} (|\partial_{\mathbf{n}} u_d^+|^2 - |\partial_{\mathbf{n}} u_n^+|^2) + |\nabla_{\tau} u_d|^2 - |\nabla_{\tau} u_n|^2 \right] h_n.$$

We now give the second order derivative of the Kohn–Vogelius criterion.

**THEOREM 4.** *Let  $\Omega$  be an open smooth subset of  $\mathbb{R}^N$  and  $\omega$  an element of  $\Omega_{d_0}$  with a  $C^{4,\alpha}$  boundary. Let  $\mathbf{h}_1$  and  $\mathbf{h}_2$  be two deformation fields in  $\mathcal{H}$ . The Kohn–Vogelius objective is twice differentiable with respect to the shape, and its second derivative in the directions  $\mathbf{h}_1$  and  $\mathbf{h}_2$  is given by*

$$(10) \quad \begin{aligned} D^2 J_{KV}(\omega)(\mathbf{h}_1, \mathbf{h}_2) &= \int_{\partial\omega} [\sigma |\nabla v|^2] (\mathbf{h}_{1\tau} \cdot \nabla_{\tau}(h_{2,n}) + \mathbf{h}_{2\tau} \cdot \nabla_{\tau}(h_{1,n}) - \mathbf{h}_{2\tau} \cdot (D\mathbf{n} \mathbf{h}_{1\tau})) \\ &\quad - \int_{\partial\omega} \partial_{\mathbf{n}} \left( [\sigma |\nabla v|^2] \right) h_{1,n} h_{2,n} + 2 [\sigma \nabla v \cdot (h_{1,n} \nabla v'_2 + h_{2,n} \nabla v'_1)] \\ &\quad + \int_{\partial\omega} \left[ \sigma \left( \partial_{\mathbf{n}}(u_n)'_1 v'_2 + \partial_{\mathbf{n}}(u_n)'_2 v'_1 - \partial_{\mathbf{n}} v'_1 (u_d)'_2 - \partial_{\mathbf{n}} v'_2 (u_d)'_1 \right) \right] \\ &\quad + 2 \int_{\partial\omega} v \left[ \sigma \partial_{\mathbf{n}}(u_n)''_{1,2} \right] - \sigma_1 \partial_{\mathbf{n}} v^+ [(u_d)''_{1,2}], \end{aligned}$$

where we have set  $v = u_d - u_n$ .

To investigate the properties of stability of this cost function, we are led to consider an admissible inclusion  $\omega^*$  to solve both (1), (2) and (1), (4) in order to obtain the corresponding measurements  $f^*$  and  $g^*$ . It is obvious that the domain  $\omega^*$  realizes the absolute minimum of the criterion  $J_{KV}$  since, by construction, we can write

$u_d = u_n$  in  $\Omega$  and hence  $J_{KV}(\omega^*) = 0$ . We will check that the Euler equation

$$DJ_{KV}(\omega^*)(\mathbf{h}) = 0$$

holds. We will also prove that

$$(11) \quad D^2 J_{KV}(\omega^*)(\mathbf{h}, \mathbf{h}) = \int_{\Omega} \sigma |\nabla v'|^2.$$

Moreover, if  $h_n \neq 0$ , then  $D^2 J_{KV}(\omega^*)(\mathbf{h}, \mathbf{h}) > 0$  holds. Nevertheless, (11) does not mean that the minimization problem is well-posed. In fact, the second important result is the following theorem that explains the instability of standard minimization algorithms.

**THEOREM 5.** *Assume that  $\omega^*$  is a critical shape of  $J_{KV}$  for which the additional condition  $u_n = u_d$  holds. Then the Riesz operator corresponding to  $D^2 J_{KV}(\omega^*)$  defined from  $H^{1/2}(\partial\omega^*)$  with values in  $H^{-1/2}(\partial\omega^*)$  is compact. Moreover, the minimization problem is severely ill-posed in the following sense: if the target domain is  $\mathcal{C}^\infty$  and if  $\lambda_n$  denotes the  $n$ th eigenvalue of  $D^2 J_{KV}(\omega^*)$ , then  $\lambda_n = o(n^{-s})$  for all  $s > 0$ .*

Theorem 5 has two main consequences. First, the shape Hessian at the global minimizer is not coercive. This means that this minimizer may not be a local strict minimum of the criterion. Moreover, the criterion provides no control of the distance between the parameter  $\omega$  and the target  $\omega^*$ . The second consequence concerns any numerical scheme used to obtain this optimal domain  $\omega^*$ . One has to face this difficulty, and this explains why frozen Newton or Levenberg–Marquard schemes have been used to solve numerically this problem [8, 3].

The paper is organized as follows. In a first section, we state some preliminary results on shape derivatives. Some are well-known facts in shape optimization and will be recalled without proof for the sake of readability. Some of them (e.g., the derivatives of a Laplace–Beltrami operator and the tangential regularity of the solution to (1), (2) along the discontinuity of the conductivity distribution) are less known and will be proved thanks to potential layer methods. Hence we will tackle the computations in section 4 that we consider as the core of this work: it is essentially devoted to proving Theorem 2. After a first part where we prove the existence of a second order derivative for the state, we propose two distinct methods to find the boundary value problem solved by this second order derivative. The method presented in the core of the paper consists in a direct differentiation of the boundary conditions. Section 5 is devoted to the analysis of the criterion, and we establish Theorems 4 and 5. We explain their consequences on the stability of critical shapes.

In the final section, we present some appendices. Section 6.1 is devoted to proving tangential regularity along the interface for solution of (1), (2). In section 6.2, we consider the bidimensional case that requires an additional size assumption on the inclusion. Section 6.3 deals gives a second proof of Theorem 2. It follows along the lines of classical proofs of shape differentiability by differentiating the weak formulation of problem (1), (2) and interpreting the result in terms of differential operator and boundary conditions.

**3. Elements of shape calculus.** Before entering the proof of Theorem 2, we recall without proof some basic facts from shape optimization (see [7] for references). Let  $\mathbf{h}$  be a deformation field in  $\mathcal{C}^2(\Omega, \mathbb{R}^d)$  with  $\|\mathbf{h}\|_{\mathcal{C}^2} < 1$ . We set  $T_t(h, \cdot) = \text{Id} + t\mathbf{h}$  and denote by  $\Omega_t$  the transported domain  $\Omega_t = T_t(\Omega)$ . To avoid heavy notations, we will use the notation  $T_t$  instead of  $T_t(h, \cdot)$ .

**Material and shape derivatives.** Classically, in the mechanics of continuous media, the material derivative is defined as being a positive limit. In our context, for any vector field  $\mathbf{h} \in \mathcal{H}$ , we define the material derivative of the domain functional  $y = y(\Omega)$  at  $\Omega$  in an admissible direction  $\mathbf{h}$  as the limit:

$$(12) \quad \dot{y}(\Omega; \mathbf{h}) = \lim_{t \rightarrow 0} \frac{y(\Omega_t) \circ T_t - y(\Omega)}{t}.$$

Similarly, one can define the material derivative  $\dot{y}(\partial\Omega, \mathbf{h})$  for any domain functional  $y = y(\partial\Omega)$  which depends on  $\partial\Omega$ . Another kind of derivative occurs, called the shape derivative of  $y(\Omega, \mathbf{h})$ . It is viewed as a first local variation. Its definition is given by the following definition.

DEFINITION 1. *The shape derivative  $y' = y'(\Omega; \mathbf{h})$  of a functional  $y(\Omega)$  at  $\Omega$  in the direction of a vector field  $\mathbf{h}$  is given by*

$$(13) \quad y' = \dot{y} - \mathbf{h} \cdot \nabla y.$$

For more details on these derivations, the reader can consult [22, 7].

**Elements of tangential derivatives.** In what follows, we will need to manipulate the tangential differential operators on a manifold. For the reader's convenience, we recall from [5, 7] some definitions and also some useful rules of calculus.

DEFINITION 2. *The tangential divergence of a vector field  $\mathbf{V} \in C^1(\mathbb{R}^d, \mathbb{R}^d)$  is given by*

$$(14) \quad \operatorname{div}_\tau(\mathbf{V}) = \operatorname{div}(\mathbf{V}) - D\mathbf{V} \cdot \mathbf{n}, \mathbf{n},$$

where the notation  $D\mathbf{V}$  denotes the Jacobian matrix of  $\mathbf{V}$ . When the vector  $\mathbf{V} \in C^1(\partial\Omega, \mathbb{R}^d)$  is defined on  $\partial\Omega$ , then the following notation is used to define the tangential divergence:

$$(15) \quad \operatorname{div}_\tau(\mathbf{V}) = \operatorname{div}(\tilde{\mathbf{V}}) - (D\tilde{\mathbf{V}} \cdot \mathbf{n}) \cdot \mathbf{n},$$

where  $\tilde{\mathbf{V}}$  stands for an arbitrary  $C^1$  extension of  $\mathbf{V}$  on an open neighborhood of  $\partial\Omega$ .

We introduce now the notion of tangential gradient  $\nabla_\tau$  of any smooth scalar function  $f$  in  $C^1(\partial\Omega, \mathbb{R}^d)$ .

DEFINITION 3. *Let an element  $f \in C^1(\partial\Omega, \mathbb{R}^d)$  be given and let  $\tilde{f}$  be an extension of  $f$  in the sense that  $\tilde{f} \in C^1(U)$  and  $\tilde{f}|_{\partial\Omega} = f$  and where  $U$  is an open neighborhood of  $\partial\Omega$ . Then the following notation is used to define the tangential gradient:*

$$(16) \quad \nabla_\tau f = \nabla \tilde{f}|_{\partial\Omega} - \nabla \tilde{f} \cdot \mathbf{n} \, \mathbf{n} \text{ on } \partial\Omega.$$

The details for the existence of such an extension can be found in [5]. Let us remark that these definitions do not depend on the choice of the extension. Furthermore, one can show the important relation

$$(17) \quad \int_{\partial\Omega} \nabla_\tau f \cdot \mathbf{F} = - \int_{\partial\Omega} f \operatorname{div}_\tau(\mathbf{F})$$

for all elements  $f \in C^1(\partial\Omega)$  and all vector fields  $F \in C^1(\partial\Omega, \mathbb{R}^d)$  satisfying  $F_n = \langle F, \mathbf{n} \rangle = 0$ .

**Integration by parts on  $\partial\Omega$ .** In general, the condition above,  $F_n = 0$ , is not always satisfied. We are then led to find another formula to extend the formula to the general case. The extension of this integration by parts formula to fields with a normal vector component involves curvature.

First, we point out that the curvature is connected to the normal vector via the tangential divergence operator. Recall that the mean curvature of  $\partial\Omega$  is defined as  $H = \operatorname{div}_\tau(\mathbf{n})$ . Making use of the form of  $\operatorname{div}_\tau(\mathbf{n})$  on the boundary, one shows straightforwardly the following statement.

**PROPOSITION 1.** *Let  $\Omega$  be an open subset of  $\mathbb{R}^N$  with a  $\mathcal{C}^2$  boundary. For any unitary extension  $\mathcal{N}$  of  $\mathbf{n}$  on a neighborhood of  $\partial\Omega$ , one has*

$$\operatorname{div}(\mathcal{N}) = H \text{ on } \partial\Omega.$$

Assume that the manifold  $\partial\Omega$  has no borders. If  $\mathbf{F} \in H^2(\partial\Omega)^3$  and  $f \in H^2(\partial\Omega)$ , then we have

$$(18) \quad \int_{\partial\Omega} \nabla_\tau f \cdot \mathbf{F} + f \operatorname{div}_\tau(\mathbf{F}) = \int_{\partial\Omega} (\nabla f \cdot \mathbf{n} + Hf) \mathbf{F} \cdot \mathbf{n}.$$

We assume now that the domain  $\Omega$  has a  $\mathcal{C}^3$  boundary. The simplest second order derivative is the Laplace–Beltrami operator; it is defined as follows (see [5, 7]), thanks to the following usual chain rule.

**DEFINITION 4.** *Let  $f \in H^2(\partial\Omega)$ . The Laplace–Beltrami  $\Delta_\tau$  of  $f$  is defined as*

$$(19) \quad \Delta_\tau f = \operatorname{div}_\tau(\nabla_\tau f).$$

There is a relation connecting the Laplace operator and the Laplace–Beltrami operator. Let us denote  $\partial_{nn}^2 f = (D^2 f \cdot \mathbf{n}) \cdot \mathbf{n}$ , where  $D^2 f$  stands for the Hessian of  $f$ .

**PROPOSITION 2.** *Let  $\Omega$  be a domain with a boundary  $\partial\Omega$  of class  $\mathcal{C}^3$ . For all functions  $f \in H^3(\Omega)$ , it holds that*

$$(20) \quad \Delta f = \Delta_\tau f + H \partial_{nn} f + \partial_{nn}^2 f \text{ on } \partial\Omega.$$

**Some useful derivatives.** We need to compute shape and material derivatives of special vector fields: the outer unit normal vector  $\mathbf{n}$ , the tangential gradient, and the Laplace–Beltrami operator applied to a function. While the derivative of the normal vector is obtained by a straightforward calculus, we have to transport from  $\partial\Omega_t$  to  $\partial\Omega$  the Laplace–Beltrami operator and the tangential gradient in order to compute the other derivatives.

*Derivatives of the normal vector.* We describe the material and shape derivatives of the normal vector. We will denote by  $\mathbf{n}$  the gradient of the signed distance to  $\partial\Omega$ . This is a unitary extension of the unitary normal vector  $\mathbf{n}$  at  $\partial\Omega$  which is smooth in the vicinity of  $\partial\Omega$ . This extension furnishes a symmetric Jacobian  $D\mathbf{n}$  that satisfies  $D\mathbf{n}\mathbf{n} = 0$  on  $\partial\Omega$ . The direction  $\mathbf{h}$  will be supposed to be in  $\mathcal{C}^2(\mathbb{R}^d, \mathbb{R}^d)$  or in  $\mathcal{C}^2(\partial\Omega, \mathbb{R}^d)$ .

**PROPOSITION 3.** *The material derivative  $\dot{\mathbf{n}}$  of the normal vector  $\mathbf{n}$  at  $\Omega$  in the direction of a vector field  $\mathbf{h} \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$  is given by*

$$\dot{\mathbf{n}} = -\nabla_\tau(\mathbf{h} \cdot \mathbf{n}) + D\mathbf{n}\mathbf{h}_\tau,$$

where  $\mathbf{h}_\tau = \mathbf{h} - \mathbf{h} \cdot \mathbf{n} \mathbf{n}$ .

Concerning its shape derivative defined as  $\mathbf{n}' = (\partial_t \mathbf{n}_t)|_{t=0}$ , where  $\mathbf{n}_t$  is any smooth unitary extension of  $\mathbf{n}$  to  $\partial\Omega_t$ , we obtain the following.

**PROPOSITION 4.** *The shape derivative  $\mathbf{n}'$  in the direction of  $\mathbf{h}$  is given by*

$$\mathbf{n}' = -\nabla_\tau(\mathbf{h} \cdot \mathbf{n}).$$

*Derivative of the tangential gradient.* For  $f \in H^3(\partial\Omega)$ , we compute the material derivative of  $\nabla_\tau f$ .

PROPOSITION 5. *For all functions  $f \in C^2(\mathbb{R}^3)$  and directions  $\mathbf{h} \in C^2(\partial\Omega, \mathbb{R}^3)$ , one has*

$$\overline{\dot{\nabla}_\tau f} = \nabla_\tau f' + (D^2 f \mathbf{h})_\tau - \nabla f \cdot \mathbf{n} \dot{\mathbf{n}} - \nabla f \cdot \dot{\mathbf{n}} \mathbf{n}.$$

*Proof.* We differentiate  $\nabla f$  and  $\nabla f \cdot \mathbf{n} \mathbf{n}$  and obtain

$$\overline{\dot{\nabla} f} = \nabla f' + D^2 f \mathbf{h},$$

while

$$\overline{\dot{\nabla f \cdot \mathbf{n} \mathbf{n}}} = \nabla f \cdot \dot{\mathbf{n}} \mathbf{n} + \nabla f \cdot \mathbf{n} \dot{\mathbf{n}} + \nabla f' \cdot \mathbf{n} \mathbf{n} + (D^2 f \mathbf{h}) \cdot \mathbf{n} \mathbf{n}.$$

The two former equations give the desired result.  $\square$

*Derivative of the Laplace–Beltrami operator.* Now, we want to compute the material derivative  $\overline{\dot{\Delta}_\tau f}$ . We begin to study how to transport the Laplace–Beltrami operator when one works on  $\partial\Omega_t$ . Let  $\Delta_{\tau,t}$  denote the Laplace–Beltrami operator on the manifold  $\partial\Omega_t$ . To compute the derivative of a Laplace–Beltrami operator, we need the following proposition, which we quote from [22].

PROPOSITION 6. *Let  $f \in H^{5/2}(\mathbb{R}^d)$ . Then*

$$(21) \quad \int_{\partial\Omega} \left[ (\Delta_{\tau,t} f) \circ T_t \gamma_\tau(t) \right] \phi \\ = - \int_{\partial\Omega} \left[ C(t) \left( \nabla(f \circ T_t) - (B(t) \mathbf{n}) \cdot \nabla(f \circ T_t) \right) \right] \cdot \nabla \phi \quad \forall \phi \in \mathcal{D}(\mathbb{R}^d).$$

In the former proposition, we set

$$(22) \quad \begin{aligned} \gamma(t) &= \det DT_t, \\ \gamma_\tau(t) &= \gamma(t) \|(DT_t^{-1})^T \cdot \mathbf{n}\|_{\mathbb{R}^d}, \\ B(t) &= \frac{D(T_t^{-1})(D(T_t)^{-1})^T}{\|(DT_t^{-1})^T \cdot \mathbf{n}\|_{\mathbb{R}^d}^2}, \\ C(t) &= \gamma_\tau(t) D(T_t^{-1})(D(T_t)^{-1})^T. \end{aligned}$$

A straightforward computation gives

$$(23) \quad \begin{aligned} \gamma'(0) &= \operatorname{div}_\tau(\mathbf{h}), \\ \gamma'_\tau(0) &= \operatorname{div}_\tau(\mathbf{h}) = \operatorname{div}_\tau(\mathbf{h}_\tau) + H\mathbf{h}_n, \\ B'(0) &= 2(D\mathbf{h}\mathbf{n}) \cdot \mathbf{n}I - (D\mathbf{h} + (D\mathbf{h})^T), \\ C'(0) &= \operatorname{div}_\tau(\mathbf{h})I - (D\mathbf{h} + (D\mathbf{h})^T). \end{aligned}$$

We then have the following result.

THEOREM 6. *Let  $f \in \mathcal{D}(\mathbb{R}^d)$ . The material derivative of  $\Delta_\tau f$  in the direction  $\mathbf{h}$  is given by*

$$(24) \quad \begin{aligned} \overline{\dot{\Delta}_\tau f} &= \Delta_\tau \dot{f} + \nabla_\tau f \cdot \nabla_\tau [\operatorname{div}_\tau(\mathbf{h}_\tau)] \\ &\quad + \nabla_\tau(H\mathbf{h}_n) \cdot \nabla_\tau f - \operatorname{div}_\tau \left( \left( (D\mathbf{h} + (D\mathbf{h})^T) \nabla_\tau f \right)_\tau \right). \end{aligned}$$



*Proof.* Formula (24) is shown in a weak sense. For each test function  $\phi \in \mathcal{C}^\infty(\partial\Omega)$ , there exists an extension  $\tilde{\phi} \in \mathcal{D}(\mathbb{R}^d)$  such that  $\partial_{\mathbf{n}}\tilde{\phi} = 0$ ; this can be done by extending  $\phi$  as a constant along the orbits of the gradient of the signed distance function to  $\partial\Omega$  and through the use of a cut-off function. For  $f \in \mathcal{D}(\mathbb{R}^d)$ , we set

$$A(t) = \int_{\partial\Omega} \frac{(\Delta_{\tau,t}f) \circ T_t - \Delta_{\tau}f}{t} \gamma_{\tau}(t) \phi.$$

After an integration by parts on  $\partial\Omega$ , we obtain

$$\begin{aligned} A(t) &= \int_{\partial\Omega} \frac{1 - \gamma_{\tau}(t)}{t} (\Delta_{\tau,t}f) \circ T_t \phi + \int_{\partial\Omega} \frac{\gamma_{\tau}}{t} ((\Delta_{\tau,t}f) \circ T_t \phi) + \frac{1}{t} \nabla_{\tau}f \cdot \nabla_{\tau}\phi, \\ &= \int_{\partial\Omega} \frac{1 - \gamma_{\tau}(t)}{t} (\Delta_{\tau,t}f) \circ T_t \phi \\ &\quad + \int_{\partial\Omega} \frac{1}{t} \left( [\nabla_{\tau}f - C(t)\nabla(f \circ T_t)] \cdot \nabla\tilde{\phi} + [(B(t)\mathbf{n} \cdot \nabla(f \circ T_t))] C(t) \mathbf{n} \cdot \nabla\tilde{\phi} \right). \end{aligned}$$

Since  $\partial_{\mathbf{n}}\tilde{\phi} = 0$  and  $C(0) = I$ , we get

$$\begin{aligned} A(t) &= \int_{\partial\Omega} \frac{1 - \gamma_{\tau}(t)}{t} (\Delta_{\tau,t}f) \circ T_t \phi + \int_{\partial\Omega} \frac{\nabla_{\tau}(f - f \circ T_t)}{t} \cdot \nabla_{\tau}\tilde{\phi} \\ &\quad + \int_{\partial\Omega} \frac{C(0) - C(t)}{t} \nabla(f \circ T_t) \cdot \nabla_{\tau}\tilde{\phi}. \end{aligned}$$

When  $t \rightarrow 0$ , then

$$\begin{aligned} \int_{\partial\Omega} \overline{\Delta_{\tau}f} \phi &= - \int_{\partial\Omega} \gamma'_{\tau}(t) \Delta_{\tau}f \phi + \nabla_{\tau}f \cdot \nabla_{\tau}\phi + (C'(0) \cdot \nabla f) \cdot \nabla_{\tau}\phi, \\ &= \int_{\partial\Omega} \left( \Delta_{\tau}f - \operatorname{div}_{\tau}(\mathbf{h}) \Delta_{\tau}f \right) \phi + \left( D\mathbf{h} + (D\mathbf{h})^T - \operatorname{div}_{\tau}(\mathbf{h}) I \right) \nabla f \cdot \nabla_{\tau}\phi, \\ &= \int_{\partial\Omega} \left[ \Delta_{\tau}f - \operatorname{div}_{\tau}(\mathbf{h}) \Delta_{\tau}f + \operatorname{div}_{\tau}(\operatorname{div}_{\tau}(h) \nabla_{\tau}f) \right. \\ &\quad \left. - \operatorname{div}_{\tau} \left( \left( (D\mathbf{h} + (D\mathbf{h})^T) \nabla f \right)_{\tau} \right) \right] \phi. \end{aligned}$$

Expanding the double divergence term, we obtain

$$\overline{\Delta_{\tau}f} = \Delta_{\tau}f + \nabla_{\tau}f \cdot \nabla_{\tau} \operatorname{div}_{\tau}(\mathbf{h}) - \operatorname{div}_{\tau} \left( \left( (D\mathbf{h} + (D\mathbf{h})^T) \nabla f \right)_{\tau} \right).$$

In order to make these derivatives explicit, we let the curvatures of  $\partial\Omega$  appear by means of

$$\nabla_{\tau}f \cdot \nabla_{\tau} \operatorname{div}_{\tau}(\mathbf{h}) = \nabla_{\tau}f \cdot \nabla_{\tau} [\operatorname{div}_{\tau}(\mathbf{h}_{\tau}) + H\mathbf{h}_n],$$

and this ends the proof of Theorem 6.  $\square$

**Structure of second order shape derivatives.** Since the fundamental work of J. Hadamard, domain derivatives of order one are well known to depend only on the normal component of the deformation field  $\mathbf{h}$ : It is written as  $l_1(\mathbf{h} \cdot \mathbf{n})$ . The reason for this is that, if the field  $\mathbf{h}$  is tangent to  $\Omega$ , then its flow preserves for small time at order one the domain  $\Phi_t(\Omega) = \Omega + o(t)$ , with abuse of notation.

This property is lost for second order derivatives as illustrated by Theorem 5.9.2 of [7]. As shown in [19], the second order derivative of a function  $E$  can be written as

$$D^2 E(\mathbf{h}_1, \mathbf{h}_2) = l_2(\mathbf{h}_1 \cdot \mathbf{n}, \mathbf{h}_2 \cdot \mathbf{n}) + l_1(Z),$$

where  $Z$  denotes

$$Z = (D_\tau \mathbf{n} \mathbf{h}_{1\tau}) \cdot \mathbf{h}_{2\tau} - \nabla_\tau(\mathbf{h}_1 \cdot \mathbf{n}) \cdot \mathbf{h}_{2\tau} - \nabla_\tau(\mathbf{h}_2 \cdot \mathbf{n}) \cdot \mathbf{h}_{1\tau}.$$

The important point is that a second order shape derivative also depends on the tangential component of the deformation fields and that this dependence disappears at critical shapes.

**4. Existence of the second order derivative of the state. Proof of Theorem 2.** This section is devoted to the proof of Theorem 2. We follow the usual strategy to derive existence in shape optimization. In the first step, we will write the weak formulation of the problem, then transport it on the reference domain, pass to the limit, and obtain existence of the material derivative. In the second step, we will seek a boundary value problem solved by the material derivative. This will provide a characterization of the second order shape derivative. Two strategies, which we will detail, are possible: The first one, explored in section 6.1, consists in working on the variational formulation, while the second one uses the tangential differential calculus by differentiating the boundary conditions. The latter approach will be presented in section 4. Both approaches require that the trace on the interface of discontinuity  $\partial\omega$  of the solution to the boundary value problem have tangential derivatives. For the sake of readability, we postponed all the needed justifications until section 6.2.

*Preliminary results.* In what follows, we will use some technical formulae. To preserve the readability of the proof of the main result, we state them in this paragraph. The tools needed for proving these results can be found in [22]. Given a smooth vector field  $\mathbf{h}$ , we denote

$$A_{\mathbf{h}} = D\mathbf{h} + D\mathbf{h}^T - \operatorname{div}(\mathbf{h}) I.$$

We begin with the following formula.

LEMMA 1. *It holds that*

$$(25) \quad \nabla u \cdot A_{\mathbf{h}} \nabla v = \nabla(\mathbf{h} \cdot \nabla u) \cdot \nabla v + \nabla(\mathbf{h} \cdot \nabla v) \cdot \nabla u - \operatorname{div}((\nabla u \cdot \nabla v) \mathbf{h}).$$

Given two smooth vector fields  $\mathbf{h}_1$  and  $\mathbf{h}_2$ , we set

$$(26) \quad \mathfrak{A} = D\mathbf{h}_2 A_{\mathbf{h}_1} + A_{\mathbf{h}_1} D\mathbf{h}_2^T - A_{\mathbf{h}_1} \operatorname{div}(\mathbf{h}_2) - (A_{\mathbf{h}_1})'(\mathbf{h}_2)$$

and

$$b = (\mathbf{h}_2 \cdot \nabla u) A_{\mathbf{h}_1} \nabla v + (\mathbf{h}_2 \cdot \nabla v) A_{\mathbf{h}_1} \nabla u - ((A_{\mathbf{h}_1} \nabla u) \cdot \nabla v) \mathbf{h}_2.$$

Here, the notation  $(A_{\mathbf{h}_1})'(\mathbf{h}_2)$  stands for the matrix defined by its elements:

$$((A_{\mathbf{h}_1})'(\mathbf{h}_2))_{k,l} = \nabla(((A_{\mathbf{h}_1})')_{k,l}) \cdot \mathbf{h}_2.$$

LEMMA 2. *One has*

$$(27) \quad \nabla u \cdot \mathfrak{A} \nabla v = \operatorname{div}(b) - (\mathbf{h}_2 \cdot \nabla u) \operatorname{div}((A_{\mathbf{h}_1} \nabla v)) - (\mathbf{h}_2 \cdot \nabla v) \operatorname{div}((A_{\mathbf{h}_1} \nabla u)).$$

We need the following crucial result.

LEMMA 3. *If  $u$  is harmonic, then*

$$(28) \quad \operatorname{div}(A_{\mathbf{h}_1} \nabla u) = \Delta(\mathbf{h}_1 \cdot \nabla u).$$

*Proof.* For any harmonic function  $u$  in  $\Omega$  and for every test function  $\phi \in \mathcal{D}(\Omega)$ , we can write

$$\int_{\Omega} \nabla \dot{u} \nabla \phi = \int_{\Omega} A_{\mathbf{h}} \nabla u \cdot \nabla \phi.$$

Then

$$\int_{\Omega} \Delta \dot{u} \cdot \phi = \int_{\Omega} \operatorname{div}(A_{\mathbf{h}} \nabla u) \cdot \phi.$$

Since  $\dot{u} = u' + \mathbf{h} \cdot \nabla u$  and since  $u'$  is harmonic in  $\Omega$ , we obtain the result.  $\square$

We follow Hettlich and Rundell [9] and Simon [21] to define the second order derivative of an operator with respect to a domain. We compute the second derivative by considering two admissible deformations fields,  $\mathbf{h}_1, \mathbf{h}_2 \in \mathcal{H}$ , that will describe the small variations of  $\partial\omega$ . Simon showed that the second derivative  $F''(\partial\omega; \mathbf{h}_1, \mathbf{h}_2)$  of  $F(\partial\omega)$  is defined as a bounded bilinear operator satisfying

$$(29) \quad F''(\partial\omega; \mathbf{h}_1, \mathbf{h}_2) = \left( F'(\partial\omega; \mathbf{h}_1) \right)' \mathbf{h}_2 - F'(\partial\omega; D\mathbf{h}_1 \mathbf{h}_2).$$

For more details, the reader can consult the appendix on page 613 of [9].

*Step 1: Existence of the second order derivative.* Let us begin the proof. Let  $\mathbf{h}_1, \mathbf{h}_2 \in \mathcal{H}$  be two vector fields. The direction  $\mathbf{h}_1$  being fixed, we consider  $\dot{u}_{1, \mathbf{h}_2}$  the variation of  $u_1$  with respect to the direction  $\mathbf{h}_2$ . We recall from [3] that the material derivative  $\dot{u}_1$  of  $u$  in the direction  $\mathbf{h}_1$  satisfies

$$\forall v \in H_0^1(\Omega), \quad \int_{\Omega} \sigma \nabla \dot{u}_1 \cdot \nabla v = \int_{\Omega} \sigma \nabla u \cdot A_{\mathbf{h}_1} \nabla v.$$

Let  $\phi_2 : \Omega \mapsto \Omega$  be the diffeomorphism defined by  $\phi_2(x) = x + \mathbf{h}_2(x)$ , and set  $\psi_2 = \phi_2^{-1}$ . Setting  $\omega_{\mathbf{h}_2} = \{x + \mathbf{h}_2(x), x \in \omega\}$ ,  $\Omega_{\mathbf{h}_2} = \{x + \mathbf{h}_2(x), x \in \Omega\} = \Omega$ , and  $\sigma_{\mathbf{h}_2} = \sigma \circ \phi_2$ , we get

$$(30) \quad \int_{\Omega_{\mathbf{h}_2}} \sigma_{\mathbf{h}_2} \nabla \dot{u}_{1, \mathbf{h}_2} \cdot \nabla v = \int_{\Omega_{\mathbf{h}_2}} \sigma_{\mathbf{h}_2} \nabla u_{\mathbf{h}_2} \cdot A_{\mathbf{h}_1} \nabla v,$$

where  $u_{\mathbf{h}_2}$  is the solution of the original problem with  $\omega_{\mathbf{h}_2}$  instead of  $\omega$ . Making the change of variables  $x = \phi_2(X)$ , we get the integral identity on the fixed domain  $\Omega$ :

$$(31) \quad \begin{aligned} \int_{\Omega} \sigma \nabla \tilde{u}_{1, \mathbf{h}_2} \cdot \left( D\psi_2 (D\psi_2)^T \det(D\phi_2) \right) \nabla v \\ = \int_{\Omega} \sigma \nabla \tilde{u}_{\mathbf{h}_2} \cdot \left( D\psi_2 \widetilde{A_{\mathbf{h}_1}} (D\psi_2)^T \det(D\phi_2) \right) \nabla v, \end{aligned}$$

with the notation  $\tilde{u} = u \circ \phi_2$  and  $\widetilde{A_{\mathbf{h}_1}} = A_{\mathbf{h}_1} \circ \phi_2$ . Since the material derivative  $\dot{u}_1$  of  $u$  with respect to the direction  $\mathbf{h}_1$  satisfies

$$\int_{\Omega} \sigma \nabla \dot{u}_1 \cdot \nabla v = \int_{\Omega} \sigma \nabla u \cdot A_{\mathbf{h}_1} \nabla v,$$

the difference of (30) and (31) gives

$$\begin{aligned} \int_{\Omega} \sigma \nabla (\tilde{u}_{1, \mathbf{h}_2} - \dot{u}_1) \cdot \nabla v &= \int_{\Omega} \sigma \nabla \tilde{u}_{1, \mathbf{h}_2} \cdot \left( I - D\psi_2 (D\psi_2)^T \det(D\phi_2) \right) \nabla v \\ &\quad + \int_{\Omega} \sigma \nabla \tilde{u}_{\mathbf{h}_2} \cdot \left( D\psi_2 \widetilde{A_{\mathbf{h}_1}} (D\psi_2)^T \det(D\phi_2) - A_{\mathbf{h}_1} \right) \nabla v \\ &\quad + \int_{\Omega} (\nabla \tilde{u}_{\mathbf{h}_2} - \nabla u) \cdot A_{\mathbf{h}_1} \nabla v. \end{aligned}$$

We quote from [13] and [9] the following asymptotic formulae:

$$\begin{aligned} \|\det(D\phi_i) - 1 - \operatorname{div}(\mathbf{h}_i)\|_{\infty} &= O(\|\mathbf{h}_i\|_{\mathcal{C}^2}^2), \\ \|D\psi_i (D\psi_i)^T \det(D\phi_i) - I + A_{\mathbf{h}_i}\|_{\infty} &= O(\|\mathbf{h}_i\|_{\mathcal{C}^2}^2), \\ \|D\psi_2 \widetilde{A_{\mathbf{h}_1}} (D\psi_2)^T \det(D\phi_2) - A_{\mathbf{h}_1} + D\mathbf{h}_2 A_{\mathbf{h}_1} \\ &\quad + A_{\mathbf{h}_1} (D\mathbf{h}_2)^T - \operatorname{div}(\mathbf{h}_2) A_{\mathbf{h}_1} - (A_{\mathbf{h}_1})'(\mathbf{h}_2)\|_{\infty} = O(\|\mathbf{h}_2\|_{\mathcal{C}^2}^2). \end{aligned}$$

Making the adequate substitutions, we easily check that the material derivative of  $\dot{u}_1$  with respect to  $\mathbf{h}_2$  exists. This derivative, denoted by  $\ddot{u}_1$ , satisfies

$$(32) \quad \int_{\Omega} \sigma \nabla \ddot{u}_1 \cdot \nabla v \, dx = \int_{\Omega} \sigma [\nabla \dot{u}_1 \cdot A_{\mathbf{h}_2} \nabla v + \nabla \dot{u}_2 \cdot A_{\mathbf{h}_1} \nabla v - \nabla u \cdot \mathfrak{A} \nabla v],$$

where  $\mathfrak{A}$  is defined in (26).

*Step 2: Derivation of (8) by differentiation of the boundary conditions on  $\partial\omega$ .* The aim of this section is to compute the second order shape derivative of the state by differentiating the boundary conditions. We start with the expression of the flux jump  $[\sigma \partial_{\mathbf{n}} u'']$  by computing the normal derivatives of each of the expressions  $\overline{[\sigma \nabla u'] \cdot \mathbf{n}}$  and  $\overline{\operatorname{div}_{\tau}(h_{1,n}[\sigma \nabla_{\tau} u])}$ . Since

$$[\sigma \nabla u'] \cdot \mathbf{n} = \operatorname{div}_{\tau}(h_{1,n}[\sigma \nabla_{\tau} u]) = h_{1,n}[\sigma \Delta_{\tau} u] + \nabla_{\tau} h_{1,n} \cdot [\sigma \nabla_{\tau} u],$$

then we get

$$\begin{aligned} (33) \quad \overline{[\sigma \nabla u'] \cdot \mathbf{n}} &= \overline{\operatorname{div}_{\tau}(h_{1,n}[\sigma \nabla_{\tau} u])} \\ &= \overline{h_{1,n}[\sigma \Delta_{\tau} u]} + h_{1,n} \overline{[\sigma \Delta_{\tau} u]} + \overline{\nabla_{\tau} h_{1,n}} \cdot [\sigma \nabla_{\tau} u] + \nabla_{\tau} h_{1,n} \cdot \overline{[\sigma \nabla_{\tau} u]}. \end{aligned}$$

In order to avoid lengthy computations, we shall concentrate on each normal derivative appearing in the above formula. Some of the results are straightforward and their proofs will be left to the reader. Combining Propositions 3 and 5, we conclude that

$$\overline{\nabla_{\tau} h_{1,n}} = -\nabla_{\tau}(\mathbf{h}_1 \cdot \nabla_{\tau} h_{2,n}) + (D^2 h_{1,n} \cdot \mathbf{h}_2)_{\tau} - \nabla h_{1,n} \cdot \dot{\mathbf{n}} \, \mathbf{n} - \nabla h_{1,n} \cdot \mathbf{n} \, \dot{\mathbf{n}}.$$

In the same manner, we also get

$$\overline{[\sigma \nabla_\tau u]} = [\sigma \nabla_\tau u'_2] + ([\sigma D^2 u] \cdot \mathbf{h}_2)_\tau - [\sigma \nabla_\tau u] \cdot \dot{\mathbf{n}} \mathbf{n} - [\sigma \nabla_\tau u] \mathbf{n} \dot{\mathbf{n}}.$$

We use Proposition 3 and obtain

$$\begin{aligned} \overline{\mathbf{h}_1 \cdot \dot{\mathbf{n}}} &= \dot{\mathbf{h}}_1 \cdot \mathbf{n} + \mathbf{h}_1 \cdot \dot{\mathbf{n}} \\ &= \mathbf{h}_2 \cdot \nabla h_{1,n} - \nabla_\tau h_{2,n} \cdot \mathbf{h}_{1\tau}. \end{aligned}$$

Formula (33) becomes

$$\begin{aligned} (34) \quad \overline{\operatorname{div}_\tau (h_{1,n} [\sigma \nabla_\tau u])} &= \overline{h_{1,n} [\sigma \Delta_\tau u]} + h_{1,n} \overline{[\sigma \Delta_\tau u]} + \overline{\nabla_\tau h_{1,n}} \cdot [\sigma \nabla_\tau u] + \nabla_\tau h_{1,n} \cdot \overline{[\sigma \nabla_\tau u]} \\ &= \left( -\nabla_\tau (\mathbf{h}_1 \cdot \nabla_\tau h_{2,n}) + (D^2 h_{1,n} \cdot \mathbf{h}_2)_\tau - \nabla h_{1,n} \cdot \dot{\mathbf{n}} \mathbf{n} \right) [\sigma \nabla_\tau u] \\ &\quad + \nabla_\tau h_{1,n} \left( [\sigma \nabla_\tau u'_2] + ([\sigma D^2 u] \cdot \mathbf{h}_2)_\tau \right) \\ &\quad + (\mathbf{h}_2 \cdot \nabla h_{1,n} - \nabla_\tau h_{2,n} \cdot \mathbf{h}_{1\tau}) \cdot [\sigma \Delta_\tau u] + h_{1,n} \overline{[\sigma \Delta_\tau u]}. \end{aligned}$$

Let us pose  $A = [\sigma (D^2 u \cdot \mathbf{h}_2)_\tau] \cdot \nabla_\tau h_{1,n}$  and  $B = [\sigma \nabla_\tau u] \cdot (D^2 h_{1,n} \cdot \mathbf{h}_2)_\tau$ ; by formula (73), we simplify  $A$  and  $B$  as follows:

$$\begin{aligned} A &= -[\sigma \nabla_\tau u] \cdot (D \mathbf{n} \nabla_\tau h_{1,n}) h_{2,n} + [\sigma D^2 u] \mathbf{h}_{2\tau} \cdot \nabla_\tau h_{1,n}, \\ B &= (D^2 h_{1,n} \cdot \mathbf{h}_{2\tau}) \cdot [\sigma \nabla_\tau u] + \nabla_\tau (\partial_n h_{1,n}) \cdot [\sigma \nabla_\tau u] h_{2,n} - [\sigma \nabla_\tau u] \cdot (D \mathbf{n} \nabla_\tau h_{1,n}) h_{2,n}. \end{aligned}$$

After this computation and the formula of  $\dot{\mathbf{n}}$ , (34) becomes

$$\begin{aligned} (35) \quad \overline{\operatorname{div}_\tau (h_{1,n} [\sigma \nabla_\tau u])} &= (-\nabla_\tau (\mathbf{h}_{1\tau} \cdot \nabla_\tau h_{2,n})) \cdot [\sigma \nabla_\tau u] - \partial_n h_{1,n} (D \mathbf{n} \mathbf{h}_{2\tau} - \nabla_\tau h_{2,n}) [\sigma \nabla_\tau u] \\ &\quad + (D^2 h_{1,n} \cdot \mathbf{h}_{2\tau}) \cdot [\sigma \nabla_\tau u] + \nabla_\tau (\partial_n h_{1,n}) \cdot [\sigma \nabla_\tau u] h_{2,n} \\ &\quad - 2[\sigma \nabla_\tau u] \cdot (D \mathbf{n} \nabla_\tau h_{1,n}) h_{2,n} \\ &\quad + \nabla_\tau h_{1,n} [\sigma \nabla_\tau u'_2] + [\sigma D^2 u] \mathbf{h}_{2\tau} \cdot \nabla_\tau h_{1,n} \\ &\quad + (h_{2,n} \partial_n (h_{1,n}) - \nabla_\tau h_{2,n} \cdot \mathbf{h}_{1\tau}) \cdot [\sigma \Delta_\tau u] \\ &\quad + \mathbf{h}_{2\tau} \cdot \nabla_\tau h_{1,n} [\sigma \Delta_\tau u] + h_{1,n} \overline{[\sigma \Delta_\tau u]}. \end{aligned}$$

We tackle the computation of  $(\partial_n u)'$ . For the sake of clarity, we subdivide the work into several steps.

*First step.* We simplify the quantity  $\overline{\operatorname{div}_\tau (h_{1,n} [\sigma \nabla_\tau u])}$ . According to (35), and by using the formula of divergence and by adding and cutting the quantity  $[\sigma \Delta_\tau u'_2]$ , we obtain

$$\begin{aligned} (36) \quad \overline{\operatorname{div}_\tau (h_{1,n} [\sigma \nabla_\tau u])} &= \operatorname{div}_\tau \left( h_{1,n} [\sigma \nabla_\tau u'_2] + (h_{2,n} \partial_n h_{1,n} - \nabla_\tau h_{2,n} \cdot \mathbf{h}_{1\tau}) [\sigma \nabla_\tau u] \right) \\ &\quad + [\sigma \Delta_\tau u] \nabla_\tau h_{1,n} \cdot \mathbf{h}_{2\tau} \\ &\quad - \partial_n h_{1,n} [\sigma \nabla_\tau u] \cdot (D \mathbf{n} \mathbf{h}_{2\tau}) + [\sigma \nabla_\tau u] \cdot (D^2 h_{1,n} \cdot \mathbf{h}_{2\tau}) \\ &\quad - 2h_{2,n} [\sigma \nabla_\tau u] \cdot (D \mathbf{n} \nabla_\tau h_{1,n}) \\ &\quad + [\sigma D^2 u] \mathbf{h}_{2\tau} \cdot \nabla_\tau h_{1,n} + h_{1,n} \left( \overline{[\sigma \Delta_\tau u]} - [\sigma \Delta_\tau u'_2] \right). \end{aligned}$$

*Second step.* We compute  $\overline{[\sigma \partial_{\mathbf{n}} u'_1]}$ . From the expression of  $\dot{\mathbf{n}}$ , we get after some straightforward computations

$$(37) \quad \begin{aligned} \overline{[\sigma \partial_{\mathbf{n}} u'_1]} &= \overline{[\sigma \nabla u'_1]} \cdot \mathbf{n} + [\sigma \nabla u'_1] \cdot \dot{\mathbf{n}} \\ &= [\sigma \partial_{\mathbf{n}}(u'_1)_2] + ([\sigma D^2 u'_1] \mathbf{h}_2) \cdot \mathbf{n} + [\sigma \nabla_{\tau} u'_1] \cdot (D\mathbf{n} \mathbf{h}_{2\tau} - \nabla_{\tau} h_{2,n}). \end{aligned}$$

*Third step.* We compute  $\sigma \partial_{\mathbf{n}}(u'_1)'_2$ . From the jump condition on the flux of the derivative (6) and from (36) and (37), we obtain

$$\begin{aligned} [\sigma \partial(u'_1)'_2] &= \overline{\operatorname{div}_{\tau}(h_{1,n}[\sigma \nabla_{\tau} u])} - ([\sigma D^2 u'_1] \mathbf{h}_2) \cdot \mathbf{n} - [\sigma \nabla_{\tau} u'_1] \cdot (D\mathbf{n} \mathbf{h}_{2\tau} - \nabla_{\tau} h_{2,n}) \\ &= \operatorname{div}_{\tau} \left( h_{1,n}[\sigma \nabla_{\tau} u'_2] + (h_{2,n} \partial_{\mathbf{n}} h_{1,n} - \nabla_{\tau} h_{2,n} \cdot \mathbf{h}_{1\tau}) [\sigma \nabla_{\tau} u] \right) \\ &\quad - ([\sigma D^2 u'_1] \mathbf{h}_2) \cdot \mathbf{n} + [\sigma \nabla_{\tau} u'_1] \cdot (\nabla_{\tau} h_{2,n} - D\mathbf{n} \mathbf{h}_{2\tau}) - \partial_{\mathbf{n}} h_{1,n} [\sigma \nabla_{\tau} u] \cdot (D\mathbf{n} \mathbf{h}_{2\tau}) \\ &\quad + (D^2 h_{1,n} \mathbf{h}_{2\tau}) \cdot [\sigma \nabla_{\tau} u] + [\sigma D^2 u] \mathbf{h}_{2\tau} \cdot \nabla_{\tau} h_{1,n} - 2h_{2,n} (D\mathbf{n} [\sigma \nabla_{\tau} u]) \cdot \nabla_{\tau} h_{1,n} \\ &\quad + \nabla_{\tau} h_{1,n} \cdot \mathbf{h}_{2\tau} [\sigma \Delta_{\tau} u] + h_{1,n} \left( \overline{[\sigma \Delta_{\tau} u]} - [\sigma \Delta_{\tau} u'_2] \right). \end{aligned}$$

Taking into account the calculation

$$\begin{aligned} &- ([\sigma D^2 u'_1] \mathbf{h}_2) \cdot \mathbf{n} + [\sigma \nabla_{\tau} u'_1] \cdot \nabla_{\tau} h_{2,n} \\ &= - \left( h_{2,n} [\sigma D^2 u'_1] \mathbf{n} + [\sigma D^2 u'_1] \mathbf{h}_{2\tau} \right) \cdot \mathbf{n} + [\sigma \nabla_{\tau} u'_1] \cdot \nabla_{\tau} h_{2,n}, \\ &= h_{2,n} \left( [\sigma \Delta_{\tau} u'_1] + H [\sigma \partial_{\mathbf{n}} u'_1] \right) + [\sigma \nabla_{\tau} u'_1] \cdot \nabla_{\tau} h_{2,n} - ([\sigma D^2 u'_1] \mathbf{h}_{2\tau}) \cdot \mathbf{n}, \\ &= \operatorname{div}_{\tau} \left( h_{2,n} [\sigma \nabla_{\tau} u'_1] \right) + H h_{2,n} [\sigma \partial_{\mathbf{n}} u'_1] - ([\sigma D^2 u'_1] \mathbf{h}_{2\tau}) \cdot \mathbf{n}, \end{aligned}$$

we get

$$(38) \quad \begin{aligned} [\sigma \partial_{\mathbf{n}}(u'_1)'_2] &= \operatorname{div}_{\tau} \left( h_{1,n} [\sigma \nabla_{\tau} u'_2] + h_{2,n} [\sigma \nabla_{\tau} u'_1] \right. \\ &\quad \left. + (h_{2,n} \partial_{\mathbf{n}} h_{1,n} - \nabla_{\tau} h_{2,n} \cdot \mathbf{h}_{1\tau}) [\sigma \nabla_{\tau} u] \right) \\ &\quad + [\sigma \Delta_{\tau} u] \nabla_{\tau} h_{1,n} \cdot \mathbf{h}_{2\tau} + H h_{2,n} [\sigma \partial_{\mathbf{n}} u'_1] - ([\sigma D^2 u'_1] \mathbf{h}_{2\tau}) \cdot \mathbf{n} \\ &\quad - \left( [\sigma \nabla_{\tau} u'_1] + \partial_{\mathbf{n}} h_{1,n} [\sigma \nabla_{\tau} u] \right) \cdot (D\mathbf{n} \mathbf{h}_{2\tau}) + (D^2 h_{1,n} \mathbf{h}_{2\tau}) \cdot [\sigma \nabla_{\tau} u] \\ &\quad - 2h_{2,n} \nabla_{\tau} h_{1,n} \cdot (D\mathbf{n} [\sigma \nabla_{\tau} u]) + [\sigma D^2 u] \mathbf{h}_{2\tau} \cdot \nabla_{\tau} h_{1,n} \\ &\quad \left. + h_{1,n} \left( \overline{[\sigma \Delta_{\tau} u]} - [\sigma \Delta_{\tau} u'_2] \right) \right). \end{aligned}$$

This formula remains hard to handle. To get a more convenient one, we decide to derive, tangentially to the direction  $\mathbf{h}_2$ , the boundary identity

$$[\sigma \partial_{\mathbf{n}} u'_1] = h_{1,n} [\sigma \Delta_{\tau} u] + \nabla_{\tau} h_{1,n} \cdot [\sigma \nabla_{\tau} u].$$

This leads to

$$\begin{aligned} &([\sigma D^2 u'_1] \mathbf{h}_{2\tau}) \cdot \mathbf{n} + (D\mathbf{n} \mathbf{h}_{2\tau}) \cdot [\sigma \nabla_{\tau} u'_1] \\ &= \nabla_{\tau} h_{1,n} \cdot \mathbf{h}_{2\tau} [\sigma \Delta_{\tau} u] + h_{1,n} \nabla_{\tau} [\sigma \Delta_{\tau} u] \cdot \mathbf{h}_{2\tau} \\ &\quad + (D^2 h_{1,n} \mathbf{h}_{2\tau}) \cdot [\sigma \nabla_{\tau} u] - \partial_{\mathbf{n}} h_{1,n} [\sigma \nabla_{\tau} u] \cdot (D\mathbf{n} \mathbf{h}_{2\tau}) + [\sigma D^2 u] \mathbf{h}_{2\tau} \cdot \nabla_{\tau} h_{1,n}. \end{aligned}$$

The preceding formula enables us to simplify the expression (38):

$$\begin{aligned}
 [\sigma \partial_{\mathbf{n}}(u'_1)'_2] &= \operatorname{div}_{\tau} \left( h_{1,n} [\sigma \nabla_{\tau} u'_2] + h_{2,n} [\sigma \nabla_{\tau} u'_1] + (h_{2,n} \partial_{\mathbf{n}} h_{1,n} - \nabla_{\tau} h_{2,n} \cdot \mathbf{h}_{1\tau}) [\sigma \nabla_{\tau} u] \right) \\
 &\quad + H h_{2,n} [\sigma \partial_{\mathbf{n}} u'_1] - 2 h_{2,n} \nabla_{\tau} h_{1,n} \cdot (D\mathbf{n} [\sigma \nabla_{\tau} u]) \\
 (39) \quad &\quad - h_{1,n} \nabla_{\tau} [\sigma \Delta_{\tau} u] \cdot \mathbf{h}_{2\tau} + h_{1,n} \left( \overline{[\sigma \Delta_{\tau} u]} - [\sigma \Delta_{\tau} u'_2] \right).
 \end{aligned}$$

From (24)

$$\begin{aligned}
 \overline{[\sigma \Delta_{\tau} u]} &= [\sigma \Delta_{\tau} \dot{u}] + \nabla_{\tau} \operatorname{div}_{\tau} (\mathbf{h}_{2\tau}) \cdot [\sigma \nabla_{\tau} u] + \nabla_{\tau} (H \mathbf{h}_{2,n}) \cdot [\sigma \nabla_{\tau} u] \\
 (40) \quad &\quad - \operatorname{div}_{\tau} \left( \left( (D\mathbf{h}_2 + (D\mathbf{h}_2)^T) [\sigma \nabla_{\tau} u] \right)_{\tau} \right),
 \end{aligned}$$

and we have

$$[\sigma \Delta_{\tau} \dot{u}] = [\sigma \Delta_{\tau} u'] + \Delta_{\tau} ([\sigma \nabla u] \cdot \mathbf{h}_2).$$

That gives

$$\begin{aligned}
 (41) \quad \left( \overline{[\sigma \Delta_{\tau} u]} - [\sigma \Delta_{\tau} u'_2] \right) &= \Delta_{\tau} ([\sigma \nabla u] \cdot \mathbf{h}_2) + \nabla_{\tau} \operatorname{div}_{\tau} (\mathbf{h}_{2\tau}) \cdot [\sigma \nabla_{\tau} u] \\
 &\quad + \nabla_{\tau} (H \mathbf{h}_{2,n}) \cdot [\sigma \nabla_{\tau} u] - \operatorname{div}_{\tau} \left( \left( (D\mathbf{h}_2 + (D\mathbf{h}_2)^T) [\sigma \nabla_{\tau} u] \right)_{\tau} \right).
 \end{aligned}$$

We break up  $\mathbf{h}_2$  into normal and tangential components and use the formula of divergence; then expression (39) becomes

$$\begin{aligned}
 &[\sigma \partial_{\mathbf{n}}(u'_1)'_2] \\
 &= \operatorname{div}_{\tau} \left( h_{1,n} [\sigma \nabla_{\tau} u'_2] + h_{2,n} [\sigma \nabla_{\tau} u'_1] + (h_{2,n} \partial_{\mathbf{n}} h_{1,n} - \nabla_{\tau} h_{2,n} \cdot \mathbf{h}_{1\tau}) [\sigma \nabla_{\tau} u] \right) \\
 &\quad + \operatorname{div}_{\tau} (h_{1,n} h_{2,n} (H I - 2 D\mathbf{n}) \cdot [\sigma \nabla_{\tau} u]) - h_{1,n} (\nabla_{\tau} [\sigma \Delta_{\tau} u] \cdot \mathbf{h}_{2\tau} - \Delta_{\tau} [\sigma \nabla_{\tau} u] \cdot \mathbf{h}_{2\tau}) \\
 &\quad + h_{1,n} \left( \nabla_{\tau} \operatorname{div}_{\tau} (\mathbf{h}_{2\tau}) \cdot [\sigma \nabla_{\tau} u] - \operatorname{div}_{\tau} \left( \left( (D\mathbf{h}_{2\tau} + (D\mathbf{h}_{2\tau})^T) [\sigma \nabla_{\tau} u] \right)_{\tau} \right) \right).
 \end{aligned}$$

LEMMA 4. *We have*

$$\begin{aligned}
 &-\nabla_{\tau} [\sigma \Delta_{\tau} u] \cdot \mathbf{h}_{2\tau} + \Delta_{\tau} [\sigma \nabla_{\tau} u] \cdot \mathbf{h}_{2\tau} \\
 &+ \nabla_{\tau} \operatorname{div}_{\tau} (\mathbf{h}_{2\tau}) \cdot [\sigma \nabla_{\tau} u] - \operatorname{div}_{\tau} \left( \left( (D\mathbf{h}_{2\tau} + (D\mathbf{h}_{2\tau})^T) [\sigma \nabla_{\tau} u] \right)_{\tau} \right) = 0.
 \end{aligned}$$

*Proof.* By using the relation between the material and shape derivatives, we obtain

$$\overline{[\sigma \Delta_{\tau} u]} = [\sigma \Delta_{\tau} u'] + \nabla ([\sigma \Delta_{\tau} u]) \cdot \mathbf{h}_2 \quad \text{and} \quad [\sigma \Delta_{\tau} \dot{u}] = [\sigma \Delta_{\tau} u'] + \Delta_{\tau} ([\sigma \nabla u] \cdot \mathbf{h}_2).$$

Let us inject these relations in (40) and while applying for the direction  $\mathbf{h}_{2\tau}$ , we obtain

$$\begin{aligned} \Delta_\tau([\sigma \nabla_\tau u] \cdot \mathbf{h}_{2\tau}) + \nabla_\tau \operatorname{div}_\tau (\mathbf{h}_{2\tau}) \cdot [\sigma \nabla_\tau u] \\ = \nabla_\tau [\sigma \Delta_\tau u] \cdot \mathbf{h}_{2\tau} + \operatorname{div}_\tau \left( \left( (D\mathbf{h}_{2\tau} + (D\mathbf{h}_{2\tau})^T) [\sigma \nabla_\tau u] \right)_\tau \right). \end{aligned}$$

This completes the proof.  $\square$

According to Lemma 4, we obtain

$$\begin{aligned} [\sigma \partial_{\mathbf{n}}(u'_1)'_2] = \operatorname{div}_\tau \left( h_{1,n} [\sigma \nabla_\tau u'_2] + h_{2,n} [\sigma \nabla_\tau u'_1] + (h_{2,n} \partial_{\mathbf{n}} h_{1,n} - \nabla_\tau h_{2,n} \cdot \mathbf{h}_{1\tau}) [\sigma \nabla_\tau u] \right) \\ + \operatorname{div}_\tau (h_{1,n} h_{2,n} (HI - 2D\mathbf{n}) \cdot [\sigma \nabla_\tau u]). \end{aligned}$$

As a conclusion, we use the relation of second derivative

$$[\sigma \partial_{\mathbf{n}} u''_{1,2}] = [\sigma \partial_{\mathbf{n}}(u'_1)'_2] - [\sigma \partial_{\mathbf{n}} u'_{D\mathbf{h}_1 \mathbf{h}_2}] = [\sigma \partial_{\mathbf{n}}(u'_1)'_2] - \operatorname{div}_\tau (D\mathbf{h}_1 \mathbf{h}_2 \cdot \mathbf{n} [\sigma \nabla_\tau u]).$$

Let us split the field  $\mathbf{h}_2$  in two parts:  $D\mathbf{h}_1 \mathbf{h}_2 \cdot \mathbf{n} = h_{2,n} \mathbf{n} \cdot D\mathbf{h}_1 \mathbf{n} + D\mathbf{h}_1 \mathbf{h}_{2\tau} \cdot \mathbf{n}$ . In the spirit of (73), we obtain

$$(42) \quad D\mathbf{h}_1 \mathbf{h}_{2\tau} \cdot \mathbf{n} = \nabla_\tau h_{1,n} \cdot \mathbf{h}_{2\tau} - \mathbf{h}_{1\tau} \cdot D\mathbf{n} \mathbf{h}_{2\tau}.$$

Thanks to (72), the jump  $[\sigma \partial_{\mathbf{n}} u'_{D\mathbf{h}_1 \mathbf{h}_2}]$  can then be written in the form

$$[\sigma \partial_{\mathbf{n}} u'_{D\mathbf{h}_1 \mathbf{h}_2}] = \operatorname{div}_\tau ((h_{2,n} \mathbf{n} \cdot \nabla_\tau h_{1,n} + \nabla_\tau h_{1,n} \cdot \mathbf{h}_{2\tau} - \mathbf{h}_{1\tau} \cdot D\mathbf{n} \mathbf{h}_{2\tau}) [\sigma \nabla_\tau u]).$$

Gathering all the terms, simplifications occur and we obtain the desired result:

$$\begin{aligned} [\sigma \partial_{\mathbf{n}} u''_{1,2}] = \operatorname{div}_\tau \left( h_{2,n} [\sigma \nabla_\tau u'_1] + h_{1,n} [\sigma \nabla_\tau u'_2] \right) \\ - \operatorname{div}_\tau ((\mathbf{h}_{1\tau} \cdot \nabla_\tau h_{2,n} + \nabla_\tau h_{1,n} \cdot \mathbf{h}_{2\tau}) [\sigma \nabla_\tau u]) \\ - \operatorname{div}_\tau (h_{2,n} h_{1,n} (2D\mathbf{n} - HI) [\sigma \nabla_\tau u]) + \operatorname{div}_\tau (\mathbf{h}_{1\tau} \cdot D\mathbf{n} \mathbf{h}_{2\tau}) [\sigma \nabla_\tau u]. \end{aligned}$$

To get the jumps of the potential, we use (29) and (68):

$$\begin{aligned} [u''_{1,2}] = [(u'_1)'_2] - [u'_{D\mathbf{h}_1 \mathbf{h}_2}] = -\mathbf{h}_1 \cdot [\nabla u'_2] - \mathbf{h}_2 \cdot [\nabla u_1] - [u'_{D\mathbf{h}_1 \mathbf{h}_2}] \\ = -h_{2,n} [\partial_{\mathbf{n}} u'_1] - h_{1,n} [\partial_{\mathbf{n}} u'_2] - \mathbf{h}_{2\tau} \cdot [\nabla_\tau u'_1] - \mathbf{h}_{1\tau} \cdot [\nabla_\tau u'_2] \\ - h_{2,n} \mathbf{n} \cdot [\nabla(\mathbf{h}_1 \cdot \nabla u)] - h_{1,n} \mathbf{n} \cdot [\nabla(\mathbf{h}_2 \cdot \nabla u)] - [u'_{D\mathbf{h}_1 \mathbf{h}_2}]. \end{aligned}$$

Thanks to the jump of the potential for the first order shape derivative given in (6), we obtain

$$\mathbf{h}_{2\tau} \cdot [\nabla_\tau u'_1] = -\mathbf{h}_{2\tau} \cdot [\nabla(\mathbf{h}_1 \cdot \nabla u)] \quad \text{and} \quad \mathbf{h}_{1\tau} \cdot [\nabla_\tau u'_2] = -\mathbf{h}_{1\tau} \cdot [\nabla(\mathbf{h}_2 \cdot \nabla u)]$$

and then

$$\begin{aligned} (43) \quad [u''_{1,2}] = -h_{2,n} [\partial_{\mathbf{n}} u'_1] - h_{1,n} [\partial_{\mathbf{n}} u'_2] - h_{2,n} \mathbf{n} \cdot [\nabla(\mathbf{h}_1 \cdot \nabla u)] \\ + \mathbf{h}_{1\tau} \cdot [\nabla(\mathbf{h}_2 \cdot \nabla u)] - [u'_{D\mathbf{h}_1 \mathbf{h}_2}]. \end{aligned}$$



Computing the other jumps that appeared in the former expression, we get

$$\begin{aligned}
 [\nabla(\mathbf{h}_2 \cdot \nabla u)] &= (D\mathbf{h}_2)^T [\nabla u] + [D^2 u] \mathbf{h}_2, \\
 \mathbf{h}_{1\tau} \cdot [\nabla(\mathbf{h}_2 \cdot \nabla u)] &= \mathbf{n} \cdot D\mathbf{h}_2 \mathbf{h}_{1\tau} [\partial_{\mathbf{n}} u] + h_{2,n} \mathbf{h}_{1\tau} \cdot [D^2 u] \mathbf{n} + \mathbf{h}_{1\tau} \cdot [D^2 u] \mathbf{h}_{2\tau}, \\
 h_{2,n} \mathbf{n} \cdot [\nabla(\mathbf{h}_1 \cdot \nabla u)] &= h_{2,n} [\partial_{\mathbf{n}} u] \mathbf{n} \cdot D\mathbf{h}_1 \mathbf{n} + h_{2,n} h_{1,n} \mathbf{n} \cdot [D^2 u] \mathbf{n} + h_{2,n} \mathbf{n} \cdot [D^2 u] \mathbf{h}_{1\tau}, \\
 [u'_{D\mathbf{h}_1 \mathbf{h}_2}] &= -D\mathbf{h}_1 \mathbf{h}_2 \cdot \mathbf{n} [\partial_{\mathbf{n}} u] = - (h_{2,n} \mathbf{n} \cdot D\mathbf{h}_1 \mathbf{n} + \mathbf{n} \cdot D\mathbf{h}_1 \mathbf{h}_{2\tau}) [\partial_{\mathbf{n}} u].
 \end{aligned}$$

With the help of formula (42), we obtain

$$\begin{aligned}
 &-h_{2,n} \mathbf{n} \cdot [\nabla(\mathbf{h}_1 \cdot \nabla u)] + \mathbf{h}_{1\tau} \cdot [\nabla(\mathbf{h}_2 \cdot \nabla u)] - [u'_{D\mathbf{h}_1 \mathbf{h}_2}] \\
 &= (\nabla_{\tau} h_{1,n} \cdot \mathbf{h}_{2\tau} + \nabla_{\tau} h_{2,n} \cdot \mathbf{h}_{1\tau}) [\partial_{\mathbf{n}} u] \\
 &\quad - 2\mathbf{h}_{1\tau} \cdot D\mathbf{n} \mathbf{h}_{2\tau} [\partial_{\mathbf{n}} u] + \mathbf{h}_{1\tau} \cdot [D^2 u] \mathbf{h}_{2\tau} - h_{2,n} h_{1,n} \mathbf{n} \cdot [D^2 u] \mathbf{n}, \\
 \mathbf{n} \cdot [D^2 u] \mathbf{n} &= -[\Delta_{\tau} u] - H [\partial_{\mathbf{n}} u] = -H [\partial_{\mathbf{n}} u], \\
 \mathbf{h}_{1\tau} \cdot [D^2 u] \mathbf{h}_{2\tau} &= \mathbf{h}_{1\tau} \cdot D([\nabla u]) \mathbf{h}_{2\tau} = \mathbf{h}_{1\tau} \cdot D([\partial_{\mathbf{n}} u] \mathbf{n}) \mathbf{h}_{2\tau} = [\partial_{\mathbf{n}} u] \mathbf{h}_{1\tau} \cdot D\mathbf{n} \mathbf{h}_{2\tau}.
 \end{aligned}$$

Finally, we gather the results of these computations to write

$$\begin{aligned}
 (44) \quad [u''_{1,2}] &= - \left( h_{2,n} [\partial_{\mathbf{n}} u'_1] + h_{1,n} [\partial_{\mathbf{n}} u'_2] \right) + (\nabla_{\tau} h_{1,n} \cdot \mathbf{h}_{2\tau} + \nabla_{\tau} h_{2,n} \cdot \mathbf{h}_{1\tau}) [\partial_{\mathbf{n}} u] \\
 &\quad + (h_{2,n} h_{1,n} H - \mathbf{h}_{1\tau} \cdot D\mathbf{n} \mathbf{h}_{2\tau}) [\partial_{\mathbf{n}} u].
 \end{aligned}$$

**Case of Neumann boundary conditions.** Since the admissible deformation fields have a support with no intersection points with the outer boundary, it is a straightforward application of the preceding computations to show that the  $u_n$  solution to (1), (4) is twice differentiable with respect to the shape. Furthermore, its second order derivative  $u''_n$  belongs to  $H^1(\Omega \setminus \bar{\omega}) \cup H^1(\omega)$  and solves

$$(45) \quad \left\{ \begin{array}{l} \Delta u''_n = 0 \text{ in } \omega \setminus \bar{\omega} \text{ and in } \omega, \\ [u''_n] = (h_{1,n} h_{2,n} H - \mathbf{h}_{1\tau} \cdot D\mathbf{n} \mathbf{h}_{2\tau}) [\partial_{\mathbf{n}} u_n] - (h_{1,n} [\partial_{\mathbf{n}}(u_n)'_2] + h_{2,n} [\partial_{\mathbf{n}}(u_n)'_1]) \\ \quad + (\mathbf{h}_{1\tau} \cdot \nabla h_{2,n} + \mathbf{h}_{2\tau} \cdot \nabla h_{1,n}) [\partial_{\mathbf{n}} u_n] \text{ on } \partial\omega, \\ [\sigma \partial_n u''_n] = \operatorname{div}_{\tau} \left( h_{2,n} [\sigma \nabla_{\tau}(u_n)'_1] + h_{1,n} [\sigma \nabla_{\tau}(u_n)'_2] + \mathbf{h}_{1\tau} \cdot D\mathbf{n} \mathbf{h}_{2\tau} [\sigma \nabla_{\tau} u_n] \right) \\ \quad - \operatorname{div}_{\tau} ((\mathbf{h}_{1\tau} \cdot \nabla_{\tau} h_{2,n} + \nabla_{\tau} h_{1,n} \cdot \mathbf{h}_{2\tau} \\ \quad + h_{2,n} h_{1,n} (2D\mathbf{n} - HI)) [\sigma \nabla_{\tau} u_n]) \text{ on } \partial\omega, \\ \partial_n u''_n = 0 \text{ on } \partial\Omega, \end{array} \right.$$

where we use the notation of Theorem 2.

### 5. Second order derivatives for the criterion.

**5.1. Proof of Theorem 4.** The differentiability of the objective is a direct application of Theorem 2. The computation we make here is based on the relation

$$(46) \quad D^2 J_{KV}(\omega)(\mathbf{h}_1, \mathbf{h}_2) = D(DJ_{KV}(\omega)\mathbf{h}_1)\mathbf{h}_2 - DJ_{KV}(\omega)D\mathbf{h}_1\mathbf{h}_2.$$

To obtain (10), we compute in the first step the shape gradient in the direction  $\mathbf{h}_1$ . Then, in the second step, we differentiate the obtained expression in the direction of  $\mathbf{h}_2$ . In what follows, we adopt the notation  $v = u_d - u_n$  to obtain concise expressions.

$$\begin{aligned} DJ_{KV}(\omega)\mathbf{h}_1 &= \sigma_1 \int_{\Omega \setminus \bar{\omega}} \operatorname{div}(|\nabla v|^2 \mathbf{h}_1) + 2\nabla v \cdot \nabla v'_1 + \sigma_2 \int_{\omega} \operatorname{div}(|\nabla v|^2 \mathbf{h}_1) + 2\nabla v \cdot \nabla v'_1 \\ &= \sigma_1(A_1 + 2B_1) + \sigma_2(A_2 + 2B_2), \end{aligned}$$

where

$$\begin{aligned} A_1 &= \int_{\Omega \setminus \bar{\omega}} \operatorname{div}(|\nabla v|^2 \mathbf{h}_1), & B_1 &= \int_{\Omega \setminus \bar{\omega}} \nabla v \cdot \nabla v'_1, \\ A_2 &= \int_{\omega} \operatorname{div}(|\nabla v|^2 \mathbf{h}_1), & B_2 &= \int_{\omega} \nabla v \cdot \nabla v'_1. \end{aligned}$$

Now, we use the classical formulae to differentiate a domain integral:

$$\begin{aligned} DA_1(\omega)\mathbf{h}_2 &= \int_{\Omega \setminus \bar{\omega}} \operatorname{div} \left( \operatorname{div}(|\nabla v|^2 \mathbf{h}_1) \mathbf{h}_2 \right) + 2 \operatorname{div}(\nabla v \cdot \nabla v'_1 \mathbf{h}_2) \\ &= - \int_{\partial\omega} \operatorname{div}(|\nabla v|^2 \mathbf{h}_1) h_{2,n} + 2\nabla v^+ \cdot \nabla (v'_1)^+ h_{1,n}, \\ DA_2(\omega)\mathbf{h}_2 &= \int_{\partial\omega} \operatorname{div}(|\nabla v|^2 \mathbf{h}_1) h_{2,n} + 2\nabla v^- \cdot \nabla (v'_1)^- h_{1,n}. \end{aligned}$$

The terms  $DB_i$ ,  $i = 1, 2$ , require more precision. First, we write

$$\begin{aligned} DB_1(\omega)\mathbf{h}_2 &= \int_{\Omega \setminus \bar{\omega}} \operatorname{div}(\nabla v \cdot \nabla v'_1 \mathbf{h}_2) + \nabla v'_1 \cdot \nabla v'_2 + \nabla v \cdot \nabla (v'_1)'_2 \\ &= - \int_{\partial\omega} \nabla v^+ \cdot \nabla (v'_1)^+ h_{2,n} + \partial_{\mathbf{n}} v^+ ((v'_1)^+)'_2 \\ &\quad + \frac{1}{2} \left( \partial_{\mathbf{n}} (v'_1)^+ (v'_2)^+ + \partial_{\mathbf{n}} (v'_2)^+ (v'_1)^+ \right) \\ &\quad - \int_{\partial\Omega} \partial_{\mathbf{n}} v ((u_n)'_1)'_2 + \frac{1}{2} \left( \partial_{\mathbf{n}} (u_d)'_1 (u_n)'_2 + \partial_{\mathbf{n}} (u_d)'_2 (u_n)'_1 \right). \end{aligned}$$

Note that we used the Green formula twice to keep the symmetry in  $\mathbf{h}_1$  and  $\mathbf{h}_2$ . We also use the fact that the derivatives  $(u_d)'_i$  are harmonic in  $\Omega \setminus \bar{\omega}$  to transform the boundary integral on the exterior boundary into an integral on the moving boundary. We obtain

$$\begin{aligned} DB_1(\omega)\mathbf{h}_2 &= - \int_{\partial\omega} \nabla v^+ \cdot \nabla (v'_1)^+ \mathbf{h}_{2,n} + \partial_{\mathbf{n}} v^+ (((u_d)'_1)'_2)^+ - v \partial_{\mathbf{n}} (((u_n)'_1)'_2)^+ \\ &\quad - \int_{\partial\omega} \frac{1}{2} \left( \partial_{\mathbf{n}} (v'_1)^+ ((u_d)'_2)^+ + \partial_{\mathbf{n}} (v'_2)^+ ((u_d)'_1)^+ \right. \\ &\quad \left. - \partial_{\mathbf{n}} ((u_n)'_1)^+ (v'_2)^+ - \partial_{\mathbf{n}} ((u_n)'_2)^+ (v'_1)^+ \right). \end{aligned}$$

By the same methods, we get

$$\begin{aligned} DB_2(\omega)\mathbf{h}_2 &= \int_{\partial\omega} \nabla v^- \cdot \nabla (v'_1)^- h_{2,n} + \partial_{\mathbf{n}} v^- ((v'_1)'_2)^- \\ &\quad + \frac{1}{2} \left( \partial_{\mathbf{n}} (v'_1)^- (v'_2)^- + \partial_{\mathbf{n}} (v'_2)^- (v'_1)^- \right). \end{aligned}$$

We regroup the different terms, and after some straightforward computations, we obtain

$$\begin{aligned} D(DJ_{KV}(\omega)\mathbf{h}_1)(\omega)\mathbf{h}_2 &= - \int_{\partial\omega} \operatorname{div} \left( \left[ \sigma |\nabla v|^2 \mathbf{h}_1 \right] \right) + 2 \left[ \sigma \nabla v \cdot \left( h_{1,n} \nabla v'_2 + h_{2,n} \nabla v'_1 \right) \right] \\ &\quad - \int_{\partial\omega} \left[ \sigma \left( (u_d)'_2 \partial_{\mathbf{n}} v'_1 + (u_d)'_1 \partial_{\mathbf{n}} v'_2 - \partial_{\mathbf{n}} (u_n)'_2 v'_1 - \partial_{\mathbf{n}} (u_n)'_1 v'_2 \right) \right] \\ &\quad + 2 \int_{\partial\omega} v \left[ \sigma \partial_{\mathbf{n}} ((u_n)'_1)'_2 \right] - \sigma_1 \partial_{\mathbf{n}} v^+ \left[ ((u_d)'_1)'_2 \right]. \end{aligned}$$

In order to compute  $D^2 J_{KV}(\omega)(\mathbf{h}_1, \mathbf{h}_2)$ , the first order derivative of the Kohn–Vogelius objective is needed. It can be written as follows:

$$DJ_{KV}(\omega)\mathbf{h} = - \int_{\partial\omega} \left[ \sigma |\nabla v|^2 \right] h_n + 2 \int_{\partial\omega} v \left[ \sigma \partial_{\mathbf{n}} (u_n)' \right] - \sigma_1 \partial_{\mathbf{n}} v^+ \left[ (u_d)' \right].$$

Gathering (46) with the jump relations for the second order derivatives, we write the second derivative of the Kohn–Vogelius criterion as

$$\begin{aligned} D^2 J_{KV}(\omega)(\mathbf{h}_1, \mathbf{h}_2) &= - \int_{\partial\omega} \operatorname{div} \left( \left[ \sigma |\nabla v|^2 \mathbf{h}_1 \right] \right) - \left[ \sigma |\nabla v|^2 \right] (D\mathbf{h}_1 \mathbf{h}_2) \cdot \mathbf{n} \\ &\quad - \int_{\partial\omega} \left[ \sigma \left( (u_d)'_2 \partial_{\mathbf{n}} v'_1 + (u_d)'_1 \partial_{\mathbf{n}} v'_2 - \partial_{\mathbf{n}} (u_n)'_2 v'_1 - \partial_{\mathbf{n}} (u_n)'_1 v'_2 \right) \right] \\ &\quad + 2 \int_{\partial\omega} \left[ \sigma \nabla v \cdot \left( h_{1,n} \nabla v'_2 + h_{2,n} \nabla v'_1 \right) \right] + v \left[ \sigma \partial_{\mathbf{n}} (u_n)''_{1,2} \right] - \sigma_1 \partial_{\mathbf{n}} v^+ \left[ (u_d)''_{1,2} \right]. \end{aligned}$$

Let us give a simplified version for the first term. We decompose the field  $\mathbf{h}_2$  into normal vector and tangential parts and use (42). After some elementary computations, we obtain

$$\begin{aligned} &- \int_{\partial\omega} \operatorname{div} \left( \left[ \sigma |\nabla v|^2 \mathbf{h}_1 \right] \right) - \left[ \sigma |\nabla v|^2 \right] (D\mathbf{h}_1 \mathbf{h}_2) \cdot \mathbf{n} \\ &= \int_{\partial\omega} \left[ \sigma |\nabla v|^2 \right] (\mathbf{h}_{1\tau} \cdot \nabla_{\tau} h_{2,n} + \mathbf{h}_{2\tau} \cdot \nabla_{\tau} h_{1,n} - \mathbf{h}_{2\tau} \cdot D\mathbf{n} \mathbf{h}_{1\tau}) \\ &\quad - \int_{\partial\omega} \partial_{\mathbf{n}} \left( \left[ \sigma |\nabla v|^2 \right] \right) h_{1,n} h_{2,n}. \end{aligned}$$

Finally, the second order derivative of the Kohn–Vogelius objective becomes

(47)

$$\begin{aligned} D^2 J_{KV}(\omega)(\mathbf{h}_1, \mathbf{h}_2) &= \int_{\partial\omega} [\sigma |\nabla v|^2] (\mathbf{h}_{1\tau} \cdot \nabla_\tau h_{2,n} + \mathbf{h}_{2\tau} \cdot \nabla_\tau h_{1,n} - \mathbf{h}_{2\tau} \cdot D\mathbf{n} \mathbf{h}_{1\tau}) \\ &\quad - \int_{\partial\omega} \partial_{\mathbf{n}} \left( [\sigma |\nabla v|^2] \right) h_{1,n} h_{2,n} + 2 \left[ \sigma \nabla v \cdot (h_{1,n} \nabla v'_2 + h_{2,n} \nabla v'_1) \right] \\ &\quad - \int_{\partial\omega} \left[ \sigma \left( (u_d)'_2 \partial_{\mathbf{n}} v'_1 + (u_d)'_1 \partial_{\mathbf{n}} v'_2 - \partial_{\mathbf{n}} (u_n)'_2 v'_1 - \partial_{\mathbf{n}} (u_n)'_1 v'_2 \right) \right] \\ &\quad + 2 \int_{\partial\omega} v \left[ \sigma \partial_{\mathbf{n}} (u_n)''_{1,2} \right] - \sigma_1 \partial_{\mathbf{n}} v^+ \left[ (u_d)''_{1,2} \right]. \end{aligned}$$

**5.2. Analysis of stability. Proof of Theorem 5.** Now, we specify the domain  $\omega$  that is assumed to be a critical shape for  $J_{KV}$ . Moreover, we assume that the additional condition  $u_d = u_n$  holds. To emphasize that we deal with such a special domain, we will denote it  $\omega^*$ . The assumptions mean that the measurements are compatible and that  $\omega^*$  is a global minimum of the criterion. From the necessary condition of order two at a minimum, the shape Hessian is positive at such a point.

Let us notice that only the normal component of  $\mathbf{h}$  appears. Let us also emphasize that there is no hope to get  $\mathbf{h} = 0$  from the structure theorem for the second order shape derivative (see [7]). The deformation field  $\mathbf{h}$  appears in  $D^2 J_{KV}(\omega^*)(\mathbf{h}, \mathbf{h})$  only through its normal component  $h_n$  since  $\omega^*$  is a critical point for  $J_{KV}$ . This remark explains why we consider in the statement of Theorem 5 the scalar Sobolev space corresponding to the normal components of the deformation field.

We now prove Theorem 5. From (47), we deduce

$$\begin{aligned} (48) \quad D^2 J_{KV}(\omega^*)[h, h] &= -2 \int_{\partial\omega^*} \left[ \sigma \left( u_d' \partial_{\mathbf{n}} v' - \partial_{\mathbf{n}} u_n' v' \right) \right] \\ &= 2 [\sigma] \int_{\partial\omega^*} \left( (u_d'^+ - u_n'^+) \operatorname{div}_\tau (h_n \nabla_\tau u_d) - \frac{\sigma_1}{\sigma_2} \partial_{\mathbf{n}} u_d^+ h_n \partial_{\mathbf{n}} (u_d' - u_n')^+ \right) \\ &= 2 [\sigma] \left( \left\langle u_d'^+ - u_n'^+, \operatorname{div}_\tau (h_n \nabla_\tau u_d) \right\rangle - \frac{\sigma_1}{\sigma_2} \left\langle \partial_{\mathbf{n}} u_d h_n, \partial_{\mathbf{n}} (u_d' - u_n')^+ \right\rangle \right), \end{aligned}$$

where  $\langle \cdot, \cdot \rangle$  denotes the duality between  $H^{1/2}(\partial\omega^*) \times H^{-1/2}(\partial\omega^*)$ . Let us introduce the operators

$$\begin{aligned} T_1 : H^{1/2}(\partial\omega^*) &\rightarrow H^{-1/2}(\partial\omega^*), & \mathbf{h} &\mapsto \operatorname{div}_\tau (h_n \nabla_\tau u_d), \\ M_1 : H^{1/2}(\partial\omega^*) &\rightarrow H^{1/2}(\partial\omega^*), & \mathbf{h} &\mapsto u_d'^+ - u_n'^+, \\ T_2 : H^{1/2}(\partial\omega^*) &\rightarrow H^{1/2}(\partial\omega^*), & \mathbf{h} &\mapsto h_n \partial_{\mathbf{n}} u_d^+, \\ M_2 : H^{1/2}(\partial\omega^*) &\rightarrow H^{-1/2}(\partial\omega^*), & \mathbf{h} &\mapsto \partial_{\mathbf{n}} (u_d'^+ - u_n'^+). \end{aligned}$$

The Hessian can then be written in the form

$$D^2 J_{KV}(\omega^*)(\mathbf{h}, \mathbf{h}) = 2 [\sigma] \left( \left\langle M_1(\mathbf{h}), T_1(\mathbf{h}) \right\rangle - \frac{\sigma_1}{\sigma_2} \left\langle T_2(\mathbf{h}), M_2(\mathbf{h}) \right\rangle \right).$$

From the classical results of Maz'ya and Shaposhnikova on multipliers [15, 24], we get easily that  $T_1$  and  $T_2$  are continuous operators. In fact, the compactness of the Hessian is a consequence of the fact that both operators  $M_1$  and  $M_2$  are compact. We use a regularity argument and remark that  $M_1$  is the composition of the operators

$$\begin{aligned} R_1 : H^{1/2}(\partial\omega^*) &\rightarrow H_\diamond^{1/2}(\partial\Omega) & \text{and} & & R_2 : H_\diamond^{1/2}(\partial\Omega) &\rightarrow H^{1/2}(\partial\omega^*) \\ \mathbf{h} &\mapsto -u'_n & & & \phi &\mapsto \psi, \end{aligned}$$

where  $\psi$  is the trace on  $\partial\omega^*$  of the  $\Psi$  solution of

$$(49) \quad \begin{cases} -\Delta\Psi = 0 \text{ in } \Omega \setminus \overline{\omega^*} \text{ and in } \omega^*, \\ [\Psi] = 0 \text{ on } \partial\omega^*, \\ [\sigma\partial_{\mathbf{n}}\Psi] = 0 \text{ on } \partial\omega^*, \\ \Psi = \phi \text{ on } \partial\Omega. \end{cases}$$

While  $R_1$  is a continuous operator, we prove that  $R_2$  is compact. Let us express  $u|_{\partial\omega^*} = \psi$ . We use the integral representation formula. The definition of the layer operators and all the justifications are postponed until section 6.1 for the lightness of the presentation. If  $u$  solves the boundary value problem (49), then it also solves the following system of integral equations:

$$\begin{bmatrix} \frac{1}{2}I + \mu K_{\omega^*} & \frac{\sigma_1}{\sigma_2 + \sigma_1} S_{\partial\Omega\partial\omega^*} \\ \mu K_{\partial\omega^*\partial\Omega} & \frac{\sigma_1}{\sigma_2 + \sigma_1} S_{\Omega} \end{bmatrix} \begin{bmatrix} (u)|_{\partial\omega^*} \\ (\partial_{\mathbf{n}}u)|_{\partial\Omega} \end{bmatrix} = \frac{\sigma_1}{\sigma_1 + \sigma_2} \begin{bmatrix} K_{\partial\Omega\partial\omega^*}\phi \\ \left(-\frac{1}{2} + K_{\Omega}\right)\phi \end{bmatrix}.$$

The matrix operator arising in this equation appears also in (55). It has a continuous inverse thanks to Theorem 7, proved in section 6.1. Let us express  $u|_{\partial\omega^*} = \psi$ :

$$\begin{aligned} (50) \quad & \left[ \left( \frac{1}{2}I + \mu K_{\omega^*} \right) - \mu S_{\partial\Omega\partial\omega^*} S_{\Omega}^{-1} K_{\partial\omega^*\partial\Omega} \right] \psi \\ &= \frac{\sigma_1}{\sigma_1 + \sigma_2} \left[ K_{\partial\Omega\partial\omega^*} - S_{\partial\Omega\partial\omega^*} S_{\Omega}^{-1} \left( \frac{1}{2}I - K_{\Omega} \right) \right] \phi. \end{aligned}$$

Since the operators  $K_{\partial\Omega\partial\omega^*}$  and  $S_{\partial\Omega\partial\omega^*}$  are compact, the operator  $R_2$  is compact, and hence  $M_1$  is compact. The proof of compactness of  $M_2$  is similar. Let us mention that a similar strategy of proof can be found in [6].

The natural question is then as to how this optimization problem is ill-posed. This question is related directly to the rate at which the singular values of the Hessian operator are decreasing. Equation (50) shows that this rate is that of the operators  $K_{\partial\Omega\partial\omega^*}$ , and  $S_{\partial\Omega\partial\omega^*}$ . Now, since for every  $u \in H^{1/2}(\partial\Omega)$ , the functions  $K_{\partial\Omega\partial\omega^*}u$  and  $S_{\partial\Omega\partial\omega^*}u$  are harmonic outside of  $\partial\Omega$  and therefore in  $\Omega$ , their restrictions on  $\partial\omega^*$  are as smooth as  $\partial\omega^*$ . We conclude that if  $\partial\omega^*$  is  $\mathcal{C}^\infty$ , then the restriction belongs to each  $H^s(\partial\omega^*)$  for  $s > 1/2$ , and that if  $\lambda_n$  denotes the  $n$ th eigenvalue of  $D^2 J_{KV}(\omega^*)$ , then  $\lambda_n = o(n^{-s})$  for all  $s > 0$ .

## 6. Appendices.

**6.1. An auxiliary boundary value problem.** We have to justify rigorously that the right-hand sides of (6)–(8) make sense. They involve tangential derivatives of  $u_n$  and  $u_d$  along the interface  $\partial\omega$  up to order three. The existence of these derivatives is not clear a priori since the gradient of the solution has a discontinuity along this interface. Our first aim is to make the tangential regularity precise along the interface  $\partial\omega$  of the solution  $u$  of (1) with either Dirichlet or Neumann boundary conditions. Our strategy is to find an integral equation solved by this trace and to use known properties on the layers operators to obtain regularity on the solution of the integral equation.

We should have access to the trace of  $u$  on the interface  $\partial\omega$ . To that end, we introduce for any  $\alpha \in H^{1/2}(\partial\omega)$  and  $\beta \in H^{-1/2}(\partial\omega)$  the following boundary value problems:

(51)

$$(D) \left\{ \begin{array}{l} \Delta v = 0 \text{ in } \Omega \setminus \bar{\omega} \text{ and in } \omega, \\ [v] = \alpha \text{ on } \partial\omega, \\ [\sigma \partial_n v] = \beta \text{ on } \partial\omega, \\ v = f_1 \text{ on } \partial\Omega, \end{array} \right. \quad (N) \left\{ \begin{array}{l} \Delta v = 0 \text{ in } \Omega \setminus \bar{\omega} \text{ and in } \omega, \\ [v] = \alpha \text{ on } \partial\omega, \\ [\sigma \partial_n v] = \beta \text{ on } \partial\omega, \\ \partial_n v = g_1 \text{ on } \partial\Omega, \end{array} \right.$$

where  $(f_1, g_1) \in H^{1/2}(\partial\Omega) \times H^{-1/2}(\partial\Omega)$ . Note that for  $\alpha = 0$ ,  $\beta = 0$ , and  $(f_1, g_1) = (f, g)$ , then  $(u_d)$  and  $u_n$  solve, respectively, (D) and (N); furthermore the choice of

$$(52) \quad \alpha = \frac{[\sigma]}{\sigma_2} h_n \partial_n u^+ \quad \text{and} \quad \beta = [\sigma] \operatorname{div}_\tau (h_n \nabla_\tau u)$$

leads to (6) and (7) when we take  $(f_1, g_1) = (0, 0)$ .

*Existence of solutions to (D) and (N).* To study these problems, we use the integral representation in terms of layer potentials.

In the first step, we recall some definitions. The fundamental solution to the Laplace equation  $\Gamma$  is defined as

$$\Gamma(x, y) = -\frac{1}{4\pi} \frac{1}{|x - y|} \quad \text{if } n = 3.$$

The integral equations applying to the direct problem will be obtained from a study of the classical single- and double-layer potentials. We begin to introduce the following operators:

$$\begin{aligned} S_{\partial\Omega\partial\omega} : u &\mapsto S_{\partial\Omega\partial\omega}u(x) := \int_{\partial\Omega} \Gamma(x, y)u(y) \, d\sigma(y), \quad x \in \partial\omega; \\ S_{\partial\omega\partial\Omega} : u &\mapsto S_{\partial\omega\partial\Omega}u(x) := \int_{\partial\omega} \Gamma(x, y)u(y) \, d\sigma(y), \quad x \in \partial\Omega; \\ K_{\partial\Omega\partial\omega} : u &\mapsto K_{\partial\Omega\partial\omega}u(x) := \int_{\partial\Omega} \partial_n \Gamma(x, y)u(y) \, d\sigma(y), \quad x \in \partial\omega; \\ K_{\partial\omega\partial\Omega} : u &\mapsto K_{\partial\omega\partial\Omega}u(x) := \int_{\partial\omega} \partial_n \Gamma(x, y)u(y) \, d\sigma(y), \quad x \in \partial\Omega. \end{aligned}$$

Note that all these operators have a smooth kernel since the boundaries  $\partial\omega$  and  $\partial\Omega$  are assumed to have no common point. We also denote

$$\begin{aligned} S_\Omega : u &\mapsto S_\Omega u(x) := \int_{\partial\Omega} \Gamma(x, y) u(y) \, d\sigma(y), \quad x \in \partial\Omega; \\ K_\Omega : u &\mapsto K_\Omega u(x) := \int_{\partial\Omega} \partial_{\mathbf{n}} \Gamma(x, y) u(y) \, d\sigma(y), \quad x \in \partial\Omega; \\ S_\omega : u &\mapsto S_\omega u(x) := \int_{\partial\omega} \Gamma(x, y) u(y) \, d\sigma(y), \quad x \in \partial\omega; \\ K_\omega : u &\mapsto K_\omega u(x) := \int_{\partial\omega} \partial_{\mathbf{n}} \Gamma(x, y) u(y) \, d\sigma(y), \quad x \in \partial\omega. \end{aligned}$$

We obtain some systems of integral equations to compute the state function and their shape derivatives. Since  $v$  is harmonic in  $\Omega \setminus \bar{\omega}$  and for all  $x \in \partial\Omega \cup \partial\omega$ , it has the classical boundary representation:

$$(53) \quad \begin{aligned} \frac{1}{2}v(x) &= \int_{\partial\Omega} \partial_{\mathbf{n}} \Gamma(x, y) v(y) - \int_{\partial\omega} \partial_{\mathbf{n}} \Gamma(x, y) v(y) \\ &\quad - \int_{\partial\Omega} \Gamma(x, y) \partial_{\mathbf{n}} v(y) + \int_{\partial\omega} \Gamma(x, y) \partial_{\mathbf{n}} v(y). \end{aligned}$$

Similarly since  $v$  is harmonic in  $\omega$ , for all  $x \in \partial\omega$  we can write

$$(54) \quad \frac{1}{2}v(x) = \int_{\partial\omega} \partial_{\mathbf{n}} \Gamma(x, y) v(y) - \int_{\partial\omega} \Gamma(x, y) \partial_{\mathbf{n}} v(y).$$

Let us denote by  $v_d$  the solution of the boundary value problem (D) in (51). Let us show how to compute their restrictions and also their normal vector derivatives on the boundaries. Incorporating the jump conditions, a straightforward computation leads to the boundary integral equations

$$(55) \quad \begin{aligned} &\begin{bmatrix} \frac{1}{2}I + \mu K_\omega & \frac{\sigma_1}{\sigma_2 + \sigma_1} S_{\partial\Omega\partial\omega} \\ \mu K_{\partial\omega\partial\Omega} & \frac{\sigma_1}{\sigma_2 + \sigma_1} S_\Omega \end{bmatrix} \begin{bmatrix} (v_d^+)_{|\partial\omega} \\ (\partial_{\mathbf{n}} v_d)_{|\partial\Omega} \end{bmatrix} \\ &= \frac{1}{\sigma_1 + \sigma_2} \begin{bmatrix} \sigma_2 \left( \frac{1}{2}I - K_\omega \right) & S_\omega \\ -\sigma_2 K_{\partial\omega\partial\Omega} & S_{\partial\omega\partial\Omega} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \frac{\sigma_1}{\sigma_1 + \sigma_2} \begin{bmatrix} K_{\partial\Omega\partial\omega} f_1 \\ \left( -\frac{1}{2} + K_\Omega \right) f_1 \end{bmatrix}, \end{aligned}$$

where  $\mu = [\sigma]/(\sigma_1 + \sigma_2)$ . Thanks to (54), the quantity  $(\partial_{\mathbf{n}} v_d)^+$  is then given by

$$S_\omega (\partial_{\mathbf{n}} v_d)_{|\partial\omega}^+ = \frac{\sigma_2}{\sigma_1} \left( -\frac{1}{2}I + K_\omega \right) \left( v_d^+(x)_{|\partial\omega} - \alpha \right) + \frac{1}{\sigma_1} S_\omega \beta.$$

Concerning  $v_n$ , the solution of the Neumann problem (N) in (51), the same kind of

computation gives

$$(56) \quad \begin{bmatrix} \frac{1}{2}I + \mu K_\omega & -\frac{\sigma_1}{\sigma_2 + \sigma_1} K_{\partial\Omega\partial\omega} \\ \mu K_{\partial\omega\partial\Omega} & -\frac{\sigma_1}{\sigma_2 + \sigma_1} \left( -\frac{1}{2}I + K_\Omega \right) \end{bmatrix} \begin{bmatrix} (v_n^+)_{|\partial\omega} \\ (v_n)_{|\partial\Omega} \end{bmatrix} \\ = \frac{1}{\sigma_1 + \sigma_2} \begin{bmatrix} \sigma_2 \left( \frac{1}{2}I - K_\omega \right) & S_\omega \\ -\sigma_2 K_{\partial\omega\partial\Omega} & S_{\partial\omega\partial\Omega} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} - \frac{\sigma_1}{\sigma_1 + \sigma_2} \begin{bmatrix} S_{\partial\Omega\partial\omega} g_1 \\ S_\Omega g_1 \end{bmatrix}.$$

Finally, the computation of  $(\partial_{\mathbf{n}} v_n)_{|\partial\omega}^+$  is given by

$$S_\omega (\partial_{\mathbf{n}} v_n)_{|\partial\omega}^+ = \frac{\sigma_2}{\sigma_1} \left( -\frac{1}{2}I + K_\omega \right) (v_n^+(x)_{|\partial\omega} - \alpha) + \frac{1}{\sigma_1} S_\omega \beta.$$

Concerning the well-posedness of (55), we can state the following result.

**THEOREM 7.** *The linear system of integral equations (55) has a unique solution in  $H^{1/2}(\partial\omega) \times H^{-1/2}(\partial\Omega)$ .*

*Proof.* Let  $A$  be the matrix operator defined on  $H^{1/2}(\partial\omega) \times H^{-1/2}(\partial\Omega)$  as

$$(57) \quad A = \begin{bmatrix} \frac{1}{2}I + \mu K_\omega & \frac{\sigma_1}{\sigma_2 + \sigma_1} S_{\partial\Omega\partial\omega} \\ \mu K_{\partial\omega\partial\Omega} & \frac{\sigma_1}{\sigma_1 + \sigma_2} S_\Omega \end{bmatrix}.$$

The main argument of the proof is based on the Fredholm alternative. We have to show that the adjoint operator  $A^*$  is injective. Since the boundaries are bounded, the adjoint operator  $A^*$  defined on  $H^{-1/2}(\partial\omega) \times H^{1/2}(\partial\Omega)$  can be written in the form

$$(58) \quad A^* = \begin{bmatrix} \frac{1}{2}I + \mu K_\omega^* & \mu K_{\partial\Omega\partial\omega}^* \\ \frac{\sigma_1}{\sigma_2 + \sigma_1} S_{\partial\omega\partial\Omega} & \frac{\sigma_1}{\sigma_2 + \sigma_1} S_\Omega \end{bmatrix}.$$

Let  $(u, v) \in H^{-1/2}(\partial\omega) \times H^{1/2}(\partial\Omega)$  be in the kernel of  $A^*$ . Consider the potential  $W$  defined for each  $x \in \mathbb{R}^d$  by

$$(59) \quad W(x) = \frac{\sigma_1}{\sigma_2 + \sigma_1} \left( \int_{\partial\omega} \Gamma(x, y) u(y) + \int_{\partial\Omega} \Gamma(x, y) v(y) \right).$$

In the first step, we show that  $W = 0$ . The function  $W$  satisfies  $\Delta W = 0$  in  $\mathbb{R}^d \setminus (\partial\omega \cup \partial\Omega)$  by construction. We check that  $W|_{\partial\Omega} = 0$  from the equation corresponding to the second line of  $A^*$ . By the properties of the single-layer potential,  $[W] = 0$  on  $\partial\omega$ . Furthermore, it holds that  $[\sigma \partial_{\mathbf{n}} W] = 0$  on  $\partial\omega$ . Indeed, we can have [14]

$$\partial_{\mathbf{n}} W^+ = \frac{\sigma_1}{\sigma_1 + \sigma_2} \left( \left( \frac{1}{2} + K_\omega^* \right) u + K_{\partial\Omega\partial\omega}^* v \right)$$



and

$$\partial_{\mathbf{n}} W^- = \frac{\sigma_1}{\sigma_1 + \sigma_2} \left( \left( -\frac{1}{2} + K_{\omega}^* \right) v + K_{\partial\Omega\partial\omega}^* v \right),$$

and hence

$$\sigma_1 \partial_{\mathbf{n}} W^+ - \sigma_2 \partial_{\mathbf{n}} W^- = \sigma_1 \left( \left( \frac{1}{2} I + \mu K_{\omega}^* \right) u + \mu K_{\partial\Omega\partial\omega}^* v \right).$$

This corresponds to the first line of  $A^*(u, v)$ . Then  $W$  solves the Laplace equation (1) with homogeneous Dirichlet boundary conditions. By the uniqueness of the solution, we get  $W = 0$  in  $\Omega$ .

In the second step, we deduce that  $u = v = 0$ . Since  $W = 0$  in  $\Omega$ , we see that  $[\partial_{\mathbf{n}} W] = 0$  on  $\partial\omega$ . Since  $[\partial_{\mathbf{n}} W] = \sigma_1 u / (\sigma_1 + \sigma_2)$  on  $\partial\omega$ , we deduce  $u = 0$ . From the second line of  $A^*(u, v) = 0$ , we see that  $S_{\Omega} v = 0$  on  $\partial\Omega$ . Since the single-layer potential operator  $S_{\Omega} : H^{-1/2}(\partial\Omega) \mapsto H^{1/2}(\partial\Omega)$  is an isomorphism,  $v = 0$  holds. The injectivity of  $A^*$  is proved. Since  $2A = I + C$ , where  $C$  is a compact operator, we conclude that  $A$  has a continuous inverse thanks to the Fredholm alternative.  $\square$

In a similar way, the problem (56) is well-posed under some additional assumptions. We define the adequate space

$$H_{\diamond}^{1/2}(\partial\Omega) = \left\{ \phi \in H^{1/2}(\partial\Omega) : \int_{\partial\Omega} \phi = 0 \right\}.$$

We can state the following result.

**THEOREM 8.** *If we impose the normalizing condition*

$$\int_{\partial\Omega} v_n = \int_{\partial\Omega} f_1,$$

*then there exists one unique couple  $((v_n)_{|\partial\omega}, (v_n)_{|\partial\Omega}) \in H^{1/2}(\partial\omega) \times H_{\diamond}^{1/2}(\partial\Omega)$  solution of (56).*

*Proof.* Set

$$(60) \quad B = \begin{bmatrix} \frac{1}{2}I + \mu K_{\omega} & -\frac{\sigma_1}{\sigma_2 + \sigma_1} K_{\partial\Omega\partial\omega} \\ \mu K_{\partial\omega\partial\Omega} & -\frac{\sigma_1}{\sigma_1 + \sigma_2} \left( -\frac{1}{2}I + K_{\Omega} \right) \end{bmatrix}$$

as the operator defined on  $H^{1/2}(\partial\omega) \times H_{\diamond}^{1/2}(\partial\Omega)$ . The adjoint  $B^*$  can be written in the form

$$(61) \quad B^* = \begin{bmatrix} \frac{1}{2}I + \mu K_{\omega}^* & \mu K_{\partial\Omega\partial\omega}^* \\ -\frac{\sigma_1}{\sigma_1 + \sigma_2} K_{\partial\omega\partial\Omega}^* & -\frac{\sigma_1}{\sigma_1 + \sigma_2} \left( -\frac{1}{2}I + K_{\Omega}^* \right) \end{bmatrix}.$$

We begin to show that  $B^*$  is injective. Let  $(u, v) \in H^{1/2}(\partial\omega) \times H^{1/2}(\partial\Omega)$  be in the kernel of  $B^*$ . We introduce the potential

$$Z(x) = -\frac{\sigma_1}{\sigma_1 + \sigma_2} \left( \int_{\partial\omega} \Gamma(x, y)u(y) + \int_{\partial\Omega} \Gamma(x, y)v(y) \right), \quad x \in \mathbb{R}^d.$$

We can see that  $Z$  is a harmonic function in  $\mathbb{R}^d \setminus (\partial\omega \cup \partial\Omega)$ , satisfying  $\partial_n Z = 0$  on  $\partial\Omega$ . By the properties of the single-layer potential,  $[Z] = 0$ . Furthermore, a straightforward calculation shows that  $[\sigma \partial_n Z] = 0$  on  $\partial\omega$ . Hence,  $Z$  solves the boundary value problem

$$-\operatorname{div}(\sigma \nabla Z) = 0 \text{ in } \Omega,$$

$$\partial_n Z = 0 \text{ on } \partial\Omega.$$

The function is therefore constant in  $\Omega$ . Writing  $[\partial_n Z] = 0$  on  $\partial\omega$ , we get easily  $u = 0$  and then  $(-\frac{1}{2} + K_\Omega^*)v = 0$ . Since the operator  $\lambda I - K_\Omega^*$  is one-to-one on  $H_\diamond^{1/2}(\partial\Omega)$ , we deduce that  $v = 0$ . We conclude the proof thanks to the Fredholm alternative.  $\square$

*Tangential regularity results.* In order to give a sense to the jump conditions arising in (6)–(8), we need to work in the space of functions of higher regularity. We choose the framework of Hölder spaces. We quote [12] to make the behavior of the layer potentials on these spaces precise.

THEOREM 9 (Kirsch [12]).

1. If  $\partial\omega$  is of class  $C^{2,\alpha}$ ,  $0 < \alpha < 1$ , then the operators  $S_\omega$  and  $K_\omega$  map  $C^\beta(\partial\omega)$  continuously into  $C^{1,\beta}$  for all  $0 < \beta \leq \alpha$ .
2. Let  $k \in \mathbb{N}$  with  $k \neq 0$ . If  $\partial\omega$  is of class  $C^{k+1,\alpha}$  with  $0 < \alpha < 1$ , then the operators  $S_\omega$  and  $K_\omega$  map  $C^{k,\beta}(\partial\omega)$  continuously into  $C^{k+1,\beta}(\partial\omega)$  for all  $0 < \beta \leq \alpha$ .
3. Let  $k$  be an integer. If  $\partial\omega$  is of class  $C^{k+2,\alpha}$ , then  $K_\omega^*$  maps  $C^{k,\beta}$  continuously into  $C^{k+1,\beta}(\partial\omega)$  for all  $0 < \beta \leq \alpha$ .

PROPOSITION 7. Assume that  $\partial\omega$  is of class  $C^{4,\alpha}$ . Then the trace of  $u_d$  solution of the boundary value problem (1), (2) belongs to  $C^{3,\alpha}(\partial\omega, \mathbb{R})$ , while  $\partial_n u_d \in C^{2,\alpha}(\partial\omega, \mathbb{R})$ .

*Proof.* We consider now the particular case where both  $\alpha$  and  $\beta$  are the zero function and  $(f_1, g_1) = (f, g)$ , where  $f$  and  $g$  are, respectively, the Dirichlet and Neumann boundary data. To recover the tangential regularity of the solution  $u$  along  $\partial\omega$ , we look at the first line of (55) to deduce that

$$(62) \quad \left[ \frac{1}{2}I + \mu K_\omega \right] (u_d)|_{\partial\omega} = -\frac{\sigma_1}{\sigma_2 + \sigma_1} S_{\partial\Omega\partial\omega} \partial_n u_d|_{\partial\Omega} + \frac{\sigma_1}{\sigma_1 + \sigma_2} K_{\partial\Omega\partial\omega} f$$

and

$$(63) \quad S_\omega(\partial_n u_d)|_{\partial\omega}^+ = \frac{\sigma_2}{\sigma_1} \left( -\frac{1}{2}I + K_\omega \right) u_d^+(x)|_{\partial\omega}.$$

It is easy to deduce that  $(u_d)|_{\partial\omega} \in C^{3,\alpha}(\partial\omega)$ . Indeed, from (62), which we consider as an equation in  $(u_d)|_{\partial\omega}$  with data  $f$  and  $(\partial_n u_d)|_{\partial\Omega} = g$ , we see that  $(f, (\partial_n u_d)|_{\partial\Omega})$  belongs to  $H^{1/2}(\partial\Omega) \times H^{-1/2}(\partial\Omega)$ , thanks to Theorem 7.

Since the two boundaries have no intersection point and since  $\partial\omega$  is of class  $C^{4,\alpha}$ , it follows that the right-hand side of the former equation is of class  $C^{3,\alpha}(\partial\omega)$ . We then conclude that the solution of (62) will be of class  $C^{3,\alpha}$  since the operator  $1/2I + \mu K_\omega$  is an isomorphism from  $C^{3,\alpha}(\partial\omega)$  into itself. With the same arguments, we show straightforwardly that  $(\partial_n u_n)|_{\partial\omega}^+ \in C^{2,\alpha}$ .  $\square$

**About the regularity of the jumps of the second derivative.** The equations giving the jump conditions  $[u'_d]$  and  $[\partial u'_d]$  show obviously that  $[u'_d]$  and  $[\partial_n u'_d]$  belong, respectively, to  $C^{2,\alpha}(\partial\omega)$  and  $C^{1,\alpha}(\partial\omega)$ . Hence, it comes straightforwardly that  $[u''_d] \in \mathcal{C}^{1,\alpha}$ . With the same arguments, we show that  $[\partial_n u''_d] \in \mathcal{C}^{0,\alpha}$  (see [22] for more details) and then that all the formal computations to get the equations describing the second derivative make sense.

**Remark on the interest of (51) for numerical schemes.** In view of a numerical discretization of the state equation, one has to emphasize that the choice of a finite elements method seems inappropriate: one should extract a tangential derivative of high order on the interface  $\partial\omega$ . The obtained numerical accuracy is not sufficient to incorporate the results in an optimization scheme. On the contrary, the systems of boundary integral equations (55) and (56) are well suited for this kind of computation. Any numerical discretization needs to compute the state, its derivatives with respect to the shape, and the normal derivatives along the interface  $\partial\omega$ . This can be done for both  $u_d$  and  $u'_d$  with just a change in the right-hand side of the same system. Nevertheless, a discussion of adapted schemes should be precise and is out of the scope of this manuscript.

**6.2. The bidimensional case.** The results obtained in the three-dimensional case can be extended to the bidimensional up to the additional assumption:

(H) *The diameter of the outer domain  $\Omega$  is strictly less than 1.*

In fact, the proofs extends to the bidimensional case on every step except the following fact: for a bidimensional domain  $D$ , the single-layer potential  $S_D$  is not in general an isomorphism from  $H^s(\partial D)$  into  $H^{s+1}(\partial D)$  for  $s \in \mathbb{R}$ . However, Hsiao and Wendland have shown in [10] that, provided that  $\text{diam}(D) < 1$ , the single-layer potential is an isomorphism from  $H^s(\partial D)$  into  $H^{s+1}(\partial D)$  for all  $s \in \mathbb{R}$ . In the previously presented results, we used the inverse of  $S_\Omega$  in the proof of Theorem 5, and the inverse of  $S_\omega$  in the proof of Theorem 7, and assumption (H) ensures that  $S_\Omega$  and  $S_\omega$  are both invertible.

The diverse notions of curvatures coincide in dimension two. Hence the expression of the second order derivative is simpler in dimension two. For example, the state  $u_d$  has a second order shape derivative  $u''_d \in H^1(\Omega \setminus \bar{\omega}) \cup H^1(\omega)$  that solves

(64)

$$\left\{ \begin{array}{l} \Delta u''_d = 0 \text{ in } \Omega \setminus \bar{\omega} \text{ and in } \omega, \\ [u''_d] = \left( (h_{1,n}h_{2,n} - h_{1,\tau}h_{2,\tau})H \right) [\partial_n u_d] - (h_{1,n}[\partial_n(u_d)'_2] + h_{2,n}[\partial_n(u_d)'_1]) \\ \quad + (\mathbf{h}_{1\tau} \cdot \nabla h_{2,n} + \mathbf{h}_{2\tau} \cdot \nabla h_{1,n}) [\partial_n u_d] \text{ on } \partial\omega, \\ [\sigma \partial_n u''_d] = \text{div}_\tau \left( h_{2,n} [\sigma \nabla_\tau(u_d)'_1] + h_{1,n} [\sigma \nabla_\tau(u_d)'_2] \right. \\ \quad \left. + (h_{1,\tau}h_{2,\tau} - h_{1,n}h_{2,n})H[\sigma \nabla_\tau u_d] \right) \\ \quad - \text{div}_\tau \left( (\mathbf{h}_{1\tau} \cdot \nabla_\tau h_{2,n} + \nabla_\tau h_{1,n} \cdot \mathbf{h}_{2\tau}) [\sigma \nabla_\tau u_d] \right) \text{ on } \partial\omega, \\ u''_d = 0 \text{ on } \partial\Omega. \end{array} \right.$$

**6.3. Another derivation of (8).** As usual in the derivation of boundary conditions satisfied by shape derivatives, two strategies are at hand. First, one can differentiate the boundary conditions satisfied by the state equation. This choice, presented in section 4, is the shortest solution but requires dealing with tangential derivations. One can alternatively deal with the weak form of the state equation

and perform as many integrations by parts as required to transform the volume integral into surface integrals that can be seen as boundary conditions. This second method will be presented now. It uses more elementary mathematical tools but requires lengthy computations. These computations are presented here only for the completeness of the presentation.

We want to make explicit the problem solved by  $(u')'$ . To achieve this, we should write the right-hand side of (32),

$$F = \int_{\Omega} \sigma \left[ \nabla \dot{u}_1 \cdot A_{\mathbf{h}_2} \nabla v + \nabla \dot{u}_2 \cdot A_{\mathbf{h}_1} \nabla v - \nabla u \cdot \mathfrak{A} \nabla v \right],$$

as the sum of an integral with  $\nabla v$  in factor and an integral of a divergence to identify the jump conditions on  $\partial\omega$ . To that end, we will use algebraic identities that involve second order derivatives of  $u, \dot{u}_i$  and of the test function  $v \in \mathcal{D}(\Omega)$ . Using Lemma 1, we obtain

$$\begin{aligned} \int_{\Omega} \sigma \nabla \dot{u}_1 \cdot A_{\mathbf{h}_2} \nabla v &= \int_{\Omega} \sigma \left[ \nabla (\mathbf{h}_2 \cdot \nabla \dot{u}_1) \cdot \nabla v + \nabla (\mathbf{h}_2 \cdot \nabla v) \cdot \nabla \dot{u}_1 - \operatorname{div} ((\nabla \dot{u}_1 \cdot \nabla v) \mathbf{h}_2) \right], \\ \int_{\Omega} \sigma \nabla \dot{u}_2 \cdot A_{\mathbf{h}_1} \nabla v &= \int_{\Omega} \sigma \left[ \nabla (\mathbf{h}_1 \cdot \nabla \dot{u}_2) \cdot \nabla v + \nabla (\mathbf{h}_1 \cdot \nabla v) \cdot \nabla \dot{u}_2 - \operatorname{div} ((\nabla \dot{u}_2 \cdot \nabla v) \mathbf{h}_1) \right]. \end{aligned}$$

Concerning the remaining terms, we use Lemma 2 to get

$$\begin{aligned} \int_{\Omega} \sigma \nabla u \cdot \mathfrak{A} \nabla v &= \int_{\Omega} \sigma \operatorname{div} ((\mathbf{h}_2 \cdot \nabla u) A_{\mathbf{h}_1} \nabla v + (\mathbf{h}_2 \cdot \nabla v) A_{\mathbf{h}_1} \nabla u - (A_{\mathbf{h}_1} \nabla u \cdot \nabla v) \mathbf{h}_2) \\ &\quad - \sigma \left[ (\mathbf{h}_2 \cdot \nabla u) \operatorname{div} (A_{\mathbf{h}_1} \nabla v) + (\mathbf{h}_2 \cdot \nabla v) \operatorname{div} (A_{\mathbf{h}_1} \nabla u) \right]. \end{aligned}$$

We apply Lemma 3 and gather the expressions obtained for  $F$ .

$$\begin{aligned} (65) \quad F &= \int_{\Omega} \sigma \left[ \nabla (\mathbf{h}_1 \cdot \nabla \dot{u}_2 + \mathbf{h}_2 \cdot \nabla \dot{u}_1) \cdot \nabla v + \nabla (\mathbf{h}_2 \cdot \nabla v) \cdot \nabla \dot{u}_1 + \nabla (\mathbf{h}_1 \cdot \nabla v) \cdot \nabla \dot{u}_2 \right] \\ &\quad + \int_{\Omega} \sigma \operatorname{div} ((A_{\mathbf{h}_1} \nabla u \cdot \nabla v - \nabla \dot{u}_1 \cdot \nabla v) \mathbf{h}_2 - (\nabla \dot{u}_2 \cdot \nabla v) \mathbf{h}_1) \\ &\quad + \int_{\Omega} \sigma \left[ (\mathbf{h}_2 \cdot \nabla v) \Delta (\mathbf{h}_1 \cdot \nabla u) - \operatorname{div} ((\mathbf{h}_2 \cdot \nabla v) A_{\mathbf{h}_1} \nabla u) - \nabla (\mathbf{h}_2 \cdot \nabla u) \cdot A_{\mathbf{h}_1} \nabla v \right]. \end{aligned}$$

Using (25), we remove the dependency on  $A_{\mathbf{h}_1} \nabla v$ :

$$\begin{aligned} \nabla (\mathbf{h}_2 \cdot \nabla u) \cdot A_{\mathbf{h}_1} \nabla v &= \nabla (\mathbf{h}_1 \cdot \nabla (\mathbf{h}_2 \cdot \nabla u)) \cdot \nabla v \\ &\quad + \nabla (\mathbf{h}_1 \cdot \nabla v) \cdot \nabla (\mathbf{h}_2 \cdot \nabla u) - \operatorname{div} ((\nabla (\mathbf{h}_2 \cdot \nabla u) \cdot \nabla v) \mathbf{h}_1). \end{aligned}$$

Therefore, we write  $F = F_1 + F_2$ , where

(66)

$$\begin{aligned} F_1 &= \int_{\Omega} \sigma \left[ \nabla (\mathbf{h}_1 \cdot \nabla \dot{u}_2 + \mathbf{h}_2 \cdot \nabla \dot{u}_1) - \nabla (\mathbf{h}_1 \cdot \nabla (\mathbf{h}_2 \cdot \nabla u)) \right] \cdot \nabla v, \\ F_2 &= \int_{\Omega} \sigma \left[ \nabla (\mathbf{h}_1 \cdot \nabla v) \cdot \nabla (\dot{u}_2 - \mathbf{h}_2 \cdot \nabla u) + \nabla (\mathbf{h}_2 \cdot \nabla v) \cdot \nabla \dot{u}_1 + (\mathbf{h}_2 \cdot \nabla v) \Delta (\mathbf{h}_1 \cdot \nabla u) \right] \\ &\quad + \int_{\Omega} \sigma \operatorname{div} \left( (A_{\mathbf{h}_1} \nabla u \cdot \nabla v - \nabla \dot{u}_1 \cdot \nabla v) \mathbf{h}_2 \right. \\ &\quad \left. + (\nabla (\mathbf{h}_2 \cdot \nabla u) \cdot \nabla v - \nabla \dot{u}_2 \cdot \nabla v) \mathbf{h}_1 - (\mathbf{h}_2 \cdot \nabla v) A_{\mathbf{h}_1} \nabla u \right). \end{aligned}$$

The connection between second order material and shape derivatives is given by

$$\ddot{u}_1 = (u'_1)'_2 + \mathbf{h}_1 \cdot \nabla \dot{u}_2 + \mathbf{h}_2 \cdot \nabla \dot{u}_1 - \mathbf{h}_1 \cdot \nabla (\mathbf{h}_2 \cdot \nabla u).$$

Incorporating this expression into (66), we rewrite (32) as

$$(67) \quad \forall v \in H_0^1(\Omega), \quad \int_{\Omega} \sigma \nabla (u'_1)'_2 \cdot \nabla v = F_2.$$

Testing it against  $v \in \mathcal{D}(\Omega \setminus \partial\omega)$ , we get  $\Delta (u'_1)'_2 = 0$  in  $\Omega \setminus \overline{\omega}$  and in  $\omega$ . We now deduce the jump conditions for  $(u'_1)'_2$ . To obtain the jump of the potential, we simply write that  $\ddot{u}_1 \in H_0^1(\Omega)$ , hence  $[\ddot{u}_1] = 0$  on  $\partial\omega$ , and then

$$(68) \quad [(u'_1)'_2] = -\mathbf{h}_1 \cdot [\nabla u'_2] - \mathbf{h}_2 \cdot [\nabla \dot{u}_1].$$

To express the jump of the flux, we then apply the Gauss formula in (67) to get

$$(69) \quad - \int_{\partial\omega} [\sigma \partial_{\mathbf{n}} (u'_1)'_2] v = F_2.$$

The second term  $F_2$  contains all the jumps of the flux on the interface  $\partial\omega$ .

**A simplified expression of  $F_2$ .** To get a simplified formula for  $F_2$  under a boundary integral, some lengthy but straightforward calculations are needed. We summarize the result by means of the following lemma.

LEMMA 5. *One has*

(70)

$$\begin{aligned} F_2 &= \int_{\partial\omega} \operatorname{div}_{\tau} (2h_{2,n} h_{1,n} D\mathbf{n} [\sigma \nabla_{\tau} u] - h_{2,n} \mathbf{n} \cdot \nabla h_{1,n} [\sigma \nabla_{\tau} u] + h_{2,n} \mathbf{h}_{1\tau} \cdot D\mathbf{n} \mathbf{n} [\sigma \nabla_{\tau} u]) v \\ &\quad + \int_{\partial\omega} \operatorname{div}_{\tau} (\mathbf{h}_{1\tau} \cdot \nabla_{\tau} (h_{2,n}) [\sigma \nabla_{\tau} u] - h_{1,n} h_{2,n} H [\sigma \nabla_{\tau} u]) v \\ &\quad - \int_{\partial\omega} \left( \operatorname{div}_{\tau} (h_{2,n} [\sigma \nabla_{\tau} u'_1]) + \operatorname{div}_{\tau} (h_{1,n} [\sigma \nabla_{\tau} u'_2]) \right) v. \end{aligned}$$

*Proof.* First, write

$$\begin{aligned} \int_{\Omega} \sigma \nabla (\mathbf{h}_1 \cdot \nabla v) \cdot \nabla (\dot{u}_2 - \mathbf{h}_2 \cdot \nabla u) &= \sigma_1 \int_{\Omega \setminus \overline{\omega}} \nabla (\mathbf{h}_1 \cdot \nabla v) \cdot \nabla u'_2 + \sigma_2 \int_{\omega} \nabla (\mathbf{h}_1 \cdot \nabla v) \cdot \nabla u'_2 \\ &= - \int_{\partial\omega} [\sigma \partial_{\mathbf{n}} u'_2] (\mathbf{h}_1 \cdot \nabla v). \end{aligned}$$

Note that the normal vector is oriented from  $\omega$  to  $\Omega \setminus \bar{\omega}$ . In the same spirit, we write

$$\begin{aligned} \nabla(\mathbf{h}_2 \cdot \nabla v) \cdot \nabla \dot{u}_1 + (\mathbf{h}_2 \cdot \nabla v) \Delta(\mathbf{h}_1 \cdot \nabla u) &= \nabla(\mathbf{h}_2 \cdot \nabla v) \cdot \nabla(\dot{u}_1 - \mathbf{h}_1 \cdot \nabla u) \\ &\quad + \operatorname{div}((\mathbf{h}_2 \cdot \nabla v) \cdot \nabla(\mathbf{h}_1 \cdot \nabla u)). \end{aligned}$$

By an argument of symmetry, we then can write

$$\int_{\Omega} \sigma \nabla(\mathbf{h}_2 \cdot \nabla v) \cdot \nabla(\dot{u}_1 - \mathbf{h}_1 \cdot \nabla u) = - \int_{\partial\omega} [\sigma \partial_{\mathbf{n}} u'_1](\mathbf{h}_2 \cdot \nabla v).$$

To drop the dependency in  $A_{\mathbf{h}_1}$ , we use (25) and get after expansion

$$\begin{aligned} \operatorname{div}((A_{\mathbf{h}_1} \nabla u \cdot \nabla v) \mathbf{h}_2) &= \operatorname{div}((\nabla(\mathbf{h}_1 \cdot \nabla v) \cdot \nabla u + \nabla(\mathbf{h}_1 \cdot \nabla u) \nabla v) \mathbf{h}_2) \\ &\quad - \operatorname{div}(\operatorname{div}((\nabla u \cdot \nabla v) \mathbf{h}_1) \mathbf{h}_2), \\ \operatorname{div}((\mathbf{h}_2 \cdot \nabla v) A_{\mathbf{h}_1} \nabla u) &= \nabla(\mathbf{h}_2 \cdot \nabla v) \cdot A_{\mathbf{h}_1} \nabla u + (\mathbf{h}_2 \cdot \nabla v) \operatorname{div}(A_{\mathbf{h}_1} \nabla u) \\ &= \nabla(\mathbf{h}_1 \cdot \nabla(\mathbf{h}_2 \cdot \nabla v)) \cdot \nabla u + \nabla(\mathbf{h}_1 \cdot \nabla u) \nabla(\mathbf{h}_2 \cdot \nabla v) \\ &\quad + (\mathbf{h}_2 \cdot \nabla v) \Delta(\mathbf{h}_1 \cdot \nabla u) - \operatorname{div}((\nabla(\mathbf{h}_2 \cdot \nabla v) \cdot \nabla u) \mathbf{h}_1) \\ &= \nabla(\mathbf{h}_1 \cdot \nabla(\mathbf{h}_2 \cdot \nabla v)) \cdot \nabla u \\ &\quad + \operatorname{div}((\mathbf{h}_2 \cdot \nabla v) \nabla(\mathbf{h}_1 \cdot \nabla u) - (\nabla(\mathbf{h}_2 \cdot \nabla v) \cdot \nabla u) \mathbf{h}_1). \end{aligned}$$

After integrating by parts, we conclude, thanks to the state equation, with

$$\int_{\Omega} \sigma \nabla(\mathbf{h}_1 \cdot \nabla(\mathbf{h}_2 \cdot \nabla v)) \cdot \nabla u = - \int_{\Omega} (\mathbf{h}_1 \cdot \nabla(\mathbf{h}_2 \cdot \nabla v)) \operatorname{div}(\sigma \nabla u) = 0.$$

We substitute the shape derivative  $u'$  to the material derivative  $\dot{u}$ :

$$\begin{aligned} F_2 &= - \int_{\partial\omega} [\sigma \partial_{\mathbf{n}} u'_1](\mathbf{h}_2 \cdot \nabla v) + [\sigma \partial_{\mathbf{n}} u'_2](\mathbf{h}_1 \cdot \nabla v) - \int_{\Omega} \sigma \operatorname{div}(\operatorname{div}((\nabla u \cdot \nabla v) \mathbf{h}_1) \mathbf{h}_2) \\ &\quad + \int_{\Omega} \sigma \operatorname{div}\left(\left((\nabla(\mathbf{h}_1 \cdot \nabla v) \cdot \nabla u) \mathbf{h}_2 + (\nabla(\mathbf{h}_2 \cdot \nabla v) \cdot \nabla u) \mathbf{h}_1\right) \right. \\ &\quad \left. - ((\nabla u'_2 \cdot \nabla v) \mathbf{h}_1 + (\nabla u'_1 \cdot \nabla v) \mathbf{h}_2)\right). \end{aligned}$$

First, we use the continuity of the flux on  $\partial\omega$ , then we integrate by parts on  $\partial\omega$ , and finally we incorporate the expressions of the jumps of the shape derivatives  $u'$  to obtain

$$\begin{aligned} &\int_{\Omega} \sigma \operatorname{div}(\mathbf{h}_1 \cdot (\nabla(\mathbf{h}_2 \cdot \nabla v) \cdot \nabla u)) \\ &= - \int_{\partial\omega} [\sigma \nabla u \cdot \nabla(\mathbf{h}_2 \cdot \nabla v)] h_{1,n} = - \int_{\partial\omega} [\sigma \nabla_{\tau} u] h_{1,n} \nabla_{\tau}(\mathbf{h}_2 \cdot \nabla v) \\ &= \int_{\partial\omega} \operatorname{div}_{\tau}([\sigma \nabla_{\tau} u] h_{1,n}) \mathbf{h}_2 \cdot \nabla v = \int_{\partial\omega} [\sigma \partial_{\mathbf{n}} u'_1] \mathbf{h}_2 \cdot \nabla v. \end{aligned}$$

This leads to a simplified expression for  $F_2$ :

$$F_2 = - \int_{\Omega} \sigma \operatorname{div} \left( \operatorname{div} ((\nabla u, \nabla v) \mathbf{h}_1) \mathbf{h}_2 + \left( (\nabla u'_1, \nabla v) \mathbf{h}_2 + (\nabla u'_2, \nabla v) \mathbf{h}_1 \right) \right).$$

Let us study each term of this sum. Using the Gauss formula and integrating by parts on the manifold  $\partial\omega$ , we obtain

$$\begin{aligned} & \int_{\Omega} \sigma \operatorname{div} \left( (\nabla u'_1, \nabla v) \mathbf{h}_2 \right) \\ &= - \int_{\partial\omega} h_{2,n} [\sigma \nabla u'_1, \nabla v] = - \int_{\partial\omega} h_{2,n} [\sigma \partial_{\mathbf{n}} u'_1] \partial_{\mathbf{n}} v - \int_{\partial\omega} h_{2,n} [\sigma \nabla_{\tau} u'_1] \nabla_{\tau} v \\ &= - \int_{\partial\omega} h_{2,n} [\sigma \partial_{\mathbf{n}} u'_1] \partial_{\mathbf{n}} v + \int_{\partial\omega} \operatorname{div}_{\tau} \left( h_{2,n} [\sigma \nabla_{\tau} u'_1] \right) v. \end{aligned}$$

By symmetry, we also get

$$\int_{\Omega} \sigma \operatorname{div} \left( (\nabla u'_2, \nabla v) \mathbf{h}_1 \right) = - \int_{\partial\omega} h_{1,n} [\sigma \partial_{\mathbf{n}} u'_2] \partial_{\mathbf{n}} v + \int_{\partial\omega} \operatorname{div}_{\tau} \left( h_{1,n} [\sigma \nabla_{\tau} u'_2] \right) v.$$

We now turn to the term with a double divergence. We first write it as a boundary integral thanks to the Gauss formula:

$$\int_{\Omega} \sigma \operatorname{div} \left( \operatorname{div} ((\nabla u, \nabla v) \mathbf{h}_1) \mathbf{h}_2 \right) = - \int_{\partial\omega} h_{2,n} \operatorname{div} \left( \mathbf{h}_1 [\sigma (\nabla u, \nabla v)] \right).$$

Then we use (14) to introduce the tangential operators:

$$\begin{aligned} \int_{\Omega} \sigma \operatorname{div} \left( \operatorname{div} ((\nabla u, \nabla v) \mathbf{h}_1) \mathbf{h}_2 \right) &= - \int_{\partial\omega} h_{2,n} \operatorname{div}_{\tau} \left( \mathbf{h}_1 [\sigma (\nabla u, \nabla v)] \right) \\ &\quad - \int_{\partial\omega} h_{2,n} D \left( \mathbf{h}_1 [\sigma (\nabla u, \nabla v)] \right) \mathbf{n} \cdot \mathbf{n}. \end{aligned}$$

We study each of these terms, starting with the one involving tangential derivatives: we expand the tangential divergence to incorporate the jump relation for the state  $u$ :

$$\begin{aligned} \operatorname{div}_{\tau} \left( \mathbf{h}_1 [\sigma (\nabla u, \nabla v)] \right) &= \operatorname{div}_{\tau} (\mathbf{h}_1) [\sigma (\nabla u, \nabla v)] + \mathbf{h}_1 \cdot \nabla_{\tau} [\sigma \nabla u, \nabla v] \\ &= \operatorname{div}_{\tau} (\mathbf{h}_1) [\sigma \nabla_{\tau} u] \cdot \nabla_{\tau} v + \mathbf{h}_1 \cdot \nabla_{\tau} [\sigma \nabla_{\tau} u, \nabla_{\tau} v]. \end{aligned}$$

Then the first term becomes

$$\begin{aligned} \int_{\partial\omega} h_{2,n} \operatorname{div}_{\tau} \left( \mathbf{h}_1 [\sigma (\nabla u, \nabla v)] \right) &= \int_{\partial\omega} h_{2,n} \operatorname{div}_{\tau} (\mathbf{h}_1) [\sigma \nabla_{\tau} u] \cdot \nabla_{\tau} v \\ &\quad + \int_{\partial\omega} h_{2,n} \mathbf{h}_1 \cdot \nabla_{\tau} [\sigma \nabla_{\tau} u, \nabla_{\tau} v]. \end{aligned}$$

We use the integration by parts formula (18) to get

$$\begin{aligned} & \int_{\partial\omega} h_{2,n} \operatorname{div}_{\tau} \left( \mathbf{h}_1 [\sigma (\nabla u, \nabla v)] \right) \\ &= \int_{\partial\omega} h_{1,n} h_{2,n} H [\sigma \nabla_{\tau} u] \cdot \nabla_{\tau} v - \operatorname{div}_{\tau} \left( \operatorname{div}_{\tau} (\mathbf{h}_1) h_{2,n} [\sigma \nabla_{\tau} u] \right) v \\ &\quad - \operatorname{div}_{\tau} (\mathbf{h}_1 h_{2,n}) [\sigma \nabla_{\tau} u] \cdot \nabla_{\tau} v \\ &= \int_{\partial\omega} \operatorname{div}_{\tau} \left( \left( \operatorname{div}_{\tau} (\mathbf{h}_1 h_{2,n}) - \operatorname{div}_{\tau} (\mathbf{h}_1) h_{2,n} - h_{1,n} h_{2,n} H \right) [\sigma \nabla_{\tau} u] \right) v. \end{aligned}$$

Expanding

$$\begin{aligned} & \operatorname{div}_\tau \left( \operatorname{div}_\tau (\mathbf{h}_1 h_{2,n}) [\sigma \nabla_\tau u] \right) v \\ &= \operatorname{div}_\tau \left( \operatorname{div}_\tau (\mathbf{h}_1) h_{2,n} [\sigma \nabla_\tau u] + \mathbf{h}_1 \cdot \nabla_\tau (h_{2,n}) [\sigma \nabla_\tau u] \right) v \\ &= \operatorname{div}_\tau \left( \operatorname{div}_\tau (\mathbf{h}_1) h_{2,n} [\sigma \nabla_\tau u] \right) v + \operatorname{div}_\tau (\mathbf{h}_1 \cdot \nabla_\tau h_{2,n} [\sigma \nabla_\tau u]) v, \end{aligned}$$

we obtain the new expression:

$$(71) \quad \int_{\partial\omega} h_{2,n} \operatorname{div}_\tau \left( \mathbf{h}_1 [\sigma (\nabla u \cdot \nabla v)] \right) = \int_{\partial\omega} \operatorname{div}_\tau \left( (\mathbf{h}_1 \cdot \nabla_\tau h_{2,n} - h_{1,n} h_{2,n} H) [\sigma \nabla_\tau u] \right) v.$$

Now, we consider the term involving normal components. We have

$$(72) \quad \begin{aligned} \mathbf{n} \cdot D(\mathbf{h}_1 [\sigma (\nabla u \cdot \nabla v)]) \mathbf{n} &= \mathbf{n} \cdot \nabla (h_{1,n} [\sigma \nabla u \cdot \nabla v]) - [\sigma \nabla u \cdot \nabla v] \mathbf{h}_{1,\tau} \cdot D\mathbf{n} \mathbf{n} \\ &= \mathbf{n} \cdot \nabla (h_{1,n}) [\sigma \nabla_\tau u] \nabla_\tau v + h_{1,n} \mathbf{n} \cdot \nabla ([\sigma \nabla u \cdot \nabla v]). \end{aligned}$$

Then we get

$$\begin{aligned} & \int_{\partial\omega} h_{2,n} D(\mathbf{h}_1 [\sigma (\nabla u \cdot \nabla v)]) \mathbf{n} \cdot \mathbf{n} \\ &= \int_{\partial\omega} h_{2,n} \mathbf{n} \cdot \nabla (h_{1,n}) [\sigma \nabla_\tau u] \cdot \nabla_\tau v + h_{2,n} h_{1,n} \mathbf{n} \cdot \nabla ([\sigma \nabla u \cdot \nabla v]) \\ &= \int_{\partial\omega} -\operatorname{div}_\tau (h_{2,n} \mathbf{n} \cdot \nabla (h_{1,n}) [\sigma \nabla_\tau u]) v + h_{2,n} h_{1,n} \mathbf{n} \cdot \nabla ([\sigma \nabla u \cdot \nabla v]). \end{aligned}$$

A straightforward calculus leads to

$$\begin{aligned} \mathbf{n} \cdot \nabla ([\sigma \nabla u \cdot \nabla v]) &= \mathbf{n} \cdot \left( [\sigma D^2 u \nabla v] + D^2 v [\sigma \nabla u] \right) \\ &= \mathbf{n} \cdot \left( \partial_{\mathbf{n}} v [\sigma D^2 u] \mathbf{n} + [\sigma D^2 u] \nabla_\tau v + D^2 v [\sigma \nabla_\tau u] \right) \\ &= \partial_{\mathbf{n}} v \left[ \sigma \frac{\partial^2 u}{\partial n^2} \right] + \mathbf{n} \cdot [\sigma D^2 u] \nabla_\tau v + \mathbf{n} \cdot D^2 v [\sigma \nabla_\tau u], \end{aligned}$$

where  $D^2 u$  is the Hessian matrix of  $u$ . From (20) and from the jump conditions for the state  $u$ , we deduce that

$$\left[ \sigma \frac{\partial^2 u}{\partial n^2} \right] = -[\sigma \Delta_\tau u].$$

When one differentiates the relation expressing the continuity of the flux for the state along the tangential direction  $\nabla_\tau v$ , one gets (see [7, p. 235])

$$0 = \nabla [\sigma \partial_{\mathbf{n}} u] \cdot \nabla_\tau v = [\sigma D^2 u] \nabla_\tau v \cdot \mathbf{n} + [\sigma \nabla u] \cdot (D\mathbf{n} \nabla_\tau v).$$

In the same spirit, it comes that

$$(73) \quad \nabla \partial_{\mathbf{n}} v \cdot [\sigma \nabla_\tau u] = D^2 v [\sigma \nabla_\tau u] \cdot \mathbf{n} + \nabla v \cdot (D\mathbf{n} [\sigma \nabla_\tau u]).$$



Since  $D\mathbf{n}$  is a symmetric matrix and  $D\mathbf{n}\mathbf{n} = 0$ , one checks  $\nabla v \cdot (D\mathbf{n}[\sigma\nabla_\tau u]) = [\sigma\nabla u] \cdot (D\mathbf{n}\nabla_\tau v)$ . Then

$$\mathbf{n} \cdot \nabla ([\sigma\nabla u \cdot \nabla v]) = -[\sigma\Delta_\tau u] \partial_{\mathbf{n}} v - 2D\mathbf{n}[\sigma\nabla_\tau u] \cdot \nabla_\tau v + [\sigma\nabla_\tau u] \nabla_\tau \partial_{\mathbf{n}} v.$$

We integrate this expression on  $\partial\omega$  and obtain after some integration by parts

$$\begin{aligned} & \int_{\partial\omega} h_{2,n} h_{1,n} \mathbf{n} \cdot \nabla ([\sigma\nabla u \cdot \nabla v]) \\ &= - \int_{\partial\omega} h_{2,n} h_{1,n} [\sigma\Delta_\tau u] \partial_{\mathbf{n}} v + \int_{\partial\omega} h_{2,n} h_{1,n} [\sigma\nabla_\tau u] \nabla_\tau \partial_{\mathbf{n}} v \\ & \quad - 2 \int_{\partial\omega} h_{2,n} h_{1,n} D\mathbf{n} [\sigma\nabla_\tau u] \cdot \nabla_\tau v, \\ &= - \int_{\partial\omega} \left[ h_{2,n} h_{1,n} [\sigma\Delta_\tau u] + \operatorname{div}_\tau (h_{2,n} h_{1,n} [\sigma\nabla_\tau u]) \right] \partial_{\mathbf{n}} v \\ & \quad + 2 \int_{\partial\omega} \operatorname{div}_\tau (h_{2,n} h_{1,n} D\mathbf{n} [\sigma\nabla_\tau u]) v. \end{aligned}$$

Hence

$$\begin{aligned} & \int_{\partial\omega} h_{2,n} D(\mathbf{h}_1 [\sigma(\nabla u \cdot \nabla v)]) \mathbf{n} \cdot \mathbf{n} \\ &= - \int_{\partial\omega} \left[ h_{2,n} h_{1,n} [\sigma\Delta_\tau u] + \operatorname{div}_\tau (h_{2,n} h_{1,n} [\sigma\nabla_\tau u]) \right] \partial_{\mathbf{n}} v \\ & \quad + \int_{\partial\omega} \operatorname{div}_\tau (2h_{2,n} h_{1,n} D\mathbf{n} [\sigma\nabla_\tau u] - h_{2,n} \mathbf{n} \cdot \nabla_\tau h_{1,n} [\sigma\nabla_\tau u]) v, \\ & \int_{\Omega} \sigma \operatorname{div} \left( \operatorname{div} ((\nabla u \cdot \nabla v) \mathbf{h}_1) \mathbf{h}_2 \right) \\ &= \int_{\partial\omega} \left[ h_{2,n} h_{1,n} [\sigma\Delta_\tau u] + \operatorname{div}_\tau (h_{2,n} h_{1,n} [\sigma\nabla_\tau u]) \right] \partial_{\mathbf{n}} v \\ & \quad - \int_{\partial\omega} \operatorname{div}_\tau (2h_{2,n} h_{1,n} D\mathbf{n} [\sigma\nabla_\tau u] - h_{2,n} \mathbf{n} \cdot \nabla h_{1,n} [\sigma\nabla_\tau u]) v \\ & \quad - \int_{\partial\omega} \operatorname{div}_\tau (\mathbf{h}_{1\tau} \cdot \nabla_\tau (h_{2,n}) [\sigma\nabla_\tau u] - h_{1,n} h_{2,n} H [\sigma\nabla_\tau u]) v. \end{aligned}$$

Gathering all the terms, we write  $F_2$  as

$$\begin{aligned} F_2 &= \int_{\partial\omega} \operatorname{div}_\tau \left( 2h_{2,n} h_{1,n} D\mathbf{n} [\sigma\nabla_\tau u] \right. \\ & \quad \left. + (\mathbf{h}_{1\tau} \cdot \nabla_\tau (h_{2,n}) - h_{2,n} \mathbf{n} \cdot \nabla h_{1,n} - h_{1,n} h_{2,n} H) [\sigma\nabla_\tau u] \right) v \\ & \quad - \int_{\partial\omega} \left( \operatorname{div}_\tau \left( h_{2,n} [\sigma\nabla_\tau u'_1] \right) + \operatorname{div}_\tau \left( h_{1,n} [\sigma\nabla_\tau u'_2] \right) \right) \\ & \quad - \int_{\partial\omega} \left( h_{2,n} h_{1,n} [\sigma\Delta_\tau u] + \operatorname{div}_\tau (h_{2,n} h_{1,n} [\sigma\nabla_\tau u]) \right) \partial_{\mathbf{n}} v \\ & \quad + \int_{\partial\omega} \left( h_{1,n} \operatorname{div}_\tau (h_{2,n} [\sigma\nabla_\tau u]) + h_{2,n} \operatorname{div}_\tau (h_{1,n} [\sigma\nabla_\tau u]) \right) \partial_{\mathbf{n}} v. \end{aligned}$$

We end the proof after expanding the tangential divergence of the last term of  $F_2$ .  $\square$

Let us return to the weak formulation (69) of the derivative. By identification, we get

$$\begin{aligned} [\sigma \partial_{\mathbf{n}}(u'_1)_2] &= \operatorname{div}_{\tau} \left( h_{2,n} [\sigma \nabla_{\tau} u'_1] \right) + \operatorname{div}_{\tau} \left( h_{1,n} [\sigma \nabla_{\tau} u'_2] \right) \\ &\quad - \operatorname{div}_{\tau} (h_{2,n} h_{1,n} (2D\mathbf{n} - HI) [\sigma \nabla_{\tau} u]) \\ &\quad - \operatorname{div}_{\tau} (\mathbf{h}_{1\tau} \cdot \nabla_{\tau} (h_{2,n}) [\sigma \nabla_{\tau} u]) \\ &\quad - h_{2,n} \mathbf{n} \cdot \nabla h_{1,n} [\sigma \nabla_{\tau} u] + h_{2,n} \mathbf{h}_{1\tau} \cdot D\mathbf{n} \mathbf{n} [\sigma \nabla_{\tau} u]. \end{aligned}$$

Finally, we conclude the expression giving the jump of the flux for the second order derivative by relation (29).

#### REFERENCES

- [1] L. AFRAITES, *Les techniques d'optimisation de forme pour un problème inverse de tomographie*, Ph.D. thesis, Université de Technologie de Compiègne, Compiègne, France, 2007.
- [2] L. AFRAITES, M. DAMBRINE, K. EPPLER, AND K. KATEB, *Detecting perfectly insulated obstacles by shape optimization techniques of order two*, Discrete Contin. Dyn. Syst. Ser. B, 8 (2007), pp. 389–416.
- [3] L. AFRAITES, M. DAMBRINE, AND D. KATEB, *Conformal mappings and shape derivatives for the transmission problem with a single measurement*, Numer. Funct. Anal. Optim., 28 (2007), pp. 519–551.
- [4] K. ASTALA AND L. PÄIVÄRINTA, *Calderón's inverse conductivity problem in the plane*, Ann. of Math. (2), 163 (2006), pp. 265–299.
- [5] M. C. DELFOUR AND J. P. ZOLESIO, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, SIAM, Philadelphia, 2001.
- [6] K. EPPLER AND H. HARBRECHT, *A regularized Newton method in electrical impedance tomography using Hessian information*, Control Cybernet., 34 (2005), pp. 203–225.
- [7] A. HENROT AND M. PIERRE, *Variation et optimisation de formes*, Math. Appl. (Berlin) 48, Springer-Verlag, Berlin, 2005.
- [8] F. HETTLICH AND W. RUNDELL, *The determination of a discontinuity in a conductivity from a single boundary measurement*, Inverse Problems, 14 (1998), pp. 67–82.
- [9] F. HETTLICH AND W. RUNDELL, *A second degree method for nonlinear inverse problems*, SIAM J. Numer. Anal., 37 (2000), pp. 587–620.
- [10] G. C. HSIAO AND W. L. WENDLAND, *A finite element method for some integral equations of the first kind*, J. Math. Anal. Appl., 58 (1977), pp. 449–481.
- [11] K. ITO, K. KUNISCH, AND Z. LI, *Level-set function approach to an inverse interface problem*, Inverse Problems, 17 (2001), pp. 1225–1242.
- [12] A. KIRSCH, *Surface gradients and continuity properties for some integral operator in classical scattering theory*, Math. Methods Appl. Sci., 11 (1989), pp. 789–804.
- [13] A. KIRSCH, *The domain derivative and two applications in inverse scattering theory*, Inverse Problems, 9 (1993), pp. 81–96.
- [14] R. KRESS, *Linear Integral Equations*, Appl. Math. Sci. 82, Springer-Verlag, Berlin, 1989.
- [15] V. G. MAZ'YA AND T. O. SHAPOSHNIKOVA, *Theory of Multipliers in Spaces of Differentiable Functions*, Monogr. Stud. Math. 23, Pitman, Boston, MA, 1985.
- [16] A. I. NACHMAN, *Reconstruction from boundary measurements*, Ann. of Math. (2), 128 (1988), pp. 531–576.
- [17] A. I. NACHMAN, *Global uniqueness for a two dimensional inverse boundary value problem*, Ann. of Math. (2), 143 (1996), pp. 71–96.
- [18] R. G. NOVIKOV, *A multidimensional inverse spectral problem for the equation  $\delta\psi + (v(x) - eu(x))\psi = 0$* , Funktsional. Anal. i Prilozhen., 22 (1988), pp. 11–22.
- [19] A. NOVRUZI AND M. PIERRE, *Structure of shape derivatives*, J. Evol. Equ., 2 (2002), pp. 365–382.
- [20] O. PANTZ, *Sensibilité de l'équation de la chaleur aux sauts de conductivité*, C.R. Math. Acad. Sci. Paris, 341 (2005), pp. 333–337.

- [21] J. SIMON, *Second variation for domain optimization problems*, in Control and Estimation of Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Internat. Ser. Numer. Math. 91, Birkhäuser Verlag, Basel, pp. 361–378.
- [22] J. SOKOLOWSKI AND J.-P. ZOLESIO, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, Springer-Verlag, Berlin, 1992.
- [23] J. SYLVESTER AND G. UHLMANN, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.
- [24] H. TRIEBEL, *Theory of Function Spaces*, Math. Anwendungen Phys. Tech. 38 Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig, 1983.

## A VIABILITY THEOREM FOR MORPHOLOGICAL INCLUSIONS\*

THOMAS LORENZ†

**Abstract.** The aim of this paper is to adapt the viability theorem from differential inclusions (governing the evolution of vectors in a finite-dimensional space) to so-called morphological inclusions (governing the evolution of nonempty compact subsets of the Euclidean space). In this morphological framework, the evolution of compact subsets of  $\mathbb{R}^N$  is described by means of flows along bounded Lipschitz vector fields (similarly to the velocity method (a.k.a. speed method) in shape analysis). Now for each compact subset, more than just one vector field is admitted—correspondingly to the set-valued map of a differential inclusion in finite dimensions. We specify sufficient conditions on the given data such that for every initial compact set, at least one of these compact-valued evolutions satisfies fixed state constraints in addition. The proofs follow an approximative track similar to the standard approach for differential inclusions in  $\mathbb{R}^N$ , but they use tools about weak compactness and weak convergence of Banach-valued functions. Finally an application to shape optimization under state constraints is sketched.

**Key words.** shape evolutions with state constraints, velocity method (speed method), morphological equations, Nagumo’s theorem, viability condition, Clarke’s generalized shape derivative

**AMS subject classifications.** 49J53, 34A60, 47N10, 49J24, 49Q10, 93C15

**DOI.** 10.1137/060670778

**1. Introduction.** State constraints provide challenging questions in any form of dynamic system. For the problem of finding sufficient and necessary conditions on the state constraints, the first complete answer for ordinary differential equations was given by Nagumo [26] in 1942, and this characterization (using Bouligand tangent cones) has been rediscovered many times during previous decades.

If solutions of any given initial value problem are not unique, then two versions of this question are to be distinguished from each other: we demand that either *all* solutions have their values in the fixed set of constraints or that (just) *at least one* solution with this property exists. In the first case, the corresponding set of constraints is called *invariant* and, in the latter case, it is *viable* or *weakly invariant*.

For autonomous differential inclusions in  $\mathbb{R}^N$ , the results are presented in Aubin’s monograph *Viability Theory* [2], for example.

The main goal of this paper is a sufficient characterization of viability for shapes.

To be more precise, we leave the familiar Euclidean space  $\mathbb{R}^N$  and consider evolutions of nonempty compact subsets of  $\mathbb{R}^N$  instead. Correspondingly, the solution  $x : [0, T] \rightarrow \mathbb{R}^N$  (of a differential inclusion) is now replaced by a curve  $K : [0, T] \rightarrow \mathcal{K}(\mathbb{R}^N)$ , with  $\mathcal{K}(\mathbb{R}^N)$  denoting the set of nonempty compact subsets of  $\mathbb{R}^N$  usually supplied with the Pompeiu–Hausdorff distance

$$\begin{aligned} d(K_1, K_2) &:= \sup_{\substack{x \in K_2, \\ y \in K_1}} \max \left\{ \text{dist}(x, K_1), \text{dist}(y, K_2) \right\} \\ &= \sup_{z \in \mathbb{R}^N} |\text{dist}(z, K_1) - \text{dist}(z, K_2)|. \end{aligned}$$

\*Received by the editors September 26, 2006; accepted for publication (in revised form) February 13, 2008; published electronically May 16, 2008. This work was supported by the European Community’s Human Potential Programme under contract HPRN-CT-2002-00281 (Evolution Equations).  
<http://www.siam.org/journals/sicon/47-3/67077.html>

†Interdisciplinary Center for Scientific Computing (IWR), Ruprecht–Karls–University of Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany (thomas.lorenz@iwr.uni-heidelberg.de).

The state constraints are again formulated as a subset, i.e., now  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  (instead of  $V \subset \mathbb{R}^N$  for differential inclusions).

**Lipschitz vector fields for specifying time derivatives of curves in  $(\mathcal{K}(\mathbb{R}^N), \mathcal{d})$ .** For formulating the viability problem in the metric space  $(\mathcal{K}(\mathbb{R}^N), \mathcal{d})$ , we have to specify how compact subsets of  $\mathbb{R}^N$  are “deformed.” The so-called *velocity method* or *speed method* has led C  a, Delfour, Zol  sio, and others to remarkable results about shape optimization (see, e.g., [9, 12, 11, 33, 38] and the references there). It is based on prescribing a vector field  $v : \mathbb{R}^N \times [0, T] \longrightarrow \mathbb{R}^N$  such that the corresponding ordinary differential equation  $\frac{d}{dt} x(\cdot) = v(x(\cdot), \cdot)$  induces a unique flow on  $\mathbb{R}^N$ . Indeed, supposing  $v$  to be sufficiently smooth, the Cauchy problem

$$\frac{d}{dt} x(\cdot) = v(x(\cdot), \cdot) \text{ in } [0, T], \quad x(0) = x_0 \in \mathbb{R}^N$$

is always well-posed, and any compact initial set  $K \subset \mathbb{R}^N$  is deformed to

$$\vartheta_v(t, K) := \left\{ x(t) \mid \exists x(\cdot) \in C^1([0, t], \mathbb{R}^N) : \frac{d}{dt} x(\cdot) = v(x(\cdot), \cdot) \text{ in } [0, t], \ x(0) \in K \right\}$$

after an arbitrary time  $t \geq 0$ . As a key advantage, this concept of set evolution does not require any regularity conditions on the compact set  $K$  or its topological boundary (but only on the vector field  $v$ ). In other words,  $v$  can be interpreted as a “direction of deformation” for  $(\mathcal{K}(\mathbb{R}^N), \mathcal{d})$ . So it is “possible to define directional derivatives and speak of shape gradient and shape Hessian with respect to (w.r.t.) the associated vector space of velocities. This second approach has been known in the literature as the *velocity method*” [12, Chapter 1, section 6]. (The “first” approach mentioned there in [12] refers to perturbations of the identity map and applying techniques of differential geometry.)

Aubin seized this notion for extending ordinary differential equations to this metric space of compact subsets. The so-called *morphological equations* are sketched in [3] and then presented in [5, 4] in more detail. (They seem to be closer to ordinary differential equations in  $\mathbb{R}^N$  than Panasyuk’s concept of “quasi-differential equations” [29, 28, 27]. In contrast to [1], morphological equations dispense with any aspects of affine-linear structure.)

For a given curve  $K(\cdot) : [0, T] \longrightarrow \mathcal{K}(\mathbb{R}^N)$ , autonomous Lipschitz vector fields  $\mathbb{R}^N \longrightarrow \mathbb{R}^N$  are used for specifying the counterparts of time derivatives. To be more precise, a Lipschitz continuous field  $g : \mathbb{R}^N \longrightarrow \mathbb{R}^N$  represents a first-order approximation of  $K(\cdot)$  at time  $t \in [0, T[$  if

$$(*) \quad \limsup_{h \downarrow 0} \frac{1}{h} \cdot \mathcal{d}(K(t+h), \vartheta_g(h, K(t))) = 0$$

(see Figure 1.1). Obviously, this limit superior being equal to 0 is even a limit because distances are always nonnegative by definition. Of course, such a field  $g(\cdot)$  does not have to be unique, and thus *all* bounded Lipschitz vector fields with property  $(*)$  form the so-called *mutation*  $\overset{\circ}{K}(t)$  of  $K(\cdot)$  at time  $t \in [0, T[$ . In particular, the mutation is a *subset* of all bounded Lipschitz functions  $\mathbb{R}^N \longrightarrow \mathbb{R}^N$  and extends the time derivative to curves in the metric space  $(\mathcal{K}(\mathbb{R}^N), \mathcal{d})$ .

**Solving a morphological equation with state constraints: Aubin’s adaptation of Nagumo’s theorem.** The step from specifying a time derivative (of a

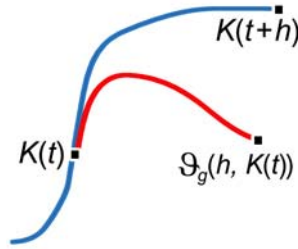


FIG. 1.1.

curve) to formulating a (generalized) differential equation is rather small. It is based just on prescribing the time derivative as a function of the current state. In connection with nonempty compact subsets of  $\mathbb{R}^N$ , a function  $f : \mathcal{K}(\mathbb{R}^N) \rightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  is given with  $\text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  denoting the set of all bounded and Lipschitz continuous functions  $\mathbb{R}^N \rightarrow \mathbb{R}^N$ .

For any initial set  $K_0 \in \mathcal{K}(\mathbb{R}^N)$ , we are looking for  $K(\cdot) : [0, T] \rightarrow \mathcal{K}(\mathbb{R}^N)$  satisfying the following:

1.  $K(\cdot)$  is Lipschitz continuous w.r.t. the Pompeiu–Hausdorff distance  $\mathcal{d}$ ;
2.  $f(K(t)) \in \mathring{K}(t)$  for a.e.  $t \in [0, T]$ , i.e.,  $\lim_{h \downarrow 0} \frac{1}{h} \cdot \mathcal{d}(K(t+h), \vartheta_{f(K(t))}(h, K(t))) = 0$ ;
3.  $K(0) = K_0$ .

Then  $K(\cdot)$  is called a *solution* of the (autonomous) *morphological equation*  $\mathring{K}(\cdot) \ni f(K(\cdot))$  in  $[0, T]$  with initial value  $K_0$ .

At first glance, the symbol  $\ni$  here seems to be contradictory to the term “equation.” The mutation  $\mathring{K}(t)$ , however, is defined as a *subset* of all transitions providing a first-order approximation of  $K(t + \cdot)$ , and so the “right-hand side”  $f(K(t)) \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  should be one of its elements. (In the classical framework of differentiable functions and vector spaces, the mutation consists of just one vector.)

Considering now additional state constraints, the question about existence of a solution has been answered completely by Aubin in [4, Theorem 0.1]. In particular, the assumptions about state constraints and  $f(\cdot)$  justify its interpretation as a counterpart of Nagumo’s theorem. Some applications and further studies are presented in [19, 17, 21]. The asymptotic condition related to the term “contingent” here is explained in more detail in the subsequent remarks after Definition 2.9 and Theorem 2.10.

**PROPOSITION 1.1** (Nagumo’s theorem for morphological equations [5, 4]). *Suppose  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  to be nonempty and closed w.r.t.  $\mathcal{d}$ .*

*Let  $f : (\mathcal{K}(\mathbb{R}^N), \mathcal{d}) \rightarrow (\text{Lip}(\mathbb{R}^N, \mathbb{R}^N), \|\cdot\|_\infty)$  be a continuous function satisfying*

1. *the uniform bound of Lipschitz constants:  $\sup_{M \in \mathcal{K}(\mathbb{R}^N)} \text{Lip } f(M) < \infty$ ;*
2. *the uniform bound of supremum norms:  $\sup_{M \in \mathcal{K}(\mathbb{R}^N)} \|f(M)\|_\infty < \infty$ .*

*For any  $M \in \mathcal{V}$ , let  $f(M) \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  be contingent to  $\mathcal{V}$  at  $M$  in the sense that*

$$\begin{aligned} 0 &= \liminf_{h \downarrow 0} \frac{1}{h} \cdot \text{dist}(\vartheta_{f(M)}(h, M), \mathcal{V}) \\ &\stackrel{\text{Def.}}{=} \liminf_{h \downarrow 0} \frac{1}{h} \cdot \inf_{C \in \mathcal{V}} \mathcal{d}(\vartheta_{f(M)}(h, M), C). \end{aligned}$$

*Then from any  $K_0 \in \mathcal{V}$  starts a solution  $K(\cdot) : [0, \infty[ \rightarrow \mathcal{K}(\mathbb{R}^N)$  of the morphological equation  $\mathring{K}(\cdot) \ni f(K(\cdot))$  which is viable in  $\mathcal{V}$ , i.e.,  $K(t) \in \mathcal{V}$  for all  $t$ .*

**The new step to morphological inclusions.** This paper focuses on the corresponding conditions (of viability) if more than one Lipschitz field is admitted for each compact set, i.e., the single-valued function  $f : \mathcal{K}(\mathbb{R}^N) \longrightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  is replaced by a set-valued map  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ . This modification of given data leads directly to the following definition: A Lipschitz continuous curve  $K(\cdot) : [0, T] \longrightarrow (\mathcal{K}(\mathbb{R}^N), d)$  is called a *solution* of the *morphological inclusion*

$$\overset{\circ}{K}(\cdot) \cap \mathcal{F}(K(\cdot)) \neq \emptyset \quad \text{in } [0, T[$$

with starting value  $K(0) \subset \mathbb{R}^N$  if  $\mathcal{F}(K(t)) \cap \overset{\circ}{K}(t) \neq \emptyset$  for almost every  $t \in [0, T[$ ; i.e., there exists  $w \in \mathcal{F}(K(t)) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  with

$$\lim_{h \downarrow 0} \frac{1}{h} \cdot d(K(t+h), \vartheta_w(h, K(t))) = 0.$$

Obviously, every morphological equation can be regarded as a morphological inclusion (just with single-valued  $\mathcal{F}$ ). So this step provides a real extension.

Considering now additional state constraints on  $K(\cdot)$ , we note that Doyen [16] has given sufficient and some necessary conditions on  $\mathcal{F}(\cdot)$  and  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  for the *invariance* of  $\mathcal{V}$  (i.e., *all continuous* solutions starting in  $\mathcal{V}$  stay in  $\mathcal{V}$ ). His key notion is first to extend Filippov's existence theorem from differential inclusions (in  $\mathbb{R}^N$ ) to morphological inclusions in  $\mathcal{K}(\mathbb{R}^N)$  [16, Theorem 7.1] and then to verify  $\text{dist}(K(t), \mathcal{V}) = 0$  for all  $t \in [0, T[$  (under the assumption that the values of  $\mathcal{F}(\cdot)$  are always contained in the corresponding *contingent cone* to  $\mathcal{V}$ ) [16, Theorem 8.2].

The main result here concerns sufficient conditions on  $\mathcal{F}(\cdot)$  and  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  for the *viability* of  $\mathcal{V}$ ; i.e., *at least one* Lipschitz continuous solution has to stay in  $\mathcal{V}$ . This question (in a more general environment) was pointed out as open in [5, section 2.3.3] and, to the best of my knowledge, has not been answered even for the special case of morphological inclusions in any article submitted before.

In fact, the following statement is very similar to the viability theorem for differential inclusions in  $\mathbb{R}^N$  (as it is discussed in [2, Theorems 3.3.2 and 3.3.4] and quoted here in Theorem 3.3). Roughly speaking,  $\mathcal{F}$  is supposed to be upper semicontinuous with closed convex values—after specifying a suitable topology on  $\text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  in a moment—and we require (at least) one “contingent direction” in the value  $\mathcal{F}(K) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  for each  $K \in \mathcal{V}$ .

**THEOREM 1.2** (viability theorem for morphological inclusions). *Let  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  be a set-valued map and  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  a nonempty closed subset satisfying the following:*

1. *all values of  $\mathcal{F}$  are nonempty and convex (i.e., for any  $\lambda \in [0, 1]$  and  $g_1, g_2 \in \mathcal{F}(K) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ , the function  $\lambda \cdot g_1 + (1 - \lambda) \cdot g_2 \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  also belongs to  $\mathcal{F}(K)$ );*
2.  *$\sup_{M \in \mathcal{K}(\mathbb{R}^N)} \sup_{f \in \mathcal{F}(M)} \text{Lip } f < \infty$  (uniformly bounded Lipschitz constants),  $\sup_{M \in \mathcal{K}(\mathbb{R}^N)} \sup_{f \in \mathcal{F}(M)} \|f\|_\infty < \infty$  (uniformly bounded supremum norms);*
3. *the graph of  $\mathcal{F}$  is closed (w.r.t. locally uniform convergence in  $\text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ );*
4. *for each  $K \in \mathcal{V}$ , some  $w \in \mathcal{F}(K) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  is contingent to  $\mathcal{V}$  at  $K$  in the sense that  $0 = \liminf_{h \downarrow 0} \frac{1}{h} \cdot \text{dist}(\vartheta_w(h, K), \mathcal{V})$ .*

*Then for every initial compact set  $K_0 \in \mathcal{V}$ , there exists at least one solution  $K(\cdot) : [0, 1] \longrightarrow \mathcal{K}(\mathbb{R}^N)$  of the morphological inclusion  $\overset{\circ}{K}(\cdot) \cap \mathcal{F}(K(\cdot)) \neq \emptyset$  with  $K(0) = K_0$  and  $K(t) \in \mathcal{V}$  for all  $t \in [0, 1]$ .*

The new analytical aspects are closely related to the proof of this theorem. Indeed, Haddad and others realized the theorem of Alaoglu as a powerful tool for constructing

solutions of differential inclusions in  $\mathbb{R}^N$  under state constraints. The counterparts of time derivatives here, however, form a bounded sequence in  $L^\infty([0, 1], \text{Lip}(\mathbb{R}^N, \mathbb{R}^N))$  which cannot be identified with a dual space in an obvious way. So results of Ülger [34] and Kisielewicz [23] come now into play for characterizing weakly compact subsets of the Bochner integrable functions  $[0, 1] \rightarrow X$  (denoted by  $L^1([0, 1], X)$ ) and the set  $C^0(K, X)$  of continuous functions  $K \rightarrow X$  with a real Banach space  $X$  and a nonempty compact set  $K \subset \mathbb{R}^N$ .

### Sketching an application to shape optimization under state constraints.

In shape optimization, the essential aim is to detect a minimizer of a given functional  $J : \mathcal{K}(\mathbb{R}^N) \rightarrow \mathbb{R}$  evaluating each nonempty compact set via a real number (see, e.g., [12, 11, 24, 33]). An additional set of constraints,  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$ , makes the problem rather complicated in general.

As an application of our viability theorem, Theorem 1.2, we suggest a set-valued map  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ , with the objective that every solution  $K(\cdot) : [0, \infty[ \rightarrow \mathcal{K}(\mathbb{R}^N)$  of the morphological inclusion  $\dot{K}(\cdot) \cap \mathcal{F}(K(\cdot)) \neq \emptyset$  satisfy the following two conditions and thus provide candidates (for the wanted minimizer):

- (i)  $J \circ K : [0, \infty[ \rightarrow \mathbb{R}$ ,  $t \mapsto J(K(t))$  is nonincreasing, and
- (ii) every compact set  $C$  being the limit of  $(K(t_n))_{n \in \mathbb{N}}$  in  $\mathcal{V}$  in the sense of Painlevé–Kuratowski for some sequence  $t_n \nearrow \infty$  satisfies a necessary condition on minimizers (in the form of Fermat’s rule).

Then Theorem 1.2 provides sufficient conditions on  $\mathcal{F}$  and  $\mathcal{V}$  for the existence of at least one solution  $K(\cdot)$  with all its values in  $\mathcal{V}$  (see Proposition 4.6).

This introduction (section 1) reflects the structure of the paper: Aubin’s theory of morphological equations is summarized in section 2; in particular, we mention the counterparts of Filippov’s and Nagumo’s theorems for evolutions in the metric space  $(\mathcal{K}(\mathbb{R}^N), d)$ . Then section 3 provides the step to morphological inclusions. It starts with the viability theorem about differential inclusions (in section 3.1), collects the tools for Banach-valued functions (in section 3.2), and verifies the viability theorem for morphological inclusions (in section 3.3). Finally, in section 4, we present the analytical details of the application to shape optimization.

**2. A brief outline of morphological equations.** Morphological equations provide typical geometric examples of so-called mutational equations. First presented in [3] and elaborated in [4, 5], mutational equations extend ordinary differential equations to a metric space  $(E, d)$ . In a word, the key idea is to describe derivatives by means of continuous maps (called *transitions*)  $\vartheta : [0, 1] \times E \rightarrow E$ ,  $(h, x) \mapsto \vartheta(h, x)$  instead of affine-linear maps  $(h, x) \mapsto x + h v$  (which are usually used in *vector spaces*). Strictly speaking, such a transition specifies the point  $\vartheta(t, x) \in E$  to which any initial point  $x \in E$  has been moved after time  $t \in [0, 1]$ .

It can be interpreted as a first-order approximation of a curve  $\xi : [0, T[ \rightarrow E$  at time  $t \in [0, T[$  if

$$\lim_{h \downarrow 0} \frac{1}{h} \cdot d(\xi(t+h), \vartheta(h, \xi(t))) = 0.$$

So-called *morphological equations* apply this concept to the set  $\mathcal{K}(\mathbb{R}^N)$  of nonempty compact subsets of  $\mathbb{R}^N$  supplied with the Pompeiu–Hausdorff distance  $d$ ,

$$\begin{aligned} d(K_1, K_2) &:= \sup_{\substack{x \in K_2, \\ y \in K_1}} \max \left\{ \text{dist}(x, K_1), \text{dist}(y, K_2) \right\} \\ &= \inf \left\{ \rho > 0 \mid K_1 \subset K_2 + \rho \mathbb{B}_1, K_2 \subset K_1 + \rho \mathbb{B}_1 \right\}. \end{aligned}$$



Here  $\mathbb{B}_1$  always denotes the closed unit ball in  $\mathbb{R}^N$ , i.e.,  $\mathbb{B}_1 := \{x \in \mathbb{R}^N \mid |x| \leq 1\}$ . This is a very general starting point for geometric evolution problems as there are no a priori restrictions in regard to the regularity of sets and their boundaries. Motivated by the velocity method (often used in shape optimization; see, e.g., [9, 12, 11, 33, 38]), ordinary differential equations are here to lay the basis for transitions.

DEFINITION 2.1.  $\text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  consists of all bounded and Lipschitz continuous functions  $\mathbb{R}^N \longrightarrow \mathbb{R}^N$ .

DEFINITION 2.2. Choosing any function  $f : \mathbb{R}^N \times [0, T] \longrightarrow \mathbb{R}^N$ , the so-called reachable set  $\vartheta_f(t, K)$  of the initial set  $K \in \mathcal{K}(\mathbb{R}^N)$  at time  $t \in [0, T]$  is defined as

$$\vartheta_f(t, K) := \left\{ x(t) \in \mathbb{R}^N \mid \exists x(\cdot) \in W^{1,1}([0, t], \mathbb{R}^N) : x(0) \in K, \right. \\ \left. \frac{d}{d\tau} x(\tau) = f(x(\tau), \tau) \text{ for a.e. } \tau \in [0, t] \right\}$$

(and correspondingly for an autonomous function  $f : \mathbb{R}^N \longrightarrow \mathbb{R}^N$ ).

The special case of constant functions  $f(\cdot) \equiv v$  (with an arbitrary vector  $v \in \mathbb{R}^N$ ) leads to the Minkowski sum  $\vartheta_f(t, K) = K + h \cdot v \subset \mathbb{R}^N$ , and for an initial set  $K = \{x\}$  with just one element, in particular, we return to the familiar affine-linear map  $(h, x) \longmapsto x + h \cdot v$  that has already been mentioned as motivation.

An essential contribution of Aubin was to specify appropriate continuity conditions on the maps  $\vartheta : [0, 1] \times E \longrightarrow E$ ,  $(h, x) \longmapsto \vartheta(h, x)$  so that the familiar track of ordinary differential equations can be followed in a metric space  $(E, d)$ . Here we quote his definition introduced in the monograph [5] (emphasizing the local features slightly more than his original version in [4]). Reachable sets of every function  $f \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  satisfy these conditions in the metric space  $(\mathcal{K}(\mathbb{R}^N), \mathcal{d})$ .

DEFINITION 2.3 (see [5, Definition 1.1.2]). Let  $(E, d)$  be a metric space. A map  $\vartheta : [0, 1] \times E \longrightarrow E$  is called a transition on  $(E, d)$  if it satisfies the following conditions:

1.  $\vartheta(0, x) = x$  for all  $x \in E$ ;
2.  $\lim_{h \downarrow 0} \frac{1}{h} \cdot d(\vartheta(t+h, x), \vartheta(h, \vartheta(t, x))) = 0$  for all  $x \in E$ ,  $t \in [0, 1[$ ;
3.  $\alpha(\vartheta) := \max(0, \sup_{x \neq y} \limsup_{h \downarrow 0} \frac{d(\vartheta(h, x), \vartheta(h, y)) - d(x, y)}{h \cdot d(x, y)}) < \infty$ ;
4.  $\beta(\vartheta) := \sup_{x \in E} \limsup_{h \downarrow 0} \frac{1}{h} \cdot d(x, \vartheta(h, x)) < \infty$ .

For any two transitions  $\vartheta_1, \vartheta_2 : [0, 1] \times E \longrightarrow E$  on the same metric space  $(E, d)$ , the transitional distance between  $\vartheta_1$  and  $\vartheta_2$  is defined by

$$d_\Lambda(\vartheta_1, \vartheta_2) := \sup_{x \in E} \limsup_{h \downarrow 0} \frac{1}{h} \cdot d(\vartheta_1(h, x), \vartheta_2(h, x)).$$

Compact reachable sets of ordinary differential equations supplied with metric  $\mathcal{d}$  satisfy all these conditions on transitions—as a consequence of the classical Cauchy–Lipschitz theorem (about solutions of ordinary differential equations).

LEMMA 2.4. For every  $f \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ , the map  $\vartheta_f : [0, 1] \times \mathcal{K}(\mathbb{R}^N) \longrightarrow \mathcal{K}(\mathbb{R}^N)$ ,  $(h, K) \longmapsto \vartheta_f(h, K)$  of reachable sets (as introduced in Definition 2.2) is a well-defined transition on the metric space  $(\mathcal{K}(\mathbb{R}^N), \mathcal{d})$  according to Definition 2.3.

To be more precise, the reachable sets satisfy the following for all initial sets  $K, K_1, K_2 \in \mathcal{K}(\mathbb{R}^N)$ , vector fields  $f, g \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ , and times  $t, h \geq 0$ :

$$\begin{aligned} \vartheta_f(0, K) &= K, \\ \vartheta_f(t+h, K) &= \vartheta_f(h, \vartheta_f(t, K)), \\ \mathcal{d}(\vartheta_f(h, K_1), \vartheta_f(h, K_2)) &\leq \mathcal{d}(K_1, K_2) \cdot e^{\text{Lip } f \cdot h}, \\ \mathcal{d}(\vartheta_f(h, K), \vartheta_g(h, K)) &\leq \|f - g\|_\infty \cdot h \cdot e^{\text{Lip } f \cdot h}, \\ \mathcal{d}(\vartheta_f(t, K), \vartheta_f(t+h, K)) &\leq \|f\|_\infty \cdot h. \end{aligned}$$

Thus,  $\alpha(\vartheta_f) \leq \text{Lip } f$ ,  $\beta(\vartheta_f) \leq \|f\|_\infty$ ,  $d_\Lambda(\vartheta_f, \vartheta_g) \leq \|f - g\|_\infty \stackrel{\text{Def.}}{=} \sup_{x \in \mathbb{R}^N} |f(x) - g(x)|$ .

In particular,  $d(\vartheta_f(h, K_1), \vartheta_g(h, K_2)) \leq e^{\text{Lip } f \cdot h} (d(K_1, K_2) + h \cdot \|f - g\|_\infty)$ .

The proof is presented in [5, Proposition 3.5.3]. In particular, this lemma justifies calling  $\vartheta_f$  a *shape transition* on  $(\mathcal{K}(\mathbb{R}^N), d)$  (or *morphological transition*, in accordance with [5, Definition 3.7.2]). For the sake of simplicity,  $f \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  is sometimes identified with its shape transition  $\vartheta_f$ .

These reachable sets provide the tools for specifying (generalized) shape derivatives of a compact-valued tube  $K(\cdot) : [0, T[ \rightsquigarrow \mathbb{R}^N$ , i.e., a curve  $K(\cdot) : [0, T[ \rightarrow \mathcal{K}(\mathbb{R}^N)$ . So the next step will be to solve equations prescribing an element of the shape mutation.

**DEFINITION 2.5** (see [5, section 3.5.4]). *For any compact-valued tube  $K(\cdot) : [0, T[ \rightsquigarrow \mathbb{R}^N$ , the shape mutation  $\mathring{K}(t)$  at time  $t \in [0, T[$  consists of all  $f \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  satisfying*

$$\lim_{h \downarrow 0} \frac{1}{h} \cdot d(\vartheta_f(h, K(t)), K(t+h)) = 0.$$

**DEFINITION 2.6.** *For any given function  $f : \mathcal{K}(\mathbb{R}^N) \times [0, T[ \rightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ , a compact-valued  $K(\cdot) : [0, T[ \rightsquigarrow \mathbb{R}^N$  is called a solution of the morphological equation*

$$\mathring{K}(\cdot) \ni f(K(\cdot), \cdot)$$

if

1.  $K(\cdot) : [0, T[ \rightsquigarrow \mathbb{R}^N$  is Lipschitz continuous with respect to  $d$ , and
2. for almost every  $t \in [0, T[$ ,  $f(K(t), t) \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  belongs to  $\mathring{K}(t)$  or, equivalently,  $\lim_{h \downarrow 0} \frac{1}{h} \cdot d(\vartheta_{f(K(t), t)}(h, K(t)), K(t+h)) = 0$ .

These conditions on a solution are in accordance with [5, Definition 1.3.1] being formulated for the autonomous case (i.e.,  $f$  not depending on time explicitly).

As an essential result of [5, 4], the Euler algorithm can be applied in the framework of morphological equations, and so the Cauchy–Lipschitz theorem (about ordinary differential equations) has the following counterpart that is proved in [5, Theorem 4.1.2] for the more general case that the values of  $f$  are bounded Lipschitz continuous *set-valued* maps.

**THEOREM 2.7.** *Suppose  $f : (\mathcal{K}(\mathbb{R}^N), d) \rightarrow (\text{Lip}(\mathbb{R}^N, \mathbb{R}^N), \|\cdot\|_\infty)$  to be Lipschitz continuous with Lipschitz constant  $\lambda$  and  $M := \sup_{K \in \mathcal{K}(\mathbb{R}^N)} \text{Lip } f(K) < \infty$ .*

*For every initial set  $K_0 \in \mathcal{K}(\mathbb{R}^N)$  and time  $T > 0$ , there exists a unique compact-valued solution  $K(\cdot) : [0, T[ \rightsquigarrow \mathbb{R}^N$  of the morphological equation  $\mathring{K}(\cdot) \ni f(K(\cdot))$  with  $K(0) = K_0$ .*

*Furthermore every Lipschitz compact-valued tube  $Q : [0, \infty[ \rightsquigarrow \mathbb{R}^N$  with  $\mathring{Q}(t) \neq \emptyset$  for every  $t \geq 0$  satisfies the following estimate at each time  $t \geq 0$ :*

$$d(K(t), Q(t)) \leq d(K_0, Q(0)) \cdot e^{(M+\lambda)t} + \int_0^t e^{(M+\lambda)(t-s)} \cdot \inf_{g \in \mathring{Q}(s)} \|f(Q(s)) - g\|_\infty ds.$$

*In particular, the solution  $K(\cdot)$  depends on the initial set  $K_0$  and the right-hand side  $f$  in a Lipschitz continuous way.*

Existence under (additional) state constraints proves to be a very interesting question for many applications. In the particular case of ordinary differential equations, Nagumo's theorem gives a necessary and sufficient condition on the set of constraints  $\mathcal{V}$  for existence of local solutions. It uses the contingent cone (in the sense of Bouligand) and has served as a key motivation for viability theory (see, e.g., [2]).

DEFINITION 2.8 (see [2, Definition 1.1.3]). *Let  $X$  be a normed vector space, let  $V \subset X$  be nonempty, and let  $x \in V$ . The contingent cone to  $V$  at  $x$  (in the sense of Bouligand) is*

$$T_V(x) := \left\{ u \in X \mid \liminf_{h \downarrow 0} \frac{1}{h} \cdot \text{dist}(x + hu, V) = 0 \right\}.$$

This classical definition of contingent cone in a normed vector space is now extended to the metric space  $(\mathcal{K}(\mathbb{R}^N), \mathbf{d})$  by using the shape transitions of  $\text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  (instead of affine-linear maps as in vector spaces).

DEFINITION 2.9 (see [5, Definition 1.5.2]). *For nonempty  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  and  $K \in \mathcal{V}$ ,*

$$T_{\mathcal{V}}(K) := \left\{ f \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N) \mid \liminf_{h \downarrow 0} \frac{1}{h} \cdot \text{dist}(\vartheta_f(h, K), \mathcal{V}) = 0 \right\}$$

*is called a contingent transition set of  $\mathcal{V}$  at  $K$ .*

*Remark.* Considering here the metric space  $(\mathcal{K}(\mathbb{R}^N), \mathbf{d})$  (instead of a normed vector space as in Definition 2.8) has an immediate consequence: By definition of the distance from a subset  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$ ,

$$\text{dist}(\vartheta_F(h, K), \mathcal{V}) \stackrel{\text{Def.}}{=} \inf_{C \in \mathcal{V}} \mathbf{d}(\vartheta_f(h, K), C),$$

with the Pompeiu–Hausdorff metric  $\mathbf{d}$  on the right-hand side. In particular, we cannot expect any trivial identities of the contingent cone to a compact subset  $V \subset \mathbb{R}^N$  and the contingent transition set to  $\mathcal{V} := \{V\} \subset \mathcal{K}(\mathbb{R}^N)$ . Furthermore, Theorem 2.10 below and the main result of this paper, i.e., Theorem 3.11, become definitely incorrect if the Pompeiu–Hausdorff distance  $\mathbf{d}$  is replaced by the one-sided distance part called Pompeiu–Hausdorff excess (as defined in [5, section 3.2.1]). A counterexample is given in the remark after Theorem 2.10.

*Remark.* The “geometric” background of reachable sets implies an additional property of shape transitions in  $T_{\mathcal{V}}(K) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ . Indeed, for any  $f \in T_{\mathcal{V}}(K)$ , every function  $g \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  with  $f(\cdot) = g(\cdot)$  in a neighborhood of  $\partial K$  is also contained in  $T_{\mathcal{V}}(K)$  because the Cauchy–Lipschitz theorem about ordinary differential equations ensures  $\partial \vartheta_f(t, K) = \vartheta_f(t, \partial K) = \vartheta_g(t, \partial K) = \partial \vartheta_g(t, K)$  for small  $t \geq 0$ . So in other words, the criterion of  $T_{\mathcal{V}}(K)$  depends only on an arbitrarily small neighborhood of the boundary  $\partial K$ .

In fact, Nagumo’s theorem also holds for morphological equations, as shown in [5, Theorem 4.1.7] (again for the more general case that  $f$  is a *single-valued* function whose values are uniformly bounded Lipschitz continuous *set-valued* maps  $\mathbb{R}^N \rightsquigarrow \mathbb{R}^N$ ).

THEOREM 2.10 (Nagumo theorem for morphological equations [5, Theorem 4.1.7]). *Suppose  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  to be nonempty and closed w.r.t.  $\mathbf{d}$ .*

*Let  $f : (\mathcal{K}(\mathbb{R}^N), \mathbf{d}) \longrightarrow (\text{Lip}(\mathbb{R}^N, \mathbb{R}^N), \|\cdot\|_{\infty})$  be a continuous function satisfying*

1.  $\sup_{M \in \mathcal{K}(\mathbb{R}^N)} \text{Lip } f(M) < \infty$ ,
2.  $\sup_{M \in \mathcal{K}(\mathbb{R}^N)} \|f(M)\|_{\infty} < \infty$ .

*Then from any initial state  $K_0 \in \mathcal{V}$  starts at least one Lipschitz solution  $K(\cdot) : [0, T[ \longrightarrow \mathcal{K}(\mathbb{R}^N)$  of  $\dot{K}(\cdot) \ni f(K(\cdot))$  viable in  $\mathcal{V}$  (i.e.,  $K(t) \in \mathcal{V}$  for all  $t$ ) if and only if  $\mathcal{V}$  is a viability domain of  $f$  in the sense of  $f(M) \in T_{\mathcal{V}}(M)$  for each  $M \in \mathcal{V}$ .*

*Remark.* A simple example proves that the one-sided Hausdorff distance—a.k.a. Pompeiu–Hausdorff excess—as defined in [5, section 3.2.1]), namely,

$$h^{\sharp}(K_1, K_2) := \sup_{x \in K_1} \inf_{y \in K_2} |x - y| \quad \text{for } K_1, K_2 \in \mathcal{K}(\mathbb{R}^N),$$

must not replace the Pompeiu–Hausdorff metric in Definition 2.9 of the contingent transition set. Consider  $\mathcal{V} := \{\mathbb{B}_1\}$  with the closed unit ball  $\mathbb{B}_1 \subset \mathbb{R}^N$  and the constant map  $f(\cdot) \equiv f_0$  with  $f_0 : \mathbb{R}^N \rightarrow \mathbb{R}^N$ ,  $x \mapsto \frac{-x}{1+|x|}$ . Indeed, the flow along  $f_0$  makes the unit ball shrink strictly, and thus  $\vartheta_{f_0}(t, \mathbb{B}_1) \subset \mathbb{R}^N$  is contained in the interior of the unit ball  $\mathbb{B}_1$  for any  $t > 0$ , i.e.,  $\vartheta_{f_0}(t, \mathbb{B}_1) \notin \mathcal{V}$  and  $h^\#(\vartheta_{f_0}(t, \mathbb{B}_1), \mathbb{B}_1) = 0$ .

$h^\#$  can be useful, however, for other types of viability problems, such as those examples discussed in [30].

**3. The step to morphological inclusions.** The main goal now is to prove a similar viability theorem for morphological *inclusions*; i.e., the single-valued function  $f : \mathcal{K}(\mathbb{R}^N) \rightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  of the right-hand side is to be replaced by a set-valued map  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ . Correspondingly to Definition 2.6, we introduce the solution of a morphological inclusion in the following way.

**DEFINITION 3.1.** *For any given function  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ , a compact-valued  $K(\cdot) : [0, T[ \rightsquigarrow \mathbb{R}^N$  is called a solution of the morphological inclusion*

$$\mathring{K}(\cdot) \cap \mathcal{F}(K(\cdot), \cdot) \neq \emptyset$$

if

1.  $K(\cdot) : [0, T[ \rightsquigarrow \mathbb{R}^N$  is Lipschitz continuous w.r.t.  $d$ , and
2.  $\mathcal{F}(K(t)) \cap \mathring{K}(t) \neq \emptyset$  for almost every  $t$ ; i.e., some  $w \in \mathcal{F}(K(t)) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  belongs to  $\mathring{K}(t)$  or, equivalently,  $\lim_{h \downarrow 0} \frac{1}{h} \cdot d(K(t+h), \vartheta_w(h, K(t))) = 0$ .

**3.1. The (well-known) viability theorem for differential inclusions.** The situation has already been investigated extensively for differential inclusions in  $\mathbb{R}^N$  (see, e.g., [2, 6]). For clarifying the new aspects of morphological inclusions, we now quote the corresponding result from [2, Theorems 3.3.2 and 3.3.5] after specifying the required terms.

**DEFINITION 3.2** (see [2, Definition 2.2.4]). *Let  $X$  and  $Y$  be normed vector spaces. A set-valued map  $F : X \rightsquigarrow Y$  is called a Marchaud map if it has the following properties:*

1.  $F$  is nontrivial, i.e.,  $\text{Graph } F \neq \emptyset$ ;
2.  $F$  is upper semicontinuous; i.e., for any  $x \in X$ , neighborhood  $V \supset F(x)$ ,  $\exists$  neighborhood  $U \subset X$  of  $x$  s.t.  $F(U) \subset V$ ;
3.  $F$  has compact convex values;
4.  $F$  has linear growth, i.e.,  $\sup_{y \in F(x)} |y| \leq C(1 + |x|)$  for all  $x \in X$ .

**THEOREM 3.3** (viability theorem for differential inclusions [2, Theorems 3.3.2 and 3.3.5]). *Consider a Marchaud map  $F : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N$  and a nonempty closed subset  $V \subset \mathbb{R}^N$  with  $F(x) \neq \emptyset$  for all  $x \in V$ .*

*Then for any  $T \in ]0, \infty[$ , the following two statements are equivalent:*

1. *For every point  $x_0 \in V$ , there is at least one solution  $x(\cdot) \in W^{1,1}([0, T], \mathbb{R}^N)$  of  $x'(\cdot) \in F(x(\cdot))$  (almost everywhere) with  $x(0) = x_0$  and  $x(t) \in V$  for all  $t$ .*
2.  *$F(x) \cap T_V(x) \neq \emptyset$  for all  $x \in V$ .*

The implication (1.)  $\implies$  (2.) is rather obvious. For proving (2.)  $\implies$  (1.), a standard approach uses an “approximating” sequence  $(x_n(\cdot))_{n \in \mathbb{N}}$  in  $W^{1,\infty}([0, 1], \mathbb{R}^N)$  such that  $\sup_t \text{dist}(x_n(t), V) \rightarrow 0$  ( $n \rightarrow \infty$ ) and  $(x_n(t), \frac{d}{dt} x_n(t))$  is close to  $\text{Graph } F \subset \mathbb{R}^N \times \mathbb{R}^N$  for almost every  $t$ . Then the theorems of Arzelà–Ascoli and Alaoglu provide a subsequence  $(x_{n_j}(\cdot))_{j \in \mathbb{N}}$  and limits  $x(\cdot) \in C^0([0, 1], \mathbb{R}^N)$ ,  $w(\cdot) \in L^\infty([0, 1], \mathbb{R}^N)$  with

$$x_{n_j}(\cdot) \rightarrow x(\cdot) \text{ uniformly, } \quad \frac{d}{dt} x_{n_j}(\cdot) \rightarrow w(\cdot) \text{ weakly}^* \text{ in } L^\infty([0, 1], \mathbb{R}^N).$$

Due to the continuous embedding  $L^\infty([0, 1], \mathbb{R}^N) \subset L^1([0, 1], \mathbb{R}^N)$ , we even obtain the convergence  $\frac{d}{dt} x_{n_j}(\cdot) \longrightarrow w(\cdot)$  weakly in  $L^1([0, 1], \mathbb{R}^N)$ . Thus,  $w(\cdot)$  is the weak derivative of  $x(\cdot)$ , and  $x(\cdot)$  is Lipschitz continuous. Finally Mazur's lemma quoted in Proposition 3.5 below implies

$$w(t) \in \bigcap_{\varepsilon > 0} \overline{\text{co}} \left( \bigcup_{z \in \mathbb{B}_\varepsilon(x(t))} F(z) \right) = F(x(t)) \quad \text{for a.e. } t.$$

Considering now morphological inclusions on  $(\mathcal{K}(\mathbb{R}^N), d)$  (instead of differential inclusions), an essential aspect changes: The derivative of a curve is no longer represented as a function in  $L^1([0, 1], \mathbb{R}^N)$ , but rather as a function  $[0, 1] \longrightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ . So the classical theorems of Arzelà–Ascoli, Alaoglu, and Mazur might have to be replaced by their counterparts concerning functions with their values in a Banach space (instead of  $\mathbb{R}^N$ ).

**3.2. Tools for functions with values in metric or Banach spaces.** Before adapting this concept for finite-dimensional differential inclusions to Banach-valued functions, we collect briefly the main tools in this framework. They consist mainly of (particularly weakly sequential) compactness criteria for both Bochner-integrable functions on a probabilistic space and continuous functions on a compact Hausdorff space.

First of all, the theorems of Arzelà–Ascoli and Mazur do not change significantly. Indeed, we always use the following general versions in this paper.

**PROPOSITION 3.4** (Arzelà–Ascoli in metric spaces [22]). *Let  $(E_1, d_1)$ ,  $(E_2, d_2)$  be precompact metric spaces; i.e., for any  $\varepsilon > 0$ , each set  $E_i$  ( $i = 1, 2$ ) can be covered by finitely many  $\varepsilon$ -balls w.r.t. metric  $d_i$ . Moreover, suppose the sequence  $(f_n)_{n \in \mathbb{N}}$  of functions  $E_1 \longrightarrow E_2$  to be uniformly equicontinuous (i.e., with a common modulus of continuity in  $E_1$ ).*

*Then there exists a subsequence  $(f_{n_j})_{j \in \mathbb{N}}$  that is a Cauchy sequence w.r.t. uniform convergence. If  $(E_2, d_2)$  is complete in addition, then  $(f_{n_j})_{j \in \mathbb{N}}$  converges uniformly to a continuous function  $E_1 \longrightarrow E_2$ .*

**PROPOSITION 3.5** (Mazur's lemma [36, section V.1, Theorem 2]). *For any weakly converging sequence  $(x_n)_{n \in \mathbb{N}}$  in a normed vector space, its weak limit is contained in the closed convex hull of  $\{x_n \mid n \in \mathbb{N}\}$ .*

The so-called Bochner integral extends the familiar concept of integration from real-valued functions to Banach-valued functions on the basis of “simple” functions.

**DEFINITION 3.6** (see [15]). *Let  $(\Omega, \Sigma, \mu)$  be a finite measure space and  $X$  a Banach space. A function  $f : \Omega \longrightarrow X$  is called simple if there exist  $x_1, x_2, \dots, x_n \in X$  and  $E_1, E_2, \dots, E_n \in \Sigma$  such that  $f = \sum_{j=1}^n x_j \chi_{E_j}$ , with  $\chi_{E_j} : \Omega \longrightarrow \{0, 1\}$  denoting the characteristic function of  $E_j \subset \Omega$ .*

*A function  $f : \Omega \longrightarrow X$  is called  $\mu$ -measurable if there exists a sequence  $(f_n)_{n \in \mathbb{N}}$  of simple functions  $\Omega \longrightarrow X$  with  $\|f - f_n\|_X \longrightarrow 0$   $\mu$ -almost everywhere for  $n \rightarrow \infty$ .*

*A  $\mu$ -measurable function  $f : \Omega \longrightarrow X$  is called Bochner integrable if there exists a sequence  $(f_n)_{n \in \mathbb{N}}$  of simple functions  $\Omega \longrightarrow X$  such that*

$$\lim_{n \rightarrow \infty} \int_{\Omega} \|f - f_n\|_X \, d\mu = 0.$$

*Then the Bochner integral of  $f$  over  $E \in \Sigma$  is defined by  $\int_E f \, d\mu := \lim_{n \rightarrow \infty} \int_E f_n \, d\mu$ . Let  $L^1(\mu, X)$  denote the Banach space of Bochner integrable functions  $\Omega \longrightarrow X$  equipped with its usual  $L^1$  norm.*

In the nineties, Ülger proved that restricting the values of Bochner integrable functions to a weakly compact subset of  $X$  implies the relative weak compactness of these functions in  $L^1(\mu, X)$ . For real-valued Lebesgue integrable functions, this is closely related with Alaoglu's theorem and a compact embedding.

**PROPOSITION 3.7** (see [34, Proposition 7]). *Let  $(\Omega, \Sigma, \mu)$  be a probabilistic space and  $X$  an arbitrary Banach space. For any weakly compact subset  $W \subset X$ , the set*

$$\{h \in L^1(\mu, X) \mid h(\omega) \in W \text{ for } \mu\text{-a.e. } \omega \in \Omega\}$$

*is relatively weakly compact in  $L^1(\mu, X)$ .*

An earlier version of this result is presented in [13], and [14] considers weak compactness of Bochner integrable functions with values in an arbitrary Banach space under weaker assumptions (see also [8]). The next proposition of Ülger provides a “weakly pointwise” characterization of weakly convergent sequences in  $L^1(\mu, X)$ .

**PROPOSITION 3.8** (see [34, Corollary 5]). *Let  $(\Omega, \Sigma, \mu)$  be a probabilistic space and  $X$  an arbitrary Banach space as in Proposition 3.7.*

*Set  $W := \{g \in L^1(\mu, X) \mid |g(\omega)| \leq 1 \text{ for } \mu\text{-almost every } \omega \in \Omega\}$ .*

*A sequence  $(g_n(\cdot))_{n \in \mathbb{N}}$  in  $W \subset L^1(\mu, X)$  converges weakly to  $g \in L^1(\mu, X)$  if and only if, for any subsequence  $(g_{n_k}(\cdot))_{k \in \mathbb{N}}$  given, there exists a sequence  $(h_k(\cdot))_{k \in \mathbb{N}}$  with  $h_k \in \text{co}\{g_{n_k}, g_{n_{k+1}}, \dots\}$  such that for  $\mu$ -almost every  $\omega \in \Omega$ ,*

$$h_k(\omega) \longrightarrow g(\omega) \quad (k \longrightarrow \infty) \quad \text{weakly in } X.$$

In fact, the classical theorem of Scorza-Dragoni [32] has a counterpart for Banach-valued functions as shown by Ricceri and Villani [31]. A so-called Carathéodory function depends on two arguments, and it is measurable w.r.t. the first one and continuous w.r.t. the second one. The key point of Scorza-Dragoni is to ensure continuity w.r.t. both arguments on “almost” the whole domain in the following sense.

**PROPOSITION 3.9** (see [31, Theorem 1]). *Let  $S$  be a compact Hausdorff topological space,  $\mu$  a Radon measure on  $S$ , and  $X, Y$  metric spaces. Suppose  $X$  to be separable.*

*Then every Carathéodory function  $g : S \times X \longrightarrow Y$  satisfies the so-called Scorza-Dragoni property; i.e., for every  $\varepsilon > 0$ , there exists a closed subset  $S_\varepsilon \subset S$  with  $\mu(S \setminus S_\varepsilon) < \varepsilon$  such that the restriction  $g|_{S_\varepsilon \times X}$  is continuous.*

So this proposition can be regarded as a counterpart of the well-known Lusin theorem (relating measurability to continuity almost everywhere)—but for functions with two arguments.

Last but not least, we quote a result of Kisielewicz characterizing weakly converging sequences of continuous functions on a compact Hausdorff space (like  $[0, T] \subset \mathbb{R}$ ).

**PROPOSITION 3.10** (see [23, Theorem 3]). *Let  $S$  be a compact Hausdorff space and  $X$  an arbitrary Banach space.  $C^0(S, X)$  denotes the Banach space of continuous functions  $S \longrightarrow X$  supplied with the supremum norm  $\|\cdot\|_\infty$ .*

*A sequence  $(g_n(\cdot))_{n \in \mathbb{N}}$  in  $C^0(S, X)$  converges weakly to  $g \in C^0(S, X)$  if and only if*

$$\bigwedge \left\{ \begin{array}{ll} \sup_n \|g_n\|_\infty < \infty & \text{and} \\ g_n(s) \longrightarrow g(s) & \text{weakly in } X \quad (n \longrightarrow \infty) \quad \text{for every } s \in S. \end{array} \right.$$

**3.3. Adapting this concept to morphological inclusions.** Now  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  and a set of constraints  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  are given.

Correspondingly to Theorem 3.3 about differential inclusions, we focus on the so-called *viability condition*, demanding from each compact set  $K \in \mathcal{V}$  that the value

$\mathcal{F}(K)$  and the contingent transition set  $T_{\mathcal{V}}(K) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  have at least one transition in common. Lacking a concrete counterpart of the Aumann integral in the metric space  $(\mathcal{K}(\mathbb{R}^N), \mathbf{d})$ , the question of its necessity (for the existence of solutions “viable in  $\mathcal{V}$ ”) is more complicated than for differential inclusions in  $\mathbb{R}^N$ , and thus we skip it here deliberately.

The main contribution of this paper is that in combination with appropriate assumptions about  $\mathcal{F}(\cdot)$  and  $\mathcal{V}$ , the viability condition is *sufficient*.

**THEOREM 3.11.** *Let  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  be a set-valued map and  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  a nonempty closed subset satisfying the following:*

1. *all values of  $\mathcal{F}$  are nonempty and convex;*
2.  *$A := \sup_{M \in \mathcal{K}(\mathbb{R}^N)} \sup_{f \in \mathcal{F}(M)} \text{Lip } f < \infty$ ,  $B := \sup_{M \in \mathcal{K}(\mathbb{R}^N)} \sup_{f \in \mathcal{F}(M)} \|f\|_{\infty} < \infty$ ;*
3. *the graph of  $\mathcal{F}$  is closed (w.r.t. locally uniform convergence in  $\text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ );*
4.  *$T_{\mathcal{V}}(M) \cap \mathcal{F}(M) \neq \emptyset$  for all  $M \in \mathcal{V}$ .*

*Then for every initial compact set  $K_0 \in \mathcal{V}$ , there exists at least one compact-valued Lipschitz continuous solution  $K(\cdot) : [0, 1] \rightsquigarrow \mathbb{R}^N$  of the morphological inclusion  $\mathring{K}(\cdot) \cap \mathcal{F}(K(\cdot)) \neq \emptyset$  with  $K(0) = K_0$  and  $K(t) \in \mathcal{V}$  for all  $t \in [0, 1]$ .*

The proof is given in several steps using the notation of Theorem 3.11 (about constants  $A, B$ ). It is based on approximative solutions.

**LEMMA 3.12** (constructing approximative solutions). *Choose any  $\varepsilon > 0$ .*

*Under the assumptions of the viability theorem, Theorem 3.11, there exist a  $B$ -Lipschitz continuous function  $K_{\varepsilon}(\cdot) : [0, 1] \rightarrow \mathcal{K}(\mathbb{R}^N)$  and a piecewise constant function  $f_{\varepsilon}(\cdot) : [0, 1[ \rightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  satisfying the following with  $R_{\varepsilon} := \varepsilon e^A$ :*

- (a)  $K_{\varepsilon}(0) = K_0$ ;
- (b)  $\text{dist}(K_{\varepsilon}(t), \mathcal{V}) \leq R_{\varepsilon}$  for all  $t \in [0, 1]$ ;
- (c)  $f_{\varepsilon}(t) \in \mathring{K}_{\varepsilon}(t) \cap \mathcal{F}(\mathbb{B}_{R_{\varepsilon}}(K_{\varepsilon}(t))) \neq \emptyset$  for all  $t \in [0, 1[$ .

*Proof.* The proof follows the same track as [5, Lemma 1.6.5] and uses Zorn’s lemma: For  $\varepsilon > 0$  fixed, let  $\mathcal{A}_{\varepsilon}(K_0)$  denote the set of all tuples  $(\tau_K, K(\cdot), f(\cdot))$  consisting of some  $\tau_K \in [0, 1]$ , a  $B$ -Lipschitz continuous function  $K(\cdot) : [0, \tau_K] \rightarrow (\mathcal{K}(\mathbb{R}^N), \mathbf{d})$ , and some piecewise constant function  $f(\cdot) : [0, 1[ \rightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  such that

- (a)  $K(0) = K_0$ ;
- (b') (1)  $\text{dist}(K(\tau_K), \mathcal{V}) \leq r_{\varepsilon}(\tau_K)$  with  $r_{\varepsilon}(t) := \varepsilon e^{At}$ ,  
(2)  $\text{dist}(K(t), \mathcal{V}) \leq R_{\varepsilon}$  for all  $t \in [0, \tau_K]$ ;
- (c)  $f(t) \in \mathring{K}(t) \cap \mathcal{F}(\mathbb{B}_{R_{\varepsilon}}(K(t))) \neq \emptyset$  for all  $t \in [0, \tau_K[$ .

Obviously,  $\mathcal{A}_{\varepsilon}(K_0)$  is not empty since it contains  $(0, K(\cdot) \equiv K_0, f(\cdot) \equiv z_0)$  with the zero function  $z_0 \equiv 0 \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ . Moreover, an order relation  $\preceq$  on  $\mathcal{A}_{\varepsilon}(K_0)$  is specified by

$$(\tau_K, K(\cdot), f(\cdot)) \preceq (\tau_M, M(\cdot), g(\cdot)) :\Longleftrightarrow \tau_K \leq \tau_M, M|_{[0, \tau_K]} = K, g|_{[0, \tau_K]} = f.$$

So Zorn’s lemma provides a maximal element  $(\tau, K_{\varepsilon}(\cdot), f_{\varepsilon}(\cdot)) \in \mathcal{A}_{\varepsilon}(K_0)$ .

As all considered functions with values in  $\mathcal{K}(\mathbb{R}^N)$  have been supposed to be  $B$ -Lipschitz continuous,  $K_{\varepsilon}(\cdot)$  is also  $B$ -Lipschitz continuous in  $[0, \tau[$ . In particular,  $K_{\varepsilon}(\cdot)$  can always be extended to the closed interval  $[0, \tau] \subset [0, 1]$  in a unique way.

Assuming  $\tau < 1$  for a moment, we obtain a contradiction if  $K_{\varepsilon}(\cdot), f_{\varepsilon}(\cdot)$  can be extended to a larger interval  $[0, \tau + \delta] \subset [0, 1]$  ( $\delta > 0$ ) preserving conditions (b’), (c).

Since closed bounded balls of  $(\mathcal{K}(\mathbb{R}^N), \mathbf{d})$  are compact, the closed set  $\mathcal{V}$  contains an element  $Z \in \mathcal{K}(\mathbb{R}^N)$  with  $\mathbf{d}(K_{\varepsilon}(\tau), Z) = \text{dist}(K_{\varepsilon}(\tau), \mathcal{V}) \leq r_{\varepsilon}(\tau)$ , and assumption 4

of Theorem 3.11 provides an element

$$g \in T_V(Z) \cap \mathcal{F}(Z) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N).$$

Due to Definition 2.9 of the contingent transition set  $T_V(Z)$ , there is a sequence  $h_m \downarrow 0$  in  $]0, 1 - \tau[$  such that  $\text{dist}(\vartheta_g(h_m, Z), V) \leq \varepsilon h_m$  for all  $m \in \mathbb{N}$ . Now set

$$K_\varepsilon(t) := \vartheta_g(t - \tau, K_\varepsilon(\tau)), \quad f_\varepsilon(t) := g \quad \text{for each } t \in [\tau, \tau + h_1].$$

Obviously, Lemma 2.4 implies  $g \in \mathring{K}_\varepsilon(t)$  for all  $t \in [\tau, \tau + h_1[$ . Moreover, it leads to

$$\begin{aligned} d(K_\varepsilon(t), Z) &\leq d(\vartheta_g(t - \tau, K_\varepsilon(\tau)), K_\varepsilon(\tau)) + d(K_\varepsilon(\tau), Z) \\ &\leq B \cdot (t - \tau) + \varepsilon e^{A\tau} \tau \leq R_\varepsilon \end{aligned}$$

for every  $t \in [\tau, \tau + \delta[$  with  $\delta := \min\{h_1, \varepsilon e^{A\frac{1-\tau}{1+B}}\}$ ; i.e., conditions (b')(2) and (c) hold in the interval  $[\tau, \tau + \delta]$ . For any index  $m \in \mathbb{N}$  with  $h_m < \delta$ ,

$$\begin{aligned} \text{dist}(K_\varepsilon(\tau + h_m), V) &\leq d(\vartheta_g(h_m, K_\varepsilon(\tau)), \vartheta_g(h_m, Z)) + \text{dist}(\vartheta_g(h_m, Z), V) \\ &\leq d(K_\varepsilon(\tau), Z) \cdot e^{Ah_m} + \varepsilon \cdot h_m \\ &\leq \varepsilon e^{A\tau} \tau \cdot e^{Ah_m} + \varepsilon \cdot h_m \leq r_\varepsilon(\tau + h_m); \end{aligned}$$

i.e., condition (b')(1) is also satisfied at time  $t = \tau + h_m$  with any large  $m \in \mathbb{N}$ .

So  $K_\varepsilon(\cdot)|_{[0, \tau+h_m]}$  and  $f_\varepsilon(\cdot)|_{[0, \tau+h_m]}$  provide the wanted contradiction, and thus  $\tau = 1$ .  $\square$

Using the abbreviation  $\tilde{K}_j := \mathbb{B}_{j+B}(K_0) \stackrel{\text{Def.}}{=} \{x \in \mathbb{R}^N \mid \text{dist}(x, K_0) \leq j + B\}$  ( $j \in \mathbb{N}$ ) for (arbitrarily large) compact neighborhoods of the initial set  $K_0$ , we obtain the following.

**LEMMA 3.13** (selecting an approximative subsequence). *Under the assumptions of the viability theorem, Theorem 3.11, there are sequences  $K_n(\cdot) : [0, 1] \rightarrow \mathcal{K}(\mathbb{R}^N)$ ,  $f_n(\cdot) : [0, 1[ \rightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  ( $n \in \mathbb{N}$ ) and functions  $K(\cdot) : [0, 1] \rightarrow \mathcal{K}(\mathbb{R}^N)$ ,  $f(\cdot) : [0, 1[ \rightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  such that for every  $j, n \in \mathbb{N}$ ,*

- (a)  $K_0 = K_n(0) = K(0)$ ;
- (b)  $K(\cdot)$  and  $K_n(\cdot)$  are  $B$ -Lipschitz continuous w.r.t.  $d$ ;
- (c)  $f_n(\cdot)$  is piecewise constant,  $\sup_{t \in [0, 1[} \text{Lip } f_n(t) \leq A < \infty$ ,  
 $\sup_{t \in [0, 1[} \|f_n(t)\|_\infty \leq B < \infty$ ;
- (d)  $\text{dist}(K_n(t), V) \leq \frac{1}{n}$  for all  $t \leq 1$ ;
- (e)  $f_n(t) \in \mathring{K}_n(t) \cap \mathcal{F}(\mathbb{B}_{1/n}(K_n(t))) \neq \emptyset$  for all  $t < 1$ ;
- (f)  $d(K_m(\cdot), K(\cdot)) \rightarrow 0$  uniformly in  $[0, 1]$  for  $m \rightarrow \infty$ ;
- (g)  $f_m(\cdot)|_{\tilde{K}_j} \rightarrow f(\cdot)|_{\tilde{K}_j}$  weakly in  $L^1([0, 1], C^0(\tilde{K}_j, \mathbb{R}^N))$  for  $m \rightarrow \infty$ ;
- (h)  $\text{Lip } f(t)(\cdot) \leq A$ ,  $\|f(t)(\cdot)\|_\infty \leq B$  for almost every  $t < 1$ .

*Proof.* The proof is based on the approximative solutions of Lemma 3.12, of course.

Indeed, for each index  $n \in \mathbb{N}$ , Lemma 3.12 provides  $K_n(\cdot) : [0, 1] \rightarrow \mathcal{K}(\mathbb{R}^N)$  and  $f_n(\cdot) : [0, 1[ \rightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  corresponding to  $\varepsilon := \frac{1}{n} e^{-A}$ . Obviously, they satisfy the properties (a)–(e) claimed here.

In particular, these features stay correct whenever we consider subsequences instead and again abbreviate them as  $(K_n(\cdot))_{n \in \mathbb{N}}$ ,  $(f_n(\cdot))_{n \in \mathbb{N}}$ , respectively.

For property (f) about uniform convergence of  $(K_n(\cdot))$  w.r.t.  $d$ . The  $B$ -Lipschitz continuity of each  $K_n(\cdot)$  has two important consequences:

1. all  $K_n(\cdot) : [0, 1] \rightarrow (\mathcal{K}(\mathbb{R}^N), d)$  ( $n \in \mathbb{N}$ ) are equicontinuous, and



2.  $\bigcup_{\substack{n \in \mathbb{N} \\ t \in [0,1]}} \{K_n(t)\}$  is contained in the compact subset  $\mathbb{B}_B(K_0)$  of  $(\mathcal{K}(\mathbb{R}^N), \mathcal{d})$ .

So, Proposition 3.4 of Arzelà–Ascoli provides a subsequence (again denoted by)  $(K_n(\cdot))_n$  converging uniformly to a continuous function  $K(\cdot) : [0, 1] \longrightarrow (\mathcal{K}(\mathbb{R}^N), \mathcal{d})$ . In particular,  $K(\cdot)$  is also  $B$ -Lipschitz continuous with  $K(0) = K_0$ ; i.e., properties (a)–(f) are fulfilled completely.

For property (g) about weak convergence of  $f_n(\cdot)|_{\tilde{K}}$  with fixed compact  $\tilde{K} \subset \mathbb{R}^N$ . We can no longer follow the same track as for differential inclusions. Indeed, the functions  $f_n(\cdot)$  of shape transitions have their values in  $\text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ , which cannot be regarded as a dual space in an obvious way. So Alaoglu’s theorem (stating that closed balls of dual Banach spaces are weakly\* compact) cannot be applied similarly to differential inclusions.

Alternatively, we restrict our considerations to a compact neighborhood  $\tilde{K}$  of  $\bigcup_{\substack{n \in \mathbb{N} \\ t \in [0,1]}} K_n(t) \subset \mathbb{R}^N$  and use a sufficient condition on relatively weakly compact sets in  $L^1([0, 1], C^0(\tilde{K}, \mathbb{R}^N))$ . Here  $C^0(\tilde{K}, \mathbb{R}^N)$  (supplied with the supremum norm  $\|\cdot\|_\infty$ ) denotes the Banach space of all continuous functions  $\tilde{K} \longrightarrow \mathbb{R}^N$ .

In fact, the set  $\{f_n(t) \mid n \in \mathbb{N}, t \in [0, 1]\} \subset C^0(\mathbb{R}^N, \mathbb{R}^N)$  is uniformly bounded by  $B$  and equicontinuous (due to property (c)). So according to Proposition 3.4 of Arzelà–Ascoli, the set of their restrictions to  $\tilde{K}$

$$W := \left\{ f_n(t)|_{\tilde{K}} \mid n \in \mathbb{N}, t \in [0, 1] \right\} \subset C^0(\tilde{K}, \mathbb{R}^N)$$

is relatively compact w.r.t.  $\|\cdot\|_\infty$ . Thus,  $\{f_n(\cdot)|_{\tilde{K}} \mid n \in \mathbb{N}\}$  is relatively weakly compact in  $L^1([0, 1], C^0(\tilde{K}, \mathbb{R}^N))$  according to Proposition 3.7, and we obtain a subsequence (again denoted by)  $(f_n(\cdot))_{n \in \mathbb{N}}$  and some  $g(\cdot) \in L^1([0, 1], C^0(\tilde{K}, \mathbb{R}^N))$  with  $f_n(\cdot)|_{\tilde{K}} \longrightarrow g(\cdot)$  weakly in  $L^1([0, 1], C^0(\tilde{K}, \mathbb{R}^N))$ .

Obviously, both the subsequence and  $g(\cdot)$  depend on  $\tilde{K}$ , however.

For property (g) about  $f_n(\cdot)|_{\tilde{K}_j}$  with every compact  $\tilde{K}_j \subset \mathbb{R}^N$  ( $j \in \mathbb{N}$ ). Now this construction of subsequences is applied to the compact subsets  $\tilde{K}_j \stackrel{\text{def.}}{=} \mathbb{B}_{j+B}(K_0)$  of  $\mathbb{R}^N$  for  $j = 1, 2, 3 \dots$  successively. By means of Cantor’s diagonal construction, we obtain a subsequence (again denoted by)  $(f_n(\cdot))_{n \in \mathbb{N}}$  and a function  $g_j(\cdot) \in L^1([0, 1], C^0(\tilde{K}_j, \mathbb{R}^N))$  (for each  $j \in \mathbb{N}$ ) such that for each  $j \in \mathbb{N}$ ,

$$f_n(\cdot)|_{\tilde{K}_j} \longrightarrow g_j(\cdot) \quad (n \longrightarrow \infty) \quad \text{weakly in } L^1([0, 1], C^0(\tilde{K}_j, \mathbb{R}^N)).$$

As restrictions to  $\tilde{K}_j$  of one and the same subsequence  $(f_n(\cdot))_{n \in \mathbb{N}}$  converge weakly for each  $j \in \mathbb{N}$ , the inclusion  $\tilde{K}_j \subset \tilde{K}_{j+1}$  implies that for any indices  $j < k$

$$g_j(t)(\cdot) = g_k(t)(\cdot)|_{\tilde{K}_j} \in C^0(\tilde{K}_j, \mathbb{R}^N) \quad \text{for a.e. } t \in [0, 1],$$

and so  $(g_j(\cdot))_{j \in \mathbb{N}}$  induces a single function  $f : [0, 1] \longrightarrow C^0(\mathbb{R}^N, \mathbb{R}^N)$  defined as

$$f(t)(x) := g_j(t)(x) \quad \text{for } x \in \tilde{K}_j \text{ and a.e. } t \in [0, 1].$$

For property (h) about Lipschitz continuity and bounds of limit function  $f(\cdot)$ . Finally, we verify  $f(t) \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ ,  $\text{Lip } f(t) \leq A$ , and  $\|f(t)\|_\infty \leq B$  for almost every  $t \in [0, 1]$ . Indeed, as in the case of differential inclusions (section 3.1), Mazur’s lemma in Proposition 3.5 ensures here for each  $j \in \mathbb{N}$  (fixed) that

$$f(\cdot)|_{\tilde{K}_j} \in \bigcap_{n \in \mathbb{N}} \overline{\text{co}} \{f_n(\cdot)|_{\tilde{K}_j}, f_{n+1}(\cdot)|_{\tilde{K}_j} \dots\} \quad \text{in } L^1([0, 1], C^0(\tilde{K}_j, \mathbb{R}^N)).$$

Thus,  $f(\cdot)|_{\widetilde{K}_j}$  can be approximated by convex combinations of  $\{f_1(\cdot)|_{\widetilde{K}_j}, f_2(\cdot)|_{\widetilde{K}_j}, \dots\}$  w.r.t. the  $L^1$  norm. A further subsequence (of these convex combinations) converges to  $f(\cdot)|_{\widetilde{K}_j}$  almost everywhere in  $[0, 1]$ . So, for almost every  $t \in [0, 1]$ ,  $f(t)|_{\widetilde{K}_j}$  belongs to the same compact convex subset of  $(C^0(\widetilde{K}_j, \mathbb{R}^N), \|\cdot\|_\infty)$  as  $f_1(t)|_{\widetilde{K}_j}, f_2(t)|_{\widetilde{K}_j}, \dots$ , namely,  $\{w \in \text{Lip}(\widetilde{K}_j, \mathbb{R}^N) \mid \text{Lip } w \leq A, \|w\|_\infty \leq B\}$ .  $\square$

LEMMA 3.14 (the limit function is a solution). *Under the assumptions of the viability theorem, Theorem 3.11, consider both  $K_n(\cdot), K(\cdot) : [0, 1] \longrightarrow \mathcal{K}(\mathbb{R}^N)$  ( $n \in \mathbb{N}$ ) and  $f_n(\cdot), f(\cdot) : [0, 1] \longrightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  specified in Lemma 3.13.*

*Then  $K(\cdot)$  is a solution of the morphological inclusion  $\dot{K}(\cdot) \cap \mathcal{F}(K(\cdot)) \neq \emptyset$  with  $K(0) = K_0$  and  $K(t) \in \mathcal{V}$  for all  $t \in [0, 1]$ .*

*Proof.*  $K(t) \in \mathcal{V}$  for all  $t \in [0, 1]$  results directly from properties (d) and (f) of Lemma 3.13 because  $\mathcal{V}$  is assumed to be a closed subset of  $(\mathcal{K}(\mathbb{R}^N), \mathcal{d})$ .

So it remains to prove  $f(t) \in \dot{K}(\cdot) \cap \mathcal{F}(K(\cdot))$  for Lebesgue-almost every  $t \in [0, 1]$ .

The Carathéodory property of each  $f_n(\cdot)$  and every  $K_n(\cdot)$  is a reachable set. As each  $f_n : [0, 1] \longrightarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  from Lemma 3.13 is piecewise constant, it can be regarded as a measurable/Lipschitz function  $[0, 1] \times \mathbb{R}^N \longrightarrow \mathbb{R}^N$ ,  $(t, x) \longmapsto f_n(t)(x)$  in the sense of [7, Definition 9.5.1], i.e.,

$$\begin{aligned} f_n(\cdot)(x) : [0, 1] &\longrightarrow \mathbb{R}^N \quad \text{is Lebesgue measurable} && \text{for every } x \in \mathbb{R}^N, \\ f_n(t)(\cdot) : \mathbb{R}^N &\longrightarrow \mathbb{R}^N \quad \text{is } A\text{-Lipschitz continuous} && \text{for every } t \in [0, 1]. \end{aligned}$$

In addition,  $\|f_n(t)(\cdot)\|_\infty \leq B$  for every  $t \in [0, 1]$ ,  $n \in \mathbb{N}$ .

Moreover, each compact set  $K_n(t) \subset \mathbb{R}^N$  coincides with the reachable set

$$\vartheta_{f_n}(t, K_0) \stackrel{\text{Def.}}{=} \{x(t) \mid \exists x \in W^{1,1}([0, t], \mathbb{R}^N) : x'(s) = f_n(s)(x(s)) \text{ for a.e. } s, x(0) \in K_0\}$$

of the initial set  $K_0$  and the function  $f_n(\cdot)(\cdot) : [0, 1] \times \mathbb{R}^N \longrightarrow \mathbb{R}^N$ . Indeed, consider a subinterval  $[s_1, s_2] \subset [0, 1]$  in which  $f_n(\cdot)$  is constant, i.e.,  $f_n(\cdot)|_{[s_1, s_2]} \equiv g \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ , and assume  $K_n(s_1) = \vartheta_{f_n}(s_1, K_0)$  (with subsequent induction in mind). Then both  $K_n(\cdot)$  and the reachable set  $\vartheta_{f_n}(\cdot, K_0)$  satisfy the morphological equation  $\dot{Q}(\cdot) \ni g$  in  $[s_1, s_2]$  and are  $B$ -Lipschitz continuous. So due to Theorem 2.7,  $K_n(\cdot) \equiv \vartheta_{f_n}(\cdot, K_0)$  in  $[s_1, s_2]$ . By means of induction, we conclude  $K_n(\cdot) \equiv \vartheta_{f_n}(\cdot, K_0)$  in  $[0, 1]$ .

For characterizing  $K(t)$  as a reachable set of  $f(\cdot)$ .  $K(t) \subset \vartheta_f(t, K_0)$  for every  $t$ . Indeed, Lemma 3.13 (f) implies the characterization as a limit with respect to  $\mathcal{d}$  (or, equivalently for compact sets here, in the sense of Painlevé–Kuratowski):

$$K(t) = \lim_{n \rightarrow \infty} K_n(t) = \lim_{n \rightarrow \infty} \vartheta_{f_n}(t, K_0).$$

For every  $x \in K(t)$ , there is a sequence  $(x_n(\cdot))_{n \in \mathbb{N}}$  of functions in  $W^{1,1}([0, t], \mathbb{R}^N)$  satisfying

$$\begin{cases} x'_n(s) = f_n(s)(x_n(s)) & \text{for a.e. } s \in [0, t], \\ x_n(0) \in K_0, \\ x_n(s) \in \vartheta_{f_n}(s, K_0) \subset \mathbb{B}_{1+B}(K_0) \stackrel{\text{Def.}}{=} \widetilde{K}_1 & \text{for each } s \in [0, t], \\ x_n(t) \longrightarrow x & \text{for } n \longrightarrow \infty. \end{cases}$$

Seizing the notions of [2, Convergence Theorem 2.4.4], the theorems of Arzelà–Ascoli and Alaoglu provide a subsequence  $(x_{n_j}(\cdot))_{j \in \mathbb{N}}$  and functions  $x(\cdot) \in C^0([0, t], \mathbb{R}^N)$ ,  $v(\cdot) \in L^1([0, t], \mathbb{R}^N)$  such that

$$x_{n_j}(\cdot) \longrightarrow x(\cdot) \quad \text{uniformly in } [0, t], \quad x'_{n_j}(\cdot) \longrightarrow v(\cdot) \quad \text{weakly in } L^1([0, t], \mathbb{R}^N),$$

implying the absolute continuity of  $x(\cdot)$  with  $x'(\cdot) = v(\cdot)$ .

For verifying  $x'(\cdot) = f(\cdot)(x(\cdot))$  (almost everywhere), we now prove  $f_{n_j}(\cdot)(x_{n_j}(\cdot)) \longrightarrow f(\cdot)(x(\cdot))$  weakly in  $L^1([0, t], \mathbb{R}^N)$  for  $j \longrightarrow \infty$ . For any  $g \in L^\infty([0, t], \mathbb{R}^N) \cong (L^1([0, t], \mathbb{R}^N))^*$ , the  $A$ -Lipschitz continuity of each  $f_{n_j}(s)$  implies

$$\int_0^t g(s)^T f_{n_j}(s)(x_{n_j}(s)) \, ds \in \int_0^t g(s)^T f_{n_j}(s)(x(s)) \, ds + c \|x(\cdot) - x_{n_j}(\cdot)\|_\infty \mathbb{B}_1.$$

As  $L^1([0, 1], C^0(\tilde{K}_1, \mathbb{R}^N)) \longrightarrow \mathbb{R}$ ,  $h \longmapsto \int_0^t g(s)^T h(s)(x(s)) \, ds$  is continuous,

$$\int_0^t g(s)^T f_{n_j}(s)(x_{n_j}(s)) \, ds \longrightarrow \int_0^t g(s)^T f(s)(x(s)) \, ds \text{ for } j \longrightarrow \infty.$$

Thus,  $x = x(t) \in \vartheta_f(t, K_0)$ .

For characterizing  $K(t)$  as a reachable set of  $f(\cdot)$ .  $\vartheta_f(t, K_0) \subset K(t)$  for every  $t$ . The next step is to verify that the tube  $K(\cdot) : [0, 1] \rightsquigarrow \mathbb{R}^N$  is invariant under  $f$ ; i.e., for every initial point  $x \in K(t)$  (with  $t \in [0, 1]$ ), the solution  $x(\cdot) \in W^{1,1}([t, 1], \mathbb{R}^N)$  of  $x'(\cdot) = f(\cdot)(x(\cdot))$  (almost everywhere) with  $x(t) = x$  satisfies  $x(\tau) \in K(\tau)$  for any  $\tau \in [t, 1]$ . Due to  $K(0) = K_0$ , this property implies  $\vartheta_f(t, K_0) \subset K(t)$  for every  $t \in [0, 1]$ .

Indeed, existence and uniqueness of this solution  $x(\cdot)$  result from (generalized) Filippov's theorem (see, e.g., [35, Theorem 2.4.3]) since  $[0, 1] \times \mathbb{R}^N$ ,  $(s, y) \longmapsto f(s)(y)$  is measurable/Lipschitz (in the sense of [7, Definition 9.5.1]). Each  $x \in K(t) = \text{Lim}_{n \rightarrow \infty} K_n(t)$  is the limit of a sequence  $(x_n)_{n \in \mathbb{N}}$  with  $x_n \in K_n(t)$ , and there exist corresponding solutions  $x_n(\cdot) \in W^{1,1}([t, 1], \mathbb{R}^N)$  ( $n \in \mathbb{N}$ ) of  $x'_n(\cdot) = f_{n_j}(\cdot)(x_n(\cdot))$  (almost everywhere) with  $x_n(t) = x_n$ . For the same reasons as before, we obtain a subsequence  $(x_{n_j}(\cdot))_{j \in \mathbb{N}}$  and a limit function  $y(\cdot) \in W^{1,1}([t, 1], \mathbb{R}^N)$  satisfying

$$\begin{cases} x_{n_j}(\cdot) \longrightarrow y(\cdot) & \text{uniformly in } [t, 1], \\ x'_{n_j}(\cdot) \longrightarrow y'(\cdot) & \text{weakly in } L^1([t, 1], \mathbb{R}^N), \\ f_{n_j}(\cdot)(x_{n_j}(\cdot)) \longrightarrow f(\cdot)(y(\cdot)) & \text{weakly in } L^1([t, 1], \mathbb{R}^N). \end{cases}$$

So  $y(\cdot)$  is identical to the uniquely determined solution  $x(\cdot)$  of  $x'(\cdot) = f(\cdot)(x(\cdot))$  (almost everywhere) with  $x(t) = x$ ; i.e., the limit  $y(\cdot)$  does not depend on the selection of the subsequence  $(x_{n_j}(\cdot))_{j \in \mathbb{N}}$ . This implies indirectly that even the whole sequence  $(x_n(\cdot))_{n \in \mathbb{N}}$  converges to  $x(\cdot)$  uniformly and all its derivatives tend weakly to  $x'(\cdot)$ . In particular,  $x(\tau) = \lim_{n \rightarrow \infty} x_n(\tau) \in \text{Lim}_{n \rightarrow \infty} K_n(\tau) = K(\tau)$  for every  $\tau \in [t, 1]$ .

Thus,  $K(t) = \vartheta_f(t, K_0)$  for every  $t \in [0, 1]$ .

$K(t)$  as a reachable set and Scorza-Dragoni ensure solution property at almost every time. That is, describing  $K(t)$  as a reachable set of  $f(\cdot)(\cdot) : [0, 1] \times \mathring{\mathbb{R}}^N \longrightarrow \mathbb{R}^N$  implies that  $f(t) \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  belongs to the shape mutation  $\mathring{K}(t)$  for almost every  $t \in [0, 1]$ :

$$\lim_{h \downarrow 0} \frac{1}{h} \cdot d(\vartheta_{f(t)}(h, K(t)), K(t+h)) = 0.$$

Indeed,  $f(\cdot)(\cdot) : [0, 1] \times \mathbb{R}^N \longrightarrow \mathbb{R}^N$  is measurable/Lipschitz, and thus Proposition 3.9 ensures the following (slightly modified) Scorza-Dragoni property:

For any  $\varepsilon > 0$ , there exists a closed subset  $J_\varepsilon \subset [0, 1 - \varepsilon]$  with  $\mathcal{L}^1([0, 1] \setminus J_\varepsilon) < 2\varepsilon$  such that the restriction of  $f(\cdot)(\cdot)$  to  $J_\varepsilon \times \mathbb{R}^N$  is continuous. Seizing an idea in the proof

of [20, Lemma 2.6], let  $\tilde{J}_\varepsilon$  be the subset of all density points of  $J_\varepsilon$  that are also Lebesgue points of the characteristic function  $\chi_{[0,1] \setminus J_\varepsilon}(\cdot)$ . Then  $\mathcal{L}^1(\tilde{J}_\varepsilon) = \mathcal{L}^1(J_\varepsilon) > 1 - 2\varepsilon$ , because Lebesgue points of each  $\mathcal{L}^1$  function always have full Lebesgue measure [37, Theorem 1.3.8] and so, in particular, density points of any measurable set also have full Lebesgue measure.

For any  $t \in \tilde{J}_\varepsilon$ ,  $x \in K(t)$ , there exist unique solutions  $x(\cdot)$ ,  $y(\cdot) \in \text{Lip}([t, 1], \mathbb{R}^N)$  of

$$x'(\cdot) = f(\cdot)(x(\cdot)), \quad y'(\cdot) = f(t)(y(\cdot)) \quad \text{a.e. in } [t, 1],$$

respectively, with  $x(t) = x = y(t)$ . Then we obtain for every  $\tau \in ]t, 1]$

$$\begin{aligned} |x(\tau) - y(\tau)| &= \left| \int_t^\tau (f(s)(x(s)) - f(t)(y(s))) \, ds \right| \\ &\leq \left| \int_{[t, \tau] \cap J_\varepsilon} (f(s)(x(s)) - f(t)(y(s))) \, ds \right| + 2B \cdot \mathcal{L}^1([t, \tau] \setminus J_\varepsilon) \\ &\leq \int_{[t, \tau] \cap J_\varepsilon} |f(s)(x(s)) - f(t)(x(s))| \, ds + 2B \cdot \mathcal{L}^1([t, \tau] \setminus J_\varepsilon) \\ &\quad + \int_{[t, \tau] \cap J_\varepsilon} A \cdot |x(s) - y(s)| \, ds. \end{aligned}$$

For  $\delta > 0$  arbitrarily small and each  $t \in \tilde{J}_\varepsilon$ , the construction of  $J_\varepsilon$  (in regard to continuity of  $f(\cdot)(\cdot)|_{J_\varepsilon \times \mathbb{R}^N}$ ) and the definition of  $\tilde{J}_\varepsilon$  (in regard to Lebesgue points of  $\chi_{[0,1] \setminus J_\varepsilon}(\cdot)$ ) provide some  $T \in ]t, 1]$  satisfying

$$\begin{cases} \sup_{s \in [t, T] \cap J_\varepsilon} \sup_{z: \text{dist}(z, K_0) \leq B} |f(s)(z) - f(t)(z)| < \delta, \\ \sup_{s \in [t, T]} \frac{1}{|s - t|} \cdot \mathcal{L}^1([t, s] \setminus J_\varepsilon) < \delta; \end{cases}$$

thus,  $|x(\tau) - y(\tau)| \leq A \int_{[t, \tau]} |x(s) - y(s)| \, ds + \delta (1 + 2B) (\tau - t)$  for any  $\tau \in ]t, T]$ .

Gronwall's lemma implies  $|x(\tau) - y(\tau)| \leq \delta (1 + 2B) e^{A \cdot (\tau - t)} (\tau - t)$  for any  $\tau$ .

As  $x \in K(t)$  is chosen arbitrarily and  $T$  does not depend on  $x$  (but only on  $\delta, \varepsilon, t$ ), the reachable sets  $\vartheta_{f(t)}(h, K(t))$  and  $K(t + h) = \vartheta_f(h, K(t))$  satisfy for any  $h \in [0, T - t]$

$$d(\vartheta_{f(t)}(h, K(t)), K(t + h)) \leq \delta (1 + 2B) e^{A h} h,$$

i.e.,

$$\lim_{h \downarrow 0} \frac{1}{h} \cdot d(\vartheta_{f(t)}(h, K(t)), K(t + h)) = 0 \quad \text{for every } t \in \tilde{J}_\varepsilon.$$

Finally,  $f(t) \in \mathcal{F}(K(t))$  for almost every  $t \in [0, 1]$ . Due to Lemma 3.13 (e), (g),  $f_n(\cdot)|_{\tilde{K}_j} \rightarrow f(\cdot)|_{\tilde{K}_j}$  weakly in  $L^1([0, 1], C^0(\tilde{K}_j, \mathbb{R}^N))$  for each compact set  $\tilde{K}_j := \mathbb{B}_{j+B}(K_0)$  ( $j \in \mathbb{N}$ ), and  $f_n(t) \in \mathcal{F}(\mathbb{B}_{1/n}(K_n(t)))$  for every  $n \in \mathbb{N}$ ,  $t \in [0, 1]$ .

Fixing the index  $j \in \mathbb{N}$  of compact sets arbitrarily, Proposition 3.8 provides a sequence  $(h_{j,n}(\cdot))_{n \in \mathbb{N}}$  with  $h_{j,n}(\cdot) \in \text{co}\{f_n(\cdot)|_{\tilde{K}_j}, f_{n+1}(\cdot)|_{\tilde{K}_j} \dots\} \subset L^1([0, 1], C^0(\tilde{K}_j, \mathbb{R}^N))$  such that for  $\mathcal{L}^1$  almost every  $t \in [0, 1]$ ,

$$h_{j,n}(t) \rightarrow f(t)|_{\tilde{K}_j} \quad (n \rightarrow \infty) \quad \text{weakly in } C^0(\tilde{K}_j, \mathbb{R}^N).$$

Proposition 3.10 and assumption 2 of Theorem 3.11, i.e.,

$$\sup_{M \in \mathcal{K}(\mathbb{R}^N)} \sup_{g \in \mathcal{F}(M)} \text{Lip } g \leq A < \infty, \quad \sup_{M \in \mathcal{K}(\mathbb{R}^N)} \sup_{g \in \mathcal{F}(M)} \|g\|_\infty \leq B < \infty,$$

imply  $h_{j,n}(t) \rightarrow f(t)|_{\widetilde{K}_j}$  uniformly in  $\widetilde{K}_j$  for  $n \rightarrow \infty$  and almost every  $t \in [0, 1[$ .

Let  $C_j \subset [0, 1[$  denote the set of full measure for which this uniform convergence holds. Then  $C := \bigcap_{j \in \mathbb{N}} C_j \subset [0, 1[$  is also a set of full measure, i.e.,  $\mathcal{L}^1([0, 1] \setminus C) = 0$ .

Choose  $t \in C$  arbitrarily. Then for each  $j \in \mathbb{N}$ , there exists an index  $n_j > j$  such that  $n_j > n_{j-1}$  and  $\|h_{j,n_j}(\cdot)|_{\widetilde{K}_j} - f(\cdot)|_{\widetilde{K}_j}\|_\infty < \frac{1}{j}$ .

Due to  $h_{j,n}(\cdot) \in co\{f_n(\cdot)|_{\widetilde{K}_j}, f_{n+1}(\cdot)|_{\widetilde{K}_j} \dots\}$ , each  $h_{j,n_j}(t)|_{\widetilde{K}_j}$  has a continuation to  $\mathbb{R}^N$  in  $co\{f_n(t), f_{n+1}(t) \dots\} \subset C^0(\mathbb{R}^N, \mathbb{R}^N)$  (that again is denoted by  $h_{j,n_j}(t)$ ), and  $h_{j,n_j}(t) \rightarrow f(t)$  locally uniformly in  $\mathbb{R}^N$  for  $j \rightarrow \infty$ .

Furthermore,  $co\{f_n(t), f_{n+1}(t) \dots\} \subset \overline{co} \mathcal{F}(\mathbb{B}_{1/n}(\bigcup_{m \geq n} K_m(t)))$ .

So finally,  $d(K_n(t), K(t)) \rightarrow 0$  and assumption 3 of Theorem 3.11 about the closed graph of  $\mathcal{F}$  (with its convex values) imply  $f(t) \in \mathcal{F}(K(t))$ .  $\square$

**4. An application to shape optimization under state constraints.** Let  $J : \mathcal{K}(\mathbb{R}^N) \rightarrow \mathbb{R}$  be a functional that is Lipschitz continuous w.r.t. the Pompeiu–Hausdorff distance  $d$ . Moreover,  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  denotes a nonempty closed set of constraints. Detecting a minimizer  $\widehat{K} \in \mathcal{V}$  of the optimization problem

$$\inf \{J(K) \mid K \in \mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)\}$$

usually proves to be rather complicated ([12, 11]; see [18, 25, 38] supplementarily). Thus, we prefer here to isolate candidates (for a minimizer) constructively by means of a necessary condition (similarly to [24]). The viability theorem, Theorem 3.11, for morphological inclusions then lays the basis for a curve  $K(\cdot) : [0, \infty[ \rightarrow \mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  such that

- (i)  $J \circ K : [0, \infty[ \rightarrow \mathbb{R}$ ,  $t \mapsto J(K(t))$  is nonincreasing, and
- (ii) every compact set  $C = \text{Lim}_{n \rightarrow \infty} K(t_n) \in \mathcal{V}$  (for some sequence  $t_n \nearrow \infty$ ) satisfies a necessary condition on minimizers (in the form of Fermat’s rule).

The first step is to specify a map  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  satisfying the following conditions on its values:

- 1. all values of  $\mathcal{F}$  are convex and closed (w.r.t. locally uniform convergence);
- 2.  $\sup_{M \in \mathcal{K}(\mathbb{R}^N)} \sup_{f \in \mathcal{F}(M)} (\|f\|_\infty + \text{Lip } f) < \infty$ ,

Essentially, the choice of  $\mathcal{F}$  is to guarantee that the composition  $t \mapsto J(K(t))$  is nonincreasing for every compact-valued solution  $K(\cdot) : [0, 1] \rightsquigarrow \mathbb{R}^N$  of the morphological inclusion  $\overset{\circ}{K}(\cdot) \cap \mathcal{F}(K(\cdot)) \neq \emptyset$ . For combining this aim with the conditions on its values, we do not use the so-called *shape semiderivative* of  $J$  in direction  $v(\cdot)$ ,

$$\delta J(K)(v) := \lim_{h \downarrow 0} \frac{1}{h} \cdot (J(\vartheta_v(h, K)) - J(K)),$$

for  $K \in \mathcal{K}(\mathbb{R}^N)$  (assuming the limit to exist) as in [12, section 7.2], [11], but we prefer the notion of Clarke’s generalized directional derivative in a Banach space (see, e.g., [10, section 1.2]) and extend it to shape transitions.

**DEFINITION 4.1.** Let  $J : (\mathcal{K}(\mathbb{R}^N), d) \rightarrow \mathbb{R}$  be Lipschitz continuous.

Clarke’s generalized shape derivative of  $J$  at  $K \in \mathcal{K}(\mathbb{R}^N)$  in direction  $v \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  is defined as

$$\delta^C J(K)(v) := \limsup_{\substack{h \downarrow 0, M \rightarrow K \\ (M \in \mathcal{K}(\mathbb{R}^N))}} \frac{1}{h} \cdot (J(\vartheta_v(h, M)) - J(M)).$$

Set  $\iota J(K) := \limsup_{\substack{M \rightarrow K \\ (M \in \mathcal{K}(\mathbb{R}^N))}} \inf \{ \delta^C J(M)(v) \mid v \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N), \|v\|_\infty + \text{Lip} v \leq 1 \}$ .

*Remark.* 1. Let  $\Lambda \geq 0$  denote the Lipschitz constant of  $J : (\mathcal{K}(\mathbb{R}^N), d) \rightarrow \mathbb{R}$ . Then, due to Lemma 2.4,  $|J(\vartheta_v(h, K)) - J(K)| \leq \Lambda \cdot d(\vartheta_v(h, K), K) \leq \Lambda \cdot \|v\|_\infty h$  for every  $v \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ , and thus  $|\delta^C J(K)(v)| \leq \Lambda \|v\|_\infty$ ,  $\iota J(K) \geq -\Lambda$ .

In particular,  $\delta^C J(K)(0) = 0$  for every  $K \in \mathcal{K}(\mathbb{R}^N)$ .

2.  $\iota J(\cdot) : \mathcal{K}(\mathbb{R}^N) \rightarrow \mathbb{R}$  is the upper semicontinuous envelope of the minimal generalized shape derivative  $\delta^C J(\cdot)(v)$  over all  $v$  in the unit ball of  $\text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ . Furthermore,  $\iota J(K) \leq \delta^C J(K)(0) = 0$  for all  $K \in \mathcal{K}(\mathbb{R}^N)$ .

DEFINITION 4.2. Using the notation of Definition 4.1, set  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ ,

$$\mathcal{F}(K) := \left\{ v \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N) \mid \|v\|_\infty + \text{Lip} v \leq 1, \quad \delta^C J(K)(v) \leq \frac{1}{2} \cdot \iota J(K) \right\}.$$

Here both the bound 1 and the factor  $\frac{1}{2}$  are rather arbitrary and can be replaced by any constants in  $[0, \infty[$  and  $]0, 1[$ , respectively. We show in Lemma 4.5 below that all values of the set-valued map  $\mathcal{F}$  are nonempty, convex, and closed w.r.t. locally uniform convergence.

LEMMA 4.3. Let  $J : (\mathcal{K}(\mathbb{R}^N), d) \rightarrow \mathbb{R}$  be Lipschitz continuous and  $K(\cdot) : [0, 1] \rightsquigarrow \mathbb{R}^N$  any compact-valued Lipschitz solution of the morphological inclusion  $\mathring{K}(\cdot) \cap \mathcal{F}(K(\cdot)) \neq \emptyset$ , with  $\mathcal{F}(\cdot)$  introduced in Definition 4.2. Then we have the following:

1.  $J \circ K : [0, 1] \rightarrow \mathbb{R}$ ,  $t \mapsto J(K(t))$  is nonincreasing.
2. Suppose  $\inf_{\mathcal{K}(\mathbb{R}^N)} J(\cdot) > -\infty$  in addition, and let  $C$  belong to the  $\omega$ -limit set of  $K(\cdot)$  in  $\mathcal{K}(\mathbb{R}^N)$ , i.e.,  $d(K(t_n), C) \rightarrow 0$  for some  $t_n \nearrow \infty$ . Then  $\iota J(C) = 0$ .

*Proof.* 1. The proof of item 1 results from the Lipschitz continuity of the composition  $[0, 1] \rightarrow \mathbb{R}$ ,  $t \mapsto J(K(t))$  and the definition of  $\mathcal{F}$ . Indeed, at almost every time  $t \in [0, 1]$ , there exists  $v \in \mathcal{F}(K(t)) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  with  $\lim_{h \downarrow 0} \frac{1}{h} \cdot d(K(t+h), \vartheta_v(h, K(t))) = 0$ . Thus,

$$\begin{aligned} & \limsup_{h \downarrow 0} \frac{1}{h} \cdot (J(K(t+h)) - J(K(t))) \\ & \leq \limsup_{h \downarrow 0} \frac{1}{h} \cdot \left( J(\vartheta_v(h, K(t))) - J(K(t)) + \text{Lip} J \cdot d(\vartheta_v(h, K(t)), K(t+h)) \right) \\ & \leq \delta^C J(K(t))(v) + 0 \\ & \leq \frac{1}{2} \cdot \iota J(K(t)) \quad (\text{since } v \in \mathcal{F}(K(t))) \\ & \leq 0 \quad (\text{due to preceding remark (item 2)}). \end{aligned}$$

2. Assume the contrary, i.e.,  $\kappa := \iota J(C) < 0$ . Then there exists some small  $\rho > 0$  such that all sets  $M \in \mathcal{K}(\mathbb{R}^N)$  with  $d(M, C) \leq 2\rho$  satisfy  $\iota J(M) < \frac{\kappa}{2} < 0$ . For all  $n \in \mathbb{N}$  sufficiently large,  $K(t_n)$  has the Pompeiu–Hausdorff distance  $< \rho$  from  $C$ . As all values of  $\mathcal{F}$  contain only functions  $v \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  with  $\|v\|_\infty + \text{Lip} v \leq 1$ , every compact-valued solution  $K(\cdot)$  of the morphological inclusion  $\mathring{K}(\cdot) \cap \mathcal{F}(K(\cdot)) \neq \emptyset$  has Lipschitz constant  $\leq 1$  (w.r.t.  $d$ ) similarly to Lemma 2.4.

Thus,  $d(K(s), C) < 2\rho$  for all  $s \in [t_n - \rho, t_n + \rho]$ . Now the same approach of estimating as in Lemma 4.3 (1) implies  $J(K(s_2)) \leq J(K(s_1)) + \frac{\kappa}{4} \cdot (s_2 - s_1)$  for all  $t_n - \rho \leq s_1 \leq s_2 \leq t_n + \rho$  and large  $n \in \mathbb{N}$ —contradicting the hypothesis  $\inf J(\cdot) > -\infty$ .  $\square$

The next lemma will be used for verifying the convexity of all values of  $\mathcal{F}$  in Lemma 4.5.

LEMMA 4.4. For every  $\lambda \in ]0, 1[$ , there exists  $\mu \in L^1([0, 1])$  satisfying

$$\frac{1}{t} \cdot \int_0^t (\mu(s) - \lambda) \, ds \longrightarrow 0 \quad (t \downarrow 0), \quad \mu(\cdot) \in \{0, 1\}$$

piecewise constant in  $]0, 1[$ .

*Proof.*  $\mu(\cdot)$  is defined to be piecewise constant in each interval  $[\frac{1}{\sqrt{n+1}}, \frac{1}{\sqrt{n}}[$  ( $n \in \mathbb{N}$ ). Set

$$\mu(t) := \begin{cases} 0 & \text{for } \frac{1}{\sqrt{n+1}} \leq t < \frac{\lambda}{\sqrt{n+1}} + \frac{1-\lambda}{\sqrt{n}}, \\ 1 & \text{for } \frac{\lambda}{\sqrt{n+1}} + \frac{1-\lambda}{\sqrt{n}} \leq t < \frac{1}{\sqrt{n}} \end{cases}$$

for each  $n \in \mathbb{N}$ . Then

$$\int_{\frac{1}{\sqrt{n+1}}}^{\frac{1}{\sqrt{n}}} (\mu(s) - \lambda) \, ds = 0,$$

and thus

$$\int_0^{\frac{1}{\sqrt{n}}} (\mu(s) - \lambda) \, ds = 0.$$

Moreover,

$$\int_{\frac{1}{\sqrt{n+1}}}^{\frac{1}{\sqrt{n}}} |\mu(s) - \lambda| \, ds = 2\lambda(1-\lambda) \left( \frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \right)$$

implies

$$\sup_{\frac{1}{\sqrt{n+1}} \leq t \leq \frac{1}{\sqrt{n}}} \frac{1}{t} \cdot \left| \int_0^t (\mu(s) - \lambda) \, ds \right| \leq \sqrt{n+1} \cdot \int_{\frac{1}{\sqrt{n+1}}}^{\frac{1}{\sqrt{n}}} |\mu(s) - \lambda| \, ds \xrightarrow{n \rightarrow \infty} 0. \quad \square$$

LEMMA 4.5. Consider  $\text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  with the topology of locally uniform convergence. All values of the set-valued map  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  introduced in Definition 4.2 are nonempty, convex, and closed.

*Proof.* For every  $K \in \mathcal{K}(\mathbb{R}^N)$ , the value  $\mathcal{F}(K)$  is a nonempty because either

- $\iota J(K) = 0$ , and then  $0 \in \mathcal{F}(K)$ , or
- $\iota J(K) < 0$ , and then there is  $v \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  with  $\|v\|_\infty + \text{Lip } v \leq 1$ ,  $\delta^C J(K)(v) \leq \frac{3}{4} \cdot \iota J(K) < 0$  (due to the definition of infimum); i.e.,  $v \in \mathcal{F}(K)$  induces a shape transition along which  $J(\cdot)$  is strictly decreasing for short times.

Each value  $\mathcal{F}(K) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  of  $\mathcal{F}$  is convex. Indeed, choose any  $v, w \in \mathcal{F}(K)$  and  $\lambda \in ]0, 1[$ . According to Lemma 4.4, there exists some  $\mu \in L^1([0, 1])$  satisfying

$$\frac{1}{t} \cdot \int_0^t (\mu(s) - \lambda) \, ds \longrightarrow 0 \quad (t \downarrow 0), \mu(\cdot) \in \{0, 1\} \text{ piecewise constant in } ]0, 1[.$$

Now we compare the evolution of an arbitrary set  $K \in \mathcal{K}(\mathbb{R}^N)$  along the autonomous Lipschitz field

$$u : \mathbb{R}^N \longrightarrow \mathbb{R}^N, \quad x \longmapsto \lambda \cdot v(x) + (1 - \lambda) \cdot w(x)$$

and along the nonautonomous vector field

$$g: \mathbb{R}^N \times [0, 1] \longrightarrow \mathbb{R}^N, \quad (x, t) \longmapsto \mu(t) \cdot v(x) + (1 - \mu(t)) \cdot w(x).$$

We verify  $d(\vartheta_u(t, M), \vartheta_g(t, M)) \leq o(t)$  for  $t \downarrow 0$  uniformly in  $M$ .

For any initial point  $x_0 \in \mathbb{R}^N$  given, let  $x(\cdot) \in C^1([0, 1], \mathbb{R}^N)$  denote the (unique) solution to  $x'(\cdot) = u(x(\cdot))$ ,  $x(0) = x_0$ . As each  $g(\cdot, t)$  is Lipschitz continuous (with Lipschitz constant  $\leq 1$ ), Filippov's theorem (applied to differential equations here) guarantees that the Cauchy problem

$$\begin{cases} y'(\cdot) = g(y(\cdot), \cdot) & \text{a.e. in } [0, 1], \\ y(0) = x_0 \in \mathbb{R}^N \end{cases}$$

has a solution  $y(\cdot) \in W^{1,1}([0, t], \mathbb{R}^N)$ . These solutions  $x(\cdot), y(\cdot)$  always satisfy

$$\begin{aligned} & |x(t) - y(t)| \\ &= \left| \int_0^t \left( \lambda v(x(s)) - \mu(s) v(y(s)) + (1 - \lambda) w(x(s)) - (1 - \mu(s)) w(y(s)) \right) ds \right| \\ &\leq \left| \int_0^t \left( (\lambda - \mu(s)) v(x(s)) + (\mu(s) - \lambda) w(x(s)) \right) ds \right| \\ &\quad + \int_0^t \mu(s) \cdot \text{Lip } v \cdot |x(s) - y(s)| ds + \int_0^t (1 - \mu(s)) \cdot \text{Lip } w \cdot |x(s) - y(s)| ds \\ &\leq \left| \int_0^t (\lambda - \mu(s)) \cdot (v(x_0) - w(x_0)) ds \right| \\ &\quad + \int_0^t |\lambda - \mu(s)| (\text{Lip } v + \text{Lip } w) |x(s) - x_0| ds + \int_0^t |x(s) - y(s)| ds \\ &\leq 2 \left| \int_0^t (\lambda - \mu(s)) ds \right| + \int_0^t 1 \cdot 2 \cdot s ds + \int_0^t |x(s) - y(s)| ds \end{aligned}$$

due to  $\|v\|_\infty + \text{Lip } v \leq 1$ ,  $\|w\|_\infty + \text{Lip } w \leq 1$ .

Gronwall's lemma ensures  $|x(t) - y(t)| \leq o(t)$  for  $t \downarrow 0$  uniformly in  $x_0 \in \mathbb{R}^N$ . (In particular, the estimate of Filippov's theorem is difficult to apply here immediately, as the integral mean of  $\mu(\cdot) - \lambda$ , but not of  $|\mu(\cdot) - \lambda|$ , is  $o(t)$  for  $t \downarrow 0$ .)

Thus, for any initial set  $M \in \mathcal{K}(\mathbb{R}^N)$ , the reachable sets satisfy

$$\text{dist}(\vartheta_u(t, M), \vartheta_g(t, M)) \leq o(t) \quad \text{for } t \downarrow 0 \quad \text{uniformly in } M \in \mathcal{K}(\mathbb{R}^N).$$

The same uniform estimates holds for  $\text{dist}(\vartheta_g(t, M), \vartheta_u(t, M))$  since the preceding solutions  $x(\cdot)$  and  $y(\cdot)$  have needed only a joint initial point  $x_0 \in \mathbb{R}^N$ .

According to the proof of Lemma 4.4, we can suppose to have a sequence  $s_n \searrow 0$  in  $]0, 1[$  such that  $\mu(\cdot) \in \{0, 1\}$  is constant in  $[s_{n+1}, s_n[$  for each  $n \in \mathbb{N}$ . So for every set  $M \in \mathcal{K}(\mathbb{R}^N)$  and time  $t \in [s_{n+1}, s_n]$ , the reachable set  $\vartheta_g(t, M)$  is either  $\vartheta_v(t - s_{n+1}, \vartheta_g(s_{n+1}, M))$  or  $\vartheta_w(t - s_{n+1}, \vartheta_g(s_{n+1}, M))$ .



Fix  $\varepsilon > 0$  arbitrarily small. For any  $h \in ]0, s_1]$  sufficiently small, we choose  $m \in \mathbb{N}$  with  $s_m < h \leq s_{m-1}$  and obtain for  $M \in \mathcal{K}(\mathbb{R}^N)$  sufficiently close to  $K$

$$\begin{aligned} & J(\vartheta_u(h, M)) - J(M) \\ & \leq J(\vartheta_g(h, M)) - J(M) + o(h) \\ & = J(\vartheta_g(h, M)) - J(\vartheta_g(s_m, M)) + \sum_{n=m}^{\infty} (J(\vartheta_g(s_n, M)) - J(\vartheta_g(s_{n+1}, M))) + o(h) \\ & \leq (h - s_m) \cdot \left( \frac{\iota J(K)}{2} + \varepsilon \right) + \sum_{n=m}^{\infty} (s_n - s_{n+1}) \cdot \left( \frac{\iota J(K)}{2} + \varepsilon \right) + o(h) \\ & = h \cdot \left( \frac{\iota J(K)}{2} + \varepsilon \right) + o(h). \end{aligned}$$

So,  $\delta^C J(K)(u) \leq \frac{\iota J(K)}{2} + \varepsilon$  with any  $\varepsilon > 0$ , i.e.,  $u \stackrel{\text{Def.}}{=} \lambda v + (1 - \lambda) w \in \mathcal{F}(K)$ .

Each value  $\mathcal{F}(K) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  is closed w.r.t. locally uniform convergence. Let  $(v_n)_{n \in \mathbb{N}}$  be a sequence in  $\mathcal{F}(K) \subset \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  converging to  $v \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  locally uniformly. Obviously, the limit holds  $\|v\|_{\infty} + \text{Lip } v \leq 1$ . Moreover, similarly to Lemma 2.4, Filippov's theorem implies

$$\sup_{\substack{M \in \mathcal{K}(\mathbb{R}^N) \\ d(K, M) \leq 1}} d(\vartheta_{v_n}(h, M), \vartheta_v(h, M)) \leq h e^h \cdot \sup_{\mathbb{B}_2(K)} |v_n(\cdot) - v(\cdot)| \longrightarrow 0$$

for  $n \longrightarrow \infty$  uniformly in  $h \in [0, 1]$  and for every set  $K \in \mathcal{K}(\mathbb{R}^N)$ . So due to the Lipschitz continuity of  $J(\cdot)$ , Clarke's generalized shape derivative satisfies

$$\begin{aligned} \delta^C J(K)(v) & \leq \limsup_{\substack{h \downarrow 0, M \rightarrow K \\ (M \in \mathcal{K}(\mathbb{R}^N))}} \frac{1}{h} \cdot (J(\vartheta_{v_n}(h, M)) - J(M)) + \text{Lip } J \cdot \sup_{\mathbb{B}_2(K)} |v_n(\cdot) - v(\cdot)| \\ & \leq \frac{1}{2} \cdot \iota J(K) + \text{Lip } J \cdot \sup_{\mathbb{B}_2(K)} |v_n(\cdot) - v(\cdot)|. \end{aligned}$$

$n \longrightarrow \infty$  reveals  $v \in \mathcal{F}(K)$ .  $\square$

*Remark.* In regard to the viability theorem, Theorem 3.11, the graph of  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  ought to be closed (still using the topology of locally uniform convergence). This feature is closely related with the lower semicontinuity of  $\delta^C J(\cdot)(v) : \mathcal{K}(\mathbb{R}^N) \longrightarrow \mathbb{R}$  (with  $v \in \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  fixed), and it will be used here as an additional assumption on  $J$ .

So now Lemmas 4.3 and 4.5 have laid the basis for applying the morphological viability theorem, Theorem 3.11. We summarize the main result of this section.

**PROPOSITION 4.6.** *Suppose  $J : \mathcal{K}(\mathbb{R}^N) \longrightarrow \mathbb{R}$  to be Lipschitz continuous w.r.t. the Pompeiu–Hausdorff distance  $d$  and bounded from below.*

*Using Definitions 4.1 and 4.2,  $\mathcal{F} : \mathcal{K}(\mathbb{R}^N) \rightsquigarrow \text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$  is assumed to have a closed graph w.r.t. locally uniform convergence on  $\text{Lip}(\mathbb{R}^N, \mathbb{R}^N)$ .*

*Let  $\mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  be nonempty and closed such that for every  $K \in \mathcal{V}$ , the intersection  $\mathcal{F}(K) \cap T_{\mathcal{V}}(K)$  is nonempty.*

*Then there exists a Lipschitz continuous solution  $K : [0, \infty[ \longrightarrow \mathcal{V} \subset \mathcal{K}(\mathbb{R}^N)$  of the morphological inclusion  $\overset{\circ}{K}(\cdot) \cap \mathcal{F}(K(\cdot)) \neq \emptyset$  such that*

- 1.  $J \circ K : [0, \infty[ \longrightarrow \mathbb{R}$ ,  $t \longmapsto J(K(t))$  is nonincreasing, and*
- 2. every element  $C \in \mathcal{K}(\mathbb{R}^N)$  of its  $\omega$ -limit set in  $\mathcal{K}(\mathbb{R}^N)$  satisfies the following necessary condition on minimizers of  $J(\cdot)$  in  $\mathcal{K}(\mathbb{R}^N)$ :  $\iota J(C) = 0$ .*

**Acknowledgments.** The author would like to thank Irina Surovtsova and Daniel Andrej for fruitful complementary discussions. He is also grateful to the anonymous referees since their very detailed reports contributed a lot to improving this article.

## REFERENCES

- [1] Z. ARTSTEIN, *A calculus for set-valued maps and set-valued evolution equations*, Set-Valued Anal., 3 (1995), pp. 213–261.
- [2] J.-P. AUBIN, *Viability Theory*, Systems Control Found. Appl., Birkhäuser, Boston, 1991.
- [3] J.-P. AUBIN, *A note on differential calculus in metric spaces and its applications to the evolution of tubes*, Bull. Pol. Acad. Sci. Math., 40 (1992), pp. 151–162.
- [4] J.-P. AUBIN, *Mutational equations in metric spaces*, Set-Valued Anal., 1 (1993), pp. 3–46.
- [5] J.-P. AUBIN, *Mutational and Morphological Analysis: Tools for Shape Evolution and Morphogenesis*, Systems Control Found. Appl., Birkhäuser, Boston, 1999.
- [6] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Grundlehren Math. Wiss. 264, Springer-Verlag, Berlin, 1984.
- [7] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Systems Control Found. Appl., Birkhäuser, Boston, 1990.
- [8] D. BÁRCENAS, *Weak compactness criteria for set-valued integrals and Radon–Nikodym theorem for vector-valued multimeasures*, Czech. Math. J., 51 (2001), pp. 493–504.
- [9] J. CÉA, *Une méthode numérique pour la recherche d'un domaine optimal*, in Computing Methods in Applied Sciences and Engineering. Part 1, R. Glowinski and J. L. Lions, eds., Lecture Notes in Econom. and Math. Systems 134, Springer-Verlag, Berlin, 1976, pp. 245–257.
- [10] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Canad. Math. Soc. Ser. Monogr. Adv. Texts, Wiley-Interscience, New York, 1983.
- [11] M. C. DELFOUR AND J.-P. ZOLÉSIO, *Velocity method and Lagrangian formulation for the computation of the shape Hessian*, SIAM J. Control Optim., 29 (1991), pp. 1414–1442.
- [12] M. C. DELFOUR AND J.-P. ZOLÉSIO, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, Adv. Des. Control 4, SIAM, Philadelphia, 2001.
- [13] J. DIESTEL, *Remarks on weak compactness in  $L^1(\mu, X)$* , Glasgow Math. J., 18 (1977), pp. 87–91.
- [14] J. DIESTEL, W. M. RUESS, AND W. SCHACHERMEYER, *Weak compactness in  $L^1(\mu, X)$* , Proc. Amer. Math. Soc., 118 (1993), pp. 443–453.
- [15] J. DIESTEL AND J. UHL, *Vector Measures*, Math. Surveys 15, AMS, Providence, RI, 1977.
- [16] L. DOYEN, *Filippov and invariance theorems for mutational inclusions of tubes*, Set-Valued Anal., 1 (1993), pp. 289–303.
- [17] L. DOYEN, *Shape Lyapunov functions and stabilization of reachable tubes of control problems*, J. Math. Anal. Appl., 184 (1994), pp. 222–228.
- [18] L. DOYEN, *Inverse function theorems and shape optimization*, SIAM J. Control Optim., 32 (1994), pp. 1621–1642.
- [19] L. DOYEN, *Mutational equations for shapes and vision-based control*, J. Math. Imaging Vis., 5 (1995), pp. 99–109.
- [20] H. FRANKOWSKA, S. PLASKACZ, AND T. RZEŻUCHOWSKI, *Measurable viability theorems and the Hamilton–Jacobi–Bellman equation*, J. Differential Equations, 116 (1995), pp. 265–305.
- [21] A. GORRE, *Evolutions of tubes under operability constraints*, J. Math. Anal. Appl., 216 (1997), pp. 1–22.
- [22] J. W. GREEN AND F. A. VALENTINE, *On the Arzelà–Ascoli theorem*, Math. Mag., 34 (1960/61), pp. 199–202.
- [23] M. KISIELEWICZ, *Weak compactness in spaces  $C(S, X)$* , in Transactions of the 11th Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, Prague, Czechoslovakia, Vol. B, 1992, pp. 101–106.
- [24] TH. LORENZ, *Set-valued maps for image segmentation*, Comput. Vis. Sci., 4 (2001), pp. 41–57.
- [25] TH. LORENZ, *Reynold’s transport theorem for differential inclusions*, Set-Valued Anal., 14 (2006), pp. 209–247.
- [26] M. NAGUMO, *Über die Lage der Integralkurven gewöhnlicher Differentialgleichungen*, Proc. Phys.-Math. Soc. Japan (3), 24 (1942), pp. 551–559.
- [27] A. I. PANASYUK, *Quasidifferential equations in metric spaces*, Differential Equations, 21 (1985), pp. 914–921.
- [28] A. I. PANASYUK, *Properties of solutions of a quasidifferential approximation equation and the equation of an integral funnel*, Differential Equations, 28 (1992), pp. 1259–1266.
- [29] A. I. PANASYUK, *Quasidifferential equations in a complete metric space under conditions of*

- the Carathéodory type*. I, Differential Equations, 31 (1995), pp. 901–910.
- [30] M. QUINCAMPOIX AND V. VELIOV, *Open-loop viable control under uncertain initial state information*, Set-Valued Anal., 7 (1999), pp. 55–87.
  - [31] B. RICCERI AND A. VILLANI, *Separability and Scorza-Dragoni's property*, Matematiche (Catania), 37 (1983), pp. 156–161.
  - [32] G. SCORZA-DRAGONI, *Un teorema sulle funzioni continue rispetto ad una e misurabili rispetto ad un'altra variabile*, Rend. Sem. Mat. Univ. Padova, 17 (1948), pp. 102–108.
  - [33] J. SOKOLOWSKI AND J.-P. ZOLÉSIO, *Introduction to Shape Optimization. Shape Sensitivity Analysis*, Springer Ser. Comput. Math. 16, Springer-Verlag, Berlin, 1992.
  - [34] A. ÜLGER, *Weak compactness in  $L^1(\mu, X)$* , Proc. Amer. Math. Soc., 113 (1991), pp. 143–149.
  - [35] R. VINTER, *Optimal Control*, Systems Control Found. Appl., Birkhäuser, Boston, 2000.
  - [36] K. YOSIDA, *Functional Analysis*, 5th ed., Grundlehren Math. Wiss. 123, Springer-Verlag, Berlin, 1978.
  - [37] W. ZIEMER, *Weakly Differentiable Functions*, Grad. Texts in Math. 120, Springer-Verlag, New York, 1989.
  - [38] J.-P. ZOLÉSIO, *Identification de domaine par déformations*, Thèse de doctorat d'état, Université de Nice, Nice, France, 1979.

# THE BREZIS–EKELAND PRINCIPLE FOR DOUBLY NONLINEAR EQUATIONS\*

ULISSE STEFANELLI†

**Abstract.** The celebrated Brezis–Ekeland principle [*C. R. Acad. Sci. Paris Ser. A-B*, 282 (1976), pp. Ai, A1197–A1198, Aii, and A971–A974] characterizes trajectories of nonautonomous gradient flows of convex functionals as solutions to suitable minimization problems. This note extends this characterization to doubly nonlinear evolution equations driven by convex potentials. The characterization is exploited in order to establish approximation results for gradient flows, doubly nonlinear equations, and rate-independent evolutions.

**Key words.** Brezis–Ekeland principle, gradient flow, doubly nonlinear equation, rate-independent evolution, Mosco convergence, Young measures, approximation

**AMS subject classification.** 35K55

**DOI.** 10.1137/070684574

**1. Introduction.** Let  $H$  denote a Hilbert space and  $T > 0$  be some final reference time. Moreover, let  $\phi : H \rightarrow (-\infty, \infty]$  be convex, proper, and lower semicontinuous,  $f \in L^2(0, T; H)$ , and  $u^0 \in D(\phi) := \{v \in H : \phi(v) \neq \infty\}$ . The variational principle formulated by Brezis and Ekeland [17, 18] and Nayroles [77, 78] (see also [9, sect. 3.4]) characterizes solutions  $u \in H^1(0, T; H)$  of the gradient flow

$$(1.1) \quad u' + \partial\phi(u) \ni f \quad \text{a.e. in } (0, T), \quad u(0) = u^0$$

(where the prime stands for time differentiation and  $\partial\phi$  is the subdifferential of  $\phi$  in the sense of convex analysis; see below) as a global minimizer of the functional  $J : H^1(0, T; H) \rightarrow [0, \infty]$  defined as

$$(1.2) \quad J(u) := \int_0^T (\phi(u) + \phi^*(f - u') - (f, u)) + \frac{1}{2}|u(T)|^2 - \frac{1}{2}|u(0)|^2 + |u(0) - u^0|^2.$$

Here  $(\cdot, \cdot)$  is the scalar product in  $H$ ,  $|\cdot|$  is the corresponding norm, and we have denoted by  $\phi^*$  the conjugate of  $\phi$ , i.e.,  $\phi^*(w) := \sup\{(w, u) - \phi(u), u \in H\}$  for all  $w \in H$ . Let us stress that  $\phi(u) + \phi^*(w) \geq (w, u)$  for all  $u, w \in H$  and that the equality holds iff  $w \in \partial\phi(u)$ . In particular, one readily checks that  $J(v) \geq 0$  for all  $v \in H^1(0, T; H)$  and that  $J(u) = \min J = 0$  iff  $u$  solves the gradient flow (1.1). Namely, the unique solution to (1.1) and the unique minimizer of  $J$  coincide, and we have the following.

**THEOREM 1.1** (Brezis and Ekeland [17, 18]).  *$u$  solves (1.1) iff  $J(u) = 0$ .*

The aim of this note is to extend the latter characterization result to the more general situation of doubly nonlinear equations. In particular, let a second convex, proper, and lower semicontinuous functional  $\psi : H \rightarrow (-\infty, \infty]$  be given. We are interested in solving for  $u \in W^{1,p}(0, T; H)$ ,  $p \in [1, \infty]$ , the equation

$$(1.3) \quad \partial\psi(u') + \partial\phi(u) \ni f \quad \text{a.e. in } (0, T), \quad u(0) = u^0,$$

where now  $f \in L^q(0, T; H)$ , with  $1/p + 1/q = 1$  (usual convention:  $1/\infty = 0$ ).

\*Received by the editors March 7, 2007; accepted for publication (in revised form) January 10, 2008; published electronically May 30, 2008. Partially supported by the CNR-STM 2006 program and by the Swiss National Science Foundation.

<http://www.siam.org/journals/sicon/47-3/68457.html>

†IMATI - CNR, v. Ferrata 1, I-27100 Pavia, Italy (ulisse.stefanelli@imati.cnr.it).

The latter equation arises in a variety of different applicative contexts. In particular, inclusion (1.3) may represent a generalized balance relation in thermomechanics. The reader is referred to Moreau [74, 75] and Germain [30] for some justification and to Colli and Visintin [23] and Colli [22] for existence results for functionals  $\psi$  of  $p$ -growth for  $1 < p < \infty$  (see also Barbu [12], Arai [3], and Senba [95] among others). The linear-growth case  $p = 1$  is strictly related with the modeling of rate-independent evolution and has been considered in connection with elasto-plasticity [25, 26, 58, 59, 60, 61], damage [68], brittle fractures [27], delamination [56], ferroelectricity [73], shape-memory alloys [67, 71, 72], and vortex pinning in superconductors [94]; see Mielke [62] for a comprehensive survey of mathematical results.

We shall make precise problem (1.3) by introducing an auxiliary function. Namely, we consider  $(u, v) \in W^{1,p}(0, T; H) \times L^q(0, T; H)$  such that

$$(1.4) \quad v \in \partial\psi(u') \quad \text{a.e. in } (0, T),$$

$$(1.5) \quad v + \partial\phi(u) \ni f \quad \text{a.e. in } (0, T),$$

$$(1.6) \quad u(0) = u^0.$$

The above relations have a clear mechanical interpretation. By letting  $u$  represent the displacement of a body from its reference configuration,  $\phi$  can be interpreted as the corresponding energy, and  $\psi$  stands for the related dissipation potential. In particular, relation (1.5) expresses the balance among the system of conservative forces  $\partial\phi(u)$ , the dissipative (viscous) force  $v$ , and the external load  $f$ . On the other hand, relation (1.4) consists of a multivalued constitutive relation for the dissipative forces.

Let now the functional  $I : W^{1,p}(0, T; H) \times L^q(0, T; H) \rightarrow [0, \infty]$  be defined as

$$(1.7) \quad \begin{aligned} I(u, v) := & \left( \int_0^T (\psi(u') + \psi^*(v) - (f, u')) + \phi(u(T)) - \phi(u^0) \right)^+ \\ & + \int_0^T (\phi(u) + \phi^*(f - v) - (f - v, u)) + |u(0) - u^0|^2. \end{aligned}$$

Note that, exactly as for  $J$ , no derivatives of the potentials appear in the definition of  $I$ , making its formulation suited for nonsmooth situations.

The key point of this note is to check that solutions of (1.4)–(1.6) are precisely the (possibly nonunique) minimizers of the nonnegative functional  $I$ . In particular, we prove the following.

**THEOREM 1.2.**  $(u, v)$  solves (1.4)–(1.6) iff  $I(u, v) = 0$ .

Let us explicitly remark that Theorem 1.2 implies the Brezis–Ekeland characterization of Theorem 1.1; namely, the present analysis extends the former. In order to prove this fact, some care has to be used since in the quadratic case  $\psi(\cdot) = |\cdot|^2/2$  the functionals  $I(u, u')$  and  $J(u)$  do not coincide. Precisely, we shall define  $K : H^1(0, T; H) \rightarrow [0, \infty]$  by

$$(1.8) \quad K(u) := I(u, u') = \left( \int_0^T (|u'|^2 - (f, u')) + \phi(u(T)) - \phi(u^0) \right)^+ + J(u).$$

Theorem 1.1 will follow from Theorem 1.2 once we check that  $K$  and  $J$  have the same minimizer. This is exactly the point of the following.

**LEMMA 1.3.**  $K(u) = 0$  iff  $J(u) = 0$ .

Indeed, by letting  $J(u) = 0$  we shall prove that  $K(u) = 0$  (the converse implication being obvious). Since  $u$  is a solution to (1.1) by Theorem 1.1, the chain rule [16, Lem. 3.3, p. 73] yields

$$\phi(u(T)) - \phi(u^0) = \int_0^T \frac{d}{dt} \phi(u) = \int_0^T (f - u', u').$$

Hence, the positive part in (1.8) is 0, and  $K(u) = J(u) = 0$ .

The functional  $J$  is convex and lower semicontinuous with respect to the weak topology of  $H^1(0, T; H)$ . Hence, one is tempted to exploit the Brezis–Ekeland characterization of Theorem 1.1 in order to obtain solutions to the gradient flow (1.1) by applying the direct method to  $J$ . This strategy is, however, much more involved than the classical maximal-monotone operator techniques (see Brezis [16]). The difficulty arises from the fact that one is not just asked to minimize  $J$  but also to prove that the minimum is 0 (the difficulty of proving the existence of a minimizer for  $J$  without using the equation was already pointed out in [18, Rem. 1]). Incidentally, note that  $J$  may fail to be coercive with respect to the weak topology of  $H^1(0, T; H)$  unless  $\phi^*$  has at least a quadratic growth and hence  $\phi$  is quadratically bounded (take  $\phi^*(\cdot) = |\cdot|$ ,  $f = 0$ ,  $u^0 = 0$ , and  $u_n$  to be any  $W^{1,1}(0, T; H)$ -bounded sequence with  $|u_n| \leq 1$  and  $u'_n$  unbounded in  $L^2(0, T; H)$ ). On the other hand,  $K$  is clearly convex, lower semicontinuous, and coercive with respect to the weak topology of  $H^1(0, T; H)$ .

Conditional existence results for the gradient flow (1.1) by means of the direct method were first obtained by Rios [85, 88] (see also [86, 87]). Later on, Auchmuty [10] proved that in the controlled-growth case the minimum problem can be reformulated as a saddle point problem for which the minimax value 0 is achieved (see also [8]). Again in the controlled-growth case and by assuming  $\phi$  to be continuously differentiable, Roubíček [90] directly checked that the optimality conditions imply (1.1) (see also the recent monograph [91, sect. 8.10]). Finally, the full extent of maximal-monotone methods has been recovered via the Brezis–Ekeland approach by Ghoussoub and Tzou [39]. In the latter paper, the authors eventually overcome the controlled-growth assumption and recast the problem within the far-reaching theory of (anti)self-dual Lagrangians by Ghoussoub [35, 34, 31, 32, 33, 37, 41, 40] (see also the monograph [36]). We mention some further results by Ghoussoub and McCann [38] for quadratic perturbations of convex functionals and the analysis of the long-time dynamics of autonomous gradient flows by Lemaire [51].

Our main focus here is, however, not on existence but rather on the application of the characterization result of Theorem 1.2 to the analysis of general approximation issues. Since solutions and minimizers of the respective functionals coincide, a quite natural idea in order to frame an abstract approach to limiting procedures is that of considering the corresponding approximating minimum problems via  $\Gamma$ -convergence [43, 24]. We shall apply this perspective and generalize some known approximation results for both gradient flows (section 6) and doubly nonlinear equations (section 7). Moreover, we obtain a new proof of a convergence result by Mielke, Roubíček, and Stefanelli [69] for the case of rate-independent problems in section 8.

The key step in the direction of approximations is a suitable  $\Gamma$ -lim inf tool settled in the frame of Young measures with values in separable and reflexive Banach spaces (see subsection 4.2). Let us mention that this perspective has already been considered by Pedregal [81, 82] and Michaille and Valadier [57] in the frame of Sobolev spaces. Here, moving from some recent version in weak topologies of the fundamental result by Balder [11, Thm. 1], we deduce a useful tool in order to pass to lower limits in

sequences of integral functionals whose integrands fulfill a suitable  $\Gamma$ -lim inf inequality. As a by-product, we obtain some generalization of former results by Salvadori [92, Thm. 3.1].

By applying the above-mentioned approximation results, we show in subsection 7.1 the existence of solutions for a class of doubly nonlinear equations by passing to the limit via Theorem 1.2 within a suitable class of regularized problems. In particular, we recover by means of a variational technique a former existence result by Colli and Visintin [23, Thm. 2.1].

Let us mention that some refined version of the functional  $I$  and Theorem 1.2 is discussed in [103] for the specific situation of linearized elastoplasticity with hardening. In particular, the (classical) well-posedness theory and the (more recent) convergence for time and space discretizations [47] is there recovered by means of a variational technique.

Let us close this introduction by observing that, besides the Brezis–Ekeland principle, a variety of global variational principles for dissipative evolutions have already been proposed. We mention Biot’s work on irreversible thermodynamics [15] and Gurtin’s principle for viscoelasticity and elastodynamics [44, 45, 46] among many others (see also the survey in Hlaváček [48]). We shall not attempt to give here a comprehensive report on the literature but rather concentrate on the specific case of doubly nonlinear evolutions. In this concern, the reader is referred to Visintin [105], where generalized solutions are obtained as minimal elements of a certain partial-order relation on the trajectories, and Mielke and Ortiz [63] (see also [70]), where solutions in the rate-independent case are recovered as suitable limits of relaxed global minimization problems.

*Remark 1.4.* The formulation in (1.2) is not the original one but is rather some modification due to Rockafellar (see again [18]) also considered in Ghoussoub and Tzou [39].

**2. Characterization.** The Brezis–Ekeland characterization of Theorem 1.1 makes no essential use of the Hilbert-space structure. Hence, let us move from the very beginning to the reflexive Banach-space framework and start by enlisting our assumptions:

- (A1)  $p \in [1, \infty]$ ,  $1/p + 1/q = 1$ , and  $H$  is a real reflexive Banach space with norm  $|\cdot|$ . We shall use the symbol  $(\cdot, \cdot)$  for the duality pairing between  $H^*$  (dual) and  $H$ .
- (A2)  $\phi, \psi : H \rightarrow (-\infty, \infty]$  are proper, convex, and lower semicontinuous.
- (A3)  $f \in L^q(0, T; H^*)$  and  $u^0 \in D(\phi) := \{v \in H : \phi(v) \neq \infty\}$ .

The subdifferentials occurring in the formulation of the Cauchy problem (1.4)–(1.6) are now acting from  $H$  to  $H^*$  being defined as

$$w \in \partial\phi(z) \quad \text{iff} \quad z \in D(\phi) \quad \text{and} \quad (w, x - z) \leq \phi(x) - \phi(z) \quad \forall x \in H$$

and analogously for  $\partial\psi$ . Notations have been chosen in such a way that the definition of the functional  $I$  in (1.7) still makes sense in the above Banach-space setting. For the sake of clarity, we shall restate here our main result.

**THEOREM 2.1.** *Under assumptions (A1)–(A3), the pair  $(u, v)$  solves (1.4)–(1.6) iff  $I(u, v) = 0$ .*

The proof of Theorem 2.1 relies on a suitable Banach version of the chain rule [16, Lem. 3.3, p. 73]. We state it here for the sake of completeness and provide a direct proof.

PROPOSITION 2.2 (chain rule). *Under assumption (A1), let  $\phi : H \rightarrow (-\infty, \infty]$  be proper, convex, and lower semicontinuous,  $u \in W^{1,p}(0, T; H)$ , and  $w \in L^q(0, T; H^*)$  be such that  $w \in \partial\phi(u)$  almost everywhere in  $(0, T)$ . Then  $t \mapsto \phi(u(t))$  is absolutely continuous and*

$$\phi(u(t)) - \phi(u(s)) = \int_s^t (z, u') \quad \forall 0 \leq s \leq t \leq T,$$

for all  $z \in L^q(0, T; H^*)$ , with  $z \in \partial\phi(u)$  almost everywhere in  $(0, T)$ .

*Proof.* Let us assume from the very beginning that both  $H$  and  $H^*$  are strictly convex ( $H$  can be equivalently renormed in such a way that this holds [5, 6]). We let  $w_\varepsilon := \partial\phi_\varepsilon(u)$  for  $\varepsilon > 0$ , where  $\phi_\varepsilon$  is the Yosida approximation of  $\phi$  (see [13, Prop. 1.1, p. 42] for definition and properties). Since  $\phi_\varepsilon$  is Gâteaux differentiable [13, Thm. 2.2, p. 57] we have

$$(2.1) \quad \phi_\varepsilon(u(t)) - \phi_\varepsilon(u(s)) = \int_s^t (w_\varepsilon, u') \quad \forall 0 \leq s \leq t \leq T.$$

Moreover, one has  $|w_\varepsilon|_* \leq |w^\circ|_* \leq |w|_*$  almost everywhere in  $(0, T)$ , where  $|\cdot|_*$  is the norm of  $H^*$  and  $w^\circ := (\partial\phi(u))^\circ$  is the element of minimal norm in  $\partial\phi(u)$ . Hence, for all  $t \in (0, T)$  such that  $|w_\varepsilon(t)|_* \leq |w^\circ(t)|_* \leq |w(t)|_*$ , one can extract a not relabeled (and a priori depending on  $t$ ) weakly convergent subsequence  $w_\varepsilon(t)$ . On the other hand, by using [13, Prop. 1.1, p. 42] we have  $w_\varepsilon(t) \rightarrow w^\circ(t)$  weakly in  $H^*$ . In particular, the whole sequence  $w_\varepsilon(t)$  converges, and we have  $w_\varepsilon \rightarrow w^\circ$  weakly in  $H^*$  pointwise almost everywhere in  $(0, T)$ . By exploiting the convergence of Yosida approximations and the dominated convergence theorem and by passing to the limit as  $\varepsilon \rightarrow 0$  in (2.1), we prove that  $t \mapsto \phi(u(t))$  is absolutely continuous on  $[0, T]$ .

Let now  $t \in (0, T)$  be a point, where  $t \mapsto \phi(u(t))$  is differentiable and  $u(t) \in D(\partial\phi) := \{v \in H : \partial\phi(v) \neq \emptyset\}$ , and let  $z \in \partial\phi(u(t))$ . Then

$$(z, x - u(t)) \leq \phi(x) - \phi(u(t)) \quad \forall x \in H.$$

By choosing  $x = u(t \pm h)$  for  $h > 0$  and passing to the limit as  $h \rightarrow 0$  we readily check that

$$\frac{d}{dt} \phi(u(t)) = (z, u'(t)),$$

and the assertion follows.  $\square$

*Proof of Theorem 2.1.* Let  $(u, v) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$  solve (1.4)–(1.6). In particular, we have

$$\phi(u) + \phi^*(f - v) = (f - v, u), \quad \psi(u') + \psi^*(v) = (v, u') \quad \text{a.e. in } (0, T)$$

so that  $I(u, v) < \infty$ . Moreover, owing to Proposition 2.2, the map  $t \mapsto \phi(u(t))$  is absolutely continuous and

$$(2.2) \quad \phi(u(T)) - \phi(u^0) = \int_0^T \frac{d}{dt} \phi(u) = \int_0^T (f - v, u').$$

Hence, since  $u(0) = u^0$ ,

$$I(u, v) = \left( \int_0^T (\psi(u') + \psi^*(v) - (v, u')) \right)^+ = 0.$$



On the other hand, let  $(u, v) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$  be such that  $I(u, v) = 0$ . The functional  $I$  results from the sum of three nonnegative contributions, namely, the positive-part term, the integral term, and the term taking into account the initial datum. As  $I(u, v) = 0$ , one has that all three terms must be 0. In particular,  $u(0) = u^0$ , and relation (1.5) holds. By using Proposition 2.2 and the already established (1.5) we have

$$\begin{aligned} & \int_0^T (\psi(u') + \psi^*(v) - (v, u')) \\ &= \int_0^T (\psi(u') + \psi^*(v) - (f, u')) + \phi(u(T)) - \phi(u^0) \leq 0, \end{aligned}$$

where the last inequality follows from the fact that the positive-part term in  $I(u, v)$  is 0. Finally, as  $\psi(u') + \psi^*(v) \geq (v, u')$  almost everywhere in  $(0, T)$ , relation (1.4) holds as well.  $\square$

Let us remark that the reflexive Banach frame of assumption (A1) includes the original Hilbert-space setting of Theorem 1.1. In particular, the Brezis–Ekeland characterization follows from Theorem 2.1 via Lemma 1.3.

Before going on, we shall comment that the alternative (and somehow more natural) choice  $\tilde{I} : W^{1,p}(0, T; H) \times L^q(0, T; H) \rightarrow [0, \infty]$  given by

$$\begin{aligned} \tilde{I}(u, v) &:= \int_0^T (\psi(u') + \psi^*(v) - (v, u')) \\ &+ \int_0^T (\phi(u) + \phi^*(f - v) - (f - v, u)) + |u(0) - u^0|^2 \end{aligned}$$

would lead to the same characterization of Theorem 2.1. On the other hand, the presence of the term  $(v, u')$  prevents the functional  $\tilde{I}$  from being lower semicontinuous with respect to the natural topologies related to the doubly nonlinear problem (1.3) (see below), making the functional  $\tilde{I}$  not interesting.

In the specific case of a Hilbert space  $H$  and a quadratic potential  $\phi(\cdot) = |\cdot|^2/2$ , relation (1.3) takes the form

$$(2.3) \quad \partial\psi(u') + u \ni f \quad \text{a.e. in } (0, T), \quad u(0) = u^0.$$

By letting the functional  $Q : W^{1,p}(0, T; H) \rightarrow [0, \infty]$  be defined as

$$Q(u) := \int_0^T (\psi(u') + \psi^*(f - u) - (f, u')) + \frac{1}{2}|u(T)|^2 - \frac{1}{2}|u(0)|^2 + |u(0) - u^0|^2,$$

one easily checks via Fenchel's duality that  $u \in W^{1,p}(0, T; H)$  solves (2.3) iff  $Q(u) = \min Q = 0$ . On the other hand, in the same spirit of (1.8), one could consider  $R : W^{1,p}(0, T; H) \rightarrow [0, \infty]$  as  $R(u) := I(u, f - u)$ , and the analogue of Lemma 1.3 holds.

The case of a quadratic potential  $\phi$  bears some relevance with respect to applications (see section 8 and [103]). Hence, we shall explicitly consider the functional  $Q$  in the following.

**3. Some extension.** As already mentioned in the introduction, Theorem 2.1 is valid in some more general frames. In particular, we shall consider the doubly nonlinear relation

$$(3.1) \quad \partial\psi(u'(t)) + \partial\varphi(t, u(t)) \ni 0 \quad \text{for a.e. } t \in (0, T), \quad u(0) = u^0,$$

where the time dependence is now included in  $\varphi$  and the second subdifferential is referred to the variable  $u$  only.

Letting  $X$  be a separable metric space, we denote by  $\mathcal{B}(X)$  its Borel  $\sigma$ -algebra, by  $\mathcal{L}$  the  $\sigma$ -algebra of the Lebesgue measurable subsets of  $(0, T)$ , and by  $\mathcal{L} \otimes \mathcal{B}(X)$  the respective product  $\sigma$ -algebra. A  $\mathcal{L} \otimes \mathcal{B}(X)$ -measurable function  $g : (0, T) \times X \rightarrow (-\infty, \infty]$  is said to be a *normal integrand* if

$$u \mapsto g(t, u) \quad \text{is lower semicontinuous for a.e. } t \in (0, T).$$

The assumptions read as follows:

- (A4)  $\psi : H \rightarrow (-\infty, \infty]$  is convex, proper, and lower semicontinuous,  
 $\varphi : [0, T] \times H \rightarrow (-\infty, \infty]$  is such that:  
 $u \mapsto \varphi(t, u)$  is proper and convex for a.e.  $t \in (0, T)$ ,  
 for all separable subspaces  $X \subset H$ , the restriction of  $\varphi$  to  $[0, T] \times X$  is a normal integrand,  
 there exists  $\pi : W^{1,p}(0, T; H) \rightarrow L^1(0, T)$  such that, for  $u \in W^{1,p}(0, T; H)$  and  $w \in L^q(0, T; H^*)$ , with  $w(t) \in \partial\varphi(t, u(t))$  for a.e.  $t \in (0, T)$ , the mapping  $t \mapsto \varphi(t, u(t))$  is absolutely continuous and satisfies

$$(3.2) \quad \varphi(t, u(t)) - \varphi(s, u(s)) = \int_s^t (w, u') + \int_s^t \pi(u) \quad \forall 0 \leq s \leq t \leq T.$$

- (A5)  $u^0 \in D(\varphi(0, \cdot)) := \{v \in H : \varphi(0, v) \neq \infty\}$ .

Assumption (A4) implies via Pettis' theorem that  $t \mapsto \varphi(t, u(t))$  is measurable for all measurable  $t \mapsto u(t)$ .

As for the generalized chain rule stated in (A4), let us mention that  $\pi$  represents a power of external actions since, at least formally,  $\pi = \partial_t \varphi$  (see also section 8 below). The chain rule (3.2) frequently holds in practice. In particular, it holds in the smooth case and if  $\varphi$  is a smooth perturbation of a convex function. The reader is referred to [64, Prop. 2.6] for a result in the nonperturbative case.

We will consider solutions  $(u, v) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$  of the Cauchy problem (see (1.4)–(1.6))

$$(3.3) \quad v \in \partial\psi(u') \quad \text{a.e. in } (0, T),$$

$$(3.4) \quad -v(t) \in \partial\varphi(t, u(t)) \quad \text{for a.e. } t \in (0, T),$$

$$(3.5) \quad u(0) = u^0.$$

Hence, let us define the functional  $\mathcal{I}$  acting on  $W^{1,p}(0, T; H) \times L^q(0, T; H^*)$  as

$$\begin{aligned} \mathcal{I}(u, v) := & \left( \int_0^T (\psi(u') + \psi^*(v) - \pi(u)) + \varphi(T, u(T)) - \varphi(0, u^0) \right)^+ \\ & + \int_0^T (\varphi(\cdot, u) + \varphi^*(\cdot, -v) + (v, u)) + |u(0) - u^0|^2, \end{aligned}$$

where the duality  $\varphi^*$  is taken with respect to the variable  $u$  only. The formulation of Theorem 1.2 in this setting reads as follows.

**THEOREM 3.1.** *Under assumptions (A1) and (A4)–(A5), the pair  $(u, v)$  solves (3.3)–(3.5) iff  $\mathcal{I}(u, v) = 0$ .*

*Sketch of the proof.* This argument follows along the same lines as that of Theorem 1.2. All solutions to (3.3)–(3.5) are easily proved to be minimizers of  $\mathcal{I}$  by means of

the chain rule (3.2). On the other hand, let  $(u, v) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$  be such that  $\mathcal{I}(u, v) = 0$ . Then one has  $u(0) = u^0$  and  $-v(t) \in \partial\varphi(t, u(t))$  for almost every  $t \in (0, T)$ . Again by (3.2) one gets

$$\begin{aligned} & \int_0^T (\psi(u') + \psi^*(v) - \langle v, u' \rangle) \\ &= \int_0^T (\psi(u') + \psi^*(v) - \pi(u)) + \varphi(T, u(T)) - \varphi(0, u^0) \leq 0, \end{aligned}$$

and the assertion follows.  $\square$

*Remark 3.2.* Let us stress that the present situation of (3.1) is not directly extending (1.3) as some extra time regularity is here exploited. As a matter of fact, the case  $\varphi(t, u) = \phi(u) - \langle f(t), u \rangle$  is included in the frame of (A4) for the smooth choice  $f \in W^{1,q}(0, T; H^*)$  only.

*Remark 3.3.* We shall mention that the generalized situation of the equation

$$(3.6) \quad \partial\psi(u(t), u'(t)) + \partial\phi(t, u(t)) \ni 0 \quad \text{for a.e. } t \in (0, T), \quad u(0) = u^0,$$

where the functional  $\psi : H \times H \rightarrow (-\infty, \infty]$  is convex in its second occurrence and subdifferentials are taken with respect to second variables, has recently attracted a good deal of attention. In particular (3.6) arises in connection with quasi-variational problems and has been considered in Aso, Frémond, and Kenmochi [4], Mielke [62], and Mielke, Rossi, and Savaré [66, 65]. The formulation of the above characterization result in Theorem 3.1 could be easily tailored to the case of (3.6) as well as to even more general situations (the nonlocal situations of [97, 98, 99, 100, 49], for instance) with no particular intricacy.

Before closing this section, let us explicitly mention that the choice  $H$  Hilbert space,  $p = 2$ , and  $\psi(\cdot) = |\cdot|^2/2$  in relation (3.1) give rise to the generalized gradient flow

$$(3.7) \quad u'(t) + \partial\varphi(t, u(t)) \ni 0 \quad \text{a.e. in } (0, T), \quad u(0) = u^0,$$

whose consideration has to be traced back to Peralba [83, 84]. We shall consider (3.7) from the point of view of approximation in section 6 below. Let us explicitly observe there that the extension of the functional  $K$  to the latter time-dependent situation is  $\mathcal{K} : H^1(0, T; H) \rightarrow [0, \infty]$  defined by

$$\begin{aligned} \mathcal{K}(u) := \mathcal{I}(u, u') &= \left( \int_0^T (|u'|^2 - \pi(u)) + \varphi(T, u(T)) - \varphi(0, u^0) \right)^+ \\ &+ \int_0^T (\varphi(\cdot, u) + \phi^*(\cdot, -u')) + \frac{1}{2}|u(T)|^2 - \frac{1}{2}|u(0)|^2 + |u(0) - u^0|^2. \end{aligned}$$

**4. Applications to approximation.** We now apply the characterization results of Theorems 2.1 and 3.1 to the approximation of solutions of the gradient flows (1.1) and (3.7) and of the doubly nonlinear equations (1.3) and (3.1). As already mentioned in the introduction, we shall make here use of the notion of  $\Gamma$ -convergence. The reader is referred to the monographs by Attouch [7] and Dal Maso [24] for some comprehensive discussion of this topic as well as to subsection 4.1 for a (minimal) selection of results to be used later.

Our aim will be to present suitable assumptions for the corresponding approximating functionals to be  $\Gamma$ -converging (or rather Mosco-converging; see below). More

precisely, since Theorems 2.1 and 3.1 directly quantify the value of the minima to be 0, what is actually needed for passing to limits are  $\Gamma$ -lim inf inequalities only. In subsection 4.2 we shall prepare a tool in order to deal with these kinds of problems in a quite general fashion. In particular, we will provide a  $\Gamma$ -lim inf result by exploiting the theory of Young measures for weak topologies in separable (and reflexive) Banach spaces. The application of the latter to the current convex situation is discussed in subsection 4.3 along with some relation with the former analysis by Salvadori [92]. Then in section 5 we systematically apply the  $\Gamma$ -lim inf tool to the functionals introduced in sections 1 and 3 and obtain the corresponding  $\Gamma$ -lim inf inequalities. Convergence results will then follow by simply checking the uniform coercivity of the approximating functionals with respect to suitable topologies. In the case of the gradient flows (1.1) and (3.7), we will show in section 6 that a natural choice is that of the weak topology in  $H^1(0, T; H)$  (note that  $K$  is lower semicontinuous and coercive with respect to the latter). This leads us to generalize some known convergence results (see [7, sect. 3.9.2, p. 386]).

In the case of the doubly nonlinear equations (1.3) and (3.1) for  $p \in (1, \infty)$ , we will provide in section 7 some conditions implying uniform coercivity with respect to the topology

$$(4.1) \quad \mathcal{T}(p) = (\text{weak-}W^{1,p}(0, T; H) + \text{strong-}L^p(0, T; H)) \times \text{weak-}L^q(0, T; H^*)$$

(recall that  $I$  is lower semicontinuous with respect to  $\mathcal{T}(p)$ ). Hence, under suitable nondegeneracy and growth-type assumptions (see (7.1)–(7.2)), section 7 leads to some generalization of former convergence results by Aizicovici and Yan [1].

The situation  $p = 1$  is much more delicate, and we shall deal with it in the specific yet relevant case of rate-independent problems in section 8. By focusing on the functional frame of the recent well-posedness theory by Mielke and Rossi [64], we shall provide a new proof of a convergence result by Mielke, Roubíček, and Stefanelli [69], although in a simplified setting. Finally, the special case of (2.3) is discussed.

**4.1. Preliminaries on Mosco convergence.** Recall that  $H$  is a real reflexive Banach space. By letting  $\phi_n, \phi : H \rightarrow (-\infty, \infty]$  be convex, proper, and lower semicontinuous, we say that the sequence  $\phi_n$  is *uniformly proper* iff there exists a bounded sequence  $u_n$  such that  $u_n \in D(\phi_n)$ . Moreover, we say that  $\phi_n \rightarrow \phi$  in the *Mosco sense* [7, 76] iff, for all  $u \in H$ ,

$$\begin{aligned} \phi(u) &\leq \liminf_{n \rightarrow \infty} \phi_n(u_n) \quad \forall u_n \rightarrow u \text{ weakly in } H, \\ \exists u_n &\rightarrow u \text{ strongly in } H \text{ such that } \phi(u) = \limsup_{n \rightarrow \infty} \phi_n(u_n). \end{aligned}$$

In particular,  $\phi_n \rightarrow \phi$  in the Mosco sense iff  $\phi_n \rightarrow \phi$  in the sense of  $\Gamma$ -convergence with respect to both the weak and the strong topology in  $H$ . Let us stress that if  $\phi_n \rightarrow \phi$  in the Mosco sense, then the sequence  $\phi_n$  is uniformly proper. One has the following characterization.

LEMMA 4.1. *The following are equivalent:*

- (i)  $\phi_n \rightarrow \phi$  in the Mosco sense in  $H$ ,
- (ii)  $\left\{ \begin{array}{l} \phi(u) \leq \inf \left\{ \liminf_{n \rightarrow \infty} \phi_n(u_n) : u_n \rightarrow u \text{ weakly in } H \right\}, \\ \phi^*(u) \leq \inf \left\{ \liminf_{n \rightarrow \infty} \phi_n^*(u_n) : u_n \rightarrow u \text{ weakly in } H^* \right\}, \\ \text{and the sequence } \phi_n^* \text{ is uniformly proper.} \end{array} \right.$

*Proof.* Assumption (i) is equivalent to  $\phi_n^* \rightarrow \phi^*$  in the Mosco sense [7, Thm. 3.18, p. 295]. Hence (ii) follows by the definition of Mosco convergence.

The converse implication (ii)  $\Rightarrow$  (i) is more involved. Let us set for convenience, for all  $u \in H$  and  $v \in H^*$ ,

$$\begin{aligned}\phi_{\sharp}(u) &:= \min \left\{ \limsup_{n \rightarrow \infty} \phi_n(u_n) : u_n \rightarrow u \text{ strongly in } H \right\}, \\ \phi_{\flat}(v) &:= \inf \left\{ \liminf_{n \rightarrow \infty} \phi_n^*(v_n) : v_n \rightarrow v \text{ weakly in } H^* \right\},\end{aligned}$$

Owing to [7, Thm. 1.13, p. 29], the minimum in the definition of  $\phi_{\sharp}$  is always attained in  $(-\infty, \infty]$ . In particular, for all  $u \in H$ ,

$$(4.2) \quad \exists u_n \rightarrow u \text{ strongly in } H \text{ such that } \phi_{\sharp}(u) = \limsup_{n \rightarrow \infty} \phi_n(u_n).$$

The sequence  $\phi_n^*$  is uniformly proper, and we can apply [7, Thm. 3.7, p. 271] in order to deduce that  $\phi_{\sharp} = (\phi_{\flat})^*$ . Hence, since by the second part of (ii) we have  $\phi^* \leq \phi_{\flat}$ , we obtain by duality that  $\phi_{\sharp} = (\phi_{\flat})^* \leq \phi$ . Finally, the first part of (ii) and (4.2) yield (i).  $\square$

Let us comment that, for the sake of establishing the following approximation results, only the implication (i)  $\Rightarrow$  (ii) is exploited. We stated the above characterization in order to underline the fact that the Mosco convergence is the natural requirement for passing to the  $\liminf$  in the sum of functionals and their respective duals (which is precisely our situation).

**4.2. A general  $\Gamma$ -lim inf result.** We shall now discuss a technical tool which will be at the basis of the forthcoming analysis. In particular, we present here a  $\Gamma$ -lim inf result in the frame of Young measures for weak topologies. The reader is referred to Castaing, Raynaud de Fitte, and Valadier [20] for a comprehensive discussion of Young measures on separable Banach spaces.

Recall that, by letting  $X$  be a separable metric space, a  $\mathcal{L} \otimes \mathcal{B}(X)$ -measurable function  $g : (0, T) \times X \rightarrow (-\infty, \infty]$  is said to be a *normal integrand* if

$$u \mapsto g(t, u) \text{ is lower semicontinuous for a.e. } t \in (0, T).$$

We denote by  $\mathcal{M}(0, T; X)$  the set of all  $\mathcal{L}$ -measurable functions  $w : (0, T) \rightarrow X$ . Following [11], a sequence  $w_n \in \mathcal{M}(0, T; X)$  is said to be *tight* if there exists a nonnegative normal integrand  $g$  such that

$$(4.3) \quad \{w \in X : g(t, w) \leq c\} \text{ is compact for a.e. } t \in (0, T) \text{ and all } c \geq 0,$$

$$(4.4) \quad \text{and } \sup_n \int_0^T g(t, w_n(t)) dt < \infty.$$

In case  $X$  is a Banach space, a  $\mathcal{L} \otimes \mathcal{B}(X)$ -measurable function  $h : (0, T) \times X \rightarrow (-\infty, \infty]$  is called a *weakly normal integrand* if

$$(4.5) \quad u \mapsto h(t, u) \text{ is weakly lower semicontinuous for a.e. } t \in (0, T),$$

and a sequence  $u_n \in \mathcal{M}(0, T; X)$  is said to be *weakly tight* if there exists a nonnegative weakly normal integrand  $h$  such that

$$(4.6) \quad \lim_{|u| \rightarrow \infty} h(t, u) = \infty \text{ for a.e. } t \in (0, T),$$

$$(4.7) \quad \text{and } \sup_n \int_0^T h(t, u_n(t)) dt < \infty.$$

We prepare a result which will turn out to be useful in the proof of Theorem 4.3.

LEMMA 4.2 (weak tightness implies tightness of the norms). *Let  $H$  be a separable and reflexive Banach space and  $u_n$  be weakly tight. Then  $|u_n|$  is tight in  $\mathbb{R}$ .*

*Proof.* Let  $h : (0, T) \times H \rightarrow [0, \infty]$  be a weakly normal integrand fulfilling (4.6)–(4.7), and define  $\tilde{h} : (0, T) \times [0, \infty] \rightarrow [0, \infty]$  as

$$(4.8) \quad \tilde{h}(t, r) := \inf_{|w| \geq r} h(t, w).$$

We shall prove that  $\tilde{h}$  is a normal integrand and that

$$(4.9) \quad \{r \geq 0 : \tilde{h}(t, r) \leq c\} \text{ is compact for a.e. } t \in (0, T) \text{ and all } c \geq 0,$$

$$(4.10) \quad \text{and} \quad \sup_n \int_0^T \tilde{h}(t, |u_n(t)|) dt < \infty.$$

Ad measurability: One directly checks that, given  $\alpha > 0$ , the set  $M = \{(t, u) \in (0, T) \times H : h(t, u) < \alpha\}$  belongs to  $\mathcal{L} \otimes \mathcal{B}(H)$  and hence  $N = \{(t, |u|) : (t, u) \in M\} \in \mathcal{L} \otimes \mathcal{B}([0, \infty))$ . Note that, for all Borel sets  $A \in \mathcal{L} \otimes \mathcal{B}([0, \infty))$ , the set  $\cup_{(t, \rho) \in A} \{t\} \times [0, \rho]$  belongs to  $\mathcal{L} \otimes \mathcal{B}([0, \infty))$  as well. Now we have

$$\tilde{h}^{-1}((-\infty, \alpha)) = \{(t, r) : \exists (t, u) \in M \text{ such that } |u| \geq r\} = \bigcup_{(t, \rho) \in N} \{t\} \times [0, \rho].$$

In particular  $\tilde{h}^{-1}((-\infty, \alpha)) \in \mathcal{L} \otimes \mathcal{B}([0, \infty))$ , and  $\tilde{h}$  is  $\mathcal{L} \otimes \mathcal{B}([0, \infty))$ -measurable.

Ad lower semicontinuity: We start by noting that  $r \mapsto \tilde{h}(t, r)$  is nondecreasing for all  $t \in (0, T)$ . Assume by contradiction that there exists  $t \in (0, T)$  such that  $u \mapsto h(t, u)$  is lower semicontinuous while  $r \mapsto \tilde{h}(t, r)$  is not, namely, that there exist an increasing sequence  $0 \leq r_n \rightarrow r$  and  $\delta > 0$  such that, for all  $n \in \mathbb{N}$ ,

$$(4.11) \quad \tilde{h}(t, r_n) + 2\delta \leq \tilde{h}(t, r).$$

Let now  $w_n \in H$  be such that  $|w_n| \geq r_n$  and  $\tilde{h}(t, r_n) + \delta \geq h(t, w_n)$  (such  $w_n$  exist by (4.8)). Then surely  $|w_n| \leq r$  as, if this was not the case, one would have

$$h(t, w_n) \stackrel{(4.8)}{\geq} \tilde{h}(t, r) \stackrel{(4.11)}{\geq} \tilde{h}(t, r_n) + 2\delta \geq h(t, w_n) + \delta.$$

Hence, we can extract a not relabeled subsequence  $w_n$  weakly converging to some  $w$  in  $H$  and compute

$$h(t, w) \stackrel{(4.5)}{\leq} \liminf_{n \rightarrow \infty} h(t, w_n) \leq \lim_{n \rightarrow \infty} \tilde{h}(t, r_n) + \delta \stackrel{(4.11)}{\leq} \tilde{h}(t, r) - 2\delta + \delta < \tilde{h}(t, r),$$

contradicting the very definition (4.8). Namely,  $r \mapsto \tilde{h}(t, r)$  is lower semicontinuous for almost every  $t \in (0, T)$ .

As a consequence, the sets  $\{(t, r) : \tilde{h}(t, r) \leq c\}$  are closed intervals for almost every  $t \in (0, T)$ . Moreover, they are also bounded for almost every  $t \in (0, T)$  due to (4.7), and we have proved (4.9). Finally, one easily checks that

$$\sup_n \int_0^T \tilde{h}(t, |u_n(t)|) dt \leq \sup_n \int_0^T h(t, u(t)) dt \stackrel{(4.7)}{<} \infty,$$

and (4.10) follows.  $\square$

A parametrized measure on  $X$  is a collection  $\nu = \{\nu_t\}_{t \in (0,T)}$  of Borel probability measures on  $X$  such that

$$t \mapsto \nu_t(B) \quad \text{is } \mathcal{L}\text{-measurable} \quad \forall B \in \mathcal{B}(X),$$

and the set of all parametrized measures is denoted by  $\mathcal{Y}(0, T; X)$ .

**THEOREM 4.3** ( $\Gamma$ -liminf tool). *Let  $H$  be a separable and reflexive Banach space and  $g_n, g_\infty : (0, T) \times H \rightarrow (-\infty, \infty]$  be weakly normal integrands such that*

$$(4.12) \quad g_\infty(t, u) \leq \inf \left\{ \liminf_{n \rightarrow \infty} g_n(t, u_n) : u_n \rightarrow u \text{ weakly in } H \right\} \\ \forall u \in H \quad \text{and for a.e. } t \in (0, T).$$

*Moreover let  $u_n$  be weakly tight. Then there exists a subsequence  $k \mapsto n_k$  and a parametrized measure  $\nu \in \mathcal{Y}(0, T; H)$  such that, for a.e.  $t \in (0, T)$ ,*

$$(4.13) \quad \nu_t \text{ is concentrated on the set } \bigcap_{j=1}^{\infty} \text{cl}_w \left( \{u_{n_k}(t) : k \geq j\} \right),$$

*where  $\text{cl}_w$  denotes the weak closure in  $H$ , and, whenever  $t \mapsto g_{n_k}^-(t, u_{n_k}(t)) = \max \{-g_{n_k}(t, u_{n_k}(t)), 0\}$  are uniformly integrable, namely,*

$$\lim_{|A| \rightarrow 0} \sup_{k \in \mathbb{N}} \int_A g_{n_k}^-(t, u_{n_k}(t)) \, dt = 0$$

*(the limit being restricted to Lebesgue measurable sets  $A \subset (0, T)$ ;  $|A|$  denotes the Lebesgue measure), we have*

$$\int_0^T \left( \int_H g_\infty(t, \xi) \, d\nu_t(\xi) \right) dt \leq \liminf_{k \rightarrow \infty} \int_0^T g_{n_k}(t, u_{n_k}(t)) \, dt.$$

*Proof.* The statement follows directly from the fundamental theorem by Balder [11, Thm. 1] adapted to weak topologies along the same lines of Rossi and Savaré [89, Thm. 3.2] (see also [64, Thm. B.1]). The idea is to rephrase the dependence of the functionals from  $n \in N := \mathbb{N} \cup \{\infty\}$  as an extra variable. As we shall see below, condition (4.12) ensures that the augmented integrand is still normal. The only difficulty arises from the fact that the weak topology of  $H$  is not globally metrizable (apart from finite-dimensional cases). By following closely the proof of [89, Thm. 3.2], we will circumvent this fact by considering the set

$$V := \{(u, r, n) \in H \times \mathbb{R} \times N : |u| \leq r\} \subset H \times \mathbb{R} \times N$$

and endowing it with the metric

$$d(v_1, v_2) := |||u_1 - u_2|||^2 + |r_1 - r_2| + |\arctan(n_1) - \arctan(n_2)| \\ \forall v_i = (u_i, r_i, n_i) \in V, \quad i = 1, 2.$$

In the latter, we have used the convention  $\arctan(\infty) = \pi/2$ , and, given a countable dense subset  $w_n$  of the unit ball in  $H^*$ , we have (classically) defined

$$|||u|||^2 := \sum_{k=0}^{\infty} 2^{-k} |(w_k, u)|^2 \quad \forall u \in H.$$

Hence,  $(V, d)$  is a separable and complete metric space since

$$\begin{aligned} v_k = (u_k, r_k, n_k) \rightarrow v = (u, r, n) \quad \text{in } V \quad \text{iff} \\ u_k \rightarrow u \text{ weakly in } H, \quad r_k \rightarrow r \text{ in } \mathbb{R}, \quad \text{and } n_k \rightarrow n \text{ in } N \end{aligned}$$

(where  $N$  is endowed with the arctan metric). Let us remark that any bounded closed set in  $(V, d)$  is compact with respect to the latter topology. Hence, all intersections of closed balls of  $H \times \mathbb{R} \times N$  with  $V$  are Borel subsets of  $V$ , namely,

$$(4.14) \quad B \in \mathcal{B}(H \times \mathbb{R} \times N) \quad \Rightarrow \quad B \cap V \in \mathcal{B}(V),$$

and any Borel measure on  $V$  can be trivially extended to a Borel measure on  $H \times \mathbb{R} \times N$ .

We apply Balder's theorem [11, Thm. 1] to the family  $v_n = (u_n, |u_n|, n)$  which turns out to be tight by Lemma 4.2 as  $N$  is compact. Hence, we find a subsequence  $k \mapsto n_k$  and a parametrized measure  $\mu = \{\mu_t\}_{t \in (0, T)} \in \mathcal{Y}(0, T; V)$  such that, for almost every  $t \in (0, T)$ ,

$$\mu_t \text{ is concentrated on the set } \Lambda(t) := \bigcap_{j=1}^{\infty} \text{cl}_V \left( \{v_{n_k}(t) : k \geq j\} \right),$$

namely, the set of  $V$ -limit points of  $v_{n_k}(t)$ .

Let now  $f : (0, T) \times V \rightarrow (-\infty, \infty]$  be defined by

$$f(t, v) := g_n(t, u) \quad \forall v = (u, r, n) \in V, \quad t \in (0, T).$$

Given any  $L \in \mathcal{L}, B \in \mathcal{B}(H)$ , and  $n \in N$ , by using (4.14) one checks that

$$(L \times B \times \mathbb{R} \times \{n\}) \cap ((0, T) \times V) \in \mathcal{L} \otimes \mathcal{B}(V).$$

Hence, we have

$$(\mathcal{L} \otimes \mathcal{B}(H) \times \mathbb{R} \times \{n\}) \cap ((0, T) \times V) \subset \mathcal{L} \otimes \mathcal{B}(V).$$

On the other hand, for all  $a \in \mathbb{R}$ ,

$$\begin{aligned} & \{(t, v) : f(t, v) \leq a\} \\ &= \bigcup_{n=1}^{\infty} \left( (\{(t, u) : g_n(t, u) \leq a\} \times \mathbb{R} \times \{n\}) \cap ((0, T) \times V) \right) \in \mathcal{L} \otimes \mathcal{B}(V), \end{aligned}$$

since all elements under the union sign are in  $\mathcal{L} \otimes \mathcal{B}(V)$ . Hence, the function  $f$  is  $\mathcal{L} \otimes \mathcal{B}(V)$ -measurable. Moreover,  $f$  is a normal integrand. Indeed, let  $v_k \rightarrow v$  in  $V$ . Then either  $n_k = n$  definitely or  $n_k \rightarrow \infty$ . In the first case, lower semicontinuity follows from the weak sequential lower semicontinuity of  $g_n$ , whereas in the second case it can be deduced from (4.12).

Since  $t \mapsto f^-(t, v_{n_k}(t)) = g_{n_k}^-(t, u_{n_k}(t))$  are uniformly integrable, again by [11, Thm. 1] we have

$$\begin{aligned} & \int_0^T \left( \int_V f(t, \zeta) \, d\mu_t(\zeta) \right) dt \leq \liminf_{k \rightarrow \infty} \int_0^T f(t, v_{n_k}(t)) \, dt \\ (4.15) \quad &= \liminf_{k \rightarrow \infty} \int_0^T g_{n_k}(t, u_{n_k}(t)) \, dt. \end{aligned}$$



By recalling that  $\Lambda(t) \subset H \times \mathbb{R} \times \{\infty\}$  for all  $t \in (0, T)$  and letting

$$\nu_t(B) := \mu_t((B \times \mathbb{R} \times N) \cap V) \quad \forall B \in \mathcal{B}(H),$$

we obtain a parametrized measure  $\nu = \{\nu_t\}_{t \in (0, T)} \in \mathcal{Y}(0, T; H)$  which fulfills (4.13) and is such that

$$\int_0^T \left( \int_V f(t, \zeta) d\mu_t(\zeta) \right) dt = \int_0^T \left( \int_H g_\infty(t, \xi) d\nu_t(\xi) \right) dt,$$

which, together with (4.15), entails the result.  $\square$

Let us remark that, under the frame of Theorem 4.3, whenever  $u_n \rightarrow u$  weakly in  $L^p(0, T; H)$  for some  $p \in [1, \infty)$  (weakly star for  $p = \infty$ ), then

$$(4.16) \quad u(t) = \int_H \xi d\nu_t(\xi) \quad \text{for a.e. } t \in (0, T).$$

This fact was already remarked in [89, Thm. 3.2] for  $p \in (1, \infty)$  (and hence, for  $p = \infty$  as well). As for  $p = 1$ , one can readily choose the weakly normal integrands

$$g(t, u) := (w(t), u) \quad \forall u \in H, \text{ a.e. } t \in (0, T), \quad \text{with } w \in L^\infty(0, T; H^*),$$

and exploit Theorem 4.3 with the constant sequence  $g_n = g$  (or [64, Thm. B.1]) in order to conclude.

**4.3. The  $\Gamma$ -lim inf result for normal convex integrands.** Let us now specify the result of Theorem 4.3 in the case of normal convex integrands  $g_n, g_\infty : [0, T] \times H \rightarrow (-\infty, \infty]$ .

**COROLLARY 4.4.** *Let  $p \in [1, \infty]$ ,  $H$  be a separable and reflexive Banach space, and  $g_n, g_\infty : (0, T) \times H \rightarrow (-\infty, \infty]$  be normal convex integrands such that (4.12) holds. Moreover, let  $u_n \rightarrow u$  weakly in  $L^p(0, T; H)$  (weakly star if  $p = \infty$ ). Then, whenever  $t \mapsto g_n^-(t, u_n(t))$  are uniformly integrable, we have*

$$(4.17) \quad \int_0^T g_\infty(t, u(t)) dt \leq \liminf_{n \rightarrow \infty} \int_0^T g_n(t, u_n(t)) dt.$$

*Proof.* Let  $j \mapsto n_j \in \mathbb{N}$  be an increasing sequence. As  $u_n$  are uniformly bounded in  $L^p(0, T; H)$ , the family  $u_{n_j}$  is weakly tight. Hence, by applying Theorem 4.3, we may extract a further subsequence  $k \mapsto n_{j_k}$  such that

$$(4.18) \quad \begin{aligned} \liminf_{k \rightarrow \infty} \int_0^T g_{n_{j_k}}(t, u_{n_{j_k}}(t)) dt &\geq \int_0^T \left( \int_H g_\infty(t, \xi) d\nu_t(\xi) \right) dt \\ &\geq \int_0^T g_\infty(t, u(t)) dt, \end{aligned}$$

where the last inequality follows from (4.16) and Jensen's inequality. Namely, for all subsequences of  $u_n$  there exist further subsequences such that (4.18) holds. An easy argument ensures that indeed (4.18) holds for the whole sequence  $g_n(\cdot, u_n(\cdot))$  as well.  $\square$

Let us now comment on the uniform integrability condition of Corollary 4.4. First of all, it is clear that the latter holds if  $g_n$  are uniformly bounded below. More generally, one can consider the case

$$(4.19) \quad \begin{aligned} g_n(t, u) &\geq -c_0|u| - \gamma(t) \quad \text{for some } c_0 > 0, \gamma \in L^1(0, T), \\ &\forall u \in H, n \in \mathbb{N}, \text{ for a.e. } t \in (0, T). \end{aligned}$$

In fact, whenever  $u_n$  converges weakly in  $L^p(0, T; H)$  (weakly star in  $p = \infty$ ), the functions  $t \mapsto |u_n(t)|$  are uniformly integrable [28, Thm. 4, p. 104], and (4.19) entails the uniform integrability of  $t \mapsto g_n^-(t, u_n(t))$ .

We shall explicitly remark that, in case  $g_n$  are independent of time, the lower bound (4.19) follows directly from the condition (4.12). Indeed, owing to [7, Thm. 3.7, p. 271], by letting  $v \in D(g_\infty^*)$  be fixed, there exist  $v_n$  such that  $v_n \rightarrow v$  strongly in  $H^*$  and  $\limsup_{n \rightarrow \infty} g_n^*(v_n) = g_\infty^*(v)$ . Then (4.19) follows with the choice

$$c_0 = \sup_n |v_n|_*, \quad \gamma(t) = g_\infty^*(v) + 1.$$

In particular, we have the following.

**COROLLARY 4.5** ( $g_n$  independent of time). *Let  $p \in [1, \infty]$ ,  $H$  be a separable and reflexive Banach space, and  $g_n, g_\infty : H \rightarrow (-\infty, \infty]$  be convex, proper, and lower semicontinuous such that (4.12) holds. Moreover, let  $u_n \rightarrow u$  weakly in  $L^p(0, T; H)$  (weakly star if  $p = \infty$ ). Then we have*

$$(4.20) \quad \int_0^T g_\infty(u(t)) \, dt \leq \liminf_{n \rightarrow \infty} \int_0^T g_n(u_n(t)) \, dt.$$

Before moving on, we remark that some result in the direction of Corollary 4.4 was already contained in the convergence analysis by Salvadori [92, Thm. 3.1]. The latter was focused on establishing conditions under which the integral functionals

$$G_n(u) = \begin{cases} \int_0^T g_n(t, u(t)) \, dt & \text{if } t \mapsto g_n^+(t, u(t)) \in L^1(0, T), \\ \infty & \text{otherwise} \end{cases}$$

would Mosco-converge to the limit functional

$$G_\infty(u) = \begin{cases} \int_0^T g_\infty(t, u(t)) \, dt & \text{if } t \mapsto g_\infty^+(t, u(t)) \in L^1(0, T), \\ \infty & \text{otherwise} \end{cases}$$

(and analogously for  $G_n^*$  and  $G_\infty^*$ , which are defined from  $g_n^*$  and  $g_\infty^*$ , respectively). The Mosco-convergence result in [92, Thm. 3.2] was obtained under some uniform quantitative properness of the functionals. In particular, both sequences  $G_n$  and  $G_n^*$  are asked to be uniformly proper on  $L^p(0, T; H)$  and  $L^q(0, T; H^*)$ , respectively. Moreover, by letting  $u_n$  and  $v_n$  be the corresponding bounded sequences, the existence of two functions  $f, f_* \in L^1(0, T)$  such that

$$|g_n(t, u_n(t))| \leq f(t) \quad \text{and} \quad |g_n^*(t, v_n(t))| \leq f_*(t) \quad \text{for a.e. } t \in (0, T)$$

is required.

The frame of Corollary 4.4 is quite weaker, since we are not assuming any control on the domains of the functionals but rather some standard uniform integrability of negative parts of the integrands. Hence, by exploiting the characterization of Lemma 4.1 and restricting to the case where  $p \in (1, \infty)$ , we can obtain a refined version of [92, Thm. 3.1] as follows.

**COROLLARY 4.6.** *Let  $p \in (1, \infty)$ ,  $H$  be a separable and reflexive Banach space, and  $g_n, g_\infty : (0, T) \times H \rightarrow (-\infty, \infty]$  be normal convex integrands such that  $g_n \rightarrow g_\infty$*

in the Mosco sense. Moreover, assume that

$$(4.21) \quad \begin{aligned} & t \rightarrow g_n^-(t, u_n(t)) \quad \text{and} \quad t \rightarrow (g_n^*)^-(t, v_n(t)) \quad \text{are uniformly integrable} \\ & \forall (u_n, v_n) \rightarrow (u, v) \text{ weakly in } L^p(0, T; H) \times L^q(0, T; H^*), \\ & G_n \quad \text{and} \quad G_n^* \quad \text{are proper on } L^p(0, T; H) \quad \text{and} \quad L^q(0, T; H^*), \text{ respectively,} \\ (4.22) \quad & \text{and either } G_n \text{ or } G_n^* \text{ is uniformly proper.} \end{aligned}$$

Then  $G_n \rightarrow G_\infty$  in the Mosco sense in  $L^p(0, T; H)$  and  $G_n^* \rightarrow G_\infty^*$  in the Mosco sense in  $L^q(0, T; H^*)$ . In particular, both sequences  $G_n$  and  $G_n^*$  are uniformly proper.

*Proof.* Owing to (4.21), Corollary 4.4 gives the  $\Gamma$ -liminf inequalities for  $G_n$  and  $G_n^*$ . Owing to the properness of  $G_n$  and  $G_n^*$ , we have that  $(G_n)^* = G_n^*$  and  $(G_n^*)^* = G_n$  [21, Thm. VII.7, p. 200]. Since  $L^p(0, T; H)$  is reflexive, the assertion follows from the uniform properness in (4.22) and Lemma 4.1.  $\square$

In the same spirit of Corollary 4.5, whenever  $g_n$  are independent of time, both the uniform integrability condition (4.21) and the uniform properness condition (4.22) are straightforward.

**5. Lim inf inequalities.** We shall now apply this functional convergence machinery to our problems. Throughout the remainder of the paper, we will assume that

$$(A6) \quad \begin{aligned} & p \in [1, \infty], \quad 1/p + 1/q = 1, \text{ and} \\ & H \text{ is a real, separable, and reflexive Banach space.} \end{aligned}$$

As for the sequences of approximating functionals we will systematically ask that

$$(5.1) \quad \begin{aligned} & t \rightarrow g_n^-(t, u_n(t)) \quad \text{and} \quad t \rightarrow (g_n^*)^-(t, v_n(t)) \quad \text{are uniformly integrable} \\ & \forall (u_n, v_n) \rightarrow (u, v) \text{ weakly (star) in } L^p(0, T; H) \times L^q(0, T; H^*) \end{aligned}$$

for the choices  $g_n = \phi_n, \psi_n, \varphi_n$  and admit the limiting cases  $p = 1, \infty$  as well.

Let us start from the case of the gradient flow (1.1). Assume that we are given data  $\phi_n$ ,  $f_n$ , and  $u_0^n$  as in section 1, and define the corresponding approximating functionals  $J_n, K_n : H^1(0, T; H) \rightarrow [0, \infty]$  for all  $u \in H^1(0, T; H)$  as

$$\begin{aligned} J_n(u) &:= \int_0^T (\phi_n(u) + \phi_n^*(f_n - u')) - (f_n, u) \\ &\quad + \frac{1}{2}|u(T)|^2 - \frac{1}{2}|u(0)|^2 + |u(0) - u_0^n|^2, \\ K_n(u) &:= \left( \int_0^T (|u'|^2 - (f_n, u')) + \phi_n(u(T)) - \phi_n(u_0^n) \right)^+ + J_n(u). \end{aligned}$$

We have the following.

**LEMMA 5.1** (lim inf inequality for  $K_n$ ). *Assume (A6), and let  $H$  be a Hilbert space,  $\phi_n \rightarrow \phi$  in the Mosco sense,  $f_n \rightarrow f$  strongly in  $L^2(0, T; H)$ ,  $u_0^n \rightarrow u^0$  in  $H$ , and  $\phi_n(u_0^n) \rightarrow \phi(u^0)$ . Then, for all  $u \in H^1(0, T; H)$ ,*

$$K(u) \leq \inf_{n \rightarrow \infty} \{ \liminf K_n(u_n) : u_n \rightarrow u \text{ weakly in } H^1(0, T; H) \}.$$

*Proof.* By applying Corollary 4.5 we readily check that

$$\int_0^T (\phi(u) + \phi^*(f - u')) \leq \liminf_{n \rightarrow \infty} \int_0^T (\phi_n(u_n) + \phi_n^*(f_n - u'_n)).$$

On the other hand, owing to the strong convergence of  $f_n$  we have

$$\int_0^T (f_n, u_n) \rightarrow \int_0^T (f, u) \quad \text{and} \quad \int_0^T (f_n, u'_n) \rightarrow \int_0^T (f, u'),$$

and, by the pointwise weak convergences  $u_n(T) \rightarrow u(T)$  and  $u_n(0) \rightarrow u(0)$  in  $H$  and the Mosco convergence of  $\phi_n$ , one obtains

$$\begin{aligned} \phi(u(T)) &\leq \liminf_{n \rightarrow \infty} \phi_n(u_n(T)) \quad \text{and} \\ \frac{1}{2}|u(T)|^2 - \frac{1}{2}|u(0)|^2 + |u(0) - u^0|^2 &= \frac{1}{2}|u(T)|^2 + \frac{1}{2}|u(0)|^2 + |u^0|^2 - 2(u(0), u^0) \\ &\leq \liminf_{n \rightarrow \infty} \left( \frac{1}{2}|u_n(T)|^2 + \frac{1}{2}|u_n(0)|^2 + |u_n^0|^2 - 2(u_n(0), u_n^0) \right) \\ &= \liminf_{n \rightarrow \infty} \left( \frac{1}{2}|u_n(T)|^2 - \frac{1}{2}|u_n(0)|^2 + |u_n(0) - u_n^0|^2 \right). \end{aligned}$$

Finally, the assertion follows by lower semicontinuity.  $\square$

In fact, an analogous result holds for  $J_n$  as well, the convergence of the initial energies  $\phi_n(u_n^0)$  not being needed. We prefer to state the  $\liminf$  inequality for  $K_n$  since, as already remarked, the sublevels of  $K_n$  are uniformly bounded in  $H^1(0, T; H)$  whereas those of  $J_n$  are not, in general.

Let us now move to the doubly nonlinear situation by fixing  $p \in [1, \infty]$  and recalling that  $1/p + 1/q = 1$ . Assume that we are given  $\phi_n, \psi_n, f_n$ , and  $u_n^0$  as in section 2, and define the corresponding approximating functionals  $I_n : W^{1,p}(0, T; H) \times L^q(0, T; H^*) \rightarrow [0, \infty]$  for all  $(u, v) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$  as

$$\begin{aligned} I_n(u, v) &:= \left( \int_0^T (\psi_n(u') + \psi_n^*(v) - (f_n, u')) + \phi_n(u(T)) - \phi_n(u_n^0) \right)^+ \\ &\quad + \int_0^T (\phi_n(u) + \phi_n^*(f_n - v) - (f_n - v, u)) + |u(0) - u_n^0|^2. \end{aligned}$$

Hence, by recalling (4.1), we have the following lemma.

**LEMMA 5.2** ( $\liminf$  inequality for  $I_n$ ). *Assume (A6), and let  $\phi_n \rightarrow \phi$  and  $\psi_n \rightarrow \psi$  in the Mosco sense and fulfill (5.1),  $f_n \rightarrow f$  strongly in  $L^q(0, T; H^*)$ ,  $u_n^0 \rightarrow u^0$  in  $H$ , and  $\phi_n(u_n^0) \rightarrow \phi(u^0)$ . Then, for all  $(u, v) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$ ,*

$$I(u, v) \leq \inf \left\{ \liminf_{n \rightarrow \infty} I_n(u_n, v_n) : (u_n, v_n) \rightarrow (u, v) \text{ in } \mathcal{T}(p) \right\}.$$

*Proof.* The proof of the latter lemma follows along the very same lines as that of Lemma 5.1 by systematically considering the  $p - q$  frame and additionally applying once again Corollary 4.5 in order to deduce that

$$\int_0^T (\psi(u) + \psi^*(v)) \leq \liminf_{n \rightarrow \infty} \int_0^T (\psi_n(u_n) + \psi_n^*(v_n)). \quad \square$$

We shall explicitly mention that, in the separable Hilbert-space setting and in the (quadratic) case of (2.3), namely, for  $\phi(\cdot) = |\cdot|^2/2$ , the above  $\liminf$  inequality holds also in some stronger form. By introducing the approximating functionals  $Q_n :$

$W^{1,p}(0, T; H) \rightarrow [0, \infty]$  as

$$Q_n(u) := \int_0^T (\psi_n(u') + \psi_n^*(f_n - u) - (f_n, u')) \\ + \frac{1}{2}|u(T)|^2 - \frac{1}{2}|u(0)|^2 + |u(0) - u_n^0|^2,$$

we have the following.

LEMMA 5.3 (lim inf inequality for  $Q_n$ ). *Assume that  $H$  is a separable Hilbert space,  $\psi_n \rightarrow \psi$  in the Mosco sense and fulfill (5.1),  $f_n \rightarrow f$  strongly in  $L^q(0, T; H)$ , and  $u_n^0 \rightarrow u^0$  in  $H$ . Then, for all  $u \in W^{1,p}(0, T; H)$ ,*

$$(5.2) \quad Q(u) \leq \inf \left\{ \liminf_{n \rightarrow \infty} Q_n(u_n) : u_n \rightarrow u \text{ weakly (star) in } W^{1,p}(0, T; H) \right\}.$$

The assertion follows via the same arguments of the proofs above, this case being indeed simplified since  $\phi$  is quadratic.

Finally, we consider the generalized situation of (3.1) by letting  $\psi_n, \varphi_n, \pi_n$ , and  $u_n^0$  be as in section 3 and defining the functionals  $\mathcal{I}_n : W^{1,p}(0, T; H) \times L^q(0, T; H^*) \rightarrow [0, \infty]$  for all  $(u, v) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$  as

$$\mathcal{I}_n(u, v) := \left( \int_0^T (\psi_n(u') + \psi_n^*(v) - \pi_n(u)) + \varphi_n(T, u(T)) - \varphi_n(0, u_n^0) \right)^+ \\ + \int_0^T (\varphi_n(\cdot, u) + \varphi_n^*(\cdot, -v) + (u, v)) + |u(0) - u_n^0|^2.$$

In order to establish a lim inf inequality result for  $\mathcal{I}_n$ , the limiting behavior of  $\pi_n$  has to be prescribed. We shall directly ask that, for all  $(u, v) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$ ,

$$(5.3) \quad \int_0^T \pi(u) \geq \sup \left\{ \limsup_{n \rightarrow \infty} \int_0^T \pi_n(u_n) : (u_n, v) \rightarrow (u, v) \text{ in } \mathcal{T}(p) \right\}.$$

The latter is readily fulfilled if  $\varphi_n(t, u) = \phi_n(u) - (f_n(t), u)$ ,  $\varphi(t, u) = \phi(u) - (f(t), u)$ , and  $f_n \rightarrow f$  strongly in  $W^{1,p}(0, T; H^*)$ ; see Remark 3.2. We have the following.

LEMMA 5.4 (lim inf inequality for  $\mathcal{I}_n$ ). *Assume (A6), and let  $\varphi_n \rightarrow \varphi$  and  $\psi_n \rightarrow \psi$  in the Mosco sense and fulfill (5.1) and (5.3),  $u_n^0 \rightarrow u^0$  in  $H$ , and  $\varphi_n(0, u_n^0) \rightarrow \varphi(0, u^0)$ . Then, for all  $(u, v) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$ ,*

$$\mathcal{I}(u, v) \leq \inf \left\{ \liminf_{n \rightarrow \infty} \mathcal{I}_n(u_n, v_n) : (u_n, v_n) \rightarrow (u, v) \text{ in } \mathcal{T}(p) \right\}.$$

*Sketch of the proof.* This proof differs from that of Lemma 5.2 for the sole use of (5.3) instead of the strong convergence of  $f_n$ .  $\square$

Before closing this section, let us explicitly consider the case of the nonautonomous gradient flow (3.7). To this aim, let the approximating functionals  $\mathcal{K}_n : H^1(0, T; H) \rightarrow [0, \infty]$  be defined as

$$\mathcal{K}_n(u) := \left( \int_0^T (|u'|^2 - \pi_n(u)) + \varphi_n(T, u(T)) - \varphi_n(0, u_n^0) \right)^+ \\ + \int_0^T (\varphi_n(\cdot, u) + \varphi_n^*(\cdot, -u')) + \frac{1}{2}|u(T)|^2 - \frac{1}{2}|u(0)|^2 + |u(0) - u_n^0|^2.$$

In order to possibly obtain some stronger convergence result in this case, we shall need some additional convergence property for  $\pi_n$ . Namely, we ask, for all  $u \in H^1(0, T; H)$ , that

$$(5.4) \quad \int_0^T \pi(u) \geq \sup \left\{ \limsup_{n \rightarrow \infty} \int_0^T \pi_n(u_n) : u_n \rightarrow u \text{ weakly in } H^1(0, T; H) \right\}.$$

Again, in the case where  $\varphi_n(t, u) = \phi_n(u) - (f_n(t), u)$ ,  $\varphi(t, u) = \phi(u) - (f(t), u)$ , the latter is fulfilled if  $f_n \rightarrow f$  strongly in  $H^1(0, T; H^*)$ . Hence, we have the following.

LEMMA 5.5 (lim inf inequality for  $\mathcal{K}_n$ ). *Assume (A6), and let  $\varphi_n \rightarrow \varphi$  in the Mosco sense and fulfill (5.1) and (5.4),  $u_n^0 \rightarrow u^0$  in  $H$ , and  $\varphi_n(0, u_n^0) \rightarrow \varphi(0, u^0)$ . Then, for all  $u \in H^1(0, T; H)$ ,*

$$\mathcal{K}(u) \leq \inf \left\{ \liminf_{n \rightarrow \infty} \mathcal{K}_n(u_n) : u_n \rightarrow u \text{ weakly in } H^1(0, T; H) \right\}.$$

The proof of the latter is obtained by easily adapting the arguments of Lemmas 5.1 and 5.4. Let us stress that, in case the weaker convergence (5.3) holds, a lim inf inequality for  $\mathcal{K}_n$  is still available as a corollary to Lemma 5.4.

**6. Approximation of gradient flows.** We shall exploit Theorems 2.1 and 3.1 and Lemmas 5.1 and 5.5 in order to recover and further generalize some approximation results for gradient flows under the separability assumption in (A6). By assuming the above notations and recalling the uniform coercivity of  $K_n$  with respect to the weak topology in  $H^1(0, T; H)$ , we obtain a first convergence result which we state below by omitting the easy proof.

LEMMA 6.1 (convergence for gradient flows). *Under the assumptions of Lemma 5.1, let  $u_n \in H^1(0, T; H)$  be such that  $K_n(u_n) \rightarrow 0$ . Then  $u_n \rightarrow u$  weakly in  $H^1(0, T; H)$  and  $K(u) = 0$ .*

Note in particular that the whole approximating sequence  $u_n$  converges since  $K$  admits a unique minimizer  $u$ .

By reducing ourselves to the case  $K_n(u_n) = 0$  (i.e., letting  $u_n$  be solutions to the respective differential problems), the above lemma recovers the result by Attouch on the approximation of gradient flows under the Mosco convergence of the functionals [7, Thm. 3.74(2), p. 388] under the separability assumption in (A6) (here no strong  $H^1(0, T; H)$  nor energy convergence is proved, though). Let us, however, remark that our result turns out to be slightly more general than the former since  $u_n$  are a priori not required to be solutions at level  $n$ . In particular, the functions  $u_n$  could be approximated solutions of the corresponding gradient flows as well. Moreover, the functional frame is here extended from (separable) Hilbert to separable reflexive Banach spaces. One has, however, to mention that the specific case of regularization by means of the Yosida approximation (in Hilbert spaces) was already discussed within the existence proof by Ghoussoub and Tzou [39]. In case the convergence of the initial energies  $\phi_n(u_n^0)$  does not hold, one is still in the position of proving a convergence result by arguing on  $J_n$  if  $J_n(u_n) = 0$  (or, more generally, in case of a weakly  $H^1(0, T; H)$ -precompact sequence  $u_n$  such that  $J_n(u_n) \rightarrow 0$ ).

Our second convergence result concerns the generalized situation of (3.7). Again, the following lemma is implied by the fact that, by assuming  $\pi_n$  to be uniformly linearly bounded,  $\mathcal{K}_n$  are uniformly coercive with respect to the weak topology of  $H^1(0, T; H)$ .

LEMMA 6.2 (convergence for generalized gradient flows). *Under the assumptions of Lemma 5.5, let  $\pi_n$  be uniformly linearly bounded and  $u_n \in H^1(0, T; H)$  be such that  $\mathcal{K}_n(u_n) \rightarrow 0$ . Then  $u_n \rightarrow u$  weakly in  $H^1(0, T; H)$  and  $\mathcal{K}(u) = 0$ .*

The latter convergence result for the nonautonomous gradient flow (3.7) is to be compared with the former results by Ortner [79, sect. 3.2] which hold in the general metric and  $\lambda$ -geodesically convex setting but under more restrictive functional convergence assumptions (see also [80]). We shall, however, stress that here the approximating  $u_n$  need not be solutions to the corresponding differential problems at level  $n$ .

Before closing this section, we shall mention the work by Mabrouk [52, 53] where the Brezis–Ekeland principle is exploited within an approximation procedure in order to establish the existence of generalized solutions to some semilinear parabolic equation with measure data (see also the results by the same author [54, 55] for some second order in time equations). Moreover, we mention that some identification result for nonlinear parabolic problems based on (a variational technique related to) the Brezis–Ekeland principle has been obtained by Barbu and Kunisch [14]. Finally, the issue of approximating nonconvex gradient flows has recently attracted some attention (see Ambrosio, Gigli, and Savaré [2] and Sandier and Serfaty [93]).

**7. Approximation of doubly nonlinear equations.** Let us now move to the situation of the doubly nonlinear relation (1.3) and fix from the very beginning and throughout this section

$$p \in (1, \infty).$$

For Lemma 5.2 to serve as the basis for a convergence result, one just needs to provide coercivity for  $I_n$  with respect to the topology  $\mathcal{T}(p)$  (see (4.1)). The latter holds, for instance, in the situation of potentials  $\psi_n$  of  $p$ -growth and functionals  $\phi_n$  with compact sublevels. In particular, we let

$$(7.1) \quad c_1|w|^p - c_2 \leq \psi(w) \leq c_3(|w|^p + 1) \quad \forall w \in H,$$

$$(7.2) \quad \phi(w) \geq c_4\|w\|^p - c_5 \quad \forall w \in V \subset H,$$

where the injection of the reflexive Banach space  $V$  into  $H$  is compact,  $\|\cdot\|$  is the norm in  $V$ , and  $c_1, c_3, c_4 > 0$ ,  $c_2, c_5 \geq 0$  are given. In this case, it may be checked that  $c_6|w|_*^q - c_7 \leq \psi^*(w)$  for all  $w \in H^*$  and with some constants  $c_6, c_7 > 0$  depending on  $c_3$  and  $p$  ( $|\cdot|_*$  is the norm in  $H^*$ ). Hence, all sublevels of  $I$  are relatively compact with respect to the topology  $\mathcal{T}(p)$  by means of well-known compactness results (see, e.g., Simon [96]). Namely, we have the following.

**LEMMA 7.1** (convergence for doubly nonlinear equations). *Under the assumptions of Lemma 5.2, let  $\phi_n$  and  $\psi_n$  fulfill (7.1)–(7.2) uniformly with respect to  $n$ . Moreover, let  $(u_n, v_n) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$  be such that  $I_n(u_n, v_n) \rightarrow 0$ . Then there exists a (not relabeled) subsequence such that  $(u_n, v_n) \rightarrow (u, v)$  in  $\mathcal{T}(p)$  and  $I(u, v) = 0$ .*

Here the convergence of the whole sequence  $(u_n, v_n)$  cannot be expected since the limiting minimum problem for  $I$  need not admit a unique minimizer.

By restricting to the case of a sequence  $(u_n, v_n)$  of solutions to the approximating doubly nonlinear problem (i.e., imposing  $I_n(u_n, v_n) = 0$ ), we recover the convergence result of Aizicovici and Yan [1, Thm. 3.1] under the extra separability assumption in (A6). The referred result is in fact slightly stronger, since the subdifferentials  $\partial\psi_n$  are replaced by general maximal monotone operators  $A_n$  and the growth and compactness requirements are weaker (although quite similar). On the other hand, under assumptions (7.1)–(7.2), our result turns out to be more general than the former, since the approximating  $(u_n, v_n)$  need not be solutions to the corresponding equations. This fact allows some possible extra freedom in the choice of the approximating sequence.

Finally, we are in the position of providing a convergence lemma which applies to the generalized situation of (3.1). The following is to our knowledge the first result in this direction.

**THEOREM 7.2** (convergence for generalized doubly nonlinear equations). *Under the assumptions of Lemma 5.4, let  $\varphi_n(t, \cdot)$  and  $\psi_n$  fulfill (7.1)–(7.2) for almost every  $t \in (0, T)$  and uniformly with respect to  $n$ . Moreover, let  $(u_n, v_n) \in W^{1,p}(0, T; H) \times L^q(0, T; H^*)$  be such that  $\mathcal{I}_n(u_n, v_n) \rightarrow 0$ . Then there exists a (not relabeled) subsequence such that  $(u_n, v_n) \rightarrow (u, v)$  in  $\mathcal{T}(p)$  and  $\mathcal{I}(u, v) = 0$ .*

**LEMMA 7.3** (convergence for case  $\phi(\cdot) = \phi_n(\cdot) = |\cdot|^2/2$ ). *Under the assumptions of Lemma 5.3, for  $p \in (1, \infty]$  let  $\psi_n$  fulfill (7.1) uniformly with respect to  $n$ . Moreover, let  $u_n \in W^{1,p}(0, T; H)$  be such that  $Q_n(u_n) \rightarrow 0$ . Then there exists a (not relabeled) subsequence such that  $u_n \rightarrow u$  weakly in  $W^{1,p}(0, T; H)$  and  $Q(u) = 0$ .*

Note that, in the frame of Lemma 7.3, if  $f \in W^{1,1}(0, T; H)$ , then the solution of (2.3) is unique and the whole sequence  $u_n$  of the statement converges.

**7.1. Existence for some doubly nonlinear equation via the Brezis–Ekeland approach.** We shall now apply the above-developed convergence theory in order to possibly (re)obtain the existence of solutions of some specific doubly nonlinear equation (1.3) via the variational characterization of Theorem 3.1. For the sake of simplicity, we reduce our attention to the Hilbert-space framework of Colli and Visintin [23]. More specifically, we shall ask that

(7.3)  $H$  is a real Hilbert space and  $p = 2$ .

**LEMMA 7.4.** *Under assumptions (A2) and (A3) and (7.1)–(7.3), there exists a solution  $(u, v)$  of (1.4)–(1.6).*

Note that the latter stands as a weaker version of [23, Thm. 2.1] where indeed  $\partial\psi$  is replaced by a general maximal monotone operator and some weaker coercivity assumption on  $\phi$  is considered.

*Proof.* For all  $n \in \mathbb{N}$  let

$$\psi_n(u) = \frac{1}{2n}|u|^2 + \psi(u) \quad \text{and} \quad \phi_n(u) = \inf_{v \in H} \left( \frac{n}{2}|v - u|^2 + \phi(v) \right).$$

Namely,  $\phi_n$  is the Yosida approximation of  $\phi$  at level  $1/n$ . The corresponding regularized problem reads now as follows:

$$v \in \frac{1}{n}u' + \partial\psi(u'), \quad v + \partial\phi_n(u) = f \quad \text{a.e. in } (0, T), \quad u(0) = u^0,$$

and clearly admits a unique solution  $(u_n, v_n) \in H^1(0, T; H) \times L^2(0, T; H)$  since  $(1/n + \partial\psi)^{-1} \circ \partial\phi_n$  is Lipschitz continuous. Hence, by Theorem 2.1 we have that  $\mathcal{I}_n(u_n, v_n) = 0$ .

We now aim at applying the  $\liminf$  result of Lemma 5.2. First of all, we have  $\psi_n \rightarrow \psi$  and  $\phi_n \rightarrow \phi$  in the Mosco sense [7, Thm. 3.26, p. 305]. Second, condition (5.1) is trivially satisfied by  $\psi_n$  since we are requiring (7.1). In particular,  $\psi \leq \psi_n$  and  $\psi_n^* \geq -\psi(0)$ . Moreover, owing to the fact that  $(\phi_n)^*(u) = \phi^*(u) + |u|^2/(2n)$  (see, for instance, [42, subsect. 4.1]), we have

$$\phi_n(u) = \sup_{v \in H} \left( (v, u) - \phi^*(v) - \frac{1}{2n}|v|^2 \right) \geq (w, u) - \phi^*(w) - \frac{1}{2}|w|^2$$

for any fixed  $w \in D(\phi^*)$ . Finally, the fact that

$$(\phi_n)^*(u) \geq \phi^*(u) \geq (u, z) - \phi(z) \quad \forall u \in H,$$

for some  $z \in D(\phi)$  fixed, entails that condition (5.1) holds for  $\phi_n$  as well.



It is a standard matter to determine new constants  $c'_1, c'_2$ , and  $c'_3$  in such a way that  $\psi_n$  fulfills the corresponding nondegeneracy and growth assumption (7.1) uniformly with respect to  $n$ . In particular, we have the fact that

$$(7.4) \quad u'_n \text{ and } v_n \text{ are bounded in } L^2(0, T; H) \text{ independently of } n.$$

Since  $D(\phi_n) = H$ , the coercivity (7.2) cannot hold at level  $n$ . In fact one can check that

$$\phi_n(u) = \frac{1}{2n} |\partial \phi_n(u)|^2 + \phi(j_n u) \geq c_4 \|j_n u\|^2 - c_5 \quad \forall u \in H,$$

where  $j_n := (1 + (1/n)\partial\phi)^{-1}$  is the resolvent. By recalling that  $j_n$  is Lipschitz continuous, uniformly with respect to  $n$ , we have the fact that

$$(7.5) \quad j_n u_n \text{ is bounded in } H^1(0, T; H) \cap L^2(0, T; V) \text{ independently of } n.$$

The bounds (7.4)–(7.5) imply that  $(u_n, v_n)$  is precompact in  $\mathcal{T}(2)$ . By extracting a not relabeled subsequence  $(u_n, v_n) \rightarrow (u, v)$  in  $\mathcal{T}(2)$  and applying Lemma 5.2, we get  $I(u, v) = 0$ , and the assertion follows from Theorem 2.1.  $\square$

Let us comment that the Hilbert-space frame of (7.3) is chosen as a possible first illustration of this technique and that the Brezis–Ekeland approach applies to the more general Banach case as well. However, in the latter case, one should consider a time-discretized problem rather than a regularized one (this was exactly the strategy in [22]). The development of a discrete version of the characterization of Theorem 3.1 is presented and applied to the convergence of time discretizations for (1.3) in [101]. As a by-product, the existence of solutions to the doubly nonlinear equation (1.3) in the Banach framework is there recovered by a purely variational technique.

**8. Approximation of rate-independent evolutions.** Let us now focus on a specific class of potentials  $\psi$  of growth  $p = 1$  (see (7.1)). In particular, in addition to (A4) we shall ask that

$$(8.1) \quad \psi \text{ is positively homogeneous of degree 1,}$$

letting the evolution problem be rate-independent (see Mielke [62]). Equivalently,  $\psi$  is required to be the support function of a convex and closed set  $C \subset H^*$  containing 0, namely,

$$(8.2) \quad \psi(w) = \sup\{(v, w) : v \in C\} \quad \forall w \in H.$$

Note in particular that  $D(\psi) = H$ . In the present rate-independent situation we are allowed to weaken the assumptions on  $\psi$  in (7.1) and ask for the upper bound only. Owing to positive homogeneity, we take with no loss of generality

$$(8.3) \quad \psi(w) \leq c_3 |w| \quad \forall w \in H,$$

which in particular says that  $C$  is contained in a ball of center 0 and radius  $c_3$ . As for  $\varphi$ , besides (A4) we require

$$(8.4) \quad t \mapsto \varphi(t, u) \text{ differentiable and } t \mapsto \partial_t \varphi(t, u) \text{ measurable } \forall u \in H.$$

Moreover, we ask for a nonnegative function  $\lambda \in L^1(0, T)$  and a constant  $c_8 > 0$  such that

$$(8.5) \quad |\partial_t \varphi(t, u)| \leq \lambda(t)(\varphi(t, u) + 1) \quad \forall u \in H,$$

$$(8.6) \quad |\partial_t \varphi(t, u) - \partial_t \varphi(t, w)| \leq c_8 |u - w| \quad \forall u, w \in H, \text{ for a.e. } t \in (0, T).$$

The latter and [64, Prop. 2.6] entail that the choice  $\pi(u(t)) := \partial_t \varphi(t, u(t))$  is admissible and fulfills the chain rule (3.2). Finally, we ask for the uniform convexity of  $\varphi$ , namely,

$$\begin{aligned} & \exists \kappa > 0 \text{ such that, } \forall u_0, u_1 \in H, \forall t \in [0, T], \forall \theta \in [0, 1], \\ & \varphi(t, (1 - \theta)u_0 + \theta u_1) \\ (8.7) \quad & \leq (1 - \theta)\varphi(t, u_0) + \theta\varphi(t, u_1) - \frac{\kappa}{2}\theta(1 - \theta)|u_0 - u_1|^2. \end{aligned}$$

In [64, subsect. 4.2] the authors discuss a nontrivial situation inspired by continuum mechanics where the latter conditions (8.2)–(8.7) are met. The crucial point now is that uniform convexity yields the Lipschitz time regularity of the solutions. In particular, under assumptions (8.2)–(8.7), all solutions  $(u, v)$  to the doubly nonlinear equation (3.1) fulfill [64, Thm. 3.2]

$$(8.8) \quad \|u'\|_{L^\infty(0, T; H)} \leq c_8/\kappa.$$

LEMMA 8.1 (convergence for rate-independent problems). *Under the assumptions of Lemma 5.4, let  $\psi_n$  fulfill (8.1) and (8.3) and  $\varphi_n$  fulfill (7.2) and (8.4)–(8.7) uniformly with respect to  $n$ . Moreover, let  $(u_n, v_n) \in W^{1,1}(0, T; H) \times L^\infty(0, T; H^*)$  be such that  $\mathcal{I}_n(u_n, v_n) = 0$ . Then there exists a (not relabeled) subsequence such that  $(u_n, v_n) \rightarrow (u, v)$  weakly star in  $W^{1,\infty}(0, T; H) \times L^\infty(0, T; H^*)$  and  $\mathcal{I}(u, v) = 0$ .*

*Proof.* First of all, one has that  $v_n \in C$  almost everywhere in  $(0, T)$  and hence are uniformly bounded in  $L^\infty(0, T; H^*)$ . Moreover, since  $(u_n, v_n)$  are solutions to (3.1), the Lipschitz continuity estimate (8.8) holds uniformly with respect to  $n$ . Additionally,  $u_n$  are uniformly bounded in  $L^1(0, T; V)$  due to (7.2). Hence, Lemma 5.4 yields the result.  $\square$

We shall mention that, differently from Lemmas 6.1 and 7.2, the latter result holds for sequences of solutions only since the estimate (8.8) is crucially used in order to obtain strong compactness in  $L^1(0, T; H)$  for  $u_n$ . In particular, we are not entitled to approximate a rate-independent situation by means of rate-dependent approximations. The reader is referred instead to Efendiev and Mielke [29] and Mielke, Rossi, and Savaré [66, 65] for some results in this direction.

The above convergence result could be alternatively obtained by applying the abstract analysis by Mielke, Roubíček, and Stefanelli [69, Thm. 3.1]. In the latter, besides the convergences of the functionals  $\psi_n \rightarrow \psi$  and  $\varphi_n \rightarrow \varphi$ , some extra closure condition, indeed fulfilled in the present situation, is crucially required [69, equation (2.11)]. Let us mention that [69] is devoted to a quite more general situation where the state space is just a Hausdorff topological space (in particular, no convexity is assumed on  $\varphi_n$ ).

We specialize further the results on rate-independent evolutions by explicitly discussing the fundamental case of (2.3), i.e., the so-called *play operator* in a Hilbert space  $H$ . The latter stands as the basic element for the construction of a relevant class of hysteresis operators, namely, the so-called Prandtl–Ishlinskii operators. The reader is referred to the classic monographs by Brokate and Sprekels [19], Krejčí [50], and Visintin [104] for a comprehensive collection of results on these operators. In particular, let us mention the convergence result [50, Thm. 3.12, p. 34] where the approximation of play operators under the Hausdorff convergence of the related characteristic convex sets  $C_n$  (see (8.2)) is discussed. Here we exploit instead the quite weaker situation of  $C_n \rightarrow C$  in the Mosco sense [7] (namely, the corresponding indicator functions Mosco-converge).

Comparing the case of the play operator with the above general result for rate-independent problems, we stress that, owing to the  $\liminf$  inequality (5.2), no strong compactness is here needed, and (7.2) can be omitted. We have the following convergence result.

LEMMA 8.2 (convergence for the play operator). *Let  $H$  be a separable Hilbert space and  $\psi_n \rightarrow \psi$  in the Mosco sense and fulfill (5.1), (8.1), and (8.3) uniformly with respect to  $n$ . Moreover, let  $f_n \rightarrow f$  strongly in  $C([0, T]; H^*)$ ,  $f_n$  be uniformly Lipschitz continuous, and  $u_n^0 \rightarrow u^0$  in  $H$ . Finally, let  $u_n \in W^{1,1}(0, T; H)$  be such that  $Q_n(u_n) = 0$ . Then  $u_n \rightarrow u$  weakly star in  $W^{1,\infty}(0, T; H)$  and  $Q(u) = 0$ .*

*Proof.* The Lipschitz regularity of  $f_n$  entails (8.5)–(8.6). Hence, the uniform control of (8.3) yields via (8.8) the uniform bound of  $u_n$  in  $W^{1,\infty}(0, T; H)$ . The assertion follows by extracting weakly star convergent subsequences and exploiting Lemma 5.3. In particular, the convergence of the whole sequence is ensured by the uniqueness of the solution of the limit problem [50, Thm. 3.1, p. 27 and Prop. 3.9, p. 33].  $\square$

The latter convergence result extends the former analysis by this author [102, Lemma 4.4] to the more natural setting  $W^{1,1}$ . One has, however, to mention that the former result was including the possibility of approximating the play operator with non-rate-independent evolutions (such as those stemming from penalizations or singular perturbations, for instance) while Lemma 8.2 is restricted to approximating plays only. On the other hand, the present convergence result is slightly more precise than the former since no strong convergence on the derivatives  $f'_n$  is required. The reader is referred to [103] for further results in this direction.

#### REFERENCES

- [1] S. AIZICOVICI AND Q. YAN, *Convergence theorems for abstract doubly nonlinear differential equations*, Panamer. Math. J., 7 (1997), pp. 1–17.
- [2] L. AMBROSIO, N. GIGLI, AND G. SAVARÉ, *Gradient flows in metric spaces and in the space of probability measures*, in Lectures Math. ETH Zürich, Birkhäuser-Verlag, Basel, 2005.
- [3] T. ARAI, *On the existence of the solution for  $\partial\varphi(u'(t)) + \partial\psi(u(t)) \ni f(t)$* , J. Fac. Sci. Univ. Tokyo Sect. IA Math., 26 (1979), pp. 75–96.
- [4] M. ASO, M. FRÉMOND, AND N. KENMOCHI, *Phase change problems with temperature dependent constraints for the volume fraction velocities*, Nonlinear Anal., 60 (2005), pp. 1003–1023.
- [5] E. ASPLUND, *Averaged norms*, Israel J. Math., 5 (1967), pp. 227–233.
- [6] E. ASPLUND, *Topics in the theory of convex functions*, in Theory and Applications of Monotone Operators (Proceedings of the NATO Advanced Study Institute, Venice, 1968), Edizioni “Oderisi”, Gubbio, 1969, pp. 1–33.
- [7] H. ATTOUCH, *Variational convergence for functions and operators*, Pitman, Boston, 1984.
- [8] J.-P. AUBIN, *Variational principles for differential equations of elliptic, parabolic and hyperbolic type*, in Mathematical Techniques of Optimization, Control and Decision, Birkhäuser, Boston, 1981, pp. 31–45.
- [9] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Grundlehren Math. Wiss. 264, Springer-Verlag, Berlin, 1984.
- [10] G. AUCHMUTY, *Saddle-points and existence-uniqueness for evolution equations*, Differential Integral Equations, 6 (1993), pp. 1161–1171.
- [11] E. J. BALDER, *A general approach to lower semicontinuity and lower closure in optimal control theory*, SIAM J. Control Optim., 22 (1984), pp. 570–598.
- [12] V. BARBU, *Existence theorems for a class of two point boundary problems*, J. Differential Equations, 17 (1975), pp. 236–257.
- [13] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, P. Noordhoff, Leyden, 1976.
- [14] V. BARBU AND K. KUNISCH, *Identification of nonlinear parabolic equations*, Control Theory Adv. Tech., 10 (1995), pp. 1959–1980.

- [15] M. A. BIOT, *Variational principles in irreversible thermodynamics with application to viscoelasticity*, Phys. Rev., 97 (1955), pp. 1463–1469.
- [16] H. BREZIS, *Opérateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert*, North-Holland Math. Stud. 5, North-Holland, Amsterdam, 1973.
- [17] H. BREZIS AND I. EKELAND, *Un principe variationnel associé à certaines équations paraboliques. Le cas dépendant du temps*, C. R. Acad. Sci. Paris Ser. A-B, 282 (1976), pp. Ai, A1197–A1198.
- [18] H. BREZIS AND I. EKELAND, *Un principe variationnel associé à certaines équations paraboliques. Le cas indépendant du temps*, C. R. Acad. Sci. Paris Ser. A-B, 282 (1976), pp. Aii, A971–A974.
- [19] M. BROKATE AND J. SPREKELS, *Hysteresis and Phase Transitions*, Appl. Math. Sci. 121, Springer-Verlag, Berlin, 1996.
- [20] C. CASTAING, P. RAYNAUD DE FITTE, AND M. VALADIER, *Young Measures on Topological Spaces*, Math. Appl. 571, Kluwer Academic, Dordrecht, 2004.
- [21] C. CASTAING AND M. VALADIER, *Convex analysis and measurable multifunctions*, in Lecture Notes in Math. 580, Springer-Verlag, Berlin, 1977.
- [22] P. COLLI, *On some doubly nonlinear evolution equations in Banach spaces*, Japan J. Indust. Appl. Math., 9 (1992), pp. 181–203.
- [23] P. COLLI AND A. VISINTIN, *On a class of doubly nonlinear evolution equations*, Comm. Partial Differential Equations, 15 (1990), pp. 737–756.
- [24] G. DAL MASO, *An Introduction to  $\Gamma$ -convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser Boston, Boston, 1993.
- [25] G. DAL MASO, A. DESIMONE, AND M. G. MORA, *Quasistatic evolution problems for linearly elastic-perfectly plastic materials*, Arch. Ration. Mech. Anal., 180 (2006), pp. 237–291.
- [26] G. DAL MASO, A. DESIMONE, M. G. MORA, AND M. MORINI, *A Vanishing Viscosity Approach to Quasistatic Evolution in Plasticity with Softening*, preprint; available online from <http://cvgmt.sns.it/cgi/get.cgi/papers/daldesmor06/> (2006).
- [27] G. DAL MASO, G. A. FRANCFORT, AND R. TOADER, *Quasistatic crack growth in nonlinear elasticity*, Arch. Ration. Mech. Anal., 176 (2005), pp. 165–225.
- [28] J. DIESTEL AND J. J. UHL, *Vector Measures*, Math. Surveys Monogr. 15, American Mathematical Society, Providence, RI, 1977.
- [29] M. A. EFENDIEV AND A. MIELKE, *On the rate-independent limit of systems with dry friction and small viscosity*, J. Convex Anal., 13 (2006), pp. 151–167.
- [30] P. GERMAIN, *Cours de Mécanique des Milieux Continus*, Tome I: Théorie générale, Masson et Cie, Éditeurs, Paris, 1973.
- [31] N. GHOUSSEUB, *A theory of anti-selfdual Lagrangians: Dynamical case*, C. R. Math. Acad. Sci. Paris, 340 (2005), pp. 325–330.
- [32] N. GHOUSSEUB, *A theory of anti-selfdual Lagrangians: Stationary case*, C. R. Math. Acad. Sci. Paris, 340 (2005), pp. 245–250.
- [33] N. GHOUSSEUB, *A variational principle for nonlinear transport equations*, Commun. Pure Appl. Anal., 4 (2005), pp. 735–742.
- [34] N. GHOUSSEUB, *Anti-selfdual Lagrangians: Variational resolutions of non self-adjoint equations and dissipative evolutions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 24 (2007), pp. 171–205.
- [35] N. GHOUSSEUB, *Antisymmetric Hamiltonians: Variational resolutions for Navier-Stokes and other nonlinear evolutions*, Comm. Pure Appl. Math., 60 (2007), pp. 619–653.
- [36] N. GHOUSSEUB, *Selfdual Partial Differential Systems and Their Variational Principles*, Universitext, Springer-Verlag, to appear.
- [37] N. GHOUSSEUB, *Maximal Monotone Operators Are Selfdual Vector Fields and Vice-Versa*, preprint; available online from <http://www.pims.math.ca/~nassif/> (2006).
- [38] N. GHOUSSEUB AND R. J. MCCANN, *A least action principle for steepest descent in a non-convex landscape*, in Partial Differential Equations and Inverse Problems, Contemp. Math. 362, American Mathematical Society, Providence, RI, 2004, pp. 177–187.
- [39] N. GHOUSSEUB AND L. TZOU, *A variational principle for gradient flows*, Math. Ann., 330 (2004), pp. 519–549.
- [40] N. GHOUSSEUB AND L. TZOU, *Iterations of anti-selfdual Lagrangians and applications to Hamiltonian systems and multiparameter gradient flows*, Calc. Var. Partial Differential Equations, 26 (2006), pp. 511–534.
- [41] N. GHOUSSEUB AND L. TZOU, *Anti-selfdual Lagrangians II: Unbounded non self-adjoint operators and evolution equations*, Ann. Mat. Pura Appl. (4), 187 (2008), pp. 323–352.
- [42] G. GILARDI AND U. STEFANELLI, *Time-discretization and global solution for a doubly nonlinear Volterra equation*, J. Differential Equations, 228 (2006), pp. 707–736.

- [43] E. DE GIORGI AND T. FRANZONI, *Su un tipo di convergenza variazionale*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8), 58 (1975), pp. 842–850.
- [44] M. E. GURTIN, *Variational principles in the linear theory of viscoelasticity*, Arch. Ration. Mech. Anal., 13 (1963), pp. 179–191.
- [45] M. E. GURTIN, *Variational principles for linear elastodynamics*, Arch. Ration. Mech. Anal., 16 (1964), pp. 34–50.
- [46] M. E. GURTIN, *Variational principles for linear initial value problems*, Quart. Appl. Math., 22 (1964), pp. 252–256.
- [47] W. HAN AND B. D. REDDY, *Plasticity, Mathematical Theory and Numerical Analysis*, Springer-Verlag, New York, 1999.
- [48] I. HLAVÁČEK, *Variational principles for parabolic equations*, Appl. Math., 14 (1969), pp. 278–297.
- [49] N. KENMOCHI AND U. STEFANELLI, *Existence for a class of nonlocal quasivariational evolution problems*, in Nonlinear Phenomena with Energy Dissipation: Mathematical Analysis, Modelling and Simulation, Chiba 2007, Mathematical Sciences and Applications, Gakuto, to appear.
- [50] P. KREJČÍ, *Hysteresis, Convexity and Dissipation in Hyperbolic Equations*, GAKUTO Internat. Ser. Math. Sci. Appl. 8, Gakkotosho, Tokyo, 1996.
- [51] B. LEMAIRE, *An asymptotical variational principle associated with the steepest descent method for a convex function*, J. Convex Anal., 3 (1996), pp. 63–70.
- [52] M. MABROUK, *Sur un problème d'évolution à données mesures; approche variationnelle*, C. R. Acad. Sci. Paris Ser. I Math., 318 (1994), pp. 47–52.
- [53] M. MABROUK, *A variational approach for a semi-linear parabolic equation with measure data*, Ann. Fac. Sci. Toulouse Math., 9 (2000), pp. 91–112.
- [54] M. MABROUK, *Un principe variationnel pour une équation non linéaire du second ordre en temps*, C. R. Acad. Sci. Paris Ser. I Math., 332 (2001), pp. 381–386.
- [55] M. MABROUK, *A variational principle for a nonlinear differential equation of second order*, Adv. in Appl. Math., 31 (2003), pp. 388–419.
- [56] A. MAINIK AND A. MIELKE, *Existence results for energetic models for rate-independent systems*, Calc. Var. Partial Differential Equations, 22 (2005), pp. 73–99.
- [57] G. MICHAILLE AND M. VALADIER, *Young measures generated by a class of integrands: A narrow epicontinuity and applications to homogenization*, J. Math. Pures Appl., 81 (2002), pp. 1277–1312.
- [58] A. MIELKE, *Finite Elastoplasticity Lie Groups and Geodesics on  $SL(d)$* , in Geometry, Mechanics, and Dynamics, Springer-Verlag, New York, 2002, pp. 61–90.
- [59] A. MIELKE, *Energetic formulation of multiplicative elasto-plasticity using dissipation distances*, Contin. Mech. Thermodyn., 15 (2003), pp. 351–382.
- [60] A. MIELKE, *Evolution of rate-independent inelasticity with microstructure using relaxation and Young measures*, in IUTAM Symposium on Computational Mechanics of Solid Materials at Large Strains (Stuttgart, 2001), Solid Mech. Appl. 108, Kluwer Academic, Dordrecht, 2003, pp. 33–44.
- [61] A. MIELKE, *Existence of minimizers in incremental elasto-plasticity with finite strains*, SIAM J. Math. Anal., 36 (2004), pp. 384–404.
- [62] A. MIELKE, *Evolution of rate-independent systems*, in Handbook of Differential Equations, Evolutionary Equations, Vol. 2, C. Dafermos and E. Feireisl, eds., Elsevier, New York, 2005, pp. 461–559.
- [63] A. MIELKE AND M. ORTIZ, *A class of minimum principles for characterizing the trajectories and the relaxation of dissipative systems*, ESAIM Control Optim. Calc. Var., to appear; preprint available online from <http://www.wias-berlin.de/main/publications/wias-publ/index.cgi.en>.
- [64] A. MIELKE AND R. ROSSI, *Existence and uniqueness results for general rate-independent hysteresis problems*, Math. Model Methods Appl. Sci., 17 (2007), pp. 81–123.
- [65] A. MIELKE, R. ROSSI, AND G. SAVARÉ, *Non parametric rate-independent flows*, manuscript, 2008.
- [66] A. MIELKE, R. ROSSI, AND G. SAVARÉ, *A metric approach to a class of doubly nonlinear evolution equations and application*, Ann. Sc. Norm. Super. Pisa Cl. Sci., to appear; preprint available online from <http://www.wias-berlin.de/main/publications/wias-publ/index.cgi.en>.
- [67] A. MIELKE AND T. ROUBÍČEK, *A rate-independent model for inelastic behavior of shape-memory alloys*, Multiscale Model. Simul., 1 (2003), pp. 571–597.
- [68] A. MIELKE AND T. ROUBÍČEK, *Rate-independent damage processes in nonlinear elasticity*, Math. Models Methods Appl. Sci., 16 (2006), pp. 177–209.

- [69] A. MIELKE, T. ROUBÍČEK, AND U. STEFANELLI,  $\Gamma$ -limits and relaxations for rate-independent evolutionary problems, *Calc. Var. Partial Differential Equations*, 31 (2008), pp. 125–164.
- [70] A. MIELKE AND U. STEFANELLI, *A Discrete Variational Principle for Rate-Independent Evolution*, WIAS preprint 1295, 2008; available online from <http://wias-berlin.de/main/publications/wias-publ/index.cgi.en>.
- [71] A. MIELKE AND F. THEIL, *A mathematical model for rate-independent phase transformations with hysteresis*, in *Proceedings of the Workshop on “Models of Continuum Mechanics in Analysis and Engineering,”* H.-D. Alber, R. Baican, and R. Farwig, eds., Shaker-Verlag, Aachen, 1999, pp. 117–129.
- [72] A. MIELKE, F. THEIL, AND V. I. LEVITAS, *A variational formulation of rate-independent phase transformations using an extremum principle*, *Arch. Ration. Mech. Anal.*, 162 (2002), pp. 137–177.
- [73] A. MIELKE AND A. M. TIMOFTE, *An energetic material model for time-dependent ferroelectric behavior: Existence and uniqueness*, *Math. Models Appl. Sci.*, 29 (2005), pp. 1393–1410.
- [74] J.-J. MOREAU, *Sur les lois de frottement, de viscosité et plasticité*, *C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre*, 271 (1970), pp. 608–611.
- [75] J.-J. MOREAU, *Sur l’évolution d’un système élasto-visco-plastique*, *C. R. Acad. Sci. Paris Sér. A-B*, 273 (1971), pp. A118–A121.
- [76] U. MOSCO, *Convergence of convex sets and of solutions of variational inequalities*, *Adv. Math.*, 3 (1969), pp. 510–585.
- [77] B. NAYROLES, *Deux théorèmes de minimum pour certains systèmes dissipatifs*, *C. R. Acad. Sci. Paris Sér. A-B*, 282 (1976), pp. Aiv, A1035–A1038.
- [78] B. NAYROLES, *Un théorème de minimum pour certains systèmes dissipatifs. Variante hilbertienne*, *Travaux Sémin. Anal. Convexe*, 6 (1976), p. 22.
- [79] C. ORTNER, *Two Variational Techniques for the Approximation of Curves of Maximal Slope*, Technical report NA05/10, Oxford University Computing Laboratory, 2005.
- [80] C. ORTNER, *Gradient flows as a selection procedure for equilibria of non-convex energies*, *SIAM J. Math. Anal.*, 38 (2006), pp. 1214–1234.
- [81] P. PEDREGAL,  $\Gamma$ -convergence through Young measures, *SIAM J. Math. Anal.*, 36 (2004), pp. 423–440.
- [82] P. PEDREGAL, *Young measures associated with homogenization*, *SIAM J. Math. Anal.*, 37 (2006), pp. 1454–1464.
- [83] J.-C. PERALBA, *Un problème d’évolution relatif à un opérateur sous-différentiel dépendant du temps*, *C. R. Acad. Sci. Paris Sér. A-B*, 275 (1972), pp. A93–A96.
- [84] J.-C. PERALBA, *Un problème d’évolution relatif à un opérateur sous-différentiel dépendant du temps*, in *Travaux du Séminaire d’Analyse Convexe*, Vol. II, Exp. 6, U.E.R. de Math., Univ. Sci. Tech. Languedoc, Montpellier, 1972, p. 17.
- [85] H. RIOS, *Étude de la question d’existence pour certains problèmes d’évolution par minimisation d’une fonctionnelle convexe*, *C. R. Acad. Sci. Paris Sér. A-B*, 283 (1976), pp. Ai, A83–A86.
- [86] H. RIOS, *La question d’existence de solutions pour certaines équations à opérateurs monotones vue comme problème de minimum ou comme problème de point-selle*, *Travaux Sémin. Anal. Convexe*, 6 (1976), p. 16.
- [87] H. RIOS, *Étude de certains problèmes paraboliques: Existence et approximation des solutions*, *Travaux Sémin. Anal. Convexe*, 8 (1978), p. 96.
- [88] H. RIOS, *Une étude d’existence sur certains problèmes paraboliques*, *Ann. Fac. Sci. Toulouse Math.* (5), 1 (1979), pp. 235–255.
- [89] G. ROSSI AND G. SAVARÉ, *Gradient flows of non convex functionals in Hilbert spaces and applications*, *ESAIM Control Optim. Calc. Var.*, 12 (2006), pp. 564–614.
- [90] T. ROUBÍČEK, *Direct method for parabolic problems*, *Adv. Math. Sci. Appl.*, 10 (2000), pp. 57–65.
- [91] T. ROUBÍČEK, *Nonlinear Partial Differential Equations with Applications*, *Internat. Ser. Numer. Math.* 153, Birkhäuser-Verlag, Basel, 2005.
- [92] A. SALVADORI, *On the  $M$ -convergence for integral functionals on  $L_X^p$* , *Atti Sem. Mat. Fis. Univ. Modena*, 33 (1984), pp. 137–154.
- [93] E. SANDIER AND S. SERFATY, *Gamma-convergence of gradient flows with applications to Ginzburg-Landau*, *Comm. Pure Appl. Math.*, 57 (2004), pp. 1627–1672.
- [94] F. SCHMID AND A. MIELKE, *Vortex pinning in super-conductivity as a rate-independent process*, *European J. Appl. Math.*, 16 (2005), pp. 799–808.
- [95] T. SENBA, *On some nonlinear evolution equation*, *Funkcial. Ekv.*, 29 (1986), pp. 243–257.
- [96] J. SIMON, *Compact sets in the space  $L^p(0, T; B)$* , *Ann. Mat. Pura Appl.*, 146 (1987), pp. 65–96.
- [97] U. STEFANELLI, *On a class of doubly nonlinear nonlocal evolution equations*, *Differential Integral Equations*, 15 (2002), pp. 897–922.

- [98] U. STEFANELLI, *On some nonlocal evolution equations in Banach spaces*, J. Evol. Equ., 4 (2004), pp. 1–26.
- [99] U. STEFANELLI, *Some quasivariational problems with memory*, Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat., 7 (2004), pp. 319–333.
- [100] U. STEFANELLI, *Nonlocal quasivariational evolution problems*, J. Differential Equations, 229 (2006), pp. 204–228.
- [101] U. STEFANELLI, *The Discrete Brezis-Ekeland Principle*, J. Convex. Anal., to appear.
- [102] U. STEFANELLI, *Some remarks on convergence and approximation for a class of hysteresis problems*, Istit. Lombardo Accad. Sci. Lett. Rend. A; preprint available online <http://www.imati.cnr.it/ulisse/pubbl.html>.
- [103] U. STEFANELLI, *A variational principle for hardening elasto-plasticity*, SIAM J. Math. Anal., to appear; preprint available from <http://www.imati.cnr.it/ulisse/pubbl.html>.
- [104] A. VISINTIN, *Differential Models of Hysteresis*, Appl. Math. Sci. 111, Springer-Verlag, Berlin, 1994.
- [105] A. VISINTIN, *A new approach to evolution*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 233–238.

## NONLINEAR OPTIMAL CONTROL VIA OCCUPATION MEASURES AND LMI-RELAXATIONS\*

JEAN B. LASSERRE<sup>†</sup>, DIDIER HENRION<sup>‡</sup>, CHRISTOPHE PRIEUR<sup>§</sup>, AND  
EMMANUEL TRÉLAT<sup>¶</sup>

**Abstract.** We consider the class of nonlinear optimal control problems (OCPs) with polynomial data, i.e., the differential equation, state and control constraints, and cost are all described by polynomials, and more generally for OCPs with smooth data. In addition, state constraints as well as state and/or action constraints are allowed. We provide a simple hierarchy of LMI- (linear matrix inequality)-relaxations whose optimal values form a nondecreasing sequence of lower bounds on the optimal value. Under some convexity assumptions, the sequence converges to the optimal value of the OCP. Preliminary results show that good approximations are obtained with few moments.

**Key words.** nonlinear control, optimal control, semidefinite programming, measures, moments

**AMS subject classifications.** 90C22, 93C10, 28A99

**DOI.** 10.1137/070685051

**1. Introduction.** Solving a general nonlinear optimal control problem (OCP) is a difficult challenge, despite the powerful theoretical tools available, e.g., the maximum principle and Hamilton–Jacobi–Bellman (HJB) optimality equation. The problem is even more difficult in the presence of state and/or control constraints. State constraints are particularly difficult to handle, and the interested reader is referred to Capuzzo-Dolcetta and Lions [8] and Soner [41] for a detailed account of HJB theory in the case of state constraints. There exist many numerical methods to compute the solution of a given OCP [37, 43]; for instance, *multiple shooting* techniques which solve two-point boundary value problems as described, e.g., in [42, 36], or *direct methods*, as, e.g., in [46, 13, 15], which use, among other things, descent or gradient-like algorithms. To deal with OCPs with state constraints, some adapted versions of the maximum principle have been developed (see [27, 34], and see [16] for a survey of this theory) but are very hard to implement in general.

On the other hand, the OCP can be written as an infinite-dimensional linear program (LP) over two spaces of measures. This is called the *weak* formulation of the OCP in Vinter [45] (stated in the more general context of differential inclusions). The two unknown measures are the state-action *occupation measure* (o.m.) *up to* the final time  $T$ , and the state o.m. *at* time  $T$ . The optimal value of the resulting LP always provides a lower bound on the optimal value of the OCP, and under some convexity

---

\*Received by the editors March 12, 2007; accepted for publication (in revised form) January 28, 2008; published electronically June 11, 2008. This work was partially supported by the French National Research Agency (ANR) under research project MOGA NT05-3-41612.

<http://www.siam.org/journals/sicon/47-4/68505.html>

<sup>†</sup>LAAS-CNRS and Institute of Mathematics, University of Toulouse, Toulouse, France (lasserre@laas.fr).

<sup>‡</sup>LAAS-CNRS, University of Toulouse, Toulouse, France, and Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic (henrion@laas.fr). The research of this author was partly supported by project 102/06/0652 of the Grant Agency of the Czech Republic and research program MSM6840770038 of the Ministry of Education of the Czech Republic.

<sup>§</sup>LAAS-CNRS, University of Toulouse, Toulouse, France (cprieur@laas.fr).

<sup>¶</sup>Université d'Orléans, Laboratoire MAPMO CNRS, UMR 6628 Fédération Denis Poisson, FR 2964 Bâtiment de Mathématiques BP 6759, 45067 Orléans cedex 2, France (emmanuel.trelat@univ-orleans.fr).



assumptions, both values coincide; see Vinter [45] and Hernandez-Hernandez et al. [20] as well. See Gaitsgory and Rossomakhine [14] for a more recent related work where, in addition, a numerical scheme is also defined for approximating an optimal control.

The dual of the original infinite-dimensional LP has an interpretation in terms of “subsolutions” of related HJB-like optimality conditions, as for the unconstrained case. The only difference with the unconstrained case is the underlying function space involved, which directly incorporates the state constraints. Namely, the functions are only defined on the state constraint set.

An interesting feature of this LP approach with o.m.’s is that state constraints, as well as state and/or action constraints, are all easy to handle; indeed they simply translate into constraints on the supports of the unknown o.m.’s. It thus provides an alternative to the use of maximum principles with state constraints. Although this LP approach is valid for any OCP, solving the corresponding (infinite-dimensional) LP is difficult in general; however, general LP approximation schemes based on grids have been proposed in, e.g., Hernández-Lerma and Lasserre [22].

This LP approach has also been used in the context of discrete-time Markov control processes, and is dual to Bellman’s optimality principle. For more details the interested reader is referred to the convex analytic approach described in Borkar [6], Hernández-Lerma and Lasserre [21, 23, 24], and the many references therein. For some continuous-time stochastic control problems (e.g., modeled by diffusions) and optimal stopping problems, the LP approach has also been used with success to prove existence of stationary optimal policies; see, for instance, Bhatt and Borkar [4], Cho and Stockbridge [9], Helmes and Stockbridge [18], Helmes, Röhl, and Stockbridge [17], Kurtz and Stockbridge [29], and also Fleming and Vermes [12]. In some of these works, the moment approach is also used to approximate the resulting infinite-dimensional LP.

**Contribution.** In this paper, we consider the particular class of nonlinear OCPs with state and/or control constraints, for which all data describing the problem (dynamics and state and control constraints) are *polynomials*. The approach also extends to the case of problems with *smooth* data and compact sets, because polynomials are dense in the space of functions considered; this point of view is detailed in section 4. In this restricted polynomial framework, the LP approach has interesting additional features that can be exploited for effective numerical computation. Indeed, for the class of OCPs considered, the following features make this LP approach attractive:

- Only the *moments* of the o.m.’s appear in the LP formulation, so that we already end up with countably many variables—a significant progress.
- Constraints on the support of the o.m.’s translate easily into either LP or SDP (semidefinite programming) *necessary* constraints on their moments. Even more, for (semialgebraic) compact supports, relatively recent powerful results from real algebraic geometry make these constraints also *sufficient*.
- When truncating to finitely many moments, the resulting LPs or SDPs are solvable and their optimal values form a monotone nondecreasing sequence (indexed by the number of moments considered) of lower bounds on the optimal value of the LP (and thus of the OCP). See [19, 30] for applications of this technique to nonconvex polynomial optimization.

Therefore, based on the above observations, we propose an approximation of the optimal value of the OCP via solving a hierarchy of SDPs (or linear matrix inequalities, LMIs) that provides a monotone nondecreasing sequence of lower bounds on the optimal value of the weak LP formulation of the OCP. In addition, under some compactness assumption of the state and control constraint sets, the sequence of lower

bounds is shown to converge to the optimal value of the LP, and thus to the optimal value of the OCP when the former and latter are equal.

As such, it could be seen as a complement to the above shooting or direct methods and, when the sequence of lower bounds converges to the optimal value, as a test of their efficiency. Finally this approach can also be used to provide a *certificate* of infeasibility. Indeed, if, in the hierarchy of LMI-relaxations of the minimum time OCP, one is infeasible, then the OCP itself is infeasible. It turns out that sometimes this certificate is provided at an early stage in the hierarchy, i.e., with very few moments. This is illustrated in two simple examples.

In a pioneering paper, Dawson [11] had suggested the use of *moments* in the LP approach with o.m.'s, but results on the  $\mathbf{K}$ -moment problem by Schmüdgen [40] and Putinar [39] were not available at that time. Later, Helmes and Stockbridge [18] and Helmes, Röhl, and Stockbridge [17] used LP moment conditions for computing some exit time moments in some diffusion model, whereas for the same models, Lasserre and Priéto-Rumeau [31] showed that SDP moment conditions are superior in terms of precision and number of moments to consider; in [32], they extended the moment approach for options pricing problems in some mathematical finance models. More recently, Lasserre, Prieur, and Henrion [33] used the o.m. approach for minimum time OCP without state constraint. Preliminary experimental results on Brockett's integrator example and on the double integrator show fast convergence with few moments.

## 2. Occupation measures and the LP approach.

**2.1. Definition of the OCP.** Let  $n$  and  $m$  be nonzero integers. Consider on  $\mathbb{R}^n$  the control system

$$(2.1) \quad \dot{x}(t) = f(t, x(t), u(t)),$$

where  $f : [0, +\infty) \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$  is smooth, and where the controls are bounded measurable functions, defined on intervals  $[0, T(\mathbf{u})]$  of  $\mathbb{R}^+$ , and taking their values in a compact subset  $\mathbf{U}$  of  $\mathbb{R}^m$ . Let  $x_0 \in \mathbb{R}^n$ , and let  $\mathbf{X}$  and  $\mathbf{K}$  be compact subsets of  $\mathbb{R}^n$ . For  $T > 0$ , a control  $u$  is said to be *admissible* on  $[0, T]$  whenever the solution  $x(\cdot)$  of (2.1), such that  $x(0) = x_0$ , is well defined on  $[0, T]$  and satisfies

$$(2.2) \quad (x(t), u(t)) \in \mathbf{X} \times \mathbf{U} \quad \text{a.e. on } [0, T]$$

and

$$(2.3) \quad x(T) \in \mathbf{K}.$$

Denote by  $\mathcal{U}_T$  the set of *admissible controls* on  $[0, T]$ .

For  $\mathbf{u} \in \mathcal{U}_T$ , the *cost* of the associated trajectory  $x(\cdot)$  is defined by

$$(2.4) \quad J(0, T, x_0, \mathbf{u}) = \int_0^T h(t, x(t), u(t)) dt + H(x(T)),$$

where  $h : [0, +\infty) \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$  and  $H : \mathbb{R}^n \rightarrow \mathbb{R}$  are smooth functions.

Consider the *optimal control problem* (OCP) of determining a trajectory solution of (2.1), starting from  $x(0) = x_0$ , satisfying the state and control constraints (2.2), the terminal constraint (2.3), and minimizing the cost (2.4). The final time  $T$  may or may not be fixed.

If the final time  $T$  is fixed, we set

$$(2.5) \quad J^*(0, T, x_0) := \inf_{\mathbf{u} \in \mathcal{U}_T} J(0, T, x_0, \mathbf{u}),$$

and if  $T$  is free, we set

$$(2.6) \quad J^*(0, x_0) := \inf_{T>0, \mathbf{u} \in \mathcal{U}_T} J(0, T, x_0, \mathbf{u}).$$

Note that a particular OCP is the minimal time problem from  $x_0$  to  $\mathbf{K}$ , by letting  $h \equiv 1$ ,  $H \equiv 0$ . In this particular case, the minimal time is the first *hitting time* of the set  $\mathbf{K}$ .

It is possible to associate a stochastic or deterministic OCP with an abstract infinite-dimensional LP problem  $\mathbf{P}$  together with its dual  $\mathbf{P}^*$  (see, for instance, Hernández-Lerma and Lasserre [21] for discrete-time Markov control problems, and Vinter [45], Hernandez et al. [20] for deterministic optimal control problems, as well as the many references therein). We next describe this LP approach in the present context.

**2.2. Notations and definitions.** For a topological space  $\mathcal{X}$ , let  $\mathcal{M}(\mathcal{X})$  be the Banach space of finite signed Borel measures on  $\mathcal{X}$ , equipped with the norm of total variation, and denote by  $\mathcal{M}(\mathcal{X})_+$  its positive cone, that is, the space of finite measures on  $\mathcal{X}$ . Let  $C(\mathcal{X})$  be the Banach space of bounded continuous functions on  $\mathcal{X}$ , equipped with the sup-norm. Notice that when  $\mathcal{X}$  is compact Hausdorff, then  $\mathcal{M}(\mathcal{X}) \simeq C(\mathcal{X})^*$ ; i.e.,  $\mathcal{M}(\mathcal{X})$  is the topological dual of  $C(\mathcal{X})$ .

Let  $\mathbb{R}[x] = [x_1, \dots, x_n]$  (resp.,  $\mathbb{R}[t, x, u] = \mathbb{R}[t, x_1, \dots, x_n, u_1, \dots, u_m]$ ) denote the set of polynomial functions of the variable  $x$  (resp., of the variables  $t, x, u$ ).

Let  $\Sigma := [0, T] \times \mathbf{X}$ ,  $\mathbf{S} := \Sigma \times \mathbf{U}$ , and let  $C_1(\Sigma)$  be the Banach space of functions  $\varphi \in C(\Sigma)$  that are continuously differentiable. For ease of exposition we use the same notation  $g$  (resp.,  $h$ ) for a polynomial  $g \in \mathbb{R}[t, x]$  (resp.,  $h \in \mathbb{R}[x]$ ) and its restriction to the compact set  $\Sigma$  (resp.,  $\mathbf{K}$ ).

Next, with  $u \in \mathbf{U}$ , let  $A : C_1(\Sigma) \rightarrow C(\mathbf{S})$  be the mapping

$$(2.7) \quad \varphi \mapsto A\varphi(t, x, u) := \frac{\partial \varphi}{\partial t}(t, x) + \langle f(t, x, u), \nabla_x \varphi(t, x) \rangle.$$

Notice that  $\partial \varphi / \partial t + \langle \nabla_x \varphi, f \rangle \in C(\mathbf{S})$  for every  $\varphi \in C_1(\Sigma)$ , because both  $\mathbf{X}$  and  $\mathbf{U}$  are compact, and  $f$  is understood as its restriction to  $\mathbf{S}$ .

Next, let  $\mathcal{L} : C_1(\Sigma) \rightarrow C(\mathbf{S}) \times C(\mathbf{K})$  be the linear mapping

$$(2.8) \quad \varphi \mapsto \mathcal{L}\varphi := (-A\varphi, \varphi_T),$$

where  $\varphi_T(x) := \varphi(T, x)$  for all  $x \in \mathbf{X}$ . Obviously,  $\mathcal{L}$  is continuous with respect to the strong topologies of  $C_1(\Sigma)$  and  $C(\mathbf{S}) \times C(\mathbf{K})$ .

Both  $(C(\mathbf{S}), \mathcal{M}(\mathbf{S}))$  and  $(C(\mathbf{K}), \mathcal{M}(\mathbf{K}))$  form a *dual pair* of vector spaces, with duality brackets

$$\langle h, \mu \rangle = \int h \, d\mu \quad \forall (h, \mu) \in C(\mathbf{S}) \times \mathcal{M}(\mathbf{S})$$

and

$$\langle g, \nu \rangle = \int g \, d\nu \quad \forall (g, \nu) \in C(\mathbf{K}) \times \mathcal{M}(\mathbf{K}).$$

Let  $\mathcal{L}^* : M(\mathbf{S}) \times M(\mathbf{K}) \rightarrow C_1(\Sigma)^*$  be the adjoint mapping of  $\mathcal{L}$ , defined by

$$(2.9) \quad \langle (\mu, \nu), \mathcal{L}\varphi \rangle = \langle \mathcal{L}^*(\mu, \nu), \varphi \rangle$$

for all  $((\mu, \nu), \varphi) \in M(\mathbf{S}) \times M(\mathbf{K}) \times C_1(\Sigma)$ .

*Remark 2.1.*

- (i) The mapping  $\mathcal{L}^*$  is continuous with respect to the weak topologies  $\sigma(\mathcal{M}(\mathbf{S}) \times \mathcal{M}(\mathbf{K}), C(\mathbf{S}) \times C(\mathbf{K}))$  and  $\sigma(C_1(\Sigma)^*, C_1(\Sigma))$ .
- (ii) Since the mapping  $\mathcal{L}$  is continuous in the strong sense, it is also continuous with respect to the weak topologies  $\sigma(C_1(\Sigma), C_1(\Sigma)^*)$  and  $\sigma(C(\mathbf{S}) \times C(\mathbf{K}), \mathcal{M}(\mathbf{S}) \times \mathcal{M}(\mathbf{K}))$ .
- (iii) In the case of a *free* terminal time  $T \leq T_0$ , it suffices to redefine  $\mathcal{L} : C_1(\Sigma) \rightarrow C(\mathbf{S}) \times C([0, T_0] \times \mathbf{K})$  by  $\mathcal{L}\varphi := (-A\varphi, \varphi)$ . The operator  $\mathcal{L}^* : M(\mathbf{S}) \times M([0, T_0] \times \mathbf{K}) \rightarrow C_1(\Sigma)^*$  is still defined by (2.9) for all  $((\mu, \nu), \varphi) \in M(\mathbf{S}) \times M([0, T_0] \times \mathbf{K}) \times C_1(\Sigma)$ .

For time-homogeneous free terminal time problems, one needs functions  $\varphi$  of  $x$  only, and so  $\Sigma = \mathbf{S} = \mathbf{X} \times \mathbf{U}$  and  $\mathcal{L} : C_1(\Sigma) \rightarrow C(\mathbf{S}) \times C(\mathbf{K})$ .

**2.3. Occupation measures and primal LP formulation.** Let  $T > 0$ , and let  $\mathbf{u} = \{u(t), 0 \leq t < T\}$  be a control such that the solution of (2.1), with  $x(0) = x_0$ , is well defined on  $[0, T]$ . Define the probability measure  $\nu^{\mathbf{u}}$  on  $\mathbb{R}^n$ , and the measure  $\mu^{\mathbf{u}}$  on  $[0, T] \times \mathbb{R}^n \times \mathbb{R}^m$ , by

$$(2.10) \quad \nu^{\mathbf{u}}(D) := \mathbf{I}_D[x(T)], \quad D \in \mathcal{B}_n,$$

$$(2.11) \quad \mu^{\mathbf{u}}(A \times B \times C) := \int_{[0, T] \cap A} \mathbf{I}_{B \times C}[(x(t), u(t))] dt$$

for all *rectangles*  $(A \times B \times C)$ , with  $(A, B, C) \in \mathcal{A} \times \mathcal{B}_n \times \mathcal{B}_m$ , and where  $\mathcal{B}_n$  (resp.,  $\mathcal{B}_m$ ) denotes the usual Borel  $\sigma$ -algebra associated with  $\mathbb{R}^n$  (resp.,  $\mathbb{R}^m$ ),  $\mathcal{A}$  the Borel  $\sigma$ -algebra of  $[0, T]$ , and  $\mathbf{I}_B(\bullet)$  the indicator function of the set  $B$ .

The measure  $\mu^{\mathbf{u}}$  is called the *occupation measure* (o.m.) of the state-action (deterministic) process  $(t, x(t), u(t))$  up to time  $T$ , whereas  $\nu^{\mathbf{u}}$  is the o.m. of the state  $x(T)$  at time  $T$ .

*Remark 2.2.* If the control  $\mathbf{u}$  is admissible on  $[0, T]$ , i.e., if the trajectory  $x(\cdot)$  satisfies the constraints (2.2) and (2.3), then  $\nu^{\mathbf{u}}$  is a probability measure supported on  $\mathbf{K}$  (i.e.,  $\nu^{\mathbf{u}} \in \mathcal{M}(\mathbf{K})_+$ ), and  $\mu^{\mathbf{u}}$  is supported on  $[0, T] \times \mathbf{X} \times \mathbf{U}$  (i.e.,  $\mu^{\mathbf{u}} \in \mathcal{M}(\mathbf{S})_+$ ). In particular,  $T = \mu^{\mathbf{u}}(\mathbf{S})$ .

Conversely, if the support of  $\mu^{\mathbf{u}}$  is contained in  $\mathbf{S} = [0, T] \times \mathbf{X} \times \mathbf{U}$  and if  $\mu^{\mathbf{u}}(\mathbf{S}) = T$ , then  $(x(t), u(t)) \in \mathbf{X} \times \mathbf{U}$  for almost every  $t \in [0, T]$ . Indeed, with (2.11),

$$\begin{aligned} T &= \int_0^T \mathbf{I}_{\mathbf{X} \times \mathbf{U}}[(x(s), u(s))] ds \\ &\Rightarrow \mathbf{I}_{\mathbf{X} \times \mathbf{U}}[(x(s), u(s))] = 1 \quad \text{a.e. in } [0, T], \end{aligned}$$

and hence  $(x(t), u(t)) \in \mathbf{X} \times \mathbf{U}$  for almost every  $t \in [0, T]$ . If moreover the support of  $\nu^{\mathbf{u}}$  is contained in  $\mathbf{K}$ , then  $x(T) \in \mathbf{K}$ . Therefore,  $\mathbf{u}$  is an admissible control on  $[0, T]$ .

Then observe that the optimization criterion (2.5) can now be written as

$$J(0, T, x_0, \mathbf{u}) = \int_{\mathbf{K}} H d\nu^{\mathbf{u}} + \int_{\mathbf{S}} h d\mu^{\mathbf{u}} = \langle (\mu^{\mathbf{u}}, \nu^{\mathbf{u}}), (h, H) \rangle,$$

and one infers from (2.1), (2.2), and (2.3) that

$$(2.12) \quad \int_{\mathbf{K}} g_T d\nu^{\mathbf{u}} - g(0, x_0) = \int_{\mathbf{S}} \left( \frac{\partial g}{\partial t} + \langle \nabla_x g, f \rangle \right) d\mu^{\mathbf{u}}$$

for every  $g \in C_1(\Sigma)$  (where  $g_T(x) \equiv g(T, x)$  for every  $x \in \mathbf{K}$ ), or equivalently, in view of (2.8) and (2.9),

$$\langle g, \mathcal{L}^*(\mu^u, \nu^u) \rangle = \langle g, \delta_{(0, x_0)} \rangle \quad \forall g \in C_1(\Sigma).$$

This in turn implies that

$$\mathcal{L}^*(\mu^u, \nu^u) = \delta_{(0, x_0)}.$$

Therefore, consider the infinite-dimensional linear program  $\mathbf{P}$ :

$$(2.13) \quad \mathbf{P} : \quad \inf_{(\mu, \nu) \in \Delta} \{ \langle (\mu, \nu), (h, H) \rangle \mid \mathcal{L}^*(\mu, \nu) = \delta_{(0, x_0)} \}$$

(where  $\Delta := \mathcal{M}(\mathbf{S})_+ \times \mathcal{M}(\mathbf{K})_+$ ). Denote by  $\inf \mathbf{P}$  its optimal value and by  $\min \mathbf{P}$  the infimum attained, in which case  $\mathbf{P}$  is said to be *solvable*. The problem  $\mathbf{P}$  is said to be *feasible* if there exists  $(\mu, \nu) \in \Delta$  such that  $\mathcal{L}^*(\mu, \nu) = \delta_{(0, x_0)}$ .

Note that  $\mathbf{P}$  is feasible whenever there exists an admissible control.

The linear program  $\mathbf{P}$  is a rephrasing of the OCP (2.1)–(2.5) in terms of the *o.m.*'s of its trajectories  $(t, x(t), u(t))$ . Its dual LP reads

$$(2.14) \quad \mathbf{P}^* : \quad \sup_{\varphi \in C_1(\Sigma)} \{ \langle \delta_{(0, x_0)}, \varphi \rangle \mid \mathcal{L}\varphi \leq (h, H) \},$$

where

$$\mathcal{L}\varphi \leq (h, H) \Leftrightarrow \begin{cases} A\varphi(t, x, u) + h(t, x, u) \geq 0 & \forall (t, x, u) \in \mathbf{S}, \\ \varphi(T, x) \leq H(x) & \forall x \in \mathbf{K}. \end{cases}$$

Denote by  $\sup \mathbf{P}^*$  its optimal value and by  $\max \mathbf{P}^*$  the supremum attained, i.e., if  $\mathbf{P}^*$  is solvable.

Discrete-time stochastic analogues of the LPs  $\mathbf{P}$  and  $\mathbf{P}^*$  are also described in Hernández-Lerma and Lasserre [21, 23] for discrete-time Markov control problems. Similarly, see Cho and Stockbridge [9], Kurtz and Stockbridge [29], and Helmes and Stockbridge [17] for some continuous-time stochastic problems.

**THEOREM 2.3.** *If  $\mathbf{P}$  is feasible, then the following hold:*

- (i)  $\mathbf{P}$  is solvable, i.e.,  $\inf \mathbf{P} = \min \mathbf{P} \leq J(0, T, x_0)$ .
- (ii) There is no duality gap, i.e.,  $\sup \mathbf{P}^* = \min \mathbf{P}$ .
- (iii) If, moreover, for every  $(t, x) \in \Sigma$  the set  $f(t, x, \mathbf{U}) \subset \mathbb{R}^n$  is convex, and the function

$$v \mapsto g_{t,x}(v) := \inf_{u \in \mathbf{U}} \{ h(t, x, u) : v = f(t, x, u) \}$$

is convex, then the OCP (2.1)–(2.5) has an optimal solution and

$$\sup \mathbf{P}^* = \inf \mathbf{P} = \min \mathbf{P} = J^*(0, T, x_0).$$

For a proof see section A.1. Theorem 2.3(iii) is due to Vinter [45].

**3. Semidefinite programming relaxations of  $\mathbf{P}$ .** The LP  $\mathbf{P}$  is infinite-dimensional, and thus not tractable as it stands. Therefore, we next present a relaxation scheme that provides a sequence of SDPs, or linear matrix inequality relaxations (in short, LMI-relaxations)  $\{\mathbf{Q}_r\}$ , each with *finitely many* constraints and variables.

Assume that  $\mathbf{X}$  and  $\mathbf{K}$  (resp.,  $\mathbf{U}$ ) are compact semialgebraic subsets of  $\mathbb{R}^n$  (resp., of  $\mathbb{R}^m$ ) of the form

$$(3.1) \quad \mathbf{X} := \{x \in \mathbb{R}^n \mid v_j(x) \geq 0, \quad j \in J\},$$

$$(3.2) \quad \mathbf{K} := \{x \in \mathbb{R}^n \mid \theta_j(x) \geq 0, \quad j \in J_T\},$$

$$(3.3) \quad \mathbf{U} := \{u \in \mathbb{R}^m \mid w_j(u) \geq 0, \quad j \in W\}$$

for some finite index sets  $J_T$ ,  $J$ , and  $W$ , where  $v_j$ ,  $\theta_j$ , and  $w_j$  are polynomial functions. Define

$$(3.4) \quad d(\mathbf{X}, \mathbf{K}, \mathbf{U}) := \max_{j \in J_1, l \in J, k \in W} (\deg \theta_j, \deg v_l, \deg w_k).$$

To highlight the main ideas, in this section we assume that  $f$ ,  $h$ , and  $H$  are polynomial functions, that is,  $h \in \mathbb{R}[t, x, u]$ ,  $H \in \mathbb{R}[x]$ , and  $f : [0, +\infty) \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  is polynomial; i.e., every component of  $f$  satisfies  $f_k \in \mathbb{R}[t, x, u]$  for  $k = 1, \dots, n$ .

**3.1. The underlying idea.** Observe the following important facts.

The restriction of  $\mathbb{R}[t, x]$  to  $\Sigma$  belongs to  $C_1(\Sigma)$ . Therefore,

$$\mathcal{L}^*(\mu, \nu) = \delta_{(0, x_0)} \quad \Leftrightarrow \quad \langle g, \mathcal{L}^*(\mu, \nu) \rangle = g(0, x_0) \quad \forall g \in \mathbb{R}[t, x],$$

because  $\Sigma$  being compact, polynomial functions are *dense* in  $C_1(\Sigma)$  for the sup-norm. Indeed, on a compact set, one may *simultaneously* approximate a function and its (continuous) partial derivatives by a polynomial and its derivatives uniformly (see [26, pp. 65–66]). Hence, the LP  $\mathbf{P}$  can be written

$$\mathbf{P} : \begin{cases} \inf_{(\mu, \nu) \in \Delta} \langle (\mu, \nu), (h, H) \rangle \\ \text{s.t.} \quad \langle g, \mathcal{L}^*(\mu, \nu) \rangle = g(0, x_0) \quad \forall g \in \mathbb{R}[t, x], \end{cases}$$

or equivalently, by linearity,

$$(3.5) \quad \mathbf{P} : \begin{cases} \inf_{(\mu, \nu) \in \Delta} \langle (\mu, \nu), (h, H) \rangle \\ \text{s.t.} \quad \langle \mathcal{L}g, (\mu, \nu) \rangle = g(0, x_0) \quad \forall g = (t^p x^\alpha); \quad (p, \alpha) \in \mathbb{N} \times \mathbb{N}^n. \end{cases}$$

The constraints of  $\mathbf{P}$ ,

$$(3.6) \quad \langle \mathcal{L}g, (\mu, \nu) \rangle = g(0, x_0) \quad \forall g = (t^p x^\alpha); \quad (p, \alpha) \in \mathbb{N} \times \mathbb{N}^n,$$

define countably many *linear* equality constraints linking the *moments* of  $\mu$  and  $\nu$ , because if  $g$  is polynomial, then so are  $\partial g / \partial t$  and  $\partial g / \partial x_k$ , for every  $k$ , and  $\langle \nabla_x g, f \rangle$ . And so,  $\mathcal{L}g$  is polynomial.

The functions  $h, H$  being also polynomial, the cost  $\langle (\mu, \nu), (h, H) \rangle$  of the OCP (2.1)–(2.5) is also a linear combination of the moments of  $\mu$  and  $\nu$ .

Therefore, the LP  $\mathbf{P}$  in (3.5) can be formulated as an LP with countably many variables (the moments of  $\mu$  and  $\nu$ ) and countably many linear equality constraints. However, it remains to express the fact that the variables should be moments of some measures  $\mu$  and  $\nu$ , with support contained in  $\mathbf{S}$  and  $\mathbf{K}$ , respectively.

At this stage, one will make some (weak) additional assumptions on the polynomials that define the compact semialgebraic sets  $\mathbf{X}, \mathbf{K}$ , and  $\mathbf{U}$ . Under such assumptions, one may then invoke recent results of real algebraic geometry on the representation of

polynomials positive on a compact set, and get necessary and sufficient conditions on the variables of  $\mathbf{P}$  to be indeed moments of two measures  $\mu$  and  $\nu$ , with appropriate support. We will use Putinar's Positivstellensatz [39] described in the next section, which yields SDP constraints on the variables.

One might also use other representation results like, e.g., Krivine [28] and Vasilescu [44] and obtain *linear* constraints on the variables (as opposed to SDP constraints). This is the approach taken by, e.g., Helmes, Röhl, and Stockbridge [17]. However, a comparison of the use of LP constraints versus SDP constraints on a related problem [31] has dictated our choice of the former.

Finally, if  $g$  in (3.6) runs only over all monomials of degree less than  $r$ , one then obtains a corresponding relaxation  $\mathbf{Q}_r$  of  $\mathbf{P}$ , which is now a finite-dimensional SDP that one may solve with public software packages. At last, one lets  $r \rightarrow \infty$ .

**3.2. Notations, definitions, and auxiliary results.** For a multi-index  $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$ , and for  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ , denote  $x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ . Consider the canonical basis  $\{x^\alpha\}_{\alpha \in \mathbb{N}^n}$  (resp.,  $\{t^p x^\alpha u^\beta\}_{p \in \mathbb{N}, \alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^m}$ ) of  $\mathbb{R}[x]$  (resp., of  $\mathbb{R}[t, x, u]$ ).

Given two sequences  $y = \{y_\alpha\}_{\alpha \in \mathbb{N}^n}$  and  $z = \{z_\gamma\}_{\gamma \in \mathbb{N} \times \mathbb{N}^n \times \mathbb{N}^m}$  of real numbers, define the linear functional  $L_y : \mathbb{R}[x] \rightarrow \mathbb{R}$  by

$$H \left( := \sum_{\alpha \in \mathbb{N}^n} H_\alpha x^\alpha \right) \mapsto L_y(H) := \sum_{\alpha \in \mathbb{N}^n} H_\alpha y_\alpha,$$

and similarly, define the linear functional  $L_z : \mathbb{R}[t, x, u] \rightarrow \mathbb{R}$  by

$$h \mapsto L_z(h) := \sum_{\gamma \in \mathbb{N} \times \mathbb{N}^n \times \mathbb{N}^m} h_\gamma z_\gamma = \sum_{p \in \mathbb{N}, \alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^m} h_{p\alpha\beta} z_{p\alpha\beta},$$

where  $h(t, x, u) = \sum_{p \in \mathbb{N}, \alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^m} h_{p\alpha\beta} t^p x^\alpha u^\beta$ .

Note that for a given measure  $\nu$  (resp.,  $\mu$ ) on  $\mathbb{R}$  (resp., on  $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m$ ), there holds, for every  $H \in \mathbb{R}[x]$  (resp., for every  $h \in \mathbb{R}[t, x, u]$ ),

$$\langle \nu, H \rangle = \int_{\mathbb{R}} H d\nu = \int_{\mathbb{R}} \sum_{\alpha \in \mathbb{N}^n} H_\alpha x^\alpha d\nu = \sum_{\alpha \in \mathbb{N}^n} H_\alpha y_\alpha = L_y(H),$$

where the real numbers  $y_\alpha = \int x^\alpha d\nu$  are the moments of the measure  $\nu$  (resp.,  $\langle \mu, h \rangle = L_z(h)$ , where  $z$  is the sequence of moments of the measure  $\mu$ ).

**DEFINITION 3.1.** For a given sequence  $z = \{z_\gamma\}_{\gamma \in \mathbb{N} \times \mathbb{N}^n \times \mathbb{N}^m}$  of real numbers, the moment matrix  $M_r(z)$  of order  $r$  associated with  $z$  has its rows and columns indexed in the canonical basis  $\{t^p x^\alpha u^\beta\}$  and is defined by

$$(3.7) \quad M_r(z)(\gamma, \beta) = z_{\gamma+\beta}, \quad \gamma, \beta \in \mathbb{N} \times \mathbb{N}^n \times \mathbb{N}^m, \quad |\gamma|, |\beta| \leq r,$$

where  $|\gamma| := \sum_j \gamma_j$ . The moment matrix  $M_r(y)$  of order  $r$  associated with a given sequence  $y = \{y_\alpha\}_{\alpha \in \mathbb{N}^n}$  has its rows and columns indexed in the canonical basis  $\{x^\alpha\}$  and is defined in a similar fashion.

Note that if  $z$  has a representing measure  $\mu$ , i.e., if  $z$  is the sequence of moments of the measure  $\mu$  on  $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m$ , then  $L_z(h) = \int h d\mu$  for every  $h \in \mathbb{R}[t, x, u]$ , and if  $\mathbf{h}$  denotes the vector of coefficients of  $h \in \mathbb{R}[t, x, u]$  of degree less than  $r$ , then

$$\langle \mathbf{h}, M_r(z) \mathbf{h} \rangle = L_z(h^2) = \int h^2 d\mu \geq 0.$$

This implies that  $M_r(z)$  is symmetric nonnegative (denoted  $M_r(z) \succeq 0$ ) for every  $r$ . The same holds for  $M_r(y)$ .

Conversely, not every sequence  $y$  that satisfies  $M_r(y) \succeq 0$  for every  $r$  has a representing measure. However, several sufficient conditions exist, in particular the following one due to Berg [3].

**PROPOSITION 3.2.** *If  $y = \{y_\alpha\}_{\alpha \in \mathbb{N}^n}$  satisfies  $|y_\alpha| \leq 1$  for every  $\alpha \in \mathbb{N}^n$ , and  $M_r(y) \succeq 0$  for every integer  $r$ , then  $y$  has a representing measure on  $\mathbb{R}^n$ , with support contained in the unit ball  $[-1, 1]^n$ .*

We next present another sufficient condition which is crucial in the proof of our main result.

**DEFINITION 3.3.** *For a given polynomial  $\theta \in \mathbb{R}[t, x, u]$ , written*

$$\theta(t, x, u) = \sum_{\delta=(p,\alpha,\beta)} \theta_\delta t^p x^\alpha u^\beta,$$

*define the localizing matrix  $M_r(\theta z)$  associated with  $z, \theta$ , and with rows and columns also indexed in the canonical basis of  $\mathbb{R}[t, x, u]$ , by*

$$(3.8) \quad M_r(\theta z)(\gamma, \beta) = \sum_{\delta} \theta_\delta z_{\delta+\gamma+\beta} \quad \gamma, \beta \in \mathbb{N} \times \mathbb{N}^n \times \mathbb{N}^m, \quad |\gamma|, |\beta| \leq r.$$

*The localizing matrix  $M_r(\theta y)$  associated with a given sequence  $y = \{y_\alpha\}_{\alpha \in \mathbb{N}^n}$  is defined similarly.*

Note that if  $z$  has a representing measure  $\mu$  on  $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m$  with support contained in the level set  $\{(t, x, u) : \theta(t, x, u) \geq 0\}$ , and if  $h \in \mathbb{R}[t, x, u]$  has degree less than  $r$ , then

$$\langle \mathbf{h}, M_r(\theta, z) \mathbf{h} \rangle = L_z(\theta h^2) = \int \theta h^2 d\mu \geq 0.$$

Hence,  $M_r(\theta z) \succeq 0$  for every  $r$ .

Let  $\Sigma^2 \subset \mathbb{R}[x]$  be the convex cone generated in  $\mathbb{R}[x]$  by all squares of polynomials, and let  $\Omega \subset \mathbb{R}^n$  be the compact basic semialgebraic set defined by

$$(3.9) \quad \Omega := \{x \in \mathbb{R}^n \mid g_j(x) \geq 0, \quad j = 1, \dots, m\}$$

for some family  $\{g_j\}_{j=1}^m \subset \mathbb{R}[x]$ .

**DEFINITION 3.4.** *The set  $\Omega \subset \mathbb{R}^n$  defined by (3.9) satisfies Putinar's condition if there exists  $u \in \mathbb{R}[x]$  such that  $u = u_0 + \sum_{j=1}^m u_j g_j$  for some family  $\{u_j\}_{j=0}^m \subset \Sigma^2$ , and the level set  $\{x \in \mathbb{R}^n \mid u(x) \geq 0\}$  is compact.*

Putinar's condition is satisfied if, e.g., the level set  $\{x : g_k(x) \geq 0\}$  is compact for some  $k$ , or if all the  $g_j$ 's are linear, in which case  $\Omega$  is a polytope. In addition, if one knows some  $M$  such that  $\|x\| \leq M$  whenever  $x \in \Omega$ , then it suffices to add the redundant quadratic constraint  $M^2 - \|x\|^2 \geq 0$  in the definition (3.9) of  $\Omega$ , and Putinar's condition is satisfied (take  $u := M^2 - \|x\|^2$ ).

**THEOREM 3.5** (Putinar's Positivstellensatz [39]). *Assume that the set  $\Omega$  defined by (3.9) satisfies Putinar's condition.*

(a) *If  $f \in \mathbb{R}[x]$  and  $f > 0$  on  $\Omega$ , then*

$$(3.10) \quad f = f_0 + \sum_{j=1}^m f_j g_j$$

*for some family  $\{f_j\}_{j=0}^m \subset \Sigma^2$ .*



(b) Let  $y = \{y_\alpha\}_{\alpha \in \mathbb{N}^n}$  be a sequence of real numbers. If

$$(3.11) \quad M_r(y) \succeq 0; \quad M_r(g_j y) \succeq 0, \quad j = 1, \dots, m; \quad \forall r = 0, 1, \dots,$$

then  $y$  has a representing measure with support contained in  $\Omega$ .

**3.3. LMI-relaxations of  $\mathbf{P}$ .** Consider the LP  $\mathbf{P}$  defined by (3.5).

Since the semialgebraic sets  $\mathbf{X}, \mathbf{K}$ , and  $\mathbf{U}$  defined, respectively, by (3.1), (3.2), and (3.3) are compact, with no loss of generality we assume (up to a scaling of the variables  $x, u$ , and  $t$ ) that  $T = 1$ ,  $\mathbf{X}, \mathbf{K} \subseteq [-1, 1]^n$ , and  $\mathbf{U} \subseteq [-1, 1]^m$ .

Next, given a sequence  $z = \{z_\gamma\}$  indexed in the basis of  $\mathbb{R}[t, x, u]$ , denote  $z(t)$ ,  $z(x)$ , and  $z(u)$  its marginals with respect to the variables  $t$ ,  $x$ , and  $u$ , respectively. These sequences are indexed in the canonical basis of  $\mathbb{R}[t]$ ,  $\mathbb{R}[x]$ , and  $\mathbb{R}[u]$ , respectively. For instance, writing  $\gamma = (k, \alpha, \beta) \in \mathbb{N} \times \mathbb{N}^n \times \mathbb{N}^m$ ,

$$\{z(t)\} = \{z_{k,0,0}\}_{k \in \mathbb{N}}; \quad \{z(x)\} = \{z_{0,\alpha,0}\}_{\alpha \in \mathbb{N}^n}; \quad \{z(u)\} = \{z_{0,0,\beta}\}_{\beta \in \mathbb{N}^m}.$$

Let  $r_0$  be an integer such that  $2r_0 \geq \max(\deg f, \deg h, \deg H, 2d(\mathbf{X}, \mathbf{K}, \mathbf{U}))$ , where  $d(\mathbf{X}, \mathbf{K}, \mathbf{U})$  is defined by (3.4). For every  $r \geq r_0$ , consider the LMI-relaxation

$$(3.12) \quad \mathbf{Q}_r : \begin{cases} \inf_{y,z} L_z(h) + L_y(H), \\ M_r(y), M_r(z) \succeq 0, \\ M_{r-\deg \theta_j}(\theta_j y) \succeq 0, \quad j \in J_1, \\ M_{r-\deg v_j}(v_j z(x)) \succeq 0, \quad j \in J, \\ M_{r-\deg w_k}(w_k z(u)) \succeq 0, \quad k \in W, \\ M_{r-1}(t(1-t)z(t)) \succeq 0, \\ L_y(g_1) - L_z(\partial g / \partial t + \langle \nabla_x g, f \rangle) = g(0, x_0) \quad \forall g = (t^p x^\alpha), \\ \text{with } p + |\alpha| - 1 + \deg f \leq 2r, \end{cases}$$

whose optimal value is denoted by  $\inf \mathbf{Q}_r$ .

**OCP with free terminal time.** For the OCP (2.6), i.e., with *free* terminal time  $T \leq T_0$ , we need to adapt the notation because  $T$  is now a variable. As already mentioned in Remark 2.1(iii), the measure  $\nu$  in the infinite-dimensional LP  $\mathbf{P}$  defined in (2.13) is now supported in  $[0, T_0] \times \mathbf{K}$  (and  $[0, 1] \times \mathbf{K}$  after rescaling) instead of  $\mathbf{K}$  previously. Hence, the sequence  $y$  associated with  $\nu$  is now indexed in the basis  $\{t^p x^\alpha\}$  of  $\mathbb{R}[t, x]$  instead of  $\{x^\alpha\}$  previously. Therefore, given  $y = \{y_{k\alpha}\}$  indexed in that basis, let  $y(t)$  and  $y(x)$  be the subsequences of  $y$  defined by

$$y(t) := \{y_{k0}\}_k, \quad k \in \mathbb{N}; \quad y(x) = \{y_{0\alpha}\}, \quad \alpha \in \mathbb{N}^n.$$

Then again (after rescaling), the LMI-relaxation  $\mathbf{Q}_r$  now reads

$$(3.13) \quad \mathbf{Q}_r : \begin{cases} \inf_{y,z} L_z(h) + L_y(H), \\ M_r(y), M_r(z) \succeq 0, \\ M_{r-r(\theta_j)}(\theta_j y) \succeq 0, \quad j \in J_1, \\ M_{r-r(v_j)}(v_j z(x)) \succeq 0, \quad j \in J, \\ M_{r-r(w_k)}(w_k z(u)) \succeq 0, \quad k \in W, \\ M_{r-1}(t(1-t)y(t)) \succeq 0, \\ M_{r-1}(t(1-t)z(t)) \succeq 0, \\ L_y(g) - L_z(\partial g / \partial t + \langle \nabla_x g, f \rangle) = g(0, x_0) \quad \forall g = (t^p x^\alpha), \\ \text{with } p + |\alpha| - 1 + \deg f \leq 2r. \end{cases}$$

The particular case of the minimal time problem is obtained with  $h \equiv 1$ ,  $H \equiv 0$ .

For *time-homogeneous* problems, i.e., when  $h$  and  $f$  do not depend on  $t$ , one may take  $\mu$  (resp.,  $\nu$ ) supported on  $\mathbf{X} \times \mathbf{U}$  (resp.,  $\mathbf{K}$ ), which simplifies the associated LMI-relaxation (3.13).

The following is the main result.

**THEOREM 3.6.** *Let  $\mathbf{X}, \mathbf{K} \subset [-1, 1]^n$  and  $\mathbf{U} \subset [-1, 1]^m$  be compact basic semi-algebraic sets defined, respectively, by (3.1), (3.2), and (3.3). Assume that  $\mathbf{X}, \mathbf{K}$ , and  $\mathbf{U}$  satisfy Putinar's condition (see Definition 3.4), and let  $\mathbf{Q}_r$  be the LMI-relaxation defined in (3.12). Then*

- (i)  $\inf \mathbf{Q}_r \uparrow \min \mathbf{P}$  as  $r \rightarrow \infty$ ;
- (ii) *if, moreover, for every  $(t, x) \in \Sigma$ , the set  $f(t, x, \mathbf{U}) \subset \mathbb{R}^n$  is convex and the function*

$$v \mapsto g_{t,x}(v) := \inf_{u \in \mathbf{U}} \{ h(t, x, u) \mid v = f(t, x, u) \}$$

*is convex, then  $\inf \mathbf{Q}_r \uparrow \min \mathbf{P} = J^*(0, T, x_0)$ , as  $r \rightarrow \infty$ .*

The proof of this result is postponed to section A.2 of the appendix.

**Remark 3.7.** It is known that the HJB optimality equation

$$(3.14) \quad \inf_{u \in \mathbf{U}} \{ A v(s, x, u) + h(s, x, u) \} = 0, \quad (s, x) \in \Sigma,$$

with boundary condition  $v_T(x) (= v(T, x)) = H(x)$  for all  $x \in \mathbf{K}$ , may have no continuously differentiable solution  $v : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$ , because of possible shocks of characteristics. On the other hand, a function  $\varphi \in C_1(\Sigma)$  is said to be a smooth *subsolution* of the HJB equation (3.14) if it is a feasible solution of  $\mathbf{P}^*$ , i.e.,

$$(3.15) \quad A \varphi(t, x, u) + h(t, x, u) \geq 0, \quad (t, x, u) \in \mathbf{S}; \quad \varphi(T, x) \leq H(x), \quad x \in \mathbf{K};$$

see, e.g., Vinter [45]. The dual of the LMI-relaxation  $\mathbf{Q}_r$ , which is also an SDP denoted  $\mathbf{Q}_r^*$ , is a reinforcement of  $\mathbf{P}^*$  in the sense that we consider only polynomial subsolutions, and, in addition, the positivity condition in (3.15) is replaced by the Putinar representation (3.10). Next, as  $\mathbf{Q}_r^*$  is an approximation of  $\mathbf{P}^*$ , a topic of further research, beyond the scope of the present paper, is how to use  $\mathbf{Q}_r^*$  to provide some information on an optimal solution of the OCP (2.1)–(2.5).

**3.4. Certificates of uncontrollability.** For minimum time OCPs, i.e., with free terminal time  $T$  and instantaneous cost  $h \equiv 1$  and  $H \equiv 0$ , the LMI-relaxations  $\mathbf{Q}_r$  defined in (3.13) may provide *certificates* of uncontrollability.

Indeed, if for a given initial state  $x_0 \in \mathbf{X}$  some LMI-relaxation  $\mathbf{Q}_r$  in the hierarchy has *no* feasible solution, then the initial state  $x_0$  *cannot* be steered to the origin in finite time. In other words,  $\inf \mathbf{Q}_r = +\infty$  provides a certificate of uncontrollability of the initial state  $x_0$ . It turns out that sometimes such certificates can be provided at cheap cost, i.e., with LMI-relaxations of low order  $r$ . This is illustrated on the Zermelo problem in section 5.3.

Moreover, one may also consider controllability in given finite time  $T$ ; that is, consider the LMI-relaxations as defined in (3.12) with  $T$  fixed and  $H \equiv 0$ ,  $h \equiv 1$ . Again, if for a given initial state  $x_0 \in \mathbf{X}$  the LMI-relaxation  $\mathbf{Q}_r$  has no feasible solution, then the initial state  $x_0$  cannot be steered to the origin in less than  $T$  units of time. And so  $\inf \mathbf{Q}_r = +\infty$  also provides a certificate of uncontrollability of the initial state  $x_0$ .

**4. Generalization to smooth OCPs.** In the previous section, we assumed, to highlight the main ideas, that  $f$ ,  $h$ , and  $H$  were polynomials. In this section, we generalize Theorem 3.6 and simply assume that  $f$ ,  $h$ , and  $H$  are smooth. Consider the LP  $\mathbf{P}$  defined in the previous section:

$$\mathbf{P} : \begin{cases} \inf_{(\mu, \nu) \in \Delta} \{ \langle (\mu, \nu), (h, H) \rangle \\ \text{s.t.} \quad \langle g, \mathcal{L}^*(\mu, \nu) \rangle = g(0, x_0) \quad \forall g \in \mathbb{R}[t, x]. \end{cases}$$

Since the sets  $\mathbf{X}$ ,  $\mathbf{K}$ , and  $\mathbf{U}$ , defined previously, are compact, it follows from [10] (see also [26, pp. 65–66]) that  $f$  (resp.,  $h$ ,  $H$ ) is the limit in  $C_1(\mathbf{S})$  (resp.,  $C_1(\mathbf{S})$ ,  $C_1(\mathbf{K})$ ) of a sequence of polynomials  $f_p$  (resp.,  $h_p$ ,  $H_p$ ) of degree  $p$  as  $p \rightarrow +\infty$ .

Hence, for every integer  $p$ , consider the LP  $\mathbf{P}_p$ ,

$$\mathbf{P}_p : \begin{cases} \inf_{(\mu, \nu) \in \Delta} \{ \langle (\mu, \nu), (h_p, H_p) \rangle \\ \text{s.t.} \quad \langle g, \mathcal{L}_p^*(\mu, \nu) \rangle = g(0, x_0) \quad \forall g \in \mathbb{R}[t, x], \end{cases}$$

to be the smooth analogue of  $\mathbf{P}$ , where the linear mapping  $\mathcal{L}_p : C_1(\Sigma) \rightarrow C(\mathbf{S}) \times C(\mathbf{K})$  is defined by

$$\mathcal{L}_p \varphi := (-A_p \varphi, \varphi_T),$$

and where  $A_p : C_1(\Sigma) \rightarrow C(\mathbf{S})$  is defined by

$$A_p \varphi(t, x, u) := \frac{\partial \varphi}{\partial t}(t, x) + \langle f_p(t, x, u), \nabla_x \varphi(t, x) \rangle.$$

For every integer  $r \geq \max(p/2, d(\mathbf{X}, \mathbf{K}, \mathbf{U}))$ , let  $\mathbf{Q}_{r,p}$  denote the LMI-relaxation (3.12) associated with the LP  $\mathbf{P}_p$ .

Recall that from Theorem 3.6 if  $\mathbf{K}$ ,  $\mathbf{X}$ , and  $\mathbf{U}$  satisfy Putinar's condition, then  $\inf \mathbf{Q}_{r,p} \uparrow \min \mathbf{P}_p$  as  $r \rightarrow +\infty$ .

The next result, generalizing Theorem 3.6, shows that it is possible to let  $p$  tend to  $+\infty$ . For convenience, set

$$v_{r,p} = \inf \mathbf{Q}_{r,p}, \quad v_p = \min \mathbf{P}_p, \quad v = \min \mathbf{P}.$$

**THEOREM 4.1.** *Let  $\mathbf{X}, \mathbf{K} \subset [-1, 1]^n$  and  $\mathbf{U} \subset [-1, 1]^m$  be compact semialgebraic sets defined, respectively, by (3.1), (3.2), and (3.3). Assume that  $\mathbf{X}, \mathbf{K}$ , and  $\mathbf{U}$  satisfy Putinar's condition (see Definition 3.4). Then*

- (i)  $v = \lim_{p \rightarrow +\infty} \lim_{\substack{r \rightarrow +\infty \\ 2r > p}} v_{r,p} = \lim_{p \rightarrow +\infty} \sup_{r > p/2} v_{r,p} \leq J^*(0, T, x_0);$
- (ii) *moreover, if for every  $(t, x) \in \Sigma$ , the set  $f(t, x, \mathbf{U}) \subset \mathbb{R}^n$  is convex and the function*

$$v \mapsto g_{t,x}(v) := \inf_{u \in \mathbf{U}} \{ h(t, x, u) \mid v = f(t, x, u) \}$$

*is convex, then  $v = J^*(0, T, x_0)$ .*

The proof of this result is in section A.3 of the appendix.

From the numerical point of view, depending on the functions  $f$ ,  $h$ ,  $H$ , the degree of the polynomials of the approximate OCP  $\mathbf{P}_p$  may be required to be large, and hence the hierarchy of LMI-relaxations  $(\mathbf{Q}_r)$  in (3.12) might not be efficiently implementable, at least in view of the performances of public SDP solvers presently available.

*Remark 4.2.* The previous construction extends to smooth OCPs on Riemannian manifolds, as follows. Let  $M$  and  $N$  be smooth Riemannian manifolds. Consider on  $M$  the control system (2.1), where  $f : [0, +\infty) \times M \times N \rightarrow TM$  is smooth, and where the controls are bounded measurable functions, defined on intervals  $[0, T(\mathbf{u})]$  of  $\mathbb{R}^+$ , and taking their values in a compact subset  $\mathbf{U}$  of  $N$ . Let  $x_0 \in M$ , and let  $\mathbf{X}$  and  $\mathbf{K}$  be compact subsets of  $M$ . Admissible controls are defined as previously. For an admissible control  $\mathbf{u}$  on  $[0, T]$ , the cost of the associated trajectory  $x(\cdot)$  is defined by (2.4), where  $h : [0, +\infty) \times M \times N \rightarrow \mathbb{R}$  and  $H : M \rightarrow \mathbb{R}$  are smooth functions.

According to the Nash embedding theorem [35], there exists an integer  $n$  (resp.,  $m$ ) such that  $M$  (resp.,  $N$ ) is smoothly isometrically embedded in  $\mathbb{R}^n$  (resp.,  $\mathbb{R}^m$ ). In this context, all previous results apply.

This remark is important for the applicability of the method described in this article. Indeed, many practical control problems (particularly in mechanics) are expressed on manifolds, and since the OCP investigated here is global, such problems cannot be expressed in general as control systems in  $\mathbb{R}^n$  (in a global chart).

**5. Illustrative examples.** We consider here the minimal time OCP, that is, we aim to approximate the *minimal time* to steer a given initial condition to the origin. We have tested the above methodology on two test OCPs, the double integrator and the Brockett integrator, for which the associated optimal value  $T^*$  can be calculated exactly. The numerical examples in this section were processed with our MATLAB package *GloptiPoly 3*.<sup>1</sup>

**5.1. The double integrator.** Consider the double integrator system in  $\mathbb{R}^2$ ,

$$(5.1) \quad \begin{aligned} \dot{x}_1(t) &= x_2(t), \\ \dot{x}_2(t) &= u(t), \end{aligned}$$

where  $x = (x_1, x_2)$  is the state and the control  $\mathbf{u} = u(t) \in \mathcal{U}$  satisfies the constraint  $|u(t)| \leq 1$  for all  $t \geq 0$ . In addition, the state is constrained to satisfy  $x_2(t) \geq -1$  for all  $t$ . In this time-homogeneous case, and with the notation of section 2, we have  $\mathbf{X} = \{x \in \mathbb{R}^2 : x_2 \geq -1\}$ ,  $\mathbf{K} = \{(0, 0)\}$ , and  $\mathbf{U} = [-1, 1]$ .

Observe that  $\mathbf{X}$  is not compact and so the convergence result of Theorem 3.6 may not hold. In fact, we may impose the additional constraint  $\|x(t)\|_\infty \leq M$  for some large  $M$  (and modify  $\mathbf{X}$  accordingly), because for initial states  $x_0$  with  $\|x_0\|_\infty$  relatively small with respect to  $M$ , the optimal trajectory remains in  $\mathbf{X}$ . However, in the numerical experiments, we have not enforced an additional constraint. We have maintained the original constraint  $x_2 \geq -1$  in the localizing constraint  $M_{r-r(v_j)}(v_j z(x)) \geq 0$ , with  $x \mapsto v_j(x) = x_2 + 1$ .

**5.1.1. Exact computation.** For this very simple system, one is able to compute exactly the optimal minimum time [43]. Denoting  $T(x)$  the minimal time to reach the origin from  $x = (x_1, x_2)$ , we have the following:

If  $x_1 \geq 1 - x_2^2/2$  and  $x_2 \geq -1$ , then  $T(x) = x_2^2/2 + x_1 + x_2 + 1$ . If  $-x_2^2/2 \text{ sign } x_2 \leq x_1 \leq 1 - x_2^2/2$  and  $x_2 \geq -1$ , then  $T(x) = 2\sqrt{x_2^2/2 + x_1 + x_2}$ . If  $x_1 < -x_2^2/2 \text{ sign } x_2$  and  $x_2 \geq -1$ , then  $T(x) = 2\sqrt{x_2^2/2 - x_1 - x_2}$ . Note that the expressions in section III.A.1 of [33] are incorrect.

**5.1.2. Numerical approximation.** Table 1 displays the values of the initial state  $x_0 \in \mathbf{X}$ , and denoting  $\inf \mathbf{Q}_r(x_0)$  the optimal value of the LMI-relaxation (3.13) for the minimum time OCP (5.1) with initial state  $x_0$ , Tables 2, 3, and 4 display

<sup>1</sup> *GloptiPoly 3* can be downloaded from [www.laas.fr/~henrion/software](http://www.laas.fr/~henrion/software).

TABLE 1  
*Double integrator: data initial state  $x_0 = (x_{01}, x_{02})$ .*

$x_{01}$	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
$x_{02}$	−1.0	−0.8	−0.6	−0.4	−0.2	0.0	0.2	0.4	0.6	0.8	1.0

TABLE 2  
*Double integrator: ratio  $\inf \mathbf{Q}_2/T(x_0)$ .*

Second LMI-relaxation: $r = 2$										
0.4598	0.3964	0.3512	0.9817	0.9674	0.9634	0.9628	0.9608	0.9600	0.9596	0.9595
0.4534	0.3679	0.9653	0.9347	0.9355	0.9383	0.9385	0.9386	0.9413	0.9432	0.9445
0.4390	0.9722	0.8653	0.8457	0.8518	0.8639	0.8720	0.8848	0.8862	0.8983	0.9015
0.4301	0.7698	0.7057	0.7050	0.7286	0.7542	0.7752	0.7964	0.8085	0.8187	0.8351
0.4212	0.4919	0.5039	0.5422	0.5833	0.6230	0.6613	0.6870	0.7121	0.7329	0.7513
0.0000	0.2238	0.3165	0.3877	0.4476	0.5005	0.5460	0.5839	0.6158	0.6434	0.6671
0.4501	0.3536	0.3950	0.4403	0.4846	0.5276	0.5663	0.5934	0.6204	0.6474	0.6667
0.4878	0.4493	0.4699	0.5025	0.5342	0.5691	0.5981	0.6219	0.6446	0.6647	0.6824
0.5248	0.5142	0.5355	0.5591	0.5840	0.6124	0.6312	0.6544	0.6689	0.6869	0.7005
0.5629	0.5673	0.5842	0.6044	0.6296	0.6465	0.6674	0.6829	0.6906	0.7083	0.7204
0.6001	0.6099	0.6245	0.6470	0.6617	0.6792	0.6972	0.7028	0.7153	0.7279	0.7369

TABLE 3  
*Double integrator: ratio  $\inf \mathbf{Q}_3/T(x_0)$ .*

Third LMI-relaxation: $r = 3$										
0.5418	0.4400	0.3630	0.9989	0.9987	0.9987	0.9985	0.9984	0.9983	0.9984	0.9984
0.5115	0.3864	0.9803	0.9648	0.9687	0.9726	0.9756	0.9778	0.9798	0.9815	0.9829
0.4848	0.9793	0.8877	0.8745	0.8847	0.8997	0.9110	0.9208	0.9277	0.9339	0.9385
0.4613	0.7899	0.7321	0.7401	0.7636	0.7915	0.8147	0.8339	0.8484	0.8605	0.8714
0.4359	0.5179	0.5361	0.5772	0.6205	0.6629	0.7013	0.7302	0.7540	0.7711	0.7891
0.0000	0.2458	0.3496	0.4273	0.4979	0.5571	0.5978	0.6409	0.6719	0.6925	0.7254
0.4556	0.3740	0.4242	0.4789	0.5253	0.5767	0.6166	0.6437	0.6807	0.6972	0.7342
0.4978	0.4709	0.5020	0.5393	0.5784	0.6179	0.6477	0.6776	0.6976	0.7192	0.7347
0.5396	0.5395	0.5638	0.5955	0.6314	0.6600	0.6856	0.7089	0.7269	0.7438	0.7555
0.5823	0.5946	0.6190	0.6453	0.6703	0.7019	0.7177	0.7382	0.7539	0.7662	0.7767
0.6255	0.6434	0.6656	0.6903	0.7193	0.7326	0.7543	0.7649	0.7776	0.7917	0.8012

the numerical values of the ratios  $\inf \mathbf{Q}_r(x_0)/T(x_0)$  for  $r = 2, 3$ , and  $5$ , respectively. Columns and rows in Tables 2, 3, and 4 are, respectively, indexed by values of  $x_{01}$  and  $x_{02}$  indicated in Table 1. A ratio near 1 indicates a good approximation in relative value.

Figures 1, 2, and 3 display the level sets of the ratios  $\inf \mathbf{Q}_r/T(x_0)$  for  $r = 2, 3$ , and  $5$ , respectively. The closer the color is to white, the closer the ratio  $\inf \mathbf{Q}_r/T(x_0)$  is to 1.

Finally, for this double integrator example we have plotted in Figure 4 the level sets of the function  $\Lambda_5(x) - T(x)$ , where  $T(x)$  is the known optimal minimum time to steer the initial state  $x$  to 0; the problem being time-homogeneous, one may take  $\Lambda_r \in \mathbb{R}[x]$  instead of  $\mathbb{R}[t, x]$ . For instance, one may verify that when the initial state is in the region where the approximation is good, then the whole optimal trajectory also lies in that region.

TABLE 4  
Double integrator: ratio  $\inf \mathbf{Q}_5/T(x_0)$ .

Fifth LMI-relaxation: $r = 5$										
0.7550	0.5539	0.3928	0.9995	0.9995	0.9995	0.9994	0.9992	0.9988	0.9985	0.9984
0.6799	0.4354	0.9828	0.9794	0.9896	0.9923	0.9917	0.9919	0.9923	0.9923	0.9938
0.6062	0.9805	0.9314	0.9462	0.9706	0.9836	0.9853	0.9847	0.9848	0.9862	0.9871
0.5368	0.8422	0.8550	0.8911	0.9394	0.9599	0.9684	0.9741	0.9727	0.9793	0.9776
0.4713	0.6417	0.7334	0.8186	0.8622	0.9154	0.9448	0.9501	0.9505	0.9665	0.9637
0.0000	0.4184	0.5962	0.7144	0.8053	0.8825	0.9044	0.9210	0.9320	0.9544	0.9534
0.4742	0.5068	0.6224	0.7239	0.7988	0.8726	0.8860	0.9097	0.9263	0.9475	0.9580
0.5410	0.6003	0.6988	0.7585	0.8236	0.8860	0.9128	0.9257	0.9358	0.9452	0.9528
0.6106	0.6826	0.7416	0.8125	0.8725	0.9241	0.9305	0.9375	0.9507	0.9567	0.9604
0.6864	0.7330	0.7979	0.8588	0.9183	0.9473	0.9481	0.9480	0.9559	0.9634	0.9733
0.7462	0.8032	0.8564	0.9138	0.9394	0.9610	0.9678	0.9678	0.9696	0.9755	0.9764

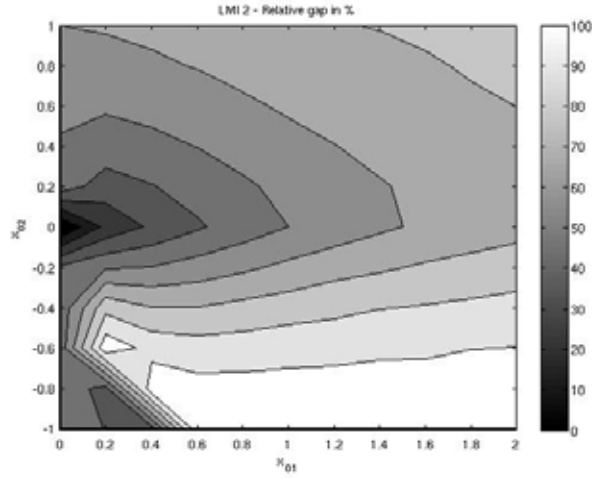


FIG. 1. Double integrator: level sets  $\inf \mathbf{Q}_2/T(x_0)$ .

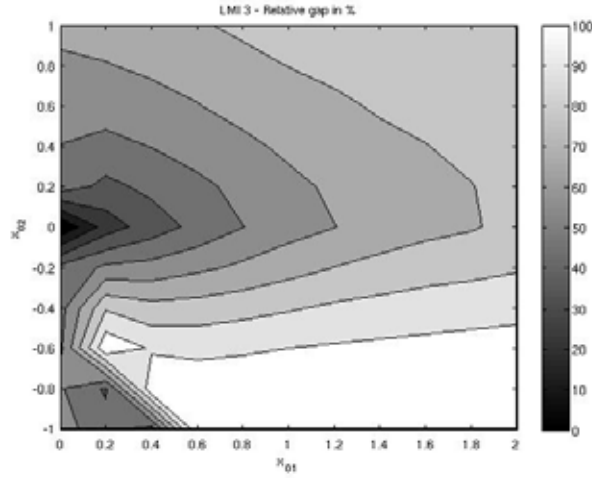


FIG. 2. Double integrator: level sets  $\inf \mathbf{Q}_3/T(x_0)$ .

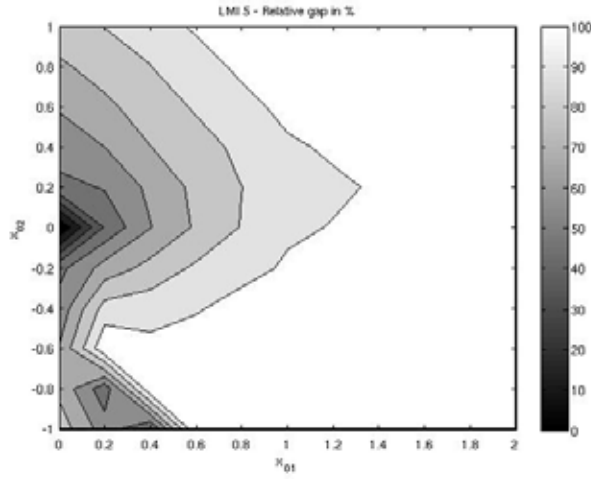


FIG. 3. Double integrator: level sets  $\inf \mathbf{Q}_5/T(x_0)$ .

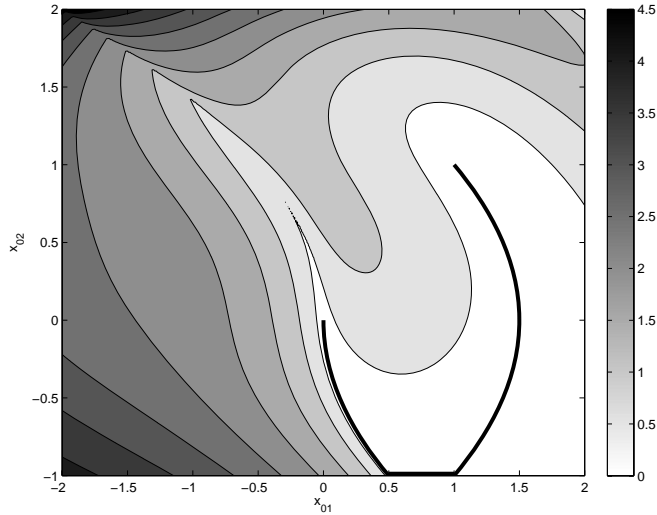


FIG. 4. Double integrator: level sets  $\Lambda_5(x) - T(x)$  and optimal trajectory starting from  $x_1(0) = x_2(0) = 1$ .

**5.2. The Brockett integrator.** Consider the so-called *Brockett system* in  $\mathbb{R}^3$  (see [7]),

$$\begin{aligned} \dot{x}_1(t) &= u_1(t), \\ \dot{x}_2(t) &= u_2(t), \\ \dot{x}_3(t) &= u_1(t)x_2(t) - u_2(t)x_1(t), \end{aligned} \tag{5.2}$$

where  $x = (x_1, x_2, x_3)$ , and the control  $\mathbf{u} = (u_1(t), u_2(t)) \in \mathcal{U}$  satisfies the constraint

$$u_1(t)^2 + u_2(t)^2 \leq 1 \quad \forall t \geq 0. \tag{5.3}$$

In this case, we have  $\mathbf{X} = \mathbb{R}^3$ ,  $\mathbf{K} = \{(0, 0, 0)\}$ , and  $\mathbf{U}$  is the closed unit ball of  $\mathbb{R}^2$ , centered at the origin.

Note that set  $\mathbf{X}$  is not compact and so the convergence result of Theorem 3.6 may not hold; see the discussion at the beginning of Example 5.1. Nevertheless, in the numerical examples, we have not enforced additional state constraints.

**5.2.1. Exact computation.** Let  $T(x)$  be the minimum time needed to steer an initial condition  $x \in \mathbb{R}^3$  to the origin. We recall the following result of [2] (given in fact for an equivalent (reachability) OCP of steering the origin to a given point  $x$ ).

**PROPOSITION 5.1.** *Consider the minimum time OCP for the system (5.2) with control constraint (5.3). The minimum time  $T(x)$  needed to steer the origin to a point  $x = (x_1, x_2, x_3) \in \mathbb{R}^3$  is given by*

$$(5.4) \quad T(x_1, x_2, x_3) = \frac{\theta \sqrt{x_1^2 + x_2^2 + 2|x_3|}}{\sqrt{\theta + \sin^2 \theta - \sin \theta \cos \theta}},$$

where  $\theta = \theta(x_1, x_2, x_3)$  is the unique solution in  $[0, \pi)$  of

$$(5.5) \quad \frac{\theta - \sin \theta \cos \theta}{\sin^2 \theta} (x_1^2 + x_2^2) = 2|x_3|.$$

Moreover, the function  $T$  is continuous on  $\mathbb{R}^3$  and is analytic outside the line  $x_1 = x_2 = 0$ .

*Remark 5.2.* Along the line  $x_1 = x_2 = 0$ , one has

$$T(0, 0, x_3) = \sqrt{2\pi|x_3|}.$$

The singular set of the function  $T$ , i.e., the set where  $T$  is not  $C^1$ , is the line  $x_1 = x_2 = 0$  in  $\mathbb{R}^3$ . More precisely, the gradients  $\partial T / \partial x_i$ ,  $i = 1, 2$ , are discontinuous at every point  $(0, 0, x_3)$ ,  $x_3 \neq 0$ . For the interested reader, the level sets  $\{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid T(x_1, x_2, x_3) = r\}$ , with  $r > 0$ , are displayed, e.g., in Prieur and Trélat [38].

**5.2.2. Numerical approximation.** Recall that the convergence result of Theorem 3.6 is guaranteed for  $\mathbf{X}$  compact only. However, in the present case  $\mathbf{X} = \mathbb{R}^3$  is not compact. One possibility is to take for  $\mathbf{X}$  a large ball of  $\mathbb{R}^3$  centered at the origin because for initial states  $x_0$  with norm  $\|x_0\|$  relatively small, the optimal trajectory remains in  $\mathbf{X}$ . However, in the numerical experiments presented below, we have chosen to maintain  $\mathbf{X} = \mathbb{R}^3$ , that is, the LMI-relaxation  $\mathbf{Q}_r$  does not include any localizing constraint  $M_{r-r(v_j)}(v_j z(x)) \succeq 0$  on the moment sequence  $z(x)$ .

Recall that  $\inf \mathbf{Q}_r \uparrow \min \mathbf{P}$  as  $r$  increases, i.e., the more moments we consider, the closer to the exact value we get.

In Table 5 we have displayed the optimal values  $\inf \mathbf{Q}_r$  for 16 different values of the initial state  $x(0) = x_0$ , in fact, all 16 combinations of  $x_{01} = 0$ ,  $x_{02} = 0, 1, 2, 3$ , and  $x_{03} = 0, 1, 2, 3$ . So, the entry (2, 3) of Table 5 for the second LMI-relaxation is  $\inf \mathbf{Q}_2$  for the initial condition  $x_0 = (0, 1, 2)$ . At some (few) places in the table, the \* indicates that the SDP solver encountered some numerical problems, which explains why one finds a lower bound  $\inf \mathbf{Q}_{r-1}$  slightly higher than  $\inf \mathbf{Q}_r$ , when practically equal to the exact value  $T^*$ .

Notice that the upper triangular part (i.e., when both first coordinates  $x_{02}, x_{03}$  of the initial condition are away from zero) displays very good approximations with low order moments. In addition, the further the coordinates from zero, the better.

For another set of initial conditions  $x_{01} = 1$  and  $x_{0i} = \{1, 2, 3\}$  Table 6 displays the results obtained at the LMI-relaxation  $\mathbf{Q}_4$ .



TABLE 5  
*Brockett integrator: LMI-relaxations:  $\inf Q_r$ .*

First LMI-relaxation: $r = 1$			
0.0000	0.9999	1.9999	2.9999
0.0140	1.0017	2.0010	3.0006
0.0243	1.0032	2.0017	3.0024
0.0295	1.0101	2.0034	3.0040
Second LMI-relaxation: $r = 2$			
0.0000	0.9998	1.9997*	2.9994*
0.2012	1.1199	2.0762	3.0453
0.3738	1.2003	2.1631	3.1304
0.4946	1.3467	2.2417	3.1943
Third LMI-relaxation: $r = 3$			
0.0000	0.9995	1.9987*	2.9984*
0.7665	1.3350	2.1563	3.0530
1.0826	1.7574	2.4172	3.2036
1.3804	2.0398	2.6797	3.4077
Fourth LMI-relaxation: $r = 4$			
0.0000	0.9992	1.9977	2.9952
1.2554	1.5925	2.1699	3.0478
1.9962	2.1871	2.5601	3.1977
2.7006	2.7390	2.9894	3.4254
Optimal time $T^*$			
0.0000	1.0000	2.0000	3.0000
2.5066	1.7841	2.1735	3.0547
3.5449	2.6831	2.5819	3.2088
4.3416	3.4328	3.0708	3.4392

TABLE 6  
*Brockett integrator:  $\inf Q_4$  with  $x_{01} = 1$ .*

Fourth LMI-relaxation: $r = 4$		
1.7979	2.3614	3.2004
2.3691	2.6780	3.3341
2.8875	3.0654	3.5337
Optimal time $T^*$		
1.8257	2.3636	3.2091
2.5231	2.6856	3.3426
3.1895	3.1008	3.5456

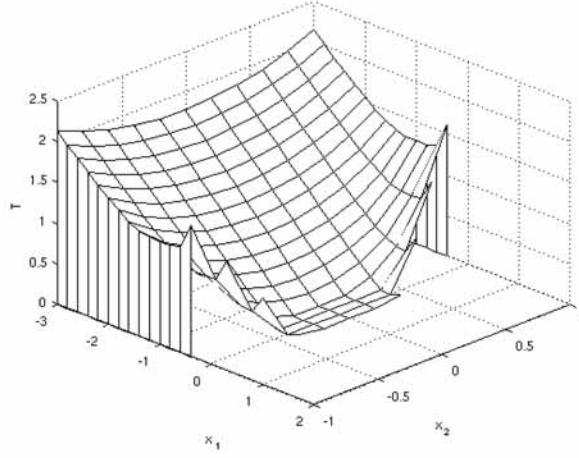
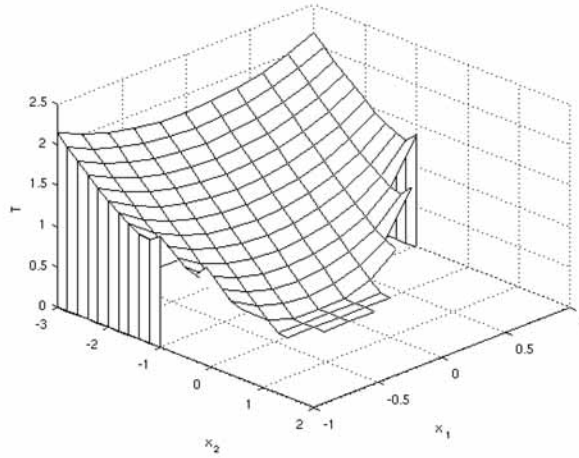
The regularity property of the minimal time function seems to be an important topic of further investigation.

**5.3. Certificate of uncontrollability in finite time.** Consider the so-called Zermelo problem in  $\mathbb{R}^2$  studied by Bokanowski et al. [5]:

$$(5.6) \quad \begin{aligned} \dot{x}_1(t) &= 1 - a x_2(t) + v \cos \theta, \\ \dot{x}_2(t) &= v \sin \theta, \end{aligned}$$

with  $a = 0.1$ . The state  $x$  is constrained to remain in the set  $\mathbf{X} := [-6, 2] \times [-2, 2] \subset \mathbb{R}^2$ , and we also have the control constraints  $0 \leq v \leq 0.44$ , as well as  $\theta \in [0, 2\pi]$ . The target  $\mathbf{K}$  to reach from an initial state  $x_0$  is the ball  $B(0, \rho)$ , with  $\rho := 0.44$ .

With the change of variable  $u_1 = v \cos \theta$ ,  $u_2 = v \sin \theta$ , and  $\mathbf{U} := \{u : u_1^2 + u_2^2 \leq \rho^2\}$ , this problem is formulated as a minimum time OCP with associated hierarchy of LMI-

FIG. 5. *Zermelo problem: uncontrollable states with  $\mathbf{Q}_1$ .*FIG. 6. *Zermelo problem: uncontrollable states with  $\mathbf{Q}_2$ .*

relaxations  $(\mathbf{Q}_r)$  defined in (3.13). Therefore, if an LMI-relaxation  $\mathbf{Q}_r$  at some stage  $r$  of the hierarchy is infeasible, then the OCP itself is infeasible; i.e., the initial state  $x_0$  cannot be steered to the target  $\mathbf{K}$  while the trajectory remains in  $\mathbf{X}$ .

Figures 5 and 6 display the uncontrollable initial states  $x_0 \in \mathbf{X}$  found with the infeasibility of the LMI-relaxations  $\mathbf{Q}_1$  and  $\mathbf{Q}_2$ , respectively. With  $\mathbf{Q}_1$ , i.e., by using only second moments, we already have a very good approximation of the controllable set displayed in [5], and  $\mathbf{Q}_2$  brings only a small additional set of uncontrollable states.

## Appendix.

**A.1. Proof of Theorem 2.3.** We first prove item (i). Consider the LP  $\mathbf{P}$  defined in (2.13), and assumed to be feasible. From the constraint  $\mathcal{L}^*(\mu, \nu) = \delta_{(0, x_0)}$ ,

one has

$$\int_{\mathbf{K}} g(T, x) d\nu - \int_{\mathbf{S}} \left( \frac{\partial g}{\partial t}(t, x) + \left\langle \frac{\partial g}{\partial x}(t, x), f(t, x, u) \right\rangle \right) d\mu = g(0, x_0) \quad \forall g \in C_1(\Sigma).$$

In particular, taking  $g(t, x) = 1$  and  $g(t, x) = T - t$ , it follows that  $\mu(\mathbf{S}) = T$  and  $\nu(\mathbf{K}) = 1$ . Hence, for every  $(\mu, \nu)$  satisfying  $\mathcal{L}^*(\mu, \nu) = \delta_{(0, x_0)}$ , the pair  $(\frac{1}{T}\mu, \nu)$  belongs to the unit ball  $B_1$  of  $(\mathcal{M}(\mathbf{S}) \times \mathcal{M}(\mathbf{K}))$ . From the Banach–Alaoglu theorem,  $B_1$  is compact for the weak\* topology, and even sequentially compact because  $B_1$  is metrizable (see, e.g., Hernández-Lerma and Lasserre [25, Lemma 1.3.2]). Since  $\mathcal{L}^*$  is continuous (see Remark 2.1), it follows that the set of  $(\mu, \nu)$  satisfying  $\mathcal{L}^*(\mu, \nu) = \delta_{(0, x_0)}$  is a closed subset of  $B_1 \cap (\mathcal{M}(\mathbf{S})_+ \times \mathcal{M}(\mathbf{K})_+)$ , and thus is compact. Moreover, since the LP  $\mathbf{P}$  is feasible, this set is nonempty. Finally, since the linear functional to be minimized is continuous,  $\mathbf{P}$  is solvable.

We next prove item (ii). Consider the set

$$D := \{(\mathcal{L}^*(\mu, \nu), \langle (h, H), (\mu, \nu) \rangle) \mid (\mu, \nu) \in \mathcal{M}(\mathbf{S})_+ \times \mathcal{M}(\mathbf{K})_+\}.$$

To prove that  $D$  is closed, consider a sequence  $\{(\mu_n, \nu_n)\}_{n \in \mathbb{N}}$  of  $\mathcal{M}(\mathbf{S})_+ \times \mathcal{M}(\mathbf{K})_+$  such that

$$(A.1) \quad (\mathcal{L}^*(\mu_n, \nu_n), \langle (h, H), (\mu_n, \nu_n) \rangle) \rightarrow (a, b)$$

for some  $(a, b) \in C_1(\Sigma)^* \times \mathbb{R}$ . This means that  $\mathcal{L}^*(\mu_n, \nu_n) \rightarrow a$  and  $\langle (h, H), (\mu_n, \nu_n) \rangle \rightarrow b$ . In particular, taking  $\varphi_0 := T - t$  and  $\varphi_1 = 1$ , there must hold

$$\mu_n(\mathbf{S}) = \langle \varphi_0, \mathcal{L}^*(\mu_n, \nu_n) \rangle \rightarrow \langle \varphi_0, a \rangle, \quad \nu_n(\mathbf{K}) = \langle \varphi_1, \mathcal{L}^*(\mu_n, \nu_n) \rangle \rightarrow \langle \varphi_1, a \rangle.$$

Hence, there exist  $n_0 \in \mathbb{N}$  and a ball  $B_r$  of  $\mathcal{M}(\mathbf{S}) \times \mathcal{M}(\mathbf{K})$ , such that  $(\mu_n, \nu_n) \in B_r$  for every  $n \geq n_0$ . Since  $B_r$  is compact, along a subsequence,  $(\mu_n, \nu_n)$  converges weakly to some  $(\mu, \nu) \in \mathcal{M}(\mathbf{S})_+ \times \mathcal{M}(\mathbf{K})_+$ . This fact, combined with (A.1) and the continuity of  $\mathcal{L}^*$ , yields  $a = \mathcal{L}^*(\mu, \nu)$  and  $b = \langle (h, H), (\mu, \nu) \rangle$ . Therefore, the set  $D$  is closed.

From Anderson and Nash [1, Theorems 3.10 and 3.22], it follows that there is no duality gap between  $\mathbf{P}$  and  $\mathbf{P}^*$ , and hence, with (i),  $\sup \mathbf{P}^* = \min \mathbf{P}$ .

Item (iii) follows from Vinter [45, Theorems 2.1 and 2.3] applied to the mappings

$$F(t, x) := f(t, x, U), \quad l(t, x, v) := \inf_{u \in U} \{ h(t, x, u) \mid v = f(t, x, u) \}$$

for  $(t, x) \in \mathbb{R} \times \mathbb{R}^n$ .  $\square$

**A.2. Proof of Theorem 3.6.** First of all, observe that  $\mathbf{Q}_r$  has a feasible solution. Indeed, it suffices to consider the sequences  $y$  and  $z$  consisting of the moments (up to order  $2r$ ) of the o.m.'s  $\nu^{\mathbf{u}}$  and  $\mu^{\mathbf{u}}$  associated with an admissible control  $\mathbf{u} \in \mathcal{U}$  of the OCP (2.1)–(2.5).

Next, observe that for every couple  $(y, z)$  satisfying all constraints of  $\mathbf{Q}_r$ , there must hold  $y_0 = 1$  and  $z_0 = 1$ . Indeed, it suffices to choose  $g(t, x) = 1$  and  $g(t, x) = 1 - t$  (or  $t$ ) in the constraint

$$L_y(g_1) - L_z(\partial g / \partial t + \langle \nabla_x g, f \rangle) = g(0, x_0).$$

We next prove that for  $r$  sufficiently large, one has  $|z(x)_\alpha| \leq 1$ ,  $|z(u)_\beta| \leq 1$ ,  $|z(t)_k| \leq 1$  for every  $k$ , and  $|y_\alpha| \leq 1$ . We provide details only of the proof for  $z(x)$ , the arguments being similar for  $z(u)$ ,  $z(t)$ , and  $y$ .

Let  $\Sigma^2 \subset \mathbb{R}[x]$  be the space of sum of squares polynomials, and let  $Q \subset \mathbb{R}[x]$  be the *quadratic modulus* generated by the polynomials  $v_j \in \mathbb{R}[x]$  that define  $\mathbf{X}$ , i.e.,

$$Q := \left\{ \sigma \in \mathbb{R}[x] \mid \sigma = \sigma_0 + \sum_{j \in J} \sigma_j v_j \quad \text{with } \sigma_j \in \Sigma^2 \quad \forall j \in \{0\} \cup J \right\}.$$

In addition, let  $Q(t) \subset Q$  be the set of elements  $\sigma$  of  $Q$  which have a representation  $\sigma_0 + \sum_{j \in J} \sigma_j v_j$  for some sum of squares family  $\{\sigma_j\} \subset \Sigma^2$  with  $\deg \sigma_0 \leq 2t$  and  $\deg \sigma_j v_j \leq 2t$  for every  $j \in J$ .

Let  $r \in \mathbb{N}$  be fixed. Since  $\mathbf{X} \subset [-1, 1]^n$ , there holds  $1 \pm x^\alpha > 0$  on  $\mathbf{X}$  for every  $\alpha \in \mathbb{N}^n$  with  $|\alpha| \leq 2r$ . Therefore, since  $\mathbf{X}$  satisfies Putinar's condition (see Definition 3.4), the polynomial  $x \mapsto 1 \pm x^\alpha$  belongs to  $Q$  (see Putinar [39]). Moreover, there exists  $l(r)$  such that  $x \mapsto 1 \pm x^\alpha \in Q(l(r))$  for every  $|\alpha| \leq 2r$ . Of course,  $x \mapsto 1 \pm x^\alpha \in Q(l)$  for every  $|\alpha| \leq 2r$  whenever  $l \geq l(r)$ .

For every feasible solution  $z$  of  $\mathbf{Q}_{l(r)}$  one has

$$|z(x)_\alpha| = |L_z(x^\alpha)| \leq z_0 = 1, \quad |\alpha| \leq 2r.$$

This follows from  $z_0 = 1$ ,  $M_{l(r)}(z) \succeq 0$ , and  $M_{l(r)-r(v_j)}(v_j z(x)) \succeq 0$  which implies

$$z_0 \pm z(x)_\alpha = L_z(1 \pm x^\alpha) = L_z(\sigma_0) + \sum_{j=1}^m L_{z(x)}(\sigma_j v_j) \geq 0.$$

With similar arguments, one redefines  $l(r)$  so that, for every couple  $(y, z)$  satisfying the constraints of  $\mathbf{Q}_{l(r)}$ , one has

$$0 \leq z_k(t) \leq 1 \quad \text{and} \quad |z(x)_\alpha|, |z(u)_\beta|, |y_\alpha| \leq 1 \quad \forall k, |\alpha|, |\beta| \leq 2r.$$

From this, and with  $l(r) := 2l(r)$ , we immediately deduce that  $|z_\gamma| \leq 1$  whenever  $|\gamma| \leq 2r$ . In particular,  $L_y(H) + L_z(h) \geq -\sum_\beta |H_\beta| - \sum_\gamma |h_\gamma|$ , which proves that  $\inf \mathbf{Q}_{l(r)} > -\infty$ , and so  $\inf \mathbf{Q}_r > -\infty$  for  $r$  sufficiently large.

Let  $\rho := \inf \mathbf{P} = \min \mathbf{P}$  (by Theorem 2.3), let  $r \geq l(r_0)$ , and let  $(z^r, y^r)$  be a nearly optimal solution of  $\mathbf{Q}_r$  with value

$$(A.2) \quad \inf \mathbf{Q}_r \leq L_{z^r}(h) + L_{y^r}(H) \leq \inf \mathbf{Q}_r + \frac{1}{r} \quad \left( \leq \rho + \frac{1}{r} \right).$$

Complete the finite vectors  $y^r$  and  $z^r$  with zeros to make them infinite sequences. Since for arbitrary  $s \in \mathbb{N}$  one has  $|y_\alpha^r|, |z_\gamma^r| \leq 1$  whenever  $|\alpha|, |\gamma| \leq 2s$ , provided  $r$  is sufficiently large, by a standard diagonal argument, there exists a subsequence  $\{r_k\}$  and two infinite sequences  $y$  and  $z$ , with  $|y_\alpha| \leq 1$  and  $|z_\gamma| \leq 1$  for all  $\alpha, \gamma$ , and such that

$$(A.3) \quad \lim_{k \rightarrow \infty} y_\alpha^{r_k} = y_\alpha \quad \forall \alpha \in \mathbb{N}^n, \quad \lim_{k \rightarrow \infty} z_\gamma^{r_k} = z_\gamma \quad \forall \gamma \in \mathbb{N} \times \mathbb{N}^n \times \mathbb{N}^m.$$

Next, let  $r$  be fixed arbitrarily. Observe that  $M_{r_k}(y^{r_k}) \succeq 0$  implies  $M_r(y^{r_k}) \succeq 0$  whenever  $r_k \geq r$ , and similarly  $M_r(z^{r_k}) \succeq 0$ . Therefore, from (A.3) and  $M_r(y^{r_k}) \succeq 0$ , we deduce that  $M_r(y) \succeq 0$ , and similarly  $M_r(z) \succeq 0$ . Since this holds for arbitrary  $r$ , and  $|y_\alpha|, |z_\gamma| \leq 1$  for all  $\alpha, \gamma$ , one infers from Proposition 3.2 that  $y$  and  $z$  are moment sequences of two measures  $\nu$  and  $\mu$  with support contained in  $[-1, 1]^n$  and  $[0, 1] \times [-1, 1]^n \times [-1, 1]^m$ , respectively. In addition, from the equalities  $y_0^{r_k} = 1$  and

$z_0^{rk} = 1$  for every  $k$ , it follows that  $\nu$  and  $\mu$  are probability measures on  $[-1, 1]^n$  and  $[0, 1] \times [-1, 1]^n \times [-1, 1]^m$ .

Next, let  $(t, \alpha) \in \mathbb{N} \times \mathbb{N}^n$  be fixed arbitrarily. From

$$L_{y^{rk}}(g_1) - g(0, x_0) - L_{z^{rk}}(\partial g / \partial t + \langle \nabla_x g, f \rangle) = 0, \quad \text{with } g = (t^p x^\alpha),$$

and the convergence (A.3), we obtain

$$L_y(g_1) - g(0, x_0) - L_z(\partial g / \partial t + \langle \nabla_x g, f \rangle) = 0, \quad \text{with } g = (t^p x^\alpha),$$

that is,  $\langle \mathcal{L}g, (\mu, \nu) \rangle = \langle g, \delta_{(0, x_0)} \rangle$ . Since  $(t, \alpha) \in \mathbb{N} \times \mathbb{N}^n$  is arbitrary, we have

$$\langle g, \mathcal{L}^*(\mu, \nu) \rangle = \langle \mathcal{L}g, (\mu, \nu) \rangle = \langle g, \delta_{(0, x_0)} \rangle \quad \forall g \in \mathbb{R}[t, x],$$

which implies that  $\mathcal{L}^*(\mu, \nu) = \delta_{(0, x_0)}$ .

Let  $z(x)$ ,  $z(u)$ , and  $z(t)$  denote the moment vectors of the marginals of  $\mu$  with respect to the variables  $x \in \mathbb{R}^n$ ,  $u \in \mathbb{R}^m$ , and  $t \in \mathbb{R}$ , respectively, i.e.,

$$z(x)^\alpha = \int x^\alpha \mu(d(t, x, u)) \quad \forall \alpha \in \mathbb{N}^n, \quad z(u)^\beta = \int u^\beta \mu(d(t, x, u)) \quad \forall \beta \in \mathbb{N}^m,$$

and  $z(t)^k = \int t^k \mu(d(t, x, u))$  for every  $k \in \mathbb{N}$ .

With  $r$  fixed arbitrarily, and using again (A.3), we also have  $M_r(\theta_j y) \succeq 0$  for every  $j \in J_T$ , and

$$M_r(v_j z(x)) \succeq 0 \quad \forall j \in J, \quad M_r(w_k z(u)) \succeq 0 \quad \forall k \in W, \quad M_r(t(1-t)z(t)) \succeq 0.$$

Since  $\mathbf{X}$ ,  $\mathbf{K}$ , and  $\mathbf{U}$  satisfy Putinar's condition (see Definition 3.4), from Theorem 3.5 (Putinar's Positivstellensatz),  $y$  is the moment sequence of a probability measure  $\nu$  supported on  $\mathbf{K} \subset [-1, 1]^n$ . Similarly,  $z(x)$  is the moment sequence of a measure  $\mu^x$  supported on  $\mathbf{X} \subset [-1, 1]^n$ ,  $z(u)$  is the moment sequence of a measure  $\mu^u$  supported on  $\mathbf{U} \subset [-1, 1]^m$ , and  $z(t)$  is the moment sequence of a measure  $\mu^t$  supported on  $[0, 1]$ . Since measures on compact sets are moment determinate, it follows that  $\mu^x$ ,  $\mu^u$ , and  $\mu^t$  are the marginals of  $\mu$  with respect to  $x$ ,  $u$ , and  $t$ , respectively. Therefore,  $\mu$  has its support contained in  $\mathbf{S}$ , and from  $\mathcal{L}^*(\mu, \nu) = \delta_{(0, x_0)}$  it follows that  $(\mu, \nu)$  satisfies all constraints of the problem  $\mathbf{P}$ .

Moreover, one has

$$\begin{aligned} \lim_{k \rightarrow \infty} \inf \mathbf{Q}_{rk} &= \lim_{k \rightarrow \infty} L_{z^{rk}}(h) + L_{y^{rk}}(H) \quad (\text{by (A.2)}) \\ &= L_z(h) + L_y(H) \quad (\text{by (A.3)}) \\ &= \int h d\mu + \int H d\nu \leq \rho = \min \mathbf{P}. \end{aligned}$$

Hence,  $(\mu, \nu)$  is an optimal solution of  $\mathbf{P}$ , and  $\min \mathbf{Q}_r \uparrow \min \mathbf{P}$  (the sequence is monotone nondecreasing). Item (i) is proved.

Item (ii) follows from Theorem 2.3(iii).  $\square$

**A.3. Proof of Theorem 4.1.** It suffices to prove that  $v_p \rightarrow v$  as  $p \rightarrow +\infty$ . For every integer  $p$ ,  $v_p = \min \mathbf{P}_p$  is attained for a couple of measures  $(\mu_p, \nu_p)$ . As in the proof of Theorem 2.3, the sequence  $\{(\mu_p, \nu_p)\}_{p \in \mathbb{N}}$  is bounded in  $\mathcal{M}(\mathbf{S})_+ \times \mathcal{M}(\mathbf{K})_+$ , and hence, along a subsequence, it converges to an element  $(\mu, \nu)$  of this space for the weak\* topology.

On the one hand, by definition,  $\mathcal{L}_p^*(\mu_p, \nu_p) = \delta_{(0, x_0)}$  for every  $p$ . On the other,  $\mathcal{L}_p^*$  tends strongly to  $\mathcal{L}^*$ , and so  $\mathcal{L}^*(\mu, \nu) = \delta_{(0, x_0)}$ . Moreover, since  $(h_p, H_p)$  tends strongly to  $(h, H)$  in  $C_1(\mathbf{S}) \times C_1(\mathbf{K})$ , one has

$$v_p = \langle (\mu_p, \nu_p), (h_p, H_p) \rangle \longrightarrow \langle (\mu, \nu), (h, H) \rangle,$$

and so  $v \leq \langle (\mu, \nu), (h, H) \rangle$ . We next prove that  $v = \langle (\mu, \nu), (h, H) \rangle$ .

Since  $(\mu_p, \nu_p)$  is an optimal solution of  $\mathbf{P}_p$ ,

$$\langle (\mu_p, \nu_p), (h_p, H_p) \rangle \leq \langle (\bar{\mu}, \bar{\nu}), (h_p, H_p) \rangle \quad \forall (\bar{\mu}, \bar{\nu}) \mid \mathcal{L}_p^*(\bar{\mu}, \bar{\nu}) = \delta_{(0, x_0)}.$$

Hence, passing to the limit,

$$\langle (\mu, \nu), (h, H) \rangle \leq \langle (\bar{\mu}, \bar{\nu}), (h, H) \rangle \quad \forall (\bar{\mu}, \bar{\nu}) \mid \mathcal{L}^*(\bar{\mu}, \bar{\nu}) = \delta_{(0, x_0)},$$

and so  $(\mu, \nu)$  is an optimal solution of  $\mathbf{P}$ , i.e.,  $v = \langle (\mu, \nu), (h, H) \rangle$ .  $\square$

**Acknowledgment.** This work benefited from comments by Carlo Savorgnan.

#### REFERENCES

- [1] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces*, John Wiley, Chichester, UK, 1987.
- [2] R. BEALS, B. GAVEAU, AND P. C. GREINER, *Hamilton-Jacobi theory and the heat kernel on Heisenberg groups*, J. Math. Pures Appl., 79 (2000), pp. 633–689.
- [3] C. BERG, *The multidimensional moment problem and semigroups*, Proc. Symp. Appl. Math., 37 (1987), pp. 110–124.
- [4] A. G. BHATT AND V. S. BORKAR, *Occupation measures for controlled Markov processes: Characterization and optimality*, Ann. Probab., 24 (1996), pp. 1531–1562.
- [5] O. BOKANOWSKI, S. MARTIN, R. MUNOS, AND H. ZIDANI, *An anti-diffusive scheme for viability problems*, Appl. Numer. Math., 56 (2006), pp. 1147–1162.
- [6] V. BORKAR, *Convex analytic methods in Markov decision processes*, in Handbook of Markov Decision Processes, E. A. Feinberg and A. Shwartz, eds., Kluwer Academic, Boston, MA, 2002, pp. 377–408.
- [7] R. W. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhäuser, Boston, MA, 1983, pp. 181–191.
- [8] I. CAPUZZO-DOLCETTA AND P. L. LIONS, *Hamilton-Jacobi equations with state constraints*, Trans. Amer. Math. Soc., 318 (1990), pp. 643–683.
- [9] M. J. CHO AND R. H. STOCKBRIDGE, *Linear programming formulation for optimal stopping problems*, SIAM J. Control Optim., 40 (2002), pp. 1965–1982.
- [10] C. COATMÉLEC, *Approximation et interpolation des fonctions différentiables de plusieurs variables*, Ann. Sci. École Norm. Sup. (3), 83 (1966), pp. 271–341.
- [11] D. A. DAWSON, *Qualitative behavior of geostochastic systems*, Stochastic Process. Appl., 10 (1980), pp. 1–31.
- [12] W. H. FLEMING AND D. VERMES, *Convex duality approach to the optimal control of diffusions*, SIAM J. Control Optim., 27 (1989), pp. 1136–1155.
- [13] R. FLETCHER, *Practical Methods of Optimization. Vol. 1. Unconstrained Optimization*, John Wiley, Chichester, UK, 1980.
- [14] V. GAITSGORY AND S. ROSSOMAKHINE, *Linear programming approach to deterministic long run average problems of optimal control*, SIAM J. Control Optim., 44 (2006), pp. 2006–2037.
- [15] P. E. GILL, W. MURRAY, AND M. H. WRIGHT, *Practical Optimization*, Academic Press, London, New York, 1981.
- [16] R. F. HARTL, S. P. SETHI, AND R. G. VICKSON, *A survey of the maximum principles for optimal control problems with state constraints*, SIAM Rev., 37 (1995), pp. 181–218.
- [17] K. HELMES, S. RÖHL, AND R. H. STOCKBRIDGE, *Computing moments of the exit time distribution for Markov processes by linear programming*, Oper. Res., 49 (2001), pp. 516–530.
- [18] K. HELMES AND R. H. STOCKBRIDGE, *Numerical comparison of controls and verification of optimality for stochastic control problems*, J. Optim. Theory Appl., 106 (2000), pp. 107–127.

- [19] D. HENRION AND J. B. LASSERRE, *Solving nonconvex optimization problems*, IEEE Control Systems Mag., 24 (2004), pp. 72–83.
- [20] D. HERNANDEZ-HERNANDEZ, O. HERNÁNDEZ-LERMA, AND M. TAKSAR, *The linear programming approach to deterministic optimal control problems*, Appl. Math., 24 (1996), pp. 17–33.
- [21] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.
- [22] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Approximation schemes for infinite linear programs*, SIAM J. Optim., 8 (1998), pp. 973–988.
- [23] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Further Topics in Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.
- [24] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *The linear programming approach*, in Handbook of Markov Decision Processes, E. A. Feinberg and A. Shwartz, eds., Kluwer Academic, Boston, MA, 2002, pp. 377–408.
- [25] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Markov Chains and Invariant Probabilities*, Birkhäuser Verlag, Basel, 2003.
- [26] M. W. HIRSCH, *Differential Topology*, Grad. Texts in Math. 33, Springer-Verlag, New York, 1976.
- [27] D. JACOBSON, M. LELE, AND J. L. SPEYER, *New necessary conditions of optimality for control problems with state-variable inequality constraints*, J. Math. Anal. Appl., 35 (1971), pp. 255–284.
- [28] J. L. KRIVINE, *Anneaux préordonnés*, J. Anal. Math., 12 (1964), pp. 307–326.
- [29] T. G. KURTZ AND R. H. STOCKBRIDGE, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim., 36 (1998), pp. 609–653.
- [30] J. B. LASSERRE, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
- [31] J. B. LASSERRE AND T. PRIÉTO-RUMEAU, *SDP vs. LP relaxations for the moment approach in some performance evaluation problems*, Stoch. Models, 20 (2004), pp. 439–456.
- [32] J. B. LASSERRE, T. PRIÉTO-RUMEAU, AND M. ZERVOS, *Pricing a class of exotic options via moments and SDP relaxations*, Math. Finance, 16 (2006), pp. 469–494.
- [33] J. B. LASSERRE, C. PRIEUR, AND D. HENRION, *Nonlinear optimal control: Numerical approximations via moments and LMI relaxations*, in Proceedings of the 44th IEEE Conference on Decision and Control, Sevilla, Spain, 2005, pp. 1648–1653.
- [34] H. MAURER, *On optimal control problems with bounded state variables and control appearing linearly*, SIAM J. Control Optim., 15 (1977), pp. 345–362.
- [35] J. NASH, *The imbedding problem for Riemannian manifolds*, Ann. of Math. (2), 63 (1956), pp. 20–63.
- [36] H. J. PESCH, *A practical guide to the solution of real-life optimal control problems*, Control Cybernet., 23 (1994), pp. 7–60.
- [37] L. S. PONTRYAGIN, V. G. BOLTYANSKIĬ, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.
- [38] C. PRIEUR AND E. TRÉLAT, *Robust optimal stabilization of the Brockett integrator via a hybrid feedback*, Math. Control Signals Systems, 17 (2005), pp. 201–216.
- [39] M. PUTINAR, *Positive polynomials on compact semi-algebraic sets*, Indiana Univ. Math. J., 42 (1993), pp. 969–984.
- [40] K. SCHMÜDGEN, *The  $K$ -moment problem for compact semi-algebraic sets*, Math. Ann., 289 (1991), pp. 203–206.
- [41] H. M. SONER, *Optimal control with state-space constraints I*, SIAM J. Control Optim., 24 (1986), pp. 552–561.
- [42] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, 3rd ed., Springer-Verlag, New York, 2002.
- [43] E. TRÉLAT, *Contrôle optimal: Théorie et applications*, Vuibert, Paris, 2005.
- [44] F.-H. VASILESCU, *Spectral measures and moment problems*, in Spectral Analysis and Its Applications, Theta Ser. Adv. Math. 2, Theta, Bucharest, 2003, pp. 173–215.
- [45] R. VINTER, *Convex duality and nonlinear optimal control*, SIAM J. Control Optim., 31 (1993), pp. 518–538.
- [46] O. VON STRYK AND R. BULIRSCH, *Direct and indirect methods for trajectory optimization*, Ann. Oper. Res., 37 (1992), pp. 357–373.

## DUALITY IN LINEAR PROGRAMMING PROBLEMS RELATED TO DETERMINISTIC LONG RUN AVERAGE PROBLEMS OF OPTIMAL CONTROL\*

LUKE FINLAY<sup>†</sup>, VLADIMIR GAITSGORY<sup>†</sup>, AND IVAN LEBEDEV<sup>‡</sup>

**Abstract.** It has been established recently that, under mild conditions, deterministic long run average problems of optimal control are “asymptotically equivalent” to infinite-dimensional linear programming problems (LPPs) and that these LPPs can be approximated by finite-dimensional LPPs. In this paper we introduce the corresponding infinite- and finite-dimensional dual problems and study duality relationships. We also investigate the possibility of using solutions of finite-dimensional LPPs and their duals for numerical construction of the optimal controls in periodic optimization problems. The construction is illustrated with a numerical example.

**Key words.** long run average optimal control, occupational measures, averaging, linear programming, duality

**AMS subject classifications.** 34E15, 34C29, 34A60, 93C70

**DOI.** 10.1137/060676398

**1. Introduction.** Infinite horizon problems of optimal control have been studied intensively in both deterministic and stochastic settings (see [1], [2], [3], [6], [7], [8], [9], [12], [14], [18], [19], [20], [21], [23], [26], [31], [37], [42], [43], [47], [56], and references therein for a sample of the literature on the subject) with linear programming formulations being one of the main tools of treating stochastic problems (see, e.g., [13], [16], [5], [38], [39], [41], [46], [51], [52], [60]).

A linear programming approach to deterministic long run average problems of optimal control was considered in [32] and [33], where it was established that these problems are “asymptotically equivalent” to infinite-dimensional (I-D) linear programming problems (LPPs) similar to those arising in stochastic control (see [13], [16], [41], [52]), and it has been shown that these I-D LPPs can be approximated by finite-dimensional (F-D) LPPs (F-D approximations of I-D LPPs arising in stochastic control problems and in deterministic problems on finite intervals of time have been studied in [38], [46], and in [48], respectively; F-D approximations of I-D LPPs arising in certain problems of calculus of variations have been considered in [24]).

In this paper we introduce problems dual to the LPPs considered in [32] and [33] (both F-D and I-D), and we study the corresponding duality relationships. We also investigate the possibility of using solutions of F-D LPPs and their duals for numerical construction of optimal controls in periodic optimization problems (thus refining the corresponding results of [32], where the option of using dual solutions was not considered).

---

\*Received by the editors November 30, 2006; accepted for publication (in revised form) February 12, 2008; published electronically June 11, 2008.

<http://www.siam.org/journals/sicon/47-4/67639.html>

<sup>†</sup>Centre for Industrial and Applied Mathematics, University of South Australia, Mawson Lakes, SA 5095, Australia (luke.finlay@unisa.edu.au, v.gaitsgory@unisa.edu.au). The second author’s work was supported by the Australian Research Council Discovery grants DP0346099 and DP0664330, IREX grant X00106494, and Linkage International grant LX0560049.

<sup>‡</sup>WorkCover Corporation of South Australia, 100 Waymouth St. Adelaide SA 5000, Australia (ilebedev@workcover.com). This author’s work was supported by the Australian Research Council Discovery grant DP0346099.



Periodic optimization problems (POPs) constitute a special family of long run average problems of optimal control, in which solutions are sought over the class of periodic regimes. POPs arise in various applications, including vibration damping, production planning, flight control, chemical engineering, ecological modeling (see [40], [45], [50], [53], [59], and references therein). A linear programming approach that we continue to develop in the present paper provides a new analytical and numerical tool for dealing with this important family of control problems.

The paper is organized as follows. Sections 2 and 3 contain some preliminaries, the purpose of which is to help the reader put the developments of the subsequent sections into perspective. In section 2, long run average problems of optimal control and their reformulations in terms of occupational measures are considered. In section 3, some results of [32] and [33] that establish connections between the long run average problems of optimal control and the I-D LPPs are restated.

Sections 4, 5, and 6 contain the main duality results. In section 4, the problem dual to the I-D LPP is introduced and duality relationships are established (Theorem 4.1). The dual problem proves to be closely related to the Hamilton–Jacobi–Bellman equation. Its solution is used to state some necessary and sufficient optimality conditions (Proposition 4.3 and Corollaries 4.4 and 4.5). In section 5, duality relationships for an I-D LPP with a finite number of constraints are discussed (Theorem 5.2) and convergence properties (as the number of constraints goes to infinity) are established (Proposition 5.1 and Theorem 5.6). In section 6, an F-D LPPs (defined on grid points) and its dual are considered, with their convergence properties (as the grid size goes to zero) being established (Theorems 6.1 and 6.2). Also, in this section, a construction of an approximate solution of the problem dual to the I-D LPPs introduced in section 4 is discussed (Proposition 6.3).

Sections 7 and 8 are devoted to construction of solutions of periodic optimization problems. In section 7, it is established that, under certain conditions, the control found with the help of an approximate solution of the problem dual to the I-D LPP (constructed in section 6) converges to the optimal control (Theorem 7.1 and Corollary 7.2). In section 8, a numerical example is considered to illustrate the construction.

In sections 9 and 10, the proofs for sections 4, 5, 6, and 7, respectively, are collected.

**2. Preliminaries I: Occupational measures formulations.** Consider the control system

$$(2.1) \quad \dot{y}(\tau) = f(u(\tau), y(\tau)), \quad \tau \in [0, S],$$

where the function  $f(u, y) : U \times \mathbb{R}^m \rightarrow \mathbb{R}^m$  is continuous in  $(u, y)$  and satisfies Lipschitz conditions in  $y$ ; the controls are Lebesgue measurable functions  $u(\tau) : [0, S] \rightarrow U$  and  $U$  is a compact metric space.

A pair  $(u(\tau), y(\tau))$  will be called *admissible* on the interval  $[0, S]$  if (2.1) is satisfied for almost all  $\tau \in [0, S]$  and  $y(\tau) \in Y \quad \forall \tau \in [0, S]$ , where  $Y$  is a given compact subset of  $\mathbb{R}^m$ . The pair will be called admissible on  $[0, \infty)$  if it is admissible on any interval  $[0, S]$ ,  $S > 0$ .

Let  $\mathcal{P}(U)$  be the space of probability measures defined on the Borel subsets of  $U$  and  $\bar{f}(\nu, y) \stackrel{\text{def}}{=} \int_U f(u, y) \nu(du)$ . Along with (2.1), let us consider a relaxed control system

$$(2.2) \quad \dot{y}(\tau) = \bar{f}(\nu(\tau), y(\tau)), \quad \tau \in [0, S],$$

in which controls are measurable functions  $\nu(\tau) \in \mathcal{P}(U)$  (see [58]).

A pair  $(\nu(\tau), y(\tau))$  will be called *relaxed admissible* on the interval  $[0, S]$  if (2.2) is satisfied for almost all  $\tau \in [0, S]$  and  $y(\tau) \in Y \forall \tau \in [0, S]$ . The pair will be called relaxed admissible on  $[0, \infty)$  if it is relaxed admissible on any interval  $[0, S]$ ,  $S > 0$ .

Let  $g(u, y) : U \times \mathbb{R}^m \rightarrow \mathbb{R}^1$  be a continuous function. We will be considering the optimal control problem

$$(2.3) \quad \frac{1}{S} \inf_{(u(\cdot), y(\cdot))} \int_0^S g(u(\tau), y(\tau)) d\tau \stackrel{\text{def}}{=} G(S)$$

and the corresponding relaxed optimal control problem

$$(2.4) \quad \frac{1}{S} \inf_{(\nu(\cdot), y(\cdot))} \int_0^S \bar{g}(\nu(\tau), y(\tau)) d\tau \stackrel{\text{def}}{=} \bar{G}(S),$$

where  $\bar{g}(\nu, u) \stackrel{\text{def}}{=} \int_U g(u, y) \nu(dy)$ , and the first (second) inf is over all admissible (respectively, relaxed admissible) pairs.

Let  $\mathcal{P}(U \times Y)$  stand for the space of probability measures defined on the Borel subsets of  $U \times Y$ . Given an admissible pair  $(u(\cdot), y(\cdot))$ , a probability measure  $\gamma \in \mathcal{P}(U \times Y)$  will be said to be generated by this pair on the interval  $[0, S]$  if

$$(2.5) \quad \int_{U \times Y} q(u, y) \gamma(du, dy) = \frac{1}{S} \int_0^S q(u(\tau), y(\tau)) d\tau$$

for any  $q(\cdot) \in C(U \times Y)$  (the space of continuous functions on  $U \times Y$ ). A probability measure  $\gamma \in \mathcal{P}(U \times Y)$  will be said to be generated by an admissible pair  $(u(\cdot), y(\cdot))$  on  $[0, \infty)$  if

$$(2.6) \quad \int_{U \times Y} q(u, y) \gamma(du, dy) = \lim_{S \rightarrow \infty} \frac{1}{S} \int_0^S q(u(\tau), y(\tau)) d\tau$$

for any  $q(\cdot) \in C(U \times Y)$  (the limit in the right-hand side of (2.6) is assumed to exist). Given a relaxed admissible pair  $(\nu(\tau), y(\tau))$ , a probability measure  $\gamma \in \mathcal{P}(U \times Y)$  will be said to be generated by this pair on the interval  $[0, S]$  if

$$(2.7) \quad \int_{U \times Y} q(u, y) \gamma(du, dy) = \frac{1}{S} \int_0^S \bar{q}(\nu(\tau), y(\tau)) d\tau$$

for any  $q(\cdot) \in C(U \times Y)$ , with

$$(2.8) \quad \bar{q}(\nu, u) \stackrel{\text{def}}{=} \int_U q(u, y) \nu(dy).$$

A probability measure  $\gamma \in \mathcal{P}(U \times Y)$  will be said to be generated by an admissible pair  $(\nu(\cdot), y(\cdot))$  on  $[0, \infty)$  if

$$(2.9) \quad \int_{U \times Y} q(u, y) \gamma(du, dy) = \lim_{S \rightarrow \infty} \frac{1}{S} \int_0^S \bar{q}(\nu(\tau), y(\tau)) d\tau$$

for any  $q(\cdot) \in C(U \times Y)$  and  $\bar{q}(\cdot)$  is as in (2.8) (the limit in the right-hand side of (2.9) is assumed to exist).

Thus defined probability measures are called occupational measures (see, e.g., [7] and [30]). In what follows  $\gamma^{(S,u(\cdot),y(\cdot))}$ ,  $\gamma^{(u(\cdot),y(\cdot))}$  will stand for the occupational measures generated by an admissible pair  $(u(\cdot), y(\cdot))$  on  $[0, S]$  and  $[0, \infty)$ , respectively, and  $\gamma^{(S,\nu(\cdot),y(\cdot))}$ ,  $\gamma^{(\nu(\cdot),y(\cdot))}$  will stand for the occupational measures generated by a relaxed admissible pair  $(\nu(\cdot), y(\cdot))$  on  $[0, S]$  and  $[0, \infty)$ , respectively.

Denote by  $\Gamma(S) \subset \mathcal{P}(U \times Y)$  and  $\bar{\Gamma}(S) \subset \mathcal{P}(U \times Y)$  the set of all occupational measures generated by the admissible (respectively, relaxed admissible) pairs on the interval  $[0, S]$ . That is,

$$(2.10) \quad \Gamma(S) \stackrel{\text{def}}{=} \bigcup_{(u(\cdot), y(\cdot))} \left\{ \gamma^{(S,u(\cdot),y(\cdot))} \right\}, \quad \bar{\Gamma}(S) \stackrel{\text{def}}{=} \bigcup_{(\nu(\cdot), y(\cdot))} \left\{ \gamma^{(S,\nu(\cdot),y(\cdot))} \right\},$$

where the unions are over all admissible and, respectively, relaxed admissible pairs on  $[0, S]$ . Using these notations, one can rewrite the problem (2.3) as

$$(2.11) \quad \inf_{\gamma \in \Gamma(S)} \int_{U \times Y} g(u, y) \gamma(du, dy) = G(S)$$

and the problem (2.4) as

$$(2.12) \quad \inf_{\gamma \in \bar{\Gamma}(S)} \int_{U \times Y} g(u, y) \gamma(du, dy) = \bar{G}(S).$$

Note that  $\Gamma(S) \subset \bar{\Gamma}(S)$  and  $G(S) \geq \bar{G}(S)$ . In a special case when  $Y$  is forward invariant with respect to the solutions of (2.1), from Filippov–Wazewski’s theorem (see, e.g., Theorem 10.4.4 in [11]) it follows that  $\bar{\Gamma}(S)$  is equal to the closure of  $\Gamma(S)$  and  $G(S) = \bar{G}(S)$  for any  $S > 0$ .

Let us introduce a metric  $\rho(\cdot, \cdot)$  on  $\mathcal{P}(U \times Y)$  by the equation

$$(2.13) \quad \rho(\gamma', \gamma'') \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \frac{1}{2^j} \left| \int_{U \times Y} q_j(u, y) \gamma'(du, dy) - \int_{U \times Y} q_j(u, y) \gamma''(du, dy) \right|,$$

where  $\gamma', \gamma'' \in \mathcal{P}(U \times Y)$ , and  $q_j(\cdot), j = 1, 2, \dots$ , is a sequence of Lipschitz continuous functions that is dense in the unit ball of  $C(U \times Y)$ . Note that this metric is consistent with the topology of weak convergence in  $\mathcal{P}(U \times Y)$ . Namely, a sequence  $\gamma^k \in \mathcal{P}(U \times Y)$  converges to  $\gamma \in \mathcal{P}(U \times Y)$  in this metric if and only if

$$(2.14) \quad \lim_{k \rightarrow \infty} \int_{U \times Y} q(u, y) \gamma^k(du, dy) = \int_{U \times Y} q(u, y) \gamma(du, dy)$$

for any  $q(\cdot) \in C(U \times Y)$ . Note also that the space  $\mathcal{P}(U \times Y)$  is known to be compact in the topology of weak convergence (see, e.g., [15]), and, hence, being equipped with the metric (2.13), it becomes a compact metric space.

Using  $\rho(\cdot, \cdot)$ , one can define the Hausdorff metric on the subsets of  $\mathcal{P}(U \times Y)$  as follows: for any  $\Gamma_1 \subset \mathcal{P}(U \times Y)$  and  $\Gamma_2 \subset \mathcal{P}(U \times Y)$ ,

$$(2.15) \quad \begin{aligned} \rho_H(\Gamma_1, \Gamma_2) &\stackrel{\text{def}}{=} \max \left\{ \sup_{\gamma \in \Gamma_1} \rho(\gamma, \Gamma_2), \sup_{\gamma \in \Gamma_2} \rho(\gamma, \Gamma_1) \right\}, \\ \rho(\gamma, \Gamma_i) &\stackrel{\text{def}}{=} \inf_{\gamma' \in \Gamma_i} \rho(\gamma, \gamma'), \quad i = 1, 2. \end{aligned}$$

Note that, although, by some abuse of terminology, we call  $\rho_H(\cdot, \cdot)$  a metric on the set of subsets of  $\mathcal{P}(U \times Y)$ , it is, in fact, a semimetric on this set (since  $\rho_H(\Gamma_1, \Gamma_2) = 0$  implies that  $\Gamma_1 = \Gamma_2$  if and only if  $\Gamma_1$  and  $\Gamma_2$  are closed).

**3. Preliminaries II: Limit linear programming problem.** Define the set  $W \subset \mathcal{P}(U \times Y)$  by the equation

$$(3.1) \quad W \stackrel{\text{def}}{=} \left\{ \gamma \in \mathcal{P}(U \times Y) : \int_{U \times Y} (\phi'(y))^T f(u, y) \gamma(du, dy) = 0 \quad \forall \phi(\cdot) \in C^1 \right\},$$

where  $C^1$  is the space of continuously differentiable functions  $\phi(y) : \mathbb{R}^m \rightarrow \mathbb{R}^1$  and  $\phi'(y)$  is a vector column of partial derivatives (the gradient) of  $\phi(y)$ .

It is easy to see that  $W$  is closed and compact in weak convergence topology of  $\mathcal{P}(U \times Y)$ . Also, it is easy to see that  $W$  is convex.

Assuming that  $W$  is not empty (see Remark 4.2 below about a necessary and sufficient condition for this to be the case), let us consider the problem

$$(3.2) \quad \min_{\gamma \in W} \int_{U \times Y} g(u, y) \gamma(du, dy) \stackrel{\text{def}}{=} G_*,$$

where  $g(\cdot)$  is the same as in (2.3) and (2.11). Since both the objective function and the constraints defining  $W$  are linear in  $\gamma$ , the problem (3.2) is that of I-D linear programming (see, e.g., [4]). Note that a problem similar to (3.2) was introduced in [41] and [52] in a much more general stochastic setting. The problem (3.2) will in what follows be referred to as the limit linear programming problem (LLPP). In the rest of this section we restate some results of [32] and [33] that are important for our further consideration.

**THEOREM 3.1.** (i) *If the set  $W$  is empty, then there exists  $S_0 > 0$  such that  $\bar{\Gamma}(S)$  (and, hence,  $\Gamma(S)$ ) are empty for  $S \geq S_0$ . If  $W$  is not empty, then  $\bar{\Gamma}(S)$  is not empty for  $S > 0$ . If  $W$  is not empty and, in addition, the set  $f(U, y) \stackrel{\text{def}}{=} \{\eta : \eta = f(u, y), u \in U\}$  is convex for all  $y \in Y$ , then  $\Gamma(S)$  is not empty for  $S > 0$ .*

(ii) *Let  $W$  be not empty. Then*

$$(3.3) \quad \lim_{S \rightarrow \infty} \rho_H(\text{co}\bar{\Gamma}(S), W) = 0$$

and

$$(3.4) \quad \lim_{S \rightarrow \infty} \bar{G}(S) = G_*.$$

(iii) *If  $Y$  is forward invariant, then*

$$(3.5) \quad \lim_{S \rightarrow \infty} \rho_H(\text{co}\Gamma(S), W) = 0$$

and

$$(3.6) \quad \lim_{S \rightarrow \infty} G(S) = G_*.$$

*Proof.* The statements (i) and (ii) have been established in [33], their proofs being similar to the proofs of Theorem 2.1 in [29] and Proposition 5 in [32]. The statement (iii) is implied by Proposition 5 in [32].  $\square$

**Remark 3.2.** In the terminology introduced in [10], the fact that the set  $\Gamma(s)$  (or  $\bar{\Gamma}(s)$ ) is not empty for all  $S > 0$  means that the *viability kernel* of the control system (2.1) (respectively, relaxed control system (2.2)) is not empty in  $Y$ . Hence, from Theorem 3.1(i) it follows that the viability kernel of the relaxed control system (2.2) is not empty in  $Y$  if and only if  $W$  is not empty, and, under the condition that

the set  $f(U, y)$  is convex for all  $y \in Y$ , the viability kernel of the control system (2.1) is not empty in  $Y$  if and only if  $W$  is not empty. Note that an alternative way of establishing this statement was proposed by Quincampoix, and it will be included in a separate publication.

**THEOREM 3.3.** *Assume that  $W$  is not empty. Then, corresponding to any extreme point  $\gamma$  of  $W$ , there exists a relaxed admissible pair  $(\nu(\cdot), y(\cdot))$  such that  $\gamma$  is generated by this pair on  $[0, \infty)$  (that is,  $\gamma = \gamma^{(\nu(\cdot), y(\cdot))}$ ).*

*Proof.* The theorem has been established in [33]. The proof is based on Proposition 4.4 in [17] and on Lemma 5.1 in [29].  $\square$

Consider the problem

$$(3.7) \quad \inf_{(u(\cdot), y(\cdot))} \lim_{S \rightarrow \infty} \frac{1}{S} \int_0^S g(u(\tau), y(\tau)) d\tau \stackrel{\text{def}}{=} G_\infty$$

and the problem

$$(3.8) \quad \inf_{(\nu(\cdot), y(\cdot))} \lim_{S \rightarrow \infty} \frac{1}{S} \int_0^S \bar{g}(\nu(\tau), y(\tau)) d\tau \stackrel{\text{def}}{=} \bar{G}_\infty,$$

where the infs in (3.7) and (3.8) are over all admissible and, respectively, over all relaxed admissible pairs on the interval  $[0, \infty)$  such that the corresponding limits exist.

**COROLLARY 3.4.** *If  $W$  is not empty, then  $\bar{G}_\infty = G_*$  and the problem (3.8) has a solution. That is, there exists  $(\nu(\cdot), y(\cdot))$ , a relaxed admissible pair on  $[0, \infty)$ , such that*

$$(3.9) \quad \lim_{S \rightarrow \infty} \frac{1}{S} \int_0^S \bar{g}(\nu(\tau), y(\tau)) d\tau = G_*.$$

*Proof.* The statement is implied by Theorem 3.3 and by the fact that LLPP (3.2) always has an extreme point solution.  $\square$

For the problem (3.7) to have a solution, one needs to introduce additional assumptions. In particular, one can use Corollary 3.4 and a standard measurable selection argument to show that, if  $(g, f)(U, y) \stackrel{\text{def}}{=} \{(\zeta, \eta) : \zeta = g(u, y), \eta = f(u, y), u \in U\}$  is convex for all  $y \in Y$ , then  $G_\infty = G_*$  and there exists  $(u(\cdot), y(\cdot))$ , an admissible pair on  $[0, \infty)$ , such that

$$(3.10) \quad \lim_{S \rightarrow \infty} \frac{1}{S} \int_0^S g(u(\tau), y(\tau)) d\tau = G_*.$$

Along with the problems considered above, let us also consider the problem

$$(3.11) \quad \inf_{(u(\cdot), y(\cdot))_{\text{per}}} \lim_{S \rightarrow \infty} \frac{1}{S} \int_0^S g(u(\tau), y(\tau)) d\tau \stackrel{\text{def}}{=} G_{\text{per}},$$

where inf is over all periodic admissible pairs, that is, over the admissible pairs such that

$$(3.12) \quad (u(\tau), y(\tau)) = (u(\tau + T), y(\tau + T)) \quad \forall \tau \geq 0,$$

for some  $T > 0$ . Note that the problem (3.11) is equivalent to a so-called periodic optimization problem

$$(3.13) \quad \inf_{(u(\cdot), y(\cdot))} \frac{1}{T} \int_0^T g(u(\tau), y(\tau)) d\tau = G_{\text{per}},$$

where  $\inf$  is over the length of the time interval  $T$  and over the admissible pairs defined on  $[0, T]$  which satisfy the periodicity condition  $y(0) = y(T)$ .

In the general case the optimal values of the problems (2.3), (3.7), and (3.11) satisfy the inequalities

$$(3.14) \quad \overline{\lim}_{S \rightarrow \infty} G(S) \leq G_\infty \leq G_{per} \Rightarrow G_* \leq G_{per},$$

and, under some additional assumptions,

$$(3.15) \quad \lim_{S \rightarrow \infty} G(S) = G_\infty = G_{per} \Rightarrow G_* = G_{per}.$$

Sufficient conditions for the equalities (3.15) to be valid have been considered in [27], [28], [35], and [36].

**LEMMA 3.5.** *If there exists a solution of the LLPP (3.2) that is generated by a periodic admissible pair, then this pair is a solution of the periodic optimization problem (3.13), and the inequalities (3.14) turn into equalities (3.15). Conversely, if equalities (3.15) are valid and if a solution of the periodic optimization problem (3.13) exists, then the occupational measure generated by this pair is a solution of the LLPP (3.2).*

*Proof.* The proof follows from (3.14) and from the definition of the occupational measures.  $\square$

**4. Dual problem and duality relationships.** Define the problem dual to LLPP (D-LLPP) by the equation

$$(4.1) \quad \sup_{(\mu, \psi(\cdot)) \in \mathcal{D}} \mu \stackrel{\text{def}}{=} \mu_*,$$

with the feasible set  $\mathcal{D} \subset \mathbb{R}^1 \times C^1$  defined as

$$(4.2) \quad \mathcal{D} \stackrel{\text{def}}{=} \{(\mu, \psi(\cdot)) : \mu = \min_{(u, y) \in U \times Y} \{g(u, y) + (\psi'(y))^T f(u, y)\}, \psi(\cdot) \in C^1\}.$$

It can be readily seen that, if  $W \neq \emptyset$  and  $\gamma \in W$ , then for any  $(\mu, \psi(\cdot)) \in \mathcal{D}$  (note that  $\mathcal{D}$  is never empty),

$$(4.3) \quad \mu \leq \int_{U \times Y} (g(u, y) + (\psi'(y))^T f(u, y)) \gamma(du, dy) = \int_{U \times Y} g(u, y) \gamma(du, dy)$$

$$(4.4) \quad \Rightarrow \mu_* \leq G_*.$$

The following statements establish more elaborate connections between LLPP (3.2) and D-LLPP (4.1).

**THEOREM 4.1.**

(i) *The optimal value of D-LLPP (4.1) is bounded (that is,  $\mu_* < \infty$ ) if and only if the set  $W$  is not empty.*

(ii) *If the optimal value of D-LLPP (4.1) is bounded, then*

$$(4.5) \quad \mu_* = G_*.$$

(iii) *The optimal value of D-LLPP (4.1) is unbounded (that is,  $\mu_* = \infty$ ) if and only if there exists a function  $\psi(\cdot) \in C^1$  such that*

$$(4.6) \quad \max_{(u, y) \in U \times Y} (\psi'(y))^T f(u, y) < 0.$$

*Proof.* The proof of Theorem 4.1 is given in section 9.  $\square$

Note that, in a stochastic setting, a duality result similar to Theorem 4.1(ii) (with  $Y = \mathbb{R}^m$ ) has been obtained in [16].

*Remark 4.2.* From the statements (i) and (ii) it follows that the set  $W$  and, hence, the set  $\bar{\Gamma}(S)$  (and also the set  $\Gamma(S)$  provided that  $f(U, y)$  is convex for all  $y \in Y$ ) are not empty (see Theorem 3.1(i)) if and only if a function  $\psi(\cdot) \in C^1$  satisfying (4.6) does not exist. Note that, if such a function  $\psi(\cdot)$  exists, then the fact that  $\bar{\Gamma}(S)$  (and, hence,  $\Gamma(S)$ ) are empty for  $S \geq S_0$  (for some  $S_0 > 0$ ) follow from the fact that this  $\psi(\cdot)$  can be used as a Liapunov function decreasing along the trajectories of the system (2.2) (respectively, (2.1)) and “forcing” them to leave  $Y$  in a finite time.

Assume that  $\mu_* < \infty$ . A function  $\psi_*(\cdot) \in C^1$  will be called a solution of D-LLPP (4.1) if

$$(4.7) \quad \mu_* = \min_{(u,y) \in U \times Y} \{g(u, y) + (\psi'_*(y))^T f(u, y)\}.$$

Defining  $H(p, y)$  by the equation

$$(4.8) \quad H(p, y) \stackrel{\text{def}}{=} \min_{u \in U} \{g(u, y) + p^T f(u, y)\}$$

and rewriting (4.7) in the form

$$(4.9) \quad \mu_* = \min_{y \in Y} H(\psi'_*(y), y) \quad \Rightarrow \quad \mu_* \leq H(\psi'_*(y), y) \quad \forall y \in Y,$$

one can come to a conclusion that a solution  $\psi_*(\cdot)$  of D-LLPP (4.1) is a smooth viscosity subsolution of the corresponding Hamilton–Jacobi–Bellman equation (see [12] and [26] for relevant definitions and developments).

Given a solution  $\psi_*(\cdot)$  of D-LLPP (4.1), define the set  $\Omega_* \subset U \times Y$  by the equation

$$(4.10) \quad \Omega_* \stackrel{\text{def}}{=} \{(u, y) \in U \times Y : g(u, y) + (\psi'_*(y))^T f(u, y) = \mu_*\}.$$

**PROPOSITION 4.3.** *Let  $\mu_* < \infty$  and  $\psi_*(\cdot)$  be a solution of D-LLPP (4.1). Then  $\gamma \in W$  will be a solution of the LLPP (3.2) if and only if*

$$(4.11) \quad \gamma(\Omega_*) = 1.$$

*Proof.* By Theorem 4.1(ii), a probability measure  $\gamma$  belonging to  $W$  will be a solution of the LLPP (3.2) if and only if

$$(4.12) \quad \mu_* = \int_{U \times Y} g(u, y) \gamma(du, dy) = \int_{U \times Y} (g(u, y) + (\psi'_*(y))^T f(u, y)) \gamma(du, dy).$$

Since (see (4.7))

$$(4.13) \quad \mu_* \leq g(u, y) + (\psi'_*(y))^T f(u, y) \quad \forall (u, y) \in U \times Y,$$

the equality (4.12) can be valid if and only if (4.11) is true. This proves the proposition.  $\square$

**COROLLARY 4.4.** *For a relaxed admissible pair  $(\nu(\cdot), y(\cdot))$  generating the occupational measure  $\gamma^{(\nu(\cdot), y(\cdot))}$  on  $[0, \infty)$  to be optimal in the problem (3.8) it is necessary and sufficient that  $\gamma^{(\nu(\cdot), y(\cdot))}(\Omega_*) = 1$ .*

*Proof.* By Corollary 3.4, the pair  $(\nu(\cdot), y(\cdot))$  will be optimal in (3.8) (that is, it satisfies (3.9)) if and only if  $\gamma^{(\nu(\cdot), y(\cdot))}$  is a solution of the LLPP (3.2). Hence, the statement is implied by Proposition 4.3.  $\square$

Let

$$(4.14) \quad \mathcal{Y}_* \stackrel{\text{def}}{=} \{y \in Y : H(\psi_*(y), y) = \mu_*\}$$

and

$$(4.15) \quad \mathcal{U}_*(y) \stackrel{\text{def}}{=} \{u \in U : g(u, y) + (\psi'_*(y))^T f(u, y) = H(\psi'_*(y), y)\}.$$

**COROLLARY 4.5.** *Let  $\psi_*(\cdot)$  be a solution of D-LLPP (4.1). A  $T$ -periodic admissible pair  $(u(\tau), y(\tau))$  will be a solution of the periodic optimization problem (3.13) if and, under the condition that the equalities (3.15) are valid, only if*

$$(4.16) \quad y(\tau) \in \mathcal{Y}_* \quad \forall \tau \in [0, T], \quad u(\tau) \in \mathcal{U}_*(y(\tau)) \quad \text{for almost all } \tau \in [0, T].$$

*Proof.* By Lemma 3.5, a  $T$ -periodic admissible pair  $(u(\tau), y(\tau))$  will be a solution of the periodic optimization problem (3.13) if and, under the condition that the equalities (3.15) are valid, only if the occupational measure  $\gamma^{(u(\cdot), y(\cdot))}$  generated by this pair is a solution of LLPP (3.2). According to Proposition 4.3, this is the case if and only if  $\gamma^{(u(\cdot), y(\cdot))}(\Omega_*) = 1$ , which is equivalent to  $(u(\tau), y(\tau)) \in \Omega_*$  for almost all  $\tau \in [0, T]$ . The latter, in turn, is equivalent to the inclusions (4.16) being satisfied for almost all  $\tau \in [0, T]$  since

$$(4.17) \quad \Omega_* = \{(u, y) : u \in \mathcal{U}_*(y), y \in \mathcal{Y}_*\}.$$

The fact that the first inclusion in (4.16) is valid for all  $\tau \in [0, T]$  follows from the fact that the function  $y(\cdot)$  is continuous and that the set  $\mathcal{Y}_*$  is compact.  $\square$

Note that a solution of D-LLPP (4.1) defined as a  $C^1$  function satisfying (4.9) may not exist, and one can consider the possibility of defining the solution as a nondifferentiable function, which satisfies (4.9) in the viscosity sense (see, e.g., [12, p. 399]). We, however, do not follow this path. Instead, in the next sections we discuss a way of constructing  $C^1$  functions that solve D-LLPP (4.1) approximately. First, in section 5, we approximate the LLPP (3.2) with I-D LLPs having a finite number of constraints  $N$  ( $N = 1, 2, \dots$ ), and we study the problem dual to the latter (section 5). Then, in section 6, we approximate the problem with  $N$  constraints and its dual by a family of F-D LPPs defined on grid points of  $U \times Y$  and the corresponding dual, and we show that, by solving these F-D LPPs, one can construct a family of functions  $\psi_{N,\Delta}(\cdot) \in C^1$  ( $\Delta$  being the parameter of the grid) such that, for any  $\delta > 0$ ,

$$(4.18) \quad \mu_* - \delta \leq \min_{(u,y) \in U \times Y} \{g(u, y) + (\psi'_{N,\Delta}(y))^T f(u, y)\} = \min_{y \in Y} H(\psi'_{N,\Delta}(y), y)$$

if  $N$  is large and  $\Delta$  is small enough (see (4.7) and (4.9)). We also show (see section 7) that, under some additional conditions, a feedback control function  $u_{N,\Delta}(\cdot) : Y \rightarrow U$  satisfying the inclusion

$$(4.19) \quad u_{N,\Delta}(y) \in \mathcal{U}_{N,\Delta}(y) \stackrel{\text{def}}{=} \{u \in U : g(u, y) + (\psi'_{N,\Delta}(y))^T f(u, y) = H(\psi'_{N,\Delta}(y), y)\}$$

can serve as an approximation to the optimal feedback control defined on the optimal periodic orbit (provided that the latter exists).



Note also that D-LLPP (4.1) can be rewritten in the form

$$(4.20) \quad \sup_{\psi(\cdot) \in C^1} \min_{y \in Y} H(\psi'(y), y) = \mu_*.$$

Problems similar to (4.20) (but in a different setting and under a different set of conditions) have been considered in the literature, particularly, in relation to finding so-called effective Hamiltonians arising in the homogenization theory (see [34] and references therein). It seems plausible that results that we discuss below could be applicable in this area, but we do not investigate this matter in the present paper.

Finally, to conclude this section, let us also mention that problems dual to I-D LPPs that arise in dealing with deterministic optimal control problems on finite time intervals and the way of how these dual problems can be used for formulating necessary and sufficient optimality conditions have been studied in [57] (see also references therein).

**5.  $N$ -approximating problem and its dual.** Let  $\{\phi_i(\cdot) \in C^1, i = 1, 2, \dots\}$  be a sequence of functions having continuous partial derivatives of second order such that any function  $\psi(\cdot) \in C^1$  and its gradient  $\psi'(\cdot)$  can be simultaneously approximated on  $Y$  by linear combinations of functions from  $\{\phi_i(\cdot), i = 1, 2, \dots\}$  and their corresponding gradients. That is, for any  $\psi(\cdot) \in C^1$  and any  $\delta > 0$ , there exist  $\beta_1, \dots, \beta_k$  (real numbers) such that

$$(5.1) \quad \max_{y \in Y} \left\{ \left| \psi(y) - \sum_1^k \beta_i \phi_i(y) \right| + \left\| \psi'(y) - \sum_1^k \beta_i \phi'_i(y) \right\| \right\} \leq \delta,$$

with  $\|\cdot\|$  being a norm in  $\mathbb{R}^m$ . An example of such an approximating sequence is the sequence of monomials  $y_1^{i_1} \dots y_m^{i_m}$ ,  $i_1, \dots, i_m = 0, 1, \dots$ , where  $y_j$  ( $j = 1, \dots, m$ ) stands for the  $j$ th component of  $y$  (see, e.g., [44]). In what follows it will be assumed that the gradients  $\phi'_i(y)$ ,  $i = 1, 2, \dots, N$ , are linearly independent on  $Y$ . That is,

$$(5.2) \quad \sum_{i=1}^N v_i \phi'_i(y) = 0 \quad \forall y \in Y \quad \Rightarrow \quad v_i = 0, \quad i = 1, 2, \dots, N.$$

Note that this is satisfied automatically if  $\phi_i(y)$  are chosen to be monomials and  $\text{int} Y$  (interior of  $Y$ ) is not empty.

Let us define the set  $W_N$  by the equation

$$(5.3) \quad W_N \stackrel{\text{def}}{=} \left\{ \gamma \mid \gamma \in \mathcal{P}(U \times Y); \int_{U \times Y} (\phi'_i(y))^T f(u, y) \gamma(du, dy) = 0, \quad i = 1, 2, \dots, N \right\}$$

and consider the linear programming problem

$$(5.4) \quad \min_{\gamma \in W_N} \int_{U \times Y} q(u, y) \gamma(du, dy) \stackrel{\text{def}}{=} G_N.$$

This problem will be referred to as  $N$ -approximating LLPP (or just  $N$ -LLPP). Note that  $W_N$  is a convex and compact subset of  $\mathcal{P}(U \times Y)$  and  $W \subset W_N$ , which implies

$$(5.5) \quad G_* \geq G_N.$$

The connection between LLPP (3.2) and  $N$ -LLPP (5.4) is established by the following proposition (see [32]).

PROPOSITION 5.1. *The set  $W$  is not empty if and only if there exists  $N_0 \geq 1$  such that  $W_N$  is not empty for  $N \geq N_0$ . If  $W$  is not empty, then*

$$(5.6) \quad \lim_{N \rightarrow \infty} \rho_H(W_N, W) = 0$$

and

$$(5.7) \quad \lim_{N \rightarrow \infty} G_N = G_*.$$

Also, if  $\gamma^N$  is a solution of the problem (5.4) and  $\lim_{N' \rightarrow \infty} \rho(\gamma^{N'}, \gamma) = 0$  for some subsequence of integers  $N'$  tending to infinity, then  $\gamma$  is a solution of (3.2). If the solution  $\gamma_*$  of the problem (3.2) is unique, then, for any solution  $\gamma^N$  of (5.4),  $\lim_{N \rightarrow \infty} \rho(\gamma^N, \gamma_*) = 0$ .

*Proof.* The proof follows from Proposition 7 in [32].  $\square$

Let us define the problem dual to  $N$ -LLPP (5.4) (referred in what follows to as  $D$ - $N$ -LLPP) by the equation

$$(5.8) \quad \sup_{(\mu, v) \in \mathcal{D}_N} \mu \stackrel{\text{def}}{=} \mu_N,$$

with the feasible set  $\mathcal{D}_N \subset \mathbb{R}^1 \times \mathbb{R}^N$  defined as

$$(5.9) \quad \mathcal{D}_N \stackrel{\text{def}}{=} \left\{ (\mu, v) : \mu = \min_{(u, y) \in U \times Y} \left\{ g(u, y) + \sum_{i=1}^N v_i (\phi'_i(y))^T f(u, y) \right\}, \quad v = (v_i) \in \mathbb{R}^N \right\}.$$

For a fixed  $N$ , the relationships between  $N$ -LLPP (5.4) and  $D$ - $N$ -LLPP (5.8) are similar to those between (3.2) and (4.1). For example, similarly to (4.3), one can obtain that, if  $W_N \neq \emptyset$  and  $\gamma \in W_N$ , then for any  $(\mu, v) \in \mathcal{D}_N$  ( $\mathcal{D}_N$  is never empty),

$$(5.10) \quad \mu \leq \int_{U \times Y} \left( g(u, y) + \sum_{i=1}^N v_i (\phi'_i(y))^T f(u, y) \right) \gamma(du, dy) = \int_{U \times Y} g(u, y) \gamma(du, dy)$$

$$(5.11) \quad \Rightarrow \quad \mu_N \leq G_N.$$

Also, the following result similar to Theorem 4.1 is valid.

THEOREM 5.2. (i) *The optimal value of  $D$ - $N$ -LLPP (5.8) is bounded (that is,  $\mu_N < \infty$ ) if and only if the set  $W_N$  is not empty.*

(ii) *If the optimal value of  $D$ - $N$ -LLPP (5.8) is bounded, then*

$$(5.12) \quad \mu_N = G_N.$$

(iii) *The optimal value of  $D$ - $N$ -LLPP (5.8) is unbounded (that is,  $\mu_N = \infty$ ) if and only if there exists  $v = (v_1, \dots, v_N)$  such that*

$$(5.13) \quad \max_{(u, y) \in U \times Y} (\psi'_v(y))^T f(u, y) < 0, \quad \psi_v(y) \stackrel{\text{def}}{=} \sum_{i=1}^N v_i \phi_i(y).$$

*Proof.* The proof of the theorem follows exactly the same steps as those in the proof of Theorem 4.1. For completeness we briefly outline these steps in section 9.  $\square$

A vector  $v = (v_i), i = 1, \dots, N$ , will be called a solution of D- $N$ -LLPP (5.8) if

$$(5.14) \quad \mu_N = \min_{(u,y) \in U \times Y} \{g(u,y) + (\psi'_v(y))^T f(u,y)\}, \quad \psi_v(y) \stackrel{\text{def}}{=} \sum_{i=1}^N v_i \phi_i(y).$$

Let us introduce the following two assumptions used in the consideration below.

*Assumption 1.* The inequality

$$(5.15) \quad (\psi'_v(y))^T f(u,y) \leq 0 \quad \forall (u,y) \in U \times Y, \quad \psi_v(y) \stackrel{\text{def}}{=} \sum_{i=1}^N v_i \phi_i(y),$$

can be valid only with  $v_i = 0, i = 1, \dots, N$ .

*Assumption 2.* The inequality

$$(5.16) \quad (\psi'(y))^T f(u,y) \leq 0 \quad \forall (u,y) \in U \times Y, \quad \psi(\cdot) \in C^1,$$

can be valid only with  $\psi'(y) = 0 \quad \forall y \in Y$ .

PROPOSITION 5.3. *Assumption 1 is equivalent to the assumption that*

$$(5.17) \quad 0 \in \text{int}(\text{co}K_N),$$

where  $\text{int}(\text{co}K_N)$  stands for the interior of the convex hull of  $K_N$ ,

$$(5.18) \quad K_N \stackrel{\text{def}}{=} \{z \in \mathbb{R}^N : z = (z_i), z_i = (\phi'_i(y))^T f(u,y), i = 1, \dots, N, (u,y) \in U \times Y\}.$$

*Proof.* The proof is in section 9.  $\square$

*Remark 5.4.* (i) It is easy to see that Assumption 2 implies the validity of Assumption 1 with any  $N = 1, 2, \dots$ . In fact, if the former is satisfied, then the inequality  $\sum_{i=1}^N v_i (\phi'_i(y))^T f(u,y) \leq 0 \quad \forall (u,y) \in U \times Y$  implies that  $\sum_{i=1}^N v_i \phi'_i(y) = 0 \quad \forall y \in Y$  and, hence, by (5.2),  $v_i = 0, i = 1, \dots, N$ . Sufficient conditions for Assumption 2 to be satisfied and special cases in which it is satisfied have been considered in [32] (see Proposition 8 and Remark 2 of [32]).

(ii) Assumption 1 implies that the function  $\psi_v(\cdot)$  satisfying (5.13) does not exist, and Assumption 2 implies that the function  $\psi(\cdot)$  satisfying (4.6) does not exist. Thus, from the former it follows that  $\mu_N$  is bounded (by Theorem 5.2(iii)) and from the latter that  $\mu_*$  is bounded (by Theorem 4.1(iii)).

THEOREM 5.5. (i) *If Assumption 1 is satisfied, then the set of solutions of D- $N$ -LLPP (5.8) is nonempty and bounded. That is (see (5.14)),*

$$(5.19) \quad \emptyset \neq V_N \stackrel{\text{def}}{=} \left\{ v = (v_i) : \mu_N = \min_{(u,y) \in U \times Y} \left\{ g(u,y) + \left( \sum_{i=1}^N v_i \phi'_i(y) \right)^T f(u,y) \right\} \right\},$$

$$(5.20) \quad \max_{v \in V_N} \|v\| \leq c_N = \text{const.}$$

(ii) *Conversely, if (5.19) and (5.20) are true, then Assumption 1 is satisfied.*

*Proof.* Proof is in section 9.  $\square$

Let us now establish that the optimal values of the D- $N$ -LLPP (5.8) converge to the optimal value of D-LLPP (4.1) as  $N \rightarrow \infty$ . Note that, as follows directly from the definitions of D- $N$ -LLPP and D-LLPP (see (5.8) and (4.1)),

$$(5.21) \quad \mu_1 \leq \mu_2 \leq \dots \leq \mu_N \leq \dots \leq \mu_*.$$

THEOREM 5.6. *The optimal value of D-N-LLPP (5.8) converges to the optimal value of D-LLPP (4.1). That is,*

$$(5.22) \quad \lim_{N \rightarrow \infty} \mu_N = \mu_*,$$

the statement being valid for both  $\mu_* < \infty$  and  $\mu_* = \infty$ .

*Proof.* If Assumption 2 is satisfied, then  $\mu_* = G_*$  and  $\mu_N = G_N$  for  $N = 1, 2, \dots$ . Hence, (5.22) follows from (5.7). The proof of the general case is in section 9.  $\square$

In conclusion of this section, let us introduce a regularity condition, which will be used in section 7 below. Let Assumption 2 be satisfied and let  $v = (v_i) \in V_N$ ,  $\psi_v(y) \stackrel{\text{def}}{=} \sum_{i=1}^N v_i \phi_i(y)$ . Define the sets  $\Omega_N^v, \mathcal{Y}_N^v$  by the equations

$$(5.23) \quad \Omega_N^v \stackrel{\text{def}}{=} \{(u, y) \in U \times Y : g(u, y) + (\psi'_v(y))^T f(u, y) = \mu_N\},$$

$$(5.24) \quad \mathcal{Y}_N^v \stackrel{\text{def}}{=} \{y : (u, y) \in \Omega_N^v\} = \{y \in Y : H(\psi'_v(y), y) = \mu_N\}$$

and the set  $\mathcal{U}_N^v(y)$  by the equation

$$(5.25) \quad \mathcal{U}_N^v(y) \stackrel{\text{def}}{=} \{u \in U : g(u, y) + (\psi'_v(y))^T f(u, y) = H(\psi'_v(y), y)\},$$

where  $H(p, y)$  is as in (4.8).

*N-regularity condition (N-RC) on Q.* Given  $Q \subset Y$ , we shall say that N-RC is satisfied on  $Q$  if, for any  $v \in V_N$  and any  $y \in \mathcal{Y}_N^v \cap Q$ , a solution of the problem

$$(5.26) \quad \max_{u \in U} \{g(u, y) + (\psi'_v(y))^T f(u, y)\}$$

is unique. That is,  $\mathcal{U}_N^v(y)$  is a singleton for  $v \in V_N$  and  $y \in \mathcal{Y}_N^v \cap Q$ .

Note that N-RC is satisfied on any  $Q \subset Y$  and with any  $N = 1, 2, \dots$  if  $g(u, y)$  is strictly convex and  $f(u, y)$  is linear in  $u$  and if  $U$  is a convex set.

**6. N $\Delta$ -approximating problem and its dual.** Assume first that  $N$  is fixed and that, for any  $\Delta > 0$ , Borel sets  $Q_{l,k}^\Delta \subset U \times Y$  ( $l = 1, \dots, L^\Delta$ ,  $k = 1, \dots, K^\Delta$ ) (called cells) are defined in such a way that two different cells do not intersect, the union of all cells is equal to  $U \times Y$ , and

$$(6.1) \quad \sup_{(u,y) \in Q_{l,k}^\Delta} \|(u, y) - (u_l, y_k)\| \leq c\Delta, \quad c = \text{const},$$

for some point  $(u_l, y_k) \in Q_{l,k}^\Delta$ . For simplicity of notation, it is assumed (from now on) that  $U$  is a compact subset of  $\mathbb{R}^n$  and  $\|\cdot\|$  stands for a norm in  $\mathbb{R}^{n+m}$ . Let us fix the points  $(u_l, y_k)$  ( $l = 1, \dots, L^\Delta$ ,  $k = 1, \dots, K^\Delta$ ) and define a polyhedral set  $W_{N,\Delta} \subset \mathbb{R}^{L^\Delta + K^\Delta}$  by the equation

$$(6.2) \quad W_{N,\Delta} \stackrel{\text{def}}{=} \left\{ \gamma = \{\gamma_{l,k}\} \geq 0 : \sum_{l,k} \gamma_{l,k} = 1, \sum_{l,k} (\phi'_i(y_k))^T f(u_l, y_k) \gamma_{l,k} = 0, i = 1, \dots, N \right\},$$

where  $\sum_{l,k} \stackrel{\text{def}}{=} \sum_{l=1}^{L^\Delta} \sum_{k=1}^{K^\Delta}$ . Consider the problem

$$(6.3) \quad \min_{\gamma \in W_{N,\Delta}} \sum_{l,k} \gamma_{l,k} g(u_l, y_k) \stackrel{\text{def}}{=} G_{N,\Delta}.$$

This is an F-D LLP which will be referred to as  $N\Delta$ -approximating LLPP (or  $N\Delta$ -LLPP).

Note that the polyhedral set  $W_{N,\Delta}$  is the set of probability measures on  $U \times Y$  which assign nonzero probabilities only to the points  $(u_l, y_k)$ , and, as such,

$$(6.4) \quad W_{N,\Delta} \subset W_N \quad \Rightarrow \quad G_{N,\Delta} \geq G_N.$$

**THEOREM 6.1.** *Let Assumption 1 be satisfied. Then the set  $W_N$  is not empty and there exists  $\Delta_0 > 0$  such that  $W_{N,\Delta}$  is not empty for  $\Delta \leq \Delta_0$ . Also*

$$(6.5) \quad \lim_{\Delta \rightarrow 0} \rho_H(W_{N,\Delta}, W_N) = 0$$

and

$$(6.6) \quad \lim_{\Delta \rightarrow 0} G_{N,\Delta} = G_N.$$

If  $\gamma^{N,\Delta}$  is a solution of the problem (6.3) and  $\lim_{\Delta' \rightarrow 0} \rho(\gamma^{N,\Delta'}, \gamma^N) = 0$  for some sequence of  $\Delta'$  tending to zero, then  $\gamma^N$  is a solution of (5.4). If the solution  $\gamma^N$  of the problem (5.4) is unique, then, for any solution  $\gamma^{N,\Delta}$  of (6.3),  $\lim_{\Delta \rightarrow 0} \rho(\gamma^{N,\Delta}, \gamma^N) = 0$ .

*Proof.* The fact that  $W_N$  is not empty if Assumption 1 is satisfied follows from Theorem 5.2(i) and from the fact that  $\mu_N$  is bounded (see Remark 5.4(ii)). The proofs of all other statements of the theorem follow from Proposition 9 in [32].  $\square$

Consider the F-D LPP

$$(6.7) \quad \max_{(\mu, \lambda) \in \mathbb{R}^1 \times \mathbb{R}^N} \left\{ \mu : \mu \leq g(u_l, y_k) + \sum_{i=1}^N \lambda_i (\phi'_i(y_k))^T f(u_l, y_k) \quad \forall (u_l, y_k) \right\} \stackrel{\text{def}}{=} \mu_{N,\Delta},$$

which is dual to  $N\Delta$ -LLPP (6.3) and which will be referred to as D- $N\Delta$ -LLPP. From the duality theory for F-D LPPs (see, e.g., [22]) it follows, in particular, that, if  $W_{N,\Delta}$  is not empty, then the optimal value of  $N\Delta$ -LLPP (6.3) is equal to the optimal value of D- $N\Delta$ -LLPP (6.7)

$$(6.8) \quad G_{N,\Delta} = \mu_{N,\Delta}$$

and the solutions set of D- $N\Delta$ -LLPP (6.7) is not empty:

$$(6.9) \quad \emptyset \neq \Lambda_{N,\Delta} \stackrel{\text{def}}{=} \left\{ \lambda = (\lambda_i) : \mu_{N,\Delta} = \min_{(u_l, y_k)} \left\{ g(u_l, y_k) + \sum_{i=1}^N \lambda_i (\phi'_i(y_k))^T f(u_l, y_k) \right\} \right\}.$$

**THEOREM 6.2.** *Let Assumption 1 be satisfied. Then*

$$(6.10) \quad \lim_{\Delta \rightarrow 0} \max_{\lambda \in \Lambda_{N,\Delta}} \text{dist}(\lambda, V_N) = 0, \quad \text{dist}(\lambda, V_N) \stackrel{\text{def}}{=} \min_{v \in V_N} \|\lambda - v\|,$$

where  $V_N$  is the solutions set of D- $N$ -LLPP (5.8) (explicitly defined in (5.19)).

*Proof.* The proof is in section 10. Let us only note here that, by Theorem 6.1, the set  $W_{N,\Delta}$  is not empty under Assumption 1, and hence, (6.8) and (6.9) are valid for  $\Delta$  small enough.  $\square$

Let us now address the issue of constructing  $\psi_{N,\Delta}(\cdot)$  that approximately solves D-LLPP (4.1) (see the end of section 4).

PROPOSITION 6.3. *Let Assumption 2 be satisfied. Then, for any  $\delta > 0$ , there exist  $N_\delta > 0$  and  $\Delta_N > 0$  such that the function  $\psi_{N,\Delta}(\cdot)$  defined by the equation*

$$(6.11) \quad \psi_{N,\Delta}(y) \stackrel{\text{def}}{=} \sum_{i=1}^N \lambda_i^{N,\Delta} \phi_i(y), \quad \lambda^{N,\Delta} = (\lambda_i^{N,\Delta}) \in \Lambda_{N,\Delta},$$

satisfies (4.18) with  $N \geq N_\delta$  and  $\Delta \leq \Delta_N$ .

*Proof.* As was mentioned above (see Remark 5.4(i)), the validity of Assumption 2 implies the validity of Assumption 1 with all  $N$ . Having this in mind, let us choose  $N_\delta$  in such a way that

$$(6.12) \quad \mu_* - \frac{\delta}{2} \leq \mu_N$$

for any  $N \geq N_\delta$  (this is possible due to Theorem 5.6). By Theorem 5.5(i), the set  $V_N$  is not empty and

$$(6.13) \quad \mu_* - \frac{\delta}{2} \leq \min_{(u,y) \in U \times Y} \left\{ g(u,y) + \left( \sum_{i=1}^N v_i \phi'_i(y) \right)^T f(u,y) \right\} \quad \forall v = (v_i) \in V_N.$$

From (6.10) it follows that, for any  $\Delta \leq \Delta_N$  ( $\Delta_N$  being positive small enough) and any  $\lambda \in \Lambda_{N,\Delta}$ , there exists  $v^{N,\Delta} = (v_i^{N,\Delta}) \in V_N$  such that

$$\begin{aligned} & \min_{(u,y) \in U \times Y} \left\{ g(u,y) + \left( \sum_{i=1}^N v_i^{N,\Delta} \phi'_i(y) \right)^T f(u,y) \right\} - \frac{\delta}{2} \\ & \leq \min_{(u,y) \in U \times Y} \left\{ g(u,y) + \left( \sum_{i=1}^N \lambda_i^{N,\Delta} \phi'_i(y) \right)^T f(u,y) \right\} \\ \Rightarrow \quad & \mu_* - \delta \leq \min_{(u,y) \in U \times Y} \left\{ g(u,y) + \left( \sum_{i=1}^N \lambda_i^{N,\Delta} \phi'_i(y) \right)^T f(u,y) \right\}. \end{aligned}$$

The latter is (4.18) with  $\psi_{N,\Delta}(\cdot)$  as in (6.11).  $\square$

**7. Convergence to the optimal periodic solution.** In this section we will show that, under certain additional conditions, the control satisfying (4.19), that is, defined as a solution of the problem

$$(7.1) \quad \min_{u \in U} \{ g(u,y) + (\psi'_{N,\Delta}(y))^T f(u,y) \},$$

where  $\psi_{N,\Delta}(\cdot)$  is as in (6.11), converges (as  $N \rightarrow \infty$  and  $\Delta \rightarrow 0$ ) to the optimal feedback control defined on the optimal periodic orbit. To ensure the existence of the latter, throughout this section it is assumed that there exists a solution  $\gamma_*$  of LLPP (3.2) that is generated by a  $T$ -periodic admissible pair  $(u_*(\cdot), y_*(\cdot))$  (note that this assumption implies that the pair is a solution of the periodic optimization problem (3.13); see Lemma 3.5).

Define sets  $\Theta_*$  and  $\mathcal{Y}_*$  by the equations

$$(7.2) \quad \Theta_* \stackrel{\text{def}}{=} \{ (u,y) : (u,y) = (u_*(\tau), y_*(\tau)) \text{ for some } \tau \in [0, T] \},$$

$$(7.3) \quad \mathcal{Y}_* \stackrel{\text{def}}{=} \{y : y = y_*(\tau) \text{ for some } \tau \in [0, T]\} = \{y : (u, y) \in \Theta_*\}.$$

The set  $\Theta_*$  can be considered as the graph of the optimal feedback control function, which is defined on the optimal orbit  $\mathcal{Y}_*$  by the equation  $u_*(y) \stackrel{\text{def}}{=} u \forall (u, y) \in \Theta_*$ . For this definition to make sense, it is assumed that from the fact that  $(u', y) \in \Theta_*$  and  $(u'', y) \in \Theta_*$  it follows that  $u' = u''$  (this assumption is satisfied if the closed curve defined by  $y_*(\tau), \tau \in [0, T]$ , does not intersect itself).

Note that the set  $\mathcal{Y}_*$  defined in (7.3) is different from the set defined in (4.14) (despite the fact that the same notation is used in both cases). Underline that in this section it is not assumed that a solution  $\psi_*(\cdot)$  of D-LLPP (4.1) (which is used in (4.14)) exists. If, however, it does exist, then by Corollary 4.5, the set defined in (7.3) is contained in the set defined in (4.14).

Let us introduce the following assumption about the set  $\Theta_*$ .

*Assumption 3.* For any  $(u, y) \in \bar{\Theta}_*$  (the closure of  $\Theta_*$ ) and any  $r > 0$ , the set  $B_r(u, y) \stackrel{\text{def}}{=} ((u, y) + rB) \cap (U \times Y)$  ( $B$  being the open unit ball in  $\mathbb{R}^{n+m}$ ) has a nonzero  $\gamma_*$ -measure:  $\gamma_*(B_r(u, y)) > 0$ .

Note that this assumption is satisfied if the optimal control function  $u_*(\cdot) : [0, T] \rightarrow U$  is piecewise continuous and at every discontinuity point  $\tau$  the value of  $u^*(\tau)$  is equal either to the limit from the left ( $u_*(\tau) = \lim_{\tau' \rightarrow \tau-} u_*(\tau')$ ) or to the limit from the right ( $u_*(\tau) = \lim_{\tau' \rightarrow \tau+} u_*(\tau')$ ).

Under the validity of Assumption 2, from Proposition 5.1 and Theorem 6.1 it follows that there exist solutions  $\gamma^{N, \Delta} = \{\gamma_{l,k}^{N, \Delta}\}$  of the  $N\Delta$ -LLPP (6.3) (considered with  $N \rightarrow \infty$  and  $\Delta \rightarrow 0$ ) such that

$$(7.4) \quad \lim_{N \rightarrow \infty} \lim_{\Delta \rightarrow 0} \rho(\gamma^{N, \Delta}, \gamma_*) = 0,$$

with the latter being valid for any solution of the  $N\Delta$ -LLPP (6.3) if  $\gamma_*$  is a unique solution of LLPP (3.2).

For any Borel subset  $\Theta$  of  $U \times Y$ , let  $\gamma^{N, \Delta}(\Theta)$  stand for the  $\gamma^{N, \Delta}$  measure of  $\Theta$ . That is,  $\gamma^{N, \Delta}(\Theta) \stackrel{\text{def}}{=} \sum_{(u_l, y_k) \in \Theta} \gamma_{l,k}^{N, \Delta}$ .

**THEOREM 7.1.** *Let Assumptions 2 and 3 be satisfied.*

(i) *Let  $Q$  be an open subset of  $Y$  such that the  $N$ -regularity condition (see the end of section 5) is satisfied on  $Q$  for all  $N$  large enough and such that the solution  $u_{N, \Delta}(\cdot)$  of the problem (7.1) is unique on  $Q$ , and it is uniformly equicontinuous on any closed subset  $\bar{Q}'$  of  $Q$  for  $N$  large and  $\Delta$  small enough. That is,*

$$(7.5) \quad \|u_{N, \Delta}(y') - u_{N, \Delta}(y'')\| \leq \omega(\|y' - y''\|) \quad \forall y', y'' \in \bar{Q}',$$

*where  $\omega(\theta)$  tends to zero as  $\theta$  tends to zero and  $N \geq N_0$ ,  $\Delta \leq \Delta_N$  ( $\omega(\cdot)$  and  $N_0$  may be different for different  $\bar{Q}' \subset Q$ ). Then*

$$(7.6) \quad \lim_{N \rightarrow \infty} \overline{\lim}_{\Delta \rightarrow 0} \|u_{N, \Delta}(y) - u_*(y)\| = 0 \quad \forall y \in \mathcal{Y}_* \cap Q,$$

*with the convergence being uniform on any closed subset of  $\mathcal{Y}_* \cap Q$ .*

(ii) *Assume, in addition, that*

$$(7.7) \quad \overline{\lim}_{N \rightarrow \infty} \overline{\lim}_{\Delta \rightarrow 0} \gamma^{N, \Delta}(\bar{Q}) = 1,$$

*where  $\bar{Q}$  is the closure of  $Q$  and that  $\gamma_*(\bar{Q}/Q) = 0$ . Then  $\gamma_*(\mathcal{Y}_* \cap Q) = 1$ . That is,  $u_{N, \Delta}(y)$  converges to  $u_*(y)$  for  $\gamma_*$ -almost all  $y \in Y$ .*

The second statement of the theorem allows an extension, which we state in the form of the following corollary.

**COROLLARY 7.2.** *Let Assumptions 2 and 3 be satisfied, and let  $Q_j, j = 1, \dots, J$ , be nonintersecting open subsets of  $Y$ , such that, for every  $j$ , the  $N$ -regularity condition is satisfied on  $Q_j$ , and (7.5) is valid on any closed  $\bar{Q}' \subset Q_j$ . Assume also that*

$$(7.8) \quad \overline{\lim}_{N \rightarrow \infty} \overline{\lim}_{\Delta \rightarrow 0} \gamma^{N, \Delta}(\overline{\cup_j Q_j}) = 1$$

and that

$$(7.9) \quad \gamma_*(\overline{\cup_j Q_j} / \cup_j Q_j) = 0,$$

where  $\overline{\cup_j Q_j}$  is the closure of  $\cup_j Q_j$ . Then

$$(7.10) \quad \lim_{N \rightarrow \infty} \overline{\lim}_{\Delta \rightarrow 0} \|u_{N, \Delta}(y) - u_*(y)\| = 0$$

for  $\gamma_*$ -almost all  $y \in Y$ .

*Proof.* The proofs of Theorem 7.1 and Corollary 7.2 are given in section 10. They are based on a result of [32] establishing that the sets

$$(7.11) \quad \Theta_{N, \Delta} \stackrel{\text{def}}{=} \{(u_l, y_k) : \gamma_{l, k}^{N, \Delta} > 0\},$$

$$(7.12) \quad \mathcal{Y}_{N, \Delta} \stackrel{\text{def}}{=} \{y : (u, y) \in \Theta_{N, \Delta}\} = \left\{ y_k : \sum_l \gamma_{l, k}^{N, \Delta} > 0 \right\}$$

are “approaching” the sets  $\Theta_*$  and  $\mathcal{Y}_*$  as  $N \rightarrow \infty$  and  $\Delta \rightarrow 0$  (see Theorem 7.3 below).  $\square$

**THEOREM 7.3.** *Let Assumption 2 be satisfied, (7.4) be valid, and  $\gamma_*, \gamma^{N, \Delta}$  be as above. Then the following hold.*

(i) *Corresponding to an arbitrary small  $r > 0$  and arbitrary small  $\delta > 0$ , there exists  $N_0$  such that, for  $N \geq N_0$  and  $\Delta \leq \Delta_N$ ,*

$$(7.13) \quad \gamma^{N, \Delta}(\Theta_{N, \Delta} / (\Theta_* + rB)) < \delta, \quad \gamma^{N, \Delta}(\mathcal{Y}_{N, \Delta} / (\mathcal{Y}_* + rD)) < \delta$$

and

$$(7.14) \quad \Theta_{N, \Delta, \delta} \subset \Theta_* + rB, \quad \mathcal{Y}_{N, \Delta, \delta} \subset \mathcal{Y}_* + rD,$$

where  $B$  and  $D$  are open unit balls in  $\mathbb{R}^{n+m}$  and  $\mathbb{R}^m$ , respectively, and

$$\Theta_{N, \Delta, \delta} \stackrel{\text{def}}{=} \{(u_l, y_k) : \gamma_{l, k}^{N, \Delta} \geq \delta\}, \quad \mathcal{Y}_{N, \Delta, \delta} \stackrel{\text{def}}{=} \left\{ y_k : \sum_l \gamma_{l, k}^{N, \Delta} \geq \delta \right\}.$$

(ii) *If, in addition, Assumption 3 is satisfied, then, corresponding to an arbitrary small  $r > 0$ , there exists  $N_0$  such that, for  $N \geq N_0$  and  $\Delta \leq \Delta_N$ ,*

$$(7.15) \quad \Theta_* \subset \Theta_{N, \Delta} + rB, \quad \mathcal{Y}_* \subset \mathcal{Y}_{N, \Delta} + rB.$$

*Proof.* The validity of (i) and (ii) is implied by Propositions 10 and 11 from [32].  $\square$



Note that from (7.4) it follows that  $\gamma^{N,\Delta} \stackrel{\text{def}}{=} \{\gamma_{l,k}^{N,\Delta}\}$  can be considered as an “approximation” of  $\gamma_*$  for  $N$  large and  $\Delta$  small enough. Due to the fact that  $\gamma_*$  is the occupational measure generated by the optimal periodic pair  $(u^*(\cdot), y^*(\cdot))$ , an element  $\gamma_{l,k}^{N,\Delta}$  of  $\gamma^{N,\Delta}$  can be interpreted as an estimate of “proportion” of time spent by the optimal pair in a “small” vicinity of the point  $(u_l, y_k)$ , while the fact that  $\gamma_{l,k}^{N,\Delta}$  is positive or zero can be interpreted as an indication of whether or not the optimal pair attends this vicinity. Theorem 7.3 can be viewed as a justification of the latter interpretation.

Based on the consideration above, one can propose the following steps to construct an approximate solution to the periodic optimization problem.

(1) Find the optimal value  $G_{N,\Delta}$  and a solution  $\gamma^{N,\Delta}$  of  $N\Delta$ -LLPP (6.3), and also find a solution  $\mu_{N,\Delta}, \lambda^{N,\Delta}$  of problem (6.7) dual to  $N\Delta$ -LLPP (6.3).

(2) Define the sets  $\Theta_{N,\Delta}$  and  $\mathcal{Y}_{N,\Delta}$  according to (7.11) and (7.12).

(3) Construct the function  $\psi_{N,\Delta}(y)$  according to (6.11) and find the control  $u_{N,\Delta}(y)$  by solving the problem (7.1) for every  $y$  in a neighborhood of  $\mathcal{Y}_{N,\Delta}$ .

(4) Integrate the system (2.1) starting from an initial point  $y(0) \in \mathcal{Y}_{N,\Delta}$  and using  $u_{N,\Delta}(y)$  as the feedback control. One can expect that the obtained solution of the system will return to a small vicinity of the starting point  $y(0)$  and it will be possible to identify the end point of the integration period,  $T_{N,\Delta}$ , as the moment of entering this vicinity.

(5) Adjust the initial condition and/or control to obtain a periodic admissible pair  $(u_{N,\Delta}(\tau), y_{N,\Delta}(\tau))$  defined on the interval  $[0, T_{N,\Delta}]$ . Calculate the integral

$$\frac{1}{T_{N,\Delta}} \int_0^{T_{N,\Delta}} g(u_{N,\Delta}(\tau), y_{N,\Delta}(\tau)) d\tau$$

and compare it with  $G_{N,\Delta}$ . If the value of the integral proves to be close to  $G_{N,\Delta}$ , then the constructed admissible pair is a “good” approximation to the solution of the periodic optimization problem.

Note that these steps are similar to the procedure described in [32]. The difference is that the approximation  $u_{N,\Delta}(\cdot)$  to the optimal control is determined as a solution of the problem (7.1). The objective function of the latter is constructed with the help of a solution of the problem (6.7) dual to  $N\Delta$ -LLPP (6.3) (the opportunity of using dual solutions for finding an approximation to the optimal control was not discussed in [32]). In the next section we consider a numerical example to illustrate the above steps.

Let us conclude this section by noting that most of the existing computational methods for periodic optimization problems are either aimed at solving the system of necessary optimality conditions (so-called dynamic optimization schemes; see, e.g., [45], [53], and references therein) or based on approximating the POP with F-D nonlinear mathematical programming problems. The latter are defined either via a discretization of the time interval (as in direct optimization schemes; see, e.g., [45]) or with the help of special parametrization techniques (as in the flatness-based algorithm of [54]). In all instances the time interval (period) is assumed to be finite, with its length being one of the optimization parameters.

In contrast to these methods, the approach of this paper is based on the fact that, if the optimal periodic regime exists, then, under mild conditions, it is a solution of the long run average optimal control problem, and, also, it generates the occupational measure that is a solution of the LLPP (3.2); the latter does not involve the time parameter at all. As has been noticed in section 4, finding a solution of the problem

dual to LLPP (3.2) is equivalent to finding a subsolution of the Hamilton–Jacobi–Bellman equation for the corresponding long run average problem of optimal control. Thus, the approach under consideration can be characterized as one belonging to the class of algorithms that solve Hamilton–Jacobi–Bellman equations for infinite time horizon optimal control problems via a discretization of the state and control spaces (see [25], [26], [43], and references therein). The obvious advantage of the approach is in its simplicity (no software except standard linear programming and ordinary differential equation solvers is needed). Its disadvantage is in computer memory requirements (especially in problems of higher dimensions). A comparison of the linear programming approach with other algorithms of the given class will be a matter of further research.

**8. A numerical example.** Let  $m = 2$ ,  $n = 1$ , and let

$$y(t) = (y_1(t), y_2(t)) \subset Y \stackrel{\text{def}}{=} [-0.25, 0.25] \times [1.3, 1.9], \quad u(t) \in U \stackrel{\text{def}}{=} [-0.4, 0.4].$$

Also, let  $f(\cdot)$  and  $g(\cdot)$  be defined by the equations

$$(8.1) \quad f^T(u, y_1, y_2) \stackrel{\text{def}}{=} (y_2 - 1.6, u),$$

$$(8.2) \quad g(u, y_1, y_2) \stackrel{\text{def}}{=} ((y_2 - 3)(y_2 - 1)^2 + 2.5)(y_2 - 1) + 3 + 1.126y_1^2 + 0.4u^2.$$

The periodic optimization problem (3.13) with this data was considered and numerically solved using the “flatness-based algorithm” in [54]. Below we present a numerical solution of the problem obtained by following the steps outlined in section 7. Define the grid of  $U \times Y$  by the equations

$$u_i^\Delta \stackrel{\text{def}}{=} i\Delta - 0.4, \quad y_{1,j}^\Delta \stackrel{\text{def}}{=} j\Delta - 0.25, \quad y_{2,k}^\Delta \stackrel{\text{def}}{=} k\Delta + 1.3,$$

where  $i = 0, 1, \dots, \frac{0.8}{\Delta}$ ,  $j = 0, 1, \dots, \frac{0.5}{\Delta}$ , and  $k = 0, 1, \dots, \frac{0.6}{\Delta}$  ( $\Delta$  is chosen in such a way that  $\frac{0.8}{\Delta}$ ,  $\frac{0.5}{\Delta}$ , and  $\frac{0.6}{\Delta}$  are integers). In this case, the  $N\Delta$ -approximating LLPP (6.3) can be written in the form

$$(8.3) \quad G_{N,\Delta} \stackrel{\text{def}}{=} \min_{\gamma \in W_{N,\Delta}} \sum_{i,j,k} g(u_i^\Delta, y_{1,j}^\Delta, y_{2,k}^\Delta) \gamma_{i,j,k},$$

with  $W_{N,\Delta}$  being the polyhedral set defined by the equation

$$(8.4) \quad \left. \begin{aligned} W_{N,\Delta} \stackrel{\text{def}}{=} \left\{ \gamma = \{\gamma_{i,j,k}\} \geq 0 : \sum_{i,j,k} \gamma_{i,j,k} = 1, \right. \\ \left. \sum_{i,j,k} (\phi'_{i_1,i_2}(y_{1,j}^\Delta, y_{2,k}^\Delta))^T f(u_i^\Delta, y_{1,j}^\Delta, y_{2,k}^\Delta) \gamma_{i,j,k} = 0, (i_1, i_2) \in I_N \right\}. \end{aligned} \right\}$$

Here,  $\phi_{i_1,i_2}(y_1, y_2) \stackrel{\text{def}}{=} y_1^{i_1} y_2^{i_2}$  and  $I_N$  is the set of multi-indices

$$I_N \stackrel{\text{def}}{=} \{i : i = (i_1, i_2), \quad i_1, i_2 = 0, 1, \dots, N, \quad i_1 + i_2 \geq 1\}.$$

Note that  $g(\cdot)$  in (8.3) and  $f(\cdot)$  in (8.4) are as in (8.2) and (8.1), respectively.

The problem (8.3) was solved using the CPLEX linear programming solver [61] for different values of  $N$  and  $\Delta$ . The obtained optimal values  $G_{N,\Delta}$  are summarized in the following table:

$\Delta$	$N$		
	4	5	6
0.05	4.1959526	4.1959709	4.1959726
0.025	4.1958774	4.1958808	4.1958807
0.0125	4.1958750	4.1958751	4.1958751

On the basis of this data, one may conclude that  $G_6 = \lim_{\Delta \rightarrow 0} G_{6,\Delta} \approx 4.1958751 \approx 4.196$ , the latter coinciding with the optimal value of the given periodic optimization problem obtained in [54]. Thus, if for some  $T$ -periodic admissible pair  $(u(\tau), y(\tau))$ ,

$$(8.5) \quad \frac{1}{T} \int_0^T \left[ ((y_2(\tau) - 3)(y_2(\tau) - 1)^2 + 2.5)(y_2(\tau) - 1) + 3 + 1.126y_1(\tau)^2 + 0.4u(\tau)^2 \right] d\tau \approx 4.196,$$

then this pair is an approximate solution of the periodic optimization problem under consideration.

Let  $\gamma^{N,\Delta} = \{\gamma_{i,j,k}^{N,\Delta}\}$  stand for the solution of (8.3), and let  $\lambda^{N,\Delta} = \{\lambda_{i_1,i_2}^{N,\Delta}\}$  stand for the solution of the problem dual to (8.3)

$$(8.6) \quad \max_{(\mu,\lambda)} \left\{ \mu : \mu \leq g(u_i^\Delta, y_{1,j}^\Delta, y_{2,k}^\Delta) - \sum_{(i_1,i_2) \in I_N} \lambda_{i_1,i_2} (\phi'_{i_1,i_2}(y_{1,j}^\Delta, y_{2,k}^\Delta))^T f(u_i^\Delta, y_{1,j}^\Delta, y_{2,k}^\Delta) \quad \forall (u_i^\Delta, y_{1,j}^\Delta, y_{2,k}^\Delta) \right\}$$

(see (6.7) for comparison). Define the sets  $\Theta_{N,\Delta}$  and  $\mathcal{Y}_{N,\Delta}$  by the equations

$$(8.7) \quad \Theta_{N,\Delta} \stackrel{\text{def}}{=} \left\{ (u_i^\Delta, y_{1,j}^\Delta, y_{2,k}^\Delta) : \gamma_{i,j,k}^{N,\Delta} \neq 0 \right\},$$

$$(8.8) \quad \mathcal{Y}_{N,\Delta} \stackrel{\text{def}}{=} \left\{ (y_{1,j}^\Delta, y_{2,k}^\Delta) : \sum_i \gamma_{i,j,k}^{N,\Delta} \neq 0 \right\}$$

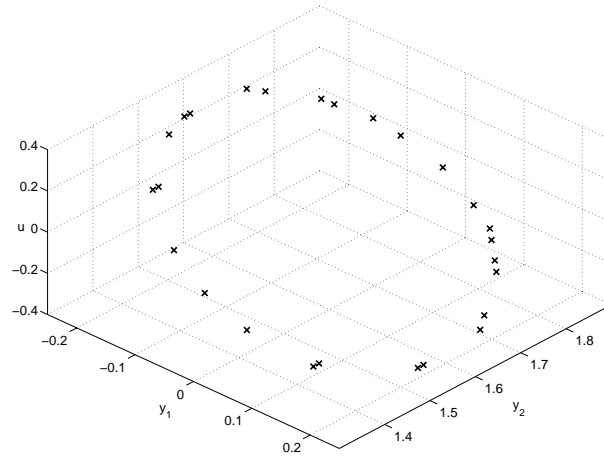
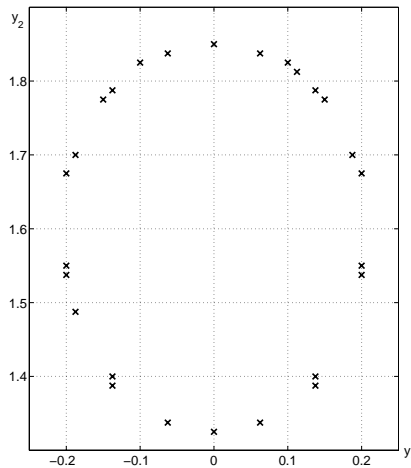
(see (7.11) and (7.12) for comparison). For  $N = 6$  and  $\Delta = 0.0125$ , the graphs of  $\Theta_{N,\Delta}$  and  $\mathcal{Y}_{N,\Delta}$  are depicted in Figures 1 and 2 below (the points belonging to these sets are marked). As follows from Theorem 7.3, these graphs can serve as approximations, respectively, to the graph of the optimal periodic control  $\Theta_*$  and to the optimal periodic orbit  $\mathcal{Y}_*$ .

Define the function  $\psi_{N,\Delta}(y_1, y_2)$  by the equation (see (6.11) for comparison)

$$(8.9) \quad \psi_{N,\Delta}(y_1, y_2) \stackrel{\text{def}}{=} \sum_{(i_1,i_2) \in I_N} \lambda_{i_1,i_2}^{N,\Delta} \phi_{i_1,i_2}(y_1, y_2).$$

The problem (7.1) is in this case of the form

$$(8.10) \quad \min_{u \in [-0.4, 0.4]} \Phi_{N,\Delta}(u, y_1, y_2),$$


 FIG. 1.  $\Theta_{N,\Delta}$ . Approximation to the graph of the optimal periodic control  $\Theta_*$ .

 FIG. 2.  $\mathcal{Y}_{N,\Delta}$ . Approximation to the optimal periodic orbit  $\mathcal{Y}_*$ .

where

$$\begin{aligned}
 \Phi_{N,\Delta}(u, y_1, y_2) = & \left[ ((y_2 - 3)(y_2 - 1)^2 + 2.5)(y_2 - 1) + 3 + 1.126y_1^2 + 0.4u^2 \right. \\
 (8.11) \quad & \left. + \left( \frac{\partial \psi_{N,\Delta}(y_1, y_2)}{\partial y_1} (y_2 - 1.6) + \frac{\partial \psi_{N,\Delta}(y_1, y_2)}{\partial y_2} u \right) \right].
 \end{aligned}$$

Note that the function  $\Phi_{N,\Delta}(u, y_1, y_2)$  is strictly convex in  $u$  and, hence, the  $N$ -regularity condition is satisfied on every subset  $Q$  of  $Y$ . Also, the solution  $u_{N,\Delta}(y_1, y_2)$

of the problem (8.10) is unique and is defined by the equation (8.12)

$$u_{N,\Delta}(y_1, y_2) = \begin{cases} -1.25 \frac{\partial \psi_{N,\Delta}(y_1, y_2)}{\partial y_2}, & -0.4 \leq -1.25 \frac{\partial \psi_{N,\Delta}(y_1, y_2)}{\partial y_2} \leq 0.4, \\ -0.4, & -1.25 \frac{\partial \psi_{N,\Delta}(y_1, y_2)}{\partial y_2} < -0.4, \\ 0.4, & -1.25 \frac{\partial \psi_{N,\Delta}(y_1, y_2)}{\partial y_2} > 0.4. \end{cases}$$

The graph of this function (with  $N$  and  $\Delta$  as in Figures 1 and 2 is depicted in Figure 3, where the points belonging to the set  $\Theta_{N,\Delta}$  are marked, for convenience, as well (the latter being visibly close to the surface defined by  $u_{N,\Delta}(\cdot)$ ).

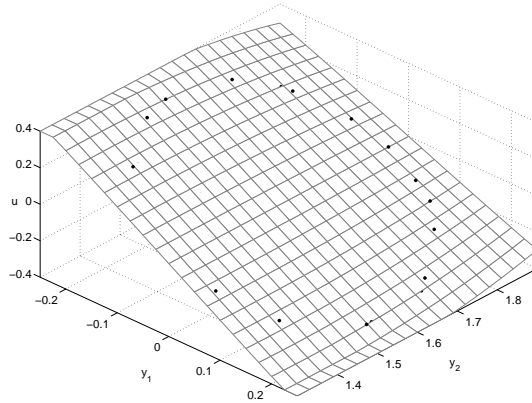


FIG. 3.  $u_{N,\Delta}(y_1, y_2)$ . Obtained feedback control function.

As can be seen from Figures 2 and 3, the function  $u_{N,\Delta}(y_1, y_2)$ , being defined by the right-hand side of (8.12), is smooth in a neighborhood  $Q$  of  $\mathcal{Y}_{N,\Delta}$ , the latter being close to the optimal periodic orbit  $\mathcal{Y}_*$ . Hence, one can assume (based on Theorem 7.1) that  $u_{N,\Delta}(y_1, y_2)$  can serve as an approximation for the optimal control  $u_*(y_1, y_2)$  on  $\mathcal{Y}_*$ . To verify that this is the case, we used MATLAB to integrate our system with the feedback control  $u_{N,\Delta}(y_1, y_2)$  and with the initial conditions  $y_1(0) = 0$ ,  $y_2(0) = 1.848$  (a point in  $\mathcal{Y}_{N,\Delta}$ ). The result of such integration is the  $T$ -periodic solution  $y(\tau) = (y_1(\tau), y_2(\tau))$  and the corresponding  $T$ -periodic control  $u(\tau) \stackrel{\text{def}}{=} u_{N,\Delta}(y_1(\tau), y_2(\tau))$ , with  $T = 4.8597$ . The graphs of these functions are shown in Figures 4 and 5 below (which look very similar to those obtained in [54]). The value of the objective function obtained on this periodic solution is evaluated to be equal to  $4.195877 \approx 4.196$ . That is, (8.5) is satisfied and, hence, an approximate solution to the periodic optimization problem is found. Note that the closed curve  $\mathcal{Y}$  defined by  $(y_1(\tau), y_2(\tau))$ ,

$$(8.13) \quad \mathcal{Y} \stackrel{\text{def}}{=} \{(y_1, y_2) : (y_1, y_2) = (y_1(\tau), y_2(\tau)) \text{ for some } \tau \in [0, T]\},$$

passes very close to the points of  $\mathcal{Y}_{N,\Delta}$  as illustrated by Figure 6.

**9. Proofs for sections 4 and 5.** Due to the approximating property of the sequence of the functions  $\phi_i(\cdot)$ ,  $i = 1, 2, \dots$ , (see (5.1)), the set  $W$  can be presented in the form

$$(9.1) \quad W = \left\{ \gamma \in \mathcal{P}(U \times Y) : \int_{U \times Y} (\phi'_i(y))^\top f(u, y) \gamma(du, dy) = 0, \quad i = 1, 2, \dots \right\},$$

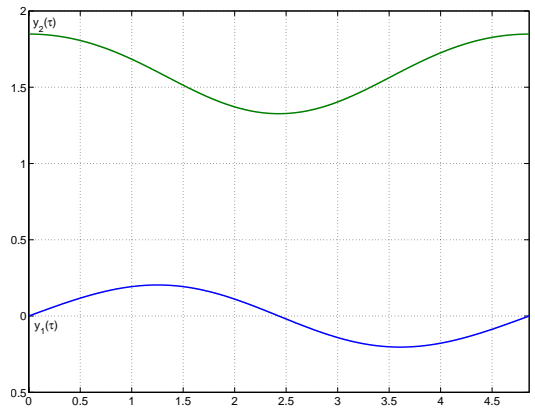


FIG. 4.  $(y_1(\tau), y_2(\tau))$ . *Obtained periodic solution.*

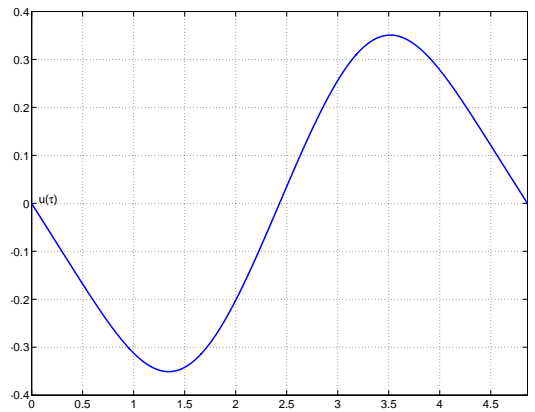


FIG. 5.  $u(\tau)$ . *Obtained periodic control.*

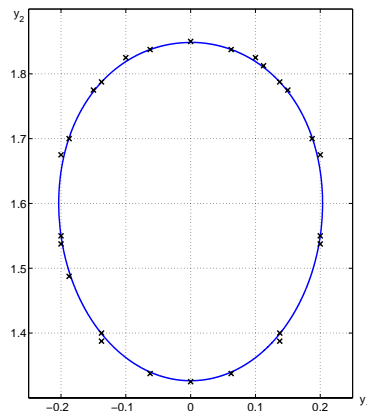


FIG. 6.  $\mathcal{Y}$ . *Obtained periodic orbit.*

where, without loss of generality, one may assume that the functions  $\phi_i(\cdot)$  satisfy the following normalization conditions:

$$(9.2) \quad \max_{y \in \bar{D}} \{|\phi_i(y)|, \|\phi'_i(y)\|, \|\phi''_i(y)\|\} \leq \frac{1}{2^i}, \quad i = 1, 2, \dots$$

In the above expression,  $\|\phi_i''(y)\|$  is a norm of the Hessian (the matrix of second derivatives of  $\phi_i(y)$ ) in  $\mathbb{R}^m \times \mathbb{R}^m$ ,  $\|\phi_i'(y)\|$  is a norm of  $\phi_i'(y)$  in  $\mathbb{R}^m$ , and  $\hat{D}$  is a closed ball in  $\mathbb{R}^m$  that contains  $Y$  in its interior.

Let  $l_1$  and  $l_\infty$  stand for the Banach spaces of infinite sequences such that, for any  $x = (x_1, x_2, \dots) \in l_1$ ,  $\|x\|_{l_1} \stackrel{\text{def}}{=} \sum_i |x_i| < \infty$  and, for any  $\lambda = (\lambda_1, \lambda_2, \dots) \in l_\infty$ ,  $\|\lambda\|_{l_\infty} \stackrel{\text{def}}{=} \sup_i |\lambda_i| < \infty$ . It is easy to see that, given an element  $\lambda \in l_\infty$ , one can define a linear continuous functional  $\lambda(\cdot) : l_1 \rightarrow \mathbb{R}^1$  by the equation

$$(9.3) \quad \lambda(x) = \sum_i \lambda_i x_i \quad \forall x \in l_1, \quad \|\lambda(\cdot)\| = \|\lambda\|_{l_\infty}.$$

It is also known (see, e.g., [49, p. 86]) that any continuous linear functional  $\lambda(\cdot) : l_1 \rightarrow \mathbb{R}^1$  can be presented in the form (9.3) with some  $\lambda \in l_\infty$ .

By (9.2),  $(\phi_1(y), \phi_2(y), \dots) \in l_1$  and  $(\frac{\partial \phi_1}{\partial y_j}, \frac{\partial \phi_2}{\partial y_j}, \dots) \in l_1$  for any  $y \in Y$ . Hence, the function  $\psi_\lambda(y)$ ,

$$(9.4) \quad \psi_\lambda(y) \stackrel{\text{def}}{=} \sum_i \lambda_i \phi_i(y), \quad \lambda = (\lambda_1, \lambda_2, \dots) \in l_\infty,$$

is continuously differentiable, with  $\psi'_\lambda(y) = \sum_i \lambda_i \phi'_i(y)$ .

*Proof of Theorem 4.1(iii).* If the function  $\psi(\cdot)$  satisfying (4.6) exists, then  $\min_{(u,y) \in U \times Y} (-\psi'(y))^T f(u, y) > 0$  and, hence,

$$(9.5) \quad \lim_{\alpha \rightarrow \infty} \min_{(u,y) \in U \times Y} \{g(u, y) + \alpha(-\psi'(y))^T f(u, y)\} = \infty.$$

This implies that the optimal value of the dual problem is unbounded ( $\mu_* = \infty$ ).

Assume now that the optimal value of the dual problem is unbounded. That is, there exists a sequence  $(\mu_k, \psi_k(\cdot))$  such that

$$(9.6) \quad \mu_k \leq g(u, y) + (\psi'_k(y))^T f(u, y) \quad \forall (u, y) \in U \times Y, \quad \lim_{k \rightarrow \infty} \mu_k = \infty$$

$$(9.7) \quad \Rightarrow 1 \leq \frac{1}{\mu_k} g(u, y) + \frac{1}{\mu_k} (\psi'_k(y))^T f(u, y) \quad \forall (u, y) \in U \times Y.$$

For  $k$  large enough,  $\frac{1}{\mu_k} |g(u, y)| \leq \frac{1}{2} \quad \forall (u, y) \in U \times Y$ . Hence

$$\frac{1}{2} \leq \frac{1}{\mu_k} (\psi'_k(y))^T f(u, y) \quad \forall (u, y) \in U \times Y.$$

That is, the function  $\psi(y) \stackrel{\text{def}}{=} -\frac{1}{\mu_k} \psi_k(y)$  satisfies (4.6).  $\square$

*Proof of Theorem 4.1(i).* From (4.4) it follows that, if  $W$  is not empty, then the optimal value of the dual problem is bounded.

Conversely, let us assume that the optimal value  $\mu_*$  of the dual problem is bounded and let us establish that  $W$  is not empty. Assume that this is not true and  $W$  is empty. Define the set  $\mathcal{Q}$  by the equation

$$(9.8) \quad \mathcal{Q} \stackrel{\text{def}}{=} \left\{ x = (x_1, x_2, \dots) : x_i = \int_{u,y} (\phi'_i(y))^T f(u, y) \gamma(du, dy), \quad \gamma \in \mathcal{P}(U \times Y) \right\}.$$

It is easy to see that the set  $\mathcal{Q}$  is a convex and compact subset of  $l_1$  (the fact that  $\mathcal{Q}$  is relatively compact in  $l_1$  is implied by (9.2); the fact that it is closed follows from that  $\mathcal{P}(U \times Y)$  is compact in weak convergence topology).

By (9.1), the assumption that  $W$  is empty is equivalent to the assumption that the set  $\mathcal{Q}$  does not contain the “zero element” ( $0 \notin \mathcal{Q}$ ). Hence, by a separation theorem (see, e.g., [49, p. 59]), there exists  $\bar{\lambda} = (\bar{\lambda}_1, \bar{\lambda}_2, \dots) \in l_\infty$  such that

$$\begin{aligned} 0 = \bar{\lambda}(0) &> \max_{x \in \mathcal{Q}} \sum_i \bar{\lambda}_i x_i = \max_{\gamma \in \mathcal{P}(U \times Y)} \int_{U \times Y} (\psi'_{\bar{\lambda}}(y))^T f(u, y) \gamma(du, dy) \\ &= \max_{(u, y) \in U \times Y} (\psi'_{\bar{\lambda}}(y))^T f(u, y), \end{aligned}$$

where  $\psi_{\bar{\lambda}}(y) = \sum_i \bar{\lambda}_i \phi_i(y)$  (see (9.4)). This implies that the function  $\psi(y) \stackrel{\text{def}}{=} \psi_{\bar{\lambda}}(y)$  satisfies (4.6), and, by Theorem 4.1(iii),  $\mu_*$  is unbounded. Thus, we have obtained a contradiction that proves that  $W$  is not empty.  $\square$

*Proof of Theorem 4.1(ii).* By Theorem 4.1(i), if the optimal value of the dual problem (4.1) is bounded, then  $W$  is not empty and, hence, a solution of the problem (3.2) exists.

Define the set  $\hat{\mathcal{Q}} \subset \mathbb{R}^1 \times l_1$  by the equation

$$\begin{aligned} \hat{\mathcal{Q}} &\stackrel{\text{def}}{=} \left\{ (\theta, x) : \theta \geq \int_{U \times Y} g(u, y) \gamma(du, dy), \right. \\ (9.9) \quad x &= (x_1, x_2, \dots), \quad x_i = \int_{U \times Y} (\phi'_i(y))^T f(u, y) \gamma(du, dy), \quad \gamma \in \mathcal{P}(U \times Y) \left. \right\}. \end{aligned}$$

The set  $\hat{\mathcal{Q}}$  is convex and closed. Also, for any  $j = 1, 2, \dots$ , the point  $(\theta_j, 0) \notin \hat{\mathcal{Q}}$ , where  $\theta_j \stackrel{\text{def}}{=} G_* - \frac{1}{j}$  and 0 is the zero element of  $l_1$ . On the basis of a separation theorem (see [49, p. 59]), one may conclude that there exists a sequence  $(\kappa^j, \lambda^j) \in \mathbb{R}^1 \times l_\infty$ ,  $j = 1, 2, \dots$  (with  $\lambda^j \stackrel{\text{def}}{=} (\lambda_1^j, \lambda_2^j, \dots)$ ) such that

$$\begin{aligned} \kappa^j \left( G_* - \frac{1}{j} \right) + \delta^j &\leq \inf_{(\theta, x) \in \hat{\mathcal{Q}}} \left\{ \kappa^j \theta + \sum_i \lambda_i^j x_i \right\} = \inf_{\gamma \in \mathcal{P}(U \times Y)} \left\{ \kappa^j \theta \right. \\ (9.10) \quad &+ \left. \int_{U \times Y} (\psi'_{\lambda^j}(y))^T f(u, y) \gamma(du, dy) \text{ s.t. } \theta \geq \int_{U \times Y} g(u, y) \gamma(du, dy) \right\}, \end{aligned}$$

where  $\delta^j > 0$  for all  $j$  and  $\psi_{\lambda^j}(y) = \sum_i \lambda_i^j \phi_i(y)$ . From (9.10) it immediately follows that  $\kappa^j \geq 0$ . Let us show that  $\kappa^j > 0$ . In fact, if this were not the case, one would obtain that

$$\begin{aligned} 0 < \delta^j &\leq \min_{\gamma \in \mathcal{P}(U \times Y)} \int_{U \times Y} (\psi'_{\lambda^j}(y))^T f(u, y) \gamma(du, dy) = \min_{(u, y) \in U \times Y} \{ (\psi'_{\lambda^j}(y))^T f(u, y) \} \\ &\Rightarrow \max_{(u, y) \in U \times Y} \{ (-\psi'_{\lambda^j}(y))^T f(u, y) \} \leq -\delta^j < 0. \end{aligned}$$

The latter would lead to the validity of the inequality (4.6) with  $\psi(y) = -\psi_{\lambda^j}(y)$ , which, by Theorem 4.1(iii), would imply that the optimal value of the dual problem is unbounded. Thus,  $\kappa^j > 0$ .



Dividing (9.10) by  $\kappa^j$  one can obtain that

$$\begin{aligned} G_* - \frac{1}{j} &< \left( G_* - \frac{1}{j} \right) + \frac{\delta^j}{\kappa_j} \\ &\leq \min_{\gamma \in \mathcal{P}(U \times Y)} \left\{ \int_{U \times Y} \left( g(u, y) + \frac{1}{\delta^j} (\psi'_{\lambda^j}(y))^T f(u, y) \right) \gamma(du, dy) \right\} \\ &= \min_{(u, y) \in U \times Y} \left\{ g(u, y) + \frac{1}{\delta^j} (\psi'_{\lambda^j}(y))^T f(u, y) \right\} \leq \mu^* \Rightarrow G_* \leq \mu_*. \end{aligned}$$

The latter and (4.4) prove (4.5).  $\square$

*Proof of Theorem 5.2.* The proof statement (iii) of the theorem follows the argument used in the proof of Theorem 4.1(iii), with the replacement of  $\psi(y)$  in (9.5) by  $\psi_v(y) = \sum_1^N v_i \phi_i(y)$ ,  $v = (v_i) \in \mathbb{R}^N$ , and with the replacement of  $\psi_k(y)$  in (9.6), (9.7) by  $\psi_{v^k}(y) = \sum_1^N v_i^k \phi_i(y)$ ,  $v^k = (v_i^k) \in \mathbb{R}^N$ .

The proofs of statements (i) and (ii) of the theorem are based on a separation theorem in F-D spaces and follow the argument used in the proofs of Theorem 4.1(i) and Theorem 4.1(ii), with the replacement of the set  $\mathcal{Q}$  defined in (9.8) by the set  $\mathcal{Q}' \subset \mathbb{R}^N$ ,  
(9.11)

$$\mathcal{Q}' \stackrel{\text{def}}{=} \left\{ x = (x_1, \dots, x_N) : x_i = \int_{u, y} (\phi'_i(y))^T f(u, y) \gamma(du, dy), \quad \gamma \in \mathcal{P}(U \times Y) \right\},$$

and with the replacement of the set  $\hat{\mathcal{Q}}$  defined in (9.9) by the set  $\hat{\mathcal{Q}}' \subset \mathbb{R}^1 \times \mathbb{R}^N$ ,

$$\hat{\mathcal{Q}}' \stackrel{\text{def}}{=} \left\{ (\theta, x) : \theta \geq \int_{U \times Y} g(u, y) \gamma(du, dy), \right.$$

(9.12)

$$\left. x = (x_1, \dots, x_N), \quad x_i = \int_{U \times Y} (\phi'_i(y))^T f(u, y) \gamma(du, dy), \quad \gamma \in \mathcal{P}(U \times Y) \right\}. \quad \square$$

*Proof of Proposition 5.3.* Let (5.17) be valid. Then, for some positive  $r$ ,

$$r\bar{B} \subset \text{co}K_N,$$

where  $\bar{B}$  is a closed unit ball in  $\mathbb{R}^N$ . By (5.18), the inequality (5.15) is equivalent to

$$(9.13) \quad \max_{z=(z_i) \in K_N} \sum_{i=1}^N v_i z_i \leq 0 \quad \Rightarrow \quad \max_{z=(z_i) \in r\bar{B}} \sum_{i=1}^N v_i z_i \leq 0.$$

Since  $\max_{z=(z_i) \in r\bar{B}} \sum_{i=1}^N v_i z_i = r\|v\|$ , the second inequality in (9.13) implies that  $v = 0$ . That is, Assumption 1 is satisfied.

Assume now that (5.17) is not valid, that is, either 0 is a boundary point of  $\text{co}K_N$  or it does not belong to the closure of  $\text{co}K_N$ . Then, by separation theorem, there exists a vector  $v = (v_i) \neq 0$  such that

$$(9.14) \quad \max_{z=(z_i) \in K_N} \sum_{i=1}^N v_i z_i = \max_{z=(z_i) \in \text{co}K_N} \sum_{i=1}^N v_i z_i \leq 0.$$

This implies that Assumption 1 is not satisfied.  $\square$

*Proof of Theorem 5.5.* The fact that Assumption 1 is satisfied implies, in particular, that the function  $\psi_v(\cdot)$  satisfying (5.13) does not exist. Hence, by Theorem 5.2(iii), the optimal value  $\mu_N$  of the D-N-LLPP (5.8) is bounded. Let  $v^k = (v_i^k) \in \mathbb{R}^N$  be such that

$$(9.15) \quad \mu_N - \frac{1}{k} \leq g(u, y) + \left( \sum_{i=1}^N v_i^k \phi'_i(y) \right)^T f(u, y) \quad \forall (u, y) \in U \times Y, \quad k = 1, 2, \dots$$

Show that the sequence  $v^k$  is bounded. That is,

$$(9.16) \quad \|v^k\| \leq c_N, \quad k = 1, 2, \dots$$

In fact, if  $v^k$ ,  $k = 1, 2, \dots$ , were not bounded, then there would exist a subsequence  $v^{k'}$  such that

$$(9.17) \quad \lim_{k' \rightarrow \infty} \|v^{k'}\| = \infty, \quad \lim_{k' \rightarrow \infty} \frac{v^{k'}}{\|v^{k'}\|} \stackrel{\text{def}}{=} \tilde{v}, \quad \|\tilde{v}\| = 1.$$

Dividing (9.15) by  $\|v^k\|$  and passing to the limit over the subsequence  $\{k'\}$ , one would obtain the following inequality:

$$(9.18) \quad 0 \leq \left( \sum_{i=1}^N \tilde{v}_i \phi'_i(y) \right)^T f(u, y) \quad \forall (u, y) \in U \times Y.$$

By Assumption 1, the fact that (9.18) is valid implies that  $\tilde{v} = (\tilde{v}_i) = 0$ , which is in contradiction to (9.17). Thus, the validity of (9.16) is established.

Due to (9.16), there exists a subsequence  $\{k'\}$  such that there exists a limit

$$(9.19) \quad \lim_{k' \rightarrow \infty} v^{k'} \stackrel{\text{def}}{=} v.$$

Passing over this subsequence to the limit in (9.15), one obtains

$$(9.20) \quad \mu_N \leq g(u, y) + \left( \sum_{i=1}^N v_i \phi'_i(y) \right)^T f(u, y) \quad \forall (u, y) \in U \times Y \quad \Rightarrow \quad v = (v_i) \in V_N.$$

The latter proves that  $V_N$  is not empty. The proof of that  $V_N$  is bounded follows the same argument as that used above to establish (9.16). This completes the proof of statement (i) of the theorem.

To prove statement (ii), let us assume that (5.19) and (5.20) are valid but Assumption 1 is not satisfied, and, hence, there exists a vector  $v = (v_i) \neq 0$  such that

$$(9.21) \quad \left( \sum_{i=1}^N v_i \phi'_i(y) \right)^T f(u, y) \leq 0 \quad \forall (u, y) \in U \times Y.$$

By (9.21), for an arbitrary  $v^0 = (v_i^0) \in V_N$  and for an arbitrary  $\alpha \geq 0$ ,

$$(9.22) \quad \mu_N \leq g(u, y) + \left( \sum_{i=1}^N (v_i^0 - \alpha v_i) \phi'_i(y) \right)^T f(u, y) \quad \forall (u, y) \in U \times Y.$$

The latter implies that  $(v^0 - \alpha v) \in V_N \forall \alpha \geq 0$ , which contradicts the fact that  $V_N$  is bounded (postulated by (5.20)). This proves statement (ii) of the theorem.  $\square$

*Proof of Theorem 5.6.* Consider first the case when  $\mu_* < \infty$ . For arbitrary small  $\delta > 0$ , there exists a function  $\psi(\cdot) \in C^1$  such that

$$\min_{(u,y) \in U \times Y} \{g(u, y) + \psi'(y)f(u, y)\} \geq \mu_* - \delta.$$

By the approximating property (5.1), there exist numbers  $v_1, v_2, \dots, v_{\bar{N}}$  (with  $\bar{N}$  being large enough) such that

$$(9.23) \quad \max_{y \in Y} \left\| \psi'(y) - \sum_{i=1}^{\bar{N}} v_i \phi'_i(y) \right\| \leq \delta$$

$$(9.24) \quad \Rightarrow \left| \min_{(u,y) \in U \times Y} \{g(u, y) + \psi'(y)f(u, y)\} - \min_{(u,y) \in U \times Y} \left\{ g(u, y) + \sum_{i=1}^{\bar{N}} v_i (\phi'_i(y))^T f(u, y) \right\} \right| \leq c\delta,$$

where  $c \stackrel{\text{def}}{=} \max_{(u,y) \in U \times Y} \|f(u, y)\|$ . From the above expressions and from the fact that

$$(9.25) \quad \min_{(u,y) \in U \times Y} \left\{ g(u, y) + \sum_{i=1}^N v_i (\phi'_i(y))^T f(u, y) \right\} \leq \mu_{\bar{N}},$$

it follows that  $\mu_* - \delta \leq \mu_{\bar{N}} + c\delta$ . This and (5.21) imply that  $0 \leq \mu_* - \mu_N \leq c(1 + \delta)$  for  $N \geq \bar{N}$ . Since  $\delta$  is arbitrarily small, the latter proves (5.22) for bounded  $\mu_*$ .

Assume now that  $\mu_* = \infty$ . Then, for an arbitrary large  $A > 0$ , there exists a function  $\psi(\cdot) \in C^1$  such that

$$\min_{(u,y) \in U \times Y} \{g(u, y) + \psi'(y)f(u, y)\} \geq A.$$

By the approximating property (5.1), there exist numbers  $v_1, v_2, \dots, v_{\bar{N}}$  (with  $\bar{N}$  being large enough), such that (9.23) and (9.24) are valid. Also, the inequality (9.25) remains valid due to the definition of  $\mu_N$  (see (5.8)). Consequently,  $\mu_{\bar{N}} \geq A - c\delta$ . Since  $\delta > 0$  can be chosen arbitrarily small, the latter implies that  $\mu_{\bar{N}} \geq A$ . Hence,  $\lim_{N \rightarrow \infty} \mu_N = \infty$ .  $\square$

## 10. Proofs for sections 6 and 7.

*Proof of Theorem 6.2.* First, let us show that the set  $\Lambda_{N,\Delta}$  is bounded for  $\Delta$  small enough. That is, show that

$$(10.1) \quad \sup_{\lambda \in \Lambda_{N,\Delta}} \|\lambda\| \leq c_N = \text{const}$$

for  $\Delta \leq \Delta_N$  ( $\Delta_N > 0$ ). Assume that this is not true and, hence, there exist sequences  $\Delta_s$  and  $\lambda^{N,\Delta_s} \in \Lambda_{N,\Delta_s}$ ,  $s = 1, 2, \dots$ , such that

$$\lim_{s \rightarrow \infty} \Delta_s = 0, \quad \lim_{s \rightarrow \infty} \|\lambda^{N,\Delta_s}\| = \infty.$$

Without loss of generality one may assume that there exists a limit

$$(10.2) \quad \lim_{s \rightarrow \infty} \frac{\lambda^{N, \Delta_s}}{\|\lambda^{N, \Delta_s}\|} \stackrel{\text{def}}{=} v, \quad \|v\| = 1.$$

From the definition of  $\Lambda_{N, \Delta}$  (see (6.9)) it follows that the inequality

$$(10.3) \quad \mu_{N, \Delta} \leq g(u_l, y_k) + \sum_{i=1}^N \lambda_i^{N, \Delta} (\phi'_i(y_k))^T f(u_l, y_k)$$

is valid for any grid point  $(u_l, y_k) \in U \times Y$ . Substituting  $\Delta_s$  for  $\Delta$  in (10.3) and then dividing the latter by  $\|\lambda^{N, \Delta_s}\|$  and passing to the limit as  $s \rightarrow \infty$ , one can prove that

$$(10.4) \quad 0 \leq \sum_{i=1}^N v_i (\phi'_i(y))^T f(u, y) \quad \forall (u, y) \in U \times Y.$$

Note that the proof of the above inequality is based on the fact that (see (5.12), (6.6), and (6.8))

$$(10.5) \quad \lim_{\Delta \rightarrow 0} \mu_{N, \Delta} = \mu_N,$$

which, in particular, implies that  $\mu_{N, \Delta_s}$  remains bounded as  $s \rightarrow \infty$ , and also on the fact that any point  $(u, y)$  in  $U \times Y$  can be presented as a limit of a sequence of grid points. From Assumption 1 it now follows that  $v = (v_i) = 0$ , which contradicts (10.2). This proves (10.1).

Let us now prove (6.10). Assuming that it is not true, one can come to a conclusion that there exist a positive number  $\alpha$  and sequences  $\Delta_s$  and  $\lambda^{N, \Delta_s} \in \Lambda_{N, \Delta_s}$ ,  $s = 1, 2, \dots$ , such that

$$(10.6) \quad \lim_{s \rightarrow \infty} \Delta_s = 0, \quad \text{dist}(\lambda^{N, \Delta_s}, V_N) \geq \alpha, \quad s = 1, 2, \dots$$

Due to (10.1), one may assume without loss of generality that there exists a limit

$$(10.7) \quad \lim_{s \rightarrow \infty} \lambda^{N, \Delta_s} \stackrel{\text{def}}{=} v^N \Rightarrow \text{dist}(v^N, V_N) \geq \alpha.$$

Substituting  $\Delta_s$  for  $\Delta$  in (10.3), taking into account (10.5), and passing to the limit as  $s \rightarrow \infty$ , one can obtain that

$$(10.8) \quad \mu_N \leq g(u, y) + \sum_{i=1}^N v_i^N (\phi'_i(y))^T f(u, y) \quad \forall (u, y) \in U \times Y \Rightarrow v^N = (v_i^N) \in V_N.$$

The latter contradicts (10.7) and, thus, proves (6.10).  $\square$

*Proofs of Theorem 7.1 and Corollary 7.2.* Fix an arbitrary  $\bar{y} \in \mathcal{Y}_* \cap Q$ . From Theorem 7.3(ii) it follows that there exists  $(u_{l, \Delta}, y_{k, \Delta}) \in \Theta_{N, \Delta}$  such that

$$(10.9) \quad \lim_{N \rightarrow \infty} \overline{\lim_{\Delta \rightarrow 0}} \|(u_{l, \Delta}, y_{k, \Delta}) - (u_*(\bar{y}), \bar{y})\| = 0.$$

Note that, since  $Q$  is open, from (10.9) it follows that there exists  $r > 0$  such that, for  $N$  large and  $\Delta$  small enough,

$$(10.10) \quad y_{k, \Delta} \in \bar{y} + r\bar{D} \subset Q, \quad \bar{D} \stackrel{\text{def}}{=} \{y : \|y\| \leq 1\}.$$

Show that the validity of (10.9) also implies the validity of

$$(10.11) \quad \lim_{N \rightarrow \infty} \overline{\lim}_{\Delta \rightarrow 0} \|u_{N,\Delta}(y_{k_{N,\Delta}}) - u_*(\bar{y})\| = 0$$

if  $N$ -RC is satisfied.

Assume that  $N$ -RC is satisfied on  $Q$  but (10.11) is not true. Then, due to (10.9), there exist a number  $\alpha > 0$  and sequences  $N_s$  ( $\lim_{s \rightarrow \infty} N_s = \infty$ ),  $\Delta_{s,j}$  ( $\lim_{j \rightarrow \infty} \Delta_{s,j} = 0$ ) such that

$$(10.12) \quad \|u_{N_s, \Delta_{s,j}}(y_{k_{N_s, \Delta_{s,j}}}) - u_{l_{N_s, \Delta_{s,j}}} \| \geq \alpha.$$

Without loss of generality, one may assume that there exist limits

$$(10.13) \quad \lim_{j \rightarrow \infty} y_{k_{N_s, \Delta_{s,j}}} \stackrel{\text{def}}{=} y_{N_s}, \quad \lim_{j \rightarrow \infty} u_{N_s, \Delta_{s,j}}(y_{k_{N_s, \Delta_{s,j}}}) \stackrel{\text{def}}{=} \tilde{u}_{N_s}, \quad \lim_{j \rightarrow \infty} u_{l_{N_s, \Delta_{s,j}}} \stackrel{\text{def}}{=} \tilde{u}_{N_s},$$

and also that there exists a limit

$$(10.14) \quad \lim_{j \rightarrow \infty} \lambda^{N_s, \Delta_{s,j}} \stackrel{\text{def}}{=} v^{N_s}, \quad v^{N_s} = (v_i^{N_s}) \in V_{N_s},$$

where the validity of the last inclusion is implied by Theorem 6.2 (remember that  $\lambda^{N_s, \Delta_{s,j}} \in \Lambda_{N_s, \Delta_{s,j}}$ ). From the duality relationships between the solutions of  $N\Delta$ -LLPP (6.3) and the solutions of  $D-N\Delta$ -LLPP (6.7) it follows that, for any  $(u_l, y_k) \in \Theta_{N,\Delta}$ ,

$$(10.15) \quad \mu_{N,\Delta} = g(u_l, y_k) + \sum_{i=1}^N \lambda_i^{N,\Delta} (\phi'_i(y_k))^T f(u_l, y_k).$$

From (10.15) and from the fact that  $u_{N,\Delta}(\cdot)$  is defined as a solution of (7.1) it also follows that, for any  $y_k \in \mathcal{Y}_{N,\Delta}$ ,

$$(10.16) \quad \mu_{N,\Delta} \geq g(u_{N,\Delta}(y_k), y_k) + \sum_{i=1}^N \lambda_i^{N,\Delta} (\phi'_i(y_k))^T f(u_{N,\Delta}(y_k), y_k).$$

Via substituting  $N_s, \Delta_{s,j}, u_{l_{N_s, \Delta_{s,j}}}, y_{k_{N_s, \Delta_{s,j}}}$  for, respectively,  $N, \Delta, u_l, y_k$  in (10.15) and then passing to the limit as  $j \rightarrow \infty$ , one can obtain (see (10.13), (10.14) and (5.12), (6.6), (6.8)) that

$$(10.17) \quad \mu_{N_s} = g(\tilde{u}_{N_s}, y_{N_s}) + \sum_{i=1}^N v_i^{N_s} (\phi'_i(y_{N_s}))^T f(\tilde{u}_{N_s}, y_{N_s}).$$

Similarly, via substituting  $N_s, \Delta_{s,j}, u_{N_s, \Delta_{s,j}}(y_{k_{N_s, \Delta_{s,j}}}), y_{k_{N_s, \Delta_{s,j}}}$  for, respectively,  $N, \Delta, u_{N,\Delta}(y_k), y_k$  in (10.16) and passing to the limit as  $j \rightarrow \infty$ , one obtains that

$$(10.18) \quad \mu_{N_s} \geq g(\tilde{u}_{N_s}, y_{N_s}) + \sum_{i=1}^N v_i^{N_s} (\phi'_i(y_{N_s}))^T f(\tilde{u}_{N_s}, y_{N_s})$$

$$(10.19) \quad \Rightarrow \quad \mu_{N_s} = g(\tilde{u}_{N_s}, y_{N_s}) + \sum_{i=1}^N v_i^{N_s} (\phi'_i(y_{N_s}))^T f(\tilde{u}_{N_s}, y_{N_s}).$$

The equality (10.19) is implied by the inequality (10.18) due to the fact that

$$(10.20) \quad \mu_{N_s} = \min_{(u,y) \in U \times Y} \left\{ g(u, y) + \sum_{i=1}^N v_i^{N_s} (\phi'_i(y))^T f(u, y) \right\},$$

which is a consequence of  $v^{N_s} \in V_{N_s}$  (see (5.19)).

From (10.17) and (10.19) it follows (see the notation introduced in (5.23), (5.24), and (5.25)) that

$$(10.21) \quad y_{N_s} \in \mathcal{Y}_{N_s}^{v_{N_s}}, \quad \tilde{u}_{N_s} \in \mathcal{U}_{N_s}^{v_{N_s}}(y_{N_s}), \quad \tilde{u}_{N_s} \in \mathcal{U}_{N_s}^{v_{N_s}}(y_{N_s}).$$

By (10.10),  $y_{N_s} \in Q \cap \mathcal{Y}_{N_s}^{v_{N_s}}$  for  $s$  large enough. Hence, due to the validity of  $N$ -RC on  $Q$  (with  $N = N_s$ ),  $\mathcal{U}_{N_s}^{v_{N_s}}(y_{N_s})$  is a singleton. Consequently, by (10.21),

$$(10.22) \quad \tilde{u}_{N_s} = \tilde{u}_{N_s}.$$

Since, on the other hand, by (10.12) and (10.13),  $\|\tilde{u}_{N_s} - \tilde{u}_{N_s}\| \geq \alpha$ , one obtains a contradiction that establishes the validity of (10.11).

To prove (7.6), let us assume that it is not valid for the given  $\bar{y}$  and, hence, there exist sequences  $N' \rightarrow \infty$  and  $\Delta' \rightarrow 0$  such that

$$(10.23) \quad \lim_{N' \rightarrow \infty} \lim_{\Delta' \rightarrow 0} u_{N', \Delta'}(\bar{y}) = \tilde{u} \neq u_*(\bar{y}).$$

The functions  $u_{N', \Delta'}(\cdot)$  are uniformly equicontinuous on  $\bar{Q}' = \bar{y} + r\bar{D}$ . Consequently, by the Arzela–Ascoli theorem, there exist subsequences  $N'' \rightarrow \infty$  and  $\Delta'' \rightarrow 0$  of  $N'$  and  $\Delta'$ , and there exist a uniformly continuous function  $u_0(\cdot) : \bar{y} + r\bar{D} \rightarrow U$  such that

$$(10.24) \quad \lim_{N'' \rightarrow \infty} \lim_{\Delta'' \rightarrow 0} \max_{x \in \bar{y} + r\bar{D}} \|u_{N'', \Delta''}(x) - u_0(x)\| = 0,$$

where, by (10.23),

$$(10.25) \quad u_0(\bar{y}) = \tilde{u} \neq u_*(\bar{y}).$$

On the other hand, passing to the limit over the subsequences  $N''$  and  $\Delta''$  in the inequality

$$\|u_0(\bar{y}) - u_*(\bar{y})\| \leq \|u_0(\bar{y}) - u_{N, \Delta}(\bar{y})\|$$

$$+ \|u_{N, \Delta}(\bar{y}) - u_{N, \Delta}(y_{k_{N, \Delta}})\| + \|u_{N, \Delta}(y_{k_{N, \Delta}}) - u_*(\bar{y})\|$$

and using (7.5), (10.11), and (10.24), one can obtain that  $\|u_0(\bar{y}) - u_*(\bar{y})\| = 0$ , which contradicts (10.25) and, thus, proves (7.6). The uniform convergence of  $u_{N, \Delta}(\cdot)$  to  $u_*(\cdot)$  on any closed subset of  $\mathcal{Y}_* \cap Q$  follows from (7.6) and the Arzela–Ascoli theorem. Thus, statement (i) of the theorem is proved.

Statement (ii) of theorem is a special case ( $J = 1$ ) of Corollary 7.2, which we now will prove. By Theorem 7.1(i) established above, (7.10) is valid for any

$$y \in \cup_j (\mathcal{Y}_* \cap Q_j) = \mathcal{Y}_* \cap (\cup_j Q_j).$$

Thus, to establish the validity of the required result, one needs to verify that

$$(10.26) \quad \gamma_*(\mathcal{Y}_* \cap (\cup_j Q_j)) = 1.$$

From (7.4) and (7.8) it follows that  $\gamma_*(\overline{\cup_j Q_j}) = 1$ , which implies that

$$y_*(\tau) \in \overline{\cup_j Q_j} \quad \forall \tau \in [0, T].$$

The latter is equivalent to  $\mathcal{Y}_* \subset \overline{\cup_j Q_j}$  and, hence,

$$(10.27) \quad \gamma_*(\mathcal{Y}_* \cap (\cup_j Q_j)) = \gamma_*(\mathcal{Y}_* \cap (\overline{\cup_j Q_j})) - \gamma_*(\mathcal{Y}_* \cap (\overline{\cup_j Q_j} / \cup_j Q_j)).$$

This and (7.9) imply (10.26).  $\square$

**Acknowledgments.** We thank M. Quincampoix for useful discussions during his visit to the University of South Australia in March 2006. We also thank J. Krawczyk for providing us with important references on the subject.

#### REFERENCES

- [1] O. ALVAREZ AND M. BARDI, *Viscosity solutions methods for singular perturbations in deterministic and stochastic control*, SIAM J. Control Optim., 40 (2001), pp. 1159–1188.
- [2] O. ALVAREZ AND M. BARDI, *Singular perturbations of nonlinear degenerate parabolic PDEs: A general convergence result*, Arch. Ration. Mech. Anal., 170 (2003), pp. 17–61.
- [3] B. D. O. ANDERSON AND P. V. KOKOTOVIC, *Optimal control problems over large time intervals*, Automatica J. IFAC, 23 (1987), pp. 355–363.
- [4] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces*, Wiley, Chichester, UK, 1987.
- [5] A. ARAPOSTATHIS, V. S. BORKAR, AND M. GHOSH, *Ergodic Control of Diffusion Processes*, Springer-Verlag, Berlin, 2008.
- [6] M. ARISAWA, H. ISHII, AND P.-L. LIONS, *A characterization of the existence of solutions for Hamilton–Jacobi equations in ergodic control problems with applications*, Appl. Math. Optim., 42 (2000), pp. 35–50.
- [7] Z. ARTSTEIN, *Invariant measures of differential inclusions applied to singular perturbations*, J. Differential Equations, 152 (1999), pp. 289–307.
- [8] Z. ARTSTEIN, *An occupational measure solution to a singularly perturbed optimal control problem*, Control Cybernet., 31 (2002), pp. 623–642.
- [9] Z. ARTSTEIN, *Invariant measures and their projections in nonautonomous dynamical systems*, Stoch. Dyn., 4 (2004), pp. 439–459.
- [10] J.-P. AUBIN, *Viability Theory*, Birkhäuser Boston, Boston, 1991.
- [11] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.
- [12] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser Boston, Boston, 1997.
- [13] G. K. BASAK, V. S. BORKAR, AND M. K. GHOSH, *Ergodic control of degenerate diffusions*, Stochastic Anal. Appl., 15 (1997), pp. 1–17.
- [14] A. BENSOUSSAN, *Perturbation Methods in Optimal Control*, John Wiley, New York, 1989.
- [15] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
- [16] A. G. BHATT AND V. S. BORKAR, *Occupation measures for controlled Markov processes: Characterization and optimality*, Ann. Probab., 24 (1996), pp. 1531–1562.
- [17] V. S. BORKAR AND V. GAITSGORY, *Averaging of singularly perturbed controlled stochastic differential equations*, Appl. Math. Optim., 56 (2007), pp. 169–209.
- [18] I. CAPPUZZO DOLCETTA AND H. ISHII, *Approximate solutions of the Bellman equation of deterministic control theory*, Appl. Math. Optim., 11 (1984), pp. 161–181.
- [19] D. A. CARLSON, A. B. HAURIE, AND A. LEIZAROWITZ, *Optimal Control on Infinite Time Horizon*, 2nd ed., Springer-Verlag, Berlin, 1991.
- [20] F. COLONIUS, *Optimal Periodic Control*, Lecture Notes in Math. 1313, Springer-Verlag, Berlin, 1988.
- [21] F. COLONIUS AND W. KLIEMANN, *Infinite time optimal control and periodicity*, Appl. Math. Optim., 20 (1989), pp. 113–130.
- [22] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.
- [23] T. D. DONCHEV AND A. L. DONTCHEV, *Singular perturbations in infinite-dimensional control systems*, SIAM J. Control Optim., 42 (2003), pp. 1795–1812.
- [24] L. C. EVANS AND D. GOMES, *Linear programming interpretations of Mather’s variational principle*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 693–702.

- [25] M. FALCONE, *Numerical solution of dynamic programming equations*, in *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser Boston, Boston, 1990, Appendix A.
- [26] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1991.
- [27] V. GAITSGORY, *Control of Systems with Fast and Slow Motions*, Nauka, Moscow, 1991 (in Russian).
- [28] V. GAITSGORY, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1240.
- [29] V. GAITSGORY, *On representation of the limit occupational measures set of a control system with applications to singularly perturbed control systems*, SIAM J. Control Optim., 43 (2004), pp. 325–340.
- [30] V. GAITSGORY AND A. LEIZAROWITZ, *Limit occupational measures set for a control system and averaging of singularly perturbed control systems*, J. Math. Anal. Appl., 233 (1999), pp. 461–475.
- [31] V. GAITSGORY AND M.-T. NGUYEN, *Multiscale singularly perturbed control systems: Limit occupational measures sets and averaging*, SIAM J. Control Optim., 41 (2002), pp. 954–974.
- [32] V. GAITSGORY AND S. ROSSOMAKHINE, *Linear programming approach to deterministic long run average problems of optimal control*, SIAM J. Control Optim., 44 (2006), pp. 2006–2037.
- [33] V. GAITSGORY AND S. ROSSOMAKHINE, *Occupational measures formulation and linear programming solution of deterministic long run average problems of optimal control*, in *Proceedings of 45th Annual IEEE Conference on Decision and Control*, San Diego, CA, 2006, pp. 5012–5017.
- [34] D. A. GOMES AND A. M. OBERMAN, *Computing the effective Hamiltonian using a variational approach*, SIAM J. Control Optim., 43 (2004), pp. 792–812.
- [35] G. GRAMMEL, *Averaging of Singularly Perturbed Systems*, Nonlinear Anal., 28 (1997), pp. 1851–1865.
- [36] G. GRAMMEL, *On nonlinear control systems with multiple time scales*, J. Dynam. Control Systems, 10 (2004), pp. 11–28.
- [37] L. GRUNE, *On the relation between discounted and average optimal value functions*, J. Differential Equations, 148 (1998), pp. 65–99.
- [38] K. HELMES AND R. H. STOCKBRIDGE, *Numerical comparison of controls and verification of optimality for stochastic control problems*, J. Optim. Theory Appl., 106 (2000), pp. 107–127.
- [39] O. HERNANDEZ-LERMA AND J. B. LASSERRE, *Markov Chains and Invariant Probabilities*, Birkhäuser-Verlag, Basel, 2003.
- [40] P. KASTURI AND P. DUPONT, *Constrained optimal control of vibration dampers*, J. Sound Vibration, 215 (1998), pp. 499–509.
- [41] T. G. KURTZ AND R. H. STOCKBRIDGE, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim., 36 (1998), pp. 609–653.
- [42] H. J. KUSHNER, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser Boston, Boston, 1990.
- [43] H. J. KUSHNER AND P. G. DUPUIS, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed. (revised), Springer-Verlag, New York, 2002.
- [44] J. G. LLAVONA, *Approximation of Continuously Differentiable Functions*, Math. Stud. 130, North-Holland, Amsterdam, 1986.
- [45] H. MAURER, CH. BUSKENS, AND G. FEICHTINGER, *Solution techniques for periodic control problems: A case study in production planning*, Optimal Control Appl. Methods, 19 (1998), pp. 185–203.
- [46] M. S. MENDIONDO AND R. H. STOCKBRIDGE, *Approximation of infinite-dimensional linear programming problems which arise in stochastic control*, SIAM J. Control Optim., 36 (1998), pp. 1448–1472.
- [47] M. QUINCAMPOIX AND F. WATBLED, *Averaging method for discontinuous Mayer’s problem of singularly perturbed control systems*, Nonlinear Anal., 54 (2003), pp. 819–837.
- [48] J. E. RUBIO, *Control and Optimization. The Linear Treatment of Nonlinear Problems*, Manchester University Press, Manchester, UK, 1985.
- [49] W. RUDIN, *Functional Analysis*, 2nd ed., McGraw-Hill, New York, 1991.
- [50] J. L. SPEYER, *Periodic optimal flight*, J. Guid. Control Dynam., 19 (1996), pp. 745–754.
- [51] R. H. STOCKBRIDGE, *Time-average control of a Martingale problem: Existence of a stationary solution*, Ann. Probab., 18 (1990), pp. 190–205.



- [52] R. H. STOCKBRIDGE, *Time-average control of a Martingale problem: A linear programming formulation*, Ann. Probab., 18 (1990), pp. 206–217.
- [53] T. L. VAN NOORDEN, S. M. VERDUYN LUNEL, AND A. BLIEK, *Optimization of cyclically operated reactors and separators*, Chem. Eng. Sci., 58 (2003), pp. 4115–4127.
- [54] S. VARIGONDA, T. GEORGIU, AND P. DAOUTIDIS, *A flatness based algorithm for optimal periodic control problems*, in Proceedings of the American Control Conference, Arlington, VA, 2001, pp. 831–836.
- [55] V. VELIOV, *A generalization of Tichonov theorem for singularly perturbed differential inclusions*, J. Dynam. Control Systems, 3 (1997), pp. 1–28.
- [56] A. VIGODNER, *Limits of Singularly Perturbed Control Problems: Dynamical Systems Approach*, Thesis for the Degree of Doctor of Philosophy, The Weizmann Institute of Science, Rehovot, Israel, 1995.
- [57] R. VINTER, *Convex duality and nonlinear optimal control*, SIAM J. Control Optim., 31 (1993), pp. 518–538.
- [58] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [59] Y. XIAO, D. CHENG, AND H. QIN, *Optimal impulsive control in periodic ecosystems*, Systems Control Lett., 55 (2006), pp. 558–565.
- [60] G. G. YIN AND Q. ZHANG, *Continuous-Time Markov Chains and Applications. A Singular Perturbation Approach*, Springer-Verlag, New York, 1997.
- [61] ILOG CPLEX, <http://ilog.com/products/cplex/>.

# **$L^\infty$ -NULL CONTROLLABILITY FOR THE HEAT EQUATION AND ITS CONSEQUENCES FOR THE TIME OPTIMAL CONTROL PROBLEM\***

GENGSHENG WANG<sup>†</sup>

**Abstract.** In this paper, we establish a certain  $L^\infty$ -null controllability for the internally controlled heat equation in  $\Omega \times [0, T]$ , with the control restricted to a product set of an open nonempty subset in  $\Omega$  and a subset of positive measure in the interval  $[0, T]$ . Based on this, we obtain a bang-bang principle for the time optimal control of the heat equation with controls taken from the set  $\mathcal{U}_{ad} = \{u(\cdot, t) : [0, \infty) \rightarrow L^2(\Omega) \text{ measurable; } u(\cdot, t) \in U, \text{ a.e. in } t\}$ , where  $U$  is a closed and bounded subset of  $L^2(\Omega)$ . Namely, each optimal control  $u^*(\cdot, t)$  of the problem satisfies necessarily the bang-bang property:  $u^*(\cdot, t) \in \partial U$  for almost all  $t \in [0, T^*]$ , where  $\partial U$  denotes the boundary of the set  $U$  and  $T^*$  is the optimal time. We also get the uniqueness of the optimal control when the target set  $S$  is convex and the control set  $U$  is a closed ball.

**Key words.** null controllability, time optimal control, bang-bang principle, heat equation

**AMS subject classifications.** 93C35, 93C05

**DOI.** 10.1137/060678191

**1. Introduction.** Let  $T$  be a positive number and let  $\Omega$  be a bounded domain in  $\mathbf{R}^n$ ,  $n \geq 1$ , with a  $C^\infty$ -smooth boundary. Let  $\omega$  stand for an open and nonempty subset of  $\Omega$ . Denote by  $\chi_\omega$  the characteristic function of  $\omega$ . Consider the following controlled heat equation:

$$(1.1) \quad \begin{cases} y_t(x, t) - \Delta y(x, t) = \chi_\omega(x)u(x, t) & \text{in } \Omega \times (0, T), \\ y(x, t) = 0 & \text{on } \partial\Omega \times (0, T), \\ y(x, 0) = y_0(x) & \text{in } \Omega, \end{cases}$$

where  $y_0(\cdot)$  is a function in  $L^2(\Omega)$  and  $u(x, t)$  is a control function taken from the space  $L^\infty(0, T; L^2(\Omega))$ .

In this paper, we establish the following  $L^\infty$ -null controllability for (1.1).

**THEOREM 1.1.** *Let  $T$  be a positive number and let  $E$  be a subset of positive measure in the interval  $[0, T]$ . For each  $\delta \geq 0$ , we write  $E_\delta$  for the set  $\{t \in \mathbf{R}^1; t + \delta \in E\}$  and denote by  $\chi_{E_\delta}$  the characteristic function of the set  $E_\delta$ . Then there exists a number  $\delta_0$  with  $0 < \delta_0 < T$  such that for each  $\delta$  with  $0 \leq \delta \leq \delta_0$  and for each element  $y_0$  in  $L^2(\Omega)$ , there is a control  $u_\delta$  in the space  $L^\infty(0, T - \delta; L^2(\Omega))$  such that the solution  $y^\delta$  to the following controlled heat equation:*

$$(1.2) \quad \begin{cases} y_t^\delta(x, t) - \Delta y^\delta(x, t) = \chi_{E_\delta}(t)\chi_\omega(x)u_\delta(x, t) & \text{in } \Omega \times (0, T - \delta), \\ y^\delta(x, t) = 0 & \text{on } \partial\Omega \times (0, T - \delta), \\ y^\delta(x, 0) = y_0(x) & \text{in } \Omega \end{cases}$$

\*Received by the editors December 20, 2006; accepted for publication (in revised form) February 13, 2008; published electronically June 13, 2008. This work was supported by the National Natural Science Foundation of China under grants 60574071 and 10471053, and by the key project of the Chinese Ministry of Education.

<http://www.siam.org/journals/sicon/47-4/67819.html>

<sup>†</sup>School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei, 430072, People's Republic of China (wanggs@public.wh.hb.cn).

reaches zero at time  $T - \delta$ , namely,  $y^\delta(x, T - \delta) = 0$  over  $\Omega$ . Moreover, the control  $u_\delta$  satisfies the following estimate:

$$(1.3) \quad \|u_\delta\|_{L^\infty(0, T-\delta; L^2(\Omega))}^2 \leq L \|y_0\|_{L^2(\Omega)}^2,$$

where  $L$  is a positive number independent of  $\delta$  and  $y_0$ .

Notice that we shall give an explicit expression for the constant  $L$  in section 2, from which one can see how the constant  $L$  depends on the set  $E$  and the number  $\delta_0$ .

The main motivation for such an investigation is, besides the novelty of it (to our best knowledge), its applications to the following time optimal control problems for the heat equation:

$$(P) \quad \text{Min } \{\tilde{t}; y(\cdot, \tilde{t}; u, y_0) \in S, u \in \mathcal{U}_{ad}\}.$$

Here  $y(\cdot, t; u, y_0)$  is the solution of the following controlled heat equation:

$$(1.4) \quad \begin{cases} y_t(x, t) - \Delta y(x, t) = \chi_\omega(x)u(x, t) & \text{in } \Omega \times (0, \infty), \\ y(x, t) = 0 & \text{on } \partial\Omega \times (0, \infty), \\ y(x, 0) = y_0(x) & \text{in } \Omega, \end{cases}$$

and

$$\mathcal{U}_{ad} = \{v : [0, \infty) \rightarrow L^2(\Omega) \text{ measurable; } v(\cdot, t) \in U \text{ for almost all } t \geq 0\},$$

with  $U$  being a bounded and closed subset of  $L^2(\Omega)$ , and  $S$  being a subset of  $L^2(\Omega)$ . In the problem (P), we shall call the set  $S$  the target set, the set  $U$  the control set, and the set  $\mathcal{U}_{ad}$  the control function set. For simplicity, we shall call a control function a control. We write  $T^*$  for the number  $\text{Min } \{\tilde{t}; y(\cdot, \tilde{t}; u, y_0) \in S, u \in \mathcal{U}_{ad}\}$  and call it the optimal time for the problem (P). A control  $u^*$  in the set  $\mathcal{U}_{ad}$  having the property

$$y(x, T^*; u^*, y_0) \in S$$

is called an optimal control for problem (P).

The most internally null controllability results for linear parabolic differential equations with control restricted on  $(0, T) \times \omega$  were obtained by making use of the equivalent observability inequality for the dual equations, which were derived by the Carleman inequality for the linear parabolic equation established in [9]. We can prove that Theorem 1.1 is equivalent to the following observability inequality (see the appendix):

(O) There exist positive numbers  $L$  and  $\delta_0$  with  $\delta_0 < T$  such that

$$(1.5) \quad d \int_{\Omega} (p^\delta(x, 0))^2 dx]^{\frac{1}{2}} \leq L \int_0^{T-\delta} \left\{ \int_{\Omega} [\chi_{E_\delta}(t) \chi_\omega(x) p^\delta(x, t)]^2 dx \right\}^{\frac{1}{2}} dt$$

for each number  $\delta$  with  $0 \leq \delta \leq \delta_0$  and each function  $p_T^\delta(x) \in L^2(\Omega)$ , where  $p^\delta(x, t)$  is the solution to the following adjoint equation:

$$(1.6) \quad \begin{cases} p_t^\delta(x, t) + \Delta p^\delta(x, t) = 0 & \text{in } \Omega \times (0, T - \delta), \\ p^\delta(x, t) = 0 & \text{on } \partial\Omega \times (0, T - \delta), \\ p^\delta(x, T - \delta) = p_T^\delta(x) & \text{in } \Omega. \end{cases}$$

However, since the control in this work is restricted in the set  $E \times \omega$ , where  $E$  is an arbitrary measurable set of positive measure in the interval  $(0, T)$ , but not in

the set  $(0, T) \times \omega$  as in most works on the internally null controllability of linear parabolic equations, we do not know how to prove the observability inequality **(O)** either directly or by making use of the Carleman inequality. We establish Theorem 1.1 by applying an iterative argument stimulated by that in [10] (see also [11] and [15]). The iterative argument used here is based on a sharp observability estimate on the eigenfunctions of the Laplacian due to Lebeau and Zuazua in [11] (see also [10]) and also based on a special result in the measure theory given in [14].

With regard to the equivalence of Theorem 1.1 and the observability inequality **(O)**, we cannot find the exact same results in the literature, but similar results, where the control is restricted in  $(0, T) \times \omega$ , were partially given in [8], [4], [2], and [17]; namely, the  $L^1$  observability implies the  $L^\infty$  controllability. For the sake of completeness of the paper, we give the proof in the appendix.

Based on Theorem 1.1, we obtain the following bang-bang principle and the uniqueness for the time optimal control for problem **(P)**.

**THEOREM 1.2.** *Let the target  $S$  be any subset in  $L^2(\Omega)$ . Assume that the problem **(P)** has at least a solution  $u^*$ , with the corresponding optimal time  $T^*$ . Also assume that the control set  $U$  is closed and bounded in  $L^2(\Omega)$ . Then it holds that  $u^*(t) \in \partial U$  for almost all  $t \in [0, T^*]$ . If we further assume that  $\chi_\omega U \subset U$ , then it holds that  $\chi_\omega u^*(t) \in \partial U$  for almost all  $t \in [0, T^*]$ .*

**THEOREM 1.3.** *Suppose that the target set  $S$  is convex and the control set  $U$  is a closed ball. Then the optimal control of problem **(P)** is unique.*

Combining Theorem 1.3 and the existence results for the time optimal control obtained in [17] (see also [20] and [22]), we can have the following unique existence result for problem **(P)**.

**THEOREM 1.4.** *Suppose that the target set  $S$  is a closed and convex subset, which contains the origin of  $L^2(\Omega)$ , and the control set  $U$  is a closed ball in  $L^2(\Omega)$ . Then there exists a unique time optimal control for problem **(P)**.*

To my best knowledge, the observation that the bang-bang principle of time optimal control of the linear parabolic equations is related to a null controllability of the equation with the control restricted to a subset of positive measure in the time interval  $[0, T]$  was first given in [16]. In [16], the authors established such a controllability result for the one-dimensional boundary controlled heat equation. For other works related to the time optimal control problems, we would like to mention [5], [6], [14], [7], [18], [19], [12], [23], [13], [22], [20], [17], and [21].

The rest of the paper is organized as follows. In section 2, we give the proof of Theorem 1.1. In section 3, we prove Theorems 1.2 and 1.3 and also give some consequences of these theorems.

**2. The null controllability (C).** In this section, we shall prove Theorem 1.1. The proof is based on a sharp estimate on the eigenfunctions of the Laplacian due to Lebeau and Zuazua (see [11]) and also based on a fundamental result in the measure theory, which will be given later. Let  $\{\lambda_i\}_{i=1}^\infty$ ,  $0 < \lambda_1 < \lambda_2 \leq \dots$ , be the eigenvalues of  $-\Delta$  with the Dirichlet boundary condition, and let  $\{X_i(x)\}_{i=1}^\infty$  be the corresponding eigenfunctions, which serve as an orthonormal basis of  $L^2(\Omega)$ . Then we have the following result (see [11]).

**THEOREM 2.1.** *Let  $\Omega$  be a bounded domain in  $\mathbf{R}^n$ ,  $n \geq 1$ , with a  $C^\infty$ -smooth boundary. Then there exist two positive constants  $C_1, C_2 > 0$  such that*

$$(2.1) \quad \sum_{\lambda_i \leq r} |a_i|^2 \leq C_1 e^{C_2 \sqrt{r}} \int_\omega \left| \sum_{\lambda_i \leq r} a_i X_i(x) \right|^2 dx$$

for every finite  $r > 0$  and every choice of the coefficients  $\{a_i\}_{\lambda_i \leq r}$  with  $a_i \in \mathbf{R}^1$ .

Notice that the reason we assume the domain  $\Omega$  has a  $C^\infty$ -smooth boundary is because Theorem 2.1 needs the domain  $\Omega$  to have such a property and the proof of Theorem 1.1 is based on Theorem 2.1.

Now, we shall first use Theorem 2.1 to derive a certain controllability result, which will help us in the proof of Theorem 1.1. For each  $r > 0$ , we set  $\mathbf{X}_r = \text{span} \{X_i(x)\}_{\lambda_i \leq r}$  and consider the following dual equation:

$$(2.2) \quad \begin{cases} \varphi_t(x, t) + \Delta \varphi(x, t) = 0 & \text{in } \Omega \times (0, T), \\ \varphi(x, t) = 0 & \text{on } \partial\Omega \times (0, T), \\ \varphi(x, T) \in \mathbf{X}_r. \end{cases}$$

Here, each element  $\varphi(x, T)$  in  $\mathbf{X}_r$  can be written as

$$\varphi(x, T) = \sum_{\lambda_i \leq r} a_i X_i(x)$$

for a certain sequence of real numbers  $\{a_i\}_{\lambda_i \leq r}$ . Then the solution  $\varphi(x, t)$  to (2.2) can be expressed by

$$\varphi(x, t) = \sum_{\lambda_i \leq r} a_i e^{-\lambda_i(T-t)} X_i(x) \quad \text{for all } t \in [0, T].$$

Set  $b_i(t) = a_i e^{-\lambda_i(T-t)}$ ,  $t \in [0, T]$ . Then by (2.1), we have

$$\begin{aligned} \sum_{\lambda_i \leq r} |b_i(t)|^2 &\leq C_1 e^{C_2 \sqrt{r}} \int_{\omega} \left| \sum_{\lambda_i \leq r} b_i(t) X_i(x) \right|^2 dx \\ &= C_1 e^{C_2 \sqrt{r}} \int_{\omega} |\varphi(x, t)|^2 dx \quad \text{for all } t \in [0, T]. \end{aligned}$$

On the other hand,

$$\begin{aligned} \sum_{\lambda_i \leq r} |b_i(t)|^2 &= \sum_{\lambda_i \leq r} a_i^2 e^{-2\lambda_i(T-t)} \geq \sum_{\lambda_i \leq r} a_i^2 e^{-2\lambda_i T} \\ &= \int_{\Omega} \varphi^2(x, 0) dx \quad \text{for all } t \in [0, T]. \end{aligned}$$

Hence,

$$\int_{\Omega} \varphi^2(x, 0) dx \leq C_1 e^{C_2 \sqrt{r}} \int_{\omega} |\varphi(x, t)|^2 dx \quad \text{for all } t \in [0, T],$$

or equivalently,

$$\left[ \int_{\Omega} \varphi^2(x, 0) dx \right]^{\frac{1}{2}} \leq (C_1 e^{C_2 \sqrt{r}})^{\frac{1}{2}} \left[ \int_{\omega} |\varphi(x, t)|^2 dx \right]^{\frac{1}{2}} \quad \text{for all } t \in [0, T],$$

from which it follows that

$$\int_E \left[ \int_{\Omega} \varphi^2(x, 0) dx \right]^{\frac{1}{2}} dt \leq (C_1 e^{C_2 \sqrt{r}})^{\frac{1}{2}} \int_E \left[ \int_{\omega} |\varphi(x, t)|^2 dx \right]^{\frac{1}{2}} dt.$$

Namely, we obtained that for each  $\varphi(\cdot, T) \in \mathbf{X}_r$ ,

$$(2.3) \quad \begin{aligned} \int_{\Omega} \varphi^2(x, 0) dx &\leq \frac{C_1 e^{C_2 \sqrt{r}}}{(m(E))^2} \left\{ \int_0^T \left[ \int_{\Omega} |\chi_E(t) \chi_{\omega}(x) \varphi(x, t)|^2 dx \right]^{\frac{1}{2}} dt \right\}^2 \\ &= \frac{C_1 e^{C_2 \sqrt{r}}}{(m(E))^2} \|\chi_E \chi_{\omega} \varphi\|_{L^1(0, T; L^2(\Omega))}^2. \end{aligned}$$

Write  $P_r$  for the orthogonal projection from  $L^2(\Omega)$  to  $\mathbf{X}_r$ . We next use (2.3) to obtain the following controllability result.

LEMMA 2.2. *For each  $r > 0$ , there exists a control  $u_r$  in the space  $L^\infty(0, T; L^2(\Omega))$  with the estimate*

$$(2.4) \quad \|u_r\|_{L^\infty(0, T; L^2(\Omega))} \leq \frac{C_1 e^{C_2 \sqrt{r}}}{(m(E))^2} \|y_0\|_{L^2(\Omega)}^2$$

such that  $P_r(y(\cdot, T)) = 0$ , where  $y(x, t)$  is the solution of (1.2) with  $\delta = 0$  and  $u = u_r$ , and where  $C_1$  and  $C_2$  are the positive constants given in Theorem 2.1.

*Proof.* Let  $y(x, t)$  be the solution of (2.1) with  $\delta = 0$ , and let  $\varphi(x, t)$  be a solution of (2.2). Then

$$\langle y(\cdot, T), \varphi(\cdot, T) \rangle - \langle y_0(\cdot), \varphi(\cdot, 0) \rangle = \int_0^T \int_{\Omega} \chi_E(t) \chi_{\omega}(x) u(x, t) \varphi(x, t) dx dt.$$

Here and in what follows,  $\langle \cdot, \cdot \rangle$  denotes the inner product in  $L^2(\Omega)$ . If we can show that  $\langle y(\cdot, T), \varphi(\cdot, T) \rangle = 0$  for all  $\varphi(x, T) \in \mathbf{X}_r$ , then  $P_r(y(\cdot, T)) = 0$ . Thus, it suffices to prove that there exists a control  $u_r \in L^\infty(0, T; L^2(\Omega))$  with the estimate (2.4) such that

$$-\langle y_0(\cdot), \varphi(\cdot, 0) \rangle = \int_0^T \int_{\Omega} \chi_E(t) \chi_{\omega}(x) u_r(x, t) \varphi(x, t) dx dt \quad \text{for all } \varphi(\cdot, T) \in \mathbf{X}_r.$$

Now, we set

$$\mathbf{Y}_r = \{\chi_E(t) \chi_{\omega}(x) \varphi(x, t); \varphi(x, t) \text{ is the solution to (2.2)}\}.$$

It is clear that  $\mathbf{Y}_r$  is a linear subspace of  $L^1(0, T; L^2(\Omega))$ . We define a linear functional  $F_r : \mathbf{Y}_r \rightarrow \mathbf{R}^1$  by  $F_r(\chi_E \chi_{\omega} \varphi) = -\langle y_0(\cdot), \varphi(\cdot, 0) \rangle$ . By inequality (2.3), we see that

$$\begin{aligned} |F_r(\chi_E \chi_{\omega} \varphi)|^2 &\leq \|y_0\|_{L^2(\Omega)}^2 \cdot \|\varphi(\cdot, 0)\|_{L^2(\Omega)}^2 \\ &\leq \frac{C_1 e^{C_2 \sqrt{r}}}{(m(E))^2} \|y_0\|_{L^2(\Omega)}^2 \cdot \|\chi_E \chi_{\omega} \varphi\|_{L^1(0, T; L^2(\Omega))}^2. \end{aligned}$$

Namely,

$$\|F_r\|_{L(\mathbf{Y}_r, \mathbf{R}^1)}^2 \leq \frac{C_1 e^{C_2 \sqrt{r}}}{(m(E))^2} \|y_0\|_{L^2(\Omega)}^2,$$

where  $\|F_r\|_{L(\mathbf{Y}_r, \mathbf{R}^1)}^2$  denotes the operator norm of  $F_r$  from the subspace  $\mathbf{Y}_r$ , endowed with the norm of  $L^1(0, T; L^2(\Omega))$ , into the space  $\mathbf{R}^1$ . Thus,  $F_r$  is a bounded linear functional on  $\mathbf{Y}_r$ . By the Hahn–Banach theorem, there is a bounded linear functional

$$G_r : L^1(0, T; L^2(\Omega)) \rightarrow \mathbf{R}^1$$

such that

$$G_r = F_r \text{ on } \mathbf{Y}_r,$$

and such that

$$\|G_r\|_{L(L^1(0,T;L^2(\Omega));\mathbf{R}^1)}^2 = \|F_r\|_{L(\mathbf{Y}_r;R^1)}^2 \leq \frac{C_1 e^{C_2 \sqrt{r}}}{(m(E))^2} \|y_0\|_{L^2(\Omega)}^2.$$

Here  $\|G_r\|_{L(L^1(0,T;L^2(\Omega));\mathbf{R}^1)}$  denotes the operator norm of  $G_r$  from the space  $L^1(0,T;L^2(\Omega))$  to  $\mathbf{R}^1$ .

Then, by making use of the Riesz representation theorem (see, [3, p. 59]), there exists a function  $u_r$  in the space  $L^\infty(0,T;L^2(\Omega))$  such that

$$G_r(f) = \int_0^T \int_\Omega f u_r dx dt \quad \text{for all } f \in L^1(0,T;L^2(\Omega)),$$

and such that

$$\|u_r\|_{L^\infty(0,T;L^2(\Omega))}^2 = \|G_r\|_{L(L^1(0,T;L^2(\Omega));\mathbf{R}^1)}^2 \leq \frac{C_1 e^{C_2 \sqrt{r}}}{(m(E))^2} \|y_0\|_{L^2(\Omega)}^2.$$

In particular,

$$F_r(\chi_E \chi_\omega \varphi) = \int_0^T \int_\Omega \chi_E \chi_\omega \varphi u_r dx dt \quad \text{for all } \chi_E \chi_\omega \varphi \in \mathbf{Y}_r.$$

Namely,

$$-\langle y_0(\cdot), \varphi(\cdot, 0) \rangle = \int_0^T \int_\Omega \chi_E \chi_\omega \varphi u_r dx dt \quad \text{for all } \varphi(\cdot, T) \in \mathbf{X}_r.$$

This completes the proof.  $\square$

The following lemma from the measure theory will be used in our later discussion, whose proof can be found in [14, pp. 256–257].

LEMMA 2.3. *For almost all  $t$  in the set  $E$ , there exists a sequence of numbers  $\{t_i\}_{i=1}^\infty$  in the interval  $[0, T]$  such that*

$$(2.5) \quad t_1 < \cdots < t_i < t_{i+1} < \cdots < \tilde{t}, \quad t_i \rightarrow \tilde{t} \text{ as } i \rightarrow \infty,$$

$$(2.6) \quad m(E \cap [t_i, t_{i+1}]) \geq \rho(t_{i+1} - t_i), \quad i = 1, 2, \dots,$$

and

$$(2.7) \quad \frac{t_{i+1} - t_i}{t_{i+2} - t_{i+1}} \leq C_0, \quad i = 1, 2, \dots,$$

where  $\rho$  and  $C_0$  are two positive constants which depend on the set  $E$ .

Before proceeding with the proof of Theorem 1.1, we introduce briefly our main strategy. By applying Lemma 2.3, there exist a number  $\tilde{t}$  and a sequence  $\{t_N\}_{N=1}^\infty$  in the interval  $(0, T)$  such that (2.5)–(2.7) hold. The main part of the proof is to show that for each  $\tilde{y}_0$  in  $L^2(\Omega)$ , there exists a control  $\tilde{u}$  in the space  $L^\infty(t_1, \tilde{t}; L^2(\Omega))$

with the estimate  $\|\tilde{u}\|_{L^\infty(t_1, \tilde{t}; L^2(\Omega))}^2 \leq L \|\tilde{y}_0\|_{L^2(\Omega)}^2$  for a certain positive constant  $L$  independent of  $\tilde{y}_0$  such that the solution  $\tilde{y}(x, t)$  to the equation

$$(2.8) \quad \begin{cases} \tilde{y}_t(x, t) - \Delta \tilde{y}(x, t) = \chi_E(t) \chi_\omega(x) \tilde{u}(x, t) & \text{in } \Omega \times (t_1, \tilde{t}), \\ \tilde{y}(x, t) = 0 & \text{on } \partial\Omega \times (t_1, \tilde{t}), \\ \tilde{y}(x, t_1) = \tilde{y}_0(x) & \text{in } \Omega \end{cases}$$

has zero value at time  $\tilde{t}$ , namely,  $\tilde{y}(x, \tilde{t}) = 0$  over  $\Omega$ . To this end, we write

$$[t_1, \tilde{t}] = \bigcup_{N=1}^{\infty} (I_N \cup J_N),$$

where  $I_N = [t_{2N-1}, t_{2N}]$  and  $J_N = [t_{2N}, t_{2N+1}]$ ,  $N = 1, 2, \dots$ . Then we choose a suitable sequence of positive numbers  $\{r_N\}_{N=1}^\infty$  having the following properties:

- (a)  $r_1 < r_2 < \dots < r_N < \dots$ ,
- (b)  $r_N \rightarrow \infty$  as  $N \rightarrow \infty$ .

On the subinterval  $I_N$ , we control the heat equation with a control  $u_N$  restricted on the subdomain  $\omega \times (I_N \cap E)$  such that  $P_{r_N}(y_N(\cdot, t_{2N})) = 0$ , where  $P_{r_N}$  denotes the orthogonal projection from  $L^2(\Omega)$  onto  $\text{span} \{X_i(x)\}_{i=1}^{r_N}$ . On the subinterval  $J_N$ , we let the heat equation freely evolve. We start by having the initial data for the equation on  $I_1$  be  $y_0$ . For the initial data on  $I_N$ ,  $N = 2, 3, \dots$ , we define it to be the ending value of the solution for the equation on  $J_{N-1}$ . The initial data of the equation on  $J_N$ ,  $N = 1, 2, \dots$ , is given by the ending value of the solution for the equation on  $I_N$ . Moreover, by making use of Lemmas 2.2 and 2.3, we will show that there is a sequence  $\{r_N\}_{N=1}^\infty$ , having the properties (a) and (b) as above, such that the  $L^\infty(I_N; L^2(\Omega))$ -norm of the control  $u_N$  is bounded by  $L^{\frac{1}{2}} \|\tilde{y}_0\|_{L^2(\Omega)}$  for a certain positive constant  $L$  independent of  $N$  and  $\tilde{y}_0$ . Then, we construct a control  $\tilde{u}$  by setting

$$\tilde{u}(x, t) = \begin{cases} u_N(x, t), & x \in \Omega, t \in I_N, N = 1, 2, \dots, \\ 0, & x \in \Omega, t \in J_N, N = 1, 2, \dots \end{cases}$$

We can show that this control  $\tilde{u}$  makes the corresponding trajectory  $\tilde{y}$  of (2.8) have zero value at time  $\tilde{t}$ .

Now, we set

$$u(x, t) = \begin{cases} \tilde{u}(x, t) & \text{in } \Omega \times (t_1, \tilde{t}), \\ 0 & \text{in } \Omega \times ((0, T) \setminus (t_1, \tilde{t})) \end{cases}$$

and take  $\tilde{y}_0$  to be  $\psi(x, t_1)$ , where  $\psi(x, t)$  is the solution of the heat equation on  $\Omega \times (0, t_1)$  with the initial data  $y_0$ . Then it is clear that this control  $u$  makes the trajectory  $y(x, t)$  of (1.2) with  $\delta = 0$  have zero value at time  $T$ . Moreover,  $\|u\|_{L^\infty(0, T; L^2(\Omega))}^2 \leq L \|y_0\|_{L^2(\Omega)}^2$ .

We next replace the sequence  $\{t_N\}_{N=1}^\infty$  and the number  $\tilde{t}$  by the sequence  $\{t_N - \delta\}_{N=1}^\infty$  and the number  $(\tilde{t} - \delta)$ , respectively, where the number  $\delta$  is such that  $0 \leq \delta \leq t_1$ . Then by making use of the same argument as above, we obtain that for each number  $\delta$  with  $0 \leq \delta \leq t_1$ , there exists a control  $u_\delta$  in the space  $L^\infty(0, T - \delta; L^2(\Omega))$  with the estimate  $\|u_\delta\|_{L^\infty(0, T - \delta; L^2(\Omega))}^2 \leq L_\delta \|y_0\|_{L^2(\Omega)}^2$  for a certain positive number  $L_\delta$ .



independent of  $y_0$ , such that the corresponding solution  $y^\delta$  to (1.2) reaches zero value at time  $T - \delta$ , namely,  $y^\delta(x, T - \delta) = 0$  over  $\Omega$ . We finally prove that  $L_\delta = L$  is independent of  $\delta$ .

Now we prove Theorem 1.1.

*Proof of Theorem 1.1.* Without loss of generality, we can assume that  $C_1 \geq 1$ , where  $C_1$  is the positive constant given in Theorem 2.1. By making use of Lemma 2.3, we can take a number  $\tilde{t}$  in the set  $E$  with  $\tilde{t} < T$  and a sequence  $\{t_N\}_{N=1}^\infty$  in the open interval  $(0, T)$  such that (2.5)–(2.7) hold for certain positive numbers  $\rho$  and  $C_0$  and such that

$$\tilde{t} - t_1 \leq \text{Min}\{\lambda_1, 1\}.$$

We shall first prove that for each  $\tilde{y}_0$  in  $L^2(\Omega)$ , there exists a control  $\tilde{u}$  in the space  $L^\infty(t_1, \tilde{t}; L^2(\Omega))$  with the estimate  $\|\tilde{u}\|_{L^\infty(t_1, \tilde{t}; L^2(\Omega))}^2 \leq L \|\tilde{y}_0\|_{L^2(\Omega)}^2$  for a certain positive constant  $L$  independent of  $\tilde{y}_0$ , such that the solution  $\tilde{y}$  to (2.8) reaches zero value at time  $\tilde{t}$ , namely,  $\tilde{y}(x, \tilde{t}) = 0$  over  $\Omega$ .

To this end, we shall use the strategy presented above. We set  $I_N = [t_{2N-1}, t_{2N}]$ ,  $J_N = [t_{2N}, t_{2N+1}]$  for  $N = 1, 2, \dots$ . Then

$$[t_1, \tilde{t}) = \bigcup_{N=1}^{\infty} (I_N \cup J_N).$$

Notice that for each  $N \geq 1$ , it holds that  $m(E \cap I_N) > 0$ .

Now, on the interval  $I_1 \equiv [t_1, t_2]$ , we consider the following controlled heat equation:

$$\begin{cases} y_1'(x, t) - \Delta y_1(x, t) = \chi_E(t) \chi_\omega(x) u_1(x, t) & \text{in } \Omega \times (t_1, t_2), \\ y_1(x, t) = 0 & \text{on } \partial\Omega \times (t_1, t_2), \\ y_1(x, t_1) = \tilde{y}_0(x) & \text{in } \Omega. \end{cases}$$

By Lemma 2.2, for any  $r_1 > 0$ , there exists a control  $u_1$  in the space  $L^\infty(t_1, t_2; L^2(\Omega))$  with the estimate

$$\|u_1\|_{L^\infty(t_1, t_2; L^2(\Omega))}^2 \leq \frac{C_1 e^{C_2 \sqrt{r_1}}}{(m(E \cap [t_1, t_2]))^2} \|\tilde{y}_0\|_{L^2(\Omega)}^2,$$

such that  $P_{r_1}(y_1(\cdot, t_2)) = 0$ . Then, by (2.6) and (2.7) in Lemma 2.3, we see that

$$\begin{aligned} \|u_1\|_{L^\infty(t_1, t_2; L^2(\Omega))}^2 &\leq \frac{C_1 e^{C_2 \sqrt{r_1}}}{\rho^2(t_2 - t_1)^2} \|\tilde{y}_0\|_{L^2(\Omega)}^2 \\ &\equiv \frac{C_1}{\rho^2(t_2 - t_1)^2} \cdot \alpha_1 \|\tilde{y}_0\|_{L^2(\Omega)}^2, \end{aligned}$$

where  $\alpha_1 = e^{C_2 \sqrt{r_1}}$ . Moreover, we have

$$\begin{aligned} \|y_1(\cdot, t_2)\|_{L^2(\Omega)}^2 &\leq \|y_1(\cdot, t_1)\|_{L^2(\Omega)}^2 + \frac{1}{\lambda_1} \int_{t_1}^{t_2} \|u_1(\cdot, s)\|_{L^2(\Omega)}^2 ds \\ &\leq \|\tilde{y}_0\|_{L^2(\Omega)}^2 + \frac{(t_2 - t_1)}{\lambda_1} \|u_1\|_{L^\infty(t_1, t_2; L^2(\Omega))}^2 \\ &\leq 2 \frac{C_1}{\rho^2(t_2 - t_1)^2} \cdot \alpha_1 \|\tilde{y}_0\|_{L^2(\Omega)}^2. \end{aligned}$$

Here we have used the facts that  $(t_2 - t_1) \leq \text{Min}(\lambda_1, 1)$ ,  $\rho < 1$ , and  $C_1 > 1$ .

On the interval  $J_1 \equiv [t_2, t_3]$ , we consider the following heat equation without control:

$$\begin{cases} z_1'(x, t) - \Delta z_1(x, t) = 0 & \text{in } \Omega \times (t_2, t_3), \\ z_1(x, t) = 0 & \text{on } \partial\Omega \times (t_2, t_3), \\ z_1(x, t_2) = y_1(x, t_2) & \text{in } \Omega. \end{cases}$$

Since  $P_{r_1}(y_1(\cdot, t_2)) = 0$ , we have

$$\begin{aligned} \|z_1(\cdot, t_3)\|_{L^2(\Omega)}^2 &\leq \exp(-2r_1(t_3 - t_2)) \cdot \|y_1(\cdot, t_2)\|_{L^2(\Omega)}^2 \\ &\leq 2 \frac{C_1}{\rho^2(t_2 - t_1)^2} \alpha_1 \cdot \exp(-2r_1(t_3 - t_2)) \cdot \|\tilde{y}_0\|_{L^2(\Omega)}^2. \end{aligned}$$

On the interval  $I_2 \equiv [t_3, t_4]$ , we consider the controlled heat equation as follows:

$$\begin{cases} y_2'(x, t) - \Delta y_2(x, t) = \chi_E(t)\chi_\omega(x)u_2(x, t) & \text{in } \Omega \times (t_3, t_4), \\ y_2(x, t) = 0 & \text{on } \partial\Omega \times (t_3, t_4), \\ y_2(x, t_3) = z_1(x, t_3) & \text{in } \Omega. \end{cases}$$

Then by Lemma 2.2, for any  $r_2 > 0$ , there exists a control  $u_2$  in the space  $L^\infty(t_3, t_4; L^2(\Omega))$  with the estimate

$$\|u_2\|_{L^\infty(t_3, t_4; L^2(\Omega))}^2 \leq \frac{C_1 e^{C_2 \sqrt{r_2}}}{m(E \cap [t_3, t_4])^2} \cdot \|z_1(\cdot, t_3)\|_{L^2(\Omega)}^2,$$

such that  $P_{r_2}(y_2(\cdot, t_4)) = 0$ . By (2.6) and (2.7) in Lemma 2.3, we get

$$\|u_2\|_{L^\infty(t_3, t_4; L^2(\Omega))}^2 \leq 2 \left( \frac{C_1}{\rho^2(t_2 - t_1)^2} \right)^2 C_0^4 \cdot \alpha_1 \cdot \alpha_2 \cdot \|\tilde{y}_0\|_{L^2(\Omega)}^2,$$

where  $\alpha_2 = \exp(C_2 \sqrt{r_2}) \exp(-2r_1(t_3 - t_2))$ . Moreover, it holds that

$$\begin{aligned} \|y_2(\cdot, t_4)\|_{L^2(\Omega)}^2 &\leq \|z_1(\cdot, t_3)\|_{L^2(\Omega)}^2 + \frac{1}{\lambda_1}(t_4 - t_3) \|u_2\|_{L^\infty(t_3, t_4; L^2(\Omega))}^2 \\ &\leq 2^2 \left( \frac{C_1}{\rho^2(t_2 - t_1)^2} \right)^2 C_0^4 \cdot \alpha_1 \cdot \alpha_2 \cdot \|\tilde{y}_0\|_{L^2(\Omega)}^2. \end{aligned}$$

On the interval  $J_2 \equiv [t_4, t_5]$ , we consider the following heat equation without control:

$$\begin{cases} z_2'(x, t) - \Delta z_2(x, t) = 0 & \text{in } \Omega \times (t_4, t_5), \\ z_2(x, t) = 0 & \text{on } \partial\Omega \times (t_4, t_5), \\ z_2(x, t_4) = y_2(x, t_4) & \text{in } \Omega. \end{cases}$$

Since  $P_{r_2}(y_2(\cdot, t_4)) = 0$ , we have

$$\begin{aligned} \|z_2(\cdot, t_5)\|_{L^2(\Omega)}^2 &\leq \exp(-2r_2(t_5 - t_4)) \|y_2(\cdot, t_4)\|_{L^2(\Omega)}^2 \\ &\leq 2^2 \left( \frac{C_1}{\rho^2(t_2 - t_1)^2} \right)^2 C_0^4 \cdot \alpha_1 \cdot \alpha_2 \cdot \|\tilde{y}_0\|_{L^2(\Omega)}^2 \cdot \exp(-2r_2(t_5 - t_4)). \end{aligned}$$

On the interval  $I_3 \equiv [t_5, t_6]$ , we consider the following controlled heat equation:

$$\begin{cases} y'_3(x, t) - \Delta y_3(x, t) = \chi_E(t) \chi_\omega(x) u_3(x, t) & \text{in } \Omega \times (t_5, t_6), \\ y_3(x, t) = 0 & \text{on } \partial\Omega \times (t_5, t_6), \\ y_3(x, t_5) = z_2(x, t_5) & \text{in } \Omega. \end{cases}$$

Then by Lemma 2.2, for any  $r_3 > 0$ , there exists a control  $u_3$  in the space  $L^\infty(t_5, t_6; L^2(\Omega))$  with the estimate

$$\|u_3\|_{L^\infty(t_5, t_6; L^2(\Omega))}^2 \leq \frac{C_1 e^{C_2 \sqrt{r_3}}}{(m(E \cap [t_5, t_6]))^2} \|z_2(\cdot, t_5)\|_{L^2(\Omega)}^2,$$

such that  $P_{r_3}(y_3(\cdot, t_6)) = 0$ . By making use of (2.6) and (2.7) again, we get

$$\|u_3\|_{L^\infty(t_5, t_6; L^2(\Omega))}^2 \leq 2^2 \left( \frac{C_1}{\rho^2(t_2 - t_1)^2} \right)^3 C_0^4 \cdot C_0^{4 \cdot 2} \cdot \alpha_1 \cdot \alpha_2 \cdot \alpha_3 \cdot \|\tilde{y}_0\|_{L^2(\Omega)}^2,$$

where  $\alpha_3 = \exp(C_2 \sqrt{r_3}) \exp(-2r_2(t_3 - t_2)C_0^{-2})$ .

Generally, on the interval  $I_N$ , we consider the controlled heat equation

$$\begin{cases} y'_N(x, t) - \Delta y_N(x, t) = \chi_E(t) \chi_\omega(x) u_N(x, t) & \text{in } \Omega \times (t_{2N-1}, t_{2N}), \\ y_N(x, t) = 0 & \text{on } \partial\Omega \times (t_{2N-1}, t_{2N}), \\ y_N(x, t_{2N-1}) = z_{N-1}(x, t_{2N-1}) & \text{in } \Omega. \end{cases}$$

On the interval  $J_N$ , we consider the following heat equation without control:

$$\begin{cases} z'_N(x, t) - \Delta z_N(x, t) = 0 & \text{in } \Omega \times (t_{2N}, t_{2N+1}), \\ z_N(x, t) = 0 & \text{on } \partial\Omega \times (t_{2N}, t_{2N+1}), \\ z_N(x, t_{2N}) = y_N(x, t_{2N}) & \text{in } \Omega. \end{cases}$$

Then by making use of the induction argument, we can obtain the following: *For each  $r_N > 0$ , there exists a control  $u_N$  in the space  $L^\infty(I_N; L^2(\Omega))$  with the following estimate:*

$$\begin{aligned} & \|u_N\|_{L^\infty(I_N; L^2(\Omega))}^2 \\ & \leq 2^{N-1} \left( \frac{C_1}{\rho^2(t_2 - t_1)^2} \right)^N C_0^4 \cdot C_0^{4 \cdot 2} \cdots C_0^{4(N-1)} \cdot \alpha_1 \cdot \alpha_2 \cdots \alpha_N \cdot \|\tilde{y}_0\|_{L^2(\Omega)}^2, \end{aligned}$$

where

$$(2.9) \quad \alpha_N = \begin{cases} \exp(C_2 \sqrt{r_1}), & N = 1, \\ \exp(C_2 \sqrt{r_N}) \exp(-2r_{N-1}(t_3 - t_2)C_0^{-2(N-2)}), & N \geq 2, \end{cases}$$

such that  $P_{r_N}(y_N(\cdot, t_{2N})) = 0$ . It is easily seen that for each  $N \geq 1$ ,

$$(2.10) \quad \|u_N\|_{L^\infty(I_N; L^2(\Omega))}^2 \leq (\tilde{C})^{N(N-1)} \alpha_1 \cdots \alpha_N \cdot \|\tilde{y}_0\|_{L^2(\Omega)}^2,$$

where

$$(2.11) \quad \tilde{C} = \frac{2C_1}{\rho^2(t_2 - t_1)^2} \cdot C_0^2.$$

Now, we set

$$(2.12) \quad r_N = \left[ \frac{2}{(t_3 - t_2)} \tilde{C}^{N-1} \right]^4 \equiv [A \cdot \tilde{C}^{N-1}]^4, \quad N \geq 1.$$

Because we have  $\tilde{C} > C_0^2 > 1$  and  $t_3 - t_2 < 1$ , it holds that

$$2^4 < r_1 < r_2 < \cdots < r_N < r_{N+1} < \cdots \quad \text{and} \quad r_N \rightarrow \infty \quad \text{as} \quad N \rightarrow \infty.$$

Moreover, we have

$$r_{N-1}^{\frac{1}{4}}(t_3 - t_2)C_0^{-2(N-2)} \geq 2 \quad \text{for each } N \geq 2.$$

Then we get

$$(2.13) \quad \exp \{ -2r_{N-1}(t_3 - t_2)C_0^{-2(N-2)} \} \leq \exp \left( -4r_{N-1}^{\frac{3}{4}} \right) \quad \text{for each } N \geq 2.$$

Since

$$\begin{aligned} \tilde{C}^{N(N-1)} \exp \left( -r_{N-1}^{\frac{3}{4}} \right) &= \frac{\tilde{C}^{N(N-1)}}{(\exp(r_{N-1}^{\frac{1}{4}}))^{r_{N-1}^{\frac{1}{2}}}} \leq \frac{\tilde{C}^{N(N-1)}}{(\exp(2\tilde{C}^{N-1}))^{r_{N-1}^{\frac{1}{2}}}} \\ &\leq \frac{\tilde{C}^{N(N-1)}}{\tilde{C}^{(N-1) \cdot 2 \cdot r_{N-1}^{\frac{1}{2}}}}, \end{aligned}$$

for each  $N \geq 2$ , we derive from (2.12) that there exists a natural number  $N_1$  with  $N_1 \geq 2$  such that for each  $N \geq N_1$ ,

$$(2.14) \quad \tilde{C}^{N(N-1)} \exp \left( -r_{N-1}^{\frac{3}{4}} \right) \leq 1.$$

By making use of (2.12) again, we obtain that for each  $N \geq 2$ ,

$$\exp(C_2 \sqrt{r_N}) \exp \left( -r_{N-1}^{\frac{3}{4}} \right) = \exp(C_2 A^2 \tilde{C}^{2(N-1)}) \exp \left( -A^3 \tilde{C}^{3(N-2)} \right).$$

Thus, there exists a natural number  $N_2$  with  $N_2 \geq 2$  such that for each  $N \geq N_2$ ,

$$(2.15) \quad \exp(C_2 \sqrt{r_N}) \exp \left( -r_{N-1}^{\frac{3}{4}} \right) \leq 1.$$

Now we set

$$(2.16) \quad N_0 = \max \{N_1, N_2\}.$$

Then by (2.13), (2.14), and (2.15), we see that for all  $N \geq N_0$ ,

$$\begin{aligned} (2.17) \quad &\tilde{C}^{N(N-1)} \alpha_N \\ &= \tilde{C}^{N(N-1)} \exp(C_2 \sqrt{r_N}) \exp \left( -2r_{N-1}(t_3 - t_2)C_0^{-2(N-2)} \right) \\ &\leq \tilde{C}^{N(N-1)} \exp(C_2 \sqrt{r_N}) \exp \left( -4r_{N-1}^{\frac{3}{4}} \right) \\ &\leq \exp \left( -2r_{N-1}^{\frac{3}{4}} \right). \end{aligned}$$

Moreover, it is obvious that

$$(2.18) \quad \alpha_N \leq 1 \quad \text{for all } N \geq N_0.$$

Now, we set

$$(2.19) \quad L = \max \{ (\tilde{C})^{N(N-1)} \alpha_1 \cdots \alpha_N, \quad 1 \leq N \leq N_0 \}.$$

It follows from (2.10), (2.17), (2.18), and (2.19) that for all  $N \geq 1$ ,

$$(2.20) \quad \|u_N\|_{L^\infty(I_N; L^2(\Omega))}^2 \leq L \|\tilde{y}_0\|_{L^2(\Omega)}^2.$$

Then we construct a control  $\tilde{u}$  by setting

$$(2.21) \quad \tilde{u}(x, t) = \begin{cases} u_N(x, t), & x \in \Omega, \quad t \in I_N, \quad N \geq 1, \\ 0, & x \in \Omega, \quad t \in J_N, \quad N \geq 1. \end{cases}$$

From (2.20) and (2.21), we easily see that the control  $\tilde{u}$  is in the space  $L^\infty(t_1, \tilde{t}; L^2(\Omega))$  and satisfies the estimate

$$\|\tilde{u}\|_{L^\infty(t_1, \tilde{t}; L^2(\Omega))}^2 \leq L \|\tilde{y}_0\|_{L^2(\Omega)}^2.$$

Let  $\tilde{y}$  be the solution of (2.8) corresponding to the control  $\tilde{u}$  constructed in (2.21). Then on the interval  $I_N$ ,  $\tilde{y}(\cdot, t) = y_N(\cdot, t)$ . Since  $P_{r_N}(y_N(\cdot, t_{2N})) = 0$  for all  $N \geq 1$  and  $r_1 < r_2 < \cdots < r_N < \cdots$ , by making use of (2.21) again, we see that

$$(2.22) \quad P_{r_N}(\tilde{y}(\cdot, t_{2M})) = 0 \quad \text{for all } M \geq N.$$

On the other hand, since  $t_{2M} \rightarrow \tilde{t}$  as  $M \rightarrow \infty$ , we obtain that

$$\tilde{y}(\cdot, t_{2M}) \rightarrow \tilde{y}(\cdot, \tilde{t}) \quad \text{strongly in } L^2(\Omega) \quad \text{as } M \rightarrow \infty.$$

This, together with (2.22), implies that  $P_{r_N}(\tilde{y}(\cdot, \tilde{t})) = 0$  for all  $N \geq 1$ . Since  $r_N \rightarrow \infty$  when  $N \rightarrow \infty$ , it holds that  $\tilde{y}(\cdot, \tilde{t}) = 0$ . Thus, we have proved that for each  $\tilde{y}_0 \in L^2(\Omega)$ , there exists a control  $\tilde{u} \in L^\infty(t_1, \tilde{t}; L^2(\Omega))$  with the estimate  $\|\tilde{u}\|_{L^\infty(t_1, \tilde{t}; L^2(\Omega))}^2 \leq L \|\tilde{y}_0\|_{L^2(\Omega)}^2$ , where the constant  $L$  is given by (2.19), such that the solution  $\tilde{y}$  to (2.8) reaches zero value at time  $\tilde{t}$ , namely,  $\tilde{y}(x, \tilde{t}) = 0$  over  $\Omega$ .

Now, we take  $\tilde{y}_0(x)$  to be  $\psi(x, t_1)$ , where  $\psi(x, t)$  is the solution to the following equation:

$$\begin{cases} \psi_t(x, t) - \Delta \psi(x, t) = 0 & \text{in } \Omega \times (0, t_1), \\ \psi(x, t) = 0 & \text{on } \partial\Omega \times (0, t_1), \\ \psi(x, 0) = y_0(x) & \text{in } \Omega, \end{cases}$$

and we construct a control  $u$  by setting

$$(2.23) \quad u(x, t) = \begin{cases} 0 & \text{in } \Omega \times (0, t_1), \\ \tilde{u}(x, t) & \text{in } \Omega \times (t_1, \tilde{t}), \\ 0 & \text{in } \Omega \times (\tilde{t}, T). \end{cases}$$

It is clear that this control  $u$  is in the space  $L^\infty(0, T; L^2(\Omega))$  and that the corresponding solution  $y$  of (1.2) with  $\delta = 0$  reaches zero value at time  $T$ , namely,  $y(x, T) = 0$  over  $\Omega$ . Moreover, the control  $u$  constructed in (2.23) satisfies the following estimate:

$$\|u\|_{L^\infty(0, T; L^2(\Omega))}^2 \leq L \|y_0\|_{L^2(\Omega)}^2,$$

where  $L$  is given by (2.19).

Next, we take  $\delta_0$  to be the number  $t_1$  given above. For each  $\delta$  with  $0 \leq \delta \leq \delta_0$ , we set

$$\tilde{t}_\delta = \tilde{t} - \delta \quad \text{and} \quad t_{N,\delta} = t_N - \delta \quad \text{for all } N = 1, 2, \dots$$

Then it holds that

$$0 \leq t_{1,\delta} < t_{2,\delta} < \dots < t_{N,\delta} \rightarrow \tilde{t}_\delta < T - \delta.$$

Moreover, we have for each  $N \geq 1$ ,

$$m(E_\delta \cap [t_{N,\delta}, t_{N+1,\delta}]) = m(E \cap [t_N, t_{N+1}]) \geq \rho(t_{N+1} - t_N)$$

and

$$\frac{t_{N+1,\delta} - t_{N,\delta}}{t_{N+2,\delta} - t_{N+1,\delta}} = \frac{t_{N+1} - t_N}{t_{N+2} - t_{N+1}} \leq C_0,$$

where  $C_0$  and  $\rho$  are the positive constants as above.

Now, we can use exactly the same argument as above to get for each  $\delta$  with  $0 \leq \delta \leq \delta_0$  the existence of a control  $u_\delta(t)$  in the space  $L^\infty(0, T - \delta; L^2(\Omega))$  such that the corresponding solution  $y^\delta$  to (1.2) reaches zero value at time  $T - \delta$ , namely,  $y^\delta(x, T - \delta) = 0$  over  $\Omega$ . Moreover, this control  $u_\delta$  satisfies the following estimate (see (2.9)–(2.12) and (2.19)):

$$\|u_\delta\|_{L^\infty(0, T - \delta; L^2(\Omega))}^2 \leq L_\delta \cdot \|y_0\|_{L^2(\Omega)}^2.$$

The constant  $L_\delta$  is given by

$$L_\delta = \max \{ (\tilde{C}_\delta)^{N(N-1)} \alpha_{1,\delta} \cdots \alpha_{N,\delta}, \quad 1 \leq N \leq N_0 \},$$

where

$$\tilde{C}_\delta = \frac{2C_1}{\rho^2(t_{2,\delta} - t_{1,\delta})^2} \cdot C_0^2$$

and

$$\alpha_{N,\delta} = \begin{cases} \exp(C_2 \sqrt{r_{1,\delta}}), & N = 1, \\ \exp(C_2 \sqrt{r_{N,\delta}}) \exp(-2r_{N-1,\delta}(t_{3,\delta} - t_{2,\delta})C_0^{-2(N-2)}), & N \geq 2, \end{cases}$$

with

$$r_{N,\delta} = \left[ \frac{2}{(t_{3,\delta} - t_{2,\delta})} \tilde{C}_\delta^{N-1} \right]^4, \quad N = 1, 2, \dots,$$

and where the natural number  $N_0$  is given by (2.16). Since

$$t_{N+1,\delta} - t_{N,\delta} = t_{N+1} - t_N \quad \text{for all } N = 1, 2, \dots,$$

we see easily that  $\tilde{C}_\delta = \tilde{C}$  and  $\alpha_{N,\delta} = \alpha_N$  for all  $N \geq 1$ . Then it holds that  $L_\delta = L$  for all  $\delta$  with  $0 \leq \delta \leq \delta_0$ . This completes the proof.  $\square$

**3. Applications to the time optimal control problem.** In this section, we shall prove Theorems 1.2 and 1.3. We also will give some consequences of these theorems. Throughout this section, we shall denote by  $y(t; u, y_0)$  the solution of (1.3) corresponding to the control  $u$  and the initial data  $y_0$ .

*Proof of Theorem 1.2.* Seeking a contradiction, we suppose that there exist a subset  $E$  of positive measure in the interval  $[0, T^*]$  and a positive number  $\varepsilon$  such that the following holds:

$$u^*(t) \in U \text{ and } d(u^*(t), \partial U) \geq \varepsilon \text{ for each } t \text{ in the set } E,$$

where  $d(u^*(t), \partial U)$  denotes the distance of the point  $u^*(t)$  to the set  $\partial U$  in  $L^2(\Omega)$ . Then we would get

$$(3.1) \quad B\left(u^*(t), \frac{\varepsilon}{2}\right) \subset U \text{ for each } t \text{ in the set } E.$$

We shall obtain from (3.1) that there exist a positive number  $\delta$  with  $\delta < T^*$  and a control  $v_\delta$  in the set  $\mathcal{U}_{ad}$  such that the following holds:

$$(3.2) \quad y(T^* - \delta; v_\delta, y_0) = y(T^*; u^*, y_0).$$

Thus,  $T^*$  could not be the optimal time for problem (P), which leads to a contradiction.

We write  $\{G(t)\}_{t \geq 0}$  for the semigroup generated by  $\Delta$  with the Dirichlet boundary condition and observe that

$$\begin{aligned} y(T^* - \delta; v_\delta, y_0) &= G(T^* - \delta)y_0 + \int_0^{T^* - \delta} G(T^* - \delta - \sigma)\chi_\omega v_\delta(\sigma)d\sigma, \\ y(T^*; u^*, y_0) &= G(T^*)y_0 + \int_0^{T^*} G(T^* - \sigma)\chi_\omega u^*(\sigma)d\sigma. \end{aligned}$$

Hence, (3.2) is equivalent to the following: There exist a positive number  $\delta$  with  $\delta < T^*$  and a control  $v_\delta$  in the set  $\mathcal{U}_{ad}$  such that the following holds:

$$(3.3) \quad \int_0^{T^* - \delta} G(T^* - \delta - \sigma)\chi_\omega v_\delta(\sigma)d\sigma = [G(T^*) - G(T^* - \delta)]y_0 + \int_0^{T^*} G(T^* - \sigma)\chi_\omega u^*(\sigma)d\sigma.$$

Notice that for any positive number  $\delta$  with  $\delta < T^*$ , we have

$$\begin{aligned} & \int_0^{T^*} G(T^* - \sigma)\chi_\omega u^*(\sigma)d\sigma \\ &= \int_0^\delta G(T^* - \sigma)\chi_\omega u^*(\sigma)d\sigma + \int_\delta^{T^*} G(T^* - \sigma)\chi_\omega u^*(\sigma)d\sigma \\ &= G(T^* - \delta) \int_0^\delta G(\delta - \sigma)\chi_\omega u^*(\sigma)d\sigma + \int_0^{T^* - \delta} G(T^* - \delta - \sigma)\chi_\omega u^*(\delta + \sigma)d\sigma \end{aligned}$$

and

$$[G(T^*) - G(T^* - \delta)]y_0 = G(T^* - \delta)[(G(\delta) - I)y_0].$$

Therefore, (3.3) is equivalent to the following: There exist a positive number  $\delta$  with  $\delta < T^*$  and a control  $v_\delta$  in the set  $\mathcal{U}_{ad}$  such that the following holds:

$$\begin{aligned}
 & \int_0^{T^*-\delta} G(T^*-\delta-\sigma)\chi_\omega v_\delta(\sigma)d\sigma \\
 &= G(T^*-\delta) \left[ \int_0^\delta G(\delta-\sigma)\chi_\omega u^*(\sigma)d\sigma + (G(\delta)-I)y_0 \right] \\
 &+ \int_0^{T^*-\delta} G(T^*-\delta-\sigma)\chi_\omega u^*(\sigma+\delta)d\sigma \\
 &\equiv G(T^*-\delta)h_\delta + \int_0^{T^*-\delta} G(T^*-\delta-\sigma)\chi_\omega u^*(\sigma+\delta)d\sigma,
 \end{aligned}
 \tag{3.4}$$

where

$$h_\delta = \int_0^\delta G(\delta-\sigma)\chi_\omega u^*(\sigma)d\sigma + (G(\delta)-I)y_0.
 \tag{3.5}$$

For each positive number  $\delta$ , we write  $E_\delta$  for the set  $\{t; t+\delta \in E\}$  and denote by  $\chi_{E_\delta}$  the characteristic function of the set  $E_\delta$ . We first claim the following: For each positive number  $\delta$  sufficiently small, there exists a control  $u_\delta$  in the space  $L^\infty(0, \infty; L^2(\Omega))$  such that

$$\|u_\delta(t)\|_{L^2(\Omega)} \leq \frac{\varepsilon}{2} \quad \text{for almost all } t \geq 0,
 \tag{3.6}$$

and such that

$$y(T^*-\delta; \chi_{E_\delta} u_\delta, 0) = G(T^*-\delta)h_\delta.
 \tag{3.7}$$

Recall that  $y(t; \chi_{E_\delta} u_\delta, 0)$  is the solution of the controlled heat equation (1.3) with  $u$  and  $y_0$  being replaced by  $\chi_{E_\delta} u_\delta$  and 0, respectively, and that  $\varphi(t) \equiv G(t)h_\delta$  is the solution of (1.3) with  $u$  and  $y_0$  being replaced by 0 and  $h_\delta$ , respectively. Then, what we claimed above is obviously equivalent to the following: For each positive number  $\delta$  sufficiently small, there exists a control  $u_\delta$  with the estimate

$$\|u_\delta(t)\|_{L^2(\Omega)} \leq \frac{\varepsilon}{2} \quad \text{for almost all } t \geq 0,$$

such that the following holds:

$$z^\delta(T^*-\delta) = 0,$$

where  $z^\delta(t)$  is the solution to the following controlled heat equation:

$$\begin{cases} z_t^\delta(t) - \Delta z^\delta(t) = \chi_\omega \chi_{E_\delta}(t) u_\delta(t) & \text{in } (0, T^*-\delta), \\ z^\delta(0) = -h_\delta. \end{cases}
 \tag{3.8}$$

However, by Theorem 1.1, there exist positive numbers  $\delta_0$  and  $L$  such that for each  $\delta$  with  $0 < \delta \leq \delta_0$ , there is a control  $u_\delta$  in the space  $L^\infty(0, T^*-\delta; L^2(\Omega))$  with the estimate

$$\|u_\delta\|_{L^\infty(0, T^*-\delta; L^2(\Omega))}^2 \leq L \|h_\delta\|_{L^2(\Omega)}^2,
 \tag{3.9}$$



such that the following holds:

$$(3.10) \quad z^\delta(T^* - \delta) = 0.$$

On the other hand, by (3.5), we can get a positive number  $\tilde{\delta}$  such that for each positive number  $\delta$  with  $\delta \leq \tilde{\delta}$ , the following holds:

$$\|h_\delta\|_{L^2(\Omega)}^2 \leq \left(\frac{\varepsilon}{2}\right)^2 / L.$$

This, together with (3.9), implies that for each positive number  $\delta$  with  $\delta \leq \min\{\delta_0, \tilde{\delta}\}$ , there is a control  $u_\delta$  with the estimate

$$(3.11) \quad \|u_\delta\|_{L^\infty(0, T^* - \delta; L^2(\Omega))} \leq \frac{\varepsilon}{2},$$

such that the corresponding solution  $z^\delta$  to (3.8) satisfies (3.10).

Next, we fix a positive number  $\delta$  and the corresponding control  $u_\delta$  such that (3.10) and (3.11) hold. Then we extend the control  $u_\delta(\cdot)$  by setting it to be zero on the interval  $(T^* - \delta, \infty)$  and still denote the extension by  $u_\delta(\cdot)$ . Clearly, this extended control  $u_\delta$  is in the space  $L^\infty(0, \infty; L^2(\Omega))$  and makes (3.6) and (3.7) hold. Thus, we have proved the above mentioned claim.

Now, we take an element  $u_0$  from the control set  $U$  and construct a control  $v_\delta$  by setting

$$(3.12) \quad v_\delta(t) = \begin{cases} u^*(t + \delta) + \chi_{E_\delta}(t)u_\delta(t) & \text{if } t \in [0, T^* - \delta], \\ u_0 & \text{if } t > T^* - \delta. \end{cases}$$

It is clear that  $v_\delta(\cdot) : [0, \infty) \rightarrow L^2(\Omega)$  is measurable. We shall prove  $v_\delta(t) \in U$  for almost all  $t \geq 0$ . Here is the argument: When  $t$  is in the set  $[0, T^* - \delta] \cap E_\delta$ , we have  $t + \delta \in E$ . Then by (3.1), we get  $B(u^*(t + \delta), \frac{\varepsilon}{2}) \in U$ . Since  $\|u_\delta(t)\|_{L^2(\Omega)} \leq \frac{\varepsilon}{2}$  for almost all  $t \geq 0$ , we have

$$\|v_\delta(t) - u^*(t + \delta)\|_{L^2(\Omega)} = \|u_\delta(t)\|_{L^2(\Omega)} \leq \frac{\varepsilon}{2} \quad \text{for almost all } t \text{ in } [0, T^* - \delta] \cap E_\delta,$$

namely,  $v_\delta(t) \in B(u^*(t + \delta), \frac{\varepsilon}{2})$  for almost all  $t$  in the set  $[0, T^* - \delta] \cap E_\delta$ . Hence,  $v_\delta(t) \in U$  for almost all  $t$  in the set  $[0, T^* - \delta] \cap E_\delta$ . On the other hand, for almost all  $t \in [0, T^* - \delta] \cap (E_\delta)^c$ , we have  $v_\delta(t) = u^*(t + \delta) \in U$ . Therefore, we have proved  $v_\delta \in \mathcal{U}_{ad}$ .

Then, by (3.7) and (3.12), we see easily that this control  $v_\delta$  makes the equality (3.4) hold, which leads to a contradiction to the optimality of  $T^*$  for problem **(P)**. Thus we have proved  $u^*(t) \in \partial U$  for almost all  $t \in [0, T^*]$ .

Finally, if the control set  $U$  has the additional property that  $\chi_\omega U \subset U$ , then we have  $\chi_\omega u^* \in \mathcal{U}_{ad}$ . It is clear that  $y(T^*; \chi_\omega u^*, y_0) = y(T^*; u^*, y_0)$ . Thus,  $\chi_\omega u^*$  is also an optimal control for problem **(P)**. Hence, it holds that  $\chi_\omega u^*(t) \in \partial U$  for almost all  $t \in [0, T^*]$ . This completes the proof.  $\square$

By Theorem 1.2, we immediately get the following consequence.

**COROLLARY 3.1.** *Suppose that the control set  $U$  is the ball  $B(0, R)$  with  $R > 0$  and that the target set  $S$  is nonempty in  $L^2(\Omega)$ . Let  $T^*$  be the optimal time and  $u^*$  be an optimal control for problem **(P)**. Then it holds that  $\|\chi_\omega u^*(\cdot, t)\|_{L^2(\Omega)} = R$  for almost all  $t \in [0, T^*]$ .*

*Remark 3.2.* From the proof of Theorem 1.2, we see that if an admissible control  $u(\cdot, t)$  does not take its value on the boundary of the control set  $U$  in a subset of positive measure in the interval  $[0, T]$ , where the number  $T$  is such that  $y(T; u, y_0) \in S$ , then there exists “room” for us to construct another admissible control  $v$  such that the corresponding trajectory  $y(t; v, y_0)$  reaches  $y(T; u, y_0)$  before the time  $T$ . Hence, such an admissible control  $u$  cannot be optimal. This idea has been used in [6], [14], [16], and [19]. The key point is how to use this “room” to construct such an admissible control  $v$ . In this work, the null controllability result established in Theorem 1.1 shows us how. It was already observed in [16] that the null controllability of the boundary controlled one-dimensional heat equation in  $(0, 1) \times (0, T)$ , with controls restricted on an arbitrary subset  $E \subset [0, T]$  of positive measure, leads to a bang-bang principle of time optimal boundary controls for the one-dimensional heat equation.

Next, we shall use Theorem 1.1 to derive the uniqueness of the optimal control for problem **(P)** with certain target sets and control sets.

*Proof of Theorem 1.3.* Let  $U$  be the closed ball  $B(v_0, R)$  in  $L^2(\Omega)$ , centered at  $v_0$  and of positive radius  $R$ . Let  $T^*$  be the optimal time for problem **(P)**. Seeking a contradiction, we suppose that there exist two different optimal controls  $u^*$  and  $v^*$  for problem **(P)**. Then there would exist a subset  $E_1$  of positive measure in the interval  $[0, T^*]$  such that  $u^*(t) \neq v^*(t)$  for every  $t \in E_1$ . We first observe that

$$y(T^*; u^*, y_0), y(T^*; v^*, y_0) \in S.$$

Then we construct a control  $w^*(t)$  by setting

$$w^*(t) = \frac{u^*(t) + v^*(t)}{2} \text{ for almost all } t \in [0, \infty).$$

It is clear that  $w^* \in \mathcal{U}_{ad}$ . Moreover, since  $S$  is convex, we have

$$y(T^*; w^*, y_0) = \frac{y(T^*; u^*, y_0) + y(T^*; v^*, y_0)}{2} \in S.$$

On the other hand, we see that for almost all  $t \in E_1$ ,

$$\begin{aligned} \|w^*(t) - v_0\|_{L^2(\Omega)}^2 &= 2 \left( \left\| \frac{u^*(t) - v_0}{2} \right\|_{L^2(\Omega)}^2 + \left\| \frac{v^*(t) - v_0}{2} \right\|_{L^2(\Omega)}^2 \right) \\ &\quad - \left\| \frac{u^*(t) - v_0}{2} - \frac{v^*(t) - v_0}{2} \right\|_{L^2(\Omega)}^2 \\ &= R^2 - \frac{1}{4} \left\| u^*(t) - v^*(t) \right\|_{L^2(\Omega)}^2 \\ &< R^2. \end{aligned}$$

Thus, there exist a positive number  $\varepsilon$  and a subset  $E$  of positive measure in the set  $E_1$  such that for each  $t \in E$ ,  $d(w^*(t), \partial B(v_0, R)) \geq \varepsilon$ . Then, we can use the same argument as in the proof of Theorem 1.2 to derive a contradiction to the optimality of  $T^*$ . This completes the proof.  $\square$

With regard to the existence of the time optimal controls for problem **(P)**, we recall that if the target set  $S$  is closed and convex in  $L^2(\Omega)$ , which contains the origin in  $L^2(\Omega)$ , and if the control set  $U$  is the ball  $B(0, R)$  with  $R > 0$ , then problem **(P)** with any initial data  $y_0 \in L^2(\Omega)$  has an optimal control (see [17] and [20]). Thus, Theorem

1.4 follows immediately from Corollary 3.1, Theorem 1.3, and the aforementioned existence result.

**Appendix.** In what follows, we shall give the proof of the equivalence of Theorem 1.1 and the observability inequality **(O)**.

We shall first prove that Theorem 1.1 implies inequality **(O)**. To this end, we multiply (1.2) by  $p^\delta$ , where  $p^\delta$  is the solution of (1.6) with  $p_T^\delta(x) \in L^2(\Omega)$  being arbitrarily given, and then integrate it over  $\Omega \times (0, T - \delta)$ . Thus we obtain

$$(A.1) \quad \int_0^{T-\delta} \int_\Omega (\partial_t y^\delta - \Delta y^\delta) p^\delta dx dt = \int_0^{T-\delta} \int_\Omega [\chi_{E_\delta}(t) \chi_\omega(x) p^\delta(x, t) u_\delta(x, t)] dx dt.$$

By using integration by parts, we see that the following equality holds for each  $y_0(\cdot) \in L^2(\Omega)$ :

$$(A.2) \quad \int_0^{T-\delta} \int_\Omega (\partial_t y^\delta - \Delta y^\delta) p^\delta dx dt = \left[ \int_\Omega y^\delta p^\delta dx \right]_0^{T-\delta} = - \int_\Omega y_0 p^\delta(x, 0) dx.$$

Now, we choose  $y_0(x) = p^\delta(x, 0)$  in the above, and then it follows from (1.3), (A.1), and (A.2) that for each  $\varepsilon > 0$ , we have

$$\begin{aligned} \int_\Omega |p^\delta(x, 0)|^2 dx &= - \int_0^{T-\delta} \int_\Omega [\chi_{E_\delta}(t) \chi_\omega(x) p^\delta(x, t) u_\delta(x, t)] dx dt \\ &\leq \int_0^{T-\delta} \|\chi_{E_\delta} \chi_\omega u_\delta\|_{L^2(\Omega)} \|\chi_{E_\delta} \chi_\omega p^\delta\|_{L^2(\Omega)} dt \\ &\leq \|u_\delta\|_{L^\infty(0, T-\delta; L^2(\Omega))} \|\chi_{E_\delta} \chi_\omega p^\delta\|_{L^1(0, T-\delta; L^2(\Omega))} \\ &\leq \frac{1}{2\varepsilon} \|u_\delta\|_{L^\infty(0, T-\delta; L^2(\Omega))}^2 + \frac{\varepsilon}{2} \|\chi_{E_\delta} \chi_\omega p^\delta\|_{L^1(0, T-\delta; L^2(\Omega))}^2 \\ &\leq \frac{1}{2\varepsilon} L \|y_0\|_{L^2(\Omega)}^2 + \frac{\varepsilon}{2} \|\chi_{E_\delta} \chi_\omega p^\delta\|_{L^1(0, T-\delta; L^2(\Omega))}^2. \end{aligned}$$

Since  $y_0 = p^\delta(\cdot, 0)$ , we can get the desired inequality (1.5) by taking  $\varepsilon = L$  in the above inequality.

Next, we prove that inequality **(O)** implies Theorem 1.1.

By (1.2) and (1.6), we have for any  $p_T^\delta \in L^2(\Omega)$  and  $y_0 \in L^2(\Omega)$ ,

$$\begin{aligned} &\langle y^\delta(\cdot, T - \delta), p^\delta(\cdot, T - \delta) \rangle - \langle y_0(\cdot), p^\delta(\cdot), p^\delta(\cdot, 0) \rangle \\ &= \int_0^{T-\delta} \int_\Omega [\chi_{E_\delta}(t) \chi_\omega(x) p^\delta(x, t) u_\delta(x, t)] dx dt. \end{aligned}$$

Here and in what follows, the notation  $\langle \cdot, \cdot \rangle$  stands for the inner product of  $L^2(\Omega)$ . If we can show that there exists a control  $u_\delta \in L^\infty(0, T - \delta; L^2(\Omega))$  with the estimate (1.3) such that

$$\langle y^\delta(\cdot, T - \delta), p^\delta(\cdot, T - \delta) \rangle = 0 \quad \text{for all } p_T^\delta \in L^2(\Omega),$$

then it holds that  $y^\delta(\cdot, T - \delta) = 0$ . Thus, it suffices to prove that there exists a control  $u_\delta \in L^\infty(0, T - \delta; L^2(\Omega))$  with the estimate (1.3) such that

$$-\langle y_0(\cdot), p^\delta(\cdot, 0) \rangle = \int_0^{T-\delta} \int_\Omega [\chi_{E_\delta}(t) \chi_\omega(x) p^\delta(x, t) u_\delta(x, t)] dx dt$$

for any  $p_T^\delta \in L^2(\Omega)$ .

Now, we set

$$X = \{\chi_{E_\delta}(t)\chi_\omega(x)p^\delta(x, t); p^\delta \text{ solves (1.6) with } p_T^\delta \in L^2(\Omega)\}.$$

It is clear that  $X$  is a linear subspace of  $L^1(0, T - \delta; L^2(\Omega))$ . We define a linear functional  $F : X \rightarrow R^1$  by setting

$$F(\chi_{E_\delta}(t)\chi_\omega(x)p^\delta) = -\langle y_0(\cdot), p^\delta(\cdot, 0) \rangle.$$

Then it follows from (1.5) that

$$\begin{aligned} & |F(\chi_{E_\delta}(t)\chi_\omega(x)p^\delta)|^2 \\ & \leq \|y_0\|_{L^2(\Omega)}^2 \|p^\delta(\cdot, 0)\|_{L^2(\Omega)}^2 \\ & \leq L \|y_0\|_{L^2(\Omega)}^2 \|\chi_{E_\delta}\chi_\omega p^\delta\|_{L^1(0, T-\delta; L^2(\Omega))}^2. \end{aligned}$$

Namely,

$$\|F\|_{L(X; R^1)} \leq L^2 \|y_0\|_{L^2(\Omega)}^2.$$

Thus,  $F$  is a bounded linear functional on  $X$ . By the Hahn–Banach theorem, there exists a bounded linear functional  $G : L^1(0, T - \delta; L^2(\Omega)) \rightarrow R^1$  such that  $G = F$  on  $X$  and such that

$$\|G\|_{L^1(0, T-\delta; L^2(\Omega)); R^1}^2 = \|F\|_{L(X, R^1)}^2 \leq L \|y_0\|_{L^2(\Omega)}^2.$$

Then by making use of the Riesz representation theorem, there exists a function  $u_\delta \in L^\infty(0, T - \delta; L^2(\Omega))$  such that for any  $f \in L^1(0, T - \delta; L^2(\Omega))$ ,

$$G(f) = \int_0^{T-\delta} \int_\Omega f u_\delta dx dt,$$

and such that

$$\|u_\delta\|_{L^\infty(0, T-\delta; L^2(\Omega))}^2 = \|G\|_{L^1(0, T-\delta; L^2(\Omega)); R^1}^2 \leq L \|y_0\|_{L^2(\Omega)}^2.$$

Hence,

$$F(\chi_{E_\delta}(t)\chi_\omega(x)p^\delta) = \int_0^{T-\delta} \int_\Omega [\chi_{E_\delta}(t)\chi_\omega(x)p^\delta(x, t)u_\delta(x, t)] dx dt.$$

Namely, for all  $p_T^\delta \in L^2(\Omega)$ , it holds that

$$-\langle y_0(\cdot), p^\delta(\cdot, 0) \rangle = \int_0^{T-\delta} \int_\Omega [\chi_{E_\delta}(t)\chi_\omega(x)p^\delta(x, t)u_\delta(x, t)] dx dt. \quad \square$$

This completes the proof.

**Acknowledgments.** The author would like to express his appreciation to professor Xu Zhang and Dr. K. D. Phung for their valuable suggestions on this work.

## REFERENCES

- [1] N. ARADA AND J.-P. RAYMOND, *Time optimal problems with Dirichlet boundary conditions*, Discrete Contin. Dyn. Syst., 9 (2003), pp. 1549–1570.
- [2] V. BARBU, *Controllability of parabolic and Navier–Stokes equations*, Sci. Math. Japonia, 6 (2002), pp. 143–211.
- [3] J. DIESTEL AND J. J. UHL, JR., *Vector Measures*, Math. Surveys 15, AMS, Providence, RI, 1977.
- [4] A. DOUBOVA, E. FERNÁNDEZ-CARA, M. GONZÁLEZ-BURGOS, AND E. ZUAZUA, *On the controllability of parabolic systems with a nonlinear term involving the state and the gradient*, SIAM J. Control Optim., 41 (2002), pp. 798–819.
- [5] YU. V. EGOROV, *Optimal control in Banach spaces*, Dokl. Nauk SSSR, 150 (1963), pp. 241–244 (in Russian).
- [6] H. O. FATTORINI, *Time optimal control of solutions of operational differential equations*, SIAM J. Control Ser. A, 2 (1964), pp. 54–59.
- [7] H. O. FATTORINI, *The time optimal controls in Banach spaces*, Appl. Math. Optim., 1 (1974), pp. 163–188.
- [8] E. FERNÁNDEZ-CARA AND E. ZUAZUA, *Null and approximate controllability for weakly blowing up semilinear heat equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 583–616.
- [9] A. V. FURSIKOV AND O. YU. IMANUVILOV, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Seoul, Korea, 1996.
- [10] G. LEBEAU AND L. ROBBIANO, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.
- [11] G. LEBEAU AND E. ZUAZUA, *Null-controllability of a system of linear thermoelasticity*, Arch. Rational Mech. Anal., 141 (1998), pp. 297–329.
- [12] X. LI AND Y. YAO, *Time optimal control for distributed parameter systems*, Sci. Sinica, 24 (1981), pp. 455–465.
- [13] X. LI AND J. YONG, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser Boston, Boston, MA, 1995.
- [14] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1971.
- [15] A. LOPEZ, X. ZHANG, AND E. ZUAZUA, *Null controllability of the heat equation as singular limit of the exact controllability of dissipative wave equations*, J. Math. Pures Appl., 79 (2000), pp. 741–808.
- [16] V. J. MIZEL AND T. I. SEIDMAN, *An abstract bang-bang principle and time-optimal boundary control of the heat equation*, SIAM J. Control Optim., 35 (1997), pp. 1204–1216.
- [17] K. D. PHUNG, G. WANG, AND X. ZHANG, *On the existence of time optimal controls for linear evolution equations*, Discrete Contin. Dyn. Syst. Ser. B, 8 (2007), pp. 925–941.
- [18] D. L. RUSSELL, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
- [19] E. J. P. G. SCHMIDT, *The “bang-bang” principle for the time-optimal problem in boundary control of the heat equation*, SIAM J. Control Optim., 18 (1980), pp. 101–107.
- [20] G. WANG, *The existence of time optimal control of semilinear parabolic equations*, Systems Control Lett., 53 (2004), pp. 171–175.
- [21] G. WANG AND L. WANG, *The bang-bang principle of time optimal controls for the heat equation with internal controls*, Systems Control Lett., 56 (2007), pp. 709–713.
- [22] L. WANG AND G. WANG, *The optimal time control of a phase-field system*, SIAM J. Control Optim., 42 (2003), pp. 1483–1508.
- [23] J. YONG, *Time optimal control for semilinear distributed parameter systems—existence theory and necessary conditions*, Kodai Math. J., 14 (1991), pp. 239–253.

## GOAL-ORIENTED ADAPTIVITY IN CONTROL CONSTRAINED OPTIMAL CONTROL OF PARTIAL DIFFERENTIAL EQUATIONS\*

M. HINTERMÜLLER<sup>†</sup> AND R. H. W. HOPPE<sup>‡</sup>

**Abstract.** Dual-weighted goal-oriented error estimates for a class of pointwise control constrained optimal control problems for second order elliptic partial differential equations are derived. It is demonstrated that the constraints give rise to a primal-dual weighted error term representing the mismatch in the complementarity system due to discretization. The paper also contains a posteriori error estimators for the  $L^2$ -norm of the error in the state and in the adjoint state.

**Key words.** adaptive finite element method, a posteriori error estimate, control constraints, goal-oriented adaptivity, PDE-constrained optimization

**AMS subject classifications.** 49K20, 65K10, 65N30, 65N50

**DOI.** 10.1137/070683891

**1. Introduction.** In many computations involving the discretization of (partial) differential equations or variational inequalities, one is interested in the accurate evaluation of some target quantity. This might be the value of the solution of a partial differential equation (PDE) at some reference point in the domain of interest, a physically relevant quantity such as the drag in airfoil design, or, in optimal control, the value of the objective function at the solution of the underlying minimization problem. Highly accurate numerical evaluations of these targets can be guaranteed by using uniform meshes with a small mesh size  $h$ . This, however, usually represents a significant computational challenge due to the resulting large scale of the discrete problems. Therefore, one seeks to adaptively refine the meshes with the goal of achieving a desired accuracy in the evaluation of the output quantity of interest while keeping the computational cost as small as possible.

For this purpose, recently for (systems of) PDEs an approach based on dual-weighted residual-based error estimates was proposed. Here we point to the pioneering work summarized in [1, 3] and the references therein; see also [7] for related literature. It essentially relies on employing the dual problem of the underlying system with the target on the right-hand side. In fact, let  $A$  denote some possibly nonlinear partial differential operator and let  $f$  be some fixed data. Then, in some abstract form, the primal problem (or PDE) is given by

$$(1.1) \quad A(y) = f.$$

Let  $y_h$  be the result of a Galerkin finite element discretization of the underlying problem. If  $G(\cdot)$  represents some desired target quantity (or goal), then the dual

---

\*Received by the editors February 27, 2007; accepted for publication (in revised form) February 13, 2008; published electronically June 13, 2008.

<http://www.siam.org/journals/sicon/47-4/68389.html>

<sup>†</sup>Department of Mathematics, University of Sussex, Mantell Building, Falmer, Brighton, BN1 9RE, UK (M.Hintermueller@sussex.ac.uk, michael.hintermueller@uni-graz.at). This author gratefully acknowledges financial support from the Austrian Science Fund FWF under START-program Y305 “Interfaces and Free Boundaries.”

<sup>‡</sup>Institute of Mathematics, University of Augsburg, D-86159 Augsburg, Germany, and Department of Mathematics, University of Houston, Houston, TX 77204-3008 (rohop@math.uh.edu). This author has been partially supported by the NSF under grants DMS-0411403 and DMS-0511611.

approach consists of considering

$$(1.2) \quad A'(y_h)^* p_h = G(\cdot)$$

from which an a posteriori error estimate of the type

$$|G(y) - G(y_h)| \leq \sum_{T \in \mathbb{T}_h} \mathbf{p}_T(y_h) \mathfrak{d}_T(p_h)$$

is derived. Above,  $A'(\cdot)^*$  is the dual operator of the Frechét-derivative  $A'(\cdot)$  of  $A(\cdot)$ . Further,  $\mathbb{T}_h = \{T\}$  denotes a computational mesh consisting of elements  $T$ , and  $\mathbf{p}_T$  and  $\mathfrak{d}_T$  stand for the primal residual and the dual weight on each cell  $T$ , respectively.

In [2] this concept was transferred to PDE-constrained optimal control problems of the type

$$(P_0) \quad \text{minimize } J(y, u) \quad \text{subject to } A(y) = f + B(u),$$

where  $(y, u)$  denotes the state-control pair and  $B$  models the control impact. The first order optimality system of  $(P_0)$  can be formally written as

$$(1.3a) \quad A(y) - B(u) = f,$$

$$(1.3b) \quad J_y(y, u) + A'(y)^* p = 0,$$

$$(1.3c) \quad J_u(y, u) - B'(u)^* p = 0.$$

Here,  $J_y$  and  $J_u$  are the partial derivatives of  $J$  with respect to  $y$  and  $u$ . The variable  $p$  is called the adjoint state. Often, (1.3c) results in an algebraic equation, while (1.3a)–(1.3b) form a primal-dual pair of PDEs similar to (1.1)–(1.2). Since (1.3a)–(1.3b) represent a system of PDEs, the dual-weighted approach can be readily carried over to this optimal control setting.

The situation, however, changes significantly if, in addition to the PDE constraint in  $(P_0)$ , one has to account for pointwise almost everywhere (a.e.) constraints on the control variable. In this case, the resulting problem becomes

$$(P_c) \quad \begin{cases} \text{minimize} & J(y, u) \\ \text{subject to} & A(y) = f + B(u), \\ & a \leq u \leq b \quad \text{a.e. on } \Omega_C \subset \Omega, \end{cases}$$

where  $\Omega \subset \mathbb{R}^n$  denotes some suitable domain with  $\Omega_C \neq \emptyset$  a measurable subset, and where  $a < b$  are given bounds. The corresponding first order necessary optimality system now involves a variational inequality as follows:

$$(1.4a) \quad A(y) - B(u) = f,$$

$$(1.4b) \quad J_y(y, u) + A'(y)^* p = 0,$$

$$(1.4c) \quad \langle J_u(y, u) - B'(u)^* p, v - u \rangle \geq 0 \quad \forall v \in U^{\text{ad}}, u \in U^{\text{ad}},$$

where the set

$$U^{\text{ad}} = \{v : a \leq v \leq b\}$$

represents the feasible controls, and  $\langle \cdot, \cdot \rangle$  denotes a suitable duality pairing. The variational inequality induces some nonsmoothness in the first order optimality system. This can be seen best when defining the Lagrange multiplier  $\lambda$  pertinent to the pointwise constraints via

$$(1.5) \quad J_u(y, u) - B'(u)^* p + \lambda = 0$$

and, assuming that  $\lambda$  permits a pointwise interpretation,

$$(1.6) \quad \lambda \geq 0 \quad \text{a.e. on } \{u = b\}, \quad \lambda \leq 0 \quad \text{a.e. on } \{u = a\}, \quad \lambda = 0 \quad \text{else.}$$

The conditions in (1.6) represent the so-called *complementarity system*. It can be written equivalently as

$$(1.7) \quad \lambda = \min\{0, \lambda + \sigma(u - a)\} + \max\{0, \lambda + \sigma(u - b)\},$$

where  $\sigma > 0$  is an arbitrarily fixed real and the max- and min-operations are understood in the pointwise sense. From (1.7) the nonsmoothness involved in the first order necessary optimality conditions becomes apparent. Of course, suitable a posteriori error concepts have to reflect this situation in order to accurately resolve the influence of the constraints on the solution of the optimal control problem.

We note that for pointwisely constrained problems such as variational inequalities of obstacle type, finite element methods based on various concepts in the a posteriori analysis have been considered in the literature. The goal-oriented dual-weighted approach was used in [4], whereas residual-type and hierarchical-type estimators were derived and analyzed in [5, 10, 13, 16]. Although the situation under consideration is different from obstacle-type problems as the pointwise constraints in our case are imposed on the control acting on the right-hand side of the PDE, a common feature in the a posteriori error analysis is the appropriate treatment of the complementarity conditions.

In this paper, our starting point will be a sufficiently general model problem class of the type  $(P_c)$ . Based on the Lagrange function

$$\mathcal{L}(y, u, p, \lambda) = J(y, u) + \langle A(y) - f - B(u), p \rangle + (u - b, \lambda)$$

of  $(P_c)$ , for convenience written here for a unilaterally constrained version of the minimization problem, and with the objective function as the goal, we derive an error representation of the type

$$\begin{aligned} J(y, u) - J(y_h, u_h) &= -\frac{1}{2} \langle \nabla_{xx} \mathcal{L}(x_h, \lambda_h)(x_h - x), x_h - x \rangle + (u_h - b, \lambda) \\ &\quad + \text{osc}_h + r(x_h, x) \end{aligned}$$

with  $x = (p, y, u)$  and its discretized version  $x_h = (p_h, y_h, u_h)$ , respectively, and  $(\cdot, \cdot)$  some inner product. Further,  $\text{osc}_h$  represents data oscillations and  $r$  is the remainder term resulting from a Taylor expansion of  $\mathcal{L}$ . In a second step we then estimate the term due to the inequality constraints and utilize the a posteriori error estimators derived in [8] in order to obtain a computable error representation.

The rest of the paper is organized as follows. In the next section we derive our new dual-weighted residual-based error estimator for a representative control constrained optimal control model problem. Section 3 is devoted to possible extensions. In fact, we study the bilaterally constrained case, a class of nonlinear governing equations, and alternative concepts for obtaining a posteriori estimates pertinent to the complementarity system. In the appendix, for our constrained optimal control problem we derive a new a posteriori error estimate with respect to the  $L^2$ -norm. Finally, in section 4 we report on numerical results due to our new error estimator.

*Notation.* Throughout we use  $\|\cdot\|_{0,\Omega}$  and  $(\cdot, \cdot)_{0,\Omega}$  for the usual  $L^2(\Omega)$ -norm and  $L^2(\Omega)$ -inner product, respectively. For convenience, with respect to the notation we



shall not distinguish between the norm (respectively, inner product) for scalar-valued or vector-valued arguments. We also use  $(\cdot, \cdot)_{0,\mathcal{S}}$ , which is the  $L^2(\mathcal{S})$ -inner product over a (measurable) subset  $\mathcal{S} \subset \Omega$ . By  $|\cdot|_{1,\Omega}$  we denote the  $H^1(\Omega)$ -seminorm  $|y|_{1,\Omega} = \|\nabla y\|_{0,\Omega}$ , which, by the Poincaré–Friedrichs inequality, is a norm on  $H_0^1(\Omega)$ . The norm in  $H^1(\Omega)$  is written as  $\|\cdot\|_{1,\Omega}$ . By  $\mathbb{T}_h = \mathbb{T}_h(\Omega)$  we denote a shape regular finite element triangulation of the domain  $\Omega$ . The subscript  $h = \max\{\text{diam}(T) \mid T \in \mathbb{T}_h\}$  indicates the mesh size of  $\mathbb{T}_h$ .

**2. Residual-based error estimate.** For deriving the structure of the new error estimate due to the inequality constraints, we consider the model problem

$$(P) \quad \begin{cases} \text{minimize} & J(y, u) := \frac{1}{2}\|y - z\|_{0,\Omega}^2 + \frac{\alpha}{2}\|u\|_{0,\Omega}^2 \\ \text{over} & (y, u) \in H_0^1(\Omega) \times L^2(\Omega) \\ \text{subject to} & -\Delta y = u + f, \\ & u \leq b \quad \text{a.e. in } \Omega, \end{cases}$$

which is a particular instance of  $(P_c)$ . The domain  $\Omega \in \mathbb{R}^2$  is assumed to be bounded and polygonal with boundary  $\Gamma := \partial\Omega$ . For the data we assume  $z, b, f \in L^2(\Omega)$  and  $\alpha > 0$ . It is well known that (P) admits a unique solution  $(y^*, u^*) \in H_0^1(\Omega) \times L^2(\Omega)$  (cf., e.g., [12]). Moreover, the optimal solution is characterized by the existence of an adjoint state  $p^* \in H_0^1(\Omega)$  and a Lagrange multiplier  $\lambda^* \in L^2(\Omega)$  which satisfy the first order necessary (and in this case, also sufficient) conditions

$$(2.1a) \quad -\Delta y^* = u^* + f,$$

$$(2.1b) \quad -\Delta p^* + y^* = z,$$

$$(2.1c) \quad \alpha u^* + \lambda^* - p^* = 0,$$

$$(2.1d) \quad u^* \leq b, \quad \lambda^* \geq 0, \quad (u^* - b, \lambda^*)_{0,\Omega} = 0.$$

We define the Lagrange functional  $\mathcal{L} : H_0^1(\Omega) \times L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega) \rightarrow \mathbb{R}$  pertinent to (P) as

$$(2.2) \quad \mathcal{L}(y, u, p, \lambda) = J(y, u) + (\nabla y, \nabla p)_{0,\Omega} - (u + f, p)_{0,\Omega} + (u - b, \lambda)_{0,\Omega}.$$

For convenience we use  $x := (p, y, u)$ ,  $x^* = (p^*, y^*, u^*)$ , and  $X = P \times Y \times L = H_0^1(\Omega) \times H_0^1(\Omega) \times L^2(\Omega)$ . Obviously, the weak form of (2.1a)–(2.1b) and (2.1c) of the optimality system (2.1) is equivalent to

$$(2.3) \quad \nabla_x \mathcal{L}(x^*, \lambda^*)(\varphi) = 0 \quad \forall \varphi \in X.$$

Let  $X_h \subset X$ , with  $X_h = P_h \times Y_h \times L_h$ , denote a finite dimensional subspace with the subscript  $h$  indicating the mesh size of the discretization obtained by a standard Galerkin method, let  $\lambda_h \in L_h \subset L^2(\Omega)$  denote the discrete (finite dimensional) counterpart of  $\lambda$  (analogously for  $\lambda^*$ ), and let  $f_h, b_h, z_h \in L_h$  be the  $L^2$ -projections of  $f, b, z$  onto  $L_h$ . The finite dimensional version of (2.1) reads

$$(2.4a) \quad \nabla_x \mathcal{L}_h(x_h^*, \lambda_h^*)(\varphi_h) = 0 \quad \forall \varphi_h \in X_h,$$

$$(2.4b) \quad u_h^* \leq b_h, \quad \lambda_h^* \geq 0, \quad (u_h^* - b_h, \lambda_h^*)_{0,\Omega} = 0,$$

where the discrete Lagrange function is given by

$$(2.5) \quad \begin{aligned} \mathcal{L}_h(x_h, \lambda_h) &= J_h(y_h, u_h) + (\nabla y_h, \nabla p_h)_{0,\Omega} - (u_h + f_h, p_h)_{0,\Omega} \\ &\quad + (u_h - b_h, \lambda_h)_{0,\Omega} \end{aligned}$$

with  $J_h(y_h, u_h) = \frac{1}{2}\|y_h - z_h\|_{0,\Omega}^2 + \frac{\alpha}{2}\|u_h\|_{0,\Omega}^2$ . Observe that the pointwise representation (2.1c) in the discrete setting reads

$$(2.6) \quad \alpha u_h^* + \lambda_h^* - M_h p_h^* = 0,$$

where  $M_h$  represents a projection operator from  $P_h$  onto  $L_h$ .

Further note that for  $x \in X$ ,  $\lambda \in L^2(\Omega)$  and  $x_h \in X_h$ ,  $\lambda_h \in L_h$ ,

$$(2.7) \quad \mathcal{L}(x, \lambda_h) = \mathcal{L}(x, \lambda) + (u - b, \lambda_h - \lambda)_{0,\Omega},$$

$$(2.8) \quad \nabla_x \mathcal{L}(x_h, \lambda_h)(\varphi_h) = \nabla_x \mathcal{L}(x_h, \lambda)(\varphi_h) + (\delta u_h, \lambda_h - \lambda)_{0,\Omega}$$

for all  $(\delta p_h, \delta y_h, \delta u_h) = \varphi_h \in X_h$ . Moreover, for our model problem (P) the second derivative of  $\mathcal{L}$  with respect to  $x$  does not depend on  $x$  and  $\lambda$ . Thus, we can write  $\nabla_{xx} \mathcal{L}(\varphi, \hat{\varphi})$  instead of  $\nabla_{xx} \mathcal{L}(x, \lambda)(\varphi, \hat{\varphi})$ . Similar observations hold true for  $\mathcal{L}_h$ . Due to  $X_h \subset X$ , we have for  $\varphi_h = (\delta p_h, \delta y_h, \delta u_h) \in X_h$ ,

$$(2.9) \quad \begin{aligned} 0 &= \nabla_x \mathcal{L}(x^*, \lambda^*)(\varphi_h) \\ &= \nabla_x \mathcal{L}(x_h^*, \lambda^*)(\varphi_h) + \nabla_{xx} \mathcal{L}(x^* - x_h^*, \varphi_h) \\ &= \nabla_x \mathcal{L}(x_h^*, \lambda_h^*)(\varphi_h) + (\delta u_h, \lambda^* - \lambda_h^*)_{0,\Omega} + \nabla_{xx} \mathcal{L}(x^* - x_h^*, \varphi_h) \\ &= \nabla_x \mathcal{L}_h(x_h^*, \lambda_h^*)(\varphi_h) - (f - f_h, \delta p_h)_{0,\Omega} - (z - z_h, \delta y_h)_{0,\Omega} \\ &\quad + (\delta u_h, \lambda^* - \lambda_h^*)_{0,\Omega} + \nabla_{xx} \mathcal{L}(x^* - x_h^*, \varphi_h) \\ &= (\delta u_h, \lambda^* - \lambda_h^*)_{0,\Omega} + \nabla_{xx} \mathcal{L}(x^* - x_h^*, \varphi_h) - (f - f_h, \delta p_h)_{0,\Omega} \\ &\quad - (z - z_h, \delta y_h)_{0,\Omega}. \end{aligned}$$

From this we further derive the relations

$$(2.10) \quad \begin{aligned} \nabla_{xx} \mathcal{L}(x_h^* - x^*, x_h^* - x^*) \\ = \nabla_{xx} \mathcal{L}(x_h^* - x^*, x_h^* - x^* + \varphi_h) - (\delta u_h, \lambda^* - \lambda_h^*)_{0,\Omega} \\ + (f - f_h, \delta p_h)_{0,\Omega} + (z - z_h, \delta y_h)_{0,\Omega}, \end{aligned}$$

$$(2.11) \quad \nabla_x \mathcal{L}(x_h^*, \lambda^*)(x^* - x_h^* - \varphi_h) = \nabla_{xx} \mathcal{L}(x_h^* - x^*, x^* - x_h^* - \varphi_h)$$

and also

$$(2.12) \quad \begin{aligned} \nabla_x \mathcal{L}(x_h^*, \lambda_h^*)(x^* - x_h^* - \varphi_h) \\ = \nabla_x \mathcal{L}(x^*, \lambda_h^*)(x^* - x_h^* - \varphi_h) + \nabla_{xx} \mathcal{L}(x_h^* - x^*, x^* - x_h^* - \varphi_h) \\ = (\lambda_h^* - \lambda^*, u^* - u_h^* - \delta u_h)_{0,\Omega} + \nabla_{xx} \mathcal{L}(x_h^* - x^*, x^* - x_h^* - \varphi_h). \end{aligned}$$

These preliminary results are now used to prove the following theorem.

**THEOREM 2.1.** *Let  $(x^*, \lambda^*) \in X \times L^2(\Omega)$  and  $(x_h^*, \lambda_h^*) \in X_h \times L_h$  denote the solution of (2.1) and its finite dimensional counterpart (2.4). Then*

$$(2.13) \quad \begin{aligned} J(y^*, u^*) - J_h(y_h^*, u_h^*) &= -\frac{1}{2} \nabla_{xx} \mathcal{L}(x_h^* - x^*, x_h^* - x^*) \\ &\quad + (u_h^* - b, \lambda^*)_{0,\Omega} + \text{osc}_h(x_h^*), \end{aligned}$$

where the oscillations  $\text{osc}_h(x_h^*)$  are given by

$$\text{osc}_h(x_h^*) = (y_h^* - z_h, z_h - z)_{0,\Omega} + \frac{1}{2} \|z - z_h\|_{0,\Omega}^2 + (f_h - f, p_h^*)_{0,\Omega}.$$

*Proof.* Observe that  $J(y^*, u^*) = \mathcal{L}(x^*, \lambda^*)$  and  $J_h(y_h^*, u_h^*) = \mathcal{L}_h(x_h^*, \lambda_h^*)$ . Using Taylor expansions and (2.7)–(2.8), we obtain

$$\begin{aligned}
 J(y^*, u^*) - J_h(y_h^*, u_h^*) &= \mathcal{L}(x^*, \lambda^*) - \mathcal{L}_h(x_h^*, \lambda_h^*) \\
 &= \mathcal{L}(x^*, \lambda^*) - \mathcal{L}_h(x^*, \lambda_h^*) - \nabla_x \mathcal{L}_h(x^*, \lambda_h^*)(x_h^* - x^*) \\
 &\quad - \frac{1}{2} \nabla_{xx} \mathcal{L}_h(x_h^* - x^*, x_h^* - x^*) \\
 &= J(y^*, u^*) - J_h(y^*, u^*) + (f_h - f, p^*)_{0,\Omega} - (u^* - b_h, \lambda_h^*)_{0,\Omega} \\
 &\quad - \nabla_x \mathcal{L}_h(x^*, \lambda_h^*)(x_h^* - x^*) - \frac{1}{2} \nabla_{xx} \mathcal{L}_h(x_h^* - x^*, x_h^* - x^*) \\
 &= \text{osc}_h(x_h^*) - (u^* - b_h, \lambda_h^*)_{0,\Omega} - \nabla_x \mathcal{L}(x^*, \lambda_h^*)(x_h^* - x^*) \\
 &\quad - \frac{1}{2} \nabla_{xx} \mathcal{L}_h(x_h^* - x^*, x_h^* - x^*) \\
 &= \text{osc}_h(x_h^*) - (u^* - u_h^*, \lambda_h^*)_{0,\Omega} + (\lambda^* - \lambda_h^*, u_h^* - u^*)_{0,\Omega} \\
 &\quad - \frac{1}{2} \nabla_{xx} \mathcal{L}_h(x_h^* - x^*, x_h^* - x^*) \\
 &= \text{osc}_h(x_h^*) + (\lambda^*, u_h^* - b)_{0,\Omega} - \frac{1}{2} \nabla_{xx} \mathcal{L}_h(x_h^* - x^*, x_h^* - x^*),
 \end{aligned}$$

where we also used the complementarity relations (2.1d) and (2.4b) as well as (2.3) and (2.4a).  $\square$

Assume, for the moment, that  $\lambda^* = 0$  and  $\lambda_h^* = 0$ ; i.e., the continuous and the discrete control constraints are inactive. Then we infer from (2.10) that for all  $\varphi_h \in X_h$  there holds

$$\begin{aligned}
 \nabla_{xx} \mathcal{L}(x_h^* - x^*, x_h^* - x^*) &= \nabla_{xx} \mathcal{L}(x_h^* - x^*, x_h^* - x^* + \varphi_h) \\
 &\quad + (f - f_h, \delta p_h)_{0,\Omega} + (z - z_h, \delta y_h)_{0,\Omega}
 \end{aligned}$$

as well as

$$\begin{aligned}
 J(y^*, u^*) - J_h(y_h^*, u_h^*) &= \frac{1}{2} \nabla_x \mathcal{L}_h(x_h, \lambda_h)(x^* - x_h^* - \varphi_h) \\
 (2.14) \quad &\quad + \frac{1}{2} (f_h - f, p^* - p_h^*)_{0,\Omega} + \frac{1}{2} (z_h - z, y^* - y_h^*)_{0,\Omega} \\
 &\quad + \text{osc}_h(x_h^*)
 \end{aligned}$$

due to (2.12). This corresponds to the result in [2, Proposition 4.1] for the unconstrained version of (P).

If  $b_h \leq b$  a.e. in  $\Omega$ , then (2.13) implies

$$J(y^*, u^*) \leq J_h(y_h^*, u_h^*) + \text{osc}_h(x_h^*).$$

Next we interpret the new, second term in the right-hand side of (2.13). For this purpose we define the active set  $\mathcal{A}^*$  and the inactive set  $\mathcal{I}^*$  at the optimal solution  $(x^*, \lambda^*)$  of (P) by

$$(2.15) \quad \mathcal{A}^* := \{x \in \Omega : u^*(x) = b(x)\}, \quad \mathcal{I}^* := \Omega \setminus \mathcal{A}^*.$$

Analogously we define the discrete counterparts  $\mathcal{A}_h^*$  and  $\mathcal{I}_h^*$ . Obviously,  $u^* < b$  a.e. in  $\mathcal{I}^*$ . By (2.1d), this implies  $\lambda^* = 0$  a.e. in  $\mathcal{I}^*$ . Therefore, the term  $(u_h^* - b, \lambda^*)_{0,\Omega}$  satisfies

$$(u_h^* - b, \lambda^*)_{0,\Omega} = (u_h^* - b_h, \lambda^*)_{0,\mathcal{A}^* \cap \mathcal{I}_h^*} + (b_h - b, \lambda^*)_{0,\mathcal{A}^*}.$$

The right-hand side above reflects the *error in complementarity*. In fact, the second term represents the data oscillation in the bound in the active set weighted by the continuous Lagrange multiplier. For this term we introduce the notation

$$\text{osc}_h^{A^*}(b; \lambda^*) := (b_h - b, \lambda^*)_{0, A^*}.$$

The first term captures a *primal-dual weighted mismatch in complementarity in  $\mathcal{A}^* \cap \mathcal{I}_h^*$* .

Let  $i_h := (i_h^p, i_h^y, i_h^u)$  be an interpolation operator such that  $i_h x \in X_h$  for  $x \in X$ . Moreover, for  $y, p \in H_0^1(\Omega)$  there exist  $i_h^p$  and  $i_h^y$  such that  $\max\{\|i_h^p p - p\|_{H^1}, \|i_h^y y - y\|_{H^1}\} \rightarrow 0$  for  $h \rightarrow 0$ . In connection with Theorem 2.1 we have the following result.

**THEOREM 2.2.** *Let the assumptions of Theorem 2.1 be satisfied. Then*

$$\begin{aligned} J(y^*, u^*) - J_h(y_h^*, u_h^*) &= -\frac{1}{2} \left( (y_h^* - z_h, i_h^y y^* - y^*)_{0, \Omega} + (\nabla(i_h^y y^* - y^*), \nabla p_h^*)_{0, \Omega} \right. \\ &\quad + (\nabla(i_h^p p^* - p^*), \nabla y_h^*)_{0, \Omega} - (u_h^* + f_h, i_h^p p^* - p^*)_{0, \Omega} \\ &\quad \left. + (M_h p_h^* - p_h^*, i_h^u u^* - u^*)_{0, \Omega} \right) \\ (2.16) \quad &+ \frac{1}{2} [(u_h^* - b, \lambda^*)_{0, \Omega} + (b_h - u^*, \lambda_h^*)_{0, \Omega}] + \frac{1}{2} (f - f_h, p_h^* - p^*)_{0, \Omega} \\ &+ \frac{1}{2} (z - z_h, y_h^* - y^*)_{0, \Omega} + \text{osc}_h(x_h^*). \end{aligned}$$

*Proof.* Utilizing (2.10)–(2.11) and considering  $\varphi_h = (\delta p_h, \delta y_h, \delta u_h) \in X_h$ , we obtain

$$\begin{aligned} J(y^*, u^*) - J_h(y_h^*, u_h^*) &= \frac{1}{2} \nabla_{xx} \mathcal{L}(x, \lambda_h^*)(x^* - x_h^*, x_h^* - x^* + \varphi_h) \\ &\quad + \frac{1}{2} (\delta u_h, \lambda^* - \lambda_h^*)_{0, \Omega} + \frac{1}{2} (f_h - f, \delta p_h)_{0, \Omega} + \frac{1}{2} (z_h - z, \delta y_h)_{0, \Omega} \\ &\quad + (u_h^* - b, \lambda^*)_{0, \Omega} + \text{osc}_h(x_h^*) \\ &= -\frac{1}{2} \nabla_x \mathcal{L}(x_h^*, \lambda_h^*)(x_h^* - x^* + \varphi_h) + \frac{1}{2} (\lambda_h^* + \lambda^*, u_h^* - u^*)_{0, \Omega} \\ &\quad + \frac{1}{2} (f_h - f, \delta p_h)_{0, \Omega} + \frac{1}{2} (z_h - z, \delta y_h)_{0, \Omega} + \text{osc}_h(x_h^*) \\ &= -\frac{1}{2} \nabla_x \mathcal{L}_h(x_h^*, \lambda_h^*)(x_h^* - x^* + \varphi_h) + \frac{1}{2} (\lambda_h^* + \lambda^*, u_h^* - u^*)_{0, \Omega} \\ &\quad + \frac{1}{2} (f - f_h, p_h^* - p^*)_{0, \Omega} + \frac{1}{2} (z - z_h, y_h^* - y^*)_{0, \Omega} + \text{osc}_h(x_h^*). \end{aligned}$$

Choosing  $\varphi_h = (i_h^p p^* - p_h^*, i_h^y y^* - y_h^*, i_h^u u^* - u_h^*) \in X_h$  and using complementary slackness, we continue with

$$\begin{aligned} J(y^*, u^*) - J_h(y_h^*, u_h^*) &= -\frac{1}{2} \nabla_x \mathcal{L}_h(x_h^*, \lambda_h^*)(i_h x^* - x^*) \\ &\quad + \frac{1}{2} [(\lambda_h^*, b_h - u^*)_{0, \Omega} + (\lambda^*, u_h^* - b)_{0, \Omega}] \\ &\quad + \frac{1}{2} (f - f_h, p_h^* - p^*)_{0, \Omega} + \frac{1}{2} (z - z_h, y_h^* - y^*)_{0, \Omega} \\ &\quad + \text{osc}_h(x_h^*). \end{aligned}$$

The assertion now follows from (2.2) and  $\alpha u_h^* - M_h p_h^* + \lambda_h^* = 0$  a.e. in  $\Omega$ .  $\square$

This result is interesting in several ways as follows:

- (i) For  $\|M_h p_h - p_h\|_{0,\Omega} \rightarrow 0$  as  $h \rightarrow 0$  sufficiently fast, only the convergence properties implied by  $i_h^p$  and  $i_h^y$  are required for obtaining an a posteriori error estimate based on (2.16). Since  $y^*$  and  $p^*$  solve elliptic PDEs, they usually enjoy more regularity than  $u^*$  and  $\lambda^*$ .
- (ii) The term in brackets on the right-hand side in (2.16) is again related to errors coming from complementary slackness. The first term of the sum can be interpreted as before, while the second term of the sum reflects the symmetric case, i.e.,

$$(b_h - u^*, \lambda_h^*)_{0,\Omega} = (b - u^*, \lambda_h^*)_{0,\mathcal{A}_h^* \cap \mathcal{I}^*} + (b_h - b, \lambda_h^*)_{0,\mathcal{A}_h^*}.$$

Hence, the first term of the right-hand side above represents the *primal-dual weighted mismatch in complementarity in  $\mathcal{I}^* \cap \mathcal{A}_h^*$* , while the second term denotes the data oscillation on  $\mathcal{A}_h^*$  weighted by the discrete multiplier, i.e.,

$$\text{osc}_h^{\mathcal{A}_h^*}(b; \lambda_h^*) := (b_h - b, \lambda_h^*)_{0,\mathcal{A}_h^*}.$$

Of course, (2.16) is not immediately amenable to numerical realization since  $u^*$  and  $\lambda^*$  are involved. Before we tackle this point, let us first state a posteriori error bounds for the control and the adjoint state which were derived in [8]. A coarser estimate was established in [14]. Recall that  $U^{ad}$  denotes the set of admissible controls, and let  $U_h^{ad}$  be its discretization. Then the following a posteriori error estimates hold true:

$$(2.17a) \quad \max(\|\lambda^* - \lambda_h^*\|_{0,\Omega}^2, \|u^* - u_h^*\|_{0,\Omega}^2) \leq C_1^2 \eta_1^2 + C_2^2 \eta_2^2 + C_b^2 \mu_h^2(b),$$

$$(2.17b) \quad \|p^* - p_h^*\|_{1,\Omega}^2 \leq C_2^2 \eta_2^2 + C_z^2 \text{osc}_h^2(z).$$

In what follows we also use

$$C_3^2 \eta_3^2 := C_1^2 \eta_1^2 + C_2^2 \eta_2^2 + C_b^2 \mu_h^2(b) \quad \text{and} \quad C_4^2 \eta_4^2 := C_2^2 \eta_2^2 + C_z^2 \text{osc}_h^2(z).$$

Here and below,  $C_i > 0$ ,  $i = 1, 2, 3, 4$ , denote constants which depend on  $\alpha$ ,  $\Omega$ , and the shape regularity of  $\mathbb{T}_h$ . The error bounds  $\eta_1$  and  $\eta_2$  are defined as

$$(2.18) \quad \eta_1^2 = \sum_T \int_T h_T^2 (p_h^* - M_h p_h^*)^2,$$

$$(2.19) \quad \eta_2^2 = \sum_T \int_T h_T^2 (f + u_h^* + \Delta y_h^*)^2 + \sum_F \int_F h_F [\nabla y_h^* \cdot n]^2 \\ + \sum_T \int_T h_T^2 (z - y_h^* + \Delta p_h^*)^2 + \sum_F \int_F h_F [\nabla p_h^* \cdot n]^2.$$

Further, the data oscillations

$$(2.20) \quad \mu_h^2(b) = \sum_{T \in \mathbb{T}_h} \|b - b_h\|_{0,T}^2,$$

$$(2.21) \quad \text{osc}_h^2(z) = \sum_{T \in \mathbb{T}_h} h_T^2 \|z - z_h\|_{0,T}^2$$

are involved.

Above,  $T$  denotes an element of the triangulation  $\mathbb{T}_h$  of  $\Omega$ . Further,  $F$  denotes a face of  $T$ , and  $h_F$  is the maximal diameter of the face  $F$ . Moreover,  $[\nabla y_h^* \cdot n]$  is the normal derivative jump over an interior face  $F$ . As noted before, the operator  $M_h$  represents the projection of a mesh function in  $P_h$  ( $= Y_h$ , typically in our context) onto  $L_h$ . If  $L_h$  is given by

$$L_h = \{u_h \in L^2(\Omega) : u_h|_T \in P_0(T), T \in \mathbb{T}_h\},$$

i.e., the function  $u_h$  is piecewise constant on  $\mathbb{T}_h$ , then the action of  $M_h$  in  $T$  is given by

$$(M_h p_h)|_T = \frac{1}{|T|} \int_T p_h(x) dx, \quad T \in \mathbb{T}_h.$$

A final observation concerns the unconstrained case, which is  $U^{ad} = L^2(\Omega)$ . In this situation we have  $\lambda^* = 0$  a.e. in  $\Omega$ . From (2.18)–(2.19) we see that the error estimator remains unaffected.

Our investigations concentrate now on the term

$$(2.22) \quad \frac{1}{2} [(u_h^* - b, \lambda^*)_{0,\Omega} + (b_h - u^*, \lambda_h^*)_{0,\Omega}] =: \Psi^*(\Omega),$$

which contains  $u^*$  and  $\lambda^*$ . A simple manipulation yields

$$\Psi^*(\Omega) = \frac{1}{2} [(\lambda_h^* - \lambda^*, b_h - u^*)_{0,\Omega} + (\lambda^* - \lambda_h^*, u_h^* - b)_{0,\Omega} + (\lambda^* + \lambda_h^*, b_h - b)_{0,\Omega}].$$

From first order optimality we recall

$$(2.23) \quad u^* \leq b, \quad \lambda^* \geq 0, \quad (u^* - b, \lambda^*)_{0,\Omega} = 0, \quad \alpha u^* - p^* + \lambda^* = 0,$$

$$(2.24) \quad u_h^* \leq b_h, \quad \lambda_h^* \geq 0, \quad (u_h^* - b_h, \lambda_h^*)_{0,\Omega} = 0, \quad \alpha u_h^* - M_h p_h^* + \lambda_h^* = 0.$$

Obviously, we have

$$(2.25a) \quad \Psi^*(\mathcal{I}^* \cap \mathcal{I}_h^*) = 0,$$

$$(2.25b) \quad \Psi^*(\mathcal{A}^* \cap \mathcal{A}_h^*) = \frac{1}{2} (\lambda^* - \lambda_h^*, b_h - b)_{0,\mathcal{A}^* \cap \mathcal{A}_h^*} + (\lambda_h^*, b_h - b)_{0,\mathcal{A}^* \cap \mathcal{A}_h^*},$$

where  $\Psi^*(\mathcal{S}) = \frac{1}{2} [(u_h^* - b, \lambda^*)_{0,\mathcal{S}} + (b_h - u^*, \lambda_h^*)_{0,\mathcal{S}}]$ . In the right-hand side of (2.25b), typically the latter term dominates. It is nonpositive if  $b_h \leq b$  a.e. in  $\mathcal{A}_h^*$ . Note that if  $b_h = b$  a.e. in  $\Omega$ , then  $\Psi^*(\mathcal{A}^* \cap \mathcal{A}_h^*) = 0$ . From the structure of  $\Psi^*(\mathcal{A}^* \cap \mathcal{A}_h^*)$  we can see that it represents a dual-weighted data oscillation on  $\mathcal{A}^* \cap \mathcal{A}_h^*$ . Subsequently we use

$$(2.26) \quad \text{osc}_h^{\mathcal{S}}(b; \lambda^* + \lambda_h^*) := (b_h - b, \lambda_h^* + \lambda^*)_{0,\mathcal{S}}.$$

Note that  $\text{osc}_h^{\mathcal{I}^* \cap \mathcal{I}_h^*}(b; \lambda^* + \lambda_h^*) = 0$ .

Utilizing (2.23)–(2.26), for  $\mathcal{C}_1^* = \mathcal{A}^* \cap \mathcal{I}_h^*$  and  $\mathcal{C}_2^* = \mathcal{I}^* \cap \mathcal{A}_h^*$  we obtain

$$(2.27a) \quad \Psi^*(\mathcal{C}_1^*) = \frac{\alpha}{2} \|u_h^* - u^*\|_{0,\mathcal{C}_1^*}^2 + \frac{1}{2} (p^* - M_h p_h^*, u_h^* - u^*)_{0,\mathcal{C}_1^*},$$

$$(2.27b) \quad \Psi^*(\mathcal{C}_2^*) = \frac{1}{2} (b_h - \alpha^{-1} p^*, \lambda_h^*)_{0,\mathcal{C}_2^*}.$$

On the respective sets we get the following estimates:

(i) In  $\mathcal{C}_1^*$  we have  $u_{|\mathcal{C}_1^*}^* = b_{|\mathcal{C}_1^*}$ . Thus,

$$|\Psi^*(\mathcal{C}_1^*)| \leq \frac{1}{2} (\|M_h p_h^* - p_h^*\|_{0,\mathcal{C}_1^*} + \|p_h^* - p^*\|_{0,\mathcal{C}_1^*} + \alpha \|u_h^* - b\|_{0,\mathcal{C}_1^*}) \|u_h^* - b\|_{0,\mathcal{C}_1^*}.$$

Given  $\mathcal{C}_1^*$  and the discrete control  $u_h^*$  and adjoint state  $p_h^*$ , the first and third terms in parentheses above are computable a posteriori. We therefore study  $\|p_h^* - p^*\|_{0,\mathcal{C}_1^*}$  next. Since  $p_h^*, p^* \in H_0^1(\Omega)$  and, for  $n \geq 2$ ,  $H_0^1(\Omega) \subset L^s(\Omega)$  for some  $s \in (2, +\infty)$ , from Hölder's inequality we obtain

$$(2.28) \quad \|p_h^* - p^*\|_{0,\mathcal{C}_1^*} \leq \text{meas}(\mathcal{C}_1^*)^{r(s)} \|p^* - p_h^*\|_{1,\mathcal{C}_1^*} \leq C_4 \text{meas}(\mathcal{C}_1^*)^{r(s)} \eta_4$$

with  $r(s) := \frac{1}{2} - \frac{1}{s} > 0$ . Hence, we get

$$(2.29) \quad \|p_h^* - p^*\|_{0,\mathcal{C}_1^*} \leq \min \left( C_0^p \eta_{0,p}, C_4 \text{meas}(\mathcal{C}_1^*)^{r(s)} \eta_4 \right) =: C^p(\mathcal{C}_1^*),$$

where  $\eta_{0,p}$  denotes the a posteriori estimator for  $\|p^* - p_h^*\|_{0,\Omega}$  (see Appendix A for its derivation) and  $C_0^p > 0$  is a constant. This yields

$$(2.30) \quad |\Psi^*(\mathcal{C}_1^*)| \leq \frac{1}{2} (\|M_h p_h^* - p_h^*\|_{0,\mathcal{C}_1^*} + C^p(\mathcal{C}_1^*) + \alpha \|u_h^* - b\|_{0,\mathcal{C}_1^*}) \cdot \|u_h^* - b\|_{0,\mathcal{C}_1^*} =: \mu_1(\mathcal{C}_1^*).$$

(ii) In  $\mathcal{C}_2^*$  we use the identities  $\lambda_h^* = M_h p_h^* - \alpha u_h^*$  and  $p^* = \alpha u^*$ . From this, and assuming  $b_h \in L^t(\Omega)$ ,  $2 \leq t \leq s$ , we infer

$$\begin{aligned} 2|\Psi^*(\mathcal{C}_2^*)| &= |(u_h^* - u^*, \lambda_h^*)_{\mathcal{C}_2^*}| \\ &\leq \text{meas}(\mathcal{C}_2^*)^{r(t)} \|b_h - \alpha^{-1} p^*\|_{t,\mathcal{C}_2^*} \|\lambda_h^*\|_{0,\mathcal{C}_2^*} \\ &\leq \text{meas}(\mathcal{C}_2^*)^{r(t)} (\|b_h - \alpha^{-1} p_h^*\|_{t,\mathcal{C}_2^*} + \alpha^{-1} \|p_h^* - p^*\|_{1,\Omega}) \|\lambda_h^*\|_{0,\mathcal{C}_2^*} \\ &\leq \text{meas}(\mathcal{C}_2^*)^{r(t)} (\|b_h - \alpha^{-1} p_h^*\|_{t,\mathcal{C}_2^*} + \alpha^{-1} C_4 \eta_4) \|\lambda_h^*\|_{0,\mathcal{C}_2^*} \end{aligned}$$

with  $r(t) \geq 0$ . Alternatively, we may use (2.17a) for estimating  $\|u_h^* - u^*\|_{0,\mathcal{M}_2^*}$ . Hence, setting

$$C^u(\mathcal{C}_2^*) := \min \left( \text{meas}(\mathcal{C}_2^*)^{r(t)} (\|b_h - \alpha^{-1} p_h^*\|_{t,\mathcal{C}_2^*} + \alpha^{-1} C_4 \eta_4), C_3 \eta_3 \right),$$

we obtain

$$(2.31) \quad |\Psi^*(\mathcal{C}_2^*)| \leq \frac{1}{2} C^u(\mathcal{C}_2^*) \|\lambda_h^*\|_{0,\mathcal{C}_2^*} =: \mu_2(\mathcal{C}_2^*).$$

Since  $\lambda_h^* = 0$  in  $\mathcal{I}_h^*$ , we obviously have  $\mu_2(\mathcal{I}_h^*) = 0$ .

In both cases above we assume  $\mu_1(\emptyset) = 0$  and  $\mu_2(\emptyset) = 0$ . Summarizing, we obtain

$$\begin{aligned} |\Psi^*(\Omega)| &= |\Psi^*(\mathcal{A}^* \cap \mathcal{A}_h^*) + \Psi^*(\mathcal{C}_1^*) + \Psi^*(\mathcal{C}_2^*)| \\ &\leq \frac{1}{2} |\text{osc}_h^{\mathcal{A}^* \cap \mathcal{A}_h^*}(b; \lambda^* + \lambda_h^*)| + \mu_1(\mathcal{C}_1^*) + \mu_2(\mathcal{C}_2^*). \end{aligned}$$

An alternative (and possibly coarse) estimate of  $\Psi^*(\Omega)$  uses only the error estimate  $\eta_3$  and  $\|\lambda_h^*\|_{0,\mathcal{A}_h^*}$  as follows:

$$|\Psi^*(\Omega)| = \frac{1}{2} |(\lambda_h^* + \lambda^*, u^* - u_h^*)| \leq \frac{1}{2} C_3 \eta_3 (C_3 \eta_3 + 2 \|\lambda_h^*\|_{0,\mathcal{A}_h^*}) =: \mu_3(\Omega).$$

If the original problem is unconstrained with respect to  $u$ , then  $\lambda^* = 0$ . As a consequence, the first order conditions yield  $\alpha u^* = p^*$ , i.e.,  $u^*$  inherits the regularity of  $p^* \in H_0^1(\Omega)$ . Then we may choose the same ansatz when discretizing  $p$  and  $u$ . Thus, we obtain  $\eta_1 = 0$ , since  $M_h$  becomes the identity operator, and—up to data oscillations— $\eta_2 = \eta_3$ , and further,  $\|M_h p_h^* - p_h^*\|_{0,C_1^*} = 0$  in  $\mu_1$ .

Finally, we express  $\mu_1$  and  $\mu_2$  such that we obtain cell-oriented error estimates. Let us first consider  $\mu_1(\mathcal{C}_1^*)$ . We have

$$\begin{aligned}\mu_1(\mathcal{C}_1^*) &= \frac{1}{2} (C^p(\mathcal{C}_1^*) + \|M_h p_h^* - p_h^*\|_{0,C_1^*} + \alpha \|u_h^* - b\|_{0,C_1^*}) \|u_h^* - b\|_{0,C_1^*} \\ &= \frac{1}{2} \left( \hat{C}^p(\mathcal{C}_1^*) + \hat{C}_5(\mathcal{C}_1^*) \|M_h p_h^* - p_h^*\|_{0,C_1^*}^2 + \alpha \|u_h^* - b\|_{0,C_1^*}^2 \right).\end{aligned}$$

Above, we use

$$\hat{C}_0^p := \begin{cases} C_0^p \frac{\|u_h^* - b\|_{0,C_1^*}}{\eta_{0,p}} & \text{if } \text{meas}(\mathcal{C}_1^*) \neq 0 \text{ and } \eta_{0,p} > 0, \\ 0 & \text{if } \text{meas}(\mathcal{C}_1^*) = 0, \end{cases}$$

as well as

$$\hat{C}_4(\mathcal{C}_1^*) := \begin{cases} C_4 \frac{\|u_h^* - b\|_{0,C_1^*}}{\eta_4} & \text{if } \text{meas}(\mathcal{C}_1^*) \neq 0 \text{ and } \eta_4 > 0, \\ 0 & \text{if } \text{meas}(\mathcal{C}_1^*) = 0, \end{cases}$$

and further,

$$\hat{C}_5(\mathcal{C}_1^*) := \begin{cases} \frac{\|u_h^* - b\|_{0,C_1^*}}{\|M_h p_h^* - p_h^*\|_{0,C_1^*}} & \text{if } \text{meas}(\mathcal{C}_1^*) \neq 0 \text{ and } \|M_h p_h^* - p_h^*\|_{0,C_1^*} > 0, \\ 0 & \text{if } \text{meas}(\mathcal{C}_1^*) = 0. \end{cases}$$

We therefore have

$$\hat{C}^p(\mathcal{C}_1^*) = \min \left( \hat{C}_0^p \eta_{0,p}^2, \hat{C}_4(\mathcal{C}_1^*) \text{meas}(\mathcal{C}_1^*)^{r(s)} \eta_4^2 \right).$$

Finally, we turn to  $\mu_2(\mathcal{C}_2^*)$ . We obtain

$$\mu_2(\mathcal{C}_2^*) = \frac{1}{2} \hat{C}^u(\mathcal{C}_2^*),$$

with

$$\hat{C}_i(\mathcal{C}_2^*) := \begin{cases} C_i \frac{\|\lambda_h^*\|_{0,C_2^*}}{\eta_i} & \text{if } \text{meas}(\mathcal{C}_2^*) \neq 0 \text{ and } \eta_i > 0, \\ 0 & \text{if } \text{meas}(\mathcal{C}_2^*) = 0, \end{cases}$$

for  $i = 3, 4$ , and

$$\begin{aligned}\hat{C}^u(\mathcal{C}_2^*) &:= \min \left( \text{meas}(\mathcal{C}_2^*)^{r(t)} (\|b_h - \alpha^{-1} p_h^*\|_{t,C_2^*} \|\lambda_h^*\|_{0,C_2^*} \right. \\ &\quad \left. + \alpha^{-1} \hat{C}_4(\mathcal{C}_2^*) \eta_4^2), \hat{C}_3(\mathcal{C}_2^*) \eta_3^2 \right).\end{aligned}$$

We summarize our above findings in the following proposition.

**PROPOSITION 2.1.** *Let the assumptions of Theorem 2.1 be satisfied. Then*

$$(2.32) \quad |\Psi^*(\Omega)| \leq \min \left( \frac{1}{2} |\text{osc}_h^{\mathcal{A}^* \cap \mathcal{A}_h^*}(b; \lambda^* + \lambda_h^*)| + \mu_1(\mathcal{C}_1^*) + \mu_2(\mathcal{C}_2^*), \mu_3(\Omega) \right).$$



We denote the right-hand side in (2.32) by  $\hat{\nu}$ . In the case where the solution of (P) satisfies  $u^* < b$  a.e. on  $\Omega$ , we expect that  $\hat{\nu} \approx 0$ . Indeed, for sufficiently small  $h$  we have  $\lambda_h^* \approx 0$  (or even  $\lambda_h^* = 0$ ). Thus,  $\mu_2(\mathcal{C}_2^*) \approx 0$  (or  $\mu_2(\mathcal{C}_2^*) = 0$ ) holds true. Further,  $\mu_1(\mathcal{C}_1^*) = 0$  since  $\mathcal{A}^* = \emptyset$ . Then (2.32) yields  $\hat{\nu} \approx 0$  (or  $\hat{\nu} = 0$ ). If (P) involves no inequality constraints on  $u$ , which means that we can set  $b \equiv +\infty$  on  $\Omega$ , then we naturally obtain  $\hat{\nu} = 0$ . Hence, we recover the error estimator for unconstrained optimal control problems; compare [2, 14].

For deriving the full error estimate, it remains to consider the first term in parentheses on the right-hand side of (2.16) in Theorem 2.2. This term is independent of the control constraints and corresponds to the usual expression obtained for (unconstrained) optimal control problems; see [2, 14]. A standard argument yields

$$\begin{aligned}
 & |(\nabla y_h^*, \nabla(i_h^p p^* - p^*))_{0,\Omega} - (u_h^* + f_h, i_h^p p^* - p^*)_{0,\Omega}| \\
 (2.33) \quad & \leq \sum_T \|-\Delta y_h^* - u_h^* - f_h\|_{0,T} \|p^* - i_h^p p^*\|_{0,T} \\
 & \quad + \sum_F \left\| \left[ \frac{\partial y_h^*}{\partial n} \right] \right\|_{0,F} \|p^* - i_h^p p^*\|_{0,F} =: \eta_2^p
 \end{aligned}$$

for the primal equation,

$$\begin{aligned}
 & |(y_h^* - z_h, i_h^y y^* - y^*)_{0,\Omega} + (\nabla(i_h^y y^* - y^*), \nabla p_h^*)_{0,\Omega}| \\
 (2.34) \quad & \leq \sum_T \|-\Delta p_h^* + y_h^* - z_h\|_{0,T} \|y^* - i_h^y y^*\|_{0,T} \\
 & \quad + \sum_F \left\| \left[ \frac{\partial p_h^*}{\partial n} \right] \right\|_{0,F} \|y^* - i_h^y y^*\|_{0,F} =: \eta_2^d
 \end{aligned}$$

for the dual equation, and

$$(2.35) \quad |(M_h p_h^* - p_h^*, i_h^u u^* - u^*)_{0,\Omega}| =: \eta_2^u.$$

The overall residual- and complementarity-based error estimate is given in the following theorem.

**THEOREM 2.3.** *Let the assumptions of Theorem 2.1 be satisfied. Then we have the following error estimate:*

$$\begin{aligned}
 (2.36) \quad & |J(y^*, u^*) - J(y_h^*, u_h^*)| \leq \frac{1}{2}(\eta_2^p + \eta_2^d + \eta_2^u) + \hat{\nu} \\
 & + \frac{1}{2}[C_0^p \eta_{0,p} \|f - f_h\|_{0,\Omega} + C_0^y \eta_{0,y} \|z - z_h\|_{0,\Omega}] \\
 & + |\text{osc}_h(x_h^*)|
 \end{aligned}$$

with  $\eta_2^p$ ,  $\eta_2^d$ ,  $\eta_2^u$ , and  $\hat{\nu}$  defined by (2.33), (2.34), (2.35), and (2.32), respectively. Further,  $C_0^y > 0$  is a constant and  $\eta_{0,y}$  denotes an error estimate for  $\|y_h^* - y^*\|_{0,\Omega}$ . For the definition of  $\eta_{0,p}$  and  $\eta_{0,y}$  see (A.10) and (A.11) in Appendix A.

The numerical evaluation of (2.36) depends on estimates of  $\|i_h^y y^* - y^*\|_{0,T}$ ,  $\|i_h^y y^* - y^*\|_{0,F}$ , and analogously, for  $i_h^p p^* - p^*$ . When discretizing the state and the adjoint state in two dimensions by continuous piecewise linear finite elements, the following

averaging technique, replacing  $\eta_2^p$  and  $\eta_2^d$  in (2.33) and (2.34), respectively, is appropriate:

$$(2.37) \quad \eta_{2,h}^p := \frac{1}{3} \sum_T \left( h_T \| -\Delta y_h^* - u_h^* - f_h \|_{0,T} \sum_{F(T)} h_F^{1/2} \left\| \left[ \frac{\partial p_h^*}{\partial n} \right] \right\|_{0,F} \right) \\ + \sum_F h_F \left\| \left[ \frac{\partial y_h^*}{\partial n} \right] \right\|_{0,F} \left\| \left[ \frac{\partial p_h^*}{\partial n} \right] \right\|_{0,F}$$

for the primal equation, and

$$(2.38) \quad \eta_{2,h}^d := \frac{1}{3} \sum_T \left( h_T \| -\Delta p_h^* + y_h^* - z_h \|_{0,T} \sum_{F(T)} h_F^{1/2} \left\| \left[ \frac{\partial y_h^*}{\partial n} \right] \right\|_{0,F} \right) \\ + \sum_F h_F \left\| \left[ \frac{\partial p_h^*}{\partial n} \right] \right\|_{0,F} \left\| \left[ \frac{\partial y_h^*}{\partial n} \right] \right\|_{0,F}$$

for the dual equation, where  $F(T)$  denotes the edges pertinent to triangle  $T$ . Notice that (2.37) and (2.38) yield typically sharper estimates than residual-based estimators for our model problem; compare (2.17) and [8]. Further observe that we can only expect boundedness of  $\|i_h^u u^* - u^*\|_{0,\Omega}$  in general. However, typically  $\|M_h p_h^* - p_h^*\|_{0,\Omega}$  is small, or, when using the same ansatz for discretizing  $u^*$  as well as  $p^*$ , it is even zero.

For the numerical evaluation of  $\hat{\nu}$  observe that  $\mathcal{I}_h^* \setminus \mathcal{A}^* \subset \mathcal{I}^*$ , and hence  $\lambda_h^* = 0$  and  $\lambda^* = 0$  on this set. Consequently, we obtain

$$\Psi^*(\mathcal{I}_h^* \setminus \mathcal{A}^*) = 0.$$

Next observe that  $\mathcal{I}_h^* = \mathcal{C}_1^* \dot{\cup} (\mathcal{I}_h^* \setminus \mathcal{A}^*)$ . Therefore, we have

$$(2.39) \quad \Psi^*(\mathcal{C}_1^*) = \Psi^*(\mathcal{I}_h^*) - \Psi^*(\mathcal{I}_h^* \setminus \mathcal{A}^*) = \Psi^*(\mathcal{I}_h^*).$$

If  $b_h = b$ , then we obtain  $\Psi^*(\mathcal{A}_h^* \setminus \mathcal{I}^*) = 0$ , and further,

$$(2.40) \quad \Psi^*(\mathcal{C}_2^*) = \Psi^*(\mathcal{A}_h^*) - \Psi^*(\mathcal{A}_h^* \setminus \mathcal{I}^*) = \Psi^*(\mathcal{A}_h^*).$$

The estimates  $\mu_1(\mathcal{C}_1^*)$  and  $\mu_1(\mathcal{C}_2^*)$ , however, do not satisfy relations analogous to (2.39)–(2.40) even when  $b_h = b$ . Hence,  $\hat{\nu}$  is not a posteriori. In order to have a fully a posteriori estimate, we replace  $\hat{\nu}$  in (2.36) by

$$(2.41) \quad \hat{\nu}^a = \min(\mu_1(\mathcal{I}_h^*), \mu_3(\Omega)) + \min(\mu_2(\mathcal{A}_h^*), \mu_3(\Omega)).$$

An alternative technique based on set estimation is the subject of section 3.2.

**3. Extensions.** Now we consider possible extensions of the concept derived in the previous section. We focus on two aspects as follows: (i) effects due to nonlinear PDEs and/or bilateral constraints; and (ii) alternative ways of making  $\hat{\nu}$  fully a posteriori.

**3.1. Semilinear PDEs and bilateral constraints.** Next we assume that the underlying PDE is semilinear as follows:

$$(3.1) \quad A(y) = Bu + f,$$

where the operators  $A$  and  $B$  induce a semilinear form  $\mathfrak{a}(\cdot)(\cdot)$  and a bilinear form  $\mathfrak{b}(\cdot, \cdot)$ , respectively. Hence, the weak form of (3.1) becomes

$$\mathfrak{a}(y)(v) = (f, v)_{0,\Omega} + \mathfrak{b}(u, v) \quad \forall v \in Y.$$

For our arguments to follow, we assume that  $A$  (respectively,  $\mathfrak{a}$ ) is sufficiently often differentiable. Further, we suppose that the control is subject to bilateral constraints, i.e.,

$$a \leq u \leq b \text{ a.e. in } \Omega.$$

The Lagrange function corresponding to the associated minimization problem has the structure

$$\mathcal{L}(x, \lambda_a, \lambda_b) = J(y, u) + \mathfrak{a}(y)(p) - (f, p)_{0,\Omega} - \mathfrak{b}(u, p) + (a - u, \lambda_a)_{0,\Omega} + (u - b, \lambda_b)_{0,\Omega},$$

where  $\lambda_a, \lambda_b \in L^2(\Omega)$  represent the Lagrange multipliers pertinent to the bilateral pointwise constraints. The first order necessary optimality conditions are given by

$$(3.2a) \quad A(y^*) - Bu^* = f,$$

$$(3.2b) \quad A'(y^*)^* p^* + J_y(y^*, u^*) = 0,$$

$$(3.2c) \quad J_u(y^*, u^*) + \lambda_b^* - \lambda_a^* - B^* p^* = 0,$$

$$(3.2d) \quad u^* \geq a, \quad \lambda_a^* \geq 0, \quad (u^* - a, \lambda_a^*)_{0,\Omega} = 0,$$

$$(3.2e) \quad u^* \leq b, \quad \lambda_b^* \geq 0, \quad (u^* - b, \lambda_b^*)_{0,\Omega} = 0.$$

As the pointwise control constraints are affine, the error estimator for the nonlinear case is similar to the linear case. This parallels the situation in [2], where the unconstrained case was considered. Due to essentially the same proof arguments as in [2, Proposition 6.1], the following result holds true. In what follows, we use

$$\mathcal{L}_0(x) = J(y, u) + \mathfrak{a}(y)(p) - (f, p)_{0,\Omega} - \mathfrak{b}(u, p),$$

and use  $\mathcal{L}_{0,h}(x)$  for its discrete counterpart.

**THEOREM 3.1.** *For a Galerkin finite element discretization of the first order necessary optimality conditions (3.2), the following relation holds true:*

$$\begin{aligned} J(y^*, u^*) - J_h(y_h^*, u_h^*) &= \frac{1}{2} \nabla_x \mathcal{L}_{0,h}(x_h^*)(x^* - i_h x^*) \\ &+ \frac{1}{2} [(u_h^* - b, \lambda_b^*)_{0,\Omega} + (b_h - u^*, \lambda_{b,h}^*)_{0,\Omega}] \\ &+ \frac{1}{2} [(a - u_h^*, \lambda_a^*)_{0,\Omega} + (u^* - a_h, \lambda_{a,h}^*)_{0,\Omega}] \\ &+ \frac{1}{2} ((f - f_h, p_h^* - p^*)_{0,\Omega} + (z - z_h, y_h^* - y^*)_{0,\Omega}) + \text{osc}_h(x_h^*) \\ &+ r(x^*, x_h^*), \end{aligned}$$

where  $r(x^*, x_h^*)$  denotes the remainder term of a Taylor expansion of  $\mathcal{L}_0$  about  $x_h^*$ . It is bounded by

$$|r(x^*, x_h^*)| \leq \sup_{\bar{x} \in [x_h^*, x^*]} |\nabla_x^3 \mathcal{L}_0(\bar{x})[x^* - x_h^*]^3|.$$

**3.2. Alternative a posteriori estimate for  $\hat{\nu}$ .** At the end of section 2 we derived an a posteriori estimate for  $\hat{\nu}$ ; recall  $\hat{\nu}^a$  in (2.41), where we replaced  $\mathcal{C}_1^*$  by  $\mathcal{I}_h^*$  and  $\mathcal{C}_2^*$  by  $\mathcal{A}_h^*$ , respectively. This may give rise to an overestimation of the error term pertinent to the complementarity system. In the following we provide an alternative approach based on set estimation.

Assuming, without loss of generality,  $b_h = b$ , we focus on the unilaterally constrained case and start by considering  $\hat{\mu}_1(\mathcal{C}_1^*)$ . For this purpose recall that  $\mathcal{C}_1^* = \mathcal{I}_h^* \cap \mathcal{A}^*$ . Similarly to [11, section 3.3] we estimate the continuous active set  $\mathcal{A}^*$  by

$$\chi_h^{\mathcal{A}^*} = 1 - \frac{b - u_h^*}{\gamma h^r + b - u_h^*},$$

where  $\gamma$  denotes some (possibly small) positive constant, and  $r > 0$  is fixed. Note that  $\chi_h^{\mathcal{A}^*} = 1$  in  $\mathcal{A}_h^*$ . Further, let  $\chi(\mathcal{S})$  denote the characteristic function of a set  $\mathcal{S} \subset \Omega$ . We briefly argue that our approximation is useful. In fact, assume that  $T \subset \mathcal{A}^*$ . Then

$$\|\chi(\mathcal{A}^*) - \chi_h^{\mathcal{A}^*}\|_{0,T} = \left\| \frac{b - u_h^*}{\gamma h^r + b - u_h^*} \right\|_{0,T} \leq \min\{1, \gamma^{-1} h^{-r} \|u^* - u_h^*\|_{0,T}\},$$

which tends to zero whenever  $\|u^* - u_h^*\|_{0,T} = \mathcal{O}(h^q)$  with  $q > r$ . If  $T \in \mathcal{I}^*$ , then we distinguish two cases as follows:

- (i)  $T \subset \{b - u_h^* > \gamma h^{\epsilon r}\}$  for some  $0 \leq \epsilon < 1$ . Then

$$\|\chi(\mathcal{A}^*) - \chi_h^{\mathcal{A}^*}\|_{0,T} = \left\| \frac{\gamma h^r}{\gamma h^r + b - u_h^*} \right\|_{0,T} \leq h^{(1-\epsilon)r} \rightarrow 0 \text{ as } h \rightarrow 0.$$

- (ii) Finally, in the case where  $T \in \{b - u_h^* \leq \gamma h^{\epsilon r}\}$ , we use  $T \subset \mathcal{I}^*$  and  $\|u^* - u_h^*\|_{0,\Omega} \rightarrow 0$  to conclude that the measure of this set tends to zero as  $h \rightarrow 0$ . We therefore use the following approximation of  $\chi(\mathcal{C}_1^*)$ :

$$\chi(\mathcal{C}_1^*) \approx \chi(\mathcal{I}_h^*) \chi_h^{\mathcal{A}^*} =: \chi_h^{\mathcal{C}_1^*}.$$

In the definition of  $\mu_1(\mathcal{C}_1^*)$ , we then use

$$\|\chi_h^{\mathcal{C}_1^*}(u_h^* - b)\|_{0,\Omega} \quad \text{instead of} \quad \|u_h^* - b\|_{0,\mathcal{C}_1^*}$$

and analogously for  $\|M_h p_h^* - p_h^*\|_{0,\mathcal{C}_1^*}$ . Further, the measure of  $\mathcal{C}_1^*$  is approximated by

$$\text{meas}(\mathcal{C}_1^*) \approx \int_{\Omega} \chi_h^{\mathcal{C}_1^*} dx.$$

The definition of  $\mu_2$  involves the set  $\mathcal{C}_2^* = \mathcal{A}_h^* \cap \mathcal{I}^*$ . Here we employ the approximation

$$\chi_h^{\mathcal{C}_2^*} := \chi(\mathcal{A}_h^*) \chi_h^{\mathcal{I}^*}$$

with  $\chi_h^{\mathcal{I}^*} = 1 - \chi_h^{\mathcal{A}^*}$ . Then we replace  $\|\lambda_h^*\|_{0,\mathcal{C}_2^*}$  by  $\|\chi_h^{\mathcal{C}_2^*} \lambda_h^*\|_{0,\Omega}$ ,  $\|b - \alpha p_h^*\|_{t,\mathcal{C}_2^*}$  by  $\|\chi_h^{\mathcal{C}_2^*} (b - \alpha^{-1} p_h^*)\|_{t,\Omega}$ , and obtain

$$\text{meas}(\mathcal{C}_2^*) \approx \int_{\Omega} \chi_h^{\mathcal{C}_2^*} dx.$$

The extension of this concept to the bilaterally constrained case is straightforward.

**4. Numerics.** For the practical realization of the goal-oriented dual-weighted approach, we follow the cycle SOLVE, ESTIMATE, MARK, and REFINES known from adaptive finite element methods. Here, SOLVE stands for the numerical solution of the discrete optimal control problem which is taken care of by a primal-dual active set strategy [9]. The following step, ESTIMATE, is devoted to the computation of the edge and element residuals of the error estimator  $\eta_h$ , the local components of the consistency error  $\hat{\nu}_h$ , and the data oscillations. We note that

$$(4.1) \quad \eta_h := \eta_{2,h}^p + \eta_{2,h}^d + \eta_{2,h}^u.$$

Here,  $\eta_{2,h}^p$  and  $\eta_{2,h}^d$  are given by (2.38) and (2.37). Moreover,  $\eta_{2,h}^u$  is given by (2.35) with  $i_h^u u^* - u^*$  replaced by  $u_h^* - \bar{u}_h^*$ , where  $\bar{u}_h^*|_T := |T|^{-1} \int_T u_h^* dx$ ,  $T \in \mathbb{T}_h$ . We refer to  $\eta_{2,T}^p$ ,  $\eta_{2,T}^d$ ,  $\eta_{2,T}^u$ ,  $T \in \mathbb{T}_h$ , as the elementwise contributions to  $\eta_{2,h}^p$ ,  $\eta_{2,h}^d$ , and  $\eta_{2,h}^u$ , respectively, so that

$$\eta_h = \sum_{T \in \mathbb{T}_h} \eta_T, \quad \eta_T := \eta_{2,T}^p + \eta_{2,T}^d + \eta_{2,T}^u.$$

Likewise, we have

$$(4.2) \quad \hat{\nu}_h = \sum_{T \in \mathbb{T}_h} \hat{\nu}_T^a,$$

where  $\hat{\nu}_T^a$ ,  $T \in \mathbb{T}_h$ , are the elementwise contributions to  $\hat{\nu}^a$  that can be easily deduced from (2.32). Finally, we summarize the remaining terms of (2.36) in Theorem 2.3 according to

$$(4.3) \quad \text{osc}_h := \frac{1}{2} [C_0^p \eta_{0,p} \|f - f_h\|_{0,\Omega} + C_0^y \eta_{0,y} \|z - z_h\|_{0,\Omega}] + |\text{osc}_h(x_h^*)|$$

and observe

$$\text{osc}_h = \sum_{T \in \mathbb{T}_h} \text{osc}_T,$$

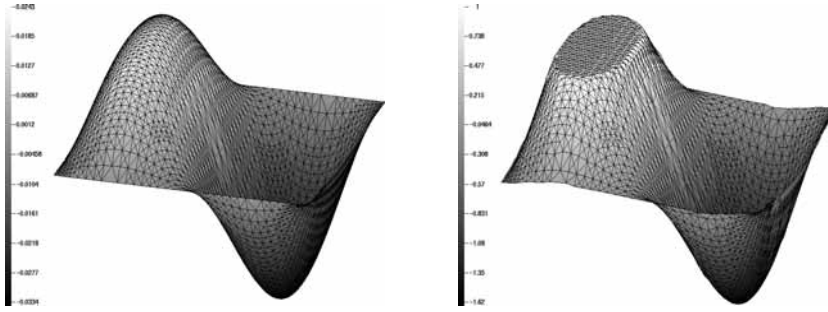
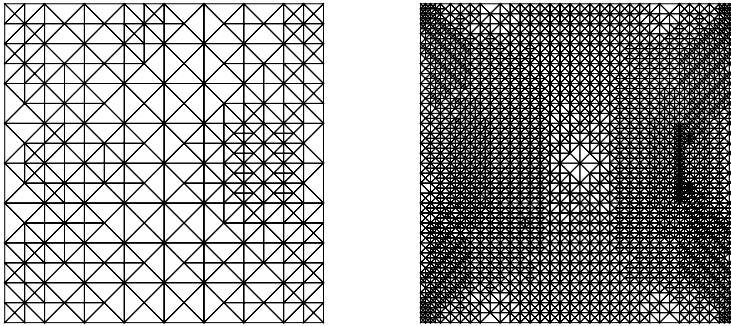
where again  $\text{osc}_T$ ,  $T \in \mathbb{T}_h$ , refers to the elementwise contribution to  $\text{osc}_h$ . In the step MARK of the adaptive cycle, we specify constants  $\Theta_i \in (0, 1)$  and select subsets  $\mathbb{M}_i \subset \mathbb{T}_h$ ,  $1 \leq i \leq 3$ , by means of the bulk criteria

$$(4.4a) \quad \Theta_1 \sum_{T \in \mathbb{T}_h} \eta_T \leq \sum_{T \in \mathbb{M}_1} \eta_T,$$

$$(4.4b) \quad \Theta_2 \sum_{T \in \mathbb{T}_h} \hat{\nu}_T^a \leq \sum_{T \in \mathbb{M}_2} \hat{\nu}_T^a,$$

$$(4.4c) \quad \Theta_3 \sum_{T \in \mathbb{T}_h} \text{osc}_T \leq \sum_{T \in \mathbb{M}_3} \text{osc}_T$$

known from the convergence analysis of adaptive finite element methods (cf., e.g., [6, 15]). The bulk criteria can be realized by a greedy algorithm (cf., e.g., [8]). The final step REFINES of the adaptive loop is devoted to the creation of a new refined mesh based on longest edge bisection of any element  $T \in \mathbb{T}_h$  that has been marked, i.e.,  $T \in \bigcup_{i=1}^3 \mathbb{M}_i$ .

FIG. 4.1. *Example 1: Optimal state (left) and optimal control (right).*FIG. 4.2. *Example 1: Adaptively refined grid after 6 (left) and 10 (right) refinement steps ( $\Theta_i = 0.6$ ,  $1 \leq i \leq 3$ ).*

Finally, we provide documentation of numerical results illustrating the performance of the goal-oriented dual-weighted approach for two representative distributed optimal control problems that have been considered in [8] in the framework of an error analysis of residual-type a posteriori error estimators for control constrained optimal control problems.

*Example 1* (constant obstacle). This first example features a constant obstacle. The data are as follows:

$$\Omega := (0, 1)^2, \quad z := \begin{cases} 200x_1x_2(x_1 - 0.5)^2(1 - x_2) & \text{if } 0 \leq x_1 \leq 0.5, \\ 200(x_1 - 1)x_2(x_1 - 0.5)^2(1 - x_2) & \text{if } 0.5 < x_1 \leq 1, \end{cases}$$

$$\alpha := 0.01, \quad b := 1, \quad f := 0.$$

Figures 4.1 and 4.2 show a visualization of the optimal state and the optimal control as well as the adaptively refined grids after 6 and 10 refinement steps in the case when  $\Theta_i = 0.6$ ,  $1 \leq i \leq 3$  in the bulk criteria (4.4). The active region is an ellipse (cf. the plateau in Figure 4.1 (right)). The convergence history of the adaptive loop is displayed in Table 4.1, containing the total number of degrees of freedom  $N_{DOF}$ , the error  $\delta_h := |J(y^*, u^*) - J_h(y_h^*, u_h^*)|$  in the objective functional, the error estimator  $\eta_h$ , the consistency error  $\hat{\nu}_h^a$ , and the data oscillations  $\text{osc}_h$ . Finally, Figure 4.3 shows the error  $\delta_h$  as a function of the total number of degrees of freedom in the case of adaptive refinement (solid line) and uniform refinement (dotted line). Since in this example the optimal state and adjoint state are smooth, there is only a slight benefit gained when using the adaptive process.

TABLE 4.1

Example 1: Convergence history of the goal-oriented dual-weighted approach.

$\ell$	$N_{\text{dof}}$	$\delta_h$	$\eta_h$	$\hat{\nu}_h^a$	$\text{osc}_h$
0	12	2.73e-03	1.47e-02	0.00e+00	1.17e-01
1	25	8.57e-04	2.03e-02	2.04e-03	6.23e-02
2	42	5.09e-04	1.42e-02	4.86e-03	3.44e-02
3	80	2.54e-04	7.63e-03	3.13e-03	2.17e-02
4	138	1.52e-04	4.61e-03	1.66e-04	1.27e-02
5	282	7.32e-05	2.30e-03	1.62e-05	7.26e-03
6	478	4.24e-05	1.35e-03	3.67e-05	4.20e-03
7	928	1.77e-05	6.45e-04	1.43e-05	5.24e-03
8	1706	9.91e-06	3.67e-04	4.27e-06	2.08e-03
9	3236	5.13e-06	1.85e-04	1.54e-06	1.20e-03
10	6237	2.52e-06	9.95e-05	3.82e-07	6.60e-04
11	11292	1.42e-06	5.25e-05	1.56e-07	3.73e-04
12	22639	5.92e-07	2.74e-05	1.63e-07	1.63e-04
13	38549	4.20e-07	1.53e-05	4.41e-08	1.12e-04
14	81325	1.57e-07	7.57e-06	7.60e-09	5.05e-05
15	136571	1.17e-07	4.38e-06	6.78e-09	3.24e-05
16	299028	4.65e-08	2.05e-06	1.32e-09	1.58e-05

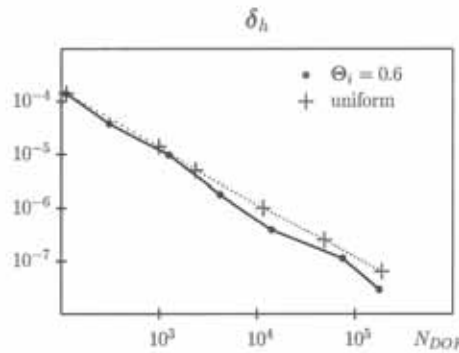


FIG. 4.3. Example 1: Adaptive refinement (solid line) versus uniform refinement (dotted line).

*Example 2* (variable obstacle). This example is constructed in such a way that there is a lack of strict complementarity. It differs from the general setting insofar as the term containing the control in the objective functional additionally includes a fixed shift control  $w \in L^2(\Omega)$  as follows:

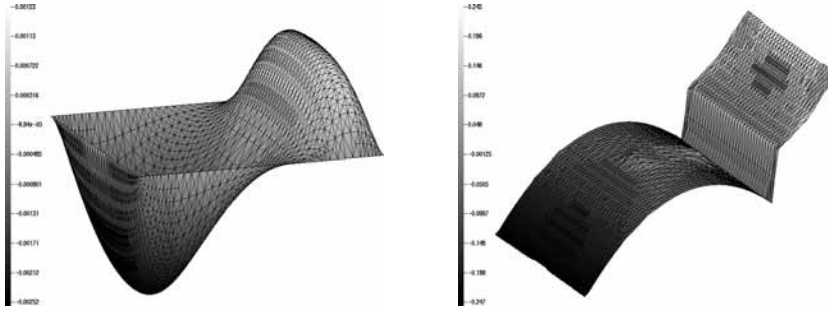
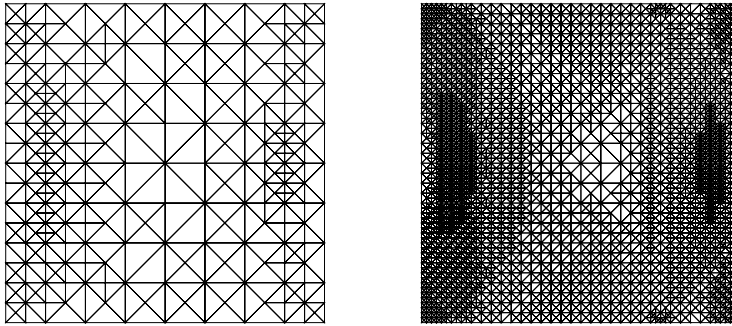
$$J(y, u) := \frac{1}{2} \|y - z\|_{0,\Omega}^2 + \frac{\alpha}{2} \|u - w\|_{0,\Omega}^2.$$

The data are as follows:

$$\begin{aligned} \Omega &:= (0, 1)^2, \quad z := 0, \quad w := \hat{u} + \alpha^{-1} (\hat{\sigma} + \Delta^{-2} \hat{u}), \\ b &:= \begin{cases} (x_1 - 0.5)^8 & \text{if } (x_1, x_2) \in \Omega_1, \\ (x_1 - 0.5)^2 & \text{otherwise,} \end{cases} \quad \alpha := 0.1, \quad f := 0. \end{aligned}$$

Here,  $\hat{u}$  and  $\hat{\sigma}$  are given by

$$\hat{u} := \begin{cases} b(x_1, x_2) & \text{if } (x_1, x_2) \in \Omega_1 \cup \Omega_2, \\ -1.01 b(x_1, x_2) & \text{otherwise} \end{cases}$$

FIG. 4.4. *Example 2: Optimal state (left) and optimal control (right).*FIG. 4.5. *Example 2: Adaptively refined grid after 6 (left) and 10 (right) refinement steps ( $\Theta_i = 0.6$ ,  $1 \leq i \leq 3$ ).*

and

$$\hat{\sigma} := \begin{cases} 2.25 (x_1 - 0.75) \cdot 10^{-4} & \text{if } (x_1, x_2) \in \Omega_2, \\ 0 & \text{otherwise} \end{cases}$$

with  $\Omega_1$  and  $\Omega_2$  specified as follows:

$$\begin{aligned} \Omega_1 &:= \{(x_1, x_2) \in \Omega \mid ((x_1 - 0.5)^2 + (x_2 - 0.5)^2)^{1/2} \leq 0.15\}, \\ \Omega_2 &:= \{(x_1, x_2) \in \Omega \mid x_1 \geq 0.75\}. \end{aligned}$$

We note that  $\Omega_2$  corresponds to the strongly active set (where strict complementarity holds true, i.e.,  $\lambda^* > 0$  a.e. in  $\Omega_2$ ), whereas the set  $\Omega_1 \cup \{(x_1, x_2) \in \Omega \mid x_1 = 0.5\}$  represents the weakly active set, where strict complementarity does not hold true, i.e.,  $\lambda^* = 0$  a.e. in this set.

The shift control  $w \in L^2(\Omega)$  is approximated by  $w_h \in L_h$ , giving rise to an additional term in the data oscillations  $\text{osc}_h(x_h^*)$ .

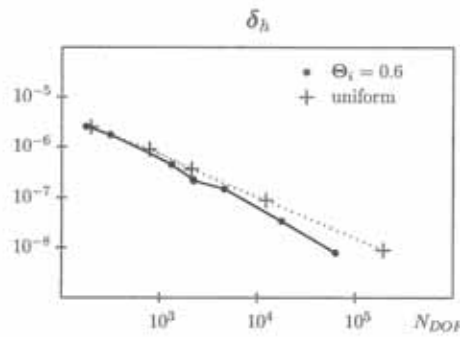
Figure 4.4 displays the computed optimal state and optimal control. Figure 4.5 shows the adaptively refined grids after 6 and 10 refinements steps, where we have chosen  $\Theta_i = 0.6$ ,  $1 \leq i \leq 3$  in the bulk criteria (4.4). Table 4.2 reflects the convergence history of the refinement process in terms of the same data as in the first example, and Figure 4.6 shows the comparison between adaptive and uniform refinement. In this example, the benefits of adaptive refinement are more pronounced than in the previous one.



TABLE 4.2

*Example 2: Convergence history of the goal-oriented dual-weighted approach.*

$\ell$	$N_{\text{dof}}$	$\delta_h$	$\eta_h$	$\hat{\nu}_h^a$	$\text{osc}_h$
0	5	2.41e-04	2.58e-06	0.00e+00	1.07e-01
1	12	1.61e-04	5.26e-06	2.71e-07	8.11e-02
2	26	7.62e-05	4.78e-06	4.19e-07	5.25e-02
3	43	3.50e-05	3.69e-06	5.82e-07	3.71e-02
4	73	1.54e-05	2.08e-06	0.00e+00	2.89e-02
5	133	8.59e-06	1.29e-06	0.00e+00	2.22e-02
6	253	4.09e-06	6.45e-07	0.00e+00	1.59e-02
7	475	2.38e-06	3.78e-07	8.08e-11	1.17e-02
8	953	1.16e-06	1.79e-07	9.86e-12	8.39e-03
9	1776	6.44e-07	9.86e-08	1.79e-12	6.05e-03
10	3507	3.41e-07	4.87e-08	2.66e-13	4.70e-03
11	6645	1.82e-07	2.64e-08	7.94e-14	3.34e-03
12	12684	1.03e-07	1.33e-08	3.08e-14	2.59e-03
13	24746	5.36e-08	7.06e-09	1.25e-14	1.91e-03
14	45486	2.99e-08	3.71e-09	2.23e-15	1.52e-03
15	90991	1.57e-08	1.91e-09	1.75e-15	1.13e-03
16	165366	8.12e-09	1.05e-09	2.65e-16	9.06e-04

FIG. 4.6. *Example 2: Adaptive refinement (solid line) versus uniform refinement (dotted line).*

It should be noted that in both examples there is comparably less refinement than in the case of the residual-type a posteriori error estimator from [8]. This does not come as a surprise. As we noted before, the error estimation derived from the goal-oriented dual-weighted approach provides a finer estimate, since the residual-type upper bound can be derived from it by further estimation. On the other hand, the residual-type estimator from [8] has been designed for an estimation of the errors in the state, the adjoint state, the control, and the adjoint control. Therefore, a more pronounced refinement has to be expected.

**Appendix A. A posteriori estimates in the  $L^2$ -norm.** In this section we derive a posteriori error estimates for  $\|p^* - p_h^*\|_{0,\Omega}$  and  $\|y^* - y_h^*\|_{0,\Omega}$ . The subsequent proof technique is based on a combination of the approaches in [8] and [17].

In what follows, we assume that  $\Omega$  is convex and that  $b = b_h$  a.e. in  $\Omega$ , and we use  $a(y, w) = (\nabla y, \nabla w)_{0,\Omega}$ . Given  $u_h^* \in L_h$ , by  $y(u_h^*), p(u_h^*) \in H_0^1(\Omega)$  we denote the solutions to

$$\begin{aligned} a(y(u_h^*), v) &= (f + u_h^*, v)_{0,\Omega}, \\ a(p(u_h^*), v) &= (z - y(u_h^*), v)_{0,\Omega} \end{aligned}$$

for all  $v \in H_0^1(\Omega)$ . The Poincaré–Friedrichs inequality yields

$$(A.1) \quad \|p(u_h^*) - p^*\|_{0,\Omega} \leq c(\Omega) \|y(u_h^*) - y^*\|_{0,\Omega},$$

$$(A.2) \quad \|y(u_h^*) - y^*\|_{0,\Omega} \leq c(\Omega) \|u_h^* - u^*\|_{0,\Omega},$$

where we assume that  $y^* \in H_0^1(\Omega)$  satisfies  $a(y^*, v) = (f + u^*, v)_{0,\Omega}$  for all  $v \in H_0^1(\Omega)$ , and  $c(\Omega)$  is a constant depending only on the domain  $\Omega$ . Hence, for  $p^* \in H_0^1(\Omega)$  satisfying  $a(p^*, v) = (z - y^*, v)_{0,\Omega}$  for all  $v \in H_0^1(\Omega)$ , we get

$$(A.3) \quad \|p^* - p_h^*\|_{0,\Omega} \leq \|p(u_h^*) - p_h^*\|_{0,\Omega} + c(\Omega)^2 \|u_h^* - u^*\|_{0,\Omega}.$$

Next let us assume that  $u^*$ , respectively,  $u_h^*$ , satisfies the system

$$\alpha u^* - p^* + \lambda^* = 0 \quad \text{and} \quad \alpha u_h^* - M_h p_h^* + \lambda_h^* = 0.$$

Then we obtain

$$(A.4) \quad \begin{aligned} \alpha \|u^* - u_h^*\|_{0,\Omega}^2 &\leq (\lambda_h^* - \lambda^*, u^* - u_h^*)_{0,\Omega} + (p^* - p_h^*, u^* - u_h^*)_{0,\Omega} \\ &\quad + \frac{\alpha}{4} \|u^* - u_h^*\|_{0,\Omega}^2 + \frac{1}{\alpha} \|p_h^* - M_h p_h^*\|_{0,\Omega}^2 \\ &\leq (p^* - p_h^*, u^* - u_h^*)_{0,\Omega} + \frac{\alpha}{4} \|u^* - u_h^*\|_{0,\Omega}^2 \\ &\quad + \frac{1}{\alpha} \|p_h^* - M_h p_h^*\|_{0,\Omega}^2 \end{aligned}$$

since  $(\lambda_h^* - \lambda^*, u^* - u_h^*)_{0,\Omega} \leq 0$ . One also has

$$(p^* - p(u_h^*), u^* - u_h^*)_{0,\Omega} \leq 0.$$

Hence, we have

$$\begin{aligned} (p^* - p_h^*, u^* - u_h^*)_{0,\Omega} &\leq (p(u_h^*) - p_h^*, u^* - u_h^*)_{0,\Omega} \\ &\leq \frac{\alpha}{4} \|u^* - u_h^*\|_{0,\Omega}^2 + \frac{1}{\alpha} \|p_h^* - p(u_h^*)\|_{0,\Omega}^2. \end{aligned}$$

This allows us to continue (A.4) as follows:

$$(A.5) \quad \|u^* - u_h^*\|_{0,\Omega}^2 \leq \frac{2}{\alpha^2} \|p_h^* - p(u_h^*)\|_{0,\Omega}^2 + \frac{2}{\alpha^2} \|p_h^* - M_h p_h^*\|_{0,\Omega}^2.$$

Combining the above estimates results in

$$(A.6) \quad \begin{aligned} \|p^* - p_h^*\|_{0,\Omega} &\leq \left(1 + \frac{\sqrt{2}}{\alpha} c(\Omega)^2\right) \|p_h^* - p(u_h^*)\|_{0,\Omega} \\ &\quad + \frac{\sqrt{2}}{\alpha} c(\Omega)^2 \|p_h^* - M_h p_h^*\|_{0,\Omega}, \end{aligned}$$

$$(A.7) \quad \begin{aligned} \|y^* - y_h^*\|_{0,\Omega} &\leq \|y(u_h^*) - y_h^*\|_{0,\Omega} + \frac{\sqrt{2}}{\alpha} c(\Omega) (\|p_h^* - p(u_h^*)\|_{0,\Omega} \\ &\quad + \|p_h^* - M_h p_h^*\|_{0,\Omega}). \end{aligned}$$

Utilizing standard  $L^2$ -estimates (see, e.g., [17, Proposition 3.8]) we infer

$$(A.8) \quad \|y(u_h^*) - y_h^*\|_{0,\Omega}^2 \leq C \left( \sum_T h_T^2 \eta_{y,T}^2 + \sum_F h_F^2 \eta_{y,F}^2 \right) =: C \tilde{\eta}_{0,y}^2,$$

$$(A.9) \quad \|p(u_h^*) - p_h^*\|_{0,\Omega}^2 \leq C \left( \sum_T h_T^2 \tilde{\eta}_{p,T}^2 + \sum_F h_F^2 \eta_{p,F}^2 \right) =: C \tilde{\eta}_{0,p}^2,$$

where the element and edge residuals are given by

$$\begin{aligned}\eta_{y,T} &:= h_T \|f + u_h^*\|_{0,T}, \\ \eta_{y,F} &:= h_F^{1/2} \|n_F \cdot [\nabla y_h^*]\|_{0,F}, \\ \tilde{\eta}_{p,T} &:= h_T \|z - y(u_h^*)\|_{0,T}, \\ \eta_{p,F} &:= h_F^{1/2} \|n_F \cdot [\nabla p_h^*]\|_{0,F}\end{aligned}$$

with  $n_F$  denoting the exterior unit normal of  $T$ . The triangle inequality yields

$$\sum_T h_T^4 \|z - y(u_h^*)\|_{0,T}^2 \leq C h^2 \tilde{\eta}_{0,y}^2 + 2 \sum_T h_T^2 \eta_{p,T}^2$$

with the element residual

$$\eta_{p,T} := h_T^2 \|z - y_h^*\|_{0,T}.$$

Finally, we derive the estimate

$$\begin{aligned}\|p^* - p_h^*\|_{0,\Omega} &\leq C \left( h^2 \tilde{\eta}_{0,y}^2 + \sum_T h_T^2 \eta_{p,T}^2 + \sum_F h_F^2 \eta_{p,F}^2 \right)^{1/2} \\ (A.10) \quad &+ \frac{\sqrt{2}}{\alpha} c(\Omega)^2 \|p_h^* - M_h p_h^*\|_{0,\Omega} + \text{osc}_{0,h}(z) + \text{osc}_{0,h}(f) \\ &=: C_0^p \eta_{0,p} + \text{osc}_{0,h}(z) + \text{osc}_{0,h}(f),\end{aligned}$$

where the data oscillations are given by

$$\begin{aligned}\text{osc}_{0,h}(z) &= \left( \sum_T h_T^2 \text{osc}_T(z)^2 \right)^{1/2}, \\ \text{osc}_T(z) &= h_T \|z - z_h\|_{0,T}\end{aligned}$$

and analogously for  $\text{osc}_{0,h}(f)$ .

The error in the state is estimated a posteriori by

$$\begin{aligned}\|y^* - y_h^*\|_{0,\Omega} &\leq C \tilde{\eta}_{0,y} + \frac{\sqrt{2}}{\alpha} c(\Omega) (\tilde{\eta}_{0,p} + \|p_h^* - M_h p_h^*\|_{0,\Omega}) \\ (A.11) \quad &+ \text{osc}_{0,h}(f) + \text{osc}_{0,h}(z) \\ &=: C_0^y \eta_{0,y} + \text{osc}_{0,h}(f) + \text{osc}_{0,h}(z).\end{aligned}$$

## REFERENCES

- [1] W. BANGERTH AND R. RANNACHER, *Adaptive Finite Element Methods for Differential Equations*, Birkhäuser, Zürich, 2003.
- [2] R. BECKER, H. KAPP, AND R. RANNACHER, *Adaptive finite element methods for optimal control of partial differential equations: Basic concept*, SIAM J. Control Optim., 39 (2000), pp. 113–132.
- [3] R. BECKER AND R. RANNACHER, *An optimal control approach to error estimation and mesh adaptation in finite element methods*, Acta Numer., 10 (2001), pp. 1–102.
- [4] H. BLUM AND F. T. SUTTMEIER, *An adaptive finite element discretization for a simplified Signorini problem*, Calcolo, 37 (2000), pp. 65–77.
- [5] D. BRAESS, C. CARSTENSEN, AND R. H. W. HOPPE, *Convergence analysis of a conforming finite element method for an obstacle problem*, Numer. Math., 107 (2007), pp. 455–471.

- [6] W. DÖRFLER, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
- [7] K. ERIKSSON, D. ESTEP, P. HANSBO, AND C. JOHNSON, *Introduction to adaptive methods for differential equations*, Acta Numer., 1995, pp. 105–158.
- [8] M. HINTERMÜLLER, R. H. W. HOPPE, Y. ILIASH, AND M. KIEWEG, *An a posteriori error analysis of adaptive finite element methods for distributed elliptic control problems with control constraints*, ESAIM Control Optim. Calc. Var., forthcoming.
- [9] M. HINTERMÜLLER, K. ITO, AND K. KUNISCH, *The primal-dual active set strategy as a semi-smooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.
- [10] R. H. W. HOPPE AND R. KORNUBER, *Adaptive multilevel methods for obstacle problems*, SIAM J. Numer. Anal., 31 (1994), pp. 301–323.
- [11] R. LI, W. LIU, H. MA, AND T. TANG, *Adaptive finite element approximation for distributed elliptic optimal control problems*, SIAM J. Control Optim., 41 (2002), pp. 1321–1349.
- [12] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, Heidelberg, New York, 1971.
- [13] W. LIU AND N. YAN, *A posteriori error estimates for a class of variational inequalities*, J. Sci. Comput., 15 (2000), pp. 361–393.
- [14] W. LIU AND N. YAN, *A posteriori error estimates for distributed convex optimal control problems*, Adv. Comput. Math., 15 (2001), pp. 285–309.
- [15] P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.
- [16] R. H. NOCHETTO, K. G. SIEBERT, AND A. VEESER, *Fully localized a posteriori error estimators and barrier sets for contact problems*, SIAM J. Numer. Anal., 42 (2005), pp. 2118–2135.
- [17] R. VERFÜRTH, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester, UK, 1996.

## PARTIAL INFORMATION LINEAR QUADRATIC CONTROL FOR JUMP DIFFUSIONS\*

YAOZHONG HU<sup>†</sup> AND BERNT ØKSENDAL<sup>‡</sup>

**Abstract.** We study a stochastic control problem in which the state process is described by a stochastic differential equation (SDE) that is driven by a Brownian motion and a Poisson random measure, and is affine in both the state and the control. The performance functional is quadratic in the state and the control. All the coefficients are allowed to be random and non-Markovian. Moreover, we may allow the control to be predictable to a given subfiltration of the filtration of the Brownian motion and the random measure (partial information control).

**Key words.** partial information, linear quadratic control, jump diffusions, backward stochastic Riccati equations

**AMS subject classifications.** Primary, 93E20, 60H05, 60G51; Secondary, 91B28

**DOI.** 10.1137/060667566

**1. Introduction.** The problem of stochastic control is always hard. Only in a few cases is there an explicit solution. There are two important approaches to the general stochastic optimal control problem. One is the Bellman dynamic programming principle, which results in the Hamilton–Jacobi–Bellman equation. This approach is applicable when the controlled system is Markovian. Another important approach is the maximum principle. For detailed accounts of the approaches to systems driven by Brownian motions, see the books [7] and [18] and the references therein.

In this paper we will consider the stochastic optimal control problems in which the controlled system is a jump-diffusion. There have been some recent developments in the case when the controlled system is Markovian. See the book [15] and the references therein. Some explicit control problems arising from finance and their solutions are also presented in this book.

Let  $(W_t, t \geq 0)$  be a Brownian motion and  $(N(ds, dz), s \geq 0, z \in \mathbb{R})$  be a Poisson random measure with the intensity measure given by  $\nu(dz)$ . The compensated Poisson random measure is denoted by  $\tilde{N}(ds, dz)$ . We will consider only the case when the state  $x_t$  at time  $t$  is described by a linear controlled jump-diffusion of the form

$$(1.1) \quad \begin{aligned} dx_t &= [A_t x_t + B_t u_t + \alpha_t] dt + [C_t x_t + D_t u_t + \beta_t] dW_t \\ &\quad + \int_{\mathbb{R}} [E_t(z)x_{t-} + F_t(z)u_t + \gamma_t(z)] \tilde{N}(dt, dz), \quad t \in [0, T], \\ x_0 &= x \in \mathbb{R}. \end{aligned}$$

Here  $u_t$  is our control process and  $A_t, B_t, \alpha_t, C_t, D_t, \beta_t, E_t(z), F_t(z)$ , and  $\gamma_t(z)$  are given  $\mathcal{F}_t$ -predictable processes, where  $\mathcal{F}_t$  is the filtration generated by the Brownian

---

\*Received by the editors August 16, 2006; accepted for publication (in revised form) February 20, 2008; published electronically June 13, 2008. This author was supported in part by the National Science Foundation under grant DMS0504783.

<http://www.siam.org/journals/sicon/47-4/66756.html>

<sup>†</sup>Department of Mathematics, University of Kansas, 405 Snow Hall, Lawrence, KS 66045-2142 (hu@math.ku.edu), and Center of Mathematics for Applications (CMA), Department of Mathematics, University of Oslo, Box 1053 Blindern, N-0316, Oslo, Norway.

<sup>‡</sup>Center of Mathematics for Applications (CMA), Department of Mathematics, University of Oslo, Box 1053 Blindern, N-0316, Oslo, Norway (oksensdal@math.uio.no), and Norwegian School of Economics and Business Administration, Helleveien 30, N-5045, Bergen, Norway.

motion  $W(s)$ ,  $s \leq t$ , and the Poisson random measures are  $\tilde{N}(ds, dz)$ ,  $s \leq t$ . The control  $u_t$  is required to be  $\mathcal{E}_t$ -predictable, where  $\mathcal{E}_t \subseteq \mathcal{F}_t$  is a given filtration representing the information available to the controller at time  $t$ . For example, we could have

$$\mathcal{E}_t = \mathcal{F}_{(t-\delta)^+}, \quad t \in [0, T],$$

where  $\delta > 0$  is a fixed *delay of information*.

The performance functional is assumed to have the form

$$(1.2) \quad J(x, u) = \mathbb{E} \left\{ H_1 x_T^2 + H_2 x_T \right\} + \mathbb{E} \left\{ \int_0^T [Q_{11}(t)x_t^2 + 2Q_{12}(t)x_t u_t + Q_{22}(t)u_t^2 + R_1(t)x_t + 2R_2(t)u_t] dt \right\},$$

where  $Q_{ij}(t)$  and  $R_i(t)$  are given bounded  $\mathcal{F}_t$ -adapted processes and  $H_i$  are given  $\mathcal{F}_T$ -measurable bounded random variables satisfying certain conditions (see section 2). Even in the absence of jumps, namely,

$$E_t(z) = F_t(z) = \gamma_t(z) = 0$$

(the diffusion case), the theory of classical linear quadratic control deals only with the case when

$$\mathcal{E}_t = \mathcal{F}_t \quad (\text{complete information case})$$

and

$$H_2 = 0, \quad \alpha_t = 0, \quad \beta_t = 0, \quad R_1(t) = R_2(t) = 0.$$

Namely, there are no first-order terms in the utility functional and there are no constant terms in the system. If the coefficients are random (but predictable) and/or  $\mathcal{E}_t \subset \mathcal{F}_t$ , then the system is no longer Markovian. The most effective method is the technique of completing squares.

However, even if  $\mathcal{E}_t = \mathcal{F}_t$ , the classical technique of completing squares is not directly applicable to the system we consider because of the appearance of the first-order terms in the utility functional and the constant terms in the controlled system. The appearance of such terms is important when we apply the results to minimum variance portfolio selection, for example.

In this paper we introduce an additional auxiliary backward Riccati equation to handle the extra terms. Thus we will have two (coupled) Riccati equations. Fortunately, they are only weakly coupled in the sense that we can solve one equation first and then substitute the solution into the other. This introduction of an additional equation, which handles the linear and constant terms, was given earlier in [17] for the constant term and in [13] for both linear and constant terms. There is a rich literature on stochastic linear quadratic control and associated Riccati equations; see, e.g., [1], [2], [5], [6], [10], [16], [17].

We will apply our results to minimum variance portfolio selection problems with or without partial information [3], [4], [8]. The results extend those in [9] (which use the Hamilton–Jacobi–Bellman dynamic programming principle) to the case of random coefficients.

It should be pointed out that the approach of the dynamic programming principle or the maximum principle cannot be applied directly here, both because of the general random coefficients in the controlled system and in the utility functional and because of partial information. Moreover, the technique of completing the square also leads us to the solution of the partial information problem.

**2. The complete information case.** Let us first consider the case with complete information, i.e.,  $\mathcal{E}_t = \mathcal{F}_t$ . Let the system be described by a one-dimensional stochastic differential equation, driven both by Brownian white noise and Poissonian random measure, as follows:

$$(2.1) \quad \begin{aligned} dx_t &= dx_t^{(u)} = [A_t x_t + B_t u_t + \alpha_t] dt + [C_t x_t + D_t u_t + \beta_t] dW_t \\ &\quad + \int_{\mathbb{R}} [E_t(z)x_{t-} + F_t(z)u_t + \gamma_t(z)] \tilde{N}(dt, dz), \quad 0 \leq t \leq T, \\ x_0 &= x \in \mathbb{R}. \end{aligned}$$

We assume that  $A_t, C_t, E_t(z), B_t, D_t, F_t(z), \alpha_t, \beta_t$ , and  $\gamma_t(z)$  are bounded  $\mathbb{R}$ -valued  $\mathcal{F}_t$ -predictable processes (they can be random). The goal is to minimize the cost functional

$$(2.2) \quad \begin{aligned} J(x, u) &= \mathbb{E} \{ H_1 x_T^2 + H_2 x_T \} \\ &\quad + \mathbb{E} \left\{ \int_0^T [Q_{11}(t)x_t^2 + 2Q_{12}(t)x_t u_t + Q_{22}(t)u_t^2 + R_1(t)x_t + 2R_2(t)u_t] dt \right\}, \end{aligned}$$

where  $Q_{ij}(t)$  and  $R_i(t)$ ,  $i, j = 1, 2$ , are given bounded  $\mathcal{F}_t$ -adapted (real-valued) stochastic processes and  $H_1$  and  $H_2$  are  $\mathcal{F}_T$ -measurable bounded random variables.

We assume throughout this paper that

$$Q_{22}(t) + \Theta_3(t) \geq 0 \quad \text{for a.a. } t, \omega,$$

where  $\Theta_3(t)$  is defined by (2.14). This is a linear system with a quadratic utility functional. We say that the control  $u_t$  is *admissible* and write  $u_t \in \mathcal{A}_{\mathcal{F}}$  if  $u_t$  is  $\mathcal{F}_t$ -predictable and equation (2.1) has a unique strong solution  $x_t = x_t^{(u)}$  for  $0 \leq t \leq T$  and

$$\mathbb{E} \left[ \int_0^T \left\{ u^2(t) + \left( x_t^{(u)} \right)^2 \right\} dt \right] < \infty.$$

We define

$$\begin{aligned} \rho_1(t) &= \int_{\mathbb{R}} E_t(z)^2 \nu(dz), & \rho_2(t) &= \int_{\mathbb{R}} \mu_t(z) [E_t(z)^2 + 2E_t(z)] \nu(dz), \\ \rho_3(t) &= \int_{\mathbb{R}} E_t(z)F_t(z) \nu(dz), & \rho_4(t) &= \int_{\mathbb{R}} \mu_t(z) [E_t(z)F_t(z) + 2F_t(z)] \nu(dz), \\ \rho_5(t) &= \int_{\mathbb{R}} F_t(z)^2 \nu(dz), & \rho_6(t) &= \int_{\mathbb{R}} \mu_t(z) F_t(z)^2 \nu(dz), \\ \rho_7(t) &= \int_{\mathbb{R}} \gamma_t(z)E_t(z) \nu(dz), & \rho_8(t) &= \int_{\mathbb{R}} \mu_t(z)\gamma_t(z) [1 + E_t(z)] \nu(dz), \\ \rho_9(t) &= \int_{\mathbb{R}} \gamma_t(z)F_t(z) \nu(dz), & \rho_{10}(t) &= \int_{\mathbb{R}} \gamma_t(z)\mu_t(z)F_t(z) \nu(dz), \end{aligned}$$

$$\begin{aligned}\rho_{11}(t) &= \int_{\mathbb{R}} \gamma_t(z)^2 \nu(dz), & \rho_{12}(t) &= \int_{\mathbb{R}} \gamma_t(z)^2 \mu_t(z) \nu(dz), \\ \rho_{13}(t) &= \int_{\mathbb{R}} \tilde{\mu}_t(z) E_t(z) \nu(dz), & \rho_{14}(t) &= \int_{\mathbb{R}} \tilde{\mu}_t(z) F_t(z) \nu(dz), \\ \rho_{15}(t) &= \int_{\mathbb{R}} \tilde{\mu}_t(z) \gamma_t(z) \nu(dz).\end{aligned}$$

We introduce the following system of backward Riccati/backward linear stochastic differential equations in the two unknown processes  $p_t$  and  $\tilde{p}_t$ :

(2.3)

$$\begin{aligned}dp_t &+ [2p_t A_t + p_t C_t^2 + 2\eta_t C_t + \rho_1(t)p_t + \rho_2(t) + Q_{11}(t)] dt \\ &- [Q_{22}(t) + p_t D_t^2 + \rho_5(t)p_t + \rho_6(t)]^{-1} [p_t B_t + p_t C_t D_t + \eta_t D_t \\ &\quad + \rho_3(t)p_t + \rho_4(t) + Q_{12}(t)]^2 dt \\ &- \eta_t dW_t - \int_{\mathbb{R}} \mu_t(z) \tilde{N}(dt, dz) = 0,\end{aligned}$$

(2.4)  $p_T = H_1$ ,

(2.5)

$$\begin{aligned}d\tilde{p}_t &+ [2p_t \alpha_t + 2\beta_t p_t C_t + 2\beta_t \eta_t + 2p_t \rho_7(t) + 2\rho_8(t)] dt \\ &+ [\tilde{p}_t A_t + C_t \tilde{\eta}_t + \rho_{13}(t) + R_1(t)] dt \\ &- 2 [Q_{22}(t) + p_t D_t^2 + \rho_5(t)p_t + \rho_6(t)]^{-1} [p_t B_t + p_t C_t D_t + \eta_t D_t + \rho_3(t)p_t + \rho_4(t) \\ &\quad + Q_{12}(t)], \\ &\left[ p_t \beta_t D_t + p_t \rho_9(t) + \rho_{10}(t) + \frac{1}{2} \tilde{p}_t B_t + \frac{1}{2} \tilde{\eta}_t D_t + \frac{1}{2} \rho_{14}(t) + R_2(t) \right] dt \\ &- \tilde{\eta}_t dW_t - \int_{\mathbb{R}} \tilde{\mu}_t(z) \tilde{N}(dt, dz) = 0,\end{aligned}$$

(2.6)  $\tilde{p}_T = H_2$ .

Here the  $\mathcal{F}_t$ -predictable, square integrable processes  $\xi_t$ ,  $\eta_t$ ,  $\mu_t(z)$  and  $\tilde{\xi}_t$ ,  $\tilde{\eta}_t$ ,  $\tilde{\mu}_t(z)$  are (implicitly) determined from  $p_t$  and  $\tilde{p}_t$ , respectively, through the semimartingale representations

$$(2.7) \quad dp_t = \xi_t dt + \eta_t dW_t + \int_{\mathbb{R}} \mu_t(z) \tilde{N}(dt, dz)$$

and

$$(2.8) \quad d\tilde{p}_t = \tilde{\xi}_t dt + \tilde{\eta}_t dW_t + \int_{\mathbb{R}} \tilde{\mu}_t(z) \tilde{N}(dt, dz).$$

We now state the first main theorem of this paper.

**THEOREM 2.1.** *Suppose the system of backward Riccati equations (2.3)–(2.6) has a solution  $p_t$  and  $\tilde{p}_t$ . Define*

$$\begin{aligned}(2.9) \quad u_t &= - [Q_{22}(t) + p_t D_t^2 + \rho_5(t)p_t + \rho_6(t)]^{-1}, \\ &\left\{ [p_t B_t + p_t C_t D_t + \eta_t D_t + \rho_3(t)p_t + \rho_4(t) + Q_{12}(t)] x_{t-} \right. \\ &\quad \left. + p_t \beta_t D_t + p_t \rho_9(t) + \rho_{10}(t) + \frac{1}{2} (\tilde{p}_t + \tilde{\eta}_t D_t + \rho_{14}(t)) + R_2(t) \right\}.\end{aligned}$$



Suppose  $u_t \in \mathcal{A}_{\mathcal{F}}$  and

$$(2.10) \quad \mathbb{E} \left[ \int_0^T \left\{ x_t^4 \eta_t^2 + (x_t^4 + u_t^4)(p_t^2 + \int_{\mathbb{R}} \mu_t^2(z) \nu(dz) \right\} dt \right] < \infty.$$

Then  $u_t$  is the unique solution of the complete information linear quadratic control problem (2.1)–(2.2). The corresponding value function is also quadratic and given by

$$(2.11) \quad \mathbb{E} (p_0) x^2 + \mathbb{E} (\tilde{p}_0) x \\ + \mathbb{E} \int_0^T \left\{ \Theta_6(t) + \Theta_9(t) - [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_5(t) + \Theta_8(t) + R_2(t)]^2 \right\} dt,$$

where  $p_t$  and  $\tilde{p}_t$  are found from solving the above backward equation and  $\Theta_i(t)$ ,  $i = 3, 5, 6, 8, 9$ , are defined by (2.12)–(2.20).

*Remark 2.2.* The existence of a solution to (2.3) has been proved recently by Hu and Song. See [11].

If all the parameters are deterministic, then we can take  $\eta_t$ ,  $\tilde{\eta}_t$ ,  $\mu_t(z)$ , and  $\tilde{\mu}_t(z)$  to be 0. In this case the stochastic Riccati equation reduces to the usual (deterministic) Riccati equation.

If at least one of the parameters is stochastic and all of them depend only on Brownian white noise  $W$ , then we may choose  $\mu_t(z)$  and  $\tilde{\mu}_t(z)$  to be 0, but  $\eta_t$  and  $\tilde{\eta}_t$  cannot both be 0. If at least one of them is stochastic and all of them depend only on Poisson noise  $N(\cdot, dz)$ , then we may choose  $\eta_t$  and  $\tilde{\eta}_t$  to be 0. But  $\mu_t(z)$  and  $\tilde{\mu}_t(z)$  cannot both be 0.

*Proof of Theorem 2.1.* We shall use the technique of completing squares.

Applying (2.7) and the integration by parts formula, we have

$$\begin{aligned} dx_t^2 &= 2x_{t-} dx_t + [C_t x_t + D_t u_t + \beta_t]^2 dt + \int_{\mathbb{R}} [E_t(z) x_{t-} + F_t(z) u_t + \gamma_t(z)]^2 N(dt, dz) \\ &= 2x_{t-} \left\{ [A_t x_t + B_t u_t + \alpha_t] dt + [C_t x_t + D_t u_t + \beta_t] dW_t \right. \\ &\quad \left. + \int_{\mathbb{R}} [E_t(z) x_{t-} + F_t(z) u_t + \gamma_t(z)] \tilde{N}(dt, dz) \right\} + [C_t x_t + D_t u_t + \beta_t]^2 dt \\ &\quad + \int_{\mathbb{R}} [E_t(z) x_{t-} + F_t(z) u_t + \gamma_t(z)]^2 \tilde{N}(dt, dz) \\ &\quad + \int_{\mathbb{R}} [E_t(z) x_t + F_t(z) u_t + \gamma_t(z)]^2 \nu(dz) dt. \end{aligned}$$

Another integration by parts yields

$$\begin{aligned} d(p_t x_t^2) &= 2p_{t-} x_{t-} \left\{ [A_t x_t + B_t u_t + \alpha_t] dt + [C_t x_t + D_t u_t + \beta_t] dW_t \right. \\ &\quad \left. + \int_{\mathbb{R}} [E_t(z) x_{t-} + F_t(z) u_t + \gamma_t(z)] \tilde{N}(dt, dz) \right\} \\ &\quad + p_t [C_t x_t + D_t u_t + \beta_t]^2 dt + \int_{\mathbb{R}} p_{t-} [E_t(z) x_{t-} + F_t(z) u_t + \gamma_t(z)]^2 N(dt, dz) \\ &\quad + x_{t-}^2 \left[ \xi_t dt + \eta_t dW_t + \int_{\mathbb{R}} \mu_t(z) \tilde{N}(dt, dz) \right] \end{aligned}$$

$$\begin{aligned}
& + 2\eta_t x_t [C_t x_t + D_t u_t + \beta_t] dt + \int_{\mathbb{R}} \mu_t(z) [E_t(z)x_{t-} + F_t(z)u_t + \gamma_t(z)]^2 N(dt, dz) \\
& + 2 \int_{\mathbb{R}} \mu_t(z)x_{t-} [E_t(z)x_{t-} + F_t(z)u_t + \gamma_t(z)] N(dt, dz).
\end{aligned}$$

Denote

$$\begin{aligned}
d\eta_1(t) &= x_{t-}^2 \left[ \eta_t dW_t + \int_{\mathbb{R}} \mu_t(z) \tilde{N}(dt, dz) \right] + 2p_t x_t [C_t x_t + D_t u_t + \beta_t] dW_t \\
&+ \int_{\mathbb{R}} \left\{ \mu_t(z)x_{t-}^2 + 2p_{t-}x_{t-} [E_t(z)x_{t-} + F_t(z)u_t + \gamma_t(z)] \right. \\
&\quad \left. + (p_{t-} + \mu_t(z)) [E_t(z)x_{t-} + F_t(z)u_t + \gamma_t(z)]^2 \right\} \tilde{N}(dt, dz) \\
&+ 2 \int_{\mathbb{R}} \mu_t(z)x_{t-} [E_t(z)x_{t-} + F_t(z)u_t + \gamma_t(z)] \tilde{N}(dt, dz)
\end{aligned}$$

and  $\eta_1(0) = 0$ . Then we see from (2.10) that  $\mathbb{E} \eta_1(t) = 0$  for all  $t \geq 0$ . We can rewrite

$$\begin{aligned}
d(p_t x_t^2) &= x_t^2 \xi_t dt + 2p_t x_t [A_t x_t + B_t u_t + \alpha_t] dt \\
&+ p_t [C_t x_t + D_t u_t + \beta_t]^2 dt + 2x_t \eta_t [C_t x_t + D_t u_t + \beta_t] dt \\
&+ \int_{\mathbb{R}} \left\{ [p_t + \mu_t(z)] [E_t(z)x_t + F_t(z)u_t + \gamma_t(z)]^2 \right\} \nu(dz) dt \\
&+ 2 \int_{\mathbb{R}} \mu_t(z)x_t [E_t(z)x_t + F_t(z)u_t + \gamma_t(z)] \nu(dz) dt + d\eta_1(t).
\end{aligned}$$

Introduce the notation

$$\begin{aligned}
\Theta_1(t) &= \xi_t + 2p_t A_t + p_t C_t^2 + 2\eta_t C_t \\
(2.12) \quad &+ \int_{\mathbb{R}} [p_t E_t(z)^2 + \mu_t(z) E_t(z)^2 + 2\mu_t(z) E_t(z)] \nu(dz);
\end{aligned}$$

$$\begin{aligned}
\Theta_2(t) &= p_t B_t + p_t C_t D_t + \eta_t D_t \\
(2.13) \quad &+ \int_{\mathbb{R}} \{p_t E_t(z) F_t(z) + \mu_t(z) E_t(z) F_t(z) + \mu_t(z) F_t(z)\} \nu(dz);
\end{aligned}$$

$$(2.14) \quad \Theta_3(t) = p_t D_t^2 + \int_{\mathbb{R}} \{p_t F_t(z)^2 + \mu_t(z) F_t(z)^2\} \nu(dz);$$

$$\begin{aligned}
\Theta_4(t) &= 2p_t \alpha_t + 2\beta_t p_t C_t + 2\beta_t \eta_t \\
(2.15) \quad &+ 2 \int_{\mathbb{R}} [(p_t + \mu_t(z)) \gamma_t(z) E_t(z) + \mu_t(z) \gamma_t(z)] \nu(dz);
\end{aligned}$$

$$(2.16) \quad \Theta_5(t) = p_t \beta_t D_t + \int_{\mathbb{R}} (p_t + \mu_t(z)) \gamma_t(z) F_t(z) \nu(dz);$$

and

$$(2.17) \quad \Theta_6(t) = p_t \beta_t^2 + \int_{\mathbb{R}} (p_t + \mu_t(z)) \gamma_t^2(z) \nu(dz).$$

Then we have

$$(2.18) \quad \mathbb{E} \{p_T x_T^2\} = \mathbb{E} \{p_0 x^2\} + \mathbb{E} \int_0^T \left\{ \Theta_1(t) x_t^2 + 2\Theta_2(t) x_t u_t + \Theta_3(t) u_t^2 + \Theta_4(t) x_t + 2\Theta_5(t) u_t + \Theta_6(t) \right\} dt.$$

To deal with the first-order terms which appeared above (2.18), we combine (2.8) with the integration by parts formula to get

$$\begin{aligned} d(\tilde{p}_t x_t) &= x_{t-} \left[ \tilde{\xi}_t dt + \tilde{\eta}_t dW_t + \int_{\mathbb{R}} \tilde{\mu}_t(z) \tilde{N}(dt, dz) \right] \\ &\quad + \tilde{p}_{t-} \left\{ [A_t x_t + B_t u_t + \alpha_t] dt + [C_t x_t + D_t u_t + \beta_t] dW_t \right. \\ &\quad \left. + \int_{\mathbb{R}} [E_t(z) x_{t-} + F_t(z) u_t + \gamma_t(z)] \tilde{N}(dt, dz) \right\} \\ &\quad + \tilde{\eta}_t [C_t x_t + D_t u_t + \beta_t] dt \\ &\quad + \int_{\mathbb{R}} \tilde{\mu}_t(z) [E_t(z) x_t + F_t(z) u_t + \gamma_t(z)] \nu(dz) dt \\ &\quad + \int_{\mathbb{R}} \tilde{\mu}_t(z) [E_t(z) x_{t-} + F_t(z) u_t + \gamma_t(z)] \tilde{N}(dt, dz). \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E} [\tilde{p}_T x_T] &= \mathbb{E} \left[ \tilde{p}_0 x + \int_0^T \left\{ x_t \tilde{\xi}_t + \tilde{p}_t [A_t x_t + B_t u_t + \alpha_t] \right. \right. \\ &\quad \left. \left. + \tilde{\eta}_t [C_t x_t + D_t u_t + \beta_t] + \int_{\mathbb{R}} \tilde{\mu}_t(z) [E_t(z) x_t + F_t(z) u_t + \gamma_t(z)] \nu(dz) \right\} dt \right] \\ (2.19) \quad &= \mathbb{E} \left[ \tilde{p}_0 x + \int_0^T \{ \Theta_7(t) x_t + 2\Theta_8(t) u_t + \Theta_9(t) \} dt \right], \end{aligned}$$

where

$$(2.20) \quad \Theta_7(t) = \tilde{\xi}_t + \tilde{p}_t A_t + C_t \tilde{\eta}_t + \int_{\mathbb{R}} \tilde{\mu}_t(z) E_t(z) \nu(dz),$$

$$(2.21) \quad \Theta_8(t) = \frac{1}{2} \left\{ \tilde{p}_t B_t + \tilde{\eta}_t D_t + \int_{\mathbb{R}} \tilde{\mu}_t(z) F_t(z) \nu(dz) \right\},$$

$$(2.22) \quad \Theta_9(t) = \tilde{p}_t \alpha_t + \tilde{\eta}_t \beta_t + \int_{\mathbb{R}} \tilde{\mu}_t(z) \gamma_t(z) \nu(dz).$$

Let

$$p_T = H_1 \quad \text{and} \quad \tilde{p}_T = H_2.$$

Therefore

$$\begin{aligned}
 & J(x, u) \\
 &= \left\{ \int_0^T \left[ Q_{11}(t)x_t^2 + 2Q_{12}(t)x_t u_t + Q_{22}(t)u_t^2 + R_1(t)x_t + 2R_2(t)u_t \right] dt + p_T x_T^2 + \tilde{p}_T x_T \right\} \\
 &= \mathbb{E} (p_0 x^2) + \mathbb{E} (\tilde{p}_0 x) + \mathbb{E} \int_0^T \left\{ [\Theta_1(t) + Q_{11}(t)] x_t^2 + 2 [\Theta_2(t) + Q_{12}(t)] x_t u_t \right. \\
 &\quad \left. + [Q_{22}(t) + \Theta_3(t)] u_t^2 + [\Theta_4(t) + \Theta_7(t) + R_1(t)] x_t \right. \\
 &\quad \left. + 2 [\Theta_5(t) + \Theta_8(t) + R_2(t)] u_t + \Theta_6(t) + \Theta_9(t) \right\} dt \\
 &= \mathbb{E} (p_0 x^2) + \mathbb{E} (\tilde{p}_0 x) \\
 &\quad + \mathbb{E} \int_0^T \left\{ \left[ \Theta_1(t) + Q_{11}(t) - [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_2(t) + Q_{12}(t)]^2 \right] x_t^2 \right. \\
 &\quad \left. + \left[ \Theta_4(t) + \Theta_7(t) + R_1(t) - 2 [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_2(t) + Q_{12}(t)] \right. \right. \\
 &\quad \left. \left. \times [\Theta_5(t) + \Theta_8(t) + R_2(t)] \right] x_t \right. \\
 &\quad \left. + [Q_{22}(t) + \Theta_3(t)] \left\{ u_t + [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_2(t) + Q_{12}(t)] x_t \right. \right. \\
 &\quad \left. \left. + [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_5(t) + \Theta_8(t) + R_2(t)] \right\}^2 \right. \\
 &\quad \left. \Theta_6(t) + \Theta_9(t) - [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_5(t) + \Theta_8(t) + R_2(t)]^2 \right\} dt.
 \end{aligned}$$

If

$$\begin{aligned}
 & (2.23) \\
 & \left\{ \begin{aligned} & \Theta_1(t) + Q_{11}(t) - [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_2(t) + Q_{12}(t)]^2 = 0, \\ & \Theta_4(t) + \Theta_7(t) + R_1(t) - 2 [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_2(t) + Q_{12}(t)] [\Theta_5(t) + \Theta_8(t) + R_2(t)] \\ & = 0, \end{aligned} \right.
 \end{aligned}$$

then

$$\begin{aligned}
 & J(x, u) = \mathbb{E} (p_0 x^2) + \mathbb{E} (\tilde{p}_0 x) + \mathbb{E} \int_0^T J_0(t) dt \\
 & \quad + \mathbb{E} \int_0^T [Q_{22}(t) + \Theta_3(t)] \left\{ u_t + [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_2(t) + Q_{12}(t)] x_t - \right. \\
 & \quad \left. + [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_5(t) + \Theta_8(t) + R_2(t)] \right\}^2 dt, \tag{2.24}
 \end{aligned}$$

where

$$J_0(t) = \Theta_6(t) + \Theta_9(t) - [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_5(t) + \Theta_8(t) + R_2(t)]^2$$

is independent of  $u_t$  and  $x_t$ . This utility functional will achieve its minimum

$$\mathbb{E} (p_0 x^2) + \mathbb{E} (\tilde{p}_0 x) + \mathbb{E} \int_0^T J_0(t) dt$$

when

$$(2.25) \quad u_t = -[Q_{22}(t) + \Theta_3(t)]^{-1} \{[\Theta_2(t) + Q_{12}(t)]x_{t-} + \Theta_5(t) + \Theta_8(t) + R_2(t)\}.$$

Thus the optimal control is also a feedback one which is linear and depends only on the state  $x_t$ .

Using the notation of  $\rho_i(t)$ , we may rewrite (2.12)–(2.17) and (2.20)–(2.22) as

$$(2.26) \quad \Theta_1(t) = \xi_t + 2p_t A_t + p_t C_t^2 + 2\eta_t C_t + \rho_1(t)p_t + \rho_2(t),$$

$$(2.27) \quad \Theta_2(t) = p_t B_t + p_t C_t D_t + \eta_t D_t + \rho_3(t)p_t + \rho_4(t),$$

$$(2.28) \quad \Theta_3(t) = p_t D_t^2 + \rho_5(t)p_t + \rho_6(t),$$

$$(2.29) \quad \Theta_4(t) = 2p_t \alpha_t + 2\beta_t p_t C_t + 2\beta_t \eta_t + 2p_t \rho_7(t) + 2\rho_8(t),$$

$$(2.30) \quad \Theta_5(t) = p_t \beta_t D_t + p_t \rho_9(t) + \rho_{10}(t),$$

$$(2.31) \quad \Theta_6(t) = p_t \beta_t^2 + p_t \rho_{11}(t) + \rho_{12}(t),$$

$$(2.32) \quad \Theta_7(t) = \tilde{\xi}_t + \tilde{p}_t A_t + C_t \tilde{\eta}_t + \rho_{13}(t),$$

$$(2.33) \quad \Theta_8(t) = \frac{1}{2} \{ \tilde{p}_t B_t + \tilde{\eta}_t D_t + \rho_{14}(t) \},$$

$$(2.34) \quad \Theta_9(t) = \tilde{p}_t \alpha_t + \tilde{\eta}_t \beta_t + \rho_{15}(t).$$

The first equation of (2.23) becomes

$$\begin{aligned} & \xi_t + 2p_t A_t + p_t C_t^2 + 2\eta_t C_t + \rho_1(t)p_t + \rho_2(t) + Q_{11}(t) \\ & + [Q_{22}(t) + p_t D_t^2 + \rho_5(t)p_t + \rho_6(t)]^{-1} \\ & \times [p_t B_t + p_t C_t D_t + \eta_t D_t + \rho_3(t)p_t + \rho_4(t) + Q_{12}(t)]^2 = 0. \end{aligned}$$

Multiplying by  $dt$ , we get

$$\begin{aligned} & \xi_t dt + [2p_t A_t + p_t C_t^2 + 2\eta_t C_t + \rho_1(t)p_t + \rho_2(t) + Q_{11}(t)] dt \\ & - [Q_{22}(t) + p_t D_t^2 + \rho_5(t)p_t + \rho_6(t)]^{-1} \\ & \times [p_t B_t + p_t C_t D_t + \eta_t D_t + \rho_3(t)p_t + \rho_4(t) + Q_{12}(t)]^2 dt = 0. \end{aligned}$$

Substituting

$$\xi_t dt = dp_t - \eta_t dW_t - \int_{\mathbb{R}} \mu_t(z) \tilde{N}(dt, dz)$$

into the equation, we have the following backward Riccati equation for  $p_t$ :

$$\begin{aligned} & dp_t + [2p_t A_t + p_t C_t^2 + 2\eta_t C_t + \rho_1(t)p_t + \rho_2(t) + Q_{11}(t)] dt \\ & - [Q_{22}(t) + p_t D_t^2 + \rho_5(t)p_t + \rho_6(t)]^{-1} [p_t B_t + p_t C_t D_t + \eta_t D_t + \rho_3(t)p_t + \rho_4(t) \\ & \quad + Q_{12}(t)]^2 dt \\ & - \eta_t dW_t - \int_{\mathbb{R}} \mu_t(z) \tilde{N}(dt, dz) = 0. \end{aligned}$$

In a similar way, we can reduce the second equation of (2.23) to

$$\begin{aligned} & d\tilde{p}_t + [2p_t \alpha_t + 2\beta_t p_t C_t + 2\beta_t \eta_t + 2p_t \rho_7(t) + 2\rho_8(t)] dt \\ & + [\tilde{p}_t A_t + C_t \tilde{\eta}_t + \rho_{13}(t) + R_1(t)] dt \end{aligned}$$

$$\begin{aligned}
& -2 \left[ Q_{22}(t) + p_t D_t^2 + \rho_5(t) p_t + \rho_6(t) \right]^{-1} \left[ p_t B_t + p_t C_t D_t + \eta_t D_t + \rho_3(t) p_t + \rho_4(t) + Q_{12}(t) \right], \\
& \left[ p_t \beta_t D_t + p_t \rho_9(t) + \rho_{10}(t) + \frac{1}{2} \tilde{p}_t B_t + \frac{1}{2} \tilde{\eta}_t D_t + \frac{1}{2} \rho_{14}(t) + R_2(t) \right] dt \\
& - \tilde{\eta}_t dW_t - \int_{\mathbb{R}} \tilde{\mu}_t(z) \tilde{N}(dt, dz) = 0.
\end{aligned}$$

**3. The partial information case.** We now study the case when our control  $u_t$  is required to be  $\mathcal{E}_t$ -predictable, where

$$\mathcal{E}_t \subseteq \mathcal{F}_t \quad \text{for all } t \in [0, T]$$

is a given subfiltration representing the information available to the controller at time  $t$ . The corresponding family of admissible controls is denoted by  $\mathcal{A}_{\mathcal{E}}$ .

**THEOREM 3.1** (partial information linear quadratic control). *Suppose the system of Riccati equations (2.3)–(2.6) has a solution  $p_t$  and  $\tilde{p}_t$ . Define*

$$\begin{aligned}
(3.1) \quad u_t^* &= - \left( \mathbb{E} \left[ \{Q_{22}(t) + \Theta_3(t)\} \middle| \mathcal{E}_t \right] \right)^{-1}, \\
& \mathbb{E} \left[ \{(\Theta_2(t) + Q_{12}(t))x_{t-} + \Theta_5(t) + \Theta_8(t) + R_2(t)\} \middle| \mathcal{E}_t \right],
\end{aligned}$$

where  $\Theta_i(t)$  are given by (2.26)–(2.34).

Suppose  $u_t^* \in \mathcal{A}_{\mathcal{E}}$  and that (2.10) holds. Then  $u_t^*$  is the unique solution of the partial information linear quadratic control problem. The value function  $J_{\mathcal{E}}(x)$  in the partial observation case is given by

$$(3.2) \quad J_{\mathcal{E}}(x) = J_{\mathcal{F}}(x) + \mathbb{E} \left[ \int_0^T \left\{ L_t M_t^2 - \mathbb{E} [L_t | \mathcal{E}_t]^{-1} (\mathbb{E} [L_t M_t | \mathcal{E}_t])^2 \right\} dt \right],$$

where  $J_{\mathcal{F}}$  is the value function in the complete information case and

$$(3.3) \quad L_t = Q_{22}(t) + \Theta_3(t)$$

and

$$(3.4) \quad M_t = L_t^{-1} [(\Theta_2(t) + Q_{12}(t))x_t + \Theta_5(t) + \Theta_8(t) + R_2(t)].$$

*Proof.* We use the computation in the proof of Theorem 2.1. By (2.24), we have

$$(3.5) \quad J(x, u) = J_{\mathcal{F}}(x) + \mathbb{E} \left[ \int_0^T L_t (u_t + M_t)^2 dt \right].$$

Note that  $L_t$  does not depend on  $X_t$  (or  $u_t$ ). For each  $t$ , define the measure  $Q_t$  by

$$(3.6) \quad dQ_t = L_t dP_t \quad \text{on } \mathcal{F}_t.$$

Then

$$\mathbb{E} \left[ \int_0^T L_t (u_t + M_t)^2 dt \right] = \int_0^T \mathbb{E}_{Q_t} [(u_t + M_t)^2] dt.$$

We can minimize this for each  $t$ . By the well-known Kallianpur–Striebel formula [12] we know that the minimum of  $\mathbb{E}_{Q_t} [(u_t + M_t)^2]$  over all  $\mathcal{E}_t$ -measurable  $u_t$  is attained at

$$\begin{aligned} u_t = u_t^* &= -\mathbb{E}_{Q_t} [M_t | \mathcal{E}_t] \\ &= -\frac{\mathbb{E} [L_t M_t | \mathcal{E}_t]}{\mathbb{E} [L_t | \mathcal{E}_t]} \\ (3.7) \quad &= -\frac{\mathbb{E} [\{(\Theta_2(t) + Q_{12}(t))x_{t-} + \Theta_5(t) + \Theta_8(t) + R_2(t)\} | \mathcal{E}_t]}{\mathbb{E} [\{Q_{22}(t) + \Theta_3(t)\} | \mathcal{E}_t]}. \end{aligned}$$

This proves (3.1). Substituting (3.7) into (3.6), we get

$$\begin{aligned} J_{\mathcal{E}}(x) &= J_{\mathcal{F}}(x) + \mathbb{E} \left[ \int_0^T L_t (u_t^* + M_t)^2 dt \right] \\ &= J_{\mathcal{F}}(x) + \mathbb{E} \left[ \int_0^T \left\{ L_t M_t^2 - (\mathbb{E} [L_t | \mathcal{E}_t])^{-1} (\mathbb{E} [L_t M_t | \mathcal{E}_t])^2 \right\} dt \right], \end{aligned}$$

which proves (3.2).  $\square$

*Remark 3.2.* We may regard the term

$$J_{\mathcal{E}}(x) - J_{\mathcal{F}}(x) = \mathbb{E} \left[ \int_0^T \left\{ L_t M_t^2 - (\mathbb{E} [L_t | \mathcal{E}_t])^{-1} (\mathbb{E} [L_t M_t | \mathcal{E}_t])^2 \right\} dt \right]$$

as the reduction of performance (or cost increase) due to the reduced information flow  $\mathcal{E}_t$ .

#### 4. Some particular cases.

**4.1. Absence of Poissonian noise.** Let us first consider the case when the system is under the influence of Brownian white noise. In the controlled system (2.1), we let

$$E_t(z) = F_t(z) = \gamma_t = 0,$$

let all the coefficients be adapted with respect to the filtration  $\mathcal{F}_t^W = \sigma(W_s, s \leq t)$ , and let  $H_1, H_2$  be  $\mathcal{F}_T^W$  measurable. Then

$$\rho_i(t) = 0 \quad \text{for all } 1 \leq i \leq 15.$$

We may assume  $\mu_t = \tilde{\mu}_t = 0$  and write (2.3)–(2.6) as

(4.1)

$$\begin{aligned} dp_t &+ [2p_t A_t + p_t C_t^2 + 2\eta_t C_t + Q_{11}(t)] dt \\ &- [Q_{22}(t) + p_t D_t^2]^{-1} [p_t B_t + p_t C_t D_t + \eta_t D_t + Q_{12}(t)]^2 dt - \eta_t dW_t = 0, \end{aligned}$$

(4.2)  $p_T = H_1,$

(4.3)

$$\begin{aligned} d\tilde{p}_t &+ [2\tilde{p}_t \alpha_t + 2\beta_t p_t C_t + 2\beta_t \eta_t] dt + [\tilde{p}_t A_t + C_t \tilde{\eta}_t + R_1(t)] dt \\ &- 2 [Q_{22}(t) + p_t D_t^2]^{-1} [p_t B_t + p_t C_t D_t + \eta_t D_t + Q_{12}(t)], \\ &\left[ p_t \beta_t D_t + \frac{1}{2} \tilde{p}_t B_t + \frac{1}{2} \tilde{\eta}_t D_t + R_2(t) \right] dt - \tilde{\eta}_t dW_t = 0, \end{aligned}$$

(4.4)  $\tilde{p}_T = H_2.$

THEOREM 4.1. Suppose the system of backward Riccati equations (4.1)–(4.4) has a solution  $p_t$  and  $\tilde{p}_t$ . Define

$$(4.5) \quad u_t = -[Q_{22}(t) + p_t D_t^2]^{-1} \left\{ [p_t B_t + p_t C_t D_t + \eta_t D_t + Q_{12}(t)] x_{t-} - p_t \beta_t D_t + \frac{1}{2} (\tilde{p}_t + \tilde{\eta}_t D_t - R_2(t)) \right\}.$$

Suppose  $u_t \in \mathcal{A}_{\mathcal{F}}$  and that (2.10) holds. Then  $u_t$  is the unique solution of the complete information linear quadratic control problem (2.1)–(2.2). The corresponding value function is also quadratic and is given by

$$\mathbb{E} (p_0) x^2 + \mathbb{E} (\tilde{p}_0) x + \mathbb{E} \int_0^T \left\{ \Theta_6(t) + \Theta_9(t) - [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_5(t) + \Theta_8(t) + R_2(t)]^2 \right\} dt,$$

where  $p_t$  and  $\tilde{p}_t$  are found from solving the above backward equations and

$$\begin{aligned} \Theta_3(t) &= p_t D_t^2, & \Theta_5(t) &= p_t \beta_t D_t, & \Theta_6(t) &= p_t \beta_t^2, \\ \Theta_8(t) &= \frac{1}{2} (\tilde{p}_t + \tilde{\eta}_t D_t), & \Theta_9(t) &= \tilde{p}_t \alpha_t + \tilde{\eta}_t \beta_t. \end{aligned}$$

**4.2. Absence of Brownian white noise.** If, in the controlled system (2.1),  $C_t = D_t = \beta_t = 0$  and all the coefficients are adapted to the filtration  $\mathcal{F}_t^P = \sigma(N(ds, dz), s \leq t)$ , and  $H_1, H_2$  are  $\mathcal{F}_T^P$  measurable, then we may consider the system

$$(4.6) \quad \begin{aligned} dp_t &+ [2p_t A_t + \rho_1(t) p_t + \rho_2(t) + Q_{11}(t)] dt \\ &- [Q_{22}(t) + \rho_5(t) p_t + \rho_6(t)]^{-1} [p_t B_t + \rho_3(t) p_t + \rho_4(t)]^2 dt - \int_{\mathbb{R}} \mu_t(z) \tilde{N}(dt, dz) = 0, \end{aligned}$$

$$(4.7) \quad p_T = H_1,$$

$$(4.8) \quad \begin{aligned} d\tilde{p}_t &+ [2\tilde{p}_t \alpha_t + 2\beta_t \eta_t + 2p_t \rho_7(t) + 2\rho_8(t)] dt + [\tilde{p}_t A_t + \rho_{13}(t) + R_1(t)] dt \\ &- 2[Q_{22}(t) + \rho_5(t) p_t + \rho_6(t)]^{-1} [p_t B_t + \rho_3(t) p_t + \rho_4(t) + Q_{12}(t)], \\ &\left[ p_t \rho_9(t) + \rho_{10}(t) + \frac{1}{2} \tilde{p}_t B_t + \frac{1}{2} \rho_{14}(t) + R_2(t) \right] dt - \int_{\mathbb{R}} \tilde{\mu}_t(z) \tilde{N}(dt, dz) = 0, \end{aligned}$$

$$(4.9) \quad \tilde{p}_T = H_2.$$

THEOREM 4.2. Suppose the system of backward Riccati equations (2.3)–(2.6) has a solution  $p_t$  and  $\tilde{p}_t$ . Define

$$(4.10) \quad u_t = -[Q_{22}(t) + \rho_5 p_t + \rho_6(t)]^{-1} \left\{ [p_t B_t + \rho_3(t) p_t + \rho_4(t) + Q_{12}(t)] x_{t-} - p_t \rho_9(t) + \rho_{10}(t) + \frac{1}{2} (\tilde{p}_t + \rho_{14}(t)) - R_2(t) \right\}.$$



Suppose  $u_t \in \mathcal{A}_{\mathcal{F}}$  and that (2.10) holds. Then  $u_t$  is the unique solution of the complete information linear quadratic control problem (2.1)–(2.2). The corresponding value function is also quadratic and is given by

$$\mathbb{E} (p_0)x^2 + \mathbb{E} (\tilde{p}_0)x + \mathbb{E} \int_0^T \left\{ \Theta_6(t) + \Theta_9(t) - [Q_{22}(t) + \Theta_3(t)]^{-1} [\Theta_5(t) + \Theta_8(t) + R_2(t)]^2 \right\} dt,$$

where  $p_t$  and  $\tilde{p}_t$  are found from solving the above backward equations and  $\Theta_i$  are given by corresponding formulas of (2.26)–(2.34).

**4.3. Classical Riccati equations.** To obtain the classical Riccati equation, we may assume that in the controlled system (2.1),

$$\alpha_t = 0, \quad \beta_t = 0, \quad \gamma_t = 0, \quad H_2 = 0, \quad Q_{12}(t) = R_1(t) = R_2(t) = 0.$$

In this case, we have

$$\rho_7(t) = \rho_9(t) = \rho_{10}(t) = \rho_{11}(t) = \rho_{12}(t) = 0.$$

The backward stochastic Riccati equation for  $\tilde{p}_t$  becomes

$$\begin{aligned} & d\tilde{p}_t + [\tilde{p}_t A_t + C_t \tilde{\eta}_t + \rho_{13}(t)] dt - 2 [Q_{22}(t) + p_t D_t^2 + \rho_5(t) p_t + \rho_6(t)]^{-1}, \\ & [p_t B_t + p_t C_t D_t + \eta_t D_t + \rho_3(t) p_t + \rho_4(t)], \\ & \left[ \frac{1}{2} \tilde{p}_t B_t + \frac{1}{2} \tilde{\eta}_t D_t + \frac{1}{2} \rho_{14}(t) \right] dt - \tilde{\eta}_t dW_t - \int_{\mathbb{R}} \tilde{\mu}_t(z) \tilde{N}(dt, dz) = 0, \\ & \tilde{p}_T = 0. \end{aligned}$$

Apparently, this equation has a solution 0. Moreover, (2.3) becomes

$$\begin{aligned} (4.11) \quad & dp_t + [2p_t A_t + p_t C_t^2 + 2\eta_t C_t + \rho_1(t) p_t + \rho_2(t) + Q_{11}(t)] dt \\ & - [Q_{22}(t) + p_t D_t^2 + \rho_5(t) p_t + \rho_6(t)]^{-1} [p_t B_t + p_t C_t D_t + \eta_t D_t + \rho_3(t) p_t + \rho_4(t) \\ & \quad + Q_{12}(t)]^2 dt \\ & - \eta_t dW_t - \int_{\mathbb{R}} \mu_t(z) \tilde{N}(dt, dz) = 0, \end{aligned}$$

$$(4.12) \quad p_T = H_1.$$

**THEOREM 4.3.** Suppose the system of backward Riccati equations (4.11)–(4.12) has a solution  $p_t$ . Define

$$\begin{aligned} (4.13) \quad u_t = & - [Q_{22}(t) + p_t D_t^2 + \rho_5 p_t + \rho_6(t)]^{-1}, \\ & \left\{ [p_t B_t + p_t C_t D_t + \eta_t D_t + \rho_3(t) p_t + \rho_4(t) + Q_{12}(t)] x_t - \right. \\ & \left. - p_t \beta_t D_t + p_t \rho_9(t) + \rho_{10}(t) - R_2(t) \right\}. \end{aligned}$$

Suppose that  $u_t \in \mathcal{A}_{\mathcal{F}}$  and that (2.10) holds. Then  $u_t$  is the unique solution of the complete information linear quadratic control problem (2.1)–(2.2). The corresponding value function is also quadratic and is given as before.

If in Theorem 4.3 we further assume

$$E_t(z) = F_t(z) = \gamma_t = 0, \quad Q_{12}(t) = R_1(t) = R_2(t) = 0,$$

then we have the following.

**COROLLARY 4.1.** *Suppose the backward Riccati equation*

$$\begin{aligned} dp_t + [2p_t A_t + p_t C_t^2 + 2\eta_t C_t + Q_{11}(t)] dt \\ - [Q_{22}(t) + p_t D_t^2]^{-1} [p_t B_t + p_t C_t D_t + \eta_t D_t]^2 dt - \eta_t dW_t = 0, \\ p_T = H_1 \end{aligned}$$

has a solution  $p_t$ . Define

$$u_t = -[Q_{22}(t) + p_t D_t^2]^{-1} \{ [p_t B_t + p_t C_t D_t + \eta_t D_t] x_{t-} - p_t \beta_t D_t + p_t \rho_9(t) + \rho_{10}(t) - R_2(t) \}.$$

Suppose  $u_t \in \mathcal{A}_{\mathcal{F}}$  and that (2.10) holds. Then  $u_t$  is the unique solution of the complete information linear quadratic control problem (2.1)–(2.2). The corresponding value function is also quadratic and is given as before.

**Remark 4.4.** This equation coincides with the equation in [13], for example.

**4.4. Deterministic linear quadratic problem.** If all the data are deterministic, then we may assume  $p_t$  and  $\tilde{p}_t$  to be deterministic, too. Hence

$$\eta_t = \mu_t(z) = \tilde{\eta}_t = \tilde{\mu}_t(z) = 0$$

and we have the following.

**THEOREM 4.5.** *Consider the following system of backward Riccati/backward linear stochastic differential equations:*

$$\begin{aligned} (4.14) \\ dp_t + [2p_t A_t + p_t C_t^2 + \rho_1(t)p_t + \rho_2(t) + Q_{11}(t)] dt \\ - [Q_{22}(t) + p_t D_t^2 + \rho_5(t)p_t + \rho_6(t)]^{-1} [p_t B_t + p_t C_t D_t + \rho_3(t)p_t + \rho_4(t) + Q_{12}(t)]^2 dt = 0, \end{aligned}$$

$$\begin{aligned} (4.15) \\ d\tilde{p}_t + [2p_t \alpha_t + 2\beta_t p_t C_t + 2p_t \rho_7(t) + \rho_8(t)] dt + [\tilde{p}_t A_t + \rho_{13}(t) + R_1(t)] dt \\ - 2 [Q_{22}(t) + p_t D_t^2 + \rho_5(t)p_t + \rho_6(t)]^{-1} [p_t B_t + p_t C_t D_t + \rho_3(t)p_t + \rho_4(t) + Q_{12}(t)], \\ \left[ p_t \beta_t D_t + p_t \rho_9(t) + \rho_{10}(t) + \frac{1}{2} \tilde{p}_t B_t + \frac{1}{2} \rho_{14}(t) + R_2(t) \right] dt = 0. \end{aligned}$$

The terminal conditions are

$$p_T = H_1 \quad \text{and} \quad \tilde{p}_T = H_2.$$

If the Riccati system (4.14)–(4.16) has solutions  $p_t$  and  $\tilde{p}_t$ , then the linear quadratic control problem (2.1)–(2.2) has a solution with the optimal control given by

(4.16)

$$u_t = - \left[ Q_{22}(t) + p_t D_t^2 + \rho_5 p_t + \rho_6(t) \right]^{-1},$$

$$\left\{ \begin{aligned} & [p_t B_t + p_t C_t D_t + \rho_3(t) p_t + \rho_4(t) + Q_{12}(t)] x_{t-} \\ & - \left[ p_t \beta_t D_t + p_t \rho_9(t) + \rho_{10}(t) + \frac{1}{2} (\tilde{p}_t + \tilde{\eta}_t D_t + \rho_{14}(t)) \right] - R_2(t) \end{aligned} \right\},$$

provided that  $u_t \in \mathcal{A}_{\mathcal{F}}$  and that (2.10) holds. The corresponding value function is given by (2.11) with

$$\begin{aligned} Q_6(t) &= p_t \beta_t^2 \int_{\mathbb{R}} p_t \gamma_t^2(z) \nu(dz), \\ Q_9(t) &= \tilde{p}_t \alpha_t, \\ Q_3(t) &= p_t D_t^2 + p_t^2 \int_{\mathbb{R}} F_t^2(z) \nu(dz), \\ Q_5(t) &= p_t \beta_t D_t + \int_{\mathbb{R}} p_t \gamma_t(z) F_t(z) \nu(dz), \\ Q_8(t) &= \frac{1}{2} \tilde{p}_t B_t. \end{aligned}$$

**5. Partial information mean-variance portfolio problem.** We now apply our results to a partial information mean-variance portfolio problem in finance.

Suppose we have a market with the following two investment possibilities:

(i) a risk free asset, whose unit price  $S_0(t)$  at time  $t$  is given by

$$(5.1) \quad dS_0(t) = \rho_t S_0(t) dt, \quad S_0(0) = 1, \quad 0 \leq t \leq T;$$

(ii) a risky asset, whose unit price  $S_1(t)$  at time  $t$  is given by

$$(5.2) \quad dS_1(t) = S_1(t-) \left[ a_t dt + b_t dW_t + \int_{\mathbb{R}} c_t(z) \tilde{N}(dt, dz) \right], \quad 0 \leq t \leq T,$$

$$S_1(0) > 0.$$

Here  $\rho_t$ ,  $a_t$ ,  $b_t$ , and  $c_t(z)$  are given  $\mathcal{F}_t$ -predictable processes. We assume that

$$(5.3) \quad \mathbb{E} \left[ \int_0^T \left\{ |\rho_t| + |a_t| + b_t^2 + \int_{\mathbb{R}} c_t(z)^2 \nu(dz) \right\} dt \right] < \infty;$$

$$(5.4) \quad \text{there exists } \varepsilon > 0 \text{ such that } c_t(z) > -1 + \varepsilon \quad \text{a.s. for a.a. } t, z;$$

$$(5.5) \quad p_t b_t^2 + \int_{\mathbb{R}} (p_t + \mu_t(z)) c_t(z)^2 \nu(dz) > 0 \quad \text{for a.a. } t, \omega,$$

where  $p_t$  is the solution of (5.14)–(5.15) following.

Conditions (5.3)–(5.4) ensure that the solution to (5.2) is given by

$$(5.6) \quad S_1(t) = S_1(0) \exp \left\{ \int_0^t \left( a_s - \frac{1}{2} b_s^2 \right) ds + \int_0^t b_s dW_s \right. \\ \left. + \int_0^t \int_{\mathbb{R}} \{ \log(1 + c_s(z)) - c_s(z) \} \nu(dz) ds + \int_0^t \int_{\mathbb{R}} \log(1 + c_s(z)) \tilde{N}(ds, dz) \right\}.$$

A portfolio in this market is a predictable process  $\phi(t) = (\phi_0(t), \phi_1(t)) \in \mathbb{R}^2$  giving the number of units of the risk-free and the risky asset, respectively, held at time  $t$ . The corresponding *wealth process*  $x(t) = x^\phi(t)$  is defined by

$$(5.7) \quad x^\phi(t) = \phi_0(t)S_0(t) + \phi_1(t)S_1(t).$$

We say that  $\phi(t)$  is *self-financing* if

$$(5.8) \quad dx^\phi(t) = \phi_0(t)dS_0(t) + \phi_1(t)dS_1(t).$$

Suppose we are given a subfiltration

$$\mathcal{E}_t \subseteq \mathcal{F}_t, \quad t \in [0, T].$$

Let

$$u_t = \phi_1(t)S_1(t)$$

be the *amount* (instead of number of shares) invested in the risky asset at time  $t$ . We say that  $u_t$  is *admissible* and write  $u_t \in \mathcal{A}_{\mathcal{E}}$  if  $u_t$  is  $\mathcal{E}_t$ -predictable,  $\phi_1(t) = \frac{u_t}{S_1(t)}$  is self-financing, and  $x^{(u)}(t) := x^{(\phi)}(t)$  is *lower bounded*. Combining the above, we see that if  $\phi \in \mathcal{A}_{\mathcal{E}}$ , then

$$(5.9) \quad dx^{(u)}(t) = \left\{ \rho_t x^{(u)}(t) + (a_t - \rho_t)u_t \right\} dt + b_t u_t dW(t) + u_t \int_{\mathbb{R}} c_t(z) \tilde{N}(dt, dz),$$

$$x^{(u)}(0) = x > 0.$$

We now consider the *partial information mean-variance portfolio problem*, which is to find the portfolio  $\hat{u} \in \mathcal{A}_{\mathcal{E}}$  which minimizes the variance

$$(5.10) \quad \mathbb{E} \left[ x^\phi(T) - \mathbb{E} x^\phi(T) \right]^2$$

under the constraint

$$(5.11) \quad \mathbb{E} \left[ x^\phi(T) \right] = K,$$

where  $K$  is a given constant.

Using the Lagrange multiplier method, we see that the problem is equivalent to minimizing

$$(5.12) \quad \mathbb{E} \left[ x^\phi(T) - \lambda \right]^2$$

for a given constant  $\lambda \in \mathbb{R}$ , without constraints. We refer the reader to [18], [14], and [8] for more information about the mean-variance portfolio problem.

For the case when  $\mathcal{E}_t = \mathcal{F}_t$  and the coefficients are all deterministic, this problem was solved in [8] by using the maximum principle for jump-diffusions.

Subsequently this was extended to the partial information case  $\mathcal{E}_t \subseteq \mathcal{F}_t$  (but still with deterministic coefficients) by Bagheri and Øksendal [3].

We now show how Theorem 3.1 gives us a solution also in the case of stochastic coefficients.

Here

$$A_t = \rho_t, \quad B_t = a_t - \rho_t, \quad \alpha_t = 0, \quad C_t = 0, \quad D_t = b_t, \quad \beta_t = 0,$$

$$E_t(z) = 0, \quad F_t(z) = c_t(z), \quad \gamma_t(z) = 0, \quad Q_{ij}(t) = R_i(t) = 0$$

and

$$H_1 = 1, \quad H_2 = -2\lambda.$$

Then (3.1) gives the following candidate for the optimal partial information portfolio:

$$(5.13) \quad u_t^* = - \left( \mathbb{E} \left[ \Theta_3(t) | \mathcal{E}_t \right] \right)^{-1} \mathbb{E} \left[ \left\{ \Theta_2(t) x^{(u)}(t-) + \Theta_5(t) + \Theta_8(t) \right\} | \mathcal{E}_t \right],$$

where  $\Theta_i(t)$ ,  $i = 2, 3, 5, 8$  are defined by (2.26)–(2.34). Hence

$$u_t^* = - \left( \mathbb{E} \left[ \left\{ p_t b_t^2 + \int_{\mathbb{R}} (p_t + \mu_t(z)) c_t(z)^2 \nu(dz) \right\} | \mathcal{E}_t \right] \right)^{-1} \\ \times \mathbb{E} \left[ \left\{ (p_t(a_t - \rho_t) + \eta_t b_t) x^{(u)}(t-) + \frac{1}{2} (\tilde{p}_t + \tilde{\eta}_t b_t) + \int_{\mathbb{R}} \tilde{\mu}_t(z) c_t(z) \nu(dz) \right\} | \mathcal{E}_t \right].$$

Here  $p_t$ ,  $\eta_t$ ,  $\mu_t(z)$  and  $\tilde{p}_t$ ,  $\tilde{\eta}_t$ ,  $\tilde{\mu}_t(z)$  are the solutions of the backward Riccati equations (2.3)–(2.6), i.e.,

$$(5.14) \quad dp_t = - \left\{ 2\rho_t p_t + \left[ p_t b_t^2 + \int_{\mathbb{R}} (p_t + \mu_t(z)) c_t^2(z) \nu(dz) \right]^{-1} \left[ p_t(a_t - \rho_t) + \eta_t b_t \right. \right. \\ \left. \left. + \int_{\mathbb{R}} \mu_t(z) c_t(z) \nu(dz) \right]^2 \right\} dt + \eta_t dW_t + \int_{\mathbb{R}} \mu_t(z) \tilde{N}(dt, dz), \quad t < T,$$

$$(5.15) \quad p_T = 1,$$

and

$$(5.16) \quad d\tilde{p}_t = - \left\{ \rho_t \tilde{p}_t + \frac{1}{2} \left[ p_t b_t^2 + \int_{\mathbb{R}} (p_t + \mu_t(z)) c_t^2(z) \nu(dz) \right]^{-1} \left[ p_t(a_t - \rho_t) + \tilde{\eta}_t b_t \right. \right. \\ \left. \left. + \int_{\mathbb{R}} \mu_t(z) c_t(z) \nu(dz) \right] \right\} dt + \tilde{\eta}_t dW_t + \int_{\mathbb{R}} \tilde{\mu}_t(z) \tilde{N}(dt, dz), \quad t < T,$$

$$(5.17) \quad \tilde{p}_T = -2\lambda.$$

Summarizing the above, we get the following.

**THEOREM 5.1.** *Suppose the system of backward Riccati equations (5.14)–(5.17) has a unique solution  $p_t$  and  $\tilde{p}_t$ . Define*

$$(5.18) \quad u_t^* = - \left( \mathbb{E} \left[ \left\{ p_t b_t^2 + \int_{\mathbb{R}} (p_t + \mu_t(z)) c_t(z)^2 \nu(dz) \right\} | \mathcal{E}_t \right] \right)^{-1} \mathbb{E} \left[ \left\{ (p_t(a_t - \rho_t) + \eta_t b_t) x^{(u)}(t-) \right. \right. \\ \left. \left. + \frac{1}{2} (\tilde{p}_t + \tilde{\eta}_t b_t) + \int_{\mathbb{R}} \tilde{\mu}_t(z) c_t(z) \nu(dz) \right\} | \mathcal{E}_t \right].$$

*Suppose  $u_t^* \in \mathcal{A}_{\mathcal{E}}$  and that (2.10) holds. Then  $u_t^*$  is the unique solution to the minimum variance problem (5.12).*

**Remark 5.2.** Suppose the conditions of Theorem 5.1 hold for each choice of  $\lambda \in \mathbb{R}$ . Let  $x_{\lambda}^*(T)$  be the optimal terminal wealth determined by the optimal control  $u_t^* = u_{\lambda,t}^*$  corresponding to  $\lambda$ . Then, in order to solve the original mean-variance portfolio problem (5.10), it remains to determine  $\lambda$  such that

$$\mathbb{E} [x_{\lambda}^*(T)] = K.$$

We omit the discussion of this equation.

**Acknowledgment.** We are grateful to Xunyu Zhou for useful comments.

## REFERENCES

- [1] M. AIT RAMI, J. B. MOORE, AND X. Y. ZHOU, *Indefinite stochastic linear quadratic control and generalized differential Riccati equation*, SIAM J. Control Optim., 40 (2001), pp. 1296–1311.
- [2] M. AIT RAMI AND X. ZHOU, *Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls*, IEEE Trans. Automat. Control, AC-45 (2000), pp. 1131–1143.
- [3] F. BAGHERY AND B. ØKSENDAL, *A maximum principle for stochastic control with partial information*, J. Stochastic Anal. Appl., 25 (2007), pp. 705–717.
- [4] F. E. BENTH, G. DI NUNNO, A. LØKKA, B. ØKSENDAL, AND F. PROSKE, *Explicit representation of the minimal variance portfolio in markets driven by Lévy processes*, Math. Finance, 13 (2003), pp. 55–72.
- [5] S. CHEN, X. LI, AND X. ZHOU, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.
- [6] X. CHEN AND X. ZHOU, *Stochastic linear quadratic regulators with indefinite control weight costs. II*, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.
- [7] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, 2nd ed., Stochastic Modelling and Applied Probability 25, Springer, Berlin, 2006.
- [8] N. C. FRAMSTAD, B. ØKSENDAL, AND A. SULEM, *Sufficient stochastic maximum principle for the optimal control of jump diffusions and applications to finance*, J. Optim. Theory Appl., 121 (2004), pp. 77–98. *Errata*, J. Optim. Theory Appl., 124 (2005), pp. 511–512.
- [9] W. GUO AND C. XU, *Optimal portfolio selection when stock prices follow a jump-diffusion process*, Math. Methods Oper. Res., 60 (2004), pp. 485–496.
- [10] Y. HU AND X. ZHOU, *Indefinite stochastic Riccati equations*, SIAM J. Control Optim., 42 (2003), pp. 123–137.
- [11] Y. HU AND X. M. SONG, *Global Solution of Backward Stochastic Differential Equation with Jumps and Application to Stochastic LQ Control*, preprint, University of Kansas, Lawrence, KS, 2007.
- [12] G. KALLIANPUR, *Stochastic Filtering Theory*, Springer, Berlin, 1980.
- [13] M. KOHLMANN AND S. TANG, *Global adapted solution of one-dimensional backward stochastic Riccati equations, with application to the mean-variance hedging*, Stochastic Process. Appl., 97 (2002), pp. 255–288.
- [14] A. E. B. LIM, *Mean-variance hedging when there are jumps*, SIAM J. Control Optim., 44 (2005), pp. 1893–1922.
- [15] B. ØKSENDAL AND A. SULEM, *Applied Stochastic Control of Jump Diffusions*, 2nd ed., Universitext, Springer, Berlin, 2007.
- [16] S. TANG, *General linear quadratic optimal stochastic control problems with random coefficients: Linear stochastic Hamilton systems and backward stochastic Riccati equations*, SIAM J. Control Optim., 42 (2003), pp. 53–75.
- [17] D. D. YAO, S. ZHANG, AND X. Y. ZHOU, *Stochastic linear-quadratic control via semidefinite programming*, SIAM J. Control Optim., 40 (2001), pp. 801–823.
- [18] J. YONG AND X. ZHOU, *Stochastic Controls. Hamiltonian Systems and HJB Equations*, Appl. Math. 43, Springer, Berlin, 1999.

## ABSTRACT SECOND ORDER HYPERBOLIC SYSTEM AND APPLICATIONS TO CONTROLLED NETWORK OF STRINGS\*

GEN QI XU<sup>†</sup>, DONG YI LIU<sup>†</sup>, AND YAN QING LIU<sup>†</sup>

**Abstract.** In this paper we study an abstract second order hyperbolic system valued in  $\mathbb{C}^N$  with appropriate boundary conditions. We prove that the system is well-posed and associates with a  $C_0$  semigroup in a Hilbert state space. Under certain conditions, we show that the spectra of the system operator are located in the vertical strip, and that there is a sequence of eigenvectors and generalized eigenvectors that forms a Riesz basis with parentheses for the Hilbert state space, and hence that the system satisfies the spectrum determined growth assumption. As applications, we investigate the exponential stability of a controlled tree-shaped network of 7-strings and a network of  $N$ -connected strings.

**Key words.** hyperbolic system, spectral distribution, Riesz basis, exponential stability, tree-shaped network

**AMS subject classifications.** 35P20, 93C20, 93D15

**DOI.** 10.1137/060649367

**1. Introduction.** Many mechanical systems (such as cables, spacecraft with flexible attachments, robots with flexible links, etc.) contain certain parts whose dynamic behavior can be rigorously described by the partial differential equations

$$(1.1) \quad M \frac{\partial^2 Y(x, t)}{\partial t^2} = T \frac{\partial^2 Y(x, t)}{\partial x^2} + P \frac{\partial Y(x, t)}{\partial x} + QY(x, t), \quad x \in (0, 1), \quad t > 0,$$

with appropriate boundary conditions, where  $Y(x, t)$  is a function valued in  $\mathbb{C}^N$ ;  $M$ ,  $T$ ,  $P$ , and  $Q$  are  $N \times N$  matrices; and  $M$  and  $T$  are positive definite matrices. For such systems, we study not only their dynamic behavior but also the control problem. To achieve their control goals, scientists have designed many passive and active controllers.

In the past decade, boundary control of systems governed by partial differential equations has become an important research field. For concrete systems many boundary feedback controllers are used to stabilize the system. For example, in [1], [2], [3], [4], and [19] the authors used boundary and pointwise feedback controllers to stabilize one-dimensional wave systems, and in [9], [10], [11], and the references therein, the boundary feedback controllers have been used in string networks; these are the simplest cases of (1.1) when  $P = Q = 0$ .

For the Timoshenko system whose motion is governed by

$$\begin{cases} \rho w_{tt}(x, t) = K(w_{xx}(x, t) - \varphi_x(x, t)), & 0 < x < \ell, \quad t > 0, \\ I_\rho \varphi_{tt}(x, t) = EI\varphi_{xx}(x, t) + K(w_x(x, t) - \varphi(x, t)), & 0 < x < \ell, \quad t > 0, \end{cases}$$

which is exactly of the form in (1.1), the authors in [12], [13], [14], and [20] used the boundary feedback controllers to exponentially stabilize the systems.

---

\*Received by the editors January 9, 2006; accepted for publication (in revised form) February 27, 2008; published electronically June 13, 2008. This research is supported by the Natural Science Foundation of China grant NSFC-60474017 and by the Liu Hui Center for Applied Mathematics, Nankai University and Tianjin University.

<http://www.siam.org/journals/sicon/47-4/64936.html>

<sup>†</sup>Department of Mathematics, Tianjin University, Tianjin, 300072, People's Republic of China (gqxu@tju.edu.cn, l\_d\_y\_000@163.com, liuyq@eyou.com).

The one-dimensional selling porous solid system, whose dynamic behavior is governed by

$$\begin{cases} \rho_0 u_{tt}(x, t) = \mu u_{xx}(x, t) + \beta \varphi_x(x, t), & 0 < x < \ell, \quad t > 0, \\ \rho_0 K \varphi_{tt}(x, t) = \alpha \varphi_{xx}(x, t) - \beta u_x(x, t) - \xi \varphi(x, t) - \gamma \varphi_t(x, t), & 0 < x < \ell, \quad t > 0, \end{cases}$$

is a case of (1.1) with damping. Within this system, [15] and [16] discussed the slow decay, and [17] used the boundary feedback controllers to achieve the uniform stabilization.

We observe that in all the works mentioned above, the stability analysis of the corresponding closed loop system is the most difficult part; it becomes much more difficult when one uses the method of spectral analysis. This is because one asserts the stability of a system from its spectral distribution only when the system satisfies the spectrum determined growth assumption, which means that the decay rate of the system is determined via the spectrum of the system operator. For a distributed parameter system, to prove that the system itself satisfies the spectrum determined growth assumption is a tough problem. Note that if the system is a Riesz one, that is, the multiplicities of eigenvalues are uniformly bounded and there is a sequence of eigenvectors and generalized eigenvectors that forms a Riesz basis for the Hilbert state space, then the spectrum determined growth assumption holds. So verification of the Riesz basis property in the literature becomes an important component (see [4], [14], [20]). Recently, we found a method for verifying the Riesz basis property (see [24], [25]). This method depends only on the distribution of eigenvalues and not on their expression and eigenvectors, which makes it possible to obtain the Riesz basis of the system by determining the asymptotic distribution of the spectrum.

A practice calculation (see, e.g., [14], [20]) shows that the asymptotic spectra of the system (1.1) are entirely determined by its principal part, i.e.,  $P = Q = 0$ . Based on this fact, in this paper we shall discuss a special case of the system (1.1) with  $P = Q = 0$  under appropriate boundary conditions. More precisely, we shall study the following system valued in  $\mathbb{C}^N$ :

$$(1.2) \quad \begin{cases} M \frac{\partial^2 Y(x, t)}{\partial t^2} = T \frac{\partial^2 Y(x, t)}{\partial x^2}, & x \in (0, 1), \quad t > 0, \\ Y(0, t) = CY(1, t), & t > 0, \\ T \frac{\partial Y(1, t)}{\partial x} - C^T T \frac{\partial Y(0, t)}{\partial x} = -\Gamma \frac{\partial Y(1, t)}{\partial t}, & t > 0, \\ Y(x, 0) = Y_0(x), \quad \frac{\partial Y(x, 0)}{\partial t} = Y_1(x), & x \in (0, 1), \end{cases}$$

where  $M$ ,  $T$ , and  $\Gamma$  are positive definite  $N \times N$  matrices;  $C$  is a real  $N \times N$  matrix satisfying  $\det(I - C) \neq 0$ ; and  $C^T$  denotes the transpose of matrix  $C$ . We shall prove that the system is well-posed and that under certain conditions, the eigenvectors and generalized eigenvectors of the system generate a Riesz basis for the Hilbert state space. Under certain conditions, the system decays exponentially.

It is worth mentioning that system (1.2) is different from those systems mentioned before because it has coupled equations and nonseparable boundary conditions. These properties cause some difficulty in practice. However, such a system has many important applications.

The contents of this paper are organized as follows. In section 2, we shall formulate the problem (1.2) into a Hilbert state space and then prove that the system associates



with a  $C_0$  semigroup and is asymptotically stable. In section 3, we shall study the spectral distribution of the system operator and prove the Riesz basis property of eigenvectors and generalized eigenvectors, which deduces that the system satisfies the spectrum determined growth assumption. Finally, in section 4, we give two examples: a tree-shaped network of 7-strings and a network of N-connected strings. We shall show that both systems are exponentially stable.

**2. Well-posedness of abstract differential equations.** In this section, we shall formulate system (1.2) into a Hilbert space and then discuss whether the system is well-posed.

Let an abstract hyperbolic system valued in  $\mathbb{C}^N$  be given by

$$(2.1) \quad \begin{cases} M \frac{\partial^2 Y(x, t)}{\partial t^2} = T \frac{\partial^2 Y(x, t)}{\partial x^2}, & x \in (0, 1), \quad t > 0, \\ Y(0, t) = CY(1, t), & t > 0, \\ T \frac{\partial Y(1, t)}{\partial x} - C^\tau T \frac{\partial Y(0, t)}{\partial x} = -\Gamma \frac{\partial Y(1, t)}{\partial t}, & t > 0, \\ Y(x, 0) = Y_0(x), \quad \frac{\partial Y(x, 0)}{\partial t} = Y_1(x), & x \in (0, 1), \end{cases}$$

where  $M$ ,  $T$ , and  $\Gamma$  are positive definite matrices, and  $C$  is a real matrix satisfying  $\det(I - C) \neq 0$ .

Set

$$V_E^k(0, 1) = \{f \in H^k([0, 1], \mathbb{C}^N) \mid f(0) = Cf(1)\},$$

where  $H^k((0, 1), \mathbb{C}^N)$  is the Sobolev space of order  $k$ .

Let

$$\mathcal{H} = V_E^1(0, 1) \times L^2([0, 1], \mathbb{C}^N)$$

be equipped with the inner product

$$\langle (f_1, f_2), (g_1, g_2) \rangle_{\mathcal{H}} = \int_0^1 (Tf_1'(x), g_1'(x))dx + \int_0^1 (Mf_2(x), g_2(x))dx,$$

and hereafter we always denote by  $(\cdot, \cdot)$  the inner product in  $\mathbb{C}^N$  and by  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$  the inner product in  $\mathcal{H}$ .

It is easy to see that for  $(f_1, f_2) \in \mathcal{H}$ ,

$$\|(f_1, f_2)\|_{\mathcal{H}} = \left( \int_0^1 (Tf_1'(x), f_1'(x))dx + \int_0^1 (Mf_2(x), f_2(x))dx \right)^{1/2}$$

is a norm on  $\mathcal{H}$ , and that  $\mathcal{H}$  is a Hilbert space.

Define an operator  $\mathcal{A}$  in  $\mathcal{H}$  by

$$(2.2) \quad \mathcal{D}(\mathcal{A}) = \{(f, g) \in V_E^2(0, 1) \times V_E^1(0, 1) \mid Tf'(1) - C^\tau Tf'(0) = -\Gamma g(1)\},$$

$$(2.3) \quad \mathcal{A}(f, g) = (g(x), M^{-1}Tf''(x)) \quad \forall (f, g) \in \mathcal{D}(\mathcal{A}).$$

With the help of the above notation we can rewrite (2.1) as an evolutionary equation in  $\mathcal{H}$  as follows:

$$(2.4) \quad \begin{cases} \frac{d}{dt}Z(t) = \mathcal{A}Z(t), & t > 0, \\ Z(t) = (Y(x, t), Y_t(x, t)), \\ Z(0) = (Y_0(x), Y_1(x)). \end{cases}$$

THEOREM 2.1. *Let  $\mathcal{H}$  and  $\mathcal{A}$  be defined as before; then  $\mathcal{A}$  is a dissipative operator,  $\mathcal{A}^{-1}$  exists and is compact on  $\mathcal{H}$ , and hence  $\mathcal{A}$  generates a  $C_0$  semigroup of contraction on  $\mathcal{H}$ .*

*Proof.* Let  $\mathcal{H}$  and  $\mathcal{A}$  be defined as before. It is easy to verify that  $\mathcal{A}$  is a closed and densely defined linear operator in  $\mathcal{H}$ . For any  $(f, g) \in \mathcal{D}(\mathcal{A})$ , we have

$$\begin{aligned} \langle \mathcal{A}(f, g), (f, g) \rangle_{\mathcal{H}} &= \int_0^1 (Tg'(x), f'(x))dx + \int_0^1 (M(M^{-1}T)f''(x), g(x))dx \\ &= (Tg(x), f'(x)) \Big|_0^1 - \int_0^1 (Tg(x), f''(x))dx + \int_0^1 (Tf''(x), g(x))dx, \\ \langle (f, g), \mathcal{A}(f, g) \rangle_{\mathcal{H}} &= \int_0^1 (Tf'(x), g'(x))dx + \int_0^1 (Mg(x), (M^{-1}Tf''(x)))dx \\ &= (Tf'(x), g(x)) \Big|_0^1 - \int_0^1 (Tf''(x), g(x))dx + \int_0^1 (Tg(x), f''(x))dx, \end{aligned}$$

and hence

$$\begin{aligned} \Re \langle \mathcal{A}(f, g), (f, g) \rangle_{\mathcal{H}} &= \Re (Tg(x), f'(x)) \Big|_0^1 \\ &= \Re (Tf'(1), g(1)) - \Re (Tf'(0), g(0)) \\ &= \Re (Tf'(1), g(1)) - \Re (Tf'(0), Cg(1)) \\ &= \Re (Tf'(1) - C^{\tau}Tf'(0), g(1)) \\ (2.5) \quad &= -(\Gamma g(1), g(1)) \leq 0, \end{aligned}$$

where we have used the conditions  $g(0) = Cg(1)$  and  $Tf'(1) - C^{\tau}Tf'(0) = -\Gamma g(1)$  and  $\Gamma$  is a positive definite matrix. So  $\mathcal{A}$  is a dissipative operator.

Now we show that  $0 \in \rho(\mathcal{A})$ . For any given  $(u, v) \in \mathcal{H}$ , we consider the solvability of equation

$$\mathcal{A}(f, g) = (u, v), \quad (f, g) \in \mathcal{D}(\mathcal{A}),$$

i.e.,

$$(2.6) \quad \begin{cases} g(x) = u(x), & x \in [0, 1], \\ M^{-1}Tf''(x) = v(x), & x \in (0, 1). \end{cases}$$

For the second equation in (2.6), integrating from  $x$  to 1 leads to

$$(2.7) \quad Tf'(1) - Tf'(x) = \int_x^1 Mv(s)ds, \quad x \in (0, 1),$$

and

$$(2.8) \quad (1-x)Tf'(1) - Tf(1) + Tf(x) = \int_x^1 dr \int_r^1 Mv(s)ds, \quad x \in (0, 1).$$

From (2.7) and (2.8) we get

$$(2.9) \quad Tf'(1) - Tf'(0) = \int_0^1 Mv(s)ds,$$

$$(2.10) \quad Tf'(1) - Tf(1) + Tf(0) = \int_0^1 dr \int_r^1 Mv(s)ds.$$

Multiplying (2.9) by  $C^\tau$  and recalling the boundary condition  $Tf'(1) - C^\tau Tf'(0) = -\Gamma g(1) = -\Gamma u(1)$  yield

$$(2.11) \quad (I - C^\tau)Tf'(1) = -\Gamma u(1) - \int_0^1 C^\tau Mv(s)ds.$$

Since  $\det(I - C^\tau) \neq 0$ , we have

$$(2.12) \quad Tf'(1) = -[I - C^\tau]^{-1} \left[ \Gamma u(1) + \int_0^1 C^\tau Mv(s)ds \right].$$

Substituting  $f(0) = Cf(1)$  into (2.10) yields

$$(2.13) \quad Tf'(1) - Tf(1) + TCf(1) = \int_0^1 dr \int_r^1 Mv(s)ds.$$

Thus we get from (2.12) and (2.13) that

$$(2.14) \quad \begin{aligned} f(1) &= -[I - C]^{-1}T^{-1}[I - C^\tau]^{-1} \left[ \Gamma u(1) + \int_0^1 C^\tau Mv(s)ds \right] \\ &\quad - [I - C]^{-1}T^{-1} \int_0^1 dr \int_r^1 Mv(s)ds. \end{aligned}$$

Therefore,

$$(2.15) \quad \begin{aligned} f(x) &= f(1) - (1-x)f'(1) + T^{-1} \int_x^1 dr \int_r^1 Mv(s)ds \\ &= -[xI + (I - C)^{-1}C]T^{-1}(I - C^\tau)^{-1} \left[ \Gamma u(1) + \int_0^1 C^\tau Mv(s)ds \right] \\ &\quad - (I - C)^{-1}CT^{-1} \int_0^1 dr \int_r^1 Mv(s)ds - T^{-1} \int_0^x dr \int_r^1 Mv(s)ds. \end{aligned}$$

Let  $f$  be given by (2.15) and  $g(x) = u(x)$ . Then  $(f, g) \in \mathcal{D}(\mathcal{A})$  and  $\mathcal{A}(f, g) = (u, v)$ . Thus the inverse operator theorem implies that  $0 \in \rho(\mathcal{A})$ . Note that  $u \in V_E^1(0, 1)$  and  $f$  has an integral representation. The Sobolev embedding theorem asserts that  $\mathcal{A}^{-1}$  is compact on  $\mathcal{H}$ . Therefore by the Lumer–Philips theorem (cf. [18]),  $\mathcal{A}$  generates a  $C_0$  semigroup of contraction on  $\mathcal{H}$ .  $\square$

**COROLLARY 2.2.** *Let  $\mathcal{A}$  be defined by (2.2) and (2.3) and  $S(t)$  be the semigroup generated by  $\mathcal{A}$ . Then  $\sigma(\mathcal{A})$  consists of all isolated eigenvalues of finite multiplicity, and  $\sigma(\mathcal{A}) \subset \{\lambda \in \mathbb{C} \mid \Re \lambda < 0\}$ . Therefore,  $S(t)$  is asymptotically stable.*

*Proof.* The first assertion follows from  $\mathcal{A}^{-1}$  being a compact operator on  $\mathcal{H}$ . For any  $\lambda \in \sigma(\mathcal{A})$ , we shall prove that  $\Re \lambda < 0$ . If it is not true, then there is at least one  $\lambda \in \sigma(\mathcal{A})$  with  $\Re \lambda = 0$ . Clearly,  $\lambda \neq 0$ . Let  $(f, g) \in \mathcal{D}(\mathcal{A})$  be a corresponding eigenvector. Then we have  $g(x) = \lambda f(x)$  and

$$0 = \Re \lambda \| (f, g) \|_{\mathcal{H}}^2 = \Re \lambda \langle (f, g), (f, g) \rangle_{\mathcal{H}} = \Re \langle \mathcal{A}(f, g), (f, g) \rangle_{\mathcal{H}} = -(\Gamma g(1), g(1)) \leq 0.$$

Since  $\Gamma$  is a positive definite matrix, we must have  $g(1) = 0$ , and hence  $f(1) = 0$ . So the vector-valued function  $f(x)$  satisfies the following differential equation:

$$(2.16) \quad \begin{cases} \lambda^2 Mf(x) = Tf''(x), & x \in (0, 1), \\ f(0) = 0 = f(1), & Tf'(1) - C^\tau Tf'(0) = 0. \end{cases}$$

Since  $M$  and  $T$  are positive definite matrices, so also is matrix  $T^{-1/2}MT^{-1/2}$ . Set  $B^2 = T^{-1/2}MT^{-1/2}$ , where  $B$  also is a positive definite matrix. Then the general solution of (2.16) has the form

$$f(x) = T^{-1/2} \sin(x\lambda B)v, \quad v \in \mathbb{C}^N.$$

Substituting the above into the boundary conditions in (2.16) leads to

$$T^{1/2} \sin(\lambda B)v = 0, \quad \lambda \left( T^{1/2}B \cos(\lambda B) - C^T T^{-1/2}B \right) v = 0.$$

Since  $\lambda \neq 0$ , we get from the above that

$$(I - C^T)T^{1/2}Bv = 0.$$

Notice that  $\det((I - C^T)T^{1/2}B) \neq 0$ . So  $v = 0$ , i.e.,  $f(x) = 0$ , and hence  $(f, g) = (0, 0)$ . This gives the contradiction that  $(f, g)$  is an eigenvector of  $\mathcal{A}$ . Therefore,  $\Re \lambda < 0$  for any  $\lambda \in \sigma(\mathcal{A})$ . The second assertion is proved. The asymptotical stability of  $S(t)$  follows from Lyubich and Phóng's theorem [21].  $\square$

**3. Spectral analysis of  $\mathcal{A}$ .** In order to investigate the properties of the semi-group  $S(t)$  generated by  $\mathcal{A}$ , we need to find some spectral properties of  $\mathcal{A}$ . In this section, we shall study the distribution of  $\sigma(\mathcal{A})$ , the completeness of eigenvectors and the generalized eigenvectors of  $\mathcal{A}$  and their Riesz basis property.

We begin with the eigenvalue problem. Let  $\lambda \in \mathbb{C}$  be an eigenvalue of  $\mathcal{A}$  and  $(f, g)$  be a corresponding eigenvector. Then we have

$$(3.1) \quad \begin{cases} g(x) = \lambda f(x), & x \in [0, 1], \\ \lambda^2 M f(x) = T f''(x), & x \in (0, 1), \\ f(0) = C f(1), \\ T f'(1) - C^T T f'(0) = -\lambda \Gamma f(1). \end{cases}$$

Set

$$\hat{f}(x) = T^{1/2} f(x), \quad \hat{g}(x) = T^{1/2} g(x), \quad B^2 = T^{-1/2} M T^{-1/2},$$

where  $B$  is a positive definite matrix. Then (3.1) is equivalent to the following equation:

$$(3.2) \quad \begin{cases} \hat{g}(x) = \lambda \hat{f}(x), & x \in [0, 1], \\ \lambda^2 B^2 \hat{f}(x) = \hat{f}''(x), & x \in (0, 1), \\ \hat{f}(0) = T^{1/2} C T^{-1/2} \hat{f}(1), \\ \hat{f}'(1) - T^{-1/2} C^T T^{1/2} \hat{f}'(0) = -\lambda T^{-1/2} \Gamma T^{-1/2} \hat{f}(1). \end{cases}$$

Clearly, the general solution of the differential equation in (3.2) is of the form

$$(3.3) \quad \hat{f}(x) = e^{x\lambda B} u + e^{-x\lambda B} v, \quad u, v \in \mathbb{C}^N.$$

Substituting this into the boundary conditions in (3.2) leads to

$$(3.4) \quad \begin{cases} (B + T^{-1/2} \Gamma T^{-1/2}) e^{\lambda B} u + (T^{-1/2} \Gamma T^{-1/2} - B) e^{-\lambda B} v \\ = T^{-1/2} C^T T^{1/2} B(u - v), \\ (u + v) = T^{1/2} C T^{-1/2} (e^{\lambda B} u + e^{-\lambda B} v). \end{cases}$$

The above algebraic equation has a nonzero solution  $(u, v)$  and this implies that the determinant of the coefficient matrix vanishes, i.e.,

$$(3.5) \quad D(\lambda) = \det \begin{bmatrix} I - \widehat{C}e^{\lambda B} & I - \widehat{C}e^{-\lambda B} \\ (B + \widehat{\Gamma})e^{\lambda B} - \widehat{C}^\tau B & (\widehat{\Gamma} - B)e^{-\lambda B} + \widehat{C}^\tau B \end{bmatrix} = 0,$$

where  $\widehat{C} = T^{1/2}CT^{-1/2}$ ,  $\widehat{\Gamma} = T^{-1/2}\Gamma T^{-1/2}$ .

Conversely, if  $\lambda \in \mathbb{C}$  such that  $D(\lambda) = 0$ , then (3.4) has a nonzero solution  $(u, v)$ . We can define function  $\widehat{f}$  as being the same as (3.3). Obviously,  $\widehat{f}$  satisfies the equation

$$\widehat{f}''(x) = \lambda^2 B^2 \widehat{f}(x).$$

Equation (3.4) implies that  $\widehat{f}$  satisfies the boundary conditions in (3.2). Consequently, the functions

$$f(x) = T^{-1/2}\widehat{f}(x), \quad g(x) = \lambda T^{-1/2}\widehat{f}(x)$$

satisfy (3.1). Therefore,  $\lambda$  is an eigenvalue of  $\mathcal{A}$ .

Since

$$\begin{aligned} D(\lambda) &= \det \begin{bmatrix} I - \widehat{C}e^{\lambda B} & I - \widehat{C}e^{-\lambda B} \\ (B + \widehat{\Gamma})e^{\lambda B} - \widehat{C}^\tau B & (\widehat{\Gamma} - B)e^{-\lambda B} + \widehat{C}^\tau B \end{bmatrix} \\ &= \det \begin{bmatrix} e^{-\lambda B} - \widehat{C} & I - \widehat{C}e^{-\lambda B} \\ (B + \widehat{\Gamma}) - \widehat{C}^\tau B e^{-\lambda B} & (\widehat{\Gamma} - B)e^{-\lambda B} + \widehat{C}^\tau B \end{bmatrix} \det \begin{bmatrix} e^{\lambda B} & 0 \\ 0 & I \end{bmatrix} \\ &= \det \begin{bmatrix} I - \widehat{C}e^{\lambda B} & e^{\lambda B} - \widehat{C} \\ (B + \widehat{\Gamma})e^{\lambda B} - \widehat{C}^\tau B & (\widehat{\Gamma} - B) + \widehat{C}^\tau B e^{\lambda B} \end{bmatrix} \det \begin{bmatrix} I & 0 \\ 0 & e^{-\lambda B} \end{bmatrix}, \end{aligned}$$

when  $\Re \lambda \rightarrow \pm\infty$ , we have

$$(3.6) \quad \lim_{\Re \lambda \rightarrow +\infty} \frac{D(\lambda)}{\det(e^{\lambda B})} = \det \begin{bmatrix} -\widehat{C} & I \\ (B + \widehat{\Gamma}) & \widehat{C}^\tau B \end{bmatrix} = (-1)^n \det[\widehat{\Gamma} + B + \widehat{C}^\tau B \widehat{C}]$$

and

$$(3.7) \quad \lim_{\Re \lambda \rightarrow -\infty} \frac{D(\lambda)}{\det(e^{-\lambda B})} = \det \begin{bmatrix} I & -\widehat{C} \\ -\widehat{C}^\tau B & (\widehat{\Gamma} - B) \end{bmatrix} = \det[\widehat{\Gamma} - B - \widehat{C}^\tau B \widehat{C}].$$

Therefore, we have the following result.

**THEOREM 3.1.** *Let  $\mathcal{A}$  be defined as in (2.2) and (2.3) and let*

$$(3.8) \quad B^2 = T^{-1/2}MT^{-1/2}, \quad \widehat{C} = T^{1/2}CT^{-1/2}, \quad \widehat{\Gamma} = T^{-1/2}\Gamma T^{-1/2},$$

*and  $D(\lambda)$  be defined by (3.5). Then*

$$(3.9) \quad \sigma(\mathcal{A}) = \{\lambda \in \mathbb{C} \mid D(\lambda) = 0\}.$$

*When  $\det[\widehat{\Gamma} - B - \widehat{C}^\tau B \widehat{C}] \neq 0$ , there is a positive constant  $h > 0$  such that*

$$(3.10) \quad \sigma(\mathcal{A}) \subset \{\lambda \in \mathbb{C} \mid -h \leq \Re \lambda < 0\}.$$

In this case,  $\sigma(\mathcal{A})$  is a union of finite many separated sets.

*Proof.* Let  $\lambda \in \mathbb{C}$  with  $\lambda \neq 0$ . For any given  $(u, v) \in \mathcal{H}$ , we consider the resolvent equation  $(\lambda I - \mathcal{A})(f, g) = (u, v)$ , i.e.,

$$(3.11) \quad \begin{cases} \lambda f - g = u, \\ \lambda g - M^{-1}Tf'' = v, \\ f(0) = Cf(1), \\ Tf'(1) - C^\tau Tf'(0) = -\Gamma g(1). \end{cases}$$

So,  $g = \lambda f - u$ , and  $f$  satisfies the following differential equation:

$$(3.12) \quad \begin{cases} \lambda^2 f(x) - M^{-1}Tf''(x) = \lambda u(x) + v(x), & x \in (0, 1), \\ f(0) = Cf(1), \\ Tf'(1) - C^\tau Tf'(0) + \lambda \Gamma f(1) = \Gamma u(1). \end{cases}$$

Set

$$\hat{f}(x) = T^{1/2}f(x), \quad \hat{g}(x) = T^{1/2}g(x), \quad \hat{u}(x) = T^{1/2}u(x), \quad \hat{v}(x) = T^{1/2}v(x).$$

Then (3.12) is changed into

$$(3.13) \quad \begin{cases} \lambda^2 B^2 \hat{f}(x) - \hat{f}''(x) = \lambda B^2 \hat{u}(x) + B^2 \hat{v}(x), & x \in (0, 1), \\ \hat{f}(0) = \hat{C}\hat{f}(1), \\ \hat{f}'(1) - \hat{C}^\tau \hat{f}'(0) + \lambda \hat{\Gamma} \hat{f}(1) = \hat{\Gamma} \hat{u}(1), \end{cases}$$

where  $B$ ,  $\hat{C}$ , and  $\hat{\Gamma}$  are defined as in (3.8).

Clearly, the differential equation in (3.13) has the general solution

$$(3.14) \quad \hat{f}(x) = e^{x\lambda B}y + e^{-x\lambda B}z - \int_0^x \sinh(\lambda(x-s)B)[B\hat{u}(s) + \lambda^{-1}B\hat{v}(s)]ds,$$

where  $y, z \in \mathbb{C}^N$ . Substituting (3.14) into the boundary conditions in (3.13) yields

$$(3.15) \quad \begin{cases} (I - \hat{C}e^{\lambda B})y + (I - \hat{C}e^{-\lambda B})z \\ = -\hat{C} \int_0^1 \sinh(\lambda(1-s)B)[B\hat{u}(s) + \lambda^{-1}B\hat{v}(s)]ds, \\ \left[ (\hat{\Gamma} + B)e^{\lambda B} - \hat{C}^\tau B \right] y + \left[ (\hat{\Gamma} - B)e^{-\lambda B} + \hat{C}^\tau B \right] z \\ = \int_0^1 \left[ B \cosh(\lambda(1-s)B) + \hat{\Gamma} \sinh(\lambda(1-s)B) \right] [B\hat{u}(s) + \lambda^{-1}B\hat{v}(s)]ds \\ + \lambda^{-1}\Gamma u(1). \end{cases}$$

Since the coefficient matrix of the above algebraic equations is

$$(3.16) \quad \tilde{G}(\lambda) = \begin{pmatrix} I - \hat{C}e^{\lambda B} & I - \hat{C}e^{-\lambda B} \\ (\hat{\Gamma} + B)e^{\lambda B} - \hat{C}^\tau B & (\hat{\Gamma} - B)e^{-\lambda B} + \hat{C}^\tau B \end{pmatrix},$$

when  $D(\lambda) = \det \tilde{G}(\lambda) \neq 0$ , we have

$$(3.17) \quad \tilde{G}^{-1}(\lambda) = \frac{\text{adj } \tilde{G}(\lambda)}{D(\lambda)} = \frac{1}{D(\lambda)} \begin{pmatrix} \hat{G}_{11}(\lambda) & \hat{G}_{12}(\lambda) \\ \hat{G}_{21}(\lambda) & \hat{G}_{22}(\lambda) \end{pmatrix},$$

where  $\text{adj}(S)$  denotes the adjoint matrix of  $S$ . We define functionals  $F_1, F_2$  on  $\mathcal{H}$  by

$$(3.18) \quad \begin{aligned} F_1(u, v, \lambda) &= -\widehat{G}_{11}(\lambda)\widehat{C} \int_0^1 \sinh(1-s)[B\widehat{u}(s) + \lambda^{-1}B\widehat{v}(s)]ds \\ &+ \widehat{G}_{12}(\lambda) \int_0^1 \left[ \widehat{\Gamma} \sinh(\lambda(1-s)B) + B \cosh(\lambda(1-s)B) \right] [B\widehat{u}(s) + \lambda^{-1}B\widehat{v}(s)]ds \\ &+ \widehat{G}_{12}(\lambda)\lambda^{-1}\widehat{\Gamma}\widehat{u}(1), \end{aligned}$$

$$(3.19) \quad \begin{aligned} F_2(u, v, \lambda) &= -\widehat{G}_{21}(\lambda)\widehat{C} \int_0^1 \sinh(\lambda((1-s)B)[B\widehat{u}(s) + \lambda^{-1}B\widehat{v}(s)]ds \\ &+ \widehat{G}_{22}(\lambda) \int_0^1 \left[ B \cosh(\lambda(1-s)B) + \widehat{\Gamma} \sinh(\lambda(1-s)B) \right] [B\widehat{u}(s) + \lambda^{-1}B\widehat{v}(s)]ds \\ &+ G_{22}(\lambda)\lambda^{-1}\widehat{\Gamma}\widehat{u}(1). \end{aligned}$$

Obviously,  $F_1$  and  $F_2$  as defined by (3.18) and (3.19), respectively, are bounded linear functionals on  $\mathcal{H}$ , and the solution to (3.15) is given by

$$(y, z) = D^{-1}(\lambda)(F_1(u, v, \lambda), F_2(u, v, \lambda)).$$

Therefore, when  $\lambda \in \mathbb{C}$  with  $D(\lambda) \neq 0$ , we have

$$(3.20) \quad \begin{aligned} \widehat{f}(x) &= D^{-1}(\lambda) [e^{x\lambda B}F_1(u, v, \lambda) + e^{-x\lambda B}F_2(u, v, \lambda)] \\ &- \int_0^x \sinh((x-s)\lambda B)[B\widehat{u}(s) + \lambda^{-1}B\widehat{v}(s)]ds. \end{aligned}$$

Set

$$(3.21) \quad f(x) = T^{-1/2}\widehat{f}(x), \quad g(x) = \lambda T^{-1/2}\widehat{f}(x) - u(x).$$

A straightforward calculation shows  $(f, g) \in \mathcal{D}(\mathcal{A})$  and

$$(\lambda I - \mathcal{A})(f, g) = (u, v).$$

So we have that  $\lambda \in \rho(\mathcal{A})$ . The first assertion follows.

Now let  $D(\lambda)$  be defined by (3.5). Then  $D(\lambda)$  is an entire function of finite exponential type on complex plane  $\mathbb{C}$ . From (3.6) and (3.7) we can see that, when  $\det[\widehat{\Gamma} - B - \widehat{C}^\tau B \widehat{C}] \neq 0$ , there are positive constants  $c_1, c_2$ , and  $h$  such that, as  $|\Re \lambda| \geq h$ ,

$$(3.22) \quad c_1 \det(e^{|\lambda|B}) \leq |D(\lambda)| \leq c_2 \det(e^{|\lambda|B}).$$

This means that  $D(\lambda)$  is a sine-type function on  $\mathbb{C}$  (see [23, Definition II.1.27, p. 61]). Levin's theorem (see [23, Proposition II.1.28]) asserts that the set of zeros of  $D(\lambda)$  is a union of finitely many separable sets. So also is  $\sigma(\mathcal{A})$ . The proof is then complete.  $\square$

In what follows we shall discuss the completeness of eigenvectors and generalized eigenvectors of  $\mathcal{A}$ . For this purpose, we begin with the following proposition.

**PROPOSITION 3.2.** *Let  $\mathcal{H}$  be defined as before. Define operator  $\mathcal{A}_0$  in  $\mathcal{H}$  by*

$$\begin{aligned} \mathcal{D}(\mathcal{A}_0) &= \{(f, g) \in V_E^2(0, 1) \times V_E^1(0, 1) \mid T f'(1) - C^\tau T f'(0) = 0\}, \\ \mathcal{A}_0(f, g) &= (g(x), M^{-1}T f''(x)). \end{aligned}$$

Then  $\mathcal{A}_0$  is a skew adjoint operator in  $\mathcal{H}$ , and for any  $(u, v) \in \mathcal{H}$ ,  $\lambda \in \mathbb{R}$ , the solution  $(f_\lambda, g_\lambda)$  of the resolvent equation

$$\lambda(f, g) - \mathcal{A}_0(f, g) = (u, v)$$

satisfies

$$\|g_\lambda(1)\| \leq M\|(u, v)\|_{\mathcal{H}},$$

where  $M > 0$  is a constant.

*Proof.* For any  $(f_i, g_i) \in \mathcal{D}(\mathcal{A}_0)$ ,  $i = 1, 2$ , it holds that

$$\begin{aligned} \langle \mathcal{A}_0(f_1, g_1), (f_2, g_2) \rangle_{\mathcal{H}} &= \int_0^1 (Tg'_1(x), f'_2(x))dx + \int_0^1 (MM^{-1}Tf''_1(x), g_2(x))dx \\ &= (Tg_1(x), f'_2(x)) \Big|_0^1 - \int_0^1 (Tg_1(x), f''_2(x))dx \\ &\quad + (Tf'_1(x), g_2(x)) \Big|_0^1 - \int_0^1 ((Tf'_1(x), g'_2(x))dx \\ &= (Tg_1(x), f'_2(x)) \Big|_0^1 + (Tf'_1(x), g_2(x)) \Big|_0^1 - \langle (f_1, g_1), \mathcal{A}_0(f_2, g_2) \rangle_{\mathcal{H}} \\ &= (g_1(1), Tf'_2(1)) - (g_1(0), Tf'_2(0)) + (Tf'_1(1), g_2(1)) - (Tf'_1(0), g_2(0)) \\ &\quad - \langle (f_1, g_1), \mathcal{A}_0(f_2, g_2) \rangle_{\mathcal{H}} \\ &= (g_1(1), Tf'_2(1) - C^T T f'(0)) + (Tf'_1(1) - C^T Tf'_1(0), g_2(1)) - \langle (f_1, g_1), \mathcal{A}_0(f_2, g_2) \rangle_{\mathcal{H}} \\ &= -\langle (f_1, g_1), \mathcal{A}_0(f_2, g_2) \rangle_{\mathcal{H}}. \end{aligned}$$

So,  $\mathcal{A}_0^* = -\mathcal{A}_0$ .

Now let  $(u, v) \in \mathcal{H}$  be given and  $\lambda \in \mathbb{R}$ . Let  $(f_\lambda, g_\lambda)$  satisfy the resolvent equation

$$(\lambda I - \mathcal{A}_0)(f, g) = (u, v), \quad (f, g) \in \mathcal{D}(\mathcal{A}_0),$$

i.e.,

$$\lambda f_\lambda(x) - g_\lambda(x) = u(x), \quad \lambda g_\lambda(x) - M^{-1}Tf''_\lambda(x) = v(x)$$

and

$$f_\lambda(0) = Cf_\lambda(1), \quad Tf'_\lambda(1) - C^T Tf'_\lambda(0) = 0.$$

Since

$$f_\lambda(1) = \int_0^1 f'_\lambda(x)dx + f_\lambda(0) = \int_0^1 f'_\lambda(x)dx + Cf_\lambda(1),$$

we have

$$f_\lambda(1) = (I - C)^{-1} \int_0^1 f'_\lambda(x)dx.$$

Similarly,

$$u(1) = (I - C)^{-1} \int_0^1 u'(x)dx.$$



Thus,

$$g_\lambda(1) = \lambda f_\lambda(1) - u(1) = (I - C)^{-1}T^{-1/2} \left[ \lambda \int_0^1 T^{1/2} f'(x) dx - \int_0^1 T^{1/2} u'(x) dx \right].$$

Consequently,

$$\begin{aligned} \|g_\lambda(1)\| &\leq \left\| (I - C)^{-1}T^{-1/2} \right\| \left[ |\lambda| \int_0^1 (T f'(x), f'(x)) dx + \int_0^1 (T u'(x), u'(x)) dx \right] \\ &\leq \left\| (I - C)^{-1}T^{-1/2} \right\| [\|\lambda\| R(\lambda, \mathcal{A}_0)(u, v)\|_{\mathcal{H}} + \|(u, v)\|_{\mathcal{H}}]. \end{aligned}$$

Since  $\mathcal{A}_0$  is a skew adjoint operator,  $\|\lambda R(\lambda, \mathcal{A}_0)\| \leq 1$ ,  $\lambda \in \mathbb{R}$ , we have

$$\|g_\lambda(1)\| \leq 2\|(I - C)^{-1}T^{-1/2}\| \|(u, v)\|_{\mathcal{H}} \quad \forall \lambda \in \mathbb{R}.$$

The desired result follows.  $\square$

**THEOREM 3.3.** *Let  $\mathcal{H}$  and  $\mathcal{A}$  be defined as before. If  $\det(\widehat{\Gamma} - B - \widehat{C}^\tau B \widehat{C}) \neq 0$ , then the system of eigenvectors and generalized eigenvectors of  $\mathcal{A}$  is complete in  $\mathcal{H}$ .*

*Proof.* Let  $\mathcal{H}$  and  $\mathcal{A}$  be defined as before, and let  $\mathcal{A}_0$  be as defined in Proposition 3.2. Denote

$$Sp(\mathcal{A}) = \overline{\text{span} \left\{ \sum y_k, y_k \in E(\lambda_k, \mathcal{A})\mathcal{H} \quad \forall \lambda_k \in \sigma(\mathcal{A}) \right\}},$$

where  $E(\lambda_k, \mathcal{A})$  is the Riesz projector corresponding to  $\lambda_k$ . We shall prove  $Sp(\mathcal{A}) = \mathcal{H}$ .

Let  $(u_0, v_0) \in \mathcal{H}$  and  $(u_0, v_0) \perp Sp(\mathcal{A})$ . Then  $R^*(\lambda, \mathcal{A})(u_0, v_0)$  is an entire function on  $\mathbb{C}$  valued in  $\mathcal{H}$ . For any  $(u, v) \in \mathcal{H}$ , we define a function on complex plane  $\mathbb{C}$  by

$$F(\lambda) = \langle (u, v), R^*(\lambda, \mathcal{A})(u_0, v_0) \rangle_{\mathcal{H}}.$$

Clearly,  $F(\lambda)$  is an entire function of finite exponential type, and

$$|F(\lambda)| \leq (\Re \lambda)^{-1} \|(u, v)\|_{\mathcal{H}} \|(u_0, v_0)\|_{\mathcal{H}}, \quad \Re \lambda > 0,$$

and hence  $\lim_{\Re \lambda \rightarrow +\infty} |F(\lambda)| = 0$ .

Now we consider the solution of equations

$$(\lambda I - \mathcal{A})(f_{1\lambda}, g_{1\lambda}) = (u, v), \quad (\lambda I - \mathcal{A}_0)(f_{2\lambda}, g_{2\lambda}) = (u, v), \quad \lambda \in \rho(\mathcal{A}) \cap \rho(\mathcal{A}_0) \cap \mathbb{R}_-.$$

Set

$$\varphi(x) = f_{1\lambda}(x) - f_{2\lambda}(x), \quad \psi(x) = g_{1\lambda}(x) - g_{2\lambda}(x).$$

Then we have

$$R(\lambda, \mathcal{A})(u, v) = (f_{1\lambda}, g_{1\lambda}) = (f_{2\lambda}, g_{2\lambda}) + (\varphi, \psi) = R(\lambda, \mathcal{A}_0)(u, v) + (\varphi, \psi),$$

$\psi(x) = \lambda \varphi(x)$ , and  $\varphi$  satisfies the following equations:

$$\begin{cases} \lambda^2 M \varphi(x) = T \varphi''(x), & x \in (0, 1), \\ \varphi(0) = C \varphi(1), \\ T \varphi'(1) - C^\tau T \varphi'(0) + \lambda \Gamma \varphi(1) = \Gamma g_{2\lambda}(1). \end{cases}$$

Setting  $\widehat{\varphi}(x) = T^{1/2}\varphi(x)$  and using the previous notation, we have

$$\widehat{\varphi}(x) = e^{x\lambda B}y + e^{-x\lambda B}z, \quad y, z \in \mathbb{C}^N,$$

where  $y$  and  $z$  solve the equations

$$\begin{cases} (I - \widehat{C}e^{\lambda B})y + (I - \widehat{C}e^{-\lambda B})z = 0, \\ [(\widehat{\Gamma} + B)e^{\lambda B} - \widehat{C}^\tau B]y + [(\widehat{\Gamma} - B)e^{-\lambda B} + \widehat{C}^\tau B]z = \lambda^{-1}T^{1/2}\Gamma g_{2\lambda}(1). \end{cases}$$

Thus

$$\begin{cases} y = (I - \widehat{C}e^{\lambda B})^{-1}(\widehat{C} - e^{\lambda B})e^{-\lambda B}z = (\widehat{C} - O_1(\lambda))e^{-\lambda B}z, \\ [\widehat{\Gamma} - B - \widehat{C}^\tau B\widehat{C} - O_2(\lambda)]e^{-\lambda B}z = \lambda^{-1}T^{1/2}\Gamma g_{2\lambda}(1), \end{cases}$$

where  $\|O_j(\lambda)\| = o(\lambda^{-1})$ ,  $j = 1, 2$ , as  $\lambda \rightarrow -\infty$ . This means that

$$e^{-\lambda B}z = \lambda^{-1}(\widehat{\Gamma} - B - \widehat{C}^\tau B\widehat{C} + O_2(\lambda))^{-1}T^{1/2}\Gamma g_{2\lambda}(1).$$

Therefore, when  $|\lambda|$  is sufficiently large, we have

$$\begin{aligned} \widehat{\varphi}(1) &= e^{\lambda B}y + e^{-\lambda B}z = e^{\lambda B}(\widehat{C} - O_1(\lambda))e^{-\lambda B}z + e^{-\lambda B}z \\ &= (I + O_3(\lambda))e^{-\lambda B}z \\ &= \lambda^{-1}(I + O_3(\lambda))(\widehat{\Gamma} - B - \widehat{C}^\tau B\widehat{C} + O_2(\lambda))^{-1}T^{1/2}\Gamma g_{2\lambda}(1). \end{aligned}$$

So, there is a constant  $M_1 > 0$  such that

$$\|\widehat{\varphi}(1)\| \leq M_1|\lambda^{-1}|\|g_{2\lambda}(1)\|.$$

Thus,

$$\begin{aligned} \|(\varphi, \psi)\|_{\mathcal{H}}^2 &= \int_0^1 (T\varphi'(x), \varphi'(x))dx + \int_0^1 (M\psi(x), \psi(x))dx \\ &= (T\varphi'(1), \varphi(1)) - (T\varphi'(0), \varphi(0)) \\ &= -\lambda(\Gamma\varphi(1), \varphi(1)) + (\Gamma g_{2\lambda}, \varphi(1)) \\ &= -\lambda(\widehat{\Gamma}\widehat{\varphi}(1), \widehat{\varphi}(1)) + (T^{-1/2}\Gamma g_{2\lambda}(1), \widehat{\varphi}(1)) \\ &\leq -\lambda\|\widehat{\Gamma}\|\|\widehat{\varphi}(1)\|^2 + \|T^{-1/2}\Gamma\|\|g_{2\lambda}\|\|\widehat{\varphi}(1)\| \\ &\leq |\lambda^{-1}|\|\Gamma\|M_1^2\|g_{2\lambda}(1)\|^2 + \|T^{-1/2}\Gamma\|\|g_{2\lambda}(1)\|M_1|\lambda^{-1}|\|g_{2\lambda}\| \\ &\leq M_2|\lambda^{-1}|\|g_{2\lambda}(1)\|^2, \end{aligned}$$

where  $M_2$  is a positive constant. According to Proposition 3.2, we have  $\|g_{2\lambda}(1)\| \leq M\|(u, v)\|_{\mathcal{H}}$ , and hence there is a positive constant  $M_3$  such that

$$\|(\varphi, \psi)\|_{\mathcal{H}} \leq M_3\sqrt{|\lambda^{-1}|}\|(u, v)\|_{\mathcal{H}}.$$

Therefore we get that, for  $\lambda \in \rho(\mathcal{A}) \cap \mathbb{R}_-$  with  $|\lambda|$  large enough,

$$\begin{aligned} |F(\lambda)| &= |\langle R(\lambda, \mathcal{A})(u, v), (u_0, v_0) \rangle_{\mathcal{H}}| = |\langle R(\lambda, \mathcal{A}_0)(u, v), (u_0, v_0) \rangle_{\mathcal{H}} + \langle (\varphi, \psi), (u_0, v_0) \rangle_{\mathcal{H}}| \\ &\leq |\lambda^{-1}|\|(u, v)\|_{\mathcal{H}}\|(u_0, v_0)\|_{\mathcal{H}} + \|(\varphi, \psi)\|_{\mathcal{H}}\|(u_0, v_0)\|_{\mathcal{H}} \\ &\leq |\lambda^{-1}|\|(u, v)\|_{\mathcal{H}}\|(u_0, v_0)\|_{\mathcal{H}} + M_3\sqrt{|\lambda^{-1}|}\|(u, v)\|_{\mathcal{H}}\|(u_0, v_0)\|_{\mathcal{H}}. \end{aligned}$$

Since  $F(\lambda)$  is an entire function of finite exponential type, the above inequality, together with the Phragmén–Lindelöf theorem (cf. [22]), implies  $F(\lambda) \equiv 0$ . So  $R^*(\lambda, \mathcal{A})(u_0, v_0) \equiv 0$ , which implies  $(u_0, v_0) = 0$ . Thus  $Sp(\mathcal{A}) = \mathcal{H}$ . The proof is then complete.  $\square$

In order to obtain the Riesz basis property of eigenvectors and generalized eigenvectors of  $\mathcal{A}$ , we need the following theorem, which comes from [25] and is an extension of the result in [24].

**THEOREM 3.4.** *Let  $\mathcal{A}$  be the generator of a  $C_0$  semigroup  $T(t)$  on a separable Hilbert space  $\mathcal{H}$ . Suppose that the following conditions are satisfied:*

(1) *The spectrum of  $\mathcal{A}$  has a decomposition*

$$(3.23) \quad \sigma(\mathcal{A}) = \sigma_1(\mathcal{A}) \cup \sigma_2(\mathcal{A}),$$

*where  $\sigma_2(\mathcal{A})$  consists of the isolated eigenvalues of  $\mathcal{A}$  of finite multiplicity (repeated many times according to its algebraic multiplicity).*

(2) *There exists a real number  $\alpha \in \mathbb{R}$  such that*

$$(3.24) \quad \sup\{\Re \lambda, \lambda \in \sigma_1(\mathcal{A})\} \leq \alpha \leq \inf\{\Re \lambda, \lambda \in \sigma_2(\mathcal{A})\}.$$

(3) *The set  $\sigma_2(\mathcal{A})$  is a union of finite many separated sets.*

*Then the following statements are true:*

(i) *There exist two  $T(t)$ -invariant closed subspaces  $\mathcal{H}_1$  and  $\mathcal{H}_2$  with the property that  $\sigma(\mathcal{A}|_{\mathcal{H}_1}) = \sigma_1(\mathcal{A})$ ,  $\sigma(\mathcal{A}|_{\mathcal{H}_2}) = \sigma_2(\mathcal{A})$ , and there exists a finite combination  $E(\Omega_k, \mathcal{A})$  of some  $\{E(\lambda_k, \mathcal{A})\}_{k=1}^\infty$ ,*

$$(3.25) \quad E(\Omega_k, \mathcal{A}) = \sum_{\lambda \in \Omega_k \cap \sigma_2(\mathcal{A})} E(\lambda, \mathcal{A}),$$

*such that  $\{E(\Omega_k, \mathcal{A})\mathcal{H}_2\}_{k \in \mathbb{N}}$  forms a Riesz basis of subspaces for  $\mathcal{H}_2$  (see [26, p. 332]). Furthermore,*

$$\mathcal{H} = \overline{\mathcal{H}_1 \oplus \mathcal{H}_2}.$$

(ii) *If  $\sup_{k \geq 1} \|E(\lambda_k, \mathcal{A})\| < \infty$ , then*

$$(3.26) \quad \mathcal{D}(\mathcal{A}) \subset \mathcal{H}_1 \oplus \mathcal{H}_2 \subset \mathcal{H}.$$

(iii)  *$\mathcal{H}$  has a decomposition of the topological direct sum,  $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ , if and only if*

$$(3.27) \quad \sup_{n \geq 1} \left\| \sum_{k=1}^n E(\Omega_k, \mathcal{A}) \right\| < \infty.$$

Combining Theorems 3.1, 3.3, and 3.4, we can get the following result.

**THEOREM 3.5.** *Let  $\mathcal{A}$  be defined by (2.2) and (2.3); let  $S(t)$  be the  $C_0$  semigroup associated with  $\mathcal{A}$ ; and let  $B$ ,  $\widehat{\Gamma}$ , and  $\widehat{C}$  be defined as in Theorem 3.1. If  $\det(\widehat{\Gamma} - B - \widehat{C}^\tau B \widehat{C}) \neq 0$ , then there is a sequence of eigenvectors and generalized eigenvectors of  $\mathcal{A}$  that forms a Riesz basis with parentheses for  $\mathcal{H}$ . Therefore,  $S(t)$  satisfies the spectrum determined growth assumption. In addition,  $S(t)$  is in fact a  $C_0$  group on  $\mathcal{H}$ .*

*Proof.* Let  $\mathcal{A}$  be defined by (2.2) and (2.3), and let  $S(t)$  be the  $C_0$  semigroup associated with  $\mathcal{A}$ . Set  $\sigma_1(\mathcal{A}) = \{-\infty\}$ ,  $\sigma_2(\mathcal{A}) = \sigma(\mathcal{A})$ . Theorem 3.1 shows that

all the conditions in Theorem 3.4 are fulfilled, so the results of Theorem 3.4 apply. Thus there is a sequence of eigenvectors and generalized eigenvectors of  $\mathcal{A}$  that forms a Riesz basis with parentheses for  $\mathcal{H}_2$ . Theorem 3.3 says that the eigenvectors and generalized eigenvectors sequence is complete in  $\mathcal{H}$ , that is,  $\mathcal{H}_2 = \mathcal{H}$ . Therefore the sequence is also a Riesz basis with parentheses for  $\mathcal{H}$ . The basis property, together with the uniform boundedness of the multiplicities of eigenvalues of  $\mathcal{A}$ , implies that  $S(t)$  satisfies the spectrum determined growth assumption. Also, the basis property of eigenvectors and generalized eigenvectors, together with the spectral distribution of  $\mathcal{A}$ , asserts that  $\mathcal{A}$  generates a  $C_0$  group on  $\mathcal{H}$ . The proof is then complete.  $\square$

As a consequence of the Riesz basis property, we have the following stability result of the system.

**THEOREM 3.6.** *Let  $\mathcal{H}$  and  $\mathcal{A}$  be defined as before, and let  $B$ ,  $\widehat{\Gamma}$ , and  $\widehat{C}$  be defined as in Theorem 3.1. Let  $\det(\widehat{\Gamma} - B - \widehat{C}^\tau B \widehat{C}) \neq 0$  and  $D(\lambda)$  be defined as (3.5). Then the following statements are true:*

- (1) *If  $\inf_{\lambda \in i\mathbb{R}} |D(\lambda)| \neq 0$ , then the system (2.4) is exponentially stable.*
- (2) *If  $\inf_{\lambda \in i\mathbb{R}} |D(\lambda)| = 0$ , then the system (2.4) is asymptotically stable but not exponentially stable.*

*Proof.* Under the above assumptions, Theorem 3.5 shows that the system (2.4) is a Riesz system and satisfies the spectrum determined growth condition. Note that

$$\sigma(\mathcal{A}) = \{\lambda \in \mathbb{C} \mid D(\lambda) = 0\}.$$

If  $\inf_{\lambda \in i\mathbb{R}} |D(\lambda)| \neq 0$ , then the imaginary axis is not an asymptote of  $\sigma(\mathcal{A})$ , which implies the system is exponentially stable. If  $\inf_{\lambda \in i\mathbb{R}} |D(\lambda)| = 0$ , then the imaginary axis is an asymptote of  $\sigma(\mathcal{A})$ , and hence the system is asymptotically stable but not exponentially stable. The proof is then complete.  $\square$

In the discussion above, we assumed that  $\Gamma$  is a positive definite matrix to ensure that there are no eigenvalues on the imaginary axis. This restriction can be relaxed—indeed, if  $\Gamma$  is a nonnegative matrix, then the results in Theorems 3.5 and 3.6 apply.

**4. Applications.** In this section we shall give two examples of this problem in practice. One is the tree-shaped network of 7-strings and the other is a network of  $N$ -connected strings. Both examples are put into the framework of (2.1), and we shall prove that both systems are exponentially stable.

**4.1. Tree-shaped network of 7-strings.** In this subsection we discuss the tree-shaped network of strings whose configuration is a simple, connected graph without closed paths.

In the past decade, the controllability and observability, as well as the stabilization, of networks of strings has been a hot topic. Many authors have obtained some nice results. For example, the authors in [5], [6], [7], and [8] discussed the problem of observability and controllability of a tree-shaped network of  $n$  strings, and the authors in [9] and [10] discussed the stabilization problem of star-shaped networks and generic trees of strings. For a more general discussion on networks of strings, and for a comprehensive list of papers on this subject, we refer the reader to the recently published book [11]. As an example, here we give the Riesz basis property and exponential stability of this system for  $n = 7$  with the shape shown in Figure 1.

First, we formulate the model into a normal form. We denote the strings by  $u_j$ ,  $j = 1, 2, \dots, 7$ . We normalize the strings to length one, and then they satisfy the equation

$$(4.1) \quad m_j u_{j,tt}(x, t) = T_j u_{j,xx}(x, t), \quad x \in (0, 1), \quad j = 1, 2, \dots, 7.$$

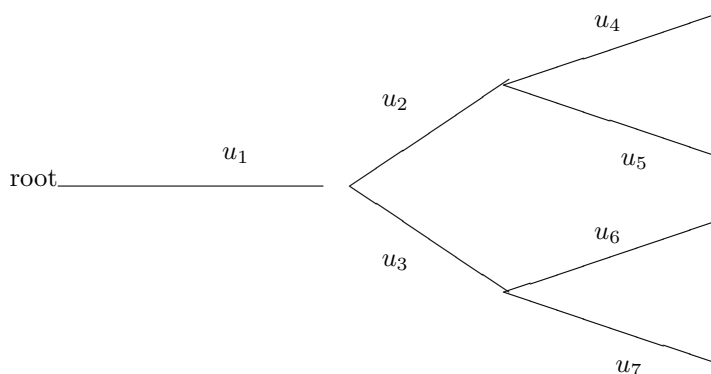


FIG. 1.

The root and the connected conditions are given by

$$(4.2) \quad \begin{aligned} u_1(0, t) &= 0, & u_1(1, t) &= u_2(0, t) = u_3(0, t), \\ u_2(1, t) &= u_4(0, t) = u_5(0, t), & u_3(1, t) &= u_6(0, t) = u_7(0, t), \end{aligned}$$

and the forced conditions at the nodes and edges are

$$(4.3) \quad \begin{aligned} T_1 u_{1,x}(1, t) - [T_2 u_{2,x}(0, t) + T_3 u_{3,x}(0, t)] &= -\alpha_1 u_{1,t}(1, t), \\ T_2 u_{2,x}(1, t) - [T_4 u_{4,x}(0, t) + T_5 u_{5,x}(0, t)] &= -\alpha_2 u_{2,t}(1, t), \\ T_3 u_{3,x}(1, t) - [T_6 u_{6,x}(0, t) + T_7 u_{7,x}(0, t)] &= -\alpha_3 u_{3,t}(1, t), \\ T_4 u_{4,x}(1, t) &= -\alpha_4 u_{4,t}(1, t), \\ T_5 u_{5,x}(1, t) &= -\alpha_5 u_{5,t}(1, t), \\ T_6 u_{6,x}(1, t) &= -\alpha_6 u_{6,t}(1, t), \\ T_7 u_{7,x}(1, t) &= -\alpha_7 u_{7,t}(1, t). \end{aligned}$$

Set

$$Y = [u_1(x, t), u_2(x, t), u_3(x, t), u_4(x, t), u_5(x, t), u_6(x, t), u_7(x, t)]^\tau$$

and denote

$$(4.4) \quad \begin{aligned} M &= \text{diag}[m_1, m_2, \dots, m_7], & T &= \text{diag}[T_1, T_2, \dots, T_7], \\ \Gamma &= \text{diag}[\alpha_1, \alpha_2, \dots, \alpha_7], \end{aligned}$$

and

$$(4.5) \quad C = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then (4.1)–(4.3) can be rewritten as

$$(4.6) \quad \begin{cases} M \frac{\partial^2 Y(x,t)}{\partial^2 t} = T \frac{\partial Y(x,t)}{\partial^2 x}, & x \in (0,1), \quad t > 0, \\ Y(0,t) = CY(1,t), & t > 0, \\ T \frac{\partial Y(1,t)}{\partial x} - C^\tau T \frac{\partial Y(0,t)}{\partial x} = -\Gamma \frac{\partial Y(1,t)}{\partial t}, & t > 0, \\ Y(x,0) = Y_0(x), \quad \frac{\partial Y(x,0)}{\partial t} = Y_1(x), \end{cases}$$

where  $Y_0(x)$  and  $Y_1(x)$  are given suitable initial value conditions.

According to Theorem 3.5, we have the following result.

THEOREM 4.1. *The system (4.6) is well-posed and asymptotically stable.*

To obtain exponential stability, we discuss the eigenvalue problem of the system.

Let  $\lambda \in \mathbb{C}$ ,  $Y(x,t) = e^{\lambda t} Y(x)$ ; then we have

$$(4.7) \quad \begin{cases} \lambda^2 MY(x) = TY_{xx}(x), & x \in (0,1), \\ Y(0) = CY(1), \\ TY_x(1) - C^\tau TY_x(0) = -\lambda \Gamma Y(1). \end{cases}$$

Let  $B^2 = \text{diag}[\rho_1^2, \rho_2^2, \dots, \rho_7^2]$ , where  $\rho_j^2 = \frac{m_j}{T_j}$ . Then we have

$$(4.8) \quad \begin{cases} \lambda^2 B^2 Y(x) = Y_{xx}(x), & x \in (0,1), \\ Y(0) = CY(1), \\ TY_x(1) - C^\tau TY_x(0) = -\lambda \Gamma Y(1). \end{cases}$$

Thus  $Y(x)$  has the following form:

$$Y(x) = \sinh(x\lambda B)u + \cosh(x\lambda B)v, \quad u, v \in \mathbb{C}^7.$$

Substituting it into the boundary conditions leads to

$$v = C[\sinh \lambda Bu + \cosh \lambda Bv],$$

$$TB[\cosh \lambda Bu + \sinh \lambda Bv] - C^\tau TBu = -\Gamma[\sinh \lambda Bu + \cosh \lambda Bv].$$

So  $\lambda$  is an eigenvalue if and only if

$$(4.9) \quad D(\lambda) = \det \begin{pmatrix} I - C \cosh \lambda B & -C \sinh \lambda B \\ TB \sinh \lambda B + \Gamma \cosh \lambda B & TB \cosh \lambda B + \Gamma \sinh \lambda B - C^\tau TB \end{pmatrix} = 0.$$

When  $\det[\Gamma - TB - C^\tau TBC] \neq 0$ , all eigenvalues are located in a strip. Note that

$$(4.10) \quad TB = \begin{pmatrix} T_1 \rho_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & T_2 \rho_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & T_3 \rho_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & T_4 \rho_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & T_5 \rho_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & T_6 \rho_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & T_7 \rho_7 \end{pmatrix},$$

$$(4.11) \quad C^\tau TBC = \begin{pmatrix} T_2 \rho_2 + T_3 \rho_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & T_4 \rho_4 + T_5 \rho_5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & T_6 \rho_6 + T_7 \rho_7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The condition  $\det[\Gamma - TB - C^T TBC] \neq 0$  means that

$$\begin{aligned} \alpha_1 &\neq T_1\rho_1 + T_2\rho_2 + T_3\rho_3, & \alpha_2 &\neq T_2\rho_2 + T_4\rho_4 + T_5\rho_5, \\ \alpha_3 &\neq T_3\rho_3 + T_6\rho_6 + T_7\rho_7, & \alpha_j &\neq T_j\rho_j, \quad j = 4, 5, 6, 7. \end{aligned}$$

Set

$$(4.12) \quad \begin{cases} w_j(\lambda) = T_j\rho_j \cosh \lambda\rho_j + \alpha_j \sinh \lambda\rho_j, \\ v_j(\lambda) = T_j\rho_j \sinh \lambda\rho_j + \alpha_j \cosh \lambda\rho_j, \end{cases} \quad j = 1, 2, 3, 4, 5, 6, 7,$$

$$(4.13) \quad \begin{cases} F_2(\lambda) = \frac{1}{T_2\rho_2} [(w_5v_4T_4\rho_4 + w_4v_5T_5\rho_5) \sinh \lambda\rho_2 + w_2w_4w_5], \\ G_2(\lambda) = \frac{1}{T_2\rho_2} [(w_5v_4T_4\rho_4 + w_4v_5T_5\rho_5) \cosh \lambda\rho_2 + v_2w_4w_5], \end{cases}$$

and

$$(4.14) \quad \begin{cases} F_3(\lambda) = \frac{1}{T_3\rho_3} [(w_7v_6T_6\rho_6 + w_6v_7T_7\rho_7) \sinh \lambda\rho_3 + w_3w_6w_7], \\ G_3(\lambda) = \frac{1}{T_3\rho_3} [(w_7v_6T_6\rho_6 + w_6v_7T_7\rho_7) \cosh \lambda\rho_3 + v_3w_6w_7]. \end{cases}$$

A complicated calculation shows that

$$(4.15) \quad \begin{aligned} D(\lambda) &= [T_2\rho_2G_2(\lambda)F_3(\lambda) + T_3\rho_3G_3(\lambda)F_2(\lambda)] \sinh \lambda\rho_1 \\ &\quad + w_1(\lambda)F_2(\lambda)F_3(\lambda). \end{aligned}$$

Now we are in a position to calculate  $\inf_{\sigma \in \mathbb{R}} |D(i\sigma)|$ . Note that

$$\inf_{\sigma \in \mathbb{R}} |w_j(i\sigma)| \neq 0, \quad \inf_{\sigma \in \mathbb{R}} |v_j(i\sigma)| \neq 0, \quad j = 1, 2, 3, 4, 5, 6, 7,$$

$$\frac{v_j(i\sigma)}{w_j(i\sigma)} = \frac{v_j(i\sigma)\overline{w_j(i\sigma)}}{|w_j(i\sigma)|^2} = \frac{\alpha_j T_j \rho_j + i \frac{T_j^2 \rho_j^2 - \alpha_j^2}{2} \sin 2\sigma \rho_j}{T_j^2 \rho_j^2 \cos^2 \sigma \rho_j + \alpha_j^2 \sin^2 \sigma \rho_j},$$

and

$$\frac{i \sin \sigma \rho_j}{w_j(i\sigma)} = \frac{i \sin \sigma \rho_j (T_j \rho_j \cos \sigma \rho_j - i \alpha_j \sin \sigma \rho_j)}{|w_j(i\sigma)|^2} = \frac{\alpha_j \sin^2 \sigma \rho_j + i \frac{T_j \rho_j}{2} \sin 2\sigma \rho_j}{T_j^2 \rho_j^2 \cos^2 \sigma \rho_j + \alpha_j^2 \sin^2 \sigma \rho_j},$$

where  $j = 1, 2, 3$ . From (4.13)–(4.14) we can get

$$\inf_{\sigma \in \mathbb{R}} |F_2(i\sigma)| > 0, \quad \inf_{\sigma \in \mathbb{R}} |F_3(i\sigma)| > 0,$$

$$\begin{aligned} \Re(G_2(i\sigma)\overline{F_2(i\sigma)}) &= [|w_5(i\sigma)|^2 \alpha_4 (T_4 \rho_4)^2 + |w_4(i\sigma)|^2 \alpha_5 (T_5 \rho_5)^2] T_2 \rho_2 \\ &\quad + \alpha_2 T_2 \rho_2 |w_4(i\sigma)|^2 |w_5(i\sigma)|^2 > 0, \end{aligned}$$

and

$$\begin{aligned} \Re(G_3(i\sigma)\overline{F_3(i\sigma)}) &= [|w_7(i\sigma)|^2 \alpha_6 (T_6 \rho_6)^2 + |w_6(i\sigma)|^2 \alpha_7 (T_7 \rho_7)^2] T_3 \rho_3 \\ &\quad + \alpha_3 T_3 \rho_3 |w_6(i\sigma)|^2 |w_7(i\sigma)|^2 > 0. \end{aligned}$$

Consequently, we have

$$\begin{aligned} \frac{D(i\sigma)}{w_1(i\sigma)F_2(i\sigma)F_3(i\sigma)} &= \left[ T_2\rho_2 \frac{G_2(i\sigma)\overline{F_2(i\sigma)}}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{G_3(i\sigma)\overline{F_3(i\sigma)}}{|F_3(i\sigma)|^2} \right] \frac{i \sin \sigma \rho_1 \overline{w_1(i\sigma)}}{|w_1(i\sigma)|^2} + 1 \\ &= \left[ T_2\rho_2 \frac{\Re(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Re(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \frac{\alpha_1 \sin^2 \sigma \rho_1}{|w_1(i\sigma)|^2} + 1 \\ &\quad - \left[ T_2\rho_2 \frac{\Im(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Im(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \frac{T_1\rho_1 \sin 2\sigma \rho_1}{2|w_1(i\sigma)|^2} \\ &\quad + i \left[ T_2\rho_2 \frac{\Re(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Re(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \frac{T_1\rho_1 \sin 2\sigma \rho_1}{2|w_1(i\sigma)|^2} \\ &\quad + i \left[ T_2\rho_2 \frac{\Im(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Im(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \frac{\alpha_1 \sin^2 \sigma \rho_1}{|w_1(i\sigma)|^2}. \end{aligned}$$

When  $\sin \sigma \rho_1 \rightarrow 0$ , we have  $D(i\sigma) \not\rightarrow 0$ . Now we assume that  $\sin \sigma \rho_1 \not\rightarrow 0$ , and then

$$\begin{aligned} &\left[ T_2\rho_2 \frac{\Re(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Re(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \frac{T_1\rho_1 \sin 2\sigma \rho_1}{2|w_1(i\sigma)|^2} \\ &+ \left[ T_2\rho_2 \frac{\Im(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Im(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \frac{\alpha_1 \sin^2 \sigma \rho_1}{|w_1(i\sigma)|^2} \rightarrow 0 \end{aligned}$$

if and only if

$$\begin{aligned} &- \left[ T_2\rho_2 \frac{\Im(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Im(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \\ &= \left[ T_2\rho_2 \frac{\Re(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Re(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \frac{T_1\rho_1 \cos \sigma \rho_1}{\alpha_1 \sin \sigma \rho_1} + o(1). \end{aligned}$$

From above we get that

$$\begin{aligned} \frac{D(i\sigma)}{w_1(i\sigma)F_2(i\sigma)F_3(i\sigma)} &= \left[ T_2\rho_2 \frac{\Re(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Re(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \frac{\alpha_1 \sin^2 \sigma \rho_1}{|w_1(i\sigma)|^2} + 1 \\ &\quad + \left[ T_2\rho_2 \frac{\Re(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Re(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \frac{T_1^2 \rho_1^2 \cos^2 \sigma \rho_1}{\alpha_1 |w_1(i\sigma)|^2} \\ &\quad + o(1) + io(1) \\ &= \left[ T_2\rho_2 \frac{\Re(G_2(i\sigma)\overline{F_2(i\sigma)})}{|F_2(i\sigma)|^2} + T_3\rho_3 \frac{\Re(G_3(i\sigma)\overline{F_3(i\sigma)})}{|F_3(i\sigma)|^2} \right] \frac{1}{\alpha_1} + 1 + o(1) + io(1). \end{aligned}$$

Therefore,

$$\inf_{\sigma \in \mathbb{R}} |D(i\sigma)| \neq 0.$$

According to Theorems 3.5 and 3.6, we have the following result.



**THEOREM 4.2.** *Let  $\mathcal{H}$  be defined as in section 3 and let  $\alpha_1 \neq T_1\rho_1 + T_2\rho_2 + T_3\rho_3$ ,  $\alpha_2 \neq T_2\rho_2 + T_4\rho_4 + T_5\rho_5$ ,  $\alpha_3 \neq T_3\rho_3 + T_6\rho_6 + T_7\rho_7$ ,  $\alpha_j \neq T_j\rho_j$ ,  $j = 4, 5, 6, 7$ . Then there is a sequence of eigenvectors and generalized eigenvectors of the system that forms a Riesz basis with parentheses for the state space  $\mathcal{H}$ . The system (4.6) is exponentially stable.*

As shown in Theorem 4.2,  $\Gamma$  being a positive definite matrix is only a sufficient condition for the system (4.1)–(4.3) to decay exponentially. In fact, if we suppose that the network of strings satisfies  $T_j = m_j = 1$ , then we can use three controllers to exponentially stabilize the network. The controllers are set up as follows:

- (1)  $\alpha_1 \neq 0$ ,  $\alpha_5\alpha_7 \neq 0$ ;
- (2)  $\alpha_1 = \alpha_3 = \alpha_5 = 0$ ; one of  $\alpha_j$ ,  $j = 4, 5, 6, 7$ , is 0.

In both cases, it always holds that  $\inf_{\sigma \in \mathbb{R}} |D(i\sigma)| > 0$ . Here we omit the details of calculation.

*Remark.* In this example, we discussed only tree-shaped networks of 7-strings (see Figure 1), the goal being to give an exact expression of the condition  $\det(\Gamma - TB - C^T TBC) \neq 0$ , and to calculate  $\inf_{\sigma \in \mathbb{R}} |D(i\sigma)|$ . Indeed, in the same manner we can discuss any tree-shaped network of strings. The matrix  $C$  is just the connective matrix, and the condition  $f(0) = Cf(1)$  is the condition of the displacement continuity. The difficulty we encounter is that the condition

$$\det[\Gamma - TB - C^T TBC] \neq 0$$

is not obvious, and there is not a general method of calculating the value  $\inf_{\sigma \in \mathbb{R}} |D(i\sigma)|$ .

**4.2.  $N$ -connected strings.** In this subsection we discuss the  $N$ -connected strings with internal nodes and boundary controls. This problem was proposed and discussed for  $N = 2$  in [2]. The exponential stability was investigated in [3]. Here, we give the exponential decay rate of the system and Riesz basis property. Let us recall the model.

Let  $y(x, t)$ , the transverse displacement of  $N$ -connected strings at location  $x$  at time  $t$ , satisfy

$$(4.16) \quad m_i \frac{\partial^2 y(x, t)}{\partial t^2} - T_i \frac{\partial^2 y(x, t)}{\partial x^2} = 0, \quad i-1 < x < i, \quad i = 1, 2, \dots, N, \quad t > 0.$$

We assume the Dirichlet condition at the left-hand  $x = 0$  and Neumann boundary condition at the right-hand  $x = N$ , where a control force  $u_N(t)$  is applied, i.e.,

$$(4.17) \quad y(0, t) = 0, \quad \frac{\partial y(N, t)}{\partial x} = u_N(t), \quad t > 0.$$

At the  $i$ th intermediate node  $x = i$ , we assume the continuity of displacement

$$(4.18) \quad y(i^+, t) = y(i^-, t), \quad i = 1, 2, \dots, N-1, \quad t > 0,$$

and discontinuity of vertical force component

$$(4.19) \quad T_i \frac{\partial y(i^-, t)}{\partial x} - T_{i+1} \frac{\partial y(i^+, t)}{\partial x} = u_i(t), \quad i = 1, 2, \dots, N-1, \quad t > 0,$$

where  $u_j(t)$ ,  $j = 1, 2, \dots, N-1$ , are applied external forces.

An important task in engineering is to design controllers  $U = (u_1(t), u_2(t), \dots, u_N(t))$  such that the system comes back to its equilibrium. The authors in [2] designed the

following feedback controllers at the intermediated point  $x = i$  and the right-hand  $x = N$ :

$$(4.20) \quad u_i(t) = -\alpha_i \frac{\partial y(i, t)}{\partial t}, \quad \alpha_i > 0, \quad i = 1, 2, \dots, N.$$

Then (4.16), together with (4.17)–(4.20), forms a closed loop system. In what follows, we shall prove that this closed loop system is a Riesz system and decays exponentially.

Let  $y_i(x, t) = y(i - 1 + x, t)$ ,  $x \in (0, 1)$ , and

$$Y(x, t) = (y_1(x, t), y_2(x, t), \dots, y_N(x, t)), \quad x \in (0, 1), \quad t > 0.$$

Then (4.16) is equivalent to an equation in  $\mathbb{C}^N$ ,

$$(4.21) \quad M \frac{\partial^2 Y(x, t)}{\partial t^2} = T \frac{\partial^2 Y(x, t)}{\partial x^2}, \quad x \in (0, 1), \quad t > 0,$$

where

$$(4.22) \quad M = \text{diag}(m_1, m_2, \dots, m_N), \quad T = \text{diag}(T_1, T_2, \dots, T_N).$$

The continuity conditions at intermediate nodes, together with the condition at the left-hand endpoint, can be written into

$$(4.23) \quad Y(0, t) = CY(1, t),$$

where

$$(4.24) \quad C = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & & \ddots & & & \ddots & & \vdots \\ \vdots & & \ddots & & & \ddots & & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

The discontinuity condition of vertical force at the intermediate nodes, together with the condition at the right-hand endpoint, can be written as

$$(4.25) \quad T \frac{\partial Y(1, t)}{\partial x} - C^T T \frac{\partial Y(0, t)}{\partial x} = U(t) = -\Gamma \frac{\partial Y(1, t)}{\partial t},$$

where  $C^T$  denotes the conjugate transpose matrix of  $C$ , and

$$(4.26) \quad \Gamma = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_N)$$

is an  $N \times N$  positive definite matrix. Thus the closed loop system can be written as

$$(4.27) \quad \begin{cases} M \frac{\partial^2 Y(x, t)}{\partial t^2} = T \frac{\partial^2 Y(x, t)}{\partial x^2}, & x \in (0, 1), \quad t > 0, \\ Y(0, t) = CY(1, t), & t > 0, \\ T \frac{\partial Y(1, t)}{\partial x} - C^T T \frac{\partial Y(0, t)}{\partial x} = -\Gamma \frac{\partial Y(1, t)}{\partial t}, & t > 0, \\ Y(x, 0) = Y_0(x), \quad \frac{\partial Y(x, 0)}{\partial t} = Y_1(x), & x \in (0, 1), \end{cases}$$

where  $Y_0(x)$  and  $Y_1(x)$  are given suitable initial data.

In this case, we can take the matrix

$$B^2 = \text{diag} \left[ \frac{m_1}{T_1}, \frac{m_2}{T_2}, \dots, \frac{m_N}{T_N} \right],$$

and then condition (3.5) is equivalent to

$$D(\lambda) = \det \begin{pmatrix} (I - C)e^{\lambda B} & (I - C)e^{-\lambda B} \\ (\Gamma + TB)e^{\lambda B} - C^T TB & (\Gamma + TB)e^{-\lambda B} + C^T TB \end{pmatrix},$$

and then condition  $\det[\widehat{\Gamma} - B - \widehat{C}^T B \widehat{C}] \neq 0$  is equivalent to  $\det[\Gamma - TB - C^T TBC] \neq 0$ .

Now let us determine  $D(\lambda)$ . Let  $\lambda \in \mathbb{C}$  be an eigenvalue, and let  $Y = (y_1, y_2, \dots, y_N)$  be a corresponding eigenfunction. Then

$$(4.28) \quad \begin{aligned} m_j \lambda^2 y_j(x) &= T_j y_{j,xx}, \\ y_1(0) &= 0, \quad y_j(1) = y_{j+1}(0), \quad j = 1, 2, \dots, N. \end{aligned}$$

$$(4.29) \quad \begin{aligned} T_j y_{j,x}(1) - T_{j+1} y_{j+1,x}(0) &= -\alpha_j \lambda y_j(1), \quad j = 1, 2, \dots, N-1, \\ T_N y_{N,x}(1) &= -\alpha_N \lambda y_N(1). \end{aligned}$$

Set  $\rho_j^2 = \frac{m_j}{T_j}$  and

$$y_1(x) = a_1 \sinh \lambda \rho_1, \quad y_j(x) = a_j \sinh \lambda \rho_j + b_j \cosh \lambda \rho_j, \quad j = 2, 3, \dots, N.$$

From (4.28) and (4.29) we get

$$(4.30) \quad \begin{cases} [T_N \rho_N \cosh \lambda \rho_N + \alpha_N \sinh \lambda \rho_N] a_N + [T_N \rho_N \sinh \lambda \rho_N + \alpha_N \cosh \lambda \rho_N] b_N = 0, \\ [T_j \rho_j \cosh \lambda \rho_j + \alpha_j \sinh \lambda \rho_j] a_j + [T_j \rho_j \sinh \lambda \rho_j + \alpha_j \cosh \lambda \rho_j] b_j = T_{j+1} \rho_{j+1} a_{j+1}, \\ \quad j = 2, \dots, N-1, \\ \sinh \lambda \rho_j a_j + \cosh \lambda \rho_j b_j = b_{j+1}, \quad j = 2, \dots, N-1, \\ [T_1 \rho_1 \cosh \lambda \rho_1 + \alpha_1 \sinh \lambda \rho_1] a_1 = T_2 \rho_2 a_2, \\ [\sinh \lambda \rho_1] a_1 = b_2. \end{cases}$$

Set  $T_{N+1} \rho_{N+1} = 1$  and

$$(4.31) \quad \begin{aligned} w_j(\lambda) &= \frac{1}{T_{j+1} \rho_{j+1}} [T_j \rho_j \cosh \lambda \rho_j + \alpha_j \sinh \lambda \rho_j], \\ v_j(\lambda) &= \frac{1}{T_{j+1} \rho_{j+1}} [T_j \rho_j \sinh \lambda \rho_j + \alpha_j \cosh \lambda \rho_j]. \end{aligned}$$

We can rewrite (4.30) as

$$(4.32) \quad \begin{aligned} (1, 0) \begin{pmatrix} w_N(\lambda) & v_N(\lambda) \\ \sinh \lambda \rho_N & \cosh \lambda \rho_N \end{pmatrix} \begin{pmatrix} a_N \\ b_N \end{pmatrix} &= 0, \\ \begin{pmatrix} w_j(\lambda) & v_j(\lambda) \\ \sinh \lambda \rho_j & \cosh \lambda \rho_j \end{pmatrix} \begin{pmatrix} a_j \\ b_j \end{pmatrix} &= \begin{pmatrix} a_{j+1} \\ b_{j+1} \end{pmatrix}, \\ \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} &= \begin{pmatrix} w_1(\lambda) & v_1(\lambda) \\ \sinh \lambda \rho_1 & \cosh \lambda \rho_1 \end{pmatrix} \begin{pmatrix} a_1 \\ 0 \end{pmatrix}. \end{aligned}$$

Therefore, for  $1 \leq k \leq N-1$ , we have

$$(4.33) \quad \begin{pmatrix} a_{k+1} \\ b_{k+1} \end{pmatrix} = \left[ \prod_{j=0}^{k-1} \begin{pmatrix} w_{k-j}(\lambda) & v_{k-j}(\lambda) \\ \sinh \lambda \rho_{k-j} & \cosh \lambda \rho_{k-j} \end{pmatrix} \right] \begin{pmatrix} a_1 \\ 0 \end{pmatrix}.$$

Note that, for  $j = 1, 2, \dots, N$ , the matrices

$$\begin{pmatrix} w_j(\lambda) & v_j(\lambda) \\ \sinh \lambda \rho_j & \cosh \lambda \rho_j \end{pmatrix} = \begin{pmatrix} \frac{T_j \rho_j}{T_{j+1} \rho_{j+1}} & \frac{\alpha_j}{T_{j+1} \rho_{j+1}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cosh \lambda \rho_j & \sinh \lambda \rho_j \\ \sinh \lambda \rho_j & \cosh \lambda \rho_j \end{pmatrix}$$

are invertible. So  $\lambda \in \mathbb{C}$  is an eigenvalue if and only if

$$(4.34) \quad D(\lambda) = (1, 0) \left[ \prod_{j=0}^{N-1} \begin{pmatrix} w_{N-j}(\lambda) & v_{N-j}(\lambda) \\ \sinh \lambda \rho_{N-j} & \cosh \lambda \rho_{N-j} \end{pmatrix} \right] \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 0$$

or

$$(4.35) \quad D(\lambda) = D(\lambda)^\tau = (1, 0) \left[ \prod_{j=1}^N \begin{pmatrix} w_j(\lambda) & \sinh \lambda \rho_j \\ v_j(\lambda) & \cosh \lambda \rho_j \end{pmatrix} \right] \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 0.$$

As before, a complicated calculation shows that

$$\inf_{\sigma \in \mathbb{R}} |D(i\sigma)| > 0.$$

The condition  $\det[\Gamma - TB - C^\tau TBC] \neq 0$  becomes

$$(4.36) \quad \alpha_j \neq T_j \rho_j + T_{j+1} \rho_{j+1}, \quad j = 1, 2, \dots, N-1, \quad \alpha_N \neq T_N \rho_N.$$

Combining this with Theorem 3.5, we have the following result.

**THEOREM 4.3.** *Let (4.36) hold. Then system (4.27) is exponentially stable. There is a sequence of eigenvectors and generalized eigenvectors of the system that forms a Riesz basis with parentheses for the space  $\mathcal{H}$  defined as in section 3.*

**Acknowledgments.** The authors would like to thank the referees for their helpful comments and suggestions.

## REFERENCES

- [1] J. L. LIONS, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.
- [2] G. CHEN, M. COLEMAN, AND H. H. WEST, *Pointwise stabilization in the middle of the span for second order systems, nonuniform and uniform exponential decay of solutions*, SIAM J. Appl. Math., 47 (1987), pp. 751–780.
- [3] K.-S. LIU, F.-L. HUANG, AND G. CHEN, *Exponential stability analysis of a long chain of coupled vibrating strings with dissipative linkage*, SIAM J. Appl. Math., 49 (1989), pp. 1694–1707.
- [4] S. COX AND E. ZUAZUA, *The rate at which energy decays in a damped string*, Comm. Partial Differential Equations, 19 (1994), pp. 213–243.
- [5] R. DÁGER, *Observation and control of vibrations in tree-shaped networks of strings*, SIAM J. Control Optim., 43 (2004), pp. 590–623.
- [6] G. LEUGERING AND E. ZUAZUA, *On exact controllability of generic trees*, in Proceedings of the Control of Systems Governed by Partial Differential Equations (Nancy, France), ESAIM Proc. 8, Soc. Math. Appl. Indust., Paris, 2000, pp. 95–105 (electronic).

- [7] R. DAGER AND E. ZUAZUA, *Controllability of star-shaped networks of strings*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 621–626.
- [8] R. DAGER AND E. ZUAZUA, *Controllability of tree-shaped networks of vibrating strings*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 1087–1092.
- [9] K. AMMARI AND M. JELLOULI, *Stabilization of star-shaped tree of elastic strings*, Differential Integral Equations, 17 (2004), pp. 1395–1410.
- [10] K. AMMARI, M. JELLOULI, AND M. KHENISSI, *Stabilization of generic trees of strings*, J. Dyn. Control Syst., 11 (2005), pp. 177–193.
- [11] R. DAGER AND E. ZUAZUA, *Wave Propagation, Observation, and Control in 1-d Flexible Multi-structures*, Math. Appl. (Berlin) 50, Springer-Verlag, Berlin, 2006.
- [12] J. U. KIM AND Y. RENARDY, *Boundary control of the Timoshenko beam*, SIAM J. Control Optim., 25 (1987), pp. 1417–1429.
- [13] Ö. MORGÚL, *Boundary control of a Timoshenko beam attached to a rigid body: Planar motion*, Internat. J. Control, 54 (1991), pp. 763–791.
- [14] G. Q. XU AND D. X. FENG, *Riesz basis property of a Timoshenko beam with boundary feedback and application*, IMA J. Appl. Math., 67 (2002), pp. 357–370.
- [15] R. QUINTANILLA, *Exponential stability for a one-dimensional problem of swelling porous elastic soils with fluid saturation*, J. Comput. Appl. Math., 145 (2002), pp. 525–533.
- [16] R. QUINTANILLA, *Slow decay for one-dimensional porous dissipation elasticity*, Appl. Math. Lett., 16 (2003), pp. 487–491.
- [17] Y. DU AND G. Q. XU, *Exponential stability of a system of linear Timoshenko type with boundary controls*, J. Systems Sci. Math. Sci., to appear (in Chinese).
- [18] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, 1983.
- [19] G.-Q. XU AND B.-Z. GUO, *Riesz basis property of evolution equations in Hilbert spaces and application to a coupled string equation*, SIAM J. Control Optim., 42 (2003), pp. 966–984.
- [20] G. Q. XU, D. X. FENG, AND S. P. YUNG, *Riesz basis property of the generalized eigenvector system of a Timoshenko beam*, IMA J. Math. Control Inform., 21 (2004), pp. 65–83.
- [21] YU. I. LYUBICH AND V. Q. PHÓNG, *Asymptotic stability of linear differential equations in Banach spaces*, Studia Math., 88 (1988), pp. 34–37.
- [22] R. M. YOUNG, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.
- [23] S. A. AVDONIN AND S. A. IVANOV, *Families of Exponentials. The Method of Moments in Controllability Problems for Distributed Parameter Systems*, Cambridge University Press, Cambridge, UK, 1995.
- [24] G. Q. XU AND S. P. YUNG, *The expansion of semigroup and criterion of Riesz basis*, J. Differential Equations, 210 (2005), pp. 1–24.
- [25] G. Q. XU, Z. J. HAN, AND S. P. YUNG, *Riesz basis property of serially connected Timoshenko beams*, Internat. J. Control, 80 (2007), pp. 470–485.
- [26] I. C. GOHBERG AND M. G. KREIN, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr. 18, AMS, Providence, RI, 1969.

## OPTIMAL REFLECTION OF DIFFUSIONS AND BARRIER OPTIONS PRICING UNDER CONSTRAINTS\*

BRUNO BOUCHARD†

**Abstract.** We introduce a new class of control problems in which the gain depends on the solution of a stochastic differential equation (SDE) reflected at the boundary of a bounded domain, along directions which are controlled by a bounded variation process. We provide a PDE characterization of the associated value function. This study is motivated by applications in mathematical finance, where such equations are related to the pricing of barrier options under portfolio constraints.

**Key words.** reflected diffusion, Skorokhod problem, viscosity solutions, barrier option, portfolio constraints

**AMS subject classifications.** 93E20, 60G99, 49L25, 91B28

**DOI.** 10.1137/070697161

**1. Introduction.** This paper is motivated by a previous work [2], where a new class of parabolic PDE with Neumann and Dirichlet conditions is introduced. The starting point of [2] is the problem of hedging a barrier option under portfolio constraints. It shows that the superhedging price is a viscosity solution of an equation of the form

$$(1.1) \quad \begin{cases} \min \left\{ -\mathcal{L}\varphi, \min_{e \in E} \mathcal{H}^e \varphi \right\} = 0 & \text{on } [0, T) \times \mathcal{O}, \\ \min \left\{ \varphi, \min_{e \in E} \mathcal{H}^e \varphi \right\} = 0 & \text{on } [0, T) \times \partial\mathcal{O}, \\ \varphi - \hat{g} = 0 & \text{on } \{T\} \times \bar{\mathcal{O}}. \end{cases}$$

Here,  $\mathcal{O}$  is an open domain of  $\mathbb{R}^d$  outside of which the option is desactivated,  $E$  is a compact subset of  $\mathbb{R}^\ell$  which depends on the constraints imposed on the portfolio,  $\mathcal{L}\varphi = \frac{\partial}{\partial t}\varphi + \frac{1}{2}\text{Tr}\sigma\sigma^*D^2\varphi$  is the Dynkin operator of the diffusion which models the evolution of the risky assets,  $\mathcal{H}^e\varphi := \delta(\cdot, e)\varphi - \langle \gamma(\cdot, e), D\varphi \rangle$  for some (oblique) inward direction  $\gamma(x, e)$ , and  $\hat{g}$  is a “smoothed” version of the payoff of the option which satisfies  $\min_{e \in E} \mathcal{H}^e \hat{g} \geq 0$  (see [2] for details and section 4 below for an example).

When the solution  $\varphi$  of the above equation is positive, the spatial boundary condition reduces to  $\min_{e \in E} \mathcal{H}^e \varphi = 0$  on  $[0, T) \times \partial\mathcal{O}$ , and, in particular cases (see [14] and [15]), the constraint  $\mathcal{H}^e \varphi \geq 0$  on the parabolic boundary of  $[0, T) \times \mathcal{O}$  propagates in the domain, which allows us to simplify the above equation in

$$(1.2) \quad \begin{cases} -\mathcal{L}\varphi = 0 & \text{on } [0, T) \times \mathcal{O}, \\ \min_{e \in E} \mathcal{H}^e \varphi = 0 & \text{on } [0, T) \times \partial\mathcal{O}, \\ \varphi - \hat{g} = 0 & \text{on } \{T\} \times \bar{\mathcal{O}}. \end{cases}$$

\*Received by the editors July 13, 2007; accepted for publication (in revised form) January 29, 2008; published electronically June 26, 2008.

<http://www.siam.org/journals/sicon/47-4/69716.html>

†Ceremade and Crest, Université Paris-Dauphine, bureau B514, place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France (bouchard@ceremade.dauphine.fr).

When  $E$  is a singleton  $\{e_0\}$ , such equations formally admit a Feynman–Kac representation of the form

$$(1.3) \quad \mathbb{E} \left[ e^{-\int_t^T \delta(X(s), e_0) dL(s)} \hat{g}(X(T)) \right],$$

where  $L$  is a nondecreasing process such that  $(X, L)$  solves on  $[t, T]$

$$(1.4) \quad \begin{aligned} X(s) &= x + \int_t^s \sigma(X(r)) dW(r) + \int_t^s \gamma(X(r), e_0) dL(r), \\ X(s) &\in \bar{\mathcal{O}} \text{ and } L(s) = \int_t^s \mathbb{I}_{\{X(r) \in \partial \mathcal{O}\}} dL(r), \quad t \leq s \leq T, \end{aligned}$$

for a given standard Brownian motion  $W$ . Thus, in this particular case, the price of the barrier option is, at least formally, given by the expectation of a functional depending on the solution of a stochastic differential equation (SDE) which is reflected at the boundary of  $\mathcal{O}$  along the direction  $\gamma(\cdot, e_0)$ . This phenomenon was already observed in [14] in a particular setting and can be easily explained when  $\hat{g} \geq 0$  and  $\hat{g}$  is nondecreasing on  $\mathcal{O}$ ; see Remark 4.4 below.

By analogy, (1.2) should be associated with a control problem of the form

$$(1.5) \quad \sup_{\epsilon \in \mathcal{E}} \mathbb{E} \left[ e^{-\int_t^T \delta(X^\epsilon(s), \epsilon(s)) dL^\epsilon(s)} \hat{g}(X^\epsilon(T)) \right],$$

where  $(X^\epsilon, L^\epsilon)$  is the solution on  $[t, T]$  of

$$(1.6) \quad \begin{aligned} X^\epsilon(s) &= x + \int_t^s \sigma(X^\epsilon(r)) dW(r) + \int_t^s \gamma(X^\epsilon(r), \epsilon(r)) dL^\epsilon(r), \\ X^\epsilon(s) &\in \bar{\mathcal{O}} \text{ and } L^\epsilon(s) = \int_t^s \mathbb{I}_{\{X^\epsilon(r) \in \partial \mathcal{O}\}} dL^\epsilon(r), \quad t \leq s \leq T, \end{aligned}$$

and  $\mathcal{E}$  is a suitable set of adapted processes with values in  $E$ . The difference with (1.3) is that the direction of reflection is now controlled by the process  $\epsilon \in \mathcal{E}$ .

This naturally leads to the introduction of a new class of control problems of the form (1.5), which, to the best of our knowledge, have not been studied so far.

In this paper, we first show that (1.6) admits a strong solution in the case where  $\mathcal{O}$  is bounded,  $|\gamma| = 1$ , and  $(\mathcal{O}, \gamma)$  satisfies the following uniform exterior cone condition:

$$(1.7) \quad \bigcup_{0 \leq \lambda \leq r} B(x - \lambda \gamma(x, e), \lambda r) \subset \mathcal{O}^c \quad \text{for all } (x, e) \in \partial \mathcal{O} \times \mathbb{R}^\ell.$$

There is a huge literature on reflected SDEs and we refer to [7] for an overview of main results. In the case where  $(X, \epsilon)$  is the solution of an SDE with Lipschitz coefficients, the existence of a strong solution under the exterior cone condition (1.7) is easily deduced from [6]. Indeed, it suffices to consider the extended system  $(X, \epsilon)$  reflected at the boundary of  $\mathcal{O} \times \tilde{E}$  for some open ball  $\tilde{E} = B(0, \tilde{r})$ , which contains the compact set  $E$  along a smooth direction  $\tilde{\gamma}$  such that  $\tilde{\gamma} = (\gamma, 0)$  on  $\mathcal{O} \times E$  and  $\tilde{\gamma} = (\gamma, -e/\tilde{r})/\sqrt{2}$  on  $\mathcal{O} \times \partial \tilde{E}$ . This system satisfies the exterior cone condition of [6]. Since  $\epsilon$  takes values in  $E$ , the reflection does not operate on this component and we deduce the existence of a solution to (1.6) from the results of [6]. However, this formulation is quite restrictive and we are interested in a more general class of controls.

We therefore come back to the initial deterministic Skorokhod problem and follow the steps of [6], which are inspired by [12]. The existence of the Skorokhod problem with directions of reflection controlled by a continuous function  $\epsilon$  with bounded variations is deduced from [6] by using the above arguments, which consists of considering an extended system. We then use suitable estimates on a family of test functions introduced in [5] to prove the existence of a solution to (1.6) in our general setting. Moreover, by considering SDEs with random coefficients, we are able to incorporate another control on the direction which takes the form of an Itô process; see section 2.

We then introduce a control problem which generalizes (1.5) and prove that its value function is a viscosity solution of an equation of the form (1.2), for which we provide a comparison result. In the case where  $\gamma(x, e)$  does not depend on  $e$ , it essentially follows from the results of [5]. In this paper, we propose a new set of conditions which is more adapted to our setting and does not seem to be covered in the existing literature; see section 3.4 below.

In the last section, we discuss the link between (1.5) and the pricing of barrier options under portfolio constraints. In a particular setting, we prove that (1.5) coincides with the superhedging price of the option, when (1.2) admits a sufficiently smooth solution. This generalizes previous results of [14]. When  $E$  is reduced to a singleton, this leads to a natural Monte-Carlo approach for its estimation. We leave the discussion of more general cases for future research.

Also, we should point out that an extension of the results of [14] to American options was carried out by [11] within a similar model as in [14], but with infinite time horizon. In mathematical terms, it essentially corresponds to mixing the problem (1.5) with an optimal stopping one. We think that our approach could be used in this setting too. We also leave this for future research.

**Notation.** Given  $E \subset \mathbb{R}^m$ ,  $m \geq 1$ , and  $E_i \subset \mathbb{R}^{m_i}$ ,  $m_i \geq 1$ , for  $i \leq I$ , we denote by  $C^{k_1, \dots, k_I}(E_1 \times \dots \times E_I, E)$  (resp.,  $C_b^{k_1, \dots, k_I}(E_1 \times \dots \times E_I, E)$ ) the set of continuous maps  $\varphi$  from  $E_1 \times \dots \times E_I$  into  $E$  that admit continuous (resp., bounded) derivatives up to order  $k_i$  in their  $i$ th component  $x_i$ . We omit  $k_i$  when it is equal to 0 and only write  $C^{k_1}(E_1 \times \dots \times E_I, E)$  when  $k_1 = k_2 = \dots = k_I$ . We omit  $E$  when  $E = \mathbb{R}$ , and, in this case, we denote by  $D_{x_i}\varphi$  and  $D_{x_i}^2\varphi$  the (partial) Jacobian and Hessian matrices with respect to  $x_i$ . We simply write  $D\varphi$  and  $D^2\varphi$  for  $D_{x_2}\varphi$  and  $D_{x_2}^2\varphi$  if  $I = 2$ . For  $T > 0$ , we define  $BV([0, T], E)$  as the set of continuous maps from  $[0, T]$  into  $E$  with a bounded total variation. For  $\epsilon \in BV([0, T], E)$ , we set  $|\epsilon| := \sum_{i \leq m} |\epsilon^i|$ , where  $|\epsilon^i|(t)$  is the total variation of  $\epsilon^i$  on  $[0, t]$ ,  $t \geq 0$ . We write  $E^c$  for  $\mathbb{R}^m \setminus E$ ,  $\partial E$  and  $\bar{E}$  denote the boundary and the closure of  $E$ ,  $\mathbb{R}_+^m = [0, \infty)^m$ , and  $\mathbb{R}_-^m = -\mathbb{R}_+^m$ . The Euclidean norm of  $x = (x^1, \dots, x^m) \in \mathbb{R}^m$  is denoted by  $|x|$ ,  $B(x, r)$  is the open ball centered on  $x$  with radius  $r$ , and  $\langle \cdot, \cdot \rangle$  is the natural scalar product on  $\mathbb{R}^m$ . We denote by  $\mathbb{M}^m$  the set of square matrices of dimension  $m$  and we extend the definition of  $|\cdot|$  with  $\mathbb{M}^m$  by identifying  $\mathbb{M}^m$  with  $\mathbb{R}^{m \times m}$ . For  $x \in \mathbb{R}^m$ ,  $\text{diag } x$  is the diagonal matrix of  $\mathbb{M}^m$  whose  $i$ th diagonal element is  $x^i$ ,  $\text{Tr } M$  is the trace of  $M \in \mathbb{M}^m$ , and  $M^*$  its transposition. All inequalities between random variables have to be taken in the a.s. sense.

**2. SDEs with controlled reflecting directions.** The goal of this section is to construct an SDE which is reflected at the boundary of some bounded open set  $\mathcal{O} \subset \mathbb{R}^d$ ,  $d \geq 1$ , along a direction which is controlled by an adapted continuous process with bounded variations taking values in a compact subset  $E$  of  $\mathbb{R}^\ell$ ,  $\ell \geq 1$ . We follow the arguments of [6] and start with the resolution of the associated (deterministic) Skorokhod problem.



**2.1. The Skorokhod problem with controlled reflecting directions.** For the sake of completeness, we first recall one of the main results of [6] which provides a solution to the Skorokhod problem for oblique reflection on general bounded sets.

**THEOREM 2.1** (Dupuis and Ishii [6]). *Fix  $\gamma \in C^2(\mathbb{R}^d, \mathbb{R}^d)$  with  $|\gamma| = 1$ . Assume that there exists some  $r \in (0, 1)$  such that*

$$(2.1) \quad \bigcup_{0 \leq \lambda \leq r} B(x - \lambda\gamma(x), \lambda r) \subset \mathcal{O}^c \quad \text{for all } x \in \partial\mathcal{O}.$$

*Then, for all  $\psi \in C([0, T], \mathbb{R}^d)$  satisfying  $\psi(0) \in \bar{\mathcal{O}}$ , there exists  $(\phi, \eta) \in C([0, T], \bar{\mathcal{O}}) \times \text{BV}([0, T], \mathbb{R}_+)$  such that  $\eta$  is nondecreasing and*

$$\phi(t) = \psi(t) + \int_0^t \gamma(\phi(s)) d\eta(s), \quad \eta(t) = \int_0^t \mathbb{I}_{\{\phi(s) \in \partial\mathcal{O}\}} d\eta(s), \quad t \leq T.$$

*Moreover,  $(\phi(t), \eta(t)) \in \sigma(\psi(s), s \leq t)$  for all  $t \leq T$ , and uniqueness holds if  $\psi \in \text{BV}([0, T], \mathbb{R}^d)$ .*

*Proof.* See Theorem 4.8 and the discussion after Corollary 5.2 in [6].  $\square$

We now fix an open bounded set  $\mathcal{O} \subset \mathbb{R}^d$ , a compact set  $E \subset \mathbb{R}^\ell$ , and  $\gamma$  satisfying

$$(2.2) \quad \gamma \in C^2(\mathbb{R}^{d+\ell}, \mathbb{R}^d), \quad |\gamma| = 1,$$

$$(2.3) \quad \exists r \in (0, 1) \text{ s.t. } \bigcup_{0 \leq \lambda \leq r} B(x - \lambda\gamma(x, e), \lambda r) \subset \mathcal{O}^c \quad \text{for all } (x, e) \in \partial\mathcal{O} \times \mathbb{R}^\ell.$$

We then deduce from Theorem 2.1 the following result.

**COROLLARY 2.1.** *Let the conditions (2.2) and (2.3) hold. Then, for all  $\psi \in \text{BV}([0, T], \mathbb{R}^d)$  satisfying  $\psi(0) \in \bar{\mathcal{O}}$  and  $\epsilon \in \text{BV}([0, T], E)$ , there exists a unique pair  $(\phi, \eta) \in C([0, T], \bar{\mathcal{O}}) \times \text{BV}([0, T], \mathbb{R}_+)$  such that  $\eta$  is nondecreasing and*

$$(2.4) \quad \phi(t) = \psi(t) + \int_0^t \gamma(\phi(s), \epsilon(s)) d\eta(s) \quad \text{and} \quad \eta(t) = \int_0^t \mathbb{I}_{\{\phi(s) \in \partial\mathcal{O}\}} d\eta(s), \quad t \leq T.$$

*Moreover,  $(\phi(t), \eta(t)) \in \sigma((\psi(s), \epsilon(s)), s \leq t)$  for all  $t \leq T$ .*

*Proof.* This is an immediate consequence of Theorem 2.1. Since  $\epsilon$  is valued in a compact set, it suffices to apply the above result to an extended fictitious reflected system  $(\psi, \epsilon)$ . We detail the proof for completeness. Fix  $\tilde{r} > 0$  so that  $\tilde{E} := B(0, \tilde{r})$  strictly contains  $E$ . Fix  $\zeta \in C^2(\mathbb{R}^\ell, [0, 1])$  such that  $\zeta(e) = 0$  for  $e \in E$  and  $\zeta(e) = 1$  for  $e \in \partial\tilde{E}$  and set, on  $\mathbb{R}^{d+\ell}$ ,  $\tilde{\gamma}(x, e) = (\gamma(x, e), -e\zeta(e)/\tilde{r})/|(\gamma(x, e), -e\zeta(e)/\tilde{r})|$ . Since  $|\gamma| = 1$ ,  $|(\gamma(x, e), -e\zeta(e)/\tilde{r})| \geq 1$  and  $\tilde{\gamma} \in C^2(\mathbb{R}^{d+\ell}, \mathbb{R}^{d+\ell})$ . Moreover,  $|(\gamma(x, e), -e\zeta(e)/\tilde{r})|^2 \leq 2$  on the closure of  $\mathcal{O} \times \tilde{E}$ ,  $|(\gamma(x, e), -e\zeta(e)/\tilde{r})|^2 = 2$  if  $e \in \partial\tilde{E}$ , and  $B(e + \lambda e/\tilde{r}, \lambda r) \cap \tilde{E} = \emptyset$  for all  $e \in \partial\tilde{E}$  and  $\lambda > 0$ ; recall that  $r < 1$ . We then deduce from (2.3) that for  $(x, e) \in \partial(\mathcal{O} \times \tilde{E})$  and  $\lambda \in [0, r/\sqrt{2}]$ ,

$$|(y, f) - ((x, e) - \lambda\tilde{\gamma}(x, e))|^2 \leq \lambda^2(r/\sqrt{2})^2 \Rightarrow (y, f) \notin \mathcal{O} \times \tilde{E}.$$

We can therefore apply Theorem 2.1 to the pair  $(\psi, \epsilon)$  reflected at the boundary of  $\mathcal{O} \times \tilde{E}$ . Since  $\epsilon$  does not reach the boundary of  $\tilde{E}$ , this leads to the required result.  $\square$

**2.2. The stochastic Skorokhod problem with controlled reflecting direction.** We now consider some probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  supporting a  $d$ -dimensional standard Brownian motion  $W$ . We denote by  $\mathbb{F} = (\mathcal{F}_t)_{t \leq T}$  the natural filtration induced by  $W$ , satisfying the usual conditions, and assume that  $\mathcal{F} = \mathcal{F}_T$ . Given two uniformly Lipschitz functions  $\mu$  and  $\sigma$  from  $\mathbb{R}^d$  into  $\mathbb{R}^d$  and  $\mathbb{M}^d$ , respectively, it is shown in [6] that, under the condition (2.1), there exists a unique pair  $(X, L)$  of  $\mathbb{F}$ -adapted continuous processes such that  $L$  is real valued, nondecreasing and

$$(2.5) \quad X(t) = x + \int_0^t \mu(X(s))ds + \int_0^t \sigma(X(s))dW(s) + \int_0^t \gamma(X(s))dL(s),$$

$$X(t) \in \bar{\mathcal{O}} \text{ and } L(t) = \int_0^t \mathbb{I}_{\{X(s) \in \partial \mathcal{O}\}} dL(s), \quad t \leq T.$$

The aim of this section is to extend this result to the case where  $\mu$  and  $\sigma$  are random and  $\gamma$  is controlled by some continuous bounded variation process  $\epsilon$  taking values in the compact set  $E$ . We refer to Remarks 2.2 and 2.3 below for comments on this a priori strong regularity assumption on the control  $\epsilon$ .

In the following, given two subsets  $E_1$  and  $E_2$  of  $\mathbb{R}^{m_1}$  and  $\mathbb{R}^{m_2}$ ,  $m_1, m_2 \geq 1$ , we denote by  $L_{\mathbb{F}}(E_1, E_2)$  the set of measurable maps

$$f : (\omega, t, x) \in \Omega \times [0, T] \times E_1 \longrightarrow f_t(\omega, x) \in E_2$$

such that  $t \mapsto f_t(\cdot, x)$  is progressively measurable for each  $x \in E_1$ , and

$$|f_t(\omega, x) - f_t(\omega, y)| \leq K|x - y| \text{ for all } x, y \in E_1 \text{ } d\mathbb{P}(\omega)\text{-a.s.}$$

for some  $K > 0$  independent of  $(t, \omega) \in [0, T] \times \Omega$ . In what follows, we shall write only  $f_t(x)$  for  $f_t(\omega, x)$ .

We denote by  $BV_{\mathbb{F}}(E_2)$  the set of  $E_2$ -valued continuous adapted processes with bounded variations. For ease of notation, we write  $\mathcal{E}$  for  $BV_{\mathbb{F}}(E)$  and set

$$\mathcal{E}_m^b := \{\epsilon \in \mathcal{E} : |\epsilon|(T) \leq m \text{ } \mathbb{P}\text{-a.s.}\}, \quad m > 0.$$

In the rest of this section, we fix  $(\mu, \sigma) \in L_{\mathbb{F}}(\mathbb{R}^d, \mathbb{R}^d \times \mathbb{M}^d)$  and assume that the conditions (2.2) and (2.3) hold. Our first result extends Theorem 5.1 in [6].

**LEMMA 2.1.** *Let  $X$  be a continuous semimartingale with values in  $\bar{\mathcal{O}}$ . Fix  $m > 0$  and  $\epsilon \in \mathcal{E}_m^b$ . Assume that  $Y$  is a continuous semimartingale with values in  $\bar{\mathcal{O}}$  satisfying, for  $0 \leq t_0 \leq t \leq T$ ,*

$$Y(t) = X(t_0) + \int_{t_0}^t \mu_s(X(s))ds + \int_{t_0}^t \sigma_s(X(s))dW(s) + \int_{t_0}^t \gamma(Y(s), \epsilon(s))dL(s),$$

where  $L$  is a nondecreasing element of  $BV_{\mathbb{F}}(\mathbb{R}_+)$  such that

$$L(t) = \int_{t_0}^t \mathbb{I}_{\{Y(s) \in \partial \mathcal{O}\}} dL(s), \quad t_0 \leq t \leq T.$$

Let  $X'$  be another continuous semimartingale with values in  $\bar{\mathcal{O}}$  and assume that  $(Y', L')$  satisfies the same properties as  $(Y, L)$  with  $X'$  in place of  $X$ . Then, there is a constant  $C_m > 0$  such that

$$\mathbb{E} \left[ \sup_{t_0 \leq s \leq t} |\Delta Y(s)|^4 \right] \leq C_m \mathbb{E} \left[ |\Delta X(t_0)|^4 + \int_{t_0}^t \sup_{t_0 \leq s \leq u} |\Delta X(s)|^4 du \right], \quad t_0 \leq t \leq T,$$

where  $\Delta Y$  and  $\Delta X$  stand for  $Y - Y'$  and  $X - X'$ .

In order to prove Lemma 2.1, we shall appeal to the following technical result. It is a simple extension of Theorem 3.2 in [6], which is based on Theorem 4.1 in [5].

LEMMA 2.2. *Given  $\theta \in (0, 1)$  there exists a family of functions  $(f_\varepsilon)_{\varepsilon>0}$  in  $C^2(\bar{\mathcal{O}} \times \bar{\mathcal{O}} \times E)$  and a constant  $K > 0$  independent of  $\varepsilon > 0$  such that, for all  $(y, y', e) \in \bar{\mathcal{O}} \times \bar{\mathcal{O}} \times E$ ,*

$$(2.6) \quad \frac{|y - y'|^2}{\varepsilon} \leq f_\varepsilon(y, y', e) \leq K \left( \varepsilon + \frac{|y - y'|^2}{\varepsilon} \right),$$

$$(2.7) \quad \langle \gamma(y, e), D_y f_\varepsilon(y, y', e) \rangle \leq K \frac{|y - y'|^2}{\varepsilon} \quad \text{if} \quad \langle y' - y, \gamma(y, e) \rangle \geq -\theta |y - y'|,$$

$$(2.8) \quad \langle \gamma(y', e), D_{y'} f_\varepsilon(y, y', e) \rangle \leq K \frac{|y - y'|^2}{\varepsilon} \quad \text{if} \quad \langle y - y', \gamma(y', e) \rangle \geq -\theta |y - y'|,$$

$$(2.9) \quad |D_y f_\varepsilon(y, y', e) + D_{y'} f_\varepsilon(y, y', e)| \vee |D_e f_\varepsilon(y, y', e)| \leq K \frac{|y - y'|^2}{\varepsilon},$$

$$(2.10) \quad |D_y f_\varepsilon(y, y', e)| \vee |D_{y'} f_\varepsilon(y, y', e)| \leq K \frac{|y - y'|}{\varepsilon},$$

$$(2.11) \quad D_{(y, y')}^2 f_\varepsilon(y, y', e) \leq \frac{C}{\varepsilon} \begin{pmatrix} I_d & -I_d \\ -I_d & I_d \end{pmatrix} + K \frac{|y - y'|^2}{\varepsilon} I_{2d}.$$

Moreover, there is  $h \in C^2(\bar{\mathcal{O}} \times E)$  with nonnegative values such that

$$(2.12) \quad \langle D_y h(y, e), \gamma(y, e) \rangle \geq 1 \quad \text{for all } (y, e) \in \partial \mathcal{O} \times E.$$

*Proof.* This follows from the proof of Theorem 4.1 in [5]. Since it is long, we provide only the main arguments. Let  $g : (p, x) \in \mathbb{R}^d \times \mathbb{R}^d$  be as in Lemma 4.4 of [5]. In particular, it satisfies

$$(2.13) \quad |D_x g(p, x)| \leq C |p|^2$$

for some  $C > 0$ . Let  $\psi \in C^2(\mathbb{R})$  be a real nondecreasing function such that  $\psi(t) = t$  for  $t \geq 2$ ,  $\psi(t) = 1$  for  $t \leq 1/2$ , and  $\psi(t) \geq t$  for all  $t \in [1/2, 2]$ . For  $\varepsilon > 0$ , we then define

$$f_\varepsilon(x, y, e) := \varepsilon \tilde{g} \left( \frac{x - y}{\varepsilon}, x, e \right), \quad (x, y, e) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^\ell,$$

with

$$\tilde{g}(p, x, e) := \psi(g(p, \gamma(x, e))), \quad (p, x, e) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^\ell.$$

All the estimates, except the one on  $|D_e f_\varepsilon(y, y', e)|$ , follow directly from the property of  $g$  stated in Lemma 4.4 of [5], as in the proof of Theorem 4.1 in [5, pp. 1136–1137], for fixed values of  $e$ . Here, the constant  $K$  can be taken independent of  $e$  because  $E$  is bounded. The estimate on  $|D_e f_\varepsilon(y, y', e)|$  follows from (2.13), the smoothness condition on  $\gamma$ , and the boundedness of  $\mathcal{O}$ , and  $E$ . The existence of the function  $h$  follows from Theorem 3.2 in [6]; see (3.20) of this paper. It suffices to repeat the argument of the proof of Corollary 2.1; i.e., consider a fictitious extended reflected system  $(x, e)$ .  $\square$

*Remark 2.1.* Observe that given  $\theta \in (0, 1)$  such that  $\theta^2 > 1 - r^2$ , we can find  $\delta \in (0, r)$  for which  $\langle y' - y, \gamma(y, e) \rangle \geq -\theta|y - y'|$  for all  $e \in E$ ,  $y \in \partial\mathcal{O}$ , and  $y' \in \bar{\mathcal{O}}$  such that  $|y - y'| \leq \delta$ . This follows from (2.3) and the observation that  $\langle y' - y, \gamma(y, e) \rangle \leq -\theta|y - y'|$ ,  $|y - y'| \leq \delta$ , and  $|\gamma| = 1$  imply that

$$|y' - (y - \lambda\gamma(y, e))|^2 \leq |y' - y|^2 - 2\lambda\theta|y - y'| + \lambda^2 = \lambda^2(1 - \theta^2) \leq \lambda^2 r^2$$

for  $\lambda := |y - y'|/\theta \leq \delta/(1 - r^2)^{\frac{1}{2}}$  with  $\delta$  small enough so that  $\lambda \leq r$ .

*Proof of Lemma 2.1.* As in [6], we first observe that we can restrict our attention to the case  $|Y - Y'| \leq \delta$ , where  $\delta$  is defined as in Remark 2.1 for  $\theta := (1 + \sqrt{1 - r^2})/2$ . Indeed, since  $\mathcal{O}$  is bounded, there is  $\tilde{r} > 0$  such that  $B(0, \tilde{r}/2) \supset \bar{\mathcal{O}}$ , and if  $\tau$  is the first time after  $t_0$  when  $|Y - Y'| \geq \delta$ , then

$$\mathbb{E} \left[ \sup_{t_0 \leq s \leq T} |\Delta Y(s)|^4 \right] \leq \frac{\tilde{r}^4}{\delta^4} \mathbb{E} \left[ \sup_{t_0 \leq s \leq \tau} |\Delta Y(s)|^4 \right].$$

From now on, we therefore assume that  $|Y - Y'| \leq \delta$ . For ease of notation, we also restrict our attention to the case where  $t_0 = 0$ ; the general case is handled similarly.

Recall from Lemma 2.2 the definitions of  $h$  and  $f_\varepsilon$  for  $\theta$  defined as above. We fix  $\varepsilon, \lambda > 0$  and define the smooth function  $\tilde{f}_\varepsilon$  on  $\bar{\mathcal{O}} \times \bar{\mathcal{O}} \times E$  by

$$(2.14) \quad \tilde{f}_\varepsilon(y, y', e) := e^{-\lambda(h(y, e) + h(y', e))} f_\varepsilon(y, y', e).$$

Fix  $\bar{K} > 0$ . Set

$$A_t := \int_0^t e^{-\bar{K}|\epsilon|(s)} \left( \left| D_e \tilde{f}_\varepsilon(Y(s), Y'(s), \epsilon(s)) \right| - \bar{K} \tilde{f}_\varepsilon(Y(s), Y'(s), \epsilon(s)) \right) d|\epsilon|(s)$$

and  $\beta_s := e^{-\bar{K}|\epsilon|(s)} e^{-\lambda(h(Y(s), \epsilon(s)) + h(Y'(s), \epsilon(s)))}$ . Since, by the estimates of Lemma 2.2,

$$\begin{aligned} & e^{-\bar{K}|\epsilon|(s)} \left| D_e \tilde{f}_\varepsilon(Y(s), Y'(s), \epsilon(s)) \right| \\ & \leq \beta_s \left( 2\lambda \sup_{(y, e) \in \bar{\mathcal{O}} \times E} |D_e h(y, e)| + 1 \right) K \left( \varepsilon + \frac{|Y(s) - Y'(s)|^2}{\varepsilon} \right) \end{aligned}$$

and

$$e^{-\bar{K}|\epsilon|(s)} \bar{K} \tilde{f}_\varepsilon(Y(s), Y'(s), \epsilon(s)) \geq \beta_s \bar{K} \frac{|Y(s) - Y'(s)|^2}{\varepsilon},$$

we can find  $C > 0$ , independent of  $\lambda$  and  $\varepsilon$ , such that  $A_t \leq \lambda C \varepsilon$  for  $\bar{K}$  large enough with respect to  $K$ ,  $\lambda$ , and  $|D_e h|$ .

Thus, applying Itô's lemma to  $\xi := (e^{-\bar{K}|\epsilon|(t)} \tilde{f}_\varepsilon(Y(t), Y'(t), \epsilon(t)))_{t \leq T}$  leads to

$$(2.15) \quad \xi_t \leq \xi_0 + T\lambda C \varepsilon + G_t + G'_t + H_t,$$

where

$$\begin{aligned} G_t &:= \int_0^t e^{-\bar{K}|\epsilon|(s)} \langle D_y \tilde{f}_\varepsilon(Y(s), Y'(s), \epsilon(s)), \gamma(Y(s), \epsilon(s)) \rangle dL(s), \\ G'_t &:= \int_0^t e^{-\bar{K}|\epsilon|(s)} \langle D_{y'} \tilde{f}_\varepsilon(Y(s), Y'(s), \epsilon(s)), \gamma(Y'(s), \epsilon(s)) \rangle dL'(s), \end{aligned}$$

and

$$\begin{aligned}
 H_t := & \int_0^t e^{-\bar{K}|\epsilon|(s)} \langle D_y \tilde{f}_\epsilon(Y(s), Y'(s), \epsilon(s)), \mu_s(X(s)) - \mu_s(X'(s)) \rangle ds \\
 & + \int_0^t e^{-\bar{K}|\epsilon|(s)} \langle D_y \tilde{f}_\epsilon(Y(s), Y'(s), \epsilon(s)), [\sigma_s(X(s)) - \sigma_s(X'(s))] dW_s \rangle \\
 & + \int_0^t e^{-\bar{K}|\epsilon|(s)} \langle D_{y'} \tilde{f}_\epsilon(Y(s), Y'(s), \epsilon(s)) + D_y \tilde{f}_\epsilon(Y(s), Y'(s), \epsilon(s)), \mu_s(X'(s)) \rangle ds \\
 & + \int_0^t e^{-\bar{K}|\epsilon|(s)} \langle D_{y'} \tilde{f}_\epsilon(Y(s), Y'(s), \epsilon(s)) + D_y \tilde{f}_\epsilon(Y(s), Y'(s), \epsilon(s)), \sigma_s(X'(s)) dW_s \rangle \\
 & + \frac{1}{2} \int_0^t e^{-\bar{K}|\epsilon|(s)} \text{Tr} \left[ D_{(y, y')}^2 \tilde{f}_\epsilon(Y(s), Y'(s), \epsilon(s)) a_s(X(s), X'(s)) \right] ds
 \end{aligned}$$

with

$$a_s(x, x') = \begin{bmatrix} \sigma_s(x) \sigma_s(x)^* & \sigma_s(x) \sigma_s(x')^* \\ \sigma_s(x') \sigma_s(x)^* & \sigma_s(x') \sigma_s(x')^* \end{bmatrix},$$

where  $*$  denotes the transposition.

Now, observe that the estimates (2.6), (2.7), and (2.12) of Lemma 2.2, Remark 2.1, and the assumption  $|Y - Y'| \leq \delta$  imply that

$$\begin{aligned}
 G_t &= \int_0^t \beta_s \langle D_y f_\epsilon(Y(s), Y'(s), \epsilon(s)), \gamma(Y(s), \epsilon(s)) \rangle dL(s) \\
 &\quad - \lambda \int_0^t \beta_s f_\epsilon(Y(s), Y'(s), \epsilon(s)) \langle D_y h(Y(s), \epsilon(s)), \gamma(Y(s), \epsilon(s)) \rangle dL(s) \\
 &\leq (K - \lambda) \int_0^t \beta_s \frac{|\Delta Y(s)|^2}{\epsilon} dL(s).
 \end{aligned}$$

Similarly,

$$G'_t \leq (K - \lambda) \int_0^t \beta_s \frac{|\Delta Y(s)|^2}{\epsilon} dL'(s).$$

Taking  $\lambda = K$ , it then follows from (2.15) that

$$(2.16) \quad \xi_t \leq \xi_0 + T\lambda C\epsilon + H_t.$$

Moreover, it follows from Doob's inequality, the estimates of Lemma 2.2, the a.s. Lipschitz continuity of  $\mu$  and  $\sigma$ , the fact that  $Y$ ,  $Y'$ ,  $X$ , and  $X'$  are bounded, and the inequality  $\alpha^2 \beta^2 \leq \alpha^4 + \beta^4$ ,  $\alpha, \beta \in \mathbb{R}$ , that

$$\mathbb{E} \left[ \sup_{s \leq t} H_s^2 \right] \leq C' \mathbb{E} \left[ \int_0^t \frac{e^{-2\bar{K}|\epsilon|(s)}}{\epsilon^2} (\epsilon^4 + |\Delta Y(s)|^4 + |\Delta X(s)|^4) ds \right],$$

where  $C'$  is a positive constant which does not depend on  $\varepsilon$ . Since  $|\epsilon|(T) \leq m$ , it follows from (2.16) and the left-hand side of (2.6) of Lemma 2.2 that

$$\mathbb{E} \left[ \sup_{s \leq t} |\Delta Y(s)|^4 \right] \leq C_m \left( \varepsilon^4 + |\Delta X(0)|^4 + \int_0^t \mathbb{E} \left[ \sup_{r \leq s} |\Delta Y(r)|^4 + \sup_{r \leq s} |\Delta X(r)|^4 \right] ds \right),$$

where  $C_m$  is a positive constant independent of  $\varepsilon$ . The required result is then obtained by sending  $\varepsilon \rightarrow 0$  and using Gronwall's lemma.  $\square$

We can now provide the main result of this section, which ensures the strong existence and uniqueness of an SDE with random coefficients and controlled reflecting directions.

**THEOREM 2.2.** *Fix  $\epsilon \in \mathcal{E}$ ,  $t \in [0, T]$ , and  $\xi$  a  $\mathcal{F}_t$ -measurable random variable with values in  $\mathcal{O}$ . Then, there exists a unique continuous adapted process  $(X, L)$  such that  $L$  is a nondecreasing element of  $\text{BV}_{\mathbb{F}}(\mathbb{R}_+)$  and*

$$(2.17) \quad \begin{aligned} X(s) &= \xi + \int_t^s \mu_r(X(r))dr + \int_t^s \sigma_r(X(r))dW(r) + \int_t^s \gamma(X(r), \epsilon(r))dL(r), \\ L(s) &= \int_t^s \mathbb{I}_{\{X(r) \in \partial \mathcal{O}\}} dL(r), \quad t \leq s \leq T. \end{aligned}$$

*Proof.* Observe that Lemma 4.7 in [6] can be easily extended to our setting by appealing to the arguments already used in the proof of Corollary 2.1. The existence and uniqueness when  $|\epsilon|(T)$  is uniformly bounded then follow from Corollary 2.1, Lemma 2.1, and the same arguments as in [6] (see the discussion after their Corollary 5.2), or as in the proof of Proposition 4.1 in [12]. In the case where  $|\epsilon|(T)$  is not uniformly bounded, we use a localization argument. For each  $m \geq 1$ , we define  $\tau_m := \inf\{s \geq t : |\epsilon|(s) \geq m\}$  and let  $(X^m, L^m)$  be the unique solution of (2.17) associated with  $\epsilon^m(\cdot) := \epsilon(\cdot \wedge \tau_m)$ . We then define  $(X, L)$  by

$$(X, L)(s) := (X^1, L^1)(s) \mathbb{I}_{\{t \leq s \leq \tau_1\}} + \sum_{m \geq 2} (X^m, L^m)(s) \mathbb{I}_{\{\tau_{m-1} < s \leq \tau_m\}}.$$

It solves (2.17) associated with  $\epsilon$ . The same argument provides uniqueness.  $\square$

**Remark 2.2.** The presence of the control  $\epsilon$  in  $\gamma$  plays a similar role as the time dependence in nonlinear Neumann-type boundary conditions of the form  $L(t, x, u, Du) = 0$  in the viscosity literature. To the best of our knowledge the papers dealing with such a time dependence impose rather strong regularity conditions. The less stringent seem to appear in [3] where, for fixed  $(x, u, p)$ , the map  $t \mapsto L(t, x, u, p)$  is absolutely continuous with respect to the Lebesgue measure; see condition (H6) of this paper. In particular,  $t \mapsto L(t, x, u, p)$  has bounded variations. It is therefore not surprising to retrieve such a condition in the definition of the set of controls  $\mathcal{E}$ .

**Remark 2.3.** Let  $(a, b)$  be a progressively measurable process with values in  $\mathbb{M}^\ell \times \mathbb{R}^\ell$  satisfying

$$\int_0^t (|b(s)| + |a(s)|^2) < \infty \quad \mathbb{P}\text{-a.s.},$$

and assume that the process  $Z$  defined on  $[t, T]$  by

$$Z(s) := z + \int_t^s b(r)dr + \int_t^s a(r)dW(r)$$

takes values in a compact set  $F$  of  $\mathbb{R}^\ell$ . Then, it follows from Theorem 2.2 that existence and uniqueness hold for

$$X(s) = x + \int_t^s \mu_r(X(r))dr + \int_t^s \sigma_r(X(r))dW(r) + \int_t^s \tilde{\gamma}(X(r), Z(r), \epsilon(r))dL(r),$$

$$L(s) = \int_t^s \mathbb{I}_{\{X(r) \in \partial\mathcal{O}\}} d|L|(r), \quad t \leq s \leq T,$$

when  $\tilde{\gamma} \in C^2(\mathbb{R}^d \times \mathbb{R}^\ell \times \mathbb{R}^\ell, \mathbb{R}^d)$  satisfies

$$\bigcup_{0 \leq \lambda \leq r} B(x - \lambda \tilde{\gamma}(x, z, e), \lambda r) \subset \mathcal{O}^c \quad \text{for all } (x, z, e) \in \partial\mathcal{O} \times \mathbb{R}^{2\ell}$$

for some  $r \in (0, 1)$ . This is easily checked by arguing as in the proof of Corollary 2.1; i.e., introduce the fictitious reflected system  $(X, Z)$  and apply Theorem 2.2. This allows us to introduce a new control on the direction of reflection which corresponds to an Itô process.

**3. Optimal control.** As in the previous section, we consider a bounded open set  $\mathcal{O} \subset \mathbb{R}^d$  and  $\gamma \in C^2(\mathbb{R}^{d+\ell}, \mathbb{R}^d)$  such that  $|\gamma| = 1$  and (2.3) holds.

**3.1. Definitions and assumptions.** We fix a compact subset  $A$  of  $\mathbb{R}^\ell$  and denote by  $\mathcal{A}$  the set of progressively measurable processes with values in  $A$ .

Let  $\mu$  and  $\sigma$  be two continuous maps on  $\mathbb{R}^d \times A$  with values in  $\mathbb{R}^d$  and  $\mathbb{M}^d$ , respectively. We assume that both are Lipschitz with respect to their first variable, uniformly in the other ones, so that  $(\mu^\alpha, \sigma^\alpha)$  defined by

$$(\mu_t^\alpha, \sigma_t^\alpha)(\cdot) := (\mu, \sigma)(\cdot, \alpha(t)), \quad t \leq T,$$

belongs to  $L_{\mathbb{F}}(\mathbb{R}^d; \mathbb{R}^d \times \mathbb{M}^d)$  for all  $\alpha \in \mathcal{A}$ . It then follows from Theorem 2.2 that, for all  $(t, x) \in [0, T] \times \bar{\mathcal{O}}$ , there exists a unique solution  $(X_{t,x}^{\alpha, \epsilon}, L_{t,x}^{\alpha, \epsilon})$  to (2.17) associated with  $(\mu^\alpha, \sigma^\alpha)$  with initial conditions given by  $(X_{t,x}^{\alpha, \epsilon}, L_{t,x}^{\alpha, \epsilon})(t) = (x, 0)$ .

The goal of this section is to provide a PDE characterization for the control problem

$$(3.1) \quad v(t, x) := \sup_{(\alpha, \epsilon) \in \mathcal{A} \times \mathcal{E}} J(t, x; \alpha, \epsilon),$$

where

$$J(t, x; \alpha, \epsilon) := \mathbb{E} \left[ \beta_{t,x}^{\alpha, \epsilon}(T) g(X_{t,x}^{\alpha, \epsilon}(T)) + \int_t^T \beta_{t,x}^{\alpha, \epsilon}(s) f(X_{t,x}^{\alpha, \epsilon}(s), \alpha(s)) ds \right],$$

$$\beta_{t,x}^{\alpha, \epsilon}(s) := e^{-\int_t^s \rho(X_{t,x}^{\alpha, \epsilon}(r), \epsilon(r)) dL_{t,x}^{\alpha, \epsilon}(r)},$$

and  $\rho, g, f$  are continuous real valued maps on  $\bar{\mathcal{O}} \times E$ ,  $\bar{\mathcal{O}}$  and  $\bar{\mathcal{O}} \times A$ , respectively. In order to ensure that  $J$  is well defined, we assume that  $\rho \geq 0$ . We also assume that

- (i)  $g$  is Lipschitz continuous;
- (ii)  $f$  is Lipschitz continuous in its first variable, uniformly in its second one;
- (iii)  $\rho$  is  $C^1$  with Lipschitz first derivative in its first variable, uniformly in its second one, and Lipschitz in its second variable, uniformly in the first one.

**3.2. Dynamic programming.** We first provide some useful estimates on  $X_{t,x}^{\alpha,\epsilon}$  and  $J$  which will be used to derive the dynamic programming principle of Lemma 3.2 below.

PROPOSITION 3.1. *For each  $m > 0$ , there is a constant  $C_m > 0$  such that for all  $(\alpha, \epsilon) \in \mathcal{A} \times \mathcal{E}_m^b$ ,  $t \leq t' \leq T$ , and  $x, x' \in \bar{\mathcal{O}}$ , we have*

$$(3.2) \quad \mathbb{E} \left[ \sup_{t' \leq s \leq T} |X_{t,x}^{\alpha,\epsilon}(s) - X_{t',x'}^{\alpha,\epsilon}(s)|^4 \right]^{\frac{1}{4}} \leq C_m \left( |x - x'| + |t' - t|^{\frac{1}{4}} \right),$$

$$(3.3) \quad \mathbb{E} \left[ \sup_{t \leq s \leq t'} |X_{t,x}^{\alpha,\epsilon}(s) - x|^4 \right]^{\frac{1}{4}} + \mathbb{E} [L_{t,x}^{\alpha,\epsilon}(t')^2]^{\frac{1}{2}} \leq C_m |t' - t|^{\frac{1}{4}},$$

$$(3.4) \quad \mathbb{E} \left[ \sup_{t' \leq s \leq T} |\ln(\beta_{t,x}^{\alpha,\epsilon}(s)) - \ln(\beta_{t',x'}^{\alpha,\epsilon}(s))| \right] \leq C_m \left( |x - x'| + |t' - t|^{\frac{1}{4}} \right).$$

*Proof.* We write  $(X, L, \beta)$  and  $(X', L', \beta')$  for  $(X_{t,x}^{\alpha,\epsilon}, L_{t,x}^{\alpha,\epsilon}, \beta_{t,x}^{\alpha,\epsilon})$  and  $(X_{t',x'}^{\alpha,\epsilon}, L_{t',x'}^{\alpha,\epsilon}, \beta_{t',x'}^{\alpha,\epsilon})$ .

1. It follows from Lemma 2.1 and Gronwall's lemma that

$$\mathbb{E} \left[ \sup_{t' \leq s \leq T} |X(s) - X'(s)|^4 \right] \leq C_m \mathbb{E} [|X(t') - x'|^4],$$

where  $C_m > 0$  denotes a generic constant independent of  $(t, t', x, x')$ . Choosing some large  $\bar{K} > 0$ , applying Itô's lemma to  $(e^{-\bar{K}|\epsilon|(t)} \tilde{f}_\epsilon(X(t), y, \epsilon(t)))_{t \leq T}$ ,  $y \in \bar{\mathcal{O}}$ , and  $\tilde{f}_\epsilon$  defined as in (2.14), and using the same arguments as in Lemma 2.1 (the terms corresponding to  $A_t$  and  $G_t$  are treated similarly; the term corresponding to  $H_t$  is bounded by using the fact that the integrands are bounded by  $C/\epsilon$  for some  $C > 0$  by Lemma 2.2), leads to

$$(3.5) \quad \mathbb{E} \left[ \sup_{t \leq s \leq t'} |X(s) - y|^4 \right] \leq C_m (|t' - t| + |x - y|^4).$$

This proves (3.2) and the bound for the first term in (3.3).

2. We now provide the bound for the second term in (3.3). Let  $h$  be defined as in Lemma 2.2. Applying Itô's lemma to  $h(X, \epsilon) - h(x, \epsilon)$  and using (2.12) leads to

$$\begin{aligned} 0 \leq L(t') &\leq \int_t^{t'} \langle D_x h(X(s), \epsilon(s)), \gamma(X(s), \epsilon(s)) \rangle dL(s) \\ &= h(X(t'), \epsilon(t')) - h(x, \epsilon(t')) \\ &\quad - \int_t^{t'} \left( \langle D_x h(X(s), \epsilon(s)), \mu_s(X(s)) \rangle + \frac{1}{2} \text{Tr}[D_x^2 h(X(s), \epsilon(s)) \sigma_s \sigma_s^*(X(s))] \right) ds \\ &\quad - \int_t^{t'} \langle D_x h(X(s), \epsilon(s)), \sigma_s(X(s)) dW_s \rangle \\ &\quad - \int_t^{t'} \langle D_\epsilon h(X(s), \epsilon(s)) - D_\epsilon h(x, \epsilon(s)), d\epsilon(s) \rangle, \end{aligned}$$



where, by the Lipschitz continuity of  $D_e h$ ,

$$\left| \int_t^{t'} \langle D_e h(X(s), \epsilon(s)) - D_e h(x, \epsilon(s)), d\epsilon(s) \rangle \right| \leq C \sup_{t \leq s \leq t'} |X(s) - x| |\epsilon|(T)$$

for some  $C > 0$  which depends only on  $h$ . Since  $|\epsilon|(T) \leq m$ , the bound for the second term in (3.3) then follows from the Lipschitz continuity of the coefficients, the previous estimates, and the boundedness of  $\mathcal{O}$  and  $E$ .

3. We finally prove (3.4). Since  $|\gamma| = 1$  and  $\rho|\gamma|^2$  is bounded, we have for  $s \in [t', T]$

$$\begin{aligned} & |\ln \beta(s) - \ln \beta'(s)| \\ &= \left| \int_t^s (\rho|\gamma|^2)(X(r), \epsilon(r)) dL(r) - \int_{t'}^s (\rho|\gamma|^2)(X'(r), \epsilon(r)) dL'(r) \right| \\ &\leq \left| \int_t^s \langle (\rho\gamma)(X(r), \epsilon(r)), \gamma(X(r), \epsilon(r)) \rangle dL(r) \right. \\ &\quad \left. - \int_{t'}^s \langle (\rho\gamma)(X(r), \epsilon(r)), \gamma(X'(r), \epsilon(r)) \rangle dL'(r) \right| \\ &\quad + \left| \int_{t'}^s \langle (\rho\gamma)(X(r), \epsilon(r)) - (\rho\gamma)(X'(r), \epsilon(r)), \gamma(X'(r), \epsilon(r)) \rangle dL'(r) \right| \\ &\leq \left| \int_{t'}^s \langle \rho\gamma(X(r), \epsilon(r)), \gamma(X(r), \epsilon(r)) \rangle dL(r) \right. \\ &\quad \left. - \int_{t'}^s \langle \rho\gamma(X(r), \epsilon(r)), \gamma(X'(r), \epsilon(r)) \rangle dL'(r) \right| \\ &\quad + C \left( L(t') + \sup_{t' \leq s \leq T} |X(s) - X'(s)| L'(T) \right) \end{aligned}$$

for some  $C > 0$  independent of  $(s, x, x', t, t')$ . If we assume that  $\rho \in C^{2,1}(\mathbb{R}^{d+\ell}, \mathbb{R})$ , then applying Itô's lemma to  $\langle X - X', \gamma(X, \epsilon) \rho(X, \epsilon) \rangle$  on  $[t', s]$  leads to

$$\begin{aligned} & \langle X(s) - X'(s), \gamma(X(s), \epsilon(s)) \rho(X(s), \epsilon(s)) \rangle \\ &= \langle X(t') - X'(t'), \gamma(X(t'), \epsilon(t')) \rho(X(t'), \epsilon(t')) \rangle \\ &\quad + \int_{t'}^s \langle \gamma(X(r), \epsilon(r)) \rho(X(r), \epsilon(r)), \mu_r(X(r)) - \mu_r(X'(r)) \rangle dr \\ &\quad + \int_{t'}^s \sum_{i \leq d} (\gamma(X(r), \epsilon(r)) \rho(X(r), \epsilon(r)))^i (\sigma_r(X(r)) - \sigma_r(X'(r)))^i dW_r \\ &\quad + \int_{t'}^s \langle \rho\gamma(X(r), \epsilon(r)), \gamma(X(r), \epsilon(r)) \rangle dL(r) \\ &\quad - \int_{t'}^s \langle \rho\gamma(X(r), \epsilon(r)), \gamma(X'(r), \epsilon(r)) \rangle dL'(r) \end{aligned}$$

$$\begin{aligned}
& + \int_{t'}^s \sum_{i \leq d} (X(r) - X'(r))^i \left( \langle D_x([\gamma\rho]^i)(X(r), \epsilon(r)), \mu_r(X(r)) \rangle \right) dr \\
& + \int_{t'}^s \sum_{i \leq d} (X(r) - X'(r))^i \left( \frac{1}{2} \text{Tr}[D_x^2([\gamma\rho]^i)(X(r), \epsilon(r)) \sigma_r \sigma_r^*(X(r))] \right) dr \\
& + \int_{t'}^s \sum_{i \leq d} (X(r) - X'(r))^i \left( \langle D_x([\gamma\rho]^i)(X(r), \epsilon(r)), \gamma(X(r), \epsilon(r)) \rangle \right) dL(r) \\
& + \int_{t'}^s \sum_{i \leq d} (X(r) - X'(r))^i \langle D_x([\gamma\rho]^i)(X(r), \epsilon(r)), \sigma_r(X(r)) dW_r \rangle \\
& + \int_{t'}^s \sum_{i \leq d} (X(r) - X'(r))^i \langle D_e([\gamma\rho]^i)(X(r), \epsilon(r)), d\epsilon(r) \rangle \\
& + \int_{t'}^s \sum_{i \leq d} \langle \sigma_r(X(r)) - (\sigma_r(X'(r)))^i, \sum_{k \leq d} (D_x([\gamma\rho]^i)(X(r), \epsilon(r)))^k \sigma_r^{k\cdot}(X(r)) \rangle dr,
\end{aligned}$$

where  $M^{j\cdot}$  denotes the  $j$ th column of a matrix  $M$ . Using the Lipschitz continuity of the coefficients and the bound  $|\epsilon|(T) \leq m$  thus leads to

$$\begin{aligned}
& \mathbb{E} \left[ \sup_{t' \leq s \leq T} \left| \int_{t'}^s \langle \rho \gamma(X(r), \epsilon(r)), \gamma(X(r), \epsilon(r)) \rangle dL(r) \right. \right. \\
& \quad \left. \left. - \int_{t'}^s \langle \rho \gamma(X(r), \epsilon(r)), \gamma(X'(r), \epsilon(r)) \rangle dL'(r) \right| \right] \\
& \leq C_m \mathbb{E} \left[ \sup_{t' \leq s \leq T} |X(s) - X'(s)|^2 \right]^{\frac{1}{2}} \left( 1 + \mathbb{E} [L(T)^2]^{\frac{1}{2}} \right),
\end{aligned}$$

where  $C_m$  depends on  $\rho$  only through the bounds on  $|\rho|$ , on the first and second derivatives in its first variable, and on the first derivative in its second variable. Thus, by the previous inequality and the Cauchy–Schwarz inequality,

$$\begin{aligned}
& \mathbb{E} \left[ \sup_{t' \leq s \leq T} |\ln \beta(s) - \ln \beta'(s)| \right] \\
& \leq C_m \left( \mathbb{E} [L(t')] + \mathbb{E} \left[ \sup_{t' \leq s \leq T} |X(s) - X'(s)|^2 \right]^{\frac{1}{2}} \left( \mathbb{E} [(L'(T))^2]^{\frac{1}{2}} + \mathbb{E} [(L(T))^2]^{\frac{1}{2}} + 1 \right) \right)
\end{aligned}$$

for some  $C_m > 0$  as above. In view of the previous estimates, the result follows for  $\rho$  smooth enough. Since the estimate of (3.3) clearly does not depend on  $\rho$ , this result is easily extended to the general case by a standard approximation argument.  $\square$

*Remark 3.1.* It follows from the pathwise uniqueness result of Theorem 2.2 and standard arguments (see, e.g., Theorems 5.3.19 and 5.4.20 of [10]), that  $X_{t,x}^{\alpha,\epsilon}$  is a strong Markov process.

LEMMA 3.1. Fix  $m > 0$  and set

$$v_m(t, x) := \sup_{(\alpha, \epsilon) \in \mathcal{A} \times \mathcal{E}_m^b} J(t, x; \alpha, \epsilon), \quad (t, x) \in [0, T] \times \bar{\mathcal{O}}.$$

Then, there is  $C_m > 0$  such that

$$|J(t, x; \alpha, \epsilon) - J(t', x'; \alpha, \epsilon)| + |v_m(t, x) - v_m(t', x')| \leq C_m \left( |t - t'|^{\frac{1}{4}} + |x - x'| \right)$$

for all  $(t, t', x, x') \in [0, T]^2 \times \bar{\mathcal{O}}^2$  and  $(\alpha, \epsilon) \in \mathcal{A} \times \mathcal{E}_m^b$ . Moreover,  $v = \lim_{m \rightarrow \infty} \uparrow v_m = \sup_{m > 0} v_m$  on  $[0, T] \times \bar{\mathcal{O}}$  and  $v$  is lower semicontinuous.

*Proof.* Since

$$|v_m(t, x) - v_m(t', x')| \leq \sup_{(\alpha, \epsilon) \in \mathcal{A} \times \mathcal{E}_m^b} |J(t, x; \alpha, \epsilon) - J(t', x'; \alpha, \epsilon)|,$$

the first assertion follows from the uniform estimates of Proposition 3.1, the Lipschitz continuity assumptions on the parameters  $g$  and  $f$ , and the fact that  $\rho \geq 0$  so that  $\beta_{t,x}^{\alpha,\epsilon} \leq 1$  for all  $(t, x) \in [0, T] \times \bar{\mathcal{O}}$  and  $(\alpha, \epsilon) \in \mathcal{A} \times \mathcal{E}$ . Clearly  $(v_m)_{m>0}$  is nondecreasing and  $v \geq \sup_{m>0} v_m$ . Thus, it remains to prove that when  $v \leq \sup_{m>0} v_m$ , the lower semicontinuity of  $v$  will then follow from the continuity of each  $v_m$ . To see this, fix  $(t, x) \in [0, T] \times \bar{\mathcal{O}}$ ,  $(\alpha, \epsilon) \in \mathcal{A} \times \mathcal{E}$  and set  $\tau_m := \inf\{s \in [t, T] : |\epsilon|(s) \geq m\}$  and  $\epsilon_m := \epsilon(\cdot \wedge \tau_m)$ ,  $m > 0$ . Since  $\tau_m \rightarrow \infty$ , we have  $\beta_{t,x}^{\alpha,\epsilon_m} f(X_{t,x}^{\alpha,\epsilon_m}, \alpha) \rightarrow \beta_{t,x}^{\alpha,\epsilon} f(X_{t,x}^{\alpha,\epsilon}, \alpha)$   $dt \times d\mathbb{P}$ -a.e. on  $[t, T]$  and  $(X_{t,x}^{\alpha,\epsilon_m}, \beta_{t,x}^{\alpha,\epsilon_m})(T) \rightarrow (X_{t,x}^{\alpha,\epsilon}, \beta_{t,x}^{\alpha,\epsilon})(T)$   $\mathbb{P}$ -a.s. as  $m \rightarrow \infty$ . By dominated convergence and the continuity of  $g$ , we then deduce that  $J(t, x; \alpha, \epsilon_m) \rightarrow J(t, x; \alpha, \epsilon)$ . This implies that, for each  $\varepsilon > 0$ , we can find  $m > 0$  such that  $v(t, x) - \varepsilon \leq v_m(t, x)$  and therefore  $v(t, x) \leq \sup_{m>0} v_m(t, x)$ .  $\square$

We can now prove the following dynamic programming principle.

LEMMA 3.2. Fix  $(t, x) \in [0, T] \times \bar{\mathcal{O}}$ . For all  $[t, T]$ -valued stopping times  $\theta$ , we have

$$v(t, x) = \sup_{(\alpha, \epsilon) \in \mathcal{A} \times \mathcal{E}} \mathbb{E} \left[ \beta_{t,x}^{\alpha,\epsilon}(\theta) v(\theta, X_{t,x}^{\alpha,\epsilon}(\theta)) + \int_t^\theta \beta_{t,x}^{\alpha,\epsilon}(s) f(X_{t,x}^{\alpha,\epsilon}(s), \alpha(s)) ds \right].$$

*Proof.* Fix  $(t_0, x_0) \in [0, T] \times \bar{\mathcal{O}}$  (the case  $t_0 = T$  is trivial). The fact that  $v(t_0, x_0)$  is bounded from above by

$$\sup_{(\alpha, \epsilon) \in \mathcal{A} \times \mathcal{E}} \mathbb{E} \left[ \beta_{t_0,x_0}^{\alpha,\epsilon}(\theta) v(\theta, X_{t_0,x_0}^{\alpha,\epsilon}(\theta)) + \int_{t_0}^\theta \beta_{t_0,x_0}^{\alpha,\epsilon}(s) f(X_{t_0,x_0}^{\alpha,\epsilon}(s), \alpha(s)) ds \right]$$

follows from the Markov feature of our model; see Remark 3.1. We now prove the converse inequality.

1. Fix  $m > 0$ . Let  $(B_n)_{n \geq 1}$  be a partition of  $[0, T] \times \bar{\mathcal{O}}$  and  $(t_n, x_n)_{n \geq 1}$  be a sequence such that  $(t_n, x_n) \in B_n$  for each  $n \geq 1$ . By definition, we can find  $\xi^n := (\alpha^n, \epsilon^n) \in \mathcal{A} \times \mathcal{E}_m^b$  such that

$$(3.6) \quad J(t_n, x_n; \xi^n) \geq v_m(t_n, x_n) - \varepsilon/3,$$

where  $\varepsilon > 0$  is a fix parameter. Moreover, by the uniform continuity of  $v_m$  and  $J(\cdot; \xi)$  for  $\xi \in \mathcal{A} \times \mathcal{E}_m^b$  (see Lemma 3.1), we can choose  $(B_n, t_n, x_n)_{n \geq 1}$  in such a way that

$$(3.7) \quad |J(\cdot; \xi^n) - J(t_n, x_n; \xi^n)| + |v_m - v_m(t_n, x_n)| \leq \varepsilon/3 \quad \text{on } B_n.$$

2. Given  $\xi \in \mathcal{A} \times \mathcal{E}_m^b$  and  $\theta$  a stopping time with values in  $[t_0, T]$ , we define  $\bar{\xi} \in \mathcal{A} \times \mathcal{E}_m^b$  by

$$\bar{\xi}(t) := \xi(t) \mathbb{I}_{\{t < \theta\}} + \mathbb{I}_{\{t \geq \theta\}} \sum_{n \geq 1} \xi^n(t) \mathbb{I}_{\{(\theta, X_{t_0,x_0}^{\xi^n}(\theta)) \in B_n\}}.$$

Using successively the Markov feature of our model (see Remark 3.1, (3.7), (3.6), (3.7)) again and the fact that  $\rho \geq 0$  (which implies that  $\beta_{t_0, x_0}(\theta) \leq 1$ ), we deduce that, for all  $\xi = (\alpha, \epsilon) \in \mathcal{A} \times \mathcal{E}^b$ ,

$$\begin{aligned}
 J(t_0, x_0; \bar{\xi}) &\geq \mathbb{E} \left[ \beta_{t_0, x_0}^\xi(\theta) J(\theta, X_{t_0, x_0}^\xi(\theta); \bar{\xi}) + \int_t^\theta \beta_{t_0, x_0}^\xi(s) f(X_{t_0, x_0}^\xi(s), \alpha(s)) ds \right] \\
 &= \mathbb{E} \left[ \beta_{t_0, x_0}^\xi(\theta) \sum_{n \geq 1} J(\theta, X_{t_0, x_0}^\xi(\theta); \xi^n) \mathbb{I}_{\{(\theta, X_{t_0, x_0}^\xi(\theta)) \in B_n\}} \right] \\
 &\quad + \mathbb{E} \left[ \int_t^\theta \beta_{t_0, x_0}^\xi(s) f(X_{t_0, x_0}^\xi(s), \alpha(s)) ds \right] \\
 &\geq \mathbb{E} \left[ \beta_{t_0, x_0}^\xi(\theta) \sum_{n \geq 1} J(t_n, x_n; \xi^n) \mathbb{I}_{\{(\theta, X_{t_0, x_0}^\xi(\theta)) \in B_n\}} \right] \\
 &\quad + \mathbb{E} \left[ \int_t^\theta \beta_{t_0, x_0}^\xi(s) f(X_{t_0, x_0}^\xi(s), \alpha(s)) ds \right] - \varepsilon/3 \\
 &\geq \mathbb{E} \left[ \beta_{t_0, x_0}^\xi(\theta) \left( \sum_{n \geq 1} v_m(t_n, x_n; \xi^n) \mathbb{I}_{\{(\theta, X_{t_0, x_0}^\xi(\theta)) \in B_n\}} \right) \right] \\
 &\quad + \mathbb{E} \left[ \int_t^\theta \beta_{t_0, x_0}^\xi(s) f(X_{t_0, x_0}^\xi(s), \alpha(s)) ds \right] - 2\varepsilon/3 \\
 &\geq \mathbb{E} \left[ \beta_{t_0, x_0}^\xi(\theta) v_m(\theta, X_{t_0, x_0}^\xi(\theta)) + \int_t^\theta \beta_{t_0, x_0}^\xi(s) f(X_{t_0, x_0}^\xi(s), \alpha(s)) ds \right] - \varepsilon.
 \end{aligned}$$

By arbitrariness of  $\varepsilon > 0$ , this shows that

$$(3.8) \quad v(t_0, x_0) \geq \mathbb{E} \left[ \beta_{t_0, x_0}^\xi(\theta) v_m(\theta, X_{t_0, x_0}^\xi(\theta)) + \int_t^\theta \beta_{t_0, x_0}^\xi(s) f(X_{t_0, x_0}^\xi(s), \alpha(s)) ds \right].$$

Since  $v_m \rightarrow v$  as  $m \rightarrow \infty$  by Lemma 3.1, it follows by dominated convergence that, for all  $\xi \in \mathcal{A} \times \mathcal{E}^b$ ,

$$v(t_0, x_0) \geq \mathbb{E} \left[ \beta_{t_0, x_0}^\xi(\theta) v(\theta, X_{t_0, x_0}^\xi(\theta)) + \int_t^\theta \beta_{t_0, x_0}^\xi(s) f(X_{t_0, x_0}^\xi(s), \alpha(s)) ds \right].$$

The same localization argument as in the proofs of Theorem 2.2 and Lemma 3.1 then implies that the above inequality actually holds for all  $\xi \in \mathcal{A} \times \mathcal{E}$ .  $\square$

**3.3. PDE characterization for the optimal control problem.** In this section, we show that  $v$  is a solution of

$$\mathcal{K}\varphi = 0,$$

where

$$\mathcal{K}\varphi := \begin{cases} \min_{a \in A} (-\mathcal{L}^a \varphi - f(\cdot, a)) = 0 & \text{on } [0, T] \times \mathcal{O}, \\ \min_{e \in E} \mathcal{H}^e \varphi = 0 & \text{on } [0, T] \times \partial \mathcal{O}, \\ \varphi - g = 0 & \text{on } \{T\} \times \bar{\mathcal{O}}, \end{cases}$$

and, for a smooth function  $\varphi$  on  $[0, T] \times \bar{\mathcal{O}}$  and  $(a, e) \in A \times E$ , we set

$$\begin{aligned} \mathcal{L}^a \varphi &:= \frac{\partial}{\partial t} \varphi + \langle \mu(\cdot, a), D\varphi \rangle + \frac{1}{2} \text{Tr} \sigma(\cdot, a) \sigma(\cdot, a)^* D^2 \varphi, \\ \mathcal{H}^e \varphi &:= \rho(\cdot, e) \varphi - \langle \gamma(\cdot, e), D\varphi \rangle. \end{aligned}$$

**3.3.1. Definitions.** Since  $v$  may not be smooth, we need to consider the above equation in the viscosity sense. Moreover, the boundary conditions may not be satisfied in a strong sense and, as usual, we have to consider a relaxed version; see, e.g., [4]. We therefore introduce the operator  $\mathcal{K}_+$  and  $\mathcal{K}_-$  defined as

$$\mathcal{K}_+ \varphi := \begin{cases} \mathcal{K}\varphi & \text{on } [0, T] \times \mathcal{O}, \\ \max \left\{ \min_{a \in A} -\mathcal{L}^a \varphi - f(\cdot, a), \min_{e \in E} \mathcal{H}^e \varphi \right\} & \text{on } [0, T] \times \partial \mathcal{O}, \\ \varphi - g & \text{on } \{T\} \times \partial \mathcal{O} \end{cases}$$

and

$$\mathcal{K}_- \varphi := \begin{cases} \mathcal{K}\varphi & \text{on } [0, T] \times \mathcal{O}, \\ \min \left\{ \min_{a \in A} -\mathcal{L}^a \varphi - f(\cdot, a), \min_{e \in E} \mathcal{H}^e \varphi \right\} & \text{on } [0, T] \times \partial \mathcal{O}, \\ \min \left\{ \varphi - g, \min_{e \in E} \mathcal{H}^e \varphi \right\} & \text{on } \{T\} \times \partial \mathcal{O}. \end{cases}$$

**DEFINITION 3.1.** We say that a lower-semicontinuous (resp., upper-semicontinuous) function  $w$  on  $[0, T] \times \bar{\mathcal{O}}$  is a viscosity supersolution (resp., subsolution) of

$$(3.9) \quad \mathcal{K}\varphi = 0$$

if for all  $\varphi \in C^{1,2}([0, T] \times \bar{\mathcal{O}})$  and all  $(t, x) \in [0, T] \times \bar{\mathcal{O}}$ , which realizes a local minimum (resp., maximum) of  $w - \varphi$  equal to 0, we have  $\mathcal{K}_+ \varphi \geq 0$  (resp.,  $\mathcal{K}_- \varphi \leq 0$ ). We say that a locally bounded function  $w$  is a (discontinuous) viscosity solution of (3.9) if  $w_*$  (resp.,  $w^*$ ) is a supersolution (resp., subsolution) of (3.9), where

$$\begin{aligned} w^*(t, x) &:= \limsup_{(t', x') \rightarrow (t, x), (t', x') \in D} w(t', x'), \\ w_*(t, x) &:= \liminf_{(t', x') \rightarrow (t, x), (t', x') \in D} w(t', x'), \quad (t, x) \in [0, T] \times \bar{\mathcal{O}}, \end{aligned}$$

with  $D := [0, T] \times \mathcal{O}$ .

**Remark 3.2.** Take  $E = \tilde{K}_1 := \tilde{K} \cap \partial B(0, 1)$ , where  $\tilde{K}$  is the domain of the support function  $\delta$  of a closed convex set  $K \subset \mathbb{R}^\ell$ , i.e.,

$$\delta(e) := \sup_{y \in K} \langle y, e \rangle, \quad e \in \mathbb{R}^\ell,$$

and assume that  $\rho(x, e) = \delta(e)$  and  $\gamma(x, e) = e$  on  $\partial\mathcal{O} \times E$ . Then, for  $\varphi \in C^1(\bar{\mathcal{O}}, (0, \infty))$ , the constraint

$$\min_{e \in E} \mathcal{H}^e \varphi = \min_{e \in E} (\delta(e) \varphi - \langle e, D\varphi \rangle) \geq 0$$

means that  $D\varphi/\varphi \in K$ ; see, e.g., [13]. In this case, the term  $\mathcal{H}^e \varphi \geq 0$  can be assimilated into a constraint on the gradient of the logarithm of the solution at the boundary of  $\mathcal{O}$ . A similar constraint appears in [2], but in the whole domain.

*Remark 3.3.* Assume that  $\mathcal{O}$  is  $C^2$  and that  $\sigma$  satisfies the noncharacteristic boundary condition

$$(3.10) \quad \min_{a \in A} |\sigma(x, a)\xi| > 0 \text{ for all } x \in \partial\mathcal{O} \text{ and } \xi \in \mathbb{R}^d \setminus \{0\}.$$

Then, it follows from the same arguments as in step 2 of the proof of Proposition 6.3 of [2] that  $w$  is a supersolution of  $\mathcal{K}_+\varphi = 0$  only if it is a supersolution of  $\bar{\mathcal{K}}_+\varphi = 0$ , where

$$\bar{\mathcal{K}}_+\varphi := \begin{cases} \mathcal{K}_+\varphi & \text{on } ([0, T] \times \mathcal{O}) \cup (\{T\} \times \bar{\mathcal{O}}), \\ \min_{e \in E} \mathcal{H}^e \varphi & \text{on } [0, T) \times \partial\mathcal{O}. \end{cases}$$

Similarly, it follows from the same arguments as in step 2 of Proposition 6.6 in [2] that  $w$  is a subsolution of  $\mathcal{K}_-\varphi = 0$  only if it is a subsolution of  $\bar{\mathcal{K}}_-\varphi = 0$ , where

$$\bar{\mathcal{K}}_-\varphi := \begin{cases} \mathcal{K}_-\varphi & \text{on } ([0, T] \times \mathcal{O}) \cup (\{T\} \times \bar{\mathcal{O}}), \\ \min_{e \in E} \mathcal{H}^e \varphi & \text{on } [0, T) \times \partial\mathcal{O}. \end{cases}$$

### 3.3.2. Super- and subsolution properties.

PROPOSITION 3.2. *The function  $v_*$  is a viscosity supersolution of (3.9).*

*Proof.* The fact that  $v_* \geq g$  on  $\{T\} \times \bar{\mathcal{O}}$  is a direct consequence of the lower semi-continuity of  $v$ ; see Lemma 3.1. Fix  $(t_0, x_0) \in [0, T) \times \bar{\mathcal{O}}$  and  $\varphi \in C^{1,2}([0, T] \times \bar{\mathcal{O}})$  such that

$$0 = (v_* - \varphi)(t_0, x_0) = \min_{[0, T] \times \bar{\mathcal{O}}} (v_* - \varphi).$$

1. We first assume that  $(t_0, x_0) \in [0, T) \times \partial\mathcal{O}$  and that

$$\max \left\{ \min_{a \in A} -\mathcal{L}^a \varphi(t_0, x_0) - f(x_0, a), \min_{e \in E} \mathcal{H}^e \varphi(t_0, x_0) \right\} =: -2\varepsilon < 0$$

and work toward a contradiction. Under the above assumption, we can find  $(a_0, e_0) \in A \times E$  and  $\delta \in (t_0, T - t_0)$  for which

$$(3.11) \quad \max \{ -\mathcal{L}^{a_0} \varphi - f(\cdot, a_0), \mathcal{H}^{e_0} \varphi \} \leq -\varepsilon$$

on  $\bar{B}_0 \cap \bar{D}_0$  where  $B_0 := B(t_0, \delta) \times B(x_0, \delta)$  and  $D_0 := (t_0 - \delta, t_0 + \delta) \times \mathcal{O}$ . Observe that we can assume, without loss of generality, that  $(t_0, x_0)$  achieves a strict local minimum so that

$$(3.12) \quad \inf_{\partial_p B_0 \cap \bar{D}_0} (v_* - \varphi) =: \zeta > 0,$$

where  $\partial_p B_0 = ([t_0 - \delta, t_0 + \delta] \times \partial B(x_0, \delta)) \cup (\{t_0 + \delta\} \times B(x_0, \delta))$ . Let  $(t_k, x_k)_{k \geq 1}$  be a sequence in  $B_0 \cap D_0$  satisfying

$$(t_k, x_k) \longrightarrow (t_0, x_0) \text{ and } v(t_k, x_k) \longrightarrow v_*(t_0, x_0) \text{ as } k \longrightarrow \infty$$

so that

$$(3.13) \quad \eta_k := v(t_k, x_k) - \varphi(t_k, x_k) \longrightarrow 0 \text{ as } k \longrightarrow \infty.$$

Let us write  $(X^k, L^k, \beta^k)$  for  $(X_{t_k, x_k}^{a_0, e_0}, L_{t_k, x_k}^{a_0, e_0}, \beta_{t_k, x_k}^{a_0, e_0})$ , where  $(a_0, e_0)$  is viewed as an element of  $\mathcal{A} \times \mathcal{E}$ . Set

$$\theta^k := \inf \{s \geq t_k : (s, X^k(s)) \notin B_0\}, \vartheta^k := \inf \{s \geq t_k : X^k(s) \notin \mathcal{O}\}.$$

It then follows from Itô's lemma, (3.11), and (3.12) that

$$\begin{aligned} v(t_k, x_k) &\leq \eta_k + \mathbb{E} \left[ \beta^k(\theta^k) v(\theta^k, X^k(\theta^k)) + \int_{t_k}^{\theta^k} \beta^k(s) f(X^k(s), a_0) ds \right] \\ &\quad - \mathbb{E} [\zeta \mathbb{I}_{\{\theta^k < \vartheta^k\}} + (\beta^k(\theta^k) \zeta + \varepsilon L^k(\theta^k)) \mathbb{I}_{\{\theta^k \geq \vartheta^k\}}], \end{aligned}$$

where we used the fact that  $\beta^k(\theta^k) = 1$  on  $\{\theta^k < \vartheta^k\}$ . Let  $c > 0$  be such that  $|\rho| \leq c$  on  $\bar{\mathcal{O}} \times E$  and observe that

$$\nu := \inf_{\ell \in [0, \infty)} e^{-c\ell} \zeta + \varepsilon \ell > 0.$$

It follows that

$$v(t_k, x_k) \leq \eta_k - \zeta \wedge \nu + \mathbb{E} \left[ \beta^k(\theta^k) v(\theta^k, X^k(\theta^k)) + \int_{t_k}^{\theta^k} \beta^k(s) f(X^k(s), a_0) ds \right],$$

which leads to a contradiction to Lemma 3.2 for  $k$  large enough; recall (3.13).

2. The case where  $(t_0, x_0) \in [0, T) \times \mathcal{O}$  is treated similarly. We assume that

$$\min_{a \in A} -\mathcal{L}^a \varphi(t_0, x_0) - f(x_0, a) =: -2\varepsilon < 0$$

and repeat the above argument with  $\delta$  small enough so that  $B(x_0, \delta) \subset \mathcal{O}$  and therefore  $\theta^k < \vartheta^k$  (so that the reflection does not operate on  $[t_0, \theta^k]$ ).  $\square$

**PROPOSITION 3.3.** *The function  $v^*$  is a viscosity subsolution of (3.9).*

*Proof.* Fix  $(t_0, x_0) \in [0, T) \times \bar{\mathcal{O}}$  and  $\varphi \in C^{1,2}([0, T] \times \bar{\mathcal{O}})$  such that

$$0 = (v^* - \varphi)(t_0, x_0) = \max_{[0, T] \times \bar{\mathcal{O}}} (v^* - \varphi).$$

The case where  $(t_0, x_0) \in [0, T) \times \bar{\mathcal{O}}$  is treated by similar arguments as in the proof of Proposition 3.2; see also what follows. We therefore assume that  $t_0 = T$ .

1. We first consider the case where  $x_0 \in \partial \mathcal{O}$ . We assume that

$$\min \left\{ \varphi - g, \min_{e \in E} \mathcal{H}^e \varphi \right\} =: 2\varepsilon > 0.$$

Set  $\phi(t, x) = \varphi(t, x) + \sqrt{T - t}$  so that  $(\partial/\partial t)\phi(t, x) \rightarrow -\infty$  as  $t \rightarrow T$  and observe that  $(T, x_0)$  also achieves a maximum for  $v^* - \phi$ . Without loss of generality, we can therefore

assume that  $(\partial/\partial t)\varphi(t, x) \rightarrow -\infty$  as  $t \rightarrow T$  and that we can find  $\delta \in (t_0, T - t_0)$  for which

$$(3.14) \quad \min \left\{ \min_{a \in A} -\mathcal{L}^a \varphi - f(\cdot, a), \varphi - g, \min_{e \in E} \mathcal{H}^e \varphi \right\} \geq \varepsilon$$

on  $\bar{B}_0 \cap \bar{D}_0$  where  $B_0 := [t_0 - \delta, T] \times B(x_0, \delta)$  and  $D_0 := (t_0 - \delta, T) \times \mathcal{O}$ . Observe that we can assume, without loss of generality, that  $(t_0, x_0)$  achieves a strict local maximum so that

$$(3.15) \quad \max_{\partial_p B_0 \cap \bar{D}_0} (v^* - \varphi) =: -\zeta < 0,$$

where  $\partial_p B_0 = ([t_0 - \delta, T] \times \partial B(x_0, \delta)) \cup (\{T\} \times B(x_0, \delta))$ . Let  $(t_k, x_k)_{k \geq 1}$  be a sequence in  $B_0 \cap D_0$  satisfying

$$(t_k, x_k) \longrightarrow (t_0, x_0) \quad \text{and} \quad v(t_k, x_k) \longrightarrow v^*(t_0, x_0) \quad \text{as } k \longrightarrow \infty$$

so that

$$(3.16) \quad \eta_k := v(t_k, x_k) - \varphi(t_k, x_k) \longrightarrow 0 \quad \text{as } k \longrightarrow \infty.$$

Let us write  $(X^k, L^k, \beta^k)$  for  $(X_{t_k, x_k}^{\alpha, \epsilon}, L_{t_k, x_k}^{\alpha, \epsilon}, \beta_{t_k, x_k}^{\alpha, \epsilon})$ , where  $(\alpha, \epsilon)$  is a given element of  $\mathcal{A} \times \mathcal{E}$ . Set

$$\theta^k := \inf \{s \geq t_k : (s, X^k(s)) \notin B_0\}, \quad \vartheta^k := \inf \{s \geq t_k : X^k(s) \notin \mathcal{O}\}.$$

It follows from Itô's lemma, (3.14), (3.15), and the identity  $v(T, \cdot) = g$  that

$$\begin{aligned} v(t_k, x_k) &\geq \eta_k + \mathbb{E} \left[ \beta^k(\theta^k) v(\theta^k, X^k(\theta^k)) + \int_{t_k}^{\theta^k} \beta^k(s) f(X^k(s), \alpha(s)) ds \right] \\ &\quad + \mathbb{E} [\zeta \mathbb{I}_{\{\theta^k < \vartheta^k\}} + (\beta^k(\theta^k)(\zeta \wedge \varepsilon) + \varepsilon L^k(\theta^k)) \mathbb{I}_{\{\theta^k \geq \vartheta^k\}}]. \end{aligned}$$

Arguing as in step 1 of the proof of Proposition 3.2, this implies that

$$\begin{aligned} v(t_k, x_k) &\geq \eta_k + \zeta \wedge \nu \\ &\quad + \mathbb{E} \left[ \beta^k(\theta^k) v(\theta^k, X^k(\theta^k)) + \int_{t_k}^{\theta^k} \beta^k(s) f(X^k(s), \alpha(s)) ds \right] \end{aligned}$$

for some  $\nu > 0$  independent of  $(\alpha, \epsilon)$ . By arbitrariness of  $(\alpha, \epsilon)$  and (3.16), this leads to a contradiction to Lemma 3.2 for  $k$  large enough.

2. The case where  $x_0 \in \mathcal{O}$  is treated similarly. It suffices to take  $\delta$  small enough so that  $B(x_0, \delta) \subset \mathcal{O}$  and therefore  $\theta^k < \vartheta^k$ .  $\square$

**3.4. A comparison result.** A lot of work has been done so far on comparison results for quasi-linear second-order parabolic PDEs with nonlinear or oblique derivative Neumann conditions; see, e.g., [1], [9], [3], or [5] and the references therein. However, as in the first three papers, they usually require additional smoothness conditions on  $\mathcal{O}$  or, as in [5], do not allow for nonlinearities at the boundary.

In this section, we provide a comparison theorem for (3.9) in the case where there exist  $\bar{e}$  and  $\underline{e}$  in  $E$  such that

$$(3.17) \quad \underline{e} \in \arg \min \{\rho(x, e), e \in E\}, \quad \bar{e} \in \arg \max \{\rho(x, e), e \in E\} \quad \text{for all } x \in \partial \mathcal{O}$$



and additional conditions on the directions of reflection are imposed as follows:

1. As in section 7.B of [4], we first make a uniform exterior ball assumption in the direction  $\gamma$ :

$$(3.18) \quad \exists b > 0 \text{ s.t. } B(x - b\gamma(x, e), b) \cap \mathcal{O} = \emptyset \text{ for all } (x, e) \in \partial\mathcal{O} \times E.$$

2. We then assume that there is a  $C^2(\bar{\mathcal{O}})$  function  $\hat{h}$  such that

$$(3.19) \quad \langle \gamma(x, e), D\hat{h}(x) \rangle \geq 1 \text{ for all } x \in \partial\mathcal{O} \text{ and } e \in E.$$

3. The direction  $\gamma(\cdot, \bar{e})$  satisfies

$$(3.20) \quad \inf_{e \in E} \langle \gamma(x, e), \gamma(x, \bar{e}) \rangle > 0 \text{ for all } x \in \partial\mathcal{O}.$$

*Remark 3.4.* The condition (3.19) holds in the case where  $E$  is a singleton; see (2.12). When  $\partial\mathcal{O}$  is  $C^2$ , i.e., the algebraic distance  $d$  to  $\partial\mathcal{O}$  is  $C^2$ , and

$$\min_{e \in E} \langle \gamma(x, e), Dd(x) \rangle \geq \varepsilon \text{ for all } x \in \partial\mathcal{O}$$

for some  $\varepsilon > 0$ , then we can choose  $\hat{h} = \varepsilon^{-1}d$ . This imposes a restriction on the direction of reflection with respect to the unit normal inward vector at  $x \in \partial\mathcal{O}$ .

Under these conditions, we can state the following comparison theorem for super- and subsolutions of (3.9).

**PROPOSITION 3.4.** *Assume that (3.17), (3.18), (3.19), and (3.20) hold. Let  $u$  (resp.,  $w$ ) be a bounded upper-semicontinuous viscosity subsolution (resp., lower-semicontinuous viscosity supersolution) of (3.9). Then,  $u \leq w$  on  $[0, T] \times \bar{\mathcal{O}}$ .*

*Proof.* We argue by contradiction and assume that  $\max_{\bar{D}}(u - w) > 0$ , with  $D := [0, T] \times \mathcal{O}$ . We can then find  $\varepsilon > 0$  small enough and  $(t_0, x_0) \in \bar{D}$  such that

$$(3.21) \quad \max_{\bar{D}}(\tilde{u} - \tilde{w} - 2\varepsilon H) = (\tilde{u} - \tilde{w} - 2\varepsilon H)(t_0, x_0) =: \eta > 0,$$

where  $\tilde{u}(t, x) = e^{\kappa t}u(t, x)$ ,  $\tilde{w}(t, x) = e^{\kappa t}w(t, x)$ , and  $H(t, x) := e^{-\kappa t - \hat{h}(x)}$ , where  $\hat{h}$  is defined as in (3.19) and  $\kappa > 0$  is a constant parameter such that

$$(3.22) \quad -\mathcal{L}^a H \geq 0 \text{ on } \bar{D} \text{ for all } a \in A.$$

We first assume that

$$(3.23) \quad u(t_0, x_0) \geq 0.$$

The case  $u(t_0, x_0) < 0$  will be treated in step 4 below.

Given  $\lambda \in \mathbb{N}$ , we next define

$$\Phi_\lambda(t, x, y) := \tilde{u}(t, x) - \tilde{w}(t, y) - \Psi_\lambda(t, x, y),$$

where

$$\begin{aligned} \Psi_\lambda(t, x, y) &:= \varepsilon(H(t, x) + H(t, y)) + \rho(x_0, \underline{e})u(t_0, x_0)\langle \gamma(x_0, \underline{e}), x - y \rangle \\ &\quad + \frac{\lambda}{2}|x - y|^2 + |t - t_0|^2 + |x - x_0|^4 \end{aligned}$$

for some  $\zeta > 0$ .

Let  $(t_\lambda, x_\lambda, y_\lambda)$  be a global maximum point for  $\Phi_\lambda$  on  $\bar{D}$ . Using standard arguments, one easily checks that

$$(3.24) \quad (t_\lambda, x_\lambda) \rightarrow (t_0, x_0), \quad \lambda |x_\lambda - y_\lambda|^2 \rightarrow 0, \quad (\tilde{u}(t_\lambda, x_\lambda), \tilde{w}(t_\lambda, y_\lambda)) \rightarrow (\tilde{u}(t_0, x_0), \tilde{w}(t_0, x_0))$$

as  $\lambda \rightarrow \infty$ ; see, e.g., Lemma 3.1 and Proposition 3.7 in [4].

Moreover, Ishii's lemma (see Theorem 8.3 in [4]) implies that we can find  $p_{\lambda,1}$ ,  $p_{\lambda,2} \in \mathbb{R}$ , and two symmetric matrices  $X_{\alpha,\lambda}$  and  $Y_{\alpha,\lambda}$ , depending on a parameter  $\alpha > 0$ , such that

$$(3.25) \quad \begin{aligned} (p_{\lambda,1}, D_x \Psi_\lambda(t_\lambda, x_\lambda, y_\lambda), X_{\alpha,\lambda}) &\in \bar{\mathcal{P}}_{\mathcal{O}}^{2,+} \tilde{u}(t_\lambda, x_\lambda), \\ (p_{\lambda,2}, -D_y \Psi_\lambda(t_\lambda, x_\lambda, y_\lambda), Y_{\alpha,\lambda}) &\in \bar{\mathcal{P}}_{\mathcal{O}}^{2,-} \tilde{w}(t_\lambda, y_\lambda) \end{aligned}$$

and

$$(3.26) \quad p_{\lambda,1} - p_{\lambda,2} = 2(t_\lambda - t_0) - \kappa \varepsilon (H(t_\lambda, x_\lambda) + H(t_\lambda, y_\lambda)),$$

$$(3.27) \quad \begin{pmatrix} X_{\alpha,\lambda} & 0 \\ 0 & -Y_{\alpha,\lambda} \end{pmatrix} \leq (A_\lambda + B_\lambda) + \alpha (A_\lambda + B_\lambda)^2,$$

where

$$\begin{aligned} A_\lambda &:= \varepsilon \begin{pmatrix} D^2 H(t_\lambda, x_\lambda) & 0 \\ 0 & D^2 H(t_\lambda, y_\lambda) \end{pmatrix} + 12(x_\lambda - x_0) \otimes (x_\lambda - x_0), \\ B_\lambda &:= \lambda \begin{pmatrix} I_d & -I_d \\ -I_d & I_d \end{pmatrix}, \end{aligned}$$

see [4] for the notations  $\bar{\mathcal{P}}_{\mathcal{O}}^{2,+}$  and  $\bar{\mathcal{P}}_{\mathcal{O}}^{2,-}$ .

1. Assume that  $x_\lambda \in \partial \mathcal{O}$ . Fix  $e \in E$ . Since  $y_\lambda \in \bar{\mathcal{O}}$ , it follows from (3.18) that  $|x_\lambda - b\gamma(x_\lambda, e) - y_\lambda|^2 \geq b^2$ . Since  $|\gamma| = 1$ , this implies

$$(3.28) \quad 2\langle \gamma(x_\lambda, e), y_\lambda - x_\lambda \rangle \geq -b^{-1}|x_\lambda - y_\lambda|^2.$$

Then, it follows from the definition of  $\varepsilon$ , the fact that  $|\gamma| = 1$ , the assumptions  $\rho \geq 0$ , (3.23), (3.17), (3.19), (3.24), and (3.28) that

$$\begin{aligned} &\rho(x_\lambda, e)u(t_\lambda, x_\lambda) - \langle \gamma(x_\lambda, e), D_x \Psi_\lambda(t_\lambda, x_\lambda, y_\lambda) \rangle \\ &= (\rho(x_0, e) - \rho(x_0, \underline{e}))u(t_0, x_0) + \rho(x_0, \underline{e})u(t_0, x_0)(1 - \langle \gamma(x_0, e), \gamma(x_0, \underline{e}) \rangle) \\ &\quad + O(\lambda^{-1}) - \langle \gamma(x_\lambda, e), \lambda(x_\lambda - y_\lambda) - \varepsilon D\hat{h}(x_\lambda)H(t_\lambda, x_\lambda) \rangle \\ &\geq O(\lambda^{-1}) + \varepsilon H(t_0, x_0). \end{aligned}$$

Arguing as above, using the inequalities  $\rho \geq 0$ , and  $u(t_0, x_0) \geq w(t_0, x_0)$ , and observing that  $\langle \gamma(y_\lambda, \underline{e}), \gamma(x_0, \underline{e}) \rangle \rightarrow 1$ , we also deduce that, if  $y_\lambda \in \partial \mathcal{O}$ ,

$$\begin{aligned} &\rho(y_\lambda, \underline{e})w(t_\lambda, y_\lambda) - \langle \gamma(y_\lambda, \underline{e}), -D_y \Psi_\lambda(t_\lambda, x_\lambda, y_\lambda) \rangle \\ &\leq \rho(x_0, \underline{e})(w(t_0, x_0) - u(t_0, x_0)) - \varepsilon H(t_0, x_0) + O(\lambda^{-1}) \\ &\leq -\varepsilon H(t_0, x_0) + O(\lambda^{-1}). \end{aligned}$$

2. We now assume that, up to a subsequence,  $t_\lambda = T$  for all  $\lambda \in \mathbb{N}$ . By step 1 and the fact that  $H(t_0, x_0) > 0$ , we must have  $u(t_\lambda, x_\lambda) \leq g(x_\lambda)$  and  $g(y_\lambda) \leq w(t_\lambda, y_\lambda)$ . Since  $g$  is continuous, we deduce from (3.24) that  $u(t_0, x_0) \leq w(t_0, w_0)$ , which contradicts (3.21); recall that  $H > 0$ .

3. The rest of the proof is standard. We first observe that  $\tilde{u}$  and  $\tilde{w}$  are viscosity super- and subsolutions of  $\tilde{\mathcal{K}}_+\varphi = 0$  and  $\tilde{\mathcal{K}}_-\varphi = 0$ , where  $\tilde{\mathcal{K}}_+$  and  $\tilde{\mathcal{K}}_-$  are defined as  $\mathcal{K}_+$  and  $\mathcal{K}_-$  with  $\mathcal{L}^a$  replaced by  $\tilde{\mathcal{L}}^a$  defined by

$$\tilde{\mathcal{L}}^a\varphi = -\kappa\varphi + \mathcal{L}^a\varphi.$$

In view of steps 1 and 2 and  $H(t_0, x_0) > 0$ , we may find  $a_\lambda$  in the compact set  $A$  such that, after possibly passing to a subsequence,

$$\begin{aligned} 0 &\geq \kappa\tilde{u}(t_\lambda, x_\lambda) - p_{\lambda,1} - \langle \mu(t_\lambda, x_\lambda, a_\lambda), D_x\Psi_\lambda(t_\lambda, x_\lambda, y_\lambda) \rangle \\ &\quad - \frac{1}{2}\text{Tr}\sigma\sigma^*(t_\lambda, x_\lambda, a_\lambda)X_{\eta,\lambda} - f(t_\lambda, x_\lambda, a_\lambda) \\ 0 &\leq \kappa\tilde{w}(t_\lambda, y_\lambda) - p_{\lambda,2} - \langle \mu(t_\lambda, y_\lambda, a_\lambda), -D_y\Psi_\lambda(t_\lambda, x_\lambda, y_\lambda) \rangle \\ &\quad - \frac{1}{2}\text{Tr}\sigma\sigma^*(t_\lambda, y_\lambda, a_\lambda)Y_{\eta,\lambda} - f(t_\lambda, y_\lambda, a_\lambda). \end{aligned}$$

Taking the difference of these two equations and using (3.21) and (3.22), the Lipschitz continuity of the coefficients, and the fact that  $A$  and  $\mathcal{O}$  are bounded, (3.26) and (3.27) leads to

$$\begin{aligned} \kappa\eta + O(\lambda^{-1}) &\leq \kappa(\tilde{u}(t_\lambda, x_\lambda) - \tilde{w}(t_\lambda, y_\lambda)) \\ &\leq O(|t_\lambda - t_0| + \lambda|x_\lambda - y_\lambda|^2 + |x_\lambda - x_0|^2 + C_\lambda\alpha), \end{aligned}$$

where  $C > 0$  is independent of  $\lambda$  and  $\alpha$ , and  $C_\lambda$  depends only on  $\lambda$ . Sending  $\alpha \rightarrow 0$  and then  $\lambda \rightarrow \infty$  thus leads to a contradiction; recall (3.24).

4. The case  $u(t_0, x_0) < 0$  is treated similarly. It suffices to consider the test function

$$\begin{aligned} \Psi_\lambda(t, x, y) &:= \varepsilon(H(t, x) + H(t, y)) + \tilde{b}^{-1}\rho(x_0, \bar{e})u(t_0, x_0)\langle \gamma(x_0, \bar{e}), x - y \rangle \\ &\quad + \frac{\lambda}{2}|x - y|^2 + |t - t_0|^2 + |x - x_0|^4, \end{aligned}$$

where  $\bar{e}$  is defined in (3.17), and  $\tilde{b} > 0$  and  $\bar{e} \in E$  satisfy

$$\min_{e \in E} \langle \gamma(x_0, \bar{e}), \gamma(x_0, e) \rangle = \langle \gamma(x_0, \bar{e}), \gamma(x_0, \tilde{e}) \rangle = \tilde{b};$$

recall (3.20). With this modification, the arguments of step 1 becomes

$$\begin{aligned} &\rho(x_\lambda, e)u(t_\lambda, x_\lambda) - \langle \gamma(x_\lambda, e), D_x\Psi_\lambda(t_\lambda, x_\lambda, y_\lambda) \rangle \\ &= (\rho(x_0, e) - \rho(x_0, \bar{e}))u(t_0, x_0) + \rho(x_0, \bar{e})u(t_0, x_0)(1 - \tilde{b}^{-1}\langle \gamma(x_0, e), \gamma(x_0, \tilde{e}) \rangle) \\ &\quad + O(\lambda^{-1}) - \langle \gamma(x_\lambda, e), \lambda(x_\lambda - y_\lambda) - \varepsilon D\hat{h}(x_\lambda)H(t_\lambda, x_\lambda) \rangle \\ &\geq O(\lambda^{-1}) + \varepsilon H(t_0, x_0) \end{aligned}$$

in the case where  $x_\lambda \in \partial\mathcal{O}$ , and becomes

$$\begin{aligned} & \rho(y_\lambda, \bar{e})w(t_\lambda, y_\lambda) - \langle \gamma(y_\lambda, \bar{e}), -D_y \Psi_\lambda(t_\lambda, x_\lambda, y_\lambda) \rangle \\ & \leq \rho(x_0, \bar{e})(w(t_0, x_0) - u(t_0, x_0)) \\ & \quad + \rho(x_0, \bar{e})u(t_0, x_0)(1 - \tilde{b}^{-1} \langle \gamma(x_0, \bar{e}), \gamma(x_0, \tilde{e}) \rangle) - \varepsilon H(t_0, x_0) + O(\lambda^{-1}) \\ & \leq -\varepsilon H(t_0, x_0) + O(\lambda^{-1}) \end{aligned}$$

in the case where  $y_\lambda \in \partial\mathcal{O}$ . The rest of the proof is similar to the above.  $\square$

*Remark 3.5.* Observe that the right-hand sides of conditions (3.17) and (3.20) are only used in step 4 of the above proof to treat the case  $u(t_0, x_0) < 0$ . It is therefore not required if  $u \geq 0$  on  $[0, T] \times \partial\mathcal{O}$ . Similarly, it can be dropped if  $w \geq 0$  on  $[0, T] \times \partial\mathcal{O}$  since, in this case, (3.21) also implies that  $u(t_0, x_0) \geq 0$ .

*Remark 3.6.* Assume that

$$\mu(x, a) = \text{diag } x\bar{\mu}(x, a), \quad \sigma(x, a) = \text{diag } x\bar{\sigma}(x, a) \quad \text{on } \mathbb{R}_+^d \times A$$

and

$$\gamma(x, e) = \text{diag } x\bar{\gamma}(x, e) \quad \text{on } (\partial\mathcal{O} \cap (0, \infty)^d) \times E$$

with  $\bar{\mu}$ ,  $\bar{\sigma}$ , and  $\bar{\gamma}$  such that  $\mu$ ,  $\sigma$ , and  $\gamma$  satisfy the general assumptions of this section. Then, the process  $X_{t,x}^{\alpha,\epsilon}$  takes values in  $(0, \infty)^d$  whenever  $x \in (0, \infty)^d$ . It is therefore natural to consider the PDE  $\mathcal{K}\varphi = 0$  on  $[0, T] \times (\bar{\mathcal{O}} \cap (0, \infty)^d)$ , with a notion of viscosity solution similar to the one of Definition 3.1 with  $\mathcal{O}$ ,  $\partial\mathcal{O}$ , and  $\bar{\mathcal{O}}$  replaced by  $\mathcal{O}^* := \mathcal{O} \cap (0, \infty)^d$ ,  $\partial\mathcal{O}^* := \partial\mathcal{O} \cap (0, \infty)^d$ , and  $\bar{\mathcal{O}}^* := \bar{\mathcal{O}} \cap (0, \infty)^d$ .

The proofs of Propositions 3.2 and 3.3 are easily adapted to this context. We therefore obtain that  $v$  is a viscosity solution of  $\mathcal{K}\varphi = 0$  on  $[0, T] \times \bar{\mathcal{O}}^*$ . Moreover, the proof of the comparison principle of Proposition 3.4 can also be extended. It suffices to add an additional penalty function of the form  $k \sum_{i \leq d} |x^i|^{-1}$ , with  $k \rightarrow \infty$ , as in [2].

*Remark 3.7.* The smoothness assumptions on  $\rho$  and  $\gamma$  are only used either to construct  $(X_{t,x}^{\alpha,\epsilon}, L_{t,x}^{\alpha,\epsilon})$  or to prove the dynamic programming principle of Lemma 3.2. We shall see, through an example in section 4.3 below, how they can be relaxed.

**4. Application to the pricing of barrier options under constraints.** As already stated in the introduction, our main motivation comes from applications in mathematical finance. More precisely, [2] provides a PDE characterization of the superhedging price of barrier options under portfolio constraints, which is very similar to the equation  $\mathcal{K}\varphi = 0$  up to an additional term inside the domain  $\mathcal{O}$  which imposes a constraint on the gradient of the logarithm of the solution.

The aim of this section is to show that the superhedging price of barrier options under portfolio constraints can actually admit a dual formulation in terms of an optimal control problem for a reflected diffusion in which the direction of reflection is controlled. Due to the additional term, which appears in the PDE of [2], we cannot expect this result to be general and we shall restrict our attention to a Black-Scholes type model; see below.

In order to simplify the presentation, we shall work under quite restrictive conditions, assuming, for instance, that the equation  $\mathcal{K}\varphi = 0$  admits a sufficiently smooth solution for a suitable choice of parameters. The general case is left for future research.

**4.1. Problem formulation.** We briefly present the hedging problem. Details can be found in [2] and the references contained in this paper.

We consider a financial market which consists of one nonrisky asset, whose price process is normalized to unity, and  $d$  risky assets  $S_{t,x} = (S_{t,x}^i)_{i \leq d}$  which solve on  $[t, T]$

$$S_{t,x}(s) = x + \int_t^s \text{diag } S_{t,x}(r) \Sigma \, dW(r),$$

where  $\Sigma$  is a  $d$ -dimensional invertible matrix. A financial strategy is described by a  $d$ -dimensional predictable process  $\pi = (\pi^1, \dots, \pi^d)$  (viewed as a line vector) satisfying the integrability condition

$$(4.1) \quad \int_0^T |\pi(s)|^2 ds < \infty \quad \mathbb{P}\text{-a.s.},$$

where  $\pi^i(s)$  is the proportion of wealth invested at time  $s$  in the risky asset  $S_{t,x}^i$ . We associate with an initial capital  $y \in \mathbb{R}$  and a financial strategy  $\pi$  the induced wealth process  $Y_{t,y}^\pi$  which solves on  $[t, T]$

$$(4.2) \quad Y(s) = y + \int_t^s Y(r) \pi(r) \text{diag } S_{t,x}(r)^{-1} dS_{t,x}(r) = y + \int_t^s Y(r) \pi(r) \Sigma \, dW(r).$$

In this paper, we restrict our attention to the case where the proportions invested in the risky asset are constrained to be bounded from below. Given  $m^i > 0$ ,  $i \leq d$ , we set

$$K := \prod_{i=1}^d [-m^i, \infty)$$

and denote by  $\Pi_K$  the set of financial strategies  $\pi$  satisfying

$$(4.3) \quad \pi \in K \quad dt \times d\mathbb{P}\text{-a.e.}$$

We consider an up-and-out type option. More precisely, we take  $\mathcal{O}$  such that

$$\mathcal{O}^* := \mathcal{O} \cap (0, \infty)^d = \left\{ x \in (0, \infty)^d : \sum_{i=1}^d x^i < \kappa \right\}, \quad \kappa > 0.$$

The “payoff” of the barrier option is a continuous map  $g$  defined on  $\mathbb{R}_+^d$  satisfying

$$(4.4) \quad g \geq 0 \quad \text{on } \mathcal{O}^* \text{ and } g = 0 \quad \text{on } \partial\mathcal{O}^* := \partial\mathcal{O} \cap (0, \infty)^d.$$

In order to apply the general results of [2], we assume that the map  $\hat{g}$  defined by

$$\hat{g}(x) = \sup_{y \in \mathbb{R}_-^d} e^{-\delta(y)} g(x^1 e^{y^1}, \dots, x^d e^{y^d}), \quad x \in \bar{\mathcal{O}}^* := \bar{\mathcal{O}} \cap (0, \infty)^d$$

is continuous. Here,  $\delta$  is the support function of  $K$ ; see Remark 3.2. We also assume that  $\hat{g}$  is almost everywhere differentiable on  $\bar{\mathcal{O}}^*$  and we denote by  $D\hat{g}$  its gradient when it is well defined.

*Remark 4.1.* One easily checks that

$$\hat{g}(x) = \sup_{y \in \mathbb{R}_-^d} e^{-\delta(y)} \hat{g}(x^1 e^{y^1}, \dots, x^d e^{y^d}), \quad x \in \bar{\mathcal{O}}^*;$$

see [2], which implies

$$\inf \left\{ \delta(e)\hat{g}(x) - \langle e, \text{diag } x D\hat{g}(x) \rangle, e \in \tilde{K}_1 \right\} \geq 0$$

for all  $x \in \bar{\mathcal{O}}^*$  where  $D\hat{g}$  is well defined. Here,  $\tilde{K}_1 := \mathbb{R}_-^d \cap \partial B(0, 1)$  is the set of unit elements of the domain of  $\delta$ ; see Remark 3.2.

The option pays  $g(S_{t,x}(T))$  at  $T$  if and only if  $S_{t,x}$  does not exit  $\mathcal{O}^*$  before  $T$ . Since  $S_{t,x}$  has positive components, this corresponds to the situation where

$$\tau_{t,x} := \inf \{s \in [t, T] : S_{t,x}(s) \notin \mathcal{O}\} > T,$$

with the usual convention  $\inf \emptyset = \infty$ .

The super-replication cost of the barrier option is then defined as the minimal initial wealth  $y$  such that  $Y_{t,y}^\pi(T) \geq g(S_{t,x}(T))\mathbb{I}_{\{T < \tau_{t,x}\}}$  for some suitable strategy  $\pi \in \Pi_K$ . This leads to the introduction of the value function defined on  $[0, T] \times \bar{\mathcal{O}}^*$  by

$$(4.5) \quad w(t, x) := \inf \{y \in \mathbb{R} : Y_{t,y}^\pi(T) \geq g(S_{t,x}(T))\mathbb{I}_{\{T < \tau_{t,x}\}} \text{ for some } \pi \in \Pi_K\}.$$

**4.2. PDE characterization.** We define  $\mathcal{L}$  as  $\mathcal{L}^0$  with  $A = \{0\}$ ,  $\mu = 0$ ,  $\sigma(x, \cdot) = \text{diag } x \Sigma$ , and  $f = 0$ . The next result is a consequence of [2].

**THEOREM 4.1** (see Bentahar and Bouchard [2]). *The value function  $w$  is the unique viscosity solution in the class of bounded functions on  $[0, T] \times (\bar{\mathcal{O}} \cap \mathbb{R}_+^d)$  of  $\mathcal{G}\varphi = 0$ , where  $\mathcal{G}\varphi$  equals*

$$\begin{cases} \min \left\{ -\mathcal{L}\varphi(t, x), \min_{e \in \tilde{K}_1} (\delta(e)\varphi(t, x) - \langle e, \text{diag } x D\varphi(t, x) \rangle) \right\} & \text{on } [0, T] \times \mathcal{O}^*, \\ \min \left\{ \varphi, \min_{e \in \tilde{K}_1} (\delta(e)\varphi(t, x) - \langle e, \text{diag } x D\varphi(t, x) \rangle) \right\} & \text{on } [0, T] \times \partial\mathcal{O}^*, \\ \varphi - \hat{g} & \text{on } \{T\} \times \bar{\mathcal{O}}^*. \end{cases}$$

In the above theorem, the notion of viscosity solution has to be taken in the classical sense.

When (4.6), (4.7), (4.8) below admit a sufficiently smooth solution, the above equation can be simplified as follows.

**PROPOSITION 4.1.** *Assume that there is a bounded nonnegative  $C^{1,3}([0, T] \times \mathcal{O}^*) \cap C^{0,1}([0, T] \times \bar{\mathcal{O}}^*) \cap C([0, T] \times \bar{\mathcal{O}}^*)$  function  $\psi$  such that  $\partial\psi/\partial t \in C^{0,1}([0, T] \times \bar{\mathcal{O}}^*)$  and satisfying*

$$(4.6) \quad -\mathcal{L}\psi(t, x) = 0 \text{ on } [0, T] \times \mathcal{O}^*,$$

$$(4.7) \quad \min_{e \in \tilde{K}_1} (\delta(e)\psi(t, x) - \langle e, \text{diag } x D\psi(t, x) \rangle) = 0 \text{ on } [0, T] \times \partial\mathcal{O}^*,$$

$$(4.8) \quad \psi = \hat{g} \text{ on } \{T\} \times \bar{\mathcal{O}}^*,$$

$$(4.9) \quad \lim_{\substack{(t', x') \rightarrow (T, x) \\ (t', x') \in [0, T] \times \mathcal{O}^*}} D\psi(t', x') = D\hat{g}(x) \text{ almost everywhere on } \bar{\mathcal{O}}^*.$$

Then,  $\psi = w$  on  $[0, T] \times \mathcal{O}^*$  and  $\psi$  is the unique bounded solution to (4.6)-(4.7)-(4.8) on  $[0, T] \times \bar{\mathcal{O}}^*$ .

*Proof.* In view of Theorem 4.1, it suffices to show that  $\psi$  is a solution of  $\mathcal{G}\varphi = 0$ . Clearly, it is a subsolution. Since  $\psi \geq 0$ , the supersolution property holds if, in addition to (4.6), (4.7), (4.8), we have

$$(4.10) \quad \min_{e \in \tilde{K}_1} (\delta(e)\psi(t, x) - \langle e, \text{diag } x D\psi(t, x) \rangle) \geq 0 \text{ on } [0, T] \times \mathcal{O}^* .$$

To see this, observe that (4.6) implies that each component  $\phi^k := (D\psi)^k$  of  $D\psi$  solves, on  $[0, T] \times \mathcal{O}^*$ ,

$$-\frac{\partial}{\partial t} \phi^k(t, x) - \frac{1}{2} \text{Tr} \text{diag } x \Sigma \Sigma' \text{diag } x D^2 \phi^k(t, x) - \langle D\phi^k(t, x)^* \text{diag } x \Sigma, \Sigma^k \rangle = 0,$$

where  $\Sigma^k$  denotes the  $k$ th line of  $\Sigma$ . Applying Itô's lemma to  $\langle e, \text{diag } S_{t,x} D\psi(\cdot, S_{t,x}) \rangle$ ,  $e \in \tilde{K}_1$ , and  $(t, x) \in [0, T] \times \mathcal{O}^*$ , and using (4.9), we deduce that

$$\begin{aligned} \langle e, \text{diag } x D\psi(t, x) \rangle &= \mathbb{E} [\langle e, \text{diag } S_{t,x}(\tau_{t,x}) D\psi(\tau_{t,x}, S_{t,x}(\tau_{t,x})) \rangle \mathbb{I}_{\{\tau_{t,x} < T\}}] \\ &\quad + \mathbb{E} [\langle e, \text{diag } S_{t,x}(T) D\hat{g}(S_{t,x}(T)) \rangle \mathbb{I}_{\{\tau_{t,x} \geq T\}}] . \end{aligned}$$

Since by (4.6) and (4.8),

$$\psi(t, x) = \mathbb{E} [\psi(\tau_{t,x}, S_{t,x}(\tau_{t,x})) \mathbb{I}_{\{\tau_{t,x} < T\}} + \hat{g}(S_{t,x}(T)) \mathbb{I}_{\{\tau_{t,x} \geq T\}}] ,$$

it follows from (4.7) and Remark 4.1 that

$$\delta(e)\psi(t, x) - \langle e, \text{diag } x D\psi(t, x) \rangle \geq 0$$

which, by arbitrariness of  $e$ , provides the required result.  $\square$

**4.3. Dual formulation.** Equations (4.6), (4.7), (4.8) are very similar to  $\mathcal{K}\varphi = 0$  with  $E = \tilde{K}_1$  and

$$\rho(x, e) := \delta(e)/|\text{diag } xe|, \quad \gamma(x, e) = \text{diag } xe/|\text{diag } xe| .$$

However, the gradient of  $\text{diag } xe/|\text{diag } xe|$  may blow up near  $\partial(0, \infty)^d$ , and it is not possible to consider a smooth extension of  $\gamma$  on  $\mathbb{R}^{2d}$  (even on  $\mathbb{R}_+^d \times \tilde{K}_1$ ).

In order to overcome this difficulty, we use the following construction. First, we define  $\mathcal{O}$  as

$$\mathcal{O} := \left\{ x \in \mathbb{R}^d : \sum_{i=1}^d |x^i| < \kappa \right\}$$

so that  $\mathcal{O}^* = \{x \in (0, \infty)^d : \sum_{i=1}^d x^i < \kappa\}$ . Let  $r \in (0, 1/2)$  be such that  $B(0, 2r) \subset \mathcal{O}$ . Then, given a nonincreasing  $C^2(\mathbb{R}, [0, 1])$  function  $\phi$  such that  $\phi(y) = 1$  if  $y \leq 1$  and  $\phi(y) = 0$  if  $y \geq 3/2$ , we set, for  $n \geq 1$ ,

$$z_n(e) := (e^i \phi(ne^i + 2) - (1 - \phi(ne^i + 2)))_{i \leq d} .$$

Observe that  $z_n(e) = e$  on  $E_n := \{e \in \tilde{K}_1 : e^i \leq -n^{-1} \text{ for all } i \leq d\}$ ,  $z_n(e) \in (-\infty, -1/(2n)]^d$  for all  $e \in \mathbb{R}^d$ , and

$$(4.11) \quad |\text{diag } xe| \geq r/(2n) := \eta_n \text{ for } (x, e) \in B(0, r)^c \times (-\infty, -1/(2n)]^d .$$

We then set (with  $1_d = (1, \dots, 1) \in \mathbb{R}^d$ )

$$\begin{aligned}\bar{\gamma}_n(x, e) &:= \text{diag } xz_n(e) \left( 1 - \phi \left( \frac{3}{2} |\text{diag } xe| / \eta_n \right) \right) - 1_d \phi \left( \frac{3}{2} |\text{diag } xe| / \eta_n \right), \\ \gamma_n(x, e) &:= \bar{\gamma}_n(x, e) / |\bar{\gamma}_n(x, e)|.\end{aligned}$$

Using (4.11), one easily checks that  $\gamma_n \in C^2(\mathbb{R}^{2d}, \mathbb{R}^d)$ . Moreover,  $\gamma_n(x, e) = \gamma(x, e) = \text{diag } xe / |\text{diag } xe|$  on  $B(0, r)^c \times E_n$ , and (2.3) holds for  $(\mathcal{O}, \gamma_n)$ .

For  $\epsilon \in \mathcal{E}_0 := \cup_{n \geq 1} \text{BV}_{\mathbb{F}}(E_n)$  and  $(t, x) \in [0, T] \times \bar{\mathcal{O}}^*$ , we can then define  $(X_{t,x}^\epsilon, L_{t,x}^\epsilon) := (X_{t,x}^{0,\epsilon}, L_{t,x}^{0,\epsilon})$  as in section 3 with  $\mu = 0$ ,  $\sigma(x, a) = \text{diag } x \Sigma$ , and  $\gamma$  defined as above. Clearly,  $X_{t,x}^\epsilon$  takes values in  $(0, \infty)^d$ .

We next define  $\rho$  on  $\mathbb{R}^d \times \tilde{K}_1$  as

$$\rho(x, e) = (\delta(e) / |\text{diag } xe|) (1 - \phi(|x|/r + 1/2))$$

so that  $\rho$  is continuous on  $\mathbb{R}^d \times \tilde{K}_1$ , satisfies the assumption of section 3 as a function on  $\mathbb{R}^d \times E_n$  for all  $n \geq 1$ , and

$$\rho(x, e) = \delta(e) / |\text{diag } xe| \quad \text{on } \partial \mathcal{O}^* \times (\cup_{n \geq 1} E_n).$$

With this construction, we can now consider the control problem

$$v(t, x) := \sup_{\epsilon \in \mathcal{E}_0} \mathbb{E} \left[ e^{-\int_t^T \rho(X_{t,x}^\epsilon(s), \epsilon(s)) dL_{t,x}^\epsilon(s)} \hat{g}(X_{t,x}^\epsilon(T)) \right], \quad (t, x) \in [0, T] \times \bar{\mathcal{O}}^*.$$

**PROPOSITION 4.2.** *The function  $v$  is a bounded viscosity solution on  $[0, T] \times \bar{\mathcal{O}}^*$  of (4.6), (4.7), (4.8).*

*Proof.* For  $n \geq 1$  and  $(t, x) \in [0, T] \times \bar{\mathcal{O}}^*$ , set

$$v_n(t, x) := \sup_{\epsilon \in \mathcal{E}_n} \mathbb{E} \left[ e^{-\int_t^T \rho(X_{t,x}^\epsilon(s), \epsilon(s)) dL_{t,x}^\epsilon(s)} \hat{g}(X_{t,x}^\epsilon(T)) \right],$$

where  $\mathcal{E}_n := \text{BV}_{\mathbb{F}}(E_n)$ . It follows from the previous discussion that we can apply Lemma 3.2 to  $v_n$ . Since  $v = \sup_{n \geq 1} v_n = \lim_{n \rightarrow \infty} \uparrow v_n$ , a monotone convergence argument shows that the dynamic programming principle of Lemma 3.2 holds for  $v$ . Following the arguments used in Propositions 3.2 and 3.3, and using the continuity of  $\rho$  and  $\gamma$  on  $(B(0, r)^c \cap (0, \infty)^d) \times \tilde{K}_1 \supset \partial \mathcal{O}^* \times \tilde{K}_1$ , we deduce that  $v$  is a viscosity solution of  $\mathcal{K}\varphi = 0$  on  $[0, T] \times \bar{\mathcal{O}}^*$ ; see Remark 3.6. Since

$$\delta(e)y - \langle e, \text{diag } xp \rangle \geq 0 \Leftrightarrow |\text{diag } xe|^{-1} (\delta(e)y - \langle e, \text{diag } xp \rangle) \geq 0$$

for  $(x, e, y, p) \in \partial \mathcal{O}^* \times \tilde{K}_1 \times \mathbb{R} \times \mathbb{R}^d$ , this implies that  $v$  is a viscosity solution on  $[0, T] \times \bar{\mathcal{O}}^*$  of (4.6), (4.7), (4.8).  $\square$

In view of Proposition 4.1, we finally obtain the main result of this section, which provides a dual formulation for the superhedging price  $w$ .

**THEOREM 4.2.** *Let the conditions of Proposition 4.1 hold. Then, for all  $(t, x) \in [0, T] \times \mathcal{O}^*$ ,*

$$(4.12) \quad w(t, x) = \sup_{\epsilon \in \mathcal{E}_0} \mathbb{E} \left[ e^{-\int_t^T \rho(X_{t,x}^\epsilon(s), \epsilon(s)) dL_{t,x}^\epsilon(s)} \hat{g}(X_{t,x}^\epsilon(T)) \right].$$

**Remark 4.2.** It follows from Theorem 4.1, Proposition 4.2, and Theorem 7.1 in [2] that

$$w(t, x) \geq \sup_{\epsilon \in \mathcal{E}_0} \mathbb{E} \left[ e^{-\int_t^T \rho(X_{t,x}^\epsilon(s), \epsilon(s)) dL_{t,x}^\epsilon(s)} \hat{g}(X_{t,x}^\epsilon(T)) \right]$$



even if the conditions of Proposition 4.1 are not satisfied.

*Remark 4.3.* When  $d = 1$ , we retrieve the results of [14]; see also [15]. In this case,  $\mathcal{E}_0 = \{-1\}$  and the right-hand side quantity in (4.12) can be computed by using Monte-Carlo methods.

*Remark 4.4.* It follows from [2] that  $w$  admits the dual formulation

$$w(t, x) = \sup_{\vartheta \in \Theta} \mathbb{E}^\vartheta \left[ e^{-\int_t^T \delta(\vartheta(s)) ds} \hat{g}(S_{t,x}(T)) \mathbb{I}_{\{\tau_{t,x} > T\}} \right],$$

where  $\Theta$  denotes the set of bounded adapted processes with values in  $\mathbb{R}_-^d$ , and  $\mathbb{E}^\vartheta$  is the expectation operator under the equivalent probability measure  $\mathbb{Q}^\vartheta$  under which the process  $W^\vartheta$  defined by

$$W^\vartheta(t) = W(t) - \int_0^t \Sigma^{-1} \vartheta(s) ds, \quad t \leq T,$$

is a Brownian motion. Since

$$S_{t,x}(s) = x + \int_t^s \text{diag } S_{t,x}(r) \Sigma dW^\vartheta(r) + \int_t^s \text{diag } S_{t,x}(r) \vartheta(r) dr,$$

and  $W^\vartheta$  has the same law under  $\mathbb{Q}^\vartheta$  as  $W$  under  $\mathbb{P}$ , this is, at least formally, equivalent to

$$w(t, x) = \sup_{\vartheta \in \Theta} \mathbb{E} \left[ e^{-\int_t^T \delta(\vartheta(s)) ds} \hat{g}(S_{t,x}^\vartheta(T)) \mathbb{I}_{\{\tau_{t,x}^\vartheta > T\}} \right]$$

with  $S_{t,x}^\vartheta$  now defined as the solution of

$$S_{t,x}^\vartheta(s) = x + \int_t^s \text{diag } S_{t,x}^\vartheta(r) \Sigma dW(r) + \int_t^s \text{diag } S_{t,x}^\vartheta(r) \vartheta(r) dr.$$

A formal change of variable ( $\vartheta = |\tilde{\vartheta}| \tilde{\vartheta} / |\text{diag } S_{t,x}^{\tilde{\vartheta}} \tilde{\vartheta}|$ ) then leads to

$$(4.13) \quad w(t, x) = \sup_{\tilde{\vartheta} \in \Theta} \mathbb{E} \left[ e^{-\int_t^T |\tilde{\vartheta}(s)| \rho(S_{t,x}^{\tilde{\vartheta}}(s), \tilde{\vartheta}(s)) ds} \hat{g}(S_{t,x}^{\tilde{\vartheta}}(T)) \mathbb{I}_{\{\tau_{t,x}^{\tilde{\vartheta}} > T\}} \right],$$

where

$$S_{t,x}^{\tilde{\vartheta}}(s) = x + \int_t^s \text{diag } S_{t,x}^{\tilde{\vartheta}}(r) \Sigma dW(r) + \int_t^s |\tilde{\vartheta}(s)| \gamma(S_{t,x}^{\tilde{\vartheta}}(r), \tilde{\vartheta}(r)) dr,$$

$\rho(x, e) = \delta(e) / |\text{diag } xe|$ ,  $\gamma(x, e) = \text{diag } xe / |\text{diag } xe|$ ,  $\tilde{\tau}_{t,x}^{\tilde{\vartheta}}$  is the first exit time of  $S_{t,x}^{\tilde{\vartheta}}$  from  $\mathcal{O}^*$ , and we use the convention  $0/0 = 0$ .

For very large values of  $|\tilde{\vartheta}|$ , the process  $S^{\tilde{\vartheta}}$  is “essentially” reflected in the direction  $\gamma(S_{t,x}^{\tilde{\vartheta}}, \tilde{\vartheta})$ .

Moreover, since  $\hat{g} \geq 0$ , we should seek a control  $\tilde{\vartheta}$  such that  $\tilde{\tau}_{t,x}^{\tilde{\vartheta}} > T$ , i.e., which “causes reflection” of  $S^{\tilde{\vartheta}}$  at least at the boundary  $\partial\mathcal{O}^*$ . The “reflection” should also be optimal so that the right-hand side of (4.13) is maximal. If  $d = 1$  and  $\hat{g}$  is non-decreasing on  $\mathcal{O}^*$ , the action of  $\tilde{\vartheta}$  should then be minimal since it decreases the value of  $S_{t,x}^{\tilde{\vartheta}}(T)$  and  $\rho(x, e) > 0$  if  $e \neq 0$ . Thus, at the limit, the process should be reflected only at the boundary  $\partial\mathcal{O}^*$ . This phenomenon, which was already observed in [14] in the one-dimensional case, naturally leads to the formulation (4.12).

## REFERENCES

- [1] G. BARLES, *Nonlinear Neumann boundary conditions for quasilinear degenerate elliptic equations and applications*, J. Differential Equations, 4 (1999), pp. 191–224.
- [2] I. BENTAHAR AND B. BOUCHARD, *Barrier option hedging under constraints: A viscosity approach*, SIAM J. Control Optim., 45 (2006), pp. 1846–1874.
- [3] M. BOURGOING, *Viscosity Solutions of Fully Nonlinear Second Order Parabolic Equations with  $L^1$ -Time Dependence and Neumann Boundary Conditions*, preprint, 2004.
- [4] M. G. CRANDALL, H. ISHII, AND P.-L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
- [5] P. DUPUIS AND H. ISHII, *On oblique derivative problems for fully nonlinear second-order elliptic partial equations on nonsmooth domains*, Nonlinear Anal. Theory Methods Appl., 15 (1990), pp. 1123–1138.
- [6] P. DUPUIS AND H. ISHII, *SDEs with oblique reflection on nonsmooth domains*, Ann. Probab., 21 (1993), pp. 554–580.
- [7] P. DUPUIS AND K. RAMANAN, *Convex duality and the Skorokhod problem*, I, II, Probab. Theory Related Fields, 115 (1999), pp. 153–195, 197–236.
- [8] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.
- [9] H. ISHII AND M.-H. SATO, *Nonlinear oblique derivative problems for singular degenerate parabolic equations on a general domain*, Nonlinear Anal., 57 (2004), pp. 1077–1098.
- [10] I. KARATZAS AND S. E. SHREVE, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York, 1991.
- [11] I. KARATZAS AND H. WANG, *A barrier option of American type*, Appl. Math. Optim., 42 (2000), pp. 259–279.
- [12] P.-L. LIONS AND A.-S. SZNITMAN, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37 (1984), pp. 511–537.
- [13] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [14] S. E. SHREVE, U. SCHMOCK, AND U. WYSTUP, *Valuation of exotic options under shortselling constraints*, Finance Stoch., 6 (2002), pp. 143–172.
- [15] S. E. SHREVE, U. SCHMOCK, AND U. WYSTUP, *Dealing with dangerous digitals*, in Foreign Exchange Risk, Risk Publications, London, 2002, pp. 327–348.

## HOMOGENEOUS APPROXIMATION, RECURSIVE OBSERVER DESIGN, AND OUTPUT FEEDBACK\*

VINCENT ANDRIEU<sup>†</sup>, LAURENT PRALY<sup>‡</sup>, AND ALESSANDRO ASTOLFI<sup>§</sup>

**Abstract.** We introduce two new tools that can be useful in nonlinear observer and output feedback design. The first one is a simple extension of the notion of homogeneous approximation to make it valid both at the origin and at infinity (homogeneity in the bi-limit). Exploiting this extension, we give several results concerning stability and robustness for a homogeneous in the bi-limit vector field. The second tool is a new recursive observer design procedure for a chain of integrator. Combining these two tools, we propose a new global asymptotic stabilization result by output feedback for feedback and feedforward systems.

**Key words.** homogeneous approximation, output feedback and observer

**AMS subject classifications.** 93B51, 93B52, 93D05, 93D15, 34D20

**DOI.** 10.1137/060675861

**1. Introduction.** The problems of designing globally convergent observers and globally asymptotically stabilizing output feedback control laws for nonlinear systems have been addressed by many authors following different routes. Many of these approaches exploit domination ideas and robustness of stability and/or convergence. In view of possibly clarifying and developing further these techniques we introduce two new tools. The first one is a simple extension of the technique of homogeneous approximation to make it valid both at the origin and at infinity. The second tool is a new recursive observer design procedure for a chain of integrator. Combining these two tools, we propose a new global asymptotic stabilization result by output feedback for feedback and feedforward systems.

To place our contribution in perspective, we consider the following system, for which we want to design a global asymptotic stabilizing output feedback:

$$(1.1) \quad \dot{x}_1 = x_2, \quad \dot{x}_2 = u + \delta_2(x_1, x_2), \quad y = x_1,$$

where (see notation (1.4))

$$(1.2) \quad \delta_2(x_1, x_2) = c_0 x_2^q + c_\infty x_2^p, \quad (c_0, c_\infty) \in \mathbb{R}^2, \quad p > q > 0.$$

In the domination's approach, the nonlinear function  $\delta_2$  is not treated per se in the design but considered as a perturbation. In this framework the output feedback controller is designed on the linear system

$$(1.3) \quad \dot{x}_1 = x_2, \quad \dot{x}_2 = u, \quad y = x_1,$$

---

\*Received by the editors November 24, 2006; accepted for publication (in revised form) February 17, 2008; published electronically June 25, 2008. The work of the first and third authors was partly supported by the Leverhulme Trust.

<http://www.siam.org/journals/sicon/47-4/67586.html>

<sup>†</sup>LAAS-CNRS, University of Toulouse, 31077 Toulouse, France (vincent.andrieu@gmail.com). This author's work was done while at Electrical and Electronic Engineering Department, Imperial College, London.

<sup>‡</sup>Centre d'Automatique et Systèmes, École des Mines de Paris, 35 Rue Saint Honoré, 77305 Fontainebleau, France (Laurent.Praly@ensmp.fr).

<sup>§</sup>Electrical and Electronic Engineering Department, Imperial College London, London, SW7 2AZ, UK (a.astolfi@ic.ac.uk), and Dipartimento di Informatica Sistemi e Produzione, University of Rome Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy.

and will be suitable for the nonlinear system (1.1), provided the global asymptotic stability obtained for the origin of the closed-loop system is robust to the nonlinear disturbance  $\delta_2$ . For instance, the design given in [13, 27] provides a linear output feedback controller which is suitable for the nonlinear system (1.1) when  $q = 1$  and  $c_\infty = 0$ . This result has been extended recently in [26] employing a homogeneous output feedback controller which allows us to deal with  $p \geq 1$  and  $c_0 = 0$ .

Homogeneity in the bi-limit and the novel recursive observer design proposed in this paper allow us to deal with the case in which  $c_0 \neq 0$  and  $c_\infty \neq 0$ . In this case, the function  $\delta_2$  is such that

1. when  $|x_2|$  is small and  $q = 1$ ,  $\delta_2(x_2)$  can be approximated by  $c_0 x_2$  and the nonlinearity can be approximated by a linear function;
2. when  $|x_2|$  is large,  $\delta_2(x_2)$  can be approximated by  $c_\infty x_2^p$ , and hence we have a polynomial growth which can be handled by a weighted homogeneous controller as in [26].

To deal with both linear and polynomial terms we introduce a generalization of weighted homogeneity which highlights the fact that a function becomes homogeneous as the state tends to the origin or to infinity but with different weights and degrees.

The paper is organized as follows. Section 2 is devoted to general properties related to homogeneity. After giving the definition of homogeneous approximation we introduce homogeneous in the bi-limit functions and vector fields (section 2.1) and list some of their properties (section 2.2). Various results concerning stability and robustness for homogeneous in the bi-limit vector fields are given in section 2.3. In section 3 we introduce a novel recursive observer design method for a chain of integrator. Section 4 is devoted to the homogeneous in the bi-limit state feedback. Finally, in section 5, using the previous tools, we establish new results on stabilization by output feedback.

### Notation.

- $\mathbb{R}_+$  denotes the set  $[0, +\infty)$ .
- For any nonnegative real number  $r$  the function  $w \mapsto w^r$  is defined as

$$(1.4) \quad w^r = \text{sign}(w) |w|^r \quad \forall w \in \mathbb{R}.$$

According to this definition,

$$(1.5) \quad \frac{dw^r}{dw} = r|w|^{r-1}, \quad w^2 = w|w|, \quad (w_1 > w_2 \text{ and } r > 0) \Rightarrow w_1^r > w_2^r.$$

- The function  $\mathfrak{H} : \mathbb{R}_+^2 \rightarrow \mathbb{R}_+$  is defined as

$$(1.6) \quad \mathfrak{H}(a, b) = \frac{a}{1+a} [1 + b].$$

- Given  $r = (r_1, \dots, r_n)^T$  in  $\mathbb{R}_+^n$  and  $\lambda$  in  $\mathbb{R}_+$ ,  $\lambda^r \diamond x = (\lambda^{r_1} x_1, \dots, \lambda^{r_n} x_n)^T$  is the dilation of a vector  $x$  in  $\mathbb{R}^n$  with weight  $r$ . Note that

$$\lambda_1^r \diamond (\lambda_2^r \diamond x) = (\lambda_1 \lambda_2)^r \diamond x.$$

- Given  $r = (r_1, \dots, r_n)^T$  in  $(\mathbb{R}_+ \setminus \{0\})^n$ ,  $|x|_r = |x_1|^{\frac{1}{r_1}} + \dots + |x_n|^{\frac{1}{r_n}}$  is the homogeneous norm with weight  $r$  and degree 1. Note that

$$|\lambda^r \diamond x|_r = \lambda |x|_r, \quad \left| \left( \frac{1}{|x|_r} \right)^r \diamond x \right|_r = 1.$$

- Given  $r$  in  $(\mathbb{R}_+ \setminus \{0\})^n$ ,  $S_r = \{x \in \mathbb{R}^n \mid |x|_r = 1\}$  is the unity homogeneous sphere. Note that each  $x$  in  $\mathbb{R}^n$  can be decomposed in polar coordinates; i.e., there exist  $\lambda$  in  $\mathbb{R}_+$  and  $\theta$  in  $S_r$  satisfying

$$(1.7) \quad x = \lambda^r \diamond \theta \quad \text{with} \quad \begin{cases} \lambda &= |x|_r, \\ \theta &= \left(\frac{1}{|x|_r}\right)^r \diamond x. \end{cases}$$

## 2. Homogeneous approximation.

**2.1. Definitions.** The use of homogeneous approximations has a long history in the study of stability of an equilibrium. It can be traced back to the Lyapunov first order approximation theorem and has been pursued by many authors; see, for example, Massera [16], Hahn [8], Hermes [9], and Rosier [29]. Similarly, this technique has been used to investigate the behavior of the solutions of dynamical systems at infinity; see, for instance, Lefschetz in [14, Chapter IX.5] and Orsi, Praly, and Mareels in [20]. In this section, we recall the definitions of homogeneous approximation at the origin and at infinity and restate and/or complete some related results.

DEFINITION 2.1 (homogeneity in the 0-limit).

- A function  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  is said to be homogeneous in the 0-limit with associated triple  $(r_0, d_0, \phi_0)$ , where  $r_0$  in  $(\mathbb{R}_+ \setminus \{0\})^n$  is the weight,  $d_0$  in  $\mathbb{R}_+$  the degree, and  $\phi_0 : \mathbb{R}^n \rightarrow \mathbb{R}$  the approximating function, if  $\phi$  is continuous,  $\phi_0$  is continuous and not identically zero, and, for each compact set  $C$  in  $\mathbb{R}^n \setminus \{0\}$  and each  $\varepsilon > 0$ , there exists  $\lambda_0$  such that

$$\max_{x \in C} \left| \frac{\phi(\lambda^{r_0} \diamond x)}{\lambda^{d_0}} - \phi_0(x) \right| \leq \varepsilon \quad \forall \quad \lambda \in (0, \lambda_0] .$$

- A vector field  $f = \sum_{i=1}^n f_i \frac{\partial}{\partial x_i}$  is said to be homogeneous in the 0-limit with associated triple  $(r_0, \mathfrak{d}_0, f_0)$ , where  $r_0$  in  $(\mathbb{R}_+ \setminus \{0\})^n$  is the weight,  $\mathfrak{d}_0$  in  $\mathbb{R}$  is the degree, and  $f_0 = \sum_{i=1}^n f_{0,i} \frac{\partial}{\partial x_i}$  is the approximating vector field, if, for each  $i$  in  $\{1, \dots, n\}$ ,  $\mathfrak{d}_0 + r_{0,i} \geq 0$  and the function  $f_i$  is homogeneous in the 0-limit with associated triple  $(r_0, \mathfrak{d}_0 + r_{0,i}, f_{0,i})$ .

This notion of local approximation of a function or of a vector field can be found in [9, 29, 2, 10].

*Example 2.2.* The function  $\delta_2 : \mathbb{R} \rightarrow \mathbb{R}$  introduced in the illustrative system (1.1) is homogeneous in the 0-limit with associated triple  $(r_0, d_0, \delta_{2,0}) = (1, q, c_0 x_2^q)$ . Furthermore, if  $q < 2$ , then the vector field  $f(x_1, x_2) = (x_2, \delta_2(x_2))$  is homogeneous in the 0-limit with associated triple

$$(2.1) \quad (r_0, \mathfrak{d}_0, f_0) = \left( (2 - q, 1), q - 1, (x_2, c_0 x_2^q) \right) .$$

DEFINITION 2.3 (homogeneity in the  $\infty$ -limit).

- A function  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  is said to be homogeneous in the  $\infty$ -limit with associated triple  $(r_\infty, d_\infty, \phi_\infty)$ , where  $r_\infty$  in  $(\mathbb{R}_+ \setminus \{0\})^n$  is the weight,  $d_\infty$  in

$\mathbb{R}_+$  is the degree, and  $\phi_\infty : \mathbb{R}^n \rightarrow \mathbb{R}$  is the approximating function, if  $\phi$  is continuous,  $\phi_\infty$  is continuous and not identically zero, and, for each compact set  $C$  in  $\mathbb{R}^n \setminus \{0\}$  and each  $\varepsilon > 0$ , there exists  $\lambda_\infty$  such that

$$\max_{x \in C} \left| \frac{\phi(\lambda^{r_\infty} \diamond x)}{\lambda^{d_\infty}} - \phi_\infty(x) \right| \leq \varepsilon \quad \forall \lambda \geq \lambda_\infty.$$

- A vector field  $f = \sum_{i=1}^n f_i \frac{\partial}{\partial x_i}$  is said to be homogeneous in the  $\infty$ -limit with associated triple  $(r_\infty, \mathfrak{d}_\infty, f_\infty)$ , where  $r_\infty$  in  $(\mathbb{R}_+ \setminus \{0\})^n$  is the weight,  $\mathfrak{d}_\infty$  in  $\mathbb{R}$  is the degree, and  $f_\infty = \sum_{i=1}^n f_{\infty,i} \frac{\partial}{\partial x_i}$  is the approximating vector field, if, for each  $i$  in  $\{1, \dots, n\}$ ,  $\mathfrak{d}_\infty + r_{\infty,i} \geq 0$  and the function  $f_i$  is homogeneous in the  $\infty$ -limit with associated triple  $(r_\infty, \mathfrak{d}_\infty + r_{\infty,i}, f_{\infty,i})$ .

*Example 2.4.* The function  $\delta_2 : \mathbb{R} \rightarrow \mathbb{R}$  given in the illustrative system (1.1) is homogeneous in the  $\infty$ -limit with associated triple  $(r_\infty, d_\infty, \delta_{2,\infty}) = (1, p, c_\infty x_2^p)$ . Furthermore, when  $p < 2$ , the vector field  $f(x_1, x_2) = (x_2, \delta_2(x_2))$  is homogeneous in the  $\infty$ -limit with associated triple

$$(2.2) \quad (r_\infty, \mathfrak{d}_\infty, f_\infty) = ((2-p, 1), p-1, (x_2, c_\infty x_2^p)).$$

**DEFINITION 2.5** (homogeneity in the bi-limit). A function  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  (or a vector field  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ) is said to be homogeneous in the bi-limit if it is homogeneous in the 0-limit and homogeneous in the  $\infty$ -limit.

*Remark 2.6.* If a function  $\phi$  (resp., a vector field  $f$ ) is homogeneous in the bi-limit, then the approximating function  $\phi_0$  or  $\phi_\infty$  (resp., the approximating vector field  $f_0$  or  $f_\infty$ ) is homogeneous in the standard sense<sup>1</sup> (with the same weight and degree).

*Example 2.7.* As a consequence of Examples 2.2 and 2.4, the vector field  $f(x_1, x_2) = (x_2, \delta_2(x_2))$  is homogeneous in the bi-limit with the associated triples given in (2.1) and (2.2) as long as  $0 < q < p < 2$ .

*Example 2.8.* The function  $x \mapsto |x|_{r_0}^{d_0} + |x|_{r_\infty}^{d_\infty}$ , where  $(d_0, d_\infty)$  are in  $\mathbb{R}_+^2$  and  $(r_0, r_\infty)$  are in  $(\mathbb{R}_+ \setminus \{0\})^{2n}$ , is homogeneous in the bi-limit with associated triples  $(r_0, d_0, |x|_{r_0}^{d_0})$  and  $(r_\infty, d_\infty, |x|_{r_\infty}^{d_\infty})$ , provided that

$$(2.3) \quad \frac{d_\infty}{r_{\infty,i}} > \frac{d_0}{r_{0,i}} \quad \forall i \in \{1, \dots, n\}.$$

*Example 2.9.* We recall (1.6) and consider two homogeneous and positive definite functions  $\phi_0 : \mathbb{R}^n \rightarrow \mathbb{R}_+$  and  $\phi_\infty : \mathbb{R}^n \rightarrow \mathbb{R}_+$  with weights  $(r_0, r_\infty)$  in  $(\mathbb{R}_+ \setminus \{0\})^{2n}$  and degrees  $(d_0, d_\infty)$  in  $(\mathbb{R}_+ \setminus \{0\})^2$ . The function  $x \mapsto \mathfrak{H}(\phi_0(x), \phi_\infty(x))$  is positive definite and homogeneous in the bi-limit with associated triples  $(r_0, d_0, \phi_0)$  and  $(r_\infty, d_\infty, \phi_\infty)$ . This way of constructing a homogeneous in the bi-limit function from two positive definite homogeneous functions is extensively used in this paper.

**2.2. Properties of homogeneous approximations.** To begin, we note that the weight and degree of a homogeneous in the 0- (resp.,  $\infty$ -) limit function are

<sup>1</sup>This is proved by noting that, for all  $x$  in  $\mathbb{R}^n$  and all  $\mu$  in  $\mathbb{R}_+ \setminus \{0\}$ ,

$$\frac{\phi_0(\mu^{r_0} \diamond x)}{\mu^{d_0}} = \frac{1}{\mu^{d_0}} \lim_{\lambda \rightarrow 0} \frac{\phi(\lambda^{r_0} \diamond (\mu^{r_0} \diamond x))}{\lambda^{d_0}} = \lim_{\lambda \rightarrow 0} \frac{\phi((\lambda\mu)^{r_0} \diamond x)}{(\lambda\mu)^{d_0}} = \phi_0(x),$$

and similarly for the homogeneous in the  $\infty$ -limit function.

not uniquely defined. Indeed, if  $\phi$  is homogeneous in the 0- (resp.,  $\infty$ -) limit with associated triple  $(r_0, d_0, \phi_0)$  (resp.,  $(r_\infty, d_\infty, \phi_\infty)$ ), then it is also homogeneous in the 0- (resp.,  $\infty$ -) limit with associated triple  $(k r_0, k d_0, \phi_0)$  (resp.,  $(k r_\infty, k d_\infty, \phi_\infty)$ ) for all  $k > 0$ . (Simply change  $\lambda$  into  $\lambda^k$ .)

It is straightforward to show that if  $\phi$  and  $\zeta$  are two functions homogeneous in the 0- (resp.,  $\infty$ -) limit, with weights  $r_{\phi,0}$  and  $r_{\zeta,0}$  (resp.,  $r_{\phi,\infty}$  and  $r_{\zeta,\infty}$ ), degrees  $d_{\phi,0}$  and  $d_{\zeta,0}$  (resp.,  $d_{\phi,\infty}$  and  $d_{\zeta,\infty}$ ), and approximating functions  $\phi_0$  and  $\zeta_0$  (resp.,  $\phi_\infty$  and  $\zeta_\infty$ ), then the following hold:

- P1: If there exists  $k \in \mathbb{R}_+$  such that  $k r_{\phi,0} = r_{\zeta,0}$  (resp.,  $k r_{\phi,\infty} = r_{\zeta,\infty}$ ), then the function  $x \mapsto \phi(x) \zeta(x)$  is homogeneous in the 0- (resp.,  $\infty$ -) limit with weight  $r_{\zeta,0}$ , degree  $k d_{\phi,0} + d_{\zeta,0}$  (resp.,  $r_{\zeta,\infty}$ ,  $k d_{\phi,\infty} + d_{\zeta,\infty}$ ) and approximating function  $x \mapsto \phi_0(x) \zeta_0(x)$  (resp.,  $x \mapsto \phi_\infty(x) \zeta_\infty(x)$ ).
- P2: If, for each  $j$  in  $\{1, \dots, n\}$ ,  $\frac{d_{\phi,0}}{r_{\phi,0,j}} < \frac{d_{\zeta,0}}{r_{\zeta,0,j}}$  (resp.,  $\frac{d_{\phi,\infty}}{r_{\phi,\infty,j}} > \frac{d_{\zeta,\infty}}{r_{\zeta,\infty,j}}$ ), then the function  $x \mapsto \phi(x) + \zeta(x)$  is homogeneous in the 0- (resp.,  $\infty$ -) limit with degree  $d_{\phi,0}$  and weight  $r_{\phi,0}$  (resp.,  $d_{\phi,\infty}$  and  $r_{\phi,\infty}$ ) and approximating function  $x \mapsto \phi_0(x)$  (resp.,  $x \mapsto \phi_\infty(x)$ ). In this case we say that the function  $\phi$  *dominates* the function  $\zeta$  in the 0-limit (resp., in the  $\infty$ -limit).
- P3: If the function  $\phi_0 + \zeta_0$  (resp.,  $\phi_\infty + \zeta_\infty$ ) is not identically zero and, for each  $j$  in  $\{1, \dots, n\}$ ,  $\frac{d_{\phi,0}}{r_{\phi,0,j}} = \frac{d_{\zeta,0}}{r_{\zeta,0,j}}$  (resp.,  $\frac{d_{\phi,\infty}}{r_{\phi,\infty,j}} = \frac{d_{\zeta,\infty}}{r_{\zeta,\infty,j}}$ ), then the function  $x \mapsto \phi(x) + \zeta(x)$  is homogeneous in the 0- (resp.,  $\infty$ -) limit with degree  $d_{\phi,0}$  and weight  $r_{\phi,0}$  (resp.,  $d_{\phi,\infty}$  and  $r_{\phi,\infty}$ ) and approximating function  $x \mapsto \phi_0(x) + \zeta_0(x)$  (resp.,  $x \mapsto \phi_\infty(x) + \zeta_\infty(x)$ ).

Some properties of the composition or inverse of functions are given in the following two propositions, the proofs of which are given in Appendices A and B.

**PROPOSITION 2.10** (composition function). *If  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $\zeta : \mathbb{R} \rightarrow \mathbb{R}$  are homogeneous in the 0- (resp.,  $\infty$ -) limit functions, with weights  $r_{\phi,0}$  and  $r_{\zeta,0}$  (resp.,  $r_{\phi,\infty}$  and  $r_{\zeta,\infty}$ ), degrees  $d_{\phi,0} > 0$  and  $d_{\zeta,0} \geq 0$  (resp.,  $d_{\phi,\infty} > 0$  and  $d_{\zeta,\infty} \geq 0$ ), and approximating functions  $\phi_0$  and  $\zeta_0$  (resp.,  $\phi_\infty$  and  $\zeta_\infty$ ), then  $\zeta \circ \phi$  is homogeneous in the 0- (resp.,  $\infty$ -) limit with weight  $r_{\phi,0}$  (resp.,  $r_{\phi,\infty}$ ), degree  $\frac{d_{\zeta,0} d_{\phi,0}}{r_{\zeta,0}}$  (resp.,  $\frac{d_{\zeta,\infty} d_{\phi,\infty}}{r_{\zeta,\infty}}$ ), and approximating function  $\zeta_0 \circ \phi_0$  (resp.,  $\zeta_\infty \circ \phi_\infty$ ).*

**PROPOSITION 2.11** (inverse function). *Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  be a bijective homogeneous in the 0- (resp.,  $\infty$ -) limit function with associated triple  $(1, d_0, \varphi_0 x^{d_0})$  with  $\varphi_0 \neq 0$  and  $d_0 > 0$  (resp.,  $(1, d_\infty, \varphi_\infty x^{d_\infty})$  with  $\varphi_\infty \neq 0$  and  $d_\infty > 0$ ). Then the inverse function  $\phi^{-1} : \mathbb{R} \rightarrow \mathbb{R}$  is a homogeneous in the 0- (resp.,  $\infty$ -) limit function with associated triple  $(1, \frac{1}{d_0}, (\frac{x}{\varphi_0})^{\frac{1}{d_0}})$  (resp.,  $(1, \frac{1}{d_\infty}, (\frac{x}{\varphi_\infty})^{\frac{1}{d_\infty}})$ .*

Despite the existence of well-known results concerning the derivative of a homogeneous function, it is not possible to say anything, in general, when dealing with homogeneity in the limit. For example, the function

$$\phi(x) = x^3 + x^2 \sin(x^2) + x^3 \sin(1/x) + x^2, \quad x \in \mathbb{R},$$

is homogeneous in the bi-limit with associated triples

$$(1, 2, x^2), \quad (1, 3, x^3).$$

However, its derivative is homogeneous in neither the 0-limit nor the  $\infty$ -limit. Nevertheless the following result holds, the proof of which is elementary.

**PROPOSITION 2.12** (integral function). *If the function  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  is homogeneous in the 0- (resp.,  $\infty$ -) limit with associated triple  $(r_0, d_0, \phi_0)$  (resp.,  $(r_\infty, d_\infty, \phi_\infty)$ ),*

then the function  $\Phi_i(x) = \int_0^{x_i} \phi(x_1, \dots, x_{i-1}, s, x_{i+1}, \dots, x_n) ds$  is homogeneous in the 0- (resp.,  $\infty$ -) limit with associated triple  $(r_0, d_0 + r_{0,i}, \Phi_{i,0})$  (resp.,  $(r_\infty, d_\infty + r_{\infty,i}, \Phi_{i,\infty})$ ), with  $\Phi_{i,0}(x) = \int_0^{x_i} \phi_0(x_1, \dots, x_{i-1}, s, x_{i+1}, \dots, x_n) ds$  (resp.,  $\Phi_{i,\infty}(x) = \int_0^{x_i} \phi_\infty(x_1, \dots, x_{i-1}, s, x_{i+1}, \dots, x_n) ds$ ).

By exploiting the definition of homogeneity in the bi-limit, it is possible to establish results which are straightforward extensions of well-known results based on the standard notion of homogeneity. These results are given as corollaries of the following key technical lemma, the proof of which is given in Appendix C.

LEMMA 2.13 (key technical lemma). *Let  $\eta : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $\gamma : \mathbb{R}^n \rightarrow \mathbb{R}_+$  be two functions homogeneous in the bi-limit, with weights  $r_0$  and  $r_\infty$ , degrees  $d_0$  and  $d_\infty$ , and approximating functions,  $\eta_0, \eta_\infty$  and  $\gamma_0, \gamma_\infty$  such that the following hold:*

$$\begin{aligned} \{x \in \mathbb{R}^n \setminus \{0\} : \gamma(x) = 0\} &\subseteq \{x \in \mathbb{R}^n : \eta(x) < 0\}, \\ \{x \in \mathbb{R}^n \setminus \{0\} : \gamma_0(x) = 0\} &\subseteq \{x \in \mathbb{R}^n : \eta_0(x) < 0\}, \\ \{x \in \mathbb{R}^n \setminus \{0\} : \gamma_\infty(x) = 0\} &\subseteq \{x \in \mathbb{R}^n : \eta_\infty(x) < 0\}. \end{aligned}$$

Then there exists a real number  $c^*$  such that, for all  $c \geq c^*$  and for all  $x$  in  $\mathbb{R}^n \setminus \{0\}$ ,

$$(2.4) \quad \eta(x) - c\gamma(x) < 0, \quad \eta_0(x) - c\gamma_0(x) < 0, \quad \eta_\infty(x) - c\gamma_\infty(x) < 0.$$

Example 2.14. To illustrate the importance of this lemma, consider, for  $(x_1, x_2)$  in  $\mathbb{R}^2$ , the functions

$$\eta(x_1, x_2) = x_1 x_2 - |x_1|^{\frac{r_1+r_2}{r_1}}, \quad \gamma(x_1, x_2) = |x_2|^{\frac{r_1+r_2}{r_2}},$$

with  $r_1 > 0$  and  $r_2 > 0$ . They are homogeneous in the standard sense, and therefore in the bi-limit, with the same weight  $r = (r_1, r_2)$  and the same degree  $d = r_1 + r_2$ . Furthermore, the function  $\gamma$  takes positive values, and for all  $(x_1, x_2)$  in  $\{(x_1, x_2) \in \mathbb{R}^2 \setminus \{0\} : \gamma(x_1, x_2) = 0\}$  we have

$$\eta(x_1, x_2) = -|x_1|^{\frac{r_1+r_2}{r_1}} < 0.$$

Thus Lemma 2.13 yields the existence of a positive real number  $c^*$  such that for all  $c \geq c^*$ , we have

$$(2.5) \quad x_1 x_2 - |x_1|^{\frac{r_1+r_2}{r_1}} - c|x_2|^{\frac{r_1+r_2}{r_2}} < 0 \quad \forall (x_1, x_2) \in \mathbb{R}^2 \setminus \{0\}.$$

This is a generalization of the procedure known as the completion of the squares in which, however, the constant  $c_1^*$  is not specified.

COROLLARY 2.15. *Let  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $\zeta : \mathbb{R}^n \rightarrow \mathbb{R}_+$  be two homogeneous in the bi-limit functions with the same weights  $r_0$  and  $r_\infty$ , degrees  $d_{\phi,0}, d_{\phi,\infty}$  and  $d_{\zeta,0}, d_{\zeta,\infty}$ , and approximating functions  $\eta_0, \phi_\infty$  and  $\zeta_0, \zeta_\infty$ . If the degrees satisfy  $d_{\phi,0} \geq d_{\zeta,0}$  and  $d_{\phi,\infty} \leq d_{\zeta,\infty}$ , and the functions  $\zeta, \zeta_0$  and  $\zeta_\infty$  are positive definite, then there exists a positive real number  $c$  satisfying*

$$\phi(x) \leq c\zeta(x) \quad \forall x \in \mathbb{R}^n.$$

*Proof.* Consider the two functions

$$\eta(x) := \phi(x) + \zeta(x), \quad \gamma(x) := \zeta(x).$$



By property P2 (or P3)<sup>2</sup> in section 2.2, they are homogeneous in the bi-limit with degrees  $d_{\zeta,0}$  and  $d_{\zeta,\infty}$ . The function  $\gamma$  and its homogeneous approximations being positive definite, all assumptions of Lemma 2.13 are satisfied. Therefore there exists a positive real number  $c$  such that

$$c\gamma(x) > \eta(x) > \phi(x) \quad \forall x \in \mathbb{R}^n \setminus \{0\}.$$

Finally, by continuity of the functions  $\phi$  and  $\zeta$  at zero, we can obtain the claim.  $\square$

**2.3. Stability and homogeneous approximation.** A very basic property of asymptotic stability is its robustness. This fact was already known to Lyapunov, who proposed his second method, in which (local) asymptotic stability of an equilibrium is established by looking at the first order approximation of the system. The case of local homogeneous approximations of higher degree has been investigated by Massera [16], Hermes [9], Rosier [29], and Kawski [12].

PROPOSITION 2.16 (see [29]). *Consider a homogeneous in the 0-limit vector field  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  with associated triple  $(r_0, \mathfrak{d}_0, f_0)$ . If the origin of the system*

$$\dot{x} = f_0(x)$$

*is locally asymptotically stable, then the origin of*

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

*is locally asymptotically stable.*

Consequently, a natural strategy to ensure local asymptotic stability of an equilibrium of a system is to design a stabilizing homogeneous control law for the homogeneous approximation in the 0-limit (see [9, 12, 5], for instance).

Example 2.17. Consider the system (1.1), with  $q = 1$  and  $p > q$ , and the linear control law

$$u = -(c_0 + 1)x_2 - x_1.$$

The closed-loop vector field is homogeneous in the 0-limit with degree  $\mathfrak{d}_0 = 0$ , weight  $(1, 1)$  (i.e., we are in the linear case), and associated vector field  $f_0(x_1, x_2) = (x_2, -x_1 - x_2)^T$ . Selecting the Lyapunov function of degree 2,

$$V_0(x_1, x_2) = \frac{1}{2}|x_1|^2 + \frac{1}{2}|x_2 + x_1|^2,$$

yields

$$\frac{\partial V_0}{\partial x}(x) f_0(x) = -|x_1|^2 - |x_2 + x_1|^2.$$

It follows, from Lyapunov's second method, that the control law locally asymptotically stabilizes the equilibrium of the system. Furthermore, local asymptotic stability is preserved in the presence of any perturbation which does not change the approximating homogeneous function, i.e., in the presence of perturbations which are dominated by the linear part (see P2 in section 2.2).

---

<sup>2</sup>If  $\phi_0(x) + \zeta_0(x) = 0$  (resp.,  $\phi_\infty(x) + \zeta_\infty(x) = 0$ ), the proof can be completed by replacing  $\zeta$  with  $2\zeta$ .

In the context of homogeneity in the  $\infty$ -limit, we have the following result.

**PROPOSITION 2.18.** *Consider a homogeneous in the  $\infty$ -limit vector field  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  with associated triple  $(r_\infty, \mathfrak{d}_\infty, f_\infty)$ . If the origin of the system*

$$\dot{x} = f_\infty(x)$$

*is globally asymptotically stable, then there exists an invariant compact subset of  $\mathbb{R}^n$ , denoted  $C_\infty$ , which is globally asymptotically stable<sup>3</sup> for the system*

$$\dot{x} = f(x).$$

The proof of the proposition is given in Appendix D.

As in the case of homogeneity in the 0-limit, this property can be used to design a feedback, ensuring boundedness of solutions.

*Example 2.19.* Consider the system (1.1) with  $0 < q < p < 2$  and the control law

$$(2.6) \quad u = -\frac{1}{2-p} x_1^{\frac{p-1}{2-p}} x_2 - x_1^{\frac{p}{2-p}} - c_\infty x_2^p - \left(x_2 + x_1^{\frac{1}{2-p}}\right)^p.$$

This control law is such that the closed-loop vector field is homogeneous in the  $\infty$ -limit with degree  $\mathfrak{d}_\infty = p - 1$ , weight  $(2 - p, 1)$ , and associated vector field  $f_\infty(x_1, x_2) = (x_2, -\frac{1}{2-p} x_1^{\frac{p-1}{2-p}} x_2 - x_1^{\frac{p}{2-p}} - (x_2 + x_1^{\frac{1}{2-p}})^p)^T$ . For the homogeneous Lyapunov function of degree 2,

$$V_\infty(x_1, x_2) = \frac{2-p}{2} |x_1|^{\frac{2}{2-p}} + \frac{1}{2} \left| x_2 + x_1^{\frac{1}{2-p}} \right|^2,$$

we get

$$\frac{\partial V_\infty}{\partial x}(x) f_\infty(x) = -|x_1|^{\frac{p+1}{2-p}} - \left| x_2 + x_1^{\frac{1}{2-p}} \right|^{p+1}.$$

It follows that the control law (2.6) guarantees boundedness of the solutions of the closed-loop system. Furthermore, boundedness of solutions is preserved in the presence of any perturbation which does not change the approximating homogeneous function in the  $\infty$ -limit, i.e., in the presence of perturbations which are negligible with respect to the dominant homogeneous part (see P2 in section 2.2).

The key step in the proof of Propositions 2.16 and 2.18 is the converse Lyapunov theorem given by Rosier in [29]. This result can also be extended to the case of homogeneity in the bi-limit.

**THEOREM 2.20** (homogeneous in the bi-limit Lyapunov functions). *Consider a homogeneous in the bi-limit vector field  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ , with associated triples  $(r_\infty, \mathfrak{d}_\infty, f_\infty)$  and  $(r_0, \mathfrak{d}_0, f_0)$  such that the origins of the systems*

$$(2.7) \quad \dot{x} = f(x), \quad \dot{x} = f_\infty(x), \quad \dot{x} = f_0(x)$$

*are globally asymptotically stable equilibria. Let  $d_{V_\infty}$  and  $d_{V_0}$  be real numbers such that  $d_{V_\infty} > \max_{1 \leq i \leq n} r_{\infty,i}$  and  $d_{V_0} > \max_{1 \leq i \leq n} r_{0,i}$ . Then there exists a  $C^1$ , positive definite, and proper function  $V : \mathbb{R}^n \rightarrow \mathbb{R}_+$  such that, for each  $i$  in  $\{1, \dots, n\}$ ,*

<sup>3</sup>See [34] for the definition of global asymptotical stability for invariant compact sets.

the function  $x \mapsto \frac{\partial V}{\partial x_i}$  is homogeneous in the bi-limit with associated triples  $(r_0, d_{V_0} - r_{0,i}, \frac{\partial V_0}{\partial x_i})$  and  $(r_\infty, d_{V_\infty} - r_{\infty,i}, \frac{\partial V_\infty}{\partial x_i})$ , and the functions  $x \mapsto \frac{\partial V}{\partial x}(x) f(x)$ ,  $x \mapsto \frac{\partial V_0}{\partial x}(x) f_0(x)$ , and  $x \mapsto \frac{\partial V_\infty}{\partial x}(x) f_\infty(x)$  are negative definite.

The proof is given in Appendix E. A direct consequence of this result is an input-to-state stability (ISS) property with respect to disturbances (see [31]). To illustrate this property, consider the system with exogenous disturbance  $\delta = (\delta_1, \dots, \delta_m)$  in  $\mathbb{R}^m$ ,

$$(2.8) \quad \dot{x} = f(x, \delta) ,$$

with  $f : \mathbb{R}^n \times \mathbb{R}^m$  a continuous vector field homogeneous in the bi-limit with associated triples  $(\mathfrak{d}_0, (r_0, \mathfrak{r}_0), f_0)$  and  $(\mathfrak{d}_\infty, (r_\infty, \mathfrak{r}_\infty), f_\infty)$ , where  $\mathfrak{r}_0$  and  $\mathfrak{r}_\infty$  in  $(\mathbb{R}_+ \setminus \{0\})^m$  are the weights associated with the disturbance  $\delta$ .

COROLLARY 2.21 (ISS property). *If the origins of the systems*

$$\dot{x} = f(x, 0), \quad \dot{x} = f_0(x, 0), \quad \dot{x} = f_\infty(x, 0)$$

*are globally asymptotically stable equilibria, then under the hypotheses of Theorem 2.20 the function  $V$  given by Theorem 2.20 satisfies,<sup>4</sup> for all  $\delta = (\delta_1, \dots, \delta_m)$  in  $\mathbb{R}^m$  and  $x$  in  $\mathbb{R}^n$ ,*

$$(2.9) \quad \begin{aligned} \frac{\partial V}{\partial x}(x) f(x, \delta) \leq & -c_V \mathfrak{H} \left( V(x)^{\frac{d_{V_0} + d_0}{d_{V_0}}}, V(x)^{\frac{d_{V_\infty} + d_\infty}{d_{V_\infty}}} \right) \\ & + c_\delta \sum_{j=1}^m \mathfrak{H} \left( |\delta_j|^{\frac{d_{V_0} + d_0}{\mathfrak{r}_{0,j}}}, |\delta_j|^{\frac{d_{V_\infty} + d_\infty}{\mathfrak{r}_{\infty,j}}} \right) , \end{aligned}$$

where  $c_V$  and  $c_\delta$  are positive real numbers.

In other words, system (2.8) with  $\delta$  as input satisfies an ISS property. The proof of this corollary is given in Appendix F.

Finally, we have also the following small-gain result for homogeneous in the bi-limit vector fields.

COROLLARY 2.22 (small-gain). *Under the hypotheses of Corollary 2.21, there exists a real number  $c_G > 0$  such that, for each class  $\mathcal{K}$  function  $\gamma_z$  and  $\mathcal{KL}$  function  $\beta_\delta$ , there exists a class  $\mathcal{KL}$  function  $\beta_x$  such that, for each function  $t \in [0, T) \mapsto (x(t), \delta(t), z(t))$ ,  $T \leq +\infty$ , with  $x$   $C^1$  and  $\delta$  and  $z$  continuous, which satisfy (2.8) on  $[0, T)$  and, for all  $0 \leq s \leq t \leq T$ ,*

$$(2.10) \quad |z(t)| \leq \max \left\{ \beta_\delta(|z(s)|, t-s), \sup_{s \leq \kappa \leq t} \gamma_z(|x(\kappa)|) \right\} ,$$

$$(2.11) \quad |\delta_i(t)| \leq \max \left\{ \beta_\delta(|z(s)|, t-s), c_G \sup_{s \leq \kappa \leq t} \left\{ \mathfrak{H}(|x(\kappa)|_{r_0}^{\mathfrak{r}_{0,i}}, |x(\kappa)|_{r_\infty}^{\mathfrak{r}_{\infty,i}}) \right\} \right\} ,$$

we have

$$(2.12) \quad |x(t)| \leq \beta_x(|(x(s), z(s))|, t-s), \quad 0 \leq s \leq t \leq T .$$

<sup>4</sup>The function  $\mathfrak{H}$  is defined in (1.6).

The proof is given in Appendix G.

*Example 2.23.* An interesting case, which can be dealt with by Corollary 2.22, is when the  $\delta_i$ 's are outputs of auxiliary systems with state  $z_i$  in  $\mathbb{R}^{n_i}$ , i.e.,

$$(2.13) \quad \delta_i(t) := \delta_i(z_i(t), x(t)), \quad \dot{z}_i = g_i(z_i, x) .$$

It can be checked that the bounds (2.11) and (2.10) are satisfied by all the solutions of (2.8) and (2.13) if there exist positive definite and radially unbounded functions  $Z_i : \mathbb{R}^{n_i} \rightarrow \mathbb{R}_+$ ; class  $\mathcal{K}$  functions  $\omega_1$ ,  $\omega_2$ , and  $\omega_3$ ; and a positive real number  $\epsilon$  in  $(0, 1)$  such that for all  $x$  in  $\mathbb{R}^n$ , for all  $i$  in  $\{1, \dots, m\}$ , and  $z_i$  in  $\mathbb{R}^{n_i}$ , we have

$$\begin{aligned} |\delta_i(z_i, x)| &\leq \omega_1(x) + \omega_2(Z_i(z_i)), & \frac{\partial Z_i}{\partial z_i}(z_i) g_i(z_i, x) &\leq -Z_i(z_i) + \omega_3(|x|) , \\ \omega_1(x) + \omega_2([1 + \epsilon] \omega_3(|x|)) &\leq c_G \mathfrak{H}(|x|_{r_0}^{r_{0,i}}, |x|_{r_\infty}^{r_{\infty,i}}) . \end{aligned}$$

Another important result exploiting Theorem 2.20 deals with finite time convergence of solutions toward a globally asymptotically stable equilibrium (see [4]). It is well known that when the origin of the homogeneous approximation in the 0-limit is globally asymptotically stable with a strictly negative degree, then solutions converge to the origin in finite time (see [3]). We extend this result by showing that if, furthermore, the origin of the homogeneous approximation in the  $\infty$ -limit is globally asymptotically stable with strictly positive degree, then the convergence time doesn't depend on the initial condition. This is expressed by the following corollary.

**COROLLARY 2.24** (uniform and finite time convergence). *Under the hypotheses of Theorem 2.20, if we have  $\mathfrak{d}_\infty > 0 > \mathfrak{d}_0$ , then all solutions of the system  $\dot{x} = f(x)$  converge in finite time to the origin, uniformly in the initial condition.*

The proof is given in Appendix H.

**3. Recursive observer design for a chain of integrators.** The notion of homogeneity in the bi-limit is instrumental in introducing a new observer design method. Throughout this section we consider a chain of integrators, with state  $\mathfrak{X}_n = (x_1, \dots, x_n)$  in  $\mathbb{R}^n$ , namely,

$$(3.1) \quad \dot{x}_1 = x_2, \dots, \dot{x}_n = u, \quad \text{or in compact form,} \quad \dot{\mathfrak{X}}_n = \mathcal{S}_n \mathfrak{X}_n + B_n u ,$$

where  $\mathcal{S}_n$  is the shift matrix of order  $n$ , i.e.,  $\mathcal{S}_n \mathfrak{X}_n = (x_2, \dots, x_n, 0)^T$  and  $B_n = (0, \dots, 0, 1)^T$ . By selecting arbitrary vector field degrees  $\mathfrak{d}_0$  and  $\mathfrak{d}_\infty$  in  $(-1, \frac{1}{n-1})$ , we see that, to possibly obtain homogeneity in the bi-limit of the associated vector field, we must choose the weights  $r_0 = (r_{0,1}, \dots, r_{0,n})$  and  $r_\infty = (r_{\infty,1}, \dots, r_{\infty,n})$  as

$$(3.2) \quad \begin{aligned} r_{0,n} &= 1, & r_{0,i} &= r_{0,i+1} - \mathfrak{d}_0 &= 1 - \mathfrak{d}_0(n-i), \\ r_{\infty,n} &= 1, & r_{\infty,i} &= r_{\infty,i+1} - \mathfrak{d}_\infty &= 1 - \mathfrak{d}_\infty(n-i). \end{aligned}$$

The goal of this section is to introduce a global homogeneous in the bi-limit observer for the system (3.1). This design follows a recursive method, which constitutes one of the main contributions of this paper.

The idea of designing an observer recursively starting from  $x_n$  and going backwards towards  $x_1$  is not new. It can be found, for instance, in [28, 26, 23, 30, 35] and in [7, Lemma 6.2.1]. Nevertheless, the procedure we propose is new and extends the results in [23, Lemmas 1 and 2] to the homogeneous in the bi-limit case.

Also, as opposed to what is proposed in [28, 26],<sup>5</sup> this observer is an exact observer (with any input  $u$ ) for a chain of integrators. The observer is given by the system<sup>6</sup>

$$(3.3) \quad \dot{\hat{\mathbf{x}}}_n = \mathcal{S}_n \hat{\mathbf{x}}_n + B_n u + K_1(\hat{x}_1 - x_1),$$

with state  $\hat{\mathbf{x}}_n = (\hat{x}_1, \dots, \hat{x}_n)$ , and where  $K_1 : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a homogeneous in the bi-limit vector field with weights  $r_0, r_\infty$  and degrees  $\mathfrak{d}_0, \mathfrak{d}_\infty$ . The output injection vector field  $K_1$  has to be selected such that the origin is a globally asymptotically stable equilibrium for the system

$$(3.4) \quad \dot{E}_1 = \mathcal{S}_n E_1 + K_1(e_1), \quad E_1 = (e_1, \dots, e_n)^T,$$

and also for its homogeneous approximations. The construction of  $K_1$  is performed via a recursive procedure whose induction argument is as follows.

Consider the system on  $\mathbb{R}^{n-i}$  given by

$$(3.5) \quad \dot{E}_{i+1} = \mathcal{S}_{n-i} E_{i+1} + K_{i+1}(e_{i+1}), \quad E_{i+1} = (e_{i+1}, \dots, e_n)^T,$$

with  $\mathcal{S}_{n-i}$  the shift matrix of order  $n-i$ , i.e.,  $\mathcal{S}_{n-i} E_{i+1} = (e_{i+2}, \dots, e_n, 0)^T$ , and  $K_{i+1} : \mathbb{R}^{n-i} \rightarrow \mathbb{R}^{n-i}$  a homogeneous in the bi-limit vector field, whose associated triples are  $((r_{0,i+1}, \dots, r_{0,n}), \mathfrak{d}_0, K_{i+1,0})$  and  $((r_{\infty,i+1}, \dots, r_{\infty,n}), \mathfrak{d}_\infty, K_{i+1,\infty})$ .

**THEOREM 3.1** (homogeneous in the bi-limit observer design). *Consider the system (3.5) and its homogeneous approximation at infinity and around the origin,*

$$\dot{E}_{i+1} = \mathcal{S}_{n-i} E_{i+1} + K_{i+1,0}(e_{i+1}), \quad \dot{E}_{i+1} = \mathcal{S}_{n-i} E_{i+1} + K_{i+1,\infty}(e_{i+1}).$$

*Suppose the origin is a globally asymptotically stable equilibrium for these systems. Then there exists a homogeneous in the bi-limit vector field  $K_i : \mathbb{R}^{n-i+1} \rightarrow \mathbb{R}^{n-i+1}$ , with associated triples  $((r_{0,i}, \dots, r_{0,n}), \mathfrak{d}_0, K_{i,0})$  and  $((r_{\infty,i}, \dots, r_{\infty,n}), \mathfrak{d}_\infty, K_{i,\infty})$ , such that the origin is a globally asymptotically stable equilibrium for the systems*

$$(3.6) \quad \begin{aligned} \dot{E}_i &= \mathcal{S}_{n-i+1} E_i + K_i(e_i), \\ \dot{E}_i &= \mathcal{S}_{n-i+1} E_i + K_{i,0}(e_i), \quad E_i = (e_i, \dots, e_n)^T, \\ \dot{E}_i &= \mathcal{S}_{n-i+1} E_i + K_{i,\infty}(e_i). \end{aligned}$$

*Proof.* We prove this result in two steps. First, we define a homogeneous in the bi-limit Lyapunov function. Then we construct the vector field  $K_i$ , depending on a parameter  $\ell$  which, if sufficiently large, renders negative definite the derivative of this Lyapunov function along the solutions of the system.

**Step 1. Definition of the Lyapunov function.** Let  $d_{W_0}$  and  $d_{W_\infty}$  be positive real numbers satisfying

$$(3.7) \quad d_{W_0} > 2 \max_{1 \leq j \leq n} r_{0,j} + \mathfrak{d}_0, \quad d_{W_\infty} > 2 \max_{1 \leq j \leq n} r_{\infty,j} + \mathfrak{d}_\infty,$$

and

$$(3.8) \quad \frac{d_{W_\infty}}{r_{\infty,i}} \geq \frac{d_{W_0}}{r_{0,i}}.$$

<sup>5</sup>Note the term  $x_i$  in (3.15) of [28], for instance.

<sup>6</sup>To simplify the presentation, we use the compact notation  $K_1(\hat{x}_1 - x_1)$  for  $K_1(\hat{x}_1 - x_1, 0, \dots, 0)$ .

The selection (3.2) implies  $r_{0,j} + \mathfrak{d}_0 > 0$  and  $r_{\infty,j} + \mathfrak{d}_\infty > 0$  for each  $j$  in  $\{1, \dots, n\}$ . Hence,

$$d_{W_0} > \max_{1 \leq j \leq n} r_{0,j}, \quad d_{W_\infty} > \max_{1 \leq j \leq n} r_{\infty,j},$$

and we can invoke Theorem 2.20 for the system (3.4) and its homogeneous approximations given in (3.5). This implies that there exists a  $C^1$ , positive definite, and proper function  $W_{i+1} : \mathbb{R}^{n-i} \rightarrow \mathbb{R}_+$  such that, for each  $j$  in  $\{i+1, \dots, n\}$ , the function  $\frac{\partial W_{i+1}}{\partial e_j}$  is homogeneous in the bi-limit with associated triples

$$\left( (r_{0,i+1}, \dots, r_{0,n}), d_{W_0} - r_{0,j}, \frac{\partial W_{i+1,0}}{\partial e_j} \right) \quad \text{and} \\ \left( (r_{\infty,i+1}, \dots, r_{\infty,n}), d_{W_\infty} - r_{\infty,j}, \frac{\partial W_{i+1,\infty}}{\partial e_j} \right).$$

Moreover, for all  $E_{i+1} \in \mathbb{R}^{n-i} \setminus \{0\}$ , we have

$$(3.9) \quad \begin{aligned} & \frac{\partial W_{i+1}}{\partial E_{i+1}}(E_{i+1}) (\mathcal{S}_{n-i} E_{i+1} + K_{i+1}(e_{i+1})) < 0, \\ & \frac{\partial W_{i+1,0}}{\partial E_{i+1}}(E_{i+1}) (\mathcal{S}_{n-i} E_{i+1} + K_{i+1,0}(e_{i+1})) < 0, \\ & \frac{\partial W_{i+1,\infty}}{\partial E_{i+1}}(E_{i+1}) (\mathcal{S}_{n-i} E_{i+1} + K_{i+1,\infty}(e_{i+1})) < 0. \end{aligned}$$

Consider the function  $q_i : \mathbb{R} \rightarrow \mathbb{R}$  defined as

$$(3.10) \quad q_i(s) = \begin{cases} \frac{r_{0,i}}{r_{0,i} + \mathfrak{d}_0} s^{\frac{r_{0,i} + \mathfrak{d}_0}{r_{0,i}}}, & |s| \leq 1, \\ \frac{r_{\infty,i}}{r_{\infty,i} + \mathfrak{d}_\infty} s^{\frac{r_{\infty,i} + \mathfrak{d}_\infty}{r_{\infty,i}}} + \frac{r_{0,i}}{r_{0,i} + \mathfrak{d}_0} - \frac{r_{\infty,i}}{r_{\infty,i} + \mathfrak{d}_\infty}, & |s| \geq 1. \end{cases}$$

Since we have  $0 < r_{0,i} + \mathfrak{d}_0$  and  $0 < r_{\infty,i} + \mathfrak{d}_\infty$ , this function is well defined and continuous on  $\mathbb{R}$ , strictly increasing and onto, and  $C^1$  on  $\mathbb{R} \setminus \{0\}$ . Furthermore, it is by construction homogeneous in the bi-limit with approximating continuous functions  $\frac{r_{0,i}}{r_{0,i} + \mathfrak{d}_0} s^{\frac{r_{0,i} + \mathfrak{d}_0}{r_{0,i}}}$  and  $\frac{r_{\infty,i}}{r_{\infty,i} + \mathfrak{d}_\infty} s^{\frac{r_{\infty,i} + \mathfrak{d}_\infty}{r_{\infty,i}}}$ . The inverse function  $q_i^{-1}$  of  $q_i$  is defined as

$$q_i^{-1}(s) = \begin{cases} \left( \frac{r_{0,i} + \mathfrak{d}_0}{r_{0,i}} s \right)^{\frac{r_{0,i}}{r_{0,i} + \mathfrak{d}_0}}, & |s| \leq \frac{r_{0,i} + \mathfrak{d}_0}{r_{0,i}}, \\ \left( \left( s - \frac{r_{0,i}}{r_{0,i} + \mathfrak{d}_0} + \frac{r_{\infty,i}}{r_{\infty,i} + \mathfrak{d}_\infty} \right) \frac{r_{\infty,i} + \mathfrak{d}_\infty}{r_{\infty,i}} \right)^{\frac{r_{\infty,i}}{r_{\infty,i} + \mathfrak{d}_\infty}}, & |s| \geq \frac{r_{0,i} + \mathfrak{d}_0}{r_{0,i}}. \end{cases}$$

By (3.8), the function

$$(3.11) \quad s \mapsto q_i^{-1}(s) \frac{d_{W_0} - r_{0,i}}{r_{0,i}} + q_i^{-1}(s) \frac{d_{W_\infty} - r_{\infty,i}}{r_{\infty,i}}$$

is homogeneous in the bi-limit with associated approximating functions  $\left( \frac{r_{0,i} + \mathfrak{d}_0}{r_{0,i}} s \right)^{\frac{d_{W_0} - r_{0,i}}{r_{0,i} + \mathfrak{d}_0}}$  and  $\left( \frac{r_{\infty,i} + \mathfrak{d}_\infty}{r_{\infty,i}} s \right)^{\frac{d_{W_\infty} - r_{\infty,i}}{r_{\infty,i} + \mathfrak{d}_\infty}}$ . Furthermore, by (3.7), it is  $C^1$  on  $\mathbb{R}$ , and its derivative is homogeneous in the bi-limit with continuous approximating functions

$$s \mapsto \frac{d_{W_0} - r_{0,i}}{r_{0,i}} \left| \frac{d_{W_0} - r_{0,i}}{r_{0,i} + \mathfrak{d}_0} s \right|^{\frac{d_{W_0} - 2r_{0,i} - \mathfrak{d}_0}{r_{0,i} + \mathfrak{d}_0}} \quad \text{and} \quad s \mapsto \frac{d_{W_\infty} - r_{\infty,i}}{r_{\infty,i}} \left| \frac{d_{W_\infty} - r_{\infty,i}}{r_{\infty,i} + \mathfrak{d}_\infty} s \right|^{\frac{d_{W_\infty} - 2r_{\infty,i} - \mathfrak{d}_\infty}{r_{\infty,i} + \mathfrak{d}_\infty}}.$$

Let  $\mathfrak{W}_i : \mathbb{R}^{n-i+1} \rightarrow \mathbb{R}_+$  be defined by

$$\begin{aligned} \mathfrak{W}_i(E_{i+1}, s) = & W_{i+1}(E_{i+1}) + \int_{q_i^{-1}(e_{i+1})}^s \left( h^{\frac{dW_0 - r_{0,i}}{r_{0,i}}} + h^{\frac{dW_\infty - r_{\infty,i}}{r_{\infty,i}}} \right) dh \\ & - \int_{q_i^{-1}(e_{i+1})}^s \left( q_i^{-1}(e_{i+1})^{\frac{dW_0 - r_{0,i}}{r_{0,i}}} + q_i^{-1}(e_{i+1})^{\frac{dW_\infty - r_{\infty,i}}{r_{\infty,i}}} \right) dh . \end{aligned}$$

This function is  $C^1$ , and by (3.8), Proposition 2.12 yields that it is homogeneous in the bi-limit with weights  $(r_{0,i+1}, \dots, r_{0,n})$  and  $(r_{\infty,i+1}, \dots, r_{\infty,n})$  for  $E_{i+1}$ ,  $r_{0,i}$  and  $r_{\infty,i}$  for  $s$ , and degrees  $d_{W_0}$  and  $d_{W_\infty}$ . Furthermore, for each  $j$  in  $\{i+1, \dots, n\}$ , the functions  $\frac{\partial \mathfrak{W}_i}{\partial e_j}(E_{i+1}, s)$  are also homogeneous in the bi-limit with the same weights and with degrees  $d_{W_0} - r_{0,j}$  and  $d_{W_\infty} - r_{\infty,j}$ .

*Step 2. Construction of the vector field  $K_i$ .* Given a positive real number  $\ell$ , we define the vector field  $K_i : \mathbb{R}^{n-i} \rightarrow \mathbb{R}^{n-i}$  as

$$K_i(e_i) = \begin{pmatrix} -q_i(\ell e_i) \\ K_{i+1}(q_i(\ell e_i)) \end{pmatrix} .$$

By Proposition 2.10 and the properties we have established for  $q_i$ ,  $K_i$  is a homogeneous in the bi-limit vector field. We show now that selecting  $\ell$  large enough yields the asymptotic stability properties. To begin with, note that for all  $E_i = (E_{i+1}, e_i)$  in  $\mathbb{R}^{n-i}$ ,

$$\frac{\partial \mathfrak{W}_i(E_{i+1}, \ell e_i)}{\partial E_i}(E_i) (\mathcal{S}_{n-i+1} E_i + K_i(e_i)) \leq T_1(E_{i+1}, \ell e_i) - \ell T_2(E_{i+1}, \ell e_i) ,$$

with the functions  $T_1$  and  $T_2$  defined as

$$\begin{aligned} T_1(E_{i+1}, \vartheta_i) &= \frac{\partial \mathfrak{W}_i}{\partial E_{i+1}}(E_{i+1}, \vartheta_i) (\mathcal{S}_{n-i} E_{i+1} + K_{i+1}(q_i(\vartheta_i))) , \\ T_2(E_{i+1}, \vartheta_i) &= \left( \vartheta_i^{\frac{dW_0 - r_{0,i}}{r_{0,i}}} - q_i^{-1}(e_{i+1})^{\frac{dW_0 - r_{0,i}}{r_{0,i}}} + \vartheta_i^{\frac{dW_\infty - r_{\infty,i}}{r_{\infty,i}}} - q_i^{-1}(e_{i+1})^{\frac{dW_\infty - r_{\infty,i}}{r_{\infty,i}}} \right) \\ &\quad \times (q_i(\vartheta_i) - e_{i+1}) . \end{aligned}$$

These functions are homogeneous in the bi-limit with weights  $(r_{\infty,i}, \dots, r_{\infty,n})$  and  $(r_{0,i}, \dots, r_{0,n})$ , degrees  $\mathfrak{d}_0 + d_{W_0}$  and  $\mathfrak{d}_\infty + d_{W_\infty}$ , and continuous approximating functions

$$\begin{aligned} T_{1,0}(E_{i+1}, \vartheta_i) &= \frac{\partial \mathfrak{W}_{i,0}}{\partial E_{i+1}}(E_{i+1}, \vartheta_i) (\mathcal{S}_{n-i} E_{i+1} + K_{i+1,0}(q_{i,0}(\vartheta_i))) , \\ T_{1,\infty}(E_{i+1}, \vartheta_i) &= \frac{\partial \mathfrak{W}_{i,\infty}}{\partial E_{i+1}}(E_{i+1}, \vartheta_i) (\mathcal{S}_{n-i} E_{i+1} + K_{i+1,\infty}(q_{i,\infty}(\vartheta_i))) , \\ T_{2,0}(E_{i+1}, \vartheta_i) &= \left( \vartheta_i^{\frac{dW_0 - r_{0,i}}{r_{0,i}}} - q_{i,0}^{-1}(e_{i+1})^{\frac{dW_0 - r_{0,i}}{r_{0,i}}} \right) (q_{i,0}(\vartheta_i) - e_{i+1}) , \end{aligned}$$

and

$$T_{2,\infty}(E_{i+1}, \vartheta_i) = \left( \vartheta_i^{\frac{dW_\infty - r_{\infty,i}}{r_{\infty,i}}} - q_{i,\infty}^{-1}(e_{i+1})^{\frac{dW_\infty - r_{\infty,i}}{r_{\infty,i}}} \right) (q_{i,\infty}(\vartheta_i) - e_{i+1}) .$$

As the function  $q_i^{-1}$  is continuous, strictly increasing and onto, the function

$$\vartheta_i^{\frac{dW_0 - r_{0,i}}{r_{0,i}}} - q_i^{-1}(e_{i+1})^{\frac{dW_0 - r_{0,i}}{r_{0,i}}} + \vartheta_i^{\frac{dW_\infty - r_{\infty,i}}{r_{\infty,i}}} - q_i^{-1}(e_{i+1})^{\frac{dW_\infty - r_{\infty,i}}{r_{\infty,i}}}$$

has a unique zero at  $q_i(\vartheta_i) = e_{i+1}$  and has the same sign as  $q_i(\vartheta_i) - e_{i+1}$ . It follows that

$$\begin{aligned} T_2(E_{i+1}, \vartheta_i) &\geq 0 & \forall (E_{i+1}, \vartheta_i) \in \mathbb{R}^{n-i}, \\ T_2(E_{i+1}, \vartheta_i) &= 0 & \Rightarrow q_i(\vartheta_i) = e_{i+1}. \end{aligned}$$

On the other hand, for all  $E_i \neq 0$ ,

$$T_1(E_{i+1}, q_i^{-1}(e_{i+1})) = \frac{\partial W_{i+1}}{\partial E_{i+1}}(E_{i+1}) (\mathcal{S}_{n-i} E_{i+1} + K_{i+1}(e_{i+1})) < 0.$$

Hence (3.9) yields

$$\begin{aligned} \{(E_{i+1}, \vartheta_i) \in \mathbb{R}^{n-i+1} \setminus \{0\} : T_2(E_{i+1}, \vartheta_i) = 0\} \\ \subseteq \{(E_{i+1}, \vartheta_i) \in \mathbb{R}^{n-i+1} : T_1(E_{i+1}, \vartheta_i) < 0\}. \end{aligned}$$

By following the same argument, it can be shown that this property holds also for the homogeneous approximations, i.e.,

$$\begin{aligned} \{(E_{i+1}, \vartheta_i) \in \mathbb{R}^{n-i+1} \setminus \{0\} : T_{2,0}(E_{i+1}, \vartheta_i) = 0\} \\ \subseteq \{(E_{i+1}, \vartheta_i) \in \mathbb{R}^{n-i+1} : T_{1,0}(E_{i+1}, \vartheta_i) < 0\}, \\ \{(E_{i+1}, \vartheta_i) \in \mathbb{R}^{n-i+1} \setminus \{0\} : T_{2,\infty}(E_{i+1}, \vartheta_i) = 0\} \\ \subseteq \{(E_{i+1}, \vartheta_i) \in \mathbb{R}^{n-i+1} : T_{1,\infty}(E_{i+1}, \vartheta_i) < 0\}. \end{aligned}$$

Therefore, by Lemma 2.13, there exists  $\ell^*$  such that, for all  $\ell \geq \ell^*$  and all  $(E_{i+1}, \vartheta_i) \neq 0$ ,

$$\begin{aligned} T_1(E_{i+1}, \vartheta_i) - \ell T_2(E_{i+1}, \vartheta_i) &< 0, \\ T_{1,0}(E_{i+1}, \vartheta_i) - \ell T_{2,0}(E_{i+1}, \vartheta_i) &< 0, \\ T_{1,\infty}(E_{i+1}, \vartheta_i) - \ell T_{2,\infty}(E_{i+1}, \vartheta_i) &< 0. \end{aligned}$$

This implies that the origin is a globally asymptotically stable equilibrium of the systems (3.6), which concludes the proof.  $\square$

To construct the function  $K_1$ , which defines the observer (3.3), it is sufficient to iterate the construction proposed in Theorem 3.1 starting from

$$K_n(e_n) = - \begin{cases} \frac{1}{1+\vartheta_0} (\ell_n e_n)^{1+\vartheta_0}, & |\ell_n e_n| \leq 1, \\ \frac{1}{1+\vartheta_\infty} (\ell_n e_n)^{1+\vartheta_\infty} + \frac{1}{1+\vartheta_0} - \frac{1}{1+\vartheta_\infty}, & |\ell_n e_n| \geq 1, \end{cases}$$

where  $\ell_n$  is any strictly positive real number. Indeed,  $K_n$  is a homogeneous in the bi-limit vector field with approximating functions  $K_{n,0}(e_n) = \frac{1}{1+\vartheta_0} (\ell_n e_n)^{1+\vartheta_0}$  and  $K_{n,\infty}(e_n) = \frac{1}{1+\vartheta_\infty} (\ell_n e_n)^{1+\vartheta_\infty}$ . This selection implies that the origin is a globally asymptotically stable equilibrium for the systems  $\dot{e}_n = K_n(e_n)$ ,  $\dot{e}_n = K_{n,0}(e_n)$ , and  $\dot{e}_n = K_{n,\infty}(e_n)$ .



Consequently the assumptions of Theorem 3.1 are satisfied for  $i + 1 = n$ . We can apply it recursively up to  $i = 1$ , obtaining the vector field  $K_1$ .

As a result of this procedure we obtain a homogeneous in the bi-limit observer, which globally asymptotically observes the state of the system (3.1), and also the state for its homogeneous approximations around the origin and at infinity. In other words, the origin is a globally asymptotically stable equilibrium of the systems

$$(3.12) \quad \dot{E}_1 = \mathcal{S}_n E_1 + K_1(e_1), \quad \dot{\bar{E}}_1 = \mathcal{S}_n E_1 + K_{1,0}(e_1), \quad \dot{\hat{E}}_1 = \mathcal{S}_n E_1 + K_{1,\infty}(e_1).$$

*Remark 3.2.* Note that when  $0 \leq \mathfrak{d}_0 \leq \mathfrak{d}_\infty$ , we have  $1 \leq \frac{r_{0,i} + \mathfrak{d}_0}{r_{0,i}} \leq \frac{r_{\infty,i} + \mathfrak{d}_\infty}{r_{\infty,i}}$  for  $i = 1, \dots, n$  and we can replace the function  $q_i$  in (3.10) by the simpler function

$$q_i(s) = s^{\frac{r_{0,i} + \mathfrak{d}_0}{r_{0,i}}} + s^{\frac{r_{\infty,i} + \mathfrak{d}_\infty}{r_{\infty,i}}},$$

which has been used already in [1].

*Example 3.3.* Consider a chain of integrators of dimension two, with the following weights and degrees:

$$(r_0, \mathfrak{d}_0) = ((2 - q, 1), q - 1), \quad (r_\infty, \mathfrak{d}_\infty) = ((2 - p, 1), p - 1).$$

When  $q \geq p$  (i.e.,  $\mathfrak{d}_0 \leq \mathfrak{d}_\infty$ ), by following the above recursive observer design we obtain two positive real numbers  $\ell_1$  and  $\ell_2$  such that the system

$$\dot{\hat{x}}_1 = \hat{x}_2 - q_1(\ell_1 e_1), \quad \dot{\hat{x}}_2 = u - q_2(\ell_2 q_1(\ell_1 e_1)), \quad e_1 = \hat{x}_1 - y,$$

with

$$(3.13) \quad q_2(s) = \begin{cases} \frac{1}{q} s^q, & |s| \leq 1, \\ \frac{1}{p} s^p + \frac{1}{q} - \frac{1}{p}, & |s| \geq 1, \end{cases} \quad q_1(s) = \begin{cases} (2 - q) s^{\frac{1}{2-q}}, & |s| \leq 1, \\ (2 - p) s^{\frac{1}{2-p}} + p - q, & |s| \geq 1, \end{cases}$$

is a global observer for the system  $\dot{x}_1 = x_2$ ,  $\dot{x}_2 = u$ ,  $y = x_1$ . Furthermore, its homogeneous approximations around the origin and at infinity are also global observers for the same system.

#### 4. Recursive design of a homogeneous in the bi-limit state feedback.

It is well known that the system (3.1) can be rendered homogeneous by using a stabilizing homogeneous state feedback which can be designed by backstepping (see [21, 25, 19, 26, 33, 10], for instance). We show in this section that this property can be extended to the case of homogeneity in the bi-limit. More precisely, we show that there exists a homogeneous in the bi-limit function  $\phi_n$  such that the system (3.1) with  $u = \phi_n(\mathfrak{X}_n)$  is homogeneous in the bi-limit, with weights  $r_0$  and  $r_\infty$  and degrees  $\mathfrak{d}_0$  and  $\mathfrak{d}_\infty$ . Furthermore, its origin and the origin of the approximating systems in the 0-limit and in the  $\infty$ -limit are globally asymptotically stable equilibria.

To design the state feedback we follow the approach of Praly and Mazenc [25]. To this end, consider the auxiliary system with state  $\mathfrak{X}_i = (x_1, \dots, x_i)$  in  $\mathbb{R}^i$ ,  $1 \leq i < n$ , and dynamics

$$(4.1) \quad \dot{x}_1 = x_2, \dots, \dot{x}_i = u \quad \text{or in compact form} \quad \dot{\mathfrak{X}}_i = \mathcal{S}_i \mathfrak{X}_i + B_i u,$$

where  $u$  is the input in  $\mathbb{R}$ ,  $\mathcal{S}_i$  is the shift matrix of order  $i$ , i.e.,  $\mathcal{S}_i \mathfrak{X}_i = (x_2, \dots, x_i, 0)^T$ , and  $B_i = (0, \dots, 1)^T$  is in  $\mathbb{R}^i$ . We show that, if there exists a homogeneous in the

bi-limit stabilizing control law for the origin of the system (4.1), then there is one for the origin of the system with state  $\mathfrak{X}_{i+1} = (x_1, \dots, x_{i+1})$  in  $\mathbb{R}^{i+1}$  defined by

$$(4.2) \quad \dot{x}_1 = x_2, \dots, \dot{x}_{i+1} = u, \text{ i.e., } \quad \dot{\mathfrak{X}}_{i+1} = \mathcal{S}_{i+1} \mathfrak{X}_{i+1} + B_{i+1} u .$$

Let  $\mathfrak{d}_0$  and  $\mathfrak{d}_\infty$  be in  $(-1, \frac{1}{n-1})$  and consider the weights and degrees defined in (3.2).

**THEOREM 4.1** (homogeneous in the bi-limit backstepping). *Suppose there exists a homogeneous in the bi-limit function  $\phi_i : \mathbb{R}^i \rightarrow \mathbb{R}$  with associated triples  $(r_0, \mathfrak{d}_0 + r_{0,i}, \phi_{i,0})$  and  $(r_\infty, \mathfrak{d}_\infty + r_{\infty,i}, \phi_{i,\infty})$  such that the following hold:*

1. *There exists  $\alpha_i \geq 1$  such that the function  $\psi_i(\mathfrak{X}_i) = \phi_i(\mathfrak{X}_i)^{\alpha_i}$  is  $C^1$  and for each  $j$  in  $\{1, \dots, i\}$  the function  $\frac{\partial \psi_i}{\partial x_j}$  is homogeneous in the bi-limit with weights  $(r_{0,1}, \dots, r_{0,i})$ ,  $(r_{\infty,1}, \dots, r_{\infty,i})$ , degrees  $\alpha_i(r_{0,i} + \mathfrak{d}_0) - r_{0,j}$ ,  $\alpha_i(r_{\infty,i} + \mathfrak{d}_\infty) - r_{\infty,j}$ , and approximating functions  $\frac{\partial \psi_{i,0}}{\partial x_j}$ ,  $\frac{\partial \psi_{i,\infty}}{\partial x_j}$ .*
2. *The origin is a globally asymptotically stable equilibrium of the systems*

(4.3)

$$\dot{\mathfrak{X}}_i = \mathcal{S}_i \mathfrak{X}_i + B_i \phi_i(\mathfrak{X}_i) , \quad \dot{\mathfrak{X}}_i = \mathcal{S}_i \mathfrak{X}_i + B_i \phi_{i,0}(\mathfrak{X}_i) , \quad \dot{\mathfrak{X}}_i = \mathcal{S}_i \mathfrak{X}_i + B_i \phi_{i,\infty}(\mathfrak{X}_i) .$$

Then there exists a homogeneous in the bi-limit function  $\phi_{i+1} : \mathbb{R}^{i+1} \rightarrow \mathbb{R}$  with associated triples  $(r_0, \mathfrak{d}_0 + r_{0,i+1}, \phi_{i+1,0})$  and  $(r_\infty, \mathfrak{d}_\infty + r_{\infty,i+1}, \phi_{i+1,\infty})$  such that the same properties hold, i.e.,

1. *there exists a real number  $\alpha_{i+1} > 1$  such that the function  $\psi_{i+1}(\mathfrak{X}_{i+1}) = \phi_{i+1}(\mathfrak{X}_{i+1})^{\alpha_{i+1}}$  is  $C^1$  and for each  $j$  in  $\{1, \dots, i+1\}$  the function  $\frac{\partial \psi_{i+1}}{\partial x_j}$  is homogeneous in the bi-limit with weights  $(r_{0,1}, \dots, r_{0,i+1})$ ,  $(r_{\infty,1}, \dots, r_{\infty,i+1})$ , degrees  $\alpha_{i+1}(r_{0,i+1} + \mathfrak{d}_0) - r_{0,j}$ ,  $\alpha_{i+1}(r_{\infty,i+1} + \mathfrak{d}_\infty) - r_{\infty,j}$ , and approximating functions  $\frac{\partial \psi_{i+1,0}}{\partial x_j}$ ,  $\frac{\partial \psi_{i+1,\infty}}{\partial x_j}$ ;*
2. *the origin is a globally asymptotically stable equilibrium of the systems*

$$(4.4) \quad \begin{aligned} \mathfrak{X}_{i+1} &= \mathcal{S}_{i+1} \mathfrak{X}_{i+1} + B_{i+1} \phi_{i+1}(\mathfrak{X}_{i+1}) , \\ \mathfrak{X}_{i+1} &= \mathcal{S}_{i+1} \mathfrak{X}_{i+1} + B_{i+1} \phi_{i+1,0}(\mathfrak{X}_{i+1}) , \\ \mathfrak{X}_{i+1} &= \mathcal{S}_{i+1} \mathfrak{X}_{i+1} + B_{i+1} \phi_{i+1,\infty}(\mathfrak{X}_{i+1}) . \end{aligned}$$

*Proof.* We prove this result in three steps. First, we construct a homogeneous in the bi-limit Lyapunov function; then we define a control law parametrized by a real number  $k$ . Finally, we show that there exists  $k$  such that the time derivative, along the trajectories of systems (4.4), of the Lyapunov function and of its approximating functions is negative definite.

*Step 1. Construction of the Lyapunov function.* Let  $d_{V_0}$  and  $d_{V_\infty}$  be positive real numbers satisfying

$$(4.5) \quad d_{V_0} > \max_{j \in \{1, \dots, n\}} \{r_{0,j}\}, \quad d_{V_\infty} > \max_{j \in \{1, \dots, n\}} \{r_{\infty,j}\} ,$$

and

$$(4.6) \quad \frac{d_{V_\infty}}{r_{\infty,i+1}} \geq \frac{d_{V_0}}{r_{0,i+1}} > 1 + \alpha_i .$$

With this selection, Theorem 2.20 gives the existence of a  $C^1$ , proper, and positive definite function  $V_i : \mathbb{R}^i \rightarrow \mathbb{R}_+$  such that, for each  $j$  in  $\{1, \dots, n\}$ , the function  $\frac{\partial V_i}{\partial x_j}$

is homogeneous in the bi-limit with weights  $(r_{0,1}, \dots, r_{0,i})$ ,  $(r_{\infty,1}, \dots, r_{\infty,i})$ , degrees  $d_{V_0} - r_{0,j}$ ,  $d_{V_\infty} - r_{\infty,j}$ , and approximating functions  $\frac{\partial V_{i,0}}{\partial x_j}$ ,  $\frac{\partial V_{i,\infty}}{\partial x_j}$ . Moreover, we have for all  $\mathfrak{X}_i \in \mathbb{R}^i \setminus \{0\}$ ,

$$(4.7) \quad \begin{aligned} \frac{\partial V_i}{\partial \mathfrak{X}_i}(\mathfrak{X}_i) [\mathcal{S}_i \mathfrak{X}_i + B_i \phi_i(\mathfrak{X}_i)] &< 0, \\ \frac{\partial V_{i,0}}{\partial \mathfrak{X}_i}(\mathfrak{X}_i) [\mathcal{S}_i \mathfrak{X}_i + B_i \phi_{i,0}(\mathfrak{X}_i)] &< 0, \\ \frac{\partial V_{i,\infty}}{\partial \mathfrak{X}_i}(\mathfrak{X}_i) [\mathcal{S}_i \mathfrak{X}_i + B_i \phi_{i,\infty}(\mathfrak{X}_i)] &< 0. \end{aligned}$$

Following [21], consider the Lyapunov function  $V_{i+1} : \mathbb{R}^{i+1} \rightarrow \mathbb{R}_+$  defined by

$$\begin{aligned} V_{i+1}(\mathfrak{X}_{i+1}) = V_i(\mathfrak{X}_i) &+ \int_{\phi_i(\mathfrak{X}_i)}^{\mathcal{X}_{i+1}} \left( h^{\frac{d_{V_0}-r_{0,i+1}}{r_{0,i+1}}} - \phi_i(\mathfrak{X}_i)^{\frac{d_{V_0}-r_{0,i+1}}{r_{0,i+1}}} \right) dh \\ &+ \int_{\phi_i(\mathfrak{X}_i)}^{\mathcal{X}_{i+1}} \left( h^{\frac{d_{V_\infty}-r_{\infty,i+1}}{r_{\infty,i+1}}} - \phi_i(\mathfrak{X}_i)^{\frac{d_{V_\infty}-r_{\infty,i+1}}{r_{\infty,i+1}}} \right) dh. \end{aligned}$$

This function is positive definite and proper. Furthermore, as  $d_{V_\infty}$  and  $d_{V_0}$  satisfy (4.6), we have

$$\frac{d_{V_\infty} - r_{\infty,i+1}}{r_{\infty,i+1}} \geq \frac{d_{V_0} - r_{0,i+1}}{r_{0,i+1}} > \alpha_i \geq 1.$$

Since the function  $\psi_i(\mathfrak{X}_i) = \phi_i(\mathfrak{X}_i)^{\alpha_i}$  is  $C^1$ , this inequality yields that the function  $V_{i+1}$  is  $C^1$ . Finally, for each  $j$  in  $\{1, \dots, n\}$ , the function  $\frac{\partial V_{i+1}}{\partial x_j}$  is homogeneous in the bi-limit with associated triples

$$\left( (r_{0,1}, \dots, r_{0,i+1}), d_{V_0} - r_{0,j}, \frac{\partial V_{i+1,0}}{\partial x_j} \right), \quad \left( (r_{\infty,1}, \dots, r_{\infty,i+1}), d_{V_\infty} - r_{\infty,j}, \frac{\partial V_{i+1,\infty}}{\partial x_j} \right).$$

*Step 2. Definition of the control law.* Recall (1.6) and consider the function  $\psi_{i+1} : \mathbb{R}^{i+1} \rightarrow \mathbb{R}$  defined by

$$\psi_{i+1}(\mathfrak{X}_{i+1}) = -k \int_0^{\mathcal{X}_{i+1}^{\alpha_{i+1}} - \phi_i(\mathfrak{X}_i)^{\alpha_i}} \mathfrak{H} \left( |s|^{\alpha_{i+1} \frac{\mathfrak{d}_0 + r_{0,i+1}}{\alpha_i r_{0,i+1}} - 1}, |s|^{\alpha_{i+1} \frac{\mathfrak{d}_\infty + r_{\infty,i+1}}{\alpha_i r_{\infty,i+1}} - 1} \right) ds,$$

where  $k$  in  $\mathbb{R}_+$  is a design parameter and  $\alpha_{i+1}$  is selected as

$$\alpha_{i+1} \geq \max \left\{ \frac{\alpha_i r_{0,i+1}}{\mathfrak{d}_0 + r_{0,i+1}}, \frac{\alpha_i r_{\infty,i+1}}{\mathfrak{d}_\infty + r_{\infty,i+1}}, 1 \right\}.$$

$\psi_{i+1}$  takes values with the same sign as  $\mathcal{X}_{i+1} - \phi_i(\mathfrak{X}_i)$ , is  $C^1$ , and, by Proposition 2.12, is homogeneous in the bi-limit. Furthermore, by Proposition 2.10, for each  $j$  in  $\{1, \dots, i+1\}$ , the function  $\frac{\partial \psi_{i+1}}{\partial x_j}$  is homogeneous in the bi-limit, with weights  $(r_{0,1}, \dots, r_{0,i+1})$ ,  $(r_{\infty,1}, \dots, r_{\infty,i+1})$ , degrees  $\alpha_{i+1}(r_{0,i+1} + \mathfrak{d}_0) - r_{0,j}$ ,  $\alpha_{i+1}(r_{\infty,i+1} + \mathfrak{d}_\infty) - r_{\infty,j}$ , and approximating functions  $\frac{\partial \psi_{i+1,0}}{\partial x_j}$ ,  $\frac{\partial \psi_{i+1,\infty}}{\partial x_j}$ . With this at hand, we choose the control law  $\phi_{i+1}$  as

$$\phi_{i+1}(\mathfrak{X}_{i+1}) = \psi_{i+1}(\mathfrak{X}_{i+1})^{\frac{1}{\alpha_{i+1}}}.$$

*Step 3. Selection of  $k$ .* Note that

$$(4.8) \quad \frac{\partial V_{i+1}}{\partial \mathfrak{X}_{i+1}}(\mathfrak{X}_{i+1}) [\mathcal{S}_{i+1} \mathfrak{X}_{i+1} + B_{i+1} \phi_{i+1}(\mathfrak{X}_{i+1})] = T_1(\mathfrak{X}_{i+1}) - k T_2(\mathfrak{X}_{i+1}) ,$$

with the functions  $T_1$  and  $T_2$  defined as

$$\begin{aligned} T_1(\mathfrak{X}_{i+1}) &= \frac{\partial V_{i+1}}{\partial \mathfrak{X}_i}(\mathfrak{X}_{i+1}) [\mathcal{S}_i \mathfrak{X}_i + B_i \mathfrak{X}_{i+1}] \\ T_2(\mathfrak{X}_{i+1}) &= \left( \mathfrak{X}_{i+1}^{\frac{dV_0 - r_{0,i+1}}{r_{0,i+1}}} - \phi_i(\mathfrak{X}_i)^{\frac{dV_0 - r_{0,i+1}}{r_{0,i+1}}} \right. \\ &\quad \left. + \mathfrak{X}_{i+1}^{\frac{dV_\infty - r_{\infty,i+1}}{r_{\infty,i+1}}} - \phi_i(\mathfrak{X}_i)^{\frac{dV_\infty - r_{\infty,i+1}}{r_{\infty,i+1}}} \right) \phi_{i+1}(\mathfrak{X}_{i+1}) . \end{aligned}$$

By the definition of homogeneity in the bi-limit and Proposition 2.10, these functions are homogeneous in the bi-limit with weights  $(r_{0,1}, \dots, r_{0,i+1})$  and  $(r_{\infty,1}, \dots, r_{\infty,i+1})$  and degrees  $dV_0 + \mathfrak{d}_0$  and  $dV_\infty + \mathfrak{d}_\infty$ . Moreover, since  $\phi_{i+1}(\mathfrak{X}_{i+1})$  has the same sign as  $\mathfrak{X}_{i+1} - \phi_i(\mathfrak{X}_i)$ ,  $T_2(\mathfrak{X}_{i+1})$  is nonnegative for all  $\mathfrak{X}_{i+1}$  in  $\mathbb{R}^{i+1}$  and, as  $\phi_{i+1}(\mathfrak{X}_{i+1}) = 0$  only if  $\mathfrak{X}_{i+1} - \phi_i(\mathfrak{X}_i) = 0$ , we get

$$\begin{aligned} T_2(\mathfrak{X}_{i+1}) = 0 &\implies \mathfrak{X}_{i+1} = \phi_i(\mathfrak{X}_i) , \\ \mathfrak{X}_{i+1} = \phi_i(\mathfrak{X}_i) &\implies T_1(\mathfrak{X}_{i+1}) = \frac{\partial V_i}{\partial \mathfrak{X}_i}(\mathfrak{X}_i) [\mathcal{S}_i \mathfrak{X}_i + B_i \phi_i(\mathfrak{X}_i)] . \end{aligned}$$

Consequently, equations (4.7) yield

$$\{\mathfrak{X}_{i+1} \in \mathbb{R}^{i+1} \setminus \{0\} : T_2(\mathfrak{X}_{i+1}) = 0\} \subseteq \{\mathfrak{X}_{i+1} \in \mathbb{R}^{i+1} : T_1(\mathfrak{X}_{i+1}) < 0\} .$$

The same implication holds for the homogeneous approximations of the two functions at infinity and around the origin, i.e.,

$$\begin{aligned} \{\mathfrak{X}_{i+1} \in \mathbb{R}^{i+1} \setminus \{0\} : T_{2,0}(\mathfrak{X}_{i+1}) = 0\} &\subseteq \{\mathfrak{X}_{i+1} \in \mathbb{R}^{i+1} : T_{1,0}(\mathfrak{X}_{i+1}) < 0\} , \\ \{\mathfrak{X}_{i+1} \in \mathbb{R}^{i+1} \setminus \{0\} : T_{2,\infty}(\mathfrak{X}_{i+1}) = 0\} &\subseteq \{\mathfrak{X}_{i+1} \in \mathbb{R}^{i+1} : T_{1,\infty}(\mathfrak{X}_{i+1}) < 0\} . \end{aligned}$$

Hence, by Lemma 2.13, there exists  $k^* > 0$  such that, for all  $k \geq k^*$ , we have for all  $\mathfrak{X}_{i+1} \neq 0$ ,

$$\begin{aligned} \frac{\partial V_{i+1}}{\partial \mathfrak{X}_{i+1}}(\mathfrak{X}_{i+1}) [\mathcal{S}_{i+1} \mathfrak{X}_{i+1} + B_{i+1} \phi_{i+1}(\mathfrak{X}_{i+1})] &< 0 , \\ \frac{\partial V_{i+1,0}}{\partial \mathfrak{X}_{i+1}}(\mathfrak{X}_{i+1}) [\mathcal{S}_{i+1} \mathfrak{X}_{i+1} + B_{i+1} \phi_{i+1,0}(\mathfrak{X}_{i+1})] &< 0 , \\ \frac{\partial V_{i+1,\infty}}{\partial \mathfrak{X}_{i+1}}(\mathfrak{X}_{i+1}) [\mathcal{S}_{i+1} \mathfrak{X}_{i+1} + B_{i+1} \phi_{i+1,\infty}(\mathfrak{X}_{i+1})] &< 0 . \end{aligned}$$

This implies that the origin is a globally asymptotically stable equilibrium of the systems (4.4).  $\square$

To construct the function  $\phi_n$  it is sufficient to iterate the construction in Theorem 4.1 starting from

$$\phi_1(x_1) = \psi_1(x_1)^{\frac{1}{\alpha_1}}, \quad \psi_1(x_1) = -k_1 \int_0^{x_1} \mathfrak{H} \left( |s|^{\alpha_1 \frac{r_{0,2}}{r_{0,1}} - 1}, |s|^{\alpha_1 \frac{r_{\infty,2}}{r_{\infty,1}} - 1} \right) ds ,$$

with  $k_1 > 0$ .

At the end of the recursive procedure, we have that the origin is a globally asymptotically stable equilibrium of the systems

$$(4.9) \quad \begin{aligned} \mathfrak{X}_n &= \mathcal{S}_n \mathfrak{X}_n + B_n \phi_n(\mathfrak{X}_n) , \\ \mathfrak{X}_n &= \mathcal{S}_n \mathfrak{X}_n + B_n \phi_{n,0}(\mathfrak{X}_n) , \\ \mathfrak{X}_n &= \mathcal{S}_n \mathfrak{X}_n + B_n \phi_{n,\infty}(\mathfrak{X}_n) . \end{aligned}$$

*Remark 4.2.* Note that if  $\mathfrak{d}_0 \geq 0$  and  $\mathfrak{d}_\infty \geq 0$ , then we can select  $\alpha_i = 1$  for all  $1 \leq i \leq n$ , and if  $\mathfrak{d}_0 \leq 0$  and  $\mathfrak{d}_\infty \geq \mathfrak{d}_0$ , then we can select  $\alpha_i = \frac{r_{0,1}}{r_{0,i+1}}$ . Finally, if  $\mathfrak{d}_\infty \leq 0$  and  $\mathfrak{d}_0 \geq \mathfrak{d}_\infty$ , then we can select  $\alpha_i = \frac{r_{\infty,1}}{r_{\infty,i+1}}$ .

*Remark 4.3.* As in the observer design, when  $\mathfrak{d}_0 \leq \mathfrak{d}_\infty$ , we have  $\frac{r_{0,i+1} + \mathfrak{d}_0}{r_{0,i+1}} \leq \frac{r_{\infty,i+1} + \mathfrak{d}_\infty}{r_{\infty,i+1}}$  for  $i = 1, \dots, n$  and we can replace the function  $\psi_i$  by the simpler function

$$(4.10) \quad \psi_{i+1}(\mathfrak{X}_{i+1}) = -k \left( |\mathfrak{X}_{i+1}^{\alpha_i} - \phi_i(\mathfrak{X}_i)|^{\alpha_{i+1} \frac{\mathfrak{d}_0 + r_{0,i+1}}{\alpha_i r_{0,i+1}}} + |\mathfrak{X}_{i+1}^{\alpha_i} - \phi_i(\mathfrak{X}_i)|^{\alpha_{i+1} \frac{\mathfrak{d}_\infty + r_{\infty,i+1}}{\alpha_i r_{\infty,i+1}}} \right) .$$

Finally, if  $0 \leq \mathfrak{d}_0 \leq \mathfrak{d}_\infty$ , then by taking  $\alpha_i = 1$  (see Remark 4.2) and  $\phi(\mathfrak{X}_{i+1}) = \psi_{i+1}(\mathfrak{X}_{i+1})$  as defined in (4.10), we recover the design in [1].

*Example 4.4.* Consider a chain of integrators of dimension two with weights and degrees

$$(r_0, \mathfrak{d}_0) = \left( (2 - q, 1), q - 1 \right), \quad (r_\infty, \mathfrak{d}_\infty) = \left( (2 - p, 1), p - 1 \right) ,$$

with  $2 > p > q > 0$ . Given  $k_1 > 0$ , using the proposed backstepping procedure we obtain a positive real number  $k_2$  such that the feedback

$$(4.11) \quad \phi_2(x_1, x_2) = -k_2 \int_0^{x_1 - \phi_1(x_1)} \mathfrak{H}(|s|^{q-1}, |s|^{p-1}) \, ds ,$$

with  $\phi_1(x_1) = -k_1 \int_0^{x_1} \mathfrak{H}(|s|^{\frac{q-1}{2-q}}, |s|^{\frac{p-1}{2-p}}) \, ds$ , renders the origin a globally asymptotically stable equilibrium of the closed-loop system. Furthermore, as a consequence of the robustness result in Corollary 2.22, there is a positive real number  $c_G$  such that, if the positive real numbers  $|c_0|$  and  $|c_\infty|$  associated with  $\delta_i$  in (1.2) are smaller than  $c_G$ , then the control law  $\phi_2$  globally asymptotically stabilizes the origin of system (1.1).

## 5. Application to nonlinear output feedback design.

**5.1. Results on output feedback.** The tools presented in the previous sections can be used to derive two new results on stabilization by output feedback for the origin of nonlinear systems. The output feedback is designed for a simple chain of integrators,

$$(5.1) \quad \dot{x} = \mathcal{S}_n x + B_n u, \quad y = x_1 ,$$

where  $x$  is in  $\mathbb{R}^n$ ,  $y$  is the output in  $\mathbb{R}$ , and  $u$  is the control input in  $\mathbb{R}$ . It is then shown to be adequate to solve the output feedback stabilization problem for the origin of systems for which this chain of integrators can be considered as the dominant part of the dynamics.

Such a domination approach has a long history. It is the cornerstone of the results in [13] (see also [27] and [24]), where a linear controller was introduced to deal with nonlinear systems. This approach has also been followed with nonlinear controllers in [22] and more recently in combination with weighted homogeneity in [35, 26, 28] and the references therein.

In the context of homogeneity in the bi-limit, we use this approach exploiting the proposed backstepping and recursive observer designs. Following the idea introduced by Qian in [26] (see also [27]), the output feedback we proposed is given by

$$(5.2) \quad \dot{\hat{\mathbf{x}}}_n = L \left( \mathcal{S}_n \hat{\mathbf{x}}_n + B_n \phi_n(\hat{\mathbf{x}}_n) + K_1(x_1 - \hat{x}_1) \right), \quad u = L^n \phi_n(\hat{\mathbf{x}}_n),$$

with  $\hat{\mathbf{x}}_n$  in  $\mathbb{R}^n$  and where  $\phi_n$  and  $K_1$  are continuous functions and  $L$  is a positive real number. Employing the recursive procedure given in sections 3 and 4, we get the following theorem, whose proof is in section 5.2.

**THEOREM 5.1.** *For all real numbers  $\mathfrak{d}_0$  and  $\mathfrak{d}_\infty$  in  $(-1, \frac{1}{n-1})$ , there exists a homogeneous in the bi-limit function  $\phi_n : \mathbb{R}^n \rightarrow \mathbb{R}$  with associated triples  $(r_0, 1 + \mathfrak{d}_0, \phi_{n,0})$  and  $(r_\infty, 1 + \mathfrak{d}_\infty, \phi_{n,\infty})$  and a homogeneous in the bi-limit vector field  $K_1 : \mathbb{R}^n \rightarrow \mathbb{R}^n$  with associated triples  $(r_0, \mathfrak{d}_0, K_{1,0})$  and  $(r_\infty, \mathfrak{d}_\infty, K_{1,\infty})$  such that for all real numbers  $L > 0$  the origin is a globally asymptotically stable equilibrium of the systems (5.1) and (5.2) and their homogeneous approximations.*

We can then apply Corollary 2.22 to get an output feedback result for nonlinear systems described by

$$(5.3) \quad \dot{x} = \mathcal{S}_n x + B_n u + \delta(t), \quad y = x_1,$$

where  $\delta : \mathbb{R}_+ \rightarrow \mathbb{R}^n$  is a continuous function related to the solutions as described in the two corollaries below and proved in section 5.2. Depending on whether  $\mathfrak{d}_0 \leq \mathfrak{d}_\infty$  or  $\mathfrak{d}_\infty \leq \mathfrak{d}_0$ , we get an output feedback result for systems in feedback or feedforward form.

**COROLLARY 5.2 (feedback form).** *If, in the design of  $\phi_n$  and  $K_1$ , we select  $\mathfrak{d}_0 \leq \mathfrak{d}_\infty$ , then for all positive real numbers  $c_0$  and  $c_\infty$  there exists a real number  $L^* > 0$  such that for every  $L$  in  $[L^*, +\infty)$ , the following holds:*

• *For every class  $\mathcal{K}$  function  $\gamma_z$  and class  $\mathcal{KL}$  function  $\beta_\delta$ , we can find two class  $\mathcal{KL}$  functions  $\beta_x$  and  $\hat{\beta}_x$  such that, for each function  $t \in [0, T) \mapsto (x(t), \hat{\mathbf{x}}_n(t), \delta(t), z(t))$ ,  $T \leq +\infty$ , with  $(x, \hat{\mathbf{x}}_n)$   $C^1$  and  $\delta$  and  $z$  continuous, which satisfies (5.3), (5.2), and for  $i$  in  $\{1, \dots, n\}$  and  $0 \leq s \leq t < T$ ,*

$$(5.4) \quad \begin{aligned} |z(t)| &\leq \max \left\{ \beta_\delta(|z(s)|, t-s), \sup_{s \leq \kappa \leq t} \gamma_z(|x(\kappa)|) \right\}, \\ |\delta_i(t)| &\leq \max \left\{ \beta_\delta(|z(s)|, t-s), \right. \\ &\quad \left. \sup_{s \leq \kappa \leq t} \left\{ c_0 \sum_{j=1}^i |x_j(\kappa)|^{\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}} + c_\infty \sum_{j=1}^i |x_j(\kappa)|^{\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}} \right\} \right\}, \end{aligned}$$

we have for all  $0 \leq s \leq t \leq T$ ,

$$|x(t)| \leq \beta_x(|(x(s), \hat{\mathbf{x}}_n(s), z(s))|, t-s), \quad |\hat{\mathbf{x}}_n(t)| \leq \hat{\beta}_x(|(x(s), \hat{\mathbf{x}}_n(s), z(s))|, t-s).$$

COROLLARY 5.3 (feedforward form). *If, in the design of  $\phi_n$  and  $K_1$ , we select  $\mathfrak{d}_\infty \leq \mathfrak{d}_0$ , then for all positive real numbers  $c_0$  and  $c_\infty$  there exists a real number  $L^* > 0$  such that for every  $L$  in  $(0, L^*]$ , the following holds:*

• *For every class  $\mathcal{K}$  function  $\gamma_z$  and class  $\mathcal{KL}$  function  $\beta_\delta$ , we can find two class  $\mathcal{KL}$  functions  $\beta_x$  and  $\beta_{\hat{x}}$  such that, for each function  $t \in [0, T) \mapsto (x(t), \hat{x}_n(t), \delta(t), z(t))$ ,  $T \leq +\infty$ , with  $(x, \hat{x}_n)$   $C^1$  and  $\delta$  and  $z$  continuous, which satisfies (5.3), (5.2), and for  $i$  in  $\{1, \dots, n\}$  and  $0 \leq s \leq t < T$ ,*

$$(5.5) \quad \begin{aligned} |z(t)| &\leq \max \left\{ \beta_\delta(|z(s)|, t-s), \sup_{s \leq \kappa \leq t} \gamma_z(|x(\kappa)|) \right\}, \\ |\delta_i(t)| &\leq \max \left\{ \beta_\delta(|z(s)|, t-s), \right. \\ &\quad \left. \sup_{s \leq \kappa \leq t} \left\{ c_0 \sum_{j=i+2}^n |x_j(\kappa)|^{\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}} + c_\infty \sum_{j=i+2}^n |x_j(\kappa)|^{\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}} \right\} \right\}, \end{aligned}$$

we have for all  $0 \leq s \leq t \leq T$ ,

$$|x(t)| \leq \beta_x(|(x(s), \hat{x}_n(s), z(s))|, t-s), \quad |\hat{x}_n(t)| \leq \beta_{\hat{x}}(|(x(s), \hat{x}_n(s), z(s))|, t-s).$$

*Example 5.4.* Following Example 2.23, we can consider the case where the  $\delta_i$ 's are outputs of auxiliary systems given in (2.13). Suppose there exist  $n$  positive definite and radially unbounded functions  $Z_i : \mathbb{R}^{n_i} \rightarrow \mathbb{R}_+$ , three class  $\mathcal{K}$  functions  $\omega_1, \omega_2, \omega_3$ , and a positive real number  $\epsilon$  in  $(0, 1)$  such that

$$|\delta_i(z_i, x)| \leq \omega_1(x) + \omega_2(Z_i(z_i)), \quad \frac{\partial Z_i}{\partial z_i}(z_i) g_i(z_i, x) \leq -Z_i(z_i) + \omega_3(|x|);$$

then, if there exist two real numbers  $\mathfrak{d}_0$  and  $\mathfrak{d}_\infty$  satisfying  $-1 < \mathfrak{d}_0 \leq \mathfrak{d}_\infty < \frac{1}{n-1}$  and

$$(5.6) \quad \omega_1(x) + \omega_2([1 + \epsilon] \omega_3(|x|)) \leq \left( \sum_{j=1}^i |x_j|^{\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}} + \sum_{j=1}^i |x_j|^{\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}} \right),$$

then Corollary 5.2 gives  $L^* > 0$  such that for all  $L$  in  $[L^*, +\infty)$ , the output feedback (5.2) is globally asymptotically stabilizing. Compared to already published results (see [13] and [26], for instance), the novelty of this case is in the simultaneous presence of the terms  $|x_j|^{\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}}$  and  $|x_j|^{\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}}$ .

On the other hand, if there exist two real numbers  $\mathfrak{d}_0$  and  $\mathfrak{d}_\infty$  satisfying  $-1 < \mathfrak{d}_\infty \leq \mathfrak{d}_0 < \frac{1}{n-1}$  and

$$\omega_1(x) + \omega_2([1 + \epsilon] \omega_3(|x|)) \leq \left( \sum_{j=i+2}^n |x_j|^{\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}} + \sum_{j=i+2}^n |x_j|^{\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}} \right),$$

then Corollary 5.3 gives  $L^* > 0$  such that for all  $L$  in  $(0, L^*]$ , the output feedback (5.2) is globally asymptotically stabilizing.

*Example 5.5.* Consider the illustrative system (1.1). The bound (5.6) gives the condition

$$(5.7) \quad 0 < q < p < 2 .$$

This is almost the least conservative condition we can obtain with the domination approach. Specifically, it is shown in [18] that, when  $p > 2$ , there is no stabilizing output feedback. However, when  $p = 2$ , (5.6) is not satisfied, although the stabilization problem is solvable (see [18]).

By Corollary 2.24, when (5.7) holds, the output feedback

$$u = L^2 \phi_2(\hat{x}_1, \hat{x}_2), \quad \begin{cases} \dot{\hat{x}}_1 &= L \hat{x}_2 - L q_1(\ell_1 e_1) , \\ \dot{\hat{x}}_2 &= \frac{u}{L} - L q_2(\ell_2 q_1(\ell_1 e_1)) , \\ e_1 &= \hat{x}_1 - y , \end{cases}$$

with  $\ell_1, \ell_2, \phi_2, q_1$ , and  $q_2$  defined in (3.13) and (4.11) and with picking  $\mathfrak{d}_0$  in  $(-1, q-1]$  and  $\mathfrak{d}_\infty$  in  $[p-1, 1)$ , globally asymptotically stabilizes the origin of the system (1.1), with  $L$  chosen sufficiently large. Furthermore, if  $\mathfrak{d}_0$  is chosen strictly negative and  $\mathfrak{d}_\infty$  strictly positive, by Corollary 2.24, convergence to the origin occurs in finite time, uniformly in the initial conditions.

*Example 5.6.* To illustrate the feedforward result consider the system<sup>7</sup>

$$\dot{x}_1 = x_2 + x_3^{\frac{3}{2}} + z^3, \quad \dot{x}_2 = x_3, \quad \dot{x}_3 = u, \quad \dot{z} = -z^4 + x_3, \quad y = x_1 .$$

For any  $\varepsilon > 0$ , there exists a class  $\mathcal{KL}$  function  $\beta_\delta$  such that

$$|z(t)|^3 \leq \max \left\{ \beta_\delta(|z(s)|, t-s), \quad (1+\varepsilon) \sup_{s \leq \kappa \leq t} |x_3(\kappa)|^{\frac{3}{4}} \right\} .$$

Therefore by letting  $\delta_1 = x_3^{\frac{3}{2}} + z^3$  we get, for all  $0 \leq s \leq t < T$  on the time of existence of the solutions,

$$|\delta_1(t)| \leq \max \left\{ \beta_\delta(|z(s)|, t-s), \quad \sup_{s \leq \kappa \leq t} (1+\varepsilon) |x_3(\kappa)|^{\frac{3}{4}} + |x_3(\kappa)|^{\frac{3}{2}} \right\} .$$

This is inequality (5.5) with  $\mathfrak{d}_0 = -\frac{1}{2}$  and  $\mathfrak{d}_\infty = \frac{1}{4}$ . Consequently, Corollary 5.3 says that it is possible to design a globally asymptotically stabilizing output feedback.

## 5.2. Proofs of output feedback results.

*Proof of Theorem 5.1.* The homogeneous in the bi-limit state feedback  $\phi_n$  and the homogeneous in the bi-limit vector field  $K_1$  involved in this feedback are obtained by following the procedures given in sections 3 and 4. They are such that the origin is a globally asymptotically stable equilibrium of the systems given in (4.9) and (3.12). To this end, as in [26], we write the dynamics of this system in the coordinates  $\hat{\mathbf{x}}_n = (\hat{x}_1, \dots, \hat{x}_n)$  and  $E_1 = (e_1, \dots, e_n)$  and in the time  $\tau$  defined by

$$(5.8) \quad e_i = \hat{x}_i - \frac{x_i}{L^{i-1}}, \quad \frac{d}{d\tau} = \frac{1}{L} \frac{d}{dt} .$$

<sup>7</sup>Recall the notation (1.4).



This yields

$$(5.9) \quad \begin{cases} \frac{d}{d\tau} \hat{\mathbf{x}}_n &= \mathcal{S}_n \hat{\mathbf{x}}_n + B_n \phi_n(\hat{\mathbf{x}}_n) + K_1(e_1), \\ \frac{d}{d\tau} E_1 &= \mathcal{S}_n E_1 + K_1(e_1) \end{cases}$$

with  $E_1 = (e_1, \dots, e_n)$ ,  $\hat{\mathbf{x}}_n = (\hat{x}_1, \dots, \hat{x}_n)$ . The right-hand side of (5.9) is a vector field which is homogeneous in the bi-limit with weights  $(r_0, r_0), (r_\infty, r_\infty)$ .

Given  $d_U > \max_j \{r_{0,j}, r_{\infty,j}\}$ , by applying Theorem 2.20 twice, we get two  $C^1$ , proper, and positive definite functions  $V : \mathbb{R}^n \rightarrow \mathbb{R}_+$  and  $W : \mathbb{R}^n \rightarrow \mathbb{R}_+$  such that for each  $i$  in  $\{1, \dots, n\}$ , the functions  $\frac{\partial V}{\partial x_i}$  and  $\frac{\partial W}{\partial e_i}$  are homogeneous in the bi-limit, with weights  $r_0$  and  $r_\infty$ , degrees  $d_U - r_{0,i}$  and  $d_U - r_{\infty,i}$ , and approximating functions  $\frac{\partial V_0}{\partial \hat{x}_j}$ ,  $\frac{\partial V_\infty}{\partial \hat{x}_j}$  and  $\frac{\partial W_0}{\partial e_j}$ ,  $\frac{\partial W_\infty}{\partial e_j}$ . Moreover, for all  $\hat{\mathbf{x}}_n \neq 0$ ,

$$(5.10) \quad \begin{aligned} \frac{\partial V}{\partial \hat{\mathbf{x}}}(\hat{\mathbf{x}}_n) [\mathcal{S}_n \hat{\mathbf{x}}_n + B_n \phi_n(\hat{\mathbf{x}}_n)] &< 0, \\ \frac{\partial V_0}{\partial \hat{\mathbf{x}}_n}(\hat{\mathbf{x}}_n) [\mathcal{S}_n \hat{\mathbf{x}}_n + B_n \phi_{n,0}(\hat{\mathbf{x}}_n)] &< 0, \\ \frac{\partial V_\infty}{\partial \hat{\mathbf{x}}_n}(\hat{\mathbf{x}}_n) [\mathcal{S}_n \hat{\mathbf{x}}_n + B_n \phi_{n,\infty}(\hat{\mathbf{x}}_n)] &< 0, \end{aligned}$$

and for all  $E_1 \neq 0$ ,

$$(5.11) \quad \begin{aligned} \frac{\partial W}{\partial E_1}(E_1) (\mathcal{S}_n E_1 + K_1(e_1)) &< 0, \\ \frac{\partial W_0}{\partial E_1}(E_1) (\mathcal{S}_n E_1 + K_{1,0}(e_1)) &< 0, \\ \frac{\partial W_\infty}{\partial E_1}(E_1) (\mathcal{S}_n E_1 + K_{1,\infty}(e_1)) &< 0. \end{aligned}$$

Consider now the Lyapunov function candidate

$$(5.12) \quad U(\hat{\mathbf{x}}_n, E_1) = V(\hat{\mathbf{x}}_n) + \mathfrak{c} W(E_1),$$

where  $\mathfrak{c}$  is a positive real number to be specified. Let

$$\begin{aligned} \eta(\hat{\mathbf{x}}_n, E_1) &= \frac{\partial V}{\partial \hat{\mathbf{x}}_n}(\hat{\mathbf{x}}_n) (\mathcal{S}_n \hat{\mathbf{x}}_n + B_n \phi_n(\hat{\mathbf{x}}_n) + K_1(e_1)), \\ \gamma(E_1) &= -\frac{\partial W}{\partial E_1}(E_1) (\mathcal{S}_n E_1 + K_1(e_1)). \end{aligned}$$

These two functions are continuous and homogeneous in the bi-limit with associated triples  $((r_0, r_0), d_U + \mathfrak{d}_0, \eta_0)$ ,  $((r_\infty, r_\infty), d_U + \mathfrak{d}_\infty, \eta_\infty)$  and  $((r_0, r_0), d_U + \mathfrak{d}_0, \gamma_0)$ ,  $((r_\infty, r_\infty), d_U + \mathfrak{d}_\infty, \gamma_\infty)$ , where  $\gamma_0, \gamma_\infty$  and  $\eta_0, \eta_\infty$  are continuous functions. Furthermore, by (5.11),  $\gamma(E_1)$  is negative definite. Hence, by (5.10), we have

$$\begin{aligned} \{(\hat{\mathbf{x}}_n, E_1) \in \mathbb{R}^{2n} \setminus \{0\} : \gamma(E_1) = 0\} &\subseteq \{(\hat{\mathbf{x}}_n, E_1) \in \mathbb{R}^{2n} : \eta(\hat{\mathbf{x}}_n, E_1) < 0\}, \\ \{(\hat{\mathbf{x}}_n, E_1) \in \mathbb{R}^{2n} \setminus \{0\} : \gamma_0(E_1) = 0\} &\subseteq \{(\hat{\mathbf{x}}_n, E_1) \in \mathbb{R}^{2n} : \eta_0(\hat{\mathbf{x}}_n, E_1) < 0\}, \\ \{(\hat{\mathbf{x}}_n, E_1) \in \mathbb{R}^{2n} \setminus \{0\} : \gamma_\infty(E_1) = 0\} &\subseteq \{(\hat{\mathbf{x}}_n, E_1) \in \mathbb{R}^{2n} : \eta_\infty(\hat{\mathbf{x}}_n, E_1) < 0\}. \end{aligned}$$

Consequently, by Lemma 2.13, there exists a positive real number  $c^*$  such that, for all  $c > c^*$  and all  $(\hat{\mathbf{x}}_n, E_1) \neq (0, 0)$ , the Lyapunov function  $U$ , defined in (5.12), satisfies

$$\begin{aligned} \frac{\partial U}{\partial \hat{\mathbf{x}}_n}(\hat{\mathbf{x}}_n, E_1) \left( \mathcal{S}_n \hat{\mathbf{x}}_n + B_n \phi_n(\hat{\mathbf{x}}_n) + K_1(e_1) \right) \\ + \frac{\partial U}{\partial E_1}(\hat{\mathbf{x}}_n, E_1)(E_1) (\mathcal{S}_n E_1 + K_1(e_1)) < 0 \end{aligned}$$

and the same holds for the homogeneous approximations in the 0-limit and in the  $\infty$ -limit; hence the claim.  $\square$

*Proof of Corollary 5.2.* We write the dynamics of the system (5.3) in the coordinates  $\hat{\mathbf{x}}_n$  and  $E_1$  and in the time  $\tau$  given in (5.8). This yields

$$(5.13) \quad \begin{cases} \frac{d}{d\tau} \hat{\mathbf{x}}_n &= \mathcal{S}_n \hat{\mathbf{x}}_n + B_n \phi_n(\hat{\mathbf{x}}_n) + K_1(e_1), \\ \frac{d}{d\tau} E_1 &= \mathcal{S}_n E_1 + K_1(e_1) + \mathfrak{D}(L) \end{cases}$$

with

$$\mathfrak{D}(L) = \left( \frac{\delta_1}{L}, \dots, \frac{\delta_n}{L^n} \right).$$

We denote the solution of this system, starting from  $(\hat{\mathbf{x}}_n(0), E_1(0))$  in  $\mathbb{R}^{2n}$  at time  $\tau$ , by  $(\hat{\mathbf{x}}_{\tau,n}(\tau), E_{\tau,1}(\tau))$ . We have

$$(5.14) \quad x_i(t) = L^{i-1} (\hat{x}_{\tau,i}(Lt) - e_{\tau,i}(Lt)).$$

The right-hand side of (5.13) is a vector field which is homogeneous in the bi-limit with weights  $(r_0, r_0), (r_\infty, r_\infty)$  for  $(\hat{\mathbf{x}}_n, E_1)$  and  $(\mathfrak{r}_0, \mathfrak{r}_\infty)$  for  $\mathfrak{D}(L)$ , where  $\mathfrak{r}_{0,i} = r_{0,i} + \mathfrak{d}_0$  and  $\mathfrak{r}_{\infty,i} = r_{\infty,i} + \mathfrak{d}_\infty$  for each  $i$  in  $\{1, \dots, n\}$ .

The time function  $\tau \mapsto \delta(\frac{\tau}{L})$  is considered as an input, and when  $\mathfrak{D}(L) = 0$ , Theorem 5.1 implies global asymptotic stability of the origin of the system (5.13) and of its homogeneous approximations. To complete the proof we show that there exists  $L^*$  such that the “input”  $\mathfrak{D}(L)$  satisfies the small-gain condition (2.11) of Corollary 2.22 for all  $L > L^*$ . Using (5.8) and (5.14), assumption (5.4) becomes, for all  $0 \leq \sigma \leq \tau < LT$  and all  $i$  in  $\{1, \dots, n\}$ ,

$$(5.15) \quad \begin{aligned} \left| \frac{\delta_i(\frac{\tau}{L})}{L^i} \right| &\leq \max \left\{ \frac{1}{L^i} \beta_\delta \left( \left| z\left(\frac{\sigma}{L}\right) \right|, \frac{\tau - \sigma}{L} \right), \right. \\ &\quad L^{-i} \sup_{\sigma \leq \kappa \leq \tau} \left\{ c_0 \sum_{j=1}^i \left| L^{(j-1)} (\hat{x}_{\tau,j}(\kappa) - e_{\tau,j}(\kappa)) \right|^{\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}} \right. \\ &\quad \left. \left. + c_\infty \sum_{j=1}^i \left| L^{(j-1)} (\hat{x}_{\tau,j}(\kappa) - e_{\tau,j}(\kappa)) \right|^{\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}} \right\} \right\}. \end{aligned}$$

Note that when  $1 \leq j \leq i \leq n$ , the function  $s \mapsto \frac{1-(n-i-1)s}{1-(n-j)s}$  is strictly increasing, mapping  $(-1, \frac{1}{n-1})$  in  $(\frac{n-i}{n+1-j}, \frac{i}{j-1})$ . As  $\mathfrak{d}_0 \leq \mathfrak{d}_\infty < \frac{1}{n-1}$ , we have for all  $1 \leq j \leq i \leq n$ ,

$$\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)} \leq \frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)} < \frac{i}{j-1}.$$

Hence, selecting  $L \geq 1$ , there exists a real number  $\epsilon > 0$  such that

$$L^{-\epsilon} \geq L^{(j-1)\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_0(n-j)}-i} \geq L^{(j-1)\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}-i}.$$

This implies

$$\begin{aligned} \frac{|\delta_i(\frac{\tau}{L})|}{L^i} &\leq \max \left\{ \frac{1}{L^i} \beta_\delta \left( \left| z \left( \frac{\sigma}{L} \right) \right|, \frac{\tau - \sigma}{L} \right), \right. \\ &\quad \left. L^{-\epsilon} \sup_{\sigma \leq \kappa \leq \tau} \left\{ c_0 \sum_{j=1}^i |(\hat{\chi}_{\tau,j}(\kappa) - e_{\tau,j}(\kappa))|^{\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}} \right. \right. \\ &\quad \left. \left. + c_\infty \sum_{j=1}^i |(\hat{\chi}_{\tau,j}(\kappa) - e_{\tau,j}(\kappa))|^{\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}} \right\} \right\}. \end{aligned}$$

On the other hand, the function

$$(\hat{\mathfrak{X}}_n, E_1) \mapsto c_0 \sum_{j=1}^i |\hat{\chi}_j - e_j|^{\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}} + c_\infty \sum_{j=1}^i |\hat{\chi}_j - e_j|^{\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}}$$

is homogeneous in the bi-limit with weights  $(r_0, r_0)$  and  $(r_\infty, r_\infty)$  and degrees  $1 - \mathfrak{d}_0(n-i-1) = r_{0,i} + \mathfrak{d}_0$  and  $1 - \mathfrak{d}_\infty(n-i-1) = r_{\infty,i} + \mathfrak{d}_\infty$  (see (3.2)). Hence, by Corollary 2.15, there exists a positive real number  $c_1$  such that

$$\begin{aligned} c_0 \sum_{j=1}^i |\hat{\chi}_j - e_j|^{\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}} + c_\infty \sum_{j=1}^i |\hat{\chi}_j - e_j|^{\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}} \\ (5.16) \quad \leq c_1 \mathfrak{H} \left( |(\hat{\mathfrak{X}}_n, E_1)|_{(r_0, r_0)}^{\mathfrak{d}_0 + r_{0,i}}, |(\hat{\mathfrak{X}}_n, E_1)|_{(r_\infty, r_\infty)}^{\mathfrak{d}_\infty + r_{\infty,i}} \right). \end{aligned}$$

Hence, by Corollary 2.22 (applied in the  $\tau$  time-scale), there exists  $c_G$  such that for any  $L^*$  large enough such that  $c_1 L^{*- \epsilon} \leq c_G$ , the conclusion holds.  $\square$

*Proof of Corollary 5.3.* The proof is similar to the previous one with the only difference being that, when  $i$  and  $j$  satisfy  $3 \leq i+2 \leq j \leq n$ , the function  $s \mapsto \frac{1-(n-i-1)s}{1-(n-j)s}$  is strictly decreasing, mapping  $(-1, \frac{1}{n-1})$  in  $(\frac{i}{j-1}, \frac{n-i}{n+1-j})$ . Moreover the condition  $-1 < \mathfrak{d}_\infty \leq \mathfrak{d}_0 < \frac{1}{n-1}$  gives the inequalities

$$\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)} \geq \frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)} > \frac{i}{j-1}.$$

Hence (5.16) holds, and by selecting  $L < 1$  we obtain the existence of a positive real number  $\epsilon$  such that

$$L^\epsilon \geq L^{(j-1)\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}-i} \geq L^{(j-1)\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}-i}.$$

From (5.5), this yields, for all  $0 \leq \sigma \leq \tau < LT$  and all  $i$  in  $\{1, \dots, n\}$ ,

$$\begin{aligned} \frac{|\delta_i(\frac{\tau}{L})|}{L^i} &\leq \max \left\{ \frac{1}{L^i} \beta_\delta \left( \left| z \left( \frac{\sigma}{L} \right) \right|, \frac{\tau - \sigma}{L} \right), \right. \\ &\quad L^\epsilon \sup_{\sigma \leq \kappa \leq \tau} \left\{ c_0 \sum_{j=i+2}^n |(\hat{\chi}_{\tau,j}(\kappa) - e_{\tau,j}(\kappa))|^{\frac{1-\mathfrak{d}_0(n-i-1)}{1-\mathfrak{d}_0(n-j)}} \right. \\ &\quad \left. \left. + c_\infty \sum_{j=i+2}^n |(\hat{\chi}_{\tau,j}(\kappa) - e_{\tau,j}(\kappa))|^{\frac{1-\mathfrak{d}_\infty(n-i-1)}{1-\mathfrak{d}_\infty(n-j)}} \right\} \right\}. \end{aligned}$$

From Corollary 2.22, the result holds for all  $L^*$  small enough to satisfy  $c_1 L^{*\varepsilon} \leq c_G$ .  $\square$

**6. Conclusion.** We have presented two new tools that can be useful in nonlinear control design. The first one is introduced to formalize the notion of homogeneous approximation valid both at the origin and at infinity. With this formalism we have given several novel results concerning asymptotic stability, robustness analysis, and also finite time convergence (uniformly in the initial conditions). The second one is a new recursive design for an observer for a chain of integrators. The combination of these two tools allows us to obtain a new result on stabilization by output feedback for systems whose dominant homogeneous in the bi-limit part is a chain of integrators.

**Appendix A. Proof of Proposition 2.10.** We give the proof only in the 0-limit case since the  $\infty$ -limit case is similar. Let  $C$  be an arbitrary compact subset of  $\mathbb{R}^n \setminus \{0\}$  and  $\epsilon$  any strictly positive real number. By the definition of homogeneity in the 0-limit, there exists  $\lambda_1 > 0$  such that we have

$$\left| \frac{\phi(\lambda^{r_{\phi,0}} \diamond x)}{\lambda^{d_{\phi,0}}} - \phi_0(x) \right| \leq 1 \quad \forall x \in C, \quad \forall \lambda \in (0, \lambda_1].$$

Hence, as  $\phi_0$  is a continuous function on  $\mathbb{R}^n$ , for all  $\lambda$  in  $(0, \lambda_1]$ , the function  $x \mapsto \frac{\phi(\lambda^{r_{\phi,0}} \diamond x)}{\lambda^{d_{\phi,0}}}$  takes its values in a compact set  $C_\phi = \phi_0(C) + B_1$ , where  $B_1$  is the unity ball.

Now, as  $\zeta_0$  is continuous on the compact subset  $C_\phi$ , it is uniformly continuous; i.e., there exists  $\nu > 0$  such that

$$|z_1 - z_2| < \nu \quad \implies \quad |\zeta_0(z_1) - \zeta_0(z_2)| < \epsilon.$$

Also there exists  $\mu_\epsilon > 0$  satisfying

$$\left| \frac{\zeta(\mu^{r_{\zeta,0}} z)}{\mu^{d_{\zeta,0}}} - \zeta_0(z) \right| \leq \epsilon \quad \forall z \in C_\phi, \quad \forall \mu \in (0, \mu_\epsilon],$$

or equivalently, since  $d_{\phi,0} > 0$ ,

$$\left| \frac{\zeta(\lambda^{d_{\phi,0}} z)}{\lambda^{\frac{d_{\phi,0} d_{\zeta,0}}{r_{\zeta,0}}}} - \zeta_0(z) \right| \leq \epsilon \quad \forall z \in C_\phi, \quad \forall \lambda \in \left(0, \mu_\epsilon^{\frac{r_{\zeta,0}}{d_{\phi,0}}}\right].$$

Similarly, there exists  $\lambda_\nu$  such that

$$\left| \frac{\phi(\lambda^{r_{\phi,0}} \diamond x)}{\lambda^{d_{\phi,0}}} - \phi_0(x) \right| \leq \nu \quad \forall x \in C, \quad \forall \lambda \in (0, \lambda_\nu].$$

It follows that

$$\begin{aligned} \left| \frac{\zeta(\phi(\lambda^{r_{\phi,0}} \diamond x))}{\lambda^{\frac{d_{\phi,0} d_{\zeta,0}}{r_{\zeta,0}}}} - \zeta_0(\phi_0(x)) \right| &\leq \left| \frac{\zeta(\phi(\lambda^{r_{\phi,0}} \diamond x))}{\lambda^{\frac{d_{\phi,0} d_{\zeta,0}}{r_{\zeta,0}}}} - \zeta_0\left(\frac{\phi(\lambda^{r_{\phi,0}} \diamond x)}{\lambda^{d_{\phi,0}}}\right) \right| \\ &\quad + \left| \zeta_0\left(\frac{\phi(\lambda^{r_{\phi,0}} \diamond x)}{\lambda^{d_{\phi,0}}}\right) - \zeta_0(\phi_0(x)) \right| \\ &\leq 2\epsilon \quad \forall x \in C, \quad \forall \lambda \in \min\left\{\lambda_1, \lambda_\nu, \mu_\epsilon^{\frac{r_{\zeta,0}}{d_{\phi,0}}}\right\}. \end{aligned}$$

This establishes homogeneity in the 0-limit of the function  $\zeta \circ \phi$ .

**Appendix B. Proof of Proposition 2.11.** We give the proof only in the 0-limit case since the  $\infty$ -limit case is similar. The function  $\phi$  being a bijection, we can assume without loss of generality that it is a strictly increasing function (otherwise we take  $-\phi$ ). This, together with homogeneity in the 0-limit, implies that  $\varphi_0$  is strictly positive. Moreover, for each  $\delta > 0$ , there exists  $t_0(\delta) > 0$  such that

$$\left| \frac{\phi(t)}{t^{d_0}} - \varphi_0 \right| \leq \delta \quad \forall t \in (0, t_0(\delta)) .$$

By letting  $\lambda = \phi(t)$ , this gives

$$\varphi_0 - \delta \leq \frac{\lambda}{\phi^{-1}(\lambda)^{d_0}} \leq \varphi_0 + \delta \quad \forall \lambda \in (0, \phi(t_0(\delta))) , \quad \forall \delta > 0 .$$

Since for  $\delta < \varphi_0$  the term on the left is strictly positive, these inequalities give

$$\left( \frac{1}{\varphi_0 + \delta} \right)^{\frac{1}{d_0}} \leq \frac{\phi^{-1}(\lambda)}{\lambda^{\frac{1}{d_0}}} \leq \left( \frac{1}{\varphi_0 - \delta} \right)^{\frac{1}{d_0}} \quad \forall \lambda \in (0, \phi^{-1}(t_0(\delta))) , \quad \forall \delta \in (0, \varphi_0) .$$

Then since the function  $\delta \mapsto \left( \frac{1}{\varphi_0 - \delta} \right)^{\frac{1}{d_0}}$  is continuous at zero, for every  $\epsilon_1 > 0$  there exists  $\delta_1(\epsilon_1) > 0$  satisfying

$$\left( \frac{1}{\varphi_0} \right)^{\frac{1}{d_0}} - \epsilon_1 \leq \left( \frac{1}{\varphi_0 + \delta_1(\epsilon_1)} \right)^{\frac{1}{d_0}} \leq \left( \frac{1}{\varphi_0 - \delta_1(\epsilon_1)} \right)^{\frac{1}{d_0}} \leq \left( \frac{1}{\varphi_0} \right)^{\frac{1}{d_0}} + \epsilon_1 .$$

This yields

$$\left| \frac{\phi^{-1}(\lambda)}{\lambda^{\frac{1}{d_0}}} - \left( \frac{1}{\varphi_0} \right)^{\frac{1}{d_0}} \right| \leq \epsilon_1 \quad \forall \lambda \in (0, \lambda_-(\epsilon_1)) ,$$

with  $\lambda_-(\epsilon_1) = \phi(t_0(\delta_1(\epsilon_1)))$ . With a similar argument, we get

$$\left| \frac{\phi^{-1}(-\lambda)}{\lambda^{\frac{1}{d_0}}} + \left( \frac{1}{\varphi_0} \right)^{\frac{1}{d_0}} \right| \leq \epsilon_1 \quad \forall \lambda \in (0, \lambda_+(\epsilon_1))$$

for some  $\lambda_+ > 0$ . Let  $\lambda_0 = \min\{\lambda_-, \lambda_+\}$ .

Now, for  $x \neq 0$  and  $\lambda > 0$ , we have

$$\left| \frac{\phi^{-1}(\lambda x)}{\lambda^{\frac{1}{d_0}}} - \left( \frac{x}{\varphi_0} \right)^{\frac{1}{d_0}} \right| = |x|^{\frac{1}{d_0}} \left| \frac{\phi^{-1}(\lambda x)}{(\lambda x)^{\frac{1}{d_0}}} - \left( \frac{1}{\varphi_0} \right)^{\frac{1}{d_0}} \right| .$$

Therefore, for any compact set  $C$  of  $\mathbb{R} \setminus \{0\}$  and any  $\epsilon > 0$ , by letting  $\epsilon_1 = \frac{\epsilon}{\max_{x \in C} |x|^{\frac{1}{d_0}}}$ , we have

$$|x|^{\frac{1}{d_0}} \epsilon_1 \leq \epsilon, \quad 0 < |\lambda x| \leq \lambda_0(\epsilon_1) \quad \forall \lambda \in \left( 0, \frac{\lambda_0(\epsilon_1)}{\max_{x \in C} |x|} \right] , \quad \forall x \in C ,$$

and therefore

$$\left| \frac{\phi^{-1}(\lambda x)}{\lambda^{\frac{1}{d_0}}} - \left( \frac{x}{\varphi_0} \right)^{\frac{1}{d_0}} \right| \leq \epsilon \quad \forall \lambda \in \left( 0, \frac{\lambda_0(\epsilon_1)}{\max_{x \in C} |x|} \right] , \quad \forall x \in C .$$

This establishes homogeneity in the 0-limit of the function  $\phi^{-1}$ .

**Appendix C. Proof of Lemma 2.13.** The proof of this lemma is divided into three parts.

1. We first show, by contradiction, that there exists a real number  $c_0$  satisfying

$$\eta_0(\theta) - c\gamma_0(\theta) < 0 \quad \forall \theta \in S_{r_0}, \quad \forall c \geq c_0.$$

Suppose there is no such  $c_0$ . This means there is a sequence  $(\theta_i)_{i \in \mathbb{N}}$  in  $S_{r_0}$  which satisfies

$$\eta_0(\theta_i) - i\gamma_0(\theta_i) \geq 0 \quad \forall i \in \mathbb{N}.$$

The sequence  $(\theta_i)_{i \in \mathbb{N}}$  lives in a compact set. Thus we can extract a convergent subsequence  $(\theta_{i_\ell})_{\ell \in \mathbb{N}}$  which converges to a point denoted  $\theta_\infty$ .

As the functions  $\eta_0$  and  $\gamma_0$  are bounded on  $S_{r_0}$  and  $\gamma_0$  takes nonnegative values,<sup>8</sup>  $\gamma_0(\theta_{i_\ell})$  must go to 0 as  $i_\ell$  goes to infinity. Since the functions  $\eta_0$  and  $\gamma_0$  are continuous, we get  $\gamma_0(\theta_\infty) = 0$  and  $\eta_0(\theta_\infty) \geq 0$ , which is impossible. Consequently, there exist  $c_0$  and  $\varepsilon_0 > 0$  such that

$$(C.1) \quad \eta_0(\theta) - c\gamma_0(\theta) \leq -\varepsilon_0 < 0 \quad \forall \theta \in S_{r_0}, \quad \forall c \geq c_0.$$

Moreover, since the functions  $\eta_0$  and  $\gamma_0$  are homogeneous in the standard sense (see Remark 2.6), we have the second inequality in (2.4).

Following the same argument, we can find positive real numbers  $c_\infty$  and  $\varepsilon_\infty$  such that

$$(C.2) \quad \eta_\infty(\theta) - c\gamma_\infty(\theta) < -\varepsilon_\infty \quad \forall \theta \in S_{r_\infty}, \quad \forall c \geq c_\infty,$$

and the third inequality in (2.4) holds.

In the rest of the proof, let

$$c_1 = \max\{c_0, c_\infty\}, \quad \varepsilon_1 = \min\{\varepsilon_0, \varepsilon_\infty\}.$$

2. Since  $\eta$  and  $\gamma$  are homogeneous in the 0-limit, there exists  $\lambda_0$  such that, for all  $\lambda \in (0, \lambda_0]$  and all  $\theta \in S_{r_0}$ , we have

$$\eta(\lambda^{r_0} \diamond \theta) \leq \lambda^{d_0} \eta_0(\theta) + \lambda^{d_0} \frac{\varepsilon_1}{4}, \quad \lambda^{d_0} \gamma_0(\theta) - \lambda^{d_0} \frac{\varepsilon_1}{4c_1} \leq \gamma(\lambda^{r_0} \diamond \theta),$$

which readily gives

$$\eta(\lambda^{r_0} \diamond \theta) - c_1 \gamma(\lambda^{r_0} \diamond \theta) \leq \lambda^{d_0} \eta_0(\theta) + \lambda^{d_0} \frac{\varepsilon_1}{2} - c_1 \lambda^{d_0} \gamma_0(\theta).$$

Using (C.1), we get

$$\eta(\lambda^{r_0} \diamond \theta) - c_1 \gamma(\lambda^{r_0} \diamond \theta) \leq -\lambda^{d_0} \frac{\varepsilon_1}{2} \quad \forall \lambda \in (0, \lambda_0], \quad \forall \theta \in S_{r_0},$$

and therefore, since  $\gamma$  takes nonnegative values,

$$\eta(\lambda^{r_0} \diamond \theta) - c\gamma(\lambda^{r_0} \diamond \theta) \leq -\lambda^{d_0} \frac{\varepsilon_1}{2} \quad \forall \lambda \in (0, \lambda_0], \quad \forall \theta \in S_{r_0}, \quad \forall c \geq c_1.$$

---

<sup>8</sup>Indeed, if we had  $\gamma_0(x) < 0$  for some  $x$  in  $\mathbb{R}^n \setminus \{0\}$ , by letting  $\epsilon = -\frac{\gamma_0(x)}{2}$ , the homogeneity in the 0-limit of  $\gamma$  would give a real number  $\lambda > 0$  satisfying  $\frac{\gamma(\lambda^{r_0} \diamond x)}{\lambda^{d_0}} \leq \gamma_0(x) + \epsilon = \frac{\gamma_0(x)}{2} < 0$ . This contradicts the fact that  $\gamma$  takes nonnegative values only. Also by continuity we have  $\gamma_0(0) \geq 0$ .

Similarly, there exists  $\lambda_\infty$  satisfying

$$\eta(\lambda^{r_\infty} \diamond \theta) - c \gamma(\lambda^{r_\infty} \diamond \theta) \leq -\lambda^{d_\infty} \frac{\varepsilon_1}{2} \quad \forall \lambda \in [\lambda_\infty, +\infty), \forall \theta \in S_{r_\infty}, \forall c \geq c_1.$$

Consequently, for each  $c \geq c_1$ , the set

$$\{x \in \mathbb{R}^n \setminus \{0\} \mid \eta(x) - c \gamma(x) \geq 0\},$$

if not empty, must be a subset of

$$C = \{x \in \mathbb{R}^n : |x|_{r_0} \geq \lambda_0\} \cup \{x \in \mathbb{R}^n : |x|_{r_\infty} \leq \lambda_\infty\},$$

which is compact and does not contain the origin.

3. Suppose now that for all  $c$  the first inequality in (2.4) is not true. This means that, for all integers  $c$  larger than  $c_1$ , there exists  $x_c$  in  $\mathbb{R}^n$  satisfying

$$\eta(x_c) - c \gamma(x_c) \geq 0,$$

and therefore  $x_c$  is in  $C$ . Since  $C$  is a compact set, there is a convergent subsequence  $(x_{c_\ell})_{\ell \in \mathbb{N}}$  which converges to a point denoted  $x^*$  different from zero. Also as above, we must have  $\gamma(x^*) = 0$  and  $\eta(x^*) \geq 0$ . But this contradicts the assumption, namely,

$$\{x \in \mathbb{R}^n \setminus \{0\}, \gamma(x) = 0\} \Rightarrow \eta(x) < 0.$$

**Appendix D. Proof of Proposition 2.18.** Because the vector field  $f$  is homogeneous in the  $\infty$ -limit, its approximating vector field  $f_\infty$  is homogeneous in the standard sense (see Remark 2.6). Let  $d_{V_\infty}$  be a positive real number larger than  $r_{\infty,i}$  for all  $i$  in  $\{1, \dots, n\}$ . Following Rosier [29], there exists a  $C^1$ , positive definite, proper, and homogeneous function  $V_\infty : \mathbb{R}^n \rightarrow \mathbb{R}_+$ , with weight  $r_\infty$  and degree  $d_{V_\infty}$ , satisfying

$$(D.1) \quad \frac{\partial V_\infty}{\partial x}(x) f_\infty(x) < 0 \quad \forall x \neq 0.$$

From P1 in section 2.2, we know that the function  $x \mapsto \frac{\partial V_\infty}{\partial x}(x) f(x)$  is homogeneous in the  $\infty$ -limit with associated triple  $(r_\infty, \mathfrak{d}_\infty + d_{V_\infty}, \frac{\partial V_\infty}{\partial x}(x) f_\infty(x))$ . Let

$$\epsilon_\infty = -\frac{1}{2} \max_{\theta \in S_{r_\infty}} \left\{ \frac{\partial V_\infty}{\partial x}(\theta) f_\infty(\theta) \right\},$$

and note that, by inequality (D.1),  $\epsilon_\infty$  is a strictly positive real number. By the definition of homogeneity in the  $\infty$ -limit, there exists  $\lambda_\infty$  such that

$$\left| \frac{\frac{\partial V_\infty}{\partial x}(\lambda^{r_\infty} \diamond \theta) f(\lambda^{r_\infty} \diamond \theta)}{\lambda^{d_{V_\infty} + \mathfrak{d}_\infty}} - \frac{\partial V_\infty}{\partial x}(\theta) f_\infty(\theta) \right| \leq \epsilon_\infty \quad \forall \theta \in S_{r_\infty}, \forall \lambda \geq \lambda_\infty.$$

This yields

$$\begin{aligned} \frac{\partial V_\infty}{\partial x}(\lambda^{r_\infty} \diamond \theta) f(\lambda^{r_\infty} \diamond \theta) &\leq \lambda^{d_{V_\infty} + \mathfrak{d}_\infty} \left( \frac{\partial V_\infty}{\partial x}(\theta) f_\infty(\theta) + \epsilon_\infty \right) \\ &\leq -\lambda^{d_{V_\infty} + \mathfrak{d}_\infty} \epsilon_\infty \quad \forall \theta \in S_{r_\infty}, \forall \lambda \geq \lambda_\infty, \end{aligned}$$

or in other words,

$$(D.2) \quad \frac{\partial V_\infty}{\partial x}(x) f(x) < 0 \quad \forall x : |x|_{r_\infty} \geq \lambda_\infty .$$

This establishes global asymptotic stability of the compact set

$$\mathcal{C}_\infty = \{x : V_\infty(x) \leq v_\infty\} ,$$

where  $v_\infty$  is given by

$$v_\infty = \max_{|x|_{r_\infty} = \lambda_\infty} \{V_\infty(x)\} .$$

**Appendix E. Proof of Theorem 2.20.** The proof is divided into three steps. First, we define three Lyapunov functions  $V_0$ ,  $V_m$ , and  $V_\infty$ . Then we build another Lyapunov function  $V$  from these three. Finally, we show that its derivative along the trajectories of the system (2.7) and its homogeneous approximations are negative definite.

1. As established in the proof of Proposition 2.18, there exist a positive real number  $\lambda_\infty$  and a  $C^1$  positive definite, proper, and homogeneous function  $V_\infty : \mathbb{R}^n \rightarrow \mathbb{R}_+$ , with weight  $r_\infty$  and degree  $d_{V_\infty}$  satisfying (D.2). Similarly, there exist a number  $\lambda_0 > 0$  and a  $C^1$  positive definite, proper, and homogeneous function  $V_0 : \mathbb{R}^n \rightarrow \mathbb{R}_+$ , with weight  $r_0$  and degree  $d_{V_0}$ , satisfying

$$(E.1) \quad \frac{\partial V_0}{\partial x}(x) f(x) < 0 \quad \forall x : 0 < |x|_{r_0} \leq \lambda_0 .$$

Finally, the global asymptotic stability of the origin of the system  $\dot{x} = f(x)$  implies the existence of a  $C^1$ , positive definite, and proper function  $V_m : \mathbb{R}^n \rightarrow \mathbb{R}_+$  satisfying

$$(E.2) \quad \frac{\partial V_m}{\partial x}(x) f(x) < 0 \quad \forall x \neq 0 .$$

2. Now we build a function  $V$  from the functions  $V_m$ ,  $V_\infty$ , and  $V_0$ . For this, we follow a technique used by Mazenc in [17] (see also [15]). Let  $v_\infty$  and  $v_0$  be two strictly positive real numbers such that  $v_0 < v_\infty$  and

$$v_\infty \geq \max_{x: |x|_{r_\infty} \leq \lambda_\infty} V_m(x), \quad v_0 \leq \min_{x: |x|_{r_0} \geq \lambda_0} V_m(x) .$$

This implies

$$\begin{aligned} \{x \in \mathbb{R}^n : V_m(x) \geq v_\infty\} &\subseteq \{x \in \mathbb{R}^n : |x|_{r_\infty} \geq \lambda_\infty\} , \\ \{x \in \mathbb{R}^n : V_m(x) \leq v_0\} &\subseteq \{x \in \mathbb{R}^n : |x|_{r_0} \leq \lambda_0\} . \end{aligned}$$

Let  $\omega_0$  and  $\omega_\infty$  be defined as

$$\omega_0 = \min_{x: \frac{1}{2} v_0 \leq V_m(x) \leq v_0} \frac{V_m(x)}{V_0(x)}, \quad \omega_\infty = \max_{x: v_\infty \leq V_m(x) \leq 2 v_\infty} \frac{V_m(x)}{V_\infty(x)} .$$

We have

$$\begin{aligned} \omega_\infty V_\infty(x) - V_m(x) &\geq 0 \quad \forall x : v_\infty \leq V_m(x) \leq 2 v_\infty , \\ V_m(x) - \omega_0 V_0(x) &\geq 0 \quad \forall x : \frac{1}{2} v_0 \leq V_m(x) \leq v_0 . \end{aligned}$$



Let

$$V(x) = \omega_\infty \varphi_\infty(V_m(x)) V_\infty(x) \\ + [1 - \varphi_\infty(V_m(x))] \varphi_0(V_m(x)) V_m(x) + \omega_0 [1 - \varphi_0(V_m(x))] V_0(x) ,$$

where  $\varphi_0$  and  $\varphi_\infty$  are  $C^1$  nondecreasing functions satisfying

$$(E.3) \quad \varphi_0(s) = 0 \quad \forall s \leq \frac{1}{2} v_0, \quad \varphi_0(s) = 1 \quad \forall s \geq v_0 ,$$

$$(E.4) \quad \varphi_\infty(s) = 0 \quad \forall s \leq v_\infty, \quad \varphi_\infty(s) = 1 \quad \forall s \geq 2v_\infty .$$

Then  $V$  is  $C^1$ , positive definite, and proper. Moreover, by construction,

$$V(x) = \begin{cases} \omega_0 V_0(x) & \forall x : V_m(x) \leq \frac{1}{2} v_0 , \\ \varphi_0(V_m(x)) V_m(x) + \omega_0 [1 - \varphi_0(V_m(x))] V_0(x) & \forall x : \frac{1}{2} v_0 \leq V_m(x) \leq v_0 , \\ V_m(x) & \forall x : v_0 \leq V_m(x) \leq v_\infty , \\ \omega_\infty \varphi_\infty(V_m(x)) V_\infty(x) + [1 - \varphi_\infty(V_m(x))] V_m(x) & \forall x : v_\infty \leq V_m(x) \leq 2v_\infty , \\ \omega_\infty V_\infty(x) & \forall x : V_m(x) \geq 2v_\infty . \end{cases}$$

Thus for each  $i$  in  $\{1, \dots, n\}$ ,

$$(E.5) \quad \frac{\partial V}{\partial x_i}(x) = \omega_\infty \frac{\partial V_\infty}{\partial x_i}(x) \quad \forall x : V_m(x) > 2v_\infty$$

and

$$(E.6) \quad \frac{\partial V}{\partial x_i}(x) = \omega_0 \frac{\partial V_0}{\partial x_i}(x) \quad \forall x : V_m(x) < \frac{1}{2} v_0 .$$

Since  $\frac{\partial V_\infty}{\partial x_i}$  and  $\frac{\partial V_0}{\partial x_i}$  are homogeneous in the standard sense, this proves that for each  $i$  in  $\{1, \dots, n\}$ ,  $\frac{\partial V}{\partial x_i}$  is homogeneous in the bi-limit, with weights  $r_0$  and  $r_\infty$  and degrees  $d_{V_0} - r_{0,i}$  and  $d_{V_\infty} - r_{\infty,i}$ .

3. It remains to show that the Lie derivative of  $V$  along  $f$  is negative definite. To this end note that, for all  $x$  such that  $\frac{1}{2} v_0 \leq V_m(x) \leq v_0$ ,

$$\frac{\partial V}{\partial x}(x) f(x) = \varphi'_0(V_m(x)) [V_m(x) - \omega_0 V_0(x)] \frac{\partial V_m}{\partial x}(x) f(x) \\ + \omega_0 [1 - \varphi_0(V_m(x))] \frac{\partial V_0}{\partial x}(x) f(x) + \varphi_0(V_m(x)) \frac{\partial V_m}{\partial x}(x) f(x)$$

and, for all  $x$  such that  $v_\infty \leq V_m(x) \leq 2v_\infty$ ,

$$\frac{\partial V}{\partial x}(x) f(x) = \varphi'_\infty(V_m(x)) [\omega_\infty V_\infty(x) - V_m(x)] \frac{\partial V_m}{\partial x}(x) f(x) \\ + \omega_\infty \varphi_\infty(V_m(x)) \frac{\partial V_\infty}{\partial x}(x) f(x) + [1 - \varphi_\infty(V_m(x))] \frac{\partial V_m}{\partial x}(x) f(x) .$$

By (D.2), (E.1), (E.2), (E.3), and (E.4), these inequalities imply

$$\frac{\partial V}{\partial x}(x) f(x) < 0 \quad \forall x \neq 0 ,$$

which proves the claim.

**Appendix F. Proof of Corollary 2.21.** Recall (1.6) and consider the functions  $\eta_1 : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$  and  $\gamma_1 : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}_+$  defined as

$$\eta_1(x, \delta) = \frac{\partial V}{\partial x}(x) \left[ f(x, \delta) - \frac{1}{2} f(x, 0) \right], \quad \gamma_1(x, \delta) = \sum_{j=1}^m \mathfrak{H} \left( |\delta_j|^{\frac{d_{V_0} + \mathfrak{d}_0}{\mathfrak{r}_{0,j}}}, |\delta_j|^{\frac{d_{V_\infty} + \mathfrak{d}_\infty}{\mathfrak{r}_{\infty,j}}} \right).$$

These functions are homogeneous in the bi-limit with weights  $r_0$  and  $r_\infty$  for  $x$  and  $\mathfrak{r}_0$  and  $\mathfrak{r}_\infty$  for  $\delta$  and degrees  $d_{V_0} + \mathfrak{d}_0$  and  $d_{V_\infty} + \mathfrak{d}_\infty$ . Since the function  $x \mapsto \frac{\partial V}{\partial x}(x) f(x, 0)$  is negative definite, then

$$\{(x, \delta) \in \mathbb{R}^{n+m} \setminus \{0\} : \gamma_1(x, \delta) = 0\} \subseteq \{(x, \delta) \in \mathbb{R}^{n+m} : \eta_1(x, \delta) < 0\}.$$

Moreover, since the homogeneous approximations of  $\eta$  are negative definite, we get

$$\begin{aligned} \{(x, \delta) \in \mathbb{R}^{n+m} \setminus \{0\} : \gamma_{1,0}(x, \delta) = 0\} &\subseteq \{(x, \delta) \in \mathbb{R}^{n+m} : \eta_{1,0}(x, \delta) < 0\}, \\ \{(x, \delta) \in \mathbb{R}^{n+m} \setminus \{0\} : \gamma_{1,\infty}(x, \delta) = 0\} &\subseteq \{(x, \delta) \in \mathbb{R}^{n+m} : \eta_{1,\infty}(x, \delta) < 0\}. \end{aligned}$$

Hence, by Lemma 2.13, there exists a positive real number  $c_\delta$  such that

$$(F.1) \quad \frac{\partial V}{\partial x}(x) \left[ f(x, \delta) - \frac{1}{2} f(x, 0) \right] \leq c_\delta \sum_{j=1}^m \mathfrak{H} \left( |\delta_j|^{\frac{d_{V_0} + \mathfrak{d}_0}{\mathfrak{r}_{0,j}}}, |\delta_j|^{\frac{d_{V_\infty} + \mathfrak{d}_\infty}{\mathfrak{r}_{\infty,j}}} \right).$$

Consider now the functions  $\eta_2 : \mathbb{R}^n \rightarrow \mathbb{R}_+$  and  $\gamma_2 : \mathbb{R}^n \rightarrow \mathbb{R}_+$  defined as

$$\eta_2(x) = \mathfrak{H} \left( V(x)^{\frac{d_{V_0} + \mathfrak{d}_0}{d_{V_0}}}, V(x)^{\frac{d_{V_\infty} + \mathfrak{d}_\infty}{d_{V_\infty}}} \right), \quad \gamma_2(x) = -\frac{1}{2} \frac{\partial V}{\partial x}(x) f(x, 0).$$

They are homogeneous in the bi-limit with weights  $r_0$  and  $r_\infty$  and degrees  $d_{V_0} + \mathfrak{d}_0$  and  $d_{V_\infty} + \mathfrak{d}_\infty$ . Since  $\gamma_2$  and its homogeneous approximations are positive definite, by Corollary 2.15 there exists a positive real number  $c_V$  such that

$$(F.2) \quad \frac{1}{2} \frac{\partial V}{\partial x}(x) f(x, 0) \leq -c_V \mathfrak{H} \left( V(x)^{\frac{d_{V_0} + \mathfrak{d}_0}{d_{V_0}}}, V(x)^{\frac{d_{V_\infty} + \mathfrak{d}_\infty}{d_{V_\infty}}} \right).$$

The two inequalities (F.1) and (F.2) yield the claim.

**Appendix G. Proof of Corollary 2.22.** Let  $d_{V_0}$  and  $d_{V_\infty}$  be such that the assumption of Theorem 2.20 holds. For each  $i$  in  $\{1, \dots, m\}$ , let  $\mu_i : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  be the strictly increasing function defined as (see (1.6))

$$(G.1) \quad \mu_i(s) = \mathfrak{H}(s^{q_i}, s^{p_i}),$$

where

$$p_i = \frac{\mathfrak{d}_\infty + d_{V_\infty}}{\mathfrak{r}_{\infty,i}}, \quad q_i = \frac{\mathfrak{d}_0 + d_{V_0}}{\mathfrak{r}_{0,i}}.$$

We first prove that the inequality given by Corollary 2.21 implies that the system (2.8), with  $\delta$  as input and  $x$  as output, is ISS with a linear gain between  $\sum_{i=1}^m \mu_i(|\delta_i|)$  and  $\mathfrak{H}(|x|_{r_0}^{\mathfrak{d}_0 + d_{V_0}}, |x|_{r_\infty}^{\mathfrak{d}_\infty + d_{V_\infty}})$ . To do so we introduce the function  $\alpha : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  as

$$\alpha(s) = \mathfrak{H} \left( s^{\frac{\mathfrak{d}_0 + d_{V_0}}{d_{V_0}}}, s^{\frac{\mathfrak{d}_\infty + d_{V_\infty}}{d_{V_\infty}}} \right), \quad s \geq 0.$$

This function is a bijection, strictly increasing, and homogeneous in the bi-limit with approximating functions  $s^{\frac{d_{V_0} + \mathfrak{d}_0}{d_{V_0}}}$  and  $s^{\frac{d_{V_\infty} + \mathfrak{d}_\infty}{d_{V_\infty}}}$ . Moreover, from Proposition 2.10, the function  $x \mapsto \alpha(V(x))$  is positive definite and homogeneous in the bi-limit with associated weights  $r_0$  and  $r_\infty$  and degrees  $\mathfrak{d}_0 + d_{V_0}$  and  $\mathfrak{d}_\infty + d_{V_\infty}$ . Moreover, its approximating homogeneous functions  $V_0(x)^{\frac{d_{V_0} + \mathfrak{d}_0}{d_{V_0}}}$  and  $V_\infty(x)^{\frac{d_{V_\infty} + \mathfrak{d}_\infty}{d_{V_\infty}}}$  are positive definite as well. Hence, we get from Corollary 2.15 the existence of a positive real number  $c_1$  satisfying

$$(G.2) \quad \mathfrak{H} \left( |x|_{r_0}^{\mathfrak{d}_0 + d_{V_0}}, |x|_{r_\infty}^{\mathfrak{d}_\infty + d_{V_\infty}} \right) \leq c_1 \alpha(V(x)) \quad \forall x \in \mathbb{R}^n.$$

On the other hand, from inequality (2.9) in Corollary 2.21, we have the property

$$(G.3) \quad \left\{ (x, \delta) \in \mathbb{R}^n \times \mathbb{R}^m : \alpha(V(x)) \geq 2 \frac{c_\delta}{c_V} \sum_{i=1}^m \mu_i(|\delta_i|) \right\} \\ \subseteq \left\{ (x, \delta) \in \mathbb{R}^n \times \mathbb{R}^m : \frac{\partial V}{\partial x}(x) f(x, \delta) \leq -\frac{c_V}{2} \alpha(V(x)) \right\}.$$

In the following, let  $t \in [0, T] \mapsto (x(t), \delta(t), z(t))$  be any function which satisfies (2.8) on  $[0, T]$  and (2.10) and (2.11) for all  $0 \leq s \leq t \leq T$ . From [32], we know the inclusion (G.3) implies the existence of a class  $\mathcal{KL}$  function  $\beta_V$  such that, for all  $0 \leq s \leq t \leq T$ ,

$$(G.4) \quad V(x(t)) \leq \max \left\{ \beta_V(V(x(s)), t-s), \sup_{s \leq \kappa \leq t} \left\{ \alpha^{-1} \left( \frac{2c_\delta}{c_V} \sum_{j=1}^m \mu_j(|\delta_j(\kappa)|) \right) \right\} \right\}.$$

With  $\alpha$  acting on both sides of inequality (G.4), (G.2) gives, for all  $0 \leq s \leq t \leq T$ ,

$$(G.5) \quad \mathfrak{H} \left( |x(t)|_{r_0}^{\mathfrak{d}_0 + d_{V_0}}, |x(t)|_{r_\infty}^{\mathfrak{d}_\infty + d_{V_\infty}} \right) \\ \leq \max \left\{ c_1 \alpha \circ \beta_V(V(x(s)), t-s), \frac{2c_1 c_\delta}{c_V} \sup_{s \leq \kappa \leq t} \left\{ \sum_{j=1}^m \mu_j(|\delta_j(\kappa)|) \right\} \right\}.$$

This is the linear gain property required. To conclude the proof it remains to show the existence of  $c_G$  such that a small gain property is satisfied.

First, note that the function  $x \mapsto \mathfrak{H}(|x|_{r_0}^{\mathfrak{d}_0 + d_{V_0}}, |x|_{r_\infty}^{\mathfrak{d}_\infty + d_{V_\infty}})$  is positive definite and homogeneous in the bi-limit with weights  $r_0$  and  $r_\infty$  and degrees  $\mathfrak{d}_0 + d_{V_0}$  and  $\mathfrak{d}_\infty + d_{V_\infty}$ . By Proposition 2.10, for  $i$  in  $\{1, \dots, m\}$  the same holds with the function  $x \mapsto \mu_i(\mathfrak{H}(|x|_{r_0}^{\mathfrak{r}_{0,i}}, |x|_{r_\infty}^{\mathfrak{r}_{\infty,i}}))$ . Hence, by Corollary 2.15, there exists a positive real number  $c_2$  satisfying

$$(G.6) \quad \mu_i(\mathfrak{H}(|x|_{r_0}^{\mathfrak{r}_{0,i}}, |x|_{r_\infty}^{\mathfrak{r}_{\infty,i}})) \leq c_2 \mathfrak{H}(|x|_{r_0}^{\mathfrak{d}_0 + d_{V_0}}, |x|_{r_\infty}^{\mathfrak{d}_\infty + d_{V_\infty}}) \quad \forall x \in \mathbb{R}^n.$$

Let  $C_i$  for  $i$  in  $\{1, \dots, m\}$  be the class  $\mathcal{K}_\infty$  functions defined as

$$C_i(c) = \max\{c^{q_i}, c^{p_i}\} + c^{\frac{p_i q_i}{q_i + p_i}} + c^{p_i + q_i}.$$

From (G.1), we get, for each  $s > 0$  and  $c > 0$ ,

$$\frac{\mu_i(cs)}{\mu_i(s)} = c^{q_i} \frac{(1 + s^{q_i})(1 + c^{p_i} s^{p_i})}{(1 + s^{p_i})(1 + c^{q_i} s^{q_i})} \leq c^{q_i} \left[ \frac{1 + c^{p_i} s^{p_i + q_i}}{1 + c^{q_i} s^{p_i + q_i}} + \frac{s^{q_i}}{1 + c^{q_i} s^{q_i + p_i}} + \frac{c^{p_i} s^{p_i}}{1 + s^{p_i}} \right],$$

where

$$c^{q_i} \frac{1 + c^{p_i} s^{p_i + q_i}}{1 + c^{q_i} s^{p_i + q_i}} \leq \max\{c^{q_i}, c^{p_i}\}, \quad \frac{c^{q_i} s^{q_i}}{1 + c^{q_i} s^{q_i + p_i}} \leq c^{\frac{p_i q_i}{q_i + p_i}}, \quad \frac{c^{q_i} c^{p_i} s^{p_i}}{1 + s^{p_i}} \leq c^{p_i + q_i}.$$

Hence, by continuity at 0, we have

$$(G.7) \quad \mu_i(cs) \leq C_i(c) \mu_i(s) \quad \forall (c, s) \in \mathbb{R}_+^2.$$

Consider the positive real numbers  $c_1$ ,  $c_2$ ,  $c_\delta$ , and  $c_V$  previously introduced, and select  $c_G$  in  $\mathbb{R}_+$  satisfying

$$(G.8) \quad c_G < \min_{1 \leq i \leq m} C_i^{-1} \left( \frac{c_V}{2m c_1 c_2 c_\delta} \right).$$

To show that such a selection for  $c_G$  is appropriate, observe that by (G.6) and (G.7) and  $\mu_i$  acting on both sides of the inequality (2.11), we get for each  $i$  in  $\{1, \dots, m\}$  and all  $0 \leq s \leq t \leq T$ ,

$$\mu_i(|\delta_i(t)|) \leq \max \left\{ \mu_i \circ \beta_\delta(|z(s)|, t-s), \right. \\ \left. C_i(c_G) c_2 \sup_{s \leq \kappa \leq t} \left\{ \mathfrak{H} \left( |x(\kappa)|_{r_0}^{\mathfrak{d}_0 + d_{V_0}}, |x(\kappa)|_{r_\infty}^{\mathfrak{d}_\infty + d_{V_\infty}} \right) \right\} \right\}.$$

Consequently

$$\sum_{i=1}^m \mu_i(|\delta_i(t)|) \leq \max \left\{ m \max_{1 \leq i \leq m} \{ \mu_i \circ \beta_\delta(|z(s)|, t-s) \}, \right. \\ (G.9) \quad \left. (m \max_{1 \leq i \leq m} C_i(c_G) c_2) \sup_{s \leq \kappa \leq t} \left\{ \mathfrak{H} \left( |x(\kappa)|_{r_0}^{\mathfrak{d}_0 + d_{V_0}}, |x(\kappa)|_{r_\infty}^{\mathfrak{d}_\infty + d_{V_\infty}} \right) \right\} \right\}.$$

Since (G.8) yields

$$\frac{2c_1 c_\delta}{c_V} m \max_{1 \leq i \leq m} C_i(c_G) c_2 < 1,$$

the existence of the function  $\beta_x$  follows from (2.10), (G.5), (G.9), and the (proof of the) small-gain theorem [11].

**Appendix H. Proof of Corollary 2.24.** First, observe that the continuity of  $f_0$ , at least, on  $\mathbb{R}^n \setminus \{0\}$  implies

$$|\mathfrak{d}_0| = -\mathfrak{d}_0 \leq \min_{1 \leq i \leq n} r_{0,i} \leq \max_{1 \leq i \leq n} r_{0,i} < d_{V_0}.$$

Then, let  $V$  be the function given in Theorem 2.20 and, since  $\mathfrak{d}_0 < 0 < \mathfrak{d}_\infty$ , the function  $\phi(x) = V(x)^{\frac{d_{V_0} + \mathfrak{d}_0}{d_{V_0}}} + V(x)^{\frac{d_{V_\infty} + \mathfrak{d}_\infty}{d_{V_\infty}}}$  is homogeneous in the bi-limit with weights  $r_0$  and  $r_\infty$ , degrees  $d_{V_0} + \mathfrak{d}_0$  and  $d_{V_\infty} + \mathfrak{d}_\infty$ , and approximating functions  $V(x)^{\frac{d_{V_0} + \mathfrak{d}_0}{d_{V_0}}}$  and  $V(x)^{\frac{d_{V_\infty} + \mathfrak{d}_\infty}{d_{V_\infty}}}$ . Moreover, the function  $\zeta(x) = -\frac{\partial V}{\partial x}(x) f(x)$  is homogeneous in the bi-limit with the same weights and degrees as  $\phi$ . Furthermore, since the function  $\zeta$  and its homogeneous approximations are positive definite, Corollary 2.15 yields a strictly positive real number  $c$  such that

$$(H.1) \quad \frac{\partial V}{\partial x}(x) f(x) \leq -c \left( V(x)^{\frac{d_{V_0} + \mathfrak{d}_0}{d_{V_0}}} + V(x)^{\frac{d_{V_\infty} + \mathfrak{d}_\infty}{d_{V_\infty}}} \right) \quad \forall x \in \mathbb{R}^n.$$

Let  $x_{ic}$  in  $\mathbb{R}^n \setminus \{0\}$  be the initial condition of a solution of the system  $\dot{x} = f(x)$ , and let  $V_{x_{ic}} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  be the function of time given by the evaluation of  $V$  along this solution. Then

$$\overline{V_{x_{ic}}(t)} \leq -c V_{x_{ic}}(t)^{\frac{d_{V_\infty} + \vartheta_\infty}{d_{V_\infty}}} \quad \forall t \geq 0 ,$$

from which we get

$$V_{x_{ic}}(t) \leq \frac{1}{\left(\frac{\vartheta_\infty}{d_{V_\infty}} ct + V(x_{ic})^{-\frac{\vartheta_\infty}{d_{V_\infty}}}\right)^{\frac{d_{V_\infty}}{\vartheta_\infty}}} \leq \frac{1}{\left(\frac{\vartheta_\infty}{d_{V_\infty}} ct\right)^{\frac{d_{V_\infty}}{\vartheta_\infty}}} \quad \forall t > 0 .$$

Therefore, setting  $T_1 = \frac{d_{V_\infty}}{c\vartheta_\infty}$ , we have

$$V_{x_{ic}}(t) \leq 1 \quad \forall t \geq T_1, \quad \forall x_{ic} \in \mathbb{R}^n$$

and

$$\overline{V_{x_{ic}}(t)} \leq -c V_{x_{ic}}(t)^{\frac{d_{V_0} - |\vartheta_0|}{d_{V_0}}} \quad \forall t \geq 0 .$$

As a result, we get

$$\begin{aligned} V_{x_{ic}}(t) &\leq \max \left\{ \left( -\frac{|\vartheta_0|}{d_{V_0}} c(t - T_1) + V_{x_{ic}}(T_1)^{\frac{|\vartheta_0|}{d_{V_0}}} \right)^{\frac{d_{V_0}}{|\vartheta_0|}}, 0 \right\} , \\ &\leq \max \left\{ \left( 1 - \frac{|\vartheta_0|}{d_{V_0}} c(t - T_1) \right)^{\frac{d_{V_0}}{|\vartheta_0|}}, 0 \right\} \quad \forall t \geq T_1 . \end{aligned}$$

Therefore, setting  $T_2 = \frac{d_{V_0}}{c|\vartheta_0|}$  yields

$$V_{x_{ic}}(t) = 0 \quad \forall t \geq T_1 + T_2 = \frac{1}{c} \left( \frac{d_{V_\infty}}{\vartheta_\infty} + \frac{d_{V_0}}{|\vartheta_0|} \right), \quad \forall x_{ic} \in \mathbb{R}^n ,$$

hence the claim.

**Acknowledgments.** The second author is extremely grateful to Wilfrid Perruquetti and Emmanuel Moulay for the many discussions about the notion of homogeneity in the bi-limit. Also, all the authors would like to thank the anonymous reviewers for their comments, which were extremely helpful in improving the quality of the paper.

#### REFERENCES

- [1] V. ANDRIEU, L. PRALY, AND A. ASTOLFI, *Nonlinear output feedback design via domination and generalized weighted homogeneity*, in Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, 2006, pp. 6391–6396.
- [2] A. BACCIOTTI AND L. ROSIER, *Liapunov Functions and Stability in Control Theory*, Lecture Notes in Control and Inform. Sci. 267, Springer, Berlin, 2001.
- [3] S. P. BHAT AND D. S. BERNSTEIN, *Geometric homogeneity with applications to finite-time stability*, Math. Control Signals Systems, 17 (2005), pp. 101–127.

- [4] S. P. BHAT AND D. S. BERNSTEIN, *Finite-time stability of continuous autonomous systems*, SIAM J. Control Optim., 38 (2000), pp. 751–766.
- [5] J.-M. CORON AND L. PRALY, *Adding an integrator for the stabilization problem*, Systems Control Lett., 17 (1991), pp. 89–104.
- [6] J.-M. CORON AND L. ROSIER, *A relation between continuous time-varying and discontinuous feedback stabilization*, J. Math. Systems Estim. Control, 4 (1994), pp. 67–84.
- [7] J. P. GAUTHIER AND I. KUPKA, *Deterministic Observation Theory And Applications*, Cambridge University Press, Cambridge, UK, 2001.
- [8] W. HAHN, *Stability of Motion*, Springer-Verlag, Berlin, 1967.
- [9] H. HERMES, *Homogeneous coordinates and continuous asymptotically stabilizing feedback controls*, in Differential Equations: Stability and Control, Lecture Notes in Pure Appl. Math. 109, S. Elaydi, ed., Marcel Dekker, New York, 1991, pp. 249–260.
- [10] Y. HONG, *Finite-time stabilization and stabilizability of a class of controllable systems*, Systems Control Lett., 46 (2002), pp. 231–236.
- [11] Z.-P. JIANG, A. TEEL, AND L. PRALY, *Small-gain theorem for ISS systems and applications*, Math. Control Signals Systems, 7 (1994), pp. 95–120.
- [12] M. KAWSKI, *Stabilization of nonlinear systems in the plane*, Systems Control Lett., 12 (1989), pp. 169–175.
- [13] H. KHALIL AND A. SABERI, *Adaptive stabilization of a class of nonlinear systems using high-gain feedback*, IEEE Trans. Automat. Control, 32 (1987), pp. 1031–1035.
- [14] S. LEFSCHETZ, *Differential Equations: Geometric Theory*, 2nd ed., Dover, New York, 1977.
- [15] W. LIU, Y. CHITOUR, AND E. SONTAG, *On finite-gain stabilizability of linear systems subject to input saturation*, SIAM J. Control Optim., 34 (1996), pp. 1190–1219.
- [16] J. L. MASSERA, *Contributions to stability theory*, Ann. of Math., 64 (1956), pp. 182–206.
- [17] F. MAZENC, *Stabilisation de trajectoires, ajout d'intégration, commandes saturées*, Mémoire de thèse en Mathématiques et Automatique de l'École Nationale Supérieure des Mines de Paris, Avril 1996.
- [18] F. MAZENC, L. PRALY, AND W. P. DAYAWANSA, *Global stabilization by output feedback: Examples and counter-examples*, Systems Control Lett., 23 (1994), pp. 119–125.
- [19] P. MORIN AND C. SAMSON, *Application of backstepping techniques to the time-varying exponential stabilisation of chained form systems*, European J. Control, 3 (1997), pp. 15–36.
- [20] R. ORSI, L. PRALY, AND I. MAREELS, *Sufficient conditions for the existence of an unbounded solution*, Automatica, 37 (2001), pp. 1609–1617.
- [21] L. PRALY, B. D'ANDRÉA-NOVEL, AND J.-M. CORON, *Lyapunov design of stabilizing controllers for cascaded systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1177–1181.
- [22] L. PRALY AND Z.-P. JIANG, *Stabilization by output feedback for systems with ISS inverse dynamics*, Systems Control Lett., 21 (1993), pp. 19–33.
- [23] L. PRALY AND Z.-P. JIANG, *Further results on robust semiglobal stabilization with dynamic input uncertainties*, in Proceedings of the 37th Annual IEEE Conference on Decision and Control, Tampa, 1998, pp. 891–896.
- [24] L. PRALY AND Z.-P. JIANG, *Linear output feedback with dynamic high gain for nonlinear systems*, Systems Control Lett., 53 (2004), pp. 107–116.
- [25] L. PRALY AND F. MAZENC, *Design of Homogeneous Feedbacks for a Chain of Integrators and Applications*, Internal report E 173, CAS, Ecole des Mines de Paris, Paris, December 11, 1995; revised April 15, 1996.
- [26] C. QIAN, *A homogeneous domination approach for global output feedback stabilization of a class of nonlinear systems*, in Proceedings of the IEEE American Control Conference, Portland, 2005, pp. 4708–4715.
- [27] C. QIAN AND W. LIN, *Output feedback control of a class of nonlinear systems: A nonseparation principle paradigm*, IEEE Trans. Automat. Control, 47 (2002), pp. 1710–1715.
- [28] C. QIAN AND W. LIN, *Recursive observer design, homogeneous approximation, and nonsmooth output feedback stabilization of nonlinear systems*, IEEE Trans. Automat. Control, 51 (2006), pp. 1457–1471.
- [29] L. ROSIER, *Homogeneous Lyapunov function for homogeneous continuous vector field*, Systems Control Lett., 19 (1992), pp. 467–473.
- [30] H. SHIM AND J. SEO, *Recursive nonlinear observer design: Beyond the uniform observability*, IEEE Trans. Automat. Control, 48 (2003), pp. 294–298.
- [31] E. D. SONTAG, *Input to state stability: Basic concepts and results*, in Nonlinear and Optimal Control Theory, P. Nistri and G. Stefani, eds., Springer-Verlag, Berlin, 2006, pp. 163–220.
- [32] E. SONTAG AND Y. WANG, *Lyapunov characterizations of input to output stability*, SIAM J. Control Optim., 39 (2000), pp. 226–249.

- [33] M. TZAMTZI AND J. TSINIAS, *Explicit formulas of feedback stabilizers for a class of triangular systems with uncontrollable linearization*, Systems Control Lett., 38 (1999), pp. 115–126.
- [34] F. W. WILSON, JR., *Smoothing derivatives of functions and applications*, Trans. Amer. Math. Soc., 139 (1969), pp. 413–428.
- [35] B. YANG AND W. LIN, *Homogeneous observers, iterative design, and global stabilization of high-order nonlinear systems by smooth output feedback*, IEEE Trans. Automat. Control, 49 (2004), pp. 1069–1080.

# INVARIANT CARNOT–CARATHEODORY METRICS ON $S^3$ , $SO(3)$ , $SL(2)$ , AND LENS SPACES\*

UGO BOSCAIN<sup>†</sup> AND FRANCESCO ROSSI<sup>‡</sup>

**Abstract.** In this paper we study the Carnot–Caratheodory metrics on  $SU(2) \simeq S^3$ ,  $SO(3)$ , and  $SL(2)$  induced by their Cartan decomposition and by the Killing form. Besides computing explicitly geodesics and conjugate loci, we compute the cut loci (globally), and we give the expression of the Carnot–Caratheodory distance as the inverse of an elementary function. We then prove that the metric given on  $SU(2)$  projects on the so-called lens spaces  $L(p, q)$ . Also for lens spaces, we compute the cut loci (globally). For  $SU(2)$  the cut locus is a maximal circle without one point. In all other cases the cut locus is a stratified set. To our knowledge, this is the first explicit computation of the whole cut locus in sub-Riemannian geometry, except for the trivial case of the Heisenberg group.

**Key words.** left-invariant sub-Riemannian geometry, Carnot–Caratheodory distance, global structure of the cut locus, lens spaces

**AMS subject classifications.** 22E30, 49J15, 53C17

**DOI.** 10.1137/070703727

**1. Introduction.** In this paper we study the global structure of the cut locus (the set of points reached optimally by more than one geodesic) for the simplest sub-Riemannian structures on three-dimensional simple Lie groups (i.e.,  $SU(2)$ ,  $SO(3)$ , and  $SL(2)$ ), namely, the left-invariant sub-Riemannian structures induced by their Cartan decomposition and by the Killing form.

Let  $G$  be a simple real Lie group of matrices with associated Lie algebra  $\mathbf{L}$  and Killing form  $\text{Kil}(\cdot, \cdot)$ . Let  $\mathbf{L} = \mathbf{k} \oplus \mathbf{p}$  be its Cartan decomposition with the usual commutation relations  $[\mathbf{k}, \mathbf{k}] \subseteq \mathbf{k}$ ,  $[\mathbf{p}, \mathbf{p}] \subseteq \mathbf{k}$ ,  $[\mathbf{k}, \mathbf{p}] \subseteq \mathbf{p}$ . If  $\mathbf{L}$  is noncompact, we also require  $\mathbf{k}$  to be the maximal compact subalgebra of  $\mathbf{L}$ . The most natural left-invariant sub-Riemannian structure that one can define on  $G$  is the one in which the distribution is generated by left translations of  $\mathbf{p}$ , and the sub-Riemannian metric  $\langle \cdot, \cdot \rangle$  at the identity is generated by a scalar multiple of the Killing form restricted to  $\mathbf{p}$ . The scalar must be chosen positive or negative in such a way that the scalar product is positive definite. We call  $G$ , endowed with such a sub-Riemannian structure, a  **$\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifold**.

$\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifolds have very special features: there are no strict abnormal minimizers, and the Hamiltonian system given by the Pontryagin maximum principle (PMP) is integrable in terms of elementary functions (products of exponentials). More precisely, if we write the distribution at a point  $g \in G$  as  $\Delta(g) = g\mathbf{p}$ , we have the following expression for geodesics parametrized by arclength, starting at time zero from  $g_0$  (see [3, 10, 14, 22, 23]):

$$(1) \quad g(t) = g_0 e^{(A_k + A_p)t} e^{-A_k t},$$

where  $A_k \in \mathbf{k}$ ,  $A_p \in \mathbf{p}$ , and we have  $\langle A_p, A_p \rangle = 1$ . Thanks to left-invariance, without loss of generality we can always assume that  $g_0$  is the identity, and we will do so throughout the paper.

\*Received by the editors September 25, 2007; accepted for publication (in revised form) February 17, 2008; published electronically June 25, 2008. The first author was partially supported by a FABER grant from Région Bourgogne.

<http://www.siam.org/journals/sicon/47-4/70372.html>

<sup>†</sup>LE2i, CNRS UMR5158, Université de Bourgogne, 9, avenue Alain Savary - BP 47870, 21078 Dijon Cedex, France (boscaïn@sissa.it).

<sup>‡</sup>SISSA, via Beirut 2-4, 34014 Trieste, Italy (rossifr@sissa.it).



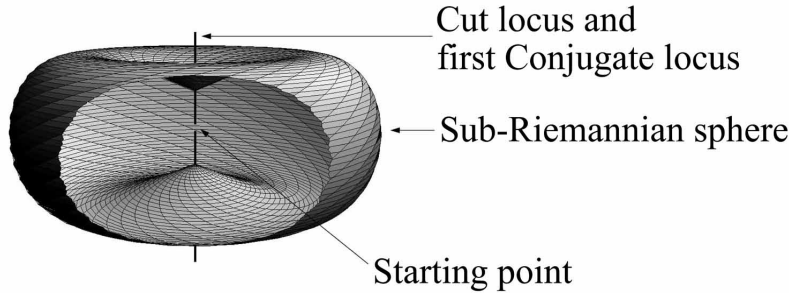


FIG. 1. Local structure of sub-Riemannian spheres and of cut and conjugate loci for 3-dim  $\mathfrak{k} \oplus \mathfrak{p}$  sub-Riemannian manifolds.

In all three-dimensional cases (i.e.,  $SU(2)$ ,  $SO(3)$ , and  $SL(2)$ ),  $\mathfrak{p}$  has dimension 2, while  $\mathfrak{k}$  has dimension 1. Writing  $\mathfrak{p} = \text{span}\{p_1, p_2\}$ , where  $\{p_1, p_2\}$  is an orthonormal frame for the sub-Riemannian structure (i.e.,  $\langle p_i, p_j \rangle = \delta_{ij}$ ) and  $\mathfrak{k} = \text{span}\{k\}$ , we can write  $A_p = \cos(\theta)p_1 + \sin(\theta)p_2$  and  $A_k = ck$  with  $\theta \in \mathbb{R}/2\pi$ ,  $c \in \mathbb{R}$ . The map associating to the triple  $(\theta, c, t)$  the final point of the corresponding geodesic starting from the identity is called the *exponential map*,

$$\begin{aligned} \text{Exp} : S^1 \times \mathbb{R} \times \mathbb{R}^+ &\rightarrow G, \\ (\theta, c, t) &\mapsto \text{Exp}(\theta, c, t) = e^{(A_k + A_p)t} e^{-A_k t}. \end{aligned}$$

For three-dimensional  $\mathfrak{k} \oplus \mathfrak{p}$  sub-Riemannian manifolds, the local structure of the sub-Riemannian spheres, cut loci, and conjugate loci starting from the identity has been described by Agrachev (unpublished), and, due to cylindrical symmetry of the Killing form in the  $\mathfrak{p}$  subspace, it is very similar to that of the Heisenberg group. Indeed, locally, the cut locus coincides with the first conjugate locus (i.e., the set where local optimality is lost) and is made by two connected one-dimensional manifolds adjacent to the identity and transversal to the distribution; see Figure 1.

However, the global structure of the cut locus was still unknown. Indeed, to our knowledge, no global structure of the cut locus is known in sub-Riemannian geometry apart from that of the Heisenberg group.

The main result of our paper is the following.

**THEOREM 1.** *Let  $K_{\text{Id}}$  be the cut locus starting from the identity. We have the following:*

- (i) *For  $SU(2)$ ,  $K_{\text{Id}}$  is a maximal circle  $S^1$  without one point (the identity).*
- (ii) *For  $SO(3)$ ,  $K_{\text{Id}}$  is a stratified set made by two manifolds glued in one point. The first manifold is  $\mathbb{RP}^2$ ; the second manifold is a maximal circle  $S^1$  without one point (the identity).*
- (iii) *For  $SL(2)$ ,  $K_{\text{Id}}$  is a stratified set made by two manifolds glued in one point. The first manifold is  $\mathbb{R}^2$ ; the second manifold is a circle  $S^1$  without one point (the identity).*

A picture of the three cut loci is given in Figure 2.

For all cases, the one-dimensional strata contain the cut locus appearing in the local analysis.

Notice that the  $\mathfrak{k} \oplus \mathfrak{p}$  sub-Riemannian manifold  $SU(2)$  has the structure of a CR (Cauchy–Riemann) manifold and is a tight structure [7, 16].

Once the cut locus is computed, one can obtain the expression of the sub-Riemannian distance from the identity. The following theorem gives the sub-Riemannian distance for  $SU(2)$ . The proof, given in section 5.1.1, can be adapted

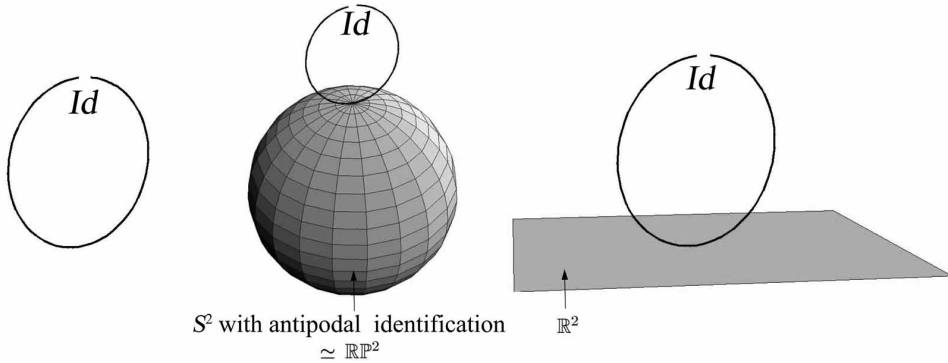


FIG. 2. The cut loci for the  $\mathfrak{k} \oplus \mathfrak{p}$  sub-Riemannian manifolds  $SU(2)$ ,  $SO(3)$ , and  $SL(2)$ .

to get similar results in the cases of  $SO(3)$  and  $SL(2)$ .

THEOREM 2. *Let*

$$SU(2) = \left\{ \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix} \mid \alpha, \beta \in \mathbb{C}, |\alpha|^2 + |\beta|^2 = 1 \right\}.$$

Consider the sub-Riemannian distance from  $\text{Id}$  defined by

$$\begin{aligned} \dot{g} &= g \left( \frac{u_1}{2} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + \frac{u_2}{2} \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \right), \\ d(\text{Id}, g_1) &:= \inf \left\{ \int_0^T \sqrt{u_1^2 + u_2^2} \mid g(0) = \text{Id}, g(T) = g_1 \right\}. \end{aligned}$$

It holds that

$$(2) \quad d \left( \text{Id}, \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix} \right) = \begin{cases} 2\sqrt{\arg(\alpha)(2\pi - \arg(\alpha))} & \text{if } \beta = 0, \\ \psi(\alpha) & \text{if } \beta \neq 0, \end{cases}$$

where  $\arg(\alpha) \in [0, 2\pi]$  and  $\psi(\alpha) = t$  is the unique solution of

$$(3) \quad \begin{cases} -\frac{ct}{2} + \arctan \left( \frac{c}{\sqrt{1+c^2}} \tan \left( \frac{\sqrt{1+c^2}t}{2} \right) \right) = \arg(\alpha), \\ \frac{\sin \left( \frac{\sqrt{1+c^2}t}{2} \right)}{\sqrt{1+c^2}} = \sqrt{1-|\alpha|^2}, \\ t \in \left( 0, \frac{2\pi}{\sqrt{1+c^2}} \right). \end{cases}$$

This theorem and its analogues for  $SO(3)$  and  $SL(2)$  are useful to give estimates for the fundamental solutions of the hypoelliptic heat equation induced by the sub-Riemannian structure (see [5, 12, 17, 19]). Moreover, this theorem can be seen as the answer, in the case of  $SU(2)$ , to the question (formulated in [14]) about the possibility of inverting the matrix equation (1); i.e., for every matrix  $g \in SU(2)$ , find a matrix  $A = A_k + A_p$ , with  $\langle A_p, A_p \rangle = 1$ , being a solution to the equation  $g = g_0 e^{(A_k + A_p)t} e^{-A_k t}$ . If  $\beta \neq 0$ , then this equation has one and only one solution; otherwise it has more than one solution (indeed, infinitely many; see sections 3 and 5).

Then we study the most natural sub-Riemannian structures on the lens spaces  $L(p, q)$  induced by the one on  $SU(2)$ . The lens space  $L(p, q)$  (with  $p, q$  coprime integers,  $p, q \neq 0$ ) is the quotient of  $SU(2)$  by the equivalence relation

$$\begin{pmatrix} \alpha_1 & \beta_1 \\ -\beta_1 & \alpha_1 \end{pmatrix} \sim \begin{pmatrix} \alpha_2 & \beta_2 \\ -\beta_2 & \alpha_2 \end{pmatrix} \text{ if } \exists \omega \in \mathbb{C} \text{ } p\text{th root of unity such that}$$

$$\begin{pmatrix} \alpha_2 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \omega & 0 \\ 0 & \omega^q \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix}.$$

The lens spaces are three-dimensional manifolds, but they are neither Lie groups nor homogeneous spaces of  $SU(2)$ , except for the case  $L(2, 1) \simeq SO(3)$ .

In the case of lens spaces, we get that the cut locus is much more complicated with respect to those on  $SU(2)$  and  $SL(2)$ . It is still a stratified set, but in general with more strata. The precise description is given in section 5.2.

Sub-Riemannian structures on the lens space  $L(4, 1)$  are particularly interesting for mechanical applications and for problems of geometry of vision on the two-dimensional sphere. Indeed,  $L(4, 1) \simeq PTS^2$ , the bundle of directions of  $S^2$ . These applications are the subject of a forthcoming paper.

The structure of this paper is the following. In section 2 we recall the definition of sub-Riemannian manifold, state the PMP (that is a first order necessary condition for optimality for problems of calculus of variations with nonholonomic constraints), and define the cut and conjugate loci. Then we define  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifolds. In section 3 we define  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian structures on  $SU(2)$ ,  $SO(3)$ , and  $SL(2)$  and compute the corresponding geodesics and conjugate loci. In section 4 we give sub-Riemannian structures on lens spaces as quotients of the  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian structure on  $SU(2)$ . The core of the paper is section 5, where we compute the cut loci and the sub-Riemannian distance. The general idea is the following: we first identify the prolongation of the cut locus arising locally, then we compute the part of the cut locus due to the symmetries of the problem, and finally, we show that there is no other cut point.

## 2. Basic definitions.

**2.1. Sub-Riemannian manifold.** An  $(n, m)$ -sub-Riemannian manifold is a triple  $(M, \Delta, \mathbf{g})$ , where

- (i)  $M$  is a connected smooth manifold of dimension  $n$ ;
- (ii)  $\Delta$  is a Lie bracket generating smooth distribution of constant rank  $m < n$ ; i.e.,  $\Delta$  is a smooth map that associates to  $q \in M$  an  $m$ -dim subspace  $\Delta(q)$  of  $T_q M$ , and  $\forall q \in M$ , we have

$$(4) \quad \text{span} \{[f_1, [\dots [f_{k-1}, f_k] \dots]](q) \mid f_i \in \text{Vec}(M) \text{ and } f_i(p) \in \Delta(p) \forall p \in M\} = T_q M.$$

Here  $\text{Vec}(M)$  denotes the set of smooth vector fields on  $M$ .

- (iii)  $\mathbf{g}_q$  is a Riemannian metric on  $\Delta(q)$ , that is, smooth as a function of  $q$ .

The Lie bracket generating condition (4) is also known as the Hörmander condition.

A Lipschitz continuous curve  $\gamma : [0, T] \rightarrow M$  is said to be *horizontal* if  $\dot{\gamma}(t) \in \Delta(\gamma(t))$  for almost every  $t \in [0, T]$ . Given a horizontal curve  $\gamma : [0, T] \rightarrow M$ , the

length of  $\gamma$  is

$$(5) \quad l(\gamma) = \int_0^T \sqrt{\mathbf{g}_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))} \, dt.$$

The distance induced by the sub-Riemannian structure on  $M$  is the function

$$(6) \quad d(q_0, q_1) = \inf\{l(\gamma) \mid \gamma(0) = q_0, \gamma(T) = q_1, \gamma \text{ horizontal}\}.$$

The hypothesis of connectedness of  $M$  and the Lie bracket generating assumption for the distribution guarantee the finiteness and the continuity of  $d(\cdot, \cdot)$  with respect to the topology of  $M$  (Chow's theorem; see, for instance, [3]).

The function  $d(\cdot, \cdot)$  is called the Carnot-Caratheodory distance and gives to  $M$  the structure of metric space (see [6, 18]).

It is a standard fact that  $l(\gamma)$  is invariant under reparametrization of the curve  $\gamma$ . Moreover, if an admissible curve  $\gamma$  minimizes the so-called *energy functional*

$$E(\gamma) = \int_0^T \mathbf{g}_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t)) \, dt$$

with  $T$  fixed (and fixed initial and final point), then  $v = \sqrt{\mathbf{g}_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}$  is constant and  $\gamma$  is also a minimizer of  $l(\cdot)$ . On the other hand, a minimizer  $\gamma$  of  $l(\cdot)$  such that  $v$  is constant is a minimizer of  $E(\cdot)$  with  $T = l(\gamma)/v$ .

A *geodesic* for the sub-Riemannian manifold is a curve  $\gamma : [0, T] \rightarrow M$  such that for every sufficiently small interval  $[t_1, t_2] \subset [0, T]$ ,  $\gamma|_{[t_1, t_2]}$  is a minimizer of  $E(\cdot)$ . A geodesic for which  $\mathbf{g}_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))$  is (constantly) equal to one is said to be parametrized by arclength.

Locally, the pair  $(\Delta, \mathbf{g})$  can be given by assigning a set of  $m$  smooth vector fields that are orthonormal for  $\mathbf{g}$ , i.e.,

$$(7) \quad \Delta(q) = \text{span}\{F_1(q), \dots, F_m(q)\}, \quad \mathbf{g}_q(F_i(q), F_j(q)) = \delta_{ij}.$$

When  $(\Delta, \mathbf{g})$  can be defined as in (7) by  $m$  vector fields defined globally, we say that the sub-Riemannian manifold is *trivializable*.

Given a  $(n, m)$ -trivializable sub-Riemannian manifold, the problem of finding a curve minimizing the energy between two fixed points  $q_0, q_1 \in M$  is naturally formulated as the optimal control problem

$$(8) \quad \dot{q} = \sum_{i=1}^m u_i F_i(q), \quad u_i \in \mathbb{R}, \quad \int_0^T \sum_{i=1}^m u_i^2(t) \, dt \rightarrow \min, \quad q(0) = q_0, \quad q(T) = q_1.$$

It is a standard fact that this optimal control problem is equivalent to the minimum time problem with controls  $u_1, \dots, u_m$  satisfying  $u_1^2 + \dots + u_m^2 \leq 1$ .

When the manifold is analytic and the orthonormal frame can be assigned through  $m$  analytic vector fields, we say that the sub-Riemannian manifold is *analytic*.

In this paper we are concerned with sub-Riemannian manifolds that are trivializable and analytic since they are given in terms of left-invariant vector fields on Lie groups.

## 2.2. First order necessary conditions, cut locus, and conjugate locus.

Consider a trivializable  $(n, m)$ -sub-Riemannian manifold. Solutions to the optimal control problem (8) are computed via the PMP (see, for instance, [3, 11, 21, 24]) that is a first order necessary condition for optimality and generalizes the Weierstraß conditions of the calculus of variations. For each optimal curve, the PMP provides a lift to the cotangent bundle that is a solution to a suitable pseudo-Hamiltonian system.

**THEOREM 3** (PMP for the problem (8)). *Let  $M$  be an  $n$ -dimensional smooth manifold, and consider the minimization problem (8), in the class of Lipschitz continuous curves, where  $F_i$ ,  $i = 1, \dots, m$  are smooth vector fields on  $M$  and the final time  $T$  is fixed. Consider the map  $\mathbf{H} : T^*M \times \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}$  defined by*

$$\mathbf{H}(q, \lambda, p_0, u) := \left\langle \lambda, \sum_{i=1}^m u_i F_i(q) \right\rangle + p_0 \sum_{i=1}^m u_i^2(t).$$

*If the curve  $q(\cdot) : [0, T] \rightarrow M$  corresponding to the control  $u(\cdot) : [0, T] \rightarrow \mathbb{R}^m$  is optimal, then there exist a never vanishing Lipschitz continuous covector  $\lambda(\cdot) : t \in [0, T] \mapsto \lambda(t) \in T_{q(t)}^*M$  and a constant  $p_0 \leq 0$  such that, for a.e.  $t \in [0, T]$ ,*

- (i)  $\dot{q}(t) = \frac{\partial \mathbf{H}}{\partial \lambda}(q(t), \lambda(t), p_0, u(t))$ ,
- (ii)  $\dot{\lambda}(t) = -\frac{\partial \mathbf{H}}{\partial q}(q(t), \lambda(t), p_0, u(t))$ , and
- (iii)  $\frac{\partial \mathbf{H}}{\partial u}(q(t), \lambda(t), p_0, u(t)) = 0$ .

**Remark 1.** A curve  $q(\cdot) : [0, T] \rightarrow M$  satisfying the PMP is said to be an *extremal*. In general, an extremal may correspond to more than one pair  $(\lambda(\cdot), p_0)$ . If an extremal satisfies the PMP with  $p_0 \neq 0$ , then it is called a *normal extremal*. If it satisfies the PMP with  $p_0 = 0$  it is called an *abnormal extremal*. An extremal can be both normal and abnormal. For normal extremals one can normalize  $p_0 = -1/2$ .

If an extremal satisfies the PMP only with  $p_0 = 0$ , then it is called a *strict abnormal extremal*. If a strict abnormal extremal is optimal, then it is called a *strict abnormal minimizer*. For a deep analysis of abnormal extremals in sub-Riemannian geometry, see [8, 15].

It is well known that all normal extremals are geodesics (see, for instance, [3]). Moreover, if there are no strict abnormal minimizers, then all geodesics are normal extremals for some fixed final time  $T$ . This always will be the case in this paper; indeed, we are concerned with sub-Riemannian manifolds of dimension 3, defined by a pair of vector fields  $F_1$  and  $F_2$  such that  $\forall q \in M$ ,  $\text{span}\{F_1(q), F_2(q), [F_1(q), F_2(q)]\} = T_q M$ , i.e., the so called three-dimensional *contact case*, for which there are no abnormal extremals (not even nonstrict).

In this case, from (iii) one gets  $u_i(t) = \langle \lambda(t), F_i(t) \rangle$ ,  $i = 1, \dots, m$ , and the PMP becomes much simpler: a curve  $q(\cdot)$  is a geodesic if and only if it is the projection on  $M$  of a solution  $(\lambda(t), q(t))$  for the Hamiltonian system on  $T^*M$  corresponding to

$$H(\lambda, q) = \frac{1}{2} \left( \sum_{i=1}^m \langle \lambda, F_i(q) \rangle^2 \right), \quad q \in M, \quad \lambda \in T_q^*M$$

satisfying  $H(\lambda(0), q(0)) \neq 0$ .

**Remark 2.** Notice that  $H$  is constant along any given solution of the Hamiltonian system. Moreover,  $H = \frac{1}{2}$  if and only if the geodesic is parametrized by arclength. In the following, for simplicity of notation we assume that all geodesics are defined for  $t \in [0, +\infty)$ .

Fix  $q_0 \in M$ . For every  $\lambda_0 \in T_{q_0}^*M$  satisfying

$$(9) \quad H(\lambda_0, q_0) = 1/2$$

and every  $t > 0$ , define the *exponential map*  $\text{Exp}(\lambda_0, t)$  as the projection on  $M$  of the solution, evaluated at time  $t$ , of the Hamiltonian system associated with  $H$ , with initial condition  $\lambda(0) = \lambda_0$  and  $q(0) = q_0$ . Notice that condition (9) defines a hypercylinder  $\Lambda_{q_0} \simeq S^{m-1} \times \mathbb{R}^{n-m}$  in  $T_{q_0}^* M$ .

**DEFINITION 4.** *The **conjugate locus from**  $q_0$  is the set  $C_{q_0}$  of critical values of the map*

$$\begin{aligned} \text{Exp} : \Lambda_{q_0} \times \mathbb{R}^+ &\rightarrow M, \\ (\lambda_0, t) &\mapsto \text{Exp}(\lambda_0, t). \end{aligned}$$

For every  $\bar{\lambda}_0 \in \Lambda_{q_0}$ , let  $t(\bar{\lambda}_0)$  be the  $n$ th positive time, if it exists, for which the map  $(\lambda_0, t) \mapsto \text{Exp}(\lambda_0, t)$  is singular at  $(\bar{\lambda}_0, t(\bar{\lambda}_0))$ . The  **$n$ th conjugate locus from**  $q_0$   $C_{q_0}^n$  is the set  $\{\text{Exp}(\bar{\lambda}_0, t(\bar{\lambda}_0)) \mid t(\bar{\lambda}_0) \text{ exists}\}$ .

The **cut locus** from  $q_0$  is the set  $K_{q_0}$  of points reached optimally by more than one geodesic, i.e., the set

$$K_{q_0} = \left\{ q \in M \mid \exists \begin{array}{l} \lambda_1, \lambda_2 \in \Lambda_{q_0}, \lambda_1 \neq \lambda_2, \\ t \in \mathbb{R}^+ \end{array} \text{ such that } \begin{array}{l} q = \text{Exp}(\lambda_1, t), \\ q = \text{Exp}(\lambda_2, t), \\ \text{Exp}(\lambda_1, \cdot) \text{ optimal in } [0, t], \\ \text{Exp}(\lambda_2, \cdot) \text{ optimal in } [0, t]. \end{array} \right\}$$

**Remark 3.** It is a standard fact that for every  $\bar{\lambda}_0$  satisfying (9), the set  $T(\bar{\lambda}_0) = \{\bar{t} > 0 \mid \text{Exp}(\lambda, t) \text{ is singular at } (\bar{\lambda}_0, \bar{t})\}$  is a discrete set (see, for instance, [3]).

**Remark 4.** Let  $(M, \Delta, \mathbf{g})$  be a sub-Riemannian manifold. Fix  $q_0 \in M$  and assume that

- (i) Each point of  $M$  is reached by an optimal geodesic starting from  $q_0$ ;
- (ii) there are no abnormal minimizers.

The following facts are well known (a proof in the three-dimensional contact case can be found in [4]):

(i) The first conjugate locus  $C_{q_0}^1$  is the set of points where the geodesics starting from  $q_0$  lose local optimality;

(ii) if  $q(\cdot)$  is a geodesic starting from  $q_0$ , and  $\bar{t}$  is the first positive time such that  $q(\bar{t}) \in K_{q_0} \cup C_{q_0}^1$ , then  $q(\cdot)$  loses optimality in  $\bar{t}$ ; i.e., it is optimal in  $[0, \bar{t}]$  and not optimal in  $[0, t]$  for any  $t > \bar{t}$ ;

(iii) if a geodesic  $q(\cdot)$  starting from  $q_0$  loses optimality at  $\bar{t} > 0$ , then  $q(\bar{t}) \in K_{q_0} \cup C_{q_0}^1$ .

As a consequence, when the first conjugate locus is included in the cut locus (as in our cases; see section 5), the cut locus is the set of points where the geodesics lose optimality.

**Remark 5.** It is well known that, while in Riemannian geometry  $K_{q_0}$  is never adjacent to  $q_0$ , in sub-Riemannian geometry this is always the case. See [2].

**2.3.  $\mathfrak{k} \oplus \mathfrak{p}$  sub-Riemannian manifolds.** For the sake of simplicity in the exposition, throughout the paper, when we deal with Lie groups and Lie algebras, we always consider that they are groups and algebras of matrices.

Let  $\mathbf{L}$  be a simple Lie algebra and  $\text{Kil}(X, Y) = \text{Tr}(ad_X \circ ad_Y)$  its Killing form. Recall that the Killing form defines a nondegenerate pseudoscalar product on  $\mathbf{L}$ . In the following we recall what we mean by a Cartan decomposition of  $\mathbf{L}$ .

**DEFINITION 5.** *A Cartan decomposition of a simple Lie algebra  $\mathbf{L}$  is any decomposition of the form*

$$(10) \quad \mathbf{L} = \mathfrak{k} \oplus \mathfrak{p}, \text{ where } [\mathfrak{k}, \mathfrak{k}] \subseteq \mathfrak{k}, \quad [\mathfrak{p}, \mathfrak{p}] \subseteq \mathfrak{k}, \quad [\mathfrak{k}, \mathfrak{p}] \subseteq \mathfrak{p}.$$

DEFINITION 6. Let  $G$  be a simple Lie group with Lie algebra  $\mathbf{L}$ . Let  $\mathbf{L} = \mathbf{k} \oplus \mathbf{p}$  be a Cartan decomposition of  $\mathbf{L}$ . In the case in which  $G$  is noncompact, assume that  $\mathbf{k}$  is the maximal compact subalgebra of  $\mathbf{L}$ .

On  $G$ , consider the distribution  $\Delta(g) = g\mathbf{p}$  endowed with the Riemannian metric  $\mathbf{g}_g(v_1, v_2) = \langle g^{-1}v_1, g^{-1}v_2 \rangle$ , where  $\langle \cdot, \cdot \rangle := \alpha \operatorname{Kil}|_{\mathbf{p}}(\cdot, \cdot)$  and  $\alpha < 0$  (resp.,  $\alpha > 0$ ) if  $G$  is compact (resp., noncompact).

In this case we say that  $(G, \Delta, \mathbf{g})$  is a  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifold.

The constant  $\alpha$  is clearly not relevant. It is chosen just to obtain good normalizations.

Remark 6. In the compact (resp., noncompact) case, the fact that  $\mathbf{g}$  is positive definite on  $\Delta$  is guaranteed by the requirement  $\alpha < 0$  (resp., by the requirements  $\alpha > 0$  and  $\mathbf{k}$  maximal compact subalgebra).

Let  $\{X_j\}$  be an orthonormal frame for the subspace  $\mathbf{p} \subset \mathbf{L}$ , with respect to the metric defined in Definition 6. Then the problem of finding the minimal energy between the identity and a point  $g_1 \in G$  in fixed time  $T$  becomes the left-invariant optimal control problem

$$\dot{g} = g \left( \sum_j u_j X_j \right), \quad u_j \in L^\infty(0, T), \quad \int_0^T \sum_j u_j^2(t) dt \rightarrow \min, \quad g(0) = \operatorname{Id}, \quad g(T) = g_1.$$

This problem admits a solution; see, for instance, Chapter 5 of [13].

For  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifolds, one can prove that strict abnormal extremals are never optimal, since the *Goh condition* (see [3]) is never satisfied. Moreover, the Hamiltonian system given by the PMP is integrable and the explicit expression of geodesics starting from the identity and parametrized by arclength is

$$(11) \quad g(t) = e^{(A_k + A_p)t} e^{-A_k t},$$

where  $A_k \in \mathbf{k}$ ,  $A_p \in \mathbf{p}$ , and  $\langle A_p, A_p \rangle = 1$ . This formula is well known in the community. It was used independently by Agrachev [1], Brockett [14], and Kupka (oral communication). The first complete proof was written by Jurdjevic in [22]. The proof that strict abnormal extremals are never optimal was first written in [10]. See also [3, 23].

Remark 7. In the three-dimensional case, the Hamiltonian system given by the PMP is indeed integrable even if the cost is not built with the Killing form (biinvariant) but is only left-invariant. For the case of  $SO(3)$  see [9].

**3.  $SU(2)$ ,  $SO(3)$ ,  $SL(2)$ , their geodesics, and their conjugate loci.** In this section we fix coordinates on  $SU(2)$ ,  $SO(3)$ ,  $SL(2)$ , and we apply formula (11) in order to get the explicit expressions for geodesics and conjugate loci.

**3.1. The  $\mathbf{k} \oplus \mathbf{p}$  problem on  $SU(2)$ .** The Lie group  $SU(2)$  is the group of unitary unimodular  $2 \times 2$  complex matrices

$$SU(2) = \left\{ \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix} \in \operatorname{Mat}(2, \mathbb{C}) \mid |\alpha|^2 + |\beta|^2 = 1 \right\}.$$

It is compact and simply connected. The Lie algebra of  $SU(2)$  is the algebra of anti-Hermitian traceless  $2 \times 2$  complex matrices

$$su(2) = \left\{ \begin{pmatrix} i\alpha & \beta \\ -\bar{\beta} & -i\alpha \end{pmatrix} \in \operatorname{Mat}(2, \mathbb{C}) \mid \alpha \in \mathbb{R}, \beta \in \mathbb{C} \right\}.$$

A basis of  $su(2)$  is  $\{p_1, p_2, k\}$ , where

$$(12) \quad p_1 = \frac{1}{2} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad p_2 = \frac{1}{2} \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \quad k = \frac{1}{2} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix},$$

whose commutation relations are  $[p_1, p_2] = k$ ,  $[p_2, k] = p_1$ , and  $[k, p_1] = p_2$ . Recall that for  $su(n)$  we have  $\text{Kil}(X, Y) = 2n\text{Tr}(XY)$  (see [20, pp. 186, 516]); thus for  $su(2)$ ,  $\text{Kil}(X, Y) = 4\text{Tr}(XY)$  and, in particular,  $\text{Kil}(p_i, p_j) = -2\delta_{ij}$ . The choice of the subspaces

$$\mathbf{k} = \text{span}\{k\} \quad \mathbf{p} = \text{span}\{p_1, p_2\}$$

provides a *Cartan decomposition* for  $su(2)$ . Moreover,  $\{p_1, p_2\}$  is an orthonormal frame for the inner product  $\langle \cdot, \cdot \rangle = -\frac{1}{2}\text{Kil}(\cdot, \cdot)$  restricted to  $\mathbf{p}$ .

Defining  $\Delta(g) = g\mathbf{p}$  and  $\mathbf{g}_g(v_1, v_2) = \langle g^{-1}v_1, g^{-1}v_2 \rangle$ , we have that  $(SU(2), \Delta, \mathbf{g})$  is a  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifold.

*Remark 8.* Observe that all the  $\mathbf{k} \oplus \mathbf{p}$  structures that one can define on  $SU(2)$  are equivalent. For instance, one could set  $\mathbf{k} = \text{span}\{p_1\}$  and  $\mathbf{p} = \text{span}\{p_2, k\}$ .

Recall that

$$SU(2) \simeq S^3 = \left\{ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in \mathbb{C}^2 \mid |\alpha|^2 + |\beta|^2 = 1 \right\}$$

via the isomorphism

$$\phi : \begin{pmatrix} SU(2) & \rightarrow & S^3, \\ \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix} & \mapsto & \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \end{pmatrix}$$

In the following we always write elements of  $SU(2)$  as pairs of complex numbers.

**3.1.1. Expression of geodesics.** We compute the explicit expression of geodesics using the formula (11). Consider an initial covector  $\lambda = \lambda(\theta, c) = \cos(\theta)p_1 + \sin(\theta)p_2 + ck \in \Lambda_{\text{Id}}$ . The corresponding exponential map is

$$\text{Exp}(\theta, c, t) := \text{Exp}(\lambda(\theta, c), t) = e^{(\cos(\theta)p_1 + \sin(\theta)p_2 + ck)t} e^{-ckt} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

with

$$\begin{aligned} \alpha &= \frac{c \sin(\frac{ct}{2}) \sin(\sqrt{1+c^2}\frac{t}{2})}{\sqrt{1+c^2}} + \cos\left(\frac{ct}{2}\right) \cos\left(\sqrt{1+c^2}\frac{t}{2}\right) \\ &\quad + i \left( \frac{c \cos(\frac{ct}{2}) \sin(\sqrt{1+c^2}\frac{t}{2})}{\sqrt{1+c^2}} - \sin\left(\frac{ct}{2}\right) \cos\left(\sqrt{1+c^2}\frac{t}{2}\right) \right), \\ \beta &= \frac{\sin(\sqrt{1+c^2}\frac{t}{2})}{\sqrt{1+c^2}} \left( \cos\left(\frac{ct}{2} + \theta\right) + i \sin\left(\frac{ct}{2} + \theta\right) \right). \end{aligned}$$

We have the following symmetry properties:

(i) *cylindrical symmetry*:

$$\text{Exp}(\theta, c, t) = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{pmatrix} \text{Exp}(0, c, t);$$



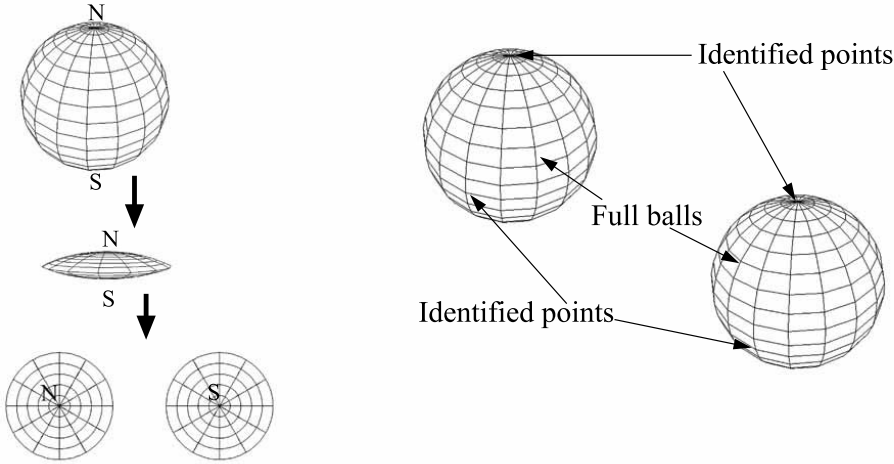


FIG. 3. Left: construction of the 2-dim picture of  $S^2$ . Right: the 3-dim picture of  $S^3$ .

(ii) *central symmetry*: Set  $\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \text{Exp}(\theta, c, t)$ . We have

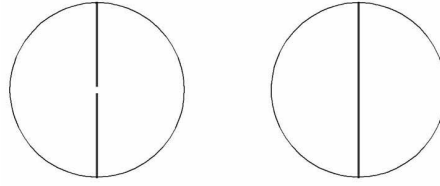
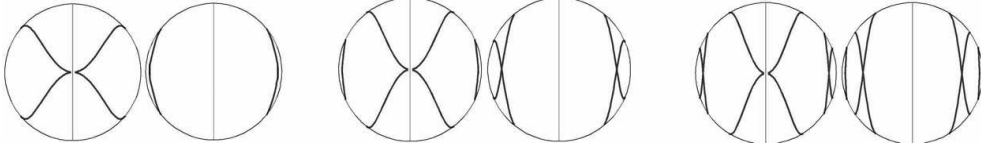
$$\text{Exp}(\theta, -c, t) = \begin{cases} \begin{pmatrix} \bar{\alpha} \\ e^{2i(\theta - \arg(\beta))}\beta \end{pmatrix} & \text{if } \beta \neq 0, \\ \begin{pmatrix} \bar{\alpha} \\ 0 \end{pmatrix} & \text{if } \beta = 0. \end{cases}$$

**3.1.2. Pictures of  $S^2$  and  $S^3$ .** We recall a standard construction for representing  $S^2$  in a two-dimensional space and  $S^3$  in a three-dimensional space. For more details, see, e.g., [26]. Consider  $S^2 \subset \mathbb{R}^3$  and flatten it on the equator plane, pushing the northern hemisphere down and the southern hemisphere up, getting two superimposed disks  $D^2$  joined along their circular boundaries. The construction is drawn in Figure 3 (left). Similarly, consider  $S^3 \subset \mathbb{C}^2 \simeq \mathbb{R}^4$ : it can be viewed as two superimposed balls joined along their boundaries. In this case the boundaries are two spheres  $S^2$ . A picture of  $S^3$  is given in Figure 3 (right).

**3.1.3. The conjugate locus.** Recall that all the partial derivatives of  $\text{Exp}$  evaluated in  $(\theta, c, t)$  lie in  $T_g SU(2) = g \cdot su(2)$  with  $g = \text{Exp}(\theta, c, t)$ . One can easily check that the three vectors  $g^{-1} \cdot \frac{\partial \text{Exp}}{\partial \theta} \big|_{(\theta, c, t)}$ ,  $g^{-1} \cdot \frac{\partial \text{Exp}}{\partial c} \big|_{(\theta, c, t)}$ ,  $g^{-1} \cdot \frac{\partial \text{Exp}}{\partial t} \big|_{(\theta, c, t)} \in su(2)$  are linearly dependent (hence  $g$  is a conjugate point) if and only if

$$\sin\left(\sqrt{1+c^2}\frac{t}{2}\right)\left(2\sin\left(\sqrt{1+c^2}\frac{t}{2}\right)-\sqrt{1+c^2}t\cos\left(\sqrt{1+c^2}\frac{t}{2}\right)\right)=0.$$

The first term is 0 if and only if  $g \in e^{\mathbf{k}} = \left\{ \begin{pmatrix} \alpha \\ 0 \end{pmatrix} \mid |\alpha| = 1 \right\}$ , while the second vanishes if and only if  $\sqrt{1+c^2}\frac{t}{2} = \tan\left(\sqrt{1+c^2}\frac{t}{2}\right)$ ; hence we have two series of conjugate


 FIG. 4.  $\mathfrak{k} \oplus \mathfrak{p}$  problem on  $SU(2)$ : projection of the odd conjugate loci.

 FIG. 5.  $\mathfrak{k} \oplus \mathfrak{p}$  problem on  $SU(2)$ : projection of the 2nd, 4th, and 6th conjugate loci.

times as follows:

(i) first series:  $t_{2n-1} = \frac{2n\pi}{\sqrt{1+c^2}}$ , to which correspond the conjugate loci  $C_{\text{Id}}^{2n-1} = e^{\mathbf{k}} \setminus \text{Id}$ ;

(ii) second series:  $t_{2n} = \frac{2x_n}{\sqrt{1+c^2}}$ , where  $\{x_1, x_2, \dots\}$  is the ordered set of the strictly positive solutions of  $x = \tan(x)$ , to which correspond the conjugate loci

$$C_{\text{Id}}^{2n} = \left\{ \left( \begin{array}{c} \frac{c \sin(x_n)}{\sqrt{1+c^2}} e^{i(\frac{\pi}{2}-y_n)} + \cos(x_n) e^{-iy_n} \\ \frac{\sin(x_n)}{\sqrt{1+c^2}} e^{i\theta} \end{array} \right) \mid \begin{array}{l} c \in \mathbb{R}, \\ \theta \in \mathbb{R}/2\pi \end{array} \right\},$$

where  $y_n = \frac{cx_n}{\sqrt{1+c^2}}$ .

*Remark 9.* Notice that all the geodesics have a countable number of conjugate times.

We present some images of conjugate loci (Figures 4 and 5). For simplicity we present images of their sections with the plane  $\text{Re}(\beta) = 0$ . The complete images can be recovered using cylindrical symmetry.

*Remark 10.* Notice that the second conjugate locus is a two-dimensional submanifold of  $SU(2)$ , while the other even conjugate loci have self-intersections.

**3.2. The  $\mathfrak{k} \oplus \mathfrak{p}$  problem on  $SO(3)$ .** The Lie group  $SO(3)$  is the group of special orthogonal  $3 \times 3$  real matrices

$$SO(3) = \{g \in \text{Mat}(3, \mathbb{R}) \mid gg^T = \text{Id}, \det(g) = 1\}.$$

It is compact and its fundamental group is  $\mathbb{Z}_2$ . The Lie algebra of  $SO(3)$  is the algebra of skew-symmetric  $3 \times 3$  real matrices

$$\mathfrak{so}(3) = \left\{ \begin{pmatrix} 0 & -a & b \\ a & 0 & -c \\ -b & c & 0 \end{pmatrix} \in \text{Mat}(3, \mathbb{R}) \right\}.$$

A basis of  $so(3)$  is  $\{p_1, p_2, k\}$ , where

$$p_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad p_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad k = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

whose commutation relations are  $[p_1, p_2] = k$ ,  $[p_2, k] = p_1$ , and  $[k, p_1] = p_2$ . Recall that  $so(3)$  and  $su(2)$  are isomorphic as Lie algebras, while  $SU(2)$  is a double covering of  $SO(3)$ .

For  $so(3)$  we have  $\text{Kil}(X, Y) = \text{Tr}(XY)$  so, in particular,  $\text{Kil}(p_i, p_j) = -2\delta_{ij}$ . The choice of the subspaces

$$\mathbf{k} = \text{span}\{k\}, \quad \mathbf{p} = \text{span}\{p_1, p_2\}$$

gives a *Cartan decomposition* for  $so(3)$ . Moreover,  $\{p_1, p_2\}$  is an orthonormal frame for the inner product  $\langle \cdot, \cdot \rangle = -\frac{1}{2}\text{Kil}(\cdot, \cdot)$  restricted to  $\mathbf{p}$ .

Defining  $\Delta(g) = g\mathbf{p}$  and  $\mathbf{g}_g(v_1, v_2) = \langle g^{-1}v_1, g^{-1}v_2 \rangle$ , we have that  $(SO(3), \Delta, \mathbf{g})$  is a  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifold. As for  $SU(2)$ , all the  $\mathbf{k} \oplus \mathbf{p}$  structures that one can define on  $SO(3)$  are equivalent.

**3.2.1. Expression of geodesics.** Consider an initial covector  $\lambda = \lambda(\theta, c) = \cos(\theta)p_1 + \sin(\theta)p_2 + ck \in \Lambda_{\text{Id}}$ . Using formula (11), we have that the exponential map is

$$\text{Exp}(\theta, c, t) := \text{Exp}(\lambda(\theta, c), t) = e^{(\cos(\theta)p_1 + \sin(\theta)p_2 + ck)t} e^{-ckt} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

with

$$\begin{aligned} a_{11} &= K_1 \cos(ct) + K_2 \cos(2\theta + ct) + K_3 c \sin(ct), \\ a_{12} &= K_1 \sin(ct) + K_2 \sin(2\theta + ct) - K_3 c \cos(ct), & a_{13} &= K_4 \cos(\theta) + K_3 \sin(\theta), \\ a_{21} &= -K_1 \sin(ct) + K_2 \sin(2\theta + ct) + K_3 c \cos(ct), \\ a_{22} &= K_1 \cos(ct) - K_2 \cos(2\theta + ct) + K_3 c \sin(ct), & a_{23} &= -K_3 \cos(\theta) + K_4 \sin(\theta), \\ a_{21} &= K_4 \cos(\theta + ct) - K_3 \sin(\theta + ct), & a_{22} &= K_3 \cos(\theta + ct) + K_4 \sin(\theta + ct), \\ a_{23} &= \frac{\cos(\sqrt{1+c^2}t) + c^2}{1+c^2}, \\ K_1 &= \frac{1 + (1+2c^2)\cos(\sqrt{1+c^2}t)}{2(1+c^2)}, & K_2 &= \frac{1 - \cos(\sqrt{1+c^2}t)}{2(1+c^2)}, \\ K_3 &= \frac{\sin(\sqrt{1+c^2}t)}{\sqrt{1+c^2}}, & K_4 &= \frac{c(1 - \cos(\sqrt{1+c^2}t))}{1+c^2}. \end{aligned}$$

The set of geodesics has symmetry properties similar to the  $SU(2)$  case. The conjugate locus can be obtained from that of the  $SU(2)$  by the canonical projection  $SU(2) \rightarrow SO(3)$ . As for  $SU(2)$ , all the geodesics have a countable number of conjugate points.

**3.3. The  $\mathbf{k} \oplus \mathbf{p}$  problem on  $SL(2)$ .** The Lie group  $SL(2)$  is the group of  $2 \times 2$  real matrices with determinant 1,

$$SL(2) = \{g \in \text{Mat}(2, \mathbb{R}) \mid \det(g) = 1\}.$$

It is a noncompact group and its fundamental group is  $\mathbb{Z}$ . The Lie algebra of  $SL(2)$  is the algebra of traceless  $2 \times 2$  real matrices

$$sl(2) = \left\{ \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \text{Mat}(2, \mathbb{R}) \right\}.$$

A basis of  $sl(2)$  is  $\{p_1, p_2, k\}$ , where

$$p_1 = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad p_2 = \frac{1}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad k = \frac{1}{2} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

whose commutation relations are  $[p_1, p_2] = -k$ ,  $[p_2, k] = p_1$ , and  $[k, p_1] = p_2$ . For  $sl(n)$  we have  $\text{Kil}(X, Y) = 2n\text{Tr}(XY)$  (see [20]); hence for  $sl(2)$ ,  $\text{Kil}(X, Y) = 4\text{Tr}(XY)$  and, in particular,  $\text{Kil}(p_i, p_j) = 2\delta_{ij}$ . The choice of the subspaces

$$\mathbf{k} = \text{span}\{k\}, \quad \mathbf{p} = \text{span}\{p_1, p_2\}$$

provides a *Cartan decomposition* for  $sl(2)$ . For  $sl(2)$  the Cartan decomposition is unique, since  $\mathbf{k}$  must be the maximal compact subalgebra. Moreover,  $\{p_1, p_2\}$  is a orthonormal frame for the inner product  $\langle \cdot, \cdot \rangle = \frac{1}{2}\text{Kil}(\cdot, \cdot)$  restricted to  $\mathbf{p}$ .

Defining  $\Delta(g) = g\mathbf{p}$  and  $\mathbf{g}_g(v_1, v_2) = \langle g^{-1}v_1, g^{-1}v_2 \rangle$ , we have that  $(SL(2), \Delta, \mathbf{g})$  is a  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifold.

**3.3.1. Expression of geodesics.** Consider an initial covector  $\lambda = \lambda(\theta, c) = \cos(\theta)p_1 + \sin(\theta)p_2 + ck \in \Lambda_{\text{Id}}$ . Using formula (11), we have that the exponential map is

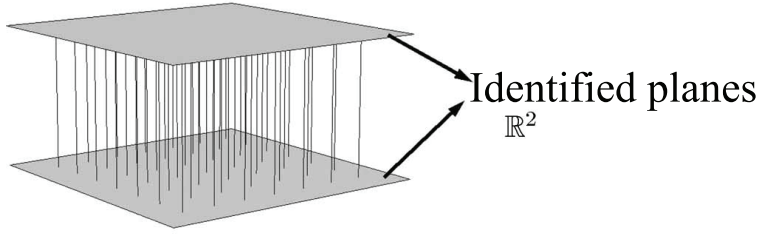
$$\text{Exp}(\theta, c, t) := \text{Exp}(\lambda(\theta, c), t) = e^{(\cos(\theta)p_1 + \sin(\theta)p_2 + ck)t} e^{-ckt} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

with

$$\begin{aligned} a_{11} &= K_1 \cos\left(c\frac{t}{2}\right) + K_2 \left( \cos\left(\theta + c\frac{t}{2}\right) + c \sin\left(c\frac{t}{2}\right) \right), \\ a_{12} &= K_1 \sin\left(c\frac{t}{2}\right) + K_2 \left( \sin\left(\theta + c\frac{t}{2}\right) - c \cos\left(c\frac{t}{2}\right) \right), \\ a_{21} &= -K_1 \sin\left(c\frac{t}{2}\right) + K_2 \left( \sin\left(\theta + c\frac{t}{2}\right) + c \cos\left(c\frac{t}{2}\right) \right), \\ a_{22} &= K_1 \cos\left(c\frac{t}{2}\right) + K_2 \left( -\cos\left(\theta + c\frac{t}{2}\right) + c \sin\left(c\frac{t}{2}\right) \right), \\ K_1 &= \begin{cases} \text{Cosh}\left(\sqrt{1-c^2}\frac{t}{2}\right), & c \in [-1, 1], \\ \cos\left(\sqrt{c^2-1}\frac{t}{2}\right), & c \in (-\infty, -1) \cup (1, +\infty), \end{cases} \\ K_2 &= \begin{cases} \frac{\text{Sinh}\left(\sqrt{1-c^2}\frac{t}{2}\right)}{\sqrt{1-c^2}}, & c \in (-1, 1), \\ \frac{t}{2}, & c \in \{-1, 1\}, \\ \frac{\sin\left(\sqrt{c^2-1}\frac{t}{2}\right)}{\sqrt{c^2-1}}, & c \in (-\infty, -1) \cup (1, +\infty). \end{cases} \end{aligned}$$

### 3.3.2. A useful decomposition of $SL(2)$ .

**PROPOSITION 7.** *For every  $g \in SL(2)$ , there exists a unique pair  $r \in e^{\mathbf{k}}$ ,  $s \in e^{\mathbf{p}}$  such that  $g = rs$ .*

FIG. 6. A picture of  $SL(2)$ .

*Proof.* First, notice that  $e^{\mathbf{K}} = SO(2)$  and that  $e^{\mathbf{P}}$  is the set of  $2 \times 2$  symmetric matrices with determinant 1 and positive trace.

Take

$$r = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \in e^{\mathbf{K}} \quad \text{and} \quad g = \begin{pmatrix} \alpha + \delta & \beta - \gamma \\ \beta + \gamma & \alpha - \delta \end{pmatrix} \in SL(2).$$

Notice that  $(\alpha, \gamma) \neq (0, 0)$ . We have to prove that there exists a unique  $\theta \in \mathbb{R}/2\pi$  such that  $s = r^{-1}g$  is symmetric with positive trace. By direct computation, one gets that  $s$  is symmetric if and only if  $\alpha \sin(\theta) = \gamma \cos(\theta)$ . For any  $(\alpha, \gamma) \in \mathbb{R}^2 \setminus (0, 0)$ , there exist two solutions of this equation  $\theta_1, \theta_2 \in \mathbb{R}/2\pi$  with  $\theta_2 = \theta_1 + \pi$ . Thus

$$\text{Tr} \left( \begin{pmatrix} \cos(\theta_1) & \sin(\theta_1) \\ -\sin(\theta_1) & \cos(\theta_1) \end{pmatrix} g \right) = -\text{Tr} \left( \begin{pmatrix} \cos(\theta_2) & \sin(\theta_2) \\ -\sin(\theta_2) & \cos(\theta_2) \end{pmatrix} g \right).$$

Observing that a symmetric matrix with determinant 1 has nonvanishing trace, either  $\theta_1$  or  $\theta_2$  provide  $\text{Tr}(s) > 0$ .  $\square$

Topologically,  $e^{\mathbf{K}} \simeq S^1$  and  $e^{\mathbf{P}} \simeq \mathbb{R}^2$ ; hence  $SL(2) \simeq S^1 \times \mathbb{R}^2$ . In the following, we represent  $SL(2)$  as the set  $\mathbb{R}^2 \times [0, 1]$  with the identification rule  $(a, b, 0) \sim (a, b, 1)$ . See Figure 6.

**3.3.3. Symmetries in the  $SL(2)$  problem.** We have the following symmetry properties:

(i) *cylindrical symmetry*:  $\text{Exp}(\theta, c, t) = e^{z_0 k} e^{x p_1 + y p_2}$ , where

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

and  $(x_0, y_0, z_0)$  are defined by  $\text{Exp}(0, c, t) = e^{z_0 k} e^{x_0 p_1 + y_0 p_2}$ ;

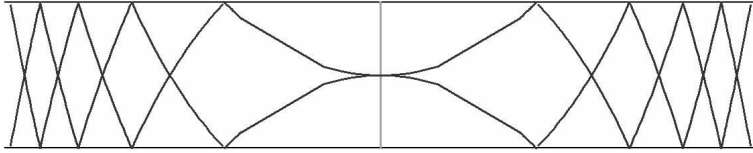
(ii) *central symmetry*:  $\text{Exp}(\theta, -c, t) = e^{-z_0 k} e^{x p_1 + y p_2}$ , where

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(2\theta) \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

and  $(x_0, y_0, z_0)$  are defined by  $\text{Exp}(\theta, c, t) = e^{z_0 k} e^{x_0 p_1 + y_0 p_2}$ .

**3.3.4. The conjugate locus.** With arguments similar to those of section 3.1.3, one checks that  $g = \text{Exp}(\theta, c, t)$  is a conjugate point if and only if

$$\begin{cases} \sinh(d) (2\sinh(d) - t\sqrt{1-c^2}\cosh(d)) = 0, & c \in (-1, 1), \\ \pm \frac{t^4}{12} = 0, & c = \pm 1, \\ \sin(d) (2\sin(d) - t\sqrt{c^2-1}\cos(d)) = 0, & c \notin [-1, 1] \end{cases}$$


 FIG. 7.  $\mathbf{k} \oplus \mathbf{p}$  problem on  $SU(2)$ : Section of the 2nd conjugate locus.

with  $d = \sqrt{1 - c^2} \frac{t}{2}$  when  $c \in (-1, 1)$  and  $d = \sqrt{c^2 - 1} \frac{t}{2}$  when  $c \notin [-1, 1]$ .

The first two equations have only the trivial solution  $t = 0$ . The third gives two series of conjugate times as follows:

(i) first series:  $t_{2n-1} = \frac{2n\pi}{\sqrt{c^2-1}}$ , to which correspond the conjugate loci  $C_{\text{Id}}^{2n-1} = e^{\mathbf{k}} \setminus \text{Id}$ ;

(ii) second series:  $t_{2n} = \frac{2x_n}{\sqrt{c^2-1}}$ , where  $\{x_1, x_2, \dots\}$  is the ordered set of the strictly positive solutions of  $x = \tan(x)$ , to which correspond the conjugate loci

$$C_{\text{Id}}^{2n} = \left\{ \begin{pmatrix} a_{11}^n(c, t) & a_{12}^n(c, t) \\ a_{21}^n(c, t) & a_{22}^n(c, t) \end{pmatrix} \middle| \begin{array}{l} c \in \mathbb{R}, \\ \theta \in \mathbb{R}/2\pi \end{array} \right\}$$

with

$$\begin{aligned} a_{11}^n(c, t) &= \cos(x_n) \cos(y_n) + \frac{\sin(x_n)}{\sqrt{c^2-1}} (\cos(\theta) + c \sin(y_n)), \\ a_{12}^n(c, t) &= \cos(x_n) \sin(y_n) + \frac{\sin(x_n)}{\sqrt{c^2-1}} (\sin(\theta) - c \cos(y_n)), \\ a_{21}^n(c, t) &= -\cos(x_n) \sin(y_n) + \frac{\sin(x_n)}{\sqrt{c^2-1}} (\sin(\theta) + c \cos(y_n)), \\ a_{22}^n(c, t) &= \cos(x_n) \cos(y_n) + \frac{\sin(x_n)}{\sqrt{c^2-1}} (-\cos(\theta) + c \sin(y_n)), \end{aligned}$$

where  $y_n = \frac{cx_n}{\sqrt{c^2-1}}$ .

*Remark 11.* Notice that not all geodesics have conjugate points. Indeed,  $\text{Exp}(\theta, c, \cdot)$  has a conjugate point if and only if  $c \in (-\infty, -1) \cup (1, +\infty)$ .

We present an image of the 2nd conjugate locus (Figure 7). For simplicity we present an image of its intersection with  $\{e^{\mathbf{k}} e^{ap_1} | a \in \mathbb{R}\}$ . The complete picture can be recovered using the cylindrical symmetry.

*Remark 12.* Notice that all even conjugate loci have self-intersection.

#### 4. A sub-Riemannian structure on lens spaces.

**4.1. Definition of  $L(p, q)$ .** Fix two coprime integers  $p, q \in \mathbb{Z}$ ,  $p, q \neq 0$ . The **lens space**  $L(p, q)$  is defined as the quotient of  $SU(2)$  with respect to the identification rule

$$\begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix} \sim \begin{pmatrix} \alpha_2 \\ \beta_2 \end{pmatrix} \text{ if } \exists \omega \in \mathbb{C} \text{ } p\text{th root of unity such that}$$

$$\begin{pmatrix} \alpha_2 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \omega & 0 \\ 0 & \omega^q \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix}.$$

Lens spaces are three-dimensional compact manifolds, but, except for  $L(2, 1) \simeq SO(3)$ , they are neither Lie groups nor homogeneous spaces of  $SU(2)$ . The following topological equivalences hold:  $\forall p, q, k \in \mathbb{Z}$ ,  $p, q$  coprime,  $p, q \neq 0$ , we have  $L(p, q) \simeq L(p, -q) \simeq L(-p, q) \simeq L(p, q + kp)$ . Lens spaces have highly nontrivial topology; for details we refer the reader to [25].

The following theorem permits us to choose a representative of  $L(p, q)$  in  $SU(2)$ .

PROPOSITION 8. *Consider the set*

$$E_p = \left\{ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in SU(2) \mid \operatorname{Re}(\alpha) > 0, \frac{\operatorname{Im}(\alpha)^2}{\sin\left(\frac{\pi}{p}\right)^2} + |\beta|^2 < 1 \right\} \subset SU(2)$$

and define  $\partial E_p^+ = \partial E_p \cap \{\operatorname{Im}(\alpha) \geq 0\}$ ,  $\partial E_p^- = \partial E_p \cap \{\operatorname{Im}(\alpha) \leq 0\}$ . Endow  $\overline{E_p}$  with the equivalence relation  $\div$  defined as follows:

1. The relation is reflexive;
2. moreover, given  $\begin{pmatrix} \alpha^+ \\ \beta^+ \end{pmatrix} \in \partial E_p^+$  and  $\begin{pmatrix} \alpha^- \\ \beta^- \end{pmatrix} \in \partial E_p^-$ , we have  $\begin{pmatrix} \alpha^+ \\ \beta^+ \end{pmatrix} \div \begin{pmatrix} \alpha^- \\ \beta^- \end{pmatrix}$  if
  - (i) either  $\operatorname{Im}(\alpha^+) = -\operatorname{Im}(\alpha^-) \neq 0$  and  $\beta^+ = e^{2\pi i \frac{q}{p}} \beta^-$ ; or
  - (ii)  $\operatorname{Im}(\alpha^+) = \operatorname{Im}(\alpha^-) = 0$  and  $\beta^+ = e^{2\pi i \frac{n}{p}} \beta^-$  for some  $n \in \{1, \dots, p\}$ .

The manifold  $\overline{E_p}/\div$  is diffeomorphic to  $L(p, q)$ .

*Proof.* Take  $\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in SU(2)$  and let us look for  $\omega$  pth root of unity such that  $\begin{pmatrix} \omega\alpha \\ \omega^q\beta \end{pmatrix} \in \overline{E_p}$ . This condition is equivalent to

$$(13) \quad \operatorname{Re}(\omega\alpha) \geq 0 \text{ and } \frac{\operatorname{Im}(\omega\alpha)^2}{\sin\left(\frac{\pi}{p}\right)^2} + |\omega^q\beta|^2 \leq 1.$$

Recalling that  $|\omega^q\beta|^2 = |\beta|^2 = 1 - |\alpha|^2$  and that  $\operatorname{Im}(\omega\alpha) = |\alpha| \sin(\arg(\omega\alpha))$  if  $\alpha \neq 0$ , we see that (13) is equivalent to

$$(14) \quad \arg(\omega\alpha) \in \left[-\frac{\pi}{p}, \frac{\pi}{p}\right] \quad \text{or} \quad \alpha = 0.$$

Thus,

(i) if  $\alpha \neq 0$ , there exists at least one solution  $\omega_1$  of  $\arg(\omega\alpha) \in \left[-\frac{\pi}{p}, \frac{\pi}{p}\right]$ . Moreover, we have two distinct solutions  $\omega_1, \omega_2$  if and only if  $\arg(\omega_1\alpha) = -\frac{\pi}{p}$  and  $\arg(\omega_2\alpha) = \frac{\pi}{p}$ . In this case,

$$\begin{pmatrix} \omega_1\alpha \\ \omega_1^q\beta \end{pmatrix} = \begin{pmatrix} |\alpha|e^{-i\frac{\pi}{p}} \\ \omega_1^q\beta \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \omega_2\alpha \\ \omega_2^q\beta \end{pmatrix} = \begin{pmatrix} |\alpha|e^{i\frac{\pi}{p}} \\ \omega_2^q\beta \end{pmatrix};$$

observe that

$$\begin{pmatrix} \omega_1\alpha \\ \omega_1^q\beta \end{pmatrix} \div \begin{pmatrix} \omega_2\alpha \\ \omega_2^q\beta \end{pmatrix}.$$

(ii) if  $\alpha = 0$ , every  $\omega$  pth root of unity satisfies  $\begin{pmatrix} 0 \\ \omega^q\beta \end{pmatrix} \in \overline{E_p}$ ; observe that for all the pairs  $\omega_1, \omega_2$  we have

$$\begin{pmatrix} 0 \\ \omega_1^q\beta \end{pmatrix} \div \begin{pmatrix} 0 \\ \omega_2^q\beta \end{pmatrix}.$$

Hence  $\forall \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in SU(2)$  we have a unique

$$\left[ \begin{pmatrix} \omega\alpha \\ \omega^q\beta \end{pmatrix} \right]_{\div} \in \overline{E_p}/_{\div};$$

i.e., the function

$$\psi : \begin{array}{ccc} L(p, q) = SU(2)/\sim & \rightarrow & \overline{E_p}/_{\div}, \\ \left[ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \right] & \mapsto & \left[ \begin{pmatrix} \omega\alpha \\ \omega^q\beta \end{pmatrix} \right]_{\div} \end{array}$$

is bijective.  $\square$

*Remark 13.* A crucial observation for what follows is that the projection

$$\Pi : \begin{array}{ccc} SU(2) & \rightarrow & L(p, q), \\ g & \mapsto & [g] \end{array}$$

is a local diffeomorphism. Moreover,  $\Pi|_{E_p} : E_p \rightarrow L(p, q) \setminus [\partial E_p]$  is a diffeomorphism. In particular,  $E_p$  contains only one representative for each equivalence classes of  $L(p, q)$ ; i.e., if  $g, h \in E_p$  and  $[g] = [h]$ , then  $g = h$ .

*Remark 14.* Proposition 8 provides a picture of  $L(p, q)$ ; recall that  $SU(2)$  is drawn as two balls in  $\mathbb{R}^3$  (see section 3.1.2). Hence  $\overline{E_p} \subset SU(2)$  is drawn as a closed ellipsoid inside one of the two balls, via the map

$$\rho : \begin{array}{ccc} \overline{E_p} & \rightarrow & \overline{B_1(0)} \subset \mathbb{R}^3, \\ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} & \mapsto & (\operatorname{Re}(\beta), \operatorname{Im}(\beta), \operatorname{Im}(\alpha)). \end{array}$$

The picture of  $E_p$  is

$$F_p = \left\{ (x_1, x_2, x_3) \in \overline{B_1(0)} \mid x_1^2 + x_2^2 + \frac{x_3^2}{\sin\left(\frac{\pi}{p}\right)^2} < 1 \right\},$$

and the one of  $\overline{E_p}$  is  $\overline{F_p}$ ; see Figure 8 (left). The identification  $\div$  induces the following identification on  $\overline{F_p}$ : given  $(x_1^+, x_2^+, x_3^+) \in \partial F_p^+ = \partial F_p \cap \{x_3 \geq 0\}$  and  $(x_1^-, x_2^-, x_3^-) \in \partial F_p^- = \partial F_p \cap \{x_3 \leq 0\}$ , they are identified when

$$x_3^+ = -x_3^- \text{ and } \begin{pmatrix} x_1^+ \\ x_2^+ \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_1^- \\ x_2^- \end{pmatrix}$$

with  $\theta = \frac{2\pi q}{p}$ ; see Figure 8 (right).

*Remark 15.* Observe that the identification rule on  $\overline{F_p}$  gives a 1-to-1 identification between  $\partial F_p \cap \{x_3 > 0\}$  and  $\partial F_p \cap \{x_3 < 0\}$ , while there are, in general, more identified points on  $\{x_1^2 + x_2^2 = 1\} \cap \{x_3 = 0\}$ ; see Figure 9.

#### 4.2. Sub-Riemannian quotient structure on $L(p, q)$ .

**PROPOSITION 9.** *The sub-Riemannian structure on  $SU(2)$  given in section 3.1 induces a 2-dim sub-Riemannian structure on  $L(p, q) = SU(2)/\sim$  via the quotient map*

$$\Pi : \begin{array}{ccc} SU(2) & \rightarrow & L(p, q), \\ x & \mapsto & [x]; \end{array}$$



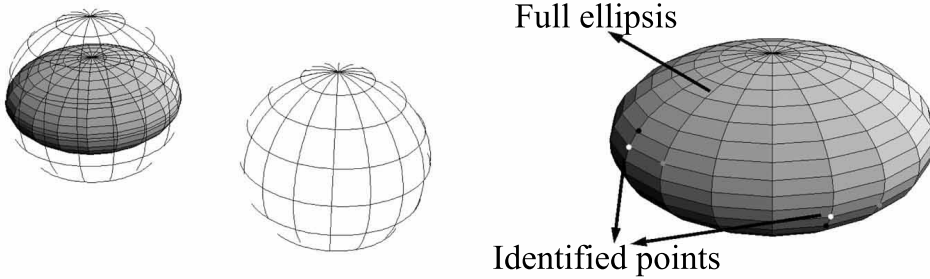


FIG. 8. Left:  $\overline{F}_4$ . Right: the representation of  $L(4,1)$ , with some examples of the identification rule.

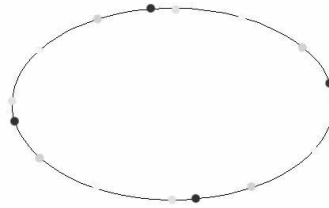


FIG. 9.  $L(4,1)$ : some examples of the identification rule on  $\{x_1^2 + x_2^2 = 1\} \cap \{x_3 = 0\}$ .

i.e.,

(i) the map

$$\tilde{\Delta} : [g] \mapsto \Pi_*(\Delta(h)) \subset T_{[g]}L(p, q) \text{ with } h \in [g]$$

is a 2-dim smooth distribution on  $L(p, q)$  that is Lie bracket generating;

(ii)  $\tilde{\mathbf{g}}_{[g]}(v_*, w_*) = \langle v_*, w_* \rangle_{[g]} := \langle v, w \rangle_h$  with  $h \in [g]$ ,  $v, w \in T_h SU(2)$ ,  $\Pi_*(v) = v_*$ ,  $\Pi_*(w) = w_*$  is a smooth positive definite scalar product on  $\tilde{\Delta}$ .

*Proof.* The role of the maps  $\Pi$  and  $\Pi_{*|_g}$  is illustrated in Figure 10.

The map  $\Pi$  is a local diffeomorphism; thus  $\Pi_{*|_g} : T_g SU(2) \rightarrow T_{[g]}L(p, q)$  is a linear isomorphism, and hence  $\Pi_{*|_g}(\Delta(g))$  is a 2-dim subspace of  $T_{[g]}L(p, q)$ .

The following two statements are consequences of Lemma 10, presented below:

(i) the distribution  $\tilde{\Delta}([g])$  is well defined; i.e.,  $\forall h_1, h_2 \in [g]$  we have

$$\Pi_{*|_{h_1}}(\Delta(h_1)) = \Pi_{*|_{h_2}}(\Delta(h_2)).$$

(ii) The positive definite scalar product  $\langle v_*, w_* \rangle_{[g]}$  is well defined; i.e.,  $\forall h_1, h_2 \in [g]$ ,  $v_1, w_1 \in T_{h_1} SU(2)$ ,  $v_2, w_2 \in T_{h_2} SU(2)$  such that  $\Pi_{*|_{h_1}}(v_1) = \Pi_{*|_{h_2}}(v_2)$  and  $\Pi_{*|_{h_1}}(w_1) = \Pi_{*|_{h_2}}(w_2)$ , we have  $\langle v_1, w_1 \rangle_{h_1} = \langle v_2, w_2 \rangle_{h_2}$ .

LEMMA 10. Let  $h_1, h_2 \in [g]$  with  $h_2 = \begin{pmatrix} \omega & 0 \\ 0 & \omega^q \end{pmatrix} h_1$ . The map

$$\phi : \begin{pmatrix} \mathbf{p} \\ n_1 \\ m_1 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} \mathbf{p} \\ \operatorname{Re}(\omega^{q-1}) & -\operatorname{Im}(\omega^{q-1}) & 0 \\ \operatorname{Im}(\omega^{q-1}) & \operatorname{Re}(\omega^{q-1}) & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} n_1 \\ m_1 \\ 0 \end{pmatrix}$$

$$\begin{array}{ccc}
 T_g SU(2) & \xrightarrow{\Pi_*|_g} & T_{[g]} L(p, q) \\
 \downarrow & & \downarrow \\
 SU(2) & \xrightarrow{\Pi} & L(p, q).
 \end{array}$$

FIG. 10. The role of the maps  $\Pi$  and  $\Pi_*$ .

is bijective. Moreover, it is an isometry with respect to the positive definite scalar product  $\langle \cdot, \cdot \rangle$  and satisfies,  $\forall \eta \in \mathfrak{p}$ ,

$$\frac{d}{dt}\Big|_{t=0} [h_1 e^{t\eta}] = \frac{d}{dt}\Big|_{t=0} [h_2 e^{t\phi(\eta)}].$$

*Proof.* Let  $h = \begin{pmatrix} a & \\ & b \end{pmatrix} \in SU(2)$  and  $\eta = (n, m, 0) \in \mathfrak{p}$ . We have

$$(15) \quad h e^{t\eta} = \begin{pmatrix} a \cos\left(\sqrt{n^2 + m^2} \frac{t}{2}\right) - b \sin\left(\sqrt{n^2 + m^2} \frac{t}{2}\right) \frac{n-im}{\sqrt{n^2+m^2}} \\ b \cos\left(\sqrt{n^2 + m^2} \frac{t}{2}\right) + a \sin\left(\sqrt{n^2 + m^2} \frac{t}{2}\right) \frac{n+im}{\sqrt{n^2+m^2}} \end{pmatrix}.$$

Take  $h_1, h_2 \in [g]$  with  $h_2 = \begin{pmatrix} \omega & 0 \\ 0 & \omega^q \end{pmatrix} h_1$  and  $\eta_1, \eta_2 \in \mathfrak{p}$  with coordinates  $\eta_1 = (n_1, m_1, 0)$  and  $\eta_2 = (n_2, m_2, 0)$ . Consider the trajectories

$$[h_1 e^{t\eta_1}] = \left[ \begin{pmatrix} a \cos\left(\sqrt{n_1^2 + m_1^2} \frac{t}{2}\right) - b \sin\left(\sqrt{n_1^2 + m_1^2} \frac{t}{2}\right) \frac{n_1-im_1}{\sqrt{n_1^2+m_1^2}} \\ b \cos\left(\sqrt{n_1^2 + m_1^2} \frac{t}{2}\right) + a \sin\left(\sqrt{n_1^2 + m_1^2} \frac{t}{2}\right) \frac{n_1+im_1}{\sqrt{n_1^2+m_1^2}} \end{pmatrix} \right]$$

and

$$\begin{aligned}
 [h_2 e^{t\eta_2}] &= \left[ \begin{pmatrix} \omega a \cos\left(\sqrt{n_2^2 + m_2^2} \frac{t}{2}\right) - \omega^q b \sin\left(\sqrt{n_2^2 + m_2^2} \frac{t}{2}\right) \frac{n_2-im_2}{\sqrt{n_2^2+m_2^2}} \\ \omega^q b \cos\left(\sqrt{n_2^2 + m_2^2} \frac{t}{2}\right) + \omega a \sin\left(\sqrt{n_2^2 + m_2^2} \frac{t}{2}\right) \frac{n_2+im_2}{\sqrt{n_2^2+m_2^2}} \end{pmatrix} \right] \\
 &= \left[ \begin{pmatrix} a \cos\left(\sqrt{n_2^2 + m_2^2} \frac{t}{2}\right) - \omega^{q-1} b \sin\left(\sqrt{n_2^2 + m_2^2} \frac{t}{2}\right) \frac{n_2-im_2}{\sqrt{n_2^2+m_2^2}} \\ b \cos\left(\sqrt{n_2^2 + m_2^2} \frac{t}{2}\right) + \omega^{1-q} a \sin\left(\sqrt{n_2^2 + m_2^2} \frac{t}{2}\right) \frac{n_2+im_2}{\sqrt{n_2^2+m_2^2}} \end{pmatrix} \right].
 \end{aligned}$$

Thus  $\frac{d}{dt}\Big|_{t=0} [h_1 e^{t\eta_1}] = \frac{d}{dt}\Big|_{t=0} [h_2 e^{t\eta_2}]$  in the case

$$\begin{cases} n_1^2 + m_1^2 = n_2^2 + m_2^2, \\ n_1 - im_1 = \omega^{q-1}(n_2 - im_2), \\ n_1 + im_1 = \omega^{1-q}(n_2 + im_2), \end{cases}$$

that is equivalent to  $\omega^{q-1}(n_1 + im_1) = n_2 + im_2$ . This equation is verified for  $\eta_2 = \phi(\eta_1)$ .  $\square$

Since  $\Pi$  is a local diffeomorphism,  $\forall g \in SU(2) \exists B(g)$  such that the map  $\Pi_{*|_{B(g)}} : T_{B(g)}SU(2) \rightarrow T_{B([g])}L(p, q)$  is a diffeomorphism, and thus  $\tilde{\Delta}$  is smooth and Lie bracket generating, and  $\langle v_*, w_* \rangle_{[g]}$  is smooth as a function of  $[g]$ .  $\square$

Proposition 9 implies that the sub-Riemannian structures on  $SU(2)$  and  $L(p, q)$  defined above are locally isometric via the map  $\Pi$ . As a consequence, the geodesics of  $(L(p, q), \tilde{\Delta}, \tilde{\mathbf{g}})$  are the projection of geodesics of  $(SU(2), \Delta, \mathbf{g})$ . The conjugate locus for  $L(p, q)$  can be obtained from that of  $SU(2)$  by the projection  $\Pi$ .

*Remark 16.* One can check that the sub-Riemannian structure induced by  $SU(2)$  on  $L(2, 1) \simeq SO(3)$  is equivalent to the  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian structure on  $SO(3)$  defined in section 3.2.

**5. Cut loci and distances.** In this section we prove the main theorems of the paper; i.e., we compute cut loci for  $SU(2)$ ,  $SO(3)$ , lens spaces, and  $SL(2)$ , and we prove the formula (3) for the sub-Riemannian distance on  $SU(2)$ .

Recall that our problems satisfy the following assumptions:

(i) Each point of  $M$  is reached by an optimal geodesic starting from  $\text{Id}$ ; see section 2.3;

(ii) we are in the three-dimensional contact case, and thus there are no abnormal minimizers. Hence Remark 4 applies.

**PROPOSITION 11.** *Let  $T(\theta, c)$  be the cut time for  $\text{Exp}(\theta, c, \cdot)$  (possibly  $+\infty$  if  $\text{Exp}(\theta, c, \cdot)$  is optimal on  $[0, +\infty)$ ). Define*

$$\mathcal{D} = \{(\theta, c, t) \in \Lambda_{\text{Id}} \times \mathbb{R}^+ \mid 0 < t < T(\theta, c)\}$$

*and  $M' = M \setminus (K_{\text{Id}} \cup \text{Id})$ . The function  $\text{Exp}|_{\mathcal{D}} : \mathcal{D} \rightarrow M'$  is a diffeomorphism from  $\mathcal{D}$  to  $M'$ .*

*Proof.* Let us first check that  $\text{Exp}(\mathcal{D}) \subset M'$ . By contradiction, let  $\text{Exp}(\theta, c, t) \in M \setminus M'$ ; thus  $t = 0$  or  $t = T(\theta, c)$  or  $\text{Exp}(\theta, c, \cdot)$  is not optimal in  $[0, t]$ , i.e.,  $t > T(\theta, c)$ . This is a contradiction. Let us verify that  $\text{Exp}|_{\mathcal{D}}$  is injective; by contradiction, let  $\text{Exp}(\theta_1, c_1, t_1) = \text{Exp}(\theta_2, c_2, t_2)$  with  $(\theta_1, c_1, t_1) \neq (\theta_2, c_2, t_2)$ . If  $t_1 \neq t_2$ , one of the two geodesics  $\text{Exp}(\theta_1, c_1, \cdot), \text{Exp}(\theta_2, c_2, \cdot)$  has already lost optimality, and thus  $t_i \geq T(\theta_i, c_i)$ ; hence  $(\theta_i, c_i, t_i) \notin \mathcal{D}$ , a contradiction. If  $t_1 = t_2$ , we have that  $\text{Exp}(\theta_1, c_1, t_1)$  is a cut point, and hence  $t_1 \geq T(\theta_1, c_1)$ , a contradiction. To verify that  $\text{Exp}|_{\mathcal{D}}$  is surjective, take  $g \in M'$  and observe that there is an optimal geodesic  $\text{Exp}(\theta, c, \cdot)$  reaching it at time  $t \leq T(\theta, c)$ . But  $t = T(\theta, c)$  implies  $g \in K_{\text{Id}}$ ; thus  $t < T(\theta, c)$ .

The smoothness of  $\text{Exp}|_{\mathcal{D}}$  and of its inverse follows from the facts that  $\text{Exp}$  is a local diffeomorphism outside the critical points (i.e., points where the differential of  $\text{Exp}$  is not of full rank) and that the critical points do not belong to  $\mathcal{D}$ . Indeed, by contradiction, let  $(\theta, c, t) \in \mathcal{D}$  be a critical point, and hence  $t$  is a conjugate time as follows: it is either the first conjugate time that coincides with the cut time (i.e.,  $t = T(\theta, c)$ ) or a greater conjugate time (i.e.,  $t > T(\theta, c)$ ). In both cases  $(\theta, c, t) \notin \mathcal{D}$ , a contradiction.  $\square$

### 5.1. The cut locus for $SU(2)$ .

**THEOREM 12.** *The cut locus for the  $\mathbf{k} \oplus \mathbf{p}$  problem on  $SU(2)$  is*

$$K_{\text{Id}} = e^{\mathbf{k}} \setminus \text{Id} = \{e^{c\mathbf{k}} \mid c \in (0, 4\pi)\}.$$

*Proof.* Let

$$g \in e^{\mathbf{k}} \setminus \text{Id} = \left\{ \begin{pmatrix} \alpha \\ 0 \end{pmatrix} \mid \alpha \in \mathbb{C}, |\alpha| = 1, \alpha \neq 1 \right\},$$

and let  $\text{Exp}(\theta, c, \cdot)$  be the minimizing geodesic steering  $\text{Id}$  to  $g$  in time  $T$ . As a consequence of the cylindrical symmetry, we have that  $\text{Exp}(\psi, c, T) = g \quad \forall \psi \in \mathbb{R}/2\pi$ ; thus  $(e^{\mathbf{k}} \backslash \text{Id}) \subset K_{\text{Id}}$ .

The core of the proof is to show that there are no cut points outside  $e^{\mathbf{k}}$ . Recall the expression of geodesics given in section 3.1.1. By contradiction, assume that  $g \in SU(2) \backslash e^{\mathbf{k}}$  is reached by two different optimal trajectories  $\text{Exp}(\theta, c, \cdot)$  and  $\text{Exp}(\psi, d, \cdot)$  at time  $T$ . Observe that  $\text{Exp}(\theta, c, \frac{2\pi}{\sqrt{1+c^2}})$  and  $\text{Exp}(\psi, d, \frac{2\pi}{\sqrt{1+d^2}}) \in e^{\mathbf{k}} \subset K_{\text{Id}}$ ; thus

$$(16) \quad 0 < T < \min \left\{ \frac{2\pi}{\sqrt{1+c^2}}, \frac{2\pi}{\sqrt{1+d^2}} \right\}.$$

Observe that  $\text{Exp}(\theta, c, T) = \text{Exp}(\psi, d, T)$  implies that  $|\beta|$  is equal in the two cases, i.e.,

$$(17) \quad \frac{\sin(\frac{\sqrt{1+c^2}T}{2})}{\sqrt{1+c^2}} = \frac{\sin(\frac{\sqrt{1+d^2}T}{2})}{\sqrt{1+d^2}}.$$

From this equation it follows that  $|c| = |d|$ . Indeed (17) is equivalent to

$$\frac{\sin(\frac{\sqrt{1+c^2}T}{2})}{\frac{\sqrt{1+c^2}T}{2}} = \frac{\sin(\frac{\sqrt{1+d^2}T}{2})}{\frac{\sqrt{1+d^2}T}{2}}.$$

From the facts that  $\frac{\sqrt{1+c^2}T}{2}, \frac{\sqrt{1+d^2}T}{2} \in (0, \pi)$  and that the function  $\frac{\sin p}{p}$  is injective for  $p \in (0, \pi)$ , it follows that  $\frac{\sqrt{1+c^2}T}{2} = \frac{\sqrt{1+d^2}T}{2}$ , and hence  $|c| = |d|$ .

Thus we consider the following two cases:

(i)  $c = d \in \mathbb{R}$ : the cylindrical symmetry implies that either  $\theta = \psi$  (so the two geodesics coincide) or  $g \in e^{\mathbf{k}}$ . This is a contradiction.

(ii)  $c = -d \in \mathbb{R} \setminus \{0\}$ : with no loss of generality we assume  $c > 0$ . Since by the central and cylindrical symmetries, we have

$$\text{Exp}(\psi, -c, t) = \left( e^{i(\psi+\theta-\arg(\beta))\beta} \bar{\alpha} \right), \quad \text{where} \quad \text{Exp}(\theta, c, t) = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

$\text{Exp}(\theta, c, t) = \text{Exp}(\psi, -c, t)$  implies  $\text{Im}(\alpha) = 0$ . Hence

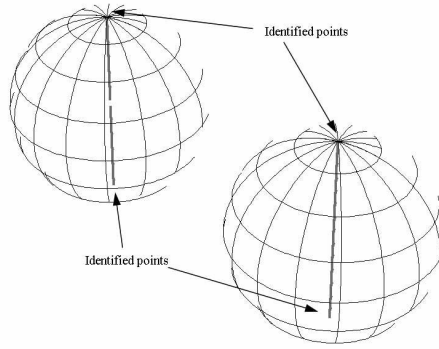
$$(18) \quad c \cos\left(\frac{ct}{2}\right) \sin\left(\frac{\sqrt{1+c^2}t}{2}\right) = \sqrt{1+c^2} \sin\left(\frac{ct}{2}\right) \cos\left(\frac{\sqrt{1+c^2}t}{2}\right).$$

The terms  $c, \sin(\frac{\sqrt{1+c^2}t}{2}), \sqrt{1+c^2}, \sin(\frac{ct}{2})$  are nonzero because of (16) and  $c < \sqrt{1+c^2}$ . Thus  $\cos(\frac{ct}{2}) = 0$  if and only if  $\cos(\frac{\sqrt{1+c^2}t}{2}) = 0$ , which is impossible because  $0 < \frac{ct}{2} < \frac{\sqrt{1+c^2}t}{2} < \pi$ . Hence we rewrite (18) as

$$\frac{\tan\left(\frac{\sqrt{1+c^2}t}{2}\right)}{\frac{\sqrt{1+c^2}t}{2}} = \frac{\tan\left(\frac{ct}{2}\right)}{\frac{ct}{2}}$$

and state that  $(0, \tan(0)), (\frac{ct}{2}, \tan(\frac{ct}{2})), (\frac{\sqrt{1+c^2}t}{2}, \tan(\frac{\sqrt{1+c^2}t}{2}))$  are three distinct points aligned on the graph of the function  $\tan$  in  $[0, \pi)$ , which is impossible. This is a contradiction.  $\square$

The cut locus for the  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifold  $SU(2)$  is given in Figure 11.

FIG. 11. The cut locus for the  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifold  $SU(2)$ .

**5.1.1. The sub-Riemannian distance in  $SU(2)$ .** In this section we compute the sub-Riemannian distance on  $SU(2)$ , i.e., we prove Theorem 2.

Let

$$g = \begin{pmatrix} \alpha \\ 0 \end{pmatrix} = \begin{pmatrix} e^{i \arg(\alpha)} \\ 0 \end{pmatrix} \in e^{\mathbf{k}}.$$

$g$  is reached by a geodesic  $\text{Exp}(\theta, c, \cdot)$  at time  $\frac{2\pi}{\sqrt{1+c^2}}$  for some  $c \in \mathbb{R}$ . Observe that

$$(19) \quad \begin{aligned} \text{Exp}\left(\theta, c, \frac{2\pi}{\sqrt{1+c^2}}\right) &= \text{Exp}\left(\theta, \pm \sqrt{\frac{4\pi^2}{t^2} - 1}, t\right) \\ &= \begin{pmatrix} -\cos\left(\sqrt{\pi^2 - \frac{t^2}{4}}\right) \mp i \sin\left(\sqrt{\pi^2 - \frac{t^2}{4}}\right) \\ 0 \end{pmatrix} = \begin{pmatrix} e^{i\left(\pi \pm \sqrt{\pi^2 - \frac{t^2}{4}}\right)} \\ 0 \end{pmatrix}. \end{aligned}$$

Thus the distance  $d(g, \text{Id})$  is the smallest  $t > 0$  such that  $e^{i\left(\pi \pm \sqrt{\pi^2 - \frac{t^2}{4}}\right)} = e^{i \arg(\alpha)}$ , whose solution is  $t = 2\sqrt{\arg(\alpha)(2\pi - \arg(\alpha))}$ , where  $\arg(\alpha)$  is chosen in  $[0, 2\pi]$ .

Let  $g = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in SU(2) \setminus e^{\mathbf{k}}$ . Applying Proposition 11, we have that  $\text{Exp}_{|\mathcal{D}}^{-1}(g)$  is well defined on  $\mathcal{D} = \left\{(\theta, c, t) \in \Lambda_{\text{Id}} \times \mathbb{R}^+ \mid 0 < t < \frac{2\pi}{\sqrt{1+c^2}}\right\}$ . Thus the sub-Riemannian distance of  $g$  from the origin is  $d(g, \text{Id}) = t$ , where  $t$  is the third component of  $\text{Exp}_{|\mathcal{D}}^{-1}(g)$ , i.e., the unique solution  $t$  of  $\text{Exp}(\theta, c, t) = g$  with  $(\theta, c, t) \in \mathcal{D}$ . Using the explicit form of  $\text{Exp}$  given in (3.1.1), one checks that the system

$$\begin{cases} \text{Exp}(\theta, c, t) = g, \\ (\theta, c, t) \in \mathcal{D} \end{cases}$$

is equivalent to

$$\begin{cases} -\frac{ct}{2} + \arctan\left(\frac{c}{\sqrt{1+c^2}} \tan\left(\frac{\sqrt{1+c^2}t}{2}\right)\right) = \arg(\alpha), \\ \frac{\sin\left(\frac{\sqrt{1+c^2}t}{2}\right)}{\sqrt{1+c^2}} = \sqrt{1-|\alpha|^2}, \\ \cos\left(\frac{ct}{2} + \theta\right) + i \sin\left(\frac{ct}{2} + \theta\right) = \arg(\beta). \end{cases}$$

The third equation has no role in the computation of distance as a consequence of the cylindrical symmetry.

*Remark 17.* The distance is a bounded function; this is due to its continuity and the compactness of  $SU(2)$ . The farthest point starting from  $\text{Id}$  is  $-\text{Id}$ , whose distance is  $2\pi$ .

Notice that  $\forall \alpha, \beta_1, \beta_2 \in \mathbb{C}, |\beta_1| = |\beta_2|$ , we have

$$d\left(\begin{pmatrix} \alpha \\ \beta_1 \end{pmatrix}, \text{Id}\right) = d\left(\begin{pmatrix} \alpha \\ \beta_2 \end{pmatrix}, \text{Id}\right) = d\left(\begin{pmatrix} \bar{\alpha} \\ \beta_1 \end{pmatrix}, \text{Id}\right).$$

This is due to the cylindrical and central symmetries.

**5.2. The cut locus for  $SO(3)$  and lens spaces.** In this section we compute the cut locus for lens spaces  $L(p, q)$ . As a particular case, we get the cut locus for  $SO(3) \simeq L(2, 1)$ .

**THEOREM 13.** *The cut locus for the sub-Riemannian problem on  $L(p, q)$  defined in section 9 is a stratification*

$$K_{[\text{Id}]} = K_{[\text{Id}]}^{\text{sym}} \cup K_{[\text{Id}]}^{\text{loc}}$$

with

$$K_{[\text{Id}]}^{\text{sym}} = [\partial E_p] = \left\{ \left[ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \right] \mid a, b \in \mathbb{C}, \text{Re}(\alpha) \geq 0, \frac{\text{Im}(\alpha)^2}{\sin\left(\frac{\pi}{p}\right)^2} + |\beta|^2 = 1 \right\},$$

$$K_{[\text{Id}]}^{\text{loc}} = [e^{\mathbf{k}}] \setminus [\text{Id}] = \left\{ \left[ \begin{pmatrix} \alpha \\ 0 \end{pmatrix} \right] \mid \alpha \in \mathbb{C}, |\alpha| = 1, \alpha^p \neq 1 \right\}.$$

*Proof.* Let us first prove the following lemma.

**LEMMA 14.** *A geodesic  $\gamma(\cdot)$  in  $L(p, q)$  steering  $[\text{Id}]$  to  $[g]$  in minimum time  $T$  admits a unique lift  $\gamma_0(\cdot)$  in  $SU(2)$  starting from  $\text{Id}$ .*

*Moreover,*  $\gamma_0(t) = \text{Exp}(\theta_0, c_0, t) \quad \forall t \in [0, T]$  *for some*  $\theta_0 \in \mathbb{R}/2\pi, c \in \mathbb{R}$ .

*Proof.* Take  $\gamma(\cdot)$  as in the hypotheses. Since  $L(p, q)$  and  $SU(2)$  are locally diffeomorphic via  $\Pi$ , there is a unique lift  $\gamma_0(\cdot)$  in  $SU(2)$  starting from  $\text{Id}$ ; i.e.,  $\gamma_0(0) = \text{Id}$  and  $[\gamma_0(t)] = \gamma(t) \quad \forall t \in [0, T]$ .

Let us prove that  $\gamma_0(\cdot)$  is an optimal trajectory reaching  $\gamma_0(T)$ . By contradiction, there exists a trajectory  $\gamma_1(\cdot)$  such that  $\gamma_1(t_1) = \gamma_0(T)$  with  $t_1 < T$ . Hence, its projection  $[\gamma_1(\cdot)]$  satisfies  $[\gamma_1(t_1)] = [g]$  with  $t_1 < T$ , a contradiction.

Since  $\gamma_0(\cdot)$  is an optimal trajectory, it is a geodesic of  $SU(2)$ , and there exist  $\theta \in \mathbb{R}/2\pi, c \in \mathbb{R}$  such that  $\gamma_0(t) = \text{Exp}(\theta_0, c_0, t) \quad \forall t \in [0, T]$ .  $\square$

Let us prove that  $K_{[\text{Id}]}^{\text{loc}} \subset K_{[\text{Id}]}$ . Consider  $[g] \in K_{[\text{Id}]}^{\text{loc}}$ , a geodesic steering  $[\text{Id}]$  to  $[g]$  in minimum time  $T$  with unique lift  $\text{Exp}(\theta_0, c_0, \cdot)$ . By the definition of  $K_{[\text{Id}]}^{\text{loc}}$ , we have  $\text{Exp}(\theta_0, c_0, T) \in e^{\mathbf{k}} \setminus \text{Id} \subset SU(2)$ ; i.e.,  $\text{Exp}(\theta_0, c_0, T)$  lies in the cut locus for the sub-Riemannian problem on  $SU(2)$ . Thus there exists another optimal geodesic  $\text{Exp}(\theta_1, c_1, \cdot)$  defined in  $[0, T]$  such that  $\text{Exp}(\theta_1, c_1, T) = \text{Exp}(\theta_0, c_0, T) \in [g]$ . Thus the geodesic  $[\text{Exp}(\theta_1, c_1, \cdot)]$  reaches  $[g]$  in minimum time. The geodesics in  $SU(2)$  are distinct in a neighborhood of  $\text{Id}$ , so their projections in a neighborhood of  $[\text{Id}]$  are distinct as well.

Let us now prove that  $K_{[\text{Id}]}^{\text{sym}} \subset K_{[\text{Id}]}$ . Consider  $[g] \in K_{[\text{Id}]}^{\text{sym}}$ , a geodesic steering  $[\text{Id}]$  to  $[g]$  in minimum time  $T$  with unique lift  $\text{Exp}(\theta_0, c_0, \cdot)$ ; call  $\text{Exp}(\theta_0, c_0, T) = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in [g]$ . If  $\beta = 0$ , we have  $[g] \in K_{[\text{Id}]}^{\text{loc}}$  or  $[g] = [\text{Id}]$ , so assume  $\beta \neq 0$ . Due to the cylindrical and central symmetries, we have

$$\text{Exp}(\theta_0 + \psi, -c_0, T) = \begin{pmatrix} \bar{\alpha} \\ e^{2i(\theta_0 - \arg(\beta)) + i\psi} \beta \end{pmatrix}.$$

Consider  $\psi^+ \in \mathbb{R}/2\pi$  being a solution of  $e^{2i(\theta_0 - \arg(\beta)) + i\psi^+} = e^{2\pi i \frac{q}{p}}$  and  $\psi^- \in \mathbb{R}/2\pi$  being a solution of  $e^{2i(\theta_0 - \arg(\beta)) + i\psi^-} = e^{-2\pi i \frac{q}{p}}$ . If  $\text{Exp}(\theta_0, c_0, T) \in \partial E_p^+$ , we have  $[\text{Exp}(\theta_0 + \psi^+, -c_0, T)] = [\text{Exp}(\theta_0, c_0, T)] = [g]$ ; if  $\text{Exp}(\theta_0, c_0, T) \in \partial E_p^-$ , we have similarly  $[\text{Exp}(\theta_0 + \psi^-, -c_0, T)] = [\text{Exp}(\theta_0, c_0, T)] = [g]$ . If  $c_0 \neq 0$ , we have found two distinct trajectories reaching  $[g]$  in optimal time; if  $c_0 = 0$ , we have  $\text{Exp}(\theta_0, c_0, T) \in \partial E_p^+ \cap \partial E_p^-$ , and thus at least one of  $\psi^+$  and  $\psi^-$  is not null, so at least one of  $\text{Exp}(\theta_0 + \psi^+, 0, \cdot)$  and  $\text{Exp}(\theta_0 + \psi^-, 0, \cdot)$  is distinct from  $\text{Exp}(\theta_0, 0, \cdot)$  in a neighborhood of  $\text{Id}$ , as are their projections in a neighborhood of  $[\text{Id}]$ .

Finally, consider  $[g] \in L(p, q) \setminus (K_{[\text{Id}]}^{\text{loc}} \cup K_{[\text{Id}]}^{\text{sym}} \cup [\text{Id}])$  and assume by contradiction that there exist two distinct geodesics steering  $[\text{Id}]$  to  $[g]$  in minimum time  $T$  with distinct lifts  $\text{Exp}(\theta_0, c_0, \cdot)$ ,  $\text{Exp}(\theta_1, c_1, \cdot)$ . There are two possibilities as follows:

(i)  $\text{Exp}(\theta_0, c_0, T) = \text{Exp}(\theta_1, c_1, T)$ . In this case,  $\text{Exp}(\theta_0, c_0, T)$  lies in the cut locus for the sub-Riemannian problem on  $SU(2)$ , and hence  $[g] \in K_{[\text{Id}]}^{\text{loc}}$ , a contradiction.

(ii)  $\text{Exp}(\theta_0, c_0, T) \neq \text{Exp}(\theta_1, c_1, T)$ . Since by hypothesis  $[g] \notin K_{[\text{Id}]}^{\text{sym}}$ , we have  $\text{Exp}(\theta_0, c_0, T), \text{Exp}(\theta_1, c_1, T) \notin \partial E_p$ . Recall that, if  $[\text{Exp}(\theta_0, c_0, T)] = [\text{Exp}(\theta_1, c_1, T)]$  and  $\text{Exp}(\theta_0, c_0, T), \text{Exp}(\theta_1, c_1, T) \in E_p$ , then  $\text{Exp}(\theta_0, c_0, T) = \text{Exp}(\theta_1, c_1, T)$ , due to Remark 13. Thus we have that  $\text{Exp}(\theta_i, c_i, T) \in SU(2) \setminus \bar{E}_p$  for  $i = 0$  or  $i = 1$ . We assume without loss of generality that  $\text{Exp}(\theta_0, c_0, T) \in SU(2) \setminus \bar{E}_p$ ; thus the geodesic  $\text{Exp}(\theta_0, c_0, t)$  with  $t \in [0, T]$  steers  $\text{Id} \in E_p$  to  $\text{Exp}(\theta_0, c_0, T) \in SU(2) \setminus \bar{E}_p$ , and hence  $\exists \tilde{t} \in (0, T)$  such that  $\text{Exp}(\theta_0, c_0, \tilde{t}) \in \partial E_p$ . Then we have that  $\gamma_0(\tilde{t}) = [\text{Exp}(\theta_0, c_0, \tilde{t})] \in K_{[\text{Id}]}^{\text{sym}}$ , and thus  $\gamma_0(t)$  is no more optimal for  $t \in [0, T]$ , a contradiction.  $\square$

*Remark 18.* Notice that  $K_{[\text{Id}]}^{\text{loc}}$  is a manifold (a circle without a point), while  $K_{[\text{Id}]}^{\text{sym}}$  is not in general. Indeed, it is an orbifold. It can be seen as  $S^2 \subset \mathbb{R}^3$  with the following identification:  $(x_1^+, x_2^+, x_3^+) \in S^2 \cap \{x_3 \geq 0\}$  and  $(x_1^-, x_2^-, x_3^-) \in S^2 \cap \{x_3 \leq 0\}$  are identified when  $x_3^+ = -x_3^-$  and

$$\begin{pmatrix} x_1^+ \\ x_2^+ \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_1^- \\ x_2^- \end{pmatrix}$$

with  $\theta = \frac{2\pi q}{p}$ . In the case  $SO(3) \simeq L(2, 1)$ , we have that  $K_{[\text{Id}]}^{\text{sym}} = \mathbb{RP}^2$  (see Figure 12 (left)), while in the other cases it is not locally Euclidean; in fact, take a neighborhood of a point  $P$  on the equator and observe that it is topologically equivalent to a set of  $p$  half-planes with a common line as the boundary.

Next we give an idea of the topology of the cut locus for  $L(4, 1)$ . Consider the space  $T_1$  made by the two intersecting strips  $\{(a, b, 0) \in \mathbb{R}^3 \mid a, b \in [-1, 1]\}$  and  $\{(a, 0, b) \in \mathbb{R}^3 \mid a, b \in [-1, 1]\}$  with the following identification:  $(-1, b, 0) \sim (1, 0, b)$  and  $(-1, 0, b) \sim (1, -b, 0)$ . The boundary of this set is topologically a circle  $S^1$ . Consider now a two-dimensional semisphere  $T_2$ . The cut locus  $K_{[\text{Id}]}^{\text{sym}}$  is topologically equivalent to the space given by gluing  $T_1$  and  $T_2$  along their boundaries  $S^1$ . The cut locus  $K_{[\text{Id}]}$  is given by gluing  $K_{[\text{Id}]}^{\text{sym}}$  with a circle  $S^1$  along a point on  $T_2$  and then removing

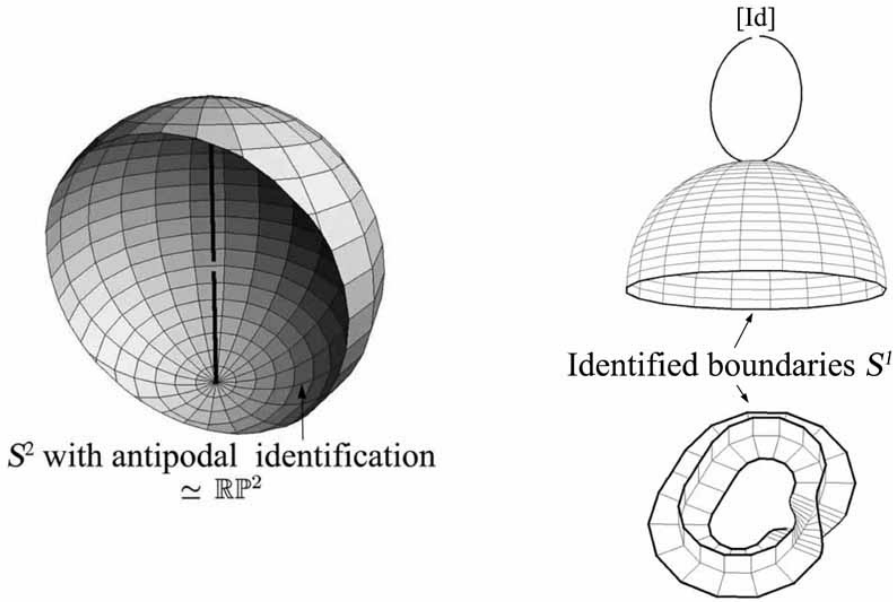


FIG. 12. Left: The cut locus for the sub-Riemannian problem on  $SO(3)$ . Right: The cut locus for the sub-Riemannian problem on  $L(4, 1)$ .

a point on  $S^1$  (the starting point). See a picture of it in Figure 12 (right).

### 5.3. The cut locus for $SL(2)$ .

THEOREM 15. The cut locus for the  $\mathbf{k} \oplus \mathbf{p}$  problem on  $SL(2)$  is a stratification

$$K_{\text{Id}} = K_{\text{Id}}^{\text{sym}} \cup K_{\text{Id}}^{\text{loc}}$$

with

$$K_{\text{Id}}^{\text{sym}} = e^{2\pi k} e^{\mathbf{p}} = \{g \in SL(2) \mid g = g^T, \text{Tr} g < 0\},$$

$$K_{\text{Id}}^{\text{loc}} = e^{\mathbf{k}} \setminus \text{Id} = \left\{ \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \mid \alpha \in R/2\pi, \alpha \neq 0 \right\}.$$

*Proof.* Let us first prove that  $K_{\text{Id}}^{\text{loc}} \subset K_{\text{Id}}$ . Let  $g \in e^{\mathbf{k}} \setminus \text{Id}$ ; it is reached optimally by a geodesic  $\text{Exp}(\theta, c, \cdot)$  at time  $T$ . Due to the cylindrical symmetry, we have  $g = \text{Exp}(\psi, c, T) \forall \psi \in \mathbb{R}/2\pi$ ; thus  $g \in K_{\text{Id}}$ .

Let us now prove that  $K_{\text{Id}}^{\text{sym}} \subset K_{\text{Id}}$ . Let  $g = e^{2\pi k} e^{x_0 p_1 + y_0 p_2} \in e^{2\pi k} e^{\mathbf{p}}$ ; it is reached optimally by a geodesic  $\text{Exp}(\theta, c, \cdot)$  at time  $T$ . If  $x_0^2 + y_0^2 = 0$ , we have  $g = e^{2\pi k} \in K_{\text{Id}}^{\text{loc}}$ ; thus it is a cut point. If  $x_0^2 + y_0^2 \neq 0$ , due to the cylindrical and central symmetry, we have  $\text{Exp}(\theta + \psi, -c, T) = e^{-2\pi k} e^{x p_1 + y p_2}$  with

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(2\theta + \psi) & \sin(2\theta + \psi) \\ \sin(2\theta + \psi) & -\cos(2\theta + \psi) \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}.$$



Choose  $\psi$  in such a way that  $\theta + \frac{\psi}{2}$  is the angle on the plane of the line passing through  $(0, 0)$  and  $(x_0, y_0)$ . In this way we have  $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ . Observing that  $e^{-2\pi k} = e^{2\pi k}$ , we finally have that  $g = \text{Exp}(\theta, c, T) = \text{Exp}(\theta + \psi, -c, T)$ . Observe that  $c \neq 0$  because  $\text{Exp}(\theta, 0, \cdot) \in e^{\mathbf{P}}$ ; thus the two geodesics  $\text{Exp}(\theta, c, \cdot)$ ,  $\text{Exp}(\theta + \psi, -c, \cdot)$  are distinct.

We now prove that there is no cut point outside  $K_{\text{Id}}^{\text{sym}} \cup K_{\text{Id}}^{\text{loc}}$ . By contradiction, let  $g \in SL(2) \setminus (K_{\text{Id}}^{\text{sym}} \cup K_{\text{Id}}^{\text{loc}} \cup \text{Id})$  be reached by two optimal trajectories  $\text{Exp}(\theta, c, \cdot)$  and  $\text{Exp}(\psi, d, \cdot)$  at time  $T$ . Writing

$$\text{Exp}(\theta, c, t) = \begin{pmatrix} g_{11}(\theta, c, t) & g_{12}(\theta, c, t) \\ g_{21}(\theta, c, t) & g_{22}(\theta, c, t) \end{pmatrix},$$

we have

$$(20) \quad r(c, t) := \sqrt{(g_{11} - g_{22})^2 + (g_{12} + g_{21})^2} = \begin{cases} \frac{\text{Sinh}(\frac{\sqrt{1-c^2}}{2} t)}{\frac{\sqrt{1-c^2}}{2}}, & c \in (-1, 1), \\ t, & c \in \{-1, 1\}, \\ \frac{\sin(\frac{\sqrt{c^2-1}}{2} t)}{\frac{\sqrt{c^2-1}}{2}}, & c \in (-\infty, -1) \cup (1, +\infty). \end{cases}$$

The identity  $\text{Exp}(\theta, c, T) = \text{Exp}(\psi, d, T)$  implies  $r(c, T) = r(d, T)$ , which implies  $c^2 = d^2$ . Indeed, observe that in the three cases described in (20) we have, respectively,  $r(c, t) > t$ ,  $r(c, t) = t$ ,  $r(c, t) < t$ ; thus  $c, d \in (-1, 1)$  or  $c, d \in \{-1, 1\}$  or  $c, d \in (-\infty, -1) \cup (1, +\infty)$ . In each of the three cases, the identity  $r(c, T) = r(d, T)$  implies  $c^2 = d^2$ . Indeed we have the following three cases.

*Case  $c, d \in (-1, 1)$ .* In this case the conclusion follows from the fact that  $\frac{\text{Sinh}(p)}{p}$  is injective for  $p \in (0, +\infty)$ .

*Case  $c, d \in \{-1, 1\}$ .* This case is straightforward.

*Case  $c, d \in (-\infty, -1) \cup (1, +\infty)$ .* Let us prove first that  $\frac{\sqrt{c^2-1}T}{2} \in (0, \pi)$ . By contradiction, assume  $\frac{\sqrt{c^2-1}T}{2} \geq \pi$ . There exists  $t \in (0, T]$  such that  $\frac{\sqrt{c^2-1}t}{2} = 0$ ; hence  $r(c, t) = 0$ , from which it follows that  $\text{Exp}(\theta, c, t) \in e^{\mathbf{K}}$ .

Hence either  $t < T$  (and  $\text{Exp}(\theta, c, \cdot)$  is not optimal on  $[0, T]$ , a contradiction) or  $t = T$  (and  $g \in K_{\text{Id}}^{\text{loc}} \cup \text{Id}$ , a contradiction). Similarly we prove that  $\frac{\sqrt{d^2-1}T}{2} \in (0, \pi)$ .

Now observe that  $r(c, T) = r(d, T)$  implies

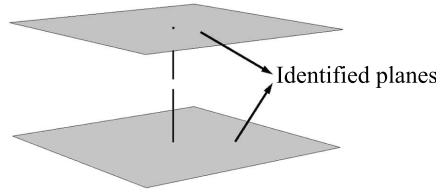
$$\frac{\sin(\frac{\sqrt{c^2-1}T}{2})}{\frac{\sqrt{c^2-1}T}{2}} = \frac{\sin(\frac{\sqrt{d^2-1}T}{2})}{\frac{\sqrt{d^2-1}T}{2}}.$$

Recalling that  $\frac{\sin p}{p}$  is injective for  $p \in (0, \pi)$ , we have  $\frac{\sqrt{c^2-1}T}{2} = \frac{\sqrt{d^2-1}T}{2}$ ; hence  $c^2 = d^2$ .

We have the following two cases:

(i)  $c = d \in \mathbb{R}$ . The identity  $g_{11}(\theta, c, T) = g_{11}(\psi, c, T)$  implies either  $\theta = \psi$  (i.e., the geodesics coincide) or  $c \in (-\infty, -1) \cup (1, +\infty)$ , and thus  $\sin(\frac{\sqrt{c^2-1}T}{2}) = 0$ , i.e.,  $\text{exp}(\theta, c, T) \in e^{\mathbf{K}}$ ; hence either  $g = \text{Id}$  or  $g \in K_{\text{Id}}^{\text{loc}}$ , a contradiction.

(ii)  $c = -d \in \mathbb{R} \setminus \{0\}$ . Writing  $g = e^{zk}e^{x_0p_1+y_0p_2}$ , we have  $\text{Exp}(\psi, -c, T) = e^{-zk}e^{xp_1+yp_2}$ . The identity  $\text{Exp}(\theta, c, T) = \text{Exp}(\psi, -c, T)$  and the uniqueness of the

FIG. 13. The cut locus for the  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifold  $SL(2)$ .

decomposition from section 3.3.2 imply  $e^{zk} = \pm \text{Id}$ . Thus  $g$  is symmetric, i.e.,  $g_{12}(\theta, c, T) = g_{21}(\theta, c, T)$ .

If  $c \in (-1, 1)$  this equation implies

$$\frac{\tan\left(c\frac{T}{2}\right)}{c\frac{T}{2}} = \frac{\text{Tanh}\left(\sqrt{1-c^2}\frac{T}{2}\right)}{\sqrt{1-c^2}\frac{T}{2}}.$$

Choosing  $c > 0$ , observe that the first positive solution  $T_1$  of the equation

$$\frac{\tan\left(c\frac{T_1}{2}\right)}{c\frac{T_1}{2}} = \frac{\text{Tanh}\left(\sqrt{1-c^2}\frac{T_1}{2}\right)}{\sqrt{1-c^2}\frac{T_1}{2}}$$

satisfies  $T_1 \in (\pi, 3\frac{\pi}{2})$ . The other cases,  $c \in \{-1, 1\}$  and  $c \in (-\infty, -1) \cup (1, +\infty)$ , are treated similarly and lead to  $T_1 \in (\pi, 3\frac{\pi}{2})$ .

Thus  $\cos\left(c\frac{T_1}{2}\right) < 0$ , and hence  $\text{Tr}(g) < 0$ . But  $\text{Exp}(\theta, c, T_1)$  symmetric and  $\text{Tr}(g) < 0$  implies  $\text{Exp}(\theta, c, T_1) \in K_{\text{Id}}^{\text{sym}}$ ; i.e.,  $T_1$  is a cut time. Thus either  $T = T_1$  (meaning that  $\text{Exp}(\theta, c, t) \in K_{\text{Id}}^{\text{sym}}$ ) or  $T > T_1$  and  $\text{Exp}(\theta, c, \cdot)$  is not optimal in  $[0, T]$ , a contradiction.  $\square$

We give a picture of the cut locus for the  $\mathbf{k} \oplus \mathbf{p}$  sub-Riemannian manifold  $SL(2)$  in Figure 13.

**Acknowledgments.** We are grateful to A. Agrachev for many illuminating discussions. We deeply thank G. Charlot for bringing to our attention some crucial properties of the cut locus, and we thank L. Paoluzzi for many explanations on lens spaces.

## REFERENCES

- [1] A. AGRACHEV, *Methods of control theory in nonholonomic geometry*, in Proceedings of the ICM-94, Birkhäuser, Zürich, 1996, pp. 1473–1483.
- [2] A. AGRACHEV, *Compactness for sub-Riemannian length-minimizers and subanalyticity*, Rend. Sem. Mat. Univ. Politec. Torino, 56 (2001), pp. 1–12.
- [3] A. A. AGRACHEV AND Y. L. SACHKOV, *Control Theory from the Geometric Viewpoint*, Encyclopaedia Math. Sci. 87, Springer-Verlag, Berlin, 2004.
- [4] A. AGRACHEV, *Exponential mappings for contact sub-Riemannian structures*, J. Dynam. Control Systems, 2 (1996), pp. 321–358.
- [5] R. BEALS, B. GAVEAU, AND P. C. GREINER, *Hamilton–Jacobi theory and the heat kernel on Heisenberg groups*, J. Math. Pures Appl., 79 (2000), pp. 633–689.
- [6] A. BELLAÏCHE, *The tangent space in sub-Riemannian geometry*, in Sub-Riemannian Geometry, A. Bellaïche and J.-J. Risler, eds., Progr. Math. 144, Birkhäuser, Basel, 1996, pp. 1–78.
- [7] A. BOGGESS, *CR Manifolds and the Tangential Cauchy–Riemann Complex*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1991.
- [8] B. BONNARD AND M. CHYBA, *Singular Trajectories and Their Role in Control Theory*, Math. Appl. 40, Springer-Verlag, Berlin, 2003.

- [9] U. BOSCAIN, T. CHAMBRION, AND G. CHARLOT, *Nonisotropic 3-level quantum systems: Complete solutions for minimum time and minimum energy*, Discrete Contin. Dyn. Syst. Ser. B, 5 (2005), pp. 957–990.
- [10] U. BOSCAIN, T. CHAMBRION, AND J.-P. GAUTHIER, *On the  $K + P$  problem for a three-level quantum system: Optimality implies resonance*, J. Dynam. Control Systems, 8 (2002), pp. 547–572.
- [11] U. BOSCAIN AND B. PICCOLI, *Optimal Synthesis for Control Systems on 2-D Manifolds*, Math. Appl. 43, Springer-Verlag, Berlin, 2004.
- [12] U. BOSCAIN AND S. POLIDORO, *Gaussian estimates for hypoelliptic operators via optimal control*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Mem. Mat. Appl., 18 (2007), pp. 333–342.
- [13] A. BRESSAN AND B. PICCOLI, *Introduction to the Mathematical Theory of Control*, AIMS Ser. Appl. Math., AIMS, Springfield, MO, 2007.
- [14] R. W. BROCKETT, *Explicitly solvable control problems with nonholonomic constraints*, in Proceedings of the 38th Annual IEEE Conference on Decision and Control, Vol. 1, IEEE, Piscataway, NJ, 1999, pp. 13–16.
- [15] Y. CHITOUR, F. JEAN, AND E. TRÉLAT, *Genericity results for singular curves*, J. Differential Geom., 73 (2006), pp. 45–73.
- [16] Y. ELIASHBERG, *Contact 3-manifolds twenty years since J. Martinet's work*, Ann. Inst. Fourier (Grenoble), 42 (1992), pp. 165–192.
- [17] B. GAVEAU, *Principe de moindre action, propagation de la chaleur et estimées sous elliptiques sur certains groupes nilpotents*, Acta Math., 139 (1977), pp. 95–153.
- [18] M. GROMOV, *Carnot–Caratheodory spaces seen from within*, in Sub-Riemannian Geometry, A. Bellaïche and J.-J. Risler, eds., Progr. Math. 144, Birkhäuser, Basel, 1996, pp. 79–323.
- [19] L. HÖRMANDER, *The analysis of linear partial differential operators*, Grundlehren Math. Wiss., Springer-Verlag, Berlin, 1983.
- [20] S. HELGASON, *Differential Geometry, Lie Groups, and Symmetric Spaces*, Pure Appl. Math., Academic Press, New York-London, 1978.
- [21] V. JURDJEVIC, *Geometric Control Theory*, Cambridge Stud. Adv. Math. 52, Cambridge University Press, Cambridge, UK, 1997.
- [22] V. JURDJEVIC, *Optimal Control, Geometry, and Mechanics*, in Mathematical Control Theory, J. Baillieul, and J. C. Willems, eds., Springer, New York, 1999, pp. 227–267.
- [23] V. JURDJEVIC, *Hamiltonian point of view on non-Euclidean geometry and elliptic functions*, Systems Control Lett., 43 (2001), pp. 25–41.
- [24] L. S. PONTRYAGIN, V. BOLTJANSKI, R. GAMKRELIDZE, AND E. MITCHTCHENKO, *The Mathematical Theory of Optimal Processes*, Wiley Interscience, New York, London, 1962.
- [25] D. ROLFSEN, *Knots and Links*, Publish or Perish, Houston, 1990.
- [26] J. R. WEEKS, *The Shape of Space*, Monogr. Textbooks Pure Appl. Math., Marcel Dekker, New York, 1985.

## ACCESSIBILITY AND CONTROLLABILITY IN THE PRESENCE OF FAST OSCILLATIONS\*

G. GRAMMEL†

**Abstract.** The accessibility properties of time-variant nonlinear control systems with fast time variables are investigated. It turns out that local accessibility can be carried from a time-invariant averaged system to the original system, provided the time variable is sufficiently fast, or equivalently, the singular perturbation parameter is sufficiently small. It is shown that the reachable sets of the averaged and the original systems contain common open sets. To this end, the singular perturbation parameter is considered as a homotopy parameter, and mapping degree methods are used. An application to controllability properties of time-variant nonlinear control systems with fast time variables is presented as well.

**Key words.** nonlinear control system, accessibility, fast oscillations, averaging

**AMS subject classifications.** 34C29, 93B03, 93B05

**DOI.** 10.1137/070706240

**1. Introduction.** We consider the singularly perturbed control system (SPCS)

$$(1) \quad \dot{z}(t) = f\left(z(t), \frac{t}{\epsilon}, u(t)\right),$$

where the parameter  $\epsilon > 0$  is small and hence reflects that the state  $z(t)$  moves slowly in comparison with the scaled time variable  $\frac{t}{\epsilon}$ . Obviously this system is time-variant and hence difficult to investigate with respect to any structural properties. As for exponential stability properties, it is common practice to construct an averaged system and to carry over information about the averaged system to the perturbed system, at least for small perturbation parameters. This method works perfectly for differential equations with slow and fast time variables whenever an averaged system can be constructed; see [11] and the references therein. It seems that much less attention has been paid to similar considerations with respect to controllability properties. In the present work we make use of local accessibility properties of the averaged control system (ACS) and show that these are transferred to the SPCS, at least for sufficiently small perturbation parameters  $\epsilon > 0$ . To this end it is shown that the reachable sets of the ACS and the SPCS have certain open sets in common for sufficiently small perturbation parameters. This is a well-known fact for regular perturbations; see [2], where a corresponding proof relies on an application of the implicit function theorem to a certain mapping obtained via geometric control theory. In the present work the method of proof is entirely different. Whereas we still make use of elementary geometric control theory, as presented in [10, 15], the transfer from the ACS to the SPCS relies on mapping degree arguments along with a particular approximation of a given trajectory to the ACS. For this purpose, the singular perturbation parameter is used to obtain a suitable homotopy. The result obtained on *robust local accessibility* is applied to show a transfer of controllability properties from the ACS to the SPCS.

---

\*Received by the editors October 24, 2007; accepted for publication (in revised form) February 28, 2008; published electronically June 25, 2008.

<http://www.siam.org/journals/sicon/47-4/70624.html>

†Center for Mathematics, Technical University of Munich, 3 Boltzmann Street, 85747 Garching, Germany (grammel@ma.tum.de).

The relation between averaging techniques and Lie brackets of vector fields has been the focus of a number of articles; see, e.g., [7, 8, 9] and the references therein. However, the main goal in these works is the construction of a meaningful limiting system, and the complications arise by an unbounded right-hand side of the differential equations involved. In the present paper we deal with a simpler situation. Since the fast oscillations are locally uniformly bounded, the construction of an ACS is straightforward and follows the pioneering work of Plotnikov [12], who, however, used a differential inclusion setting. The approximation results on nonlinear systems with two time scales, as presented in [1, 4, 13, 16], could be used just as well, if tailored in an appropriate way. It turns out that the Lie brackets for the ACS are averages of the corresponding Lie brackets for the SPCS.

The paper is organized as follows. The construction of the ACS is performed in section 2. The setting is fixed there as well. Section 3 contains the core of our presentation. Here it is shown that one can deduce local accessibility from the ACS to the SPCS. An application to a transfer of controllability properties is discussed in section 4.

## 2. Preliminaries. The setting is as follows.

*Assumption 2.1* (regularity). The state space of the singularly perturbed control system is the Euclidean space  $\mathbf{R}^n$ . The control range is a compact metric space  $\Omega$ . The controls are measurable functions  $u : \mathbf{R} \rightarrow \Omega$ , and the set of those controls is denoted by  $\mathcal{U}$ . For fixed  $\omega \in \Omega$  and  $s \in \mathbf{R}$ , the vector field  $\mathbf{R}^n \ni z \mapsto f(z, s, \omega) \in \mathbf{R}^n$  is of class  $C^\infty$ . The mapping  $\mathbf{R}^n \times \mathbf{R} \times \Omega \ni (z, s, \omega) \mapsto f(z, s, \omega) \in \mathbf{R}^n$  is continuous and uniformly Lipschitz continuous in the first argument; i.e., there is a constant  $L \geq 0$  such that one has

$$\|f(z_1, s, \omega) - f(z_2, s, \omega)\| \leq L\|z_1 - z_2\|$$

for all  $z_1, z_2 \in \mathbf{R}^n$ ,  $s \in \mathbf{R}$ , and  $\omega \in \Omega$ .

*Assumption 2.2* (periodicity). There is a time  $S > 0$ , the period, such that we have

$$f(z, s, \omega) = f(z, s + S, \omega)$$

for all  $(z, s, \omega) \in \mathbf{R}^n \times \mathbf{R} \times \Omega$ .

For  $\epsilon > 0$ ,  $z^0 \in \mathbf{R}^n$ ,  $t^0 \in \mathbf{R}$ , and  $u \in \mathcal{U}$ , let  $\mathbf{R} \ni t \mapsto z_\epsilon(t, z^0, t^0, u) \in \mathbf{R}^n$  be the unique trajectory to the SPCS (1) with initial condition  $z(t^0) = z^0$ . For  $t \geq t^0$ , the reachable set of the SPCS (1) is defined by

$$\mathcal{R}_\epsilon(t, z^0, t^0) := \bigcup_{u \in \mathcal{U}} \{z_\epsilon(t, z^0, t^0, u)\}.$$

Note that the reachable sets  $\mathcal{R}_\epsilon(t, z^0, t^0) \subset \mathbf{R}^n$  are bounded.

In what follows, we make use of the periodicity in order to construct an ACS reflecting the situation when the singular perturbation parameter tends to zero. An  $S$ -periodic control  $u \in \mathcal{U}$  generates an averaged vector field  $\omega_0^u$  on  $\mathbf{R}^n$  by

$$\omega_0^u(z) := \frac{1}{S} \int_0^S f(z, s, u(s)) ds.$$

Let  $\Omega_0 \subset C^\infty(\mathbf{R}^n; \mathbf{R}^n)$  be the collection of all averaged vector fields. In order to produce a control system with the averaged vector fields, we define a mapping  $f_0 :$

$\mathbf{R}^n \times \Omega_0 \rightarrow \mathbf{R}^n$  by

$$f_0(z, \omega_0) := \omega_0(z).$$

Then, the ACS on  $\mathbf{R}^n$  is given by

$$(2) \quad \dot{z}(t) = f_0(z(t), u_0(t)),$$

where  $u_0 : \mathbf{R} \rightarrow \Omega_0$  is measurable. Let  $\mathcal{U}_0$  be the set of measurable controls  $u_0 : \mathbf{R} \rightarrow \Omega_0$ .

For  $z^0 \in \mathbf{R}^n$ ,  $t^0 \in \mathbf{R}$ ,  $t \in \mathbf{R}$ , and  $u_0 \in \mathcal{U}_0$ , let  $\mathbf{R} \ni t \mapsto z_0(t, z^0, t^0, u_0) \in \mathbf{R}^n$  be the unique trajectory to the ACS (2) with initial condition  $z(t^0) = z^0$ . For  $t \geq t^0$ ,  $z^0 \in \mathbf{R}^n$ , the reachable set of the ACS (2) is defined by

$$\mathcal{R}_0(t, z^0, t^0) := \bigcup_{u_0 \in \mathcal{U}_0} \{z_0(t, z^0, t^0, u_0)\}.$$

The relation between the SPCS (1) and the ACS (2), in terms of reachable sets, is as follows.

**LEMMA 2.3.** *Let Assumptions 2.1 and 2.2 be satisfied. For any initial state  $z^0 \in \mathbf{R}^n$  and any time horizon  $T > 0$ , there is a constant  $K = K(z^0, T) \geq 0$  such that for all  $t^0 \in \mathbf{R}$ ,  $t \in [t^0, t^0 + T]$ , and all  $\epsilon > 0$ , the estimation*

$$d_H(\mathcal{R}_0(t, z^0, t^0), \mathcal{R}_\epsilon(t, z^0, t^0)) \leq K \epsilon$$

*holds true, where  $d_H(\cdot, \cdot)$  denotes the Hausdorff semimetric for bounded sets in  $\mathbf{R}^n$ .*

*Proof.* The proof follows from [5, Lemma 2.2 and its proof]. The result also can be deduced from [6, Theorem 2.3 and its proof].  $\square$

Note that in case of ordinary differential equations, i.e.,  $\Omega = \{\omega\}$  consists of one point only, the ACS (2) is reduced to

$$\dot{z}(t) = f_0(z(t)),$$

where the averaged vector field  $f_0 : \mathbf{R}^n \rightarrow \mathbf{R}^n$  is given by

$$f_0(z) := \frac{1}{S} \int_0^S f(z, s, \omega) ds.$$

In this case a proof of Lemma 2.3 can be found in [14].

**3. Accessibility via the averaged system.** In what follows we make use of local accessibility properties of the ACS.

**Assumption 3.1** (accessibility). The ACS (2) is locally accessible on  $\mathbf{R}^n$ ; i.e., for any  $z^0 \in \mathbf{R}^n$  and any  $T > 0$  the reachable set  $\mathcal{R}_0([0, T], z^0, 0)$  has interior points in  $\mathbf{R}^n$ .

Note that for any initial value  $z^0 \in \mathbf{R}^n$  and for any open neighborhood  $\mathcal{V} \subset \mathbf{R}^n$  about  $z^0 \in \mathbf{R}^n$ , there is a time  $T_{\mathcal{V}} > 0$  with

$$\mathcal{R}_0([0, T_{\mathcal{V}}], z^0, 0) \subset \mathcal{V}.$$

This property easily can be deduced from Assumptions 2.1 and 2.2 using Gronwall's lemma. Hence, the interior points of the reachable set  $\mathcal{R}_0([0, T], z^0, 0)$  are produced

locally about  $z^0 \in \mathbf{R}^n$ , and standard results on the relationship between local accessibility and Lie algebras generated by vector fields, such as Theorem 11 in [15], can be used.

The SPCS (1) is uniformly locally accessible on  $\mathbf{R}^n$  in the following sense.

**THEOREM 3.2.** *Let Assumptions 2.1, 2.2, and 3.1 be satisfied. For any initial value  $z^0 \in \mathbf{R}^n$ ,  $T > 0$ , there are  $\epsilon_0 > 0$ ,  $z^1 \in \mathbf{R}^n$ ,  $r > 0$  such that for any  $\epsilon \in [0, \epsilon_0]$  and any initial time  $t^0 \in \mathbf{R}$ , the inclusion*

$$B(r; z^1) \subset \mathcal{R}_\epsilon([t^0, t^0 + T], z^0, t^0)$$

is valid.

*Proof.* The proof is based on the homotopy invariance of the mapping degree. To this end we consider the perturbation parameter as a homotopy parameter.

*Step 1.* By the local accessibility of the ACS there is an open and dense set  $X^0 \subset \mathbf{R}^n$  such that for all  $z^0 \in X^0$  the Lie algebra generated by the vector fields  $\omega_0 \in \Omega_0$  has full rank  $n = \dim(\mathbf{R}^n)$ ; see, e.g., [15, Theorem 11, p. 177]. Let  $z^0 \in X^0$  (the case  $z^0 \notin X^0$  easily can be reduced to the case  $z^0 \in X^0$ ). Then (see, e.g., [15, Lemma 4.2.8, p. 152]), there are averaged vector fields  $\omega_0^1, \dots, \omega_0^n \in \Omega_0$  and times  $(t_1^0, \dots, t_n^0) \in (0, \frac{T}{n})^n$  such that the derivative of the mapping

$$\left(0, \frac{T}{n}\right)^n \ni (t_1, \dots, t_n) \mapsto (e^{\omega_0^n t_n} \circ \dots \circ e^{\omega_0^1 t_1})(z^0) \in \mathbf{R}^n$$

has full rank  $n = \dim(\mathbf{R}^n)$  at  $(t_1^0, \dots, t_n^0) \in (0, \frac{T}{n})^n$ . Here, the mapping  $\mathbf{R} \times \mathbf{R}^n \ni (t, z^0) \mapsto e^{t\omega} z^0 \in \mathbf{R}^n$  denotes the flow generated by the vector field  $\omega$  on  $\mathbf{R}^n$ . Let  $\mathcal{O} \subset (0, \frac{T}{n})^n$  be an open ball about  $(t_1^0, \dots, t_n^0) \in (0, \frac{T}{n})^n$  such that the mapping

$$H_0: \overline{\mathcal{O}} \rightarrow \mathbf{R}^n, \quad (t_1, \dots, t_n) \mapsto (e^{\omega_0^n t_n} \circ \dots \circ e^{\omega_0^1 t_1})(z^0)$$

is regular and injective. We set

$$z^1 := H_0(t_1^0, \dots, t_n^0).$$

Then we have

$$\deg(H_0, \mathcal{O}, z^1) \neq 0,$$

where  $\deg$  denotes the mapping degree of Brouwer. For an introduction to elementary properties of Brouwer's mapping degree, consult, for instance, [3]. It is convenient to describe the mapping  $H_0$  via a certain control function for the ACS (2). For  $(t_1, \dots, t_n) \in \overline{\mathcal{O}}$  we define

$$u_0^{(t_1, \dots, t_n)}(t) := \begin{cases} \omega_0^1 & \text{for } t \in (-\infty, t_1), \\ \omega_0^i & \text{for } t \in [t_1 + \dots + t_{i-1}, t_1 + \dots + t_i), \quad 2 \leq i \leq n-1, \\ \omega_0^n & \text{for } t \in [t_1 + \dots + t_{n-1}, \infty). \end{cases}$$

Then we have

$$H_0(t_1, \dots, t_n) = z_0(t_1 + \dots + t_n, z^0, 0, u_0^{(t_1, \dots, t_n)}).$$

*Step 2.* For  $\epsilon > 0$ ,  $t^0 \in \mathbf{R}$ , we define a control  $u_\epsilon^{(t_1, \dots, t_n)} \in \mathcal{U}$  for the SPCS (1) in the following way. We set

$$u_\epsilon^{(t_1, \dots, t_n)}(t) := \begin{cases} u_1(\frac{t}{\epsilon}) & \text{for } t \in (-\infty, t^0 + t_1), \\ u_i(\frac{t}{\epsilon}) & \text{for } t \in [t^0 + t_1 + \dots + t_{i-1}, t^0 + t_1 + \dots + t_i), \quad 2 \leq i \leq n-1, \\ u_n(\frac{t}{\epsilon}) & \text{for } t \in [t^0 + t_1 + \dots + t_{n-1}, \infty), \end{cases}$$

where, for  $i = 1, \dots, n$ ,  $u_i \in \mathcal{U}$  is an  $S$ -periodic control function generating  $\omega_0^i$ . We define a mapping

$$H_\epsilon : \overline{\mathcal{O}} \rightarrow \mathbf{R}^n, \quad (t_1, \dots, t_n) \mapsto z_\epsilon(t^0 + t_1 + \dots + t_n, z^0, t^0, u_\epsilon^{(t_1, \dots, t_n)}).$$

The continuity of  $H_\epsilon$  is obvious.

*Step 3.* Next we show that the mapping

$$[0, \infty) \times \overline{\mathcal{O}} \ni (\epsilon, t_1, \dots, t_n) \mapsto H_\epsilon(t_1, \dots, t_n) \in \mathbf{R}^n$$

is continuous as well. This is obviously the case at  $\epsilon > 0$  but has to be shown at  $\epsilon = 0$ . To this end we prove that there is a constant  $C = C(z^0, T) \geq 0$  with

$$(3) \quad \max_{t \in [0, T]} \|z_\epsilon(t^0 + t, z^0, t^0, u_\epsilon^{(t_1, \dots, t_n)}) - z_0(t, z^0, 0, u_0^{(t_1, \dots, t_n)})\| \leq C \epsilon$$

for all  $t^0 \in \mathbf{R}$ ,  $(t_1, \dots, t_n) \in \overline{\mathcal{O}}$  and sufficiently small  $\epsilon > 0$ . We set  $s_k := k\epsilon S$  for  $k = 0, 1, \dots, \lceil \frac{T}{\epsilon S} \rceil$  and use the abbreviations  $u_0 = u_0^{(t_1, \dots, t_n)}$ ,  $z_0(\cdot) := z_0(\cdot, z^0, 0, u_0)$ . Then we have

$$z_0(s_{k+1}) = z_0(s_k) + \int_{s_k}^{s_{k+1}} f_0(z_0(s), u_0(s)) ds.$$

We can also write

$$z_0(s_{k+1}) = z_0(s_k) + \int_{s_k}^{s_{k+1}} f_0(z_0(s_k), \omega_0^{i(k)}) ds + \epsilon S L E_{k,1} + E_{k,2},$$

where the errors  $E_{k,1} \in \mathbf{R}^n$  are caused by the fact that possibly  $z_0(s_k) \neq z_0(s)$  for some  $s \in (s_k, s_{k+1})$ . Note that we can estimate

$$\|E_{k,1}\| \leq \max_{s_k \leq s \leq s_{k+1}} \|z_0(s) - z_0(s_k)\| \leq \epsilon S P,$$

where  $P = P(T, z^0) \geq 0$  is an upper bound for the norm of the vector fields  $f$  (and hence for  $f_0$  as well) along the trajectories starting in  $z^0 \in \mathbf{R}^n$ . The errors  $E_{k,2} \in \mathbf{R}^n$  are of a different kind and appear because the discretization times  $s_k$  might not be the switching times of the control function  $u_0$ . Hence, the errors  $E_{k,2}$  appear for at most  $n = \dim(\mathbf{R}^n)$  different indices  $k$ . Alternatively, we can write

$$z_0(s_{k+1}) = z_0(s_k) + \int_{s_k}^{s_{k+1}} f\left(z_0(s_k), \frac{t^0 + s}{\epsilon}, u_\epsilon(t^0 + s)\right) ds + \epsilon S L E_{k,1} + E_{k,2},$$

where  $u_\epsilon = u_\epsilon^{(t_1, \dots, t_n)}$ . As for the corresponding trajectory of the SPCS (1), we use the abbreviations  $z_\epsilon(\cdot) = z_\epsilon(\cdot, z^0, t^0, u_\epsilon)$  and obtain

$$z_\epsilon(t^0 + s_{k+1}) = z_\epsilon(t^0 + s_k) + \int_{s_k}^{s_{k+1}} f\left(z_\epsilon(t^0 + s), \frac{t^0 + s}{\epsilon}, u_\epsilon(t^0 + s)\right) ds.$$

Replacing  $z_\epsilon(t^0 + s)$  by  $z_\epsilon(t^0 + s_k)$ , we possibly produce errors which are expressed by  $E_{k,3} \in \mathbf{R}^n$  in the following equality:

$$z_\epsilon(t^0 + s_{k+1}) = z_\epsilon(t^0 + s_k) + \int_{s_k}^{s_{k+1}} f\left(z_\epsilon(t^0 + s_k), \frac{t^0 + s}{\epsilon}, u_\epsilon(t^0 + s)\right) ds + \epsilon S L E_{k,3}.$$



Note that we can estimate

$$\|E_{k,3}\| \leq \max_{s_k \leq s \leq s_{k+1}} \|z_\epsilon(t^0 + s) - z_\epsilon(t^0 + s_k)\| \leq \epsilon SP.$$

Defining  $\Delta_k := \|z_\epsilon(t^0 + s_k) - z_0(s_k)\|$ , we obtain  $\Delta_0 = 0$  along with the recursive relation

$$\begin{aligned} \Delta_{k+1} &\leq \Delta_k + \int_{s_k}^{s_{k+1}} L\Delta_k ds + \epsilon SLE_{k,1} + E_{k,2} + \epsilon SLE_{k,3} \\ &\leq \Delta_k(1 + \epsilon SL) + \epsilon SLE_{k,1} + E_{k,2} + \epsilon SLE_{k,3}, \end{aligned}$$

where the number of indices  $k \in \{0, 1, \dots, [\frac{T}{\epsilon S}]\}$  with  $\|E_{k,2}\| > 0$  is not exceeding  $n = \dim(\mathbf{R}^n)$ . For those indices, we have the error estimations

$$\|E_{k,2}\| \leq 2\epsilon SP.$$

In the recursive relation above, the error  $\Delta_k$  is amplified by the factor  $1 + \epsilon SL \geq 1$  to obtain an estimation for  $\Delta_{k+1}$ . Hence, an early appearance of the errors  $E_{k,2} \geq 0$  has a large effect. In the worst case, we have

$$\|E_{k,2}\| = \begin{cases} 2\epsilon SP & \text{for } k = 0, 1, \dots, n-1, \\ 0 & \text{for } k = n, n+1, \dots, [\frac{n}{\epsilon S}]. \end{cases}$$

This yields the estimation

$$\Delta_k \leq (\epsilon SL\epsilon SP + 2\epsilon SP + \epsilon SL\epsilon SP) \sum_{l=0}^{k-1} (1 + \epsilon SL)^l \leq 2(\epsilon^2 S^2 LP + \epsilon SP)ne^{LT}$$

for  $k = 0, 1, \dots, n-1$ . For  $k = n-1, n, \dots, [\frac{T}{\epsilon S}]$ , we obtain the recursive relation

$$\Delta_{k+1} \leq \Delta_k(1 + \epsilon SL) + \epsilon SL\|E_{k,1}\| + \epsilon SL\|E_{k,3}\|,$$

which yields

$$\begin{aligned} \Delta_k &\leq \Delta_{n-1}(1 + \epsilon S)^{k-n+1} + 2\epsilon^2 S^2 PL \sum_{l=0}^{k-n} (1 + \epsilon SL)^l \\ &\leq 2(\epsilon^2 S^2 LP + \epsilon SP)ne^{2LT} + 2\epsilon SPL e^{LT}, \end{aligned}$$

from which (3) immediately follows. Now we take  $(t_1, \dots, t_n), (\hat{t}_1, \dots, \hat{t}_n) \in \overline{\mathcal{O}}$  and an  $\epsilon > 0$  sufficiently small. Then we can estimate

$$\begin{aligned} &\|H_\epsilon(t_1, \dots, t_n) - H_0(\hat{t}_1, \dots, \hat{t}_n)\| \\ &\leq \|H_\epsilon(t_1, \dots, t_n) - H_0(t_1, \dots, t_n)\| + \|H_0(t_1, \dots, t_n) - H_0(\hat{t}_1, \dots, \hat{t}_n)\| \\ &\leq C\epsilon + \|H_0(t_1, \dots, t_n) - H_0(\hat{t}_1, \dots, \hat{t}_n)\|, \end{aligned}$$

and the claim follows from the continuity of  $H_0$ .

*Step 4.* There is a ball  $B(r; z^1)$  about  $z^1 \in \mathbf{R}^n$  and an  $\epsilon_0 > 0$  such that

$$H_\epsilon(\partial\mathcal{O}) \cap B(r; z^1) = \emptyset$$

for all  $\epsilon \in [0, \epsilon_0]$ . By the homotopy invariance property of the mapping degree, we conclude that

$$\deg(H_\epsilon, \mathcal{O}, z) \neq 0$$

for all  $\epsilon \in [0, \epsilon_0]$  and all  $z \in B(r; z^1)$ , and the proof is finished.  $\square$

The Lie algebraic condition on local accessibility of the ACS is specified in the following remark.

*Remark 3.3.* Let  $\omega_0^1, \omega_0^2 \in \Omega_0$  be two averaged vector fields on  $\mathbf{R}^n$ ; i.e., there are  $u^1, u^2 \in \mathcal{U}$  such that we can write

$$\omega_0^i(z) = \frac{1}{S} \int_0^S f(z, s, u^i(s)) ds, \quad i = 1, 2,$$

for all  $z \in \mathbf{R}^n$ . A straightforward calculation shows that we can represent the Lie brackets of the averaged vector fields in the following way:

$$[\omega_0^1, \omega_0^2](z) = \frac{1}{S^2} \int_0^S \int_0^S [f(\cdot, s_1, u^1(s_1)), f(\cdot, s_2, u^2(s_2))](z) ds_1 ds_2$$

for all  $z \in \mathbf{R}^n$ . Similarly, higher order Lie brackets of the averaged vector fields are obtained by averaging higher order Lie brackets of the original vector fields.

**4. Controllability via the averaged system.** Theorem 3.2 has immediate applications to controllability properties of the singularly perturbed systems. Since complete controllability is an uncommon property of nonlinear control systems, the notion of a *control set* is presented in [2]. Roughly speaking a control set is a subset of the state space in which complete controllability holds. We give a similar notion of controllability for the time-variant singularly perturbed systems.

**DEFINITION 4.1.** Let  $M \subset \mathbf{R}^n$ . We say that the ACS (2) is controllable on  $M$  if

$$M \subset \mathcal{R}_0([0, \infty), z^0, 0)$$

for any initial state  $z^0 \in M$ . For a fixed  $\epsilon > 0$ , we say that the SPCS (1) is controllable on  $M$  if

$$M \subset \mathcal{R}_\epsilon([t^0, \infty), z^0, t^0)$$

for any pair of initial conditions  $(z^0, t^0) \in M \times \mathbf{R}$ .

**THEOREM 4.2.** Let Assumptions 2.1, 2.2, and 3.1 be satisfied. Let  $M \subset \mathbf{R}^n$  be compact with  $\text{int}M \neq \emptyset$ . If the ACS (2) is controllable on  $M$ , then there is an  $\epsilon_M > 0$  such that the SPCS (1) is controllable on  $M$  for all  $\epsilon \in (0, \epsilon_M]$ . In particular, there is a time  $T_M > 0$  such that

$$M \subset \mathcal{R}_\epsilon([t^0, t^0 + T_M], z^0, t^0)$$

for any pair of initial conditions  $(z^0, t^0) \in M \times \mathbf{R}$  and  $\epsilon \in (0, \epsilon_M]$ .

*Proof.* Let  $z^0 \in \text{int}M$ . There are  $\epsilon_0 > 0$ ,  $z^1 \in M$ ,  $r > 0$ ,  $T > 0$  such that for any  $\epsilon \in [0, \epsilon_0]$ , any initial time  $t^0 \in \mathbf{R}$ , and any initial state  $x^0 \in B(r; z^1)$ , we have the inclusion

$$z^0 \in \mathcal{R}_\epsilon([t^0, t^0 + T], x^0, t^0) \subset M.$$

This follows immediately from Theorem 3.2 via time reversal. Since the ACS (2) is controllable on  $M$ , for any  $z \in M$  there is an averaged control  $u_0 \in \mathcal{U}_0$  and a time  $T_z > 0$  with

$$z_0(t^0 + T_z, z, t^0, u_0) = z^1.$$

Additionally, by Lemma 2.3 there are  $\epsilon_z > 0$  and  $r_z > 0$  such that for any  $\epsilon \in [0, \epsilon_z]$ , any  $t^0 \in \mathbf{R}$ , and any  $x \in B(r_z; z)$ , we have

$$\mathcal{R}_\epsilon([t^0, t^0 + T_z], x, t^0) \cap B(r; z^1) \neq \emptyset.$$

By compactness of  $M$ , there are a time  $T_M^+ > 0$  and an  $\epsilon_M^+ > 0$  such that

$$z^0 \in \mathcal{R}_\epsilon([t^0, t^0 + T_M^+], z, t^0)$$

for all  $\epsilon \in (0, \epsilon_M^+]$ , all  $t^0 \in \mathbf{R}$ , and all  $z \in M$ . Note that the ACS (2) is controllable on  $M$  if we replace  $f$  by  $-f$ , i.e., after a time reversal. Hence, similar arguments show that there are a time  $T_M^- > 0$  and an  $\epsilon_M^- > 0$  such that

$$z \in \mathcal{R}_\epsilon([t^0, t^0 + T_M^-], z^0, t^0)$$

for all  $\epsilon \in (0, \epsilon_M^-]$  and all  $z \in M$ . Overall, we obtain the required result with

$$\epsilon_M := \min(\epsilon_M^+, \epsilon_M^-), \quad T_M := T_M^+ + T_M^-,$$

and the proof is finished.  $\square$

Clearly, the compactness of  $M$  is necessary in order to obtain a uniform maximal controllability time  $T_M \geq 0$  and a uniform bound  $\epsilon_M > 0$  for the singularly perturbed systems. Compactness does not mean an enormous restriction, since for noncompact  $M$  one still can consider compact subsets. A more serious restriction to the applicability of Theorem 4.2 is given by the presence of interior points in  $M$ . However, the following counterexample shows that it is indispensable for the transfer of controllability.

*Example 4.3.* Consider the SPCS in  $\mathbf{R}^2$  given by

$$\dot{z}_1(t) = z_1(t) \sin\left(\frac{t}{\epsilon}\right) - z_2(t) + z_1(t)u(t),$$

$$\dot{z}_2(t) = z_1(t) + z_2(t) \sin\left(\frac{t}{\epsilon}\right) + z_2(t)u(t),$$

where  $u(t) \in \Omega := \{0, 1\}$  for a.a.  $t \in \mathbf{R}$ . The ACS can be written as

$$\dot{z}_1(t) = -z_2(t) + z_1(t)w(t),$$

$$\dot{z}_2(t) = z_1(t) + z_2(t)w(t),$$

where  $w(t) \in [0, 1]$  for a.a.  $t \in \mathbf{R}$ . Clearly, the ACS is controllable on

$$M := S^1 = \{(z_1, z_2) \in \mathbf{R}^2 : z_1^2 + z_2^2 = 1\}.$$

Just take the constant control  $w \equiv 0$  in order to keep the radius  $r(t) = \sqrt{z_1(t)^2 + z_2(t)^2}$  constant. Using polar coordinates via the transformation

$$z_1(t) = r(t) \cos(\phi(t)), \quad z_2(t) = r(t) \sin(\phi(t)),$$

the SPCS becomes

$$\dot{r}(t) = r(t) \left( \sin \left( \frac{t}{\epsilon} \right) + u(t) \right), \quad \dot{\phi}(t) = 1.$$

The variation of constants yields the solution

$$r(t) = e^{\epsilon(1 - \cos(\frac{t}{\epsilon}))} e^{\int_0^t u(s) ds}$$

to the initial value  $r(0) = 1$ . Hence, for  $t > 0$ , we have  $r(t) = 1$  if and only if  $t \in \epsilon 2\pi\mathbb{N}$  and  $u(s) = 0$  for a.a.  $s \in [0, t]$ . Accordingly, for rational  $\epsilon > 0$ , the set of points in  $M = S^1$  that can be reached from an initial value  $(z_1^0, z_2^0) \in M$  is only finite. Clearly, the SPCS is not controllable on  $M = S^1$ . Notice that the ACS does not exactly meet Assumption 3.1, since it is not accessible at the origin. However, we can add additional controls locally about the origin in order to achieve local accessibility of the ACS without affecting the described behavior near the unit sphere  $S^1$ .

#### REFERENCES

- [1] Z. ARTSTEIN AND A. VIGODNER, *Singularly perturbed ordinary differential equations with dynamic limits*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 541–569.
- [2] F. COLONIUS AND W. KLIEMANN, *The Dynamics of Control*, Systems Control Found. Appl., Birkhäuser Boston, Boston, 2000.
- [3] K. DEIMLING, *Nonlinear Functional Analysis*, Springer, Berlin, 1980.
- [4] V. GAITSGORY AND A. LEIZAROWITZ, *Limit occupational measures set for a control system and averaging of singularly perturbed control systems*, J. Math. Anal. Appl., 233 (1999), pp. 461–475.
- [5] G. GRAMMEL, *On the van-der-Pol oscillator with noisy nonlinearity*, Nonlinearity, 13 (2000), pp. 1343–1355.
- [6] G. GRAMMEL, *Exponential stability via the averaged system*, J. Dynam. Control Systems, 7 (2001), pp. 327–338.
- [7] J. KURZWEIL AND J. JARNIK, *Iterated Lie brackets in limit processes in ordinary differential equations*, Results Math., 14 (1988), pp. 125–137.
- [8] W. LIU, *Averaging theorems for highly oscillatory differential equations and iterated Lie brackets*, SIAM J. Control Optim., 35 (1997), pp. 1989–2020.
- [9] S. MARTINEZ, J. CORTEZ, AND F. BULLO, *Analysis and design of oscillatory control systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 1164–1177.
- [10] H. NIJMEIJER AND A. J. VAN DER SCHAFT, *Nonlinear Dynamical Control Systems*, Springer, Berlin, 1990.
- [11] J. PEUTEMAN AND D. AEYELS, *Exponential stability of nonlinear time-varying differential equations and partial averaging*, Math. Control Signals Systems, 15 (2002), pp. 42–70.
- [12] V. A. PLOTNIKOV, *Averaging method for differential inclusions and its application to optimal control problems*, Differ. Equ., 15 (1980), pp. 1013–1018.
- [13] M. QUINCAMPOIX AND F. WATBLED, *Averaging method for discontinuous Mayer’s problem of singularly perturbed control systems*, Nonlinear Anal. Theory Methods Appl., 54 (2003), pp. 819–837.
- [14] J. A. SANDERS AND F. VERHULST, *Averaging Methods in Nonlinear Dynamical Systems*, Springer, New York, 1985.
- [15] E. SONTAG, *Mathematical Control Theory*, 2nd ed., Springer, New York, 1998.
- [16] A. VIGODNER, *Limits of singularly perturbed control problems with statistical dynamics of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.

## ERGODIC CONTROL OF CONTINUOUS-TIME MARKOV CHAINS WITH PATHWISE CONSTRAINTS\*

TOMÁS PRIETO-RUMEAU<sup>†</sup> AND ONÉSIMO HERNÁNDEZ-LERMA<sup>‡</sup>

**Abstract.** This paper deals with unichain ergodic continuous-time controlled Markov chains (CMCs) with a denumerable state space and possibly unbounded reward rates as well as unbounded transition rates. The problem we are concerned with is to find control policies that maximize a sample-path average reward over the family of admissible policies for which a certain pathwise average cost is below a given value with probability one. To study this problem, first, we give conditions for the existence of sample-path average optimal policies. Then we analyze constrained average reward CMCs with “expected” constraints, and, finally, we apply these results to our original control problem with pathwise constraints. Examples on the control of a queueing system and an epidemic process illustrate the feasibility of our approach.

**Key words.** continuous-time controlled Markov chains, Markov decision processes, dynamic programming, ergodic control, constrained control problems

**AMS subject classifications.** 93E20, 90C40, 60J27

**DOI.** 10.1137/060668857

**1. Introduction.** This paper studies unichain ergodic continuous-time denumerable-state controlled Markov chains (CMCs), also known as Markov decision processes (MDPs). The corresponding transition and reward rates can both be unbounded. The problem we are concerned with is to maximize a long-run sample-path (or pathwise) average reward for the given CMC, subject to constraints on a given long-run pathwise average cost. (For expositional convenience we consider a single constraint, but it should be evident how to extend our results to any finite number of constraints.)

To analyze our problem we proceed in three steps. In the first one, we study *unconstrained* pathwise average reward CMCs. In the second step we give conditions for the existence of optimal policies for constrained CMCs with *expected constraints*; see (4.1). In the third and final step we extend the results in the former steps to our problem with *pathwise constraints* (also known in the literature as *hard constraints*); see (3.1). Our approach is illustrated with examples on the control of a queueing system and the control of an epidemic process.

As can be seen in recent papers, e.g., [3, 16, 19, 30] and their references, stochastic control problems with constraints form a very active area of research because they naturally arise in many important applications. However, almost all of the literature is concentrated on problems with *expected* constraints. In contrast, for problems with *pathwise* constraints, there is, to the best of our knowledge, just a handful of papers. For instance, Haviv [14] and Ross and Varadarajan [26, 27] study finite-state, finite-action, discrete-time Markov decision processes, mainly in the multichain case. The paper by Haviv is particularly interesting because it clearly shows, by means

---

\*Received by the editors September 5, 2006; accepted for publication (in revised form) March 3, 2008; published electronically June 25, 2008. This research was partially supported by CONACyT grant 45693-F.

<http://www.siam.org/journals/sicon/47-4/66885.html>

<sup>†</sup>Corresponding author. Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Educación a Distancia, Calle Senda del Rey 9, 28040 Madrid, Spain (tprieto@ccia.uned.es).

<sup>‡</sup>Departamento de Matemáticas, CINVESTAV-IPN, Apartado Postal 14-740, México D.F. 07000, Mexico (ohernand@math.cinvestav.mx).

of examples, that pathwise constraints are in general more “natural” than expected constraints—see also Ross and Varadarajan [26, Example 1]. We should also mention Wu, Arapostathis, and Shakkottai [28] who begin with a constrained queue in heavy traffic and end up with the ergodic control of a diffusion process with pathwise constraints—see [28, equations (6a), (6b)].

As can be seen in the related literature, there are several standard techniques for analyzing constrained control problems, such as convex analysis, Lagrange multipliers, linear programming, and dynamic programming. Our approach in this paper uses, mainly, a combination of the last three techniques.

The remainder of this paper is organized as follows. In section 2 we introduce the control model we will be dealing with. In section 3 we state our main results on CMCs with pathwise constraints (Theorem 3.3). The corresponding proofs are postponed to section 4. These proofs require several preliminary results which are of valuable interest by themselves. More precisely, we study the existence and characterization of (unconstrained) pathwise average optimal policies; see Theorem 4.3. Also, we consider problems with *expected* constraints which, as far as we know, have not been systematically studied for average reward CMCs; see Theorems 4.9 and 4.10. In section 5 we introduce the examples mentioned above, and we conclude in section 6 with some general remarks. Finally, in section 7 we give the proofs of two interesting results: one concerns the so-called Lyapunov conditions, and the other gives an easily verifiable condition for uniform exponential ergodicity of CMCs.

**2. Model definition and main assumptions.** We will deal with the control model

$$\{S, (A(i), i \in S), q_{ij}(a), r(i, a), u(i, a)\},$$

where

- $S$  is the state space, which we assume to be a denumerable set. Without loss of generality we will assume that  $S$  is the set of nonnegative integers, i.e.,  $S := \{0, 1, 2, \dots\}$ . (See the second example in section 5 for a bidimensional state space that can be suitably enumerated in the form  $0, 1, 2, \dots$ )
- for each  $i \in S$ ,  $A(i)$  is the set of admissible control actions in the state  $i$ , and it is assumed to be a Borel space, that is, a Borel subset of a complete and separable metric space. The action space  $A(i)$  is endowed with its Borel  $\sigma$ -algebra, denoted by  $\mathbb{B}(A(i))$ . We also define the set  $K := \{(i, a) : i \in S, a \in A(i)\}$ .
- for every  $i, j \in S$  and  $a \in A(i)$ ,  $q_{ij}(a)$  denotes the transition rate from  $i$  to  $j$  under the action  $a \in A(i)$ . The function  $a \mapsto q_{ij}(a)$  is measurable on  $\mathbb{B}(A(i))$  (measurability on the real numbers set  $\mathbb{R}$  is always with respect to the usual Borel  $\sigma$ -algebra) for each  $i, j \in S$ . By definition, the transition rates verify  $q_{ij}(a) \geq 0$  for every  $(i, a) \in K$  and  $j \in S$  such that  $i \neq j$ . We will also assume that they are *conservative*, i.e.,

$$(2.1) \quad \sum_{j \in S} q_{ij}(a) = 0 \quad \forall (i, a) \in K,$$

and *stable*, i.e.,

$$(2.2) \quad q(i) := \sup_{a \in A(i)} \{-q_{ii}(a)\} < \infty \quad \forall i \in S.$$

- the reward rate function  $r : K \rightarrow \mathbb{R}$  and the cost rate function  $u : K \rightarrow \mathbb{R}$  are measurable on  $A(i)$  for each fixed  $i \in S$ . We interpret  $r$  as a reward to be (somehow) maximized with the restriction that the cost  $u$  does not exceed (in a suitably defined sense) a given value.

This control model is the same as in, e.g., [9, 10, 12, 25], except that these references deal with *unconstrained* MDPs and so the control model does not include the cost rate  $u(i, a)$ .

**Control policies.** Let  $\Phi_m$  be the family of functions  $\varphi \equiv \{\varphi_t(B|i)\}$ , where  $t \geq 0$ ,  $i \in S$ , and  $B \in \mathbb{B}(A(i))$ , such that

- for each  $t \geq 0$  and  $i \in S$ ,  $B \mapsto \varphi_t(B|i)$  is a probability measure on  $(A(i), \mathbb{B}(A(i)))$ ;
- for each  $i \in S$  and  $B \in \mathbb{B}(A(i))$ , the function  $t \mapsto \varphi_t(B|i)$  is measurable on  $[0, \infty)$ .

We say that  $\varphi \in \Phi_m$  is a (*randomized*) *Markov policy*. Given  $\varphi \in \Phi_m$ , we will use the following notation:

$$(2.3) \quad q_{ij}(t, \varphi) := \int_{A(i)} q_{ij}(a) \varphi_t(da|i) \quad \forall i, j \in S, t \geq 0,$$

which is finite as a consequence of (2.1) and (2.2). We also define

$$(2.4) \quad r(t, i, \varphi) := \int_{A(i)} r(i, a) \varphi_t(da|i) \quad \text{and} \quad u(t, i, \varphi) := \int_{A(i)} u(i, a) \varphi_t(da|i)$$

for  $i, j \in S$  and  $t \geq 0$ . In what follows, we will impose conditions ensuring that the integrals in (2.4) are well defined and finite.

In the family of Markov policies, we will consider two especially relevant classes of policies: stationary policies and deterministic stationary policies.

Given a Markov policy  $\varphi \equiv \{\varphi_t(B|i)\}$ , we say that  $\varphi$  is *stationary* if it does not depend on  $t \geq 0$ , and we will write  $\varphi \equiv \{\varphi(B|i)\}$ . The set of stationary (randomized) policies is denoted by  $\Phi_s$ .

A stationary policy  $\varphi$  is *deterministic* if  $\varphi(\cdot|i)$  is a Dirac probability measure for each  $i \in S$ . The set of deterministic stationary policies can be identified with the class  $\mathbb{F}$  of functions  $f : S \rightarrow \cup_{i \in S} A(i)$  such that  $f(i) \in A(i)$  for every  $i \in S$ . Therefore, we have

$$\mathbb{F} \subseteq \Phi_s \subseteq \Phi_m.$$

When considering stationary policies  $\varphi \in \Phi_s$ , the expressions in (2.3) and (2.4) will be simply written as  $q_{ij}(\varphi)$ ,  $r(i, \varphi)$ , and  $u(i, \varphi)$ .

For each  $\varphi \in \Phi_m$  there exists a nonhomogeneous  $Q$ -process with transition rates given by  $q_{ij}(t, \varphi)$ ; for a proof, we refer to [29]. To ensure the regularity of the  $Q$ -process, we must impose further conditions on the control model.

**Assumptions.** We next state the assumptions on our control model. Our Assumptions A, B, and C below are mainly taken from [10, 25].

*Assumption A.* There exists a nondecreasing function  $w \geq 1$  on  $S$  such that

- $\lim_{i \rightarrow \infty} w(i) = \infty$ ;
- there exist constants  $b \geq c > 0$  and a finite set  $S_0 \subset S$  such that

$$\sum_{j \in S} q_{ij}(a) w(j) \leq -cw(i) + bI_{S_0}(i) \quad \forall (i, a) \in K,$$

where  $I$  denotes the indicator function;

- (c) there exists a constant  $M > 0$  such that  $|r(i, a)| \leq Mw(i)$  and  $|u(i, a)| \leq Mw(i)$  for every  $(i, a) \in K$ .

The function  $w$  in Assumption A is often referred to as a Lyapunov function or a moment-like function. Assumptions A(a) and A(b) ensure that, for each  $\varphi \in \Phi_m$ , there exists a Markov process  $\{x^\varphi(t)\}$ , with transition function  $p^\varphi(s, i, t, j)$ , for  $i, j \in S$  and  $t \geq s \geq 0$ , whose transition rates are given by  $q_{ij}(t, \varphi)$ . For a proof we refer to [9, Theorem 3.1].

Given an initial state  $i \in S$  at time  $s \geq 0$  and a Markov policy  $\varphi \in \Phi_m$ , we will denote by  $P_{s,i}^\varphi$  and  $E_{s,i}^\varphi$  the probability measure and the corresponding expectation operator defined by  $p^\varphi(s, i, t, j)$ . When  $s = 0$ ,  $P_{0,i}^\varphi$ ,  $E_{0,i}^\varphi$ , and  $p^\varphi(0, i, t, j)$  will be simply written as  $P_i^\varphi$ ,  $E_i^\varphi$ , and  $p^\varphi(i, t, j)$ , respectively.

DEFINITION 2.1. *Given  $\varphi \in \Phi_m$ ,  $i \in S$ , and  $T \geq 0$ , we define the total pathwise reward and the total expected reward on  $[0, T]$  as*

$$(2.5) \quad J_T^0(i, \varphi) := \int_0^T r(t, x(t), \varphi) dt \quad \text{and} \quad J_T(i, \varphi) := E_i^\varphi \left[ \int_0^T r(t, x(t), \varphi) dt \right],$$

*respectively. Replacing the reward rate  $r$  with the cost rate  $u$ , we obtain the definition of  $J_{u,T}^0(i, \varphi)$  and  $J_{u,T}(i, \varphi)$ .*

*The long-run pathwise average reward (or pathwise average reward, for short) and the long-run expected average reward (also referred to as gain or expected average reward) are given by*

$$J^0(i, \varphi) := \liminf_{T \rightarrow \infty} \frac{1}{T} J_T^0(i, \varphi) \quad \text{and} \quad J(i, \varphi) := \liminf_{T \rightarrow \infty} \frac{1}{T} J_T(i, \varphi),$$

*respectively. Similarly, the pathwise average cost and the expected average cost are, respectively, defined as*

$$J_u^0(i, \varphi) := \limsup_{T \rightarrow \infty} \frac{1}{T} J_{u,T}^0(i, \varphi) \quad \text{and} \quad J_u(i, \varphi) := \limsup_{T \rightarrow \infty} \frac{1}{T} J_{u,T}(i, \varphi).$$

These definitions follow a standard convention for CMCs: since  $r$  is a reward rate,  $J^0(i, \varphi)$  and  $J(i, \varphi)$  are defined as a “lim inf”; when  $u$  is a cost rate,  $J_u^0(i, \varphi)$  and  $J_u(i, \varphi)$  are defined as a “lim sup.”

We note that in (2.5) the Markov process is denoted by  $x(t)$  rather than  $x^\varphi(t)$ . This is because, for instance, in the notation  $J_T^0(i, \varphi)$  the initial state  $i$  and the policy  $\varphi$  completely determine the probability measure  $P_i^\varphi$ , and thus writing  $x^\varphi(t)$  is somewhat redundant.

Remark 2.2. Observe that, if  $\varphi \in \Phi_m$  is not randomized, then the random variables  $J^0(i, \varphi)$  and  $J_u^0(i, \varphi)$  denote the long-run average reward and cost of a given sample path of  $\{x^\varphi(t)\}_{t \geq 0}$ , while if  $\varphi \in \Phi_m$  is randomized, then the corresponding reward and cost are integrated with respect to  $\varphi_t(da|x(t))$  (recall (2.4)). The reason for this is that the transition rates themselves of the Markov process  $\{x^\varphi(t)\}$  are defined by integration with respect to  $\varphi_t(da|x(t))$ ; see (2.3). This is, however, the usual definition of the long-run pathwise average reward and cost for a randomized policy; see, e.g., [8].

The next lemma, which follows from [9, Theorem 3.1(a)], [10, Lemma 3.2(a)], and [15, Lemma 2.1], yields an estimate on the expected growth of  $w(x(t))$ .



LEMMA 2.3. *If Assumptions A(a) and A(b) hold, then*

$$(2.6) \quad E_i^\varphi w(x(t)) \leq e^{-ct}w(i) + \frac{b}{c}(1 - e^{-ct}) \quad \forall \varphi \in \Phi_m, i \in S, t \geq 0.$$

*Proof.* We give a sketch of the proof. The inequality

$$(2.7) \quad \sum_{j \in S} q_{ij}(a)w(j) \leq -cw(i) + bI_{S_0}(i) \quad \forall (i, a) \in K$$

in Assumption A(b) implies the inequality

$$(2.8) \quad \sum_{j \in S} q_{ij}(t, \varphi)w(j) \leq -cw(i) + b \quad \forall \varphi \in \Phi_m, i \in S, t \geq 0.$$

To see this, integrate (2.7) with respect to  $\varphi_t(da|i)$ . Then, the interchange of sum and integral is derived from the conservative and stability properties (2.1) and (2.2). Now, (2.8) is similar to the condition (a<sub>2</sub>) in [10, Lemma 3.2(a)]. Therefore, the result follows.  $\square$

Furthermore, if Assumption A(c) holds, then we derive from Lemma 2.3 that, for every  $\varphi \in \Phi_m$  and  $i \in S$ ,

$$(2.9) \quad |J(i, \varphi)| \leq bM/c \quad \text{and} \quad |J_u(i, \varphi)| \leq bM/c.$$

Our next assumption imposes the usual continuity-compactness conditions. The function  $w$  is taken from Assumption A.

*Assumption B.*

- (a) For each  $i \in S$ , the action space  $A(i)$  is compact.
- (b) For every  $i \in S$ ,  $q(i) \leq w(i)$ .
- (c) For each  $i, j \in S$ , the functions  $r(i, a)$ ,  $u(i, a)$ ,  $q_{ij}(a)$ , and  $\sum_{k \in S} q_{ik}(a)w(k)$  are continuous on  $A(i)$ .
- (d) There exist constants  $c' > 0$  and  $b' \geq 0$  such that

$$\sum_{j \in S} q_{ij}(a)w^2(j) \leq c'w^2(i) + b' \quad \forall (i, a) \in K.$$

Observe that the continuity of the  $q_{ij}(a)$ , together with Assumptions A(a) and B(d), implies that the series  $\sum_{k \in S} q_{ik}(a)w(k)$  is uniformly convergent on  $A(i)$  and, hence, continuous (cf. the last statement of Assumption B(c)).

Assumption B(b) is not strictly necessary. Indeed, it suffices that the function  $i \mapsto w(i)q(i)$  is bounded by a function  $w'$  satisfying a Lyapunov condition as the one stated in Assumption B(d). However, in practice,  $q(i)$  and  $w(i)$  are of the same order, and there is no loss of generality in assuming that  $q(i) \leq w(i)$ .

*Assumption C(a).* For each  $\varphi \in \Phi_s$ , the Markov process  $\{x^\varphi(t)\}_{t \geq 0}$  is irreducible.

By Assumptions A and C(a), for each  $\varphi \in \Phi_s$  the Markov process has a unique invariant probability measure, denoted by  $\mu_\varphi$ , and

$$\mu_\varphi(w) := \int_S w d\mu_\varphi$$

is finite; see [23, Theorem 4.2]. Moreover, by integration of (2.6) with respect to  $\mu_\varphi$ , we obtain  $\int_S w d\mu_\varphi \leq b/c$  for all  $\varphi \in \Phi_s$ . It follows also that  $J(i, \varphi)$  and  $J_u(i, \varphi)$  in

Definition 2.1 are constant if  $\varphi \in \Phi_s$  (i.e., they do not depend on the initial state of the system), and they verify

$$J(i, \varphi) = \lim_{T \rightarrow \infty} \frac{1}{T} J_T(i, \varphi) = \sum_{j \in S} r(j, \varphi) \mu_\varphi\{j\} =: g(\varphi)$$

and

$$(2.10) \quad J_u(i, \varphi) = \lim_{T \rightarrow \infty} \frac{1}{T} J_{u,T}(i, \varphi) = \sum_{j \in S} u(j, \varphi) \mu_\varphi\{j\} =: g_u(\varphi)$$

for all  $i \in S$ .

The following strong law of large numbers holds. Its proof is a direct consequence of the results in [2] and is therefore omitted. (Next, we will use the standard abbreviation “a.s.” for “almost surely.”)

**PROPOSITION 2.4.** *Suppose that Assumptions A and C(a) hold. For each  $\varphi \in \Phi_s$  and every initial state  $i \in S$ ,*

$$J^0(i, \varphi) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T r(x(t), \varphi) dt = g(\varphi) \quad P_i^\varphi\text{-a.s.}$$

and

$$J_u^0(i, \varphi) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T u(x(t), \varphi) dt = g_u(\varphi) \quad P_i^\varphi\text{-a.s.}$$

**3. Main results.** Now we are ready to define the ergodic control problem with pathwise constraints. Our goal is to maximize with probability one the long-run pathwise average reward  $J^0(i, \varphi)$  over the family of policies  $\varphi \in \Phi_m$  that satisfy the following constraint on the long-run pathwise average cost (which can be interpreted as a budgetary restriction):

$$J_u^0(i, \varphi) \leq \theta_0 \quad P_i^\varphi\text{-a.s.}$$

for  $i \in S$ , where  $\theta_0 \in \mathbb{R}$  is a given constant. In short (writing “s.t.” for “subject to”),

$$(3.1) \quad \max J^0(i, \varphi) \quad \text{s.t.} \quad \varphi \in \Phi_m \quad \text{and} \quad J_u^0(i, \varphi) \leq \theta_0 \quad \text{for } i \in S.$$

It is worth noting that if the stationary policy  $\varphi \in \Phi_s$  verifies  $g_u(\varphi) \leq \theta_0$ , then, as a consequence of Proposition 2.4,  $J_u^0(i, \varphi) \leq \theta_0$  with  $P_i^\varphi$ -probability one for every  $i \in S$ . Our definition of an optimal stationary policy for (3.1) is the following.

**DEFINITION 3.1.** *We say that a policy  $\varphi^* \in \Phi_s$  such that  $g_u(\varphi^*) \leq \theta_0$  is optimal for the pathwise constrained CMC (3.1), or pathwise constrained optimal, if for each  $\varphi \in \Phi_m$  and every  $i \in S$  such that  $J_u^0(i, \varphi) \leq \theta_0$   $P_i^\varphi$ -a.s., we have  $J^0(i, \varphi) \leq g(\varphi^*)$   $P_i^{\varphi^*}$ -a.s.*

If  $\varphi^* \in \Phi_s$  is a pathwise constrained optimal policy, then we define the optimal value of the constrained CMC (3.1) as  $V^* := g(\varphi^*)$ .

**Further assumptions.** Before stating our next assumption, we introduce some notation. We define the following norm on the space of functions  $v : S \rightarrow \mathbb{R}$ :

$$\|v\|_w := \sup_{i \in S} \frac{|v(i)|}{w(i)},$$

and we denote by  $\mathcal{B}_w(S)$  the Banach space of functions on  $S$  with finite  $w$ -norm.

Sufficient conditions for Assumption C(b) below are given in [10, 25]. More precisely, a monotonicity condition is proposed in [10], and a uniform integrability condition is given in [25]. Moreover, in Theorem 7.2 in the appendix we propose a new condition yielding Assumption C(b).

*Assumption C(b).* The control model is  $w$ -exponentially ergodic on  $\mathbb{F}$ ; that is, there exist constants  $\delta > 0$  and  $R > 0$  such that

$$\sup_{f \in \mathbb{F}} |E_i^f v(x(t)) - \mu_f(v)| \leq R e^{-\delta t} \|v\|_w w(i)$$

for every  $i \in S$ ,  $t \geq 0$ , and  $v \in \mathcal{B}_w(S)$ , where  $\mu_f(v) := \int_S v d\mu_f$ .

We state our last assumption.

*Assumption D.* There exist constants  $\tilde{b} \geq \tilde{c} > 0$  and a finite set  $\tilde{S}_0 \subset S$  such that

$$\sum_{j \in S} q_{ij}(a) w^2(j) \leq -\tilde{c} w^2(i) + \tilde{b} I_{\tilde{S}_0}(i) \quad \forall (i, a) \in K.$$

Obviously, Assumption D implies Assumption B(d). Theorem 7.1 in the appendix proves that Assumption D implies Assumption A(b). For future reference we note that, under Assumptions A(a) and D, a result similar to Lemma 2.3 holds. Namely,

$$(3.2) \quad E_i^\varphi w^2(x(t)) \leq e^{-\tilde{c}t} w^2(i) + \frac{\tilde{b}}{\tilde{c}} (1 - e^{-\tilde{c}t}) \quad \forall \varphi \in \Phi_m, i \in S, t \geq 0.$$

If in addition Assumption C(a) is satisfied, then, by integration of (3.2) with respect to  $\mu_\varphi$ ,

$$(3.3) \quad \int_S w^2 d\mu_\varphi \leq \tilde{b}/\tilde{c} \quad \forall \varphi \in \Phi_s.$$

Before stating our main theorem, we need to recall some results on *expected* average reward optimality taken from [10].

**Average reward optimality.** We say that a policy  $\varphi^* \in \Phi_m$  is *average reward optimal* or *gain optimal* for the reward rate function  $r$  if

$$J(i, \varphi^*) = \sup_{\varphi \in \Phi_m} J(i, \varphi) =: J^*(i) \quad \forall i \in S,$$

where the above supremum is finite as a consequence (2.9). Our next theorem characterizes the optimal value function  $J^*$  and the class of average reward optimal stationary policies; for a proof, see [10, Theorem 4.1].

**THEOREM 3.2.** *Suppose that Assumptions A, B, and C hold. Then*

- (i) *there exists a solution  $(g^*, h^*) \in \mathbb{R} \times \mathcal{B}_w(S)$  to the average reward optimality equation (AROE)*

$$(3.4) \quad g^* = \max_{a \in A(i)} \left\{ r(i, a) + \sum_{j \in S} q_{ij}(a) h^*(j) \right\} \quad \forall i \in S;$$

- (ii) *a deterministic stationary policy  $f^* \in \mathbb{F}$  is gain optimal if and only if  $f^*(i)$  attains the maximum in (3.4) for every  $i \in S$ . A stationary policy  $\varphi \in \Phi_s$  is gain optimal if and only if  $\varphi(\cdot|i)$  is supported on the set of maxima of (3.4) for each  $i \in S$ ;*

(iii) the constant  $g^*$  is the optimal gain, i.e.,  $J^*(i) = g^*$  for each  $i \in S$ .

The policies  $f \in \mathbb{F}$  attaining the maximum in the AROE (3.4) are usually referred to as *canonical*. The set of canonical policies is denoted by  $\mathbb{F}_{ca}$ . Since it is straightforward to see that  $\mathbb{F}_{ca}$  is nonempty, Theorem 3.2(ii) tacitly states the *existence of gain optimal policies in  $\mathbb{F}$* .

Obviously, a result similar to Theorem 3.2 holds when the reward rate function  $r$  is replaced with any reward rate function satisfying Assumptions A(c) and B(c).

**Pathwise constrained optimality.** Now we state our main results on the existence of pathwise constrained optimal policies and the characterization of the corresponding value function. (At this point, recall Definition 3.1.)

Under our standing assumptions (in particular, see (2.9) and Theorem 3.2) both

$$(3.5) \quad \theta_{\min} := \min_{\varphi \in \Phi_s} g_u(\varphi) \quad \text{and} \quad \theta_{\max} := \max_{\varphi \in \Phi_s} g_u(\varphi)$$

are finite. When dealing with the constrained problem (3.1), to avoid trivial situations we will assume that the constant  $\theta_0$  verifies

$$(3.6) \quad \theta_{\min} < \theta_0 < \theta_{\max}.$$

**THEOREM 3.3.** *Suppose that Assumptions A, B, C, and D hold, and consider the pathwise constrained CMC (3.1) with  $\theta_0 \in \mathbb{R}$  as in (3.6). Then*

- (i) *there exists a pathwise constrained optimal policy  $\varphi^* \in \Phi_s$ , that is,  $g(\varphi^*) = V^*$  and  $g_u(\varphi^*) \leq \theta_0$ ; in addition,  $\varphi^*$  is such that it randomizes in at most one state;*
- (ii) *there exist  $\lambda_0 \leq 0$  and  $h \in \mathcal{B}_w(S)$  such that*

$$V^* = \max_{a \in A(i)} \left\{ r(i, a) + \lambda_0(u(i, a) - \theta_0) + \sum_{j \in S} q_{ij}(a)h(j) \right\} \quad \forall i \in S;$$

- (iii) *for each  $\lambda \in (-\infty, 0]$ , let  $(g(\lambda), h_\lambda) \in \mathbb{R} \times \mathcal{B}_w(S)$  be a solution to the following AROE (such a solution indeed exists as a consequence of Theorem 3.2):*

$$g(\lambda) = \max_{a \in A(i)} \left\{ r(i, a) + \lambda(u(i, a) - \theta_0) + \sum_{j \in S} q_{ij}(a)h_\lambda(j) \right\} \quad \forall i \in S.$$

*Then  $V^* = \min_{\lambda \leq 0} g(\lambda)$ .*

For the proof of Theorem 3.3, see section 4 below.

**4. Proofs.** To prove the main result of this paper, Theorem 3.3, we need several preliminary results. These concern sample-path average optimality and CMCs with expected constraints.

**Sample-path average optimality.** In what follows, we will suppose that Assumptions A, B, C, and D are satisfied. Theorem 3.2 proves the existence of a gain (or *expected* average reward) optimal policy. Now we prove that there exist policies that are *pathwise* average reward optimal as in Definition 4.1 below.

Similar results for discrete-time Markov decision processes can be found in [18, Chapter 11]. Pathwise average optimality for continuous-time controlled Markov chains, under hypotheses similar to ours, has also been studied in [8]. For a discussion of the results in [8] and ours, see Remark 4.4 below.

DEFINITION 4.1. A policy  $\varphi^* \in \Phi_s$  is said to be sample-path average optimal if for every  $\varphi \in \Phi_m$  and  $i \in S$ ,  $J^0(i, \varphi) \leq g(\varphi^*)$   $P_i^\varphi$ -a.s.

In order to obtain a characterization of sample-path average optimal policies, we need a preliminary result that uses the notation (2.3). We also use  $\xrightarrow{p}$  to denote convergence in probability.

LEMMA 4.2. Suppose that Assumptions A(a), B(b), and D are verified. Given a Markov policy  $\varphi \in \Phi_m$ , an initial state  $i \in S$ , and an arbitrary  $h \in \mathcal{B}_w(S)$ , we have

$$\frac{1}{n} \int_0^n L^{\varphi, t} h(x(t)) dt \xrightarrow{p} 0 \quad \text{as } n \rightarrow \infty$$

with respect to  $P_i^\varphi$ , where  $(L^{\varphi, t} h)(i) := \sum_{j \in S} q_{ij}(t, \varphi) h(j)$  for  $i \in S$ .

*Proof.* By the theory of Markov processes we know that  $Z_t^\varphi := h(x(t)) - h(x(0)) - \int_0^t L^{\varphi, s} h(x(s)) ds$ , for  $t \geq 0$ , is a  $P_i^\varphi$ -martingale. Consider now the discrete-time martingale difference

$$Z_{n+1}^\varphi - Z_n^\varphi = h(x(n+1)) - h(x(n)) - \int_n^{n+1} L^{\varphi, s} h(x(s)) ds \quad \text{for } n \geq 0.$$

Observe that, by Assumptions A(b) (which is implied by Assumption D; see Theorem 7.1) and B(b),  $|(L^{\varphi, t} h)(i)| \leq \|h\|_w(2+b)w^2(i)$  and thus, using (3.2), we deduce that

$$\sup_{n \in \mathbb{N}} E_i^\varphi |Z_{n+1}^\varphi - Z_n^\varphi| < \infty.$$

As a consequence of [13, Theorem 2.18],  $Z_n^\varphi/n$  converges to 0 in probability. Moreover, it is easily proved that  $h(x(n))/n$  also converges in probability to 0. Hence, the result follows.  $\square$

The next result, together with Theorem 3.2, gives a characterization of sample-path average optimal stationary policies and the corresponding optimal value.

THEOREM 4.3. Suppose that Assumptions A, B, C, and D hold. A stationary policy is sample-path average optimal if and only if it is gain optimal.

*Proof.* As a consequence of Proposition 2.4, sample-path average optimal policies are necessarily gain optimal. Let us prove the converse result. Let  $(g^*, h^*) \in \mathbb{R} \times \mathcal{B}_w(S)$  be a solution of the AROE (3.4). We will show that, given a Markov policy  $\varphi \in \Phi_m$  and an initial state  $i \in S$ ,  $J^0(i, \varphi) \leq g^*$   $P_i^\varphi$ -a.s.

To this end, note that by the AROE, for every  $s \geq 0$ ,

$$r(s, x(s), \varphi) + L^{\varphi, s} h^*(x(s)) \leq g^*,$$

and thus

$$\frac{1}{t} \int_0^t r(s, x(s), \varphi) ds + \frac{1}{t} \int_0^t L^{\varphi, s} h^*(x(s)) ds \leq g^* \quad \forall t > 0.$$

By Lemma 4.2, there exists a sequence  $\{t_n\}_{n \geq 0}$  such that  $t_n \rightarrow \infty$  and

$$\frac{1}{t_n} \int_0^{t_n} L^{\varphi, s} h^*(x(s)) ds \xrightarrow{P_i^\varphi\text{-a.s.}} 0 \quad \text{as } n \rightarrow \infty.$$

Therefore, recalling Definition 2.1,

$$J^0(i, \varphi) \leq \liminf_{n \rightarrow \infty} \frac{1}{t_n} \int_0^{t_n} r(s, x(s), \varphi) ds \leq g^* \quad P_i^\varphi\text{-a.s.},$$

as we wanted to prove.  $\square$

*Remark 4.4.* As we have already mentioned, sample-path average optimality has been analyzed in [8]. In that paper, it is proved that canonical policies are sample-path average optimal under the “lim sup criterion” by imposing a Lyapunov condition on  $w^4$  (see [8, Assumption C(2)]), whereas our Theorem 4.3 shows that under a Lyapunov condition on  $w^2$  we reach sample-path average optimality for the “lim inf criterion.”

It is worth noting that, when dealing with stationary policies, the “lim inf” and the “lim sup” criteria are equivalent (see Proposition 2.4). When dealing with *nonstationary* policies, however, proving an inequality such as, e.g.,  $J^0(i, \varphi) \leq g^*$   $P_i^\varphi$ -a.s., is more difficult for the lim sup than for the lim inf criterion.

**CMCs with expected constraints.** As was remarked in section 1, there are several approaches to analyzing constrained CMCs. Here, we combine the Lagrange multipliers and the linear programming techniques with the dynamic programming equation approach. We assume that Assumptions A, B, C, and D hold.

We want to maximize, for every initial state  $i \in S$ ,  $J(i, \varphi)$  over the set of policies  $\varphi \in \Phi_m$  such that  $J_u(i, \varphi) \leq \theta_0$ , where  $\theta_0$  is as in (3.6). In short, we consider the following constrained problem:

$$(4.1) \quad \max J(i, \varphi) \quad \text{s.t.} \quad \varphi \in \Phi_m \quad \text{and} \quad J_u(i, \varphi) \leq \theta_0, \quad \text{for } i \in S.$$

**DEFINITION 4.5.** We say that a policy  $\varphi^* \in \Phi_m$  such that  $J_u(i, \varphi^*) \leq \theta_0$  for every  $i \in S$  is optimal for the constrained CMC (4.1) if, for each  $i \in S$  and every  $\varphi \in \Phi_m$  such that  $J_u(i, \varphi) \leq \theta_0$ , we have  $J(i, \varphi) \leq J(i, \varphi^*)$ .

We denote by  $V^*(i, \theta_0)$  the optimal value function of (4.1), i.e.,

$$V^*(i, \theta_0) := \sup_{\varphi \in \Phi_m} \{J(i, \varphi) : J_u(i, \varphi) \leq \theta_0\} \quad \text{for } i \in S.$$

Now we need to introduce some notation. We define the following norm on the space of functions  $v : S \rightarrow \mathbb{R}$ :

$$\|v\|_1 := \sup_{i \in S} |v(i)|,$$

and we denote by  $\mathcal{B}_1(S)$  the Banach space of functions on  $S$  with finite 1-norm. We also define  $\mathcal{C}_w(K)$  as the set of functions  $v : K \rightarrow \mathbb{R}$  such that  $i \mapsto \sup_{a \in A(i)} |v(i, a)|$  is in  $\mathcal{B}_w(S)$ , and, in addition,  $a \mapsto v(i, a)$  is continuous on  $A(i)$  for each  $i \in S$ . By Assumptions A(c) and B(c), the functions  $r$  and  $u$  are in  $\mathcal{C}_w(K)$ .

We denote by  $\mathcal{P}_w(S)$  the family of probability measures on  $S$  for which the integral of  $w$  is finite. In particular, for every  $\varphi \in \Phi_s$ ,  $\mu_\varphi$  is in  $\mathcal{P}_w(S)$ . Similarly, we denote by  $\mathcal{P}_w(K)$  the set of probability measures  $\mu$  on  $K$  such that  $\sum_{i \in S} w(i) \mu(\{i\} \times A(i)) < \infty$ . Finally, for each  $\varphi \in \Phi_s$ , we define  $\hat{\mu}_\varphi \in \mathcal{P}_w(K)$  as follows: for  $i \in S$  and measurable  $B \subseteq A(i)$ ,

$$\hat{\mu}_\varphi(\{i\} \times B) := \mu_\varphi\{i\} \varphi(B|i).$$

We consider the  $w$ -weak topology on  $\mathcal{P}_w(K)$ , i.e., the smallest topology for which

$$\hat{\mu} \mapsto \int_K v d\hat{\mu}$$

is continuous for every  $v \in \mathcal{C}_w(K)$ . The space  $\mathcal{P}_w(K)$  with the  $w$ -weak topology is a Borel space [5, Appendix A.5].

In what follows, we will analyze several properties of the set

$$\Gamma := \{\hat{\mu}_\varphi : \varphi \in \Phi_s\} \subseteq \mathcal{P}_w(K).$$

Lemma 4.6 below is a standard result (see, e.g., [20, section 3]). However, the denumerability of the state space  $S$  simplifies its proof and weakens its hypotheses. Actually, we use the well-known fact that if  $\mu_\varphi$  is an invariant probability measure for the Markov process with transition rates  $\{q_{ij}(\varphi)\}$ , then  $\sum_{i \in S} \mu_\varphi(i) q_{ij}(\varphi) = 0$  (see [1, Chapter 5]).

LEMMA 4.6. *Given  $\hat{\mu} \in \mathcal{P}_w(K)$ , a necessary and sufficient condition for  $\hat{\mu} \in \Gamma$  is that*

$$(4.2) \quad \int_K Lv \, d\hat{\mu} = 0 \quad \text{for every } v \in \mathcal{B}_1(S),$$

where  $(Lv)(i, a) := \sum_{j \in S} q_{ij}(a)v(j)$  for  $(i, a) \in K$ .

*Proof of the necessity.* Fix  $\hat{\mu} \in \Gamma$  and  $v \in \mathcal{B}_1(S)$ . Hence,  $\hat{\mu} = \hat{\mu}_\varphi$  for some  $\varphi \in \Phi_s$ . We have

$$\begin{aligned} \int_K Lv \, d\hat{\mu} &= \sum_{i \in S} \int_{A(i)} \left[ \sum_{j \in S} q_{ij}(a)v(j) \right] \varphi(da|i) \mu_\varphi\{i\} \\ &= \sum_{i \in S} \sum_{j \in S} q_{ij}(\varphi)v(j) \mu_\varphi\{i\} \\ (4.3) \quad &= \sum_{j \in S} v(j) \sum_{i \in S} \mu_\varphi\{i\} q_{ij}(\varphi) = 0, \end{aligned}$$

where the interchange of the sums in (4.3) follows from Assumptions A(b), B(b), and (3.3). This proves (4.2).

*Proof of the sufficiency.* Suppose that (4.2) holds for some  $\hat{\mu} \in \mathcal{P}_w(K)$ . Therefore, by a standard result on the disintegration of measures [17, Proposition D.8], there exists  $\varphi \in \Phi_s$  such that, for each  $i \in S$  and measurable  $B \subseteq A(i)$ ,

$$\hat{\mu}(\{i\} \times B) = \mu\{i\} \varphi(B|i),$$

where  $\mu \in \mathcal{P}_w(S)$  denotes the marginal of  $\hat{\mu}$  on  $S$ . Hence, as in the proof of the necessary condition, and letting  $v(\cdot) := I_{\{j\}}(\cdot)$ , we can show that  $\sum_{i \in S} \mu\{i\} q_{ij}(\varphi) = 0$  for every  $j \in S$ , and then  $\mu$  is necessarily the invariant probability measure of  $\{x^\varphi(t)\}_{t \geq 0}$ . Thus  $\hat{\mu} = \hat{\mu}_\varphi$ , as we wanted to prove.  $\square$

LEMMA 4.7. *The set  $\Gamma$  is convex and compact for the  $w$ -weak topology.*

*Proof.* It follows directly from Lemma 4.6 that any convex combination of measures in  $\Gamma$  lies in  $\Gamma$ .

For the compactness statement, we derive from Assumptions A(a) and B(a) that the sets  $\{(i, a) \in K : w^2(i) \leq nw(i)\}$  are compact in  $K$  for every  $n \geq 1$ . We also have, by (3.3),

$$\sup_{\hat{\mu} \in \Gamma} \int_K w^2 d\hat{\mu} = \sup_{\varphi \in \Phi_s} \int_S w^2 d\mu_\varphi < \infty.$$

Therefore, from [5, Corollary A.30(c)],  $\Gamma$  is compact.  $\square$

Observe that, when dealing with constrained CMCs, convexity is usually referred to the class of policies (see, e.g., [11, Lemma 3.3]). In this paper, however, we exploit convexity properties of the class of invariant state-action occupation measures, and thus our approach is quite different from that in [11].

Our next result is a direct consequence of the convexity property of  $\Gamma$  proved in Lemma 4.7, and it is stated without proof.

LEMMA 4.8. *The function  $\theta \mapsto V(\theta) := \sup\{\int_K r d\hat{\mu} : \hat{\mu} \in \Gamma, \int_K u d\hat{\mu} \leq \theta\}$  is concave and nondecreasing on  $[\theta_{\min}, \theta_{\max}]$ . (Recall the notation (3.5).)*

In particular, since  $V^*(i, \theta_0)$  is a supremum over  $\Phi_m$  (see Definition 4.5) and  $V(\theta_0)$  is a supremum over  $\Phi_s$ , we have

$$(4.4) \quad V(\theta_0) \leq V^*(i, \theta_0) \quad \forall i \in S.$$

A standard result for control problems with  $n$  constraints is that there exists an optimal randomized stationary policy that is a convex combination of at most  $n + 1$  deterministic stationary policies. (See, e.g., [7].) Part (iii) in the following theorem gives a more explicit characterization of this optimal policy. We use the notation introduced in Definition 4.5.

THEOREM 4.9. *Suppose that Assumptions A, B, C, and D hold, and consider the constrained CMC (4.1) with  $\theta_0$  as in (3.6). Then*

- (i) *the value function  $V^*(i, \theta_0)$  does not depend on  $i \in S$ , and  $V(\theta_0) \equiv V^*(i, \theta_0)$ ;*
- (ii) *there exist  $\lambda_0 \leq 0$  and  $h \in \mathcal{B}_w(S)$  such that*

$$V(\theta_0) = \max_{a \in A(i)} \left\{ r(i, a) + \lambda_0(u(i, a) - \theta_0) + \sum_{j \in S} q_{ij}(a)h(j) \right\} \quad \forall i \in S;$$

- (iii) *there exists an optimal randomized stationary policy which randomizes in at most one state.*

*Proof of (i).* By Lemma 4.8, the function  $V$ , defined on a closed bounded interval, is concave, and therefore its hypograph is a convex set. Notice that for every  $\hat{\mu} \in \Gamma$ , the point with coordinates  $(\int_K u d\hat{\mu}, \int_K r d\hat{\mu})$  belongs to the hypograph of  $V$ .

Let  $-\lambda_0$ , with  $\lambda_0 \leq 0$ , be a superdifferential (in analogy to a subdifferential) of  $V$  at  $\theta_0$ . The hypograph of  $V$  is contained in the half-space

$$\{(x, y) \in \mathbb{R}^2 : \lambda_0(x - \theta_0) + (y - V(\theta_0)) \leq 0\}.$$

In particular, for every  $\hat{\mu} \in \Gamma$ ,

$$\int_K (r + \lambda_0(u - \theta_0)) d\hat{\mu} \leq V(\theta_0),$$

and thus

$$(4.5) \quad \max_{\hat{\mu} \in \Gamma} \int_K (r + \lambda_0(u - \theta_0)) d\hat{\mu} \leq \max_{\hat{\mu} \in \Gamma} \left\{ \int_K r d\hat{\mu} : \int_K u d\hat{\mu} \leq \theta_0 \right\} = V(\theta_0).$$

Note that the two maxima above are attained because the functions  $\int_K r d\hat{\mu}$  and  $\int_K u d\hat{\mu}$  are continuous on  $\Gamma$ , by the definition of the  $w$ -weak topology, and because the sets  $\Gamma$  and  $\{\hat{\mu} \in \Gamma : \int_K u d\hat{\mu} \leq \theta_0\}$  are compact; recall Lemma 4.7.

Now, let us prove that (4.5) holds with equality. Let  $\hat{\mu}^* \in \Gamma$  attain the maximum in the right-hand equality of (4.5). Since  $\lambda_0 \leq 0$ ,

$$V(\theta_0) = \int_K r d\hat{\mu}^* \leq \int_K (r + \lambda_0(u - \theta_0)) d\hat{\mu}^* \leq \max_{\hat{\mu} \in \Gamma} \int_K (r + \lambda_0(u - \theta_0)) d\hat{\mu}.$$



Hence, we have shown that (4.5) holds with equality or, equivalently,

$$\max_{\varphi \in \Phi_s} \int_K (r + \lambda_0(u - \theta_0)) d\hat{\mu}_\varphi = \max_{\varphi \in \Phi_s} \left\{ \int_K r d\hat{\mu}_\varphi : \int_K u d\hat{\mu}_\varphi \leq \theta_0 \right\} = V(\theta_0).$$

The left-most term of this equality equals the optimal value of an expected average reward control problem with reward rate given by  $r + \lambda_0(u - \theta_0)$ . Therefore, by Theorem 3.2, there exists  $h \in \mathcal{B}_w(S)$  such that  $(V(\theta_0), h)$  is a solution of the corresponding AROE, i.e.,

$$(4.6) \quad V(\theta_0) = \max_{a \in A(i)} \left\{ r(i, a) + \lambda_0(u(i, a) - \theta_0) + \sum_{j \in S} q_{ij}(a) h(j) \right\} \\ = \sup_{\varphi \in \Phi_m} \left\{ \liminf_{T \rightarrow \infty} \frac{1}{T} E_i^\varphi \int_0^T [r(t, x(t), \varphi) + \lambda_0(u(t, x(t), \varphi) - \theta_0)] dt \right\}$$

for every  $i \in S$ .

Now, fix an initial state  $i \in S$  and an arbitrary policy  $\varphi \in \Phi_m$  such that  $J_u(i, \varphi) \leq \theta_0$ . We have

$$J(i, \varphi) \leq J(i, \varphi) + \liminf_{T \rightarrow \infty} \frac{1}{T} E_i^\varphi \int_0^T \lambda_0[u(t, x(t), \varphi) - \theta_0] dt \\ \leq \liminf_{T \rightarrow \infty} \frac{1}{T} E_i^\varphi \int_0^T [r(t, x(t), \varphi) + \lambda_0(u(t, x(t), \varphi) - \theta_0)] dt \\ \leq V(\theta_0),$$

and thus  $V^*(i, \theta_0) \leq V(\theta_0)$  for every  $i \in S$ , which together with (4.4) shows that  $V^*(i, \theta_0) = V(\theta_0)$  for every  $i \in S$ . This proves statement (i).

*Proof of (ii).* Statement (ii) follows now from (4.6).

*Proof of (iii).* We know from Theorem 7.2 that the control model is  $w$ -exponentially ergodic on  $\Phi_s$ , and it follows that for every  $\varphi \in \Phi_s$  the gain  $g(\varphi)$  verifies the corresponding Poisson equation (see [10, Lemma 5.1]). Then, by considering the topology of weak convergence on  $\Phi_s$  (see the proof of Theorem 7.2), one can extend the proof of Lemmas 5.3 and 5.4 in [24] and Lemma 5.16 in [12] and prove that the “gain function” (recall (2.10))

$$(4.7) \quad \varphi \mapsto g_u(\varphi)$$

is continuous on  $\Phi_s$ .

From the feasibility condition (3.6) and by the same argument as in the proof of [11, Theorem 2.1], we deduce that there exist two canonical policies (for the AROE (4.6))  $f_1^*$  and  $f_2^*$  in  $\mathbb{F}$ , which differ in at most one state  $i_0 \in S$ , and such that  $g_u(f_1^*) \leq \theta_0 \leq g_u(f_2^*)$ . Fix  $0 \leq \alpha \leq 1$ , and let  $\varphi_\alpha^* \in \Phi_s$  coincide with  $f_1^*$  and  $f_2^*$  in the states  $j \neq i_0$ , and let  $\varphi_\alpha^*(\cdot | i_0)$  randomize between  $f_1^*(i_0)$  and  $f_2^*(i_0)$  with probabilities  $\alpha$  and  $1 - \alpha$ , respectively.

We know that (see Theorem 3.2(ii)) for every  $\alpha \in [0, 1]$ , the policy  $\varphi_\alpha^*$  is expected average optimal for the reward rate  $r + \lambda_0(u - \theta_0)$ . By the continuity property of  $\alpha \mapsto g_u(\varphi_\alpha^*)$  (recall (4.7)), it follows that for some  $\alpha^* \in [0, 1]$ ,  $g_u(\varphi_{\alpha^*}^*) = \theta_0$ , and thus  $\varphi_{\alpha^*}^*$  is an optimal policy for the constrained CMC (4.1).  $\square$

Theorem 4.9 shows that the constrained control problem (4.1) is equivalent to a nonconstrained CMC depending on a constant  $\lambda_0 \leq 0$ . However, the constant  $\lambda_0$  is unknown. Moreover, its value is derived from the function  $V$ , which is precisely the function that we want to determine! To overcome this situation, we propose our next result, based on Ky-Fan's minimax theorem [6].

**THEOREM 4.10.** *Suppose that Assumptions A, B, C, and D are satisfied and consider the constrained CMC (4.1) with  $\theta_0$  as in (3.6).*

*For each  $\lambda \in (-\infty, 0]$ , let  $(g(\lambda), h_\lambda) \in \mathbb{R} \times \mathcal{B}_w(S)$  be a solution to the AROE*

$$g(\lambda) = \max_{a \in A(i)} \left\{ r(i, a) + \lambda(u(i, a) - \theta_0) + \sum_{j \in S} q_{ij}(a) h_\lambda(j) \right\} \quad \forall i \in S.$$

*Then  $V(\theta_0) = \min_{\lambda \leq 0} g(\lambda)$ .*

*Proof.* Consider the function  $H : \Gamma \times (-\infty, 0] \rightarrow \mathbb{R}$  defined by

$$H(\hat{\mu}, \lambda) := \int_K (r + \lambda(u - \theta_0)) d\hat{\mu}.$$

Obviously,  $H$  is concave (as it is linear) on  $\Gamma$  for fixed  $\lambda$ , and it is convex on  $(-\infty, 0]$  for every  $\hat{\mu}$ . Moreover,  $\Gamma$  is convex and compact (recall Lemma 4.7), and  $H$  is continuous on  $\Gamma$  for fixed  $\lambda$  (by the definition of the  $w$ -weak topology). It follows from [6, Theorem 8] that

$$\max_{\hat{\mu} \in \Gamma} \inf_{\lambda \leq 0} H(\hat{\mu}, \lambda) = \inf_{\lambda \leq 0} \max_{\hat{\mu} \in \Gamma} H(\hat{\mu}, \lambda).$$

Hence, on the one hand,  $\inf_{\lambda \leq 0} H(\hat{\mu}, \lambda)$  equals  $-\infty$  if  $\int_K u d\hat{\mu} > \theta_0$  and equals  $\int_K r d\hat{\mu}$  otherwise. Therefore, by Theorem 4.9,

$$V(\theta_0) = \max_{\hat{\mu} \in \Gamma} \inf_{\lambda \leq 0} H(\hat{\mu}, \lambda).$$

On the other hand, by arguments similar to those in Theorem 4.9,

$$\max_{\hat{\mu} \in \Gamma} H(\hat{\mu}, \lambda) = g(\lambda).$$

Therefore,  $V(\theta_0) = \inf_{\lambda \leq 0} g(\lambda)$ . As a consequence of Theorem 4.9, this infimum is attained at  $\lambda_0$ . This completes the proof.  $\square$

Theorem 4.10 shows that the constrained CMC can be solved by means of a parametric family of AROEs, which do not depend on unknown parameters, as it was the case for Theorem 4.9.

Finally, we give the proof of the main result in this paper.

*Proof of Theorem 3.3. Proof of (i).* By Theorem 4.9, we know that there exists an optimal stationary policy  $\varphi^* \in \Phi_s$  for the “expected” constrained problem (4.1), which randomizes in at most one state. Let us show that  $\varphi^*$  is also optimal for (3.1).

We know that

$$(4.8) \quad g(\varphi^*) = V(\theta_0) \quad \text{and} \quad g_u(\varphi^*) \leq \theta_0.$$

Let  $\lambda_0$  be as in Theorem 4.9. Fix an initial state  $i \in S$  and an arbitrary policy  $\varphi \in \Phi_m$ . We know from Theorems 4.3 and 4.9 that

$$\liminf_{T \rightarrow \infty} \frac{1}{T} \int_0^T [r(t, x(t), \varphi) + \lambda_0(u(t, x(t), \varphi) - \theta_0)] dt \leq V(\theta_0) \quad P_i^\varphi\text{-a.s.}$$

Therefore (recalling that  $\lambda_0 \leq 0$ ),

$$V(\theta_0) \geq J^0(i, \varphi) + \lambda_0(J_u^0(i, \varphi) - \theta_0) \quad P_i^\varphi\text{-a.s.}$$

Hence, if  $\varphi \in \Phi_m$  verifies  $J_u^0(i, \varphi) \leq \theta_0$   $P_i^\varphi$ -a.s., then  $J^0(i, \varphi) \leq V(\theta_0)$   $P_i^\varphi$ -a.s.

Therefore, by (4.8),  $g(\varphi^*) = V^* = V(\theta_0)$ , and thus  $\varphi^*$  is optimal for (3.1).

*Proof of (ii) and (iii).* Since  $V^* = V(\theta_0)$ , these results easily follow from Theorems 4.9 and 4.10. This completes the proof of Theorem 3.3.  $\square$

**5. Examples. A single-server controlled queueing system.** Consider a queueing system with state space  $S = \{0, 1, 2, \dots\}$ . The state variable is interpreted as the number of jobs (waiting or being served) in the system. For each  $i \in S$ , the action space  $A(i)$  is a compact metric space. Jobs arrive at a rate  $\lambda > 0$  and they are served at a rate  $\mu > 0$ . The decision-maker controls a service parameter  $h_1(i, a)$  (which allows either an increase or decrease in the service rate) and an arrival parameter  $h_2(i, a)$ , which is interpreted similarly. We suppose that the functions  $h_1$  and  $h_2$  are bounded and continuous on  $K$ . The transition rates of the system are given by

$$q_{01}(a) = -q_{00}(a) = h_2(0, a) \quad \forall a \in A(0),$$

assumed to be positive. In addition, for  $i \geq 1$ ,

$$q_{i,i-1}(a) = \mu i + h_1(i, a) \quad \text{and} \quad q_{i,i+1}(a) = \lambda i + h_2(i, a),$$

both assumed to be positive, and  $q_{ii}(a) = -q_{i,i-1}(a) - q_{i,i+1}(a)$  for every  $a \in A(i)$ .

The decision-maker obtains a reward with rate  $r(i, a) = pi$ , where  $p > 0$ , when there are  $i$  jobs in the system, and incurs a cost with rate  $c(i, a)$  depending on the control actions, where  $c(i, a)$  is assumed to be continuous on  $A(i)$  for each fixed  $i \in S$  and bounded on  $K$ .

This is the same controlled queueing system as in [10, section VI]. In that reference, the authors consider the net reward rate  $r - c$ . It seems plausible, however, that the controller may have an a priori estimate of the maximal cost per time unit that can be incurred when controlling the system. Therefore, the controller should maximize the long-run average reward with rate  $r$  subject to the restriction that the long-run average cost with rate  $c$  is less than or equal to, say,  $\theta_0$ , assumed to belong to the range of  $c$  (see (3.6)). Moreover, in this case, pathwise constraints are very realistic because, for instance, fitting a budget in terms of an “expected value” does not make much sense.

**PROPOSITION 5.1.** *Consider the controlled queueing system defined above, and suppose that  $\mu > \lambda$ . Then the controller has a pathwise constrained optimal policy.*

*Proof.* Let  $w(i) := C(i + 1)$ , where  $C \geq 1$ . Obviously, Assumptions A(a), B(a), B(c), and C(a) hold. Choosing  $C$  large enough, one can easily prove that Assumptions A(c) and B(b) are also satisfied. By Theorems 7.1 and 7.2, if we prove that Assumption D holds, that is, there exist constants  $\tilde{c} > 0$  and  $\tilde{b} \geq 0$ , and a finite set  $\tilde{S}_0 \subset S$  such that

$$(5.1) \quad \sum_{j \in S} q_{ij}(a) w^2(j) \leq -\tilde{c} w^2(i) + \tilde{b} I_{\tilde{S}_0}(i) \quad \forall (i, a) \in K,$$

then Assumptions A, B, and C will be satisfied.

To obtain (5.1) we first note that, by a simple calculation,

$$\sum_{j \in S} q_{ij}(a)w^2(j) \leq w^2(i)(-2(\mu - \lambda) + F(i)) \quad \forall (i, a) \in K,$$

where  $F$  is a positive function such that  $F(i) \rightarrow 0$  as  $i \rightarrow \infty$ . Therefore, letting  $\tilde{c} := \mu - \lambda$ , there exists some  $i^*$  such that

$$(5.2) \quad \sum_{j \in S} q_{ij}(a)w^2(j) \leq -\tilde{c}w^2(i) \quad \text{for } i > i^* \text{ and } (i, a) \in K.$$

Also, for large enough  $\tilde{b}$ ,

$$(5.3) \quad \sum_{j \in S} q_{ij}(a)w^2(j) \leq -\tilde{c}w^2(i) + \tilde{b} \quad \text{for } i \leq i^* \text{ and } (i, a) \in K.$$

We can now derive (5.1) from (5.2) and (5.3), with  $\tilde{S}_0 = \{0, 1, \dots, i^*\}$ . The stated result follows from Theorem 3.3.  $\square$

Observe that we have solved the control problem under hypotheses that are weaker than those in [10]. Indeed, we impose only the condition  $\mu > \lambda$ , whereas in [10] the authors require further assumptions (see Condition E<sub>4</sub> in [10, page 244]).

The reason for this is that, in [10], the authors derive  $w$ -exponential ergodicity from the monotonicity conditions in [10, Assumption C]. In particular, those assumptions require the set  $S_0$  in our Assumption A(b) to be necessarily  $S_0 = \{0\}$ . However, under our hypotheses, we derive  $w$ -exponential ergodicity from a Lyapunov condition on  $w^2$  (Theorem 7.2), and thus the set  $S_0$  in Assumption A(b) may be any petite set, in particular, any finite set, and thus Assumption A(b) becomes easier to verify. Similarly, the condition E<sub>1</sub>(b) in [8, page 43] is not needed. This shows that Assumption D appears to be of a great applicability in models of practical interest.

The following example consists of a bidimensional denumerable state space that can be enumerated in such a way that the Lyapunov condition in Assumption A is easily verified, in particular, the monotonicity requirement.

**A controlled epidemic process.** We analyze a birth and death epidemic process similar to the one described in [21]. We divide the population into three disjoint classes: the susceptibles (individuals who are not infected, but who are exposed to contagion), the infectives (who can transmit the infection to the susceptibles), and the immunized. For expositional clarity, we will focus on the dynamic evolution of the susceptibles and the infectives.

We assume that there are two sources of infection: either by propagation due to the presence of infectives or by an external source. To consider a more realistic model, we will suppose that the infectives and the susceptibles are subject to “natural” birth and death rates. We suppose that an infected person who recovers from the disease is immunized.

Finally, the controller wishes to minimize the average number of infectives, subject to a certain restriction on the expenses of a quarantine program (in order to avoid the propagation of the infection) and the level of medical treatment applied to the population.

We consider the state space  $S = \{0, 1, 2, \dots\} \times \{0, 1, 2, \dots\}$ , where  $(i, j) \in S$  denotes the number of infectives and susceptibles, respectively. When the population is at state  $(i, j)$ , the possible transitions are to the following states:

- $(i+1, j-1)$  (that is, a susceptible has been infected) with a rate  $\lambda_1(a_1)\sqrt{ij} + \lambda_2(a_2)j$ , where  $\sqrt{ij}$  corresponds to propagation of the infection (this expression was proposed in [22]), the control  $a_1 \in A_1$  stands for the level of the quarantine program, and where  $\lambda_2(a_2)j$  is the rate of infection from an external source (corrected according to the level  $a_2 \in A_2$  of the medical treatment);
- $(i, j+1)$  with a rate  $\lambda_S(i+j+1)$ , where  $\lambda_S > 0$  is the natural birth rate of a susceptible individual;
- $(i-1, j)$  with a rate  $\mu_I i + \mu(a_2)i$ , where  $\mu_I > 0$  is the natural death rate of an infected person, and where  $\mu(a_2)i$  is a recovery rate (depending on the level  $a_2$  of the medical treatment);
- $(i, j-1)$  with a rate  $\mu_S j$ , where  $\mu_S > 0$  is the natural death rate of the susceptibles;
- $(i+1, j)$  with a rate  $\lambda_I(i+j+1)$ , where  $\lambda_I > 0$  is the natural birth rate of an infective.

We suppose that the action space is  $A = A_1 \times A_2$ , where  $A_1$  and  $A_2$  are closed bounded intervals in  $[0, \infty)$ . We also suppose that the functions  $\lambda_1$ ,  $\lambda_2$ , and  $\mu$  are positive and continuous.

The reward rate to be maximized by the controller is  $r(i, j, a_1, a_2) = -i$ . The cost rate in the state  $(i, j) \in S$  is  $c_1(a_1)i + c_2(a_2)i + c_3(a_2)j$ . The first term corresponds to the quarantine program (applied to the infectives) and the other two to the medical treatment (applied to both the infectives and the susceptibles). Suppose that  $c_1$ ,  $c_2$ , and  $c_3$  are continuous.

We denote the state process under the policy  $\varphi \in \Phi_m$  by

$$\{x^\varphi(t)\}_{t \geq 0} \equiv \{(i^\varphi(t), j^\varphi(t))\}_{t \geq 0}.$$

The controller wishes to determine a policy  $\varphi \in \Phi_m$  that minimizes

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T i^\varphi(t) dt \quad \text{a.s.}$$

subject to the budgetary restriction

$$\begin{aligned} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T [(c_1(t, x^\varphi(t), \varphi) + c_2(t, x^\varphi(t), \varphi))i^\varphi(t) \\ + c_3(t, x^\varphi(t), \varphi)j^\varphi(t)] dt \leq \theta_0 \quad \text{a.s.}, \end{aligned}$$

for a given  $\theta_0$  that is assumed to be in the corresponding range (see (3.6)).

We choose a Lyapunov function of the form  $w(i, j) = C(i+j+1)$ , where  $C \geq 1$ . Observe that this function is nondecreasing on  $S$ , provided that the state space  $S$  is “enumerated” in the order

$$(0, 0), (1, 0), (0, 1), (2, 0), (1, 1), (0, 2), (3, 0), (2, 1), (1, 2), (0, 3), (4, 0), \dots$$

**PROPOSITION 5.2.** *Consider the controlled epidemic process described above, and suppose that*

$$(5.4) \quad \min\{\mu_I + \mu_S, \mu(a_2)\} > \lambda_I + \lambda_S \quad \forall a_2 \in A_2.$$

*Then there exists a stationary pathwise constrained optimal policy.*

*Proof.* As in the proof of Proposition 5.1, we can show that Assumptions A, B, C(a), and D hold, provided that the constant  $C$  (in the definition of  $w$ ) is large enough. The stated result follows from Theorem 3.3.

Note that the condition in the statement of this result depends neither on  $\lambda_1$  nor on  $\lambda_2$ . This is because the corresponding transition (from  $(i, j)$  to  $(i + 1, j - 1)$ ) does not modify the total size  $i + j$  of the population, and thus, in terms of stability, it does not affect the behavior of the dynamic system.  $\square$

Observe that, since the state space  $S$  is bidimensional, the total population can decrease along either  $i$  or  $j$ . Hence, the condition (5.4) involves the minimum of the two corresponding death rates.

Once again, we see that Theorem 7.2 allows us to show the existence of optimal policies for the corresponding control problem under fairly general assumptions. Indeed, the condition  $\mu > \lambda$  (Proposition 5.1) and  $\min\{\mu_I + \mu_S, \mu(a_2)\} > \lambda_I + \lambda_S$  for  $a_2 \in A_2$  (Proposition 5.2) are just the usual stability conditions for a noncontrolled dynamic system; that is, the service rate (respectively, the death rate) is greater than the arrival rate (respectively, the birth rate).

**6. Concluding remarks.** Our main motivation for this paper was to study average reward continuous-time CMCs with pathwise constraints, but, as shown in the previous sections, on our way to the final result (Theorem 3.3) we provided a detailed analysis of pathwise average optimality (Theorem 4.3) and constrained CMCs with *expected* constraints (Theorems 4.9 and 4.10). Furthermore, the examples in section 5 clearly show that our assumptions are verifiable with no special degree of difficulty.

An important open question is whether our approach here—for *denumerable-state* CMCs—can be systematically extended to more general Markov control problems, for instance, controlled diffusions [28].

**7. Appendix.** The next result is useful for the verification of the Lyapunov conditions. In particular, it shows that Assumption D implies Assumption A(b).

**THEOREM 7.1.** *Suppose that  $w : S \rightarrow [1, \infty)$  is a nondecreasing function such that  $\lim_{i \rightarrow \infty} w(i) = \infty$ , and fix an arbitrary  $0 < \alpha < 1$ . Suppose also that there exist constants  $b \geq c > 0$  and a finite set  $S_0 \subset S$  such that*

$$(7.1) \quad \sum_{j \in S} q_{ij}(a)w(j) \leq -cw(i) + bI_{S_0}(i) \quad \forall (i, a) \in K,$$

where the transition rates  $q_{ij}(a)$  verify the conditions (2.1) and (2.2). Then there exist constants  $\hat{b} \geq \hat{c} > 0$  such that

$$(7.2) \quad \sum_{j \in S} q_{ij}(a)w^\alpha(j) \leq -\hat{c}w^\alpha(i) + \hat{b}I_{S_0}(i) \quad \forall (i, a) \in K.$$

*Proof.* Fix  $i \in S$  and choose a constant  $C > c$ . Let  $x := C + q(i)$ . The inequality (7.1) can be rewritten as

$$\frac{1}{x} \left( \sum_{j \neq i} q_{ij}(a)w(j) + (x + q_{ii}(a))w(i) \right) \leq \frac{x - c}{x}w(i) + \frac{b}{x}I_{S_0}(i).$$

By Jensen's inequality, it follows that

$$\sum_{j \in S} q_{ij}(a)w^\alpha(j) + xw^\alpha(i) \leq x \left( (x - c)x^{-1}w(i) + bx^{-1}I_{S_0}(i) \right)^\alpha.$$

Using the inequality  $(y + z)^\alpha \leq y^\alpha + z^\alpha$ , we obtain

$$\sum_{j \in S} q_{ij}(a) w^\alpha(j) \leq [(x - c)^\alpha x^{1-\alpha} - x] w^\alpha(i) + b^\alpha x^{1-\alpha} I_{S_0}(i).$$

The concavity of the function  $y \mapsto y^\alpha$  implies that  $y^\alpha - z^\alpha \leq \alpha(y - z)z^{\alpha-1}$  whenever  $0 \leq y < z$ . Hence,

$$\sum_{j \in S} q_{ij}(a) w^\alpha(j) \leq -\alpha c w^\alpha(i) + b^\alpha x^{1-\alpha} I_{S_0}(i) \quad \forall (i, a) \in K.$$

Therefore, (7.2) holds with  $\hat{c} := \alpha c$  and  $\hat{b} := b^\alpha \max_{i \in S_0} \{(C + q(i))^{1-\alpha}\}$ . The fact that  $\hat{b} \geq \hat{c}$  follows from the inequality (2.6) when  $w$  is replaced with  $w^\alpha$ . This completes the proof.  $\square$

In our next result, we propose an alternative sufficient condition for uniform  $w$ -exponential ergodicity of CMCs (in particular, for Assumption C(b)).

**THEOREM 7.2.** *Suppose that Assumptions A(a), B(a), C(a), and D, as well as the continuity condition on  $a \mapsto q_{ij}(a)$  in Assumption B(c), are satisfied. Then the control model is  $w$ -exponentially ergodic on  $\Phi_s$ , that is, there exist constants  $\delta > 0$  and  $R > 0$  such that*

$$\sup_{\varphi \in \Phi_s} |E_i^\varphi v(x(t)) - \mu_\varphi(v)| \leq R e^{-\delta t} \|v\|_w w(i)$$

for every  $i \in S$ ,  $t \geq 0$ , and  $v \in \mathcal{B}_w(S)$ .

*Proof.* Fix an initial state  $i \in S$ , an integer  $k \geq 0$ , and  $t \geq 0$ . By (3.2),

$$\begin{aligned} w^2(i) + \tilde{b}/\tilde{c} &\geq \sup_{\varphi \in \Phi_s} \sum_{j \in S} p^\varphi(i, t, j) w^2(j) \geq \sup_{\varphi \in \Phi_s} \sum_{j \geq k} p^\varphi(i, t, j) w^2(j) \\ &\geq w(k) \cdot \sup_{\varphi \in \Phi_s} \sum_{j \geq k} p^\varphi(i, t, j) w(j). \end{aligned}$$

Therefore,  $\lim_{k \rightarrow \infty} \sup_{\varphi \in \Phi_s} \sum_{j \geq k} p^\varphi(i, t, j) w(j) = 0$ , which is precisely the uniform integrability condition in [25, Theorem 2.5] extended to the class of randomized stationary policies.

However, in the proof of [25, Theorem 2.5], the set  $S_0$  in Assumption A(b) is  $S_0 = \{0\}$ , and  $w$ -exponential ergodicity reduces to the class of deterministic stationary policies  $\mathbb{F}$ . This is not a real problem because the proof can be modified in order to account for the case of a finite set  $S_0$  and the class  $\Phi_s$ . Let us mention the main steps of this modified proof:

1. Observe that the corresponding discrete-time result (Key Theorem II in [4]) is stated for any finite so-called “taboo” set. In particular, choose  $S_0$  as the taboo set. The proof of Theorem 2.5 in [25] is the same, except for [25, Proposition 4.9]. However, using the same reasoning, one can prove a result similar to [25, Proposition 4.9], but now with a taboo set  $S_0$ .
2. In the set of stationary policies  $\Phi_s$  we consider the topology of the weak convergence, that is,  $\varphi_n \rightarrow \varphi$  when  $\varphi_n(\cdot|i)$  weakly converges to  $\varphi(\cdot|i)$  for every  $i \in S$ . With this topology,  $\Phi_s$  is compact and metrizable. The proof of Theorem 2.5 in [25] is now easily extended from  $\mathbb{F}$  to  $\Phi_s$ .

This completes the proof.  $\square$

Theorem 7.2 is of valuable interest by itself. In view of this theorem we now know that there exist (at least) *three* different sets of sufficient conditions for the  $w$ -exponential ergodicity in Assumption C(b), namely, monotonicity conditions [10], uniform integrability conditions [25], and a Lyapunov condition on  $w^2$  as in Assumption D (in fact, we can replace  $w^2$  with  $w^{1+\varepsilon}$  for some  $\varepsilon > 0$ ).

## REFERENCES

- [1] W. J. ANDERSON, *Continuous-Time Markov Chains*, Springer, New York, 1991.
- [2] R. N. BHATTACHARYA, *On the functional central limit theorem and the law of the iterated logarithm for Markov processes*, Z. Wahrsch. Verw. Gebiete, 60 (1982), pp. 185–201.
- [3] R. C. CHEN AND G. L. BLANKENSHIP, *Dynamic programming equations for discounted constrained stochastic control*, IEEE Trans. Automat. Control, 49 (2004), pp. 699–709.
- [4] R. DEKKER, A. HORDIJK, AND F. M. SPIEKSMAN, *On the relation between recurrence and ergodicity properties in denumerable Markov decision chains*, Math. Oper. Res., 19 (1994), pp. 539–559.
- [5] H. FÖLLMER AND A. SCHIED, *Stochastic Finance. An Introduction in Discrete Time*, De Gruyter Stud. Math. 27, Walter de Gruyter & Co., Berlin, 2002.
- [6] J. B. G. FRENK, G. KASSAY, AND J. KOLUMBÁN, *On equivalent results in minimax theory*, European J. Oper. Res., 157 (2004), pp. 46–58.
- [7] J. GONZÁLEZ-HERNÁNDEZ AND O. HERNÁNDEZ-LERMA, *Extreme points of sets of randomized strategies in constrained optimization and control problems*, SIAM J. Optim., 15 (2005), pp. 1085–1104.
- [8] X. P. GUO AND X.-R. CAO, *Optimal control of ergodic continuous-time Markov chains with average sample-path rewards*, SIAM J. Control Optim., 44 (2005), pp. 29–48.
- [9] X. P. GUO AND O. HERNÁNDEZ-LERMA, *Continuous-time controlled Markov chains with discounted rewards*, Acta Appl. Math., 79 (2003), pp. 195–216.
- [10] X. P. GUO AND O. HERNÁNDEZ-LERMA, *Drift and monotonicity conditions for continuous-time controlled Markov chains*, IEEE Trans. Automat. Control, 48 (2003), pp. 236–244.
- [11] X. P. GUO AND O. HERNÁNDEZ-LERMA, *Constrained continuous-time Markov control processes with discounted criteria*, Stoch. Anal. Appl., 21 (2003), pp. 379–399.
- [12] X. P. GUO, O. HERNÁNDEZ-LERMA, AND T. PRIETO-RUMEAU, *A survey of recent results on continuous-time Markov decision processes*, Top, 14 (2006), pp. 177–261.
- [13] P. HALL AND C. C. HEYDE, *Martingale Limit Theory and Its Applications*, Academic Press, New York, 1980.
- [14] M. HAVIV, *On constrained Markov decision processes*, Oper. Res. Lett., 19 (1996), pp. 25–28.
- [15] O. HERNÁNDEZ-LERMA, *Lectures on Continuous-Time Markov Control Processes*, Sociedad Matemática Mexicana, Mexico City, 1994.
- [16] O. HERNÁNDEZ-LERMA, J. GONZÁLEZ-HERNÁNDEZ, AND R. R. LÓPEZ-MARTÍNEZ, *Constrained average cost Markov control processes in Borel spaces*, SIAM J. Control Optim., 42 (2003), pp. 442–468.
- [17] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer, New York, 1996.
- [18] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Further Topics on Discrete-Time Markov Control Processes*, Springer, New York, 1999.
- [19] L. A. KORF, *Approximating infinite horizon stochastic optimal control in discrete time with constraints*, Ann. Oper. Res., 142 (2006), pp. 165–186.
- [20] T. G. KURTZ AND R. H. STOCKBRIDGE, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim., 36 (1998), pp. 609–653.
- [21] C. LEFÈVRE, *Optimal control of a birth and death epidemic process*, Oper. Res., 29 (1981), pp. 971–982.
- [22] D. R. MCNEIL, *On the simple stochastic epidemic*, Biometrika, 59 (1972), pp. 494–497.
- [23] S. P. MEYN AND R. L. TWEEDIE, *Stability of Markovian processes III: Foster-Lyapunov criteria for continuous-time processes*, Adv. Appl. Prob., 25 (1993), pp. 518–548.
- [24] T. PRIETO-RUMEAU AND O. HERNÁNDEZ-LERMA, *The Laurent series, sensitive discount and Blackwell optimality for continuous-time controlled Markov chains*, Math. Methods Oper. Res., 61 (2005), pp. 123–145.
- [25] T. PRIETO-RUMEAU AND O. HERNÁNDEZ-LERMA, *Bias optimality for continuous-time controlled Markov chains*, SIAM J. Control Optim., 45 (2006), pp. 51–73.



- [26] K. W. ROSS AND R. VARADARAJAN, *Markov decision processes with sample path constraints: The communicating case*, Oper. Res., 37 (1989), pp. 780–790.
- [27] K. W. ROSS AND R. VARADARAJAN, *Multichain Markov decision processes with a sample path constraint: A decomposition approach*, Math. Oper. Res., 16 (1991), pp. 195–207.
- [28] W. WU, A. ARAPOSTATHIS, AND S. SHAKKOTTAI, *Optimal power allocation for a time-varying wireless channel under heavy-traffic approximation*, IEEE Trans. Automat. Control, 51 (2006), pp. 580–594.
- [29] L. YE, X. P. GUO, AND O. HERNÁNDEZ-LERMA, *Existence and regularity of a nonhomogeneous transition matrix under measurability conditions*, J. Theoret. Probab., to appear. Available online at <http://arxiv.org/abs/0804.4441>.
- [30] A. ZADOROJNIY AND A. SHWARTZ, *Robustness of policies in constrained Markov decision processes*, IEEE Trans. Automat. Control, 51 (2006), pp. 635–638.

## CONTROLLABILITY OF THE KIRCHHOFF SYSTEM FOR BEAMS AS A LIMIT OF THE MINDLIN–TIMOSHENKO SYSTEM\*

F. D. ARARUNA<sup>†</sup> AND E. ZUAZUA<sup>‡</sup>

**Abstract.** We consider the dynamical one-dimensional Mindlin–Timoshenko system for beams. We analyze how its controllability properties depend on the modulus of elasticity in shear  $k$ . In particular we prove that the exact boundary controllability property of the Kirchhoff system may be obtained as a singular limit, as  $k \rightarrow \infty$ , of the partial controllability of projections over a sharp subspace of solutions generated by the eigenfunctions that converge, as  $k \rightarrow \infty$ , towards the spectrum of the Kirchhoff system.

**Key words.** vibrating beams, controllability, observability, Mindlin–Timoshenko, Kirchhoff, Ingham inequality, Fourier decomposition, singular limit

**AMS subject classifications.** 73K05, 93B05, 93B07

**DOI.** 10.1137/060659934

**1. Introduction.** The Mindlin–Timoshenko system of equations is a widely used and, physically, fairly complete mathematical model for describing the transverse vibrations of beams. For a beam of length  $L$  this one-dimensional system reads as follows:

$$(1.1) \quad \begin{cases} \frac{\rho h^3}{12} u'' - u_{xx} + k(u + v_x) = 0 & \text{in } Q, \\ \rho h v'' - k(u + v_x)_x = 0 & \text{in } Q, \end{cases}$$

where  $Q = (0, L) \times (0, T)$ ,  $(0, L)$  being the segment occupied by the beam with  $L > 0$  and  $T$  a given positive time. In this coupled system of two second order hyperbolic equations, the prime  $'$  stands for the partial derivative in time  $t$  and the subscript  $x$  for the space derivative. The unknown  $u = u(x, t)$  represents the angle of rotation and  $v = v(x, t)$  the vertical displacement at time  $t$  of the cross section located  $x$  units from the end-point  $x = 0$ . The constant  $h > 0$  represents the thickness of the beam that, for this model, is considered to be small and uniform, independent of  $x$ . The constant  $\rho$  is the mass density per unit volume of the beam, and the parameter  $k$  is the so-called modulus of elasticity in shear. It is given by the formula  $k = \widehat{k} E h / 2(1 + \mu)$ , where  $\widehat{k}$  is a shear correction coefficient,  $E$  is the Young's modulus, and  $\mu$  is the Poisson's ratio,  $0 < \mu < 1/2$ .

We impose the following boundary conditions on the left-hand side:

$$(1.2) \quad u(0, \cdot) = 0, \quad v_x(0, \cdot) = \Theta_k \quad \text{on } (0, T),$$

\*Received by the editors May 16, 2006; accepted for publication (in revised form) March 5, 2008; published electronically June 25, 2008.

<http://www.siam.org/journals/sicon/47-4/65993.html>

<sup>†</sup>Departamento de Matemática, Universidade Federal da Paraíba, 58051-900 João Pessoa - PB, Brazil (fagner@mat.ufpb.br). The research of this author was partially supported by PDEE-CAPES (MCT-Brasil) grant BEX0412/02-3.

<sup>‡</sup>IMDEA-Matemáticas and Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain (enrique.zuazua@uam.es). The research of this author was supported by grant MTM2005-00714, the DOMINO Project CIT-370200-2005-10 in the PROFIT Program, the i-MATH project of the CONSOLIDER program of the Spanish MEC, and the SIMUMAT Project of the CAM (Spain).

and on the right-hand side we impose

$$(1.3) \quad u(L, \cdot) = v_x(L, \cdot) = 0 \quad \text{on} \quad (0, T).$$

According to these conditions, the angle of rotation is kept fixed both at  $x = 0$  and  $x = L$ , and the boundary control  $\Theta_k$  is a lateral force applied on the vertical displacement at the extreme  $x = 0$ . In particular, no control is applied at  $x = L$ . To make the system complete, let us include the initial conditions

$$(1.4) \quad u(\cdot, 0) = u_0, \quad u'(\cdot, 0) = u_1, \quad v(\cdot, 0) = v_0, \quad v'(\cdot, 0) = v_1 \quad \text{in} \quad (0, L).$$

When assuming that the linear filament of the beam remains perpendicular to the deformed middle surface, the transverse shear effects are neglected and one obtains the so-called Kirchhoff system (see Lagnese and Lions [7])

$$(1.5) \quad \begin{cases} \rho h v'' - \frac{\rho h^3}{12} v''_{xx} + v_{xxxx} = 0 & \text{in } Q, \\ v_x(0, \cdot) = v_x(L, \cdot) = 0 & \text{on } (0, T), \\ v_{xxx}(0, \cdot) = \Xi, \quad v_{xxx}(L, \cdot) = 0 & \text{on } (0, T), \\ v(\cdot, 0) = v_0, \quad v'(\cdot, 0) = v_1 & \text{in } (0, L). \end{cases}$$

The control  $\Xi$  enters this system through the third derivative of the state at  $x = 0$ .

Note that neglecting the shear effects of the beam is formally equivalent to making the modulus  $k$  tend to infinity in (1.1), since  $k$  is inversely proportional to the shear angle.

The connections of these two systems and the singular perturbation problem of passing to the limit as  $k$  tends to infinity have been recently intensively investigated. We refer, for instance, to [11], [12], and [13], where these issues are addressed for a number of nonlinear models and under various boundary conditions.

This paper is devoted to the analysis of the controllability properties of these systems and the corresponding singular perturbation problem (as  $k \rightarrow \infty$ ). Our main goal is analyzing whether the exact controllability property of the Kirchhoff system (1.5) for beams may be obtained as a limit of those of the Mindlin–Timoshenko system (1.1)–(1.4) when the singular parameter  $k$  tends to infinity.

The problem of exact controllability for the Mindlin–Timoshenko system can be formulated as follows: given  $T > 0$ , large enough, and initial data  $\{u_0, u_1, v_0, v_1\}$ , to find a control  $\Theta_k$  such that the solution of system (1.1)–(1.4) satisfies the conditions

$$u(\cdot, T) = u'(\cdot, T) = v(\cdot, T) = v'(\cdot, T) = 0 \quad \text{in} \quad (0, L).$$

According to the Hilbert uniqueness method (HUM) introduced by Lions (see [8]), this property is equivalent to a suitable observability inequality for the adjoint system that, after reversing the sense of time, can be written as follows:

$$(1.6) \quad \begin{cases} \frac{\rho h^3}{12} \phi'' - \phi_{xx} + k(\phi + \psi_x) = f & \text{in } Q, \\ \rho h \psi'' - k(\phi + \psi_x)_x = g & \text{in } Q, \\ \phi(0, t) = \phi(L, t) = \psi_x(0, t) = \psi_x(L, t) = 0 & \text{on } (0, T), \\ \phi(x, 0) = \phi_0(x), \quad \phi'(x, 0) = \phi_1(x) & \text{in } (0, L), \\ \psi(x, 0) = \psi_0(x), \quad \psi'(x, 0) = \psi_1(x) & \text{in } (0, L). \end{cases}$$

To be more precise, in the observability problem,  $f$  and  $g$  vanish ( $f \equiv g \equiv 0$ ) so that the problem consists of estimating the energy of the initial data in terms of

boundary measurements. The general system (1.6) with nonvanishing right-hand side terms is useful when analyzing the well-posedness of the nonhomogeneous boundary value problem (1.1)–(1.4) by transposition.

We are mainly interested in the behavior of the controls  $\Theta_k$ , as  $k \rightarrow \infty$ , and whether in the limit as  $k \rightarrow \infty$  one obtains a control  $\Xi$  such that the solution of system (1.5) verifies

$$v(\cdot, T) = v'(\cdot, T) = 0 \quad \text{in} \quad (0, L).$$

This problem was treated initially in [7] with different boundary conditions. The goals in [7] were

- (i) to show that the control time  $T$  is independent of  $k$ , for any given initial state, and to find, for each  $k$ , a control  $\Theta_k$  driving the system (1.1)–(1.4) to rest at time  $T$ , and
- (ii) to study the behavior of  $\Theta_k$  as  $k \rightarrow \infty$ .

Combining the HUM and multiplier inequalities, the authors of [7] obtained a control time independent of  $k$ . But, to prove (ii), they imposed the physically unrealistic extra assumption that  $L < h$ . Moreover, in [7, Remarks 3.4 and 3.5, p. 109], it was conjectured that *as  $k \rightarrow \infty$ ,  $\Theta_k$  converges, in some appropriate sense, towards a control driving the system (1.5) to equilibrium in time  $T$ .*

In this paper we obtain the following main results:

- The controls  $\Theta_k$  of the Mindlin–Timoshenko system may diverge exponentially as  $k \rightarrow \infty$ .
- By analyzing the underlying spectrum, it is possible to decompose the adjoint system (1.6) into two subsystems. It is sufficient to obtain a uniform (with relation to  $k$ ) observability inequality for one of these subsystems.
- Accordingly, the exact controllability requirement on system (1.1)–(1.4) is relaxed to a partial controllability property over a suitable projection of solutions, and the controls  $\Theta_k$  remain bounded as  $k \rightarrow \infty$ .
- The partial controls  $\Theta_k$  obtained this way converge to an exact control for the limit system (1.5).

With these results, we conclude that the exact controllability property of the Kirchhoff system may be obtained as a limit of the partial controllability property of the Mindlin–Timoshenko system. This solves the problem proposed by Lagnese and Lions in [7] for the present boundary conditions. The uniform (with respect to the parameter  $k$ ) partial controllability result is taken over the subspace of the solutions generated by the eigenfunctions that, in the limit, cover the whole spectrum of the limit Kirchhoff model.

The rest of the paper is organized as follows. In section 2 we briefly mention some elementary properties (existence, uniqueness, and regularity) of solutions for system (1.6) and we rigorously study its limit behavior as  $k \rightarrow \infty$  towards the Kirchhoff system. In section 3 we analyze the properties of the spectrum of system (1.6), finding two families of eigenvalues. As  $k \rightarrow \infty$ , one of these families of eigenvalues tends to those of the limit Kirchhoff system, while the other one diverges, disappearing in the limit in the sense that, since they diverge, they do not lead to eigenvalues of the limit system. This fact occurs due to (and it is in agreement with) the asymptotic simplification that is produced when passing from a system of two equations and two dependent variables to a scalar equation with only one unknown variable. In section 4 we discuss some elementary properties of system (1.1)–(1.4) with nonhomogeneous boundary conditions, i.e., in the absence of controls. We also analyze the convergence,

as  $k \rightarrow \infty$ , towards the solution of the nonhomogeneous Kirchhoff system. Section 5 is devoted to the problem of observability for system (1.6). We show that the observability constant may blow up exponentially as  $k \rightarrow \infty$ . In section 6, applying the Ingham inequality in the Fourier decomposition of solutions, we get a uniform observability result filtering the eigenfunction components corresponding to eigenvalues that diverge as  $k \rightarrow \infty$ . Filtering corresponds, in other words, to projecting solutions over the subspace of eigencomponents that are well behaved. In section 7, combining the results of the previous section with the HUM, we derive the uniform partial controllability result. More precisely, we prove that the projection over the subspace of solutions of (1.1)–(1.4) generated by the eigenvalues convergent (as  $k \rightarrow \infty$ ) and their corresponding eigenfunctions is uniformly controllable with respect to  $k$ . In the limit we obtain the exact boundary controllability property of the Kirchhoff system (1.5). Therefore, we see that it suffices to consider only the solutions in a suitable subspace to ensure that the conjecture in [7] is true.

The analysis in this paper depends on the boundary conditions we have chosen that make possible the explicit computation of the spectrum. Similar results are expected for other boundary conditions, but a further analysis of this issue is needed.

The results in this paper are related to previous ones on the behavior of controls for systems of vibrations under singular perturbations. We refer to [2] and [3] for the problem of control and homogenization of the wave equation, and to [5] and [17] for the behavior of controls under numerical approximations. We also refer to [16] for a discussion and comparison of these two topics. Similar methods have also been used in [10] to analyze the partial controllability of a model for spherical shells.

**2. Asymptotic limit of the homogeneous system.** For the sake of completeness, in this section we study the asymptotic limit of the solutions of the homogeneous system (1.6) as  $k$  tends to infinity. Before, we mention some elementary properties of these solutions.

System (1.6) is well-posed in the energy space  $\mathcal{X} = H_0^1(0, L) \times L^2(0, L) \times H^1(0, L) \times L^2(0, L)$ . More precisely, for any  $\{\phi_0, \phi_1, \psi_0, \psi_1\} \in \mathcal{X}$  and  $\{f, g\} \in L^1(0, T; [L^2(0, L)]^2)$  there exists a unique solution in the class

$$(2.1) \quad \{\phi, \psi\} \in C^0([0, T]; H_0^1(0, L) \times H^1(0, L)) \cap C^1([0, T]; [L^2(0, L)]^2)$$

satisfying the inequality

$$(2.2) \quad \|\{\phi(t), \phi'(t), \psi(t), \psi'(t)\}\|_k \leq C_1 e^{C_2 T} \left( \|\{\phi_0, \phi_1, \psi_0, \psi_1\}\|_k + \|\{f, g\}\|_{L^1(0, T; [L^2(0, L)]^2)} \right)$$

for all  $t \in [0, T]$ , where the norm  $\|\cdot\|_k$  is defined by

$$\begin{aligned} \|\{u_1, u_2, v_1, v_2\}\|_k^2 &= \int_0^L |u_{1x}|^2 dx + \frac{\rho h^3}{12} \int_0^L |u_2|^2 dx + k \int_0^L |u_1 + v_{1x}|^2 dx \\ &\quad + \int_0^L |v_1|^2 dx + \rho h \int_0^L |v_2|^2 dx. \end{aligned}$$

On the other hand, the energy  $E_k(t)$  of the system

$$(2.3) \quad \begin{aligned} E_k(t) &= \frac{1}{2} \int_0^L \left\{ \frac{\rho h^3}{12} |\phi'(x, t)|^2 + \rho h |\psi'(x, t)|^2 + |\phi_x(x, t)|^2 \right. \\ &\quad \left. + k |\phi(x, t) + \psi_x(x, t)|^2 \right\} dx \end{aligned}$$

satisfies

$$(2.4) \quad \frac{dE_k}{dt}(t) = \int_0^L [f(x, t) \phi'(x, t) + g(x, t) \psi'(x, t)] dx.$$

In particular, when  $f \equiv g \equiv 0$ , the energy  $E_k$  is conserved along time.

Estimate (2.2) holds as a consequence of this energy identity and Gronwall's inequality because of the obvious relation between the energy  $E_k$  and the norm  $\|\cdot\|_k$ . The norm  $\|\cdot\|_k$  is equivalent to the square root of the sum  $E_k + \|v_1\|_{L^2(0,L)}^2$ . Note that the canonical norm in  $\mathcal{X}$  can be bounded above uniformly in terms of the norm  $\|\cdot\|_k$  for all  $k \geq 1$ ; i.e., there exists  $C > 0$  independent of  $k$  such that

$$(2.5) \quad \|\cdot\|_{\mathcal{X}} \leq C \|\cdot\|_k \quad \forall k \geq 1,$$

where  $\|\cdot\|_{\mathcal{X}}$  stands for the canonical norm in  $\mathcal{X}$ .

Let us note that the energy  $E_k$  does not define a norm in  $\mathcal{X}$ . Accordingly, it is natural to introduce the norm  $\|\cdot\|_k$  and the following decomposition of the energy space:  $\mathcal{X} = \mathcal{X}_0 \oplus \mathcal{X}_1$ , with

$$\mathcal{X}_0 = H_0^1(0, L) \times L^2(0, L) \times V \times H \quad \text{and} \quad \mathcal{X}_1 = \{ \{0, 0, c_1, c_2\} \in \mathcal{X}; c_i \in \mathbb{R}, i = 1, 2 \},$$

where

$$V = H^1(0, L) \cap H \quad \text{and} \quad H = \left\{ v \in L^2(0, L); \int_0^L v(x) dx = 0 \right\}.$$

In  $\mathcal{X}_0$  the energy defines a norm which is equivalent to  $\|\cdot\|_k$ . On the other hand, the spaces  $\mathcal{X}_0$  and  $\mathcal{X}_1$  are invariant under the flow generated by system (1.6) in the sense given in the following result.

**PROPOSITION 2.1.** *Given data  $\{\phi_0, \phi_1, \psi_0, \psi_1\}$  and  $\{0, f, 0, g\}$  belonging to  $\mathcal{X}_i$  and  $L^1(0, T; \mathcal{X}_i)$ , respectively, with  $i = 0, 1$ , the associated solution belongs to  $\mathcal{X}_i$  for all  $t \in [0, T]$ .*

*Proof.* First, we show that if  $\{\phi_0, \phi_1, \psi_0, \psi_1\} \in \mathcal{X}_0$  and  $\{0, f, 0, g\} \in L^1(0, T; \mathcal{X}_0)$ , then the corresponding solution belongs to  $\mathcal{X}_0$  for all  $t \in [0, T]$ . In fact, integrating  $(1.6)_2$  ( $(1.6)_2$  means the second equation in (1.6)) on  $]0, L[$  and using the conditions  $(1.6)_3$ , we have

$$\frac{d^2}{dt^2} \int_0^L \psi(x, t) dx = 0,$$

that is,

$$\int_0^L \psi(x, t) dx = \int_0^L \psi_0(x) dx + t \int_0^L \psi_1(x) dx = 0.$$

Now, if  $\{\phi_0, \phi_1, \psi_0, \psi_1\} \in \mathcal{X}_1$  and  $\{0, f, 0, g\} \in L^1(0, T; \mathcal{X}_1)$ , we have  $\phi_0 = \phi_1 = f = 0$ ,  $\psi_0 = c_1$ ,  $\psi_1 = c_2$ , and  $g = g(t) \in L^1(0, T)$ . In this way, we get, directly from the system (1.6), that  $\phi \equiv \psi_x \equiv 0$ . So  $\psi = \psi(t)$  and, by  $(1.6)_2$ , we obtain

$$\rho h \psi''(t) = g(t) \quad \forall t \in [0, T]$$

and, therefore,

$$\psi(t) = \int_0^t w(s) ds + c_2 t + c_1 \quad \forall t \in [0, T],$$

where  $w(\sigma) = (1/\rho h) \int_0^\sigma g(\xi) d\xi$ .  $\square$

We also have the following “hidden regularity” result.

PROPOSITION 2.2. *For any  $T > 0$ , there exists a constant  $C = C(T) > 0$ , independent of  $k$ , such that the solution  $\{\phi, \psi\}$  of (1.6) satisfies the inequality*

$$(2.6) \quad \|\{\phi_x(0, \cdot), \psi(0, \cdot)\}\|_{L^2(0,T) \times H^1(0,T)}^2 \leq C \left\{ \|\{\phi_0, \phi_1, \psi_0, \psi_1\}\|_k^2 + \|\{f, g\}\|_{L^1(0,T;[L^2(0,L)]^2)}^2 \right\}$$

for any  $\{\phi_0, \phi_1, \psi_0, \psi_1\} \in \mathcal{X}$  and  $\{f, g\} \in L^1(0, T; [L^2(0, L)]^2)$ .

*Proof.* It is enough to consider smooth solutions since a classical density argument allows us to extend the inequality (2.6) to finite-energy solutions. We use a multiplier method (see [8]). We multiply (1.6)<sub>1</sub> by  $(L - x)\phi_x$  and (1.6)<sub>2</sub> by  $(L - x)\psi_x$ , and after integrating by parts over  $Q$  we get

$$(2.7) \quad \begin{aligned} & \frac{L}{2} \int_0^T \left\{ \rho h |\psi'(0, t)|^2 + |\phi_x(0, t)|^2 \right\} dt = \frac{1}{2} \int_Q \left\{ \frac{\rho h^3}{12} |\phi'(x, t)|^2 + \rho h |\psi'(x, t)|^2 \right. \\ & \quad \left. + |\phi_x(x, t)|^2 + k |\phi(x, t) + \psi_x(x, t)|^2 \right\} dx dt - \int_Q \left\{ \frac{\rho h^3}{12} |\phi'(x, t)|^2 - |\phi_x(x, t)|^2 \right\} dx dt \\ & \quad - \left[ \frac{\rho h^3}{12} \int_0^L \phi'(x, t) (L - x) \phi_x(x, t) dx + \rho h \int_0^L \psi'(x, t) (L - x) \psi_x(x, t) dx \right] \Big|_0^T \\ & \quad + \int_Q f(x, t) (L - x) \phi_x(x, t) dx dt + \int_Q g(x, t) (L - x) \psi_x(x, t) dx dt. \end{aligned}$$

Using (2.2), from (2.7) we obtain the estimate

$$(2.8) \quad \|\{\phi_x(0, \cdot), \psi'(0, \cdot)\}\|_{[L^2(0,T)]^2}^2 \leq C \left\{ \|\{\phi_0, \phi_1, \psi_0, \psi_1\}\|_k^2 + \|\{f, g\}\|_{L^1(0,T;[L^2(0,L)]^2)}^2 \right\},$$

with  $C = C(T) > 0$  a constant independent of  $k$ .

On the other hand, by the trace theorem, there exists a constant  $C_\gamma > 0$ , independent of  $k$ , such that

$$\|\psi(0, \cdot)\|_{L^2(0,T)} \leq C_\gamma \|\psi(\cdot, t)\|_{H^1(Q)}.$$

Thus by (2.2) it follows that

$$(2.9) \quad \|\psi(0, \cdot)\|_{L^2(0,T)} \leq C \left( \|\{\phi_0, \phi_1, \psi_0, \psi_1\}\|_k + \|\{f, g\}\|_{L^1(0,T;[L^2(0,L)]^2)} \right),$$

where  $C = C(T) > 0$  is a constant independent of  $k$ .

Combining (2.8) and (2.9), we deduce the inequality (2.6), uniformly on  $k \geq 1$ .  $\square$

Concerning the asymptotic behavior of the solutions of the homogeneous Mindlin–Timoshenko system (1.6), as  $k$  tends to infinity, the following result holds.

THEOREM 2.1. *Let  $\{\phi_k, \psi_k\}$  be the unique solution of (1.6) with data  $\{\phi_{0k}, \phi_{1k}, \psi_{0k}, \psi_{1k}\} \in \mathcal{X}$  and  $\{f, g\} \in L^1(0, T; H_0^1(0, L) \times L^2(0, L))$ .*

(a) *Weak convergence. Assume that the initial data satisfy*

$$(2.10) \quad \|\{\phi_{0k}, \phi_{1k}, \psi_{0k}, \psi_{1k}\}\|_k^2 \leq C \quad \forall k \geq 1,$$

with  $C$  being a positive constant independent of  $k$  and

$$(2.11) \quad \{\phi_{0k}, \phi_{1k}, \psi_{0k}, \psi_{1k}\} \rightarrow \{\phi_0, \phi_1, \psi_0, \psi_1\} \text{ weakly in } \mathcal{X}.$$

Then, as  $k \rightarrow \infty$ , the following convergence property holds:

$$(2.12) \quad \{\phi_k, \phi'_k, \psi_k, \psi'_k\} \rightarrow \{-\psi_x, -\psi'_x, \psi, \psi'\} \text{ weakly* in } L^\infty(0, T; \mathcal{X}),$$

where  $\psi$  solves the homogeneous Kirchhoff system

$$(2.13) \quad \begin{cases} \rho h \psi'' - \frac{\rho h^3}{12} \psi''_{xx} + \psi_{xxxx} = f_x + g & \text{in } Q, \\ \psi_x(0, \cdot) = \psi_x(L, \cdot) = \psi_{xxx}(0, \cdot) = \psi_{xxx}(L, \cdot) = 0 & \text{on } (0, T), \\ \psi(\cdot, 0) = \psi_0, \quad \left[ \psi(\cdot, 0) - \frac{h^2}{12} \psi_{xx}(\cdot, 0) \right]' = \psi_1 + \frac{h^2}{12} \phi_{1x} & \text{in } (0, L). \end{cases}$$

(b) Strong convergence. If the initial data satisfy the additional conditions

$$(2.14) \quad \phi_0 + \psi_{0x} = 0, \quad \lim_{k \rightarrow \infty} E_k(0) = \mathcal{E}(0),$$

where  $\mathcal{E}(t)$  is the energy of (2.13) given by

$$(2.15) \quad \mathcal{E}(t) = \frac{1}{2} \int_0^L \left\{ \rho h |\psi'(x, t)|^2 + \frac{\rho h^3}{12} |\psi'_x(x, t)|^2 + |\psi_{xx}(x, t)|^2 \right\} dx,$$

then, for  $1 < p < \infty$ , the following strong convergence holds as  $k \rightarrow \infty$ :

$$(2.16) \quad \{\phi_k, \phi'_k, \psi_k, \psi'_k\} \rightarrow \{-\psi_x, -\psi'_x, \psi, \psi'\} \text{ strongly in } L^p(0, T; \mathcal{X}).$$

*Remark 2.1.*

- The existence and uniqueness of weak solutions of the limit system (2.13) can be obtained by classical methods. More precisely, when  $\{\psi_0, \psi_1, f_x + g\} \in W \times H^1(0, L) \times L^1(0, T; L^2(0, L))$ , where  $W = \{v \in H^2(0, L); v_x(0) = v_x(L) = 0\}$ , there exists a unique finite energy solution  $\psi$  in the class

$$\psi \in C^0([0, T]; W) \cap C^1([0, T]; H^1(0, L))$$

satisfying the variational formulation of (2.13)

$$\rho h \frac{d}{dt} (\psi'(t), w) + \frac{\rho h^3}{12} \frac{d}{dt} (\psi'_x(t), w_x) + (\psi_{xx}(t), w_{xx}) = (f_x(t) + g(t), w)$$

for all  $w \in W$ , the boundary conditions (2.13)<sub>2</sub>, and the initial conditions (2.13)<sub>3</sub>. Here  $(\cdot, \cdot)$  represents the inner product in  $L^2(0, L)$ . Furthermore the energy  $\mathcal{E}(t)$  in (2.15) satisfies

$$\mathcal{E}'(t) = \int_0^L [f_x(x, t) + g(x, t)] \psi'(x, t) dx.$$

If  $f_x + g \equiv 0$ , the energy is conserved.



- Note, however, that, in order to identify fully the initial data of the solutions of the limit system (2.13) and, more precisely, to determine the initial data of  $\psi'$ , an elliptic equation has to be solved. Namely, the initial datum for the velocity  $\psi'$  in (2.13)<sub>3</sub> is determined by solving the elliptic equation

$$(2.17) \quad \psi'(\cdot, 0) \in H^1(0, L) : \quad \rho h \psi'(0) - \frac{\rho h^3}{12} \psi'_{xx}(0) = \rho h \psi_1 + \frac{\rho h^3}{12} \phi_{1x},$$

as the proof of the theorem will show.

To be more precise, this elliptic equation can be written in the variational form

$$(2.18) \quad -\frac{\rho h^3}{12} (\psi'_x(0), w_x) - \rho h (\psi'(0), w) = \frac{\rho h^3}{12} (\phi_1, w_x) - \rho h (\psi_1, w) \quad \forall w \in H^1(0, L)$$

in which the term  $\phi_{1x}$ , which is an element of  $(H^1(0, L))'$ , is not the derivative of  $\phi_1$  in the sense of transposition but rather the linear mapping so that, when acting on any element  $w$  of  $H^1(0, L)$ , yields the value  $-(\phi_1, w_x)$ . The same can be said about  $\psi'_{xx}(0)$ , which represents the element of  $(H^1(0, L))'$  yielding  $-(\psi'_x(0), w_x)$ .

- Similar results hold when the right-hand side terms  $\{f_k, g_k\}$  depend on  $k$  and converge in a suitable sense. But we shall not discuss this issue since it is not needed for the purpose of this paper.

*Proof of Theorem 2.1.* We will prove the theorem in two steps.

*Step 1. Weak convergence.* Considering  $\{f, g\} \in L^2(0, T; H_0^1(0, L) \times L^2(0, L))$  and the sequence of initial data  $\{\phi_{0k}, \phi_{1k}, \psi_{0k}, \psi_{1k}\} \in \mathcal{X}$  satisfying (2.10), the right-hand side of (2.2) and the energies  $E_k$  are uniformly bounded on  $k$ . Consequently

$$\left| \begin{array}{l} (\{\phi_k, \phi'_k, \psi_k, \psi'_k\}) \text{ is bounded in } L^\infty(0, T; \mathcal{X}), \\ (\sqrt{k}[\phi_k + \psi_{kx}]) \text{ is bounded in } L^\infty(0, T; L^2(0, L)). \end{array} \right|$$

This immediately yields a uniform bound for  $\phi$ ,  $\phi'$ , and  $\psi'$  in the corresponding spaces. We also get a uniform bound on  $\psi$  in  $L^2(0, L)$ . The uniform bound on  $\psi$  in  $H^1(0, L)$  can be easily obtained from the bound in  $L^2(0, L)$  and in  $\|\cdot\|_k$  and the fact that

$$(2.19) \quad \begin{aligned} \|\psi_x\|_{L^2(0, L)} &\leq \|\phi + \psi_x\|_{L^2(0, L)} + \|\phi\|_{L^2(0, L)} \leq k \|\phi + \psi_x\|_{L^2(0, L)} + \|\phi\|_{L^2(0, L)} \\ &\leq C \|\{\phi, \phi', \psi, \psi'\}\|_k. \end{aligned}$$

Extracting subsequences, which we still denote by  $(\{\phi_k, \psi_k\})$ , we get

$$(2.20) \quad \{\phi_k, \phi'_k, \psi_k, \psi'_k\} \rightarrow \{\phi, \phi', \psi, \psi'\} \text{ weakly } * \text{ in } L^\infty(0, T; \mathcal{X})$$

with

$$(2.21) \quad \phi + \psi_x = 0.$$

For test functions  $\{z, w\} \in H_0^1(0, L) \times H^1(0, L)$  satisfying

$$(2.22) \quad z + w_x = 0,$$

the variational formulation of (1.6) reduces to

$$(2.23) \quad \frac{\rho h^3}{12} \frac{d}{dt} (\phi'_k(t), z) + \rho h \frac{d}{dt} (\psi'_k(t), w) + (\phi_{kx}(t), z_x) = (f(t), z) + (g(t), w).$$

Using the convergences (2.20) in (2.23) and applying identities (2.21) and (2.22), the limit weak formulation can be written in terms of  $\psi$  as follows:

$$(2.24) \quad \rho h \frac{d}{dt} (\psi'(t), w) + \frac{\rho h^3}{12} \frac{d}{dt} (\psi'_x(t), w_x) + (\psi_{xx}(t), w_{xx}) = (f_x(t) + g(t), w) \quad \forall w \in W.$$

This identity is a weak form of (2.13)<sub>1</sub>. The two boundary conditions

$$\psi_x(0, t) = \psi_x(L, t) = 0 \quad \text{on} \quad (0, T)$$

are deduced from the facts that  $\psi_x = -\phi$  and that  $\phi$  satisfies the Dirichlet boundary conditions. The other two in (2.13), namely,

$$\psi_{xxx}(0, t) = \psi_{xxx}(L, t) = 0 \quad \text{on} \quad (0, T),$$

are implicit in the weak form of the equation since the test function  $w$  does not vanish on the boundary.

To conclude our result, it remains to identify the initial data of the limit system. In view of the convergences (2.20), and classical compactness arguments,  $\psi_k \rightarrow \psi$  in  $C^0([0, T]; L^2(0, L))$ . Then  $\psi_k(\cdot, 0) \rightarrow \psi(\cdot, 0)$  in  $L^2(0, L)$ , which, combined with (2.11), guarantees that  $\psi(\cdot, 0) = \psi_0$ . In order to identify  $\psi'(\cdot, 0)$ , we multiply both sides of (2.23) by the function  $\theta_\delta \in H^1(0, T)$  defined by

$$\theta_\delta(t) = \begin{cases} -\frac{t}{\delta} + 1 & \text{if } 0 \leq t \leq \delta, \\ 0 & \text{if } \delta < t \leq T \end{cases}$$

and we integrate by parts to obtain

$$\begin{aligned} & -\frac{\rho h^3}{12} (\phi'_k(0), z) + \frac{\rho h^3}{12\delta} \int_0^\delta (\phi'_k(t), z) dt - \rho h (\psi'_k(0), w) + \frac{\rho h}{\delta} \int_0^\delta (\psi'_k(t), w) dt \\ & + \int_0^\delta (\phi_{kx}(t), z_x) \theta_\delta(t) dt = -\frac{\rho h^3}{12} (\phi_{1k}, z) - \rho h (\psi_{1k}, w) + \frac{\rho h^3}{12\delta} \int_0^\delta (\phi'_k(t), z) dt \\ & + \frac{\rho h}{\delta} \int_0^\delta (\psi'_k(t), w) dt + \int_0^\delta (\phi_{kx}(t), z_x) \theta_\delta(t) dt = \int_0^\delta (f(t), z) \theta_\delta(t) dt \\ & + \int_0^\delta (g(t), w) \theta_\delta(t) dt. \end{aligned}$$

Passing to the limit in the last equality as  $k \rightarrow \infty$  and using (2.21) and (2.22), we get

$$\begin{aligned} & \frac{\rho h^3}{12} (\phi_1, w_x) - \rho h (\psi_1, w) + \frac{\rho h^3}{12\delta} \int_0^\delta (\psi'_x(t), w_x) dt + \frac{\rho h}{\delta} \int_0^\delta (\psi'(t), w) dt \\ & + \int_0^\delta (\psi_{xx}(t), w_{xx}) \theta_\delta(t) dt = \int_0^\delta (f_x(t), w) \theta_\delta(t) dt + \int_0^\delta (g(t), w) \theta_\delta(t) dt. \end{aligned}$$

On the other hand, multiplying in (2.24) by  $\theta_\delta$  and integrating in time, we obtain an expression that, compared with the previous one, yields the identity (2.18). This completes the proof of part (a) of the theorem.

*Step 2. Strong convergence.* We know by (2.4) that the energy  $E_k(t)$  associated with  $\{\phi_k, \psi_k\}$  of (1.6) satisfies

$$(2.25) \quad E_k(t) = E_k(0) + \int_0^t \int_0^L [f(x, s) \phi'_k(x, s) + g(x, s) \psi'_k(x, s)] dx ds.$$

On the other hand, in view of Remark 2.1 and (2.21), it follows that the energy of system (2.13) satisfies

$$\mathcal{E}(t) = \mathcal{E}(0) + \int_0^t \int_0^L [-f(x, s) \psi'_x(x, s) + g(x, s) \psi'(x, s)] dx dt.$$

Therefore, combining (2.14), (2.20), (2.21), and (2.25), we get

$$(2.26) \quad \lim_{k \rightarrow \infty} E_k(t) = \mathcal{E}(t).$$

As a consequence of (2.26), we have the norm convergence which, together with the weak convergence result (2.12), yields the strong convergence one, (b), of the theorem.

Let us develop this last argument in more detail. In view of the weak convergence of solutions and the structure of the energy  $E_k$ , it follows that

$$\liminf_{k \rightarrow \infty} \int_0^T E_k(t) dt \geq \frac{1}{2} \int_0^T \int_0^L \left[ \frac{\rho h^3}{12} |\psi'_x|^2 + \rho h |\psi'|^2 + |\psi_{xx}|^2 \right] dx dt = \int_0^T \mathcal{E}(t) dt.$$

This fact, together with (2.20), (2.21), and (2.26), implies

$$\begin{aligned} \lim_{k \rightarrow \infty} \int_0^T \int_0^L |\phi'_k|^2 dx dt &= \int_0^T \int_0^L |\psi'_x|^2 dx dt, \\ \lim_{k \rightarrow \infty} \int_0^T \int_0^L |\psi'_k|^2 dx dt &= \int_0^T \int_0^L |\psi'|^2 dx dt, \\ \lim_{k \rightarrow \infty} \int_0^T \int_0^L |\phi_{kx}|^2 dx dt &= \int_0^T \int_0^L |\psi_{xx}|^2 dx dt, \end{aligned}$$

and

$$(2.27) \quad \lim_{k \rightarrow \infty} k \int_0^T \int_0^L |\phi_k + \psi_{kx}|^2 dx dt \rightarrow 0.$$

This, combined with the weak convergence, implies the strong convergence of  $\{\phi'_k, \psi'_k, \phi_{kx}\}$  to  $\{-\psi'_x, \psi', -\psi_{xx}\}$  in  $[L^2(Q)]^3$ . The strong convergence of  $\psi_k$  in  $L^2(0, T; H^1(0, L))$  is then a consequence of (2.27) and the fact that  $\phi_k$  strongly converges to  $-\psi_x$  in  $L^2(Q)$ .

Strong convergence in  $L^2(0, T; \mathcal{X})$ , together with the uniform boundedness in  $L^\infty(0, T; \mathcal{X})$ , implies strong convergence in  $L^p(0, T; \mathcal{X})$  for all  $1 < p < \infty$ .  $\square$

**3. Spectral analysis.** This section is devoted to analyzing the asymptotic behavior, as  $k$  tends to infinity, of the spectrum of the Mindlin–Timoshenko system. With this goal in mind, we write system (1.6) (with  $f = g = 0$ ) in the following abstract form:

$$\Phi' = -i\mathcal{A}\Phi,$$

where  $\Phi = [\phi, \phi', \psi, \psi']^T$  and the operator  $\mathcal{A} : D(\mathcal{A}) \subset \mathcal{X} \rightarrow \mathcal{X}$  is given by

$$\mathcal{A} = i \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{12}{\rho h^3} \left( \frac{\partial^2}{\partial x^2} - k \right) & 0 & -\frac{12k}{\rho h^3} \frac{\partial}{\partial x} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{k}{\rho h} \frac{\partial}{\partial x} & 0 & \frac{k}{\rho h} \frac{\partial^2}{\partial x^2} & 0 \end{bmatrix}$$

with domain

$$D(\mathcal{A}) = [H_0^1(0, L) \cap H^2(0, L)] \times H_0^1(0, L) \times W \times H^1(0, L).$$

The eigenvalue problem for the operator  $\mathcal{A}$  reads

$$(3.1) \quad \mathcal{A}\Phi = \lambda\Phi.$$

Let us compute the eigenvalues and the corresponding eigenfunctions. In view of the various equations involved in (3.1) and the boundary conditions satisfied by the components  $\phi$  and  $\psi$ , the solutions  $\Phi = [\phi, \phi', \psi, \psi']^T$  associated with the eigenfunctions are such that

$$\{\phi(x, t), \psi(x, t)\} = e^{-i\lambda t} \{\sin(m\pi x/L), c \cos(m\pi x/L)\},$$

where the constant  $c$  is to be determined in terms of  $m$  and  $\lambda$ . In particular, computing the components  $\phi$  and  $\psi$  suffices to identify the 4-component vector.

From (3.1) we have

$$(3.2) \quad \begin{cases} \lambda^2 \frac{\rho h^3}{12} \phi - \phi_{xx} + k(\phi + \psi_x) = 0, \\ \lambda^2 \rho h \psi - k(\phi + \psi_x)_x = 0. \end{cases}$$

Taking the derivative of (3.2)<sub>1</sub> with respect to  $x$  and substituting in (3.2)<sub>2</sub>, we get

$$(3.3) \quad \psi = \frac{1}{\lambda^2 \rho h} \left( \phi_{xxx} - \frac{\lambda^2 \rho h^3}{12} \phi_x \right).$$

Now, doing the same in (3.3) and substituting in (3.2)<sub>1</sub>, it follows that

$$\phi_{xxxx} - \left( \frac{\rho h \lambda^2}{k} + \frac{\lambda^2 \rho h^3}{12} \right) \phi_{xx} + \left( \frac{\lambda^4 \rho^2 h^4}{12k} + \lambda^2 \rho h \right) \phi = 0.$$

Since  $\phi(x, t) = e^{-i\lambda t} \sin(m\pi x/L)$ , we obtain, for  $\lambda$ , the fourth degree equation

$$(3.4) \quad \lambda^4 - \left( \frac{12\pi^2 m^2}{\rho h^3 L^2} + \frac{\pi^2 k m^2}{\rho h L^2} + \frac{12k}{\rho h^3} \right) \lambda^2 + \frac{12\pi^4 k m^4}{\rho^2 h^4 L^4} = 0,$$

while  $c$  satisfies

$$(3.5) \quad c = \frac{\pi^3 m^3}{\lambda^2 \rho h L^3} - \frac{h^2 m \pi}{12L}.$$

Solving (3.4), we find the eigenvalues

$$\begin{aligned} \tilde{\lambda}_{k,m}^{\pm} = \pm & \left[ \frac{6\pi^2 m^2}{\rho h^3 L^2} + \frac{\pi^2 k m^2}{2\rho h L^2} + \frac{6k}{\rho h^3} \right. \\ & \left. + \frac{1}{2} \sqrt{\frac{144k^2}{\rho^2 h^6} + \frac{288\pi^2 k m^2}{\rho^2 h^6 L^2} + \frac{24\pi^2 k^2 m^2}{\rho^2 h^4 L^2} + \left( \frac{12\pi^2 m^2}{\rho h^3 L^2} - \frac{\pi^2 k m^2}{\rho h L^2} \right)^2} \right]^{\frac{1}{2}} \end{aligned}$$

and

$$\lambda_{k,m}^{\pm} = \pm \left[ \frac{6\pi^2 m^2}{\rho h^3 L^2} + \frac{\pi^2 k m^2}{2\rho h L^2} + \frac{6k}{\rho h^3} - \frac{1}{2} \sqrt{\frac{144k^2}{\rho^2 h^6} + \frac{288\pi^2 k m^2}{\rho^2 h^6 L^2} + \frac{24\pi^2 k^2 m^2}{\rho^2 h^4 L^2} + \left( \frac{12\pi^2 m^2}{\rho h^3 L^2} - \frac{\pi^2 k m^2}{\rho h L^2} \right)^2} \right]^{\frac{1}{2}}.$$

We denote by  $c_m$  and  $\tilde{c}_m$  the corresponding values of  $c$  according to the definition (3.5).

For  $m$  fixed, we see easily that, as  $k$  tends to infinity,

$$(3.6) \quad \tilde{\lambda}_{k,m}^{\pm} \rightarrow \pm\infty.$$

This corresponds to that half of the spectrum that disappears when letting  $k$  tend to infinity, in the sense that, since  $\tilde{\lambda}_{k,m}^{\pm}$  diverge as  $k \rightarrow \infty$ , do not lead to any eigenvalue of the limit system.

The following result describes the asymptotic behavior of the other family of eigenvalues.

PROPOSITION 3.1. *For fixed  $m \in \mathbb{N}$ , as  $k \rightarrow \infty$ ,*

$$(3.7) \quad \lambda_{k,m}^{\pm} \rightarrow \lambda_m^{\pm} = \pm \sqrt{\frac{12\pi^4 m^4}{12\rho h L^4 + \pi^2 \rho h^3 L^2 m^2}}.$$

*These are the eigenvalues of the limit Kirchhoff system (2.13) (with  $f_x + g = 0$ ) for which the corresponding eigenfunctions are  $\cos(m\pi x/L)$ .*

*Proof.* It is sufficient to prove convergence for the  $+$  sign. To simplify the notation we denote by  $\lambda_{k,m}$  the eigenvalues  $\lambda_{k,m}^+$ . We have to observe that

$$(3.8) \quad \left| \lambda_{k,m} - \sqrt{\frac{12\pi^4 m^4}{12\rho h L^4 + \pi^2 \rho h^3 L^2 m^2}} \right| = \left| \frac{\lambda_{k,m}^2 - \frac{12\pi^4 m^4}{12\rho h L^4 + \pi^2 \rho h^3 L^2 m^2}}{\lambda_{k,m} + \sqrt{\frac{12\pi^4 m^4}{12\rho h L^4 + \pi^2 \rho h^3 L^2 m^2}}} \right| \\ = \left| \frac{(12\rho h L^4 + \pi^2 \rho h^3 L^2 m^2) \lambda_{k,m}^2 - 12\pi^4 m^4}{(12\rho h L^4 + \pi^2 \rho h^3 L^2 m^2) \lambda_{k,m} + \sqrt{(12\rho h L^4 + \pi^2 \rho h^3 L^2 m^2) (12\pi^4 m^4)}} \right|.$$

Let us now analyze separately the numerator and denominator of this expression.

Using the algebraic identity  $a - b = (a^2 - b^2) / (a + b)$ , we get

$$(3.9) \quad |\lambda_{k,m}|^2 = \frac{12\pi^4 k m^4}{\rho^2 h^4 L^4} \left[ \frac{6\pi^2 m^2}{\rho h^3 L^2} + \frac{\pi^2 k m^2}{2\rho h L^2} + \frac{6k}{\rho h^3} \right. \\ \left. + \frac{1}{2} \sqrt{\frac{144k^2}{\rho^2 h^6} + \frac{288\pi^2 k m^2}{\rho^2 h^6 L^2} + \frac{24\pi^2 k^2 m^2}{\rho^2 h^4 L^2} + \frac{144\pi^4 m^4}{\rho^2 h^6 L^4} - \frac{24\pi^4 k m^4}{\rho^2 h^4 L^4} + \frac{\pi^4 k^2 m^4}{\rho^2 h^2 L^4}} \right]^{-1}.$$

Then the numerator  $\mathcal{N}$  on the right-hand side of (3.8) can be rewritten as

$$\mathcal{N} = \left| -12\pi^4 m^4 + (12\rho h L^4 + \pi^2 \rho h^3 L^2 m^2) \frac{12\pi^4 k m^4}{\rho^2 h^4 L^4} \left[ \frac{6\pi^2 m^2}{\rho h^3 L^2} + \frac{\pi^2 k m^2}{2\rho h L^2} + \frac{6k}{\rho h^3} \right. \right. \\ \left. \left. + \frac{1}{2} \sqrt{\frac{144k^2}{\rho^2 h^6} + \frac{288\pi^2 k m^2}{\rho^2 h^6 L^2} + \frac{24\pi^2 k^2 m^2}{\rho^2 h^4 L^2} + \frac{144\pi^4 m^4}{\rho^2 h^6 L^4} - \frac{24\pi^4 k m^4}{\rho^2 h^4 L^4} + \frac{\pi^4 k^2 m^4}{\rho^2 h^2 L^4}} \right]^{-1} \right|,$$

that is,

$$\mathcal{N} = \left| -12\pi^4 m^4 + \frac{(12L^2 + \pi^2 h^2 m^2) 12\pi^4 m^4}{\frac{6\pi^2 m^2}{k} + \frac{\pi^2 h^2 m^2}{2} + 6L^2 + \frac{\rho h^3 L^2}{2} \sqrt{r}} \right|,$$

where

$$r = \frac{144}{\rho^2 h^6} + \frac{288\pi^2 m^2}{\rho^2 h^6 L^2 k} + \frac{24\pi^2 m^2}{\rho^2 h^4 L^2} + \frac{144\pi^4 m^4}{\rho^2 h^6 L^4 k^2} - \frac{24\pi^4 m^4}{\rho^2 h^4 L^4 k} + \frac{\pi^4 m^4}{\rho^2 h^2 L^4}.$$

Thus, we have the following estimate:

$$\begin{aligned} \mathcal{N} &= \left| \frac{72\pi^4 L^2 m^4 + 6\pi^6 h^2 m^6 - \frac{72\pi^6 m^6}{k} - 6\pi^4 \rho h^3 L^2 m^4 \sqrt{r}}{\frac{6\pi^2 m^2}{k} + \frac{\pi^2 h^2 m^2}{2} + 6L^2 + \frac{\rho h^3 L^2}{2} \sqrt{r}} \right| \\ &\leq \left| \left( \frac{144\pi^2 L^2 m^2}{h^2} + 12\pi^4 m^4 - \frac{144\pi^4 m^4}{h^2 k} \right) - 12\pi^2 \rho h L^2 m^2 \sqrt{r} \right| \\ (3.10) \quad &= \left| \frac{\left( \frac{144\pi^2 L^2 m^2}{h^2} + 12\pi^4 m^4 - \frac{144\pi^4 m^4}{h^2 k} \right)^2 - (12\pi^2 \rho h L^2 m^2)^2 r}{\frac{144\pi^2 L^2 m^2}{h^2} + 12\pi^4 m^4 - \frac{144\pi^4 m^4}{h^2 k} + 12\pi^2 \rho h L^2 m^2 \sqrt{r}} \right| \\ &= \left| \frac{4(12)^3 \pi^4 L^2 m^4}{12h^2 L^2 k + \pi^2 h^4 m^2 k - 12\pi^2 h^2 m^2 + \rho h^5 L^2 k \sqrt{r}} \right|. \end{aligned}$$

Dividing the numerator of the last fraction of (3.10) by the second term of its denominator, it follows that

$$(3.11) \quad \mathcal{N} \leq \frac{4(12)^3 \pi^2 L^2 m^2}{h^4 k}.$$

On the other hand, the denominator  $\mathbf{D}$  of the last term of (3.8) is bounded below by

$$(3.12) \quad \mathbf{D} \geq 12\sqrt{\rho h \pi^2 L^2 m^2}.$$

From (3.8), (3.11), and (3.12) we obtain that

$$(3.13) \quad \left| \lambda_{k,m} - \sqrt{\frac{12\pi^4 m^4}{12\rho h L^4 + \pi^2 \rho h^3 L^2 m^2}} \right| \leq \frac{c}{k},$$

where  $c = 4(12)^2/h^4\sqrt{\rho h}$ . The estimate (3.13) immediately implies the statement (3.7) of the proposition.  $\square$

*Remark 3.1.* As we mentioned above, the eigenvalues  $\tilde{\lambda}_{k,m}^\pm$  tend to  $\pm\infty$ . In other words, they disappear as  $k$  tends to infinity. This fact is intimately related to the asymptotic simplification that the system undergoes when passing from a system of two equations and two dependent variables to a scalar equation with only one dependent variable. Obviously, for a complete description of the space of solutions of (2.13), the eigenpairs  $(\lambda_m^\pm, \cos(m\pi x/L))$  obtained in the limit as  $k$  tends to infinity suffice.

**4. Asymptotic limit of the controlled system.** Our interest now is to study the asymptotic behavior of the solutions  $\{u_k, v_k\}$  of the system (1.1)–(1.4) when  $k$  tends to infinity. They are defined by transposition (see [9]) as follows. First, we consider the solution of the adjoint system

$$(4.1) \quad \begin{cases} \frac{\rho h^3}{12} \phi'' - \phi_{xx} + k(\phi + \psi_x) = f & \text{in } Q, \\ \rho h \psi'' - k(\phi + \psi_x)_x = g & \text{in } Q, \\ \phi(0, \cdot) = \phi(L, \cdot) = \psi_x(0, \cdot) = \psi_x(L, \cdot) = 0 & \text{on } (0, T), \\ \phi(\cdot, T) = \phi'(\cdot, T) = \psi(\cdot, T) = \psi'(\cdot, T) = 0 & \text{in } (0, L). \end{cases}$$

As indicated in the introduction, although, normally, the adjoint system is taken to be homogeneous (i.e.,  $f \equiv g \equiv 0$ ), we consider the case where  $f$  and  $g$  are arbitrary since this is useful to define the solution of (1.1)–(1.4) by transposition.

This system may be reduced to (1.6) by the change of variables  $t \rightarrow T - t$ . Then, when  $\{f, g\} \in L^1(0, T; H_0^1(0, L) \times L^2(0, L))$ , it admits a unique solution in the class (2.1) satisfying (2.2) and the hidden regularity property (2.6). Moreover, the conditions of Theorem 2.1 on the initial data and right-hand side terms are satisfied for (4.1). Therefore, in the limit as  $k \rightarrow \infty$ ,

$$(4.2) \quad \phi + \psi_x = 0.$$

Multiplying both sides of (1.1)<sub>1</sub> by  $\phi$  and of (1.1)<sub>2</sub> by  $\psi$  and integrating, formally, by parts in  $Q$ , we obtain the identity

$$(4.3) \quad \begin{aligned} & \int_Q [f(x, t)u(x, t) + g(x, t)v(x, t)] dx dt = \frac{\rho h^3}{12} \int_0^L \phi(x, 0)u_1(x) dx \\ & - \frac{\rho h^3}{12} \int_0^L \phi'(x, 0)u_0(x) dx + \rho h \int_0^L [\psi(x, 0)v_1(x) - \psi'(x, 0)v_0(x)] dx \\ & - k \int_0^T \Theta_k \psi(0, t) dt. \end{aligned}$$

In view of (2.1) and (2.6), the right-hand side of (4.3) makes sense, provided

$$(4.4) \quad \{u_0, u_1, v_0, v_1\} \in \mathcal{X}' = L^2(0, L) \times H^{-1}(0, L) \times L^2(0, L) \times [H^1(0, L)]'$$

and

$$(4.5) \quad \Theta_k \in [H^1(0, T)]'.$$

Assuming that  $\Theta_k$  is of the form

$$(4.6) \quad \Theta_k = \Theta'_{1k}/k \text{ with } \Theta_{1k} \in L^2(0, T) \text{ of compact support in } (0, T),$$

the identity (4.3) may be rewritten as

$$(4.7) \quad \int_Q [f(x, t) u(x, t) + g(x, t) v(x, t)] dx dt = \frac{\rho h^3}{12} [\langle u_1, \phi(\cdot, 0) \rangle_0 - (u_0, \phi'(\cdot, 0))] + \rho h [\langle v_1, \psi(\cdot, 0) \rangle_1 - (v_0, \psi'(\cdot, 0))] + \int_0^T \Theta_{1k} \psi'(0, t) dt,$$

where  $\langle \cdot, \cdot \rangle_0$  (resp.,  $\langle \cdot, \cdot \rangle_1$ ) represents the duality between  $H^{-1}(0, L)$  (resp.,  $(H^1(0, L))'$ ) and  $H_0^1(0, L)$  (resp.,  $H^1(0, L)$ ).

Note that in (4.6) the prime ' stands for the classical derivative in the sense of distributions.

We adopt (4.7) as a definition of the solution of (1.1)–(1.4) in the sense of transposition. Arguing as in [8] and in view of the hidden regularity properties in Proposition 2.2, we deduce that system (1.1)–(1.4) has a unique solution in the class

$$\{u, v\} \in C^0([0, T]; [L^2(0, L)]^2).$$

Moreover, there exists a constant  $C > 0$ , independent of  $k$ , such that

$$(4.8) \quad \|\{u, v\}\|_{L^\infty(0, T; [L^2(0, L)]^2)} \leq C \left( \|\{u_0, u_1, v_0, v_1\}\|_{\mathcal{X}'} + \|\Theta_{1k}\|_{L^2(0, T)} \right).$$

Similarly one can show that

$$(4.9) \quad \{u, v\} \in C^1([0, T]; H^{-1}(0, L) \times [H^1(0, L)]')$$

and show an estimate of the form

$$(4.10) \quad \|\{u, v\}\|_{W^{1, \infty}(0, T; H^{-1}(0, L) \times [H^1(0, L)]')} \leq C \left( \|\{u_0, u_1, v_0, v_1\}\|_{\mathcal{X}'} + \|\Theta_{1k}\|_{L^2(0, T)} \right).$$

The solution by transposition of the system (1.5) can be defined in a similar way. Indeed, multiplying system (1.5) by the weak solution  $\psi$  of the backward problem (that can be transformed into (2.13) by time-reversal)

$$(4.11) \quad \begin{cases} \rho h \psi'' - \frac{\rho h^3}{12} \psi_{xx}'' + \psi_{xxxx} = f_x + g & \text{in } Q, \\ \psi_x(0, \cdot) = \psi_x(L, \cdot) = \psi_{xxx}(0, \cdot) = \psi_{xxx}(L, \cdot) = 0 & \text{on } (0, T), \\ \psi(\cdot, T) = \psi'(\cdot, T) = 0 & \text{in } (0, L) \end{cases}$$

and after integrating by parts in  $Q$ , we get

$$(4.12) \quad \int_Q [g(x, t) + f_x(x, t)] v(x, t) dx dt = \frac{\rho h^3}{12} [\langle v_{1x}, \psi_x(\cdot, 0) \rangle_0 - (v_{0x}, \psi'_x(\cdot, 0))] + \rho h [(v_1, \psi(\cdot, 0)) - (v_0, \psi'(\cdot, 0))] + \int_0^T \Xi \psi(0, t) dt.$$

We adopt identity (4.12) as a definition of the solution of (1.5) in the sense of transposition. In this sense, when  $\{v_0, v_1\} \in H^1(0, L) \times L^2(0, L)$ , system (1.5) possesses a unique solution in the class  $v \in C^0([0, T]; H^1(0, L)) \cap C^1([0, T]; L^2(0, L))$ .

The following result describes the asymptotic behavior as  $k \rightarrow \infty$ .



THEOREM 4.1. Consider initial data  $\{u_0, u_1, v_0, v_1\} \in \mathcal{X}'$  independent of  $k$  such that

$$(4.13) \quad u_0 + v_{0x} = 0, \quad u_1 + v_{1x} = 0,$$

and  $\Theta_k$  satisfying (4.6) and

$$(4.14) \quad \Theta_{1k} \rightarrow \Theta_1 \text{ weakly in } L^2(0, T), \quad \Theta_1 \text{ being of compact support in } (0, T).$$

Let  $\{u_k, v_k\}$  be the solution of (1.1)–(1.4). Then, as  $k \rightarrow \infty$ , the convergence

$$(4.15) \quad \{u_k, v_k\} \rightarrow \{-v_x, v\} \text{ weakly* in } L^\infty(0, T; L^2(0, L) \times L^2(0, L))$$

holds, where  $v$  is the solution of system (1.5) with  $\Xi = -\Theta'_1$ .

Remark 4.1. As we shall see in the application to controllability, the controls for both the Midlin–Timoshenko and Kirchhoff systems can be taken to be of compact support in  $(0, T)$ .

Proof of Theorem 4.1. For data  $\{u_0, u_1, v_0, v_1, \Theta_k\}$  in the conditions of Theorem 4.1, we consider, for each  $k > 0$ ,  $\{u_k, v_k\}$  the unique solution of (1.1) – (1.4) in the sense of transposition.

Using (2.1), (2.2), (2.6), and (4.14), it follows, by (4.7), that

$$\{u_k, v_k\} \text{ is bounded in } L^\infty(0, T; L^2(0, L) \times L^2(0, L)).$$

Then we can extract a subsequence, that we still denote in the same form, such that

$$(4.16) \quad \{u_k, v_k\} \rightarrow \{u, v\} \text{ weakly* in } L^\infty(0, T; L^2(0, L) \times L^2(0, L)).$$

Applying (4.2), (4.14), and (4.16) in (4.7), we obtain, in the limit,

$$(4.17) \quad \int_Q [f(x, t) u(x, t) + g(x, t) v(x, t)] dx dt = \frac{\rho h^3}{12} [-\langle u_1, \psi_x(\cdot, 0) \rangle_0 + (u_0, \psi'_x(\cdot, 0))] \\ + \rho h [\langle v_1, \psi(\cdot, 0) \rangle_1 - (v_0, \psi'(\cdot, 0))] + \int_0^T \Theta_1 \psi'(0, t) dt,$$

where  $\psi$  is the weak solution of system (4.11). Note that here we have used the fact that the weak convergence property in Theorem 2.1 is also true for the solutions of the adjoint system evaluated at time  $t = 0$ . This result has not been explicitly stated in Theorem 2.1 but can be derived in view of the properties stated there and standard arguments.

On the other hand, from (1.1)<sub>1</sub> we have that

$$u_k + v_{kx} = -\frac{1}{k} \left( \frac{\rho h^3}{12} u''_k - u_{kxx} \right).$$

Then, in the limit as  $k \rightarrow \infty$  (convergence takes place in a very weak topology),

$$u + v_x = 0.$$

In this way, using the last equation, and the compatibility conditions on the initial data (4.13), identity (4.17) can be written as in (4.12) with  $\Xi = -\Theta'_1$ . Thus,  $v$  is the unique solution, by transposition, of system (1.5).  $\square$

**5. Nonuniform observability.** In this section we consider the adjoint system (1.6) in the particular case where  $f \equiv g \equiv 0$ . More precisely, assume that  $\{\phi, \psi\}$  solves

$$(5.1) \quad \begin{cases} \frac{\rho h^3}{12} \phi'' - \phi_{xx} + k(\phi + \psi_x) = 0 & \text{in } Q, \\ \rho h \psi'' - k(\phi + \psi_x)_x = 0 & \text{in } Q, \\ \phi(0, t) = \phi(L, t) = \psi_x(0, t) = \psi_x(L, t) = 0 & \text{on } (0, T), \\ \phi(x, 0) = \phi_0(x), \quad \phi'(x, 0) = \phi_1(x) & \text{in } (0, L), \\ \psi(x, 0) = \psi_0(x), \quad \psi'(x, 0) = \psi_1(x) & \text{in } (0, L). \end{cases}$$

We have the following observability result.

**THEOREM 5.1.** *For  $T > 2\alpha L$ , with*

$$\alpha = \max \left\{ \sqrt{\frac{\rho h^3}{12}}, \sqrt{\frac{\rho h}{k}} \right\}, \quad k \geq 1, \quad \text{and } h \leq \min \left\{ \sqrt[3]{\frac{3}{\rho}}, \frac{1}{4\rho} \right\},$$

*there exists a constant  $C_k^* > 0$  such that, for any solution of (5.1),*

$$(5.2) \quad \|\{\phi_0, \phi_1, \psi_0, \psi_1\}\|_k^2 \leq C_k^* \int_0^T \left\{ |\phi_x(0, t)|^2 + |\psi(0, t)|^2 + \rho h |\psi'(0, t)|^2 \right\} dt.$$

*More precisely,*

$$C_k^* = AC_k,$$

*where  $A = A(T, L, \rho, h)$  is a positive constant and*

$$(5.3) \quad C_k = \frac{L}{2(T - 2\alpha L)} \exp \left( \sqrt{k}L + \frac{2\sqrt{3}L}{h} + L^3 + 3L \right).$$

*Remark 5.1.*

- The observability time in Theorem 5.1 is optimal and uniform in the sense that, for  $k$  large enough, or more precisely, for  $k \geq 12/h^2$ , we can take  $T > 2L\sqrt{\rho h^3/12}$  independent of  $k$ . However, the observability constant  $C_k^*$  diverges exponentially as  $k \rightarrow \infty$ . Therefore it is not of use for getting uniform controllability results as  $k$  tends to infinity.
- The hypotheses of Theorem 5.1 on  $h$  and  $k$  are natural since, as it was said in the introduction, the Mindlin–Timoshenko model was deduced for thin beams (which makes the smallness assumption on  $h$  natural) and also because we are interested in the singular limit  $k \rightarrow \infty$ .

*Proof of Theorem 5.1.*

*Step 1.* First, we consider  $\{\phi_0, \phi_1, \psi_0, \psi_1\} \in \mathcal{X}_0$ . In this case the energy defines a norm equivalent to the usual one  $\|\cdot\|_k$ . We will prove that

$$(5.4) \quad E_k(0) \leq C_k \int_0^T \left\{ |\phi_x(0, t)|^2 + |\psi(0, t)|^2 + \rho h |\psi'(0, t)|^2 \right\} dt.$$

For this, we use a genuinely one-dimensional method which consists roughly of viewing (5.1)<sub>1</sub> and (5.1)<sub>2</sub> as evolution equations with respect to  $x$ , while  $t$  plays the role of the space variable. This argument was used in [15] when studying the controllability of the semilinear wave equation in one space dimension.

Let us define the functional

$$(5.5) \quad F_k(x) = \frac{1}{2} \int_{\alpha x}^{T-\alpha x} \left\{ \frac{\rho h^3}{12} |\phi'(x, t)|^2 + \rho h |\psi'(x, t)|^2 + |\phi_x(x, t)|^2 + k |\phi(x, t) + \psi_x(x, t)|^2 + |\psi(x, t)|^2 \right\} dt.$$

Note that

$$(5.6) \quad F_k(0) = \frac{1}{2} \int_0^T \left\{ |\phi_x(0, t)|^2 + |\psi(0, t)|^2 + \rho h |\psi'(0, t)|^2 \right\} dt.$$

The derivative of the functional  $F_k$  is

$$(5.7) \quad F'_k(x) = \int_{\alpha x}^{T-\alpha x} \left\{ \frac{\rho h^3}{12} \phi'(x, t) \phi'_x(x, t) + \rho h \psi'(x, t) \psi'_x(x, t) + \phi_x(x, t) \phi_{xx}(x, t) + k (\phi(x, t) + \psi_x(x, t)) (\phi(x, t) + \psi_x(x, t))_x + \psi(x, t) \psi_x(x, t) \right\} dt - \frac{1}{2} \sum_{t=T-\alpha x, \alpha x} \left\{ \frac{\rho h^3}{12} |\phi'(x, t)|^2 + \rho h |\psi'(x, t)|^2 + |\phi_x(x, t)|^2 + k |\phi(x, t) + \psi_x(x, t)|^2 + |\psi(x, t)|^2 \right\}.$$

Integrating by parts and using (5.1)<sub>1</sub>, we get

$$(5.8) \quad \begin{aligned} & \int_{\alpha x}^{T-\alpha x} \frac{\rho h^3}{12} \phi'(x, t) \phi'_x(x, t) dt = - \int_{\alpha x}^{T-\alpha x} \frac{\rho h^3}{12} \phi''(x, t) \phi_x(x, t) dt \\ & + \left[ \frac{\rho h^3}{12} \phi'(x, t) \phi_x(x, t) \right]_{\alpha x}^{T-\alpha x} = - \int_{\alpha x}^{T-\alpha x} \phi_{xx}(x, t) \phi_x(x, t) dt \\ & + \int_{\alpha x}^{T-\alpha x} k (\phi(x, t) + \psi_x(x, t)) \phi_x(x, t) dt + \left[ \frac{\rho h^3}{12} \phi'(x, t) \phi_x(x, t) \right]_{\alpha x}^{T-\alpha x}. \end{aligned}$$

Since  $h \leq (3/\rho)^{\frac{1}{3}}$ , we have

$$(5.9) \quad \begin{aligned} & \left[ \frac{\rho h^3}{12} \phi'(x, t) \phi_x(x, t) \right]_{\alpha x}^{T-\alpha x} \leq \frac{1}{4} \sum_{t=T-\alpha x, \alpha x} \left\{ \left( \frac{\rho h^3}{6} \right)^2 |\phi'(x, t)|^2 + |\phi_x(x, t)|^2 \right\} \\ & \leq \frac{1}{4} \sum_{t=T-\alpha x, \alpha x} \left\{ \frac{\rho h^3}{12} |\phi'(x, t)|^2 + \rho h |\psi'(x, t)|^2 + |\phi_x(x, t)|^2 + k |\phi(x, t) + \psi_x(x, t)|^2 + |\psi(x, t)|^2 \right\}. \end{aligned}$$

Using (5.1)<sub>2</sub> and integrating by parts, it follows that

$$(5.10) \quad \begin{aligned} & \int_{\alpha x}^{T-\alpha x} k (\phi(x, t) + \psi_x(x, t)) (\phi(x, t) + \psi_x(x, t))_x dt \\ & = \int_{\alpha x}^{T-\alpha x} \rho h \psi''(x, t) [\phi(x, t) + \psi_x(x, t)] dt = - \int_{\alpha x}^{T-\alpha x} \rho h \phi'(x, t) \psi'(x, t) dt \\ & - \int_{\alpha x}^{T-\alpha x} \rho h \psi'(x, t) \psi'_x(x, t) dt + \{\rho h \psi'(x, t) [\phi(x, t) + \psi_x(x, t)]\}_{\alpha x}^{T-\alpha x}. \end{aligned}$$

We also get

$$\begin{aligned}
 (5.11) \quad & \{\rho h \psi'(x, t) [\phi(x, t) + \psi_x(x, t)]\}_{|\alpha x}^{T-\alpha x} \leq \frac{1}{4} \sum_{t=T-\alpha x, \alpha x}^{T-\alpha x} \{(2\rho h)^2 |\psi'(x, t)|^2 \\
 & + |\phi(x, t) + \psi_x(x, t)|^2\} \leq \frac{1}{4} \sum_{t=T-\alpha x, \alpha x} \left\{ \frac{\rho h^3}{12} |\phi'(x, t)|^2 + \rho h |\psi'(x, t)|^2 \right. \\
 & \left. + |\phi_x(x, t)|^2 + k |\phi(x, t) + \psi_x(x, t)|^2 + |\psi(x, t)|^2 \right\}
 \end{aligned}$$

because  $h \leq 1/4\rho$  and  $k \geq 1$ .

Thus, substituting (5.8)–(5.11) in (5.7), we deduce

$$\begin{aligned}
 (5.12) \quad F'_k(x) & \leq \int_{\alpha x}^{T-\alpha x} k (\phi(x, t) + \psi_x(x, t)) \phi_x(x, t) dt + \int_{\alpha x}^{T-\alpha x} \psi(x, t) \psi_x(x, t) dt \\
 & - \int_{\alpha x}^{T-\alpha x} \rho h \phi'(x, t) \psi'(x, t) dt \leq \left( \sqrt{k} + \frac{2\sqrt{3}}{h} + L^2 + 3 \right) L F_k(x)
 \end{aligned}$$

and, therefore,

$$(5.13) \quad F_k(x) \leq \exp \left( \sqrt{k}L + \frac{2\sqrt{3}L}{h} + L^3 + 3L \right) F_k(0).$$

Integrating (5.13) in  $(0, L)$ , we have

$$(5.14) \quad \int_0^L F_k(x) dx \leq L \exp \left( \sqrt{k}L + \frac{2\sqrt{3}L}{h} + L^3 + 3L \right) F_k(0).$$

Since  $T > 2\alpha L$ , we obtain, by conservation of energy and (5.14),

$$\begin{aligned}
 (5.15) \quad (T - 2\alpha L) E_k(0) & = \int_{\alpha L}^{T-\alpha L} E_k(0) dt = \int_{\alpha L}^{T-\alpha L} E_k(t) dt \leq \int_0^L F_k(x) dx \\
 & \leq L \exp \left( \sqrt{k}L + \frac{2\sqrt{3}L}{h} + L^3 + 3L \right) F_k(0),
 \end{aligned}$$

which implies (5.4).

*Step 2.* We consider now  $\{\phi_0, \phi_1, \psi_0, \psi_1\} \in \mathcal{X}$  and decompose it in the following way:

$$\{\phi_0, \phi_1, \psi_0, \psi_1\} = \{\phi_0, \phi_1, \psi_0 - c_1, \psi_1 - c_2\} + \{0, 0, c_1, c_2\},$$

where  $c_1 = (1/L) \int_0^L \psi_0(x) dx$  and  $c_2 = (1/L) \int_0^L \psi_1(x) dx$ . In this way, according to inequality (5.4), for the initial data  $\{\tilde{\phi}_0, \tilde{\phi}_1, \tilde{\psi}_0, \tilde{\psi}_1\} = \{\phi_0, \phi_1, \psi_0 - c_1, \psi_1 - c_2\} \in \mathcal{X}_0$ , the corresponding solution  $\{\tilde{\phi}, \tilde{\psi}\}$  of (5.1) satisfies

$$\left\| \{\tilde{\phi}_0, \tilde{\phi}_1, \tilde{\psi}_0, \tilde{\psi}_1\} \right\|_k^2 \leq C_{1k} \int_0^T \left\{ \left| \tilde{\phi}_x(0, t) \right|^2 + \left| \tilde{\psi}(0, t) \right|^2 + \rho h \left| \tilde{\psi}'(0, t) \right|^2 \right\} dt,$$

with  $C_{1k} = 2(1 + L + L^2) C_k$  and  $C_k$  as in (5.3).

Taking into account that  $\tilde{\psi} = \psi - c_2 t - c_1$ , it follows that  $\{\phi, \psi\}$  verifies

$$\begin{aligned}
 \|\{\phi_0, \phi_1, \psi_0, \psi_1\}\|_k^2 &\leq 2\|\{\phi_0, \phi_1, \psi_0, \psi_1\}\|_k^2 + 2\|\{0, 0, c_1, c_2\}\|_k^2 \\
 &\leq 2C_{1k} \int_0^T \left\{ |\phi_x(0, t)|^2 + |(\psi - c_2 t - c_1)(0, t)|^2 + \rho h |(\psi' - c_2)(0, t)|^2 \right\} dt \\
 (5.16) \quad &+ 2L \left[ (c_1)^2 + (c_2)^2 \right] \leq 4C_{1k} \int_0^T \left\{ |\phi_x(0, t)|^2 + |\psi(0, t)|^2 + \rho h |\psi'(0, t)|^2 \right\} dt \\
 &+ 2L \left[ (1 + 4TC_{1k})(c_1)^2 + \left( 1 + 2\rho h TC_{1k} + \frac{4T^3}{3} C_{1k} \right) (c_2)^2 \right].
 \end{aligned}$$

We now need to estimate the last term of (5.16). Integrating (5.1)<sub>2</sub> from 0 to  $L$  and using the initial and boundary conditions of system (5.1), we have

$$\int_0^L \psi'(x, t) dx = c_2 L \quad \text{and} \quad \int_0^L \psi(x, t) dx = (c_1 + tc_2)L.$$

Hence

$$(c_2)^2 \leq \frac{1}{L} \int_0^L |\psi'(x, t)|^2 dx \leq \frac{2}{\rho h L} E_k(t)$$

and

$$\begin{aligned}
 (c_1)^2 &\leq \frac{2}{L} \int_0^L |\psi(x, t)|^2 dx + 2T^2 (c_2)^2 \\
 (5.17) \quad &\leq \frac{4}{L} \max \left\{ \frac{T^2}{\rho h}, 1 \right\} \left[ E_k(t) + \frac{1}{2} \int_0^L |\psi(x, t)|^2 dx \right].
 \end{aligned}$$

Combining these estimates, we see that the observability inequality in the theorem follows easily.  $\square$

In the following section we show how the uniform (with respect to  $k$ ) observability inequality can be proved in a subspace of solutions that, as  $k \rightarrow \infty$ , covers the whole energy space for the limit system.

**6. Uniform observability.** To prove the uniform observability of filtered solutions of the Mindlin–Timoshenko system (5.1), we need the following refined version of the classical Ingham inequality on the theory of nonharmonic Fourier series (see Haraux [4] and Micu and Zuazua [14]).

**THEOREM 6.1** (see [4], [14]). *Let  $f = f(t)$  be of the form  $f(t) = \sum_{n \in \mathbb{Z}} a_n e^{i\lambda_n t}$ , where  $(\lambda_n)_n$  is a sequence of real numbers such that there exist  $N \in \mathbb{N}$ ,  $\gamma > 0$ , and  $\gamma_\infty > 0$  such that*

$$(6.1) \quad \lambda_{n+1} - \lambda_n \geq \gamma_\infty > 0 \quad \text{if} \quad |n| > N,$$

$$(6.2) \quad \lambda_{n+1} - \lambda_n \geq \gamma > 0 \quad \forall n \in \mathbb{Z}.$$

*Let  $T > 0$  be such that  $T > 2\pi/\gamma_\infty$ . Then, there exist two positive constants  $C^{(1)}$  and  $C^{(2)}$  such that*

$$(6.3) \quad C^{(1)} \sum_{n \in \mathbb{Z}} |a_n|^2 \leq \int_0^T |f(t)|^2 dt \leq C^{(2)} \sum_{n \in \mathbb{Z}} |a_n|^2$$

for all  $(a_n)_n \in l^2$ . More precisely,  $C^{(1)} = C^{(1)}(2N+1)$  and  $C^{(2)} = C^{(2)}(2N+1)$ , where  $C^{(i)}(j)$ ,  $i = 1, 2$ , are given by the following recurrent formulas:

$$\begin{cases} C^{(1)}(j+1) = \left[ \left( \frac{2C^{(2)}(j)}{|J|} + 1 \right) \frac{4}{C^{(1)}(j)(|J|\gamma_\infty - 2\pi)^2 \gamma^2} + \frac{2}{|J|} \right]^{-1}, \\ C^{(2)}(j+1) = 2[|J|(j+1) + C^{(2)}(0)], \quad j = 0, 1, \dots, \end{cases}$$

and  $C^{(1)}(0)$ ,  $C^{(2)}(0)$  are such that (6.3) holds in the particular case in which  $\gamma_\infty = \gamma > 0$ .

*Remark 6.1.* The particular case when  $\gamma_\infty = \gamma$  corresponds to the classical result by Ingham [6] which shows the existence of positive constants  $C^{(1)}$  and  $C^{(2)}$  such that (6.3) holds when  $T > 2\pi/\gamma$ . Theorem 6.1 allows us to deduce that, for general sequences  $(\lambda_n)_n$ , inequality (6.3) holds when  $T$  is smaller, because the asymptotic gap  $\gamma_\infty$  is in general larger than  $\gamma$ .

To apply Theorem 6.1 and deduce the uniform observability of system (5.1), we need precise estimates on the gap of the spectrum of (5.1). For this, we will look for solutions of this system in separated variables.

According to the asymptotic properties of the two families of eigenvalues  $(\lambda_{k,m})_{m \in \mathbb{N}}$  and  $(\tilde{\lambda}_{k,m})_{m \in \mathbb{N}}$  of (5.1) given by Proposition 3.1 and by (3.6), respectively, we consider only the family  $(\lambda_{k,m})_{m \in \mathbb{N}}$ , because it is precisely this one and its corresponding eigenfunctions that generate the solutions that converge to the solutions of the limit Kirchhoff system, while the other one disappears, as  $k$  tends to infinity, in the sense that it does not lead to the eigenvalues of the limit system.

Let us now consider the class of solutions of (5.1) generated by the eigenfunctions associated with the eigenvalues  $\lambda_{k,m}$ :

$$W_\lambda = \left\{ \{\phi, \psi\} \text{ solution of (5.1) such that} \right. \\ \left. \{\phi, \psi\} = \sum_{m \in \mathbb{N}} \left( a_{k,m}^+ e^{-i\lambda_{k,m}^+ t} + a_{k,m}^- e^{-i\lambda_{k,m}^- t} \right) \left\{ \sin\left(\frac{m\pi x}{L}\right), c_m \cos\left(\frac{m\pi x}{L}\right) \right\} \right\}$$

with  $c_m$  being as in (3.5) and  $a_{k,m}^\pm = (a_{k,m}^0 - ia_{k,m}^1/\lambda_{k,m}^\pm)/2$ , where  $a_{k,m}^0$  and  $a_{k,m}^1$  are the Fourier coefficients of the initial data  $\{\phi_0, \phi_1\}$  on the basis of sinusoidal eigenfunctions,

$$\{\phi_0, \phi_1\} = \left\{ \sum_{m \in \mathbb{N}} a_{k,m}^0 \sin\left(\frac{m\pi x}{L}\right), \sum_{m \in \mathbb{N}} a_{k,m}^1 \sin\left(\frac{m\pi x}{L}\right) \right\}.$$

Obviously, this is a strict subspace of the whole space of solutions. Indeed, in this subspace we have excluded all the eigencomponents associated with the eigenvalues  $\tilde{\lambda}_{k,m}^\pm$ . In this subspace there is a one-to-one correspondence between the initial data  $\{\phi_0, \phi_1\}$  of  $\phi$  and the initial data  $\{\psi_0, \psi_1\}$  of  $\psi$ . More precisely, the Fourier coefficients of the latter are related to the previous ones by the relations

$$(6.4) \quad b_{k,m}^0 = c_m a_{k,m}^0, \quad b_{k,m}^1 = c_m a_{k,m}^1.$$

Let us analyze the gap between consecutive eigenvalues  $\lambda_{k,m}^\pm$ . For this, we address the following result.

PROPOSITION 6.1. *Given*

$$0 < \epsilon < \pi^2 \sqrt{12/(12\rho hL^2 + \pi^2\rho h^3)}/L$$

and

$$k \geq 8(12)^3 L/\epsilon \pi \sqrt{6\rho h^3}$$

we have

$$(6.5) \quad \left| \lambda_{k,m+1}^\pm - \lambda_{k,m}^\pm \right| \geq \gamma_\infty > 0 \quad \text{with} \quad \gamma_\infty = \frac{\pi}{L} \sqrt{\frac{12}{\rho h^3}} - \epsilon \quad \forall m \geq m_0,$$

where

$$(6.6) \quad m_0 = \frac{2}{h} \sqrt[4]{\frac{9L^2}{\epsilon^2(12\rho hL^2 + \pi^2\rho h^3)}}.$$

On the other hand,

$$(6.7) \quad \left| \lambda_{k,m+1}^\pm - \lambda_{k,m}^\pm \right| \geq \gamma > 0 \quad \text{with} \quad \gamma = \frac{\pi^2}{L} \sqrt{\frac{12}{12\rho hL^2 + \pi^2\rho h^3}} - \epsilon \quad \forall m \geq 1.$$

*Proof.* To simplify the notation, we denote by  $\lambda_{km}$  both  $\lambda_{k,m}^+$  and  $\lambda_{k,m}^-$ . In view of (3.13) we get

$$\begin{aligned} |\lambda_{k,m+1} - \lambda_{k,m}| &\geq \frac{(m+1)\pi^2}{L} \sqrt{\frac{12}{\frac{12\rho hL^2}{(m+1)^2} + \pi^2\rho h^3}} - \frac{m\pi^2}{L} \sqrt{\frac{12}{\frac{12\rho hL^2}{m^2} + \pi^2\rho h^3}} - \frac{2c}{k} \\ &\geq \frac{\pi^2}{L} \sqrt{\frac{12}{\frac{12\rho hL^2}{m^2} + \pi^2\rho h^3}} - \frac{2c}{k} \geq \frac{\pi}{L} \sqrt{\frac{12}{\rho h^3}} - \left( \frac{24\sqrt{3}L}{h^2m\sqrt{12\rho hL^2 + \pi^2\rho h^3m^2}} + \frac{2c}{k} \right), \end{aligned}$$

with  $c$  being as (3.13), that is,  $c = 4(12)^2/h^4\sqrt{\rho h}$ . It is easy to see that when  $m \geq m_0$ , with  $m_0$  as in (6.6), and  $k \geq 4c/\epsilon$ , then

$$\frac{24\sqrt{3}L}{h^2m\sqrt{12\rho hL^2 + \pi^2\rho h^3m^2}} + \frac{2c}{k} \leq \epsilon.$$

This implies the asymptotic gap condition (6.5).

Let us analyze now the behavior of the gap for all  $m \geq 1$ . Proceeding as before we have

$$|\lambda_{k,m+1} - \lambda_{k,m}| \geq \frac{\pi^2}{L} \sqrt{\frac{12}{\frac{12\rho hL^2}{m^2} + \pi^2\rho h^3}} - \frac{2c}{k} \geq \frac{\pi^2}{L} \sqrt{\frac{12}{12\rho hL^2 + \pi^2\rho h^3}} - \frac{2c}{k}$$

for all  $m \geq 1$ . In this way we obtain the gap (6.7), and this concludes the proof of the proposition.  $\square$

In view of the gap conditions (6.5) and (6.7) we have all the ingredients we need to prove the following result.

THEOREM 6.2. *Let  $T > 2L\sqrt{\rho h^3/12}$ . Then there exist positive constants  $c = c(T)$  and  $C = C(T)$  such that*

$$(6.8) \quad c \|\{\phi_0, \phi_1, \psi_0, \psi_1\}\|_k^2 \leq \int_0^T |\psi'(0, t)|^2 dt \leq C \|\{\phi_0, \phi_1, \psi_0, \psi_1\}\|_k^2$$

for all solutions  $\{\phi, \psi\}$  of (5.1) in the class  $W_\lambda$  with initial data  $\{\phi_0, \phi_1, \psi_0, \psi_1\}$  satisfying the condition

$$(6.9) \quad \phi_0 + \psi_{0x} = 0.$$

*Proof.* We consider  $\{\phi, \psi\} \in W_\lambda$  as the solution of (5.1) with initial data  $\{\phi_0, \phi_1, \psi_0, \psi_1\}$  satisfying (6.9). Thus

$$(6.10) \quad \psi'(0, t) = -i \sum_{m \in \mathbb{N}} c_m \left( a_{k,m}^+ \lambda_{k,m}^+ e^{-i\lambda_{k,m}^+ t} + a_{k,m}^- \lambda_{k,m}^- e^{-i\lambda_{k,m}^- t} \right).$$

Let  $T > 2L\sqrt{\rho h^3/12}$ . Applying Theorem 6.1 to the series (6.10) and using the gap conditions, we deduce the existence of positive constants  $C^{(1)} = C^{(1)}(T, \gamma)$  and  $C^{(2)} = C^{(2)}(T, \gamma)$  such that

$$(6.11) \quad \begin{aligned} & C^{(1)} \sum_{m \in \mathbb{N}} c_m^2 \left[ \left( a_{k,m}^+ \lambda_{k,m}^+ \right)^2 + \left( a_{k,m}^- \lambda_{k,m}^- \right)^2 \right] \\ & \leq \int_0^T |\psi'(0, t)|^2 dt \leq C^{(2)} \sum_{m \in \mathbb{N}} c_m^2 \left[ \left( a_{k,m}^+ \lambda_{k,m}^+ \right)^2 + \left( a_{k,m}^- \lambda_{k,m}^- \right)^2 \right]. \end{aligned}$$

Since the initial data  $\{\phi_0, \phi_1, \psi_0, \psi_1\}$  satisfy (6.9), it is easy to see that, for the family of solutions under consideration, the following equivalence holds true:

$$(6.12) \quad \sum_{m \in \mathbb{N}} c_m^2 \left[ \left( a_{k,m}^+ \lambda_{k,m}^+ \right)^2 + \left( a_{k,m}^- \lambda_{k,m}^- \right)^2 \right] \sim \|\{\phi_0, \phi_1, \psi_0, \psi_1\}\|_k^2$$

uniformly on  $k$  for all data  $\{\phi_0, \phi_1, \psi_0, \psi_1\}$  whose solution  $\{\phi, \psi\} \in W_\lambda$ . Combining (6.11) and (6.12) we complete the proof of the theorem.  $\square$

*Remark 6.2.* Let us compare the observability inequalities in (6.8) with Theorem 5.1 as follows:

- In Theorem 5.1, the observed quantity in the right-hand side term of (5.2) depends on both  $\phi$  and  $\psi$ . This would imply a controllability result for system (1.1)–(1.4), but with an extra control entering on  $u$  at  $x = 0$ . In (6.8) the observed quantity depends only on  $\psi$  (more precisely on  $\psi'(0, t)$ ) and this corresponds to using one simple control in (1.1)–(1.4).
- The time of controllability in (6.8) is smaller and the observability constant remains bounded as  $k \rightarrow \infty$ .
- We obtained inequalities (6.8) only for the solutions in the subspace  $W_\lambda$ , since the other family of eigenvalues  $(\tilde{\lambda}_{k,m})_{m \in \mathbb{N}}$  diverges (as  $k \rightarrow \infty$ ) and, consequently, the subspace they generate does not contribute to our main goal, which is to show the controllability of the Kirchhoff system as a limit of the Mindlin–Timoshenko one, as we shall see in the following section.



*Remark 6.3.* Let us finally mention a variant of the observability result in (6.8) that will be used in what follows. Consider a function  $\beta : (0, T) \rightarrow [0, 1]$  in the class  $C^\infty$  such that

$$(6.13) \quad \beta(t) = \begin{cases} 1 & \text{if } t \in (2\epsilon, T - 2\epsilon), \\ 0 & \text{if } t \in (0, \epsilon) \cup (T - \epsilon, T) \end{cases}$$

with  $\epsilon > 0$  sufficiently small such that  $T - 2\epsilon > 2L\sqrt{\rho h^3/12}$ . In view of the time invariance of system (5.1), we deduce

$$(6.14) \quad c \|\{\phi(\cdot, \epsilon), \phi'(\cdot, \epsilon), \psi(\cdot, \epsilon), \psi'(\cdot, \epsilon)\}\|_k^2 \leq \int_0^T \beta(t) |\psi'(0, t)|^2 dt$$

for all solutions  $\{\phi, \psi\}$  of (5.1) in the class  $W_\lambda$ .

**7. Uniform controllability in optimal time.** Due to the results of uniform observability obtained in the previous section, we can apply the HUM to obtain a uniform (with respect to  $k$ ) controllability result for suitable projections of solutions of the Mindlin–Timoshenko system. To be more precise, since only the eigenvalues of the family  $(\lambda_{k,m})_{m \in \mathbb{N}}$  tend to eigenvalues of the limit Kirchhoff system, it is sufficient to obtain the control result on the projections  $\Pi_\lambda$  over the eigencomponents entering in the subspace  $W_\lambda^0$  of  $W_\lambda$  as follows:

$W_\lambda^0 = \{\{\phi, \psi\} \in W_\lambda \text{ such that the initial data } \{\phi_0, \phi_1, \psi_0, \psi_1\} \text{ satisfy (6.9)}\}.$

The partial controllability condition we shall achieve at the final time  $t = T$  reads

$$(7.1) \quad \Pi_\lambda \{u_k(\cdot, T), u'_k(\cdot, T), v_k(\cdot, T), v'_k(\cdot, T)\} = 0.$$

This means that

$$(7.2) \quad \begin{aligned} \frac{\rho h^3}{12} \langle u'_k(\cdot, T), \sin(m\pi x/L) \rangle_0 + \rho h \langle v'_k(\cdot, T), c_m \cos(m\pi x/L) \rangle_1 &= 0, \\ \frac{\rho h^3}{12} (u_k(\cdot, T), \sin(m\pi x/L)) + \rho h (v_k(\cdot, T), c_m \cos(m\pi x/L)) &= 0 \quad \forall m \in \mathbb{N}. \end{aligned}$$

Furthermore, we also describe the asymptotic behavior of the controls, as  $k \rightarrow \infty$ . As we shall see, they converge to exact controls for the limit system. The following holds.

**THEOREM 7.1.** *Let  $T > 2L\sqrt{\rho h^3/12}$ . Then, for all initial data  $\{u_0, u_1, v_0, v_1\} \in \mathcal{X}'$  satisfying the compatibility condition (4.13), there exists a control  $\Theta_k \in H^{-1}(0, T)$ , with*

$$k\Theta_k = \Theta'_{1k} : \Theta_{1k} \in L^2(0, T) \text{ of compact support in } (0, T),$$

such that the solution  $\{u_k, v_k\}$  of (1.1)–(1.4) satisfies (7.1).

Moreover, the function  $\Theta_{1k}$  may be written in the form  $\Theta_{1k} = -\rho h \beta(\cdot) \widehat{\psi}'_k(0, \cdot)$ , where  $\{\widehat{\phi}_k, \widehat{\psi}_k\} \in W_\lambda^0$  is the solution of system (5.1) with initial data  $\{\widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k}\}$  minimizing the functional

$$(7.3) \quad \begin{aligned} J_k \{\phi_{0k}, \phi_{1k}, \psi_{0k}, \psi_{1k}\} &= \frac{\rho h}{2} \int_0^T \beta(t) |\psi'_k(0, t)|^2 dt - \frac{\rho h^3}{12} \langle u_1, \phi_{0k} \rangle_0 \\ &\quad + \frac{\rho h^3}{12} (u_0, \phi_{1k}) - \rho h [\langle v_1, \psi_{0k} \rangle_1 - (v_0, \psi_{1k})] \end{aligned}$$

over  $W_\lambda^0$ , where  $\{\phi_k, \psi_k\} \in W_\lambda^0$  is the solution of (5.1) with initial data  $\{\phi_{0k}, \phi_{1k}, \psi_{0k}, \psi_{1k}\}$ . Furthermore, as  $k \rightarrow \infty$ ,

$$\Theta_{1k} \rightarrow \Theta_1 \text{ strongly in } L^2(0, T)$$

with  $\Theta_1$  of compact support in  $(0, T)$ . The limit control  $\Xi = -\Theta_1' \in H^{-1}(0, T)$  is an exact control driving system (1.5) to equilibrium in time  $T$ . Moreover, the function  $\Theta_1$  may be written in the form  $\Theta_1 = -\rho h \beta(\cdot) \hat{\psi}'(0, \cdot)$ , where  $\hat{\psi}$  is the solution of the adjoint system

$$(7.4) \quad \begin{cases} \rho h \psi'' - \frac{\rho h^3}{12} \psi_{xx}'' + \psi_{xxxx} = 0 & \text{in } Q, \\ \psi_x(0, \cdot) = \psi_x(L, \cdot) = \psi_{xxx}(0, \cdot) = \psi_{xxx}(L, \cdot) = 0 & \text{on } (0, T), \\ \psi(\cdot, 0) = \psi_0, \quad \psi'(\cdot, 0) = \psi_1 & \text{in } (0, L), \end{cases}$$

with initial data  $\{\hat{\psi}_0, \hat{\psi}_1\} \in W \times H^1(0, L)$  minimizing the functional

$$(7.5) \quad J\{\psi_0, \psi_1\} = \frac{\rho h}{2} \int_0^T \beta(t) |\psi'(0, t)|^2 dt - \frac{\rho h^3}{12} [\langle v_{1x}, \psi_{0x} \rangle_0 - \langle v_{0x}, \psi_{1x} \rangle] - \rho h [\langle v_1, \psi_0 \rangle - \langle v_0, \psi_1 \rangle],$$

where  $\psi$  is the solution of (7.4) with initial data  $\{\psi_0, \psi_1\}$ .

*Remark 7.1.* In the hypotheses of Theorem 7.1, there are many possible controls  $\Theta_k \in H^{-1}(0, T)$  and  $\Xi \in H^{-1}(0, T)$  fulfilling the controllability requirements. The construction we develop below, presented in the statement of the theorem, provides controls of the form  $\Theta_k = \mu_k'$  and  $\Xi = \mu'$ , with  $\mu_k, \mu \in L^2(0, T)$  having compact support in time and a minimal  $L_\beta^2$ -norm. The weight function  $\beta$  is chosen as in (6.13).

*Proof of Theorem 7.1.* We proceed in several steps.

*Step 1. Existence of the control.* Consider  $\{\phi_k, \psi_k\} \in W_\lambda^0$  the unique solution of the adjoint system (5.1) with initial data  $\{\phi_{0k}, \phi_{1k}, \psi_{0k}, \psi_{1k}\}$ . Multiplying (1.1)<sub>1</sub> and (1.1)<sub>2</sub> by  $\phi_k$  and  $\psi_k$ , respectively, and integrating by parts in  $Q$ , we get

$$\begin{aligned} & \frac{\rho h^3}{12} \{ [\langle u_k'(\cdot, T), \phi_k(\cdot, T) \rangle_0 - \langle u_k(\cdot, T), \phi_k'(\cdot, T) \rangle] - [\langle u_1, \phi_{0k} \rangle_0 - \langle u_0, \phi_{1k} \rangle] \} \\ & + \rho h \{ [\langle v_k'(\cdot, T), \psi_k(\cdot, T) \rangle_1 - \langle v_k(\cdot, T), \psi_k'(\cdot, T) \rangle] - [\langle v_1, \psi_{0k} \rangle_1 - \langle v_0, \psi_{1k} \rangle] \} \\ & - \int_0^T \Theta_{1k} \psi_k'(0, t) dt = 0. \end{aligned}$$

Thus to prove (7.1) in the sense of (7.2) it is sufficient to prove the existence of  $\Theta_{1k} \in L^2(0, T)$  such that

$$(7.6) \quad -\frac{\rho h^3}{12} [\langle u_1, \phi_{0k} \rangle_0 - \langle u_0, \phi_{1k} \rangle] - \rho h [\langle v_1, \psi_{0k} \rangle_1 - \langle v_0, \psi_{1k} \rangle] - \int_0^T \Theta_{1k} \psi_k'(0, t) dt = 0$$

for all data  $\{\phi_{0k}, \phi_{1k}, \psi_{0k}, \psi_{1k}\}$  whose solution  $\{\phi_k, \psi_k\} \in W_\lambda^0$ .

In view of the structure of  $\beta$  and due to (6.8), the quadratic functional  $J_k$  defined in (7.3) is continuous, strictly convex, and coercive. So, there exists a unique

minimizer  $\{\widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k}\}$ , whose solution  $\{\widehat{\phi}_k, \widehat{\psi}_k\} \in W_\lambda^0$  can be characterized by the formula

$$(7.7) \quad \begin{aligned} & -\frac{\rho h^3}{12} [\langle u_1, \phi_{0k} \rangle_0 - \langle u_0, \phi_{1k} \rangle] - \rho h [\langle v_1, \psi_{0k} \rangle_1 - \langle v_0, \psi_{1k} \rangle] \\ & + \rho h \int_0^T \beta(t) \widehat{\psi}'_k(0, t) \psi'_k(0, t) dt = 0 \end{aligned}$$

for all data  $\{\phi_{0k}, \phi_{1k}, \psi_{0k}, \psi_{1k}\}$  whose solution  $\{\phi_k, \psi_k\} \in W_\lambda^0$ .

According to (7.7), the function  $\Theta_{1k} = -\rho h \beta(\cdot) \widehat{\psi}'_k(0, \cdot) \in L^2(0, T)$ , where  $\{\widehat{\phi}_k, \widehat{\psi}_k\} \in W_\lambda^0$  solves (5.1) with the minimizer  $\{\widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k}\}$  as data, verifies (7.6). Therefore

$$(7.8) \quad \Theta_k = -\frac{\rho h}{k} \left[ \beta(\cdot) \widehat{\psi}'_k(0, \cdot) \right]' \in H^{-1}(0, T)$$

is the control we were looking for.

*Step 2. Uniform bound of the control.* Let us observe that, since  $\{\widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k}\}$  is the minimizer of  $J_k$ , we have

$$J_k \left\{ \widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k} \right\} \leq J_k \{0, 0, 0, 0\} = 0.$$

Consequently

$$(7.9) \quad \int_0^T |\Theta_{1k}(t)|^2 dt \leq C \|\{u_0, u_1, v_0, v_1\}\|_{\mathcal{X}'} \left\| \left\{ \widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k} \right\} \right\|_k.$$

In view of the first inequality of (6.8), we can estimate the last term in (7.9) by

$$(7.10) \quad C \|\{u_0, u_1, v_0, v_1\}\|_{\mathcal{X}'} \left( \int_0^T \rho h \beta(t) \left| \widehat{\psi}'_k(0, t) \right|^2 dt \right)^{\frac{1}{2}}.$$

Combining (7.9) and (7.10), we obtain

$$(7.11) \quad \|\Theta_{1k}\|_{L^2(0, T)} \leq C \|\{u_0, u_1, v_0, v_1\}\|_{\mathcal{X}'}.$$

*Step 3. Convergence of controls.* Thanks to (7.11) there exists a subsequence of  $(\Theta_{1k})$  (still denoted by the index  $k$  to simplify the notation) such that

$$(7.12) \quad \Theta_{1k} \rightharpoonup \Theta_1 \text{ weakly in } L^2(0, T).$$

We now consider  $\{u_k, v_k\}$  as the solution of (1.1)–(1.4) with  $\Theta_k$  given in (7.8). Thus, we are in the conditions of Theorem 4.1 and we can assert that the convergence (4.15) holds.

It remains to prove that  $\Xi = -\Theta'_1$  is the control such that the solution  $v$  of (1.5) satisfies

$$(7.13) \quad v(\cdot, T) = v'(\cdot, T) = 0 \quad \text{in } (0, L),$$

with

$$(7.14) \quad \Theta_1 = -\rho h \beta(\cdot) \widehat{\psi}'(0, \cdot),$$

where  $\widehat{\psi}$  is the solution of (7.4) with initial data  $\{\widehat{\psi}_0, \widehat{\psi}_1\} \in W \times H^1(0, L)$  minimizing the functional (7.5). For this, it is sufficient to prove that

$$(7.15) \quad \begin{aligned} & -\frac{\rho h^3}{12} [\langle v_{1x}, \psi_{0x} \rangle_0 - (v_{0x}, \psi_{1x})] - \rho h [(v_1, \psi_0) - (v_0, \psi_1)] \\ & + \rho h \int_0^T \beta(t) \widehat{\psi}'(0, t) \psi'(0, t) dt = 0 \quad \forall \{\psi_0, \psi_1\} \in W \times H^1(0, L), \end{aligned}$$

where  $\psi$  is the solution of (7.4) with initial data  $\{\psi_0, \psi_1\}$ .

We know that, for  $\Theta_{1k} = -\rho h \beta(\cdot) \widehat{\psi}'_k(0, \cdot)$ , where  $\{\widehat{\phi}_k, \widehat{\psi}_k\} \in W_\lambda^0$  is the solution of (5.1) with the minimizer  $\{\widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k}\}$  as data, the solution of system (1.1)–(1.4) satisfies (7.1). Hence, we get

$$(7.16) \quad -\frac{\rho h^3}{12} [\langle u_1, \phi_{0k} \rangle_0 - (u_0, \phi_{1k})] - \rho h [\langle v_1, \psi_{0k} \rangle_1 - (v_0, \psi_{1k})] - \int_0^T \Theta_{1k} \psi'_k(0, t) dt = 0$$

for all data  $\{\phi_{0k}, \phi_{1k}, \psi_{0k}, \psi_{1k}\}$  whose solution  $\{\phi_k, \psi_k\} \in W_\lambda^0$ .

Combining the first inequalities in (6.8) and (7.11) we deduce that the sequence  $(\{\widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k}\})$  is uniformly bounded in  $\mathcal{X}$ . So, extracting a subsequence, that we still denote by  $(\{\widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k}\})$ , we get

$$\{\widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k}\} \rightarrow \{\bar{\phi}_0, \bar{\phi}_1, \bar{\psi}_0, \bar{\psi}_1\} \text{ weakly in } \mathcal{X},$$

and, by (2.2) (in this case  $f = g = 0$ ), we can pass to the limit as  $k \rightarrow \infty$  on the corresponding solutions and see that the limit  $\bar{\psi}$  is the weak solution of (7.4) with initial data  $\{\bar{\psi}_0, \bar{\psi}_1\}$ .

Multiplying (5.1)<sub>1</sub> and (5.1)<sub>2</sub> by  $(-\eta \sigma_{xx})$  and  $\eta \sigma_x$ , respectively, where  $\eta$  belongs to  $C_0^1(2\epsilon, T - 2\epsilon)$  and  $\sigma(x) = (L - x)e^{(x-L)^3 x^3}$ , and integrating in  $(0, L) \times (2\epsilon, T - 2\epsilon)$ , with  $\epsilon > 0$  small enough, we get the following identity for the solution  $\{\widehat{\phi}_k, \widehat{\psi}_k\}$ :

$$(7.17) \quad \begin{aligned} & \frac{\rho h^3}{12} \int_{2\epsilon}^{T-2\epsilon} \int_0^L \widehat{\phi}'_k(x, t) \eta'(t) \sigma_{xx}(x) dx dt - \int_{2\epsilon}^{T-2\epsilon} \int_0^L \widehat{\phi}_{kx}(x, t) \eta(t) \sigma_{xxx}(x) \\ & + \rho h \int_{2\epsilon}^{T-2\epsilon} \int_0^L \widehat{\psi}'_{xk}(x, t) \eta'(t) \sigma(x) dx dt + \rho h L \int_{2\epsilon}^{T-2\epsilon} \beta(t) \widehat{\psi}'_k(0, t) \eta'(t) dt = 0, \end{aligned}$$

with  $\beta$  being the function in (6.13).

Passing to the limit in (7.17) we deduce that  $\bar{\psi}$  satisfies

$$(7.18) \quad \begin{aligned} & -\frac{\rho h^3}{12} \int_{2\epsilon}^{T-2\epsilon} \int_0^L \bar{\psi}'_x(x, t) \eta'(t) \sigma_{xx}(x) dx dt - L \int_{2\epsilon}^{T-2\epsilon} \Theta_1 \eta'(t) dt \\ & + \int_{2\epsilon}^{T-2\epsilon} \int_0^L \bar{\psi}_{xx}(x, t) \eta(t) \sigma_{xxx}(x) + \rho h \int_{2\epsilon}^{T-2\epsilon} \int_0^L \bar{\psi}'_x(x, t) \eta'(t) \sigma(x) dx dt = 0. \end{aligned}$$

On the other hand, multiplying (7.4)<sub>1</sub> by  $\eta \sigma_x$ , we have the following identity for the limit solution  $\bar{\psi}$ :

$$(7.19) \quad \begin{aligned} & \rho h \int_{2\epsilon}^{T-2\epsilon} \int_0^L \overline{\psi}'_x(x, t) \eta'(t) \sigma(x) dx dt + \rho h L \int_{2\epsilon}^{T-2\epsilon} \beta(t) \overline{\psi}'(0, t) \eta'(t) dt \\ & - \frac{\rho h^3}{12} \int_{2\epsilon}^{T-2\epsilon} \int_0^L \overline{\psi}'_x(x, t) \eta'(t) \sigma_{xx}(x) dx dt + \int_{2\epsilon}^{T-2\epsilon} \int_0^L \overline{\psi}_{xx}(x, t) \eta(t) \sigma_{xxx}(x) dx dt = 0. \end{aligned}$$

Combining (7.18) and (7.19) we finally deduce

$$(7.20) \quad \int_{2\epsilon}^{T-2\epsilon} \left[ \Theta_1 + \rho h \beta(t) \overline{\psi}'(0, t) \right] \eta'(t) dt = 0 \quad \forall \eta \in C_0^1(2\epsilon, T-2\epsilon)$$

and then

$$(7.21) \quad \Theta_1 = -\rho h \beta(\cdot) \overline{\psi}'(0, \cdot),$$

where  $\overline{\psi}$  is the solution of the adjoint system (7.4).

To show that (7.15) is satisfied it is sufficient to pass to the limit in (7.16) using as test functions the solutions of the corresponding adjoint systems in separated variables. In this way, one rigorously reproduces at the variational level the proof that, heuristically, would consist of passing to the limit in (7.2) and, using the fact that  $u = -v_x$  in the limit, of deducing

$$(7.22) \quad -\frac{\rho h^3}{12} [\langle v_{1x}, \psi_{0x} \rangle_0 - \langle v_{0x}, \psi_{1x} \rangle] - \rho h [\langle v_1, \psi_0 \rangle - \langle v_0, \psi_1 \rangle] - \int_0^T \Theta_1 \psi'(0, t) dt = 0$$

and, consequently,

$$(v(\cdot, T), \cos(m\pi x/L)) = (v'(\cdot, T), \cos(m\pi x/L)) = 0 \quad \forall m \in \mathbb{N}.$$

To conclude the proof of the theorem, it remains to prove that the function  $\Theta_1$  can be identified as in (7.14). In fact, it follows from (7.15), (7.21), and (7.22) that

$$\int_0^T \beta(t) \left[ \widehat{\psi}'(0, t) - \overline{\psi}'(0, t) \right] \psi'(0, t) dt = 0$$

for all solutions  $\psi$  of the adjoint problem (7.4).

Taking  $\psi = \widehat{\psi} - \overline{\psi}$ , it follows that

$$\int_0^T \beta(t) \left[ \widehat{\psi}'(0, t) - \overline{\psi}'(0, t) \right]^2 dt = 0$$

and, therefore, we obtain (7.14). Considering (7.15) and (7.16) with data  $\{\widehat{\psi}_0, \widehat{\psi}_1\}$  and  $\{\widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k}\}$ , respectively, we get

$$\begin{aligned} & -\frac{\rho h^3}{12} \left[ \langle u_1, \widehat{\phi}_{0k} \rangle_0 - \langle u_0, \widehat{\phi}_{1k} \rangle \right] - \rho h \left[ \langle v_1, \widehat{\psi}_{0k} \rangle_1 - \langle v_0, \widehat{\psi}_{1k} \rangle \right] \\ & + \rho h \int_0^T \beta(t) \left| \widehat{\psi}'_k(0, t) \right|^2 dt = 0 \end{aligned}$$

and

$$\begin{aligned} & -\frac{\rho h^3}{12} \left[ \langle v_{1x}, \widehat{\psi}_{0x} \rangle_0 - \langle v_{0x}, \widehat{\psi}_{1x} \rangle \right] - \rho h \left[ \langle v_1, \widehat{\psi}_0 \rangle - \langle v_0, \widehat{\psi}_1 \rangle \right] \\ & + \rho h \int_0^T \beta(t) \left| \widehat{\psi}'(0, t) \right|^2 dt = 0. \end{aligned}$$

It follows from the last two equations that

$$\int_0^T \beta(t) \left| \widehat{\psi}'_k(0, t) \right|^2 dt \rightarrow \int_0^T \beta(t) \left| \widehat{\psi}'(0, t) \right|^2 dt$$

which, together with the weak convergence (7.12), yields

$$\Theta_{1k} \rightarrow \Theta_1 \text{ strongly in } L^2(0, T)$$

and

$$\lim_{k \rightarrow \infty} J_k \left\{ \widehat{\phi}_{0k}, \widehat{\phi}_{1k}, \widehat{\psi}_{0k}, \widehat{\psi}_{1k} \right\} = J \left\{ \widehat{\psi}_0, \widehat{\psi}_1 \right\},$$

proving the theorem.  $\square$

*Remark 7.2.* According to Theorem 7.1 we can recover the exact controllability property of the Kirchhoff system as a limit of the partial controllability properties of the Mindlin–Timoshenko one.

*Remark 7.3.* Let us observe also that these results are obtained for the optimal control time  $T > 2L\sqrt{\rho h^3/12}$  which is the best possible one for both Mindlin–Timoshenko and Kirchhoff systems.

**Acknowledgments.** This paper was part of F. D. Araruna’s doctoral thesis [1] at Universidade Federal do Rio de Janeiro in 2004. Most of this research was done while this author was visiting Universidad Autónoma de Madrid, and he expresses his thanks for their kind hospitality.

#### REFERENCES

- [1] F. D. ARARUNA, *Controlabilidade Exata do Sistema de Kirchhoff como Limite do Sistema de Mindlin–Timoshenko*, Ph.D. thesis, Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 2004.
- [2] C. CASTRO AND E. ZUAZUA, *Concentration and lack of observability of waves in highly heterogeneous media*, Arch. Rational Mech. Anal., 164 (2002), pp. 39–72.
- [3] C. CASTRO AND E. ZUAZUA, *Low frequency asymptotic analysis of a string with rapidly oscillating density*, SIAM J. Appl. Math., 60 (2000), pp. 1205–1233.
- [4] A. HARAUX, *Séries lacunaires et contrôle semi-interne des vibrations d’une plaque rectangulaire*, J. Math. Pures Appl. (9), 68 (1989), pp. 457–465.
- [5] J. A. INFANTE AND E. ZUAZUA, *Boundary observability for the space semi-discretizations of the 1-D wave equation*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 407–438.
- [6] A. E. INGHAM, *Some trigonometrical inequalities with applications to the theory of series*, Math. Z., 41 (1936), pp. 367–379.
- [7] J. E. LAGNESE AND J. L. LIONS, *Modelling Analysis and Control of Thin Plates*, Rech. Math. Appl. 6, Masson, Paris, 1988.
- [8] J. L. LIONS, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués. Tome I. Contrôlabilité Exacte*, Rech. Math. Appl. 8, Masson, Paris, 1988.
- [9] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non-homogènes et applications*, Vol. 1, Dunod, Gauthier-Villars, Paris, 1968.
- [10] P. LORETI AND V. VALENTE, *Partial exact controllability for spherical membranes*, SIAM J. Control Optim., 35 (1997), pp. 641–653.
- [11] G. P. MENZALA AND E. ZUAZUA, *The beam equation as a limit of a 1-D nonlinear von Kármán model*, Appl. Math. Lett., 12 (1999), pp. 47–52.
- [12] G. P. MENZALA AND E. ZUAZUA, *Timoshenko’s beam equation as limit of a nonlinear one-dimensional von Kármán system*, Proc. Roy. Soc. Edinburgh Sect. A, 130 (2000), pp. 855–875.
- [13] G. P. MENZALA AND E. ZUAZUA, *Timoshenko’s plate equation as a singular limit of the dynamical Von Kármán system*, J. Math. Pures Appl. (9), 79 (2000), pp. 73–94.
- [14] S. MICU AND E. ZUAZUA, *Boundary controllability of a linear hybrid system arising in the control of noise*, SIAM J. Control Optim., 35 (1997), pp. 1614–1637.

- [15] E. ZUAZUA, *Exact controllability for the semilinear wave equation in one space dimension*, Ann. Inst. H. Poincaré Anal. Non-Linéaire, 10 (1993), pp. 109–129.
- [16] E. ZUAZUA, *Observability of 1-D waves in heterogeneous and semi-discrete media*, in Advances in Structural Control, J. Rodellar, A. Barbat, and F. Casciati, eds., CIMNE, Barcelona, 1999, pp. 1–30.
- [17] E. ZUAZUA, *Propagation, observation, and control of waves approximated by finite difference methods*, SIAM Rev., 47 (2005), pp. 197–243.

## AN INTRINSIC BEHAVIORAL APPROACH TO THE GAP METRIC\*

WENMING BIAN<sup>†</sup>, MARK FRENCH<sup>†</sup>, AND HARISH K. PILLAI<sup>‡</sup>

**Abstract.** An intrinsic trajectory level approach without any recourse to an algebraic structure of a representation is utilized to develop a behavioral approach to robust stability. In particular it is shown how the controllable behavior can be constructed at the trajectory level via Zorn's lemma, and this is utilized to study the controllable-autonomous decomposition. Stability concepts are defined, and the relation between this framework and the well-known difficulties of classical input-output approaches to systems over the doubly infinite time axis are discussed. The gap distance is generalized to the behavioral setting via a trajectory level definition; and a basic robust stability theorem is established for linear shift invariant behaviors. The robust stability theorem is shown to provide an explicit robustness interpretation to the behavioral  $\mathcal{H}^\infty$  synthesis of Willems and Trentelmann.

**Key words.** linear shift invariant behaviors, maximal controllable subbehavior, gap metric, robust stability

**AMS subject classifications.** 93B05, 93D09, 93D25, 93C05

**DOI.** 10.1137/060656681

**1. Introduction.** We begin by observing that the graph topology with its various metrizations plays a fundamental role in the theory of robust stability for classical linear time invariant systems [1, 4, 19]. The contribution of this paper is to develop the basic theory of robust stability involving the gap distance directly from a behavioral perspective, observing that recent approaches to generalizations of the gap metric [4] have been purely trajectory based and hence are easily amenable to such an approach. There has been previous interest in developing behavioral notions of the gap metric; see, e.g., [11] for an example.

From a behavioral point of view [9, 14, 15, 16], the approach is especially fundamental. Much has been made of the intrinsic nature of behavioral definitions and the need for “representation-free” approaches. In this note, we do not have recourse to representations at all; indeed all proofs are at the intrinsic trajectory level and are not restricted, for example, to differential systems. This gives this paper a different “flavor” to much of the recent behavioral literature which is predominately of an algebraic nature. We illustrate our results by considering a system with a pair of (in general) noncommensurate delays: such a system class falls out of the scope of the existing algebraic techniques of the existing behavioral theory for delay systems where, to achieve an algebraic structure, delays are assumed to be commensurate [5].

There are two interrelated reasons for this approach. The first is mathematical: only a limited set-theoretic/analytic structure is required to obtain the required results; hence it is inappropriate to utilize any further structures (e.g., of an algebraic type); this in turn yields greater generality. The second is of applied consequence: a powerful robust stability result should impose as little structure as possible on the

---

\*Received by the editors April 7, 2006; accepted for publication (in revised form) March 7, 2008; published electronically June 25, 2008.

<http://www.siam.org/journals/sicon/47-4/65668.html>

<sup>†</sup>Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, United Kingdom (wb@ecs.soton.ac.uk, mcf@ecs.soton.ac.uk).

<sup>‡</sup>Department of Electrical Engineering, Indian Institute of Technology, Bombay 400076, India (hp@ee.iitb.ac.in).



structure of the perturbations permitted: the perturbed systems may not be representable by differential systems; and they may, for example, arise as a delay or distributed system or even defy direct representation. A set-theoretic treatment of the perturbed system is therefore the appropriate treatment.

Our basic robust stability theorem provides a self-contained basis for the robustness interpretation of the behavioral  $\mathcal{H}^\infty$  results in [17, 12]. The composition of the set-theoretic/analytic treatment in this paper of the resulting perturbed systems is thus set against the controller synthesis for the nominal system, which appropriately is in the context of systems with greater structure (i.e., differential systems).

In relation to the classical approaches, we remark that the standard  $\mathcal{H}^2$  gap is a metric on transfer functions and does not directly apply to systems which either have nonzero initial conditions or which are not minimal (i.e., have uncontrollable modes). The  $\nu$ -gap [13] metric also induces the graph topology on transfer functions and can handle nonzero initial conditions at zero by its definition on the doubly infinite time axis. However, the  $\nu$ -gap is also directly applicable only to controllable systems. By defining systems to be limits of Cauchy sequences in the graph topology [13], the standard gap approaches can also be extended to nonminimal cases; a contribution of this paper from a classical perspective is to provide an alternate and slightly more general approach to these cases. The trajectory formulation considered is also directly applicable to infinite-dimensional systems both in the context of the nominal and perturbed plant and controller: for example delay-differential systems are directly handled, compared to the less direct classical techniques (e.g., the Cauchy sequence approach). We observe also that within the classical framework there has been a move towards representation-free approaches to the gap, e.g., especially for approaches to nonlinear systems [4]. The behavioral approach considered here is one natural extension of this viewpoint.

We emphasize that the main contribution is not in the minor increase in scope of the resulting theorem (as discussed above); rather the contribution is in an alternative and direct derivation of the results within a behavioral framework. Furthermore, we argue that the behavioral approach taken overcomes the well-known difficulties [3] of input-output systems theories on the doubly infinite time axis.

The paper is structured as follows. In section 3 we introduce and review fundamental definitions in the behavioral setting and develop an intrinsic set-theoretic approach to the controllable part of a behavior. Section 4 discusses the controllable-autonomous decomposition, again from a purely trajectory level viewpoint. Section 5 introduces stability definitions and the notion of a closed loop interconnection and considers the concept of stabilizability. From this intrinsic trajectory level framework, two independent issues are discussed. Namely, in section 6 we consider the Georgiou–Smith double time axis “paradox” and show that the Georgiou–Smith example can be satisfactorily treated within the framework considered. Second, in section 7, we utilize the developed framework to develop a fundamental robust stability result. Section 8 shows how this result provides an appropriate interpretation of the robustness guarantee achieved by the behavioral  $\mathcal{H}_\infty$  synthesis of Willems and Trentelmann, and section 9 concludes with an example.

**2. An intrinsic approach to controllable subbehaviors.** Let  $\mathcal{T}$  denote the time set, taken throughout to be either  $\mathbb{Z}$  or  $\mathbb{R}$ , and let  $\mathcal{T}_+ = \mathbb{N}$  if  $\mathcal{T} = \mathbb{Z}$  and  $\mathcal{T}_+ = \mathbb{R}_+$  if  $\mathcal{T} = \mathbb{R}$ . An interval, say  $[a, b]$ , is understood as  $[a, b] = \{t \in \mathcal{T}, a \leq t \leq b\}$ . For  $n \geq 1$ , let  $\text{map}(\mathcal{T}, \mathbb{R}^n)$  be the set of all maps from  $\mathcal{T}$  to  $\mathbb{R}^n$ . An  $n$ -valued behavior  $\mathfrak{B}$  is a subset of  $\text{map}(\mathcal{T}, \mathbb{R}^n)$ , i.e.,  $\mathfrak{B} \subset \text{map}(\mathcal{T}, \mathbb{R}^n)$ . The shift operator  $\sigma_t$ ,  $t \in \mathcal{T}$ , is

defined:  $\sigma_t w(\cdot) = w(\cdot + t)$ .

DEFINITION 2.1. *Let  $\mathfrak{B}$  be a behavior. Then*

1.  *$\mathfrak{B}$  is said to be linear if  $\mathfrak{B}$  is a vector space;*
2.  *$\mathfrak{B}$  is said to be shift invariant (time invariant) if  $w \in \mathfrak{B}$  implies  $\sigma_t w \in \mathfrak{B}$  for all  $t \in \mathcal{T}$ .*

Smooth differential behaviors are linear, shift invariant, continuous-time behaviors which can be expressed as the kernel of a differential operator, i.e., those for which there exists a polynomial valued matrix  $R$  such that

$$(2.1) \quad \mathfrak{B} = \left\{ w \in C^\infty \mid R \left( \frac{d}{dt} \right) w = 0 \right\}.$$

Equivalently in the discrete-time setting, the operator is that of unit shifts:

$$(2.2) \quad \mathfrak{B} = \{ w \in \text{map}(\mathbb{Z}, \mathbb{R}^n) \mid R(\sigma_1) w = 0 \},$$

and such behaviors are called difference behaviors.

Observe that in this note we will be interested in nondifferential/difference behaviors, for example, systems incorporating a time delay.

DEFINITION 2.2. *A behavior  $\mathfrak{B}$  is said to have memory  $l \geq 0$  if for any  $w_1, w_2 \in \mathfrak{B}$  with  $w_1|_{[a, a+l]} = w_2|_{[a, a+l]}$  and  $a \in \mathcal{T}$  the trajectory*

$$w_3(t) = \begin{cases} w_1(t) & \text{if } t \leq a, \\ w_2(t) & \text{if } t \geq a \end{cases}$$

*also lies in  $\mathfrak{B}$ .*

Clearly, a shift invariant behavior  $\mathfrak{B}$  has memory  $l \geq 0$  if and only if for any  $w_1, w_2 \in \mathfrak{B}$  with  $w_1|_{[0, l]} = w_2|_{[0, l]}$  the trajectory

$$(2.3) \quad w_3(t) = \begin{cases} w_1(t) & \text{if } t \leq 0, \\ w_2(t) & \text{if } t \geq 0 \end{cases}$$

*also lies in  $\mathfrak{B}$ .*

If a behavior has memory  $0 \leq l < \infty$ , it is said to have finite memory; if  $l = 0$ , then it is memoryless. Note that a nonmemoryless continuous-time differential behavior has finite memory, and  $l > 0$  can be taken to be arbitrarily small; a discrete time behavior also has finite memory, and here  $l \geq 0$  depends on the system order. The minimal memory  $l_0 \geq 0$  of a behavior  $\mathfrak{B}$  is the largest number such that  $\mathfrak{B}$  has memory  $l$  for all  $l > l_0$ . Note that the minimum is not necessarily attained.

The standard definition of autonomy is that behavior  $\mathfrak{B}$  is said to be autonomous if, for any  $w_1, w_2 \in \mathfrak{B}$ ,  $w_1|_{(-\infty, 0]} = w_2|_{(-\infty, 0]}$  implies  $w_1 = w_2$ . Note that as far as differential systems are concerned, autonomous behaviors have finite memory. We therefore relax the definition for autonomy as follows.

DEFINITION 2.3. *A behavior  $\mathfrak{B}$  is said to be autonomous if there exists  $0 \leq l_0 < \infty$  such that, for any  $w_1, w_2 \in \mathfrak{B}$  and any interval  $V$  of length greater than  $l_0$ ,  $w_1|_V = w_2|_V$  implies  $w_1 = w_2$ .*

Nonautonomy of a behavior is thus just the existence of a trajectory in the behavior whose support has a complement containing an interval of length greater than  $l_0$ , e.g., a compactly supported trajectory.

It should be observed that it is possible that if  $\mathfrak{B}$  has minimal memory  $l_0$ , then, e.g., an autonomous subbehavior can have a minimal memory (i)  $l = 0$ , (ii)  $0 < l < l_0$ , or (iii)  $l = l_0$ ; for example, consider the following:

$$(2.4) \quad \dot{y}(t) = ay(t - 2\tau) + by(t - \tau) + cy(t) + du(t - 2\tau);$$

then  $l_0 = 2\tau$  and the autonomous behavior  $\mathfrak{B}_{u=0}$  corresponds to (i) if  $a = b = 0$ , (ii) if  $a = c = 0, b \neq 0$ , and (iii) if  $b = c = 0, a \neq 0$ .

The behavioral notion of controllability is defined in [9] for differential behaviors. However, this definition is restrictive since it assumes shift invariance of the behavior concerned. We therefore give a modified definition for behavioral controllability that is applicable to more general behaviors but coincides with the notion of controllability in [9] for the case of shift invariant behaviors.

**DEFINITION 2.4.** *A behavior  $\mathfrak{B}$  is said to be controllable if, given  $w_1, w_2 \in \mathfrak{B}$  and  $s \in \mathcal{T}$ , there exist  $w_3 \in \mathfrak{B}$  and  $\tau \in \mathcal{T}_+$  such that*

$$(2.5) \quad w_3(t) = \begin{cases} w_1(t) & \text{if } t \leq s, \\ w_2(t) & \text{if } t \geq s + \tau. \end{cases}$$

This definition requires that the concatenating function  $w_3$  lies in  $\mathfrak{B}$ . This can be hard to guarantee in certain cases, including generalizations to multidimensional systems. Therefore we next introduce the notion of  $\mathfrak{B}$ -controllability.

**DEFINITION 2.5.** *Given a behavior  $\mathfrak{B}$ , a subbehavior  $\mathfrak{B}^* \subset \mathfrak{B}$  is said to be  $\mathfrak{B}$ -controllable if for all  $w_1, w_2 \in \mathfrak{B}^*$  and  $s \in \mathcal{T}$  there exist  $w_3 \in \mathfrak{B}$  and  $\tau \in \mathcal{T}_+$  such that*

$$(2.6) \quad w_3(t) = \begin{cases} w_1(t) & \text{if } t \leq s, \\ w_2(t) & \text{if } t \geq s + \tau. \end{cases}$$

We remark that if  $\mathfrak{B}, \mathfrak{B}^*$  are both shift invariant behaviors and  $\mathfrak{B}^* \subset \mathfrak{B}$ , then  $\mathfrak{B}^*$  is  $\mathfrak{B}$ -controllable if and only if, given any  $w_1, w_2 \in \mathfrak{B}^*$ , there exist  $w_3 \in \mathfrak{B}$  and  $\tau \in \mathcal{T}_+$  such that

$$(2.7) \quad w_3(t) = \begin{cases} w_1(t) & \text{if } t \leq 0, \\ w_2(t) & \text{if } t \geq \tau. \end{cases}$$

If  $\mathfrak{B}^* = \mathfrak{B}$ , then the  $\mathfrak{B}$ -controllability of  $\mathfrak{B}^*$  is the same as the controllability defined by Definition 2.4. So controllability in the sense of Definition 2.4 implies  $\mathfrak{B}$ -controllability. But the following example shows that the converse does not hold.

*Example.* Let  $\mathfrak{B} = C^\infty(\mathbb{R}, \mathbb{R})$  and  $\mathfrak{B}^*$  be the set of all constant functions. Then both  $\mathfrak{B}$  and  $\mathfrak{B}^*$  are linear shift invariant behaviors with finite memory and  $\mathfrak{B}^* \subset \mathfrak{B}$ . It is straightforward to check that  $\mathfrak{B}^*$  is  $\mathfrak{B}$ -controllable. However, it is neither  $\mathfrak{B}^*$ -controllable nor controllable in the sense of Definition 2.4. Similarly, let  $\mathfrak{B} = C^0(\mathbb{R}, \mathbb{R})$ . Then the subbehavior  $\mathfrak{B}^* = \text{span}\{e^t\}$  is  $\mathfrak{B}$ -controllable, but it is not  $\mathfrak{B}^*$ -controllable.

We now consider the properties of  $\mathfrak{B}$ -controllable behaviors.

**LEMMA 2.6.** *Suppose  $\mathfrak{B}$  is a behavior. Then there exists at least one maximal  $\mathfrak{B}$ -controllable subbehavior.*

*Proof.* First, any behavior has at least one  $\mathfrak{B}$ -controllable subbehavior, i.e.,  $\{0\}$ . Second, set inclusion defines a partial order on the set of all  $\mathfrak{B}$ -controllable subbehaviors. Any chain of  $\mathfrak{B}$ -controllable subbehaviors,

$$(2.8) \quad \mathfrak{B}_\alpha \subset \mathfrak{B}_\beta \subset \cdots \subset \mathfrak{B}_\gamma \subset \cdots,$$

where  $\alpha, \beta, \gamma \cdots \in \Gamma$  and  $\Gamma$  is the index set, has an upper bound:

$$(2.9) \quad \mathfrak{B}_\alpha \subset \bigcup_{\beta \in \Gamma} \mathfrak{B}_\beta = \mathfrak{B}^*.$$

$\mathfrak{B}^* \subset \mathfrak{B}$  is  $\mathfrak{B}$ -controllable since, given any  $w_1, w_2 \in \mathfrak{B}^*$ , we have  $w_1 \in B_\alpha$ ,  $w_2 \in \mathfrak{B}_\beta$  for some  $\alpha, \beta \in \Gamma$ , hence  $w_1, w_2 \in \mathfrak{B}_\gamma$ ,  $\gamma = \max\{\alpha, \beta\}$ , and by the  $\mathfrak{B}$ -controllability of  $\mathfrak{B}_\gamma$  it follows that there exists  $w_3 \in \mathfrak{B}$  satisfying (2.6) as required. Zorn's lemma then gives the existence of a maximal  $\mathfrak{B}$ -controllable subbehavior as required.  $\square$

Note that this set-theoretic construction is extremely general: we do not require any linearity, memory, or differential/difference structure on  $\mathfrak{B}$ . In general, maximal  $\mathfrak{B}$ -controllable subbehaviors are not unique. However, if the behavior  $\mathfrak{B}$  is linear, then there exists a unique maximal  $\mathfrak{B}$ -controllable linear subbehavior.

**THEOREM 2.7.** *Suppose  $\mathfrak{B}$  is a linear behavior. Then there exists a unique maximal linear  $\mathfrak{B}$ -controllable subbehavior.*

*Proof.* We consider the set of all linear  $\mathfrak{B}$ -controllable subbehaviors. With the relation induced by subset inclusion, this set is partially ordered and a maximal subbehavior  $\mathfrak{B}_{\text{cont}}$  exists which is also linear.

To show the uniqueness, let  $\mathfrak{B}_1$  be another linear maximal  $\mathfrak{B}$ -controllable subbehavior, and let  $\mathfrak{B}_2$  denote the linear span of  $\mathfrak{B}_{\text{cont}}$  and  $\mathfrak{B}_1$ :  $\mathfrak{B}_2 = \mathfrak{B}_{\text{cont}} + \mathfrak{B}_1$ . Let  $w_1, w_2 \in \mathfrak{B}_2$ . Without loss of generality, we may suppose that  $w_i = \alpha_i x_i + \beta_i y_i$  with  $\alpha_i, \beta_i \in \mathbb{R}$ ,  $x_i \in \mathfrak{B}_{\text{cont}}$ ,  $y_i \in \mathfrak{B}_1$ , and  $i = 1, 2$ . Since  $0 \in \mathfrak{B}_{\text{cont}} \cap \mathfrak{B}_1$  and by the definition of  $\mathfrak{B}$ -controllability, we have the following: for all  $s \in \mathcal{T}$ , there exist  $\tau_1, \tau_2 > 0$  and  $z_1, v_1 \in \mathfrak{B}$  such that  $z_1|_{(-\infty, s]} = x_1|_{(-\infty, s]}$ ,  $z_1|_{[s+\tau_1, \infty)} = 0|_{[s+\tau_1, \infty)}$ ,  $v_1|_{(-\infty, s]} = y_1|_{(-\infty, s]}$ ,  $v_1|_{[s+\tau_2, \infty)} = 0|_{[s+\tau_2, \infty)}$ . Let  $\tau_3 = \max\{\tau_1, \tau_2\}$  and  $w_3 = \alpha_1 z_1 + \beta_1 v_1 \in \mathfrak{B}$ . Then we have  $w_3|_{(-\infty, s]} = \alpha_1 x_1|_{(-\infty, s]} + \beta_1 y_1|_{(-\infty, s]} = w_1|_{(-\infty, s]}$  and  $w_3|_{[s+\tau_3, \infty)} = 0|_{[s+\tau_3, \infty)}$ . This shows that  $w_1$  is switched to 0 in  $\mathfrak{B}$ . Similarly we can prove that there exist  $\tau_4 > 0, w_4 \in \mathfrak{B}$  such that  $w_4|_{(-\infty, s]} = 0|_{(-\infty, s]}$  and  $w_4|_{[s+\tau_4, \infty)} = w_2|_{[s+\tau_4, \infty)}$ . Write  $\tau_5 = \max\{\tau_3, \tau_4\}$ ,  $w_5 = w_3 + w_4$ . Then we see that  $w_5 \in \mathfrak{B}$ ,  $w_5|_{(-\infty, s]} = w_1|_{(-\infty, s]}$  and  $w_5|_{[s+\tau_5, \infty)} = w_2|_{[s+\tau_5, \infty)}$ . This shows that  $\mathfrak{B}_2$  is  $\mathfrak{B}$ -controllable. The maximality of  $\mathfrak{B}_{\text{cont}}$  along with  $\mathfrak{B}_{\text{cont}} \subset \mathfrak{B}_2$  implies that the two behaviors are the same. Hence  $\mathfrak{B}_{\text{cont}}$  is unique.  $\square$

In the rest of this paper, for a linear behavior  $\mathfrak{B}$ , we always use  $\mathfrak{B}_{\text{cont}}$  to denote its unique maximal  $\mathfrak{B}$ -controllable subbehavior.

The above proof also shows that the sum of any two linear  $\mathfrak{B}$ -controllable subbehaviors is  $\mathfrak{B}$ -controllable and, therefore, so is the sum of all linear  $\mathfrak{B}$ -controllable subbehaviors. Hence

$$\mathfrak{B}_{\text{cont}} = \text{span}\{B \subset \mathfrak{B} : B \text{ is linear and } \mathfrak{B}\text{-controllable}\}.$$

Next we show that shift invariance is preserved for the unique maximal  $\mathfrak{B}$ -controllable subbehaviors. In particular, in the shift invariant linear setting,  $\mathfrak{B}_{\text{cont}}$  is linear and shift invariant.

**LEMMA 2.8.** *Suppose  $\mathfrak{B}$  is shift invariant and has a unique maximal  $\mathfrak{B}$ -controllable subbehavior  $\mathfrak{B}^*$ . Then  $\mathfrak{B}^*$  is shift invariant.*

*Proof.* Let  $r, s \in \mathcal{T}$ ,  $\sigma_r w_1, \sigma_r w_2 \in \sigma_r \mathfrak{B}^*$  with  $w_1, w_2 \in \mathfrak{B}^*$ . Then there exist  $w_3 \in \mathfrak{B}$  and  $\tau > 0$  such that  $w_3(t) = w_1(t)$  for  $t \leq s + r$  and  $w_3(t) = w_2(t)$  for  $t \geq s + r + \tau$ . Hence

$$\sigma_r w_3(t) = \begin{cases} \sigma_r w_1(t) & \text{if } t \leq s, \\ \sigma_r w_2(t) & \text{if } t \geq s + \tau. \end{cases}$$

Since  $\sigma_r w_3 \in \sigma_r \mathfrak{B} = \mathfrak{B}$ , we see that  $\sigma_r \mathfrak{B}^*$  is  $\mathfrak{B}$ -controllable and hence  $\sigma_r \mathfrak{B}^* \subset \mathfrak{B}^*$  as  $\mathfrak{B}^*$  is the unique maximal  $\mathfrak{B}$ -controllable subbehavior.  $\square$

**COROLLARY 2.9.** *Suppose  $\mathfrak{B}$  is linear and shift invariant. Then  $\mathfrak{B}_{\text{cont}}$  is linear and shift invariant.*

We conclude this section by showing that  $\mathfrak{B}$ -controllable linear subbehaviors inherit memory properties from the original behavior  $\mathfrak{B}$ .

**LEMMA 2.10.** *Let  $\mathfrak{B}$  be a linear shift invariant behavior with finite memory  $l \geq 0$ . Then  $\mathfrak{B}_{\text{cont}}$  has memory  $l \geq 0$ .*

*Proof.* First, we need a new notion: a subbehavior  $\mathfrak{B}^*$  is 0- $\mathfrak{B}$ -controllable if it is shift invariant and, given any  $w_1, w_2 \in \mathfrak{B}^*$ , there exist  $w_3 \in \mathfrak{B}$  and  $\tau \in \mathcal{T}_+$  satisfying (2.7). Using the same procedure as used in Theorem 2.7, we can see that a maximal linear 0- $\mathfrak{B}$ -controllable subbehavior of  $\mathfrak{B}$  exists, denoted by  $\mathfrak{B}_{\text{cont}}^0$ , which is shift invariant. By the remark following Definition 2.5,  $\mathfrak{B}_{\text{cont}} = \mathfrak{B}_{\text{cont}}^0$ .

Now let  $w_1, w_2 \in \mathfrak{B}_{\text{cont}}$  with  $w_1|_{[0,l]} = w_2|_{[0,l]}$ . Then

$$w_3(t) = \begin{cases} w_1(t) & \text{if } t < 0 \\ w_2(t) & \text{if } t \geq 0 \end{cases} \in \mathfrak{B}.$$

Since  $\mathfrak{B}_{\text{cont}}$  is  $\mathfrak{B}$ -controllable, for any  $w \in \mathfrak{B}_{\text{cont}}$ , there exist  $\tau_1 > 0$  and  $v_1 \in \mathfrak{B}$  such that  $v_1|_{(-\infty,0]} = w|_{(-\infty,0]}$  and  $v_1|_{[\tau_1,\infty)} = w|_{[\tau_1,\infty)}$ ; that is,  $w_3$  can be switched to  $w$ . Similarly, there exist  $\tau_2 > 0$  and  $v_2 \in \mathfrak{B}$  such that  $v_2|_{(-\infty,0]} = w|_{(-\infty,0]}$  and  $v_2|_{[\tau_2,\infty)} = w_2|_{[\tau_2,\infty)} = w_3|_{[\tau_2,\infty)}$ ; that is,  $w$  can be switched to  $w_3$ . Hence  $\text{span}\{w_3\} + \mathfrak{B}_{\text{cont}}^0$  is 0- $\mathfrak{B}$ -controllable. By its maximality  $w_3 \in \mathfrak{B}_{\text{cont}}^0$  and therefore  $w_3 \in \mathfrak{B}_{\text{cont}}$  as shown above. This completes the proof.  $\square$

We remark that all results in this section remain valid if the  $\mathfrak{B}$ -controllability is replaced by the controllability defined in Definition 2.4.

Given a linear differential behavior  $\mathfrak{B}$ , a unique maximal controllable (as per Definition 2.4) subbehavior exists, denoted by  $\mathfrak{B}_c$ . Since this controllable subbehavior is  $\mathfrak{B}$ -controllable,  $\mathfrak{B}_c \subset \mathfrak{B}_{\text{cont}}$ . The next lemma shows that for certain behaviors (that can be represented as kernels of certain classes of operator) the two notions of controllability (Definitions 2.4 and 2.5) coincide. In particular, for differential/difference behaviors,  $\mathfrak{B}_{\text{cont}}$  equals  $\mathfrak{B}_c$ .

We introduce the following notation. Let  $R : \text{dom}(R) \rightarrow \text{im}(R) \subset \text{map}(\mathcal{T}, \mathbb{R}^{n_2})$  be an operator, where  $\text{dom}(R) \subset \text{map}(\mathcal{T}, \mathbb{R}^{n_1})$  is the domain of  $R$  and  $\text{im}(R) \subset \text{map}(\mathcal{T}, \mathbb{R}^{n_2})$  denotes the image of  $R$ .

**DEFINITION 2.11.** *Let  $R : \text{dom}(R) \rightarrow \text{im}(R) \subset \text{map}(\mathcal{T}, \mathbb{R}^{n_2})$  be an operator, where  $\text{dom}(R) \subset \text{map}(\mathcal{T}, \mathbb{R}^{n_1})$ .  $R$  is said to have local action if there exists  $k_2 \geq k_1 > 0$  such that, for all  $t \in \mathcal{T}$ ,  $w_1, w_2 \in \text{dom}(R)$  with  $w_1|_{[t-k_2, t+k_2]} = w_2|_{[t-k_2, t+k_2]}$  implies  $(Rw_1)|_{[t-k_1, t+k_1]} = (Rw_2)|_{[t-k_1, t+k_1]}$ .*

Differential operators, (backward) difference operators, and delay-differential operators all have local action property. For differential operators, the constants  $k_1, k_2$  can be as small as possible. For difference/delay operators,  $k_2 - k_1$  should be greater than or equal to the maximum of differences/delays involved.

**LEMMA 2.12.** *Let  $\mathfrak{B}$  be a linear behavior and  $\mathfrak{B}_c$  be the maximal controllable subbehavior of  $\mathfrak{B}$  (as per Definition 2.4). Let  $R : \text{dom}(R) \rightarrow \text{map}(\mathcal{T}, \mathbb{R}^{n_1})$ ,  $R_c : \mathfrak{B} \rightarrow \text{map}(\mathcal{T}, \mathbb{R}^{n_2})$  and  $S : \text{im}(R_c) \rightarrow \text{map}(\mathcal{T}, \mathbb{R}^{n_1})$  be linear operators, where  $\mathfrak{B} \subset \text{dom}(R) \subset \text{map}(\mathcal{T}, \mathbb{R}^n)$ . Suppose*

- (i)  $R = SR_c$ ;
- (ii)  $\mathfrak{B} = \{w \in \text{dom}(R), Rw = 0\}$  and  $\mathfrak{B}_c = \{w \in \mathfrak{B} : R_c w = 0\}$ ;
- (iii) the behavior  $\mathfrak{B}_S = \{w \in \text{im}(R_c) : Sw = 0\}$  is autonomous; and

(iv)  $R_c$  has local action.

Then  $\mathfrak{B}_{\text{cont}} = \mathfrak{B}_c$ .

*Proof.* Let

$$\mathfrak{B}_1 = \left\{ w \in \mathfrak{B} : \begin{array}{l} \text{for any } s \in \mathcal{T}, \text{ there exist } \tau_1, \tau_2 \geq 0, v_1, v_2 \in \mathfrak{B} \text{ such that} \\ v_1|_{(-\infty, s]} = w|_{(-\infty, s]} \text{ and } v_1|_{[s+\tau_1, \infty)} = 0|_{[s+\tau_1, \infty)}, \\ v_2|_{(-\infty, s]} = 0|_{(-\infty, s]} \text{ and } v_2|_{[s+\tau_2, \infty)} = w|_{[s+\tau_2, \infty)}. \end{array} \right\}.$$

We claim that  $\mathfrak{B}_{\text{cont}} = \mathfrak{B}_1$  and  $\mathfrak{B}_1 = \mathfrak{B}_c$ , so  $\mathfrak{B}_{\text{cont}} = \mathfrak{B}_c$ .

We first prove  $\mathfrak{B}_{\text{cont}} = \mathfrak{B}_1$ . Since  $0 \in \mathfrak{B}_{\text{cont}}$ , any  $w \in \mathfrak{B}_{\text{cont}}$  can be patched to 0 in  $\mathfrak{B}$  and vice versa. This shows that  $\mathfrak{B}_{\text{cont}} \subset \mathfrak{B}_1$ . To show the reverse inclusion, let  $w_1, w_2 \in \mathfrak{B}_1$ . Then for any  $s \in \mathcal{T}$  there exist  $\tau_i \geq 0$  and  $u_i, v_i \in \mathfrak{B}$  ( $i = 1, 2$ ) such that

$$\begin{aligned} u_i|_{(-\infty, s]} &= w_i|_{(-\infty, s]}, & u_i|_{[s+\tau_i, \infty)} &= 0|_{[s+\tau_i, \infty)}, \\ v_i|_{(-\infty, s]} &= 0|_{(-\infty, s]} & \text{and } v_i|_{[s+\tau_i, \infty)} &= w_i|_{[s+\tau_i, \infty)}. \end{aligned}$$

From this it follows that  $u_1 + v_2$  patches  $w_1$  to  $w_2$  and  $u_2 + v_1$  patches  $w_2$  to  $w_1$ . So  $\text{span}\{w_1, w_2\}$  is  $\mathfrak{B}$ -controllable and therefore  $w_1, w_2 \in \mathfrak{B}_{\text{cont}}$ , which proves that  $\mathfrak{B}_1 \subset \mathfrak{B}_{\text{cont}}$ .

We now prove  $\mathfrak{B}_1 = \mathfrak{B}_c$ . Since  $\mathfrak{B}_c \subset \mathfrak{B}_{\text{cont}} = \mathfrak{B}_1$ , we need only prove  $\mathfrak{B}_1 \subset \mathfrak{B}_c$ . Suppose it is not the case. Then there exists  $w \in \mathfrak{B}_1 \setminus \mathfrak{B}_c$  such that  $R_c w \neq 0$ ; that is, there exists  $t_0 \in \mathcal{T}$  with  $R_c w(t_0) \neq 0$ . Since  $R_c$  has local action, there exists  $k_2 > k_1 > 0$  such that for all  $t \in \mathcal{T}$

$$w_1, w_2 \in \mathfrak{B} \quad \text{with } w_1|_{[t-k_2, t+k_2]} = w_2|_{[t-k_2, t+k_2]}$$

implies

$$(R_c w_1)|_{[t-k_1, t+k_1]} = (R_c w_2)|_{[t-k_1, t+k_1]}.$$

Let  $s \in \mathcal{T}$  such that  $s > t_0 + k_2$ . By the definition of  $\mathfrak{B}_1$ , there exist  $\tau > 0$  and  $w_s \in \mathfrak{B}$  such that

$$w_s|_{(-\infty, s]} = w|_{(-\infty, s]} \quad \text{and} \quad w_s|_{[s+\tau, \infty)} = 0|_{[s+\tau, \infty)}.$$

Let  $v_s = R_c w_s$ . Since  $w_s \in \mathfrak{B} = \{w : R_c w = 0\}$ , by assumption (i),  $Sv_s = SR_c w_s = R_c w_s = 0$  and therefore  $v_s \in \mathfrak{B}_S$ . Since  $R_c$  has local action and  $w_s|_{[s+\tau, \infty)} = 0|_{[s+\tau, \infty)}$ , there exist  $a, b \geq s + \tau, a < b$  such that  $v_s|_{[a, b]} = R_c w_s|_{[a, b]} = R_c 0|_{[a, b]} = 0|_{[a, b]}$ . By the autonomy of  $\mathfrak{B}_S$  (choosing  $a, b$  such that  $b - a > \text{minimum memory of } \mathfrak{B}_S$ ),  $v_s \equiv 0$ , i.e.,  $R_c w_s = 0$ . Since  $w_s|_{(-\infty, s]} = w|_{(-\infty, s]}$  and  $t_0 < s - k_2$ , we see that  $w_s|_{[t_0-k_2, t_0+k_2]} = w|_{[t_0-k_2, t_0+k_2]}$ . Therefore, by the local action assumption on  $R_c$ ,  $(R_c w)(t_0) = (R_c w_s)(t_0) = 0$ . This is a contradiction and completes the proof.  $\square$

We may now apply this lemma to linear differential, delay-differential, and difference behaviors, that is, behaviors defined by systems of differential/delay-differential/difference equations. As shown in [5, 9], those behaviors are kernels of linear operators governed by matrices of polynomials.

**THEOREM 2.13.** *For a differential/delay-differential/difference behavior  $\mathfrak{B}$ , its maximal  $\mathfrak{B}$ -controllable subbehavior  $\mathfrak{B}_{\text{cont}}$  is the same as its maximal controllable subbehavior  $\mathfrak{B}_c$ .*

*Proof.* We first suppose that  $\mathfrak{B}$  is a differential behavior. Both  $\mathfrak{B}$  and  $\mathfrak{B}_c$  have the representations

$$\mathfrak{B} = \left\{ w \in \mathfrak{C}^\infty : R \left( \frac{d}{dt} \right) w = 0 \right\}, \quad \mathfrak{B}_c = \left\{ w \in \mathfrak{B} : R_c \left( \frac{d}{dt} \right) w = 0 \right\},$$

where  $R(\xi), R_c(\xi)$  are  $m \times n$  matrices of polynomials of  $\xi$ . Since  $\mathfrak{B}_c \subset \mathfrak{B}$ , there exists a nonsingular polynomial matrix  $S$  such that  $R = SR_c$ . Moreover, the kernel of  $S$  is an autonomous behavior. So all assumptions of Lemma 2.12 hold and  $\mathfrak{B}_{\text{cont}} = \mathfrak{B}_c$ .

If  $\mathfrak{B}$  is a delay-differential behavior, as shown in [5], the proof is almost the same except that the matrix operators are  $R(\frac{d}{dt}, \sigma), R_c(\frac{d}{dt}, \sigma)$  with  $\sigma$  the delay operation. Where no differentiation operators are present, we obtain the proof for the difference behavior case.  $\square$

Earlier, we have shown that

$$\mathfrak{B}_{\text{cont}} = \Sigma \{ B \subset \mathfrak{B} : B \text{ is linear and } \mathfrak{B}\text{-controllable} \}.$$

Since  $\mathfrak{B}_c = \mathfrak{B}_{\text{cont}}$  for a differential/difference behavior  $\mathfrak{B}$ , this gives us a direct set-theoretic construction of  $\mathfrak{B}_c$ . To the best of the authors' knowledge, this direct set-theoretic construction of  $\mathfrak{B}_c$  does not appear in the literature. Within the behavioral literature, the controllable subbehavior is typically constructed algebraically given the equations governing the behavior, and it is shown via the duality between the behavior and the algebraic structure that the controllable subbehavior is the "largest" such subset. It is noteworthy to observe that in some settings (e.g., both one-dimensional (1D) and  $n$ -dimensional ( $n$ D) differential systems), the existence of the corresponding maximal algebraic object appears constructively; see [9, 8].

**3. The autonomous-controllable decomposition.** For 1D differential/difference behaviors  $\mathfrak{B}$  it is well known that  $\mathfrak{B}$  can be split into a direct sum of the controllable and an autonomous part:

$$\mathfrak{B} = \mathfrak{B}_c \oplus \mathfrak{B}_a,$$

where  $\mathfrak{B}_c \subset \mathfrak{B}$  is the maximal controllable subbehavior of  $\mathfrak{B}$  as per Definition 2.4 and  $\mathfrak{B}_a$  is an autonomous subbehavior. Because of Theorem 2.13, we in fact have

$$(3.1) \quad \mathfrak{B} = \mathfrak{B}_{\text{cont}} \oplus \mathfrak{B}_a.$$

This direct sum decomposition is a special feature which holds only for certain classes of systems (such as the differential case [9]). For example, in the  $n$ D differential/difference setting, it is known that this sum is not direct for  $n > 1$  (see [18]), and in the context of delay differential systems it is known only that the sum is direct for commensurate delays [5]. However, since an additive decomposition is critical to what follows, we do not restrict our attention to direct sums but treat it as an important special case.

Therefore this section examines both direct and nondirect sum decompositions at the trajectory level and examines the relationship between autonomy of any direct summand to the  $\mathfrak{B}$ -controllable part and the corresponding lack of controllability of this part. We wish to show that, under certain conditions,  $\mathfrak{B} = \mathfrak{B}_{\text{cont}} \oplus \mathfrak{B}^*$  or  $\mathfrak{B} = \mathfrak{B}_{\text{cont}} + \mathfrak{B}^*$  implies that  $\mathfrak{B}^*$  is autonomous. Behaviors with such decomposition will be studied later on for stability and robustness. We have already established a connection between  $\mathfrak{B}_{\text{cont}}$  and  $\mathfrak{B}_c$  (the maximal controllable subbehavior of  $\mathfrak{B}$  as

per Definition 2.4). Note that in this paper we are considering only behaviors defined over the time set  $\mathcal{T}$  which is either  $\mathbb{Z}$  or  $\mathbb{R}$ .

We begin the study of these decompositions first by considering linear, shift invariant, autonomous behaviors.

LEMMA 3.1. *Let  $\mathfrak{B}$  be linear, shift invariant, autonomous behavior. Let  $\mathfrak{B}_{\text{cont}}$  be the maximal linear  $\mathfrak{B}$ -controllable subbehavior of  $\mathfrak{B}$ . Then  $\mathfrak{B}_{\text{cont}} = \{0\}$ .*

*Proof.* First, it is obvious that  $0 \in \mathfrak{B}_{\text{cont}}$ . Suppose there exists  $w \neq 0, w \in \mathfrak{B}_{\text{cont}}$ . By  $\mathfrak{B}$ -controllability, there exist trajectories  $w_1, w_2 \in \mathfrak{B}$  and  $\tau_1, \tau_2 > 0$  such that  $w_1|_{(-\infty, 0]} = 0, w_1|_{(\tau_1, \infty)} = w|_{(\tau_1, \infty)}$  and  $w_2|_{(-\infty, 0]} = w|_{(-\infty, 0]}, w_2|_{(\tau_2, \infty)} = 0$ . By shift invariance,  $\sigma_{-\tau_2-l_0} w_1 \in \mathfrak{B}$ , where  $l_0$  is the minimum finite memory. Since  $\sigma_{-\tau_2-l_0} w_1|_{[\tau_2, \tau_2+l_0]} = 0 = w_2|_{[\tau_2, \tau_2+l_0]}$ , it follows from the autonomous assumption that  $\sigma_{-\tau_2-l_0} w_1(t) = w_2(t)$  for all  $t \in \mathcal{T}$ . This tells us that  $w_2(t) = 0$  for all  $t \in \mathcal{T}$  and therefore  $w_1 = \sigma_{\tau_2+l_0} w_2 = 0$ . So  $w \equiv 0$ , which is a contradiction.  $\square$

Note that the above lemma is only in one direction.

For any  $V \subset \mathcal{T}$ , let  $P_V$  denote the natural projection (restriction) of signals defined on  $\mathcal{T}$  onto signals defined on  $V$ . As shorthand we write  $P_+$  for  $P_{\mathcal{T}_+}$  and  $P_-$  for  $P_{\mathcal{T} \setminus \mathcal{T}_+}$ .

LEMMA 3.2. (1)  $P_V(\mathfrak{B}_1 + \mathfrak{B}_2) = P_V\mathfrak{B}_1 + P_V\mathfrak{B}_2$  for any two behaviors  $\mathfrak{B}_1, \mathfrak{B}_2$ .

(2) If  $\mathfrak{B}$  is a linear, shift invariant behavior with finite memory  $l > 0$  and  $\mathfrak{B} = \mathfrak{B}_{\text{cont}} \oplus \mathfrak{B}^*$ , then  $P_V\mathfrak{B} = P_V\mathfrak{B}_{\text{cont}} \oplus P_V\mathfrak{B}^*$  for all intervals  $V$  of length greater than  $l$ .

*Proof.* Claim (1) is rather obvious. To establish claim (2), we need only prove that the sum  $P_V\mathfrak{B}_{\text{cont}} \oplus P_V\mathfrak{B}^*$  is direct. Let  $w_1 \in \mathfrak{B}^*, w_2 \in \mathfrak{B}_{\text{cont}}$  such that  $w_1|_V = w_2|_V$ , i.e.,  $w_1|_V = w_2|_V \in P_V\mathfrak{B}_{\text{cont}} \cap P_V\mathfrak{B}^*$ . Consider any  $w_3 \in \mathfrak{B}_{\text{cont}}$ . By the memory property  $w_1$  can be patched to  $w_2$  and conversely, and by  $\mathfrak{B}$ -controllability  $w_2$  can be patched to  $w_3$  and conversely. This tells that  $w_1$  can be patched to  $w_3$  in  $\mathfrak{B}$  and conversely. Hence  $\mathfrak{B}' := \mathfrak{B}_{\text{cont}} + \text{span}(w_1)$  is  $\mathfrak{B}$ -controllable. Since  $\mathfrak{B}_{\text{cont}}$  is the unique maximal  $\mathfrak{B}$ -controllable subbehavior of  $\mathfrak{B}$ , it must contain  $\mathfrak{B}'$ . Hence  $w_1 \in \mathfrak{B}_{\text{cont}}$ . By the direct sum property it follows that  $w_1 = 0$ , and hence  $w_1|_V = w_2|_V = 0$ .  $\square$

PROPOSITION 3.3. *Let  $\mathfrak{B}$  be a linear, shift invariant behavior with finite memory. Then the following are equivalent:*

1.  $\mathfrak{B}_{\text{cont}}$  splits  $\mathfrak{B}$ ; i.e.,  $\mathfrak{B} = \mathfrak{B}_{\text{cont}} \oplus \mathfrak{B}^*$  for some subbehavior  $\mathfrak{B}^* \subset \mathfrak{B}$ .
2. There exists a behavior  $\mathfrak{B}^*$  for which  $(\mathfrak{B}^*)_{\text{cont}} = \{0\}$  (where  $(\mathfrak{B}^*)_{\text{cont}}$  is the maximal linear  $\mathfrak{B}$ -controllable subbehavior of  $\mathfrak{B}^*$ ) and  $\mathfrak{B} = \mathfrak{B}_{\text{cont}} \oplus \mathfrak{B}^*$ .

*Proof.* That 2 implies 1 is obvious. It remains to show that 1 implies 2.

Since  $\mathfrak{B}_{\text{cont}}$  splits  $\mathfrak{B}$ , there exists  $\mathfrak{B}^*$  such that  $\mathfrak{B} = \mathfrak{B}_{\text{cont}} \oplus \mathfrak{B}^*$ . Then  $(\mathfrak{B}^*)_{\text{cont}} \subset \mathfrak{B}_{\text{cont}} \cap \mathfrak{B}^* = \{0\}$ .  $\square$

Note that if one considers the maximal  $\mathfrak{B}^*$ -controllable subbehavior of  $\mathfrak{B}^*$ , instead of  $(\mathfrak{B}^*)_{\text{cont}}$ , then the conclusion of Proposition 3.3 continues to hold. At this juncture, we would like to comment that since Lemma 3.1 is only in one direction, we cannot conclude that  $\mathfrak{B}^*$  in Proposition 3.3 is autonomous. We further observe that if  $\mathfrak{B}_c$  (the controllable part of a behavior  $\mathfrak{B}$  as per Definition 2.4) has finite codimension (as in the differential [10] and commensurate delay [5] settings), then it is known that  $\mathfrak{B}_c$  splits  $\mathfrak{B}$ . In these cases, we know that the summand  $\mathfrak{B}^*$  is autonomous (by the traditional definition).

Unfortunately the autonomy of  $\mathfrak{B}^*$  in the present situation remains a problem. So we introduce the following definition for the rest of this paper.

DEFINITION 3.4. *A behavior  $\mathfrak{B}$  is said to have a controllable-autonomous decomposition if there exists an autonomous subbehavior of  $\mathfrak{B}$ , denoted by  $\mathfrak{B}_{\text{aut}}$ , such*



that

$$\mathfrak{B} = \mathfrak{B}_{\text{cont}} + \mathfrak{B}_{\text{aut}}.$$

**4. Stability.** Stability is determined by the signal spaces involved. We will consider the spaces  $L^p(\mathcal{T}, \mathbb{R}^n)$  with  $0 \leq p \leq \infty$ . In the case when  $\mathcal{T} = \mathbb{R}$ , it is the standard  $L^p$  spaces such as  $L^2(\mathbb{R}, \mathbb{R}^n)$  and  $L^\infty(\mathbb{R}, \mathbb{R}^n)$  for continuous-time signals. In the case when  $\mathcal{T} = \mathbb{Z}$ , it becomes the standard  $l^p$  spaces used for discrete signals.

Given a general normed signal space (say)  $Y$  of signals from  $\mathcal{T}$  or  $\mathcal{T}_+$  to  $\mathbb{R}^n$ , the corresponding extended space  $Y_e$  is defined as

$$Y_e = \{y : \mathcal{I} \rightarrow \mathbb{R}^n : T_\tau y \in Y \text{ for all } \tau \in \mathcal{I}_+\},$$

where  $\mathcal{I} = \mathcal{T}$  or  $\mathcal{T}_+$  subject to on which set the space  $Y$  is defined, and  $T_\tau$  is the truncation operator, that is,  $(T_\tau y)(t) = y(t)$  for  $t \leq \tau$  and 0 for  $t > \tau$ .

In this section, the behaviors considered will be restricted to be within the extended signal spaces  $L_e^p := L_e^p(\mathcal{T}, \mathbb{R}^n)$ ,  $1 \leq p \leq \infty$ , i.e.,  $L^p$  behaviors or subsets of  $L_e^p$ .

As shorthand we write  $X = L^p(\mathcal{T}_+) =: L^p(\mathcal{T}_+, \mathbb{R}^n)$ ,  $1 \leq p \leq \infty$ . So  $X_e = L_e^p(\mathcal{T}_+)$ . We remark that when the results do not need a normed structure on the signal spaces, our discussions and definitions also remain valid for  $C^\infty$  behaviors (with  $X = C^\infty(\mathcal{T}_+)$ ).

We generalize the standard behavioral definition of stability for autonomous systems as follows.

**DEFINITION 4.1.** *An autonomous system  $\mathfrak{B}_{\text{aut}}$  is said to be  $X$ -stable if, for any  $w \in \mathfrak{B}_{\text{aut}}$ ,  $w|_{[0, \infty)} \in X$ .*

This notion of stability can be equivalently expressed as the statement that  $\mathfrak{B}_{\text{aut}}$  is stable if and only if  $P_+ \mathfrak{B}_{\text{aut}} \subset X$ . For nonautonomous systems, we adopt the following stability concept for behaviors with input-output partition (see [9]), which captures the notion of “whatever the past, given a bounded future input, the future output is bounded.”

**DEFINITION 4.2.** *A behavior  $\mathfrak{B}$  with input-output partition  $u|y$  is  $X$ -stable if for all  $(u, y) \in \mathfrak{B}$  with  $u|_{\mathcal{T}_+} \in X$  we have  $y|_{\mathcal{T}_+} \in X$ .*

When  $X$  is given, throughout the paper we refer to the notion of “ $X$ -stability” simply as “stability.”

Associated with any behavior are the stable subbehaviors which correspond to the behavior taking zero values up to time  $t = 0$ .

**DEFINITION 4.3.** *The graph  $\mathcal{G}_{\mathfrak{B}}$  of a behavior  $\mathfrak{B}$  is defined to be*

$$(4.1) \quad \mathcal{G}_{\mathfrak{B}} := \{w \in X \mid \text{there exists } v \in \mathfrak{B} \text{ such that } v|_{\mathcal{T}_-} = 0, v|_{\mathcal{T}_+} = w|_{\mathcal{T}_+}\}.$$

The extended graph  $\mathcal{Z}_{\mathfrak{B}}$  of  $\mathfrak{B}$  is defined to be

$$\mathcal{Z}_{\mathfrak{B}} := \{w \in X_e \mid \text{there exists } v \in \mathfrak{B} \text{ such that } v|_{\mathcal{T}_-} = 0, v|_{\mathcal{T}_+} = w|_{\mathcal{T}_+}\}.$$

Note that when  $\mathcal{T} = \mathbb{R}$ ,  $X = L^2(\mathbb{R}_+)$ ,  $\mathcal{G}_{\mathfrak{B}}$  corresponds to the classical  $\mathcal{H}^2$  graph [13].

**LEMMA 4.4.** *Let  $\mathfrak{B} = \mathfrak{B}_{\text{cont}} \oplus \mathfrak{B}_{\text{aut}}$  be a linear, shift invariant behavior with finite memory. Then  $\mathcal{Z}_{\mathfrak{B}_{\text{cont}}} = \mathcal{Z}_{\mathfrak{B}}$  and  $\mathcal{G}_{\mathfrak{B}_{\text{cont}}} = \mathcal{G}_{\mathfrak{B}}$ .*

*Proof.* Since  $\mathfrak{B}_{\text{cont}} \subset \mathfrak{B}$  it follows that  $\mathcal{Z}_{\mathfrak{B}_{\text{cont}}} \subset \mathcal{Z}_{\mathfrak{B}}$  and  $\mathcal{G}_{\mathfrak{B}_{\text{cont}}} \subset \mathcal{G}_{\mathfrak{B}}$ . Conversely, let  $w \in \mathcal{Z}_{\mathfrak{B}}$ . By the direct sum, there exist  $w_1 \in \mathfrak{B}_{\text{cont}}$ ,  $w_2 \in \mathfrak{B}_{\text{aut}}$  such that

$w = w_1 + w_2$ . By definition of  $\mathcal{Z}_{\mathfrak{B}}$ , it follows that  $(w_1 + w_2)|_{\mathcal{T}_-} = 0$ , so by Lemma 3.2(2),  $w_1|_{\mathcal{T}_-} = w_2|_{\mathcal{T}_-} = 0$ . By the autonomy of  $\mathfrak{B}_{\text{aut}}$ , it follows that  $w_2 = 0$ . Hence  $\mathcal{Z}_{\mathfrak{B}} \subset \mathcal{Z}_{\mathfrak{B}_{\text{cont}}}$  and  $\mathcal{G}_{\mathfrak{B}} \subset \mathcal{G}_{\mathfrak{B}_{\text{cont}}}$ .  $\square$

We now introduce a notion of uniform stability, which captures the property that in addition to stability there is a uniform gain between future inputs and outputs when the past is zero. We will discuss the relation between this notion of stability and dissipativity descriptions of stability in section 7.

DEFINITION 4.5. *A linear behavior with input-output partition  $u|y$  is uniformly stable if*

1.  $\mathfrak{B}$  is stable;
2. *there exists a bounded operator  $\Psi: X \rightarrow X$  such that for all  $(u, y) \in \mathfrak{B}$  such that  $u|_{\mathcal{T}_+} \in X$ ,  $(u, y)|_{\mathcal{T}_-} = 0$  it follows that  $y = \Psi(u)$ .*

Note that the existence of a single stable autonomous subbehavior  $\mathfrak{B}_{\text{aut}}$  such that  $\mathfrak{B} = \mathfrak{B}_{\text{cont}} \oplus \mathfrak{B}_{\text{aut}}$  does *not* imply stability. For an example, consider  $\dot{x} = x + u$ ,  $\dot{z} = -z$ ,  $y = x + z$ . Then the subbehavior generated by  $\dot{z} = -z$ ,  $u = x = 0$ ,  $y = z$  is stable and has the direct sum property, and yet the behavior is not stable. However, in the context of differential systems, this property characterizes *stabilizability* (see [9] for the case of  $X = C_0(\mathbb{R}_+)$ ).

DEFINITION 4.6. *A behavior  $\mathfrak{B}$  is said to be stabilizable if for all  $w_1 \in \mathfrak{B}$  there exists  $w_2 \in \mathfrak{B}$  such that  $w_1|_{(-\infty, 0]} = w_2|_{(-\infty, 0]}$  and  $w_2|_{[0, \infty)} \in X$ .*

A useful sufficient condition for stabilizability is as follows.

LEMMA 4.7. *Let  $\mathfrak{B}$  be a linear shift invariant behavior with finite memory. If  $\mathfrak{B} = \mathfrak{B}_{\text{cont}} + \mathfrak{B}_{\text{aut}}$  and  $\mathfrak{B}_{\text{aut}}$  is stable, then  $\mathfrak{B}$  is stabilizable.*

*Proof.* Suppose there exists a stable autonomous subbehavior  $\mathfrak{B}_{\text{aut}}$  such that  $\mathfrak{B} = \mathfrak{B}_{\text{cont}} + \mathfrak{B}_{\text{aut}}$ . Then, given any  $w \in \mathfrak{B}$ , there exist  $w_1 \in \mathfrak{B}_{\text{cont}}$ ,  $w_2 \in \mathfrak{B}_{\text{aut}}$  such that  $w = w_1 + w_2$ . By controllability and shift invariance of  $\mathfrak{B}_{\text{cont}}$ ,  $w_1$  can be patched with  $0 \in \mathfrak{B}_{\text{cont}}$  so that there exists  $w'_1 \in X$  such that  $w_1|_{(-\infty, 0]} = w'_1|_{(-\infty, 0]}$ . By the stability of  $\mathfrak{B}_{\text{aut}}$ ,  $w_2 \in X$ . Hence  $w' = w'_1 + w_2 \in X$  and  $w|_{(-\infty, 0]} = w'|_{(-\infty, 0]}$ , and thus  $\mathfrak{B}$  is stabilizable as required.  $\square$

It is natural to ask whether the converse holds, that is, whether stabilizability of  $\mathfrak{B}$  implies stability of  $\mathfrak{B}_{\text{aut}}$ . For differential systems, Theorem 5.2.30 in [9] establishes the equivalence. For delay-differential systems with commensurate delays, the equivalence has been conjectured in [5, p. 117]. It is thus useful to define the weaker notion of “soundly stabilizable” and to formally note the equivalence of “soundly stabilizable” and “stabilizable” for differential systems as follows.

DEFINITION 4.8. *A behavior  $\mathfrak{B}$  is said to be soundly stabilizable if there exists a stable  $\mathfrak{B}_{\text{aut}}$  such that  $\mathfrak{B} = \mathfrak{B}_{\text{aut}} + \mathfrak{B}_{\text{cont}}$ .*

PROPOSITION 4.9. *For a differential behavior  $\mathfrak{B}$ ,  $\mathfrak{B}$  is stabilizable if and only if it is soundly stabilizable.*

*Proof.* See Theorem 5.2.30 in [9].  $\square$

We are primarily interested in the standard feedback interconnections shown in Figures 1 and 2.

DEFINITION 4.10. *Given a plant behavior  $\mathfrak{B}^P$ , a controller behavior  $\mathfrak{B}^C$ , and interconnection behavior  $\mathfrak{B}^I$ ,*

$$(4.2) \quad \mathfrak{B}^I = \{(w_0, w_1, w_2)^T \in X_e \mid w_0 = w_1 + w_2\},$$

*we define the closed-loop behavior  $\mathfrak{B}^{P \wedge C}$  as follows:*

$$\mathfrak{B}^{P \wedge C} = \{(w_0, w_1, w_2)^T \in \mathfrak{B}^I \mid w_1 \in \mathfrak{B}^P, w_2 \in \mathfrak{B}^C\}.$$

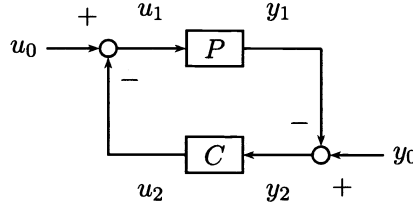
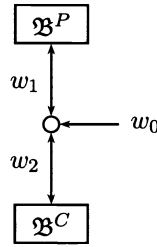


FIG. 1. The closed loop.

FIG. 2. The interconnected behaviors:  $w_i = (u_i, y_i)^T$ ,  $i = 0, 1, 2$ .

To ensure uniqueness of solutions of the closed loop (modulo the autonomous part of the behavior) we adopt the following definition.

DEFINITION 4.11. Given a plant behavior  $\mathfrak{B}^P$ , a controller behavior  $\mathfrak{B}^C$ , and interconnection behavior  $\mathfrak{B}^I$  (4.2), the behavior  $\mathfrak{B}^{P \wedge I C}$  is said to be well-posed if

$$(4.3) \quad X_e = \mathcal{Z}_{\mathfrak{B}^P} \oplus \mathcal{Z}_{\mathfrak{B}^C}.$$

This captures the idea that for the interconnection of behaviors with a zero past, “ $w_0$  is an input, and for any input  $w_0$ , there exist unique internal signals  $w_1, w_2$ .”

By (4.3), any  $w_0 \in X_e$  has a unique decomposition  $w_0 = w_1 + w_2$  with  $w_1 \in \mathcal{Z}_{\mathfrak{B}^P}$  and  $w_2 \in \mathcal{Z}_{\mathfrak{B}^C}$ . Hence two projection operators can be defined as below:

$$(4.4) \quad \begin{aligned} \Pi_{P//C}: X_e &\rightarrow \mathcal{Z}_{\mathfrak{B}^P}, w_0 \mapsto w_1, \\ \Pi_{C//P}: X_e &\rightarrow \mathcal{Z}_{\mathfrak{B}^C}, w_0 \mapsto w_2. \end{aligned}$$

Due to the interconnection behavior, we have

$$(4.5) \quad \Pi_{P//C} + \Pi_{C//P} = I.$$

DEFINITION 4.12. A controller behavior  $\mathfrak{B}^C$  is said to be a stabilizing controller for a plant behavior  $\mathfrak{B}^P$  if  $\mathfrak{B}^{P \wedge I C}$  is stable.

**5. Double axis time theories.** In [3] it is shown that classical notions of stability and causality lead to problematic inconsistencies when input-output systems defined over a doubly infinite time axis are considered. In particular a causal and stable system  $P_h$  was considered, defined by the following convolution:

$$(5.1) \quad P_h: u \mapsto y \quad : \quad y(t) = \int_{-\infty}^{\infty} h(t - \tau)u(\tau) d\tau = (h * u)(t),$$

where  $h(t) = \exp(t)$  for  $t \geq 0$  and 0 otherwise. The interconnection of  $P_h$  with a controller  $C$  implementing negative unity feedback with gain greater than one as in

Figure 1 was considered. It was shown [3] that if such a closed loop is considered to be well-posed and stable (in the sense that the (single valued) map  $\Pi: L^2(\mathbb{R}_+) \rightarrow L^2(\mathbb{R}_+)$  defined by  $\begin{pmatrix} u_0 \\ y_0 \end{pmatrix} \mapsto \begin{pmatrix} u_1 \\ y_1 \end{pmatrix}$  exists (and necessarily has finite induced norm)), then necessarily the classical  $L^2(\mathbb{R})$  graph of  $P_h$  is closed. It was further shown in [3] that the trajectories

$$(5.2) \quad \tilde{u}(t) = \begin{cases} \exp(-t), & t \geq 0, \\ 0, & t < 0, \end{cases} \quad \tilde{y}(t) = \begin{cases} -\frac{1}{2} \exp(-t), & t \geq 0, \\ -\frac{1}{2} \exp(t), & t < 0, \end{cases}$$

can be obtained as the limit of a sequence of trajectories lying in the  $L^2(\mathbb{R})$  graph of  $P_h$ , which is a contradiction since this solution does not satisfy the relation (5.1). Furthermore, it can be shown [3] that if  $w_0 = (\tilde{u}, \tilde{y})^T$  acts as the disturbance to the closed loop  $[P_h, C]$ , where  $C$  is negative unity feedback in Figure 1, then there is no solution  $w_1 = (u_1, y_1)^T$ .

We view the above observations as indicating an inadequacy of (5.1) as a complete physical model. By linearity the input-output model (5.1) enforces that  $u = 0$  implies  $y = 0$ , and hence there is no nontrivial autonomous subbehavior. By the natural inclusion of such autonomous subbehaviors, and with the corresponding relaxations of the notions of causality, well-posedness, and stability, the example can be reconsidered as follows.

Let  $\mathfrak{B}^{P_h}$  denote the smallest differential behavior containing all trajectories  $(u, y)$  satisfying (5.1). It can be observed that  $\mathfrak{B}^{P_h}$  can be expressed by the following (minimal) kernel representation:

$$\mathfrak{B}^{P_h} = \{w \in L^1_{\text{loc}}(\mathbb{R}) \mid [-1 \ s - 1] w = 0\},$$

where  $s = \frac{d}{dt}$ , where solutions are interpreted in the weak sense in  $L^1_{\text{loc}}(\mathbb{R})^1$  (rather than  $L^1_e(\mathbb{R})$ ) to avoid any possible implicit imposition of a time direction. The behavior  $\mathfrak{B}^{P_h}$  can be explicitly expressed as

$$\mathfrak{B}^{P_h} = \left\{ \begin{pmatrix} u \\ y \end{pmatrix} \in L^1_{\text{loc}}(\mathbb{R}) \mid \begin{array}{l} y(t) = y_0 \exp(t) + \int_{-\infty}^t \exp(t - \tau) u(\tau) d\tau, \\ y_0 \in \mathbb{R}, u \in L^1_{\text{loc}}(\mathbb{R}). \end{array} \right\}.$$

In terms of the definitions given in this paper, this behavior is indeed stabilizable, and negative unity feedback with a gain greater than one provides a well-posed stabilizing interconnection: if  $\mathfrak{B}^C = \{w \in L^2_e(\mathbb{R}_+) \mid w = (u, y)^T, u(t) = -ky(t)\}$ , then  $\mathcal{B}^{P_h \wedge_I C}$  is (uniformly) stable for  $k > 1$ . It is important to observe that

$$\begin{pmatrix} \tilde{u} \\ \tilde{y} \end{pmatrix} = \begin{pmatrix} \tilde{u} \\ -\frac{1}{2} \exp(t) + \int_{-\infty}^t \exp(t - \tau) \tilde{u}(\tau) d\tau \end{pmatrix} \in \mathcal{B}^{P_h} \cap L^2(\mathbb{R}),$$

since  $(\tilde{u}, \tilde{y})^T$  can be explained by the sum of the unforced solution  $(0, \exp(\cdot))^T$  and the forced solution  $(\tilde{u}, \int_{-\infty}^t \exp(t - \tau) \tilde{u}(\tau) d\tau)^T$ .

It is important to observe that the approach developed in this paper provides an alternative approach to addressing the classical problems of a doubly infinite time axis. The approach taken here is perhaps halfway between a double and a half line time axis, in that signals are defined over the whole of  $\mathbb{R}$ , but stability notions are related to boundedness of signals when restricted to  $\mathbb{R}_+$ . Only in the case of uniform

<sup>1</sup>  $f \in L^1_{\text{loc}}(\mathbb{R})$  if, for all compact  $\Omega \subset \mathbb{R}$ ,  $f|_{\Omega} \in L^1(\Omega)$ .

stability do we consider an induced norm and a zero past. The notion of well-posedness again restricts our attention to the subbehavior with a zero past and importantly does not impose uniqueness of solutions:  $(w_0, w_1, w_2), (w_0, v_1, v_2) \in \mathfrak{B}^{P_h \wedge C}$  does not imply  $(w_1, w_2) = (v_1, v_2)$  in general.

The approach considered in [7] identifies the operator  $P_h$  defined by (5.1) with its  $L^2(\mathbb{R})$  closure  $\bar{P}_h$ . In this case, the closure exists and is the (stable) anticausal operator  $\bar{P}_h = P_g: L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R})$ , where

$$(5.3) \quad P_g: u \mapsto y \quad : \quad y(t) = \int_{-\infty}^{\infty} g(t - \tau)u(\tau) \, d\tau = (g * u)(t),$$

where  $g(t) = -\exp(t)$  for  $t \leq 0$  and zero otherwise. In [3], the identification of these two input-output systems is interpreted as “more or less amount[ing] to abandoning any notion of causality,” and it was stated that “this is not a natural option, however, if the direction of time is well-defined.” It has the additional problem that  $\bar{P}_h$  is stable, whose response on the bounded input  $u(t) = 1$  if  $0 \leq t \leq 1$ ,  $u(t) = 0$  otherwise is the following bounded output signal:

$$y(t) = \begin{cases} 0 & \text{if } t > 1, \\ 1 - \exp(1 - t) & \text{if } 0 \leq t \leq 1, \\ \exp(t) - \exp(1 + t) & \text{if } t < 0. \end{cases}$$

We view this as problematic, since  $P_h$  itself is defined as an operator  $L^2(\mathbb{R}) \rightarrow L^2_{\text{loc}}(\mathbb{R})$ , with the following unbounded output response to the above input:

$$y(t) = \begin{cases} \exp(t) - \exp(1 + t) & \text{if } t > 1, \\ 1 - \exp(t) & \text{if } 0 \leq t \leq 1, \\ 0 & \text{if } t < 0. \end{cases}$$

These problems are avoided only in the case whereby the input-output operator has causal closure. In the discrete setting, the class of such transfer functions has been precisely characterized in [6] as the class of all Smirnoff functions, a class which includes all causal stable operators and excludes all causal unstable operators (as in the example considered), thus indicating the intrinsic difficulties with the input-output theory over  $\mathbb{R}$ .

In common with these other approaches to resolving the so-called Georgiou–Smith paradox, the procedure of identifying the convolution system (5.1) with the smallest differential behavior containing the same input-output pairs also identifies the same behavior  $\mathcal{B}^P = \mathcal{B}^{P_h} = \mathcal{B}^{P_g}$  to the anticausal input-output system (5.3), as it is easily verified that  $\mathfrak{B}^P$  contains all trajectories  $(u, y) \in L^1_{\text{loc}}$  satisfying (5.3).<sup>2</sup> However, the consideration of  $\mathcal{B}^P$  permits us to maintain a sensible notion of causality as follows.

**DEFINITION 5.1.** *A behavior  $\mathfrak{B}$  with input-output partition  $(u, y)$  is said to be causal if*

$$T_\tau u_1 = T_\tau u_2 \quad \implies \quad T_\tau \mathfrak{B}_{u_1} = T_\tau \mathfrak{B}_{u_2},$$

where  $\mathfrak{B}_u = \{w \in \mathfrak{B} \mid \exists y \text{ such that } w = (u, y) \in \mathfrak{B}\}$ .

<sup>2</sup>Importantly, however,  $\mathcal{B}^P$  also includes unbounded trajectories such as  $(0, \exp(\cdot))^T$  which are neither of the form  $(u, P_h u)^T$  nor  $(u, P_g u)^T$ .

This can be interpreted as stating that the set of all past trajectories which can be generated from a particular past input cannot be affected by changing the future input, and it represents a generalization of the notion of a causal operator (where the nonuniqueness of the output given the input is suitably accounted for). We can now observe that  $\mathfrak{B}^P$  indeed preserves causality, and we have thus arrived at a position whereby we can consider a suitable treatment of the system (5.1) in which the physical object under study can be thought of as causal and stabilizable. We find the explanation of trajectories of the system as a combination of an autonomous unforced subbehavior and a causal input to be more in line with physical thinking than the interpretation of the trajectories arising from a noncausal input to a single valued operator.

### 6. A behavioral generalization of the gap metric and robust stability.

In this section we will be concerned with deriving the behavioral version of the central robust stability theorem for linear time invariant systems. Our concern, for now, is with behaviors whose underlying signal space is equipped with a norm  $\|\cdot\|$ ; that is,  $X$  is a vector space and all behaviors  $\mathfrak{B}$  are such that  $P_+\mathfrak{B} \subset X_e$ . Furthermore, we assume that  $(X, \|\cdot\|)$  has the property that  $\|T_\tau x\| \leq a$  with  $a > 0$  for all  $\tau \geq 0$  implies  $x \in X$ . The classical Lebesgue and Sobolev signal spaces, e.g.,  $X = L^p$ ,  $l^p$ ,  $W^{p,m}$   $1 \leq m, p \leq \infty$ , satisfy this condition.

DEFINITION 6.1. *A mapping  $\Psi : \text{dom}(\Psi) \subset X_e \rightarrow X_e$  is said to be causal if  $T_\tau \Psi w = T_\tau \Psi T_\tau w$  for all  $w \in \text{dom}(\Psi)$  and  $\tau > 0$  and with  $T_\tau w \in \text{dom}(\Psi)$ . Its induced norm, denoted by  $\|\Psi\|$ , is defined as*

$$\|\Psi\| = \sup \left\{ \frac{\|T_\tau \Psi w\|}{\|T_\tau w\|} : w \in \text{dom}(\Psi), \tau > 0, T_\tau w \neq 0 \text{ and } T_\tau w \in \text{dom}(\Psi) \right\}.$$

Observe that  $\|\Pi_{P//C}\| \geq 1$  since, for any  $w_0 \in \mathcal{G}_{\mathfrak{B}^P}$ ,  $\Pi_{P//C} w_0 = w_0$ . Motivated by [4] and the considerations in section 4 we define the following notion of a gap distance between behaviors.

DEFINITION 6.2. *Given two behaviors  $\mathfrak{B}^1, \mathfrak{B}^2$  define a gap functional:*

$$(6.1) \quad \vec{\delta}(\mathfrak{B}^1, \mathfrak{B}^2) = \begin{cases} \inf_{\Phi \in \mathcal{O}} \|(I - \Phi)|_{\mathcal{G}_{\mathfrak{B}^1}}\| & \text{if } \mathfrak{B}^2 \text{ is soundly stabilizable,} \\ 1 & \text{if not,} \end{cases}$$

$$(6.2) \quad \delta(\mathfrak{B}^1, \mathfrak{B}^2) = \max \left\{ \vec{\delta}(\mathfrak{B}^1, \mathfrak{B}^2), \vec{\delta}(\mathfrak{B}^2, \mathfrak{B}^1) \right\},$$

where

$$\mathcal{O} = \{\Phi : \text{dom}(\Phi) \subset \mathcal{G}_{\mathfrak{B}^1} \rightarrow \mathcal{G}_{\mathfrak{B}^2} \mid \Phi \text{ bijective, causal, } \Phi(0) = 0\}.$$

In the case of differential systems, the above definition of the gap can be related to the classical definitions as follows. Let  $P_1(s)$ ,  $P_2(s)$  denote transfer functions corresponding to  $\mathfrak{B}_{\text{cont}}^{P_1}$ ,  $\mathfrak{B}_{\text{cont}}^{P_2}$  respectively, and let  $(N_i, D_i) \in \mathcal{RH}_\infty$  form normalized coprime factorizations of  $P_i$ ,  $i = 1, 2$ . It follows that the classical graphs for  $P_1(s)$  and  $P_2(s)$  correspond (in the frequency domain) to the time domain graphs  $\mathcal{G}_{\mathfrak{B}_{\text{cont}}^{P_1}}$ ,  $\mathcal{G}_{\mathfrak{B}_{\text{cont}}^{P_2}}$ . In the case where  $X = L^2$ , it has been shown in [4] and the references therein that

$$(6.3) \quad \inf_{\Phi \in \mathcal{O}} \|(I - \Phi)|_{\mathcal{G}_{P_1(s)}}\| = \vec{\delta}_0(P_1(s), P_2(s)),$$

where the classical gap  $\vec{\delta}_0$  can be expressed in a number of equivalent manners. Here we adopt an expression [13] which shows that the gap corresponds to the size of smallest stable coprime factor perturbation between the two plants:

$$\vec{\delta}_0(P_1(s), P_2(s)) = \inf \left\{ \left\| \begin{pmatrix} \Delta_N \\ \Delta_D \end{pmatrix} \right\|_{\mathcal{H}^\infty} \mid \begin{pmatrix} \Delta_N \\ \Delta_D \end{pmatrix} \in \mathcal{RH}^\infty, P_2 = (N_1 + \Delta_N)(D_1 + \Delta_D)^{-1} \right\}.$$

In the context of differential systems, Lemma 4.4 shows that  $\mathcal{G}_{\mathfrak{B}_{\text{cont}}^1} = \mathcal{G}_{\mathfrak{B}^1}$  and  $\mathcal{G}_{\mathfrak{B}_{\text{cont}}^2} = \mathcal{G}_{\mathfrak{B}^2}$ , and Proposition 4.9 shows the equivalence between the concepts of sound stabilizability and stabilizability. Therefore it can be easily shown that the gap functional is determined as follows:

$$(6.4) \quad \vec{\delta}(\mathfrak{B}^1, \mathfrak{B}^2) = \begin{cases} \vec{\delta}_0(P_1(s), P_2(s)) & \text{if } \mathfrak{B}^2 \text{ is stabilizable,} \\ 1 & \text{if not.} \end{cases}$$

We also remark, for completeness, that the directed gap can be computed via a standard  $\mathcal{H}^\infty$  optimization [2]:

$$\vec{\delta}_0(P_1, P_2) = \inf_{Q \in \mathcal{H}^\infty} \left\| \begin{pmatrix} D_1 \\ N_1 \end{pmatrix} - \begin{pmatrix} D_2 \\ N_2 \end{pmatrix} Q \right\|.$$

Observe that Definition 6.2 is a “real” behavioral definition: everything is defined in terms of trajectories, and all subbehaviors involved can be expressed in set-theoretic terms from the original behavior  $\mathfrak{B}$ . From a behavioral perspective, it should also be noted that the definition does not require a distinguished input-output partition. It is natural to wish to substitute the condition of stabilizability for that of sound stabilizability in the definition of the gap functional, but as we have discussed previously the equivalence of these concepts is known only in the case of differential systems (see Proposition 4.9).

The central reason for consideration of gap distances in systems theory is to obtain robust stability results (see Theorem 6.6 below). In particular we want  $\delta$  to capture the idea that any sensible stabilizing controller for  $\mathfrak{B}^P$  will also stabilize  $\mathfrak{B}^{P_1}$ , provided  $\delta(\mathfrak{B}^P, \mathfrak{B}^{P_1})$  is small. By definition, the distance between  $\mathfrak{B}$  and  $\mathfrak{B}_{\text{cont}}$  is zero if  $\mathfrak{B}$  is soundly stabilizable—this is reasonable since *any* stabilizing controller for  $\mathfrak{B}$  will automatically stabilize  $\mathfrak{B}_{\text{cont}}$  since  $\mathfrak{B}_{\text{cont}} \subset \mathfrak{B}$ . Consequently,  $\delta$  is necessarily at most a pseudometric; indeed the distance between two stabilizable differential systems with the same transfer function will be 0 (since the graphs of the behaviors are identical, the minimizing  $\Phi$  in the definition of the gap distance can be taken to be the identity, and hence the gap distance is zero).

For the controllable-autonomous decomposition of the interconnected behavior, we have the following lemma.

LEMMA 6.3. *Suppose  $\mathfrak{B}^P, B^C$  are linear behaviors with controllable-autonomous decompositions  $\mathfrak{B}^P = \mathfrak{B}_{\text{cont}}^P + \mathfrak{B}_{\text{aut}}^P$ ,  $\mathfrak{B}^C = \mathfrak{B}_{\text{cont}}^C + \mathfrak{B}_{\text{aut}}^C$ . Then  $\mathfrak{B}^{P \wedge I C}$  has the following controllable-autonomous decomposition:  $\mathfrak{B}^{P \wedge I C} = \mathfrak{B}_{\text{cont}}^{P \wedge I C} + \mathfrak{B}_{\text{aut}}^{P \wedge I C}$  and*

$$(6.5) \quad \begin{aligned} \mathfrak{B}_{\text{cont}}^{P \wedge I C} &= \{(w_1 + w_2, w_1, w_2) \mid w_1 \in \mathfrak{B}_{\text{cont}}^P, w_2 \in \mathfrak{B}_{\text{cont}}^C\}, \\ \mathfrak{B}_{\text{aut}}^{P \wedge I C} &= \{(w_1 + w_2, w_1, w_2) \mid w_1 \in \mathfrak{B}_{\text{aut}}^P, w_2 \in \mathfrak{B}_{\text{aut}}^C\}. \end{aligned}$$

*If, in addition,  $\mathfrak{B}^P = \mathfrak{B}_{\text{cont}}^P \oplus \mathfrak{B}_{\text{aut}}^P$ ,  $\mathfrak{B}^C = \mathfrak{B}_{\text{cont}}^C \oplus \mathfrak{B}_{\text{aut}}^C$ , then  $\mathfrak{B}^{P \wedge I C} = \mathfrak{B}_{\text{cont}}^{P \wedge I C} \oplus \mathfrak{B}_{\text{aut}}^{P \wedge I C}$ .*

*Proof.* It is straightforward to verify that  $\mathfrak{B}_{\text{cont}}^{P \wedge I C}$  is the maximal controllable behavior and that  $\mathfrak{B}_{\text{aut}}^{P \wedge I C}$  is autonomous. Let  $w \in \mathfrak{B}^{P \wedge I C}$ . Then  $w = (v_1 + v_2, v_1, v_2)$ , and by the decompositions of  $\mathfrak{B}^P$ ,  $\mathfrak{B}^C$ , there exist elements  $x_1 \in \mathfrak{B}_{\text{cont}}^P$ ,  $x_2 \in \mathfrak{B}_{\text{aut}}^P$  and  $y_1 \in \mathfrak{B}_{\text{cont}}^C$ ,  $y_2 \in \mathfrak{B}_{\text{aut}}^C$  such that  $v_1 = x_1 + x_2$ ,  $v_2 = y_1 + y_2$ . Consequently, there exists a decomposition of  $w = z_1 + z_2$ , where  $z_1 \in (\mathfrak{B}_{\text{cont}}^P + \mathfrak{B}_{\text{cont}}^C) \times \mathfrak{B}_{\text{cont}}^P \times \mathfrak{B}_{\text{cont}}^C$  and  $z_2 \in (\mathfrak{B}_{\text{aut}}^P + \mathfrak{B}_{\text{aut}}^C) \times \mathfrak{B}_{\text{aut}}^P \times \mathfrak{B}_{\text{aut}}^C$ , namely  $z_1 = (x_1 + y_1, x_1, y_1)$ ,  $z_2 = (x_2 + y_2, x_2, y_2)$ .

When  $\mathfrak{B}^P = \mathfrak{B}_{\text{cont}}^P \oplus \mathfrak{B}_{\text{aut}}^P$ ,  $\mathfrak{B}^C = \mathfrak{B}_{\text{cont}}^C \oplus \mathfrak{B}_{\text{aut}}^C$ , the existence for each of  $x_1, x_2, y_1, y_2$  and  $z_1, z_2$  is unique. Hence  $\mathfrak{B}^{P \wedge I C} = \mathfrak{B}_{\text{cont}}^{P \wedge I C} \oplus \mathfrak{B}_{\text{aut}}^{P \wedge I C}$ .  $\square$

The following key results relate a condition of stability of a particular half-line projection to stability of the entire system behavior.

LEMMA 6.4. *Let  $\mathfrak{B}^{P \wedge I C}$  be well-posed, and let*

$$\hat{\mathfrak{B}} = \left\{ w = (w_0, w_1, w_2) \in \mathfrak{B}_{\text{cont}}^{P \wedge I C} : w_0|_{[0, \infty)} = 0 \right\}.$$

*Suppose  $X = \mathcal{G}_{\mathfrak{B}^P} \oplus \mathcal{G}_{\mathfrak{B}^C}$ . Then, for any  $(w_0, w_1, w_2) \in \hat{\mathfrak{B}}$ ,  $w_1|_{[0, \infty)}, w_2|_{[0, \infty)} \in X$ .*

*Proof.* Let  $w = (w_0, w_1, w_2) \in \hat{\mathfrak{B}}$ . By controllability, there exist  $\tau \geq 0$ ,  $\bar{w} = (\bar{w}_0, \bar{w}_1, \bar{w}_2) \in \mathfrak{B}^{P \wedge I C}$  such that  $\bar{w}|_{(-\infty, -\tau]} = 0$  and  $\bar{w}|_{[0, \infty)} = w|_{[0, \infty)}$ . Therefore  $\sigma_\tau \bar{w}_1 \in \mathcal{Z}_{\mathfrak{B}^P}$ ,  $\sigma_\tau \bar{w}_2 \in \mathcal{Z}_{\mathfrak{B}^C}$ , and  $\bar{w}_0|_{\mathbb{R} \setminus (-\tau, 0]} = 0$ , and hence  $\bar{w}_0|_{[0, \infty)}, \sigma_\tau \bar{w}_0|_{[0, \infty)} \in X$ . Since  $X = \mathcal{G}_{\mathfrak{B}^P} \oplus \mathcal{G}_{\mathfrak{B}^C}$ ,  $\sigma_\tau \bar{w}_0|_{[0, \infty)} = z_1 + z_2$  for some  $z_1 \in \mathcal{G}_{\mathfrak{B}^P}$ ,  $z_2 \in \mathcal{G}_{\mathfrak{B}^C}$ . By the well-posedness assumption, it follows that  $z_1 = \sigma_\tau \bar{w}_1$ ,  $z_2 = \sigma_\tau \bar{w}_2$ . Since  $\bar{w}|_{[0, \infty)} = w|_{[0, \infty)}$  and  $z_1, z_2 \in X$ , we see that  $w_1|_{[0, \infty)}, w_2|_{[0, \infty)} \in X$ .  $\square$

PROPOSITION 6.5. *Let  $\mathfrak{B}^P, \mathfrak{B}^C$  be linear, shift invariant behaviors with finite memory and  $\mathfrak{B}^P = \mathfrak{B}_{\text{cont}}^P \oplus \mathfrak{B}_{\text{aut}}^P$ ,  $\mathfrak{B}^C = \mathfrak{B}_{\text{cont}}^C \oplus \mathfrak{B}_{\text{aut}}^C$ . Suppose  $\mathfrak{B}^P$  and  $\mathfrak{B}^C$  are soundly stabilizable and  $\mathfrak{B}^{P \wedge I C}$  is well-posed. Suppose further that  $X = \mathcal{G}_{\mathfrak{B}^P} \oplus \mathcal{G}_{\mathfrak{B}^C}$ . Then  $\mathfrak{B}^{P \wedge I C}$  is stable.*

*Proof.* Suppose  $w \in \mathfrak{B}^{P \wedge I C}$ , and  $w = (w_0, w_1, w_2)$ . We have to show that if  $w_0|_{[0, \infty)} \in X$ , then  $w_1|_{[0, \infty)}, w_2|_{[0, \infty)} \in X$ . Let  $w_0 \in X$ . Since  $\mathfrak{B}^{P \wedge I C} = \mathfrak{B}_{\text{cont}}^{P \wedge I C} + \mathfrak{B}_{\text{aut}}^{P \wedge I C}$ , it follows that  $w = x + y$ , where  $x = (x_0, x_1, x_2) \in \mathfrak{B}_{\text{cont}}^{P \wedge I C}$  and  $y = (y_0, y_1, y_2) \in \mathfrak{B}_{\text{aut}}^{P \wedge I C}$ . By stability of  $\mathfrak{B}_{\text{aut}}^P, \mathfrak{B}_{\text{aut}}^C$  we know that  $x_0|_{[0, \infty)} = w_0 - y_0|_{[0, \infty)} \in X$ .

Let

$$\tilde{x}_0 = \begin{cases} 0, & t \leq 0, \\ x_0, & t > 0. \end{cases}$$

Then  $\tilde{x}_0|_{[0, \infty)} \in X$ . Since  $X = \mathcal{G}_{\mathfrak{B}^P} \oplus \mathcal{G}_{\mathfrak{B}^C}$ , we see that

$$(6.6) \quad \tilde{x}_0|_{[0, \infty)} = \tilde{x}_1|_{[0, \infty)} + \tilde{x}_2|_{[0, \infty)} \quad \text{for some } \tilde{x}_1 \in \mathcal{G}_{\mathfrak{B}^P}, \tilde{x}_2 \in \mathcal{G}_{\mathfrak{B}^C}.$$

Consider  $v = (v_0, v_1, v_2) := (x_0 - \tilde{x}_0, x_1 - \tilde{x}_1, x_2 - \tilde{x}_2)$ . Since  $x_1, \tilde{x}_1 \in \mathfrak{B}_{\text{cont}}^P$ ,  $x_2, \tilde{x}_2 \in \mathfrak{B}_{\text{cont}}^C$ , we see that  $x_1 - \tilde{x}_1 \in \mathfrak{B}_{\text{cont}}^P$ ,  $x_2 - \tilde{x}_2 \in \mathfrak{B}_{\text{cont}}^C$ , and hence  $v \in \mathfrak{B}_{\text{cont}}^{P \wedge I C}$ . Since  $v_0|_{[0, \infty)} = 0, v \in \hat{\mathfrak{B}}$ . By the above lemma,  $v_1|_{[0, \infty)}, v_2|_{[0, \infty)} \in X$ . Since  $\tilde{x}_1, \tilde{x}_2 \in X$ , we see that  $x_1|_{[0, \infty)}, x_2|_{[0, \infty)} \in X$ . By the stability assumption of  $\mathfrak{B}_{\text{aut}}^P$  and  $\mathfrak{B}_{\text{aut}}^C$ , it follows that  $w_1|_{[0, \infty)}, w_2|_{[0, \infty)} \in X$ . This completes the proof.  $\square$

We can now give the proof of the main robust stability result. Before giving the proof we remark that the result follows straightforwardly from Proposition 6.5 once it has been shown that  $X = \mathcal{G}_{\mathfrak{B}_{\text{cont}}^P} \oplus \mathcal{G}_{\mathfrak{B}_{\text{cont}}^C}$  and that this classical condition is obtained directly using the technique of [1]: we have included this part of the proof from [1] for completeness.

THEOREM 6.6. *Suppose  $\mathfrak{B}^P, \mathfrak{B}^{P_1}, \mathfrak{B}^C$  are linear, shift invariant behaviors with finite memory. If*



1.  $\mathfrak{B}^P, \mathfrak{B}^C$  are soundly stabilizable,
2.  $\mathfrak{B}^{P \wedge I^C}$  is well-posed, causal, and uniformly stable,
3.  $\mathfrak{B}^{P_1 \wedge I^C}$  is well-posed, causal, and
4.  $\tilde{\delta}(\mathfrak{B}^P, \mathfrak{B}^{P_1}) \|\Pi_{P//C}\| < 1$ ,

then  $\mathfrak{B}^{P_1 \wedge I^C}$  is uniformly stable.

*Proof.* Condition 6.6 implies that there exists a stable  $\mathfrak{B}_{\text{aut}}^{P_1}$  such that  $\mathfrak{B}^{P_1} = \mathfrak{B}_{\text{cont}}^{P_1} + \mathfrak{B}_{\text{aut}}^{P_1}$  by definition of the gap and since  $\|\Pi_{P//C}\| \geq 1$ .

By condition 6.6, there exists a surjective mapping  $\Phi : D \subset \mathcal{G}_{\mathfrak{B}^P} \rightarrow \mathcal{G}_{\mathfrak{B}^{P_1}}$  such that

$$\|\Phi - I\| \|\Pi_{P//C}\| < 1.$$

Let  $(w_0, w_1, w_2) \in \mathfrak{B}^{P_1 \wedge I^C}$  with  $(w_0, w_1, w_2)|_{\mathcal{T}_-} = 0, (w_0, w_1, w_2)|_{\mathcal{T}_+} \in X$ . By definition of the extended graph,  $w_1 \in \mathcal{Z}_{\mathfrak{B}^{P_1}}, w_2 \in \mathcal{Z}_{\mathfrak{B}^C}$ .

For any  $\tau > 0$ , by sound stabilizability and Lemma 4.7, there exist  $\bar{w}_1 \in \mathfrak{B}^{P_1}, \bar{w}_2 \in \mathfrak{B}^C$  such that  $\bar{w}_i|_{\mathcal{T}_-} = \sigma_\tau w_i|_{\mathcal{T}_-}, \bar{w}_i|_{\mathcal{T}_+} \in X$  for  $i = 1, 2$ . This shows that  $w_i|_{(-\infty, \tau]} = w_i^\tau|_{(-\infty, \tau]}$ , where  $w_i^\tau = \sigma_{-\tau} \bar{w}_i$ , and  $w_i^\tau|_{\mathcal{T}_-} = 0, w_i^\tau|_{\mathcal{T}_+} \in X$ . By shift invariance,  $w_1^\tau|_{\mathcal{T}_+} \in \mathcal{G}_{\mathfrak{B}^{P_1}}, w_2^\tau|_{\mathcal{T}_+} \in \mathcal{G}_{\mathfrak{B}^C}$  and  $T_\tau w_i = T_\tau w_i^\tau$ .

Since  $\Phi$  is surjective from  $\mathcal{G}_{\mathfrak{B}^P}$  to  $\mathcal{G}_{\mathfrak{B}^{P_1}}$ , there exist  $w_3^\tau \in \mathcal{G}_{\mathfrak{B}^P}$  and  $w_1^\tau|_{\mathcal{T}_+} = \Phi w_3^\tau$ . Write  $x_\tau = w_3^\tau + w_2^\tau|_{\mathcal{T}_+}$ . Then by condition 6.6,  $\Pi_{P//C} x_\tau = w_3^\tau \in X, \Pi_{C//P} x_\tau = w_2^\tau|_{\mathcal{T}_+} \in X$  and

$$\begin{aligned} T_\tau(w_0|_{\mathcal{T}_+}) &= T_\tau(w_1|_{\mathcal{T}_+}) + T_\tau(w_2|_{\mathcal{T}_+}) = T_\tau(w_1^\tau|_{\mathcal{T}_+}) + T_\tau(w_2^\tau|_{\mathcal{T}_+}) = T_\tau \Phi w_3^\tau + T_\tau(w_2^\tau|_{\mathcal{T}_+}) \\ &= T_\tau \Phi \Pi_{P//C} x_\tau + T_\tau \Pi_{C//P} x_\tau = T_\tau \Phi \Pi_{P//C} T_\tau x_\tau + T_\tau \Pi_{C//P} T_\tau x_\tau \\ (6.7) \quad &= T_\tau(\Phi - I) \Pi_{P//C} T_\tau x_\tau + T_\tau x_\tau \end{aligned}$$

and

$$(6.8) \quad T_\tau \Pi_{P_1//C}(w_0|_{\mathcal{T}_+}) = T_\tau(w_1|_{\mathcal{T}_+}) = T_\tau \Phi w_3^\tau = T_\tau \Phi \Pi_{P//C} x_\tau = T_\tau \Phi \Pi_{P//C} T_\tau x_\tau.$$

By (6.7), we have

$$\|T_\tau x_\tau\| \leq \|T_\tau(w_0|_{\mathcal{T}_+})\| + \|T_\tau(\Phi - I) \Pi_{P//C} T_\tau x_\tau\| \leq \|w_0|_{\mathcal{T}_+}\| + \|\Phi - I\| \|\Pi_{P//C}\| \|T_\tau x_\tau\|,$$

which gives

$$\|T_\tau x_\tau\| \leq \frac{\|w_0|_{\mathcal{T}_+}\|}{1 - \|\Phi - I\| \|\Pi_{P//C}\|}.$$

By (6.8), we have

$$\begin{aligned} \|T_\tau \Pi_{P_1//C}(w_0|_{\mathcal{T}_+})\| &\leq \|T_\tau(\Phi - I) \Pi_{P//C} T_\tau x_\tau\| + \|T_\tau \Phi \Pi_{P//C} T_\tau x_\tau\| \\ &\leq (1 + \|\Phi - I\|) \|\Pi_{P//C}\| \frac{\|w_0|_{\mathcal{T}_+}\|}{1 - \|\Phi - I\| \|\Pi_{P//C}\|}. \end{aligned}$$

Hence  $w_1 = \Pi_{P_1//C}(w_0|_{\mathcal{T}_+}) \in X$  and, therefore, by (4.5),  $w_2 = \Pi_{C//P_1}(w_0) \in X$ .

So, for any  $w_0$  with  $w_0|_{\mathcal{T}_+} \in X, w_0|_{\mathcal{T}_-} = 0$ , we have shown that there exist  $w_1 \in \mathcal{G}_{\mathfrak{B}^{P_1}}, w_2 \in \mathcal{G}_{\mathfrak{B}^C}$  such that  $w_0|_{\mathcal{T}_+} = w_1 + w_2$ , i.e.,  $X = \mathcal{G}_{\mathfrak{B}^{P_1}} + \mathcal{G}_{\mathfrak{B}^C}$ . By the well-posedness assumption, the sum is direct. Applying Proposition 6.5, we see that  $\mathfrak{B}^{P_1 \wedge I^C}$  is stable. The above proof shows that both  $\Pi_{P_1//C}$  and  $\Pi_{C//P_1}$  are bounded. Hence  $\mathfrak{B}^{P_1 \wedge I^C}$  is uniformly stable.  $\square$

Due to Proposition 4.9, we have the following corollary.

**COROLLARY 6.7.** *Suppose  $\mathfrak{B}^P, \mathfrak{B}^{P_1}, B^C$  are linear, shift invariant differential behaviors with finite memory. If  $\mathfrak{B}^P, \mathfrak{B}^C$  are stabilizable,  $\mathfrak{B}^{P \wedge I^C}$  is uniformly stable, and  $\mathfrak{B}^{P_1 \wedge I^C}$  is well-posed, causal, and  $\tilde{\delta}(\mathfrak{B}^P, \mathfrak{B}^{P_1}) \|\Pi_{P//C}\| < 1$ , then  $\mathfrak{B}^{P_1 \wedge I^C}$  is uniformly stable.*

### 7. Relation to the behavioral $\mathcal{H}^\infty$ synthesis of Trentelman and Willems.

Within the context of  $L^2$  signal spaces, classical  $\mathcal{H}^\infty$  synthesis [20] provides constructions for controllers  $C$  which achieve  $\|\Pi_{P//C}\| \leq 1$ , i.e., solve the normalized version of the inequality required in our robustness theorems. The classical gap robustness results then provide an explicit description of plant uncertainties tolerated in the closed loop. In direct counterpart, and in the interests of a self-contained behavioral theory, it is relevant to relate the results of this paper to the behavioral approach to  $\mathcal{H}^\infty$  synthesis found in [12, 17], for then our basic robust stability theorem completes a “behavioral robust control theory” by providing an explicit robustness interpretation of the behavioral  $\mathcal{H}^\infty$  synthesis results.

Therefore we explicitly describe the relationship between the problem formulation of [12, 17] and this paper. We first consider Proposition 1 of [12]. By choosing the exogenous variable  $d$  to be  $w_0$  and the endogenous “to be controlled” variable  $f$  to be  $w_1$ , we have

$$\mathcal{K} = \{(w_0, w_1) \in C^\infty \mid \exists w_2 \in C^\infty \text{ such that } (w_0, w_1, w_2) \in \mathfrak{B}^{P \wedge_I C}\},$$

and  $G_{w_0 \rightarrow w_1}$  is the transfer function corresponding to  $\Pi_{P//C}$ . Proposition 1 asserts that if  $\mathfrak{B}^P$ ,  $\mathfrak{B}^C$  are smooth differential behaviors and  $\mathfrak{B}^{P \wedge_I C}$  is controllable, then the following are equivalent:

1. in  $\mathcal{K}$ ,  $w_0$  is the input,  $w_1$  is the output, and  $\|G_{w_0 \rightarrow w_1}\|_{\mathcal{H}_\infty} \leq 1$ ;
2.  $\mathcal{K}$  is  $\Sigma$ -dissipative on  $\mathbb{R}_-$ , and  $m(\mathcal{K}) = \sigma_+(\Sigma)$ ;
3.  $\|w_1\|_{\mathcal{L}^2(\mathbb{R}, \mathbb{R}^f)} \leq \|w_0\|_{\mathcal{L}^2(\mathbb{R}, \mathbb{R}^d)}$ ,  $w_0$  is free in  $\mathcal{K}$ , and  $(0, w_1) \in \mathcal{K}$  implies that  $\lim_{t \rightarrow \infty} w_1(t) = 0$ .

Here  $\Sigma = \text{diag}(I_d, -I_f)$ ,  $\sigma_+(\Sigma)$  is the number of positive eigenvalues of  $\Sigma$ , and  $m(\mathcal{K})$  is the number of “free” input variables: in the context of this paper the free variables enter additively in both the input and output channels, and hence  $m(\mathcal{K}) = \dim(\mathcal{U} \times \mathcal{Y})$ . We refer the reader to [12, 17] for the definition of  $\Sigma$ -dissipativity on  $\mathbb{R}_-$ .

We now relate the above stability concepts to the notion of uniform stability, within an  $L^2$  context, as considered in this paper. Consider the following condition:

4.  $\mathfrak{B}^{P \wedge_I C}$  is uniformly stable, and  $\|\Pi_{P//C}\|_{L^2(\mathbb{R}_+)} \leq 1$ .

Then we have the following proposition.

**PROPOSITION 7.1.** *Let  $X = L^2(\mathbb{R}_+)$ . Suppose  $\mathfrak{B}^P$ ,  $\mathfrak{B}^C$  are differential behaviors and  $\mathfrak{B}^{P \wedge_I C}$  is controllable. Then 1, 2, 3, and 4 are equivalent.*

*Proof.* Proposition 1 of [12] establishes the equivalence between 1, 2, and 3. It is well known that  $\|\Pi_{P//C}\|_{L^2(\mathbb{R}_+)} = \|G_{w_0 \rightarrow w_1}\|_{\mathcal{H}_\infty}$ . Hence 4 implies 1. On the other hand, suppose 1 holds. Consider  $(w_0, w_1, w_2) \in \mathfrak{B}^{P \wedge_I C}$ , and suppose  $w_0|_{\mathbb{R}_+} \in X$ . Let  $w_0 = x_0 + y_0$ , where  $x_0|_{\mathbb{R}_+} = 0$ ,  $y_0|_{\mathbb{R}_-} = 0$ . Then there exist  $x_1, x_2, y_1, y_2$  such that  $(x_0, x_1, x_2) \in \widehat{\mathfrak{B}^{P \wedge_I C}} \subset \mathfrak{B}^{P \wedge_I C}$ ,  $(y_0, y_1, y_2) \in \mathfrak{B}^{P \wedge_I C}$ , and  $x_1 + y_1 = w_1$ ,  $x_2 + y_2 = w_2$ . By Lemma 6.4, and since  $x_0|_{\mathbb{R}_+} = 0 \in X$ , it follows that  $x_1|_{\mathbb{R}_+}, x_2|_{\mathbb{R}_+} \in X$ . Since  $G_{w_0 \rightarrow w_1} \in \mathcal{H}_\infty$  and  $y_0|_{\mathbb{R}_+} \in X$  it follows that  $y_1|_{\mathbb{R}_+}, y_2|_{\mathbb{R}_+} \in X$ . Hence  $w_1|_{\mathbb{R}_+}, w_2|_{\mathbb{R}_+} \in X$ , and consequently  $\mathfrak{B}^{P \wedge_I C}$  is stable. The inequality in 1 implies the inequality in 4, and hence 1 implies 4 as required.  $\square$

Within the context of disturbance attenuation for linear controllable differential systems in an  $L^2$  setting, the results of [12, 17] establish conditions under which there exists a controllable differentiable behavior  $\mathfrak{B}^C$  which renders  $\Sigma$ -dissipativity on  $\mathbb{R}_-$  of the closed-loop interconnection  $\mathfrak{B}^{P \wedge_I C}$ . Here  $\mathfrak{B}^P$  is also required to be a controllable differential behavior. Since the resulting interconnection  $\mathfrak{B}^{P \wedge_I C}$  is controllable, and by the above, it follows that this synthesis yields the uniform stability condition 4

above, and, in turn, the robust stability theorem (Theorem 6.6) provides an explicit description of a set of plants for which stability can be guaranteed.

It is worth noting that it is observed in [17] that the synthesis can be extended in an ad hoc manner from the controllable case to the general case by introducing appropriate stabilizability assumptions in the analysis. Theorem 6.6 can be utilized to achieve these observations directly. Given a (soundly) stabilizable plant behavior  $\mathfrak{B}^P$ , follow the  $\mathcal{H}^\infty$  synthesis to derive a controller  $\mathfrak{B}^C$  (which is controllable) for the controllable plant subbehavior  $\mathfrak{B}_{\text{cont}}^P$ . Then, since  $\vec{\delta}(\mathfrak{B}_{\text{cont}}^P, \mathfrak{B}^P) = 0$ , Theorem 6.6 can be applied to establish the required uniform stability for the interconnection of the derived controller behavior  $\mathfrak{B}^C$  and the original plant  $\mathfrak{B}^P$ .

**8. An illustrative example.** We consider an example in an  $L^\infty$  setting. Let the behaviors for the nominal system  $P$ , the perturbed system  $P_1$ , and the controller  $C$  be given by

$$\mathfrak{B}^P = \left\{ \begin{pmatrix} u_1 \\ v_1 \\ x \end{pmatrix} \left| \begin{array}{l} u_1, v_1, x \in L_e^\infty(\mathbb{R}, \mathbb{R}), \dot{x} = au_1 + bv_1 \end{array} \right. \right\},$$

$$\mathfrak{B}^{P_1} = \left\{ \begin{pmatrix} u_1 \\ v_1 \\ x \end{pmatrix} \left| \begin{array}{l} u_1, v_1, x \in L_e^\infty(\mathbb{R}, \mathbb{R}), \dot{x} = a\sigma_{\tau_1}u_1 + b\sigma_{\tau_2}v_1 \end{array} \right. \right\},$$

and

$$\mathfrak{B}^C = \left\{ \begin{pmatrix} cy \\ dy \\ y \end{pmatrix} \left| \begin{array}{l} y \in L_e^\infty(\mathbb{R}, \mathbb{R}), \end{array} \right. \right\},$$

where  $a, b, c, d \in \mathbb{R}$  such that  $ac + bd = -1$ .  $\mathfrak{B}^P, \mathfrak{B}^{P_1}$ , and  $\mathfrak{B}^C$  are all linear, shift invariant behaviors with finite memory. Moreover, they are all controllable so that  $\mathfrak{B}^P = \mathfrak{B}_{\text{cont}}^P + \{0\}$ ,  $\mathfrak{B}^{P_1} = \mathfrak{B}_{\text{cont}}^{P_1} + \{0\}$ , and  $\mathfrak{B}^C = \mathfrak{B}_{\text{cont}}^C + \{0\}$ . We also have

$$\begin{aligned} \mathcal{G}_{\mathfrak{B}^P} &= \left\{ \begin{pmatrix} u_1 \\ v_1 \\ x \end{pmatrix} \left| \begin{array}{l} u_1, v_1, x \in L^\infty(\mathbb{R}, \mathbb{R}), \dot{x} = au_1 + bv_1, \\ u_1|_{\mathbb{R}_-} = v_1|_{\mathbb{R}_-} = x|_{\mathbb{R}_-} = 0, \end{array} \right. \right\}, \\ \mathcal{G}_{\mathfrak{B}^{P_1}} &= \left\{ \begin{pmatrix} u_1 \\ v_1 \\ x \end{pmatrix} \left| \begin{array}{l} u_1, v_1, x \in L^\infty(\mathbb{R}, \mathbb{R}), \dot{x} = a\sigma_{\tau_1}u_1 + b\sigma_{\tau_2}v_1, \\ u_1|_{\mathbb{R}_-} = v_1|_{\mathbb{R}_-} = x|_{\mathbb{R}_-} = 0, \end{array} \right. \right\}, \\ \mathcal{Z}_{\mathfrak{B}^P} &= \left\{ \begin{pmatrix} u_1 \\ v_1 \\ x \end{pmatrix} \left| \begin{array}{l} u_1, v_1, x \in L_e^\infty(\mathbb{R}, \mathbb{R}), \dot{x} = au_1 + bv_1, \\ u_1|_{\mathbb{R}_-} = v_1|_{\mathbb{R}_-} = x|_{\mathbb{R}_-} = 0, \end{array} \right. \right\}, \\ \mathcal{Z}_{\mathfrak{B}^{P_1}} &= \left\{ \begin{pmatrix} u_1 \\ v_1 \\ x \end{pmatrix} \left| \begin{array}{l} u_1, v_1, x \in L_e^\infty(\mathbb{R}, \mathbb{R}), \dot{x} = a\sigma_{\tau_1}u_1 + b\sigma_{\tau_2}v_1, \\ u_1|_{\mathbb{R}_-} = v_1|_{\mathbb{R}_-} = x|_{\mathbb{R}_-} = 0. \end{array} \right. \right\}. \end{aligned}$$

The parallel projection  $\Pi_{P//C}$  operator is given by

$$\Pi_{P//C} \begin{pmatrix} u_0 \\ v_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} u_0 - cy_0 + cx \\ v_0 - dy_0 + dx \\ x \end{pmatrix} \quad \text{for any } u_0, v_0, y_0 \in L_e^\infty(\mathbb{R}_+, \mathbb{R}),$$

where  $x$  is the unique solution to the equation

$$\dot{x} = -x + au_0 + bv_0 + y_0, x(0) = 0.$$

After straightforward calculation, we see that  $\|T_\tau x\| \leq \|T_\tau(au_0 + bv_0 + y_0)\|$  for any  $\tau > 0$  and

$$\|\Pi_{P/JC}\| \leq 1 + |c| + |d|(1 + |a| + |b|).$$

This shows that  $\mathfrak{B}^{P \wedge I^C}$  is uniformly stable.

To estimate the gap between  $\mathfrak{B}^P$  and  $\mathfrak{B}^{P_1}$ , we let

$$\Phi \begin{pmatrix} u \\ v \\ x \end{pmatrix} = \begin{pmatrix} u \\ v \\ y \end{pmatrix} \quad \text{for any } \begin{pmatrix} u \\ v \\ x \end{pmatrix} \in \mathfrak{B}^P,$$

where

$$\begin{pmatrix} u \\ v \\ y \end{pmatrix} \in \mathfrak{B}^{P_1}.$$

Then

$$\begin{aligned} |x(t) - y(t)| &= \left| \int_0^t (au(s) + bv(s))ds - \int_0^t (au(s - \tau_1) + bv(t - \tau_2))ds \right| \\ &\leq |a| \left| \int_{t-\tau}^t u(s)ds \right| + |b| \left| \int_{t-\tau}^t v(s)ds \right| \\ &\leq \tau_1 |a| \|u\| + \tau_2 |b| \|v\| \end{aligned}$$

and so

$$\begin{aligned} \|(I - \Phi)|_{\mathcal{G}_{\mathfrak{B}^P}}\| &= \sup \left\{ \frac{\|T_\tau(x - y)\|}{\|T_\tau(u, v, x)^\top\|} \mid \tau > 0, \begin{pmatrix} u \\ v \\ x \end{pmatrix} \in \mathcal{G}_{\mathfrak{B}^P}, \begin{pmatrix} u \\ v \\ x \end{pmatrix} \neq 0 \right\} \\ &\leq \tau_1 |a| + \tau_2 |b|. \end{aligned}$$

Hence,  $\vec{\delta}(\mathfrak{B}^P, \mathfrak{B}^{P_1}) \leq |a|\tau_1 + |b|\tau_2$  and, by Theorem 6.6,  $\mathfrak{B}^{P_1 \wedge I^C}$  is uniformly stable, provided

$$(\tau_1 |a| + \tau_2 |b|)[1 + |c| + |d|(1 + |a| + |b|)] < 1.$$

By definition of uniform stability, this means that

(i) for any  $u_1, v_1, x, y \in L_e^\infty(\mathbb{R}, \mathbb{R})$  with  $\dot{x} = a\sigma_{\tau_1}u_1 + b\sigma_{\tau_2}v_1$  on  $\mathbb{R}$  and  $(u_1 + cy)|_{\mathbb{R}_+}, (v_1 + dy)|_{\mathbb{R}_+}, (x + y)|_{\mathbb{R}_+} \in L^\infty(\mathbb{R}_+, \mathbb{R})$ , we have  $u_1|_{\mathbb{R}_+}, v_1|_{\mathbb{R}_+}, x|_{\mathbb{R}_+}, y|_{\mathbb{R}_+} \in L^\infty(\mathbb{R}_+, \mathbb{R})$ ;

(ii) the mappings which map  $(u_0, v_0, x_0)|_{\mathbb{R}_+}$  to  $(u_1, v_1, x_1)$  and  $(cy, dy, y)$ , respectively, are both bounded, where  $u_1 + cy = u_0, v_1 + dy = v_0, x_1 + y = x_0, (cy, dy, y)|_{\mathbb{R}_-} = 0, (u_1, v_1, x_1)|_{\mathbb{R}_-} = 0$ , and

$$\dot{x}_1 = a\sigma_{\tau_1}u_1 + b\sigma_{\tau_2}v_1.$$

We remark, as in the introduction, that such a system lies outside the scope of the existing algebraic behavioral theory for delay systems where delays  $\tau_1, \tau_2$  would additionally be required to be commensurate [5].

**Acknowledgments.** The authors would like to thank the reviewers and the associate editor for comments which have greatly improved the manuscript and led to the development of the discussion in section 5.

## REFERENCES

- [1] T. GEORGIU AND M. C. SMITH, *Optimal robustness in the gap metric*, IEEE Trans. Automat. Control, 35 (1990), pp. 673–686.
- [2] T. T. GEORGIU, *On the computation of the gap metric*, Systems Control Lett., 11 (1988), pp. 253–257.
- [3] T. GEORGIU AND M. C. SMITH, *Intrinsic difficulties in using the doubly-infinite time axis for input-output control theory*, IEEE Trans. Automat. Control, 40 (1995), pp. 516–518.
- [4] T. GEORGIU AND M. C. SMITH, *Robustness analysis of nonlinear feedback systems: An input-output approach*, IEEE Trans. Automat. Control, 42 (1997), pp. 1200–1221.
- [5] H. GLUESING-LUERSSEN, *Linear Delay-Differential Systems with Commensurate Delays: An Algebraic Approach*, Lecture Notes in Math. 1770, Springer, Berlin, 2002.
- [6] B. JACOB AND J. PARTINGTON, *Graphs, closability, and causality of linear time-invariant discrete-time systems*, Internat. J. Control, 73 (2000), pp. 1051–1060.
- [7] P. M. MÄKILÄ, *On three puzzles in robust control*, IEEE Trans. Automat. Control, 45 (2000), pp. 552–556.
- [8] H. K. PILLAI AND S. SHANKAR, *A behavioral approach to the control of distributed systems*, SIAM J. Control. Optim., 37 (1998), pp. 388–408.
- [9] J. W. POLDERMAN AND J. C. WILLEMS, *Introduction to Mathematical Systems Theory*, Springer, New York, 1998.
- [10] P. ROCHA AND J. WOOD, *Trajectory control and interconnection of 1D and nD systems*, SIAM J. Control. Optim., 40 (2001), pp. 107–134.
- [11] A. J. SASANE, *Distance between behaviors*, Internat. J. Control, 72 (2003), pp. 1214–1223.
- [12] H. L. TRENTelman AND J. C. WILLEMS, *Synthesis of dissipative systems using quadratic differential forms: Part II*, IEEE Trans. Automat. Control, 47 (2002), pp. 70–86.
- [13] G. VINNICOMBE, *Uncertainty and Feedback:  $\mathcal{H}_\infty$ -shaping Control System Synthesis*, Imperial College Press, London, 2001.
- [14] J. C. WILLEMS, *From time series to linear system - Part I. Finite dimensional linear time invariant systems*, Automatica J. IFAC, 22 (1986), pp. 561–580.
- [15] J. C. WILLEMS, *From time series to linear system - Part II. Exact modelling*, Automatica J. IFAC, 22 (1986), pp. 675–694.
- [16] J. C. WILLEMS, *From time series to linear system - Part III. Approximate modelling*, Automatica J. IFAC, 23 (1987), pp. 87–115.
- [17] J. C. WILLEMS AND H. L. TRENTelman, *Synthesis of dissipative systems using quadratic differential forms: Part I*, IEEE Trans. Automat. Control, 47 (2002), pp. 53–69.
- [18] J. WOOD, E. ROGERS, AND D. H. OWENS, *Controllable and autonomous nD linear systems*, Multidimens. Systems Signal Process., 10 (1999), pp. 33–69.
- [19] G. ZAMES AND A. K. EL-SAKKARY, *Unstable systems and feedback: The gap metric*, in Proceedings of the Allerton Conference, 1980, pp. 380–385.
- [20] K. ZHOU, J. DOYLE, AND K. GLOVER, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1996.

## ALGEBRAIC STRUCTURES IN NONLINEAR SYSTEMS OVER RINGS OBTAINED BY IMMERSION\*

TOSHIYUKI OHTSUKA<sup>†</sup>

**Abstract.** An immersion of a system is a mapping of the initial state exactly preserving the input-output map. As has already been shown, a system is immersible into a rational-in-the-state representation (RSR) and a polynomial-in-the-state representation (PSR) if and only if the field generated by the observation space is finitely generated, which is true in most practical systems. In this paper, some algebraic structures and their geometric counterparts associated with an RSR and a PSR obtained via an immersion are discussed. First, RSRs and PSRs are viewed as systems over rings in a unified framework, and the notions of an invariant ideal and an invariant variety, which are related to a differential algebraic equation, are introduced. Then, it is shown that an RSR and a PSR have invariant ideals and invariant varieties associated with an immersion. In particular, an invariant variety of an RSR or a PSR is the Zariski closure of the image of the immersion, i.e., the smallest variety containing the image of the immersion. The degrees of freedom in RSRs and PSRs obtained via immersion are also investigated and are characterized in terms of invariant ideals.

**Key words.** system immersion, systems over rings, input-output map

**AMS subject classifications.** 37C10, 93B11, 93B25, 93C10

**DOI.** 10.1137/070698233

**1. Introduction.** An immersion [9, 15, 18, 22] of a system is a mapping of the initial state from the original state space to another state space, so as to preserve the input-output map exactly, and it is usually a mapping to a higher-dimensional space. The model structure of the given system can be simplified while preserving the input-output map with an immersion. Immersions into linear [1, 5, 12, 15], bilinear [16], quasi-state-affine [2], and rational and polynomial [22] systems have been discussed in the literature.

Immersion was used in [1, 2, 12, 15] to design an observer for a nonlinear system that is immersible into a linear system or other particular forms. Similarly, immersion into a polynomial system is potentially applicable to observer design for a broad class of nonlinear systems if the general methodology of observer design is established for polynomial systems. Often in system analysis and control design we assume polynomial systems [4, 14, 20, 23, 26, 27]. These techniques may be applicable to a wider class of nonlinear systems through the use of immersion. In contrast to the polynomial approximation of a nonlinear system, immersion does not raise problems due to approximation errors and can preserve the input-output map over an unbounded region in the state space, which is useful when applying theoretical results, at the expense of an increase in the dimension.

Although only a restricted class of nonlinear systems is immersible into linear or bilinear systems, most practical systems are immersible into rational systems and polynomial systems, as shown in [22]. More precisely, a nonlinear system is immersible

---

\*Received by the editors July 24, 2007; accepted for publication (in revised form) March 7, 2008; published electronically June 25, 2008. This work was partially supported by a Grant-in-Aid for Scientific Research (19560442) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

<http://www.siam.org/journals/sicon/47-4/69823.html>

<sup>†</sup>Department of Systems Innovation, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan (ohtsuka@sys.es.osaka-u.ac.jp).

into a rational-in-the-state representation (RSR), a polynomial-in-the-state representation (PSR), and a quadratic-in-the-state representation (QSR) if and only if a field generated by the observation space is finitely generated over the field of real numbers. Moreover, it is sufficient for immersibility into these representations that all functions in the given system are differentially algebraic functions, which most practical systems consist of. Therefore, an RSR, a PSR, or a QSR can be used as a general model structure for a wide class of nonlinear systems.

In contrast to previous work on immersion, this paper focuses on structures in systems obtained via immersion, rather than immersibility conditions. Although RSRs and PSRs are general forms of nonlinear systems on their own, they have particular structures of rational functions and polynomials, respectively. Therefore, some properties of the original system or the immersion may be reflected in additional algebraic structures in the RSRs and PSRs. Moreover, those algebraic structures may have some geometric interpretations. In fact, it is shown in this paper that the algebraic structures of RSRs and PSRs can be characterized in terms of rings, and their geometric counterparts are expressed naturally in terms of affine varieties rather than a manifold in differential geometry. These algebraic and geometric structures can be found in rational or polynomial systems aside from immersion-related problems.

This paper is organized as follows. In section 2, definitions of immersions and immersibility conditions are briefly reviewed, and an algebraic characterization is given for the minimal dimension of RSRs. In section 3, we show that RSRs and PSRs can be viewed as systems over rings in a unified framework and introduce the notions of an invariant ideal and an invariant variety. Then, we show the existence of an invariant ideal and an invariant variety in a system over a ring obtained by immersion. The notions of an invariant ideal and an invariant variety can be used as fundamental tools in analysis and the control of systems over rings in general. As a further application of algebraic tools, we characterize the degrees of freedom in systems over a ring in section 4, which again highlights the importance of invariant ideals in the analysis of nonlinear systems over rings. Finally, concluding remarks are given in section 5.

## 2. System immersion.

**2.1. Immersion and invariant immersion.** We treat an input-affine nonlinear system,

$$\Sigma \begin{cases} \dot{x} = g_0(x) + \sum_{i=1}^m g_i(x)u_i, \\ y = h(x), \end{cases}$$

where  $x(t) \in U \subset \mathbf{R}^n$  denotes the state vector,  $U$  an open set,  $u(t) = [u_1(t), \dots, u_m(t)]^T \in \mathbf{R}^m$  the input vector, and  $y(t) \in \mathbf{R}^p$  the output vector. The system is denoted by  $\Sigma(g_0, g_1, \dots, g_m, h)$  (by  $\Sigma$  for short hereafter). System  $\Sigma$  is said to be analytic on  $U$  if  $g_i : U \rightarrow \mathbf{R}^n$  ( $i \in I_{0,m}$ ) and  $h : U \rightarrow \mathbf{R}^p$  are analytic functions on  $U$ . Sets of indices are denoted as  $I_{i_1, i_2} = \{i \in \mathbf{Z} : i_1 \leq i \leq i_2\}$  and  $I_{i_1, \infty} = \{i \in \mathbf{Z} : i \geq i_1\}$ . The admissible set  $\Omega$  of the input function  $u : [0, \infty) \rightarrow \mathbf{R}^m$  is a set of bounded piecewise continuous functions with a common upper bound. We assume  $|u_i(t)| \leq 1$  ( $t \geq 0, i \in I_{1,m}$ ) without loss of generality. The trajectory of the state equation of system  $\Sigma$  starting from an initial state in  $U$  at  $t = 0$  and driven by an input function  $u$  is denoted by  $\Phi_t^{\Sigma, u} : U \rightarrow U$ . That is, for a given initial state  $x_0 \in U$  and an input function  $u \in \Omega$ , the solution of the state equation is given by  $x(t) = \Phi_t^{\Sigma, u}(x_0)$ . A set

of all possible Lie derivatives of the output map,  $\mathcal{L}_\Sigma$ , and the observation space  $\mathcal{O}_\Sigma$  are defined, respectively, by

$$\mathcal{L}_\Sigma = \{L_{g_{i_1}} \dots L_{g_{i_k}} h_j : j \in I_{1,p}, (i_1, \dots, i_k) \in I_{0,m}^k, k \in I_{0,\infty}\},$$

$$\mathcal{O}_\Sigma = \text{span}_{\mathbf{R}} \mathcal{L}_\Sigma,$$

where  $h = [h_1, \dots, h_p]^T$  and  $L_{g_i} h_j = (\partial h_j / \partial x) g_i$ . The observability codistribution at  $x_0 \in U$  is a vector space over  $\mathbf{R}$  given by

$$d\mathcal{O}_\Sigma(x_0) = \text{span}_{\mathbf{R}} \left\{ \frac{\partial \eta}{\partial x}(x_0) : \eta \in \mathcal{O}_\Sigma \right\}.$$

First, we define an immersion of a system [9, 15, 18].

**DEFINITION 2.1.** *An analytic system  $\Sigma(g_0, \dots, g_m, h)$  defined on an open set  $U \subset \mathbf{R}^n$  is said to be immersible on an open set  $U' \subset U$  into another  $\tilde{n}$ -dimensional system  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$  if there exists an analytic mapping  $\alpha : U' \rightarrow \mathbf{R}^{\tilde{n}}$  such that  $\tilde{\Sigma}$  is analytic on an open set containing  $\alpha(U')$ , and for every  $x_0 \in U'$  and for every  $u \in \Omega$ ,*

$$h \circ \Phi_t^{\Sigma, u}(x_0) = \tilde{h} \circ \Phi_t^{\tilde{\Sigma}, u}(\alpha(x_0))$$

*holds for every sufficiently small  $t > 0$ . Such a mapping  $\alpha$  is called an immersion of  $\Sigma$  on  $U'$  into  $\tilde{\Sigma}$ . We omit  $U'$  when it is obvious in the context or  $U' = U$ .*

The following proposition [22] is useful for checking whether or not a given mapping  $\alpha$  is an immersion.

**PROPOSITION 2.2.** *Let  $U \subset \mathbf{R}^n$  be an open set, let  $\alpha : U \rightarrow \mathbf{R}^{\tilde{n}}$  be an analytic mapping, let  $\Sigma(g_0, \dots, g_m, h)$  be an analytic system on  $U$ , and let  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$  be an analytic system on an open set containing  $\alpha(U)$ . The mapping  $\alpha$  is an immersion of  $\Sigma$  into  $\tilde{\Sigma}$  if and only if the following holds for all  $x_0 \in U$ :*

$$L_{g_{i_1}} \dots L_{g_{i_k}} h(x_0) = L_{\tilde{g}_{i_1}} \dots L_{\tilde{g}_{i_k}} \tilde{h}(\alpha(x_0)),$$

$$(i_1, \dots, i_k) \in I_{0,m}^k, k \in I_{0,\infty}.$$

Next, we define a particular form of an immersion [22].

**DEFINITION 2.3.** *An analytic system  $\Sigma(g_0, \dots, g_m, h)$  defined on an open set  $U \subset \mathbf{R}^n$  is said to be invariantly immersible on an open set  $U' \subset U$  into another  $\tilde{n}$ -dimensional system  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$  if there exists an analytic mapping  $\alpha : U' \rightarrow \mathbf{R}^{\tilde{n}}$  such that  $\tilde{\Sigma}$  is analytic on an open set containing  $\alpha(U')$ , and, for all  $x \in U'$ ,*

$$L_{g_i} \alpha(x) = \tilde{g}_i(\alpha(x)), \quad i \in I_{0,m},$$

$$h(x) = \tilde{h}(\alpha(x))$$

*hold. Such a mapping  $\alpha$  is called an invariant immersion of  $\Sigma$  on  $U'$  into  $\tilde{\Sigma}$ .*

Often, the invariant immersion is simply called an *immersion* in the literature. However, we distinguish invariant immersions from immersions in this work because the former not only preserves the input-output map but also has additional geometric properties. That is, for every state trajectory  $x(t) \in U'$  of the original system  $\Sigma$ ,  $\alpha(x(t)) \in \alpha(U')$  is a state trajectory of  $\tilde{\Sigma}$ .



**2.2. Immersibility conditions.** Rational and polynomial structures with respect to the state are discussed in this paper, for which the notions of fields and rings are suitable. Let  $C^\omega(U)$  be the ring of all real-valued analytic functions on an open set  $U \subset \mathbf{R}^n$ . For a subset  $A \subset C^\omega(U)$ ,  $\mathbf{R}[A]$  denotes the ring generated by  $A$  over  $\mathbf{R}$ . If  $U$  is a domain (connected open set),  $\mathbf{R}[A]$  is an integral domain and its fraction field  $\mathbf{R}(A)$  is well defined and called the field generated by  $A$  over  $\mathbf{R}$ . If  $A = \{\alpha_1, \dots, \alpha_\nu\}$  and  $\alpha = [\alpha_1, \dots, \alpha_\nu]^T$ ,  $\mathbf{R}[A]$  and  $\mathbf{R}(A)$  are also denoted by  $\mathbf{R}[\alpha]$  and  $\mathbf{R}(\alpha)$ , respectively. For a state vector  $\tilde{x} \in \mathbf{R}^{\tilde{n}}$ , its elements  $\tilde{x}_1, \dots, \tilde{x}_{\tilde{n}}$  can be regarded as analytic functions on  $\mathbf{R}^{\tilde{n}}$ , and  $\mathbf{R}[\tilde{x}]$  and  $\mathbf{R}(\tilde{x})$  denote a polynomial ring and a rational function field, respectively. We denote the subset of  $\mathbf{R}[\tilde{x}]$  with the total degree less than or equal to  $\ell$  as  $\mathbf{R}[\tilde{x}]_{\leq \ell}$ . Then, an RSR, a PSR, and a QSR are formally defined as follows [22].

**DEFINITION 2.4.** Consider a system  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$  with a state vector  $\tilde{x} \in \mathbf{R}^{\tilde{n}}$ . System  $\tilde{\Sigma}$  is said to be an RSR if  $\tilde{g}_i(\tilde{x}) \in \mathbf{R}(\tilde{x})^{\tilde{n}}$  ( $i \in I_{0,m}$ ) and  $\tilde{h}(\tilde{x}) \in \mathbf{R}(\tilde{x})^p$ . System  $\tilde{\Sigma}$  is said to be a PSR if  $\tilde{g}_i(\tilde{x}) \in \mathbf{R}[\tilde{x}]^{\tilde{n}}$  ( $i \in I_{0,m}$ ) and  $\tilde{h}(\tilde{x}) \in \mathbf{R}[\tilde{x}]^p$ . In particular, a PSR  $\tilde{\Sigma}$  is said to be a QSR if  $\tilde{g}_i(\tilde{x}) \in \mathbf{R}[\tilde{x}]_{\leq 2}^{\tilde{n}}$  ( $i \in I_{0,m}$ ) and  $\tilde{h}(\tilde{x}) \in \mathbf{R}[\tilde{x}]_{\leq 1}^p$ .

It has already been proved that immersibilities are equivalent between an RSR, a PSR, and a QSR [22].

**PROPOSITION 2.5.** For an analytic system  $\Sigma$  defined on an open set  $U$ , the following three claims are equivalent:

- (i) System  $\Sigma$  is immersible (resp., invariantly immersible) on  $U$  into an RSR.
- (ii) System  $\Sigma$  is immersible (resp., invariantly immersible) on  $U$  into a PSR.
- (iii) System  $\Sigma$  is immersible (resp., invariantly immersible) on  $U$  into a QSR.

Since it suffices to consider an RSR consisting of rational functions, immersibility and invariant immersibility are well characterized in terms of fields [22].

**PROPOSITION 2.6.** For an analytic system  $\Sigma$  defined on a domain  $U$ , the following three claims are equivalent:

- (i) On an open and dense subset of  $U$ , system  $\Sigma$  is invariantly immersible into an RSR with an analytic mapping defined on  $U$ .
- (ii) On an open and dense subset of  $U$ , system  $\Sigma$  is immersible into an RSR with an analytic mapping defined on  $U$ .
- (iii) The field  $\mathbf{R}(\mathcal{O}_\Sigma)$  is finitely generated over  $\mathbf{R}$ .

Moreover, if (iii) holds, every set of analytic generators of  $\mathbf{R}(\mathcal{O}_\Sigma)$  gives an invariant immersion in (i).

Many types of nonlinear systems satisfy condition (iii) in Proposition 2.6 and are (invariantly) immersible into an RSR, a PSR, and a QSR. Moreover, as shown in [22], it is sufficient for invariant immersibility into an RSR that all functions in a given system are differentially algebraic functions, which most practical systems consist of.

In advance of ring-theoretic discussions in subsequent sections, we give a field-theoretic characterization for the minimal dimension of RSRs obtained by immersion.

**THEOREM 2.7.** Suppose condition (iii) in Proposition 2.6 holds, and let  $\rho$  be the degree of transcendency of  $\mathbf{R}(\mathcal{O}_\Sigma)$  over  $\mathbf{R}$ . If  $\mathbf{R}(\mathcal{O}_\Sigma)$  is a purely transcendental extension of  $\mathbf{R}$ , the minimal dimension of RSRs into which  $\Sigma$  is invariantly immersible on an appropriate open subset of  $U$  is  $\rho$ . Otherwise, the minimal dimension of the RSRs is  $\rho + 1$ .

*Proof.* Note that  $\mathbf{R}(\mathcal{O}_\Sigma)$  has at least  $\rho$  generators, which are algebraically independent of each other. If  $\mathbf{R}(\mathcal{O}_\Sigma)$  is a purely transcendental extension of  $\mathbf{R}$ , there exists a transcendence basis  $\{\alpha_1, \dots, \alpha_\rho\} \subset \mathbf{R}(\mathcal{O}_\Sigma)$  such that  $\mathbf{R}(\mathcal{O}_\Sigma) = \mathbf{R}(\alpha_1, \dots, \alpha_\rho)$ . Since

every element in  $\mathcal{O}_\Sigma \subset C^\omega(U)$  is analytic, it is nonzero on an open and dense subset of  $U$ . Therefore, the mapping  $\alpha = [\alpha_1, \dots, \alpha_\rho]^T$ , consisting of elements in  $\mathbf{R}(\mathcal{O}_\Sigma)$ , is analytic on an open and dense subset of  $U$ . Then, there exists an open set  $U'' \subset U$  such that  $\Sigma$  is invariantly immersible on  $U''$  into an RSR with  $\alpha$ .

If  $\mathbf{R}(\mathcal{O}_\Sigma)$  is not a purely transcendental extension of  $\mathbf{R}$ ,  $\mathbf{R}(\mathcal{O}_\Sigma) \neq \mathbf{R}(\alpha_1, \dots, \alpha_\rho)$  for any transcendence basis  $\{\alpha_1, \dots, \alpha_\rho\} \subset \mathbf{R}(\mathcal{O}_\Sigma)$ . However, since  $\mathbf{R}(\mathcal{O}_\Sigma)$  is finitely generated over  $\mathbf{R}$ , there exist a finite number of elements  $\beta_1, \dots, \beta_\mu \in \mathbf{R}(\mathcal{O}_\Sigma)$  ( $\mu \geq 1$ ) such that  $\mathbf{R}(\mathcal{O}_\Sigma) = \mathbf{R}(\alpha_1, \dots, \alpha_\rho, \beta_1, \dots, \beta_\mu)$ , where  $\mathbf{R}(\alpha_1, \dots, \alpha_\rho, \beta_1, \dots, \beta_\mu)$  is a finite algebraic extension of  $\mathbf{R}(\alpha_1, \dots, \alpha_\rho)$  [19, Theorem 3.1.3]. Moreover, since the characteristic of  $\mathbf{R}(\alpha_1, \dots, \alpha_\rho)$  is zero, its finite extension is always a simple extension [19, Theorem 2.5.2]. That is, there exists a primitive element  $\gamma \in \mathbf{R}(\mathcal{O}_\Sigma)$  such that  $\mathbf{R}(\alpha_1, \dots, \alpha_\rho, \beta_1, \dots, \beta_\mu) = \mathbf{R}(\alpha_1, \dots, \alpha_\rho, \gamma)$ , which implies  $\mathbf{R}(\mathcal{O}_\Sigma) = \mathbf{R}(\alpha_1, \dots, \alpha_\rho, \gamma)$ . The mapping  $\alpha' = [\alpha_1, \dots, \alpha_\rho, \gamma]^T$  is analytic on an open and dense subset of  $U$ , and there exists an open set  $U'' \subset U$  such that  $\Sigma$  is invariantly immersible on  $U''$  into a RSR with  $\alpha'$ .  $\square$

### 3. Algebraic structures after immersion.

**3.1. Ideals associated with an immersion.** Suppose, for a given system  $\Sigma(g_0, \dots, g_m, h)$  on a domain  $U \subset \mathbf{R}^n$ , we have  $\mathbf{R}(\mathcal{O}_\Sigma) = \mathbf{R}(\alpha)$  with an analytic mapping  $\alpha : U \rightarrow \mathbf{R}^{\tilde{n}}$ . Then, Proposition 2.6 implies that  $\Sigma$  is invariantly immersible into an RSR  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$  with  $\alpha$  on an open and dense subset  $U' \subset U$ . Let the state vector of  $\tilde{\Sigma}$  be  $\tilde{x} \in \mathbf{R}^{\tilde{n}}$ . Since all functions in  $\tilde{\Sigma}$  are rational functions of  $\tilde{x}$ , we have  $\mathcal{O}_{\tilde{\Sigma}} \subset \mathbf{R}(\tilde{x})$ . Moreover, since  $\tilde{\Sigma}$  is defined on an open set containing  $\alpha(U')$ , denominators of the rational functions in  $\tilde{\Sigma}$  do not vanish identically on  $\alpha(U)$ , which implies a particular algebraic structure in  $\tilde{\Sigma}$ , as discussed below. A basic tool for analyzing an algebraic structure in a system after immersion is a relation ideal of a mapping [7, 19].

DEFINITION 3.1. Given a subset  $S \subset \mathbf{R}^{\tilde{n}}$ , denote by

$$\mathcal{I}(S) = \{\tilde{f} \in \mathbf{R}[\tilde{x}] : \tilde{f}(\tilde{x}) = 0 \text{ for all } \tilde{x} \in S\}$$

the ideal of polynomials vanishing on  $S$ . When  $S$  is an image of a mapping  $\alpha : U \rightarrow \mathbf{R}^{\tilde{n}}$ ,  $\mathcal{I}(\alpha(U))$  is called the relation ideal of  $\alpha$ .

Let  $P = \mathcal{I}(\alpha(U))$  be the relation ideal of the immersion of  $\Sigma$  into the RSR  $\tilde{\Sigma}$ , and let  $\alpha^* : \mathbf{R}[\tilde{x}] \rightarrow \mathbf{R}[\alpha]$  be a *substitution mapping* (or a pull back) defined by  $\alpha^*(f) = f(\alpha)$  for  $f \in \mathbf{R}[\tilde{x}]$ . Then,  $\alpha^*$  is a surjective ring homomorphism such that  $\text{Ker } \alpha^* = P$ , which induces a ring isomorphism  $\mathbf{R}[\alpha] \cong \mathbf{R}[\tilde{x}]/P$ . Since  $\mathbf{R}[\alpha]$  is an integral domain,  $P$  is a prime ideal.

As mentioned previously, every denominator of functions in the RSR  $\tilde{\Sigma}$  is not identically zero on  $\alpha(U)$  or, equivalently, does not belong to the prime ideal  $P$ . Therefore, every element of  $\tilde{g}_0, \dots, \tilde{g}_m$  and  $\tilde{h}$  belongs to not only the rational function field  $\mathbf{R}(\tilde{x})$  but also the *localization* of  $\mathbf{R}[\tilde{x}]$  at  $P$ , which is given by

$$\mathbf{R}[\tilde{x}]_P = \{f/g \in \mathbf{R}(\tilde{x}) : f, g \in \mathbf{R}[\tilde{x}], \text{ and } g \notin P\}.$$

Note that  $\mathbf{R}[\tilde{x}]_P$  is a *local ring* with the unique maximal ideal

$$P\mathbf{R}[\tilde{x}]_P = \{f/g \in \mathbf{R}(\tilde{x}) : f, g \in \mathbf{R}[\tilde{x}], g \notin P, \text{ and } f \in P\}.$$

It should also be noted that an element of  $\mathbf{R}[\tilde{x}]_P$  does not necessarily define a rational function on the whole of  $\alpha(U)$  because its denominator can vanish at some points.

However, there exists an open and dense subset  $U' \subset U$  such that an element of  $\mathbf{R}[\tilde{x}]_P$  is analytic on an open set containing  $\alpha(U')$ .

We can naturally extend the substitution mapping  $\alpha^* : \mathbf{R}[\tilde{x}] \rightarrow \mathbf{R}[\alpha]$  to a mapping  $\alpha^* : \mathbf{R}[\tilde{x}]_P \rightarrow \mathbf{R}(\alpha)$ , which is also a surjective ring homomorphism with  $\text{Ker } \alpha^* = PR[\tilde{x}]_P$  and induces a ring isomorphism  $\mathbf{R}(\alpha) \cong \mathbf{R}[\tilde{x}]_P / PR[\tilde{x}]_P$ . We use the same symbol to denote the substitution mappings for  $\mathbf{R}[\tilde{x}]$  and  $\mathbf{R}[\tilde{x}]_P$  because their domains are obvious from the context.

An immediate application of the maximal ideal  $PR[\tilde{x}]_P$  is the test of the observability rank condition for the system after immersion.

**THEOREM 3.2.** *Let  $\alpha$  be an invariant immersion of system  $\Sigma$  defined on a domain  $U \subset \mathbf{R}^n$  into an RSR  $\tilde{\Sigma}$  with the state vector  $\tilde{x} \in \mathbf{R}^{\tilde{n}}$  such that  $\mathbf{R}(\mathcal{O}_\Sigma) = \mathbf{R}(\alpha)$ , and let  $P \subset \mathbf{R}[\tilde{x}]$  be the relation ideal of  $\alpha$ . Then, there exist  $r_1, \dots, r_{\tilde{n}} \in PR[\tilde{x}]_P$  such that  $\tilde{x}_i + r_i \in \mathbf{R}(\mathcal{O}_{\tilde{\Sigma}})$  ( $i \in I_{1,\tilde{n}}$ ). Moreover, the observability rank condition  $\dim d\mathcal{O}_{\tilde{\Sigma}}(\tilde{x}) = \tilde{n}$  holds if*

$$\det \left( I_{\tilde{n}} + \frac{\partial r}{\partial \tilde{x}}(\tilde{x}) \right) \neq 0,$$

where  $r = [r_1, \dots, r_{\tilde{n}}]^T$  and  $I_{\tilde{n}}$  denotes the  $\tilde{n} \times \tilde{n}$  identity matrix.

*Proof.* Since every  $\alpha_i$  belongs to  $\mathbf{R}(\mathcal{O}_\Sigma)$ , there exist a finite number of elements  $H_1, \dots, H_\mu \in \mathcal{O}_\Sigma$  and rational functions  $\phi_i(X_1, \dots, X_\mu)$  ( $i \in I_{0,\tilde{n}}$ ) such that

$$\alpha_i = \phi_i(H_1, \dots, H_\mu).$$

Then, by Proposition 2.2, there also exist a finite number of elements  $\tilde{H}_1, \dots, \tilde{H}_\mu \in \mathcal{O}_{\tilde{\Sigma}}$  such that  $H_i = \tilde{H}_i \circ \alpha$ , and, consequently,  $\tilde{\alpha}_i = \phi_i(\tilde{H}_1, \dots, \tilde{H}_\mu) \in \mathbf{R}(\mathcal{O}_{\tilde{\Sigma}})$  also satisfy  $\alpha_i = \tilde{\alpha}_i \circ \alpha$ . Therefore,  $\tilde{\alpha}_i$  belongs to  $\mathbf{R}[\tilde{x}]_P$  and  $\alpha^*(\tilde{\alpha}_i) = \alpha_i$ . Since  $\tilde{x}_i \in \mathbf{R}[\tilde{x}]_P$  also satisfies  $\alpha^*(\tilde{x}_i) = \alpha_i$ , and  $\alpha^* : \mathbf{R}[\tilde{x}]_P \rightarrow \mathbf{R}(\alpha)$  is a ring homomorphism, we have  $\alpha^*(\tilde{\alpha}_i - \tilde{x}_i) = 0$ , which implies  $r_i = \tilde{\alpha}_i - \tilde{x}_i \in \text{Ker } \alpha^* = PR[\tilde{x}]_P$ .

Let  $\phi = [\phi_1, \dots, \phi_{\tilde{n}}]^T$ ,  $\tilde{H} = [\tilde{H}_1, \dots, \tilde{H}_\mu]^T$ , and  $X = [X_1, \dots, X_\mu]^T$ . From the definition of  $\tilde{\alpha}_i$ , we have  $\tilde{x} + r = \phi(\tilde{H})$ . By taking partial differentiation with respect to  $\tilde{x}$ , we have

$$I_{\tilde{n}} + \frac{\partial r}{\partial \tilde{x}}(\tilde{x}) = \frac{\partial \phi}{\partial X}(\tilde{H}) \frac{\partial \tilde{H}}{\partial \tilde{x}}(\tilde{x}).$$

If the determinant of the left-hand side is nonzero, the rank of  $\partial \tilde{H}(\tilde{x}) / \partial \tilde{x}$  on the right-hand side must be  $\tilde{n}$ , which implies  $\dim d\mathcal{O}_{\tilde{\Sigma}}(\tilde{x}) = \tilde{n}$ .  $\square$

It should be noted that Theorem 3.2 is applicable not only to RSRs but also to PSRs because PSRs are also RSRs.

**3.2. Nonlinear systems over rings.** Now, we can characterize some algebraic structures of RSRs and PSRs after immersion in terms of such a subring of the rational function field  $\mathbf{R}(\tilde{x})$  as  $\mathbf{R}[\tilde{x}]$  and  $\mathbf{R}[\tilde{x}]_P$  rather than  $\mathbf{R}(\tilde{x})$  itself. To this end, we prepare some ring-theoretic notions of nonlinear system theory in place of the usual differential geometric settings.

Let  $R$  be a partial differential subring of  $\mathbf{R}(\tilde{x})$ , such as  $\mathbf{R}[\tilde{x}]$  and  $\mathbf{R}[\tilde{x}]_P$ , satisfying  $(\partial / \partial \tilde{x}_i)R \subset R$  for all  $i \in I_{1,\tilde{n}}$ . Then, a PSR or an RSR  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$  is regarded as a *system over a ring* such that  $\tilde{g}_0, \dots, \tilde{g}_m \in R^{\tilde{n}}$  and  $\tilde{h} \in R^p$  for an appropriate ring  $R$ . Note that the state  $\tilde{x}$  belongs to Euclidean space as usual in the present notion of a system over a ring, which is different from *linear systems over rings* [6, 13]. It

should also be noted that we consider a partial differential ring, while the conventional differential algebraic system theory [8, 10, 24] is based on ordinary differential fields with the single derivation operator  $d/dt$ .

The vector fields  $\tilde{g}_0, \dots, \tilde{g}_m$  can be viewed as elements of a free  $R$ -module  $R^{\tilde{n}}$  rather than sections of a tangent bundle. Moreover, the Lie derivative  $L_{\tilde{g}_i}$  is regarded as a mapping  $L_{\tilde{g}_i} : R \rightarrow R$ , which is not a ring endomorphism of  $R$  in general but a derivation of  $R$  regarded as an  $\mathbf{R}$ -module. The differential  $d$  is also regarded as a mapping  $d : R \rightarrow (R^{\tilde{n}})^*$ , where  $(R^{\tilde{n}})^*$  denotes the dual module of  $R^{\tilde{n}}$ , and  $(R^{\tilde{n}})^*$  can be identified with  $R^{1 \times \tilde{n}}$ . Through the use of componentwise application of the substitution mapping  $\alpha^* : R \rightarrow \alpha^*(R) \subset \mathbf{R}(\alpha)$ , we can define  $\bar{\alpha}^* : R^{\tilde{n}} \rightarrow \alpha^*(R)^{\tilde{n}}$  and  $\underline{\alpha}^* : (R^{\tilde{n}})^* \rightarrow (\alpha^*(R)^{\tilde{n}})^*$ .

An important algebraic structure in a system over a ring is the invariance of an ideal under the Lie derivative.

**DEFINITION 3.3.** *An ideal  $I \subset R$  is said to be an invariant ideal of a system  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$  over  $R$  if*

$$L_{\tilde{g}_i} I \subset I, \quad i \in I_{0,m},$$

*holds.*

If the ring  $R$  is Noetherian, every ideal is finitely generated. In other words, for any ideal  $I \subset R$ , there exist a finite number of elements  $\tilde{f}_1, \dots, \tilde{f}_s \in R$  such that  $I = \{\sum_{i=1}^s a_i \tilde{f}_i : a_i \in R\}$ , which is denoted by  $I = (\tilde{f}_1, \dots, \tilde{f}_s)$ . For example,  $\mathbf{R}[\tilde{x}]$  is Noetherian by Hilbert's basis theorem, and its localization  $\mathbf{R}[\tilde{x}]_P$  is also Noetherian [17]. It is readily shown that a finitely generated ideal  $I = (\tilde{f}_1, \dots, \tilde{f}_s)$  is an invariant ideal if and only if

$$(3.1) \quad L_{\tilde{g}_i} \tilde{f}_j \in (\tilde{f}_1, \dots, \tilde{f}_s), \quad i \in I_{0,m}, \quad j \in I_{1,s},$$

holds. Then, it is meaningful to consider a differential algebraic equation (DAE) over  $R$ :

$$(3.2) \quad \tilde{\Sigma}_I \begin{cases} \dot{\tilde{x}} = \tilde{g}_0(x) + \sum_{i=1}^m \tilde{g}_i(\tilde{x}) u_i, \\ 0 = \tilde{f}(\tilde{x}), \end{cases}$$

where  $\tilde{f}(\tilde{x}) = [\tilde{f}_1(\tilde{x}), \dots, \tilde{f}_s(\tilde{x})]^T \in R^s$ . In this paper, we call  $\tilde{x}_0 \in \mathbf{R}^{\tilde{n}}$  a *regular point* of DAE  $\tilde{\Sigma}_I$  if all denominators in  $\tilde{g}_i$  and  $\tilde{f}_i$  are nonzero at  $x_0$ . Equation (3.1) implies that if a regular point  $x_0$  satisfies  $\tilde{f}(x_0) = 0$ , the trajectory starting from  $x_0$  satisfies  $\tilde{f}(\Phi_t^{\Sigma, u}(x_0)) = 0$  for every admissible input  $u \in \Omega$ , as long as the trajectory  $\Phi_t^{\Sigma, u}(x_0)$  is defined. If  $\tilde{\Sigma}$  is a PSR,  $R = \mathbf{R}[\tilde{x}]$  and every point in  $\mathbf{R}^{\tilde{n}}$  is a regular point of DAE  $\tilde{\Sigma}_I$ .

Since a geometric counterpart of an ideal is an affine variety [7], a geometric object related to an invariant ideal is naturally defined as an invariant variety.

**DEFINITION 3.4.** *Let  $I$  be an ideal of  $\mathbf{R}[\tilde{x}]$ . An affine variety (or an algebraic set) defined by  $I$  is a subset of  $\mathbf{R}^{\tilde{n}}$  given by*

$$\mathcal{V}(I) = \{\tilde{x} \in \mathbf{R}^{\tilde{n}} : \tilde{f}(\tilde{x}) = 0 \text{ for all } \tilde{f} \in I\}.$$

When  $I = (\tilde{f}_1, \dots, \tilde{f}_s)$ ,  $\mathcal{V}(I)$  is also denoted by  $\mathcal{V}(\tilde{f}_1, \dots, \tilde{f}_s)$ .

**DEFINITION 3.5.** *Let  $R$  be a partial differential ring such that  $\mathbf{R}[\tilde{x}] \subset R \subset \mathbf{R}(\tilde{x})$ , and let  $P$  be an ideal of  $\mathbf{R}[\tilde{x}]$ .  $\mathcal{V}(P)$  is called an invariant variety of a system over  $R$  if the ideal generated by  $P$  in  $R$ ,  $PR$  is an invariant ideal of the system.*

Note that if  $R = \mathbf{R}[\tilde{x}]$ , then  $PR = P$ , and if  $R = \mathbf{R}[\tilde{x}]_P$ , then  $PR = P\mathbf{R}[\tilde{x}]_P$ . Since  $\mathbf{R}[\tilde{x}]$  is Noetherian,  $\mathcal{V}(P)$  can be expressed as a set of the common zeros of a finite number of polynomials. By the invariance of the ideal  $PR$ , these polynomials are identically zero along a trajectory starting from a regular point  $\tilde{x}_0 \in \mathcal{V}(P)$  of  $\tilde{\Sigma}$ . That is, the trajectory stays in  $\mathcal{V}(P)$  as long as it is defined, which motivates the notion of the invariant variety.

The notions of an invariant ideal and an invariant variety can be regarded as generalizations of an *algebraic particular integral* and *invariant algebraic surface* [3, 25] of a polynomial vector field. It should be noted that an invariant variety may have a singular point as a variety and is not necessarily a manifold globally. In particular, the existence of an invariant variety does not necessarily imply the existence of a foliation of manifolds.

**3.3. Invariance in a system after immersion.** Through the use of the algebraic and geometric notions defined above, we can characterize an invariance in a system after immersion. Hereafter, we assume  $R$  to be a partial differential ring such that  $\mathbf{R}[\tilde{x}] \subset R \subset \mathbf{R}(\tilde{x})$ .

**THEOREM 3.6.** *Let  $\alpha$  be an invariant immersion of system  $\Sigma(g_0, \dots, g_m, h)$  defined on a domain  $U \subset \mathbf{R}^n$  into system  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$  over  $R$ , and let  $P \subset \mathbf{R}[\tilde{x}]$  be the relation ideal of  $\alpha$ . Then,  $PR$  is an invariant ideal of  $\tilde{\Sigma}$ .*

*Proof.* For  $r \in PR$ , we relate two Lie derivatives  $L_{\tilde{g}_i}r$  and  $L_{g_i}(r \circ \alpha)$  through the definition of an invariant immersion to show  $L_{\tilde{g}_i}r \in PR$ . First,  $L_{\tilde{g}_i}r$  belongs to  $R$  because  $R$  is a partial differential ring and  $\tilde{g}_i$  belongs to  $R^{\tilde{n}}$ . Moreover, since  $\text{Ker } \alpha^* = PR$ , there exists an open and dense subset  $U' \subset U$  such that  $r$  is analytic on an open set containing  $\alpha(U')$  and  $r(\alpha(x)) = 0$  for all  $x \in U'$ . Therefore, we have  $L_{g_i}r(\alpha(x)) = 0$  ( $i \in I_{0,m}$ ) on  $U'$ . Meanwhile, from the definition of an invariant immersion, we have

$$\begin{aligned} L_{g_i}r(\alpha(x)) &= \frac{\partial r(\alpha(x))}{\partial \tilde{x}} \frac{\partial \alpha(x)}{\partial x} g_i(x) \\ &= \frac{\partial r(\alpha(x))}{\partial \tilde{x}} \tilde{g}_i(\alpha(x)) = (L_{\tilde{g}_i}r)(\alpha(x)) \end{aligned}$$

for all  $x \in U'$ . In summary,  $(L_{\tilde{g}_i}r)(\tilde{x}) = 0$  for all  $\tilde{x} \in \alpha(U')$ , which means  $L_{\tilde{g}_i}r \in PR$ .  $\square$

Then, the invariance of an ideal implies the invariance of a variety in a system after immersion, which is characterized in terms of the Zariski closure [7].

**DEFINITION 3.7.** *The Zariski closure of a subset  $S \subset \mathbf{R}^{\tilde{n}}$  is the smallest variety containing  $S$ .*

**PROPOSITION 3.8.** *The Zariski closure of  $S \subset \mathbf{R}^{\tilde{n}}$  is given by  $\mathcal{V}(\mathcal{I}(S))$ .*

**THEOREM 3.9.** *Let  $\alpha$ ,  $\tilde{\Sigma}$ , and  $P$  be the same as in Theorem 3.6. Then,  $\mathcal{V}(P)$  is the Zariski closure of  $\alpha(U)$  and, moreover, an invariant variety of  $\tilde{\Sigma}$ .*

*Proof.* Since  $P$  is the relation ideal of  $\alpha$ , i.e.,  $P = \mathcal{I}(\alpha(U))$ , Proposition 3.8 implies that  $\mathcal{V}(P)$  is the Zariski closure of  $\alpha(U)$ . Moreover, since  $PR$  is an invariant ideal of  $\tilde{\Sigma}$  by Theorem 3.6,  $\mathcal{V}(P)$  is also an invariant variety.  $\square$

For an invariant immersion  $\alpha : U \rightarrow \mathbf{R}^{\tilde{n}}$  of a given system  $\Sigma$  on an open and dense subset  $U' \subset U$  into a system over  $R$ ,  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$ , its relation ideal  $P = \mathcal{I}(\alpha(U)) \subset \mathbf{R}[\tilde{x}]$  is always finitely generated according to Hilbert's basis theorem, and its generators are also generators of the ideal  $PR$  in  $R$ . If  $P = (\tilde{f}_1, \dots, \tilde{f}_s)$  holds, then an image of any trajectory of the original system by the immersion  $\alpha(\Phi_t^{\Sigma, u}(x_0))$

$(x_0 \in U')$  is always a solution of the DAE given in (3.2). That is,  $\alpha(\Phi_t^{\Sigma, u}(x_0))$  ( $x_0 \in U'$ ) belongs to  $\mathcal{V}(\tilde{f}_1, \dots, \tilde{f}_s)$ . Moreover, Theorem 3.9 implies that not only the image of a trajectory of the original system but also any trajectory starting from a point on  $\mathcal{V}(\tilde{f}_1, \dots, \tilde{f}_s)$  remains on  $\mathcal{V}(\tilde{f}_1, \dots, \tilde{f}_s)$ , as long as it is defined.

*Example 3.1.* Consider a one-dimensional analytic system on  $\mathbf{R}$ :

$$\Sigma \left\{ \begin{array}{l} \dot{x} = \frac{\sin x}{x}, \\ y = x. \end{array} \right.$$

The observation space of this system is given by

$$\mathcal{O}_\Sigma = \text{span}_{\mathbf{R}} \left\{ x, \frac{\sin x}{x}, \frac{\cos x \cdot x - \sin x}{x^2} \cdot \frac{\sin x}{x}, \dots \right\}.$$

We have  $\mathbf{R}(\mathcal{O}_\Sigma) = \mathbf{R}(x, \sin x, \cos x)$ , and  $\Sigma$  is immersible into an RSR with  $\alpha(x) = [x, \sin x, \cos x]^T$ . In fact, an RSR  $\tilde{\Sigma}(\tilde{g}_0, \tilde{h})$  can be readily constructed from  $(\partial\alpha/\partial x)\dot{x} = [\sin x/x, \sin x \cdot \cos x/x, -\sin^2 x/x]^T$  as follows:

$$\tilde{\Sigma} \left\{ \begin{array}{l} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \end{bmatrix} = \begin{bmatrix} \tilde{x}_2/\tilde{x}_1 \\ \tilde{x}_2\tilde{x}_3/\tilde{x}_1 \\ -\tilde{x}_2^2/\tilde{x}_1 \end{bmatrix}, \\ y = \tilde{x}_1. \end{array} \right.$$

The RSR  $\tilde{\Sigma}$  is analytic on an open set  $\tilde{U} = \{\tilde{x} \in \mathbf{R}^3 : \tilde{x}_1 \neq 0\} \supset \alpha(\mathbf{R} \setminus \{0\})$ . Therefore,  $\Sigma$  is immersible into the RSR  $\tilde{\Sigma}$  on  $\mathbf{R} \setminus \{0\}$ . The immersion  $\alpha$  itself is analytic on  $\mathbf{R}$  and its image is a helix along the  $\tilde{x}_1$  axis.

From the algebraic relation between the trigonometric functions,  $\sin^2 x + \cos^2 x - 1 = 0$ , the relation ideal of  $\alpha$  is  $P = (\tilde{x}_2^2 + \tilde{x}_3^2 - 1)$ . The maximal ideal of the local ring  $\mathbf{R}[\tilde{x}]_P$  has the form

$$P\mathbf{R}[\tilde{x}]_P = \left\{ \frac{n(\tilde{x})}{d(\tilde{x})}(\tilde{x}_2^2 + \tilde{x}_3^2 - 1) : n, d \in \mathbf{R}[\tilde{x}], \text{ and } d \notin P \right\}.$$

Note that every element of  $P\mathbf{R}[\tilde{x}]_P$  vanishes when  $\tilde{x} = \alpha(x)$  is substituted.

Theorem 3.6 implies that  $P\mathbf{R}[\tilde{x}]_P$  is an invariant ideal of  $\tilde{\Sigma}$ . In fact, for any  $n(\tilde{x})(\tilde{x}_2^2 + \tilde{x}_3^2 - 1)/d(\tilde{x}) \in P\mathbf{R}[\tilde{x}]_P$ , we have

$$\begin{aligned} & L_{\tilde{g}_0} \left[ \frac{n(\tilde{x})}{d(\tilde{x})}(\tilde{x}_2^2 + \tilde{x}_3^2 - 1) \right] \\ &= \left[ L_{\tilde{g}_0} \frac{n(\tilde{x})}{d(\tilde{x})} \right] (\tilde{x}_2^2 + \tilde{x}_3^2 - 1) + \frac{n(\tilde{x})}{d(\tilde{x})} L_{\tilde{g}_0} (\tilde{x}_2^2 + \tilde{x}_3^2 - 1) \\ &= \left[ L_{\tilde{g}_0} \frac{n(\tilde{x})}{d(\tilde{x})} \right] (\tilde{x}_2^2 + \tilde{x}_3^2 - 1) + \frac{n(\tilde{x})}{d(\tilde{x})} \left( 2\tilde{x}_2 \cdot \frac{\tilde{x}_2\tilde{x}_3}{\tilde{x}_1} + 2\tilde{x}_3 \cdot \frac{-\tilde{x}_2^2}{\tilde{x}_1} \right) \\ &= \left[ L_{\tilde{g}_0} \frac{n(\tilde{x})}{d(\tilde{x})} \right] (\tilde{x}_2^2 + \tilde{x}_3^2 - 1). \end{aligned}$$

Note that the Lie derivative of  $n(\tilde{x})(\tilde{x}_2^2 + \tilde{x}_3^2 - 1)/d(\tilde{x}) \in P\mathbf{R}[\tilde{x}]_P$  vanishes even when  $\tilde{x}$  does not belong to the image of  $\alpha$ .

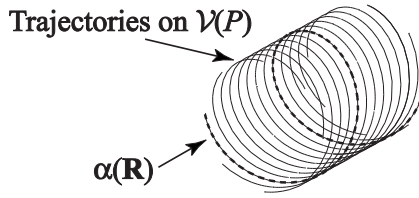


FIG. 3.1. Trajectories on the invariant variety.

Finally, Theorem 3.9 claims that a variety  $\mathcal{V}(\tilde{x}_2^2 + \tilde{x}_3^2 - 1)$ , a cylinder along the  $\tilde{x}_1$  axis, is the Zariski closure of  $\alpha(\mathbf{R})$ , a helix, and an invariant variety of  $\tilde{\Sigma}$ . Then, for any solution  $x(t)$  of  $\Sigma$ , its image  $\alpha(x(t))$  satisfies the DAE

$$\tilde{\Sigma}_P \left\{ \begin{array}{l} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \end{bmatrix} = \begin{bmatrix} \tilde{x}_2/\tilde{x}_1 \\ \tilde{x}_2\tilde{x}_3/\tilde{x}_1 \\ -\tilde{x}_2^2/\tilde{x}_1 \end{bmatrix}, \\ 0 = \tilde{x}_2^2 + \tilde{x}_3^2 - 1. \end{array} \right.$$

Moreover, any solution of the state equation of  $\tilde{\Sigma}$  starting from a point on  $\mathcal{V}(\tilde{x}_2^2 + \tilde{x}_3^2 - 1)$  satisfies the DAE, as long as it is defined (see Figure 3.1).

*Example 3.2.* According to Proposition 2.5,  $\Sigma$  in Example 3.1 is invariantly immersible on  $\mathbf{R} \setminus \{0\}$  into not only an RSR but also a PSR. In fact, an invariant immersion into a PSR is readily obtained by augmenting  $\alpha(x)$  with  $1/x$  ( $= 1/\tilde{x}_1$  on  $\alpha(\mathbf{R})$ ) as  $\beta(x) = [x, \sin x, \cos x, 1/x]^T$ , and a PSR  $\bar{\Sigma}(\bar{g}_0, \bar{h})$  is obtained from  $(\partial\beta/\partial x)\dot{x} = [\sin x/x, \sin x \cdot \cos x/x, -\sin^2 x/x, -\sin x/x^3]^T$  as follows:

$$\bar{\Sigma} \left\{ \begin{array}{l} \begin{bmatrix} \dot{\bar{x}}_1 \\ \dot{\bar{x}}_2 \\ \dot{\bar{x}}_3 \\ \dot{\bar{x}}_4 \end{bmatrix} = \begin{bmatrix} \bar{x}_2\bar{x}_4 \\ \bar{x}_2\bar{x}_3\bar{x}_4 \\ -\bar{x}_2^2\bar{x}_4 \\ -\bar{x}_2\bar{x}_4^3 \end{bmatrix}, \\ y = \bar{x}_1. \end{array} \right.$$

The PSR  $\bar{\Sigma}$  is analytic on  $\mathbf{R}^4 \supset \beta(\mathbf{R} \setminus \{0\})$ .

From an additional algebraic relation  $x \cdot (1/x) - 1 = 0$ , the relation ideal of  $\beta$  is  $Q = (\bar{x}_2^2 + \bar{x}_3^2 - 1, \bar{x}_1\bar{x}_4 - 1)$ . Then, Theorem 3.6 implies that  $Q$  is an invariant ideal of  $\bar{\Sigma}$ . In fact, it is readily confirmed that  $L_{\bar{g}_0}Q \subset Q$ . For example, we have, for any  $p \in \mathbf{R}[\bar{x}]$ ,

$$\begin{aligned} L_{\bar{g}_0}[p(\bar{x})(\bar{x}_1\bar{x}_4 - 1)] &= [L_{\bar{g}_0}p(\bar{x})](\bar{x}_1\bar{x}_4 - 1) + p(\bar{x})[\bar{x}_4 \cdot \bar{x}_2\bar{x}_4 + \bar{x}_1 \cdot (-\bar{x}_2\bar{x}_4)] \\ &= [L_{\bar{g}_0}p(\bar{x}) - p(\bar{x})\bar{x}_2\bar{x}_4^2](\bar{x}_1\bar{x}_4 - 1) \in Q. \end{aligned}$$

Finally, Theorem 3.9 states that  $\mathcal{V}(\bar{x}_2^2 + \bar{x}_3^2 - 1, \bar{x}_1\bar{x}_4 - 1)$  is the Zariski closure of  $\beta(\mathbf{R} \setminus \{0\})$  and an invariant variety of  $\bar{\Sigma}$ . Then, for any solution  $x(t)$  of  $\Sigma$ , its image

$\beta(x(t))$  satisfies the DAE

$$\bar{\Sigma}_Q \left\{ \begin{array}{l} \begin{bmatrix} \dot{\bar{x}}_1 \\ \dot{\bar{x}}_2 \\ \dot{\bar{x}}_3 \\ \dot{\bar{x}}_4 \end{bmatrix} = \begin{bmatrix} \bar{x}_2 \bar{x}_4 \\ \bar{x}_2 \bar{x}_3 \bar{x}_4 \\ -\bar{x}_2^2 \bar{x}_4 \\ -\bar{x}_2 \bar{x}_4^3 \end{bmatrix}, \\ 0 = \bar{x}_2^2 + \bar{x}_3^2 - 1, \\ 0 = \bar{x}_1 \bar{x}_4 - 1. \end{array} \right.$$

Moreover, any solution of  $\bar{\Sigma}$  starting from a point on  $\mathcal{V}(\bar{x}_2^2 + \bar{x}_3^2 - 1, \bar{x}_1 \bar{x}_4 - 1)$  satisfies the DAE, as long as it is defined.

**4. Degrees of freedom in systems over a ring.** For a given nonlinear system, an immersion into a system over a ring is not unique, and it is meaningful to discuss how to choose an immersion, as in Theorem 2.7. Moreover, even if an immersion is fixed, a system over a ring is not unique. In order to choose an appropriate representation for system analysis and control design, it is important to clarify the degrees of freedom in systems over a ring obtained by one immersion.

From the definition of an invariant immersion and the fact  $\text{Ker } \alpha^* = PR$ , it is straightforward to show the following parameterization of all systems over  $R$  into which a given system is invariantly immersible with the same immersion  $\alpha$ .

**THEOREM 4.1.** *Let  $\Sigma$  be an analytic system defined on a domain  $U \subset \mathbf{R}^n$ , let  $\alpha : U \rightarrow \mathbf{R}^{\tilde{n}}$  be an invariant immersion of  $\Sigma$  on an open and dense subset of  $U$  into a system over a ring  $R$ ,  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$ , and let  $P \subset \mathbf{R}[\tilde{x}]$  be the relation ideal of  $\alpha$ . Then all systems over  $R$  into which  $\Sigma$  is invariantly immersible on an open and dense subset of  $U$  with  $\alpha$  can be parameterized as  $\tilde{\Sigma}'(\tilde{g}_0 + v_0, \dots, \tilde{g}_m + v_m, \tilde{h} + r)$  with  $v_i \in PR^{\tilde{n}}$  ( $i \in I_{0,m}$ ) and  $r \in PR_P^p$ .*

*Proof.* First, if  $\tilde{\Sigma}$ ,  $v_i \in PR^{\tilde{n}}$  ( $i \in I_{0,m}$ ), and  $r \in PR^p$  are analytic on the image of an open and dense subset  $U' \subset U$  by  $\alpha$ , then it is obvious from the definition of an invariant immersion and  $v_i(\alpha(x)) = 0$ ,  $r(\alpha(x)) = 0$  (for all  $x \in U'$ ) that  $\alpha$  is an invariant immersion of  $\Sigma$  on  $U'$  into a system over  $R$ ,  $\tilde{\Sigma}'(\tilde{g}_0 + v_0, \dots, \tilde{g}_m + v_m, \tilde{h} + r)$ .

Conversely, if  $\alpha$  is an invariant immersion of  $\Sigma$  on an open and dense subset  $U' \subset U$  into not only  $\tilde{\Sigma}$  but also another system over  $R$ ,  $\tilde{\Sigma}'(\tilde{g}'_0, \dots, \tilde{g}'_m, \tilde{h}')$ , then all elements of  $\tilde{g}'_i$  ( $i \in I_{0,m}$ ) and  $\tilde{h}'$  also belong to  $R$ , and we have, for all  $x \in U'$ ,

$$L_{g_i} \alpha(x) = \tilde{g}_i(\alpha(x)) = \tilde{g}'_i(\alpha(x)), \quad i \in I_{0,m},$$

$$h(x) = \tilde{h}(\alpha(x)) = \tilde{h}'(\alpha(x)).$$

Therefore, for all  $x \in U'$ ,

$$\alpha^*(\tilde{g}_i - \tilde{g}'_i)(x) = \tilde{g}_i(\alpha(x)) - \tilde{g}'_i(\alpha(x)) = 0, \quad i \in I_{0,m},$$

$$\alpha^*(\tilde{h} - \tilde{h}')(x) = \tilde{h}(\alpha(x)) - \tilde{h}'(\alpha(x)) = 0,$$

which implies  $\tilde{g}_i - \tilde{g}'_i \in PR^{\tilde{n}}$  ( $i \in I_{0,m}$ ) and  $\tilde{h} - \tilde{h}' \in RR^p$ .  $\square$

Note that, as mentioned previously,  $\tilde{\Sigma}$  and  $\tilde{\Sigma}'$  are RSRs if  $R = \mathbf{R}[\tilde{x}]_P$  and PSRs if  $R = \mathbf{R}[\tilde{x}]$ . It is obvious that if  $\tilde{\Sigma}$  is a QSR, the parameterization of all QSRs is obtained with such additional constraints to the case of PSRs as  $\tilde{g}_i + v_i \in \mathbf{R}[\tilde{x}]_{\leq 2}^{\tilde{n}}$  ( $i \in I_{0,m}$ ) and  $\tilde{h} + r \in \mathbf{R}[\tilde{x}]_{\leq 1}^p$ .



Now, the remaining problem is the characterization of all systems over  $R$  into which a given system is immersible with the same immersion  $\alpha$  that is not necessarily an invariant immersion.

**THEOREM 4.2.** *Let  $\Sigma$  be an analytic system defined on a domain  $U \subset \mathbf{R}^n$ , let  $\alpha : U \rightarrow \mathbf{R}^{\tilde{n}}$  be an invariant immersion of  $\Sigma$  on an open and dense subset of  $U$  into a system over  $R$ ,  $\tilde{\Sigma}(\tilde{g}_0, \dots, \tilde{g}_m, \tilde{h})$ , and let  $P \subset \mathbf{R}[\tilde{x}]$  be the relation ideal of  $\alpha$ . Let  $\tilde{\Sigma}'(\tilde{g}'_0, \dots, \tilde{g}'_m, \tilde{h}')$  be another system over  $R$ , and let  $v_i = \tilde{g}'_i - \tilde{g}_i$  ( $i \in I_{0,m}$ ) and  $r = \tilde{h}' - \tilde{h}$ . Then,  $\Sigma$  is immersible on an open and dense subset of  $U$  with  $\alpha$  into  $\tilde{\Sigma}'$  if and only if*

$$L_{v_i} \mathcal{L}_{\tilde{\Sigma}'} \subset PR, \quad i \in I_{0,m},$$

$$r \in PR^p.$$

*Proof.* Let  $U'$  be an open and dense subset of  $U$  such that both  $\tilde{\Sigma}$  and  $\tilde{\Sigma}'$  are analytic on  $\alpha(U')$ . From Proposition 2.2,  $\alpha$  is also an immersion of  $\Sigma$  on  $U'$  into  $\tilde{\Sigma}'$  if and only if, for all  $x \in U'$ ,

$$L_{\tilde{g}_{i_1}} \dots L_{\tilde{g}_{i_k}} \tilde{h}(\alpha(x)) - L_{\tilde{g}'_{i_1}} \dots L_{\tilde{g}'_{i_k}} \tilde{h}'(\alpha(x)) = 0,$$

$$(i_1, \dots, i_k) \in I_{0,m}^k, \quad k \in I_{0,\infty},$$

which is equivalent to

$$L_{\tilde{g}_{i_1}} \dots L_{\tilde{g}_{i_k}} \tilde{h}_j - L_{\tilde{g}'_{i_1}} \dots L_{\tilde{g}'_{i_k}} \tilde{h}'_j \in PR,$$

$$(4.1) \quad j \in I_{1,p}, \quad (i_1, \dots, i_k) \in I_{0,m}^k, \quad k \in I_{0,\infty}.$$

We will show that the claimed condition is equivalent to condition (4.1).

(If) We show the sufficiency by induction. First, for  $k = 0$ , condition (4.1) holds because  $\tilde{h}' - \tilde{h} = r \in PR^p$  by the assumption.

Next, suppose condition (4.1) holds for a certain  $k$ , and let  $\tilde{H} \in \mathcal{L}_{\tilde{\Sigma}}$  and  $\tilde{H}' \in \mathcal{L}_{\tilde{\Sigma}'}$  be  $k$ th order Lie derivatives corresponding to each other such that  $\tilde{H} - \tilde{H}' \in PR$ . Since  $PR$  is an invariant ideal of  $\tilde{\Sigma}$  by Theorem 3.6, we have, for any  $l \in I_{0,m}$ ,  $L_{\tilde{g}_l}(\tilde{H} - \tilde{H}') \in PR$ , which implies

$$\begin{aligned} L_{\tilde{g}_l} \tilde{H} - L_{\tilde{g}_l} \tilde{H}' &= L_{\tilde{g}_l} \tilde{H} - L_{\tilde{g}'_l - v_l} \tilde{H}' \\ &= L_{\tilde{g}_l} \tilde{H} - L_{\tilde{g}'_l} \tilde{H}' + L_{v_l} \tilde{H}' \in PR. \end{aligned}$$

Since  $L_{v_l} \tilde{H}' \in PR$  from the assumption, we have  $L_{\tilde{g}_l} \tilde{H} - L_{\tilde{g}'_l} \tilde{H}' \in PR$ , which means that condition (4.1) also holds for  $k + 1$ .

(Only if) First, condition (4.1) with  $k = 0$  implies  $\tilde{h}' - \tilde{h} = r \in PR^p$ . Condition (4.1) also implies

$$\begin{aligned} L_{\tilde{g}_{i_1}} \dots L_{\tilde{g}_{i_k}} \tilde{h}_j - L_{\tilde{g}'_{i_1}} \dots L_{\tilde{g}'_{i_k}} \tilde{h}'_j &= L_{\tilde{g}_{i_1}} (L_{\tilde{g}_{i_2}} \dots L_{\tilde{g}_{i_k}} \tilde{h}_j - L_{\tilde{g}_{i_2}} \dots L_{\tilde{g}'_{i_k}} \tilde{h}'_j) \\ &\quad - L_{v_{i_1}} (L_{\tilde{g}'_{i_2}} \dots L_{\tilde{g}'_{i_k}} \tilde{h}'_j) \in PR, \end{aligned}$$

$$j \in I_{1,p}, \quad (i_1, \dots, i_k) \in I_{0,m}^k, \quad k \in I_{0,\infty},$$

where  $L_{\tilde{g}_{i_2}} \dots L_{\tilde{g}_{i_k}} \tilde{h}_j - L_{\tilde{g}_{i_2}} \dots L_{\tilde{g}'_{i_k}} \tilde{h}'_j \in PR$  by condition (4.1) and its Lie derivative also belongs to  $PR$  by Theorem 3.6. Therefore,  $L_{v_{i_1}}(L_{\tilde{g}'_{i_2}} \dots L_{\tilde{g}'_{i_k}} \tilde{h}'_j)$  should also belong to  $PR$ , which means  $L_{v_i} \mathcal{L}_{\tilde{\Sigma}'} \in PR$ .  $\square$

It should be noted that the state equation of  $\tilde{\Sigma}'(\tilde{g}'_0, \dots, \tilde{g}'_m, \tilde{h}')$  can be rewritten as

$$\begin{aligned} \dot{\tilde{x}} &= \tilde{g}'_0(\tilde{x}) + \sum_{i=1}^m \tilde{g}'_i(\tilde{x}) u_i \\ &= \tilde{g}_0(\tilde{x}) + \sum_{i=1}^m \tilde{g}_i(\tilde{x}) u_i + \sum_{i=0}^m v_i(\tilde{x}) w_i, \end{aligned}$$

with  $w_0 = 1$ ,  $w_1 = u_1, \dots, w_m = u_m$ . That is,  $v_i$  can be viewed as vector fields corresponding to disturbances on  $\tilde{\Sigma}$ . Then, since every element of  $PR$  vanishes on  $\mathcal{V}(P)$ , Theorem 4.2 can be interpreted as a generalization of the condition for the output invariance [11, 21] with the initial state restricted to  $\mathcal{V}(P)$ .

According to Theorem 4.2, the output mapping of a system over  $R$  has a degree of freedom in  $PR^p$ , as in the case of an invariant immersion. However, the condition  $L_{v_i} \mathcal{L}_{\tilde{\Sigma}'} \subset PR$  is not useful for finding the free parameter  $v_i$  in the vector fields because  $\mathcal{L}_{\tilde{\Sigma}'}$  itself contains  $v_i$ . The following sufficient conditions give explicit expressions for  $v_i$  such that  $L_{v_i} \mathcal{L}_{\tilde{\Sigma}'} \subset PR$  holds. Note that  $\mathcal{L}_{\tilde{\Sigma}}$  in the following does not contain  $v_i$ .

**THEOREM 4.3.** *Let  $\alpha$ ,  $\tilde{\Sigma}$ ,  $P$ , and  $\tilde{\Sigma}'$  be the same as in Theorem 4.2.  $\Sigma$  is immersible on an open and dense subset of  $U$  with  $\alpha$  into  $\tilde{\Sigma}'$  if*

$$\begin{aligned} L_{v_i} \mathcal{L}_{\tilde{\Sigma}} &\subset PR, \quad L_{v_i} PR \subset PR, \quad i \in I_{0,m}, \\ r &\in PR^p, \end{aligned}$$

or, equivalently,

$$\begin{aligned} v_i &\in \bar{\alpha}^{*-1}(\text{Ker } \underline{\alpha}^*(d\mathcal{L}_{\tilde{\Sigma}} \cup d(PR))), \quad i \in I_{0,m}, \\ r &\in PR^p. \end{aligned}$$

*Proof.* We show that condition (4.1) holds by induction. First, for  $k = 0$ , condition (4.1) holds because  $\tilde{h}' - \tilde{h} = r \in PR^p$  by the assumption. Next, suppose condition (4.1) holds for a certain  $k$ , and let  $\tilde{H} \in \mathcal{L}_{\tilde{\Sigma}}$  and  $\tilde{H}' \in \mathcal{L}_{\tilde{\Sigma}'}$  be  $k$ th order Lie derivatives corresponding to each other such that  $\tilde{H} - \tilde{H}' \in PR$ . Then, we have, for any  $l \in I_{0,m}$ ,

$$\begin{aligned} L_{\tilde{g}_l} \tilde{H} - L_{\tilde{g}'_l} \tilde{H}' &= L_{\tilde{g}_l} \tilde{H} - L_{\tilde{g}_l + v_l}(\tilde{H} + (\tilde{H}' - \tilde{H})) \\ &= -L_{\tilde{g}_l}(\tilde{H}' - \tilde{H}) - L_{v_l} \tilde{H} - L_{v_l}(\tilde{H}' - \tilde{H}), \end{aligned}$$

where every term on the right-hand side belongs to  $PR$  by the invariance of  $PR$  and the assumptions. Therefore, we have  $L_{\tilde{g}_l} \tilde{H} - L_{\tilde{g}'_l} \tilde{H}' \in PR$ , which means that condition (4.1) also holds for  $k + 1$ .

Next, we derive an explicit expression for  $v_i$  from the conditions  $L_{v_i} \mathcal{L}_{\tilde{\Sigma}} \subset PR$  and  $L_{v_i} PR \subset PR$ . The conditions mean that, for any  $\tilde{H} \in \mathcal{L}_{\tilde{\Sigma}}$  and any  $s \in PR$ , we have

$$d\tilde{H}(v_i) = \frac{\partial \tilde{H}}{\partial \tilde{x}} v_i \in PR, \quad ds(v_i) = \frac{\partial s}{\partial \tilde{x}} v_i \in PR,$$

or, equivalently,

$$\begin{aligned}\alpha^*(d\tilde{H}(v_i)) &= \underline{\alpha}^*(d\tilde{H})(\bar{\alpha}^*(v_i)) = 0, \\ \alpha^*(ds(v_i)) &= \underline{\alpha}^*(ds)(\bar{\alpha}^*(v_i)) = 0.\end{aligned}$$

That is,

$$\bar{\alpha}^*(v_i) \in \text{Ker } \underline{\alpha}^*(d\tilde{H}), \quad \bar{\alpha}^*(v_i) \in \text{Ker } \underline{\alpha}^*(ds).$$

Therefore,

$$\begin{aligned}\bar{\alpha}^*(v_i) &\in \bigcap_{\tilde{H} \in \mathcal{L}_{\tilde{\Sigma}}} \text{Ker } \underline{\alpha}^*(d\tilde{H}) = \text{Ker } \underline{\alpha}^*(d\mathcal{L}_{\tilde{\Sigma}}), \\ \bar{\alpha}^*(v_i) &\in \bigcap_{s \in PR} \text{Ker } \underline{\alpha}^*(ds) = \text{Ker } \underline{\alpha}^*(d(PR)),\end{aligned}$$

and, moreover,

$$\bar{\alpha}^*(v_i) \in \text{Ker } \underline{\alpha}^*(d\mathcal{L}_{\tilde{\Sigma}}) \cap \text{Ker } \underline{\alpha}^*(d(PR)) = \text{Ker } \underline{\alpha}^*(d\mathcal{L}_{\tilde{\Sigma}} \cup d(PR)),$$

which is equivalent to  $v_i \in \bar{\alpha}^{*-1}(\text{Ker } \underline{\alpha}^*(d\mathcal{L}_{\tilde{\Sigma}} \cup d(PR)))$ .  $\square$

*Example 4.1.* Consider  $\Sigma$  in Example 3.1, which is invariantly immersible with  $\alpha$  into system  $\tilde{\Sigma}$  over the local ring  $\mathbf{R}[\tilde{x}]_P$ . Theorem 4.1 implies that all RSRs into which  $\Sigma$  is invariantly immersible with  $\alpha$  are parameterized as

$$(4.2) \quad \tilde{\Sigma}' \left\{ \begin{array}{l} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \end{bmatrix} = \begin{bmatrix} \tilde{x}_2/\tilde{x}_1 + v_{01}(\tilde{x}) \\ \tilde{x}_2\tilde{x}_3/\tilde{x}_1 + v_{02}(\tilde{x}) \\ -\tilde{x}_2^2/\tilde{x}_1 + v_{03}(\tilde{x}) \end{bmatrix}, \\ y = \tilde{x}_1 + r(\tilde{x}), \end{array} \right.$$

where  $v_{01}$ ,  $v_{02}$ ,  $v_{03}$ , and  $r$  belong to  $PR[\tilde{x}]_P$ . It can readily be shown that  $PR[\tilde{x}]_P$  is also the invariant ideal of  $\tilde{\Sigma}'$ .

In order to demonstrate Theorems 4.2 and 4.3, let us consider another mapping  $\beta(x) = [x, \sin x, \cos x, \tan x]^T$ . It is straightforward to show that  $\mathbf{R}(\mathcal{O}_{\Sigma}) = \mathbf{R}(\beta)$  and that  $\beta$  is an invariant immersion of  $\Sigma$  into the following RSR,  $\bar{\Sigma}(\bar{g}_0, \bar{h})$ :

$$\bar{\Sigma} \left\{ \begin{array}{l} \begin{bmatrix} \dot{\bar{x}}_1 \\ \dot{\bar{x}}_2 \\ \dot{\bar{x}}_3 \\ \dot{\bar{x}}_4 \end{bmatrix} = \begin{bmatrix} \bar{x}_2/\bar{x}_1 \\ \bar{x}_2\bar{x}_3/\bar{x}_1 \\ -\bar{x}_2^2/\bar{x}_1 \\ \bar{x}_2/(\bar{x}_1\bar{x}_3^2) \end{bmatrix}, \\ y = \bar{x}_1. \end{array} \right.$$

The relation ideal of  $\beta$  is  $\bar{P} = (\bar{x}_2^2 + \bar{x}_3^2 - 1, \bar{x}_2 - \bar{x}_3\bar{x}_4) \subset \mathbf{R}[\bar{x}]$ , and the maximal ideal  $\bar{P}\mathbf{R}[\bar{x}]_{\bar{P}}$  is an invariant ideal of  $\bar{\Sigma}$ . For  $\bar{v} \in \bar{P}\mathbf{R}[\bar{x}]_{\bar{P}}^4$  and  $\bar{r} \in \bar{P}\mathbf{R}[\bar{x}]_{\bar{P}}$ , all the conditions in Theorems 4.1, 4.2, and 4.3 hold, and  $\beta$  is an invariant immersion of  $\Sigma$  into an RSR  $\bar{\Sigma}'(\bar{g}_0 + \bar{v}, \bar{h} + \bar{r})$ .

If  $\bar{v} = [0, 0, 0, \bar{v}_4]^T$  with  $\bar{v}_4 \in \mathbf{R}[\bar{x}]_{\bar{P}}$  and  $\bar{r} \in PR[\bar{x}]_P \cap \mathbf{R}(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ , the conditions in Theorem 4.2 hold, while the conditions in Theorems 4.1 and 4.3 do not hold. In

fact,  $\bar{v} \notin \bar{P}\mathbf{R}[\bar{x}]_{\bar{P}}^4$  implies that  $\beta$  is not an invariant immersion but an immersion into  $\bar{\Sigma}'(\bar{g}_0 + \bar{v}, \bar{h} + \bar{r})$ . Moreover,  $\bar{P}\mathbf{R}[\bar{x}]_{\bar{P}}$  is not invariant with respect to the Lie derivative along  $\bar{v}$ , which implies that Theorem 4.3 is not applicable. In summary, Theorem 4.2 allows an immersion that is not an invariant immersion, and there is a gap between the necessary and sufficient conditions in Theorem 4.2 and the sufficient conditions in Theorem 4.3.

**5. Conclusion.** In this paper, some algebraic structures and their geometric counterparts in nonlinear systems over rings obtained by immersion have been discussed. In a system over a ring, vector fields are viewed as elements of a free module rather than as sections of a tangent bundle, and an important algebraic structure is the invariance of an ideal under the Lie derivative, which corresponds to the invariance of an affine variety. In particular, it has been shown that the relation ideal of an invariant immersion is an invariant ideal of a system over a ring obtained by that immersion, and all systems over a ring obtained via the same invariant immersion can be parameterized with the relation ideal. In the case of general immersions that are not necessarily invariant immersions, it has also been shown in this paper that the degrees of freedom in systems over a ring correspond to a generalization of the output invariance. Aside from immersion-related problems, the notions of nonlinear systems over rings, an invariant ideal, and an invariant variety can be used as fundamental tools for the analysis and control of rational or polynomial systems.

## REFERENCES

- [1] J. BACK AND J. H. SEO, *Immersion of non-linear systems into linear systems up to output injection: Characteristic equation approach*, Internat. J. Control, 77 (2004), pp. 723–734.
- [2] G. BESANÇON AND A. ȚICLEA, *Immersion-based observer design for nonlinear systems*, in Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, CA, 2006, pp. 4615–4620.
- [3] L. CAIRÓ, *Darboux integrability for 3D Lotka-Volterra systems*, J. Phys. A Mathematical, 33 (2000), pp. 2395–2406.
- [4] Z. CHEN AND J. HUANG, *Global robust stabilization of cascaded polynomial systems*, Systems Control Lett., 47 (2002), pp. 445–453.
- [5] D. CLAUDE, M. FLIESS, AND A. ISIDORI, *Immersion, directe et par bouclage, d'un système non linéaire dans un linéaire*, C. R. Acad. Sci. Paris Sér. I Math., 296 (1983), pp. 237–240.
- [6] G. CONTE AND A. M. PERDON, *Systems over rings: Geometric theory and applications*, Annual Review in Control, 24 (2000), pp. 113–124.
- [7] D. COX, J. LITTLE, AND D. O'SHEA, *Ideals, Varieties, and Algorithms*, 2nd ed., Springer-Verlag, Tokyo, 2000 (in Japanese); Springer-Verlag, New York, 1997 (in English).
- [8] M. FLIESS AND S. T. GLAD, *An algebraic approach to linear and nonlinear control*, in Essays on Control: Perspectives in the Theory and Its Applications, H. L. Trentelman and J. C. Willems, eds., Birkhäuser, Boston, 1993, pp. 223–267.
- [9] M. FLIESS AND I. KUPKA, *A finiteness criterion for nonlinear input-output differential systems*, SIAM J. Control Optim., 21 (1983), pp. 721–728.
- [10] M. FLIESS, J. LÉVINE, P. MARTIN, AND P. ROUCHON, *Flatness and defect of non-linear systems: Introductory theory and examples*, Internat. J. Control, 61 (1995), pp. 1327–1361.
- [11] A. ISIDORI, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, New York, 1995.
- [12] P. JOUAN, *Immersion of nonlinear systems into linear systems modulo output injection*, SIAM J. Control Optim., 41 (2003), pp. 1756–1778.
- [13] E. W. KAMEN, *Linear systems over rings: From R. E. Kalman to the present*, in Mathematical System Theory: The Influence of R. E. Kalman, A. C. Antoulas, ed., Springer-Verlag, New York, 1991, pp. 311–324.
- [14] A. LEVIN, *An analytical method of estimating the domain of attraction for polynomial differential equations*, IEEE Trans. Automat. Control, 39 (1994), pp. 2471–2475.
- [15] J. LEVINE AND R. MARINO, *Nonlinear system immersion, observers and finite-dimensional filters*, Systems Control Lett., 7 (1986), pp. 133–142.

- [16] J. T.-H. LO, *Global bilinearization of systems with control appearing linearly*, SIAM J. Control, 13 (1975), pp. 879–885.
- [17] H. MATSUMURA, *Commutative Ring Theory*, Kyoritsu Publishing, Tokyo, Japan, 1980 (in Japanese).
- [18] S. MONACO AND D. NORMAND-CYROT, *On the immersion of a discrete-time polynomial analytic system into a polynomial affine one*, Systems Control Lett., 3 (1983), pp. 83–90.
- [19] M. NAGATA, *Commutative Field Theory*, Shokabo, Tokyo, Japan, 1985 (in Japanese).
- [20] D. NEŠIĆ, *A note on observability tests for general polynomial and simple Wiener-Hammerstein systems*, Systems Control Lett., 35 (1998), pp. 219–227.
- [21] H. NIJMEIJER AND A. J. VAN DER SCHAFT, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
- [22] T. OHTSUKA, *Model structure simplification of nonlinear systems via immersion*, IEEE Trans. Automat. Control, 50 (2005), pp. 607–618.
- [23] P. A. PARRILO, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 2000.
- [24] P. S. PEREIRA DA SILVA AND E. DELALEAU, *Algebraic necessary and sufficient conditions of input-output linearization*, Forum Math., 13 (2001), pp. 335–357.
- [25] D. SCHLOMIUK, *Algebraic particular integrals, integrability and the problem of the center*, Trans. Amer. Math. Soc., 338 (1993), pp. 799–841.
- [26] A. TESI, F. VILLORESI, AND R. GENESIO, *On the stability domain estimation via a quadratic Lyapunov function: Convexity and optimality properties for polynomial systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 1650–1657.
- [27] B. TIBKEN, *Observability of nonlinear systems—an algebraic approach*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas, 2004, pp. 4824–4825.

## CONVEX DUALITY AND ENTROPY-BASED MOMENT CLOSURES: CHARACTERIZING DEGENERATE DENSITIES\*

CORY D. HAUCK<sup>†</sup>, C. DAVID LEVERMORE<sup>‡</sup>, AND ANDRÉ L. TITS<sup>§</sup>

**Abstract.** A common method for constructing a function from a finite set of moments is to solve a constrained minimization problem. The idea is to find, among all functions with the given moments, that function which minimizes a physically motivated, strictly convex functional. In the kinetic theory of gases, this functional is the kinetic entropy; the given moments are macroscopic densities; and the solution to the constrained minimization problem is used to formally derive a closed system of partial differential equations which describe how the macroscopic densities evolve in time. Moment equations are useful because they simplify the kinetic, phase-space description of a gas, and with entropy-based closures, they retain many of the fundamental properties of kinetic transport. Unfortunately, in many situations, macroscopic densities can take on values for which the constrained minimization problem does not have a solution. Essentially, this is because the moments are not continuous functionals with respect to the  $L^1$  topology. In this paper, we give a geometric description of these so-called *degenerate densities* in the most general possible setting. Our key tool is the complementary slackness condition that is derived from a dual formulation of a minimization problem with relaxed constraints. We show that the set of degenerate densities is a union of convex cones and, under reasonable assumptions, that this set is small in both a topological and a measure-theoretic sense. This result is important for further assessment and implementation of entropy-based moment closures.

**Key words.** convex duality, convex optimization, optimization in function spaces, kinetic theory, entropy-based closures, moment equations, gas dynamics

**AMS subject classifications.** 49N15, 90C25, 82C40, 35A35, 94A17, 76N15, 14P10

**DOI.** 10.1137/070691139

**1. Introduction.** In gas dynamics, the kinetic description of a gas is often simplified by using moment equations. In this reduced setting, a gas is characterized by a finite-dimensional vector  $\rho$  of densities that are moments of the kinetic distribution function  $F$  with respect to polynomials of the microscopic velocity. Evolution equations for  $\rho$  are derived by taking moments of the Boltzmann equation which governs the evolution of  $F$ . The derivation requires that an approximation for  $F$  be reconstructed from the densities  $\rho$ , giving what is called a *closure*.

One well-known method for prescribing a closure is to find a function that minimizes the kinetic entropy subject to the constraint that its moments agree with  $\rho$ . Such closures are called *entropy-based closures*. In recent years, they have generated substantial interest due to important structural properties which they inherit from the Boltzmann equation. These properties were first brought to light in [24].

---

\*Received by the editors May 10, 2007; accepted for publication (in revised form) December 3, 2007; published electronically July 2, 2008.

<http://www.siam.org/journals/sicon/47-4/69113.html>

<sup>†</sup>Computational Physics and Methods (CCS-2) and Center for Nonlinear Studies (CNLS), Los Alamos National Laboratory, Los Alamos, NM 87545 (cdhauck@lanl.gov). The work of this author was funded in part by the Department of Mathematics at the University of Maryland, College Park under National Science Foundation VIGRE grant DMS-0240049 and in part by the U.S. Department of Energy at Los Alamos Laboratory under contract DE-AC52-06NA25396 and the DOE Office of Science Advanced Computing Research (ASCR) Program.

<sup>‡</sup>Department of Mathematics and Institute for Physical Sciences and Technology, University of Maryland, College Park, MD 20742 (lvrnr@math.umd.edu).

<sup>§</sup>Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, College Park, MD 20742 (andre@umd.edu). The work of this author was supported in part by the National Science Foundation under Grant DMI-0422931 and by the U.S. Department of Energy under Grant DEFG0204ER25655.

In cases where the moments are continuous with respect to the relevant topology, there is always an entropy minimizer [6, 21]. Unfortunately, in classical gas dynamics, this is not usually the case. As a result, there are often physically relevant densities for which the constrained entropy minimization problem does not have a solution. In such cases, entropy-based closures are not well-defined, and these densities are called *degenerate*. In this paper, we provide a geometric description for the set of degenerate densities in the most general possible setting. We believe that this description is an important step in assessing the practical usefulness of entropy-based closures.

**1.1. Moment systems and entropy-based closures.** Consider a gas that is enclosed in a container, represented mathematically by the set  $\Omega \subset \mathbb{R}^d$  (typically  $d = 3$ ). The kinetic distribution function  $F = F(v, x, t)$  which describes the kinetic state of the gas is a nonnegative function that is defined for positions  $x \in \Omega$ , velocities  $v \in \mathbb{R}^d$ , and times  $t \geq 0$  so that, for any measurable set  $\Lambda \subset \Omega \times \mathbb{R}^d$ ,

$$(1) \quad \int_{\Lambda} F(v, x, t) \, dv dx$$

gives the number of particles at time  $t$  with positions  $x$  and velocities  $v$  such that  $(v, x) \in \Lambda$ . The evolution of  $F$  is governed by the Boltzmann transport equation

$$(2) \quad \partial_t F + v \cdot \nabla_x F = \mathcal{C}(F),$$

where  $\mathcal{C}$  is an integral operator that describes the collisions between particles which drive the system to local thermal equilibrium.

Solutions of (2) formally satisfy the local balance law [9]

$$(3) \quad \partial_t \mathcal{H}(F) + \nabla_x \cdot \mathcal{J}(F) = \mathcal{S}(F),$$

where the functionals

$$(4) \quad \mathcal{H}(g) \equiv \langle g \log(g) - g \rangle \quad \text{and} \quad \mathcal{J}(g) \equiv \langle v(g \log(g) - g) \rangle$$

are the *kinetic entropy* and *kinetic entropy flux*, respectively, and

$$(5) \quad \mathcal{S}(g) \equiv \langle \log(g) \mathcal{C}(g) \rangle$$

is the *kinetic entropy dissipation*. Here and throughout this paper,  $\langle \cdot \rangle$  denotes Lebesgue integration over all  $v \in \mathbb{R}^d$ , and we assume for the moment that the integrals in (4) and (5) are well-defined. According to Boltzmann's "H-theorem" [9],

$$(6) \quad \mathcal{S}(g) \leq 0,$$

with equality if and only if  $\mathcal{C}(g) = 0$ . In such cases,  $g$  is said to be in a state of local thermal equilibrium, and it takes the form of a Maxwellian distribution

$$(7) \quad \mathcal{M}_{\rho, u, \theta}(v) \equiv \frac{\rho}{(2\pi\theta)^{d/2}} \exp\left(-\frac{|v - u|^2}{2\theta}\right),$$

where  $\rho$  and  $\theta$  are positive scalars and  $u \in \mathbb{R}^d$ . In this way,  $\mathcal{H}$  acts as a Lyapunov functional for (2).

In order to reduce computational cost, the kinetic description of a gas provided by  $F$  is often simplified by retaining only a finite number of its velocity averages,

or *moments*. Equations which govern the evolution of these moments are derived by integrating (2) with respect to a vector

$$(8) \quad \mathbf{m} = (m_0, \dots, m_{n-1})^T$$

whose components are polynomials in  $v$ . Since  $v$  commutes with the spatial gradient, these equations take the form

$$(9) \quad \partial_t \boldsymbol{\rho} + \nabla_x \cdot \langle v \mathbf{m} F \rangle = \langle \mathbf{m} \mathcal{C}(F) \rangle,$$

where the moments

$$(10) \quad \boldsymbol{\rho} = \boldsymbol{\rho}(x, t) \equiv \langle \mathbf{m} F \rangle$$

are the *spatial densities* associated with  $F$ . Here again, we assume that the integrals in (9) and (10) are well-defined.

In general, (9) is not a closed system because there is no way to express the flux terms  $\langle v \mathbf{m} F \rangle$  and collision terms  $\langle \mathbf{m} \mathcal{C}(F) \rangle$  in terms of  $\boldsymbol{\rho}$ . Furthermore, in a moment description, an exact expression for  $F$  is not available. An alternative is to approximate  $F$  by an ansatz of the form

$$(11) \quad \mathcal{F}[\boldsymbol{\rho}] = \mathcal{F}(v, \boldsymbol{\rho}(x, t)).$$

By substituting  $\mathcal{F}$  for  $F$  in (9), the evolution of  $\boldsymbol{\rho}$  can be approximated by the closed system of balance laws

$$(12) \quad \partial_t \boldsymbol{\rho} + \nabla_x \cdot \mathbf{f}(\boldsymbol{\rho}) = \mathbf{c}(\boldsymbol{\rho}),$$

where the flux term  $\mathbf{f}$  and collision term  $\mathbf{c}$  are given by

$$(13) \quad \mathbf{f}(\boldsymbol{\rho}) = \langle v \mathbf{m} \mathcal{F}[\boldsymbol{\rho}] \rangle \quad \text{and} \quad \mathbf{c}(\boldsymbol{\rho}) = \langle \mathbf{m} \mathcal{C}(\mathcal{F}[\boldsymbol{\rho}]) \rangle.$$

One way to specify  $\mathcal{F}$  is to invoke the principle of entropy minimization (or maximization in the physics community, where the term “entropy” refers to  $-\mathcal{H}$  and has been widely used for over a century). The probabilistic interpretation of entropy dates back to Boltzmann [4, 5], who argued that the entropy of a system of identical particles depends on the number of microstates (particle arrangements in phase space) that are consistent with the macroscopic state of the system. This dependence is expressed by the famous logarithmic relationship known as *Boltzmann’s entropy formula* [8] (and also as Boltzmann’s equation, although distinct from (2)) and was first presented in its popular form by Planck [28, 29]. The practical application of entropy as a tool for statistical inference was championed by Jaynes although, in [19], Jaynes himself attributes the original mathematical concepts to Gibbs, who generalized Boltzmann’s entropy formula [16]. Jaynes also credits Shannon [32] for illuminating the central role that entropy plays in the theory of information. The relationship between statistics and information theory was further pursued by Kullback [23]. Many of the first rigorous results concerning entropy minimization can be found in [10] and references therein.

Closures which are based on the entropy minimization principle use the ansatz

$$(14) \quad \mathcal{F}[\boldsymbol{\rho}] = \arg \min_{g \in \mathbb{F}_{\mathbf{m}}} \{ \mathcal{H}(g) : \langle \mathbf{m} g \rangle = \boldsymbol{\rho} \}$$



at each  $x$  and  $t$  to formally close (9). Here

$$(15) \quad \mathbb{F}_{\mathbf{m}} \equiv \{g \in L^1(\mathbb{R}^d) : g \geq 0 \text{ and } |\mathbf{m}g| \in L^1(\mathbb{R}^d)\},$$

and  $|\cdot|$  is the standard Euclidean norm.

It is readily checked that  $\mathcal{H}$  is strictly convex over  $\mathbb{F}_{\mathbf{m}}$ . Thus if the minimizer in (14) exists, it is unique and the closure is well-defined. In such cases, (12) is a hyperbolic system of PDEs whose solutions satisfy the local dissipation law

$$(16) \quad \partial_t h(\boldsymbol{\rho}) + \nabla_x \cdot j(\boldsymbol{\rho}) = s(\boldsymbol{\rho}),$$

where

$$(17) \quad h(\boldsymbol{\rho}) \equiv \mathcal{H}(\mathcal{F}[\boldsymbol{\rho}])$$

is a strictly convex function of  $\boldsymbol{\rho}$  and

$$(18) \quad j(\boldsymbol{\rho}) \equiv \mathcal{J}(\mathcal{F}[\boldsymbol{\rho}]), \quad s(\boldsymbol{\rho}) \equiv \mathcal{S}(\mathcal{F}[\boldsymbol{\rho}]) \leq 0.$$

Although any choice for the ansatz  $\mathcal{F}[\boldsymbol{\rho}]$  will yield a system of the form (12), the entropy ansatz ensures that  $s(\boldsymbol{\rho}) \leq 0$  and that  $h$  is strictly convex. These conditions are important for two reasons. First, the dissipation law for a strictly convex function of  $\boldsymbol{\rho}$ , as given by (16), implies the existence of a well-posed linear  $L^2$  (Hilbert space) theory for (12) [33]. Second,  $h$  acts as a Lyapunov function for (12). To see this, note that (16) is simply (3) evaluated at  $F = \mathcal{F}[\boldsymbol{\rho}]$ ; and like in Boltzmann's H-theorem,  $s(\boldsymbol{\rho})$  vanishes if and only if  $\mathcal{C}(\mathcal{F}[\boldsymbol{\rho}]) = 0$ , in which case  $\mathcal{F}[\boldsymbol{\rho}]$  takes the form of a Maxwellian distribution [24].

The entropy minimization procedure yields an entire hierarchy of systems with the aforementioned properties whose members are generated by appending an initial choice of  $\mathbf{m}$  with additional polynomial components. For this reason, entropy-based closures have been applied to other areas of kinetic theory such as radiation transport [12, 13] and charge transport in semiconductors [1, 11, 22]. (Additional references for charge transport can be found in [1].) In the case of gas dynamics, the moment hierarchy begins with the canonical choice  $\mathbf{m} = (1, v, \frac{1}{2}|v|^2)^T$ . For this choice,  $\mathcal{F}[\boldsymbol{\rho}]$  is always a Maxwellian, and the entropy-based closure generates Euler's equations for a compressible gas.

**1.2. Realizability and degenerate densities.** A density  $\boldsymbol{\rho}$  is said to be *realized* by a function  $g \in \mathbb{F}_{\mathbf{m}}$  if  $\boldsymbol{\rho} = \langle \mathbf{m}g \rangle$ . The set of all such *realizable* densities will be denoted by  $\mathcal{R}_{\mathbf{m}}$ . An entropy-based closure is applicable only to those realizable densities for which the minimization problem (14) with equality constraints has a solution. If the moments in (14) were continuous with respect to the  $L^1$  topology, then there would always be a minimizer. Indeed, for such cases, Borwein and Lewis have shown in [6] that a constrained minimizer exists for a large class of convex functionals that include the classical entropy  $\mathcal{H}$ . However, in gas dynamics, the moments are typically *not continuous* in the  $L^1$  topology. As a result, there are realizable densities for which the minimizer in (14) does not exist. For such densities, which we term *degenerate*, modifications must be made to the entropy-based procedure. There are essentially two approaches:

1. Show that the set of nondegenerate densities is invariant under the dynamics of the balance law (12) with the entropy-based closure (as discussed in [20]) or impose such a condition in a way that is physically reasonable and mathematically justifiable.

2. Develop a modified closure that (i) is well-posed for *all* physically realizable values of  $\rho$ , (ii) recovers the minimum entropy-based closures whenever the minimizer in (14) exists, and (iii) generates systems of hyperbolic PDEs that dissipate a physically meaningful, convex entropy. This is the approach taken in [31].

We define  $\mathcal{D}_{\mathbf{m}}$  to be the set of all degenerate densities. In general, the set  $\mathcal{D}_{\mathbf{m}}$  depends on  $\mathbf{m}$ , and understanding its geometry is critical to determining whether entropy-based closures can be used in practice. In either of the modified approaches listed above, it is important—at the very least—to show that  $\mathcal{D}_{\mathbf{m}}$  is small in some sense, thereby minimizing the number of physically realizable spatial densities which require special treatment. In the first approach, this means limiting the number of initial conditions which must be discarded; in the second, it means limiting the number of physically realizable densities which require a modified closure.

Another reason to study  $\mathcal{D}_{\mathbf{m}}$  is that the equilibrium densities, i.e., those densities which are moments of a Maxwellian distribution (7), lie on its boundary [20, 21, 24, 31]. Because the kinetic entropy drives solutions of (3) toward local thermal equilibrium, we expect that trajectories defined by solutions to (16) will, at times, come very close to  $\mathcal{D}_{\mathbf{m}}$ . Thus it is very important to have a detailed understanding of its geometry.

Previous studies of the set  $\mathcal{D}_{\mathbf{m}}$  can be found in [20, 21, 31]. In [20], Junk provides a geometric description for  $\mathcal{D}_{\mathbf{m}}$  in a one-dimensional setting ( $d = 1$ ) with  $\mathbf{m} = (1, v, v^2, v^3, v^4)^T$ . It turns out in this case that  $\mathcal{D}_{\mathbf{m}}$  is a codimension one manifold. This result was discovered, in part, by extending the definition of  $h$  given by (18) to include cases where the minimizer in (14) does not exist. This is done simply by replacing the minimum in (14) with an infimum, viz.,

$$(19) \quad h_J(\rho) \equiv \inf_{g \in \mathbb{F}_{\mathbf{m}}} \{\mathcal{H}(g) : \langle \mathbf{m}g \rangle = \rho\}.$$

Later, in [21], Junk considers a more general case in which  $\mathbf{m}$  consists of a radial component  $|v|^N$ , for some even integer  $N \geq 2$ , plus polynomial components of lower degree. For such cases, he provides an integrability condition to determine whether  $\mathcal{D}_{\mathbf{m}}$  is nonempty. In practice, this condition is easily checked and extensible to more general choices of  $\mathbf{m}$ . However, a description of the geometry of  $\mathcal{D}_{\mathbf{m}}$ , as given in [20], is still lacking for the general setting.

In [31], Schneider introduces a different extension for  $h$  by relaxing the constraints in (14):

$$(20) \quad h_S(\rho) \equiv \min_{g \in \mathbb{F}_{\mathbf{m}}} \{\mathcal{H}(g) : \langle \mathbf{m}g \rangle \preceq^{\circ} \rho\}.$$

Here the notation  $\langle \mathbf{m}g \rangle \preceq^{\circ} \rho$  means—roughly speaking—that inequalities between certain components are allowed. (See section 3.2 for a precise definition.) The key difference between (14) and (20) is that the constraint set of the latter is closed in the weak- $L^1(\mathbb{R}^d)$  topology, whereas the constraint set of the former is not. Schneider uses this fact to prove that the minimizer in (20) with relaxed constraints always exists and is equal to the minimizer with equality constraints (14) when that minimizer exists (see our Theorem 3 and Corollary 4 below). In doing so, he provides a necessary and sufficient condition to determine whether a given density  $\rho$  is an element of  $\mathcal{D}_{\mathbf{m}}$ . However, this condition gives little insight into the geometry of  $\mathcal{D}_{\mathbf{m}}$ .

The main contribution of the present paper is a geometrical description of the set  $\mathcal{D}_{\mathbf{m}}$  in the most general possible setting. Our results are based on a dual formulation of (20) and are summarized in the following theorems.

- In Theorems 14 and 16, we prove strong duality for both the equality constraint problem (19) and the relaxed constraint problem (20). One consequence of these theorems is that  $h_S = h_J$ , even when the infimum in (20) is not attained. In Theorem 14, we also prove *complementary slackness conditions* which relate the density  $\rho$  in (20) to the dual variable and serve as the basis of our geometrical description.
- In Theorem 25, we show that the set  $\mathcal{D}_{\mathbf{m}}$  is a union of convex cones. The vertices of these cones are *nondegenerate* densities that lie on the boundary between the degenerate and nondegenerate densities in  $\mathcal{R}_{\mathbf{m}}$ . This conical description is based on the complementary slackness condition from Theorem 14.
- In Theorem 28, we show that, under reasonable assumptions, the set  $\mathcal{D}_{\mathbf{m}}$  is a nowhere dense subset of  $\mathcal{R}_{\mathbf{m}}$  that has Lebesgue measure zero and is restricted to the boundary of the nondegenerate, realizable densities. The assumptions we employ hold in all known cases. Whether they hold in general is an interesting and (to our knowledge) open question in analysis and algebraic geometry.

In the process of investigating  $\mathcal{D}_{\mathbf{m}}$ , we also recover and extend many previous results from both [20, 21] and [31].

The organization of the paper is as follows. In section 2 we introduce some notation and background information. In section 3 we review the entropy minimization problem. In section 4 we give a dual formulation of the minimization problem with relaxed constraints (20) and prove duality theorems for both (19) and (20). We use these theorems to show that  $h_S = h_J$  (even when the infimum in (19) is not attained) and to establish a complementary slackness condition. In section 5 we review the formal structure of entropy-based closures for nondegenerate densities and determine how that structure differs for degenerate cases. In section 6 we use the complementary slackness condition to describe the geometry of  $\mathcal{D}_{\mathbf{m}}$ . We then introduce the assumptions that allow us to make further assertions about the “smallness” of  $\mathcal{D}_{\mathbf{m}}$ . At the end of the section, we present two examples. In section 7 we give conclusions and discuss future work. Finally, in the appendix we provide a diagram and tables to assist the reader with notation.

**2. Preliminaries.** In this section, we introduce notation and present preliminary results. We refer the reader to the appendix for help in recalling the notation and useful properties for sets and mappings given throughout the paper.

**2.1. Admissible spaces.** For a given moment system, the choice of  $\mathbf{m}$  must satisfy criteria based on physical considerations. We require that components of  $\mathbf{m}$  form a basis for an  $n$ -dimensional linear space  $\mathbb{M}$  of multivariate polynomials over the field of real numbers that satisfies the following conditions:

- I.  $\mathbb{M} \supset \text{span}\{1, v_1, \dots, v_d, |v|^2\}$ ;
- (21) II.  $\mathbb{M}$  is invariant under translation and rotation;
- III. the set  $\mathbb{M}_c \equiv \{p \in \mathbb{M} : \langle |p| \exp(p) \rangle < \infty\}$  has a nonempty interior.

Our definition of  $\mathbb{M}_c$  is slightly different than the original definition given in [24]. However, its interior is the same under both definitions.

Spaces that satisfy conditions I–III are called *admissible*. According to condition I, any set of moment equations will incorporate the conservation laws for mass, momentum, and energy which are given by the moments of the kinetic distribution

function with respect to  $1$ ,  $v$ , and  $\frac{1}{2}|v|^2$ , respectively. In condition II, invariance under translation and rotation means that, for every  $u \in \mathbb{R}^d$  and every orthogonal matrix  $O$ , the mappings  $v \mapsto v - u$  and  $v \mapsto O^T v$  map  $\mathbb{M}$  onto itself. These properties will ensure that the moment equations are Galilean invariant—that is, invariant under the transformations  $x \mapsto x - ut$  and  $x \mapsto O^T x$ . Condition III applies specifically to entropy-based closures. It turns out that the minimizer (14), if it exists, has the form  $e^p$ , where  $p \in \mathbb{M}_c$ . Hence a nonempty interior for  $\mathbb{M}_c$  is a necessary requirement for any practical applications.

Typically an admissible space  $\mathbb{M}$  is generated by the span of polynomial functions whose moments are physical quantities of specific interest. (Here the canonical examples are the polynomials  $1$ ,  $v$ , and  $\frac{1}{2}|v|^2$ .) It may be that additional polynomial components are added to  $\mathbf{m}$  to ensure that  $\mathbb{M}$  is admissible. It should be noted that the vector  $\mathbf{m}$  that generates a given  $\mathbb{M}$  is not unique.

For convenience, we will assume, without loss of generality, that the components of  $\mathbf{m}$  are homogeneous. We decompose  $\mathbf{m}$  into subvectors:

$$(22) \quad \mathbf{m} = (\mathbf{m}_0^T, \mathbf{m}_1^T, \mathbf{m}_2^T, \dots, \mathbf{m}_N^T)^T,$$

where the  $n_j$  components of  $\mathbf{m}_j$  are the  $j$ th degree polynomial components of  $\mathbf{m}$ . Consistency requires that  $\sum_{j=0}^N n_j = n$ . Any polynomial  $p \in \mathbb{M}$  can be expressed as the sum of its homogeneous components:

$$(23) \quad p = \boldsymbol{\alpha}^T \mathbf{m} = \sum_{j=1}^N \boldsymbol{\alpha}_j^T \mathbf{m}_j,$$

where  $\boldsymbol{\alpha} \in \mathbb{R}^n$  is a vector of constant coefficients that decomposes into subvectors

$$(24) \quad \boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \dots, \boldsymbol{\alpha}_N^T)^T.$$

We briefly outline how one can generate a space  $\mathbb{M}$ . Given the even integer  $N \geq 2$  and  $j < N$ , let  $\mathbb{Q}_j$  be the space of all homogeneous polynomials from  $\mathbb{R}^d$  to  $\mathbb{R}$  of degree  $j$ . For each  $j$ ,  $\mathbb{Q}_j$  can be composed into rotationally invariant subspaces in the following way [15, Corollary 2.60]:

$$(25) \quad \mathbb{Q}_j = \begin{cases} \mathbb{H}_j \oplus |v|^2 \mathbb{H}_{j-2} \oplus |v|^4 \mathbb{H}_{j-4} \oplus \dots \oplus |v|^j \mathbb{H}_0, & j \text{ even,} \\ \mathbb{H}_j \oplus |v|^2 \mathbb{H}_{j-2} \oplus |v|^4 \mathbb{H}_{j-4} \oplus \dots \oplus |v|^{j-1} \mathbb{H}_1, & j \text{ odd.} \end{cases}$$

Here  $\mathbb{H}_k$  is the space of harmonic polynomials of degree  $k$  given by

$$(26) \quad \mathbb{H}_k = |v|^k \text{span} \left\{ Y^k \left( \frac{v}{|v|} \right) \right\},$$

and  $Y^k$  maps vectors on the unit sphere  $\mathbb{S}^{d-1}$  to the  $k$ -fold spherical harmonic tensor, which is unique modulo constant multiples. (Here the term “span” refers to all real linear combinations of the scalar components of the tensor.)

The decomposition in (25) is unique in the sense that no proper subset of the subspaces in (25) is rotationally invariant [15]. Thus, in order to be rotationally invariant, an admissible space  $\mathbb{M}$  must be a direct sum of some combination of the subspaces in (25) taken from each  $\mathbb{Q}_j$ ,  $j \leq N$ . In addition, the condition of translational invariance implies that choices for larger values of  $j$  will directly affect choices for smaller values

of  $j$ . For example, inclusion of the term  $|v|^j$  requires inclusion of the lower degree terms in the expansion of  $|v - u|^j$ .

To satisfy condition III,  $\mathbb{M}$  must include polynomials from  $\mathbb{Q}_N$  which dominate the behavior of odd degree polynomials of lower degree for large  $|v|$ . In particular,  $\mathbb{M}$  must include multiples of  $|v|^N$ . This is because spherical harmonics (both odd and even) other than  $Y^0 \equiv 1$  take on both positive and negative values on the unit sphere. Excluding  $|v|^N$  would therefore lead to polynomials  $p$ , all of which satisfy  $\lim_{r \rightarrow \infty} p(r\omega) = \infty$  for all  $\omega$  contained in some subset of  $\mathbb{S}^{d-1}$  with positive Lebesgue measure. In such cases  $\exp(p)$  is not integrable for any  $p \in \mathbb{M}$ , and condition III is violated.

In applications it is sometimes convenient to represent components of  $\mathbf{m}$  in tensor format. There are two reasons for this. The first reason is the convenience with which one can express  $\mathbf{m}$  given (25) and (26). The second reason is that the evolution of the moment of  $\langle TF \rangle$  for any  $j$ -fold tensor  $T = T(v)$  depends on the divergence of the  $(j + 1)$ -fold tensor  $\langle vTF \rangle$  (refer to (9)).

The tensors in which we are interested are often symmetric and sometimes traceless. For example, the Gaussian closure which will be described in section 5.3 is based on the vector

$$(27) \quad \mathbf{m} = \begin{pmatrix} \mathbf{m}_0 \\ \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ v \\ v \vee v \end{pmatrix} = \begin{pmatrix} 1 \\ v \\ (v \vee v - \frac{1}{3}|v|^2 I) + \frac{1}{3}|v|^2 I \end{pmatrix},$$

where  $v \vee v$  is the symmetric tensor product of  $v$  with itself.<sup>1</sup> In the strict vector representation,  $\mathbf{m}_2$  is composed only of the  $d(d+1)/2$  linearly independent components of the tensor  $v \vee v$ . The components have the form  $v_i v_j$ , where  $1 \leq i \leq d$  and  $i \leq j \leq d$ .

Vectors  $\alpha \in \mathbb{R}^n$  can also be represented by tensors, in which case the product in (23) is interpreted as a sum of tensor inner products.<sup>2</sup> For a given a polynomial  $p$ , the tensor form of  $\alpha$  in (23) is unique under the additional requirement that it has the same symmetry properties as  $\mathbf{m}$ .

**2.2. Cones.** Many of the sets that we will encounter in this paper are cones [3, 30]. A subset  $C$  of  $\mathbb{R}^k$  is a *cone* if, for all real numbers  $\lambda > 0$ ,  $y \in C$  if and only if  $\lambda y \in C$ . A cone is *solid* if it has a nonempty interior. A closed cone  $C$  is *pointed* if  $-C \cap C$  is the origin. A closed cone that is convex, pointed, and solid is called *proper*. For example, the set  $\mathbb{F}_{\mathbf{m}}$  is a solid, convex cone, whose closure in  $L^1(\mathbb{R}^d)$  is proper. Several other cones will be introduced in the subsections that follow, and eventually we will see that the set  $\mathcal{D}_{\mathbf{m}}$  is also a cone.

Associated with every cone  $C$  is its polar cone<sup>3</sup>

$$(28) \quad C^\circ \equiv \{z \in \mathbb{R}^k : z^T y \leq 0 \quad \forall y \in C\}.$$

It is readily checked that the polar of a proper cone is proper.

A vector  $z \in \mathbb{R}^k$  is *tangent* to a subset  $\Omega \subset \mathbb{R}^k$  at a point  $y \in \Omega$  if  $z = 0$  or if

$$(29) \quad \lim_{j \rightarrow \infty} \frac{y_j - y}{|y_j - y|} = \frac{z}{|z|}$$

<sup>1</sup>Given a symmetric  $j$ -fold tensor  $S$  and a symmetric  $k$ -fold tensor  $T$ , the symmetric tensor product of  $S$  and  $T$  is  $S \vee T = T \vee S \equiv \frac{1}{(j+k)!} \sum_{\pi \in \Pi} S_{i_{\pi(1)}, \dots, i_{\pi(j)}} T_{i_{\pi(j+1)}, \dots, i_{\pi(j+k)}}$ , where  $\Pi$  is the set of all permutation of the integers  $1, \dots, j+k$ .

<sup>2</sup>For  $k > j$ , the symmetric inner product (or contraction) of a symmetric  $j$ -fold tensor  $S$  and a symmetric  $k$ -fold tensor  $T$  is  $(S \cdot T)_{i_{j+1}, \dots, i_{j+k}} \equiv \sum_{i_1, \dots, i_j} S_{i_1, \dots, i_j} T_{i_1, \dots, i_j, i_{j+1}, \dots, i_{j+k}}$ .

<sup>3</sup>The polar cone is the negative of the dual cone  $C^* \equiv \{z \in \mathbb{R}^k : z^T y \geq 0 \quad \forall y \in C\}$ .

for some sequence  $\{y_j\}_{j=1}^\infty \subset \Omega$  such that  $y_j \rightarrow y$ , but  $y_j \neq y$  for all  $j$ . The *tangent cone of  $\Omega$  at  $y$* , which we denote  $\mathcal{TC}(\Omega, y)$ , is the set of all vectors that are tangent to  $\Omega$  at  $y$ . A vector  $w \in \mathbb{R}^k$  is *normal* to  $\Omega$  at  $y \in \Omega$  if there exist sequences  $\{y_j\}_{j=1}^\infty \subset \Omega$  and  $\{w_j\}_{j=1}^\infty \subset \mathbb{R}^k$  such that

$$(30) \quad y_j \rightarrow y, \quad w_j \rightarrow w, \quad w_j \in (\mathcal{TC}(\Omega, y_j))^\circ \quad \forall j.$$

The *normal cone of  $\Omega$  at  $y$* , which we denote  $\mathcal{NC}(\Omega, y)$ , is the set of all vectors that are normal to  $\Omega$  at  $y$ . For the important case that  $\Omega$  is convex,

$$(31) \quad \mathcal{NC}(\Omega, y) = \{z \in \mathbb{R}^k : z^T(y' - y) \leq 0 \quad \forall y' \in \Omega\}.$$

In particular,  $\mathcal{NC}(\Omega, y)$  is convex. If  $\partial\Omega$  is a  $C^1$  (continuously differentiable) manifold containing  $y$ , then  $\mathcal{NC}(\Omega, y)$  is a ray with base point at the origin that points in the outward normal direction to  $\partial\Omega$  at  $y$ . More generally, given any  $C^1$  manifold  $M \ni y$  of dimension  $j$ ,  $\mathcal{NC}(M, y)$  is a subspace of dimension  $n - j$ . If  $M \subset \Omega$ , then  $\mathcal{NC}(\Omega, y) \subset \mathcal{NC}(M, y)$ . In sections 5.4 and 6.4, we will use the notation  $\mathcal{NC}_0(\Omega, y)$  to denote the normal cone without the origin:

$$(32) \quad \mathcal{NC}_0(\Omega, y) \equiv \mathcal{NC}(\Omega, y) \setminus \{0\}.$$

A particularly useful application of cones is to provide a partial ordering of elements in  $\mathbb{R}^k$  (or, more generally, in any vector space). Given a pointed, convex cone  $C$  and  $y_1$  and  $y_2$  in  $\mathbb{R}^k$ , we say that  $y_1 \leq_C y_2$ , or  $y_2 \geq_C y_1$ , if and only if  $y_2 - y_1 \in C$ .

**2.3. Realizable densities.** Our motivation for solving (14), (19), or (20) is to find a closure for the moment equations (9). Thus we are interested only in constraints based on densities which are realizable, i.e., elements of the set

$$(33) \quad \mathcal{R}_{\mathbf{m}} \equiv \{\rho \in \mathbb{R}^n : \rho = \langle \mathbf{m}g \rangle, g \in \mathbb{F}_{\mathbf{m}}\}.$$

With this notation we formally define the set  $\mathcal{D}_{\mathbf{m}}$ :

$$(34) \quad \mathcal{D}_{\mathbf{m}} \equiv \{\rho \in \mathcal{R}_{\mathbf{m}} : \text{the minimizer in (14) does not exist}\}.$$

A density  $\rho \in \mathcal{R}_{\mathbf{m}}$  has a natural decomposition based on the decomposition of  $\mathbf{m}$  in (22):

$$(35) \quad \rho = (\rho_0^T, \rho_1^T, \rho_2^T, \dots, \rho_N^T)^T,$$

where  $\rho_j = \langle \mathbf{m}_j g \rangle$  for some  $g \in \mathbb{F}_{\mathbf{m}}$ . The set  $\mathcal{R}_{\mathbf{m}}$  has several important properties, one of which is its relation to the cone

$$(36) \quad A_{\mathbf{m}} \equiv \{\alpha \in \mathbb{R}^n : \alpha^T \mathbf{m} \leq 0\}.$$

It is straightforward to verify that  $A_{\mathbf{m}}$  is a proper cone.

**THEOREM 1** (Junk [21]). *The set  $\mathcal{R}_{\mathbf{m}}$  is an open, convex, solid cone, and its closure is proper. In fact,  $\mathcal{R}_{\mathbf{m}} = \text{int } A_{\mathbf{m}}^\circ$ , and every vector in  $\mathcal{R}_{\mathbf{m}}$  is realized by a bounded, nonnegative function with compact support.*

*Proof.* We refer the reader to Theorem A.2 of [21] for a proof (which applies to the case  $\mathbf{m}_N = |v|^N$  but can be modified to the general case with little effort). However, to provide the reader with some intuition, we show here that  $\mathcal{R}_{\mathbf{m}} \subset \text{int } A_{\mathbf{m}}^\circ$ . Let  $\rho \in \mathcal{R}_{\mathbf{m}}$ . Then  $\rho = \langle \mathbf{m}g \rangle$  for some  $g \in \mathbb{F}_{\mathbf{m}}$  and according to (36)

$$(37) \quad \alpha^T \rho = \langle \alpha^T \mathbf{m}g \rangle \leq 0$$

for all  $\alpha \in A_{\mathbf{m}}$ . Further, since  $\alpha^T \mathbf{m}$  is a polynomial, it can be zero only on a set of zero Lebesgue measure. Hence  $\alpha^T \rho < 0$ , which proves that  $\rho \in \text{int } A_{\mathbf{m}}^\circ$ .  $\square$

**2.4. Exponentially realizable densities.** We will see below that the minimizer of (20) has the form

$$(38) \quad G_{\alpha} \equiv \exp(\alpha^T \mathbf{m}),$$

where  $\alpha$  solves the dual problem to (20). The integral of  $G_{\alpha}$  is the *density potential*

$$(39) \quad h^*(\alpha) \equiv \langle G_{\alpha} \rangle,$$

which was introduced in [25] as a tool for elucidating the formal structure of entropy-based closures. As the notation suggests,  $h^*$  is the Legendre dual of  $h$ . In section 5, we will discuss this relationship in more detail. The name “density potential” is derived from the fact that its formal derivative  $\mathbf{r}$  generates the moments of  $G_{\alpha}$ . Given the set

$$(40) \quad \mathcal{A}_{\mathbf{m}} \equiv \{\alpha \in \mathbb{R}^n : G_{\alpha} \in \mathbb{F}_{\mathbf{m}}\},$$

$\mathbf{r} : \mathcal{A}_{\mathbf{m}} \rightarrow \mathbb{R}^n$  is defined by

$$(41) \quad \mathbf{r}(\alpha) \equiv \langle \mathbf{m} G_{\alpha} \rangle.$$

It should be noted in (40) that the condition  $\alpha \in \mathcal{A}_{\mathbf{m}}$  is stronger than  $G_{\alpha} \in L^1(\mathbb{R}^d)$ , since the latter can still yield moments that are infinite. The image of  $\mathcal{A}_{\mathbf{m}}$  under  $\mathbf{r}$  is the set of *exponentially realizable densities*:

$$(42) \quad \mathcal{R}_{\mathbf{m}}^{\text{exp}} \equiv \mathbf{r}(\mathcal{A}_{\mathbf{m}}).$$

The set  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$  is a solid cone. It need not be convex, nor is it necessarily open.

Since  $\mathbf{r}(\mathcal{A}_{\mathbf{m}}) = \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ , it is important to understand the structure of  $\mathcal{A}_{\mathbf{m}}$ . Its interior has a rather simple expression:

$$(43) \quad \text{int } \mathcal{A}_{\mathbf{m}} = \{\alpha \in \mathbb{R}^n : \alpha_N^T \mathbf{m}_N(v) < 0 \quad \forall v \neq 0\} = \{\alpha \in \mathbb{R}^n : \alpha_N \in \text{int } A_{\mathbf{m}_N}\},$$

where

$$(44) \quad A_{\mathbf{m}_j} \equiv \{\alpha_j \in \mathbb{R}^{n_j} : \alpha_j^T \mathbf{m}_j \leq 0\}, \quad 1 \leq j \leq N,$$

is a proper cone for  $j$  even. (It can be checked that condition III of section 2.1 is equivalent to  $\text{int } \mathcal{A}_{\mathbf{m}}$  being nonempty.) If  $\alpha \in \text{int } \mathcal{A}_{\mathbf{m}}$ , then the behavior of  $p = \alpha^T \mathbf{m}$  is dominated for large  $|v|$  by the homogeneous component  $p_N = \alpha_N^T \mathbf{m}_N$ , and

$$(45) \quad \lim_{|v| \rightarrow \infty} p(v) = \lim_{|v| \rightarrow \infty} p_N(v) = \lim_{|v| \rightarrow \infty} |v|^N p_N(v/|v|) = -\infty.$$

For such  $\alpha$ ,  $G_{\alpha}$  decays exponentially, and the moments  $\mathbf{r}(\alpha)$  are finite.

From (43), one can easily show that

$$(46) \quad \text{cl } \mathcal{A}_{\mathbf{m}} = \{\alpha \in \mathbb{R}^n : \alpha_N \in A_{\mathbf{m}_N}\} \quad \text{and} \quad \partial \mathcal{A}_{\mathbf{m}} \subset \{\alpha \in \mathbb{R}^n : \alpha_N \in \partial A_{\mathbf{m}_N}\}.$$

Even so, the boundary component  $\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$  is, in general, very complicated. If  $\alpha \in \partial \mathcal{A}_{\mathbf{m}}$ , then  $\alpha_N \in \partial A_{\mathbf{m}_N}$  and  $p_N(\lambda v) = 0$  for some  $v \neq 0$  and all  $\lambda \in \mathbb{R}$ , and it may be that there are unbounded sequences  $\{v_i\}_{i=1}^{\infty}$  such that  $\lim_{i \rightarrow \infty} p(v_i) > -\infty$ . In such cases, it is not clear whether the moments  $\mathbf{r}(\alpha)$  are finite, i.e., whether  $\alpha \in \mathcal{A}_{\mathbf{m}}$ . We will revisit this issue in section 6.3. For now, we turn our attention to the entropy minimization problem (20).

**3. Entropy minimization.** Most of this section reproduces and discusses the main result from [31]. In this setting, we then state Theorem 9, which is the basis for our new results.

**3.1. The entropy functional.** Recall that the strictly convex *entropy functional*  $\mathcal{H} : \mathbb{F}_{\mathbf{m}} \mapsto \mathbb{R} \cup \{\infty\}$  is given by

$$(47) \quad \mathcal{H}(g) \equiv \langle g \log g - g \rangle.$$

By employing the convention  $0 \log 0 = 0$ —which is consistent with the fact that  $\lim_{z \rightarrow 0} z \log z = 0$ —one can make sense of the integrand for those values of  $v$  where  $g(v) = 0$ . There are functions  $g \in \mathbb{F}_{\mathbf{m}}$  such that  $\mathcal{H}(g) = +\infty$ ; however, in order for  $\mathcal{H}(g)$  to be well-defined, the negative contribution to the integral  $\mathcal{H}^-(g)$  must be finite. We show that this is indeed the case.

LEMMA 2. *For each  $g \in \mathbb{F}_{\mathbf{m}}$ , let  $K_g = \{v \in \mathbb{R}^d : g(v) \log(g(v)) - g(v) < 0\}$ . Then*

$$(48) \quad \mathcal{H}^-(g) \equiv - \int_{K_g} (g(v) \log(g(v)) - g(v)) \, dv \leq \int_{\mathbb{R}^d} (|v|^2 g(v) + e^{-|v|^2}) \, dv.$$

*In particular,  $\mathcal{H}^-(g)$  is finite.*

The proof of this lemma is based on Young's inequality: For all  $z, y > 0$ ,

$$(49) \quad z \log z - z \geq y \log y - y + (\log y)(z - y)$$

or, equivalently,

$$(50) \quad z \log z - z \geq z \log y - y.$$

These two inequalities follow immediately from the convexity of the mapping  $z \mapsto z \log z - z$ .

*Proof.* Letting  $z = g(v)$  and  $y = e^{-|v|^2}$  in (50) gives, after integration over  $K_g$ ,

$$(51) \quad \mathcal{H}^-(g) \leq \int_{K_g} (|v|^2 g(v) + e^{-|v|^2}) \, dv \leq \int_{\mathbb{R}^d} (|v|^2 g(v) + e^{-|v|^2}) \, dv,$$

which is finite since  $|v|^2 \in \mathbb{M}$ .  $\square$

**3.2. Schneider's problem.** Given  $\boldsymbol{\rho} = (\boldsymbol{\rho}_0, \dots, \boldsymbol{\rho}_N) \in \mathcal{R}_{\mathbf{m}}$ , we seek a solution of (20), where the relation  $\langle \mathbf{m}g \rangle \preceq^\circ \boldsymbol{\rho}$  (or, equivalently,  $\boldsymbol{\rho} \succeq^\circ \langle \mathbf{m}g \rangle$ ) is shorthand for

$$(52a) \quad \langle \mathbf{m}_j g \rangle = \boldsymbol{\rho}_j, \quad 0 \leq j \leq N-1,$$

$$(52b) \quad \langle \mathbf{m}_N g \rangle \leq_{A_{\mathbf{m}_N}^\circ} \boldsymbol{\rho}_N,$$

and  $A_{\mathbf{m}_N}^\circ \equiv (A_{\mathbf{m}_N})^\circ$ . Note that (52b) means that

$$(53) \quad \boldsymbol{\alpha}_N^T \langle \mathbf{m}_N g \rangle \leq \boldsymbol{\alpha}_N^T \boldsymbol{\rho}_N \quad \text{whenever} \quad \boldsymbol{\alpha}_N^T \mathbf{m}_N \geq 0.$$

The components of  $\langle \mathbf{m}_j g \rangle$ ,  $0 \leq j < N$ , will be referred to as *lower-order moments*, and the components of  $\langle \mathbf{m}_N g \rangle$  will be referred to as *higher-order moments*.

The main result from [31] concerning the minimization problem with relaxed constraints (20) is the following theorem.



THEOREM 3 (Schneider [31]). *For any  $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ , there is a unique minimizer for the minimization problem (20). This minimizer has the form  $G_{\boldsymbol{\alpha}}$  given by (38), where  $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}}$ . Conversely, for each  $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}}$ ,*

$$(54) \quad \mathcal{H}(G_{\boldsymbol{\alpha}}) = \min_{g \in \mathbb{F}_{\mathbf{m}}} \{ \mathcal{H}(g) : \langle \mathbf{m}g \rangle \preceq^{\circ} \mathbf{r}(\boldsymbol{\alpha}) \},$$

where  $\mathbf{r}(\boldsymbol{\alpha})$  is given by (41). Moreover,  $G_{\boldsymbol{\alpha}}$  also satisfies the equality constraint problem (14) with  $\boldsymbol{\rho} = \mathbf{r}(\boldsymbol{\alpha})$ .

We define  $\mathbf{a} : \mathcal{R}_{\mathbf{m}} \rightarrow \mathcal{A}_{\mathbf{m}}$  as the mapping which assigns to  $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$  the vector  $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}}$  such that  $G_{\boldsymbol{\alpha}}$  solves (20)—that is,

$$(55) \quad G_{\mathbf{a}(\boldsymbol{\rho})} \equiv \arg \min_{g \in \mathbb{F}_{\mathbf{m}}} \{ \mathcal{H}(g) : \langle \mathbf{m}g \rangle \preceq^{\circ} \boldsymbol{\rho} \}.$$

The converse statement of Theorem 3 implies the following.

COROLLARY 4. *Let  $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ . Then  $G_{\mathbf{a}(\boldsymbol{\rho})}$  is the unique minimizer of the entropy minimization problem with equality constraints (14).*

To help the reader's intuition, we provide a proof for Theorem 3 with the use of three lemmas. The first lemma is used to prove the existence of a minimizer for the minimization problem with relaxed constraints (20), and the first item of this lemma is a direct consequence of Lemma 2.

LEMMA 5 (Schneider [31]). *The entropy functional  $\mathcal{H}$  satisfies the following:*

1.  $\mathcal{H}(g) > -\infty$  for all  $g \in \mathbb{F}_{\mathbf{m}}$ .
2.  $\mathcal{H}$  is convex and lower semicontinuous with respect to the norm  $\|g\|_{L_{\mathbf{m}}^1(\mathbb{R}^d)} \equiv \langle \mathbf{m}g \rangle$ .
3. Subsets of  $\mathbb{F}_{\mathbf{m}}$  which are bounded in the  $L_{\mathbf{m}}^1(\mathbb{R}^d)$  topology and on which  $\mathcal{H}$  is bounded are weakly relatively compact in  $L^1(\mathbb{R}^d)$ .

The second lemma is a statement about the constraint set

$$(56) \quad C_{\mathbf{m}}(\boldsymbol{\rho}) \equiv \{ g \in \mathbb{F}_{\mathbf{m}} : \langle \mathbf{m}g \rangle \preceq^{\circ} \boldsymbol{\rho} \}.$$

LEMMA 6. *For each  $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ , the set  $C_{\mathbf{m}}(\boldsymbol{\rho})$  is closed in the weak- $L^1$  topology.*

*Proof.* Let  $\{g_k\}_{k=1}^{\infty}$  be any sequence in  $C_{\mathbf{m}}(\boldsymbol{\rho})$  that converges in weak- $L^1(\mathbb{R}^d)$  to a function  $g_*$ . For the highest-order moments, Fatou's lemma implies that if  $\boldsymbol{\alpha}_N^T \mathbf{m}_N \geq 0$ , then

$$(57) \quad \boldsymbol{\alpha}_N^T \langle \mathbf{m}_N g_* \rangle \leq \lim_{k \rightarrow \infty} \boldsymbol{\alpha}_N^T \langle \mathbf{m}_N g_k \rangle = \boldsymbol{\alpha}_N^T \boldsymbol{\rho}_N.$$

For  $j < N$ , more can be said. We break up the integral  $\langle \mathbf{m}_j g_k \rangle$  into two pieces:

$$(58) \quad \boldsymbol{\rho}_j = \langle \mathbf{m}_j g_k \rangle = \int_{|v| < R} \mathbf{m}_j g_k \, dv + \int_{|v| > R} \mathbf{m}_j g_k \, dv,$$

where  $R > 0$  is an arbitrary constant. For the first term in (58), weak- $L^1(\mathbb{R}^d)$  convergence implies that

$$(59) \quad \int_{|v| < R} \mathbf{m}_j g_k \, dv \xrightarrow{k \rightarrow \infty} \int_{|v| < R} \mathbf{m}_j g_* \, dv, \quad 0 \leq j \leq N.$$

Meanwhile, in the second term

$$(60) \quad \frac{|\mathbf{m}_j|}{|v|^N} < \frac{C_0}{R^{N-j}}, \quad |v| > R, \quad 0 \leq j < N,$$

for some constant  $C_0$  that is independent of  $R$ . Hence

$$(61) \quad \int_{|v|>R} \mathbf{m}_j g_k dv \leq \int_{|v|>R} \frac{|\mathbf{m}_j|}{|v|^N} |v|^N g_k dv < \frac{C_0}{R^{N-j}} \sup_k |\langle |v|^N g_k \rangle|.$$

Since  $\{g_k\}_{k=1}^\infty \subset C_{\mathbf{m}}$ , the sequence  $\{\langle |v|^N g_k \rangle\}_{k=1}^\infty$  is uniformly bounded in  $k$ , and it follows from (59) and (61) that

$$(62) \quad \begin{aligned} \lim_{k \rightarrow \infty} |\rho_j - \langle \mathbf{m}_j g_k \rangle| &\leq \lim_{k \rightarrow \infty} \int_{|v|>R} |\mathbf{m}_j g_k - \mathbf{m}_j g_*| dv \\ &\leq \frac{C_0}{R^{N-j}} \left( \sup_k \langle |v|^N g_k \rangle + \langle |v|^N g_* \rangle \right) < \frac{C_1}{R^{N-j}} \end{aligned}$$

for some constant  $C_1 > 0$  that is independent of  $R$ . Since  $R$  can be arbitrarily large, we conclude that  $\langle \mathbf{m}_j g_* \rangle = \rho_j$  for all  $j < N$ . Hence  $g_* \in C_{\mathbf{m}}(\rho)$ .  $\square$

The third lemma is used to prove the form of the minimizer. For any bounded measurable set  $K \subset \mathbb{R}^d$  and any locally integrable function  $g$ , let

$$(63) \quad \begin{aligned} \langle g \rangle_K &\equiv \int_K g(v) dv \quad \text{and} \quad \mathbb{F}_{\mathbf{m}}^K \equiv \{g \in L_1(\mathbb{R}^d) : g \geq 0 \\ &\text{and } \langle m_i g \rangle_K < \infty, i = 0, \dots, n-1\}. \end{aligned}$$

On  $\mathbb{F}_{\mathbf{m}}^K$ , we define

$$(64) \quad \mathcal{H}^K(g) \equiv \langle g \log g - g \rangle_K.$$

As with  $\mathcal{H}$ , the negative contribution to  $\mathcal{H}^K$  must be finite (see Lemma 2) in order for it to be well-defined, and restricting  $\text{Dom}(\mathcal{H}^K)$  to  $\mathbb{F}_{\mathbf{m}}^K$  ensures that this will be the case.

LEMMA 7 (Junk [20, 21], Borwein and Lewis [6]). *For any bounded set  $K \subset \mathbb{R}^d$  and any function  $f \in \mathbb{F}_{\mathbf{m}}^K$ , the problem*

$$(65) \quad \min_{g \in \mathbb{F}_{\mathbf{m}}^K} \{ \mathcal{H}^K(g) : \langle \mathbf{m} g \rangle_K = \langle \mathbf{m} f \rangle_K \}$$

*has a unique minimizer, which takes the form  $G_{\alpha}$  for some  $\alpha \in \mathbb{R}^n$ .*

*Proof of Theorem 3.* The proof has three parts.

1. *Existence and uniqueness.* Let  $\rho \in \mathcal{R}_{\mathbf{m}}$ . By Theorem 1, the set

$$(66) \quad C_{\mathbf{m}}(\rho) \equiv \{g \in \mathbb{F}_{\mathbf{m}} : \langle \mathbf{m} g \rangle \preceq^{\circ} \rho\}$$

contains bounded functions with compact support. Because such functions have finite entropy, the subset of  $C_{\mathbf{m}}(\rho)$  on which  $\mathcal{H}$  is finite is nonempty. Moreover, by Lemma 2,  $\mathcal{H}$  is bounded below on  $C_{\mathbf{m}}(\rho)$ . Hence  $h_S(\rho)$  is finite, and there exists  $\{g_i\}_{i=1}^\infty \subset C_{\mathbf{m}}(\rho)$  such that  $\mathcal{H}(g_i) \rightarrow h_S(\rho)$ . By Lemma 5, there is a subsequence  $\{g_{i_k}\}_{k=1}^\infty$  that converges in weak- $L^1$  to a function  $\hat{g}_\rho$ , and since  $C_{\mathbf{m}}(\rho)$  is closed (Lemma 6),  $\hat{g}_\rho \in C_{\mathbf{m}}(\rho)$ . Finally, since  $\mathcal{H}$  is lower semicontinuous (Lemma 5),

$$(67) \quad \mathcal{H}(\hat{g}_\rho) \leq \lim_{k \rightarrow \infty} \mathcal{H}(g_{i_k}) = h_S(\rho).$$

Thus  $\hat{g}_\rho$  attains the minimum in (20), and strict convexity of  $\mathcal{H}$  implies that the minimizer is unique.

2. *Form of the minimizer.* According to Lemma 7, for any bounded set  $K \subset \mathbb{R}^d$ ,

$$(68) \quad \min \{ \mathcal{H}^K(g) : \langle \mathbf{m}g \rangle_K = \langle \mathbf{m}\hat{g}_\rho \rangle_K \}$$

has a solution of the form  $G_\alpha$ . We conclude then that  $\hat{g}_\rho = G_\alpha$  on  $K$ ; otherwise, the function

$$(69) \quad g_\rho^*(v) = \begin{cases} G_\alpha(v), & v \in K, \\ \hat{g}_\rho, & v \notin K, \end{cases}$$

would satisfy  $\mathcal{H}(g_\rho^*) \leq \mathcal{H}(\hat{g}_\rho)$ , an obvious contradiction. Since  $K$  is arbitrary, we conclude that  $\hat{g}_\rho = G_\alpha$  and, in order to satisfy to constraints in (20), that  $\alpha \in \mathcal{A}_\mathbf{m}$ .

3. *Converse statement.* Applying Young's inequality (49) to  $z = g$  and  $y = G_\alpha$  and integrating over all velocity space gives

$$(70) \quad \mathcal{H}(g) \geq \mathcal{H}(G_\alpha) + \alpha^T \langle \mathbf{m}(g - G_\alpha) \rangle.$$

By hypothesis,  $\alpha \in \mathcal{A}_\mathbf{m}$ , which implies that  $\alpha_N \in \mathcal{A}_{\mathbf{m}_N}$ . Thus if  $g \in \mathbb{F}_\mathbf{m}$  satisfies  $\langle \mathbf{m}g \rangle \preceq^\circ \langle \mathbf{m}G_\alpha \rangle$ , then according to (52) and (53),

$$(71) \quad \alpha^T \langle \mathbf{m}(g - G_\alpha) \rangle = \sum_{j=1}^N \alpha_j^T \langle \mathbf{m}_j(g - G_\alpha) \rangle = \alpha_N^T \langle \mathbf{m}_N(g - G_\alpha) \rangle \geq 0.$$

Thus, from (70),  $\mathcal{H}(g) \geq \mathcal{H}(G_\alpha)$ . This concludes the proof.  $\square$

The existence part of this proof provides some intuition as to why the optimization problem with equality constraints (14) may not always have a minimizer. Suppose that the minimizing sequence  $\{g_{i_k}\}_{k=1}^\infty$  were restricted to the set

$$(72) \quad C_\mathbf{m}^0(\rho) \equiv \{g \in \mathbb{F}_\mathbf{m} : \langle \mathbf{m}g \rangle = \rho\}$$

rather than merely lying in  $C_\mathbf{m}(\rho)$ . Then  $\{g_{i_k}\}_{k=1}^\infty$  would still converge in the weak- $L^1(\mathbb{R}^d)$  topology to  $\hat{g}_\rho$ , with  $\langle \mathbf{m}_j \hat{g}_\rho \rangle = \rho_j$  for  $j < N$ . However, the bound in (61) does not help when  $j = N$ . Hence there is no way to ensure that  $\langle \mathbf{m}_N \hat{g}_\rho \rangle = \rho_N$ —only that  $\langle \mathbf{m}_N \hat{g}_\rho \rangle \leq_{A_\mathbf{m}} \rho_N$ . This is precisely why Schneider introduces the inequality constraint:  $C_\mathbf{m}(\rho)$  is closed in the weak- $L^1$  topology, whereas  $C_\mathbf{m}^0(\rho)$  is not.

Such behavior begs the following question: For what values of  $\rho$  does a minimizing sequence for (14) *not* converge inside  $C_\mathbf{m}^0(\rho)$ ? These will be the densities which make up the set  $\mathcal{D}_\mathbf{m}$ . In [31], Schneider attempts to address this question in the following corollary to Theorem 3.

**COROLLARY 8** (Schneider [31]). *Given  $\rho \in \mathcal{R}_\mathbf{m}$ , the minimizer in (14) exists if and only if there is no function of the form  $G_\alpha$  in  $C_\mathbf{m}(\rho) \setminus C_\mathbf{m}^0(\rho)$ .*

Unfortunately, this result provides little understanding of the geometry of  $\mathcal{D}_\mathbf{m}$ . A more insightful point of view is given by the following theorem.

**THEOREM 9.** *Given  $\rho \in \mathcal{R}_\mathbf{m}$ , the minimization problem with equality constraints (14) has a minimizer if and only if  $\rho \in \mathcal{R}_\mathbf{m}^{\text{exp}}$ . In other words,*

$$(73) \quad \mathcal{D}_\mathbf{m} = \mathcal{R}_\mathbf{m} \setminus \mathcal{R}_\mathbf{m}^{\text{exp}}.$$

*Proof.* The “if” part of this theorem is just Corollary 4. The “only if” part will be proved at the end of section 4.3.  $\square$

An immediate consequence of Theorem 9 is that  $\mathcal{D}_{\mathbf{m}}$  is a cone. However, the essential point of the theorem is that when  $\mathcal{D}_{\mathbf{m}}$  is nonempty there are realizable densities  $\boldsymbol{\rho}$  that *cannot* be realized by a functions of the form  $G_{\boldsymbol{\alpha}}$ . In other words,  $\boldsymbol{\rho} \notin \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  even though  $\mathbf{a}(\boldsymbol{\rho}) \in \mathcal{A}_{\mathbf{m}}$ . It is this idea which lays the foundation for the results in [20, 21], where a description of  $\mathcal{D}_{\mathbf{m}}$  is given for the case  $\mathbf{m}_N = |v|^N$ . Theorem 9 will also be the basis for the new results of this paper. However, for a general admissible space  $\mathbb{M}$ , we will need to formulate the dual for relaxed constraint problem (20) and derive complementary slackness conditions in order to find a useful geometric description for  $\mathcal{D}_{\mathbf{m}}$ . In the process, we will recover and extend many of the results from [20, 21, 31].

**4. Dual formulation.** Because  $\mathcal{H}$  is convex on  $\mathbb{F}_{\mathbf{m}}$  and the constraints in (20) are linear, it is reasonable to apply a dual treatment to the relaxed-constraint problem, e.g., [3, 7, 26]. In this section, we prove two important duality theorems and the complementary slackness conditions that accompany them. We also give an alternate proof of the form of the minimizer in Theorem 3 and a proof of the “only if” part of Theorem 9.

**4.1. The dual function.** We define the Lagrangian function  $\mathcal{L} : \mathbb{F}_{\mathbf{m}} \times \mathbb{R}^n \times \mathcal{R}_{\mathbf{m}} \rightarrow \mathbb{R} \cup \{\infty\}$  associated to (20) by

$$(74) \quad \mathcal{L}(g, \boldsymbol{\alpha}, \boldsymbol{\rho}) \equiv \mathcal{H}(g) + \boldsymbol{\alpha}^T(\boldsymbol{\rho} - \langle \mathbf{m}g \rangle)$$

and the dual function  $\psi : \mathbb{R}^n \times \mathcal{R}_{\mathbf{m}} \rightarrow \mathbb{R} \cup \{-\infty\}$  by

$$(75) \quad \psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) \equiv \inf_{g \in \mathbb{F}_{\mathbf{m}}} \mathcal{L}(g, \boldsymbol{\alpha}, \boldsymbol{\rho}).$$

The dual function is closely related to the density potential  $h^*$ . In fact, we have the following.

**THEOREM 10.** *For all  $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}}$  and  $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ ,*

$$(76) \quad \psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) = \mathcal{L}(G_{\boldsymbol{\alpha}}, \boldsymbol{\alpha}, \boldsymbol{\rho}) = \boldsymbol{\alpha}^T \boldsymbol{\rho} - h^*(\boldsymbol{\alpha}).$$

*Proof.* We apply Young’s inequality (50) and make the identification  $z = g$  and  $y = G_{\boldsymbol{\alpha}}$  to derive the pointwise inequality

$$(77) \quad (g \log g - g) - \boldsymbol{\alpha}^T \mathbf{m}g \geq -G_{\boldsymbol{\alpha}}.$$

Integration of (77) over  $\mathbb{R}^d$  and addition of  $\boldsymbol{\alpha}^T \boldsymbol{\rho}$  to both sides give a lower bound on  $\mathcal{L}$  and hence  $\psi$ :

$$(78) \quad \psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) \geq \boldsymbol{\alpha}^T \boldsymbol{\rho} - h^*(\boldsymbol{\alpha}).$$

For  $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}}$ , the definitions of  $\mathcal{H}$ ,  $G_{\boldsymbol{\alpha}}$ , and  $h^*$  (given in (47), (38), and (39), respectively) imply that

$$(79) \quad \mathcal{H}(G_{\boldsymbol{\alpha}}) = \boldsymbol{\alpha}^T \langle \mathbf{m}G_{\boldsymbol{\alpha}} \rangle - \langle G_{\boldsymbol{\alpha}} \rangle = \boldsymbol{\alpha}^T \langle \mathbf{m}G_{\boldsymbol{\alpha}} \rangle - h^*(\boldsymbol{\alpha}).$$

Thus by (74),

$$(80) \quad \mathcal{L}(G_{\boldsymbol{\alpha}}, \boldsymbol{\alpha}, \boldsymbol{\rho}) = \boldsymbol{\alpha}^T \boldsymbol{\rho} - h^*(\boldsymbol{\alpha}),$$

so that, from (75),

$$(81) \quad \psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) \leq \boldsymbol{\alpha}^T \boldsymbol{\rho} - h^*(\boldsymbol{\alpha}).$$

Together (78), (80), and (81) imply (76).  $\square$

**4.2. Smoothness properties of the dual function.** The following smoothness properties of  $\psi$  will be used throughout the remainder of the paper.

THEOREM 11. *Let  $\rho \in \mathcal{R}_{\mathbf{m}}$ . Then*

1.  $\psi(\cdot, \rho)$  is strictly concave on  $\mathcal{A}_{\mathbf{m}}$  and infinitely Fréchet differentiable on  $\text{int } \mathcal{A}_{\mathbf{m}}$ , with derivatives

$$(82a) \quad \frac{\partial \psi}{\partial \alpha}(\alpha, \rho) = \rho - \mathbf{r}(\alpha),$$

$$(82b) \quad \frac{\partial^{(i)} \psi}{\partial \alpha^{(i)}}(\alpha, \rho) = - \left\langle \mathbf{m}^{\vee(i)} G_{\alpha} \right\rangle, \quad i > 1,$$

where  $\mathbf{m}^{\vee(i)}$  is the  $i$ th tensor power of  $\mathbf{m}$ ,<sup>4</sup>

2. for any  $\alpha, \beta \in \mathcal{A}_{\mathbf{m}}$ , the function

$$(83) \quad \phi(\tau) \equiv \psi(\tau\alpha + (1 - \tau)\beta, \rho)$$

is twice differentiable at each  $\tau \in [0, 1]$  (one-sided at end points) with derivatives

$$(84a) \quad \phi'(\tau) = (\alpha - \beta)^T [\rho - \mathbf{r}(\tau\alpha + (1 - \tau)\beta)],$$

$$(84b) \quad \phi''(\tau) = - \left\langle \left( (\alpha - \beta)^T \mathbf{m} \right)^2 G_{\tau\alpha + (1 - \tau)\beta} \right\rangle.$$

In particular, the function  $\phi'(\tau)$  is a decreasing function of  $\tau$ ;

3. the function  $\psi(\cdot, \rho)$  is upper semicontinuous on  $\mathcal{A}_{\mathbf{m}}$ .

*Proof.* For the proofs of the first two statements above, we refer the reader to Lemmas 5.1 and 5.2 in [21] along with a few comments. First, the lemmas in [21] refer to  $h^*$  rather than to  $\psi(\cdot, \rho)$ . This makes little difference since the two functions differ only by a linear factor (see Theorem 10). Also, the proofs in [21] are constructed specifically for the special case when  $m_N = |v|^N$ ; however, modifications to the general setting are straightforward. To prove the third statement we simply invoke Fatou's lemma. Given a sequence  $\{\alpha_{(i)}\}_{i=1}^{\infty} \subset \mathcal{A}_{\mathbf{m}}$  with limit  $\alpha \in \mathcal{A}_{\mathbf{m}}$ ,

$$(85) \quad \langle G_{\alpha} \rangle \leq \lim_{i \rightarrow \infty} \langle G_{\alpha_{(i)}} \rangle.$$

Hence  $\lim_{i \rightarrow \infty} \psi(\alpha_{(i)}, \rho) \leq \psi(\alpha, \rho)$ .  $\square$

COROLLARY 12. *For all  $\alpha \in \text{int } \mathcal{A}_{\mathbf{m}}$ ,  $h_{\alpha}^*(\alpha) = \mathbf{r}(\alpha)$  and  $h_{\alpha\alpha}^*(\alpha) = \langle \mathbf{m}\mathbf{m}^T G_{\alpha} \rangle$ , which is positive-definite on  $\alpha \in \text{int } \mathcal{A}_{\mathbf{m}}$ .*

Several remarks should be made concerning Theorem 11. First, statement 1 implies statement 2, but only for  $\alpha, \beta \in \text{int } \mathcal{A}_{\mathbf{m}}$ . Second, for  $\alpha, \beta \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ ,  $\phi''$  need not be continuous and higher derivatives may not exist. Finally, in spite of the smoothness properties given by Theorem 11, the dual function need not even be continuous on  $\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ . Indeed, given a sequence  $\{\alpha_{(i)}\}_{i=1}^{\infty} \in \mathcal{A}_{\mathbf{m}}$  with limit  $\alpha \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ , it is possible that  $h^*(\alpha) < \lim_{i \rightarrow \infty} h^*(\alpha_{(i)})$ . As an example, consider the one-dimensional case ( $d = 1$ ) when  $\mathbf{m} = (1, v, v^2, v^3, v^4)^T$ . This case has been studied in detail in [20]. Given the following five points in the  $(v, w)$  plane:

$$\begin{aligned} (v_0, w_0) &= (0, 0), & (v_1, w_1) &= (1, 0), & (v_2, w_2) &= (i, -i^2), \\ (v_3, w_3) &= (2i, i), & (v_4, w_4) &= (2i + 1, 0), \end{aligned}$$

<sup>4</sup>The tensor power of a symmetric tensor  $S$  is defined recursively. For  $n > 1$ ,  $S^{\vee(n)} \equiv S \vee S^{\vee(n-1)}$  while  $S^{\vee(1)} \equiv S$ .

the unique degree-four polynomial interpolating these points is

$$(86) \quad p_i(v) = \alpha_{(i)}^T \mathbf{m}(v) = \sum_{j=0}^4 \alpha_{(i)j} v^j,$$

where

$$\begin{aligned} \alpha_{0(i)} &= 0, & \alpha_{1(i)} &= \frac{2i+1}{4i-2} + \frac{4i^2+2i}{i^2-1}, & \alpha_{2(i)} &= -\frac{4i^2+6i+1}{i^2-1} - \frac{2i^2+4i+1}{4i^2-2i}, \\ \alpha_{3(i)} &= \frac{4i+2}{i^2-1} + \frac{3i+2}{4i^2-2i}, & \alpha_{4(i)} &= -\frac{1}{i^2-1} - \frac{1}{4i^2-2i}. \end{aligned}$$

(The notation  $\alpha_{(i)}$  denotes a sequence of vectors rather than the usual notation  $\alpha_i$ , which denotes the components of a single vector  $\alpha$  corresponding to polynomials of degree  $i$ .) As  $i \rightarrow \infty$ ,

$$(87) \quad \alpha_{(i)} \rightarrow \alpha_* = \left(0, \frac{9}{2}, -\frac{9}{2}, 0, 0\right)^T \quad \text{and} \quad G_{\alpha_*} = \exp\left(-\frac{9}{2}v^2 + \frac{9}{2}v\right).$$

The density potential  $h^*(\alpha_*)$  moments  $\mathbf{r}(\alpha_*)$  are finite. Therefore  $\alpha_* \in \mathcal{A}_{\mathbf{m}}$ , but clearly  $\alpha_* \notin \text{int } \mathcal{A}_{\mathbf{m}}$ . Moreover, one may readily check that  $p_i$  is positive and concave on the interval  $[2i, 2i+1]$ , and hence

$$(88) \quad h^*(\alpha_{(i)}) = \langle G_{\alpha_{(i)}} \rangle > \int_{2i}^{2i+1} e^{p_i(v)} dv > \int_{2i}^{2i+1} (1 + p_i(v)) dv > 1 + \frac{i}{2} \rightarrow \infty \quad \text{as } i \rightarrow \infty.$$

Note that the second inequality above follows from the fact that  $e^x > 1+x$ , while the concavity of  $p_i$  on  $[2i, 2i+1]$  implies that the graph of  $p_i$  lies above the line segment  $\ell$  joining the points  $(2i, i)$  and  $(2i+1, 0)$  in the  $(v, w)$  plane. Therefore the integral of  $p_i$  over  $[2i, 2i+1]$  is bounded below by the area of the triangle formed by  $\ell$ , the  $v$ -axis, and the line  $\{v = 2i\}$ . The area of this triangle is  $i/2$ . A similar argument shows that, for any  $j \geq 0$ ,  $\langle |v|^j G_{\alpha_{(i)}} \rangle \rightarrow \infty$  as  $i \rightarrow \infty$  while  $\langle v^j G_{\alpha_*} \rangle$  is finite.

The reason that  $\psi(\cdot, \rho)$  is discontinuous at the boundary of  $\mathcal{A}_{\mathbf{m}}$  is the same reason that the minimization problem (14) with equality constraints fails: because mass at the tails of the functions escapes as  $i \rightarrow \infty$ . In the example above, this is precisely what happens to the mass of  $G_{\alpha_{(i)}}$  that is supported on the interval  $[2i, 2i+1]$ . The same thing occurs with the minimizing sequence  $\{g_{i_k}\}_{k=1}^\infty$  in the proof of Theorem 3. The difference is that, for  $\{g_{i_k}\}_{k=1}^\infty$ , only the highest moments fail to converge in the minimizing sequence, whereas none of the moments in this example converge. The reason for this difference is that the moments  $\langle \mathbf{m} g_{i_k} \rangle$  are all bounded. The moments of  $\{G_{\alpha_{(i)}}\}_{i=1}^\infty$  would converge if higher-order moments were controlled in some way. Controlling the moments is, in effect, the same as requiring  $\alpha_i \rightarrow \alpha_*$  along a specified path. In fact, we will see at the very end of section 5.4 that the map  $\rho \mapsto \psi(\mathbf{a}(\rho), \rho)$  is continuous on  $\mathcal{R}_{\mathbf{m}}$ .

**4.3. Duality theorems.** The main results of this subsection are based on the following strong duality theorem where, for a given cone  $C$ , the notations “ $\leq_C$ ” and “ $\geq_C$ ” are defined in the last paragraph of section 2.2.

**THEOREM 13** (see [26]). *Consider the problem*

$$(89) \quad \begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && f_i(x) \leq_{K_i} 0, \quad i = 1, \dots, m; \quad Ax = b, \end{aligned}$$

where the functions  $f_0, \dots, f_m : \mathbb{X} \rightarrow \mathbb{R} \cup +\infty$  are convex over a vector space  $\mathbb{X}$ ,  $A : \mathbb{X} \rightarrow \mathbb{R}^k$  is a linear mapping,  $b \in \mathbb{R}^k$ , and each  $K_i$  is a proper cone for  $i = 1, \dots, m$ . Let  $D$  be the intersection of the domains of  $f_0, \dots, f_m$  (i.e.,  $D$  is a convex set over which each  $f_i$  is finite). Suppose there exists  $\tilde{x} \in D$ , with  $f_i(\tilde{x}) < 0$ ,  $i = 1, \dots, m$ , and  $A\tilde{x} = b$ . Further suppose that the set  $\{Ax - b : x \in D\}$  contains a neighborhood of the origin. Then strong duality holds, i.e.,

$$(90) \quad \inf\{f_0(x) : f_i(x) \leq_{K_i} 0, i = 1, \dots, m; Ax = b\} \\ = \sup_{\substack{\lambda_i \geq_{K_i^\circ} 0 \\ \nu \in \mathbb{R}^k}} \inf_{x \in D} \left\{ f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \nu^T (Ax - b) \right\},$$

and the dual optimal value is attained whenever it is not  $-\infty$ .

Theorem 13 follows from [26, Exercise 8.7] and can be proven by using arguments found in [26, Chapter 8]. It can also be proven along the lines of similar results found in [7, sections 5.3.2 and 5.9.1]. However, whereas those results require the existence of some  $\tilde{x}$  in the relative interior of  $D$ , Theorem 13 requires only that  $\tilde{x} \in D$ . A side benefit of this is that there is no need to specify a topology on  $\mathbb{X}$ . In return, our condition that  $\{Ax - b : x \in D\}$  contains a neighborhood of the origin is not present in the statements in [7].

To prove Theorem 13, one may repeat the arguments found in [7, section 5.3.2] with the notation “ $\leq$ ” changed to curly “ $\preceq$ .” The only difference from that proof is in the contradiction argument showing (in the notation of [7]) that  $\mu = 0$  is not possible. The proof in [7] first shows, with logic that remains valid under our weaker assumptions on  $\tilde{x}$ , that if  $\mu = 0$ , then there must exist  $\nu \neq 0$  such that  $\nu^T (Ax - b) \geq 0$  for all  $x \in D$ . At that point, our assumption that  $\{Ax - b : x \in D\}$  contains a neighborhood of the origin immediately implies that  $\nu = 0$ , which yields the requisite contradiction.

The statement of Theorem 13 is of much interest in the present context for two reasons: (i) Our primal decision variable  $g$  lies in an infinite-dimensional vector space, and (ii) it is not straightforward to show that the relative interior condition on  $\tilde{x}$  (or in our case  $\tilde{g}$ ) actually applies. (However, see [6, Definition 2.1], where the authors introduce the notion of a *pseudo relative interior*.) On the other hand, that our additional condition on  $\{Ax - b : x \in D\}$  holds is a direct consequence of the openness of  $\mathcal{R}_m$ .

Direct application of Theorem 13 leads to the following results.

**THEOREM 14.** Let  $\rho \in \mathcal{R}_m$ , and let  $h_S$  and  $\psi$  be given by (20) and (75), respectively. Then

$$(91) \quad h_S(\rho) = \max_{\alpha \in \mathcal{A}_m} \psi(\alpha, \rho),$$

where the maximum on the right is attained by a unique  $\hat{\alpha} \in \mathcal{A}_m$ . If  $\hat{g}_\rho$  solves (20), then  $\hat{g}_\rho = G_{\hat{\alpha}}$ . Furthermore,  $\hat{g}_\rho$  and  $\hat{\alpha}$  satisfy the complementary slackness condition

$$(92) \quad \hat{\alpha}^T \rho = \hat{\alpha}^T \langle m \hat{g}_\rho \rangle = \hat{\alpha}^T \langle m G_{\hat{\alpha}} \rangle,$$

and  $\hat{g}_\rho$  minimizes  $\mathcal{L}(g, \hat{\alpha}, \rho)$  over  $\mathbb{F}_m$ , i.e.,

$$(93) \quad \psi(\hat{\alpha}, \rho) = \mathcal{L}(\hat{g}_\rho, \hat{\alpha}, \rho).$$

*Proof.* Theorem 14 may be recast in the form of Theorem 13 by setting  $m = 1$  and introducing the following mapping of notation:

$$\begin{aligned} \mathbb{X} &\mapsto L_{\mathbf{m}}^1(\mathbb{R}^d); & f_1(x) &\mapsto \boldsymbol{\rho}_N - \langle \mathbf{m}_N g \rangle; & Ax &\mapsto \langle \mathbf{m}_j g \rangle, & j = 1, \dots, N-1; \\ x &\mapsto g; & K_1 &\mapsto A_{\mathbf{m}_N}^\circ; & b &\mapsto \boldsymbol{\rho}_j, & j = 1, \dots, N-1; \\ f_0 &\mapsto \mathcal{H}; & \lambda &\mapsto \boldsymbol{\alpha}_N; & \nu &\mapsto \boldsymbol{\alpha}_j, & j = 1, \dots, N-1. \end{aligned}$$

All of the conditions of Theorem 13 hold. However, we must be careful to ensure that  $\mathcal{H}$  is restricted to a domain on which it is finite. Thus we consider the minimization problem over the set

$$(94) \quad \tilde{\mathbb{F}}_{\mathbf{m}} = \{g \in \mathbb{F}_{\mathbf{m}} : \mathcal{H}(g) < \infty\}.$$

This set is convex and includes all bounded functions in  $\mathbb{F}_{\mathbf{m}}$  with compact support. Thus by Theorem 1, the moment mapping  $g \mapsto \langle \mathbf{m} g \rangle$  maps  $\tilde{\mathbb{F}}_{\mathbf{m}}$  onto  $\mathcal{R}_{\mathbf{m}}$ , and since  $\mathcal{R}_{\mathbf{m}}$  is open, the set

$$(95) \quad \left\{ \langle \mathbf{m}_i g \rangle - \boldsymbol{\rho}_i : g \in \tilde{\mathbb{F}}_{\mathbf{m}}, i < N \right\}$$

contains a neighborhood of the origin. By the polar cone theorem [3, page 162],  $(A_{\mathbf{m}_N}^\circ)^\circ = A_{\mathbf{m}_N}$  so that strong duality holds, i.e.,

$$(96) \quad h_s(\boldsymbol{\rho}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \{\psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) : \boldsymbol{\alpha}_N \in A_{\mathbf{m}_N}\}.$$

Moreover, because  $\psi$  is strictly concave, the maximum in (96) is attained by a *unique*  $\hat{\boldsymbol{\alpha}} \in \{\boldsymbol{\alpha} \in \mathbb{R}^n : \boldsymbol{\alpha}_N \in A_{\mathbf{m}_N}\}$ . According to the constraint conditions in (52),

$$(97) \quad \langle \mathbf{m}_j \hat{g}_{\boldsymbol{\rho}} \rangle = \boldsymbol{\rho} \text{ for } j < N \quad \text{and} \quad \hat{\boldsymbol{\alpha}}_N^T \langle \mathbf{m}_N \hat{g}_{\boldsymbol{\rho}} \rangle \geq \hat{\boldsymbol{\alpha}}_N^T \boldsymbol{\rho}_N.$$

Thus  $\hat{\boldsymbol{\alpha}}^T(\boldsymbol{\rho} - \langle \mathbf{m} \hat{g}_{\boldsymbol{\rho}} \rangle) \leq 0$  and

$$(98) \quad \begin{aligned} h_s(\boldsymbol{\rho}) &= \psi(\hat{\boldsymbol{\alpha}}, \boldsymbol{\rho}) = \inf_{g \in \tilde{\mathbb{F}}_{\mathbf{m}}} \left\{ \mathcal{H}(g) + \hat{\boldsymbol{\alpha}}^T(\boldsymbol{\rho} - \langle \mathbf{m} g \rangle) \right\} \\ &\leq \mathcal{H}(\hat{g}_{\boldsymbol{\rho}}) + \hat{\boldsymbol{\alpha}}^T(\boldsymbol{\rho} - \langle \mathbf{m} \hat{g}_{\boldsymbol{\rho}} \rangle) \leq \mathcal{H}(\hat{g}_{\boldsymbol{\rho}}) = h_s(\boldsymbol{\rho}). \end{aligned}$$

Equations (92) and (93) follow immediately.

To finish the proof, we need only show that  $\hat{\boldsymbol{\alpha}} \in \mathcal{A}_{\mathbf{m}}$  and  $\hat{g}_{\boldsymbol{\rho}} = G_{\hat{\boldsymbol{\alpha}}}$ . For any nonnegative function  $g$ , straightforward calculation verifies that

$$(99) \quad g \log g - g - \hat{\boldsymbol{\alpha}}^T \mathbf{m} g = \phi(g) - G_{\hat{\boldsymbol{\alpha}}},$$

where

$$(100) \quad \phi(g) \equiv \left[ g \log \left( \frac{g}{G_{\hat{\boldsymbol{\alpha}}}} \right) + (G_{\hat{\boldsymbol{\alpha}}} - g) \right].$$

Applying (49) with  $z = g/G_{\hat{\boldsymbol{\alpha}}}$  and  $y = 1$  shows that, for each  $v \in \mathbb{R}^d$ ,  $\phi(g(v)) \geq 0$ , with equality if and only if  $G_{\hat{\boldsymbol{\alpha}}}(v) = g(v)$ . Now, for each  $R > 0$ , define the set  $B_R \equiv \{v \in \mathbb{R}^d : |v| < R\}$ . Setting  $g = \hat{g}_{\boldsymbol{\rho}}$  in (99) and integrating over  $B_R$  gives

$$(101) \quad \mathcal{H}^{B_R}(\hat{g}_{\boldsymbol{\rho}}) - \left\langle \hat{\boldsymbol{\alpha}}^T \mathbf{m} \hat{g}_{\boldsymbol{\rho}} \right\rangle_{B_R} = \langle \phi(\hat{g}_{\boldsymbol{\rho}}) \rangle_{B_R} - \langle G_{\hat{\boldsymbol{\alpha}}} \rangle_{B_R}.$$



(Note that, since  $\hat{g}_\rho \in \mathbb{F}_\mathbf{m}$ , all of the integrals above are well-defined.) From (74), (101), (76), and (39), it follows that

$$\begin{aligned}
 \mathcal{L}(\hat{g}_\rho, \hat{\alpha}, \rho) &= \mathcal{H}(\hat{g}_\rho) + \hat{\alpha}^T(\rho - \langle \mathbf{m}\hat{g}_\rho \rangle) \\
 &= \mathcal{H}^{B_R}(\hat{g}_\rho) + \mathcal{H}^{\mathbb{R}^d \setminus B_R}(\hat{g}_\rho) + \hat{\alpha}^T \left( \rho - \langle \mathbf{m}\hat{g}_\rho \rangle_{B_R} - \langle \mathbf{m}\hat{g}_\rho \rangle_{\mathbb{R}^d \setminus B_R} \right) \\
 (102) \quad &= \langle \phi(\hat{g}_\rho) \rangle_{B_R} - \langle G_{\hat{\alpha}} \rangle_{B_R} + \mathcal{H}^{\mathbb{R}^d \setminus B_R}(\hat{g}_\rho) + \hat{\alpha}^T \left( \rho - \langle \mathbf{m}\hat{g}_\rho \rangle_{\mathbb{R}^d \setminus B_R} \right) \\
 &= \mathcal{L}(G_{\hat{\alpha}}^{B_R}, \hat{\alpha}, \rho) + \langle \phi(\hat{g}_\rho) \rangle_{B_R} + \mathcal{H}^{\mathbb{R}^d \setminus B_R}(\hat{g}_\rho) - \hat{\alpha}^T \langle \mathbf{m}\hat{g}_\rho \rangle_{\mathbb{R}^d \setminus B_R},
 \end{aligned}$$

where

$$(103) \quad G_{\hat{\alpha}}^{B_R}(v) = \begin{cases} G_{\hat{\alpha}}(v), & v \in B_R, \\ 0, & v \notin B_R. \end{cases}$$

Now since  $\phi(g(v)) \geq 0$ , the function  $R \mapsto \Phi(R) \equiv \langle \phi(\hat{g}_\rho) \rangle_{B_R}$  is nonnegative and nondecreasing, and  $\Phi(R) = 0$  if and only if  $G_{\hat{\alpha}}$  and  $g$  agree on  $B_R$ . On the other hand, since  $\mathcal{H}(\hat{g}_\rho)$  and  $\langle \mathbf{m}\hat{g}_\rho \rangle$  are finite,

$$(104) \quad \mathcal{H}^{\mathbb{R}^d \setminus B_R}(\hat{g}_\rho) - \hat{\alpha}^T \langle \mathbf{m}\hat{g}_\rho \rangle_{\mathbb{R}^d \setminus B_R} \rightarrow 0 \quad \text{as } R \rightarrow \infty.$$

It follows then from (102) that, for  $R$  is sufficiently large,

$$(105) \quad \mathcal{L}(\hat{g}_\rho, \hat{\alpha}, \rho) > \mathcal{L}(G_{\hat{\alpha}}^{B_R}, \hat{\alpha}, \rho)$$

unless  $G_{\hat{\alpha}}^{B_R}$  agrees with  $\hat{g}_\rho$  on  $B_R$ . Since  $\hat{g}_\rho$  minimizes  $\mathcal{L}(\cdot, \hat{\alpha}, \rho)$ , we conclude that this exception is indeed the case. Moreover, since  $R$  is arbitrary, it follows that  $\hat{g}_\rho = G_{\hat{\alpha}}$ . Finally, the fact that  $\hat{g}_\rho \in \mathbb{F}_\mathbf{m}$  implies that  $\hat{\alpha} \in \mathcal{A}_\mathbf{m}$ .  $\square$

Several remarks are in order here.

1. If  $\rho \in \mathcal{R}_\mathbf{m}^{\text{exp}}$ , then Theorem 14 can be proven more directly using by Theorem 3.
3. Indeed, weak duality is easy to show: If  $g \in \mathbb{F}_\mathbf{m}$  satisfies the constraint conditions from (52), then

$$(106) \quad \mathcal{L}(g, \alpha, \rho) = \mathcal{H}(g) + \alpha^T(\rho - \langle \mathbf{m}g \rangle) \leq \mathcal{H}(g)$$

for all  $\alpha \in \mathcal{A}_\mathbf{m}$ . By invoking the definitions of  $\psi$  (75) and  $h_S$  (20), we find that

$$\begin{aligned}
 \psi(\alpha, \rho) &= \inf_{g \in \mathbb{F}_\mathbf{m}} \mathcal{L}(g, \alpha, \rho) \leq \inf_{g \in \mathbb{F}_\mathbf{m}} \{ \mathcal{L}(g, \alpha, \rho) : \langle \mathbf{m}g \rangle \preceq^\circ \rho \} \\
 (107) \quad &\leq \inf_{g \in \mathbb{F}_\mathbf{m}} \{ \mathcal{H}(g) : \langle \mathbf{m}g \rangle \preceq^\circ \rho \} = h_S(\rho).
 \end{aligned}$$

On the other hand, if  $\rho = \mathbf{r}(\hat{\alpha})$  for some  $\hat{\alpha} \in \mathcal{A}_\mathbf{m}$ , then it follows from Theorem 3, (76), and the definition of  $\mathcal{H}$  (47) that

$$(108) \quad h_S(\rho) = \mathcal{H}(G_{\hat{\alpha}}) = \psi(\hat{\alpha}, \rho).$$

From (107) and (108), one can easily deduce strong duality (91) and the complementary slackness condition (92).

2. If it is known a priori that the maximum in (96) is attained by  $\hat{\alpha} \in \mathcal{A}_\mathbf{m}$ , then the form of the minimizer follows almost immediately. In this case,  $G_{\hat{\alpha}} \in \mathbb{F}_\mathbf{m}$  (which is needed for  $\mathcal{L}$  to be well-defined) so that (76) and (93) imply (108). Because  $\mathcal{L}$  is strictly convex in its first argument, its minimizer is unique, and consequently  $\hat{g}_\rho = G_{\hat{\alpha}}$ .

3. Since  $\rho_j = \langle \mathbf{m}_j G_{\hat{\alpha}} \rangle$  for  $j < N$ , the only nontrivial part of the complementary slackness condition (92) is

$$(109) \quad \hat{\alpha}_N^T \rho_N = \hat{\alpha}_N^T \langle \mathbf{m}_N \hat{g}_\rho \rangle = \hat{\alpha}_N^T \langle \mathbf{m}_N G_{\hat{\alpha}} \rangle.$$

This relationship between  $\hat{\alpha}_N$  and  $\rho_N$  will be the key to characterizing the set  $\mathcal{D}_{\mathbf{m}}$ .

The following corollary will be used in section 6. It is an immediate consequence of the complementary slackness condition.

COROLLARY 15. *Let  $\rho \in \mathcal{R}_{\mathbf{m}}$ , and let  $\hat{\alpha} \in \mathcal{A}_{\mathbf{m}}$  solve (91). Then*

$$(110) \quad h_S(\rho) = \min_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) : \hat{\alpha}^T \langle \mathbf{m}g \rangle = \hat{\alpha}^T \rho \right\},$$

and  $G_{\hat{\alpha}}$  is the unique minimizer.

*Proof.* Let  $g \in \mathbb{F}_{\mathbf{m}}$  be given. Using Young's inequality (49) with  $z = g$  and  $y = G_{\hat{\alpha}}$  gives

$$(111) \quad \mathcal{H}(g) \geq \mathcal{H}(G_{\hat{\alpha}}) + \hat{\alpha}^T \langle \mathbf{m}(g - G_{\hat{\alpha}}) \rangle,$$

which, given the complementary slackness condition (92), implies that

$$(112) \quad \mathcal{H}(g) \geq \mathcal{H}(G_{\hat{\alpha}}) + \hat{\alpha}^T (\langle \mathbf{m}g \rangle - \rho).$$

Thus if  $g$  satisfies the constraints in (110), then  $\mathcal{H}(g) \geq \mathcal{H}(G_{\hat{\alpha}}) = h_S(\rho)$ .  $\square$

A duality theorem similar to Theorem 14 holds for the minimization problem in (19) that defines  $h_J(\rho)$ . Like Theorem 14, it is a consequence of Theorem 13, and its proof is essentially the same.

THEOREM 16. *Let  $\rho \in \mathcal{R}_{\mathbf{m}}$ , and let  $h_J(\rho)$  and  $\psi$  be given by (19) and (75), respectively. Then*

$$(113) \quad h_J(\rho) = \max_{\alpha \in \mathcal{A}_{\mathbf{m}}} \psi(\alpha, \rho),$$

where the maximum on the right is attained by a unique  $\tilde{\alpha} \in \mathcal{A}_{\mathbf{m}}$ . Furthermore, if the infimum in (19) is attained by some function  $\tilde{g}_\rho \in \mathbb{F}_{\mathbf{m}}$  which satisfies the equality constraints of (19), then  $\tilde{g}_\rho = G_{\tilde{\alpha}}$  and  $\tilde{g}_\rho$  minimizes  $\mathcal{L}(g, \tilde{\alpha}, \rho)$ , i.e.,  $\psi(\tilde{\alpha}, \rho) = \mathcal{L}(\tilde{g}_\rho, \tilde{\alpha}, \rho)$ .

The careful reader may note that application of Theorem 13 to proving Theorem 16 initially gives a statement similar to (96) but without any constraint on  $\alpha$ . However, the arguments which follow (96) show that  $\alpha \in \mathcal{A}_{\mathbf{m}}$  independently of this initial restriction.

Theorems 14 and 16 prove that the infima in (19) and (20) are equal—that is,

$$(114) \quad h_S(\rho) = h_J(\rho) = \max_{\alpha \in \mathcal{A}_{\mathbf{m}}} \psi(\alpha, \rho),$$

even if the infimum in (19) is not attained. In light of (114), the definition of  $h$  given in (17), which applies only to  $\rho \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ , can be extended to all of  $\mathcal{R}_{\mathbf{m}}$  by setting

$$(115) \quad h(\rho) \equiv \max_{\alpha \in \mathcal{A}_{\mathbf{m}}} \psi(\alpha, \rho).$$

In addition, we can now complete the proof of Theorem 9.

*Proof of Theorem 9.* We have already proven the “if” statement in Theorem 9. We now prove the “only if” statement. To this end, let  $\rho \in \mathcal{R}_{\mathbf{m}}$  be such that (14) has a minimizer. According to (114) this minimizer is also the minimizer of (20) and is therefore given by  $G_{\mathbf{a}(\rho)}$ . Hence, the equality constraint conditions in (14) imply that  $\rho = \langle \mathbf{m}G_{\mathbf{a}(\rho)} \rangle$ , which means  $\rho \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ .  $\square$

**5. The relationship between  $\alpha$  and  $\rho$ .** The formal structure of entropy-based closures depends heavily on the Legendre dual relationship between the functions  $h$  and  $h^*$  and their derivatives. In this section, we review this relationship for nondegenerate densities and show how Legendre duality ensures that the resulting system of PDEs is symmetric hyperbolic. We then discuss what aspects of the dual relationship hold in degenerate densities. A similar analysis can be found in [21] for the case  $\mathbf{m}_N = |v|^N$ .

**5.1. Properties for nondegenerate cases.** Recall that the function  $\mathbf{a}$  maps each  $\rho \in \mathcal{R}_{\mathbf{m}}$  to the unique vector  $\hat{\alpha} \in \mathcal{A}_{\mathbf{m}}$  that solves (91). In particular,

$$(116) \quad \hat{g}_{\rho} = G_{\mathbf{a}(\rho)} \quad \text{and} \quad h(\rho) = \psi(\mathbf{a}(\rho), \rho).$$

It turns out that  $\mathbf{a}$ , when restricted to  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$ , is the inverse of the function  $\mathbf{r}$  defined in (41).

**THEOREM 17.** *The function  $\mathbf{r}$  is one-to-one from  $\mathcal{A}_{\mathbf{m}}$  onto  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$  with inverse  $\mathbf{a}$ . It is a diffeomorphism between  $\text{int } \mathcal{A}_{\mathbf{m}}$  and  $\text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ .*

*Proof.* We first identify  $\mathbf{a}$  as the inverse of  $\mathbf{r}$ . Since  $\mathbf{r}$  is (by definition) onto  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$ , we need only to show that  $\mathbf{a}(\mathbf{r}(\alpha)) = \alpha$  for each  $\alpha \in \mathcal{A}_{\mathbf{m}}$ . By the definition of  $\mathbf{a}$ ,

$$(117) \quad \mathcal{H}(G_{\mathbf{a}(\mathbf{r}(\alpha))}) = \min_{g \in \mathbb{F}_{\mathbf{m}}} \{ \mathcal{H}(g) : \langle \mathbf{m}g \rangle \preceq^{\circ} \mathbf{r}(\alpha) \}.$$

However, Theorem 3 implies that

$$(118) \quad \mathcal{H}(G_{\alpha}) = \min_{g \in \mathbb{F}_{\mathbf{m}}} \{ \mathcal{H}(g) : \langle \mathbf{m}g \rangle \preceq^{\circ} \mathbf{r}(\alpha) \}.$$

Since this minimizer is unique, it follows that  $\mathbf{a}(\mathbf{r}(\alpha)) = \alpha$ . If  $\alpha \in \text{int } \mathcal{A}_{\mathbf{m}}$ , then, according to Corollary 12,  $\mathbf{r}$  is the derivative of the density potential  $h^*$  on  $\text{int } \mathcal{A}_{\mathbf{m}}$  and its Jacobian

$$(119) \quad \frac{\partial \mathbf{r}}{\partial \alpha}(\alpha) = \frac{\partial^2 h^*}{\partial \alpha^2}(\alpha, \rho) = \langle \mathbf{m} \mathbf{m}^T G_{\alpha} \rangle$$

is a positive-definite matrix. The inverse function theorem implies then that  $\mathbf{r}$  is a diffeomorphism from  $\text{int } \mathcal{A}_{\mathbf{m}}$  onto  $\text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ .  $\square$

The following corollary implies that  $\mathcal{D}_{\mathbf{m}}$  cannot divide  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$  into disjoint subsets.

**COROLLARY 18.** *The set  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$  is pathwise-connected.*

*Proof.* Given  $\rho_{(0)}, \rho_{(1)} \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ , we seek a continuous function  $\Gamma : [0, 1] \rightarrow \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  such that

$$(120) \quad \Gamma(0) = \rho_{(0)} \quad \text{and} \quad \Gamma(1) = \rho_{(1)}.$$

Convexity of  $\mathcal{A}_{\mathbf{m}}$  implies that

$$(121) \quad \alpha_{\lambda} \equiv \lambda \mathbf{a}(\rho_{(0)}) + (1 - \lambda) \mathbf{a}(\rho_{(1)}) \in \mathcal{A}_{\mathbf{m}} \quad \forall \lambda \in [0, 1].$$

Thus, in view of Theorem 11, the function  $\Gamma(\lambda) = \mathbf{r}(\alpha_{\lambda})$  satisfies (120).  $\square$

An immediate consequence of Theorem 14 is that  $h$  (as the maximum of a family of linear functions in  $\rho$ ) is convex on  $\mathcal{R}_{\mathbf{m}}$ . However, more can be said if we restrict  $h$  to convex subsets of  $\text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ .

**THEOREM 19.** *When restricted to  $\text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  and  $\text{int } \mathcal{A}_{\mathbf{m}}$ , respectively, the functions  $h$  and  $h^*$  are locally strictly convex, Legendre duals of one another.*

*Proof.* We first show that  $h$  is the Legendre transform of  $h^*$ . From (76),

$$(122) \quad h(\mathbf{r}(\boldsymbol{\alpha})) + h^*(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \mathbf{r}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}},$$

where, according to Corollary 12,

$$(123) \quad \mathbf{r}(\boldsymbol{\alpha}) = h_{\boldsymbol{\alpha}}^*(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} \in \text{int } \mathcal{A}_{\mathbf{m}}.$$

We next show that the Legendre transform of  $h^*$  recovers  $h$ . The inverse relationship between  $\mathbf{a}$  and  $\mathbf{r}$  (Theorem 17) implies that (122) may be rewritten in terms of  $\boldsymbol{\rho} = \mathbf{r}(\boldsymbol{\alpha})$ :

$$(124) \quad h(\boldsymbol{\rho}) + h^*(\mathbf{a}(\boldsymbol{\rho})) = \mathbf{a}(\boldsymbol{\rho})^T \boldsymbol{\rho}, \quad \boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}.$$

Differentiating (124) and using (123) again gives

$$(125) \quad \mathbf{a}(\boldsymbol{\rho}) = h_{\boldsymbol{\rho}}(\boldsymbol{\rho}), \quad \boldsymbol{\rho} \in \text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}.$$

Finally, for  $\boldsymbol{\rho} \in \text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ ,

$$(126) \quad h_{\boldsymbol{\rho}\boldsymbol{\rho}}(\boldsymbol{\rho}) = \frac{\partial \mathbf{a}}{\partial \boldsymbol{\rho}}(\boldsymbol{\rho}) = \left[ \frac{\partial \mathbf{r}}{\partial \boldsymbol{\alpha}}(\mathbf{a}(\boldsymbol{\rho})) \right]^{-1} = [h_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^*(\mathbf{a}(\boldsymbol{\rho}))]^{-1},$$

which, by Corollary 12, is positive-definite. Thus  $h$  and  $h^*$  are strictly convex.  $\square$

**5.2. Application to kinetic moment closures.** The dual relationship between  $h$  and  $h^*$  is used in [24] to show that entropy-based closures formally produce hyperbolic systems which dissipate a convex entropy and satisfy an H-theorem. Indeed, if  $\boldsymbol{\rho} \in \text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  and  $\hat{\boldsymbol{\alpha}} = \mathbf{a}(\boldsymbol{\rho})$ , then, according to (123), the moment system (12) can be expressed in terms of  $\hat{\boldsymbol{\alpha}}$ :

$$(127) \quad \partial_t h_{\boldsymbol{\alpha}}^*(\hat{\boldsymbol{\alpha}}) + \nabla_x \cdot j_{\boldsymbol{\alpha}}^*(\hat{\boldsymbol{\alpha}}) = \mathbf{c}(h_{\boldsymbol{\alpha}}^*(\hat{\boldsymbol{\alpha}})),$$

where  $j^*(\boldsymbol{\alpha}) \equiv \langle v G_{\boldsymbol{\alpha}} \rangle$  is the flux potential and

$$(128) \quad j_{\boldsymbol{\alpha}}^*(\hat{\boldsymbol{\alpha}}) = \mathbf{f}(\boldsymbol{\rho}).$$

Carrying out the time and space derivatives in (127) gives

$$(129) \quad h_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^*(\hat{\boldsymbol{\alpha}}) \partial_t \hat{\boldsymbol{\alpha}} + j_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^*(\hat{\boldsymbol{\alpha}}) \cdot \nabla_x \hat{\boldsymbol{\alpha}} = \mathbf{c}(h_{\boldsymbol{\alpha}}^*(\hat{\boldsymbol{\alpha}})),$$

which has the form of a symmetric hyperbolic system [14]. Furthermore, by multiplying (12) by  $h_{\boldsymbol{\rho}}$  and applying relations (125) and (128), we find that  $h(\boldsymbol{\rho})$  satisfies:

$$(130) \quad \partial_t h(\boldsymbol{\rho}) + \nabla_x \cdot j(\boldsymbol{\rho}) = \mathbf{a}(\boldsymbol{\rho})^T \mathbf{c}(\boldsymbol{\rho}),$$

where  $j(\boldsymbol{\rho}) \equiv \mathbf{a}(\boldsymbol{\rho})^T \mathbf{f}(\boldsymbol{\rho}) - j^*(\mathbf{a}(\boldsymbol{\rho}))$ . Then by (5) and (6),

$$(131) \quad \mathbf{a}(\boldsymbol{\rho})^T \mathbf{c}(\boldsymbol{\rho}) = \mathcal{S}(G_{\mathbf{a}(\boldsymbol{\rho})}) \leq 0$$

with equality if and only if  $G_{\mathbf{a}(\boldsymbol{\rho})}$  is a local Maxwellian (7). This is a direct analogue of Boltzmann's H-theorem for (2). (See [24] for details.)

**5.3. Nondegenerate examples.** For  $N = 2$ , there are two possible closures: Maxwellian and Gaussian. Both are well-known, and in both cases  $\mathcal{A}_{\mathbf{m}} = \text{int } \mathcal{A}_{\mathbf{m}}$  and  $\mathcal{R}_{\mathbf{m}} = \mathcal{R}_{\mathbf{m}}^{\text{exp}} = \text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ . Traditionally, these closures are expressed by using so-called fluid variables:

(132)

$$\begin{aligned} \text{density: } \rho &= \langle F \rangle, & \text{temperature matrix: } \Theta &= \frac{\langle (v-u) \vee (v-u) F \rangle}{\langle F \rangle}, \\ \text{bulk velocity: } u &= \frac{\langle v F \rangle}{\langle F \rangle}, & \text{temperature: } \theta &= \frac{1}{3} \text{trace}(\Theta) = \frac{\langle |v-u|^2 F \rangle}{3 \langle F \rangle}. \end{aligned}$$

1. *Maxwellian closure.* If  $\mathbf{m} = (1, v, \frac{1}{2}|v|^2)^T$ , the ansatz  $\mathcal{F}[\rho]$  in (14) is a Maxwellian distribution:

$$(133) \quad \mathcal{M}_{\rho, u, \theta}(v) \equiv \frac{\rho}{(2\pi\theta)^{d/2}} \exp\left(-\frac{|v-u|^2}{2\theta}\right).$$

The fluid variables are related to the densities  $\rho_i$  by

$$(134) \quad \rho_0 = \rho, \quad \rho_1 = \rho u, \quad \rho_2 = \frac{1}{2}\rho u^2 + \frac{3}{2}\rho\theta$$

and to the vectors  $\hat{\alpha}_i$  by

$$(135) \quad \hat{\alpha}_0 = \log\left(\frac{\rho}{(2\pi\theta)^{d/2}}\right) - \frac{|u|^2}{2\theta}, \quad \hat{\alpha}_1 = \frac{u}{\theta}, \quad \hat{\alpha}_2 = -\frac{1}{\theta}.$$

The moment equations in this case are the compressible Euler equations for a gas of point particles:

$$(136a) \quad \partial_t \rho + \nabla_x \cdot (\rho u) = 0,$$

$$(136b) \quad \partial_t (\rho u) + \nabla_x \cdot (\rho u \vee u + \rho \theta I) = 0,$$

$$(136c) \quad \partial_t \left( \frac{1}{2} \rho |u|^2 + \frac{d}{2} \rho \theta \right) + \nabla_x \cdot \left( \frac{1}{2} \rho |u|^2 u + \frac{d+2}{2} \rho \theta u \right) = 0.$$

The spatial entropy

$$(137) \quad h(\rho) = \langle \mathcal{M}_{\rho, u, \theta} \log \mathcal{M}_{\rho, u, \theta} - \mathcal{M}_{\rho, u, \theta} \rangle = \rho \left[ \log \left( \frac{\rho}{(2\pi\theta)^{d/2}} \right) - \frac{d+2}{2} \right]$$

is locally conserved by smooth solutions for (136) but is dissipated along shocks.

2. *Gaussian closure.* If  $\mathbf{m} = (1, v, v \vee v)^T$ , the ansatz  $\mathcal{F}[\rho]$  in (14) is a Gaussian distribution:

$$(138) \quad \mathcal{G}_{\rho, u, \Theta}(v) = \frac{\rho}{\sqrt{\det(2\pi\Theta)}} \exp\left(-\frac{1}{2}(v-u) \cdot \Theta^{-1} \cdot (v-u)\right).$$

The fluid variables are related to the densities  $\rho_i$  by

$$(139) \quad \rho_0 = \rho, \quad \rho_1 = \rho u, \quad \rho_2 = \rho u \vee u + \rho \Theta$$

and to the vectors  $\hat{\alpha}_i$  by

$$(140) \quad \hat{\alpha}_0 = \log \left( \frac{\rho}{\sqrt{\det(2\pi\Theta)}} \right) - \frac{1}{2} u \cdot \Theta^{-1} \cdot u, \quad \hat{\alpha}_1 = \Theta^{-1} \cdot u, \quad \hat{\alpha}_2 = -\frac{1}{2} \Theta^{-1}.$$

The moment equations in this case are

$$(141a) \quad \partial_t \rho + \nabla_x \cdot (\rho u) = 0,$$

$$(141b) \quad \partial_t (\rho u) + \nabla_x \cdot (\rho u \vee u + \rho \Theta) = 0,$$

$$(141c) \quad \partial_t (\rho u \vee u + \rho \Theta) + \nabla_x \cdot (\rho u \vee u \vee u + 3\rho \Theta \vee u) = \langle v \vee v \mathcal{C}(\mathcal{G}_{\rho,u,\Theta}) \rangle,$$

and solutions to this system satisfy a local dissipation law for the spatial entropy

$$(142) \quad h(\rho) = \langle \mathcal{G}_{\rho,u,\Theta} \log \mathcal{G}_{\rho,u,\Theta} - \mathcal{G}_{\rho,u,\Theta} \rangle = \rho \left[ \log \left( \frac{\rho}{\sqrt{\det(2\pi\Theta)}} \right) - \frac{d+2}{2} \right].$$

Note that, in both of the examples above, the expressions for  $\hat{\alpha}$  and  $\rho$  can be used to determine  $\mathbf{a}(\rho)$  explicitly. However, generally speaking, an analytical solution is not available, and a numerical solution must be computed via (91).

**5.4. Properties for degenerate cases.** If  $\rho \in \mathcal{D}_{\mathbf{m}}$ , then the minimizer with equality constraints (14) does not exist, and the entropy-based closure is not well-defined. Although it is possible to recover a well-defined closure by using the relaxed constraints in (20), much of the formal structure is lost. For example, if  $\rho \in \mathcal{D}_{\mathbf{m}}$ , then (123) and (126) no longer hold because  $\mathbf{r}(\mathbf{a}(\rho)) \neq \rho$  and, as shown in Corollary 22 below,  $h_S$  fails to be strictly convex on  $\mathcal{R}_{\mathbf{m}}$  whenever  $\mathcal{D}_{\mathbf{m}}$  is nonempty. Since many of the properties of entropy-based closures require  $h$  to be strictly convex, this fact is critical.

The situation for degenerate densities may be best understood via the projection operator  $\pi : \mathcal{R}_{\mathbf{m}} \rightarrow \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ , which assigns to each vector  $\rho \in \mathcal{R}_{\mathbf{m}}$  the density which is realized by the minimizer of (20):

$$(143) \quad \pi(\rho) \equiv \mathbf{r}(\mathbf{a}(\rho)) = \langle \mathbf{m} G_{\mathbf{a}(\rho)} \rangle.$$

Before discussing  $\pi$  further, we introduce some notation that will be useful for the remainder of the paper. First we have the natural decompositions for  $\mathbf{r}$ ,  $\mathbf{a}$ , and  $\pi$  based on the decomposition of  $\mathbf{m}$  in (22):

$$(144) \quad \mathbf{r} = (\mathbf{r}_0^T, \mathbf{r}_1^T, \dots, \mathbf{r}_N^T)^T, \quad \mathbf{a} = (\mathbf{a}_0^T, \mathbf{a}_1^T, \dots, \mathbf{a}_N^T)^T, \quad \pi = (\pi_0^T, \pi_1^T, \dots, \pi_N^T)^T.$$

With this notation,

$$(145) \quad G_{\mathbf{a}(\rho)} = \exp \left( \sum_{j=1}^N \mathbf{a}_j(\rho)^T \mathbf{m}_j \right), \quad \mathbf{r}_j(\alpha) = \langle \mathbf{m}_j G_{\alpha} \rangle, \quad \pi_j(\rho) = \mathbf{r}_j(\mathbf{a}(\rho)).$$

Next, for any  $\rho \in \mathbb{R}^n$  and any  $\zeta \in \mathbb{R}^{n_N}$ , we define

$$(146) \quad \rho +_N \zeta \equiv (\rho_0^T, \rho_1^T, \dots, \rho_N^T + \zeta^T)^T.$$

This notation will often be applied to subsets of  $\mathbb{R}^n$  and  $\mathbb{R}^{n_N}$  in the context of set addition.

PROPOSITION 20. Let  $\bar{\rho} \in \mathcal{R}_m^{\text{exp}}$ , and let  $\bar{\alpha} = \mathbf{a}(\bar{\rho})$ . Then for any  $\rho \in \mathbb{R}^n$ , the following are equivalent:

$$(147a) \quad (i) \quad \rho_N - \bar{\rho}_N \in \mathcal{NC}(A_{m_N}, \bar{\alpha}_N);$$

$$(147b) \quad (ii) \quad (\alpha_N - \bar{\alpha}_N)^T (\rho_N - \bar{\rho}_N) \leq 0 \quad \forall \alpha_N \in A_{m_N};$$

$$(147c) \quad (iii) \quad \bar{\alpha}_N^T (\rho_N - \bar{\rho}_N) = 0 \quad \text{and} \quad \alpha_N^T (\rho_N - \bar{\rho}_N) \leq 0 \quad \forall \alpha_N \in A_{m_N}.$$

*Proof.* Here (i)  $\Leftrightarrow$  (ii) is just the definition of a normal cone (31), and the implication that (iii)  $\Rightarrow$  (ii) is clear. To prove that (ii)  $\Rightarrow$  (iii), we use the freedom to choose any  $\alpha_N \in A_{m_N}$ . Setting  $\alpha_N = 0$  and then  $\alpha_N = 2\bar{\alpha}_N$  in (147b) gives

$$(148) \quad \bar{\alpha}_N^T (\rho_N - \bar{\rho}_N) \geq 0 \quad \text{and} \quad \bar{\alpha}_N^T (\rho_N - \bar{\rho}_N) \leq 0,$$

respectively. We conclude that  $\bar{\alpha}_N^T (\rho_N - \bar{\rho}_N) = 0$ , which, when substituted back into (147b), gives the inequality in (iii).  $\square$

LEMMA 21. The projection  $\pi$  satisfies the following relations:

$$(149a) \quad (i) \quad \pi_j(\rho) = \rho_j \quad \forall \rho \in \mathcal{R}_m \text{ and } j < N;$$

$$(149b) \quad (ii) \quad \mathbf{a}_N(\rho)^T \pi_N(\rho) = \mathbf{a}_N(\rho)^T \rho_N \quad \forall \rho \in \mathcal{R}_m;$$

$$(149c) \quad (iii) \quad \pi(\rho) = \rho \text{ if and only if } \rho \in \mathcal{R}_m^{\text{exp}};$$

$$(149d) \quad (iv) \quad \pi(\{\bar{\rho} +_N \mathcal{NC}(A_{m_N}, \bar{\alpha}_N)\} \cap O) = \bar{\rho} \\ \forall \bar{\rho} \in \mathcal{R}_m^{\text{exp}} \text{ and any } O \subset \mathcal{R}_m \text{ containing } \bar{\rho};$$

$$(149e) \quad (v) \quad \pi(\mathcal{D}_m) = \mathbf{r}(\mathcal{A}_m \cap \partial \mathcal{A}_m) = \mathcal{R}_m^{\text{exp}} \cap \partial \mathcal{R}_m^{\text{exp}};$$

$$(149f) \quad (vi) \quad \mathbf{a}(\pi(\rho)) = \mathbf{a}(\rho) \quad \forall \rho \in \mathcal{R}_m;$$

$$(149g) \quad (vii) \quad h(\pi(\rho)) = h(\rho) \quad \forall \rho \in \mathcal{R}_m.$$

*Proof.* We prove each statement in order.

1. Equation (149a) follows from the constraint conditions in (52a).
2. Equation (149b) is just a restatement of the nontrivial component of the complementary slackness condition (109) with  $\hat{\alpha} = \mathbf{a}(\rho)$ .
3. By Theorem 17,  $\pi = \mathbf{r} \circ \mathbf{a}$  is the identity map on  $\mathcal{R}_m^{\text{exp}}$ . Thus  $\pi(\rho) = \rho$  if  $\rho \in \mathcal{R}_m^{\text{exp}}$ . However, the range of  $\pi$  is  $\pi(\mathcal{R}_m) = \mathbf{r}(\mathcal{A}_m) = \mathcal{R}_m^{\text{exp}}$ . Thus if  $\rho \notin \mathcal{R}_m^{\text{exp}}$ , then  $\pi(\rho)$  cannot equal  $\rho$ .
4. Let  $\bar{\rho} \in \mathcal{R}_m^{\text{exp}}$ , let  $O \subset \mathcal{R}_m$  be an open set containing  $\bar{\rho}$ , and let  $\bar{\alpha} = \mathbf{a}(\bar{\rho})$ . Choose any  $\rho \in \{\bar{\rho} +_N \mathcal{NC}(A_{m_N}, \bar{\alpha}_N)\}$ . Then  $\rho = \bar{\rho}$  for  $j < N$ , so by Proposition 20,  $\bar{\alpha}^T \rho = \bar{\alpha}^T \bar{\rho}$  and  $\alpha^T \rho \leq \alpha^T \bar{\rho}$  for all  $\alpha \in \mathcal{A}_m$ . Therefore

$$(150) \quad \psi(\bar{\alpha}, \bar{\rho}) = \psi(\bar{\alpha}, \rho) \leq \psi(\mathbf{a}(\rho), \rho) \leq \psi(\mathbf{a}(\rho), \bar{\rho}) \leq \psi(\bar{\alpha}, \bar{\rho}).$$

Here the equality in (150) follows immediately from the definition of  $\psi$  (75) and the fact that  $\bar{\alpha}^T \rho = \bar{\alpha}^T \bar{\rho}$ . The first inequality in (150) uses the fact that  $\psi(\mathbf{a}(\rho), \rho)$  maximizes  $\psi(\cdot, \rho)$  over all  $\alpha \in \mathcal{A}_m$ ; the second uses the fact that  $\alpha^T \rho \leq \alpha^T \bar{\rho}$  for all  $\alpha \in \mathcal{A}_m$ ; and the third uses the fact that  $\psi(\bar{\alpha}, \bar{\rho})$  maximizes  $\psi(\cdot, \bar{\rho})$  over all  $\alpha \in \mathcal{A}_m$ . We conclude from (150) that  $\psi(\mathbf{a}(\rho), \bar{\rho}) = \psi(\bar{\alpha}, \bar{\rho})$ . Since  $\bar{\alpha}$  is the *unique* maximizer of  $\psi(\cdot, \bar{\rho})$  over all  $\alpha \in \mathcal{A}_m$ , it follows that  $\mathbf{a}(\rho) = \bar{\alpha}$ . Therefore  $\pi(\rho) = \mathbf{r}(\mathbf{a}(\rho)) = \mathbf{r}(\bar{\alpha}) = \bar{\rho}$ .

5. We first argue by contraction to show that  $\pi(\mathcal{D}_m) \subset \mathbf{r}(\mathcal{A}_m \cap \partial \mathcal{A}_m)$ . Thus, suppose there exist  $\rho \in \mathcal{D}_m$  and  $\alpha \in \text{int } \mathcal{A}_m$  such that  $\pi(\rho) = \mathbf{r}(\alpha)$ . We know that

$$(151) \quad \psi(\pi(\rho), \mathbf{a}(\pi(\rho))) = \max_{\alpha \in \mathcal{A}_m} \psi(\pi(\rho), \alpha),$$

and since  $\psi$  is differentiable on  $\text{int } \mathcal{A}_{\mathbf{m}}$ , first order optimality conditions imply that

$$(152) \quad \frac{\partial \psi}{\partial \alpha}(\pi(\rho), \mathbf{a}(\pi(\rho))) = \rho - \pi(\pi(\rho)) = 0.$$

However,  $\pi$  is a projection; therefore, (152) implies that  $\rho = \pi(\rho)$ . According to (149c), this contradicts the assumption that  $\rho \in \mathcal{D}_{\mathbf{m}}$ .

We next show that  $\mathbf{r}(\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}) \subset \pi(\mathcal{D}_{\mathbf{m}})$ . Let  $\bar{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ , and let  $O \subset \mathcal{R}_{\mathbf{m}}$  be an open set containing  $\mathbf{r}(\bar{\alpha}) \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ . Then choose (see (32))

$$(153) \quad \rho \in \{\mathbf{r}(\bar{\alpha}) +_N \mathcal{NC}_0(\mathcal{A}_{\mathbf{m}_N}, \bar{\alpha}_N)\} \cap O.$$

Since  $\bar{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ ,  $\bar{\alpha}_N \in \partial \mathcal{A}_{\mathbf{m}_N}$  (see (46)), and this set is nonempty. By (149d),  $\pi(\rho) = \mathbf{r}(\bar{\alpha})$ . Thus we need only show that  $\rho \in \mathcal{D}_{\mathbf{m}}$ . If it is not, then  $\rho \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  and  $\pi(\rho) = \rho = \mathbf{r}(\bar{\alpha})$ , which contradicts (153). Thus  $\rho \in \mathcal{D}_{\mathbf{m}}$ .

Finally, we show that  $\mathbf{r}(\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}) = \mathcal{R}_{\mathbf{m}}^{\text{exp}} \cap \partial \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ . Because  $\mathbf{r}$  is one-to-one on  $\mathcal{A}_{\mathbf{m}}$  (Theorem 17),

$$(154) \quad \mathbf{r}(\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}) = \mathbf{r}(\mathcal{A}_{\mathbf{m}}) \setminus \mathbf{r}(\text{int } \mathcal{A}_{\mathbf{m}}) = \mathcal{R}_{\mathbf{m}}^{\text{exp}} \setminus \text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}} = \mathcal{R}_{\mathbf{m}}^{\text{exp}} \cap \partial \mathcal{R}_{\mathbf{m}}^{\text{exp}}.$$

6. Given that  $\mathbf{a} \circ \mathbf{r}$  is the identity map on  $\mathcal{A}_{\mathbf{m}}$  (Theorem 17),  $\mathbf{a}(\pi(\rho)) = (\mathbf{a} \circ \mathbf{r})(\mathbf{a}(\rho)) = \mathbf{a}(\rho)$ .

7. The proof is a simple calculation. For any  $\rho \in \mathcal{R}_{\mathbf{m}}$ , (149a), (149b), and (149f) give

$$(155) \quad h(\pi(\rho)) = \psi(\mathbf{a}(\pi(\rho)), \pi(\rho)) = \psi(\mathbf{a}(\pi(\rho)), \rho) = \psi(\mathbf{a}(\rho), \rho) = h(\rho). \quad \square$$

**COROLLARY 22.** *The set  $\mathcal{D}_{\mathbf{m}}$  is empty if and only if  $\mathcal{A}_{\mathbf{m}}$  is open. If  $\mathcal{D}_{\mathbf{m}}$  is nonempty, then  $h$  fails to be strictly convex.*

*Proof.* The first statement is an immediate consequence of (149e). The second statement is a consequence of (149d) and (149g), which together imply that  $h$  is constant on the cone  $\{\bar{\rho} +_N \mathcal{NC}(\mathcal{A}_{\mathbf{m}_N}, \mathbf{a}_N(\bar{\rho}))\}$  for any  $\bar{\rho} \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ . If  $\mathcal{D}_{\mathbf{m}}$  is nonempty, then by (149e),  $\mathcal{R}_{\mathbf{m}}^{\text{exp}} \cap \partial \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  is also nonempty; if  $\bar{\rho} \in \mathcal{R}_{\mathbf{m}}^{\text{exp}} \cap \partial \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ , then  $\mathbf{a}(\bar{\rho}) \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$  and, consequently,  $\mathbf{a}_N(\bar{\rho}) \in \partial \mathcal{A}_{\mathbf{m}_N}$  (see (46)). As a result,  $\{\bar{\rho} +_N \mathcal{NC}(\mathcal{A}_{\mathbf{m}_N}, \mathbf{a}(\bar{\rho}))\}$  is nontrivial, and  $h$  cannot be strictly convex on all of  $\mathcal{R}_{\mathbf{m}}$ .  $\square$

It turns out that  $\mathcal{A}_{\mathbf{m}}$  is open only for  $N = 2$ . (To see this fact, one need only realize that, for  $N > 2$ , the vector  $\alpha \in \mathcal{A}_{\mathbf{m}}$  corresponding to any Maxwellian  $\mathcal{M}_{\rho, u, \theta}$  lies on the boundary  $\partial \mathcal{A}_{\mathbf{m}}$ .) Thus Corollary 22 shows that the Maxwellian and Gaussian closures are the exception rather than the rule. However, in spite of the difficulties encountered for  $\alpha \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ , (124) and (125) extend to all of  $\mathcal{R}_{\mathbf{m}}$ .

**THEOREM 23.** *For all  $\rho \in \mathcal{R}_{\mathbf{m}}$ ,*

$$(156) \quad h(\rho) + h^*(\mathbf{a}(\rho)) = \mathbf{a}(\rho)^T \rho,$$

*and the function  $\mathbf{a}$  is the continuous Fréchet derivative of  $h$  everywhere on  $\mathcal{R}_{\mathbf{m}}$ , i.e.,*

$$(157) \quad \mathbf{a}(\rho) = h_{\rho}(\rho).$$

*Proof.* Let  $\rho \in \mathcal{R}_{\mathbf{m}}$ , and set  $\bar{\rho} = \pi(\rho) \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ . By (124),

$$(158) \quad h(\pi(\rho)) + h^*(\mathbf{a}(\pi(\rho))) = \mathbf{a}(\pi(\rho))^T \pi(\rho).$$



However, according to Lemma 21,  $\mathbf{a}(\boldsymbol{\rho}) = \mathbf{a}(\bar{\boldsymbol{\rho}})$ ,  $h(\boldsymbol{\rho}) = h(\bar{\boldsymbol{\rho}})$ , and  $\mathbf{a}(\boldsymbol{\rho})^T \boldsymbol{\rho} = \mathbf{a}(\boldsymbol{\rho})^T \bar{\boldsymbol{\rho}}$ . Therefore (158) and (156) are equivalent.

We now move on to proving (157). By using (116), we find that

$$(159) \quad \begin{aligned} h(\boldsymbol{\rho} + \boldsymbol{\delta}) &= \psi(\mathbf{a}(\boldsymbol{\rho} + \boldsymbol{\delta}), \boldsymbol{\rho} + \boldsymbol{\delta}) \geq \psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho} + \boldsymbol{\delta}) \\ &= \psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}) + \mathbf{a}(\boldsymbol{\rho})^T \boldsymbol{\delta} = h(\boldsymbol{\rho}) + \mathbf{a}(\boldsymbol{\rho})^T \boldsymbol{\delta} \end{aligned}$$

and, similarly, that

$$(160) \quad \begin{aligned} h(\boldsymbol{\rho}) &= \psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}) \geq \psi(\mathbf{a}(\boldsymbol{\rho} + \boldsymbol{\delta}), \boldsymbol{\rho}) = \psi(\mathbf{a}(\boldsymbol{\rho} + \boldsymbol{\delta}), \boldsymbol{\rho} + \boldsymbol{\delta}) - [\mathbf{a}(\boldsymbol{\rho} + \boldsymbol{\delta})]^T \boldsymbol{\delta} \\ &= h(\boldsymbol{\rho} + \boldsymbol{\delta}) - [\mathbf{a}(\boldsymbol{\rho} + \boldsymbol{\delta})]^T \boldsymbol{\delta}. \end{aligned}$$

Together (159) and (160) imply that

$$(161) \quad 0 \leq h(\boldsymbol{\rho} + \boldsymbol{\delta}) - h(\boldsymbol{\rho}) - \mathbf{a}(\boldsymbol{\rho})^T \boldsymbol{\delta} \leq |\boldsymbol{\delta}| |\mathbf{a}(\boldsymbol{\rho} + \boldsymbol{\delta}) - \mathbf{a}(\boldsymbol{\rho})|.$$

Hence, to complete the proof, we need only to show that  $\mathbf{a}$  is continuous.

Equation (159) implies also that  $\mathbf{a}(\boldsymbol{\rho})$  is a *subgradient* of  $h$  at  $\boldsymbol{\rho}$  [30, section 23, page 214]. The set of all subgradients is called the *subdifferential* of  $h$  at  $\boldsymbol{\rho}$  and is denoted by  $\partial h(\boldsymbol{\rho})$ . It is a general result from convex analysis [30, Theorem 24.7] that, because  $h$  is convex, the set  $\partial h(S) \equiv \bigcup_{\boldsymbol{\rho} \in K} \partial h(\boldsymbol{\rho})$  is bounded whenever  $K \subset \mathbb{R}^n$  is bounded. In particular, if  $\{\boldsymbol{\rho}_{(i)}\}_{i=1}^\infty \subset \mathcal{R}_{\mathbf{m}}$  converges to  $\boldsymbol{\rho}_* \in \mathcal{R}_{\mathbf{m}}$ , then  $\{\mathbf{a}(\boldsymbol{\rho}_{(i)})\}_{i=1}^\infty$  is a bounded sequence. Let  $\boldsymbol{\alpha}_*$  be any subsequential limit for this sequence. Then

$$(162) \quad \begin{aligned} \psi(\mathbf{a}(\boldsymbol{\rho}_*), \boldsymbol{\rho}_*) &= \lim_{i \rightarrow \infty} \psi(\mathbf{a}(\boldsymbol{\rho}_*), \boldsymbol{\rho}_{(i_k)}) \\ &\leq \lim_{i \rightarrow \infty} \psi(\mathbf{a}(\boldsymbol{\rho}_{(i_k)}), \boldsymbol{\rho}_{(i_k)}) \leq \psi(\boldsymbol{\alpha}_*, \boldsymbol{\rho}_*) \leq \psi(\mathbf{a}(\boldsymbol{\rho}_*), \boldsymbol{\rho}_*), \end{aligned}$$

where  $\{i_k\}_{k=1}^\infty$  is any sequence of integers such that  $\boldsymbol{\alpha}_* = \lim_{i \rightarrow \infty} \mathbf{a}(\boldsymbol{\rho}_{(i_k)})$ . The first and last inequalities in (162) follow because  $\psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho})$  maximizes  $\psi(\cdot, \boldsymbol{\rho})$ , whereas the middle inequality is a consequence of the fact that  $\psi(\cdot, \boldsymbol{\rho})$  is upper semicontinuous (Theorem 11).

From (162), we deduce that  $\psi(\boldsymbol{\alpha}_*, \boldsymbol{\rho}) = \psi(\mathbf{a}(\boldsymbol{\rho}_*), \boldsymbol{\rho})$ , and, since  $\mathbf{a}(\boldsymbol{\rho}_*)$  is the *unique* minimizer of  $\psi(\cdot, \boldsymbol{\rho})$ , it follows that  $\boldsymbol{\alpha}_* = \mathbf{a}(\boldsymbol{\rho}_*)$ . Because  $\{\mathbf{a}(\boldsymbol{\rho}_{(i)})\}$  is bounded and all of its converging subsequences converge to  $\mathbf{a}(\boldsymbol{\rho}_*)$ , it follows then that

$$(163) \quad \lim_{i \rightarrow \infty} \mathbf{a}(\boldsymbol{\rho}_{(i)}) = \mathbf{a}(\boldsymbol{\rho}_*).$$

Thus  $\mathbf{a}$  is continuous, and  $h$  is continuously differentiable.  $\square$

Note that, as a consequence of Theorem 23,  $h(\boldsymbol{\rho}) = \psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho})$  is a differentiable on all of  $\mathcal{R}_{\mathbf{m}}$  even though  $\psi(\cdot, \boldsymbol{\rho})$  may not be continuous for  $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ . We alluded to this fact earlier in section 4.2.

**6. Geometry of  $\mathcal{D}_{\mathbf{m}}$ .** In this section, we give a description of the geometry of the set  $\mathcal{D}$ . The main results are given in Theorem 25, which shows that  $\mathcal{D}$  is a union of cones, and in Theorem 28, which concludes that, with additional assumptions,  $\mathcal{D}$  is small in both a topological and a measure-theoretic sense. We begin with some motivation for why such results are important.

**6.1. Motivation: Behavior of the closure near degeneracy.** Even though  $\mathcal{D}_{\mathbf{m}}$  is usually nonempty, there is evidence to suggest that if  $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  initially, then

densities in  $\mathcal{D}_{\mathbf{m}}$  might never be attained. To investigate this possibility, we introduce the function  $\chi : \mathcal{R}_{\mathbf{m}} \rightarrow \mathbb{R}$ , defined by

$$(164) \quad \chi(\boldsymbol{\rho}) \equiv \int_{\mathbb{R}^d} |v \mathbf{m}(v)| G_{\mathbf{a}(\boldsymbol{\rho})}(v) dv.$$

For the entropy-based closure,  $\chi$  is closely related to the flux  $\mathbf{f}$  in (13), and we show below that  $\chi$  becomes unbounded as  $\boldsymbol{\rho}$  approaches  $\mathcal{D}_{\mathbf{m}}$ . As pointed out in [20], such divergent behavior raises the possibility that  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$  is invariant under the dynamics of the closure.

**PROPOSITION 24.** *Let  $\{\boldsymbol{\rho}_{(j)}\}_{j=1}^{\infty}$  be a sequence in  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$  such that  $\boldsymbol{\rho}_{(j)} \rightarrow \boldsymbol{\rho}_* \in \mathcal{D}_{\mathbf{m}}$ , and for each  $j$ , let  $\chi_j \equiv \chi(\boldsymbol{\rho}_{(j)})$ . Then  $\{\chi_j\}_{j=1}^{\infty}$  is unbounded.*

*Proof.* Since  $\{\boldsymbol{\rho}_{(j)}\}_{j=1}^{\infty} \subset \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ ,

$$(165) \quad \mathbf{r}(\mathbf{a}(\boldsymbol{\rho}_{(j)})) = \boldsymbol{\rho}_{(j)}, \quad j = 1, 2, \dots,$$

and taking limits on both sides gives

$$(166) \quad \lim_{j \rightarrow \infty} \mathbf{r}(\mathbf{a}(\boldsymbol{\rho}_{(j)})) = \boldsymbol{\rho}_*.$$

We proceed by showing that if  $\{\chi_j\}_{j=1}^{\infty}$  is bounded, then

$$(167) \quad \lim_{j \rightarrow \infty} \mathbf{r}(\mathbf{a}(\boldsymbol{\rho}_{(j)})) = \mathbf{r}(\mathbf{a}(\boldsymbol{\rho}_*)).$$

Together (166)–(167) will then imply that  $\boldsymbol{\rho}_* \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  which, by contradicting our hypothesis, proves the claim. Hence, suppose that  $\{\chi_j\}_{j=1}^{\infty}$  is bounded. To conclude (167), we calculate

$$(168) \quad \begin{aligned} \left| \mathbf{r}(\mathbf{a}(\boldsymbol{\rho}_*)) - \mathbf{r}(\mathbf{a}(\boldsymbol{\rho}_{(j)})) \right| &= \left| \langle \mathbf{m} G_{\mathbf{a}(\boldsymbol{\rho}_*)} \rangle - \langle \mathbf{m} G_{\mathbf{a}(\boldsymbol{\rho}_{(j)})} \rangle \right| \\ &\leq \int_{\mathbb{R}^d} |\mathbf{m}(v)| \left| G_{\mathbf{a}(\boldsymbol{\rho}_*)}(v) - G_{\mathbf{a}(\boldsymbol{\rho}_{(j)})}(v) \right| dv \\ &= \int_{|v| > R} |\mathbf{m}(v)| \left| G_{\mathbf{a}(\boldsymbol{\rho}_*)}(v) - G_{\mathbf{a}(\boldsymbol{\rho}_{(j)})}(v) \right| dv \end{aligned}$$

$$(169) \quad + \int_{|v| < R} |\mathbf{m}(v)| \left| G_{\mathbf{a}(\boldsymbol{\rho}_*)}(v) - G_{\mathbf{a}(\boldsymbol{\rho}_{(j)})}(v) \right| dv,$$

where  $R > 0$  is an arbitrary constant. We handle the integrals for  $|v| > R$  and  $|v| < R$  in (168) separately. For  $|v| > R$ ,

$$(170) \quad \begin{aligned} &\int_{|v| > R} |\mathbf{m}(v)| \left| G_{\mathbf{a}(\boldsymbol{\rho}_*)}(v) - G_{\mathbf{a}(\boldsymbol{\rho}_{(j)})}(v) \right| dv \\ &\leq \int_{|v| > R} \frac{|v \mathbf{m}(v)|}{R} \left| G_{\mathbf{a}(\boldsymbol{\rho}_*)}(v) - G_{\mathbf{a}(\boldsymbol{\rho}_{(j)})}(v) \right| dv \\ &\leq \frac{1}{R} \int_{\mathbb{R}^d} |v \mathbf{m}(v)| \left| G_{\mathbf{a}(\boldsymbol{\rho}_*)}(v) - G_{\mathbf{a}(\boldsymbol{\rho}_{(j)})}(v) \right| dv \leq \frac{C}{R}, \end{aligned}$$

where

$$(171) \quad C \equiv 2 \max \left\{ \chi(\boldsymbol{\rho}_*), \sup_j \{\chi_j\} \right\}.$$

For  $|v| < R$ , continuity of  $\mathbf{a}$  (see Theorem 23) implies that  $\mathbf{a}(\boldsymbol{\rho}_{(j)}) \rightarrow \mathbf{a}(\boldsymbol{\rho}_*)$ . Hence the sequence  $G_{\mathbf{a}(\boldsymbol{\rho}_{(j)})}$  is uniformly bounded on  $\{v \in \mathbb{R}^d : |v| \leq R\}$ . By the Lebesgue bounded convergence theorem,

$$(172) \quad \lim_{j \rightarrow \infty} \int_{|v| < R} |\mathbf{m}(v)| G_{\mathbf{a}(\boldsymbol{\rho}_j)}(v) dv = \int_{|v| < R} |\mathbf{m}(v)| G_{\mathbf{a}(\boldsymbol{\rho}_*)}(v) dv.$$

Together (168), (170), and (172) imply that

$$(173) \quad \lim_{j \rightarrow \infty} |\mathbf{r}(\mathbf{a}(\boldsymbol{\rho}_*)) - \mathbf{r}(\mathbf{a}(\boldsymbol{\rho}_j))| \leq \frac{C}{R}.$$

Because  $R$  can be arbitrarily large, we conclude that (167) holds, which proves the claim.  $\square$

Note that, by uniformly bounding  $\chi_j$  in the proof above, we are providing uniform control on the highest-order moments in  $\boldsymbol{\rho}_{(j)}$ . In general, such control is not possible, which is why the minimizer in (14) with equality constraints does not always exist. (See the discussion following the proof of Theorem 3.)

The behavior of  $\chi$  expressed in Proposition 24 was first observed by Junk for the one-dimensional example in [20]. In particular, for a sequence  $\{\boldsymbol{\rho}_{(j)}\}_{j=1}^\infty \in \text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ , it was found that  $\langle v \mathbf{m}_N G_{\mathbf{a}(\boldsymbol{\rho})} \rangle$  diverges to either positive or negative infinity as  $\boldsymbol{\rho}_{(j)} \rightarrow \boldsymbol{\rho}_* \in \mathcal{D}_{\mathbf{m}}$ , with the sign depending on the direction of approach.

Suppose now that it can be proven that  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$  is invariant under the dynamics of the balance law (12) with the entropy-based closure. Then if  $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  initially, the entropy minimization problem with equality constraints (14) will always have a solution, and the formal properties of the closure based on the Legendre duality between  $h$  and  $h^*$  will be maintained. However, it must be shown—at a minimum—that  $\mathcal{D}_{\mathbf{m}}$  is small in some sense, thereby limiting the number of initial conditions in  $\mathcal{R}_{\mathbf{m}}$  which must be discarded in order to maintain a well-defined closure. In the following subsections, we use the complementary slackness conditions (92) to show that, under reasonable hypotheses,  $\mathcal{D}_{\mathbf{m}}$  is indeed a Lebesgue measure zero set.

**6.2. The complementary slackness condition and normal cones.** From the complementary slackness condition (149b), we obtain the following result.

**THEOREM 25.** *The set  $\mathcal{R}_{\mathbf{m}}$  can be expressed as the following union of cones:*

$$(174) \quad \mathcal{R}_{\mathbf{m}} = \bigcup_{\bar{\boldsymbol{\rho}} \in \mathcal{R}_{\mathbf{m}}^{\text{exp}}} \bar{\boldsymbol{\rho}} +_N \mathcal{NC}(A_{\mathbf{m}_N}, \mathbf{a}_N(\bar{\boldsymbol{\rho}})).$$

The proof of this theorem uses the following lemma.

**LEMMA 26.** *Let  $\mathbf{m}$  be a vector whose polynomial components form the basis for an admissible space  $\mathbb{M}$ . Then  $A_{\mathbf{m}} \subset \{\boldsymbol{\alpha} \in \mathbb{R}^n : \boldsymbol{\alpha}_N \in A_{\mathbf{m}_N}\}$ , where  $A_{\mathbf{m}}$  and  $A_{\mathbf{m}_N}$  are defined in (36) and (44), respectively.*

*Proof.* Let  $\boldsymbol{\alpha} \in A_{\mathbf{m}}$ , and let  $v_* \in \mathbb{R}^d$  be fixed. Because the components of  $\mathbf{m}_i$  are homogeneous polynomials of degree  $i$ , for any  $\lambda > 0$ ,

$$(175) \quad 0 \geq \frac{1}{\lambda^N} \boldsymbol{\alpha}^T \mathbf{m}(\lambda v_*) = \sum_{i=0}^N \frac{\lambda^i}{\lambda^N} \boldsymbol{\alpha}_i^T \mathbf{m}_i(v_*).$$

Taking the limit  $\lambda \rightarrow \infty$  in (175) gives  $\boldsymbol{\alpha}_N^T \mathbf{m}_N(v_*) \leq 0$ , and since  $v_*$  is arbitrary, we conclude that  $\boldsymbol{\alpha}_N \in A_{\mathbf{m}_N}$ .  $\square$

*Proof of Theorem 25.* Suppose that  $\rho \in \mathbb{R}^n$  and that  $\bar{\rho} \in \mathcal{R}_m^{\text{exp}}$ . Before making any further assumptions about  $\rho$  or any relationship between  $\rho$  and  $\bar{\rho}$ , we note that from Proposition 20 we obtain the following set of equivalent statements:

$$(176a) \quad (i) \quad \rho_N - \bar{\rho}_N \in \mathcal{NC}(A_{m_N}, \mathbf{a}_N(\bar{\rho}));$$

$$(176b) \quad (ii) \quad (\alpha_N - \mathbf{a}_N(\bar{\rho}))^T (\rho_N - \bar{\rho}_N) \leq 0 \quad \forall \alpha_N \in A_{m_N};$$

$$(176c) \quad (iii) \quad \mathbf{a}_N(\bar{\rho})^T (\rho_N - \bar{\rho}_N) = 0 \quad \text{and} \quad \alpha_N^T (\rho_N - \bar{\rho}_N) \leq 0 \quad \forall \alpha_N \in A_{m_N}.$$

We first show containment of the left-hand side of (174). Given  $\rho \in \mathcal{R}_m$ , let  $\bar{\rho} = \pi(\rho) \in \mathcal{R}_m^{\text{exp}}$ . Then  $\rho_j = \bar{\rho}_j$  for  $j < N$  and, by (149f),  $\mathbf{a}(\bar{\rho}) = \mathbf{a}(\rho)$ . Thus, from (149b),

$$(177) \quad \mathbf{a}_N(\bar{\rho})^T \rho_N = \mathbf{a}_N(\bar{\rho})^T \bar{\rho}_N.$$

Meanwhile, the constraint conditions in (52) imply that

$$(178) \quad \alpha_N^T \rho_N \leq \alpha_N^T \bar{\rho}_N \quad \forall \alpha \in A_m.$$

We conclude from (176)–(178) that  $\rho_N - \bar{\rho}_N \in \mathcal{NC}(A_{m_N}, \mathbf{a}_N(\bar{\rho}))$ .

Next we show containment in the other direction. Suppose that  $\rho_j = \bar{\rho}_j$  for  $j < N$  and that  $\rho_N - \bar{\rho}_N \in \mathcal{NC}(A_{m_N}, \mathbf{a}_N(\bar{\rho}))$ . By Theorem 1,  $\mathcal{R}_m = \text{int } A_m^\circ$ , so it is sufficient to prove that  $\rho \in \text{int } A_m^\circ$ . Because  $\bar{\rho} \in \mathcal{R}_m^{\text{exp}} \subset \mathcal{R}_m = \text{int } A_m^\circ$ , it follows that  $\alpha^T \bar{\rho} < 0$  for all  $\alpha \in A_m$ . Furthermore, by Lemma 26,  $\alpha_N \in A_{m_N}$  for all such  $\alpha$ . Hence from (178),

$$(179) \quad \alpha^T \rho = \alpha^T \bar{\rho} + \alpha_N^T (\rho_N - \bar{\rho}_N) < 0 \quad \forall \alpha \in A_m.$$

This shows that  $\rho \in \text{int } A_m^\circ$  and concludes the proof.  $\square$

For  $\bar{\rho} \in \text{int } \mathcal{R}_m^{\text{exp}}$ ,  $\mathcal{NC}(A_{m_N}, \mathbf{a}_N(\bar{\rho}))$  is just the origin in  $\mathbb{R}^{n_N}$ . In such cases, Theorem 25 is trivial, and the construction  $\bar{\rho} +_N \mathcal{NC}(A_{m_N}, \mathbf{a}_N(\bar{\rho}))$  does not generate any new densities. Therefore  $\mathcal{D}_m$  is constructed entirely by convex cones attached to  $\bar{\rho} \in \mathcal{R}_m^{\text{exp}} \cap \partial \mathcal{R}_m^{\text{exp}}$ . Recall from (32) that  $\mathcal{NC}_0(A_{m_N}, \mathbf{a}_N(\bar{\rho})) = \mathcal{NC}(A_{m_N}, \mathbf{a}_N(\bar{\rho})) \setminus \{0\}$ . We have the following corollary.

**COROLLARY 27.** *The degenerate densities are*

$$(180) \quad \mathcal{D}_m = \bigcup_{\bar{\rho} \in \mathcal{R}_m^{\text{exp}} \cap \partial \mathcal{R}_m^{\text{exp}}} \{\bar{\rho} +_N \mathcal{NC}_0(A_{m_N}, \mathbf{a}_N(\bar{\rho}))\} = \bigcup_{\bar{\alpha} \in \mathcal{A}_m \cap \partial \mathcal{A}_m} \{\mathbf{r}(\bar{\alpha}) +_N \mathcal{NC}_0(A_{m_N}, \bar{\alpha}_N)\}.$$

**6.3. Smoothness assumptions on  $\mathcal{A}_m \cap \partial \mathcal{A}_m$ .** Corollary 27 gives the degenerate densities associated with each  $\bar{\rho} \in \mathcal{R}_m^{\text{exp}} \cap \partial \mathcal{R}_m^{\text{exp}}$ . However, a clean description of  $\mathcal{D}_m$  requires also that  $\mathcal{R}_m^{\text{exp}} \cap \partial \mathcal{R}_m^{\text{exp}}$  itself have a nice structure. In particular, we would like to say that  $\mathcal{R}_m^{\text{exp}} \cap \partial \mathcal{R}_m^{\text{exp}}$  is a finite union of disjoint manifolds. At this point we are unable to prove such a result in general, in part due to the complicated structure of  $\mathcal{A}_m \cap \partial \mathcal{A}_m$  to which we alluded in section 2.4. We therefore make two assumptions. The first assumption says that  $\mathcal{A}_m \cap \partial \mathcal{A}_m$  is a union of disjoint manifolds with dimensional restrictions that are related to the dimensions of the normal cones in (180) in such a way as to ensure that  $\mathcal{D}_m$  is a lower-dimensional subset of  $\mathcal{R}_m$ . The second assumption says that the mapping  $\mathbf{r}$  is diffeomorphic when restricted to each of these manifolds. Thus each dimension  $k$  manifold in  $\mathcal{A}_m \cap \partial \mathcal{A}_m$  will map to

a dimension  $k$  manifold in  $\mathcal{R}_{\mathbf{m}}^{\text{exp}} \cap \partial\mathcal{R}_{\mathbf{m}}^{\text{exp}}$ . Before stating our assumptions, we define the orthogonal projections  $\mathcal{P}_N : \mathbb{R}^n \mapsto \mathbb{R}^{n_N}$  and  $\mathcal{P}_{\tilde{N}} : \mathbb{R}^n \mapsto \mathbb{R}^{n-n_N}$  by

$$(181) \quad \mathcal{P}_N(\boldsymbol{\alpha}) \equiv (0, \dots, 0, 0, \boldsymbol{\alpha}_N^T)^T \quad \text{and} \quad \mathcal{P}_{\tilde{N}}(\boldsymbol{\alpha}) \equiv \boldsymbol{\alpha} - \mathcal{P}_N(\boldsymbol{\alpha}) = (\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T, \dots, \boldsymbol{\alpha}_{N-1}^T, 0)^T.$$

*Assumption I.* The vector  $\mathbf{m}$  is such that the set  $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$  can be decomposed into a finite collection  $\mathcal{S}$  of disjoint, smooth ( $C^\infty$ ) manifolds in  $\mathbb{R}^n$ . Furthermore, if  $S$  is one such manifold, then  $\mathcal{P}_N$  projects  $S$  onto a manifold  $S_N \subset \partial\mathcal{A}_{\mathbf{m}_N}$  with codimension at least one in  $\mathbb{R}^{n_N}$  and  $\mathcal{P}_{\tilde{N}}$  projects  $S$  onto a manifold  $S_{\tilde{N}}$  of codimension at least one in  $\mathbb{R}^{n-n_N}$ .

We call  $\mathcal{S}$  a *stratification* of  $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ ; the manifolds  $S$  that make up  $\mathcal{S}$  are called *strata*. We fully expect that  $\mathcal{S}$  can be chosen so that, for each  $S \in \mathcal{S}$ , the projection  $S_N$  is indeed a manifold. If so,  $S_N$  will certainly have codimension of one or more since, by (46),  $S_N \subset \partial\mathcal{A}_{\mathbf{m}_N}$ . Furthermore, if  $\boldsymbol{\alpha}_N \in \partial\mathcal{A}_{\mathbf{m}_N}$ , then  $\boldsymbol{\alpha}_N^T \mathbf{m}_N(\lambda\omega) = 0$  for some  $\omega \in \mathbb{S}^{d-1}$  and all  $\lambda \in \mathbb{R}$ , which means that  $\mathbf{m}_N$  no longer provides uniform control over lower-degree polynomials. Thus, in order to maintain the integrability condition (40) that defines  $\mathcal{A}_{\mathbf{m}}$ , we expect further restrictions on the components  $\boldsymbol{\alpha}_j$  for  $j < N$ . This is the motivation for the codimension one restriction on the manifold  $S_{\tilde{N}}$  in Assumption I. Since, in general,  $\dim(S) \leq \dim(S_N) + \dim(S_{\tilde{N}})$ , these restrictions together imply that  $S$  itself has codimension of at least two in  $\mathbb{R}^n$ .

It should be noted that Assumption I is known to hold for at least two cases:

$$(182a) \quad (i) \quad d = 1 \text{ and } N \geq 2;$$

$$(182b) \quad (ii) \quad d > 1, \quad N = 4, \text{ and } \mathbf{m}_4 = |v|^4.$$

(Whether or not Assumption I holds in any other case is, to our knowledge, an open question.) For the first case above,  $\boldsymbol{\alpha}_j = \alpha_j$  and  $n = N + 1$ . For  $i = 1, \dots, N/2$ , we define the sets

$$(183) \quad \mathcal{A}_{\mathbf{m}}^{2i} = \{\boldsymbol{\alpha} \in \mathbb{R}^n : \alpha_j = 0 \text{ for } 2i < j \leq N \text{ and } \alpha_{2i} < 0\}.$$

Clearly each  $\mathcal{A}_{\mathbf{m}}^{2i}$  is a manifold of dimension  $2i + 1$  such that

$$(184) \quad \mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}} = \bigcup_{i=1}^{N/2-1} \mathcal{A}_{\mathbf{m}}^{2i} \quad \text{and} \quad \mathcal{A}_{\mathbf{m}}^N = \text{int } \mathcal{A}_{\mathbf{m}}.$$

For the second case,  $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}} = \{\boldsymbol{\alpha} \in \mathbb{R}^n : \boldsymbol{\alpha}_2^T \mathbf{m}_2 < 0\}$ . If  $\mathbf{m}_2 = |v|^2$ , then  $G_{\boldsymbol{\alpha}}$  has the form of a Maxwellian distribution (133) on  $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ ; if  $\mathbf{m}_2 = v \vee v$ , then  $G_{\boldsymbol{\alpha}}$  has the form of a Gaussian distribution (138) on  $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ .

One possible way to prove that Assumption I always holds is to show that the integrability condition which defines  $\mathcal{A}_{\mathbf{m}}$  can be expressed as a family of polynomial equalities and inequalities for  $\boldsymbol{\alpha}$ . Sets expressed in this way are called *semialgebraic* and are known to have a stratification with special properties [2, 27]. One can show, for example, that the sets  $\mathcal{A}_{\mathbf{m}_j}$  ( $j$  even) and  $\text{cl } \mathcal{A}_{\mathbf{m}}$  are semialgebraic. One can also show that the interiors and boundaries of these sets are semialgebraic. See [17] for details.

*Assumption II.* The vector  $\mathbf{m}$  is such that if Assumption I holds and if  $S$  is an element of the stratification of  $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ , then for each  $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ , the restriction of  $\psi(\cdot, \boldsymbol{\rho})$  to  $S$  is infinitely Fréchet differentiable on  $S$ .

One may easily verify that Assumption II also holds for the cases in (182). When both Assumptions I and II hold,  $\mathbf{r}$  is a smooth diffeomorphism with inverse  $\mathbf{a}$  when restricted to any manifold in the stratification of  $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ .

**6.4. Fiber bundles.** The construction of  $\mathcal{D}_{\mathbf{m}}$  by attaching cones to the densities  $\mathcal{R}_{\mathbf{m}}^{\text{exp}} \cap \partial \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  is very similar to the construction of a fiber bundle. A (*continuous*) *fiber bundle*  $(\mathcal{B}, B, F, \mathcal{P})$  [18] consists of topological spaces  $\mathcal{B}$ ,  $B$ , and  $F$  along with a projection  $\mathcal{P} : \mathcal{B} \rightarrow B$  such that, for every  $y \in B$ , there is a neighborhood  $O \subset B$  containing  $y$  such that  $\mathcal{P}^{-1}(O)$  is homeomorphic to  $O \times F$ . In addition, if  $\phi$  is this homeomorphism and  $\Pi$  is the natural projection of  $O \times F$  onto  $O$  (i.e.,  $\Pi(y \times F) = y$  for all  $y \in O$ ), then  $\Pi(\phi(\mathcal{P}^{-1}))$  is the identity on  $O$ . The space  $B$  is called the *base space*,  $F$  is called the *fiber space*, and often  $\mathcal{B}$  itself is called the bundle. Roughly speaking,  $\mathcal{B}$  is constructed by attaching to each point in  $B$  a (topologically equivalent) copy of  $F$  that varies continuously from point to point in the base space. If Assumptions I and II hold, then for each manifold  $S$  in a stratification  $\mathcal{S}$  of  $\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ , the manifold  $\mathbf{r}(S)$  acts like a base space; the cones  $\mathcal{NC}(A_{\mathbf{m}_N}, \alpha_N)$ ,  $\alpha \in S$ , are like fibers; and  $\pi$  is the projection onto the base space. The entire structure is

$$(185) \quad \mathcal{B}(S) = \bigcup_{\alpha \in S} \{\mathbf{r}(\alpha) +_N \mathcal{NC}(A_{\mathbf{m}_N}, \alpha_N)\},$$

and, in view of Corollary 27,  $\mathcal{D}_{\mathbf{m}} = \bigcup_{S \in \mathcal{S}} \mathcal{B}_0(S)$ , where

$$(186) \quad \mathcal{B}_0(S) = \mathcal{B}(S) \setminus \mathbf{r}(S) = \bigcup_{\alpha \in S} \{\mathbf{r}(\alpha) +_N \mathcal{NC}_0(A_{\mathbf{m}_N}, \alpha_N)\}.$$

Unfortunately, we cannot conclude that  $\mathcal{B}(S)$  is a bundle even with Assumptions I and II. In short, we have been unable to show a local homeomorphism between the base-fiber product space and the inverse image  $\pi^{-1}(S)$ . However, the sets taken from the examples in section 6.6 below are all fiber bundles. This is fairly easy to check because, in these examples, the convex cones  $A_{\mathbf{m}_N}$  and  $\mathcal{NC}(A_{\mathbf{m}_N}, \alpha_N)$ ,  $\alpha_N \in \partial A_{\mathbf{m}_N}$ , have explicit expressions that are (relatively) simple.

**6.5. Smallness of  $\mathcal{D}_{\mathbf{m}}$ .** If Assumptions I and II hold, we can show that  $\mathcal{D}_{\mathbf{m}}$  is small in the following sense.

**THEOREM 28.** *Suppose that Assumptions I and II hold. Then  $\mathcal{D}_{\mathbf{m}}$  has zero Lebesgue measure,  $\text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  is a dense subset of  $\mathcal{R}_{\mathbf{m}}$ , and  $\mathcal{D}_{\mathbf{m}} \subset \partial \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ .*

*Proof.* The basic idea of the argument is that the image of a smooth map from a lower-dimensional space to a higher-dimensional space has zero Lebesgue measure. We will construct such a map  $F$  whose image covers a portion of  $\mathcal{D}_{\mathbf{m}}$ . We can then cover  $\mathcal{D}_{\mathbf{m}}$  with the images from a countable number of similar maps.

Let  $\mathcal{S}$  be a stratification of  $\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$  as provided by Assumption I, and let  $S \in \mathcal{S}$  have dimension  $j$ . According to Assumption I,  $S_N \equiv \mathcal{P}_N S \subset \partial A_{\mathbf{m}_N}$  and  $S_{\tilde{N}} \equiv \mathcal{P}_{\tilde{N}} S$  are smooth manifolds with dimensions which we denote by  $j_N$  and  $j_{\tilde{N}}$ , respectively. In general,  $\dim(S) \leq \dim(S_N) + \dim(S_{\tilde{N}})$ , and in view of Assumption I,  $\dim(S_{\tilde{N}}) < n - n_N$ . Therefore

$$(187) \quad j \leq j_N + j_{\tilde{N}} < j_N + (n - n_N).$$

This inequality is the key to our result. For any  $\alpha \in S$ , the normal cone  $\mathcal{NC}(S_N, \alpha_N)$  is a subspace of dimension  $n_N - j_N$ , and one can readily show that

$$(188) \quad \mathcal{NC}(A_{\mathbf{m}_N}, \alpha_N) \subset \mathcal{NC}(S_N, \alpha_N), \quad \alpha \in S.$$

We can therefore proceed with the proof by considering the set

$$(189) \quad \mathcal{K}_{\mathbf{m}} \equiv \bigcup_{S \in \mathcal{S}} \bigcup_{\alpha \in S} K(\alpha) \supset (\mathcal{R}_{\mathbf{m}}^{\text{exp}} \cap \partial \mathcal{R}_{\mathbf{m}}^{\text{exp}}) \cup \mathcal{D}_{\mathbf{m}},$$

where the affine spaces

$$(190) \quad K(\alpha) \equiv \mathbf{r}(\alpha) +_N \mathcal{NC}(S_N, \alpha_N), \quad \alpha \in S,$$

are constructed by attaching  $\mathcal{NC}(S_N, \alpha_N)$  to  $\mathbf{r}(\alpha) \in \mathbf{r}(S) \subset \mathcal{R}_{\mathbf{m}}^{\text{exp}} \cap \partial \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ .

Let  $U \subset S$  be the nonempty intersection of  $S$  with a bounded open ball in  $\mathbb{R}^n$ . Because  $S$  is a manifold, there exists a smooth diffeomorphism  $\tau : U \rightarrow \mathbb{R}^j$  such that  $\tau(U)$  is the open unit disk  $\mathbb{D}^j$ . Define a second mapping  $\mathbf{V} : U \rightarrow \mathbb{R}^{n_N \times (n_N - j_N)}$  such that  $\mathbf{V}(\alpha)$  is a matrix whose  $(n_N - j_N)$  columns are vectors in  $\mathbb{R}^{n_N}$  that form a basis for  $\mathcal{NC}(S, \alpha_N)$ . Since  $S$  is smooth, this basis can be chosen to vary smoothly over  $\alpha \in U$ . Then by using  $\tau$  and  $\mathbf{V}$ , define  $F : \mathbb{R}^j \times \mathbb{R}^{n_N - j_N} \rightarrow \mathbb{R}^n$  by

$$(191) \quad F(\mathbf{y}, \mathbf{b}) \equiv \mathbf{r}(\tau^{-1}(\mathbf{y})) +_N \mathbf{V}(\tau^{-1}(\mathbf{y})) \cdot \mathbf{b}.$$

In view of Assumption II,  $F$  is smooth, and by (187),  $j + (n_N - j_N) < n$ . Thus, by [18, Proposition 1.2], the image  $F(\mathbb{D}^j \times \mathbb{R}^{n_N - j_N}) = \bigcup_{\alpha \in U} K(\alpha)$  has zero Lebesgue measure. Because measure is countably subadditive, repeating this argument for each  $j$ -ball  $U$  in a countable cover of  $S$  and then for each  $S \in \mathcal{S}$  shows that  $\mathcal{K}_{\mathbf{m}}$  has zero Lebesgue measure. Since  $\mathcal{D}_{\mathbf{m}} \subset \mathcal{K}_{\mathbf{m}}$ ,  $\mathcal{D}_{\mathbf{m}}$  also has zero Lebesgue measure, and since  $\mathcal{R}_{\mathbf{m}} \setminus \mathcal{K}_{\mathbf{m}} \subset \text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ ,  $\text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$  and  $\mathcal{R}_{\mathbf{m}}$  have the same closure. (Otherwise, there would exist an open set of positive measure contained in  $\mathcal{K}_{\mathbf{m}}$ .) Therefore  $\mathcal{D}_{\mathbf{m}} \subset \partial \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ .  $\square$

**6.6. Examples.** We will assume that Assumptions I and II hold in the following examples.

1. *Junk's example.* The case  $m_N = |v|^N$  has been studied in [20, 21, 31], particularly when  $N = 4$ . For general  $N$ ,

$$(192) \quad \mathcal{A}_{\mathbf{m}_N} = \{\alpha_N \in \mathbb{R} : \alpha_N \leq 0\} \quad \text{and} \quad \partial \mathcal{A}_{\mathbf{m}_N} = \{0\}.$$

If  $\rho \in \mathcal{R}_{\mathbf{m}}$  and  $\mathbf{a}_N(\rho) = 0$ , then  $\mathbf{a}_{N-1}(\rho) = 0$  as well; otherwise,  $G_{\mathbf{a}(\rho)} \notin \mathbb{F}_{\mathbf{m}}$ . With this fact in mind, we conclude from Corollary 15 that  $G_{\mathbf{a}(\rho)}$  is actually the minimizer of  $\mathcal{H}$  subject to fewer constraints:

$$(193) \quad \mathcal{H}(G_{\mathbf{a}(\rho)}) = \min_{g \in \mathbb{F}_{\mathbf{m}}} \{\mathcal{H}(g) : \langle \mathbf{m}_j g \rangle = \rho_j, \ j \leq N-2\}.$$

Let  $\bar{\mathbf{m}}$  contain the components of  $\mathbf{m}$  of degree  $\bar{N} \equiv N-2$  and less:

$$(194) \quad \bar{\mathbf{m}} \equiv (\mathbf{m}_0, \mathbf{m}_1, \dots, \mathbf{m}_{N-2})^T,$$

and let the variables  $\bar{\rho}$  and  $\bar{\alpha}$  and the functions  $\bar{\mathbf{r}}$  and  $\bar{\mathbf{a}}$  be defined similarly. For this example,

$$(195) \quad \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}} \subset \{\alpha \in \mathbb{R}^n : \bar{\alpha} \in \mathcal{A}_{\bar{\mathbf{m}}}, \ \alpha_{N-1} = 0, \ \alpha_N = 0\},$$

but these two sets are not necessarily equal, since the latter may include  $\alpha$  for which  $G_{\alpha} \in \mathbb{F}_{\bar{\mathbf{m}}}$ , but  $G_{\alpha} \notin \mathbb{F}_{\mathbf{m}}$ . However, one may readily conclude that  $G_{\alpha} \in \mathbb{F}_{\mathbf{m}}$  for all  $\bar{\alpha} \in \text{int } \mathcal{A}_{\bar{\mathbf{m}}}$ . Hence,

$$(196) \quad \{\alpha \in \mathbb{R}^n : \bar{\alpha} \in \text{int } \mathcal{A}_{\bar{\mathbf{m}}}, \ \alpha_{N-1} = 0, \ \alpha_N = 0\} \subset \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}.$$

Let  $\mathcal{S}$  be a stratification of  $\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ . The projection of any manifold  $S \in \mathcal{S}$  onto  $\partial \mathcal{A}_{\mathbf{m}_N}$  is the origin in  $\mathbb{R}^{n_N}$ , so the normal cone attached to  $\alpha \in S$  is just the nonnegative axis:

$$(197) \quad \mathcal{NC}(\mathcal{A}_{\mathbf{m}_N}, \alpha_N) = \{\sigma_N \in \mathbb{R} : \sigma_N \geq 0\} = A_{\mathbf{m}_N}^{\circ}.$$

Therefore

$$(198) \quad \mathcal{D}_{\mathbf{m}} = \{\boldsymbol{\rho} : \rho_N > \mathbf{r}_N(\boldsymbol{\alpha}), \boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}\}.$$

Because  $A_{\mathbf{m}_N}$  is one-dimensional, the inequality in (198) is scalar.

If  $N = 4$ , the situation simplifies further, because  $\text{int } \mathcal{A}_{\bar{\mathbf{m}}} = \mathcal{A}_{\bar{\mathbf{m}}}$  and the inclusion in (195) becomes an equality. In addition,  $\mathcal{R}_{\bar{\mathbf{m}}} = \mathcal{R}_{\bar{\mathbf{m}}}^{\text{exp}}$  and  $\bar{\mathbf{r}}$  is a diffeomorphism on all of  $\mathcal{A}_{\bar{\mathbf{m}}}$ . Therefore

$$(199) \quad \begin{aligned} \mathcal{D}_{\mathbf{m}} &= \{\boldsymbol{\rho} : \rho_N > \mathbf{r}_N(\boldsymbol{\alpha}), \bar{\boldsymbol{\alpha}} \in \mathcal{A}_{\bar{\mathbf{m}}}, \boldsymbol{\alpha}_N = \boldsymbol{\alpha}_{N-1} = 0\} \\ &= \{\boldsymbol{\rho} : \rho_N > \mathbf{r}_N(0, 0, \bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})), \rho_{N-1} = \mathbf{r}_{N-1}(0, 0, \bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})), \bar{\boldsymbol{\rho}} \in \mathcal{R}_{\bar{\mathbf{m}}}\}. \end{aligned}$$

The components  $\mathbf{r}_N(0, 0, \bar{\mathbf{a}}(\bar{\boldsymbol{\rho}}))$  and  $\mathbf{r}_{N-1}(0, 0, \bar{\mathbf{a}}(\bar{\boldsymbol{\rho}}))$  are simple to compute since  $\mathbf{r}(0, 0, \bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})) = \langle \mathbf{m} G_{\bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})} \rangle$  and  $\bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})$  has an explicit formula when  $\bar{N} = 2$ . (See the examples in section 5.3.)

2. *A non-Junkian example.* The situation becomes more complicated when  $\mathbf{m}_N$  includes polynomials other than  $|v|^N$ , because the inequality constraints from the relaxed minimization problem (20) are no longer scalar. The simplest example of this type occurs when

$$(200) \quad \mathbf{m}_N = (v \vee v) |v|^{N-2}.$$

We examine in detail the two-dimensional case ( $d = 2$ ) and write  $\boldsymbol{\alpha}_N \in A_{\mathbf{m}_N}$  in the form of a symmetric matrix:

$$(201) \quad \boldsymbol{\alpha}_N = \begin{pmatrix} (\boldsymbol{\alpha}_N)_{11} & (\boldsymbol{\alpha}_N)_{12} \\ (\boldsymbol{\alpha}_N)_{21} & (\boldsymbol{\alpha}_N)_{22} \end{pmatrix} = \begin{pmatrix} a+b & c \\ c & a-b \end{pmatrix}.$$

As a matrix,  $\boldsymbol{\alpha}_N$  must be negative-definite. Thus, with respect to the  $(a, b, c)$  coordinates, the set  $A_{\mathbf{m}_N}$  is a cone in  $\mathbb{R}^3$  that can be found in a high school geometry text:

$$(202) \quad \begin{aligned} A_{\mathbf{m}_N} &= \{(a, b, c) \in \mathbb{R}^3 : a \leq -\sqrt{b^2 + c^2}\} \\ \text{and} \quad \partial A_{\mathbf{m}_N} &= \{(a, b, c) \in \mathbb{R}^3 : a = -\sqrt{b^2 + c^2}\}. \end{aligned}$$

Let  $\mathcal{S}$  be the stratification of  $\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ , and let  $S \in \mathcal{S}$  so that  $S_N \in \partial A_{\mathbf{m}_N}$ . The set  $\partial A_{\mathbf{m}_N}$  itself has a stratification  $\mathcal{T}$  consisting of two manifolds:  $T_1$  is the origin in  $\mathbb{R}^3$ , and  $T_2$  is the remainder of the cone. We consider  $S_N$  as a subset of each manifold separately.

- (a)  $\boldsymbol{\alpha}_N \in T_1$ . In this case,  $a = b = c = 0$  and

$$(203) \quad \mathcal{NC}(A_{\mathbf{m}_N}, \boldsymbol{\alpha}_N) = A_{\mathbf{m}_N}^\circ = \{\boldsymbol{\alpha}_N : \boldsymbol{\alpha}_N \geq 0\}.$$

The situation essentially reduces to the Junkian case. The fiber bundle associated with  $S \subset \{\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}} : \boldsymbol{\alpha}_N = 0\}$  is

$$(204) \quad \mathcal{B}(S) = \{\boldsymbol{\rho} : \rho_N \geq A_{\mathbf{m}_N}^\circ \mathbf{r}_N(\boldsymbol{\alpha}), \boldsymbol{\alpha} \in S\},$$

and if  $N = 4$ ,

$$(205) \quad \begin{aligned} \mathcal{B}(S) &= \{\boldsymbol{\rho} : \rho_N \geq A_{\mathbf{m}_N}^\circ \mathbf{r}_N(0, 0, \bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})), \rho_{N-1} \\ &= \mathbf{r}_{N-1}(0, 0, \bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})), \bar{\boldsymbol{\rho}} \in \mathcal{R}_{\bar{\mathbf{m}}}\}, \end{aligned}$$



where  $\bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})$  has an explicit formula. (See the examples in section 5.3.) However, unlike the Junkian case, the inequalities in (204) and (205) are no longer scalar. Rather, it must be understood in terms of the polar cone  $A_{\mathbf{m}_N}^\circ$ .

- (b)  $\boldsymbol{\alpha}_N \in T_2$ . In this case  $a \leq -|b| < 0$ . In the  $(a, b, c)$  coordinates,  $\mathcal{NC}(A_{\mathbf{m}_N}, \boldsymbol{\alpha}_N)$  is a ray:

$$(206) \quad \mathcal{NC}(A_{\mathbf{m}_N}, \boldsymbol{\alpha}_N) = \left\{ \lambda \left( \sqrt{b^2 + c^2}, b, c \right) : \lambda \geq 0 \right\},$$

which can then be reexpressed in terms of the components of  $\boldsymbol{\alpha}_N$  by inverting (201). The bundle associated with any  $S \subset \{\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}} : \boldsymbol{\alpha}_N \neq 0\}$  is

$$(207) \quad \mathcal{B}(S) = \{\boldsymbol{\rho} : \boldsymbol{\rho}_N = \mathbf{r}_N(\boldsymbol{\alpha}) + \mathcal{NC}(A_{\mathbf{m}_N}, \boldsymbol{\alpha}_N), \boldsymbol{\rho}_j = \mathbf{r}_j(\boldsymbol{\alpha}), j < N, \boldsymbol{\alpha} \in S\}.$$

The set  $\mathcal{D}_{\mathbf{m}}$  is the union of sets of the form  $\mathcal{B}_0(S) = \mathcal{B}(S) \setminus \mathbf{r}(S)$ , where  $\mathcal{B}(S)$  is a bundle of the type given in (204) or (207).

One should note from these examples that our ability to identify degenerate densities is currently limited by our inability to explicitly identify the elements of  $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ . However, because the set  $\mathcal{A}_{\mathbf{m}_N}$  is semialgebraic, one should presumably be able to compute  $\mathcal{NC}(A_{\mathbf{m}_N}, \boldsymbol{\alpha}_N)$  for any given  $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ , even though such computations will likely be much more tedious than in the examples given above.

**7. Conclusions and discussion.** We have given in this paper a description of the set  $\mathcal{D}_{\mathbf{m}}$  of degenerate densities based on a geometric interpretation of the complementary slackness conditions associated with the dual formulation of (20). Roughly speaking, the set  $\mathcal{D}_{\mathbf{m}}$  is constructed by attaching a convex cone to every point in the boundary component  $\partial\mathcal{R}_{\mathbf{m}}^{\text{exp}} \cap \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ . This description recovers and extends previous results concerning the constrained entropy minimization problem.

Analytically, we see three important open questions that must be solved. First, one must determine if Assumptions I and II hold in a setting that is more general than the examples in (182). Concerning Assumption I, this means understanding the structure of the set of polynomials  $p$  for which  $v \mapsto p(v)e^{p(v)}$  is Lebesgue integrable. For example, do the coefficients of such polynomials form a semialgebraic set? Second, it must be determined whether the sets  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$  and  $\mathcal{R}_{\mathbf{m}}$  are invariant under the dynamics of the balance law (12) with the entropy-based closure. (Although not discussed in this paper, such a condition on  $\mathcal{R}_{\mathbf{m}}$  is obviously necessary for entropy-based closures to have any practical application.) Finally, it must be determined whether the existence of degenerate densities and the dynamics of (12) near such densities are simply artifacts of the entropy-based closure or if they actually reflect some physically relevant properties of the original Boltzmann equation (2).

Numerically speaking, a full implementation of entropy-based closures for gas dynamics faces many challenges. (An implementation has been attempted in [34], although the issue of degenerate densities was not addressed.) Clearly a discretization of (12) must preserve any invariant properties of  $\mathcal{R}_{\mathbf{m}}$  and  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$  with respect to the balance law (12). As pointed out in [20], even if  $\mathcal{R}_{\mathbf{m}}^{\text{exp}}$  is invariant under (12), solving the dual optimization problem (74) becomes extremely difficult for  $\boldsymbol{\rho}$  near  $\mathcal{D}_{\mathbf{m}}$  because the function  $h^*$  is very hard to evaluate. The reason for this is that, as  $\boldsymbol{\alpha}$  approaches  $\partial\mathcal{A}_{\mathbf{m}}$ , the function  $G_{\boldsymbol{\alpha}}$  can develop isolated modes that are often overlooked in a numerical quadrature. The result is a regularization effect in which accuracy is lost. In addition, the matrix  $\langle \mathbf{m}\mathbf{m}^T G_{\boldsymbol{\alpha}} \rangle$  becomes poorly conditioned near the boundary of

$\mathcal{A}_{\mathbf{m}}$ . Any minimization algorithm for (74) must be carefully formulated in order to overcome these challenges. Furthermore, as with the degenerate densities themselves, one must determine if these difficulties are by-products of the closure or related in some way to the dynamics of the Boltzmann equation.

**Appendix.** The purpose of this appendix is to provide the reader with a reference for important notation used in the main body of the paper. We includes tables of important sets and mappings (Tables 1 and 2) and also a diagram (Figure 1)

TABLE 1

*A list of important sets and properties used in this paper. Properties in brackets are known to hold under Assumptions I and II.*

Set	Lies in. . .	Defining equation(s)	Important properties
$\mathbb{F}_{\mathbf{m}}$	$L^1(\mathbb{R}^d)$	(15)	Convex cone; closure is proper
$\mathcal{A}_{\mathbf{m},j}$	$\mathbb{R}^{n_j}$	(44)	Proper cone for $j$ even
$\mathcal{A}_{\mathbf{m}}$	$\mathbb{R}^n$	(36)	Proper cone
$\mathcal{A}_{\mathbf{m}}$	$\mathbb{R}^n$	(40)	$\text{int}(\mathcal{A}_{\mathbf{m}}) = \mathbb{R}^{n-n_N} \times \text{int } \mathcal{A}_{\mathbf{m},N}$ ; $\text{cl}(\mathcal{A}_{\mathbf{m}}) = \mathbb{R}^{n-n_N} \times \mathcal{A}_{\mathbf{m},N}$ ; $\partial \mathcal{A}_{\mathbf{m}} \subset \mathbb{R}^{n-n_N} \times \partial \mathcal{A}_{\mathbf{m},N}$
$\mathcal{R}_{\mathbf{m}}$	$\mathbb{R}^n$	(33)	Open, solid, convex cone; $\mathcal{R}_{\mathbf{m}} = \text{int } \mathcal{A}_{\mathbf{m}}^\circ$
$\mathcal{R}_{\mathbf{m}}^{\text{exp}}$	$\mathbb{R}^n$	(42)	Solid cone; in general, not convex or open; $\mathcal{R}_{\mathbf{m}}^{\text{exp}} \subset \mathcal{R}_{\mathbf{m}}$ ; $[\mathcal{R}_{\mathbf{m}} \subset \text{cl}(\text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}})]$
$\mathcal{D}_{\mathbf{m}}$	$\mathbb{R}^n$	(34), (73)	$\mathcal{D}_{\mathbf{m}} = \mathcal{R}_{\mathbf{m}} \setminus \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ ; cone; [zero Lebesgue measure]

TABLE 2

*A list of important functions and properties used in this paper.*

Function	Domain/Range	Defining equation(s)	Important properties
$\mathbf{m}$	$\mathbb{R}^d \rightarrow \mathbb{R}^n$	(8)	Polynomial components; see (21)
$\mathcal{H}$	$\mathbb{F}_{\mathbf{m}} \rightarrow \mathbb{R} \cup \{\infty\}$	(4), (47)	Strictly convex and bounded below on $\mathbb{F}_{\mathbf{m}}$
$G_{\alpha}$	$\mathcal{A}_{\mathbf{m}} \rightarrow \mathbb{F}_{\mathbf{m}}$	(38)	Positive; convex on $\mathcal{A}_{\mathbf{m}}$
$\mathbf{r}$	$\mathcal{A}_{\mathbf{m}} \rightarrow \mathcal{R}_{\mathbf{m}}^{\text{exp}}$	(41)	Bijjective on $\mathcal{A}_{\mathbf{m}}$ ; diffeomorphic on $\text{int } \mathcal{A}_{\mathbf{m}}$ ; derivative of $h^*$ on $\text{int } \mathcal{A}_{\mathbf{m}}$
$\mathbf{a}$	$\mathcal{R}_{\mathbf{m}} \rightarrow \mathcal{A}_{\mathbf{m}}$	(55)	Continuous on $\mathcal{R}_{\mathbf{m}}$ ; diffeomorphic on $\text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ ; $\mathbf{a} \circ \mathbf{r}$ is identity on $\mathcal{A}_{\mathbf{m}}$
$h$	$\mathcal{R}_{\mathbf{m}} \rightarrow \mathbb{R}$	(14), (19), (20), (115)	Convex, differentiable on $\mathcal{R}_{\mathbf{m}}$ ; strictly convex on $\text{int } \mathcal{R}_{\mathbf{m}}^{\text{exp}}$ ; Legendre dual of $h^*$ on $\text{int } \mathcal{A}_{\mathbf{m}}$ ;
$h^*$	$\mathcal{A}_{\mathbf{m}} \rightarrow \mathbb{R}$	(39)	Strictly convex; directionally differentiable on $\mathcal{A}_{\mathbf{m}}$ ; differentiable on $\text{int } \mathcal{A}_{\mathbf{m}}$ ; Legendre dual of $h$ on $\text{int } \mathcal{A}_{\mathbf{m}}$ ; generally not continuous at $\partial \mathcal{A}_{\mathbf{m}} \cap \mathcal{A}_{\mathbf{m}}$
$\mathcal{L}$	$\mathbb{F}_{\mathbf{m}} \times \mathbb{R}^n \times \mathcal{R}_{\mathbf{m}} \rightarrow \mathbb{R} \cup \{\infty\}$	(74)	Strictly convex with respect to first argument
$\psi$	$\mathbb{R}^n \times \mathcal{R}_{\mathbf{m}} \rightarrow \mathbb{R} \cup \{-\infty\}$	(75)	Strictly concave; $\psi(\alpha, \rho) = \alpha^T \rho - h^*(\alpha)$

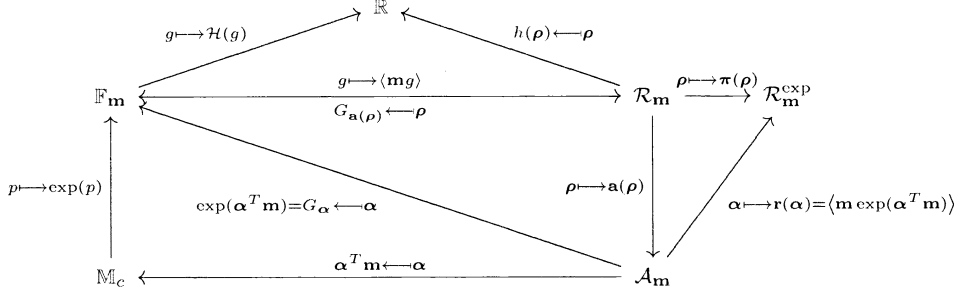


FIG. 1. A commutative diagram summarizing mappings and relationships between important sets.

emphasizing the relationships between different sets. Recall that a cone is proper when it is closed, pointed, convex, and has a nonempty interior (see section 2.2).

**Acknowledgments.** The authors thank Dr. Jack Calcut, who was a valuable reference for many technical items in sections 6.3–6.6 related to algebraic geometry, including the proof of Theorem 28. They also thank the referees, who provided insightful comments and suggestions. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Energy.

#### REFERENCES

- [1] A. M. ANILE, G. MASCALI, AND V. ROMANO, *Recent Developments in Hydrodynamical Modeling of Semiconductors*, Lecture Notes in Math. 1823, Springer-Verlag, Berlin, 2003, pp. 1–56.
- [2] R. BENEDETTI AND J.-J. RISLER, *Real Algebraic and Semi-Algebraic Sets*, Actualités Mathématiques, Hermann, Paris, 1990.
- [3] D. P. BERTSEKAS, A. NEDIĆ, AND A. E. OZDAGLAR, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.
- [4] L. BOLTZMANN, *Studien über das Gleichgewicht der lebendigen Kraft zwischen bewegten materiellen Punkten*, Wien. Ber., 58 (1868), pp. 517–560.
- [5] L. BOLTZMANN, *Über die Beziehung zwischen dem zweiten Hauptsatz der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen*, Wien. Ber., 76 (1877), pp. 373–435.
- [6] J. M. BORWEIN AND A. S. LEWIS, *Duality relationships for entropy-like minimization problems*, SIAM J. Control Optim., 29 (1991), pp. 325–338.
- [7] S. BOYD AND L. VANDENDERGHE, *Convex Optimization*, Cambridge University Press, New York, 2004.
- [8] H. B. CALLEN, *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed., John Wiley and Sons, New York, 1985.
- [9] C. CERCIGNANI, R. ILNER, AND M. PULVIRENTI, *The Mathematical Theory of Dilute Gases*, Appl. Math. Sci. 106, Springer-Verlag, New York, 1994.
- [10] I. CSISZAR, *I-divergence geometry of probability distributions and minimization problems*, Ann. Probab., 3 (1975), pp. 146–158.
- [11] P. DEGOND AND C. RINGHOFER, *Quantum moment hydrodynamics and the entropy principle*, J. Stat. Phys., 112 (2003), pp. 587–627.
- [12] B. DUBROCA AND J.-L. FUEGAS, *Étude théorique et numérique d’une hiérarchie de modèles aus moments pour le transfert radiatif*, C. R. Acad. Sci. Paris I, 329 (1999), pp. 915–920.
- [13] B. DUBROCA AND A. KLAR, *Half-moment closure for radiative transfer equations*, J. Comput. Phys., 180 (2002), pp. 584–596.
- [14] L. C. EVANS, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 2002.
- [15] G. B. FOLLAND, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

- [16] J. W. GIBBS, *Elementary Principles in Statistical Mechanics*, Charles Scribner's Sons, New York, 1902.
- [17] C. D. HAUCK, *Entropy-Based Moment Closures in Semiconductor Models*, Ph.D. thesis, University of Maryland, College Park, MD, 2006.
- [18] M. W. HIRSCH, *Differential Topology*, Grad. Texts Math. 33, Springer-Verlag, New York, 1997.
- [19] E. T. JAYNES, *Information theory and statistical mechanics*, Phys. Rev., 106 (1957), pp. 620–630.
- [20] M. JUNK, *Domain of definition of Levermore's five moment system*, J. Stat. Phys., 93 (1998), pp. 1143–1167.
- [21] M. JUNK, *Maximum entropy for reduced moment problems*, Math. Models Methods Appl. Sci., 10 (2000), pp. 1001–1025.
- [22] M. JUNK AND V. ROMANO, *Maximum entropy systems of the semiconductor Boltzmann equation using Kane's dispersion relation*, Contin. Mech. Thermodyn., 17 (2004), pp. 247–267.
- [23] S. KULLBACK, *Information Theory and Statistics*, John Wiley and Sons, New York, 1959.
- [24] C. D. LEVERMORE, *Moment closure hierarchies for kinetic theory*, J. Stat. Phys., 83 (1996), pp. 1021–1065.
- [25] C. D. LEVERMORE, *Moment closure hierarchies for the Boltzmann-Poisson equation*, VLSI Design, 6 (1998), pp. 97–101.
- [26] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1969.
- [27] J. MILNOR, *Singular Points of Complex Hypersurfaces*, Princeton University Press and the University Press of Tokyo, Princeton, NJ, 1968.
- [28] M. PLANCK, *Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum*, Verhandlungen der Deutschen Physikalischen Gesellschaft, 2 (1900), pp. 237–245.
- [29] M. PLANCK, *Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum*, Ann. Phys., 4 (1901), pp. 553–563.
- [30] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [31] J. SCHNEIDER, *Entropic approximation in kinetic theory*, Math. Model. Numer. Anal., 38 (2004), pp. 541–561.
- [32] C. E. SHANNON, *A mathematical theory of communication*, Bell System Tech. J., 27 (1948), pp. 379–423 and 623–656.
- [33] J. C. STRIKWERDA, *Finite Difference Schemes and Partial Differential Equations*, 2nd ed., SIAM, Philadelphia, 2004.
- [34] P. L. TALLEC AND J. P. PERLAT, *Numerical Analysis of Levermore's Moment System*, Preprint 3124, INRIA, Le Chesnay Cedex, France, 1997, pp. 1–36.

# INPUT-OUTPUT MODEL EQUIVALENCE OF SPIN SYSTEMS: A CHARACTERIZATION USING LIE ALGEBRA HOMOMORPHISMS\*

FRANCESCA ALBERTINI<sup>†</sup> AND DOMENICO D'ALESSANDRO<sup>‡</sup>

**Abstract.** In this paper, we consider the problem of model equivalence for quantum systems. Two models are said to be (input-output) equivalent if they give the same output for every admissible input. In the case of quantum systems, we take as output the expectation value of a given observable or, more generally, a probability distribution for the result of a quantum measurement. We link the input-output equivalence of two models to the existence of a homomorphism of the underlying Lie algebra. In several cases, a Cartan decomposition of the Lie algebra  $su(N)$  is useful to find such a homomorphism and to determine the classes of equivalent models. We consider in detail the important cases of two level systems with a Cartan structure and of spin networks. In the latter case, complete results are given generalizing previous results to networks of spin particles with arbitrary values of the spins. In treating this problem, we give independent proofs of some instrumental results on the subalgebras of  $su(N)$ .

**Key words.** quantum control systems, parameter identification, Lie algebraic methods, spin systems

**AMS subject classifications.** 93B30, 17B45, 17B81

**DOI.** 10.1137/050632671

**1. Introduction.** Structural properties of quantum systems recently have been the subject of investigation with methods of control theory. Appropriate definitions of controllability and observability of quantum systems have been given and practical conditions for checking these properties have been proposed (see, e.g., [1], [10], [17], [23]). In many cases, the tools used are those of Lie algebra and Lie group theory. Information on the properties of the dynamics is obtained by a study of the structure of a Lie algebra associated with the system and how this relates to the particular equations at hand. This geometric approach has proved useful not only to analyze the dynamics but also to design control laws (see, e.g., [3], [21], [22], [13]). This approach can also be used to study problems of parameter identification of quantum systems, and this is the subject of the present paper.

The problem we shall study is the classification of models of quantum systems whose behavior cannot be distinguished by an external observer. We shall call these models (*input-output*) *equivalent*. This problem is motivated by several experimental scenarios. In particular consider a molecule which is a network of particles with spin with all the other degrees of freedom neglected. A model Hamiltonian is associated with this system in which parameters modeling the interaction between particles as well as the interaction with an external electromagnetic field are unknown. Also, the initial state of the system might be unknown. In experimental scenarios such as nuclear magnetic resonance and electron paramagnetic resonance, it is possible to drive the system with a magnetic field and measure the expectation value of a given observable, for example, the total spin in a given direction. The question of

---

\*Received by the editors May 30, 2005; accepted for publication (in revised form) February 22, 2008; published electronically July 2, 2008.

<http://www.siam.org/journals/sicon/47-4/63267.html>

<sup>†</sup>Department of Pure and Applied Mathematics, University of Padova, ViaTrieste 63, 35121 Padova, Italy (albertin@math.unipd.it).

<sup>‡</sup>Department of Mathematics, Iowa State University, Ames, IA 50011 (daless@iastate.edu). This author's research was supported by the NSF under Career grant ECS-0237925.

fundamental and practical importance is to what extent, with this type of experiments, it is possible to distinguish between different models. As we shall see in this paper, this question is related to the existence of a particular Lie algebra homomorphism which maps one into the other the equations of the two models (definitions of Lie algebra theory are given in section 3). Further motivation for this research can be found in [27] where it was shown that thermodynamic methods commonly used to identify the parameters of spin networks such as in molecular magnets [5], [8] are not always adequate.

The main results of this paper are the solution of the model equivalence problem for a class of two level systems in Theorem 2 and Theorem 7 where we completely solve the problem of characterizing equivalent models for networks of spin. The latter result generalizes results previously obtained in [2] and [11], which were proven only for networks of spin  $\frac{1}{2}$  and 1's, to networks of interacting spins of any value and where the spin itself is an unknown parameter to be identified. The generalization given in this paper is obtained through a Cartan decomposition technique recently presented in [12] which helps in determining the homomorphism between equivalent models in the form of a Cartan involution. In the process we shall prove a number of auxiliary results (Theorems 4–6) on the structure of the Lie algebra  $su(N)$  and in particular on its subalgebras. These results could be formulated in terms of representation theory of Lie algebras (see, e.g., [18]), but we give here independent proofs which use only linear algebra arguments.

The paper is organized as follows. In section 2, we describe the problem of model equivalence for quantum systems. In section 3 we provide some background definitions of Lie algebra theory and describe how the quantum mechanical models we consider are related to mathematical notions in this theory. In section 4, we link the equivalence of two models to the existence of an appropriate Lie algebra homomorphism. This is the content of Theorem 1. In several cases the structure of the dynamics is related to a Cartan decomposition of  $su(N)$  and suggests the form of such a homomorphism as well as of the classes of equivalent models. We give a two level example in section 5 and treat the case of general spin networks in section 6. Instrumental to the solution of the model equivalence problem for spin networks are some results of independent interest concerning the existence of subalgebras of  $su(N)$  with specific features. The proofs of these results are presented in section 7. Concluding remarks are given in section 8.

**2. The problem of model equivalence for quantum systems.** Consider a model Hamiltonian for a quantum system,  $H(t) := H(u(t))$ , where, in a semiclassical description, the dependence on time is due to the interaction with classical external fields,  $u := u(t)$ , which play the role of *controls*. For every  $t$ ,  $H(t)$  is a Hermitian operator on a Hilbert space  $\mathcal{H}$ . The state of a quantum system is described by a density matrix  $\rho$ , i.e., a positive, trace one, Hermitian operator on  $\mathcal{H}$ . The evolution of the state of the system,  $\rho := \rho(t)$ , is determined, other than by  $H$ , by the initial state  $\rho(0) = \rho_0$ . In particular,  $\rho$  is the solution of the *Liouville's equation*,

$$(1) \quad \dot{\rho} = [-iH, \rho],$$

with initial condition  $\rho(0) = \rho_0$ . Here and in the following, for two matrices  $A$  and  $B$ , the commutator  $[A, B]$  is defined as  $[A, B] := AB - BA$ . According to the measurement postulate of quantum mechanics, with any measured quantity there is associated an observable  $S$  which is a Hermitian operator on  $\mathcal{H}$ . There are various

types of measurements (see, e.g., [6]). Considering, for simplicity, a Von Neumann–Lüders measurement,<sup>1</sup> writing  $S$  in terms of orthogonal projections

$$(2) \quad S := \sum_j \lambda_j \Pi_j,$$

the probability of having a result  $\lambda_j$ , when the state is  $\rho$ , is given by

$$(3) \quad P_j := \text{Tr}(\Pi_j \rho).$$

As the probabilities  $P_j$  are the only information that can be gathered by an external observer, we are motivated to ask what classes of models  $\{H, \rho_0\}$  will give the same probabilities, for any functional form of the control  $u$ . In other words, we ask what classes of models are indistinguishable by experiments that involve driving the system with controls, in a given set of functions, and measuring a given observable. These models will be called (*input-output*) *equivalent*.

*Remark 2.1.* The term *output* for the probabilities (3) or for the expectation value (4) below needs to be clarified. We have called output the information that can be gathered measuring a given observable  $S$ . In particular, consider a large number of identical systems. By measuring the observable  $S$  at a fixed time  $t$  and recording the various results, we obtain the probabilities (3). Two models that have the same probabilities (output) at any time  $t$  are exactly equivalent from the input-output point of view, because a measurement of  $S$  on any of them will give exactly the same result with exactly the same probability (or will have exactly the same expectation value if that is what interests us). Therefore, while it is convenient to call the functions (3) and (4) output and make a mathematical analogy with classical systems, the situation is not the same as for classical systems. In particular, for classical systems it is possible to monitor the output continuously. For quantum systems we consider measurements that in practice happen only at one instant of time. However, we can still ask what the classes of models would give *at any time* the same expectation value (output (4)) or the same result with the same probability (output (3)).<sup>2</sup>

It is appropriate to treat the case where the result of the measurement is the *expectation value* of the measurable  $S$ , i.e.,

$$(4) \quad y := \text{Tr}(S\rho).$$

Therefore, we take (4) as the *output* of the system. Not only is this the case in several experimental situations, such as nuclear magnetic resonance, but it is not a significant restriction as compared to the case where the probabilities (3) are considered. As the structure of the output (4) is the same as that of the outputs (3), the passage from the treatment for the expectation value to the one for probabilities corresponds to extending a single output treatment to a multiple output treatment. This can be accomplished without difficulties.

We need to assume some structure on the Hamiltonian  $H$ , in order to render the problem of characterizing the classes of equivalent models tractable. This corresponds to the passage from *unstructured uncertainty* to *parametric uncertainty*, often

<sup>1</sup>Natural extensions of what we shall say can be made to general measurements.

<sup>2</sup>Considering continuous measurements in quantum mechanics is possible and of interest in several experimental scenarios. In these cases, however, the equations of the dynamics (1) have to be modified to take into account the back-action of the measurement (see, e.g., [6]). These types of models will not be considered here.

discussed in identification theory (see, e.g., [26]). In particular, it is often the case that the Hamiltonian  $H = H(u)$  has the *bilinear* form

$$(5) \quad H := H_0 + \sum_{j=1}^m H_j u_j(t)$$

for some control functions  $u_1, \dots, u_m$ , and *internal Hamiltonian*  $H_0$  and *interaction Hamiltonians*  $H_j$ 's,  $j = 1, \dots, m$ . In this paper we shall consider only *finite dimensional models*, and therefore  $H_0$ ,  $H_j$ ,  $j = 1, \dots, m$ , and  $\rho$  are Hermitian matrices of finite dimension  $N \times N$ .

*Remark 2.2.* A classical problem in nonlinear control theory related to the one considered here is the *realization problem*. In the relevant literature, one considers a map from a space of input functions to a space of output functions. This map can be given in an abstract way [20], [28], or through Volterra [7], [9], [14], [19] or Fliess [15] series. Then conditions are given for the *existence* and *uniqueness* of a dynamical model which implements the given input-output relations. An algorithm for the construction of this model for bilinear systems on  $\mathbb{R}^N$  is given, for example, in [24]. In our case, the input-output map is already given as realizable with a given class of models where only the parameters and the dimensions are unknown. The problem is to characterize the class of equivalence of models giving the specified input-output map. Therefore the problem considered here is essentially a *uniqueness* problem in realization theory while the existence is already assumed. In this sense the question treated here is more in the spirit of the work done in [4] for neural networks. The a priori assumptions on the structure of the system allow us to obtain stronger results. In particular, while in [20], [28] (under a suitable hypothesis) the realization is proved to be unique up to a diffeomorphism, in the results of sections 5 and 6 we shall give the explicit map between two different realizations of the same input-output map and therefore the explicit construction of all the input-output equivalent models. We notice that since the model (1), (5), (4) is bilinear, we could have followed a different approach by looking at the input-output map associated with the system, transforming the system into a system on  $\mathbb{R}^{N^2}$ , where  $N^2$  is the dimension of the space of Hermitian matrices, and applying the results of [24]. This, however, does not seem to be the most natural approach. In fact, in doing this, we would have hidden some of the structure of the problem, as, for example, the fact that the linear map for the dynamics in (1) is given in terms of the commutator with a Hermitian matrix.

**3. Lie algebra theory and modeling of quantum systems.** In this paper we are interested in matrix *Lie algebras* over the real field, i.e., real vector spaces of matrices closed under the commutator. Particularly important Lie algebras for us are the Lie algebra of *skew-Hermitian*  $N \times N$  matrices, which is denoted by  $u(N)$ , and the Lie algebra of *skew-Hermitian*  $N \times N$  matrices with trace equal to zero, which is denoted by  $su(N)$ . Accordingly the spaces of Hermitian matrices and Hermitian matrices with zero trace will be denoted by  $iu(N)$  and  $isu(N)$ , as their elements are obtained from those of  $u(N)$  or  $su(N)$  by multiplication by the imaginary unit  $i := \sqrt{-1}$ . In general, we shall often use the notation  $i\mathcal{L}$  to denote a subspace of  $iu(N)$  corresponding to a subspace  $\mathcal{L}$  of  $u(N)$  (or vice versa). All the spaces are inner product spaces when equipped with the inner product  $\langle A, B \rangle := \text{Tr}(AB^*)$ . A subalgebra  $\mathcal{L}$  of  $u(N)$  is a subspace of  $u(N)$  which is also a Lie algebra. To every matrix Lie algebra  $\mathcal{L}$  is associated a *Lie group* which is the group generated by elements  $e^A$  with  $A$  in  $\mathcal{L}$  equipped with the structure of an analytic manifold. We shall denote this Lie



group by  $e^{\mathcal{L}}$ . The Lie group associated with  $u(N)$  ( $su(N)$ ) is the Lie group of unitary matrices (with determinant equal to 1) and is denoted by  $U(N)$  ( $SU(N)$ ).

In the model Hamiltonian (5) it is often true that  $H_0$  and the  $H_j$ 's belong to two orthogonal complementary subspaces of  $iu(N)$  corresponding to a Cartan decomposition [16] of  $u(N)$ . These are two orthogonal subspaces  $i\mathcal{K}$  and  $i\mathcal{P}$  such that the corresponding subspaces of  $u(N)$ ,  $\mathcal{K}$  and  $\mathcal{P}$ , satisfy<sup>3</sup>

$$(6) \quad u(N) = \mathcal{K} \oplus \mathcal{P}$$

and the commutation relations

$$(7) \quad [\mathcal{K}, \mathcal{K}] \subseteq \mathcal{K}, \quad [\mathcal{K}, \mathcal{P}] \subseteq \mathcal{P}, \quad [\mathcal{P}, \mathcal{P}] \subseteq \mathcal{K}.$$

If  $\mathcal{L}$  and  $\mathcal{L}'$  are two (matrix) Lie algebras, a *homomorphism*  $\phi$  is a linear map  $\phi : \mathcal{L} \rightarrow \mathcal{L}'$ , which preserves the commutation operation, i.e.,  $\phi([A, B]) = [\phi(A), \phi(B)]$ , where the commutators on the left and right are calculated in  $\mathcal{L}$  and  $\mathcal{L}'$ , respectively. If  $\mathcal{L}$  and  $\mathcal{L}'$  are two inner product spaces, associated with every linear map  $\phi$  is a dual map  $\phi^*$  which is a linear map  $\mathcal{L}' \rightarrow \mathcal{L}$  defined by the property  $\langle A, \phi(B) \rangle_{\mathcal{L}'} = \langle \phi^*(A), B \rangle_{\mathcal{L}}$ , where the inner product  $\langle \cdot, \cdot \rangle_{\mathcal{L}'}$  ( $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ ) is defined on  $\mathcal{L}'$  ( $\mathcal{L}$ ). A bijective homomorphism is called an *isomorphism*. An isomorphism is called an *automorphism* if  $\mathcal{L} = \mathcal{L}'$ . An automorphism  $\theta$  is called a *Cartan involution* if  $\theta^2$  is the identity map. Associated with every Cartan decomposition satisfying (7) is a Cartan involution  $\theta$  such that  $\mathcal{K}$  and  $\mathcal{P}$  are the  $+1$  and  $-1$  eigenspaces of  $\theta$ . The following simple example illustrates these concepts.

*Example 3.1.* Consider the decomposition (6) of  $u(N)$  where  $\mathcal{K}$  and  $\mathcal{P}$  are the subspaces of purely real and purely imaginary matrices, respectively. It is clear that the commutation relations (7) are verified. Moreover, the associated Cartan involution  $\theta$  is the complex conjugation which leaves unchanged the elements in  $\mathcal{K}$  and changes the sign of the elements in  $\mathcal{P}$ .

Another type of automorphism of (subalgebras of)  $u(N)$  is a *conjugacy* which is defined as  $\phi_P(A) := PAP^*$ , where  $P$  is a unitary matrix and in particular such that  $PP^* = P^*P = \mathbf{1}$ , where  $\mathbf{1}$  is the identity matrix. Notice that if  $\mathcal{K}$  and  $\mathcal{P}$  define a Cartan decomposition of  $u(N)$  as in (6),  $P\mathcal{K}P^*$  and  $P\mathcal{P}P^*$  also define a decomposition.

Up to conjugacies, there are three types of Cartan decompositions of  $u(N)$  labeled **AI**, **AII**, and **AIII**. In a decomposition **AI**,  $\mathcal{K} = so(N)$  and  $\mathcal{P} = (so(N))^{\perp}$  up to conjugacy.  $so(N)$  is the subalgebra of  $u(N)$  of real skew-symmetric matrices. This is the decomposition presented in Example 3.1. In a decomposition **AII**,  $N$  is even and  $\mathcal{K} = sp(N/2)$  and  $\mathcal{P} = (sp(N/2))^{\perp}$  up to conjugacy.  $sp(N/2)$  is the subalgebra of  $u(N)$  of symplectic matrices, namely, matrices  $A$  satisfying  $AJ + JA^T = 0$ , where  $J$  is the matrix

$$J := \begin{pmatrix} 0 & \mathbf{1}_{N/2 \times N/2} \\ -\mathbf{1}_{N/2 \times N/2} & 0 \end{pmatrix}.$$

The third type labeled **AIII** is not relevant here (see [16] for details).

<sup>3</sup>In the literature (see, e.g., [16]) Cartan decompositions are defined for general semisimple Lie algebras, and the results of Cartan parametrize all the possible decompositions for  $su(N)$ . From decompositions of  $su(N)$  one can obtain decompositions of  $u(N)$  by including multiples of the identity in one of the two subspaces  $\mathcal{K}$  or  $\mathcal{P}$ . Although in the literature one often refers to decompositions of  $su(N)$ , we have preferred to define decompositions in terms of  $u(N)$  to ease notation.

If the system is a multipartite system, every  $H_j$ ,  $j = 1, \dots, m$ , in (5) is a linear combination of Hamiltonians modeling the interaction of each individual system with the external field. In matrix notation,  $H_j$  is a linear combination of elements of the type  $\mathbf{1} \otimes \mathbf{1} \otimes \dots \otimes \mathbf{1} \otimes L \otimes \mathbf{1} \otimes \dots \otimes \mathbf{1}$ , where  $L$  is a Hermitian matrix of appropriate dimensions and all the other places are occupied by identities  $\mathbf{1}$ . Also,  $H_0$  is very often a linear combination of elements modeling the interaction between two subsystems, which can be written as tensor products of matrices equal to the identity, except in two locations. In these cases, using the Cartan decomposition (6) described in the recent paper [12] one finds that  $iH_0 \in \mathcal{P}$  and  $iH_j \in \mathcal{K}$ ,  $j = 1, \dots, m$ . Also if  $S$  is a sum of observables on each individual subsystem, i.e., total spin angular momentum (see, e.g., [25]), it can always be written as a sum of tensor products all equal to the identity, except in one position. In these cases,  $iS \in \mathcal{K}$  belongs to the subspace  $\mathcal{K}$  of the above Cartan decomposition of [12].<sup>4</sup> We shall describe this structure in detail in section 6 because it is the main feature we use for the solution of the input-output equivalence problem.

According to the results in [1], [10], [17], [23], the controllability and observability properties of system (1), (4), (5) depend on the Lie algebra  $\mathcal{L}$  generated by  $iH_0$  and  $iH_j$ ,  $j = 1, \dots, m$ , i.e., the smallest subalgebra of  $u(N)$  containing these matrices. In particular the set of states reachable from a density matrix  $\rho_0$  is given by

$$(8) \quad \mathcal{O}_{\rho_0} := \{X\rho_0X^* \mid X \in e^{\mathcal{L}}\}.$$

If, for every  $\rho_0$ ,  $\mathcal{O}_{\rho_0}$  is the set of all the density matrices with the same spectrum as  $\rho_0$ , the system is called *controllable*. This is the case if and only if  $\mathcal{L} = u(N)$  or  $\mathcal{L} = su(N)$ . The system is *observable* if there are no pairs of states indistinguishable by input-output experiments. This is the case if and only if

$$(9) \quad isu(N) \subseteq \mathcal{O}_S := \{XSX^* \mid X \in e^{\mathcal{L}}\}.$$

Controllability implies observability for nonscalar  $S$  [10].

In the following, we shall consider, as a standing assumption, only finite dimensionality of the Hamiltonian  $H$  and the bilinear form (5) and will make precise the assumptions on the (Cartan) structure of the Hamiltonian when needed.

**4. Model equivalence and Lie algebra homomorphisms.** Consider two models with a Hamiltonian of the form (5) and an output of the form (4):

$$(10) \quad \dot{\rho} = \left[ -i \left( H_0 + \sum_{j=1}^m H_j u_j \right), \rho \right], \quad \rho(0) = \rho_0, \quad y = \text{Tr}(S\rho),$$

$$(11) \quad \dot{\rho}' = \left[ -i \left( H'_0 + \sum_{j=1}^m H'_j u_j \right), \rho' \right], \quad \rho'(0) = \rho'_0, \quad y' = \text{Tr}(S'\rho').$$

The following theorem links the existence of an appropriate Lie algebra homomorphism to the equivalence of the two models.

**THEOREM 1.** *Let  $N$  and  $N'$  be the dimensions of the two models (10), (11), respectively. Let  $\phi$  be a homomorphism,  $\phi : u(N) \rightarrow u(N')$ , and  $\phi^*$  its dual with respect to the standard inner product  $\langle A, B \rangle := \text{tr}(AB^*)$ . Assume*

$$(12) \quad -iH'_0 = \phi(-iH_0), \quad -iH'_j = \phi(-iH_j), \quad \phi^*(iS') = iS.$$

<sup>4</sup>Notice that the situation may be different if we consider the case of a single output given by the expectation value (4) and the case of several outputs given by the probabilities in (3).

Then if

$$(13) \quad i\rho'_0 = \phi(i\rho_0),$$

the models are equivalent. Vice versa, if the models are equivalent and (11) is observable, then (13) holds.

*Proof.* Multiply (10) and (11) by  $i$  and then apply  $\phi$  to the equation obtained from (10). Combining the two resulting equations, using the first two of (12), we obtain

$$(14) \quad \frac{d}{dt}(i\rho' - \phi(i\rho)) = \left[ \phi(-iH_0) + \sum_j \phi(-iH_j)u_j, i\rho' - \phi(i\rho) \right].$$

If (13) is verified, then  $i\rho'(t) = \phi(i\rho(t))$  for every  $t$  and for every control. Therefore we have from the third one in (12),

$$(15) \quad \text{Tr}(S'\rho') = \text{Tr}(-iS'i(\rho')) = \text{Tr}(-iS'\phi(i\rho)) = \text{Tr}(\phi^*(-iS')i\rho) = \text{Tr}(S\rho),$$

and the two models are equivalent. Vice versa, assume that the two models are equivalent. From (15), we have

$$(16) \quad \text{Tr}(iS'(i\rho' - \phi(i\rho))(t)) = 0$$

for every  $t$ . Writing the solution of (14) as  $(i\rho' - \phi(i\rho))(t) = X(i\rho' - \phi(i\rho))(0)X^*$ , where  $X$  is the solution of the (Schrödinger) operator equation

$$(17) \quad \dot{X} = \left( \phi(-iH_0) + \sum_j \phi(-iH_j)u_j \right) X, \quad X(0) = \mathbf{1},$$

we have

$$(18) \quad \text{Tr}(X^*iS'X(i\rho'_0 - \phi(i\rho_0))) = 0.$$

As the system (11) is observable, we have that  $X^*iS'X$  span all of  $su(n')$ , which implies  $i\rho'_0 = \phi(i\rho_0)$ .  $\square$

Summarizing, the theorem says that if the equation describing the dynamics is related through a Lie algebra homomorphism  $\phi$ , and under an observability condition, then the two models are equivalent if and only if the initial states are related through the same Lie algebra homomorphism  $\phi$ . As we shall show in the remainder of the paper (cf. also [2]), it is possible for cases of physical interest to give a stronger version of Theorem 1. In particular, it is possible to show that the existence of a homomorphism  $\phi$  satisfying (12) is also necessary for equivalence of two models. Moreover, it is possible to construct such a homomorphism. This way, we can characterize the class of equivalent models. We shall do this for a two level example in the next section and for general spin networks in section 6. In both cases we exploit a Cartan decomposition underlying the dynamics of the models.

In general, more structure will have to be assumed to avoid trivial cases. For example, if  $S = S'$  is a scalar matrix, then every two models are equivalent. To avoid this case, an appropriate extra assumption is the *observability* of the two models. Also, we need to assume that the initial states are not both perfect mixtures (i.e., multiples of the identity); otherwise, with  $S = S'$ , the output for any two equivalent models will be the same, independently of the dynamics. Moreover,  $-iH_j$  and  $-iH'_j$ ,  $j = 0, \dots, m$ , may be generally assumed traceless, as the trace only adds an extra common phase factor to the dynamics, which cannot be detected. We shall use these assumptions in the following.

**5. Model equivalence of two level systems.** Consider a spin  $\frac{1}{2}$  particle which is driven by an electromagnetic control field along the  $z$  axis, interacts with a constant unknown magnetic field along an (unknown) direction in the  $x$ - $y$  plane, and has an unknown initial state. The practical question is to what extent, by driving the system with the control field and measuring the average value of the spin magnetization in the  $z$  direction, it is possible to obtain information about the unknown parameters of the system. This type of model has a Cartan structure which is shared by several other models of physical interest and is instrumental in finding a homomorphism between equivalent models. We describe this below.

The Lie algebra  $su(2)$ , which is the relevant Lie algebra in the two level case, has, up to conjugacy, only one Cartan decomposition which corresponds to the classical Euler decomposition of the Lie group  $SU(2)$  [16]. This extends to a decomposition of  $u(2)$  which can always be written as

$$(19) \quad u(2) = \mathcal{K} \oplus \mathcal{P}.$$

Here  $\mathcal{K}$  and  $\mathcal{P}$  satisfy the commutation relations in (7) and are given, up to conjugacy, by

$$(20) \quad \mathcal{K} := \text{span}\{i\sigma_z\}, \quad \mathcal{P} := \text{span}\{i\sigma_x, i\sigma_y, i\mathbf{1}_{2 \times 2}\}.$$

Here,  $\mathbf{1}_{2 \times 2}$  is the  $2 \times 2$  identity matrix and  $\sigma_x$ ,  $\sigma_y$ , and  $\sigma_z$  are the *Pauli matrices*

$$(21) \quad \sigma_x := \frac{1}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y := \frac{1}{2} \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}, \quad \sigma_z := \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

which satisfy the commutation relations

$$(22) \quad [i\sigma_x, i\sigma_y] = i\sigma_z, \quad [i\sigma_y, i\sigma_z] = i\sigma_x, \quad [i\sigma_z, i\sigma_x] = i\sigma_y.$$

The dynamical and output equations, for the above model of a spin  $\frac{1}{2}$  particle in an electromagnetic field, can be written as

$$(23) \quad \dot{\rho} = [A + i\sigma_z u(t), \rho], \quad y = \text{Tr}(\sigma_z \rho), \quad \rho(0) = \rho_0,$$

where  $\rho_0$  is an unknown initial density matrix and  $A := x i\sigma_x + y i\sigma_y$ , with  $x$  and  $y$  unknown. This model has a *Cartan structure* in that  $A$  is in  $\mathcal{P}$  and  $i\sigma_z$  (the control and observation part) is in  $\mathcal{K}$ , with  $\mathcal{K}$  and  $\mathcal{P}$  defined as in (20). We assume  $x^2 + y^2 \neq 0$  which implies controllability and therefore observability [10] for this model. The following result characterizes all the classes of equivalent models in terms of Lie algebra homomorphisms.

**THEOREM 2.** *Consider two models*

$$(24) \quad \dot{\rho} = [A + i\sigma_z u(t), \rho], \quad y = \text{Tr}(\sigma_z \rho), \quad \rho(0) = \rho_0,$$

$$(25) \quad \dot{\rho}' = [A' + i\sigma_z u(t), \rho], \quad y = \text{Tr}(\sigma_z \rho'), \quad \rho'(0) = \rho'_0,$$

with  $\rho_0$  and  $\rho'_0$  not both equal to scalar matrices (representing perfect mixtures) and  $A$  and  $A'$  of the form

$$(26) \quad A := x i\sigma_x + y i\sigma_y \quad \text{and} \quad A' := x' i\sigma_x + y' i\sigma_y$$

for real parameters  $x, y, x', y'$ . Assume

$$(27) \quad x^2 + y^2 \neq 0 \quad \text{and} \quad x'^2 + y'^2 \neq 0.$$

Then the two models are equivalent if and only if there exists an automorphism  $\phi : u(2) \rightarrow u(2)$  with

$$(28) \quad \phi^*(i\sigma_z) = i\sigma_z$$

and

$$(29) \quad A' = \phi(A), \quad \phi(i\sigma_z) = i\sigma_z, \quad i\rho'_0 = \phi(i\rho_0).$$

*Proof.* It is clear that if the automorphism  $\phi$  exists, satisfying (28) and (29), the two models are equivalent. This follows from a direct application of Theorem 1, with (28) and (29) replacing (12) and (13). To prove the opposite, first notice that, from the equivalence assumption, we have

$$(30) \quad y(t) := \text{Tr}(\sigma_z \rho(t)) = \text{Tr}(\sigma_z \rho'(t)) := y'(t)$$

for every  $t \geq 0$  and every admissible control.

We consider an automorphism  $\phi$  of the type

$$(31) \quad \phi(L) := e^{-i\alpha\sigma_z} L e^{i\alpha\sigma_z}, \quad L \in u(2),$$

as  $\alpha$  varies in  $\mathbb{R}$ .

Clearly (28) and the second equation of (29) are verified for any  $\alpha \in \mathbb{R}$ . Moreover,

$$(32) \quad \phi(A) = \bar{x}i\sigma_x + \bar{y}i\sigma_y,$$

with

$$(33) \quad \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = K_\alpha \begin{pmatrix} x \\ y \end{pmatrix},$$

and

$$(34) \quad K_\alpha := \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}.$$

Also, if we write

$$(35) \quad \begin{aligned} i\rho_0 &:= \rho_x i\sigma_x + \rho_y i\sigma_y + \rho_z i\sigma_z + \frac{1}{2}i\mathbf{1}, \\ i\rho'_0 &:= \rho'_x i\sigma_x + \rho'_y i\sigma_y + \rho'_z i\sigma_z + \frac{1}{2}i\mathbf{1}, \end{aligned}$$

we have

$$(36) \quad \phi(i\rho_0) = \bar{\rho}_x i\sigma_x + \bar{\rho}_y i\sigma_y + \bar{\rho}_z i\sigma_z + \frac{1}{2}i\mathbf{1},$$

with

$$(37) \quad \begin{pmatrix} \bar{\rho}_x \\ \bar{\rho}_y \end{pmatrix} = K_\alpha \begin{pmatrix} \rho_x \\ \rho_y \end{pmatrix}.$$

Using the equivalence assumption (30) at  $t = 0$ , we obtain

$$(38) \quad \rho_z = \rho'_z.$$

Moreover, differentiating (30), using the dynamical equations (24) and (25), we obtain

$$(39) \quad \text{Tr}(\rho[\sigma_z, A]) = \text{Tr}(\rho'[\sigma_z, A']).$$

Writing this at time  $t = 0$  and using the definitions (26) and (35) along with the commutation relation for the Pauli matrices (22), we obtain

$$(40) \quad \rho_y x - \rho_x y = \rho'_y x' - \rho'_x y'.$$

Differentiating (39) and using the fact that the resulting equation has to be valid for every value of the control, we obtain the two equations

$$(41) \quad \text{Tr}(\sigma_z[A, [A, \rho]]) = \text{Tr}(\sigma_z[A', [A', \rho']])$$

and

$$(42) \quad \text{Tr}(i\sigma_z[A, [\sigma_z, \rho]]) = \text{Tr}(i\sigma_z[A', [\sigma_z, \rho']]).$$

From (42), as for (40), we obtain

$$(43) \quad x\rho_x + y\rho_y = x'\rho'_x + y'\rho'_y.$$

From (41), we obtain

$$(44) \quad (x^2 + y^2)\text{Tr}(\sigma_z\rho) = (x'^2 + y'^2)\text{Tr}(\sigma_z\rho').$$

Using the fact that  $\text{Tr}(\sigma_z\rho)$  is not always zero (because of the controllability condition (27))<sup>5</sup> and (30), we have

$$(46) \quad x^2 + y^2 = x'^2 + y'^2.$$

Therefore, for some  $\alpha$ , we can write

$$(47) \quad \begin{pmatrix} x' \\ y' \end{pmatrix} = K_\alpha \begin{pmatrix} x \\ y \end{pmatrix},$$

with  $K_\alpha$  in (34), and this, compared with (33) and (32), gives the first term of (29). To obtain the third term (with the same  $\phi$ ), we recall from (27) that  $x^2 + y^2 \neq 0$ . Letting  $J := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$  and using (47), we can write (40) and (43), respectively, as

$$(48) \quad [x, y]J[\rho_x, \rho_y]^T = [x, y]K_\alpha^T J[\rho'_x, \rho'_y]^T,$$

$$(49) \quad [x, y][\rho_x, \rho_y]^T = [x, y]K_\alpha^T [\rho'_x, \rho'_y]^T.$$

Since  $K_\alpha^T$  commutes with  $J$ , we can write these as

$$(50) \quad \begin{pmatrix} [x, y]J \\ [x, y] \end{pmatrix} [\rho_x, \rho_y]^T = \begin{pmatrix} [x, y]J \\ [x, y] \end{pmatrix} K_\alpha^T [\rho'_x, \rho'_y]^T.$$

---

<sup>5</sup>From controllability (27), we cannot have

$$(45) \quad \text{Tr}(\sigma_z\rho(t)) \equiv \text{Tr}(\sigma_z\rho'(t)) \equiv 0$$

for every control. This would mean that, for every reachable evolution operator  $X$ , the solution of the (Schrödinger) operator equation  $\dot{X} = (A + i\sigma_z u)X$ , with  $X(0) = \mathbf{1}$ ,  $X^* \sigma_z X$ , would be orthogonal to  $\rho_0$ . However, because of controllability,  $X$  may attain all the values in  $SU(2)$ , and therefore  $X^* \sigma_z X$  span, as  $X$  varies, may attain all of  $isu(2)$ . Therefore,  $X^* \sigma_z X$  is always orthogonal to  $\rho_0$  only if  $\rho_0$  is a multiple of the identity, which we have excluded.

Since

$$x^2 + y^2 = -\det \begin{pmatrix} [x, y]^J \\ [x, y] \end{pmatrix} \neq 0,$$

we can write

$$(51) \quad [\rho'_x, \rho'_y]^T = K_\alpha [\rho_x, \rho_y]^T,$$

and therefore

$$(52) \quad [\rho'_x, \rho'_y] = [\bar{\rho}_x, \bar{\rho}_y],$$

which along with  $\rho'_z = \rho_z$  gives

$$(53) \quad i\rho' = \phi(i\rho).$$

This concludes the proof of the theorem.  $\square$

*Remark 5.1.* The proof of the above theorem also gives the explicit form of the homomorphism relating two equivalent models.

## 6. Model equivalence of spin networks.

**6.1. A set of models of spin networks.** We consider a network of  $n$  particles with spin that interact according to Heisenberg interaction. In particular, we denote the spin of the  $j$ th particle by  $l_j$  and by  $N_j := 2l_j + 1$  the dimension of the Hilbert space for the state of the  $j$ th particle. The dimension of the Hilbert state space associated with the entire network is  $N := \prod_{j=1}^n N_j$ . The class of Hamiltonians we consider is of the form

$$(54) \quad H(t) := i(A + B_x u_x(t) + B_y u_y(t) + B_z u_z(t)),$$

where  $A$ , modeling the *Heisenberg interaction* among the particles, and  $B_{x,y,z}$ , modeling the interaction with external fields, are given by

$$(55) \quad \begin{aligned} A &:= -i \sum_{k < l, k, l=1}^n J_{kl} (I_{kx, lx} + I_{ky, ly} + I_{kz, lz}), \\ B_v &:= -i \left( \sum_{k=1}^n \gamma_k I_{kv} \right) \quad \text{for } v = x, y, \text{ or } z, \end{aligned}$$

respectively. Here and in the following we denote by  $I_{k_1 v_1, \dots, k_r v_r}$ , for  $1 \leq k_1 < \dots < k_r \leq n$  and  $v_j \in \{x, y, z\}$ , the  $N \times N$  matrix which is the Kronecker product of  $n$  matrices where in the  $j$ th position we have the  $N_j \times N_j$  identity if  $j \notin \{k_1, \dots, k_r\}$ , while if  $j = k_s$  we have the  $N_j \times N_j$  representation of the  $v_s$  component of spin angular momentum for a particle with spin  $l_j$ . Such matrices are given by the Pauli matrices (21) in the case where  $l_j = \frac{1}{2}$  and can be calculated for every value of the spin (see, e.g., [25, section 3.5]). For convenience of the reader, and since these matrices will be used several times in the following, we give their explicit form in Appendix B. With some abuse of notation, we shall continue denoting these matrices by  $\sigma_x$ ,  $\sigma_y$ , and  $\sigma_z$  without explicit reference to the value of the spin. Therefore, for instance, we have that for a system of three spin,

$$(56) \quad I_{1x, 3y} := \sigma_x \otimes \mathbf{1} \otimes \sigma_y$$

is the Kronecker product of three matrices.  $\sigma_x$  is the representation of the spin angular momentum in the  $x$  direction for the first spin (dimension  $N_1 \times N_1$ ),  $\sigma_y$  is the representation of the spin angular momentum in the  $y$  direction for the third spin (dimension  $N_3 \times N_3$ ), and the matrix  $\mathbf{1}$  is the identity matrix for the second spin (dimension  $N_2 \times N_2$ ). In all dimensions, the matrices  $\sigma_x$ ,  $\sigma_y$ , and  $\sigma_z$  have several properties we shall use in the following. In particular, they satisfy the commutation relations (22) and

$$(57) \quad \sigma_x^2 + \sigma_y^2 + \sigma_z^2 = l_j(l_j + 1)\mathbf{1}_{N_j \times N_j}$$

(see, e.g., formula (3.5.34a) in [25]).

The real scalar parameter  $J_{kl}$  in (55) is the *exchange constant* between particle  $k$  and particle  $l$ , and the real scalar parameter  $\gamma_k$  is the gyromagnetic ratio of particle  $k$ . We assume that the spins of the network have all nonzero and different gyromagnetic ratios. We can associate a graph with the model, where each node represents a particle and an edge connects two nodes if and only if the corresponding exchange constant is different from zero. It is not difficult to see that if the model is controllable, then, necessarily, this graph is connected. We shall assume this to be the case. Moreover, controllability implies observability for every output of the form (4) where  $S$  is a nonscalar matrix [10]. In our case, we presume to measure the expectation values of the total magnetization in the  $x$ ,  $y$ , and  $z$  directions given as in (4), where  $S$  is one of the matrices:

$$(58) \quad S_v := \sum_{k=1}^n I_{kv} \quad \text{with} \quad v \in \{x, y, z\}.$$

*Remark 6.1.* We have chosen to model the interaction between spins with Heisenberg interaction of the form  $A$  in (55) because we have in mind applications to molecular magnets as described in [27], which was the first motivation for this research. The Heisenberg interaction is the most common model of interaction between spins when relativistic effects (and higher order terms) are neglected (see, e.g., [29]). However, other types of interaction can be more appropriate for modeling the interaction between spins for other systems. If we consider two-body interaction, the proof of Theorem 7 still holds in the first half but the computations in the second half use explicitly the form of the interaction.

A model of the type described above will be denoted by  $\Sigma := \Sigma(n, l_j, J_{kl}, \gamma_k, \rho_0)$ , where the parameters  $n, l_j, J_{kl}, \gamma_k, \rho_0$ , which determine the model, are unknown. We will assume to have two controllable models  $\Sigma$  and  $\Sigma' := \Sigma'(n', l'_j, J'_{kl}, \gamma'_k, \rho'_0)$  which satisfy the previous requirements, and we look for necessary and sufficient conditions for these two models to be equivalent. We shall mark with a prime,  $'$ , all the quantities concerning the system  $\Sigma'$ .

**6.2. Relevant homomorphism of  $u(N)$ .** In [12] a method was described to construct a Cartan decomposition of the Lie algebra  $su(N)$  for a multipartite system, starting from decompositions of the Lie algebras  $su(N_j)$  associated with the single subsystems, each of dimension  $N_j$ , with  $N := \prod_{j=1}^n N_j$ . In particular, we have the following result.

**THEOREM 3** (see [12, section 5]). *Consider a multipartite system with  $n$  subsystems of dimensions  $N_1, \dots, N_n$ . Consider the Lie algebra  $u(N_j)$  related to the  $j$ th subsystem and a Cartan decomposition*

$$(59) \quad u(N_j) = \mathcal{K}_j \oplus \mathcal{P}_j$$



of type **AI** or **AII**. Denote by  $\sigma_j$  ( $S_j$ ) a generic element of an orthogonal basis of  $i\mathcal{K}_j$  ( $i\mathcal{P}_j$ ). Let the (total) Lie algebra  $u(N_1 N_2 \cdots N_n)$  be decomposed as

$$(60) \quad iu(N_1 N_2 \cdots N_n) = \mathcal{I}_o \oplus \mathcal{I}_e.$$

$\mathcal{I}_o$  ( $\mathcal{I}_e$ ) is the vector space spanned by matrices which are the tensor products of an odd (even) number of elements of type  $\sigma_j$ . Then  $u(N_1 N_2 \cdots N_n) = i\mathcal{I}_o \oplus i\mathcal{I}_e$  is a Cartan decomposition, i.e.,

$$(61) \quad [i\mathcal{I}_o, i\mathcal{I}_o] \subseteq i\mathcal{I}_o, \quad [i\mathcal{I}_o, i\mathcal{I}_e] \subseteq i\mathcal{I}_e, \quad [i\mathcal{I}_e, i\mathcal{I}_e] \subseteq i\mathcal{I}_o.$$

The decomposition (60) is called a decomposition of the *odd-even type*.

Associated with Cartan decomposition (61) is a Cartan involution  $\phi$  which is the identity on  $i\mathcal{I}_o$  and multiplication by  $-1$  on  $i\mathcal{I}_e$ . The structure of system (54) and (55) suggests that it is possible to choose this Cartan involution as a homomorphism mapping the equations of two equivalent models as in (12). In fact, assume that there is the same number of subsystems (spin particles) in the two models and that corresponding subsystems have the same dimensions (namely, the same spin). If we can display a decomposition (59) of type **AI** or **AII** for every (spin)  $u(N_j)$  such that  $i\sigma_{x,y,z} \in \mathcal{K}_j$ , then, for every value of the parameters, it holds that  $B_{x,y,z}(\cdot) \in i\mathcal{I}_o$  and  $A(\cdot) \in i\mathcal{I}_e$ . As shown in Theorems 4–6 following, decompositions of this type exist. We shall see in the following subsection that the Cartan involution associated with an odd-even type of Cartan decomposition is the correct homomorphism to describe classes of equivalent spin networks. In fact, not only are models which are related by such a homomorphism equivalent (according to Theorem 1) but the opposite is true as well. In other words, two equivalent models are either exactly the same or are related through such a homomorphism.

The following three theorems show the existence of a decomposition (59) of  $u(N_j)$  of type **AI** or **AII** where the subalgebra  $\mathcal{K}_j$  contains the matrices  $i\sigma_x$ ,  $i\sigma_y$ , and  $i\sigma_z$ . Equivalently, they show the existence of a subalgebra of  $sp(\frac{N_j}{2})$  (type **AII**) or  $so(N_j)$  (type **AI**) conjugate to the Lie algebra spanned by  $i\sigma_x$ ,  $i\sigma_y$ , and  $i\sigma_z$ . The proofs are presented in the following section. We shall see that the situation is different for integer and half-integer spins.

**THEOREM 4.** *If the dimension  $N_j$  of the system is even (half-integer spin (Fermions)), there exists a subalgebra of  $sp(\frac{N_j}{2})$  conjugate to the Lie algebra spanned by  $i\sigma_x$ ,  $i\sigma_y$ , and  $i\sigma_z$ .*

**THEOREM 5.** *If the dimension  $N_j$  of the system is odd (integer spin (Bosons)), there exists a subalgebra of  $so(N_j)$  conjugate to the Lie algebra spanned by  $i\sigma_x$ ,  $i\sigma_y$ , and  $i\sigma_z$ .*

**THEOREM 6.** *If the dimension  $N_j$  of the system is even (half-integer spin (Fermions)), there is no subalgebra of  $so(N_j)$  conjugate to the Lie algebra spanned by  $i\sigma_x$ ,  $i\sigma_y$ , and  $i\sigma_z$ .*

**6.3. Necessary and sufficient conditions for model equivalence.** In this subsection we shall prove the equivalence result concerning models of spin networks. This is given by the following theorem.

**THEOREM 7.** *Let  $\Sigma := \Sigma(n, l_j, J_{kl}, \gamma_k, \rho_0)$  and  $\Sigma' := \Sigma(n', l'_j, J'_{kl}, \gamma'_k, \rho'_0)$  be two given models (see (54), (55)). Assume that both models are controllable, that for model  $\Sigma$  ( $\Sigma'$ ), all the  $\gamma_k$  ( $\gamma'_k$ ) are nonzero and different from each other, and that  $\rho_0$  and  $\rho'_0$  are not both scalar matrices. Then  $\Sigma$  is equivalent to  $\Sigma'$  i.e.,*

$$(62) \quad y_v(t) := \text{Tr}(S_v \rho(t)) \equiv y'_v(t) := \text{Tr}(S'_v \rho'(t)) \quad \text{for } v \in \{x, y, z\}$$

and for every control  $u_x, u_y, u_z$ , if and only if the following condition holds:

**Condition (\*):**

1.  $n = n'$ ,

Up to a permutation of the set  $\{1, \dots, n\}$  (i.e., a permutation of the indices for the particles), the following three conditions hold:

2.

$$\gamma_k = \gamma'_k,$$

3.

$$(63) \quad l_k = l'_k,$$

and

4. one of the following two conditions holds:

(a)

$$(64) \quad A = A' \text{ and } \rho_0 = \rho'_0.$$

(b) Given the Cartan involution  $\phi$  associated with the decomposition of the odd-even type as from Theorem 3 (see also Remark 6.2 below),

$$(65) \quad A' = \phi(A) \text{ and } i\rho'_0 = \phi(i\rho_0).$$

*Remark 6.2.* The simplest way to describe the Cartan involution  $\phi$  associated with the odd-even decomposition is in terms of its action on the subspaces  $i\mathcal{I}_o$  and  $i\mathcal{I}_e$  defined in Theorem 3; that is,  $\phi$  leaves unchanged the elements of  $i\mathcal{I}_o$  and multiplies by  $-1$  the elements of  $i\mathcal{I}_e$ .  $\mathcal{I}_o$  and  $\mathcal{I}_e$  are defined in terms of the Cartan decompositions on the single subsystems. In our case, this is done according to Theorems 4 and 5. In particular consider the  $j$ th subsystem and assume that it has even dimension. Then with the unitary transformation  $U$  defined in (87),  $u(N_j)$  has the Cartan decomposition (59) of type **AII**, with  $\mathcal{K}_j$ , given by  $\mathcal{K}_j := U^* sp(N_j/2)U$  and  $\mathcal{P}_j = U^* sp(N_j/2)^\perp U$ . Analogously if the  $j$ th system has odd dimension, one considers the unitary transformation  $U$  defined in (92). Then  $u(N_j)$  has the Cartan decomposition (59) of the type **AI**, with  $\mathcal{K}_j$ , given by  $\mathcal{K}_j := U^* so(N_j)U$ , and  $\mathcal{P}_j = U^* so(N_j)^\perp U$ . According to Theorems 4 and 5, in both cases  $i\sigma_{x,y,z}$  are in  $\mathcal{K}_j$ . If we call  $\sigma_j$  a generic element of  $i\mathcal{K}_j$ ,  $\mathcal{I}_o$  ( $\mathcal{I}_e$ ) is spanned by an odd (even) number of  $\sigma_j$ 's. Notice in particular that in the model (55),  $A \in i\mathcal{I}_e$  and  $B_v \in i\mathcal{I}_o$ , so that  $\phi$  changes the sign of  $A$  and leaves the  $B_v$ 's unchanged.

Theorem 7 says that, under appropriate controllability assumptions, two equivalent models for spin networks are equivalent if and only they have the same number of particles, corresponding particles have the same spin, and their dynamical models and initial states are either exactly the same or are related through the Cartan involution associated with a decomposition of the odd-even type. In practical terms, given a general spin network, by driving the network with an external electromagnetic field and measuring the total spin in the  $x$ ,  $y$ , and  $z$  direction, it is, in principle, possible to identify the number of particles, their spin, the gyromagnetic ratios of every spin, and the exchange constants only up to a common sign factor, if the initial state is not known. The proof that Condition (\*) implies equivalence is an application of the

general property of Theorem 1. The proof that equivalence implies Condition (\*) is considerably longer. However, several results can be obtained with proofs that are formal modifications of the ones presented in [2] for the special case of spin  $\frac{1}{2}$  particles. We shall focus on the new part of the proof needed to generalize to the case of unknown spins:

*Condition (\*) implies equivalence.*

It is clear that if Condition (\*) holds with (64), then the two models differ possibly only by a permutation of the indices of the particles. So they are equivalent. Assume now that Condition (\*) holds with (65) and assume for simplicity (and without loss of generality) that the permutation of indices is the trivial permutation. Let  $\phi$  be the Cartan involution associated with the decomposition of the odd-even type. We notice that

$$(66) \quad \phi^*(iS_v) = iS_v = iS'_v, \quad v = x, y, z.$$

In fact, given any  $C \in \mathfrak{u}(N)$ , we can write  $C = C_o + C_e$ , with  $C_o \in i\mathcal{I}_o$  and  $C_e \in i\mathcal{I}_e$ . It holds that

$$\text{Tr}(\phi^*(iS_v)C) := \text{Tr}((iS_v)\phi(C)) = \text{Tr}((iS_v)(C_o - C_e)) = \text{Tr}((iS_v)C_o) = \text{Tr}((iS_v)C),$$

which, since it has to hold for every  $C$ , gives (66). Equations (65) and (66) imply that (12) of Theorem 1 holds. Since we also have (13), from (65) we conclude that the two models are equivalent using Theorem 1 as follows.

*Equivalence implies Condition (\*).*

The technique used in [2] to prove this result for networks of spin  $\frac{1}{2}$  particles extends to the general case treated here. However, further analysis is required in this case, in particular to prove that equivalent spin networks have the same values of the spins, while in [2] it was assumed that the networks were composed by all spin  $\frac{1}{2}$ 's. The main reason why the proof in [2] can be extended to this case is that the basic commutation relations, which were the essential ingredient of the proofs in [2], still hold. More precisely, the matrices  $\sigma_x$ ,  $\sigma_y$ , and  $\sigma_z$  still satisfy, for every value of the spin, the commutation relations (22). This fact implies that it also holds that

$$(67) \quad [I_{k_1 v_1, \dots, k_r v_r}, I_{\bar{k} v_{\bar{k}}}] = \begin{cases} 0 & \text{if } \bar{k} \notin \{k_1, \dots, k_r\}, \\ 0 & \text{if } \exists j \text{ with } \bar{k} = k_j \text{ and } v_{\bar{k}} = v_j, \\ iI_{k_1 v_1, \dots, k_j [v_j v_{\bar{k}}], \dots, k_r v_r} & \text{if } \exists j \text{ with } \bar{k} = k_j \text{ and } v_{\bar{k}} \neq v_j \end{cases}$$

independently of the values of the spins.<sup>6</sup>

Assume now that the two models  $\Sigma$  and  $\Sigma'$  are equivalent. Then, using exactly the same arguments as in the proof of Proposition 4.1 of [2], we obtain that the number of the spin particle must be the same, namely,  $n = n'$ , and, up to a permutation of the indices,  $\gamma_k = \gamma'_k$  for all  $k \in \{1, \dots, n\}$ , which is parts 1 and 2 of Condition (\*). Moreover, as in Proposition 4.1 of [2], we obtain

$$(68) \quad \text{Tr}(I_{kv}\rho(t)) = \text{Tr}(I'_{kv}\rho'(t)) \quad \forall k \in \{1, \dots, n\}, \quad \forall v \in \{x, y, z\}.$$

<sup>6</sup>In the notation used here  $[v\bar{v}]$  give a result in agreement with the commutator of  $i\sigma_v$  and  $i\sigma_{\bar{v}}$  (cf. (22)). Therefore, for example,  $[xy] := z$ . Moreover,  $I_{k_1 v_1, \dots, k_j (-v_j), \dots, k_r v_r} := -I_{k_1 v_1, \dots, k_j (v_j), \dots, k_r v_r}$ .

Here  $I'_{kv}$  is defined as  $I_{kv}$  but for  $\Sigma'$  and, at this point, it may be different from  $I_{kv}$  since we have not shown yet that corresponding spins must be equal. To prove this fact, we shall use Lemma 8 below. The proof of this lemma is a generalization of the proof of Lemma 5.2 in [2], where we use the general property (57) instead of the corresponding property for spin  $\frac{1}{2}$ 's. We postpone this proof to Appendix A.

LEMMA 8. *Assume that for all  $t \geq 0$ , all possible trajectories  $\rho(t)$  of  $\Sigma$ , and corresponding  $\rho'(t)$  of  $\Sigma'$ , for fixed values  $1 \leq k_1, \dots, k_r \leq n$ ,  $v_j \in \{x, y, z\}$  and for given constants  $\beta$  and  $\beta'$ , we have*

$$(69) \quad \beta \text{Tr} (I_{k_1 v_1, \dots, k_r v_r} \rho(t)) = \beta' \text{Tr} (I'_{k_1 v_1, \dots, k_r v_r} \rho'(t)).$$

Then

1. *for any pair of indices  $\bar{k}, \bar{d} \in \{1, \dots, n\}$  with  $\bar{k} \in \{k_1, \dots, k_r\}$  and  $\bar{d} \notin \{k_1, \dots, k_r\}$ ,*

$$(70) \quad \beta J_{\bar{k}\bar{d}} \text{Tr} (I_{k_1 v_1, \dots, k_r v_r, \bar{d} \bar{v}} \rho(t)) = \beta' J'_{\bar{k}\bar{d}} \text{Tr} (I'_{k_1 v_1, \dots, k_r v_r, \bar{d} \bar{v}} \rho'(t))$$

*for any value  $\bar{v} \in \{x, y, z\}$ .*

2. *For any pair of indices  $\bar{k}, \bar{d}$  both in  $\{k_1, \dots, k_r\}$  (for example,  $\bar{k} = k_1, \bar{d} = k_2$ ),*

$$(71) \quad \begin{aligned} & \beta (l_{\bar{d}}(l_{\bar{d}} + 1)) J_{\bar{k}\bar{d}} \text{Tr} (I_{k_1 v_1, k_3 v_3, \dots, k_r v_r} \rho(t)) \\ &= \beta' (l'_{\bar{d}}(l'_{\bar{d}} + 1)) J'_{\bar{k}\bar{d}} \text{Tr} (I'_{k_1 v_1, k_3 v_3, \dots, k_r v_r} \rho'(t)). \end{aligned}$$

In other words, formula (71) means that from (69), it is possible to derive a new formula as follows. Select two indices in the set  $\{k_1, \dots, k_r\}$ ,  $\bar{k}$  and  $\bar{d}$ . One of the two indices (say,  $\bar{d}$ ) disappears from the subscript in the matrices  $I$  and corresponding  $I'$ . However, a coefficient  $l_{\bar{d}}(l_{\bar{d}} + 1)$  and  $l'_{\bar{d}}(l'_{\bar{d}} + 1)$  appears in the left- and right-hand side, respectively, as well as a coefficient  $J_{\bar{k}\bar{d}}$  and  $J'_{\bar{k}\bar{d}}$ .

Analogously, for formula (70) one selects two indices  $\bar{k}$  and  $\bar{d}$  corresponding to two given particles, with  $\bar{k} \in \{k_1, \dots, k_r\}$  and  $\bar{d} \notin \{k_1, \dots, k_r\}$ .  $J_{\bar{k}\bar{d}}$  and  $J'_{\bar{k}\bar{d}}$  denote the coupling constants between the  $\bar{k}$ th and  $\bar{d}$ th particle in the two models. Formula (70) is in fact three formulas, one for each value of  $\bar{v} = x, y, z$ .

We shall now prove that, under the assumption of equivalence, the squares of the exchange constants  $J_{dk}$  and  $J'_{dk}$  must be proportional, with a proportionality factor common to all pairs of indices  $d$  and  $k$ , and this will also be instrumental in the proof of part 3 of Condition (\*).

Fix any  $1 \leq k_1 < k_2 \leq n$ . Then, by applying statement 1 of Lemma 8, i.e., (70) with  $\bar{k} = k_1, \bar{d} = k_2$ , to (68) with  $k = k_1$ , we have

$$(72) \quad J_{k_1 k_2} \text{Tr} (I_{k_1 v_1, k_2 v_2} \rho(t)) = J'_{k_1 k_2} \text{Tr} (I'_{k_1 v_1, k_2 v_2} \rho'(t)) \quad \forall v_1, v_2 \in \{x, y, z\}.$$

Now, to the previous equality we apply statement 2 of Lemma 8, i.e., (71) with  $\bar{k} = k_1$  and  $\bar{d} = k_2$ , to get

$$(l_{k_2}(l_{k_2} + 1)) J_{k_1 k_2}^2 \text{Tr} (I_{k_1 v_1} \rho(t)) = (l'_{k_2}(l'_{k_2} + 1)) J_{k_1 k_2}'^2 \text{Tr} (I'_{k_1 v_1} \rho'(t)),$$

which, by (68), implies

$$(73) \quad (l_{k_2}(l_{k_2} + 1)) J_{k_1 k_2}^2 = (l'_{k_2}(l'_{k_2} + 1)) J_{k_1 k_2}'^2.$$

Using the facts that the two indices  $k_1$  and  $k_2$  above are arbitrary and that the graph associated with the network is connected, by the controllability assumption (cf. the

discussion at the end of subsection 6.1) it is easy to see that there exists a positive constant  $\alpha \in \mathbb{R}$  such that, for all  $1 \leq d < k \leq n$ ,

$$(74) \quad J_{dk}^2 = \alpha^2 J'_{dk}{}^2 \quad \text{and} \quad l_k(l_k + 1) = \frac{1}{\alpha^2} l'_k(l'_k + 1).$$

Using (74), we can now prove part 3 of Condition (\*). We will do this using some lemmas and arguing by contradiction. First, notice that from (74), we have that if there exists a  $\bar{k} \in \{1, \dots, n\}$  such that  $l_{\bar{k}} = l'_{\bar{k}}$ , then necessarily  $\alpha^2 = 1$ , and thus  $l_j = l'_j$  for all  $j = 1, \dots, n$ , namely, all the particles have the same spin. So if we assume that (63) does not hold, without loss of generality, we can assume  $l_1 > l'_1$ . Using (74), we get that  $l_j > l'_j$  for all  $j = 1, \dots, n$ , and thus also  $N_j > N'_j$ . Let  $R := \frac{N}{N_1} = \prod_{j=2}^n N_j$  and  $R' := \frac{N'}{N'_1} = \prod_{j=2}^n N'_j$ .

LEMMA 9. *For all  $t \in \mathbb{R}$  and all the admissible trajectories  $\rho$  and corresponding trajectories  $\rho'$ , we have*

$$(75) \quad \begin{aligned} & Tr \left( (e^{i\sigma_z t} \otimes \mathbf{1}_{R \times R}) I_{1v} (e^{-i\sigma_z t} \otimes \mathbf{1}_{R \times R}) \rho(s) \right) \\ &= Tr \left( (e^{i\sigma_z t} \otimes \mathbf{1}_{R' \times R'}) I'_{1v} (e^{-i\sigma_z t} \otimes \mathbf{1}_{R' \times R'}) \rho'(s) \right) \end{aligned}$$

for all  $s \geq 0$ .

*Proof.* First, we notice that from the Campbell–Baker–Hausdorff formula, we have

$$(76) \quad (e^{i\sigma_z t} \otimes \mathbf{1}_{R \times R}) I_{1v} (e^{-i\sigma_z t} \otimes \mathbf{1}_{R \times R}) = \sum_{k=0}^{\infty} \left( ad_{i\sigma_z \otimes \mathbf{1}_{R \times R}}^k I_{1v} \right) \frac{t^k}{k!} \quad \forall v \in \{x, y, z\}$$

and an analogous equation for  $\Sigma'$ . Moreover, by applying Lemma 11 in Appendix A, with  $W = I_{1v}$ ,  $W' = I'_{1v}$ , and  $k = 1$ ,  $v = z$ , we have

$$Tr \left( ad_{i\sigma_z \otimes \mathbf{1}_{R \times R}} I_{1v} \rho(s) \right) = Tr \left( ad_{i\sigma_z \otimes \mathbf{1}_{R' \times R'}} I'_{1v} \rho'(s) \right).$$

Now we can apply again Lemma 11 to the previous equality to get

$$Tr \left( ad_{i\sigma_z \otimes \mathbf{1}_{R \times R}}^2 I_{1v} \rho(s) \right) = Tr \left( ad_{i\sigma_z \otimes \mathbf{1}_{R' \times R'}}^2 I'_{1v} \rho'(s) \right).$$

By applying this procedure repeatedly we obtain

$$Tr \left( ad_{i\sigma_z \otimes \mathbf{1}_{R \times R}}^k I_{1v} \rho(s) \right) = Tr \left( ad_{i\sigma_z \otimes \mathbf{1}_{R' \times R'}}^k I'_{1v} \rho'(s) \right)$$

for all  $k \geq 0$ . Using this in (76), equation (75) follows.  $\square$

The proof of the following lemma is given in Appendix A.

LEMMA 10. *The following formula holds:*

$$(77) \quad (e^{i\sigma_z t} \otimes \mathbf{1}_{R \times R}) I_{1x} (e^{-i\sigma_z t} \otimes \mathbf{1}_{R \times R}) := P_{N_1}(t) \otimes \mathbf{1}_{R \times R},$$

where the matrix  $P_{N_1}(\cdot)$  is periodic with period  $2\pi$ . Moreover,

$$(78) \quad P_{N_1}(\pi) = -P_{N_1}(0) = -\sigma_x.$$

Using Lemmas 9 and 10, we can now conclude the proof that the spins are the same. Let  $\bar{\rho}(s) \otimes \mathbf{1}_{R \times R}$  (resp.,  $\bar{\rho}'(s) \otimes \mathbf{1}_{R' \times R'}$ ) be the orthogonal component of  $\rho(s)$

(resp.,  $\rho'(s)$ ) along  $\sigma_x \otimes \mathbf{1}_{R \times R}$  (resp.,  $\sigma_x \otimes \mathbf{1}_{R' \times R'}$ ). Using (77), equality (75) with  $v = x$  can be written as

$$(79) \quad \text{Tr} (P_{N_1}(t)\bar{\rho}(s)) R = \text{Tr} (P_{N'_1}(t)\bar{\rho}'(s)) R'.$$

Since we have assumed by contradiction  $R > R'$ , from (79) we have for every  $t$ ,

$$(80) \quad \text{Tr} (P_{N_1}(t)\bar{\rho}(s)) < \text{Tr} (P_{N'_1}(t)\bar{\rho}'(s)).$$

Now we will derive a contradiction by evaluating the previous inequality at  $t = 0$  and  $t = \pi$  and using (78). In fact we have

$$\text{Tr} (P_{N_1}(0)\bar{\rho}(s)) < \text{Tr} (P_{N'_1}(0)\bar{\rho}'(s)),$$

and thus

$$\text{Tr} (P_{N_1}(\pi)\bar{\rho}(s)) = -\text{Tr} (P_{N_1}(0)\bar{\rho}(s)) > -\text{Tr} (P_{N'_1}(0)\bar{\rho}'(s)) = \text{Tr} (P_{N'_1}(\pi)\bar{\rho}'(s)).$$

The previous inequality contradicts (80). Thus we conclude that  $l_1 = l'_1$ , which implies that (63) holds.

Since the two equivalent models  $\Sigma$  and  $\Sigma'$  have the same spin, the positive constant  $\alpha$  in (74) is equal to one. Therefore, for every pair  $d, k \in \{1, \dots, n\}$ ,  $J_{dk}$  and  $J'_{dk}$  only differ possibly by the sign factor. Using the same argument as in the main theorem of [2] we can in fact conclude that there are only two possible cases: The case where  $J_{dk} = J'_{dk}$  for every pair  $d, k$ , and the case where  $J_{dk} = -J'_{dk}$  for every pair  $d, k$ . If we are in the first case, then from the observability (which follows from controllability) of the model, we must have  $\rho_0 = \rho'_0$ , and thus (64) holds. This is case (a) of part 4 of Condition (\*). On the other hand, if  $J'_{kd} = -J_{kd}$  for every pair  $1 \leq k < d \leq n$ , we may conclude using Theorem 1. In fact we consider the homomorphism  $\phi$  given by the Cartan involution associated with the odd-even decomposition as in the previous part of the proof. Conditions (12) hold, and thus, since the models are equivalent and observable, we get that

$$i\rho'_0 = \phi(i\rho_0),$$

and thus (65) holds. This concludes the proof of the theorem.

**7. Proofs of Theorems 4–6.** In the proofs of Theorems 4 and 5, we shall use the following two types of elementary  $k \times k$  matrices:

$$(81) \quad C_k = \text{diag}(-1, 1, -1, \dots, (-1)^k), \quad T_k = \text{adiag}(1, 1, 1, \dots, 1) = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}.$$

Thus, the matrix  $C_k$  is diagonal with alternating elements while  $T_k$  is antidiagonal with all ones on the secondary diagonal and zeros everywhere else. Obvious properties of these matrices are the following:

$$(82) \quad C_k^2 = T_k^2 = \mathbf{1}_{k \times k}, \quad T_k = T_k^T.$$

We are interested in the action of these matrices by similarity transformation on diagonal and tridiagonal  $k \times k$  matrices.<sup>7</sup> In particular, let us denote by  $D$  a generic, real, diagonal,  $k \times k$  matrix and by  $F$  a generic, real,  $k \times k$ , tridiagonal matrix which is also symmetric and has zero diagonal; thus  $F$  will be of the type

$$F = \begin{pmatrix} 0 & a_1 & 0 & 0 & \cdots & 0 \\ a_1 & 0 & a_2 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & a_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{k-2} & 0 & a_{k-1} \\ 0 & 0 & \cdots & 0 & a_{k-1} & 0 \end{pmatrix}.$$

If  $M^a$  denotes the antitransposed of  $M$ , namely, the matrix obtained by reflecting about the secondary diagonal, we can easily verify the following properties:

1.

$$(83) \quad C_k D C_k = D, \quad C_k F C_k = -F,$$

2.

$$(84) \quad T_k D T_k = D^a, \quad T_k F T_k = F^a.$$

*Remark 7.1.* In the proofs of Theorems 4 and 5 below we do not need to write explicitly the matrices  $\sigma_x$ ,  $\sigma_y$ , and  $\sigma_z$ , which are the representation of the  $x$ ,  $y$ , and  $z$  components of spin angular momentum; nevertheless we need to write their structure. This structure is given by (85) and (86) for half-integer spin and by (90) and (91) for integer spin (see, e.g., [25, section 3.5] and Appendix B). In both proofs we make the explicit computation only for  $i\sigma_x$  and  $i\sigma_z$ ; this is possible since the third of the commutation relations (22) gives  $i\sigma_y = [i\sigma_z, i\sigma_x]$ .

Now we are ready to prove Theorems 4 and 5.

*Proof of Theorem 4.* The matrices  $i\sigma_z$  and  $i\sigma_x$  have (for every value of the spin) the following structure:

$$(85) \quad i\sigma_z = i \begin{pmatrix} D & 0 \\ 0 & -D^a \end{pmatrix},$$

$$(86) \quad i\sigma_x = i \begin{pmatrix} F & P \\ P^T & F^a \end{pmatrix},$$

where  $F$  and  $D$  have the structure above specified with  $k := \frac{N_j}{2}$ , and  $P$  is a  $k \times k$  real matrix of all zeros except in the  $(k, 1)$ st position. Now use

$$(87) \quad U := \begin{pmatrix} C_k & 0 \\ 0 & T_k \end{pmatrix},$$

which is orthogonal and therefore unitary.

We calculate, using the first terms of (83) and (84),

$$(88) \quad i\tilde{\sigma}_z := U i\sigma_z U^* = i \begin{pmatrix} D & 0 \\ 0 & -D \end{pmatrix}.$$

<sup>7</sup>A matrix  $F$  is tridiagonal if  $f_{ij} = 0$  when  $|i - j| > 1$ .

Moreover, using the second terms of (83) and (84), we have

$$(89) \quad i\tilde{\sigma}_x := Ui\sigma_xU^* = i \begin{pmatrix} -F & C_kPT_k \\ T_kPC_k & F \end{pmatrix}.$$

It is easily seen that  $i\tilde{\sigma}_z$  and  $i\tilde{\sigma}_x$  are symplectic by observing that  $C_kPT_k$  is a real symmetric matrix (only the  $(k, k)$ th element is different from zero). Therefore  $\text{sp}(\frac{N_j}{2})$  contains a subalgebra conjugate to the one spanned by  $i\sigma_x$  and  $i\sigma_z$ , and therefore  $i\sigma_y$  and the theorem is proved.  $\square$

We now proceed to the proof of Theorem 5.

*Proof of Theorem 5.* In this case we set  $k := \frac{N_j-1}{2}$ . The matrix  $i\sigma_z$  has the form

$$(90) \quad i\sigma_z := i \begin{pmatrix} -D & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & D^a \end{pmatrix},$$

with  $D$  of dimension  $k \times k$ . Moreover,  $i\sigma_x$  has the form

$$(91) \quad i\sigma_x := i \begin{pmatrix} F & v & 0 \\ v^T & 0 & w^t \\ 0 & w & F^a \end{pmatrix},$$

where  $F$  is as above and  $v$  ( $w$ ) is a vector of dimension  $k$  with only the last (the first) component different from zero, and the components different from zero are equal for  $v$  and  $w$ . We use the unitary matrix

$$(92) \quad U := \begin{pmatrix} \frac{i}{\sqrt{2}}C_k & 0 & (-1)^k \frac{i}{\sqrt{2}}T_k \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}}T_k & 0 & \frac{1}{\sqrt{2}}C_k \end{pmatrix},$$

which is easily seen to be unitary by (81) and (82). We calculate

$$(93) \quad \begin{aligned} i\tilde{\sigma}_z &:= Ui\sigma_zU^* \\ &= i \begin{pmatrix} \frac{1}{2}(T_kD^aT_k - C_kDC_k) & 0 & \frac{i}{2}((-1)^kT_kD^aC_k - C_kDT_k) \\ 0 & 0 & 0 \\ \frac{i}{2}(T_kDC_k - (-1)^kC_kD^2T_k) & 0 & \frac{1}{2}(C_kD^aC_k - T_kDT_k) \end{pmatrix}. \end{aligned}$$

Using the first terms of (83) and (84), we find that the diagonal blocks are zero. Moreover, the remaining elements of the matrix are real so that  $i\tilde{\sigma}_z$  is real. Analogously, we calculate

$$(94) \quad \begin{aligned} i\tilde{\sigma}_x &:= Ui\sigma_xU^* \\ &= i \begin{pmatrix} \frac{1}{2}(C_kFC_k + (-1)^{2k}T_kF^aT_k) & \frac{i}{\sqrt{2}}(C_kv + (-1)^kT_kw) & \frac{i}{2}(C_kFT_k + (-1)^kT_kF^aC_k) \\ * & 0 & \frac{1}{\sqrt{2}}(v^TT_k + w^TC_k) \\ * & * & \frac{1}{2}(T_kFT_k + C_kF^aC_k) \end{pmatrix}, \end{aligned}$$

where we have denoted by  $*$  the components that can be obtained from the requirement that the matrix is skew-Hermitian. Now the  $(1, 1)$  and  $(3, 3)$  blocks are zero from the second terms of properties (83) and (84), while the  $(2, 3)$  block is zero because of



the structure of the vectors  $v$  and  $w$ . All the other blocks are purely real matrices so that  $i\tilde{\sigma}_x$  is also in  $so(N)$ , and this completes the proof.  $\square$

We now give the proof of the negative result in Theorem 6.

*Proof of Theorem 6.* Assume that there exists a matrix  $X \in SU(N_j)$  such that

$$(95) \quad Xi\sigma_x X^* := \tilde{R}_x,$$

$$(96) \quad Xi\sigma_y X^* := \tilde{R}_y,$$

$$(97) \quad Xi\sigma_z X^* := \tilde{R}_z,$$

with  $\tilde{R}_x$ ,  $\tilde{R}_y$ , and  $\tilde{R}_z$  in  $so(N_j)$ . Then we can use the **AI** Cartan decomposition of  $SU(N_j)$  [16] to write  $X$  as

$$(98) \quad X = K_1 A K_2,$$

with  $K_1$  and  $K_2$  in  $SO(N_j)$  and  $A$  diagonal, i.e.,

$$(99) \quad A := \text{diag} \left( e^{i\phi_1}, \dots, e^{i\phi_{\frac{N_j}{2}}} \right).$$

Therefore we can write

$$(100) \quad K_1 A K_2 i\sigma_{x,y,z} K_2^T \bar{A} K_1^T = \tilde{R}_{x,y,z},$$

or, defining  $R_{x,y,z} := K_1^T \tilde{R}_{x,y,z} K_1$ , which is also real skew-symmetric, we can write

$$(101) \quad K_2 i\sigma_{x,y,z} K_2^T = \bar{A} R_{x,y,z} A.$$

The real matrices  $R_{x,y,z}$  must satisfy the same basic commutation relations (22) of  $i\sigma_x$ ,  $i\sigma_y$ , and  $i\sigma_z$  and have the same eigenvalues of  $i\sigma_x$ ,  $i\sigma_y$ , and  $i\sigma_z$ , namely, for a (half-integer) spin  $j$ ,  $\pm ji$ ,  $\pm(j+1)i$ ,  $\dots$ ,  $\pm\frac{1}{2}i$ . We now study the structure of  $R_{x,y,z}$  in (101) and get a contradiction with these facts.

First notice that, since  $A$  is diagonal as in (99), the action of  $A$  on the right-hand side of (101), namely,  $R \rightarrow \bar{A} R A$ , changes the (real) element  $r_{jk}$  of  $R$  into  $r_{jk} e^{-i(\phi_j - \phi_k)}$ . Since the entries on the left-hand side of (101) are either all purely imaginary or all purely real, if  $\phi_j - \phi_k$  is not a multiple of  $\frac{\pi}{2}$ , then we must have  $r_{jk} = 0$ . Consider the indices  $1, \dots, N_j$  and let  $\mathcal{O}$  be the set of indices  $k$  such that  $\phi_1 - \phi_k$  is an odd multiple of  $\frac{\pi}{2}$  and  $\mathcal{E}$  the set of indices  $k$  such that  $\phi_1 - \phi_k$  is an even multiple of  $\frac{\pi}{2}$ , and  $\mathcal{N}$  the set of indices  $k$  such that  $\phi_1 - \phi_k$  is not an integer multiple of  $\frac{\pi}{2}$ .

From (101) it follows that since  $i\sigma_y$  is real, the terms  $r_{jk}$  of  $R_y$ , where  $j$  and  $k$  belong to different sets, must be zero because in that case  $e^{i(\phi_j - \phi_k)}$  in (99) has a nonzero imaginary part. Therefore only the elements  $r_{jk}$ , where both  $j$  and  $k$  belong to  $\mathcal{O}$  or  $\mathcal{E}$  or  $\mathcal{N}$ , are possibly different from zero. Therefore after possibly reordering rows and columns, which corresponds to a similarity transformation by a permutation matrix,  $R_y$  must be of block diagonal form, and without loss of generality and for simplicity we shall assume only two blocks (rather than three). Therefore we write

$$(102) \quad R_y := \begin{pmatrix} Y_{11} & 0 \\ 0 & Y_{22} \end{pmatrix},$$

where  $Y_{11}$  has dimensions  $n_o \times n_o$  with  $n_o$  the cardinality of  $\mathcal{O}$ , and  $Y_{22}$  has dimension  $(N_j - n_o) \times (N_j - n_o)$ . Both  $Y_{11}$  and  $Y_{22}$  are skew-symmetric matrices. An analogous argument shows that, after possibly the same reordering of column and row indices,  $R_z$  can be written as

$$(103) \quad R_z := \begin{pmatrix} 0 & Z_{12} \\ -Z_{12}^T & 0 \end{pmatrix},$$

where  $Z_{12}$  is a general matrix of dimensions  $n_o \times (N_j - n_o)$ .

Now consider the possible values for  $n_o$ .  $n_o$  odd is to be excluded because this would cause  $\det Y_{11} = 0$  in (102), and this contradicts the fact that  $R_y$  has no zero eigenvalues. Moreover,  $n_o \neq (N_j - n_o)$  (i.e.,  $n_o \neq \frac{N_j}{2}$ ) would cause  $R_z$  to have a determinant equal to zero. This can be easily verified by calculating

$$(104) \quad \det(R_z^2) = (\det R_z)^2 = \det \begin{pmatrix} -Z_{12}Z_{12}^T & 0 \\ 0 & -Z_{12}^TZ_{12} \end{pmatrix}$$

since, in this case, at least one of the matrices on the diagonal blocks does not have full rank. These considerations already exclude the cases where  $\frac{N_j}{2}$  is an odd number as for spins  $\frac{1}{2}$ ,  $\frac{5}{2}$ ,  $\frac{9}{2}$ , etc., and we can assume  $R_y$  and  $R_z$  of the form (102) and (103) with  $n_o = \frac{N_j}{2}$ . To obtain a contradiction in this case too, we first notice that, since  $Y_{11}$  and  $Y_{22}$  have even dimension and are skew-symmetric, we can apply a similarity transformation  $T := \begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix}$ , with  $T_1$  and  $T_2$  orthogonal so that  $TR_yT^T$  is block diagonal,

$$(105) \quad TR_yT^T = \left( D_1, D_2, \dots, D_{\frac{N_j}{2}} \right),$$

where the  $2 \times 2$  block  $D_k$  has the form

$$(106) \quad D_k := \begin{pmatrix} 0 & l_k \\ -l_k & 0 \end{pmatrix},$$

where each  $l_k$  corresponds to a pair of complex conjugate eigenvalues of  $R_y$  so that  $l_k = \frac{p}{2}$  with  $p$  odd corresponds to the pair  $\pm \frac{p}{2}i$ . Moreover, we choose  $T$  so that the first  $\frac{N_j}{4}$  blocks are ordered according to the increasing value of  $l_k$ , and the same holds for the last  $\frac{N_j}{4}$  blocks. We shall therefore assume this structure of  $R_y$  in the remainder of the proof. We notice also that the transformation  $TR_zT^T$  does not change the structure of  $R_z$ , as  $Z_{12}$  in (103) was chosen to be a general  $\frac{N_j}{2} \times \frac{N_j}{2}$  real matrix. Express  $Z_{12}$  in terms of  $2 \times 2$  blocks  $L_{fk}$ ,  $f, k = 1, \dots, \frac{N_j}{4}$ ,  $k = \frac{N_j}{4} + 1, \dots, \frac{N_j}{2}$ , which is possible since  $\frac{N_j}{2}$  is an even number. Now, we impose the fact that  $R_y$  and  $R_z$  have to satisfy the same commutation relations as  $i\sigma_y$  and  $i\sigma_z$ . In particular, we must have

$$(107) \quad [[R_y, R_z], R_y] = R_z.$$

This equation gives the following for the  $L_{fk}$  block:

$$(108) \quad 2D_f L_{fk} D_k - L_{fk} D_k^2 - D_f^2 L_{fk} = L_{fk}.$$

If we write the generic  $L_{fk}$  as

$$(109) \quad L_{fk} := \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix},$$

and recall the structure of  $D_f$  and  $D_k$ ,

$$(110) \quad D_f := \begin{pmatrix} 0 & l_f \\ -l_f & 0 \end{pmatrix}, \quad D_k := \begin{pmatrix} 0 & l_k \\ -l_k & 0 \end{pmatrix},$$

we obtain the following equations for  $a_2$  and  $a_3$  (and analogous equations for  $a_1$  and  $a_4$ ):

$$(111) \quad 2l_k l_f a_3 = (1 - l_f^2 - l_k^2) a_2,$$

$$(112) \quad 2l_k l_f a_2 = (1 - l_f^2 - l_k^2) a_3.$$

Combining these, we obtain

$$(113) \quad 4l_k^2 l_f^2 a_3 = (1 - l_k^2 - l_f^2)^2 a_3,$$

which shows, taking the square root of both sides, that the only possible ways to have  $a_3$  and therefore  $a_2$  different from zero are the cases  $l_f + l_k = \pm 1$ . In these cases we can easily see that

$$(114) \quad a_3 = -a_2.$$

Similarly, one finds that we have  $a_4$  and  $a_1$  in (109) different from zero if and only if  $l_f + l_k = \pm 1$ , and in these cases we have

$$(115) \quad a_1 = a_4.$$

In conclusion, all the blocks  $L_{fk}$  are zero except the ones corresponding to indices  $f$  and  $k$  with neighboring values of  $l_f$  and  $l_k$ , which have the structure

$$(116) \quad L_{fk} := \begin{pmatrix} x & y \\ -y & x \end{pmatrix}.$$

Therefore  $R_z$  has the form in (103) where the  $f$ th block row of  $Z_{12}$  has at most two blocks different from zero and with the structure in (116). We denote these blocks by  $P_f$  and  $S_f$ , where  $P$  ( $S$ ) stands for “predecessor” (“successor”) and corresponds to the index  $k$  such that  $l_k = l_f - 1$  and  $l_k = l_f + 1$ , respectively. Now, we argue that a matrix  $R_z$  with this structure must necessarily have all the (purely imaginary) eigenvalues with multiplicity at least two, and this gives the desired contradiction because  $R_z$  should have the same spectrum of  $i\sigma_{x,y,z}$ , which consists of all simple eigenvalues. In order to see this fact, reconsider the block structure of  $R_y$  in (105). If the blocks corresponding to eigenvalues  $\pm \frac{1}{2}i$  and  $\pm \frac{3}{2}i$  belong to the same half, then the corresponding matrix  $R_z$  will have a two-dimensional block row (or column) equal to zero, and therefore 0 will be an eigenvalue with multiplicity at least 2. Therefore we can assume that these two blocks belong to two different halves, and by the ordering we have imposed they must be the first ones of each half. Assume that the block corresponding to  $\pm \frac{1}{2}i$  is in the first half. If this is not the case, consider the transpose of  $R_z$  and repeat the arguments that follow. It is possible to choose a block diagonal similarity transformation

$$(117) \quad U := \text{diag} \left( G_1, G_2, \dots, G_{\frac{N_j}{4}}, F_1, F_2, \dots, F_{\frac{N_j}{4}} \right),$$

with all the  $G_f$ 's and  $F_f$ 's being  $2 \times 2$  orthogonal matrices so that  $UR_zU^T$  has the same structure as before, but all the matrices  $P_j$  and  $S_j$  are *scalar* matrices. We construct the matrix  $U$  proceeding by block rows. The first block row contains only  $S_1$ , as  $\frac{1}{2}$  has no predecessors. All the zero blocks remain zero and  $S_1$  is transformed into

$$(118) \quad G_1 S_1 F_1^T.$$

We choose  $F_1 = \mathbf{1}_{2 \times 2}$  and  $G_1$ , which has the general form

$$(119) \quad G_1 := \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix},$$

so that  $\sin(\theta)x + \cos(\theta)y = 0$  if  $S_1 := \begin{pmatrix} x & y \\ -y & x \end{pmatrix}$ . This will give a scalar matrix. At the generic  $f$ th block row, we have at most two nonzero blocks  $P_{fk}$  and  $S_{fb}$ , where we now use an extra index  $k$  and  $b$  to indicate the block column to which they belong. They transform into

$$(120) \quad P_{fk} \rightarrow G_f P_{fk} F_k^T$$

and

$$(121) \quad S_{fb} \rightarrow G_f S_{fb} F_b^T,$$

respectively, while all the other blocks remain zero. If  $F_k$  has not been chosen before, we set  $F_k = \mathbf{1}_{2 \times 2}$ . In any case, we choose  $G_f$  as before to make  $G_f P_{fk} F_k^T$  a scalar matrix. We then choose  $F_b$  to make  $G_f S_{fb} F_b^T$  a scalar matrix.  $G_f$  and  $F_b$  had not been chosen at previous steps. This is obvious for  $G_f$  and follows by an induction argument for  $F_b$  since all the  $F$  matrices chosen before the  $f$ th step correspond to predecessors and successors with (column) indices strictly less than  $b$  (recall that in the two halves of the matrix  $R_y$  the blocks are arranged in increasing order of (absolute value of) eigenvalue). In conclusion, modulo the similarity transformation defined by  $U$  in (117), we can assume that  $R_z$  has the form

$$(122) \quad R_z = K \otimes I_{2 \times 2},$$

where  $K$  is a skew-symmetric  $\frac{N_j}{2} \times \frac{N_j}{2}$  matrix. By known results on the eigenvalues of the Kronecker products of two matrices, it follows that the eigenvalues of  $R_z$  are the same as those of  $K$ , each with multiplicity at least 2. This gives the desired contradiction and concludes the proof of the theorem.  $\square$

**8. Conclusions.** This paper has presented a collection of mathematical results concerning the input-output equivalence of quantum systems. Models that are equivalent cannot be distinguished by an external observer and therefore the determination of parameters in a quantum Hamiltonian can be obtained only up to equivalent models. Motivated by recent results on the isospectrality of quantum Hamiltonians [27] in molecular magnets, we have completely characterized the classes of spin networks which are equivalent. In several cases, the characterization of equivalent models can be obtained through a Lie algebra homomorphism, which is suggested by a Cartan structure of the underlying dynamics.

We believe many of the results and the concepts presented in this paper for quantum systems could be generalized to classes of systems relevant in other applications with both dynamics and output linear in the state. This will be the subject of further research.

**Appendix A: Additional results and proofs.** The proof of the following lemma can be obtained with a formal modification of the proof of Lemma 4.4 in [2] and is therefore omitted.

LEMMA 11. *Let  $\Sigma$  and  $\Sigma'$  be two equivalent models. If  $W$  and  $W'$  are two given Hermitian matrices such that*

$$(123) \quad \text{Tr}(W\rho(t)) = \text{Tr}(W'\rho'(t))$$

*for every pair of corresponding trajectories  $\rho(t)$  and  $\rho'(t)$ , then it also holds that*

$$(124) \quad \text{Tr}([W, I_{kv}]\rho(t)) = \text{Tr}([W', I'_{(k)v}]\rho'(t)) \quad \forall k \in \{1, \dots, n\}, \quad \forall v \in \{x, y, z\},$$

*up to a permutation of the indices.*<sup>8</sup>

*Proof of Lemma 8.* We first state a lemma whose proof can be obtained from the proof of Lemma 5.1 in [2] and then proceed to the proof of Lemma 8.

LEMMA 12. *Assume that for all  $t \geq 0$ , all the possible trajectories  $\rho(t)$  of  $\Sigma$  and corresponding  $\rho'(t)$  of  $\Sigma'$ , for fixed values  $1 \leq k_1, \dots, k_r \leq n$ , and fixed  $v_j \in \{x, y, z\}$ , we have*

$$(125) \quad \text{Tr}(I_{k_1 v_1, \dots, k_r v_r} \rho(t)) = \text{Tr}(I'_{k_1 v_1, \dots, k_r v_r} \rho'(t)).$$

*Then*

1. *equation (125) holds for any possible choice of the values of  $v_j \in \{x, y, z\}$ ;*
- 2.

$$(126) \quad \begin{aligned} & \text{Tr} \left( \left[ \left[ iI_{\bar{d}v_{\bar{d}}}, [iI_{\bar{k}v_{\bar{k}}}, A] \right], I_{k_1 v_1, \dots, k_r v_r} \right] \rho(t) \right) \\ &= \text{Tr} \left( \left[ \left[ iI'_{\bar{d}v_{\bar{d}}}, [iI'_{\bar{k}v_{\bar{k}}}, A'] \right], I'_{k_1 v_1, \dots, k_r v_r} \right] \rho'(t) \right) \end{aligned}$$

*for all the indices  $1 \leq \bar{d} \neq \bar{k} \leq n$  and every  $\{v_{\bar{d}} \neq v_{\bar{k}}\} \in \{x, y, z\}$ .*

We now proceed to the proof of Lemma 8. First notice that from Lemma 12, it is enough to prove (70) and (71) for a particular choice of  $\{v_j\}$  and  $\bar{v}$ . Moreover, we have, for  $\bar{d} > \bar{k}$ ,

$$(127) \quad [iI_{\bar{d}z}, [iI_{\bar{k}x}, A]] = -J_{\bar{k}\bar{d}} iI_{\bar{k}z, \bar{d}x}.$$

1. By applying Lemma 12 (equation (126)) to (69) and using (127) we get:

$$(128) \quad \beta \text{Tr} \left( [-J_{\bar{k}\bar{d}} iI_{\bar{k}z, \bar{d}x}, I_{k_1 v_1, \dots, k_r v_r}] \rho(t) \right) = \beta' \text{Tr} \left( \left[ -J'_{\bar{k}\bar{d}} iI'_{\bar{k}z, \bar{d}x}, I'_{k_1 v_1, \dots, k_r v_r} \right] \rho'(t) \right).$$

We may assume, without loss of generality, that  $\bar{k} = k_j$  and  $v_j = x$ . In this case we have

$$-J_{\bar{k}\bar{d}} [iI_{\bar{k}z, \bar{d}x}, I_{k_1 v_1, \dots, k_r v_r}] = J_{\bar{k}\bar{d}} iI_{k_1 v_1, \dots, k_j y, \dots, k_r v_r, \bar{d}x}.$$

Combining the previous equality with (128), equation (70) follows easily.

2. Using the same procedure, we obtain again (128), but now both indices  $\bar{k}$  and  $\bar{d}$  are in  $\{k_1, \dots, k_r\}$ . Assume, for example that  $k_1 = \bar{k}$  and  $k_2 = \bar{d}$ , and take  $v_{k_1} = v_{k_2} = x$ .

<sup>8</sup>This permutation is the same and fixed for all the results in which it is mentioned.

Now we have

$$(129) \quad [I_{k_1 z, k_2 x}, I_{k_1 x, k_2 x, \dots, k_r v_r}] = I_{k_1 y, k_2 x^2, k_3 v_3, \dots, k_r v_r},$$

where, with this notation, we mean that in the  $k_2$ th position we have the matrix  $\sigma_x^2$ . Thus, combining (128) and (129), we get

$$(130) \quad \beta J_{k_1 k_2} \text{Tr} (I_{k_1 y, k_2 x^2, k_3 v_3, \dots, k_r v_r} \rho(t)) = \beta' J'_{k_1 k_2} \text{Tr} (I'_{k_1 y, k_2 x^2, k_3 v_3, \dots, k_r v_r} \rho'(t)).$$

Using the same procedure, we conclude that

$$(131) \quad \beta J_{k_1 k_2} \text{Tr} (I_{k_1 y, k_2 y^2, k_3 v_3, \dots, k_r v_r} \rho(t)) = \beta' J'_{k_1 k_2} \text{Tr} (I'_{k_1 y, k_2 y^2, k_3 v_3, \dots, k_r v_r} \rho'(t))$$

and

$$(132) \quad \beta J_{k_1 k_2} \text{Tr} (I_{k_1 y, k_2 z^2, k_3 v_3, \dots, k_r v_r} \rho(t)) = \beta' J'_{k_1 k_2} \text{Tr} (I'_{k_1 y, k_2 z^2, k_3 v_3, \dots, k_r v_r} \rho'(t)).$$

Adding together (130), (131), and (132) and using (57), we get

$$\beta(l_{k_2}(l_{k_2}+1))J_{k_1 k_2} \text{Tr} (I_{k_1 y, k_3 v_3, \dots, k_r v_r} \rho(t)) = \beta'(l'_{k_2}(l'_{k_2}+1))J'_{k_1 k_2} \text{Tr} (I'_{k_1 y, k_3 v_3, \dots, k_r v_r} \rho'(t)),$$

as desired.

*Proof of Lemma 10.* We recall the formulas (77) and (78) to be proved, i.e.,

$$(133) \quad (e^{i\sigma_z t} \otimes \mathbf{1}_{R \times R}) I_{1x} (e^{-i\sigma_z t} \otimes \mathbf{1}_{R \times R}) := P_{N_1}(t) \otimes \mathbf{1}_{R \times R},$$

where the matrix  $P(\cdot)$  is periodic with period  $2\pi$ , and

$$(134) \quad P_{N_1}(\pi) = -P_{N_1}(0) = \sigma_x.$$

The proof can be done directly by computing the matrix above. This is simplified by the fact that the matrix  $\sigma_z$  is always a diagonal matrix. We will give an outline of the argument when  $l_1$  is a half-integer spin. The idea is to use the representations for the matrices  $\sigma_z$  and  $\sigma_x$  given by (85) and (86). The case of integer spin can be derived similarly starting with the representations given by (90) and (91).

Using (85) and (86), we obtain

$$(135) \quad e^{i\sigma_z t} i\sigma_x e^{-i\sigma_z t} = \begin{pmatrix} e^{iDt} F e^{-iDt} & e^{iDt} P e^{iD^a t} \\ e^{iD^a t} P^T e^{-iDt} & e^{-iD^a t} F^a e^{iD^a t} \end{pmatrix}.$$

The properties of the matrices  $D$ ,  $P$ , and  $F$  are described in section 7. Moreover,  $D = \text{diag}(j, j-1, \dots, \frac{1}{2})$  for a half-integer spin  $j$ . By using these properties, it follows that all the time-dependent terms in (135) are of the form  $e^{it}$ . Thus matrix (135) is periodic of period  $2\pi$ . The fact that the dependence is of type  $e^{it}$ , in turn, implies that (133) and (134) hold.

## Appendix B: Matrix elements of spin angular momentum operator.

For a spin  $l$ , the matrices  $\sigma_{x,y,z}$  are of dimensions  $2l+1$ . It is convenient to label the rows and columns by the index  $-l, -l+1, \dots, l-1, l$ . With this convention we have that  $\sigma_z$  is diagonal, and in particular,

$$(\sigma_z)_{m,s} := m\delta_{ms}, \quad m, s = -l, -l+1, \dots, l-1, l.$$

$\sigma_x$  and  $\sigma_y$  are defined through the matrices  $J_+$  and  $J_-$  as

$$\sigma_x := \frac{J_+ + J_-}{2}, \quad \sigma_y := \frac{J_+ - J_-}{2i}$$

with

$$(J_+)_{m,s} := \sqrt{(l-m)(l+m+1)}\delta_{s(m+1)}, \quad m, s = -l, -l+1, \dots, l-1, l,$$

$$(J_-)_{m,s} := \sqrt{(l+m)(l-m+1)}\delta_{(s+1)(m)}, \quad m, s = -l, -l+1, \dots, l-1, l.$$

#### REFERENCES

- [1] F. ALBERTINI AND D. D'ALESSANDRO, *Notions of controllability for bilinear multilevel quantum systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 1399–1403.
- [2] F. ALBERTINI AND D. D'ALESSANDRO, *Model identification for spin networks*, Linear Algebra Appl., 394 (2005), pp. 237–256.
- [3] F. ALBERTINI AND D. D'ALESSANDRO, *Control of the evolution of Heisenberg spin systems*, Eur. J. Control, 10 (2004), pp. 497–504.
- [4] F. ALBERTINI AND E. D. SONTAG, *For neural networks, function determines form*, Neural Netw., 6 (1993), pp. 975–990.
- [5] B. BARBARA AND L. GUNTHER, *Magnets, molecules and quantum mechanics*, Physics World, (1999), pp. 35–39.
- [6] H.-P. BREUER AND F. PETRUCCIONE, *The Theory of Open Quantum Systems*, Oxford University Press, New York, 2002.
- [7] R. W. BROCKETT, *Volterra series and geometric control theory*, Automatica—J. IFAC, 12 (1976), pp. 167–176.
- [8] G. CHRISTOU, D. GATTESCHI, D. N. HENDRICKSON, AND R. SESSOLI, *Single-molecule magnets*, MRS Bull., (2000), pp. 66–71.
- [9] P. E. CROUCH, *Dynamical realizations of finite Volterra series*, SIAM J. Control Optim., 19 (1981), pp. 177–202.
- [10] D. D'ALESSANDRO, *On quantum state observability and measurement*, J. Phys. A, 36 (2003), pp. 9721–9735.
- [11] D. D'ALESSANDRO, *Controllability, observability and parameter identification of two coupled spin 1's*, IEEE Trans. Automat. Control, 50 (2005), pp. 1054–1058.
- [12] D. D'ALESSANDRO AND F. ALBERTINI, *Quantum symmetries and Cartan decompositions in arbitrary dimensions*, J. Phys. A, 40 (2007), pp. 2439–2453.
- [13] D. D'ALESSANDRO, *Introduction to Quantum Control and Dynamics*, CRC Press, Boca Raton, FL, 2007.
- [14] P. D'ALESSANDRO, A. ISIDORI, AND A. RUBERTI, *Realization and structure theory of bilinear dynamical systems*, SIAM J. Control, 12 (1974), pp. 517–535.
- [15] M. FLIESS, *Functionelles causales non lineaires et indeterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.
- [16] S. HELGASON, *Differential Geometry, Lie Groups and Symmetric Spaces*, Academic Press, London, 1978.
- [17] G. M. HUANG, T. J. TARN, AND J. W. CLARK, *On the controllability of quantum mechanical systems*, J. Math. Phys., 24 (1983), pp. 2608–2618.
- [18] J. E. HUMPHREYS, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag, New York, 1972.
- [19] A. ISIDORI AND A. RUBERTI, *Realization theory of bilinear systems*, in Geometric Methods in System Theory, D. Q. Mayne and R. W. Brockett, eds., D. Reidel Pub. Co., Dordrecht, Boston, 1973, pp. 83–130.
- [20] B. JAKUBCZYK, *Existence and uniqueness of realizations of nonlinear systems*, SIAM J. Control Optim., 18 (1980), pp. 455–471.
- [21] N. KHANEJA, S. GLASER, AND R. BROCKETT, *Sub-Riemannian geometry and time optimal control of three spin systems: Quantum gates and coherence transfer*, Phys. Rev. A (3), 65 (2002), 032301.
- [22] V. RAMAKRISHNA, R. J. OBER, AND H. RABITZ, *Control of a coupled two-spin system without hard pulses*, Phys. Rev. A (3), 65 (2002), 063405.
- [23] V. RAMAKRISHNA, M. SALAPAKA, M. DAHLEH, H. RABITZ, AND A. PEIRCE, *Controllability of molecular systems*, Phys. Rev. A (3), 51 (1995), pp. 960–966.

- [24] A. RUBERTI, A. ISIDORI, AND P. D'ALESSANDRO, *Theory of Bilinear Dynamical Systems*, International Center for Mechanical Sciences, Courses and Lectures 158, Springer-Verlag, New York, 1972.
- [25] J. J. SAKURAI, *Modern Quantum Mechanics*, Addison-Wesley, Reading, MA, 1994.
- [26] S. SASTRY AND M. BODSON, *Adaptive Control: Stability, Convergence and Robustness*, Advanced Reference Series, Prentice-Hall, Upper Saddle River, NJ, 1989.
- [27] H.-J. SCHMIDT AND M. LUBAN, *Continuous families of isospectral Heisenberg spin systems and the limits of inference from measurements*, J. Phys. A, 34 (2001), pp. 2839–2858.
- [28] H. J. SUSSMANN, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263–284.
- [29] R. M. WHITE, *Quantum Theory of Magnetism*, Springer Ser. Solid-State Sci. 32, Springer-Verlag, Berlin, New York, 1983.



## GLOBAL SMOOTH SOLUTIONS OF THE QUASI-LINEAR WAVE EQUATION WITH INTERNAL VELOCITY FEEDBACK\*

ZHI-FEI ZHANG<sup>†</sup> AND PENG-FEI YAO<sup>†</sup>

**Abstract.** We study the existence of global smooth solutions for the quasi-linear wave equation by an internal local damping when initial data are close to a given equilibrium. Our interest is in studying the structure of the damping region, which guarantees the existence of global solutions. Our results show that the structure of the damping region depends on the geometric properties of a Riemannian metric, based on the coefficients and the equilibrium of the system. Some geometrical conditions are presented to obtain the damping region.

**Key words.** quasi-linear wave equation, Riemannian metric, internal velocity feedback

**AMS subject classifications.** 49B, 49E, 35B35, 35L65, 35L70, 38J45

**DOI.** 10.1137/070679454

**1. Introduction and main results.** Let  $n \geq 2$  be an integer,  $\Omega \subset R^n$  be a bounded open set with smooth boundary  $\Gamma$ , and

$$\mathbf{a}(x, y) = (a_1(x, y), a_2(x, y), \dots, a_n(x, y))$$

be a smooth mapping from  $\overline{\Omega} \times R^n$  to  $R^n$  with

$$(1.1) \quad \mathbf{a}(x, 0) = 0, \quad x \in \overline{\Omega},$$

such that  $(a_{ij}(x, y))$  is symmetrical and

$$(1.2) \quad (a_{ij}(x, y)) > 0 \quad \forall (x, y) \in \overline{\Omega} \times R^n,$$

where  $a_{ij} = a_{i y_j}$  are the partial derivatives of  $a_i$  with respect to the variable  $y$ . Let  $f(x, s): \overline{\Omega} \times R \rightarrow R$  be a smooth function such that

$$(1.3) \quad f(x, 0) = 0, \quad x \in \overline{\Omega}.$$

We consider the following problem:

$$(1.4) \quad \ddot{u}(t, x) - \operatorname{div} \mathbf{a}(x, \nabla u) + f(x, \dot{u}) = 0 \quad \text{in } (0, +\infty) \times \Omega,$$

$$u = w \quad \text{on } (0, +\infty) \times \Gamma,$$

$$u(0, x) = u_0(x), \quad \dot{u}(0, x) = u_1(x) \quad \text{on } (0, +\infty) \times \Gamma,$$

where  $w$  is an equilibrium solution, which satisfies

$$(1.5) \quad \operatorname{div} \mathbf{a}(x, \nabla w) = 0, \quad x \in \Omega.$$

\*Received by the editors January 7, 2007; accepted for publication (in revised form) February 27, 2008; published electronically July 2, 2008. This work was supported by National Natural Science Foundation of China grants 60225003, 60334040, 60221301 and 60774025.

<http://www.siam.org/journals/sicon/47-4/67945.html>

<sup>†</sup>Key Laboratory of Systems and Control, Institute of System Science, Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing 100080, People's Republic of China (zhangzf@amss.ac.cn, pfyao@iss.ac.cn).

It is well known that solutions to the problem (1.4) usually develop singularities after some time even if initial data  $(u_0, u_1)$  are close enough to  $(w, 0)$  [11]. In this paper we study what condition on  $f$  can guarantee that the problem (1.4) admits global smooth solutions when initial data  $(u_0, u_1)$  are close enough to  $(w, 0)$ . For this purpose, we assume that  $f$  satisfies

$$(1.6) \quad f_s(x, s) \geq 0, \quad x \in \overline{\Omega}, \quad |s| \leq 1,$$

$$\text{where } f_s(x, s) = \frac{\partial f(x, s)}{\partial s}.$$

Let

$$(1.7) \quad G = \{ x \in \overline{\Omega} \mid f_s(x, 0) > 0 \}.$$

We call  $G$  the damping region of the system (1.4). The structure of  $G$  reflects the effect of the internal dissipation  $f$ . We seek geometric conditions on  $G$  such that the problem (1.4) has global smooth solutions in time; moreover, we aim to establish energy decay estimates. We note that  $G = \overline{\Omega}$  is one of the choices for such purposes (see Theorems 1.1 and 1.2). However, we are particularly interested in the case where  $G$  is not the whole domain  $\Omega$  and is as small as possible in a technical sense. We show that such geometric conditions depend closely on a Riemannian metric, given by (1.10) below; see Corollaries 1.1–1.3. An example is presented at the end of this section.

Problems like (1.4) have been extensively studied, and a wealth of results on this subject is available in the literature. For bounded domains, see [2], [6], [8], [10], [12], [13], [14], [15], [19], [20], [24], [25], [26], [28], [39]. For the Cauchy problem, see [29], [30], [31], [40]. For exterior domains, see [21], [22], and the references therein. Reference [14] proved the global existence of solutions for the Kirchhoff wave equation with a boundary feedback. The piecewise multiplier method was introduced in [19] to study the structure of the damping region. To cope with a principal part with variable coefficients, the geometrical method was used in [8]. This method was introduced in [34] for the controllability of the wave equation with variable coefficients and was extended in [3], [4], [5], [16], [17], [18], [23], [32], [35], [36], and many others. For a recent survey on the geometric method, see [9]. Very recently, this method was used to study problems with quasi-linear principal parts in [37], [38], and the present paper.

We state our main results in detail, and their proofs will be given in the subsequent sections.

Let  $m \geq [\frac{n}{2}] + 3$  be a given positive integer. Let  $w \in H^m(\Omega)$  be an equilibrium of the system (1.4).

**DEFINITION 1.1.** *We say that  $(u_0, u_1) \in H^m(\Omega) \times H^{m-1}(\Omega)$  satisfies the compatibility conditions of  $m$  order with  $w$  if*

$$(1.8) \quad u_0|_{\Gamma} = w|_{\Gamma}, \quad u_k|_{\Gamma} = 0, \quad 1 \leq k \leq m-1,$$

where for  $k \geq 2$ ,

$$(1.9) \quad u_k = u^{(k)}(0),$$

as computed formally (and recursively) in terms of  $u_0$  and  $u_1$ , using (1.4).

Inspired by [7], we seek to find suitable geometric conditions on  $G$  such that the problem (1.4) has solutions  $u(t, x)$  in

$$\bigcap_{k=0}^m C^k((0, +\infty), H^{m-k}(\Omega)),$$

if

$$(u_0, u_1) \in H^m(\Omega) \times H^{m-1}(\Omega)$$

are close to  $(w, 0) \in H^m(\Omega) \times H^{m-1}(\Omega)$ , and satisfies the compatibility conditions of  $m$  order with  $w$ .

Let  $A(x, y) = (a_{ij}(x, y))$  be the  $n \times n$  matrix, given in (1.2), for each  $(x, y) \in \overline{\Omega} \times R^n$ . We define

$$(1.10) \quad g = A^{-1}(x, \nabla w(x)), \quad x \in \overline{\Omega},$$

as a Riemannian metric on  $\overline{\Omega}$  and consider the couple  $(\overline{\Omega}, g)$  as a Riemannian manifold with the boundary  $\Gamma$ . Here the metric  $g$  depends not only on the coefficients  $a_{ij}(\cdot, \cdot)$  but also on the equilibrium  $w$ . For each  $x \in \Omega$ , the metric  $g$  induces the inner product and the norm on the tangent space  $R_x^n = R^n$  by

$$(1.11) \quad g(X, Y) = \sum_{ij} g_{ij}(x) \alpha_i \beta_j, \quad |X|_g = g(X, X)^{1/2},$$

$$(1.12) \quad (g_{ij}(x)) = (a_{ij}(x, \nabla w(x)))^{-1}, \quad X = \sum_{i=1}^n \alpha_i \frac{\partial}{\partial x_i}, \quad Y = \sum_{i=1}^n \beta_i \frac{\partial}{\partial x_i}.$$

We denote the covariant differential of the metric  $g$  by  $D_g$ . If  $H$  is a vector field on  $\overline{\Omega}$ , then the covariant differential  $D_g H$  of  $H$  is a two order tensor field defined by

$$D_g H(X, Y) = g(D_{g_Y} H, X), \quad X, Y \in R_x^n, \quad x \in \overline{\Omega}.$$

We make the following assumption:

(H) There exist  $\varepsilon > 0$ ,  $\alpha > 0$ ,  $\Omega_i \subseteq \Omega$  with  $C^\infty$  boundary  $\partial\Omega_i$  and vector fields  $H^i$ ,  $i = 1, 2, \dots, I$ , such that  $\Omega_i \cap \Omega_j = \emptyset$  for  $1 \leq i < j \leq I$  and

$$(1.13) \quad D_g H^i(X, X) \geq \alpha |X|_g^2 \quad \forall X \in R_x^n, \quad x \in \Omega_i,$$

$$(1.14) \quad G \supseteq \overline{\Omega} \cap \mathcal{N}_\varepsilon \left[ \bigcup_{i=1}^I \Gamma_0^i \cup \left( \Omega \setminus \bigcup_{i=1}^I \Omega_i \right) \right],$$

where

$$\mathcal{N}_\varepsilon(S) = \bigcup_{x \in S} \{y \in R^n \mid |y - x| < \varepsilon\}, \quad S \subset R^n, \quad \Gamma_0^i = \{x \in \partial\Omega_i \mid H^i(x) \cdot \nu^i(x) > 0\},$$

and  $\nu^i(x)$  is the unit normal of  $\partial\Omega_i$  at  $x$  in the metric  $g$ , pointing towards the exterior of  $\Omega_i$ .

The results on the existence of short time solutions to the quasi-linear wave equation have been established in [7], [27]. Our first goal is to obtain the global smooth solution in time and the decay of the energy under the above assumption (H).

THEOREM 1.1. *Let equilibrium  $w \in H^m(\Omega)$  be such that assumption (H) holds. Suppose that  $(u_0, u_1) \in H^m(\Omega) \times H^{m-1}(\Omega)$  satisfies the compatibility conditions of  $m$  order with  $w$ . Then for  $(u_0 - w, u_1)$  small in  $H^m(\Omega) \times H^{m-1}(\Omega)$ , the system (1.4) has a global solution  $u$  in*

$$\bigcap_{k=0}^m C^k((0, +\infty), H^{m-k}(\Omega)).$$

THEOREM 1.2. *Under the same assumptions as in Theorem 1.1, we have that the energy of the problem (1.4) decays exponentially; that is, there exist positive constants  $C > 0$ ,  $k > 0$  such that*

$$(1.15) \quad \mathcal{E}(t) \leq Ce^{-kt}, \quad t \geq 0,$$

where  $\mathcal{E}(t) = \|u - w\|_{H^m(\Omega)}^2 + \sum_{k=1}^m \|u^{(k)}\|_{H^{m-k}(\Omega)}^2$ .

Since the set in the right-hand side of (1.14) is always a subset of  $\overline{\Omega}$ , the results in Theorems 1.1 and 1.2 hold if the damping region is the closure of the whole domain, that is,  $G = \overline{\Omega}$ . However, we are particularly interested in the case  $G \neq \overline{\Omega}$ .

If there is just one vector field  $H$  such that the inequality (1.13) holds for all  $x \in \Omega$ , we can take  $I = 1$  and  $\Omega_1 = \Omega$ . In this case, to obtain global smooth solutions and the decay of the energy, the damping region  $G$  need only be supported in a neighborhood of  $\Gamma_0$ , where

$$(1.16) \quad \Gamma_0 = \{x \in \Gamma \mid H(x) \cdot \nu(x) > 0\}.$$

This roughly means that we can dig out almost the whole  $\Omega$  from the domain  $\overline{\Omega}$  and that the damping region is supported only in a neighborhood of a subset of the boundary.

Suppose that the principal part in the problem (1.4) is linear, that is,  $\operatorname{div} \mathbf{a}(x, \nabla u) = \Delta u$ ; then

$$g_{ij}(x) = \delta_{ij}, \quad x \in \overline{\Omega},$$

$g = \cdot$  is the standard metric of  $R^n$ , and  $D_g = \nabla$  is the gradient of the standard metric. Let  $x_0 \in R^n$  be given. Then the vector field  $H(x) = x - x_0$  on  $\overline{\Omega}$  meets condition (1.13) with  $\alpha = 1$ . Thus, the damping region need only be supported in a neighborhood of

$$\Gamma_0 = \{x \in \Gamma \mid (x - x_0) \cdot \nu(x) > 0\}.$$

This is the case studied in [28].

The structure of the sets in (1.14) was given in [19] for the linear wave equation with constant coefficients, where  $\operatorname{div} \mathbf{a}(x, \nabla u) = \Delta u$ ,  $g_{ij}(x) = \delta_{ij}$ , and  $D_g = \nabla$ . Reference [19] also presents several examples where the number of the vector fields is larger than one. The present form in a general metric  $g$  was used in [8] for the linear wave equation, but with variable coefficients.

In general, whether there exists one vector field  $H$  satisfying estimate (1.13) on the whole domain  $\Omega$  largely depends on the sectional curvature of the Riemannian metric  $g$ . We show that if the sectional curvature of the Riemannian metric  $g$  is nonpositive, then such a vector field exists. More generally, we can always find a finite number of vector fields  $H^i$  and suitable subsets  $\Omega_i \subset \Omega$  such that the estimate (1.13) holds. (See Corollary 1.2 below.)

For  $x \in \overline{\Omega}$  let  $\Pi \subset R_x^n$  be a two-dimensional subspace. Denote by  $k_x(\Pi)$  the sectional curvature of the subspace  $\Pi$  at  $x$  under the Riemannian metric  $g$ . Let

$$(1.17) \quad \kappa = \sup_{x \in \overline{\Omega}, \Pi \subset R_x^n} k_x(\Pi).$$

Let  $x_0 \in \overline{\Omega}$  be given. Let  $\rho(x) = \rho(x, x_0)$  be the distance function from  $x \in \overline{\Omega}$  to  $x_0$  in the metric  $g$ . If  $g = \cdot$  is the standard metric of  $R^n$ , then  $\rho(x) = |x - x_0|$ . For a general metric  $g$ , such as (1.10), the structure of  $\rho(x)$  is very complicated. For the properties of this function, see any Riemannian geometry book, for example, [33]. We define a vector field  $H(x, x_0)$  on  $\overline{\Omega}$  by

$$(1.18) \quad H(x, x_0) = \rho(x) D_g \rho(x), \quad x \in \overline{\Omega}.$$

If  $g = \cdot$ , then  $D_g = \nabla$  and  $H(x, x_0) = x - x_0$ .

We have the following.

**COROLLARY 1.1.** *Suppose that  $\kappa \leq 0$ . Then the results in Theorems 1.1 and 1.2 hold for initial data close to  $(w, 0)$  if the damping region  $G$  is supported in a neighborhood of*

$$(1.19) \quad \Gamma_0 = \{x \in \Gamma \mid H(x, x_0) \cdot \nu(x) > 0\}.$$

If  $\kappa > 0$ , one vector field satisfying (1.13) on the whole  $\overline{\Omega}$  does not exist in general. A counterexample was given in [34]. But we can dig out a finite number of geodesic balls with the radius  $\leq \pi/2\sqrt{\kappa}$  from  $\overline{\Omega}$  and let the feedback region be supported in a neighborhood of the remaining part of  $\overline{\Omega}$ . This is the following.

**COROLLARY 1.2.** *Let  $\kappa > 0$ . Suppose that there are points  $x_i \in \overline{\Omega}$  for  $1 \leq i \leq I$  such that*

$$\rho(x_i, x_j) > \frac{\pi}{\sqrt{\kappa}}, \quad 1 \leq i, j \leq I, \quad i \neq j.$$

*Let  $B(x_i, r_0)$  and  $S(x_i, r_0)$  be the geodesic ball and the geodesic sphere centered at  $x_i$  with radius  $r_0$  in the metric  $g$ , respectively, where  $r_0 < \pi/2\sqrt{\kappa}$ . Let  $\Gamma_0^i = \{x \mid x \in S(x_i, r_0), H(x, x_i) \cdot \nu(x) > 0\}$ . If*

$$G \supseteq \overline{\Omega} \cap \mathcal{N}_\varepsilon \left[ \bigcup_{i=1}^I \Gamma_0^i \cup \left( \Omega \setminus \bigcup_{i=1}^I B(x_i, r_0) \right) \right],$$

*then the results in Theorems 1.1 and 1.2 hold for initial data close to  $(w, 0)$ .*

Since we can dig out from  $\Omega$  as many as possible geodesic balls with radii small, the following results are immediate.

**COROLLARY 1.3.** *For  $\varepsilon > 0$  given, we can choose a damping region  $G \subset \overline{\Omega}$  with*

$$\text{meas}(G) < \varepsilon,$$

*where  $\text{meas}(G)$  is the  $n$ -dimensional Lebesgue measure of  $G$  such that the results in Theorems 1.1 and 1.2 hold for initial data close to  $(w, 0)$ .*

We consider next a dissipation much weaker than those studied in most papers. See [14] for boundary feedbacks and see [24] for internal ones. Let  $f$  satisfy (1.6) and let there exist constants  $r > 0$  and  $c > 0$  such that

$$(1.20) \quad \left| \frac{\partial f}{\partial x_i}(x, s) \right| \leq c|s|^{2r-1}, \quad x \in \Omega, \quad s \in R, \quad 1 \leq i \leq n.$$

Let  $G$  be defined in the following way:  $G$  consists of all the points  $x \in \overline{\Omega}$  for which there are constants  $c(x) > 0$  such that

$$(1.21) \quad f(x, s)s \geq c(x)|s|^{\beta+2}, \quad s \in R,$$

where  $\beta \geq 0$  is a fixed number; that is,

$$(1.22) \quad G = \{x \in \overline{\Omega} \mid f(x, s)s \geq c(x)|s|^{\beta+2}, \ c(x) > 0\}.$$

For the Kirchhoff wave system, as was studied in [14] and [24], the following estimates are obtained. The energy of the first order decays as

$$(1.23) \quad \|\dot{u}(t)\|_{L^2(\Omega)} + \|u(t)\|_{H^1(\Omega)} \leq \frac{C}{(1+t)^{2/\beta}},$$

and the second order energy is just bounded as

$$(1.24) \quad \|\ddot{u}(t)\|_{L^2(\Omega)} + \|u(t)\|_{H^2(\Omega)} \leq C.$$

For the Kirchhoff system it is not clear whether or not a dissipation such as (1.21) can make the energy of the second order decay at the same rate as in (1.23).

Since the Kirchhoff system has a local regularity in time in the space  $H^2(\Omega) \times H^1(\Omega)$ , a dissipation such as (1.21) ensures the existence of global  $H^2(\Omega)$  solutions (1.24) and the decay of the first order energy (1.23). But, for the system (1.4) with a general quasi-linear principal part, we have only a local regularity in  $H^m(\Omega) \times H^{m-1}(\Omega)$  that is much higher,  $m \geq [\frac{n}{2}] + 3$ . Roughly speaking, for existence of global solutions to such systems with a local regularity in  $H^m(\Omega) \times H^{m-1}(\Omega)$ , the energies of all the orders from 1 to  $m-1$  should decay in the same way and the energy of order  $m$  should be bounded. In the case (1.7), the energy of order  $m$  actually decays exponentially. Since we do not know whether or not a dissipation such as (1.21) makes the energies of order larger than one decay, we consider some special quasi-linear principal parts to lower the local regularity to let (1.21) work.

We consider the problem

$$(1.25) \quad \ddot{u}(t, x) - \sum_{ij=1}^n \left( a_{ij}(x, \|\nabla u\|^{2r}) u_{x_i} \right)_{x_j} + f(x, \dot{u}) = 0 \quad \text{in } (0, +\infty) \times \Omega,$$

$$u = 0 \quad \text{on } \partial D,$$

$$u(0, x) = u_0(x), \dot{u}(0, x) = u_1(x) \quad \text{on } (0, +\infty) \times \Gamma,$$

where  $\|\cdot\|$  is the  $L^2(\Omega)$  norm and  $r$  is a constant satisfying

$$(1.26) \quad r > \frac{\beta+1}{2}.$$

In the problem (1.25), we assume that  $a_{ij}(x, s)$  are smooth functions on  $\Omega \times R$  such that the matrices  $A(x, s) = (a_{ij}(x, s))$  are symmetrical on  $\Omega \times R$  satisfying

$$(1.27) \quad \sum_{i,j=1}^n (a_{ij}(x, s)) \xi_i \xi_j \geq a_0 |\xi|^2 \quad \forall (x, s) \in \overline{\Omega} \times R, \quad \forall \xi \in R^n,$$

where  $a_0$  is a positive constant. We define a Riemann metric on  $\Omega$  by

$$(1.28) \quad g = A^{-1}(x, 0).$$

If  $a_{ij}(x, s) = a(s)$  for all  $1 \leq i, j \leq n$ , the problem (1.25) represents a Kirchhoff wave model; see [14] or [24].

We make the following assumption.

**(H1)** For any  $(u_0, u_1) \in H^2(\Omega) \times H^1(\Omega)$  with appropriate compatibility, the system (1.25) has a solution of short time in  $H^2(\Omega) \times H^1(\Omega)$ .

We have the following.

**THEOREM 1.3.** *Let the assumptions **(H)** and **(H1)** hold where  $G$  is given by (1.22). Suppose that  $(u_0, u_1) \in H^2(\Omega) \times H^1(\Omega)$  satisfies the compatibility conditions of order 2. Then for  $(u_0, u_1)$  small in  $H^2(\Omega) \times H^1(\Omega)$ , system (1.25) has a global solution  $u$  in*

$$\bigcap_{k=0}^2 C^k((0, +\infty), H^{2-k}(\Omega)).$$

Moreover, the estimates in (1.23) and (1.24) hold, where  $u(t)$  is the solution to the system (1.25).

Finally, we give an example to verify Corollary 1.1.

Let

$$\mathbf{a}(y) = \begin{pmatrix} a_1(y_1), a_2(y_2) \end{pmatrix}: \quad R^2 \rightarrow R^2,$$

where  $a_i$  are smooth functions on  $R$  for  $i = 1, 2$  such that

$$a'_i(s) > 0, \quad s \in R.$$

Let

$$w(x) = x_1 x_2, \quad x = (x_1, x_2) \in R^2.$$

Then

$$\operatorname{div} \mathbf{a}(\nabla w) = 0, \quad x \in R^2.$$

The metric is

$$g = A^{-1}(x, \nabla w) = \begin{pmatrix} 1/a'_1(x_2) & 0 \\ 0 & 1/a'_2(x_1) \end{pmatrix}.$$

We let

$$a_1(s) = a_2(s) = \arctan s, \quad s \in R.$$

Then

$$g = \begin{pmatrix} 1 + x_2^2 & 0 \\ 0 & 1 + x_1^2 \end{pmatrix}.$$

By [34, Lemma 3.2], the Gauss curvature of  $(R^2, g)$  is

$$(1.29) \quad \kappa(x) = -\frac{2 + |x|^2}{(1 + x_1^2)^2(1 + x_2^2)^2} \leq 0, \quad x = (x_1, x_2) \in R^2.$$

**Conclusion.** The results in Corollary 1.1 hold true for the following problem:

$$\begin{aligned}
 (1.30) \quad & \ddot{u}(t, x) - \frac{u_{x_1 x_1}(t, x)}{1 + u_{x_1}^2} - \frac{u_{x_2 x_2}(t, x)}{1 + u_{x_2}^2} + f(x, \dot{u}) = 0 \quad \text{in } (0, +\infty) \times \Omega, \\
 & u = 0 \quad \text{on } \partial D, \\
 & u(0, x) = u_0(x), \dot{u}(0, x) = u_1(x) \quad \text{on } (0, +\infty) \times \Gamma
 \end{aligned}$$

for any domain  $\Omega \subset \mathbb{R}^2$  bounded and  $m = 4$ .

**2. Energy estimates.** We assume that solutions to the system (1.4) exist for some  $T > 0$  when initial data  $(u_0, u_1)$  are in  $H^m(\Omega) \times H^{m-1}(\Omega)$  satisfying the  $m$  order compatibility conditions with a given equilibrium  $w \in H^m(\Omega)$ , and we write those as  $u = w + \phi$ . Since  $w$  does not depend on  $t$ , the system (1.4) is equivalent to

$$\begin{aligned}
 (2.1) \quad & \ddot{\phi}(t, x) - \operatorname{div} \mathbf{b}(x, \nabla \phi) + f(x, \dot{\phi}) = 0 \quad \text{in } (0, +\infty) \times \Omega, \\
 & \phi = 0 \quad \text{on } (0, +\infty) \times \Gamma, \\
 & \phi(0, x) = \phi_0(x), \dot{\phi}(0, x) = \phi_1(x) \quad \text{on } (0, +\infty) \times \Gamma,
 \end{aligned}$$

where we have set

$$(2.2) \quad \mathbf{b}(x, y) = \mathbf{a}(x, \nabla w + y), \quad (x, y) \in \overline{\Omega} \times \mathbb{R}^n,$$

$$(2.3) \quad \phi_0 = u_0 - w, \quad \phi_1 = u_1, \quad x \in \Omega.$$

Let  $\phi$  solve the problem (2.1) for some  $T > 0$ . Denote

$$(2.4) \quad B(x, y) = (b_{ij}(x, y)), \quad b_{ij}(x, y) = a_{ij}(x, \nabla w + y), \quad (x, y) \in \overline{\Omega} \times \mathbb{R}^n;$$

then

$$(2.5) \quad \dot{\mathbf{b}}(x, \nabla \phi) = B_\phi(t) \nabla \dot{\phi},$$

and for  $j \geq 2$ ,

$$(2.6) \quad \mathbf{b}^{(j)}(x, \nabla \phi) = B_\phi(t) \nabla \phi^{(j)} + \sum_{k=1}^{j-1} B_\phi^{(k)}(t) \nabla \phi^{(j-k)},$$

where

$$(2.7) \quad B_\phi(t) = B(x, \nabla \phi).$$

We compute the  $j$  order derivatives of  $f(x, \dot{\phi})$  with respect to  $t$  and obtain

$$\begin{aligned}
 (2.8) \quad & (f(x, \dot{\phi}))^{(j)} = \sum_{i=1}^j \sum_{l_1 + \dots + l_i = j} f_y^{(i)}(x, \dot{\phi}) \dot{\phi}^{(l_1)} \dot{\phi}^{(l_2)} \dots \dot{\phi}^{(l_i)} \\
 & = f_y(x, \dot{\phi}) \dot{\phi}^{(j)} + \sum_{i=2}^j \sum_{l_1 + \dots + l_i = j} f_y^{(i)}(x, \dot{\phi}) \dot{\phi}^{(l_1)} \dot{\phi}^{(l_2)} \dots \dot{\phi}^{(l_i)}.
 \end{aligned}$$

We define

$$(2.9) \quad \mathcal{B}_\phi(t)v = \operatorname{div} B_\phi(t) \nabla v, \quad v \in H^2(\Omega),$$



and

$$(2.10) \quad v_{\nu_B} = \langle B_\phi(t) \nabla v, \nu \rangle, \quad v \in H^2(\Omega), \quad x \in \Gamma;$$

then

$$(2.11) \quad (\mathcal{B}_\phi(t)v, \phi)_{L^2(\Omega)} = -(B_\phi(t) \nabla v, \nabla \phi)_{L^2(\Omega)}, \quad v, \phi \in H^2(\Omega) \cap H_\Gamma^1(\Omega),$$

where

$$H_\Gamma^1(\Omega) = \{v \in H^1(\Omega), v|_{\Gamma} = 0\}.$$

For  $T > 0$ , differentiating the system (2.1)  $j$  times with respect to  $t$ , we get

$$(2.12) \quad \begin{aligned} \phi^{(j)}(t, x) - \mathcal{B}_\phi(t) \phi^{(j)} + f_y(x, \dot{\phi}) \dot{\phi}^{(j)} + r_j(t) &= 0 \quad \text{in } (0, +\infty) \times \Omega, \\ \phi^{(j)} &= 0 \quad \text{on } (0, +\infty) \times \Gamma, \end{aligned}$$

where  $r_j(t)$  is defined by

$$(2.13) \quad r_j(t) = \sum_{i=2}^j \sum_{l_1+\dots+l_i=j} f_y^{(i)}(x, \dot{\phi}) \dot{\phi}^{(l_1)} \dot{\phi}^{(l_2)} \dots \dot{\phi}^{(l_i)} - \sum_{k=1}^{j-1} \operatorname{div} B_\phi^{(k)}(t) \nabla \phi^{(j-k)}$$

for  $2 \leq j \leq m-1$  and  $r_1(t) = 0$ .

For the system (2.12), we define the corresponding energy as

$$(2.14) \quad V_j(t) = \|\dot{\phi}^{(j)}(t, x)\|_{L^2(\Omega)}^2 + (B_\phi(t) \nabla \phi^{(j)}(t, x), \nabla \phi^{(j)}(t, x))_{L^2(\Omega)}.$$

Moreover, we introduce an operator

$$(2.15) \quad \mathcal{N}_\phi(t)v = \operatorname{div} N_\phi(t) \nabla v, \quad v \in H^2(\Omega),$$

where

$$N_\phi(t) = \int_0^1 B(x, s \nabla \phi) ds.$$

Then the system (2.1) can be rewritten as

$$(2.16) \quad \begin{aligned} \ddot{\phi}(t, x) - \mathcal{N}_\phi(t) \phi + f(x, \dot{\phi}) &= 0 \quad \text{in } (0, +\infty) \times \Omega, \\ \phi &= 0 \quad \text{on } (0, +\infty) \times \Gamma, \end{aligned}$$

with the energy defined as

$$(2.17) \quad V_0(t) = \|\dot{\phi}(t, x)\|_{L^2(\Omega)}^2 + (N_\phi(t) \nabla \phi(t, x), \nabla \phi(t, x))_{L^2(\Omega)}.$$

Next, we define

$$(2.18) \quad \mathcal{E}(t) = \sum_{k=0}^m \|\phi^{(k)}(t, x)\|_{H^{m-k}(\Omega)}^2,$$

$$(2.19) \quad \mathcal{Q}(t) = \sum_{j=0}^{m-1} V_j(t),$$

$$(2.20) \quad P(t) = \|\phi(t, x)\|_{L^2(\Omega)}^2 + \|\dot{\phi}(t, x)\|_{L^2(\Omega)}^2 + (N_\phi(t) \nabla \phi(t, x), \nabla \phi(t, x))_{L^2(\Omega)},$$

$$(2.21) \quad R(t) = \sum_{j=0}^{m-1} \|\phi^{(j)}(t, x)\|_{L^2(\Omega)}^2,$$

$$(2.22) \quad \mathcal{L}(t) = \sum_{k=3}^{2m} \mathcal{E}^{k/2}(t).$$

**THEOREM 2.1.** *Let  $\gamma > 0$  be given and let  $\phi$  be a solution to the problem (2.1) on the interval  $[0, T]$  for some  $T > 0$  such that*

$$(2.23) \quad \sup_{0 \leq t \leq T} \|\phi(t, x)\|_{H^m(\Omega)} \leq \gamma, \quad \sup_{0 \leq t \leq T} \|\dot{\phi}(t, x)\|_{H^{m-1}(\Omega)} \leq \gamma.$$

*Then there are constants  $C_{0,\gamma} > 0$  and  $C_\gamma > 0$ , which depend only on  $\gamma$ , such that for  $0 \leq t \leq T$ ,*

$$(2.24) \quad C_{0,\gamma} \mathcal{Q}(t) \leq \mathcal{E}(t) \leq C_\gamma \mathcal{Q}(t) + C_\gamma \mathcal{L}(t) + C_\gamma \|\phi(t, x)\|_{L^2(\Omega)}^2,$$

$$(2.25) \quad -\dot{\mathcal{Q}}(t) \leq C_\gamma \mathcal{L}(t) + 2(f_y(x, \dot{\phi}) \phi^{(j+1)}, \phi^{(j+1)})_{L^2(\Omega)} + 2(f(x, \dot{\phi}), \dot{\phi})_{L^2(\Omega)},$$

$$(2.26) \quad \dot{\mathcal{Q}}(t) \leq C_\gamma \mathcal{L}(t),$$

$$(2.27) \quad P(t) \leq \left( P(0) + C_\gamma \int_0^t \mathcal{E}^{3/2}(\tau) d\tau \right) e^t.$$

We collect here a few basic properties of Sobolev spaces.

(1) Let  $s_1 \geq s_2 > 0$ . For any  $\varepsilon > 0$  there exists  $C_\varepsilon > 0$  such that

$$\|v\|_{H^{s_2}(\Omega)}^2 \leq \varepsilon \|v\|_{H^{s_1}(\Omega)}^2 + C_\varepsilon \|v\|_{L^2(\Omega)}^2 \quad \forall v \in H^{s_1}(\Omega).$$

(2) If  $s > n/2$ , then for each  $k = 0, \dots$ , we have  $H^{s+k}(\Omega) \subset C^k(\bar{\Omega})$  with continuous embedding.

(3) If  $r := \min\{s_1, s_2, s_1 + s_2 - [n/2] - 1\} \geq 0$ , then there exists a constant  $C > 0$  such that

$$\|f_1 f_2\|_{H^r(\Omega)} \leq C \|f_1\|_{H^{s_1}(\Omega)} \|f_2\|_{H^{s_2}(\Omega)} \quad \forall f_1 \in H^{s_1}(\Omega), f_2 \in H^{s_2}(\Omega).$$

(4) Let  $s_j \geq 0$ ,  $j = 1, \dots, k$ , and  $r := \min_{1 \leq i \leq k} \min_{j_1 \leq \dots \leq j_i} \{s_{j_1} + \dots + s_{j_i} - (i-1)([n/2] + 1)\} \geq 0$ ; then there exists a constant  $C > 0$  such that, for  $1 \leq j \leq k$ ,

$$(2.28) \quad \|f_1 \cdots f_k\|_{H^r(\Omega)} \leq C \|f_1\|_{H^{s_1}(\Omega)} \cdots \|f_k\|_{H^{s_k}(\Omega)} \quad \forall f_j \in H^{s_j}(\Omega).$$

LEMMA 2.1. *Let  $\gamma > 0$  be given. Let  $\phi \in H^m(\Omega)$  satisfy the conditions (2.23). Let  $p(\cdot, \cdot)$  and  $\lambda(\cdot, \cdot)$  be smooth functions on  $\bar{\Omega} \times R^n$  and on  $\bar{\Omega} \times R$ , respectively. Set*

$$P(x) = p(x, \nabla \phi), \quad \Lambda(x) = \lambda(x, \dot{\phi}).$$

*Then there exists a constant  $C_\gamma > 0$ , depending on  $\gamma$ , such that for  $0 \leq k \leq m-1$ ,*

$$(2.29) \quad \|P\|_{H^k(\Omega)} \leq C_\gamma \sum_{j=0}^k (1 + \|\phi\|_{H^m(\Omega)})^j,$$

$$(2.30) \quad \|\Lambda\|_{H^k(\Omega)} \leq C_\gamma \sum_{j=0}^k (1 + \|\dot{\phi}\|_{H^{m-1}(\Omega)})^j.$$

*Proof.* We prove the estimates by induction on  $k$ .

For  $k = 0$ , since  $\sup |\nabla \phi| \leq C\|\phi\|_{H^m(\Omega)} \leq C_\gamma$ , we have

$$\|\Lambda\|_{L^2(\Omega)} = \left( \int_{\Omega} \lambda^2(x, \nabla \phi) dx \right)^{\frac{1}{2}} \leq \sup_{x \in \bar{\Omega}, |y| \leq C_\gamma} |\lambda(x, y)| \sqrt{\text{meas}(G)} \leq C_\gamma.$$

We assume that the estimate (2.30) is true for  $0 \leq k \leq m-2$ . We aim to prove that it is true with  $k$  replaced by  $k+1$ . Note that

$$\Lambda_{x_i}(x) = \lambda_{x_i}(x, \dot{\phi}) + \sum_{j=1}^n \lambda_s(x, \dot{\phi}) \dot{\phi}_{x_i}, \quad i = 1, 2, \dots, n,$$

and that we have

$$\begin{aligned} \|\Lambda\|_{H^{k+1}(\Omega)} &= \left( \sum_{i=1}^n \|\Lambda_{x_i}\|_{H^k(\Omega)}^2 + \|\Lambda\|_{L^2(\Omega)}^2 \right)^{1/2} \\ &\leq C_\gamma + C_\gamma \sum_{i=1}^n \|\lambda_{x_i}(x, \dot{\phi})\|_{H^k(\Omega)} + \sum_{i=1}^n \|\lambda_s(x, \dot{\phi}) \dot{\phi}_{x_i}\|_{H^k(\Omega)} \\ &\leq C_\gamma + C_\gamma \sum_{j=1}^k (1 + \|\dot{\phi}\|_{H^{m-1}(\Omega)})^j \\ &\quad + C_\gamma \sum_{i,j=1}^n \|\lambda_s(x, \dot{\phi})\|_{H^k(\Omega)} \cdot \|\dot{\phi}_{x_i}\|_{H^{m-2}(\Omega)} \\ &\leq C_\gamma \sum_{j=0}^{k+1} (1 + \|\dot{\phi}\|_{H^{m-1}(\Omega)})^j. \end{aligned}$$

The same procedure yields the estimate on  $P(x)$ .  $\square$

LEMMA 2.2. *Let  $\gamma > 0$  be given and let  $\phi$  be a solution to the problem (2.1) on  $[0, T)$  such that the condition (2.23) holds. Then for  $2 \leq j \leq m-1$ , there exists a constant  $C_\gamma > 0$ , depending on  $\gamma$ , such that*

$$(2.31) \quad \|r_j(t)\|_{H^{m-1-j}(\Omega)}^2 \leq C_\gamma \sum_{k=2}^{m-1} \mathcal{E}^k(t),$$

where  $r_j(t)$  is given by (2.13).

*Proof.* Let  $b(\cdot, \cdot)$  be a function on  $\Omega \times R^n$ . For  $1 \leq k \leq j-1$ , we have the formulas

$$(2.32) \quad b^{(k)}(x, \nabla \phi) = \sum_{s=1}^k \sum_{r_1+\dots+r_s=k} D_y^s b \left( \nabla \phi^{(r_1)}(t, x), \dots, \nabla \phi^{(r_s)}(t, x) \right),$$

where  $D_y^s b$  denotes the covariant differential of order  $s$  of the function  $b(x, y)$ , given in (2.4), with respect to the variable  $y$  in the standard metric of  $R^n$ . Then  $r_j(t)$  are the sum of such functions in the form

$$(2.33) \quad \left( f(x, \nabla \phi) \phi_{x_{j_1}}^{(r_1)} \dots \phi_{x_{j_s}}^{(r_s)} \phi_{x_p}^{(j-k)} \right)_{x_q} \quad \text{or} \quad g(x, \dot{\phi}) \dot{\phi}^{(l_1)} \dots \dot{\phi}^{(l_i)},$$

with  $r_1 + \dots + r_s = k$  for  $1 \leq s \leq k \leq j-1$  and with  $l_1 + \dots + l_i = j$  for  $2 \leq i \leq j$ , respectively. Using the estimates (2.28)–(2.30), we have

$$(2.34) \quad \left\| \left( f(x, \nabla \phi) \phi_{x_{j_1}}^{(r_1)} \dots \phi_{x_{j_s}}^{(r_s)} \phi_{x_p}^{(j-k)} \right)_{x_q} \right\|_{H^{m-1-j}(\Omega)}^2 \\ \leq \|f(x, \nabla \phi) \phi_{x_{j_1}}^{(r_1)} \dots \phi_{x_{j_s}}^{(r_s)} \phi_{x_p}^{(j-k)}\|_{H^{m-j}(\Omega)}^2 \\ \leq C_\gamma \|\phi_{x_{j_1}}^{(r_1)}\|_{H^{m-r_1-1}(\Omega)}^2 \dots \|\phi_{x_{j_s}}^{(r_s)}\|_{H^{m-r_s-1}(\Omega)}^2 \|\phi_{x_p}^{(j-k)}\|_{H^{m-1-j+k}(\Omega)}^2 \\ \leq C_\gamma \|\phi^{(r_1)}\|_{H^{m-r_1}(\Omega)}^2 \dots \|\phi^{(r_s)}\|_{H^{m-r_s}(\Omega)}^2 \|\phi^{(j-k)}\|_{H^{m-j+k}(\Omega)}^2 \leq C_\gamma \mathcal{E}^{s+1}(t).$$

For the second term in (2.33), noticing  $2 \leq i \leq j \leq m-1$ , we have

$$(2.35) \quad \|g(x, \dot{\phi}) \dot{\phi}^{(l_1)} \dots \dot{\phi}^{(l_i)}\|_{H^{m-j-1}(\Omega)}^2 \\ \leq \|g(x, \dot{\phi})\|_{H^{m-j-1}(\Omega)}^2 \|\dot{\phi}^{(l_1)}\|_{H^{m-l_1-1}(\Omega)}^2 \dots \|\dot{\phi}^{(l_i)}\|_{H^{m-l_i-1}(\Omega)}^2 \\ \leq C_\gamma \mathcal{E}^i(t).$$

The estimate (2.31) follows from (2.34) and (2.35).  $\square$

LEMMA 2.3 (see [37]). *Let  $\gamma > 0$  be given and let  $\phi$  be a solution to the problem (2.1) on  $[0, T]$  for some  $T > 0$  such that the condition (2.23) holds. Then there exists a constant  $C_\gamma > 0$ , depending on  $\gamma$ , such that*

$$(2.36) \quad \|v\|_{H^{k+1}(\Omega)}^2 \leq C_\gamma (\|\mathcal{B}_\phi(t)v\|_{H^{k-1}(\Omega)}^2 + \|v\|_{H^k(\Omega)}^2)$$

for  $v \in H^{k+1}(\Omega) \cap H_0^1(\Omega)$ ,  $0 \leq k \leq m-1$ ,  $t \in [0, T]$ .

*Proof of Theorem 2.1.* It is clear that there exists a constant  $C_{0,\gamma}$  such that

$$C_{0,\gamma} \mathcal{Q}(t) \leq \mathcal{E}(t).$$

Let us prove the right-hand side of the inequality (2.24).

First, we prove

$$(2.37) \quad \|\phi^{(j)}\|_{H^{m-j}(\Omega)}^2 \leq C_\gamma \mathcal{Q}(t) + C_\gamma \mathcal{L}(t) + \|\phi(t)\|_{L^2(\Omega)}^2$$

for  $1 \leq j \leq m$  by induction on  $j$ . We have

$$\begin{aligned} & \|\phi^{(m)}\|_{L^2(\Omega)}^2 + \|\phi^{(m-1)}\|_{H^1(\Omega)}^2 \\ & \leq \|\phi^{(m)}\|_{L^2(\Omega)}^2 + C_\gamma(B_\phi(t)\nabla\phi^{(m-1)}, \nabla\phi^{(m-1)}) + \|\phi^{(m-1)}\|_{L^2(\Omega)}^2 \\ & \leq C_\gamma\mathcal{Q}(t), \end{aligned}$$

which means that the inequality (2.37) holds for  $j = m$  or  $j = m - 1$ . Assume the inequality (2.37) is valid for  $j = l$  and  $j = l - 1$ . Using the formulas (2.11), (2.12), (2.28), and (2.36), we obtain

$$\begin{aligned} \|\phi^{(l-2)}\|_{H^{m-l+2}(\Omega)}^2 & \leq C_\gamma\|\mathcal{B}_\phi(t)\phi^{(l-2)}\|_{H^{m-l}(\Omega)}^2 + C_\gamma\|\phi^{(l-2)}\|_{H^{m-l+1}(\Omega)}^2 \\ & \leq C_\gamma\|\phi^{(l)}\|_{H^{m-l}(\Omega)}^2 + C_\gamma\|f_y(x, \dot{\phi})\phi^{(l-1)}\|_{H^{m-l}(\Omega)}^2 + C_\gamma\mathcal{L}(t) \\ & \quad + \varepsilon\|\phi^{(l-2)}\|_{H^{m-l+2}(\Omega)}^2 + C_{\varepsilon,\gamma}\|\phi^{(l-2)}\|_{L^2(\Omega)}^2 \\ & \leq C_\gamma\|\phi^{(l)}\|_{H^{m-l}(\Omega)}^2 + C_\gamma\|\phi^{(l-1)}\|_{H^{m-l+1}(\Omega)}^2 + C_\gamma\mathcal{L}(t) \\ & \quad + \varepsilon\|\phi^{(l-2)}\|_{H^{m-l+2}(\Omega)}^2 + C_{\varepsilon,\gamma}\|\phi^{(l-2)}\|_{L^2(\Omega)}^2, \end{aligned}$$

where the following inequality is used:

$$\|f_y(x, \dot{\phi})\phi^{(l-1)}\|_{H^{m-l}(\Omega)}^2 \leq C_\gamma\|f_y(x, \dot{\phi})\|_{H^{m-2}(\Omega)}^2\|\phi^{(l-1)}\|_{H^{m-l+1}(\Omega)}^2.$$

By induction the estimate (2.37) follows.

On the other hand, using an ellipticity estimate for the operator  $\mathcal{N}_\phi(t)$  in (2.15) similar to that for  $\mathcal{B}_\phi(t)$  in Lemma 2.3 and (2.16), we obtain

$$\begin{aligned} \|\phi(t, x)\|_{H^m(\Omega)}^2 & \leq C_\gamma\|\mathcal{N}_\phi(t)\phi\|_{H^{m-2}(\Omega)}^2 + C_\gamma\|\phi(t, x)\|_{H^{m-1}(\Omega)}^2 \\ (2.38) \quad & \leq C_\gamma\|\phi^{(2)}\|_{H^{m-2}(\Omega)}^2 + C_\gamma\|f(x, \dot{\phi})\|_{H^{m-2}(\Omega)}^2 + C_\gamma\mathcal{L}(t) \\ & \quad + \varepsilon\|\phi\|_{H^m(\Omega)}^2 + C_{\varepsilon,\gamma}\|\phi\|_{L^2(\Omega)}^2. \end{aligned}$$

Combining (2.37) and (2.38), the proof of the estimate (2.24) is complete.

Let  $V_j(t)$  and  $V_0(t)$  be defined by (2.14) and (2.17), respectively. We differentiate  $V_j(t)$  and have

$$\begin{aligned} (2.39) \quad \dot{V}_j(t) & = 2(\phi^{(j+2)}(t), \phi^{(j+1)}(t))_{L^2(\Omega)} \\ & \quad + (\dot{B}_\phi(t)\nabla\phi^{(j)}, \nabla\phi^{(j)})_{L^2(\Omega)} + 2(B_\phi(t)\nabla\phi^{(j)}, \nabla\phi^{(j+1)})_{L^2(\Omega)}. \end{aligned}$$

Using the formulas (2.12) and (2.11) in (2.39), and by Lemma 2.2, we obtain

$$\begin{aligned} (2.40) \quad -\dot{V}_j(t) & = -(\dot{B}_\phi(t)\nabla\phi^{(j)}, \nabla\phi^{(j)})_{L^2(\Omega)} + 2(r_j(t), \phi^{(j+1)}(t, x))_{L^2(\Omega)} \\ & \quad + 2(f_y(x, \dot{\phi})\phi^{(j+1)}, \phi^{(j+1)})_{L^2(\Omega)} \\ & \leq C_\gamma\mathcal{L}(t) + 2(f_y(x, \dot{\phi})\phi^{(j+1)}, \phi^{(j+1)})_{L^2(\Omega)} \\ & \leq C_\gamma\mathcal{L} + 2(f_y(x, \dot{\phi})\phi^{(j+1)}, \phi^{(j+1)})_{L^2(\Omega)}. \end{aligned}$$

Moreover, we also have, by (2.16),

$$(2.41) \quad -\dot{V}_0(t) = -(\dot{\mathcal{N}}_\phi(t) \nabla \phi, \nabla \phi)_{L^2(\Omega)} + 2(f(x, \dot{\phi}), \dot{\phi})_{L^2(\Omega)}.$$

The inequalities (2.26) and (2.25) follow from (2.40) and (2.41).

Similarly we have

$$\dot{P}(t) \leq C_\gamma \mathcal{E}^{3/2}(t) + P(t), \quad 0 \leq t \leq T,$$

which yields (2.27) by Gronwall's inequality.

**3. Energy estimates in integral form.** In this section, we establish the following.

**THEOREM 3.1.** *Let all assumptions in Theorem 1.1 hold. Let  $\gamma > 0$  be given and let  $\phi$  be a solution to the problem (2.1) on the interval  $[0, T]$  for some  $T > 0$  such that the condition (2.23) holds. Then there are constants  $C_\gamma > 0$  and  $T_\gamma > 3 \sup_{x \in \bar{\Omega}} |H|_g / \alpha$  such that for  $0 \leq s \leq t \leq T$ ,  $t - s \geq T_\gamma$ ,*

$$(3.1) \quad \int_s^t \mathcal{Q}(\tau) d\tau + C_\gamma \mathcal{Q}(t) \leq C_\gamma \mathcal{Q}(s) + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau,$$

where  $\mathcal{L}(t)$  and  $\alpha$  are as given in (2.22) and (1.13), respectively.

Let  $\gamma > 0$  be given and let  $\phi$  satisfy the problem (2.1) on the interval  $[0, T]$  for some  $T > 0$  such that the condition (2.23) holds. For  $t \in [0, T]$ , let  $g_\phi$  be the metric on  $\bar{\Omega}$  given by

$$(3.2) \quad g_\phi = B_\phi^{-1}(t),$$

where the matrix  $B_\phi(t)$  is defined by (2.7). Consider the pair  $(\bar{\Omega}, g_\phi)$  as a Riemannian manifold for fixed  $t$ . Let  $X, Y$  be vector fields on  $\bar{\Omega}$  and let  $f$  be a function. Then

$$(3.3) \quad \langle X, Y \rangle_{g_\phi} = \langle B_\phi^{-1}(t) X, Y \rangle, \quad \nabla_{g_\phi} f = B_\phi(t) \nabla f,$$

where  $\langle \cdot, \cdot \rangle$  and  $\langle \cdot, \cdot \rangle_{g_\phi}$  are the products in the standard metric and the metric  $g_\phi$ , respectively, and where  $\nabla$  and  $\nabla_{g_\phi}$  are the gradients in the standard metric and the metric  $g_\phi$ , respectively. In addition, it is easy to check from (3.3) that there are constants  $C_{0,\gamma} > 0$  and  $C_\gamma > 0$  such that, for  $t \in [0, T]$ ,

$$(3.4) \quad C_{0,\gamma} |\nabla_g f|_g^2 \leq |\nabla_{g_\phi} f|_{g_\phi}^2 = \langle B_\phi(t) \nabla f, \nabla f \rangle \leq C_\gamma |\nabla_g f|_g^2, \quad f \in C^\infty(\Omega),$$

under the condition (2.23). Let

$$(3.5) \quad (b_{ij}(x, y))^{-1} = (b^{ij}(x, y)), \quad (x, y) \in \bar{\Omega} \times \mathbb{R}^n,$$

where  $(b_{ij}(x, y))$  are as given in (2.4). Then

$$(3.6) \quad A^{-1}(x, \nabla w) = (b^{ij}(x, 0)).$$

Let  $D_{g_\phi}$  and  $D_g$  be the Levi-Civita connections in the Riemannian metrics  $g_\phi$  and  $g$ , respectively. Let  $H$  be a vector field on  $\bar{\Omega}$ . Denote by  $D_{g_\phi} H$  and  $D_g H$  the covariant differentials in the metrics  $g_\phi$  and  $g$ , respectively. They are two order tensor fields on  $\bar{\Omega}$ . We define

$$(3.7) \quad \zeta = D_{g_\phi} H - D_g H.$$

LEMMA 3.1 (see [37]). Let  $H$  be a vector field on  $\overline{\Omega}$ . Suppose that the tensor field  $\zeta = \zeta(\cdot, \cdot)$  of order two is given in the formula (3.7). Let  $\gamma > 0$  be given and let  $\phi$  be such that  $\sup_{x \in \overline{\Omega}} |\nabla \phi| \leq \gamma$ . Then there exists a constant  $C_\gamma > 0$  such that

$$(3.8) \quad |\zeta(X, Y)| \leq C_\gamma(|D\phi| + |D^2\phi|)|X||Y| \quad \forall X, Y \in R_x^n, \quad x \in \overline{\Omega},$$

where  $D$  is the covariant differential of the standard product of the Euclidean space  $R^n$ .

LEMMA 3.2. Let  $\phi^{(j)}$  be a solution to (2.12) and let  $\hat{\Omega} \subseteq \Omega$  be a subset.

(1) Suppose that  $\mathcal{H}$  is a vector field on  $\hat{\Omega}$ . Then

$$(3.9) \quad \begin{aligned} & \int_s^t \int_{\partial \hat{\Omega}} \left[ \mathcal{H}(\phi^{(j)}) \phi_{\nu_B}^{(j)} + \frac{1}{2}((\phi^{(j)})^2 - |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2) \langle \mathcal{H}, \nu \rangle \right] d\sigma d\tau \\ &= (\phi^{(j)}, \mathcal{H}(\phi^{(j)}))_{L^2(\hat{\Omega})} \Big|_s^t + \int_s^t \int_{\hat{\Omega}} [r_j \mathcal{H}(\phi^{(j)}) + f_y(x, \dot{\phi}) \phi^{(j)} \mathcal{H}(\phi^{(j)})] dx d\tau \\ & \quad + \int_s^t \int_{\hat{\Omega}} \left\{ D_{g_\phi} \mathcal{H}(\nabla_{g_\phi} \phi^{(j)}, \nabla_{g_\phi} \phi^{(j)}) + \frac{1}{2}(\phi^{(j)})^2 - |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 \right\} \operatorname{div} \mathcal{H} \Big\} dx d\tau. \end{aligned}$$

(2) Let  $h \in C^2(\hat{\Omega})$ . We have

$$(3.10) \quad \begin{aligned} & \int_s^t \int_{\hat{\Omega}} h[(\phi^{(j)})^2 - |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2] dx d\tau \\ &= (\phi^{(j)}, h\phi^{(j)})_{L^2(\hat{\Omega})} \Big|_s^t + \int_s^t \int_{\partial \hat{\Omega}} \left( \frac{1}{2}(\phi^{(j)})^2 h_{\nu_B} - h\phi^{(j)} \phi_{\nu_B}^{(j)} \right) d\sigma d\tau \\ & \quad + \int_s^t \int_{\hat{\Omega}} \left[ r_j(\tau) h\phi^{(j)} + f_y(x, \dot{\phi}) \phi^{(j)} h\phi^{(j)} - \frac{1}{2}(\phi^{(j)})^2 \mathcal{B}_\phi(\tau) h \right] dx d\tau. \end{aligned}$$

LEMMA 3.3. Let all assumptions in Theorem 1.1 hold. Let  $\gamma > 0$  be given and let  $\phi$  be a solution to the problem (2.1) on the interval  $[0, T]$  for some  $T > 0$  such that the condition (2.23) holds. Then there exist constants  $C_\gamma > 0$  and  $T_\gamma > 3 \sup_{x \in \overline{\Omega}} |H|_g / \alpha$  such that, if  $0 \leq s \leq t \leq T$ ,  $t - s \geq T_\gamma$ , then

$$(3.11) \quad \begin{aligned} \int_s^t \mathcal{Q}(\tau) d\tau &\leq C_\gamma \int_s^t \mathcal{L}(\tau) d\tau + C_\gamma \int_s^t R(\tau) d\tau \\ & \quad + C_\gamma \int_s^t \int_{\Omega} (f_y(x, \dot{\phi}) \phi^{(j+1)} \phi^{(j+1)} + f(x, \dot{\phi}) \dot{\phi}) dx d\tau, \end{aligned}$$

where  $R(t)$  and  $\mathcal{L}(t)$  are as given in (2.21) and (2.22), respectively.

*Proof.* For  $0 < \varepsilon_2 < \varepsilon_1 < \varepsilon_0 < \varepsilon$ , set

$$(3.12) \quad Q_k = \mathcal{N}_{\varepsilon_k} \left[ \bigcup_{i=1}^I \Gamma_0^i \cup \left( \Omega \setminus \bigcup_{j=1}^I \Omega_i \right) \right], \quad k = 0, 1, 2.$$

Obviously we have

$$(3.13) \quad Q_2 \subset Q_1 \subset \overline{Q_0} \subset G.$$

Let  $\beta^i, i = 1, \dots, I$ , satisfy

$$(3.14) \quad \begin{aligned} \beta^i &\in C_0^\infty(R^n), \quad 0 \leq \beta^i \leq 1, \\ \beta^i &= \begin{cases} 1 & \text{on } \overline{\Omega_i} \setminus Q_1, \\ 0 & \text{on } Q_2. \end{cases} \end{aligned}$$

For each  $i, 1 \leq i \leq I$ , set

$$(3.15) \quad \hat{\Omega} := \Omega_i, \quad \mathcal{H} := \beta^i H^i, \quad h := \frac{1}{2} \operatorname{div}(\beta^i H^i).$$

*Step 1.* For  $1 \leq j \leq m-1$ , we estimate the energy  $V_j(t)$ .

(1) We estimate the integral

$$\int_s^t \int_{\Omega \setminus Q_1} |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 dx d\tau,$$

where  $|\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 = \langle B_\phi(t) \nabla \phi^{(j)}, \nabla \phi^{(j)} \rangle$ .

Applying the identity (3.9), we obtain

$$(3.16) \quad \begin{aligned} &\int_s^t \int_{\partial\Omega_i} \left[ \beta^i H^i(\phi^{(j)}) \phi_{\nu_B}^{(j)} + \frac{1}{2} (\dot{\phi}^{(j)})^2 - |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 \langle \beta^i H^i, \nu \rangle \right] d\sigma d\tau \\ &\quad - \frac{1}{2} \int_s^t \int_{\partial\Omega_i} \left[ \frac{1}{2} (\phi^{(j)})^2 (\operatorname{div} \beta^i H^i)_{\nu_B} - \operatorname{div}(\beta^i H^i) \phi^{(j)} \phi_{\nu_B}^{(j)} \right] d\sigma d\tau \\ &= \left( \phi^{(j)}, \frac{1}{2} \operatorname{div}(\beta^i H^i) \phi^{(j)} \right)_{L^2(\Omega_i)} \Big|_s^t + \int_s^t \int_{\Omega_i} D_{g_\phi}(\beta^i H^i)(\nabla_{g_\phi} \phi^{(j)}, \nabla_{g_\phi} \phi^{(j)}) dx d\tau \\ &\quad + (\phi^{(j)}, \beta^i H^i(\phi^{(j)}))_{L^2(\Omega_i)} \Big|_s^t - \frac{1}{4} \int_s^t \int_{\Omega_i} (\phi^{(j)})^2 \mathcal{B}_\phi(\tau) \operatorname{div}(\beta^i H^i) dx d\tau \\ &\quad + \int_s^t \int_{\Omega_i} [r_j(\tau) + f_y(x, \dot{\phi}) \phi^{(j)}] \left[ \beta^i H^i(\phi^{(j)}) + \frac{1}{2} \operatorname{div}(\beta^i H^i) \phi^{(j)} \right] dx d\tau. \end{aligned}$$

Next, we show that the left-hand side of (3.16) is nonpositive. In fact, we notice that

$$\partial\Omega_i = \Gamma_0^i \cup (\partial\Omega_i \setminus \Gamma) \cup ((\partial\Omega_i \setminus \Gamma_0^i) \cap \Gamma) \equiv I_1 \cup I_2.$$

Since  $\partial\Omega_i \setminus \Gamma \subseteq \Omega \setminus \cup_{i=1}^I \Omega_i$ , we have  $\beta^i = 0$  in  $I_1 \subseteq Q_2$ . Since  $I_2 \subseteq \Gamma$ , it follows from [34] that

$$\beta^i H^i(\phi^{(j)}) \phi_{\nu_B}^{(j)} = |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 \langle \beta^i H^i, \nu^i \rangle, \quad \phi^{(j)} = 0 \quad \text{on } I_2.$$

We get

$$(3.17) \quad \begin{aligned} &\int_s^t \int_{\partial\Omega_i} \left[ \beta^i H^i(\phi^{(j)}) \phi_{\nu_B}^{(j)} + \frac{1}{2} (\dot{\phi}^{(j)})^2 - |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 \langle \beta^i H^i, \nu \rangle \right] d\sigma d\tau \\ &= \frac{1}{2} \int_s^t \int_{(\partial\Omega_i \setminus \Gamma_0^i) \cap \Gamma} |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 \beta^i H^i \nu d\sigma d\tau \leq 0. \end{aligned}$$



Furthermore, we decompose  $\partial\Omega_i = (\partial\Omega_i \setminus \Omega) \cup (\partial\Omega_i \cap \Omega) \equiv J_1 \cup J_2$ . Since  $J_1 \subseteq \Gamma$ , by the boundary condition, we get

$$(3.18) \quad -\frac{1}{2} \int_s^t \int_{J_1} \left[ \frac{1}{2} \phi^{(j)2} (\operatorname{div} \beta^i H^i)_{\nu_B} - \operatorname{div}(\beta^i H^i) \phi^{(j)} \phi_{\nu_B}^{(j)} \right] dx d\tau = 0.$$

It is easy to verify that  $J_2 = \partial\Omega_i \cap \Omega \subseteq \Omega \setminus \cup_{i=1}^I \Omega_i \subseteq Q_2$ ; thus

$$(3.19) \quad -\frac{1}{2} \int_s^t \int_{J_2} \left[ \frac{1}{2} \phi^{(j)2} (\operatorname{div} \beta^i H^i)_{\nu_B} - \operatorname{div}(\beta^i H^i) \phi^{(j)} \phi_{\nu_B}^{(j)} \right] dx d\tau = 0.$$

Combining (3.17)–(3.19), it follows that the left-hand side of (3.16) is nonpositive.

We sum up (3.16) from  $i = 1$  to  $i = I$  and obtain

$$(3.20) \quad \begin{aligned} & \sum_{i=1}^I \int_s^t \int_{\Omega_i \setminus Q_1} D_{g_\phi}(\beta^i H^i)(\nabla_{g_\phi} \phi^{(j)}, \nabla_{g_\phi} \phi^{(j)}) dx d\tau \\ & \leq - \sum_{i=1}^I (\dot{\phi}^{(j)}, M^i(\phi^{(j)}))_{L^2(\Omega_i)} \Big|_s^t + \frac{1}{4} \sum_{i=1}^I \int_s^t \int_{\Omega_i} (\phi^{(j)})^2 \mathcal{B}_\phi(\tau) \operatorname{div}(\beta^i H^i) dx d\tau \\ & \quad - \sum_{i=1}^I \int_s^t \int_{\Omega_i} r_j(\tau) M^i(\phi^{(j)}) dx d\tau - \sum_{i=1}^I \int_s^t \int_{\Omega_i} f_y(x, \dot{\phi}) \dot{\phi}^{(j)} M^i(\phi^{(j)}) dx d\tau \\ & \quad - \sum_{i=1}^I \int_s^t \int_{\Omega_i \cap Q_1} D_{g_\phi}(\beta^i H^i)(\nabla_{g_\phi} \phi^{(j)}, \nabla_{g_\phi} \phi^{(j)}) dx d\tau, \end{aligned}$$

where  $M^i(\phi^{(j)}) = \beta^i H^i(\phi^{(j)}) + \frac{1}{2} \operatorname{div}(\beta^i H^i) \phi^{(j)}$ . We estimate every item on the right-hand side of the inequality (3.20) as

$$(3.21) \quad \sum_{i=1}^I \left| (\dot{\phi}^{(j)}, M^i(\phi^{(j)}))_{L^2(\Omega_i)} \Big|_s^t \right| \leq C_\gamma V_j(t) + C_\gamma V_j(s),$$

$$(3.22) \quad \sum_{i=1}^I \left| - \int_s^t \int_{\Omega_i} \frac{1}{4} \phi^{(j)2} \mathcal{B}_\phi(t) \operatorname{div}(\beta^i H^i) \right| \leq C_r \int_s^t \int_{\Omega} (\phi^{(j)})^2 dx d\tau,$$

$$(3.23) \quad \begin{aligned} & \sum_{i=1}^I \left| \int_s^t \int_{\Omega_i} r_j(t) M^i(\phi^{(j)}) dx dt \right| \\ & \leq C_\gamma \int_s^t \left( \int_{\Omega} r_j^2(t) dx \right)^{\frac{1}{2}} \cdot \left( \int_{\Omega_i} (M^i(\phi^{(j)}))^2 dx \right)^{\frac{1}{2}} dt \\ & \leq C_\gamma \int_s^t V_j^{\frac{1}{2}}(\tau) \cdot \left( \sum_{k=2}^{m-1} \mathcal{E}^{\frac{k}{2}}(\tau) \right) d\tau \leq \varepsilon \int_s^t V_j(\tau) d\tau + C_\varepsilon \int_s^t \mathcal{L}(\tau) d\tau. \end{aligned}$$

Due to (2.40), there exists  $C_\gamma > 0$  such that

$$(3.24) \quad \dot{V}_j(t) - C_\gamma \mathcal{L}(t) \leq -2 \int_{\Omega} f_y(x, \dot{\phi}) (\phi^{(j+1)})^2 dx.$$

Thus we have

$$\begin{aligned}
 (3.25) \quad & \sum_{i=1}^I \left| \int_s^t \int_{\Omega_i} f_y(x, \dot{\phi}) \dot{\phi}^{(j)} \left[ \beta^i H^i(\phi^{(j)}) + \frac{1}{2} \operatorname{div}(\beta^i H^i) \phi^{(j)} \right] \right| \\
 & \leq C_\gamma \int_s^t V_j^{\frac{1}{2}}(\tau) \left[ \int_{\Omega} f_y^2(x, \dot{\phi}) (\dot{\phi}^{(j)})^2 dx \right]^{\frac{1}{2}} d\tau \\
 & \leq C_\gamma \int_s^t V_j^{\frac{1}{2}}(\tau) (C_\gamma \mathcal{L}(\tau) - \dot{V}_j(\tau))^{\frac{1}{2}} d\tau \\
 & \leq \varepsilon \int_s^t V_j(\tau) d\tau + C_\varepsilon \int_s^t (C_\gamma \mathcal{L}(\tau) - \dot{V}_j(\tau)) d\tau \\
 & \leq \varepsilon \int_s^t V_j(\tau) d\tau + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau + C_\varepsilon V_j(s).
 \end{aligned}$$

We also have

$$(3.26) \quad \sum_{i=1}^I \left| \int_s^t \int_{\Omega_i \cap Q_1} D_{g_\phi}(\beta^i H^i)(\nabla_{g_\phi} \phi^{(j)}, \nabla_{g_\phi} \phi^{(j)}) dx d\tau \right| \leq C_\gamma \int_s^t \int_{\Omega \cap Q_1} |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 dx d\tau.$$

On the other hand, by (3.7)–(3.8) and (1.13), (3.4), (3.14), we have

$$\begin{aligned}
 (3.27) \quad & \sum_{i=1}^I \int_s^t \int_{\Omega_i \setminus Q_1} D_{g_\phi}(\beta^i H^i)(\nabla_{g_\phi} \phi^{(j)}, \nabla_{g_\phi} \phi^{(j)}) dx d\tau \\
 & = \sum_{i=1}^I \int_s^t \int_{\Omega_i \setminus Q_1} [D_g(\beta^i H^i)(\nabla_{g_\phi} \phi^{(j)}, \nabla_{g_\phi} \phi^{(j)}) + \eta(\nabla_{g_\phi} \phi^{(j)}, \nabla_{g_\phi} \phi^{(j)})] dx d\tau \\
 & \geq \alpha \int_s^t \int_{\Omega \setminus Q_1} |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 dx d\tau - C_\gamma \int_s^t \mathcal{L}(\tau) d\tau \\
 & \geq \alpha C_{0,\gamma} \int_s^t \int_{\Omega \setminus Q_1} |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 dx d\tau - C_\gamma \int_s^t \mathcal{L}(\tau) d\tau.
 \end{aligned}$$

Combining (3.20)–(3.27), we obtain

$$\begin{aligned}
 (3.28) \quad & \alpha C_{0,\gamma} \int_s^t \int_{\Omega \setminus Q_1} |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 dx d\tau \leq C_\gamma V_j(t) + C_\gamma V_j(s) + \varepsilon \int_s^t V_j(\tau) d\tau \\
 & + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau + C_\gamma \int_s^t \int_{\Omega} (\phi^{(j)})^2 dx d\tau + C_\gamma \int_s^t \int_{\Omega \cap Q_1} |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 dx d\tau.
 \end{aligned}$$

(2) We estimate

$$\int_s^t \int_{\Omega \cap Q_1} |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 dx d\tau.$$

Let  $\xi \in C_0^\infty(R^n)$  such that  $0 \leq \xi \leq 1$ , and let

$$(3.29) \quad \xi = \begin{cases} 0 & \text{on } R^n \setminus Q_0, \\ 1 & \text{on } Q_1. \end{cases}$$

We have

$$(3.30) \quad 0 = \int_s^t \int_\Omega [\ddot{\phi}^{(j)} - \mathcal{B}_\phi(\tau)\phi^{(j)} + r_j(\tau) + f_y(x, \dot{\phi})\dot{\phi}^{(j)}] \xi \phi^{(j)} dx d\tau,$$

that is,

$$\begin{aligned} 0 &= \int_\Omega \dot{\phi}^{(j)} \xi \phi^{(j)} dx \Big|_s^t - \int_s^t \int_\Omega \xi (\dot{\phi}^{(j)})^2 dx d\tau + \int_s^t \int_\Omega f_y(x, \dot{\phi}) \dot{\phi}^{(j)} \xi \phi^{(j)} dx d\tau \\ &\quad + \int_s^t \int_\Omega [\xi |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 + \phi^{(j)} \langle \nabla_{g_\phi} \phi^{(j)}, \nabla_{g_\phi} \xi \rangle_{g_\phi}] dx d\tau + \int_s^t \int_\Omega r_j(\tau) \xi \phi^{(j)} dx d\tau. \end{aligned}$$

We can easily get

$$\begin{aligned} (3.31) \quad &\int_s^t \int_{\Omega \cap Q_1} |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 = \int_s^t \int_{\Omega \cap Q_1} \xi |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 \leq \int_s^t \int_\Omega \xi |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 \\ &\leq C_\gamma V_j(s) + C_\gamma V_j(t) + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau + C_\varepsilon \int_s^t \int_\Omega (\phi^{(j)})^2 dx d\tau \\ &\quad + \varepsilon \int_s^t \int_\Omega |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 + \varepsilon \int_s^t \int_\Omega (\dot{\phi}^{(j)})^2 dx d\tau + C_\gamma \int_s^t \int_{Q_0} (\phi^{(j)})^2 dx d\tau. \end{aligned}$$

(3) We need to further estimate

$$\int_s^t \int_{Q_0} (\dot{\phi}^{(j)})^2 dx d\tau.$$

By (1.7) and (3.13), we see that there exists  $c_0 > 0$  such that

$$(3.32) \quad f_s(x, 0) \geq c_0 > 0, \quad x \in \overline{Q_0}.$$

We can take  $\gamma$  small in (2.23), which can guarantee that

$$(3.33) \quad f_s(x, \dot{\phi}) \geq \frac{1}{2}c_0, \quad x \in \overline{Q_0},$$

by the embedding theorem. Since  $\dot{\phi}^{(j)}(x) = 0$  on  $\partial\Omega$ , we obtain

$$\begin{aligned} (3.34) \quad &\frac{1}{2}c_0 \int_s^t \int_{Q_0} (\dot{\phi}^{(j)})^2 dx d\tau \leq \int_s^t \int_\Omega f_y(x, \dot{\phi}) (\dot{\phi}^{(j)})^2 dx d\tau \\ &\leq C_\gamma V_j(s) + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau. \end{aligned}$$

(4) Taking  $h := \frac{1}{2}\alpha C_{0,\gamma}$  and  $\hat{\Omega} = \Omega$  in the identity (3.10) yields

$$(3.35) \quad \begin{aligned} \frac{1}{2}\alpha C_{0,\gamma} \int_s^t \int_{\Omega} [(\dot{\phi}^{(j)})^2 - |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2] dx d\tau &\leq C_\gamma V_j(s) + C_\gamma V_j(t) + \varepsilon \int_s^t V_j(\tau) d\tau \\ &+ C_\varepsilon \int_s^t \int_{\Omega} (\phi^{(j)})^2 dx d\tau + C_\gamma \int_s^t \int_{Q_0} (\dot{\phi}^{(j)})^2 dx d\tau + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau. \end{aligned}$$

Substituting (3.31) and (3.34) into (3.28), we obtain

$$(3.36) \quad \begin{aligned} \alpha C_{0,\gamma} \int_s^t \int_{\Omega} |\nabla_{g_\phi} \phi^{(j)}|_{g_\phi}^2 dx d\tau &\leq C_\gamma V_j(s) + C_\gamma V_j(t) + C_\varepsilon \int_s^t \int_{\Omega} (\phi^{(j)})^2 dx d\tau \\ &+ \varepsilon \int_s^t V_j(\tau) d\tau + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau. \end{aligned}$$

Combining the inequalities (3.34)–(3.36), for  $1 \leq j \leq m-1$ , we finally get

$$(3.37) \quad \begin{aligned} \frac{1}{2}\alpha C_{0,\gamma} \int_s^t V_j(\tau) d\tau \\ \leq C_\gamma V_j(t) + C_\gamma V_j(s) + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau + C_\gamma \int_s^t \int_{\Omega} (\phi^{(j)})^2 dx d\tau. \end{aligned}$$

*Step 2.* We shall estimate  $V_0(t)$ . We define

$$(3.38) \quad h_\phi = N_\phi^{-1}(t),$$

and replace  $g_\phi$  in Step 1 by  $h_\phi$ . By a similar process, there exists a constant denoted by  $C_\gamma$  such that

$$(3.39) \quad \dot{V}_0(t) - C_\gamma \mathcal{L}(t) \leq -2 \int_{\Omega} f(x, \dot{\phi}) \dot{\phi} dx.$$

In light of

$$(3.40) \quad f(x, \dot{\phi}) - f(x, 0) = \left( \int_0^1 f_y(x, s \nabla \phi) ds \right) \dot{\phi},$$

we have

$$(3.41) \quad \begin{aligned} \sum_{i=1}^I \left| \int_s^t \int_{\Omega_i} f(x, \dot{\phi}) M^i(\phi) \right| &\leq C_\gamma \int_s^t V_0^{1/2}(\tau) \left( \int_{\Omega} f^2(x, \dot{\phi}) dx \right)^{1/2} d\tau \\ &\leq C_\gamma \int_s^t V_0^{1/2}(\tau) \left[ \int_{\Omega} f(x, \dot{\phi}) \dot{\phi} dx \right]^{1/2} d\tau \\ &\leq C_\gamma \int_s^t V_0^{1/2}(\tau) \left[ C_\gamma \mathcal{L}(\tau) - \dot{V}_0(\tau) \right]^{1/2} d\tau \\ &\leq \varepsilon \int_s^t V_0(\tau) d\tau + C_\varepsilon \int_s^t \mathcal{L}(\tau) d\tau + C_\gamma V_0(t) + C_\gamma V_0(s). \end{aligned}$$

Using a method similar to that in Step 1, we obtain

$$(3.42) \quad \frac{1}{2}\alpha C_{0,\gamma} \int_s^t V_0(\tau) d\tau \leq C_\gamma \left( V_0(t) + V_0(s) + \int_s^t \mathcal{L}(\tau) d\tau \right) + C_\gamma \int_s^t \int_\Omega \phi^2 dx d\tau.$$

Step 3. Combining (3.37) and (3.42), we have

$$(3.43) \quad \frac{1}{2}\alpha C_{0,\gamma} \int_s^t Q(\tau) d\tau \leq C_\gamma \left( Q(t) + Q(s) + \int_s^t \mathcal{L}(\tau) d\tau \right) + C_\gamma \int_s^t R(\tau) dx d\tau,$$

where  $R(t)$  is given in (2.21). Moreover, the inequalities (2.26) in section 2 imply that

$$(3.44) \quad \max\{Q(t), Q(s)\} \leq \frac{1}{t-s} \int_s^t Q(\tau) d\tau + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau \\ + 2 \int_s^t \int_\Omega (f_y(x, \dot{\phi})(\phi^{(j+1)})^2 + f(x, \dot{\phi})\dot{\phi}) dx d\tau.$$

Substituting (3.44) into (3.43), we finally obtain  $C_\gamma$  and  $T_\gamma > 3 \sup_{x \in \bar{\Omega}} |H|_g / \alpha$  such that the inequality (3.11) holds.  $\square$

We now absorb the lower order terms in (3.11). This will be accomplished by applying the following nonlinear version of the compactness-uniqueness argument.

LEMMA 3.4. *Let all assumptions in Theorem 3.1 hold. Then for  $1 \leq j \leq m-1$  we have*

$$(3.45) \quad \int_s^t \int_\Omega (\phi^{(j)})^2 dx d\tau \leq C_\gamma \int_s^t \mathcal{L}(\tau) d\tau + C_\gamma \int_s^t \int_\Omega f_y(x, \dot{\phi})(\phi^{(j+1)})^2 dx d\tau,$$

and for  $j = 0$ ,

$$(3.46) \quad \int_s^t \int_\Omega \phi^2 dx d\tau \leq C_\gamma \int_s^t \mathcal{L}(\tau) d\tau + C_\gamma \int_s^t \int_\Omega f(x, \dot{\phi})\dot{\phi} dx d\tau,$$

respectively, for  $t - s \geq T_0$ , where  $T_0 \geq 3 \sup_{x \in \bar{\Omega}} |H|_g / \alpha$ .

*Proof.* By Lemma 4 in [8], we need only prove that for some  $T_0$  large enough there is  $C_\gamma > 0$  such that

$$(3.47) \quad \int_0^{T_0} \int_\Omega (\phi^{(j)})^2 dx d\tau \leq C_\gamma \int_0^{T_0} \mathcal{L}(\tau) d\tau + C_\gamma \int_0^{T_0} \int_\Omega f_y(x, \dot{\phi})(\phi^{(j+1)})^2 dx d\tau.$$

For convenience, we denote  $\psi := \phi^{(j)}$  and  $V(t) := V_j(t)$  for  $1 \leq j \leq m-1$ . Recall that there exists a constant  $C_\gamma > 0$  such that

$$\dot{V}(t) - C_\gamma \mathcal{L}(t) \leq -2 \int_\Omega f_y(x, \dot{\phi})\dot{\psi}^2 dx < 0,$$

which means that  $V(t) - C_\gamma \int_0^t \mathcal{L}(\tau) d\tau$  is decreasing. Hence we have

$$T_0 \left( V(T_0) - C_\gamma \int_0^{T_0} \mathcal{L}(\tau) d\tau \right) \leq \int_0^{T_0} \left( V(t) - C_\gamma \int_0^t \mathcal{L}(\tau) d\tau \right) dt,$$

that is,

$$T_0 \left( V(T_0) - C_\gamma \int_0^{T_0} \mathcal{L}(\tau) d\tau \right) \leq \int_0^{T_0} V(\tau) d\tau.$$

Using the inequality (3.37), we obtain from the above inequality that

$$(3.48) \quad V(T_0) \leq \frac{C_\gamma}{T_0} V(0) + \left( C_\gamma + \frac{C_\gamma}{T_0} \right) \int_0^{T_0} \mathcal{L}(\tau) d\tau + \frac{C_\gamma}{T_0} \int_0^{T_0} \int_\Omega \psi^2 dx d\tau.$$

On the other hand, there exists  $C_\gamma > 0$  such that

$$-\dot{V}(t) \leq 2 \int_\Omega f_y(x, \dot{\phi}) \dot{\psi}^2 dx + C_\gamma \mathcal{L}(t).$$

Integrating the above inequality on interval  $[0, T_0]$ , we easily get

$$(3.49) \quad V(0) \leq V(T_0) + C_\gamma \int_0^{T_0} \mathcal{L}(\tau) d\tau + 2 \int_0^{T_0} \int_\Omega f_y(x, \dot{\phi}) \dot{\psi}^2 dx d\tau.$$

For  $T_0$  large enough, the inequalities (3.48) and (3.49) imply that

$$(3.50) \quad V(0) \leq C_\gamma \int_0^{T_0} \int_\Omega \psi^2 dx d\tau + C_\gamma \int_0^{T_0} \mathcal{L}(\tau) d\tau + 2 \int_0^{T_0} \int_\Omega f_y(x, \dot{\phi}) \dot{\psi}^2 dx d\tau.$$

Now we prove (3.47) by contradiction. Let  $\phi_k(t)$  be a sequence of solutions to the system (2.1) such that

$$(3.51) \quad \int_0^{T_0} \int_\Omega \psi_k^2 dx d\tau = 1,$$

$$(3.52) \quad \lim_{k \rightarrow \infty} \int_0^{T_0} \int_\Omega f_y(x, \dot{\phi}_k) \dot{\psi}_k^2 dx d\tau = 0, \quad \lim_{k \rightarrow \infty} \int_0^{T_0} \mathcal{L}_k(\tau) d\tau = 0,$$

where  $V_k$  is the energy of  $\psi_k$ .

By (3.50)–(3.52), we get

$$(3.53) \quad V_k(0) \text{ is bounded } \forall k.$$

Then there is a subsequence of  $\psi_k$  (denoted by  $\psi_k$  again) such that

$$(3.54) \quad \psi_k(0) \rightarrow \text{some } \bar{\psi}_0 \text{ weakly in } H^1(\Omega),$$

$$(3.55) \quad \dot{\psi}_k(0) \rightarrow \text{some } \bar{\psi}_1 \text{ weakly in } L^2(\Omega).$$

Denote by  $\bar{\psi}$  the solution corresponding to the initial data  $(\bar{\psi}_0, \bar{\psi}_1)$ :

$$(3.56) \quad \bar{\psi}(0, x) = \bar{\psi}_0(x), \quad \dot{\bar{\psi}}(0, x) = \bar{\psi}_1(x) \quad \text{in } \Omega.$$

Then

$$(3.57) \quad \{\psi_k, \dot{\psi}_k\} \rightarrow \{\bar{\psi}, \dot{\bar{\psi}}\} \text{ in } L^\infty(0, T_0; H^1(\Omega) \times L^2(\Omega)) \text{ weakly } \star.$$

By Aubin's compactness results (see [1]), it follows that

$$(3.58) \quad \psi_k \rightarrow \bar{\psi} \text{ strongly in } L^\infty(0, T_0; L^2(\Omega)).$$

Since

$$\int_0^{T_0} \mathcal{E}_k(\tau) d\tau \leq \sqrt{T_0} \left( \int_0^{T_0} \mathcal{L}_k(\tau) d\tau \right)^{\frac{1}{2}} \rightarrow 0,$$

we have

$$\phi_k \rightarrow 0 \text{ in } H^m((0, T_0) \times \Omega), \quad \dot{\phi}_k \rightarrow 0 \text{ in } H^{m-1}((0, T_0) \times \Omega).$$

By using the imbedding theorem, we have

$$(3.59) \quad \sup_{(\tau, x) \in [0, T_0] \times \bar{\Omega}} |\phi_k| \rightarrow 0, \quad \sup_{(\tau, x) \in [0, T_0] \times \Omega} |\dot{\phi}_k| \rightarrow 0,$$

as  $k$  goes to infinity. The relationships (3.52), (3.59), and (1.7) imply that

$$(3.60) \quad \dot{\psi}_k \rightarrow 0 \text{ strongly in } L^2(0, T_0; L^2(G)).$$

Then passing to the limit in (2.12), for  $\bar{\psi}$  we get

$$(3.61) \quad \begin{aligned} \ddot{\bar{\psi}} - \mathcal{A}\bar{\psi} &= 0 \quad \text{in } (0, T_0) \times \Omega, \\ \dot{\bar{\psi}} &= 0 \quad \text{on } (0, T_0) \times G, \\ \bar{\psi} &= 0 \quad \text{on } (0, +\infty) \times \Gamma, \end{aligned}$$

where  $\mathcal{A}\psi = \operatorname{div} A(x, \nabla w) \nabla \psi$ .

Let  $\dot{\bar{\psi}} = z$ . The above system implies

$$(3.62) \quad \begin{aligned} \ddot{z} - \mathcal{A}z &= 0 \quad \text{in } (0, T_0) \times \Omega, \\ z &= 0 \quad \text{on } (0, T_0) \times G, \\ z &= 0 \quad \text{on } (0, +\infty) \times \Gamma. \end{aligned}$$

For  $T_0 \geq 3 \sup_{x \in \bar{\Omega}} |H|_g / \alpha$ , the problem (3.62) implies  $z = 0$ . Since  $\bar{\psi}|_\Gamma = 0$ , we have  $\bar{\psi} = 0$  on  $(0, T) \times \Omega$ , which contradicts (3.51). We use the same method to absorb the lower term in (3.42), which yields (3.46).  $\square$

*Proof of Theorem 3.1.* From (3.24) we obtain

$$2 \int_\Omega f_y(x, \dot{\phi}) (\phi^{(j+1)})^2 dx \leq -\dot{V}_j(t) + C_\gamma \mathcal{L}(t)$$

for  $1 \leq j \leq m$ , and consequently

$$2 \int_s^t \int_\Omega f_y(x, \dot{\phi}) (\phi^{(j+1)})^2 dx d\tau \leq V_j(s) - V_j(t) + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau.$$

Similarly we have

$$\int_s^t \int_\Omega 2f(x, \dot{\phi}) \dot{\phi} dx d\tau \leq V_0(s) - V_0(t) + C_\gamma \int_s^t \mathcal{L}(\tau) d\tau.$$

The inequality (3.1) follows from Lemmas 3.3 and 3.4.

By an argument similar to Lemma 3.2, the inequality in (2.24) in Theorem 2.1 of section 2 can be improved into the following.

LEMMA 3.5. *Let all assumptions in Theorem 1.1 hold. Let  $\gamma > 0$  be given and let  $\phi$  satisfy the problem (2.1) on the interval  $[0, T]$  for some  $T > 0$  such that the condition (2.23) holds. Then there exist constants  $C_\gamma > 0$  and  $T_\gamma > 0$  such that*

$$(3.63) \quad \mathcal{E}(t) \leq C_\gamma \mathcal{Q}(t) + C_\gamma \mathcal{L}(t)$$

for  $t \geq T_\gamma \geq 3 \sup_{x \in \bar{\Omega}} |H|_g / \alpha$ .

#### 4. Global smooth solutions.

*Proof of Theorem 1.1.* All the notations remain the same as before. Let  $(\phi_0, \phi_1) \in H^m(\Omega) \times H^{m-1}(\Omega)$  be given such that the problem (2.1) has a short time solution. We look for a constant  $\eta_0$  such that if

$$(4.1) \quad \mathcal{E}(0) \leq \eta_0,$$

then solutions of the system (2.1) are global.

Throughout this section we take  $\gamma = 1$ . Let  $C_{0,1}, C_1 \geq 1$ , and  $T_1 > 3 \sup_{x \in \bar{\Omega}} |H|_g / \alpha$  be given according to  $\gamma = 1$  in the theorems of sections 2 and section 3. Let

$$(4.2) \quad \Xi(\eta) = \sum_{k=0}^{2m-3} \eta^{k/2},$$

$$(4.3) \quad \theta(\eta) = 2e^\eta \max\{C_1 C_{0,1}^{-1}, 2C_1 \Xi(1), 2C_1^2 \Xi(1)\eta\}$$

for  $\eta \in (0, \infty)$ .

Let

$$(4.4) \quad 0 < \eta < \min\{1, \theta^{-2}(2T_1)/4\}$$

be given. We assume that

$$(4.5) \quad \mathcal{E}(0) \leq \eta^{3/2} < \eta.$$

Then there is some  $\delta > 0$  such that

$$(4.6) \quad \mathcal{E}(t) < \eta$$

for  $t \in [0, \delta)$ . Let  $\delta_0 > 0$  be the largest number such that inequality (4.6) holds for  $t \in [0, \delta_0)$ .

We show that  $\delta_0 = +\infty$ .

*Step 1.* We claim  $\delta_0 > 2T_1$ . Suppose that this is not true, i.e.,  $\delta_0 \leq 2T_1$ . We have a contradiction as follows.

Using the inequalities (2.24), (3.44), and (2.27), we obtain

$$Q(0) \leq C_{0,1}^{-1} \mathcal{E}(0), \quad Q(t) \leq C_{0,1}^{-1} \mathcal{E}(0) + 2C_1 T_1 \eta^{3/2} \Xi(1),$$

$$P(0) \leq Q(0) + \mathcal{E}(0) \leq (C_{0,1}^{-1} + 1) \mathcal{E}(0),$$

and

$$C_1 \|\phi\|_{L^2(\Omega)}^2 \leq C_1 P(t) \leq C_1 e^{2T_1} \{(C_{0,1}^{-1} + 1) \mathcal{E}(0) + 2C_1 T_1 \eta^{3/2} \Xi(1)\}.$$



Since

$$\mathcal{L}(t) = \sum_{k=3}^{2m} \mathcal{E}^{k/2}(t) \leq \eta^{3/2} \Xi(\eta) \leq \eta^{3/2} \Xi(1),$$

we have, via (4.5) and (4.6),

(4.7)

$$\begin{aligned} \mathcal{E}(t) &\leq C_1 \mathcal{Q}(t) + C_1 \|\phi(t)\|_{L^2(\Omega)}^2 + C_1 \mathcal{L}(t) \\ &\leq [C_1 C_{0,1}^{-1} + e^{2T_1} C_1 (C_{0,1}^{-1} + 1)] \mathcal{E}(0) + [2C_1^2 T_1 \Xi(1) + 2C_1^2 T_1 e^{2T_1} \Xi(1) + C_1 \Xi(1)] \eta^{3/2} \\ &\leq 2e^{2T_1} C_1 (C_{0,1}^{-1} + 1) \mathcal{E}(0) + 2e^{2T_1} [2C_1^2 T_1 \Xi(1) + C_1 \Xi(1)] \eta^{3/2} \\ &\leq 2e^{2T_1} \max\{C_1 (C_{0,1}^{-1} + 1), 4C_1^2 T_1 \Xi(1), 2C_1 \Xi(1)\} [\mathcal{E}(0) + \eta^{3/2}] \\ &\leq \Theta(2T_1) [\mathcal{E}(0) + \eta^{3/2}] \leq 2\Theta(2T_1) \eta^{1/2} \eta < \eta \end{aligned}$$

for all  $t \in [0, \delta_0]$ . This contradicts the definition of  $\delta_0$ .

*Step 2.* Next, we prove that  $\delta_0 = +\infty$  by contradiction. Let  $T_1 \leq s < t < \delta_0$  with  $t - s \geq T_1$ . Integrating the inequality (3.63) over interval  $(s, t)$  yields

$$\int_s^t \mathcal{E}(\tau) d\tau \leq C_1 \int_s^t \mathcal{Q}(\tau) d\tau + C_1 \Xi(1) \eta^{1/2} \int_s^t \mathcal{E}(\tau) d\tau,$$

that is,

$$(4.8) \quad \int_s^t \mathcal{E}(\tau) d\tau \leq C_1 [1 - C_1 \Xi(1) \eta^{1/2}]^{-1} \int_s^t \mathcal{Q}(\tau) d\tau.$$

The inequality (4.8), in turn, shows

$$(4.9) \quad \int_s^t \mathcal{L}(\tau) d\tau \leq \Xi(1) \eta^{1/2} \int_s^t \mathcal{E}(\tau) d\tau \leq h(\eta^{1/2}) \int_s^t \mathcal{Q}(\tau) d\tau,$$

where

$$(4.10) \quad h(x) = C_1 \Xi(1) x [1 - C_1 \Xi(1) x]^{-1}, \quad 0 \leq x \leq 1.$$

Furthermore, combining the inequalities (3.1) and (4.9), we obtain

$$C_1 \mathcal{Q}(t) + \int_s^t \mathcal{Q}(\tau) d\tau \leq C_1 \mathcal{Q}(s) + C_1 h(\eta^{1/2}) \int_s^t \mathcal{Q}(\tau) d\tau,$$

that is,

$$(4.11) \quad \mathcal{Q}(t) + w(\eta) \int_s^t \mathcal{Q}(\tau) d\tau \leq \mathcal{Q}(s),$$

where  $w(\eta) = C_1^{-1} - h(\eta^{1/2})$ .

We fix  $\eta > 0$  small such that

$$(4.12) \quad \eta^{1/2} \leq \left\{ \max\{1, 4C_1 C_{0,1}^{-1} \Theta(2T_1), 2\Theta(2T_1), 2C_1 \Xi(1)\} \right\}^{-1},$$

$$w(\eta) > 0.$$

If  $\delta_0 < \infty$ , we find a contradiction as follows. Combining (3.63) and (4.11) yields

$$\begin{aligned}
 (4.13) \quad \mathcal{E}(t) &\leq C_1 \mathcal{Q}(t) + C_1 \mathcal{L}(t) \\
 &\leq C_1 \mathcal{Q}(T_1) + C_1 \Xi(1) \eta^{3/2} \\
 &\leq C_1 C_{0,1}^{-1} \mathcal{E}(T_1) + C_1 \Xi(1) \eta^{3/2} \\
 &\leq 2C_1 C_{0,1}^{-1} \Theta(2T_1) \eta^{3/2} + C_1 \Xi(1) \eta^{3/2} < \eta,
 \end{aligned}$$

for  $0 \leq t \leq \delta_0$ , where the estimate  $\mathcal{E}(T_1) \leq 2\Theta(2T_1)\eta^{3/2}$  is as obtained in Step 1. This again contradicts the definition of  $\delta_0$ .

*Proof of Theorem 1.2.* The inequality (4.11) yields

$$(4.14) \quad \mathcal{Q}(t) - \mathcal{Q}(s) + w(\eta) \int_s^t \mathcal{Q}(\tau) d\tau \leq 0$$

for  $T_1 \leq s < t < \delta_0$  with  $t - s \geq T_1$ . Thus we have

$$(4.15) \quad \int_0^{t-T_1} s^k \mathcal{Q}(s) ds \geq \frac{(t-T_1)^{k+1}}{k+1} \mathcal{Q}(t) + \frac{w(\eta)}{k+1} \int_0^{t-T_1} \tau^{k+1} \mathcal{Q}(\tau) d\tau, \quad k \geq 0,$$

which yields

$$(4.16) \quad \mathcal{Q}(t) \leq \mathcal{Q}(0) e^{-w(\eta)(t-T_1)}, \quad t \geq T_1.$$

The estimate (1.15) follows from (4.16).

*Proof of the corollaries.* We show that assumption **(H)** holds under the assumption of Corollaries 1.1 and 1.2.

(1) Let  $\kappa \leq 0$ . We will verify that the vector field  $H(x, x_0) = \rho(x) D_g \rho(x)$  meets the inequality (1.13) for all  $x \in \overline{\Omega}$  where  $\rho(x) = \rho(x, x_0)$  is the distance function of the metric  $g$  in (1.10). To this end, we consider a space form  $M$  of constant curvature  $\kappa$  with a Riemannian metric  $\langle \cdot, \cdot \rangle_M$  and use the Hessian comparison theorem.

Let  $\tilde{D}$  be the Levi-Civita connection on  $M$ . Let  $p \in M$ . Let  $\tilde{\rho}(q)$  be the distance function on  $M$  from  $p$  to  $q \in M$ . Also let  $\tilde{r} : [0, b] \rightarrow M$  be a normal geodesic from  $p$  to  $q$  such that there is no conjugate point of  $p$  on  $\tilde{r}$ ; then  $\tilde{\rho}(q) = b$ .  $\tilde{X} \in M_q$  is a vector such that  $\langle \tilde{X}, \tilde{r}'(b) \rangle_M = 0$  and  $|\tilde{X}|_M = 1$ . We have (see [33])

$$(4.17) \quad \tilde{D}^2 \tilde{\rho}(q)(\tilde{X}, \tilde{X}) = \begin{cases} \sqrt{\kappa} \cot(\sqrt{\kappa} b), & \kappa > 0, \\ \frac{1}{b}, & \kappa = 0, \\ \sqrt{-\kappa} \coth(\sqrt{-\kappa} b), & \kappa < 0. \end{cases}$$

Let  $x \in \Omega$  and  $r : [0, b] \rightarrow R^n$  be a normal minimal geodesic from  $x_0$  to  $x$  with  $\rho(x) = b$ . Let  $X \in R_x^n$  such that  $g(X, r'(b)) = 0$ , where  $g$  is defined as in (1.11).

Since  $\kappa \leq 0$ , the exponential map  $\exp_x : R^n \rightarrow R^n$  is a diffeomorphism by the Cartan-Hadamard theorem. Then  $\rho(x) = \rho(x, x_0)$  is smooth on  $\Omega \setminus x_0$ . For all  $x \in \Omega$ ,

we obtain from the Hessian comparison theorem that

$$\begin{aligned}
 (4.18) \quad D_g^2 \rho(X, X)(x) &= |X|_g^2 D_g^2 \rho \left( \frac{X}{|X|_g}, \frac{X}{|X|_g} \right) \geq |X|_g^2 \tilde{D}^2 \tilde{\rho}(\tilde{X}, \tilde{X}) \\
 &= \begin{cases} \frac{1}{\rho(x)} |X|_g^2, & \kappa = 0, \\ \sqrt{-\kappa} \coth(\sqrt{-\kappa} \rho(x)) |X|_g^2, & \kappa < 0 \end{cases} \\
 &\geq \frac{1}{\rho(x)} |X|_g^2, \quad \kappa \leq 0,
 \end{aligned}$$

where we have used the inequality

$$\sqrt{-\kappa} \rho (e^{\sqrt{-\kappa} \rho} + e^{-\sqrt{-\kappa} \rho}) \geq e^{\sqrt{-\kappa} \rho} - e^{-\sqrt{-\kappa} \rho}, \quad \kappa < 0, \quad \rho > 0.$$

For any vector  $Y \in R_x^n$ , we have the decomposition as

$$(4.19) \quad Y = X + g(Y, r'(b))r'(b),$$

where  $g(X, r'(b)) = 0$ .

Define a vector field on  $\bar{\Omega}$  by  $H(x, x_0) = \rho(x)D_g \rho(x)$ . Combining (4.18) and (4.19), we obtain

$$\begin{aligned}
 (4.20) \quad DH(Y, Y)(x) &= \rho D_g^2 \rho(Y, Y) + g^2(Y, r'(b)) \\
 &= \rho D_g^2 \rho(X, X) + g^2(Y, r'(b)) \geq |X|_g^2 + g^2(Y, r'(b)) = |Y|_g^2;
 \end{aligned}$$

that is, (1.13) is true for  $I = 1$ , and  $\Omega_1 = \Omega$ . Also the damping region  $G$  need only be supported in a neighborhood of  $\Gamma_0$ , which is defined in (1.19).

(2) We assume that  $\kappa > 0$ . Then  $\rho(x) = \rho(x, x_i)$  is smooth on each  $B(x_i, r_0) \setminus x_i$ , where  $\rho(x) \leq r_0 < \pi/2\sqrt{\kappa}$ . Using a procedure similar to (4.18), we have

$$(4.21) \quad D_g^2 \rho(X, X)(x) \geq |X|_g^2 \tilde{D}^2 \tilde{\rho}(\tilde{X}, \tilde{X}) = \sqrt{\kappa} \cot(\sqrt{\kappa} \rho(x)) |X|_g^2$$

for  $\kappa > 0$  and  $x \in \bar{B}(x_i, r_0)$ . In general, the inequality (4.21) does not hold true on the whole  $\bar{\Omega}$ . It is easily checked that the function  $t\sqrt{\kappa} \cot(\sqrt{\kappa} t)$  is monotonically decreasing in  $t \in (0, \infty)$ , and thus

$$\rho(x)\sqrt{\kappa} \cot(\sqrt{\kappa} \rho) > \sqrt{\kappa} r_0 \cot(\sqrt{\kappa} r_0) \quad \text{for } \rho(x) < r_0,$$

and

$$t\sqrt{\kappa} \cot(\sqrt{\kappa} t) \leq 1 \quad \forall t \in (0, \infty).$$

For  $\kappa > 0$ , we set

$$\alpha = \sqrt{\kappa} r_0 \cot(\sqrt{\kappa} r_0).$$

Thus, for  $x \in \bar{B}(x_i, r_0)$ , we have

$$DH^i(Y, Y)(x) = \rho D_g^2 \rho(X, X) + g^2(Y, r'(b)) \geq \alpha(|X|_g^2 + g^2(Y, r'(b))) = \alpha|Y|_g^2,$$

where  $H^i = H(x, x_i) = \rho(x, x_i)D_g \rho(x, x_i)$  and  $\alpha \leq 1$ . Since the radius of the geodesic balls is fixed, there exists a finite number of vector fields  $H^i$  satisfying the assumption **(H)**.

**5. Proof of Theorem 1.3.** Let  $T > 0$  be given and let  $u \in C([0, T], H^2(\Omega))$  satisfy the problem (1.25) on the interval  $[0, T]$ . Let  $\gamma > 0$  be given. We assume that the solution  $u$  of (1.25) satisfies

$$(5.1) \quad \|\nabla u\| \leq \gamma, \quad 0 \leq t < T.$$

Let  $g_u$  be the metric on  $\bar{\Omega}$  given by

$$(5.2) \quad g_u = A^{-1}(x, \|\nabla u\|^{2r}),$$

where  $A(x, s) = (a_{ij}(x, s))$ .

Consider the pair  $(\bar{\Omega}, g_u)$  as a Riemannian manifold for fixed  $t \in [0, T]$ . Let  $X, Y$  be vector fields on  $\bar{\Omega}$  and  $f$  be a function. Then

$$(5.3) \quad \langle X, Y \rangle_{g_u} = \langle A^{-1}(x, \|\nabla u\|^{2r})X, Y \rangle, \quad \nabla_{g_u} f = A(x, \|\nabla u\|^{2r})\nabla f,$$

where  $\langle \cdot, \cdot \rangle$  and  $\langle \cdot, \cdot \rangle_{g_u}$  are the inner products in the standard metric and the metric  $g_u$ , and  $\nabla$  and  $\nabla_{g_u}$  are the gradients in the standard metric and the metric  $g_u$ , respectively.

Let the metric  $g$  be given as in (1.28). Then under the assumptions (1.27) and (5.1) there are constants  $c_0 > 0$  and  $c_\gamma > 0$  such that

$$(5.4) \quad c_0 |\nabla_g f|_g^2 \leq |\nabla_{g_u} f|_{g_u}^2 = \langle A(x, \|\nabla u\|^{2r})\nabla f, \nabla f \rangle \leq c_\gamma |\nabla_g f|_g^2,$$

for  $t \in [0, T]$ ,  $f \in C^\infty(\Omega)$ .

Let  $D_{g_u}$  and  $D_g$  be the Levi-Civita connections in the Riemannian metrics  $g_u$  and  $g$ , respectively. Let  $H$  be a vector field on  $\bar{\Omega}$ . Denote by  $D_{g_u}H$  and  $D_gH$  the covariant differentials in the metrics  $g_u$  and  $g$ . They are two order tensor fields on  $\bar{\Omega}$ . We define

$$(5.5) \quad \zeta = D_{g_u}H - D_gH.$$

We have the following.

**LEMMA 5.1** (see [37]). *Let  $H$  be a vector field on  $\bar{\Omega}$ . Suppose that the tensor field of order two  $\zeta = \zeta(\cdot, \cdot)$  is given in the formula (5.5). Let  $\gamma > 0$  be given and let  $u$  satisfy the assumption (5.1). Then there exists a constant  $c_\gamma > 0$ , which depends only on  $\gamma$ , such that*

$$(5.6) \quad |\zeta(X, Y)| \leq c_\gamma \|\nabla u\|^{2r} |X| |Y| \quad \forall X, Y \in R_x^n, \quad x \in \bar{\Omega}.$$

We define the energy of order one associated with the problem (1.25) by

$$(5.7) \quad E_1(t) = \|\dot{u}\|_{L^2(\Omega)}^2 + (A(x, \|\nabla u\|^{2r})\nabla u, \nabla u)_{L^2(\Omega)},$$

and define the second order energy by

$$(5.8) \quad E_2(t) = (A(x, \|\nabla u\|^{2r})\nabla \dot{u}, \nabla \dot{u})_{L^2(\Omega)} + \|\operatorname{div}(A(x, \|\nabla u\|^{2r}) \cdot \nabla u)\|_{L^2(\Omega)}^2.$$

We obtain the following, which is similar to Lemma 3 in [8].

**LEMMA 5.2.** *Let  $u$  be a solution of the problem (1.25) and let  $\hat{\Omega} \subseteq \Omega$  be a subset with a boundary  $\partial\hat{\Omega}$ .*

(1) Suppose that  $\mathcal{H}$  is a vector field on  $\hat{\Omega}$ . Then

$$\begin{aligned} & \int_s^t E_1^{\frac{\beta}{2}} \int_{\partial\hat{\Omega}} \left[ \mathcal{H}(u) \frac{\partial u}{\partial \nu_a} + \frac{1}{2} ((\dot{u})^2 - |\nabla_{g_u} u|_{g_u}^2) \langle \mathcal{H}, \nu \rangle \right] d\sigma d\tau \\ &= E_1^{\frac{\beta}{2}} (\dot{u}, \mathcal{H}(u))_{L^2(\hat{\Omega})} \Big|_s^t + \int_s^t E_1^{\frac{\beta}{2}} \int_{\hat{\Omega}} f(x, \dot{u}) \mathcal{H}(u) dx d\tau \\ &\quad - \frac{\beta}{2} \int_s^t E_1^{\frac{\beta-2}{2}} E_1'(\tau) \int_{\hat{\Omega}} \dot{u} \mathcal{H}(u) dx d\tau \\ &\quad + \int_s^t E_1^{\frac{\beta}{2}} \int_{\hat{\Omega}} \left\{ D_{g_u} \mathcal{H}(\nabla_{g_u} u, \nabla_{g_u} u) + \frac{1}{2} (\dot{u}^2 - |\nabla_{g_u} u|_{g_u}^2) \operatorname{div} \mathcal{H} \right\} dx d\tau, \end{aligned}$$

where  $\nu$  is the unit normal of  $\partial\hat{\Omega}$  pointing towards the exterior of  $\hat{\Omega}$ , and

$$\frac{\partial u}{\partial \nu_a} = \sum_{i,j=1}^n a_{ij}(x, \|\nabla u\|^{2r}) \frac{\partial u}{\partial x_j} \nu_i$$

is the conormal derivative.

(2) Let  $h \in C^2(\hat{\Omega})$ . We have

$$\begin{aligned} & \int_s^t E_1^{\frac{\beta}{2}} \int_{\hat{\Omega}} h(\dot{u}^2 - |\nabla_{g_u} u|_{g_u}^2) dx d\tau \\ &= \int_s^t E_1^{\frac{\beta}{2}} \int_{\partial\hat{\Omega}} \left( \frac{1}{2} u^2 \frac{\partial h}{\partial \nu_a} - h u \frac{\partial u}{\partial \nu_a} \right) d\sigma d\tau - \frac{\beta}{2} \int_s^t E_1^{\frac{\beta-2}{2}} E_1'(\tau) \int_{\hat{\Omega}} \dot{u} h(u) dx d\tau \\ &\quad + E_1^{\frac{\beta}{2}}(\tau) (\dot{u}, h u)_{L^2(\hat{\Omega})} \Big|_s^t + \int_s^t E_1^{\frac{\beta}{2}} \int_{\hat{\Omega}} \left( f(x, \dot{u}) h u - \frac{1}{2} u^2 \operatorname{div} \nabla_{g_u} h \right) dx d\tau. \end{aligned}$$

*Proof of Theorem 1.3.* From (1.25), we know that

$$(5.9) \quad (\ddot{u}, \dot{u})_{L^2(\Omega)} + \int_{\Omega} \sum_{i,j=1}^n a_{ij}(x, \|\nabla u\|^{2r}) \dot{u}_{x_i} u_{x_j} + \int_{\Omega} f(x, \dot{u}) \dot{u} = 0,$$

which gives

$$E_1'(t) = -2 \int_{\Omega} f(x, \dot{u}) \dot{u} + 2 \int_{\Omega} \sum_{i,j=1}^n a_{ijs}(x, \|\nabla u\|^{2r}) \|\nabla u\|^{2r-2} (\nabla u, \nabla \dot{u}) u_{x_i} u_{x_j},$$

where  $a_{ijs}$  is the partial differential of  $a_{ij}$  with respect to  $s$ .

There exists a constant  $c_\gamma > 0$  such that

$$(5.10) \quad \left| \sum_{i,j=1}^n a_{ijs}(x, \|\nabla u\|^{2r}) \|\nabla u\|^{2r-2} (\nabla u, \nabla \dot{u}) u_{x_i} u_{x_j} \right| \leq \frac{1}{2} c_\gamma E_1^{r+1/2}(t) E_2^{1/2}(t).$$

Obviously we have

$$(5.11) \quad E_1'(t) - c_\gamma E_1^{r+1/2}(t) E_2^{1/2}(t) \leq -2 \int_{\Omega} f(x, \dot{u}) \dot{u} \leq 0.$$

Notice that

$$(5.12) \quad \int_s^t E_1^{\frac{\beta+2}{2}}(\tau) d\tau = \int_s^t E_1^{\frac{\beta}{2}} \int_{\Omega} (\dot{u}^2 + |\nabla_{g_u} u|_{g_u}^2) dx d\tau.$$

After a procedure similar to that in the proofs of Lemmas 3.3 and 3.4, we have

$$\begin{aligned} \int_s^t E_1^{\frac{\beta+2}{2}}(\tau) d\tau &\leq c_{\gamma} E_1^{\frac{\beta+2}{2}}(s) + c_{\gamma} E_1^{\frac{\beta+2}{2}}(t) + c_{\gamma} E_1(s) - c_{\gamma} E_1(t) \\ &\quad + c_{\gamma} \int_s^t E_1^{\frac{\beta}{2}} \int_G (\dot{u}^2 + f^2(x, \dot{u})) dx d\tau + c_{\gamma} \int_s^t E_1^{r+1/2}(\tau) E_2^{1/2}(\tau) d\tau. \end{aligned}$$

By (1.21) and (5.12) we have

$$\begin{aligned} \int_G (\dot{u}^2 + f^2(x, \dot{u})) dx &\leq c \int_G (\dot{u} f(x, \dot{u}))^{\frac{2}{\beta+2}} dx \leq c \left( \int_G \dot{u} f(x, \dot{u}) dx \right)^{\frac{2}{\beta+2}} \\ &\leq c(c_{\gamma} E_1^{r+1/2} E_2^{1/2} - E_1')^{\frac{2}{\beta+2}}. \end{aligned}$$

Hence, using the Young inequality, for any  $\varepsilon > 0$  we have

$$\begin{aligned} \int_s^t E_1^{\frac{\beta}{2}} \int_G (\dot{u}^2 + f^2(x, \dot{u})) dx d\tau &\leq c \int_s^t E_1^{\frac{\beta}{2}}(\tau) (c_{\gamma} E_1^{r+1/2} E_2^{1/2} - E_1'(\tau))^{\frac{\beta+2}{2}} d\tau \\ &\leq \int_s^t (\varepsilon E_1^{\frac{\beta+2}{2}} + c(\gamma, \varepsilon) E_1^{r+1/2} E_2^{1/2} - c(\varepsilon) E_1') d\tau \\ &\leq \varepsilon \int_s^t E_1^{\frac{\beta+2}{2}} d\tau + c(\gamma, \varepsilon) \int_s^t E_1^{r+1/2}(\tau) E_2^{1/2}(\tau) d\tau + C(\varepsilon) E_1(s) - C(\varepsilon) E_1(t). \end{aligned}$$

Similar to (3.44), we have

$$\begin{aligned} \max \left\{ E_1^{\frac{\beta+2}{2}}(t), E_1^{\frac{\beta+2}{2}}(s) \right\} &\leq \frac{1}{t-s} \int_s^t E_1^{\frac{\beta+2}{2}}(\tau) d\tau + c_{\gamma} \int_s^t E_1^{r+1/2+\beta/2}(\tau) E_2^{1/2}(\tau) d\tau \\ &\quad + c_{\gamma} \int_s^t E_1^{3/2+\beta/2}(\tau) d\tau + \int_s^t E_1^{\frac{\beta}{2}}(\tau) f(x, \dot{u}) \dot{u} dx d\tau, \end{aligned}$$

where the following estimate has been used:

$$-E_1'(t) \leq c E^{3/2}(t) + c_{\gamma} E_1^{r+1/2} E_2^{1/2}(t) + \int_{\Omega} f(x, \dot{u}) \dot{u} dx.$$

Then there exist constants  $c_{\gamma} > 0$  and  $T_{\gamma} > 3 \sup_{x \in \Omega} |H|_g / \alpha$  such that, if  $0 \leq s \leq t \leq T$ ,  $t - s \geq T_{\gamma}$ , then

$$\int_s^t E_1^{\frac{\beta+2}{2}}(\tau) d\tau + c_{\gamma} E_1(t) + c_{\gamma} E_1^{\frac{\beta+2}{2}}(t) \leq c_{\gamma} E_1(s) + c_{\gamma} E_1^{\frac{\beta+2}{2}}(s) + c_{\gamma} \int_s^t \mathcal{L}(\tau) d\tau,$$

where  $\mathcal{L}(t)$  is the high order of the energy defined as

$$\mathcal{L}(\tau) = E_1^{r+1/2+\beta/2}(\tau) E_2^{1/2}(\tau) + E_1^{3/2+\beta/2}(\tau) + E_1^{r+1/2}(\tau) E_2^{1/2}(\tau).$$

After a discussion similar to that in section 4, we obtain the existence of global solutions to (1.25). Furthermore, we have

$$(5.13) \quad E_1(t) + E_1^{\frac{\beta+2}{2}}(t) - E_1(s) - E_1^{\frac{\beta+2}{2}}(s) + w(\eta) \int_s^t E_1^{\frac{\beta+2}{2}}(\tau) d\tau \leq 0$$

for  $0 \leq s \leq t \leq T$ ,  $t - s \geq T_\gamma$ , where

$$w(\eta) > 0$$

for  $\eta > 0$  small. From (5.13) we know that

$$E_1(t) \leq E_1(s), \quad 0 \leq s \leq t \leq T, \quad t - s \geq T_\gamma,$$

which yields

$$(5.14) \quad E_1^{\frac{\beta+2}{2}}(s) - E_1^{\frac{\beta+2}{2}}(t) \leq c(E_1(s) - E_1(t))$$

for  $\eta \leq (\frac{2c}{2+\beta})^{2/\beta}$ .

Combining (5.13) and (5.14), we get

$$E_1(t) - E_1(s) + w(\eta) \int_s^t E_1^{\frac{\beta+2}{2}}(\tau) d\tau \leq 0$$

for  $0 \leq s \leq t \leq T$ ,  $t - s \geq T_\gamma$ , which yields

$$(5.15) \quad E_1(t) \leq E_1(0) \left( \frac{1}{1 + \frac{2}{\beta} w(\eta)(t - T_\gamma) E_1^{\beta/2}(0)} \right)^{2/\beta}, \quad t \geq T_\gamma.$$

We can choose the suitable constant  $c$  such that the estimate (1.23) holds for all  $t > 0$ .

Next, we prove that the second energy is bounded. From the estimate (5.15), the following corollary is immediate.

**COROLLARY 5.1.** *If  $w > \beta/2$ , then*

$$(5.16) \quad \int_0^t E_1^w(s) ds \leq \frac{2w}{2w - \beta} C E_1(0)^{w-\beta/2}, \quad t \geq 0.$$

The equation can be written as

$$(5.17) \quad \ddot{u} - \operatorname{div}(A \cdot \nabla u) + f(x, \dot{u}) = 0.$$

A computation yields

$$\begin{aligned} (5.18) \quad E_2'(t) &= (\dot{A} \cdot \nabla \dot{u}, \nabla \dot{u})_{L^2(\Omega)} + 2(A \cdot \nabla \dot{u}, \nabla \ddot{u})_{L^2(\Omega)} \\ &\quad + (2\operatorname{div}(A \cdot \nabla u), (\operatorname{div}(\dot{A} \cdot \nabla u) + \operatorname{div}(A \cdot \nabla \dot{u})))_{L^2(\Omega)} \\ &= (\dot{A} \cdot \nabla \dot{u}, \nabla \dot{u})_{L^2(\Omega)} - 2(\operatorname{div}(A \cdot \nabla \dot{u}), \ddot{u})_{L^2(\Omega)} \\ &\quad + (2\operatorname{div}(A \cdot \nabla u), \operatorname{div}(\dot{A} \cdot \nabla u)) + (2\operatorname{div}(A \cdot \nabla u), \operatorname{div}(A \cdot \nabla \dot{u}))_{L^2(\Omega)} \\ &= (\dot{A} \cdot \nabla \dot{u}, \nabla \dot{u})_{L^2(\Omega)} + 2(\operatorname{div}(A \cdot \nabla \dot{u}), f)_{L^2(\Omega)} + 2(\operatorname{div}(A \cdot \nabla u), \operatorname{div}(\dot{A} \cdot \nabla u))_{L^2(\Omega)} \\ &\equiv I_1 + I_2 + I_3, \end{aligned}$$

where  $I_2$  can be written as

$$I_2 = 2(\operatorname{div}(A \cdot \nabla \dot{u}), f)_{L^2(\Omega)} = -2 \int_{\Omega} \sum_{i,j=1}^n \{a_{ij} \dot{u}_{x_j} f_{x_i}(x, \dot{u}) + a_{ij} \dot{u}_{x_j} f_s(x, \dot{u}) \dot{u}_{x_i}\}.$$

We need to estimate the term  $E'_2(t)$ . First, it is easily seen that

$$(5.19) \quad I_1(t) \leq CE^{r-1/2}(t)E_2^{3/2}(t),$$

$$I_2(t) \leq CE^{r-1/2}(t)E_2^{1/2}(t),$$

$$I_3(t) \leq CE^{r-1/2}(t)E_2(t),$$

where the condition (1.20) is used.

Combining (5.18) and (5.19), we know that at least one of the following three inequalities is true:

$$(5.20) \quad E'_2(t) \leq cE^{r-1/2}(t)E_2^{3/2}(t),$$

$$(5.21) \quad E'_2(t) \leq CE^{r-1/2}(t)E_2^{1/2}(t),$$

$$(5.22) \quad E'_2(t) \leq CE^{r-1/2}(t)E_2(t).$$

Letting  $\omega = r - 1/2$  in (5.16) yields

$$(5.23) \quad \int_0^t E^{r-1/2}(s)ds \leq \frac{2r-1}{2r-1-\beta} CE(0)^{r-1/2-\beta/2}.$$

The inequalities (5.20)–(5.23) imply that the second energy  $E_2(t)$  is bounded in  $t \in [0, \infty)$ .

**Acknowledgments.** The authors thank the reviewers for their suggestion to include the case (1.21).

#### REFERENCES

- [1] J. P. AUBIN, *Un théorème de compacité*, C. R. Acad. Sci. Paris, 256 (1963), pp. 5042–5044.
- [2] V. BARBU, I. LASIECKA, AND M. A. RAMMAHA, *Blow-up of generalized solutions to wave equations with nonlinear degenerate damping and source terms*, Indiana Univ. Math. J., 56 (2007), pp. 995–1021.
- [3] M. M. CAVALCANTI, A. KHEMMOUDJ, AND M. MEDJDEN, *Uniform stabilization of the damped Cauchy–Ventcel problem with variable coefficients and dynamic boundary conditions*, J. Math. Anal. Appl., 328 (2007), pp. 900–930.
- [4] S. CHAI, Y. GUO, AND P.-F. YAO, *Boundary feedback stabilization of shallow shells*, SIAM J. Control Optim., 42 (2003), pp. 239–259.
- [5] S. CHAI AND P. F. YAO, *Observability inequalities for thin shells*, Sci. China Ser. A, 46 (2003), pp. 300–311.
- [6] S. G. CHAI AND K. LIU, *Boundary stabilization of the transmission of wave equations with variable coefficients*, Chinese Ann. Math. Ser. A, 26 (2005), pp. 605–612 (in Chinese).
- [7] C. M. DAFERMOS AND W. J. HRUSA, *Energy methods for quasilinear hyperbolic initial-boundary value problems. Applications to elastodynamics*, Arch. Rational Mech. Anal., 87 (1985), pp. 267–292.
- [8] S. J. FENG AND D. X. FENG, *Nonlinear internal damping of wave equations with variable coefficients*, Acta Math. Sin. (Engl. Ser.), 20 (2004), pp. 1057–1072.



- [9] R. GULLIVER, I. LASIECKA, W. LITTMAN, AND R. TRIGGIANI, *The case for differential geometry in the control of single and coupled PDEs: The structural acoustic chamber*, in Geometric Methods in Inverse Problems and PDE Control, IMA Vol. Math. Appl. 137, Springer, New York, 2004, pp. 73–181.
- [10] A. HARAUX, *Stabilization of trajectories for some weakly damped hyperbolic equations*, J. Differential Equations, 59 (1985), pp. 145–154.
- [11] S. KLAINERMAN AND A. MAJDA, *Formation of singularities for wave equations including the nonlinear vibrating string*, Comm. Pure Appl. Math., 33 (1980), pp. 241–263.
- [12] V. KOMORNIK, *Exact Controllability and Stabilization: The Multiplier Method*, John Wiley, Chichester, UK, 1994.
- [13] J. LAGNESE, *Control of wave processes with distributed controls supported on a subregion*, SIAM J. Control Optim., 21 (1983), pp. 68–85.
- [14] I. LASIECKA AND J. ONG, *Global solvability and uniform decays of solutions to quasilinear equation with nonlinear boundary dissipation*, Comm. Partial Differential Equations, 24 (1999), pp. 2069–2107.
- [15] I. LASIECKA AND R. TRIGGIANI, *Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometrical conditions*, Appl. Math. Optim., 25 (1992), pp. 189–224.
- [16] I. LASIECKA AND R. TRIGGIANI, *Uniform stabilization of a shallow shell model with nonlinear boundary feedbacks*, J. Math. Anal. Appl., 269 (2002), pp. 642–688.
- [17] I. LASIECKA, R. TRIGGIANI, AND P. F. YAO, *Inverse/observability estimates for second-order hyperbolic equations with variable coefficients*, J. Math. Anal. Appl., 235 (1999), pp. 13–57.
- [18] I. LASIECKA, R. TRIGGIANI, AND P. F. YAO, *Carleman estimates for a plate equation on a Riemann manifold with energy level terms*, in Analysis and Applications—ISAAC 2001 (Berlin), Int. Soc. Anal. Appl. Comput. 10, Kluwer, Dordrecht, The Netherlands, 2003, pp. 199–236.
- [19] K. LIU, *Locally distributed control and damping for the conservative system*, SIAM J. Control Optim., 35 (1997), pp. 1574–1590.
- [20] M. NAKAO, *Decay of solutions of the wave equation with a local nonlinear dissipation*, Math. Ann., 305 (1996), pp. 403–417.
- [21] M. NAKAO, *Energy decay for the linear and semilinear wave equations in exterior domains with some localized dissipations*, Math. Z., 238 (2001), pp. 781–797.
- [22] M. NAKAO, *Existence of global solutions for the Kirchhoff-type quasilinear wave equation in exterior domains with a half-linear dissipation*, Kyushu J. Math., 58 (2004), pp. 373–391.
- [23] S. NICAISE AND C. PIGNOTTI, *Internal and boundary observability estimates for the heterogeneous Maxwell's system*, Appl. Math. Optim., 54 (2006), pp. 47–70.
- [24] K. ONO, *On global solutions and blow-up solutions of nonlinear Kirchhoff strings with nonlinear dissipation*, J. Math. Anal. Appl., 216 (1997), pp. 321–342.
- [25] M. A. RAMMAHA AND T. A. STREI, *Global existence and nonexistence for nonlinear wave equations with damping and source terms*, Trans. Amer. Math. Soc., 354 (2002), pp. 3621–3637.
- [26] M. SLEMROD, *Weak asymptotic decay via a “related invariance principle” for a wave equation with nonlinear, nonmonotone damping*, Proc. Roy. Soc. Edinburgh Sect. A, 113 (1989), pp. 87–97.
- [27] M. E. TAYLOR, *Partial Differential Equations I: Basic Theory*, Springer-Verlag, New York, 1996.
- [28] L. R. TCHEUGOUÉ TÉBOU, *Stabilization of the wave equation with localized nonlinear damping*, J. Differential Equations, 145 (1998), pp. 502–524.
- [29] G. TODOROVA, *Cauchy problem for a nonlinear wave equation with nonlinear damping and source terms*, Nonlinear Anal. Ser. A Theory Methods, 41 (2000), pp. 891–905.
- [30] G. TODOROVA AND B. YORDANOV, *The energy decay problem for wave equations with nonlinear dissipative terms in  $R^n$* , Indiana Univ. Math. J., 56 (2007), pp. 389–416.
- [31] G. TODOROVA AND B. YORDANOV, *Critical exponent for a nonlinear wave equation with damping*, J. Differential Equations, 174 (2001), pp. 464–489.
- [32] R. TRIGGIANI AND P. F. YAO, *Carleman estimates with no lower-order terms for general Riemann wave equations. Global uniqueness and observability in one shot*, Appl. Math. Optim., 46 (2002), pp. 331–375.
- [33] H. WU, C. L. SHEN, AND Y. L. YU, *An Introduction to Riemannian Geometry*, Beijing University Press, Beijing, 1989 (in Chinese).
- [34] P.-F. YAO, *On the observability inequalities for exact controllability of wave equations with variable coefficients*, SIAM J. Control Optim., 37 (1999), pp. 1568–1599.
- [35] P.-F. YAO, *Observability inequalities for shallow shells*, SIAM J. Control Optim., 38 (2000), pp. 1729–1756.

- [36] P. F. YAO, *Observability inequalities for the Euler-Bernoulli plate with variable coefficients*, in Differential Geometric Methods in the Control of Partial Differential Equations, Contemp. Math. 268, Amer. Math. Soc., Providence, RI, 2000, pp. 383–406.
- [37] P. F. YAO, *Global smooth solutions for the quasilinear wave equation with boundary dissipation*, J. Differential Equations, 241 (2007), pp. 62–93.
- [38] P. F. YAO, *Boundary Controllability for the Quasilinear Wave Equation*, <http://arxiv.org/abs/math/0603280> (2006).
- [39] E. ZUAZUA, *Exponential decay for the semilinear wave equation with locally distributed damping*, Comm. Partial Differential Equations, 15 (1990), pp. 205–235.
- [40] E. ZUAZUA, *Exponential decay for the semilinear wave equation with localized damping in unbounded domains*, J. Math. Pures Appl. (9), 70 (1991), pp. 513–529.

## NECESSARY AND SUFFICIENT OPTIMALITY CONDITIONS FOR RELAXED AND STRICT CONTROL PROBLEMS\*

SEID BAHMALI†

**Abstract.** We consider a stochastic control problem where the set of strict (classical) controls is not necessarily convex, and the system is governed by a nonlinear stochastic differential equation, in which the control enters both the drift and the diffusion coefficients. By introducing a new approach, we establish necessary as well as sufficient conditions of optimality for two models. The first concerns the relaxed controls, which are measure-valued processes in which an optimal solution exists. The second is a particular case of the first and relates to strict control problems. These results are given in the form of global stochastic maximum principle by using only the first-order expansion and the associated adjoint equation. This improves all of the previous works on the subject.

**Key words.** stochastic differential equation, strict control, relaxed control, maximum principle, adjoint process, variational inequality

**AMS subject classification.** 93Exx

**DOI.** 10.1137/070681053

**1. Introduction.** We study a stochastic control problem where the system is governed by a nonlinear stochastic differential equation (SDE) of the type

$$\begin{cases} dx_t^v = b(t, x_t^v, v_t) dt + \sigma(t, x_t^v, v_t) dW_t, \\ x_0^v = \xi, \end{cases}$$

where  $b$  and  $\sigma$  are given deterministic functions,  $\xi$  is the initial data, and  $W = (W_t)_{t \geq 0}$  is a standard  $d$ -dimensional Brownian motion, defined on a filtered probability space  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathcal{P})$  satisfying the usual conditions.

The control variable  $v = (v_t)$ , called strict (classical) control, is an  $\mathcal{F}_t$  adapted process with values in some set  $U$  of  $\mathbb{R}^k$ . We denote by  $\mathcal{U}$  the class of all strict controls.

The criteria to be minimized, over the set  $\mathcal{U}$ , has the form

$$J(v) = \mathbb{E} \left[ g(x_T^v) + \int_0^T h(t, x_t^v, v_t) dt \right],$$

where  $g$  and  $h$  are given maps and  $x_t^v$  is the trajectory of the system controlled by  $v$ .

A control  $u \in \mathcal{U}$  is called optimal if it satisfies

$$J(u) = \inf_{v \in \mathcal{U}} J(v).$$

This kind of stochastic control problems have been studied extensively, both by the dynamic programming approach and by the Pontryagin stochastic maximum principle. In this paper, we are concerned with the second approach, whose objective is to establish necessary as well as sufficient conditions for optimality of controls. There are

---

\*Received by the editors January 26, 2007; accepted for publication (in revised form) March 18, 2008; published electronically July 16, 2008. This work is partially supported by Algerian-French cooperation, Tassili 07 MDU 705.

<http://www.siam.org/journals/sicon/47-4/68105.html>

†Laboratory of Applied Mathematics, University Med Khider, P.O. Box 145, Biskra 07000, Algeria (sbahlali@yahoo.fr).

many works concerning this subject. The first contribution in this direction is made by Kushner [18]. The other fundamental advance was developed by Haussmann [11, 12]. Versions of the stochastic maximum principle in which the diffusion coefficient is allowed to depend explicitly on the control variable were derived by Arkin and Saksonov [1], Bensoussan [5], Elliot [8], Elliot and Kohlmann [9], and Peng [23]. Necessary (as well as sufficient) optimality conditions for linear systems with random coefficients, where no  $L^p$ -bounds are imposed on the controls, are established by Cadellinas and Karatzas [6].

The common fact in most of these works is that an optimal solution in the class of strict controls may fail to exist. Existence of such a strict optimal control follows from the Filippov convexity condition, which is the convexity of the image of the action space  $U$  by the map  $(b(t, x, \cdot), \sigma\sigma^*(t, x, \cdot), h(t, x, \cdot))$ ; see [4], [7], [13], [14], [19]. Without this convexity condition, an optimal strict control does not necessarily exist in  $\mathcal{U}$ . To overcome this problem of existence without imposing the Filippov condition, the idea is then to introduce a bigger new class  $\mathcal{R}$  of processes in which the controller chooses at time  $t$  a probability measure  $q_t(da)$  on the control set  $U$ , rather than an element  $v_t$  of  $U$ . This new class of processes is called relaxed controls and has a richer topological structure, for which the control problem becomes solvable.

In the relaxed model, the system is governed by the SDE

$$\begin{cases} dx_t^q = \int_U b(t, x_t^q, a) q_t(da) dt + \int_U \sigma(t, x_t^q, a) q_t(da) dW_t, \\ x_0^q = \xi. \end{cases}$$

The functional cost to be minimized, over the class  $\mathcal{R}$  of relaxed controls, is defined by

$$J(q) = \mathbb{E} \left[ g(x_T^q) + \int_0^T \int_U h(t, x_t^q, a) q_t(da) dt \right].$$

A relaxed control  $\mu$  is called optimal if it solves

$$J(\mu) = \inf_{q \in \mathcal{R}} J(q).$$

The relaxed control problem finds its interest in two essential points. The first is that an optimal solution exists. Fleming [10] derived an existence result of an optimal relaxed control with uncontrolled diffusion coefficient. The existence of an optimal solution, where the drift and the diffusion coefficients depend explicitly on the relaxed control variable, has been solved by El Karoui, Nguyen, and Piqué [7]. The relaxed optimal control in this general case is shown to be Markovian. See Bahlali, Mezerdi, and Djehiche [2] for an alternative proof for existence of an optimal relaxed control. The second is that it is a generalization of the strict control problem. Indeed, if  $q_t(da) = \delta_{v_t}(da)$  is a Dirac measure concentrated at a single point  $v_t$ , then we get a strict control problem as a particular case of the relaxed one.

Motivated by the existence of an optimal solution, the unique versions of stochastic maximum principle for relaxed controls were established by Mezerdi and Bahlali [22] in the case of uncontrolled diffusion, Bahlali, Mezerdi, and Djehiche [2], where the drift and the diffusion coefficients depend explicitly on the relaxed control variable, and Bahlali, Djehiche, and Mezerdi [3] for the problem of mixed singular-relaxed controls. All these results are obtained by using the previous works on strict controls, Ekeland's variational principle, and some stability properties of the trajectories and adjoint processes with respect to the control variable.

The general stochastic maximum principle for strict controls established by Peng [23] and its extension to the class of measure-valued processes developed by Bahlali, Mezerdi, and Djehiche [2], have been both obtained by using the second-order expansion. Then, these two results are given with two adjoint processes and a variational inequality of the second-order.

Our aim in this paper is to establish necessary as well as sufficient conditions of optimality in the form of global stochastic maximum principle, for relaxed and strict controls, without using the second-order expansion. To achieve this goal, we introduce a new approach and derive these two main results as follows.

First, we give the optimality conditions for relaxed controls. The main idea is to use the fact that the set of relaxed controls is convex. Then, we establish necessary optimality conditions by using the classical way of the convex perturbation method. More precisely, if we denote by  $\mu$  an optimal relaxed control and  $q$  an arbitrary element of  $\mathcal{R}$ , then with a sufficiently small  $\theta > 0$  and for each  $t \in [0, T]$ , we can define a perturbed control as follows:

$$\mu_t^\theta = \mu_t + \theta(q_t - \mu_t).$$

We derive the variational equation from the state equation, and the variational inequality from the inequality

$$0 \leq J(\mu^\theta) - J(\mu).$$

By using the fact that the drift, the diffusion, and the running cost coefficients are linear with respect to the relaxed control variable, necessary optimality conditions are obtained directly in the global form. This result improves significantly that of Bahlali, Mezerdi, and Djehiche [2], in the sense where we use only the first-order expansion with only one adjoint process.

To achieve the first result of this paper, we prove under minimal additional hypothesis, that these necessary optimality conditions for relaxed controls are also sufficient.

The second main result in this paper characterizes the optimality for strict control processes. It is directly derived from the above result by restricting from relaxed to strict controls. The main idea is to replace the relaxed controls by a Dirac measures charging a strict controls. Thus, we reduce the set  $\mathcal{R}$  of relaxed controls and minimize the cost  $J$  over the subset  $\delta(\mathcal{U}) = \{q \in \mathcal{R} / q = \delta_v; \ v \in \mathcal{U}\}$ . Then, we derive necessary optimality conditions by using only the first-order expansion and the associated adjoint equation. Therefore we no longer need the second-order expansion. This result improves considerably the Peng stochastic maximum principle [23]. Moreover, we prove that these necessary conditions becomes sufficient, without imposing either the convexity of  $U$  or that of the Hamiltonian  $H$  in  $v$ .

This paper is organized as follows. In section 2, we formulate the strict and relaxed control problems and give the various assumptions used throughout this paper. Section 3 is devoted to study the relaxed control problems and we establish necessary as well as sufficient conditions of optimality for relaxed controls. In the last section, we derive directly from the results of section 3, the optimality conditions for strict controls.

Throughout this paper, we denote by  $C$  some positive constant and we need the following matrix notations. We denote by  $\mathcal{M}_{n \times d}(\mathbb{R})$  the space of  $n \times d$  real matrices and by  $\mathcal{M}_{n \times n}^d(\mathbb{R})$  the linear space of vectors  $M = (M_1, \dots, M_d)$ , where

$M_i \in \mathcal{M}_{n \times n}(\mathbb{R})$ .

For any  $M, N \in \mathcal{M}_{n \times n}^d(\mathbb{R})$ ,  $L, S \in \mathcal{M}_{n \times d}(\mathbb{R})$ ,  $\alpha, \beta \in \mathbb{R}^n$ , and  $\gamma \in \mathbb{R}^d$ , we use the following notations:

$$\alpha\beta = \sum_{i=1}^n \alpha_i \beta_i \in \mathbb{R} \text{ is the product scalar in } \mathbb{R}^n;$$

$$LS = \sum_{i=1}^d L_i S_i \in \mathbb{R}, \text{ where } L_i \text{ and } S_i \text{ are the } i\text{th columns of } L \text{ and } S;$$

$$ML = \sum_{i=1}^d M_i L_i \in \mathbb{R}^n;$$

$$M\alpha\gamma = \sum_{i=1}^d (M_i \alpha) \gamma_i \in \mathbb{R}^n;$$

$$MN = \sum_{i=1}^d M_i N_i \in \mathcal{M}_{n \times n}(\mathbb{R});$$

$$MLN = \sum_{i=1}^d M_i L N_i \in \mathcal{M}_{n \times n}(\mathbb{R});$$

$$ML\gamma = \sum_{i=1}^d M_i L \gamma_i \in \mathcal{M}_{n \times n}(\mathbb{R}).$$

We denote by  $L^*$  the transpose of the matrix  $L$  and  $M^* = (M_1^*, \dots, M_d^*)$ .

**2. Formulation of the problem.** Let  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathcal{P})$  be a filtered probability space satisfying the usual conditions, on which a  $d$ -dimensional Brownian motion  $W = (W_t)_{t \geq 0}$  is defined. We assume that  $(\mathcal{F}_t)$  is the  $\mathcal{P}$ -augmentation of the natural filtration of  $W$ .

Let  $T$  be a strictly positive real number and  $U$  a nonempty compact set of  $\mathbb{R}^k$ .

### 2.1. The strict control problem.

**DEFINITION 1.** An admissible strict control is an  $\mathcal{F}_t$ -adapted process  $v = (v_t)$  with values in  $U$  such that

$$\mathbb{E} \left[ \sup_{0 \leq t \leq T} |v_t|^2 \right] < \infty.$$

We denote by  $\mathcal{U}$  the set of all admissible strict controls.

For any  $v \in \mathcal{U}$ , we consider the following controlled SDE:

$$(1) \quad \begin{cases} dx_t^v = b(t, x_t^v, v_t) dt + \sigma(t, x_t^v, v_t) dW_t, \\ x_0^v = \xi, \end{cases}$$

where

$$\begin{aligned} b &: [0, T] \times \mathbb{R}^n \times U \longrightarrow \mathbb{R}^n, \\ \sigma &: [0, T] \times \mathbb{R}^n \times U \longrightarrow \mathcal{M}_{n \times d}(\mathbb{R}), \end{aligned}$$

and  $\xi$  is an  $n$ -dimensional  $\mathcal{F}_0$ -measurable random variable such that

$$\mathbb{E} |\xi|^2 < \infty.$$

The expected cost to be minimized is defined from  $\mathcal{U}$  into  $\mathbb{R}$  by

$$(2) \quad J(v) = \mathbb{E} \left[ g(x_T^v) + \int_0^T h(t, x_t^v, v_t) dt \right],$$

where

$$\begin{aligned} g : \mathbb{R}^n &\longrightarrow \mathbb{R}, \\ h : [0, T] \times \mathbb{R}^n \times U &\longrightarrow \mathbb{R}. \end{aligned}$$

A strict control  $u$  is called optimal if it satisfies

$$(3) \quad J(u) = \inf_{v \in \mathcal{U}} J(v).$$

The following assumptions will be in force throughout this paper:

$$(4) \quad b, \sigma, g, h \text{ are continuously differentiable with respect to } x.$$

They and all their derivatives  $b_x, \sigma_x, g_x, h_x$  are continuous in  $(x, v)$ .

$b_x, \sigma_x, g_x, h_x$  are uniformly bounded.  $b, \sigma, g, h$  are bounded by  $C(1 + |x| + |v|)$ .

Under the above assumptions, for every  $v \in \mathcal{U}$ , (1) has a unique strong solution, and the functional cost  $J$  is well defined from  $\mathcal{U}$  into  $\mathbb{R}$ .

**2.2. The relaxed model.** The strict control problem  $\{(1), (2), \text{ and } (3)\}$  formulated in the last subsection may fail to have an optimal solution without the Filippov convexity condition, which is the convexity of the image of the action space  $U$  by the map  $(b(t, x, \cdot), \sigma\sigma^*(t, x, \cdot), h(t, x, \cdot))$ ; see [4], [7], [13], [14], [19]. Let us begin by a deterministic example which shows that even in simple cases, existence of a strict optimal control is not ensured (see Fleming [10] and Yong and Zhou [24] for other examples).

The problem is to minimize, over the set of measurable functions  $v : [0, T] \rightarrow \{-1, 1\}$ , the following functional cost:

$$J(v) = \int_0^T (x_t^v)^2 dt,$$

where  $x_t^v$  denotes the solution of

$$\begin{cases} dx_t^v = v_t dt, \\ x_0^v = 0. \end{cases}$$

We then have

$$\inf_{v \in \mathcal{U}} J(v) = 0.$$

Indeed, consider the following sequence of controls:

$$v_t^n = (-1)^k \quad \text{if} \quad \frac{k}{n}T \leq t \leq \frac{k+1}{n}T, \quad 0 \leq k \leq n-1.$$

Then, clearly

$$\begin{aligned} \left| x_t^{v^n} \right| &\leq \frac{T}{n}, \\ |J(v^n)| &\leq \frac{T^3}{n^2}, \end{aligned}$$

which implies that

$$\inf_{v \in \mathcal{U}} J(v) = 0.$$

There is, however, no control  $v$  such that  $J(v) = 0$ . If this were the case, then for every  $t$ ,  $x_t^v = 0$ . This in turn would imply that  $v_t = 0$ , which is impossible. The problem is that the sequence  $(v^n)$  has no limit in the space of strict controls. This limit, if it exists, will be the natural candidate for optimality. If we identify  $v_t^n$  with the Dirac measure  $\delta_{v_t^n}(da)$  and set  $q_n(dt, dv) = \delta_{v_t^n}(dv) dt$ , then we get a measure on  $[0, 1] \times U$ . Then, the sequence  $(q_n(dt, dv))_n$  converges weakly to  $\frac{1}{2} dt [\delta_{-1} + \delta_1](da)$ .

This suggests that the set  $\mathcal{U}$  of strict controls is too narrow and should be embedded into a wider class with a richer topological structure, for which the control problem becomes solvable.

The idea of relaxed controls is to replace the  $U$ -valued process  $(v_t)$  with  $\mathbb{P}(U)$ -valued process  $(q_t)$ , where  $\mathbb{P}(U)$  is the space of probability measures equipped with the topology of weak convergence.

**DEFINITION 2.** A relaxed control is the term  $q = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, P, W_t, q_t, x_t, \xi)$  such that (1)  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, P)$  is a filtered probability space satisfying the usual conditions. (2)  $(q_t)_t$  is a  $\mathbb{P}(\bar{U})$ -valued process, progressively measurable with respect to  $(\mathcal{F}_t)_t$  and such that for each  $t$ ,  $\mathbf{1}_{[0, t]} \cdot q$  is  $\mathcal{F}_t$ -measurable. (3)  $(x_t)_t$  is  $\mathbb{R}^n$ -valued and  $\mathcal{F}_t$ -adapted with continuous paths such that  $x_0 = \xi$  and for each  $f \in C_b^2(\mathbb{R}^n, \mathbb{R})$

$$f(x_t) - f(\xi) - \int_0^t \int_U Lf(s, x_s, a) q_s(\omega, da) ds$$

is a  $P$ -martingale, where  $L$  is the infinitesimal generator associated with (5), acting on a map  $f$  in  $C_b^2(\mathbb{R}^n, \mathbb{R})$ .

By a slight abuse of notation, we will often denote a relaxed control by  $q$  instead of specifying all of the components.

**Remark 3.** The set of strict controls is embedded into the set of relaxed controls by the mapping

$$f : v \mapsto f_v(dt, da) = dt \delta_{v_t}(da),$$

where  $\delta_v$  is the atomic measure concentrated at a single point  $v$ .

For more details on relaxed controls, see [2], [3], [7], [10], [20], [21], [22].

**DEFINITION 4.** An admissible relaxed control is a relaxed control  $q = (q_t)$  such that

$$\mathbb{E} \left[ \sup_{0 \leq t \leq T} |q_t|^2 \right] < \infty.$$

We denote by  $\mathcal{R}$  the set of all admissible relaxed controls.



For any  $q \in \mathcal{R}$ , we consider the following relaxed SDE:

$$(5) \quad \begin{cases} dx_t^q = \int_U b(t, x_t^q, a) q_t(da) dt + \int_U \sigma(t, x_t^q, a) q_t(da) dW_t, \\ x_0^q = \xi. \end{cases}$$

The expected cost to be minimized, in the relaxed model, is defined from  $\mathcal{R}$  into  $\mathbb{R}$  by

$$(6) \quad J(q) = \mathbb{E} \left[ g(x_T^q) + \int_0^T \int_U h(t, x_t^q, a) q_t(da) dt \right].$$

A relaxed control  $\mu$  is called optimal if it solves

$$(7) \quad J(\mu) = \inf_{q \in \mathcal{R}} J(q).$$

Existence of an optimal solution for the problem  $\{(5), (6), (7)\}$  has been solved by El Karoui, Nguyen, and Piqué [7] by using a compactification method. The relaxed optimal control in this general case is shown to be Markovian. See Bahlali, Mezerdi, and Djehiche [2] for an alternative proof for existence of an optimal relaxed control.

*Remark 5.* If we put

$$\begin{aligned} \bar{b}(t, x_t^q, q_t) &= \int_U b(t, x_t^q, a) q_t(da), \\ \bar{\sigma}(t, x_t^q, q_t) &= \int_U \sigma(t, x_t^q, a) q_t(da), \\ \bar{h}(t, x_t^q, q_t) &= \int_U h(t, x_t^q, a) q_t(da), \end{aligned}$$

then (5) becomes

$$(5') \quad \begin{cases} dx_t^q = \bar{b}(t, x_t^q, q_t) dt + \bar{\sigma}(t, x_t^q, q_t) dW_t, \\ x_T^q = \xi, \end{cases}$$

with a functional cost given by

$$J(q) = \mathbb{E} \left[ g(x_T^q) + \int_0^T \bar{h}(t, x_t^q, q_t) dt \right].$$

Hence, by introducing relaxed controls, we have replaced  $U$  by a larger space  $\mathbb{P}(U)$ . We have gained the advantage that  $\mathbb{P}(U)$  is both compact and convex. Furthermore, the new coefficients of (5) and the running cost are linear with respect to the relaxed control variable.

*Remark 6.* The coefficients  $\bar{b}$  and  $\bar{\sigma}$  (defined in the above remark) check, respectively, the same assumptions as  $b$  and  $\sigma$ . Then, under assumptions (4),  $\bar{b}$  and  $\bar{\sigma}$  are uniformly Lipschitz and with linear growth. Then, by classical results on SDEs (the Itô theorem; see Ikeda and Watanabe [15], Karatzas and Shreve [17]), for every  $q \in \mathcal{R}$ , (5') admits a unique strong solution. Consequently, for every  $q \in \mathcal{R}$ , (5) has a unique strong solution. On the other hand, it is easy to see that  $\bar{h}$  checks the same assumptions as  $h$ . Then, the functional cost  $J$  is well defined from  $\mathcal{R}$  into  $\mathbb{R}$ .

*Remark 7.* If  $q_t = \delta_{v_t}$  is an atomic measure concentrated at a single point  $v_t \in U$ , then for each  $t \in [0, T]$  we have

$$\begin{aligned}\int_U b(t, x_t^q, a) q_t(da) &= \int_U b(t, x_t^q, a) \delta_{v_t}(da) = b(t, x_t^q, v_t), \\ \int_U \sigma(t, x_t^q, a) q_t(da) &= \int_U \sigma(t, x_t^q, a) \delta_{v_t}(da) = \sigma(t, x_t^q, v_t), \\ \int_U h(t, x_t^q, a) q_t(da) &= \int_U h(t, x_t^q, a) \delta_{v_t}(da) = h(t, x_t^q, v_t).\end{aligned}$$

In this case  $x^q = x^v$ ,  $J(q) = J(v)$ , and we get a strict control problem. Thus the problem of strict controls  $\{(1), (2), (3)\}$  is a particular case of relaxed control problem  $\{(5), (6), (7)\}$ .

*Remark 8.* We note that the relaxed equation (5) can be expressed in terms of martingale measure (see El Karoui, Nguyen, and Piqué [7] and Bahlali, Mezerdi, and Djehiche [2]). If we follow this formulation, the state equation is governed by a martingale measure and is given by

$$\begin{cases} dx_t^q = \int_U b(t, x_t^q, a) q_t(da) dt + \int_U \sigma(t, x_t^q, a) M(da, dt), \\ x_0^q = \xi. \end{cases}$$

where  $M(da, dt)$  is a martingale measure with intensity the relaxed control  $q_t(da) dt$ .

In our formulation of the relaxed stochastic control problem, the state equation (5) is governed by the Brownian motion  $W$ . This formulation was used by Ma and Yong [20] for a relaxed control problem of forward-backward systems. See Ma and Yong [20] for more details.

### 3. Necessary and sufficient optimality conditions for relaxed controls.

In this section, we study the problem  $\{(5), (6), (7)\}$  and we establish necessary as well as sufficient conditions of optimality for relaxed controls.

**3.1. Preliminary results.** Since the set of relaxed controls  $\mathcal{R}$  is convex, a classical way of treating such a problem is to use the convex perturbation method. More precisely, let  $\mu$  be an optimal relaxed control and  $x_t^\mu$  the solution of (5) controlled by  $\mu$ . Then, we can define a perturbed relaxed control as follows:

$$(8) \quad \mu_t^\theta = \mu_t + \theta(q_t - \mu_t),$$

where  $\theta > 0$  is sufficiently small and  $q$  is an arbitrary element of  $\mathcal{R}$ .

Denote by  $x_t^\theta$  the solution of (5) associated with  $\mu^\theta$ .

From the optimality of  $\mu$ , the variational inequality will be derived from the fact that

$$(9) \quad 0 \leq J(\mu^\theta) - J(\mu).$$

To this end, we need the following classical lemmas.

LEMMA 9. *Under assumptions (4), we have*

$$(10) \quad \lim_{\theta \rightarrow 0} \left[ \sup_{0 \leq t \leq T} \mathbb{E} |x_t^\theta - x_t^\mu|^2 \right] = 0.$$

*Proof.* We have

$$\begin{aligned}
 x_t^\theta - x_t^\mu &= \int_0^t \left[ \int_U b(s, x_s^\theta, a) \mu_s^\theta(da) - \int_U b(s, x_s^\mu, a) \mu_s(da) \right] ds \\
 &\quad + \int_0^t \left[ \int_U \sigma(s, x_s^\theta, a) \mu_s^\theta(da) - \int_U \sigma(s, x_s^\mu, a) \mu_s(da) \right] dW_s \\
 &= \int_0^t \left[ \int_U b(s, x_s^\theta, a) \mu_s^\theta(da) - \int_U b(s, x_s^\mu, a) \mu_s^\theta(da) \right] ds \\
 &\quad + \int_0^t \left[ \int_U b(s, x_s^\mu, a) \mu_s^\theta(da) - \int_U b(s, x_s^\mu, a) \mu_s(da) \right] ds \\
 &\quad + \int_0^t \left[ \int_U \sigma(s, x_s^\theta, a) \mu_s^\theta(da) - \int_U \sigma(s, x_s^\mu, a) \mu_s^\theta(da) \right] dW_s \\
 &\quad + \int_0^t \left[ \int_U \sigma(s, x_s^\mu, a) \mu_s^\theta(da) - \int_U \sigma(s, x_s^\mu, a) \mu_s(da) \right] dW_s.
 \end{aligned}$$

By using the definition of  $\mu_t^\theta$  and taking expectation, we have

$$\begin{aligned}
 \mathbb{E} |x_t^\theta - x_t^\mu|^2 &\leq C \mathbb{E} \int_0^t \left| \int_U b(s, x_s^\theta, a) \mu_s(da) - \int_U b(s, x_s^\mu, a) \mu_s(da) \right|^2 ds \\
 &\quad + C \theta^2 \mathbb{E} \int_0^t \left| \int_U b(s, x_s^\theta, a) q_s(da) - \int_U b(s, x_s^\theta, a) \mu_s(da) \right|^2 ds \\
 &\quad + C \mathbb{E} \int_0^t \left| \int_U \sigma(s, x_s^\theta, a) \mu_s(da) - \int_U \sigma(s, x_s^\mu, a) \mu_s(da) \right|^2 ds \\
 &\quad + C \theta^2 \mathbb{E} \int_0^t \left| \int_U \sigma(s, x_s^\theta, a) q_s(da) - \int_U \sigma(s, x_s^\theta, a) \mu_s(da) \right|^2 ds.
 \end{aligned}$$

By (4),  $b$  and  $\sigma$  are uniformly Lipschitz with respect to  $x$ . Hence

$$\mathbb{E} |x_t^\theta - x_t^\mu|^2 \leq C \mathbb{E} \int_0^t |x_s^\theta - x_s^\mu|^2 ds + C \theta^2.$$

By using Gronwall's lemma and the Buckholder–Davis–Gundy inequality, we obtain the desired result.  $\square$

LEMMA 10. Let  $z_t$  be the solution of the following linear equation (called variational equation):

$$(11) \quad \begin{cases} dz_t = \int_U b_x(t, x_t^\mu, a) \mu_t(da) z_t dt + \int_U \sigma_x(t, x_t^\mu, a) \mu_t(da) z_t dW_t \\ \quad + \left[ \int_U b(t, x_t^\mu, a) \mu_t(da) - \int_U b(t, x_t^\mu, a) q_t(da) \right] dt \\ \quad + \left[ \int_U \sigma(t, x_t^\mu, a) \mu_t(da) - \int_U \sigma(t, x_t^\mu, a) q_t(da) \right] dW_t, \\ z_0 = 0. \end{cases}$$

Then, we have

$$(12) \quad \lim_{\theta \rightarrow 0} \mathbb{E} \left| \frac{x_t^\theta - x_t^\mu}{\theta} - z_t \right|^2 = 0.$$

*Proof.* For simplicity, we put

$$(13) \quad X_t = \frac{x_t^\theta - x_t^\mu}{\theta} - z_t.$$

Then, we have

$$\begin{aligned}
 X_t = & \frac{1}{\theta} \int_0^t \int_U b(s, x_s^\theta, a) \mu_s^\theta(da) - \int_U b(s, x_s^\mu, a) \mu_s^\theta(da) ds \\
 & + \frac{1}{\theta} \int_0^t \left[ \int_U b(s, x_s^\mu, a) \mu_s^\theta(da) - \int_U b(s, x_s^\mu, a) \mu_s(da) \right] ds \\
 & + \frac{1}{\theta} \int_0^t \left[ \int_U \sigma(s, x_s^\theta, a) \mu_s^\theta(da) - \int_U \sigma(s, x_s^\mu, a) \mu_s^\theta(da) \right] dW_s \\
 & + \frac{1}{\theta} \int_0^t \left[ \int_U \sigma(s, x_s^\mu, a) \mu_s^\theta(da) - \int_U \sigma(s, x_s^\mu, a) \mu_s(da) \right] dW_s \\
 & - \int_0^t \int_U b_x(s, x_s^\mu, a) \mu_s(da) z_s ds - \int_0^t \int_U \sigma_x(s, x_s^\mu, a) \mu_s(da) z_s dW_s \\
 & - \int_0^t \left[ \int_U b(s, x_s^\mu, a) \mu_s(da) - \int_U b(s, x_s^\mu, a) q_s(da) \right] ds \\
 & - \int_0^t \left[ \int_U \sigma(s, x_s^\mu, a) \mu_s(da) - \int_U \sigma(s, x_s^\mu, a) q_s(da) \right] dW_s.
 \end{aligned}$$

By using the definition of  $\mu^\theta$  and taking expectation, we get

$$\begin{aligned}
 \mathbb{E} |X_t|^2 \leq & C \mathbb{E} \int_0^t \int_0^1 \int_U |b_x(s, x_s^\mu + \lambda \theta(X_s + z_s), a) X_s|^2 \mu_s(da) d\lambda ds \\
 & + C \mathbb{E} \int_0^t \int_0^1 \int_U |\sigma_x(s, x_s^\mu + \lambda \theta(X_s + z_s), a) X_s|^2 \mu_s(da) d\lambda ds \\
 & + C \mathbb{E} |\alpha_t^\theta|^2,
 \end{aligned}$$

where  $\alpha_t^\theta$  is given by

$$\begin{aligned}
 \alpha_t^\theta = & \int_0^t \int_0^1 \int_U b_x(s, x_s^\mu + \lambda \theta(X_s + z_s), a) (x_s^\theta - x_s^\mu) q_s(da) d\lambda ds \\
 & - \int_0^t \int_0^1 \int_U b_x(s, x_s^\mu + \lambda \theta(X_s + z_s), a) (x_s^\theta - x_s^\mu) \mu_s(da) d\lambda ds \\
 & + \int_0^t \int_0^1 \int_U \sigma_x(s, x_s^\mu + \lambda \theta(X_s + z_s), a) (x_s^\theta - x_s^\mu) q_s(da) d\lambda dW_s \\
 & - \int_0^t \int_0^1 \int_U \sigma_x(s, x_s^\mu + \lambda \theta(X_s + z_s), a) (x_s^\theta - x_s^\mu) \mu_s(da) d\lambda dW_s \\
 & + \int_0^t \int_0^1 \int_U b_x(s, x_s^\mu + \lambda \theta(X_s + z_s), a) z_s \mu_s(da) d\lambda ds \\
 & + \int_0^t \int_0^1 \int_U \sigma_x(s, x_s^\mu + \lambda \theta(X_s + z_s), a) z_s \mu_s(da) d\lambda dW_s \\
 & - \int_0^t \int_U b_x(s, x_s^\mu, a) z_s \mu_s(da) ds - \int_0^t \int_U \sigma_x(s, x_s^\mu, a) z_s \mu_s(da) dW_s.
 \end{aligned}$$

Since  $b_x$  and  $\sigma_x$  are continuous and bounded, then

$$\mathbb{E} |X_t|^2 \leq C \mathbb{E} \int_0^t |X_s|^2 ds + C \mathbb{E} |\alpha_t^\theta|^2,$$

$$\lim_{\theta \rightarrow 0} \mathbb{E} |\alpha_t^\theta|^2 = 0.$$

We conclude by using Gronwall's lemma in the above inequality.  $\square$

LEMMA 11. *Let  $\mu$  be an optimal relaxed control minimizing the cost  $J$  over  $\mathcal{R}$  and  $x_t^\mu$  the associated optimal trajectory. Then, for any  $q \in \mathcal{R}$ , we have*

$$(14) \quad 0 \leq \mathbb{E} [g(x_T^\mu) z_T] + \mathbb{E} \int_0^T \int_U h_x(t, x_t^\mu, a) \mu_t(da) z_t dt \\ + \mathbb{E} \int_0^T \left[ \int_U h(t, x_t^\mu, a) q_t(da) - \int_U h(t, x_t^\mu, a) \mu_t(da) \right] dt.$$

*Proof.* By (9), we have

$$0 \leq \mathbb{E} [g(x_T^\theta) - g(x_T^\mu)] \\ + \mathbb{E} \int_0^T \int_U h(t, x_t^\theta, a) \mu_t^\theta(da) dt - \mathbb{E} \int_0^T \int_U h(t, x_t^\mu, a) \mu_t(da) dt \\ = \mathbb{E} [g(x_T^\theta) - g(x_T^\mu)] \\ + \mathbb{E} \int_0^T \int_U h(t, x_t^\theta, a) \mu_t^\theta(da) dt - \mathbb{E} \int_0^T \int_U h(t, x_t^\mu, a) \mu_t^\theta(da) dt \\ + \mathbb{E} \int_0^T \int_U h(t, x_t^\mu, a) \mu_t^\theta(da) dt - \mathbb{E} \int_0^T \int_U h(t, x_t^\mu, a) \mu_t(da) dt.$$

By using the definition of  $\mu_t^\theta$ , we have

$$0 \leq \mathbb{E} [g(x_T^\theta) - g(x_T^\mu)] \\ + \mathbb{E} \int_0^T \left[ \int_U h(t, x_t^\theta, a) \mu_t(da) - \int_U h(t, x_t^\mu, a) \mu_t(da) \right] dt \\ + \theta \mathbb{E} \int_0^T \left[ \int_U h(t, x_t^\theta, a) q_t(da) - \int_U h(t, x_t^\theta, a) \mu_t(da) \right] dt.$$

Hence,

$$0 \leq \mathbb{E} \int_0^1 g_x(x_T^\mu + \lambda \theta (X_T + z_T)) z_T d\lambda \\ + \mathbb{E} \int_0^T \int_U \int_0^1 h_x(t, x_t^\mu + \lambda \theta (X_t + z_t), a) \mu_t(da) z_t d\lambda dt \\ + \mathbb{E} \int_0^T \left[ \int_U h(t, x_t^\mu, a) q_t(da) - \int_U h(t, x_t^\mu, a) \mu_t(da) \right] dt \\ + \rho_t^\theta,$$

where  $X$  is defined in (13) and  $\rho_t^\theta$  is given by

$$\rho_t^\theta = \mathbb{E} \int_0^1 g_x(x_T^\mu + \lambda \theta (X_T + z_T)) X_T d\lambda \\ + \mathbb{E} \int_0^T \int_U \int_0^1 h_x(t, x_t^\mu + \lambda \theta (X_t + z_t), a) \mu_t(da) X_t d\lambda dt.$$

Using the Cauchy–Schwartz inequality, Lemma 10, and the fact that  $g_x$  and  $h_x$  are continuous and bounded, we get

$$\lim_{\theta \rightarrow 0} \rho_t^\theta = 0.$$

The proof is completed by letting  $\theta$  go to 0 in the above inequality.  $\square$

**3.2. Variational inequality and adjoint equation.** In this subsection, we introduce the adjoint process. With this process, we derive the variational inequality from (14). The linear terms in (11) may be treated in the following way (see Bensoussan [5]). Let  $\Phi$  be the fundamental solution of the linear equation

$$(15) \quad \begin{cases} d\Phi_t = \int_U b_x(t, x_t^\mu, a) \mu_t(da) \Phi_t dt + \int_U \sigma_x(t, x_t^\mu, a) \mu_t(da) \Phi_t dW_t, \\ \Phi_0 = I_d. \end{cases}$$

This equation is linear with bounded coefficients. Hence, it admits a unique strong solution which is invertible, and its inverse  $\Psi_t$  is the unique solution of

$$(16) \quad \begin{cases} d\Psi_t = \left[ \int_U \sigma_x(t, x_t^\mu, a) \mu_t(da) \Psi_t \int_U \sigma_x^*(t, x_t^\mu, a) \mu_t(da) \right] dt \\ \quad - \int_U b_x(t, x_t^\mu, a) \mu_t(da) \Psi_t dt - \int_U \sigma_x(t, x_t^\mu, a) \mu_t(da) \Psi_t dW_t, \\ \Psi_0 = I_d. \end{cases}$$

Moreover,  $\Phi$  and  $\Psi$  satisfy

$$(17) \quad \mathbb{E} \left[ \sup_{0 \leq t \leq T} |\Phi_t|^2 \right] + \mathbb{E} \left[ \sup_{0 \leq t \leq T} |\Psi_t|^2 \right] < \infty.$$

We introduce the following processes:

$$(18) \quad \alpha_t = \Psi_t z_t,$$

$$(19) \quad X = \Phi_T^* g_x(x_T^\mu) + \int_0^T \left[ \Phi_t^* \int_U h_x(t, x_t^\mu, a) \mu_t(da) \right] dt,$$

$$(20) \quad Y_t = \mathbb{E}[X / \mathcal{F}_t] - \int_0^t \left[ \Phi_s^* \int_U h_x(s, x_s^\mu, a) \mu_s(da) \right] ds.$$

We remark from (18), (19), (20) that

$$(21) \quad \mathbb{E}[\alpha_T Y_T] = \mathbb{E}[g_x(x_T^\mu) z_T].$$

Since  $g_x$  and  $h_x$  are bounded, then by (17),  $X$  is square integrable. Hence, the process  $(\mathbb{E}[X / \mathcal{F}_t])_{t \geq 0}$  is a square integrable martingale with respect to the natural filtration of the Brownian motion  $W$ . Then, by Itô's representation theorem we have

$$Y_t = \mathbb{E}[X] + \int_0^t Q_s dW_s - \int_0^t \int_U \Phi_s^* h_x(s, x_s^\mu, a) \mu_s(da) ds,$$

where  $Q$  is an adapted process such that  $\mathbb{E} \int_0^T |Q_s|^2 ds < \infty$ .

By applying Itô's formula to  $\alpha_t$ , then with  $\alpha_t Y_t$  and using (21), the variational inequality (14) becomes

$$(22) \quad 0 \leq \mathbb{E} \int_0^T [\mathcal{H}(t, x_t^\mu, q_t, p_t^\mu, P_t^\mu) - \mathcal{H}(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu)] dt,$$

where the Hamiltonian  $\mathcal{H}$  is defined from  $[0, T] \times \mathbb{R}^n \times \mathbb{P}(U) \times \mathbb{R}^n \times \mathcal{M}_{n \times d}(\mathbb{R})$  into  $\mathbb{R}$  by

$$\mathcal{H}(t, x, q, p, P) = \int_U h(t, x, a) q(da) + \int_U b(t, x, a) q(da) p + \int_U \sigma(t, x, a) q(da) P,$$

$(p^\mu, P^\mu)$  is a pair of adapted processes given by

$$(23) \quad p_t^\mu = \Psi_t^* Y_t; \quad p^\mu \in \mathcal{L}^2([0, T]; \mathbb{R}^n),$$

$$(24) \quad P_t^\mu = \Psi_t^* Q_t - \int_U \sigma_x^*(t, x_t^\mu, a) \mu_t(da) p_t^\mu; \quad P^\mu \in \mathcal{L}^2([0, T]; \mathbb{R}^{n \times d}),$$

and the process  $Q$  satisfies

$$(25) \quad \int_0^t Q_s dW_s = \mathbb{E} \left[ \Phi_T^* g_x(x_T^\mu) + \int_0^T \Phi_t^* \int_U h_x(t, x_t^\mu, a) \mu_t(da) dt / \mathcal{F}_t \right] \\ - \mathbb{E} \left[ \Phi_T^* g_x(x_T^\mu) + \int_0^T \Phi_t^* \int_U h_x(t, x_t^\mu, a) \mu_t(da) dt \right].$$

The process  $p^\mu$  is called the adjoint process and from (19), (20), (23), it is given explicitly by

$$p_t^\mu = \mathbb{E} \left[ \Psi_t^* \Phi_T^* g_x(x_T^\mu) + \Psi_t^* \int_t^T \Phi_s^* \int_U h_x(s, x_s^\mu, a) \mu_s(da) ds / \mathcal{F}_t \right].$$

By applying Itô's formula to the adjoint processes  $p^\mu$  in (23), we obtain the adjoint equation, which is a linear backward stochastic differential equation, given by

$$(26) \quad \begin{cases} -dp_t^\mu = \mathcal{H}_x(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu) dt - P_t^\mu dW_t, \\ p_T^\mu = g_x(x_T^\mu). \end{cases}$$

**3.3. Necessary optimality conditions for relaxed controls.** Starting from the variational inequality (22), we can now state the necessary optimality conditions, for the relaxed control problem  $\{(5), (6), (7)\}$ , in the global form.

**THEOREM 12** (necessary optimality conditions for relaxed controls in global form). *Let  $\mu$  be an optimal relaxed control minimizing the cost  $J$  over  $\mathcal{R}$  and let  $x^\mu$  denote the corresponding optimal trajectory. Then, there exists a unique pair of adapted processes*

$$(p^\mu, P^\mu) \in \mathcal{L}^2([0, T]; \mathbb{R}^n) \times \mathcal{L}^2([0, T]; \mathbb{R}^{n \times d}),$$

which are solution of the backward stochastic differential equations (26) such that

$$(27) \quad \mathcal{H}(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu) = \inf_{q_t \in \mathbb{P}(U)} \mathcal{H}(t, x_t^\mu, q_t, p_t^\mu, P_t^\mu); \quad a.e., \quad a.s.$$

*Proof.* The result follows immediately from (22).  $\square$

**Remark 13.** Bahlali, Mezerdi, and Djehiche [2] established necessary optimality conditions for relaxed controls of the second-order with two adjoint processes. The result of the above theorem improves that of [2], in the sense where we consider the same relaxed control problem, in which the control variable enters both the drift and the diffusion coefficients, and we establish necessary optimality conditions of the first-order with only one adjoint process.

**3.4. Sufficient optimality conditions for relaxed controls.** In this subsection, we study when the necessary optimality conditions (27) become sufficient. We recall assumptions (4) and the adjoint equation (26). For any  $q \in \mathcal{R}$ , we denote by  $x^q$  the solution of (5) controlled by  $q$ .

**THEOREM 14** (sufficient optimality conditions for relaxed controls). *If we assume that the functions  $g$  and  $x \mapsto \mathcal{H}(t, x, q, p, P)$  are convex, then  $\mu$  is an optimal solution of the problem  $\{(5), (6), (7)\}$  if it satisfies (27).*

*Proof.* We know that the set of relaxed controls  $\mathcal{R}$  is convex and the Hamiltonian  $\mathcal{H}$  is linear in  $q$ . Let  $\mu$  be an arbitrary element of  $\mathcal{R}$  (candidate to be optimal). For any  $q \in \mathcal{R}$ , we have

$$\begin{aligned} J(\mu) - J(q) &= \mathbb{E}[g(x_T^\mu) - g(x_T^q)] \\ &\quad + \mathbb{E} \int_0^T \left[ \int_U h(t, x_t^\mu, a) \mu_t(da) - \int_U h(t, x_t^q, a) q_t(da) \right] dt. \end{aligned}$$

Since  $g$  is convex, we get

$$g(x_T^q) - g(x_T^\mu) \geq g_x(x_T^\mu)(x_T^q - x_T^\mu).$$

Thus,

$$g(x_T^\mu) - g(x_T^q) \leq g_x(x_T^\mu)(x_T^\mu - x_T^q).$$

Hence,

$$\begin{aligned} J(\mu) - J(q) &\leq \mathbb{E}[g_x(x_T^\mu)(x_T^\mu - x_T^q)] \\ &\quad + \mathbb{E} \int_0^T \left[ \int_U h(t, x_t^\mu, a) \mu_t(da) - \int_U h(t, x_t^q, a) q_t(da) \right] dt. \end{aligned}$$

We remark that  $p_T^\mu = g_x(x_T^\mu)$ ; then we have

$$\begin{aligned} J(\mu) - J(q) &\leq \mathbb{E}[p_T^\mu(x_T^\mu - x_T^q)] \\ &\quad + \mathbb{E} \int_0^T \left[ \int_U h(t, x_t^\mu, a) \mu_t(da) - \int_U h(t, x_t^q, a) q_t(da) \right] dt. \end{aligned}$$

Applying Itô's formula to  $p_t^\mu(x_t^\mu - x_t^q)$  and taking expectation, we obtain

$$\begin{aligned} (28) \quad J(\mu) - J(q) &\leq \mathbb{E} \int_0^T [\mathcal{H}(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu) - \mathcal{H}(t, x_t^q, q_t, p_t^\mu, P_t^\mu)] dt \\ &\quad - \mathbb{E} \int_0^T \mathcal{H}_x(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu)(x_t^\mu - x_t^q) dt. \end{aligned}$$

Since  $\mathcal{H}$  is convex in  $x$  and linear in  $\mu$ , then by using the Clarke generalized gradient of  $\mathcal{H}$  evaluated at  $(x_t, \mu_t)$  and the necessary optimality conditions (27), it follows by [25, Lemmas 2.2 and 2.3] that

$$\mathcal{H}(t, x_t^q, q_t, p_t^\mu, P_t^\mu) - \mathcal{H}(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu) \geq \mathcal{H}_x(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu)(x_t^q - x_t^\mu),$$

or equivalently

$$0 \geq \mathcal{H}(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu) - \mathcal{H}(t, x_t^q, q_t, p_t^\mu, P_t^\mu) - \mathcal{H}_x(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu)(x_t^\mu - x_t^q).$$

By the above inequality and (28), we have

$$J(\mu) - J(q) \leq 0.$$

The theorem is then proved.  $\square$



**4. Necessary and sufficient optimality conditions for strict controls.** In this section, we study the strict control problem  $\{(1), (2), (3)\}$ , and from the results of section 3, we derive the optimality conditions for strict controls. To this end, consider the following subset of  $\mathcal{R}$ :

$$(29) \quad \delta(\mathcal{U}) = \{q \in \mathcal{R} \mid q = \delta_v; \ v \in \mathcal{U}\}.$$

The set  $\delta(\mathcal{U})$  is the collection of all relaxed controls in the form of Dirac measure charging a strict control.

Denote by  $\delta(U)$  the action set of all relaxed controls in  $\delta(\mathcal{U})$ .

If  $q \in \delta(\mathcal{U})$ , then  $q = \delta_v$  with  $v \in \mathcal{U}$ . In this case we have for each  $t$ ,  $q_t \in \delta(U)$  and  $q_t = \delta_{v_t}$ .

*Remark 15.* The necessary and sufficient optimality conditions for relaxed controls, respectively Theorems 12 and 14, hold if we replace  $\mathcal{R}$  by  $\delta(\mathcal{U})$  and  $\mathbb{P}(U)$  by  $\delta(U)$ .

**LEMMA 16.** *The relaxed control  $\mu = \delta_u$  minimizes  $J$  over  $\delta(\mathcal{U})$  if and only if the strict control  $u$  minimizes  $J$  over  $\mathcal{U}$ .*

*Proof.* Let  $\mu = \delta_u$  be an optimal relaxed control minimizing the cost  $J$  over  $\delta(\mathcal{U})$ ; we then have

$$(30) \quad J(\mu) \leq J(q) \quad \forall q \in \delta(\mathcal{U}).$$

Since  $q \in \delta(\mathcal{U})$ , then there exists  $v \in \mathcal{U}$  such that  $q = \delta_v$ . It is easy to see that

$$(31) \quad \begin{cases} x^\mu = x^u, \\ x^q = x^v, \\ J(\mu) = J(u), \\ J(q) = J(v). \end{cases}$$

By (30), we get

$$J(u) \leq J(v) \quad \forall v \in \mathcal{U}.$$

Conversely, let  $u$  be a strict control minimizing the cost  $J$  over  $\mathcal{U}$ . Then

$$J(u) \leq J(v) \quad \forall v \in \mathcal{U}.$$

Since the controls  $u, v \in \mathcal{U}$ , then there exist  $\mu, q \in \delta(\mathcal{U})$  such that

$$\mu = \delta_u, \quad q = \delta_v.$$

This implies that relations (31) hold. Consequently, we get

$$J(\mu) \leq J(q) \quad \forall q \in \delta(\mathcal{U}).$$

The proof is completed.  $\square$

*Remark 17.* (1) The relaxed optimal control  $\mu$  exists, but it is not necessarily an element of  $\delta(\mathcal{U})$ . A strict optimal control  $u$  does not necessarily exist.

(2) If the relaxed optimal control  $\mu \in \delta(\mathcal{U})$ , then we get existence of a strict optimal control. In this case, if we reduce the class of relaxed controls  $\mathcal{R}$  to the set  $\delta(\mathcal{U})$ , then the relaxed control problem simply becomes a strict control problem in which an optimal solution exists.

(3) We know that existence of an optimal solution of strict control problem is ensured by the Filippov condition. It is interesting to see that if we have the Filippov condition, then the relaxed optimal control is an element of  $\delta(\mathcal{U})$ .

**4.1. Necessary optimality conditions for strict controls.** Define the Hamiltonian in the strict case from  $[0, T] \times \mathbb{R}^n \times U \times \mathbb{R}^n \times \mathcal{M}_{n \times d}(\mathbb{R})$  into  $\mathbb{R}$  by

$$H(t, x, v, p, P) = h(t, x, v) + b(t, x, v)p + \sigma(t, x, v)P.$$

**THEOREM 18** (necessary optimality conditions for strict controls in global form). *Suppose that  $u$  is an optimal strict control minimizing the cost  $J$  over  $\mathcal{U}$  and  $x^u$  denotes the solution of (1) controlled by  $u$ . Then, there exists a unique pair of adapted processes*

$$(p^u, P^u) \in \mathcal{L}^2([0, T]; \mathbb{R}^n) \times \mathcal{L}^2([0, T]; \mathbb{R}^{n \times d}),$$

which is a solution of the following backward stochastic differential equation:

$$(32) \quad \begin{cases} -dp_t^u = H_x(t, x_t^u, u_t, p_t^u, P_t^u) dt - P_t^u dW_t, \\ p_T^u = g_x(x_T^u), \end{cases}$$

such that

$$(33) \quad H(t, x_t^u, u_t, p_t^u, P_t^u) = \inf_{v_t \in U} H(t, x_t^u, v_t, p_t^u, P_t^u); \quad a.e., \quad a.s.$$

*Proof.* Let  $u$  be an optimal solution of the strict control problem  $\{(1), (2), (3)\}$  and  $v$  be an arbitrary element of  $\mathcal{U}$ . Then, there exist  $\mu, q \in \delta(\mathcal{U})$  such that

$$(34) \quad \mu = \delta_u, \quad q = \delta_v.$$

Since  $u$  minimizes the cost  $J$  over  $\mathcal{U}$ , then by Lemma 16,  $\mu$  minimizes  $J$  over  $\delta(\mathcal{U})$ . Hence, by the necessary optimality conditions for relaxed controls (Theorem 12), there exists a unique pair of adapted processes  $(p_t^\mu, P_t^\mu)$ , which is a solution of (26) such that

$$\mathcal{H}(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu) = \inf_{q_t \in \delta(U)} \mathcal{H}(t, x_t^\mu, q_t, p_t^\mu, P_t^\mu); \quad a.e., \quad a.s.$$

By (34) we can easily see that

$$\begin{aligned} x^\mu &= x^u, \\ \mathcal{H}(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu) &= H(t, x_t^u, u_t, p_t^u, P_t^u), \\ \mathcal{H}(t, x_t^\mu, q_t, p_t^\mu, P_t^\mu) &= H(t, x_t^u, v_t, p_t^u, P_t^u), \end{aligned}$$

where the pair  $(p^u, P^u)$  is the unique solution of (32). The theorem is then proved.  $\square$

*Remark 19.* Peng [23] established necessary optimality conditions for strict controls of the second-order with two adjoint processes. The result of the above theorem improves that of Peng, in the sense where we consider the same strict control problem, with nonconvex control domain and a general state equation in which the control variable enters both the drift and the diffusion coefficients, and we establish necessary optimality conditions of the first-order with only one adjoint process.

**4.2. Sufficient optimality conditions for strict controls.** We recall assumptions (4) and the adjoint equation (32).

**THEOREM 20** (sufficient optimality conditions for strict controls). *If we assume that  $g$  and the map  $x \mapsto H(t, x, v, p, P)$  are convex, then  $u$  is an optimal solution of the problem  $\{(1), (2), (3)\}$ , if it satisfies (33).*

*Proof.* Let  $u$  be a strict control (candidate to be optimal) such that the necessary optimality conditions for strict controls (33) hold. Then, we have

$$H(t, x_t^u, u_t, p_t^u, P_t^u) = \inf_{v_t \in U} H(t, x_t^u, v_t, p_t^u, P_t^u); \quad a.e., \quad a.s.$$

The controls  $u, v$  are elements of  $\mathcal{U}$ ; then there exist  $\mu, q \in \delta(\mathcal{U})$  such that

$$\mu = \delta_u, \quad q = \delta_v.$$

This implies that

$$\begin{aligned} x^\mu &= x^u, \\ \mathcal{H}(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu) &= H(t, x_t^u, u_t, p_t^u, P_t^u), \\ \mathcal{H}(t, x_t^\mu, q_t, p_t^\mu, P_t^\mu) &= H(t, x_t^u, v_t, p_t^u, P_t^u). \end{aligned}$$

By the above equalities and the necessary optimality conditions for strict controls (33), we deduce that

$$\mathcal{H}(t, x_t^\mu, \mu_t, p_t^\mu, P_t^\mu) = \inf_{q_t \in \delta(U)} \mathcal{H}(t, x_t^\mu, q_t, p_t^\mu, P_t^\mu); \quad a.e., \quad a.s.$$

Since  $H$  is convex in  $x$ , it is easy to see that  $\mathcal{H}$  is convex in  $x$ , and since  $g$  is convex, then from the sufficient optimality conditions for relaxed controls (Theorem 14),  $\mu$  minimizes the cost  $J$  over  $\delta(\mathcal{U})$ . By Lemma 16, we deduce that  $u$  minimizes the cost  $J$  over  $\mathcal{U}$ . The theorem is then proved.  $\square$

*Remark 21.* The sufficient optimality conditions for strict controls are proved without assuming either the convexity of  $U$  or that of  $H$  in  $v$ .

**Acknowledgment.** The author thanks the referee who offered many useful remarks and suggestions that improved the first version of this paper.

## REFERENCES

- [1] V. I. ARKIN AND M. T. SAKSONOV, *Necessary optimality conditions for stochastic differential equations*, Soviet. Math. Dokl., 20 (1979), pp. 1–5.
- [2] S. BAHALALI, B. MEZERDI, AND B. DJEHICHE, *Approximation and optimality necessary conditions in relaxed stochastic control problems*, J. Appl. Math. Stoch. Anal., (5) 2006, pp. 1–23.
- [3] S. BAHALALI, B. DJEHICHE, AND B. MEZERDI, *The relaxed stochastic maximum principle in singular control of diffusions*, SIAM J. Control Optim., 46 (2007), pp. 427–444.
- [4] H. BECKER AND V. MANDREKAR, *On the existence of optimal random controls*, J. Math. Mech., 18 (1969), pp. 1151–1166.
- [5] A. BENSOUSSAN, *Lectures on stochastic control*, in Nonlinear Filtering and Stochastic Control, Lecture Notes in Math. 972, Springer-Verlag, Berlin, 1982.
- [6] A. CADENILLAS AND I. KARATZAS, *The stochastic maximum principle for linear convex optimal control with random coefficients*, SIAM J. Control. Optim., 33 (1995), pp. 590–624.
- [7] N. EL KAROUI, N. H. NGUYEN, AND M. JEANBLANC-PICQUÉ, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics, 20 (1987), pp. 169–219.
- [8] R. J. ELLIOTT, *The optimal control of diffusions*, Appl. Math. Optim., 22 (1990), pp. 229–240.

- [9] R. J. ELLIOTT AND M. KOHLMANN, *The second order minimum principle and adjoint process*, Stochastics Stochastics Rep., 46 (1994), pp. 25–39.
- [10] W. H. FLEMING, *Generalized solutions in optimal stochastic control*, in Differential Games and Control Theory 2, (Proc. 2nd Kingston Conference, Kingston, RI, 1976), Lecture Notes in Pure and Appl. Math., 30 (1977), pp. 147–165.
- [11] U. G. HAUSSMANN, *General necessary conditions for optimal control of stochastic systems*, Math. Programming Stud., 6 (1976), pp. 30–48.
- [12] U. G. HAUSSMANN, *A stochastic maximum principle for optimal control of diffusions*, Pitman Research Notes in Math Series 151, John Wiley & Sons, New York, 1986.
- [13] U. G. HAUSSMANN, *Existence of optimal Markovian controls for degenerate diffusions*, in Stochastic Differential Systems, Lecture Notes in Control and Inform. Sci. 78, Springer-Verlag, Berlin, 1986, pp. 171–186.
- [14] U. G. HAUSSMANN AND J. P. LEPELTIER, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.
- [15] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, 2nd ed., Kodansha, Tokyo, North-Holland, Amsterdam, 1989.
- [16] J. JACOD AND J. MÉMIN, *Sur un type de convergence intermédiaire entre la convergence en loi et la convergence en probabilité*, Seminar on Probability XV, Lecture Notes in Math. 850, Springer-Verlag, Berlin, 1981.
- [17] I. KARATZAS AND S. SHREVE, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
- [18] H. J. KUSHNER, *Necessary conditions for continuous parameter stochastic optimization problems*, SIAM J. Control Optim., 10 (1973), pp. 550–565.
- [19] H. J. KUSHNER, *Existence results for optimal stochastic controls*, J. Optim. Theory Appl., 15 (1975), pp. 347–359.
- [20] J. MA AND J. YONG, *Solvability of forward-backward SDEs and the nodal set of Hamilton-Jacobi-Bellman equations*, A Chinese summary appears in Chinese Ann. Math. Ser. A, 16 (1995), p. 532, Chinese Ann. Math. Ser. B, 16 (1995), pp. 279–298.
- [21] B. MEZERDI AND S. BAHLALI, *Approximation in optimal control of diffusion processes*, Random Oper. Stochastic Equations, 8 (2000), pp. 365–372.
- [22] B. MEZERDI AND S. BAHLALI, *Necessary conditions for optimality in relaxed stochastic control problems*, Stochastics Stochastics Rep., 73 (2002), pp. 201–218.
- [23] S. PENG, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.
- [24] J. YONG AND X. Y. ZHOU, *Stochastic Controls, Hamilton Systems and HJB Equations*, Appl. Math. 43, Springer-Verlag, New York, 1999.
- [25] X. Y. ZHOU, *Sufficient conditions of optimality for stochastic systems with controllable diffusions*, IEEE Trans. Automat. Control, 41 (1996), pp. 1176–1179.

## A PROXIMAL-PROJECTION METHOD FOR FINDING ZEROS OF SET-VALUED OPERATORS\*

DAN BUTNARIU<sup>†</sup> AND GÁBOR KASSAY<sup>‡</sup>

**Abstract.** In this paper we study the convergence of an iterative algorithm for finding zeros with constraints for not necessarily monotone set-valued operators in a reflexive Banach space. This algorithm, which we call the proximal-projection method is, essentially, a fixed point procedure, and our convergence results are based on new generalizations of the Browder's demiclosedness principle. We show how the proximal-projection method can be applied for solving ill-posed variational inequalities and convex optimization problems with data given or computable by approximations only. The convergence properties of the proximal-projection method we establish also allow us to prove that the proximal point method (with Bregman distances), whose convergence was known to occur for maximal monotone operators, still converges when the operator involved in it is monotone with sequentially weakly closed graph.

**Key words.** Bregman distance, Bregman projection,  $D_f$ -antiresolvent,  $D_f$ -nonexpansivity pole,  $D_f$ -inverse strongly monotone operator,  $D_f$ -firm operator,  $D_f$ -nonexpansive operator,  $D_f$ -resolvent, firmly nonexpansive operator, Legendre function, maximal monotone operator, monotone operator, nonexpansive operator, proximal mapping, proximal point method, proximal projection method, relative projection, sequentially consistent function, projected subgradient method, Tikhonov–Browder regularization, strongly monotone operator, uniformly convex function, variational inequality

**AMS subject classifications.** Primary, 90C25, 47J25, 47J20; Secondary, 90C30, 90C48, 47N10

**DOI.** 10.1137/070682071

**1. Introduction.** In what follows,  $X$  denotes a real reflexive Banach space with norm  $\|\cdot\|$ , and  $X^*$  denotes the (topological) dual of  $X$  with the dual norm  $\|\cdot\|_*$ . Let  $f : X \rightarrow (-\infty, +\infty]$  be a proper, lower semicontinuous convex function with domain  $\text{dom } f$ . Then, the Fenchel conjugate  $f^* : X^* \rightarrow (-\infty, +\infty]$  is also a proper lower semicontinuous convex function and  $f^{**} := (f^*)^* = f$ . We assume that  $f$  is a Legendre function in the sense given to this term in [20, Definition 5.2]; that is,  $f$  is essentially smooth and essentially strictly convex. Then, according to [20, Theorem 5.4], the function  $f^*$  is a Legendre function too. Moreover, by [20, Theorems 5.6 and 5.10], both functions  $f$  and  $f^*$  have domains with nonempty interior, are (Gâteaux) differentiable on the interiors of their respective domains,

$$(1.1) \quad \text{ran } \nabla f = \text{dom } \nabla f^* = \text{int dom } f^* = \text{dom } \partial f^*,$$

$$(1.2) \quad \text{ran } \nabla f^* = \text{dom } \nabla f = \text{int dom } f = \text{dom } \partial f,$$

---

\*Received by the editors February 4, 2007; accepted for publication (in revised form) March 27, 2008; published electronically July 16, 2008.

<http://www.siam.org/journals/sicon/47-4/68207.html>

<sup>†</sup>Department of Mathematics, University of Haifa, 31905 Haifa, Israel (dbutnaru@math.haifa.ac.il). This author's work was partially done during his visit to The Graduate Center of the City University of New York. He gratefully acknowledges Professor Gabor Herman's support and the informative discussions and activities in which he was invited to take part.

<sup>‡</sup>Faculty of Mathematics and Computer Science, Babes-Bolyai University, 400084 Cluj-Napoca, Romania (kassay@math.ubbcluj.ro). This author gratefully acknowledges the support of the Cesarea Benjamin de Rothschild Foundation during his May 2006 visit to the Department of Mathematics of the University of Haifa, Israel, where part of this research was completed.

and

$$(1.3) \quad \nabla f = (\nabla f^*)^{-1}.$$

With the function  $f$  we associate the function  $W_f : X^* \times X \rightarrow (-\infty, +\infty]$  defined by

$$(1.4) \quad W_f(\xi, x) = f(x) - \langle \xi, x \rangle + f^*(\xi).$$

By the Young–Fenchel inequality the function  $W_f$  is nonnegative and  $\text{dom } W_f = (\text{dom } f^*) \times (\text{dom } f)$ . It is known (see [1], [2], [19], [38] and see also section 2 below) that, for any nonempty closed convex set  $E$  contained in  $X$  such that  $E \cap \text{int dom } f \neq \emptyset$ , the function  $\text{Proj}_E^f : \text{int dom } f^* \rightarrow X$ , given by

$$(1.5) \quad \text{Proj}_E^f \xi = \arg \min \{W_f(\xi, x) : x \in E\},$$

is well defined and its range is contained in  $E \cap \text{int dom } f$ . In fact, this function is a particular proximal projection, in the sense given to this term in [21], and in [38] is called the projection onto  $E$  relative to  $f$  because, in the particular case when  $X$  is a Hilbert space and  $f(x) = \frac{1}{2} \|x\|^2$ , the vector  $\text{Proj}_E^f \xi$  coincides with the usual (metric) projection of  $\xi$  onto  $E$ .

In this paper we are interested in the following problem.

**PROBLEM 1.1.** *Given an operator  $A : X \rightarrow 2^{X^*}$  and a nonempty closed subset  $C$  of  $X$  such that*

$$(1.6) \quad \emptyset \neq C \cap \text{dom } A \subseteq \text{int dom } f,$$

*find  $x \in C$  such that  $0^* \in Ax$ , where  $0^*$  denotes the null vector in  $X^*$ .*

Our purpose is to discover sufficient conditions for the following iterative procedure, which we call *the proximal-projection method*:

$$(1.7) \quad \begin{aligned} x^0 &\in C_0 \cap \text{dom } A \cap \text{int dom } f \quad \text{and} \\ x^{k+1} &\in \text{Proj}_{C_{k+1} \cap \text{dom } A}^f (\nabla f(x^k) - Ax^k) \quad \forall k \in \mathbb{N}, \end{aligned}$$

to generate approximations of solutions to Problem 1.1 when  $\{C_k\}_{k \in \mathbb{N}}$  is a sequence of closed convex subsets of  $X$  approximating weakly (see Definition 4.5) the set  $C$  under the following conditions of compatibility of  $f$  and  $C_k$  with the data of Problem 1.1.

**ASSUMPTION 1.2.** *For each  $k \in \mathbb{N}$ , the set  $C_k \cap \text{dom } A$  is convex, closed, and satisfies*

$$(1.8) \quad C \subseteq C_k \quad \text{and} \quad (\nabla f - A)(C_k) \subseteq \text{int dom } f^*.$$

In this paper (1.6) and Assumption 1.2 are standing assumptions, even if not explicitly mentioned, whenever we refer to Problem 1.1 or to the proximal-projection method. In view of (1.6) and (1.8) the sets  $C_k \cap \text{dom } A \cap \text{int dom } f$  are necessarily nonempty and, consequently,  $(\nabla f - A)(C_k)$  is nonempty too. This fact, Assumption 1.2, and Lemma 2.1, which ensures that

$$\text{dom Proj}_{C_k \cap \text{dom } A}^f = \text{int dom } f^* \quad \text{and} \quad \text{ran Proj}_{C_k \cap \text{dom } A}^f \subseteq C_k \cap \text{dom } A \cap \text{int dom } f,$$

taken together guarantee that the procedure of generating sequences in (1.7) is well defined.

The proximal-projection method described above is a natural generalization of Landweber's method of finding zeros of linear operators [53], of Shor's [75] (see also [76]) and Ermoliev's [46] "gradient descent" methods for finding unconstrained minima of convex functions, and of Polyak's "projected-subgradient" method for finding constrained minima of convex functions [66]. These methods inspired the construction of a plethora of algorithms for finding zeros of various operators as well as for other purposes. Among them are the algorithms presented in [1], [3], [4], [6], [7], [8], [9], [10], [11], [12], [13], [14], [17], [21], [30], [37], [38], [67], [74] which, in turn, inspired this research. The main formal differences between the proximal-projection method (1.7) and its classical counterparts developed in the 1950s and 1960s consist of the use of proximal projections, instead of metric projections, and of projecting not on the set  $C$  involved in Problem 1.1 but on some convex approximations  $C_k$  of the set  $C$ . The use of proximal projections instead of metric projections is mostly due to the fact that metric projections in Banach spaces which are not Hilbertian do not have many of those properties (like single valuedness and nonexpansivity) which make them so useful in a Hilbert space setting for establishing convergence of algorithms based on them. As far as we know, the idea of using proximal projections instead of metric projections in projected-subgradient type algorithms goes back to Alber's works [1], [2].

As we make clear in section 3.3, there is an intimate connection between the proximal-projection method and the well-known proximal point method with Bregman distances—see (3.23). The proximal point method with Bregman distances considered in this paper is itself a generalization of the classical proximal point algorithm developed in the 1950s by Krasnoselskii [51]; see also Moreau [59], [60], [61], Yosida [78], Martinet [57], [58], and Rockafellar [70], [71], among others (see [54] for a survey of the literature concerning the classical proximal point method). It emerged from the works of Erlander [45], Eriksson [44], Eggermont [43], and Eckstein [42], who studied various instances of the algorithm in  $\mathbb{R}^n$ . Its convergence analysis in Banach spaces which are not necessarily Hilbertian was initiated in [31], [32], and [50] (see [33] and [31] for related references on this topic). Lemma 3.8 shows that, in our setting, the proximal point method with Bregman distances is a particular instance of the proximal-projection method.

Computing projections, metric or proximal, onto a closed convex set with complicated geometry is, in itself, a challenging problem. It requires (see (1.5)) solving convex nonlinear programming problems with convex constraints. Specific techniques for finding proximal projections are presented in [5], [22], and [34]. It is obvious from these works that it is much easier to find proximal projections onto sets with simple geometry such as, for instance, hyperplanes, half spaces, or finite intersections of such sets. These facts naturally led to the question of whether it is possible, in the process of computing iterates of metric or proximal projections algorithms, to replace the constraint set  $C$  by some approximations  $C_k$  whose geometry is simple enough to allow relatively easy calculation of the required metric or proximal projections at each iterative step  $k$ . That this approach is sound is quite clear from the works of Mosco [62], its subsequent developments due to Attouch [15], Aubin and Frankowska [16], Bonnans and Shapiro [25], Dontchev and Zolezzi [41], and from the studies of Liskovets [55], [56]. Its main difficulty in the case of the proximal-projection method is that the approximations  $C_k$  one uses should converge to  $C$  in a manner that ensures stable convergence of the algorithm to solutions of the problem. For some variants of the proximal-projection method, types of convergence of the sets  $C_k$  to  $C$  which

are sufficiently good for this purpose are presented in [4], [7], [9], [12], [13]. They are mostly relaxed forms of Hausdorff metric convergence. It was shown in [8] that, in some circumstances, fast Mosco convergence of the sets  $C_k$  to  $C$  (see [8, Definition 2.1]), a form of convergence significantly less demanding than Hausdorff convergence, is sufficient to make the proximal-projection method (1.7) applied to variational inequalities converge. As we show below, these convergence requirements in the case of the proximal-projection method (1.7) can be further weakened. In fact, in our convergence theorems for the proximal-projection method we require only weak Mosco convergence of the sets  $C_k$  to  $C$  (see Definition 4.5(a)), and this is significantly less demanding than Hausdorff metric or fast Mosco convergence.

The purpose of this work is to find general conditions which guarantee that the proximal-projection method converges weakly or strongly to solutions of Problem 1.1. Observe that there is no apparent connection between the data of Problem 1.1 and the function  $f$  involved in the definition of the proximal-projection method. Our main question is how the function  $f$  should be chosen in order to ensure (weak or strong) convergence of the proximal-projection method to solutions of Problem 1.1 without excessively conditioning the problem data. The function  $f$  is a parameter of the proximal-projection method whose appropriate choice, as we show below, can make the procedure converge to solutions of Problem 1.1 even if the problem data are quite “bad” in the sense that they do not have properties that are usually difficult to verify in practice, such as maximal monotonicity, strict monotonicity, various forms of nonexpansivity, continuity, or closedness properties of some kind or another. Theorems 4.7 and 5.2 are our responses to the question posed above.

Theorem 4.7 shows that for guaranteeing that the proximal-projection method produces weak approximations of solutions for Problem 1.1, it is sufficient to choose a function  $f$  which, besides the conditions (1.6) and (1.8) which are meant to make the procedure consistent with the problem data, should satisfy some requirements, the majority of which are common features of the powers of the norm  $\|\cdot\|^p$  with  $p > 1$  in uniformly convex and smooth Banach spaces. The only somewhat outstanding condition we require for  $f$  is that the operator  $A$  involved in the problem be  $D_f$ -inverse strongly monotone on  $C$  or, if the set  $C$  is approximated by sets  $C_k$ , that  $D_f$ -inverse strong monotonicity of  $A$  should occur on the union of those sets.  $D_f$ -inverse strong monotonicity, a notion introduced in this paper (see Definition 3.3), is a generalization of the notion of firm nonexpansivity for operators in a Hilbert space (see (4.34)). Although in Hilbert spaces provided with the function  $f = \frac{1}{2} \|\cdot\|^2$  this notion coincides with the notion of firm nonexpansivity and, also, with the notion of  $D_f$ -firmness introduced in [21] (see Definition 3.4 below), outside this particular setting the notions of  $D_f$ -inverse strong monotonicity and  $D_f$ -firmness complement each other (cf. section 3.2). In section 4.3 we present several corollaries of Theorem 4.7 and examples which clearly show that fitting a function  $f$  to the specific data of Problem 1.1 may come naturally in many situations. If  $X$  is a Hilbert space and  $A$  is a firmly nonexpansive operator, then the natural choice is  $f = \frac{1}{2} \|\cdot\|^2$  (see Corollary 4.8). In this case the convergence of the proximal-projection method happens to be strong if the set of solutions of the problem has nonempty interior. If in Problem 1.1 we have  $C = X$ , then the problem is equivalent to that of finding a zero for the operator  $\nabla f - \nabla f \circ A_f$ , where  $A_f$  is the  $D_f$ -resolvent of  $A$  (a notion introduced in [21]—see also (3.19)), and application of the proximal-projection method with  $C_k = X$  to  $\nabla f - \nabla f \circ A_f$  is no more and no less than the proximal point method with Bregman distances mentioned above. The operator  $\nabla f - \nabla f \circ A_f$  is  $D_f$ -inverse strongly



monotone whenever the operator  $A$  is monotone (cf. Lemma 3.7). This leads us to the application of Theorem 4.7 to the operator  $\nabla f - \nabla f \circ A_f$ , which is Corollary 4.9. It shows that the proximal point algorithm with Bregman distances converges subsequentially weakly (and when  $A$  has a single zero, sequentially weakly) for a large class of functions  $f$  whenever  $A$  is monotone and provided that its graph is sequentially weakly closed (as happens, for instance, when graph  $A$  is convex and closed in  $X \times X^*$ ). It seems to us that this is the first time when weak convergence of the proximal point algorithm with Bregman distances is proved without requiring maximal monotonicity of  $A$ . The proximal-projection method is also a tool for solving monotone variational inequalities via their Tikhonov–Browder regularization. This is shown by Corollary 4.10, another consequence of Theorem 4.7. Corollary 4.10 also asks for the monotone operator  $B : X \rightarrow 2^{X^*}$  involved in the variational inequality to be such that  $\nabla f - \nabla f \circ \text{Proj}_C^f \circ [(1 - \alpha)\nabla f - B]$  is  $D_f$ -inverse strongly monotone (or, equivalently, such that  $\text{Proj}_C^f \circ [(1 - \alpha)\nabla f - B]$  is  $D_f$ -firm). This happens in many situations of practical interest. Several such situations are described in Examples 4.2 and 4.11.

A careful analysis of the proof of Theorem 4.7 reveals the fact that the proximal-projection method (1.7) is a procedure of approximating fixed points for the operator  $\text{Proj}_C^f \circ (\nabla f - A)$  by iterating the operator. Among the customary tools of proving convergence of such algorithms are Browder’s demiclosedness principle [29, Theorem 8.4] and, related to it, Opial’s convergence results for orbits of asymptotically regular operators [63, Theorems 1 and 2] and their generalizations. Our Theorem 4.7 is based on Proposition 4.4, a generalization of Browder’s demiclosedness principle, which works in reflexive Banach spaces and which is of interest by itself. If the Banach space  $X$  has finite dimension, Proposition 4.4 can be substantially improved—see Proposition 5.1. Thus, in spaces with finite dimension we can also improve Theorem 4.7 by dropping some of the requirements made on the problem data. This is, in fact, our Theorem 5.2, which guarantees convergence of the proximal-projection method with less demanding conditions than closedness of the graph of  $A$ . Accordingly, in finite-dimensional spaces the conclusions of Corollaries 4.9 and 4.10 can be reached at lesser cost for the operators involved in them, as shown by Corollaries 5.3 and 5.4, respectively.

This paper continues and develops a series of concepts, methods, and techniques initiated in [1], [19], [21], [33], [38], and [50]. In sections 2 and 3 we present in a unified approach the notions, notations, and preliminary results on which our convergence analysis of the proximal-projection method is based. It should be noted that, in section 2, some of the notions and results are presented in a more general setting than strictly needed in the subsequent parts of the material. This is done so because we hope to use the framework created in the current paper as a base for a forthcoming study of methods for solving nonclassical variational inequalities which are only tangentially approached here.

**2. Proximal mappings, relative projections, and variational inequalities.** In this section we present the notions, notations, and results concerning proximal projections and variational inequalities which are essential for the convergence analysis of the proximal-projection method done in what follows.

**2.1. Proximal mappings and relative projections.** Throughout this paper we denote by  $\mathcal{F}_f$  the set of proper, lower semicontinuous, convex functions  $\varphi : X \rightarrow$

$(-\infty, +\infty]$  which satisfy the conditions

$$(2.1) \quad \text{dom } \varphi \cap \text{int dom } f \neq \emptyset$$

and

$$(2.2) \quad \varphi_f := \inf \{ \varphi(x) : x \in \text{dom } \varphi \cap \text{dom } f \} > -\infty.$$

With every  $\varphi \in \mathcal{F}_f$  we associate the function  $\text{Env}_\varphi^f : X^* \rightarrow (-\infty, +\infty]$  given by

$$(2.3) \quad \text{Env}_\varphi^f(\xi) = \inf \{ \varphi(x) + W_f(\xi, x) : x \in X \}.$$

This is a natural generalization of the notion of the *Moreau envelope function* (see [72, Definition 1.22]). By (2.1) and (2.2) it results that the function  $\text{Env}_\varphi^f$  is proper and  $\text{dom } \text{Env}_\varphi^f = \text{dom } f^*$ . Another generalization of the *Moreau envelope function* is the function  $\text{env}_\varphi^f := \text{Env}_\varphi^f \circ \nabla f$  introduced and studied in [23]. Using Fenchel's duality theorem, it is easy to deduce that if  $\varphi \in \mathcal{F}_f$ , then

$$\text{Env}_\varphi^f(\xi) = f^*(\xi) - (\varphi + f)^*(\xi) = f^*(\xi) - (\varphi^* \square f^*)(\xi),$$

where  $\varphi^* \square f^*$  denotes the infimal convolution of  $\varphi^*$  and  $f^*$ .

The next result shows a way of generalizing the notion of Moreau proximal mapping (in the sense given to this term in [72]) whose study was initiated in [59], [60], [61] and further developed in [70], [71]. As we will make clear below, the generalization we propose here slightly differs from the notion of  $D_f$ -proximal mapping introduced and studied in [21]. In fact, most of the next lemma can also be deduced from [21, Propositions 3.22 and 3.23] due to the equality (2.7) established below. We prefer to present it here with a direct proof for the sake of completeness.

**LEMMA 2.1.** *Suppose that  $\varphi \in \mathcal{F}_f$ . For any  $\xi \in \text{int dom } f^*$  there exists a unique global minimizer, denoted  $\text{Prox}_\varphi^f(\xi)$ , of the function  $\varphi(\cdot) + W_f(\xi, \cdot)$ . The vector  $\text{Prox}_\varphi^f(\xi)$  is contained in  $\text{dom } \partial \varphi \cap \text{int dom } f$  and we have*

$$(2.4) \quad \text{Prox}_\varphi^f(\xi) = [\partial(\varphi + f)]^{-1}(\xi) = (\partial \varphi + \nabla f)^{-1}(\xi).$$

*Proof.* Let  $\xi \in \text{int dom } f^*$ . Then  $\text{Env}_\varphi^f(\xi)$  is finite and the function  $f - \langle \xi, \cdot \rangle$  is coercive (see [68, Theorem 7A] or [20, Fact 3.1]); that is, its sublevel sets

$$\text{lev}_{\leq}^f(\alpha) := \{x \in X : f(x) \leq \alpha\}$$

are bounded for all  $\alpha \in \mathbb{R}$ . Consequently, the function  $W_f(\xi, \cdot)$  is coercive too. Let  $\{x^k\}_{k \in \mathbb{N}}$  be a sequence contained in  $\text{dom } \varphi \cap \text{dom } f$  and such that

$$\lim_{k \rightarrow \infty} [\varphi(x^k) + W_f(\xi, x^k)] = \text{Env}_\varphi^f(\xi).$$

The sequence  $\{\varphi(x^k) + W_f(\xi, x^k)\}_{k \in \mathbb{N}}$  being convergent is also bounded. So, for some real number  $M > 0$  we have

$$W_f(\xi, x^k) \leq M - \varphi(x^k) \leq M - \varphi_f,$$

showing that the sequence  $\{x^k\}_{k \in \mathbb{N}}$  is contained in the sublevel set  $\text{lev}_{\leq}^\psi(M - \varphi_f)$  of the function  $\psi := W_f(\xi, \cdot)$ . By the coercivity of  $W_f(\xi, \cdot)$  it follows that the sequence

$\{x^k\}_{k \in \mathbb{N}}$  is bounded. Since the space  $X$  is reflexive, it results that  $\{x^k\}_{k \in \mathbb{N}}$  has a weakly convergent subsequence  $\{x^{i_k}\}_{k \in \mathbb{N}}$ . Let  $\bar{x} = \text{w-lim}_{k \rightarrow \infty} x^{i_k}$ . The functions  $f$  and  $\varphi$  are sequentially weakly lower semicontinuous because they are lower semicontinuous and convex. Hence,  $\varphi(\cdot) + W_f(\xi, \cdot)$  is also sequentially weakly lower semicontinuous and, thus, we have

$$\varphi(\bar{x}) + W_f(\xi, \bar{x}) \leq \liminf_{k \rightarrow \infty} [\varphi(x^{i_k}) + W_f(\xi, x^{i_k})] = \text{Env}_\varphi^f(\xi) < +\infty.$$

This implies that  $\bar{x} \in \text{dom } \varphi \cap \text{dom } f$  and that  $\bar{x}$  is a minimizer of  $\varphi(\cdot) + W_f(\xi, \cdot)$ .

Suppose that  $y$  is any minimizer of  $\varphi(\cdot) + W_f(\xi, \cdot)$ . Then  $y$  is also a minimizer of  $\varphi + f - \xi$ . Therefore, we have that  $0 \in \partial(\varphi + f - \xi)(y)$ , that is,  $\xi \in \partial(\varphi + f)(y)$ . The function  $f$  is continuous on  $\text{int dom } f$  because function  $f$  is convex and lower semicontinuous. This and (2.1) imply (see [69]) that  $\partial(\varphi + f)(y) = \partial\varphi(y) + \nabla f(y)$ . Hence,

$$(2.5) \quad \xi \in \partial\varphi(y) + \nabla f(y).$$

Since  $\text{dom } \nabla f = \text{int dom } f$  (see (1.2)), this implies that

$$y \in \text{dom } \partial\varphi \cap \text{dom } \nabla f = \text{dom } \partial\varphi \cap \text{int dom } f.$$

Hence, all minimizers of  $\varphi(\cdot) + W_f(\xi, \cdot)$  are contained in  $\text{dom } \partial\varphi \cap \text{int dom } f \subseteq \text{dom } \varphi \cap \text{int dom } f$ . The Legendre function  $f$  is strictly convex on the convex subsets of  $\text{dom } \partial f$  and, in particular, on the convex set  $\text{dom } \varphi \cap \text{int dom } f = \text{dom } \varphi \cap \text{dom } \partial f$ . Thus,  $\varphi(\cdot) + W_f(\xi, \cdot)$  is strictly convex on this set too. Consequently, there is at most one minimizer of  $\varphi(\cdot) + W_f(\xi, \cdot)$  on the convex set  $\text{dom } \varphi \cap \text{int dom } f$  and this proves that the minimizer  $\bar{x}$ , whose existence was established above, is unique. Formula (2.4) follows from (2.5) when  $y = \bar{x}$ .  $\square$

Lemma 2.1 ensures the well definedness of the function

$$(2.6) \quad \text{Prox}_\varphi^f : \text{int dom } f^* \rightarrow \text{dom } \partial\varphi \cap \text{int dom } f, \quad \xi \rightarrow \text{Prox}_\varphi^f(\xi)$$

when  $\varphi \in \mathcal{F}_f$ . We call this function *the proximal mapping relative to  $f$  associated to  $\varphi$* . The well definedness of the proximal mappings relative to the Legendre function  $f$  can also be deduced from [21, Theorem 3.18], where the well definedness of the resolvent

$$(2.7) \quad \text{prox}_\varphi^f := \text{Prox}_\varphi^f \circ \nabla f$$

was established. In more particular circumstances for  $f$  and  $\varphi$ , the well definedness of  $\text{Prox}_\varphi^f$  was proved in [1], [32], and [38].

Let  $E$  be a closed convex subset of  $X$  satisfying

$$(2.8) \quad E \cap \text{int dom } f \neq \emptyset.$$

Then the indicator function of the set  $E$ , that is, the function  $\iota_E : X \rightarrow (-\infty, +\infty]$  defined by  $\iota_E(x) = 0$  if  $x \in E$ , and  $\iota_E(x) = +\infty$  otherwise, is contained in  $\mathcal{F}_f$ . The operator  $\text{Prox}_{\iota_E}^f$  is called the *projection onto  $E$  relative to  $f$*  (cf. [38]) and is denoted  $\text{Proj}_E^f$  in what follows. According to Lemma 2.1, we have

$$\text{Proj}_E^f = (N_E + \nabla f)^{-1},$$

where  $N_E$  denotes the normal cone operator associated to the set  $E$ . The operator

$$(2.9) \quad \text{proj}_E^f := \text{prox}_{\iota_E}^f = \text{Proj}_E^f \circ \nabla f$$

is exactly the *Bregman projection onto  $E$  relative to  $f$* , whose importance in convex optimization was first emphasized in [24]. To see that, it is sufficient to recall that the *Bregman distance*  $D_f : X \times \text{int dom } f \rightarrow (-\infty, +\infty]$  is the function defined by

$$(2.10) \quad D_f(y, x) := f(y) - f(x) - \langle \nabla f(x), y - x \rangle = W_f(\nabla f(x), y),$$

with  $\text{dom } D_f = (\text{dom } f) \times (\text{int dom } f)$ .

**2.2. Variational inequalities.** There is an intimate connection between proximal mappings and variational inequalities. It is based on the following result, which extends the variational characterization of  $\text{Proj}_E^f$  given in [38] to a variational characterization of  $\text{Prox}_\varphi^f$ .

LEMMA 2.2. *Suppose that  $\varphi \in \mathcal{F}_f$  and  $\xi \in \text{int dom } f^*$ . If  $\hat{x} \in \text{dom } \partial\varphi \cap \text{int dom } f$ , then the following conditions are equivalent:*

- (a)  $\hat{x} = \text{Prox}_\varphi^f(\xi)$ ;
- (b)  $\hat{x}$  is a solution of the variational inequality

$$(2.11) \quad \langle \xi - \nabla f(x), y - x \rangle \leq \varphi(y) - \varphi(x) \quad \forall y \in \text{dom } \varphi \cap \text{dom } f;$$

- (c)  $\hat{x}$  is a solution of the variational inequality

$$W_f(\xi, x) + W_f(\nabla f(x), y) - W_f(\xi, y) \leq \varphi(y) - \varphi(x) \quad \forall y \in \text{dom } \varphi \cap \text{dom } f.$$

*Proof.* (a)  $\Rightarrow$  (b). Suppose that  $\hat{x} = \text{Prox}_\varphi^f(\xi)$ . Take  $y \in \text{dom } \varphi \cap \text{dom } f$  and  $t \in (0, 1)$ . Then  $(1 - t)\hat{x} + ty \in \text{dom } \varphi \cap \text{dom } f$  and we have

$$\varphi(\hat{x}) + W_f(\xi, \hat{x}) \leq \varphi((1 - t)\hat{x} + ty) + W_f(\xi, (1 - t)\hat{x} + ty),$$

that is,

$$(2.12) \quad \varphi((1 - t)\hat{x} + ty) - \varphi(\hat{x}) \geq W_f(\xi, \hat{x}) - W_f(\xi, (1 - t)\hat{x} + ty).$$

Since  $\hat{x} \in \text{int dom } f$ , there exists  $t_0 \in (0, 1)$  such that for any  $t \in (0, t_0)$  we have that  $(1 - t)\hat{x} + ty \in \text{int dom } f$ . Consequently, for any  $t \in (0, t_0)$  the function  $W_f(\xi, \cdot)$  is differentiable at  $(1 - t)\hat{x} + ty$ . Clearly, we also have

$$\nabla W_f(\xi, \cdot)(u) = \nabla f(u) - \xi \quad \forall u \in \text{int dom } f.$$

Therefore, by the convexity of  $W_f(\xi, \cdot)$  and (2.12), we deduce

$$\begin{aligned} t^{-1} [\varphi((1 - t)\hat{x} + ty) - \varphi(\hat{x})] &\geq \langle \nabla W_f(\xi, \cdot)((1 - t)\hat{x} + ty), \hat{x} - y \rangle \\ &= \langle \nabla f((1 - t)\hat{x} + ty) - \xi, \hat{x} - y \rangle \end{aligned}$$

for any  $t \in (0, t_0)$ . The function  $\nabla f(\cdot)$  is norm-to-weak continuous (see, for instance, [65, Proposition 2.8]). Hence, letting  $t \rightarrow 0^+$  in the last inequality, we get

$$\varphi^\circ(\hat{x}; y - \hat{x}) \geq \langle \nabla f(\hat{x}) - \xi, \hat{x} - y \rangle,$$

where  $\varphi^\circ$  stands for the right-hand side derivative of  $\varphi$ , that is,

$$\varphi^\circ(x; d) = \lim_{s \rightarrow 0^+} \frac{\varphi(x + sd) - \varphi(x)}{s}.$$

Taking into account that

$$\varphi(y) - \varphi(\hat{x}) \geq \varphi^\circ(\hat{x}; y - \hat{x})$$

we obtain (2.11).

(b) $\Rightarrow$ (a). Suppose that

$$\langle \xi - \nabla f(\hat{x}), y - \hat{x} \rangle \leq \varphi(y) - \varphi(\hat{x}) \quad \forall y \in \text{dom } \varphi \cap \text{dom } f.$$

Observe that

$$(2.13) \quad \nabla W_f(\xi, \cdot) = \nabla f(\cdot) - \xi.$$

Then, by the convexity of  $W_f(\xi, \cdot)$  and (2.13), for any  $y \in \text{dom } \varphi \cap \text{dom } f$ , we have that

$$\begin{aligned} W_f(\xi, y) - W_f(\xi, \hat{x}) &\geq \langle \nabla W_f(\xi, \cdot)(\hat{x}), y - \hat{x} \rangle \\ &= \langle \nabla f(\hat{x}) - \xi, y - \hat{x} \rangle \\ &\geq \varphi(\hat{x}) - \varphi(y). \end{aligned}$$

This shows that  $\hat{x} = \text{Prox}_\varphi^f(\xi)$ . The equivalence (b)  $\Leftrightarrow$  (c) results immediately by observing that

$$W_f(\xi, x) + W_f(\nabla f(x), y) - W_f(\xi, y) = \langle \xi - \nabla f(x), y - x \rangle$$

whenever  $y \in \text{dom } f$ ,  $\xi \in \text{int dom } f^*$ , and  $x \in \text{int dom } f$ .  $\square$

A consequence of Lemma 2.2 is the following generalization of the variational characterization of the Bregman projections originally given in [5].

**COROLLARY 2.3.** *Let  $x \in \text{int dom } f$  and let  $E$  be a nonempty, closed, and convex set such that  $E \cap \text{int dom } f \neq \emptyset$ . If  $\hat{x} \in E$ , then the following conditions are equivalent:*

- (i) *The vector  $\hat{x}$  is the Bregman projection of  $x$  onto  $E$  with respect to  $f$ ;*
- (ii) *the vector  $z = \hat{x}$  is the unique solution of the variational inequality*

$$\langle \nabla f(x) - \nabla f(z), z - y \rangle \geq 0 \quad \forall y \in E;$$

- (iii) *the vector  $z = \hat{x}$  is the unique solution of the variational inequality*

$$D_f(y, z) + D_f(z, x) \leq D_f(y, x) \quad \forall y \in E.$$

Now we are in position to establish the connection between the proximal mappings and a class of variational inequalities. It extends similar results known to hold in less general settings (see, for instance, [1] and [47, Proposition 1.5.8]). Here we consider the following variational inequality:

$$(2.14) \quad \text{Find } x \in \text{int dom } f \text{ such that}$$

$$\exists \xi \in Bx : [\langle \xi, y - x \rangle \geq \varphi(x) - \varphi(y) \quad \forall y \in \text{dom } f],$$

where  $\varphi \in \mathcal{F}_f$  and  $B : X \rightarrow 2^{X^*}$  is an operator which satisfies the condition

$$(2.15) \quad \emptyset \neq \text{dom } B \cap \text{dom } \varphi \cap \text{int dom } f \quad \text{and} \quad \text{ran } (\nabla f - B) \subseteq \text{int dom } f^*.$$

Condition (2.15) guarantees that the operator

$$\text{Prox}_\varphi^f(\nabla f - B) := \text{Prox}_\varphi^f \circ (\nabla f - B)$$

is well defined. Therefore, the following statement makes sense.

LEMMA 2.4. *Let  $\varphi \in \mathcal{F}_f$  and  $\hat{x} \in \text{dom } \partial\varphi \cap \text{int dom } f$ . Suppose that  $B : X \rightarrow 2^{X^*}$  is an operator which satisfies (2.15). Then  $\hat{x}$  is a solution of the variational inequality (2.14) if and only if it is a fixed point of the operator  $\text{Prox}_\varphi^f(\nabla f - B)$ .*

*Proof.* Note that  $\hat{x}$  is a solution of (2.14) if and only if there exists  $\xi \in B\hat{x}$  such that

$$(2.16) \quad \langle (\nabla f(\hat{x}) - \xi) - \nabla f(\hat{x}), y - \hat{x} \rangle \leq \varphi(y) - \varphi(\hat{x}) \quad \forall y \in \text{dom } \varphi \cap \text{dom } f.$$

According to Lemma 2.2, this is equivalent to  $\hat{x} = \text{Prox}_\varphi^f(\nabla f(\hat{x}) - \xi)$  for some  $\xi \in B\hat{x}$  which, in turn, is equivalent to  $\hat{x} \in \text{Prox}_\varphi^f(\nabla f(\hat{x}) - B\hat{x})$ , i.e., to the condition that  $\hat{x}$  is a fixed point of  $\text{Prox}_\varphi^f(\nabla f - B)$ .  $\square$

Let  $B : X \rightarrow 2^{X^*}$  be an operator and suppose that the closed convex subset  $C$  of  $X$  satisfies

$$(2.17) \quad \emptyset \neq \text{dom } B \cap C \cap \text{int dom } f \quad \text{and} \quad \text{ran } (\nabla f - B) \subseteq \text{int dom } f^*.$$

Note that if  $\varphi := \iota_C$ , then the variational inequality (2.14) is precisely the following problem, which we call *classical variational inequality*:

$$(2.18) \quad \text{Find } x \in C \cap \text{int dom } f \text{ such that}$$

$$\exists \xi \in Bx : [\langle \xi, y - x \rangle \geq 0 \quad \forall y \in C \cap \text{dom } f].$$

Applying Lemma 2.4 and (1.3) in this case, we rediscover the following known result (cf. [2]).

LEMMA 2.5. *Suppose that the condition (2.17) is satisfied and that  $\hat{x} \in C \cap \text{int dom } f$ . Then the following statements are equivalent:*

- (a) *The vector  $\hat{x}$  is a solution of the classical variational inequality (2.18);*
- (b) *the vector  $\hat{x}$  is a fixed point of the operator  $\text{Proj}_C^f(\nabla f - B)$ ;*
- (c) *the vector  $\hat{x}$  is a zero for the operator  $V[B; C; f] : X \rightarrow 2^{X^*}$  given by*

$$(2.19) \quad V[B; C; f] := \nabla f - \nabla f \circ \text{Proj}_C^f(\nabla f - B).$$

Lemmas 2.4 and 2.5 provided the initial motivation for this research. Observe that Lemma 2.4 reduces the problem of finding a solution for the classical variational inequality (2.18) to the problem of finding a fixed point for the operator  $\text{Proj}_C^f(\nabla f - B)$ . It is well known that, in many instances, by iterating an operator starting from initial points located in its domain, one produces sequences which converge to fixed points of the operator. This suggests that for solving the classical variational inequality (2.18), we would have to produce sequences defined by the iterative rule

$$(2.20) \quad x^{k+1} \in \text{Proj}_C^f(\nabla f(x^k) - Bx^k)$$

in the hope that such sequences will converge to fixed points of  $\text{Proj}_C^f(\nabla f - B)$ . Note that any fixed point of the operator  $\text{Proj}_C^f(\nabla f - B)$  is a zero for the operator  $V[B; C; f]$  given by (2.19), and conversely. According to (2.9), we have that

$$\begin{aligned}
 (2.21) \quad \text{Proj}_C^f(\nabla f - V[B; C; f]) &= \text{Proj}_C^f \circ \nabla f \circ \text{Proj}_C^f \circ (\nabla f - B) \\
 &= \text{proj}_C^f \circ \text{proj}_C^f \circ (\nabla f)^{-1} \circ (\nabla f - B) \\
 &= \text{proj}_C^f \circ \nabla f^* \circ (\nabla f - B) \\
 &= \text{Proj}_C^f(\nabla f - B).
 \end{aligned}$$

Thus, we are naturally led to the question of whether, and in which conditions, the sequences generated according to the rule  $x^{k+1} \in \text{Proj}_C^f(\nabla f(x^k) - V[B; C; f]x^k)$ , which are the same (see (2.21)) as the sequences generated according to rule (2.20), converge to zeros of the operator  $V[B; C; f]$ . This is, in fact, a particular instance of the more general question of whether, and in which conditions, the proximal-projection method (1.7) approximates zeros of an operator  $A$  (in our specific case  $A := V[B; C; f]$ ), provided that such zeros exist. In what follows we present answers to this question. It is interesting to observe that by focusing, in our convergence analysis, on conditions concerning the operator  $V[B; C; f]$  instead of  $B$  we do not mean that computing values of  $\text{Proj}_C^f(\nabla f - V[B; C; f])$  is easier than computing values of the same operator via the formula  $\text{Proj}_C^f(\nabla f - B)$ . However, from a theoretical point of view, the operator  $V[B; C; f]$  associated with  $B$  via formula (2.19) may happen to be better conditioned than  $B$  for a convergence analysis of the proximal-projection method. This aspect can be clearly seen after a careful dissection of the considerations which lead to our main convergence results presented in this paper. It should be taken into account that the operator  $B$  may not have zeros in  $C$  even if the operator  $V[B; C; f]$ , associated to  $B$  by (2.19), does. For example, take  $X = \mathbb{R}$ ,  $f(x) = \frac{1}{2}x^2$ ,  $C = [1, 2]$ , and  $Bx = x$  for all  $x \in X$ . Then  $V[B; C; f]x = x - 1$  vanishes at  $x = 1$ , but  $B$  does not have any zero in  $C$ .

**3.  $D_f$ -nonexpansivity poles,  $D_f$ -inverse strong monotonicity, and  $D_f$ -firmness of operators in Banach spaces.** In this section we introduce the notion of  $D_f$ -inverse strong monotonicity for operators from  $X$  to  $2^{X^*}$ . We clarify how this notion is related to the notions of  $D_f$ -nonexpansivity pole introduced in [33] and  $D_f$ -firm operator introduced in [21]. Using these relations we show that the proximal point method with Bregman distances can be seen as a particular instance of the proximal-projection method applied to a  $D_f$ -inverse strongly monotone operator.

**3.1.  $D_f$ -nonexpansivity poles.** In what follows, we associate with the function  $f$  described in section 1, and with any operator  $A : X \rightarrow 2^{X^*}$ , the operator  $A^f : X \rightarrow 2^X$  given by

$$(3.1) \quad A^f := \nabla f^* \circ (\nabla f - A).$$

We call this operator the  $D_f$ -antiresolvent of  $A$ . Observe that

$$(3.2) \quad \text{dom } A^f \subseteq \text{dom } A \cap \text{int dom } f \quad \text{and} \quad \text{ran } A^f \subseteq \text{int dom } f,$$

and that, if  $x \in \text{int dom } f$ , then  $0^* \in Ax$  if and only if  $x \in \text{Fix } A^f$ , where  $\text{Fix } A^f$  denotes the set of fixed points of  $A^f$ . Therefore, the (possibly empty) set of solutions of Problem 1.1 situated in  $\text{int dom } f$ , denoted  $\mathcal{S}_f(A, C)$ , is exactly

$$(3.3) \quad \mathcal{S}_f(A, C) = C \cap \text{Fix } A^f.$$

We are going to prove that, for operators  $A$  which are  $D_f$ -inverse strongly monotone, the set  $\mathcal{S}_f(A, C)$  is exactly the set of  $D_f$ -nonexpansivity poles of  $A^f$  over the set  $C$ , and this fact will be later used in our convergence analysis of the proximal-projection method. To this end, recall the following notion.

DEFINITION 3.1 (cf. [33]). *Let  $T : X \rightarrow 2^X$  be an operator and let  $Y$  be a subset of  $X$  such that*

$$(3.4) \quad \emptyset \neq T(Y \cap \text{int dom } f) \subseteq \text{int dom } f.$$

*The vector  $z \in X$  is called a  $D_f$ -nonexpansivity pole of  $T$  over  $Y$  if the following conditions are satisfied:*

$$(3.5) \quad z \in Y \cap \text{int dom } f,$$

$$(3.6) \quad (x \in Y \cap \text{int dom } f \quad \text{and} \quad u \in Tx) \Rightarrow \langle \nabla f(x) - \nabla f(u), z - u \rangle \leq 0.$$

*We denote by  $\text{Nexp}_Y^f T$  the set of  $D_f$ -nonexpansivity poles of  $T$  over  $Y$ .*

Operators having  $D_f$ -nonexpansivity poles were termed *totally nonexpansive operators* in [33]. Operators  $T$  such that  $\text{ran } T \subseteq \text{dom } T = \text{int dom } f$  and having  $\text{Nexp}_{\text{int dom } f}^f T \supseteq \text{Fix } T$  were called  *$\mathcal{B}$ -class operators* in [20] and [21].  $\mathcal{B}$ -class operators necessarily have  $\text{Nexp}_{\text{int dom } f}^f T = \text{Fix } T$  (cf. [21, Proposition 3.3]). However, not every operator having  $D_f$ -nonexpansivity poles over some subset  $Y$  of  $X$  is  $\mathcal{B}$ -class. For example, the operator  $Tx = \{x^2\}$  when  $X = \mathbb{R}$ ,  $f = \frac{1}{2}|\cdot|^2$ , and  $Y = \text{int dom } f = \mathbb{R}$  has  $z \in \text{Nexp}_{\text{int dom } f}^f T$  if and only if

$$x(1-x)(z-x^2) \leq 0 \quad \forall x \in \mathbb{R},$$

and this last inequality cannot hold because, irrespective of  $z$ , we have  $\lim_{x \rightarrow \infty} x(1-x)(z-x^2) = \infty$ . Hence,  $\text{Nexp}_{\text{int dom } f}^f T = \emptyset$ . In spite of that,  $\text{Fix } T = \{0, 1\}$  and, if  $Y = [0, 1]$ , then (3.6) holds for  $z = 0$  only, i.e.,  $\text{Nexp}_{[0,1]}^f T = \{0\} \neq \text{Fix } T$ .

The following lemma summarizes several properties of operators having nonexpansivity poles which are used in this work.

LEMMA 3.2. *Let the operator  $T : X \rightarrow 2^X$  and the set  $Y \subseteq X$  be such that condition (3.4) holds. Then the following statements are true:*

(a) *The (possibly empty) set  $\text{Nexp}_Y^f T$  is convex and closed when  $Y \subseteq \text{int dom } f$  is convex and closed;*

(b) *a vector  $z \in Y \cap \text{int dom } f$  is a  $D_f$ -nonexpansivity pole of  $T$  over  $Y$  if and only if*

$$(3.7) \quad (x \in Y \cap \text{int dom } f \quad \text{and} \quad u \in Tx) \Rightarrow D_f(z, u) + D_f(u, x) \leq D_f(z, x).$$

(c)  *$\text{Nexp}_Y^f T \subseteq \text{Fix } T$  and  $T$  is single-valued at any  $D_f$ -nonexpansivity pole.*

*Proof.* Statement (a) results from the fact that the function  $z \rightarrow \langle \nabla f(x) - \nabla f(u), z - u \rangle$  is linear and continuous. Statement (b) follows from (3.6) and (2.10). Now, by taking in (3.7)  $x = z \in \text{Nexp}_Y^f T$ , one gets

$$(3.8) \quad D_f(z, u) = D_f(u, z) = 0 \quad \forall u \in Tx.$$



By (3.4), if  $u \in Tz$ , then  $u \in \text{int dom } f$ . The function  $f$  being essentially strictly convex is strictly convex on  $\text{int dom } f$ . Hence, the equalities in (3.8) cannot hold unless  $u = z$  (cf. [33, Proposition 1.1.4]). In other words, we have the following implication:

$$(3.9) \quad z \in \text{Nexp}_Y^f T \Rightarrow Tz = \{z\}.$$

It shows that  $\text{Nexp}_Y^f T \subseteq \text{Fix } T$  and that  $T$  is single-valued at  $D_f$ -nonexpansivity poles.  $\square$

**3.2.  $D_f$ -inverse strong monotonicity and  $D_f$ -firmness.** The notion of a  $D_f$ -inverse strongly monotone operator, introduced in this section, and the notion of a  $D_f$ -firm operator, originally introduced in [21, Definition 3.4], are generalizations of the notion of firmly nonexpansive operator in a Hilbert space. Recall (cf. [48, pp. 41–42]) that if  $X$  is a Hilbert space (which we always identify with its dual  $X^*$ ), then an operator  $A : X \rightarrow X$  is firmly nonexpansive on a subset  $Y$  of  $X$  if and only if

$$(3.10) \quad \langle Ax - Ay, x - y \rangle \geq \|Ax - Ay\|^2 \quad \forall x, y \in Y.$$

It can be easily seen from the definitions given below that if  $X$  is a Hilbert space and if  $f = \frac{1}{2} \|\cdot\|^2$ , then the operator  $A$  is  $D_f$ -inverse strongly monotone on its domain if and only if it is firmly nonexpansive on its domain, and this happens if and only if  $A$  is  $D_f$ -firm.

It is interesting to note that if, instead of the function  $f = \frac{1}{2} \|\cdot\|^2$ , we provide the Hilbert space  $X$  with the Legendre function  $f_q := \frac{1}{2q} \|\cdot\|^2$  for some positive real number  $q$ , then the inequality (3.12) below with  $f_q$  instead of  $f$  becomes the usual  $q$ -inverse strong monotonicity condition (see [49, p. 256])

$$\langle Ax - Ay, x - y \rangle \geq q \|Ax - Ay\|^2 \quad \forall x, y \in X,$$

a generalization of the firm nonexpansivity which describes operators  $A : X \rightarrow X$  whose inverses  $A^{-1}$  are strongly monotone. This explains why we have chosen the term “ $D_f$ -inverse strongly monotone” in order to designate operators satisfying the requirements of the next definition.

Returning to the general context in which  $X$  and  $f$  are as described in section 1, we introduce the following notion.

**DEFINITION 3.3.** *Let  $Y$  be a subset of the space  $X$ . The operator  $A : X \rightarrow 2^{X^*}$  is called  $D_f$ -inverse strongly monotone on the set  $Y$  if*

$$(3.11) \quad Y \cap (\text{dom } A) \cap (\text{int dom } f) \neq \emptyset$$

and

$$(3.12) \quad \left. \begin{array}{l} x, y \in Y \cap \text{int dom } f \\ \xi \in Ax \text{ and } \eta \in Ay \end{array} \right\} \Rightarrow \langle \xi - \eta, \nabla f^*(\nabla f(x) - \xi) - \nabla f^*(\nabla f(y) - \eta) \rangle \geq 0.$$

Operators satisfying a somewhat less restrictive condition than (3.12) were studied in [38, section 5] under the name of inverse-monotone operators relative to  $f$ .

In our further considerations it is important to keep in mind that, in general, an operator  $A$  (even in a Hilbert space, provided that  $f$  is not the function  $f = \frac{1}{2} \|\cdot\|^2$ ) does not have to satisfy (3.10) in order to be  $D_f$ -inverse strongly monotone on  $Y$ . For instance, if the function  $f$  has a minimizer in  $\text{int dom } f$  (i.e., if the equation

$\nabla f(x) = 0$  has a solution), and if  $\alpha \in (0, 1)$ , then the operator  $A = \alpha \nabla f$  is  $D_f$ -inverse strongly monotone on  $Y = \text{dom } A = \text{int dom } f$  without necessarily satisfying (3.10) on  $Y = \text{int dom } f$ . Indeed, in this case, if  $x \in \text{int dom } f$ , then  $\{\nabla f(x), 0^*\} \subset \text{ran } \nabla f = \text{int dom } f^*$  and, due to the convexity of  $\text{int dom } f^*$ , we have that

$$(1 - \alpha) \nabla f(x) = (1 - \alpha) \nabla f(x) + \alpha 0^* \in \text{int dom } f^* = \text{dom } \nabla f^*,$$

and, consequently, the operator

$$A^f x = \nabla f^*((1 - \alpha) \nabla f(x))$$

has  $\text{dom } A^f = \text{int dom } f$ . If  $x, y \in \text{int dom } f$ , and if  $\beta = 1 - \alpha$ , then, by the monotonicity of  $\nabla f^*$ , we deduce that

$$\langle Ax - Ay, \nabla f^*(\nabla f(x) - Ax) - \nabla f^*(\nabla f(y) - Ay) \rangle$$

$$= \alpha \beta^{-1} \langle \beta \nabla f(x) - \beta \nabla f(y), \nabla f^*(\beta \nabla f(x)) - \nabla f^*(\beta \nabla f(y)) \rangle \geq 0;$$

i.e., the operator  $A = \alpha \nabla f$  is  $D_f$ -inverse strongly monotone on  $\text{int dom } f$ . However, the operator  $A = \alpha \nabla f$  does not have to be firmly nonexpansive on its domain even if  $X$  is a Hilbert space. For example, take the considerations above  $X$  to be any nonzero Hilbert space and  $f = \frac{1}{3} \|\cdot\|^3$ . Then,  $\nabla f(x) = \|x\|x$  and, for  $y = 0$ , (3.10) reduces to  $\|x\| \leq \alpha$ , a relation that does not hold for any  $x \in X$ .

We are going to show that there are strong connections between the  $D_f$ -inverse strong monotonicity of the operator  $A$  and the  $D_f$ -firmness of its  $D_f$ -antiresolvent  $A^f$ . For this purpose we recall the following.

DEFINITION 3.4 (cf. [21, Definition 3.4]). *An operator  $T : X \rightarrow 2^X$  is called  $D_f$ -firm if it satisfies the conditions*

$$(3.13) \quad \emptyset \neq \text{dom } T \cup \text{ran } T \subseteq \text{int dom } f$$

and

$$(3.14) \quad u \in Tx \quad \text{and} \quad v \in Ty \Rightarrow \langle \nabla f(u) - \nabla f(v), u - v \rangle \leq \langle \nabla f(x) - \nabla f(y), u - v \rangle.$$

We start with the following result which summarizes some basic properties of  $D_f$ -inverse strongly monotone operators.

LEMMA 3.5. *Let  $A : X \rightarrow 2^{X^*}$  be an operator and let  $Y$  be a subset of  $X$  which satisfies (3.11). The following statements are true:*

(a) *The operator  $A$  is  $D_f$ -inverse strongly monotone on  $Y$  if and only if it satisfies the following condition for any  $x, y \in Y \cap \text{int dom } f$ :*

$$\left. \begin{array}{l} u \in A^f x \\ v \in A^f y \end{array} \right\} \Rightarrow D_f(u, v) + D_f(v, u) + D_f(u, x) + D_f(v, y) \leq D_f(v, x) + D_f(u, y);$$

(b) *if  $A$  is  $D_f$ -inverse strongly monotone on  $Y$ , then  $A$  is  $D_f$ -inverse strongly monotone on any subset of  $Y$  which intersects  $\text{dom } A \cap \text{int dom } f$ ;*

(c) *the operator  $A$  is  $D_f$ -inverse strongly monotone on its domain if and only if its  $D_f$ -antiresolvent  $A^f$  is  $D_f$ -firm;*

(d) *if  $A$  is  $D_f$ -inverse strongly monotone on its domain, then  $A^f$  is single valued and all its fixed points are  $D_f$ -nonexpansivity poles on  $\text{int dom } f$ .*

*Proof.* Statements (a), (b), and (c) result from (3.1), (3.2), (3.13), (3.14), and (2.10). The single valuedness of  $A^f$  in statement (d) is a consequence of (c) and of [21, Proposition 3.5(iii)]. Letting  $y = z \in \text{Fix } A^f$  in the inequality of (a), and taking into account the single valuedness of  $A^f$ , one obtains that  $z$  satisfies (3.7) for  $T = A^f$ .  $\square$

Whenever the operator  $A$  involved in Problem 1.1 is  $D_f$ -inverse strongly monotone on the domain  $C$  of the problem, we have

$$(3.15) \quad \mathcal{S}_f(A, C) = C \cap \text{Fix } A^f = \text{Nexp}_C^f A^f.$$

This immediately follows from the next result.

LEMMA 3.6. *The following statements are true:*

- (a) *If  $z \in \text{Nexp}_C^f A^f$ , then  $Az = \{0^*\}$ ;*
- (b) *if the operator  $A$  is  $D_f$ -inverse strongly monotone on  $C$  and  $z \in \text{int dom } f$  is a solution of Problem 1.1, then  $z \in \text{Nexp}_C^f A^f$ .*

*Proof.* Statement (a) results from (3.1) and Lemma 3.2(c). In order to prove (b), note that for any  $x \in C \cap \text{int dom } f$  and  $y \in A^f x$ , we have

$$D_f(z, x) - D_f(z, y) = D_f(y, x) - \langle \nabla f(x) - \nabla f(y), z - y \rangle$$

and  $y = \nabla f^*(\nabla f(x) - \xi)$  for some  $\xi \in Ax$ . Thus,  $\nabla f(y) = \nabla f(x) - \xi$  and

$$(3.16) \quad D_f(z, x) - D_f(z, y) = D_f(y, x) + \langle 0^* - \xi, z - y \rangle.$$

If  $z$  is a solution of Problem 1.1, then  $z \in C$ ,  $0^* \in Az$ , and  $z = \nabla f^*(\nabla f(z) - 0^*)$ . As a consequence,

$$(3.17) \quad \langle 0^* - \xi, z - y \rangle = \langle 0^* - \xi, \nabla f^*(\nabla f(z) - 0^*) - \nabla f^*(\nabla f(x) - \xi) \rangle.$$

Since  $A$  is  $D_f$ -inverse strongly monotone on  $C$ , it results that the right-hand side of (3.17) is nonnegative (see (3.12)) for any  $x \in C \cap \text{int dom } f$ . Hence, by (3.16) and (3.17), the inequality in (3.7) results, and it shows that  $z \in \text{Nexp}_C^f A^f$ .  $\square$

The class of operators which are  $D_f$ -inverse strongly monotone contains some meaningful operators. Among them are all operators  $A[T] : X \rightarrow 2^{X^*}$  given by

$$(3.18) \quad A[T] = \nabla f - \nabla f \circ T,$$

where  $T : X \rightarrow 2^X$  is a  $D_f$ -firm operator. In particular,  $A[T]$  is  $D_f$ -inverse strongly monotone when  $T = B_f$ , where  $B_f : X \rightarrow 2^X$  is the  $D_f$ -resolvent (cf. [21]) of a monotone operator  $B : X \rightarrow 2^{X^*}$ , i.e.,

$$(3.19) \quad B_f := (\nabla f + B)^{-1} \circ \nabla f.$$

According to [21, Proposition 3.8], the operator  $T = B_f$  satisfies the condition (3.13). These facts are summarized in the following lemma.

LEMMA 3.7. *Let  $T : X \rightarrow 2^X$  be an operator which satisfies (3.13). Then the following statements are true:*

- (a)  *$\text{dom } A[T] = \text{dom } T$ , and its antiresolvent is  $(A[T])^f = T$ ;*
- (b) *the operator  $T$  is  $D_f$ -firm if and only if the operator  $A[T] : X \rightarrow 2^{X^*}$  defined by (3.18) is  $D_f$ -inverse strongly monotone on its domain;*
- (c) *if  $B : X \rightarrow 2^{X^*}$  is a monotone operator with  $\text{dom } B \cap \text{int dom } f \neq \emptyset$ , then  $B_f$  is  $D_f$ -firm, single-valued on its domain,*

$$(3.20) \quad A[B_f] = \nabla f - \nabla f \circ (\nabla f + B)^{-1} \circ \nabla f,$$

and  $A[B_f]$  is  $D_f$ -inverse strongly monotone on its domain.

*Proof.* Statement (a) results from (3.1) and (3.18). To prove (b), observe that for any  $x, y \in \text{dom } T$ , for any  $\xi \in A[T]x$ , and for any  $\eta \in A[T]y$ , we have

$$(3.21) \quad \xi = \nabla f(x) - \nabla f(u) \quad \text{and} \quad \eta = \nabla f(y) - \nabla f(v)$$

for some  $u \in Tx$  and for some  $v \in Ty$ . Therefore,

$$(3.22) \quad \begin{aligned} &\langle \xi - \eta, \nabla f^*(\nabla f(x) - \xi) - \nabla f^*(\nabla f(y) - \eta) \rangle \\ &= \langle \nabla f(x) - \nabla f(y), u - v \rangle - \langle \nabla f(u) - \nabla f(v), u - v \rangle. \end{aligned}$$

If  $T$  is  $D_f$ -firm, then the right-hand side of (3.22) is nonnegative and this implies that  $A[T]$  is  $D_f$ -inverse strongly monotone on  $\text{dom } T$ . Conversely, suppose that  $A[T]$  is  $D_f$ -inverse strongly monotone on  $\text{dom } T$ . If  $u \in Tx$  and  $v \in Ty$ , then the vectors  $\xi$  and  $\eta$  given by (3.21) satisfy (3.22) and the left-hand side of this equality is nonnegative. Hence,  $T$  is  $D_f$ -firm. This proves (b). In order to prove (c) recall that, since  $B$  is monotone, its resolvent,  $B_f$ , is necessarily  $D_f$ -firm and single-valued on its domain (cf. [21, Proposition 3.8]) and, as noted above, it satisfies (3.13). Thus, according to (b), the operator  $A[B_f]$  is  $D_f$ -inverse strongly monotone on its domain.  $\square$

**3.3. Connection between the proximal-projection method and the proximal point method.** Lemma 3.7 helps to establish a connection between the proximal-projection method and the *proximal point method (with Bregman distances)*. The proximal point method we are referring to in this paper is the iterative procedure which, in our setting, can be described by

$$(3.23) \quad x^0 \in \text{dom } B_f \quad \text{and} \quad x^{k+1} = B_f(x^k) \quad \forall k \in \mathbb{N},$$

where  $B : X \rightarrow 2^{X^*}$  is a monotone operator with  $\text{dom } B \cap \text{int dom } f \neq \emptyset$ . Its well definedness is guaranteed when

$$(3.24) \quad \emptyset \neq \text{ran } B_f \subseteq \text{dom } B_f.$$

To ensure that the inclusion in this condition holds, it is sufficient to make sure that  $\text{dom } B_f = X$ . This implicitly happens when one considers the classical proximal point method (see [70]) where  $X$  is a Hilbert space,  $f = \frac{1}{2} \|\cdot\|^2$ , and  $B$  is presumed to be maximal monotone. Alternative conditions which imply that  $\text{dom } B_f = X$  when  $B$  is maximal monotone are presented in [31] in a more general setting. In particular, those conditions hold if  $X$  is a uniformly convex and uniformly smooth Banach space,  $B$  is maximal monotone, and  $f = \frac{1}{2} \|\cdot\|^2$ .

A reason for which maximal monotonicity of  $B$  is a commonly used condition for ensuring well definedness of the proximal point method is that, if  $B$  is maximal monotone and  $f = \frac{1}{2} \|\cdot\|^2$ , then  $\nabla f + B$  and  $\nabla f^*$  are surjective and, thus,  $\nabla f^* \circ (\nabla f + B)$  is surjective too, that is,  $\text{dom } B_f = \text{ran } [\nabla f^* \circ (\nabla f + B)] = X$ . An important observation, for which we are indebted to a referee, shows that in cases when  $B$  is maximal monotone with  $0^* \in \text{ran } B$  and  $f$  is a Legendre function (as presumed in the setting of this work), condition (3.24) is satisfied whenever we have that (i)  $\text{dom } B \subseteq \text{int dom } f$  or (ii)  $B$  is  $3^*$ -monotone and  $\text{dom } B \cap \text{int dom } f \neq \emptyset$ . That is so because, if (i) or (ii) is true, then, according to [21, Corollary 3.14], the operator  $B_f$  is  $\mathcal{B}$ -class, and this implies that (3.24) holds by the definition of  $\mathcal{B}$ -class operators (see [21, Definition 3.1]).

Well definedness of the proximal point method can be sometimes ensured for operators  $B$  which are monotone without being maximal monotone. In such cases, it is interesting to know whether the proximal point method preserves the convergence properties which make it so useful in applications requiring finding zeros of maximal monotone operators. Here is an example of a monotone operator which is not maximal and for which (3.24) holds (in spite of the fact that  $\nabla f + B$  is not surjective). Take  $X = \mathbb{R}$ ,  $f = \frac{1}{2}|\cdot|^2$  and let  $B : X \rightarrow 2^{X^*}$  be given by  $Bx = \{0\}$  if  $x \leq 0$ , and  $Bx = \emptyset$  if  $x > 0$ . The operator  $B$  is monotone, but it is not maximal monotone since the operator defined by  $B'x = \{0\}$  for all  $x \in X$  is a proper monotone extension of  $B$ . Obviously, in this case  $\nabla f$  is the identity,  $(\nabla f + B)x = \{x\}$  if  $x \leq 0$ ,  $(\nabla f + B)x = \emptyset$  if  $x > 0$ , and, therefore,  $\nabla f + B$  is not surjective. However,  $B_f = (\nabla f + B)^{-1} = \nabla f + B$  and, hence,  $\text{ran } B_f = \text{dom } B_f = (-\infty, 0]$ , showing that (3.24) is satisfied, that is, the proximal point algorithm is well defined.

The next result establishes the connection between the proximal-projection method and the proximal point method. It requires that  $\text{dom } B_f$  should be convex and closed. This necessarily happens if  $\nabla f + B$  is surjective and  $\text{dom } f = X$ . However,  $\text{dom } B_f$  may happen to be convex and closed even if  $\nabla f + B$  is not surjective, as one can see from the example above. Other instances in which  $\text{dom } B_f$  is convex and closed are described in the remarks preceding Corollary 4.9 as well as in the body of that corollary.

**LEMMA 3.8.** *Let  $B : X \rightarrow 2^{X^*}$  be a monotone operator such that  $\text{dom } B_f$  is convex and closed and suppose that (3.24) is satisfied. Then the proximal point method (3.23) is exactly the proximal-projection method applied to the  $D_f$ -inverse strongly monotone operator  $A[B_f]$  with  $C_k = X$  for all  $k \in \mathbb{N}$ .*

*Proof.* Observe that, by (3.20), we have

$$\begin{aligned} \text{Proj}_{\text{dom } B_f}^f (\nabla f - A[B_f]) &= \text{Proj}_{\text{dom } B_f}^f \left[ \nabla f \circ (\nabla f + B)^{-1} \circ \nabla f \right] \\ &= \text{proj}_{\text{dom } B_f}^f \left[ (\nabla f + B)^{-1} \circ \nabla f \right] = B_f, \end{aligned}$$

where the last equality result holds because (3.24) is satisfied. This shows that the proximal point method and the proximal-projection method are overlapping when one takes  $A = A[B_f]$  and  $C_k = X$  for all  $k \in \mathbb{N}$  in (1.7).  $\square$

**4. Convergence analysis of the proximal-projection method.** In this section we present a convergence theorem for the proximal-projection method in reflexive Banach spaces. Our convergence analysis is based on a generalization of Lemma 5.7 in [38] which, in turn, is a generalization of a result also known as the Browder's demiclosedness principle [29, Theorem 8.4]. Throughout this section we assume that the function  $f$  and the Banach space  $X$  are as described in section 1.

**4.1. A generalization of Browder's demiclosedness principle.** The demiclosedness principle of Browder says that if  $X$  is a Banach space and if  $T : Y \rightarrow Y$  is a nonexpansive mapping on the nonempty closed convex subset  $Y$  of  $X$ , then for any sequence  $\{z^k\}_{k \in \mathbb{N}} \subseteq Y$  which is weakly convergent and has  $\lim_{k \rightarrow \infty} \|Tz^k - z^k\| = 0$ , the vector  $z = w - \lim_{k \rightarrow \infty} z^k$  is necessarily a fixed point of  $T$ . In [38, Lemma 5.7] a similar result was shown to hold for operators  $T : X \rightarrow X$  which are nonexpansive relative to  $f$  (in the sense given to this term in [39]), i.e., such that

$$(4.1) \quad D_f(Tx, Ty) \leq D_f(x, y) \quad \forall x, y \in Y,$$

provided that  $\text{dom } f = \text{dom } \nabla f = X$  and that  $f$  is not only Legendre but also totally convex (see [33]) and bounded on bounded sets, while  $T$  satisfies  $\lim_{k \rightarrow \infty} D_f(Tz^k, z^k) = 0$ .

Our current generalization of Browder's demiclosedness principle concerns set-valued operators  $T$  satisfying a somehow less stringent nonexpansivity condition than (4.1) with respect to a function  $f$  subjected to weaker requirements than those involved in [38, Lemma 5.7]. In what follows we use the following notion which generalizes that of a nonexpansive operator relative to  $f$ .

**DEFINITION 4.1.** *The operator  $T : X \rightarrow 2^X$  is said to be  $D_f$ -nonexpansive if it satisfies (3.13) and for any  $x \in \text{dom } T$  there exists  $u \in Tx$  such that*

$$(4.2) \quad (\forall y \in \text{dom } T) : [v \in Ty \Rightarrow D_f(v, u) \leq D_f(y, x)].$$

In what follows (see Theorem 4.7 below) we will be interested in operators whose antiresolvents are simultaneously  $D_f$ -firm and  $D_f$ -nonexpansive. It should be noted that the notions of  $D_f$ -nonexpansivity and  $D_f$ -firmness are not equivalent, although some operators may have both properties. If  $X$  is a Hilbert space provided with  $f = \frac{1}{2} \|\cdot\|^2$ , then it is obvious that any  $D_f$ -firm operator (i.e., any firmly nonexpansive operator) is  $D_f$ -nonexpansive (i.e., nonexpansive). Even in this context, the converse implication does not generally hold. Take, for example, the case where the Hilbert space is  $X = \mathbb{R}$  and  $Ax = 2x$ . Its antiresolvent is  $A^f x = -x$  and it is  $D_f$ -nonexpansive without being  $D_f$ -firm. However, operators whose antiresolvents are simultaneously  $D_f$ -nonexpansive and  $D_f$ -firm are not specific to the setting of Hilbert spaces provided with  $f = \frac{1}{2} \|\cdot\|^2$ . For instance, if  $X = \mathbb{R}$  and  $f(x) = \frac{1}{4}x^4$ , then the antiresolvent  $A^f$  of the operator  $A = \alpha \nabla f$  with  $\alpha \in (0, 1)$  is  $D_f$ -nonexpansive and  $D_f$ -firm at the same time. Indeed, an easy calculation shows that  $f^*(y) = \frac{3}{4}y^{\frac{4}{3}}$ , thus  $\nabla f^*(y) = y^{\frac{1}{3}}$ , and one obtains  $A^f x = (1 - \alpha)^{\frac{1}{3}}x$ . Therefore,

$$D_f(A^f y, A^f x) = (1 - \alpha)^{\frac{4}{3}} D_f(y, x) \leq D_f(y, x) \quad \forall x, y \in X,$$

showing that  $A^f$  is  $D_f$ -nonexpansive. On the other hand, by replacing the data above in relation (3.14) we obtain  $(1 - \alpha)^{\frac{1}{3}} \leq 1$ , which is true since  $\alpha \in (0, 1)$ . This shows that  $A^f$  is also  $D_f$ -firm.

One still may hope that (as happens in the particular situation noted above when  $X$  is a Hilbert space provided with  $f = \frac{1}{2} \|\cdot\|^2$ ) a  $D_f$ -firm operator is always  $D_f$ -nonexpansive. The following example shows that this is not the case.

**Example 4.2** (a  $D_f$ -firm operator which is not  $D_f$ -nonexpansive). Let  $X = \mathbb{R}$  and let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be the Legendre function given by  $f(x) = |x|^{3/2}$ . Take the continuous, single-valued operator  $T : \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$T(x) = \begin{cases} \frac{1}{4} & \text{if } x < \frac{1}{16}, \\ \sqrt{x} & \text{if } x \geq \frac{1}{16}. \end{cases}$$

We first show that  $T$  is  $D_f$ -firm; that is, we verify that for every  $x, y \in \mathbb{R}$  one has

$$(4.3) \quad [f'(T(x)) - f'(T(y))][T(x) - T(y)] \leq [f'(x) - f'(y)][T(x) - T(y)].$$

For symmetry reasons we can assume that  $x > y$ . The case  $x, y < 1/16$  being trivial, we shall consider the following two cases.

*Case 1.*  $x, y \geq 1/16$ . In this case (4.3) reduces to  $[\sqrt[4]{x} - \sqrt[4]{y}][\sqrt{x} - \sqrt{y}] \leq (\sqrt{x} - \sqrt{y})^2$ , which is equivalent to the obviously true inequality  $1 \leq \sqrt[4]{x} + \sqrt[4]{y}$ .

Case 2.  $x \geq 1/16$  and  $y < 1/16$ . In this case we distinguish two subcases as follows:

(i)  $y \geq 0$ . In this situation (4.3) can be rewritten as  $[\sqrt[4]{x} - \sqrt{1/4}][\sqrt{x} - 1/4] \leq [\sqrt{x} - \sqrt{y}][\sqrt{x} - 1/4]$ , which, in turn, is equivalent to  $\sqrt[4]{x} + \sqrt{y} \leq \sqrt{x} + 1/2$ . This last inequality holds since for  $x \geq 1/16$  and  $y < 1/16$  we obviously have  $\sqrt[4]{x} + \sqrt{y} \leq \sqrt[4]{x} + 1/4$ , and it is easy to verify that

$$(4.4) \quad \sqrt[4]{x} + 1/4 \leq \sqrt{x} + 1/2 \quad \forall x \in [0, \infty).$$

(ii)  $y < 0$ . In this case, the inequality (4.3) is equivalent to  $[\sqrt[4]{x} - \sqrt{1/4}][\sqrt{x} - 1/4] \leq [\sqrt{x} + \sqrt{-y}][\sqrt{x} - 1/4]$ , that is, to  $\sqrt[4]{x} - 1/2 \leq \sqrt{x} + \sqrt{-y}$ . This last inequality is true because  $\sqrt{x} + \sqrt{-y} \geq \sqrt{x}$  and, by (4.4), we also have  $\sqrt[4]{x} \leq \sqrt{x} + 1/4 < \sqrt{x} + 1/2$  for every  $x > 0$ .

These show that the operator  $T$  is  $D_f$ -firm. Now we verify that  $T$  is not  $D_f$ -nonexpansive, that is, that there exist two real numbers  $x$  and  $y$  such that

$$(4.5) \quad f(T(y)) - f(T(x)) - f'(T(x))(T(y) - T(x)) > f(y) - f(x) - f'(x)(y - x).$$

Taking  $x = 1/8$  and  $y = 1/16$ , a simple computation shows that the left-hand side of (4.5) equals  $(1/2)^{15/4}[2^{3/4} + 2^{1/2} - 3]$ , while the right-hand side equals  $(1/2)^{13/2}[2^{1/2} + 2 - 3]$ . Therefore, for the given  $x$  and  $y$ , relation (4.5) becomes

$$2^{13/2}[2^{3/4} + 2^{1/2} - 3] > 2^{15/4}[2^{1/2} - 1].$$

Since

$$2^{13/2}[2^{3/4} + 2^{1/2} - 3] \simeq 8.6895$$

and

$$2^{15/4}[2^{1/2} - 1] \simeq 5.5730,$$

we deduce that (4.5) holds.  $\square$

Before proceeding with the presentation of our generalization of Browder's demiclosedness principle, several observations concerning its hypothesis are in order.

*Remark 4.3.* (a) The next result is a proper generalization of Lemma 5.7 in [38] which, in turn, is a proper generalization of Browder's demiclosedness principle. To see this, observe that when the operator  $T : X \rightarrow X$  is  $D_f$ -nonexpansive, it satisfies (4.7) below.

(b) The hypotheses of the next result implicitly require that the space  $X$  should be reflexive. The fact is that a function  $f$  which is lower semicontinuous with  $\text{int dom } f \neq \emptyset$  and uniformly convex on bounded subsets of  $\text{int dom } f$  exists on a Banach space  $X$  only if  $X$  is reflexive (cf. [35, Corollary 4.3] in conjunction with [38, Theorem 2.10(ii)]).

With these facts in mind we now proceed with the presentation of the generalization of Browder's demiclosedness principle.

**PROPOSITION 4.4.** *Suppose that the function  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ . Let  $T : X \rightarrow 2^X$  be an operator satisfying (3.13) and suppose that  $\nabla f$  is bounded on bounded subsets of  $\text{dom } T \cup \text{ran } T$ . If  $\{z^k\}_{k \in \mathbb{N}} \subseteq \text{dom } T$  is a sequence which converges weakly to a vector  $z \in \text{dom } T$  and if, for some sequence  $\{u^k\}_{k \in \mathbb{N}}$  satisfying*

$$(4.6) \quad (\forall k \in \mathbb{N} : u^k \in Tz^k) \quad \text{and} \quad \lim_{k \rightarrow \infty} D_f(u^k, z^k) = 0,$$

there exists  $u \in T(z)$  such that

$$(4.7) \quad \liminf_{k \rightarrow \infty} D_f(u^k, u) \leq \liminf_{k \rightarrow \infty} D_f(z^k, z),$$

then the vector  $z$  is a fixed point of  $T$ .

*Proof.* For any  $x \in \text{int dom } f$ , one has

$$D_f(z^k, x) - D_f(z^k, z) = D_f(z, x) + \langle \nabla f(x) - \nabla f(z), z - z^k \rangle.$$

This implies that

$$(4.8) \quad \liminf_{k \rightarrow \infty} D_f(z^k, x) = D_f(z, x) + \liminf_{k \rightarrow \infty} D_f(z^k, z)$$

because, since  $\{z^k\}_{k \in \mathbb{N}}$  converges weakly to  $z$ , we have

$$\lim_{k \rightarrow \infty} \langle \nabla f(x) - \nabla f(z), z - z^k \rangle = 0.$$

Since the function  $f$  is Legendre, it is strictly convex on  $\text{int dom } f$ . This implies that  $D_f(z, x) > 0$  whenever  $x \neq z$  (cf. [33, Proposition 1.1.4]) and, consequently, by (4.8) we obtain

$$(4.9) \quad x \neq z \Rightarrow \liminf_{k \rightarrow \infty} D_f(z^k, x) > \liminf_{k \rightarrow \infty} D_f(z^k, z).$$

We claim that

$$(4.10) \quad \liminf_{k \rightarrow \infty} D_f(u^k, u) = \liminf_{k \rightarrow \infty} D_f(z^k, u).$$

To prove this claim, observe that

$$(4.11) \quad D_f(u^k, u) = D_f(z^k, u) + [f(u^k) - f(z^k)] - \langle \nabla f(u), u^k - z^k \rangle.$$

The sequence  $\{z^k\}_{k \in \mathbb{N}}$  is bounded because it is weakly convergent. Since the function  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ , it is also sequentially consistent (cf. [38, Theorem 2.10]). Therefore, by (4.6), we deduce that

$$(4.12) \quad \lim_{k \rightarrow \infty} \|u^k - z^k\| = 0.$$

Hence,  $\{u^k\}_{k \in \mathbb{N}}$  is bounded and

$$(4.13) \quad \lim_{k \rightarrow \infty} \langle \nabla f(u), u^k - z^k \rangle = 0.$$

The convexity of  $f$  on  $\text{int dom } f$  implies

$$(4.14) \quad \langle \nabla f(u^k), u^k - z^k \rangle \geq f(u^k) - f(z^k) \geq \langle \nabla f(z^k), u^k - z^k \rangle \quad \forall k \in \mathbb{N}.$$

By hypothesis,  $\nabla f$  is bounded on bounded subsets of  $\text{dom } T \cup \text{ran } T$ . Therefore, the sequences  $\{\nabla f(z^k)\}_{k \in \mathbb{N}}$  and  $\{\nabla f(u^k)\}_{k \in \mathbb{N}}$  are bounded. Thus, by (4.12) and (4.14), we deduce that

$$\lim_{k \rightarrow \infty} [f(u^k) - f(z^k)] = 0.$$

This, combined with (4.11) and (4.13), implies (4.10), and the claim above is proved.



Suppose by contradiction that  $z \notin Tz$ . Then the vector  $u \in Tz$  whose existence is guaranteed by hypothesis has  $u \neq z$  and then, by (4.9), we deduce

$$(4.15) \quad \liminf_{k \rightarrow \infty} D_f(z^k, u) > \liminf_{k \rightarrow \infty} D_f(z^k, z).$$

On the other hand, by (4.7) and (4.10), we have that

$$(4.16) \quad \liminf_{k \rightarrow \infty} D_f(z^k, z) \geq \liminf_{k \rightarrow \infty} D_f(u^k, u) = \liminf_{k \rightarrow \infty} D_f(z^k, u),$$

which contradicts (4.15). This completes the proof.  $\square$

**4.2. A convergence theorem for the proximal-projection algorithm.** At this stage we are in position to consider the question of convergence of the procedure (1.7) towards solutions of Problem 1.1. For this purpose, we recall the following.

**DEFINITION 4.5.** (a) (Cf. [62, p. 519]). *The weak upper limit of the sequence  $\{E_k\}_{k \in \mathbb{N}}$  of subsets of  $X$  is the set denoted  $w\text{-}\overline{\lim}_{k \rightarrow \infty} E_k$  and consisting of all  $x \in X$  such that there exists a subsequence  $\{E_{i_k}\}_{k \in \mathbb{N}}$  of  $\{E_k\}_{k \in \mathbb{N}}$  and a sequence  $\{x^k\}_{k \in \mathbb{N}}$  in  $X$  which converges weakly to  $x$  and has the property that  $x^k \in E_{i_k}$  for each  $k \in \mathbb{N}$ .*

(b) *The operator  $A : X \rightarrow 2^{X^*}$  is sequentially weakly-strongly closed if its graph is sequentially closed in  $X \times X^*$  provided with the (weak)  $\times$  (strong)-topology, that is,*

$$(4.17) \quad \left. \begin{array}{l} \forall k \in \mathbb{N} : \xi^k \in Av^k \\ v^k \rightarrow v \text{ and } \xi^k \rightarrow \xi \end{array} \right\} \Rightarrow \xi \in Av.$$

Among the classes of sequentially weakly-strongly closed operators  $A : X \rightarrow 2^{X^*}$  is the class of maximal monotone operators (see [64]). Also, it is obvious that, if  $X$  has finite dimension, then the class of sequentially weakly-strongly closed operators  $A : X \rightarrow 2^{X^*}$  consists of all operators  $A$  with closed graph. Another class of sequentially weakly-strongly closed operators of special interest in applications (related to Fredholm integral equations; see [52, section 8.6]) is the class of compact linear operators  $A : X \rightarrow X^*$ . Recall (see, for instance, [52, p. 405]) that a linear operator  $T$  between two Banach spaces is called compact if the image through  $T$  of the open unit ball is a relatively compact set. A compact linear operator transforms weakly convergent sequences into strongly convergent sequences (cf. [52, Theorem 8.1.7]). So, if  $A : X \rightarrow X^*$  is a linear compact operator, it is necessarily sequentially weakly-strongly closed.

Before proceeding towards the main result of this paper, the following observations may be of use.

**Remark 4.6.** (a) A question which naturally comes to mind in the process of analyzing our next theorem is whether, given a sequence  $\{E_k\}_{k \in \mathbb{N}}$  of nonempty, closed, convex subsets of  $X$ , the set  $E = w\text{-}\overline{\lim}_{k \rightarrow \infty} E_k$  is necessarily convex. The fact is that, in general, the answer to this question is negative. To see this, take  $X = \mathbb{R}$  and

$$E_k := \begin{cases} \left[ -1 - \frac{1}{k+1}, -1 + \frac{1}{k+1} \right] & \text{if } k \text{ is even,} \\ \left[ 1 - \frac{1}{k+1}, 1 + \frac{1}{k+1} \right] & \text{if } k \text{ is odd.} \end{cases}$$

It is easy to verify that the weak upper limit of this sequence of closed convex sets is the nonconvex set  $\{-1, +1\}$ .

(b) It is meaningful to note that the answer to the question posed above is affirmative when  $E = \overline{\text{w-lim}_{k \rightarrow \infty} E_k}$  is the Mosco limit of  $\{E_k\}_{k \in \mathbb{N}}$ . To make things precise, recall (cf. [62, p. 519]) that a sequence of subsets  $\{E_k\}_{k \in \mathbb{N}}$  of  $X$  is called (Mosco) convergent to  $E \subseteq X$  if

$$\text{s-lim}_{k \rightarrow \infty} \overline{E_k} = E = \overline{\text{w-lim}_{k \rightarrow \infty} E_k},$$

where  $\text{s-lim}_{k \rightarrow \infty} \overline{E_k}$  stands for the collection of limit points of the convergent sequences  $\{v^k\}_{k \in \mathbb{N}}$  in  $X$  such that  $v^k \in E_k$  for all  $k \in \mathbb{N}$ . According to [62, p. 520], if  $E$  is the (Mosco) limit of a sequence  $\{E_k\}_{k \in \mathbb{N}}$  of closed and convex subsets of  $X$ , then  $E$  is necessarily closed and convex too.

(c) An essential part of condition (b) of Theorem 4.7 below is the requirement that the gradient  $\nabla f$  of the Legendre function  $f$  should be bounded on bounded subsets of  $\text{int dom } f$ . It is important to observe that *if the Legendre function  $f$  has this property, then  $\text{dom } f = X$* . Indeed, since  $f$  is essentially smooth, it follows that  $\text{int dom } f \neq \emptyset$  and any sequence  $\{x^k\}_{k \in \mathbb{N}}$  contained in  $\text{int dom } f$  and converging to a point of the boundary of  $\text{int dom } f$  has the property that  $\lim_{k \rightarrow \infty} \|\nabla f(x^k)\|_* = \infty$  (cf. [20, Theorem 5.6]). Now, suppose that  $\{x^k\}_{k \in \mathbb{N}}$  is a convergent sequence contained in  $\text{int dom } f$  and denote by  $x$  its limit. Then the sequence  $\{\nabla f(x^k)\}_{k \in \mathbb{N}}$  is bounded because the sequence  $\{x^k\}_{k \in \mathbb{N}}$  is bounded and  $\nabla f$  is bounded on bounded subsets of  $\text{int dom } f$ . We claim that  $x \in \text{int dom } f$ . Assume by contradiction that  $x \notin \text{int dom } f$ . Then  $x$  belongs to the boundary of  $\text{int dom } f$ . Hence,  $\lim_{k \rightarrow \infty} \|\nabla f(x^k)\|_* = \infty$ , and this contradicts the boundedness of  $\{\nabla f(x^k)\}_{k \in \mathbb{N}}$ . Since  $\text{int dom } f$  contains the limit of any convergent sequence of vectors contained in it, it follows that  $\text{int dom } f$  is, simultaneously, a closed and open set. The space  $X$ , being a Banach space, is necessarily arcways connected and, thus, a connected space (cf. [40, Theorem 10.3.2]). Consequently,  $X$  is the only nonempty subset of  $X$  which is open and closed at the same time (cf. [40, Theorem 10.1.8]), that is,  $\text{int dom } f = X$ .

(d) By remark (c) above, the Legendre function  $f$  has  $\nabla f$  bounded on bounded subsets of  $\text{int dom } f$  only if  $\text{dom } f = X$ . Now, suppose that the Legendre function  $f$  has  $\text{dom } f = X$ . Then, according to [18, Proposition 7.8],  $\nabla f$  is bounded on bounded subsets of  $X$  if and only if  $f$  is bounded on bounded subsets of  $X$ . In the particular case of a space  $X$  of finite dimension, this implies that any Legendre function  $f$  with  $\text{dom } f = X$  has  $\nabla f$  bounded on bounded subsets of  $X$  because, in these circumstances, as being a convex function,  $f$  is continuous on  $X$  and, hence, bounded on bounded subsets of  $X$ . In other words, if  $X$  has finite dimension, then the Legendre function  $f$  has  $\nabla f$  bounded on bounded subsets of  $\text{int dom } f$  if and only if  $\text{dom } f = X$ .

The following theorem establishes the basic convergence properties of the proximal-projection method. It should be observed that, in view of Remark 4.6(d), the hypothesis of point (b) of the theorem implicitly requires that  $\text{dom } f = X$ . It is interesting to note that, in general, this theorem guarantees subsequential weak convergence of the sequences generated by the proximal-projection method to solutions of Problem 1.1. However, the theorem also shows that, in some situations of special practical interest, the proximal-projection method produces sequences which converge weakly to such solutions. This is the case when Problem 1.1 has a unique solution (see (ii1)) and this is the typical situation when one has to solve regularized variational inequalities (as emphasized in Corollary 4.10 below). Problem 1.1 may not have a unique solution, but the sequences generated by the proximal-projection method still may converge weakly to such solutions, provided that the Legendre function  $f$  involved in the procedure has sequentially weak-to-weak continuous gradient  $\nabla f$ —see

(ii2). This condition is obviously satisfied by any Legendre function  $f$  on  $X$  when the space  $X$  has finite dimension. If  $X$  has infinite dimension, then finding Legendre functions with a sequentially weak-to-weak continuous gradient is somehow more challenging. We briefly describe here a class of such functions which is presented in more detail in [36]. To this end, let  $G : X \rightarrow X^*$  be a linear continuous operator and let  $G^*$  be its adjoint, that is, the unique linear continuous operator  $G^* : X(\simeq X^{**}) \rightarrow X^*$  such that  $\langle Gx, y \rangle = \langle G^*y, x \rangle$  for all  $x, y \in X$  (see [73, Theorem 4.10]). Define the function  $f_G : X \rightarrow \mathbb{R}$  by  $f_G(x) = \langle Gx, x \rangle$ . It is easy to see that  $\nabla f_G = G + G^*$ . Consequently,  $\nabla f_G$  is linear and continuous and, therefore, weak-to-weak continuous (and, thus, satisfies (ii2)). If, in addition, the operator  $G$  is positive definite and has the property (which necessarily holds when  $\nabla f_G$  is onto and, in particular, when  $G$  is strongly monotone) that  $\lim_{\|x\| \rightarrow \infty} \|Gx + G^*x\|_* = +\infty$ , then  $f_G$  is a Legendre function and, hence, it is a Legendre function with a sequentially weak-to-weak continuous gradient. The functions  $f_G$  presented here have linear gradients. However, there are Legendre functions whose sequentially weak-to-weak continuous gradient is not linear. For instance, this is the case of the function  $f : \ell^p \rightarrow \mathbb{R}$  defined by  $f(x) = \frac{1}{p} \|x\|^p$  with  $p \in (1, +\infty) \setminus \{2\}$  which has  $\nabla f(x)_j = \|x\|^{p-1} \text{sign}(x_j)$  for all  $j \in \mathbb{N}$ —see [29, Proposition 8.2]. Identifying more classes of Legendre functions with a sequentially weak-to-weak continuous gradient is of interest not only for the implementation of the proximal-projection method, but also for the application of other algorithms (see [33] and the references therein).

**THEOREM 4.7.** *Suppose that the function  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ ,  $\nabla f^*$  is bounded on bounded subsets of  $\nabla f(\text{dom } A)$ , and that, in addition to (1.6) and Assumption 1.2, the subsets  $C_k$  of  $X$  satisfy*

$$(4.18) \quad C = \text{w-}\overline{\lim}_{k \rightarrow \infty} C_k.$$

*If Problem 1.1 has at least one solution, if the operator  $A$  is  $D_f$ -inverse strongly monotone on the set  $Q := \bigcup_{k \in \mathbb{N}} C_k$ , and if at least one of the following two conditions is satisfied:*

(a)  *$\nabla f$  is uniformly continuous on bounded subsets of  $\text{int dom } f$  and  $A$  is sequentially weakly-strongly closed;*

(b)  *$\nabla f$  is bounded on bounded subsets of  $\text{int dom } f$ ,  $A^f$  is  $D_f$ -nonexpansive and  $C \subseteq \text{dom } A$ ,*

*then any sequence  $\{x^k\}_{k \in \mathbb{N}}$  generated by the proximal-projection method (1.7) has the following properties:*

(i) *it is bounded and has weak accumulation points, and any such point is a solution of Problem 1.1;*

(ii) *it converges weakly and its weak limit is solution to Problem 1.1 in each of the following situations;*

(ii1) *if Problem 1.1 has a unique solution;*

(ii2) *if  $\nabla f$  is sequentially weak-to-weak continuous.*

(iii) *If the Banach space  $X$  has finite dimension, then  $\{x^k\}_{k \in \mathbb{N}}$  converges in norm to a solution of Problem 1.1.*

*Proof.* Let  $z \in C$  be a solution of Problem 1.1. Then, clearly  $z \in \text{dom } A$  and, by (1.6), we deduce that  $z \in \text{int dom } f$ . For each  $k \in \mathbb{N}$ , let  $\zeta^k \in Ax^k$  be such that

$$(4.19) \quad x^{k+1} = \text{Proj}_{C_{k+1} \cap \text{dom } A}^f(\nabla f(x^k) - \zeta^k).$$

Denote

$$(4.20) \quad u^k := \nabla f^*(\nabla f(x^k) - \zeta^k).$$

Observe that

$$(4.21) \quad u^k \in A^f x^k \quad \text{and} \quad x^{k+1} = \text{proj}_{C_{k+1} \cap \text{dom } A}^f u^k \quad \forall k \in \mathbb{N}.$$

By hypothesis, the operator  $A$  is  $D_f$ -inverse strongly monotone on the set  $Q$  and  $z$  is a solution of Problem 1.1. Note that, since  $0^* \in Az$  and  $z \in C \subseteq Q$ , Lemma 3.6(b) applies with  $C$  replaced by  $Q$ . It implies that  $z \in \text{Nexp}_Q^f A^f$ . By Lemma 2.1 we have that

$$(4.22) \quad x^k \in C_k \cap \text{dom } A \cap \text{int dom } f \subseteq Q \cap \text{dom } A \cap \text{int dom } f \quad \forall k \in \mathbb{N}.$$

First we prove the following.

*Claim 1.* The sequence  $\{x^k\}_{k \in \mathbb{N}}$  is bounded.

In order to show this, notice that, by applying Lemma 3.2(b) to  $z \in \text{Nexp}_Q^f A^f$  we deduce

$$(4.23) \quad D_f(z, u^k) + D_f(u^k, x^k) \leq D_f(z, x^k) \quad \forall k \in \mathbb{N}.$$

This implies

$$(4.24) \quad D_f(z, u^k) \leq D_f(z, x^k) \quad \forall k \in \mathbb{N}.$$

By Assumption 1.2, we have that  $z \in C \subseteq C_k$  for all  $k \in \mathbb{N}$ . Thus, taking into account (2.10), (4.21), and Lemma 2.2 applied to  $\varphi = \iota_{C_{k+1} \cap \text{dom } A}$ , we obtain that

$$(4.25) \quad D_f(z, x^{k+1}) + D_f(x^{k+1}, u^k) \leq D_f(z, u^k) \quad \forall k \in \mathbb{N}.$$

Combining (4.25) with (4.24) yields

$$(4.26) \quad D_f(z, x^{k+1}) \leq D_f(z, x^k) \quad \forall k \in \mathbb{N},$$

showing that the nonnegative sequence  $\{D_f(z, x^k)\}_{k \in \mathbb{N}}$  is nonincreasing and, therefore, bounded. Let  $\beta$  be an upper bound of  $\{D_f(z, x^k)\}_{k \in \mathbb{N}}$ . According to (1.4), (2.10), and (4.26), we deduce that

$$f^*(\nabla f(x^k)) - \langle \nabla f(x^k), z \rangle + f(z) = W_f(\nabla f(x^k), z) = D_f(z, x^k) \leq \beta \quad \forall k \in \mathbb{N}.$$

This implies that the sequence  $\{\nabla f(x^k)\}_{k \in \mathbb{N}}$  is contained in the sublevel set  $\text{lev}_{\leq}^{\psi}(\beta - f(z))$  of the function  $\psi := f^* - \langle \cdot, z \rangle$ . Since  $z \in \text{int dom } f = \text{int dom } (f^*)^*$ , and since the function  $f^*$  is proper and lower semicontinuous, an application of the Moreau–Rockafellar theorem (see [68, Theorem 7(A)] or [20, Fact 3.1]) shows that  $f^* - \langle \cdot, z \rangle$  is coercive. Consequently, all sublevel sets of  $\psi$  are bounded. Hence, the sequence  $\{\nabla f(x^k)\}_{k \in \mathbb{N}}$  is bounded. By hypothesis,  $\nabla f^*$  is bounded on bounded subsets of  $\nabla f(\text{dom } A)$  and, according to (1.7),  $\{\nabla f(x^k)\}_{k \in \mathbb{N}}$  is contained in  $\nabla f(\text{dom } A)$ . Hence, the sequence  $x^k = \nabla f^*(\nabla f(x^k))$ ,  $k \in \mathbb{N}$ , is bounded. This proves Claim 1.

Now we are going to prove the following.

*Claim 2.* The sequence  $\{x^k\}_{k \in \mathbb{N}}$  has weak accumulation points and any such point is a solution of Problem 1.1.

The space  $X$  being reflexive, there exists a weakly convergent subsequence  $\{x^{i_k}\}_{k \in \mathbb{N}}$  of  $\{x^k\}_{k \in \mathbb{N}}$ . Let  $\bar{x} = \text{w-lim}_{k \rightarrow \infty} x^{i_k}$ . In order to show that  $\bar{x} \in C$ , denote  $y^k = x^{i_k}$  for each  $k \in \mathbb{N}$ . According to (4.22), we have that  $y^k \in C_{i_k}$  for every  $k \in \mathbb{N}$ . By hypothesis,  $\text{w-lim}_{k \rightarrow \infty} C_k = C$ , and this implies that  $\bar{x} \in C$  (see Definition 4.5). It remains to

prove that  $0^* \in A\bar{x}$ . To this end, observe that, according to (2.10), (4.25), and (4.24), we have

$$(4.27) \quad 0 \leq D_f(z, u^k) - D_f(z, x^{k+1}) \leq D_f(z, x^k) - D_f(z, x^{k+1}) \quad \forall k \in \mathbb{N}.$$

As noted above, the sequence  $\{D_f(z, x^k)\}_{k \in \mathbb{N}}$  is nonincreasing and nonnegative and, therefore, it converges. By (4.27), this implies that the sequence  $\{D_f(z, u^k)\}_{k \in \mathbb{N}}$  converges and has the same limit as  $\{D_f(z, x^k)\}_{k \in \mathbb{N}}$ . By (4.23) we also have that

$$(4.28) \quad D_f(u^k, x^k) \leq D_f(z, x^k) - D_f(z, u^k) \quad \forall k \in \mathbb{N},$$

and, thus,

$$(4.29) \quad \lim_{k \rightarrow \infty} D_f(u^k, x^k) = 0.$$

Since  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ , it results that it is sequentially consistent too (cf. [38, Theorem 2.10]). Therefore, the equality (4.29) implies that

$$(4.30) \quad \lim_{k \rightarrow \infty} \|u^k - x^k\| = 0.$$

Now, we distinguish two possible situations. First, suppose that condition (a) is satisfied. Note that, according to (4.20), we have

$$\zeta^k = \nabla f(x^k) - \nabla f(u^k) \quad \forall k \in \mathbb{N}.$$

Since  $\nabla f$  is uniformly continuous on bounded subsets of its domain, we deduce by (4.30) that  $\lim_{k \rightarrow \infty} \zeta^k = 0^*$ . The operator  $A$  being sequentially weakly-strongly closed, this and the fact that  $\{x^{i_k}\}_{k \in \mathbb{N}}$  converges weakly to  $\bar{x}$  imply that  $0^* \in A\bar{x}$ . Hence,  $\bar{x}$  is a solution of Problem 1.1 when condition (a) is satisfied.

Alternatively, suppose that condition (b) is satisfied. Recall that, in this case, we necessarily have  $\text{dom } f = X$  (cf. Remark 4.6(d)). By hypothesis (b), we have that  $C \subseteq \text{dom } A$ . By Assumption 1.2, we have that

$$(\nabla f - A)(C) \subseteq (\nabla f - A)(C_0) \subseteq \text{int dom } f^*,$$

and, as shown above,  $\bar{x} \in C$ . Hence,

$$\emptyset \neq (\nabla f - A)(\bar{x}) \subseteq \text{int dom } f^*,$$

which implies that  $\bar{x} \in \text{dom } A^f$ . From (4.2) written with  $x^{i_k}$  instead of  $y$  and  $u^{i_k}$  instead of  $v$ , we obtain that there exists  $\bar{u} \in A^f \bar{x}$ , such that

$$(4.31) \quad D_f(u^{i_k}, \bar{u}) \leq D_f(x^{i_k}, \bar{x}) \quad \forall k \in \mathbb{N}.$$

This implies

$$(4.32) \quad \liminf_{k \rightarrow \infty} D_f(u^{i_k}, \bar{u}) \leq \liminf_{k \rightarrow \infty} D_f(x^{i_k}, \bar{x}).$$

Proposition 4.4, (4.29), and (4.32) imply that  $\bar{x}$  is a fixed point of  $A^f$ , that is,  $0^* \in A\bar{x}$ . Hence,  $\bar{x}$  is a solution of Problem 1.1 in this case too and Claim 2 is established. Also, this completes the proof of (i).

The fact that if Problem 1.1 has a unique solution, then the sequence  $\{x^k\}_{k \in \mathbb{N}}$  converges weakly to that solution is an immediate consequence of (i). Assume that the function  $f$  has a sequentially weak-to-weak continuous gradient. We show next that, in this case, the sequence  $\{x^k\}_{k \in \mathbb{N}}$  cannot have more than one weak accumulation point. Suppose by contradiction that this is not the case and that  $x'$  and  $x''$  are two different weak accumulation points of  $\{x^k\}_{k \in \mathbb{N}}$ . Let  $\{x^{i_k}\}_{k \in \mathbb{N}}$  and  $\{x^{j_k}\}_{k \in \mathbb{N}}$  be subsequences of  $\{x^k\}_{k \in \mathbb{N}}$  converging weakly to  $x'$  and  $x''$ , respectively. By (i) combined with Lemma 3.6(b) it results that  $\{x', x''\} \subseteq \mathcal{S}_f(A, C) = \text{Nexp}_C^f(A^f)$ . Hence, the inequality (4.26) still holds for any  $z \in \{x', x''\}$ . It implies that the sequences  $\{D_f(x', x^k)\}_{k \in \mathbb{N}}$  and  $\{D_f(x'', x^k)\}_{k \in \mathbb{N}}$  are convergent. Let  $a$  and  $b$  be their respective limits. For any  $k \in \mathbb{N}$  we have

$$D_f(x', x^k) - D_f(x'', x^k) = D_f(x', x'') + \langle \nabla f(x'') - \nabla f(x^k), x' - x'' \rangle$$

because of (2.10). Replacing in this equation  $x^k$  by  $x^{j_k}$  and letting  $k \rightarrow \infty$ , we deduce that

$$(4.33) \quad a - b = D_f(x', x'')$$

because  $\nabla f$  is sequentially weak-to-weak continuous. A similar reasoning with  $x'$  and  $x''$  interchanged shows that

$$b - a = D_f(x'', x').$$

Adding this equality to (4.33) we obtain that  $D_f(x', x'') = 0$ . This cannot happen unless  $x' = x''$  because  $f$ , being a Legendre function, is strictly convex on  $C \cap \text{int dom } f$ . Thus, we reached a contradiction and this completes the proof of (ii). It is clear that (iii) follows from (i) and (ii) since the gradient of any convex function in a finite-dimensional space is continuous on the interior of its domain (see, for instance, [65, Proposition 2.8]). This completes the proof of the theorem.  $\square$

**4.3. Consequences of Theorem 4.7.** If  $X$  is a Hilbert space provided with the function  $f = \frac{1}{2} \|\cdot\|^2$ , then Theorem 4.7 has a somewhat simpler form, and even strong convergence of the sequence generated by the proximal-projection method can sometimes be ensured. Note that in this case the operator  $A : X \rightarrow 2^X$  is  $D_f$ -inverse strongly monotone if and only if it is *firmly nonexpansive* in the sense that

$$(4.34) \quad \left. \begin{array}{l} x, y \in X \\ \xi \in Ax, \eta \in Ay \end{array} \right\} \Rightarrow \langle \xi - \eta, x - y \rangle \geq \|\xi - \eta\|^2.$$

Clearly, if  $A$  has this property, then  $A$  is necessarily single-valued on  $X$ , and the operator  $A^f$  (which is exactly  $I - A$ ) is nonexpansive and, thus,  $D_f$ -nonexpansive. Since in this situation  $\text{Proj}_{C_k}^f$  is exactly the metric projection operator  $\text{Proj}_{C_k}$ , we obtain the following result.

**COROLLARY 4.8.** *Let  $X$  be a Hilbert space. Suppose that  $A : X \rightarrow X$  is a firmly nonexpansive operator (i.e., it satisfies (4.34)). If  $C$  is a nonempty, closed, and convex subset of  $X$ , if  $\{C_k\}_{k \in \mathbb{N}}$  is a sequence of subsets of  $X$  satisfying (4.18) and such that  $C_k \cap \text{dom } A$  is convex and closed and contains  $C$  for each  $k \in \mathbb{N}$ , and if Problem 1.1 has at least one solution, then the sequence  $\{x^k\}_{k \in \mathbb{N}}$  generated according to the rule*

$$(4.35) \quad x^0 \in C_0 \cap \text{dom } A \quad \text{and} \quad x^{k+1} = \text{Proj}_{C_{k+1} \cap \text{dom } A}(x^k - Ax^k)$$

converges weakly to a solution of Problem 1.1. If  $\text{int } \mathcal{S}_f(A, C) \neq \emptyset$ , then  $\{x^k\}_{k \in \mathbb{N}}$  converges strongly.

*Proof.* As noted above, the operator  $A$  satisfying (4.34) is  $D_f$ -inverse strongly monotone and  $D_f$ -nonexpansive. Applying Theorem 4.7(b) to  $f = \frac{1}{2} \|\cdot\|^2$ , which has  $\nabla f = I$  (and, hence, has  $\nabla f$  sequentially weak-to-weak continuous), and taking into account that  $A^f = I - A$ , one deduces that the sequence  $\{x^k\}_{k \in \mathbb{N}}$  converges weakly to a vector in  $\mathcal{S}_f(A, C)$ . The set  $\mathcal{S}_f(A, C) = \text{Nexp}_C^f A^f$  is convex and closed (cf. Lemma 3.2(a)). In the current circumstances, the inequality (4.26) still holds for all  $z \in \mathcal{S}_f(A, C)$ . It is equivalent to the condition

$$\|z - x^{k+1}\| \leq \|z - x^k\| \quad \forall z \in \mathcal{S}_f(A, C), \quad \forall k \in \mathbb{N}.$$

Therefore, one can apply Theorem 2.16(iii) in [18], or Theorem 4.5.10 in [26], in order to deduce that the sequence  $\{x^k\}_{k \in \mathbb{N}}$  converges strongly when  $\text{int } \mathcal{S}_f(A, C) \neq \emptyset$ .  $\square$

It was pointed out in subsection 3.3 that there is a strong connection between the proximal-projection method (1.7) and the proximal point method (3.23)—see Lemma 3.8. As far as we know, convergence of the proximal point method in reflexive Banach spaces was established only for maximal monotone operators. We use the connection between the proximal point method and the proximal-projection method to obtain convergence of the proximal point method for operators which are monotone with sequentially weak-to-weak closed graphs (but are not necessarily maximal monotone). Clearly, in spaces with finite dimension, any monotone operator with a closed graph and, in general, monotone operators with closed convex graphs, have this property. The other requirement of the next corollary that  $\text{dom } B_f$  should be convex is necessarily satisfied if  $\nabla f^*$  and  $\nabla f + B$  are surjective because, in this case,  $\text{dom } B_f = \text{ran } \nabla f^* \circ (\nabla f + B) = X$ . This condition is sufficient without being necessary, as the example preceding Lemma 3.8 shows. It can be easily verified that this also happens whenever  $\text{Graph } B$  is convex and  $\nabla f$  is linear. Since the corollary is based on Theorem 4.7(a), the remarks preceding Theorem 4.7 concerning the implications of the hypothesis on the domains of  $f$  and  $f^*$  still apply here.

**COROLLARY 4.9.** *Suppose that the following conditions are satisfied:*

- (a)  *$f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ ;*
- (b)  *$\nabla f$  is uniformly continuous on bounded subsets of  $\text{int dom } f$  as well as sequentially weak-to-weak continuous;*
- (c)  *$\nabla f^*$  is bounded on bounded subsets of  $\text{int dom } f^*$ .*

*If  $B : X \rightarrow 2^{X^*}$  is a monotone operator with sequentially weakly-closed graph in  $X \times X^*$ , satisfying (3.24) and such that  $\text{dom } B_f$  is convex, if  $B$  has at least one zero in  $\text{int dom } f$ , and if*

- (d)  *$\text{dom } B_f$  is closed in  $X$*

*or*

- (e)  *$\nabla f$  is bounded on bounded subsets of  $\text{int dom } f$ ,*

*then the sequences generated by the proximal point method (3.23) converge weakly to zeros of the operator  $B$ .*

*Proof.* We start by observing that, due to the boundedness on bounded subsets of its domain of  $\nabla f^*$ , we have that  $\text{dom } f^* = X^*$ —see Remark 4.6(c). Our proof is done in two stages. In Stage 1 we prove that the conclusion of the corollary holds under assumption (d), i.e., when  $\text{dom } B_f$  is closed in  $X$ . In Stage 2 we show that, if  $\nabla f$  is bounded on bounded subsets of  $\text{int dom } f$  (that is, if condition (e) holds), then  $\text{dom } B_f$  is necessarily closed in  $X$  and, thus, by the result of Stage 1, the conclusion of the corollary is true in this case too.

*Stage 1.* Assume that  $\text{dom } B_f$  is closed in  $X$ . By Lemma 3.8, the proximal point method is identical to the proximal-projection method applied to the operator  $A := A[B_f]$  given by (3.20) and to the sets  $C_k = X$ . Therefore, for proving the conclusion of the corollary in this case, it is sufficient to show that the operator  $A := A[B_f]$  satisfies the requirements of Theorem 4.7(a). In order to do that it is sufficient to ensure that the operator  $A[B_f]$  is  $D_f$ -inverse strongly monotone on its domain and has a sequentially weakly-strongly closed graph. Observe that, according to Lemma 3.7, the operator  $A[B_f]$  is  $D_f$ -inverse strongly monotone on its domain. It remains to show that  $A[B_f]$  is sequentially weakly-strongly closed. Let  $\{y^k\}_{k \in \mathbb{N}}$  be a weakly convergent sequence contained in  $\text{dom } A[B_f]$  and denote  $\xi^k = A[B_f]y^k$ . Suppose that  $\{\xi^k\}_{k \in \mathbb{N}}$  converges strongly in  $X^*$  to some vector  $\xi$ . Let  $y = \text{w-lim}_{k \rightarrow \infty} y^k$ . By hypothesis (b), the sequence  $\{\nabla f(y^k)\}_{k \in \mathbb{N}}$  converges weakly in  $X^*$  to  $\nabla f(y)$ . Thus, the sequence

$$(4.36) \quad \nabla f(y^k) - \xi^k = \left[ \nabla f \circ (\nabla f + B)^{-1} \circ \nabla f \right] (y^k)$$

converges weakly in  $X^*$  to  $\nabla f(y) - \xi$ . Denote  $u^k := \nabla f^* (\nabla f(y^k) - \xi^k)$  and observe that, by (4.36), we have that

$$(4.37) \quad u^k = \left[ (\nabla f + B)^{-1} \circ \nabla f \right] (y^k) = B_f y^k \quad \forall k \in \mathbb{N}.$$

According to hypothesis (c), the sequence  $\{u^k\}_{k \in \mathbb{N}}$  is bounded because the sequence  $\{\nabla f(y^k) - \xi^k\}_{k \in \mathbb{N}}$  is bounded (as shown above, this sequence is weakly convergent). Let  $\{u^{i_k}\}_{k \in \mathbb{N}}$  be a weakly convergent subsequence of  $\{u^k\}_{k \in \mathbb{N}}$  and let  $u$  be the weak limit of this subsequence. By (4.37) we deduce that  $\nabla f(y^{i_k}) \in (\nabla f + B) u^{i_k}$  for all  $k \in \mathbb{N}$ , and thus we obtain

$$(4.38) \quad \nabla f(y^{i_k}) - \nabla f(u^{i_k}) \in B u^{i_k} \quad \forall k \in \mathbb{N}.$$

By hypothesis (b), we have that

$$(4.39) \quad \text{w-} \lim_{k \rightarrow \infty} [\nabla f(y^{i_k}) - \nabla f(u^{i_k})] = \nabla f(y) - \nabla f(u).$$

Since  $\text{Graph } B$  is sequentially weak-to-weak closed and  $\{u^{i_k}\}_{k \in \mathbb{N}}$  converges weakly to  $u$ , the relations (4.38) and (4.39) imply that  $\nabla f(y) - \nabla f(u) \in Bu$ , i.e.,  $\nabla f(y) \in \nabla f(u) + Bu$ . Consequently, we have that

$$(4.40) \quad u = (\nabla f + B)^{-1} (\nabla f(y)) = B_f y$$

because the operator  $B_f$  is single-valued (cf. Lemma 3.7). On the other hand, by (4.36), (4.37), (4.39), and (4.40), we have that

$$\xi = \nabla f(y) - \text{w-} \lim_{k \rightarrow \infty} \nabla f(u^{i_k}) = \nabla f(y) - \nabla f(u) = \nabla f(y) - (\nabla f \circ B_f) y,$$

showing that  $(y, \xi) \in \text{Graph } A[B_f]$ . Hence,  $A := A[B_f]$  is sequentially weakly-strongly closed and, by Theorem 4.7(a) combined with Lemma 3.8, it results that any sequence generated by the proximal point method converges weakly to a zero of  $A[B_f]$ . Noting that

$$0^* \in A[B_f] \bar{x} \Leftrightarrow \nabla f(\bar{x}) = (\nabla f \circ B_f)(\bar{x}) \Leftrightarrow \bar{x} = B_f \bar{x} \Leftrightarrow 0^* \in B \bar{x},$$



we deduce that any sequence generated by the proximal point method converges weakly to a zero of  $B$ . This completes the proof in Stage 1.

*Stage 2.* Assume that  $\nabla f$  is bounded on bounded subsets of  $\text{int dom } f$ . Then, by Remark 4.6(c),  $\text{dom } f = X$ . We are going to show that, in this case, the set  $\text{dom } B_f$  is closed. To this end, let  $\{z^k\}_{k \in \mathbb{N}}$  be a sequence contained in  $\text{dom } B_f$  and converging in  $X$  to some vector  $\bar{z}$ . Denote  $w^k = B_f z^k$ . We claim that the sequence  $\{w^k\}_{k \in \mathbb{N}}$  is bounded. To show that, note that since the operator  $A := A[B_f]$  is  $D_f$ -inverse strongly monotone (cf. Lemma 3.7(c)), all solutions of Problem 1.1 (in which  $A = A[B_f]$ ) are in the set  $\text{Nexp}_{\text{dom } A[B_f]}^f (A[B_f])^f$  (cf. Lemma 3.6). Taking into account that, by Lemma 3.7(a,b),  $B_f = (A[B_f])^f$  and  $\text{dom } B_f = \text{dom } A[B_f]$ , we deduce that all solutions of Problem 1.1 are contained in  $\text{Nexp}_{\text{dom } B_f}^f B_f$ . Let  $z$  be such a solution. Then, by Lemma 3.2(b), we have

$$D_f(z, w^k) + D_f(w^k, z^k) \leq D_f(z, z^k) \quad \forall k \in \mathbb{N}.$$

According to the definition of the modulus of total convexity of  $f$  on the bounded set  $\{z^k\}_{k \in \mathbb{N}}$  denoted  $\nu_f(\{z^k\}_{k \in \mathbb{N}}, \cdot)$ —see [38]—we obtain

$$(4.41) \quad 0 \leq \nu_f\left(\{z^k\}_{k \in \mathbb{N}}; \|w^k - z^k\|\right) \leq D_f(w^k, z^k) \leq D_f(z, z^k) \quad \forall k \in \mathbb{N}.$$

Since  $\nabla f$  is bounded on bounded subsets of  $X$ , the function  $f$  is also bounded on bounded subsets of  $X$ . Consequently, taking into account (2.10), we deduce that the sequence  $\{D_f(z, z^k)\}_{k \in \mathbb{N}}$  is bounded. Let  $M$  be an upper bound of this sequence. Suppose by contradiction that the sequence  $\{w^k\}_{k \in \mathbb{N}}$  contains a subsequence  $\{w^{j_k}\}_{k \in \mathbb{N}}$  such that  $\lim_{k \rightarrow \infty} \|w^{j_k}\| = \infty$ . Then there exists a positive integer  $k_0$  such that for all integers  $k \geq k_0$  we have  $\|w^{j_k} - z^{j_k}\| \geq 1$ . By [38, Proposition 2.1(ii)] and (4.41), we deduce that for any  $k \geq k_0$  we have

$$(4.42) \quad \|w^{j_k} - z^{j_k}\| \nu_f\left(\{z^k\}_{k \in \mathbb{N}}; 1\right) \leq \nu_f\left(\{z^k\}_{k \in \mathbb{N}}; \|w^{j_k} - z^{j_k}\|\right) \leq M \quad \forall k \in \mathbb{N}.$$

The function  $f$  is, by hypothesis, uniformly convex on bounded subsets of  $\text{int dom } f$  and, consequently, it is also totally convex on bounded subsets of  $\text{int dom } f$ ; cf. [38, Theorem 2.10]. Therefore,  $\nu_f(\{z^k\}_{k \in \mathbb{N}}; 1) > 0$ . Taking this into account together with the fact that  $\{z^{j_k}\}_{k \in \mathbb{N}}$  is bounded (since it is convergent) and letting  $k \rightarrow \infty$  in (4.42), we reach a contradiction. Hence, the sequence  $\{w^k\}_{k \in \mathbb{N}}$  is bounded. Let  $\{w^{s_k}\}_{k \in \mathbb{N}}$  be a weakly convergent subsequence of  $\{w^k\}_{k \in \mathbb{N}}$  and let  $\bar{w}$  be the weak limit of this subsequence. According to (3.19), we have that

$$(4.43) \quad \nabla f(z^k) - \nabla f(w^k) \in Bw^k \quad \forall k \in \mathbb{N}.$$

Since  $\nabla f$  is sequentially weak-to-weak continuous, and since  $B$  has a sequentially weak-to-weak closed graph, the relation (4.43), written with  $s_k$  instead of  $k$ , implies that  $\nabla f(\bar{z}) - \nabla f(\bar{w}) \in B\bar{w}$ . This shows that  $\bar{w} = B_f \bar{z}$ , that is,  $\bar{z} \in \text{dom } B_f$ . Hence,  $\text{dom } B_f$  is closed and the proof of the corollary is complete.  $\square$

Another result which follows from Theorem 4.7 concerns a method of regularizing and solving classical variational inequalities in the form (2.18). Since the problem of solving (2.18) may be ill-posed (in the sense that it may not have solutions or it may have multiple solutions) and, therefore, many algorithms for approximating solutions

may not converge, or may converge only subsequentially, to solutions of the problem, one “regularizes” the original problem by solving an auxiliary problem which has a unique solution and whose solution is in the vicinity of the solution set of the original problem, provided that the latter is not empty. A regularization technique, which originates in the works of Tikhonov [77] and Browder [27], [28], consists of replacing the original variational inequality (2.18) by the following regularized variational inequality:

$$(4.44) \quad \text{Find } x \in C \cap \text{int dom } f \text{ such that}$$

$$\exists \xi \in Bx : [\langle \xi + \alpha \nabla f(x), y - x \rangle \geq 0 \quad \forall y \in C \cap \text{dom } f]$$

for some real number  $\alpha > 0$ . If  $B$  is a monotone operator, then  $B + \alpha \nabla f$  is strictly monotone and, therefore, the variational inequality (4.44) cannot have more than one solution. Moreover, in many practically interesting situations, the variational inequality (4.44) has a solution even if the original variational inequality (2.18) does not and, if  $\alpha$  is sufficiently small, then the solution of (4.44) is close to the solution set of the unperturbed variational inequality (2.18) whenever the latter has solutions. This is, for instance, the case (cf. [7, Theorem 3.2]) when the Banach space  $X$  is simultaneously uniformly convex and uniformly smooth and endowed with the Legendre function  $f := \frac{1}{p} \|\cdot\|^p$  for some  $p > 1$  and  $B$  is maximal monotone. Theorem 4.7 allows us to prove the next corollary which extends the applicability of this regularization technique to reflexive Banach spaces which are not necessarily uniformly convex and uniformly smooth and to produce a weakly convergent algorithm for solving (4.44) in this more general setting, even if  $B$  is not maximal monotone. This is of interest because closeness of the solution of (4.44) with small  $\alpha > 0$  to the (presumed nonempty) solution set of the original variational inequality (2.18) can be guaranteed even if  $B$  is not maximal monotone (cf. [9, Theorem 2.1]).

**COROLLARY 4.10.** *Let  $B : X \rightarrow 2^{X^*}$  be a monotone operator and let  $C$  be a nonempty, convex, and closed subset of  $\text{dom } B \cap \text{int dom } f$ . Suppose that  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ ,  $\nabla f^*$  is bounded on bounded subsets of  $\text{int dom } f^*$ , and that, for some real number  $\alpha > 0$ , we have that*

$$(4.45) \quad \emptyset \neq ((1 - \alpha)\nabla f - B)(C) \subseteq \text{int dom } f^*,$$

*and the operator  $\text{Proj}_C^f [(1 - \alpha)\nabla f - B]$  is  $D_f$ -firm. The iterative procedure defined by*

$$(4.46) \quad x^0 \in C \text{ and } x^{k+1} \in \text{Proj}_C^f [(1 - \alpha)\nabla f(x^k) - Bx^k] \quad \forall k \in \mathbb{N}$$

*is well defined and converges weakly to the necessarily unique solution of the variational inequality (4.44), provided that such a solution exists, if one of the following conditions is satisfied:*

(a)  *$X$  has finite dimension,  $\text{dom } f = X$ ,  $\nabla f$  is uniformly continuous on bounded subsets of  $X$ , and  $B$  has a closed graph and is bounded on bounded subsets of its domain;*

(b)  *$\nabla f$  is bounded on bounded subsets of  $\text{int dom } f$ , and the operator*

$$\text{Proj}_C^f [(1 - \alpha)\nabla f - B]$$

*is  $D_f$ -nonexpansive.*

*Proof.* Well definedness of the procedure results from (4.45). The variational inequality (4.44) cannot have more than one solution since the operator  $B' := B + \alpha \nabla f$  is strictly monotone. According to Lemma 2.5, finding a solution of (4.44) is equivalent to finding a zero for the operator  $V := V[B'; C; f]$  defined by (2.19). Note that

$$(4.47) \quad V = \nabla f - \nabla f \circ \text{Proj}_C^f((1 - \alpha)\nabla f - B)$$

and

$$(4.48) \quad V^f = \text{Proj}_C^f((1 - \alpha)\nabla f - B)$$

and that, by (2.21) applied to  $B'$  instead of  $B$ , we have

$$(4.49) \quad \text{Proj}_C^f(\nabla f - V) = \text{Proj}_C^f((1 - \alpha)\nabla f - B).$$

Therefore, we can equivalently rewrite the procedure (4.46) as

$$(4.50) \quad x^0 \in C \quad \text{and} \quad x^{k+1} \in \text{Proj}_C^f(\nabla f(x^k) - Vx^k) \quad \forall k \in \mathbb{N}.$$

This is exactly (1.7) applied to  $V$  instead of  $A$  with the sequence of sets  $C_k = C$  for all  $k \in \mathbb{N}$ . Now, suppose that condition (a) of our corollary is satisfied. In this case, if we show that the graph of  $V$  is closed in  $X \times X^*$  and that  $V$  is  $D_f$ -inverse strongly monotone, then Theorem 4.7(a) applies and leads to the conclusion of the corollary. Also, Theorem 4.7(b) implies that, if condition (b) of the corollary holds, then the procedure (4.50) is weakly convergent to the unique solution of (4.44), provided that  $V$  is  $D_f$ -inverse strongly monotone.  $D_f$ -inverse strong monotonicity of  $V$  results in both cases from Lemma 3.5 combined with (4.48) and with our hypothesis that  $V^f = \text{Proj}_C^f[(1 - \alpha)\nabla f - B]$  is  $D_f$ -firm. So, it remains to prove that, under assumption (a) of the corollary, the graph of  $V$  is closed. To this end, let  $\{y^k\}_{k \in \mathbb{N}}$  be a sequence in  $\text{dom } V$  and assume that this sequence converges to  $y \in X$ . Let  $\{\xi^k\}_{k \in \mathbb{N}}$  be the sequence

$$\xi^k = \nabla f(y^k) - \nabla f \circ \text{Proj}_C^f((1 - \alpha)\nabla f(y^k) - \zeta^k),$$

where  $\zeta^k \in By^k$  for all  $k \in \mathbb{N}$ . Suppose that  $\lim_{k \rightarrow \infty} \xi^k = \xi$ . Then, by Lemma 2.1, we have

$$\nabla f(y^k) - \xi^k = \left[ \nabla f \circ (\nabla f + N_C)^{-1} \right] ((1 - \alpha)\nabla f(y^k) - \zeta^k).$$

According to the hypothesis,  $\nabla f^*$  is bounded on bounded subsets of  $\text{int dom } f^*$ . This and Remark 4.6(c) combined imply that  $\text{dom } f^* = X^*$ . Therefore, for each  $k \in \mathbb{N}$ , we have that  $\nabla f(y^k) - \xi^k \in X^* = \text{int dom } f^*$  and, thus, from the previous equality we deduce

$$(4.51) \quad (1 - \alpha)\nabla f(y^k) - \zeta^k \in (\nabla f + N_C) [\nabla f^*(\nabla f(y^k) - \xi^k)] \quad \forall k \in \mathbb{N}.$$

Since  $B$  is bounded on the bounded set  $\{y^k\}_{k \in \mathbb{N}}$ , it follows that the sequence  $\{\zeta^k\}_{k \in \mathbb{N}}$  is bounded. Let  $\{\zeta^{i_k}\}_{k \in \mathbb{N}}$  be a convergent subsequence of  $\{\zeta^k\}_{k \in \mathbb{N}}$  and let  $\zeta$  be its limit. Since  $\nabla f$  and  $\nabla f^*$  are continuous on their respective domains,  $\text{dom } f^* = X^*$  (as shown above), and the normality operator  $N_C$  is maximal monotone (and, hence, has a closed graph), (4.51) implies

$$(1 - \alpha)\nabla f(y) - \zeta \in (\nabla f + N_C) [\nabla f^*(\nabla f(y) - \xi)].$$

Hence, we have

$$\nabla f^*(\nabla f(y) - \xi) = \text{Proj}_C^f [(1 - \alpha)\nabla f(y) - \zeta]$$

showing that

$$\xi = \nabla f(y) - \nabla f \circ \text{Proj}_C^f [(1 - \alpha)\nabla f(y) - \zeta] \in Vy.$$

This completes the proof.  $\square$

The requirement made in Corollary 4.10 that the operator  $V^f = \text{Proj}_C^f [(1 - \alpha)\nabla f - B]$  should be  $D_f$ -firm may seem unusual. Here are several examples which show that this condition is quite often satisfied without excessively costly demands on the operator  $B$  or the function  $f$ . The next example shows that Corollary 4.10(b) is applicable to solve variational inequalities in the form (4.44) when  $X$  is any Hilbert space,  $f = \frac{1}{2} \|\cdot\|^2$ ,  $C = \text{dom } B = X$ , and  $B$  is a monotone operator which is either contractive with some constant  $\gamma > 0$  or strongly monotone with some constant  $\delta > 0$  and even in more general conditions (when (4.52) holds for some  $\alpha \in (0, \frac{1}{2})$  and  $\beta > 0$ ). Obviously, in this setting, (2.18) is exactly the problem of finding a zero of  $B$ . To follow the considerations in the examples below one should first note that by replacing in (2.18) the operator  $B$  by  $\beta B$ , where  $\beta$  is a positive constant, one obtains a variational inequality which is equivalent to the original one.

*Example 4.11.* Let  $X$  be a Hilbert space and let  $f = \frac{1}{2} \|\cdot\|^2$ . Suppose that  $C = \text{dom } B = X$ . If  $B$  is contractive with some constant  $\gamma > 0$  or strongly monotone with some constant  $\delta > 0$ , then the operator  $V^f = \text{Proj}_C^f [(1 - \alpha)\nabla f - \beta B]$  is  $D_f$ -firm and  $D_f$ -nonexpansive whenever  $\alpha \in (0, \frac{1}{2})$  with  $\beta = \sqrt{\alpha(1 - \alpha)}\gamma^{-1}$  in the contractive case and  $\beta = (1 - 2\alpha)\delta$  in the strongly monotone case.

In order to prove this observe that, in the current setting,  $\nabla f = \text{Proj}_C^f = I$ . Also, the  $D_f$ -firmness condition (3.14) for  $T = V^f$  is equivalent to (4.34) and this is exactly

$$\begin{aligned} & \langle [(1 - \alpha)x - \beta\xi] - [(1 - \alpha)y - \beta\eta], x - y \rangle \\ & \geq \|[(1 - \alpha)x - \beta\xi] - [(1 - \alpha)y - \beta\eta]\|^2, \end{aligned}$$

which can be equivalently rewritten as

$$(4.52) \quad \alpha(1 - \alpha) \|x - y\|^2 + (1 - 2\alpha)\beta \langle \xi - \eta, x - y \rangle \geq \beta^2 \|\xi - \eta\|^2$$

for all pairs  $(x, \xi), (y, \eta) \in \text{Graph } B$ . Moreover, if  $V^f$  is  $D_f$ -firm, then it is also  $D_f$ -nonexpansive. Due to the monotonicity of  $B$ , the inequality (4.52) holds whenever  $\alpha \in (0, \frac{1}{2})$  and

$$(4.53) \quad \alpha(1 - \alpha) \|x - y\|^2 \geq \beta^2 \|\xi - \eta\|^2 \quad \forall (x, \xi) \in \text{Graph } B, \quad \forall (y, \eta) \in \text{Graph } B.$$

If  $B$  is contractive with constant  $\gamma$ , then we have

$$\gamma \|x - y\| \geq \|\xi - \eta\| \quad \forall (x, \xi) \in \text{Graph } B, \quad \forall (y, \eta) \in \text{Graph } B,$$

and multiplying this inequality by  $\beta = \sqrt{\alpha(1 - \alpha)}\gamma^{-1}$  we deduce that (4.53) holds. Similarly, note that (4.52) is satisfied when

$$(4.54) \quad (1 - 2\alpha) \langle \xi - \eta, x - y \rangle \geq \beta \|\xi - \eta\|^2 \quad \forall (x, \xi) \in \text{Graph } B, \quad \forall (y, \eta) \in \text{Graph } B.$$

If  $B$  is strongly monotone with constant  $\delta$ , then

$$\langle \xi - \eta, x - y \rangle \geq \delta \|\xi - \eta\|^2 \quad \forall (x, \xi) \in \text{Graph} B, \quad \forall (y, \eta) \in \text{Graph} B.$$

Multiplying this inequality by  $(1 - 2\alpha)$  we deduce that (4.54) holds in this case.  $\square$

In situations in which  $X$  is not a Hilbert space, or  $X$  is a Hilbert space but the monotone operator  $B$  does not satisfy (4.52) for some  $\alpha \in (0, \frac{1}{2})$  and  $\beta > 0$ , the question of how to choose the “regularization function”  $f$  and the “regularization parameter”  $\alpha$  in the perturbed variational inequality (4.44) in order to force  $D_f$ -firmness and/or  $D_f$ -nonexpansivity on  $V^f = \text{Proj}_C^f((1 - \alpha)\nabla f - B)$  is relevant. Here are examples of situations when  $V^f$  is  $D_f$ -firm even if  $X$  is not a Hilbert space.

*Example 4.12.* If the monotone operator  $B : X \rightarrow 2^{X^*}$  and the nonempty closed convex subset  $C$  of  $\text{dom } B \cap \text{int dom } f$  have the property (4.45), then  $V^f = \text{Proj}_C^f((1 - \alpha)\nabla f - B)$  is  $D_f$ -firm in any of the following situations:

(a)  $C = X$ ,  $\alpha \in (0, 1)$ , and the operator  $\alpha\nabla f + B$  is  $D_f$ -inverse strongly monotone on its domain;

(b)  $X$  is a Hilbert space,  $C = X$ ,  $f = \frac{1}{2} \|\cdot\|^2$ ,  $\alpha \in (0, \frac{1}{2}]$ , and  $B$  is contractive with constant  $\alpha(1 - \alpha)$ .

In order to show this, observe that in case (a), according to Lemma 2.1, we have  $V^f = (\alpha\nabla f + B)^f$  and, by Lemma 3.5(c), the conclusion follows. In case (b) we have that both  $\nabla f$  and  $\text{Proj}_C^f$  coincide with the identity operator  $I$ . Also,  $B$  is necessarily single-valued because it is contractive. These imply (see section 3) that verification of the  $D_f$ -firmness of  $V^f$  amounts to verifying the firm nonexpansivity of  $(1 - \alpha)I - B$ . Note that  $(1 - \alpha)I - B$  is firmly nonexpansive (see (3.10)) if and only if the operator  $\alpha I + B$  is firmly nonexpansive, that is, if and only if

$$\begin{aligned} & \alpha^2 \|x - y\|^2 + 2\alpha \langle x - y, Bx - By \rangle + \|Bx - By\|^2 \\ & \leq \alpha \|x - y\|^2 + \langle x - y, Bx - By \rangle \quad \forall x, y \in X. \end{aligned}$$

Since  $\alpha \in (0, \frac{1}{2}]$  and  $B$  is monotone, the last inequality holds whenever

$$(4.55) \quad \alpha^2 \|x - y\|^2 + \|Bx - By\|^2 \leq \alpha \|x - y\|^2 \quad \forall x, y \in X.$$

By the contractivity of  $B$  with constant  $\alpha(1 - \alpha)$  we deduce that  $\|Bx - By\|^2 \leq \alpha^2(1 - \alpha)^2 \|x - y\|^2$ , and this shows that (4.55) is obviously satisfied.  $\square$

Example 4.12(a), in conjunction with Corollary 4.10, leads to an algorithm for solving the variational inequality (4.44) whenever it is possible to find a Legendre function  $f$  and a number  $\alpha \in (0, 1)$  such that  $\alpha\nabla f + B$  is  $D_f$ -inverse strongly monotone on its domain. Example 4.12(b) points out a situation in which the assumptions of Corollary 4.10(b) hold and, thus, weak convergence of procedure (4.46) to the (presumed existing) solution of the variational inequality (4.44) is guaranteed.

*Example 4.13.* If the monotone operator  $B : X \rightarrow 2^{X^*}$  and the nonempty closed convex subset  $C$  of  $\text{dom } B \cap \text{int dom } f$  have the property (4.45), and if

$$(4.56) \quad \begin{aligned} & \left\langle [\alpha\nabla f(x) + \xi] - [\alpha\nabla f(y) + \eta], \right. \\ & \left. \text{Proj}_C^f [(1 - \alpha)\nabla f(x) - \xi] - \text{Proj}_C^f [(1 - \alpha)\nabla f(y) - \eta] \right\rangle \geq 0, \end{aligned}$$

for any  $(x, \xi)$  and  $(y, \eta)$  in  $\text{Graph} B$ , then the operator  $V^f = \text{Proj}_C^f[(1 - \alpha)\nabla f - B]$  is  $D_f$ -firm. In particular, this happens in any of the following situations:

- (a)  $B = (\frac{1}{2} - \alpha) \nabla f$  and  $\alpha \in (0, \frac{1}{2})$ ;  
 (b) the operator  $P : X \times X^* \rightarrow 2^{X^* \times X}$ , defined by

$$(4.57) \quad P(z, \zeta) = (0^*, \text{Proj}_C^f((1 - \alpha)\nabla f(z) - \zeta)),$$

for some  $\alpha \in (0, 1)$  is monotone when  $X \times X^*$  is provided with the norm  $\|(z, \zeta)\| = (\|z\|^2 + \|\zeta\|_*^2)^{1/2}$  and with the duality pairing  $\langle (z, \zeta), (\zeta', z') \rangle = \langle \zeta', z \rangle + \langle \zeta, z' \rangle$  (and, therefore, its dual is isometric with  $X^* \times X$ ).

Let  $x, y \in \text{dom } V^f$  and let  $\xi \in Bx$  and  $\eta \in By$ . Denote

$$x' = \text{Proj}_C^f((1 - \alpha)\nabla f(x) - \xi) \text{ and } y' = \text{Proj}_C^f((1 - \alpha)\nabla f(y) - \eta).$$

By Lemma 2.1 we have

$$(1 - \alpha)\nabla f(x) - \xi - \nabla f(x') \in N_C(x') \text{ and } (1 - \alpha)\nabla f(y) - \eta - \nabla f(y') \in N_C(y'),$$

which imply

$$\langle (1 - \alpha)\nabla f(x) - \xi, y' - x' \rangle \leq \langle \nabla f(x'), y' - x' \rangle$$

and, respectively,

$$\langle (1 - \alpha)\nabla f(y) - \eta, x' - y' \rangle \leq \langle \nabla f(y'), x' - y' \rangle.$$

Summing up the last two inequalities, we obtain that

$$(4.58) \quad \begin{aligned} & \langle [(1 - \alpha)\nabla f(x) - \xi] - [(1 - \alpha)\nabla f(y) - \eta], x' - y' \rangle \\ & \geq \langle \nabla f(x') - \nabla f(y'), x' - y' \rangle. \end{aligned}$$

The  $D_f$ -firmness condition (3.14) for the operator  $T = V^f$  is exactly

$$\langle \nabla f(x) - \nabla f(y), x' - y' \rangle \geq \langle \nabla f(x') - \nabla f(y'), x' - y' \rangle.$$

According to (4.58), this is satisfied when

$$\langle \nabla f(x) - \nabla f(y), x' - y' \rangle \geq \langle [(1 - \alpha)\nabla f(x) - \xi] - [(1 - \alpha)\nabla f(y) - \eta], x' - y' \rangle,$$

and this last inequality is equivalent to (4.56). Hence, if (4.56) holds, then the operator  $V^f$  is  $D_f$ -firm. In case (a) we have that  $\alpha \nabla f + B = (1 - \alpha)\nabla f - B$ , and using the monotonicity of  $\text{Proj}_C^f$  (cf. [38, Theorem 4.6]) one deduces that (4.56) holds. Suppose that we are in case (b) and the operator  $P$ , given by (4.57), is monotone. Then observe that

$$\begin{aligned} & \left\langle [\alpha \nabla f(x) + \xi] - [\alpha \nabla f(y) + \eta], \right. \\ & \left. \text{Proj}_C^f[(1 - \alpha)\nabla f(x) - \xi] - \text{Proj}_C^f[(1 - \alpha)\nabla f(y) - \eta] \right\rangle \\ & = \langle (x, \alpha \nabla f(x) + \xi) - (y, \alpha \nabla f(y) + \eta), P(x, \xi) - P(y, \eta) \rangle, \end{aligned}$$

and that the last expression is nonnegative due to the monotonicity of  $P$  (see [38, Proposition 4.7]). Hence, (4.56) holds in this case too.  $\square$

Note that problem (4.44), in which  $B = \beta \nabla f$  for some  $\beta > 0$ , is equivalent to the problem of finding the minimizer of  $f$  over  $C$ . The facts observed in Example 4.13(a), in conjunction with Corollary 4.10, lead to a proximal-projection method of finding that minimizer, provided that  $f$  satisfies the other requirements there. Obviously, the effectiveness of that method, as well as of the other methods discussed in this work, depends on the possibility of computing proximal projections onto  $C$ . Algorithms for computing proximal projections are presented in [5], [22], and [38].

Observe that the definition of the operator  $P$ , given by (4.57), does not involve the operator  $B$  but only the Legendre function  $f$ , the closed and convex set  $C$ , and the real number  $\alpha \in (0, 1)$ . Thus, Example 4.13(b) shows that if  $P$  is monotone, then  $\text{Proj}_C^f [(1 - \alpha) \nabla f - B]$  is  $D_f$ -firm for all monotone operators  $B : X \rightarrow 2^{X^*}$ . In other words, if for the Legendre function  $f$ , the closed and convex set  $C$  and the real number  $\alpha \in (0, 1)$  monotonicity of  $P$  can be ensured, then Corollary 4.10 guarantees solvability of a large class of variational inequalities via the particular variant of proximal-projection method (4.46). Since in the process of solving variational inequalities the set  $C$  is a priori given, one should ask whether on some Banach spaces one can find Legendre functions  $f$  for which monotonicity of  $P$  is ensured for some  $\alpha \in (0, 1)$  no matter how the closed and convex set  $C$  is chosen. We do not have any example to prove or disprove existence of spaces  $X$  on which such a Legendre function  $f$  exists.

**5. Convergence of the proximal-projection method in spaces of finite dimension.** Theorem 4.7 and its corollaries ensure weak and, sometimes, strong convergence of the proximal-projection method to solutions of Problem 1.1 under conditions which, besides the  $D_f$ -inverse strong monotonicity of the operator  $A$ , require sequential weak-strong closedness of the  $\text{Graph} A$  or, alternatively,  $D_f$ -nonexpansivity of  $A^f$ . In this section we show that, when the space  $X$  has finite dimension, some of these requirements can be dropped or weakened. This is possible due to the validity in spaces of finite dimension of another generalization of Browder's demiclosedness principle, which we present below.

**5.1. Another variant of the generalized demiclosedness principle.** The following result applies to operators  $T : X \rightarrow 2^X$  which are not necessarily  $D_f$ -nonexpansive but satisfy condition (5.2) below, which is more general than  $D_f$ -firmness (compare condition (5.2) with Definition 3.4). It is interesting to observe that, if  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$  and if  $T$  has a closed graph, then the conclusion of the next result holds even if the hypothesis that  $u$  satisfies (5.2) is removed. This happens because the equality in (5.1) implies that the sequences  $\{z^k\}_{k \in \mathbb{N}}$  and  $\{u^k\}_{k \in \mathbb{N}}$  converge to the same limit  $z$  and, then, closedness of the graph of  $T$  guarantees that  $z \in Tz$ .

**PROPOSITION 5.1.** *Suppose that the space  $X$  has finite dimension,  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ , and  $T : X \rightarrow 2^X$  is an operator satisfying condition (3.13). Let  $\{z^k\}_{k \in \mathbb{N}}$  be a sequence in  $\text{dom } T$  converging to an element  $z \in \text{dom } T$ . If for some sequence  $\{u^k\}_{k \in \mathbb{N}}$  satisfying*

$$(5.1) \quad (\forall k \in \mathbb{N} : u^k \in Tz^k) \quad \text{and} \quad \lim_{k \rightarrow \infty} D_f(u^k, z^k) = 0,$$

*there exists  $u \in Tz$  such that*

$$(5.2) \quad \liminf_{k \rightarrow \infty} \langle \nabla f(u^k) - \nabla f(u), u^k - u \rangle \leq 0,$$

*then the vector  $z$  is a fixed point of  $T$ .*

*Proof.* Since the function  $f$  is convex and differentiable on  $\text{int dom } f$ , the gradient  $\nabla f$  is continuous on  $\text{int dom } f$ . This fact and the strict convexity of  $f$  on  $\text{int dom } f$  imply that

$$(5.3) \quad \lim_{k \rightarrow \infty} \langle \nabla f(z^k) - \nabla f(x), z^k - x \rangle = \langle \nabla f(z) - \nabla f(x), z - x \rangle > 0$$

whenever  $x \in (\text{int dom } f) \setminus \{z\}$ . Because the function  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ , it is also sequentially consistent (cf. [38, Theorem 2.10]). Therefore, the equality in (5.1) implies that

$$(5.4) \quad \lim_{k \rightarrow \infty} \|z^k - u^k\| = 0.$$

Consequently, the sequences  $\{z^k\}_{k \in \mathbb{N}}$  and  $\{u^k\}_{k \in \mathbb{N}}$  converge to the same limit  $z$ . By condition (3.13), the boundedness of  $\{u^k\}_{k \in \mathbb{N}}$ , and the continuity of  $\nabla f$ , we deduce that

$$(5.5) \quad \lim_{k \rightarrow \infty} \langle \nabla f(z^k) - \nabla f(z), z^k - z \rangle = 0 = \lim_{k \rightarrow \infty} \langle \nabla f(z^k) - \nabla f(z), u^k - u \rangle.$$

Note that

$$(5.6) \quad \begin{aligned} \langle \nabla f(u^k) - \nabla f(u), u^k - u \rangle &= \langle \nabla f(u^k) - \nabla f(u), u^k - z^k \rangle \\ &\quad + \langle \nabla f(u^k) - \nabla f(u), z^k - u \rangle \\ &= \langle \nabla f(u^k) - \nabla f(u), u^k - z^k \rangle \\ &\quad + \langle \nabla f(z^k) - \nabla f(u), z^k - u \rangle \\ &\quad + \langle \nabla f(u^k) - \nabla f(z^k), z^k - u \rangle. \end{aligned}$$

Since the sequence  $\{\nabla f(u^k)\}_{k \in \mathbb{N}}$  is bounded, it follows from (5.4) that the first term of the last sum in (5.6) converges to zero as  $k \rightarrow \infty$ . The continuity of  $\nabla f$  and the fact noted above that the sequences  $\{u^k\}_{k \in \mathbb{N}}$  and  $\{z^k\}_{k \in \mathbb{N}}$  converge to the same limit  $z$  imply that the third term of the last sum converges to zero as  $k \rightarrow \infty$ . Taking the limit as  $k \rightarrow \infty$  on both sides of (5.6), we obtain that

$$(5.7) \quad \lim_{k \rightarrow \infty} \langle \nabla f(u^k) - \nabla f(u), u^k - u \rangle = \lim_{k \rightarrow \infty} \langle \nabla f(z^k) - \nabla f(u), z^k - u \rangle.$$

In order to conclude the proof, suppose by contradiction that  $z \notin T(z)$ . Then,  $u \neq z$  and, therefore, by (5.5), (5.2), (5.7), and (5.3), respectively, we obtain

$$\begin{aligned} 0 &= \lim_{k \rightarrow \infty} \langle \nabla f(z^k) - \nabla f(z), z^k - z \rangle = \lim_{k \rightarrow \infty} \langle \nabla f(z^k) - \nabla f(z), u^k - u \rangle \\ &\geq \lim_{k \rightarrow \infty} \langle \nabla f(u^k) - \nabla f(u), u^k - u \rangle = \lim_{k \rightarrow \infty} \langle \nabla f(z^k) - \nabla f(u), z^k - u \rangle > 0, \end{aligned}$$

which is a contradiction.  $\square$

**5.2. A convergence theorem for the proximal-projection method in spaces of finite dimension.** The following theorem shows that, in finite-dimensional spaces, convergence of the proximal-projection method to solutions of Problem 1.1 can be ensured with lesser requirements on the operator  $A$ , in addition to the  $D_f$ -inverse strong monotonicity, than those involved in Theorem 4.7 and its corollaries.

**THEOREM 5.2.** *Suppose that the space  $X$  has finite dimension,  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ ,  $\nabla f^*$  is bounded on bounded subsets of  $\nabla f(\text{dom } A)$ , that (1.6) and Assumption 1.2 hold, and*

$$(5.8) \quad C \cap \text{dom } A = \text{w-}\overline{\lim}_{k \rightarrow \infty} (C_k \cap \text{dom } A).$$



If Problem 1.1 has at least one solution, if the operator  $A : X \rightarrow 2^{X^*}$  is  $D_f$ -inverse strongly monotone on  $Q = \bigcup_{k=0}^{\infty} C_k$ , and if  $C \cap \text{dom } A$  is closed, then the sequences generated by the proximal-projection method (1.7) are well defined and converge to solutions of Problem 1.1.

*Proof.* Well definedness of the sequences generated by (1.7) follows from (1.6) and Assumption 1.2. Suppose that, for each  $k \in \mathbb{N}$ ,  $\zeta^k$  and  $u^k$  are as in (4.19) and (4.20), respectively. Then, clearly, condition (4.21) holds too. Because the operator  $A$  is  $D_f$ -inverse strongly monotone on its domain, the operator  $A^f$  is  $D_f$ -firm (cf. Lemma 3.5). This means that

$$(5.9) \quad \langle \nabla f(u^k) - \nabla f(u), u^k - u \rangle \leq \langle \nabla f(x^k) - \nabla f(x), u^k - u \rangle$$

for any pair  $(x, u) \in \text{Graph } A^f$  and for any  $k \in \mathbb{N}$ . Now, repeating without change the arguments in the proof of Theorem 4.7, one can see that Claim 1 proved there still holds in our setting and implies that the sequence  $\{x^k\}_{k \in \mathbb{N}}$  is bounded. Let  $\{x^{i_k}\}_{k \in \mathbb{N}}$  be a convergent subsequence of  $\{x^k\}_{k \in \mathbb{N}}$  and let  $\bar{x}$  be its limit. An argument identical to that made in the proof of Theorem 4.7 (Claim 2) for the same purpose shows that  $\bar{x} \in C$  and (4.30) holds. According to (5.8), since  $x^{i_k} \in C_{i_k} \cap \text{dom } A$ , it results that

$$(5.10) \quad \bar{x} \in C \cap \text{dom } A.$$

This and (1.6) imply that  $(\nabla f - A)(\bar{x}) \neq \emptyset$ . So, by Assumption 1.2, we deduce

$$\emptyset \neq (\nabla f - A)(\bar{x}) \subseteq (\nabla f - A)(C) \subseteq (\nabla f - A)(C_k) \subseteq \text{int dom } f^*,$$

which clearly implies that  $\bar{x} \in \text{dom } A^f$ . Now, taking into account that, by (5.10),  $A\bar{x} \neq \emptyset$ , writing (5.9) for  $i_k$  instead of  $k$  and  $\bar{x}$  instead of  $x$ , and for any  $u \in A\bar{x}$ , and letting in the resulting inequality  $k \rightarrow \infty$ , we deduce that

$$(5.11) \quad \liminf_{k \rightarrow \infty} \langle \nabla f(u^{i_k}) - \nabla f(u), u^{i_k} - u \rangle \leq 0$$

because  $\nabla f$  is continuous on  $\text{int dom } f$ . This shows that the sequence  $\{u^{i_k}\}_{k \in \mathbb{N}}$  satisfies (5.2) for  $T = A^f$ ,  $z = \bar{x}$ , and any  $u \in A^f \bar{x}$ . Also, by (4.30) and (5.11), the sequences  $\{x^{i_k}\}_{k \in \mathbb{N}}$  and  $\{u^{i_k}\}_{k \in \mathbb{N}}$  satisfy (5.1). Hence, Proposition 5.1 applies to  $T = A^f$ ,  $z = \bar{x}$ , and  $u \in A^f \bar{x}$ . By consequence, we have that  $\bar{x}$  is a fixed point of  $A^f$  and, hence, a zero of  $A$ . It remains to prove that  $\bar{x}$  is the only accumulation point of the sequence  $\{x^k\}_{k \in \mathbb{N}}$ . The proof in this respect reproduces without modifications the arguments made in the proof of Theorem 4.7 in order to show that the sequence  $\{x^k\}_{k \in \mathbb{N}}$  has a single weak accumulation point when  $\nabla f$  is sequentially weak-to-weak continuous.  $\square$

**5.3. Consequences of Theorem 5.2.** Using Theorem 5.2 instead of Theorem 4.7 we can prove again Corollaries 4.9 and 4.10 in a finite-dimensional setting with different, and less demanding, conditions on  $A$ . Here is the new version of Corollary 4.9.

**COROLLARY 5.3.** *Suppose that the space  $X$  has finite dimension,  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ , and  $\text{dom } f^* = X^*$ . If  $B : X \rightarrow 2^{X^*}$  is a monotone operator satisfying (3.24) and having at least one zero, and if any of the following conditions holds:*

- (a)  $\text{ran } (\nabla f + B)$  is closed in  $X^*$  and  $\nabla f^*(\text{ran } (\nabla f + B))$  is convex;
- (b)  $\text{ran } (\nabla f + B) = X^*$ ,

then the sequences generated by the proximal point method (3.23) converge to zeros of the operator  $B$ .

*Proof.* Recall that in this setting  $\nabla f^*$  is bounded on bounded subsets of  $\text{int dom } f^*$  (cf. Remark 4.6(d)). Observe that, according to Lemma 3.7, we have that

$$\text{dom } A[B_f] = \text{dom } B_f = \nabla f^*(\text{ran } (\nabla f + B)).$$

Therefore, if condition (a) holds, the set  $\text{dom } A[B_f] = \text{dom } B_f$  is closed and convex. Since condition (b) implies (a), this remains true when (b) holds. Again by Lemma 3.7, the operator  $A[B_f]$  is  $D_f$ -inverse strongly monotone on its domain. Consequently, the operator  $A := A[B_f]$  satisfies the requirements of Theorem 5.2 when  $C = C_k = X$  for all  $k \in \mathbb{N}$ . Applying Theorem 5.2 to  $A[B_f]$  and taking into account Lemma 3.8, the conclusion follows.  $\square$

Now we give another variant of Corollary 4.10(a) in which the condition that  $A$  should have a closed graph is replaced by less demanding requirements.

**COROLLARY 5.4.** *Let  $B : X \rightarrow 2^{X^*}$  be a monotone operator. Suppose that the space  $X$  has finite dimension,  $f$  is uniformly convex on bounded subsets of  $\text{int dom } f$ , and  $\text{dom } f^* = X^*$ . If  $C$  is a closed convex subset of  $\text{dom } B \cap \text{int dom } f$  such that, for some real number  $\alpha > 0$ ,*

$$(5.12) \quad \emptyset \neq ((1 - \alpha)\nabla f - B)(C) \subseteq \text{int dom } f^*,$$

*and the operator  $\text{Proj}_C^f \circ [(1 - \alpha)\nabla f - B]$  is  $D_f$ -firm, then the iterative procedure given by (4.46) is well defined and converges to the necessarily unique solution of the variational inequality (4.44), provided that such a solution exists.*

*Proof.* Since (4.48) and (4.49) still hold, the operator  $V$  given by (4.47) is  $D_f$ -inverse strongly monotone on its domain (cf. Lemma 3.7). By (5.12) and by the fact that  $C \subseteq \text{dom } B \cap \text{int dom } f$ , it results that  $C \subseteq \text{dom } V$ . Hence, Theorem 5.2 applies to the operator  $A = V$  and the sets  $C_k = C$ , and the conclusion follows.  $\square$

**Acknowledgments.** The authors are grateful to two anonymous referees for helpful comments and suggestions which led to improvement of the originally submitted version of this work.

## REFERENCES

- [1] Y. I. ALBER, *Generalized projection operators in Banach spaces: Properties and applications*, in Functional-Differential Equations, M. E. Draklin and E. Litsyn, eds., Differential Equations Israel Sem. 1, Coll. Judea Samaria, Ariel, 1993, pp.1–21.
- [2] Y. I. ALBER, *Metric and generalized projection operators in Banach spaces: Properties and applications*, in Theory and Applications of Nonlinear Operators of Accretive and Monotone Type, A. G. Kartsatos, ed., Lecture Notes in Pure and Appl. Math. 178, Marcel Dekker, New York, 1996, pp. 15–50.
- [3] Y. I. ALBER, *Generalized projections, decompositions, and the Pythagorean-type theorem in Banach spaces*, Appl. Math. Lett., 11 (1998), pp. 115–121.
- [4] Y. I. ALBER, *Stability of the proximal projection algorithm for nonsmooth convex optimization problems with perturbed constraint sets*, J. Nonlinear Convex Anal., 4 (2003), pp. 1–14.
- [5] Y. I. ALBER AND D. BUTNARIU, *Convergence of Bregman projection methods for solving consistent convex feasibility problems in reflexive Banach spaces*, J. Optim. Theory Appl., 92 (1997), pp. 33–61.
- [6] Y. I. ALBER, D. BUTNARIU, AND G. KASSAY, *On the convergence and stability of a regularization method for maximal monotone inclusions and its applications to convex optimization*, in Variational Inequalities and Applications, F. Giannessi and G. Maugeri, eds., Springer-Verlag, New York, 2005, pp. 89–132.

- [7] Y. I. ALBER, D. BUTNARIU, AND I. RYAZANTSEVA, *Regularization methods for ill-posed inclusions and variational inequalities with domain perturbations*, J. Nonlinear Convex Anal., 2 (2001), pp. 53–79.
- [8] Y. I. ALBER, D. BUTNARIU, AND I. RYAZANTSEVA, *Regularization of monotone variational inequalities with Mosco approximations of the constraint sets*, Set-Valued Anal., 13 (2005), pp. 265–290.
- [9] Y. I. ALBER, D. BUTNARIU, AND I. RYAZANTSEVA, *Regularization and resolution of monotone variational inequalities with operators given by hypomonotone approximations*, J. Nonlinear Convex Anal., 6 (2005), pp. 23–53.
- [10] Y. I. ALBER AND S. GUERRE-DELABRIERE, *On the projection method for fixed point problems*, Analysis (Munich), 21 (2001), pp. 17–39.
- [11] Y. I. ALBER, A. N. IUSEM, A. N. SOLODIV, AND M. V. SOLODOV, *Minimization of nonsmooth convex functionals in Banach spaces*, J. Convex Anal., 4 (1997), pp. 235–255.
- [12] Y. I. ALBER, A. G. KARTSATOS, AND E. LITSYN, *Iterative solutions of unstable variational inequalities on approximately given sets*, Abstr. Appl. Anal., 1 (1996), pp. 45–54.
- [13] Y. I. ALBER AND Z. M. NASHED, *Iterative-projection regularization of ill-posed variational inequalities*, Analysis (Munich), 24 (2004), pp. 19–39.
- [14] Y. I. ALBER AND S. REICH, *An iterative method for solving a class of nonlinear operator equations in Banach spaces*, Panamer. Math. J., 4 (1994), pp. 39–54.
- [15] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, Boston, 1984.
- [16] J. P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.
- [17] A. B. BAKUSHINSKII AND B. T. POLJAK, *On the solution of variational inequalities*, Soviet Math. Dokl., 15 (1974), pp. 1705–1710.
- [18] H. H. BAUSCHKE AND J. M. BORWEIN, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.
- [19] H. H. BAUSCHKE AND J. M. BORWEIN, *Legendre functions and the method of random Bregman projections*, J. Convex Anal., 4 (1997), pp. 27–67.
- [20] H. H. BAUSCHKE, J. M. BORWEIN, AND P. L. COMBETTES, *Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces*, Commun. Contemp. Math., 3 (2001), pp. 615–647.
- [21] H. H. BAUSCHKE, J. M. BORWEIN, AND P. L. COMBETTES, *Bregman monotone optimization algorithms*, SIAM J. Control Optim., 42 (2003), pp. 596–636.
- [22] H. H. BAUSCHKE AND P. L. COMBETTES, *Construction of best Bregman approximations in reflexive Banach spaces*, Proc. Amer. Math. Soc., 131 (2003), pp. 3757–3766.
- [23] H. H. BAUSCHKE, P. L. COMBETTES, AND D. NOLL, *Joint minimization with alternating Bregman proximity operators*, Pac. J. Optim., 2 (2006), pp. 401–424.
- [24] L. M. BREGMAN, *The relaxation method for finding common points of convex sets and its application to the solution of problems in convex programming*, Comput. Math. Math. Phys., 7 (1967), pp. 200–217.
- [25] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
- [26] J. M. BORWEIN AND Q. J. ZHU, *Techniques of Variational Analysis*, Springer, New York, 2005.
- [27] F. E. BROWDER, *Multivalued monotone nonlinear mappings and duality mappings in Banach spaces*, Trans. Amer. Math. Soc., 118 (1965), pp. 338–351.
- [28] F. E. BROWDER, *Existence and approximation of solutions of nonlinear variational inequalities*, Proc. Nat. Acad. Sci. U.S.A., 56 (1966), pp. 1080–1086.
- [29] F. E. BROWDER, *Nonlinear operators and nonlinear equations of evolution in Banach spaces*, in Nonlinear Functional Analysis (Proc. Sympos. Pure Math., Vol. XVIII, Part 2, Chicago, IL, 1968), American Mathematical Society, Providence, RI, 1976, pp. 1–308.
- [30] R. E. BRUCK, *An iterative solution for a variational inequality for certain monotone operators in Hilbert space*, Bull. Amer. Math. Soc., 81 (1975), pp. 890–892. [Corrigendum: Bull. Amer. Math. Soc., 81 (1976), p. 353.]
- [31] R. S. BURACHIK AND S. SCHEIMBERG, *A proximal point method for the variational inequality problem in Banach spaces*, SIAM J. Control Optim., 39 (2000), pp. 1633–1649.
- [32] D. BUTNARIU AND A. N. IUSEM, *On a proximal point method for convex optimization in Banach spaces*, Numer. Funct. Anal. Optim., 18 (1997), pp. 723–744.
- [33] D. BUTNARIU AND A. N. IUSEM, *Totally Convex Functions for Fixed Points Computation and Infinite Dimensional Optimization*, Kluwer Academic Publishers, Dordrecht, 2000.
- [34] D. BUTNARIU, A. N. IUSEM, AND E. RESMERITA, *Total Convexity for Powers of the Norm in Uniformly Convex Banach Spaces*, J. Convex Anal., 7 (2000), pp. 319–334.
- [35] D. BUTNARIU, A. N. IUSEM, AND C. ZĂLINESCU, *On uniform convexity, total convexity and convergence of the proximal point and outer Bregman projection algorithms in Banach spaces*, J. Convex Anal., 10 (2003), pp. 35–61.

- [36] D. BUTNARIU AND G. KASSAY, *On classes of Legendre functions with special properties*, in preparation.
- [37] D. BUTNARIU AND E. RESMERITA, *Averaged subgradient methods for constrained convex optimization and Nash equilibria computation*, Optimization, 51 (2002), pp. 863–888.
- [38] D. BUTNARIU AND E. RESMERITA, *Bregman distances, totally convex functions, and a method for solving operator equations in Banach spaces*, Abstr. Appl. Anal., 2006, 84919.
- [39] D. BUTNARIU, S. REICH, AND A. ZASLAVSKI, *Weak convergence for orbits of nonlinear operators in reflexive Banach spaces*, Numer. Funct. Anal. Optim., 24 (2003), pp. 489–508.
- [40] A. CSASZAR, *General Topology*, Akademiai Kiado, Budapest, 1978.
- [41] A. L. DONTCHEV AND T. ZOLEZZI, *Well-Posed Optimization Problems*, Springer-Verlag, Berlin, 1993.
- [42] J. ECKSTEIN, *Nonlinear proximal point algorithms using Bregman functions, with application to convex programming*, Math. Oper. Res., 18 (1993), pp. 202–226.
- [43] P. P. B. EGGERMONT, *Multiplicative iterative algorithms for convex programming*, Linear Algebra Appl., 130 (1990), pp. 25–42.
- [44] J. ERIKSSON, *An Interval Primal-Dual Algorithm for Linear Programming*, Technical report 85-10, Department of Mathematics, Linköping University, Linköping, Sweden, 1985.
- [45] S. ERLANDER, *Entropy in linear programs*, Math. Program., 21 (1981), pp. 137–151.
- [46] Y. M. ERMOLIEV, *Methods for solving nonlinear extremal problems*, Kibernetika (Kiev), 1 (1966), pp. 1–17 (Russian).
- [47] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vol. 1, Springer-Verlag, New York, 2003.
- [48] K. GOEBEL AND S. REICH, *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings*, Marcel Dekker, New York, Basel, 1984.
- [49] E. G. GOLHSTEIN AND N. TRETYAKOV, *Modified Lagrangians and Monotone Maps in Optimization*, Wiley, New York, 1996.
- [50] G. KASSAY, *The proximal point algorithm for reflexive Banach spaces*, Studia Math., 30 (1985), pp. 9–17.
- [51] M. A. KRASNOSELSKII, *Two observations about the method of successive approximations*, Uspekhi Mat. Nauk, 10 (1955), pp. 123–127 (in Russian).
- [52] E. KREYSZIG, *Introductory Functional Analysis*, John Wiley & Sons, New York, 1978.
- [53] L. LANDWEBER, *An iterative formula for Fredholm integral equations of the first kind*, Amer. J. Math., 73 (1951), pp. 615–624.
- [54] B. LEMAIRE, *The proximal algorithm*, in New Methods in Optimization and Their Industrial Uses, Internat. Schriftenreihe Numer. Math. 87, J. P. Penot, ed., Birkhäuser, Basel, 1989, pp. 73–87.
- [55] O. A. LISKOVETS, *External approximations for the regularization of monotone variational inequalities*, Soviet Math. Dokl., 36 (1988), pp. 220–224.
- [56] O. A. LISKOVETS, *Regularized variational inequalities with pseudo-monotone operators on approximately given sets*, Differential Equations, 11 (1989), pp. 1970–1977 (in Russian).
- [57] B. MARTINET, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–158.
- [58] B. MARTINET, *Algorithmes pour la résolution des problèmes d'optimisation et minimax*, Thèse d'état, Université de Grenoble, Grenoble, France, 1972.
- [59] J.-J. MOREAU, *Fonctions convexes duales et points proximaux dans un espace Hilbertien*, C. R. Acad. Sci. Paris, 255 (1962), pp. 2897–2899.
- [60] J.-J. MOREAU, *Propriétés des applications 'prox'*, C. R. Acad. Sci. Paris, 256 (1963), pp. 1069–1071.
- [61] J.-J. MOREAU, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
- [62] U. MOSCO, *Convergence of convex sets and of solutions of variational inequalities*, Adv. Math., 3 (1969), pp. 510–585.
- [63] Z. OPJAL, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.
- [64] D. PASCALI AND S. SBURLAN, *Nonlinear Mappings of Monotone Type*, Martinus Nijhoff, The Hague; Sijthoff & Noordhoff International Publishers, Alphen ann den Rijn, 1978.
- [65] R. R. PHELPS, *Convex Functions, Monotone Operators, and Differentiability*, 2nd ed., Springer-Verlag, Berlin, 1993.
- [66] B. T. POLYAK, *A general method for solving extremum problems*, Dokl. Akad. Nauk SSSR, 174 (1967), pp. 593–597.
- [67] B. T. POLYAK, *Introduction to Optimization*, Optimization Software, Inc., New York, 1987.

- [68] R. T. ROCKAFELLAR, *Level sets and continuity of conjugate convex functions*, Trans. Amer. Math. Soc., 123 (1966), pp. 46–63.
- [69] R. T. ROCKAFELLAR, *On the maximality of sums of nonlinear operators*, Trans. Amer. Math. Soc., 49 (1970), pp. 75–88.
- [70] R. T. ROCKAFELLAR, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
- [71] R. T. ROCKAFELLAR, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
- [72] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
- [73] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1991.
- [74] A. RUSZCZYŃSKI, *A merit function approach of the subgradient method*, Optim. Methods Softw., 23 (2008), pp. 161–172.
- [75] N. Z. SHOR, *Application of the method of gradient descent to the solution of the network transportation problem*, in “Materials of the Scientific Seminar on Theoretical and Applied Questions of Cybernetics and Operation Research”, Ukrainian Academy of Sciences, Kiev, 1962, pp. 1–17 (Russian).
- [76] N. Z. SHOR, *Minimization Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin, 1985.
- [77] A. N. TIKHONOV, *Regularization of incorrectly posed problems*, Dokl. Akad. Nauk, 4 (1963), pp. 1035–1038.
- [78] K. YOSIDA, *Lectures on Differential and Integral Equations*, Interscience Publishers, London, 1960.

# ALMOST SURE STABILITY OF DISCRETE-TIME SWITCHED LINEAR SYSTEMS: A TOPOLOGICAL POINT OF VIEW\*

XIONGPING DAI<sup>†</sup>, YU HUANG<sup>‡</sup>, AND MINGQING XIAO<sup>§</sup>

**Abstract.** In this paper, we study the stability of discrete-time switched linear systems via symbolic topology formulation and the multiplicative ergodic theorem. A sufficient and necessary condition for  $\mu_A$ -almost sure stability is derived, where  $\mu_A$  is the Parry measure of the topological Markov chain with a prescribed transition (0,1)-matrix  $A$ . The obtained  $\mu_A$ -almost sure stability is invariant under small perturbations of the system. The topological description of stable processes of switched linear systems in terms of Hausdorff dimension is given, and it is shown that our approach captures the maximal set of stable processes for linear switched systems. The obtained results cover the stochastic Markov jump linear systems, where the measure is the natural Markov measure defined by the transition probability matrix. Two examples are provided to illustrate the theoretical outcomes of the paper.

**Key words.** discrete-time switched linear system, topological Markov chain, almost sure stability, Lyapunov exponent, Hausdorff dimension

**AMS subject classifications.** 93C55, 93D05, 93D09

**DOI.** 10.1137/070699676

## 1. Introduction.

**1.1. Motivation.** A switched linear system consists of a family of linear subsystems and a rule that governs the switching among them. These types of models are found in many practical systems in which switching is necessary and essential as the system dynamics evolve. More specifically, we consider the discrete-time dynamical system in the form of

$$(1.1) \quad x_{\ell+1} = H_{\omega_\ell} x_\ell, \quad \ell \geq 0,$$

where  $x_\ell \in \mathbb{R}^n$  and  $n \geq 2$  is a fixed integer,  $\omega_\ell$  takes a value in a given finite-symbolic set, say  $\mathcal{A} = \{1, \dots, \kappa\}$ , and  $H_i \in \mathbb{R}^{n \times n}$  for  $i \in \mathcal{A}$ . Let us denote the nonnegative integer set by  $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$  and the set of all mappings  $\mathbb{Z}_+ \rightarrow \mathcal{A}$  by

$$(1.2) \quad \Sigma_\kappa = \{\omega: \mathbb{Z}_+ \rightarrow \mathcal{A}\}.$$

Then switching can be classified into two situations: (i) arbitrary switching; i.e., the switching rule is characterized by  $\Sigma_\kappa$  defined by (1.2); (ii) switching is subject to certain constraints; i.e., the switching rule is characterized by a subset of  $\Sigma_\kappa$ .

Stability is the primary concern for switched systems. The analysis of its stability is much more difficult and challenging than that of linear systems. When arbitrary

---

\*Received by the editors August 8, 2007; accepted for publication (in revised form) March 16, 2008; published electronically July 16, 2008. This project was supported in part by NSFC (10671088, 10771222) and 973 project (2006CB805903), in part by NSF of Guangdong, and in part by NSF DMS-0605181 of the U.S.

<http://www.siam.org/journals/sicon/47-4/69967.html>

<sup>†</sup>Department of Mathematics, Nanjing University, Nanjing 210093, People's Republic of China (xpdai@nju.edu.cn).

<sup>‡</sup>Department of Mathematics, Zhongshan (Sun Yat-Sen) University, Guangzhou 510275, People's Republic of China (stshyu@mail.sysu.edu.cn).

<sup>§</sup>Corresponding author. Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408 (mxiao@math.siu.edu).

switching is considered, a switched system is said to be asymptotically stable if *all* its trajectories converge to the origin. This is also called absolute stability, and it requires that all infinite products of matrices taken from  $\{H_1, H_2, \dots, H_\kappa\}$  converge to zero. That is,  $\lim_{\ell \rightarrow \infty} \prod_{j=0}^{\ell-1} H_{\omega_j} = 0$  for any index sequence  $\{\omega_0, \omega_1, \dots\}$  with  $\omega_j \in \mathcal{A}$ . This can be equivalently stated by requiring the joint spectral radius of  $\{H_1, H_2, \dots, H_\kappa\}$  to be strictly less than one [3]. To the best of our knowledge, the study of stability of switched linear systems (1.1) has been focused on absolute stability (for example, see [13, 19, 29, 30, 32, 34, 35] and references therein). Currently available approaches for showing absolute stability of switched systems are essentially based on the search of common Lyapunov functions or variations of the same framework. The existence of a common Lyapunov function for a given switched system is quite restrictive, since an expecting common Lyapunov function has to guarantee that the energy of the overall system decreases to zero along all possible state trajectories governed by switches. Moreover, some critical situations are not able to be addressed by using the Lyapunov function approach. For instance, it is well known that the system (1.1) may not be absolutely stable even if each  $H_i, i \in \mathcal{A}$ , is asymptotically stable (i.e., all eigenvalues of  $H_i$  are inside the unit circle) (e.g., see [9]). A switched system that is not absolutely stable does not imply the end of stability analysis of the system. For example, for the stochastic Markov jump systems, almost sure stability (instead of absolute stability) plays a key role in the study of these types of systems, since it provides important convergence information in an “average” sense (under appropriate probability measures) which has been proved to be very useful and effective in applications (see [5, 6, 15, 16, 17, 20, 21, 23, 22, 26, 24, 27, 31, 37] and references therein). Another situation is when switching is subject to a subset of  $\Sigma_\kappa$  (so-called admissible switching set); in this case, how to identify the stability of (1.1) has not been clearly characterized yet.

The main challenge for the study of switched systems results from the switched paths that are arbitrary, although it may be subject to some constraints. The switched mechanism basically is uncertain, and the stability analysis has to cover all possible switchings (jumps). With such an uncertainty, the condition of absolute stability of switched systems is hardly met as the number of switches increases (i.e.,  $\kappa$  is getting large). Thus, to look for condition(s) of “almost” stability instead of absolute stability becomes more realistic in real applications.

In this paper we apply the ergodic theorem from the topology point of view to discuss the stability of (1.1), an approach that is not available in the current literature. More specifically, we translate the problem (1.1) into a finite state topological Markov chain setting under the framework of symbolic topology formulation. Then we study the dynamics (1.1) by the Lyapunov exponents based on the corresponding topological Markov chain. The main mathematical tool is the multiplicative ergodic theorem, which is a fundamental theory for describing the qualitative behaviors of dynamical systems in terms of the statistical characterization. We derive a necessary and sufficient condition for  $\mu_A$ -almost sure stability of (1.1) in which  $\mu_A$  is the Parry measure, i.e., the unique measure with maximal entropy for the underlying setting. Moreover, we have shown that the almost sure stability is not altered under small linear perturbation of the system (1.1), which is important and critical for real applications. Furthermore, a topological description of stable processes of (1.1) in terms of Hausdorff dimension is given, and this finding illustrates the significance for choosing the measure  $\mu_A$  (the Parry measure). Our obtained results cover the stochastic Markov jump linear systems, where the measure is the natural Markov measure defined by the transition probability matrix. The proposed approach and the

obtained results of this paper provide a fresh point of view for the study of switched systems.

**1.2. Outlines.** The paper is organized as follows. In section 2, we transform the switched system (1.1) to a symbolic dynamical system under the framework of topology. Then system (1.1) with constraints is expressed equivalently as a one-sided topological Markov chain with a prescribed transition  $(0,1)$ -matrix  $A$ . The concept of almost sure stability is introduced, and two preliminary propositions are provided in this section. In section 3, a necessary and sufficient condition for  $\mu_A$ -almost sure stability of (1.1) is presented, and a topological description of stable processes of (1.1) in terms of Hausdorff dimension is given. Section 4 addresses the connection between our obtained results and those for stochastic Markov jump linear systems. Two examples, one for arbitrary switching and another for restricted switching, are provided to illustrate theoretical results of the paper in section 5. The paper ends with concluding remarks.

**1.3. Matrix norms.** Let  $H = (h_{ij})$  be an  $n \times n$  matrix of real numbers. Throughout this paper the *norm*,  $\|H\|$ , of  $H$  can be either  $\|H\|_F$ , or  $\|H\|_1$ , or  $\|H\|_\infty$ , whose definitions are, respectively,

$$\|H\|_F = \sqrt{\sum_{i,j=1}^n |h_{ij}|^2}, \quad \|H\|_1 = \sum_{i,j=1}^n |h_{ij}|, \quad \text{and} \quad \|H\|_\infty = \max_{1 \leq i,j \leq n} |h_{ij}|.$$

It is known that these norms satisfy the following properties:

$$\begin{aligned} (1.3a) \quad & \|H + H'\| \leq \|H\| + \|H'\|, \\ (1.3b) \quad & \|HH'\| \leq \|H\|\|H'\|, \\ (1.3c) \quad & \|Hx\| \leq \|H\|\|x\|, \end{aligned}$$

where  $H'$  is also an  $n \times n$  matrix, and  $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$  is an  $n$ -dimensional vector. In particular (1.3c) holds for  $\|H\|$  in terms of corresponding induced norms in  $\mathbb{R}^n$  by  $\|x\| = \sqrt{\sum_{i=1}^n |x_i|^2}$ , or  $\sum_{i=1}^n |x_i|$ , or  $\max_{1 \leq i \leq n} |x_i|$ , respectively.

**2. Symbolic topology formulation.** We use a symbolic string  $\omega = (\omega_0 \omega_1 \dots)$  with  $\omega_j \in \mathcal{A} = \{1, \dots, \kappa\}$  to represent a specific switching path of (1.1). The set of all possible switching paths  $\omega = (\omega_0 \omega_1 \dots)$  is the  $\kappa$ -dimensional one-sided symbolic space given by (1.2), that is,

$$\Sigma_\kappa = \{\omega = (\omega_0 \omega_1 \dots) \mid \omega_i \in \mathcal{A} \text{ for } i = 0, 1, 2, \dots\},$$

which is a compact metric space endowed with the usual distance function

$$(2.1a) \quad \text{dist}(\omega, \omega') = \rho^{-n(\omega, \omega')} \quad \forall \omega, \omega' \in \Sigma_\kappa,$$

where  $\rho > 1$  is any prescribed constant and

$$(2.1b) \quad n(\omega, \omega') = \inf\{\ell \in \mathbb{Z}_+ \mid \omega_\ell \neq \omega'_\ell\}.$$

If  $\omega_\ell = \omega'_\ell$  for all nonnegative integers  $\ell$ , then  $n(\omega, \omega') := +\infty$ .

In order to include those cases in which switching constraints exist, let  $A = [a_{ij}]$  be an irreducible  $(0,1)$ -matrix of size  $\kappa \times \kappa$ , which is predefined. That is,  $a_{ij}$  equals



either 0 or 1, and for any pair  $(i, j)$  there is some integer  $n > 0$  such that  $a_{ij}^{(n)} > 0$ , where  $a_{ij}^{(n)}$  is the  $(i, j)$ th element of  $A^n$ . The *admissible set*  $\Sigma_A$  is defined to be

$$\Sigma_A = \{ \omega = (\omega_0 \omega_1 \cdots) \in \Sigma_\kappa \mid a_{\omega_\ell \omega_{\ell+1}} = 1 \text{ for } \ell = 0, 1, \dots \}.$$

Thus it is clear that in general  $\Sigma_A \subset \Sigma_\kappa$ . For the arbitrary switching case where the matrix  $A$  satisfies  $a_{ij} \equiv 1$  for any  $1 \leq i, j \leq \kappa$ , we have  $\Sigma_A = \Sigma_\kappa$ . Clearly the matrix  $A$  contains transition information of all admissible paths, and thus it is usually called a transition matrix.

It is not difficult to see that if  $(\omega_0 \omega_1 \omega_2 \cdots) \in \Sigma_A$ , then we have  $(\omega_1 \omega_2 \cdots) \in \Sigma_A$ . The  $\Sigma_A$ -invariant mapping

$$\sigma_A: \Sigma_A \rightarrow \Sigma_A; \quad (\omega_0 \omega_1 \cdots) \mapsto (\omega_1 \omega_2 \cdots)$$

is called the one-sided shift defined by the transition matrix  $A$ . The dynamical system  $(\Sigma_A, \sigma_A)$  is said to be a *one-sided topological Markov chain* with the transition matrix  $A$ , which is a compact subsystem of the one-sided full-shift dynamical system  $(\Sigma_\kappa, \sigma)$ , where  $\sigma: \Sigma_\kappa \rightarrow \Sigma_\kappa$  is similarly defined by  $(\omega_0 \omega_1 \cdots) \mapsto (\omega_1 \omega_2 \cdots)$  for any  $\omega = (\omega_0 \omega_1 \cdots) \in \Sigma_\kappa$ .

We next define a random matrix over  $\Sigma_A$  associated with the system (1.1) by

$$S: \Sigma_A \rightarrow \{H_1, \dots, H_\kappa\}; \quad \omega \mapsto S(\omega) = H_{\omega_0} \quad \forall \omega = (\omega_0 \omega_1 \cdots) \in \Sigma_A.$$

Let us denote for any  $t \in \mathbb{N}$  and for any  $\omega \in \Sigma_A$

$$(2.2a) \quad \sigma_A^t = \overbrace{\sigma_A \circ \cdots \circ \sigma_A}^{t \text{ times}} \quad \text{and} \quad \sigma_A^0 = id: \Sigma_A \rightarrow \Sigma_A$$

and

$$(2.2b) \quad S(\omega, t) = S(\sigma_A^{t-1} \omega) \cdots S(\omega): \mathbb{R}^n \rightarrow \mathbb{R}^n.$$

Here  $S(\cdot, \cdot)$  is called a *linear cocycle* based on  $(\Sigma_A, \sigma_A)$ . Notice that  $S(\sigma_A^\ell \omega) = H_{\omega_\ell}$  for any  $\omega = (\omega_0 \omega_1 \cdots)$  and  $\ell \in \mathbb{Z}_+$ . Then the system

$$(2.3) \quad x_{\ell+1} = S(\sigma_A^\ell \omega) x_\ell, \quad \text{where } \omega \in \Sigma_A,$$

can be regarded as a hybrid linear system with Markovian switchings. Thus for a specific path  $\omega$  and an initial condition  $x_0$ , according to (2.2), the state of (2.3) can be expressed as

$$(2.4a) \quad x_{\ell+1} = S(\omega, \ell + 1) x_0.$$

In fact  $S(\cdot, \cdot)$  describes the discrete-time linear skew-product flow  $\varphi_{\sigma_A, S}$  based on ergodic system  $(\Sigma_A, \sigma_A)$  in the following sense:

$$(2.4b) \quad \varphi_{\sigma_A, S}: \Sigma_A \times \mathbb{R}^n \times \mathbb{Z}_+ \rightarrow \Sigma_A \times \mathbb{R}^n; \quad (\omega, x, t) \mapsto (\sigma_A^t \omega, S(\omega, t)x).$$

For a given one-sided topological Markov chain  $(\Sigma_A, \sigma_A)$  with a transition matrix  $A$ , one always can define an invariant measure  $\mu$  under the one-sided shift  $\sigma_A$  from the classical Krylov–Bogolioubov theorem [36]. Now we are ready to introduce the definition of  $\mu$ -almost sure stability of (1.1).

DEFINITION 1. Let  $A$  be an irreducible transition matrix and  $\mu$  be an ergodic  $\sigma_A$ -invariant Borel probability measure on  $\Sigma_A$ ; namely,  $\mu(\sigma_A^{-1}B) = \mu(B)$  for any Borel subset  $B$  of  $\Sigma_A$ , and  $\mu(B) = 0$  or  $1$  whenever  $\sigma_A^{-1}B = B$  holds  $\mu$ -mod  $0$ .<sup>1</sup> The switched linear system (1.1) is said to be “ $\mu$ -almost surely stable” with respect to an admissible set  $\Sigma_A$  if (1.1) is exponentially stable for  $\mu$ -almost all switching sequences  $\omega$  in  $\Sigma_A$ . This is equivalent to saying that, for  $\mu$ -a.e. (almost everywhere)  $\omega \in \Sigma_A$ , we have

$$\lim_{\ell \rightarrow \infty} \frac{1}{\ell} \ln \|x_\ell\| = \lim_{\ell \rightarrow \infty} \frac{1}{\ell} \ln \|S(\omega, \ell)x_0\| < 0 \quad \forall x_0 \in \mathbb{R}^n.$$

Since  $\Sigma_A$  is compact under the metric given by (2.1), the set of all  $\sigma_A$ -invariant Borel probability measures on  $\Sigma_A$  is nonempty and is a compact convex set. It contains many elements. The question arises as to which is the most suitable member for the study of almost sure stability. In section 3, we shall show that the Parry measure, which has a strong ergodic property, is what we look for.

In order to introduce the concept of Lyapunov exponent, we need to show that the following proposition holds.

PROPOSITION 1. Let  $S(\cdot, \cdot)$  be the linear cocycle as in (2.2b) based on the ergodic system  $(\Sigma_A, \mu, \sigma_A)$ . Then we have

$$(2.5) \quad \lim_{t \rightarrow \infty} \frac{1}{t} \int_{\Sigma_A} \ln \|S(\omega, t)\| d\mu(\omega) = \inf_{t \in \mathbb{N}} \frac{1}{t} \int_{\Sigma_A} \ln \|S(\omega, t)\| d\mu(\omega).$$

*Proof.* Let us first define

$$f_t = \int_{\Sigma_A} \ln \|S(\omega, t)\| d\mu(\omega).$$

For any  $t_1, t_2 \in \mathbb{Z}_+$ , according to (2.2), for any  $\omega \in \Sigma_A$  one has

$$\begin{aligned} \|S(\omega, t_1 + t_2)\| &= \|S(\sigma_A^{t_1}\omega, t_2)S(\omega, t_1)\| \\ &\leq \|S(\sigma_A^{t_1}\omega, t_2)\| \|S(\omega, t_1)\|. \end{aligned}$$

The property  $\sigma_A$ -invariance of  $\mu$  implies

$$\int_{\Sigma_A} \ln \|S(\sigma_A^{t_1}\omega, t_2)\| d\mu(\omega) = \int_{\Sigma_A} \ln \|S(\omega, t_2)\| d\mu(\omega) = f_{t_2}.$$

Hence we have

$$\begin{aligned} f_{t_1+t_2} &\leq \int_{\Sigma_A} \ln \|S(\sigma_A^{t_1}\omega, t_2)\| d\mu(\omega) + \int_{\Sigma_A} \ln \|S(\omega, t_1)\| d\mu(\omega) \\ &= f_{t_1} + f_{t_2}. \end{aligned}$$

Therefore, the real sequence  $(f_t)_{t=0}^\infty$  is subadditive. From Kingman's subadditive ergodic theorem [36], it follows that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_{\Sigma_A} \ln \|S(\omega, t)\| d\mu(\omega) = \inf_{t \in \mathbb{N}} \frac{1}{t} \int_{\Sigma_A} \ln \|S(\omega, t)\| d\mu(\omega),$$

which completes the proof.  $\square$

<sup>1</sup>Two Borel sets “ $B = C$   $\mu$ -mod  $0$ ” means that  $\mu((B \setminus C) \cup (C \setminus B)) = 0$ . In addition, the  $\sigma_A$ -invariance of  $\mu$  is equivalent to  $\int_{\Sigma_A} f d\mu = \int_{\Sigma_A} f \circ \sigma_A d\mu$  for all  $f \in C(\Sigma_A)$ , which we will use in the proof of Proposition 1, where  $C(\Sigma_A)$  denotes the set of all continuous real functions defined on  $\Sigma_A$ .

As a convention, the real number

$$\lambda(\sigma_A, S, \mu) := \lim_{t \rightarrow \infty} \frac{1}{t} \int_{\Sigma_A} \ln \|S(\omega, t)\| d\mu(\omega)$$

is called the *Lyapunov exponent* of the cocycle  $S(\cdot, \cdot)$  based on the ergodic system  $(\Sigma_A, \mu, \sigma_A)$  [18]. According to the *multiplicative ergodic theorem* (see, for example, [33]), for  $\mu$ -a.e.  $\omega \in \Sigma_A$  we know

$$(2.6) \quad \lambda(\sigma_A, S, \mu) = \lim_{t \rightarrow \infty} \frac{1}{t} \ln \|S(\omega, t)\|,$$

where the limit does not depend on the choice of the path  $\omega$ . This leads to a necessary and sufficient condition of  $\mu$ -almost sure stability for system (2.3) (or, equivalently, (1.1) subject to the constraint  $\Sigma_A$ ).

**PROPOSITION 2.** *Let  $\mu$  be an ergodic probability measure with respect to  $(\Sigma_A, \sigma_A)$ . Then the system*

$$x_{\ell+1} = S(\sigma_A^\ell \omega) x_\ell \quad \forall \omega \in \Sigma_A$$

*is  $\mu$ -almost surely stable if and only if*

$$\lambda(\sigma_A, S, \mu) < 0.$$

*Here  $\sigma_A^\ell$  is defined in (2.2a).*

### 3. Stability and topological description of stable processes.

**3.1. Almost sure stability.** Following convention, we now define a *canonical Markov measure* generated by an irreducible  $(0, 1)$ -matrix  $A$  of size  $\kappa \times \kappa$ . Let us denote the spectral radius of the nonnegative matrix  $A$  by  $\rho_A$ . Then from the Perron–Frobenius theorem it follows that there are two positive vectors

$$(3.1a) \quad v = (v_1, \dots, v_\kappa)^T \quad \text{and} \quad u = (u_1, \dots, u_\kappa)$$

in  $\mathbb{R}^\kappa$  such that

$$(3.1b) \quad Av = \rho_A v \quad \text{and} \quad uA = \rho_A u$$

with

$$(3.1c) \quad \sum_{i=1}^{\kappa} u_i v_i = 1.$$

Let

$$(3.2a) \quad p_A = (p_1, \dots, p_\kappa) \quad \text{with} \quad p_i = u_i v_i \quad \text{for} \quad 1 \leq i \leq \kappa$$

and

$$(3.2b) \quad P_A = [p_{ij}], \quad \text{where} \quad p_{ij} = \frac{a_{ij} v_j}{\rho_A v_i} \quad \text{for} \quad 1 \leq i, j \leq \kappa.$$

Then we have  $p_A P_A = p_A$ . The matrix  $P_A = [p_{ij}]$  can be viewed as a transition probability matrix with  $p_{ij} = 0$  if and only if  $a_{ij} = 0$ . The canonical  $\sigma_A$ -invariant Markov measure  $\mu_A$  on  $\Sigma_A$  is derived as follows:

$$(3.3a) \quad \mu_A([i_0 \cdots i_\ell]_A) = p_{i_0} p_{i_0 i_1} \cdots p_{i_{\ell-1} i_\ell},$$

where

$$(3.3b) \quad [i_0 \cdots i_\ell]_A = \{\omega \in \Sigma_A \mid \omega_0 = i_0, \dots, \omega_\ell = i_\ell\}$$

is the *cylinder* defined by the word of length  $\ell + 1$  for  $(i_0 \cdots i_\ell) \in \mathcal{A}^{\ell+1}$  with any  $\ell + 1 \in \mathbb{N}$ . According to [36, Theorem 1.13] we know that

- (1)  $\mu_A$  is supported on  $\Sigma_A$  with  $\text{supp}(\mu_A) = \Sigma_A$ , where  $\text{supp}(\mu_A)$  means the minimal  $\sigma_A$ -invariant closed subset of  $\Sigma_A$  with  $\mu_A$ -measure 1.
- (2)  $\mu_A$  is an ergodic  $\sigma_A$ -invariant Borel probability measure on  $\Sigma_A$ .

*Remark 1.* Compared to the stochastic Markovian, the above formulation is more general since the transition rates (or the generator) of the Markov chain can be arbitrary due to the fact that it is not necessary for the transition matrix  $A$  to be a probability matrix. We now derive the first main result of this paper.

**THEOREM 1.** *Consider the switched linear system (1.1) with the switching sequence belonging to a topological Markov shift  $(\Sigma_A, \sigma_A)$ . For any  $t \in \mathbb{N}$ , write*

$$(3.4) \quad \lambda_{i_0 i_1 \cdots i_{t-1}} = \|H_{i_{t-1}} \cdots H_{i_0}\| \quad \forall (i_0 \cdots i_{t-1}) \in \mathcal{A}^t,$$

where  $\|\cdot\|$  is the matrix norm and where  $\mathcal{A}^t = \overbrace{\mathcal{A} \times \cdots \times \mathcal{A}}^{t \text{ times}}$ . Let  $p_A = (p_1, \dots, p_\kappa)$ ,  $P_A = [p_{ij}]$  and the measure  $\mu_A$  be defined by (3.2) and (3.3), respectively. Then (1.1) is  $\mu_A$ -almost surely stable if and only if there is at least one  $\hat{t} \in \mathbb{N}$  such that

$$(3.5) \quad \prod_{(i_0 \cdots i_{\hat{t}-1}) \in \mathcal{A}^{\hat{t}}} \lambda_{i_0 i_1 \cdots i_{\hat{t}-1}}^{p_{i_0} p_{i_0 i_1} \cdots p_{i_{\hat{t}-2} i_{\hat{t}-1}}} < 1.$$

Moreover, if (1.1) is  $\mu_A$ -almost surely stable, then there exists some  $\varepsilon > 0$  such that every switched linear system

$$(1.1)' \quad x_{\ell+1} = H'_{\omega_\ell} x_\ell, \quad \ell \geq 0,$$

is also  $\mu_A$ -almost surely stable on  $(\Sigma_A, \sigma_A)$  whenever

$$(3.6) \quad \|H_i - H'_i\| \leq \varepsilon, \quad 1 \leq i \leq \kappa.$$

*Proof.* Consider the discrete-time flow generated by system (2.3)

$$\varphi_{\sigma_A, S}: \Sigma_A \times \mathbb{R}^n \times \mathbb{Z}_+ \rightarrow \Sigma_A \times \mathbb{R}^n$$

defined by

$$(\omega, x, t) \mapsto (\sigma_A^t \omega, S(\omega, t)x) = (\sigma_A^t \omega, H_{\omega_{t-1}} \cdots H_{\omega_0} x)$$

for any  $t \in \mathbb{Z}_+$ ,  $x \in \mathbb{R}^n$  and for any  $\omega = (\omega_0 \omega_1 \cdots) \in \Sigma_A$ .

Since (1.1) subject to the constraint  $\Sigma_A$  and (2.3) represent the same systems, by Proposition 2, the  $\mu_A$ -almost sure stability of (1.1) means that the Lyapunov exponent  $\lambda(\sigma_A, S, \mu_A)$  of  $\varphi_{\sigma_A, S}$  associated with  $\mu_A$  is strictly negative.

Assume that  $\lambda(\sigma_A, S, \mu_A) < 0$  holds. Then, it follows from Proposition 1 that there is some  $\hat{t} \in \mathbb{N}$  which is such that

$$\frac{1}{\hat{t}} \int_{\Sigma_A} \ln \|S(\omega, \hat{t})\| d\mu_A(\omega) < 0.$$

Thus, as  $\Sigma_A = \bigcup \{[i_0 \cdots i_{\hat{t}-1}]_A : (i_0 \cdots i_{\hat{t}-1}) \in \mathcal{A}^{\hat{t}}\}$ , where  $[i_0 \cdots i_{\hat{t}-1}]_A$  is defined with respect to  $A$ , and  $[i_0 \cdots i_{\hat{t}-1}]_A$  are disjoint of each other, (3.5) follows from (2.2) and (3.3).

Conversely, let (3.5) hold for some  $\hat{t} \in \mathbb{N}$ . Then we obtain

$$\begin{aligned} 0 &> \sum_{(i_0 \cdots i_{\hat{t}-1}) \in \mathcal{A}^{\hat{t}}} p_{i_0} p_{i_0 i_1} \cdots p_{i_{\hat{t}-2} i_{\hat{t}-1}} \ln \lambda_{i_0 i_1 \cdots i_{\hat{t}-1}} \\ &= \sum_{(i_0 \cdots i_{\hat{t}-1}) \in \mathcal{A}^{\hat{t}}} \int_{[i_0 \cdots i_{\hat{t}-1}]_A} \ln \|S(\omega, \hat{t})\| d\mu_A(\omega) \\ &= \int_{\Sigma_A} \ln \|S(\omega, \hat{t})\| d\mu_A(\omega), \end{aligned}$$

which implies that

$$0 > \frac{1}{\hat{t}} \int_{\Sigma_A} \ln \|S(\omega, \hat{t})\| d\mu_A(\omega) \geq \lambda(\sigma_A, S, \mu_A)$$

holds.

Since  $p_A = (p_1, \dots, p_\kappa)$ ,  $P_A = [p_{ij}]$ , and  $\mu_A$  all are independent of  $\{H_1, \dots, H_\kappa\}$ , and the index

$$\lambda_{i_0 i_1 \cdots i_{\hat{t}-1}} = \|H_{i_{\hat{t}-1}} \cdots H_{i_0}\| \quad \forall (i_0 \cdots i_{\hat{t}-1}) \in \mathcal{A}^{\hat{t}}$$

is continuous with respect to  $(H_1, \dots, H_\kappa)$ , the robustness follows directly from the criterion (3.5). Thus the proof is completed.  $\square$

*Remark 2.* The ergodic  $\sigma_A$ -invariant measure  $\mu_A$  defined by (3.2)–(3.3) is called the “Parry measure” of the topological Markov chain  $(\Sigma_A, \sigma_A)$ . It is a Gibbs measure which has the maximal entropy, namely,  $h_{top}(\sigma_A) = h_{\mu_A}(\sigma_A)$ , and such a property can characterize other measures in a “maximal” way (for more detail see section 3.3 below). The proof of Theorem 1 holds as long as  $\mu_A$  is an ergodic  $\sigma_A$ -invariant Borel probability measure on  $\Sigma_A$  with  $\text{supp}(\mu_A) = \Sigma_A$ .

Theorem 1 gives a necessary and sufficient condition for  $\mu_A$ -almost sure stability of topological Markov jump system (2.3). In particular, in view of (3.5), sufficient condition can be tested by finding  $\hat{t}$  satisfying (3.5).

The following subsections provide a topological description of the set of stable processes.

**3.2. Hausdorff dimension.** Let  $X$  be a compact metric space with a metric  $d(\cdot, \cdot)$ . We denote

$$B(x, r) = \{y \in X \mid d(x, y) \leq r\},$$

which stands for a closed ball centered at  $x \in X$  with radius  $r > 0$ . Recall that for  $s \geq 0$  the  $s$ -Hausdorff measure of  $Y \subseteq X$  for a given metric  $d(\cdot, \cdot)$  is defined as

$$\mathcal{H}_d^s(Y) = \liminf_{\delta \rightarrow 0} \left\{ \sum_i |A_i|^s : \bigcup_i A_i \supseteq Y \text{ and } \sup_i |A_i| < \delta \right\},$$

where  $\{A_i\}$  is a countable cover of  $Y$  and  $|A_i|$  represents the diameter of  $A_i$  in terms of  $d(\cdot, \cdot)$ . Furthermore, the *Hausdorff dimension* of  $Y$  is defined by

$$HD_d(Y) = \inf \{s \in \mathbb{R}_+ : \mathcal{H}_d^s(Y) = 0\}.$$

For a given Borel probability measure  $\mu$  on  $X$ , the *Hausdorff dimension* of  $\mu$  is given by

$$HD_d(\mu) = \inf \{HD_d(Y) : Y \text{ is a Borel subset of } X \text{ with } \mu(Y) = 1\}.$$

The nonnegative number  $HD_d(\cdot)$ , which depends upon the given metric  $d(\cdot, \cdot)$ , in general satisfies only

$$(3.7a) \quad HD_d(\mu) \leq HD_d(\text{supp}(\mu))$$

and

$$(3.7b) \quad HD_d(Y) \leq HD_d(Y') \quad \text{if } Y \subseteq Y';$$

for instance, we refer readers to [14] for more details. Here we state some important facts (see [38]) that will be needed for the proof of our next theorem.

- (a) Let  $\mu$  be a Borel probability measure on  $X$  and  $E$  a Borel subset of  $X$  with  $\mu(E) > 0$ . If  $\underline{\delta} \leq \liminf_{r \rightarrow 0} \frac{\ln \mu(B(x, r))}{\ln r} \leq \limsup_{r \rightarrow 0} \frac{\ln \mu(B(x, r))}{\ln r} \leq \bar{\delta}$  for all  $x \in E$ , then  $\underline{\delta} \leq HD_d(E) \leq \bar{\delta}$ .

For any  $\delta > 0$ , let  $N(X, \delta)$  denote the minimal number of  $\delta$ -ball  $B(x, \delta)$  needed to cover  $X$ . We define the *lower* and *upper box-dimension* of  $X$ , respectively, as

$$\underline{\dim}_B(X) = \liminf_{\delta \rightarrow 0} \frac{\ln N(X, \delta)}{-\ln \delta}$$

and

$$\overline{\dim}_B(X) = \limsup_{\delta \rightarrow 0} \frac{\ln N(X, \delta)}{-\ln \delta}.$$

If  $\underline{\dim}_B(X) = \overline{\dim}_B(X)$ , denoted by  $\dim_B(X)$ , then we call it the *box-dimension* of  $X$  with respect to  $d(\cdot, \cdot)$ . It is well known from [14] that the following holds.

- (b)  $HD_d(X) \leq \underline{\dim}_B(X)$ .

**3.3. Topological entropy.** Let  $f: (X, \mu) \rightarrow (X, \mu)$  be a continuous transformation of a compact metrizable space  $X$  which preserves a Borel probability measure  $\mu$ . The topological dynamical system  $(X, f)$  is said to be *Li-Yorke chaotic* [28, 2] provided that there is an uncountable subset  $X_0$  of  $X$  such that

$$\limsup_{\ell \rightarrow \infty} d(f^\ell x, f^\ell y) > 0 \quad \forall x, y \in X_0, \quad x \neq y,$$

and

$$\liminf_{\ell \rightarrow \infty} d(f^\ell x, f^\ell y) = 0 \quad \forall x, y \in X_0, \ x \neq y.$$

In order to quantitatively measure the uncertainty, randomness, or disorder of the system  $(X, f)$ , in 1958, Kolmogorov [25] introduced the concept of *measure-theoretic entropy*, denoted by  $h_\mu(f)$ , to the measure-preserving system  $(X, \mu, f)$ . In 1965, Adler, Konheim, and McAndrew [1] introduced topological entropy, which is the analogous invariant for topological dynamical systems. It soon turned out that there is a simple relationship between these quantities: maximizing the metric entropy over a suitable class of measures defined on a dynamical system gives its topological entropy (see, for example, [36]). For the convenience of later discussion, we let

$$B_\ell^f(x, \delta) = \{y \in X \mid \text{dist}(f^i x, f^i y) \leq \delta \text{ for } 0 \leq i < \ell\}$$

for any  $x \in X$  and any  $\ell \in \mathbb{N}, \delta > 0$ . Clearly,  $B(x, \delta) = B_1^f(x, \delta)$ . Due to Brin and Katok [8], we know that the following holds.

- (c)  $h_\mu(f, x) := \lim_{\delta \rightarrow 0} \liminf_{\ell \rightarrow \infty} \frac{\ln \mu(B_\ell^f(x, \delta))}{-\ell} = \lim_{\delta \rightarrow 0} \limsup_{\ell \rightarrow \infty} \frac{\ln \mu(B_\ell^f(x, \delta))}{-\ell}$  holds for  $\mu$ -a.e.  $x \in X$ .
- (d)  $h_\mu(f, x)$  is  $f$ -invariant; namely,  $h_\mu(f, x) = h_\mu(f, f^\ell x)$  for all  $\ell > 0$  for  $\mu$ -a.e.  $x \in X$ .
- (e)  $h_\mu(f) = \int_X h_\mu(f, x) d\mu$ .
- (f) In particular, if  $\mu$  is ergodic, then  $h_\mu(f) = h_\mu(f, x)$  for  $\mu$ -a.e.  $x \in X$ .

Next, by  $N(X, \ell, \delta)$  we denote the minimal number of Bowen balls  $B_\ell^f(x, \delta)$  needed to cover  $X$ . Then, the *topological entropy* of  $(X, f)$ , written  $h_{\text{top}}(f)$ , is defined by the following:

$$(g) \quad h_{\text{top}}(f) = \lim_{\delta \rightarrow 0} \limsup_{\ell \rightarrow \infty} \frac{\ln N(X, \ell, \delta)}{\ell}.$$

According to the well-known variational principle of entropy [36], we have the following:

- (h)  $h_{\text{top}}(f) = \sup_\mu h_\mu(f)$ , where  $\mu$  ranges over all  $f$ -invariant ergodic Borel probability measures on  $X$ .

Note that if  $h_{\text{top}}(f) = h_\mu(f)$ ,  $\mu$  is called a *maximal entropy measure*. By virtue of (f) above, the maximal entropy measure  $\mu$  contains almost all complexity information of the system  $(X, f)$ .

Chaos and entropy are characteristics of the complexity of system  $(X, f)$  from two different viewpoints. Chaos is closely related to system behavior, while entropy focuses on “physical principles.” In general, they have the following relationships:

- (i)  $h_{\text{top}}(f) > 0$  implies that  $(X, f)$  is Li–Yorke chaotic [4].
- (j)  $h_{\text{top}}(\sigma_A) > 0$  if and only if  $(\Sigma_A, \sigma_A)$  is Li–Yorke chaotic [39].

**3.4. Stable processes.** Theorem 1 gives only a measure description of the set of stable switching sequences of system (1.1) on  $\Sigma_A$ . The next theorem provides a precisely topological description of the set of stable processes of (1.1), mainly motivated by [10, 11, 12].

**THEOREM 2.** *We consider the topological Markov jump linear system (2.3). Let*

$$\Sigma_{\text{stab}}(S; A) = \{\omega \in \Sigma_A \mid x_{\ell+1} = S(\sigma_A^\ell \omega) x_\ell \text{ is exponentially stable}\}.$$

*If  $\lambda(\sigma_A, S, \mu_A) < 0$ , then we have*

$$(3.8) \quad HD_\rho(\Sigma_{\text{stab}}(S; A)) = HD_\rho(\Sigma_A) = \frac{h_{\text{top}}(\sigma_A)}{\ln \rho}.$$

Here  $HD_{\rho}(\cdot)$  means the Hausdorff dimension under the metric  $\rho(\cdot, \cdot)$  defined by (2.1). Moreover,  $HD_{\rho}(\Sigma_{stab}(S; A)) > 0$  if and only if  $(\Sigma_A, \sigma_A)$  is Li-Yorke chaotic.

*Proof.* First note  $\Sigma_A \supseteq \Sigma_{stab}(S; A)$ . This implies  $HD_{\rho}(\Sigma_A) \geq HD_{\rho}(\Sigma_{stab}(S; A))$  from section 3.2. It follows from Proposition 2 and Definition 1 that  $\Sigma_{stab}(S; A)$  has  $\mu_A$ -measure 1. Hence the definition of  $HD_{\rho}(\mu_A)$  yields  $HD_{\rho}(\Sigma_{stab}(S; A)) \geq HD_{\rho}(\mu_A)$ . We thus obtain

$$(3.9) \quad HD_{\rho}(\Sigma_A) \geq HD_{\rho}(\Sigma_{stab}(S; A)) \geq HD_{\rho}(\mu_A).$$

Since  $\mu_A$  is the Parry measure, this implies  $h_{top}(\sigma_A) = h_{\mu_A}(\sigma_A)$ . In order to finish the proof, it is sufficient to show the following two equalities:

$$(3.10a) \quad HD_{\rho}(\mu_A) = \frac{h_{\mu_A}(\sigma_A)}{\ln \rho}$$

and

$$(3.10b) \quad HD_{\rho}(\Sigma_A) \leq \frac{h_{top}(\sigma_A)}{\ln \rho},$$

where  $h_{top}(\sigma_A)$  and  $h_{\mu_A}(\sigma_A)$  stand for the topological entropy and the measure-theoretic entropy with respect to  $\mu_A$  of the Markov chain  $(\Sigma_A, \sigma_A)$ , respectively.

To prove (3.10a), we first claim that for any ergodic  $\sigma_A$ -invariant Borel probability measure  $\mu$  on  $\Sigma_A$  we have

$$(3.11) \quad HD_{\rho}(\mu) = \frac{h_{\mu}(\sigma_A)}{\ln \rho}.$$

From (2.1a) we see that  $\sigma_A$  possesses the following so called “similarity property”:

$$(3.12) \quad \text{dist}(\sigma_A \omega, \sigma_A \omega') = \rho \text{dist}(\omega, \omega') \quad \text{if } \text{dist}(\omega, \omega') < \rho^{-2}.$$

In fact, given any  $\omega = (\omega_0 \omega_1 \omega_2 \cdots)$ ,  $\omega' = (\omega'_0 \omega'_1 \omega'_2 \cdots) \in \Sigma_A$ . If we have  $\text{dist}(\omega, \omega') < \rho^{-2}$ , then  $n(\omega, \omega') \geq 3$ , which implies  $\omega_0 = \omega'_0$ ,  $\omega_1 = \omega'_1$ , and  $\omega_2 = \omega'_2$ . Thus,  $n(\sigma_A \omega, \sigma_A \omega') = n(\omega, \omega') - 1$ , and thus  $\text{dist}(\sigma_A \omega, \sigma_A \omega') = \rho^{-n(\sigma_A \omega, \sigma_A \omega')} = \rho \rho^{-n(\omega, \omega')} = \rho \text{dist}(\omega, \omega')$ .

Now for given any  $i \geq 3$ , we have

$$(3.13) \quad [\omega_0 \cdots \omega_{i+\ell}]_A = B(\omega, \rho^{-(i+\ell)}) = B_{\ell+1}^{\sigma_A}(\omega, \rho^{-i}) \quad \forall \omega \in \Sigma_A \text{ and } \forall \ell \in \mathbb{N},$$

because of  $B(\sigma_A^j \omega, \rho^{-i}) = [\omega_j \cdots \omega_{j+i}]_A$  and

$$\begin{aligned} B_{\ell+1}^{\sigma_A}(\omega, \rho^{-i}) &= \bigcap_{j=0}^{\ell} \sigma_A^{-j} B(\sigma_A^j \omega, \rho^{-i}) \\ &= \bigcap_{j=0}^{\ell} \sigma_A^{-j} [\omega_j \cdots \omega_{j+i}]_A. \end{aligned}$$

Hence  $\mu(B(\omega, \rho^{-i-\ell})) = \mu(B_{\ell+1}^{\sigma_A}(\omega, \rho^{-i}))$ , and thus we have

$$\begin{aligned} \frac{\ln \mu(B(\omega, \rho^{-i-\ell}))}{\ln \rho^{-i-\ell}} &= \frac{\ln \mu(B_{\ell+1}^{\sigma_A}(\omega, \rho^{-i}))}{\ln \rho^{-i-\ell}} \\ &= \frac{\ln \mu(B_{\ell+1}^{\sigma_A}(\omega, \rho^{-i}))}{-(i+\ell) \ln \rho}, \end{aligned}$$



which implies that

$$\limsup_{\ell \rightarrow \infty} \frac{\ln \mu(B(\omega, \boldsymbol{\rho}^{-i-\ell}))}{\ln \boldsymbol{\rho}^{-i-\ell}} = \frac{1}{\ln \boldsymbol{\rho}} \limsup_{\ell \rightarrow \infty} \frac{\ln \mu(B_{\ell+1}^{\sigma_A}(\omega, \boldsymbol{\rho}^{-i}))}{-\ell}$$

and

$$\liminf_{\ell \rightarrow \infty} \frac{\ln \mu(B(\omega, \boldsymbol{\rho}^{-i-\ell}))}{\ln \boldsymbol{\rho}^{-i-\ell}} = \frac{1}{\ln \boldsymbol{\rho}} \liminf_{\ell \rightarrow \infty} \frac{\ln \mu(B_{\ell+1}^{\sigma_A}(\omega, \boldsymbol{\rho}^{-i}))}{-\ell}.$$

Thus we obtain

$$(3.14a) \quad \limsup_{\delta \rightarrow 0} \frac{\ln \mu(B(\omega, \delta))}{\ln \delta} = \frac{1}{\ln \boldsymbol{\rho}} \limsup_{\ell \rightarrow \infty} \frac{\ln \mu(B_{\ell+1}^{\sigma_A}(\omega, \boldsymbol{\rho}^{-i}))}{-\ell}$$

and

$$(3.14b) \quad \liminf_{\delta \rightarrow 0} \frac{\ln \mu(B(\omega, \delta))}{\ln \delta} = \frac{1}{\ln \boldsymbol{\rho}} \liminf_{\ell \rightarrow \infty} \frac{\ln \mu(B_{\ell+1}^{\sigma_A}(\omega, \boldsymbol{\rho}^{-i}))}{-\ell}.$$

Since  $i \geq 3$  is arbitrary, from statements (c) and (f) in section 3.3 and (3.14), we arrive at

$$(3.15) \quad \begin{aligned} \lim_{\delta \rightarrow 0} \frac{\ln \mu(B(\omega, \delta))}{\ln \delta} &= \frac{1}{\ln \boldsymbol{\rho}} \lim_{i \rightarrow \infty} \liminf_{\ell \rightarrow \infty} \frac{\ln \mu(B_{\ell+1}^{\sigma_A}(\omega, \boldsymbol{\rho}^{-i}))}{-\ell} \\ &= \frac{1}{\ln \boldsymbol{\rho}} \lim_{i \rightarrow \infty} \limsup_{\ell \rightarrow \infty} \frac{\ln \mu(B_{\ell+1}^{\sigma_A}(\omega, \boldsymbol{\rho}^{-i}))}{-\ell} \\ &= \frac{h_\mu(\sigma_A)}{\ln \boldsymbol{\rho}} \quad \mu\text{-a.e. } \omega \in \Sigma_A. \end{aligned}$$

Therefore, (3.11) follows directly from statement (a) in section 3.2.

On the other hand, from (3.13) we see that

$$N(\Sigma_A, \boldsymbol{\rho}^{-i-\ell}) = N(\Sigma_A, \ell + 1, \boldsymbol{\rho}^{-i}) \quad \forall \ell \in \mathbb{N}, i \geq 3.$$

Thus

$$(3.16) \quad \begin{aligned} \limsup_{\delta \rightarrow 0} \frac{\ln N(\Sigma_A, \delta)}{-\ln \delta} &= \limsup_{\ell \rightarrow 0} \frac{\ln N(\Sigma_A, \boldsymbol{\rho}^{-i-\ell})}{-\ln \boldsymbol{\rho}^{-i-\ell}} \\ &= \limsup_{\ell \rightarrow \infty} \frac{\ln N(\Sigma_A, \ell + 1, \boldsymbol{\rho}^{-i})}{(i + \ell) \ln \boldsymbol{\rho}} \\ &= \frac{1}{\ln \boldsymbol{\rho}} \limsup_{\ell \rightarrow \infty} \frac{\ln N(\Sigma_A, \ell + 1, \boldsymbol{\rho}^{-i})}{\ell}. \end{aligned}$$

For arbitrary  $i$ , according to statement (g) in section 3.3, (3.16) yields

$$(3.17) \quad \begin{aligned} \limsup_{\delta \rightarrow 0} \frac{\ln N(\Sigma_A, \delta)}{-\ln \delta} &= \frac{1}{\ln \boldsymbol{\rho}} \lim_{i \rightarrow \infty} \limsup_{\ell \rightarrow \infty} \frac{\ln N(\Sigma_A, \ell + 1, \boldsymbol{\rho}^{-i})}{\ell} \\ &= \frac{h_{top}(\sigma_A)}{\ln \boldsymbol{\rho}}, \end{aligned}$$

which implies that (3.10b) holds from statement (b) in section 3.2. Therefore, (3.10) holds.

Finally, from (3.8) and statement (j) in section 3.3, it follows directly that the Hausdorff dimension  $HD_{\rho}(\Sigma_{stab}(S; A))$  is strictly positive if and only if  $(\Sigma_A, \sigma_A)$  is Li–Yorke chaotic.

Thus, the proof is completed.  $\square$

*Remark 3.* Identity (3.8) implies that the  $\mu_A$ -measure defined in (3.3) (which is the Parry measure of  $(\Sigma_A, \sigma_A)$ ) is a desirable measure from the topological point of view since the “size” of the set of all stable paths is the same as the set of all admissible paths in the sense of Hausdorff dimension.

*Remark 4.* Suppose that there are two Parry measures of  $(\Sigma_{\kappa}, \sigma)$  such that system (1.1) is  $\mu_{A'}$ - and  $\mu_A$ -almost surely stable with respect to two admissible sets  $\Sigma_{A'}$  and  $\Sigma_A$ , respectively. Assume that  $h_{top}(\sigma_A) \geq h_{top}(\sigma_{A'})$ . Then according to Theorem 2, we have

$$(3.18) \quad HD_{\rho}(\Sigma_{stab}(S; A)) = \frac{h_{top}(\sigma_A)}{\ln \rho} \geq \frac{h_{top}(\sigma_{A'})}{\ln \rho} = HD_{\rho}(\Sigma_{stab}(S; A')).$$

Hence, larger topological entropy of  $\sigma_A$  results in a larger set of stable paths of (1.1).

*Remark 5.* According to [39], if the underlying setting  $(\Sigma_A, \sigma_A)$  is Li–Yorke chaotic, then it has positive entropy. Thus the more Li–Yorke chaotic  $(\Sigma_A, \sigma_A)$  behaves, the larger the set of  $\mu_A$ -almost surely stable paths the switched linear system (1.1) has.

The largest topological entropy of  $(\Sigma_A, \sigma_A)$  is attained when  $\Sigma_A = \Sigma_{\kappa}$  and  $\sigma_A = \sigma$ . We thus consider the switched system (1.1) with the switching paths allowed to be the whole symbol space  $\Sigma_{\kappa}$ . The corresponding transition matrix  $A = [a_{ij}]$  satisfies  $a_{ij} = 1$  for all  $1 \leq i, j \leq \kappa$ . In this case, the unique maximal entropy measure  $\mu_A$  is the  $(\frac{1}{\kappa}, \dots, \frac{1}{\kappa})$ -product measure  $\mu_{\kappa}$  with  $h_{top}(\sigma) = \ln \kappa$  (Theorem 8.9 in [36]). The following corollary characterizes the  $\mu_{\kappa}$ -almost sure stability of (1.1) without constraints.

**COROLLARY 1.** *Let us consider the switched linear system (1.1) with the switching sequences belonging to the whole symbol space  $\Sigma_{\kappa}$ . Then we have the following.*

- (1) *The system is  $\mu_{\kappa}$ -almost stable if and only if there is at least one  $\hat{t} \in \mathbb{N}$  such that*

$$\prod_{(i_0 \dots i_{\hat{t}-1}) \in \mathcal{A}^{\hat{t}}} \lambda_{i_0 i_1 \dots i_{\hat{t}-1}} < 1,$$

where  $\lambda_{i_0 i_1 \dots i_{\hat{t}-1}}$  is defined by (3.4).

- (2) *Let  $\Sigma_{stab}(S; \Sigma_{\kappa}) = \{\omega \in \Sigma_{\kappa} \mid x_{\ell+1} = S(\sigma^{\ell} \omega) x_{\ell} \text{ is exponentially stable}\}$ . Then we have*

$$HD_{\rho}(\Sigma_{stab}(S; \Sigma_{\kappa})) = HD_{\rho}(\Sigma_{\kappa}) = \frac{\ln \kappa}{\ln \rho},$$

provided that the system (1.1) is  $\mu_{\kappa}$ -almost stable.

*Proof.* The proof is straightforward based on Theorems 1 and 2. Note that (3.5) is equivalent to

$$\left[ \prod_{(i_0 \dots i_{\hat{t}-1}) \in \mathcal{A}^{\hat{t}}} \lambda_{i_0 i_1 \dots i_{\hat{t}-1}} \right]^{(\frac{1}{\kappa})^{\hat{t}}} < 1,$$

which is equivalent to

$$\prod_{(i_0 \cdots i_{t-1}) \in \mathcal{A}^t} \lambda_{i_0 i_1 \cdots i_{t-1}} < 1. \quad \square$$

**COROLLARY 2.** *Suppose that  $\|H_i\| \leq 1$  for  $i = 1, 2, \dots, \kappa$  and that there is at least  $j$  with  $1 \leq j \leq \kappa$  such that  $\|H_j\| < 1$ . Then the system (1.1) is  $\mu_\kappa$ -almost surely stable.*

The next corollary describes the “strongest”  $\mu$ -almost surely stable case.

**COROLLARY 3.** *Suppose that  $\|H_i\| < 1$  for  $i = 1, 2, \dots, \kappa$ . Then the system (1.1) is  $\mu$ -almost stable for any  $\sigma$ -invariant probability measure  $\mu$ , and, consequently, (1.1) is absolutely stable.*

**4. Stochastic Markov jump linear systems.** In this section, we consider the discrete-time system (1.1), where  $\omega_k$  is a discrete-time Markovian stochastic process taking value in  $\mathcal{A} = \{1, \dots, \kappa\}$ , with transition probabilities  $p_{ij} = \Pr\{\omega_{k+1} = j | \omega_k = i\}$ . The matrix  $P = [p_{ij}]$  is called the transition probability distribution. Now the switched system given by

$$(4.1) \quad x_{\ell+1} = H_{\omega_\ell} x_\ell, \quad \ell \geq 0,$$

is a standard stochastic Markov jump linear system where  $\omega_\ell$  is a random variable. The almost sure stability of this type of system has been discussed by many authors. Recently a new necessary and sufficient condition for almost sure stability by the approach of using a so-called lifted version of the system was proposed [5, Proposition 3.4]. We shall show that our results obtained in the previous section can cover such a case.

Assume that the transition probability matrix  $P$  is irreducible. According to the Perron–Frobenius theorem [7], there exists a unique (invariant) distribution  $p = (p_1, \dots, p_\kappa)$  such that

- (a)  $0 < p_i < 1$ ,  $\sum_{i=1}^\kappa p_i = 1$ ;
- (b) (invariance)  $p \cdot P = p$ .

Now a natural Markov measure  $\mu_{p,P}$  on  $\Sigma_\kappa$  defined by  $(p, P)$  can be obtained as follows:

$$(4.2) \quad \mu_{p,P}([i_0 \cdots i_\ell]) = p_{i_0} p_{i_0 i_1} \cdots p_{i_{\ell-1} i_\ell}.$$

Next, we define an irreducible  $(0, 1)$ -matrix  $A_P = [a_{ij}]_{\kappa \times \kappa}$  associated with  $P$  as follows:

$$(4.3) \quad a_{ij} = \begin{cases} 1 & \text{if } p_{ij} > 0, \\ 0 & \text{if } p_{ij} = 0. \end{cases}$$

Let  $(\Sigma_{A_P}, \sigma_{A_P})$  be the one-sided Markov shift defined by the transition matrix  $A_P$ .

One can verify directly that

- $\mu_{p,P}$  is supported on  $\Sigma_{A_P}$  with  $\text{supp}(\mu_{p,P}) = \Sigma_{A_P}$ ;
- $\mu_{p,P}$  is an ergodic  $\sigma_{A_P}$ -invariant Borel probability measure on  $\Sigma_{A_P}$  [36, Theorem 1.13].

Thus, according to the proofs of Theorems 1 and 2, we immediately have the following corollary.

**COROLLARY 4.** *Consider the stochastic Markov jump linear system (4.1) with the probability transition matrix  $P$ . Let the corresponding ergodic  $\sigma_{A_P}$ -invariant Borel*

probability measure  $\mu_{p,P}$  be defined as (4.2). Then the system is  $\mu_{p,P}$ -almost surely stable if and only if there is at least one  $\hat{t} \in \mathbb{N}$  such that

$$(4.4) \quad \prod_{(i_0 \dots i_{\hat{t}-1}) \in \mathcal{A}^{\hat{t}}} \lambda_{i_0 i_1 \dots i_{\hat{t}-1}}^{p_{i_0} p_{i_0 i_1} \dots p_{i_{\hat{t}-2} i_{\hat{t}-1}}} < 1,$$

where  $\lambda_{i_0 i_1 \dots i_{\hat{t}-1}}$  is given by (3.4). Moreover, if (4.1) is  $\mu_{p,P}$ -almost surely stable, then there exists some  $\varepsilon > 0$  such that every switched linear system

$$(1.1)' \quad x_{\ell+1} = H'_{\omega_\ell} x_\ell, \quad \ell \geq 0,$$

is also  $\mu_{p,P}$ -almost surely stable based on  $(\Sigma_{A_P}, \sigma_{A_P})$  whenever

$$\|H_i - H'_i\| \leq \varepsilon, \quad 1 \leq i \leq \kappa.$$

COROLLARY 5. Over a  $(p, P)$ -Markov shift  $(\Sigma_{A_P}, \sigma_{A_P})$ , write

$$\Sigma_{stab}(p, P) = \left\{ \omega \in \Sigma_{A_P} \mid x_{\ell+1} = S(\sigma_{A_P}^\ell \omega) x_\ell \text{ with } \lim_{\ell \rightarrow \infty} \frac{1}{\ell} \ln \|x_\ell\| < 0 \ \forall x_0 \in \mathbb{R}^n \right\}.$$

If (4.4) holds for some  $\hat{t} \in \mathbb{N}$ , then

$$(4.5) \quad HD_\rho(\Sigma_{A_P}) \geq HD_\rho(\Sigma_{stab}(p, P)) \geq HD_\rho(\mu_{p,P}).$$

Remark 6. For the natural Markov measure defined above, in general, we have only

$$HD_\rho(\Sigma_{A_P}) \geq HD_\rho(\Sigma_{stab}(\mu_{p,P})),$$

since  $\mu_{p,P}$  usually is not the Parry measure, and

$$HD_\rho(\mu_{p,P}) = \frac{h_{\mu_{p,P}}(\sigma_{A_P})}{\ln \rho} \neq HD_\rho(\Sigma_{A_P}).$$

This is mainly due to  $\mu_{p,P}$  not needing to be a maximal entropy. Hence the “size” of the set of all stable paths is possibly smaller than that of the admissible set  $\Sigma_{A_P}$ .

We close this section with two more remarks.

Remark 7. Criterion (4.4) for almost sure stability is an extension of the Fang–Loparo–Feng criterion [15]. In fact, if  $\hat{t} = 1$ , then we have

$$\prod_{i \in \mathcal{A}} \|S_i\|^{p_i} < 1,$$

which coincides with the Fang–Loparo–Feng sufficient criterion.

Remark 8. Compared with the Bolzern–Colaneri–Nicolao criterion [5], our approach in getting (4.4) does not involve any lifting of the system (4.1). For an  $m$ -lifting of the system, one easily obtains from a corresponding  $\kappa^m$ -dimensional  $(\tilde{p}, \tilde{P})$ -Markov shift  $(\Sigma_{\tilde{A}_{\tilde{P}}}, \sigma_{\tilde{A}_{\tilde{P}}})$  [5, Propositions 3.1 and 3.2], and thus it can be included in our framework.

**5. Illustrative examples.** In this section, we give two examples (with  $\kappa = 2$ ) to show how to apply the criteria obtained in this paper. In what follows, we denote by  $\mu_2$  the Parry measure of the full shift system  $(\Sigma_2, \sigma)$ , and  $\|\cdot\|$  stands for  $\|\cdot\|_F$ , which is defined in section 1.3.

*Example 1.* Consider the switched system (1.1) with  $\mathcal{A} = \{1, 2\}$ , and

$$H_1 = \begin{bmatrix} 0.2 & 1 \\ 0 & 0.2 \end{bmatrix}, \quad H_2 = \begin{bmatrix} 0.9 & 0.4 \\ 0.5 & 0.2 \end{bmatrix}.$$

It is obvious that this system is not stable for all switching sequences  $\omega \in \Sigma_2$  since the spectral radius of  $H_2$  is greater than 1. However, it is  $\mu_2$ -almost surely stable, where  $\mu_2$  is the maximal entropy measure of  $(\Sigma_2, \sigma)$ , since

$$\begin{aligned} \lambda_{11} &= \|H_1 H_1\| = 0.4040, \quad \lambda_{12} = \|H_1 H_2\| = 0.7432, \\ \lambda_{21} &= \|H_2 H_1\| = 1.1377, \quad \lambda_{22} = \|H_2 H_2\| = 1.2545, \\ \prod_{(i_0 i_1) \in \{1,2\} \times \{1,2\}} \lambda_{i_0 i_1} &= 0.4285 < 1. \end{aligned}$$

Hence, this demonstrates that the system is  $\mu_2$ -almost surely stable and the Hausdorff dimension of the set of all stable sequences  $\omega \in \Sigma_2$  equals 1 (under the metric constant  $\rho = 2$ ) by Corollary 1.

In [5], the system with stochastic Markov chains was considered, where the transition probability distribution was

$$P = \begin{bmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{bmatrix}.$$

The unique invariant distribution in this case is  $p = (0.2, 0.8)$ . It is not difficult to see that condition (4.4) holds true for  $\hat{t} = 3$ . Hence, the system is  $\mu_{p,P}$ -almost surely stable by Corollary 4. However, one can directly verify that the Hausdorff dimension of the set of all stable sequences in the sense of  $\mu_{p,P}$ -almost sure stability is strictly less than 1.

We have known that the system is both  $\mu_2$ -almost surely stable and  $\mu_{p,P}$ -almost surely stable. Nevertheless, to check  $\mu_{p,P}$ -almost surely stable, one needs to use Corollary 4 since it requires the information of transition probability distribution. It is possible for a stochastic Markov jump system to be  $\mu_{p,P}$ -almost surely stable but not to be  $\mu_2$ -almost surely stable since the maximal entropy measure  $\mu_2$  and the ergodic invariant measure  $\mu_{p,P}$  are mutually singular.

According to Theorem 1, we know that the almost sure stability of (1.1) is robust; that is, there exists  $\varepsilon > 0$  such that whenever  $\|H'_i - H_i\| \leq \varepsilon$ ,  $i = 1, 2$ , the switched system

$$(5.1) \quad x_{\ell+1} = H'_{\omega_\ell} x_\ell, \quad \omega_\ell \in \{1, 2\},$$

is also  $\mu_2$ -almost surely stable and the Hausdorff dimension of the set of all stable sequences  $\omega \in \Sigma_2$  equals 1, too. Now as an example we estimate the upper bound of admissible perturbation constant  $\varepsilon$ . Let

$$\lambda_{ij} = \|H_i H_j\| \quad \forall i, j \in \{1, 2\}.$$

We first solve the inequality

$$\begin{aligned}
 & (\lambda_{11} + \delta)(\lambda_{12} + \delta)(\lambda_{21} + \delta)(\lambda_{22} + \delta) \\
 & \leq \lambda_{11}\lambda_{12}\lambda_{21}\lambda_{22} + |\delta|(\lambda_{11}\lambda_{12}\lambda_{21} + \lambda_{11}\lambda_{12}\lambda_{22} + \lambda_{11}\lambda_{21}\lambda_{22} + \lambda_{12}\lambda_{21}\lambda_{22}) \\
 & \quad + \delta^2(\lambda_{11}\lambda_{12} + \lambda_{11}\lambda_{21} + \lambda_{11}\lambda_{22} + \lambda_{12}\lambda_{21} + \lambda_{12}\lambda_{22} + \lambda_{21}\lambda_{22}) \\
 & \quad + |\delta|^3(\lambda_{11} + \lambda_{12} + \lambda_{21} + \lambda_{22}) + \delta^4 \\
 & < 1.
 \end{aligned}$$

Substituting the values of  $\lambda_{ij}$  into the above inequality, one can get approximately

$$|\delta| < 0.1542.$$

Next, we denote  $G_i = H'_i - H_i$ ,  $i = 1, 2$ . A sufficient condition for the  $\mu_2$ -almost sure stability of the perturbation system (5.1) is

$$\begin{aligned}
 & \prod_{(i_0 i_1) \in \{1,2\} \times \{1,2\}} \|(H_{i_0} + G_{i_0})(H_{i_1} + G_{i_1})\| \\
 & \leq [\|H_1 H_1\| + 2\varepsilon\|H_1\| + \varepsilon^2] [\|H_1 H_2\| + \varepsilon(\|H_1\| + \|H_2\|) + \varepsilon^2] \\
 & \quad [\|H_2 H_1\| + \varepsilon(\|H_1\| + \|H_2\|) + \varepsilon^2] [\|H_2 H_2\| + 2\varepsilon\|H_2\| + \varepsilon^2] \\
 & < 1,
 \end{aligned}$$

where

$$\varepsilon = \max\{\|G_1\|, \|G_2\|\}.$$

Since

$$\|H_1\| = 1.0392, \quad \|H_2\| = 1.1225,$$

it follows that a sufficient condition for the inequality

$$\prod_{(i_0 i_1) \in \{1,2\}^2} \|(H_{i_0} + G_{i_0})(H_{i_1} + G_{i_1})\| \leq \prod_{(i_0 i_1) \in \{1,2\}^2} (\lambda_{i_0 i_1} + 2.245\varepsilon + \varepsilon^2) < 1$$

to hold is

$$2.245\varepsilon + \varepsilon^2 = \delta < 0.1542,$$

which yields

$$\varepsilon < 0.0667.$$

*Example 2.* Let us consider the system given in Example 1 with the switching sequences  $\omega$  belonging to the topological Markov chain  $(\Sigma_A, \sigma_A)$  with the topological transition matrix  $A$  as

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

This means that during the switching process the subsystem  $H_2$  cannot be allowed to follow itself.

Direct computations show that the spectral radius of  $A$  is given by

$$\rho_A = \frac{1 + \sqrt{5}}{2},$$

and the Perron vectors are

$$v^T = u = \left( \frac{1 + \sqrt{5}}{\sqrt{10 + 2\sqrt{5}}}, \sqrt{\frac{2}{5 + \sqrt{5}}} \right).$$

Thus the Parry distribution is

$$p_A = (0.7236, 0.2764),$$

and the transition probability matrix is given by

$$P_A = \begin{bmatrix} \frac{2}{1+\sqrt{5}} & \frac{2}{3+\sqrt{5}} \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0.6180 & 0.3820 \\ 1 & 0 \end{bmatrix}.$$

So we have

$$\prod_{(i_0 i_1) \in \{1,2\}^2} \lambda_{i_0 i_1}^{p_{i_0 p_{i_0 i_1}}} = 0.6366 < 1.$$

This verifies that condition (3.5) holds for  $\hat{t} = 2$ . The system thus is  $\mu_A$ -almost surely stable, where  $\mu_A$  is the Parry measure of  $(\Sigma_A, \sigma_A)$ , and

$$HD_{\boldsymbol{\rho}}(\Sigma_{stab}(S; A)) = HD_{\boldsymbol{\rho}}(\Sigma_A) = \frac{\ln(1 + \sqrt{5})}{\ln 2} - 1 > 0,$$

where  $\boldsymbol{\rho} = 2$  (see (2.1a)).

**6. Concluding remarks.** By viewing switching sequences as the elements in symbolic topology space, we have established a necessary and sufficient condition for almost sure stability of discrete-time switched linear systems by using the multiplicative ergodic theorem. Among all ergodic probability measures, Parry measure has been shown to be able to capture the maximal set of stable processes for linear switched systems in the sense of Hausdorff dimension. The  $\mu_A$ -almost sure stability is unchanged under small linear perturbations of the system. Furthermore, a connection between the switched system (1.1) and its corresponding symbolic dynamical system  $(\Sigma_A, \sigma_A)$  is identified; that is, the more Li–Yorke chaotic  $(\Sigma_A, \sigma_A)$  behaves, the larger the set of  $\mu_A$ -almost surely stable paths (1.1) has. Some recent results for the stochastic Markov jump linear systems can be adopted in our framework. Future research will be concentrated on the continuous-time case as well as on nonlinear switched systems.

**Acknowledgment.** The authors would like to thank the anonymous reviewers for their suggestions which have been very helpful in the improvement of the paper.

REFERENCES

[1] R. L. ADLER, A. G. KONHEIM, AND M. H. MCANDREW, *Topological entropy*, Trans. Amer. Math. Soc., 114 (1965), pp. 309–319.  
[2] V. AFRAIMOVICH AND S.-B. HSU, *Lectures on Chaotic Dynamical Systems*, AMS/IP Stud. Adv. Math. 28, AMS, Providence, RI; International Press, Somerville, MA, 2003.

- [3] M. BERGER AND Y. WANG, *Bounded semigroups of matrices*, Linear Algebra Appl., 166 (1992), pp. 21–27.
- [4] F. BLANCHARD, E. GLASNER, S. KOLYADA, AND A. MAASS, *On Li-Yorke pairs*, J. Reine Angew. Math., 547 (2002), pp. 51–68.
- [5] P. BOLZERN, P. COLANERI, AND G. DE NICOLAO, *On almost sure stability of discrete-time Markov jump linear systems*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Atlantis, Paradise Island, Bahamas, 2004, pp. 3204–3208.
- [6] P. BOLZERN, P. COLANERI, AND G. DE NICOLAO, *On almost sure stability of continuous-time Markov jump linear systems*, Automatica J. IFAC, 42 (2006), pp. 983–988.
- [7] P. BREMAUD, *Markov Chains. Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer-Verlag, New York, 1999.
- [8] M. BRIN AND A. KATOK, *On local entropy*, in Geometric Dynamics, Lecture Notes in Math. 1007, Springer-Verlag, Berlin, 1983, pp. 30–38.
- [9] O. L. V. COSTA, M. D. FRAGOSO, AND R. P. MARQUES, *Discrete-Time Markov Jump Linear Systems*, Probab. Appl. (N.Y.), Springer-Verlag, London, 2005.
- [10] X. DAI, Z.-L. ZHOU, AND X.-Y. GENG, *Some relations between Hausdorff-dimensions and entropies*, Sci. China Ser. A, 41 (1998), pp. 1068–1075.
- [11] X. DAI, *Existence of full-Hausdorff-dimension invariant measures of dynamical systems with dimension metrics*, Arch. Math. (Basel), 85 (2005), pp. 470–480.
- [12] X. DAI AND Y.-P. JIANG, *Distance entropy of dynamical systems on noncompact-phase spaces*, Discrete Contin. Dyn. Syst., 20 (2008), pp. 313–333.
- [13] W. P. DAYAWANSA AND C. F. MARTIN, *A converse Lyapunov theorem for a class of dynamical systems which undergo switching*, IEEE Trans. Automat. Control, 44 (1999), pp. 751–760.
- [14] K. FALCONER, *Fractal Geometry: Mathematical Foundations and Applications*, John Wiley & Sons, New York, 1990.
- [15] Y. FANG, K. A. LOPARO, AND X. FENG, *Almost sure and  $\delta$ -moment stability of jump linear systems*, Internat. J. Control, 59 (1994), pp. 1281–1307.
- [16] Y. FANG, *A new general sufficient condition for almost sure stability of jump linear systems*, IEEE Trans. Automat. Control, 42 (1997), pp. 378–382.
- [17] X. FENG, K. A. LOPARO, Y. JI, AND H. J. CHIZECK, *Stochastic stability properties of jump linear systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 38–53.
- [18] H. FURSTENBERG AND H. KESTEN, *Products of random matrices*, Ann. Math. Statist., 31 (1960), pp. 457–469.
- [19] J. P. HESPAHNA, *Uniform stability of switched linear systems: Extensions of LaSalle’s invariance principle*, IEEE Trans. Automat. Control, 49 (2004), pp. 470–462.
- [20] Y. JI AND H. J. CHIZECK, *Controllability, stabilizability, and continuous-time Markovian jump linear quadratic control*, IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.
- [21] Y. JI, H. J. CHIZECK, X. FENG, AND K. A. LOPARO, *Stability and control of jump linear systems*, Control Theory Adv. Tech., 7 (1991), pp. 247–270.
- [22] I. A. KAC AND N. N. KRASOVSKII, *On the stability of systems with random parameters*, J. Appl. Math. Mech., 24 (1960), pp. 1225–1246.
- [23] R. Z. KHASMINSKII, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1980.
- [24] R. Z. KHASMINSKII, C. ZHU, AND G. YIN, *Stability of regime-switching diffusions*, Stochastic Process. Appl., 117 (2007), pp. 1037–1051.
- [25] A. N. KOLMOGOROV, *A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces*, Dokl. Akad. Nauk. SSSR, 119 (1958), pp. 861–864.
- [26] H. J. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.
- [27] Z. G. LI, Y. C. SOH, AND C. Y. WEN, *Sufficient conditions for almost sure stability of jump linear systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1325–1329.
- [28] T.-Y. LI AND J. A. YORKE, *Period three implies chaos*, Amer. Math. Monthly, 82 (1975), pp. 985–992.
- [29] D. LIBERZON, *Switching in Systems and Control*, Birkhäuser Boston, Boston, 2003.
- [30] D. LIBERZON, J. P. HESPAHNA, AND A. S. MORSE, *Stability of switched systems: A Lie-algebraic condition*, Systems Control Lett., 37 (1999), pp. 117–122.
- [31] X. MAO, *Stability of stochastic differential equations with Markovian switching*, Stochastic Process. Appl., 79 (1999), pp. 45–67.
- [32] P. MASON, U. BOSCAIN, AND Y. CHITOUR, *Common polynomial Lyapunov functions for linear switched systems*, SIAM J. Control Optim., 45 (2006), pp. 226–245.
- [33] V. I. OSELEDEC, *A multiplicative ergodic theorem, Lyapunov characteristic numbers for dynamical systems*, Trudy Mosk. Mat. Obs., 19 (1968), pp. 119–210.
- [34] Z. SUN AND S. S. GE, *Analysis and synthesis of switched linear control systems*, Automatica J. IFAC, 41 (2005), pp. 181–195.



- [35] Z. SUN AND S. S. GE, *Switched Linear Systems, Control and Design*, Springer-Verlag, London, 2005.
- [36] P. WALTERS, *An Introduction to Ergodic Theory*, Grad. Texts in Math. 79, Springer-Verlag, New York, 1982.
- [37] M. XIAO, *Optimal control of nonlinear systems with controlled transitions*, Nonlinear Dyn. Syst. Theory, 5 (2005), pp. 177–188.
- [38] L. S. YOUNG, *Dimension, entropy and Lyapunov exponents*, Ergodic Theory Dynam. Systems, 2 (1982), pp. 109–124.
- [39] Z.-L. ZHOU, *The topological Markov chain*, Acta Math. Sinica (N.S.), 4 (1988), pp. 330–337.

## HAMILTON–JACOBI EQUATIONS ARISING FROM BOUNDARY CONTROL PROBLEMS WITH STATE CONSTRAINTS\*

SILVIA FAGGIAN†

**Abstract.** The analysis of a class of infinite-dimensional Hamilton–Jacobi–Bellman (HJB) equations is undertaken related to linear convex boundary control problems for PDEs with constraints on the state. A definition of generalized solution (namely, weak solution) of the HJB equation is provided (weak is the limit of strong, while strong is the limit of classical, as defined in [S. Faggian, *Appl. Math. Optim.*, 51 (2005), pp. 123–162]). Consequently an existence and uniqueness result is provided for weak solutions of the HJB equation, as well as existence of an optimal control, being the limit of optimal controls of approximating problems. The study then describes an economic application to optimal investment with vintage capital, having positivity constraints on the capital.

**Key words.** linear convex control, boundary control, state constraints, age-structured systems, vintage models

**AMS subject classifications.** 49J20, 49J27, 35B37

**DOI.** 10.1137/070683738

**1. Introduction.** The present paper is devoted to the analysis of the Hamilton–Jacobi–Bellman (HJB) equation as related to linear convex boundary control in Hilbert spaces with constraints on the state. More precisely, we let  $H$  and  $U$  be separable real Hilbert spaces with scalar products  $(\cdot|\cdot)_H$  and  $(\cdot|\cdot)_U$ , respectively, and consider a dynamical system of the type

$$(1.1) \quad \begin{cases} y'(\tau) = A_0 y(\tau) + B u(\tau), & \tau \in [t, T], \\ y(t) = x \in H, \end{cases}$$

where  $H$  is the state space,  $y : [t, T] \rightarrow H$  is the trajectory,  $U$  is the control space and  $u : [t, T] \rightarrow U$  is the control,  $A_0 : D(A_0) \subset H \rightarrow H$  is the infinitesimal generator of a strongly continuous semigroup of linear operators  $\{e^{\tau A_0}\}_{\tau \geq 0}$  on  $H$ , and the control operator  $B$  is linear and *unbounded*, say  $B : U \rightarrow [D(A_0^*)]'$ . In addition, we consider a cost functional given by

$$(1.2) \quad J(t, x, u) = \int_t^T [g_0(\tau, y(\tau)) + h(\tau, u(\tau))] d\tau + \varphi_0(y(T)),$$

where the functions  $\varphi_0$  and, for all fixed  $\tau$  in  $[0, T]$ ,  $g_0(\tau, \cdot)$  and  $h(\tau, \cdot)$  are lower semi-continuous (l.s.c.) and convex, and possibly infinite valued, as explained later. Our problem is that of minimizing  $J(t, x, u)$  with respect to  $u$  over a suitable subset of  $L^p(t, T; U)$ ,  $p > 1$ , of admissible controls.

More precisely we are interested in deriving an existence and uniqueness result for the HJB equation associated to the problem, that is,

$$(1.3) \quad \begin{cases} \phi_t(t, x) + \mathcal{H}(T - t, -B^* \phi_x(t, x)) - (A_0 x | \phi_x(t, x))_H = g_0(T - t, x), \\ \phi(0, x) = \varphi_0(x) \end{cases}$$

\*Received by the editors February 26, 2007; accepted for publication (in revised form) March 8, 2008; published electronically July 23, 2008.

<http://www.siam.org/journals/sicon/47-4/68373.html>

†LUM “Jean Monnet,” Casamassima, Bari, I-70010, Italy (faggian@lum.it).

for all  $x \in H$  and  $t \in [0, T]$ , where

$$\mathcal{H}(t, u) = \sup_{v \in U} \{ (u|v)_U - h(t, v) \}.$$

The treatment of such a problem with not necessarily continuous  $g_0$  and  $\varphi_0$  costs allows us to study a rather general class of problems with state constraints. Indeed one may show that if the set of admissible controls is of the type

$$\mathcal{U}(t, x) := \{ u \in L^p(t, T; U) : y(\tau; t, x, u) \in \mathcal{C} \ \forall \tau \in [t, T] \},$$

where we denoted by  $y(\tau; t, x, u)$  the trajectory of the system at time  $\tau$  which started in  $x$  at time  $t$ , and is driven by the control  $u$ , and where

$$(1.4) \quad \mathcal{C} \subset H \text{ is a closed convex set}$$

possibly having an empty interior, then the problem is equivalent to a problem with modified costs  $g$  and  $\varphi$  (set equal to  $+\infty$  outside  $\mathcal{C}$ ), and with no constraints on the state variable.

After recalling the results for the unconstrained problem contained in [33] and [34], we discuss in sections 4 and 5 existence and uniqueness of a *weak* solution of the HJB equation, that is, the pointwise limit of regular *strong* (as defined in [33]) solutions of approximating equations. Moreover we show that the value function  $W$  defined by

$$W(t, x) := \inf_{u \in \mathcal{U}(t, x)} J(t, x, u)$$

is the unique weak solution of the backward HJB equation, and it can be used in some applied cases to provide a feedback formula for optimal controls.

It is well known that control problems with unbounded control operator  $B$  arise when we rephrase into abstract terms some boundary control problem for PDEs (or, more generally, problems with control on a subdomain). Indeed, our framework with the application to the economic problem of optimal investment with vintage capital is motivated by that of Barucci and Gozzi [13, 12], which we describe in detail in section 6. Similar problems with an unbounded control operator have been discussed in a series of papers by this author and others. The unconstrained case has been studied for both finite and infinite horizons [33, 34, 36], while [35] contains the finite horizon case with constrained controls. Then the important novelty of this paper is that the (finite horizon) case with both *boundary control* and *state constraints* are undertaken. In particular the paper extends to the case of boundary control the results of existence and uniqueness for HJB equation contained in the paper by Cannarsa and Di Blasio [17] for distributed control and with different assumptions. Moreover, our model also takes into account the case of unconstrained problems as those in [33, 34], but with merely l.s.c. data  $\varphi_0$  and  $g_0$  rather than differentiable ones, as in [33, 34].

In our opinion this paper contributes to the study of a subject, that of optimal control problems in infinite dimensions with boundary control and/or state constraints, which is difficult to treat and whose literature is rather poor. Indeed, for the linear convex problem with *state constraints* but bounded  $B$ , we mention again the work by Cannarsa and Di Blasio [17]; for the case of viscosity solutions (with both bounded  $B$  and  $A$ ) we mention the papers by Gozzi, Cannarsa, and Soner [18] and Kocan and Soravia [47]. Finally we mention the book by Fattorini [37] on maximum principle.

Some further references on *boundary control* in infinite dimensions follow. We recall that such problems have been studied in the framework of classical/strong solutions and in that of viscosity solutions. Regarding dynamic programming in the classical/strong framework, the available results mainly regard the case of linear systems and quadratic costs (where the HJB equation reduces to the operator Riccati equation). The reader is then referred, e.g., to the books by Lasiecka and Triggiani [48, 49], to the book by Bensoussan et al. [14], and, for the case of nonautonomous systems, to the papers by Acquistapace, Flandoli, and Terreni [1] and Acquistapace and Terreni [2, 3, 4]. For the case of a linear system and a general convex cost, we mention the papers by this author [31, 32, 33, 34]. On the Pontryagin maximum principle for boundary control problems, see the book by Barbu and Precupanu (Chapter 4 in [11]).

For viscosity solutions and HJB equations in infinite dimensions we mention the series of papers by Crandall and Lions [21, 22, 23, 24, 25, 26, 27], where also some boundary control problem arises. Moreover, for boundary control we mention Gozzi, Cannarsa, and Soner [19] and the paper by Cannarsa, and Tessitore [20] on existence and uniqueness of viscosity solutions of the HJB equation. We note also that a verification theorem in the case of viscosity solutions has been proved in some finite-dimensional case in the book by Yong and Zhou [55]. We finally mention the paper by Fabbri [30], where the author derives an existence and uniqueness result for the viscosity solution of the HJB equation associated to optimal investment with vintage capital (with infinite horizon and without constraints), which is the application of section 6 of the present paper, and some continuity property of the value function,<sup>1</sup> obtaining the results by making use of the specific properties of the state equation, while no result is provided there for the general problem.<sup>2</sup>

We mention also some fundamental papers and books on the case of *distributed control* in the classical/strong framework, such as the works by Barbu and Da Prato [7, 8, 9] for some linear convex problems, Di Blasio [28, 29] for the case of constrained control, Cannarsa and Di Blasio [17] for the case of state constraints, and Barbu, Da Prato, and Popa [10] and Gozzi [42, 44, 43] for semilinear systems.

Regarding applications, on control on a subdomain (boundary or point control) we refer the reader to the many examples contained in the books by Lasiecka and Triggiani [48, 49] and by Bensoussan et al. [14, 15]. Moreover, for economic models with vintage capital the reader may see the papers by Barucci and Gozzi [13, 12] Feichtinger et al. [41, 38, 39, 40], and for population dynamic the book by Iannelli [46], the paper by Anița et al. [6], and the papers by Almeder et al. [5] and the references therein.

The paper is organized as follows. In section 2 we recall the definition of strong solutions and the results on existence and uniqueness strong solutions with unconstrained state. In section 3 we formulate the problem with state constraints, and in sections 4 and 5 we discuss existence and uniqueness for weak solutions of the HJB equation associated to the optimal control problem; finally in section 6 we apply the theory to optimal investment with vintage capital and positivity constraints.

**2. Preliminaries: The unconstrained case.** We recall here all the relevant results in the *unconstrained* case that are needed in what follows. According to the notation in [33], if  $X$  and  $Y$  are Banach spaces, we denote by  $|\cdot|_X$  the norm on  $X$ ,

<sup>1</sup>More precisely, the author proves Lipschitz  $B$ -continuity, where  $B = (A^* - \lambda I)^{-1}(A - \lambda I)^{-1}$ .

<sup>2</sup>No comparison between strong and viscosity solutions is yet available for the unconstrained problem, even for optimal investment with vintage capital.

by  $|\cdot|$  the euclidean norm in  $\mathbb{R}$ , and we set

$$\begin{aligned} Lip(X; Y) &= \left\{ f : X \rightarrow Y : [f]_L := \sup_{x, y \in X, x \neq y} \frac{|f(x) - f(y)|_Y}{|x - y|_X} < +\infty \right\}, \\ C_{Lip}^1(X) &:= \{f \in C^1(X) : [f']_L < +\infty\}, \\ \mathcal{B}_r(X, Y) &:= \left\{ f : X \rightarrow Y : |f|_{\mathcal{B}_r} := \sup_{x \in X} \frac{|f(x)|_Y}{1 + |x|_X^r} < +\infty \right\}, \quad \mathcal{B}_r(X) := \mathcal{B}_r(X, \mathbb{R}). \end{aligned}$$

Moreover we set

$$\Sigma_0(X) := \{w \in \mathcal{B}_2(X) : w \text{ is convex, } w \in C_{Lip}^1(X)\}$$

and, for  $T > 0$ ,

$$\begin{aligned} \mathcal{Y}([0, T] \times X) &= \{w : [0, T] \times X \rightarrow \mathbb{R} : w \in C([0, T], \mathcal{B}_2(X)), \\ &\quad w(t, \cdot) \in \Sigma_0(X) \ \forall t \in [0, T], \ w_x \in C([0, T], \mathcal{B}_1(X, X'))\}. \end{aligned}$$

All the spatial derivatives above have to be intended as Fréchet differentials.

Then we consider two Hilbert spaces  $V, V'$ , being dual spaces, which we do not identify for reasons stated in Remark 2.2, and we denote the duality pairing by  $\langle \cdot, \cdot \rangle$ . We set  $V'$  as the state space of the problem and denote by  $U$  the control space, with  $U$  being another Hilbert space.

Given an initial time  $t \geq 0$ , an initial state  $x \in V'$ , a finite horizon  $T > t$ , a number  $p \geq 2$ , and a control  $u \in L^p(t, T; U)$ , we consider a state equation of type

$$(2.1) \quad \begin{cases} y'(\tau) = Ay(\tau) + Bu(\tau), & \tau \in ]t, \tau[, \\ y(t) = x \end{cases}$$

and an objective functional of type

$$(2.2) \quad J_T(t, x, u) = \int_t^T [g(\tau, y(\tau)) + h(\tau, u(\tau))] d\tau + \varphi(y(T)).$$

We deal with the problem of minimizing  $J_T(t, x, \cdot)$  over all  $u \in L^p(t, T; U)$ , taking the following set of assumptions on the data.

*Assumptions 2.1.*

1.  $A : D(A) \subset V' \rightarrow V'$  is the infinitesimal generator of a strongly continuous semigroup  $\{e^{\tau A}\}_{\tau \geq 0}$  on  $V'$ ;
2. there exists  $\omega \geq 0$  such that  $|e^{\tau A}x|_{V'} \leq e^{\omega\tau}|x|_{V'}$  for all  $\tau \geq 0$ ;
3.  $g \in \mathcal{Y}([0, T] \times V')$ ,  $t \mapsto [g_x(t, \cdot)]_L \in L^1(0, T)$ ;
4.  $\varphi \in \Sigma_0(V')$ ;
5.  $B \in L(U, V')$ ;
6.  $h(t, \cdot)$  is convex and l.s.c.;  $\partial_u h(t, \cdot)$  is injective for all  $t \in [0, T]$ ;
7. if is set  $\mathcal{H}(t, u) := [h(\tau, \cdot)]^*(u)$ , then we assume  $\mathcal{H} \in \mathcal{Y}([0, T] \times U)$ ,  $\mathcal{H}(t, 0) = 0$ , and  $\sup_{t \in [0, T]} [\mathcal{H}_u(t, \cdot)]_L < +\infty$ .

We will often refer to the trajectory as to the mild solution of (2.1), that is,

$$(2.3) \quad y(\tau) = e^{(\tau-t)A}x + \int_t^\tau e^{(\tau-\sigma)A}Bu(\sigma)d\sigma, \quad \tau \in [t, T].$$

*Remark 2.2.* We do not identify  $V$  and  $V'$ , for in the applications the problem is naturally set in a Hilbert space  $H$ , such that  $V \subset H \equiv H' \subset V'$  (with all bounded inclusions). Indeed, in order to avoid the discontinuities due to the presence of  $B$ , as they appear in (1.1)–(1.2), we work in the extended state space  $V'$  related to  $H$  in the following way:  $V$  is the Hilbert space  $D(A_0^*)$  endowed with the scalar product  $(v|w)_V := (v|w)_H + (A_0^*v|A_0^*w)_H$ , and  $V'$  is the dual space of  $V$  endowed with the operator norm. Then assume that  $B \in L(U, V')$  and extend the semigroup  $\{e^{tA_0}\}_{t \geq 0}$  on  $H$  to a semigroup  $\{e^{tA}\}_{t \geq 0}$  on the space  $V'$ , having infinitesimal generator  $A$ , a proper extension of  $A_0$ . The reader is referred to [34] for a detailed treatment. The coefficient  $\omega$  could be any real number, but is assumed positive in order to avoid double proofs for positive and negative signs.

*Remark 2.3.* Note that  $g$  and  $\varphi$  arising from applications are, as functions of the  $x$  variable, often naturally defined in  $H$ , not on the larger space  $V'$ , and belong to the class  $C^1(H)$ . Then we need to *assume* here that they can be extended to functions in  $C^1(V')$ , which is a nontrivial issue. We refer the reader to section 6 to see how such an extension is obtained in the specific case of the economic example, and to [33] and [34] for a thorough discussion of this issue.

*Remark 2.4.* In Assumption 2.1(7), we assumed  $\mathcal{H}(t, 0) = 0$ . Such an assumption is not restrictive since  $\mathcal{H}(t, 0) = -\inf_{v \in U} h(t, v)$ , and if this value is not 0, we may reduce to this case simply setting  $\bar{g} = g + \inf_{v \in U} h(t, v)$  and  $\bar{h} = h - \inf_{v \in U} h(t, v)$  and treating the problem with  $\bar{g}$  and  $\bar{h}$  in place of  $g$  and  $h$ . Note also that the assumption  $\partial h(t, \cdot)$  injective is intended to yield a good definition for  $\mathcal{H}_u$  as it is, roughly speaking,  $\mathcal{H}_u = (\partial h)^{-1}$ . Note also that once one has the datum  $h$ , its convex conjugate  $\mathcal{H}$  is very often explicitly computed. Then the assumptions on  $\mathcal{H}$  are essentially assumptions on its convex conjugate  $h$ , but more conveniently stated to ensure  $\mathcal{H}$  has the desired properties.

Such optimal control problems can be associated by means of dynamic programming to the HJB equation

$$(2.4) \quad \begin{cases} v_t(t, x) - \mathcal{H}(t, -B^*v_x(t, x)) + \langle Ax|v_x(t, x) \rangle + g(t, x) = 0, & (t, x) \in [0, T] \times V', \\ v(T, x) = \varphi(x), \end{cases}$$

which can be written, by the change of variable  $v(t, x) = \phi(T - t, x)$ , as

$$(2.5) \quad \begin{cases} \phi_t(t, x) + \mathcal{H}(T - t, -B^*\phi_x(t, x)) - \langle Ax, \phi_x(t, x) \rangle = g(T - t, x), & (t, x) \in [0, T] \times V', \\ \phi(0, x) = \varphi(x). \end{cases}$$

Finally, the value function of the problem is defined as

$$(2.6) \quad W_T(t, x) = \inf_{u \in L^p(t, T; U)} J_T(t, x, u).$$

Indeed in [33] Faggian proved existence and uniqueness of *strong* solutions, as defined in Definition 2.5 below, for a class of more general HJB equations, that is,

$$(2.7) \quad \begin{cases} \phi_t(t, x) + F(t, \phi_x(t, x)) - \langle Ax, \phi_x(t, x) \rangle = g(T - t, x), & (t, x) \in [0, T] \times V', \\ \phi(0, x) = \varphi(x), \end{cases}$$

where  $F$  satisfies

$$(2.8) \quad F \in \mathcal{Y}([0, T] \times V), \quad F(t, 0) = 0, \quad \sup_{t \in [0, T]} [F_p(t, \cdot)]_L < +\infty.$$

Note indeed that if we set

$$F(t, p) := \mathcal{H}(t, -B^*p) = \sup_{u \in U} \{|u| - B^*p\}_U - h(t, u),$$

then  $F$  satisfies (2.8) and it is well defined for  $p$  in  $V$ , to which  $\phi_x(t, x)$  belongs.

DEFINITION 2.5. *Let Assumptions 2.1(1–4) and (2.8) be satisfied. We say that  $\phi \in C([0, T], \mathcal{B}_2(V'))$  is a strong solution of (2.7) if there exists a family  $\{\phi^\varepsilon\}_\varepsilon \subset C([0, T], \mathcal{B}_2(V'))$  such that*

- (i)  $\phi^\varepsilon(t, \cdot) \in C_{Lip}^1(V')$  and  $\phi^\varepsilon(t, \cdot)$  is convex for all  $t \in [0, T]$ ;  $\phi^\varepsilon(0, x) = \varphi(x)$  for all  $x \in V'$ ;
- (ii) there exist constants  $\Gamma_1, \Gamma_2 > 0$  such that

$$\sup_{t \in [0, T]} [\phi_x^\varepsilon(t, \cdot)]_L \leq \Gamma_1, \quad \sup_{t \in [0, T]} |\phi_x^\varepsilon(t, 0)|_V \leq \Gamma_2 \quad \forall \varepsilon > 0;$$

- (iii) for all  $x \in D(A)$ ,  $t \mapsto \phi^\varepsilon(t, x)$  is continuously differentiable;
- (iv)  $\phi^\varepsilon \rightarrow \phi$ , as  $\varepsilon \rightarrow 0+$ , in  $C([0, T], \mathcal{B}_2(V'))$ ;
- (v) there exists  $g_\varepsilon \in C([0, T]; \mathcal{B}_2(V'))$  such that, for all  $t \in [0, T]$  and  $x \in D(A)$ ,

$$\phi_t^\varepsilon(t, x) + F(t, \phi_x^\varepsilon(t, x)) - \langle Ax, \phi_x^\varepsilon(t, x) \rangle = g_\varepsilon(T - t, x),$$

with  $g_\varepsilon(t, x) \rightarrow g(t, x)$  pointwise, and  $\int_0^T |g_\varepsilon(s, \cdot) - g(s, \cdot)|_{\mathcal{B}_2} ds \rightarrow 0$ , as  $\varepsilon \rightarrow 0+$ .

The main result contained in [33] is as follows.

THEOREM 2.6. *Let Assumptions 2.1(1–4) and (2.8) be satisfied. There exists a unique strong solution  $\phi$  of (2.7) in the class  $C([0, T], \mathcal{B}_2(V'))$  with the following properties:*

- (i) for all  $x \in D(A)$ ,  $\phi(\cdot, x)$  is Lipschitz continuous;
- (ii)  $\phi \in \mathcal{Y}([0, T] \times V')$ ; moreover the following estimate is satisfied for all  $t \in [0, T]$ :

$$(2.9) \quad [\phi_x(t)]_L \leq e^{2\omega t} [\varphi']_L + \int_0^t e^{2\omega(t-s)} [g_x(T-s, \cdot)]_L ds.$$

Regarding applications to the optimal control problem, in [34] we were able to prove the following theorem.

THEOREM 2.7. *Let Assumptions 2.1(1–7) be satisfied, and let  $\phi$  be the strong solution of (2.5) described in Theorem 2.6. Then*

$$W_T(t, x) = \phi(T - t, x) \quad \forall t \in [0, T], \quad \forall x \in V',$$

that is, the value function  $W_T$  of the optimal control problem is the unique strong solution of the backward HJB equation (2.4).

Finally, we recall the result on optimal control in feedback form.

THEOREM 2.8. *Under Assumptions 2.1, the unique optimal control for problem*

$$\inf \left\{ J(t, x, u) : u \in L^p(t, T; U), \quad y(\tau) := e^{(\tau-t)A}x + \int_t^\tau e^{(\tau-\sigma)A}Bu(\sigma)d\sigma \right\}$$

is given by

$$u^*(\tau) = \mathcal{H}_u(\tau, -B^*W_x(\tau, y^*(\tau))),$$

where  $y^*$  is the unique solution of the closed loop equation

$$y^*(\tau) := e^{(\tau-t)A}x + \int_t^s e^{(\tau-\sigma)A}B\mathcal{H}_u(\sigma, -B^*W_x(\sigma, y^*(\sigma)))d\sigma.$$

**3. Setting of the optimal control problem with state constraints.** Next we want to perform dynamic programming for a problem in  $V'$  that contains, as a subclass, the problem set in  $H$  through (1.1)–(1.3), in the spirit of the previous section.

Then if  $K$  is a convex closed subset of  $V'$ , we consider the general problem

$$\inf\{J(t, x, u) : u \in L^p(t, T; U)\}, \quad p > 1,$$

for any  $(t, x) \in [0, T] \times K$ , where the trajectory is given by (2.3), the cost functional by (2.2), but with no regularity assumptions on  $g$  and  $\varphi$ , which are assumed merely l.s.c. and convex in the  $x$  variable.

More precisely, we make the following assumptions on the data.

*Assumptions 3.1.* We define

$$\Sigma_K \equiv \Sigma_K(V') := \{\phi : V' \rightarrow (-\infty, +\infty] : \phi \text{ is convex and l.s.c., } K \subset D(\phi)\},$$

where  $D(\phi) = \{x \in V' : \phi(x) < +\infty\}$ , and assume the following:

1.  $A : D(A) \subset V' \rightarrow V'$  is the infinitesimal generator of a strongly continuous semigroup  $\{e^{\tau A}\}_{\tau \geq 0}$  on  $V'$ ;
2. there exists  $\omega \geq 0$  such that  $|e^{\tau A}x|_{V'} \leq e^{\omega\tau}|x|_{V'}$  for all  $\tau \geq 0$ ;
3.  $B \in L(U, V')$ ;
4.  $g(t, \cdot) \in \Sigma_K$  for all  $t \in [0, T]$ ;  $g(\cdot, x)$  is l.s.c. and  $L^1(0, T)$  for all  $x \in V'$ ;
5.  $\varphi \in \Sigma_K$ ;
6.  $h(t, \cdot)$  is convex and l.s.c.;  $\partial_u h(t, \cdot)$  is injective for all  $t \in [0, T]$ ; moreover  $h(t, u) \geq a(t)|u|_U^p + b(t)$ , with  $a(t) \geq \alpha_T > 0$ ,  $b \in L^1(0, T; \mathbb{R})$ ,  $p > 1$ ;
7. if it is set  $\mathcal{H}(t, u) := \sup_{v \in U} \{(u|v)_U - h(t, v)\}$ , then we assume  $\mathcal{H} \in \mathcal{Y}([0, T] \times U)$ ,  $\mathcal{H}(t, 0) = 0$ , and  $\sup_{t \in [0, T]} [\mathcal{H}_u(t, \cdot)]_L < +\infty$ .

The functional  $J(t, x, u)$  has to be minimized with respect to  $u$  over the set of admissible controls given by

$$\mathcal{U}(t, x) = \{u \in L^p(t, T; U) : J(t, x, u) < +\infty\}.$$

The value function defined as

$$W(t, x) := \inf_{u \in \mathcal{U}(t, x)} J(t, x, u)$$

is the candidate solution to the HJB equation given by (2.4).

*Remark 3.2.* Again, Assumptions 3.1(6–7) imply that if we set  $F(t, p) = \mathcal{H}(t, -B^*p)$ , then  $F$  satisfies (2.8).

We end this section by explaining to which extent such a framework applies to both the unconstrained problem (i.e., the case  $K = V'$ ) with l.s.c. data and the case with state constraints ( $K$  a proper subset of  $V'$ ). Indeed, not all problems such as (1.1), (1.2), (1.3) are going to fit into this framework.



We need to assume that  $g_0$  and  $\varphi_0$  appearing in (1.2) admit l.s.c. extensions to  $V'$ , as shown more precisely in the following.

*Assumptions 3.3.* Let  $K = cl_{V'}(C)$ , the smallest closed convex set containing  $C$ , with respect to the topology of  $V'$ . We assume there exist l.s.c. functions  $g : [0, T] \times V' \rightarrow \mathbb{R}$  and  $\varphi : V' \rightarrow \mathbb{R}$ , finite on  $[0, T] \times K$ , such that  $g(t, x) = g_0(t, x)$ , and  $\varphi(x) = \varphi_0(x)$  for all  $t$  in  $[0, T]$  and for all  $x$  in  $C$ .

Consequently, we make the following remarks.

*Remark 3.4.* (a) If  $K$  is any closed convex subset of  $V'$  and  $g$  and  $\varphi$  are l.s.c. functions of  $x$  in  $V'$ , then the functions

$$\tilde{g}(t, x) = \begin{cases} g(t, x), & x \in K, \\ +\infty, & x \in V' \setminus K, \end{cases}$$

and

$$\tilde{\varphi}(x) = \begin{cases} \varphi(x), & x \in K, \\ +\infty, & x \in V' \setminus K, \end{cases}$$

are still l.s.c. in the  $x$  variable. Moreover, with such a choice, we have

$$D(\tilde{g}(t, \cdot)) = D(\tilde{\varphi}) = K \quad \forall t.$$

(b) Let us assume we have a dynamical system in the space  $V'$  with trajectory given by (2.3), and a cost functional of type (2.2), where  $g$  and  $\varphi$  are l.s.c. functions, and that we want to minimize such a functional over  $L^p(t, T; U)$  with respect to the control  $u$ , with the constraint

$$y(\tau) \text{ lies in } K, \text{ where } K \text{ is a closed and convex subset of } V',$$

that is, minimize  $J$  over the set of admissible controls

$$(3.1) \quad \mathcal{U}(t, x) := \{u \in L^p(t, T; U) : y(\tau; t, x, u) \in K \quad \forall \tau \in [t, T]\}$$

for any  $(t, x) \in [0, T] \times K$ . Then it is easy to check that such a problem is equivalent to minimizing  $J$  over the whole class  $L^p(t, T; U)$  of controls (no state constraints), with costs  $\tilde{g}$  and  $\tilde{\varphi}$  as defined above, in place of  $g$  and  $\varphi$ .

(c) In the applications, as in the problem set in  $H$ , the constraints on the state are naturally given by means of some convex closed subset  $\mathcal{C}$  of the space  $H$ . Then one may easily show that

$$\mathcal{C} = K \cap H.$$

Then if one knows by dynamic programming that the optimal trajectory  $y^*$  lies in  $K$ , and knows by some other means—as it often happens—that whenever the initial datum  $x$  is in  $H$  also the whole trajectory lies in  $H$ , one may derive that

$$x \in \mathcal{C} \Rightarrow y^*(s) \in \mathcal{C} \quad \forall s \in [t, T],$$

that is, the state constraint is satisfied in the original sense. The reader will find such an application in the last section of the paper.

**4. Weak solutions of the HJB equation.** In order to study the optimal control problem just set and the associated HJB equation (2.4), we need to extend the theory described in the previous section to equations with more general convex coefficients, possibly taking the value  $+\infty$ . We answer this issue by introducing *weak* solutions of the HJB equation in the next section. The definition is given for a wider class of equations that includes the HJB equation of the optimal control problem as a subclass by assuming (2.8), which is implied by Assumptions 3.1.

DEFINITION 4.1. *Let Assumptions 3.1(1–5) and let (2.8) be satisfied. Then  $\phi : [0, T] \times V' \rightarrow (-\infty, +\infty]$  is a weak solution of (2.7) if*

- (i)  $\phi(t, \cdot) \in \Sigma_K$  for all  $t \in [0, T]$ ;
- (ii) *there exist sequences  $\{\varphi_n\}_n \subset \Sigma_0$  and  $\{g_n\} \subset \mathcal{Y}([0, T] \times V')$ , such that*

$$\varphi_n(x) \uparrow \varphi(x), \quad g_n(t, x) \uparrow g(t, x) \quad \forall x \in V', \quad \forall t \in [0, T] \text{ as } n \rightarrow +\infty,$$

*and, moreover, if  $\phi_n$  is the unique strong solution of*

$$\begin{cases} \phi_t(t, x) + F(t, \phi_x(t, x)) - \langle Ax, \phi_x(t, x) \rangle = g_n(T - t, x), & (t, x) \in [0, T] \times V', \\ \phi(0, x) = \varphi_n(x) \end{cases}$$

*in  $C([0, T], \mathcal{B}_2(V'))$ , then*

$$\phi_n(t, x) \uparrow \phi(t, x) \quad \forall (t, x) \in [0, T] \times V'.$$

Remark 4.2. Since strong solutions were proved in [34] to be Lipschitz with respect to the time variable and  $C^1$  with respect to the space variable, and the weak solution  $\phi$  is a sup-envelope of strong solutions  $\phi_n$ , then  $\phi$  is l.s.c. in  $[0, t] \times V'$ . For the same reason  $\phi_n$  convex in the  $x$  variable implies that  $\phi$  is convex in  $x$  as well.

Note also that by (i) in the definition, the weak solution  $\phi$  is finite on  $K$ .

THEOREM 4.3. *Let Assumptions 3.1 be satisfied. Assume that there exists a weak solution  $\phi$  of (2.5). Then, for any  $(t, x) \in [0, T] \times K$ , the following hold:*

- (i) *The solution  $\phi$  satisfies*

$$(4.1) \quad \phi(T - t, x) = \inf_{u(t, x)} J(t, x, u).$$

- (ii) *There exists a pair  $\{u^*, y^*\}$ , optimal at  $(t, x)$ , such that*

$$\phi(T - t, x) = J(t, x, u^*).$$

- (iii) *If  $\varphi_n$  and  $g_n$  are as described in Definition 4.1, then there exists a sequence of pairs  $\{(u_n^*, y_n^*)\}$ , optimal at  $(t, x)$  for the functional*

$$J_n(t, x, u) = \int_t^T [h(\tau, u(\tau)) + g_n(\tau, y(\tau))] d\tau + \varphi_n(y(T)),$$

*such that*

$$u^* = w - \lim_{n \rightarrow +\infty} u_n^* \quad \text{in } L^p(t, T; U)$$

*and*

$$y^*(s) = w - \lim_{n \rightarrow +\infty} y_n^*(s) \quad \text{in } V' \quad \forall s \in [t, T],$$

$$y^* = w - \lim_{n \rightarrow +\infty} y_n^* \quad \text{in } L^p(t, T; V').$$

*Proof.* First we recall that, due to Remark 3.2, weak solutions of (2.5) are well defined. Let  $\phi$  be a weak solution of (2.5) and let  $\{g_n\}$ ,  $\{\varphi_n\}$ , and  $\{\phi_n\}$  be as defined in Definition 4.1. Since  $\phi_n$  are strong solutions of (2.5), then

$$\phi_n(T-t, x) = \inf_{u \in L^p(t, T; U)} J_n(t, x, u),$$

as stated by Theorem 2.7. Moreover,  $g_n \leq g$  and  $\varphi_n \leq \varphi$  so that

$$\phi_n(T-t, x) \leq J_n(t, x, u) \leq J(t, x, u) \quad \forall t, x, u.$$

By taking the limit in the preceding estimate as  $n \rightarrow +\infty$  and then the infimum with respect to the control  $u$ , we obtain

$$(4.2) \quad \phi(T-t, x) \leq \inf_{u \in L^p(t, T; U)} J(t, x, u).$$

The reverse inequality can be proved as follows. Let  $t \in [0, T]$  and  $x \in K$ , and let  $\{u_n^*\}$  be the sequence of (unique) optimal controls for  $J_n(t, x, u)$  as described in Theorem 2.8, so that

$$\phi_n(T-t, x) = J_n(t, x, u_n^*).$$

Also let  $\{y_n^*\}$  be the associated trajectories.

We claim that

$$(4.3) \quad \sup_{n \in \mathbb{N}} \|u_n^*\|_{L^p(t, x; U)} < +\infty.$$

Indeed, note that

$$\|y_n^*\|_{L^\infty(t, T; V')} \leq C_0(|x|_{V'} + \|u_n^*\|_{L^p(t, T; U)}),$$

where

$$C_0 = C_0(T) = \max \left\{ e^{\omega T}, \frac{e^{q\omega T} - 1}{q\omega} \|B\|_{L(U, V')} \right\} \quad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1.$$

Moreover, since  $g_n$  and  $\varphi_n$  are increasing and  $g_n \in \mathcal{Y}([0, T], V')$ , there exist positive constants  $C_1 = C_1(T)$  and  $C_2$  (independent of  $x$  and  $t$ ) such that

$$g_n(t, x) \geq g_0(t, x) \geq -C_1(1 + |x|_{V'})$$

and

$$\varphi_n(x) \geq \varphi_0(x) \geq -C_2(1 + |x|_{V'}).$$

Recall also that  $h$  satisfies Assumptions 3.1. Then, for all  $n$ ,  $x$ , and  $t$ , we have

$$(4.4) \quad \begin{aligned} J_n(t, x, u_n^*) &\geq -C_1 \int_t^T (1 + |y_n^*(s)|_{V'}) ds + \int_t^T a(s) |u_n^*(s)|^p ds \\ &\quad + \int_t^T b(s) ds - C_2(1 + |y_n^*(T)|_{V'}) \\ &\geq \alpha \|u_n^*\|_{L^p(t, T; U)}^p - C \|u_n^*\|_{L^p(t, T; U)} - D, \end{aligned}$$

with  $\alpha = \alpha_T$  as in Assumptions 3.1, and

$$C := C(T) = C_0(C_2 + C_1T),$$

$$D := D(T, x) = |x|_{V'}(C_1T + C_2) + C_1T + C_2 + \|b\|_{L^1(0,T;\mathbb{R})}.$$

On the other hand, we have

$$(4.5) \quad J_n(t, x, u_n^*) \leq \phi(T - t, x) =: C_3,$$

with  $C_3 = C_3(t, x) < +\infty$  by definition for  $x \in K$ . Then from (4.4) and (4.5) we derive

$$\alpha \|u_n^*\|_{L^p(t,T;U)}^p - C \|u_n^*\|_{L^p(t,T;U)} \leq C_3 + D,$$

which implies (4.3).

Now (4.3) implies that there exists a subsequence  $\{u_{n_k}^*\}$  of  $\{u_n^*\}$  and  $u^* \in L^p(t, T; U)$  such that

$$u_{n_k}^* \rightharpoonup u^* \text{ weakly in } L^p(t, T; U).$$

Consequently, if  $y^*$  is the trajectory associated to the control  $u^*$ ,

$$y_{n_k}^*(s) \rightharpoonup y^*(s) \text{ weakly in } V' \quad \forall s \in [t, T]$$

and

$$y_{n_k}^* \rightharpoonup y^* \text{ in } L^p(t, T; V').$$

The proof of such facts is standard and we omit it.

Next we want to show that

$$(4.6) \quad \liminf_{k \rightarrow +\infty} J_{n_k}(t, x, u_{n_k}^*) \geq J(t, x, u^*).$$

To this purpose, note that  $u \mapsto \int_t^T h(s, u(s))ds$  being a convex function, it is also weakly l.s.c., so that

$$(4.7) \quad \liminf_{k \rightarrow +\infty} \int_t^T h(s, u_{n_k}^*(s))ds \geq \int_t^T h(s, u^*(s))ds.$$

Moreover, let  $m \in \mathbb{N}$  be fixed. Since  $\varphi_n$  is an increasing sequence, we have

$$\varphi_{n_k}(y_{n_k}^*(T)) \geq \varphi_m(y_{n_k}^*(T)) \quad \forall n_k \geq m.$$

Now

$$\liminf_{k \rightarrow +\infty} \varphi_{n_k}(y_{n_k}^*(T)) \geq \liminf_{k \rightarrow +\infty} \varphi_m(y_{n_k}^*(T)) \geq \varphi_m(y^*(T))$$

since  $\varphi_m$  is convex (and then weakly l.s.c.). By letting  $m \rightarrow +\infty$  we then obtain

$$(4.8) \quad \liminf_{k \rightarrow +\infty} \varphi_{n_k}(y_{n_k}^*(T)) \geq \varphi(y^*(T)).$$

In the same way we find

$$\liminf_{k \rightarrow +\infty} g_{n_k}(s, y_{n_k}^*(s)) \geq g(s, y^*(s)) \quad \forall s \in [t, T].$$

Moreover since the sequence  $g_n$  is increasing, then

$$g_{n_k}(s, y_{n_k}^*(s)) \geq g_{n_1}(s, y_{n_k}^*(s)) \geq -C(1 + |y_{n_k}^*(s)|_{V'}^2)$$

for some  $C > 0$ , as  $g_{n_1} \in \mathcal{Y}([0, T] \times V')$ . Since  $y_n \in L^p(t, T; V')$ , by Fatou's lemma we derive

$$(4.9) \quad \liminf_{k \rightarrow +\infty} \int_t^T g_{n_k}(s, y_{n_k}^*(s)) ds \geq \int_t^T g(s, y^*(s)) ds.$$

By means of (4.7)–(4.9) we then obtain (4.6), which implies

$$(4.10) \quad \phi(T - t, x) = \liminf_{k \rightarrow +\infty} J_{n_k}(t, x, u_{n_k}^*) \geq J(t, x, u^*).$$

Then, by means of (4.2), we have also

$$(4.11) \quad \phi(T - t, x) = \lim_{n \rightarrow +\infty} J_n(t, x, u_n^*) = \liminf_{k \rightarrow +\infty} J_{n_k}(t, x, u_{n_k}^*) = J(t, x, u^*).$$

Hence (4.1) holds and  $u^*$  is optimal.  $\square$

The following corollary easily follows.

**COROLLARY 4.4.** *If there exists a weak solution of (2.5), then it is unique.*

*Remark 4.5.* If the function  $h$  is strictly convex in  $u$ , then the optimal control  $u^*$  is unique and the whole sequence of optimal controls  $u_n^*$  converges weakly to  $u^*$  in  $L^p(t, T; U)$ . Moreover, if, for example,  $h(t, u) = h(u) = |u|_U^p$ , then

$$(4.12) \quad u_n^* \rightarrow u^* \text{ strongly in } L^p(t, T; U)$$

and

$$(4.13) \quad y_n^* \rightarrow y^* \text{ in } C([t, T], V').$$

Indeed we have just showed that  $J_n(t, x, u_n^*) \rightarrow J(t, x, u^*)$ , as  $n \rightarrow +\infty$ , so that

$$(4.14) \quad \int_t^T |u_n^*(\tau)|_U^p d\tau = \int_t^T |u^*(\tau)|_U^p d\tau + \int_t^T [g(\tau, y^*(\tau)) - g_n(\tau, y_n^*(\tau))] d\tau \\ + \varphi(y^*(\tau)) - \varphi_n(y_n^*(\tau)) + \sigma(n),$$

with  $\sigma(n) \rightarrow 0$  as  $n \rightarrow 0$ . Hence from (4.8)–(4.9) we obtain

$$\limsup_{k \rightarrow +\infty} \int_t^T |u_{n_k}^*(\tau)|_U^p d\tau \leq \int_t^T |u^*(\tau)|_U^p d\tau,$$

along a subsequence, which together with (4.7) implies  $u_{n_k}^* \rightarrow u^*$  strongly in  $L^p(t, T; U)$  as  $k \rightarrow +\infty$ . To show that the whole sequence of optimal controls  $u_n^*$  converges strongly in  $L^p$ , one assumes by contradiction that there exist an  $\varepsilon > 0$  and a subsequence  $\{u_{n_l}^*\}$  such that  $\|u_{n_l}^* - u^*\|_p \geq \varepsilon$  for all  $l \in \mathbb{N}$ . Then such a subsequence is itself bounded and has a subsequence converging weakly and, possibly passing to a further subsequence, strongly to some optimal control  $v^*$ . But as the optimal control is unique, this yields a contradiction. The proof of (4.13) is a straightforward consequence of the Lebesgue theorem.

*Remark 4.6.* Besides existence and uniqueness for the HJB equation, Cannarsa and Di Blasio [17] provide optimal feedback maps in the case when  $B$  is (bounded and) invertible—which is not true in the case of the economic application of section 6. Nevertheless, we show in an example how one can derive feedback controls when dealing with explicitly given data  $B$  and  $h$ .

**5. An existence result.** We may prove also existence of weak solutions when  $g$  has a particular dependence on the time variable, as stated and proved in the theorem below. To this extent we first prove the following lemma.

LEMMA 5.1. *If  $\phi_1$  and  $\phi_2$  are strong solutions of (2.5) associated to data  $\varphi_1 \leq \varphi_2$  and  $g_1 \leq g_2$ , respectively, then  $\phi_1 \leq \phi_2$ .*

*Proof.* Note that according to the properties of strong solutions proved in [33], for  $i = 1, 2$  we have

$$\phi_i(T - t, x) = \inf_{L^p(t, T; U)} J_i(t, x, u),$$

where

$$J_i(t, x, u) = \int_t^T [h(\tau, u(\tau)) + g_i(\tau, y(\tau))] d\tau + \varphi_i(y(T)).$$

Since the assumptions on the data imply

$$J_1(t, x, u) \leq J_2(t, x, u) \quad \forall u \in L^p(t, T; U),$$

by taking the infimum over  $L^p(t, T; U)$ , we get the desired inequality.  $\square$

THEOREM 5.2. *Assume that Assumptions 3.1 are satisfied, with  $g$  of the following type:*

$$(5.1) \quad g(t, x) = \gamma(x)\eta(t), \quad \gamma \in \Sigma_K, \quad \eta \in C([0, T], \mathbb{R}).$$

*Assume also that for each  $(t, x) \in [0, T] \times V'$  there exists an admissible control. Then there exists a weak solution of (2.5).*

*Proof.* We construct a sequence such as that in Definition 4.1 by using Yosida approximations. For given  $\lambda > 0$ , we define

$$(5.2) \quad \begin{aligned} \varphi_\lambda(x) &:= \inf_{y \in V'} \left\{ \varphi(y) + \frac{|x - y|_{V'}^2}{2\lambda} \right\}, \quad x \in V', \\ \gamma_\lambda(x) &:= \inf_{y \in V'} \left\{ \gamma(y) + \frac{|x - y|_{V'}^2}{2\lambda} \right\}, \quad x \in V'. \end{aligned}$$

It is well known (see, for instance, [11, Chap. 2]) that  $\varphi_\lambda, \gamma_\lambda$  are in  $\Sigma_0$ , with  $[\varphi'_\lambda]_L \leq \frac{1}{\lambda}$  and  $[\gamma'_\lambda]_L \leq \frac{1}{\lambda}$ . Moreover

$$\varphi_\lambda(x) \uparrow \varphi(x), \quad \gamma_\lambda(x) \uparrow \gamma(x) \quad \forall x \in V', \quad \forall t \in [0, T] \text{ as } \lambda \downarrow 0^+.$$

Then both  $\varphi_\lambda$  and  $\gamma_\lambda(\cdot)\eta(\cdot)$  satisfy Assumptions 3.1, so that there exists a unique strong solution  $\phi_\lambda$  of equation

$$\begin{cases} \phi_t(t, x) + \mathcal{H}(T - t, \phi_x(t, x)) - \langle Ax, \phi_x(t, x) \rangle = \gamma_\lambda(x)\eta(T - t), & (t, x) \in [0, T] \times V', \\ \phi(0, x) = \varphi_\lambda(x). \end{cases}$$

Moreover,  $\phi_\lambda$  satisfies

$$(5.3) \quad \phi_\lambda(T - t, x) = \inf_{L^p(t, T; U)} J_\lambda(t, x, u),$$

where

$$J_\lambda(t, x, u) = \int_t^T [h(\tau, u(\tau)) + g_\lambda(y(\tau))\eta(\tau)]d\tau + \varphi_\lambda(y(T)).$$

Since  $\varphi_\lambda(x) \uparrow \varphi(x)$  and  $\gamma_\lambda(x) \uparrow \gamma(x)$ , by Lemma 5.1 we have that for all  $x$  and  $t$  the sequence  $\{\phi_\lambda(t, x)\}_\lambda$  is increasing. We then set

$$\phi(t, x) := \sup_{\lambda > 0} \phi_\lambda(t, x) = \lim_{\lambda \rightarrow 0^+} \phi_\lambda(t, x).$$

By definition, the function  $\phi$  is l.s.c. in  $(t, x)$  and convex in  $x$  for all fixed  $t$ . It is then left to show that  $K \subset D(\phi(t, \cdot))$  for all fixed  $t \in [0, T]$ . Let  $x \in K$ ,  $t \in [0, T]$ , and let  $\bar{u}$  be an admissible control for the problem starting at  $(t, x)$ . Then

$$\phi_\lambda(T - t, x) \leq J_\lambda(t, x, \bar{u}) \leq J(t, x, \bar{u}) < +\infty,$$

and so

$$\phi(T - t, x) \leq J(t, x, \bar{u}) < +\infty. \quad \square$$

*Remark 5.3.* If  $g$  depends on time, one may set

$$g_\lambda(t, x) := \inf_{y \in V'} \left\{ g(t, y) + \frac{|x - y|_{V'}^2}{2\lambda} \right\}, \quad x \in V'.$$

Note that if  $x_0 \in D(g(t, \cdot))$ , then

$$g_\lambda(t, x) \leq g(t, x_0) + \frac{|x - x_0|_{V'}^2}{2\lambda},$$

so that  $g_\lambda(t, \cdot) \in \mathcal{B}_2(V')$  for all  $t$ . If one is able to show that

$$g_\lambda \in C([0, T], \mathcal{B}_2(V')) \quad \text{and} \quad (g_\lambda)_x \in C([0, T], \mathcal{B}_1(V')),$$

then the proof above may be performed as before, as both  $\varphi_\lambda$  and  $g_\lambda$  satisfy the assumptions in Theorem 2.6 (see also Remark 5.2 in [33]).

Theorems 5.2 and 4.3 imply the following result.

**COROLLARY 5.4.** *Let Assumptions 3.1 be satisfied, with  $g$  as in (5.1). Then the following properties are equivalent:*

- (i) *there exists a unique weak solution of (2.5);*
  - (ii) *at each  $(t, x) \in [0, T] \times K$  there exists an admissible control.*
- Moreover if (i) or (ii) holds, there exists an optimal pair  $(u^*, y^*)$  and*

$$\phi(T - t, x) = J(t, x, u^*).$$

**6. Applications to vintage capital problems with state constraints.** We here describe our motivating example: the problem of optimal investment with vintage capital in the setting introduced by Barucci and Gozzi [13, 12], and later reprised and generalized by Feichtinger et al. [41, 38, 39] and Faggian [33, 34].

The capital accumulation is described by the system

$$(6.1) \quad \begin{cases} \frac{\partial y(\tau, s)}{\partial \tau} + \frac{\partial y(\tau, s)}{\partial s} + \mu y(\tau, s) = u_1(\tau, s), & (\tau, s) \in ]t, T[ \times ]0, \bar{s}], \\ y(\tau, 0) = u_0(\tau), & \tau \in ]t, T[, \\ y(t, s) = x(s), & s \in [0, \bar{s}], \end{cases}$$

with  $t > 0$  the initial time,  $\bar{s} \in [0, +\infty[$  the maximal allowed age, and  $\tau \in [0, T[$  with horizon  $T < +\infty$ . The unknown  $y(\tau, s)$  represents the amount of capital goods of age  $s$  accumulated at time  $\tau$ , the initial datum is a function  $x \in L^2(0, \bar{s})$ , and  $\mu > 0$  is a depreciation factor. Moreover,  $u_0 : [t, T[ \rightarrow \mathbb{R}$  is the investment in new capital goods ( $u_0$  is the boundary control) while  $u_1 : [t, T[ \times [0, \bar{s}] \rightarrow \mathbb{R}$  is the investment at time  $\tau$  in capital goods of age  $s$  (hence, the distributed control). Investments are jointly referred to as the control  $u = (u_0, u_1)$ . Note that such problems are known as *vintage capital* problems, for the capital goods depend jointly on time  $\tau$  and on age  $s$ , which is equivalent to their dependence on time and vintage  $\tau - s$ .

In addition, we consider the firm profits represented by the functional

$$I(t, x; u_0, u_1) = \int_t^T e^{-\lambda\tau} [R(Q(\tau)) - c(u(\tau))] d\tau + e^{-\lambda T} R_0(Q(T)),$$

where, for some given measurable coefficient  $\alpha$ , we have that

$$Q(\tau) = \int_0^{\bar{s}} \alpha(s) y(\tau, s) ds$$

is the output rate (linear in  $y(\tau)$ ), and  $R, R_0$  are concave revenues from the output  $Q$ . Moreover we set

$$c(u_0, u_1) = \beta_0 |u_0|^p + \gamma_0 u_0 + \beta_1 |u_1|_H^p + (u_1 | \gamma_1)_H, \quad (u_0, u_1) \in \mathbb{R} \times L^2(0, \bar{s}),$$

with  $\beta_1 |u_1|_H^p + (u_1 | \gamma_1)_H$  indicating the running investment cost for technologies of age greater than 0, and with  $\beta_0 |u_0|^p + \gamma_0 u_0$  the investment cost in new technologies, including adjustment innovation.

The entrepreneur's problem is that of maximizing  $I(t, x; u_0, u_1)$  over all controls  $(u_0, u_1)$  chosen in a set such that the corresponding trajectories are positive, that is, such that the following constraint on the state is satisfied:

$$(6.2) \quad y(\tau, s; t, x, u_0, u_1) \geq 0 \quad \forall \tau, s.$$

When rephrased in an infinite-dimensional setting, with  $H := L^2(0, \bar{s})$  as state space, the state equation (6.1) can be reformulated as a linear control system with an unbounded control operator, that is,

$$(6.3) \quad \begin{cases} y'(\tau) = A_0 y(\tau) + B u(\tau), & \tau \in ]t, T[, \\ y(t) = x, \end{cases}$$

where  $y : [t, T] \rightarrow H$ ,  $x \in H$ ,  $A_0 : D(A_0) \subset H \rightarrow H$  is the infinitesimal generator of a strongly continuous semigroup  $\{e^{A_0 t}\}_{t \geq 0}$  on  $H$  with domain  $D(A_0) = \{f \in H^1(0, \bar{s}) : f(0) = 0\}$  and defined as  $A_0 f(s) = -f'(s) - \mu f(s)$ , the control space is  $U = \mathbb{R} \times H$ , the control function is a couple  $u \equiv (u_0, u_1) : [t, T] \rightarrow \mathbb{R} \times H$ , and the control operator is given by  $Bu \equiv B(u_0, u_1) = u_1 + u_0 \delta_0$  for all  $(u_0, u_1) \in \mathbb{R} \times H$ ,  $\delta_0$  being the Dirac delta at the point 0. Although  $B \notin L(U, H)$ , note that  $B \in L(U, D(A_0^*))$ .

Following Remark 2.2, we then set

$$V = D(A_0^*) = \{f \in H^1(0, \bar{s}) : f(\bar{s}) = 0\}$$

and  $V' = D(A_0^*)'$ . Such abstract rephrasing is discussed in detail in [34], to which the reader is referred.



Regarding the target functional, we set

$$J_T(t, x; u) := -I(t, x; u_0, u_1),$$

where

- $g_0 : [0, T] \times H \rightarrow \mathbb{R}$ ,  $g(\tau, x) = -e^{-\lambda\tau} R((\alpha|x)_H)$ ;
- $h : [0, T] \times U \rightarrow \mathbb{R}$ ,  $h(t, u_0, u_1) = e^{-\lambda t} [\beta_0 |u_0|^p + \gamma_0 u_0 + \beta_1 |u_1|_H^p + (u_1|\gamma_1)_H]$ ;
- $\phi_0 : H \rightarrow \mathbb{R}$ ,  $\phi_0(x) = -e^{-\lambda T} R_0((\alpha|x)_H)$ .

Moreover

$$\mathcal{C} := \{x \in L^2(0, \bar{s}) : x(s) \geq 0 \text{ for a.e. } s \in [0, \bar{s}]\},$$

and, for any  $(t, x) \in [0, T] \times \mathcal{C}$ , the control  $u$  is chosen in the set of admissible controls

$$\mathcal{U}(t, x) := \{u \in L^p(0, T; U) : y(\tau; t, x, u) \in \mathcal{C}\}.$$

Note that  $\mathcal{C}$  is a convex closed subset of  $H$ , whose closure in  $V'$  is

$$K = cl_{V'}(\mathcal{C}) = \{z \in V' : \langle z, v \rangle \geq 0 \ \forall v \in V, v \geq 0\}.$$

*Remark 6.1.* Note that  $K$  is invariant with respect to the semigroup, that is,

$$e^{\tau A}(K) \subset K \ \forall \tau \in [t, T].$$

Indeed  $\langle e^{(\tau-t)A}x, v \rangle = \langle x, e^{(\tau-t)A^*}v \rangle \geq 0$ , as (see [34]) we have

$$[e^{(\tau-t)A^*}v](s) = e^{-\mu(\tau-t)}v(s + \tau - t)\chi_{[0, \bar{s}-\tau+t]}(s) \geq 0 \ \forall s \in [0, \bar{s}], \ \forall v \in V.$$

*Remark 6.2.* If we want to ensure Assumptions 3.1 are satisfied, we need to better specify the properties of the cost and the revenue functions. For instance, if we assume  $\alpha$  to be more regular, say  $\alpha \in V$ , and  $R$  and  $R_0$  concave in  $\mathbb{R}$  (hence continuous), then Assumptions 3.3 are satisfied by setting  $g(\tau, x) = -e^{-\lambda\tau} R(\langle a, x \rangle)$  when  $x \in K$  and  $g(\tau, x) = +\infty$  elsewhere, and similarly  $\varphi(x) = -e^{-\lambda T} R_0(\langle a, x \rangle)$  when  $x \in K$  and  $\varphi(x) = +\infty$  elsewhere. Assumptions 3.1(4–7) are then also satisfied.

Then the following existence and uniqueness result for the problem holds.

**THEOREM 6.3.** *Assume that the data of the problem described above satisfy Assumptions 3.1, with  $p \geq 2$ . Assume also  $R$  bounded on bounded subsets, and  $\alpha \in V'$ . Let  $t \in [0, T]$  and  $x \in K$ . Then*

- (i) *there exists a unique weak solution  $\phi$  to (2.5);*
- (ii) *for every  $t$  in  $[0, T]$  and every  $x$  in  $K$ , there exists  $(u^*, y^*)$  optimal at  $(t, x)$  for the optimal control problem, with  $y^*(\tau) \in K$  for every  $\tau$  in  $[t, T]$ ;*
- (iii)  $J(t, x, u^*) = \phi(T - t, x)$ .

*Proof.* For all  $t$  and  $x$  in  $K$  the null control is admissible. Indeed

$$|J(t, x, 0)| = \left| \int_t^T e^{-\lambda\tau} R(\langle \alpha, e^{(\tau-t)A}x \rangle) d\tau \right| \leq M \frac{e^{-\lambda t} - e^{-\lambda T}}{\lambda} < +\infty,$$

where  $M = \sup\{|R(y)| : |y| \leq |\alpha|_V e^{\omega T} |x|_{V'}\}$ . The proof is then a straightforward application of Corollary 5.4.  $\square$

Once the problem is solved in the extended state space  $V'$ , one has to go back to the original problem set in  $H$ . In order to do so we recall that whenever the initial datum  $x$  is in  $H$ , the whole trajectory lies in  $H$ , as Barucci and Gozzi prove in their paper [13].

THEOREM 6.4. *Given any initial datum  $x \in H$  and control  $u \in L^p(t, T; U)$ , the mild solution of (6.3)*

$$y(s) = e^{(s-t)A}x + \int_t^s e^{(s-\tau)A}Bu(\tau)d\tau$$

*belongs to  $C([t, T]; H)$ .*

More precisely, it is shown in [34] that whenever  $x$  is in  $H$ , we may write the solution also as

$$(6.4) \quad y(\tau) = e^{-(\tau-t)A_0}x + \int_t^\tau e^{-(\tau-r)A_0}u_1(r)dr - A_0 \int_t^\tau e^{-(\tau-r)A_0}w(r)dr,$$

where  $w(r)(s) = e^{-\mu s}u_0(r)$ , or even more explicitly as

$$(6.5) \quad \begin{aligned} y(\tau, s) &= e^{-\mu(\tau-t)}x(s-\tau+t)\chi_{[\tau-t, \bar{s}]}(s) \\ &+ u_0(\tau-s)e^{-\mu s}\chi_{[0, (\tau-t) \wedge \bar{s}]}(s) + \int_0^{(\tau-t) \wedge s} e^{-\mu q}u_1(\tau-q, s-q)dq. \end{aligned}$$

Then by means of the preceding theorem and Remark 3.4(c), we may infer that the state constraints are satisfied by the optimal trajectory any time the initial datum lies in  $\mathcal{C}$ , as stated next.

THEOREM 6.5. *Let  $(t, x) \in [0, T] \times K$  and let  $(u^*, y^*)$  be optimal at  $(t, x)$ . Then*

$$y^*(\tau) \in \mathcal{C} \quad \forall \tau \in [t, T].$$

**6.1. Feedback controls.** We intend to show here that once the running cost (or profit, as in the economic example) is explicitly given as in the example we are treating, one may expect to derive a more meaningful feedback formula for optimal controls than the one obtained by the limiting procedure in Theorem 4.3.

After some easy but tedious computations, one obtains

$$(6.6) \quad \mathcal{H}(t, u_0, u_1) = (q-1)^{q-1}q^{-q}e^{-\lambda t}[\beta_0|e^{\lambda t}u_0 - \gamma_0|^q + \beta_1|e^{\lambda t}u_1 - \gamma_1|_H^q],$$

where  $q = p/(p-1)$ . Note that as  $h$  is strictly convex, the optimal control is unique, as we remarked in section 4. Moreover it is easily shown that  $\mathcal{H}$  satisfies Assumptions 3.1(7).

LEMMA 6.6. *Let the assumptions of Theorem 6.3 be satisfied, and let  $p \geq 2$ . Let  $(u^*, y^*)$  be the unique optimal couple at  $(t, x)$ , and let  $(u_n^*, y_n^*)$  be the associated sequence of optimal couples of the approximating problems. Then*

- (i)  $u_{0n}^* \rightarrow u_0^*$  strongly in  $L^p(t, T)$ ,<sup>3</sup> and  $u_{1n}^* \rightarrow u_1^*$  strongly in  $L^p(t, T; H)$ ;
- (ii) if  $x \in \mathcal{C}$ , then  $y_n^*(\tau) \rightarrow y^*(\tau)$  strongly in  $H$  for all  $\tau \in [t, T]$ .

*Proof.* By making use of the fact that the optimal control is unique and  $J_n(u_n^*) \rightarrow J(u^*)$  as  $n \rightarrow +\infty$ , by weak convergence of both  $u_{0n}^*$  and  $u_{1n}^*$  to  $u_0^*$  and  $u_1^*$ , respectively, and proceeding as in Remark 4.5, one may show that along a subsequence

$$(6.7) \quad \begin{aligned} \limsup_{k \rightarrow +\infty} \int_t^T e^{-\lambda \tau} [\beta_0|u_{0n_k}^*(\tau)|^p + \beta_1|u_{1n_k}^*(\tau)|_H^p] d\tau \\ \leq \int_t^T e^{-\lambda \tau} [\beta_0|u_0^*(\tau)|^p + \beta_1|u_1^*(\tau)|_H^p] d\tau. \end{aligned}$$

<sup>3</sup> $L^p(t, T) \equiv L^p(t, T; \mathbb{R})$ .

Since also

$$(6.8) \quad \begin{aligned} \liminf_{n \rightarrow +\infty} \int_t^T e^{-\lambda\tau} [\beta_0 |u_{0n}^*(\tau)|^p + \beta_1 |u_{1n}^*(\tau)|_H^p] d\tau \\ \geq \int_t^T e^{-\lambda\tau} [\beta_0 |u_0^*(\tau)|^p + \beta_1 |u_1^*(\tau)|_H^p] d\tau, \end{aligned}$$

then the limit exists along the subsequence  $\{u_{n_k}^*\}$  and coincides with the right-hand side. Again as in Remark 4.5, one shows that the same assertion yields for the whole sequence  $\{u_n^*\}$ , that is,

$$(6.9) \quad \lim_{n \rightarrow +\infty} \int_t^T e^{-\lambda\tau} [\beta_0 |u_{0n}^*(\tau)|^p + \beta_1 |u_{1n}^*(\tau)|_H^p] d\tau = \int_t^T e^{-\lambda\tau} [\beta_0 |u_0^*(\tau)|^p + \beta_1 |u_1^*(\tau)|_H^p] d\tau.$$

Note that (6.9) and the weak convergence of  $u_n^*$  to  $u_0^*$  imply

$$\begin{aligned} \liminf_{n \rightarrow \infty} \int_t^T \beta_0 e^{-\lambda\tau} |u_{0n}^*(\tau)|^p d\tau &= \int_t^T \beta_0 e^{-\lambda\tau} |u_0^*(\tau)|^p d\tau, \\ \liminf_{n \rightarrow \infty} \int_t^T \beta_1 e^{-\lambda\tau} |u_{1n}^*(\tau)|_H^p d\tau &= \int_t^T \beta_1 e^{-\lambda\tau} |u_1^*(\tau)|_H^p d\tau, \end{aligned}$$

so that one derives that (i) is satisfied at least along a subsequence and, by a standard argument, along the whole sequence. The proof of (ii) is performed by making use of the particular structure of the solution of the economic problem, given by (6.4) and (6.5). Then

$$(6.10) \quad \begin{aligned} &|y_n^*(\tau) - y^*(\tau)|_H \\ &\leq \int_t^\tau e^{(\tau-r)A_0} |u_{1n}^*(r) - u_1^*(r)|_H dr + \left| A_0 \int_t^\tau e^{-(\tau-r)A_0} (w_n(r) - w(r)) dr \right|_H \\ &\equiv I_1^n(\tau) + I_2^n(\tau), \end{aligned}$$

with  $w_n(\tau)(s) = e^{-\mu s} u_{0n}^*(\tau)$ . Now

$$I_1^n(\tau) \leq C \|u_{1n}^* - u_1^*\|_{L^1(t, T; H)} \leq C_1 \|u_{1n}^* - u_1^*\|_{L^p(t, T; H)}$$

for suitable positive constants  $C$  and  $C_1$ , while by means of (6.5) we have

$$(6.11) \quad \begin{aligned} I_2^n(\tau)^2 &\leq \int_0^{\bar{s} \wedge (\tau-t)} e^{-2\mu s} |(u_{0n}^* - u^*)(\tau-s)|^2 ds \\ &\leq \|u_{0n}^* - u^*\|_{L^2(t, T)}^2 \leq C \|u_{0n}^* - u^*\|_{L^p(t, T)}^2, \end{aligned}$$

which together yields (ii).  $\square$

**THEOREM 6.7.** *In the assumptions of Theorem 6.3, let  $\psi$  be the restriction of  $\phi$  to the set  $[0, T] \times H$ , that is,*

$$\psi : [0, T] \times H \rightarrow \mathbb{R}, \quad \psi(t, x) := \phi(t, x),$$

and let

$$\partial_x \psi(t, x) = \{h \in H : \psi(t, y) - \psi(t, x) \geq (h|y - x)_H \quad \forall y \in H\}$$

be its spatial subgradient. Also let  $(u^*, y^*)$  be the optimal couple at  $(t, x) \in [t, T] \times H$ , and  $\tau \in [t, T]$ . Then

(i)  $u_0^*(\tau) = \lim_{n \rightarrow +\infty} u_{0n}^*(\tau)$ , where

$$u_{0n}^*(\tau) = -\beta_0 p^{1-q} |e^{\lambda\tau} \phi_x^n(T - \tau, y_n^*(\tau))(0) + \gamma_0|^{q-2} (e^{\lambda\tau} \phi_x^n(T - \tau, y_n^*(\tau))(0) + \gamma_0);$$

(ii) there exists a selection  $\xi(\tau) \in \partial_x \psi(T - \tau, y^*(\tau))$  such that

$$u_1^*(\tau) = -\beta_1 p^{1-q} |e^{\lambda\tau} \xi(\tau) + \gamma_1|_H^{q-2} (e^{\lambda\tau} \xi(\tau) + \gamma_1).$$

*Proof.* Let  $(t, x)$  be fixed in  $[0, T] \times K$ , and let  $u_n^*$  be the sequence of controls, optimal at  $(t, x)$ , of the approximating problems, as in Theorem 4.3. It is easy to show that the operator  $B : U \rightarrow V'$ ,  $Bu = u_1 + \delta_0 u_0$  has adjoint

$$B^* : V \rightarrow U, \quad B^*v := (v(0), v) \in \mathbb{R} \times V,$$

so that by means of the feedback formula in Theorem 2.8 applied to optimal controls  $u_n^*$ , one has

(6.12)

$$u_{0n}^*(\tau) = -\beta_0 p^{1-q} |e^{\lambda\tau} \phi_x^n(T - \tau, y_n^*(\tau))(0) + \gamma_0|^{q-2} (e^{\lambda\tau} \phi_x^n(T - \tau, y_n^*(\tau))(0) + \gamma_0),$$

which proves the first assertion in the statement of the theorem, while

$$(6.13) \quad u_{1n}^*(\tau) = -\beta_1 p^{1-q} |e^{\lambda\tau} \phi_x^n(T - \tau, y_n^*(\tau)) + \gamma_1|_H^{q-2} (e^{\lambda\tau} \phi_x^n(T - \tau, y_n^*(\tau)) + \gamma_1).$$

By reversing (6.13) one obtains

$$(6.14) \quad \phi_x^n(T - \tau, y_n^*(\tau)) = -\beta_1^{1-p} p e^{-\lambda\tau} |u_{1n}^*(\tau)|_H^{p-2} u_{1n}^*(\tau) + e^{-\lambda\tau} \gamma_1,$$

so that by Lemma 6.6(i) there exists  $k \mapsto n_k$  such that for almost every  $\tau$  in  $[t, T]$  one has

$$(6.15) \quad \xi(\tau) := H - \lim_{k \rightarrow \infty} \phi_x^{n_k}(T - \tau, y_{n_k}^*(\tau)) = -\beta_1^{1-p} p e^{-\lambda\tau} |u_1^*(\tau)|_H^{p-2} u_1^*(\tau) + e^{-\lambda\tau} \gamma_1.$$

From

$$\phi^{n_k}(T - \tau, y) - \phi^{n_k}(T - \tau, y_{n_k}^*(\tau)) \geq \langle \phi_x^{n_k}(T - \tau, y_{n_k}^*(\tau)), y - y_{n_k}^*(\tau) \rangle_{V, V'} \quad \forall y \in V',$$

and recalling that when  $x$  is in  $H$  the whole trajectory  $y_{n_k}^*(\tau)$  lies in  $H$  for all  $\tau$ , we derive

$$(6.16) \quad \phi^{n_k}(T - \tau, y) - \phi^{n_k}(T - \tau, y_{n_k}^*(\tau)) \geq (\phi_x^{n_k}(T - \tau, y_{n_k}^*(\tau)) | y - y_{n_k}^*(\tau))_H \quad \forall y \in H.$$

We now would like to derive from the preceding inequality the following estimate:

$$(6.17) \quad \psi(T - \tau, y) - \psi(T - \tau, y^*(\tau)) \geq (\xi(\tau) | y - y^*(\tau))_H \quad \forall y \in H \text{ for a.e. } \tau \in [t, T].$$

Indeed, by means of the same argument that led to (4.8), we get

$$\phi(T - \tau, y) - \phi(T - \tau, y^*(\tau)) \geq \limsup_{k \rightarrow \infty} (\phi^{n_k}(T - \tau, y) - \phi^{n_k}(T - \tau, y_{n_k}^*(\tau))).$$

On the other hand, for almost every  $\tau$  in  $[t, T]$ ,  $\phi_x^{n_k}(T - \tau, y_n^*(\tau))$  converges to  $\xi(\tau)$  in  $H$ , and  $y_n^*(\tau)$  converges to  $y^*(\tau)$ , both strongly in  $H$  (Lemma 6.6(ii)), so that

$$(\phi_x^{n_k}(T - \tau, y_n^*(\tau)) \mid y - y_n^*(\tau))_H \rightarrow (\xi(\tau) \mid y - y^*(\tau))_H \text{ for a.e. } \tau \in [t, T].$$

Hence,

$$\xi(\tau) \in \partial_x \psi(T - \tau, y^*(\tau)) \text{ for a.e. } \tau \in [t, T],$$

and passing to limits in (6.13),

$$u_1^*(\tau) = -\beta_1 p^{1-q} |e^{\lambda\tau} \xi(\tau) + \gamma_1|_H^{q-2} (e^{\lambda\tau} \xi(\tau) + \gamma_1). \quad \square$$

#### REFERENCES

- [1] P. ACQUISTAPACE, F. FLANDOLI, AND B. TERRENI, *Initial boundary value problems and optimal control for nonautonomous parabolic systems*, SIAM J. Control Optim., 29 (1991), pp. 89–118.
- [2] P. ACQUISTAPACE AND B. TERRENI, *Infinite-horizon linear-quadratic regulator problems for nonautonomous parabolic systems with boundary control*, SIAM J. Control Optim., 34 (1996), pp. 1–30.
- [3] P. ACQUISTAPACE AND B. TERRENI, *Classical solutions of nonautonomous Riccati equations arising in parabolic boundary control problems*, Appl. Math. Optim., 39 (1999), pp. 361–409.
- [4] P. ACQUISTAPACE AND B. TERRENI, *Classical solutions of nonautonomous Riccati equations arising in parabolic boundary control problems II*, Appl. Math. Optim., 41 (2000), pp. 199–226.
- [5] CH. ALMEDER, J. P. CAULKINS, G. FEICHTINGER, AND G. TRAGLER, *Age-structured single-state drug initiation model cycles of drug epidemics and optimal prevention programs*, Socio-Economic Planning Sciences, 38 (2004), pp. 91–109.
- [6] S. ANIȚA, M. IANNELLI, M.-Y. KIM, AND E.-J. PARK, *Optimal harvesting for periodic age-dependent population dynamics*, SIAM J. Appl. Math., 58 (1998), pp. 1648–1666.
- [7] V. BARBU AND G. DA PRATO, *Hamilton–Jacobi Equations in Hilbert Spaces*, Pitman, London, 1983.
- [8] V. BARBU AND G. DA PRATO, *Hamilton–Jacobi equations in Hilbert spaces; variational and semigroup approach*, Ann. Mat. Pura Appl. (4), 42 (1985), pp. 303–349.
- [9] V. BARBU AND G. DA PRATO, *A note on a Hamilton–Jacobi equation in Hilbert space*, Nonlinear Anal., 9 (1985), pp. 1337–1345.
- [10] V. BARBU, G. DA PRATO, AND C. POPA, *Existence and uniqueness of the dynamic programming equation in Hilbert spaces*, Nonlinear Anal., 7 (1983), pp. 283–299.
- [11] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Editura Academiei, Bucharest, 1986.
- [12] E. BARUCCI AND F. GOZZI, *Investment in a vintage capital model*, Res. Economics, 52 (1998), pp. 159–188.
- [13] E. BARUCCI AND F. GOZZI, *Technology adoption and accumulation in a vintage capital model*, J. Economics, 74 (2001), pp. 1–30.
- [14] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite-Dimensional Systems*, Vol. 1, Birkhäuser, Boston, 1992.
- [15] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite-Dimensional Systems*, Vol. 2, Birkhäuser, Boston, 1993.
- [16] R. BOUCEKKINE, O. LICANDRO, L. A. PUCH, AND F. DEL RIO, *Vintage capital and the dynamics of the AK model*, J. Econom. Theory, 120 (2005), pp. 39–72.
- [17] P. CANNARSA AND G. DI BLASIO, *A direct approach to infinite dimensional Hamilton–Jacobi equations and applications to convex control with state constraints*, Differential Integral Equations, 8 (1995), pp. 225–246.
- [18] P. CANNARSA, F. GOZZI, AND H. M. SONER, *A boundary value problem for Hamilton–Jacobi equations in Hilbert spaces*, Appl. Math. Optim., 24 (1991), pp. 197–220.
- [19] P. CANNARSA, F. GOZZI, AND H. M. SONER, *A dynamic programming approach to nonlinear boundary control problems of parabolic type*, J. Funct. Anal., 117 (1993), pp. 25–61.

- [20] P. CANNARSA AND M. E. TESSITORE, *Infinite-dimensional Hamilton–Jacobi equations and Dirichlet boundary control problems of parabolic type*, SIAM J. Control Optim., 34 (1996), pp. 1831–1847.
- [21] M. G. CRANDALL AND P.-L. LIONS, *Hamilton–Jacobi equations in infinite dimensions. Part I: Uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379–396.
- [22] M. G. CRANDALL AND P.-L. LIONS, *Hamilton–Jacobi equations in infinite dimensions. Part II: Existence of viscosity solutions*, J. Funct. Anal., 65 (1986), pp. 368–405.
- [23] M. G. CRANDALL AND P.-L. LIONS, *Hamilton–Jacobi equations in infinite dimensions. Part III*, J. Funct. Anal., 68 (1986), pp. 214–247.
- [24] M. G. CRANDALL AND P.-L. LIONS, *Hamilton–Jacobi equations in infinite dimensions. Part IV: Hamiltonians with unbounded linear terms*, J. Funct. Anal., 90 (1990), pp. 237–283.
- [25] M. G. CRANDALL AND P.-L. LIONS, *Hamilton–Jacobi equations in infinite dimensions. Part V: Unbounded linear terms and B-continuous solutions*, J. Funct. Anal., 97 (1991), pp. 417–465.
- [26] M. G. CRANDALL AND P.-L. LIONS, *Hamilton–Jacobi equations in infinite dimensions. Part VI: Nonlinear A and Tataru’s method refined*, in Evolution Equations, Control Theory, and Biomathematics (Han sur Lesse, 1991), Lecture Notes in Pure and Appl. Math. 155, Dekker, New York, 1994, pp. 51–89.
- [27] M. G. CRANDALL AND P.-L. LIONS, *Hamilton–Jacobi equations in infinite dimensions. Part VII: The HJB equation is not always satisfied*, J. Funct. Anal., 125 (1994), pp. 111–148.
- [28] G. DI BLASIO, *Global solutions for a class of Hamilton–Jacobi equations in Hilbert spaces*, Numer. Funct. Anal. Optim., 8 (1985/1986), pp. 261–300.
- [29] G. DI BLASIO, *Optimal control with infinite horizon for distributed parameter systems with constrained controls*, SIAM J. Control Optim., 29 (1991), pp. 909–925.
- [30] G. FABBRI, *A viscosity solution approach to the infinite-dimensional HJB equation related to a boundary control problem in a transport equation*, SIAM J. Control Optim., 47 (2008), pp. 1022–1052.
- [31] S. FAGGIAN, *Boundary control problems with convex cost and dynamic programming in infinite dimension Part I: The maximum principle*, Differential Integral Equations, 17 (2004), pp. 1149–1174.
- [32] S. FAGGIAN, *Boundary control problems with convex cost and dynamic programming in infinite dimension Part II: Hamilton–Jacobi–Bellman equation*, Discrete Contin. Dyn. Syst., 12 (2005), pp. 323–346.
- [33] S. FAGGIAN, *Regular solutions of Hamilton–Jacobi equations arising in Economics*, Appl. Math. Optim., 51 (2005), pp. 123–162.
- [34] S. FAGGIAN, *Applications of dynamic programming to economic problems with vintage capital*, Dyn. Contin. Discrete Impuls. Syst., to appear.
- [35] S. FAGGIAN AND F. GOZZI, *On the dynamic programming approach for optimal control problems of PDE’s with age structure*, Math. Popul. Stud., 11 (2004), pp. 233–270.
- [36] S. FAGGIAN AND F. GOZZI, *Dynamic Programming for Infinite Horizon Boundary Control Problems of PDE’s with Vintage Capital*, preprint, Dipartimento di Matematica, Università di Pisa, Italy, 2006.
- [37] H. O. FATTORINI, *Infinite-Dimensional Optimization and Control Theory*, Encyclopedia Math. Appl. 62, Cambridge University Press, Cambridge, UK, 1999.
- [38] G. FEICHTINGER, R. F. HARTL, P. M. KORT, AND V. M. VELIOV, *Dynamic investment behavior taking into account ageing of the capital goods*, in Dynamical Systems and Control, Stability Control Theory Methods Appl. 22, Chapman & Hall/CRC, Boca Raton, FL, 2004, pp. 379–391.
- [39] G. FEICHTINGER, R. F. HARTL, P. M. KORT, AND V. M. VELIOV, *Anticipation effects of technological progress on capital accumulation: A vintage capital approach*, J. Econom. Theory, 126 (2006), pp. 143–164.
- [40] G. FEICHTINGER, R. HARTL, AND S. SETHI, *Dynamical optimal control models in advertising: Recent developments*, Management Sci., 40 (1994), pp. 195–226.
- [41] G. FEICHTINGER, G. TRAGLER, AND V. M. VELIOV, *Optimality conditions for age-structured control systems*, J. Math. Anal. Appl., 288 (2003), pp. 47–68.
- [42] F. GOZZI, *Some results for an optimal control problem with a semilinear state equation I*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8), 82 (1988), pp. 423–429.
- [43] F. GOZZI, *Some results for an infinite horizon control problem governed by a semilinear state equation*, in Control and Estimation of Distributed Parameter Systems (Vorau, 1988), F. Kappel, K. Kunisch, and W. Schappacher, eds., Internat. Ser. Numer. Math. 91, Birkhäuser-Verlag, Basel, 1989, pp. 145–163.

- [44] F. GOZZI, *Some results for an optimal control problem with semilinear state equation*, SIAM J. Control Optim., 29 (1991), pp. 751–768.
- [45] F. GOZZI AND C. MARINELLI, *Optimal Advertising under Uncertainty with Memory and Lags*, manuscript.
- [46] M. IANNELLI, *Mathematical Theory of Age-Structured Population Dynamics*, Giardini Editore, Pisa, 1995.
- [47] M. KOCAN AND P. SORAVIA, *A viscosity approach to infinite-dimensional Hamilton–Jacobi equations arising in optimal control with state constraints*, SIAM J. Control Optim., 36 (1998), pp. 1348–1375.
- [48] I. LASIECKA AND R. TRIGGIANI, *Control Theory for Partial Differential Equations: Continuous and Approximation Theory*, I. *Abstract Parabolic Systems*, Encyclopedia Math. Appl. 74, Cambridge University Press, Cambridge, UK, 2000.
- [49] I. LASIECKA AND R. TRIGGIANI, *Control Theory for Partial Differential Equations: Continuous and Approximation Theory*, II. *Abstract Hyperbolic-Like Systems over a Finite Time Horizon*, Encyclopedia Math. Appl. 75, Cambridge University Press, Cambridge, UK, 2000.
- [50] J.-L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vol. 1, Dunod, Paris, 1968.
- [51] J.-L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vol. 2, Dunod, Paris, 1968.
- [52] J.-L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vol. 3, Dunod, Paris, 1969.
- [53] C. MARINELLI, *Optimal Advertising under Uncertainty*, Ph.D. thesis, draft.
- [54] M. NERLOVE AND J. K. ARROW, *Optimal advertising policy under dynamic conditions*, Economica, 29 (1962), pp. 129–142.
- [55] J. YONG AND X. Y. ZHOU, *Stochastic Controls. Hamiltonian Systems and HJB Equations*, Appl. Math. 43, Springer-Verlag, New York, 1999.

# CONTROLLABILITY PROBLEMS FOR THE STRING EQUATION ON A HALF-AXIS WITH A BOUNDARY CONTROL BOUNDED BY A HARD CONSTANT\*

L. V. FARDIGOLA<sup>†</sup>

**Abstract.** In this paper necessary and sufficient conditions for null-controllability and approximate null-controllability are obtained for the control system  $w_{tt} = w_{xx} - q^2 w$ ,  $w_x(0, t) = u(t)$ ,  $x > 0$ ,  $t \in (0, T)$ , where  $q \geq 0$ ,  $T > 0$ , and  $u$  is a control bounded by a hard constant. These problems are considered in the Sobolev spaces. Controls solving these problems are found explicitly. Bang-bang controls solving the approximate null-controllability problem are found by means of the solutions of the Markov power moment problem. Continuous controls solving this problem are constructed with the aid of the Cesàro means of a Fourier series generated by the data of the control system.

**Key words.** wave equation, half-axis, controllability problem, control bounded by a hard constant, Neumann control, Fourier transform, Sobolev space, Cesàro mean

**AMS subject classifications.** 93B05, 35B37, 35L05

**DOI.** 10.1137/070684057

**1. Introduction.** Controllability problems for hyperbolic partial differential equations were investigated in a number of papers (see, e.g., the references in [10]).

One of the most generally accepted ways to study control systems with distributed parameters is their interpretation in the form

$$(1.1) \quad \frac{dW}{dt} = \mathbf{A}W + \mathbf{B}u, \quad t \in (0, T),$$

where  $T > 0$ ,  $W : (0, T) \rightarrow \mathcal{H}$  is an unknown function,  $u : (0, T) \rightarrow H$  is a control,  $\mathcal{H}$ ,  $H$  are Banach spaces,  $A$  is an infinitesimal operator in  $\mathcal{H}$ , and  $B : H \rightarrow \mathcal{H}$  is a linear bounded operator. An important advantage of this approach is the possibility of employing ideas and techniques of semigroup operator theory. At the same time it should be noticed that the most substantial results on operator semigroups that are the most important for applications deal with the case when the semigroup generator  $A$  has a discrete spectrum or a compact resolvent and therefore the semigroup may be treated by means of eigenelements of  $A$ . These assumptions correspond to differential equations in bounded domains only.

In this paper we consider the wave equation on a half-axis. We should note that most of the papers investigating controllability of the wave equation deal with bounded domains and consider  $L^2$ -controllability or, more generally,  $L^p$ -controllability ( $2 \leq p < +\infty$ ); see [1], [4], [5], [6], [7], [11], and many others. But only  $L^\infty$ -controls can be realized practically. Moreover, such controls should be bounded by a hard constant (like in restriction (1.4)) for practical purposes. Furthermore, classical control theory started precisely from this point of view as switching controls are the ones realized in a concrete system. That is why we construct bang-bang controls solving the approximate null-controllability problem in this paper. However, sometimes to control a system

---

\*Received by the editors March 1, 2007; accepted for publication (in revised form) April 1, 2008; published electronically July 23, 2008.

<http://www.siam.org/journals/sicon/47-4/68405.html>

<sup>†</sup>Mathematical Division, Institute for Low Temperature Physics and Engineering, 47 Lenin Ave., Kharkiv 61103, Ukraine (fardigola@ukr.net).



arising in nature and technology we do not necessarily have to stress the system and drive it to a desired state immediately and directly. We often need to control a system by letting it fluctuate while trying to drive it to a desired state without forcing it too much. Therefore we construct continuous controls solving the approximate null-controllability problem and satisfying (1.4).

Consider the wave equation on a half-axis,

$$(1.2) \quad w_{tt} = w_{xx} - q^2 w, \quad x > 0, \quad t \in (0, T),$$

controlled by the Neumann boundary condition

$$(1.3) \quad w_x(0, t) = u(t), \quad t \in (0, T),$$

where  $T > 0$ ,  $q \geq 0$  is a real number. We also assume that the control  $u$  satisfies the restriction

$$(1.4) \quad u \in \mathcal{B}^U(0, T) = \{v \in L^2(0, T) \mid |v(t)| \leq U \text{ a.e. on } (0, T)\},$$

where  $U > 0$  is given.

Controllability problems for the wave equation on a half-axis in the context of controls bounded by a hard constant were investigated in [14], [15]. In these papers the wave equation on a half-axis controlled by the Dirichlet boundary condition was studied for  $q = 0$ . In the present paper most of the results of [14], [15] are extended on the case of the Neumann control for  $q \geq 0$ . If  $q = 0$ , the cases of the Dirichlet control and the Neumann control are rather similar. If  $q > 0$ , investigation of the null-controllability problem for (1.2) is essentially more complicated and requires additional methods. In the present paper the operator  $\Phi$  describing the influence of a control on a target state is introduced and studied in the Sobolev spaces  $H_0^s$ ,  $s \leq 1$ . In fact, application of  $\Phi$  is the most essential new point of the paper. In addition, continuous controls  $u \in \mathcal{B}^U(0, T)$  solving the approximate null-controllability problem are constructed in the present paper. They are not considered in [14], [15].

In section 3 we obtain necessary and sufficient conditions for null-controllability and approximate null-controllability of system (1.2), (1.3) with restrictions (1.4) on the control. Controls solving the problems of null-controllability and approximate null-controllability are found explicitly. If  $q = 0$ , they coincide with the initial states, and if  $q > 0$ , they are given in the form of an integral transform of the initial state such that its kernel is a modified Bessel function. In both cases these controls may be of a rather complicated form.

Construction of bang-bang controls that solve the approximate null-controllability problem is the main goal of the section 4. We show that this problem can be reduced to a system of the Markov power moment problems. They may be solved by the algorithm given in [14]. Further we prove that solutions of the Markov power moment problems give us solutions of the approximate null-controllability problem for  $s < 1/2$ .

Bang-bang controls are the simplest by their structure, but they do not allow us to solve the approximate null-controllability problem for  $s \geq 1/2$ . Moreover, in some cases we should avoid to stress of the control system and drive it to a desired state in a continuous way. That is why we construct continuous controls that solve the approximate null-controllability problem for  $s \leq 1$  and satisfy (1.4) in section 5. They are obtained with the aid of the Cesàro means for a Fourier series determined by the data of the considered system. These controls generate a continuous steering state if an initial state is continuous. The problem of steering a finite string from a continuous

initial state with a continuous control function that is at  $t = 0$  compatible with the initial state to the zero state such that the generated system state is continuous has been considered in [4].

In the appendix (section 6) some auxiliary statements are proved. In particular, the operators describing the influence of a control on a target state of the considered control system are constructed and investigated. The results of this section are applied in sections 3–5.

If we replace  $q$  by  $iq$  in (1.2), i.e., consider the equation

$$(1.2') \quad w_{tt} = w_{xx} + q^2 w, \quad x > 0, \quad t \in (0, T),$$

instead of (1.2), then replacing  $q$  by  $iq$  throughout this paper gives us results analogous to ones obtained in sections 3–6 for the control system (1.2'), (1.3).

**2. Notation.** Let us give definitions of the spaces used in the work. Let  $\mathcal{S}$  be the Schwartz space [12], [13]

$$\mathcal{S} = \left\{ \varphi \in C^\infty(\mathbb{R}) \mid \forall m \in \mathbb{N} \right. \\ \left. \forall l \in \mathbb{N} \sup \left\{ \left| D^m \varphi(x) \right| (1 + |x|^2)^l \mid x \in \mathbb{R} \right\} < +\infty \right\},$$

and let  $\mathcal{S}'$  be the dual space, where  $D = -i\partial/\partial x$  and  $|\cdot|$  is the Euclidean norm.

Denote by  $H_l^s$  ( $s, l \in \mathbb{R}$ ) the following Sobolev spaces:

$$H_l^s = \left\{ \varphi \in \mathcal{S}' \mid (1 + |x|^2)^{l/2} (1 + |D|^2)^{s/2} \varphi \in L^2(\mathbb{R}) \right\}, \\ \|\varphi\|_l^s = \left( \int_{-\infty}^{\infty} \left| (1 + |x|^2)^{l/2} (1 + |D|^2)^{s/2} \varphi(x) \right|^2 dx \right)^{1/2}.$$

Let  $\mathcal{F} : \mathcal{S}' \longrightarrow \mathcal{S}'$  be the Fourier transform operator. For  $\varphi \in \mathcal{S}$  we have

$$(\mathcal{F}\varphi)(\sigma) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{-ix\sigma} \varphi(x) dx$$

and  $(\mathcal{F}f, \psi) = (f, \mathcal{F}^{-1}\psi)$  for  $f \in \mathcal{S}'$ ,  $\psi \in \mathcal{S}$ . It is well known [3, Chapter 1] that  $\mathcal{F}H_0^s = H_s^0$  and  $\|\varphi\|_0^s = \|\mathcal{F}\varphi\|_s^0$  if  $\varphi \in H_0^s$ .

A distribution  $f \in \mathcal{S}'$  is said to be *odd* if  $(f, \varphi(\xi)) = -(f, \varphi(-\xi))$ ,  $\varphi \in \mathcal{S}$ . A distribution  $f \in \mathcal{S}'$  is said to be *even* if  $(f, \varphi(\xi)) = (f, \varphi(-\xi))$ ,  $\varphi \in \mathcal{S}$ .

Let  $\Omega$  be the odd extension operator,  $\Omega : \mathcal{S}' \longrightarrow \mathcal{S}'$ ,  $(\Omega f)(x) = f(x) - f(-x)$ ,  $f \in \mathcal{S}'$ , and let  $\Xi$  be the even extension operator,  $\Xi : \mathcal{S}' \longrightarrow \mathcal{S}'$ ,  $(\Xi f)(x) = f(x) + f(-x)$ ,  $f \in \mathcal{S}'$ .

Further, we use the spaces

$$\mathcal{H}^s = \left\{ \varphi \in \mathcal{S}' \times \mathcal{S}' \mid \text{supp } \varphi \subset [0, +\infty) \text{ and } \Xi\varphi \in H_0^s \times H_0^{s-1} \right\}, \\ \tilde{H}^s = \left\{ \varphi \in H_0^s \times H_0^{s-1} \mid \varphi \text{ is even} \right\}$$

with the norm  $\|\varphi\|^s = ((\|\varphi_0\|_0^s)^2 + (\|\varphi_1\|_0^{s-1})^2)^{1/2}$ ,  $\varphi = (\varphi_0^s)$ , and also the space

$$\hat{H}_s = \left\{ \varphi \in H_s^0 \times H_{s-1}^0 \mid \varphi \text{ is even} \right\}$$

with the norm  $\|\varphi\|_s = ((\|\varphi_0\|_s^0)^2 + (\|\varphi_1\|_{s-1}^0)^2)^{1/2}$ ,  $\varphi = (\varphi_0)$ . We consider control system (1.2), (1.3) in the spaces  $H_0^s$ ,  $s \in \mathbb{R}$ .

Applying the even extension operator to system (1.2), (1.3) we reduce it to the form (1.1), where

$$(2.1) \quad \mathbf{A} = \begin{pmatrix} 0 & 1 \\ \left(\left(\frac{d}{dx}\right)^2 - q^2\right) & 0 \end{pmatrix}, \quad \mathbf{A} : \tilde{H}^{s-2} \longrightarrow \tilde{H}^{s-2}, \quad D(\mathbf{A}) = \tilde{H}^s,$$

$$(2.2) \quad \mathbf{B} = \begin{pmatrix} 0 \\ -2\delta(x) \end{pmatrix}, \quad \mathbf{B} : \mathbb{R} \longrightarrow \tilde{H}^{s-2}, \quad D(\mathbf{B}) = \mathbb{R}.$$

Here  $\delta$  is the Dirac distribution,  $\delta = H'$ ,  $H$  is the Heaviside function  $H(\xi) = 1$  if  $\xi > 0$ , and  $H(\xi) = 0$  otherwise.

**3. Conditions for (approximate) null-controllability.** Consider control system (1.2), (1.3) with the initial conditions

$$(3.1) \quad \begin{cases} w(x, 0) = w_0^0(x), \\ w_t(x, 0) = w_1^0(x), \end{cases} \quad x > 0,$$

where  $w^0 = \begin{pmatrix} w_0^0 \\ w_1^0 \end{pmatrix} \in \mathcal{H}^s$ . We assume throughout this section that  $s \leq 1$ .

Let  $W^0 = \Xi w^0$ ,  $W(\cdot, t) = \Xi \begin{pmatrix} w(\cdot, t) \\ \partial w(\cdot, t) / \partial t \end{pmatrix}$ . Evidently,  $W^0 \in \tilde{H}^s$ ,  $W(\cdot, t) \in \tilde{H}^s$  ( $t \in (0, T)$ ). It is easy to see that control problem (1.2), (1.3), (3.1) is equivalent to the Cauchy problem

$$(3.2) \quad \frac{dW}{dt} = \begin{pmatrix} 0 & 1 \\ \left(\left(\frac{d}{dx}\right)^2 - q^2\right) & 0 \end{pmatrix} W - \begin{pmatrix} 0 \\ 2\delta(x) \end{pmatrix} u, \quad t \in (0, T),$$

$$(3.3) \quad W(x, 0) = W^0,$$

where  $W(\cdot, t) \in \tilde{H}^s$ ,  $t \in [0, T]$ ,  $W^0 \in \tilde{H}^s$ ,  $u \in \mathcal{B}^U(0, T)$ .

Thus control problem (1.2), (1.3), (3.1) can be reduced to the Cauchy problem with initial condition (3.3) for a system of the form (1.1) with the parameter (control)  $u$ .

Consider for (3.2), (3.3) the steering condition

$$(3.4) \quad W(\cdot, T) = W^T,$$

where  $W^T \in \tilde{H}^s$ .

For given  $T > 0$ ,  $w^0 \in \mathcal{H}^s$  denote by  $\mathcal{R}_T^U(w^0)$  the set of the states  $W^T \in \tilde{H}^s$  for which there exists a control  $u \in \mathcal{B}^U(0, T)$  such that problem (3.2)–(3.4) has a unique solution.

**DEFINITION 3.1.** A state  $w^0 \in \mathcal{H}^s$  is called *null-controllable at a given time  $T > 0$  if 0 belongs to  $\mathcal{R}_T^U(w^0)$*  and *approximately null-controllable at a given time  $T > 0$  if 0 belongs to the closure of  $\mathcal{R}_T^U(w^0)$  in  $\tilde{H}^s$* .

Let  $\Phi : \mathcal{S}' \longrightarrow \mathcal{S}'$  with  $D(\Phi) = \{g \in \mathcal{S}' \mid g \text{ is odd and } \text{supp } g \subset [-T, T]\}$  such that

$$(\Phi g)(x) = \mathcal{F}_{\sigma \rightarrow x}^{-1} \left( \frac{-i}{\sqrt{\sigma^2 + q^2}} (\mathcal{F}g) \left( \sqrt{\sigma^2 + q^2} \right) \right) (x), \quad g \in D(\Phi).$$

In the appendix it is proved that  $\Phi$  is invertible,  $\Phi^{-1} : \mathcal{S}' \longrightarrow \mathcal{S}'$ ,  $D(\Phi^{-1}) = R(\Phi) = \{g \in \mathcal{S}' \mid g \text{ is even and } \text{supp } g \subset [-T, T]\}$  (Lemma 6.3). Moreover,  $\Phi$  is bounded from  $H_0^{s-1}$  to  $H_0^s$  (Lemma 6.1) and  $\Phi^{-1}$  is bounded from  $H_0^s$  to  $H_0^{s-1}$  (Lemma 6.3).

PROPOSITION 3.2. *Let  $W^0 \in \mathcal{S}' \times \mathcal{S}'$ ,  $u \in \mathcal{B}^U(0, T)$ . Then*

$$(3.5) \quad W(x, T) = E(x, T) * \left[ W^0(x) - \Phi \begin{pmatrix} \Omega \mathcal{U} \\ \Omega \mathcal{U}' \end{pmatrix} (x) \right], \quad t \in [0, T],$$

where  $\mathcal{U}(t) = u(t)(H(t) - H(t - T))$ ,  $W$  is a unique solution of (3.2)–(3.3),  $*$  is the convolution with respect to  $x$ , and

$$(3.6) \quad E(x, t) = \frac{1}{2} \begin{pmatrix} \partial/\partial t & 1 \\ (\partial/\partial t)^2 & \partial/\partial t \end{pmatrix} \left( J_0 \left( q\sqrt{t^2 - |x|^2} \right) (\text{sgn } t) H(t^2 - x^2) \right).$$

In particular, if  $W^0 \in \tilde{H}^p$ ,  $\Omega \mathcal{U} \in H^{p-1}$ , and  $\Xi \mathcal{U} \in H^{p-1}$ , then  $W(\cdot, T) \in \tilde{H}^p$ ,  $p \in \mathbb{R}$ .

*Proof.* Applying the Fourier transform with respect to  $x$  to problem (3.2)–(3.4) we obtain the following problem in  $\mathcal{S}' \times \mathcal{S}'$ :

$$(3.7) \quad \frac{dV}{dt} = \begin{pmatrix} 0 & 1 \\ -(\sigma^2 + q^2) & 0 \end{pmatrix} V - \sqrt{\frac{2}{\pi}} \begin{pmatrix} 0 \\ 1 \end{pmatrix} u, \quad t \in (0, T),$$

$$(3.8) \quad V(\cdot, 0) = V^0,$$

$$(3.9) \quad V(\cdot, T) = V^T,$$

where  $V(\cdot, t) = \mathcal{F}W(\cdot, t)$ ,  $t \in [0, T]$ ,  $V^0 = \mathcal{F}W^0$ ,  $V^T = \mathcal{F}W^T$ . Then the function

$$(3.10) \quad V(\sigma, t) = \Sigma(\sigma, t) \left( V^0(\sigma) - \sqrt{\frac{2}{\pi}} \int_0^t \Sigma(\sigma, -\tau) \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(\tau) d\tau \right), \quad t \in [0, T],$$

is a unique solution of (3.7)–(3.8) in  $\mathcal{S}' \times \mathcal{S}'$ . Here

$$\Sigma(\sigma, t) \equiv \begin{pmatrix} \cos(\rho t) & \frac{\sin(\rho t)}{\rho} \\ -\rho \sin(\rho t) & \cos(\rho t) \end{pmatrix} \Big|_{\sqrt{\sigma^2 + q^2}} \equiv \begin{pmatrix} \partial/\partial t & 1 \\ (\partial/\partial t)^2 & \partial/\partial t \end{pmatrix} \frac{\sin \left( t\sqrt{\sigma^2 + q^2} \right)}{\sqrt{\sigma^2 + q^2}}.$$

Put  $E(x, t) = \mathcal{F}_{\sigma \rightarrow x}^{-1} \Sigma(\sigma, t) / \sqrt{2\pi}$ . Let us prove that

$$(3.11) \quad \mathcal{F}_{\sigma \rightarrow x}^{-1} \left( \frac{\sin \left( t\sqrt{\sigma^2 + q^2} \right)}{\sqrt{\sigma^2 + q^2}} \right) (x) = \sqrt{\frac{\pi}{2}} J_0 \left( q\sqrt{t^2 - x^2} \right) H(t^2 - x^2) \text{sgn } t,$$

where  $J_\nu(\xi) = \sum_{m=0}^{\infty} \frac{(-1)^m}{\Gamma(m+1)\Gamma(m+\nu+1)} \left(\frac{\xi}{2}\right)^{2m+\nu}$  is the Bessel function ( $\nu \in \mathbb{R}$ ) and  $\Gamma$  is the Euler gamma function. It is well known that

$$(3.12) \quad \left( F_{(\sigma, q) \rightarrow (x, y)}^{-1} \frac{\sin \left( t\sqrt{\sigma^2 + q^2} \right)}{\sqrt{\sigma^2 + q^2}} \right) (x) = \frac{\text{sgn } t H(t^2 - x^2 - y^2)}{\sqrt{t^2 - x^2 - y^2}}.$$

Since  $\mathcal{F}_{y \rightarrow q}^{-1} \frac{H(a^2 - y^2)}{\sqrt{a^2 - y^2}} = \sqrt{\frac{\pi}{2}} J_0(qa)$  then (3.11) follows from (3.12). Hence  $E$  is of the form (3.6). It follows from (3.10) that

$$(3.13) \quad W(x, T) = E(x, T) * \left[ W^0(x) - \sqrt{\frac{2}{\pi}} \mathcal{F}^{-1} \int_0^\infty \left( -\frac{\sin(t\sqrt{\sigma^2 + q^2})}{\sqrt{\sigma^2 + q^2}} \right) \mathcal{U}(t) dt \right].$$

Therefore (3.5) holds. With regard to Lemmas 6.1 and 6.7 we conclude that if  $W^0 \in \tilde{H}^p$ ,  $\Omega \mathcal{U} \in H^{p-1}$ , and  $\Xi \mathcal{U} \in H^{p-1}$ , then  $W(\cdot, T) \in \tilde{H}^p$ ,  $p \in \mathbb{R}$ . The proposition is proved.

Thus we have

$$(3.14) \quad \mathcal{R}_T^U(w^0) = \left\{ E(x, T) * \left[ \Xi w^0(x) - \Phi \left( \frac{\Omega \mathcal{U}}{\Omega \mathcal{U}'} \right) (x) \right] \mid u \in \mathcal{B}^U(0, T) \right\}.$$

The following theorem gives us sufficient conditions for (approximate) null-controllability.

**THEOREM 3.3.** *For a state  $w^0 \in \mathcal{H}^s$  and a given time  $T > 0$  assume that*

$$(3.15) \quad \text{supp } w_0^0 \subset [0, T];$$

$$(3.16) \quad \Xi w_0^0 \in H_0^1 \text{ and } |(\Phi^{-1} \Xi w_0^0)(\xi)| \leq U \text{ a.e. on } [0, T];$$

$$(3.17) \quad \Xi w_1^0 = \Phi \left( \frac{d}{d\xi} (\text{sgn } \xi (\Phi^{-1} \Xi w_0^0)) \right).$$

*Then the state  $w^0$  is null-controllable at the time  $T$ . Moreover, the solution of the null-controllability problem (the control  $u$ ) is unique and*

$$\begin{aligned} u(t) &= -\frac{d}{dt} \int_t^T I_0(q\sqrt{x^2 - t^2}) w_0^{0'}(x) dx \\ &= w_0^{0'}(t) + \int_t^T \frac{qt I_1(q\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} w_0^{0'}(x) dx \\ (3.18) \quad &= w_1^0(t) + \int_t^T \frac{qx I_1(q\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} w_1^0(x) dx \quad \text{a.e. on } (0, T). \end{aligned}$$

Here  $I_\nu(\xi) = i^{-\nu} J_\nu(i\xi)$  is the modified Bessel function. Under conditions (3.15)–(3.17)  $T_* = \max \text{supp } w_0^0$  is the optimal time and  $u$  of the form (3.18) (for  $T = T_*$ ) is the time-optimal control for the null-controllability problem.

*Proof.* Put

$$(3.19) \quad \widehat{\mathcal{U}} = \Phi^{-1} \Xi w_0^0.$$

It follows from Lemma 6.3 and (3.16) that  $\widehat{\mathcal{U}}$  is odd and  $\widehat{\mathcal{U}} \in L^\infty(\mathbb{R})$ . Put  $\mathcal{U}(t) = H(t)\widehat{\mathcal{U}}(t)$  and denote by  $u(t)$  its restriction on  $[0, T]$ . Then  $\widehat{\mathcal{U}}(t) = (\Omega \mathcal{U})(t)$ ,  $\mathcal{U}(t) = u(t)(H(t) - H(t - T))$ . Using again Lemma 6.4 we conclude that  $u \in \mathcal{B}^U(0, T)$  and the first part of assertion (3.18) is true for it. Taking into account (3.17), (3.19) we

get

$$\begin{aligned} W_1^0 &= \Xi w_1^0 = \Phi \left( \frac{d}{dt} (\operatorname{sgn} t (\Phi^{-1} \Xi w_0^0)) \right) \\ (3.20) \quad &= \Phi \left( \frac{d}{dt} (\operatorname{sgn} t (\Phi^{-1} \Phi \Omega \mathcal{U})) \right) = \Phi \left( \frac{d}{dt} (\operatorname{sgn} t \Omega \mathcal{U}) \right) = \Phi \Omega \mathcal{U}'. \end{aligned}$$

With regard to (3.19), (3.20) we have from (3.5) that  $W(\cdot, T) = 0$ . Here  $W$  is the solution of (3.2), (3.3). Thus  $w_0^0$  is null-controllable.

It follows from (3.20) and Lemma 6.5 that  $(\Xi \mathcal{U})' = \widetilde{W}_1^{0'}$ , where  $\widetilde{W}_1^0 \in H_0^0$  is even,  $\operatorname{supp} \widetilde{W}_1^0 \subset [-T, T]$ , and  $\widetilde{W}_1^0(t) = W_1^0(t) + \int_{|t|}^T \frac{qxI_1(q\sqrt{x^2-t^2})}{\sqrt{x^2-t^2}} W_1^0(x) dx$ . Since  $\Xi \mathcal{U}$  and  $\widetilde{W}_1^0$  have compact supports, then  $\Xi \mathcal{U} = \widetilde{W}_1^0$ . Therefore the second part of (3.18) is true.

Put  $T_* = \max \operatorname{supp} w_0^0$ . With regard to (3.5) we conclude that  $T = T_*$  is the optimal time and  $u$  of the form (3.18) (for  $T = T_*$ ) is the time-optimal control for null-controllability problem. The theorem is proved.

The following theorem asserts that conditions (3.15)–(3.17) are not only sufficient but also necessary for (approximate) null-controllability.

**THEOREM 3.4.** *If a state  $w^0 \in \mathcal{H}^s$  is approximately null-controllable at a given time  $T > 0$ , then assertions (3.15)–(3.17) are valid.*

*Proof.* For each  $n \in \mathbb{N}$  there exists a state  $W^n \in \mathcal{R}_T^U(w^0)$  such that  $\|W^n\|^s < 1/n$ . With regard to (3.14) we have

$$E(x, -T) * W^n(x) = W^0(x) - \Phi \begin{pmatrix} \Omega \mathcal{U}_n \\ \Omega \mathcal{U}_n' \end{pmatrix} (x)$$

for some  $u_n \in \mathcal{B}^U(0, T)$ , where  $\mathcal{U}_n(t) = u_n(t)(H(t) - H(t - T))$ . Using Lemma 6.7 we get

$$(3.21) \quad \Phi \begin{pmatrix} \Omega \mathcal{U}_n \\ \Omega \mathcal{U}_n' \end{pmatrix} \longrightarrow W^0 \quad \text{as } n \longrightarrow \infty \text{ in } \widetilde{H}^s.$$

According to Lemma 6.1  $\operatorname{supp} W_0^0 \subset [0, T]$ . Applying Lemma 6.3 we have

$$(3.22) \quad \begin{pmatrix} \Omega \mathcal{U}_n \\ \Omega \mathcal{U}_n' \end{pmatrix} \longrightarrow \Phi^{-1} W^0 \quad \text{as } n \longrightarrow \infty \text{ in } H_{s-1}^0 \times H_{s-2}^0.$$

Put  $\widehat{\mathcal{U}} = \Phi^{-1} W_0^0$ . With regard to Lemma 6.3 we get  $\operatorname{supp} \widehat{\mathcal{U}} \subset [-T, T]$ ,  $\widehat{\mathcal{U}}$  is odd. Let us prove that

$$(3.23) \quad |\widehat{\mathcal{U}}(t)| \leq U \quad \text{a.e. on } [-T, T].$$

Since  $\Omega \mathcal{U}_n \longrightarrow \widehat{\mathcal{U}}$  as  $n \longrightarrow \infty$  in  $H_0^{s-1}$  then the sequence  $\{\Omega \mathcal{U}_n\}_{n=1}^\infty$  converges to  $\widehat{\mathcal{U}}$  as  $n \longrightarrow \infty$  in  $\mathcal{S}'$  and consequently in  $(L^2(\mathbb{R}))'$  because  $\{\Omega \mathcal{U}_n\}_{n=1}^\infty$  is uniformly bounded on  $\mathbb{R}$  and  $\mathcal{S}$  is dense in  $L^2(\mathbb{R})$ . By the Riesz theorem  $\widehat{\mathcal{U}} \in L^2(\mathbb{R})$ . Since  $u_n \in \mathcal{B}^U(0, T)$  then (3.23) holds. With regard to Lemma 6.1 we obtain from here that (3.15), (3.16) are true. Setting  $\mathcal{U}(t) = H(t)\widehat{\mathcal{U}}(t)$  and taking into account (3.22) we obtain

$$\begin{pmatrix} \Omega \mathcal{U}_n \\ \Omega \mathcal{U}_n' \end{pmatrix} \longrightarrow \begin{pmatrix} \Omega \mathcal{U} \\ \Omega \mathcal{U}' \end{pmatrix} = \Phi^{-1} W^0 \quad \text{as } n \longrightarrow \infty \text{ in } H_{s-1}^0 \times H_{s-2}^0.$$

Hence (3.17) holds. The theorem is proved.

*Remark 3.1.* According to Lemma 6.5 condition (3.16),

$$\Xi w_0^0 \in H_0^1 \quad \text{and} \quad \left| w_0^{0'}(t) + \int_t^T \frac{qt I_1(q\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} w_0^{0'}(x) dx \right| \leq U \quad \text{a.e. on } [0, T],$$

is equivalent to

$$(3.16') \quad w_1^0 \in H_0^0 \quad \text{and} \quad \left| w_1^0(t) + \int_t^T \frac{qx I_1(q\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} w_1^0(x) dx \right| \leq U \quad \text{a.e. on } [0, T]$$

if (3.15) and (3.17) hold.

*Remark 3.2.* It follows from the proof of Theorem 3.3 that condition (3.17) is equivalent to

$$(3.17') \quad w_0^{0'}(x) - w_1^0(x) + \int_x^\infty \frac{q I_1(q\sqrt{\xi^2 - x^2})}{\sqrt{\xi^2 - x^2}} (x w_0^{0'}(\xi) - \xi w_1^0(\xi)) d\xi = 0, \quad x > 0.$$

And according to Lemma 6.6 condition (3.17) is equivalent to

$$(3.17'') \quad w_1^0(x) = w_0^{0'}(x) + q \int_x^\infty w_0^0(\xi) \frac{I_1(q(\xi - x))}{\xi - x} d\xi, \quad x > 0.$$

Other conditions equivalent to (3.17) are given below in Remarks 4.1 and 5.2.

*Remark 3.3.* Let  $q = 0$ . Then (3.17) is of the form  $w_1^0 = w_0^{0'}$  and (3.18) is of the form  $u(t) = w_0^{0'}(t) = w_1^0(t)$  a.e. on  $[0, T]$ .

*Remark 3.4.* With regard to Lemma 6.5

$$(3.24) \quad w_1^0 \in H_0^0 \quad \text{and} \quad |w_1^0(x)| \leq \frac{U}{I_0(qT)} \quad \text{a.e. on } [0, T]$$

is sufficient for (3.16') (and (3.16)) and

$$(3.25) \quad w_1^0 \in H_0^0 \quad \text{and} \quad |w_1^0(x)| \leq U(1 + qT) \quad \text{a.e. on } [0, T]$$

is necessary for (3.16') (and (3.16)). If  $q = 0$ , we obtain from here that (3.16') and (3.16) are equivalent to

$$(3.26) \quad w_1^0 \in H_0^0 \quad \text{and} \quad |w_1^0(x)| \leq U \quad \text{a.e. on } [0, T].$$

It follows from Example 3.1 (see below) that (3.24) is not necessary and (3.25) is not sufficient for (3.16') (and (3.16)) in the case  $q > 0$  for each  $T > 0$ .

*Example 3.1.* Let  $q > 0$ ,  $T > 0$ ,  $\alpha > 0$ . Put  $f(x) = \alpha(H(x + T) - H(x - T))$ ,  $h(x) = f(x) - \int_{|x|}^\infty \frac{qt J_1(\sqrt{t^2 - x^2})}{\sqrt{t^2 - x^2}} f(t) dt = \alpha J_0(\sqrt{T^2 - x^2})(H(x + T) - H(x - T))$ . With regard to Lemma 6.5 we obtain that  $f(t) = h(t) + \int_{|t|}^\infty \frac{qx I_1(\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} h(x) dx$ . Evidently,  $\sup\{|f(x)| \mid x \in (-T, T)\} = \sup\{|h(t)| \mid t \in (-T, T)\} = \alpha$ . Setting  $\alpha = U$  we conclude that (3.24) is not necessary for (3.16') (and (3.16)) for each  $T > 0$ . Setting  $\alpha = (1 + qT)U$  we conclude that (3.25) is not sufficient for (3.16') (and (3.16)) for each  $T > 0$ .

**4. Bang-bang controls.** The solution found in section 3 for the null-controllability problem (i.e., the control) may be too complicated for practical purposes. In this section we find bang-bang controls solving the approximate null-controllability problem. We consider a system of the Markov power moment problems and show that their bang-bang solutions are solutions of the approximate null-controllability problem if  $s < 1/2$ .

Consider control system (3.2), (3.3) and assume that for  $T > 0$  and  $w^0 \in \mathcal{H}^s$ ,  $s \leq 1$ , conditions (3.15)–(3.17) hold. According to Theorem 3.3 there exists  $\tilde{u} \in \mathcal{B}^U(0, T)$  such that  $W^0 = \Xi w^0 = \Phi\left(\frac{\Omega \tilde{\mathcal{U}}}{\Omega \mathcal{U}'}\right)$ , where  $\tilde{\mathcal{U}}(t) = \tilde{u}(t)[H(t) - H(t - T)]$ . With regard to (3.5) we obtain

$$(4.1) \quad W(x, T) = E(x, T) * \Phi \begin{pmatrix} \Omega(\tilde{\mathcal{U}} - \mathcal{U}) \\ \Omega(\tilde{\mathcal{U}}' - \mathcal{U}') \end{pmatrix}$$

for a given  $u \in \mathcal{B}^U(0, T)$ , where  $\mathcal{U}(t) = u(t)[H(t) - H(t - T)]$  and  $W$  is the solution of (3.2), (3.3).

Let  $n \in \mathbb{N}$ . Taking into account (3.18) we conclude that

$$(4.2) \quad \begin{aligned} \int_0^T t^n \tilde{u}(t) dt &= - \int_0^\infty t^n \frac{d}{dt} \int_t^T I_0(q\sqrt{x^2 - t^2}) w_0^{0'}(x) dx dt \\ &= -n \int_0^T t^{n-1} \left( w_0^0(t) + \int_t^\infty \frac{qx I_1(q\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} w_0^0(x) dx \right) dt. \end{aligned}$$

For  $f \in \mathcal{H}^s$  and  $m \in \mathbb{N}$  we have

$$\begin{aligned} \int_0^\infty t^m \left( f(t) + \int_t^\infty \frac{qx I_1(q\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} f(x) dx \right) dt \\ = \int_0^\infty x^m f(x) dx + \int_0^\infty x f(x) \int_0^x \frac{qt^m I_1(q\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} dt dx. \end{aligned}$$

Since

$$\int_0^x t^m (x^2 - t^2)^k dt = \frac{x^{2k+m+1}}{2} \int_0^1 \xi^{(m-1)/2} (1-\xi)^k d\xi = \frac{x^{2k+m+1}}{2} \frac{\Gamma(\frac{m+1}{2}) \Gamma(k+1)}{\Gamma(k + \frac{m+1}{2} + 1)}$$

then

$$\begin{aligned} \int_0^\infty \frac{qt^m I_1(q\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} dt &= \Gamma\left(\frac{m+1}{2}\right) \sum_{k=0}^\infty \frac{x^{2k+m+1}}{(k+1)! \Gamma(k + \frac{m+1}{2} + 1)} \left(\frac{q}{2}\right)^{2k+2} \\ &= -x^{m-1} + \Gamma\left(\frac{m+1}{2}\right) \left(\frac{2}{q}\right)^{(m-1)/2} x^{(m-1)/2} I_{(m-1)/2}(qx). \end{aligned}$$

Hence

$$(4.3) \quad \begin{aligned} \int_0^\infty t^m \left( f(t) + \int_t^\infty \frac{qx I_1(q\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} f(x) dx \right) dt \\ = \Gamma\left(\frac{m+1}{2}\right) \left(\frac{2}{q}\right)^{(m-1)/2} \int_0^\infty x^{(m+1)/2} I_{(m-1)/2}(qx) f(x) dx \end{aligned}$$



for  $f \in \mathcal{H}^s$  and  $m \in \mathbb{N}$ . For  $n = \overline{0, \infty}$  and  $j = 0, 1$  put

$$(4.4) \quad \omega_j^n = -n\Gamma\left(\frac{n}{2}\right)\left(\frac{2}{q}\right)^{(n-2)/2} \int_0^\infty x^{n/2} I_{(n-2)/2}(qx) w_j^0(x) dx.$$

In particular, if  $q = 0$ , then  $\omega_j^n = -n \int_0^\infty x^{n-1} w_j^0(x) dx$ ,  $n = \overline{0, \infty}$ . Then

$$(4.5) \quad \int_0^T t^n \tilde{u}(t) dt = \omega_0^n, \quad n = \overline{0, \infty}.$$

Taking into account (3.17), (3.18), and (4.3) we get

$$\int_0^T t^n \tilde{u}(t) dt = \int_0^T t^n \left( w_1^0(x) + \int_t^\infty \frac{qx I_1(q\sqrt{x^2 - t^2})}{\sqrt{x^2 - t^2}} w_1^0(x) dx \right) dt = -\frac{\omega_1^{n+1}}{n+1}$$

for  $n = \overline{0, \infty}$ . Summarizing we get the following.

*Remark 4.1.* Condition (3.17) is equivalent to  $\omega_1^n + n\omega_0^{n-1} = 0$ ,  $n = \overline{1, \infty}$ .

**DEFINITION 4.1.** *The problem of determination of a function  $u \in \mathcal{B}^U(0, T)$  such that*

$$(4.6) \quad \int_0^T t^n u(t) dt = \omega_0^n, \quad n = \overline{0, \infty},$$

*for a given  $\{\omega_0^n\}_{n=0}^\infty$  and  $T > 0$  is called the Markov power moment problem on  $(0, T)$  for the infinite sequence  $\{\omega_0^n\}_{n=0}^\infty$ .*

With regard to (4.1) we obtain  $W(x, T) \equiv 0 \Leftrightarrow \Omega \tilde{\mathcal{U}}(t) \equiv \Omega \mathcal{U}(t)$ . Due to the Paley–Wiener theorem we get that  $\tilde{\nu} = \mathcal{F}\Omega \tilde{\mathcal{U}}$  and  $\nu = \mathcal{F}\Omega \mathcal{U}$  are entire functions. Therefore  $W(x, T) \equiv 0 \Leftrightarrow [\tilde{\nu}^{(m)}(0) = \nu^{(m)}(0)]$ ,  $m = \overline{0, \infty}$ . Taking into account (4.5) and

$$\tilde{\nu}^{(m)}(0) = (-i)^m \sqrt{\frac{2}{\pi}} \int_0^T t^m \tilde{u}(t) dt, \quad \nu^{(m)}(0) = (-i)^m \sqrt{\frac{2}{\pi}} \int_0^T t^m u(t) dt, \quad m = \overline{0, \infty},$$

we conclude that  $u \in \mathcal{B}^U(0, T)$  is a solution of the Markov power moment problem (4.6) for  $\{\omega_0^n\}_{n=0}^\infty$  given by (4.4) iff it is a solution of the null-controllability problem for (3.2), (3.3).

With regard to Theorem 3.3 this gives us the following.

**THEOREM 4.2.** *Assume that  $T > 0$  and for a state  $w^0 \in \mathcal{H}^s$  conditions (3.15)–(3.17) hold. Assume also that  $\{\omega_0^n\}_{n=0}^\infty$  is defined by (4.4). Then the Markov power moment problem (4.6) has a unique solution on  $(0, T)$  for  $\{\omega_0^n\}_{n=0}^\infty$ . Moreover, this solution is the solution of the null-controllability problem for  $w^0$  at the time  $T$  and is of the form (3.18).*

Consider (4.6) for a finite set of  $n$ :

$$(4.7) \quad \int_0^T t^n u(t) dt = \omega_0^n, \quad n = \overline{0, N},$$

where  $N \in \mathbb{N}$ .

**DEFINITION 4.3.** *The problem of determination of a function  $u \in \mathcal{B}^U(0, T)$  satisfying condition (4.7) for a given  $\{\omega_0^n\}_{n=0}^N$  and  $T > 0$  is called the Markov power moment problem on  $(0, T)$  for the finite sequence  $\{\omega_0^n\}_{n=0}^N$ .*

Obviously,  $u$  of the form (3.18) is a solution of this problem for  $\{\omega_0^n\}_{n=0}^\infty$  given by (4.4), but it is not unique.

Let us show that solutions of moment problem (4.7) for various  $N$  give us controls solving the approximate null-controllability problem.

**THEOREM 4.4.** *Let  $T > 0$ ,  $w^0 \in \mathcal{H}^s$ ,  $s < 1/2$ . Also, let conditions (3.15)–(3.17) be fulfilled and  $\{\omega_0^n\}_{n=0}^\infty$  be defined by (4.4). Then for all  $\varepsilon > 0$  there exists  $N > 0$  such that for each solution  $u \in \mathcal{B}^U(0, T)$  of moment problem (4.7) the corresponding solution  $W$  of control system (3.2), (3.3) satisfies  $\|W(\cdot, T)\|^s < \varepsilon$ .*

*Proof.* Taking into account (4.5) we have

$$\begin{aligned} \left( \mathcal{F}(\tilde{u} - u) \right)(\sigma) &= \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{(-i\sigma)^n}{n!} \int_0^T t^n (\tilde{u}(t) - u(t)) dt \\ &= \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{(-i\sigma)^n}{n!} \left( \omega_0^n - \int_0^T t^n u(t) dt \right). \end{aligned}$$

Let  $N \in \mathbb{N}$  be fixed and also let  $u \in \mathcal{B}^U(0, T)$  be a solution of (4.7). Then

$$(4.8) \quad \left( \mathcal{F}(\tilde{u} - u) \right)(\sigma) = \frac{1}{\sqrt{2\pi}} \sum_{n=N+1}^{\infty} \frac{(-i\sigma)^n}{n!} \left( \omega_0^n - \int_0^T t^n u(t) dt \right).$$

Since

$$\begin{aligned} \left| \left( \frac{\partial}{\partial \sigma} \right)^m \left( \mathcal{F}(\tilde{u} - u) \right)(\sigma) \right| &= \frac{1}{\sqrt{2\pi}} \left| \int_0^T (ix)^m e^{-i\sigma x} (\tilde{u}(t) - u(t)) dt \right| \\ &\leq \frac{2UT^{m+1}}{\sqrt{2\pi}(m+1)!}, \quad \sigma \in \mathbb{R}, \quad m = \overline{0, \infty}, \end{aligned}$$

then

$$(4.9) \quad \left| \left( \mathcal{F}(\tilde{u} - u) \right)(\sigma) \right| \leq \frac{2UT^{N+2}a^{N+1}}{\sqrt{2\pi}(N+2)!}, \quad |\sigma| \leq a,$$

$$(4.10) \quad \left| \left( \mathcal{F}(\tilde{u} - u) \right)(\sigma) \right| \leq \frac{2UT}{\sqrt{2\pi}}, \quad \sigma \in \mathbb{R},$$

where  $a > 0$ . Therefore

$$\begin{aligned} \left( \left\| \mathcal{F}(\tilde{u} - u) \right\|_{s-1}^0 \right)^2 &= \int_{-\infty}^{\infty} (1 + \sigma^2)^{s-1} \left| \left( \mathcal{F}(\tilde{u} - u) \right)(\sigma) \right|^2 d\sigma \\ &\leq \frac{4U^2(Ta)^{2(N+2)}}{\pi a((N+2)!)^2} + \frac{4(UT)^2}{\pi} \int_a^{\infty} (1 + \sigma^2)^{s-1} d\sigma \\ &\leq \frac{4U^2(Ta)^{2(N+2)}}{\pi a((N+2)!)^2} - \frac{4(UT)^2}{\pi} \frac{a^{2s-1}}{2s-1}. \end{aligned}$$

Applying the Stirling formula we have

$$\frac{(Ta)^{N+2}}{(N+2)!} \leq \left( \frac{Tae}{N+2} \right)^{N+2} \frac{1}{\sqrt{2\pi(N+2)}}.$$

Setting  $a = (N + 2)/(2Te)$  we get

$$\left(\|\tilde{u} - u\|_0^{s-1}\right)^2 = \left(\|\mathcal{F}(\tilde{u} - u)\|_{s-1}^0\right)^2 \leq \frac{U^2Te}{\pi^2(N+2)^22^{2N+2}} - \frac{(N+2)^{2s-1}}{\pi(2s-1)(2Te)^{2s-1}}.$$

Applying Proposition 3.2 and Lemmas 6.7 and 6.1 we conclude that

$$(4.11) \quad |||W(\cdot, T)|||^p \leq M_q^T L_q^p \left\| \begin{pmatrix} \Omega(\tilde{u} - u) \\ \Omega(\tilde{u}' - u') \end{pmatrix} \right\|^{p-1} \leq 2M_q^T L_q^p \|\tilde{u} - u\|_0^{p-1},$$

since  $\|f'\|_0^{p-1} \leq \|f\|_0^p$ ,  $f \in H_0^p$ ,  $p \leq 1$ . Therefore

$$(4.12) \quad |||W(\cdot, T)|||^s \leq 2M_q^T L_q^s \left(\|\tilde{u} - u\|_0^{s-1}\right)^2 \leq 2M_q^T L_q^s \left(\frac{U^2Te}{\pi^2(N+2)^22^{2N+2}} - \frac{(N+2)^{2s-1}}{\pi(2s-1)(2Te)^{2s-1}}\right)^{1/2} \longrightarrow 0 \quad \text{as } N \longrightarrow \infty.$$

The theorem is proved.

Denote

$$\begin{aligned} \mathcal{B}_N^U(0, T) = \{u \in \mathcal{B}^U(0, T) \mid \exists T_* \in (0, T) (|u(t)| = U \text{ a.e. on } (0, T_*)) \\ \wedge (u(t) = 0 \text{ a.e. on } (T_*, T)) \\ \wedge (u \text{ has no more than } N \text{ discontinuities on } (0, T_*))\}. \end{aligned}$$

It is well known [8, 9] that if the Markov power moment problem (4.7) is solvable, then there exists its solution  $u \in \mathcal{B}_N^U(0, T)$ . Taking into account Theorem 4.4 we conclude that under the conditions of this theorem we can find a solution  $u_N \in \mathcal{B}_N^U(0, T)$  of the Markov power moment problem (4.7) for  $N \in \mathbb{N}$ , and such solutions  $\{u_N\}_{N=1}^\infty$  give us bang-bang controls solving the approximate null-controllability problem (see also (4.12)).

Thus the following theorem is true.

**THEOREM 4.5.** *Let  $T > 0$ ,  $w^0 \in \mathcal{H}^s$ ,  $s < 1/2$ . Also, let conditions (3.15)–(3.17) be fulfilled and  $\{\omega_0^n\}_{n=0}^\infty$  be defined by (4.4). Then for all  $N \in \mathbb{N}$  there exists a solution  $u_N \in \mathcal{B}_N^U(0, T)$  of moment problem (4.7). Moreover, for this  $u_N$  the corresponding solution  $W^N$  of control system (3.2), (3.3) satisfies the estimate*

$$\begin{aligned} & |||W^N(\cdot, T)|||^s \\ & \leq 2M_q^T L_q^s \left(\frac{U^2Te}{\pi^2(N+2)^22^{2N+2}} - \frac{(N+2)^{2s-1}}{\pi(2s-1)(2Te)^{2s-1}}\right)^{1/2} \longrightarrow 0 \quad \text{as } N \longrightarrow \infty. \end{aligned}$$

Bang-bang controls solving the approximate null-controllability problem; i.e., solutions (4.7) ( $N \in \mathbb{N}$ ) can be found by the algorithm given in [14].

Let us show that the condition  $s < 1/2$  of Theorems 4.4 and 4.5 is essential. Precisely, if  $s = 1$ , then  $\exists w^0 \in \mathcal{H}^s \forall u \in \cup_{N \in \mathbb{N}} \mathcal{B}_N^U(0, T) \exists \varepsilon_0 > 0$  such that for the solution  $W$  of (3.2), (3.3) corresponding to the control  $u$  we have  $|||W(\cdot, T)|||^s \geq \varepsilon_0$ . Thus the state  $w^0$  is not approximately null-controllable at the time  $T$  by bang-bang controls if  $s = 1$ .

*Example 4.1.* Let  $s = 1$ ,  $T > 0$ . Also, let  $\tilde{\mathcal{U}}(t) = \frac{U}{2} [H(t) - H(t - T)]$ ,  $w^0 = H(x)(\Phi(\frac{\Omega \tilde{\mathcal{U}}}{\Omega \tilde{\mathcal{U}}'}))(x)$ . Evidently,  $w^0 \in \mathcal{H}^s$  and (3.15)–(3.17) hold for it. Let  $N \in \mathbb{N}$ ,  $u \in \mathcal{B}_N^U(0, T)$ . Hence  $u(t) = \alpha \sum_{k=0}^N (-1)^k [H(t - t_k) - H(t - t_{k+1})]$ , where  $\alpha = \pm U$ ,  $0 = t_0 < t_1 < t_2 < \dots < t_{N+1} = T_* \leq T$ ,  $\mathcal{U}(t) = (H(t) - H(t - T))$ . Let  $W$  be a solution of (3.2), (3.3) corresponding to the control  $u$ . Taking into account (4.1) and Lemmas 6.7 and 6.3 we get

$$\| \| W(\cdot, T) \| \| \geq \frac{1}{M_q^T K_q^0} \left\| \left( \begin{array}{c} \Omega(\tilde{\mathcal{U}} - \mathcal{U}) \\ \Omega(\tilde{\mathcal{U}} - \mathcal{U})' \end{array} \right) \right\| \geq \frac{1}{M_q^T K_q^0} \left\| \Omega(\tilde{\mathcal{U}} - \mathcal{U}) \right\|_0^0.$$

Since  $\| \Omega(\tilde{\mathcal{U}} - \mathcal{U}) \|_0^0 \geq (\int_{-T}^T |\Omega(\tilde{\mathcal{U}} - \mathcal{U})(t)|^2 dt) \geq U \sqrt{\frac{T}{2}}$  then  $\| \| W(\cdot, T) \| \| \geq \frac{U \sqrt{T}}{\sqrt{2} M_q^T K_q^0} = \varepsilon_0$ . That was to be proved.

**5. Continuous controls.** Note again that the solution of the null-controllability problem (i.e., the control) found in section 3 may be too complicated for practical purposes. That is why bang-bang controls solving the approximate null-controllability problem have been constructed in section 4. They also solve the Markov power moment problem. Bang-bang controls are the simplest by their structure. But, first, they do not allow us to solve approximate null-controllability problem for  $s \geq 1/2$ , and second, the algorithm for solving the Markov power moment problems (and constructing bang-bang controls) is rather sensitive to computational errors. Moreover, in some cases we need to avoid straining the control system and drive it to a desired state in a continuous way. That is why we construct continuous controls that solve the approximate null-controllability problem for  $s \leq 1$  and satisfy restriction (1.4). They are obtained with the help of the Cesàro means for a Fourier series determined by the data of the considered system. These controls generate a continuous steering state if an initial state is continuous. The problem of steering a finite string from a continuous initial state with a continuous control function that is at  $t = 0$  compatible with the initial state to the zero state such that the generated system state is continuous was considered in [4].

Consider control system (3.2), (3.3) and assume that for  $s \leq 1$ ,  $T > 0$ , and  $w^0 \in \mathcal{H}^s$  conditions (3.15)–(3.17) hold. Set

$$(5.1) \quad \tilde{\mathcal{U}}(t) = H(t) (\Phi^{-1} \Xi w_0^0)(t)$$

and denote by  $\tilde{u}$  the restriction of  $\tilde{\mathcal{U}}$  on  $[0, T]$ . By Theorem 3.3  $\tilde{u} \in \mathcal{B}^U(0, T)$ . Put

$$\nu_n = \frac{2}{T} \int_0^T \tilde{u}(t) \sin \frac{\pi n t}{T} dt, \quad n = \overline{0, \infty}.$$

Taking into account (5.1), (6.4), and Lemma 6.3 we get

$$(5.2) \quad \nu_n = \frac{\sqrt{2\pi i}}{T} (\mathcal{F} \Omega \tilde{\mathcal{U}}) \left( \frac{\pi n}{T} \right) = \frac{2\pi n}{T^2} \int_0^T \cos \left( x \sqrt{(\pi n/T)^2 - q^2} \right) w_0^0(x) dx.$$

Put

$$(5.3) \quad u_n(t) = \frac{1}{n+1} \sum_{k=0}^n \sum_{l=0}^k \nu_l \sin \frac{\pi l t}{T} = \sum_{l=0}^n \frac{n+1-l}{n+1} \nu_l \sin \frac{\pi l t}{T}, \quad n = \overline{0, \infty}.$$

Evidently,  $u_n$  is a Cesàro mean for the Fourier series  $\sum_{l=0}^{\infty} \nu_l \sin \frac{\pi l t}{T}$ ,  $n = \overline{0, \infty}$ . Set  $\nu(t) = \sum_{k=-\infty}^{\infty} (\Omega \tilde{\mathcal{U}})(t - 2Tk)$ . It is well known [16, Chapter 3, section 3] that  $u_n(t) = \frac{1}{T} \int_{-T}^T \nu(t + \xi) F_n(\xi) d\xi$ , where  $F_n(\xi) = \frac{1}{2(n+1)} \left( \frac{\sin((n+1)\pi\xi/T)}{\sin(\pi\xi/T)} \right)^2$  is the Fejér kernel and  $\frac{1}{T} \int_{-T}^T F_n(\xi) d\xi = 1$ . Hence  $|u_n(t)| \leq \left| \frac{1}{T} \int_{-T}^T \nu(t + \xi) F_n(\xi) d\xi \right| \leq \frac{U}{T} \int_{-T}^T F_n(\xi) d\xi = U$  since  $F_n(\xi) \geq 0$ ,  $\xi \in \mathbb{R}$ . Therefore

$$(5.4) \quad u_n \in \mathcal{B}^U(0, T).$$

For  $\mathcal{U}_l^0(t) = \sin \frac{\pi l t}{T} (H(t) - H(t - T))$ ,  $l = \overline{1, \infty}$ , we have

$$(\mathcal{F}\mathcal{U}_l^0)(\sigma) = \frac{i}{\sqrt{2\pi}} \left( \frac{\sin[T(\sigma + \frac{\pi l}{T})]}{\sigma + \frac{\pi l}{T}} - \frac{\sin[T(\sigma - \frac{\pi l}{T})]}{\sigma - \frac{\pi l}{T}} \right) = \sqrt{\frac{2}{\pi}} \frac{(-1)^{l+1} \frac{\pi l}{T} i \sin(\sigma T)}{\sigma^2 - (\frac{\pi l}{T})^2}.$$

Hence  $\mathcal{F}\mathcal{U}_l^0 \in H_p^0$ ,  $p < 3/2$ ,  $l = \overline{1, \infty}$ . Therefore

$$(5.5) \quad \mathcal{U}_n \in H_0^p, \quad p < 3/2, \quad n = \overline{0, \infty},$$

where  $\mathcal{U}_n(t) = u_n(t) (H(t) - H(t - T))$ .

Let us prove that

$$(5.6) \quad u_n \longrightarrow \tilde{u} \quad \text{as } n \longrightarrow \infty \text{ in } L^2(0, T).$$

Since  $\tilde{u} \in \mathcal{B}^U(0, T) \subset L^2(0, T)$  then

$$(5.7) \quad \frac{T}{2} \sum_{l=0}^{\infty} \nu_l^2 = \left( \|\tilde{u}\|_{L^2(0, T)} \right)^2.$$

Taking this into account we have from [16, Chapter 3, section 1] that the first and the second Cesàro means tend to  $\|\tilde{u}\|_{L^2(0, T)}$  as  $n \rightarrow \infty$ , i.e.,

$$(5.8) \quad \sum_{l=0}^n \frac{n+1-l}{n+1} \nu_l^2 \longrightarrow \frac{T}{2} \sum_{l=0}^{\infty} \nu_l^2 \quad \text{as } n \longrightarrow \infty,$$

$$(5.9) \quad \sum_{l=0}^n \frac{(n+1-l)(n+2-l)}{(n+1)(n+2)} \nu_l^2 \longrightarrow \frac{T}{2} \sum_{l=0}^{\infty} \nu_l^2 \quad \text{as } n \longrightarrow \infty.$$

Since (5.8) implies  $\sum_{l=0}^n \frac{l}{n+1} \nu_l^2 \rightarrow 0$  as  $n \rightarrow \infty$  then (5.9) yields  $\sum_{l=0}^n \frac{l^2}{(n+1)^2} \nu_l^2 \rightarrow 0$  as  $n \rightarrow \infty$ . Summarizing, we obtain

$$(5.10) \quad \begin{aligned} \|\tilde{u} - u_n\|_{L^2(0, T)} &= \left\| \sum_{l=0}^{\infty} \nu_l \sin \frac{\pi l t}{T} - \sum_{l=0}^n \frac{n+1-l}{n+1} \nu_l \sin \frac{\pi l t}{T} \right\|_{L^2(0, T)} \\ &= \sqrt{\frac{T}{2}} \left( \sum_{l=n+1}^{\infty} \nu_l^2 + \sum_{l=0}^n \frac{l^2}{(n+1)^2} \nu_l^2 \right)^{1/2} \longrightarrow 0 \quad \text{as } n \longrightarrow \infty, \end{aligned}$$

and hence (5.6) holds.

Taking into account (4.11), (5.4), and (5.10) we get the following.

**THEOREM 5.1.** *Let  $T > 0$ ,  $w^0 \in \mathcal{H}^s$ ,  $s \leq 1$ . Also, let conditions (3.15)–(3.17) be fulfilled,  $\{\nu_n\}_{n=0}^\infty$  and  $\{u_n\}_{n=0}^\infty$  be defined by (5.2), (5.3), respectively. Then  $u_n \in \mathcal{B}^U(0, T)$ ,  $n = \overline{0, \infty}$ , and*

$$\|W^n(\cdot, T)\|^s \leq \sqrt{2T} M_q^T L_q^s \left( \sum_{l=n+1}^\infty \nu_l^2 + \sum_{l=0}^n \frac{l^2}{(n+1)^2} \nu_l^2 \right)^{1/2} \longrightarrow 0 \text{ as } n \longrightarrow \infty,$$

where  $W^n$  is the solution of (3.2), (3.3) corresponding to the control  $u_n$ . Moreover,  $u_n \in C^\infty(0, T)$  and  $\mathcal{U}_n \in H_0^p$ ,  $p < 3/2$ ,  $n = \overline{0, \infty}$ , where  $\mathcal{U}_n(t) = u_n(t)(H(t) - H(t - T))$ .

Thus the continuous controls  $u_n$ ,  $n = \overline{0, \infty}$ , defined by (5.3) are solutions of the approximate null-controllability problem.

*Remark 5.1.* It follows from Proposition 3.2 and (5.5) that  $W^n(\cdot, T) \in H_0^p$  if  $\Xi w_0^0 \in H_0^p$  and  $p < 5/2$ . It is well known [3, Chapter 1] that  $\tilde{H}^p \subset C^1(\mathbb{R}) \times C(\mathbb{R})$  if  $p > 3/2$ . Thus  $W^n(\cdot, T) \in C^1(\mathbb{R}) \times C(\mathbb{R})$  if  $\Xi w_0^0 \in H_0^p$ ,  $3/2 < p < 5/2$ .

For  $n \in \mathbb{N}$  denote  $\lambda_n = \sqrt{(\pi n/T)^2 - q^2}$ ,

$$\tilde{\omega}_j^n = \sqrt{\frac{\pi}{2}} (\mathcal{F} \Xi w_0^0)(\lambda_n), \quad j = 1, 2.$$

Then  $\nu_n = \frac{2i}{T} \sqrt{\frac{\pi}{2}} (\mathcal{F} \tilde{\mathcal{U}}) \left( \frac{\pi n}{T} \right) = \frac{2i}{T} \sqrt{\frac{\pi}{2}} (\mathcal{F} (\Phi^{-1} \Xi w_0^0)) \left( \frac{\pi n}{T} \right) = -\frac{2\pi n}{T^2} \tilde{\omega}_0^n$ . Condition (3.17) is equivalent to  $\Omega w_1^0 = \Phi \tilde{\mathcal{U}}'$ , where  $\tilde{\mathcal{U}}$  is given by (5.1). Therefore

$$\begin{aligned} \tilde{\omega}_1^n &= \frac{Ti}{\pi n} \sqrt{\frac{\pi}{2}} \left( \mathcal{F} (\Xi \tilde{\mathcal{U}})' \right) \left( \frac{\pi n}{T} \right) = - \int_0^T \cos \frac{\pi n t}{T} \tilde{u}(t) dt \\ &= - \sum_{l=0}^\infty \nu_l \int_0^T \cos \frac{\pi n t}{T} \sin \frac{\pi l t}{T} dt = \frac{2}{T} \sum_{l=0}^\infty ((-1)^{l+n} - 1) \frac{l^2}{l^2 - n^2} \tilde{\omega}_0^l, \quad n = \overline{0, \infty}. \end{aligned}$$

Thus the following assertion is true.

*Remark 5.2.* Condition (3.17) is equivalent to

$$\tilde{\omega}_1^n = \frac{2}{T} \sum_{l=0}^\infty ((-1)^{l+n} - 1) \frac{l^2}{l^2 - n^2} \tilde{\omega}_0^l, \quad n = \overline{0, \infty}.$$

*Example 5.1.* Let  $s \leq 1$ ,  $q \geq 0$ ,  $T > 0$ ,  $w_0^0(x) = \frac{1}{2} (x^2 - T^2) (H(x) - H(x - T))$ ,  $w_1^0(x) = (\Phi \frac{d}{dt} (\operatorname{sgn} t \Phi^{-1} \Xi w_0^0)) (x)$ . Evidently,  $w^0 \in \mathcal{H}^s$ . With regard to Lemma 6.4 we get  $(\Phi^{-1} \Xi w_0^0)(t) = t I_0 (q \sqrt{T^2 - t^2}) (H(t + T) - H(t - T))$ .

By Theorem 3.3 the state  $w^0$  is null-controllable (and hence approximately null-controllable) if  $U \geq \sup \{ |t I_0 (q \sqrt{T^2 - t^2})| \mid t \in [0, T] \}$ .

Let us find the continuous controls  $u_n$ ,  $n = \overline{1, \infty}$ , of the form (5.3) that solve approximate null-controllability problem and estimate  $\|W^n(\cdot, T)\|^s$ , where  $W^n$  is the solution of (3.2), (3.3) corresponding to  $u_n$ .

For  $l \in \mathbb{N}$  we have  $\nu_n = -\frac{2\pi n}{T^2} \tilde{\omega}_0^n$ ,

$$\begin{aligned} \tilde{\omega}_0^l &= \frac{1}{2} \int_0^T \cos(\lambda_l x) (x^2 - T^2) dx = \frac{1}{\lambda_l^2} \left( T \cos(\lambda_l T) - \frac{\sin(\lambda_l T)}{\lambda_l} \right), \\ \nu_l &= \frac{2\pi l}{T^2 \lambda_l^2} \left( \frac{\sin(\lambda_l T)}{\lambda_l} - T \cos(\lambda_l T) \right), \end{aligned}$$

and

$$u_n(t) = \frac{2\pi}{T^2} \sum_{l=1}^n \frac{(n+1-l)l}{(n+1)\lambda_l^2} \left( \frac{\sin(\lambda_l T)}{\lambda_l} - T \cos(\lambda_l T) \right) \sin \frac{\pi l t}{T}, \quad t \in [0, T].$$

We have  $|\lambda_l| \leq q \Leftrightarrow 1 \leq l \leq \sqrt{2}Tq/\pi$ . Set  $n_0 \in \mathbb{N}$  such that  $n_0 \leq \sqrt{2}Tq/\pi$  and  $n_0 + 1 > \sqrt{2}Tq/\pi$ . Therefore

$$|\nu_l| \leq \frac{8\pi T l \sinh^2(Tq/2)}{(Tq)^2} \quad \text{if } 1 \leq l \leq n_0,$$

$$|\nu_l| \leq \frac{8T}{\pi l} \quad \text{if } l \geq n_0 + 1.$$

Let  $n > n_0$ . With regard to (5.10) we obtain

$$\begin{aligned} & \sqrt{2T} \left( \sum_{l=n+1}^{\infty} \nu_l^2 + \sum_{l=1}^n \frac{l^2}{(n+1)^2} \nu_l^2 \right)^{1/2} \\ & \leq 4(2T)^{3/2} \left( \frac{1}{\pi^2} \sum_{l=n+1}^{\infty} \frac{1}{l^2} + \frac{1}{\pi^2(n+1)^2} \sum_{l=n_0+1}^n 1 + \frac{\pi^2 \sinh^4(Tq/2)}{(Tq)^4(n+1)^2} \sum_{l=1}^{n_0} l^4 \right)^{1/2} \\ & \leq 4(2T)^{3/2} \left( \frac{2}{\pi^2 n} + (1 - \delta_{0,n_0}) \frac{\pi^2 \sinh^4(Tq/2)}{5(Tq)^4} \frac{(n_0+1)^5}{(n+1)^2} \right) = \varepsilon_n, \end{aligned}$$

where  $\delta_{km}$  is the Kronecker delta  $\delta_{km} = 1$  if  $k = m$ , and  $\delta_{km} = 0$  otherwise. According to Theorem 5.1  $\|W^n(\cdot, T)\|^s \leq M_q^T L_q^s \varepsilon_n$ . For example, for various  $q \geq 0$  and  $T > 0$  we have

$$\begin{aligned} q = 0, \quad T > 0, \quad n_0 = 0, \quad \varepsilon_n &= \frac{8(2T)^{3/2}}{\pi\sqrt{n}}, \\ q > 0, \quad Tq = 1, \quad n_0 = 0, \quad \varepsilon_n &= \frac{8(2T)^{3/2}}{\pi\sqrt{n}}, \\ q > 0, \quad Tq = 4, \quad n_0 = 1, \quad \varepsilon_n &= 4(2T)^{3/2} \sqrt{\frac{2}{\pi^2 n} + \frac{\pi^2 \sinh^4 2}{40(n+1)^2}}, \\ q > 0, \quad Tq = 6, \quad n_0 = 2, \quad \varepsilon_n &= 4(2T)^{3/2} \sqrt{\frac{2}{\pi^2 n} + \frac{3\pi^2 \sinh^4 3}{80(n+1)^2}}, \\ q > 0, \quad Tq = 8, \quad n_0 = 3, \quad \varepsilon_n &= 4(2T)^{3/2} \sqrt{\frac{2}{\pi^2 n} + \frac{\pi^2 \sinh^4 4}{20(n+1)^2}}, \\ q > 0, \quad Tq = 10, \quad n_0 = 4, \quad \varepsilon_n &= 4(2T)^{3/2} \sqrt{\frac{2}{\pi^2 n} + \frac{\pi^2 \sinh^4 5}{16(n+1)^2}}. \end{aligned}$$

**6. Appendix.** We assume throughout this section that  $p \in \mathbb{R}$ .

Denote  $\partial : \mathcal{S}' \longrightarrow \mathcal{S}'$  with  $D(\partial) = \{f \in \mathcal{S}' \mid f \text{ is even and } \text{supp } f \subset [-T, T]\}$  such that  $\partial f = f'$ ,  $f \in D(\partial)$ . Evidently,  $R(\partial) = \{g \in \mathcal{S}' \mid f \text{ is odd and } \text{supp } g \subset [-T, T]\}$ . Then

$$(6.1) \quad \|\partial f\|_0^{p-1} = \|i\sigma \mathcal{F}f\|_{p-1}^0 \leq \|\mathcal{F}f\|_p^0 = \|f\|_0^p, \quad f \in D(\partial) \cap H_0^p;$$

i.e.,  $\partial$  is bounded linear operator from  $H_0^p$  to  $H_0^{p-1}$ . Due to the Paley–Wiener theorem we obtain that  $\partial$  is invertible and  $D(\partial^{-1}) = R(\partial)$ . According to the inverse operator theorem we conclude that  $\exists N^p > 0$  such that

$$(6.2) \quad \|\partial^{-1}g\|_0^p \leq N^p \|g\|_0^{p-1}, \quad g \in R(\partial) \cap H_0^{p-1} = D(\partial^{-1}) \cap H_0^{p-1}.$$

LEMMA 6.1. *Let  $g \in D(\Phi)$ ,  $f = \Phi g$ . Then*

- (i)  $\text{supp } f \subset [-T, T]$ ;
- (ii)  $f$  is even;
- (iii)  $\|f\|_0^p \leq L_q^p \|g\|_0^{p-1}$ ,  $g \in D(\Phi) \cap H_0^p$ , where  $L_q^p > 0$ ;
- (iv) if  $q = 0$ , then  $\Phi = \partial^{-1}$ .

*Proof.* Taking into account the generalized Paley–Wiener theorem [2, Chapter 3] we conclude that (1)  $G = \mathcal{F}g$  is a regular functional; (2)  $G$  is of a polynomial growth on  $\mathbb{R}$ ; and (3)  $G$  can be extended to an entire function of the order  $\leq 1$  and the type  $\leq T$ . Evidently,  $G$  is odd. Therefore  $\frac{1}{\sqrt{\sigma^2 + q^2}} G(\sqrt{\sigma^2 + q^2})$  is an even entire function of the order  $\leq 1$  and the type  $\leq T$  with a polynomial growth on  $\mathbb{R}$ . Using again the generalized Paley–Wiener theorem we obtain that  $f \in \mathcal{S}'$  and (i), (ii) are true. Let us prove (iii). Since  $\|G\|_{p-1}^0 = \|g\|_0^{p-1}$  we then have

$$\begin{aligned} (\|f'\|_0^{p-1})^2 &= \left( \left\| \frac{\sigma G(\sqrt{\sigma^2 + q^2})}{\sqrt{\sigma^2 + q^2}} \right\|_{p-1}^0 \right)^2 \\ &= 2 \int_0^\infty (1 + \sigma^2)^{p-1} \left| \frac{G(\sqrt{\sigma^2 + q^2})}{\sqrt{\sigma^2 + q^2}} \right|^2 \sigma^2 d\sigma \\ &= 2 \int_q^\infty (1 - q^2 + \mu^2)^{p-1} |G(\mu)|^2 \sqrt{1 - \frac{q^2}{\mu^2}} d\mu \leq (\tilde{L}_q^p)^2 (\|g\|_0^{p-1})^2, \end{aligned}$$

where  $\tilde{L}_q^p = (1 + q^2)^{-(p-1)/2}$  if  $p \leq 1$ , and  $\tilde{L}_q^p = 1$  otherwise. With regard to (6.2) we obtain  $\|f\|_0^p \leq N^p \|f'\|_0^{p-1} \leq L_q^p \|g\|_0^{p-1}$ , where  $L_q^p = N^p \tilde{L}_q^p$ . Therefore (iii) holds. If  $q = 0$ , then  $f = \Phi g = \mathcal{F}^{-1} \left( \frac{(\mathcal{F}g)(\sigma)}{(i\sigma)} \right) = \partial^{-1}g$ ; i.e., (iv) is true. The lemma is proved.

LEMMA 6.2. *Let  $g \in D(\Phi) \cap H_0^0$ ,  $f = \Phi g$ . Then*

$$(6.3) \quad f(x) = - \int_{|x|}^\infty J_0(q\sqrt{t^2 - x^2}) g(t) dt.$$

*Proof.* From the definition of  $\Phi$  we get

$$f = -i\mathcal{F}_{\sigma \rightarrow x}^{-1} \left( \frac{(\mathcal{F}g)(\sqrt{\sigma^2 + q^2})}{\sqrt{\sigma^2 + q^2}} \right) = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty g(t) \mathcal{F}_{\sigma \rightarrow x}^{-1} \left( \frac{\sin(t\sqrt{\sigma^2 + q^2})}{\sqrt{\sigma^2 + q^2}} \right) dt.$$

With regard to (3.11) we conclude that

$$f(x) = -\frac{1}{2} \int_{-\infty}^\infty g(t) J_0(q\sqrt{t^2 - x^2}) H(t^2 - x^2) \text{sgn } t dt.$$

Therefore 6.3 is true. The lemma is proved.



Denote  $\Psi : \mathcal{S}' \longrightarrow \mathcal{S}'$  with  $D(\Psi) = \{f \in \mathcal{S}' \mid f \text{ is even and } \text{supp } g \subset [-T, T]\}$  such that

$$(6.4) \quad (\Psi f)(t) = \mathcal{F}_{\mu \rightarrow t}^{-1} \left( i\mu \left( \mathcal{F}f \right) \left( \sqrt{\mu^2 - q^2} \right) \right) (t), \quad f \in D(\Psi).$$

LEMMA 6.3. *Let  $f \in D(\Psi)$ ,  $g = \Psi f$ . Then*

- (i)  $\text{supp } g \subset [-T, T]$ ;
- (ii)  $g$  is odd;
- (iii)  $\|g\|_0^{p-1} \leq K_q^{p-1} \|f\|_0^p$ ,  $f \in D(\Psi) \cap H_0^p$ , where  $K_q^p > 0$ ;
- (iv) if  $q = 0$ , then  $\Psi = \partial$ .

In addition,  $R(\Psi) = D(\Phi)$ ,  $D(\Psi) = R(\Phi)$ , and  $\Psi = \Phi^{-1}$ .

*Proof.* Reasoning as in the proof of Lemma 6.1 we conclude that  $f \in \mathcal{S}'$  and (i), (ii) are true. Hence  $R(\Psi) \subset D(\Phi)$ . Let us prove (iii). We have  $R(\Phi) \subset D(\Psi)$  and

$$(6.5) \quad \Psi \Phi g = \mathcal{F}_{\mu \rightarrow t}^{-1} \left( i\mu \left( \mathcal{F}_{x \rightarrow \xi} (\Phi g)(x) \right) (\xi) \Big|_{\xi = \sqrt{\mu^2 - q^2}} \right) = \mathcal{F}_{\mu \rightarrow t}^{-1} \left( \frac{\mu}{\mu} (\mathcal{F}g)(\mu) \right) = g;$$

i.e.,  $\Phi : D(\Phi) \longrightarrow R(\Phi)$  is invertible and  $\Phi^{-1} = \Psi$ .

Denote by  $\Phi_p$  the restriction of  $\Phi$  on  $H_0^{p-1}$ :  $\Phi_p : D(\Phi) \cap H_0^{p-1} \longrightarrow R(\Phi) \cap H_0^p$ . According to Lemma 6.1  $\Phi_p$  is bounded. It follows from (6.5) that  $\Phi_p$  is invertible and  $\Phi_p^{-1} = \Psi_{p-1}$ , where  $\Psi_{p-1}$  is the restriction of  $\Psi$  on  $H_0^p$ :  $\Psi_p : D(\Psi) \cap H_0^p \longrightarrow R(\Psi) \cap H_0^{p-1}$ . Applying the inverse operator theorem we conclude that  $\Psi_{p-1}$  is bounded, i.e., (iii) holds. If  $q = 0$ , then  $g = \Phi f = \frac{d}{dt} \mathcal{F}^{-1}((\mathcal{F}g)(\sigma)) = \partial g$ , i.e., (iv) is true. The lemma is proved.

LEMMA 6.4. *Let  $f \in D(\Phi^{-1}) \cap H_0^1$ ,  $g = \Phi^{-1}f$ . Then*

$$(6.6) \quad g(t) = -\frac{d}{dt} \int_{|t|}^{\infty} I_0 \left( q\sqrt{x^2 - t^2} \right) f'(x) dx = f'(t) + qt \int_{|t|}^{\infty} \frac{I_1 \left( q\sqrt{x^2 - t^2} \right)}{\sqrt{x^2 - t^2}} f'(x) dx.$$

*Proof.* We have  $g(t) = \frac{d}{dt} \mathcal{F}_{\mu \rightarrow t}^{-1} \left( \frac{-i}{\sqrt{\mu^2 - q^2}} (\mathcal{F}f')(\sqrt{\mu^2 - q^2}) \right)$ . Since  $\frac{\sin(t\sqrt{\sigma^2 + q^2})}{\sqrt{\sigma^2 + q^2}}$  is an entire function with respect to  $(\sigma, q)$ , then replacing  $q$  by  $iq$  we get from (3.11)

$$(6.7) \quad \mathcal{F}_{\sigma \rightarrow x}^{-1} \left( \frac{\sin \left( t\sqrt{\sigma^2 - q^2} \right)}{\sqrt{\sigma^2 - q^2}} \right) (x) = \sqrt{\frac{\pi}{2}} I_0 \left( q\sqrt{t^2 - x^2} \right) H(t^2 - x^2) \text{sgn } t.$$

Reasoning as in the proof of Lemma 6.2 and using (6.7) instead of (3.11) we conclude that (6.6) holds. The lemma is proved.

LEMMA 6.5. *Let  $f \in D(\Phi^{-1}) \cap H_0^0$ . Then*

$$(6.8) \quad \Phi^{-1}f = \tilde{g}',$$

where  $\tilde{g} \in H_0^0$  is even,  $\text{supp } \tilde{g} \subset [-T, T]$ , and

$$(6.9) \quad \tilde{g}(t) = f(t) + \int_{|t|}^{\infty} \frac{qx I_1 \left( q\sqrt{x^2 - t^2} \right)}{\sqrt{x^2 - t^2}} f(x) dx,$$

$$(6.10) \quad f(x) = \tilde{g}(x) - \int_{|x|}^{\infty} \frac{qt J_1 \left( q\sqrt{t^2 - x^2} \right)}{\sqrt{t^2 - x^2}} \tilde{g}(t) dt.$$

Moreover,

- (i) if  $|f(x)| \leq \mathcal{F}$  a.e. on  $[-T, T]$ , then  $|\tilde{g}(t)| \leq \mathcal{F}I_0(qT)$  a.e. on  $[-T, T]$ ;
- (ii) if  $|\tilde{g}(t)| \leq \mathcal{G}$  a.e. on  $[-T, T]$ , then  $|f(x)| \leq \mathcal{G}(1 + qT)$  a.e. on  $[-T, T]$ .

*Proof.* It follows from Lemma 6.4 that

$$\tilde{g}(t) = - \int_{|t|}^{\infty} I_0 \left( q\sqrt{x^2 - t^2} \right) f'(x) dx + C.$$

Evidently  $\text{supp } \tilde{g} \subset [-T, T]$  iff  $C = 0$ . Hence (6.9) is true. It follows from (6.8) and Lemma 6.2 that (6.10) is also true. It remains to prove (i) and (ii). If  $|f(x)| \leq \mathcal{F}$  a.e. on  $[-T, T]$ , then  $|\tilde{g}(t)| \leq \mathcal{F} \int_{|t|}^T \left| \frac{\partial}{\partial x} I_0 \left( q\sqrt{x^2 - t^2} \right) \right| dx \leq \mathcal{F}I_0(qT)$ . If  $|\tilde{g}(t)| \leq \mathcal{G}$  a.e. on  $[-T, T]$ , then  $|f(x)| \leq \mathcal{G} + \mathcal{G} \int_{|x|}^T \frac{qt}{\sqrt{t^2 - x^2}} dt \leq \mathcal{G}(1 + qT)$ . Thus (i) and (ii) hold and the lemma is proved.

LEMMA 6.6. Let  $f \in H_0^1$  be even,  $\text{supp } f \subset [-T, T]$ ,  $h = \Phi \frac{d}{dt} (\text{sgn } t \Phi^{-1} f)$ . Then

$$(6.11) \quad h(x) = \text{sgn } x f'(x) + q \int_{|x|}^{\infty} f(\xi) \frac{I_1(q(\xi - |x|))}{\xi - |x|} d\xi.$$

*Proof.* Put  $F = \mathcal{F}f$ . Denote  $G(\xi) = \frac{1}{i\xi} \mathcal{F}_{t \rightarrow \xi} \left( \frac{d}{dt} (\text{sgn } t (\Phi^{-1} f)(t)) \right) (\xi)$ . Then

$$\begin{aligned} G &= (\mathcal{F}_{t \rightarrow \xi} (\text{sgn } t (\Phi^{-1} f)(t))) = \frac{1}{\pi} \mathcal{P} \frac{1}{\xi} * \xi F \left( \sqrt{\xi^2 - q^2} \right) \\ &= \frac{2}{\pi} \text{V.p.} \int_0^{\infty} \frac{1}{\xi^2 - \mu^2} \mu^2 F \left( \sqrt{\mu^2 - q^2} \right) d\mu. \end{aligned}$$

Therefore  $h = -\Delta_q \mathcal{F}_{\sigma \rightarrow x} \frac{G(\sqrt{\sigma^2 + q^2})}{\sigma^2 + q^2} = -\frac{1}{2} \Delta_q \left( \frac{e^{-q|x|}}{q} * (\mathcal{F}_{\sigma \rightarrow x} G(\sqrt{\sigma^2 + q^2})) \right)$ , where  $\Delta_q = \left( \frac{d}{dx} \right)^2 - q^2$ . Hence

$$\begin{aligned} h &= \frac{1}{2} \sqrt{\frac{2}{\pi}} \Delta_q \int_0^{\infty} \left( \left( \frac{e^{-q|x|}}{q} * \frac{\sin(|x|\sqrt{\mu^2 - q^2})}{\sqrt{\mu^2 - q^2}} \right) H(\mu^2 - q^2) \right. \\ &\quad \left. - \left( \frac{e^{-q|x|}}{q} * \frac{e^{|x|\sqrt{q^2 - \mu^2}}}{\sqrt{q^2 - \mu^2}} \right) H(q^2 - \mu^2) \right) \mu^2 F \left( \sqrt{\mu^2 - q^2} \right) d\mu \\ &= \frac{2}{\pi} \Delta_q \left( -\frac{e^{-q|x|}}{q} \int_0^{\infty} f'(\nu) \int_0^{\infty} \frac{\sin(\nu\sqrt{\mu^2 - q^2})}{\sqrt{\mu^2 - q^2}} d\mu d\nu \right. \\ &\quad \left. + \int_0^{\infty} f(\nu) \int_0^{\infty} \frac{\cos(\nu\sqrt{\mu^2 - q^2}) \sin(|x|\sqrt{\mu^2 - q^2})}{\sqrt{\mu^2 - q^2}} d\mu d\nu \right. \\ &\quad \left. - \int_0^{\infty} f(\nu) \int_0^q \frac{\cosh(\nu\sqrt{q^2 - \mu^2}) \cosh(x\sqrt{q^2 - \mu^2})}{\sqrt{q^2 - \mu^2}} d\mu d\nu \right). \end{aligned}$$

Taking into account (6.7) and  $\frac{2}{\pi} \int_0^1 \frac{\cosh(\xi t)}{\sqrt{1-t^2}} dt = I_0(\xi)$  we obtain

$$\begin{aligned} h(x) &= \frac{1}{2} \Delta_q \left( (f'(x) * (I_0(qx) \operatorname{sgn} x)) \Big|_{x=0} \frac{e^{-q|x|}}{q} \right. \\ &\quad \left. + \operatorname{sgn} x (f(x) * (I_0(qx) \operatorname{sgn} x)) - f(x) * I_0(qx) \right) \\ &= \frac{1}{2} (\operatorname{sgn} x \Delta_q (f(x) * (\operatorname{sgn} x I_0(qx))) - \Delta_q (f(x) * I_0(qx))) \\ &= \operatorname{sgn} x f'(x) + \frac{1}{2} \operatorname{sgn} x (f(x) * (\operatorname{sgn} x (\Delta_q I_0(qx)))) - \frac{1}{2} f(x) * (\Delta_q I_0(qx)) \\ &= \operatorname{sgn} x f'(x) - \int_0^\infty f(\xi + |x|) (\Delta_q I_0(q\xi)) d\xi. \end{aligned}$$

Since  $\Delta_q I_0(q\xi) = (q^2/2) (I_2(q\xi) - I_0(q\xi)) = -qI_1(q\xi)/\xi$ , we conclude that (6.11) holds and the lemma is proved.

LEMMA 6.7. *If  $f \in H_0^p \times H_0^{p-1}$  and  $g = \mathcal{F}f$ , then*

$$(6.12) \quad \|E(x, t) * f\|^p = \|\Sigma(|\sigma|, t)g\|_p^p \leq M_q^T \|g\|_p^p = M_q^T \|f\|^p, \quad t \in \mathbb{R},$$

where  $M_q^T = \sqrt{(2t^2 + 6)(1 + q^2)}$ .

*Proof.* For all  $t \in \mathbb{R}$  we have

$$\begin{aligned} \|E(x, t) * f\|^p &= \|\Sigma(\sigma, t)g\|_p^p \\ &\leq \left\| \begin{pmatrix} \cos(t\sqrt{\sigma^2 + q^2}) \\ -\sqrt{\sigma^2 + q^2} \sin(t\sqrt{\sigma^2 + q^2}) \end{pmatrix} g_0 \right\|_p^p + \left\| \begin{pmatrix} \sin(t\sqrt{\sigma^2 + q^2}) \\ \sqrt{\sigma^2 + q^2} \cos(t\sqrt{\sigma^2 + q^2}) \end{pmatrix} g_1 \right\|_p^p \\ &\leq \sqrt{2}\sqrt{1 + q^2} \|g_0\|_p^0 + \left( \left( \left\| \frac{\sin(t\sqrt{\sigma^2 + q^2})}{\sqrt{\sigma^2 + q^2}} g_1 \right\|_p^0 \right)^2 + \left( \|g_1\|_{p-1}^0 \right)^2 \right)^{1/2}. \end{aligned}$$

Since  $(1 + |\sigma|^2) \left| \frac{\sin(t\sqrt{\sigma^2 + q^2})}{\sqrt{\sigma^2 + q^2}} \right|^2 \leq t^2 + 2$  we obtain (6.12). The lemma is proved.

## REFERENCES

- [1] H. O. FATTORINI, *Infinite Dimensional Optimization and Control Theory*, Cambridge University Press, Cambridge, UK, 1999.
- [2] I. M. GELFAND AND G. E. SHILOV, *Generalized Functions*, 3, Fismatgiz, Moscow, 1958 (in Russian).
- [3] S. G. GINDIKIN AND L. R. VOLEVICH, *Distributions and Convolution Equations*, Gordon and Breach, Philadelphia, 1992.
- [4] M. GUGAT, *Optimal boundary control of a string to rest in finite time with continuous state*, ZAMM Z. Angew. Math. Mech., 86 (2006), pp. 134–150.
- [5] M. GUGAT AND G. LEUGERING, *Solutions of  $L^p$ -norm-minimal control problems for the wave equation*, Comput. Appl. Math., 21 (2002), pp. 227–244.
- [6] M. GUGAT, G. LEUGERING, AND G. SKLYAR,  *$L^p$ -optimal boundary control for the wave equation*, SIAM J. Control Optim., 44 (2005), pp. 49–74.
- [7] W. KRABS AND G. LEUGERING, *On boundary controllability of one-dimension vibrating systems by  $W_0^{1,p}$ -controls for  $p \in [0, \infty)$* , Math. Methods Appl. Sci., 17 (1994), pp. 77–93.

- [8] M. G. KREIN AND A. A. NUDEL'MAN, *The Markov Moment Problem and Extremal Problems*, Nauka, Moscow, 1973 (in Russian); English translation, AMS, Providence, RI, 1977.
- [9] V. I. KOROBV AND G. M. SKLYAR, *Time optimality and the power moment problem*, Mat. Sb., 134 (1987), pp. 186–206.
- [10] I. LASIECKA AND R. TRIGGIANI, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories. 2: Abstract Hyperbolic-Like Systems over a Finite Time Horizon*, Cambridge University Press, Cambridge, UK, 2000.
- [11] M. NEGREANU AND E. ZUAZUA, *Convergence of a multigrid method for the controllability of a 1-d wave equation*, C.R. Math. Acad. Sci. Paris, 338 (2004), pp. 413–418.
- [12] L. SCHWARTZ, *Théorie des Distributions*, Vol. 1, Hermann, Paris, 1950.
- [13] L. SCHWARTZ, *Théorie des Distributions*, Vol. 2, Hermann, Paris, 1951.
- [14] G. M. SKLYAR AND L. V. FARDIGOLA, *The Markov power moment problem in problems of controllability and frequency extinguishing for the wave equation on a half-axis*, J. Math. Anal. Appl., 276 (2002), pp. 109–134.
- [15] G. M. SKLYAR AND L. V. FARDIGOLA, *The Markov trigonometric moment problem in controllability problems for the wave equation on a half-axis*, Matem. Fizika, Analiz, Geometriya, 9 (2002), pp. 233–242.
- [16] A. ZIGMUND, *Trigonometric Series*, Vol. 1, Cambridge University Press, Cambridge, UK, 1959.

# OPTIMAL DIVIDEND PAYMENTS AND REINVESTMENTS OF DIFFUSION PROCESSES WITH BOTH FIXED AND PROPORTIONAL COSTS\*

JOSTEIN PAULSEN†

**Abstract.** Assets are assumed to follow a diffusion process subject to some conditions. The owners can pay dividends at their discretion, but whenever assets reach zero, they have to reinvest money so that assets never go negative. With each dividend payment there is a fixed and a proportional cost, and so with reinvestments. The goal is to maximize the expected value of discounted net cash flow, i.e., dividends paid minus reinvestments. It is shown that there can be two different solutions depending on the model parameters and the costs as follows. (1) Whenever assets reach a barrier  $y^*$  they are reduced to  $y^* - \delta^*$  through a dividend payment, and whenever they reach 0 they are increased to  $\gamma^* \leq y^* - \delta^*$  by a reinvestment. (2) There is no optimal policy, but the value function is approximated by policies of the form described in (1) for increasing barriers. We provide criteria to decide whether an optimal solution exists, and when it does not, we show how to calculate the value function. We discuss how the problem can be solved numerically and give numerical examples.

**Key words.** optimal dividends, diffusion models, impulse control, barrier strategy

**AMS subject classifications.** 49N25, 93E20, 91B28, 60J70

**DOI.** 10.1137/070691632

**1. Introduction and model formulation.** In this paper the value of a company is defined as the expected present value of dividends paid to the owners minus reinvestments made. Assuming no transaction costs and other market imperfections, the Modigliani–Miller theorem says that the value of the company is independent of its dividend and reinvestment policy. However, this assumption is not likely to hold in practice. Transactions such as dividend payments and reinvestments are not likely to be free of costs. There may be tax differences between dividends received and asset appreciations, and these can be formulated as costs. Investment of the company's assets may be expensive, which is an issue we review in section 3. When transaction costs are taken into account, dividend and reinvestment policies do matter, and the valuation problem becomes more complex. A justification for our valuation method in the presence of transaction costs can be found in Sethi and Taksar [11].

In addition to finding the value of the company, it is of course also useful to find the optimal dividend as well as reinvestment policies, when they exist, and to know when there are no such optimal policies. These problems are addressed in this paper for a rather general diffusion model when transaction costs consist of a constant part and a part that is proportional to the amount transacted.

To get the mathematical formalism right, let  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, P)$  be a probability space satisfying the usual conditions; i.e., the filtration  $(\mathcal{F}_t)_{t \geq 0}$  is right continuous and  $P$ -complete. The income process without dividends is assumed to follow the dynamics

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t,$$

where  $W$  is a Brownian motion on the probability space.

---

\*Received by the editors May 14, 2007; accepted for publication (in revised form) April 3, 2008; published electronically July 30, 2008.

<http://www.siam.org/journals/sicon/47-5/69163.html>

†Department of Mathematics, University of Bergen, Johs. Brunsgt. 12, 5008 Bergen, Norway (jostein@math.uib.no).

From this income process dividends can be paid out to the owners, but at a cost. Total dividends paid out until time  $t$  are denoted by  $D_t$ . The costs incurred up to time  $t$  from paying out  $D_t$  are denoted by  $\bar{D}_t$ .

In addition, whenever the income process reaches zero, the owners are under the obligation to reinvest in order to prevent it from going negative. Total reinvestments until time  $t$  are denoted by  $C_t$ . Again costs are incurred with reinvestments, and the costs incurred up to time  $t$  from reinvesting  $C_t$  are denoted by  $\bar{C}_t$ .

If the processes  $C$ ,  $D$ ,  $\bar{C}$ , and  $\bar{D}$  are nondecreasing, adapted, and right continuous, we say they are admissible.

After dividends and reinvestments, the capital  $Y$  retained in the company has the dynamics

$$(1.1) \quad dY_t = \mu(Y_t)dt + \sigma(Y_t)dW_t + dC_t - dD_t - d\bar{C}_t - d\bar{D}_t,$$

with  $Y_{0-} = y$ .

Let  $\mathcal{D}$  be the set of all admissible dividend and reinvestment policies. For  $(C, D) \in \mathcal{D}$ , let  $\nu_n = \inf\{t : C_t \vee D_t > n\}$  and define

$$V_{C,D}(y) = \limsup_{n \rightarrow \infty} E^y \left[ \int_{0-}^{\nu_n-} e^{-rt} dA_t \right],$$

where  $A_t = D_t - C_t$  is net payout. Here  $r$  is a properly chosen discount factor and  $E^y$  is expectation with  $Y_{0-} = y$ . The reason for introducing the stopping times  $\{\nu_n\}$  is that both  $C_t$  and  $D_t$  can go to infinity as  $t$  goes to infinity, and therefore  $\lim_{t \rightarrow \infty} A_t$  may not exist, and we are a priori not guaranteed that  $\lim_{t \rightarrow \infty} e^{-rt} A_t$  exists. Therefore the somewhat complicated definition of  $V_{C,D}(y)$  is required.

Our task is to find

$$(1.2) \quad V^*(y) = \sup_{\mathcal{D}} V_{C,D}(y).$$

Furthermore, if it exists, we also want to find the optimal policy  $(C^*, D^*) \in \mathcal{D}$ .

In this paper it is assumed that with each dividend payment there is a fixed cost  $d_0$ , independent of the size of the payment, plus a part that is proportional to the size of the payment. Similarly, with each reinvestment there is a fixed cost  $c_0$  plus a proportional part. Because of the fixed costs, there can be only a finite number of payments on each finite time interval, so we can write

$$(1.3) \quad \begin{aligned} \bar{C}_t &= c_0 \sum_{s \leq t} 1_{\{\Delta C_s > 0\}} + c_1 C_t, \quad 0 \leq c_1 < 1, \\ \bar{D}_t &= d_0 \sum_{s \leq t} 1_{\{\Delta D_s > 0\}} + d_1 D_t, \end{aligned}$$

where  $c_1$  and  $d_1$  are the proportionality factors, assumed nonnegative. We shall call such costs linear costs. Note that  $c_1 \geq 1$  is meaningless since, in that case, the net effect of a reinvestment is zero or negative. Clearly  $\bar{C}$  and  $\bar{D}$  are admissible whenever  $C$  and  $D$  are. From (1.1) and (1.3) we easily get

$$(1.4) \quad \begin{aligned} \Delta C_t &= \frac{\Delta Y_t + c_0}{1 - c_1} 1_{\{\Delta C_t > 0\}}, \quad 0 \leq c_1 < 1, \\ \Delta D_t &= -\frac{\Delta Y_t + d_0}{1 + d_1} 1_{\{\Delta D_t > 0\}}. \end{aligned}$$

Richard [10], Constantinides and Richard [3], and Harrison, Sellke, and Taylor [4] considered optimality of a storage system where, in addition to linear costs connected with increasing or decreasing the storage, there is also a holding cost. In all these papers the underlying process  $X$  is a Brownian motion with drift; i.e.,  $\mu$  and  $\sigma$  are constants. When holding costs are proportional, it was shown in [4] that for the linear Brownian motion this storage problem is equivalent to the dividend and reinvestment problem studied here. For more general diffusions, this equivalence no longer holds. Although the object of [4] was the storage problem, using this equivalence, those authors studied the dividend and reinvestment problem instead. They showed that the optimal policy is such that when assets reach a barrier  $y^*$ , reduce them to  $y^* - \delta^*$ , and when they are at zero, increase them to  $\gamma^*$ , where  $y^*$ ,  $\delta^*$ , and  $\gamma^*$  are uniquely given. This is an impulse control problem since payments are in lumps. We extend these results to the more general diffusion model given in section 2 and review the Brownian motion with drift in Example 2.1. This process has the advantage that an optimal solution always exists, but we shall see that this is not always the case for more general processes.

In Porteus [9] the optimal dividend and reinvestment problem is analyzed in a discrete time setting. This paper is also interesting as a general background and motivation of the problem.

The case with  $c_0 = d_0 = 0$  has been studied in several papers. Under the assumptions given in section 2, the problem was solved in Shreve, Lehoczky, and Gaver [12]. In this case both the optimal dividend process and the optimal reinvestment process, if they exist, are singular processes. Under somewhat different assumptions on the diffusion coefficients, the problem was solved in [11]. Avram, Palmowski, and Pistorius [1] solved the same problem, but with  $X$  as a spectrally negative Lévy process. All these papers assume that  $d_1 = 0$ , but letting  $d_1 > 0$  causes no extra problem.

Most of the attention in the optimal dividend literature has been on the case where activity stops whenever  $Y$  reaches zero; i.e., the task is to maximize

$$(1.5) \quad V_0^*(y) = \sup_{\mathcal{D}_0} E^y \left[ \int_{0-}^{\tau_y} e^{-rt} dD_t \right],$$

where  $\tau_y = \inf\{t : Y_t < 0\}$  with  $\tau_y = \infty$  if  $Y_t \geq 0$  for all  $t$  is the time of ruin. Here  $\mathcal{D}_0$  is the set of all admissible dividend policies (there are no reinvestments in this case). When  $d_0 > 0$ , this problem was first studied by Jeanblanc-Picqué and Shiryaev [5] for a Brownian motion with drift. Cadenillas, Sarkar, and Zapatero [2] worked with a different model, and also a more general utility function for payouts. Under the assumptions of the present paper, a complete solution to this problem is given in Paulsen [8], where it is shown that the optimal dividend process, if it exists, is a jump process, where an amount  $\delta_0^*$  is paid in dividends when the capital reaches a barrier  $y_0^*$ . Under the same assumptions, but with  $d_0 = 0$ , this problem was solved in [12] (at even some higher level of generality).

It may well be that when  $Y$  reaches zero the owners can choose between investing new money or terminating the business. In this case we can calculate  $V^*(y)$  and  $V_0^*(y)$ , and the one that is largest determines whether termination or reinvestment is optimal.

## 2. Results. We start with a list of assumptions.

- A1.  $|\mu(y)| + |\sigma(y)| \leq K(1 + y)$  for all  $y \geq 0$  and some  $K > 0$ .
- A2.  $\mu$  and  $\sigma$  are continuously differentiable and the derivatives  $\mu'$  and  $\sigma'$  are Lipschitz continuous for all  $y \geq 0$ .
- A3.  $\sigma^2(y) > 0$  for all  $y \geq 0$ .

A4.  $\mu'(y) \leq r$  for all  $y \geq 0$ .

The operator  $L$  is defined as

$$Lf(y) = \frac{1}{2}\sigma^2(y)f''(y) + \mu(y)f'(y) - rf(y)$$

for  $f \in C^2(0, \infty)$ . Note that under A2 and A3, any solution of  $Lf(y) = 0$  has  $f \in C^3[0, \infty)$  and  $f'''(y)$  is Lipschitz continuous (see, e.g., Krylov [7, Theorem 6.5.3]).

*Remark 2.1.* Assumption A4 may seem a bit unnatural and restrictive. To give an intuitive idea of what it means, we consider the special case

$$dX_t = (\mu_0 + \mu_1 X_t)dt + \sigma(X_t)dW_t, \quad X_0 = x.$$

Here  $\mu'(x) = \mu_1$ , and furthermore,

$$E^x[e^{-rt}X_t] = \left(x + \frac{\mu_0}{\mu_1}\right)e^{(\mu_1-r)t} - \frac{\mu_0}{\mu_1}e^{-rt}.$$

If  $\mu_1 \leq r$ , then this stabilizes, but if  $\mu_1 > r$ , it grows to infinity and this indicates that it is better to wait. The right quantities to compare are therefore  $\mu'(x)$  and  $r$ , one representing the geometric growth rate and the other the geometric discounting rate. The condition  $\mu'(x) \leq r$  just says that in no state should growth rate exceed discounting rate. Relaxing condition A4 by allowing for  $\mu'(x) > r$  for  $x$  in some subset of the positive real line is clearly possible, but it does complicate matters and is therefore a topic for future research.  $\square$

Before we continue let us briefly recapitulate the general solution when  $d_0 = 0$ , given in [12, Theorem 4.4]. According to this result we should look for a solution of  $LV(y) = 0$ ,  $y > 0$ , and  $y^*$  which satisfy

$$(2.1) \quad V'(0) = \frac{1}{1-c_1}, \quad V'(y^*) = \frac{1}{1+d_1}, \quad \text{and} \quad V''(y^*) = 0.$$

On  $y > y^*$ , set

$$V(y) = V(y^*) + \frac{y - y^*}{1 + d_1}.$$

The optimal solution is singular control at barrier  $y^*$ , and  $V = V^*$  is the value function.

Furthermore, if (2.1) has no solution, then there is no optimal control, but the value function is the limit of singular controls at barrier  $\bar{y}$  for increasing  $\bar{y}$ .

The next definition is connected with (1.4).

**DEFINITION 2.1.** *A lump sum reinvestment strategy with jump-size  $\gamma$  satisfies*

$$\Delta C_t = \frac{\gamma + c_0}{1 - c_1} 1_{\{Y_{t-}=0\}}.$$

*A lump sum dividend barrier strategy at  $\bar{y}$  with jump-size  $\delta \in (0, \bar{y}]$  satisfies for  $Y_{0-} = y$ ,*

$$\begin{aligned} \Delta D_0 &= \frac{y - (\bar{y} - \delta) - d_0}{1 + d_1} 1_{\{y \geq \bar{y}\}}, \\ \Delta D_t &= \frac{\delta - d_0}{1 + d_1} 1_{\{Y_{t-} = \bar{y}\}}. \end{aligned}$$



The corresponding value function is denoted by  $V_{\bar{y},\gamma,\delta}(y)$ . Also we set  $V_{\bar{y},\gamma(\bar{y}),\delta(\bar{y})}(y) = \sup_{\gamma,\delta \in (0,\bar{y})} V_{\bar{y},\gamma,\delta}(y)$ .

PROPOSITION 2.2. Assume A1–A4 and that the process  $Y$  is a strong solution of (1.1) with  $\bar{C}$  and  $\bar{D}$  as in (1.3). Assume that  $0 < c_0 < 1$  and  $d_0 > 0$ , and for fixed  $\bar{y}$ , consider the variational problem for unknown  $V$ ,  $\bar{\gamma} \in (0, \bar{y})$ , and  $\bar{\delta} \in (0, \bar{y})$ ,

$$\begin{aligned}
 (2.2) \quad & LV(y) = 0, \quad 0 < y < \bar{y}, \\
 & V(\bar{\gamma}) = V(0) + \frac{\bar{\gamma} + c_0}{1 - c_1}, \\
 & V'(\bar{\gamma}) = \frac{1}{1 - c_1}, \\
 & V(\bar{y}) = V(\bar{y} - \bar{\delta}) + \frac{\bar{\delta} - d_0}{1 + d_1}, \\
 & V'(\bar{y} - \bar{\delta}) = \frac{1}{1 + d_1}, \\
 & V(y) = V(\bar{y}) + \frac{y - \bar{y}}{1 + d_1}, \quad y > \bar{y}.
 \end{aligned}$$

This problem has a unique solution with  $0 < \bar{\gamma} \leq \bar{y} - \bar{\delta} < \bar{y}$ . Furthermore  $V(y) = V_{\bar{y},\gamma(\bar{y}),\delta(\bar{y})}(y)$  for all  $y \geq 0$ , and so  $\bar{\gamma} = \gamma(\bar{y})$  and  $\bar{\delta} = \delta(\bar{y})$ .

Here the equation  $V(\bar{\gamma}) = V(0) + (\bar{\gamma} + c_0)/(1 - c_1)$  is just an accounting equality; the value after raising the capital from 0 to  $\bar{\gamma}$  equals the value when capital is 0 plus the capital needed, including costs, for the raise. The equation  $V'(\bar{\gamma}) = 1/(1 - c_1)$  says that at  $\bar{\gamma}$  the marginal increase of the value of the company by raising capital by one unit equals the marginal cost of that raise. Hence when capital is raised to  $\bar{\gamma}$ , the owners are indifferent to a further marginal raise. The next two equations have a similar interpretation, while the last equation in (2.2) is only a definition of  $V(y)$  for  $y > \bar{y}$ .

From an intuitive and economic point of view, Proposition 2.2 is very reasonable. Since termination is not an option, for any given  $\bar{y}$  there must necessarily be some strategy that is better than, or not as bad as, any other. A mathematical (and rather long and technical) proof is given in the appendix.

Proposition 2.2 gives optimality for a barrier strategy at a fixed barrier  $\bar{y}$ . The general optimality problem is more complicated, but here is the solution.

THEOREM 2.3. Assume A1–A4 and that the process  $Y$  is a strong solution of (1.1) with  $\bar{C}$  and  $\bar{D}$  as in (1.3). Assume that  $0 < c_0 < 1$  and  $d_0 > 0$ , and consider the following variational problem for unknown  $V$ ,  $y^*$ ,  $\gamma^* \in (0, y^*)$ , and  $\delta^* \in (0, y^*)$ :

$$\begin{aligned}
 (2.3) \quad & LV(y) = 0, \quad 0 < y < y^*, \\
 & V(\gamma^*) = V(0) + \frac{\gamma^* + c_0}{1 - c_1}, \\
 & V'(\gamma^*) = \frac{1}{1 - c_1}, \\
 & V(y^*) = V(y^* - \delta^*) + \frac{\delta^* - d_0}{1 + d_1}, \\
 & V'(y^* - \delta^*) = \frac{1}{1 + d_1}, \\
 & V'(y^*) = \frac{1}{1 + d_1},
 \end{aligned}$$

$$V(y) = V(y^*) + \frac{y - y^*}{1 + d_1}, \quad y > y^*.$$

(a) If (2.3) has a solution, then this solution is unique and

$$V(y) = V_{y^*, \gamma(y^*), \delta(y^*)}(y) = V^*(y), \quad y \geq 0.$$

Therefore  $\gamma^* = \gamma(y^*)$  and  $\delta^* = \delta(y^*)$ .

(b) If (2.3) has no solution, then there is no optimal policy, but

$$V^*(y) = \lim_{\bar{y} \rightarrow \infty} V_{\bar{y}, \gamma(\bar{y}), \delta(\bar{y})}(y)$$

and this limit exists and is finite for every  $y \geq 0$ .

The proof is given in the appendix.

Note that (2.3) is essentially the same as (2.2) with the addition of the “smooth fit” condition at  $y^*$ ,  $V'(y^*) = (1 + d_1)^{-1}$ . This is necessary for finding the optimal  $y^*$ , if it exists. The economic interpretation is that at  $y^*$  the marginal value of keeping money in the company equals the marginal value of paying money out as dividends.

Theorem 2.3 shows that if an optimal strategy exists, then it is a lump sum dividend barrier strategy at some  $y^*$  with jump-size  $\delta^* = \delta(y^*)$  in conjunction with a lump sum reinvestment strategy with jump-size  $\gamma^* = \gamma(y^*)$ . If no optimal strategy exists, then the value function is finite and can be approximated by barrier strategies of the form described in Proposition 2.2 for increasing barriers  $\bar{y}$ .

*Remark 2.2.* A natural question is whether there is an  $x^* > 0$  so that, instead of waiting until assets reach zero, it would be better to reinvest at  $x^*$ . An intuitive economic argument that this is not the case can be drawn directly from assumption A4, since the growth rate of assets is never higher than the discounting rate, so there is no reason to reinvest money before it is necessary. Therefore, letting reinvestments take place only when assets hit zero represents no loss of generality. A more mathematical argument is that if there is such an  $x^*$ , then the equation

$$V(\gamma^*) = V(0) + \frac{\gamma^* + c_0}{1 - c_1}$$

in (2.3) must be replaced by the two equations

$$\begin{aligned} V(x^* + \gamma^*) &= V(x^*) + \frac{\gamma^* + c_0}{1 - c_1}, \\ V'(x^*) &= \frac{1}{1 - c_1}. \end{aligned}$$

But then a solution of this new system would have (at least) two distinct points  $y_{c,i}$ ,  $i = 1, 2$ , where  $V''(y_{c,i}) = 0$ , which is impossible by Lemma A.1.

The following result is useful for deciding whether (2.3) has a solution. Note that, except for the rather mild extra condition in (b), parts (a) and (b) yield an equivalence relation.

**PROPOSITION 2.4.** *Given the same assumptions as in Theorem 2.3, the following hold:*

(a) *Assume that (2.3) has no solution. Then there exists a solution  $g_2$  of  $Lg = 0$  so that*

$$\lim_{y \rightarrow \infty} g_2(y) = \lim_{y \rightarrow \infty} g'_2(y) = 0.$$

Furthermore, for any other independent solution  $g_1$ ,

$$\lim_{y \rightarrow \infty} g'_1(y) = \lim_{y \rightarrow \infty} \frac{g_1(y)}{y} = g'_1$$

for some positive and finite  $g'_1$ .

(b) Assume that there are two solutions  $g_1$  and  $g_2$  of  $Lg = 0$  so that

$$\begin{aligned} \lim_{y \rightarrow \infty} g'_1(y) &= g'_1, \\ \lim_{y \rightarrow \infty} g_2(y) &= 0, \end{aligned}$$

where  $g'_1$  is finite and nonzero. Assume in addition that

$$\lim_{y \rightarrow \infty} \left( \frac{g_1(y)}{g'_1} - y \right) > \frac{\mu(0)}{r} - d_0.$$

Then (2.3) has no solution.

(c) Assume there is a solution  $g$  of  $Lg = 0$  so that

$$\lim_{y \rightarrow \infty} \frac{g(y)}{y} = \infty$$

or equivalently,

$$\lim_{y \rightarrow \infty} g'(y) = \infty.$$

Then (2.3) has a solution.

*Example 2.1.* Let the income process without dividends follow

$$dX_t = \mu dt + \sigma dW_t,$$

with costs as in (1.3). This is a linear Brownian motion, and it is easy to verify that  $Lg(y) = 0$  has the independent solutions  $g_i(y) = e^{\theta_i y}$ ,  $i = 1, 2$ , where

$$\theta_1 = \frac{1}{\sigma^2} \left( \sqrt{\mu^2 + 2r\sigma^2} - \mu \right) \quad \text{and} \quad \theta_2 = -\frac{1}{\sigma^2} \left( \sqrt{\mu^2 + 2r\sigma^2} + \mu \right).$$

Clearly  $\theta_1 > 0$ , and hence an optimal solution exists by Proposition 2.4(c). This is the main result of [4]. To find a numerical solution, the general method of Example 3.1 can be used.

Here is another result that, in conjunction with Proposition 2.4, may be helpful in deciding whether (2.3) has a solution. A proof can be found in [8].

LEMMA 2.5. Assume A2 and A3 and let  $f_i(y)$ ,  $i = 1, 2$ , solve

$$\frac{1}{2}\sigma^2(y)f''_i(y) + \mu_i(y)f'_i(y) - rf_i(y) = 0, \quad y \geq 0,$$

where  $\mu_1(y) > \mu_2(y)$  for all  $y \geq 0$  and

$$f_i(0) = f_0 \quad \text{and} \quad f'_i(0) = f_1 \geq 0, \quad i = 1, 2.$$

Then  $f'_1(y) < f'_2(y)$  for all  $y > 0$ , which in turn implies that  $f_1(y) < f_2(y)$  for all  $y > 0$ .

The following result can be used to find the value function when (2.3) has no solution.

PROPOSITION 2.6. *Given the same assumptions as in Theorem 2.3, assume that (2.3) has no solution, and let  $V$  be the value function. Consider the set of equations (in  $\bar{\gamma}$ )*

$$(2.4) \quad \begin{aligned} V'(\bar{\gamma}) &= \frac{1}{1-c_1}, \\ V(\bar{\gamma}) &= V(0) + \frac{\bar{\gamma} + c_0}{1-c_1}. \end{aligned}$$

Furthermore, with  $g_1$  and  $g_2$  as in Proposition 2.4(a), write

$$V(y) = a_1 g_1(y) + a_2 g_2(y).$$

(a) We have

$$\lim_{y \rightarrow \infty} V'(y) = \frac{1}{1+d_1}.$$

(b) If  $c_1 + d_1 > 0$ , then (2.4) has a unique solution. Furthermore

$$\begin{aligned} a_1 &= \frac{1}{1+d_1} \frac{1}{g'_1}, \\ a_2 &= \frac{1}{1-c_1} \frac{1}{g'_2(\bar{\gamma})} - \frac{1}{1+d_1} \frac{1}{g'_1} \frac{g'_1(\bar{\gamma})}{g'_2(\bar{\gamma})}. \end{aligned}$$

Here  $g'_1 = \lim_{y \rightarrow \infty} g'_1(y)$  and  $\bar{\gamma}$  is the solution of

$$\frac{1-c_1}{1+d_1} \frac{1}{g'_1} (g_1(y) - g_1(0)) + \left( \frac{1}{g'_2(y)} - \frac{1-c_1}{1+d_1} \frac{1}{g'_1} \frac{g'_1(y)}{g'_2(y)} \right) (g_2(y) - g_2(0)) - y = c_0.$$

(c) If  $c_1 = d_1 = 0$  there are two possibilities as follows:

- (i) The equations in (2.4) have a unique solution, and then  $a_1$ ,  $a_2$ , and  $\bar{\gamma}$  are in (b) above. Also, denoting the barriers in (2.2) as  $\bar{y}_n$  and the corresponding reinvestment levels as  $\gamma_n$ , we then have  $\gamma_n \rightarrow \bar{\gamma}$  as  $\bar{y}_n \rightarrow \infty$ .
- (ii) The equations in (2.4) have no solution, but

$$\begin{aligned} a_1 &= \frac{1}{g'_1}, \\ a_2 &= \frac{\lim_{y \rightarrow \infty} \left( \frac{g_1(y)}{g'_1} - y \right) - \frac{g'_1(0)}{g'_1} - c_0}{g_2(0)}. \end{aligned}$$

**3. A financial example.** Assume that the income process without dividends is a linear Brownian motion with drift  $\mu$  and diffusion  $\sigma$ , but that money can be invested in risk-free assets with return  $r$ . Assume in addition that there are investment costs incurred with intensity  $\alpha(Y_t)$ . The dynamics (1.1) then becomes

$$(3.1) \quad \begin{aligned} dY_t &= (\mu + (r - \alpha(Y_t))Y_t)dt + \sigma dW_t + (1 - c_1)\Delta C_t - (1 + d_1)\Delta D_t \\ &\quad - c_0 1_{\{\Delta C_t > 0\}} - d_0 1_{\{\Delta D_t > 0\}}, \quad Y_{0-} = y. \end{aligned}$$

For this model  $\mu(y) = \mu + (r - \alpha(y))y$ , and so  $\mu'(y) \leq r$  for all  $y \geq 0$  if and only if

$$(3.2) \quad \alpha(y) + \alpha'(y)y \geq 0, \quad y \geq 0.$$

The total cost of investment intensity is  $\alpha(Y_t)Y_t$ , and a reasonable assumption is that this consists of a fixed part  $\alpha_0$  and a part that is proportional to the amount invested,  $\alpha_1$ , i.e.,

$$(3.3) \quad \alpha(y) = \frac{\alpha_0}{y} + \alpha_1.$$

Clearly (3.2) is satisfied, and (3.1) becomes

$$(3.4) \quad \begin{aligned} dY_t &= (\mu_0 + (r - \alpha_1)Y_t)dt + \sigma dW_t + (1 - c_1)\Delta C_t - (1 + d_1)\Delta D_t \\ &\quad - c_0 1_{\{\Delta C_t > 0\}} - d_0 1_{\{\Delta D_t > 0\}}, \quad Y_{0-} = y, \end{aligned}$$

where  $\mu_0 = \mu - \alpha_0$ . We shall assume that  $\mu_0 > 0$  and  $0 \leq \alpha_1 < r$ . When  $\alpha_1 = r$ , this is the linear Brownian motion analyzed in Example 2.1.

We need to solve

$$(3.5) \quad Lg(y) = \frac{1}{2}\sigma^2 g''(y) + (\mu_0 + (r - \alpha_1)y)g'(y) - rg(y) = 0.$$

Substituting  $z = -k(y)$  and  $f(z) = g(y)$  with

$$k(y) = \frac{r - \alpha_1}{\sigma^2} \left( y + \frac{\mu_0}{r - \alpha_1} \right)^2$$

brings it into the confluent hypergeometric form

$$zf''(z) + \left( \frac{1}{2} - z \right) f'(z) + \frac{r}{2(r - \alpha_1)} f(z) = 0.$$

Using the forms  $y_3$  and  $y_5$  from [13, p. 5] gives

$$\begin{aligned} g_1(y) &= e^{-k(y)} F(a, b, k(y)), \\ g_2(y) &= e^{-k(y)} U(a, b, k(y)), \end{aligned}$$

where

$$a = \frac{1}{2} + \frac{r}{2(r - \alpha_1)}, \quad b = \frac{1}{2},$$

and  $F$  and  $U$  are the first and second forms of Kummer's solution.

Consider first the case with  $\alpha_1 = 0$ , i.e., investment costs are constant. Then  $a = 1$ , and instead of the  $g_1$  above we can use the simpler  $g_1(y) = ry + \mu_0$ . By [13, p. 60] asymptotically as  $y \rightarrow \infty$ ,

$$e^{-k(y)} U(a, b, k(y)) \sim e^{-k(y)} (k(y))^{-a},$$

and hence  $\lim_{y \rightarrow \infty} g_2(y) = 0$ . Furthermore,  $g'_1(y) = r$  and

$$\lim_{y \rightarrow \infty} \left( \frac{g_1(y)}{r} - y \right) = \frac{\mu_0}{r} = \frac{\mu(0)}{r}.$$

Therefore the conditions of Proposition 2.4(b) are satisfied, and consequently there is no optimal control. The value function is, by Proposition 2.6(b),

$$V^*(y) = \frac{1}{1+d_1} \left( y + \frac{\mu_0}{r} \right) - a_2 e^{-k(y)} U(a, b, k(y)).$$

Here the first term is the value if money could be reinvested without costs when reaching zero, and the second term is a reduction in this value due to reinvestment costs. For this reason,  $a_2 > 0$ . A straightforward differentiation gives that

$$V^{*''}(y) = a_2 e^{-k(y)} [(k''(y) - (k'(y))^2)(U(a, b, k(y)) - U'(a, b, k(y))) + (k'(y))^2(U'(a, b, k(y)) - U''(a, b, k(y)))].$$

Trivially for the  $m$ th derivative,  $\text{sgn}(U^{(m)}(a, b, k(y))) = (-1)^m$ , and for all  $y$  sufficiently large,  $(k'(y)^2) > k''(y)$ , and hence  $g''(y) < 0$  for all  $y$  sufficiently large. But then it follows from Lemma A.1(f) that  $V^*$  is concave. If  $c_1 + d_1 > 0$ ,  $a_2$  can be calculated as in Proposition 2.6(b), and we find that  $\bar{\gamma}$  solves

$$(y + c_0)(1 + d_1) - (c_1 + d_1) \frac{g_2(y) - g_2(0)}{g_2'(y)} = (1 - c_1)r,$$

and then

$$a_2 = \left( \frac{1}{1+d_1} - \frac{1}{1-c_1} \right) \frac{1}{g_2'(\bar{\gamma})}.$$

When  $c_1 = d_1 = 0$ , concavity brings us into Proposition 2.6(c), case (ii), and we get

$$V^*(y) = y + \frac{\mu_0}{r} - \frac{c_0}{U(a, b, k(0))} e^{-(k(y)-k(0))} U(a, b, k(y)).$$

Assume now that  $0 < \alpha_1 < r$ . By [13, p. 60], asymptotically as  $y \rightarrow \infty$ ,

$$(3.6) \quad e^{-k(y)} F(a, b, k(y)) \sim \frac{\Gamma(b)}{\Gamma(a)} \left( \frac{r - \alpha_1}{r} \right)^{\frac{r}{2(r-\alpha_1)}} \left( y + \frac{\mu_0}{r - \alpha_1} \right)^{\frac{r}{r-\alpha_1}}.$$

Therefore, Proposition 2.4(c) applies, and an optimal solution exists.

As a byproduct let us comment on a paper by Cadenillas, Sarkar, and Zapatero [2]. In their paper the income process without dividends follows the mean reverting cash reservoir model

$$dX_t = (\alpha(\rho - X_t) - \beta)dt + \sigma dW_t, \quad \alpha, \beta, \rho \geq 0,$$

with costs as in (1.3). Although in the above it was assumed that  $\mu_0 > 0$ , an easy application of Lemma 2.5, as in the proof of Proposition 3.1 below, shows that an optimal solution exists for this problem as well.

When  $\alpha(y)$  has a more complex structure, there will be no analytical solution of  $Lg(y) = 0$ , but a numerical solution can readily be found. This is outlined in Example 3.1, but before attempting to use this procedure it is useful to know if an optimal strategy exists. To this end, the following result may be helpful.

**PROPOSITION 3.1.** *In addition to the assumptions in Theorem 2.3 let  $\alpha(y)$  satisfy (3.2). Furthermore, assume that  $\alpha(y) \geq \alpha_1$  for all  $y \geq 0$ , where  $\alpha_1 > 0$ . Then (2.3) has a solution.*

TABLE 1

Values of  $y^*$ ,  $\gamma^*$ ,  $y^* - \delta^*$  and values of  $V^*(y)$  for various  $y$  as a function of  $c_0$  when  $\sigma^2 = \mu_0 = 1$ ,  $d_0 = 0.1$ ,  $c_1 = d_1 = 0.05$ ,  $r = 0.1$ , and  $\alpha_1 = 0.02$ .

$c_0$	0	0.1	0.5	1	3	5	7.763	10
$y^*$	4.50	5.14	5.63	5.89	6.33	6.54	6.73	6.84
$\gamma^*$	0	0.61	1.06	1.31	1.72	1.92	3.10	2.20
$y^* - \delta^*$	0.47	1.06	1.51	1.75	2.15	2.35	2.52	2.62
$V^*(0)$	8.81	8.52	7.96	7.36	5.13	2.96	0	-2.39
$V^*(0.2)$	9.02	8.80	8.41	8.01	6.55	5.14	3.21	1.67
$V^*(1)$	9.77	9.66	9.54	9.44	9.15	8.90	8.56	8.29
$V^*(5)$	13.50	13.38	13.28	13.23	13.16	13.11	13.08	13.07

*Proof.* Let  $f_1(y)$  and  $f_2(y)$  satisfy  $L_i f_i(y) = 0$  together with  $f_i(0) = f_0$  and  $f'_i(0) = f_1 \geq 0$ , where

$$L_i f(y) = \frac{1}{2} \sigma^2(y) f''(y) + \mu_i(y) f'(y) - r f(y), \quad i = 1, 2,$$

and

$$\mu_1(y) = \mu_1 + (r - \alpha_1)y \quad \text{and} \quad \mu_2(y) = \mu + (r - \alpha(y))y,$$

where  $\mu_1$  is chosen large enough so that  $\mu_1(y) > \mu_2(y)$  for all  $y \geq 0$ . By Lemma 2.5,  $f_1(y) < f_2(y)$  for all  $y \geq 0$ , and by (3.6) asymptotically,

$$f_1(y) \sim C \left( y + \frac{\mu_1}{r - \alpha_1} \right)^{\frac{r}{r - \alpha_1}}$$

for some  $C > 0$ . Therefore, a solution of (2.3) exists by Proposition 2.4(c).  $\square$

*Example 3.1.* In this example,  $Y$  is given by (3.4) with  $\mu_0 = 1$ ,  $\sigma = 1$ ,  $c_1 = 0.05$ ,  $d_0 = 0.1$ ,  $d_1 = 0.05$ ,  $r = 0.1$ , and  $\alpha_1 = 0.02$ . Two independent solutions  $g_1$  and  $g_2$  with  $g_1(0) = 0$ ,  $g'_1(0) = 1$  and  $g_2(0) = 1$ ,  $g'_2(0) = 0$  were found using the Runge–Kutta method with stepsize  $h = 0.01$  and linear interpolation between the grid points. The function `fsolve` in MATLAB was used to solve (2.3) and this worked well, provided the initial values were not too far away from the true values. Some results are given in Table 1 for varying  $c_0$ . With  $c_0 = 7.763$ ,  $V^*(0) = 0$ , and then  $V^*(y) = V_0^*(y)$  due to uniqueness. Here  $V_0^*$  is the value function when there is absorption at zero; see (1.3). Hence  $c_0 = 7.763$  is the break-even point, where the owners are indifferent to investing new money or letting the company go bankrupt.

One notable observation from Table 1 is that  $V^*(5)$  decreases very little in  $c_0$ ; the reason for this must be that the probability of hitting zero when restarting at  $\eta^* = y^* - \delta^*$  must decrease rather rapidly with  $\eta^*$ .

The limiting case  $c_0 = 0$  is actually not covered here, but since this means that there is no lump sum cost at 0, the optimal policy at 0 is singular, and hence  $\gamma^* = 0$  in this case. In fact, the values when  $c_0 = 0$  were obtained by solving and

$$V'(0) = \frac{1}{1 - c_1}, \quad V(y^*) - V(y^* - \delta^*) = \frac{\delta^* - d_0}{1 + d_1}, \quad \text{and} \quad V'(y^* - \delta^*) = V'(y^*) = \frac{1}{1 + d_1}.$$

Table 2 is similar to Table 1, but with the difference that in Table 2  $d_0$  varies while  $c_0 = 0.1$  is fixed. Comparing the two tables, we see that the effects of varying  $c_0$  and  $d_0$  are very different. For example,  $V^*(0)$  is most sensitive to a decrease in  $c_0$ ,

TABLE 2

Values of  $y^*$ ,  $\gamma^*$ ,  $y^* - \delta^*$  and values of  $V^*(y)$  for various  $y$  as a function of  $d_0$  when  $\sigma^2 = \mu_0 = 1$ ,  $c_0 = 0.1$ ,  $c_1 = d_1 = 0.05$ ,  $r = 0.1$ , and  $\alpha_1 = 0.02$ .

$d_0$	0	0.1	0.5	1	3	5	10	20
$y^*$	1.94	5.14	10.22	14.83	29.28	41.80	70.53	124.46
$\gamma^*$	0.67	0.61	0.54	0.50	0.45	0.43	0.40	0.38
$y^* - \delta^*$	1.94	1.06	0.83	0.73	0.61	0.57	0.52	0.48
$V^*(0)$	8.95	8.52	7.93	7.53	6.67	6.19	5.48	4.73
$V^*(0.2)$	9.22	8.80	8.21	7.81	6.96	6.48	5.77	5.02
$V^*(1)$	10.10	9.66	9.05	8.64	7.75	7.26	6.51	5.74
$V^*(5)$	13.92	13.38	12.55	11.98	10.76	10.08	9.06	8.99

TABLE 3

Values of  $y^*$ ,  $\gamma^*$ ,  $y^* - \delta^*$  and values of  $V^*(y)$  for various  $y$  as a function of  $c_0 = d_0$  when  $\sigma^2 = \mu_0 = 1$ ,  $c_1 = d_1 = 0.05$ ,  $r = 0.1$ , and  $\alpha_1 = 0.02$ .

$c_0 = d_0$	0	0.1	0.5	1	3	5	5.416	10
$y^*$	1.30	5.14	10.75	15.68	30.84	43.80	46.71	73.28
$\gamma^*$	0	0.61	0.97	1.15	1.45	1.60	1.62	1.80
$y^* - \delta^*$	1.30	1.06	1.25	1.37	1.60	1.73	1.75	1.91
$V^*(0)$	9.24	8.52	7.37	6.35	3.22	0.53	0	-5.61
$V^*(0.2)$	9.45	8.80	7.83	7.03	4.70	2.80	2.43	-1.42
$V^*(1)$	10.22	9.66	8.95	8.45	7.32	6.59	6.46	5.28
$V^*(5)$	14.04	13.38	12.47	11.88	10.66	10.00	9.88	8.99

while the opposite is the case with  $V^*(5)$ . Here the case with  $d_0 = 0$  gives singular control at  $y^*$ , and the values when  $d_0 = 0$  were found by solving

$$V(\gamma^*) - V(0) = \frac{\gamma^* + c_0}{1 - c_0}, \quad V'(\gamma^*) = \frac{1}{1 - c_1}, \quad V'(y^*) = \frac{1}{1 + d_1}, \quad \text{and} \quad V''(y^*) = 0.$$

In Table 3 we let  $c_0 = d_0$  increase, while the other parameters are as before. Comparing with Tables 1 and 2, we see that for small  $y$  the values of  $V^*(y)$  are much like those in Table 1, while for large  $y$  they are more like those in Table 2. This is not unreasonable; for small  $y$  there is a big chance that the process will hit zero, and so an investment has to be made, while for big  $y$  the main cost is with dividend payments. When  $c_0 = d_0 = 0$  the control is singular, and this case is found by solving

$$V^*(0) = \frac{1}{1 - c_1}, \quad V'(y^*) = \frac{1}{1 + d_1}, \quad \text{and} \quad V''(y^*) = 0.$$

**Appendix. Proof of results in section 2.** The following result will be useful throughout.

LEMMA A.1. *Let  $\mu$  and  $\sigma$  satisfy A2–A4 and let  $f$  be a solution of  $Lf(y) = 0$ . On the interval  $[0, \infty)$ , the following hold:*

- (a) *If  $f$  has a zero, then  $f'$  has no zero.*
- (b) *If  $f'(\tilde{y}) = 0$  for some  $\tilde{y}$ , then  $(y - \tilde{y})f(y)f'(y) > 0$  for  $y \neq \tilde{y}$ .*
- (c) *If for some  $\tilde{y}$ ,  $f'(\tilde{y}) > 0$  and  $f''(\tilde{y}) \leq 0$ , then  $f$  is a concave function on  $[0, \tilde{y}]$ .*
- (d) *If  $f'$  has a zero, then  $f''$  has no zero.*
- (e) *If  $f''(\tilde{y}) = 0$  for some  $\tilde{y}$ , then  $(y - \tilde{y})f'(y)f''(y) > 0$  for  $y \neq \tilde{y}$ .*
- (f) *Assume that for some  $\tilde{y}$ ,  $f'(\tilde{y}) > 0$ . Then there are three possibilities as follows:*
  - (i)  *$f'$  is positive for all  $y$  and there is a  $y_c \in [0, \infty]$  so that  $f$  is concave on  $[0, y_c)$  (if  $y_c > 0$ ) and convex on  $(y_c, \infty)$  (if  $y_c < \infty$ ).*



- (ii)  $f$  is negative and concave for all  $y$  and there is a  $y_m \in [0, \infty)$  so that  $f'(y) > 0$  on  $(0, y_m)$  and  $f'(y) < 0$  on  $(y_m, \infty)$ .
- (iii)  $f$  is positive and convex for all  $y$  and there is a  $y_m \in [0, \infty)$  so that  $f'(y) < 0$  on  $(0, y_m)$  and  $f'(y) > 0$  on  $(y_m, \infty)$ .

*Proof.* Parts (a)–(c) are from [12, Lemma 4.2]. They consider bounded intervals, but there is no difference if we use the interval  $[0, \infty)$ . In the proof of their Lemma 4.2, Shreve, Lehoczky, and Gaver show that  $L_1 f' = 0$ , where the operator  $L_1$  satisfies the conditions for (a) and (b). Parts (d) and (e) are therefore (a) and (b) used on  $f'$  instead on  $f$ .

Now to (f). If  $f'$  is positive for all  $y$ , let  $y_c = \sup\{y : f''(y) \leq 0\}$ . Then  $f$  is concave on  $(0, y_c)$  and convex on  $(y_c, \infty)$  by (c). In particular if  $f(y_0) = 0$  for some  $y_0$ , then  $f'$  is positive by (a). For case (ii) assume that  $f$  is negative for all  $y$  and that  $f'(y_0) = 0$  for some  $y_0$ . By (b),  $(y - y_0)f'(y) < 0$  for  $y \neq y_0$ , and hence  $f'(y) > 0$  on  $(0, y_0)$  and  $f'(y) < 0$  on  $(y_0, \infty)$ . By (d),  $f''$  has no zero, and hence  $f''$  must be negative by what was just proved, i.e.,  $f$  is concave. Case (iii) follows from (ii) applied on  $-f$ .  $\square$

LEMMA A.2. Assume A2 and A3 and let  $g_1$  and  $g_2$  be two independent solutions of  $Lg = 0$  so that  $g'_2(y) \neq 0$  on  $[0, \bar{y}]$ . Let

$$v(y) = g'_1(y)g''_2(y) - g''_1(y)g'_2(y)$$

and

$$U(x, y) = g'_1(x)g'_2(y) - g'_1(y)g'_2(x).$$

Then

- (a)  $v(y) \neq 0$  on  $[0, \bar{y}]$ .
- (b)  $U(x, x) = 0$  on  $[0, \bar{y}]$  and  $U(x, y) \neq 0$  on  $0 \leq x < y \leq \bar{y}$ .

*Proof.* For (a), assume that  $v(y_0) = 0$  so that

$$\frac{g''_1(y_0)}{g'_1(y_0)} = \frac{g''_2(y_0)}{g'_2(y_0)}.$$

Multiplying both sides by  $\frac{1}{2}\sigma^2(y_0)$  and using that  $Lg_i = 0$  gives that this is equivalent to

$$\frac{g_1(y_0)}{g'_1(y_0)} = \frac{g_2(y_0)}{g'_2(y_0)},$$

or equivalently, that the Wronskian

$$W(g_1, g_2)(y_0) = g_1(y_0)g'_2(y_0) - g_2(y_0)g'_1(y_0) = 0.$$

But it is well known that the Wronskian of two independent solutions never vanishes, and hence we have arrived at a contradiction.

For (b), note that  $U(x_0, y_0) = 0$  is equivalent to  $h(x_0) = h(y_0)$ , where

$$h(x) = \frac{g'_1(x)}{g'_2(x)}.$$

But to obtain this result, we must have that  $h'(z) = 0$  for some  $z \in (x_0, y_0)$ . But

$$h'(z) = \frac{v(z)}{(g'_2(z))^2},$$

and this never vanishes according to (a). This proves the result.  $\square$

LEMMA A.3. Assume A2–A4 and let  $\bar{y} > 0$  be given. Then there exists two independent solutions  $g_1$  and  $g_2$  of  $Lg = 0$  so that

(i)  $g_1(0) = g_1(\bar{y}) = 1$ ,  $g'_1(0) < 0$ ,  $g'_1(\bar{y}) > 0$ , and  $g''_1(y) > 0$  on  $[0, \bar{y}]$ . Also  $g(y) > 0$  on  $[0, \bar{y}]$ .

(ii)  $g_2(0) = 1$ ,  $g'_2(0) = 0$ ,  $g'_2(y) > 0$  on  $(0, \bar{y}]$ , and  $g''_2(y) > 0$  on  $[0, \bar{y}]$ .

*Proof.* Clearly  $g_2$  is defined by the initial conditions, and by Lemma A.1(b),  $yg_2(y)g'_2(y) > 0$  for  $y > 0$ ; hence  $g'_2(y) > 0$  for  $y \in (0, \varepsilon)$  for some  $\varepsilon > 0$ . But by Lemma A.1(d),  $g''_2(y) \neq 0$ , and since  $g'_2(y) > g'_2(0)$  on  $(0, \varepsilon)$ , it follows that  $g''_2(y) > 0$ . But then  $g'_2(y) > 0$  for  $y > 0$  as well.

Next define  $g_0$  by  $Lg_0 = 0$ ,  $g_0(0) = 0$ , and  $g'_0(0) = 1$ , which implies by Lemma A.1(a) that  $g'_0(y) > 0$  for all  $y \geq 0$ . Let

$$g_1(y) = \frac{1 - g_2(\bar{y})}{g_0(\bar{y})}g_0(y) + g_2(y).$$

Clearly  $g_1(0) = g_1(\bar{y}) = 1$ . Also  $g'_1(0) = (1 - g_2(\bar{y}))/g_0(\bar{y}) < 0$ . There must exist a  $y_0 \in (0, \bar{y})$  so that  $g'_0(y_0) = 0$ , but this implies by Lemma A.1(d) that  $g''_0(y) \neq 0$ . Since  $g'_1(0) < 0$  it is necessary that  $g'_1(y) > 0$ . Finally, let  $\tau_0^y$  be the first time the process  $Y$ , subject to no control, hits the value 0 when  $Y_0 = y$ , and similarly let  $\tau_{\bar{y}}^y$  be the first time it hits  $\bar{y}$ . Also let  $T_{\bar{y}}^y = \min\{\tau_0^y, \tau_{\bar{y}}^y\}$ . By standard results in diffusion theory (see, e.g., [6, Chapter 15.3]),

$$g_1(y) = E^y \left[ e^{-rT_{\bar{y}}^y} \right],$$

and hence  $g_1(y) > 0$  on  $[0, \bar{y}]$ .  $\square$

LEMMA A.4. There exists a solution to the variational problem (2.2) with  $0 < \bar{\gamma} \leq \bar{y} - \bar{\delta} < \bar{y}$ .

*Proof.* Throughout the proof,  $g_1$  and  $g_2$  are as in Lemma A.3 so that a general solution to  $LV = 0$  is

$$V(y) = a_1g_1(y) + a_2g_2(y).$$

Assume first that  $c_1 + d_1 > 0$ . In order to satisfy  $V'(\gamma) = (1 - c_1)^{-1}$  and  $V'(\eta) = (1 + d_1)^{-1}$  (with  $\eta = \bar{y} - \delta$ ), a straightforward computation shows that

$$(A.1) \quad \begin{aligned} a_1 &= a_1(\gamma, \eta) = \frac{1}{U(\gamma, \eta)} \left( \frac{g'_2(\eta)}{1 - c_1} - \frac{g'_2(\gamma)}{1 + d_1} \right), \\ a_2 &= a_2(\gamma, \eta) = -\frac{1}{U(\gamma, \eta)} \left( \frac{g'_1(\eta)}{1 - c_1} - \frac{g'_1(\gamma)}{1 + d_1} \right), \end{aligned}$$

where  $U(\gamma, \eta) = g'_1(\gamma)g'_2(\eta) - g'_1(\eta)g'_2(\gamma)$ . Since  $U(0, \eta) = g'_1(0)g'_2(\eta) < 0$ , it follows by Lemma A.3(b) that  $U(\gamma, \eta) < 0$  when  $0 \leq \gamma < \eta \leq \bar{y}$ . Also, since both  $g_1$  and  $g_2$  are convex,  $a_1(\gamma, \eta) < 0$  and  $a_2(\gamma, \eta) > 0$ . Furthermore,  $a_1(\gamma, \eta) \rightarrow -\infty$  and  $a_2(\gamma, \eta) \rightarrow \infty$  as  $\eta - \gamma \rightarrow 0$ .

We will now prove that with  $a_1$  and  $a_2$  as in (A.1), the equation  $\alpha_\eta(\gamma) = 0$ , where

$$\begin{aligned} \alpha_\eta(\gamma) &= V(\gamma) - V(0) - \frac{\gamma + c_0}{1 - c_1} \\ &= a_1(\gamma, \eta)(g_1(\gamma) - 1) + a_2(\gamma, \eta)(g_2(\gamma) - 1) - \frac{\gamma + c_0}{1 - c_1}, \end{aligned}$$

has a solution  $\gamma < \eta$  for any  $\eta \in (0, \bar{y}]$ , and similarly that the equation  $\beta_\gamma(\eta) = 0$  with

$$\begin{aligned}\beta_\gamma(\eta) &= V(\bar{y}) - V(\eta) - \frac{\bar{y} - \eta - d_0}{1 + d_1} \\ &= a_1(\gamma, \eta)(g_1(\bar{y}) - g_1(\eta)) + a_2(\gamma, \eta)(g_2(\bar{y}) - g_2(\eta)) - \frac{\bar{y} - \eta - d_0}{1 + d_1}\end{aligned}$$

has a solution  $\eta > \gamma$  for any  $\gamma \in [0, \bar{y})$ . Making two simple graphs then shows that there is a point  $(\bar{\gamma}, \bar{\eta})$  so that  $\alpha_{\bar{\eta}}(\bar{\gamma}) = \beta_{\bar{\gamma}}(\bar{\eta}) = 0$ , and so existence is established.

To prove that  $\alpha_\eta(\gamma) = 0$  has a solution for some  $\gamma < \eta$ , note first that  $\alpha_\eta(0) = -c_0/(1 - c_1)$ . Furthermore, by the above asymptotic properties of  $a_i(\gamma, \eta)$ ,  $i = 1, 2$ ,  $\alpha_\eta(\gamma) \rightarrow \infty$  as  $\gamma \rightarrow \eta$ , and hence  $\alpha_\eta(\gamma) = 0$  has a solution.

The proof that  $\beta_\gamma(\eta) = 0$  for some  $\eta > \gamma$  is a bit trickier. Since  $\beta_\gamma(\bar{y}) = d_0/(1 + d_1)$ , it is sufficient to prove that  $\beta_\gamma(\eta) \rightarrow -\infty$  as  $\eta \rightarrow \gamma$ . Write

$$\beta_\gamma(\eta) = \frac{1}{U(\gamma, \eta)} k_\gamma(\eta) - \frac{\bar{y} - \eta - d_0}{1 + d_1},$$

where

$$k_\gamma(\eta) = \left( \frac{g'_2(\eta)}{1 - c_1} - \frac{g'_2(\gamma)}{1 + d_1} \right) (1 - g_1(\eta)) - \left( \frac{g'_1(\eta)}{1 - c_1} - \frac{g'_1(\gamma)}{1 + d_1} \right) (\bar{g}_2 - g_2(\eta))$$

and  $\bar{g}_2 = g_2(\bar{y})$ . Since  $1/U(\gamma, \eta) \rightarrow -\infty$  as  $\eta \rightarrow \gamma$ , it is sufficient to prove that  $k_\gamma(\gamma) > 0$ , or equivalently, since  $(1 - c_1)^{-1} - (1 + d_1)^{-1} > 0$ , that

$$(A.2) \quad h(\gamma) = g'_2(\gamma)(1 - g_1(\gamma)) - g'_1(\gamma)(\bar{g}_2 - g_2(\gamma)) > 0, \quad \gamma \in [0, \bar{y}).$$

Now  $h(0) = -g'_1(0)(\bar{g}_2 - 1) > 0$  and  $h(\bar{y}) = 0$ . Furthermore,

$$(A.3) \quad h'(\gamma) = g''_2(\gamma)(1 - g_1(\gamma)) - g''_1(\gamma)(\bar{g}_2 - g_2(\gamma))$$

so that  $h'(\bar{y}) = 0$ . But a Taylor expansion of  $g_i(\gamma)$  around  $\bar{y}$  and then of  $g'_i(\bar{y})$  around  $\gamma$  gives

$$h'(\gamma) = (\bar{y} - \gamma)v(\gamma) + o(\bar{y} - \gamma),$$

where  $v$  is as in Lemma A.2. Since  $v(0) = g''_2(0)g'_1(0) < 0$ , it follows from Lemma A.2(a) that  $v(y) < 0$  for  $y \in [0, \bar{y}]$ , and therefore  $h(y) > 0$  for  $y \in (\bar{y} - \varepsilon, \bar{y})$  for some  $\varepsilon > 0$ . As a consequence, if  $h(\gamma) = 0$  has a root in  $[0, \bar{y})$ , there will necessarily be at least two roots, and at two of these roots  $h'$  will have opposite signs. But  $h(\gamma) = 0$  implies that

$$\bar{g}_2 - g_2(\gamma) = \frac{g'_2(\gamma)}{g'_1(\gamma)}(1 - g_1(\gamma)),$$

so that in particular  $g'_1(\gamma) > 0$ , and furthermore using (A.3),

$$h'(\gamma) = \frac{1 - g_1(\gamma)}{g'_1(\gamma)} v(\gamma) < 0,$$

which is a contradiction. Hence (A.2) holds and this ends the proof for the case with  $c_1 + d_1 > 0$ .

Assume now that  $c_1 + d_1 = 0$ , which clearly implies that  $\bar{\gamma} = \bar{\eta}$ ; hence we must prove the existence of

$$(A.4) \quad V'(\gamma) = a_1 g_1'(\gamma) + a_2 g_2'(\gamma) = 1$$

as well as of

$$(A.5) \quad V(\gamma) - V(0) = a_1(g_1(\gamma) - 1) + a_2(g_2(\gamma) - 1) = \gamma + c_0$$

and

$$(A.6) \quad V(\bar{y}) - V(\gamma) = a_1(1 - g_1(\gamma)) + a_2(\bar{g}_2 - g_2(\gamma)) = \bar{y} - \gamma - d_0.$$

Here  $g_1$  and  $g_2$  are again as in Lemma A.3. Adding (A.5) and (A.6) gives

$$(A.7) \quad a_2 = \frac{\bar{y} + c_0 - d_0}{\bar{g}_2 - 1}.$$

It remains to prove that there is an  $a_1$  and a  $\gamma$  so that (A.4) and (A.5) are satisfied with  $a_2$  given by (A.7).

By the properties of  $g_1$ , there is a unique  $\gamma_0 \in (0, \bar{y})$  so that  $g_1'(\gamma_0) = 0$ . If  $a_2 g_2'(\gamma_0) = 1$ ,  $a_1$  can be chosen so that (A.5) is satisfied at the point  $\gamma_0$ , and we have a solution. If  $a_2 g_2'(\gamma_0) \neq 1$ , then by (A.4),

$$a_1 = a_1(\gamma) = \frac{1 - a_2 g_2'(\gamma)}{g_1'(\gamma)},$$

and inserting this into (A.5) gives the problem  $h(\gamma) = 0$ , where

$$h(\gamma) = \frac{g_1(\gamma) - 1}{g_1'(\gamma)} + a_2 \left( g_2(\gamma) - 1 - \frac{g_2'(\gamma)}{g_1'(\gamma)} (g_2(\gamma) - 1) \right) - \gamma - c_0.$$

It is easy to see that  $h(0) = -c_0$  and that  $h(\bar{y}) = -d_0$ . But  $h(\gamma) \rightarrow \infty$  as  $\gamma$  approaches  $\gamma_0$  from one side, while  $h(\gamma) \rightarrow -\infty$  as  $\gamma$  approaches  $\gamma_0$  from the other side. Therefore  $h(\gamma) = 0$  must have a solution, and we are done.  $\square$

LEMMA A.5. *The variational problem (2.2) has a unique solution with  $0 < \bar{\gamma} \leq \bar{y} - \bar{\delta} < \bar{y}$ .*

*Proof.* It remains only to prove uniqueness since existence was proved in Lemma A.4. We content ourselves with the case  $c_1 + d_1 > 0$ , since the case  $c_1 + d_1 = 0$  is easier. Assume there are two different sets of solutions  $V_i, \bar{\gamma}_i, \bar{\delta}_i$ ,  $i = 1, 2$ , and let  $\eta_i = \bar{y} - \bar{\delta}_i$ . For ease of notation, we replace  $\bar{\gamma}_i$  and  $\bar{\eta}_i$  by  $\gamma_i$  and  $\eta_i$ . Let  $V_0(y) = V_2(y) - V_1(y)$ , and assume without loss of generality that  $V_0(0) \geq 0$ .

Assume that  $V_0'(y) > 0$  for  $y \in (0, \bar{y})$ . The fact that  $V_1'(\gamma_1) = V_2'(\gamma_2)$  implies that  $\gamma_1 < \gamma_2$ . But then

$$\begin{aligned} \frac{c_0}{1 - c_1} &= \int_0^{\gamma_1} (V_1'(y) - (1 - c_1)^{-1}) dy \\ &< \int_0^{\gamma_1} (V_2'(y) - (1 - c_1)^{-1}) dy \\ &< \int_0^{\gamma_2} (V_2'(y) - (1 - c_1)^{-1}) dy = \frac{c_0}{1 - c_1}, \end{aligned}$$

which is a contradiction. The same argument can be used if  $V'_0(y) < 0$  for  $y \in (0, \bar{y})$ . So by Lemma A.1(a) we can conclude that  $V_0(y) > 0$  for  $y \in [0, \bar{y}]$  and that  $V'_0(y_0) = 0$  for some  $y_0 \in (0, \bar{y})$ . By Lemma A.1(b),

$$(A.8) \quad \begin{aligned} V'_2(y) &< V'_1(y), & y \in (0, y_0), \\ V'_2(y) &> V'_1(y), & y \in (y_0, \bar{y}). \end{aligned}$$

Assume now that  $\gamma_2 \leq y_0$ , which implies that  $\gamma_1 > \gamma_2$  by (A.8). Then

$$\begin{aligned} \frac{c_0}{1-c_1} &= \int_0^{\gamma_1} (V'_1(y) - (1-c_1)^{-1}) dy \\ &> \int_0^{\gamma_2} (V'_1(y) - (1-c_1)^{-1}) dy \\ &> \int_0^{\gamma_2} (V'_2(y) - (1-c_1)^{-1}) dy = \frac{c_0}{1-c_1}, \end{aligned}$$

which is a contradiction. Hence we must have that

$$(A.9) \quad y_0 < \gamma_1 < \gamma_2.$$

By (A.9),  $\eta_i > y_0$ , and hence by (A.8),

$$(A.10) \quad y_0 < \eta_1 < \eta_2.$$

But this gives

$$\begin{aligned} -\frac{d_0}{1+d_1} &= \int_{\eta_2}^{\bar{y}} (V'_2(y) - (1+d_1)^{-1}) dy \\ &> \int_{\eta_2}^{\bar{y}} (V'_1(y) - (1+d_1)^{-1}) dy \\ &> \int_{\eta_1}^{\bar{y}} (V'_1(y) - (1+d_1)^{-1}) dy = -\frac{d_0}{1+d_1}, \end{aligned}$$

which is a contradiction. In the last inequality we used that

$$\max_{y \in [\eta_1, \eta_2]} V'_1(y) = \max\{V'_1(\eta_1), V'_1(\eta_2)\} \leq \max\{V'_1(\eta_1), V'_2(\eta_2)\} = (1+d_1)^{-1}.$$

We have thus arrived at a contradiction, and so there must be uniqueness.  $\square$

LEMMA A.6. Assume that (2.2) holds. Then

$$V(y) - V(0) \leq \frac{y + c_0}{1 - c_1},$$

with equality if and only if  $y = \bar{\gamma}$ . Furthermore

$$V(\bar{y}) - V(y) \geq \frac{\bar{y} - y - d_0}{1 + d_1},$$

with equality if and only if  $y = \bar{y} - \bar{\delta}$ .

*Proof.* For simplicity we write  $\gamma = \bar{\gamma}$  and  $\delta = \bar{\delta}$ . Also set  $\eta = \bar{y} - \bar{\delta}$ . Now from  $V'(\eta) = (1 + d_1)^{-1}$  and

$$V(\bar{y}) - V(\eta) - \frac{\bar{y} - \eta}{1 + d_1} = \int_{\eta}^{\bar{y}} (V'(x) - (1 + d_1)^{-1}) dx = -\frac{d_0}{1 + d_1},$$

it follows that  $V'(y_0) < V'(\eta)$  for some  $y_0 > \eta$ , and hence  $V''(y_1) < 0$  for some  $y_1 > \eta$ . Therefore, by Lemma A.1(f), either  $V$  is concave on  $[0, \bar{y}]$  or there is a  $y_c$  with  $\eta < y_c < \bar{y}$  so that  $V$  is concave on  $[0, y_c)$  and convex on  $(y_c, \bar{y}]$ .

Now define

$$h(y) = V(y) - V(0) - \frac{y}{1 - c_1} = \int_0^y (V'(x) - (1 - c_1)^{-1}) dx.$$

Then  $h(0) = 0$ ,  $h(\gamma) = c_0/(1 - c_1)$ , and  $h'(\gamma) = 0$ . If  $V$  is concave on  $[0, \bar{y}]$ ,  $h$  will be increasing on  $[0, \gamma)$  and decreasing on  $(\gamma, \bar{y}]$ , and so  $h$  takes its maximum at  $\gamma$ . If  $y_c < \bar{y}$ , where  $y_c$  is as above, then since  $y_c > \eta$ ,  $h(\eta) \leq h(\gamma)$ , and so  $h(\bar{y}) - h(\eta) < 0$  will again imply that  $h$  takes its maximum at  $\gamma$ . But

$$h(\bar{y}) - h(\eta) = \frac{\bar{y} - \eta - d_0}{1 + d_1} - \frac{\bar{y} - \eta}{1 - c_1} < 0.$$

This gives the first part. The second part is proved similarly by defining

$$h(y) = V(\bar{y}) - V(y) - \frac{\bar{y} - y}{1 + d_1} = \int_y^{\bar{y}} (V'(x) - (1 + d_1)^{-1}) dx$$

and proceeding as above to prove that  $h$  takes a minimum at  $\eta$  with  $h(\eta) = -d_0/(1 + d_1)$ .  $\square$

*Proof of Proposition 2.2.* Using Lemmas A.5 and A.6, the proof follows along the same lines as the proof of Theorem 2.3(a), but is simpler. We therefore omit the details.  $\square$

With this result we can continue to the proof of Theorem 2.3, but first we give some lemmas.

LEMMA A.7. Assume that (2.3) has a solution. Then

- (a)  $LV(y) \leq 0$  for all  $y > 0$ .
- (b)  $\frac{y-x-d_0}{1+d_1} \leq V(y) - V(x) \leq \frac{y-x+c_0}{1-c_1}$ ,  $0 \leq x \leq y$ . There is left equality when  $y \geq y^*$ ,  $x = \delta^*$ , and right equality when  $y = \gamma^*$  and  $x = 0$ .
- (c)  $y^*$ ,  $\gamma^*$ ,  $\delta^*$ , and  $V$  are uniquely given.

*Proof.* Note first that since  $V'(y^*) = V'(y^* - \delta^*)$ , this is necessarily case (i) of Lemma A.1(f), with  $y^* - \delta^* < y_c < y^*$ . Therefore  $V''(y^* -) \geq 0$ , and so by continuity,

$$rV(y^*) = \frac{1}{2}\sigma^2(y^*)V''(y^* -) + \mu(y^*)V'(y^*) \geq \mu(y^*)(1 + d_1)^{-1}.$$

Using this inequality, we get for  $y > y^*$ ,

$$\begin{aligned} LV(y) &= \frac{\mu(y)}{1 + d_1} - r \left( V(y^*) + \frac{y - y^*}{1 + d_1} \right) \\ &= (1 + d_1)^{-1} (\mu(y) - \mu(y^*) - r(y - y^*)) + \mu(y^*)(1 + d_1)^{-1} - rV(y^*) \\ &\leq (1 + d_1)^{-1} \int_{y^*}^y (\mu'(x) - r) dx \leq 0. \end{aligned}$$

To prove (b), set  $\eta^* = y^* - \delta^*$ . From the observation at the beginning of the proof,  $V'(y) \geq (1 + d_1)^{-1}$  on  $(0, \eta^*)$  and  $V'(y) \leq (1 + d_1)^{-1}$  on  $(\eta^*, \infty)$ . Therefore

$$V(y) - V(x) - \frac{y - x}{1 + d_1} = \int_x^y (V'(u) - (1 + d_1)^{-1}) du$$

is smallest for  $x = \eta^*$  and  $y \geq y^*$ , and then it equals  $-d_0/(1 + d_1)$ . This gives the first inequality. For the second inequality, note that  $V'(y) \geq (1 - c_1)^{-1}$  on  $(0, \gamma^*)$  and  $V'(y) \leq (1 - c_1)^{-1}$  on  $(\gamma^*, \infty)$ . Therefore,

$$V(y) - V(x) - \frac{y - x}{1 - c_1} = \int_x^y (V'(u) - (1 - c_1)^{-1}) du$$

takes its maximum at  $x = 0$  and  $y = \gamma^*$ , and the maximum value is  $c_0/(1 - c_1)^{-1}$ .

For (c), let  $V_i, \gamma_i^*, \delta_i^*, y_i^*$ ,  $i = 1, 2$ , be two sets of solutions. Let  $V_0(y) = V_2(y) - V_1(y)$  and assume that  $V_0(0) \geq 0$ . The same arguments as in the proof of uniqueness in Lemma A.5 still hold up to and including (A.10). Furthermore, by (A.8) and the fact that  $V_1'(y_1^*) = V_2'(y_2^*)$ , it is necessary that  $y_2^* < y_1^*$ . This, together with (A.8) and (A.10), gives with  $\eta_i^* = y_i^* - \delta_i^*$ ,

$$\begin{aligned} -\frac{d_0}{1 + d_1} &= \int_{\eta_2^*}^{y_2^*} (V_2'(y) - (1 + d_1)^{-1}) dy \\ &> \int_{\eta_2^*}^{y_2^*} (V_1'(y) - (1 + d_1)^{-1}) dy \\ &> \int_{\eta_1^*}^{y_1^*} (V_1'(y) - (1 + d_1)^{-1}) dy = -\frac{d_0}{1 + d_1}, \end{aligned}$$

which is a contradiction. Hence uniqueness follows.  $\square$

LEMMA A.8. Assume that (2.3) has no solution. Denote a solution of (2.2) for given  $\bar{y}$  by  $V_{\bar{y}}(y)$ . Then  $V_{\bar{y}}'(\bar{y}-) < (1 + d_1)^{-1}$  for all  $\bar{y} > 0$ .

*Proof.* We denote  $\gamma = \bar{\gamma}$ ,  $\delta = \bar{\delta}$ , and  $\eta = \bar{y} - \bar{\delta}$ , and when there is no misunderstanding, sometimes we also write  $V$  for  $V_{\bar{y}}$ .

Let  $\bar{y} < d_0$ , giving  $V(\bar{y}) - V(\eta) = (\delta - d_0)/(1 + d_1) < 0$ . Therefore we have case (ii) of Lemma A.1(f), implying that  $V'(\bar{y}-) < 0$ . If we can prove that  $V_{\bar{y}}'(\bar{y}-)$  as a function of  $\bar{y}$  is continuous, then necessarily  $V_{\bar{y}}'(\bar{y}-) < (1 + d_1)^{-1}$  for all  $\bar{y} > 0$  since otherwise  $V_{\bar{y}_0}'(\bar{y}_0-) = (1 + d_1)^{-1}$  for some  $\bar{y}_0$ , in which case (2.3) would hold with  $y^* = \bar{y}_0$ .

To this end let  $g_1$  and  $g_2$  be two linearly independent solutions of  $Lg = 0$  with  $g_1(0) = 0$  and  $g_2(0) = 1$ . Also let the constants  $a_1$  and  $a_2$  be such that  $V_{\bar{y}}(y) = a_1 g_1(y) + a_2 g_2(y)$ ,  $y \leq \bar{y}$ . The second through the fifth equations of (2.2) then become

$$\begin{aligned} a_1 g_1(\gamma) + a_2 g_2(\gamma) - \frac{\gamma + c_0}{1 - c_1} &= 0, \\ a_1 g_1'(\gamma) + a_2 g_2'(\gamma) - \frac{1}{1 - c_1} &= 0, \\ a_1 (g_1(\bar{y}) - g_1(\eta)) + a_2 (g_2(\bar{y}) - g_2(\eta)) - \frac{\bar{y} - \eta - d_0}{1 + d_1} &= 0, \\ a_1 g_1'(\eta) + a_2 g_2'(\eta) - \frac{1}{1 + d_1} &= 0. \end{aligned}$$

This can be written as  $\mathbf{h}(\mathbf{x}, \bar{y}) = \mathbf{0}$  with  $\mathbf{x} = (a_1, a_2, \gamma, \eta)$ . Using these equations, the

Jacobian  $J = \frac{\partial \mathbf{h}}{\partial \mathbf{x}}$  becomes

$$J = \begin{bmatrix} g_1(\gamma) & g_2(\gamma) - 1 & 0 & 0 \\ g'_1(\gamma) & g'_2(\gamma) & V''(\gamma) & 0 \\ g_1(\bar{y}) - g_1(\eta) & g_2(\bar{y}) - g_2(\eta) & 0 & 0 \\ g'_1(\eta) & g'_2(\eta) & 0 & V''(\eta) \end{bmatrix}.$$

We will prove that the determinant  $|J| \neq 0$ . If  $V''(\eta) = 0$ , by Lemma A.1(f),  $V$  would be convex on  $[\eta, \bar{y}]$ , but then  $V(\bar{y}) - V(\eta) > (\bar{y} - \eta - d_0)/(1 + d_1)$ , which is a contradiction. Hence  $V''(\eta) < 0$  and then  $V''(\gamma) < 0$  as well. Therefore, for  $|J|$  to be zero, it is necessary that for nonzero constants  $b$  and  $c$ ,

$$\begin{aligned} bg_1(\gamma) + c(g_2(\gamma) - 1) &= 0, \\ b(g_1(\bar{y}) - g_1(\eta)) + c(g_2(\bar{y}) - g_2(\eta)) &= 0. \end{aligned}$$

Letting  $\tilde{V} = bg_1(y) + cg_2(y)$ , this gives

$$(A.11) \quad \tilde{V}(\gamma) = \tilde{V}(0) \quad \text{and} \quad \tilde{V}(\bar{y}) = \tilde{V}(\eta).$$

It follows that  $\tilde{V}$  must satisfy Lemma A.1(f), but according to this, (A.11) is impossible, and therefore  $|J| \neq 0$ . By the implicit function theorem,  $a_1 = a_1(\bar{y})$  and  $a_2 = a_2(\bar{y})$  are continuous functions of  $\bar{y}$ , and then so is  $V'_{\bar{y}}(\bar{y}-) = a_1(\bar{y})g'_1(\bar{y}) + a_2(\bar{y})g'_2(\bar{y})$ .  $\square$

LEMMA A.9. Assume that (2.3) has no solution, and let  $V_{\bar{y}_i}(y)$ ,  $i = 1, 2$ , be solutions of (2.2) for barriers  $\bar{y}_1 < \bar{y}_2$ . Then  $V_{\bar{y}_1}(y) < V_{\bar{y}_2}(y)$  for all  $y \geq 0$ . Furthermore,

$$V(y) = \lim_{\bar{y} \rightarrow \infty} V_{\bar{y}}(y)$$

is finite for all  $y \geq 0$ . In fact  $V$  is three times continuously differentiable with

$$V^{(i)}(y) = \lim_{\bar{y} \rightarrow \infty} V_{\bar{y}}^{(i)}(y), \quad i = 1, 2, 3,$$

so, in particular,

$$LV(y) = 0, \quad y > 0.$$

Finally,

$$\frac{y - x - d_0}{1 + d_1} < V(y) - V(x) \leq \frac{y - x + c_0}{1 - c_1}, \quad 0 \leq x \leq y.$$

*Proof.* For simplicity we write  $V_i = V_{\bar{y}_i}$ ,  $\gamma_i = \bar{\gamma}_i$ ,  $\delta_i = \bar{\delta}_i$ , and  $\eta_i = \bar{y}_i - \bar{\delta}_i$ .

Let  $V_0(y) = V_2(y) - V_1(y)$ . If  $V'_0(y) > 0$  on  $[0, \bar{y}_1]$ , i.e.,  $V'_2(y) > V'_1(y)$ , then  $\gamma_2 > \gamma_1$  and so

$$\begin{aligned} \frac{c_0}{1 - c_1} &= \int_0^{\gamma_1} (V'_1(y) - (1 - c_1)^{-1}) dy \\ &< \int_0^{\gamma_1} (V'_2(y) - (1 - c_1)^{-1}) dy \\ &< \int_0^{\gamma_2} (V'_2(y) - (1 - c_1)^{-1}) dy = \frac{c_0}{1 - c_1}, \end{aligned}$$



which is a contradiction. In a similar way it is easy to show that  $V'_0(y) < 0$  on  $[0, \bar{y}_1]$  also gives a contradiction.

By this and Lemma A.1(a) it is therefore necessary that  $V_0(y) \neq 0$  on  $[0, \bar{y}_1]$  and that there is a  $y_0 \in [0, \bar{y}_1]$  so that  $V'_0(y_0) = 0$ . Assuming that  $V_0(y) < 0$  on  $[0, \bar{y}_1]$ , Lemma A.1(b) gives

$$(A.12) \quad \begin{aligned} V'_1(y) &< V'_2(y), & y \in (0, y_0), \\ V'_1(y) &> V'_2(y), & y \in (y_0, \bar{y}_1). \end{aligned}$$

If  $\eta_1 < y_0$ , then  $\gamma_1 < y_0$  as well, and then from (A.12) it follows that  $\gamma_1 < \gamma_2$ . But this was proved above to give a contradiction, and hence  $\eta_1 \geq y_0$ . But then  $V'_2(\eta_1) \leq V'_1(\eta_1)$ , and hence  $\eta_2 \leq \eta_1$  by (A.12) and Lemma A.8, the latter saying that  $V'_2(\bar{y}_2-) < V'_1(\eta_1)$ . Therefore, again by Lemma A.8, the second inequality in (A.12) holds for  $y \in (0, \bar{y}_2)$  and so

$$\begin{aligned} -\frac{d_0}{1+d_1} &= \int_{\eta_1}^{\bar{y}_1} (V'_1(y) - (1+d_1)^{-1}) dy \\ &= \int_{\eta_1}^{\bar{y}_2} (V'_1(y) - (1+d_1)^{-1}) dy \\ &> \int_{\eta_1}^{\bar{y}_2} (V'_2(y) - (1+d_1)^{-1}) dy \\ &> \int_{\eta_2}^{\bar{y}_2} (V'_2(y) - (1+d_1)^{-1}) dy = -\frac{d_0}{1+d_1}, \end{aligned}$$

which is a contradiction. Hence we have proved that  $V_0(y) > 0$  on  $[0, \bar{y}_1]$  and that  $V'_0(y_0) = 0$  for some  $y_0 \in [0, \bar{y}_1]$ . Therefore (A.8) holds, where the second inequality is for  $y \in (y_0, \bar{y}_1)$ .

We will extend the result to  $(\bar{y}_1, \bar{y}_2]$ , and since  $V'_1(y) = V'_2(y)$  on  $[\bar{y}_2, \infty)$ , we are then in fact done. By Lemma A.8,  $V'_2(y) < (1+d_1)^{-1}$  on  $(\eta_2, \bar{y}_2)$ , while by definition  $V'_1(y) = (1+d_1)^{-1}$  on  $(\bar{y}_1, \bar{y}_2)$ . We have shown that  $V_2(\bar{y}_1) > V_1(\bar{y}_1)$ , and hence to finish the first part of the lemma it is sufficient to show that  $V_2(\bar{y}_2) > V_1(\bar{y}_2)$ . To this end, note first that if  $\eta_2 \leq y_0$ , then  $\gamma_2 \leq y_0$  as well, and so by (A.8),  $\gamma_1 > \gamma_2$ . However, this is proved to give a contradiction as before, and so  $\eta_2 > y_0$ , implying by (A.8) that  $\eta_1 < \eta_2$ . Now  $V'_2(y) > (1+d_1)^{-1}$  on  $(\eta_1, \eta_2)$ , and using that  $V_1(\bar{y}_2) = V_1(\eta_1) + (\bar{y}_2 - \eta_1 - d_0)/(1+d_1)$ , we get

$$\begin{aligned} V_2(\bar{y}_2) - V_1(\bar{y}_2) &= V_2(\eta_2) - V_1(\eta_1) - \frac{\eta_2 - \eta_1}{1+d_1} \\ &= \int_{\eta_1}^{\eta_2} (V'_2(y) - (1+d_1)^{-1}) dy + V_2(\eta_1) - V_1(\eta_1) > 0. \end{aligned}$$

For the second part let  $\tilde{V}_n$  be the value function with a barrier  $\bar{y}$  and  $c_0 = c_1 = d_0 = d_1 = 0$ , and let  $V_n = V_{\bar{y}_n}$  be as above. Then clearly  $V_n(y) \leq \tilde{V}_n(y)$ , but by [12, Theorem 4.4],  $\tilde{V}_n(y) \leq y + \mu(0)/r$ . Letting  $\bar{y}_n \rightarrow \infty$  gives boundedness. To prove that the derivatives converge, as in the proof of Lemma A.8, let

$$V_n(y) = a_1(\bar{y}_n)g_1(y) + a_2(\bar{y}_n)g_2(y),$$

where  $g_1(0) = 0$  and  $g_2(0) = 1$ . Then  $a_2(\bar{y}_n) = V_n(0) \rightarrow V(0) \stackrel{\text{def}}{=} a_2$  as  $\bar{y}_n \rightarrow \infty$ . Also using, e.g., that  $V_n(1) \rightarrow V(1)$  gives that  $a_1(\bar{y}_n) \rightarrow a_1$ , and so

$$V(y) = a_1g_1(y) + a_2g_2(y).$$

Convergence of the derivatives follows.

For the last part we use the notation above. We know that  $V'_n(y) > (1 - c_1)^{-1}$  on  $(0, \gamma_n)$ , and from Lemma A.8 we know that  $V'_n(y) < (1 - c_1)^{-1}$  on  $(\gamma_n, \bar{y}_n)$ . Therefore,

$$\begin{aligned} V_n(y) - V_n(x) - \frac{y-x}{1-c_0} &= \int_x^y (V'_n(u) - (1-c_1)^{-1}) du \\ &\leq \int_0^{\gamma_n} (V'_n(u) - (1-c_1)^{-1}) du = \frac{c_0}{1-c_1}. \end{aligned}$$

Letting  $\bar{y}_n \rightarrow \infty$  gives the second inequality.

For the first inequality we know likewise that  $V'_n(y) > (1 + d_1)^{-1}$  on  $(0, \eta_n)$  and that  $V'_n(y) < (1 + d_1)^{-1}$  on  $(\eta_n, \bar{y}_n)$ . We saw in the first part of the proof that  $\{\eta_n\}$  is increasing in  $\bar{y}_n$ . If  $\eta_n \rightarrow \infty$  as  $\bar{y}_n \rightarrow \infty$ , then with  $\eta_n > y$ ,

$$V_n(y) - V_n(x) - \frac{y-x}{1+d_1} = \int_x^y (V'_n(u) - (1+d_1)^{-1}) du > 0.$$

Letting  $\bar{y}_n \rightarrow \infty$  gives that  $V(y) - V(x) - (y-x)/(1+d_1) \geq 0$ . Assume instead that  $\eta_n \rightarrow \eta$  as  $\bar{y}_n \rightarrow \infty$ , and choose  $\bar{y}_n > y$ . If  $x < \eta$ , then for  $\bar{y}_n$  sufficiently large  $V_n(\eta) - V_n(x) - (\eta-x)/(1+d_1) \geq 0$ , and letting  $\bar{y}_n \rightarrow \infty$  shows that the contribution from the interval  $[x, \eta]$  is nonnegative. We may therefore assume that  $x \geq \eta$ . But for  $u \geq \eta$  we saw in the proof of the first part of the lemma that (A.8) applies, and hence  $\{V'_n(u)\}$  is increasing in  $\bar{y}_n$ , and in particular  $V'(u) > V'_n(u)$  for all  $n$  by convergence of the derivatives. Therefore  $V(y) - V(x) > V_n(y) - V_n(x)$  and so

$$V(y) - V(x) - \frac{y-x}{1+d_1} > V_n(y) - V_n(x) - \frac{y-x}{1+d_1} > -\frac{d_0}{1+d_1}. \quad \square$$

*Proof of Theorem 2.3.* Let  $\rho_n = \inf\{t : Y_t \vee C_t \vee D_t \geq n\}$  and let  $\nu_n = \inf\{t : C_t \vee D_t \geq n\}$ . Furthermore, let  $0 = S_0 < S_1 < S_2 < \dots$  be the times investments are made, and similarly let  $0 = T_0 < T_1 < T_2 < \dots$  be the times dividends are paid. If  $\lim_{n \rightarrow \infty} S_n < \infty$  or  $\lim_{n \rightarrow \infty} T_n < \infty$ , it is clear that  $E^y[\int_0^\infty e^{-rt} dA_t] = -\infty$ . Hence we may assume that  $\lim_{n \rightarrow \infty} S_n = \lim_{n \rightarrow \infty} T_n = \infty$ .

Let  $V$  be as in (2.3) if this has a solution and be as in Lemma A.9 if (2.3) has no solution. By the generalized Itô formula, the fact that  $C^c = D^c = 0$ , together with Lemmas A.7 and A.9, we have

$$\begin{aligned} e^{-r(t \wedge \rho_n)} V(Y_{t \wedge \rho_n}) &= V(y) + \int_0^{t \wedge \rho_n} e^{-rs} \sigma(Y_s) V'(Y_s) dW_s \\ &\quad + \sum_{S_i \leq t \wedge \rho_n} e^{-rS_i} (V(Y_{S_i-} + \Delta Y_{S_i}) - V(Y_{S_i-})) \\ &\quad + \sum_{T_i \leq t \wedge \rho_n} e^{-rT_i} (V(Y_{T_i-} + \Delta Y_{T_i}) - V(Y_{T_i-})) \\ &\leq V(y) + \int_0^{t \wedge \rho_n} e^{-rs} \sigma(Y_s) V'(Y_s) dW_s \\ &\quad + \sum_{S_i \leq t \wedge \rho_n} e^{-rS_i} \frac{\Delta Y_{S_i} + c_0}{1-c_1} \\ &\quad + \sum_{T_i \leq t \wedge \rho_n} e^{-rT_i} \frac{\Delta Y_{T_i} + d_0}{1+d_1} \end{aligned}$$

$$\begin{aligned}
&= V(y) + \int_0^{t \wedge \rho_n} e^{-rs} \sigma(Y_s) V'(Y_s) dW_s \\
&\quad + \int_0^{t \wedge \rho_n} e^{-rs} dC_s - \int_0^{t \wedge \rho_n} e^{-rs} dD_s.
\end{aligned}$$

Therefore,

$$V(y) \geq E^y \left[ \int_0^{t \wedge \rho_n} e^{-rs} dD_s \right] - E^y \left[ \int_0^{t \wedge \rho_n} e^{-rs} dC_s \right] + E^y \left[ e^{-r(t \wedge \rho_n)} V(Y_{t \wedge \rho_n}) \right].$$

As in the proof of [12, Lemma 3.3] it follows that

$$M(t) = \int_0^{t-} e^{-rs} dA_s + e^{-rt} V(Y_{t-})$$

is a supermartingale. Again following the arguments in [12, Lemma 3.3] gives

$$\begin{aligned}
V(y) &\geq E^y \left[ \int_0^{\nu_n-} e^{-rs} dA_s \right] + E^y \left[ e^{-r\nu_n} V(Y_{\nu_n-}) \right] \\
&\geq E^y \left[ \int_0^{\nu_n-} e^{-rs} dA_s \right] + V(0) E^y \left[ e^{-r\nu_n} \right].
\end{aligned}$$

Clearly  $\nu_n \rightarrow \infty$  as  $n \rightarrow \infty$ ; hence,

$$V(y) \geq \limsup_{n \rightarrow \infty} E^y \left[ \int_0^{\nu_n-} e^{-rs} dA_s \right].$$

If (2.3) has a solution, then using the strategy of Theorem 2.3(a) gives the equalities above so that

$$V(y) = \int_0^t e^{-rs} dD_s - \int_0^t e^{-rs} dC_s - \int_0^t e^{-rs} \sigma(Y_s) V'(Y_s) dW_s + e^{-rt} V(Y_t).$$

But  $0 \leq Y_s \leq y^*$ , and hence the stochastic integral has expectation zero. Furthermore,

$$E^y \left[ \int_0^t e^{-rs} dD_s \right] < y + \frac{\delta^* - d_0}{1 + d_1} E^y \left[ \sum_{i=1}^{\infty} e^{-rT_i} \right] < \infty$$

since  $T_1, T_2, \dots$  is a renewal process. Therefore, taking expectations and using the monotone convergence theorem together with the dominated convergence theorem, we get

$$V(y) = E^y \left[ \int_0^{\infty} e^{-rt} dA_t \right].$$

Assume now that (2.3) does not have a solution, and let  $(C^*, D^*)$  be an optimal strategy. Let  $D'$  be “the more generous payout,”

$$\Delta D'_s = V(Y_{s-}) - V(Y_{s-} + \Delta Y_s) > \Delta D_s^* \quad \text{when} \quad \Delta D_s^* > 0,$$

where the inequality follows from Lemma A.9. Therefore,  $P(D'_{t \wedge \tau_y} > D^*_{t \wedge \tau_y}) > 0$  for  $t$  sufficiently large (otherwise  $E[\int_0^{\nu_n-} e^{-rs} dD_s^*] = 0$  for all  $n$ , and  $D^*$  is not optimal). Consequently, as in the first part of the proof with  $A^* = D^* - C^*$  and  $A' = D' - C'$ ,

$$V(y) \geq \limsup_{n \rightarrow \infty} E^y \left[ \int_0^{\nu_n-} e^{-rs} dA'_s \right] > \limsup_{n \rightarrow \infty} E^y \left[ \int_0^{\nu_n-} e^{-rs} dA_s^* \right].$$

Let

$$\varepsilon = V(y) - \limsup_{n \rightarrow \infty} E \left[ \int_{0-}^{\nu_n-} e^{-rs} dA_s^* \right].$$

By Lemma A.9, there exists a  $\bar{y}$  so that  $V(y) - V_{\bar{y}, \gamma(\bar{y}), \delta(\bar{y})}(y) < \varepsilon$ , and so  $(C^*, D^*)$  cannot be optimal. That  $V(y) = \lim_{\bar{y} \rightarrow \infty} V_{\bar{y}, \gamma(\bar{y}), \delta(\bar{y})}(y)$  follows from Lemma A.9. This ends the proof.  $\square$

*Proof of Proposition 2.4.* Let  $g(y) = V(y)$ , which is a solution by Lemma A.9. Also let

$$S(y) = \int_0^y e^{-\int_0^x \frac{2\mu(v)}{\sigma^2(v)} dv} dx$$

be the scale function. By the method of reduction of order (see, e.g., [14, p. 31]), another independent solution is

$$\tilde{g}_2(y) = g(y) \int_l^y \frac{1}{g^2(x)} S'(x) dx \stackrel{\text{def}}{=} g(y)(k(y) - k(l)),$$

where  $l$  is chosen large enough so that  $g(l) > 0$ .

Let  $\tau_0^y$  and  $\tau_{\bar{y}}^y$  be as in the proof of Lemma A.3. By results in [6, Chapter 15.1],

$$P^y(\tau_{\bar{y}}^y < \tau_0^y) = \frac{S(y)}{S(\bar{y})}.$$

This means that if  $S(\bar{y}) \rightarrow \infty$  as  $\bar{y} \rightarrow \infty$ , then  $P^y(\tau_0^y < \infty) = 1$ . But then it cannot be optimal to never pay out dividends, since this will result in a negative value function for all  $y$ , which clearly is wrong. Hence  $\lim_{y \rightarrow \infty} S(y) < \infty$ , and since  $S'(y) > 0$ , this implies in particular that  $\lim_{y \rightarrow \infty} S'(y) = 0$ .

By Lemma A.9 and what we have already proved, it is clear that  $k_\infty = \lim_{y \rightarrow \infty} k(y)$  exists and is finite. Therefore  $g_2(y) = g(y)(k(y) - k_\infty)$  is also a solution, and by L'Hospital's rule,

$$\lim_{y \rightarrow \infty} g_2(y) = \lim_{y \rightarrow \infty} \frac{k(y) - k_\infty}{\frac{1}{g(y)}} = \lim_{y \rightarrow \infty} \frac{S'(y)}{g'(y)} = 0,$$

since  $V' = \lim_{y \rightarrow \infty} g'(y) = 1/(1 + d_1)$  by Lemmas A.1(f) and A.9. Finally, any solution is a linear combination of  $V(y)$  and  $g_2(y)$ , and the last result follows.

To prove (b), observe first that  $V(y) \leq \tilde{V}(y)$ , where  $\tilde{V}$  is the value function when  $c_0 = c_1 = d_0 = 0$ . As in [12, Theorem 4.4],

$$\tilde{V}(y) = \frac{1}{1 + d_1} \left( y + \frac{\mu(0)}{r} \right).$$

Assume that there is an optimal policy. Then by Theorem 2.3(a) and the above, we obtain for  $y \geq y^*$ ,

$$V(y) = V(\eta^*) + \frac{y - \eta^* - d_0}{1 + d_1} \leq \frac{1}{1 + d_1} \left( y + \frac{\mu(0)}{r} - d_0 \right).$$

Let  $\bar{y} > y$  and let  $T_{\bar{y}}^y = \min\{\tau_0^y, \tau_{\bar{y}}^y\}$ . Using a barrier strategy at  $\bar{y}$  that pays the whole amount  $\bar{y}$  yields the value function

$$\begin{aligned} \hat{V}_{\bar{y}}(y) &= E^y \left[ e^{-rT_{\bar{y}}^y} 1_{\{Y_{T_{\bar{y}}^y-} = \bar{y}\}} \right] \left( \frac{\bar{y} - d_0}{1 + d_1} + \hat{V}_{\bar{y}}(0) \right) + E^y \left[ e^{-rT_{\bar{y}}^y} 1_{\{Y_{T_{\bar{y}}^y-} = 0\}} \right] \hat{V}_{\bar{y}}(0) \\ &= h_{\bar{y}}(y) \frac{\bar{y} - d_0}{1 + d_1} + k_{\bar{y}}(y) \hat{V}_{\bar{y}}(0), \end{aligned}$$

where

$$\begin{aligned} h_{\bar{y}}(y) &= E^y \left[ e^{-rT_{\bar{y}}^y} 1_{\{Y_{T_{\bar{y}}^y} = \bar{y}\}} \right], \\ k_{\bar{y}}(y) &= E^y \left[ e^{-rT_{\bar{y}}^y} \right]. \end{aligned}$$

Again by results from Chapter 15.3 in [6],  $h_{\bar{y}}$  is the solution of

$$Lh_{\bar{y}}(y) = 0, \quad h_{\bar{y}}(\bar{y}) = 1, \quad \text{and} \quad h_{\bar{y}}(0) = 0.$$

This gives

$$\hat{V}_{\bar{y}}(y) = \frac{\bar{y} - d_0}{1 + d_1} \left( \frac{g_1(y)}{g_1(\bar{y}) - \frac{g_1(0)}{g_2(0)}g_2(\bar{y})} - \frac{g_2(y)}{\frac{g_2(0)}{g_1(0)}g_1(\bar{y}) - g_2(\bar{y})} \right) + k_{\bar{y}}(y)\hat{V}_{\bar{y}}(0).$$

Clearly  $\hat{V}_{\bar{y}}(0)$  is bounded, and since  $Y$  increases only exponentially, letting  $\bar{y} = e^{y^2}$  implies that  $k_{\bar{y}}(y) \rightarrow 0$  as  $y \rightarrow \infty$ . Therefore, for any positive  $\varepsilon < d_0$  and  $y$  sufficiently large,

$$\hat{V}_{\bar{y}}(y) \geq \frac{1}{1 + d_1} \frac{g_1(y)}{g_1'} - \frac{\varepsilon}{2} \geq \frac{1}{1 + d_1} \left( y + \frac{\mu(0)}{r} \right) - \varepsilon > V(y),$$

which is a contradiction. Hence (2.3) cannot have a solution, and this proves (b).

Part (c) is the same as (a) since all solutions are a linear combination of  $g_1$  and  $g_2$  from (a).  $\square$

*Proof of Proposition 2.6.* As in the proof of Lemma A.9, let  $V_n(y)$  be the value function with barrier  $\bar{y}_n$ . Then  $V_n(y) = a_1(\bar{y}_n)g_1(y) + a_2(\bar{y}_n)g_2(y)$ , and by the proof of Lemma A.9,  $a_i(\bar{y}_n) \rightarrow a_i$  as  $\bar{y}_n \rightarrow \infty$ , where  $V(y) = a_1g_1(y) + a_2g_2(y)$ .

Let  $V' = \lim_{y \rightarrow \infty} V'(y)$ . By Lemma A.9,  $V' \geq (1 + d_1)^{-1}$ , and by Lemma A.8,

$$\frac{1}{1 + d_1} > V'_n(\bar{y}_n) = a_1(\bar{y}_n)g'_1(\bar{y}_n) + a_2(\bar{y}_n)g'_2(\bar{y}_n).$$

Letting  $\bar{y}_n \rightarrow \infty$  gives that  $(1 + d_1)^{-1} \geq V'$ , hence  $V' = (1 + d_1)^{-1}$ , and this proves (a). For (b) define

$$\begin{aligned} h(y) &= V(y) - V(0) - \frac{y + c_0}{1 - c_1}, \\ h_n(y) &= V_n(y) - V_n(0) - \frac{y + c_0}{1 - c_1}. \end{aligned}$$

Since  $V'_n(\gamma_n) = (1 - c_1)^{-1}$  and  $V_n$  is concave on  $[0, \gamma_n]$ , we have that  $h'_n(0) > 0$ , implying that  $h'(0) \geq 0$ . By (a),  $h'(y) < 0$  for all  $y$  sufficiently large, and so  $h$  must have a maximum at some  $\bar{\gamma}$ , and then  $h'(\bar{\gamma}) = 0$ , i.e.,  $V'(\bar{\gamma}) = (1 - c_1)^{-1}$ . By Lemma A.1(f),  $V'(y) > (1 - c_1)^{-1}$  for  $y < \bar{\gamma}$  and  $V'(y) < (1 - c_1)^{-1}$  for  $y > \bar{\gamma}$ . Since  $V$  is an optimal value function and all reinvestments at 0 have a marginal cost of  $(1 - c_1)^{-1}$ , it is optimal to reinvest as long as the marginal return on reinvestments is at least as high as the marginal cost of reinvesting, which by the above arguments just means that

$$(A.13) \quad V(\bar{\gamma}) = V(0) + \frac{\bar{\gamma} + c_0}{1 - c_1},$$

and this is the same as the second equation in (2.4).

Assume that  $\{\gamma_n\}$  is unbounded so that  $\gamma_n \rightarrow \infty$  along some sequence  $\bar{y}_n$ . Then for  $\varepsilon > 0$  and  $n$  large enough, since  $V_n$  is concave on  $[0, \gamma_n]$ ,

$$V'_n(\bar{\gamma} + \varepsilon) > (1 - c_1)^{-1} > V'(\bar{\gamma} + \varepsilon).$$

Letting  $n \rightarrow \infty$  gives a contradiction, hence  $\{\gamma_n\}$  is bounded, and so (along a subsequence)  $\gamma_n \rightarrow \gamma_0$  for some  $\gamma_0$ . But

$$V'(\gamma_0) = V'(\gamma_0) - V'_n(\gamma_0) + V'_n(\gamma_0) - V'_n(\gamma_n) + (1 - c_1)^{-1}.$$

However,

$$|V'(\gamma_0) - V'_n(\gamma_0)| + |V'_n(\gamma_0) - V'_n(\gamma_n)| \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

hence  $V'(\gamma_0) = (1 - c_1)^{-1}$ , and so  $\gamma_0 = \bar{\gamma}$ . Also  $h_n(\gamma_n) = 0$  so  $h(\bar{\gamma}) = 0$  as well, giving another proof of (A.13).

Finally for (c), if  $c_1 = d_1 = 0$  but the equations (2.4) have a solution, the argument is as above. So assume that (2.4) has no solution. Then  $V'(y) > 1$  for all  $y$ , and hence at  $y = 0$  it is desirable to reinvest as much as possible since the marginal return on reinvestment is higher than the marginal cost of reinvestment. Therefore, as in (A.13),

$$(A.14) \quad V(0) = \lim_{y \rightarrow \infty} (V(y) - y) - c_0.$$

Then (a) and (A.14), together with the limit results for  $g_2(y)$  given in Proposition 2.4(a), yield the result.  $\square$

#### REFERENCES

- [1] F. AVRAM, Z. PALMOWSKI, AND M. R. PISTORIUS, *On the optimal dividend problem for a spectrally negative Lévy process*, Ann. Appl. Probab., 17 (2007), pp. 156–180.
- [2] A. CADENILLAS, S. SARKAR, AND F. ZAPATERO, *Optimal dividend policy with mean-reverting cash reservoir*, Math. Finance, 17 (2007), pp. 81–109.
- [3] G. M. CONSTANTINIDES AND S. F. RICHARD, *Existence of optimal simple policies for discounted-cost inventory and cash management in continuous time*, Oper. Res., 26 (1978), pp. 620–636.
- [4] J. M. HARRISON, T. M. SELLKE, AND A. J. TAYLOR, *Impulse control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 454–466.
- [5] M. JEANBLANC-PICQUÉ AND A. N. SHIRYAEV, *Optimization of the flow of dividends*, Russian Math. Surveys, 50 (1995), pp. 257–277.
- [6] S. KARLIN AND H. M. TAYLOR, *A Second Course in Stochastic Processes*, Academic Press, New York, 1981.
- [7] N. V. KRYLOV, *Lectures on Elliptic and Parabolic Equations in Hölder Spaces*, Grad. Stud. Math. 12, American Mathematical Society Providence, RI, 1996.
- [8] J. PAULSEN, *Optimal dividend payments until ruin of diffusion processes when payments are subject to both fixed and proportional costs*, Adv. Appl. Probab., 39 (2007), pp. 669–689.
- [9] E. L. PORTEUS, *On optimal dividend, reinvestment, and liquidation policies for the firm*, Oper. Res., 25 (1977), pp. 818–834.
- [10] S. F. RICHARD, *Optimal impulse control of a diffusion process with both fixed and proportional costs of control*, SIAM J. Control Optim., 15 (1977), pp. 79–91.
- [11] S. P. SETHI AND M. TAKSAR, *Optimal financing of a corporation subject to random returns*, Math. Finance, 12 (2002), pp. 155–172.
- [12] S. E. SHREVE, J. P. LEHOCZKY, AND D. P. GAVAR, *Optimal consumption for general diffusions with absorbing and reflecting barriers*, SIAM J. Control Optim., 22 (1984), pp. 55–75.
- [13] L. J. SLATER, *Confluent Hypergeometric Functions*, Cambridge University Press, London, 1960.
- [14] K. YOSIDA, *Lectures on Differential and Integral Equations*, Dover Publications, New York, 1990.

## ON LOCAL TRANSVERSE FEEDBACK LINEARIZATION\*

CHRISTOPHER NIELSEN<sup>†</sup> AND MANFREDI MAGGIORE<sup>†</sup>

**Abstract.** Given a control-affine system and a controlled invariant submanifold, we present necessary and sufficient conditions for local feedback equivalence to a system whose dynamics transversal to the submanifold are linear and controllable. A key ingredient used in the analysis is the new notion of transverse controllability indices of a control system with respect to a set.

**Key words.** set stabilization, feedback linearization, controlled invariant sets, zero dynamics, multi-input systems, nonlinear geometric control

**AMS subject classifications.** 37N35, 93B10, 93B29, 93B27

**DOI.** 10.1137/070682125

**1. Introduction.** Ever since Poincaré’s seminal work [22], the problem of equivalence of vector fields has been a central question in the field of dynamics. In his 1879 work, Poincaré found sufficient conditions for an analytic vector field to be locally equivalent to a linear one by means of an analytic transformation. Poincaré’s key insight in formulating this problem was that, rather than trying to solve a differential equation, it is convenient to seek a coordinate transformation reducing the associated vector field to its “simplest” form, the normal form. In control theory, the problem of equivalence of a control system to a linear controllable system by means of smooth coordinate transformations was first formulated by Krener in 1973 [14]. In 1978, Brockett [3] formulated and solved the feedback linearization problem for single-input, single-output systems, whereby the equivalence to a linear controllable system is established by means of a smooth coordinate transformation and a regular feedback transformation; this is referred to as feedback equivalence. The multi-input, multi-output extension of Brockett’s work was carried out by Jakubczyk and Respondek in [12] and, independently, by Hunt, Su, and Meyer in [8]; see also [26]. When a control system is not feedback linearizable, it is natural to ask whether it admits a feedback linearizable subsystem. This problem, first posed by Isidori and Krener in [10], is referred to as partial feedback linearization. For single-input systems, Krener, Isidori, and Respondek [13] investigated partial feedback linearization yielding a linear subsystem of maximal dimension. This result was extended by Marino in [15] to the multi-input case; see also [16], [23]. For systems with outputs, Xu and Hunt [30], [31], consider a similar problem.

In [2], Banaszuk and Hauser formulated and solved the transverse feedback linearization problem (TFLP) for periodic orbits of single-input control-affine systems. If  $\Gamma^*$  is a periodic orbit of the open-loop system, the problem entails finding conditions for feedback equivalence to a control system whose dynamics transversal to  $\Gamma^*$  are linear and controllable. In [19], we generalized Banaszuk and Hauser’s results to the case when  $\Gamma^*$  is an arbitrary controlled invariant embedded submanifold of the

---

\*Received by the editors February 7, 2007; accepted for publication (in revised form) March 21, 2008; published electronically July 30, 2008. Both authors were supported by the National Science and Engineering Research Council (NSERC) of Canada. The first author was partially supported by the Ontario Graduate Scholarship (OGS).

<http://www.siam.org/journals/sicon/47-5/68212.html>

<sup>†</sup>Department of Electrical and Computer Engineering, University of Toronto, 10 King’s College Road, Toronto, ON, M5S 3G4, Canada (nielsen@control.utoronto.ca, maggiore@control.utoronto.ca).

state space. In this paper we present the complete solution to the local TFLP for multi-input systems, relying on a mild regularity assumption. A key ingredient used in the analysis of the problem is the new notion of transverse controllability indices of a control system with respect to a set. The transverse controllability indices are an adaptation of those introduced by Marino [15].

When the set  $\Gamma^*$  is an equilibrium point, the problem considered in this paper (see section 3) reduces to the classical state-space exact linearization problem. In this special case our conditions coincide with those of the classical results on feedback equivalence to linear, time-invariant, controllable systems [8], [12], and the transverse controllability indices coincide with the controllability indices introduced by Marino [15].

We now discuss some of the applications of transverse feedback linearization (TFL). While classical feedback linearization is used to stabilize equilibria of non-linear systems, TFL is applicable to the more general set stabilization problem. Indeed, if a system is transversely feedback linearizable with respect to a controlled invariant manifold  $\Gamma^*$ , then designing a controller that locally stabilizes  $\Gamma^*$  amounts to designing a stabilizer for the origin of a linear time-invariant system, and so the set stabilization problem is greatly simplified (see section 3 for a more precise discussion). In light of the above, TFL is relevant to all those problems where the control objective is the stabilization of a manifold, rather than an equilibrium. Consider, for instance, the simplest synchronization (or state agreement) problem: make the states of two coupled dynamical systems converge to one another. This is equivalent to stabilizing the diagonal subspace. In the more general case, when one wants to make the outputs of several coupled dynamical systems converge to one another, then, generally, the set to be stabilized is a manifold. As other relevant applications of TFL we mention path following (make the output of a dynamical system approach and follow a path) [2], [18], [19], and the stabilization of virtual constraints in mechanical systems [24].

Another important application of our main result is the solution of the following *zero dynamics assignment problem with relative degree*. Given a control-affine system and a controlled invariant manifold  $\Gamma^*$ , does there exist an output function yielding a well-defined vector relative degree whose associated zero dynamics manifold locally coincides with  $\Gamma^*$ ? Our main result in Theorem 3.2 gives checkable necessary and sufficient conditions that completely answer this question.

This paper is organized as follows. Section 2 contains mathematical preliminaries. Section 3 presents the formal problem statement, the statement of our main result, Theorem 3.2, and a comparison of our result to the solution of the classical state-space exact linearization problem [8], [12]. A relationship to the partial feedback linearization problem is established in Theorem 3.5. In section 4 we introduce the notion of transverse indices, compare them to the controllability indices of Marino [15] in Lemma 4.1, and establish their feedback invariance. The proof of the main result is presented in section 5, and section 6 contains concluding remarks.

**2. Preliminaries.** Consider a control system  $\Sigma$  modeled by equations of the form

$$(2.1) \quad \Sigma : \quad \dot{x} = f(x) + \sum_{i=1}^m g_i(x)u_i =: f(x) + g(x)u.$$

Here  $x \in \mathbb{R}^n$  is the state, and  $u = (u_1, \dots, u_m) \in \mathbb{R}^m$  is the control input. The vector fields  $f, g_1, \dots, g_m : \mathbb{R}^n \longrightarrow T\mathbb{R}^n$  are smooth ( $C^\infty$ ). We assume throughout this paper



that  $g_1, \dots, g_m$  are linearly independent.

When we talk of a manifold  $M$ , we mean a smooth manifold without boundary. By a submanifold we mean an embedded submanifold. All objects are presumed to be smooth. In this paper we consider submanifolds of  $\mathbb{R}^n$ , where  $\mathbb{R}^n$  is identified with Euclidean  $n$ -space.

**2.1. Notation.** If  $k$  is a positive integer,  $\mathbf{k}$  denotes the set of integers  $\{0, 1, \dots, k-1\}$ . We let  $\text{col}(x_1, \dots, x_k) := [x_1 \ \dots \ x_k]^\top$  and, given two column vectors  $a$  and  $b$ , we let  $\text{col}(a, b) := [a^\top \ b^\top]^\top$ . If  $U$  is an open set of  $\mathbb{R}^n$ , let  $\text{Diff}(U)$  denote the collection of diffeomorphisms from  $U$  to some open set  $\tilde{U} \subset \mathbb{R}^n$ . If  $F : M \rightarrow N$  is a map between manifolds, then  $dF_x : T_x M \rightarrow T_{F(x)} N$  denotes its differential. If  $M$  and  $N$  are vector spaces, then use  $dF_x$  to denote the Jacobian matrix of  $F$  at  $x$ . If  $F : M \rightarrow N$  is a diffeomorphism between two manifolds, and if  $v$  is a vector field on  $M$ , then the differential of  $F$  can be used to define a vector field on  $N$  by means of the push-forward map  $F_*$ , defined as  $F_* v(q) = (dF_p v(p))|_{p=F^{-1}(q)}$ . This corresponds to the usual change of coordinates in a differential equation. We denote by  $I_m$  the  $m \times m$  identity matrix. The direct sum of two matrices  $A$  and  $B$  is the block diagonal matrix

$$A \oplus B = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix},$$

where the zeros denote matrices of suitable size. Given two subspaces  $V$  and  $W$  of the same vector space, the notation  $V \oplus W$  (internal direct sum) is used to represent the subspace  $V + W$  when  $V$  and  $W$  are linearly independent.

**DEFINITION 2.1.** *Given an open set  $U \subseteq \mathbb{R}^n$ , a regular static feedback, denoted  $(\alpha, \beta)$ , on  $U$  for the control system (2.1) is a relation*

$$u = \alpha(x) + \beta(x)v,$$

where  $u = (u_1, \dots, u_m)$  and  $\alpha : U \rightarrow \mathbb{R}^m$ ,  $\beta : U \rightarrow \text{GL}(m, \mathbb{R})$  are smooth mappings. We denote by  $\tilde{f} := f + g\alpha$  and  $\tilde{g} := g\beta$  the vector fields obtained after the application of  $(\alpha, \beta)$ .

**DEFINITION 2.2.** *Two control systems,  $\Sigma : \dot{x} = f(x) + g(x)u$  and  $\hat{\Sigma} : \dot{\hat{x}} = \hat{f} + \hat{g}\hat{u}$ , are feedback equivalent on an open set  $U \subseteq \mathbb{R}^n$  if there exist a regular static feedback  $(\alpha, \beta)$  defined on  $U$  and a diffeomorphism  $\Xi \in \text{Diff}(U)$  such that*

$$\hat{f} = \Xi_*(f + g\alpha), \quad \hat{g} = \Xi_*(g\beta)$$

on  $U$ .

On a manifold  $M$ ,  $\mathbf{V}(M)$  will denote the set of all smooth vector fields on  $M$ , and  $C^\infty(M)$  will denote the ring of smooth real-valued functions on  $M$ . Given  $v \in \mathbf{V}(M)$ ,  $\phi_t^v(p)$  denotes the solution of  $\dot{x} = v(x)$  with initial condition  $x(0) = p$  at time  $t$ . A closed connected set  $N \subset M$  is said to be invariant for  $f \in \mathbf{V}(M)$  if

$$(p_0 \in N) \Rightarrow (\forall t \in \mathbb{R})(\phi_t^f(p_0) \in N).$$

A closed connected set  $N \subset \mathbb{R}^n$  is called controlled invariant for (2.1) if there exists a smooth feedback  $\bar{u} : N \rightarrow \mathbb{R}^m$  making  $N$  an invariant set for the closed-loop system. Following [28], we denote the class of closed, connected, embedded submanifolds of  $\mathbb{R}^n$  which are controlled invariant for (2.1) by  $\mathcal{I}(f, g, \mathbb{R}^n)$ . If  $N \in \mathcal{I}(f, g, \mathbb{R}^n)$ , we write

$\mathcal{F}(f, g, N)$  for the collection of maps that render  $N$  controlled invariant, i.e., maps  $\bar{u} : N \rightarrow \mathbb{R}^m$  such that  $f + g\bar{u}$  is tangent to  $N$ , i.e.,

$$(f + g\bar{u})|_N : N \rightarrow TN.$$

If  $f \in \mathbf{V}(M)$  and  $\lambda \in C^\infty(M)$ , then

$$L_f \lambda(p) = \lim_{h \rightarrow 0} \frac{1}{h} \left[ \lambda(\phi_h^f(p)) - \lambda(p) \right]$$

is the Lie derivative of  $\lambda$  with respect to  $f$  at  $p$ ; it is also an element of  $C^\infty(M)$ . If  $f, g \in \mathbf{V}(M)$ , then the Lie bracket of  $f$  and  $g$  is defined by the relation

$$L_{[f,g]} \lambda = L_f(L_g \lambda) - L_g(L_f \lambda) \quad \forall \lambda \in C^\infty(M).$$

We will use the following standard notation for iterated Lie derivatives and Lie brackets:

$$\begin{aligned} L_g L_f \lambda &:= L_g(L_f \lambda), \\ L_g^0 \lambda &:= \lambda, \quad L_g^k \lambda := L_g(L_g^{k-1} \lambda), \\ ad_f^0 g &:= g, \quad ad_f^k g := \left[ f, ad_f^{k-1} g \right], \quad k \geq 1. \end{aligned}$$

The set  $\mathbf{V}(M)$  can be equipped with different algebraic structures. For our purposes it suffices to consider  $\mathbf{V}(M)$  as either (i) a vector space (infinite dimensional) over  $\mathbb{R}$  which, when endowed with the Lie bracket  $[\cdot, \cdot] : \mathbf{V}(M) \times \mathbf{V}(M) \rightarrow \mathbf{V}(M)$ , becomes a Lie algebra, or (ii) a module over the ring  $C^\infty(M)$ . Given a Lie algebra  $\mathfrak{g}$ , a subset  $\mathfrak{h} \subset \mathfrak{g}$  is called a subalgebra if  $h_1, h_2 \in \mathfrak{h}$  implies  $[h_1, h_2] \in \mathfrak{h}$ .

Finally, if  $A \subset M$  is any subset, then a smooth map  $r : M \rightarrow A$  such that  $r|_A = 1_A$ , where  $1_A$  is the identity map on  $A$ , is called a smooth retraction of  $M$  onto  $A$ . The following lemma regarding retractions is a simpler, local version of the tubular neighborhood theorem [5].

**LEMMA 2.3.** *Let  $N \subset \mathbb{R}^n$  be an  $n^*$ -dimensional submanifold of  $\mathbb{R}^n$ . Then, for every  $p \in N$  there exist a neighborhood  $U$  of  $p$  in  $\mathbb{R}^n$  and a smooth retraction  $r : U \rightarrow N \cap U$ .*

**2.2. Vector bundles.** The discussion, terminology, and notation of this section is standard and can be found in [6] or [25]. A smooth  $n$ -dimensional vector bundle (or  $n$ -plane bundle)  $\xi = (\pi, E, B)$  can be thought of as<sup>1</sup> a family  $\{E_p\}_{p \in B}$  of disjoint  $n$ -dimensional vector spaces parameterized by a space  $B$ . The union of these vector spaces is the space  $E$ , and  $B$  is called the base space. The map  $\pi : E \rightarrow B$ ,  $E_p \mapsto p$  is a smooth surjective submersion and is called the vector bundle projection. Moreover,  $\xi$  is “locally trivial” in the sense that, locally (with respect to  $B$ ),  $E$  looks like a product with  $\mathbb{R}^n$ : for each  $p \in B$ , there is a neighborhood  $U$  of  $p$  and a diffeomorphism  $t : \pi^{-1}(U) \rightarrow U \times \mathbb{R}^n$ ,  $v_q \mapsto (t_1(q), t_2(q)v)$ , which is an isomorphism from each fiber  $\pi^{-1}(q)$  onto  $q \times \mathbb{R}^n$  for each  $q \in U$ . This property is exhibited by the commutative diagram below, where  $\tilde{\pi}(q, v) = q$ :

$$\begin{array}{ccc} \pi^{-1}(U) & \xrightarrow{t} & U \times \mathbb{R}^n \\ \pi \downarrow & \swarrow \tilde{\pi} & \\ U & & \end{array}$$

<sup>1</sup>The discussion here is informal. In particular we define a vector bundle as a triple  $(\pi, E, B)$ , where more formally one defines a vector bundle as a 5-tuple by augmenting the above triple with two operations  $\oplus$  and  $\otimes$ . See [6] and [25] for more details.

The pair  $(\pi^{-1}(U), t)$  is called a vector bundle chart with domain  $U$  and dimension  $n$ . The collection of all vector bundle charts of  $\xi$  is a vector bundle atlas. A vector bundle is a manifold in its own right with an atlas of compatible vector bundle charts. We will usually refer to a vector bundle as simply  $\xi$ , or  $\pi : E \rightarrow B$ , or even denote the bundle by  $E$  alone.

If  $A \subset B$  is any subset and  $\xi = (\pi, E, B)$ , then we denote  $\pi^{-1}(A)$  by  $E|_A$ , the restriction of  $\xi$  to  $A$ . It is a well-defined vector bundle. A subbundle of the bundle  $\xi = (\pi, E, B)$  is a bundle  $\xi_0 = (\pi_0, E_0, B)$ , over the same base space  $B$  such that  $E_0 \subset E$ , and  $\pi|_{E_0} = \pi_0$ . Additionally, there must exist a vector bundle atlas  $\Phi$  for  $\xi$  such that if  $(\pi^{-1}(U), t) \in \Phi$ , the following diagrams commute:

$$\begin{array}{ccc} \pi^{-1}(U) & \xrightarrow{t} & U \times \mathbb{R}^k \times \mathbb{R}^{n-k} \\ \pi \downarrow & \nearrow \tilde{\pi} & \\ U & & \end{array} \quad \begin{array}{ccc} \pi_0^{-1}(U) & \xrightarrow{t} & U \times \mathbb{R}^k \times \{0\} \\ \pi_0 \downarrow & \nearrow \tilde{\pi} & \\ U & & \end{array}$$

The prime example of a vector bundle is the tangent bundle  $\pi : TM \rightarrow M$  of a manifold  $M$ . If  $N \subset M$  is a submanifold of  $M$ , then  $TN$  is a subbundle of  $TM|_N$ :

$$TN = \{v_x \in TM|_N : v(x) \in T_x N\}.$$

The algebraic normal bundle of  $N$  in  $M$  is the subbundle over  $N$  whose fibers are the quotient spaces  $T_x M / T_x N$ . It is denoted  $TM|_N / TN$ .

Let  $\xi = (\pi, E, B)$  be a smooth vector bundle. A  $C^\infty$  inner product or orthogonal structure on  $\xi$  is a family  $\{\alpha_p\}_{p \in B}$ , where each  $\alpha_p$  is an inner product on  $E_p$  and the map  $(p, y, z) \mapsto \alpha_p(y, z)$  defined on  $\{(p, y, z) \in B \times E \times E : p = \pi(y) = \pi(z)\}$  is  $C^\infty$ . The pair  $(\xi, \alpha)$  is called an orthogonal vector bundle. If  $M$  is a manifold, a  $C^\infty$  orthogonal structure on  $TM$  is called a Riemannian metric. In this paper, orthogonal structures will always arise in subbundles of  $T\mathbb{R}^n|_V$ , where  $V$  is a submanifold of  $\mathbb{R}^n$ , whereby the standard inner product on  $\mathbb{R}^n$  is used. Suppose  $(\xi, \alpha)$  is an orthogonal vector bundle. If  $y, z$  are in the same fiber  $E_p$ , we write  $\langle y, z \rangle$  or  $\langle y, z \rangle_p$  for  $\alpha_p(y, z)$ . If  $\xi = (\pi^0, E^0, B) \subset \eta = (\pi, E, B)$  is a subbundle, the orthogonal complement  $\xi^\perp \subset \eta$  is the subbundle defined fiberwise as

$$(\xi^\perp)_p = (\xi_p)^\perp = \{y \in E_p : \langle y, z \rangle = 0, z \in E_p^0\}.$$

Note that  $\xi^\perp$  is isomorphic to  $\eta/\xi$ . Of particular interest to us will be the case when  $N \subset M$  is a submanifold and  $M$  has a Riemannian metric. In this case  $TN^\perp \subset TM|_N$  is called the geometric normal bundle of  $N$  in  $M$ .

**2.3. Distributions.** A smooth distribution  $D$  on a manifold  $M$  is an assignment to each  $p \in M$  of a subspace  $D(p) \subseteq T_p M$  which varies smoothly as  $p$  varies. Locally, a smooth distribution is spanned by a collection of smooth vector fields which are called local generators. A point  $p \in M$  is a regular point of  $D$  if there exists a neighborhood  $U$  containing  $p$  for which  $\dim(D(q))$  is constant for all  $q \in U$ . In this case  $D$  is said to be nonsingular on  $U$ . If  $p$  is a regular point of a distribution  $D$  with  $\dim D(p) = d$ , then there exist an open neighborhood  $U^0$  of  $p$  and  $d$  smooth local generators  $f_1, \dots, f_d$  defined on  $U^0$  such that for each  $q \in U^0$ ,  $D(q) = \text{span}\{f_1(q), \dots, f_d(q)\}$ . We will write  $D = \text{span}\{f_1, \dots, f_d\}$  when such a finite set of local generators exists.

A nonsingular distribution can be viewed as a subbundle of  $TM$ . As such, when  $TM$  has an orthogonal structure, we will use the notation  $D^\perp$  to indicate the orthogonal complement of  $D$  in  $TM$ . This stands in contrast to the notation  $\text{ann}(D)$  which

we use to denote the annihilator of  $D$  contained in  $TM^*$ , the cotangent bundle. If  $D$  is a distribution defined on a manifold  $M$  and  $N \subset M$  is a submanifold we will at times consider the subbundles  $TN + D$  and  $TN \cap D$  of  $TM|_N$  defined fiberwise, for each  $p \in N$ , by  $T_p N + D(p)$  and  $T_p N \cap D(p)$ , respectively. The following fact is used in what follows.

LEMMA 2.4. *Let  $N \subset M$  be a  $C^\infty$ -submanifold of the  $C^\infty$ -manifold  $M$ . Let  $p \in N$  be a regular point of a  $C^\infty$ -distribution  $D$  on  $M$ . Suppose there exists an open neighborhood  $V$  in  $N$  such that  $\dim(T_p N \cap D(p))$  is constant for all  $p \in V$ . Then there exists a neighborhood  $U$  of  $p$  in  $V$  such that  $TN \cap D$  and  $(TN \cap D)^\perp$  are smooth in  $U$ .*

Given a smooth distribution  $D$  on  $M$ , we denote by  $\bar{D}$  the involutive closure of  $D$ , i.e., the intersection of all involutive distributions containing  $D$ . We denote by  $\text{Lie}(D)$  the smallest subalgebra of  $\mathcal{V}(M)$  containing the vector fields in  $D$ . We will use  $\text{Lie}_{C^\infty(M)}(D)$  to denote the smooth distribution spanned by vector fields in  $\text{Lie}(D)$ . Then  $\text{Lie}_{C^\infty(M)}(D) \subseteq \bar{D}$ , and generally the inclusion is proper.

If  $\Delta$  and  $\Lambda$  are distributions and  $f$  is a vector field, then we use the following notation:

$$\begin{aligned} [\Delta, \Lambda] &= \text{span}\{[X, Y] : X \in \Delta, Y \in \Lambda\}, \\ [f, \Delta] &= \text{span}\{[f, \tau] : \tau \in \Delta\}. \end{aligned}$$

A smooth distribution  $\Delta$  is invariant under a vector field  $f$  if  $[f, \Delta] \subseteq \Delta$ . A distribution  $\Delta$  defined on an open set  $U$  is said to be locally controlled invariant for the dynamics (2.1) if for each  $x_0 \in U$  there exist a neighborhood  $U^0$  of  $x_0$  and a regular static feedback  $(\alpha, \beta)$  on  $U^0$  such that

$$\begin{aligned} [\tilde{f}, \Delta] &\subseteq \Delta, \\ [\tilde{g}_i, \Delta] &\subseteq \Delta, \quad i \in \{1, \dots, m\}, \end{aligned}$$

where  $\tilde{f} = f + g\alpha$  and  $\tilde{g} = g\beta$ .

THEOREM 2.5 (see [7], [9], [11], [20], [21]). *Let  $\Delta$  be an involutive distribution. Suppose  $\Delta$  and  $\Delta + \text{span}\{g_1, \dots, g_m\}$  are nonsingular on  $U$ . Then  $\Delta$  is locally controlled invariant for the dynamics (2.1) if and only if*

$$\begin{aligned} [f, \Delta] &\subseteq \Delta + \text{span}\{g_1, \dots, g_m\}, \\ [g_i, \Delta] &\subseteq \Delta + \text{span}\{g_1, \dots, g_m\}, \quad i \in \{1, \dots, m\}. \end{aligned}$$

**3. Local transverse feedback linearization problem and solution.** In this section we present the main problem studied in this paper. Suppose we are given a pair  $(\Gamma^*, u^*)$ , where  $\Gamma^* \in \mathcal{A}(f, g, \mathbb{R}^n)$ ,  $\dim \Gamma^* = n^*$ , and  $u^* \in \mathcal{F}(f, g, \Gamma^*)$ . In this presentation we consider the controlled invariant set  $\Gamma^*$  as given data. For example, in a mechanical system this might correspond to a motion planning task being solved in order to obtain the shortest path between two given points. In most situations, however, one is given a set  $\Gamma$ , perhaps defined by virtual constraints or design goals, and then one must pare away pieces of  $\Gamma$  until all which remains is the largest controlled invariant submanifold  $\Gamma^*$  contained in  $\Gamma$ . Various tools in the literature exist for this purpose (see, for instance, the zero dynamics algorithm [9] or the constrained dynamics algorithm [21]). Generally, the existing tools generate a local characterization of  $\Gamma^*$  about the initialization point of the algorithms. In some cases, viability

theory [1] can be used to obtain global characterizations of invariant sets for dynamical systems. The main problem investigated in this paper, stated next, concerns the decomposition of the system dynamics into a subsystem describing the motion on  $\Gamma^*$  and one describing the motion transversal to  $\Gamma^*$ , with the essential requirement that the transversal subsystem be feedback linearizable.

**Local transverse feedback linearization problem.** Given a pair  $(\Gamma^*, u^*) \in \mathcal{S}(f, g, \mathbb{R}^n) \times \mathcal{T}(f, g, \mathbb{R}^n)$  and a point  $p_0 \in \Gamma^*$  find, if possible, a neighborhood  $U$  of  $p_0$  in  $\mathbb{R}^n$ , a transformation  $\Xi \in \text{Diff}(U)$ ,  $\Xi : U \rightarrow \mathbb{R}^{n^*} \times \mathbb{R}^{n-n^*}$ ,  $x \mapsto (z, \xi)$ , and a feedback transformation  $(\alpha, \beta)$ , such that (2.1) is feedback equivalent on  $U$  to

$$(3.1) \quad \begin{aligned} \dot{z} &= f^0(z, \xi) + g^1(z, \xi)v_1 + g^2(z, \xi)v_2, \\ \dot{\xi} &= A\xi + Bv_1, \end{aligned}$$

where  $v = \text{col}(v_1, v_2) \in \mathbb{R}^m$ ;  $B$  is full rank, the pair  $(A, B)$  is controllable, and  $\Xi(\Gamma^* \cap U) = \{(z, \xi) \in \mathbb{R}^{n^*} \times \mathbb{R}^{n-n^*} : \xi = 0\}$ .

In words, we seek to characterize conditions under which (2.1) is feedback equivalent to a system whose dynamics transversal to the set  $\Gamma^*$  are linear, time-invariant, and controllable. LTFLP asks for a coordinate and feedback transformation valid on  $U$  which generates a normal form with two types of decompositions. On the one hand, system dynamics near  $\Gamma^* \cap U$  are decomposed into a tangential subsystem, the  $z$ -dynamics, and a transversal subsystem, the  $\xi$ -dynamics. On the other hand, the original  $m$  control inputs are decomposed into transversal and tangential components  $v_1$  and  $v_2$ , respectively.

The terminology “transverse feedback linearization” should not be mistaken for “transverse linearization,” a technique consisting of the Jacobian linearization of the dynamics transversal to a periodic orbit. In the special case when  $\Gamma^*$  is a periodic orbit, the notions of “transverse feedback linearization” and “transverse linearization” differ similarly to the way that “feedback linearization” around an equilibrium differs from “linearization” around the equilibrium.

Transverse feedback linearization finds application in the stabilization of  $\Gamma^*$ . For, if a transversal controller  $v_1$  is designed that stabilizes  $\xi = 0$ , and the trajectories of the closed-loop system are bounded, then the controller stabilizes  $\Gamma^*$  in original coordinates. If, on the other hand, the trajectories of the closed-loop system are not all bounded, then stabilization of  $\xi = 0$  implies the stabilization of  $\Gamma^*$  in original coordinates if there exists a class- $\mathcal{K}$  function  $\alpha$  such that  $\|\xi(x)\| \geq \alpha(\|x\|_{\Gamma^*})$ , where  $\|x\|_{\Gamma^*}$  is the point-to-set distance of a point  $x$  to the set  $\Gamma^*$ , defined as  $\|x\|_{\Gamma^*} := \inf_{p \in \Gamma^*} \|x - p\|$ . Hereafter, we assume that the preliminary regular feedback  $(u^*, I_m)$  is applied to (2.1) so that  $f|_{\Gamma^*}$  is tangent to  $\Gamma^*$ . Next, we present a technical result which is useful in proving the main theorem.

**THEOREM 3.1.** *LTFLP is solvable if and only if there exist  $\rho_0$  smooth  $\mathbb{R}$ -valued functions  $\alpha_1, \dots, \alpha_{\rho_0}$ , defined on an open neighborhood  $U$  of  $p_0$  in  $\mathbb{R}^n$ , such that*

1.  $U \cap \Gamma^* \subset \{x \in U : \alpha_i(x) = 0, i = 1, \dots, \rho_0\}$ ; and
2. the system

$$(3.2) \quad \begin{aligned} \dot{x} &= f(x) + \sum_{i=1}^m g_i(x)u_i, \\ y' &= \text{col}(\alpha_1(x), \dots, \alpha_{\rho_0}(x)) \end{aligned}$$

has vector relative degree  $\{k_1, \dots, k_{\rho_0}\}$  with  $k_1 + \dots + k_{\rho_0} = n - n^*$  at  $p_0$ .

Moreover, the zero dynamics  $\mathcal{Z}^*$  of (3.2) coincide with  $\Gamma^*$  on  $U$ :  $\mathcal{Z}^* \cap U = \Gamma^* \cap U$ .

We omit the proof of Theorem 3.1 and instead refer the reader to [19, Theorem 4.1] whose proof is identical. System (3.2) has  $m$  inputs and  $\rho_0$  outputs and hence is not square. The notion of vector relative degree of a nonsquare system is the same as that of a square system given in section 5.1 of [9], with the difference that the  $\rho_0 \times m$  decoupling matrix  $A(x)$  with components  $a_{ij}(x) = L_{g_j} L_f^{k_i-1} \alpha_i(x)$  is assumed to be full-rank, rather than nonsingular, at  $p_0$ .

In section 4 we give the coordinate-free definition of transverse controllability indices. It turns out (see Lemma 4.3) that  $\{k_1, \dots, k_{\rho_0}\}$  in Theorem 3.1 are precisely the transverse controllability indices of (2.1) with respect to  $\Gamma^*$ .

Theorem 3.1 characterizes the solvability of LTFLP in terms of the existence of a virtual output function  $\alpha : U \rightarrow \mathbb{R}^{\rho_0}$  satisfying (1) and (2). Once the output function is known, a coordinate and feedback transformation yielding (3.1) is found constructively using [9, Proposition 5.1.2] and additional elementary manipulations. The theorem also shows that LTFLP is equivalent to the *zero dynamics assignment problem with relative degree* mentioned in the introduction. For, the theorem states that LTFLP is solvable if and only if  $\Gamma^*$  can be made into the zero dynamics manifold of (2.1) induced by a suitable output yielding a well-defined vector relative degree. On the other hand, Theorem 3.1 does not give any way of finding the output function or even to determine whether it exists. Hence, it has limited value for constructing the coordinate and feedback transformation.

Consider the distributions

$$(3.3) \quad G_i := \text{span}\{ad_f^j g_k : 0 \leq j \leq i, 1 \leq k \leq m\}$$

and recall from section 2 that  $\bar{G}_i$  denotes the involutive closure of  $G_i$ . Now we give the main result of this paper.

**THEOREM 3.2** (main result). *Suppose that  $\bar{G}_i$ ,  $i \in \mathbf{n} - \mathbf{n}^* - \mathbf{1}$ , are regular at  $p_0 \in \Gamma^*$ . Then, LTFLP is solvable at  $p_0$  if and only if*

- (a)  $\dim(T_{p_0} \Gamma^* + G_{\mathbf{n} - \mathbf{n}^* - \mathbf{1}}(p_0)) = n$ , and
- (b) *there exists an open neighborhood  $U$  of  $p_0$  in  $\mathbb{R}^n$  such that for all  $i \in \mathbf{n} - \mathbf{n}^* - \mathbf{1}$  ( $\forall p \in \Gamma^* \cap U$ )  $\dim(T_p \Gamma^* + G_i(p)) = \dim(T_p \Gamma^* + \bar{G}_i(p)) = \text{constant}$ .*

It is useful to specialize Theorem 3.2 to the case when  $\Gamma^*$  is an equilibrium, because in this special case LTFLP coincides with the state-space exact linearization problem whose solution was given in [8], [12].

**COROLLARY 3.3.** *Assume that  $\Gamma^* = \{p_0\}$  is an equilibrium point of the open-loop system  $\dot{x} = f(x)$  and that  $\bar{G}_i$ ,  $i \in \mathbf{n} - \mathbf{1}$ , are regular at  $p_0$ . Then, LTFLP is solvable at  $p_0$  if and only if*

- (a)  $\dim G_{\mathbf{n} - \mathbf{1}}(p_0) = n$ , and
- (b)'  $G_i$ ,  $i \in \mathbf{n} - \mathbf{1}$ , are involutive and regular at  $p_0$ .

*Proof.* It suffices to show that, under the assumption that the distributions  $\bar{G}_i$  are regular, (b)' is equivalent to condition (b) in Theorem 3.2. Assume that (b) holds, i.e.,  $G_i(p_0) = \bar{G}_i(p_0)$ . For all  $p$  in a neighborhood of  $p_0$ , one has

$$\dim(G_i(p_0)) \leq \dim(G_i(p)) \leq \dim(\bar{G}_i(p)) = \dim(\bar{G}_i(p_0)) = \dim(G_i(p_0)),$$

and so all inequalities above are equalities. Therefore,  $\dim(G_i(p)) = \dim(G_i(p_0))$  and  $\dim(G_i(p)) = \dim(\bar{G}_i(p))$ , proving that (b)  $\implies$  (b)'. The converse implication is obvious.  $\square$

We recall the following classical result.

**THEOREM 3.4** (state-space exact linearization [8], [12]). *Assume that  $\Gamma^* = \{p_0\}$  is an equilibrium point of the open-loop system  $\dot{x} = f(x)$ . Then, LTFLP is solvable at  $p_0$  if and only if conditions (a) and (b)' in Corollary 3.3 hold.*

Note that conditions (a) and (b)' in Corollary 3.3 imply that the distributions  $\bar{G}_i$ ,  $i \in \mathbf{n} - \mathbf{1}$ , are regular at  $p_0$ . Thus, in the special case when  $\Gamma^*$  is an equilibrium point, the conditions of our main result coincide with those of the state-space exact linearization problem. The difference between Corollary 3.3 and Theorem 3.4 is that the former relies on the *preliminary assumption* that the distributions  $\bar{G}_i$  are regular at  $p_0$ , while the latter shows that regularity of  $\bar{G}_i$  at  $p_0$  is actually necessary for the solvability of LTFLP, and hence there is no need to impose it as a preliminary requirement.

The assumptions of Theorem 3.2 are checkable; however, its proof does not provide a constructive procedure for finding the virtual outputs described in Theorem 3.1. The next result sheds additional light on LTFLP by relating it to the partial feedback linearization problem. The result isn't a viable solution to LTFLP because its assumptions are not checkable. On the other hand, the theorem provides guidelines for finding the output function in Theorem 3.1, as discussed below.

**THEOREM 3.5.** *Suppose that  $\bar{G}_i$ ,  $i \in \mathbf{n} - \mathbf{n}^* - \mathbf{1}$ , are regular at  $p_0 \in \Gamma^*$ . Then, LTFLP is solvable at  $p_0$  if and only if there exist a neighborhood  $U$  of  $p_0$  and a smooth, involutive, and regular distribution  $\Delta$  on  $U$  such that*

- (i)  $\Delta|_{\Gamma^*} = T\Gamma^*$ .
- (ii)  $\Delta$  is locally controlled invariant under (2.1).
- (iii)  $(\forall p \in \Gamma^* \cap U) \dim(T_p\Gamma^* + G_{n-n^*-1}(p)) = n$ .
- (iv)  $(\forall i \in \mathbf{n} - \mathbf{n}^* - \mathbf{1}) \Delta + G_i$  is regular and involutive on  $U$ .

*Proof.* Suppose that LTFLP is solvable at  $p_0$ . The necessity of conditions (i)–(iv) can be easily shown by considering the normal form (3.1) and taking

$$(3.4) \quad \Delta = \text{span} \left\{ \frac{\partial}{\partial z_1}, \dots, \frac{\partial}{\partial z_{n^*}} \right\}.$$

Conversely, suppose conditions (i)–(iv) hold. These conditions imply the conditions of [10, Theorem 2.1]. In particular, condition (iv) implies conditions (a) and (b) of [10, Theorem 2.1]. Therefore, by [10, Theorem 2.1] we obtain a system whose dynamics in transformed coordinates is given by (3.1) and where  $\Delta$  is given by (3.4). The integral submanifolds of  $\Delta$  foliate a neighborhood  $U$  of  $p_0$  and are locally given by the sets  $\{(z, \xi) : \xi = \xi^0 = \text{constant}\}$ . Condition (i) means that one of the leaves of the foliation is precisely  $\Gamma^* \cap U$ . Without loss of generality this leaf is taken as the zero level set  $\{(z, \xi) : \xi = 0\}$ .  $\square$

Note that the distribution  $\Delta$  in Theorem 3.5 is not unique. Also note that this theorem involves an interaction between the concepts of controlled invariant distributions and controlled invariant manifolds. Together, Theorems 3.1, 3.2, and 3.5 can be used to find solutions to LTFLP. The following steps outline the typical procedure one may follow in searching for the output function.

1. Represent  $\Gamma^*$  in a neighborhood of  $p_0$  as the zero level set of  $n - n^*$   $\mathbb{R}$ -valued functions. Using Theorem 3.1, check if there exists a subset of  $\rho_0$  of these functions, with  $\rho_0$  defined in (4.1), yielding the correct vector relative degree.
2. If the above step fails, check the conditions of Theorem 3.2 to verify whether or not the problem is solvable. In simple cases, the procedure described in the proof of Theorem 3.2 may yield the desired output functions.
3. If Theorem 3.2 establishes that the problem is solvable, then there exists a

distribution  $\Delta$  satisfying the conditions in Theorem 3.5. If  $\Delta$  is found then, after computing the controllability indices defined in (4.2), the output functions are obtained by finding those exact one-forms that span the codistributions  $\text{ann}(\Delta + G_i)$ ,  $i = 0, \dots, k_1 - 2$ , and arranging them in the order illustrated below, with the integers  $\rho_i$  defined in (4.1).

$\text{ann}(\Delta + G_{k_1-2})$	$d\alpha_1$			
$\vdots$	$\vdots$			
$\text{ann}(\Delta + G_{k_1-2-j})$	$dL_f^j \alpha_1$			
$\vdots$	$\vdots$			
$\text{ann}(\Delta + G_{k_2-2})$	$dL_f^{k_1-k_2} \alpha_1$	$d\alpha_2$		
$\vdots$	$\vdots$	$\vdots$		
$\text{ann}(\Delta + G_{k_2-2-j})$	$dL_f^{k_1-k_2+j} \alpha_1$	$dL_f^j \alpha_2$		
$\vdots$	$\vdots$	$\vdots$		
$\text{ann}(\Delta + G_{k_3-2})$	$dL_f^{k_1-k_3} \alpha_1$	$dL_f^{k_2-k_3} \alpha_2$	$d\alpha_3$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$\text{ann}(\Delta + G_{k_{\rho_0}-2})$	$dL_f^{k_1-k_{\rho_0}} \alpha_1$	$dL_f^{k_2-k_{\rho_0}} \alpha_2$	$\dots$	$d\alpha_{\rho_0}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\text{ann}(\Delta + G_0)$	$dL_f^{k_1-2} \alpha_1$	$dL_f^{k_2-2} \alpha_2$	$\dots$	$dL_f^{k_{\rho_0}-2} \alpha_{\rho_0}$

In each row of the above table, the codistribution in the left column is locally spanned by all the differentials in that row *plus* all the differentials in the rows above. For example, locally we have that

$$\text{ann}(\Delta + G_{k_2-2}) = \text{span}\{d\alpha_1, dL_f \alpha_1, \dots, dL_f^{k_1-k_2} \alpha_1, d\alpha_2\}.$$

The functions  $\alpha_1, \dots, \alpha_{\rho_0}$ , resulting from the integration of the exact one-forms  $d\alpha_1, \dots, d\alpha_{\rho_0}$  along the diagonal of the table are the required outputs. Next we present an example to illustrate the use of Theorems 3.1, 3.2, and 3.5.

*Example.* Consider the system

$$(3.5) \quad \dot{x} = \begin{bmatrix} -x_2 \\ x_1 \\ x_3 x_4 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ x_3 \\ 1 \end{bmatrix} u_1 + \begin{bmatrix} -x_2 \\ x_1 \\ 0 \\ 0 \end{bmatrix} u_2$$

along with the pair  $(\Gamma^*, u^*) \in \mathcal{S}(f, g, \mathbb{R}^4) \times \mathcal{F}(f, g, \mathbb{R}^4)$ ,

$$\begin{aligned} \Gamma^* &= \{x \in \mathbb{R}^4 : x_1^2 + x_2^2 - x_3 = x_4 = 0\}, \\ u^* &= \text{col}(0, 0). \end{aligned}$$

The set  $\Gamma^*$  is an elliptic paraboloid embedded in the subspace  $\{x \in \mathbb{R}^4 : x_4 = 0\}$ . We want to perform TFL of (3.5) with respect to  $\Gamma^*$  near  $p_0 = \text{col}(4, 0, 2, 0)$ . In this example  $n = 4$  and  $n^* = 2$ , so we seek to feedback linearize a subsystem of dimension  $n - n^* = 2$ . The natural approach to solving this problem is to check if one of the two constraints which define  $\Gamma^*$  satisfies the conditions of Theorem 3.1. In this case, both constraints, taken individually as scalar outputs, yield a well-defined



relative degree near  $p_0$  of 1 which does not equal  $n - n^*$ . Taken together, as a vector output, the constraints do not yield a well-defined vector relative degree. In both cases, the conditions of Theorem 3.1 are not satisfied by these constraints. Next, we check whether or not LTFLP is solvable for (3.5) using Theorem 3.2. Checking condition (a) one finds

$$\dim(T_{p_0}\Gamma^* + G_1(p_0)) = 4.$$

Also, since  $[g_1, g_2] = 0$ , it follows that  $G_0 = \bar{G}_0$  everywhere. It is then an easy matter to check that for any  $p \in \Gamma^*$ ,  $\dim(T_p\Gamma^* + G_0(p)) = 3$ . Thus condition (b) holds and LTFLP is solvable despite the fact that the constraints which locally define  $\Gamma^*$  do not satisfy Theorem 3.1.

The fact that Theorem 3.2 holds for this system implies that there exists a distribution  $\Delta$  satisfying Theorem 3.5. After some trial and error, one finds that the distribution

$$\Delta = \text{span} \left\{ \begin{bmatrix} -x_2 \\ x_1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} x_1 \\ x_2 \\ 2x_3 \\ 0 \end{bmatrix} \right\}$$

satisfies the conditions of Theorem 3.5. Then, by the Frobenius theorem, there exists an exact one-form  $d\alpha$  which spans  $\text{ann}(\Delta + G_0)$ . The corresponding function  $\alpha$  is given by

$$\alpha(x) = \ln \left( \frac{x_3}{x_1^2 + x_2^2} \right) - x_4$$

which yields a well-defined relative degree of 2 at  $p_0$  and satisfies Theorem 3.1. As a result, the coordinate transformation  $\Xi: x \mapsto \text{col}(z_1, z_2, \xi_1, \xi_2)$  defined as  $z_1 = x_1$ ,  $z_2 = x_2$ ,  $\xi_1 = \alpha$ ,  $\xi_2 = L_f\alpha = x_4$ , directly yields the normal form (3.1), with  $v_1 = u_1$  and  $v_2 = u_2$ . The coordinate transformation  $\Xi(x)$  is valid on  $\mathbb{R}^4 \setminus (\{x \in \mathbb{R}^4 : x_1 = x_2 = 0\} \cup \{x \in \mathbb{R}^4 : x_3 = 0\})$ .

**4. Transverse controllability indices and preliminary results.** In linear systems theory, controllability indices (see [4] and [29]) describe certain properties which are invariant under coordinate and nonsingular feedback transformations and serve to categorize controllable linear systems. Controllability indices have been ported to the nonlinear setting. They have been used to characterize the largest feedback linearizable subsystem of a nonlinear system [15], and conditions under which a system is feedback linearizable; see [17], [21]. Here, we adapt these ideas to the framework of TFL. Let  $V$  be an open subset of  $\Gamma^*$ . For each  $p \in V$ , let

$$(4.1) \quad \begin{aligned} \rho_0(p) &:= \dim(T_p\Gamma^* + G_0(p)) - n^*, \\ \rho_i(p) &:= \dim(T_p\Gamma^* + \bar{G}_{i-1}(p) + \text{ad}_f^i G_0(p)) - \dim(T_p\Gamma^* + \bar{G}_{i-1}(p)), \end{aligned}$$

$i = 1, 2, \dots$ , where  $G_i$  are as defined in (3.3), and

$$\text{ad}_f^i G_0 = \text{span}\{\text{ad}_f^i X : X \in G_0\}, \quad i = 0, 1, \dots$$

Geometrically, at each  $p \in \Gamma^*$ , the integers  $\rho_i(p)$  represent the number of linearly independent vectors in  $\text{ad}_f^i G_0(p)$  which are not in  $T_p\Gamma^* + \bar{G}_{i-1}(p)$ . Associated to the

list  $\{\rho_0(p), \dots, \rho_i(p), \dots\}$  is a set of  $\rho_0(p)$  integers,  $\{k_1(p), \dots, k_{\rho_0(p)}\}$ , which we refer to as the *transverse controllability indices of (2.1) with respect to  $\Gamma^*$* , defined as (we omit the argument  $p$ )

$$(4.2) \quad k_i := \text{card} \{\rho_j \geq i, j \geq 0\}, \quad i \in \{1, \dots, \rho_0\}.$$

Note that  $k_1 \geq k_2 \geq \dots \geq k_{\rho_0}$ . We show in Corollary 4.2 that the transverse controllability indices are invariant under coordinate and feedback transformations.

Condition (b) of Theorem 3.2 implies that  $\rho_0, \rho_1, \dots, \rho_{n-n^*-2} = \text{constant}$ , while condition (a) implies that  $\sum_i \rho_i = n - n^*$ . In the special case when  $\Gamma^*$  is an equilibrium point, it is useful to compare our definition of controllability indices with the definition by Marino in [15]. Marino's definition relies on the distributions

$$\begin{aligned} \mathcal{G}_f &= f + G_0 = \{f + g : g \in G_0\}, \\ \mathcal{G}_i &= \mathcal{G}_{i-1} + [\mathcal{G}_f, \mathcal{G}_{i-1}], \quad \mathcal{G}_0 = G_0, \quad i = 1, 2, \dots, \\ \mathcal{S}_i &= \bar{\mathcal{G}}_{i-1} + \text{ad}_f^i G_0, \quad \mathcal{S}_0 = G_0, \quad i = 1, 2, \dots, \end{aligned}$$

and uses the integers

$$\begin{aligned} r_0 &= \dim \mathcal{G}_0, \\ r_i &= \dim \mathcal{S}_i - \dim \bar{\mathcal{G}}_{i-1} \end{aligned}$$

in place of the integers  $\rho_i$  in the definition of controllability indices. We now show that, when  $\Gamma^*$  is an equilibrium point, the integers  $\rho_i$  and  $r_i$  are identical, and thus the notion of transverse controllability indices reduces to the classical notion of controllability indices.

LEMMA 4.1. *For all nonnegative integers  $i$ ,  $\bar{G}_i = \bar{\mathcal{G}}_i$ . Thus, when  $\Gamma^* = \{p_0\}$  is an equilibrium point,  $\rho_i = r_i$ .*

*Proof.* By definition,  $G_0 = \mathcal{G}_0$ , so the lemma trivially holds for  $i = 0$ . We now show that  $\bar{G}_i \subseteq \bar{\mathcal{G}}_i$  for all  $i \in \mathbb{N}$ . By definition,

$$\mathcal{G}_i = \mathcal{G}_{i-1} + [\mathcal{G}_f, \mathcal{G}_{i-1}].$$

Since  $f \in \mathcal{G}_f$ , it follows that  $G_i \subseteq \mathcal{G}_i$  for all nonnegative integers  $i$ , which implies  $\bar{G}_i \subseteq \bar{\mathcal{G}}_i$ .

Next, we show that  $\mathcal{G}_i \subseteq \bar{G}_i$  for all  $i \in \mathbb{N}$  which implies  $\bar{\mathcal{G}}_i \subseteq \bar{G}_i$ . To this end, it suffices to prove that  $\mathcal{G}_i \subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$  since  $\text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i) \subseteq \bar{G}_i$ . It is obvious that  $\mathcal{G}_0 \subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_0)$  and

$$\begin{aligned} \mathcal{G}_1 &= G_0 + [\mathcal{G}_f, G_0] \\ &= \text{span}\{g_1, \dots, g_m, \text{ad}_f g_1, \dots, \text{ad}_f g_m, [g, g_1], \dots, [g, g_m] : g \in G_0\} \\ &\subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_1). \end{aligned}$$

For the induction, assume that, for some positive integer  $I \geq 2$ ,

$$\mathcal{G}_{i-1} \subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_{i-1}), \quad i \in \{2, \dots, I\}.$$

We must show that  $\mathcal{G}_i = \mathcal{G}_{i-1} + [\mathcal{G}_f, \mathcal{G}_{i-1}] \subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$  for  $i \in \{2, \dots, I\}$ . It is enough to prove that  $[\mathcal{G}_f, \mathcal{G}_{i-1}] = [f, \mathcal{G}_{i-1}] + [G_0, \mathcal{G}_{i-1}] \subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$ . However, since  $[G_0, \mathcal{G}_{i-1}] \subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_{i-1}) \subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$ , all we are left to show is that  $[f, \mathcal{G}_{i-1}] \subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$ .

Let  $\tau_1 = g_1$ ,  $\tau_2 = g_2, \dots, \tau_{im-1} = ad_f^{i-1} g_{m-1}$ ,  $\tau_{im} = ad_f^{i-1} g_m$ . Then a general vector field in  $\text{Lie}_{C^\infty(\mathbb{R}^n)}(G_{i-1})$  is a  $C^\infty(\mathbb{R}^n)$ -linear combination of vector fields of the form

$$(4.3) \quad \vartheta = [\tau_{j_k}, [\tau_{j_{k-1}}, \dots, [\tau_{j_2}, \tau_{j_1}]]],$$

$1 \leq j_k \leq im$ ,  $1 \leq k < \infty$ . By assumption, any vector field in  $\mathcal{G}_{i-1}$  can also be expressed in this way. Take any vector field  $h \in \mathcal{G}_{i-1}$  and consider

$$[f, h] = \left[ f, \sum_{i \in \mathcal{I}} c_i \vartheta_i \right] = \sum_{i \in \mathcal{I}} [f, c_i \vartheta_i],$$

where  $\mathcal{I}$  is some finite index set,  $c_i \in C^\infty(\mathbb{R}^n)$ , and  $\vartheta_i$  are of the form (4.3). Each term in the above summation can be expressed as  $[f, c_i \vartheta_i] = c_i[f, \vartheta_i] + (L_f c_i) \vartheta_i$ , so it is enough to show that  $[f, \vartheta] \in \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$ , where  $\vartheta$  is of the form (4.3).

When  $k = 1$ , i.e.,  $\vartheta = \tau_{j_1}$ , then  $[f, \tau_{j_1}] \in G_i \subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$ . Next assume that

$$\vartheta = [\tau_{j_{k-1}}, [\tau_{j_{k-2}}, \dots, [\tau_2, \tau_1]]]$$

is such that  $[f, \vartheta] \in \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$ . We will show that  $[f, [\tau_{j_k}, \vartheta]] \in \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$ . Clearly,  $[\tau_{j_k}, \vartheta] \in \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_{i-1})$  for any  $1 \leq j_k \leq im$ . By the Jacobi identity,

$$[f, [\tau_{j_k}, \vartheta]] = [[\vartheta, f], \tau_{j_k}] + [[f, \tau_{j_k}], \vartheta],$$

and since  $[\vartheta, f] \in \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$  and  $\tau_{j_k} \in G_{i-1}$ , it follows that  $[[\vartheta, f], \tau_{j_k}] \in \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$ . Also,  $[f, \tau_{j_k}] \in G_i$  so that  $[[f, \tau_{j_k}], \vartheta] \in \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i)$ . This induction argument shows that  $[f, \mathcal{G}_{i-1}] \subseteq \text{Lie}_{C^\infty(\mathbb{R}^n)}(G_i) \subseteq G_i$ , as required.  $\square$

**COROLLARY 4.2.** *The transverse controllability indices of system (2.1) with respect to a set  $\Gamma^*$  are invariant under coordinate and feedback transformations.*

*Proof.* The push-forward map  $F_*$  associated with any  $F \in \text{Diff}(U)$  is an isomorphism at each  $p \in \Gamma^* \cap U$ . It follows from the definition of the integers  $\rho_0, \dots, \rho_i, \dots$  that they do not change under coordinate transformations. By Lemma 4.1,

$$\rho_i(p) = \dim(T_p \Gamma^* + \mathcal{S}_i(p)) - \dim(T_p \Gamma^* + \bar{\mathcal{G}}_{i-1}(p)).$$

In [15, Proposition 2] it is shown that  $\mathcal{S}_i$  and  $\bar{\mathcal{G}}_{i-1}$  are feedback invariant, and so the integers  $\rho_0, \dots, \rho_i, \dots$  are also invariant under feedback transformations.  $\square$

**LEMMA 4.3.** *Suppose that LTFLP is solvable at  $p_0 \in \Gamma^*$ . Then the transverse controllability indices of (2.1) with respect to  $\Gamma^*$  coincide with the controllability indices of  $(A, B)$  in (3.1).*

*Proof.* This lemma will be proved by direct calculation of the integers  $\rho_i$  in  $(z, \xi)$  coordinates. Let  $V = \Xi(\Gamma^* \cap U)$ . By the properties of the normal form (3.1), for any  $\tilde{p} \in \Gamma^* \cap U$ ,  $\Xi(\tilde{p}) = \text{col}(p, 0)$ . Hence, in  $(z, \xi)$  coordinates we have that for any  $p \in V$  and any  $i \in \mathbf{n} - \mathbf{n}^*$ ,

$$(4.4) \quad T_p V + G_i(\text{col}(p, 0)) = \text{Im} \left( \begin{bmatrix} I_{n^*} & \star & \star & \cdots & \star \\ 0_{n-n^* \times n^*} & B & AB & \cdots & A^i B \end{bmatrix} \right).$$

In  $(z, \xi)$  coordinates, consider the collection of constant distributions  $\Delta_i$ ,  $i \in \mathbf{n} - \mathbf{n}^*$ , given by

$$\Delta_i = \text{Im} (I_{n^*} \oplus [B \quad AB \quad \cdots \quad A^i B]).$$

At each  $p \in V$ ,  $\Delta_i(p) = T_p V + G_i(\text{col}(p, 0))$ . Furthermore, since each  $\Delta_i$  is (trivially) involutive and  $G_i|_V \subset \Delta_i$ , it follows that  $\bar{G}_i|_V \subseteq \Delta_i$ . This shows that for all  $i \in \mathbf{n} - \mathbf{n}^*$ ,

$$TV + \bar{G}_i \subseteq \Delta_i = TV + G_i.$$

On the other hand,  $TV + G_i \subseteq TV + \bar{G}_i$  always holds, and so we have shown that  $\Delta_i = TV + G_i = TV + \bar{G}_i$ . Calculating the integers  $\rho_i$ , we have

$$\begin{aligned} \rho_i(p) &= \dim(T_p \Gamma^* + \bar{G}_{i-1}(p) + \text{ad}_{\tilde{f}}^i G_0(p)) - \dim(T_p \Gamma^* + \bar{G}_{i-1}(p)) \\ &= \dim(T_p \Gamma^* + G_{i-1}(p) + \text{ad}_{\tilde{f}}^i G_0(p)) - \dim(T_p \Gamma^* + G_{i-1}(p)) \\ &= \dim(T_p \Gamma^* + G_i(p)) - \dim(T_p \Gamma^* + G_{i-1}(p)) \\ &= \text{rank}(\Delta_i) - \text{rank}(\Delta_{i-1}) \\ &= \text{rank}([B \ \cdots \ A^i B]) - \text{rank}([B \ \cdots \ A^{i-1} B]). \end{aligned}$$

The claim follows from the definition of the integers  $\{k_1, \dots, k_{\rho_0}\}$ .  $\square$

When the transverse controllability indices are constant on an open subset of  $\Gamma^*$  and the distributions  $\bar{G}_i$  are regular, the next two lemmas establish the existence of a feedback transformation yielding a particularly useful set of local generators for each  $\bar{G}_i$ .

LEMMA 4.4. *Let  $\tilde{U}$  be an open subset of  $\mathbb{R}^n$  such that  $\tilde{V} := \tilde{U} \cap \Gamma^* \neq \emptyset$ . Assume that, for all  $i \in \mathbf{n} - \mathbf{n}^*$ ,*

$$\begin{aligned} (\forall p \in \tilde{V}) \quad \dim(T_p \Gamma^* + G_i(p)) &= \dim(T_p \Gamma^* + \bar{G}_i(p)) = \text{constant}, \\ (\forall p \in \tilde{U}) \quad \dim(\bar{G}_i(p)) &= \nu_i = \text{constant}. \end{aligned}$$

*Then,  $\rho_0 \geq \rho_1 \geq \dots \geq \rho_{n-n^*-1}$  and there exist an open set  $U \subseteq \tilde{U}$  and a regular static feedback  $(\alpha, \beta)$  on  $U$  such that, letting  $V := U \cap \Gamma^*$ , for all  $p \in V$  and for all  $i \in \mathbf{n} - \mathbf{n}^*$ , the following holds:*

$$(4.5) \quad T_p \Gamma^* + \bar{G}_i(p) = T_p \Gamma^* \oplus \left( \bigoplus_{j=0}^i \text{span} \left\{ \text{ad}_{\tilde{f}}^j \tilde{g}_k : 1 \leq k \leq \rho_j \right\} (p) \right).$$

*Proof.* Choose an open set  $U \subseteq \tilde{U}$  such that  $V := U \cap \Gamma^* \neq \emptyset$  and such that  $V$  is covered by a coordinate chart in the atlas of  $\Gamma^*$ . Apply the preliminary feedback transformation  $(u^*, I_m)$  defined on  $V$ . Let  $\tilde{f} = f + gu^*$ . On  $V$ , define the distribution (i.e., a subbundle of  $T\mathbb{R}^n|_V$  defined using the natural orthogonal structure on  $\mathbb{R}^n$ )

$$\mathcal{G}_0 = [\bar{G}_0 \cap TV]^\perp \cap \bar{G}_0.$$

On  $V$ ,  $\bar{G}_0 \cap TV$  is constant dimensional since

$$\dim(\bar{G}_0 \cap TV) = \dim(TV) + \dim(\bar{G}_0) - \dim(TV + \bar{G}_0).$$

Since  $\bar{G}_0$  and  $TV$  are regular distributions and their intersection is constant dimensional, it follows from Lemma 2.4 that, by possibly shrinking  $U$  (and hence  $V$ ),  $\bar{G}_0 \cap TV$  is smooth, and so too is  $[\bar{G}_0 \cap TV]^\perp$ . Thus,  $\mathcal{G}_0$  is the intersection of smooth, regular distributions. Furthermore,  $\mathcal{G}_0$  has constant dimension on  $V$  since, for each  $p \in V$ ,

$$\begin{aligned} \dim(\mathcal{G}_0(p)) &= n - \dim(\bar{G}_0(p) \cap T_p V) + \dim(\bar{G}_0(p)) - \dim([\bar{G}_0(p) \cap T_p V]^\perp + \bar{G}_0(p)) \\ &= \dim(\bar{G}_0(p)) - \dim(\bar{G}_0(p) \cap T_p V) \\ &= \rho_0. \end{aligned}$$

Since  $\mathcal{G}_0 \subseteq \bar{G}_0$  and  $\mathcal{G}_0 \cap TV = (\bar{G}_0^\perp + TV^\perp) \cap (\bar{G}_0^\perp + TV^\perp)^\perp = 0$ , we have

$$(\forall p \in V) \ T_p V \oplus \mathcal{G}_0(p) = T_p V + \bar{G}_0(p) = T_p V + G_0(p).$$

By construction,  $V$  is covered by a coordinate chart, so there exist  $n^*$  vector fields on  $V$  such that at each  $p \in V$ ,  $T_p V = \text{span}\{v_1, \dots, v_{n^*}\}(p)$ . Moreover, by possibly shrinking  $U$  (and hence  $V$ ), there exist  $\rho_0$  vector fields  $w_1, \dots, w_{\rho_0} : V \rightarrow T\mathbb{R}^n|_V$  such that  $\mathcal{G}_0 = \text{span}\{w_1, \dots, w_{\rho_0}\}$  so that, on  $V$ ,

$$TV \oplus \mathcal{G}_0 = \text{span}\{v_1, \dots, v_{n^*}\} \oplus \text{span}\{w_1, \dots, w_{\rho_0}\}.$$

Using the fact that  $\mathcal{G}_0 \subset TV + G_0$ , we write

$$(4.6) \quad w_j = \sum_{k=1}^{n^*} \alpha_k^j v_k + \sum_{k=1}^m \beta_k^j g_k, \quad j = 1, \dots, \rho_0,$$

where  $\alpha_k^j : V \rightarrow \mathbb{R}$ ,  $\beta_k^j : V \rightarrow \mathbb{R}$  are  $C^\infty(V)$  functions. Let  $\beta_0$  be the  $m \times \rho_0$  matrix of real-valued functions whose  $(k, j)$ th element is  $\beta_k^j$  and let

$$\begin{bmatrix} \tilde{g}_1 & \cdots & \tilde{g}_{\rho_0} \end{bmatrix} = \begin{bmatrix} g_1 & \cdots & g_m \end{bmatrix} \beta_0.$$

We now show that  $\tilde{g}_1, \dots, \tilde{g}_{\rho_0}$  are linearly independent, which implies that  $\beta_0$  is full rank. Suppose there exist  $\rho_0$  functions  $c_i \in C^\infty(V)$  such that  $\sum_{i=1}^{\rho_0} c_i \tilde{g}_i = 0$ . Then by (4.6),  $\sum_{i=1}^{\rho_0} c_i w_i \in TV$ , which implies  $c_i = 0$ ,  $i = 1, \dots, \rho_0$ , since  $\mathcal{G}_0 \cap TV = 0$ . Note that this argument also shows that  $\text{span}\{\tilde{g}_1, \dots, \tilde{g}_{\rho_0}\} \cap TV = 0$ . For if this were false, then there would exist a linear combination of the  $w_i$ 's in (4.6) which, pointwise, belongs to  $T_p V$ .

Next, we seek  $m - \rho_0$  vector fields  $\tilde{g}_{\rho_0+1}, \dots, \tilde{g}_m$  which belong to  $TV$  and such that, on  $V$ ,  $\text{span}\{\tilde{g}_1, \dots, \tilde{g}_m\} = G_0$ . By possibly shrinking  $U$  (and hence  $V$ ), there exists a set of smooth local generators,  $\tilde{g}_{\rho_0+1}, \dots, \tilde{g}_m$ , for  $G_0 \cap TV$ . We now have the desired decomposition on  $V$ ,

$$TV + G_0 = TV + \bar{G}_0 = TV \oplus \text{span}\{\tilde{g}_1, \dots, \tilde{g}_{\rho_0}\},$$

where, in the new basis for  $G_0$ , at each  $p \in V$ ,

$$(4.7) \quad \begin{aligned} \tilde{g}_1(p), \dots, \tilde{g}_{\rho_0}(p) &\in G_0(p), \\ \tilde{g}_{\rho_0+1}(p), \dots, \tilde{g}_m(p) &\in T_p V. \end{aligned}$$

On  $V$ ,  $\tilde{f}(p) \in T_p V$ , so we have that  $ad_{\tilde{f}}^j \tilde{g}_k \in TV + \bar{G}_{j-1}$ ,  $\rho_0 + 1 \leq k \leq m$ ,  $j = 0, 1, \dots$ , and hence  $\rho_0 \geq \rho_1, \dots, \rho_{n-n^*-1}$ .

Now we perform the induction step. Assume that, for some positive integer  $I$ , and any  $i \in \{0, \dots, I\}$ , there exists a basis  $\{\hat{g}_1, \dots, \hat{g}_m\}$  for  $G_0$  such that

$$\begin{aligned} \text{(a)} \quad TV + \bar{G}_{i-1} &= TV \oplus \left( \bigoplus_{j=0}^{i-1} \text{span} \left\{ ad_{\tilde{f}}^j \hat{g}_k : 1 \leq k \leq \rho_j \right\} \right), \\ \text{(b)} \quad (\forall k \in \{\rho_{i-1} + 1, \dots, m\}) \quad ad_{\tilde{f}}^{i-1} \hat{g}_k &\in TV + \bar{G}_{i-2}. \end{aligned}$$

Property (b) implies that  $\rho_{i-1} \geq \rho_i, \dots, \rho_{n-n^*-1}$ . We now seek a basis  $\{\tilde{g}_1, \dots, \tilde{g}_m\}$  for  $G_0$  such that for any  $i \in \{0, \dots, I\}$ ,

$$\begin{aligned} \text{(a)'} \quad TV + \bar{G}_i &= TV \oplus \left( \bigoplus_{j=0}^i \text{span}\{ad_{\tilde{f}}^j \tilde{g}_k : 1 \leq k \leq \rho_j\} \right), \\ \text{(b)'} \quad (\forall k \in \{\rho_i + 1, \dots, m\}) \quad ad_{\tilde{f}}^i \tilde{g}_k &\in TV + \bar{G}_{i-1}. \end{aligned}$$

On  $V$ , define the distribution

$$\mathcal{G}_i = [\bar{G}_i \cap (TV + \bar{G}_{i-1})]^\perp \cap \bar{G}_i.$$

Note that  $\bar{G}_i \cap (TV + \bar{G}_{i-1})$  is constant dimensional since

$$\dim(\bar{G}_i \cap (TV + \bar{G}_{i-1})) = \dim(\bar{G}_i) + \dim(TV + \bar{G}_{i-1}) - \dim(TV + \bar{G}_i).$$

The distribution  $\bar{G}_i$  is regular on  $U$  and  $TV + \bar{G}_{i-1}$  is constant dimensional on  $V$ . Since their intersection is constant dimensional, it follows from Lemma 2.4 that the orthogonal complement of their intersection is smooth, and thus  $\mathcal{G}_i$ , being the intersection of two smooth and regular distributions, is smooth. Furthermore,

$$\begin{aligned} \dim \mathcal{G}_i &= n - \dim(\bar{G}_i) - \dim(TV + \bar{G}_{i-1}) + \dim(TV + \bar{G}_i) \\ &\quad + \dim(\bar{G}_i) - \dim([\bar{G}_i \cap (TV + \bar{G}_{i-1})]^\perp + \bar{G}_i) \\ &= \dim(TV + \bar{G}_i) - \dim(TV + \bar{G}_{i-1}) \\ &= \rho_i. \end{aligned}$$

By construction,  $\mathcal{G}_i \subseteq \bar{G}_i$  and  $\mathcal{G}_i \cap (TV + \bar{G}_{i-1}) = 0$ , so by dimensionality we have that

$$(4.8) \quad (TV + \bar{G}_{i-1}) \oplus \mathcal{G}_i = TV + \bar{G}_i = TV + G_i.$$

By possibly shrinking  $U$  (and hence  $V$ ), there exist  $\rho_i$  smooth vector fields  $w_1, \dots, w_{\rho_i}$  such that on  $V$ ,  $\mathcal{G}_i = \text{span}\{w_1, \dots, w_{\rho_i}\}$ . Hence, by (4.8) we can write

$$(4.9) \quad w_j = \bar{w} + \sum_{k=1}^{\rho_{i-1}} \beta_k^j ad_{\tilde{f}}^i \hat{g}_k + \sum_{k=\rho_{i-1}+1}^m \beta_k^j ad_{\tilde{f}}^i \hat{g}_k, \quad j \in \{1, \dots, \rho_i\},$$

where  $\bar{w} \in TV + \bar{G}_{i-1}$  and each  $\beta_k^j : V \rightarrow \mathbb{R}$  is a  $C^\infty(V)$  function. By property (b), for all  $k \in \{\rho_{i-1} + 1, \dots, m\}$ ,  $ad_{\tilde{f}}^i \hat{g}_k \in TV + \bar{G}_{i-1}$ . Let

$$(4.10) \quad \hat{w}_j := \sum_{k=1}^{\rho_{i-1}} \beta_k^j ad_{\tilde{f}}^i \hat{g}_k, \quad j = 1, \dots, \rho_i.$$

Notice that  $\text{span}\{\hat{w}_1, \dots, \hat{w}_{\rho_i}\} \cap (TV + \bar{G}_{i-1}) = 0$ . For if this were false, then there would exist a  $C^\infty(V)$ -linear combination of the  $w_j$  belonging to  $TV + \bar{G}_{i-1}$  which, by (4.8), is not possible. Furthermore,  $\hat{w}_1, \dots, \hat{w}_{\rho_i}$  are linearly independent because if there exist  $\rho_i$  functions  $c_i \in C^\infty(V)$  such that on  $V$   $c_1 \hat{w}_1 + \dots + c_{\rho_i} \hat{w}_{\rho_i} = 0$ , then, for some  $w^* \in TV + \bar{G}_{i-1}$ ,  $c_1 w_1 + \dots + c_{\rho_i} w_{\rho_i} - w^* = 0$ , thus  $c_1 w_1 + \dots + c_{\rho_i} w_{\rho_i} = 0$ ,

implying  $c_1 = \cdots = c_{\rho_i} = 0$ . Now let  $\beta_i$  be the  $\rho_{i-1} \times \rho_i$  matrix of smooth functions whose  $(k, j)$ th element is  $\beta_k^j$  obtained from (4.10) so that

$$\begin{bmatrix} \hat{w}_1 & \cdots & \hat{w}_{\rho_i} \end{bmatrix} = \begin{bmatrix} ad_{\tilde{f}}^i \hat{g}_1 & \cdots & ad_{\tilde{f}}^i \hat{g}_{\rho_{i-1}} \end{bmatrix} \beta_i.$$

The vector fields  $\hat{w}_1, \dots, \hat{w}_{\rho_i}$  are linearly independent and are generated as the image of  $\rho_{i-1}$  linearly independent vector fields under  $\beta_i$ . Therefore  $\beta_i$  is full rank. We can now write

$$TV + \bar{G}_i = (TV + \bar{G}_{i-1}) \oplus \text{span}\{\hat{w}_1, \dots, \hat{w}_{\rho_i}\}.$$

Let

$$\begin{bmatrix} \tilde{g}_1 & \cdots & \tilde{g}_{\rho_i} \end{bmatrix} = \begin{bmatrix} \hat{g}_1 & \cdots & \hat{g}_{\rho_{i-1}} \end{bmatrix} \beta_i.$$

The vector fields  $\{\tilde{g}_1, \dots, \tilde{g}_{\rho_i}\}$  are linearly independent because the vector fields  $\{\hat{g}_1, \dots, \hat{g}_{\rho_{i-1}}\}$  are linearly independent and  $\beta_i$  is full rank. Moreover, since  $\text{span}\{\tilde{g}_1, \dots, \tilde{g}_{\rho_i}\} \subseteq \text{span}\{\hat{g}_1, \dots, \hat{g}_{\rho_{i-1}}\}$  and both are constant dimensional, we can find (making  $U$ , and hence  $V$ , smaller if necessary)  $\rho_{i-1} - \rho_i$  vector fields  $\tilde{g}_{\rho_{i-1}+1}, \dots, \tilde{g}_{\rho_{i-1}}$  such that  $\text{span}\{\tilde{g}_1, \dots, \tilde{g}_{\rho_{i-1}}\} = \text{span}\{\hat{g}_1, \dots, \hat{g}_{\rho_{i-1}}\}$  (hence preserving property (a) from the induction assumption) and  $ad_{\tilde{f}}^i \tilde{g}_{\rho_{i-1}+1}, \dots, ad_{\tilde{f}}^i \tilde{g}_{\rho_{i-1}} \in TV + \bar{G}_{i-1}$ . This is done by finding local generators for the smooth distribution

$$\text{span}\{\tilde{g}_1, \dots, \tilde{g}_{\rho_i}\}^\perp \cap \text{span}\{\hat{g}_1, \dots, \hat{g}_{\rho_{i-1}}\},$$

which has constant dimension  $\rho_{i-1} - \rho_i$ . Finally, let  $\tilde{g}_{\rho_{i-1}+1} = \hat{g}_{\rho_{i-1}+1}, \dots, \tilde{g}_m = \hat{g}_m$ . We have therefore obtained a basis  $\{\tilde{g}_1, \dots, \tilde{g}_m\}$  for  $G_0$  in which properties (a)' and (b)' hold. In summary, the induction process gives a basis for  $G_0$  in which the input vector fields are arranged in such a way that for  $i \in \mathbf{n} - \mathbf{n}^*$ ,

$$(TV + \bar{G}_{i-1} + ad_{\tilde{f}}^i G_0) / (TV + \bar{G}_{i-1}) \simeq \text{span}\{ad_{\tilde{f}}^i \tilde{g}_1, \dots, ad_{\tilde{f}}^i \tilde{g}_{\rho_i}\}.$$

We are left to show that this arrangement can be achieved using a regular static feedback and that the arrangement is valid on  $U$ , and not just on  $V$ , as is presently the case. To this end, let  $\tilde{g} = [\tilde{g}_1 \cdots \tilde{g}_m]$  and define a regular static feedback defined on  $V$  by  $(\hat{\alpha}, \hat{\beta})$ , where  $\hat{\alpha} = u^*$  and  $\hat{\beta} = (g^\top g)^{-1} g^\top \tilde{g}$ . To obtain a feedback transformation defined off of  $\Gamma^*$ , we can, by possibly shrinking  $U$  (and hence  $V$ ) and applying Lemma 2.3, introduce a retraction  $r : U \rightarrow V$  of  $U$  onto  $V$ . Then, let  $\alpha = \hat{\alpha} \circ r$  and  $\beta = \hat{\beta} \circ r$ . The regular static feedback  $(\alpha, \beta)$  has the desired properties.  $\square$

In order to identify directions in the intersection  $T_p V \cap \bar{G}_i(p)$  which are not contained in the intersection  $T_p V \cap \bar{G}_{i-1}(p)$ , it is useful to define the integers

$$\begin{aligned} \mu_0(p) &:= \dim(T_p V \cap \bar{G}_0(p)), \\ \mu_i(p) &:= \dim(T_p V \cap \bar{G}_i(p)) - \dim(T_p V \cap \bar{G}_{i-1}(p)) \end{aligned}$$

for  $i \in \mathbb{N}$ , and let

$$n_i(p) := \sum_{j=0}^i \mu_j(p),$$

so that  $\dim(T_p V \cap \bar{G}_i(p)) = n_i(p)$ . Under the assumption of Lemma 4.4, we have that  $\dim(T_p V \cap \bar{G}_i(p)) = n^* + \nu_i - \dim(T_p V + \bar{G}_i(p))$ , and hence the  $\mu_i$  are constant for all  $i \in \mathbf{n} - \mathbf{n}^*$ , and we have the following result.

LEMMA 4.5. *Let  $\tilde{U}$  be an open subset of  $\mathbb{R}^n$  such that  $\tilde{V} := \tilde{U} \cap \Gamma^* \neq \emptyset$ . Assume that, for all  $i \in \mathbf{n} - \mathbf{n}^*$ , the conditions of Lemma 4.4 hold. Then, there exists an open set  $U \subseteq \tilde{U}$  and  $n_{n-\mathbf{n}^*-1}$  vector fields  $v_\ell^j \in \mathbf{V}(U)$ ,  $0 \leq j \leq n - \mathbf{n}^* - 1$  such that, after the feedback transformation of Lemma 4.4, letting  $V := U \cap \Gamma^*$ , and*

$$G_i^\parallel := \text{span}\{v_1^0, \dots, v_{\mu_0}^0, \dots, v_1^i, \dots, v_{\mu_i}^i\}$$

*one has that, for all  $i \in \mathbf{n} - \mathbf{n}^*$ , on  $U$ ,*

$$\bar{G}_i = G_i^\parallel \oplus \left( \bigoplus_{j=0}^i \text{span} \left\{ \text{ad}_{\tilde{f}}^j \tilde{g}_k : 1 \leq k \leq \rho_j \right\} \right)$$

and  $G_i^\parallel|_V = TV \cap \bar{G}_i$ .

*Proof.* Suppose the feedback transformation  $(\alpha, \beta)$  of Lemma 4.4 has been applied, which is valid on  $U \subseteq \tilde{U}$  as defined therein. Since every point  $p$  of  $U$  is a regular point for the distributions  $\bar{G}_i$ , we can, by possibly shrinking  $U$  (and hence  $V$ ), find a set of local generators  $X_1^i, \dots, X_{\nu_i}^i$  valid on  $U$  for  $\bar{G}_i$ ,  $i \in \mathbf{n} - \mathbf{n}^*$ .

On  $V$ , define the distribution  $Q_0 = TV \cap \bar{G}_0$ . By assumption,  $Q_0$  has constant dimension  $\mu_0$ . Moreover, since  $Q_0$  is the intersection of two smooth, regular distributions and is constant dimensional, it is, by Lemma 2.4, smooth. By shrinking  $U$  (and hence  $V$ ) we can find a basis such that, on  $V$ ,  $Q_0 = \text{span}\{\hat{v}_1, \dots, \hat{v}_{\mu_0}\}$ . By construction,  $Q_0 \subset \bar{G}_0$  so that each  $\hat{v}_k \in Q_0$  can be expressed as

$$\hat{v}_k = \sum_{j=1}^{\nu_0} \hat{c}_{j0}^k X_j^0, \quad k \in \{1, \dots, \mu_0\},$$

where each  $\hat{c}_{j0}^k : V \rightarrow \mathbb{R}$  is a  $C^\infty(V)$  function. Next, we apply Lemma 2.3 and, by possibly shrinking  $U$  (and hence  $V$ ), introduce a retraction  $r : U \rightarrow V$  of  $U$  onto  $V$ . Let  $c_{j0}^k = \hat{c}_{j0}^k \circ r$  so that

$$v_k^0 := \sum_{i=1}^{\nu_0} c_{i0}^k X_i^0, \quad k \in \{1, \dots, \mu_0\},$$

are now vector fields defined on  $U$ , and let  $G_0^\parallel := \text{span}\{v_1^0, \dots, v_{\mu_0}^0\}$ . It follows that  $G_0^\parallel \subset \bar{G}_0$  in  $U$  and  $G_0^\parallel|_V = Q_0$ . By Lemma 4.4,  $TV + \bar{G}_0 = TV \oplus \text{span}\{\tilde{g}_1, \dots, \tilde{g}_{\rho_0}\}$  so that  $Q_0 \cap \text{span}\{\tilde{g}_1, \dots, \tilde{g}_{\rho_0}\} = 0$ . The distribution  $\bar{G}_0$  has dimension  $\nu_0 = n_0 + \rho_0$  throughout  $U$  so that

$$\begin{aligned} (\forall p \in U) \quad \bar{G}_0(p) &\supseteq \text{span}\{v_1^0, \dots, v_{\mu_0}^0\}(p) + \text{span}\{\tilde{g}_1, \dots, \tilde{g}_{\rho_0}\}(p), \\ (\forall p \in V) \quad \bar{G}_0(p) &= \text{span}\{v_1^0, \dots, v_{\mu_0}^0\}(p) \oplus \text{span}\{\tilde{g}_1, \dots, \tilde{g}_{\rho_0}\}(p), \\ \text{where} \quad \text{span}\{v_1^0, \dots, v_{\mu_0}^0\} &\subseteq TV. \end{aligned}$$

The vector fields  $v_1^0, \dots, v_{\mu_0}^0, \tilde{g}_1, \dots, \tilde{g}_{\rho_0}$  are linearly independent on  $V$ ; therefore they remain linearly independent in some open neighborhood of  $V$  in  $\mathbb{R}^n$ , without loss of generality,  $U$ . Therefore, they are local generators for  $\bar{G}_0$  on  $U$ . Next, we perform the



induction step. Assume that, for some positive integer  $I$ , and any  $i \in \{0, \dots, I\}$ , on  $U$ ,

$$(4.11) \quad \begin{aligned} \bar{G}_{i-1} &= G_{i-1}^{\parallel} \oplus \left( \bigoplus_{j=0}^{i-1} \text{span} \left\{ ad_{\tilde{f}}^j \tilde{g}_k : 1 \leq k \leq \rho_j \right\} \right) \\ \text{and } G_{i-1}^{\parallel} \Big|_V &= TV \cap \bar{G}_{i-1}. \end{aligned}$$

We want to show the existence of  $\mu_i$  vector fields  $v_1^i, \dots, v_{\mu_i}^i$  such that, for any  $i \in \{0, \dots, I\}$ , letting  $G_i^{\parallel} = G_{i-1}^{\parallel} \oplus \text{span}\{v_1^i, \dots, v_{\mu_i}^i\}$ , one has that, on  $U$ ,

$$(4.12) \quad \begin{aligned} \bar{G}_i &= G_i^{\parallel} \oplus \left( \bigoplus_{j=0}^i \text{span} \left\{ ad_{\tilde{f}}^j \tilde{g}_k : 1 \leq k \leq \rho_j \right\} \right) \\ \text{and } G_i^{\parallel} \Big|_V &= TV \cap \bar{G}_i. \end{aligned}$$

On  $V$ , define the distribution  $Q_i$  by  $Q_i = (TV \cap \bar{G}_i) \cap (TV \cap \bar{G}_{i-1})^{\perp}$ . The distribution  $Q_i$  is the intersection of two smooth, regular distributions. Furthermore, for all  $p \in V$ ,

$$\begin{aligned} \dim(Q_i(p)) &= \dim(T_p V \cap \bar{G}_i(p)) - \dim(T_p V \cap \bar{G}_{i-1}(p)) \\ &= \mu_i \end{aligned}$$

is constant by assumption, and thus, by Lemma 2.4,  $Q_i$  is a smooth regular distribution. Locally, by making  $U$  (and hence  $V$ ) smaller if necessary, there exist local generators  $\hat{v}_k$ ,  $k \in \{1, \dots, \mu_i\}$ , for  $Q_i$ . By construction,  $Q_i \subset \bar{G}_i$  so that each  $\hat{v}_k \in Q_i$  can be expressed as

$$\hat{v}_k = \sum_{j=1}^{\nu_i} \hat{c}_{ji}^k X_j^i, \quad k \in \{1, \dots, \mu_i\},$$

where each  $\hat{c}_{ji}^k : V \rightarrow \mathbb{R}$  is a  $C^\infty(V)$  function. Let  $c_{ji}^k = \hat{c}_{ji}^k \circ r$  so that

$$v_k^i := \sum_{j=1}^{\nu_i} c_{ji}^k X_j^i, \quad k \in \{1, \dots, \mu_i\}$$

are vector fields defined on  $U$ , and let  $G_{i/i-1}^{\parallel} := \text{span}\{v_1^i, \dots, v_{\mu_i}^i\}$ . It follows that  $G_{i/i-1}^{\parallel} \subset \bar{G}_i$  and  $G_{i/i-1}^{\parallel} \Big|_V = Q_i$ . By the definition of  $Q_i$  and by (4.11), it follows that  $G_{i-1}^{\parallel} \Big|_V \cap G_{i/i-1}^{\parallel} \Big|_V = (TV \cap \bar{G}_{i-1}) \cap Q_i = 0$ . Furthermore, since  $(TV \cap \bar{G}_{i-1}) \subset (TV \cap \bar{G}_i)$ , we have that  $TV \cap \bar{G}_i = G_{i-1}^{\parallel} \oplus G_{i/i-1}^{\parallel} =: G_i^{\parallel}$ . In addition, by Lemma 4.4, on  $V$ ,

$$G_i^{\parallel} \cap \left( \bigoplus_{j=0}^i \text{span} \left\{ ad_{\tilde{f}}^j \tilde{g}_k : 1 \leq k \leq \rho_j \right\} \right) = 0.$$

Finally, since  $\dim(\bar{G}_i) = n_i + \sum_{j=0}^i \rho_j$  we have that (4.12) holds on  $V$ . Thus, since  $G_i^{\parallel} \subset \bar{G}_i$  on  $U$ , (4.12) also holds in a neighborhood of  $V$ , without loss of generality,  $U$ .  $\square$



**5. Proof of the main result (Theorem 3.2).** Suppose that LTFLP is solvable at  $p_0 \in \Gamma^*$ . Let  $V = \Xi(\Gamma^* \cap U)$  and consider the expression (4.4) for  $TV + G_i$  in local coordinates. It is clear from (4.4) that the subspace  $T_p V + G_i(\text{col}(p, 0))$  has constant dimension  $n^* + \text{rank}([B \ \cdots \ A^i B])$ . Since the pair  $(A, B)$  is controllable, we have that  $\text{rank}([B \ \cdots \ A^{n-n^*-1} B]) = n - n^*$  and condition (a) holds. As far as condition (b) is concerned, we have already shown in the proof of Lemma 4.3 that  $TV + G_i = TV + \tilde{G}_i$ .

Conversely, suppose conditions (a) and (b) hold. These two conditions, along with the regularity of  $\tilde{G}_i$ ,  $i \in \mathbf{n} - \mathbf{n}^* - 1$ , allow one to invoke Lemmas 4.4 and 4.5. Specifically, there exist a neighborhood  $\tilde{U}$  of  $p_0$  in  $\mathbb{R}^n$ , a regular static feedback  $(\alpha, \beta)$  on  $\tilde{U}$ , and  $n_{k_1-1}$  vector fields defined on  $\tilde{U}$  such that, letting  $\tilde{V} := \tilde{U} \cap \Gamma^*$ , the distributions  $\tilde{G}_i$  have the representation given in (4.13), (4.14) on  $\tilde{U}$ , and at each  $p \in \tilde{V}$  the  $n$  vector fields of the array (4.16) are linearly independent.

Having applied the static feedback  $(\alpha, \beta)$ , we will denote  $\tilde{f}$  and  $\tilde{g}$  by, respectively,  $f$  and  $g$  to simplify notation. We construct  $\rho_0$  functions  $\alpha_i : U \rightarrow \mathbb{R}$  satisfying Theorem 3.1. Pick the point  $p_0$  as the origin for the  $s$ -coordinate system to be generated from the vector fields in (4.16) (see [25], [27]). Compose the flows generated by the vector fields in (4.16) starting from the bottom row. Consider the mapping

$$\begin{aligned} F_\emptyset : W_\emptyset \subset \mathbb{R}^{n^*-n_{k_1-1}} &\rightarrow V \\ : S_\emptyset = (s_1^\parallel, \dots, s_{n^*-n_{k_1-1}}^\parallel) &\mapsto \phi_{s_{n^*-n_{k_1-1}}^\parallel}^{v_{n^*-n_{k_1-1}}} \circ \cdots \circ \phi_{s_1^\parallel}^{v_1}(p_0). \end{aligned}$$

We continue by moving upwards in the array (4.16) to generate a sequence of mappings similar to  $F_\emptyset$ . To each pair  $(G_{i/i-1}^\nabla, G_{i/i-1}^\parallel)$ ,  $0 \leq i \leq k_1 - 1$ , we associate a set of times<sup>3</sup>  $S_{i/i-1} = (S_{i/i-1}^\nabla; S_{i/i-1}^\parallel) := (s_{(i/i-1,1)}^\nabla, \dots, s_{(i/i-1,\rho_i)}^\nabla; s_{(i/i-1,1)}^\parallel, \dots, s_{(i/i-1,\mu_i)}^\parallel)$  and a mapping

$$\begin{aligned} F_{i/i-1} : W_{i/i-1} \subset \mathbb{R}^{\mu_i + \rho_i} &\rightarrow \mathbb{R}^n \\ : (S_{i/i-1}^\parallel, S_{i/i-1}^\nabla) &\mapsto \Phi_{i/i-1}^\parallel \circ \Phi_{i/i-1}^\nabla(p). \end{aligned}$$

Here  $\Phi_{i/i-1}^\parallel$  is the composition of flows generated by vector fields spanning  $G_{i/i-1}^\parallel$ , and  $\Phi_{i/i-1}^\nabla$  is the composition of flows generated by the vector fields spanning  $G_{i/i-1}^\nabla$ . Specifically,

$$\begin{aligned} \Phi_{i/i-1}^\parallel &= \phi_{s_{(i/i-1,\mu_i)}^\parallel}^{v_{\mu_i}} \circ \cdots \circ \phi_{s_{(i/i-1,1)}^\parallel}^{v_1}, \\ \Phi_{i/i-1}^\nabla &= \phi_{s_{(i/i-1,\rho_i)}^\nabla}^{ad_f^i g_{\rho_i}} \circ \cdots \circ \phi_{s_{(i/i-1,1)}^\nabla}^{ad_f^i g_1}. \end{aligned}$$

Let

$$(5.1) \quad s = \text{col}(S_\emptyset; S_{k_1-1/k_1-2}; \dots; S_{1/0}; S_0)$$

and let  $W \subset \mathbb{R}^n$  be a neighborhood of  $s = 0$ , sufficiently small, to ensure that the map

$$(5.2) \quad \begin{aligned} F : W &\rightarrow F(W) \\ s &\mapsto F_0 \circ F_{1/0} \circ \cdots \circ F_{k_1-2/k_1-3} \circ F_{k_1-1/k_1-2} \circ F_\emptyset(p_0) \end{aligned}$$

<sup>3</sup>We define  $i/(i-1) := 0$  when  $i = 0$  to be consistent with the array (4.16).

is a diffeomorphism onto its image and that  $F(W) \subset \tilde{U}$ . The existence of  $W$  is guaranteed by the inverse function theorem and the fact that the differential of  $F$  at  $s = 0$ ,

$$(5.3) \quad dF_0 = \begin{bmatrix} v_1 & \cdots & v_{n^*-n_{k_1-1}} & ad_f^{k_1-1}g_1 & \cdots & g_1 & \cdots & g_{\rho_0} & v_1^0 & \cdots & v_{\mu_0}^0 \end{bmatrix} (p_0),$$

is an  $n \times n$  square matrix whose columns span the subspace  $T_{p_0}\Gamma^* + G_{k_1-1}(p_0)$  which, by condition (a), has dimension  $n$ . As candidate (virtual) output functions, let  $\alpha_i$ ,  $i \in \{1, \dots, \rho_0\}$  be the time spent flowing along  $ad_f^{k_i-1}g_i$ , i.e.,

$$(5.4) \quad \alpha_i(x) = s_{(k_i-1/k_i-2, i)}^\dagger(x), \quad i \in \{1, \dots, \rho_0\}.$$

The image of  $\tilde{V}$  under  $F^{-1}$  is the hyperplane

$$F^{-1}(\tilde{V}) = \{s \in W : S_0^\dagger = 0, S_{1/0}^\dagger = 0, \dots, S_{k_1-1/k_1-2}^\dagger = 0\}.$$

Since the chosen functions  $\alpha_1, \dots, \alpha_{\rho_0}$  are a subset of the functions whose zero level set defines  $F(\tilde{V})$ , the  $\alpha_i$  are identically zero on  $F(\tilde{V})$ , and hence condition (1) of Theorem 3.1 is satisfied. Next, we must show that  $\alpha = \text{col}(\alpha_1, \dots, \alpha_{\rho_0})$  yields a well-defined vector relative degree (VRD) of  $(k_1, \dots, k_{\rho_0})$  at  $p_0 = F^{-1}(0)$ . As per [9], this entails showing that

(VRD1)  $L_{ad_f^{k_j}g_j}\alpha_i(x) = 0$  for all  $1 \leq j \leq \rho_0$ , for all  $0 \leq k \leq k_i-2$ , for all  $1 \leq i \leq \rho_0$ , and for all  $x$  in a neighborhood of  $p_0$ .

(VRD2) the  $\rho_0 \times \rho_0$  matrix

$$(5.5) \quad \begin{pmatrix} L_{ad_f^{k_1-1}g_1}\alpha_1(p_0) & \cdots & L_{ad_f^{k_1-1}g_{\rho_0}}\alpha_1(p_0) \\ L_{ad_f^{k_2-1}g_1}\alpha_2(p_0) & \cdots & L_{ad_f^{k_2-1}g_{\rho_0}}\alpha_2(p_0) \\ \vdots & \vdots & \vdots \\ L_{ad_f^{k_{\rho_0}-1}g_1}\alpha_{\rho_0}(p_0) & \cdots & L_{ad_f^{k_{\rho_0}-1}g_{\rho_0}}\alpha_{\rho_0}(p_0) \end{pmatrix}$$

is nonsingular at  $p = p_0$  (if this matrix is nonsingular, then the decoupling matrix has full rank).

First we show that VRD1 holds. Fix a set of times  $S_\emptyset = c_\emptyset$ ,  $S_{k_1-1/k_1-2} = c_{k_1-1/k_1-2}$ ,  $\dots$ ,  $S_{k_i-1/k_i-2} = c_{k_i-1/k_i-2}$ , where each  $c_j$  is a constant vector, to uniquely determine the hyperplane

$$H_i = \{s \in W : S_\emptyset = c_\emptyset, S_{k_1-1/k_1-2} = c_{k_1-1/k_1-2}, \dots, S_{k_i-1/k_i-2} = c_{k_i-1/k_i-2}\}.$$

Consider the point  $s = \text{col}(c_\emptyset, \dots, c_{k_i-1/k_i-2}, 0, \dots, 0) \in H_i$  and let  $x = F(s) \in \tilde{U}$ . Through  $x$  there passes an integral submanifold of each  $\tilde{G}_i$ ,  $i \in \mathbf{k}_1 - \mathbf{1}$ , which we denote by  $L_i(x)$ . Consider the map  $F_0 \circ F_{1/0} \circ \cdots \circ F_{k_i-2/k_i-3}(x)$ . It is the composition of the flows defined by vector fields which are local generators for  $\tilde{G}_{k_i-2}$ . Therefore the image of this map is the  $\nu_{k_i-2}$ -dimensional manifold  $L_{k_i-2}(x) \cap \tilde{U}$ . On the other hand, the image of this map in  $s$ -coordinates is the hyperplane  $H_i$ , i.e.,  $H_i = F^{-1}(L_{k_i-2}(x) \cap \tilde{U})$ . Therefore for each  $s \in H_i$ ,  $T_s H_i = (F^{-1})_* \tilde{G}_{k_i-2}(s) = \text{Im}(\text{col}(0, I_{\nu_{k_i-2}}))$ . The function  $\alpha_i$  is among those fixed times which define the hyperplane  $H_i$ . Therefore,  $d\alpha_i \in \text{ann}(\tilde{G}_{k_i-2}) \subset \cdots \subset \text{ann}(G_0)$ , and hence VRD1 holds in a sufficiently small neighborhood of  $p_0$ .

Next we show that VRD2 holds. Treating  $T\mathbb{R}^n$  as an orthogonal bundle with the usual inner product, we see that the value of the  $(i, j)$ th entry of (5.5) is equal to

$$\left\langle d\alpha_i, ad_f^{k_i-1} g_j \right\rangle (p_0).$$

From the expression (5.3) for  $dF_0$  it follows that

$$ad_f^{k_i-1} g_j(p_0) = \left[ F_\star \left( \frac{\partial}{\partial s_{(k_i-1/k_i-2, j)}^\flat} \right) \right] \Big|_{s=0}, \quad 1 \leq i \leq j \leq \rho_0,$$

so that

$$\frac{\partial}{\partial s_{(k_i-1/k_i-2, j)}^\flat} = \left[ F_\star^{-1} \left( ad_f^{k_i-1} g_j(x) \right) \right] \Big|_{x=p_0}, \quad 1 \leq i \leq j \leq \rho_0.$$

In light of this and the definition of  $\alpha_i$ ,  $i \in \{1, \dots, \rho_0\}$ , given by (5.4), in  $s$ -coordinates the values of the entries of (5.5), along and below the diagonal, at  $p_0$  are

$$\left\langle ds_{k_i-1/k_i-2, i}^\flat, \frac{\partial}{\partial s_{(k_i-1/k_i-2, j)}^\flat} \right\rangle = \delta_{ij}, \quad 1 \leq i \leq j \leq \rho_0,$$

where  $\delta_{ij}$  is the Kronecker delta function. Thus, at  $0 = F^{-1}(p_0)$  the matrix (5.5), in  $s$ -coordinates, has ones along its diagonal and zeros below. Therefore it is nonsingular at  $s = 0$ , which is equivalent to being nonsingular at  $p_0 = F(0)$ .  $\square$

**6. Conclusions.** We have determined necessary and sufficient conditions under which a multi-input nonlinear control-affine system is locally transversally feedback linearizable with respect to a given invariant submanifold. Our main conditions are checkable, though we do not present a constructive procedure for finding the coordinate and feedback transformations.

One can similarly pose the global transverse feedback linearization problem (GTFLP) in which one, roughly speaking, seeks a single coordinate and feedback transformation such that (2.1) is feedback equivalent to the normal form (3.1) in a tubular neighborhood of  $\Gamma^*$ . Clearly the geometry of  $\Gamma^*$  will play an increased role in characterizing the solution. In [19], we provided sufficient conditions for the solvability of GTFLP in the single-input case. GTFLP for multi-input systems remains an open problem.

## REFERENCES

- [1] J. P. AUBIN, *Viability Theory*, Birkhäuser, Boston, 1991.
- [2] A. BANASZUK AND J. HAUSER, *Feedback linearization of transverse dynamics for periodic orbits*, Systems Control Lett., 26 (1995), pp. 95–105.
- [3] R. W. BROCKETT, *Feedback invariants for nonlinear systems*, in Proceedings of the IFAC World Congress, Helsinki, 1978, pp. 1115–1120.
- [4] P. BRUNOVSKÝ, *A classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–187.
- [5] V. GUILLEMIN AND A. POLLACK, *Differential Topology*, Prentice-Hall, Englewood Cliffs, NJ, 1974.
- [6] M. W. HIRSCH, *Differential Topology*, Graduate Texts in Math. 33, Springer-Verlag, New York, 1976.
- [7] R. M. HIRSCHORN, *(A,  $\mathcal{B}$ )-invariant distributions and disturbance decoupling of nonlinear systems*, SIAM J. Control Optim., 19 (1981), pp. 1–19.

- [8] L. R. HUNT, R. SU, AND G. MEYER, *Design for multi-input nonlinear systems*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 268–298.
- [9] A. ISIDORI, *Nonlinear Control Systems*, 3rd ed., Springer, New York, 1995.
- [10] A. ISIDORI AND A. J. KRENER, *On feedback equivalence of nonlinear systems*, Systems Control Lett., 2 (1982), pp. 118–121.
- [11] A. ISIDORI, A. J. KRENER, C. GORI-GIORGI, AND S. MONACO, *Nonlinear decoupling via feedback: A differential geometric approach*, IEEE Trans. Automat. Control, 26 (1981), pp. 331–345.
- [12] B. JAKUBCZYK AND W. RESPONDEK, *On linearization of control systems*, Bull. Acad. Polon. Sci. Sér. Sci. Math., 28 (1980), pp. 517–522.
- [13] A. J. KRENER, A. ISIDORI, AND W. RESPONDEK, *Partial and robust linearization by feedback*, in Proceedings of the 22nd IEEE Conference on Decision and Control, San Antonio, 1983, pp. 126–130.
- [14] A. J. KRENER, *On the equivalence of control systems and the linearization of nonlinear systems*, SIAM J. Control Optim., 11 (1973), pp. 670–676.
- [15] R. MARINO, *On the largest feedback linearizable subsystem*, Systems Control Lett., 6 (1986), pp. 345–351.
- [16] R. MARINO, W. M. BOOTHBY, AND D. L. ELLIOTT, *Geometric properties of linearizable control systems*, Math. Systems Theory, 18 (1985), pp. 97–123.
- [17] R. MARINO AND P. TOMEI, *Nonlinear Control Design*, Prentice-Hall, Toronto, 1995.
- [18] C. NIELSEN AND M. MAGGIORE, *Maneuver regulation via transverse feedback linearization: Theory and examples*, in Proceedings of the IFAC Symposium on Nonlinear Control Systems (NOLCOS), Stuttgart, Germany, 2004, pp. 59–66.
- [19] C. NIELSEN AND M. MAGGIORE, *Output stabilization and maneuver regulation: A geometric approach*, Systems Control Lett., 55 (2006), pp. 418–427.
- [20] H. NIJMEIJER, *Controlled invariance for affine control systems*, Internat. J. Control, 34 (1981), pp. 825–833.
- [21] H. NIJMEIJER AND A. VAN DER SCHAFT, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
- [22] H. POINCARÉ, *Sur les propriétés des fonctions définies par les équations aux différences partielles*, in Œuvres, Vol. 1, Gauthier-Villars, Paris, 1929, pp. xcix–cx.
- [23] W. RESPONDEK, *Partial linearization, decompositions and fibre linear systems*, in Theory and Applications of Nonlinear Control Systems, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 137–154.
- [24] A. SHIRIAEV, J. W. PERRAM, AND C. CANUDAS DE WIT, *Constructive tool for orbital stabilization of underactuated nonlinear systems: Virtual constraints approach*, IEEE Trans. Automat. Control, 50 (2005), pp. 1164–1176.
- [25] M. SPIVAK, *A Comprehensive Introduction to Differential Geometry*, Vol. 1, 3rd ed., Publish or Perish, Houston, 2005.
- [26] R. SU, *On the linear equivalents of nonlinear systems*, Systems Control Lett., 2 (1982), pp. 48–52.
- [27] R. SU AND L. R. HUNT, *A canonical expansion for nonlinear systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 670–673.
- [28] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer-Verlag, New York, 1985.
- [29] W. M. WONHAM AND A. S. MORSE, *Feedback invariants of linear multivariable systems*, Automatica, 8 (1972), pp. 93–100.
- [30] Z. XU AND L. R. HUNT, *On the largest input-output linearizable subsystem*, IEEE Trans. Automat. Control, 41 (1996), pp. 128–132.
- [31] Z. XU AND L. R. HUNT, *Finding maximal linear subsystems of nonlinear systems with outputs*, IEEE Trans. Automat. Control, 45 (2000), pp. 1701–1704.

## NECESSARY CONDITIONS FOR MULTIOBJECTIVE OPTIMAL CONTROL PROBLEMS WITH FREE END-TIME\*

B. T. KIEN<sup>†</sup>, N.-C. WONG<sup>‡</sup>, AND J.-C. YAO<sup>‡</sup>

**Abstract.** Necessary conditions of optimality are derived for multiobjective optimal control problems with free end-time, in which the dynamics constraint is modeled as a nonconvex differential inclusion. The obtained results cover some previous results on necessary conditions for multiobjective and single objective optimal control problems.

**Key words.** multiobjective optimal control problems, free end-time, preference, nonconvex differential inclusion, nonsmooth analysis

**AMS subject classifications.** 49K24, 90C29

**DOI.** 10.1137/080714683

**1. Introduction.** The derivation of necessary conditions for multiobjective optimal control (MOC for short) problems in which the dynamic constraint is modeled as a differential inclusion has been a recent area of research. Problems of MOC naturally arise, for example, in economics (see [6]), in chemical engineering (see [3]), and in multiobjective control design (see [26]). Let us assume that  $\prec$  is a preference in  $R^m$ . We are interested in deriving necessary conditions for the following problem with free end-times and state constraints:

$$\begin{aligned} \text{(P)} \quad & \text{Minimize } g(a, x(a), b, x(b)) \\ & \text{over intervals } [a, b] \text{ and arcs } x \in W^{1,1}([a, b], R^n) \text{ which satisfy} \\ & \dot{x}(t) \in F(t, x(t)), \text{ a.e., } t \in [a, b], \\ & (a, x(a), b, x(b)) \in C, \end{aligned}$$

where  $g : R \times R^n \times R \times R^n \rightarrow R^m$  is a given mapping,  $F : R \times R^n \rightrightarrows R^n$  is a given multifunction,  $C$  is a closed set in  $R \times R^n \times R \times R^n$ , and  $W^{1,1}([a, b], R^n)$  is the space of absolutely continuous functions  $x : [a, b] \rightarrow R^n$ .

Given  $x \in W^{1,1}([a, b], R^n)$  we define  $x^e$  to be an extension of  $x$  obtained by constants extension for the left endpoint on  $(-\infty, a)$  and for the right endpoint on  $(b, +\infty)$ . A feasible process  $([a, b], x)$  comprises a closed interval  $[a, b]$  and an arc  $x \in W^{1,1}([a, b], R^n)$  which satisfy the constraints of (P). A feasible process  $([a_*, b_*], x_*)$  is said to be a local solution of (P) if there does not exist any feasible process  $([a, b], x)$  with  $d([a, b], x), ([a_*, b_*], x_*) \leq \epsilon$  such that  $g(a, x(a), b, x(b)) \prec g(a_*, x_*(a_*), b_*, x_*(b_*))$  for some  $\epsilon > 0$ . Here

$$d([a, b], x), ([a', b'], y) := |a - a'| + |b - b'| + |x(a) - y(a')| + \int_{a \wedge a'}^{b \vee b'} |\dot{x}^e(s) - \dot{y}^e(s)| ds,$$

in which  $a \wedge a' := \min\{a, a'\}$  and  $b \vee b' := \max\{b, b'\}$ . We remark that the notion of  $W^{1,1}$  local optimizers to differential inclusions was first introduced and studied

\*Received by the editors July 27, 2007; accepted for publication (in revised form) March 18, 2008; published electronically August 13, 2008. This research was partially supported by a grant from the National Science Council of Taiwan, R.O.C.

<http://www.siam.org/journals/sicon/47-5/71468.html>

<sup>†</sup>Department of Information and Technology, Hanoi University of Civil Engineering, 55 Giai Phong, Hanoi, Vietnam (kienbt@uce.edu.vn).

<sup>‡</sup>Department of Applied Mathematics, National Sun Yat-Sen University, Kaohsiung 80424, Taiwan, R.O.C. (wong@math.nsysu.edu.tw, yaojc@math.nsysu.edu.tw).

in [15] under the name of “intermediate local minimizers,” which are different from the classical notions of weak and strong local minimizers in variational and optimal control problems.

In the scalar case ( $m = 1$ ), there are several papers dealing with necessary conditions of the Euler–Lagrange type for (P). The generalized Euler–Lagrange condition was first established by Mordukhovich [15] for problems governed by nonconvex, compact-valued, Lipschitzian differential inclusions on the fixed time interval and then was extended to free-time problems in [14]. Further extensions for unbounded differential inclusions were given by Ioffe [8] and Loewen and Rockafellar [10], by Vinter and Zheng [25] for problems with unbounded differential inclusions on the fixed time interval, and by Vinter and Zheng [23] and Vinter [22] for free-time problems.

Particularly, Vinter [22] provided an efficient scheme for deriving necessary conditions of local optimization solutions of (P) (see [22, Theorem 8.4.1]). A notable feature of the new free end-time necessary conditions is that they cover problems with measurable time-dependent data. For such problems, standard analytical techniques for deriving free-time necessary conditions, which depend on a transformation of the time variable, no longer work.

It is natural to ask whether the conclusions of theorems in [22] are still valid for the case of MOC problems. The goal of this paper is to obtain such results for (P).

Unfortunately, the scheme of the proof given by [22] fails to apply to our problem. The reason is that in this case we cannot use scalar estimations nor differentiable property of functions for the problem. However, that scheme helps us derive necessary conditions for the Bolza problem with finite Lagrangian which plays an important role in the establishment of necessary conditions for (P).

In a close connection, recently Zhu [28] established a result on the Hamiltonian necessary conditions for a nonsmooth MOC problem with endpoint constraints involving regular preferences. This result was extended later by Bellaassali and Jourani [2]. Based on an analysis of Ioffe’s scheme [8], as it was mentioned, Bellaassali and Jourani [2] obtained an interesting result on necessary conditions for MOC problems. However, [2] and [28] considered only optimal problems with the fixed time interval.

In order to derive necessary conditions of the Euler–Lagrange type for (P), we use a variant of Ioffe’s scheme [8] to reduce the problem to the scalar case, as has been done in [2] and [24]. We then use the Ekeland principle and necessary conditions for the Bolza problem. Together with the maximum theorem and some analytical techniques of nonsmooth analysis we finally obtain desired results.

The rest of the paper contains three sections. In section 2 we present some notions and auxiliary results involving generalized differentiation. In section 3 we derive necessary conditions for the Bolza problems. The final section is devoted to deriving necessary conditions for problem (P).

**2. Preliminaries and auxiliary results.** Throughout the paper  $B$  stands for the closed unit ball in  $R^n$ , and  $R_\infty$  stands for  $R \cup \{+\infty\}$ .

In what follows we often deal with set-valued mappings  $\Gamma : R^n \rightrightarrows R^n$ , for which the notation

$$\text{Limsup}_{x \rightarrow \bar{x}} \Gamma(x) := \{x^* \in R^n : \exists x_k \rightarrow \bar{x}, x_k^* \rightarrow x^* \text{ with } x_k^* \in \Gamma(x_k)\}$$

denotes the sequential Painlevé–Kuratowski upper limit of  $\Gamma$  at a point  $\bar{x} \in R^n$ . The



set

$$\text{Gph}\Gamma := \{(x, y) \in R^n \times R^n : y \in \Gamma(x)\}$$

is called the graph of  $\Gamma$ .

Take a closed set  $A \subset R^n$  and point  $x \in A$ . The set

$$\hat{N}_A(x) := \left\{ x^* \in R^n : \limsup_{u \xrightarrow{A} x} \frac{\langle x^*, u - x \rangle}{\|u - x\|} \leq 0 \right\}$$

is called the Fréchet normal cone to  $A$  at  $x$ . Let  $\bar{x} \in A$ ; the set

$$N_A(\bar{x}) := \text{Limsup}_{x \rightarrow \bar{x}} \hat{N}_A(x)$$

is the limiting normal cone to  $A$  at  $\bar{x}$ .

Given a lower semicontinuous (l.s.c.) function  $f : R^n \rightarrow R_\infty$  and a point  $x \in R^n$  such that  $f(x) < \infty$ , the limiting subdifferential of  $f$  at  $x$  is the set

$$\partial f(x) = \{x^* : (x^*, -1) \in N_{\text{epi}f}(x, f(x))\}.$$

It is well known that if  $f$  is Lipschitz continuous around  $x$  with rank  $K$ , then for any  $x^* \in \partial f(x)$ , one has  $\|x^*\| \leq K$ . The limiting normal cone and limiting subdifferential were introduced by Mordukhovich [18]. We refer the reader to Chapter 1 in [12] for comprehensive commentaries. Further properties of limiting normal cone and limiting subdifferential can be founded in [12] and [4].

Let  $\Gamma : X \subset R^n \rightarrow 2^{R^n}$  be a multifunction. We now assume that  $\Gamma$  has closed values and define the function  $\rho_\Gamma : X \times R^n \rightarrow R$  by

$$\rho_\Gamma(x, y) = d(y, \Gamma(x)) := \inf_{v \in \Gamma(x)} \|y - v\|.$$

The following property of the subdifferential of  $\rho_\Gamma$ , which was first established in [21], will be needed in section 4.

LEMMA 2.1. *Assume that  $\text{Gph}\Gamma$  is closed and  $(\bar{x}, \bar{y}) \in \text{Gph}\Gamma$ . Then one has*

$$N_{\text{Gph}\Gamma}(\bar{x}, \bar{y}) = \bigcup_{\lambda \geq 0} \lambda \partial \rho_\Gamma(\bar{x}, \bar{y}).$$

Moreover, if  $\rho_\Gamma(x, y) > 0$  and  $v \in \partial_y \rho_\Gamma(x, y)$ , then there exists a point  $z \in \Pi_{\Gamma(x)}(y)$  such that  $v = \frac{y-z}{\|y-z\|}$ . Here  $\Pi_{\Gamma(x)}(y)$  is the set of metric projections of  $y$  onto  $\Gamma(x)$ .

The proof of Lemma 2.1 can also be found in [8], [12], and [24].

Recall that the multifunction  $\Gamma : X \subset R^n \rightrightarrows R^n$  is said to be l.s.c. on  $X$  if for each  $x_0 \in X$  and an open set  $V$  satisfying  $F(x_0) \cap V \neq \emptyset$ , there exists a neighborhood  $U$  of  $x_0$  such that  $F(x) \cap V \neq \emptyset$  for all  $x \in U \cap X$ .  $F$  is said to be upper semicontinuous (u.s.c.) on  $X$  if for each  $x_0 \in X$  and an open set  $V$  in  $R^n$  satisfying  $F(x_0) \subset V$ , there exists a neighborhood  $U$  of  $x_0$  such that  $F(x) \subset V$  for all  $x \in U \cap X$ .  $F$  is said to be continuous on  $X$  if it is both l.s.c. and u.s.c. on  $X$ .

In what follows we shall need the next lemma.

LEMMA 2.2. *Let  $X \subset R^n$ ,  $Y \subset R^n$  be nonempty sets,  $\phi : Y \times R^n \rightarrow R$  be a continuous function, and  $\Gamma : X \subset R^n \rightrightarrows R^n$  be a multifunction with compact values. Assume that  $\Gamma$  is Lipschitz continuous on  $X$ ; that is, there exists a constant  $k > 0$  such that*

$$\Gamma(x') \subset \Gamma(x) + k|x' - x|B$$

for all  $x, x' \in X$ . Then the function  $M$  defined by

$$M(x, y) = \max\{\phi(y, u) : u \in \Gamma(x)\}$$

is continuous on  $X \times Y$ .

*Proof.* We first show that  $\Gamma$  is l.s.c. on  $X$ . Indeed, take any point  $x_0 \in X$  and a open set  $V$  such that  $\Gamma(x_0) \cap V \neq \emptyset$ . We want to prove that there exists a neighborhood  $U$  of  $x_0$  such that  $\Gamma(x) \cap V \neq \emptyset$  for all  $x \in U$ . Otherwise, there is a sequence  $x_n \rightarrow x_0$  satisfying  $\Gamma(x_n) \cap V = \emptyset$ . Take  $y_0 \in \Gamma(x_0) \cap V$ . By the property of  $\Gamma$ ,  $d(y_0, \Gamma(x_n)) \leq k|x_0 - x_n|$ . Hence for each  $n$ , there exists  $y_n \in \Gamma(x_n)$  such that  $|y_0 - y_n| \leq k|x_0 - x_n|$ . Consequently,  $y_n \rightarrow y_0$  and so  $y_n \in V$  for  $n$  sufficiently large. It follows that  $y_n \in \Gamma(x_n) \cap V$  for  $n$  sufficiently large, which is a contradiction. Thus  $\Gamma$  is l.s.c. on  $X$ . By the standard arguments, we can also show that  $\Gamma$  is u.s.c. on  $X$ .

For each  $(x, y) \in X \times Y$  we put  $z = (x, y)$ . Define mappings  $\hat{\phi} : R^n \times Y \times R^n \rightarrow R$  and  $\hat{\Gamma} : X \times Y \rightarrow R^n$  by  $\hat{\phi}(z, u) = \phi(y, u)$  and  $\hat{\Gamma}(z) = \Gamma(x)$ . Then we have

$$M(x, y) = M(z) = \max\{\hat{\phi}(z, u) : u \in \hat{\Gamma}(z)\}.$$

Since  $\hat{\Gamma}$  is continuous on  $X \times Y$  with compact values and  $\hat{\phi}$  is a continuous function, the maximum theorem (see [1, Maximum theorem, p. 116]) implies that  $M$  is continuous on  $X \times Y$ .  $\square$

We remark that in [13] Mordukhovich and Nam showed that under certain conditions,  $M$  is locally Lipschitz continuous (see [13, Theorem 5.2]). However, they required that the cost function  $\phi$  is locally Lipschitzian. As we need only the continuity of  $M$ , in Lemma 2.2, we did not require that  $\phi$  is locally Lipschitzian.

The rest of this section is destined for some notion of preferences in  $R^m$ . The concept of a preference first appeared in the value theory of economics. In the area of multiobjective optimization and optimal control much research has been devoted to the weak Pareto solution and its generalizations. The preference relation between vectors  $x, y \in R^m$  in the sense of weak Pareto is defined by  $x \prec y$  if and only if  $x_i \leq y_i$  for  $i = 1, \dots, m$  and at least one of the inequalities is strict. In other words,  $x \prec y$  if and only if  $x - y \in R_-^m$  and  $x \neq y$ , where  $R_-^m := \{z \in R^m : z_i \leq 0, i = 1, 2, \dots, m\}$ . In this paper we use more general preference relations for which necessary conditions of the weak Pareto solution and its generalization can be derived and refined from our necessary conditions.

Let  $\prec$  be a preference in  $R^m$  and  $r \in R^m$ . We will call the set  $\mathcal{L}[r] := \{s \in R^m : s \prec r\}$  a level set at  $r$ , and  $\bar{\mathcal{L}}[r]$  is the closure of  $\mathcal{L}[r]$ .

We shall use the following definition (see [12, Definition 5.55] and [28]).

**DEFINITION 2.3.** A preference  $\prec$  is closed provided that

- (a) for any  $r \in R^n$ ,  $r \in \bar{\mathcal{L}}[r]$ ; and
- (b) for any  $r \prec s$ ,  $t \in \bar{\mathcal{L}}[r]$  implies that  $t \prec s$ .

We say that  $\prec$  is regular at  $\bar{r}$  (in the sense of [28]) provided that

- (c)

$$\text{Limsup}_{r, \theta \rightarrow \bar{r}} N_{\bar{\mathcal{L}}[r]}(\theta) \subset N_{\bar{\mathcal{L}}[\bar{r}]}(\bar{r}).$$

It is noted that the regularity notion for preference was introduced in [17] as *normal semicontinuity*, the name under which it is studied in Chapter 5 of [12]. In the above definition, the regularity is somewhat different from that in Definition 5.69 of [12], where a preference  $\prec$  is regular at  $(\bar{\theta}, \bar{r}) \in \text{Gph} \mathcal{L}$  if

$$\text{Limsup}_{(r, \theta) \xrightarrow{\text{Gph} \mathcal{L}} (\bar{\theta}, \bar{r})} \hat{N}_{\mathcal{L}[r]}(\theta) = N_{\mathcal{L}[\bar{\theta}]}(\bar{r}).$$

Let us give some examples for Definition 2.3.

*Example 2.4* (single objective problem). When  $m = 1$  the relation  $r \prec s$  becomes  $r < s$ . It is obvious that this relation satisfies conditions (a)–(c). Therefore necessary conditions for (P) are true generalizations of necessary conditions for single objective optimal control (see Corollary 4.2).

*Example 2.5* (weak Pareto optimal control problem). In a weak Pareto optimal control problem we define the preference by  $r \prec s$  if and only if  $r_i \leq s_i$ ,  $i = 1, 2, \dots, m$ , and at least one of the inequalities is strict. It is easy to check that this  $\prec$  satisfies (a) and (b) at any  $r \in R^n$ . Moreover, for any  $r \in R^m$ ,  $\mathcal{L}[r] = r + R_-^m$ , where  $R_-^m := \{s \in R^m : s_i \leq 0, i = 1, 2, \dots, m\}$ . It follows that  $N_{\mathcal{L}[r]}(\theta) \subset R_+^m = N_{\mathcal{L}[r]}(r)$  for all  $r$  and  $\theta$ . Hence (c) is also satisfied. Thus the necessary conditions for (P) with respect to  $\prec$  are true for weak Pareto optimal control problems (see Corollary 4.3).

**3. The Bolza problem with finite Lagrangian.** In this section we derive necessary conditions of the Bolza problem

$$(BP) \quad \text{Minimize } J(a, b, x) := l(a, x(a), b, x(b)) + \int_a^b L(t, x(t), \dot{x}(t)) dt$$

over intervals  $[a, b]$  and arcs  $x \in W^{1,1}([a, b], R^n)$ ,

where  $l : R \times R^n \times R \times R^n \rightarrow R_\infty$  and  $L : R \times R^n \times R^n \rightarrow R$  are given functions.

A triple  $([a, b], x)$  which satisfies the constraint of (PB) is called a feasible process. A feasible process  $([a_*, b_*], x_*)$  is a local solution of (BP) if there exists  $\epsilon > 0$  such that  $J(a, b, x) \geq J(a_*, b_*, x_*)$  for all feasible process satisfying  $d(([a, b], x), ([a_*, b_*], x_*)) \leq \epsilon$ .

We now fix a feasible process  $([a_*, b_*], x_*)$  for the problem and assume the following assumptions which involve positive numbers  $\delta, \delta_0, \delta_1$ :

(BH1)  $l$  is Lipschitz continuous near  $(a_*, x_*(a_*), b_*, x_*(b_*))$  with rank  $k_l$ .

(BH2)  $L(\cdot, x, \cdot)$  is  $\mathcal{L} \times \mathcal{B}$  measurable for each  $x \in R^n$ , and  $L(t, \cdot, \cdot)$  is l.s.c. for a.e.  $t \in [a_*, b_*]$ .

(BH3) For all  $N$  there exists  $k_N \in L^1[a_*, b_*]$  such that

$$|L(t, x, v) - L(t, x', v)| \leq k_N(t)|x - x'|, \quad L(t, x_*(t), v) \geq -k_N(t)$$

for all  $x, x' \in x_*(t) + \delta B$  and  $v \in \dot{x}_*(t) + NB$  a.e.  $t \in [a_*, b_*]$ .

(BH4) There exist essentially bounded functions  $\bar{u} : [a_* - \delta_0, a_*] \rightarrow R^n$  and  $\tilde{u} : [b_*, b_* + \delta_1] \rightarrow R^n$  such that the functions  $t \mapsto L(t, x_*(a_*), \bar{u}(t))$  and  $t \mapsto L(t, x_*(b_*), \tilde{u}(t))$  are essentially bounded on  $[a_* - \delta_0, a_*]$  and  $[b_*, b_* + \delta_1]$ , respectively. Moreover, there exist positive constants  $k_0, k_1$  such that for all  $u \in R^n$ , one has

$$|L(t, x, u) - L(t, x', u)| \leq k_0|x - x'| \quad \forall x, x' \in x_*(a_*) + \delta B \text{ a.e. } t \in [a_* - \delta_0, a_*]$$

and

$$|L(t, x, u) - L(t, x', u)| \leq k_1|x - x'| \quad \forall x, x' \in x_*(b_*) + \delta B \text{ a.e. } t \in [b_*, b_* + \delta_1].$$

Define

$$\mathcal{H}_\lambda(t, x, v, p) = \langle p, v \rangle - \lambda L(t, x, v).$$

We have the following result on necessary conditions for (BP).

**THEOREM 3.1.** Assume that  $([a_*, b_*], x_*)$  is a local minimizer of (BP), for which  $J(a_*, x_*, b_*) < \infty$  and (BH1)–(BH3) are satisfied.

Then there exist an arc  $p \in W^{1,1}([a_*, b_*], R^n)$  and real numbers  $\xi$ ,  $\eta$ , and  $\lambda \geq 0$  such that

- (i)  $\lambda + \|p\|_\infty = 1$ ;
- (ii)  $\dot{p}(t) \in \text{co}\{(\alpha, p(t)) \in \lambda \partial L(t, x_*(t), \dot{x}_*(t))\}$  a.e.  $t \in [a_*, b_*]$ ;
- (iii)  $(-\xi, p(a_*), \eta, -p(b_*)) \in \lambda \partial l(a_*, x_*(a_*), b_*, x_*(b_*))$ ;
- (iv)  $\langle p(t), \dot{x}_*(t) \rangle - \lambda L(t, x_*(t), \dot{x}_*(t)) \geq \langle p(t), v \rangle - \lambda L(t, x_*(t), v)$  for all  $v \in R^n$  a.e.; and
- (v)

$$\xi \leq \lim_{\sigma \rightarrow 0} \text{ess sup}_{t \in [a_* - \sigma, a_* + \sigma]} \mathcal{H}_\lambda(t, x_*(t), \dot{x}_*(t), p(a_*))$$

and

$$\eta \leq \lim_{\sigma \rightarrow 0} \text{ess sup}_{t \in [b_* - \sigma, b_* + \sigma]} \mathcal{H}_\lambda(t, x_*(t), \dot{x}_*(t), p(b_*)).$$

Moreover, if (BH4) holds, then

$$\xi \geq \lim_{\sigma \rightarrow 0} \text{ess inf}_{t \in [a_* - \sigma, a_*]} \mathcal{H}_\lambda(t, x_*(a_*), \bar{u}(t), p(a_*))$$

and

$$\eta \geq \lim_{\sigma \rightarrow 0} \text{ess inf}_{t \in [b_*, b_* + \sigma]} \mathcal{H}_\lambda(t, x_*(b_*), \tilde{u}(t), p(b_*)).$$

*Proof.* To prove the theorem, we use a variant of the scheme in [22, Theorem 8.4.1].

*Step 1.* Take  $a_* \in R$ ,  $g_1 : R^n \rightarrow R_\infty$ ,  $g_2 : R \times R^n \rightarrow R_\infty$ , and  $g_3 : R \rightarrow R_\infty$ . Let  $([a_*, b_*], x_*)$  be a  $W^{1,1}$  local minimizer for the following problem:

$$\begin{aligned} & \text{Minimize } g_1(x(a_*)) + g_2(b, x(b)) + g_3(b) + \int_{a_*}^b L(t, x(t), \dot{x}(t)) dt \\ & \text{over processes } ([a_*, b], x) \text{ which satisfy } x \in W^{1,1}([a_*, b]). \end{aligned}$$

Assume that (BH2) and (BH3) are satisfied,  $g_1$  is Lipschitz continuous near  $x_*(a_*)$ ,  $g_2$  is twice continuously differentiable near  $(b_*, x_*(b_*))$ , and  $g_3$  is Lipschitz continuous near  $b_*$ .

We show that there exist  $p \in W^{1,1}$  and  $\lambda \geq 0$  such that

- (A1)  $\lambda + \|p\|_\infty = 1$ ;
- (B1)  $\dot{p}(t) \in \text{co}\{(\alpha, p(t)) \in \lambda \partial L(t, x_*(t), \dot{x}_*(t))\}$  a.e.;
- (C1)  $p(a_*) \in \lambda \partial g_1(x_*(a_*))$ ,  $-p(b_*) = \lambda \nabla_x g_2(b_*, x_*(b_*))$ ;
- (D1)  $\langle p(t), \dot{x}_*(t) \rangle - \lambda L(t, x_*(t), \dot{x}_*(t)) \geq \langle p(t), v \rangle - \lambda L(t, x_*(t), v)$  for all  $v \in R^n$  a.e.;
- and
- (E1)

$$\lambda \nabla_b g_2(b_*, x_*(b_*)) \leq \text{ess sup}_{[b_* - \sigma, b_* + \sigma]} \mathcal{H}_\lambda(t, x_*(t), \dot{x}_*(t), p(b_*)) + \lambda k_3,$$

in which  $k_3$  is a Lipschitz constant for  $g_3$ . Moreover, if (BH4) holds, then

$$-\lambda k_3 + \lim_{\sigma \rightarrow 0} \text{ess inf}_{[b_* - \sigma, b_* + \sigma]} \mathcal{H}_\lambda(t, x_*(b_*), \bar{u}(t), p(b_*)) \leq \lambda \nabla_b g_2(b_*, x_*(b_*)).$$

Conditions (A1)–(D1) follow directly from the fixed end-time conditions in [24, Theorem 3]. It remains to prove (E1). For  $\sigma > 0$  sufficiently small,  $([a_*, b_* - \sigma], x_*)$  must have cost not less than that of  $([a_*, b_*], x_*)$ . Hence we have

$$g_2(b_*, x_*(b_*)) + g_3(b_*) + \int_{a_*}^{b_*} L_*(t) dt \leq g_2(b_* - \sigma, x_*(b_* - \sigma)) + g_3(b_* - \sigma) + \int_{a_*}^{b_* - \sigma} L_*(t) dt,$$

where  $L_*(t) := L(t, x_*(t), \dot{x}_*(t))$ . Since  $g_2$  is  $C^2$ , we get

$$0 \leq -\nabla_b g_2(b_*, x_*(b_*))\sigma - \nabla_x g_2(b_*, x_*(b_*)) \int_{b_* - \sigma}^{b_*} \dot{x}_*(t) dt + o(\sigma) + k_3\sigma - \int_{b_* - \sigma}^{b_*} L_*(t) dt.$$

Consequently,

$$\begin{aligned} 0 &\leq -\lambda \nabla_b g_2(b_*, x_*(b_*))\sigma - \lambda \nabla_x g_2(b_*, x_*(b_*)) \\ &\quad \times \int_{b_* - \sigma}^{b_*} \dot{x}_*(t) dt + \lambda o(\sigma) + k_3\sigma - \lambda \int_{b_* - \sigma}^{b_*} L_*(t) dt \\ &= -\lambda \nabla_b g_2(b_*, x_*(b_*))\sigma + \int_{b_* - \sigma}^{b_*} [\langle p(b_*), \dot{x}_*(t) \rangle - \lambda L_*(t)] dt + \lambda o(\sigma) + \lambda k_3\sigma. \end{aligned}$$

Hence

$$\begin{aligned} \lambda \nabla_b g_2(b_*, x_*(b_*)) &\leq \lim_{\sigma \rightarrow 0} \frac{1}{\sigma} \int_{b_* - \sigma}^{b_*} [\langle p(b_*), x_*(t) \rangle - \lambda L_*(t)] dt + \lambda k_3 \\ &\leq \lim_{\sigma \rightarrow 0} \operatorname{ess\,sup}_{[b_* - \sigma, b_* + \sigma]} \mathcal{H}_\lambda(t, x_*(t), \dot{x}_*(t), p(b_*)) + \lambda k_3. \end{aligned}$$

We now assume that (BH4) is fulfilled. Define a multifunction

$$F : [b_*, b_* + \delta_1] \times R^n \times R \rightarrow R^n \times R$$

by setting  $F(t, x, y) = \{(u, v) \in R^n \times R : v = L(t, x, u)\}$ . It is clear that for a.e.  $t \in [b_*, b_* + \delta_1]$ , the multifunction  $(x, y) \mapsto F(t, x, y)$  is Lipschitz continuous with rank  $k_1$  in a neighborhood of  $(x_*(b_*), y(b_*))$ , where  $y$  is a given constant function. Define the function  $\hat{z} : [b_*, b_* + \delta_1] \rightarrow R^n \times R$  by  $\hat{z}(t) = (\hat{x}(t), \hat{y}(t))$ , where

$$\hat{x}(t) = x_*(b_*) + \int_{b_*}^t \tilde{u}(s) ds, \quad \hat{y}(t) = y(b_*) + \int_{b_*}^t L(s, x_*(b_*), \tilde{u}(s)) ds.$$

We see that  $\hat{z}(t) \in F(t, x_*(b_*), y(b_*))$ . For each  $\sigma < \delta_1$  we put  $K_\sigma = \exp(\int_{b_*}^{b_* + \sigma} k_1 dt)$ ,  $\rho_\sigma(\hat{z}) = \int_{b_*}^{b_* + \sigma} \rho_F(t, \hat{z}(t), \dot{\hat{z}}(t)) dt$ , where  $\rho_F(t, z(t), \dot{z}(t)) := d(\dot{z}(t), F(t, z(t)))$  and  $z(t) = (x(t), y(t))$ . Since  $t \mapsto \tilde{u}(t)$  and  $t \mapsto L(t, x_*(b_*), \tilde{u}(t))$  are essentially bounded, there exists a constant  $M > 0$  such that

$$|\hat{z}(t) - (x_*(b_*), y(b_*))| \leq M|t - b_*|.$$

Hence  $\hat{z}(t) \rightarrow (x_*(b_*), y(b_*))$  as  $t \rightarrow b_*$ . By the Lipschitz continuity of  $F$ , we have

$$F(t, x_*(b_*), y(b_*)) \subset F(t, \hat{z}(t)) + k_1 |\hat{z}(t) - (x_*(b_*), y(b_*))|$$

for a.e.  $t \in [b_*, b_* + \sigma_1]$  for some  $\sigma_1 < \delta_1$ . This implies that  $\rho(t, \hat{z}(t), \dot{\hat{z}}(t)) \leq k_1 M |t - b_*|$  for a.e.  $t \in [b_*, b_* + \sigma_1]$ . Hence for all  $\sigma \in (0, \sigma_1)$  we have

$$\rho_\sigma(\hat{z}) = \int_{b_*}^{b_* + \sigma} \rho_F(t, \hat{z}(t), \dot{\hat{z}}(t)) dt \leq M k_1 \sigma^2.$$

Consequently,  $K_\sigma \rho_\sigma(\hat{z}) \rightarrow 0$  as  $\sigma \rightarrow 0$ . By Theorem 3.16 in [5], for each  $\sigma \in (0, \sigma_1)$  there exists a solution  $z_\sigma(t) = (x_\sigma(t), y_\sigma(t))$ ,  $t \in [b_*, b_* + \sigma]$ ,  $\dot{z}_\sigma(t) \in F(t, z_\sigma(t))$  with  $z_\sigma(b_*) = \hat{z}(b_*)$  satisfying

$$\int_{b_*}^{b_* + \sigma} |\dot{z}_\sigma(t) - \dot{\hat{z}}(t)| dt \leq K_\sigma \int_{b_*}^{b_* + \sigma} \rho_F(t, \hat{z}(t), \dot{\hat{z}}(t)) dt \leq K_\sigma M k_1 \sigma^2.$$

This implies that

$$\int_{b_*}^{b_* + \sigma} |\dot{x}_\sigma(t) - \tilde{u}(t)| dt \leq K_\sigma M k_1 \sigma^2$$

and

$$\int_{b_*}^{b_* + \sigma} |L(t, x_\sigma(t), \dot{x}_\sigma(t)) - L(t, x_*(b_*), \tilde{u}(t))| dt \leq K_\sigma M k_1 \sigma^2.$$

Fixing any  $\sigma \in (0, \sigma_1)$ , we define a function  $x$  by concatenating  $x_*(t)$ ,  $a_* \leq t \leq b_*$ , and  $x_\sigma(t)$ ,  $b_* \leq t \leq b_* + \sigma$ . We therefore obtain a feasible process  $([a_*, b_* + \sigma], x)$ . Since  $([a_*, b_* + \sigma], x)$  must have cost not less than that of  $([a_*, b_*], x_*)$ , we conclude that

$$\begin{aligned} g_2(b_*, x_*(b_*)) + g_3(b_*) + \int_{a_*}^{b_*} L_*(t) dt &\leq g_2(b_* + \sigma, x(b_* + \sigma)) + g_3(b_* + \sigma) \\ &+ \int_{a_*}^{b_* + \sigma} L(t, x(t), \dot{x}(t)) dt. \end{aligned}$$

Hence

$$\begin{aligned} 0 &\leq \nabla_b g_2(b_*, x_*(b_*)) \sigma + o(\sigma) + k_3 \sigma + \int_{b_*}^{b_* + \sigma} \nabla_x g_2(b_*, x_*(b_*)) \dot{x}_\sigma(t) dt \\ &+ \int_{b_*}^{b_* + \sigma} L(t, x_\sigma(t), \dot{x}_\sigma(t)) dt \\ &\leq \nabla_b g_2(b_*, x_*(b_*)) \sigma + o(\sigma) + k_3 \sigma \\ &+ \int_{b_*}^{b_* + \sigma} \nabla_x \langle g_2(b_*, x_*(b_*)), \tilde{u}(t) \rangle dt + |\nabla_x g_2(b_*, x_*(b_*))| K_\sigma M k_1 \sigma^2 \\ &+ \int_{b_*}^{b_* + \sigma} L(t, x_*(b_*), \dot{\hat{z}}(t)) dt + K_\sigma M k_1 \sigma^2. \end{aligned}$$

Multiplying the latter inequality by  $\lambda \geq 0$  and dividing by  $\sigma > 0$  yields

$$\begin{aligned} & - \lambda k_3 + \frac{1}{\sigma} \int_{b_*}^{b_* + \sigma} [p(b_*) \tilde{u}(t) - \lambda L(t, x_*(b_*), \tilde{u}(t))] dt \\ & \leq K_\sigma M k_1 \sigma (\nabla_x g_2(b_*, x_*(b_*)) + 1) + \lambda \nabla_b g_2(b_*, x_*(b_*)). \end{aligned}$$

This implies that

$$-\lambda k_3 + \lim_{\sigma \rightarrow 0} \operatorname{ess\,inf}_{[b_* - \sigma, b_* + \sigma]} \mathcal{H}_\lambda(t, x_*(b_*), \tilde{u}(t), p(b_*)) \leq \lambda \nabla_b g_2(b_*, x_*(b_*)).$$

Thus assertions of Step 1 are obtained.

*Step 2.* Take  $a_* \in R$  and  $g : R \times R^n \times R^n \rightarrow R_\infty$ . Let  $([a_*, b_*], x_*)$  be a  $W^{1,1}$  local minimizer for the following problem:

$$\begin{aligned} & \text{Minimize } g(x(a_*), b, x(b)) + \int_{a_*}^b L(t, x(t), \dot{x}(t)) dt \\ & \text{over processes } ([a_*, b], x) \text{ which satisfy } x \in W^{1,1}([a_*, b]). \end{aligned}$$

Assume that (BH2) and (BH3) are satisfied and  $g$  is Lipschitz continuous near  $(x_*(a_*), b_*, x_*(b_*))$  with a rank  $k_g$ . We show that there exist  $p \in W^{1,1}$ , and real numbers  $\eta$  and  $\lambda \geq 0$  such that

$$\begin{aligned} & \text{(A2) } \lambda + \|p\|_\infty + |\eta| = 1, \\ & \text{(B2) } \dot{p}(t) \in \operatorname{co}\{\alpha : (\alpha, p(t)) \in \lambda \partial L(t, x_*(t), \dot{x}_*(t))\} \text{ a.e.,} \\ & \text{(C2) } (p(a_*), \eta, -p(b_*)) \in \lambda \partial g(x_*(a_*), b_*, x_*(b_*)), \\ & \text{(D2) } \langle p(t), \dot{x}_*(t) \rangle - \lambda L(t, x_*(t), \dot{x}_*(t)) \geq \langle p(t), v \rangle - \lambda L(t, x_*(t), v) \text{ for all } v \in R^n \text{ a.e.,} \\ & \text{and} \\ & \text{(E2)} \end{aligned}$$

$$\eta \leq \lim_{\sigma \rightarrow 0} \operatorname{ess\,sup}_{[b_* - \sigma, b_*]} \mathcal{H}_\lambda(t, x_*(t), \dot{x}_*(t), p(b_*)).$$

Moreover, if (BH4) holds, then

$$\eta \geq \lim_{\sigma \rightarrow 0} \operatorname{ess\,inf}_{[b_*, b_* + \sigma]} \mathcal{H}_\lambda(t, x_*(b_*), \tilde{u}(t), p(b_*)).$$

Take a sequence  $K_i \rightarrow \infty$  and define

$$\begin{aligned} J_i(b, x, \tau, y) &:= g(x(a_*), \tau(a_*), y(a_*)) \\ &+ \int_{a_*}^b L(t, x(t), \dot{x}(t)) dt + K_i(|\tau(b) - b|^2 + |y(b) - x(b)|^2), \end{aligned}$$

where  $\tau$  and  $y$  are constant functions. Denote by  $W$  the set of all  $([a_*, b], z = (x, \tau, y))$  such that  $x \in W^{1,1}([a_*, b], R^n)$ ,  $\tau \in R$ ,  $y \in R^n$ . With respect to the metric

$$\begin{aligned} & d([a_*, b], (x, \tau, y), [a_*, b'], (x', \tau', y')) \\ &= |b - b'| + |x(a_*) - x'(a_*)| + \|\dot{x}^e - \dot{x}'^e\|_{L^1} + |\tau - \tau'| + |y - y'|, \end{aligned}$$

$W$  is complete and  $J_i$  is continuous. Let us define a sequence  $\epsilon_i$  by

$$\epsilon_i^2 := J_i(b_*, x_*, b_*, x_*(b_*)) - \inf_W J_i(b, x, \tau, y).$$

By similar arguments as in Step 5 to follow, we can show that  $\epsilon_i \rightarrow 0$ . The Ekeland principle now gives us, for each  $i$ , a point  $(b_i, x_i, \tau_i, y_i)$  in  $W$  such that

$$(1) \quad d[(b_i, x_i, \tau_i, y_i), (b_*, x_*, b_*, x_*(b_*))] \leq \epsilon_i,$$

$$(2) \quad J_i(b_i, x_i, \tau_i, y_i) \leq J_i(b, x, \tau, y) + \epsilon_i d[(b_i, x_i, \tau_i, y_i), (b, x, \tau, y)] \quad \forall (b, x, \tau, y) \in W.$$

From (1), it follows that  $b_i \rightarrow b_*$ ,  $\tau_i \rightarrow b_*$ ,  $y_i \rightarrow x_*(b_*)$ ,  $x_i^e \rightarrow x_*^e$  uniformly,  $\dot{x}_i^e \rightarrow \dot{x}_*^e$  a.e. and in  $L^1$ . Also, (2) implies that  $(b_i, \tau_i, x_i, y_i)$  is a  $W$  minimizer of the functional

$$\begin{aligned} \tilde{J}_i(b, z) &:= g(x(a_*), \tau(a_*), y(a_*)) + \epsilon_i(|\tau(a_*) - \tau_i| + |x(a_*) - x_i(a_*)| + |y(a_*) - y_i|) \\ &+ K_i(|\tau(b) - b|^2 + |y(b) - x(b)|^2) + \epsilon_i(|b - b_i| + \int_b^{b \vee b_i} |\dot{x}_i^e(t)| dt) \\ &+ \int_{a_*}^b (L(t, x(t), \dot{x}(t)) + \epsilon_i|\dot{x} - \dot{x}_i^e|) dt. \end{aligned}$$

Put

$$g_1(z(a_*)) = g(x(a_*), \tau(a_*), y(a_*)) + \epsilon_i(|\tau(a_*) - \tau_i| + |x(a_*) - x_i(a_*)| + |y(a_*) - y_i|),$$

$$g_2(b, z(b)) = K_i(|\tau(b) - b|^2 + |y(b) - x(b)|^2),$$

and

$$g_3(b) = \epsilon_i \left( |b - b_i| + \int_b^{b \vee b_i} |\dot{x}_i^e(t)| dt \right).$$

According to Step 1, there exist  $p_i$  and real numbers  $\lambda_i \geq 0$ ,  $\eta_i$ , and  $r_i$  such that

- (i)  $\lambda_i + |\eta_i| + \|p_i\|_\infty = 1$ ,
- (ii)  $\dot{p}_i(t) \in \text{co}\{\alpha : (\alpha, p_i(t)) \in \lambda_i \partial L(t, x_i(t), \dot{x}_i(t)) + \epsilon_i \lambda_i \{0\} \times B\}$  a.e.  $t \in [a_*, b_i]$ ,
- (iii)  $(p_i(a_*), \eta_i, r_i) \in \lambda_i \partial g(x_i(a_*), \tau_i, y_i) + \lambda_i \epsilon_i B \times B \times B$  and  $-(p_i(b_i), \eta_i, r_i) = \lambda_i \nabla_z g_2(b_i, x_i(b_i), \tau_i(b_i), y_i(b_i))$ ,
- (iv)  $\langle p_i(t), \dot{x}_i(t) \rangle - \lambda_i L(t, x_i(t), \dot{x}_i(t)) \geq \langle p_i(t), v \rangle - \lambda L(t, x_i(t), v) - \lambda_i \epsilon_i |v - \dot{x}_i|$  for all  $v \in R^n$  and a.e.  $t \in [a_*, b_i]$ , and
- (v)

$$\begin{aligned} &\lambda_i \nabla_b g_2(b_i, x_i(b_i), \tau_i(b_i), y_i(b_i)) \\ &\leq \lim_{\sigma \rightarrow 0} \text{ess sup}_{[b_i - \sigma, b_i]} \mathcal{H}_{\lambda_i}(t, x_i(t), \dot{x}_i(t), p_i(b_i)) + \lambda_i \epsilon_i k_3. \end{aligned}$$

Assume that (BH4) is fulfilled. Putting  $\tilde{u}_i = \tilde{u}^e$ , we see that functions  $u_i$  and

$$t \mapsto L(t, x_i(b_i), u_i(t)) + \epsilon_i |\tilde{u}_i(t)|$$

are essentially bounded on  $[b_i, b_i + \delta_1]$ . Moreover, for  $i$  sufficiently large, the function

$$x \mapsto L(t, x, u) + \epsilon_i |u - x_i^e(t)|$$

is Lipschitz continuous with rank  $k_1$  for a.e.  $t \in [b_i, b_i + \delta_1]$ .

By the conclusion of Step 1, one has

$$\begin{aligned} &\lambda_i \nabla_b g_2(b_i, x_i(b_i), \tau_i(b_i), y_i(b_i)) \\ &\geq -\lambda_i \epsilon_i k_3 + \lim_{\sigma \rightarrow 0} \text{ess inf}_{[b_i, b_i + \sigma]} [\langle p_i(b_i), u_i(t) \rangle - \lambda_i L(t, x_i(b_i), \tilde{u}_i(t)) - \lambda_i \epsilon_i |\tilde{u}_i(t)|]. \end{aligned}$$

From (iii) we have  $-p_i(b_i) = -2\lambda_i K_i(y_i(b_i) - x_i(b_i))$ ,  $-\eta_i = 2\lambda_i K_i(\tau_i - b_i)$ ,  $-r_i = 2\lambda_i K_i(y_i(b_i) - x_i(b_i))$ . Hence  $-p_i(b_i) = r_i$  and  $\lambda_i \nabla_b g_2(b_i, x_i(b_i), \tau_i(b_i), y_i(b_i)) = \eta_i$ .

Since the  $p_i$ 's are bounded and their derivatives are bounded by an integrable function,  $p_i \rightarrow p$  uniformly and  $\dot{p}_i \rightarrow \dot{p}$  weakly in  $L^1$  for some  $p \in W^{1,1}$ . A further



subsequence extraction ensures that  $\lambda_i \rightarrow \lambda$ ,  $\eta_i \rightarrow \eta$  for some  $\lambda \geq 0$  and  $\eta$ . By passing to the limits as  $i \rightarrow \infty$  in (i)–(v), we obtain (B2)–(E2).

Since  $\lambda_i + \|p_i\|_\infty + |\eta_i| \neq 0$ , by scaling multipliers we can arrange so that  $\lambda_i + \|p_i\|_\infty + |\eta_i| = 1$ . Letting  $i \rightarrow \infty$  we obtain (A2). The proof of Step 2 is complete.

*Step 3* (necessary conditions for fixed right end-time problem). Take  $b_* \in R$  and  $g : R \times R^n \times R^n \rightarrow R_\infty$ . Let  $([a_*, b_*], x_*)$  be a  $W^{1,1}$  local solution of the following problem:

$$\begin{aligned} & \text{Minimize } g(a, x(a), x(b_*)) + \int_a^{b_*} L(t, x(t), \dot{x}(t)) dt \\ & \text{over processes } ([a, b_*], x) \text{ which satisfy } x \in W^{1,1}([a, b_*], R^n). \end{aligned}$$

Assume that (BH2) and (BH3) are satisfied and  $g$  is Lipschitz continuous in a neighborhood of  $(a_*, x_*(a_*), x_*(b_*))$ . We show that there exist  $p$  and real numbers  $\xi$  and  $\lambda \geq 0$  such that

$$\begin{aligned} \text{(A3)} \quad & \lambda + \|p\|_\infty + |\xi| = 1, \\ \text{(B3)} \quad & \dot{p}(t) \in \text{co}\{\alpha : (\alpha, p(t)) \in \lambda \partial L(t, x_*(t), \dot{x}_*(t))\} \text{ a.e. } t \in [a_*, b_*], \\ \text{(C3)} \quad & (-\xi, p(a_*), -p(b_*)) \in \lambda \partial g(a_*, x_*(a_*), x_*(b_*)), \\ \text{(D3)} \quad & \langle p(t), \dot{x}_*(t) \rangle - \lambda L(t, x_*(t), \dot{x}_*(t)) \geq \langle p(t), v \rangle - \lambda L(t, x_*(t), v) \text{ for all } v \in R^n \text{ a.e.} \\ & t \in [a_*, b_*], \text{ and} \\ \text{(E3)} \quad & \end{aligned}$$

$$\xi \leq \lim_{\sigma \rightarrow 0} \text{ess sup}_{[a_*, a_* + \sigma]} \mathcal{H}_\lambda(t, x_*(t), \dot{x}_*(t), p(a_*)).$$

Moreover,

$$\xi \geq \lim_{\sigma \rightarrow 0} \text{ess inf}_{[a_* - \sigma, a_*]} \mathcal{H}_\lambda(t, x_*(a_*), \bar{u}(t), p(a_*))$$

whenever (BH4) is fulfilled.

Put  $a'_* = -b_*$ ,  $b' = -a$ ,  $b'_* = -a_*$ ,  $x'(s) = x(-s)$ ,  $\bar{u}'(s) = -\bar{u}(-s)$ ,  $x'_*(s) = x_*(-s)$ ,  $g'(t, x, y) = g(-t, x, y)$ , and  $L'(s, x, y) = L(-s, x, -y)$ . By considering a change of independent variable  $s = -t$ , it follows that  $([a'_*, b'_*], x'_*)$  is a solution of the following problem:

$$\begin{aligned} & \text{Minimize } g'(b', x(b'), x(a'_*)) + \int_{a'_*}^{b'} L'(s, x'(s), \dot{x}'(s)) ds \\ & \text{over processes } ([a'_*, b'], x') \text{ which satisfy } x' \in W^{1,1}([a'_*, b'], R^n). \end{aligned}$$

According to Step 2, there exist  $p', \mu', \gamma', \lambda'$ , and  $\eta'$  such that

$$\begin{aligned} \text{(i)} \quad & \lambda' + \|p'\|_\infty + |\eta'| = 1, \\ \text{(ii)} \quad & \dot{p}'(s) \in \text{co}\{\alpha : (\alpha, p'(s)) \in \lambda' \partial L'(s, x'_*(s), \dot{x}'_*(s))\} \text{ a.e. } s \in [a', b'_*], \\ \text{(iii)} \quad & (\eta', -p'(b'_*), p'(a'_*)) \in \lambda' \partial g'(b'_*, x'_*(b'_*), x'_*(a'_*)), \\ \text{(iv)} \quad & \langle p'(s), \dot{x}'_*(s) \rangle - \lambda' L'(s, x'_*(s), \dot{x}'_*(s)) \geq \langle p'(s), v \rangle - \lambda' L'(s, x'_*(s), v) \text{ for all } v \in R^n \\ & \text{a.e., and} \\ \text{(v)} \quad & \end{aligned}$$

$$\eta' \leq \lim_{\sigma \rightarrow 0} \text{ess sup}_{[b'_* - \sigma, b'_*]} [\langle p'(b'_*), \dot{x}'_*(s) \rangle - \lambda' L'(s, x'_*(s), \dot{x}'_*(s))].$$

Moreover,

$$\eta' \geq \lim_{\sigma \rightarrow 0} \text{ess} \inf_{[b'_*, b'_* + \sigma]} [\langle p'(b'_*), \bar{u}'(s) \rangle - \lambda L'(s, x'_*(s), \bar{u}'(s))]$$

whenever (BH6) is fulfilled.

Put  $\xi = \eta'$ ,  $\lambda = \lambda'$ , and  $p(s) = -p'(-s)$ . By simple computation we obtain (A3)–(E3) from assertions (i)–(v).

*Step 4.* Take  $g_1, g_2 : R \times R^n \rightarrow R_\infty$ ,  $g_3 : R \rightarrow R_\infty$ . Let  $([a_*, b_*], x_*)$  be a solution of the following problem:

$$\begin{aligned} & \text{Minimize } g_1(a, x(a)) + g_2(b, x(b)) + g_3(b) + \int_a^b L(t, x(t), \dot{x}(t)) dt \\ & \text{over processes } ([a, b], x) \text{ which satisfy } x \in W^{1,1}([a, b], R^n). \end{aligned}$$

Assume that (BH2) and (BH3) are satisfied,  $g_1$  is Lipschitz continuous near  $(a_*, x_*(a_*))$ ,  $g_2$  is twice differentiable near  $(b_*, x_*(b_*))$ , and  $g_3$  is Lipschitz continuous near  $b_*$  with rank  $k_3$ .

Fixing  $b = b_*$ , we see that  $([a_*, b_*], x_*)$  is a solution of the following problem:

$$\begin{aligned} & \text{Minimize } g_1(a, x(a)) + g_2(b_*, x(b_*)) + g_3(b_*) + \int_a^{b_*} L(t, x(t), \dot{x}(t)) dt \\ & \text{over processes } ([a, b_*], x) \text{ which satisfy } x \in W^{1,1}([a, b_*], R^n). \end{aligned}$$

According to Step 3, there exist  $p$  and real numbers  $\lambda \geq 0$  and  $\xi$  such that

- (A4)  $\lambda + \|p\|_\infty + |\xi| = 1$ ;
- (B4)  $\dot{p}(t) \in \text{co}\{\alpha : (\alpha, p(t)) \in \lambda \partial L(t, x_*(t), \dot{x}_*(t))\}$  a.e.  $t \in [a_*, b_*]$ ;
- (C4)  $(-\xi, p(a_*)) \in \lambda \partial g_1(a_*, x_*(a_*))$ ,  $-p(b_*) = \lambda \nabla_x g_2(b_*, x_*(b_*))$ ;
- (D4)  $\langle p(t), \dot{x}_*(t) \rangle - \lambda L(t, x_*(t), \dot{x}_*(t)) \geq \langle p(t), v \rangle - \lambda L(t, x_*(t), v)$  for all  $v \in R^n$  a.e.  $t \in [a_*, b_*]$ ; and
- (E4)

$$\xi \leq \lim_{\sigma \rightarrow 0} \text{ess} \sup_{[a_*, a_* + \sigma]} \mathcal{H}_\lambda(t, x_*(t), \dot{x}_*(t), p(a_*)).$$

Moreover,

$$\xi \geq \lim_{\sigma \rightarrow 0} \text{ess} \inf_{[a_* - \sigma, a_*]} \mathcal{H}_\lambda(t, x_*(a_*), \bar{u}(t), p(a_*)).$$

Since  $([a_*, b_*], x_*)$  is also a solution of the following problem:

$$\begin{aligned} & \text{Minimize } g_1(a_*, x(a_*)) + g_2(b, x(b)) + g_3(b) + \int_{a_*}^b L(t, x(t), \dot{x}(t)) dt \\ & \text{over processes } ([a_*, b], x) \text{ which satisfy } x \in W^{1,1}([a_*, b], R^n), \end{aligned}$$

a similar argument as in Step 1 shows that

$$\begin{aligned} & -\lambda k_3 + \lim_{\sigma \rightarrow 0} \text{ess} \inf_{[b_*, b_* + \sigma]} \mathcal{H}_\lambda(t, x_*(b_*), \tilde{u}(t), p(b_*)) \leq \lambda \nabla_b g_2(b_*, x_*(b_*)) \\ & \leq \lim_{\sigma \rightarrow 0} \text{ess} \sup_{[b_* - \sigma, b_*]} \mathcal{H}_\lambda(t, x_*(t), \dot{x}_*(t), p(b_*)) + \lambda k_3. \end{aligned}$$

*Step 5.* We now return to the problem (BP). Let  $([a_*, b_*], x_*)$  be a solution of (BP), which we reiterate here:

$$\begin{aligned} &\text{Minimize } J(a, b, x) := l(a, x(a), b, x(b)) + \int_a^b L(t, x(t), \dot{x}(t)) dt \\ &\text{over intervals } [a, b] \text{ and arcs } x \in W^{1,1}([a, b], R^n). \end{aligned}$$

We want to show that there exist  $p$ , and real numbers  $\lambda \geq 0$ ,  $\xi$ , and  $\eta$  which satisfy the conclusion of Theorem 3.1.

Take a sequence  $K_i \rightarrow \infty$ . For each  $i$  we put

$$\begin{aligned} (3) \quad J_i(a, b, x, \tau, y) &= l(a, x(a), \tau(a), y(a)) \\ &\quad + \int_a^b L(t, x(t), \dot{x}(t)) dt + K_i(|\tau(b) - b|^2 + |y(b) - x(b)|^2), \end{aligned}$$

where  $\tau$  and  $y$  are constant functions. Denote by  $\tilde{W}$  the set of all  $(a, b, z = (x, \tau, y))$  such that  $x \in W^{1,1}([a, b], R^n)$ ,  $\tau \in R$ ,  $y \in R^n$ . It is clear that  $\tilde{W}$  is a metric space with respect to metric  $d$  induced by the norm

$$|(a, b, x, \tau, y)| = |a| + |b| + |x(a)| + \|\dot{x}^e\|_{L^1} + |\tau| + |y|.$$

Moreover,  $J_i$  is continuous on  $\tilde{W}$ . Define a sequence  $\epsilon_i$  by

$$\epsilon_i^2 := J_i(a_*, b_*, x_*, b_*, x_*(b_*)) - \inf_{\tilde{W}} J_i(a, b, x, \tau, y).$$

We claim that  $\epsilon_i \rightarrow 0$ . In fact, from (BH1) we get

$$l(a, x(a), \tau, y) \geq l(a, x(a), b, x(b)) - k_l(|\tau - b| + |y - x(b)|).$$

Hence

$$\begin{aligned} J_i(a, b, x, \tau, y) &\geq l(a, x(a), b, x(b)) + \int_a^b L(t, x(t), \dot{x}(t)) dt \\ &\quad - k_l(|\tau - b| + |y - x(b)|) + K_i(|\tau(b) - b|^2 + |y(b) - x(b)|^2) \\ &\geq J_i(a_*, b_*, x_*, b_*, x_*(b_*)) - k_l^2/2K_i. \end{aligned}$$

This implies that  $\epsilon_i \leq \frac{k_l}{\sqrt{2K_i}} \rightarrow 0$ . Since  $(a_*, b_*, x_*, b_*, x_*(b_*))$  is an  $\epsilon_i$  minimizer, Ekeland's principle gives us, for each  $i$ , a point  $(a_i, b_i, x_i, \tau_i, y_i)$  such that

$$(4) \quad d[(a_i, b_i, x_i, \tau_i, y_i), (a_*, b_*, x_*, b_*, x_*(b_*))] \leq \epsilon_i,$$

$$\begin{aligned} (5) \quad J_i(a_i, b_i, x_i, \tau_i, y_i) &\leq J_i(a, b, x, \tau, y) \\ &\quad + \epsilon_i d[(a, b, x, \tau, y), (a_i, b_i, x_i, \tau_i, y_i)] \quad \forall (a, b, x, \tau, y) \in \tilde{W}. \end{aligned}$$

From (4) we get  $a_i \rightarrow a_*$ ,  $b_i \rightarrow b_*$ ,  $\tau_i \rightarrow b_*$ ,  $y_i \rightarrow x_*(b_*)$ ,  $x_i \rightarrow x_*$  uniformly,  $\dot{x}_i^e \rightarrow \dot{x}_*^e$  a.e. and in  $L^1$ . It follows from (5) that  $(a_i, b_i, x_i, \tau_i, y_i)$  is a  $\tilde{W}$  minimizer of the functional

$$\begin{aligned} \tilde{J}_i(a, z) &:= l(a, x(a), \tau(a), y(a)) + \epsilon_i(|a - a_i| + |x(a) - x_i(a_i)| + |\tau - \tau_i| + |y - y_i|) \\ &\quad + \epsilon_i \int_{a \wedge a_i}^a |\dot{x}_i^e(t)| dt + \int_a^b (L(t, x(t), \dot{x}(t)) + \epsilon_i |\dot{x}(t) - \dot{x}_i^e(t)|) dt \\ &\quad + \epsilon_i K_i(|\tau(b) - b|^2 + |y(b) - x(b)|^2) + \epsilon_i \left( |b - b_i| + \int_b^{b \vee b_i} |\dot{x}_i^e(t)| dt \right), \end{aligned}$$

where  $z := (x, \tau, y)$ .

According to Step 4, there exist  $p_i$  and real numbers  $\lambda_i \geq 0$ ,  $\xi_i$ ,  $\eta_i$ , and  $r_i$  such that

- (A5)  $\lambda_i + \|p_i\|_\infty + |\xi_i| + |\eta_i| + |r_i| = 1$ ,
- (B5)  $\dot{p}_i(t) \in \text{co}\{\alpha : (\alpha, p_i(t)) \in \lambda \partial L(t, x_i(t), \dot{x}_i(t)) + \epsilon_i \lambda_i \{0\} \times B\}$  a.e.  $t \in [a_i, b_i]$ ,
- (C5)  $(-\xi_i, p_i(a_i), \eta_i, r_i) \in \lambda_i \partial l(a_i, x_i(a_i), \tau_i, y_i) + \lambda_i \epsilon_i k_3(B \times \{0\} \times \{0\} \times \{0\}) + \lambda_i \epsilon_i B^4$   
and  $-(p_i(b_i), \eta_i, r_i) = 2\epsilon_i K_i(-y_i + x_i(b_i), \tau_i - b_i, y_i - x_i(b_i))$ ,
- (D5)  $\langle p_i(t), \dot{x}_i(t) \rangle - \lambda_i L(t, x_i(t), \dot{x}_i(t)) \geq \langle p_i(t), v \rangle - \lambda L(t, x_i(t), v) - \lambda_i \epsilon_i |v - \dot{x}_i|$  for all  $v \in R^n$  and a.e.  $t \in [a_i, b_i]$ , and
- (E5)

$$\begin{aligned} & \lim_{\sigma \rightarrow 0} \text{ess} \inf_{[a_i - \sigma, a_i]} [\langle p_i(a_i), \bar{u}_i(t) \rangle - \lambda_i L(t, x_i(t), \dot{x}_i(t)) - \lambda_i \epsilon_i |\bar{u}_i(t)|] \leq \xi_i \\ & \leq \lim_{\sigma \rightarrow 0} \text{ess} \sup_{[a_i, a_i + \sigma]} [\langle p_i(a_i), \dot{x}_i(t) \rangle - \lambda_i L(t, x_i(t), \dot{x}_i(t))] \end{aligned}$$

and

$$\begin{aligned} & -\lambda_i \epsilon_i (1 + M) + \lim_{\sigma \rightarrow 0} \text{ess} \inf_{[b_i, b_i + \sigma]} [\langle p_i(b_i), \tilde{u}_i(t) \rangle - \lambda_i L(t, x_i(t), \tilde{u}_i(t)) - \lambda_i \epsilon_i |\tilde{u}_i(t)|] \\ & \leq 2K_i \epsilon_i (b_i - \tau_i) = \eta_i \leq \lim_{\sigma \rightarrow 0} \text{ess} \sup_{[b_i - \sigma, b_i]} [\langle p_i(b_i), \dot{x}_i(t) \rangle - \lambda L(t, x_i(t), \dot{x}_i(t))] \\ & \quad + \lambda_i \epsilon_i (1 + M), \end{aligned}$$

where  $\bar{u}_i = \bar{u}^c$  and  $\tilde{u}_i = \tilde{u}^c$ .

Since  $p_i$ 's are bounded and their derivatives are bounded by an integrable function,  $p_i \rightarrow p$  uniformly and  $\dot{p}_i \rightarrow \dot{p}$  weakly in  $L^1$  for some  $p \in W^{1,1}$ . A further subsequence extraction ensures that  $\lambda_i \rightarrow \lambda$ ,  $\eta_i \rightarrow \eta$ ,  $\xi_i \rightarrow \xi$ , and  $r_i \rightarrow -p(b_*)$ . Note that since  $-p_i(b_i) = r_i$ , it follows that  $\lambda + \|p\| \neq 0$ . By passing to the limit and standard arguments we can show that  $\lambda$  and  $p$  satisfy the conclusion of the theorem. The proof of the theorem is complete.  $\square$

*Remark 3.2.* Theorem 8.4.1 in [22] gave necessary conditions for problem (P) in the scalar case. It is possible to reduce (BP) to (P) (in the case  $m = 1$ ) and process it as in [23]. However, this transformation causes the structure of the problem to deteriorate and so it is difficult to obtain the desired conclusions. In the above argument, we exploited the structure of (BP) and gave a direct proof to establish our result.

**4. Necessary conditions for MOC.** In this section we derive necessary conditions for (P). Fix a feasible triple  $([a_*, b_*], x_*)$  and assume the following hypotheses which involve positive number  $\delta$ , a nonnegative function  $k_F \in L^1[a_*, b_*]$ , and a number  $\beta \geq 0$ :

- (H1)  $g$  is Lipschitz continuous on a neighborhood of  $(a_*, x_*(a_*), b_*, x_*(b_*))$  with rank  $k_g$  and  $C$  is a closed set.
- (H2)  $F$  is  $\mathcal{L} \times \mathcal{B}$  measurable with nonempty values and  $\text{Gph} F(t, \cdot)$  is closed.
- (H3)  $F$  has the integrable sub-Lipschitzian property (see [10]), that is,

$$F(t, x') \cap (x_*(t) + NB) \subset F(t, x) + (k_F(t) + \beta N)|x' - x|B$$

for all  $N \geq 0$ ,  $x', x \in x_*(t) + \delta B$ , a.e.  $t \in [a_*, b_*]$ .

- (H4) There exist positive constants  $c_0, c_1$ ,  $k_0$ , and  $k_1$  such that

$$\begin{cases} F(t, x) \subset c_0 B, \\ F(t, x') \subset F(t, x) + k_0 |x' - x| B \end{cases}$$

for a.e.  $t \in [a_* - \delta, a_*]$  and for all  $x, x' \in x_*(a_*) + \delta B$ ; and

$$\begin{cases} F(t, x) \subset c_1 B, \\ F(t, x') \subset F(t, x) + k_1 |x' - x| B \end{cases}$$

for a.e.  $t \in [b_*, b_* + \delta]$  and for all  $x, x' \in x_*(b_*) + \delta B$ .

In what follows,  $H(t, x, p) := \sup\{\langle p, v \rangle : v \in F(t, x(t))\}$  and  $\text{ess}_{\tau \rightarrow t} f(\tau)$  is the essential value of a real value function  $f$  at  $t \in I \subset \mathbb{R}$ , that is,  $\text{ess}_{\tau \rightarrow t} f(\tau) := [a_-, a_+]$ , where

$$a_- := \lim_{\delta \rightarrow 0} \text{ess inf}_{\tau \in [t-\delta, t+\delta]} f(\tau) \text{ and } a_+ := \lim_{\delta \rightarrow 0} \text{ess sup}_{\tau \in [t-\delta, t+\delta]} f(\tau).$$

We refer the reader to [22, Proposition 8.3.2] for properties of essential values.

We are ready to state our main result.

**THEOREM 4.1.** *Suppose  $([a_*, b_*], x_*)$  is a local minimizer of (P), preference  $\prec$  is regular at  $g(a_*, x_*(a_*), b_*, x_*(b_*))$ , and assumptions (H1)–(H4) are satisfied. Then there exist an arc  $p \in W^{1,1}([a_*, b_*], \mathbb{R}^n)$ , a vector  $w \in N_{\bar{L}[g(a_*, x_*(a_*), b_*, x_*(b_*))]}(g(a_*, x_*(a_*), b_*, x_*(b_*)))$  with  $|w| = 1$ , and real numbers  $\lambda \geq 0$ ,  $\xi$ , and  $\eta$  such that*

- (i)  $\lambda + \|p\|_\infty = 1$ ,
- (ii)  $\dot{p}(t) \in \text{co}\{\alpha : (\alpha, p(t)) \in N_{\text{Gph}F(t, \cdot)}(x_*(t), \dot{x}_*(t))\}$  a.e.  $t \in [a_*, b_*]$ ,
- (iii)  $(-\xi, p(a_*), \eta, -p(b_*)) \in \lambda \partial \langle w, g(a_*, x_*(a_*), b_*, x_*(b_*)) \rangle + N_C(a_*, x_*(a_*), b_*, x_*(b_*))$ ,
- (iv)  $\langle p(t), \dot{x}_*(t) \rangle = H(t, x_*(t), p(t))$  a.e.  $t \in [a_*, b_*]$ ,
- (v)  $\xi \in \text{ess}_{t \rightarrow a_*} H(t, x_*(a_*), p(a_*))$  and  $\eta \in \text{ess}_{t \rightarrow b_*} H(t, x_*(b_*), p(b_*))$ .

*Proof.* Define a mapping  $\rho_F : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  by

$$\rho_F(t, x, \dot{x}) = \inf\{|\dot{x} - v| : v \in F(t, x)\}.$$

According to Lemma 7 in [24] it follows from (H3) that  $\rho_F(t, \cdot, \cdot)$  satisfies condition (BH3) for a.e.  $t \in [a_*, b_*]$ . Put

$$W_\epsilon = \{([a, b], x) : x \in W^{1,1}([a, b]), d((([a, b], x), [a_*, b_*], x_*)) \leq \epsilon\}$$

and

$$S_\epsilon = \{([a, b], x) \in W_\epsilon : (a, x(a), b, x(b)) \in C, \dot{x}(t) \in F(t, x(t)) \text{ a.e.}\}.$$

It is clear that  $W_\epsilon$  is a complete metric space and  $S_\epsilon$  is a closed set in  $W_\epsilon$ .

Fix  $N$  and reduce the size of  $\epsilon$  such that  $([a_*, b_*], x_*)$  is the solution of (P) in  $S_\epsilon$ .

As in [24] and [2], we use a variant of Ioffe's scheme [8] by considering the following two possible situations:

- (a) There exist  $\epsilon' \in (0, \epsilon)$  and  $K > 0$  such that for any  $([a, b], x) \in W_{\epsilon'}$ , one has

$$(6) \quad d((([a, b], x), S_\epsilon) \leq K \left[ \int_a^b \rho_F(t, x(t), \dot{x}(t)) dt + K d_C(a, x(a), b, x(b)) \right].$$

- (b) There exist a sequence  $\epsilon'_k \rightarrow 0$  and a sequence  $([a_k, b_k], x_k) \in W_{\epsilon'_k}$  such that

$$(7) \quad d((([a_k, b_k], x_k), S_\epsilon) > 2k \left[ \int_{a_k}^{b_k} \rho_F(t, x_k(t), \dot{x}_k(t)) dt + 2k d_C(a_k, x_k(a_k), b_k, x_k(b_k)) \right].$$

Case (a). Since  $g(a_*, x_*(a), b_*, x_*(b_*)) \in \overline{\mathcal{L}}[g(a_*, x_*(a_*), b_*, x_*(b_*))]$ , there exists a sequence  $\theta_k \in \mathcal{L}[g(a_*, x_*(a_*), b_*, x_*(b_*))]$  such that  $|\theta_k - g(a_*, x_*(a_*), b_*, x_*(b_*))| \leq 1/k^2$ . Put  $\Omega_k = \overline{\mathcal{L}}[\theta_k]$  and define the function

$$\varphi(a, b, x, \theta) = \begin{cases} |g(a, x(a), b, x(b)) - \theta| & \text{if } (a, b, x, \theta) \in S_{\epsilon'} \times \Omega_k, \\ +\infty & \text{otherwise.} \end{cases}$$

We claim that  $\varphi$  is l.s.c. on  $W_{\epsilon'} \times \Omega_k$ . Indeed, assume that  $((a, b, x), \theta) \in W_{\epsilon'} \times \Omega_k$  and  $((a_n, b_n, x_n), \theta_n) \xrightarrow{W_{\epsilon'}} ((a, b, x), \theta)$ . If  $(a, b, x, \theta) \in S_{\epsilon'} \times \Omega_k$ , then it follows from Lipschitzian continuity of  $g$  that

$$\begin{aligned} |\varphi(a_n, b_n, x_n, \theta_n) - \varphi(a, b, x, \theta)| &\leq k_g(|a_n - a| + |b_n - b| + |x_n(a_n) - x(a)| \\ &\quad + |x_n(b_n) - x(b)|) + |\theta_n - \theta| \\ &\leq k_g(|a_n - a| + |b_n - b| + 2\|x_n - x\|_{\infty}) + |\theta_n - \theta| \rightarrow 0. \end{aligned}$$

If  $(a, b, x, \theta) \notin S_{\epsilon'} \times \Omega_k$ , then  $(a_n, b_n, x_n, \theta_n) \notin S_{\epsilon'} \times \Omega_k$  for  $n$  sufficiently large because  $S_{\epsilon'} \times \Omega_k$  is closed in  $W_{\epsilon'} \times R^m$ . Hence  $\lim_{n \rightarrow \infty} \varphi(a_n, b_n, x_n, \theta_n) = +\infty \geq \varphi(a, b, x, \theta)$ . Thus  $\varphi$  is l.s.c. Since  $\varphi(a, b, x, \theta) \geq 0$ , one has

$$\varphi(a_*, b_*, x_*, \theta_k) \leq \inf_{(x, \theta) \in W_{\epsilon'} \times \Omega_k} \varphi(a, b, x, \theta) + 1/k^2.$$

The Ekeland principle gives us, for each  $k$ , a point  $(a_k, b_k, x_k, \chi_k) \in W_{\epsilon'} \times \Omega_k$  such that

$$(8) \quad \varphi(a_k, b_k, x_k, \chi_k) \leq \varphi(a_*, b_*, x_*, \theta_k) < \frac{1}{k^2},$$

$$(9) \quad |a_k - a_*| + |b_k - b_*| + |x_k(a_k) - x_*(a_*)| + \|\dot{x}_k^e - \dot{x}_*^e\|_{L^1} + |\chi_k - \theta_k| \leq 1/k,$$

$$(10) \quad \varphi(a_k, b_k, x_k, \chi_k) \leq \varphi(a, b, x, \theta) + \frac{1}{k} [d([a, b], x), ([a_k, b_k], x_k) + |\theta - \chi_k|]$$

for all  $((a, b, x), \theta) \in W_{\epsilon'} \times \Omega_k$ . From (8), we get  $(a_k, b_k, x_k) \in S_{\epsilon'}$ . Inequality (9) implies that  $a_k \rightarrow a_*$ ,  $b_k \rightarrow b_*$ ,  $x_k(a_k) \rightarrow x_*(a_*)$ ,  $x_k^e \rightarrow x_*^e$  uniformly,  $\dot{x}_k^e \rightarrow \dot{x}_*^e$  a.e., and  $\chi_k \rightarrow g(a_*, x_*(a_*), b_*, x_*(b_*))$ . We claim that  $\chi_k \neq g(a_k, x_k(a_k), b_k, x_k(b_k))$ . Indeed, suppose that  $\chi_k = g(a_k, x_k(a_k), b_k, x_k(b_k))$ . Since  $\prec$  is closed, the relations  $\chi_k \in \overline{\mathcal{L}}[\chi_k]$  and  $\chi_k \prec g(a_*, x_*(a_*), b_*, x_*(b_*))$  imply that

$$g(a_k, x_k(a_k), b_k, x_k(b_k)) = \chi_k \prec g(a_*, x_*(a_*), b_*, x_*(b_*)).$$

This contradicts the fact that  $([a_*, b_*], x_*)$  is a minimizer.

Put

$$w_k = \frac{\chi_k - g(a_k, x_k(a_k), b_k, x_k(b_k))}{|\chi_k - g(a_k, x_k(a_k), b_k, x_k(b_k))|}.$$

We can assume that  $w_k \rightarrow w$  with  $|w| = 1$ . Substituting  $(a, b, x) = (a_k, b_k, x_k)$  into (10), it follows that

$$0 \in \partial \left( |g(a_k, x_k(a_k), b_k, x_k(b_k)) - \cdot| + \frac{1}{k} |\cdot - \chi_k| \right) (\chi_k) + N_{\Omega_k}(\chi_k).$$

This implies that  $w_k \in \frac{1}{k}B + N_{\Omega_k}(\chi_k)$ . Hence

$$w \in \lim_{k \rightarrow \infty} N_{\Omega_k}(\chi_k) \subset N_{\bar{\mathcal{E}}[g(a_*, x_*(a_*), b_*, x_*(b_*))]}(g(a_*, x_*(a_*), b_*, x_*(b_*))).$$

Also, substituting  $\theta = \chi_k$  into (10), it follows that

$$\varphi(a_k, b_k, x_k, \chi_k) \leq \varphi(a, b, x, \chi_k) + \frac{1}{k}d([a, b], x), ([a_k, b_k], x_k).$$

Combining this with (6) yields

$$\begin{aligned} \varphi(a_k, b_k, x_k, \chi_k) &\leq \varphi(a, b, x, \chi_k) + \frac{1}{k}d([a, b], x), ([a_k, b_k], x_k) \\ &\quad + \left[ \int_a^b \rho_F(t, x(t), \dot{x}(t)) dt + Kd_C(a, x(a), b, x(b)) \right] \end{aligned}$$

for all  $([a, b], x) \in W_{\epsilon'}$ . This implies that  $([a_k, b_k], x_k)$  is a  $W_{\epsilon'}$  minimizer of the following Bolza problem:

$$\begin{aligned} J(a, b, x) &:= \int_a^b \left( \rho_F(t, x(t), \dot{x}(t)) + \frac{1}{k}|\dot{x} - \dot{x}_k^e| \right) dt + |g(a, x(a), b, x(b)) - \chi_k| \\ &\quad + Kd_C(a, x(a), b, x(b)) + \frac{1}{k}(|a - a_k| + |b - b_k| + |x(a) - x_k(a_k)| \\ &\quad + \int_{a_k \wedge a}^a |\dot{x}_k^e| dt + \int_b^{b \vee b_k} |\dot{x}_k^e| dt. \end{aligned}$$

Put

$$\begin{aligned} L(t, x(t), \dot{x}(t)) &= \rho_F(t, x(t), \dot{x}(t)) + \frac{1}{k}|\dot{x} - \dot{x}_k^e|, \\ l(a, x(a), b, x(b)) &= |g(a, x(a), b, x(b)) - \chi_k| + Kd_C(a, x(a), b, x(b)) \\ &\quad + \frac{1}{k} \left( |a - a_k| + |b - b_k| + |x(a) - x_k(a_k)| + \int_{a_k \wedge a}^a |\dot{x}_k^e| dt + \int_b^{b \vee b_k} |\dot{x}_k^e| dt \right). \end{aligned}$$

It is easy to check that hypotheses (BH1)–(BH3) hold for  $l$  and  $L$ . By Theorem 3.1, there exist  $p_k \in W^{1,1}$ , and real numbers  $\lambda_k \geq 0$ ,  $\xi_k$ ,  $\eta_k$  such that

- (A)  $\lambda_k + |p_k|_\infty = 1$ ,
- (B)  $\dot{p}_k(t) \in \text{co}\{\alpha : (\alpha, p_k(t)) \in \lambda_k \partial \rho_F(t, x_k(t), \dot{x}_k(t)) + \frac{\lambda_k}{k} \{0\} \times B\}$  a.e.  $t \in [a_k, b_k]$ ,
- (C)  $(-\xi_k, p_k(a_k), \eta_k, -p_k(b_k)) \in \lambda_k \partial \langle w_k, g(a_k, x_k(a_k), b_k, x_k(b_k)) \rangle + \frac{\lambda_k}{k} B^3 \times \{0\} + \frac{\lambda_k}{k} M(B \times \{0\} \times B \times \{0\}) + \lambda_k K \partial d_C(a_k, x_k(a_k), b_k, x_k(b_k))$ ,
- (D)  $\langle p_k(t), \dot{x}_k(t) \rangle - \lambda_k \rho_F(t, x_k(t), \dot{x}_k(t)) \geq \langle p_k(t), \dot{v} \rangle - \lambda_k \rho_F(t, x_k(t), v) - \frac{\lambda_k}{k} |v - \dot{x}_k^e|$  for all  $v \in R^n$  a.e., and
- (E)

$$(11) \quad \xi_k \leq \lim_{\sigma \rightarrow 0} \text{ess sup}_{t \in [a_k - \sigma, a_k]} (\langle p_k(a_k), \dot{x}_k(t) \rangle - \lambda_k \rho_F(t, x_k(t), \dot{x}_k(t)))$$

and

$$(12) \quad \eta_k \leq \lim_{\sigma \rightarrow 0} \text{ess sup}_{t \in [b_k, b_k + \sigma]} (\langle p_k(b_k), \dot{x}_k(t) \rangle - \lambda_k \rho_F(t, x_k(t), \dot{x}_k(t))).$$

Fix any  $\sigma < \delta$ . By (H4) we can find essentially bounded selections  $\bar{u}_k$  and  $\tilde{u}_k$  of  $F(\cdot, x_k(a_k))$  and  $F(\cdot, x_k(b_k))$ , respectively, such that

$$(13) \quad \langle p_k(a_k), \bar{u}_k(t) \rangle = \max_{u \in F(t, x_k(a_k))} \langle p_k(a_k), u \rangle = H(t, x_k(a_k), p_k(a_k)) \text{ a.e. } t \in [a_k - \sigma, a_k]$$

and

$$(14) \quad \langle p_k(b_k), \tilde{u}_k(t) \rangle = \max_{u \in F(t, x_k(b_k))} \langle p_k(b_k), u \rangle = H(t, x_k(b_k), p_k(b_k)) \text{ a.e. } t \in [b_k, b_k + \sigma].$$

Since  $L(t, x_k(a_k), \bar{u}_k(t)) = \frac{1}{k} |\bar{u}_k(t)|$ , the function  $t \mapsto L(t, x_k(a_k), \bar{u}_k(t))$  is essentially bounded on  $[a_k - \sigma, a_k]$ . Moreover, the function  $x \mapsto L(t, x, u)$  is Lipschitz continuous with rank  $k_0$  in a neighborhood  $x_k(a_k)$  for  $k$  sufficiently large and for a.e.  $t \in [a_k - \sigma, a_k]$ . Also, the function  $t \mapsto L(t, x_k(b_k), \tilde{u}_k(t))$  is essentially bounded on  $[b_k, b_k + \sigma]$ , and  $x \mapsto L(t, x, u)$  is Lipschitz continuous with rank  $k_1$  for a.e.  $t \in [b_k, b_k + \sigma]$ . Hence (BH4) is fulfilled. By the conclusion of Theorem 3.1 we have

$$(15) \quad \xi_k \geq \lim_{\sigma \rightarrow 0} \text{ess} \inf_{t \in [a_k - \sigma, a_k]} \left( H(t, x_k(a_k), p_k(a_k)) - \frac{\lambda_k}{k} |\bar{u}_k(t)| \right)$$

and

$$(16) \quad \eta_k \geq \lim_{\sigma \rightarrow 0} \text{ess} \inf_{t \in [b_k, b_k + \sigma]} \left( H(t, x_k(b_k), p_k(b_k)) - \frac{\lambda_k}{k} |\tilde{u}_k(t)| \right).$$

Since  $p_k$ 's are bounded and their derivatives are bounded by an integrable function,  $p_k \rightarrow p$  uniformly and  $\dot{p}_k \rightarrow \dot{p}$  weakly in  $L^1$  for some  $p \in W^{1,1}$ . A further subsequence extraction ensures that  $\lambda_k \rightarrow \lambda$ ,  $\eta_k \rightarrow \eta$ ,  $\xi_k \rightarrow \xi$ .

By passing to the limit as  $k \rightarrow \infty$  in (A) we obtain (i). From (B) and Lemma 2.1, we have

$$\dot{p}_k(t) \in \text{co} \left\{ \alpha : (\alpha, p_k(t)) \in N_{\text{Grph}F(t, \cdot)}(x_k(t), \dot{x}_k(t)) + \frac{\lambda_k}{k} \{0\} \times B \right\}.$$

Passing to the limit as  $k \rightarrow \infty$  yields

$$\dot{p}(t) \in \text{co} \{ \alpha : (\alpha, p(t)) \in N_{\text{Grph}F(t, \cdot)}(x_*(t), \dot{x}_*(t)) \}.$$

Hence (ii) follows. As

$$\lambda_k K \partial d_C(a_k, x_k(a_k), b_k, x_k(b_k)) \subset N_C(a_k, x_k(a_k), b_k, x_k(b_k)),$$

passing to the limit in (C) and (D), we obtain (iii) and (iv), respectively.

By Lemma 2.2,  $H(t, \cdot, \cdot)$  is continuous for a.e.  $t$ . Passing to the limit in (11) and (15), and using properties of essential values (see [22, Proposition 8.3.2]), we get

$$(17) \quad \lim_{\sigma \rightarrow 0} \text{ess} \inf_{t \in [a_* - \sigma, a_*]} H(t, x_*(a_*), p(a_*)) \leq \xi \leq \lim_{\sigma \rightarrow 0} \text{ess} \sup_{t \in [a_* - \sigma, a_*]} \langle p(a_*), \dot{x}_*(t) \rangle.$$

By (H4), we have

$$F(t, x_*(t)) \subset F(t, x_*(a_*)) + k_0 |x_*(t) - x_*(a_*)| B \text{ for a.e. } t \in [a_* - \sigma, a_*].$$



Hence

$$\sup_{u \in F(t, x_*(t))} \langle p(a_*), u \rangle \leq \sup_{u \in F(t, x_*(a_*))} (\langle p(a_*), u \rangle + k_0 |x_*(t) - x_*(a_*)|).$$

This implies that

$$(18) \quad \lim_{\sigma \rightarrow 0} \operatorname{ess\,sup}_{t \in [a_* - \sigma, a_*]} H(t, x_*(t), p(a_*)) \leq \lim_{\sigma \rightarrow 0} \operatorname{ess\,sup}_{t \in [a_* - \sigma, a_*]} H(t, x_*(a_*), p(a_*)).$$

Combining (17) with (18) yields

$$\lim_{\sigma \rightarrow 0} \operatorname{ess\,inf}_{t \in [a_* - \sigma, a_* + \sigma]} H(t, x_*(a_*), p(a_*)) \leq \xi \leq \lim_{\sigma \rightarrow 0} \operatorname{ess\,sup}_{t \in [a_* - \sigma, a_* + \sigma]} H(t, x_*(a_*), p(a_*)),$$

which means that  $\xi \in \operatorname{ess}_{t \rightarrow a_*} H(t, x_*(a_*), p(a_*))$ . By similar arguments, we can show that  $\eta \in \operatorname{ess}_{t \rightarrow b_*} H(t, x_*(b_*), p(b_*))$ . Thus (v) follows.

Case (b). Putting  $\epsilon_k = d([a_k, b_k, x_k], S_\epsilon)$ , we have

$$0 < \epsilon_k \leq d([a_k, b_k], x_k), ([a_*, b_*], x_*) \leq \epsilon'_k \rightarrow 0.$$

From (7) it follows that

$$\inf_{(a, b, x) \in W_\epsilon} \tilde{J}(a, b, x) + \frac{\epsilon_k}{2k} > \tilde{J}(a_k, b_k, x_k),$$

where  $\tilde{J}(a, b, x) := \int_a^b \rho_F(t, x(t), \dot{x}(t)) dt + 2kd_C(a, x(a), b, x(b))$ . By the Ekeland principle, for each  $k$  there exists a triple  $([\bar{a}_k, \bar{b}_k], \bar{x}_k) \in W_\epsilon$  such that

$$(19) \quad d([\bar{a}_k, \bar{b}_k], \bar{x}_k), ([a_k, b_k], x_k) \leq \epsilon_k/2$$

and  $([\bar{a}_k, \bar{b}_k], \bar{x}_k)$  is a  $W_\epsilon$  minimizer of the functional

$$(20) \quad J_*(a, b, x) := \tilde{J}(a, b, x) + \frac{1}{k} d([a, b], x), ([\bar{a}_k, \bar{b}_k], \bar{x}_k).$$

It is clear that (19) implies  $([\bar{a}_k, \bar{b}_k], \bar{x}_k) \xrightarrow{W_\epsilon} ([a_*, b_*], x_*)$  and  $([\bar{a}_k, \bar{b}_k], \bar{x}_k) \notin S_\epsilon$ . Rewrite (20) in the form

$$\begin{aligned} J_*(a, b, x) = & \int_a^b \rho_F(t, x(t), \dot{x}(t)) dt + |\dot{x}(t) - \dot{\bar{x}}^e(t)| dt + 2kd_C(a, x(a), b, x(b)) \\ & + \frac{1}{k} \left( |a - \bar{a}_k| + |b - \bar{b}_k| + \int_{a \wedge \bar{a}_k}^a |\dot{\bar{x}}_k^e| dt + \int_b^{b \vee \bar{b}_k} |\dot{\bar{x}}_k^e| dt \right). \end{aligned}$$

According to Theorem 3.1, there exist  $p_k$  and real numbers  $\lambda_k \geq 0$ ,  $\xi_k$ , and  $\eta_k$  such that

- (A)'  $\lambda_k + |p_k|_\infty = 1$ ,
- (B)'  $\dot{p}_k(t) \in \operatorname{co}\{\alpha : (\alpha, p_k(t)) \in \lambda_k \partial \rho_F(t, \bar{x}_k(t), \dot{\bar{x}}_k(t)) + \frac{\lambda_k}{k} \{0\} \times B\}$  a.e.  $t \in [\bar{a}_k, \bar{b}_k]$ .
- (C)'  $(-\xi_k, p_k(\bar{a}_k), \eta_k, -p_k(\bar{b}_k)) \in \lambda_k 2k \partial d_C(\bar{a}_k, \bar{x}_k(a_k), \bar{b}_k, \bar{x}_k(b_k)) + \frac{\lambda_k}{k} B^2 \times \{0\} \times \{0\} + \frac{\lambda_k}{k} M(B \times B \times \{0\} \times \{0\})$ ,
- (D)'  $\langle p_k(t), \dot{\bar{x}}_k(t) \rangle - \lambda_k \rho_F(t, \bar{x}_k(t), \dot{\bar{x}}_k(t)) \geq \langle p_k(t), v \rangle - \lambda_k \rho_F(t, \bar{x}_k(t), v) - \frac{\lambda_k}{k} |v - \dot{\bar{x}}_k^e|$  for all  $v \in R^n$  a.e., and

(E)'

$$\begin{aligned} & \lim_{\sigma \rightarrow 0} \operatorname{ess} \inf_{t \in [\bar{a}_k - \sigma, \bar{a}_k]} \left( H(t, \bar{x}_k(\bar{a}_k), p_k(\bar{a}_k), \cdot) - \frac{\lambda_k}{k} |\bar{u}_k(t)| \right) \leq \xi_k \\ & \leq \lim_{\sigma \rightarrow 0} \operatorname{ess} \sup_{t \in [\bar{a}_k - \sigma, \bar{a}_k]} (\langle p_k(\bar{a}_k), \dot{\bar{x}}_k(t) \rangle - \lambda_k \rho_F(t, \bar{x}_k(t), \dot{\bar{x}}_k(t))) \end{aligned}$$

and

$$\begin{aligned} & \lim_{\sigma \rightarrow 0} \operatorname{ess} \inf_{t \in [\bar{b}_k - \sigma, \bar{b}_k]} \left( H(t, \bar{x}_k(\bar{b}_k), p_k(\bar{b}_k), \cdot) - \frac{\lambda_k}{k} |\tilde{u}_k(t)| \right) \leq \eta_k \\ & \leq \lim_{\sigma \rightarrow 0} \operatorname{ess} \sup_{t \in [\bar{b}_k - \sigma, \bar{b}_k]} (\langle p_k(\bar{b}_k), \dot{\bar{x}}_k(t) \rangle - \lambda_k \rho_F(t, \bar{x}_k(t), \dot{\bar{x}}_k(t))), \end{aligned}$$

where  $\bar{u}_k$  and  $\tilde{u}_k$  are essentially bounded selections of  $F(\cdot, \bar{x}_k(\bar{a}_k))$  and  $F(\cdot, \bar{x}_k(\bar{b}_k))$ , respectively, which satisfy

$$\langle p_k(\bar{a}_k), \bar{u}_k(t) \rangle = H(t, \bar{x}_k(\bar{a}_k), p_k(\bar{a}_k)) \text{ a.e. } t \in [\bar{a}_k - \sigma, \bar{a}_k]$$

and

$$\langle p_k(\bar{b}_k), \tilde{u}_k(t) \rangle = H(t, \bar{x}_k(\bar{b}_k), p_k(\bar{b}_k)) \text{ a.e. } t \in [\bar{b}_k, \bar{b}_k + \sigma].$$

Note that  $\operatorname{esssup} \bar{u}_k \leq c_0$  and  $\operatorname{esssup} \tilde{u}_k \leq c_1$  for  $k$  sufficiently large. By using similar arguments as in part (a), we can assume that  $p_k \rightarrow p$  uniformly and  $\dot{p}_k \rightarrow \dot{p}$  weakly in  $L^1$ ,  $\lambda_k \rightarrow \lambda_0$ ,  $\eta_k \rightarrow \eta$ ,  $\xi_k \rightarrow \xi$ . By passing to the limits from (A)'–(E)' we get

- (i)  $\lambda_0 + \|p\|_\infty = 1$ .
- (ii)  $\dot{p}(t) \in \operatorname{co}\{\alpha : (\alpha, p(t)) \in N_{\operatorname{Gph} F(t, \cdot)}(x_*(t), \dot{x}_*(t))\}$  a.e.
- (iii)  $(-\xi, p(a_*), \eta, -p(b_*)) \in N_C(a_*, x_*(a_*), b_*, x_*(b_*))$ .
- (iv)  $\langle q(t), \dot{x}(t) \rangle \geq \langle q(t), v \rangle$  for all  $v \in F(t, x_*(t))$  a.e.
- (v)  $\xi \in \operatorname{ess}_{t \rightarrow a_*} H(t, x_*(a_*), p(a_*))$  and  $\eta \in \operatorname{ess}_{t \rightarrow a_*} H(t, x_*(b_*), p(b_*))$ .

We now claim that  $\|p\| \neq 0$ . Indeed, suppose that  $p = 0$ . Then from the fact  $([\bar{a}_k, \bar{b}_k], \bar{x}_k) \notin S_\epsilon$  we have either  $(\bar{a}_k, \bar{x}_k(\bar{a}_k), \bar{b}_k, \bar{x}_k(\bar{b}_k)) \notin C$  or  $(\bar{x}_k(t), \dot{\bar{x}}_k(t)) \notin \operatorname{Gph} F(t, \cdot)$ . If  $(\bar{a}_k, \bar{x}_k(\bar{a}_k), \bar{b}_k, \bar{x}_k(\bar{b}_k)) \notin C$ , then (C)' implies

$$|\xi_k| + |p_k(\bar{a}_k)| + |\eta_k| + |p_k(\bar{b}_k)| \geq 2k\lambda_k - \frac{\lambda_k}{k}(1 + M).$$

Hence

$$\frac{|\xi_k| + |p_k(\bar{a}_k)| + |\eta_k| + |p_k(\bar{b}_k)|}{2k} \geq \lambda_k - \frac{\lambda_k}{2k^2}(1 + M).$$

By letting  $k \rightarrow \infty$  we get  $\lambda_0 = 0$ . This contradicts  $\lambda_0 = 1$ .

If  $(\bar{x}_k(t), \dot{\bar{x}}_k(t)) \notin \operatorname{Grap} F(t, \cdot)$  then (D)' implies that

$$p_k(t) \in \lambda_k \partial_{\dot{x}} \rho_F(t, \bar{x}_k(t), \dot{\bar{x}}_k(t)) + \frac{\lambda_k}{k} B.$$

By Lemma 2.1,  $|p_k(t)| \geq \lambda_k - \frac{\lambda_k}{k}$ . This implies that

$$\lambda_k - \frac{\lambda_k}{k} \leq \|p_k\|.$$

By letting  $k \rightarrow \infty$  we obtain  $\lambda_0 = 0$ , which is absurd. Thus it must have  $\|p\| = 1 - \lambda_0 \neq 0$ . By scaling multipliers, we can assume that  $\|p\| = 1$ . Hence we obtain the conclusion of the theorem by putting  $\lambda = 0$ . The proof is complete.  $\square$

We remark, as pointed out by a referee, that actually the “regularity” (normal semicontinuity) assumption on the preference is not needed in the main theorem, Theorem 4.1, if we use the extended limiting normal cone mentioned above for the level set instead of the basic/limiting one. Let us give some corollaries of Theorem 4.1.

When  $m = 1$ , (P) becomes a single objective problem. In this case, we have the following.

**COROLLARY 4.2** (see [22, Theorem 8.4.1]). *Suppose  $([a_*, b_*], x_*)$  is a local minimizer of (P) and assumptions (H1)–(H4) are satisfied. Then there exist an arc  $p \in W^{1,1}$ , and real numbers  $\lambda \geq 0$ ,  $\xi$ , and  $\eta$  such that*

- (i)  $\lambda + \|p\|_\infty = 1$ ,
- (ii)  $\dot{p}(t) \in \text{co}\{\alpha : (\alpha, p(t)) \in N_{\text{Grph}F(t, \cdot)}(x_*(t), \dot{x}_*(t))\}$  a.e.,
- (iii)  $(-\xi, p(a_*), \eta, -p(b_*)) \in \lambda \partial g(a_*, x_*(a_*), b_*, x_*(b_*)) + N_C(a_*, x_*(a_*), b_*, x_*(b_*))$ ,
- (iv)  $\langle p(t), \dot{x}_*(t) \rangle = H(t, x_*(t), p(t))$  a.e.  $t \in [a_*, b_*]$ , and
- (v)  $\xi \in \text{ess}_{t \rightarrow a_*} H(t, x_*(a_*), p(a_*))$  and  $\eta \in \text{ess}_{t \rightarrow b_*} H(t, x_*(b_*), p(b_*))$ .

When (P) is a weak Pareto optimal problem, we have the following.

**COROLLARY 4.3.** *Suppose  $([a_*, b_*], x_*)$  is a weak Pareto solution to the multiobjective optimal problem (P) and assumptions (H1)–(H4) are satisfied. Then there exist an arc  $p \in W^{1,1}$ , real numbers  $\lambda \geq 0$ ,  $\xi$ ,  $\eta$  and a vector  $w \in R_+^m$  with  $\sum_{i=1}^m w_i = 1$  such that*

- (i)  $\lambda + \|p\|_\infty = 1$ ,
- (ii)  $\dot{p}(t) \in \text{co}\{\alpha : (\alpha, p(t)) \in N_{\text{Grph}F(t, \cdot)}(x_*(t), \dot{x}_*(t))\}$  a.e.,
- (iii)  $(-\xi, p(a_*), \eta, -p(b_*)) \in \lambda \partial \langle w, g(a_*, x_*(a_*), b_*, x_*(b_*)) \rangle + N_C(a_*, x_*(a_*), b_*, x_*(b_*))$ ,
- (iv)  $\langle p(t), \dot{x}_*(t) \rangle = H(t, x_*(t), p(t))$  for a.e.  $t \in [a_*, b_*]$ , and
- (v)  $\xi \in \text{ess}_{t \rightarrow a_*} H(t, x_*(a_*), p(a_*))$  and  $\eta \in \text{ess}_{t \rightarrow b_*} H(t, x_*(b_*), p(b_*))$ .

To provide some perspective on what we have obtained, in the rest of the paper we give an illustrative example.

**Example 4.4.** Consider the following weak Pareto optimal control problem:

$$\text{Minimize } g(x(b)) = (x_1(b) - x_2(b), x_1(b))$$

over intervals  $[0, b]$  and arcs  $x = (x_1, x_2) \in W^{1,1}([0, b], R^2)$  which satisfy

$$\begin{cases} (\dot{x}_1(t), \dot{x}_2(t)) \in F(t, x(t)), \\ b \leq 2, \\ (x_1(0), x_2(0)) = (0, -2), \end{cases}$$

where

$$F(t, x) := \begin{cases} [-1, 1] \times \{1\} & \text{if } t \leq 1, \\ \{1, t\} \times \{1\} & \text{if } t > 1. \end{cases}$$

Evidently, this is problem (P) with the initial time fixed ( $a = 0$ ) and

$$C = \{0\} \times \{(0, -2)\} \times (-\infty, 2] \times R^2.$$

For each  $w = (w_1, w_2)$ ,  $w_1 + w_2 = 1$ , we have  $\langle w, g(x(b)) \rangle = x_1(b) - w_1 x_2(b)$ . By simple computation, we have

$$H(t, (x_1, x_2), (p_1, p_2)) = \begin{cases} |p_1| + p_2 & \text{if } t \leq 1, \\ \max\{p_1 + p_2, tp_1 + p_2\} & \text{if } t > 1. \end{cases}$$

We now assume that  $([0, b], x)$  is a solution of the problem. By Corollary 4.3, there exist  $p$ , real numbers  $\lambda \geq 0$  and  $\eta$  and vector  $w = (w_1, w_2) \in R_+^2$ ,  $w_1 + w_2 = 1$  such that assertions (i)–(v) of Corollary 4.3 are satisfied.

Since

$$\text{Gph}F(t, \cdot) = \begin{cases} R^2 \times ([-1, 1] \times \{1\}) & \text{if } t \leq 1, \\ R^2 \times \{1, t\} \times \{1\} & \text{if } t > 1, \end{cases}$$

we get

$$N_{\text{Gph}F(t, \cdot)}(x(t), \dot{x}(t)) = \begin{cases} \{(0, 0)\} \times N_{[-1, 1] \times \{1\}}(\dot{x}(t)) & \text{if } t \leq 1, \\ \{(0, 0)\} \times N_{\{1, t\} \times \{1\}}(\dot{x}(t)) & \text{if } t > 1. \end{cases}$$

Hence (ii) implies  $\dot{p} = (0, 0)$ . Consequently,  $p = (p_1, p_2)$ , where  $p_1$  and  $p_2$  are constants. From (iii) we get

$$(\eta, -p(b)) \in \lambda\{0\} \times \{(1, -w_1)\} + N_{(-\infty, 2]}(b) \times \{(0, 0)\}.$$

This implies that

$$(21) \quad \eta \in N_{(-\infty, 2]}(b) \text{ and } p(b) = (p_1, p_2) = (-\lambda, \lambda w_1).$$

From (iv) of Corollary 4.3, we obtain the equation

$$p_1 \dot{x}_1 + p_2 \dot{x}_2 = \begin{cases} |p_1| + p_2 & \text{if } t \leq 1, \\ \max\{p_1 + p_2, tp_1 + p_2\} & \text{if } t > 1 \end{cases}$$

for a.e.  $t \in [0, b]$ . Since  $\dot{x}_2 = 1$ , we get  $x_2 = t - 2$  and

$$p_1 \dot{x}_1 = \begin{cases} |p_1| & \text{if } t \leq 1, \\ \max\{p_1, tp_1\} & \text{if } t > 1 \end{cases}$$

for a.e.  $t \in [0, b]$ . Since  $p_1 = -\lambda \leq 0$ , we obtain the equation

$$p_1 \dot{x}_1 = \begin{cases} -p_1 & \text{if } t \leq 1, \\ p_1 & \text{if } t > 1 \end{cases}$$

for a.e.  $t \in [0, b]$ .

We now consider the following cases.

*Case 1.* Consider  $b < 2$ . Then we have  $\eta = 0$  because of (21). By (v) we have  $0 = H(b, x(b), p(b))$ . This implies that  $|p_1| + p_2 = 0$  if  $0 \leq b \leq 1$  and  $p_1 + p_2 = 0$  if  $1 < b < 2$ . Hence if  $0 \leq b \leq 1$ , then  $p = \lambda = 0$ , which is a contradiction. Thus we must have  $1 < b < 2$  and  $0 = p_1 + p_2 = \lambda(w_1 - 1)$ . It follows that  $w_1 = 1$  and  $\lambda \neq 0$ . From above we obtain

$$x_1 = \begin{cases} -t & \text{if } 0 \leq t \leq 1, \\ t - 2 & \text{if } 1 < t \leq b. \end{cases}$$

Case 2. Consider  $b = 2$ . From (21) it follows that  $\eta \geq 0$ . By (v) we get

$$\eta \in H(2, x(2), p(2)) = p_1 + p_2 = \lambda(w_1 - 1).$$

In this case we also have  $p_1 \neq 0$ . So it yields

$$x_1 = \begin{cases} -t & \text{if } 0 \leq t \leq 1, \\ t - 2 & \text{if } 1 < t \leq 2. \end{cases}$$

Thus we showed that if  $([0, b_*], x_* = (x_{1*}, x_{2*}))$  is a solution, then  $1 < b_* \leq 2$ ,  $x_{2*} = t - 2$ , and

$$x_{1*} = \begin{cases} -t & \text{if } 0 \leq t \leq 1, \\ t - 2 & \text{if } 1 < t \leq b_*. \end{cases} \quad \square$$

**Acknowledgments.** The authors sincerely thank the referees for their helpful comments and suggestions which improved this manuscript greatly.

#### REFERENCES

- [1] C. BERGE, *Topological Spaces*, Oliver and Boyd Ltd., Edinburgh, 1963; Dover, Mineola, NY, 1987.
- [2] S. BELLAASSALI AND A. JOURANIS, *Necessary optimality conditions in multiobjective dynamic optimization*, SIAM J. Control Optim., 42 (2004), pp. 2043–2061.
- [3] V. BHASKAR, S. K. GUPTA, AND A. K. RAY, *Applications of multiobjective optimization in chemical engineering*, Rev. Chem. Eng., 16 (2000), pp. 1–54.
- [4] J. M. BORWEIN AND Q. J. ZHU, *Techniques of Variational Analysis*, Springer-Verlag, New York, 2005.
- [5] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, 1990.
- [6] G. DEBREU, *Theory of Value*, John Wiley and Sons, New York, 1959.
- [7] R. GABASOV, F. M. KIRILLOVA, AND B. MORDUKHOVICH, *The discrete maximum principle*, Dokl. Akad. Nauk SSSR, 213 (1973), pp. 19–22 (in Russian); English translation in Soviet Math. Dokl. 14 (1973), pp. 1624–1627.
- [8] A. IOFFE, *Euler–Lagrange and Hamiltonian formalisms in dynamic optimization*, Trans. AMS., 349 (1997), pp. 2871–2900.
- [9] A. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.
- [10] P. D. LOEWEN AND R. T. ROCKAFELLAR, *Optimal control of unbounded differential inclusions*, SIAM J. Control Optim., 32 (1994), pp. 442–470.
- [11] P. D. LOEWEN AND R. T. ROCKAFELLAR, *Bolza problems with general time constraints*, SIAM J. Control Optim., 35 (1997), pp. 2050–2069.
- [12] B. S. MORDUKHOVICH, *Variational Analysis and Generalized Differentiation I, II*, Springer-Verlag, Berlin, 2006.
- [13] B. S. MORDUKHOVICH AND N. M. NAM, *Variational stability and marginal functions via generalized differentiation*, Math. Oper. Res., 30 (2005), pp. 800–816.
- [14] B. S. MORDUKHOVICH, *Optimization and finite difference approximations of nonconvex differential inclusions with free time*, in Nonsmooth Analysis and Geometric Method in Deterministic Optimal Control, B. S. Mordukhovich and H. J. Sussmann, eds., Springer, New York, 1996, pp. 153–202.
- [15] B. S. MORDUKHOVICH, *Discrete approximations and refined Euler–Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.
- [16] B. S. MORDUKHOVICH, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian).
- [17] B. S. MORDUKHOVICH, *Nonsmooth analysis with nonconvex generalized differentials and conjugate mappings*, Dokl. Akad. Nauk BSSR, 28 (1984), pp. 976–979 (in Russian).
- [18] B. S. MORDUKHOVICH, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.

- [19] R. T. ROCKAFELLAR, *Hamilton–Jacobi theory and parametric analysis in fully convex problems of optimal control*, J. Global Optim., 248 (2004), pp. 419–431.
- [20] J. D. L. ROWLAND AND R. B. VINTER, *Dynamic optimization problems with free-time and active state constraints*, SIAM J. Control Optim., 31 (1993), pp. 677–697.
- [21] L. THIBAUT, *On subdifferentials of optimal value functions*, SIAM J. Control Optim., 29 (1991), pp. 1019–1036.
- [22] R. B. VINTER, *Optimal Control*, Birkhäuser, Boston, 2000.
- [23] R. B. VINTER AND H. ZHENG, *Necessary conditions for free end-time measurably time dependent optimal control problems with state constraints*, Set-Valued Anal., 8 (2000), pp. 11–29.
- [24] R. B. VINTER AND H. ZHENG, *Necessary conditions for optimal control problems with state constraints*, Trans. AMS, 350 (1998), pp. 1181–1204.
- [25] R. B. VINTER AND H. ZHENG, *The extended Euler–Lagrange condition for nonconvex variational problems*, SIAM J. Control Optim., 35 (1997), pp. 56–77.
- [26] B. VROEMEN AND B. DE JAGER, *Multiobjective control: An overview*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 440–445.
- [27] Q. J. ZHU, *Necessary optimality conditions for nonconvex differential inclusions with endpoint constraints*, J. Differential Equations, 124 (1996), pp. 186–204.
- [28] Q. J. ZHU, *Hamiltonian necessary conditions for a multiobjective optimal control problem with endpoint constraints*, SIAM J. Control Optim., 39 (2000), pp. 97–112.

# A RIESZ BASIS METHODOLOGY FOR PROPORTIONAL AND INTEGRAL OUTPUT REGULATION OF A ONE-DIMENSIONAL DIFFUSIVE-WAVE EQUATION\*

BOUMEDIÈNE CHENTOUF<sup>†</sup> AND JUN-MIN WANG<sup>‡</sup>

**Abstract.** In this article, we consider a dam-river system modeled by a diffusive-wave equation. This model is commonly used in hydraulic engineering to describe dynamic behavior of the unsteady flow in a river for shallow water when the flow variations are not important. In order to stabilize and regulate the system, we propose a proportional and integral boundary controller. Contrary to many physical systems, we end up with a nondissipative closed-loop system with noncollocated actuators and sensors. We show that the closed-loop system is a Riesz spectral system and generates an analytic semigroup. Then, we shall be able to assign the spectrum of the closed-loop system in the open left half-plane to ensure its exponential stability as well as the output regulation independently of any known or unknown constant perturbation. These results are illustrated by several numerical examples.

**Key words.** dam-river system, proportional and integral boundary control, analytic semigroup, Riesz basis, stability

**AMS subject classifications.** 35B37, 35K20, 35P10, 35P20, 47D06, 93D15, 93C20

**DOI.** 10.1137/060671188

**1. Introduction and background.** It is well known that the flow dynamics in an irrigation river are generally described by nonlinear coupled hyperbolic partial differential equations and are called de Saint-Venant equations [37]. Nevertheless, by adopting several assumptions such as (i) neglecting lateral inflow and inertia terms, (ii) assuming small flow variations as well as small bed slope of the river, (iii) observing that the flow can be reasonably represented by a one-dimensional model, and (iv) observing that the range of flow values is somewhat reduced, one can consider the following system (see [3], [18], or [31] for more details):

$$\frac{\partial Q(x, t)}{\partial t} = \alpha \frac{\partial^2 Q(x, t)}{\partial x^2} - \beta \frac{\partial Q(x, t)}{\partial x} + w, \quad 0 < x < \ell,$$

where  $Q$  is the water flow and  $w$  is the lateral discharge, which is assumed to be constant for sake of simplicity ( $w > 0$  represents the lateral inflow due to rains, for example, whereas  $w < 0$  is the lateral outflow due to water withdrawals). The positive constants  $\alpha$ ,  $\beta$ ,  $\ell$  and the variables  $x$  and  $t$  denote, respectively, the diffusion, the celerity, the length of the river, the distance in the downstream direction, and the time. This model is used in hydraulic engineering to describe dynamic behavior of the unsteady flow in a river for shallow water when the flow variations are not important (see [18] and the references therein).

---

\*Received by the editors October 1, 2006; accepted for publication (in revised form) March 31, 2008; published electronically August 13, 2008.

<http://www.siam.org/journals/sicon/47-5/67118.html>

<sup>†</sup>Department of Mathematics and Statistics, Sultan Qaboos University, P.O. Box 36, Al Khodh 123, Muscat, Sultanate of Oman (chentouf@squ.edu.om). The research of this author was supported by Sultan Qaboos University.

<sup>‡</sup>Department of Mathematics, Beijing Institute of Technology, Beijing 100081, China (wangjc@graduate.hku.hk). The research of this author was supported by the National Natural Science Foundation of China and the Program for New Century Excellent Talents in University of China.

In practice, an irrigation system often consists of natural rivers to convey water released from an upstream dam to consumption locations, which are distributed along the reach. The system under consideration consists of a dam and one river reach with a measuring station at its downstream end. The control action variable is the upstream water flow, which means that the control acts so that the desired discharge is delivered. Therefore, we shall take the upstream flow as a control input variable, take the downstream flow as an output observation, and leave the downstream boundary condition as free (no control). In other words, the downstream influence is negligible for the mass transfer, that is to say, the flow variations are negligible, which is realistic in the case of long river reaches. This leads us to the following system:

$$(1.1) \quad \begin{cases} \frac{\partial Q(x, t)}{\partial t} = \alpha \frac{\partial^2 Q(x, t)}{\partial x^2} - \beta \frac{\partial Q(x, t)}{\partial x} + w, & 0 < x < \ell, \\ Q(0, t) = u(t), \quad Q_x(\ell, t) = 0, \\ y(t) = Q(\ell, t), \end{cases}$$

where  $u(t)$  is the input (actuator) control and  $y(t)$  is the output (sensor) observation. Obviously, the actuator and sensor are not implemented at the same “location.” Hence the duality condition  $C = B^*$  does not arise, where  $B$  and  $C$  are, respectively, the input and output operators. This leads to a concrete example of a system with non-collocated actuators and sensors. As the reader may know, one of the main objectives in the management of irrigation systems is to keep the flow rate at the downstream end of the river close to a reference flow rate (target) fixed by the administration in charge of the river and calculated in order to ensure good environmental conditions for wildlife in the river.

Based on the above reasons, we shall investigate the stabilization and regulation problem of the system (1.1) with the following boundary proportional and integral controllers:

$$(1.2) \quad \begin{cases} u(t) = k_P y(t) + k_I \xi(t), \\ \dot{\xi}(t) = y(t) - y_r, \end{cases}$$

where  $k_P, k_I \in \mathbb{R} \setminus \{0\}$  are, respectively, the proportional and the integral gains, whereas  $y_r$  is the constant reference signal to track. The role of the proportional gain  $k_P$  is to speed up the exponential decay rate for stability if necessary, while the role of integral gain  $k_I$  is to obtain the regulation property, namely, (i) reject disturbances  $w$ , such as rains or withdrawals; (ii) make the output  $y(t) = Q(\ell, t)$  track a given constant reference signal  $y_r$  in spite of the constant perturbation  $w$ ; and (iii) achieve the exponential stability of the closed-loop system (1.1)–(1.2).

Note that the system (1.1) has been studied by many authors by means of approximation methods such as discretization (see, for instance, [18], [19], and [31] and the references therein), and hence the distributed character of (1.1) is not preserved. Recently, a qualitative analysis of (1.1) has been carried out in [2] by using the tools of infinite-dimensional systems theory [4]. In fact, stability and regulation results for the closed-loop system are proved when only an integral controller is applied in (1.2), that is,  $k_P = 0$ . Note also that there is a considerable literature devoted to the design theory of integral and/or proportional controllers for infinite-dimensional systems. Indeed, in [34] and [41], the authors have been concerned with the existence of proportional and integral controllers for a class of infinite-dimensional systems with *distributed controls*. Moreover, the control operators are assumed to be bounded, and



hence their results cannot be applied to our problem. Furthermore, there exists a few research papers (see [35] and [36]) where the authors proposed an integral controller for a class of infinite-dimensional systems with boundary and distributed controls. However, in these works, the authors have been primarily interested in designing *only integral controllers* for systems with *bounded boundary input and output operators* via *smooth* (twice differentiable) controls. Once again, these restrictions do not permit us to use this approach. In a related area, there is a vast literature devoted to the frequency-domain robust controller design approach for systems described by transfer functions (see, for instance, [5], [11], [12], [21], [23], [27], [28] for low-gain control; [22] and [29] for high-gain control; and [24], [25], and [26] for regular systems subject to actuator nonlinearities). Unfortunately, we have met with some technical difficulties when using this approach. Finally, the reader may also find many articles where the general theoretical results obtained in the works cited above are applied or adapted to physical systems such as heat-exchangers and linearized Saint-Venant systems (see [6], [7], [8], [9], and [39]).

The main contribution of this paper is to adopt the Riesz basis approach and the shooting method in order to extend the results obtained in [2] on the system (1.1)–(1.2) without proportional gain ( $k_P = 0$ ) to the case when the feedback control (1.2) involves a proportional gain  $k_P \neq 0$ . Contrary to the work in [2], where  $k_P = 0$  in (1.2), the advantage of the presence of the proportional gain  $k_P$  is to allow us to have more freedom in designing the stabilizing controller. Moreover, by a suitable choice of the proportional gain  $k_P$ , one can improve, if necessary, the decay rate for the stability of the closed-loop system (1.1)–(1.2) compared to the case when only an integral controller is applied. It is important to note that it is not obvious to deduce neither the well-posedness nor stability results of the closed-loop system (1.1)–(1.2) from those obtained in [2]. This is due to the fact that the domain of the operator studied in [2] is perturbed, and hence the classical theory of perturbation of operators [13] fails. We also point out that the system (1.1)–(1.2) is nondissipative and has noncollocated actuator and sensor. Hence two major difficulties arise: first, how to show the  $C_0$ -semigroup generation [33], and second, the stability for the system. To overcome this situation, the Riesz basis methodology [15], [16], [40] is used to show that the closed-loop system is a Riesz spectral system and generates an analytic semigroup. Concerning the exponential stability of the system (1.1)–(1.2), the shooting method is applied to assign the spectrum of the closed-loop system in the open left half-plane by means of an appropriate choice of the proportional and integral gains  $k_P$  and  $k_I$ . We should note that the existence results of such gains are theoretic, whereas their tuning is not dealt with here and still remain an open problem. Furthermore, the output regulation is guaranteed independently of any constant (known or unknown) perturbation.

The paper is organized as follows. In the next section, some preliminary results are stated for the uncontrolled system. We also convert the closed-loop system (1.1)–(1.2) into an evolution equation in an appropriate Hilbert space and then state the main results of this article. Sections 3 and 4 are devoted to the proof of the main results. First, we deal with the eigenvalue problem, and asymptotic expansions of both eigenvalues and eigenfunctions of the system are explicitly presented. Next, we use the Green's function approach to obtain an estimate of the resolvent which leads to the completeness of the root subspace. Then, the Riesz basis property and the  $C_0$ -semigroup generation of the system are proved. Finally, we establish, under certain conditions on the feedback gains  $k_P$  and  $k_I$ , the exponential stability and deduce the regulation of the closed-loop system (1.1)–(1.2). These results are illustrated by some

numerical applications in the last section.

**2. Preliminaries and main results.** First, let us consider the uncontrolled system ( $u(t) = 0$ ) with no disturbances:

$$(2.1) \quad \begin{cases} \frac{\partial Q(x, t)}{\partial t} = \alpha \frac{\partial^2 Q(x, t)}{\partial x^2} - \beta \frac{\partial Q(x, t)}{\partial x}, & 0 < x < \ell, \\ Q(0, t) = Q_x(\ell, t) = 0. \end{cases}$$

Taking the Hilbert state space  $\mathcal{H}_0 = L^2(0, \ell)$  equipped with the usual inner product, the system (2.1) can be written in the following abstract form:

$$\dot{Q}(t) = A_0 Q(t),$$

where  $A_0$  is an unbounded linear operator defined by

$$(2.2) \quad \mathcal{D}(A_0) := \left\{ Q \in H^2(0, \ell); Q(0) = Q'(\ell) = 0 \right\} \quad \text{and} \quad A_0 := \alpha \frac{\partial^2}{\partial x^2} - \beta \frac{\partial}{\partial x}.$$

Clearly,  $\lambda^0$  is an eigenvalue of  $A_0$  if and only if the system

$$(2.3) \quad \alpha f_0'' - \beta f_0' - \lambda^0 f_0 = 0,$$

$$(2.4) \quad f_0(0) = f_0'(\ell) = 0$$

has a nonzero solution.

The principal properties of the operator  $A_0$  are summarized as follows (see [2] for details).

LEMMA 2.1. *The operator  $A_0$  generates an exponentially stable  $C_0$ -semigroup of contractions  $S_0(t)$  on  $\mathcal{H}_0$ . Moreover, the spectrum  $\sigma(A_0)$  of  $A_0$  consists of negative real numbers of the form  $-\frac{\alpha\tau^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , where  $\tau$  is a nonzero solution to*

$$(2.5) \quad \beta\ell \sin \tau + 2\alpha\tau \cos \tau = 0.$$

Consider now the function  $x \mapsto f_0(x, \lambda)$  as the solution of (2.3) subject to the conditions

$$(2.6) \quad f_0'(\ell, \lambda) = 0, \quad f_0(\ell, \lambda) = 1.$$

Then, it is obvious that the zeros of  $f_0(0, \lambda)$  are the eigenvalues of  $A_0$ , which are real and negative by Lemma 2.5, namely,

$$\dots < \lambda_n^0 < \dots < \lambda_1^0 < 0.$$

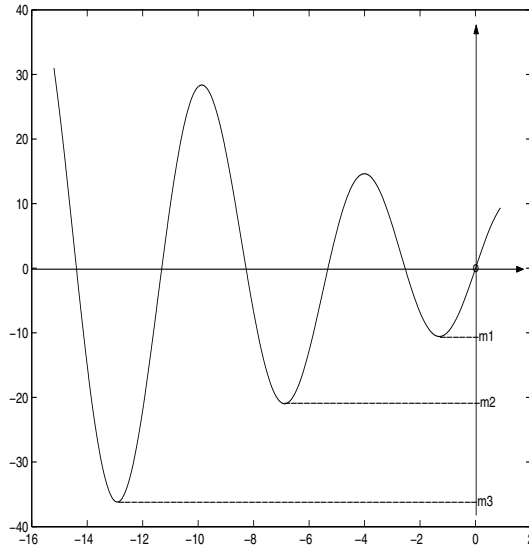
We have the following result, whose proof is given in [2].

PROPOSITION 2.2. *Consider the normalized solution of (2.3)–(2.4) by  $f_0(\ell, \lambda) = 1$ . Then the zeros of  $f_0(0, \lambda)$  are the eigenvalues  $\lambda_i^0$  of the operator  $A_0$ , which are negative real numbers and simple. In addition, we have*

$$(2.7) \quad f_0(0, \lambda) = \prod_{i=1}^{\infty} \left( 1 - \frac{\lambda}{\lambda_i^0} \right).$$

Denote

$$B(\lambda) := \lambda f_0(0, \lambda) = \lambda \prod_{i=1}^{\infty} \left( 1 - \frac{\lambda}{\lambda_i^0} \right).$$

FIG. 1. The polynomial  $B_n(\lambda)$ .

The infinite product  $B(\lambda)$  will play a crucial role in the stability of the closed-loop system (1.1)–(1.2). It follows from Lemma 2.5 and Proposition 2.2 that the zeros of  $B$ , namely,  $\lambda_i^0$ , are simple and negative. Furthermore, between two consecutive zeros, the function  $B$  attains its local maxima and minima. Thus, let

$$M_k = \min_{\lambda_{2k+1}^0 < \lambda < \lambda_{2k}^0} B(\lambda)$$

denote negative minima when  $B$  is negative between two zeros, and let  $m_k$  be the ordered set of values  $M_k$  as follows (see Figure 1):

$$(2.8) \quad \cdots < m_2 < m_1 < 0.$$

Using the same arguments as in [2], one can prove the following lemma.

**LEMMA 2.3.** *The polynomials  $B_n(\lambda) := \lambda \prod_{i=1}^n (1 - \frac{\lambda}{\lambda_i^0})$  converge uniformly to  $B(\lambda)$  in any compact domain of the complex plane. Furthermore, any compact domain of the complex plane, not containing the zeros of  $B(\cdot) - ((\cdot)k_p + k_I)$  on its boundary, contains the same number of zeros of  $B(\cdot) - ((\cdot)k_p + k_I)$  and  $B(\cdot) - ((\cdot)k_p + k_I)$  for  $n$  large enough.*

Now, let us convert the closed-loop system (1.1)–(1.2) into an evolution equation in an appropriate Hilbert space and then state the basic properties of the system operator. Clearly, the closed-loop system (1.1)–(1.2) is governed in the “augmented” state space

$$\mathcal{H} := L^2(0, \ell) \times \mathbb{C},$$

equipped with the inner product induced by the norm

$$\|(f, \xi)\|^2 = \int_0^\ell |f(x)|^2 dx + |\xi|^2 \quad \forall (f, \xi) \in \mathcal{H},$$

by the system

$$(2.9) \quad \dot{\phi}(t) = \mathcal{A}_{PI}\phi(t) + (w, -y_r).$$

Here  $\phi = (f, \xi)$  and  $\mathcal{A}_{PI}$  is an unbounded linear operator defined by

$$(2.10) \quad \mathcal{D}(\mathcal{A}_{PI}) := \left\{ (f, \xi) \in H^2(0, \ell) \times \mathbb{C}; f(0) = k_P f(\ell) + k_I \xi; f'(\ell) = 0 \right\}$$

and

$$(2.11) \quad \mathcal{A}_{PI}(f, \xi) := \left( \alpha f'' - \beta f', f(\ell) \right) \text{ for any } (f, \xi) \in \mathcal{D}(\mathcal{A}_{PI}).$$

Recall that  $\alpha$ ,  $\beta$ , and  $\ell$  are positive constants and the feedback gains  $k_P$  and  $k_I$  are nonzero numbers.

Now, one can show that, given  $(g, \eta) \in \mathcal{H}$ , the equation  $\mathcal{A}_{PI}(f, \xi) = (g, \eta)$  has a unique solution  $(f, \xi) \in \mathcal{D}(\mathcal{A}_{PI})$  given by

$$\begin{cases} f(x) = \eta + \frac{1}{\beta} \left[ \int_x^\ell (1 - e^{\omega(x-\xi)}) g(\xi) d\xi \right], \\ \xi = \frac{1}{k_I} (f(0) - k_P f(\ell)), \end{cases}$$

where  $\omega = \beta/\alpha$ . Therefore, the operator  $(\mathcal{A}_{PI})^{-1} \in \mathcal{L}(\mathcal{H})$ . Moreover, by the Sobolev embedding theorem [1], we deduce that  $(\mathcal{A}_{PI})^{-1}$  is compact on  $\mathcal{H}$ , and hence the spectrum  $\sigma(\mathcal{A}_{PI})$  consists of isolated eigenvalues only [13].

Before stating our main results, let us recall that a nonzero  $Y$ , of a Hilbert space  $H$ , is called a generalized eigenvector of a linear operator  $\mathcal{A}$ , corresponding to an eigenvalue  $\lambda$  (with finite algebraic multiplicity) of  $\mathcal{A}$ , if there is a positive integer  $n$  such that  $(\lambda - \mathcal{A})^n Y = 0$ .

Let  $\text{Sp}(\mathcal{A})$  be the root subspace of a linear operator  $\mathcal{A}$  which is defined as the closed subspace spanned by all generalized eigenvectors of  $\mathcal{A}$ .

A sequence in  $H$  is said to be complete if its linear span is dense in  $\mathcal{H}$ . Also, a sequence in  $H$  is called *minimal* if each element of this sequence lies outside the closed linear span of the remaining elements. In turn, two sequences  $\{e_i\}$  and  $\{e_i^*\}$  are said to be *biorthogonal* in  $H$  if

$$\langle e_i, e_j^* \rangle = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

for every  $i$  and  $j$ . It is well known that for a given sequence  $\{e_i\}$ , a biorthogonal sequence  $\{e_i^*\}$  exists if and only if  $\{e_i\}$  is minimal, and  $\{e_i^*\}$  is uniquely determined if and only if  $\{e_i\}$  is complete.

Now, a sequence  $\{e_i\}_{i=1}^\infty$  is called a Bessel sequence in  $H$  if for any  $x \in H$ , the series  $\{\langle x, e_i \rangle\}_{i=1}^\infty \in \ell^2$ . On the other hand, a sequence  $\{e_i\}_{i=1}^\infty$  is called a basis for  $H$  if any element  $x \in H$  has a unique representation,

$$(2.12) \quad x = \sum_{i=1}^{\infty} a_i e_i,$$

and the convergence of the series is in the norm of  $H$ . Finally, a sequence  $\{e_i\}_{i=1}^\infty$  with a biorthogonal sequence  $\{e_i^*\}_{i=1}^\infty$  is called a Riesz basis for  $H$  if  $\{e_i\}_{i=1}^\infty$  is an

approximately normalized basis of  $H$  and the series in (2.12) converges unconditionally in the norm of  $H$ . Another different, but equivalent, condition for a Riesz basis property has the following form (see [42, p. 27]):

- (a) both  $\{e_i\}_{i=1}^\infty$  and  $\{e_i^*\}_{i=1}^\infty$  are complete in  $H$ ; and
- (b) both  $\{e_i\}_{i=1}^\infty$  and  $\{e_i^*\}_{i=1}^\infty$  are Bessel sequences in  $H$ .

It is also well known that  $\{e_i\}_{i=1}^\infty$  is a Riesz basis for  $H$  if and only if its biorthogonal sequence  $\{e_i^*\}_{i=1}^\infty$  is a Riesz basis for  $H$ .

Now, we are able to state the main results of this work. Indeed, the first main result is related to the Riesz basis property.

**THEOREM 2.4.** *Let  $|\tilde{k}_P| \neq 1$ , where  $\tilde{k}_P := k_P e^{\frac{1}{2} \frac{\ell \beta}{\alpha}}$ . Then, the generalized eigenfunctions of the operator  $\mathcal{A}_{PI}$  form a Riesz basis in  $\mathcal{H}$ . In turn, if  $|\tilde{k}_P| = 1$ , then the generalized eigenfunctions of  $\mathcal{A}_{PI}$  form, in  $\mathcal{H}$ , a Riesz basis with parentheses.*

The second result is mainly concerned with the properties of the semigroup generated by the operator  $\mathcal{A}_{PI}$ .

**THEOREM 2.5.** *The operator  $\mathcal{A}_{PI}$  defined by (2.10)–(2.11) generates a  $C_0$ -semigroup  $S_{PI}(t)$  in  $\mathcal{H}$ . Therefore,  $S_{PI}(t)$  satisfies the spectrum-determined growth condition  $S(\mathcal{A}_{PI}) = \omega(\mathcal{A}_{PI})$ , where  $S(\mathcal{A}_{PI}) = \sup_{\lambda \in \sigma(\mathcal{A}_{PI})} \operatorname{Re} \lambda$  is the spectral bound and  $\omega(\mathcal{A}_{PI})$  is the growth order of the semigroup  $S_{PI}(t)$ . Moreover,  $S_{PI}(t)$  is an analytic semigroup in  $\mathcal{H}$ .*

Finally, regarding the third main result, we use notation from (2.8) and Figure 1 to state the stability and regulation properties of the closed-loop system.

**THEOREM 2.6.** (i) *If  $m_1 < k_I < 0$  and  $k_P$  is a sufficiently small positive number, then the spectrum  $\sigma(\mathcal{A}_{PI})$  consists of negative real numbers.*

(ii) *If  $m_2 < k_I < m_1 < 0$  and  $k_P$  is a positive number sufficiently small, then the spectrum  $\sigma(\mathcal{A}_{PI})$  of the operator  $\mathcal{A}_{PI}$  consists of negative real numbers, except two which are complex conjugate numbers with positive or negative real parts.*

*Therefore, in both cases where the spectrum of the operator  $\mathcal{A}_{PI}$ , defined by (2.10)–(2.11) has negative real part, the analytic semigroup  $S_{PI}(t)$  is exponentially stable. Moreover, for any initial data  $\phi_0 = (Q_0, \xi_0) \in \mathcal{D}(\mathcal{A}_{PI})$  and for any constant perturbation  $w$  in  $\mathcal{H}$ , we have the output regulation*

$$\lim_{t \rightarrow \infty} y(t) = \lim_{t \rightarrow \infty} Q(\ell, t) = y_r.$$

*Furthermore, the closed-loop system (1.1)–(1.2) is exponentially stable in  $\mathcal{H}$  in spite of the constant perturbation  $w \in \mathcal{H}$ .*

**3. Proof of Theorems 2.4 and 2.5.** In this section, we are going to show Theorems 2.4 and 2.5, namely, the well-posedness and the Riesz basis generation of the system. For sake of clarity, we divide this section into three parts.

**3.1. Eigenvalue problem.** We shall establish, in this subsection, the basic results of the eigenvalue problem related to the operator  $\mathcal{A}_{PI}$ . From the eigenvalue equation  $\mathcal{A}_{PI}(f, \xi) = \lambda(f, \xi)$ , we have the characteristic equation in  $\lambda$ :

$$(3.1) \quad \begin{cases} f''(x) - \frac{\beta}{\alpha} f'(x) - \frac{1}{\alpha} \lambda f(x) = 0, & 0 < x < \ell, \\ (\lambda k_P + k_I) f(\ell) = \lambda f(0), \quad f'(\ell) = 0. \end{cases}$$

In order to solve the above equation, we introduce the following transformation that can translate the interval  $[0, \ell]$  into  $[0, 1]$ :

$$(3.2) \quad x = z\ell, \quad f(x) = g(z), \quad z \in [0, 1].$$

Then, (3.1) changes into

$$(3.3) \quad \begin{cases} g''(z) - \frac{\ell\beta}{\alpha}g'(z) - \frac{\ell^2}{\alpha}\lambda g(z) = 0, & 0 < z < 1, \\ (\lambda k_P + k_I)g(1) = \lambda g(0), \quad g'(1) = 0. \end{cases}$$

Let

$$(3.4) \quad \tau_1(\lambda) = \frac{\ell\beta + \ell\sqrt{\beta^2 + 4\alpha\lambda}}{2\alpha}, \quad \tau_2(\lambda) = \frac{\ell\beta - \ell\sqrt{\beta^2 + 4\alpha\lambda}}{2\alpha}.$$

Clearly,  $e^{\tau_1 z}$  and  $e^{\tau_2 z}$  are two independent solutions of  $g''(z) - \frac{\ell\beta}{\alpha}g'(z) - \frac{\ell^2}{\alpha}\lambda g(z) = 0$ , and thus the general solution form of (3.3) can be given by

$$g(z) = c_1 e^{\tau_1 z} + c_2 e^{\tau_2 z},$$

where  $c_1$  and  $c_2$  satisfy the following system:

$$\begin{cases} (\lambda k_P + k_I)(c_1 e^{\tau_1} + c_2 e^{\tau_2}) = \lambda(c_1 + c_2), \\ c_1 \tau_1 e^{\tau_1} + c_2 \tau_2 e^{\tau_2} = 0. \end{cases}$$

Hence, (3.3) has a nontrivial solution if and only if

$$\det(\Delta(\lambda)) = 0,$$

where  $\Delta(\lambda)$  is the coefficient matrix given by

$$(3.5) \quad \Delta(\lambda) := \begin{bmatrix} (\lambda k_P + k_I)e^{\tau_1} - \lambda & (\lambda k_P + k_I)e^{\tau_2} - \lambda \\ \tau_1 e^{\tau_1} & \tau_2 e^{\tau_2} \end{bmatrix}.$$

By a direct computation, we have

$$\begin{aligned} \det(\Delta(\lambda)) &= (\tau_2 - \tau_1)(\lambda k_P + k_I)e^{\tau_1 + \tau_2} - \lambda(\tau_2 e^{\tau_2} - \tau_1 e^{\tau_1}) \\ &= -\frac{\ell\sqrt{\beta^2 + 4\alpha\lambda}}{\alpha}(\lambda k_P + k_I)e^{(\ell\beta)/\alpha} - \lambda(\tau_2 e^{\tau_2} - \tau_1 e^{\tau_1}). \end{aligned}$$

Let  $\lambda := \frac{\alpha\rho^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , with  $\rho \in \mathbb{C}$  and  $\arg \rho \in [-\pi/2, \pi/2)$ . Then  $\tau_1(\lambda)$  and  $\tau_2(\lambda)$ , defined by (3.4), become

$$(3.6) \quad \tau_1(\rho) = \frac{\ell\beta}{2\alpha} + \rho, \quad \tau_2(\rho) = \frac{\ell\beta}{2\alpha} - \rho.$$

We also have

$$\begin{aligned} (3.7) \quad e^{-\frac{\ell\beta}{2\alpha}} \det(\Delta(\rho)) &= -2\rho \left( \frac{\alpha\rho^2}{\ell^2} - \frac{\beta^2}{4\alpha} \right) \tilde{k}_P - 2\rho \tilde{k}_I \\ &\quad - \left( \frac{\alpha\rho^2}{\ell^2} - \frac{\beta^2}{4\alpha} \right) \left( \frac{\ell\beta}{2\alpha} e^{-\rho} - \rho e^{-\rho} - \frac{\ell\beta}{2\alpha} e^{\rho} - \rho e^{\rho} \right) \\ &= \frac{\alpha\rho^3}{\ell^2} (e^{-\rho} + e^{\rho} - 2\tilde{k}_P) - \frac{\beta\rho^2}{2\ell} (e^{-\rho} - e^{\rho}) \\ &\quad - \frac{\beta^2}{4\alpha} \rho \left( e^{-\rho} + e^{\rho} - 2\tilde{k}_P + \frac{8\alpha}{\beta^2} \tilde{k}_I \right) + \frac{\ell\beta^3}{8\alpha^2} (e^{-\rho} - e^{\rho}), \end{aligned}$$

where

$$(3.8) \quad \tilde{k}_P := k_P e^{\frac{1}{2} \frac{\ell\beta}{\alpha}}, \quad \tilde{k}_I := k_I e^{\frac{1}{2} \frac{\ell\beta}{\alpha}}.$$

Let us summarize the previous results in the following lemma.

LEMMA 3.1. *Let  $\lambda := \frac{\alpha\rho^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , with  $\arg \rho \in [-\pi/2, \pi/2)$ . Then the characteristic determinant  $\det(\Delta(\rho))$  has the following form:*

$$(3.9) \quad e^{-\frac{\ell\beta}{2\alpha}} \det(\Delta(\rho)) = \frac{\alpha\rho^3}{\ell^2} (e^{-\rho} + e^{\rho} - 2\tilde{k}_P) - \frac{\beta\rho^2}{2\ell} (e^{-\rho} - e^{\rho}) \\ - \frac{\beta^2}{4\alpha} \rho \left( e^{-\rho} + e^{\rho} - 2\tilde{k}_P + \frac{8\alpha}{\beta^2} \tilde{k}_I \right) + \frac{\ell\beta^3}{8\alpha^2} (e^{-\rho} - e^{\rho}),$$

with  $\tilde{k}_P$  and  $\tilde{k}_I$  being given by (3.8).

Then, we have the following.

THEOREM 3.2. *Let  $|\tilde{k}_P| > 1$ . Then the eigenvalues  $\lambda_n$  of the operator  $\mathcal{A}_{PI}$  are the solutions of the eigenvalue problem (3.3) and have the following asymptotic expansion: for  $s = 1, 2$ ,*

$$(3.10) \quad \lambda_{ns} = - \left( \frac{\beta}{\ell} + \frac{\beta^2}{4\alpha} \right) + \frac{(\ln |\tilde{k}_{Ps}|)^2 - (2n\pi + \delta)^2 + i(4n\pi + 2\delta) \ln |\tilde{k}_{Ps}|}{\alpha^{-1}\ell^2} + \mathcal{O}(n^{-1}),$$

where  $n$  are positive integers,

$$(3.11) \quad \tilde{k}_{P1} := \tilde{k}_P + \sqrt{\tilde{k}_P^2 - 1}, \quad \tilde{k}_{P2} := \tilde{k}_P - \sqrt{\tilde{k}_P^2 - 1},$$

and

$$(3.12) \quad \delta := \begin{cases} 0 & \text{when } \tilde{k}_P \geq 1; \\ \pi & \text{when } \tilde{k}_P \leq -1. \end{cases}$$

In turn, if  $|\tilde{k}_P| = 1$ , then the eigenvalues  $\lambda_{n1}$  and  $\lambda_{n2}$  are not separable when their moduli are large enough and have the following asymptotic expression:

$$(3.13) \quad \lambda_{ns} = -\frac{\beta^2}{4\alpha} - \frac{\alpha(2n\pi + \delta)^2}{\ell^2} + \mathcal{O}(n^{-1}), \quad s = 1, 2.$$

*Proof.* Using (3.9), the characteristic equation  $\det(\Delta(\rho)) = 0$  means that  $\rho$ , with  $\arg \rho \in [-\pi/2, \pi/2)$ , satisfies the following:

$$(3.14) \quad e^{\rho} + e^{-\rho} - 2\tilde{k}_P - \frac{1}{2} \frac{\ell\beta}{\alpha} (e^{-\rho} - e^{\rho}) \rho^{-1} + \mathcal{O}(\rho^{-2}) = 0,$$

which leads to

$$(3.15) \quad e^{\rho} + e^{-\rho} - 2\tilde{k}_P + \mathcal{O}(\rho^{-1}) = 0.$$

Assume now that  $|\tilde{k}_P| > 1$ . Then, it follows that the solutions of the equation

$$e^{\rho} + e^{-\rho} - 2\tilde{k}_P = 0$$

are given by

$$(3.16) \quad \tilde{\rho}_{ns} = \ln |\tilde{k}_{Ps}| + 2n\pi i + \delta i, \quad s = 1, 2, \quad n \in \mathbb{N},$$

where  $\delta$  is as defined in (3.12). Rouché's theorem can be applied to (3.15) to obtain

$$(3.17) \quad \rho_{ns} = \tilde{\rho}_{ns} + \alpha_{ns}, \quad \alpha_{ns} = \mathcal{O}(n^{-1}), \quad s = 1, 2,$$

for sufficiently large positive integers  $n$ . Substituting  $\rho_{ns}$  into (3.14), we get

$$e^{\tilde{\rho}_{ns} + \alpha_{ns}} + e^{-\tilde{\rho}_{ns} - \alpha_{ns}} - 2\tilde{k}_P - \frac{1}{2} \frac{\ell\beta}{\alpha} (e^{-\tilde{\rho}_{ns} - \alpha_{ns}} - e^{\tilde{\rho}_{ns} + \alpha_{ns}}) \rho_{ns}^{-1} + \mathcal{O}(\rho_{ns}^{-2}) = 0.$$

Note that

$$e^{\tilde{\rho}_{ns}} + e^{-\tilde{\rho}_{ns}} - 2\tilde{k}_P = 0, \quad e^{\tilde{\rho}_{ns}} = \tilde{k}_{P_s},$$

and hence

$$\begin{aligned} e^{\tilde{\rho}_{ns} + \alpha_{ns}} + e^{-\tilde{\rho}_{ns} - \alpha_{ns}} &= e^{\alpha_{ns}} \left[ 2\tilde{k}_P - e^{-\tilde{\rho}_{ns}} \right] + e^{-\tilde{\rho}_{ns} - \alpha_{ns}} \\ &= 2\tilde{k}_P e^{\alpha_{ns}} - e^{\alpha_{ns}} e^{-\tilde{\rho}_{ns}} + e^{-\tilde{\rho}_{ns} - \alpha_{ns}} \\ &= 2\tilde{k}_P [1 + \alpha_{ns} + \mathcal{O}(n^{-2})] - e^{-\tilde{\rho}_{ns}} [e^{\alpha_{ns}} - e^{-\alpha_{ns}}] \\ &= 2\tilde{k}_P + 2\alpha_{ns} \tilde{k}_P + \mathcal{O}(n^{-2}) - e^{-\tilde{\rho}_{ns}} [2\alpha_{ns} + \mathcal{O}(n^{-2})] \\ &= 2\tilde{k}_P + 2\alpha_{ns} \tilde{k}_P + \mathcal{O}(n^{-2}) - \tilde{k}_{P_s}^{-1} [2\alpha_{ns} + \mathcal{O}(n^{-2})] \\ &= 2\tilde{k}_P + 2\alpha_{ns} (\tilde{k}_P - \tilde{k}_{P_s}^{-1}) + \mathcal{O}(n^{-2}), \end{aligned}$$

where we have also expanded the exponential functions according to their Taylor series. Similarly,

$$\begin{aligned} e^{\tilde{\rho}_{ns} + \alpha_{ns}} - e^{-\tilde{\rho}_{ns} - \alpha_{ns}} &= e^{\alpha_{ns}} \left[ 2\tilde{k}_P - e^{-\tilde{\rho}_{ns}} \right] - e^{-\tilde{\rho}_{ns} - \alpha_{ns}} \\ &= 2\tilde{k}_P e^{\alpha_{ns}} - e^{\alpha_{ns}} e^{-\tilde{\rho}_{ns}} - e^{-\tilde{\rho}_{ns} - \alpha_{ns}} \\ &= 2\tilde{k}_P [1 + \alpha_{ns} + \mathcal{O}(n^{-2})] - e^{-\tilde{\rho}_{ns}} [e^{\alpha_{ns}} + e^{-\alpha_{ns}}] \\ &= 2\tilde{k}_P + 2\alpha_{ns} \tilde{k}_P + \mathcal{O}(n^{-2}) - e^{-\tilde{\rho}_{ns}} [2 + \mathcal{O}(n^{-2})] \\ &= 2\tilde{k}_P + 2\alpha_{ns} \tilde{k}_P + \mathcal{O}(n^{-2}) - \tilde{k}_{P_s}^{-1} [2 + \mathcal{O}(n^{-2})] \\ &= 2 (\tilde{k}_P - \tilde{k}_{P_s}^{-1}) + 2\tilde{k}_P \alpha_{ns} + \mathcal{O}(n^{-2}). \end{aligned}$$

These two estimates lead us to write

$$\begin{aligned} 0 &= e^{\tilde{\rho}_{ns} + \alpha_{ns}} + e^{-\tilde{\rho}_{ns} - \alpha_{ns}} - 2\tilde{k}_P - \frac{1}{2} \frac{\ell\beta}{\alpha} (e^{-\tilde{\rho}_{ns} - \alpha_{ns}} - e^{\tilde{\rho}_{ns} + \alpha_{ns}}) \rho_{ns}^{-1} + \mathcal{O}(\rho_{ns}^{-2}) \\ &= 2\tilde{k}_P + 2\alpha_{ns} (\tilde{k}_P - \tilde{k}_{P_s}^{-1}) - 2\tilde{k}_P \\ &\quad + \frac{1}{2} \frac{\ell\beta}{\alpha} \left( 2 (\tilde{k}_P - \tilde{k}_{P_s}^{-1}) + 2\tilde{k}_P \alpha_{ns} + \mathcal{O}(n^{-2}) \right) \tilde{\rho}_{ns}^{-1} + \mathcal{O}(n^{-2}) \\ &= 2\alpha_{ns} (\tilde{k}_P - \tilde{k}_{P_s}^{-1}) + \frac{1}{2} \frac{\ell\beta}{\alpha} \left( 2 (\tilde{k}_P - \tilde{k}_{P_s}^{-1}) + 2\tilde{k}_P \alpha_{ns} \right) \tilde{\rho}_{ns}^{-1} + \mathcal{O}(n^{-2}) \\ &= 2\alpha_{ns} (\tilde{k}_P - \tilde{k}_{P_s}^{-1}) + \frac{\ell\beta}{\alpha} (\tilde{k}_P - \tilde{k}_{P_s}^{-1}) \tilde{\rho}_{ns}^{-1} + \mathcal{O}(n^{-2}). \end{aligned}$$



Thus, we obtain

$$\alpha_{ns} = -\frac{\ell\beta}{2\alpha\tilde{\rho}_{ns}} + \mathcal{O}(n^{-2}) = -\frac{\ell\beta}{2\alpha\left(\ln|\tilde{k}_{Ps}| + 2n\pi i + \delta i\right)} + \mathcal{O}(n^{-2}).$$

This, together with (3.16)–(3.17), yields

$$(3.18) \quad \rho_{ns} = \ln|\tilde{k}_{Ps}| + 2n\pi i + \delta i - \frac{\ell\beta}{2\alpha\left(\ln|\tilde{k}_{Ps}| + 2n\pi i + \delta i\right)} + \mathcal{O}(n^{-2}),$$

and so

$$\rho_{ns}^2 = (\ln|\tilde{k}_{Ps}|)^2 - (2n\pi + \delta)^2 + i(4n\pi + 2\delta)\ln|\tilde{k}_{Ps}| - \frac{\ell\beta}{\alpha} + \mathcal{O}(n^{-1}).$$

Now, using the fact that  $\lambda_{ns} = \frac{\alpha\rho_{ns}^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , the desired result (3.10) is directly obtained.

Finally, when  $|\tilde{k}_P| = 1$ , similar arguments permit us to claim that  $\lambda_{n1}$  and  $\lambda_{n2}$ , for  $n \in \mathbb{N}$ , are not separable when their moduli are large enough and have the asymptotic expression given by (3.13). We omit the details here.  $\square$

The following result deals with the case  $|\tilde{k}_P| < 1$ .

**THEOREM 3.3.** *Let  $|\tilde{k}_P| < 1$ . Then the eigenvalues  $\lambda_n$  of the operator  $\mathcal{A}_{PI}$  are the solutions of the eigenvalue problem (3.3) and have the following asymptotic expansion: for  $s = 1, 2$ ,*

$$(3.19) \quad \lambda_{ns} = -\left(\frac{\beta}{\ell} + \frac{\beta^2}{4\alpha}\right) - \frac{\alpha(2n\pi + \theta_s)^2}{\ell^2} + \mathcal{O}(n^{-1}),$$

where  $n$  are positive integers and

$$(3.20) \quad \theta_1 := \tan^{-1}\left(\frac{\sqrt{1 - \tilde{k}_P^2}}{\tilde{k}_P}\right), \quad \theta_2 := -\theta_1.$$

*Proof.* Let  $\rho \in \mathbb{C}$  with  $\arg \rho \in [-\pi/2, \pi/2)$ . When  $|\tilde{k}_P| < 1$ , the equation

$$e^\rho + e^{-\rho} - 2\tilde{k}_P = 0$$

has solutions

$$(3.21) \quad \tilde{\rho}_{ns} = \theta_s i + 2n\pi i, \quad s = 1, 2, \quad n \in \mathbb{N}.$$

Applying Rouché's theorem to (3.15) yields

$$(3.22) \quad \rho_{ns} = \tilde{\rho}_{ns} + \alpha_{ns}, \quad \alpha_{ns} = \mathcal{O}(n^{-1}), \quad s = 1, 2,$$

for sufficiently large positive integers  $n$ . Next, inserting  $\rho_{ns}$  into (3.14), we obtain, after a careful computation,

$$e^{\tilde{\rho}_{ns} + \alpha_{ns}} + e^{-\tilde{\rho}_{ns} - \alpha_{ns}} - 2\tilde{k}_P - \frac{1}{2} \frac{\ell\beta}{\alpha} (e^{-\tilde{\rho}_{ns} - \alpha_{ns}} - e^{\tilde{\rho}_{ns} + \alpha_{ns}}) \rho_{ns}^{-1} + \mathcal{O}(\rho_{ns}^{-2}) = 0.$$

Then, since

$$e^{\tilde{\rho}_{ns}} + e^{-\tilde{\rho}_{ns}} - 2\tilde{k}_P = 0, \quad e^{\tilde{\rho}_{ns}} = e^{i\theta_s} = \tilde{k}_P + i\sqrt{1 - \tilde{k}_P^2},$$

one can get, as for the case  $|\tilde{k}_P| > 1$ , the estimates

$$\begin{aligned}
 e^{\tilde{\rho}_{ns} + \alpha_{ns}} + e^{-\tilde{\rho}_{ns} - \alpha_{ns}} &= e^{\alpha_{ns}} \left[ 2\tilde{k}_P - e^{-\tilde{\rho}_{ns}} \right] + e^{-\tilde{\rho}_{ns} - \alpha_{ns}} \\
 &= 2\tilde{k}_P e^{\alpha_{ns}} - e^{\alpha_{ns}} e^{-\tilde{\rho}_{ns}} + e^{-\tilde{\rho}_{ns} - \alpha_{ns}} \\
 &= 2\tilde{k}_P \left[ 1 + \alpha_{ns} + \mathcal{O}(n^{-2}) \right] - e^{-\tilde{\rho}_{ns}} \left[ e^{\alpha_{ns}} - e^{-\alpha_{ns}} \right] \\
 &= 2\tilde{k}_P + 2\alpha_{ns}\tilde{k}_P + \mathcal{O}(n^{-2}) - e^{-\tilde{\rho}_{ns}} \left[ 2\alpha_{ns} + \mathcal{O}(n^{-2}) \right] \\
 &= 2\tilde{k}_P + 2\alpha_{ns}\tilde{k}_P + \mathcal{O}(n^{-2}) - e^{-i\theta_s} \left[ 2\alpha_{ns} + \mathcal{O}(n^{-2}) \right] \\
 &= 2\tilde{k}_P + 2\alpha_{ns}\tilde{k}_P + \mathcal{O}(n^{-2}) - \left( \tilde{k}_P - i\sqrt{1 - \tilde{k}_P^2} \right) \left[ 2\alpha_{ns} + \mathcal{O}(n^{-2}) \right] \\
 &= 2\tilde{k}_P + 2i\alpha_{ns}\sqrt{1 - \tilde{k}_P^2} + \mathcal{O}(n^{-2})
 \end{aligned}$$

and

$$\begin{aligned}
 e^{\tilde{\rho}_{ns} + \alpha_{ns}} - e^{-\tilde{\rho}_{ns} - \alpha_{ns}} &= e^{\alpha_{ns}} \left[ 2\tilde{k}_P - e^{-\tilde{\rho}_{ns}} \right] - e^{-\tilde{\rho}_{ns} - \alpha_{ns}} \\
 &= 2\tilde{k}_P e^{\alpha_{ns}} - e^{\alpha_{ns}} e^{-\tilde{\rho}_{ns}} - e^{-\tilde{\rho}_{ns} - \alpha_{ns}} \\
 &= 2\tilde{k}_P \left[ 1 + \alpha_{ns} + \mathcal{O}(n^{-2}) \right] - e^{-\tilde{\rho}_{ns}} \left[ e^{\alpha_{ns}} + e^{-\alpha_{ns}} \right] \\
 &= 2\tilde{k}_P + 2\alpha_{ns}\tilde{k}_P + \mathcal{O}(n^{-2}) - e^{-\tilde{\rho}_{ns}} \left[ 2 + \mathcal{O}(n^{-2}) \right] \\
 &= 2\tilde{k}_P + 2\alpha_{ns}\tilde{k}_P + \mathcal{O}(n^{-2}) - e^{-i\theta_s} \left[ 2 + \mathcal{O}(n^{-2}) \right] \\
 &= 2\tilde{k}_P + 2\alpha_{ns}\tilde{k}_P + \mathcal{O}(n^{-2}) - \left( \tilde{k}_P - i\sqrt{1 - \tilde{k}_P^2} \right) \left[ 2 + \mathcal{O}(n^{-2}) \right] \\
 &= 2i\sqrt{1 - \tilde{k}_P^2} + 2\tilde{k}_P\alpha_{ns} + \mathcal{O}(n^{-2}).
 \end{aligned}$$

Hence

$$\begin{aligned}
 0 &= e^{\tilde{\rho}_{ns} + \alpha_{ns}} + e^{-\tilde{\rho}_{ns} - \alpha_{ns}} - 2\tilde{k}_P - \frac{1}{2} \frac{\ell\beta}{\alpha} \left( e^{-\tilde{\rho}_{ns} - \alpha_{ns}} - e^{\tilde{\rho}_{ns} + \alpha_{ns}} \right) \rho_{ns}^{-1} + \mathcal{O}(\rho_{ns}^{-2}) \\
 &= 2\tilde{k}_P + 2i\alpha_{ns}\sqrt{1 - \tilde{k}_P^2} - 2\tilde{k}_P + \frac{1}{2} \frac{\ell\beta}{\alpha} \left( 2i\sqrt{1 - \tilde{k}_P^2} + 2\tilde{k}_P\alpha_{ns} + \mathcal{O}(n^{-2}) \right) \tilde{\rho}_{ns}^{-1} + \mathcal{O}(n^{-2}) \\
 &= 2i\alpha_{ns}\sqrt{1 - \tilde{k}_P^2} + \frac{1}{2} \frac{\ell\beta}{\alpha} \left( 2i\sqrt{1 - \tilde{k}_P^2} + 2\tilde{k}_P\alpha_{ns} \right) \tilde{\rho}_{ns}^{-1} + \mathcal{O}(n^{-2}) \\
 &= 2i\alpha_{ns}\sqrt{1 - \tilde{k}_P^2} + i\frac{\ell\beta}{\alpha} \sqrt{1 - \tilde{k}_P^2} \tilde{\rho}_{ns}^{-1} + \mathcal{O}(n^{-2}),
 \end{aligned}$$

which, together with (3.21), gives

$$\alpha_{ns} = -\frac{\ell\beta}{2\alpha} \tilde{\rho}_{ns}^{-1} + \mathcal{O}(n^{-2}) = -\frac{\ell\beta}{2\alpha(\theta_s i + 2n\pi i)} + \mathcal{O}(n^{-2}).$$

Therefore

$$(3.23) \quad \rho_{ns} = \theta_s i + 2n\pi i - \frac{\ell\beta}{2\alpha(\theta_s i + 2n\pi i)} + \mathcal{O}(n^{-2}),$$

and so

$$\rho_{ns}^2 = -(2n\pi + \theta_s)^2 - \frac{\ell\beta}{\alpha} + \mathcal{O}(n^{-1}).$$

Finally, since  $\lambda_{ns} = \frac{\alpha\rho_{ns}^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , one can deduce (3.19). This completes the proof.  $\square$

We are now in a position to investigate the asymptotic behavior of the eigenfunctions. The result is described as follows.

**THEOREM 3.4.** *Let  $\sigma(\mathcal{A}_{PI}) = \{\lambda_{n1}, \lambda_{n2}, n \in \mathbb{N}\}$  be the eigenvalues of  $\mathcal{A}_{PI}$ . If  $|\tilde{k}_P| > 1$  (respectively,  $|\tilde{k}_P| < 1$ ), then  $\lambda_n = \frac{\alpha\rho_n^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , with  $\rho_n \in \mathbb{C}$  and  $\arg \rho_n \in [-\pi/2, \pi/2)$ , are given by (3.10) and (3.18) (respectively, (3.19) and (3.23)). Furthermore, the corresponding eigenfunctions  $\{(f_{n1}, \xi_{n1}), (f_{n2}, \xi_{n2})\}$  have the following asymptotics: for  $s = 1, 2$ ,*

$$(3.24) \quad \begin{cases} f_{ns}(x) = g_{ns}(z) = e^{\frac{1}{2}\frac{\ell\beta}{\alpha}z} e^{-\rho_{ns}(1-z)} + e^{\frac{1}{2}\frac{\ell\beta}{\alpha}z} e^{\rho_{ns}(1-z)} + \mathcal{O}(n^{-1}) \\ \quad = e^{\frac{1}{2}\frac{\beta}{\alpha}x} e^{-\rho_{ns}(1-\frac{x}{\ell})} + e^{\frac{1}{2}\frac{\beta}{\alpha}x} e^{\rho_{ns}(1-\frac{x}{\ell})} + \mathcal{O}(n^{-1}), \quad x = z\ell, \\ \xi_{ns} = \frac{g_{ns}(0) - k_P g_{ns}(1)}{k_I} = \mathcal{O}(n^{-1}) \end{cases}$$

for sufficiently large positive integer  $n$ , where  $f_{ns}(x) = g_{ns}(z)$  with  $x = z\ell$  given in (3.2). Moreover,  $\{(f_{n1}, \xi_{n1}), (f_{n2}, \xi_{n2})\}$  is approximately normalized in  $\mathcal{H}$  in the sense that there exist positive constants  $c_1, c_2$  independent of  $n$ , such that

$$(3.25) \quad c_1 \leq \|f_{ns}\|_{L^2(0,\ell)} = \sqrt{\ell} \|g_{ns}\|_{L^2(0,1)}, \quad |\xi_{ns}| \leq c_2, \quad n \in \mathbb{N}, \quad s = 1, 2,$$

*Proof.* We consider only the case  $|\tilde{k}_P| > 1$ , and the same arguments can be applied when  $|\tilde{k}_P| < 1$ . From (3.3), (3.5), (3.6), and linear algebra theory, the function  $g$ , with respect to the eigenvalue  $\lambda = \frac{\alpha\rho^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , where  $\rho \in \mathbb{C}$  and  $\arg \rho \in [-\pi/2, \pi/2)$ , is given by

$$(3.26) \quad e^{-\frac{1}{2}\frac{\ell\beta}{\alpha}z} g(z, \rho) = e^{-\frac{1}{2}\frac{\ell\beta}{\alpha}z} \begin{vmatrix} e^{\tau_1 z} & e^{\tau_2 z} \\ \tau_1 e^{\tau_1} & \tau_2 e^{\tau_2} \end{vmatrix} = \tau_2 e^{\frac{1}{2}\frac{\ell\beta}{\alpha}z} e^{-\rho(1-z)} - \tau_1 e^{\frac{1}{2}\frac{\ell\beta}{\alpha}z} e^{\rho(1-z)}.$$

The first estimate of (3.24) is a consequence of (3.26) by setting

$$g_{ns}(z) = -\rho_{ns}^{-1} e^{-\frac{1}{2}\frac{\ell\beta}{\alpha}z} g(z, \rho_{ns}), \quad s = 1, 2,$$

in (3.26). Regarding the second one, it can be proved as follows:

$$\xi_{ns} = \frac{g_{ns}(0) - k_P g_{ns}(1)}{k_I} = \frac{2\tilde{k}_P - 2k_P e^{\frac{1}{2}\frac{\ell\beta}{\alpha}}}{k_I} + \mathcal{O}(n^{-1}) = \mathcal{O}(n^{-1}),$$

where we have used the fact that

$$e^{\rho_{ns}} + e^{-\rho_{ns}} - 2\tilde{k}_P = \mathcal{O}(n^{-1})$$

and the definition of  $\tilde{k}_P$  given in (3.8). Finally, in order to prove (3.25), we use (3.18) to obtain that  $\|e^{\frac{1}{2}\frac{\beta}{\alpha}x} e^{\rho_{ns}(1-\frac{x}{\ell})}\|_{L^2(0,1)}^2$  and  $\|e^{\frac{1}{2}\frac{\beta}{\alpha}x} e^{-\rho_{ns}(1-\frac{x}{\ell})}\|_{L^2(0,1)}^2$  are uniformly bound in  $(0, 1)$ . This, together with  $|\xi_{ns}|^2 = \mathcal{O}(n^{-2})$ , gives

$$\int_0^\ell |f_{ns}(x)|^2 dx = \ell \int_0^1 |g_{ns}(z)|^2 dz$$

and the desired inequalities (3.25)  $\square$

Now, one can notice that the same process can also produce asymptotic expansions for the eigenpairs of  $\mathcal{A}_{PI}^*$ , which is the adjoint operator of  $\mathcal{A}_{PI}$  and is given by

$$(3.27) \quad \mathcal{A}_{PI}^* \begin{pmatrix} h \\ \eta \end{pmatrix} := \begin{pmatrix} \alpha h'' + \beta h' \\ \alpha k_I h'(0) \end{pmatrix} \vee \begin{pmatrix} h \\ \eta \end{pmatrix} \in \mathcal{D}(\mathcal{A}_{PI}^*)$$

and

$$(3.28) \quad \mathcal{D}(\mathcal{A}_{PI}^*) := \{(z, \eta) \in H^2(0, \ell) \times \mathbb{C} : h(0) = 0, \eta = \alpha h'(\ell) + \beta h(\ell) - \alpha k_P h'(0)\}.$$

Since  $\mathcal{A}_{PI}$  is a discrete operator, then so is  $\mathcal{A}_{PI}^*$  (see [10, p. 2354]). Moreover, if  $\lambda$  is an eigenvalue of  $\mathcal{A}_{PI}$ , then  $\bar{\lambda}$  is an eigenvalue of  $\mathcal{A}_{PI}^*$  (see [20, p. 26]). Hence, we can get the eigenvalues of  $\mathcal{A}_{PI}^*$  directly from (3.10) and (3.19) for  $|\tilde{k}_P| > 1$  and  $|\tilde{k}_P| < 1$ , respectively, with the same algebraic multiplicity (see [10, p. 2354]). Also, the same arguments used in the proof in Theorem 3.4 will yield the counterpart of Theorem 3.4 for  $\mathcal{A}_{PI}^*$ , namely, the following.

**THEOREM 3.5.** *Let  $\sigma(\mathcal{A}_{PI}^*) = \{\bar{\lambda}_{n1}, \bar{\lambda}_{n2}, n \in \mathbb{N}\}$  be the eigenvalues of  $\mathcal{A}_{PI}^*$ . If  $|\tilde{k}_P| > 1$  (respectively,  $|\tilde{k}_P| < 1$ ),  $\lambda_n = \frac{\alpha \rho_n^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , with  $\rho_n \in \mathbb{C}$  and  $\arg \rho_n \in [-\pi/2, \pi/2)$ , are given by (3.10) and (3.18) (respectively, (3.19) and (3.23)). Furthermore, the corresponding eigenfunctions  $\{(h_{n1}, \eta_{n1}), (h_{n2}, \eta_{n2})\}$  have the following asymptotics: for  $s = 1, 2$ ,*

$$(3.29) \quad \begin{cases} h_{ns}(x) = \phi_{ns}(z) = e^{-\frac{1}{2}\frac{\ell\beta}{\alpha}z} e^{\overline{\rho_{ns}}z} - e^{-\frac{1}{2}\frac{\ell\beta}{\alpha}z} e^{-\overline{\rho_{ns}}z} = e^{-\frac{1}{2}\frac{\beta}{\alpha}x} e^{\overline{\rho_{ns}}\frac{x}{\ell}} - e^{-\frac{1}{2}\frac{\beta}{\alpha}x} e^{-\overline{\rho_{ns}}\frac{x}{\ell}}, \\ \eta_{ns} = \mathcal{O}(n^{-1}) \end{cases}$$

for sufficiently large positive integer  $n$ , where

$$(3.30) \quad x = z\ell, \quad h(x) = \phi(z), \quad z \in [0, 1].$$

Moreover,  $\{(h_{ns}, \eta_{ns}), s = 1, 2\}$  is approximately normalized in  $\mathcal{H}$ .

*Proof.* From (3.27) and (3.28), the eigenvalue problem of  $\mathcal{A}_{PI}^*$  is

$$(3.31) \quad \begin{cases} h''(x) + \frac{\beta}{\alpha}h'(x) - \frac{1}{\alpha}\lambda h(x) = 0, & 0 < x < \ell, \\ h(0) = 0, \quad \alpha(\lambda k_P + k_I)h'(0) = \lambda\alpha h'(\ell) + \lambda\beta h(\ell). \end{cases}$$

By the transformation (3.30), which is similar to (3.2), the above equation changes into

$$(3.32) \quad \begin{cases} \phi''(z) + \frac{\ell\beta}{\alpha}\phi'(z) - \frac{\ell^2}{\alpha}\lambda\phi(z) = 0, & 0 < z < 1, \\ \phi(0) = 0, \quad \alpha(\lambda k_P + k_I)\phi'(0) = \lambda\alpha\phi'(1) + \lambda\beta\ell\phi(1). \end{cases}$$

Then  $e^{-\tau_1 z}$  and  $e^{-\tau_2 z}$  are two independent solutions of  $\phi''(z) + \frac{\ell\beta}{\alpha}\phi'(z) - \frac{\ell^2}{\alpha}\lambda\phi(z) = 0$ . As in the proof of Theorem 3.4, the function  $\phi$  with respect to the eigenvalue  $\bar{\lambda}$  of  $\mathcal{A}_{PI}^*$ , where  $\lambda = \frac{\alpha \rho^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , with  $\rho \in \mathbb{C}$  and  $\arg \rho \in [-\pi/2, \pi/2)$ , is given by

$$(3.33) \quad \phi(z, \rho) = \begin{vmatrix} 1 & 1 \\ e^{-\tau_1 z} & e^{-\tau_2 z} \end{vmatrix} = e^{-\tau_2 z} - e^{-\tau_1 z} = e^{-\frac{1}{2}\frac{\ell\beta}{\alpha}z} e^{\rho z} - e^{-\frac{1}{2}\frac{\ell\beta}{\alpha}z} e^{-\rho z}.$$

The first expression of (3.29) then follows from (3.33) by setting

$$\phi_{ns}(z) = \phi(z, \overline{\rho_{ns}}), \quad s = 1, 2,$$

in (3.33), and  $\eta_{ns}$  can be obtained by

$$\eta_{ns} = \alpha h'_{ns}(\ell) + \beta h_{ns}(\ell) - \alpha k_P h'_{ns}(0) = \frac{\alpha}{\ell} \phi'_{ns}(1) + \beta \phi_{ns}(1) - \frac{\alpha}{\ell} k_P \phi'_{ns}(0) = \mathcal{O}(n^{-1}).$$

Here we have used the fact that

$$e^{\overline{\rho_{ns}}} + e^{-\overline{\rho_{ns}}} - 2\tilde{k}_P = \mathcal{O}(n^{-1})$$

and the definition of  $\tilde{k}_P$  in (3.8). A direct computation can further show that  $\{(h_{ns}, \eta_{ns}), s = 1, 2\}$  is approximately normalized.  $\square$

**3.2. Completeness of the root subspace.** First, we have the following.

**THEOREM 3.6.** *Let  $\sigma(\mathcal{A}_{PI}) = \{\lambda_{n1}, \lambda_{n2}, n \in \mathbb{N}\}$  be the eigenvalues of  $\mathcal{A}_{PI}$ , and let  $\lambda_n = \frac{\alpha \rho_n^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , with  $\rho_n \in \mathbb{C}$  and  $\arg \rho_n \in [-\pi/2, \pi/2)$ . Then there exists a constant  $\widetilde{M} > 0$  independent of  $\lambda$  such that*

$$(3.34) \quad \|R(\lambda, \mathcal{A}_{PI})\| \leq \widetilde{M} |\lambda|^{-1/2}$$

for all  $\lambda = \frac{\alpha \rho^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , with  $\rho \in \mathbb{C}$ , and  $\arg \rho \in [-\pi/2, \pi/2)$  lies outside all circles of radius  $\varepsilon > 0$  and the circles are centered at the zeros of  $\det(\Delta(\rho)) = 0$  (see (3.9)).

*Proof.* Let  $\lambda = \alpha \rho^2 / \ell^2 - \frac{1}{4} \beta^2 / \alpha \in \rho(\mathcal{A}_{PI})$ , with  $\rho \in \mathbb{C}$  and  $\arg \rho \in [-\pi/2, \pi/2)$ , and let  $(\phi, c) \in \mathcal{H}$ . Our aim is to solve the resolvent equation

$$(\lambda I - \mathcal{A}_{PI}) \begin{pmatrix} f \\ \xi \end{pmatrix} = \begin{pmatrix} \phi \\ c \end{pmatrix},$$

or, equivalently,

$$\begin{cases} \lambda f(x) - \alpha f''(x) + \beta f'(x) = \phi(x), \\ \lambda \frac{f(0) - k_P f(\ell)}{k_I} - f(\ell) = c, \end{cases}$$

or

$$(3.35) \quad \begin{cases} f''(x) - \frac{\beta}{\alpha} f'(x) - \frac{1}{\alpha} \lambda f(x) = -\frac{1}{\alpha} \phi(x), & 0 < x < \ell, \\ \lambda f(0) - \lambda k_P f(\ell) - k_I f(\ell) = k_I c, & f'(\ell) = 0. \end{cases}$$

By using the transformation (3.2), we obtain

$$(3.36) \quad \begin{cases} g''(z) - \frac{\ell \beta}{\alpha} g'(z) - \frac{\ell^2}{\alpha} \lambda g(z) = -\frac{\ell^2}{\alpha} \phi(z\ell), & 0 < z < 1, \\ \lambda g(0) - (\lambda k_P + k_I) g(1) = k_I c, & g'(1) = 0. \end{cases}$$

Set

$$(3.37) \quad \Phi(z, \lambda) := g(z) - \frac{k_I c}{\lambda(1 - k_P) - k_I}, \quad \lambda(1 - k_P) \neq k_I.$$

Then,  $\Phi(z, \lambda)$  satisfies

$$(3.38) \quad \begin{cases} \Phi''(z, \lambda) - \frac{\ell\beta}{\alpha}\Phi'(z, \lambda) - \frac{\ell^2}{\alpha}\lambda\Phi(z, \lambda) = -\frac{\ell^2}{\alpha}\phi(z\ell) + \frac{\ell^2}{\alpha}\frac{\lambda k_{Ic}}{\lambda(1-k_P) - k_I}, \\ \lambda\Phi(0, \lambda) - (\lambda k_P + k_I)\Phi(1, \lambda) = 0, \quad \Phi'(1, \lambda) = 0. \end{cases}$$

Therefore, every solution  $\Phi(z, \lambda)$  of (3.38) can be represented as (see, e.g., [32, p. 31, Theorem 2])

$$(3.39) \quad \Phi(z, \lambda) = \int_0^1 G(z, \xi, \lambda) \left( -\frac{\ell^2}{\alpha}\phi(\xi\ell) + \frac{\ell^2}{\alpha}\frac{\lambda k_{Ic}}{\lambda(1-k_P) - k_I} \right) d\xi,$$

and hence the solution of (3.36) can be written as follows:

$$(3.40) \quad g(z) = \int_0^1 G(z, \xi, \lambda) \left( -\frac{\ell^2}{\alpha}\phi(\xi\ell) + \frac{\ell^2}{\alpha}\frac{\lambda k_{Ic}}{\lambda(1-k_P) - k_I} \right) d\xi + \frac{k_{Ic}}{\lambda(1-k_P) - k_I}.$$

Here  $G(x, \xi, \lambda)$  is the Green's function given by

$$G(z, \xi, \lambda) := \frac{1}{\det(\Delta(\lambda))} H(z, \xi, \lambda),$$

with

$$(3.41) \quad H(z, \xi, \lambda) := \begin{vmatrix} e^{\tau_1 z} & e^{\tau_2 z} & \eta(z, \xi, \lambda) \\ U_1(e^{\tau_1 z}) & U_1(e^{\tau_2 z}) & U_2(e^{\tau_1 z}) \\ U_2(e^{\tau_2 z}) & U_2(y_2) & U_2(\eta) \end{vmatrix},$$

$$(3.42) \quad \eta(z, \xi, \lambda) := \frac{1}{4\rho} \text{sign}(z - \xi) \left( e^{\tau_1(z-\xi)} - e^{\tau_2(z-\xi)} \right),$$

where we have used the fact that  $\lambda = \frac{\alpha\rho^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ ,  $U_1(g) := (\lambda k_P + k_I)g(1) - \lambda g(0)$ , and  $U_2(g) := g'(1)$ . Based on the above equations, we can claim that for  $\lambda \in \rho(\mathcal{A}_{PI})$ , with  $|\lambda|$  large enough, there exists a constant  $M$  independent of  $z, \xi \in [0, 1]$  so that

$$(3.43) \quad |H(z, \xi, \lambda)| \leq M e^{\frac{\ell\beta}{2\alpha}} |\rho|^2 e^{|\rho|}, \quad \rho \in \mathbb{C}, \quad \arg \rho \in [-\pi/2, \pi/2].$$

Thus we conclude from (3.9) that

$$(3.44) \quad |G(z, \xi, \lambda)| \leq M_1 |\rho|^{-1}$$

holds for all  $\rho \in \mathbb{C}$ , with  $\arg \rho \in [-\pi/2, \pi/2]$  outside those circles of radius  $\varepsilon > 0$  and centered at the zeros of  $\det(\Delta(\rho)) = 0$ , where  $M_1$  is some constant independent of  $z, \xi \in [0, 1]$ . This will, in turn, yield estimates for  $g(z)$ ,  $f(x)$ , and  $\|R(\lambda, \mathcal{A}_{PI})\|$ , respectively, as follows:

$$(3.45) \quad |f(x)| = |g(z)| \leq M_2 |\rho|^{-1}$$

and

$$(3.46) \quad \|R(\lambda, \mathcal{A}_{PI})\| \leq M_3 |\rho|^{-1}$$

for all  $\rho \in \mathbb{C}$ , with  $\arg \rho \in [-\pi/2, \pi)$  outside those circles of radius  $\varepsilon > 0$  and centered at the zeros of  $\det(\Delta(\rho)) = 0$ , where  $M_2$  and  $M_3$  are some constants independent of  $z, \xi \in [0, 1]$ . This achieves the proof of the desired result by taking  $\widetilde{M} = M_3$ .  $\square$

Note that we can obtain a more precise estimate for  $\|R(\lambda, \mathcal{A}_{PI})\|$  as follows (for more details, the reader is referred to [14]).

**THEOREM 3.7.** *There exist positive constants  $r$  and  $M_r$  such that for an arbitrary  $\kappa \in (\pi/2, \pi)$ , we have*

$$\Sigma_{\kappa, r} := \{\lambda \in \mathbb{C} : |\arg(\lambda - r)| < \kappa, \lambda \neq 0\} \subset \rho(\mathcal{A}_{PI})$$

and

$$(3.47) \quad \|R(\lambda, \mathcal{A}_{PI})\| \leq \frac{M_r}{|\lambda - r|} \quad \forall \lambda \in \Sigma_{\kappa, r}.$$

We have the following lemma.

**LEMMA 3.8.** *Let  $\sigma(\mathcal{A}_{PI}) = \{\lambda_{n1}, \lambda_{n2}, n \in \mathbb{N}\}$  be the eigenvalues of  $\mathcal{A}_{PI}$ . If  $|\widetilde{k}_P| > 1$  (respectively,  $0 < |\widetilde{k}_P| < 1$ ), then  $\lambda_n = \frac{\alpha \rho_n^2}{\ell^2} - \frac{\beta^2}{4\alpha}$ , with  $\rho_n \in \mathbb{C}$  and  $\arg \rho_n \in [-\pi/2, \pi/2)$ , are given by (3.10) and (3.18) (respectively, (3.19) and (3.23)). In addition, for any sufficiently large  $n$ , each eigenvalue  $\lambda_{ni}$ ,  $i = 1, 2$ , of  $\mathcal{A}_{PI}$  is algebraically simple.*

*Proof.* From (3.40), the multiplicity of each  $\lambda \in \sigma(\mathcal{A}_{PI})$  with sufficiently large modulus, as a pole of  $R(\lambda, \mathcal{A}_{PI})$ , is less than or equal to the multiplicity of  $\lambda$  as a zero of the entire function  $\det(\Delta(\rho))$  with respect to  $\rho$ . On the other hand, it is a routine exercise to verify that  $\lambda$  is geometrically simple. Since from (3.14), under the assumption  $|\widetilde{k}_P| \neq 1$ , all zeros of  $\det(\Delta(\rho)) = 0$  with large moduli are simple, the result then follows from the general formula  $m_a \leq p \cdot m_g$  (see, e.g., [30, p. 148]), where  $p$  denotes the order of the pole of the resolvent operator and  $m_a, m_g$  denote, respectively, the algebraic and geometric multiplicity.  $\square$

The next result follows.

**PROPOSITION 3.9.** *Both root subspaces of the operators  $\mathcal{A}_{PI}$  and  $\mathcal{A}_{PI}^*$  are complete in  $\mathcal{H}$ , that is to say,  $\text{Sp}(\mathcal{A}_{PI}) = \text{Sp}(\mathcal{A}_{PI}^*) = \mathcal{H}$ .*

*Proof.* It follows from Lemma 5 on page 2355 of [10] that the following orthogonal decomposition holds:

$$\mathcal{H} = \sigma_\infty(\mathcal{A}_{PI}^*) \oplus \text{Sp}(\mathcal{A}_{PI}),$$

where  $\sigma_\infty(\mathcal{A}_{PI}^*)$  consists of those  $Y \in \mathcal{H}$  so that  $R(\lambda, \mathcal{A}_{PI}^*)Y$  is an analytic function of  $\lambda$  in the whole complex plane. Hence  $\text{Sp}(\mathcal{A}_{PI}) = \mathcal{H}$  if and only if  $\sigma_\infty(\mathcal{A}_{PI}^*) = \{0\}$ . Now suppose that  $Y \in \sigma_\infty(\mathcal{A}_{PI}^*)$ . Since  $R(\lambda, \mathcal{A}_{PI}^*)Y$  is an analytic function of  $\lambda$ , it is also for  $\rho$ , and hence by the maximum modulus principle (or the Phragmén-Lindelöf theorem) of analytic functions and the fact that  $\|R(\lambda, \mathcal{A}_{PI}^*)\| = \|R(\bar{\lambda}, \mathcal{A}_{PI})\|$ , it follows from Theorem 3.6 that

$$\|R(\lambda, \mathcal{A}_{PI}^*)Y\| \leq M|\lambda|^{-1/2}\|Y\| \quad \forall \lambda \in \mathbb{C}$$

for some constant  $M > 0$ . By Theorem 1 on page 3 of [17], we conclude that  $R(\lambda, \mathcal{A}_{PI}^*)Y$  is a constant with respect to  $\lambda$ , i.e.,

$$R(\lambda, \mathcal{A}_{PI}^*)Y = Y_0 \quad \text{for some } Y_0 \in \mathcal{H}.$$

Thus

$$Y = (\lambda - \mathcal{A}_{PI}^*)Y_0 = -\mathcal{A}_{PI}^*Y_0 + \lambda Y_0 \quad \forall \lambda \in \mathbb{C}.$$

Finally, comparing the coefficients of  $\lambda^j$ , one can readily find that  $Y_0 = 0$ . This concludes the proof of the result.  $\square$

**3.3. Riesz basis generation.** In order to establish the Riesz basis property of the root subspace of  $\mathcal{A}_{PI}$ , we need the following lemma from [38].

LEMMA 3.10. *Suppose that a sequence  $\{\nu_n\}$  has asymptotics*

$$(3.48) \quad \nu_n = \alpha(n + i\beta \ln n) + \mathcal{O}(1), \quad \alpha \neq 0, \quad n = 1, 2, 3, \dots,$$

where  $\beta$  is a real number and  $\sup_{n \geq 1} \operatorname{Re} \nu_n < \infty$ . Then the sequence  $\{e^{\nu_n x}\}_{n=1}^\infty$  is a Bessel sequence in  $L^2(0, 1)$ .

Then, we have the following.

LEMMA 3.11. *Let  $\rho_{ns}$ ,  $s = 1, 2$ , be given by (3.18) and (3.23) according to  $|\tilde{k}_P| > 1$  and  $|\tilde{k}_P| < 1$ , respectively. Then  $\{e^{\rho_{n1}z}, e^{\rho_{n2}z}\}_{n=1}^\infty$  and  $\{e^{-\rho_{n1}z}, e^{-\rho_{n2}z}\}_{n=1}^\infty$  are two Bessel sequences in  $L^2(0, 1)$ . Hence,  $\{e^{\rho_{n1}x/\ell}, e^{\rho_{n2}x/\ell}\}_{n=1}^\infty$  and  $\{e^{-\rho_{n1}x/\ell}, e^{-\rho_{n2}x/\ell}\}_{n=1}^\infty$  are also two Bessel sequences in  $L^2(0, \ell)$ .*

*Proof.* Let  $\nu_{ns} := \rho_{ns}$ ,  $s = 1, 2$ . Then the result concerning the sequence  $\{e^{\rho_{n1}z}, e^{\rho_{n2}z}\}_{n=1}^\infty$  can be directly obtained from Lemma 3.10 by setting  $\beta = 0$  and  $\alpha = 2\pi i$  in (3.48). Similarly, for the sequence  $\{e^{-\rho_{n1}z}, e^{-\rho_{n2}z}\}_{n=1}^\infty$ , one can take  $\nu_n := -\rho_{ns}$ ,  $s = 1, 2$ ,  $\beta = 0$ , and  $\alpha = -2\pi i$  in (3.48). The last result is obvious since it can be directly verified from the definition of Bessel sequences (see [42, p. 122]).  $\square$

Now, we are able to prove the Riesz basis property for the operator  $\mathcal{A}_{PI}$  stated in Theorem 2.4.

*Proof of Theorem 2.4.* Let  $|\tilde{k}_P| \neq 1$  and let  $\sigma(\mathcal{A}_{PI}) = \{\lambda_{n1}, \lambda_{n2}, n \in \mathbb{N}\}$  be the eigenvalues of  $\mathcal{A}_{PI}$ . From Lemma 3.8, we have that each eigenvalue of  $\mathcal{A}_{PI}$  with sufficient large modulus is simple, and hence there exists an integer  $N > 0$  so that for all  $n > N$ ,  $\lambda_{ns}$ ,  $s = 1, 2$ , is simple. For  $n \leq N$ , if the algebraic multiplicity of each  $\lambda_{ns}$  is  $m_{ns}$ , we can find the highest order generalized eigenfunction  $\Phi_{n,s,1}$  from

$$(\mathcal{A}_{PI} - \lambda_{ns})^{m_{ns}} \Phi_{n,s,1} = 0 \quad \text{but} \quad (\mathcal{A}_{PI} - \lambda_{ns})^{m_{ns}-1} \Phi_{n,s,1} \neq 0, \quad s = 1, 2.$$

The other lower order linearly independent generalized eigenfunctions associated with  $\lambda_{ns}$  can be found through  $\Phi_{n,s,j} = (\mathcal{A}_{PI} - \lambda_{ns})^{j-1} \Phi_{n,s,1}$ ,  $j = 2, 3, \dots, m_{ns}$ . Assume  $\Phi_{ns}$  is an eigenfunction of  $\mathcal{A}_{PI}$  corresponding to  $\lambda_{ns}$  with  $n > N$ . Then  $\{\{\{\Phi_{n,s,j}\}_{j=1}^{m_{ns}}\}_{n \leq N} \cup \{\Phi_{ns}\}_{n > N}\}_{s=1}^2$  are all linearly independent generalized eigenfunctions of  $\mathcal{A}_{PI}$ . Let  $\{\{\{\Psi_{n,s,j}\}_{j=1}^{m_{ns}}\}_{n \leq N} \cup \{\Psi_{ns}\}_{n > N}\}$  be the biorthogonal sequence of  $\{\{\{\Phi_{n,s,j}\}_{j=1}^{m_{ns}}\}_{n \leq N} \cup \{\Phi_{ns}\}_{n > N}\}$ . Then  $\{\{\{\Psi_{n,s,j}\}_{j=1}^{m_{ns}}\}_{n \leq N} \cup \{\Psi_{ns}\}_{n > N}\}_{s=1}^2$  are all linearly independent generalized eigenfunctions of  $\mathcal{A}_{PI}^*$ . It is well known that these two sequences are minimal in  $\mathcal{H}$ , and from Proposition 3.9 they are also complete in  $\mathcal{H}$ .

Hence, in order to prove the Riesz basis of the system, it suffices to show that both eigenfunctions  $\{\Phi_{ns}\}_{n > N, s=1,2}$  and  $\{\Psi_{ns}\}_{n > N, s=1,2}$  of the operators  $\mathcal{A}_{PI}$  and  $\mathcal{A}_{PI}^*$  are, respectively, Bessel sequences in  $\mathcal{H}$ . Since  $1 \leq \|\Phi_{ns}\| \|\Psi_{ns}\| \leq M$  for some constant  $M$  independent of  $n$  (see [42, p. 19]), we may assume without loss of generality that  $\Phi_{ns} = (f_{ns}, \xi_{ns})$  given by (3.24) and  $\Psi_{ns} = (h_{ns}, \eta_{ns})$  given by (3.29) for all  $n > N$ . Then it follows from Lemma 3.11 and the expansions (3.24) and (3.29) that both sequences  $\{f_{n1}, f_{n2}\}_{n > N}$  and  $\{h_{n1}, h_{n2}\}_{n > N}$  are Bessel sequences in  $L^2(0, \ell)$  and both  $\{\xi_{n1}, \xi_{n2}\}_{n > N}$  and  $\{\eta_{n1}, \eta_{n2}\}_{n > N}$  are Bessel sequences in  $\mathbb{C}$ . Therefore both of  $\{\Phi_{ns}\}_{n > N, s=1,2}$  and  $\{\Psi_{ns}\}_{n > N, s=1,2}$  are also Bessel sequences in  $\mathcal{H}$ , and the result follows.



Now, let  $|\tilde{k}_P| = 1$ . From (3.13), we already know that the eigenvalues  $\lambda_{n1}$  and  $\lambda_{n2}$  are not separable when their moduli are large enough, and hence the system may have the same eigenvalues depending on the feedback constants  $k_P$  and  $k_I$ . For such a case, the generalized eigenfunctions of the operator  $\mathcal{A}_{PI}$  form, in  $\mathcal{H}$ , a Riesz basis with parentheses, and each eigenspace, corresponding to eigenvalues with enough large modulus, has dimension two [38]. We omit the details here and leave this as an exercise for the reader.  $\square$

Finally, we reach the proof of the second main result of this work.

*Proof of Theorem 2.5.* We shall assume that  $|\tilde{k}_P| \neq 1$  (the case  $|\tilde{k}_P| = 1$  can be treated similarly). First, the existence of the  $C_0$ -semigroup  $S_{PI}(t)$  follows from Lemma 3.8 and Theorem 2.4. Indeed, since  $\{\{\Phi_{n,s,j}\}_{j=1}^{m_n}\}_{n \leq N, s=1,2} \cup \{\Phi_{ns}\}_{n > N, s=1,2}$  forms a Riesz basis for  $\mathcal{H}$ , then any  $Y \in \mathcal{H}$  can be expanded as follows:

$$Y = \sum_{s=1}^2 \sum_{n=1}^N \sum_{j=1}^{m_n} a_{nsj} \Phi_{n,s,j} + \sum_{s=1}^2 \sum_{n=N+1}^{\infty} a_{ns} \Phi_{ns},$$

where  $a_{nsj}$  and  $a_{ns}$  are constants. Moreover, for such  $Y$ , the semigroup  $S_{PI}(t)$  satisfies

$$(3.49) \quad S_{PI}(t)Y = \sum_{s=1}^2 \sum_{n=1}^N e^{\lambda_{ns}t} \sum_{j=1}^{m_n} a_{nsj} \sum_{i=0}^{m_{ns}-j} \frac{t^i}{i!} \Phi_{n,s,i} + \sum_{s=1}^2 \sum_{n=N+1}^{\infty} e^{\lambda_{ns}t} a_{ns} \Phi_{ns}.$$

Then, using the estimate (3.47) of  $\|R(\lambda, \mathcal{A}_{PI})\|$  on  $\Sigma_{\kappa,r}$ , one can deduce the analytic property of the semigroup (see [33]). Finally, the spectrum-determined growth condition is a direct consequence of the Riesz basis property (also from the analyticity of the semigroup).  $\square$

**4. Proof of Theorem 2.6.** It is well known that although the spectrum of the uncontrolled operator  $A_0$  lies in the open left half-plane (see Lemma 2.5), the operator  $\mathcal{A}_{PI}$  (see (2.10)–(2.11)) of the closed-loop system may have eigenvalues in the right half-complex-plane, and hence the semigroup  $S_{PI}(t)$  will be unstable if the proportional and/or integral gains  $k_P$  and/or  $k_I$  are not properly chosen. We thus need to propose a design method for the proportional and integral gains  $k_P$  and  $k_I$  so that the closed-loop system (2.9) will be exponentially stable.

Using Theorem 4.3 of [33], one can claim that in order to get the exponential stability of the analytic semigroup  $S_{PI}(t)$ , it suffices to show that all the eigenvalues of the operator  $\mathcal{A}_{PI}$  defined by (2.10)–(2.11) have negative real part. Unfortunately, this property turns out to be very difficult to establish in our case since (i) it is not obvious to prove the dissipativity of the operator  $\mathcal{A}_{PI}$  even if under conditions on the proportional and integral gains  $k_P$  and  $k_I$ , and (ii) the problem of obtaining an explicit expression of the eigenvalues of  $\mathcal{A}_{PI}$  is equivalent to solving an unusual transcendental equation. Therefore, it is not easy to get an explicit condition on the proportional and integral gains  $k_P$  and  $k_I$ , which involves the system parameters  $\alpha$ ,  $\beta$ , and  $\ell$  so that the eigenvalues lie in the open left half-plane. In return, we are able to give some implicit conditions on  $k_P$  and  $k_I$  to resolve this hard problem.

First, we have the following result.

**PROPOSITION 4.1.** *If the proportional and integral gains  $k_P$  and  $k_I$  are negative, then any real eigenvalue of the operator  $\mathcal{A}_{PI}$ , defined by (2.10)–(2.11), must be necessarily negative.*

*Proof.* Let  $\lambda$  be an eigenvalue of  $\mathcal{A}_{PI}$  and  $\phi = (f, \xi)$  an associated eigenfunction. Then its eigenvalue problem (3.1) has two characteristic roots given by

$$\eta_{1,2} = \frac{\beta \pm \sqrt{\Delta_\lambda}}{2\alpha}, \quad \text{where } \Delta_\lambda = \beta^2 + 4\alpha\lambda.$$

Consider the following two cases.

*Case (i).*  $\eta_1 = \eta_2$ . This implies that  $\Delta_\lambda = 0$ , i.e.  $\lambda = -\frac{\beta^2}{4\alpha}$ . Next, one can readily prove that  $\lambda = -\frac{\beta^2}{4\alpha}$  is an eigenvalue if and only if

$$\frac{\beta^2}{4\alpha} k_P - k_I = \frac{\beta^2}{4\alpha} \left( 1 + \frac{\beta\ell}{2\alpha} \right) e^{-\frac{\beta\ell}{2\alpha}}.$$

Hence for appropriate negative proportional and integral gains,  $\lambda = -\frac{\beta^2}{4\alpha} < 0$  is an eigenvalue.

*Case (ii).*  $\eta_1 \neq \eta_2$ . In this case, let us consider the normalized solution, of (3.1), at  $x = \ell$ , by  $f(\ell) = 1$ . This, together with the boundary condition  $f'(\ell) = 0$ , implies that the solution of (3.1) is

$$(4.1) \quad f(x) = \frac{e^{-\frac{\beta\ell}{\alpha}}}{2\sqrt{\Delta_\lambda}} \left[ \left( \beta + \sqrt{\Delta_\lambda} \right) e^{\eta_1 L + \eta_2 x} - \left( \beta - \sqrt{\Delta_\lambda} \right) e^{\eta_2 L + \eta_1 x} \right].$$

Now let  $k_P, k_I < 0$  and assume that  $\lambda = \gamma^2$ , where  $\gamma \in \mathbb{R} \setminus \{0\}$ , is an eigenvalue of  $\mathcal{A}_{PI}$ . This is equivalent to claiming that the other boundary condition  $(\lambda k_P + k_I) = \lambda f(0)$  holds. In return, using (4.1) and the fact that  $f(\ell) = 1$ , it follows after a simple calculation that the following equation holds:

$$\left( \beta + \sqrt{\Delta(\gamma)} \right) e^{\frac{\sqrt{\Delta(\gamma)}\ell}{2\alpha}} - \left( \beta - \sqrt{\Delta(\gamma)} \right) e^{-\frac{\sqrt{\Delta(\gamma)}\ell}{2\alpha}} = 2e^{\frac{\beta\ell}{\alpha}} \frac{\sqrt{\Delta(\gamma)}}{\gamma^2} (k_P \gamma^2 + k_I),$$

where  $\Delta(\gamma) = \beta^2 + 4\alpha\gamma^2$ . Finally, since  $\sqrt{\Delta(\gamma)} > \beta$  and  $k_P, k_I < 0$ , the above equation leads to a contradiction.  $\square$

Let us rewrite the spectral system (3.1) in the following form:

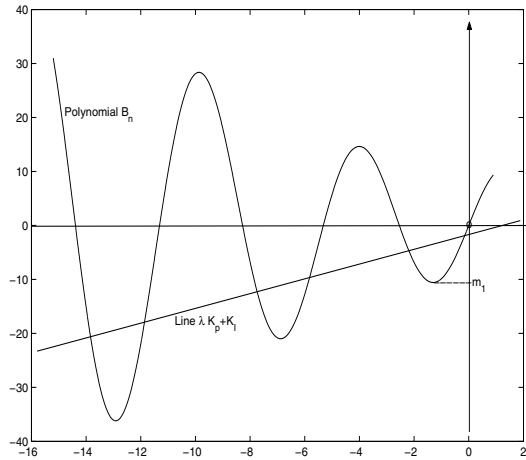
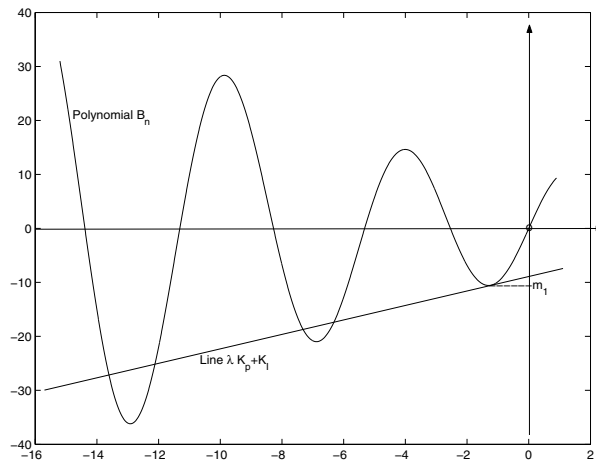
$$(4.2) \quad \begin{cases} \alpha f''(x, \lambda) - \beta f'(x, \lambda) = \lambda f(x, \lambda), & 0 \leq x \leq \ell, \\ \lambda f(0, \lambda) = (\lambda k_P + k_I) f(\ell, \lambda), & f'(\ell, \lambda) = 0. \end{cases}$$

Then, by considering the normalized solution at  $x = \ell$  via  $f(\ell, \lambda) = 1$ , it follows that all eigenvalues of (4.2) are the roots of the left-end boundary condition  $\lambda f(0, \lambda) = \lambda k_P + k_I$ . Since  $f(\cdot, \lambda)$  satisfies the same equation, as in (2.3) and (2.6), it necessarily has the same representation given by (2.7), and hence the condition  $\lambda f(0, \lambda) = \lambda k_P + k_I$  yields

$$(4.3) \quad B(\lambda) := \lambda \prod_{i=1}^{\infty} \left( 1 - \frac{\lambda}{\lambda_i^0} \right) = \lambda k_P + k_I.$$

Recall that we have to exclude the zero solution  $\lambda = 0$  of the above equation as  $0 \in \rho(\mathcal{A}_{PI})$ , the resolvent set of  $\mathcal{A}_{PI}$ . We are now able to prove the third main result stated as Theorem 2.6.

*Proof of Theorem 2.6.* (i) First, one can claim from the definition of  $B(\lambda)$  that  $B(\lambda) > 0$  whenever  $\lambda$  is a positive real number, and hence any real root of  $B(\lambda)$  should

FIG. 2.  $B_n(\lambda) = \lambda k_P + k_I$  has  $n + 1$  negative distinct real roots.FIG. 3.  $n - 1$  negative distinct real roots and one double negative real roots to  $B_n(\lambda) = \lambda k_P + k_I$ .

be negative. Second, since both  $\lambda_i^0$  and  $k_I$  are negative and  $k_P$  is positive, it follows that  $B(\lambda_i^0) \neq k_P \lambda_i^0 + k_I$ , and thus for  $n$  large enough, the zeros of  $B(\cdot) - ((\cdot)k_P + k_I)$  and  $B_n(\cdot) - ((\cdot)k_P + k_I)$  coincide. in the finite strip  $\lambda_i^0 \leq \operatorname{Re}(z) \leq 0$  and  $|\operatorname{Im}(z)| \leq \varepsilon$  which contains a piece of the negative real axis.

Now if  $m_1 < k_I < 0$  and  $k_P$  is a positive number sufficiently small (see Figure 2), then the  $n + 1$  roots of  $B_n(\lambda) = \lambda k_P + k_I$  are all real and negative. Finally, note that in this case, we may have  $n - 1$  negative distinct real roots and one double negative real root (see Figure 3). Since these properties are valid for all  $B_n$ , they also hold true for  $B$  by Lemma 2.3.

(ii) Consider now a large enough strip that would include the points where  $B_n$  attains the minima  $m_1$  and  $m_2$  (see Figure 4). Then using the graph of  $B_n$ , one can see that if  $m_2 < k_I < m_1$  and  $k_P$  is a positive number sufficiently small, the equation  $B_n(\lambda) = \lambda k_P + k_I$  has only  $n - 1$  real roots instead of  $n + 1$ . Hence two roots are missing. Since  $B_n(\lambda) = \lambda k_P + k_I$  must have  $n + 1$  roots, the two missing roots should

go conjugate complex (which is a bifurcation) with positive or negative real part. These properties still remain true for  $B$ , by Lemma 2.3.

Finally, we are going to show the output regulation and the exponential stability of the closed-loop system (1.1)–(1.2). Although the proof is similar to that of Theorem 4.1 in [2], we would rather give some details for sake of completeness.

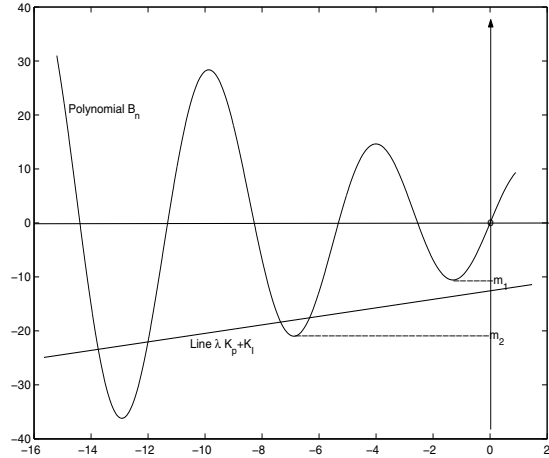


FIG. 4.  $n - 1$  negative distinct real roots and two complex conjugate roots to  $B_n(\lambda) = \lambda k_P + k_I$ .

Due to Theorem 2.5 and the exponential stability of the semigroup  $S_{PI}(t)$ , the solution  $\phi(t) = (Q(\cdot, t), \xi(t))$  of the closed-loop system (2.9), or alternatively (1.1)–(1.2), stemming from the initial data  $\phi_0 = (Q_0, \xi_0) \in \mathcal{D}(\mathcal{A}_{PI})$ , can be written as

$$\phi(t) = S_{PI}(t)\phi_0 + \int_0^t S_{PI}(t-s)(w, -y_r)ds = S_{PI}(t)\phi_0 + \mathcal{A}_{PI}^{-1}(S_{PI}(t) - I)(w, -y_r)$$

and satisfies

$$(4.4) \quad \lim_{t \rightarrow \infty} \phi(t) = \lim_{t \rightarrow \infty} (Q(\cdot, t), \xi(t)) = -\mathcal{A}_{PI}^{-1}(w, -y_r).$$

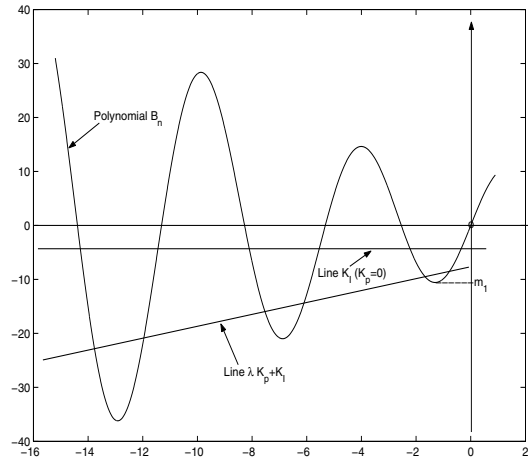
Let  $\phi^* := (Q^*, y_r^*) = -\mathcal{A}_{PI}^{-1}(w, -y_r)$  and hence  $\mathcal{A}_{PI}(Q^*, y_r^*) = (-w, y_r)$ . Using (2.10)–(2.11), it follows that  $Q^*(\ell) = y_r$ . This, together with (4.4), implies that  $\lim_{t \rightarrow \infty} y(t) = \lim_{t \rightarrow \infty} Q(\ell, t) = y_r$ . Concerning the stability of the closed-loop system (2.9), we first note that the steady state is  $\phi^* = \mathcal{A}_{PI}^{-1}(-w, y_r)$ . Then, setting  $\tilde{\phi} := \phi - \phi^* = \phi - \mathcal{A}_{PI}^{-1}(-w, y_r)$ , the closed-loop system (2.9) can be written as

$$(4.5) \quad \dot{\tilde{\phi}}(t) = \mathcal{A}_{PI}(\tilde{\phi}(t) + \phi^*) + (w, -y_r) = \mathcal{A}_{PI}\tilde{\phi}(t) + \mathcal{A}_{PI}\phi^* + (w, -y_r) = \mathcal{A}_{PI}\tilde{\phi}(t).$$

Consequently, the exponential stability of the closed-loop system (2.9) is equivalent to that of the semigroup  $S_{PI}(t)$ , independently of the perturbation  $w$ . The proof is complete.  $\square$

*Remark 4.1.*

1. As mentioned in the introduction, one of the advantages of introducing the proportional gain  $k_P$  in our control feedback law (contrary to the work [2] where  $k_P = 0$ ) is that this gain improves the stability and the regulation of the closed-loop system (1.1)–(1.2), where  $k_P \neq 0$ , in comparison with the

FIG. 5. Stability improvement for  $k_P > 0$ .

closed-loop system (1.1)–(1.2) with  $k_P = 0$ . To see how this goes, let us sketch the argument from the proof of Theorem 2.6. Suppose that  $k_P = 0$ . Then if  $m_1 < k_I < 0$  (the other cases can be treated similarly), the equation  $B_n(\lambda) = k_I$  has  $n + 1$  distinct negative real roots (see Figure 5), namely,  $\mu_i$ ,  $i = 1, \dots, n$ . Now, one can always choose  $k_P$  sufficiently small such that the new equation  $B_n(\lambda) = \lambda k_P + k_I$  has not only  $n + 1$  distinct negative real roots  $\eta_i$ ,  $i = 1, \dots, n$ , but also  $\max_i \{\eta_i\} < \max_i \{\mu_i\}$ , and hence the spectrum is moved to the left as shown in Figure 5.

- According to Proposition 4.1, one can claim that in order to recover all negative real eigenvalues of the operator  $\mathcal{A}_{PI}$ , it suffices to take both  $k_P$  and  $k_I$  negative and then to try adding further conditions on  $k_P$  and  $k_I$  to conclude the stability. Unfortunately, we have tried in this direction but without much success. This is due to the fact that when  $k_P$  and  $k_I$  are negative, the arguments of the proof of Theorem 2.6 fail. Indeed, one can easily check that for  $k_P, k_I < 0$  and for given  $\alpha, \beta$ , and  $\ell$ , the spectrum  $\sigma(\mathcal{A}_{PI})$  may contain many complex eigenvalues, and we cannot control their real parts by means of our approach. For instance, consider a river whose characteristics are  $\ell = 2700$ ,  $\alpha = 2000$ , and  $\beta = 0.9$  (see [31] for more details about this model). Then, let  $k_I = -0.01$  and  $k_P = -0.05$ . Using MAPLE, one can verify that  $\lambda = 0.001150641022 + i0.003238193679$  is an eigenvalue among others. This physical example shows that for  $k_P < 0$ , we may have complex eigenvalues with positive real part, and hence the closed-loop system is unstable. However, it is clear that this does not mean that the system is unstable whenever  $k_P < 0$ , but as mentioned above, this is a drawback of our approach. This is why we had to choose  $k_I < 0$  and  $k_P > 0$  in Theorem 2.6 to overcome this difficulty and conclude some results on the spectrum  $\sigma(\mathcal{A}_{PI})$ .

**5. Numerical applications.** Consider a river reach (1.1)–(1.2) with length  $\ell = 2000m$  and a reference discharge  $Q_0 = 2m^3/s$ . The coefficients are hence  $\alpha = 664m^2/s$  and  $\beta = 7.7854m/s$  (for more details about the model, the reader is referred to [18]). Now, assume the constant perturbation  $w = 1/2$  and the reference  $y_r = 1$ . Next, we apply the finite difference method for the space variable  $x$

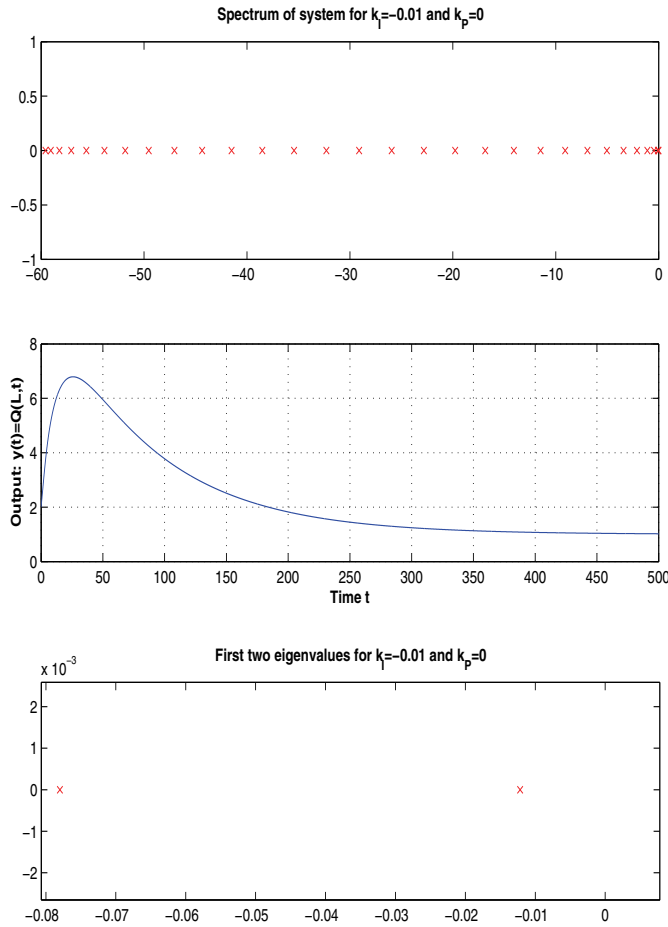


FIG. 6. Output regulation for  $K_I = -0.05$ ,  $K_P = 0$ .

to transform the distributed parameter system (1.1)–(1.2) to a first order system of differential equations and then use MATLAB. Taking the proportional gain  $k_P = 0$  and the integral gain  $k_I = -0.01$ , we observe that the system spectrum consists of negative real numbers, and the output  $y(t) = Q(\ell, t)$  is regulated to  $y_r = 1$  (see Figure 6). Furthermore, we notice that for sufficiently small values of  $k_P$ , the spectrum moves to the left in the sense that the first eigenvalue of the system is shifted to the left (see the third parts of Figures 6 and 7) and thus the regulation is guaranteed with smaller values of time  $t$  (see Figures 6 and 7). In other words, less time is needed to regulate the system, and hence both exponential stability and regulation are sped up. For  $k_P = 0.259$ , one complex conjugate pair, with negative real part, appears but we still have the stability as well as the regulation of the system (see Figure 8). However, although  $k_I$  is unchanged, i.e.,  $k_I = -0.01$ , if the proportional gain  $k_P$  is not “sufficiently small,” for instance,  $k_P = 0.9$ , there are two complex eigenvalues with positive real part, and hence neither the stability nor the output regulation is guaranteed (see Figure 9).

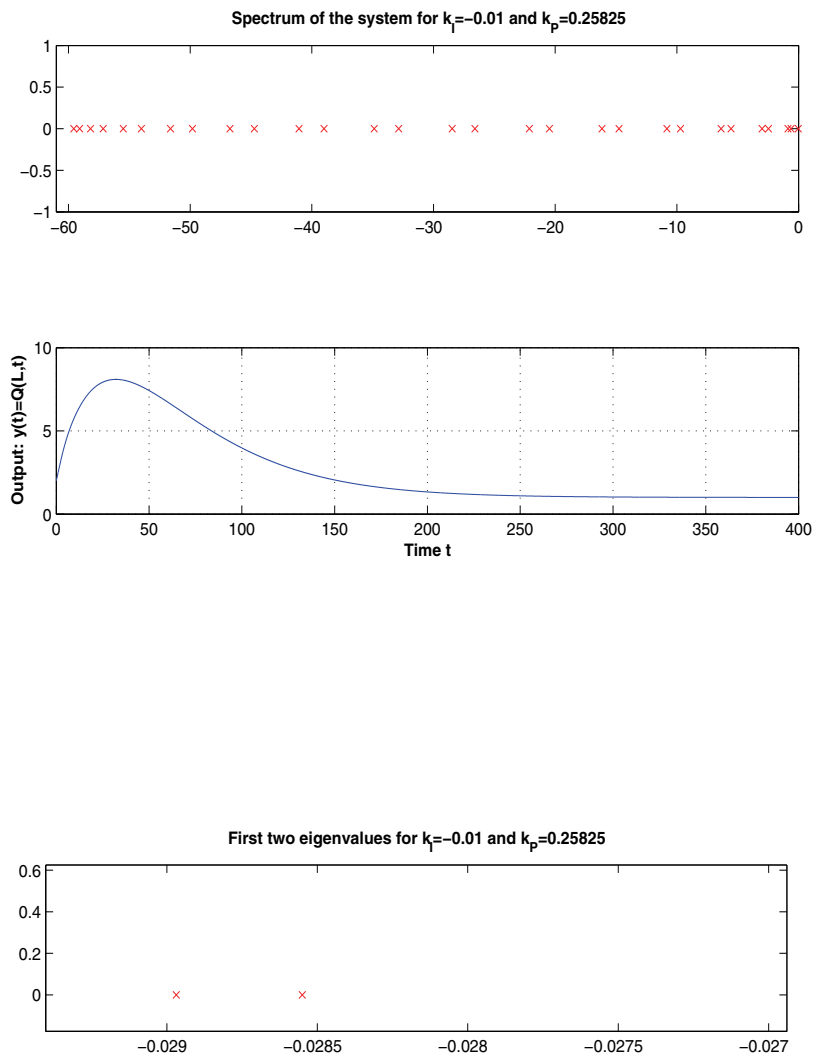


FIG. 7. Negative real spectrum: Stability and output regulation for  $k_I = -0.01$ ,  $k_P = 0.25825$ .

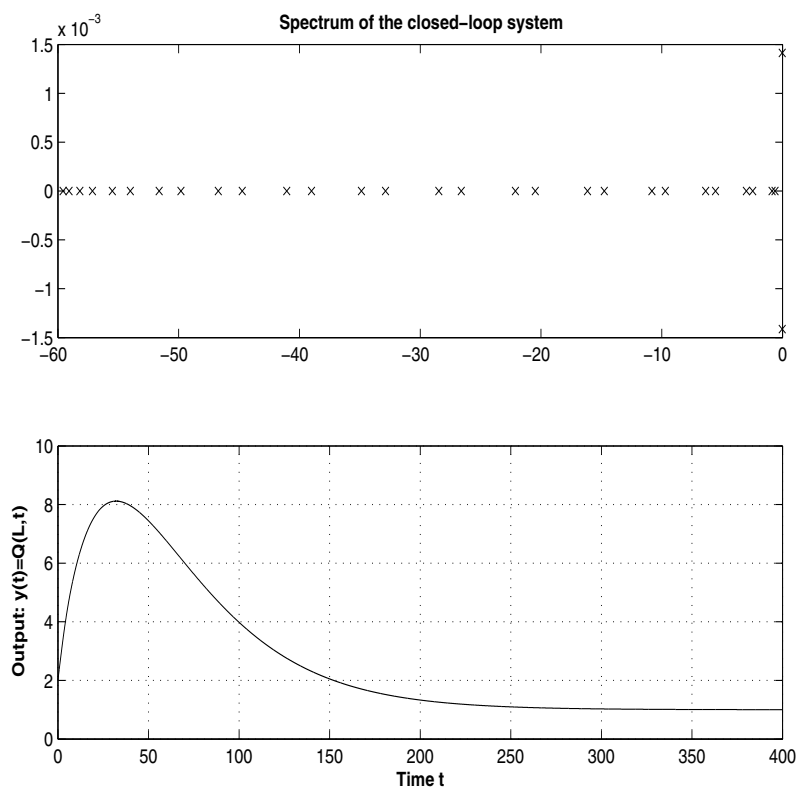


FIG. 8. One complex conjugate pair of eigenvalues with negative real part: Stability and output regulation for  $k_I = -0.01$ ,  $k_P = 0.259$ .

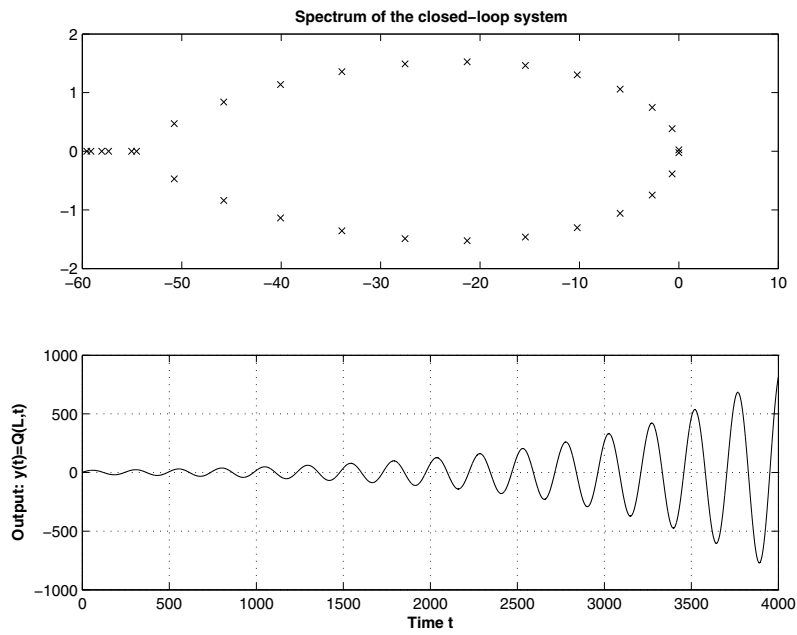


FIG. 9. One complex conjugate pair of eigenvalues with positive real part: Nonregulation for  $k_I = -0.01$ ,  $k_P = 0.9$ .



**Acknowledgments.** The authors are grateful to the associate editor and the referees for their constructive criticism and valuable suggestions for improving the paper.

## REFERENCES

- [1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] B. CHENTOUF AND J. M. WANG, *Stabilization of a one-dimensional dam-river system: A nondissipative and noncollocated case*, J. Optim. Theory Appl., 134 (2007), pp. 223–239.
- [3] V. T. CHOW, *Open Channel Hydraulics*, McGraw-Hill, New York, 1985.
- [4] R. F. CURTAIN AND H. ZWART, *An introduction to infinite-dimensional linear systems theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.
- [5] R. F. CURTAIN, H. LOGEMANN, AND O. STAFFANS, *Stability results of Popov-type for infinite-dimensional systems with applications to integral control*, Proc. London Math. Soc. (3), 86 (2003), pp. 779–816.
- [6] V. DOS-SANTOS, Y. TOURÉ, AND N. CISLO, *Régulation de canaux d'irrigation: Approche par contrôle frontière multivariable, et modèle interne d'EDP*, E-revue des Sciences et Technologies de l'Automatique, 1 (2004), article 13.
- [7] V. DOS-SANTOS AND Y. TOURÉ, *Irrigation multireaches regulation problem by internal model boundary control*, in Proceedings of the 44th IEEE Conference on Decision and Control and 2005 European Control Conference, Seville, Spain, pp. 1905–1910.
- [8] V. DOS-SANTOS, G. BASTIN, AND Y. TOURÉ, *Regulation in multi-reach open channels by internal model boundary control*, Int. J. Tomogr. Stat., 5 (2007), pp. 91–96.
- [9] P. DUFOUR, L. JOSSERAND, AND Y. TOURÉ, *Commande par actions frontières d'un système d'échangeurs de chaleur*, RAIRO-APII-JESA, 30 (1996), pp. 1375–1391.
- [10] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. Part III. Spectral Operators*, John Wiley, New York, 1971.
- [11] T. FLIEGNER, H. LOGEMANN, AND E. P. RYAN, *Low-gain integral control of well-posed linear infinite-dimensional systems with input and output nonlinearities*, J. Math. Anal. Appl., 261 (2001), pp. 307–336.
- [12] T. FLIEGNER, H. LOGEMANN, AND E. P. RYAN, *Low-gain integral control of continuous-time linear systems subject to input and output nonlinearities*, Automatica J. IFAC, 39 (2003), pp. 455–462.
- [13] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1976.
- [14] J. M. GALLARDO, *Generation of analytic semigroups by second-order differential operators with nonseparated boundary conditions*, Rocky Mountain J. Math., 30 (1967), pp. 869–899.
- [15] B.-Z. GUO, *Riesz basis approach to the stabilization of a flexible beam with a tip mass*, SIAM J. Control Optim., 39 (2001), pp. 1736–1747.
- [16] B.-Z. GUO, J.-M. WANG, AND S.-P. YUNG, *Boundary stabilization of a flexible manipulator with rotational inertia*, Differential Integral Equations, 18 (2005), pp. 1013–1038.
- [17] B. YA. LEVIN, *Lectures on Entire Functions*, Transl. Math. Monogr. 150, AMS, Providence, RI, 1996.
- [18] X. LITRICO AND D. GEORGES, *Robust continuous-time and discrete-time flow control of a dam-river system. (I) Modeling*, Appl. Math. Model., 23 (1999), pp. 809–827.
- [19] X. LITRICO AND D. GEORGES, *Robust continuous-time and discrete-time flow control of a dam-river system. (II) Controller design*, Appl. Math. Model., 23 (1999), pp. 829–846.
- [20] J. LOCKER, *Spectral Theory of Non-Self-Adjoint Two-Point Differential Operators*, Math. Surveys Monogr. 73, AMS, Providence, RI, 2000.
- [21] H. LOGEMANN AND R. F. CURTAIN, *Absolute stability results for well-posed infinite-dimensional systems with applications to low-gain integral control*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 395–424.
- [22] H. LOGEMANN AND D. H. OWENS, *Robust high-gain feedback control of infinite-dimensional minimum-phase systems*, IMA J. Math. Control Inform., 4 (1987), pp. 195–220.
- [23] H. LOGEMANN AND D. H. OWENS, *Low-gain control of unknown infinite-dimensional systems: A frequency-domain approach*, Dynam. Stability Systems, 4 (1989), pp. 13–29.
- [24] H. LOGEMANN AND E. P. RYAN, *Time-varying and adaptive integral control of infinite-dimensional regular linear systems with input nonlinearities*, SIAM J. Control Optim., 38 (2000), pp. 1120–1144.
- [25] H. LOGEMANN, E. P. RYAN, AND S. TOWNLEY, *Integral control of infinite-dimensional linear systems subject to input saturation*, SIAM J. Control Optim., 36 (1998), pp. 1940–1961.

- [26] H. LOGEMANN, E. P. RYAN, AND S. TOWNLEY, *Integral control of linear systems with actuator nonlinearities: Lower bounds for the maximal regulating gain*, IEEE Trans. Automat. Control, 44 (1999), pp. 1315–1319.
- [27] H. LOGEMANN AND S. TOWNLEY, *Low-gain control of uncertain regular linear systems*, SIAM J. Control Optim., 35 (1997), pp. 78–116.
- [28] H. LOGEMANN AND S. TOWNLEY, *Adaptive low-gain integral control of multivariable well-posed linear systems*, SIAM J. Control Optim., 41 (2003), pp. 1722–1732.
- [29] H. LOGEMANN AND H. ZWART, *On robust PI-control of infinite-dimensional systems*, SIAM J. Control Optim., 30 (1992), pp. 573–593.
- [30] Z. H. LUO, B. Z. GUO, AND O. MORGÜL, *Stability and Stabilization of Infinite Dimensional Systems with Applications*, Springer-Verlag, London, 1999.
- [31] P. O. MALATERRE, *Modélisation, Analyse et Commande Optimale LQR d'un Canal d'Irrigation*, Ph.D. thesis, ENGREF, Montpellier, France, 1994.
- [32] M. A. NAIMARK, *Linear Differential Operators I*, Frederick Ungar Publishing Company, New York, 1967.
- [33] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
- [34] S. A. POHJOLAINEN, *Robust multivariable PI-controllers for infinite dimensional systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 17–30.
- [35] S. A. POHJOLAINEN AND I. LÄTTI, *Robust controllers for boundary control systems*, Internat. J. Control, 38 (1983), pp. 1189–1197.
- [36] S. A. POHJOLAINEN, *Robust controller for systems with exponentially stable strongly continuous semigroups*, J. Math. Anal. Appl., 111 (1985), pp. 622–636.
- [37] B. DE SAINT-VENANT, *Théorie du mouvement non permanent des eaux avec application aux crues des rivières et à l'introduction des marrées dans leur lit*, C. R. Acad. Sci., 73 (1871), pp. 148–154.
- [38] A. A. SHKALIKOV, *Boundary value problems for ordinary differential equations with a parameter in the boundary conditions*, J. Soviet Math., 33 (1986), pp. 1311–1342.
- [39] Y. TOURÉ AND L. JOSSERAND, *Semigroup formalism and internal model for a heat exchanger*, E-Revues des Sciences et Technologies de l'Automatique, 1 (2004), article 5.
- [40] J.-M. WANG, G.-Q. XU, AND S.-P. YUNG, *Exponential stabilization of laminated beams with structural damping and boundary feedback controls*, SIAM J. Control Optim., 44 (2005), pp. 1575–1597.
- [41] C. Z. XU AND H. JERBI, *A robust PI-controller for infinite-dimensional systems*, Internat. J. Control, 61 (1995), pp. 33–45.
- [42] R. M. YOUNG, *An Introduction to Nonharmonic Fourier Series*, Academic Press, London, 2001.

# APPROXIMATE FIXED POINT ITERATION WITH AN APPLICATION TO INFINITE HORIZON MARKOV DECISION PROCESSES\*

ANTHONY ALMUDEVAR†

**Abstract.** Many approximate iterative algorithms can be represented by the form  $V_n = TV_{n-1} + U_n$ ,  $n \geq 1$ , where  $V_{n-1}, U_n$ ,  $n \geq 1$ , are elements of a seminormed linear space  $(\mathcal{V}, \|\cdot\|)$  and  $T$  is a contractive operator. The objective is usually to calculate a fixed point  $V^*$  of  $T$ . Here, the quantities  $U_n$  may be interpreted as the errors obtained when the operator  $T$  can be evaluated only approximately. As is well known, in the absence of such an approximation error, the algorithm converges to a fixed point with a geometric rate under general conditions. In this article the convergence properties of such algorithms in the presence of an approximation error  $U_n$  are studied. It is shown that the error  $\|V_n - V^*\|$  is dominated by  $\|U_n\|$ , so that convergence of  $\|U_n\|$  to zero implies the convergence of  $\|V_n - V^*\|$  to zero, at a rate determined by  $\|U_n\|$ . The results are naturally extended to a relative error of the form  $\|U_n\|/\|V_{n-1}\|$ , as well as to  $J$ -stage contractions. The utility of this general theory is then demonstrated by an extended application to the problem of model-based approximate and adaptive control of Markov decision processes. The theory is shown to permit a sharpening of known convergence rates under more general conditions. Additionally, bounds on regret for adaptive controls with forced exploration are calculated in terms of a stagewise exploration rate. This permits the determination of an optimal choice of exploration rate within the class of certainty-equivalence adaptive control policies.

**Key words.** fixed point, contraction mappings, dynamic programming, Markov decision processes

**AMS subject classifications.** 65C05, 90C40, 93E35

**DOI.** 10.1137/S0363012904441520

**1. Introduction.** Iterative algorithms play a central role in many areas of control theory, optimization, and numerical analysis. Such algorithms generate successive solutions  $V_n = TV_{n-1}$ ,  $n \geq 1$ , of a fixed point equation  $V = TV$  induced by a contractive operator  $T$ . Practical considerations often permit only an approximate evaluation of  $T$ , so characterizing the impact of such approximations is a problem of some importance. However, few general principles concerning this problem have emerged. The simplest case of a single uniform bound on approximations of  $T$  for one-dimensional real valued operators was considered in Isaacson and Keller [21]. More typically, the problem has been analyzed for specific applications. An important example is approximate value iteration for Markov decision processes (MDPs), discussed variously in Federgruen and Schweitzer [11], Hernández-Lerma [17], Moore and Atkeson [28], Bertsekas and Tsitsiklis [5], and Munos [29]. Additionally, approximate Monte Carlo schemes for iterative solutions of large linear and nonlinear systems of equations have been analyzed in Halton [14, 15].

The present article divides naturally into two parts. In the first, the basis for a general theory of approximate iterative algorithms is introduced. In the second, the utility of the approach is demonstrated with a detailed development of some

---

\*Received by the editors February 25, 2004; accepted for publication (in revised form) April 6, 2008; published electronically August 13, 2008. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

<http://www.siam.org/journals/sicon/47-5/44152.html>

†Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642 (anthony\_almudevar@urmc.rochester.edu).

problems associated with the approximate control of MDPs. In the remainder of this introduction a detailed outline is presented.

**1.1. General approximate iteration processes.** The general algorithm considered is of the form  $V_n = T_n V_{n-1} = TV_{n-1} + U_n$ , where  $V_{n-1}, U_n$  are elements of a seminormed linear space  $(\mathcal{V}, \|\cdot\|)$  on which the operator  $T$  is defined. It is assumed that  $T$  has a fixed point  $V^*$ , or an equivalence class of fixed points, to which the algorithm is designed to converge. This operator is assumed to be  $J$ -stage contractive of modulus  $\rho < 1$ . Then  $U_n$  may be interpreted as the error resulting from an approximate evaluation of  $TV_{n-1}$  in which  $T$  is replaced by an approximate operator  $T_n$ .

The main theoretical results are presented in section 2, consisting of a comprehensive study of the convergence properties of the general algorithm. In particular it is found that  $\|V_n - V^*\| \leq O(\max(\|U_n\|, (\rho + \epsilon)^n))$ ,  $\epsilon > 0$ , under general conditions. This result extends naturally to  $J$ -stage contractions and to the case of vanishing relative error  $\|U_n\|/\|V_{n-1}\| \rightarrow_n 0$ .

**1.2. Approximate and adaptive control of MDPs.** The remainder of the article consists of an application of the theoretical results to the problem of the approximate control of MDPs with infinite horizon and discounted costs. In the standard formulation, an MDP navigates indefinitely through a *state space*  $\mathcal{X}$  by discrete stages. At each state  $x$  an *action*  $a$  from *action space*  $\mathcal{A}$  is taken, resulting in *cost*  $R(x, a)$ . The process transfers to a subsequent state according to probability measure  $Q(\cdot | x, a)$ . The state-action space  $\mathcal{K} \subset \mathcal{X} \times \mathcal{A}$  consists of all state-action pairs  $(x, a)$  for which action  $a$  is available from state  $x$ . If  $\beta$  is the discount factor, we refer to the object  $\pi = (\mathcal{K}, Q, R, \beta)$  as a Markov control model (MCM). A *control policy*  $\Phi$  is a rule (which may be probabilistic) for choosing an action based on the current state and process history. An MCM  $\pi$  together with a control policy  $\Phi$  defines an MDP. The objective is to minimize the expected total discounted cost from any given initial state.

Calculation of an optimal control usually proceeds by considering the *value function*  $V_\pi^*$ , which gives the lowest achievable expected discounted cost  $V_\pi^*(x)$  for initial state  $x$  under MCM  $\pi$ . Under general conditions,  $V_\pi^*$  is the fixed point of the *Bellman operator*  $T_\pi$ , which is defined for a specific MCM  $\pi$ . Then  $V_\pi^*$  may be calculated by the iterative algorithm  $V_n = T_\pi V_{n-1}$ ,  $n \geq 1$ , usually referred to as the *value iteration* (VI), and it is easily established that the algorithm converges to  $V_\pi^*$ . The optimal control can then be derived by first calculating  $V_\pi^*$ , then implicitly through an evaluation of  $T_\pi V_\pi^*$ . Therefore,  $\pi$  must be known in order to optimize cost.

Often, the evaluation of  $T_\pi V$  is impossible or impractical. Two important cases may be identified. The first arises when the evaluation of  $T_\pi V$  is computationally unfeasible and is carried out using approximation methods [5]. The second case arises when the MCM  $\pi$  on which  $T_\pi$  depends is unknown but may be estimated (see, for example, Kumar and Varaiya [24] or Hernández-Lerma [17]). In each case, a VI algorithm may be carried out approximately, resulting in some estimate of  $V_\pi^*$ . What is needed is to first characterize the effect of the approximation on the evaluation of  $T_\pi V$  and then to deduce the behavior of the approximate VI algorithm.

New issues arise when considering *adaptive control policies*, definable as policies which refine the applied control using data collected while the MDP is active. Ideally, these policies approach optimal cost performance when  $\pi$  is unknown. The analytical object is to estimate *regret*, which is the excess in costs obtainable from the adaptive control over the optimal costs obtainable when  $\pi$  is known. Because many such

policies rely on sequential refinements of estimates of  $V_\pi^*$ , the theory of approximate VI has direct bearing on the estimation of regret. Thus, the problem of approximating the value function and the problem of adaptive control are naturally considered in sequence, as in Hernández-Lerma and Marcus [16] and Hernández-Lerma [17], and which is the approach taken here.

The overall strategy will be to consider approximate VI as a special case of the general algorithm discussed in section 2. If  $\hat{\pi}$  is an estimate of MCM  $\pi$ , then the Bellman operator  $T_\pi$  can be approximated by  $T_{\hat{\pi}}$ . The analysis then proceeds by first defining a distance between models  $\pi$  and  $\hat{\pi}$ , then bounding  $\|T_\pi V - T_{\hat{\pi}} V\|$  with this distance for a suitable norm. Once this is done, the properties of approximate VI algorithms, and of approximate and adaptive controls, may be deduced directly from the model estimation process. Because the main results follow from properties which may be characterized purely in terms of normed linear spaces, it becomes straightforward to define a single general model. In the development which follows,  $\mathcal{X}$  and  $\mathcal{A}$  will be general Borel spaces. We also permit the cost function  $R(x, a)$  to be unbounded. The essential condition is the existence of a norm with respect to which the Bellman operator  $T_\pi$  is  $J$ -stage contractive. In this article we employ the class of weighted supremum norms. Much of the technical details presented here follow approaches taken in previous work in the literature but are needed to accommodate the use of this type of norm, as well as the related model distance, as will now be discussed.

**1.2.1. Approximate VI with general bounds.** When VI can be implemented only approximately, this can be modelled by a sequence of approximate operators  $T_n$ ,  $n \geq 1$ , close to the Bellman operator  $T_\pi$  in some sense, which defines an *approximate value iteration* algorithm (AVIA)  $V_n = T_n V_{n-1}$ ,  $n \geq 1$ . The technical problem is then to bound  $\|V_n - V_\pi^*\|$  asymptotically, or for finite  $n$  as needed. Such a bound will follow from a uniform bound on evaluation error  $\|T_n V_{n-1} - T_\pi V_{n-1}\| \leq \epsilon$  as discussed in, for example, [5, 29].

Such uniform bounds may also be obtained directly from the theory in section 2, and they rely primarily on the contractive properties of  $T_\pi$ . The analysis may then be extended to nonstationary bounds  $\|T_n V_{n-1} - T_\pi V_{n-1}\| \leq \epsilon_n$ . When these bounds converge to zero, it follows that the AVIA converges to  $V_\pi^*$  with a rate determined by  $\epsilon_n$ . The result holds when only relative error  $\|T_n V_{n-1} - T_\pi V_{n-1}\|/\|V_{n-1}\| \leq \epsilon_n$  can be bounded. The results are easily extended to  $J$ -stage contractive operators.

**1.2.2. Approximate VI using model estimates.** In section 3 the general MDP model is defined, and the problem of approximate VI based on model estimation is considered. A model distance between  $\pi$  and  $\hat{\pi}$  is defined and used to bound evaluation error  $\|T_{\hat{\pi}} V - T_\pi V\|$ , allowing for weighted supremum norms.

If we are given a sequence of estimates  $\hat{\pi}_n$ ,  $n \geq 1$ , of a model  $\pi$ , then an AVIA can be defined by the approximate operators  $T_n = T_{\hat{\pi}_n}$ . In [11] it was shown for finite  $\mathcal{K}$  that if the model estimates are consistent, this AVIA will converge to  $V_\pi^*$ . A convergence rate was also determined, which depends directly on the rate of convergence of the model estimate. In [17] this result was extended to general (Borel) state-action spaces with bounded costs. We improve on these results in two ways, first by allowing unbounded costs and general state-action spaces, then by strictly improving the rates of convergence reported in [11, 17].

As an alternative to the AVIA, we may determine the fixed point of each operator  $T_{\hat{\pi}_n}$ , yielding a sequence of value functions  $V_n$  for each approximate MCM  $\hat{\pi}_n$ . We refer to this as a *certainty-equivalence* algorithm (CEA). Intuitively, this would seem to be

the superior approach, provided it is computationally feasible. Rates of convergence of  $V_n$  to  $V_\pi^*$  are reported in [17] for the CEA, and equivalent rates for the more general model are derived in section 3 below. Interestingly, the rates of convergence for the AVIA reported in [11, 17] are strictly slower than for the CEA, whereas the sharper rates derived in section 3 below are equivalent to those for the CEA, except that they are asymptotic.

**1.2.3. Adaptive control policies.** A model-based adaptive control policy generates a sequence of model estimates  $\hat{\pi}_n$ ,  $n \geq 1$ , typically using process history, while exploiting these estimates to approximate optimal control. One natural approach, known as the *principle of estimation and control* (PEC) [17] (alternatively, *certainty-equivalence* [24]), is to refine at regular intervals the applied control policy based on the most recent model estimate. This may be done using a synchronous AVIA or CEA.

In section 4, the problem of estimating the regret of an adaptive control policy is discussed. Here we follow the approach taken in Schäl [30], in which the concept of *asymptotic discount optimality* was introduced. If the initial state  $x$  of the MDP is fixed, we may consider the expected remaining discounted cost at stage  $n$  under a policy  $\Phi$  and compare it to the expected remaining discounted cost when  $\Phi$  is used up to stage  $n - 1$  and the optimal control is used subsequently. In [30] costs may be unbounded and the action space is Borel, whereas the state space is assumed to be countable. The concept has been extended to more general models in [17] and in Hernández-Lerma and Lasserre [18, 19]. In section 4, bounds are developed for the general model which employ the weighted supremum norm and the model distances developed in section 3.

In section 5, the parameterization of the MCM is discussed, with the objective of verifying that model distances developed in section 3 are Lipschitz continuous on the parameter space. Similar concepts are reported in, for example, [24, 17], but are here extended to accommodate weighted supremum norms.

**1.2.4. Forced exploration.** It is important to note that for a general PEC adaptive control policy there is no guarantee that the applied control approaches the optimal one, since the model estimate need not be consistent under these conditions [24]. The issue is a fundamental one, since if optimal control is not approached, the MDP will indefinitely accrue excess costs (i.e., nonzero regret). Technically, the issue is straightforward. Suppose  $\mathcal{K}$  is discrete. In the absence of additional structure, a consistent estimate of the MCM is not achievable unless each available action is selected from each state infinitely often, which is not guaranteed using a PEC policy.

When convergence of a PEC policy to optimal performance cannot be guaranteed, one alternative is the use of *forced exploration*, in which exploratory selection of actions alternates with the application of an approximately optimal policy. If the exploration frequency is decreased at a suitable rate, then consistency conditions may be met, while achieving convergence to optimal performance. Such policies have been proposed in Kumar and Varaiya [24] and Cybenko, Gray, and Moizumi [7], with convergence to optimality verified. In Kearns and Singh [23] a policy which directs the MDP to underexplored states was proposed, with reported regret bounds of order  $n^{-1/5}$  after the  $n$ th stage. This applies to a discounted and average cost criterion for finite  $\mathcal{K}$ .

In section 6 a form of adaptive control is presented which accommodates forced exploration in the form of scheduled randomized exploration. A rate of exploration  $\alpha_n$  is defined, interpreted as the frequency of exploratory stages near stage  $n$ . This rate

then determines the rate of convergence for model estimates. The theory of sections 3, 4, and 5 then permits a careful calculation of a bound on regret, which may be decomposed into regret due to suboptimal control and due to forced exploration. This in turn permits selection of an optimal rate of exploration. In particular, we find that among exploration rates proportional to  $n^{-r}$ , the optimal choice is  $r = 1/3$ , resulting in a bound on regret of order  $n^{-1/3}$ , whereas if forced exploration were not needed, a bound of order  $n^{-1/2}$  would be obtained.

**1.2.5. Model-free adaptive control policies.** Although the methods considered in this paper rely on model estimation, it is important to note that a number of model-free methods exist (see Sutton and Barto [31]). A well-known example is *Q-learning*, due to Watkins [37] (see also Watkins and Dayan [38]), which is a method of directly estimating the value function, and hence the optimal control, from online data. Conditions for convergence include visitation of each state-action pair  $(x, a)$  infinitely often so, as for a PEC policy, forced exploration may be necessary. See Tsitsiklis [34] for rigorous convergence criteria. Q-learning has a convergence rate that is strictly greater than the PEC policy. See Szepesvári [32], Kearns and Singh [22], or Evan-Dar and Mansour [10] for a discussion of this issue. On the other hand, the principal advantage of Q-learning is that value function estimates are updated without integration over  $\mathcal{X}$ . Thus Q-learning may be preferable to a PEC policy when updating the control using VI is not feasible online.

A bandit theory approach to the adaptive control of MDPs is proposed in Lai and Yakowitz [25]. This adaptive control policy proceeds by applying to the MDP in turn a sequence of control policies selected in random order. The selection procedure tends to remain longer in controls with lower observed costs. For general  $\mathcal{X}$  and countably many control policies, this approach achieves a growth in cumulative regret for undiscounted costs of  $O(a_n \log(n))$  for any increasing unbounded sequence  $a_n$ . This is equivalent to a stage  $n$  regret less than  $n^{-1+\epsilon}$ ,  $\epsilon > 0$ . However, exploratory randomization is applied over the entire space of control policies rather than the action space, which is the more common approach. Unless the set of control policies is of low dimension, the resulting performance will be subject to a high degree of variability attributable to the randomization scheme, an effect reported in [25]. Despite this, the much faster theoretical rate of convergence suggests that a careful comparison of this approach with the one described in section 6 is warranted.

**2. Approximate fixed point algorithms.** Suppose  $(\mathcal{V}, \|\cdot\|)$  is a seminormed linear space on which a mapping  $T : \mathcal{V} \rightarrow \mathcal{V}$  is given. Define the operator  $T^J : \mathcal{V} \rightarrow \mathcal{V}$  to be the  $J$ th iteration of  $T$  on any element of  $\mathcal{V}$ . Consider the following two assumptions:

- (A1) *Pseudocontraction.* There exists  $V^* \in \mathcal{V}$  with  $\|V^*\| < \infty$ , a positive integer  $J$ , and a constant  $\rho \in (0, 1)$  such that  $TV^* = V^*$  and  $\|T^J V - T^J V^*\| \leq \rho \|V - V^*\|$  for all  $V \in \mathcal{V}$ .
- (A2) *Lipschitz continuity.* There exists finite  $L$  such that  $\|TV_1 - TV_2\| \leq L\|V_1 - V_2\|$  for all  $V_1, V_2 \in \mathcal{V}$ .

If (A1) holds and  $\|\cdot\|$  is a true norm,  $V^*$  is the unique fixed point of  $T$ . If  $\|\cdot\|$  is a seminorm, any other fixed point is a member of an equivalence class  $\{V \in \mathcal{V} : \|V - V^*\| = 0\}$ . We will study convergence properties of sequences of the form  $\|V_n - V^*\|$  exclusively. Whether this implies convergence of  $V_n$  to  $V^*$  in some other sense depends on the particular structure of the seminorm.

For most of the theory presented in this article the *pseudocontraction* condition of (A1) suffices. The stronger conditions under which the Banach fixed point theorem

holds ( $T$  is ( $J$ -stage) contractive and  $(\mathcal{V}, \|\cdot\|)$  is closed and complete) imply (A1). In either case for the fixed point algorithm

$$(2.1) \quad \begin{aligned} V_0 &= v_0 \in \mathcal{V}, \\ V_n &= TV_{n-1}, \quad n \geq 1, \end{aligned}$$

$\|V_n - V^*\|$  converges to zero. Assumption (A2) will be employed in addition to (A1) when  $J > 1$ .

The purpose of this article is to study the related algorithm

$$(2.2) \quad \begin{aligned} V_0 &= v_0 \in \mathcal{V}, \\ V_n &= TV_{n-1} + U_n, \quad n \geq 1, \end{aligned}$$

where  $U_n \in \mathcal{V}$  is in some sense small and typically represents the error term of an approximate evaluation of  $TV_{n-1}$ . It will be natural to think of (2.2) as arising from the replacement of operator  $T$  with an approximate operator  $T_n$  for the  $n$ th iteration, setting  $T_n V_{n-1} = TV_{n-1} + U_n$ . We will generally impose the following assumption on any approximate algorithm (2.2).

(A3) In (2.2),  $\|v_0\| < \infty$  and  $\|U_n\| < \infty$  for  $n \geq 1$ .

Condition (A1) with  $J = 1$  and (A3) guarantees that each iterate has a finite seminorm, since

$$(2.3) \quad \begin{aligned} \|V_n\| &\leq \|V^*\| + \|V_n - V^*\| \leq \|V^*\| + \rho\|V_{n-1} - V^*\| + \|U_n\| \\ &\leq \rho\|V_{n-1}\| + (1 + \rho)\|V^*\| + \|U_n\|, \quad n \geq 1, \end{aligned}$$

and we have  $\|V_0\| < \infty$ . Conditions (A2) and (A3) also imply  $\|V_n\| < \infty$ ,  $n \geq 1$ , by the same argument with  $\rho$  replaced by  $L$ .

Our immediate objective is to show that  $\|U_n\| \rightarrow_n 0$  implies  $\|V_n - V^*\| \rightarrow_n 0$  in (2.2), or, when  $\|U_n\|$  does not converge to zero, to bound the quantity  $\|V_n - V^*\|$  in terms of the error magnitudes  $\|U_n\|$ .

**2.1. Introductory lemmas.** The results given in this section depend on the properties of summations of the form  $\sum_{i=1}^n \rho^{n-i} \alpha_i$ , where  $\rho \in (0, 1)$  and the sequence  $\alpha_1, \alpha_2, \dots$  is some representation of the magnitude of the error sequence  $U_1, U_2, \dots$ .

LEMMA 2.1. *If  $\alpha_1, \alpha_2, \dots$  is a sequence of real numbers in  $[0, \infty)$  and  $\rho \in (0, 1)$ , then*

$$(2.4) \quad \limsup_{n \rightarrow \infty} \sum_{i=1}^n \rho^{n-i} \alpha_i \leq \limsup_{n \rightarrow \infty} \frac{\alpha_n}{1 - \rho}.$$

*Proof.* We may assume  $\limsup_{n \rightarrow \infty} \alpha_n = K$  is finite; otherwise (2.4) holds trivially. Then for any  $\epsilon > 0$  we may select finite  $N_\epsilon$  such that  $\sup_{i > N_\epsilon} \alpha_i \leq K + \epsilon$ . Then for  $n > N_\epsilon$ ,

$$\sum_{i=1}^n \rho^{n-i} \alpha_i = \sum_{i=1}^{N_\epsilon} \rho^{n-i} \alpha_i + \sum_{i=N_\epsilon+1}^n \rho^{n-i} \alpha_i \leq \rho^n \sum_{i=1}^{N_\epsilon} \rho^{-i} \alpha_i + (K + \epsilon) \sum_{i=0}^n \rho^{n-i}.$$

The first term of this upper bound is  $\rho^n$  multiplied by a finite summation which does not depend on  $n$ ; hence this term vanishes as  $n \rightarrow \infty$ . The second term converges to  $(K + \epsilon)/(1 - \rho)$ . This is true for any  $\epsilon$ , from which (2.4) follows.  $\square$



The following lemma is a version of L'Hôpital's rule, the proof of which follows Theorem 9.1 in Fischer [13].

LEMMA 2.2. *Suppose  $\{a_n\}$ ,  $\{b_n\}$  are two real valued sequences such that  $b_{n+1} > b_n > 0$  for all  $n$ , and  $\lim_{n \rightarrow \infty} b_n = \infty$ . Then for any finite  $L$ ,*

$$(2.5) \quad \liminf_{n \rightarrow \infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n} \geq L \text{ implies } \liminf_{n \rightarrow \infty} \frac{a_n}{b_n} \geq L.$$

*Proof.* Suppose for each  $\epsilon > 0$  there exists  $N_\epsilon$  such that  $(a_{n+1} - a_n)/(b_{n+1} - b_n) > L - \epsilon$  for  $n \geq N_\epsilon$ . It follows that, for  $k > 0$ ,

$$\frac{a_{N_\epsilon+k} - a_{N_\epsilon}}{b_{N_\epsilon+k} - b_{N_\epsilon}} > L - \epsilon,$$

which is equivalent to

$$\frac{a_{N_\epsilon+k}}{b_{N_\epsilon+k}} > \frac{a_{N_\epsilon}}{b_{N_\epsilon+k}} + \left(1 - \frac{b_{N_\epsilon}}{b_{N_\epsilon+k}}\right)(L - \epsilon).$$

Letting  $k \rightarrow \infty$  gives

$$\liminf_{k \rightarrow \infty} \frac{a_{N_\epsilon+k}}{b_{N_\epsilon+k}} > L - \epsilon,$$

from which (2.5) follows.  $\square$

If  $\alpha_i$  converges at a well-defined rate, then we can express the convergent behavior of  $\sum_{i=1}^n \rho^{n-i} \alpha_i$  in terms of that of  $\alpha_i$ .

LEMMA 2.3. *Suppose  $\rho > 0$ . Let  $\{\alpha_n\}$  be a positive sequence of real numbers. Let*

$$r_u = \limsup_{n \rightarrow \infty} \frac{\alpha_{n-1}}{\alpha_n}.$$

*Suppose  $\sum_{i=1}^n \rho^{-i} \alpha_i = \infty$ . Then*

$$(2.6) \quad \liminf_{n \rightarrow \infty} \frac{\rho^{-n} \alpha_n}{\sum_{i=1}^n \rho^{-i} \alpha_i} \geq 1 - r_u \rho.$$

*Proof.* Using the notation of Lemma 2.2, set  $a_n = \rho^{-n} \alpha_n$  and  $b_n = \sum_{i=1}^n \rho^{-i} \alpha_i$ . Then

$$\frac{a_n - a_{n-1}}{b_n - b_{n-1}} = \frac{\rho^{-n} \alpha_n - \rho^{-(n-1)} \alpha_{n-1}}{\rho^{-n} \alpha_n} = 1 - \rho \frac{\alpha_{n-1}}{\alpha_n}.$$

Taking the limit supremum and applying Lemma 2.2 gives (2.6).  $\square$

We deal separately with the case of geometric rate  $\rho$ .

LEMMA 2.4. *If, for  $\rho \in (0, 1)$ ,*

$$\limsup_{n \rightarrow \infty} n^{-1} \log(\alpha_n) \leq \log(\rho),$$

*then*

$$(2.7) \quad \limsup_{n \rightarrow \infty} n^{-1} \log \left( \sum_{i=1}^n \rho^{n-i} \alpha_i \right) \leq \log(\rho).$$

*Proof.* By hypothesis, given  $\epsilon > 0$  there exists finite  $K_\epsilon$  such that  $\alpha_n \leq K_\epsilon(\rho + \epsilon)^n$  for all  $n$ . Then

$$\begin{aligned} n^{-1} \log \left( \sum_{i=1}^n \rho^{n-i} \alpha_i \right) &\leq n^{-1} \log \left( K_\epsilon \rho^n \sum_{i=1}^n (1 + \epsilon/\rho)^i \right) \\ &\leq \log(\rho) + n^{-1} \log \left( K_\epsilon \frac{(1 + \epsilon/\rho)^{n+1}}{\epsilon/\rho} \right) \\ &= \log(\rho) + \log(1 + \epsilon/\rho) + n^{-1} \log(K_\epsilon(\rho/\epsilon + 1)). \end{aligned}$$

Taking the limit supremum gives

$$\limsup_{n \rightarrow \infty} n^{-1} \log \left( \sum_{i=1}^n \rho^{n-i} \alpha_i \right) \leq \log(\rho) + \log(1 + \epsilon/\rho),$$

which gives (2.7) by making  $\epsilon$  arbitrarily small.  $\square$

**2.2. Single stage contractions.** The aim of this subsection is to study those convergence properties of  $\|V_n - V^*\|$  of (2.2) which follow from the convergence properties of  $U_n$ ,  $n \geq 1$ , when (A1) holds with  $J = 1$ . We additionally assume (A3) holds, which guarantees that the iterates have finite seminorm. We first show that the algorithmic accuracy  $\|V_n - V^*\|$  can be asymptotically bounded by the error magnitude  $\|U_n\|$ .

LEMMA 2.5. *If in (2.2) (A1) ( $J=1$ ) and (A3) hold, then*

$$(2.8) \quad \limsup_{n \rightarrow \infty} \|V_n - V^*\| \leq (1 - \rho)^{-1} \limsup_{n \rightarrow \infty} \|U_n\|.$$

*Proof.* Fix any initial point  $v_0 \in \mathcal{V}$ . Subtract  $V^*$  from each side of (2.2), giving for  $n \geq 1$ ,

$$(2.9) \quad \begin{aligned} \|V_n - V^*\| &= \|TV_{n-1} - V^* + U_n\| \leq \|TV_{n-1} - V^*\| + \|U_n\| \\ &= \|TV_{n-1} - TV^*\| + \|U_n\| \leq \rho \|V_{n-1} - V^*\| + \|U_n\|. \end{aligned}$$

Applying (2.9) iteratively gives

$$(2.10) \quad \|V_n - V^*\| \leq \rho^n \|V_0 - V^*\| + \sum_{i=1}^n \rho^{n-i} \|U_i\|.$$

Then (2.8) follows by applying Lemma 2.1 and noting that  $\rho < 1$ .  $\square$

By Lemma 2.5 the boundedness of (2.2) follows from the boundedness of the sequence  $\|U_n\|$ . In some applications it may be more natural to bound the relative error  $\|U_n\|/\|V_{n-1}\|$ . Lemma 2.6 deals with a somewhat more general case in which bounds are assumed on the sequence  $\|U_n\|/\max(s, \|V_{n-1}\|)$ , where  $s$  is a nonnegative constant.

LEMMA 2.6. *If in (2.2) (A1) ( $J = 1$ ) and (A3) hold, then for any  $s \geq 0$  and  $\delta \in [0, 1 - \rho)$ ,*

$$\limsup_{n \rightarrow \infty} \|U_n\|/\max(s, \|V_{n-1}\|) = \delta,$$

*then*

$$\limsup_{n \rightarrow \infty} \|V_n - V^*\| \leq \delta(1 - \rho - \delta)^{-1} \max(s, \|V^*\|).$$

*Proof.* Fix  $\epsilon > 0$  such that  $\rho + \delta + \epsilon < 1$ . Then there exists  $N_\epsilon$  such that  $\|U_n\| \leq (\delta + \epsilon) \max(s, \|V_{n-1}\|)$  for all  $n \geq N_\epsilon$ . Then

$$\begin{aligned}
 \|V_n - V^*\| &\leq \|TV_{n-1} - TV^*\| + \|U_n\| \\
 &\leq \rho\|V_{n-1} - V^*\| + (\delta + \epsilon) \max(s, \|V_{n-1}\|) \\
 &\leq \rho\|V_{n-1} - V^*\| + (\delta + \epsilon) \max(s, \|V_{n-1} - V^*\| + \|V^*\|) \\
 &= \max(\rho\|V_{n-1} - V^*\| + (\delta + \epsilon)s, (\rho + \delta + \epsilon)\|V_{n-1} - V^*\| + (\delta + \epsilon)\|V^*\|) \\
 (2.11) \quad &\leq (\rho + \delta + \epsilon)\|V_{n-1} - V^*\| + (\delta + \epsilon) \max(s, \|V^*\|), \quad n \geq N_\epsilon.
 \end{aligned}$$

Applying (2.11) iteratively,

$$\begin{aligned}
 \|V_{N_\epsilon+n} - V^*\| &\leq (\rho + \delta + \epsilon)^n \|V_{N_\epsilon} - V^*\| + (\delta + \epsilon) \sum_{i=0}^{n-1} (\rho + \delta + \epsilon)^i \max(s, \|V^*\|) \\
 &\leq (\rho + \delta + \epsilon)^n \|V_{N_\epsilon} - V^*\| + (\delta + \epsilon)(1 - \rho - \delta - \epsilon)^{-1} \max(s, \|V^*\|);
 \end{aligned}$$

hence

$$\limsup_{n \rightarrow \infty} \|V_n - V^*\| \leq (\delta + \epsilon)(1 - \rho - \delta - \epsilon)^{-1} \max(s, \|V^*\|),$$

which proves the theorem by making  $\epsilon$  arbitrarily small.  $\square$

The next lemma gives some conditions under which the rate of convergence of  $\|V_n - V^*\|$  may be bounded.

LEMMA 2.7. *If in (2.2) (A1) ( $J = 1$ ) and (A3) hold, then*

(i)  $\sum_{i=1}^{\infty} \rho^{-i} \|U_i\| < \infty$  *implies*

$$(2.12) \quad \limsup_{n \rightarrow \infty} \rho^{-n} \|V_n - V^*\| < \infty;$$

(ii)  $\limsup_{n \rightarrow \infty} n^{-1} \log(\|U_n\|) \leq \log(\rho)$  *implies*

$$(2.13) \quad \limsup_{n \rightarrow \infty} n^{-1} \log(\|V_n - V^*\|) \leq \log(\rho);$$

(iii) *for any sequence  $\{d_i\}$ ,  $d_i \in (0, \infty)$  satisfying  $\|U_i\| \leq d_i$ ,  $i \geq 1$ , if  $r_u < \rho^{-1}$ , where  $r_u = \limsup_{n \rightarrow \infty} d_{n-1}/d_n$ , then*

$$(2.14) \quad \limsup_{n \rightarrow \infty} d_n^{-1} \|V_n - V^*\| \leq (1 - r_u \rho)^{-1}.$$

*Proof.* We treat the three cases in order.

(i) If  $\sum_{i=1}^{\infty} \rho^{-i} \|U_i\| = K < \infty$ , then from (2.10) we have

$$\|V_n - V^*\| \leq \rho^n \|V_0 - V^*\| + \rho^n K,$$

from which (2.12) follows.

(ii) From (2.10) we may write

$$n^{-1} \log(\|V_n - V^*\|) \leq n^{-1} \log \left( \rho^n \|V_0 - V^*\| + \sum_{i=1}^n \rho^{n-i} \|U_i\| \right).$$

Apply Lemma 2.4 to the above inequality, setting  $\alpha_i = \|U_i\|$  for  $i > 1$  and  $\alpha_1 = \|U_1\| + \rho\|V_0 - V^*\|$ . This yields (2.13).

(iii) From (2.10) we may write

$$(2.15) \quad d_n^{-1} \|V_n - V^*\| \leq \frac{\|V_0 - V^*\| + \sum_{i=1}^n \rho^{-i} d_i}{\rho^{-n} d_n}.$$

Note that  $r_u < \rho^{-1}$  implies  $\sum_{i=1}^{\infty} \rho^{-i} d_i = \infty$ ; hence taking the limit supremum of (2.15) while applying Lemma 2.3, noting that  $(1 - r_u \rho) > 1$ , yields (2.14).  $\square$

*Remark 2.1.* The sequence  $\{d_i\}$  in Lemma 2.7(iii) may be chosen so as to readily calculate  $r_u$ . If  $r_u = \rho^{-1}$ , then part (ii) is used, since the upper bound in (2.14) is  $\infty$ . As an example, if the sequence  $\{U_n\}$  is a mapping of independent and identically distributed (i.i.d.) sequences we may be able to deduce from the law of the iterated logarithm  $\|U_n\| \leq d_n = Kn^{-1/2} \log \log n$  for some finite  $K$ , in which case  $r_u = 1$ .

Conversely, we may devise a bounding sequence  $\{d_i\}$  which converges to zero, but which can't be used in Lemma 2.7 to obtain a meaningful rate under the assumption that  $\|U_n\| \leq d_n$ . For some  $\rho \in (0, 1)$  let  $d_n = \sup\{\rho^i : \rho^i \leq 1/n, i \in \mathbb{Z}^+\}$ . For large enough  $n$ ,  $d_n > \rho^n$  and so  $\sum_{i=1}^{\infty} \rho^{-i} d_i = \infty$ ; thus condition (i) is not applicable. Then note that there is an increasing subsequence  $n_i$ ,  $i \geq 1$ , such that  $d_{n_i} > 1/(n_i + 1)$ , from which it follows that  $\limsup_{n \rightarrow \infty} n^{-1} \log(d_n) = 0$ , so that condition (ii) is not applicable. In addition,  $r_u = \rho^{-1}$  so condition (iii) is not applicable. The situation is easily remedied by noting that  $d_n \leq 1/n$ , for which  $r_u = 1$ , so that a finite rate can be deduced using Lemma 2.7(iii). In general, if a sequence does not satisfy  $r_u < \rho^{-1}$ , it may be dominated by one that does.

Clearly, algorithm (2.2) will not converge more quickly than the error term. This is expressed formally in the next theorem. Note that in the following theorem  $T$  need not be a contraction map. In place, a weaker form of (A2) suffices.

**THEOREM 2.8.** *In algorithm (2.2) let*

$$d_n = \sup_{n' \geq n} \|U_{n'}\|$$

for  $n \geq 1$ . If  $V^*$  is a fixed point of  $T$ , and there exists a finite constant  $L$  such that  $\|TV - TV^*\| \leq L\|V - V^*\|$  for all  $V \in \mathcal{V}$ , then

$$(2.16) \quad \limsup_{n \rightarrow \infty} \frac{\|V_{n-1} - V^*\|}{d_n} > 0.$$

*Proof.* Suppose (2.16) does not hold. Fix  $\epsilon > 0$ . Then there exists  $N$  such that

$$(2.17) \quad \|V_{n-1} - V^*\| \leq \epsilon d_n \quad \forall n \geq N.$$

Let  $N_1$  be the smallest integer not less than  $N$  such that  $\|U_{N_1}\| \geq (1 - \epsilon)d_N$ , which exists by the definition of  $d_n$ . This implies  $\|U_{N_1}\| \geq (1 - \epsilon)d_{N_1}$ . Applying (2.17) gives

$$\begin{aligned} \|U_{N_1}\| &= \|(V_{N_1} - V^*) - (TV_{N_1-1} - TV^*)\| \\ &\leq \|V_{N_1} - V^*\| + \|TV_{N_1-1} - TV^*\| \\ &\leq \|V_{N_1} - V^*\| + L\|V_{N_1-1} - V^*\| \\ &\leq \epsilon d_{N_1+1} + L\epsilon d_{N_1} \\ &\leq (1 + L)\epsilon d_{N_1}. \end{aligned}$$

We can always set  $\epsilon$  small enough to force  $(1 + L)\epsilon d_{N_1} < (1 - \epsilon)d_{N_1}$ , in which case (2.17) leads to a contradiction, since  $\|U_{N_1}\| \geq (1 - \epsilon)d_{N_1}$ . Hence (2.16) follows.  $\square$

**2.3.  $J$ -stage contractions.** The results of the previous subsection extend naturally to  $J$ -stage contractions with the additional assumption that the operator is first order Lipschitz. Note that if  $L < 1$  in (A2), then  $T$  is contractive, and the previous results apply directly.

Consider the  $J$ -stage error term of (2.2),

$$W_n^J = V_n - T^J V_{n-J}, \quad n \geq J.$$

Under assumptions (A2) and (A3),  $\|V_n\|$  and  $\|T^J V_n\|$  are finite; hence so is  $\|W_n^J\|$ . The subsequence

$$(2.18) \quad V_{mJ+k} = T^J V_{(m-1)J+k} + W_{mJ+k}^J, \quad m \geq 1,$$

is then of type (2.2) with operator  $T^J$  satisfying (A1) and (A3) for  $J = 1$ , after a suitable relabelling, and the results of the previous subsection apply directly. However, bounds may be more naturally obtained for  $U_n$  as originally defined in (2.2). Hence the strategy in this subsection will be to derive generalizations of Lemmas 2.5–2.7 to the general  $J$ -stage contraction operator (Theorems 2.10, 2.11, and 2.12, respectively) while maintaining bounding conditions expressed in terms of the single stage error  $U_n$ . Note that Theorem 2.8 already applies to the general case.

We will need the following lemma.

LEMMA 2.9. *For algorithm (2.2) if (A2) holds, then for any  $I \geq 1$ ,  $n \geq 0$ ,*

$$(2.19) \quad \|V_{n+I} - T^I V_n\| \leq \sum_{i=1}^I L^{I-i} \|U_{n+i}\|.$$

*Proof.* We proceed by induction. Suppose (2.19) holds for some  $I \geq 1$ ; then

$$\begin{aligned} \|V_{n+I+1} - T^{I+1} V_n\| &= \|TV_{n+I} - T^{I+1} V_n + U_{n+I+1}\| \\ &\leq \|TV_{n+I} - T^{I+1} V_n\| + \|U_{n+I+1}\| \\ &\leq L\|V_{n+I} - T^I V_n\| + \|U_{n+I+1}\| \\ &\leq L \sum_{i=1}^I L^{I-i} \|U_{n+i}\| + \|U_{n+I+1}\| \\ &\leq \sum_{i=1}^{I+1} L^{I+1-i} \|U_{n+i}\|, \end{aligned}$$

and (2.19) holds for  $I+1$ . That (2.19) holds for  $I = 1$  follows directly from (2.2).  $\square$

By Lemma 2.9  $\|U_n\| \rightarrow_n 0$  directly implies  $\|W_n^J\| \rightarrow_m 0$  and therefore  $\|V_{mJ+k} - V^*\| \rightarrow_m 0$ . This holds for  $k = 0, \dots, J-1$ , from which it follows that  $\|V_n - V^*\| \rightarrow_n 0$ .

Define a *block error*

$$B_n^{J,L} = \sum_{j=1}^J L^{J-j} \|U_{n-J+j}\|, \quad n \geq J \text{ and } B_n^{J,L} = 0, \quad n < J,$$

for  $J \geq 1$ ,  $L \in \mathbb{R}^+$ . Note that for  $J = 1$ ,  $B_n^{J,L} = \|U_n\|$  for any value of  $L$ . Directly from Lemma 2.9, the error term in (2.18) can be bounded by

$$(2.20) \quad \|W_{mJ+k}^J\| \leq B_{mJ+k}^{J,L}, \quad m \geq 1.$$

Theorem 2.10 generalizes Lemma 2.5 to  $J$ -stage contractions.

THEOREM 2.10. *If in (2.2) (A1) and (A3) hold, and (A2) holds when  $J > 1$ , then*

$$(2.21) \quad \limsup_{n \rightarrow \infty} \|V_n - V^*\| \leq (1 - \rho)^{-1} \limsup_{n \rightarrow \infty} B_n^{J,L}.$$

*Proof.* If  $J = 1$ , set  $L$  to be any positive number. Then by (2.20) and Lemma 2.5 applied to (2.18) for some fixed  $k \geq 0$ ,

$$\limsup_{m \rightarrow \infty} \|V_{mJ+k} - V^*\| \leq \limsup_{m \rightarrow \infty} B_{mJ+k}^{J,L},$$

which holds for  $k = 0, \dots, J-1$ , from which (2.21) follows.  $\square$

Similarly, Theorem 2.11 generalizes Lemma 2.6.

THEOREM 2.11. *If in (2.2) (A1) and (A3) hold, and (A2) holds when  $J > 1$ , then for any  $s \geq 0$ ,  $\delta \in [0, 1 - \rho)$ ,*

$$(2.22) \quad \limsup_{n \rightarrow \infty} B_{n+J}^{J,L} / \max(s, \|V_n\|) = \delta$$

*implies*

$$\limsup_{n \rightarrow \infty} \|V_n - V^*\| \leq \delta(1 - \rho - \delta)^{-1} \max(s, \|V^*\|).$$

*Proof.* If  $J = 1$ , set  $L$  to be any positive number. For algorithm (2.18) for fixed  $k \geq 0$ , using (2.20), then (2.22), we obtain

$$\begin{aligned} \limsup_{m \rightarrow \infty} \|W_{mJ+k}^J\| / \max(s, \|V_{(m-1)J+k}\|) &\leq \limsup_{m \rightarrow \infty} B_{mJ+k}^{J,L} / \max(s, \|V_{(m-1)J+k}\|) \\ &= \limsup_{m \rightarrow \infty} B_{mJ+k}^{J,L} / \max(s, \|V_{(m-1)J+k}\|) \\ &\leq \delta. \end{aligned}$$

We may therefore apply Lemma 2.6 to (2.18), yielding

$$\limsup_{m \rightarrow \infty} \|V_{mJ+k} - V^*\| \leq \delta(1 - \rho - \delta)^{-1} \max(s, \|V^*\|).$$

Letting  $k = 0, \dots, J-1$  completes the proof.  $\square$

Finally, Theorem 2.12 generalizes Lemma 2.7. It will be convenient to define the function

$$\sigma(r, J) = \begin{cases} \frac{1-r^J}{1-r}; & r \neq 1, \\ J; & r = 1, \end{cases}$$

which is continuous in  $r \in \Re$  for any fixed  $J$ .

THEOREM 2.12. *If for algorithm (2.2) (A1) and (A3) hold and (A2) holds when  $J > 1$ , then*

(i)  $\sum_{i=1}^{\infty} (\rho^{1/J})^{-i} \|U_i\| < \infty$  *implies*

$$(2.23) \quad \limsup_{n \rightarrow \infty} (\rho^{1/J})^{-n} \|V_n - V^*\| < \infty;$$

(ii)  $\limsup_{n \rightarrow \infty} n^{-1} \log(\|U_n\|) \leq \log(\rho^{1/J})$  *implies*

$$(2.24) \quad \limsup_{n \rightarrow \infty} n^{-1} \log(\|V_n - V^*\|) \leq \log(\rho^{1/J});$$

(iii) for any sequence  $\{d_i\}$ ,  $d_i \in (0, \infty)$ , satisfying  $\|U_i\| \leq d_i$ ,  $i \geq 1$ , if  $r_u < \rho^{-1/J}$ , where  $r_u = \limsup_{n \rightarrow \infty} d_{n-1}/d_n$ , then

$$(2.25) \quad \limsup_{n \rightarrow \infty} d_n^{-1} \|V_n - V^*\| \leq \sigma(Lr_u, J)(1 - r_u^J \rho)^{-1}.$$

*Proof.* For case (i) fix  $k \geq 0$  in algorithm (2.18). Consider the series

$$(2.26) \quad \begin{aligned} \sum_{m=1}^{\infty} \rho^{-m} \|W_{mJ+k}^J\| &\leq \sum_{m=1}^{\infty} \rho^{-m} B_{mJ+k}^{J,L} \\ &\leq \max(L^J, 1) \rho^{-1} \sum_{i=1}^{\infty} (\rho^{1/J})^{-i} \|U_i\| \\ &< \infty, \end{aligned}$$

where the inequalities follow from (2.20) and condition (i). Then given (2.26) we may apply Lemma 2.7(i) to (2.18), giving

$$\limsup_{m \rightarrow \infty} \rho^{-m} \|V_{mJ+k} - V^*\| < \infty,$$

which then implies (2.23) after taking  $k = 0, \dots, J-1$ , noting that

$$(\rho^{1/J})^{-(Jm+k)} \|V_{mJ+k} - V^*\| = (\rho^{-k/J}) \rho^{-m} \|V_{mJ+k} - V^*\|.$$

For case (ii) fix  $k \geq 0$  in algorithm (2.18). Then by (2.20),

$$(2.27) \quad \begin{aligned} m^{-1} \log(\|W_{mJ+k}^J\|) &\leq m^{-1} \log(B_{mJ+k}^{J,L}) \\ &\leq m^{-1} \max_{(m-1)J+k+1 \leq n \leq mJ+k} (\log(\|U_n\|) + \log(J \max(L^J, 1))) \\ &\leq (J + km^{-1}) \left( \sup_{n \geq (m-1)J+k+1} n^{-1} \log(\|U_n\|) \right) \\ &\quad + m^{-1} \log(J \max(L^J, 1)), \quad m \geq 1. \end{aligned}$$

Taking the limit supremum of (2.27) with condition (ii) gives

$$\limsup_{m \rightarrow \infty} m^{-1} \log(\|W_{mJ+k}^J\|) \leq J \log(\rho^{1/J}) = \log(\rho).$$

This allows us to apply Lemma 2.7(ii) to (2.18), giving

$$\limsup_{m \rightarrow \infty} m^{-1} \log(\|V_{mJ+k} - V^*\|) \leq \log(\rho),$$

or equivalently,

$$\begin{aligned} \limsup_{m \rightarrow \infty} \frac{(J + km^{-1})}{(mJ + k)} \log(\|V_{mJ+k} - V^*\|) &= J \limsup_{m \rightarrow \infty} (mJ + k)^{-1} \log(\|V_{mJ+k} - V^*\|) \\ &\leq \log(\rho), \end{aligned}$$

which then implies (2.24) after taking  $k = 0, \dots, J-1$ .

For case (iii) fix  $k \geq 0$  in algorithm (2.18). From (2.20) we have

$$\|W_{mJ+k}^J\| \leq B_{mJ+k}^{J,L} \leq D_{mJ+k}^{J,L}, \quad m \geq 1,$$

where

$$D_n^{J,L} = \sum_{j=1}^J L^{J-j} d_{n-J+j}, \quad n \geq J, \text{ and } D_n^{J,L} = 0, \quad n < J.$$

Then

$$\begin{aligned} \sum_{m=1}^{\infty} \rho^{-m} D_{mJ+k}^{J,L} &\geq \min(1, L^J) \sum_{m=1}^{\infty} \rho^{-m} \sum_{j=1}^J d_{(m-1)J+k+j} \\ (2.28) \qquad \qquad \qquad &\geq \min(1, L^J) \rho \sum_{i=k}^{\infty} (\rho^{1/J})^{-i} d_i = \infty \end{aligned}$$

by applying condition (iii). Then for any  $\epsilon > 0$  there is  $N_\epsilon$  such that  $d_n \leq (r_u + \epsilon)d_{n+1}$  when  $n > N_\epsilon$ , and hence  $d_n \leq (r_u + \epsilon)^j d_{n+j}$ . Then, for all large enough  $m$ ,

$$\begin{aligned} D_{(m-1)J+k}^{J,L} &= \sum_{j=1}^J L^{J-j} d_{(m-2)J+k+j} \leq (r_u + \epsilon)^J \sum_{j=1}^J L^{J-j} d_{(m-1)J+k+j} \\ &= (r_u + \epsilon)^J D_{mJ+k}^{J,L} \end{aligned}$$

so that taking the limit supremum in  $m$  and letting  $\epsilon$  approach zero gives

$$(2.29) \qquad \qquad \qquad \limsup_{m \rightarrow \infty} \frac{D_{(m-1)J+k}^{J,L}}{D_{mJ+k}^{J,L}} = r_u^J.$$

We may then apply Lemma 2.7(iii) to (2.18) using bounding sequence  $D_{mJ+k}^{J,L}, m \geq 1$ , which with (2.28) and (2.29) yields

$$(2.30) \qquad \qquad \qquad \limsup_{m \rightarrow \infty} (D_{mJ+k}^{J,L})^{-1} \|V_{mJ+k} - V^*\| \leq (1 - r_u^J \rho)^{-1}.$$

Then for large enough  $m$ ,

$$\begin{aligned} D_{mJ+k}^{J,L} &= \sum_{j=1}^J L^{J-j} d_{(m-1)J+k+j} \leq \sum_{j=1}^J L^{J-j} (r_u + \epsilon)^{J-j} d_{mJ+k} \\ (2.31) \qquad \qquad &= d_{mJ+k} \sigma(L(r_u + \epsilon), J), \end{aligned}$$

which, with (2.30), letting  $\epsilon$  approach zero, gives

$$(2.32) \qquad \qquad \qquad \limsup_{m \rightarrow \infty} d_{mJ+k}^{-1} \|V_{mJ+k} - V^*\| \leq \sigma(r_u L, J) (1 - r_u^J \rho)^{-1}.$$

Then (2.25) follows by setting  $k = 0, \dots, J-1$ .  $\square$

**2.4. A general error model.** The structure of (2.2) suggests that a bound on  $\|U_n\|$  which directly depends on  $\|V_{n-1}\|$  may be available. This motivates a general error model, given as the following assumption:

(A4) In (2.2),

$$\|U_n\| \leq a_n + b_n \|V_{n-1}\|, \quad n \geq 1,$$



for some sequence of nonnegative finite constants  $\{a_n; n \geq 1\}$ ,  $\{b_n; n \geq 1\}$ .

It will be convenient to define, for a constant  $L$ , a sequence  $\{b_n; n \geq 1\}$  and integers  $n_2, n_1$ ,

$$(2.33) \quad \bar{L}_{n_1}^{n_2} = \begin{cases} \prod_{n=n_1}^{n_2} (L + b_n); & n_2 \geq n_1, \\ 1; & n_2 < n_1. \end{cases}$$

Theorem 2.11 can be used to establish the convergence of (2.2) given the error model in (A4). We will first need the following lemma.

LEMMA 2.13. *If (A4) holds for (2.2), in addition to (A1), and (A2) if  $J > 1$ , then*

$$(2.34) \quad \|V_{n+I} - V^*\| \leq \bar{L}_{n+1}^{n+I} \|V_n - V^*\| + \sum_{i=1}^I \bar{L}_{n+i+1}^{n+I} (a_{n+i} + b_{n+i} \|V^*\|)$$

for  $n \geq 1$ ,  $I \geq 1$ , with  $\bar{L}_{n_1}^{n_2}$  defined as in (2.33) for sequence  $\{b_n; n \geq 1\}$  and constant  $L$  satisfying  $\|TV - V^*\| \leq L\|V - V^*\|$ .

*Proof.* We proceed by induction. Fix  $n \geq 1$  and suppose (2.34) holds for  $I$ . Then directly from (2.2),

$$\|V_{n+I+1} - V^*\| \leq \|TV_{n+I} - V^*\| + \|U_{n+I+1}\|.$$

Applying (A4) gives

$$(2.35) \quad \begin{aligned} \|V_{n+I+1} - V^*\| &\leq L\|V_{n+I} - V^*\| + a_{n+I+1} + b_{n+I+1}\|V_{n+I}\| \\ &\leq (L + b_{n+I+1})\|V_{n+I} - V^*\| + a_{n+I+1} + b_{n+I+1}\|V^*\|; \end{aligned}$$

then applying the induction hypothesis to  $\|V_{n+I} - V^*\|$  gives

$$\begin{aligned} \|V_{n+I+1} - V^*\| &\leq (L + b_{n+I+1})\bar{L}_{n+1}^{n+I}\|V_n - V^*\| \\ &\quad + (L + b_{n+I+1}) \sum_{i=1}^I \bar{L}_{n+i+1}^{n+I} (a_{n+i} + b_{n+i}\|V^*\|) \\ &\quad + a_{n+I+1} + b_{n+I+1}\|V^*\| \\ &= \bar{L}_{n+1}^{n+I+1}\|V_n - V^*\| + \sum_{i=1}^{I+1} \bar{L}_{n+i+1}^{n+I+1} (a_{n+i} + b_{n+i}\|V^*\|) \end{aligned}$$

so that if (2.34) holds for  $I$ , it holds for  $I + 1$ . That (2.34) holds for  $I = 1$  follows from a simple relabelling of (2.35).  $\square$

Remark 2.2. In the proof of Lemma 2.13 the Lipschitz condition (A2) is used only in inequalities of the form  $\|TV - V^*\| \leq L\|V - V^*\|$ , (i.e., involving fixed point  $V^*$ ) so that (A2) may be replaced with a weaker condition.

We can now establish convergence properties of (2.2) under (A4).

THEOREM 2.14. *If (A1) and (A3) hold, with (A2) holding when  $J > 1$ , and if (A4) holds with  $a_n \rightarrow_n 0$  and  $b_n \rightarrow_n 0$ , then for algorithm (2.2)  $\|V_n - V^*\| \rightarrow_n 0$ .*

*Proof.* Suppose  $J = 1$ . Then

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{\|U_{n+1}\|}{\max(1, \|V_n\|)} &\leq \limsup_{n \rightarrow \infty} \frac{a_{n+1} + b_{n+1}\|V_n\|}{\max(1, \|V_n\|)} \\ &\leq \limsup_{n \rightarrow \infty} a_{n+1} + b_{n+1} = 0. \end{aligned}$$

Noting that  $B_n^{J,L} = \|U_n\|$  for  $J = 1$ , and setting  $s = 1$ , the theorem is proved directly from Theorem 2.11.

Now suppose  $J > 1$ . For  $I > 1$  we have, using (A4) and Lemma 2.13,

$$\begin{aligned}
 \|U_{n+I}\| &\leq a_{n+I} + b_{n+I}\|V_{n+I-1}\| \\
 &\leq a_{n+I} + b_{n+I}\|V^*\| + b_{n+I}\|V_{n+I-1} - V^*\| \\
 &\leq a_{n+I} + b_{n+I}\|V^*\| + b_{n+I}\bar{L}_{n+1}^{n+I-1}\|V_n - V^*\| \\
 (2.36) \quad &\quad + b_{n+I} \sum_{i=1}^{I-1} \bar{L}_{n+i+1}^{n+I-1} (a_{n+i} + b_{n+i}\|V^*\|).
 \end{aligned}$$

Additionally, (2.36) is also true for  $I = 1$  (directly by (A4) and the triangular inequality) if the final summation is interpreted as zero. Note that  $\lim_{n \rightarrow \infty} \bar{L}_{n+n_1}^{n+n_2} = L^{\max(0, n_2 - n_1 + 1)}$ . Examining the terms of the upper bound of (2.36) and dividing by  $\max(1, \|V_n\|)$  gives

$$\frac{\|U_{n+I}\|}{\max(1, \|V_n\|)} \leq a_{n+I} + b_{n+I}(K_1\|V^*\| + K_2)$$

for some finite constants  $K_1, K_2$  which do not depend on  $n$ ; hence,

$$(2.37) \quad \limsup_{n \rightarrow \infty} \frac{\|U_{n+I}\|}{\max(1, \|V_n\|)} = 0.$$

Then consider

$$(2.38) \quad \frac{\|B_{n+J}^{J,L}\|}{\max(1, \|V_n\|)} = \sum_{j=1}^J L^{J-j} \frac{\|U_{n+j}\|}{\max(1, \|V_n\|)}.$$

We may then use (2.37) to conclude that the limit supremum of each term on the right-hand side of (2.38) is zero. The theorem is proved following a direct application of Theorem 2.11.  $\square$

**2.5. Finite bounds.** The convergence results given above are based on asymptotic bounds. However, inequality (2.10) provides a means to calculate finite upper bounds on  $\|V_n - V^*\|$ . In this section we suppose we may bound errors with constants  $\|U_n\| \leq d_n$ , where we assume  $d_n > 0$ .

We first deal with the single stage contraction of modulus  $\rho$ . The first term in the upper bound is composed of the exponentially decreasing error of the exact algorithm. The second term may be bounded by

$$\sum_{i=1}^n \rho^{n-i} \|U_i\| \leq d_n I_n, \quad n \geq 1,$$

where  $I_n = \sum_{i=1}^n \rho^{n-i} d_i / d_n$ . If we assume  $r_u < \rho^{-1}$ , then from Lemma 2.3 we have

$$\limsup_{n \rightarrow \infty} I_n \leq (1 - r_u \rho)^{-1};$$

hence we rewrite (2.10) as

$$\|V_n - V^*\| \leq \|V_0 - V^*\| \rho^n + d_n I_n, \quad n \geq 1,$$

and consider the problem of usefully bounding  $I_n$ . If  $d_n$  has a tractable form, we may simply calculate  $I_n$  numerically. In this case, it will be useful to know something of the iterative properties of  $I_n$ . We show in the following lemma that, under a type of convexity assumption on the sequence  $d_n$ , once  $I_n$  decreases in  $n$ , it decreases indefinitely.

LEMMA 2.15. *If a positive sequence  $\{d_n; n \geq 1\}$  satisfies*

$$(2.39) \quad d_{n+1}/d_{n+2} \leq d_n/d_{n+1}, \quad n \geq 1,$$

*then  $I_{n+1} < I_n$  implies  $I_{n+2} < I_{n+1}$ .*

*Proof.* We may write

$$I_{n+1} - I_n = 1 + \left( \rho \frac{d_n}{d_{n+1}} - 1 \right) I_n, \quad n \geq 1,$$

from which it follows that

$$(2.40) \quad I_{n+1} - I_n < 0 \text{ if and only if } \left( 1 - \rho \frac{d_n}{d_{n+1}} \right) I_n > 1.$$

Then if  $I_{n+1} - I_n < 0$ , we have  $(1 - \rho \frac{d_n}{d_{n+1}}) > 0$  and hence  $(1 - \rho \frac{d_{n+1}}{d_{n+2}}) > 0$  from condition (2.39). This in turn implies by (2.40) that

$$\begin{aligned} \left( 1 - \rho \frac{d_{n+1}}{d_{n+2}} \right) I_{n+1} &= \left( 1 - \rho \frac{d_{n+1}}{d_{n+2}} \right) \left( 1 + \rho \frac{d_n}{d_{n+1}} I_n \right) \\ &> \left( 1 - \rho \frac{d_{n+1}}{d_{n+2}} \right) \left( 1 - \rho \frac{d_n}{d_{n+1}} \right)^{-1} \geq 1, \end{aligned}$$

which proves the lemma.  $\square$

*Remark 2.3.* Condition (2.39) is quite general and is satisfied by any polynomially decreasing bound  $1/n^k$ .

The same calculation can be made in the case of a general  $J$ -stage contraction of modulus  $\rho$ . Combining Lemma 2.9 with (2.10) gives

$$\|V_n - V^*\| \leq K \sum_{i=1}^n (\rho^{1/J})^{n-i} d_i + O\left((\rho^{1/J})^n\right), \quad n \geq 1,$$

for some finite constant  $K$  which does not depend on  $n$ . The finite bound may be treated in the same way as for the single stage contraction.

**2.6. Stochastic error terms.** The convergence results obtained in this section depend only on the properties of  $U_n$ , which can be defined in terms of the underlying seminormed linear space. If in (2.2), under (A1),  $U_n$  is stochastic, then almost sure convergence properties of  $\|V_n - V^*\|$  are deduced from those of  $\|U_n\|$ , and no other stochastic properties are needed.

To obtain  $L_p$  convergence or bounds, we need to regard  $\|U_n\|$  as a stochastic quantity. We start with the original seminormed linear space  $(\mathcal{V}, \|\cdot\|)$ , with contractive operator  $T$  as defined in (A1). We then define a new seminormed linear space based on  $(\mathcal{V}, \|\cdot\|)$ . Suppose we have a probability space  $(\Omega, \mathcal{F}, \mathcal{P})$  and can define a measurable space on  $\mathcal{V}$ . Let  $\mathcal{V}_\Omega$  be the linear space under the convolution operation of measurable mappings  $V : \Omega \rightarrow \mathcal{V}$ . Let  $\mathcal{U}$  be the space of random variables on  $(\Omega, \mathcal{F}, \mathcal{P})$ . Then for

$V \in \mathcal{V}_\Omega$  we have  $\|V\| \in \mathcal{U}$ . If  $\|\cdot\|_p$  is an  $L_p$  norm on random variables  $\mathcal{U}$ , we define a new seminorm  $\|V\|_\Omega = \|(\|V\|)\|_p$ . Then define the operators  $T_\Omega^I : \mathcal{V}_\Omega \rightarrow \mathcal{V}_\Omega$  by  $(T_\Omega^I V)(\omega) = T^I(V(\omega))$  for  $I \geq 1$ . Set  $V^*(\omega) \equiv v^*$ , where  $v^*$  is a fixed point of  $T$  in  $\mathcal{V}$ . Then  $(T_\Omega V^*)(\omega) = T(V^*(\omega)) \equiv T v^* = v^*$ , so that  $V^*$  is a fixed point of  $T_\Omega$ . For any  $V \in \mathcal{V}_\Omega$  we have  $\|T_\Omega^J V - V^*\|(\omega) = \|T^J(V(\omega)) - v^*\| \leq \rho \|V(\omega) - v^*\| = \rho \|V - V^*\|(\omega)$ . This leads to  $\|T_\Omega^J V - V^*\|_\Omega = \| \|T_\Omega^J V - V^*\| \|_p \leq \|\rho\| V - V^* \|_p = \rho \|V - V^*\|_\Omega$ , so that the seminormed linear space  $(\mathcal{V}_\Omega, \|\cdot\|_\Omega)$  also satisfies (A1) for  $J$ , modulus  $\rho$ , and operator  $T_\Omega$ .

To summarize, convergence with probability one (w.p.1) (or in  $L_p$ ) of  $\|V_n - V^*\|$  in (2.2) is a consequence of convergence w.p.1 (or in  $L_p$ ) of  $\|U_n\|$ .

**3. The dynamic programming operator for MDPs.** Following section 1.2, calculation of the optimal control of discrete time MDPs usually reduces to a VI algorithm based on the Bellman operator  $T_\pi$  associated with MCM  $\pi$ . In practice, exact knowledge of  $\pi$  is a special case, but it may be possible to construct an estimate  $\hat{\pi}$  of  $\pi$ , or a sequence of estimates  $\hat{\pi}_1, \hat{\pi}_2, \dots$ . In such a case, it is natural to consider substituting within a VI algorithm the estimates for the actual model  $\pi$ . The expectation would then be that the resulting value function solution would be close to the true one, to the degree that the model estimates are close to  $\pi$ . In this section we apply the theory of section 2 to this problem, with the objective of describing the behavior of approximate VI algorithms in terms of model estimates.

The general MDP considered is defined in section 3.1. It is important to emphasize that, because the theory is given largely in terms of the properties of normed linear spaces, the structure of the MDP can be left quite general. The essential condition is the existence of a well-behaved contractive VI algorithm coupled with a suitable norm. In the presence of unbounded costs, the weighted supremum norm may be used, which may be reduced to the standard supremum norm for bounded costs. See [19] for extensive discussion on this topic. We will employ this norm after some further discussion in section 3.1.

Section 3.2 will deal with two issues. A model distance is first defined. This distance incorporates the weight function used to define the weighted supremum norm. This permits a bound on the error due to the approximate evaluation of Bellman's operator to be given in terms of model distance.

We then note that one of the advantages of the general theory is that contractive properties need to be verified only for the exact operator  $T_\pi$ . However, it still needs to be verified that the approximate operators  $T_n$  are closed on the set of elements of finite norm, which will then imply assumption (A3). Theorem 3.5 below allows this question to be resolved on the basis of model distances. As a general rule, regularity conditions need only be verified for the exact model  $\pi$ . They follow for any model estimate  $\hat{\pi}$  as long as the model distance is finite.

Sections 3.3 and 3.4 then present the main results for, respectively, the AVIA and CEA forms of approximate VI as discussed in section 1.2.2.

**3.1. Model definition.** We adopt the following conventions. For any Borel space  $\mathcal{U}$ ,  $\mathcal{B}(\mathcal{U})$  is the class of Borel sets, and  $\mathcal{M}(\mathcal{U})$  is the set of probability measures on  $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$ . Then  $F(\mathcal{U})$  is the set of measurable real valued functions on  $\mathcal{U}$ . If  $\|\cdot\|$  is a seminorm defined on  $F(\mathcal{U})$ , then  $F(\mathcal{U}, \|\cdot\|) = \{V \in F(\mathcal{U}) : \|V\| < \infty\}$ . Let  $\sigma(H)$  denote the  $\sigma$ -algebra generated by a mapping  $H$  defined on a probability space.

A Markov decision process will be made of the following elements (see, for example, [18, 19]):

(M1) A Borel space  $\mathcal{X}$ . We refer to  $\mathcal{X}$  as the *state space*.

- (M2) A Borel space  $\mathcal{A}$ . We refer to  $\mathcal{A}$  as the *action space*.
- (M3) With each  $x \in \mathcal{X}$  associate  $\mathcal{K}_x \in \mathcal{B}(\mathcal{A})$  with  $\mathcal{K}_x \neq \emptyset$ . The *state-action space*  $\mathcal{K} = \{(x, a) \in \mathcal{X} \times \mathcal{A} : a \in \mathcal{K}_x\}$  is assumed to be a measurable subset of  $\mathcal{X} \times \mathcal{A}$ .
- (M4) A *transition kernel*  $Q : \mathcal{K} \rightarrow \mathcal{M}(\mathcal{X})$ . We assume for each fixed  $E_x \in \mathcal{B}(\mathcal{X})$  that  $Q(E_x \mid \cdot)$  is a measurable function from  $\mathcal{K}$  to  $\mathfrak{R}$ . Furthermore, we assume there is a measure  $\mu_Q$  on  $\mathcal{X}$  with respect to which  $Q(\cdot \mid x, a)$  has a density  $f(\cdot \mid x, a)$  for each  $(x, a) \in \mathcal{K}$ .
- (M5) A measurable mapping  $R : \mathcal{K} \rightarrow (-\infty, \infty)$ , referred to as the *cost function*.
- (M6) A *discount factor*  $\beta \geq 0$ .

A reference to state-action space  $\mathcal{K}$  will implicitly include  $(\mathcal{X}, \mathcal{A})$ . A reference to the stochastic kernel  $Q$ , defined on  $\mathcal{K}$ , will implicitly include  $\mu_Q$ . An MCM will be the object  $\pi = (\mathcal{K}, Q, R, \beta)$ . Let  $\mathcal{K}^f$  be the set of all measurable mappings  $f : \mathcal{X} \rightarrow \mathcal{A}$  for which  $f(x) \in \mathcal{K}_x$  for all  $x \in \mathcal{X}$ . Assume  $\mathcal{K}^f$  is not empty.

An MDP will consist of an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  coupled with a control policy.

**DEFINITION 3.1.** A control policy consists of a sequence of measurable mappings  $\Phi = \{\Phi_n, n \geq 1\}$  of the form  $\Phi_n : (\mathcal{K})^{n-1} \times \mathcal{X} \rightarrow \mathcal{M}(\mathcal{A})$ . We assume  $\Phi_n(\mathcal{K}_{x_n} \mid x_1, a_1, \dots, x_{n-1}, a_{n-1}, x_n) = 1$  for  $n \geq 1$ .

Intuitively, the MDP is realized as a random process  $(X_1, A_1, X_2, A_2, \dots)$  on the Borel space  $\mathcal{K}^\infty$ . Stage  $n$  is taken to refer to  $(X_n, A_n)$ , and the history at stage  $n$  up to state  $X_n$  is denoted  $H_n^x = (X_1, A_1, \dots, X_n) \in \mathcal{K}^{n-1} \times \mathcal{X}$ . Similarly, the history at stage  $n$  up to action  $A_n$  is denoted  $H_n^a = (X_1, A_1, \dots, X_n, A_n) \in \mathcal{K}^n$ . The stage  $n$  cost is taken to be  $R(X_n, A_n)$ . It has been shown (see [17]) that for each  $x \in \mathcal{X}$  there exists a unique probability measure  $P_x^\Phi$  on  $\mathcal{K}^\infty$  which satisfies

$$\begin{aligned} P_x^\Phi(X_1 = x) &= 1, \\ P_x^\Phi(A_n \in E_a \mid H_n^x) &= \Phi_n(E_a \mid H_n^x) \quad \forall E_a \in \mathcal{B}(\mathcal{A}), H_n^x \in \mathcal{K}^{n-1} \times \mathcal{X}, \\ (3.1) \quad P_x^\Phi(X_{n+1} \in E_x \mid H_n^a) &= Q(E_x \mid X_n, A_n) \quad \forall E_x \in \mathcal{B}(\mathcal{X}), H_n^a \in \mathcal{K}^n, \end{aligned}$$

for  $n \geq 1$  and each admissible history  $H_n^x, H_n^a$ . The criterion for choosing a policy  $\Phi$  is given as a  $\beta$ -discounted cost from initial point  $X_1 = x$ ,

$$V^\Phi(x) = E_x^\Phi \left[ \sum_{n=1}^{\infty} \beta^{n-1} R(X_n, A_n) \right], \quad x \in \mathcal{X},$$

where  $E_x^\Phi$  is the expectation operator of  $P_x^\Phi$ . In this case interest is in control processes, which assume finite total expected cost. The *value function* is

$$V^*(x) = \inf_{\Phi} V^\Phi(x), \quad x \in \mathcal{X},$$

and any policy  $\Phi^*$  satisfying  $V^{\Phi^*}(x) = V^*(x)$  for all  $x \in \mathcal{X}$  is the minimum  $\beta$ -discounted cost policy. Of special interest will be deterministic policies.

**DEFINITION 3.2.** Given a sequence of policy functions  $\tilde{\phi} = \{\phi_n \in \mathcal{K}^f; n \geq 1\}$ , where  $\phi_n \equiv \phi_n(x \mid H_n^x)$  may depend on  $H_n^x$ ,  $\Phi$  is a deterministic policy based on  $\tilde{\phi}$  if  $\Phi_n(E_a \mid H_n^x) = I\{\phi_n(X_n \mid H_n^x) \in E_a\}$  for all  $n \geq 1$ . If  $\phi_n(\cdot \mid H_n^x) \equiv \phi(\cdot)$  for some single  $\phi \in \mathcal{K}^f$ , then  $\Phi$  is a stationary deterministic policy.

For a stochastic kernel  $Q$  on  $\mathcal{K}$  let  $\mathcal{L}_Q$  be the set of functions  $V \in F(\mathcal{X})$  for which the integral  $\int_{y \in \mathcal{X}} V(y) dQ(y \mid x, a)$  exists and is finite for all  $(x, a) \in \mathcal{K}$ . The set of *weight functions*  $\mathcal{W}(\mathcal{X}) \subset F(\mathcal{X})$  is taken to be the set of all measurable functions

$w : \mathcal{X} \rightarrow (0, \infty)$ . For any weight function  $w$  we define the weighted supremum norm  $\|\cdot\|_w$  on  $F(\mathcal{X})$  by  $\|V\|_w = \sup_{x \in \mathcal{X}} w^{-1}(x)|V(x)|$ . Then define quantities

$$\eta_Q^w(x, a) = w(x)^{-1} \int_{y \in \mathcal{X}} w(y) dQ(y | x, a), \quad (x, a) \in \mathcal{K},$$

$$\eta_Q^w = \sup_{(x, a) \in \mathcal{K}} \eta_Q^w(x, a).$$

We next define for an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  the Bellman operator, or dynamic programming operator (DPO),  $T_\pi : \mathcal{L}_Q \rightarrow F(\mathcal{X})$ ,

$$(3.2) \quad T_\pi V(x) = \inf_{a \in \mathcal{K}_x} R(x, a) + \beta \int_{y \in \mathcal{X}} V(y) dQ(y | x, a).$$

Removing the infimum operation in (3.2) gives the quantity

$$(3.3) \quad T_\pi^a V(x, a) = R(x, a) + \beta \int_{y \in \mathcal{X}} V(y) dQ(y | x, a),$$

also defined for all  $V \in \mathcal{L}_Q$ . Let  $V_\pi^*$  denote the value function for  $\pi$ .

This leads to the definition of a solution space for an MCM.

**DEFINITION 3.3.** A solution space for MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  consists of vector space  $\mathcal{V} \subset F(\mathcal{X})$  and a weight function  $w$  which defines norm  $\|\cdot\|_w$ , together denoted  $\mathcal{V}^w = (\mathcal{V}, w)$ , which satisfies

- (i)  $\mathcal{V} \subset \mathcal{L}_Q$ ,
- (ii)  $T_\pi \mathcal{V} \subset \mathcal{V}$  for DPO  $T_\pi$ , and
- (iii)  $V_\pi^* \in \mathcal{V}$ .

Additionally,  $\mathcal{V}^w$  is a contractive solution space if  $T_\pi V_\pi^* = V_\pi^*$  and if  $(\mathcal{V}, \|\cdot\|_w, T_\pi, V_\pi^*, \rho, J)$  satisfies (A1) and (A2) if  $J > 1$ .

**Remark 3.1.** We may have  $\rho \neq \beta$ . This will generally be the case for unbounded costs. See Lippman [27], Van Nunen and Wessels [35] or Hernández-Lerma and Lasserre [19].

The solution to the MCM control problem is ideally of the following form.

**DEFINITION 3.4.** MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  possesses an optimal stationary deterministic policy (OSDP) if

- (i) there exists  $\phi_\pi^* \in \mathcal{K}^f$  such that

$$(3.4) \quad \phi_\pi^*(x) = \arg \min_{a \in \mathcal{K}_x} R(x, a) + \beta \int_{y \in \mathcal{X}} V_\pi^*(y) dQ(y | x, a);$$

- (ii) the stationary deterministic policy implied by  $\phi_\pi^*$  is a minimum  $\beta$ -discounted cost policy.

The existence of a solution space and an OSDP can be guaranteed under a wide variety of conditions, which are discussed in detail in, for example, Bertsekas and Shreve [4] and Hernández-Lerma and Lasserre [18, 19]. Typically  $R(x, a)$  is assumed to be nonnegative and lower semicontinuous, and the action space projection  $\mathcal{K}_x$  is assumed to possess some compactness property. Additionally, a continuity condition on the kernel  $Q$  on  $\mathcal{K}$  is usually imposed. The contraction property for  $T$  is easy to verify when  $\beta < 1$  and  $R$  is bounded, in which case the unweighted supremum norm (i.e.,  $w \equiv 1$ ) suffices. On the other hand, if  $R$  is unbounded, the supremum norm is usually unsuitable, but may be replaced with a weighted supremum norm

with respect to which  $T$  is contractive. Conditions under which this is the case are given in [27], and in fact these conditions also imply (A2). Informally, the weight function  $w$  is required to bound  $R$  in some sense. See also Van Nunen and Wessels [35], Bhattacharya and Majumdar [6], and Hernández-Lerma and Lesserre [19].

For  $\beta \geq 1$ , it will generally be required that the process terminate after some finite time. In this case a weighted supremum norm which induces contractivity may exist. Some conditions are given in Bertsekas [3]. In general, if the process may terminate at any stage with a probability bounded away from zero, then it is easily established that  $T$  is contractive in the unweighted supremum norm under bounded costs (see Almedevar [1]).

It is important to note that there will be no advantage in specifying any single set of regularity conditions on the model elements (M1)–(M6). The existence of a contractive solution space and an OSDP will be our point of departure.

**3.2. Approximate evaluation of the DPO.** If MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  possesses a contractive solution space  $\mathcal{V}^w$ , we may define a VI algorithm to be a sequence of the form

$$(3.5) \quad \begin{aligned} V_0 &= v_0 \in \mathcal{V}, \\ V_n &= T_\pi V_{n-1}, \quad n \geq 1, \end{aligned}$$

for which we may easily conclude  $\|V_n - V_\pi^*\|_w \rightarrow_n 0$  at a geometric rate. If the model pair  $(Q, R)$  is unknown, but we have an estimate  $\hat{\pi} = (\mathcal{K}, \hat{Q}, \hat{R}, \beta)$ , we may construct an approximate DPO  $T_{\hat{\pi}}$  to use in place of  $T_\pi$ . In this case we will need to define some notion of distance between two models. Additionally, we need to establish that the approximate DPO possesses a suitable range and domain, and that any evaluation  $T_{\hat{\pi}}V$  is close to  $T_\pi V$ , with the evaluation error expressed naturally in terms of model distance.

Recall that the definition of a stochastic kernel in (M4) includes a notion of a density on a common measure over state-action space  $\mathcal{K}$ . Distances between probability measures will ultimately be derived from these densities, so stochastic kernels must be absolutely continuous with respect to some common measure to be comparable. Accordingly, we give the following definition.

**DEFINITION 3.5.** *Two stochastic kernels  $Q_1$  and  $Q_2$  of type (M4) defined on a common state-action space conform if  $\mu_{Q_1} = \mu_{Q_2}$ .*

The distance between two conforming stochastic kernels on  $\mathcal{K}$  with weight function  $w$  will be defined as

$$D_q^w(Q_1, Q_2) = \sup_{(x,a) \in \mathcal{K}} w^{-1}(x) \int_{y \in \mathcal{X}} w(y) |f_1(y | x, a) - f_2(y | x, a)| d\mu_Q(y),$$

where  $f_1(y | x, a)$ ,  $f_2(y | x, a)$  are the densities of  $Q_1(\cdot | x, a)$  and  $Q_2(\cdot | x, a)$  with respect to common measure  $\mu_Q$ . Similarly, the distance between two cost functions  $R_1, R_2$  of type (M5) will be defined as

$$D_r^w(R_1, R_2) = \sup_{(x,a) \in \mathcal{K}} w(x)^{-1} |R_1(x, a) - R_2(x, a)|.$$

If the model distances  $D_r^w(R, \hat{R}), D_q^w(Q, \hat{Q})$  between actual and approximate models  $\pi = (\mathcal{K}, Q, R, \beta)$  and  $\hat{\pi} = (\mathcal{K}, \hat{Q}, \hat{R}, \beta)$  can be bounded, then it suffices to verify contractive, Lipschitz, or ergodic properties for  $\pi$  alone. This idea is developed in the following four lemmas and Theorem 3.5.

LEMMA 3.1. Suppose for stochastic kernel  $Q$  on  $\mathcal{K}$  and weight function  $w \in \mathcal{W}(\mathcal{X})$  we have  $\eta_Q^w(x, a) < \infty$  for all  $(x, a) \in \mathcal{K}$ . Then  $F(\mathcal{X}, \|\cdot\|_w) \subset \mathcal{L}_Q$ .

*Proof.* Note that  $|V(y)| \leq \|V\|_w w(y)$  for each  $y \in \mathcal{X}$ . For  $V \in F(\mathcal{X}, \|\cdot\|_w)$ ,  $\|V\|_w < \infty$ . By hypothesis,  $\|V\|_w w \in \mathcal{L}_Q$ , and hence so is  $V$ .  $\square$

LEMMA 3.2. Suppose we have some MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  and weight function  $w \in \mathcal{W}(\mathcal{X})$  such that  $\eta_Q^w(x, a) < \infty$  for all  $(x, a) \in \mathcal{K}$ . If  $\hat{Q}$  is a stochastic kernel on  $\mathcal{K}$  which conforms to  $Q$ , and if  $D_q^w(Q, \hat{Q}) < \infty$ , then

- (i)  $\eta_{\hat{Q}}^w(x, a) \leq \eta_Q^w(x, a) + D_q^w(Q, \hat{Q}) < \infty$  for all  $(x, a) \in \mathcal{K}$ ,
- (ii)  $\eta_{\hat{Q}}^w \leq \eta_Q^w + D_q^w(Q, \hat{Q})$ , and
- (iii)  $F(\mathcal{X}, \|\cdot\|_w) \subset \mathcal{L}_{\hat{Q}}$ .

*Proof.* Denote the density kernels for  $Q, \hat{Q}$  by  $f, \hat{f}$ . We have

$$(3.6) \quad w(y)\hat{f}(y | x, a) \leq w(y)f(y | x, a) + w(y)|f(y | x, a) - \hat{f}(y | x, a)|, \quad y \in \mathcal{X}.$$

Integrating with respect to  $y$  over measure  $\mu_Q$ , multiplying by  $w^{-1}(x)$ , then taking the supremum of the final term over  $\mathcal{K}$  gives (i), from which (ii) follows by taking the supremum of the remaining terms over  $\mathcal{K}$ . Then (iii) follows by applying Lemma 3.1.  $\square$

LEMMA 3.3. Let  $f_1, f_2$  be real valued functions on a set  $E$ . Suppose  $\inf_x f_2(x) > -\infty$  and  $|f_2(x)| < \infty$  for all  $x \in E$ . Then  $|\inf_x f_1(x) - \inf_x f_2(x)| \leq \sup_x |f_1(x) - f_2(x)|$ .

*Proof.* First, note that since  $f_2$  is everywhere finite,  $\sup_x |f_1(x) - f_2(x)|$  is well defined. If  $|\inf_x f_1(x)| = \infty$ , it follows that  $\sup_x |f_1(x) - f_2(x)| = \infty$ . Then suppose  $\inf_x f_1(x)$  is finite. For any  $\epsilon > 0$  there exists  $x^* \in E$  such that  $f_2(x^*) \leq \inf_x f_2(x) + \epsilon$ . Then

$$\inf_x f_1(x) - \inf_x f_2(x) \leq f_1(x^*) - f_2(x^*) + \epsilon \leq \sup_x |f_1(x) - f_2(x)| + \epsilon.$$

A similar argument gives  $\inf_x f_2(x) - \inf_x f_1(x) \leq \sup_x |f_1(x) - f_2(x)| + \epsilon$ . The lemma follows by letting  $\epsilon$  approach 0.  $\square$

Remark 3.2. This lemma is similar to one in Hinderer [20, p. 17].

LEMMA 3.4. Suppose we have an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  with solution space  $\mathcal{V}^w$ , and an MCM  $\hat{\pi} = (\mathcal{K}, \hat{Q}, \hat{R}, \beta)$  for which  $\hat{Q}$  conforms to  $Q$ . Furthermore, suppose  $T_\pi : \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w) \rightarrow \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$ . Then

$$(3.7) \quad \|T_{\hat{\pi}}V - T_\pi V\|_w \leq D_r^w(R, \hat{R}) + \beta\|V\|_w D_q^w(Q, \hat{Q})$$

for any  $V \in \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w) \cap \mathcal{L}_{\hat{Q}}$ .

*Proof.* Denote the density kernels for  $Q, \hat{Q}$  by  $f, \hat{f}$ . By assumption,  $T_\pi V(x)$  is finite and  $V \in \mathcal{L}_Q$ , so for fixed  $x$ ,  $T_\pi^a V(x, a)$  as a function of  $a$  satisfies the conditions imposed on  $f_2$  in Lemma 3.3, so that

$$\begin{aligned} & w(x)^{-1} |T_{\hat{\pi}}V(x) - T_\pi V(x)| \\ & \leq \sup_{a \in \mathcal{K}_x} w(x)^{-1} \left| \hat{R}(x, a) - R(x, a) \right| \\ & \quad + \sup_{a \in \mathcal{K}_x} w(x)^{-1} \left| \beta \int_{y \in \mathcal{X}} w(y)^{-1} V(y) w(y) \left( \hat{f}(y|x, a) - f(y|x, a) \right) d\mu_Q(y) \right| \\ & \leq D_r^w(R, \hat{R}) + \beta\|V\|_w D_q^w(Q, \hat{Q}). \end{aligned}$$



Then taking the supremum over  $\mathcal{X}$  gives (3.7).  $\square$

The required properties of an approximate DPO are established in the next theorem.

**THEOREM 3.5.** *Suppose we have an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  with solution space  $\mathcal{V}^w$ , and an MCM  $\hat{\pi} = (\mathcal{K}, \hat{Q}, \hat{R}, \beta)$  for which  $\hat{Q}$  conforms to  $Q$ . Furthermore, suppose*

- (a)  $\eta_Q^w(x, a) < \infty$ ,
- (b)  $T_\pi : \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w) \rightarrow \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$ ,
- (c)  $D_q^w(Q, \hat{Q})$ ,  $D_r^w(R, \hat{R})$  are finite, and
- (d)  $T_{\hat{\pi}} : \mathcal{V} \cap \mathcal{L}_{\hat{Q}} \rightarrow \mathcal{V}$ .

Then  $T_{\hat{\pi}} : \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w) \rightarrow \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$ .

*Proof.* By Lemma 3.2, noting (a) and (c) we have  $F(\mathcal{X}, \|\cdot\|_w) \subset \mathcal{L}_{\hat{Q}}$ , and hence  $T_{\hat{\pi}} : \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w) \rightarrow \mathcal{V}$ . Then suppose  $V \in \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$ . By Lemma 3.4  $\|T_{\hat{\pi}}V - T_\pi V\|_w < \infty$ . Condition (b) implies  $\|T_\pi V\|_w < \infty$ , hence  $\|T_{\hat{\pi}}V\|_w < \infty$ , which completes the proof.  $\square$

**Remark 3.3.** If  $\mathcal{V}^w$  is a contractive solution space, then condition (b) is easily verified.

**3.3. AVIAs.** We are now in a position to investigate the convergence properties of iterative algorithms based on approximate DPOs. We begin with the following definition.

**DEFINITION 3.6.** *Given an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  with solution space  $\mathcal{V}^w$ , and a sequence of MCMs  $\hat{\pi}_n = (\mathcal{K}, \hat{Q}_n, \hat{R}_n, \beta)$  for which  $(\pi, \mathcal{V}^w)$  and  $\hat{\pi}_n$  satisfy the conditions of Theorem 3.5 for each  $n \geq 1$ , an AVIA associates with each  $v_0 \in \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$  a sequence*

$$\begin{aligned} V_0 &= v_0, \\ V_n &= T_{\hat{\pi}_n} V_{n-1}, \quad n \geq 1. \end{aligned}$$

**Remark 3.4.** Theorem 3.5 ensures that the sequence of an AVIA is well defined and exists in  $\mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$ .

This is the type of algorithm considered in [11] for finite state-action spaces. See also [17] for a more general model with bounded costs. In that work an AVIA is referred to as a type of *nonstationary VI*.

The following theorem summarizes the convergence properties of an AVIA and is a direct application of Theorems 2.12 and 2.14. The notation conforms by setting  $U_n = T_{\hat{\pi}_n} V_{n-1} - T_\pi V_{n-1}$ .

**THEOREM 3.6.** *Suppose we have some MCM  $(\mathcal{K}, Q, R, \beta)$  with contractive solution space  $\mathcal{V}^w$  with an AVIA (Definition 3.6). Suppose there exist two sequences of positive constants  $a_n, b_n$ ,  $n \geq 1$ , such that*

$$\begin{aligned} D_r^w(R, \hat{R}_n) &\leq a_n, \\ D_q^w(Q, \hat{Q}_n) &\leq b_n, \quad n \geq 1, \end{aligned}$$

with  $\lim_{n \rightarrow \infty} a_n = 0$  and  $\lim_{n \rightarrow \infty} b_n = 0$ . Then

$$(3.8) \quad \lim_{n \rightarrow \infty} \|V_n - V_\pi^*\|_w = 0.$$

Additionally, define  $d_n = a_n + b_n \beta \|V_{n-1}\|$ ,  $n \geq 1$ . Then

- (i) if  $\sum_{n=1}^\infty (\rho^{1/J})^{-i} d_n < \infty$ , then

$$(3.9) \quad \limsup_{n \rightarrow \infty} (\rho^{1/J})^{-n} \|V_n - V_\pi^*\|_w < \infty;$$

(ii) if  $\limsup_{n \rightarrow \infty} n^{-1} \log(d_n) \leq \log(\rho^{1/J})$ , then

$$(3.10) \quad \limsup_{n \rightarrow \infty} n^{-1} \log(\|V_n - V_\pi^*\|_w) \leq \log(\rho^{1/J});$$

(iii) if  $r_u = \limsup_{n \rightarrow \infty} d_{n-1}/d_n < \rho^{-1/J}$ , then

$$(3.11) \quad \limsup_{n \rightarrow \infty} d_n^{-1} \|V_n - V_\pi^*\|_w < (1 - r_u^J \rho)^{-1}.$$

*Proof.* Assumption (A3) follows from Definition 3.6 and the contractive property of  $T_\pi$ . By Lemma 3.4 we directly conclude that assumption (A4) holds. Since  $D_q^w(Q, \hat{Q}_n)$  and  $D_q^w(Q, \hat{Q}_n)$  converge to zero, (3.8) follows directly from Theorem 2.14. Then by Lemma 3.4  $\|U_n\|_w \leq d_n$ , and (3.9)–(3.11) follow by applying Theorem 2.12.  $\square$

*Remark 3.5.* If  $\|V_\pi^*\| > 0$ , we may replace  $d_n$  in Theorem 3.6 with  $d'_n = a_n + b_n \beta \|V_\pi^*\|$ . In this case there exists finite  $K$  such that  $d'_n \leq K d_n$  with  $\lim_{n \rightarrow \infty} d_n/d'_n = 1$ . It is easily shown that the conditions and results for cases (i), (ii), and (iii) hold for  $d_n$  if and only if they hold for  $d'_n$ .

*Remark 3.6.* If  $D_q^w(Q, \hat{Q}_n)$  or  $D_r^w(R, \hat{R}_n)$  do not converge to zero but can be bounded, and if  $\|V_n\|_w$  can be bounded (assuming, for example, that  $R(x, a)$  can be bounded), Lemma 3.4 and Theorem 2.10 can be used directly to bound  $\limsup_{n \rightarrow \infty} \|V_n - V_\pi^*\|_w$ .

A different bound for an AVIA is obtained in [17] for the discounted ( $\beta < 1$ ) model with cost bounded by  $R^b < \infty$  and supremum norm (see also [11]) in addition to some other model regularity conditions. Using the notation of Theorem 3.6 the reported bound (obtained from the proof in Theorem 4.8, Chapter 2 of [17]) is

$$(3.12) \quad \|V_n - V_\pi^*\|_w \leq \frac{a_{[n/2]} + \beta R^b (1 - \beta)^{-1} b_{[n/2]}}{1 - \beta} + \frac{2R^b \beta^{[n/2]}}{1 - \beta}, \quad n \geq 1,$$

where  $[t]$  is the largest integer less than or equal to  $t$ . The asymptotic bound from Theorem 3.6 is

$$(3.13) \quad \|V_n - V_\pi^*\|_w \leq \frac{a_n + \beta \|V_\pi^*\|_w b_n}{1 - \beta} + o(a_n + b_n), \quad n \geq 1,$$

when  $a_n$  and  $b_n$  decrease nongeometrically (i.e.,  $r_u = 1$  in Theorem 3.6). If we then follow [17] by bounding  $\|V_\pi^*\|_w$  with  $R^b(1 - \beta)^{-1}$ , the asymptotic bound in (3.12) is obtained by replacing  $(a_n, b_n)$  in (3.13) with  $(a_{[n/2]}, b_{[n/2]})$ . This will result in a faster rate of convergence for (3.13). Additionally, this implies that the finite bound obtained in section 2.5 will be uniformly smaller than (3.12) for all large enough  $n$ .

**3.4. Certainty-equivalence methods.** We have not assumed that the approximate DPOs of Definition 3.6 are contractive. If we may verify that repeated iteration of  $T_{\hat{\pi}_n}$  in a suitable starting point for any  $n$  results in some limit  $V_{\hat{\pi}_n}^*$ , we may wish to investigate whether  $V_{\hat{\pi}_n}^*$  converges to  $V_\pi^*$ . Such an algorithm differs from the AVIA, and we reserve the term *certainty-equivalence* for this approach. In effect, the MDP is controlled as though the current model estimate is correct. See [24] or [17] for further discussion.

**DEFINITION 3.7.** Suppose we are given an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  with solution space  $\mathcal{V}^w$ , and a sequence of MCMs  $\hat{\pi}_n = (\mathcal{K}, \hat{Q}_n, \hat{R}_n, \beta)$  for which  $(\pi, \mathcal{V}^w)$  and  $\hat{\pi}_n$  satisfy the conditions of Theorem 3.5 for each  $n \geq 1$ . Suppose there exists a sequence

$v_{0,n} \in \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$ ,  $n \geq 1$ , such that for each  $n$  we have  $\lim_{i \rightarrow \infty} \|T_{\hat{\pi}_n}^i v_{0,n} - V_{\hat{\pi}_n}^*\|_w = 0$  for some  $V_{\hat{\pi}_n}^* \in \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$ . Then a CEA is defined as an algorithm which calculates the sequence  $(V_{\hat{\pi}_1}^*, V_{\hat{\pi}_2}^*, \dots)$ .

Conditions for the convergence of a CEA are given in the next theorem.

**THEOREM 3.7.** *Suppose for a CEA (Definition 3.7)  $\mathcal{V}^w$  is a contractive solution space. Then for  $n \geq 1$ ,*

$$(3.14) \quad \|V_{\hat{\pi}_n}^* - V_{\pi}^*\|_w \leq \frac{\sigma(L, J)}{(1 - \rho)} (D_r^w(R, \hat{R}_n) + \beta \|V_{\hat{\pi}_n}^*\|_w D_q^w(Q, \hat{Q}_n)).$$

If in addition  $D_r^w(R, \hat{R}_n) \rightarrow_n 0$  and  $D_q^w(Q, \hat{Q}_n) \rightarrow_n 0$ , then  $\sup_n \|V_{\hat{\pi}_n}^*\|_w < \infty$ , and hence  $\|V_{\hat{\pi}_n}^* - V_{\pi}^*\|_w \rightarrow_n 0$ .

*Proof.* Fix  $n \geq 1$  and apply the VI algorithm for operator  $T_{\hat{\pi}_n}$  with starting value  $v_{0,n}$ ; that is, define algorithm

$$\begin{aligned} V_0 &= v_{0,n}, \\ V_i &= T_{\hat{\pi}_n} V_{i-1}, \quad i \geq 1. \end{aligned}$$

This is an AVIA satisfying Definition 3.6 with  $(\hat{R}_i, \hat{Q}_i) \equiv (R, \hat{Q}_n)$ . The conditions of Lemma 3.4 apply, giving

$$\|T_{\hat{\pi}_n} V_{i-1} - T_{\pi} V_{i-1}\| \leq D_r^w(R, \hat{R}_n) + \beta \|V_{i-1}\| D_q^w(Q, \hat{Q}_n), \quad i \geq 1.$$

Using Theorem 2.10, and noting (2.31) in the proof of Theorem 2.12, we obtain

$$\begin{aligned} \|V_{\hat{\pi}_n}^* - V_{\pi}^*\|_w &= \limsup_{i \rightarrow \infty} \|V_i - V^*\| \\ &\leq \frac{\sigma(L, J)}{(1 - \rho)} \limsup_{i \rightarrow \infty} D_r^w(R, \hat{R}_n) + \beta \|V_i\| D_q^w(Q, \hat{Q}_n), \end{aligned}$$

where  $L = 1$  for  $J = 1$ . Under the assumption  $\|V_i - V_{\hat{\pi}_n}^*\|_w \rightarrow_i 0$  the above inequality evaluates to (3.14).

Then assume  $D_r^w(R, \hat{R}_n) \rightarrow_n 0$  and  $D_q^w(Q, \hat{Q}_n) \rightarrow_n 0$ . Suppose there is an increasing sequence of integers  $n_1, n_2, \dots$  such that  $\|V_{\hat{\pi}_{n_i}}^*\|_w$  is increasing and unbounded as  $i \rightarrow \infty$ . For any  $\epsilon > 0$  we have  $I_\epsilon$  such that  $i > I_\epsilon$  implies

$$(3.15) \quad \frac{\|V_{\hat{\pi}_{n_i}}^* - V_{\pi}^*\|_w}{\|V_{\hat{\pi}_{n_i}}^*\|_w} \geq \frac{\|V_{\hat{\pi}_{n_i}}^*\|_w - \|V_{\pi}^*\|_w}{\|V_{\hat{\pi}_{n_i}}^*\|_w} \geq 1 - \epsilon$$

after applying the triangle inequality. Divide (3.14) for  $n = n_i$  by  $\|V_{\hat{\pi}_{n_i}}^*\|_w$ . As  $i \rightarrow \infty$ , the upper bound of the resulting inequality approaches 0, which combined with (3.15) leads to a contradiction. Hence  $\|V_{\hat{\pi}_n}^*\|_w$  must remain bounded. The boundedness of  $\|V_{\hat{\pi}_n}^*\|_w$  then implies  $\|V_{\hat{\pi}_n}^* - V_{\pi}^*\|_w \rightarrow_n 0$ .  $\square$

Clearly, the AVIA has the advantage that only a single VI algorithm is involved while achieving the same convergence rate. On the other hand, if an efficient specialized algorithm for calculating  $V_{\hat{\pi}_n}^*$  is available, a CEA may be preferable. Of course, intermediate options are available. We may modify an AVIA so that some number of VIs greater than one are performed using  $T_{\hat{\pi}_n}$  once  $(\hat{Q}_n, \hat{R}_n)$  is available. In this case, let  $j_n$  be the highest model index available for the approximate operator  $T_{\hat{\pi}_n}$ . The convergence rate is given in terms of  $(\hat{Q}_{j_n}, \hat{R}_{j_n})$  as given in Theorem 3.6, with the results remaining otherwise directly applicable.

**4. Value functions for approximate optimal policies.** The methods of section 3 are used to estimate the optimal achievable cost of an MDP but do not explicitly describe the properties of any particular control. Of course, the process of evaluating an approximate DPO on value function  $V$  yields a control function  $\phi(x) = \arg \min_{a \in \mathcal{K}_x} T_\pi^a V(x, a)$ , and we would expect that if  $\hat{\pi}$  is close to the true model  $\pi$ , and if  $V$  is close to the value function  $V_\pi^*$ , then  $\phi$  would be close to the optimal control function  $\phi_\pi^*$  in some sense. The purpose of this section is to formalize this idea.

Given an MDP  $(\mathcal{K}, Q, R, \beta, \Phi)$  with measure (3.1), construct, when well defined, the sequence

$$(4.1) \quad \Lambda_n^\Phi(H_n^x) = E_x^\Phi \left[ \sum_{i=n}^{\infty} \beta^{i-n} R(X_n, A_n) \mid H_n^x \right], \quad n \geq 1,$$

which can be interpreted as the expected remaining discounted cost calculated from stage  $n$  given history  $H_n^x$ . We wish to show that  $\Lambda_n^\Phi(H_n^x)$  exists, is finite, and is bounded from below by  $V_\pi^*(X_n)$ . This gives us a method of evaluating directly the consequence of using a suboptimal control and allows us to develop some notion of the convergence of a control to optimality. This approach is similar to that used in [30], in which the deviation from optimality at stage  $n$  is taken to be  $|E_x^\Phi[\Lambda_n^\Phi(H_n^x)] - E_x^\Phi[V_\pi^*(X_n)]|$ . In that terminology, a control policy is described as *asymptotically discount optimal* if this deviation vanishes as  $n \rightarrow \infty$  for all  $x \in \mathcal{X}$ . In this article, we do not employ the expectation but describe convergence directly in terms of  $\Lambda_n^\Phi(H_n^x)$ .

The measure of deviation of a control from optimality will be given by the following quantity. Define for MCM  $\pi$  and  $(x, a) \in \mathcal{K}$ ,

$$(4.2) \quad \lambda_\pi(x, a) = T_\pi^a V_\pi^*(x, a) - V_\pi^*(x).$$

This quantity is also employed in [30]. Note that if  $V_\pi^*$  is a fixed point of  $T_\pi$ , then it is easily verified that  $\lambda_\pi(x, a) \geq 0$ . We will need to impose the following conditions on MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  with solution space  $\mathcal{V}^w$ :

(B1) There exists a finite constant  $b_Q$  such that  $\int_{y \in \mathcal{X}} w(y) dQ(y \mid x, a) \leq w(x) + b_Q$  for all  $(x, a) \in \mathcal{K}$ .

(B2) There exists a finite constant  $b_\pi$  such that  $\lambda_\pi(x, a) \leq b_\pi w(x)$  for all  $(x, a) \in \mathcal{K}$ .

Condition (B1) is similar to regularity conditions developed for MCMs with unbounded rewards in, for example, [27, 35]. See also [6] for further discussion. Condition (B2) follows from (B1),  $\|V_\pi^*\| < \infty$ , and a suitable restriction on cost function  $R$  and state-action projection  $\mathcal{K}_x$ .

The important properties of (4.1) are given in the following theorem.

**THEOREM 4.1.** *Suppose an MDP  $\pi = (\mathcal{K}, Q, R, \beta)$  with control  $\Phi$  has contractive solution space  $\mathcal{V}^w$ . Suppose additionally that (B1) holds and  $\beta < 1$ . Then the quantity  $\Lambda_n^\Phi(H_n^x)$  of (4.1) is well defined, and*

$$(4.3) \quad \Lambda_n^\Phi(H_n^x) = V_\pi^*(X_n) + E_x^\Phi \left[ \sum_{i=0}^{\infty} \beta^i \lambda_\pi(X_{n+i}, A_{n+i}) \mid H_n^x \right].$$

*If, in addition, (B2) holds, then  $\Lambda_n^\Phi(H_n^x) < \infty$  w.p.1.*

*Proof.* Fix  $n$ , and note that

$$E_x^\Phi[w(X_{n+1}) \mid H_n^x] \leq w(X_n) + b_Q$$

by condition (B1). Then for any  $m \geq 1$ ,

$$\begin{aligned} E_x^\Phi[w(X_{n+m}) \mid H_n^x] &= E_x^\Phi[E_x^\Phi[w(X_{n+m}) \mid H_{n+m-1}^x] \mid H_n^x] \\ &\leq E_x^\Phi[w(X_{n+m-1}) + b_Q \mid H_n^x]. \end{aligned}$$

Repeated application of this argument gives

$$(4.4) \quad E_x^\Phi[w(X_{n+m}) \mid H_n^x] \leq w(X_n) + mb_Q.$$

Under the assumption that  $\mathcal{V}^w$  is a contractive solution space,  $V_\pi^*$  is a fixed point of  $T_\pi$ , and hence  $\lambda_\pi(x, a) \geq 0$ , so that the expectation  $E_x^\Phi[\sum_{i=0}^\infty \beta^i \lambda_\pi(X_{n+i}, A_{n+i}) \mid H_n^x]$  is well defined. Then

$$\begin{aligned} \lambda_\pi(X_{n+i}, A_{n+i}) &= R(X_{n+i}, A_{n+i}) + \beta \int_{y \in \mathcal{X}} V_\pi^*(y) dQ(y \mid X_{n+i}, A_{n+i}) - V_\pi^*(X_{n+i}) \\ &= R(X_{n+i}, A_{n+i}) + \beta E_x^\Phi[V_\pi^*(X_{n+i+1}) \mid H_{n+i}^a] - V_\pi^*(X_{n+i}), \end{aligned}$$

which, using the monotone convergence theorem and noting that  $\sigma(H_n^x) \subset \sigma(H_{n+i}^a)$  for  $i \geq 0$ , leads to

$$\begin{aligned} E_x^\Phi \left[ \sum_{i=0}^\infty \beta^i \lambda_\pi(X_{n+i}, A_{n+i}) \mid H_n^x \right] &= \lim_{N \rightarrow \infty} E_x^\Phi \left[ \sum_{i=0}^N \beta^i \lambda_\pi(X_{n+i}, A_{n+i}) \mid H_n^x \right] \\ &= \lim_{N \rightarrow \infty} E_x^\Phi \left[ \sum_{i=0}^N \beta^i R(X_{n+i}, A_{n+i}) \mid H_n^x \right] \\ (4.5) \quad &\quad -V_\pi^*(X_n) + E_x^\Phi[\beta^N \beta V_\pi^*(X_{n+N}) \mid H_n^x]. \end{aligned}$$

By assumption,  $\|V_\pi^*\|_w < \infty$ , so that

$$\begin{aligned} E_x^\Phi[\beta^N V_\pi^*(X_{n+N}) \mid H_n^x] &\leq \beta^N \|V_\pi^*\|_w E_x^\Phi[w(X_{n+N}) \mid H_n^x] \\ &\leq \beta^N \|V_\pi^*\|_w (w(X_n) + Nb_Q) \end{aligned}$$

using (4.4). This upper bound approaches 0 as  $N \rightarrow \infty$ . From (4.5) this implies that  $\Lambda_n^\Phi(H_n^x)$  of (4.1) is well defined, from which (4.3) follows. Applying (4.4) and condition (B2) gives

$$\begin{aligned} E_x^\Phi \left[ \sum_{i=0}^\infty \beta^i \lambda_\pi(X_{n+i}, A_{n+i}) \mid H_n^x \right] &\leq E_x^\Phi \left[ \sum_{i=0}^\infty \beta^i b_\pi w(X_{n+i}) \mid H_n^x \right] \\ (4.6) \quad &\leq \sum_{i=0}^\infty \beta^i (b_\pi w(X_n) + i b_\pi b_Q), \end{aligned}$$

the upper bound of which is finite, which completes the proof.  $\square$

The following theorem bounds the control error defined in (4.2) for a control function calculated (perhaps approximately) by

$$\phi(x) = \arg \min_{a \in \mathcal{K}_x} T_{\hat{\pi}}^a V(x, a)$$

when  $\hat{\pi}$  approximates true model  $\pi$  and  $V$  approximates value function  $V_\pi^*$ .

**THEOREM 4.2.** *Suppose we are given an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  with contractive solution space  $\mathcal{V}^w$  for which  $\eta_Q^w < \infty$ . Suppose we are given a second MCM  $\hat{\pi} =$*

$(\mathcal{K}, \hat{Q}, \hat{R}, \beta)$  such that  $\hat{Q}$  conforms to  $Q$ ,  $D_r^w(\hat{R}, R), D_q^w(\hat{Q}, Q)$  are finite, and  $T_{\hat{\pi}} : \mathcal{V} \rightarrow \mathcal{V}$ . For any  $V \in \mathcal{V}$  and  $\phi \in \mathcal{K}^f$  let

$$\varepsilon_V(x) = T_{\hat{\pi}}^a V(x, \phi(x)) - T_{\hat{\pi}} V(x)$$

when well defined, and set  $\varepsilon_V(x) \equiv 0$  otherwise. Then,

$$(4.7) \quad \begin{aligned} w^{-1}(x) \lambda_{\pi}(x, \phi(x)) / 2 &\leq \|V - V_{\pi}^*\|_w (\eta_Q^w + D_q^w(\hat{Q}, Q)) \\ &+ D_r^w(\hat{R}, R) + \|V_{\pi}^*\|_w D_q^w(\hat{Q}, Q) + \|\varepsilon_V\|_w / 2. \end{aligned}$$

*Proof.* If  $V \notin F(\mathcal{X}, \|\cdot\|_w)$ , then the upper bound in (4.7) is equal to  $\infty$ . Otherwise, assume  $V \in F(\mathcal{X}, \|\cdot\|_w)$ . By definition,  $V \in \mathcal{L}_Q$ , and by Lemma 3.2,  $V \in \mathcal{L}_{\hat{Q}}$ . This means the quantities  $T_{\pi} V_{\pi}^*(x)$ ,  $T_{\pi}^a V_{\pi}^*(x, a)$ ,  $T_{\hat{\pi}} V(x)$ ,  $T_{\hat{\pi}}^a V(x, a)$  are all well defined. We may write

$$(4.8) \quad \begin{aligned} w(x)^{-1} |T_{\hat{\pi}}^a V(x, a) - T_{\pi}^a V_{\pi}^*(x, a)| &\leq w(x)^{-1} |\hat{R}(x, a) - R(x, a)| \\ &+ w(x)^{-1} \beta \int w(y)^{-1} |V(y) - V_{\pi}^*(y)| w(y) f(y|x, a) d\mu_Q(y) \\ &+ w(x)^{-1} \beta \int w(y)^{-1} |V(y) - V_{\pi}^*(y)| w(y) |\hat{f}(y|x, a) - f(y|x, a)| d\mu_Q(y) \\ &+ w(x)^{-1} \beta \int w(y)^{-1} V_{\pi}^*(y) w(y) |\hat{f}(y|x, a) - f(y|x, a)| d\mu_Q(y) \\ &\leq \|V - V_{\pi}^*\|_w (\eta_Q^w + D_q^w(\hat{Q}, Q)) + D_r^w(\hat{R}, R) + \|V_{\pi}^*\|_w D_q^w(\hat{Q}, Q). \end{aligned}$$

Using the triangle inequality gives

$$\begin{aligned} w(x)^{-1} |T_{\pi}^a V^*(x, \phi(x)) - V_{\pi}^*(x)| &\leq w(x)^{-1} |T_{\pi}^a V^*(x, \phi(x)) - T_{\hat{\pi}}^a V(x, \phi(x))| \\ &+ w(x)^{-1} |T_{\hat{\pi}}^a V(x, \phi(x)) - T_{\hat{\pi}} V(x)| \\ &+ w(x)^{-1} |T_{\hat{\pi}} V(x) - V_{\pi}^*(x)| \\ &= K_1 + K_2 + K_3. \end{aligned}$$

Note that  $K_1$  can be bounded by (4.8) and  $K_2$  is equivalent to  $w(x)^{-1} \varepsilon_V(x)$ . By assumption,  $V_{\pi}^*(x)$  is finite and  $V_{\pi}^* \in \mathcal{L}_Q$ , so Lemma 3.3 applied to  $T_{\hat{\pi}}^a V(x, a)$  and  $T_{\pi}^a V_{\pi}^*(x, a)$  for fixed  $x$  may be used to bound  $K_3$  as follows:

$$\begin{aligned} w(x)^{-1} |T_{\hat{\pi}} V(x) - V_{\pi}^*(x)| &\leq w(x)^{-1} \sup_{a \in \mathcal{A}_x} |T_{\hat{\pi}}^a V(x, a) - T_{\pi}^a V_{\pi}^*(x, a)| \\ &\leq \sup_{(x, a) \in \mathcal{K}} w(x)^{-1} |T_{\hat{\pi}}^a V(x, a) - T_{\pi}^a V_{\pi}^*(x, a)|, \end{aligned}$$

which may be bounded using (4.8). Then (4.7) follows.  $\square$

We have shown that the realized cost of a policy may be estimated using two steps. First, the deviation of a particular control function  $\lambda_{\pi}(x, \phi(x))$  is bounded using Theorem 4.2. This bound is then incorporated into the expression (4.3) given in Theorem 4.1. We will illustrate this procedure on a type of adaptive control in section 6.

**5. Parametric models.** In this section we consider parametric representations of model approximations. A class of model elements  $(Q^{\theta}, R^{\theta})$  is indexed by a parameter  $\theta \in \Theta$  for some metric space  $\Theta$ . A model approximation assumes the form

$\hat{\pi} = (\mathcal{K}, \hat{Q}, \hat{R}, \beta) = (\mathcal{K}, Q^{\hat{\theta}}, R^{\hat{\theta}}, \beta)$  for some  $\hat{\theta} \in \Theta$ . Model identification then reduces to an estimation problem on  $\Theta$ , which is especially advantageous when  $\Theta$  is naturally a finite-dimensional space. Of course, any model with finite  $\mathcal{K}$  may be represented this way. This approach is especially suitable when the source of randomness of the MCM is a single underlying stochastic process, for example, some finite collection of arrival processes.

We proceed by first defining a parametric family of models, and then imposing some notion of Lipschitz continuity for model distance terms  $(D_q^w(Q^{\theta_1}, Q^{\theta_2}), D_r^w(R^{\theta_1}, R^{\theta_2}))$  with respect to  $\Theta$ . We will show that this is easily done for densities of the exponential family type, which includes many commonly used densities such as the Poisson, gamma, and Gaussian. It is important to note that little reference is made in this section to any specific MCM. The objective is rather to define a useful class of putative models from which  $(\hat{Q}, \hat{R})$  may be selected, and which permit a convenient calculation of model distances. Once this is done, the theory of sections 3 and 4 applies.

DEFINITION 5.1. *Given the state-action space  $\mathcal{K}$ , a parametric model (PM) consists of*

- (i) *a parameter space  $\Theta$  with metric  $d(\cdot, \cdot)$ ,*
- (ii) *a measure  $\mu_{\Theta} \in \mathcal{M}(\mathcal{X})$ , and*
- (iii) *a set  $M_{qr}^{\Theta} = \{(Q^{\theta}, R^{\theta}) : \theta \in \Theta\}$  of pairs in which  $Q^{\theta}$  satisfies (M4) with respect to measure  $\mu_{\Theta}$  and  $R^{\theta}$  satisfies (M5).*

*The PM is denoted  $M^{\Theta} = (\Theta, d, M_{qr}^{\Theta}, \mu_{\Theta})$ . Additionally, the PM conforms to an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  with solution space  $\mathcal{V}^w$  if  $\mu_{\Theta} = \mu_Q$ , if there exists  $\theta' \in \Theta$  such that  $(Q^{\theta'}, R^{\theta'}) = (Q, R)$ , and if the DPO for MCM  $\pi^{\theta} = (\mathcal{K}, Q^{\theta}, R^{\theta}, \beta)$  satisfies  $T_{\pi^{\theta}} : \mathcal{V} \cap \mathcal{L}_{Q^{\theta}} \rightarrow \mathcal{V}$  for each  $\theta \in \Theta$ .*

Remark 5.1. That  $T_{\pi^{\theta}} : \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w) \rightarrow \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$  may be established under the conditions of Theorem 3.5. For the types of problems considered in this article, it will generally suffice to replace  $\mathcal{V}$  with  $\mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$ .

For convenience we denote the density of  $Q^{\theta}(\cdot \mid x, a)$  with respect to  $\mu_{\Theta}$  by  $f^{\theta}(\cdot \mid x, a)$  for which  $E_{x,a}^{\theta}$  is the expectation operator.

**5.1. Lipschitz continuity.** Given a natural PM, we would like to establish a form of Lipschitz continuity which will allow us to express model distance measures  $D_q^w(Q^{\theta_1}, Q^{\theta_2})$  and  $D_r^w(R^{\theta_1}, R^{\theta_2})$  in terms of  $d(\theta_1, \theta_2)$ . In practice, this will have to incorporate the weight function  $w$  used in Definition 3.3 of solution space  $\mathcal{V}^w$ . We begin with the following definition.

DEFINITION 5.2. *For a given metric space  $(\Theta, d)$ , Borel space  $\mathcal{U}$ , and a measurable function  $t : \mathcal{U} \rightarrow \mathbb{R}$ , we say a parametric family  $\{g^{\theta} : \theta \in \Theta\}$  of density functions on  $\mathcal{U}$  is  $t$ -weighted Lipschitz on  $B \subset \Theta$  if*

$$(5.1) \quad |g^{\theta_1}(u) - g^{\theta_2}(u)| \leq d(\theta_1, \theta_2)t(u) (g^{\theta_1}(u) + g^{\theta_2}(u)) \quad \forall u \in \mathcal{U}$$

*for any  $\theta_1, \theta_2 \in B$ .*

Given  $M^{\Theta} = (\Theta, d, M_{qr}^{\Theta}, \mu_{\Theta})$  defined for state-action space  $\mathcal{K}$ , a weight function  $w \in \mathcal{W}(\mathcal{X})$  and a subset  $B \in \Theta$ , define the following conditions:

- (C1) There exists a finite constant  $C_B^r$  such that

$$w(x)^{-1}|R^{\theta_1}(x, a) - R^{\theta_2}(x, a)| \leq C_B^r d(\theta_1, \theta_2)$$

for all  $(x, a) \in \mathcal{K}$ ,  $\theta_1, \theta_2 \in B$ .

- (C2) There exists a finite constant  $C_B^q$  and a measurable mapping  $t(y|x, a) : \mathcal{X} \times \mathcal{K} \rightarrow (0, \infty)$  such that for each  $(x, a) \in \mathcal{K}$  the parametric family  $\{f^\theta(\cdot|x, a) : \theta \in \Theta\}$  is  $t(\cdot|x, a)$ -weighted Lipschitz on  $B$ , with

$$(5.2) \quad \sup_{\theta \in B} \sup_{(x, a) \in \mathcal{K}} w(x)^{-1} E_{x, a}^\theta [w(Y)t(Y|x, a)] \leq C_B^q.$$

That (C1) and (C2) imply Lipschitz continuity of  $M^\Theta$  is verified in the following theorem.

**THEOREM 5.1.** *Suppose conditions (C1)–(C2) hold for a PM  $M^\Theta = (\Theta, d, M_{qr}^\Theta, \mu_\Theta)$  defined on a state-action space  $\mathcal{K}$ , a weight function  $w \in \mathcal{W}(\mathcal{X})$ , and a subset  $B \subset \Theta$ . Then*

$$(5.3) \quad D_q^w(Q^{\theta_1}, Q^{\theta_2}) \leq 2C_B^q d(\theta_1, \theta_2),$$

$$(5.4) \quad D_r^w(R^{\theta_1}, R^{\theta_2}) \leq C_B^r d(\theta_1, \theta_2)$$

for all  $\theta_1, \theta_2 \in B$ .

*Proof.* For any  $\theta_1, \theta_2 \in B$ ,  $(x, a) \in \mathcal{K}$  we have by (C2),

$$\begin{aligned} & \int_{\mathcal{X}} w(y) |f^{\theta_1}(y|x, a) - f^{\theta_2}(y|x, a)| d\mu_\Theta(y) \\ & \leq d(\theta_1, \theta_2) \int_{\mathcal{X}} w(y) t(y|x, a) (f^{\theta_1}(y|x, a) + f^{\theta_2}(y|x, a)) d\mu_\Theta(y) \\ & = d(\theta_1, \theta_2) (E_{x, a}^{\theta_1} [w(Y)t(Y|x, a)] + E_{x, a}^{\theta_2} [w(Y)t(Y|x, a)]), \end{aligned}$$

where we employ inequality (5.1). Inequality (5.3) follows by applying (5.2). Then (5.4) follows directly from assumption (C1).  $\square$

**5.2. Exponential family densities.** It is important to establish that the form of the inequality (5.1) is one that might be encountered in practice, with some natural choice of weighting function  $t(u)$ . It will, in fact, hold under general conditions for an exponential family type of density commonly employed in statistical inference. This form of density usually provides a natural relationship between a parameter space and a probability measure. See, for example, Lehmann and Casella [26].

**DEFINITION 5.3.** *An exponential family of densities on  $\mathcal{U} \subset \mathbb{R}^p$  (with respect to the Lebesgue measure) with parameter space  $\Theta \subset \mathbb{R}^k$  takes the form*

$$g^\theta(u) = \exp(v(u, \theta) - b(\theta) + h(u)), \quad u \in \mathcal{U},$$

where

$$v(u, \theta) = \sum_{i=1}^k \theta_i v_i(u),$$

for real valued functions  $v_1, \dots, v_k, h$  defined on  $\mathcal{U}$  and  $b$  defined on  $\Theta$ .

Definition 5.2 is motivated by the following theorem.

**THEOREM 5.2.** *Suppose we have an exponential family of densities given in Definition 5.3. Define metric  $d(x, y) = \|x - y\|_1$  on  $\mathbb{R}^k$ , where  $\|\cdot\|_1$  is  $\ell^1$  norm, and suppose  $b(\theta)$  is locally Lipschitz. Let  $B$  be a compact subset of  $\Theta$ . Then the exponential family is  $t$ -weighted Lipschitz on  $B$  with respect to  $d$ , with  $t(u) = \|(v_1(u), \dots, v_k(u))\|_1 + M_B$  for some finite constant  $M_B$ .*



*Proof.* We will make use of the inequality

$$(5.5) \quad |e^x - 1| \leq |x|e^{\max(0,x)} \leq |x|(1 + e^x), \quad x \in \mathbb{R}.$$

Let  $\theta_1, \theta_2 \in B$ . Setting  $\delta = \theta_2 - \theta_1$ ,  $b^\delta = b(\theta_2) - b(\theta_1)$ , we may write directly

$$|g^{\theta_2}(u) - g^{\theta_1}(u)| = |\exp(v(u, \delta) - b^\delta) - 1| |g^{\theta_1}(u)|,$$

and then using (5.5) gives

$$\begin{aligned} |g^{\theta_2}(u) - g^{\theta_1}(u)| &\leq |v(u, \delta) - b^\delta| (1 + \exp(v(u, \delta) - b^\delta)) |g^{\theta_1}(u)| \\ &= |v(u, \delta) - b^\delta| (g^{\theta_2}(u) + g^{\theta_1}(u)). \end{aligned}$$

Finally, if  $b$  is locally Lipschitz, there exists a constant  $M_B$  such that  $|b(\theta_2) - b(\theta_1)| \leq M_B d(\theta_1, \theta_2)$  for  $\theta_1, \theta_2 \in B$ . Also,  $|v(u, \delta)| \leq \|(v_1(u), \dots, v_k(u))\|_1 d(\theta_1, \theta_2)$ , giving

$$|g^{\theta_2}(u) - g^{\theta_1}(u)| \leq d(\theta_1, \theta_2) (\|(v_1(u), \dots, v_k(u))\|_1 + M_B) (g^{\theta_1}(u) + g^{\theta_2}(u))$$

for all  $u \in \mathcal{U}$ ,  $\theta_1, \theta_2 \in B$ , which concludes the proof.  $\square$

This result would be used to establish the property given in Definition 5.2, after which it would remain to be verified that the expectations in condition (C2) can be suitably bounded.

**6. Adaptive control.** We now consider the problem of adaptive control. Suppose an MDP operates indefinitely according to an MCM with unknown  $(Q, R)$ . The process history is available for constructing a sequence of model estimates  $\{\hat{Q}_n, \hat{R}_n; n \geq 1\}$ , synchronized with the stages of the MDP, which are in turn used to refine the control policy using the methods discussed in sections 3 and 4. Estimation is based on a PM of the form discussed in section 5, with parameter space  $\Theta \subset \mathbb{R}^k$ . If we cannot guarantee that statistical information regarding each component of the parameter is available at each stage, then we resort to forced exploration.

This, however, leaves us with two sources of regret. The *estimation regret* is that inherent in using a model estimate instead of the true model to calculate a control policy, which can be bounded using the methods of sections 3 and 4. The second source, *exploratory regret*, is that attributable to exploration. We take as given that a nonzero regret accrues from any exploratory behavior. We also note that by using Theorems 4.1 and 4.2 an upper bound can be placed on exploratory regret. Hence, we may place a bound on the rate at which exploratory regret is accrued by characterizing the exploration rate itself, which we take to be the rate (expressed as a proportion of stages) at which the MDP adopts an exploratory control.

Clearly, these two forms of regret represent a type of trade-off. A higher exploration rate results in lower estimation regret at the cost of higher exploration regret. The purpose of this section is to give an analytical bound for the combined regret in terms of the exploration rate. This can then be used to design an efficient exploratory control policy.

We will need to accomplish three preliminary tasks. We first need to expand our definition of an MDP in order to accommodate exploratory behavior (section 6.1). Fortunately, we can leave the basic definition given in section 3.1 essentially unaltered. We then need to develop a notion of model estimation in an online setting using process history (section 6.2). We will assume a finite-dimensional PM of the type defined in section 5. The objective will be to develop regularity conditions which permit a well-defined rate of model estimate convergence. We will also need to define precisely an exploration rate, which will permit direct calculation of estimation and exploratory regret (section 6.3).

**6.1. Definition of an adaptive MDP.** We will need to expand on the definition of an MDP to include exploratory behavior, as well as observable data available for model estimation. Throughout this section we assume the existence of an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  which possesses a contractive solution space  $\mathcal{V}^w = (\mathcal{V}, w)$  as well as an OSDP based on control function  $\phi_\pi^* \in \mathcal{K}^f$  with value function  $V_\pi^*$ . We add two new elements as follows:

- (M7) A Borel space  $\mathcal{O}$ , called the *observation space*, and a stochastic kernel  $Q^{o,x} : \mathcal{K} \rightarrow \mathcal{M}(\mathcal{O} \times \mathcal{X})$  for which  $Q^{o,x}(\mathcal{O} \times E_x \mid x, a) = Q(E_x \mid x, a)$  for all  $E_x \in \mathcal{B}(\mathcal{X})$ ,  $(x, a) \in \mathcal{K}$ .
- (M8) A binary outcome  $\mathcal{Z} = \{0, 1\}$  and a sequence of measurable mappings  $p_n^e : (\mathcal{X}\mathcal{Z}\mathcal{A}\mathcal{O})^{n-1} \times \mathcal{X} \rightarrow [0, 1]$ .

We will let  $E_{x,a}^{Q^{o,x}}$  be the expectation operator associated with  $Q^{o,x}(\cdot \mid x, a)$ . In addition to the state-action pair  $(X_n, A_n)$  and cost  $R(X_n, A_n)$ , we associate with stage  $n$  two new quantities. First, we include a random observation  $O_n$  defined on  $\mathcal{O}$ , with distribution calculable from  $Q^{o,x}(\cdot \mid X_n, A_n)$  according to (M7). This represents the information available to the controller pertaining to the  $n$ th stage transition from  $X_n$  to  $X_{n+1}$  and the realization of the  $n$ th stage cost. This is assumed to be available in time to influence the control applied at the  $n + 1$ st stage. Estimates are assumed to be sample averages of functions of  $O_n$ . This formulation suffices, for example, when the parameter represents a state transition rate, in which case  $O_n = (X_n, X_{n+1})$ . Alternatively, if the parameters represent arrival rates, then  $O_n$  may be a set of arrival counts coupled with an interobservation time. Additionally,  $O_n$  may include an observation of  $R(X_n, A_n)$ , and may incorporate observation error, although some conditions on state observability are necessary for the adaptive control policy defined below. It should be pointed out, however, that this construction does not accommodate many important recursive estimation procedures. Additional discussion of this topic will be deferred to section 6.7.

Given state  $X_n$ , a binary randomization quantity  $Z_n \in \mathcal{Z} = \{0, 1\}$  is observed, then an action  $A_n$  is selected according to a distribution dependent on the process history up to that point (including  $X_n$  and  $Z_n$ ). The role of  $Z_n$  is to define an *exploration schedule* (that is, exploratory behavior is forced when  $Z_n = 1$ ), and will be controlled by  $p_n^e$  given in (M8). It will be helpful to think of the order of realization of the random quantities as  $X_n \rightarrow Z_n \rightarrow A_n \rightarrow O_n \rightarrow X_{n+1} \rightarrow Z_{n+1} \rightarrow \dots$ . We accordingly define the Borel space  $\mathcal{S} \subset \mathcal{X}\mathcal{Z}\mathcal{A}\mathcal{O}$  to be all elements  $(x, w, a, o) \in \mathcal{X}\mathcal{Z}\mathcal{A}\mathcal{O}$  for which  $(x, a) \in \mathcal{K}$ .

It will be convenient to redefine the history vectors

$$\begin{aligned} H_n^a &= (X_1, Z_1, A_1, O_1, \dots, X_n, Z_n, A_n), \\ H_n^z &= (X_1, Z_1, A_1, O_1, \dots, X_n, Z_n), \\ H_n^x &= (X_1, Z_1, A_1, O_1, \dots, X_n). \end{aligned} \tag{6.1}$$

The adaptive control will be a mixture of two policies,  $\Phi^e = (\Phi_1^e, \Phi_2^e, \dots)$  and  $\Phi^o = (\Phi_1^o, \Phi_2^o, \dots)$ , where  $\Phi_n^e$  and  $\Phi_n^o$  are measurable mappings of  $H_n^x$  to  $\mathcal{M}(\mathcal{A})$ . The randomization variable  $Z_n$  is used to select the policy according to the form

$$\Phi_n(E_a \mid H_n^z) = (1 - Z_n)\Phi_n^o(E_a \mid H_n^x) + Z_n\Phi_n^e(E_a \mid H_n^x), \quad E_a \in \mathcal{B}(\mathcal{A}). \tag{6.2}$$

Essentially the purpose of  $\Phi^e$  is to explore, while  $\Phi^o$  in some sense converges to the OSDP. By regarding the action space as incorporating the randomization variable  $Z_n$  and the state space incorporating the observation variable  $O_n$  (while keeping  $R$  and

$Q$  independent of  $Z_n$  and  $O_n$ ), we retain the original definition of an MCM of the type defined in section 3, for which, given starting state  $X_1 = x$ , a unique measure  $P_x^\Phi$  exists on the Borel space  $\mathcal{S}^\infty$  satisfying

$$\begin{aligned} P_x^\Phi(X_1 = x) &= 1, \\ P_x^\Phi((O_n, X_{n+1}) \in E_{ox} \mid H_n^a) &= Q^{o,x}(E_{ox} \mid X_n, A_n), \quad E_{ox} \in \mathcal{B}(\mathcal{OX}), \\ P_x^\Phi(X_{n+1} \in E_x \mid H_n^a) &= Q(E_x \mid X_n, A_n), \quad E_x \in \mathcal{B}(\mathcal{X}), \\ P_x^\Phi(Z_n = 1 \mid H_n^x) &= p_n^e(H_n^x), \\ (6.3) \quad P_x^\Phi(A_n \in E_a \mid H_n^z) &= \Phi_n(E_a \mid H_n^z), \quad E_a \in \mathcal{B}(\mathcal{A}), \end{aligned}$$

for  $n \geq 1$  and each admissible history  $H_n^x, H_n^z, H_n^a$ . As above, we let  $E_x^\Phi$  be the expectation operator of  $P_x^\Phi$ , and  $x$  may be any initial state.

We additionally assume we have a PM  $M^\Theta$  which conforms to  $\pi = (\mathcal{K}, Q, R, \beta)$ , where  $\Theta \subset \mathbb{R}^k$  and the metric  $d$  on  $\Theta$  is taken to be  $d(\theta, \theta') = \|\theta - \theta'\|_1$ , where  $\|\cdot\|_1$  is  $\ell_1$  norm in  $\mathbb{R}^k$ . The true parameter is denoted  $\theta'$ , that is,  $(Q, R) = (Q^{\theta'}, R^{\theta'})$ . Estimates  $(\hat{\theta}_n, \hat{\phi}_n)$  of parameter  $\theta'$  and optimal control function  $\phi_\pi^*$  will be associated with each stage  $n$ . We therefore assume that  $(\hat{\theta}_n, \hat{\phi}_n)$  can be calculated by the end of stage  $n$ .

**6.2. Online parameter estimation.** We now consider the problem of constructing the estimators  $\hat{\theta}_n$ . It may be that a specific observation  $O_n$  contains statistical information about some, but not all, parameter components  $(\theta_1, \dots, \theta_k)$ . This will depend on the current state-action pair  $(X_n, A_n)$  through the measure  $Q^{o,x}(\cdot \mid x, a)$ . We may therefore associate with each parameter component  $\theta_j$  the elements of  $\mathcal{K}$  which admit estimation.

**DEFINITION 6.1.** Suppose we have an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$ , observation space  $\mathcal{O}$ , kernel  $Q^{o,x}$  as defined in (M7), and a conforming PM  $M^\Theta$  with  $\Theta \subset \mathbb{R}^k$  and  $(Q, R) = (Q^{\theta'}, R^{\theta'})$ . The informative subset of  $\mathcal{K}$  for component  $\theta_j$  of  $\theta = (\theta_1, \dots, \theta_k)$ , which we denote  $\mathcal{K}(\theta_j)$ , consists of all  $(x, a) \in \mathcal{K}$  for which a measurable estimator  $\bar{\theta}_j(O)$ ,  $O \in \mathcal{O}$ , exists satisfying

$$(6.4) \quad E_{x,a}^{Q^{o,x}}[\bar{\theta}_j(O)] = \theta'_j \text{ and } E_{x,a}^{Q^{o,x}}[(\bar{\theta}_j(O) - \theta'_j)^2] \leq \nu$$

for some constant  $0 \leq \nu < \infty$ , where  $\theta' = (\theta'_1, \dots, \theta'_k)$  is the true parameter.

For convenience set  $I_n(\theta_j) = I\{(X_n, A_n) \in \mathcal{K}(\theta_j)\}$ . Given measure (6.3) consider the quantities

$$W_n(\theta_j) = \sum_{i=1}^n (\bar{\theta}_j(O_i) - \theta'_j) I_i(\theta_j), \quad n \geq 1.$$

Note that the process  $(W_1(\theta_j), W_2(\theta_j), \dots)$  is adapted to the filtration  $(\sigma(H_2^a), \sigma(H_3^a), \dots)$ . From (6.3) and (6.4) we have

$$\begin{aligned} E_x^\Phi[W_n(\theta_j) \mid H_n^a] &= E_x^\Phi[(\bar{\theta}_j(O_n) - \theta'_j) I_n(\theta_j) \mid H_n^a] + W_{n-1}(\theta_j) \\ &= E_x^\Phi[(\bar{\theta}_j(O_n) - \theta'_j) I_n(\theta_j) \mid X_n, A_n] + W_{n-1}(\theta_j) \\ &= W_{n-1}(\theta_j), \quad n \geq 1, \end{aligned}$$

so that  $\{W_n(\theta_j); n \geq 0\}$  is a martingale, where we set  $W_0(\theta_j) = 0$ . Similarly define

$$\begin{aligned} \Delta_n(\theta_j) &= E_x^\Phi[(W_n(\theta_j) - W_{n-1}(\theta_j))^2 \mid H_n^a] \\ &= E_x^\Phi[(\bar{\theta}_j(O_n) - \theta'_j)^2 I_n(\theta_j) \mid H_n^a] \\ &\leq \nu I_n(\theta_j), \quad n \geq 1. \end{aligned}$$

Let  $S_n(\theta_j) = \sum_{i=1}^n \Delta_i(\theta_j)$  and define the counting process

$$M_n(\theta_j) = \sum_{i=1}^n I_i(\theta_j), \quad n \geq 1,$$

and set  $M_0(\theta_j) = 0$ . It follows that  $S_n(\theta_j) \leq \nu M_n(\theta_j)$ . Since  $\{W_n(\theta_j); n \geq 1\}$  is clearly square-integrable, we may apply a suitable martingale law of large numbers (for example, Theorem 1.3.15 in Duflo [9]) to conclude

$$\left| \frac{W_n(\theta_j)}{S_n(\theta_j)} \right| = o\left(S_n(\theta_j)^{-1/2+\epsilon}\right),$$

or equivalently,

$$(6.5) \quad \left| \frac{W_n(\theta_j)}{M_n(\theta_j)} \right| = o\left((S_n(\theta_j)/M_n(\theta_j))^{1/2+\epsilon} M_n(\theta_j)^{-1/2+\epsilon}\right) \leq o\left(M_n(\theta_j)^{-1/2+\epsilon}\right)$$

for any small  $\epsilon > 0$ . This leads to component estimates

$$(6.6) \quad \hat{\theta}_{n,j} = \begin{cases} M_n(\theta_j)^{-1} \sum_{i=1}^n \bar{\theta}_j(O_i) I_i(\theta_j); & M_n(\theta_j) \geq 1, \\ \hat{\theta}_{0,j}; & M_n(\theta_j) = 0 \end{cases}$$

for  $n \geq 1, j = 1, \dots, k$ , where  $\hat{\theta}_0 = (\hat{\theta}_{0,1}, \dots, \hat{\theta}_{0,k})$  is a suitably chosen starting value. The parameter estimate sequence is then  $\hat{\theta}_n = (\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,k})$ . From (6.5) we have

$$(6.7) \quad \begin{aligned} |\hat{\theta}_{n,j} - \theta'_j| &= o\left(M_n(\theta_j)^{-1/2+\epsilon}\right), \\ d(\hat{\theta}_n, \theta') &= o\left(M_n(\theta)^{-1/2+\epsilon}\right) \end{aligned}$$

for any  $\epsilon > 0$ , where

$$M_n(\theta) = \min_{1 \leq j \leq k} M_n(\theta_j), \quad n \geq 1.$$

Hence, convergence of  $\hat{\theta}_n$  to  $\theta'$  follows from  $M_n(\theta) \rightarrow \infty$  at a rate implied by  $M_n(\theta)$ .

Finally, the following lemma will be useful in establishing a rate of convergence for an adaptive control. We will need to impose the following condition:

(D1) For a given weight function  $w \in \mathcal{W}(\mathcal{X})$ ,

$$E_{x,a}^{Q^{o,x}}[w(X)|\bar{\theta}_j(O) - \theta'_j|] \leq \tau w(x)$$

for a finite constant  $\tau$  when  $(x, a) \in \mathcal{K}(\theta_j)$ , for each  $j = 1, \dots, k$ .

This condition supplements (B1).

LEMMA 6.1. *Suppose, under the conditions of Definition 6.1, given a weight function  $w \in \mathcal{W}(\mathcal{X})$ , conditions (B1) and (D1) hold. Then*

$$(6.8) \quad \begin{aligned} E_x^\Phi[w(X_{n+m})|\hat{\theta}_{n+m-1,j} - \theta'_j| \mid H_n^a] &\leq (w(X_n) + mb_Q)|\hat{\theta}_{n-1,j} - \theta'_j| \\ &\quad + \frac{m\tau w(X_n) + b_Q(\nu^{1/2} + \tau)m(m-1)/2}{M_n(\theta_j)} \end{aligned}$$

for  $m \geq 0, n \geq 1$ .

*Proof.* First note that  $X_n$  and  $\theta'_j$  are both  $\sigma(H_n^a)$ -measurable, so that (6.8) is easily verified for  $m = 0$ . Next assume  $m \geq 1$ . For  $n \geq 1$ , if  $M_n(\theta_j) \geq 1$  we may write

$$\begin{aligned} |\hat{\theta}_{n,j} - \theta'_j| &= \frac{|\sum_{i=1}^n (\bar{\theta}_j(O_i) - \theta'_j) I_i(\theta_j)|}{M_n(\theta_j)} \\ &\leq \frac{|(\bar{\theta}_j(O_n) - \theta'_j) I_n(\theta_j)|}{M_n(\theta_j)} + \frac{|\sum_{i=1}^{n-1} (\bar{\theta}_j(O_i) - \theta'_j) I_i(\theta_j)|}{M_n(\theta_j)}. \end{aligned}$$

Note that  $M_n(\theta_j) \geq M_{n-1}(\theta_j)$ , and that the second term of this upper bound is zero if  $M_{n-1}(\theta_j) = 0$ . We may therefore write

$$(6.9) \quad |\hat{\theta}_{n,j} - \theta'_j| \leq \frac{|(\bar{\theta}_j(O_n) - \theta'_j) I_n(\theta_j)|}{M_n(\theta_j)} + |\hat{\theta}_{n-1,j} - \theta'_j|.$$

We will need the following three inequalities:

$$(6.10) \quad E_x^\Phi \left[ w(X_{n+1}) |\hat{\theta}_{n,j} - \theta'_j| \mid H_n^a \right] \leq \frac{\tau w(X_n)}{M_n(\theta_j)} + (w(X_n) + b_Q) |\hat{\theta}_{n-1,j} - \theta'_j|,$$

$$(6.11) \quad E_x^\Phi \left[ |\hat{\theta}_{n,j} - \theta'_j| \mid H_n^a \right] \leq \frac{\nu^{1/2}}{M_n(\theta_j)} + |\hat{\theta}_{n-1,j} - \theta'_j|,$$

$$(6.12) \quad E_x^\Phi \left[ w(X_{n+1}) \mid H_n^a \right] \leq w(X_n) + b_Q.$$

Inequalities (6.10) and (6.11) make use of (6.9) and the fact that  $\hat{\theta}_{n-1,j}$  and  $M_n(\theta_j)$  are  $\sigma(H_n^a)$ -measurable. Then, in particular, (6.10) follows from (D1) and (B1), (6.11) follows from Definition 6.1 with Jensen's inequality, and (6.12) follows from (B1). Note that (6.10)–(6.12) hold trivially for  $M_n(\theta_j) = 0$ .

We then complete the argument by induction. Suppose (6.8) holds for  $n \geq 1$  for some  $m \geq 1$ . Then, considering  $m + 1$ , we write

$$\begin{aligned} &E_x^\Phi \left[ w(X_{n+m+1}) |\hat{\theta}_{n+m,j} - \theta'_j| \mid H_n^a \right] \\ &= E_x^\Phi \left[ E_x^\Phi [w(X_{n+m+1}) |\hat{\theta}_{n+m,j} - \theta'_j| \mid H_{n+1}^a] \mid H_n^a \right] \\ &\leq E_x^\Phi \left[ (w(X_{n+1}) + mb_Q) |\hat{\theta}_{n,j} - \theta'_j| + \frac{m\tau w(X_{n+1}) + b_Q(\nu^{1/2} + \tau)m(m-1)/2}{M_{n+1}(\theta_j)} \mid H_n^a \right] \\ &\leq E_x^\Phi \left[ (w(X_{n+1}) + mb_Q) |\hat{\theta}_{n,j} - \theta'_j| \mid H_n^a \right] \\ &\quad + \frac{E_x^\Phi [m\tau w(X_{n+1}) + b_Q(\nu^{1/2} + \tau)m(m-1)/2 \mid H_n^a]}{M_n(\theta_j)} \end{aligned}$$

after applying the induction hypothesis and using the fact that  $M_{n+1}(\theta_j) \geq M_n(\theta_j)$  and that  $M_n(\theta_j)$  is  $\sigma(H_n^a)$ -measurable. Direct application of inequalities (6.10)–(6.12) yields, after some algebra, (6.8) with  $m$  incremented to  $m + 1$ . The proof follows by noting that for  $m = 1$ , (6.8) follows directly from (6.10).  $\square$

**6.3. Exploration rates.** In order to verify convergence of  $\hat{\theta}_n$  to  $\theta'$  we must have each informative subset  $\mathcal{K}(\theta_j)$  visited infinitely often. This is not problematic if  $\mathcal{K}(\theta_j) \equiv \mathcal{K}$  for each  $\theta_j$ , or if we may otherwise verify that  $M_n(\theta) \rightarrow_n \infty$  for a sufficiently rich class of control policies. In the absence of any such guarantee, one

method of ensuring sufficient exploration is to employ blocks of exploratory control, which is the motivation for the mixed policy defined in (6.2). Define

$$\begin{aligned} B_n &= I\{Z_n = 1\}I\{Z_{n-1} = 0\}, \quad n \geq 1, \\ S_n &= \sum_{i=1}^n B_i, \quad n \geq 1, \\ J_k &= \inf\{j : S_j = k\}, \quad k \geq 1, \\ I_k &= \inf\{m \geq 1 : Z_{J_k+m} = 0\}I\{J_k < \infty\}, \quad k \geq 1, \end{aligned}$$

setting for convenience  $Z_0 = 0$ . A block begins at stage  $n$  if  $B_n = 1$ ,  $S_n$  is the number of blocks begun by stage  $n$ ,  $J_k$  is the stage at which the  $k$ th block begins, and  $I_k$  is the duration of the  $k$ th block. We will employ the following conditions:

- (D2)  $S_n \rightarrow_n \infty$ .
- (D3) For some constant  $\delta > 0$ , for each  $\theta_j$  of  $\theta$ ,

$$P_x^\Phi(K_n(\theta_j)I\{B_n = 1\} \mid H_n^z) \geq \delta I\{B_n = 1\},$$

where

$$K_n(\theta_j) = \cup_{j=1}^{I_{S_n}} \{(X_{n+j-1}, A_{n+j-1}) \in \mathcal{K}(\theta_j)\},$$

for  $n \geq 1$ .

*Remark 6.1.* Under (D2) the sequence  $\{J_k : k \geq 1\}$  satisfies  $J_k < \infty$  and forms an increasing sequence of stopping times.

*Remark 6.2.* Condition (D3) states that given that a block begins at stage  $n$ , conditional on the history up to that block, the probability of visiting  $\mathcal{K}(\theta_j)$  within the block is at least  $\delta$ . It holds uniformly over all  $X_n$  and  $\theta_j$ . This requirement was devised primarily for mathematical convenience, and in practice may need to be relaxed. The use of an exploratory control in an actual control setting would presumably be subject to many practical constraints. See Thrun [33] for an interesting discussion of this issue.

We make use of the following theorem due to Dubins and Freedman [8].

**THEOREM 6.2.** *Suppose a sequence of events  $E_j$  is adapted to filtration  $\mathcal{F}_j$ ,  $j \geq 0$ , defined on probability measure  $P$ ; then*

$$L_n = \frac{\sum_{j=1}^n I\{E_j\}}{\sum_{j=1}^n P(E_j|\mathcal{F}_{j-1})}$$

*converges to a finite limit  $L$  w.p.1, with  $L = 1$  on  $\{\sum_{j \geq 1} P(E_j|\mathcal{F}_{j-1}) = \infty\}$ .*

Under conditions (D2) and (D3), the exploration rate follows from the counting process  $S_n$ , as formalized in the following theorem.

**THEOREM 6.3.** *Given an MDP with measure (6.3), if (D2)–(D3) hold, then*

$$\liminf_{n \rightarrow \infty} \frac{M_n(\theta_j)}{S_n} \geq \delta$$

*w.p.1 for all  $\theta_j$ .*

*Proof.* By (D2)  $J_1, J_2, \dots$  forms a sequence of increasing, finite stopping times. We may define the  $\sigma$ -algebra  $\mathcal{F}_k^J \subset \mathcal{S}^\infty$  to be that generated by Borel sets resolvable

by  $(X_1, Z_1, A_1, O_1, \dots, X_{J_k}, Z_{J_k})$ . This defines filtration  $\mathcal{F}^J = (\mathcal{F}_1^J, \mathcal{F}_2^J, \dots)$ . We then have

$$(6.13) \quad M_n(\theta_j) \geq \sum_{k=1}^{S_n-1} I\{K_{J_k}(\theta_j)\}, \quad n \geq 1,$$

where we set the summation to be zero for  $S_n - 1 < 1$ . We then argue that  $K_{J_k}(\theta_j) \in \mathcal{F}_{k+1}^J$ , since the occurrence of  $K_{J_k}(\theta_j)$  is resolved before stopping time  $J_{k+1}$ . On the other hand, by (D3) we have  $P_x^\Phi(K_{J_k}(\theta_j) \mid \mathcal{F}_k^J) \geq \delta$ ,  $k \geq 1$ . Applying Theorem 6.2 and noting (D2) and (6.13) gives

$$1 = \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^{S_n-1} I\{K_{J_k}(\theta_j)\}}{\sum_{k=1}^{S_n-1} P_x^\Phi(K_{J_k}(\theta_j) \mid \mathcal{F}_k^J)} \leq \liminf_{n \rightarrow \infty} \frac{M_n(\theta_j)}{\delta(S_n - 1)} \quad \text{w.p.1,}$$

which proves the theorem.  $\square$

The consequence of Theorem 6.3 is that under conditions (D2)–(D3), given (6.7), the exploration schedule will guarantee

$$d(\hat{\theta}_n, \theta') = o\left(S_n^{-1/2+\epsilon}\right)$$

for any  $\epsilon > 0$ .

**6.3.1. Randomized exploration schedules.** It is straightforward to devise a completely deterministic schedule that satisfies condition (D2). Alternatively, the exploration schedule may be randomized, in which case it remains to define conditions on block transition probabilities which ensure (D2). Define the sequence of mappings  $\alpha_n(H_n^x)$  by

$$P_x^\Phi(B_n = 1 \mid H_n^x) = \alpha_n(H_n^x) I\{Z_{n-1} = 0\}, \quad n \geq 1.$$

We may interpret  $\alpha_n(H_n^x)$  as the probability of entering an exploration block at stage  $n$ , given that the MDP was not in an exploration block at stage  $n-1$ . In fact,  $\alpha_n(H_n^x)$  follows from  $p_n^e$  by  $\alpha_n(H_n^x) I\{Z_{n-1} = 0\} = p_n^e(H_n^x) I\{Z_{n-1} = 0\}$ . The following theorem gives a well-defined rate for  $S_n$  in terms of  $\alpha_n(H_n^x)$ .

**THEOREM 6.4.** *Given an MDP with measure (6.3), if w.p.1*

- (i)  $\sum_{i=1}^n \alpha_i(H_i^x) \rightarrow_n \infty$ ,
- (ii)  $\alpha_n(H_n^x) \rightarrow_n 0$ , and
- (iii)  $\limsup_{k \rightarrow \infty} k^{-1}(I_1 + \dots + I_k) = \mu_I < \infty$ ,

then

$$\lim_{n \rightarrow \infty} \frac{S_n}{\sum_{i=1}^n \alpha_i(H_i^x)} = 1.$$

*Proof.* Suppose  $S_n$  is bounded. Then either  $Z_n = 1$  for all large enough  $n$ , or  $Z_n = 0$  for all large enough  $n$ . The first case contradicts (iii), since  $I_k = \infty$  for some finite  $k$ . For the second case, there exists  $N$  such that for  $n > N$ , we have

$$\sum_{i=1}^n P_x^\Phi(B_i = 1 \mid H_i^x) = \sum_{i=1}^n \alpha_n(H_i^x) I\{Z_{n-1} = 0\} \geq \sum_{i=N}^n \alpha_i(H_i^x).$$

The lower bound approaches  $\infty$  as  $n \rightarrow \infty$  by (i). Noting that the process  $(B_1, B_2, \dots)$  is adapted to  $(\sigma(H_2^x), \sigma(H_3^x), \dots)$  we may use Theorem 6.2 to conclude that  $S_n \rightarrow \infty$ ,

leading to a contradiction. We must therefore have  $S_n \rightarrow_n \infty$ , which in turn implies  $\sum_{i=1}^{\infty} P_x^\Phi(B_i = 1 \mid H_i^x) = \infty$ . We then have

$$(6.14) \quad \begin{aligned} \sum_{i=1}^n P_x^\Phi(B_i = 1 \mid H_i^x) &= \sum_{i=1}^n \alpha_i(H_i^x) I\{Z_{i-1} = 0\} \\ &= \sum_{i=1}^n \alpha_i(H_i^x) - \sum_{i=1}^n \alpha_i(H_i^x) I\{Z_{i-1} = 1\}. \end{aligned}$$

Fix  $\epsilon > 0$ . From (ii), (iii) there is  $N_\epsilon$  such that  $S_n^{-1} \sum_{i=1}^n I\{Z_i = 1\} < \mu_I + \epsilon$  and  $\alpha_n(H_n^x) < \epsilon$  for  $n > N_\epsilon$ . The final summation in (6.14) satisfies

$$\sum_{i=1}^n \alpha_i(H_i^x) I\{Z_{i-1} = 1\} < K_\epsilon + \epsilon(\mu_I + \epsilon)S_n$$

for  $n > N_\epsilon$  and some finite  $K_\epsilon$ . Dividing (6.14) by  $S_n$  and letting  $n \rightarrow \infty$  gives

$$\limsup_{n \rightarrow \infty} \left| \frac{\sum_{i=1}^n P_x^\Phi(B_i \mid H_i^x)}{S_n} - \frac{\sum_{i=1}^n \alpha_i(H_i^x)}{S_n} \right| < \epsilon(\mu_I + \epsilon),$$

which proves the theorem using Theorem 6.2, letting  $\epsilon \rightarrow 0$ .  $\square$

It will be convenient to define a class of *randomized* exploration schedules. The essential feature is that transitions into and out of exploration blocks conditioned on the current history remain essentially random.

DEFINITION 6.2. An exploration schedule  $\{p_n^e(H^x); n \geq 1\}$  is randomized if

- (i) there exists a constant  $\gamma < 1$  such that  $p_n^e(H_n^x) \leq \gamma$ ,  $n \geq 1$ , w.p.1;
- (ii) there exists a sequence of constants  $\alpha_n^u$  such that  $\alpha_n(H_n^x) \leq \alpha_n^u$ , and we have

$$\liminf_{n \rightarrow \infty} \frac{\sum_{i=1}^n \alpha_i(H_i^x)}{\sum_{i=1}^n \alpha_i^u} = K_\alpha$$

w.p.1 for a finite constant  $K_\alpha > 0$ .

For convenience set

$$\xi_n = \sum_{i=1}^n \alpha_i^u \text{ and } \bar{\alpha}_n^u = \sup_{m \geq 0} \alpha_{n+m}^u, \quad n \geq 1.$$

The advantages of a randomized exploration schedule might be considered largely mathematical. It easily guarantees sufficient variety of state-action pairs while obeying a well-defined exploration rate. But Remark 6.2 may be equally relevant here. We summarize the rate of a randomized exploration schedule in the following theorem.

THEOREM 6.5. Suppose a randomized exploration schedule (Definition 6.2) satisfies  $\alpha_n^u \rightarrow_n 0$  and  $\xi_n \rightarrow_n \infty$ . Then w.p.1,

$$(6.15) \quad \liminf_{n \rightarrow \infty} \frac{S_n}{\xi_n} > 0$$

and

$$(6.16) \quad \liminf_{n \rightarrow \infty} \frac{\alpha_n^u}{P_x^\Phi(Z_n = 1)} \geq (1 - r_u \gamma),$$

where  $r_u = \limsup_{n \rightarrow \infty} \alpha_{n-1}^u / \alpha_n^u$ .



*Proof.* By hypothesis, and condition (ii) of Definition 6.2, conditions (i)–(ii) of Theorem 6.4 are satisfied. Additionally, condition (i) of Definition 6.2 implies condition (iii) of Theorem 6.4 for some  $\mu_I < \infty$ . Then (6.15) follows directly from Theorem 6.4.

We may then write for  $n \geq 1$ , under Definition 6.2,

$$\begin{aligned} P_x^\Phi(Z_n = 1) &= P_x^\Phi(Z_n = 1 \mid Z_{n-1} = 0)P_x^\Phi(Z_{n-1} = 0) \\ &\quad + P_x^\Phi(Z_n = 1 \mid Z_{n-1} = 1)P_x^\Phi(Z_{n-1} = 1) \\ &\leq \alpha_n^u P_x^\Phi(Z_{n-1} = 0) + \gamma P_x^\Phi(Z_{n-1} = 1) \\ &\leq \alpha_n^u + \gamma P_x^\Phi(Z_{n-1} = 1), \end{aligned}$$

which, applied iteratively, gives

$$(6.17) \quad P_x^\Phi(Z_n = 1) \leq \sum_{i=1}^n \gamma^{n-i} \alpha_i^u.$$

Then (6.16) follows from a direct application of Lemma 2.3 to the upper bound of (6.17).  $\square$

An example of a randomized exploration schedule is easy to construct. We may generate a production schedule  $Z_1, Z_2, \dots$  with  $Z_0 = 0$  from a nonhomogenous Markov chain with transition matrix

$$K_n = \begin{bmatrix} 1 - \alpha_n & \alpha_n \\ 1 - \gamma & \gamma \end{bmatrix}$$

with suitably chosen  $\alpha_n, \gamma$ , governing the transition from  $Z_{n-1}$  to  $Z_n$ , and assume transitions occur independently of process history.

Finally, we will make use of the following lemma.

LEMMA 6.6. *Given an MDP (6.3), if a weight function  $w \in \mathcal{W}(\mathcal{X})$  satisfies (B1), for a randomized exploration schedule (Definition 6.2) we have for  $n \geq 1$ ,  $m \geq 0$ ,*

$$(6.18) \quad \begin{aligned} &E_x^\Phi[w(X_{n+m})I\{Z_{n+m} = 1\} \mid H_n^x] \\ &\leq (w(X_n) + mb_Q)I\{Z_{n-1} = 1\} + \bar{\alpha}_n^u(m+1)(w(X_n) + mb_Q). \end{aligned}$$

*Proof.* We have

$$\{Z_{n+m} = 1\} \subset \{Z_{n-1} = 1\} \cup \left(\bigcup_{j=0}^m \{B_{n+j} = 1\}\right),$$

which implies

$$(6.19) \quad \begin{aligned} E_x^\Phi[w(X_{n+m})I\{Z_{n+m} = 1\} \mid H_n^x] &\leq E_x^\Phi[w(X_{n+m}) \mid H_n^x]I\{Z_{n-1} = 1\} \\ &\quad + \sum_{j=0}^m E_x^\Phi[w(X_{n+m})I\{B_{n+j} = 1\} \mid H_n^x]. \end{aligned}$$

To analyze the terms in (6.19), we write, for  $j \geq 0$ ,

$$(6.20) \quad \begin{aligned} E_x^\Phi[w(X_{n+j})I\{B_{n+j} = 1\} \mid H_n^x] &= E_x^\Phi[E_x^\Phi[w(X_{n+j})I\{B_{n+j} = 1\} \mid H_{n+j}^x] \mid H_n^x] \\ &= E_x^\Phi[E_x^\Phi[I\{B_{n+j} = 1\} \mid H_{n+j}^x]w(X_{n+j}) \mid H_n^x] \\ &\leq E_x^\Phi[\alpha_{n+j}^u w(X_{n+j}) \mid H_n^x] \end{aligned}$$

from Definition 6.2. First suppose  $m = 0$ . Then (6.18) follows by applying (6.20), with  $j = 0$ , to (6.19) and then noting that  $X_n$  is  $\sigma(H_n^x)$ -measurable and that  $\alpha_n^u \leq \bar{\alpha}_n^u$ . Then suppose  $m \geq 1$ . Setting  $j = 0, \dots, m-1$ , we have

$$\begin{aligned} E_x^\Phi[w(X_{n+m})I\{B_{n+j} = 1\} \mid H_n^x] &= E_x^\Phi[E^\Phi[w(X_{n+m})I\{B_{n+j} = 1\} \mid H_{n+m-1}^a] \mid H_n^x] \\ &= E_x^\Phi[E^\Phi[w(X_{n+m}) \mid H_{n+m-1}^a]I\{B_{n+j} = 1\} \mid H_n^x] \\ (6.21) \qquad \qquad \qquad &\leq E_x^\Phi[(w(X_{n+m-1}) + b_Q)I\{B_{n+j} = 1\} \mid H_n^x] \end{aligned}$$

by (B1), noting that  $B_{n+j}$  is  $\sigma(H_{n+m-1}^a)$ -measurable if  $j \leq m-1$ . Iterating (6.21) and applying (6.20) gives

$$\begin{aligned} E_x^\Phi[w(X_{n+m})I\{B_{n+j} = 1\} \mid H_n^x] &\leq E_x^\Phi[(w(X_{n+j}) + (m-j)b_Q)I\{B_{n+j} = 1\} \mid H_n^x] \\ &\leq E_x^\Phi[\alpha_{n+j}^u(w(X_{n+j}) + (m-j)b_Q) \mid H_n^x]; \end{aligned}$$

hence (6.19) implies

$$\begin{aligned} E_x^\Phi[w(X_{n+m})I\{Z_{n+m} = 1\} \mid H_n^x] &\leq E_x^\Phi[w(X_{n+m}) \mid H_n^x]I\{Z_{n-1} = 1\} \\ (6.22) \qquad \qquad \qquad &+ \sum_{j=0}^m \alpha_{n+j}^u (E_x^\Phi[w(X_{n+j}) \mid H_n^x] + (m-j)b_Q). \end{aligned}$$

Then apply (4.4) to (6.22), which yields (6.18).  $\square$

**6.4. Control policy definition.** We are now in a position to define an adaptive control policy for which a bound on total regret may be calculated, which consists of the following elements:

- (E1) *Control model.* We have an MCM  $\pi = (\mathcal{K}, Q, R, \beta)$  with  $\beta < 1$ . There is a contractive solution space  $\mathcal{V}^w$  and an OSDP based on control function  $\phi_\pi^* \in \mathcal{K}^f$ . The definition includes the observation process defined in (M7) and an exploration schedule defined in (M8). The policy  $\Phi$  assumes the form given in (6.2), and hence the process is governed by the measure defined by (6.3).
- (E2) *Weight function.* The weight function associated with  $\mathcal{V}^w$  satisfies (B1)–(B2). We will further assume that  $\inf_x w(x) > 0$ . If not, we may add a positive constant to  $w(x)$ , after which (B1)–(B2) will still hold. In this case (B1) implies  $\eta_Q^w < \infty$ .
- (E3) *PM.* We assume the existence of a PM  $M^\Theta$  which conforms to  $\pi$  and  $\mathcal{V}^w$ . We assume that  $\Theta \subset \mathbb{R}^k$  for finite dimension  $k$  and take the metric  $d$  to be based on the  $\ell_1$  norm. The true parameter is denoted by  $\theta' = (\theta'_1, \dots, \theta'_k)$ . We further assume that  $D_r^w(R, R^\theta) < \infty$  and  $D_q^w(Q, Q^\theta) < \infty$  for all  $\theta \in \Theta$ . In addition, we assume that Lipschitz continuity conditions (5.3) and (5.4) hold for finite constants  $C_B^q, C_B^r$ . By Theorem 5.1, this follows from (C1)–(C2).
- (E4) *CEA.* For each  $\theta \in \Theta$ , the conditions of Theorem 3.5 follow from (E1)–(E3), so we may assert  $T_{\pi^\theta} : \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w) \rightarrow \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$ . Suppose for each  $\theta \in \Theta$  there exists elements  $v_{0,\theta}, V_{\pi^\theta}^* \in \mathcal{V} \cap F(\mathcal{X}, \|\cdot\|_w)$  for which  $\|T_{\pi^\theta}^i v_{0,\theta} - V_{\pi^\theta}^*\|_w \rightarrow_i 0$ . Then we may construct a CEA (Definition 3.7) from any sequence  $\hat{\theta}_1, \hat{\theta}_2, \dots$ .
- (E5) *Parametric estimation.* Assume that the informative subset  $\mathcal{K}(\theta_j)$  is nonempty for each  $\theta_j$ . At stage  $n$  we may calculate estimate  $\hat{\theta}_n = (\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,k})$ , using the estimators defined by (6.6). We also assume that condition (D1) holds.

- (E6) *Exploratory control.* We assume  $p_n^e$  is a randomized exploration schedule (Definition 6.2) which satisfies the conditions of Theorem 6.5, from which (D2) follows. We assume  $\Phi^e$  satisfies condition (D3), hence Theorem 6.3 applies, so that given (6.7) we may conclude

$$d(\hat{\theta}_n, \theta') = o\left(S_n^{-1/2+\epsilon}\right)$$

for any  $\epsilon > 0$ . We further assume that in Theorem 6.5  $(1 - r_u \gamma) > 0$ .

- (E7) *Optimal control.* At stage  $n$  an estimate  $\hat{\phi}_n$  of  $\phi_\pi^*$  is calculated based on model  $\hat{\pi}_n = \pi^{\hat{\theta}_{n-1}}$ . The lag of one stage is due to the fact that  $\hat{\theta}_n$  is dependent on observation  $O_n$ , which is observed after control  $A_n$  is applied, while  $\hat{\theta}_{n-1}$  may be observed before. A default estimate for  $\hat{\theta}_0$  is assumed to be available. The value function  $V_{\hat{\pi}_n}^*$  given in (E4) is calculated. For each  $x \in \mathcal{X}$  a minimization operation is undertaken for  $T_{\hat{\pi}_n}^a V_{\hat{\pi}_n}^*(x, a)$  over  $a \in \mathcal{K}_x$ , with  $\hat{\phi}_n(x)$  set to the solution. The actual minimum may or may not be achievable. In either case we set

$$\varepsilon_n(x) = T_{\hat{\pi}_n}^a V_{\hat{\pi}_n}^*(x, \hat{\phi}_n(x)) - T_{\hat{\pi}_n} V_{\hat{\pi}_n}^*(x).$$

In principle, this quantity may be made arbitrarily small if the minimization cannot be exact. The optimal component of the control  $\Phi$  is then given by

$$\Phi_n^o(E_a \mid H_n^x) = I\{\hat{\phi}_n(X_n) \in E_a\}$$

for any  $E_a \in \mathcal{B}(\mathcal{A})$ .

Our final task is to estimate the combined regret of the MDP defined by (E1)–(E7). This is done in the form of an upper bound on  $\Lambda_n^\Phi(H_n^x)$ . In the discussion which follows,  $c_1, \dots, c_{12}$  are finite positive constants which do not depend on  $n$ . Their exact values are omitted for the sake of clarity.

For any MDP with a contractive solution space satisfying (B1)–(B2), an upper bound is directly obtainable from (4.3) and (4.6) of Theorem 4.1 and its proof. In particular, there exist constants  $c_1, c_2$  for which

$$(6.23) \quad \Lambda_n^\Phi(H_n^x) \leq V_\pi^*(X_n) + c_1 w(X_n) + c_2 \leq V_\pi^*(X_n) + \left(c_1 + \frac{c_2}{w^*}\right) w(X_n),$$

where  $w^* = \inf_x w(x)$ . A refinement of this bound for the adaptive policy discussed here follows.

**THEOREM 6.7.** *For the MDP defined in (E1)–(E7) the following bound on regret holds:*

$$(6.24) \quad \begin{aligned} \Lambda_x^\Phi(H_n^x) \leq & V_\pi^*(X_n) + (w(X_n) + 1)(c_9 I\{Z_{n-1} = 1\} + c_{10} \bar{\alpha}_n^u \\ & + c_{11} d(\hat{\theta}_{n-1}, \theta') + c_{12} (M_{n-1}(\theta))^{-1}) \end{aligned}$$

for some finite constants  $c_9, c_{10}, c_{11}, c_{12}$ .

*Proof.* First, bound the error of the value function estimate of the CEA. Under the Lipschitz property of  $M^\Theta$  we have

$$(6.25) \quad \max\left(D_q^w(Q^{\hat{\theta}_n}, Q), D_r^w(R^{\hat{\theta}_n}, R)\right) \leq C_B d(\hat{\theta}_n, \theta'), \quad n \geq 1,$$

for some finite  $C_B$ . Then Theorem 3.7 may be re-expressed as

$$(6.26) \quad \begin{aligned} & \|V_{\hat{\pi}_n}^* - V_{\pi}^*\|_w \left(1 - c_3 D_q^w(Q^{\hat{\theta}_{n-1}}, Q)\right) \\ & \leq c_4 \max \left(D_r^w(R^{\hat{\theta}_{n-1}}, R), D_q^w(Q^{\hat{\theta}_{n-1}}, Q)\right), \quad n \geq 1, \end{aligned}$$

for finite positive constants  $c_3, c_4$ , making use of the inequality  $\|V_{\hat{\pi}_n}^*\|_w \leq \|V_{\hat{\pi}_n}^* - V_{\pi}^*\|_w + \|V_{\pi}^*\|_w$ , and the fact that  $\|V_{\pi}^*\|_w < \infty$ . Given (6.25) and (6.26) we may then identify constants  $c_5, c_6$  such that

$$(6.27) \quad \|V_{\hat{\pi}_n}^* - V_{\pi}^*\|_w \leq c_5 d(\hat{\theta}_{n-1}, \theta') \quad \text{when } d(\hat{\theta}_{n-1}, \theta') \leq c_6.$$

With sufficient regularity conditions we can set the quantity  $\varepsilon_n \equiv 0$  defined in (E7); otherwise, we assume that  $\varepsilon_n$  may be made as small as we wish. Then by Theorem 4.2 and (6.27) there is a constant  $c_7$  such that

$$\lambda_{\pi}(x, \hat{\phi}_n(x)) \leq c_7 w(x) d(\hat{\theta}_{n-1}, \theta') \quad \text{when } d(\hat{\theta}_{n-1}, \theta') \leq c_6$$

if we may ensure, say, that  $\|\varepsilon_n\|_w \leq d(\hat{\theta}_{n-1}, \theta')$ . Then for fixed  $n$ ,  $m \geq 0$  consider a term of the form

$$(6.28) \quad \begin{aligned} \lambda_{\pi}(X_{n+m}, A_{n+m}) & \leq \lambda_{\pi}(X_{n+m}, A_{n+m}) I\{Z_{n+m} = 1\} \\ & + \lambda_{\pi}(X_{n+m}, A_{n+m}) I\{Z_{n+m} = 0 \wedge d(\hat{\theta}_{n+m-1}, \theta') \leq c_6\} \\ & + \lambda_{\pi}(X_{n+m}, A_{n+m}) I\{d(\hat{\theta}_{n+m-1}, \theta') > c_6\} \\ & = B_{n+m}^1 + B_{n+m}^2 + B_{n+m}^3 \end{aligned}$$

and consider the problem of estimating  $E_x^{\Phi}[\lambda_{\pi}(X_{n+m}, A_{n+m}) \mid H_n^x]$ . For term  $B_{n+m}^1$ , by condition (B2) and Lemma 6.6 we may write

$$(6.29) \quad \begin{aligned} E_x^{\Phi}[B_{n+m}^1 \mid H_n^x] & \leq b_{\pi} E_x^{\Phi}[w(X_{n+m}) I\{Z_{n+m} = 1\} \mid H_n^x] \\ & \leq b_{\pi} ((w(X_n) + mb_Q) I\{Z_{n-1} = 1\} + \bar{\alpha}_n^u(m+1)(w(X_n) + mb_Q)). \end{aligned}$$

For term  $B_{n+m}^2$  note that  $Z_{n+m} = 0$  implies  $A_{n+m} = \hat{\phi}_{n+m}(X_{n+m})$ , so that

$$(6.30) \quad B_{n+m}^2 \leq c_7 w(X_{n+m}) d(\hat{\theta}_{n+m-1}, \theta').$$

Finally, we may write

$$(6.31) \quad B_{n+m}^3 \leq b_{\pi} c_6^{-1} w(X_{n+m}) d(\hat{\theta}_{n+m-1}, \theta').$$

By applying Lemma 6.1 to each of the  $k$  terms of  $d(\hat{\theta}_{n+m-1}, \theta')$  we have, for  $m \geq 0$ ,

$$(6.32) \quad \begin{aligned} E_x^{\Phi}[w(X_{n+m}) d(\hat{\theta}_{n+m-1}, \theta') \mid H_n^x] & = E_x^{\Phi}[E_x^{\Phi}[w(X_{n+m}) d(\hat{\theta}_{n+m-1}, \theta') \mid H_n^a] \mid H_n^x] \\ & \leq E_x^{\Phi} \left[ (w(X_n) + mb_Q) d(\hat{\theta}_{n-1}, \theta') \right. \\ & \quad \left. + \frac{m\tau w(X_n) + b_Q(\nu^{1/2} + \tau)m(m-1)/2}{M_n(\theta)} \mid H_n^x \right] \\ & \leq (w(X_n) + mb_Q) d(\hat{\theta}_{n-1}, \theta') \\ & \quad + \frac{m\tau w(X_n) + b_Q(\nu^{1/2} + \tau)m(m-1)/2}{M_{n-1}(\theta)}, \end{aligned}$$

noting that  $M_{n-1}(\theta) \leq M_n(\theta) \leq M_n(\theta_j)$  and that  $X_n, \hat{\theta}_{n-1}$  and  $M_{n-1}(\theta)$  are each  $\sigma(H_n^x)$ -measurable.

Combining (6.28)–(6.32) gives, for some finite constant  $c_8$ ,

$$(6.33) \quad \begin{aligned} E_x^\Phi[\lambda_\pi(X_{n+m}, A_{n+m}) \mid H_n^x] &\leq b_\pi(w(X_n) + mb_Q)I\{Z_{n-1} = 1\} \\ &\quad + b_\pi \bar{\alpha}_n^u(m+1)(w(X_n) + mb_Q) \\ &\quad + c_8(w(X_n) + mb_Q)d(\hat{\theta}_{n-1}, \theta') \\ &\quad + c_8 \frac{m\tau w(X_n) + b_Q(\nu^{1/2} + \tau)m(m-1)/2}{M_{n-1}(\theta)}. \end{aligned}$$

Applying Theorem 4.1 to inequality (6.33) yields after taking suitable summations (6.24).  $\square$

The importance of (6.24) is that it decomposes the bound on regret into that attributable to exploration and estimation. Furthermore, each source of regret can be explicitly related to the exploration rate, permitting selection of that rate. By Theorems 6.3 and 6.5 we have, for  $\epsilon > 0$ ,

$$(6.34) \quad d(\hat{\theta}_n, \theta') = o\left(\xi_n^{-1/2+\epsilon}\right).$$

For the sake of argument, suppose  $\alpha_n^u \propto n^{-r}$  for  $0 < r \leq 1$ . Then  $\xi_n \propto n^{1-r}$  for  $r < 1$  and  $\xi_n \propto \log(n)$  for  $r = 1$ . This gives  $d(\hat{\theta}_{n-1}, \theta') = o(n^{(r-1)/2+\epsilon})$ , and  $M_n(\theta)^{-1} = O(n^{r-1})$  for  $r < 1$  and  $d(\hat{\theta}_{n-1}, \theta') = o(\log(n)^{-1/2+\epsilon})$ , and  $M_n(\theta)^{-1} = O(\log(n)^{-1})$  for  $r = 1$ , which yields

$$\Lambda^\Phi(H_n^x) = V_\pi^*(X_n) + (w(X_n) + 1) \left( c_9 I\{Z_{n-1} = 1\} + O(n^{-r}) + o(n^{(r-1)/2+\epsilon}) \right)$$

for  $r < 1$  and

$$\Lambda^\Phi(H_n^x) = V_\pi^*(X_n) + (w(X_n) + 1) \left( c_9 I\{Z_{n-1} = 1\} + O(n^{-1}) + o(\log(n)^{-1/2+\epsilon}) \right)$$

for  $r = 1$ . By Theorem 6.5 we have  $P_x^\Phi(Z_n = 1) = O(\alpha_n^u) = O(n^{-r})$ ; thus we may take the overall contribution to regret due to exploration to be  $O(n^{-r})$ , and that due to estimation to be  $o(n^{(r-1)/2+\epsilon})$  for  $r < 1$  or  $o(\log(n)^{-1/2+\epsilon})$  for  $r = 1$ . The upper bound is optimized by setting  $r = 1/3$ ; that is, we may devise an adaptive policy at which regret is accrued at rate  $o(n^{-1/3+\epsilon})$ .

**7. Conclusion and further work.** We have formulated the basis for a general theory of approximate iterative algorithms, giving a comprehensive set of convergence results following from properties given in terms of normed linear spaces.

The value of such a theory was illustrated by an extended application to the problem of model-based approximate and adaptive control of Markov decision processes. The theory permitted a sharpening of known convergence rates, applied to a more general model. Additionally, bounds on regret for adaptive controls with forced exploration are calculated in terms of a stagewise exploration rate. This permits the determination of an optimal choice of exploration rate.

Although this work establishes a number of important mathematical principles, there are a number of areas in which further refinement would make the theory more relevant to the practical design of control policies.

Attention was restricted to discounted cost models. Alternative norms or operators which induce the contraction property have been developed for the average cost

criterion control model (see, for example, [18, 19, 3, 36]). Application of this theory to average cost models would seem to be a logical extension.

In this work, parametric estimators have been restricted to a class of sample averages, whereas a much broader class of recursive estimators is required for general applications. A suitable extension should be possible for estimators with martingale properties, permitting natural conditions under which (6.7) will hold. The next task would be to derive a result comparable to Lemma 6.1. Once these conditions are met, Theorem 6.7 could then be argued as before. The theory could then conform more to the standard *partially observed control model* discussed in [17, 12, 2].

We briefly reiterate a number of points made earlier. The exploratory control presented here is mathematically convenient, but in practice one would expect any number of stringent constraints or specialized structures. For example, in [12] estimation of an unknown parameter within a repair system model is made specifically at replacement times, providing a natural example of the informative subset discussed in section 6.2. It is hoped that the present theory can be used to carefully analyze alternative exploratory controls of a variety of structures.

The rate for regret reported in section 6.4 is optimal for the type of adaptive control considered here. The fact that a strictly better cumulative regret of order  $a_n \log(n)$ , for any  $a_n \rightarrow \infty$ , can be achieved for some models using a bandit theory approach [25] suggests a number of possibilities. Perhaps the regret rate reported here can be further sharpened for some subset of the models presented. Failing that, it would still be of interest to consider the possibility that an alternative model-based control can be designed which approaches the faster bandit theory rate. It could be expected that such a control would be more stable than one based on randomization over control policies.

**Acknowledgments.** The author wishes to thank the referees for their careful reading of the manuscript and for their guidance during the revision process. The author also wishes to thank Ronald Shonkwiler and Franklin Mendivil for their generous advice.

#### REFERENCES

- [1] A. ALMUDEVAR, *A dynamic programming algorithm for the optimal control of piecewise deterministic Markov processes*, SIAM J. Control Optim., 40 (2001), pp. 525–539.
- [2] A. ARAPOSTATHIS, V. S. BORKAR, E. FERNÁNDEZ-GAUCHERAND, M. K. GHOSH, AND S. I. MARCUS, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344.
- [3] D. P. BERTSEKAS, *Dynamic Programming and Optimal Control*, Volumes I and II, Athena Scientific, Belmont, 2001.
- [4] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
- [5] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, 1996.
- [6] R. N. BHATTACHARYA AND M. MAJUMDAR, *Controlled semi-Markov models: The discounted case*, J. Statist. Plann. Inference, 21 (1989), pp. 365–381.
- [7] G. CYBENKO, R. GRAY, AND K. MOIZUMI, *Q-learning: A tutorial and extensions*, in Proceedings of the First International Conference on Mathematics of Neural Networks (Oxford University, 1995), Kluwer, Norwell, MA, 1997, pp. 24–33.
- [8] L. E. DUBINS AND D. A. FREEDMAN, *A sharper form of the Borel-Cantelli lemma and the strong law*, Ann. Math. Statist., 36 (1965), pp. 800–807.
- [9] M. DUFLO, *Random Iterative Models*, Appl. Math., 34, Springer-Verlag, Berlin, Heidelberg, 1997.

- [10] E. EVAN-DAR AND Y. MANSOUR, *Learning rates for Q-learning*, J. Mach. Learn. Res., 5 (2004), pp. 1–25.
- [11] A. FEDERGRUEN AND P. J. SCHWEITZER, *Nonstationary Markov decision problems with converging parameters*, J. Optim. Theory Appl., 34 (1981), pp. 207–241.
- [12] E. FERNÁNDEZ-GAUCHERAND, A. ARAPOSTATHIS, AND S. I. MARCUS, *Analysis of an adaptive control scheme for a partially observed controlled Markov chain*, IEEE Trans. Automat. Control, 38 (1993), pp. 987–993.
- [13] E. FISCHER, *Intermediate Real Analysis*, Springer-Verlag, New York, 1983.
- [14] J. HALTON, *Sequential Monte Carlo techniques for the solution of linear systems*, J. Sci. Comput., 9 (1994), pp. 213–257.
- [15] J. HALTON, *Sequential Monte Carlo techniques for solving non-linear systems*, Monte Carlo Methods Appl., 12 (2006), pp. 113–141.
- [16] O. HERNÁNDEZ-LERMA AND S. I. MARCUS, *Adaptive control of discounted Markov decision chains*, J. Optim. Theory Appl., 46 (1985), pp. 227–235.
- [17] O. HERNÁNDEZ-LERMA, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.
- [18] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Discrete-Time Markov Control Processes: Basic Optimality*, Springer, New York, 1996.
- [19] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Further Topics on Discrete-Time Markov Control Processes*, Springer, New York, 1999.
- [20] K. HINDERER, *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Oper. Res. Math. Systems 33, Springer-Verlag, New York, 1970.
- [21] E. ISAACSON AND H. B. KELLER, *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.
- [22] M. KEARNS AND S. SINGH, *Finite-sample convergence rates for Q-learning and indirect algorithms*, Neural Information Processing Systems, 11 (1998), pp. 996–1002.
- [23] M. KEARNS AND S. SINGH, *Near-optimal reinforcement learning in polynomial time*, Machine Learning, 49 (2003), pp. 209–232.
- [24] P. R. KUMAR AND P. VARAIYA, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.
- [25] T. L. LAI AND S. YAKOWITZ, *Machine learning and nonparametric bandit theory*, IEEE Trans. Automat. Control, 30 (1995), pp. 1199–1209.
- [26] E. L. LEHMANN AND G. CASELLA, *Theory of Point Estimation*, 2nd ed., Springer, New York, 1998.
- [27] S. A. LIPPMAN, *On dynamic programming with unbounded rewards*, Management Sci., 21 (1975), pp. 1225–1233.
- [28] A. W. MOORE AND C. G. ATKESON, *Prioritized sweeping: Reinforcement learning with less data and less real time*, Machine Learning, 13 (1993), pp. 103–130.
- [29] R. MUNOS, *Performance bounds in  $L_p$ -norm for approximate value iteration*, SIAM J. Control Optim., 46 (2007), pp. 541–561.
- [30] M. SCHÄL, *Estimation and control in discounted stochastic dynamic programming*, Stochastics, 20 (1987), pp. 51–71.
- [31] R. S. SUTTON AND A. G. BARTO, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.
- [32] CS. SZEPESVÁRI, *The asymptotic convergence rate of Q-learning*, Neural Information Processing Systems, 10 (1997), pp. 1064–1070.
- [33] S. B. THRUN, *Efficient Exploration in Reinforcement Learning*, Technical report CMU-CS-92-102, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1992.
- [34] J. N. TSITSIKLIS, *Asynchronous stochastic approximation and Q-learning*, Machine Learning, 16 (1994), pp. 185–202.
- [35] J. A. E. E. VAN NUNEN AND J. WESSELS, *A note on dynamic programming with unbounded rewards*, Management Sci., 24 (1978), pp. 576–580.
- [36] O. VEGA-AMAYA, *The average cost optimality equation: A fixed point approach*, Bol. Soc. Mat. Mexicana, 9 (2003) pp. 185–195.
- [37] C. I. C. H. WATKINS, *Learning from Delayed Rewards*, Ph.D. thesis, Cambridge University, Cambridge, UK, 1989.
- [38] C. I. C. H. WATKINS AND P. DAYAN, *Q-learning*, Machine Learning, 8 (1992), pp. 279–292.

## DIRICHLET PROBLEMS FOR SOME HAMILTON–JACOBI EQUATIONS WITH INEQUALITY CONSTRAINTS\*

JEAN-PIERRE AUBIN<sup>†</sup>, ALEXANDRE M. BAYEN<sup>‡</sup>, AND PATRICK SAINT-PIERRE<sup>§</sup>

**Abstract.** We use viability techniques for solving Dirichlet problems with inequality constraints (obstacles) for a class of Hamilton–Jacobi equations. The hypograph of the “solution” is defined as the “capture basin” under an auxiliary control system of a target associated with the initial and boundary conditions, viable in an environment associated with the inequality constraint. From the tangential condition characterizing capture basins, we prove that this solution is the unique “upper semicontinuous” solution to the Hamilton–Jacobi–Bellman partial differential equation in the Barron–Jensen/Frankowska sense. We show how this framework allows us to translate properties of capture basins into corresponding properties of the solutions to this problem. For instance, this approach provides a representation formula of the solution which boils down to the Lax–Hopf formula in the absence of constraints.

**Key words.** Hamilton–Jacobi equations, viability theory, optimal control, traffic modeling

**AMS subject classifications.** 49J24, 49L25, 90B20, 58C06

**DOI.** 10.1137/060659569

### 1. Introduction.

**1.1. Motivation.** This article is motivated by macroscopic fluid models of highway traffic, following the pioneering work of Lighthill and Whitham [64] and Richards [78]. In their original work, the authors modeled highway traffic flow with a first order hyperbolic *partial differential equation* with concave flux function, called the Lighthill–Whitham–Richards (partial differential) equation. This model is the seminal model for numerous highway traffic flow studies available in the literature today [2, 45, 46, 63, 33, 87, 31]. It models the evolution of the density of vehicles on a highway by a conservation law, in which the mathematical model of the flux function inside the conservation law results from empirical measurements [60].

Solutions to such equations may have shocks (they are set-valued maps), which model abrupt changes in vehicle density on the highway [2], and only model physical phenomena to a certain degree. Hence discontinuous selections of these solutions are investigated, for instance, the *entropy solution* [2] of Oleinik [73], which is acknowledged to be the proper weak solution of this problem. There has been an extensive literature on this problem, of which we single out the work of Bardos, Leroux, and Nédélec [24]; see also Strub and Bayen [83].

Very few results applicable to highway traffic are available for control of first order hyperbolic conservation laws. Differential flatness [50] has been successfully applied to the Burgers equation (and therefore to the Lighthill–Whitham–Richards equation) in [75] order to avoid the formation of such shockwaves. This analysis does not so far

---

\*Received by the editors May 10, 2006; accepted for publication (in revised form) April 3, 2008; published electronically September 8, 2008.

<http://www.siam.org/journals/sicon/47-5/65956.html>

<sup>†</sup>LASTRE (Laboratoire d’Applications des Systèmes Tychastiques Régulés), 14, rue Domat, F-75005 Paris, France (aubin.jp@gmail.com, <http://lastre.asso.fr/aubin>).

<sup>‡</sup>Corresponding author. Department of Civil and Environmental Engineering, University of California at Berkeley, Davis Hall 711, Berkeley, CA 94720-1710 (bayen@ce.berkeley.edu).

<sup>§</sup>Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France (patrick.saint.pierre@gmail.com).



extend to the presence of shocks. Lyapunov-based techniques have also been applied to the Burgers equation [62]. Adjoint-based methods have been successfully applied to networks of Lighthill–Whitham–Richards equations in [57]; these results seem so far the most promising, but they do not have guarantees to provide an optimal control policy. Questions of interest in controlling first order partial differential equations [66, 74, 82, 86], and in particular, Lighthill–Whitham–Richards equations, are still open and difficult to solve due to the presence of shocks occurring in the solutions of these partial differential equations [3, 20, 21, 32, 36, 42, 47, 48, 49, 58, 59, 61].

In order to alleviate the technical difficulties resulting from shocks present in solutions of the Lighthill–Whitham–Richards equation, an alternate formulation consists in considering the *cumulated number of vehicles*, widely used in the transportation literature as well [70, 71, 72]. The cumulative number of vehicles can be thought of as a primitive of the density over space. Formally, the evolution of the cumulated number  $\mathbf{N}(t, x)$  of vehicles is the solution of a *Hamilton–Jacobi (partial differential) equation* of the form

$$\frac{\partial \mathbf{N}(t, x)}{\partial t} + \psi \left( \frac{\partial \mathbf{N}(t, x)}{\partial x} \right) = \psi(v(t)),$$

where the flux function  $\psi$  appearing in this Hamilton–Jacobi equation is in fact concave as shown by the empirically measured flux function of the Lighthill–Whitham–Richards equation [64, 78, 24, 83]. The function  $v(\cdot)$  will be regarded as a control of the Hamilton–Jacobi equation in forthcoming studies. It could, for example, model the inflow of vehicles at the entrance of a stretch of highway. It is a given datum in this paper.

The solution of this Hamilton–Jacobi equation has no shocks but is not necessarily differentiable. It is only upper semicontinuous. Actually, the nondifferentiability of the cumulated number of vehicles is closely related to the presence of the shocks of the solution to the Lighthill–Whitham–Richards equation (see, for instance, [39, 40, 41]).

Since the Lighthill–Whitham–Richards equation and the Hamilton–Jacobi equation model the same physical phenomenon, and since both formulations are equivalently used in the highway transportation literature, we single out in this paper the study of the evolution of the cumulated number of vehicles in order to leverage the extensive knowledge of Hamilton–Jacobi equations for which control and viability techniques can be applied [65, 67, 68, 69, 76, 77, 81].

**1.2. Contributions of the paper.** We shall revisit this Hamilton–Jacobi equation by answering new questions as follows:

- introducing a nontrivial right-hand side;
- involving Dirichlet conditions;
- and, above all, imposing *inequality constraints* on the solution, for instance, upper bounds on the cumulated number of vehicles, depending on time and space variables.

For this purpose, we suggest using a novel point of view based on the concept of capture basin of a target viable in an environment extensively studied in the framework of viability theory; i.e., given a closed subset of a finite dimensional vector space regarded as an environment, a closed subset of this environment considered as a target and a control system, the viable capture basin is the subset of initial states of the environment from which starts at least one evolution governed by the control system viable in the environment until the finite time when it reaches the target (see Definition 3.3). It happens that the *hypograph* of the solution to the Hamilton–Jacobi

equation satisfying initial and Dirichlet conditions as well as inequality constraints is the capture basin of an auxiliary target (involving initial and boundary conditions) viable in an auxiliary environment (involving inequality constraints) under an auxiliary control system (involving the flux function of the Hamilton–Jacobi equation).

Hence, anticipating this property, we define the *viability hypsolution* of the Dirichlet problem for this Hamilton–Jacobi equation with constraints from this property as being a viable capture basin (see Definition 4.1). Then we proceed by translating properties of viable capture basins (see [7], for instance) in the language of partial differential equations for this particular case. We shall prove that the viability hypsolution

1. is the *unique* generalized solution in the Barron–Jensen/Frankowska sense<sup>1</sup> (a weaker concept of viscosity solution introduced by Crandall, Evans, and Lions in [44, 43] for continuous solutions adapted to the case when solution is only semicontinuous): Theorem 9.1;
2. is equivalently the *unique* upper semicontinuous solution in the contingent Frankowska sense:<sup>2</sup> Theorem 8.1;
3. satisfies the sup-linearity property and depends “hypocontinuously” on the initial and Dirichlet conditions;
4. is represented by the Lax–Hopf formula [1] (see Theorem 5.1) in the absence of inequality constraints, a more involved representation formula (see Theorem 5.5) in the presence of inequality constraints, upper estimates (maximum principle; see Proposition 5.3), and lower estimates (see Proposition 5.4).

The results presented in this article have since been applied to highway traffic data [30, 29], using available algorithms to solve, in particular, viability problems numerically [80, 37, 38].

**1.3. Outline of the paper.** In order to make the paper more readable, section 3 gathers some definitions, notations, and basic prerequisites of viability theory and convex analysis for the convenience of readers who are not familiar with these topics. We then state the problem and the main assumptions, which will not be repeated. We next define the viability hypsolution to the nonhomogeneous Dirichlet/initial value problem for our class of Hamilton–Jacobi equations under inequality constraints as the capture basin of a target summarizing the Dirichlet/initial data viable in a target associated with inequality constraints. Then, we translate the properties of capture basins into the viability hypsolution, starting with a general representation formula providing Lax–Hopf formulas in the absence of inequality constraints. We next check that the viability hypsolution satisfies the Dirichlet and initial conditions as well as the inequality constraints. The last three sections are devoted to the proof that the viability hypsolution is a solution to the Hamilton–Jacobi partial differential equation

---

<sup>1</sup>Frankowska proved that the epigraph of the value function of an optimal control problem—assumed to be only lower semicontinuous—is semipermeable (i.e., invariant and backward viable) under a (natural) auxiliary system. Furthermore, when it is continuous, its epigraph is viable and its hypograph invariant [53, 54, 56]. By duality, the latter property is equivalent to the fact that the value function is a viscosity solution of the associated Hamilton–Jacobi equation in the sense of Crandall and Lions. See also [26, 22, 8] for more details. Such concepts have been extended to solutions of systems of first order partial differential equations without boundary conditions by Frankowska and the first author (see [14, 15, 16, 17, 18, 19] and Chapter 8 of [5]). See also [11, 12].

<sup>2</sup>Contingent inequalities were first introduced in [4] for characterizing Lyapunov functions and value functions of a class of control problems and later, used in [84, 85] to investigate infinitesimal properties of Lyapunov and value functions in differential games. The “backward inequality” was introduced for the first time in [55, 56] to prove uniqueness of lower semicontinuous solutions of Hamilton–Jacobi–Bellman equations. See also [23, 25, 27, 28, 34, 35].

in two equivalent dual generalized senses by translating both the viability theorem and the invariance theorem characterizing the capture basin in terms of either tangential conditions or normal conditions, as it was done in a long series of papers by Frankowska. Using tangential conditions, we express the viability hyposolution as a solution to the Hamilton–Jacobi partial differential equation couched in terms of contingent hypoderivatives, whereas using normal conditions, we characterize it in terms of superdifferentials, as it was done independently by Barron–Jensen and Frankowska, in the spirit of nonsmooth analysis and viscosity solutions. The presence of inequality constraints complicates the technical formulation of the concept of solution at points where the solution touches the boundary of the constraint, above all in the superdifferential formulation, justifying the reason why we conclude this paper with this dual characterization.

**2. Statement of the problem.** This section states the problem of interest for this article. Section 3 provides all prerequisites for the concepts used in the later sections for a reader not familiar with viability theory and convex analysis. No prerequisites from viability theory are required to read this section.

**2.1. Notation.** For notational convenience, and in order to avoid multiplication of the letters used in the article, we have used the letters  $\sigma$  and  $\tau$  in several different ways, which depend on context; i.e., for  $\sigma$ , we have the following definitions, based on context:

- Support function for some compact convex subset  $A \subset X$ , where  $\sigma_A(v) := \sigma(A, v) := \sup_{u \in A} \langle u, v \rangle$  is the support function of  $A$ . Note that the first argument of  $\sigma$  is a set, while the second is a vector.
- Auxiliary min inf function  $\sigma(t, x, u) := \min(t, \tau(x, u))$ , defined in Theorem 5.1. Note that this function has three arguments, which are one scalar  $t$  and two vectors  $x$  and  $u$  of  $X$ .
- Auxiliary min inf functional  $\sigma(t, x, u(\cdot)) = \min(t, \tau(x, u(\cdot)))$ , defined in the proof of Theorem 5.1. Note that this function has three arguments, which are one scalar  $t$ , one vector  $x \in X$ , and function  $u(\cdot)$  (measurable, integrable).

Similarly for the notation  $\tau$  is used as a

- Dummy variable  $\tau$ , for example, in integrals. Note that  $\tau$  has no argument.
- Pseudotime  $\tau(t)$ , for example, in (13). Note that  $\tau(t)$  has one argument  $t$  which corresponds to the running time of the corresponding differential inclusion.
- Auxiliary inf function  $\tau(x, u) := \inf_{x+tu \notin K} t$ , defined in Theorem 5.1. Note that this function has two inputs, which are vectors  $x$  and  $u$  of  $X$ .
- Auxiliary inf functional  $\tau(x, u(\cdot)) := \inf_{x+\int_0^t u(\tau) d\tau \notin K} t$ , defined in the proof of Theorem 5.1. Note that this function has two inputs, which is one vector  $x \in X$  and function  $u(\cdot)$  (measurable, integrable).

We will use this notation in the rest of the article, and in each of the cases of interest, the context, i.e., the number of arguments of  $\tau$ , provides the proper definition.

**2.2. Assumptions.** We set  $X := \mathbb{R}^n$ . Let us consider

1. a concave function  $\psi : X \mapsto \mathbb{R}$  satisfying growth conditions

$$\forall v \in X, \quad \beta - \sigma_A(v) \leq \psi(v) \leq \delta - \sigma_A(v)$$

for some compact convex subset  $A \subset X$ , where  $\sigma_A(v) := \sup_{u \in A} \langle u, v \rangle$  is the support function of  $A$  and where  $\beta \leq \delta$ .

2. a bounded continuous function  $v : \mathbb{R}_+ \mapsto \text{Dom}(\psi)$ .

3. an upper semicontinuous initial datum  $\mathbf{N}_0 : X \mapsto \mathbb{R}_+$ . We set  $\mathbf{N}_0(0, x) := \mathbf{N}_0(x)$  and  $\mathbf{N}_0(t, x) := -\infty$  if  $t > 0$ .
4. a closed subset  $K \subset X$  with nonempty interior  $\text{Int}(K) =: \Omega$  and boundary  $\partial K =: \Gamma$ .
5. an upper semicontinuous boundary datum  $\gamma : \mathbb{R}_+ \times X \mapsto \mathbb{R}$ , satisfying<sup>3</sup>

$$\forall x \in \partial K, \quad \mathbf{N}_0(x) = \gamma(0, x) \quad \text{and} \quad \forall t \geq 0, \quad \forall x \in \text{Int}(K), \quad \gamma(t, x) = -\infty.$$

6. a Lipschitz function  $\mathbf{b} : \mathbb{R}_+ \times X \mapsto \mathbb{R} \cup \{-\infty\}$  setting the upper constraint.

We shall also assume in this paper that the data satisfy the following consistency conditions:

$$(1) \quad \left\{ \begin{array}{l} \text{(i)} \quad \forall x \in \partial K, \quad \mathbf{N}_0(x) = \gamma(0, x); \\ \text{(ii)} \quad \forall t \geq 0, \quad \forall x \in K, \quad \max(\mathbf{N}_0(t, x), \gamma(t, x)) \leq \mathbf{b}(t, x); \\ \text{(iii)} \quad \forall 0 \leq r \leq s, \quad \forall x \in \partial K, \quad \forall y \in \partial K, \quad \gamma(r, x) - \gamma(s, y) \leq \left\langle \frac{1}{s-r} \int_r^s v(\tau) d\tau, x - y \right\rangle; \\ \text{(iv)} \quad \forall x \in K, \quad \forall y \in \partial K, \quad \mathbf{N}_0(x) \leq \inf_{s \geq 0} \left( \gamma(s, y) + \left\langle \frac{1}{s} \int_0^s v(\tau) d\tau, x - y \right\rangle \right), \end{array} \right.$$

which are needed only to prove that the Dirichlet/initial conditions are satisfied (see Theorem 6.1). When the function  $v(\cdot) \equiv v$  is constant, they boil down to

$$\left\{ \begin{array}{l} \text{(i)} \quad \forall x \in \partial K, \quad \mathbf{N}_0(x) = \gamma(0, x); \\ \text{(ii)} \quad \forall t \geq 0, \quad \forall x \in K, \quad \max(\mathbf{N}_0(t, x), \gamma(t, x)) \leq \mathbf{b}(t, x); \\ \text{(iii)} \quad \forall 0 \leq r \leq s, \quad \forall x \in \partial K, \quad y \in \partial K, \quad \gamma(r, x) - \gamma(s, y) \leq \langle v, x - y \rangle; \\ \text{(iv)} \quad \forall x \in K, \quad y \in \partial K, \quad \mathbf{N}_0(x) \leq \inf_{s \geq 0} \gamma(s, y) + \langle v, x - y \rangle. \end{array} \right.$$

Under the above mentioned assumptions that are assumed throughout this paper, we shall solve the existence of a solution to the *nonhomogenous Hamilton–Jacobi equation*

$$(2) \quad \forall t > 0, \quad x \in \text{Int}(K), \quad \frac{\partial \mathbf{N}(t, x)}{\partial t} + \psi \left( \frac{\partial \mathbf{N}(t, x)}{\partial x} \right) = \psi(v(t))$$

satisfying the *initial and Dirichlet conditions*

$$(3) \quad \left\{ \begin{array}{l} \text{(i)} \quad \forall x \in K, \quad \mathbf{N}(0, x) = \mathbf{N}_0(x) \text{ (initial condition),} \\ \text{(ii)} \quad \forall t \geq 0, \quad \forall x \in \partial K, \quad \mathbf{N}(t, x) = \gamma(t, x) \text{ (Dirichlet boundary condition)} \end{array} \right.$$

and the *viability constraints*

$$(4) \quad \forall t \geq 0, \quad x \in K, \quad \mathbf{N}(t, x) \leq \mathbf{b}(t, x) \text{ (upper inequality constraint).}$$

<sup>3</sup>This is not mandatory. We can take any function such that  $\text{Dom}(\gamma) \subset K$  is strictly contained in  $K$ , an instance which may be useful for defining “guards” in impulse or hybrid systems, for instance. Boundary conditions are obtained when  $\text{Dom}(\gamma) = \partial K$ .

*Example.* This equation is motivated by a commonly used first order model equation in highway traffic (*Lighthill–Whitham–Richards* equation) when  $X := \mathbb{R}$  and  $K := [\xi, +\infty[$ ,  $\psi$  a concave flux function vanishing at density 0 and at a jam density  $\omega > 0$  and  $\mathbf{N}(t, x)$  is the cumulated number of vehicles at time  $t$  and at location  $x \in K$ . Consistency conditions (1) read in this case:  $\mathbf{N}_0(\xi) = \gamma(0, \xi)$  and

$$(5) \quad \left\{ \begin{array}{l} \text{(i)} \quad \forall t \geq 0, \forall x \in K, \max(\mathbf{N}_0(t, x), \gamma(t, x)) \leq \mathbf{b}(t, x); \\ \text{(ii)} \quad \forall 0 \leq r \leq s, \gamma(r, \xi) - \gamma(s, \xi) \leq 0 \text{ (monotonocity)}; \\ \text{(iii)} \quad \forall x \in K, \mathbf{N}_0(x) \leq \inf_{s \geq 0} \left( \gamma(s, \xi) + \left\langle \frac{1}{s} \int_0^s v(\tau) d\tau, x - \xi \right\rangle \right). \end{array} \right.$$

Then the trapezoidal flux function (such as the one proposed by Daganzo [45, 46]) defined by

$$\psi(v) = \begin{cases} \nu^b v & \text{if } v \leq \gamma^b, \\ \delta & \text{if } v \in [\gamma^b, \gamma^\sharp], \\ \nu^\sharp(\omega - v) & \text{if } v \geq \gamma^\sharp, \end{cases}$$

and the Greenshield flux function

$$\psi(v) = \begin{cases} \nu v & \text{if } v \leq 0, \\ \frac{\nu}{\omega} v(\omega - v) & \text{if } v \in [0, \omega], \\ \nu(\omega - v) & \text{if } v \geq \omega \end{cases}$$

satisfy the assumptions on the function  $\psi$  with  $A := [-\nu^b, +\nu^\sharp]$  and  $A := [-\nu, +\nu]$ , respectively (see Lemma 7.2 in section 7).

We characterize the solution to this nonhomogenous Dirichlet/initial value problem with inequality constraints through the capture basin of a target defined by the Dirichlet/initial conditions viable in an environment defined by inequality constraints under an adequate control system.

**3. Prerequisite from viability theory and convex analysis.** Readers familiar with convex analysis and viability theory can skip this section and proceed directly to section 4.

**3.1. Some prerequisites from viability theory.** Here,  $X := \mathbb{R}^n$  and  $Y := \mathbb{R}^m$  denote finite dimensional vector spaces. Let  $f : X \times Y \mapsto X$  be a single-valued map describing the dynamics of a control system and  $U : X \rightsquigarrow Y$  the set-valued map describing the state-dependent constraints on the controls.

First, any solution to a control system with state-dependent constraints on the controls

$$\begin{cases} \text{(i)} & x'(t) = f(x(t), u(t)), \\ \text{(ii)} & u(t) \in U(x(t)) \end{cases}$$

can be regarded as a solution to the differential inclusion  $x'(t) \in F(x(t))$ , where the right-hand side is defined by  $F(x) := f(x, U(x)) := \{f(x, u)\}_{u \in U(x)}$ .

We denote by  $\mathcal{S}(x) \subset \mathcal{C}(0, \infty; X)$  the set of *absolutely continuous functions*  $t \mapsto x(t) \in X$  satisfying

$$\text{for almost all } t \geq 0, \quad x'(t) \in F(x(t))$$

starting at time 0 at  $x$ :  $x(0) = x$ . The set-valued map  $\mathcal{S} : X \rightsquigarrow \mathcal{C}(0, \infty; X)$  is called the *solution map* associated with  $F$ .

Therefore, from now on, as long as we do not need to implicate explicitly the controls in our study, we shall replace control problems by differential inclusions.

We shall say that  $K$  is *locally viable under  $F$*  if from every  $x \in K$  starts a solution  $x(\cdot)$  to the differential inclusion  $x' \in F(x)$  viable in  $K$  on the nonempty interval  $[0, T_x[$  in the sense that

$$\forall t \in [0, T_x[, \quad x(t) \in K$$

and that  $K$  is *viable* if we can take  $T_x = +\infty$ . It is *locally backward invariant under  $F$*  if for every  $t_0 \in ]0, +\infty[$ ,  $x \in K$ , for all solutions  $x(\cdot)$  to the differential inclusion  $x' \in F(x)$  arriving at  $x$  at time  $t_0$ , there exists  $s \in [0, t_0[$  such that  $x(\cdot)$  is viable in  $K$  on the interval  $[s, t_0]$ , and *backward invariant* if we can take  $s = 0$ .

We denote by

$$\text{Graph}(F) := \{(x, y) \in X \times Y \mid y \in F(x)\}$$

the *graph* of a set-valued map  $F : X \rightsquigarrow Y$  and by  $\text{Dom}(F) := \{x \in X \mid F(x) \neq \emptyset\}$  its *domain*.

Most of the results of viability theory are true whenever we assume that the dynamics is Marchaud as follows.

DEFINITION 3.1 (Marchaud map). *We shall say that  $F$  is a Marchaud map if*

$$\left\{ \begin{array}{ll} \text{(i)} & \text{the graph of } F \text{ is closed,} \\ \text{(ii)} & \text{the values } F(x) \text{ of } F \text{ are convex,} \\ \text{(iii)} & \text{the growth of } F \text{ is linear:} \\ & \exists c > 0 \mid \forall x \in X, \|F(x)\| := \sup_{v \in F(x)} \|v\| \leq c(\|x\| + 1). \end{array} \right.$$

*We shall say that  $F$  is  $\lambda$ -Lipschitz if*

$$\forall x, y \in X, \quad F(x) \subset F(y) + \lambda\|x - y\|B,$$

*where  $B$  is the unit ball.*

This covers the case of *Marchaud control systems*, where  $(x, u) \mapsto f(x, u)$  is continuous, affine with respect to the controls  $u$  and with linear growth, and when  $U$  is Marchaud.

We recall the following version of the important Theorem 3.5.2 of [5].

THEOREM 3.2 (the stability theorem). *Assume that  $F : X \rightsquigarrow X$  is Marchaud. Then the solution map  $\mathcal{S}$  is upper semicompact with nonempty values; this means that whenever  $x_n \in X$  converge to  $x$  in  $X$  and  $x_n(\cdot) \in \mathcal{S}(x_n)$  is a solution to the differential inclusion  $x' \in F(x)$  starting at  $x_n$ , there exists a subsequence (again denoted by  $x_n(\cdot)$ ) converging to a solution  $x(\cdot) \in \mathcal{S}(x)$  uniformly on compact intervals.*

We shall also need some other prerequisites from [5].

DEFINITION 3.3 (capture basin of a target). *Let  $C \subset K \subset X$  be two subsets,  $C$  being regarded as a target,  $K$  as a constrained set. The subset  $\text{Capt}(K, C)$  of initial*

states  $x_0 \in K$  such that  $C$  is reached in finite time before possibly leaving  $K$  by at least one solution  $x(\cdot) \in \mathcal{S}(x_0)$  starting at  $x_0$  is called the viable-capture basin of  $C$  in  $K$ . A subset  $K$  is a repeller under  $F$  if all solutions starting from  $K$  leave  $K$  in finite time. A subset  $D$  is locally backward invariant relative to  $K$  if all backward solutions starting from  $D$  viable in  $K$  are actually viable in  $K$ .

We recall the following result of [10].

**THEOREM 3.4** (fixed-point characterization of capture basins). *The viable-capture basin  $\text{Capt}(K, C)$  of a target  $C$  viable in  $K$  is*

1. the largest subset  $D$  satisfying  $C \subset D \subset K$  and  $D \subset \text{Capt}(D, C)$ ,
2. the smallest subset  $D$  satisfying  $C \subset D \subset K$  and  $\text{Capt}(K, D) \subset D$ , and
3. the unique subset  $D$  satisfying  $C \subset D \subset K$  and

$$D = \text{Capt}(K, D) = \text{Capt}(D, C).$$

The subset  $K \setminus C$  denotes the intersection of  $K$  and the complement of  $C$ ; i.e., it is the set of elements of  $K$  which do not belong to  $C$ . We can derive the following characterization of capture basin (see [7]).

**THEOREM 3.5** (viability characterization of capture basins). *Let us assume that  $F$  is Marchaud and that the subsets  $C \subset K$  and  $K$  are closed. If  $K \setminus C$  is a repeller (this is the case when  $K$  itself is a repeller), then the viable-capture basin  $\text{Capt}(K, C)$  of the target  $C$  under  $\mathcal{S}$  is the unique closed subset satisfying  $C \subset D \subset K$  and*

$$(6) \quad \begin{cases} \text{(i)} & D \setminus C \text{ is locally viable under } \mathcal{S}, \\ \text{(ii)} & D \text{ is locally backward invariant relative to } K. \end{cases}$$

The contingent cone  $T_L(x)$  to  $L \subset X$  at  $x \in L$  is the set of directions  $v \in X$  such that there exist sequences  $h_n > 0$  converging to 0 and  $v_n$  converging to  $v$  satisfying  $x + h_n v_n \in L$  for every  $n$  (see, for instance, [13] or [79] for more details). The (regular) normal cone is the polar cone  $N_L(x) := (T_L(x))^\circ$  of the contingent cone.

**DEFINITION 3.6** (Frankowska property). *Let us consider a set-valued map  $F : X \rightrightarrows X$  and two subsets  $C \subset K$  and  $K$ . We shall say that a subset  $D$  between  $C$  and  $K$  satisfies the Frankowska property with respect to  $F$  if*

$$(7) \quad \begin{cases} \text{(i)} & \forall x \in D \setminus C, \quad F(x) \cap T_D(x) \neq \emptyset; \\ \text{(ii)} & \forall x \in D \cap \text{Int}(K), \quad -F(x) \subset T_D(x); \\ \text{(iii)} & \forall x \in D \cap \partial K, \quad -F(x) \cap T_K(x) \subset T_D(x). \end{cases}$$

Actually, conditions (7)(ii), (iii) boil down to the same condition,

$$\forall x \in D, \quad -F(x) \cap T_K(x) \subset T_D(x).$$

When  $K$  is further assumed to be backward locally invariant, the above conditions (7) boil down to

$$(8) \quad \begin{cases} \text{(i)} & \forall x \in D \setminus C, \quad F(x) \cap T_D(x) \neq \emptyset; \\ \text{(ii)} & \forall x \in D, \quad -F(x) \subset T_D(x). \end{cases}$$

Theorem 3.5 and the viability<sup>4</sup> and invariance theorems imply the following.

<sup>4</sup>See, for instance, Theorems 3.2.4, 3.3.2, and 3.5.2 of [5].

**THEOREM 3.7** (tangential characterization of capture basins). *Let us assume that  $F$  is Marchaud, that  $K$  is closed, and that a closed subset  $C$  satisfies  $\text{Viab}_F(K \setminus C) = \emptyset$ . Then the viable-capture basin  $\text{Capt}_F^K(C)$  is*

1. the largest closed subset  $D$  satisfying  $C \subset D \subset K$  and

$$(9) \quad \forall x \in D \setminus C, \quad F(x) \cap T_D(x) \neq \emptyset;$$

2. the unique closed subset  $D$  satisfying the Frankowska property (7) if  $F$  is Lipschitz.

We provide the dual characterization of the capture basin in terms of normal cones due to Frankowska.

**LEMMA 3.8** (normal characterization of capture basins). *Let us assume that*

$$\forall x \in K, \quad 0 \in \text{Int}(F(x) + T_K(x)).$$

*Then property (7) is equivalent to the dual property*

$$(10) \quad \begin{cases} \text{(i)} & \forall x \in D \setminus C, \quad \forall p \in N_D(x), \quad \sigma(F(x), -p) \geq 0; \\ \text{(ii)} & \forall x \in D \cap \text{Int}(K), \quad \forall p \in N_D(x), \quad \sigma(F(x), -p) \leq 0; \\ \text{(iii)} & \forall x \in D \cap \partial K, \quad \forall p \in N_D(x), \quad \inf_{q \in N_K(x)} \sigma(F(x), q - p) \leq 0. \end{cases}$$

*Proof.* Whenever  $0 \in \text{Int}(F(x) + T_K(x))$ , Proposition 3.9 on page 50 of [6] implies that the support function of  $-F(x) \cap T_K(x)$  is the inf-convolution of the support functions of  $-F(x)$  and  $T_K(x)$  as follows:

$$\sigma(-F(x) \cap T_K(x), p) = \inf_{q \in N_K(x)} \sigma(F(x), q - p).$$

Consequently, inclusion  $-F(x) \cap T_K(x) \subset T_D(x)$  is equivalent to

$$\forall p, \quad \inf_{q \in N_K(x)} \sigma(F(x), q - p) \leq \sigma(T_D(x), p),$$

which can be written

$$\forall p \in N_D(x), \quad \inf_{q \in N_K(x)} \sigma(F(x), q - p) \leq 0.$$

This concludes the proof.  $\square$

**3.2. Some prerequisites of convex analysis.** We gather in this section notations and some results on convex analysis for the convenience of the reader not familiar with this topic. Since the authors of most books on convex analysis have chosen to study convex functions rather than concave ones, we have chosen to associate with the concave function  $\psi$  the Fenchel transform  $\varphi^*$  of  $\varphi := -\psi$  rather than the “concave Fenchel” transform  $\psi^\boxtimes$  defined by the concave function

$$\psi^\boxtimes(u) := \inf_{p \in \text{Dom}(\psi)} [\langle p, u \rangle - \psi(p)] = -\varphi^*(-u).$$

The basic theorem of convex analysis states that  $\psi = \psi^{\boxtimes\boxtimes}$  if and only if  $\psi$  is concave, upper semicontinuous, and nontrivial (i.e.,  $\text{Dom}(\psi) := \{p \mid \varphi(p) > -\infty\} \neq \emptyset$ ).



The epigraph  $\mathcal{E}p(\varphi)$  of an extended function  $\varphi$  is the set of pairs  $(x, \lambda) \in X \times \mathbb{R}$  such that  $\varphi(x) \leq \lambda$ , and the hypograph  $\mathcal{H}yp(\psi)$  of a function  $\psi$  is the set of pairs  $(p, \mu) \in X \times \mathbb{R}$  such that  $\mu \leq \psi(p)$ . Note that the hypograph of  $\psi$  is related to the epigraph of  $\varphi$  by the relation

$$(p, \lambda) \in \mathcal{H}yp(\psi) \text{ if and only if } (p, -\lambda) \in \mathcal{E}p(\varphi).$$

An extended function is lower semicontinuous if and only if its epigraph is closed and is upper semicontinuous if and only if its hypograph is closed.

DEFINITION 3.9 (hypoderivatives and superdifferentials). *The hypoderivative  $D_{\downarrow}\psi(p)$  and the epiderivative  $D_{\uparrow}\varphi(p)$  are related to the tangent cones of the hypograph of  $\psi$  and epigraph of  $\varphi$  by the relations*

$$\mathcal{H}yp(D_{\downarrow}\psi(p)) := T_{\mathcal{H}yp(\psi)}(p, \psi(p)) \text{ and } \mathcal{E}p(D_{\uparrow}\varphi(p)) := T_{\mathcal{E}p(\varphi)}(p, \varphi(p)).$$

*The superdifferential  $\partial_+\psi(p)$  of the concave function  $\psi$  at  $p$  is defined by*

$$u \in \partial_+\psi(p) \text{ if } \forall v \in X, \langle u, v \rangle \geq D_{\downarrow}\psi(p)(v),$$

*and the subdifferential  $\partial_-\varphi(p)$  is defined by*

$$u \in \partial_-\varphi(p) \text{ if } \forall v \in X, \langle u, v \rangle \leq D_{\uparrow}\varphi(p)(v).$$

We infer that

$$\forall v \in X, D_{\downarrow}\psi(p)(v) = -D_{\uparrow}\varphi(p)(v)$$

and that

$$u \in \partial_+\psi(p) \text{ if and only if } u \in -\partial_-\varphi(p).$$

The polar cone  $P^-$  of a given set  $P$  is defined by

$$P^- = \{p \in X^* \mid \forall x \in P, \langle p, x \rangle \leq 0\},$$

where  $X^*$  is the dual space of  $X$ , and the normal cone  $N_K(x) := T_K(x)^-$  to  $K$  at  $x \in K$  we use in this paper is the polar cone to the contingent cone to  $K$  at  $x \in K$ . The superdifferential  $\partial_+\psi(p)$  and the subdifferential  $\partial_-\varphi(p)$  are related to the normal cones of the hypograph of  $\psi$  and epigraph of  $\varphi$  by the relations

$$u \in \partial_+\psi(p) \text{ if and only if } (-u, 1) \in N_{\mathcal{H}yp(\psi)}(p, \psi(p))$$

and

$$u \in \partial_-\varphi(p) \text{ if and only if } (u, -1) \in N_{\mathcal{E}p(\varphi)}(p, \varphi(p)).$$

Recall the Legendre inversion formula

$$u \in -\partial_+\psi(p) \text{ if and only if } p \in \partial_-\varphi^*(u)$$

and the (decreasing) monotonicity property of superdifferential maps  $p \rightsquigarrow \partial_+\psi(p)$  of a concave function,

$$\forall u_i \in \partial_+\psi(p_i), i = 1, 2, \langle u_1 - u_2, p_1 - p_2 \rangle \leq 0.$$

The subdifferential  $\partial_-\sigma(K, p)$  of the support function is defined by the support zone  $\{u \in K \text{ such that } \sigma(K, p) = \langle p, u \rangle\}$  of  $p$  in  $K$ . See [6] or [79] for more details.

**4. The viability hyposolution.** The assumption that the flux function  $\psi$  is concave and upper semicontinuous plays a crucial role for defining the viability hyposolution. Indeed, the Fenchel theorem allows us to characterize it by

$$(11) \quad \psi(p) = \inf_{u \in \text{Dom}(\varphi^*)} [\varphi^*(u) - \langle p, u \rangle],$$

where  $\varphi^*$  is the Fenchel conjugate function, which is the convex lower semicontinuous function defined by

$$(12) \quad \varphi^*(u) := \sup_{p \in \text{Dom}(\psi)} [\langle p, u \rangle + \psi(p)].$$

We introduce the auxiliary characteristic control system,

$$(13) \quad \begin{cases} \tau'(t) = -1, \\ x'(t) = u(t), \\ y'(t) = \varphi^*(u(t)) - \psi(v(\tau(t))), \text{ where } u(t) \in \text{Dom}(\varphi^*). \end{cases}$$

The function  $\tau(t)$  corresponds to a *countdown*, i.e., a pseudotime decaying at rate  $-1$ . This technique of augmentation of a dynamics by  $\tau'(t) = -1$  is common in the Hamilton–Jacobi partial differential equation literature; see, for example, [55, 56]. To be rigorous, we have to mention *once and for all* that the controls  $u(\cdot)$  are measurable integrable functions with values in  $\text{Dom}(\varphi^*)$ , and thus, ranging over  $L^1(0, \infty; \text{Dom}(\varphi^*))$ , and that the above system of differential equations is valid for almost all  $t \geq 0$ .

We set  $\mathbf{c}(t, x) := \max(\mathbf{N}_0(t, x), \gamma(t, x))$ , defined by

$$\mathbf{c}(t, x) := \begin{cases} -\infty & \text{if } t > 0 \text{ and } x \in \Omega := \text{Int}(K), \\ \mathbf{N}_0(x) & \text{if } t = 0 \text{ and } x \in K, \\ \gamma(t, x) & \text{if } t \geq 0 \text{ and } x \in \Gamma := \partial K. \end{cases}$$

We introduce the environment  $\mathcal{K} := \mathcal{Hyp}(\mathbf{b})$  is the subset of triples  $(T, x, y) \subset \mathbb{R}_+ \times X \times \mathbb{R}$  such that  $y \leq \mathbf{b}(T, x)$  (this is the *hypograph* of the function  $\mathbf{b}$ ) and the target  $\mathcal{C} := \mathcal{Hyp}(\mathbf{c})$  defined as the subset of triples  $(T, x, y) \subset \mathbb{R}_+ \times X \times \mathbb{R}$  such that  $y \leq \mathbf{c}(T, x)$  (which is the *hypograph* of the function  $\mathbf{c}$ ).

**DEFINITION 4.1** (the viability hyposolution). *The capture basin  $\text{Capt}_{(13)}(\mathcal{K}, \mathcal{C})$  of a target  $\mathcal{C}$  viable in the environment  $\mathcal{K}$  under control system (13) is the subset of initial states  $(t, x, y)$  such that there exists a measurable control  $u(\cdot)$  such that the associated solution*

$$s \mapsto \left( t - s, x + \int_0^s u(\tau) d\tau, y + \int_0^s (\varphi^*(u(\tau)) - \psi(v(t - \tau))) d\tau \right)$$

*is viable in  $\mathcal{K}$  until it reaches the target  $\mathcal{C}$ .*

*The viability hyposolution  $\mathbf{N}$  is defined by*

$$(14) \quad \mathbf{N}(t, x) := \sup_{(t, x, y) \in \text{Capt}_{(13)}(\mathcal{K}, \mathcal{C})} y.$$

Note that  $\mathcal{Hyp}(\mathbf{M}) \subset \mathcal{Hyp}(\mathbf{N})$  if and only if  $\mathbf{N}$  is pointwise larger than  $\mathbf{M}$ . Therefore, using hypographs, the two order relations coincide.

We shall prove the following.

**THEOREM 4.2** (nonhomogenous Dirichlet/initial value problem with inequality constraints). *The viability hypsolution  $\mathbf{N}$  defined by (14) is the largest upper semicontinuous solution to Hamilton–Jacobi equation (2) satisfying initial and Dirichlet conditions (3) and inequality constraints (4) in both the contingent solution sense (see (28)) and in the contingent normal sense (see (31)). If the functions  $\psi$ ,  $\varphi^*$ , and  $v$  are furthermore Lipschitz, then the viability hypsolution  $\mathbf{N}$  is its unique upper semicontinuous solution in both the contingent Frankowska sense (see (29) and (30)) and in the Barron-Jensen/Frankowska sense (see (32), (33) and Theorems 8.1 and 9.1 for the precise statement).*

*Remark.* Note that the concept of “largest solution” coincides with the pointwise one. Inequalities (32) and (33) defining the concept of generalized solutions depend on the type of assumption made on the flux function  $\psi$ . The present work uses a standard assumption in transportation engineering, namely that the flux  $\psi$  is *concave*, whereas a majority of mathematical studies of Hamilton–Jacobi partial differential equations assume that  $\psi$  is *convex*. This change induces an unusual modification of the signs in the inequalities defining the concept of Barron-Jensen/Frankowska solutions. Under the assumption of convex fluxes, this solution would be lower semicontinuous (and sometimes called the lower semicontinuous solution to Hamilton–Jacobi equations). Under the assumption imposed by transportation engineering considerations, the present solution is upper semicontinuous and the signs in inequalities (32) and (33) are changed. The mathematical formulation of the engineering problem thus led to a slightly unusual framework for solving this Hamilton–Jacobi equation. The convex version of this paper will appear in the forthcoming book [9].

We shall derive this theorem and other results from the properties of capture basins gathered in [7, 10]. Since the capture basin of a union of targets is the union of the capture basins of these targets, we infer that whenever  $\mathbf{c} := \sup_i \mathbf{c}_i$  is the upper envelope of a family of functions  $\mathbf{c}_i$ , then the viability hypsolution is the upper envelope

$$\forall t \geq 0, x \in X, \quad \mathbf{N}(t, x) = \sup_i \mathbf{N}_{\mathbf{c}_i}(t, x)$$

of the solutions  $\mathbf{N}_{\mathbf{c}_i}$  (sup-linearity property).

In particular, since  $\mathbf{c}(t, x) := \max(\mathbf{N}_0(t, x), \gamma(t, x))$  (extended to  $-\infty$  when  $t > 0$  or  $x \in \text{Int}(K)$ ), we obtain the decomposition formula

$$(15) \quad \mathbf{N}(t, x) = \max(\mathbf{N}_{\mathbf{N}_0}(t, \mathbf{x}), \mathbf{N}_\gamma(t, \mathbf{x}))$$

in terms of initial condition component  $\mathbf{N}_{\mathbf{N}_0}$  and the Dirichlet component  $\mathbf{N}_\gamma$  of the viability hypsolution  $\mathbf{N}$  defined by

$$\begin{cases} \mathbf{N}_{\mathbf{N}_0}(t, x) := \sup_{(t, x, y) \in \text{Capt}_{(13)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{N}_0))} y, \\ \mathbf{N}_\gamma(t, x) := \sup_{(t, x, y) \in \text{Capt}_{(13)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\gamma))} y. \end{cases}$$

The viability hypsolution depends continuously on the data in the following sense: If the hypographs of a sequence of initial data  $\mathbf{c}_j$  converge in the upper Painlevé–Kuratowski sense (see, for instance, [13]) to the hypograph of data  $\mathbf{c}$ , then the upper

Painlevé–Kuratowski limit of the hypographs of the solutions  $\mathbf{N}_j$  associated with data  $\mathbf{c}_j$  is contained in the hypograph of the hyposolution  $\mathbf{N}$  associated with data  $\mathbf{c}$  (upper hypocontinuity property). If the functions  $\psi$ ,  $\varphi^*$ , and  $v$  are furthermore Lipschitz, the hypograph of the hyposolution  $\mathbf{N}$  associated with data  $\mathbf{c}$  is contained in the lower Painlevé–Kuratowski limit of the hypographs of the solutions  $\mathbf{N}_j$  associated with data  $\mathbf{c}_j$  (lower hypocontinuity property), so that both the upper and lower limits coincide with the hypograph of the hyposolution  $\mathbf{N}$  (hypoconvergence of the solutions; see [13] or [79] for a definition). These statements follow from Theorem 6.6 of [7] stating that if the system is both Marchaud and Lipschitz, the capture basin of a Painlevé–Kuratowski limit of targets is the Painlevé–Kuratowski limit of the capture basins of the targets.

5. Lax–Hopf formula and estimates of the solution.

**5.1. The Lax–Hopf formula for Dirichlet problems.** When there is no inequality constraint, we prove that the viability hyposolution can be represented explicitly as a simple maximization problem involving the Fenchel conjugate  $\varphi^*$  defined by (12).

**THEOREM 5.1** (the Lax–Hopf formula). *Let us consider the case without inequality constraints and set*

$$\tau(x, u) := \inf_{x+tu \notin K} t \quad \text{and} \quad \sigma(t, x, u) := \min(t, \tau(x, u)).$$

Then the viability hyposolution (17) can be written

$$(16) \quad \left\{ \begin{array}{l} \mathbf{N}(t, x) \\ = \sup_{\{u \in \text{Dom}(\varphi^*)\}} \left[ \mathbf{c}(t - \sigma(t, x, u), x + \sigma(t, x, u)u) - \sigma(t, x, u)\varphi^*(u) \right. \\ \qquad \qquad \qquad \left. + \int_{t-\sigma(t, x, u)}^t (\psi(v(\tau)))d\tau \right]. \end{array} \right.$$

Using the decomposition  $\mathbf{N}(t, x) = \max(\mathbf{N}_{\mathbf{N}_0}(t, x), \mathbf{N}_\gamma(t, x))$ , we derive the more explicit formula

$$(17) \quad \left\{ \begin{array}{l} \mathbf{N}_{\mathbf{N}_0}(t, x) = \sup_{u \in \text{Dom}(\varphi^*)} (\mathbf{N}_0(x + tu) - t\varphi^*(u)) + \int_0^t \psi(v(\tau))d\tau, \\ \mathbf{N}_\gamma(t, x) = \sup_{\{u \in \text{Dom}(\varphi^*) \mid \tau(x, u) \leq t\}} \left[ \gamma(t - \tau(x, u), x + \tau(x, u)u) - \tau(x, u)\varphi^*(u) \right. \\ \qquad \qquad \qquad \left. + \int_{t-\tau(x, u)}^t \psi(v(\tau))d\tau \right] \end{array} \right.$$

involving the initial and Dirichlet conditions.

*Proof.* Let us associate the following with  $u(\cdot)$ :

$$\tau(x, u(\cdot)) := \inf_{x + \int_0^t u(\tau)d\tau \notin K} t \quad \text{and} \quad \sigma(t, x, u(\cdot)) = \min(t, \tau(x, u(\cdot))).$$

The formula is derived from the general representation formula

$$\left\{ \begin{array}{l} \mathbf{N}(t, x) = \sup_{u(\cdot)} \left[ \mathbf{c} \left( t - \sigma(t, x, u(\cdot)), x + \int_0^{\sigma(t, x, u(\cdot))} u(\tau) d\tau \right) - \int_0^{\sigma(t, x, u(\cdot))} \varphi^*(u(\tau)) d\tau \right. \\ \left. + \int_{t-\sigma(t, x, u(\cdot))}^t \psi(v(\tau)) d\tau \right] \end{array} \right.$$

of the viability hypsolution without constraints given by Corollary 5.6.

We proceed in two steps.

1. Taking constant controls  $u(\cdot) \equiv u$  and observing that  $\tau(x, u) = \tau(x, u(\cdot))$ , we infer that

$$\begin{aligned} \sup_{u \in \text{Dom}(\varphi^*)} & \left( \mathbf{c}(t - \sigma(t, x, u), x + \sigma(t, x, u)u) - \sigma(t, x, u)\varphi^*(u) \right) \\ & + \int_{t-\sigma(t, x, u)}^t \psi(v(\tau)) d\tau \leq \mathbf{N}(t, x). \end{aligned}$$

2. Let us associate with  $u(\cdot)$  the function  $\hat{u}$  defined by  $\hat{u}(s) := \frac{1}{s} \int_0^s u(\tau) d\tau$ . We first observe that

$$\tau(x, u(\cdot)) = \tau(x, \hat{u}(\tau(x, u(\cdot)))).$$

Since  $\varphi^*$  is convex and lower semicontinuous and  $\psi$  is concave and upper semicontinuous, the Jensen inequality implies

$$\varphi^* \left( \frac{1}{s} \int_0^s u(\tau) d\tau \right) \leq \frac{1}{s} \int_0^s \varphi^*(u(\tau)) d\tau$$

and

$$\frac{1}{s} \int_0^s \psi(v(t - \tau)) d\tau \leq \psi \left( \frac{1}{s} \int_{t-s}^t v(\tau) d\tau \right),$$

and thus

$$(18) \quad \int_0^s \psi(v(t - \tau)) d\tau - \int_0^s \varphi^*(u(\tau)) d\tau \leq s \left( \psi \left( \frac{1}{s} \int_{t-s}^t v(\tau) d\tau \right) - \varphi^*(\hat{u}(s)) \right).$$

Consequently, setting  $t^\# := \sigma(t, x, u(\cdot)) = \tau(x, \hat{u}(\sigma(t, x, u(\cdot))))$  and  $u^\# := \hat{u}(t^\#)$ , we obtain inequalities

$$\left\{ \begin{array}{l} \mathbf{c} \left( t - t^\#, x + \int_0^{t^\#} u(\tau) d\tau \right) - \int_0^{t^\#} \varphi^*(u(\tau)) d\tau + \int_{t-t^\#}^t \psi(v(\tau)) d\tau \\ \leq \mathbf{c}(t - t^\#, x + t^\# u^\#) - t^\# \varphi^*(u^\#) + \int_{t-t^\#}^t \psi(v(\tau)) d\tau \\ \leq \sup_{\{u \in \text{Dom}(\varphi^*)\}} \left( \mathbf{c}(t - \sigma(t, x, u), x + \sigma(t, x, u)u) - \sigma(t, x, u)\varphi^*(u) \right) \\ \quad + \int_{t-\sigma(t, x, u)}^t \psi(v(\tau)) d\tau. \end{array} \right.$$

Therefore, by taking the supremum, we obtain

$$\begin{aligned} \mathbf{N}(t, x) \leq \sup_{u \in \text{Dom}(\varphi^*)} & \left( \mathbf{c}(t - \sigma(t, x, u), x + \sigma(t, x, u)u) - \sigma(t, x, u)\varphi^*(u) \right) \\ & + \int_{t-\sigma(t, x, u)}^t \psi(v(\tau))d\tau. \end{aligned}$$

This completes the proof of the Lax–Hopf inequality.

**COROLLARY 5.2** (case of the traffic model). *When  $X := \mathbb{R}$ ,  $K := [\xi, +\infty[$ ,  $\psi$  is a concave flux function vanishing at density 0 and at a jam density  $\omega > 0$ , and  $\mathbf{N}(t, x)$  is the cumulated number of vehicles at time  $t$  and at location  $x \in K$ . Consistency conditions (5) imply the existence of a unique upper semicontinuous solution  $\mathbf{N}(t, x) = \max(\mathbf{N}_{\mathbf{N}_0}(t, x), \mathbf{N}_\gamma(t, x))$  to this problem in the Barron–Jensen/Frankowska sense satisfying the Lax–Hopf formula*

$$(19) \quad \begin{cases} \mathbf{N}_{\mathbf{N}_0}(t, x) = \sup_{u \in \text{Dom}(\varphi^*)} \left( \mathbf{N}_0(x + tu) - t\varphi^*(u) + \int_0^t \psi(v(t - \tau))d\tau \right), \\ \mathbf{N}_\gamma(t, x) = \sup_{\{u \in \text{Dom}(\varphi^*) \mid u \leq \frac{\xi - x}{t}\}} \left( \gamma\left(t - \frac{\xi - x}{u}, \xi\right) - \frac{\xi - x}{u}\varphi^*(u) + \int_0^{\frac{\xi - x}{u}} \psi(v(t - \tau))d\tau \right). \end{cases}$$

**5.2. A posteriori estimates.** The maximum principle, an a priori upper estimate of a solution of a partial differential equation (whether it exists or not) is here obtained as an a posteriori estimate, a property of the *viability hypsolution*.

**PROPOSITION 5.3** (upper estimate of the viability hypsolution). *The viability hypsolution satisfies*

$$\mathbf{N}(t, x) \leq \sup_{u \in \text{Dom}(\varphi^*)} \left( \mathbf{c}(t - \sigma(t, x, u), x + \sigma(t, x, u)u) - \left\langle u, \int_{t-\sigma(t, x, u)}^t v(\tau)d\tau \right\rangle \right).$$

Consequently, the viability hypsolution satisfies the following (a posteriori instead of a priori) estimate

$$\mathbf{N}(t, x) \leq \sup_{t \geq 0, x \in K} \mathbf{c}(t, x) + t \text{Diam}(\text{Dom}(\varphi^*)) \sup_{t \geq 0} \|v(t)\|$$

(maximum principle).

*Proof.* Fix  $u \in \text{Dom}(\varphi^*)$  and set  $\sigma(t, x, u) =: s$ . Definition (12) of the conjugate function implies

$$(20) \quad \psi\left(\frac{1}{s} \int_{t-s}^t v(\tau)d\tau\right) - \varphi^*(u) \leq -\left\langle \frac{1}{s} \int_{t-s}^t v(\tau)d\tau, u \right\rangle.$$

Consequently,

$$\begin{cases} \mathbf{c}(t - \sigma(t, x, u), x + \sigma(t, x, u)u) - \sigma(t, x, u)\varphi^*(u) + \int_{t-\sigma(t, x, u)}^t (\psi(v(\tau)))d\tau \\ \leq \mathbf{c}(t - \sigma(t, x, u), x + \sigma(t, x, u)u) - \left\langle u, \int_{t-\sigma(t, x, u)}^t v(\tau)d\tau \right\rangle \\ \leq \sup_{w \in \text{Dom}(\varphi^*)} \left( \mathbf{c}(t - \sigma(t, x, w), x + \sigma(t, x, w)w) - \left\langle w, \int_{t-\sigma(t, x, w)}^t v(\tau)d\tau \right\rangle \right). \end{cases}$$

Taking the supremum over  $u \in \text{Dom}(\varphi^*)$ , Lax-Hopf formula (16) implies the upper estimate

$$\mathbf{N}(t, x) \leq \sup_{u \in \text{Dom}(\varphi^*)} \left( \mathbf{c}(t - \sigma(t, x, u), x + \sigma(t, x, u)u) - \left\langle u, \int_{t-\sigma(t, x, u)}^t v(\tau) d\tau \right\rangle \right).$$

This completes the proof.  $\square$

In the same way, we provide a lower estimate of the solution.

**PROPOSITION 5.4** (lower estimate). *Assume that  $v(t) := v$  is constant and, for simplicity, that the function  $\psi$  is differentiable. Then*

$$\mathbf{c}(t - \sigma(t, x, -\psi'(v)), x - \sigma(t, x, -\psi'(v))\psi'(v)) + \sigma(t, x, \psi'(v)) \langle v, \psi'(v) \rangle \leq \mathbf{N}(t, x).$$

Consequently, the hyposolution is nonnegative on its positivity domain  $\text{Dom}_+(\mathbf{N})$ , defined as the subset of pairs  $(t, x) \in \mathbb{R}_+ \times K$  such that

$$\mathbf{c}(t - \sigma(t, x, -\psi'(v)), x - \sigma(t, x, -\psi'(v))\psi'(v)) + \sigma(t, x, -\psi'(v)) \langle v, \psi'(v) \rangle \geq 0.$$

*Proof.* By Definition 3.9 of the superdifferential,

$$\forall u \in \partial_+ \psi(v), \quad \psi(v) - \varphi^*(-u) = \langle v, u \rangle.$$

Therefore, if  $\psi$  is differentiable, taking  $u := -\psi'(v)$  as the unique element of  $-\partial_+ \psi(v) = \partial_- \varphi(v)$ , the Legendre equality  $\psi(v) - \varphi^*(-\psi'(v)) = \langle v, \psi'(v) \rangle$  yields

$$\begin{cases} \mathbf{c}(t - \sigma(t, x, -\psi'(v)), x - \sigma(t, x, -\psi'(v))\psi'(v)) + \sigma(t, x, -\psi'(v)) \langle v, \psi'(v) \rangle \\ = \mathbf{c}(t - \sigma(t, x, u), x + \sigma(t, x, u)u) - \sigma(t, x, u) \langle v, u \rangle \\ = \mathbf{c}(t - \sigma(t, x, u), x + \sigma(t, x, u)u) + \sigma(t, x, u)(\psi(v) - \varphi^*(u)) \leq \mathbf{N}(t, x) \end{cases}$$

thanks to the Lax-Hopf formula.  $\square$

**5.3. General representation formula.** We have derived Lax-Hopf formula (16) from a general representation formula (21) valid when there is viability constraints.

**THEOREM 5.5** (representation formula of the viability solution (the case with constraints)). *We already set*

$$\tau(x, u(\cdot)) := \inf_{x + \int_0^t u(\tau) d\tau \notin K} t \quad \text{and} \quad \sigma(t, x, u(\cdot)) = \min(t, \tau(x, u(\cdot))).$$

The viability hyposolution can be represented in the form

$$(21) \quad \begin{cases} \mathbf{N}(t, x) = \sup_{u(\cdot)} \left[ \min \left( \mathbf{c} \left( t - \sigma(t, x, u(\cdot)), x + \int_0^{\sigma(t, x, u(\cdot))} u(\tau) d\tau \right) \right. \right. \\ \left. \left. - \int_0^{\sigma(t, x, u(\cdot))} \varphi^*(u(\tau)) d\tau + \int_{t-\sigma(t, x, u(\cdot))}^t \psi(v(\tau)) d\tau, \right. \right. \\ \left. \left. \inf_{s \in [0, \sigma(t, x, u(\cdot))]} \left( \mathbf{b} \left( t - s, x + \int_0^s u(\tau) d\tau \right) - \int_0^s \varphi^*(u(\tau)) d\tau + \int_{t-s}^t \psi(v(\tau)) d\tau \right) \right) \right]. \end{cases}$$

Using the decomposition  $\mathbf{N}(t, x) = \max(\mathbf{N}_{N_0}(t, x), \mathbf{N}_\gamma(t, x))$ , this formula boils down to

$$\left\{ \begin{array}{l} \mathbf{N}_{N_0}(t, x) = \sup_{u(\cdot)} \left[ \min \left( \mathbf{N}_0 \left( x + \int_0^t u(\tau) d\tau \right) - \int_0^t \varphi^*(u(\tau)) d\tau + \int_0^t \psi(v(\tau)) d\tau, \right. \right. \\ \left. \left. \inf_{s \in [0, t]} \left( \mathbf{b} \left( t - s, x + \int_0^s u(\tau) d\tau \right) - \int_0^s \varphi^*(u(\tau)) d\tau + \int_{t-s}^t \psi(v(\tau)) d\tau \right) \right] \end{array} \right.$$

and

$$\left\{ \begin{array}{l} \mathbf{N}_\gamma(t, x) = \sup_{\{u(\cdot) | \tau(x, u(\cdot)) \leq t\}} \left[ \min \left( \gamma \left( t - \tau(x, u(\cdot)), x + \int_0^{\tau(x, u(\cdot))} u(\tau) d\tau \right) \right. \right. \\ \left. \left. - \int_0^{\tau(x, u(\cdot))} \varphi^*(u(\tau)) d\tau + \int_{t-\tau(x, u(\cdot))}^t \psi(v(\tau)) d\tau, \right. \right. \\ \left. \left. \inf_{s \in [0, \tau(x, u(\cdot))]} \left( \mathbf{b} \left( t - s, x + \int_0^s u(\tau) d\tau \right) - \int_0^s \varphi^*(u(\tau)) d\tau + \int_{t-s}^t \psi(v(\tau)) d\tau \right) \right] \end{array} \right.$$

*Proof.* We begin by observing that a solution  $(\tau(\cdot), x(\cdot), y(\cdot))$  to control system (13) starting from  $(t, x, y)$  is given by  $\tau(s) = t - s$ ,  $x(s) = x + \int_0^s u(r) dr$  and

$$y(s) = y + \int_0^s (\varphi^*(u(r)) - \psi(v(t-r))) dr$$

for some  $u(\cdot)$ .

Therefore, to say that  $(t, x, y)$  belongs to the capture basin  $\text{Capt}_{(13)}(\mathcal{K}, \mathcal{C})$  amounts to saying that there exists a solution  $(\tau(\cdot), x(\cdot), y(\cdot))$  to the characteristic control system (13) starting from  $(t, x, y)$  and  $t^* \in [0, t]$  such that

1.  $(t - t^*, x(t^*), y(t^*))$  belongs to the target  $\mathcal{C}$ , i.e., such that

$$\begin{aligned} y(t^*) &:= y + \int_0^{t^*} (\varphi^*(u(\tau)) - \psi(v(t-\tau))) d\tau \leq \mathbf{c}(t - t^*, x(t^*)) \\ &= \mathbf{c} \left( t - t^*, x + \int_0^{t^*} u(\tau) d\tau \right). \end{aligned}$$

2. For all  $s \in [0, t^*]$ ,  $(t - s, x(s), y(s))$  belongs to the environment  $\mathcal{K}$ , i.e., such that

$$\begin{aligned} y(s) &= y + \int_0^s (\varphi^*(u(\tau)) - \psi(v(t-\tau))) d\tau \leq \mathbf{b}(t - s, x(s)) \\ &= \mathbf{b} \left( t - s, x + \int_0^s u(\tau) d\tau \right). \end{aligned}$$

This implies that

$$y \leq \min \left( \begin{array}{l} \mathbf{c} \left( t - t^*, x + \int_0^{t^*} u(\tau) d\tau \right) - \int_0^{t^*} (\varphi^*(u(\tau)) - \psi(v(t-\tau))) d\tau, \\ \inf_{s \in [0, t^*]} \mathbf{b} \left( t - s, x + \int_0^s u(\tau) d\tau \right) - \int_0^s (\varphi^*(u(\tau)) - \psi(v(t-\tau))) d\tau \end{array} \right).$$



Since  $y$  is finite, this implies that  $\mathbf{c}(t - t^*, x + \int_0^{t^*} u(\tau) d\tau)$  must be finite, and thus, that

1. either  $t - t^* = 0$ , in which case  $\mathbf{c}(t - t^*, x + \int_0^{t^*} u(\tau) d\tau) = \mathbf{N}_0(x + \int_0^t u(\tau) d\tau)$ ;
2. or  $x(t^*) \in \partial K$ , which means that  $t^* = \tau(x, u(\cdot)) = \sigma(t, x, u(\cdot)) \leq t$ , in which case

$$\mathbf{c}\left(t - t^*, x + \int_0^{t^*} u(\tau) d\tau\right) = \gamma\left(t - \sigma(t, x, u(\cdot)), x + \int_0^{\sigma(t, x, u(\cdot))} u(\tau) d\tau\right).$$

This implies that  $\mathbf{N}(t, x) \leq \mathbf{V}(t, x)$ , where

$$\left\{ \begin{array}{l} \mathbf{V}(t, x) = \sup_{u(\cdot)} \left( \min \left( \mathbf{c}\left(t - \sigma(t, x, u(\cdot)), x + \int_0^{\sigma(t, x, u(\cdot))} u(\tau) d\tau\right) \right. \right. \\ \quad \left. \left. - \int_0^{\sigma(t, x, u(\cdot))} \varphi^*(u(\tau)) d\tau + \int_{t - \sigma(t, x, u(\cdot))}^t \psi(v(\tau)) d\tau, \right. \right. \\ \left. \inf_{s \in [0, \sigma(t, x, u(\cdot))]} \left( \mathbf{b}\left(t - s, x + \int_0^s u(\tau) d\tau\right) - \int_0^s \varphi^*(u(\tau)) d\tau + \int_{t-s}^t \psi(v(\tau)) d\tau \right) \right) \end{array} \right\}.$$

For proving the converse inequality, we associate with every  $\varepsilon > 0$  a control  $t \mapsto u_\varepsilon(t) \in \text{Dom}(\varphi^*)$  such that

$$\left\{ \begin{array}{l} \mathbf{V}(t, x) - \varepsilon \leq \min \left( \mathbf{c}\left(t - \sigma(t, x, u_\varepsilon(\cdot)), x + \int_0^{\sigma(t, x, u_\varepsilon(\cdot))} u_\varepsilon(\tau) d\tau\right) \right. \\ \quad \left. - \int_0^{\sigma(t, x, u_\varepsilon(\cdot))} \varphi^*(u_\varepsilon(\tau)) d\tau + \int_{t - \sigma(t, x, u_\varepsilon(\cdot))}^t \psi(v(\tau)) d\tau, \right. \\ \left. \inf_{s \in [0, \sigma(t, x, u_\varepsilon(\cdot))]} \left( \mathbf{b}\left(t - s, x + \int_0^s u_\varepsilon(\tau) d\tau\right) - \int_0^s \varphi^*(u_\varepsilon(\tau)) d\tau + \int_{t-s}^t \psi(v(\tau)) d\tau \right) \right) \end{array} \right\}.$$

Therefore, setting  $x_\varepsilon(t) := x + \int_0^t u_\varepsilon(s) ds$  and

$$y_\varepsilon(t) := \mathbf{V}(t, x) - \varepsilon + \int_0^t (\varphi^*(u_\varepsilon(r)) - \psi(v(t - r))) dr$$

we observe that the function  $s \mapsto (t - s, x_\varepsilon(s), y_\varepsilon(s))$  starts from  $(t, x, \mathbf{V}(t, x) - \varepsilon)$ , is a solution to characteristic control system (13), viable in  $\mathcal{K}$  for  $s \leq \sigma(t, x, u_\varepsilon(\cdot))$  because

$$y_\varepsilon(s) = \mathbf{V}(t, x) - \varepsilon + \int_0^s (\varphi^*(u_\varepsilon(r)) - \psi(v(t - r))) dr \leq \mathbf{b}(t - s, x_\varepsilon(s)),$$

and reaches the target  $\mathcal{C} := \text{Hyp}(\mathbf{c})$  at time  $t_\varepsilon := \sigma(t, x, u_\varepsilon(\cdot))$ ,

$$y_\varepsilon(t_\varepsilon) = \mathbf{V}(t, x) - \varepsilon + \int_0^{t_\varepsilon} (\varphi^*(u_\varepsilon(r)) - \psi(v(t - r))) dr \leq \mathbf{c}(t - t_\varepsilon, x_\varepsilon(t_\varepsilon)).$$

This implies that  $(t, x, \mathbf{V}(t, x) - \varepsilon)$  belongs to the capture basin  $\text{Capt}_{(13)}(\mathcal{K}, \mathcal{C})$ , and thus, that  $\mathbf{V}(t, x) - \varepsilon \leq \mathbf{N}(t, x)$ . Letting  $\varepsilon$  converge to 0 provides the converse inequality, and thus, the representation formula we were looking for.  $\square$

COROLLARY 5.6 (representation formula of the viability solution (the case without constraints)). *Without inequality constraints, the viability hypsolution can be represented in the form*

$$\left\{ \begin{array}{l} N(t, x) = \sup_{u(\cdot)} \left( \int_{t-\sigma(t, x, u(\cdot))}^t \psi(v(\tau)) d\tau \right. \\ \left. + \mathbf{c} \left( t - \sigma(t, x, u(\cdot)), x + \int_0^{\sigma(t, x, u(\cdot))} u(\tau) d\tau \right) - \int_0^{\sigma(t, x, u(\cdot))} \varphi^*(u(\tau)) d\tau \right). \end{array} \right.$$

**6. Dirichlet/initial conditions and inequality constraints.** We begin by checking that the viability hypsolution satisfies the initial condition, the Dirichlet condition, and the inequality constraints.

THEOREM 6.1 (Dirichlet/initial conditions and inequality constraints). *Consistency conditions (1) imply that the viability hypsolution satisfies the initial and Dirichlet conditions (3) and inequality constraints (4).*

*Proof.* Inclusions

$$\mathcal{C} := \mathcal{Hyp}(\mathbf{c}) \subset \text{Capt}_{(13)}(\mathcal{K}, \mathcal{C}) \subset \mathcal{K} := \mathcal{Hyp}(\mathbf{b})$$

imply that

$$\forall t \geq 0, \forall x \in K, \quad \mathbf{c}(t, x) \leq \mathbf{N}(t, x) \leq \mathbf{b}(t, x),$$

and thus inequality constraint  $\mathbf{N}(t, x) \leq \mathbf{b}(t, x)$  and inequalities  $\mathbf{N}_0(x) \leq \mathbf{N}(0, x)$  for all  $x \in K$  and  $\gamma(t, x) \leq \mathbf{N}(t, x)$  for all  $t \geq 0$  and  $x \in \partial K$ . We now prove by contradiction that consistency conditions (1) imply converse inequalities  $\mathbf{N}_0(x) \geq \mathbf{N}(0, x)$  for all  $x \in K$  and  $\gamma(t, x) \geq \mathbf{N}(t, x)$  for all  $t \geq 0$  and  $x \in \partial K$  that we summarize in

$$\forall (t, x) \in \text{Dom}(\mathbf{c}), \quad \mathbf{N}(t, x) \leq \mathbf{c}(t, x).$$

Assume that there exist  $(t, \xi) \in \text{Dom}(\mathbf{c})$  and  $\varepsilon > 0$  such that

$$\mathbf{N}(t, \xi) = \mathbf{c}(t, \xi) + \varepsilon.$$

Since  $(t, \xi, \mathbf{N}(t, \xi))$  belongs to the capture basin  $\text{Capt}_{(13)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c}))$ , there exists a solution  $(\tau(\cdot), x(\cdot), y(\cdot))$  to the characteristic control system (13) starting from  $(t, \xi, \mathbf{N}(t, \xi))$  and  $t^* > 0$  such that  $(t - t^*, x(t^*), y(t^*))$  belongs to the hypograph  $\mathcal{Hyp}(\mathbf{c})$ ; i.e., setting  $x(t^*) = \xi + \int_0^{t^*} u(\tau) d\tau = \eta$ , we obtain

$$y(t^*) = \mathbf{N}(t, \xi) + \int_0^{t^*} \varphi^*(u(\tau)) d\tau - \int_0^{t^*} \psi(v(t - \tau)) d\tau \leq \mathbf{c}(t - t^*, \eta).$$

Inequality (18) and definition (20) imply

$$(22) \quad \int_0^s \psi(v(t - \tau)) d\tau - \int_0^s \varphi^*(u(\tau)) d\tau \leq - \left\langle \frac{1}{s} \int_{t-s}^t v(\tau) d\tau, \widehat{u}(s) \right\rangle.$$

Piecing these inequalities together and taking  $s = t^*$ , we infer that

$$\begin{cases} \mathbf{c}(t, \xi) + \varepsilon + \left\langle \frac{1}{t^*} \int_{t-t^*}^t v(\tau) d\tau, \eta - \xi \right\rangle \\ \leq \mathbf{N}(t, \xi) + \int_0^{t^*} \varphi^*(u(\tau)) d\tau - \int_0^{t^*} \psi(v(t - \tau)) d\tau \leq \mathbf{c}(t - t^*, \eta), \end{cases}$$

from which we deduce that

$$\varepsilon \leq \mathbf{c}(t - t^*, \eta) - \mathbf{c}(t, \xi) - \left\langle \frac{1}{t^*} \int_{t-t^*}^t v(\tau) d\tau, \eta - \xi \right\rangle.$$

Consistency conditions (1) can be written in the form

$$\forall 0 \leq r \leq s, \forall x \in K, y \in \partial K, \mathbf{c}(r, x) - \mathbf{c}(s, y) \leq \left\langle \frac{1}{s-r} \int_r^s v(\tau) d\tau, x - y \right\rangle.$$

Taking  $r := t - t^*$ ,  $s := t$ ,  $x := \eta$ , and  $y := \xi$ , we obtain the contradiction  $\varepsilon \leq 0$ , and thus, we proved that for any  $(t, \xi) \in \text{Dom}(\mathbf{c})$ ,  $\mathbf{N}(t, \xi) = \mathbf{c}(t, \xi)$ .  $\square$

**7. Other auxiliary systems.** For proving that the viability hypsolution is the solution, in a generalized sense, to the Hamilton-Jacobi partial differential equation derived from the tangential or normal conditions characterizing capture basins, we need assumptions that control system (13) does not satisfy.

The two inequalities characterizing the Barron-Jensen/Frankowska solution follow from the two inclusions characterizing the Frankowska property of the capture basin (Definition 3.6). One is derived from the viability theorem and requires the assumption that  $F$  is Marchaud (upper semicontinuous, linear growth, with convex images); the other one is derived from the invariance theorem, valid whenever  $F$  is Lipschitz with closed values, without bounds on the size of their images (see Theorems 3.7 and 3.8). This is the reason why we introduce below two new systems, (23) and (24). The first one complies with the “Marchaud assumptions” of the viability theorem, so that the capture basin under it will satisfy the first inclusion of the Frankowska property, the second one to the “Lipschitz assumptions” of the invariance theorem, so that the capture basin under it will satisfy the second inclusion of the Frankowska property. The aim of this section is to derive from the inclusions of the Frankowska property the corresponding inequalities defining the Barron-Jensen/Frankowska property. However, to conclude, we need to prove that the capture basin is the same under the original system and the two new ones. This is achieved by our proof; i.e., the capture basin being the same under the three systems, it captures these two properties, and thus, these two inequalities, each valid under the assumptions made (convexity *with* bounds for the one deriving from the viability theorem and Lipschitz property *without* bounds for the other one deriving from the invariance theorem).

It happens that the capture basin of the hypograph of  $\mathbf{c}$  viable in the hypograph of  $\mathbf{b}$  under control system (13) is still the capture basin under other auxiliary systems which satisfy these assumptions, so we shall be able to transfer the theorems concerning capture basins.

The function  $\psi$  being concave and finite, it is then continuous so that, the function  $v(\cdot)$  being bounded, the constant

$$\alpha := \sup_{u \in \text{Dom}(\varphi^*)} \varphi^*(u) - \inf_{\tau \geq 0} \psi(v(\tau))$$

is finite by Lemma 7.3. The new characteristic control systems are defined by

$$(23) \quad \begin{cases} \tau'(t) = -1, \\ x'(t) = u(t) \\ y'(t) = -\psi(v(\tau(t))) + \varphi^*(u(t)) + \pi(t) \end{cases} \quad \begin{array}{l} \text{where } u(t) \in \text{Dom}(\varphi^*), \\ \text{where } \pi(t) \in [0, \alpha + \psi(v(\tau(t)) - \varphi^*(u(t)))] \end{array}$$

and

$$(24) \quad \begin{cases} \tau'(t) = -1, \\ x'(t) = u(t) \\ y'(t) = -\psi(v(\tau(t))) + \varphi^*(u(t)) + \pi(t) \end{cases} \quad \begin{array}{l} \text{where } u(t) \in \text{Dom}(\varphi^*), \\ \text{where } \pi(t) \geq 0, \end{array}$$

where we added a new control  $\pi$  ranging over different intervals.

LEMMA 7.1 (equality between capture basins). *The capture basins of the hypograph of the function  $\mathbf{c}$  by systems (13), (23), and (24) coincide as follows:*

$$\text{Capt}_{(24)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c})) = \text{Capt}_{(23)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c})) = \text{Capt}_{(13)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c})).$$

Furthermore,

$$\text{Capt}_{(23)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c})) = \text{Capt}_{(23)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c})) - \{0\} \times \{0\} \times \mathbb{R}_+$$

(where in a vector space,  $A - B := \{a - b\}_{a \in A, b \in B}$ ).

*Proof.* Inclusions

$$\text{Capt}_{(13)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c})) \subset \text{Capt}_{(23)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c})) \subset \text{Capt}_{(24)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c}))$$

are obvious. For proving that

$$\text{Capt}_{(24)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c})) \subset \text{Capt}_{(13)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c})),$$

let us consider an element  $(t, x, y) \in \text{Capt}_{(24)}(\text{Hyp}(\mathbf{b}), \text{Hyp}(\mathbf{c}))$ . This means that there exist  $u(\cdot) \in L^1(0, +\infty; \text{Dom}(\varphi^*))$  and a corresponding solution  $(\tau(\cdot), x(\cdot), y(\cdot))$  to the characteristic control system (24) starting from  $(t, x, y)$  given by  $\tau(s) = t - s$ ,  $x(s) = x + \int_0^s u(r) dr$ , and

$$(25) \quad y(s) \geq y - \int_0^s (\psi(v(t-r)) - \varphi^*(u(r))) dr$$

and there exists  $t^* \in [0, t]$  such that  $(t-t^*, x(t^*), y(t^*)) \in \text{Hyp}(\mathbf{c})$  and, for all  $s \in [0, t^*]$ ,  $(t-s, x(s), y(s)) \in \text{Hyp}(\mathbf{b})$ . Setting

$$y_0(s) := y + \int_0^s (\varphi^*(u(r)) - \psi(v(t-r))) dr$$

we infer that  $(\tau(\cdot), x(\cdot), y_0(\cdot))$  is a solution to the characteristic control system (13) starting from  $(t, x, y)$  viable in the environment  $\mathcal{Hyp}(\mathbf{b})$  because

$$\forall s \in [0, t^*], \quad y_0(s) \leq y(s) \leq \mathbf{b}(t-s, x(s))$$

until time  $t^*$ , where it reaches the target  $\mathcal{Hyp}(\mathbf{c})$  because

$$y_0(t^*) \leq y(t^*) \leq \mathbf{c}(t-t^*, x(t^*)).$$

This means that  $(t, x, y) \in \text{Capt}_{(13)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c}))$ .

We also observe that whenever  $(t, x, y) \in \text{Capt}_{(24)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c}))$  and  $z \leq y$ , inequality (25) implies that

$$y(s) \geq y - \int_0^s (\psi(v(t-r)) - \varphi^*(u(r))) dr \geq z - \int_0^s (\psi(v(t-r)) - \varphi^*(u(r))) dr$$

and thus that  $(t, x, z)$  also belongs to the capture basin, so that,

$$\text{Capt}_{(23)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c})) = \text{Capt}_{(23)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c})) - \{0\} \times \{0\} \times \mathbb{R}_+.$$

The proof is completed.  $\square$

We also need the following.

**LEMMA 7.2.** *Let  $\psi : X \mapsto \mathbb{R}$  be an upper semicontinuous concave function. The domain of its Fenchel transform  $\varphi^*$  is contained in a closed convex subset  $A$  if and only if the function  $\psi$  satisfies inequality*

$$\exists \beta \in \mathbb{R} \text{ such that } \forall v \in X, \quad \beta - \sigma_A(v) \leq \psi(v).$$

*Its Fenchel transform  $\varphi^*$  is bounded on a convex subset  $A$  if and only if the function  $\psi$  satisfies*

$$\exists \delta \in \mathbb{R} \text{ such that } \forall v \in X^*, \quad \psi(v) \leq \delta - \sigma_A(v).$$

*Proof.* Since  $\psi(0) = \inf_{u \in \text{Dom}(\varphi^*)} \varphi^*(u)$ , we infer that

$$\forall v \in X, \quad \forall u \in \text{Dom}(\varphi^*), \quad \psi(0) - \sigma_{\text{Dom}(\varphi^*)}(v) \leq \varphi^*(u) - \langle u, v \rangle$$

so that, by taking the infimum over  $u$ , we obtain inequality  $\psi(0) - \sigma_{\text{Dom}(\varphi^*)}(v) \leq \psi(v)$ . It is enough to set  $\beta := \psi(0)$  and to take  $A := \text{Dom}(\varphi^*)$ . Conversely, assume that for all  $v \in X$ ,  $\psi(v) \geq \beta - \sigma_A(v)$ . We shall prove that  $\text{Dom}(\varphi^*) \subset A$ . If not, there would exist  $u \in \text{Dom}(\varphi^*) \setminus A$ . The separation theorem states there exist  $p_0 \in X$  and  $\varepsilon > 0$  such that  $\varepsilon \leq \langle p_0, u \rangle - \sigma_A(p_0)$ . Consequently, for every  $\lambda > 0$ ,

$$\lambda \varepsilon \leq \langle \lambda p_0, u \rangle - \sigma_A(\lambda p_0) \leq \langle \lambda p_0, u \rangle + \psi(\lambda p_0) - \beta \leq \varphi^*(u) - \beta$$

by assumption and by the definition of  $\varphi^*$ . Letting  $\lambda \mapsto +\infty$  implies that  $\varphi^*(u) = +\infty$ , i.e., that  $u \notin \text{Dom}(\varphi^*)$ , a contradiction.

For proving the second statement, we observe that if  $\delta := \sup_{u \in \text{Dom}(\varphi^*)} \varphi^*(u) < +\infty$  is finite, then

$$\psi(v) \leq \delta + \inf_{u \in \text{Dom}(\varphi^*)} \langle v, -u \rangle = \delta - \sigma_{\text{Dom}(\varphi^*)}(v)$$

so that the inequality holds true with  $A := \text{Dom}(\varphi^*)$ . Conversely, inequality  $\psi(v) \leq \delta - \sigma_A(v)$  implies that

$$\forall u \in A, \quad \varphi^*(u) \leq \sup_{v \in \text{Dom}(\psi)} [\langle v, u \rangle + \delta - \sigma_A(v)] < +\infty$$

is bounded on  $A$ , and thus, on  $\text{Dom}(\varphi^*)$  whenever this domain is contained in  $A$ .  $\square$

Control systems (23) and (24) are actually differential inclusions

$$(\tau'(t), x'(t), y'(t)) \in F(\tau(t), x(t), y(t)),$$

where

$$(26) \quad F(\tau, x, y) := \{(-1, u, -\psi(v(\tau)) + \varphi^*(u) + \pi)\}_{u \in \text{Dom}(\varphi^*), \pi \in [0, \alpha + \psi(v(\tau)) - \varphi^*(u)]}$$

and

$$(\tau'(t), x'(t), y'(t)) \in F_\infty(\tau(t), x(t), y(t)),$$

where

$$(27) \quad F_\infty(\tau, x, y) := \{(-1, u, -\psi(v(\tau)) + \varphi^*(u) + \pi)\}_{u \in \text{Dom}(\varphi^*), \pi \geq 0},$$

respectively.

LEMMA 7.3. *The set-valued map  $F$  is Marchaud and, if the functions  $\psi$ ,  $\varphi^*$ , and  $v$  are Lipschitz, the set-valued map  $F_\infty$  is Lipschitz with closed images.*

*Proof.* For proving that the set-valued map  $F$  is Marchaud, we shall check successively that

1. *the values  $F(\tau, x, y)$  of the set-valued map  $F$  are convex.* Indeed, for convex weight  $\lambda_i \geq 0$  such that  $\sum \lambda_i = 1$ , we can write

$$\sum \lambda_i (-1, u_i, -\psi(v(\tau)) + \varphi^*(u_i) + \pi_i) = (-1, \bar{u}, \varphi^*(\bar{u}) - \psi(v(\tau)) + \bar{\pi}),$$

where  $\bar{u} := \sum \lambda_i u_i$  and

$$\bar{\pi} := \sum \lambda_i \varphi^*(u_i) - \varphi^*\left(\sum \lambda_i u_i\right) + \sum \lambda_i \pi_i.$$

Since the domain of  $\varphi^*$  is convex,  $\bar{u} \in \text{Dom}(\varphi^*)$ . We observe that  $\bar{\pi}$  is non-negative and smaller than or equal to  $\alpha + \psi(v(\tau)) - \varphi^*(\bar{u})$  because

$$\begin{cases} \bar{\pi} \leq \sum \lambda_i \varphi^*(u_i) - \varphi^*(\sum \lambda_i u_i) + \sum \lambda_i (\alpha + \psi(v(\tau)) - \varphi^*(u_i)) \\ = \alpha + \psi(v(\tau)) - \varphi^*(\sum \lambda_i u_i). \end{cases}$$

2. *the graph of the set-valued map  $F$  is closed.* Indeed, let us consider a sequence of elements  $((\tau_n, x_n, y_n), (-1, u_n, \lambda_n))$  of the graph of  $F$  converging to  $((\tau, x, y), (-1, u, \lambda))$ , where  $\lambda_n := -\psi(v(\tau_n)) + \varphi^*(u_n) + \pi_n$  and where  $\pi_n \in [0, \alpha + \psi(v(\tau_n)) - \varphi^*(u_n)]$ .

Since the function  $(\tau, x, y, u) \mapsto \varphi^*(u) - \psi(v(\tau))$  is lower semicontinuous and since

$$(\tau_n, x_n, y_n, u_n, \lambda_n) = (\tau_n, x_n, y_n, u_n, -\psi(v(\tau_n)) + \varphi^*(u_n) + \pi_n)$$

belongs to the epigraph of this function (because  $\pi_n$  is positive by construction), which is closed, we deduce that the limit  $(\tau, x, y, u, \lambda)$  also belongs to this epigraph, i.e., that  $\lambda \geq \varphi^*(u) - \psi(v(\tau))$ . It is enough to set  $\pi := \lambda - \varphi^*(u) - \psi(v(\tau)) \geq 0$ , which from now on defines  $\pi$ . Recall that  $\pi_n = \lambda_n + \psi(v(\tau_n)) - \varphi^*(u_n) \leq \alpha + \mathbf{I}(\tau_n, x_n) - \varphi^*(u_n)$  by construction of  $\pi_n$ . Therefore,  $\lambda_n \leq \alpha$ . Therefore, taking the limit,  $\lambda = \pi + \varphi^*(u) - \psi(v(\tau)) \leq \alpha$ . In summary, the limit  $((\tau, x, y), (-1, u, \lambda))$  of elements  $((\tau_n, x_n, y_n), (-1, u_n, \lambda_n))$  belongs to the graph of  $F$  since  $\lambda = -\psi(v(\tau)) + \varphi^*(u) + \pi$ , where  $\pi \in [0, \alpha + \psi(v(\tau)) - \varphi^*(u)]$ .

3. *the images  $F(\tau, x, y)$  of  $F$  are bounded.* This follows from Lemma 7.2 because  $\text{Dom}(\varphi^*)$  is bounded and

$$\varphi^*(u) - \psi(v(\tau)) + \pi \leq \alpha := \sup_{u \in \text{Dom}(\varphi^*)} \varphi^*(u) - \inf_{\tau \geq 0} \psi(v(\tau))$$

is finite since  $\varphi^*$  is bounded above. Therefore

$$\|(-1, u, \varphi^*(u) - \psi(v(\tau)) + \pi)\| \leq \max(1, \|\text{Dom}(\varphi^*)\|, \alpha).$$

Hence, we have proved that the set-valued map  $F$  is Marchaud. The fact that  $F_\infty$  is Lipschitz is obvious since the functions  $\psi$ ,  $\varphi^*$ , and  $v$ , are assumed to be Lipschitz and since the controls  $u$  range over  $\mathbb{R}_+$  which, being constant, is Lipschitz.  $\square$

We thus deduce the following.

**PROPOSITION 7.4** (upper semicontinuity of the solution). *The viability hyposolution is upper semicontinuous and its hypograph satisfies*

$$\mathcal{Hyp}(\mathbf{N}) = \text{Capt}_{(23)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c})) = \text{Capt}_{(24)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c})).$$

*The viability hyposolution is concave whenever the functions  $\mathbf{b}$  and  $\mathbf{c}$  are concave.*

*Proof.* The first statement follows from Proposition 4.3 of [5] stating that under a Marchaud control system, the capture basin of a target is closed whenever the target  $\mathcal{Hyp}(\mathbf{c})$  and the environment  $\mathcal{Hyp}(\mathbf{b})$  are closed and the complement of the target in the environment is a repeller; this is the case because the first component of the system is  $\tau'(t) = -1$  which implies that all solutions  $(t - s, x(s), y(s))$  starting from any  $(t, x, y)$  leave  $\mathbb{R}_+ \times X \times \mathbb{R}$ , and thus,  $\mathcal{Hyp}(\mathbf{b}) \subset \mathbb{R}_+ \times X \times \mathbb{R}$ . Since we have proved that

$$\text{Capt}_{(23)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c})) = \text{Capt}_{(23)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c})) - \{0\} \times \{0\} \times \mathbb{R}_+,$$

we infer that  $\text{Capt}_{(23)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c}))$  is a hypograph, and thus, the hypograph of the viability hyposolution.  $\square$

**8. Contingent solution to the Hamilton–Jacobi equation.** We shall prove that the viability hyposolution to Hamilton–Jacobi equation (2) (see Definition 4.1) is the *contingent solution* by characterizing them in terms of tangent cones and translating them into terms of contingent hyposolutions.

**THEOREM 8.1** (contingent Frankowska solution). *The viability hyposolution  $\mathbf{N}$  is the largest upper semicontinuous solution satisfying*

$$(28) \quad \psi(v(t)) \geq \inf_{u \in \text{Dom}(\varphi^*)} (\varphi^*(u) - D_\downarrow \mathbf{N}(t, x)(-1, u))$$

*and the initial/Dirichlet conditions and the inequality constraints. If the functions  $\psi$ ,  $\varphi^*$ , and  $v$  are furthermore Lipschitz, then  $\mathbf{N}$  is the smallest upper semicontinuous solution satisfying the following:*

1. If  $\mathbf{N}(t, x) < \mathbf{b}(t, x)$ , then

$$(29) \quad \psi(v(t)) \leq \inf_{u \in \text{Dom}(\varphi^*)} (D_{\downarrow} \mathbf{N}(t, x)(1, -u) + \varphi^*(u)).$$

2. If  $\mathbf{N}(t, x) = \mathbf{b}(t, x)$ , then

$$(30) \quad \psi(v(t)) \leq \inf_{\{u | \psi(v(t)) \leq D_{\downarrow} \mathbf{b}(t, x)(1, -u) + \varphi^*(u)\}} (D_{\downarrow} \mathbf{N}(t, x)(1, -u) + \varphi^*(u)).$$

We need the following technical lemma on tangent cones to hypographs for proving Theorem 8.1.

LEMMA 8.2 (tangent cones to hypographs). *If  $\psi : X \mapsto \mathbf{R}_+ \cup \{-\infty\}$  is an extended function and if  $D_{\downarrow} \psi(p)(dp)$  is finite, then, for every  $w < \psi(p)$  and every  $\mu \in \mathbf{R}$ , the pair  $(dp, \mu)$  belongs to the contingent cone  $T_{\mathcal{Hyp}(\psi)}(p, w)$  to the hypograph of  $\psi$  at  $(p, w)$ .*

*Proof.* Let  $(dp, \lambda)$  belong to  $T_{\mathcal{Hyp}(\psi)}(p, \psi(p))$ . Then we know that there exist sequences  $h_n > 0$  converging to 0,  $dp_n$  converging to  $dp$ , and  $\lambda_n$  converging to  $\lambda$  such that  $(p + h_n dp_n, \psi(p) + h_n \lambda_n)$  belongs to  $\mathcal{Hyp}(\psi)$ . Therefore, for  $w < \psi(p)$  and  $\mu \in \mathbf{R}$  and  $h_n$  small enough,

$$(p + h_n dp_n, w + h_n \mu) = (p + h_n dp_n, \psi(p) + h_n \lambda_n) + (0, w - \psi(p) + h_n(\mu - \lambda_n)) \in \mathcal{Hyp}(\psi)$$

belongs to the hypograph of  $\psi$  because  $w - \psi(p) + h_n(\mu - \lambda_n) \leq 0$  for  $h_n$  small enough. Therefore, since  $dp_n \rightarrow dp$  and  $\mu_n := \mu \rightarrow \mu$ , we infer that  $(dp, \mu) \in T_{\mathcal{Hyp}(\psi)}(p, w)$ .  $\square$

*Proof of Theorem 8.1.* Observe first that

$$(t, x, y) \in \mathcal{Hyp}(\mathbf{N}) \setminus \mathcal{Hyp}(\mathbf{c}) \text{ if and only if } t > 0, x \in \text{Int}(K), \text{ and } y \leq \mathbf{N}(t, x).$$

Indeed,  $\mathcal{Hyp}(\mathbf{N}) \setminus \mathcal{Hyp}(\mathbf{c})$  is the set of  $(t, x, y)$  such that  $\mathbf{c}(t, x) < y \leq \mathbf{N}(t, x)$ . This is automatically satisfied when  $t > 0$  and  $x \in \text{Int}(K)$  whenever  $y \leq \mathbf{N}(t, x)$  since in this case,  $\mathbf{c}(t, x) = -\infty$ . It is impossible otherwise since, by Theorem 6.1  $\mathbf{N}(t, x) = \mathbf{c}(t, x)$ .

Theorem 4.6 of [7] states that since  $F$  is Marchaud by Lemma 7.3, the capture basin is the largest closed subset between the hypograph of  $\mathbf{c}$  and  $\mathbb{R}_+ \times X \times \mathbb{R}$  such that  $\mathcal{Hyp}(\mathbf{N}) \setminus \mathcal{Hyp}(\mathbf{c})$  is locally viable under  $F$ .

Theorems 3.2.4 and 3.3.4 of [7] state that  $\mathcal{Hyp}(\mathbf{N}) \setminus \mathcal{Hyp}(\mathbf{c})$  is locally viable under  $F$  if and only if for all  $t > 0$ , for all  $x \in X$ , for all  $y \leq \mathbf{N}(t, x)$ ,  $\exists u \in \text{Dom}(\varphi^*)$ ,  $\exists \pi \in [0, \alpha + \psi(v(t)) - \varphi^*(u)]$ , such that

$$(-1, u, -\psi(v(t)) + \varphi^*(u) + \pi) \in T_{\mathcal{Hyp}(\mathbf{N})}(t, x, y).$$

If  $y = \mathbf{N}(t, x)$ , then

$$T_{\mathcal{Hyp}(\mathbf{N})}(t, x, \mathbf{N}(t, x)) =: \mathcal{Hyp}(D_{\downarrow} \mathbf{N}(t, x))$$

so that we infer that there exists  $u \in \text{Dom}(\varphi^*)$

$$-\psi(v(t)) + \varphi^*(u) + \pi \leq D_{\downarrow} \mathbf{N}(t, x)(-1, u)$$

from which inequality (28) ensues.

Conversely, since  $D_{\downarrow} \mathbf{N}(t, x)(-1, \cdot)$  is upper semicontinuous and the domain of  $\varphi^*$  is compact, inequality (28) implies the existence of  $u \in \text{Dom}(\varphi^*)$  such that

$$(-1, u, -\psi(v(t)) + \varphi^*(u) + \pi) \in T_{\mathcal{Hyp}(\mathbf{N})}(t, x, \mathbf{N}(t, x)).$$



When  $y < \mathbf{N}(t, x)$ , then Lemma 8.2 implies that

$$(-1, u, -\psi(v(t)) + \varphi^*(u) + \pi) \in T_{\mathcal{Hyp}(\mathbf{N})}(t, x, y)$$

because  $(-1, u)$  belongs to the domain of  $D_{\downarrow} \mathbf{N}(t, x)$ .

By Theorems 4.7 and 4.10 of [7], the capture basin is the smallest closed subset between the hypographs of  $\mathbf{c}$  and  $\mathbf{b}$  such that  $\mathcal{Hyp}(\mathbf{N})$  is backward invariant with respect to  $\mathcal{Hyp}(\mathbf{b})$ . Since  $F_{\infty}$  is Lipschitz by Lemma 7.3 whenever the functions  $\psi$ ,  $\varphi^*$ , and  $v$  are Lipschitz, the invariance theorem (Theorem 5.3.4 in [5]) states that  $\mathcal{Hyp}(\mathbf{N})$  is backward invariant with respect to  $\mathcal{Hyp}(\mathbf{b})$  under  $F_{\infty}$  if and only if

$$\forall (t, x, y) \in \mathcal{Hyp}(\mathbf{N}), \quad F_{\infty}(t, x, y) \cap T_{\mathcal{Hyp}(\mathbf{b})}(t, x, y) \subset T_{\mathcal{Hyp}(\mathbf{N})}(t, x, y).$$

Since the function  $\mathbf{b}$  is assumed to be continuous,

$$\text{Int}(\mathcal{Hyp}(\mathbf{b})) = \{(t, x, y) \text{ such that } y < \mathbf{b}(t, x)\}.$$

Therefore, we have to investigate the following two cases:

1. For all  $(t, x, y) \in \mathcal{Hyp}(\mathbf{N}) \cap \text{Int}(\mathcal{Hyp}(\mathbf{b}))$ . Then

$$\begin{aligned} \forall t \geq 0, \quad \forall x \in X, \quad \forall y \leq \mathbf{N}(t, x), \quad \forall u \in \text{Dom}(\varphi^*), \quad \forall \pi \geq 0, \\ (1, -u, \psi(v(t)) - \varphi^*(u) - \pi) \in T_{\mathcal{Hyp}(\mathbf{N})}(t, x, y). \end{aligned}$$

If  $y = \mathbf{N}(t, x)$ , then we infer that for all  $u \in \text{Dom}(\varphi^*)$ ,

$$\psi(v(t)) - \varphi^*(u) \leq D_{\downarrow} \mathbf{N}(t, x)(1, -u)$$

from which we derive inequality (29). Conversely, since for all  $u \in \text{Dom}(\varphi^*)$ ,  $(1, -u)$  belongs to the domain of  $D_{\downarrow} \mathbf{N}(t, x)$ , we derive that

$$(1, -u, \psi(v(t)) - \varphi^*(u) - \pi) \in T_{\mathcal{Hyp}(\mathbf{N})}(t, x, y)$$

holds true.

2. For all  $(t, x, y) \in \mathcal{Hyp}(\mathbf{N}) \cap \partial(\mathcal{Hyp}(\mathbf{b}))$ , and in this case,  $y = \mathbf{N}(t, x) = \mathbf{b}(t, x)$ . Then,  $\forall t \geq 0, \forall x \in X, \forall u \in \text{Dom}(\varphi^*), \forall \pi \geq 0$  such that

$$(1, -u, \psi(v(t)) - \varphi^*(u) - \pi) \in T_{\mathcal{Hyp}(\mathbf{b})}(t, x, y)$$

we have

$$(1, -u, \psi(v(t)) - \varphi^*(u) - \pi) \in T_{\mathcal{Hyp}(\mathbf{N})}(t, x, y).$$

This means that whenever

$$\psi(v(t)) \leq D_{\downarrow} \mathbf{b}(t, x)(1, -u) + \varphi^*(u),$$

then

$$\psi(v(t)) \leq D_{\downarrow} \mathbf{N}(t, x)(1, -u) + \varphi^*(u),$$

which is (30).

Theorem 5.5 states that the viability hyposolution is the valuation function (21) of the underlying optimal control problem (13).

The *associated regulation map*  $R$  for regulating the optimal evolutions is thus defined by

$$\forall t > 0, x \in X, R(t, x) := \{u \mid 0 \leq D_{\downarrow} \mathbf{N}(t, x)(-1, u) - \varphi^*(u) + \psi(v(t))\}.$$

One can prove that the optimal solutions of the control problem are governed by the control system

$$\begin{cases} \tau'(s) = -1, \\ x'(s) = u(s) \in R(\tau(s), x(s)), \\ y'(s) = \varphi^*(u(s)) - \psi(v(\tau(s))). \end{cases}$$

This motivates a further study of the regulation map. If the solution  $\mathbf{N}$  is differentiable, the regulation map can be written in the form

$$R(t, x) := \left\{ u \mid 0 \leq -\frac{\partial \mathbf{N}(t, x)}{\partial t} + \frac{\partial \mathbf{N}(t, x)}{\partial x} u - \varphi^*(u) + \psi(v(t)) \right\}.$$

The elements  $u$  maximizing the right-hand side are the elements belonging to

$$-\partial_+ \psi \left( \frac{\partial \mathbf{N}(t, x)}{\partial x} \right).$$

Consequently,

$$-\partial_+ \psi \left( \frac{\partial \mathbf{N}(t, x)}{\partial x} \right) \subset R(t, x).$$

Actually, approximations of the regulation map and thus, optimal evolutions, as well as the solution to the Hamilton–Jacobi–Bellman equation are provided by the capture basin algorithm.

**9. Barron-Jensen/Frankowska solution to the Hamilton–Jacobi equation.** Instead of characterizing capture basins in terms of tangent cones and translating them into terms of contingent Frankowska hypolutions, we translate them into the equivalent formulation of Barron-Jensen/Frankowska solutions, a weaker concept of viscosity solutions requiring only the upper semicontinuity of the solution instead of its continuity.

**THEOREM 9.1** (Barron-Jensen/Frankowska solution). *The viability hyposolution  $\mathbf{N}$  is the largest upper semicontinuous solution between  $\mathbf{c}$  and  $\mathbf{b}$  satisfying*

$$(31) \quad \begin{cases} \text{(i)} & \forall t > 0, \forall x \in \text{Int}(K), \forall (p_t, p_x) \in \partial_+ \mathbf{N}(t, x), p_t + \psi(p_x) \leq \psi(v(t)), \\ \text{(ii)} & \forall t > 0, \forall x \in \text{Int}(K), \forall (p_t, p_x) \in (\text{Dom}(D_{\downarrow} \mathbf{N}(t, x)))^-, \\ & p_t - \sigma(\text{Dom}(\varphi^*), p_x) \leq 0. \end{cases}$$

*If the functions  $\psi$ ,  $\varphi^*$ , and  $v$  are furthermore Lipschitz, then  $\mathbf{N}$  is the smallest upper semicontinuous solution between  $\mathbf{c}$  and  $\mathbf{b}$  satisfying the following:*

1. *If  $\mathbf{N}(t, x) < \mathbf{b}(t, x)$ , then*

$$(32) \quad \begin{cases} \text{(i)} & \forall t \geq 0, \forall x \in K \text{ such that} \\ & \mathbf{N}(t, x) < \mathbf{b}(t, x), \forall (p_t, p_x) \in \partial_+ \mathbf{N}(t, x), \\ & p_t + \psi(p_x) \geq \psi(v(t)); \\ \text{(ii)} & \forall t \geq 0, \forall x \in K \text{ such that} \\ & \mathbf{N}(t, x) < \mathbf{b}(t, x), \forall (p_t, p_x) \in (\text{Dom}(D_{\downarrow} \mathbf{N}(t, x)))^-, \\ & p_t - \sigma(\text{Dom}(\varphi^*), p_x) \geq 0. \end{cases}$$

2. If  $\mathbf{N}(t, x) = \mathbf{b}(t, x)$ , then

$$(33) \quad \begin{cases} \forall (p_t, p_x) \in \partial_+ \mathbf{N}(t, x), \exists (q_t, q_x) \in \partial_+ \mathbf{b}(t, x) \text{ and } 0 < \mu < 1 \text{ such that} \\ \text{either } p_t - q_t - \sigma(\text{Dom}(\varphi^*), p_x - q_x) \geq 0 \\ \text{or } \frac{p_t - \mu q_t}{1 - \mu} + \psi\left(\frac{p_x - \mu q_x}{1 - \mu}\right) \geq \psi(v(t)). \end{cases}$$

Thus, the unique upper semicontinuous solution satisfies all these properties.

Observe that under the Lipschitz assumptions, the viability hypsolution satisfies

$$(34) \quad \begin{cases} \text{(i)} & \forall t > 0, \forall x \in \text{Int}(K) \text{ such that } \mathbf{N}(t, x) < \mathbf{b}(t, x), \\ & \forall (p_t, p_x) \in \partial_+ \mathbf{N}(t, x), p_t + \psi(p_x) = \psi(v(t)), \\ \text{(ii)} & \forall (p_t, p_x) \in (\text{Dom}(D_\downarrow \mathbf{N}(t, x)))^-, p_t - \sigma(\text{Dom}(\varphi^*), p_x) = 0. \end{cases}$$

We need the following technical lemma on normal cones to hypographs for proving Theorem 9.1.

LEMMA 9.2 (normal cones to hypographs). *A pair  $(u, \lambda)$  belongs to the normal cone  $N_{\mathcal{Hyp}(\psi)}(p, w)$  to the hypograph of  $\psi$  at  $(p, w)$  if and only*

1. *if  $w = \psi(p)$ ; then either*

- $\lambda = 0$  and  $u \in (\text{Dom}(D_\downarrow \psi(p)))^-$  or
- $\lambda > 0$  and  $u \in -\lambda \partial_+ \psi(p)$ .

2. *if  $w < \psi(p)$ ; then  $\lambda = 0$  and  $u \in (\text{Dom}(D_\downarrow \psi(p)))^-$ .*

*In particular, if the domain of  $D_\downarrow \psi(p)$  is dense in  $X$ , then  $(u, \lambda)$  belongs to the normal cone  $N_{\mathcal{Hyp}(\psi)}(p, w)$  to the hypograph of  $\psi$  at  $(p, w)$  if and only if  $\lambda = 0$  and  $u = 0$ . This is the case whenever  $\psi$  is Lipschitz around  $p$ .*

*Proof.* Let us consider now a pair  $(u, \lambda)$  belonging to the normal cone  $N_{\mathcal{Hyp}(\psi)}(p, w) := (T_{\mathcal{Hyp}(\psi)}(p, w))^-$  to the hypograph of  $\psi$  at  $(p, w)$ . Therefore,

$$\forall (dp, \mu) \in T_{\mathcal{Hyp}(\psi)}(p, w), \quad \langle (dp, \mu), (u, \lambda) \rangle = \langle u, dp \rangle + \lambda \mu \leq 0.$$

Examine first the case when  $w = \psi(p)$ , for which  $(dp, \mu) \in T_{\mathcal{Hyp}(\psi)}(p, \psi(p))$  if and only if  $dp \in \text{Dom}(D_\downarrow \psi(p))$  and  $\mu \leq D_\downarrow \psi(p)(dp)$ . If  $\lambda < 0$ , we obtain a contradiction because, when  $\mu \rightarrow -\infty$ ,  $\langle u, dp \rangle + \lambda \mu \rightarrow +\infty$ . Hence

- either  $\lambda > 0$ , and thus, dividing by  $\lambda$  and taking  $\mu := D_\downarrow \psi(p)(dp)$ , we obtain

$$\forall dp \in \text{Dom}(D_\downarrow \psi(p)), \quad \left\langle \frac{u}{\lambda}, dp \right\rangle + D_\downarrow \psi(p)(dp) \leq 0$$

which means that  $-\frac{u}{\lambda} \in \partial_+ \psi(p)$ ;

- or  $\lambda = 0$  and we obtain

$$\forall dp \in \text{Dom}(D_\downarrow \psi(p)), \quad \langle u, dp \rangle \leq 0,$$

which means that  $u \in (\text{Dom}(D_\downarrow \psi(p)))^-$  by definition of the polar cone.

When  $w < \psi(p)$ , inequalities

$$\forall (dp, \mu) \in T_{\mathcal{Hyp}(\psi)}(p, w), \quad \langle (dp, \mu), (u, \lambda) \rangle = \langle u, dp \rangle + \lambda \mu \leq 0$$

imply that  $\lambda = 0$  thanks to Lemma 8.2; otherwise,  $\lambda \mu$  converges to  $+\infty$  when  $\mu \rightarrow +\infty$  when  $\lambda > 0$ , and when  $\mu \rightarrow -\infty$  when  $\lambda < 0$  since  $\mu$  is allowed to range over  $\mathbb{R}$ . Therefore  $u \in (\text{Dom}(D_\downarrow \psi(p)))^-$  because whenever  $dp \in \text{Dom}(D_\downarrow \psi(p))$  and  $\mu \in \mathbb{R}$ ,

then  $(dp, \mu) \in T_{\mathcal{Hyp}(\psi)}(p, w)$ . If the domain  $D_{\downarrow}\psi(p)$  is dense in  $X$ , then the polar cone  $(\text{Dom}(D_{\downarrow}\psi(p)))^-$  is  $\{0\}$ , and thus  $u = 0$ .  $\square$

*Proof of Theorem 9.1.* Proposition 7.4 states that the hypograph of the viability hyposolution satisfies

$$\mathcal{Hyp}(\mathbf{N}) = \text{Capt}_{(23)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c})) = \text{Capt}_{(24)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c})).$$

Theorem 3.4 states that  $\text{Capt}_{(23)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c}))$  is the *largest* subset  $\mathcal{D}$  between  $\mathcal{C}$  and  $\mathcal{K}$  such that  $\mathcal{D} \setminus \mathcal{C}$  is locally viable.

Taking  $\mathcal{D} := \mathcal{Hyp}(\mathbf{N})$ , Theorems 3.2.4 and 3.3.4 of [5] state that  $\mathcal{Hyp}(\mathbf{N}) \setminus \mathcal{Hyp}(\mathbf{c})$  is locally viable under  $F$  if and only if for all  $t > 0$ , for all  $x \in \text{Int}(K)$ , for all  $y \leq \mathbf{N}(t, x)$ ,  $\exists u \in \text{Dom}(\varphi^*)$ ,  $\exists \pi \in [0, \alpha + \psi(v(t)) - \varphi^*(u)]$ , such that for all  $(-p_t, -p_x, \lambda) \in N_{\mathcal{Hyp}(\mathbf{N})}(t, x, y)$ ,

$$(35) \quad \begin{cases} \langle (-p_t, -p_x, \lambda), (-1, u, -\psi(v(t)) + \varphi^*(u) + \pi) \rangle \\ = p_t - \langle p_x, u \rangle + \lambda(-\psi(v(t)) + \varphi^*(u) + \pi) \leq 0. \end{cases}$$

By Lemma 9.2, if  $y = \mathbf{N}(t, x)$ ,  $(-p_t, -p_x, \lambda) \in N_{\mathcal{Hyp}(\mathbf{N})}(t, x, y)$  means that either  $\lambda > 0$ , and that, taking  $\lambda = 1$ ,  $(p_t, p_x) \in \partial_+ \mathbf{N}(t, x)$ , or that  $\lambda = 0$ , and that  $(p_t, p_x) \in (\text{Dom}(D_{\downarrow} \mathbf{N}(t, x)))^-$ . If  $y < \mathbf{N}(t, x)$ ,  $(-p_t, -p_x, \lambda) \in N_{\mathcal{Hyp}(\mathbf{N})}(t, x, y)$  means also that  $\lambda = 0$ , and that  $(p_t, p_x) \in (\text{Dom}(D_{\downarrow} \mathbf{N}(t, x)))^-$ .

Consequently, condition (35) can be written in the following form:

- The case when  $y = \mathbf{N}(t, x)$  and  $\lambda = 1$ :

$$\begin{cases} \forall t > 0, \forall x \in \text{Int}(K), \forall (p_t, p_x) \in \partial_+ \mathbf{N}(t, x), \text{ then} \\ p_t - \psi(v(t)) + \inf_{u \in \text{Dom}(\varphi^*)} [\varphi^*(u) - \langle p_x, u \rangle] \\ = p_t - \psi(v(t)) + \psi(p_x) \leq 0. \end{cases}$$

- The case when  $y \leq \mathbf{N}(t, x)$  and  $\lambda = 0$ :

$$\begin{cases} \forall t > 0, \forall x \in \text{Int}(K), \forall (p_t, p_x) \in (\text{Dom}(D_{\downarrow} \mathbf{N}(t, x)))^-, \text{ then} \\ p_t - \sup_{u \in \text{Dom}(\varphi^*)} \langle p_x, u \rangle = p_t - \sigma(\text{Dom}(\varphi^*), p_x) \leq 0. \end{cases}$$

(Recall that this condition disappears whenever the viability hyposolution  $\mathbf{N}$  is hypodifferentiable, and, in particular, when the hyposolution is Lipschitz.)

*Proof of inequalities (32) and (33).* Theorem 3.4 states that  $\text{Capt}_{(24)}(\mathcal{Hyp}(\mathbf{b}), \mathcal{Hyp}(\mathbf{c}))$  is the *smallest* subset  $\mathcal{D}$  between  $\mathcal{C}$  and  $\mathcal{K}$  such that  $\mathcal{D}$  is backward invariant with respect to  $\mathcal{K}$ . Theorem 3.7 and Lemma 3.8 state that  $\mathcal{D} := \mathcal{Hyp}(\mathbf{N})$  is backward invariant with respect to  $\mathcal{Hyp}(\mathbf{b})$  under (24) if and only if one of the following holds:

1. For all  $(t, x, y) \in \mathcal{Hyp}(\mathbf{N}) \cap \text{Int}(\mathcal{Hyp}(\mathbf{b}))$ ,

$$\forall (-p_t - p_x, \lambda) \in N_{\mathcal{Hyp}(\mathbf{N})}(t, x, y), \quad \sigma(F_{\infty}(x), (p_t, p_x, -\lambda)) \leq 0.$$

Since the function  $\mathbf{b}$  is assumed to be continuous,

$$\text{Int}(\mathcal{Hyp}(\mathbf{b})) = \{(t, x, y) \text{ such that } y < \mathbf{b}(t, x)\},$$

the first case means that  $y \leq \mathbf{N}(t, x) < \mathbf{b}(t, x)$  and the above condition implies that

$$(36) \quad \begin{cases} \forall (-p_t - p_x, \lambda) \in N_{\mathcal{Hyp}(\mathbf{N})}(t, x, y), \\ \langle (p_t, p_x, -\lambda), (-1, u, -\psi(v(t)) + \varphi^*(u) + \pi) \rangle \\ = -p_t + \langle p_x, u \rangle + \lambda(\psi(v(t)) - \varphi^*(u) - \pi) \leq 0. \end{cases}$$

This implies that  $\lambda \geq 0$ .

Consequently, condition (36) can be written in the following form:

- The case when  $y = \mathbf{N}(t, x) < \mathbf{b}(t, x)$  and  $\lambda = 1$ :

$$\begin{cases} \forall t > 0, \forall x \in X, \forall (p_t, p_x) \in \partial_+ \mathbf{N}(t, x), \text{ then} \\ -p_t + \psi(v(t)) + \sup_{u \in \text{Dom}(\varphi^*)} [\langle p_x, u \rangle - \varphi^*(u)] \\ = -p_t + \psi(v(t)) - \psi(p_x) \leq 0. \end{cases}$$

- The case when  $y \leq \mathbf{N}(t, x)$  and  $\lambda = 0$ :

$$\begin{cases} \forall t > 0, \forall x \in X, \forall (p_t, p_x) \in (\text{Dom}(D_+ \mathbf{N}(t, x)))^-, \text{ then} \\ -p_t + \sup_{u \in \text{Dom}(\varphi^*)} \langle p_x, u \rangle = -p_t + \sigma(\text{Dom}(\varphi^*), p_x) \leq 0. \end{cases}$$

2. For all  $(t, x, y) \in \mathcal{Hyp}(\mathbf{N}) \cap \partial(\mathcal{Hyp}(\mathbf{b}))$ , and in this case,  $y = \mathbf{N}(t, x) = \mathbf{b}(t, x)$  and

$$\begin{cases} \forall (-p_t - p_x, \lambda) \in N_{\mathcal{Hyp}(\mathbf{N})}(t, x, y), \exists (-q_t - q_x, \mu) \in N_{\mathcal{Hyp}(\mathbf{b})}(t, x, y) \\ \text{such that } \sigma(F_\infty(x), (p_t - q_t, p_x - q_x, \mu - \lambda)) \leq 0, \end{cases}$$

where  $\lambda \geq 0$  and  $\mu > 0$  since we have assumed that  $\mathbf{b}$  is Lipschitz, and thus hypodifferentiable. This can be translated into the following form:

$$-p_t + q_t + \sup_u (\langle p_x - q_x, u \rangle + (\mu - \lambda)(\varphi^*(u)) + \sup_{\pi \geq 0} (\mu - \lambda)[\pi - \psi(v(t))]) \leq 0.$$

This implies that  $\lambda \geq \mu > 0$ .

- The case when  $\lambda - \mu = 0$ . It happens when both  $(p_t, p_x) \in \partial_+ \mathbf{N}(t, x)$  and  $(q_t, q_x) \in \partial_+ \mathbf{b}(t, x)$ . In this case, the above inequality boils down to

$$-p_t + q_t + \sigma(\text{Dom}(\varphi^*), p_x - q_x) \leq 0.$$

- The case when  $\lambda - \mu > 0$ . The condition states that for every  $\lambda > 0$  and  $(p_t, p_x) \in \partial_+ \mathbf{N}(t, x)$ , there exist  $0 < \mu < \lambda$  and  $(q_t, q_x) \in \partial_+ \mathbf{b}(t, x)$  such that

$$-\frac{\lambda p_t - \mu q_t}{\lambda - \mu} + \sup_u \left( \left\langle \frac{\lambda p_x - \mu q_x}{\lambda - \mu}, u \right\rangle - \varphi^*(u) \right) + \psi(v(t)) \leq 0,$$

which can be written

$$-\frac{\lambda p_t - \mu q_t}{\lambda - \mu} - \psi \left( \frac{\lambda p_x - \mu q_x}{\lambda - \mu} \right) + \psi(v(t)) \leq 0.$$

This completes the proof.  $\square$

## REFERENCES

- [1] O. ALVAREZ, E. N. BARRON, and H. ISHII, *Hopf-Lax formulas for semicontinuous data*, Indiana Univ. Math. J., 48 (1999), pp. 993–1035.
- [2] R. ANSORGE, *What does the entropy condition mean in traffic flow theory?*, Transportation Res. Part B, 24 (1990), pp. 133–143.
- [3] J. A. ATWELL, J. T. BORGGAARD, and B. B. KING, *Reduced order controllers for Burgers' equation with a nonlinear observer*, Appl. Math. Comput. Sci., 11 (2001), pp. 1311–1330.
- [4] J.-P. AUBIN, *Contingent derivatives of set-valued maps and existence of solutions to nonlinear inclusions and differential inclusions*, in Mathematical Analysis and Applications, Part A, Adv. in Math. Suppl. Stud. 7a, L. Nachbin, ed., Academic Press, New York, 1981, pp. 159–229.

- [5] J.-P. AUBIN, *Viability Theory*, Systems and Control: Foundations and Applications, Birkhäuser Boston, Boston, MA, 1991.
- [6] J.-P. AUBIN, *Optima and Equilibria*, Springer-Verlag, New York, 1993.
- [7] J.-P. AUBIN, *Viability kernels and capture basins of sets under differential inclusions*, SIAM J. Control Optim., 40 (2001), pp. 853–881.
- [8] J.-P. AUBIN, *Boundary-value problems for systems of first-order partial differential inclusions with constraints*, Progr. Nonlinear Differential Equations Appl., 55 (2003), pp. 25–60.
- [9] J.-P. AUBIN, A. M. BAYEN, N. BONNEUIL, and P. SAINT-PIERRE, *Viability, Control, and Games*, Springer-Verlag, New York, to appear.
- [10] J.-P. AUBIN and F. CATTE, *Bilateral fixed-point and algebraic properties of viability kernels and capture basins of sets*, Set-Valued Anal., 10 (2002), pp. 379–416.
- [11] J.-P. AUBIN and G. DA PRATO, *Solutions contingentes de l'équation de la variété centrale*, C. R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 295–300.
- [12] J.-P. AUBIN and G. DA PRATO, *Contingent solutions to the center manifold equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 9 (1992), pp. 13–28.
- [13] J.-P. AUBIN and H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser Boston, Boston, MA, 1990.
- [14] J.-P. AUBIN and H. FRANKOWSKA, *Inclusions aux dérivées partielles gouvernant des contrôles de rétroaction*, C. R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 851–856.
- [15] J.-P. AUBIN and H. FRANKOWSKA, *Systèmes hyperboliques d'inclusions aux dérivées partielles*, C. R. Acad. Sci. Paris Sér. I Math., 312 (1991), pp. 271–276.
- [16] J.-P. AUBIN and H. FRANKOWSKA, *Hyperbolic systems of partial differential inclusions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 18 (1991), pp. 541–562.
- [17] J.-P. AUBIN and H. FRANKOWSKA, *Partial differential inclusions governing feedback controls*, J. Convex Anal., 2 (1995), pp. 19–40.
- [18] J.-P. AUBIN and H. FRANKOWSKA, *The viability kernel algorithm for computing value functions of infinite horizon optimal control problems*, J. Math. Anal. Appl., 201 (1996), pp. 555–576.
- [19] J.-P. AUBIN and H. FRANKOWSKA, *Set-valued solutions to the Cauchy problem for hyperbolic systems of partial differential inclusions*, NoDEA Nonlinear Differential Equations Appl., 4 (1999), pp. 149–168.
- [20] J. BAKER, A. ARMAOU, and P. D. CHRISTOFIDES, *Nonlinear control of incompressible fluid flow: Application to Burgers' equation and 2d channel flow*, J. Math. Anal. Appl., 252 (2000), pp. 230–255.
- [21] A. BALOGH and M. KRSTIC, *Burgers' equation with nonlinear boundary feedback:  $H_1$  stability, well posedness, and simulation*, Math. Prob. in Engrg., 6 (2000), pp. 189–200.
- [22] M. BARDI and I. CAPUZZO-DOLCETTA, *Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, MA, 1997.
- [23] M. BARDI and L. EVANS, *On Hopf's formulas for solutions of Hamilton-Jacobi equations*, Nonlinear Anal. Theory Methods Appl., 8 (1984), pp. 1373–1381.
- [24] C. BARDOS, A. Y. LEROUX, and J. C. NEDELEC, *First order quasilinear equations with boundary conditions*, Commun. Partial Differential Equations, 4 (1979), pp. 1017–1034.
- [25] G. BARLES, *Solutions de viscosité des équations de Hamilton-Jacobi*, Springer, Paris, 1994.
- [26] E. N. BARRON and R. JENSEN, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians*, Commun. Partial Differential Equations, 15 (1990), pp. 1713–1742.
- [27] E. N. BARRON and R. JENSEN, *Optimal control and semicontinuous viscosity solutions*, Proc. Amer. Math. Soc., 113 (1991), pp. 393–402.
- [28] E. N. BARRON, R. JENSEN, and W. LIU, *Hopf-Lax-type formula for  $u_t + H(u, Du) = 0$* , J. Differential Equations, 126 (1996), pp. 48–61.
- [29] A. M. BAYEN, C. CLAUDEL, and P. SAINT-PIERRE, *Computations of solutions to the Moskowitz Hamilton-Jacobi-Bellman equation under viability constraints*, in Proceedings of the 46th IEEE Conference on Decision and Control (CDC), New Orleans, LA, 2007, pp. 4737–4742.
- [30] A. M. BAYEN, C. CLAUDEL, and P. SAINT-PIERRE, *Viability-based computations of solutions to the Hamilton-Jacobi-Bellman equation*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 4416, G. Buttazzo, A. Bemporad, and A. Bicchi, eds., Springer-Verlag, New York, 2007, pp. 645–649.
- [31] A. M. BAYEN, R. L. RAFFARD, and C. TOMLIN, *Network congestion alleviation using adjoint hybrid control: Application to highways*, in Hybrid Systems, Computation and Control, Lecture Notes in Comput. Sci. 1790, Springer-Verlag, New York, 2004, pp. 95–110.
- [32] T. R. BEWLEY, *Flow control: New challenges for a new renaissance*, Progress in Aerospace Science, 37 (2001), pp. 21–58.
- [33] E. BOURREL and J.-B. LESORT, *Mixing micro and macro representations of traffic flow: A hybrid model based on the LWR theory*, in Proceedings of the 82nd Meeting of the Transportation Research Board, Washington, D.C., 2003.

- [34] C. BYRNES and H. FRANKOWSKA, *Unicité des solutions optimales et absence de chocs pour les équations d'Hamilton–Jacobi–Bellman et de Riccati*, C. R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 427–431.
- [35] C. BYRNES and H. FRANKOWSKA, *Uniqueness of optimal trajectories and the nonexistence of shocks for Hamilton–Jacobi–Bellman and Riccati partial differential equations*, in Differential Inclusions and Optimal Control, Lecture Notes in Nonlinear Anal., L. Gorniewicz, J. Andres, and P. Nistri, eds., J. Schauder Center for Nonlinear Studies, Los Alamos, NM, 1998, pp. 89–112.
- [36] C. I. BYRNES, D. S. GILLIAM, and V. I. SHUBOV, *Semiglobal stabilization of a boundary controlled viscous Burgers' equation*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 680–681.
- [37] P. CARDALIAGUET, M. QUINCAMPOIX, and P. SAINT-PIERRE, *Set-valued numerical analysis for optimal control and differential games*, in Stochastic and Differential Games: Theory and Numerical Methods, Ann Internat. Soc. Dynam. Games 4, M. Bardi, T. E. S. Raghavan, and T. Parthasarathy, eds., Birkhäuser Boston, Boston, MA, 1999, pp. 177–247.
- [38] P. CARDALIAGUET, M. QUINCAMPOIX, and P. SAINT-PIERRE, *Set-valued numerical analysis for optimal control and differential games*, in Stochastic and Differential Games: Theory and Numerical Methods, Annals of the International Society of Dynamic Games, M. Bardi, T. E. S. Raghavan, and T. Parthasarathy, eds., Birkhäuser Boston, Boston, MA, 1999, pp. 177–247.
- [39] N. CAROFF and H. FRANKOWSKA, *Optimality and characteristics of Hamilton–Jacobi–Bellman equations*, Int. Ser. Numer. Math., 107 (1992), pp. 169–180.
- [40] N. CAROFF and H. FRANKOWSKA, *A note on conjugate points and shocks in nonlinear optimal control*, Bull. Polish Acad. Sci., 42 (1994), pp. 115–128.
- [41] N. CAROFF and H. FRANKOWSKA, *Conjugate points and shocks in nonlinear optimal control*, Trans. Amer. Math. Soc., 348 (1996), pp. 3133–3153.
- [42] P. D. CHRISTOFIDES, *Nonlinear and Robust Control of Partial Differential Equation Systems: Methods and Applications to Transport-Reaction Processes*, Birkhäuser Boston, Boston, MA, 2001.
- [43] M. G. CRANDALL, L. C. EVANS, and P.-L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
- [44] M. G. CRANDALL and P.-L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
- [45] C. DAGANZO, *The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory*, Transportation Res. Part B, 28 (1994), pp. 269–287.
- [46] C. DAGANZO, *The cell transmission model, part II: Network traffic*, Transportation Res. Part B, 29 (1995), pp. 79–93.
- [47] W. B. DUNBAR, N. PETIT, P. ROUCHON, and P. MARTIN, *Boundary control of a nonlinear Stefan problem*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 1309–1314.
- [48] W. B. DUNBAR, N. PETIT, P. ROUCHON, and P. MARTIN, *Motion planning for a nonlinear Stefan problem*, ESAIM: Control Optim. Calc. Var., 9 (2003), pp. 275–296.
- [49] L. C. EVANS, *Partial Differential Equations*, AMS, Providence, RI, 1998.
- [50] M. FLIESS, J. LÉVINE, Ph. MARTIN, and P. ROUCHON, *Flatness and defect of nonlinear systems: Introductory theory and examples*, Int. J. Control, 60 (1995), pp. 1327–1361.
- [51] H. FRANKOWSKA, *L'équation d'Hamilton–Jacobi contingente*, C. R. Acad. Sci. Paris Sér. I Math., 304 (1987), pp. 295–298.
- [52] H. FRANKOWSKA, *Optimal trajectories associated to a solution of contingent Hamilton–Jacobi equations*, in Proceedings of the 26th IEEE Conference on Decision and Control, Los Angeles, CA, 1987, pp. 727–732.
- [53] H. FRANKOWSKA, *Optimal trajectories associated to a solution of contingent Hamilton–Jacobi equations*, Appl. Math. Optim., 19 (1989), pp. 291–311.
- [54] H. FRANKOWSKA, *Hamilton–Jacobi equation: Viscosity solutions and generalized gradients*, J. Math. Anal. Appl., 141 (1989), pp. 21–26.
- [55] H. FRANKOWSKA, *Lower semicontinuous solutions to Hamilton–Jacobi–Bellman equations*, in Proceedings of the 30th IEEE Conference on Decision and Control, Brighton, UK, 1991, pp. 265–270.
- [56] H. FRANKOWSKA, *Lower semicontinuous solutions of Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.
- [57] M. GUGAT, M. HERTY, A. KLAR, and G. LEUGERING, *Optimal control for traffic flow networks*, J. Optim. Theory Appl., 126 (2005), pp. 589–616.
- [58] A. JAMESON, *Analysis and design of numerical schemes for gas dynamics 1: Artificial diffusion, upwind biasing, limiters and their effect on accuracy and multigrid convergence*, Int. J. Comput. Fluid Dynam., 4 (1995), pp. 171–218.

- [59] A. JAMESON, *Analysis and design of numerical schemes for gas dynamics 2: Artificial diffusion and discrete shock structure*, Int. J. Comput. Fluid Dynam., 4 (1995), pp. 1–38.
- [60] Z. JIA, C. CHEN, B. COIFMAN, and P. VARAIYA, *The PeMS algorithms for accurate, real time estimates of g-factors and speeds from single loop detectors*, in Proceedings of the IEEE Intelligent Transportation Systems Conference, Oakland, CA, 2001, pp. 536–541.
- [61] T. KOBAYASHI, *Adaptive regulator design of a viscous Burgers' system by boundary control*, IMA J. Math. Control Inform., 18 (2001), pp. 427–437.
- [62] M. KRSTIC, *On global stabilization of Burgers' equation by boundary control*, Systems Control Lett., 37 (1999), pp. 123–142.
- [63] J.-P. LEBACQUE and J.-B. LESORT, *Macroscopic traffic flow models: A question of order*, in Proceedings of the 14th Transportation and Traffic Theory (ISTTT), Jerusalem, Israel, 1999, pp. 3–25.
- [64] M. J. LIGHTHILL and G. B. WHITHAM, *On kinematic waves. II. A theory of traffic flow on long crowded roads*, Proc. Roy. Soc. London, 229 (1956), pp. 317–345.
- [65] P.-L. LIONS and J.-C. ROCHET, *Hopf formula and multitime Hamilton–Jacobi equations*, Proc. Amer. Math. Soc., 96 (1986), pp. 79–84.
- [66] H. LY, K. D. MEASE, and E. S. TITI, *Distributed and boundary control of the viscous Burgers' equation*, Numer. Funct. Anal. Optim., 18 (1997), pp. 143–188.
- [67] I. MITCHELL, *A Toolbox of Level Set Methods*, <http://www.cs.ubc.ca/~mitchell> (2005).
- [68] I. MITCHELL, *Application of Level Set Methods to Control and Reachability Problems in Continuous and Hybrid Systems*, Ph.D. thesis, Stanford University, Stanford, CA, 2000.
- [69] I. MITCHELL, A. M. BAYEN, and C. J. TOMLIN, *A time-dependent Hamilton–Jacobi formulation of reachable sets for continuous dynamic games*, IEEE Trans. Automat. Control, 50 (2005), pp. 947–957.
- [70] G. F. NEWELL, *A simplified theory of kinematic waves in highway traffic, part I: General theory*, Transportation Res. Part B, 27 (1993), pp. 281–287.
- [71] G. F. NEWELL, *A simplified theory of kinematic waves in highway traffic, part II: Queueing at freeway bottlenecks*, Transportation Res. Part B, 27 (1993), pp. 289–303.
- [72] G. F. NEWELL, *A simplified theory of kinematic waves in highway traffic, part III: Multi-destination flows*, Transportation Res. Part B, 27 (1993), pp. 303–313.
- [73] O. A. OLEINIK, *Discontinuous solutions of nonlinear differential equations*, Uspekhi Mat. Nauk, 12 (1957), pp. 3–73 (in Russian). English translation in Amer. Math. Soc. Transl. (2), 26 (1963), pp. 95–172.
- [74] M. ÖNDER and E. ÖZBAY, *Low dimensional modelling and Dirichlet boundary controller design for Burgers equation*, Internat. J. Control, 77 (2004), pp. 895–906.
- [75] N. PETIT, *Delay Systems. Flatness in Process Control and Control of Some Wave Equations*, Ph.D. thesis, Ecole des Mines de Paris, Paris, France, 2000.
- [76] S. PLASKACZ and M. QUINCAMPOIX, *Oleinik–Lax formulas and multitime Hamilton–Jacobi systems*, to appear.
- [77] S. PLASKACZ and M. QUINCAMPOIX, *On representation formulas for Hamilton–Jacobi equations related to calculus of variation problems*, to appear.
- [78] P. I. RICHARDS, *Shock waves on the highway*, Oper. Res., 4 (1956), pp. 42–51.
- [79] R. T. ROCKAFELLAR and R. WETS, *Variational Analysis*, Springer-Verlag, New York, 1997.
- [80] P. SAINT-PIERRE, *Approximation of the viability kernel*, Appl. Math. Optim., 29 (1994), pp. 187–209.
- [81] P. SAINT-PIERRE, *Approximation of capture basins for hybrid systems*, in Hybrid Systems: Computation and Control, Proceedings of the HSCC 2002 Conference, Lecture Notes in Comput. Sci. 2034, Springer-Verlag, New York, 2002, pp. 378–392.
- [82] R. C. SMITH and M. A. DEMETRIOU, *Research Directions in Distributed Parameter Systems*, SIAM, Philadelphia, 2003.
- [83] I. S. STRUB and A. M. BAYEN, *Weak formulation of boundary conditions for scalar conservation laws: An application to highway modeling*, Int. J. Robust Nonlinear Control, 16 (2006), pp. 733–748.
- [84] A. I. SUBBOTIN, *Generalization of the main equation of differential game theory*, J. Optim. Theory Appl., 43 (1984), pp. 103–133.
- [85] A. I. SUBBOTIN, *Generalized Solutions of First Order PDEs: The Dynamical Optimization Perspective*, Birkhäuser Boston, Boston, MA, 1995.
- [86] J. YI, H. LIN, L. ALVAREZ, and R. HOROWITZ, *Stability of macroscopic traffic flow through wavefront expansion*, in American Control Conference, Anchorage, AK, 2002, pp. 1484–1490.
- [87] J. YI, H. LIN, L. ALVAREZ, and R. HOROWITZ, *Stability of macroscopic traffic flow modeling through wavefront expansion*, Transportation Res. Part B, 37 (2003), pp. 661–679.



## MEAN-VARIANCE HEDGING UNDER PARTIAL INFORMATION\*

MICHAEL MANIA<sup>†</sup>, REVAZ TEVZADZE<sup>‡</sup>, AND TEIMURAZ TORONJADZE<sup>†</sup>

**Abstract.** We consider the mean-variance hedging problem under partial information. The underlying asset price process follows a continuous semimartingale, and strategies have to be constructed when only part of the information in the market is available. We show that the initial mean-variance hedging problem is equivalent to a new mean-variance hedging problem with an additional correction term, which is formulated in terms of observable processes. We prove that the value process of the reduced problem is a square trinomial with coefficients satisfying a triangle system of backward stochastic differential equations and the filtered wealth process of the optimal hedging strategy is characterized as a solution of a linear forward equation.

**Key words.** backward stochastic differential equation, semimartingale market model, incomplete markets, mean-variance hedging, partial information

**AMS subject classifications.** 90A09, 60H30, 90C39

**DOI.** 10.1137/070700061

**1. Introduction.** In the problem of derivative pricing and hedging it is usually assumed that the hedging strategies have to be constructed by using all market information. However, in reality, investors acting in a market have limited access to the information flow. For example, an investor may observe just stock prices, but stock appreciation rates depend on some unobservable factors; one may think that stock prices can be observed only at some time intervals or up to some random moment before an expiration date, or an investor would like to price and hedge a contingent claim whose payoff depends on an unobservable asset, and he observes the prices of an asset correlated with the underlying asset. Besides, investors may not be able to use all available information even if they have access to the full market flow. In all such cases, investors are forced to make decisions based on only a part of the market information.

We study a mean-variance hedging problem under partial information when the asset price process is a continuous semimartingale and the flow of observable events do not necessarily contain all information on prices of the underlying asset.

We assume that the dynamics of the price process of the asset traded on the market is described by a continuous semimartingale  $S = (S_t, t \in [0, T])$  defined on a filtered probability space  $(\Omega, \mathcal{A}, (\mathcal{A}_t, t \in [0, T]), P)$ , satisfying the usual conditions, where  $\mathcal{A} = \mathcal{A}_T$  and  $T < \infty$  is the fixed time horizon. Suppose that the interest rate is equal to zero and the asset price process satisfies the structure condition; i.e., the process  $S$  admits the decomposition

$$(1.1) \quad S_t = S_0 + N_t + \int_0^t \lambda_u d\langle N \rangle_u, \quad \langle \lambda \cdot N \rangle_T < \infty \quad \text{a.s.},$$

where  $N$  is a continuous  $\mathcal{A}$ -local martingale and  $\lambda$  is an  $\mathcal{A}$ -predictable process.

---

\*Received by the editors August 14, 2007; accepted for publication (in revised form) April 28, 2008; published electronically September 8, 2008. This work was supported by Georgian National Science Foundation grant STO07/3-172.

<http://www.siam.org/journals/sicon/47-5/70006.html>

<sup>†</sup>Business School, Georgian American University, 3, Alleyway II, Chavchavadze Ave. 17, A, Tbilisi, Georgia, and A. Razmadze Mathematical Institute, 1, M. Aleksidze St., Tbilisi, Georgia (misha.mania@gmail.com, toronj333@yahoo.com).

<sup>‡</sup>Business School, Georgian American University, 3, Alleyway II, Chavchavadze Ave. 17, A, Tbilisi, Georgia, and Institute of Cybernetics, 5, S. Euli St., Tbilisi, Georgia (tevdzadze@cybernet.ge).

Let  $G$  be a filtration smaller than  $\mathcal{A}$ :

$$G_t \subseteq \mathcal{A}_t \quad \text{for every } t \in [0, T].$$

The filtration  $G$  represents the information that the hedger has at his disposal; i.e., hedging strategies have to be constructed using only information available in  $G$ .

Let  $H$  be a  $P$ -square integrable  $\mathcal{A}_T$ -measurable random variable, representing the payoff of a contingent claim at time  $T$ .

We consider the mean-variance hedging problem

$$(1.2) \quad \text{to minimize } E[(X_T^{x,\pi} - H)^2] \quad \text{over all } \pi \in \Pi(G),$$

where  $\Pi(G)$  is a class of  $G$ -predictable  $S$ -integrable processes. Here  $X_t^{x,\pi} = x + \int_0^t \pi_u dS_u$  is the wealth process starting from initial capital  $x$ , determined by the self-financing trading strategy  $\pi \in \Pi(G)$ .

In the case  $G = \mathcal{A}$  of complete information, the mean-variance hedging problem was introduced by Föllmer and Sondermann [8] in the case when  $S$  is a martingale and then developed by several authors for a price process admitting a trend (see, e.g., [6], [12], [25], [26], [24], [10], [11]).

Asset pricing with partial information under various setups has been considered. The mean-variance hedging problem under partial information was first studied by Di Masi, Platen, and Runggaldier [3] when the stock price process is a martingale and the prices are observed only at discrete time moments. For general filtrations and when the asset price process is a martingale, this problem was solved by Schweizer [27] in terms of  $G$ -predictable projections. Pham [22] considered the mean-variance hedging problem for a general semimartingale model, assuming that the observable filtration contains the augmented filtration  $F^S$  generated by the asset price process  $S$

$$(1.3) \quad F_t^S \subseteq G_t \quad \text{for every } t \in [0, T].$$

In this paper, using the variance-optimal martingale measure with respect to the filtration  $G$  and suitable Kunita–Watanabe decomposition, the theory developed by Gouriéroux, Laurent, and Pham [10] and Rheinländer and Schweizer [23] to the case of partial information was extended.

If  $G$  is not containing  $F^S$ , then  $S$  is not a  $G$ -semimartingale and the problem is more involved. Let us introduce an additional filtration  $F = (F_t, t \in [0, T])$ , which is an augmented filtration generated by  $F^S$  and  $G$ .

Then the price process  $S$  is a continuous  $F$ -semimartingale, and the canonical decomposition of  $S$  with respect to the filtration  $F$  is of the form

$$(1.4) \quad S_t = S_0 + \int_0^t \widehat{\lambda}_u^F d\langle M \rangle_u + M_t,$$

where  $\widehat{\lambda}^F$  is the  $F$ -predictable projection of  $\lambda$  and

$$M_t = N_t + \int_0^t [\lambda_u - \widehat{\lambda}_u^F] d\langle N \rangle_u$$

is a continuous  $F$ -local martingale. Besides  $\langle M \rangle = \langle N \rangle$ , and these brackets are  $F^S$ -predictable.

Throughout the paper we shall make the following assumptions:

(A)  $\langle M \rangle$  is  $G$ -predictable and  $d\langle M \rangle_t dP$  a.e.  $\hat{\lambda}^F = \hat{\lambda}^G$ ; hence  $P$ -a.s. for each  $t$

$$E(\lambda_t | F_{t-}^S \vee G_t) = E(\lambda_t | G_t);$$

(B) any  $G$ -martingale is an  $F$ -local martingale;

(C) the filtration  $G$  is continuous; i.e., all  $G$ -local martingales are continuous;

(D) there exists a martingale measure for  $S$  (on  $F_T$ ) that satisfies the reverse Hölder condition.

*Remark.* It is evident that if  $F^S \subseteq G$ , then  $\langle M \rangle$  is  $G$ -predictable. Besides, in this case  $G = F$ , and conditions (A) and (B) are satisfied.

We shall use the notation  $\hat{Y}_t$  for the process of the  $G$ -projection of  $Y$ . Condition (A) implies that

$$\hat{S}_t = E(S_t | G_t) = S_0 + \int_0^t \hat{\lambda}_u d\langle M \rangle_u + \widehat{M}_t.$$

Let

$$H_t = E(H | F_t) = EH + \int_0^t h_u dM_u + L_t$$

and

$$H_t = EH + \int_0^t h_u^G d\widehat{M}_u + L_t^G$$

be the Galtchouk–Kunita–Watanabe (GKW) decompositions of  $H_t = E(H | F_t)$  with respect to local martingales  $M$  and  $\widehat{M}$ , where  $h$  and  $h^G$  are  $F$ -predictable processes and  $L$  and  $L^G$  are local martingales strongly orthogonal to  $M$  and  $\widehat{M}$ , respectively.

We show (Theorem 3.1) that the initial mean-variance hedging problem (1.2) is equivalent to the problem to minimize the expression

$$(1.5) \quad E \left[ \left( x + \int_0^T \pi_u d\hat{S}_u - \hat{H}_T \right)^2 + \int_0^T \left( \pi_u^2 (1 - \rho_u^2) + 2\pi_u \tilde{h}_u \right) d\langle M \rangle_u \right]$$

over all  $\pi \in \Pi(G)$ , where

$$\tilde{h}_t = \widehat{h}_t^G \rho_t^2 - \hat{h}_t \quad \text{and} \quad \rho_t^2 = \frac{d\langle \widehat{M} \rangle_t}{d\langle M \rangle_t}.$$

Thus, the problem (1.5), equivalent to (1.2), is formulated in terms of  $G$ -adapted processes. One can say that (1.5) is the mean-variance hedging problem under complete information with an additional correction term.

Let us introduce the value process of the problem (1.5):

$$(1.6) \quad V^H(t, x) = \operatorname{ess\,inf}_{\pi \in \Pi(G)} E \left[ \left( x + \int_t^T \pi_u d\hat{S}_u - \hat{H}_T \right)^2 + \int_t^T \left[ \pi_u^2 (1 - \rho_u^2) + 2\pi_u \tilde{h}_u \right] d\langle M \rangle_u | G_t \right].$$

We show in Theorem 4.1 that the value function of the problem (1.5) admits a representation

$$V^H(t, x) = V_t(0) - 2V_t(1)x + V_t(2)x^2,$$

where the coefficients  $V_t(0)$ ,  $V_t(1)$ , and  $V_t(2)$  satisfy a triangle system of backward stochastic differential equations (BSDEs). Besides, the filtered wealth process of the optimal hedging strategy is characterized as a solution of the linear forward equation

$$(1.7) \quad \begin{aligned} \widehat{X}_t^* = x &- \int_0^t \frac{\rho_u^2 \varphi_u(2) + \widehat{\lambda}_u V_u(2)}{1 - \rho_u^2 + \rho_u^2 V_u(2)} \widehat{X}_u^* d\widehat{S}_u \\ &+ \int_0^t \frac{\rho_u^2 \varphi_u(1) + \widehat{\lambda}_u V_u(1) + \widetilde{h}_u}{1 - \rho_u^2 + \rho_u^2 V_u(2)} d\widehat{S}_u. \end{aligned}$$

Note that if  $F^S \subseteq G$ , then

$$(1.8) \quad \rho = 1, \quad \widetilde{h} = 0, \quad \widehat{M} = M, \quad \text{and} \quad \widehat{S} = S.$$

In the case of complete information ( $G = \mathcal{A}$ ), in addition to (1.8) we have  $\widehat{\lambda} = \lambda$  and  $\widehat{M} = N$ , and (1.7) gives equations for the optimal wealth process from [19].

In section 5 we consider a diffusion market model, which consists of two assets  $S$  and  $\eta$ , where  $S_t$  is a state of a process being controlled and  $\eta_t$  is the observation process. Suppose that  $S_t$  and  $\eta_t$  are governed by

$$\begin{aligned} dS_t &= \mu_t dt + \sigma_t dw_t^0, \\ d\eta_t &= a_t dt + b_t dw_t, \end{aligned}$$

where  $w^0$  and  $w$  are Brownian motions with correlation  $\rho$  and the coefficients  $\mu, \sigma, a$ , and  $b$  are  $\mathcal{F}^\eta$ -adapted. In this case  $\mathcal{A}_t = \mathcal{F}_t = \mathcal{F}_t^{S, \eta}$ , and the flow of observable events is  $\mathcal{G}_t = \mathcal{F}_t^\eta$ . As an application of Theorem 4.1 we also consider a diffusion market model with constant coefficients and assume that an investor observes the price process  $S$  only up to a random moment  $\tau$  before the expiration date  $T$ . In this case we give an explicit solution of (1.2).

**2. Main definitions and auxiliary facts.** Denote by  $\mathcal{M}^e(F)$  the set of equivalent martingale measures for  $S$ , i.e., the set of probability measures  $Q$  equivalent to  $P$  such that  $S$  is a  $F$ -local martingale under  $Q$ .

Let

$$\mathcal{M}_2^e(F) = \{Q \in \mathcal{M}^e(F) : EZ_T^2(Q) < \infty\},$$

where  $Z_t(Q)$  is the density process (with respect to the filtration  $F$ ) of  $Q$  relative to  $P$ . We assume that  $\mathcal{M}_2^e(F) \neq \emptyset$ .

*Remark 2.1.* Note that  $\mathcal{M}_2^e(\mathcal{A}) \neq \emptyset$  implies that  $\mathcal{M}_2^e(F) \neq \emptyset$  (see Remark 2.1 from Pham [22]).

It follows from (1.4) and condition (A), that the density process  $Z_t(Q)$  of any element  $Q$  of  $\mathcal{M}^e(F)$  is expressed as an exponential martingale of the form

$$\mathcal{E}_t(-\widehat{\lambda} \cdot M + L),$$

where  $L$  is a  $F$ -local martingale strongly orthogonal to  $M$  and  $\mathcal{E}_t(X)$  is the Doleans-Dade exponential of  $X$ .

If the local martingale  $Z_t^{min} = \mathcal{E}_t(-\hat{\lambda} \cdot M)$  is a true martingale,  $dQ^{min}/dP = Z_T^{min}$  defines the minimal martingale measure for  $S$ .

Recall that a measure  $Q$  satisfies the reverse Hölder inequality  $R_2(P)$  if there exists a constant  $C$  such that

$$E\left(\frac{Z_T^2(Q)}{Z_\tau^2(Q)}|F_\tau\right) \leq C, \quad P\text{-a.s.}$$

for every  $F$ -stopping time  $\tau$ .

*Remark 2.2.* If there exists a measure  $Q \in \mathcal{M}^e(F)$  that satisfies the reverse Hölder inequality  $R_2(P)$ , then according to Theorem 3.4 of Kazamaki [14] the martingale  $M^Q = -\hat{\lambda} \cdot M + L$  belongs to the class  $BMO$  and hence  $-\hat{\lambda} \cdot M$  also belongs to  $BMO$ , i.e.,

$$(2.1) \quad E\left(\int_\tau^T \hat{\lambda}_u^2 d\langle M \rangle_u | F_\tau\right) \leq \text{const}$$

for every stopping time  $\tau$ . Therefore, it follows from Theorem 2.3 of [14] that  $\mathcal{E}_t(-\hat{\lambda} \cdot M)$  is a true martingale. So, condition (D) implies that the minimal martingale measure exists (but  $Z^{min}$  is not necessarily square integrable).

Let us make some remarks on conditions (B) and (C).

*Remark 2.3.* Condition (B) is satisfied if and only if the  $\sigma$ -algebras  $F_t^S \vee G_t$  and  $G_T$  are conditionally independent given  $G_t$  for all  $t \in [0, T]$  (see Theorem 9.29 from Jacod [13]).

*Remark 2.4.* Condition (C) is weaker than the assumption that the filtration  $F$  is continuous. The continuity of the filtration  $F$  and condition (B) imply the continuity of the filtration  $G$ , but the converse is not true in general. Note that filtrations  $F$  and  $F^S$  can be discontinuous. Recall that the continuity of a filtration means that all local martingales with respect to this filtration are continuous.

By  $\mu^K$  we denote the Dolean measure of an increasing process  $K$ . For all unexplained notations concerning the martingale theory used below, we refer the reader to [5], [18], [13].

Let  $\Pi(F)$  be the space of all  $F$ -predictable  $S$ -integrable processes  $\pi$  such that the stochastic integral

$$(\pi \cdot S)_t = \int_0^t \pi_u dS_u, \quad t \in [0, T],$$

is in the  $\mathcal{S}^2$  space of semimartingales, i.e.,

$$E\left(\int_0^T \pi_s^2 d\langle M \rangle_s\right) + E\left(\int_0^T |\pi_s \hat{\lambda}_s| d\langle M \rangle_s\right)^2 < \infty.$$

Denote by  $\Pi(G)$  the subspace of  $\Pi(F)$  of  $G$ -predictable strategies.

*Remark 2.5.* Since  $\hat{\lambda} \cdot M \in BMO$  (see Remark 2.2), it follows from the proof of Theorem 2.5 of Kazamaki [14] that

$$\begin{aligned} E\left(\int_0^T |\pi_u \hat{\lambda}_u| d\langle M \rangle_u\right)^2 &= E\langle |\pi| \cdot M, |\hat{\lambda}| \cdot M \rangle_T^2 \\ &\leq 2\|\hat{\lambda} \cdot M\|_{BMO} E\int_0^T \pi^2 d\langle M \rangle_u < \infty. \end{aligned}$$

Therefore, under condition (D) the  $G$ -predictable (resp.,  $F$ -predictable) strategy  $\pi$  belongs to the class  $\Pi(G)$  (resp.,  $\Pi(F)$ ) if and only if  $E \int_0^T \pi_s^2 d\langle M \rangle_s < \infty$ .

Define  $J_T^2(F)$  and  $J_T^2(G)$  as spaces of terminal values of stochastic integrals, i.e.,

$$J_T^2(F) = \{(\pi \cdot S)_T : \pi \in \Pi(F)\}.$$

$$J_T^2(G) = \{(\pi \cdot S)_T : \pi \in \Pi(G)\}.$$

For convenience we give some assertions from [4], which establishes necessary and sufficient conditions for the closedness of the space  $J_T^2(F)$  in  $L^2$ .

PROPOSITION 2.1. *Let  $S$  be a continuous semimartingale. Then the following assertions are equivalent:*

- (1) *There is a martingale measure  $Q \in \mathcal{M}^e(F)$ , and  $J_T^2(F)$  is closed in  $L^2$ .*
- (2) *There is a martingale measure  $Q \in \mathcal{M}^e(F)$  that satisfies the reverse Hölder condition  $R_2(P)$ .*
- (3) *There is a constant  $C$  such that for all  $\pi \in \Pi(F)$  we have*

$$\|\sup_{t \leq T} (\pi \cdot S)_t\|_{L^2(P)} \leq C \|(\pi \cdot S)_T\|_{L^2(P)}.$$

- (4) *There is a constant  $c$  such that for every stopping time  $\tau$ , every  $A \in \mathcal{F}_\tau$ , and every  $\pi \in \Pi(F)$ , with  $\pi = \pi I_{[\tau, T]}$ , we have*

$$\|I_A - (\pi \cdot S)_T\|_{L^2(P)} \geq cP(A)^{1/2}.$$

Note that assertion (4) implies that for every stopping time  $\tau$  and for every  $\pi \in \Pi(G)$  we have

$$(2.2) \quad E \left( \left( 1 + \int_\tau^T \pi_u dS_u \right)^2 / F_\tau \right) \geq c.$$

Now we recall some known assertions from the filtering theory. The following proposition can be proved similarly to [18].

PROPOSITION 2.2. *If conditions (A), (B), and (C) are satisfied, then for any continuous  $F$ -local martingale  $M$ , with  $M_0 = 0$ , and any  $G$ -local martingale  $m^G$*

$$(2.3) \quad \widehat{M}_t = E(M_t | G_t) = \int_0^t \frac{d\langle \widehat{M}, m^G \rangle_u}{d\langle m^G \rangle_u} dm_u^G + L_t^G,$$

where  $L^G$  is a local martingale orthogonal to  $m^G$ .

It follows from this proposition that for any  $G$ -predictable,  $M$ -integrable process  $\pi$  and any  $G$ -martingale  $m^G$

$$\begin{aligned} \langle \widehat{(\pi \cdot M)}, m^G \rangle_t &= \int_0^t \pi_u \frac{d\langle \widehat{M}, m^G \rangle_u}{d\langle m^G \rangle_u} d\langle m^G \rangle_u \\ &= \int_0^t \pi_u d\langle \widehat{M}, m^G \rangle_u = \langle \pi \cdot \widehat{M}, m^G \rangle_t. \end{aligned}$$

Hence, for any  $G$ -predictable,  $M$ -integrable process  $\pi$

$$(2.4) \quad (\widehat{(\pi \cdot M)})_t = E \left( \int_0^t \pi_s dM_s | G_t \right) = \int_0^t \pi_s d\widehat{M}_s.$$

Since  $\pi, \lambda$ , and  $\langle M \rangle$  are  $G$ -predictable, from (2.4) we have

$$(2.5) \quad \widehat{(\pi \cdot S)}_t = E \left( \int_0^t \pi_u dS_u | G_t \right) = \int_0^t \pi_u d\widehat{S}_u,$$

where

$$\widehat{S}_t = S_0 + \int_0^t \widehat{\lambda}_u d\langle M \rangle_u + \widehat{M}_t.$$

**3. Separation principle: The optimality principle.** Let us introduce the value function of the problem (1.2) defined as

$$(3.1) \quad U^H(t, x) = \operatorname{ess\,inf}_{\pi \in \Pi(G)} E \left( \left( x + \int_t^T \pi_u dS_u - H \right)^2 | G_t \right).$$

By the GKW decomposition

$$(3.2) \quad H_t = E(H | F_t) = EH + \int_0^t h_u dM_u + L_t$$

for a  $F$ -predictable,  $M$ -integrable process  $h$  and a local martingale  $L$  strongly orthogonal to  $M$ . We shall use also the GKW decompositions of  $H_t = E(H | F_t)$  with respect to the local martingale  $\widehat{M}$

$$(3.3) \quad H_t = EH + \int_0^t h_u^G d\widehat{M}_u + L_t^G,$$

where  $h^G$  is a  $F$ -predictable process and  $L^G$  is a  $F$ -local martingale strongly orthogonal to  $\widehat{M}$ .

It follows from Proposition 2.2 (applied for  $m^G = \widehat{M}$ ) and Lemma A.1 that

$$(3.4) \quad \langle E(H | G.), \widehat{M} \rangle_t = \int_0^t \widehat{h}_u^G \rho_u^2 d\langle M \rangle_u.$$

We shall use the notation

$$(3.5) \quad \widetilde{h}_t = \widehat{h}_t^G \rho_t^2 - \widehat{h}_t.$$

Note that  $\widetilde{h}$  belongs to the class  $\Pi(G)$  by Lemma A.2.

Let us introduce now a new optimization problem, equivalent to the initial mean-variance hedging problem (1.2), to minimize the expression

$$(3.6) \quad E \left[ \left( x + \int_0^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 + \int_0^T \left( \pi_u^2 (1 - \rho_u^2) + 2\pi_u \widetilde{h}_u \right) d\langle M \rangle_u \right]$$

over all  $\pi \in \Pi(G)$ . Recall that  $\widehat{S}_t = E(S_t | G_t) = S_0 + \int_0^t \widehat{\lambda}_u d\langle M \rangle_u + \widehat{M}_t$ .

**THEOREM 3.1.** *Let conditions (A), (B), and (C) be satisfied. Then the initial mean-variance hedging problem (1.2) is equivalent to the problem (3.6). In particular,*

for any  $\pi \in \Pi(G)$  and  $t \in [0, T]$

$$(3.7) \quad E \left[ \left( x + \int_t^T \pi_u dS_u - H \right)^2 | G_t \right] = E \left[ \left( H - \widehat{H}_T \right)^2 | G_t \right] \\ + E \left[ \left( x + \int_t^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 + \int_t^T \left( \pi_u^2 (1 - \rho_u^2) + 2\pi_u \tilde{h}_u \right) d\langle M \rangle_u | G_t \right].$$

*Proof.* We have

$$(3.8) \quad E \left[ \left( x + \int_t^T \pi_u dS_u - H \right)^2 | G_t \right] \\ = E \left[ \left( x + \int_t^T \pi_u d\widehat{S}_u - H + \int_t^T \pi_u d(M_u - \widehat{M}_u) \right)^2 | G_t \right] \\ = E \left[ \left( x + \int_t^T \pi_u d\widehat{S}_u - H \right)^2 | G_t \right] \\ + 2E \left[ \left( x + \int_t^T \pi_u d\widehat{S}_u - H \right) \left( \int_t^T \pi_u d(M_u - \widehat{M}_u) \right) | G_t \right] \\ + E \left[ \left( \int_t^T \pi_u d(M_u - \widehat{M}_u) \right)^2 | G_t \right] = I_1 + 2I_2 + I_3.$$

It is evident that

$$(3.9) \quad I_1 = E \left[ \left( x + \int_t^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 | G_t \right] + E \left[ \left( H - \widehat{H}_T \right)^2 | G_t \right].$$

Since  $\pi, \widehat{\lambda}$ , and  $\widehat{\langle M \rangle}$  are  $G_T$ -measurable and the  $\sigma$ -algebras  $F_t^S \vee G_t$  and  $G_T$  are conditionally independent given  $G_t$  (see Remark 2.3), it follows from (2.4) that

$$(3.10) \quad E \left[ \int_t^T \pi_u \widehat{\lambda}_u d\langle M \rangle_u \int_t^T \pi_u d(M_u - \widehat{M}_u) | G_t \right] \\ = E \left[ \int_t^T \pi_u \widehat{\lambda}_u d\langle M \rangle_u \int_0^T \pi_u d(M_u - \widehat{M}_u) | G_t \right] \\ - E \left[ \int_t^T \pi_u \widehat{\lambda}_u d\langle M \rangle_u \int_0^t \pi_u d(M_u - \widehat{M}_u) | G_t \right] \\ = E \left[ \int_t^T \pi_u \widehat{\lambda}_u d\langle M \rangle_u E \left( \int_0^T \pi_u d(M_u - \widehat{M}_u) | G_T \right) | G_t \right] \\ - E \left[ \int_t^T \pi_u \widehat{\lambda}_u d\langle M \rangle_u | G_t \right] E \left[ \int_0^t \pi_u d(M_u - \widehat{M}_u) | G_t \right] \\ = 0.$$



On the other hand, by using decomposition (3.2), equality (3.4), properties of square characteristics of martingales, and the projection theorem, we obtain

$$\begin{aligned}
 & E \left[ H \int_t^T \pi_u d(M_u - \widehat{M}_u) | G_t \right] \\
 &= E \left[ H \int_t^T \pi_u dM_u | G_t \right] - E \left[ \widehat{H}_T \int_t^T \pi_u d\widehat{M}_u | G_t \right] \\
 &= E \left[ \int_t^T \pi_u d\langle M, E(H|F) \rangle_u | G_t \right] - E \left[ \int_t^T \pi_u d\langle \widehat{H}, \widehat{M} \rangle_u | G_t \right] \\
 &= E \left[ \int_t^T \pi_u h_u d\langle M \rangle_u | G_t \right] - E \left[ \int_t^T \pi_u \widehat{h}_u^G \rho_u^2 d\langle M \rangle_u | G_t \right] \\
 (3.11) \quad &= E \left[ \int_t^T \pi_u (\widehat{h}_u - \widehat{h}_u^G \rho_u^2) d\langle M \rangle_u | G_t \right] = -E \left[ \int_t^T \pi_u \widetilde{h}_u d\langle M \rangle_u | G_t \right].
 \end{aligned}$$

Finally, it is easy to verify that

$$\begin{aligned}
 (3.12) \quad & 2E \left[ \int_t^T \pi_u \widehat{M}_u \int_t^T \pi_u d(M_u - \widehat{M}_u) | G_t \right] + E \left[ \left( \int_t^T \pi_u d(M_u - \widehat{M}_u) \right)^2 | G_t \right] \\
 &= E \left[ \left( \int_t^T \pi_u^2 d\langle M \rangle_u - \int_t^T \pi_u^2 d\langle \widehat{M} \rangle_u \right) | G_t \right] = E \left[ \int_t^T \pi_u^2 (1 - \rho_u^2) d\langle M \rangle_u | G_t \right].
 \end{aligned}$$

Therefore (3.8), (3.9), (3.10), (3.11), and (3.12) imply the validity of equality (3.7).  $\square$

Thus, it follows from Theorem 3.1 that the optimization problems (1.2) and (3.6) are equivalent. Therefore it is sufficient to solve the problem (3.6), which is formulated in terms of  $G$ -adapted processes. One can say that (3.6) is a mean-variance hedging problem under complete information with a correction term and can be solved by using methods for complete information.

Let us introduce the value process of the problem (3.6)

$$\begin{aligned}
 V^H(t, x) &= \operatorname{ess\,inf}_{\pi \in \Pi(G)} E \left[ \left( x + \int_t^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 \right. \\
 (3.13) \quad & \left. + \int_t^T \left[ \pi_u^2 (1 - \rho_u^2) + 2\pi_u \widetilde{h}_u \right] d\langle M \rangle_u | G_t \right].
 \end{aligned}$$

It follows from Theorem 3.1 that

$$(3.14) \quad U^H(t, x) = V^H(t, x) + E[(H - \widehat{H}_T)^2 | G_t].$$

The optimality principle takes in this case the following form.

**PROPOSITION 3.1** (optimality principle). *Let conditions (A), (B) and (C) be satisfied. Then*

(a) *for all  $x \in R$ ,  $\pi \in \Pi(G)$ , and  $s \in [0, T]$  the process*

$$V^H \left( t, x + \int_s^t \pi_u d\widehat{S}_u \right) + \int_s^t \left[ \pi_u^2 (1 - \rho_u^2) + 2\pi_u \widetilde{h}_u \right] d\langle M \rangle_u$$

is a submartingale on  $[s, T]$ , admitting an right continuous with left limits (RCLL) modification.

(b)  $\pi^*$  is optimal if and only if the process

$$V^H \left( t, x + \int_s^t \pi_u^* d\widehat{S}_u \right) + \int_s^t \left[ (\pi_u^*)^2 (1 - \rho_u^2) + 2\pi_u^* \tilde{h}_u \right] d\langle M \rangle_u$$

is a martingale.

This assertion can be proved in a standard manner (see, e.g., [7], [15]). The proof more adapted to this case one can see in [19].

Let

$$V(t, x) = \operatorname{ess\,inf}_{\pi \in \Pi(G)} E \left[ \left( x + \int_t^T \pi_u d\widehat{S}_u \right)^2 + \int_t^T \pi_u^2 (1 - \rho_u^2) d\langle M \rangle_u | G_t \right]$$

and

$$V_t(2) = \operatorname{ess\,inf}_{\pi \in \Pi(G)} E \left[ \left( 1 + \int_t^T \pi_u d\widehat{S}_u \right)^2 + \int_t^T \pi_u^2 (1 - \rho_u^2) d\langle M \rangle_u | G_t \right].$$

It is evident that  $V(t, x)$  (resp.,  $V_t(2)$ ) is the value process of the optimization problem (3.6) in the case  $H = 0$  (resp.,  $H = 0$  and  $x = 1$ ), i.e.,

$$V(t, x) = V^0(t, x) \quad \text{and} \quad V_t(2) = V^0(t, 1).$$

Since  $\Pi(G)$  is a cone, we have

$$\begin{aligned} V(t, x) &= x^2 \operatorname{ess\,inf}_{\pi \in \Pi(G)} E \left[ \left( 1 + \int_t^T \frac{\pi_u}{x} d\widehat{S}_u \right)^2 \right. \\ (3.15) \quad &\quad \left. + \int_t^T \left( \frac{\pi_u}{x} \right)^2 (1 - \rho_u^2) d\langle M \rangle_u | G_t \right] = x^2 V_t(2). \end{aligned}$$

Therefore from Proposition 3.1 and equality (3.15) we have the following.

COROLLARY 3.1. (a) *The process*

$$V_t(2) \left( 1 + \int_s^t \pi_u d\widehat{S}_u \right)^2 + \int_s^t (\pi_u)^2 (1 - \rho_u^2) d\langle M \rangle_u,$$

$t \geq s$ , is a submartingale for all  $\pi \in \Pi(G)$  and  $s \in [0, T]$ .

(b)  $\pi^*$  is optimal if and only if

$$V_t(2) \left( 1 + \int_s^t \pi_u^* d\widehat{S}_u \right)^2 + \int_s^t (\pi_u^*)^2 (1 - \rho_u^2) d\langle M \rangle_u,$$

$t \geq s$ , is a martingale.

Note that in the case  $H = 0$  from Theorem 3.1 we have

$$\begin{aligned} (3.16) \quad & E \left[ \left( 1 + \int_t^T \pi_u dS_u \right)^2 \middle| G_t \right] \\ &= E \left[ \left( 1 + \int_t^T \pi_u d\widehat{S}_u \right)^2 + \int_t^T \pi_u^2 (1 - \rho_u^2) d\langle M \rangle_u \middle| G_t \right] \end{aligned}$$

and, hence,

$$(3.17) \quad V_t(2) = U^0(t, 1).$$

LEMMA 3.1. *Let conditions (A)–(D) be satisfied. Then there is a constant  $1 \geq c > 0$  such that  $V_t(2) \geq c$  for all  $t \in [0, T]$  a.s. and*

$$(3.18) \quad 1 - \rho_t^2 + \rho_t^2 V_t(2) \geq c \quad \mu^{(M)} \text{ a.e.}$$

*Proof.* Let

$$V_t^F(2) = \operatorname{ess\,inf}_{\pi \in \Pi(F)} E \left[ \left( 1 + \int_t^T \pi_u dS_u \right)^2 \middle| F_t \right].$$

It follows from assertion (4) of Proposition 2.1 that there is a constant  $c > 0$  such that  $V_t^F(2) \geq c$  for all  $t \in [0, T]$  a.s. Note that  $c \leq 1$  since  $V^F \leq 1$ . Then by (3.17)

$$\begin{aligned} V_t(2) = U^0(t, 1) &= \operatorname{ess\,inf}_{\pi \in \Pi(G)} E \left[ \left( 1 + \int_t^T \pi_u dS_u \right)^2 \middle| G_t \right] \\ &= \operatorname{ess\,inf}_{\pi \in \Pi(G)} E \left[ E \left( \left( 1 + \int_t^T \pi_u dS_u \right)^2 \middle| F_t \right) \middle| G_t \right] \\ &\geq E(V_t^F(2) | G_t) \geq c. \end{aligned}$$

Therefore, since  $\rho_t^2 \leq 1$  by Lemma A.1,

$$1 - \rho_t^2 + \rho_t^2 V_t(2) \geq 1 - \rho_t^2 + \rho_t^2 c \geq \inf_{r \in [0, 1]} (1 - r + rc) = c. \quad \square$$

**4. BSDEs for the value process.** Let us consider the semimartingale backward equation

$$(4.1) \quad Y_t = Y_0 + \int_0^t f(u, Y_u, \psi_u) d\langle m \rangle_u + \int_0^t \psi_u dm_u + L_t$$

with the boundary condition

$$(4.2) \quad Y_T = \eta,$$

where  $\eta$  is an integrable  $G_T$ -measurable random variable,  $f : \Omega \times [0, T] \times R^2 \rightarrow R$  is  $\mathcal{P} \times \mathcal{B}(R^2)$  measurable, and  $m$  is a local martingale. A solution of (4.1)–(4.2) is a triple  $(Y, \psi, L)$ , where  $Y$  is a special semimartingale,  $\psi$  is a predictable  $m$ -integrable process, and  $L$  a local martingale strongly orthogonal to  $m$ . Sometimes we call  $Y$  alone the solution of (4.1)–(4.2), keeping in mind that  $\psi \cdot m + L$  is the martingale part of  $Y$ .

Backward stochastic differential equations have been introduced in [1] for the linear case as the equations for the adjoint process in the stochastic maximum principle. The semimartingale backward equation, as a stochastic version of the Bellman equation in an optimal control problem, was first derived in [2]. The BSDE with more general nonlinear generators was introduced in [21] for the case of Brownian filtration, where the existence and uniqueness of a solution of BSDEs with generators satisfying

the global Lipschitz condition was established. These results were generalized for generators with quadratic growth in [16], [17] for BSDEs driven by a Brownian motion and in [20], [28] for BSDEs driven by martingales. But conditions imposed in these papers are too restrictive for our needs. We prove here the existence and uniqueness of a solution by directly showing that the unique solution of the BSDE that we consider is the value of the problem.

In this section we characterize optimal strategies in terms of solutions of suitable semimartingale backward equations.

**THEOREM 4.1.** *Let  $H$  be a square integrable  $F_T$ -measurable random variable, and let conditions (A), (B), (C), and (D) be satisfied. Then the value function of the problem (3.6) admits a representation*

$$(4.3) \quad V^H(t, x) = V_t(0) - 2V_t(1)x + V_t(2)x^2,$$

where the processes  $V_t(0)$ ,  $V_t(1)$ , and  $V_t(2)$  satisfy the following system of backward equations:

$$(4.4) \quad \begin{aligned} Y_t(2) &= Y_0(2) + \int_0^t \frac{(\psi_s(2)\rho_s^2 + \widehat{\lambda}_s Y_s(2))^2}{1 - \rho_s^2 + \rho_s^2 Y_s(2)} d\langle M \rangle_s \\ &\quad + \int_0^t \psi_s(2) d\widehat{M}_s + L_t(2), \quad Y_T(2) = 1, \end{aligned}$$

$$(4.5) \quad \begin{aligned} Y_t(1) &= Y_0(1) + \int_0^t \frac{(\psi_s(2)\rho_s^2 + \widehat{\lambda}_s Y_s(2))(\psi_s(1)\rho_s^2 + \widehat{\lambda}_s Y_s(1) - \tilde{h}_s)}{1 - \rho_s^2 + \rho_s^2 Y_s(2)} d\langle M \rangle_s \\ &\quad + \int_0^t \psi_s(1) d\widehat{M}_s + L_t(1), \quad Y_T(1) = E(H|G_T), \end{aligned}$$

$$(4.6) \quad \begin{aligned} Y_t(0) &= Y_0(0) + \int_0^t \frac{(\psi_s(1)\rho_s^2 + \widehat{\lambda}_s Y_s(1) - \tilde{h}_s)^2}{1 - \rho_s^2 + \rho_s^2 Y_s(2)} d\langle M \rangle_s \\ &\quad + \int_0^t \psi_s(0) d\widehat{M}_s + L_t(0), \quad Y_T(0) = E^2(H|G_T), \end{aligned}$$

where  $L(2)$ ,  $L(1)$ , and  $L(0)$  are  $G$ -local martingales orthogonal to  $\widehat{M}$ .

Besides, the optimal filtered wealth process  $\widehat{X}_t^{x, \pi^*} = x + \int_0^t \pi_u^* d\widehat{S}_u$  is a solution of the linear equation

$$(4.7) \quad \begin{aligned} \widehat{X}_t^* &= x - \int_0^t \frac{\rho_u^2 \psi_u(2) + \widehat{\lambda}_u Y_u(2)}{1 - \rho_u^2 + \rho_u^2 Y_u(2)} \widehat{X}_u^* d\widehat{S}_u \\ &\quad + \int_0^t \frac{\psi_u(1)\rho_u^2 + \widehat{\lambda}_u Y_u(1) - \tilde{h}_u}{1 - \rho_u^2 + \rho_u^2 Y_u(2)} d\widehat{S}_u. \end{aligned}$$

*Proof.* Similarly to the case of complete information one can show that the optimal strategy exists and that  $V^H(t, x)$  is a square trinomial of the form (4.3) (see, e.g., [19]). More precisely the space of stochastic integrals

$$J_{t,T}^2(G) = \left\{ \int_t^T \pi_u dS_u : \pi \in \Pi(G) \right\}$$

is closed by Proposition 2.1, since  $\langle M \rangle$  is  $G$ -predictable. Hence there exists optimal strategy  $\pi^*(t, x) \in \Pi(G)$  and  $U^H(t, x) = E[|H - x - \int_t^T \pi_u^*(t, x) dS_u|^2 | G_t]$ . Since

$\int_t^T \pi_u^*(t, x) dS_u$  coincides with the orthogonal projection of  $H - x \in L^2$  on the closed subspace of stochastic integrals, then the optimal strategy is linear with respect to  $x$ , i.e.,  $\pi_u^*(t, x) = \pi_u^0(t) + x\pi_u^1(t)$ . This implies that the value function  $U^H(t, x)$  is a square trinomial. It follows from the equality (3.14) that  $V^H(t, x)$  is also a square trinomial, and it admits the representation (4.3).

Let us show that  $V_t(0)$ ,  $V_t(1)$ , and  $V_t(2)$  satisfy the system (4.4)–(4.6). It is evident that

$$(4.8) \quad \begin{aligned} V_t(0) = V^H(t, 0) = \operatorname{ess\,inf}_{\pi \in \Pi(G)} E \left[ \left( \int_t^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 \right. \\ \left. + \int_t^T [\pi_u^2 (1 - \rho_u^2) + 2\pi_u \tilde{h}_u] d\langle M \rangle_u | G_t \right] \end{aligned}$$

and

$$(4.9) \quad \begin{aligned} V_t(2) = V^0(t, 1) = \operatorname{ess\,inf}_{\pi \in \Pi(G)} E \left[ \left( 1 + \int_t^T \pi_u d\widehat{S}_u \right)^2 \right. \\ \left. + \int_t^T \pi_u^2 (1 - \rho_u^2) d\langle M \rangle_u | G_t \right]. \end{aligned}$$

Therefore, it follows from the optimality principle (taking  $\pi = 0$ ) that  $V_t(0)$  and  $V_t(2)$  are RCLL  $G$ -submartingales and

$$\begin{aligned} V_t(2) &\leq E(V_T(2)|G_t) \leq 1, \\ V_t(0) &\leq E(E^2(H|G_T)|G_t) \leq E(H^2|G_t). \end{aligned}$$

Since

$$(4.10) \quad V_t(1) = \frac{1}{2}(V_t(0) + V_t(2) - V^H(t, 1)),$$

the process  $V_t(1)$  is also a special semimartingale, and since  $V_t(0) - 2V_t(1)x + V_t(2)x^2 = V^H(t, x) \geq 0$  for all  $x \in R$ , we have  $V_t^2(1) \leq V_t(0)V_t(2)$ ; hence

$$V_t^2(1) \leq E(H^2|G_t).$$

Expressions (4.8), (4.9), and (3.13) imply that  $V_T(0) = E^2(H|G_T)$ ,  $V_T(2) = 1$ , and  $V^H(T, x) = (x - E(H|G_T))^2$ . Therefore from (4.10) we have  $V_T(1) = E(H|G_T)$ , and  $V(0)$ ,  $V(1)$ , and  $V(2)$  satisfy the boundary conditions.

Thus, the coefficients  $V_t(i)$ ,  $i = 0, 1, 2$ , are special semimartingales, and they admit the decomposition

$$(4.11) \quad V_t(i) = V_0(i) + A_t(i) + \int_0^t \varphi_s(i) d\widehat{M}_s + m_t(i), \quad i = 0, 1, 2,$$

where  $m(0)$ ,  $m(1)$ , and  $m(2)$  are  $G$ -local martingales strongly orthogonal to  $\widehat{M}$  and  $A(0)$ ,  $A(1)$ , and  $A(2)$  are  $G$ -predictable processes of finite variation.

There exists an increasing continuous  $G$ -predictable process  $K$  such that

$$\langle M \rangle_t = \int_0^t \nu_u dK_u, \quad A_t(i) = \int_0^t a_u(i) dK_u, \quad i = 0, 1, 2,$$

where  $\nu$  and  $a(i)$ ,  $i = 0, 1, 2$ , are  $G$ -predictable processes.

Let  $\widehat{X}_{s,t}^{x,\pi} \equiv x + \int_s^t \pi_u d\widehat{S}_u$  and

$$Y_{s,t}^{x,\pi} \equiv V^H\left(t, \widehat{X}_{s,t}^{x,\pi}\right) + \int_s^t \left[\pi_u^2(1 - \rho_u^2) + 2\pi_u \tilde{h}_u\right] d\langle M \rangle_u.$$

Then by using (4.3), (4.11), and the Itô formula for any  $t \geq s$  we have

$$\begin{aligned} \left(\widehat{X}_{s,t}^{x,\pi}\right)^2 &= x + \int_s^t \left[2\pi_u \widehat{\lambda}_u \widehat{X}_{s,u}^{x,\pi} + \pi_u^2 \rho_u^2\right] d\langle M \rangle_u \\ &\quad + 2 \int_s^t \pi_u \widehat{X}_{s,u}^{x,\pi} d\widehat{M}_u \end{aligned} \quad (4.12)$$

and

$$\begin{aligned} Y_{s,t}^{x,\pi} - V^H(s, x) &= \int_s^t \left[ \left(\widehat{X}_{s,u}^{x,\pi}\right)^2 a_u(2) - 2\widehat{X}_{s,u}^{x,\pi} a_u(1) + a_u(0) \right] dK_u \\ &\quad + \int_s^t \left[ \pi_u^2(1 - \rho_u^2 + \rho_u^2 V_{u-}(2)) + 2\pi_u \widehat{X}_{s,u}^{x,\pi} \left(\widehat{\lambda}_u V_{u-}(2) + \varphi_u(2)\rho_u^2\right) \right. \\ &\quad \left. - 2\pi_u \left(V_{u-}(1)\widehat{\lambda}_u + \varphi_u(1)\rho_u^2 - \tilde{h}_u\right) \right] \nu_u dK_u + m_t - m_s, \end{aligned} \quad (4.13)$$

where  $m$  is a local martingale.

Let

$$\begin{aligned} G(\pi, x) &= G(\omega, u, \pi, x) = \pi^2(1 - \rho_u^2 + \rho_u^2 V_{u-}(2)) + 2\pi x \left(\widehat{\lambda}_u V_{u-}(2) + \varphi_u(2)\rho_u^2\right) \\ &\quad - 2\pi(V_{u-}(1)\widehat{\lambda}_u + \varphi_u(1)\rho_u^2 - \tilde{h}_u). \end{aligned}$$

It follows from the optimality principle that for each  $\pi \in \Pi(G)$  the process

$$\begin{aligned} &\int_s^t \left[ \left(\widehat{X}_{s,u}^{x,\pi}\right)^2 a_u(2) - 2\widehat{X}_{s,u}^{x,\pi} a_u(1) + a_u(0) \right] dK_u \\ &\quad + \int_s^t G\left(\pi_u, \widehat{X}_{s,u}^{x,\pi}\right) \nu_u dK_u \end{aligned} \quad (4.14)$$

is increasing for any  $s$  on  $s \leq t \leq T$ , and for the optimal strategy  $\pi^*$  we have the equality

$$\begin{aligned} &\int_s^t \left[ \left(\widehat{X}_{s,u}^{x,\pi^*}\right)^2 a_u(2) - 2\widehat{X}_{s,u}^{x,\pi^*} a_u(1) + a_u(0) \right] dK_u \\ &= - \int_s^t G\left(\pi_u^*, \widehat{X}_{s,u}^{x,\pi^*}\right) \nu_u dK_u. \end{aligned} \quad (4.15)$$

Since  $\nu_u dK_u = d\langle M \rangle_u$  is continuous, without loss of generality one can assume that the process  $K$  is continuous (see [19] for details). Therefore, by taking in (4.14)  $\tau_s(\varepsilon) = \inf\{t \geq s : K_t - K_s \geq \varepsilon\}$  instead of  $t$ , we have that for any  $\varepsilon > 0$  and  $s \geq 0$

$$\begin{aligned} &\frac{1}{\varepsilon} \int_s^{\tau_s(\varepsilon)} \left[ \left(\widehat{X}_{s,u}^{x,\pi}\right)^2 a_u(2) - 2\widehat{X}_{s,u}^{x,\pi} a_u(1) + a_u(0) \right] dK_u \\ &\geq -\frac{1}{\varepsilon} \int_s^{\tau_s(\varepsilon)} G\left(\pi_u, \widehat{X}_{s,u}^{x,\pi}\right) \nu_u dK_u. \end{aligned} \quad (4.16)$$

By passing to the limit in (4.16) as  $\varepsilon \rightarrow 0$ , from Proposition B of [19] we obtain

$$x^2 a_u(2) - 2x a_u(1) + a_u(0) \geq -G(\pi_u, x) \nu_u, \quad \mu^K\text{-a.e.},$$

for all  $\pi \in \Pi(G)$ . Similarly from (4.15) we have that  $\mu^K$ -a.e.

$$x^2 a_u(2) - 2x a_u(1) + a_u(0) = -G(\pi_u^*, x) \nu_u$$

and hence

$$(4.17) \quad x^2 a_u(2) - 2x a_u(1) + a_u(0) = -\nu_u \operatorname{ess\,inf}_{\pi \in \Pi(G)} G(\pi_u, x).$$

The infimum in (4.17) is attained for the strategy

$$(4.18) \quad \hat{\pi}_t = \frac{V_t(1)\hat{\lambda}_t + \varphi_t(1)\rho_t^2 - \tilde{h}_t - x(V_t(2)\hat{\lambda}_t + \varphi_t(2)\rho_t^2)}{1 - \rho_t^2 + \rho_t^2 V_t(2)}.$$

From here we can conclude that

$$(4.19) \quad \begin{aligned} \operatorname{ess\,inf}_{\pi \in \Pi(G)} G(\pi_t, x) &\geq G(\hat{\pi}_t, x) \\ &= -\frac{\left(V_t(1)\hat{\lambda}_t + \varphi_t(1)\rho_t^2 - \tilde{h}_t - x(V_t(2)\hat{\lambda}_t + \varphi_t(2)\rho_t^2)\right)^2}{1 - \rho_t^2 + \rho_t^2 V_t(2)}. \end{aligned}$$

Let  $\pi_t^n = I_{[0, \tau_n]}(t) \hat{\pi}_t$ , where  $\tau_n = \inf\{t : |V_t(1)| \geq n\}$ .

It follows from Lemmas A.2, 3.1, and A.3 that  $\pi^n \in \Pi(G)$  for every  $n \geq 1$  and hence

$$\operatorname{ess\,inf}_{\pi \in \Pi(G)} G(\pi_t, x) \leq G(\pi_t^n, x)$$

for all  $n \geq 1$ . Therefore

$$(4.20) \quad \operatorname{ess\,inf}_{\pi \in \Pi(G)} G(\pi_t, x) \leq \lim_{n \rightarrow \infty} G(\pi_t^n, x) = G(\hat{\pi}_t, x).$$

Thus (4.17), (4.19), and (4.20) imply that

$$(4.21) \quad \begin{aligned} &x^2 a_t(2) - 2x a_t(1) + a_t(0) \\ &= \nu_t \frac{(V_t(1)\hat{\lambda}_t + \varphi_t(1)\rho_t^2 - \tilde{h}_t - x(V_t(2)\hat{\lambda}_t + \varphi_t(2)\rho_t^2))^2}{1 - \rho_t^2 + \rho_t^2 V_t(2)}, \quad \mu^K\text{-a.e.}, \end{aligned}$$

and by equalizing the coefficients of square trinomials in (4.21) (and integrating with respect to  $dK$ ) we obtain

$$(4.22) \quad A_t(2) = \int_0^t \frac{(\varphi_s(2)\rho_s^2 + \hat{\lambda}_s V_s(2))^2}{1 - \rho_s^2 + \rho_s^2 V_s(2)} d\langle M \rangle_s,$$

$$(4.23) \quad A_t(1) = \int_0^t \frac{(\varphi_s(2)\rho_s^2 + \hat{\lambda}_s V_s(2))(\varphi_s(1)\rho_s^2 + \hat{\lambda}_s V_s(1) - \tilde{h}_s)}{1 - \rho_s^2 + \rho_s^2 V_s(2)} d\langle M \rangle_s,$$

$$(4.24) \quad A_t(0) = \int_0^t \frac{(\varphi_s(1)\rho_s^2 + \hat{\lambda}_s V_s(1) - \tilde{h}_s)^2}{1 - \rho_s^2 + \rho_s^2 V_s(2)} d\langle M \rangle_s,$$

which, together with (4.11), implies that the triples  $(V(i), \varphi(i), m(i))$ ,  $i = 0, 1, 2$ , satisfy the system (4.4)–(4.6).

Note that  $A(0)$  and  $A(2)$  are integrable increasing processes and relations (4.22) and (4.24) imply that the strategy  $\hat{\pi}$  defined by (4.18) belongs to the class  $\Pi(G)$ .

Let us show now that if the strategy  $\pi^* \in \Pi(G)$  is optimal, then the corresponding filtered wealth process  $\hat{X}_t^{\pi^*} = x + \int_0^t \pi_u^* d\hat{S}_u$  is a solution of (4.7).

By the optimality principle the process

$$Y_t^{\pi^*} = V^H\left(t, \hat{X}_t^{\pi^*}\right) + \int_0^t \left[ (\pi_u^*)^2 (1 - \rho_u^2) + 2\pi_u^* \tilde{h}_u \right] d\langle M \rangle_u$$

is a martingale. By using the Itô formula we have

$$\begin{aligned} Y_t^{\pi^*} &= \int_0^t \left( \hat{X}_u^{\pi^*} \right)^2 dA_u(2) - 2 \int_0^t \hat{X}_u^{\pi^*} dA_u(1) + A_t(0) \\ &\quad + \int_0^t G\left(\pi_u^*, \hat{X}_u^{\pi^*}\right) d\langle M \rangle_u + N_t, \end{aligned}$$

where  $N$  is a martingale. Therefore by applying equalities (4.22), (4.23), and (4.24) we obtain

$$\begin{aligned} Y_t^{\pi^*} &= \int_0^t \left( \pi_u^* - \frac{V_u(1)\hat{\lambda}_u + \varphi_u(1)\rho_u^2 - \tilde{h}_u}{1 - \rho_u^2 + \rho_u^2 V_u(2)} \right. \\ &\quad \left. + \hat{X}_u^{\pi^*} \frac{V_u(2)\hat{\lambda}_u + \varphi_u(2)\rho_u^2}{1 - \rho_u^2 + \rho_u^2 V_u(2)} \right)^2 (1 - \rho_u^2 + \rho_u^2 V_u(2)) d\langle M \rangle_u + N_t, \end{aligned}$$

which implies that  $\mu^{\langle M \rangle}$ -a.e.

$$\pi_u^* = \frac{V_u(1)\hat{\lambda}_u + \varphi_u(1)\rho_u^2 - \tilde{h}_u}{1 - \rho_u^2 + \rho_u^2 V_u(2)} - \hat{X}_u^{\pi^*} \frac{(V_u(2)\hat{\lambda}_u + \varphi_u(2)\rho_u^2)}{1 - \rho_u^2 + \rho_u^2 V_u(2)}.$$

By integrating both parts of this equality with respect to  $d\hat{S}$  (and adding then  $x$  to the both parts), we obtain that  $\hat{X}^{\pi^*}$  satisfies (4.7).  $\square$

The uniqueness of the system (4.4)–(4.6) we shall prove under following condition (D\*), stronger than condition (D).

Assume that

(D\*)

$$\int_0^T \frac{\hat{\lambda}_u^2}{\rho_u^2} d\langle M \rangle_u \leq C.$$

Since  $\rho^2 \leq 1$  (Lemma A.1), it follows from (D\*) that the mean-variance tradeoff of  $S$  is bounded, i.e.,

$$\int_0^T \hat{\lambda}_u^2 d\langle M \rangle_u \leq C,$$

which implies (see, e.g., Kazamaki [14]) that the minimal martingale measure for  $S$  exists and satisfies the reverse Hölder condition  $R_2(P)$ . So, condition (D\*) implies



condition (D). Besides, it follows from condition (D\*) that the minimal martingale measure  $\widehat{Q}^{min}$  for  $\widehat{S}$

$$d\widehat{Q}^{min} = \mathcal{E}_T \left( -\frac{\widehat{\lambda}}{\rho^2} \cdot \widehat{M} \right)$$

also exists and satisfies the reverse Hölder condition. Indeed, condition (D\*) implies that  $\mathcal{E}_t(-2\frac{\widehat{\lambda}}{\rho^2} \cdot \widehat{M})$  is a  $G$ -martingale and hence

$$E \left( \mathcal{E}_{tT}^2 \left( -\frac{\widehat{\lambda}}{\rho^2} \cdot \widehat{M} \right) | G_t \right) = E \left( \mathcal{E}_{tT} \left( -2\frac{\widehat{\lambda}}{\rho^2} \cdot \widehat{M} \right) e^{\int_t^T \frac{\widehat{\lambda}_u^2}{\rho_u^2} d\langle M \rangle_u} G_t \right) \leq e^C.$$

Recall that the process  $Z$  belongs to the class  $D$  if the family of random variables  $Z_\tau I_{(\tau \leq T)}$  for all stopping times  $\tau$  is uniformly integrable.

**THEOREM 4.2.** *Let conditions (A), (B), (C), and (D\*) be satisfied. If a triple  $(Y(0), Y(1), Y(2))$ , where  $Y(0) \in D$ ,  $Y^2(1) \in D$ , and  $c \leq Y(2) \leq C$  for some constants  $0 < c < C$ , is a solution of the system (4.4)–(4.6), then such a solution is unique and coincides with the triple  $(V(0), V(1), V(2))$ .*

*Proof.* Let  $Y(2)$  be a bounded strictly positive solution of (4.4), and let

$$\int_0^t \psi_u(2) d\widehat{M}_u + L_t(2)$$

be the martingale part of  $Y(2)$ .

Since  $Y(2)$  solves (4.4), it follows from the Itô formula that for any  $\pi \in \Pi(G)$  the process

$$(4.25) \quad Y_t^\pi = Y_t(2) \left( 1 + \int_s^t \pi_u d\widehat{S}_u \right)^2 + \int_s^t \pi_u^2 (1 - \rho_u^2) d\langle M \rangle_u,$$

$t \geq s$ , is a local submartingale.

Since  $\pi \in \Pi(G)$ , from Lemma A.1 and the Doob inequality we have

$$(4.26) \quad \begin{aligned} & E \sup_{t \leq T} \left( 1 + \int_0^t \pi_u d\widehat{S}_u \right)^2 \\ & \leq \text{const} \left( 1 + E \int_0^T \pi_u^2 \rho_u^2 d\langle M \rangle_u \right) + E \left( \int_0^T |\pi_u \widehat{\lambda}_u| d\langle M \rangle_u \right)^2 < \infty. \end{aligned}$$

Therefore, by taking in mind that  $Y(2)$  is bounded and  $\pi \in \Pi(G)$  we obtain

$$E \left( \sup_{s \leq u \leq T} Y_u^\pi \right)^2 < \infty,$$

which implies that  $Y^\pi \in D$ . Thus  $Y^\pi$  is a submartingale (as a local submartingale from the class  $D$ ), and by the boundary condition  $Y_T(2) = 1$  we obtain

$$Y_s(2) \leq E \left( \left( 1 + \int_s^T \pi_u d\widehat{S}_u \right)^2 + \int_s^T \pi_u^2 (1 - \rho_u^2) d\langle M \rangle_u | G_s \right)$$

for all  $\pi \in \Pi(G)$  and hence

$$(4.27) \quad Y_t(2) \leq V_t(2).$$

Let

$$\tilde{\pi}_t = -\frac{\hat{\lambda}_t Y_t(2) + \psi_t(2) \rho_t^2}{1 - \rho_t^2 + \rho_t^2 Y_t(2)} \mathcal{E}_t \left( -\frac{\hat{\lambda} Y(2) + \psi(2) \rho^2}{1 - \rho^2 + \rho^2 Y(2)} \cdot \hat{S} \right).$$

Since  $1 + \int_0^t \tilde{\pi}_u d\hat{S}_u = \mathcal{E}_t \left( -\frac{\hat{\lambda} Y(2) + \psi(2) \rho^2}{1 - \rho^2 + \rho^2 Y(2)} \cdot \hat{S} \right)$ , it follows from (4.4) and the Itô formula that the process  $Y^{\tilde{\pi}}$  defined by (4.25) is a positive local martingale and hence a supermartingale. Therefore

$$(4.28) \quad Y_s(2) \geq E \left( \left( 1 + \int_s^T \tilde{\pi}_u d\hat{S}_u \right)^2 + \int_s^T \tilde{\pi}_u^2 (1 - \rho_u^2) d\langle M \rangle_u | G_s \right).$$

Let us show that  $\tilde{\pi}$  belongs to the class  $\Pi(G)$ .

From (4.28) and (4.27) we have for every  $s \in [0, T]$

$$(4.29) \quad E \left( \left( 1 + \int_s^T \tilde{\pi}_u d\hat{S}_u \right)^2 + \int_s^T \tilde{\pi}_u^2 (1 - \rho_u^2) d\langle M \rangle_u | G_s \right) \leq Y_s(2) \leq V_s(2) \leq 1$$

and hence

$$(4.30) \quad E \left( 1 + \int_0^T \tilde{\pi}_u d\hat{S}_u \right)^2 \leq 1,$$

$$(4.31) \quad E \int_0^T \tilde{\pi}_u^2 (1 - \rho_u^2) d\langle M \rangle_u \leq 1.$$

By (D\*) the minimal martingale measure  $\hat{Q}^{min}$  for  $\hat{S}$  satisfies the reverse Hölder condition, and hence all conditions of Proposition 2.1 are satisfied. Therefore the norm

$$E \left( \int_0^T \tilde{\pi}_s^2 \rho_s^2 d\langle M \rangle_s \right) + E \left( \int_0^T |\tilde{\pi}_s \hat{\lambda}_s| d\langle M \rangle_s \right)^2$$

is estimated by  $E(1 + \int_0^T \tilde{\pi}_u d\hat{S}_u)^2$  and hence

$$E \int_0^T \tilde{\pi}_u^2 \rho_u^2 d\langle M \rangle_u < \infty, \quad E \left( \int_0^T |\tilde{\pi}_s \hat{\lambda}_s| d\langle M \rangle_s \right)^2 < \infty.$$

It follows from (4.31) and the latter inequality that  $\tilde{\pi} \in \Pi(G)$ , and from (4.28) we obtain

$$Y_t(2) \geq V_t(2),$$

which together with (4.27) gives the equality  $Y_t(2) = V_t(2)$ .

Thus  $V(2)$  is a unique bounded strictly positive solution of (4.4). Besides,

$$(4.32) \quad \int_0^t \psi_u(2) d\widehat{M}_u = \int_0^t \varphi_u(2) d\widehat{M}_u, \quad L_t(2) = m_t(2)$$

for all  $t$ ,  $P$ -a.s.

Let  $Y(1)$  be a solution of (4.5) such that  $Y^2(1) \in D$ . By the Itô formula the process

$$(4.33) \quad \begin{aligned} R_t = Y_t(1) \mathcal{E}_t \left( -\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2V(2)} \cdot \widehat{S} \right) \\ + \int_0^t \mathcal{E}_u \left( -\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2V(2)} \cdot \widehat{S} \right) \frac{(\varphi_u(2)\rho_u^2 + \widehat{\lambda}_uV_u(2))\tilde{h}_u}{1 - \rho_u^2 + \rho_u^2V_u(2)} d\langle M \rangle_u \end{aligned}$$

is a local martingale. Let us show that  $R_t$  is a martingale.

As was already shown, the strategy

$$\tilde{\pi}_u = \frac{\psi_u(2)\rho_u^2 + \widehat{\lambda}_uY_u(2)}{1 - \rho^2 + \rho^2Y_u(2)} \mathcal{E}_u \left( -\frac{\psi(2)\rho^2 + \widehat{\lambda}Y(2)}{1 - \rho^2 + \rho^2Y(2)} \cdot \widehat{S} \right)$$

belongs to the class  $\Pi(G)$ .

Therefore (see (4.26)),

$$(4.34) \quad E \sup_{t \leq T} \mathcal{E}_t^2 \left( -\frac{\psi(2)\rho^2 + \widehat{\lambda}Y(2)}{1 - \rho^2 + \rho^2Y(2)} \cdot \widehat{S} \right) = E \sup_{t \leq T} \left( 1 + \int_0^t \tilde{\pi}_u d\widehat{S} \right)^2 < \infty,$$

and hence

$$Y_t(1) \mathcal{E}_t \left( -\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2V(2)} \cdot \widehat{S} \right) \in D.$$

On the other hand, the second term of (4.33) is the process of integrable variation, since  $\tilde{\pi} \in \Pi(G)$  and  $\tilde{h} \in \Pi(G)$  (see Lemma A.2) imply that

$$\begin{aligned} E \int_0^T \left| \mathcal{E}_u \left( -\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2V(2)} \cdot \widehat{S} \right) \frac{(\varphi_u(2)\rho_u^2 + \widehat{\lambda}_uV_u(2))\tilde{h}_u}{1 - \rho_u^2 + \rho_u^2V_u(2)} \right| d\langle M \rangle_u \\ = E \int_0^T |\tilde{\pi}_u \tilde{h}_u| d\langle M \rangle_u \leq E^{1/2} \int_0^T \tilde{\pi}_u^2 d\langle M \rangle_u E^{1/2} \int_0^T \tilde{h}_u^2 d\langle M \rangle_u < \infty. \end{aligned}$$

Therefore, the process  $R_t$  belongs to the class  $D$ , and hence it is a true martingale. By using the martingale property and the boundary condition we obtain

$$(4.35) \quad \begin{aligned} Y_t(1) = E \left( \widehat{H}_T \mathcal{E}_{tT} \left( -\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2V(2)} \cdot \widehat{S} \right) \right. \\ \left. + \int_t^T \mathcal{E}_{tu} \left( -\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2V(2)} \cdot \widehat{S} \right) \frac{(\varphi_u(2)\rho_u^2 + \widehat{\lambda}_uV_u(2))\tilde{h}_u}{1 - \rho_u^2 + \rho_u^2V_u(2)} d\langle M \rangle_u \middle| G_t \right). \end{aligned}$$

Thus, any solution of (4.5) is expressed explicitly in terms of  $(V(2), \varphi(2))$  in the form (4.35). Hence the solution of (4.5) is unique, and it coincides with  $V_t(1)$ .

It is evident that the solution of (4.6) is also unique.  $\square$

*Remark 4.1.* In the case  $F^S \subseteq G$  we have  $\rho_t = 1, \tilde{h}_t = 0$ , and  $\hat{S}_t = S_t$ , and (4.7) takes the form

$$\begin{aligned} \hat{X}_t^* &= x - \int_0^t \frac{\psi_u(2) + \hat{\lambda}_u Y_u(2)}{Y_u(2)} \hat{X}_u^* dS_u \\ &\quad + \int_0^t \frac{\psi_u(1) + \hat{\lambda}_u Y_u(1)}{Y_u(2)} dS_u. \end{aligned}$$

**COROLLARY 4.1.** *In addition to conditions (A)–(C) assume that  $\rho$  is a constant and the mean-variance tradeoff  $\langle \hat{\lambda} \cdot M \rangle_T$  is deterministic. Then the solution of (4.4) is the triple  $(Y(2), \psi(2), L(2))$ , with  $\psi(2) = 0, L(2) = 0$ , and*

$$(4.36) \quad Y_t(2) = V_t(2) = \nu \left( \rho, 1 - \rho^2 + \langle \hat{\lambda} \cdot M \rangle_T - \langle \hat{\lambda} \cdot M \rangle_t \right),$$

where  $\nu(\rho, \alpha)$  is the root of the equation

$$(4.37) \quad \frac{1 - \rho^2}{x} - \rho^2 \ln x = \alpha.$$

Besides,

$$(4.38) \quad \begin{aligned} Y_t(1) &= E \left( H \mathcal{E}_{tT} \left( -\frac{\hat{\lambda} V(2)}{1 - \rho^2 + \rho^2 V(2)} \cdot \hat{S} \right) \right. \\ &\quad \left. + \int_t^T \mathcal{E}_{tu} \left( -\frac{\hat{\lambda} V(2)}{1 - \rho^2 + \rho^2 V(2)} \cdot \hat{S} \right) \frac{\lambda_u V_u(2) \tilde{h}_u}{1 - \rho^2 + \rho^2 V_u(2)} d\langle M \rangle_u | G_t \right) \end{aligned}$$

uniquely solves (4.5), and the optimal filtered wealth process satisfies the linear equation

$$(4.39) \quad \begin{aligned} \hat{X}_t^* &= x - \int_0^t \frac{\hat{\lambda}_u V_u(2)}{1 - \rho^2 + \rho^2 V_u(2)} \hat{X}_u^* d\hat{S}_u \\ &\quad + \int_0^t \frac{\varphi_u(1)\rho^2 + \hat{\lambda}_u V_u(1) - \tilde{h}_u}{1 - \rho^2 + \rho^2 V_u(2)} d\hat{S}_u. \end{aligned}$$

*Proof.* The function  $f(x) = \frac{1-\rho^2}{x} - \rho^2 \ln x$  is differentiable and strictly decreasing on  $]0, \infty[$  and takes all values from  $] -\infty, +\infty[$ . So (4.37) admits a unique solution for all  $\alpha$ . Besides, the inverse function  $\alpha(x)$  is differentiable. Therefore  $Y_t(2)$  is a process of finite variation, and it is adapted since  $\langle \hat{\lambda} \cdot M \rangle_T$  is deterministic.

By definition of  $Y_t(2)$  we have that for all  $t \in [0, T]$

$$\frac{1 - \rho^2}{Y_t(2)} - \rho^2 \ln Y_t(2) = 1 - \rho^2 + \langle \hat{\lambda} \cdot M \rangle_T - \langle \hat{\lambda} \cdot M \rangle_t.$$

It is evident that for  $\alpha = 1 - \rho^2$  the solution of (4.37) is equal to 1, and it follows from (4.36) that  $Y(2)$  satisfies the boundary condition  $Y_T(2) = 1$ . Therefore

$$\begin{aligned} &\frac{1 - \rho^2}{Y_t(2)} - \rho^2 \ln Y_t(2) - (1 - \rho^2) \\ &= -(1 - \rho^2) \int_t^T d \frac{1}{Y_u(2)} + \rho^2 \int_t^T d \ln Y_u(2) \\ &= \int_t^T \left( \frac{1 - \rho^2}{Y_u^2(2)} + \frac{\rho^2}{Y_u(2)} \right) dY_u(2) \end{aligned}$$

and

$$\int_t^T \frac{1 - \rho^2 + \rho^2 Y_u(2)}{Y_u^2(2)} dY_u(2) = \langle \hat{\lambda} \cdot M \rangle_T - \langle \hat{\lambda} \cdot M \rangle_t$$

for all  $t \in [0, T]$ . Hence

$$\int_0^t \frac{1 - \rho^2 + \rho^2 Y_u(2)}{Y_u^2(2)} dY_u(2) = \langle \hat{\lambda} \cdot M \rangle_t,$$

and, by integrating both parts of this equality with respect to  $Y(2)/(1 - \rho^2 + \rho^2 Y(2))$ , we obtain that  $Y(2)$  satisfies

$$(4.40) \quad Y_t(2) = Y_0(2) + \int_0^t \frac{Y_u^2(2) \hat{\lambda}_u^2}{1 - \rho^2 + \rho^2 Y_u(2)} d\langle M \rangle_u,$$

which implies that the triple  $(Y(2), \psi(2) = 0, L(2) = 0)$  satisfies (4.4) and  $Y(2) = V(2)$  by Theorem 4.2. Equations (4.38) and (4.39) follow from (4.35) and (4.7), respectively, by taking  $\varphi(2) = 0$ .  $\square$

*Remark 4.2.* In case  $F^S \subseteq G$  we have  $\widehat{M} = M$  and  $\rho = 1$ . Therefore (4.40) is linear and  $Y_t(2) = e^{\langle \hat{\lambda} \cdot M \rangle_t - \langle \hat{\lambda} \cdot M \rangle_T}$ . In the case  $\mathcal{A} = G$  of complete information,  $Y_t(2) = e^{\langle \lambda \cdot N \rangle_t - \langle \lambda \cdot N \rangle_T}$ .

## 5. Diffusion market model.

*Example 1.* Let us consider the financial market model

$$\begin{aligned} d\tilde{S}_t &= \tilde{S}_t \mu_t(\eta) dt + \tilde{S}_t \sigma_t(\eta) dw_t^0, \\ d\eta_t &= a_t(\eta) dt + b_t(\eta) dw_t, \end{aligned}$$

subjected to initial conditions. Here  $w^0$  and  $w$  are correlated Brownian motions with  $Edw_t^0 dw_t = \rho dt, \rho \in (-1, 1)$ .

Let us write

$$w_t = \rho w_t^0 + \sqrt{1 - \rho^2} w_t^1,$$

where  $w^0$  and  $w^1$  are independent Brownian motions. It is evident that  $w^\perp = -\sqrt{1 - \rho^2} w^0 + \rho w^1$  is a Brownian motion independent of  $w$ , and one can express Brownian motions  $w^0$  and  $w^1$  in terms of  $w$  and  $w^\perp$  as

$$(5.1) \quad w_t^0 = \rho w_t - \sqrt{1 - \rho^2} w_t^\perp, \quad w_t^1 = \sqrt{1 - \rho^2} w_t + \rho w_t^\perp.$$

Suppose that  $b^2 > 0$ ,  $\sigma^2 > 0$ , and coefficients  $\mu, \sigma, a$ , and  $b$  are such that  $F_t^{S, \eta} = F_t^{w^0, w}$  and  $F_t^\eta = F_t^w$ .

We assume that an agent would like to hedge a contingent claim  $H$  (which can be a function of  $S_T$  and  $\eta_T$ ) using only observations based on the process  $\eta$ . So the stochastic basis will be  $(\Omega, \mathcal{F}, F_t, P)$ , where  $F_t$  is the natural filtration of  $(w^0, w)$  and the flow of observable events is  $G_t = F_t^w$ .

Also denote  $dS_t = \mu_t dt + \sigma_t dw_t^0$ , so that  $d\tilde{S}_t = \tilde{S}_t dS_t$  and  $S$  is the return of the stock.

Let  $\tilde{\pi}_t$  be the number of shares of the stock at time  $t$ . Then  $\pi_t = \tilde{\pi}_t \tilde{S}_t$  represents an amount of money invested in the stock at the time  $t \in [0, T]$ . We consider the mean-

variance hedging problem

(5.2)

$$\text{to minimize } E \left[ \left( x + \int_0^T \tilde{\pi}_t d\tilde{S}_t - H \right)^2 \right] \quad \text{over all } \tilde{\pi} \text{ for which } \tilde{\pi}\tilde{S} \in \Pi(G),$$

which is equivalent to studying the mean-variance hedging problem

$$\text{to minimize } E \left[ \left( x + \int_0^T \pi_t dS_t - H \right)^2 \right] \quad \text{over all } \pi \in \Pi(G).$$

*Remark 5.1.* Since  $S$  is not  $G$ -adapted,  $\tilde{\pi}_t$  and  $\tilde{\pi}_t\tilde{S}_t$  cannot be simultaneously  $G$ -predictable and the problem

$$\text{to minimize } E \left[ \left( x + \int_0^T \tilde{\pi}_t d\tilde{S}_t - H \right)^2 \right] \quad \text{over all } \tilde{\pi} \in \Pi(G)$$

is not equivalent to the problem (5.2). In this setting, condition (A) is not satisfied, and it needs separate consideration.

By comparing with (1.1) we get that in this case

$$M_t = \int_0^t \sigma_s dw_s^0, \quad \langle M \rangle_t = \int_0^t \sigma_s^2 ds, \quad \lambda_t = \frac{\mu_t}{\sigma_t^2}.$$

It is evident that  $w$  is a Brownian motion also with respect to the filtration  $F^{w^0, w^1}$  and condition (B) is satisfied. Therefore by Proposition 2.2

$$\widehat{M}_t = \rho \int_0^t \sigma_s dw_s.$$

By the integral representation theorem the GKW decompositions (3.2) and (3.3) take the following forms:

$$(5.3) \quad c_H = EH, \quad H_t = c_H + \int_0^t h_s \sigma_s dw_s^0 + \int_0^t h_s^1 dw_s^1,$$

$$(5.4) \quad H_t = c_H + \rho \int_0^t h_s^G \sigma_s dw_s + \int_0^t h_s^\perp dw_s^\perp.$$

By putting expressions (5.1) for  $w^0$  and  $w^1$  in (5.3) and equalizing integrands of (5.3) and (5.4), we obtain

$$h_t = \rho^2 h_t^G - \sqrt{1 - \rho^2} \frac{h_t^\perp}{\sigma_t}$$

and hence

$$\widehat{h}_t = \rho^2 \widehat{h}_t^G - \sqrt{1 - \rho^2} \frac{\widehat{h}_t^\perp}{\sigma_t}.$$

Therefore by the definition of  $\tilde{h}$

$$(5.5) \quad \tilde{h}_t = \rho^2 \widehat{h}_t^G - \widehat{h}_t = \sqrt{1 - \rho^2} \frac{\widehat{h}_t^\perp}{\sigma_t}.$$

By using notations

$$Z_s(0) = \rho\sigma_s\varphi_s(0), \quad Z_s(1) = \rho\sigma_s\varphi_s(1), \quad Z_s(2) = \rho\sigma_s\varphi_s(2), \quad \theta_s = \frac{\mu_s}{\sigma_s},$$

we obtain the following corollary of Theorem 4.1.

**COROLLARY 5.1.** *Let  $H$  be a square integrable  $F_T$ -measurable random variable. Then the processes  $V_t(0), V_t(1)$ , and  $V_t(2)$  from (4.3) satisfy the following system of backward equations:*

$$(5.6) \quad V_t(2) = V_0(2) + \int_0^t \frac{(\rho Z_s(2) + \theta_s V_s(2))^2}{1 - \rho^2 + \rho^2 V_s(2)} ds + \int_0^t Z_s(2) dw_s, \quad V_T(2) = 1,$$

$$(5.7) \quad \begin{aligned} V_t(1) &= V_0(1) + \int_0^t \frac{(\rho Z_s(2) + \theta_s V_s(2)) (\rho Z_s(1) + \theta_s V_s(1) - \sqrt{1 - \rho^2} \hat{h}_s^\perp)}{1 - \rho^2 + \rho^2 V_s(2)} ds \\ &\quad + \int_0^t Z_s(1) dw_s, \quad V_T(1) = E(H|G_T), \end{aligned}$$

$$(5.8) \quad \begin{aligned} V_t(0) &= V_0(0) + \int_0^t \frac{(\rho Z_s(1) + \theta_s V_s(1) - \sqrt{1 - \rho^2} \hat{h}_s^\perp)^2}{1 - \rho^2 + \rho^2 V_s(2)} ds \\ &\quad + \int_0^t Z_s(0) dw_s, \quad V_T(0) = E^2(H|G_T). \end{aligned}$$

Besides, the optimal wealth process  $\hat{X}^*$  satisfies the linear equation

$$(5.9) \quad \begin{aligned} \hat{X}_t^* &= x - \int_0^t \frac{\rho Z_s(2) + \theta_s V_s(2)}{1 - \rho^2 + \rho^2 V_s(2)} \hat{X}_s^* (\theta_s ds + \rho dw_s) \\ &\quad + \int_0^t \frac{\rho Z_s(1) + \theta_s V_s(1) - \sqrt{1 - \rho^2} \hat{h}_s^\perp}{1 - \rho^2 + \rho^2 V_s(2)} (\theta_s ds + \rho dw_s). \end{aligned}$$

Suppose now that  $\theta_t$  and  $\sigma_t$  are deterministic. Then the solution of (5.6) is the pair  $(V_t(2), Z_t(2))$ , where  $Z(2) = 0$  and  $V(2)$  satisfies the ordinary differential equation

$$(5.10) \quad \frac{dV_t(2)}{dt} = \frac{\theta_t^2 V_t^2(2)}{1 - \rho^2 + \rho^2 V_t(2)}, \quad V_T(2) = 1.$$

By solving this equation we obtain

$$(5.11) \quad V_t(2) = \nu \left( \rho, 1 - \rho^2 + \int_t^T \theta_s^2 ds \right) \equiv \nu_t^{\theta, \rho},$$

where  $\nu(\rho, \alpha)$  is the solution of (4.37). From (5.10) it follows that

$$(5.12) \quad \left( \ln \nu_t^{\theta, \rho} \right)' = \frac{\theta_t^2 \nu_t^{\theta, \rho}}{1 - \rho^2 + \rho^2 \nu_t^{\theta, \rho}} \quad \text{and} \quad \ln \frac{\nu_s^{\theta, \rho}}{\nu_t^{\theta, \rho}} = \int_t^s \frac{\theta_r^2 \nu_r^{\theta, \rho} dr}{1 - \rho^2 + \rho^2 \nu_r^{\theta, \rho}}.$$

If we solve the linear BSDE (5.7) and use (5.12), we obtain

$$\begin{aligned} V_t(1) &= E \left[ \widehat{H}_T(w) \mathcal{E}_{tT} \left( - \int_0^\cdot \frac{\theta_r \nu_r^{\theta, \rho}}{1 - \rho^2 + \rho^2 \nu_r^{\theta, \rho}} (\theta_r dr + \rho dw_r) \right) \middle| G_t \right], \\ \int_t^T \frac{\theta_s \nu_s^{\theta, \rho} \sigma_s}{1 - \rho^2 + \rho^2 \nu_s^{\theta, \rho}} E \left[ \tilde{h}_s(w) \mathcal{E}_{ts} \left( - \int_0^\cdot \frac{\theta_r \nu_r^{\theta, \rho}}{1 - \rho^2 + \rho^2 \nu_r^{\theta, \rho}} (\theta_r dr + \rho dw_r) \right) \middle| G_t \right] ds \\ &= \nu_t^{\theta, \rho} E \left[ \widehat{H}_T(w) \mathcal{E}_{tT} \left( - \int_0^\cdot \frac{\theta_r \nu_r^{\theta, \rho}}{1 - \rho^2 + \rho^2 \nu_r^{\theta, \rho}} \rho dw_r \right) \middle| G_t \right] \\ &+ \nu_t^{\theta, \rho} \int_t^T \frac{\mu_s}{1 - \rho^2 + \rho^2 \nu_s^{\theta, \rho}} E \left[ \tilde{h}_s(w) \mathcal{E}_{ts} \left( - \int_0^\cdot \frac{\theta_r \nu_r^{\theta, \rho}}{1 - \rho^2 + \rho^2 \nu_r^{\theta, \rho}} \rho dw_r \right) \middle| G_t \right] ds. \end{aligned}$$

By using the Girsanov theorem we finally get

$$\begin{aligned} V_t(1) &= \nu_t^{\theta, \rho} E \left[ \widehat{H}_T \left( \rho \int_0^\cdot \frac{\theta_r \nu_r^{\theta, \rho}}{1 - \rho^2 + \rho^2 \nu_r^{\theta, \rho}} dr + w \right) \middle| G_t \right] \\ (5.13) \quad &+ \nu_t^{\theta, \rho} \int_t^T \frac{\mu_s}{1 - \rho^2 + \rho^2 \nu_s^{\theta, \rho}} E \left[ \tilde{h}_s \left( \rho \int_0^\cdot \frac{\theta_r \nu_r^{\theta, \rho}}{1 - \rho^2 + \rho^2 \nu_r^{\theta, \rho}} dr + w \right) \middle| G_t \right] ds. \end{aligned}$$

Besides, the optimal strategy is of the form

$$\begin{aligned} \pi_t^* &= - \frac{\theta_t V_t(2)}{(1 - \rho^2 + \rho^2 V_t(2)) \sigma_t} \widehat{X}_t^* \\ &+ \frac{\rho Z_t(1) + \theta_t V_t(1) - \sqrt{1 - \rho^2} \widehat{h}_t^\perp}{(1 - \rho^2 + \rho^2 V_t(2)) \sigma_t}. \quad \square \end{aligned}$$

If in addition  $\mu$  and  $\sigma$  are constants and the contingent claim is of the form  $H = \mathcal{H}(S_T, \eta_T)$ , then one can give an explicit expressions also for  $\tilde{h}$ ,  $\widehat{h}^\perp$ ,  $\widehat{H}$ , and  $Z(1)$ .

*Example 2.* In Frey and Runggaldier [9] the incomplete-information situation arises, assuming that the hedger is unable to monitor the asset continuously but is confined to observations at discrete random points in time  $\tau_1, \tau_2, \dots, \tau_n$ . Perhaps it is more natural to assume that the hedger has access to price information on full intervals  $[\sigma_1, \tau_1], [\sigma_2, \tau_2], \dots, [\sigma_n, \tau_n]$ . For the models with nonzero drifts, even the case  $n = 1$  is nontrivial. Here we consider this case in detail.

Let us consider the financial market model

$$d\tilde{S}_t = \mu \tilde{S}_t dt + \sigma \tilde{S}_t dW_t, \quad S_0 = S,$$

where  $W$  is a standard Brownian motion and the coefficients  $\mu$  and  $\sigma$  are constants. Assume that an investor observes only the returns  $S_t - S_0 = \int_0^t \frac{1}{S_u} d\tilde{S}_u$  of the stock prices up to a random moment  $\tau$  before the expiration date  $T$ . Let  $\mathcal{A}_t = F_t^S$ , and let  $\tau$  be a stopping time with respect to  $F^S$ . Then the filtration  $G_t$  of observable events is equal to the filtration  $F_{t \wedge \tau}^S$ .

Consider the mean-variance hedging problem

$$\text{to minimize } E \left[ \left( x + \int_0^T \pi_t dS_t - H \right)^2 \right] \quad \text{over all } \pi \in \Pi(G),$$

where  $\pi_t$  is a dollar amount invested in the stock at time  $t$ .



By comparing with (1.1) we get that in this case

$$N_t = M_t = \sigma W_t, \quad \langle M \rangle_t = \sigma^2 t, \quad \lambda_t = \frac{\mu}{\sigma^2}.$$

Let  $\theta = \frac{\mu}{\sigma}$ . The measure  $Q$  defined by  $dQ = \mathcal{E}_T(\theta W)dP$  is a unique martingale measure for  $S$ , and it is evident that  $Q$  satisfies the reverse Hölder condition. It is also evident that any  $G$ -martingale is  $F^S$ -martingale and that conditions (A)–(C) are satisfied. Besides,

$$(5.14) \quad E(W_t|G_t) = W_{t \wedge \tau}, \quad \hat{S}_t = \mu t + \sigma W_{t \wedge \tau} \quad \text{and} \quad \rho_t = I_{\{t \leq \tau\}}.$$

By the integral representation theorem

$$(5.15) \quad E(H|F_t^S) = EH + \int_0^t h_u \sigma dW_u$$

for  $F$ -predictable  $W$ -integrable process  $h$ . On the other hand, by the GKW decomposition with respect to the martingale  $W^\tau = (W_{t \wedge \tau}, t \in [0, T])$ ,

$$(5.16) \quad E(H|F_t^S) = EH + \int_0^t h_u^G \sigma dW_u^\tau + L_t^G$$

for  $F^S$ -predictable process  $h^G$  and  $F^S$  martingale  $L^G$  strongly orthogonal to  $W^\tau$ . Therefore, by equalizing the right-hand sides of (5.15) and (5.16) and taking the mutual characteristics of both parts with  $W^\tau$ , we obtain  $\int_0^{t \wedge \tau} (h_u^G \rho_u^2 - h_u) du = 0$  and hence

$$(5.17) \quad \int_0^t \tilde{h}_u du = \int_0^t (\hat{h}_u^G I_{(u \leq \tau)} - \hat{h}_u) du = - \int_0^t I_{(u > \tau)} E(h_u | F_\tau^S) du.$$

Therefore, by using notations

$$Z_s(0) = \rho \sigma \varphi_s(0), \quad Z_s(1) = \rho \sigma \varphi_s(1), \quad Z_s(2) = \rho \sigma \varphi_s(2),$$

it follows from Theorem 4.1 that the processes  $(V_t(2), Z_t(2))$  and  $(V_t(1), Z_t(1))$  satisfy the following system of backward equations:

$$(5.18) \quad \begin{aligned} V_t(2) &= V_0(2) + \int_0^{t \wedge \tau} \frac{(Z_s(2) + \theta V_s(2))^2}{V_s(2)} ds + \int_{t \wedge \tau}^t \theta^2 V_s^2(2) ds \\ &+ \int_0^{t \wedge \tau} Z_s(2) dW_s, \quad V_T(2) = 1, \end{aligned}$$

$$(5.19) \quad \begin{aligned} V_t(1) &= V_0(1) + \int_0^{t \wedge \tau} \frac{(Z_s(2) + \theta V_s(2))(Z_s(1) + \theta V_s(1))}{V_s(2)} ds \\ &+ \int_{t \wedge \tau}^t \theta V_s(2) (\theta V_s(1) + E(h_s | F_\tau^S)) ds \\ &+ \int_0^{t \wedge \tau} Z_s(1) dW_s, \quad V_T(1) = E(H|G_T). \end{aligned}$$

Equation (5.18) admits in this case an explicit solution. To obtain the solution one should solve first the equation

$$(5.20) \quad U_t = U_0 + \int_0^t \theta^2 U_s^2 ds, \quad U_T = 1,$$

in the time interval  $[\tau, T]$  and then the BSDE

$$(5.21) \quad V_t(2) = V_0(2) + \int_0^t \frac{(Z_s(2) + \theta V_s(2))^2}{V_s(2)} ds + \int_0^t Z_s(2) dW_s$$

in the interval  $[0, \tau]$ , with the boundary condition  $V_\tau(2) = U_\tau$ . The solution of (5.20) is

$$U_t = \frac{1}{1 + \theta^2(T - t)},$$

and the solution of (5.21) is expressed as

$$V_t(2) = \frac{1}{E((1 + \theta^2(T - \tau))\mathcal{E}_{t,\tau}^2(-\theta W)|F_t^S)}$$

(this can be verified by applying the Itô formula for the process  $V_t^{-1}(2)\mathcal{E}_t^2(-\theta W)$  and by using the fact that this process is a martingale). Therefore

$$(5.22) \quad V_t(2) = \begin{cases} \frac{1}{1 + \theta^2(T - t)} & \text{if } t \geq \tau, \\ \frac{1}{E((1 + \theta^2(T - \tau))\mathcal{E}_{t,\tau}^2(-\theta W)|F_t^S)} & \text{if } t \leq \tau. \end{cases}$$

According to (4.37), taking in mind (5.14), (5.17), and the fact that  $e^{-\int_t^T \theta^2 V_u(2) du} = \frac{1}{1 + \theta^2(T - t)}$  on the set  $t \geq \tau$ , the solution of (5.19) is equal to

(5.23)

$$\begin{aligned} V_t(1) = & E\left(\frac{H}{1 + \theta^2(T - t)} + \int_t^T \frac{\theta V_u(2) h_u du}{1 + \theta^2(T - u)} | F_\tau^S\right) I_{(t > \tau)} \\ & + E\left(\mathcal{E}_{t,\tau}\left(-\frac{\varphi(2) + \lambda V(2)}{V(2)} \cdot S\right) \left(\frac{H}{1 + \theta^2(T - \tau)} + \int_\tau^T \frac{\theta V_u(2) h_u du}{1 + \theta^2(T - u)}\right) | F_t^S\right) I_{(t \leq \tau)}. \end{aligned}$$

By Theorem 4.1 the optimal filtered wealth process is a solution of a linear SDE, which takes in this case the following form:

$$(5.24) \quad \begin{aligned} \hat{X}_t^* = & x - \int_0^{t \wedge \tau} \frac{\varphi_u(2) + \theta V_u(2)}{V_u(2)} \hat{X}_u^* (\theta du + dW_u) - \int_{t \wedge \tau}^t \theta^2 V_u(2) \hat{X}_u^* du \\ & + \int_0^{t \wedge \tau} \frac{\varphi_u(1) + \theta V_u(1)}{V_u(2)} (\theta du + dW_u) + \int_{t \wedge \tau}^t (\theta^2 V_u(1) + \mu E(h_u | F_\tau^S)) du. \end{aligned}$$

The optimal strategy is equal to

$$(5.25) \quad \begin{aligned} \pi_t^* = & \left[ -\frac{\varphi_t(2) + \theta V_t(2)}{V_t(2)} I_{(t \leq \tau)} - \theta^2 V_t(2) I_{(t > \tau)} \right] \hat{X}_t^* \\ & + \frac{\varphi_t(1) + \theta V_t(1)}{V_t(2)} I_{(t \leq \tau)} + (\theta^2 V_t(1) + \mu E(h_t | F_\tau^S)) I_{(t > \tau)}, \end{aligned}$$

where  $\hat{X}_t^*$  is a solution of the linear equation (5.24),  $V(2)$  and  $V(1)$  are given by (5.22) and (5.23), and  $\varphi(2)$  and  $\varphi(1)$  are integrands of their martingale parts, respectively. In particular the optimal strategy in time interval  $[\tau, T]$  (i.e., after interrupting observations) is of the form

$$(5.26) \quad \pi_t^* = -\theta^2 V_t(2) \hat{X}_t^* + \theta^2 V_t(1) + \mu E(h_t | F_\tau^S),$$

where

$$\widehat{X}_t^* = \frac{\widehat{X}_\tau^*}{1 + \theta^2(t - \tau)} - \int_\tau^t (\theta^2 V_u(1) - \mu E(h_u | F_\tau^S)) \frac{1}{1 + \theta^2(t - u)} du.$$

For instance, if  $\tau$  is deterministic, then  $V_t(2)$  is also deterministic:

$$V_t(2) = \begin{cases} \frac{1}{1 + \theta^2(T - t)} & \text{if } t \geq \tau, \\ \frac{1}{1 + \theta^2(T - t)} e^{-\theta^2(\tau - t)} & \text{if } t \leq \tau, \end{cases}$$

and  $\varphi(2) = 0$ .

Note that it is not optimal to do nothing after interrupting observations, and in order to act optimally one should change the strategy deterministically as it is given by (5.26).

**Appendix A.** For convenience we give the proofs of the following assertions used in the paper.

LEMMA A.1. *Let conditions (A)–(C) be satisfied and  $\widehat{M}_t = E(M_t | G_t)$ . Then  $\langle \widehat{M} \rangle$  is absolutely continuous w.r.t.  $\langle M \rangle$  and  $\mu^{(M)}$  a.e.*

$$\rho_t^2 = \frac{d\langle \widehat{M} \rangle_t}{d\langle M \rangle_t} \leq 1.$$

*Proof.* By (2.4) for any bounded  $G$ -predictable process  $h$

$$\begin{aligned} E \int_0^t h_s^2 d\langle \widehat{M} \rangle_s &= E \left( \int_0^t h_s d\widehat{M}_s \right)^2 = E \left( E \left( \int_0^t h_s dM_s | G_t \right) \right)^2 \\ (A.1) \qquad \qquad \qquad &\leq E \left( \int_0^t h_s dM_s \right)^2 = E \int_0^t h_s^2 d\langle M \rangle_s, \end{aligned}$$

which implies that  $\langle \widehat{M} \rangle$  is absolutely continuous w.r.t.  $\langle M \rangle$ , i.e.,

$$\langle \widehat{M} \rangle_t = \int_0^t \rho_s^2 d\langle M \rangle_s$$

for a  $G$ -predictable process  $\rho$ .  $\square$

Moreover (A.1) implies that the process  $\langle M \rangle - \langle \widehat{M} \rangle$  is increasing and hence  $\rho^2 \leq 1$   $\mu^{(M)}$  a.e.

LEMMA A.2. *Let  $H \in L^2(P, F_T)$ , and let conditions (A)–(C) be satisfied. Then*

$$E \int_0^T \tilde{h}_u^2 d\langle M \rangle_u < \infty.$$

*Proof.* It is evident that

$$E \int_0^T (h_u^G)^2 d\langle \widehat{M} \rangle_u < \infty, \quad E \int_0^T h_u^2 d\langle M \rangle_u < \infty.$$

Therefore, by the definition of  $\tilde{h}$  and Lemma A.1,

$$\begin{aligned} & E \int_0^T \tilde{h}_u^2 d\langle M \rangle_u \\ & \leq 2E \int_0^T \hat{h}_u^2 d\langle M \rangle_u + 2E \int_0^T \left( \widehat{h}_u^G \right)^2 \rho_u^4 d\langle M \rangle_u \\ & \leq 2E \int_0^T h_u^2 d\langle M \rangle_u + 2E \int_0^T \left( h_u^G \right)^2 \rho_u^2 d\langle \widehat{M} \rangle_u < \infty. \end{aligned}$$

Thus  $\tilde{h} \in \Pi(G)$  by Remark 2.5.  $\square$

LEMMA A.3. (a) Let  $Y = (Y_t, t \in [0, T])$  be a bounded positive submartingale with the canonical decomposition

$$Y_t = Y_0 + B_t + m_t,$$

where  $B$  is a predictable increasing process and  $m$  is a martingale. Then  $m \in BMO$ .

(b) In particular the martingale part of  $V(2)$  belongs to  $BMO$ . If  $H$  is bounded, then martingale parts of  $V(0)$  and  $V(1)$  also belong to the class  $BMO$ , i.e., for  $i = 0, 1, 2$ ,

$$(A.2) \quad E \left( \int_{\tau}^T \varphi_u^2(i) \rho_u^2 d\langle M \rangle_u | G_{\tau} \right) + E (\langle m(i) \rangle_T - \langle m(i) \rangle_{\tau} | G_{\tau}) \leq C$$

for every stopping time  $\tau$ .

*Proof.* By applying the Itô formula for  $Y_T^2 - Y_{\tau}^2$  we have

$$(A.3) \quad \langle m \rangle_T - \langle m \rangle_{\tau} + 2 \int_{\tau}^T Y_u dB_u + 2 \int_{\tau}^T Y_u dm_u = Y_T^2 - Y_{\tau}^2 \leq \text{const}$$

Since  $Y$  is positive and  $B$  is an increasing process, by taking conditional expectations in (A.3) we obtain

$$E(\langle m \rangle_T - \langle m \rangle_{\tau} | F_{\tau}) \leq \text{const}$$

for any stopping time  $\tau$ , and hence  $m \in BMO$ .

(A.2) follows from assertion (a) applied for positive submartingales  $V(0)$ ,  $V(2)$ , and  $V(0) + V(2) - 2V(1)$ . For the case  $i = 1$  one should take into account also the inequality

$$\langle m(1) \rangle_t \leq \text{const}(\langle m(0) + m(2) - 2m(1) \rangle_t + \langle m(0) \rangle_t + \langle m(2) \rangle_t). \quad \square$$

**Acknowledgments.** We would like to thank N. Lazrieva and anonymous referees for useful remarks and comments.

#### REFERENCES

- [1] J. M. BISMUT, *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. Appl., 44 (1973), pp. 384–404.
- [2] R. CHITASHVILI, *Martingale Ideology in the Theory of Controlled Stochastic Processes*, Lecture Notes in Math. 1021, Springer-Verlag, New York, 1983, pp. 73–92.
- [3] G. B. DI MASI, E. PLATEN, AND W. J. Runggaldier, *Hedging of options under discrete observation on assets with stochastic volatility*, in Seminar on Stochastics Analysis and Random Fields Applications, Progr. Probab. 36, Birkhäuser-Verlag, Basel, Switzerland, 1995, pp. 359–364.

- [4] F. DELBAEN, P. MONAT, W. SCHACHERMAYER, W. SCHWEIZER, AND C. STRICKER, *Weighted norm inequalities and hedging in incomplete markets*, Finance Stoch., 1 (1997), pp. 181–227.
- [5] C. DELLACHERIE AND P. A. MEYER, *Probabilités et Potentiel*, II. Hermann, Paris, 1980.
- [6] D. DUFFIE AND H. R. RICHARDSON, *Mean-variance hedging in continuous time*, Ann. Appl. Probab., 1 (1991), pp. 1–15.
- [7] N. EL KAROUI AND M. C. QUENEZ, *Dynamic programming and pricing of contingent claims in an incomplete market*, SIAM J. Control Optim., 33 (1995), pp. 29–66.
- [8] H. FÖLLMER AND D. SONDERMANN, *Hedging non-redundant contingent claims*, in Contrib. Math. Econ., Hon. G. Debreu, W. Hildenbrand, and A. Mas-Collel, eds., North-Holland, Amsterdam, 1986, pp. 205–223.
- [9] R. FREY AND W. J. Runggaldier, *Risk-minimizing hedging strategies under restricted information: The case of stochastic volatility models observable only at discrete random times. Financial optimization*, Math. Methods Oper. Res., 50 (1999), pp. 339–350.
- [10] C. GOURIEROUX, J. P. LAURENT, AND H. PHAM, *Mean-variance hedging and numeraire*, Math. Finance, 8 (1998), pp. 179–200.
- [11] D. HEATH, E. PLATEN, AND M. SCHWEIZER, *A comparison of two quadratic approaches to hedging in incomplete markets*, Math. Finance, 11 (2001), pp. 385–413.
- [12] C. HIPPI, *Hedging general claims*, in Proceedings of the 3rd AFIR Colloquium, 2, Rome, 1993, pp. 603–613.
- [13] J. JACOD, *Calcul Stochastique et Problèmes des Martingales*, Lecture Notes in Math. 714, Springer-Verlag, Berlin, 1979.
- [14] N. KAZAMAKI, *Continuous Exponential Martingales and BMO*, Lecture Notes in Math. 1579, Springer-Verlag, New York, 1994.
- [15] D. O. KRAMKOV, *Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets*, Probab. Theory Related Fields, 105 (1996), pp. 459–479.
- [16] M. KOBYLANSKI, *Backward stochastic differential equation and partial differential equations with quadratic growth*, Ann. Probab., 28 (2000), pp. 558–602.
- [17] J. P. LEPELTIER AND J. SAN MARTIN, *Existence for BSDE with superlinear-quadratic coefficient*, Stoch. Stoch. Rep., 63 (1998), pp. 227–240.
- [18] R. SH. LIPTZER AND A. N. SHIRYAYEV, *Martingale Theory*, Nauka, Moscow, 1986.
- [19] M. MANIA AND R. TEVZADZE, *Backward stochastic PDE and imperfect hedging*, Int. J. Theor. Appl. Finance, 6 (2003), pp. 663–692.
- [20] M. A. MORLAIS, *Quadratic Backward Stochastic Differential Equations (BSDEs) Driven by a Continuous Martingale and Application to the Utility Maximization Problem*, <http://hal.ccsd.cnrs.fr/ccsd-00020254/>
- [21] E. PARDOUX AND S. G. PENG, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.
- [22] H. PHAM, *Mean-variance hedging for partially observed drift processes*, Int. J. Theor. Appl. Finance, 4 (2001), pp. 263–284.
- [23] T. RHEINLÄNDER AND M. SCHWEIZER, *On  $L^2$ -projections on a space of stochastic integrals*, Ann. Probab., 25 (1997), pp. 1810–1831.
- [24] M. SCHÄL, *On quadratic cost criteria for option hedging*, Math. Oper. Res., 19 (1994), pp. 121–131.
- [25] M. SCHWEIZER, *Mean-variance hedging for general claims*, Ann. Appl. Probab., 2 (1992), pp. 171–179.
- [26] M. SCHWEIZER, *Approximating random variables by stochastic integrals*, Ann. Probab., 22 (1994), pp. 1536–1575.
- [27] M. SCHWEIZER, *Risk-minimizing hedging strategies under restricted information*, Math. Finance, 4 (1994), pp. 327–342.
- [28] R. TEVZADZE, *Solvability of backward stochastic differential equations with quadratic growth*, Stochastic Proc. Appl., 118 (2008), pp. 503–515.

## A KNOWLEDGE-GRADIENT POLICY FOR SEQUENTIAL INFORMATION COLLECTION\*

PETER I. FRAZIER<sup>†</sup>, WARREN B. POWELL<sup>†</sup>, AND SAVAS DAYANIK<sup>†</sup>

**Abstract.** In a sequential Bayesian ranking and selection problem with independent normal populations and common known variance, we study a previously introduced measurement policy which we refer to as the knowledge-gradient policy. This policy myopically maximizes the expected increment in the value of information in each time period, where the value is measured according to the terminal utility function. We show that the knowledge-gradient policy is optimal both when the horizon is a single time period and in the limit as the horizon extends to infinity. We show furthermore that, in some special cases, the knowledge-gradient policy is optimal regardless of the length of any given fixed total sampling horizon. We bound the knowledge-gradient policy's suboptimality in the remaining cases, and show through simulations that it performs competitively with or significantly better than other policies.

**Key words.** ranking and selection, Bayesian statistics, sequential decision analysis

**AMS subject classifications.** 62F07, 62F15, 62L05

**DOI.** 10.1137/070693424

**1. Introduction.** We consider a ranking and selection problem in which we are faced with  $M \geq 2$  alternatives, each of which can be measured sequentially to estimate its constant but unknown underlying average performance. The measurements are noisy, and as we obtain more measurements, our estimates become more accurate. We assume normally distributed measurement noise and independent normal Bayesian priors for each alternative's underlying average performance. We have a budget of  $N$  measurements to spread over the  $M$  alternatives before deciding which is best. The goal is to choose the alternative with the best underlying average performance.

Information collection problems of this type arise in a number of applications:

- (i) Choosing the chemical compound from a library of existing test compounds that has the greatest effectiveness against a particular disease. A compound's effectiveness may be measured by exposing cultured cells infected with the disease to the compound and observing the result. The compound found most effective will be developed into a drug for treating the disease.
- (ii) Choosing the most efficient of several alternative assembly line configurations. We may spend a certain short amount of time testing different configurations, but once we put one particular configuration into production, that choice will remain in production for a period of several years.
- (iii) Selecting the best of several policies applied to a stochastic Markov decision process. The policies may be evaluated only through Monte Carlo simulation, so a method of ranking and selection is needed to determine which policy is best. This selection may be as part of a larger algorithm for finding the optimal policy as in evolutionary policy iteration [3].

---

\*Received by the editors May 31, 2007; accepted for publication (in revised form) April 29, 2008; published electronically September 8, 2008.

<http://www.siam.org/journals/sicon/47-5/69342.html>

<sup>†</sup>Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ 08544 (pfrazier@princeton.edu, powell@princeton.edu, sdayanik@princeton.edu). The second author's research was partially supported by AFOSR contract FA9550-08-1-0195. The third author's research was partially supported by the Center for Dynamic Data Analysis for Homeland Security, ONR Award N00014-07-1-0150.

In this article we study a measurement policy introduced in [16] under the name of the  $(R_1, \dots, R_1)$  policy, and referred to herein as the knowledge-gradient (KG) policy. We briefly describe this policy and leave further description for section 4.1. Let  $\mu_x^n$  and  $(\sigma_x^n)^2$  denote the mean and variance of the posterior predictive distribution for the unknown value of alternative  $x$  after the first  $n$  measurements. Then the KG policy is the policy that chooses its  $(n+1)$ st measurement  $X^{KG}((\mu_1^n, \sigma_1^n), \dots, (\mu_M^n, \sigma_M^n))$  from within  $\{1, \dots, M\}$  to maximize the single-period expected increase in value,  $\mathbb{E}_n[(\max_{x'} \mu_{x'}^{n+1}) - (\max_{x'} \mu_{x'}^n)]$ , where  $\mathbb{E}_n$  indicates the conditional expectation with respect to what is known after the first  $n$  measurements. That is,

$$X^{KG}((\mu_1^n, \sigma_1^n), \dots, (\mu_M^n, \sigma_M^n)) \in \arg \max_{x^n \in \{1, \dots, M\}} \mathbb{E}_n \left[ (\max_{x'} \mu_{x'}^{n+1}) - (\max_{x'} \mu_{x'}^n) \right].$$

In this expression the expectation is implicitly a function of  $x^n$ , the measurement decision at time  $n$ . If the maximum is attained by more than one alternative, then we choose the one with the smallest index. As the terminal reward is given by  $\max_{x=1, \dots, M} \mu_x^N$ , this policy is like a gradient ascent algorithm on a utility surface with domain parameterized by the state of knowledge  $((\mu_1, \sigma_1), \dots, (\mu_M, \sigma_M))$ . It may also be viewed as a single-step Bayesian look-ahead policy.

In this work we continue the analysis of [16]. We demonstrate that the KG policy, introduced there as the most rudimentary of a collection of potential policies and studied for its simplicity but neglected thereafter, is actually a powerful and efficient tool for ranking and selection that should be considered for application alongside current state-of-the-art policies. As discussed in detail in section 2, a number of other sequential Bayesian look-ahead policies have been derived in recent years by solving a sequence of single-stage optimization problems just as the KG policy does, and, among these, the optimal computing budget allocation for linear loss of [18] and the LL(S) policy of [12] assume situations most similar to the one assumed here. The KG policy differs, however, from these other policies in that it solves its single-stage problem exactly, while the other policies must use approximations. We believe that solving the look-ahead problem exactly offers an advantage.

After formulating the problem in section 3 and defining the policy in section 4, we show in section 5 that the KG policy is optimal in the limit as  $N \rightarrow \infty$  in the sense that the policy incurs no opportunity cost in the limit as infinitely many measurements are allowed. Also, by its construction and as noted in [16], KG is optimal when there is only one measurement remaining. This provides optimality guarantees at two extremes:  $N$  large and  $N$  small. While many policies are asymptotically optimal without performing particularly well in the finite sample case, a policy with both kinds of optimality satisfies a more stringent performance check. For example, the equal-allocation policy is asymptotically optimal, but it is not optimal when  $N = 1$ , except in certain special cases, and performs poorly overall. In the other extreme, myopic policies for generic Markov decision processes often perform poorly because they ignore long-term rewards. By being optimal for both  $N = 1$  and  $N = \infty$ , KG avoids the problem that most afflicts other myopic policies, while retaining single-sample optimality.

In accordance with our belief that optimality at two extremes suggests good performance in the region between, we provide a bound on the policy's suboptimality for finite  $N$  in section 6. In section 7 we introduce the KG persistence property and use it to show both optimality for the case when  $M = 2$  and for a further special case in which the means and variances are ordered. Our proof that KG is optimal when

$M = 2$  confirms a claim made by Gupta and Miescke [15], who showed its optimality among deterministic policies for  $M = 2$ , but did not offer a formal proof for optimality among sequential policies. Finally, in section 8, we demonstrate in numerical experiments that KG performs competitively against the other policies discussed here. In particular, the KG policy is best according to the measure of average performance across a number of randomly generated problems, and the margin by which it outperforms the best competing policies on the most favorable problems is significantly larger than the margin by which it is outperformed on the most unfavorable problems.

**2. Literature review.** The KG policy was introduced in [16] as the simplest of a collection of look-ahead policies and was studied because its simplicity provided tractability, but this simple policy has seldom been studied or applied in the years since. Instead, a number of more complex Bayesian look-ahead policies have been introduced. A series of researchers beginning with [4] and continuing with [5], [9], [7], [8], [6] proposed and then refined a family of policies known as the optimal computing budget allocation (OCBA). These policies are derived by formulating a static optimization problem in which one chooses the measurements to maximize the probability of later correctly selecting the best alternative. OCBA policies solve this optimization problem by approximating the objective function with various bounds and relaxations, and by assuming that the predictive mean will remain unchanged by measurement. They then solve the approximate problem using gradient ascent or greedy heuristics, or with an asymptotic solution that is exact in the limit as the number of measurements in the second stage is large. All OCBA policies assume normal samples with known sampling variance, but in practice one may estimate this variance through sampling.

Any OCBA policy can be extended to multistage or fully sequential problems by performing the second stage of the two-stage policy repeatedly, at each stage calling all previous measurements the first stage and the set of measurements to be taken next the second stage. It is in this extension that one sees the similarity to the one-step Bayesian look-ahead approach of KG, which extends the one-stage policy which is optimal with one measurement remaining to a sequential policy by supposing at each point in time that the current measurement will be the last.

The OCBA policies mentioned above are designed to maximize the probability of correctly selecting the best alternative, while KG is designed to maximize the expected value of the chosen alternative. These different objective functions are also termed 0–1 loss and linear loss, respectively. They are similar but not identical, 0–1 loss perhaps being more appropriate when knowledge of the identity of the best is intrinsically valuable (and where accidentally choosing the second best is nearly as harmful as choosing the worst), and linear loss being more appropriate when value is obtained directly by implementing the chosen alternative.

Recently [18] introduced an OCBA policy designed to minimize expected linear loss. Although more similar to KG than other OCBA policies, it differs in that it uses the Bonferroni inequality to approximate the linear loss objective function for a single stage, and then solves the approximate problem using a second approximation which is accurate in the limit as the second stage is large. This is in contrast to KG, which solves the single-stage problem exactly. The OCBA policy in [18] does not assume, as the other OCBA approaches do, that the posterior predictive mean is equal to the prior predictive mean, and in this regard it is more similar to the approach of [12] discussed below.

A set of Bayesian look-ahead ranking and selection policies distinct from OCBA



were introduced in [12]. They differ by not assuming the predictive means equal through time and by allowing the sampling variance to be unknown. This causes the posterior predictive mean to be student- $t$  distributed, inducing an optimization problem governing the second-stage allocation with an objective function that is somewhat different from that in OCBA formulations. This objective function, corresponding to expected loss, is bounded below, and this lower bound is then approximately minimized. The resulting solution minimizes the lower bound exactly in the limit as sampling costs are small, or as the number of second-stage measurements is large.

Six policies are derived in total by considering both 0–1 and linear loss under three different settings: two-stage measurements with a budget constraint; two-stage without a budget constraint; and sequential. Among these policies, the one most similar to KG is LL(S), which uses linear loss in a sequential setting, allocating  $\tau$  measurements at a time.

In [10] an unknown-variance version of the KG policy was developed under the name LL<sub>1</sub>. The authors compared LL<sub>1</sub> to LL(S) using Monte Carlo simulations and found that LL<sub>1</sub> performed well for a small sampling budget, but degraded in performance as the sampling budget increased. We briefly discuss how these results relate to our own in section 8.

In addition to the Bayesian approaches to sequentially ranking and selecting normal populations described thus far, a substantial amount of progress has been made using a frequentist approach. We do not review this literature in detail, but state only that an overview may be found in [1] and that a more recent policy which performs quite well in the multistage setting with normal rewards is given in [23], [22]. Other sequential and staged policies for independent normal rewards with frequentist guarantees include those in [25], [27], [17], [26], and [24].

Sequential tests also exist which choose measurements based upon confidence bounds for the value  $Y_x$ . Such tests include interval estimation [19], which was developed for on-line bandit-style learning in a reinforcement learning setting, and upper confidence bound estimation [3], which was developed for estimating value functions for Markov decision processes. Both tests form frequentist confidence intervals for each  $Y_x$  and then select the alternative with the largest upper bound on its confidence interval for measurement. Such policies have general applicability beyond the independent normal setting discussed here.

**3. Problem formulation.** We state a formal model for our problem, including transition and objective functions. We then formulate the problem as a dynamic program.

**3.1. A formal model.** Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space and let  $\{1, \dots, M\}$  be the set of alternatives. For each  $x \in \{1, \dots, M\}$  define a random variable  $Y_x$  to be the true underlying value of alternative  $x$ . We assume a Bayesian setting for the problem in which we have a multivariate normal prior predictive distribution for the random vector  $Y$ , and we further assume that the components of  $Y$  are independent under the prior and that  $\max_{x=1, \dots, M} |Y_x|$  is integrable. We will be allotted exactly  $N$  measurements, and time will be indexed using  $n$  with the first measurement decision made at time 0. At each time  $0 \leq n < N$ , we choose an alternative  $x^n$  to measure. Let  $\varepsilon^{n+1}$  be the measurement error, which we assume is normally distributed with mean 0 and a finite known variance  $(\sigma^\varepsilon)^2$  that is the same across all alternatives. We also assume that errors are independent of each other and of the random vector  $Y$ . Then define  $\hat{y}^{n+1} = Y_x + \varepsilon^{n+1}$  to be the measurement value observed. At time  $N$ , we choose an implementation decision  $x^N$  based on the measurements recorded, and we

receive an implementation reward  $\hat{y}^{N+1}$ . We assume that the reward is unbiased, so that  $\hat{y}^{N+1}$  satisfies  $\mathbb{E}[\hat{y}^{N+1}|Y, x^N] = Y_{x^N}$ . Define the filtration  $(\mathcal{F}^n)_{n=0}^N$  by letting  $\mathcal{F}^n$  be the sigma-algebra generated by  $x^0, \hat{y}^1, x^1, \dots, x^{n-1}, \hat{y}^n$ . We will use the notation  $\mathbb{E}_n[\cdot]$  to indicate  $\mathbb{E}[\cdot | \mathcal{F}^n]$ , the conditional expectation taken with respect to  $\mathcal{F}^n$ . Measurement and implementation decisions  $x^n$  are restricted to be  $\mathcal{F}^n$ -measurable so that decisions may depend only on measurements observed and decisions made in the past.

Let  $\mu^0 := \mathbb{E}[Y]$  and  $\Sigma^0 := \text{Cov}[Y]$  be the mean and covariance of the predictive distribution for  $Y$  so that  $Y$  has prior predictive distribution  $\mathcal{N}(\mu^0, \Sigma^0)$  and  $\Sigma^0$  is a diagonal covariance matrix. Note that our assumed integrability of  $\max_x |Y_x|$  is equivalent to assuming integrability of every  $Y_x$  because  $|Y_{x'}| \leq \max_x |Y_x|$  and  $\max_x |Y_x| \leq |Y_1| + \dots + |Y_M|$ , which is equivalent to assuming  $\Sigma_{xx}^0$  finite for every  $x$ .

We will use the Bayes rule to form a sequence of posterior predictive distributions for  $Y$  from this prior and the successive measurements. Let  $\mu^n := \mathbb{E}_n[Y]$  be the mean vector and  $\Sigma^n := \text{Cov}[Y | \mathcal{F}^n]$  the covariance matrix of the predictive distribution after  $n$  measurements have been made. Because the error term  $\varepsilon^{n+1}$  is independent and normally distributed, the predictive distribution for  $Y$  will remain normal with independent components, and  $\Sigma^n$  will be diagonal almost surely. We write  $(\sigma_x^n)^2$  to refer to the diagonal component  $\Sigma_{xx}^n$  of the covariance matrix. Then  $Y_x \sim \mathcal{N}(\mu_x^n, (\sigma_x^n)^2)$  conditionally on  $\mathcal{F}^n$ . We will also write  $\beta_x^n := (\sigma_x^n)^{-2}$  to refer to the precision of the predictive distribution for  $Y_x$ ,  $\beta^n := (\beta_1^n, \dots, \beta_M^n)$  to refer to the vector of precisions, and  $\beta^\varepsilon := (\sigma^\varepsilon)^{-2}$  to refer to the measurement precision. Note that  $\sigma^\varepsilon < \infty$  implies  $\beta^\varepsilon > 0$ .

Our goal will be to choose the measurement policy  $(x^0, \dots, x^{N-1})$  and implementation decision  $x^N$  that maximizes  $\mathbb{E}[Y_{x^N}]$ . The implementation decision  $x^N$  that maximizes  $\mathbb{E}_N[Y_{x^N}] = \mu_x^N$  is any element of  $\arg \max_x \mu_x^N$ , and the value achieved is  $\max_x \mu_x^N$ . Thus, letting  $\Pi$  be the set of measurement strategies  $\pi = (x^0, \dots, x^{N-1})$  adapted to the filtration, we may write our problem's objective function as

$$(1) \quad \sup_{\pi \in \Pi} \mathbb{E}^\pi \left[ \max_x \mu_x^N \right].$$

**3.2. State space and transition function.** Our state space is the space of all possible predictive distributions for  $Y$ . It can be shown by induction that these are all multivariate normal with independent components. We formally define the state space  $\mathbb{S}$  by  $\mathbb{S} := \mathbb{R}^M \times (0, \infty]^M$ , and it consists of points  $s = (\mu, \beta)$  where, for each  $x \in \{1, \dots, M\}$ ,  $\mu_x$  and  $\beta_x$  are, respectively, the mean and precision of a normal distribution. We will write  $S^n := (\mu^n, \beta^n)$  to refer to the state at time  $n$ . The notation  $S^n$  will refer to a random variable, while  $s$  will refer to a fixed point in the state space.

Fix a time  $n$ . We use the Bayes rule to update the predictive distribution of  $Y_x$  conditioned on  $\mathcal{F}^n$  to reflect the observation  $\hat{y}^{n+1} = Y_x + \varepsilon^{n+1}$ , obtaining a posterior predictive distribution conditioned on  $\mathcal{F}^{n+1}$ . Since  $\varepsilon^{n+1}$  is an independent normal random variable and the family of normal distributions is closed under sampling, the posterior predictive distribution is also normal. Thus our posterior predictive distribution for  $Y_x$  is  $\mathcal{N}(\mu_x^{n+1}, 1/\beta_x^{n+1})$ , and writing it as a function of the prior and the observation reduces to writing  $\mu^{n+1}$  and  $\beta^{n+1}$  as functions of  $\mu^n$ ,  $\beta^n$ , and  $\hat{y}^{n+1}$ . The Bayes rule tells us that these functions are

$$(2) \quad \mu_x^{n+1} = \begin{cases} [\beta_x^n \mu_x^n + \beta^\varepsilon \hat{y}^{n+1}] / \beta_x^{n+1} & \text{if } x^n = x, \\ \mu_x^n & \text{otherwise,} \end{cases}$$

$$(3) \quad \beta_x^{n+1} = \begin{cases} \beta_x^n + \beta^\epsilon & \text{if } x^n = x, \\ \beta_x^n & \text{otherwise.} \end{cases}$$

Conditionally on  $\mathcal{F}^n$ , the random variable  $\mu^{n+1}$  has a multivariate normal distribution whose mean and variance we can compute. First, we use the tower property of conditional expectation and the definitions of  $\mu^n$  and  $\mu^{n+1}$  as the predictive means of  $Y$  given  $\mathcal{F}^n$  and  $\mathcal{F}^{n+1}$ , respectively, to write  $\mathbb{E}_n[\mu^{n+1}] = \mathbb{E}_n[\mathbb{E}_{n+1}[Y]] = \mathbb{E}_n[Y] = \mu^n$ . Then we compute the variance of  $\mu^{n+1}$  componentwise. For those alternatives  $x \neq x^n$  that we do not measure, our posterior is equal to our prior and  $\mu^{n+1} = \mu^n$ . This shows that  $\text{Var}[\mu_x^{n+1} | \mathcal{F}^n] = 0$  if  $x \neq x^n$ . For  $x = x^n$  this variance is generally positive. Let us define

$$(4) \quad \tilde{\sigma}_x^n := \sqrt{\text{Var}[\mu_x^{n+1} | \mathcal{F}^n, x^n = x]},$$

so that  $(\tilde{\sigma}_x^n)^2$  is equal to  $\text{Var}[\mu_x^{n+1} | \mathcal{F}^n, x^n = x]$ . This variance may be interpreted as the variance of the *change* in the predictive mean  $\mu_x^{n+1} - \mu_x^n$  caused by a measurement as  $\text{Var}[\mu_x^{n+1} | \mathcal{F}^n, x^n = x] = \text{Var}[\mu_x^{n+1} - \mu_x^n | \mathcal{F}^n, x^n = x]$ . As shown in the following proposition, it is also equal to the reduction in predictive variance, i.e., the reduction in “uncertainty,” caused by a measurement.

**PROPOSITION 3.1.** *For every  $x = 1, \dots, M$ , we have  $(\tilde{\sigma}_x^n)^2 = (\sigma_x^n)^2 - (\sigma_x^{n+1})^2$ .*

*Proof.* We begin with the relation

$$(\mu_x^{n+1} - Y_x) = (\mu_x^{n+1} - \mu_x^n) + (\mu_x^n - Y_x).$$

Squaring both sides, taking the expectation with respect to  $\mathcal{F}^{n+1}$ , and noting that  $(\sigma_x^{n+1})^2 = \mathbb{E}_{n+1}[(Y_x - \mu_x^{n+1})^2]$  gives

$$\begin{aligned} (\sigma_x^{n+1})^2 &= \mathbb{E}_{n+1}[(\mu_x^n - Y_x)^2] \\ &\quad + 2\mathbb{E}_{n+1}[(\mu_x^n - Y_x)(\mu_x^{n+1} - \mu_x^n)] + \mathbb{E}_{n+1}[(\mu_x^{n+1} - \mu_x^n)^2] \\ &= \mathbb{E}_{n+1}[(\mu_x^n - Y_x)^2] + 2(\mu_x^n - \mu_x^{n+1})(\mu_x^{n+1} - \mu_x^n) + (\mu_x^{n+1} - \mu_x^n)^2 \\ &= \mathbb{E}_{n+1}[(\mu_x^n - Y_x)^2] - (\mu_x^{n+1} - \mu_x^n)^2. \end{aligned}$$

Since  $\sigma_x^{n+1} \in \mathcal{F}^n$ , we may take the expectation with respect to  $\mathcal{F}^n$  to get

$$\begin{aligned} (\sigma_x^{n+1})^2 &= \mathbb{E}_n[\mathbb{E}_{n+1}[(\mu_x^n - Y_x)^2]] - \mathbb{E}_n[(\mu_x^{n+1} - \mu_x^n)^2] \\ &= \mathbb{E}_n[(\mu_x^n - Y_x)^2] - \mathbb{E}_n[(\mu_x^{n+1} - \mu_x^n)^2] \\ &= (\sigma_x^n)^2 - (\tilde{\sigma}_x^n)^2. \quad \square \end{aligned}$$

To more easily compute  $\tilde{\sigma}_x^n$ , define a function  $\tilde{\sigma} : (0, \infty] \mapsto [0, \infty)$  by

$$(5) \quad \tilde{\sigma}(\beta_x) = \sqrt{(\beta_x)^{-1} - (\beta_x + \beta^\epsilon)^{-1}}.$$

Then we have that  $\tilde{\sigma}_x^n = \tilde{\sigma}(\beta_x^n)$  by Proposition 3.1 applied to the identities  $(\sigma_x^{n+1})^2 = (\beta_x^{n+1})^{-1} = (\beta_x^n + \beta^\epsilon)^{-1}$  and  $(\sigma_x^n)^2 = (\beta_x^n)^{-1}$ .

**Remark 3.1.** For  $\beta_x \in (0, \infty)$ , we have that  $(\tilde{\sigma}(\beta_x))^2 = \beta^\epsilon / [(\beta_x + \beta^\epsilon)\beta_x]$  is strictly decreasing in  $\beta_x$ , and thus so is  $\tilde{\sigma}(\beta_x)$ .

Since  $\mu_{x^n}^{n+1}$  is a normal random variable with conditional mean  $\mu_{x^n}^n$  and conditional variance  $(\tilde{\sigma}(\beta_{x^n}^n))^2$  under  $\mathcal{F}^n$ , we can write in terms of an  $\mathcal{F}^n$  adapted sequence  $Z^1, \dots, Z^N$  of standard normal random variables,

$$(6) \quad \mu^{n+1} = \mu^n + \tilde{\sigma}(\beta_{x^n}^n) Z^{n+1} e_{x^n},$$

$$(7) \quad \beta^{n+1} = \beta^n + \beta^\epsilon e_{x^n},$$

where  $e_x$  is a vector in  $\mathbb{R}^M$  with all components zero except for component  $x$ , which is equal to 1. We also define a function  $T : \mathbb{S} \times \{1, \dots, M\} \times \mathbb{R} \mapsto \mathbb{S}$  by

$$(8) \quad T((\mu, \beta), x, z) := (\mu + \tilde{\sigma}(\beta_x) z e_x, \beta + \beta^\epsilon e_x),$$

so that  $S^{n+1} = T(S^n, x^n, Z^{n+1})$ . This is our transition function.

We briefly recall and summarize the random variables which play a role in the measurement process. The underlying and unknown value of alternative  $x$  is denoted  $Y_x$  and is randomly fixed at the beginning of the measurement process. At time  $n$ ,  $\mu_x^n$  is our best estimate of  $Y_x$ , and  $\beta_x^n$  is the precision with which we make this estimate. The result of our time  $n$  measurement causes us to update this estimate to  $\mu_x^{n+1}$ , which we now know with precision  $\beta_x^{n+1}$ . This change from  $\mu_x^n$  to  $\mu_x^{n+1}$  is random, and furthermore is normally distributed with mean 0 and standard deviation  $\tilde{\sigma}(\beta_x^n)$  when we measure alternative  $x$ .

One may think of  $Y_x$  as fixed and of  $\mu_x^n$  as converging toward  $Y_x$  while  $\beta_x^n$  converges to infinity under some appropriately exploratory sampling strategy. It is also appropriate, however, to fix  $\mu_x^n$  and  $\beta_x^n$  (this is the essential content of conditioning on  $\mathcal{F}^n$ ) and think of  $Y_x$  as an unknown quantity. From this viewpoint,  $Y_x$  is random and, furthermore, is normally distributed with predictive mean  $\mu_x^n$  and precision  $\beta_x^n$ . This randomness does not imply that  $Y_x$  must be chosen again according to the predictive normal distribution, but instead the predictive normal distribution only quantifies our uncertain knowledge of the value  $Y_x$  adopted when it was first chosen.

**3.3. Dynamic program.** We apply a dynamic programming approach to our problem. In this approach, the value function is defined as the value of the optimal policy given a particular state  $S^n$  at a particular time  $n$ , and may also be determined recursively through Bellman's equation. If the value function can be computed efficiently, the optimal policy may then also be computed from it. Although in this problem the "curse of dimensionality" makes direct computation of the value function difficult even for  $M$  as small as 3, the dynamic programming principle still provides a valuable method for studying the problem.

The terminal value function  $V^N : \mathbb{S} \mapsto \mathbb{R}$  is given by (1) as

$$(9) \quad V^N(s) := \max_{x \in \{1, \dots, M\}} \mu_x \quad \text{for every } s = (\mu, \beta) \in \mathbb{S}.$$

The dynamic programming principle tells us that the value function at any other time  $0 \leq n < N$  is given recursively by

$$(10) \quad V^n(s) = \max_{x \in \{1, \dots, M\}} \mathbb{E} [V^{n+1}(T(s, x, Z^{n+1}))], \quad s \in \mathbb{S}.$$

We define the Q-factors,  $Q^n : \mathbb{S} \times \{1, \dots, M\} \mapsto \mathbb{R}$ , as

$$(11) \quad Q^n(s, x) := \mathbb{E} [V^{n+1}(T(s, x, Z^{n+1}))], \quad s \in \mathbb{S},$$

and the dynamic programming principle tells us that any policy whose measurement decisions satisfy

$$(12) \quad X^{*n}(s) \in \arg \max_{x \in \{1, \dots, M\}} Q^n(s, x), \quad s \in \mathbb{S},$$

is optimal. Finally, we define the value of a measurement policy  $\pi \in \Pi$  as

$$(13) \quad V^{n, \pi}(s) := \mathbb{E}^\pi [V^N(S^N) \mid S^n = s], \quad s \in \mathbb{S}.$$

This same object might also be thought of as the reward-to-go from state  $s$  at time  $n$  under policy  $\pi$ .

Later we will need several preliminary results concerning the benefit of measurement. First, the following proposition states that, under the optimal policy, it is always better to make a measurement than to measure nothing at all. Here, the value of measuring alternative  $x$  when  $S^n = s$  at time  $n$  is  $Q^n(s, x)$ , and the value of making no measurement is  $V^{n+1}(s)$ . The proof is left until Appendix A.

**PROPOSITION 3.2.**  $Q^n(s, x) \geq V^{n+1}(s)$  for every  $0 \leq n < N$ ,  $s \in \mathbb{S}$ , and  $x \in \{1, \dots, M\}$ .

We see as a corollary to this proposition that the optimal policy will never measure an alternative with zero variance (i.e., with infinite precision) unless all the other alternatives also have zero variance. In other words, there is no value to measuring something that we know perfectly. This is stated precisely in the following corollary.

**COROLLARY 3.1.** Let  $i, j \in \{1, \dots, M\}$ ,  $n < N$ , and  $s = (\mu, \beta) \in \mathbb{S}$ . If  $\beta_j = \infty$ , then  $Q^n(s, i) \geq Q^n(s, j)$ .

*Proof.* Since  $\tilde{\sigma}(\beta_j) = \tilde{\sigma}(\infty) = 0$  and  $\beta_j + \beta^\epsilon = \beta_j$ ,

$$T(s, j, Z^{n+1}) = (\mu + \tilde{\sigma}(\beta_j)Z^{n+1}e_j, \beta + \beta^\epsilon e_j) = (\mu, \beta) = s.$$

Then, by Proposition 3.2,

$$Q^n(s, j) = \mathbb{E} [V^{n+1}(T(s, j, Z^{n+1}))] = V^{n+1}(s) \leq Q^n(s, i). \quad \square$$

We also have a second corollary to the proposition. Proposition 3.2 allowed arbitrarily specifying the alternative to which the extra measurement would be applied, while this corollary points out that the extra measurement may be made according to the optimal policy, in which case  $Q^n(s, x)$  is equal to  $V^n(s)$ . We will use this corollary in section 6 to bound the suboptimality of KG.

**COROLLARY 3.2.**  $V^{n+1}(s) \leq V^n(s)$  for all states  $s \in \mathbb{S}$ .

*Proof.* In Proposition 3.2, take the extra measurement  $x$  to be the measurement made by the optimal policy in state  $s$ .  $\square$

Let us say that a policy  $\pi$  is *stationary* if  $X^{\pi, n}(s) = X^{\pi, 0}(s)$  for all  $s \in \mathbb{S}$  and all  $n = 1, \dots, N - 1$ . In this case we denote  $X^{\pi, n}$  simply by  $X^\pi$ . Corollary 3.2 showed that the value of the optimal policy increases as more measurements are allowed, and we will see in Theorem 3.1 below that this monotonicity also holds for stationary policies.

**THEOREM 3.1.**  $V^{\pi, n}(s) \geq V^{\pi, n+1}(s)$  for every stationary policy  $\pi$  and every state  $s \in \mathbb{S}$ .

The proof is left until Appendix A. We will need this theorem when showing both asymptotic optimality and bounded suboptimality of KG.

**4. The knowledge-gradient policy.** In our problem, the entire reward is received after the final measurement. We may formulate an equivalent problem in which the reward is given in pieces over time, but the total reward given is identical. We define the KG policy as that policy which maximizes the single period reward under this alternate formulation. We will see later that this KG policy is optimal in several cases and has bounded suboptimality in all others. This policy was first introduced in [16] under the name of the  $(R_1, \dots, R_1)$  policy.

**4.1. Definition.** The problem given by (1) has a terminal reward  $V^N(S^N) := \max_x \mu_x^N$ , but no rewards at any other times. We restructure these rewards by writing  $V^N(S^N)$  as a telescoping sequence,

$$\max_x \mu_x^N = [V^N(S^N) - V^N(S^{N-1})] + \dots + [V^N(S^{n+1}) - V^N(S^n)] + V^N(S^n).$$

Thus, the problem that provides single period reward  $V^N(S^n)$  at time  $n$  and  $V^N(S^k) - V^N(S^{k-1})$  at times  $k = n + 1, \dots, N$  is equivalent to problem (1) because the total reward provided is the same in each case. The KG policy  $\pi^{KG}$  is defined as the policy that chooses its measurements to maximize the expectation of the single period reward provided under this alternate formulation,  $\mathbb{E}_n [V^N(T(S^n, x, Z^{n+1})) - V^N(S^n)]$ . Since the  $(Z^n)_{n=1}^N$  are independent and identically distributed normal random variables, we may take  $Z$  to be a generic standard normal random variable and write the decision function of the KG policy  $X^{KG} : \mathbb{S} \mapsto \{1, \dots, M\}$  as

$$(14) \quad X^{KG}(s) \in \arg \max_{x \in \{1, \dots, M\}} \mathbb{E} [V^N(T(s, x, Z)) - V^N(s)] \quad \text{for every } s \in \mathbb{S},$$

where ties in the  $\arg \max$  are broken by choosing the alternative with the smaller index. Note that KG is stationary in time so we drop the time index  $n$  when we write  $X^{KG}$ . Since  $V^N(s)$  does not depend on  $x$ , the KG policy may be rewritten as

$$(15) \quad X^{KG}(s) \in \arg \max_{x \in \{1, \dots, M\}} \mathbb{E} [V^N(T(s, x, Z))] = \arg \max_{x \in \{1, \dots, M\}} Q^{N-1}(s, x).$$

*Remark 4.1.* As noted in [16], KG is optimal by construction when  $N = 1$ . This is because  $V^{N-1} = V^{KG, N-1}$  by (12) and (15), where  $V^{KG, n}$  denotes the value of the KG policy at time  $n$  and is defined according to (13) with the policy  $\pi$  fixed to KG.

If we think of  $V^N(\cdot)$  as a utility function, or as a measure of the amount of “knowledge” contained in a state, we see from (14) that the KG policy chooses its decisions in the direction of steepest expected ascent of this measure. This is the reason behind the name *knowledge-gradient policy*. One may also view it as a single-step look-ahead policy.

**4.2. Computation.** It was already known in [16] that an exact and computationally tractable expression exists for  $X^{KG}$ . We present it here.

For each  $x \in \{1, \dots, M\}$  define a function  $\zeta_x : \mathbb{S} \mapsto [0, \infty)$  by

$$(16) \quad \zeta_x(s) := - \left| \frac{\mu_x - \max_{x' \neq x} \mu_{x'}}{\tilde{\sigma}(\beta_x)} \right|.$$

Except for the sign,  $\zeta_x(S^n)$  is the minimum distance, in terms of the number of standard deviations  $\tilde{\sigma}(\beta_x^n)$ , that a measurement of alternative  $x$  must alter  $\mu_x^{n+1}$  from its premeasurement value of  $\mu_x^n$  to make  $\arg \max_{x'} \mu_{x'}^{n+1} \neq \arg \max_{x'} \mu_{x'}^n$ —that is, to

change the identity of the alternative with the largest conditional expected value. In addition, define the function  $f : \mathbb{R} \mapsto \mathbb{R}$  as

$$(17) \quad f(z) := z\Phi(z) + \varphi(z),$$

where  $\Phi(z)$  is the normal cumulative distribution function and  $\varphi(z)$  is the normal probability density function. Then the following theorem provides an efficient way to compute KG's decisions. The proof may be found in Appendix A.

**THEOREM 4.1.** *For every  $s = (\mu, \beta) \in \mathbb{S}$ , we have*

$$(18) \quad Q^{N-1}(s, x) = \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)),$$

$$(19) \quad X^{KG}(s) \in \arg \max_{x \in \{1, \dots, M\}} \tilde{\sigma}(\beta_x) f(\zeta_x(s))$$

with ties broken by choosing the alternative with the smallest index.

The term  $Q^{N-1}(s, x) - \max_{x'} \mu_{x'} = \tilde{\sigma}(\beta_x) f(\zeta_x(s))$  is in some sense the expected value of the information that would be obtained by measuring alternative  $x$  and is sometimes called the “expected value of information,” or EVI, e.g., in [12] and [10].

Computation of the KG policy via (19) scales linearly with the number of alternatives  $M$ . This compares well with other policies that might be used on this problem. To compute the KG policy at time  $n$ , we must first find the largest and second largest  $\mu_x^n$  across all alternatives  $x$ , which will be used to compute  $\zeta_x^n := \zeta_x(S^n)$ . This may be implemented either by an initial pass through the alternatives at each time period, or by storing and updating the two values across time periods. Once we have the largest and second largest  $\mu_x^n$ , we iterate through the alternatives, calculating  $\tilde{\sigma}(\beta_x^n) f(\zeta_x^n)$  for each one and returning the alternative with the largest value for this expression. This iteration may be streamlined by recomputing the expression only for those alternatives that changed  $\zeta_x^n$  or  $\beta_x^n$  from the previous iteration.

The following remark, which is an easily obtained consequence of Theorems 1 and 2 in [16] and may also be obtained directly from (18), may also be used to accelerate the computation of the KG policy by eliminating some alternatives from consideration. It is also useful for proving later results. It states that if an alternative dominates another in both mean and variance, then of the two, KG prefers the dominating alternative.

**Remark 4.2.** For every  $s = (\mu, \beta) \in \mathbb{S}$  such that  $\mu_j \geq \mu_i$  and  $\beta_j \leq \beta_i$  we have  $Q^{N-1}(s, j) \geq Q^{N-1}(s, i)$ .

Finally, during computation, we may also use the following remark to eliminate some alternatives from consideration, again improving the speed with which we may compute the KG policy.

**Remark 4.3.** Take  $n = N - 1$  in Corollary 3.1. If  $\beta_j = \infty$  for some  $j \in \{1, \dots, M\}$  (that is, if the predictive distribution  $\mathcal{N}(\mu_j, 1/\beta_j)$  for  $Y_j$  is a point mass), then  $Q^{N-1}(S, i) \geq Q^{N-1}(S, j)$  for every  $i \in \{1, \dots, M\}$ .

Thus, KG will never measure an alternative with zero variance unless every alternative has zero variance. Corollary 3.1 shows that the optimal policy shares this behavior of preferring not to measure any alternative whose true value is known perfectly.

**4.3. Behavior.** KG balances two considerations when it chooses its measurement decisions. First, it prefers to measure those alternatives about which comparatively little is known. These alternatives  $x$  are the ones whose predictive distributions

have large variance  $(\sigma_x^n)^2$  or, equivalently, have small precision  $\beta_x^n$ . Thus, we have that if KG prefers to measure some alternative  $i$  over another alternative  $j$ , then it would still prefer to measure alternative  $i$  over  $j$  if the predictive variance of  $i$  were increased.

Second, KG prefers to measure alternatives  $x$  with  $|\mu_x^n - \max_{x' \neq x} \mu_{x'}^n|$  close to 0. We call  $-|\mu_x^n - \max_{x' \neq x} \mu_{x'}^n|$  the *unnormalized influence* and  $\zeta_x^n = -|\mu_x^n - \max_{x' \neq x} \mu_{x'}^n|/\tilde{\sigma}(\beta_x^n)$  the *normalized influence*, or simply the *influence*, of alternative  $x$ , where  $\tilde{\sigma}(\beta_x^n)$  is understood as a normalization term because predictions for different alternatives have different variances and comparison does not make sense unless we standardize these differences. Measurements of alternatives with large influence are more likely to cause a change in the optimal implementation decision; that is, to cause  $\arg \max_{x'} \mu_{x'}^n \neq \arg \max_{x'} \mu_{x'}^{n+1}$ . KG's preference for small predictive precision and large influence are formalized in Propositions 4.1 and 4.2, but first we calculate the derivative of  $f$ , as defined in (17), in a lemma.

LEMMA 4.1. *We have  $f'(z) = \Phi(z) \geq 0$  for every  $z \in \mathbb{R}$ .*

*Proof.* First note that  $\frac{d}{dz} e^{-z^2/2} = -ze^{-z^2/2}$ , showing that  $\varphi'(z) = -z\varphi(z)$ . From this we see that  $f$  has nonnegative derivative  $f'(z) = \Phi(z) + z\varphi(z) - z\varphi(z) = \Phi(z)$ , which completes the proof.  $\square$

PROPOSITION 4.1. *Let states  $s = (\mu, \beta) \in \mathbb{S}$ ,  $s' = (\mu', \beta') \in \mathbb{S}$  and alternatives  $i, j \in \{1, \dots, M\}$  satisfy the following criteria:  $\zeta_i(s') > \zeta_i(s)$ ,  $\zeta_j(s') = \zeta_j(s)$ ,  $\beta'_i < \beta_i$ , and  $\beta'_j = \beta_j$ . If  $Q^{N-1}(s, i) > Q^{N-1}(s, j)$ , then  $Q^{N-1}(s', i) > Q^{N-1}(s', j)$ .*

*Proof.* First,  $\tilde{\sigma}(\beta'_i) \geq \tilde{\sigma}(\beta_i)$  by Remark 3.1 and  $f(\zeta_i(s')) \geq f(\zeta_i(s))$  by Lemma 4.1. By (18),  $Q^{N-1}(s', i) > Q^{N-1}(s, i)$ . Also, the equalities  $\tilde{\sigma}(\beta'_j) = \tilde{\sigma}(\beta_j)$  and  $f(\zeta_j(s')) = f(\zeta_j(s))$  imply through (18) that  $Q^{N-1}(s', j) = Q^{N-1}(s, j)$ . Thus, if  $Q^{N-1}(s, i) > Q^{N-1}(s, j)$ , then  $Q^{N-1}(s', i) \geq Q^{N-1}(s, i) > Q^{N-1}(s, j) = Q^{N-1}(s', j)$ .  $\square$

PROPOSITION 4.2. *If alternative  $i$  and state  $s = (\mu, \beta)$  are such that  $\zeta_i(s) \geq \zeta_j(s)$  and  $\beta_i < \beta_j$  for every alternative  $j \neq i$ , then  $X^{KG}(s) = i$ .*

*Proof.* Let  $j$  be an alternative different from  $i$ . Then  $\tilde{\sigma}(\beta_i) > \tilde{\sigma}(\beta_j)$  by Remark 3.1 and  $f(\zeta_i(s)) \geq f(\zeta_j(s))$  by Lemma 4.1. This implies that  $Q^{N-1}(s, i) > Q^{N-1}(s, j)$  by Proposition 4.1. Since this is true for all  $j \neq i$ , we have that  $i = \arg \max_j Q^{N-1}(s, j) = X^{KG}(s)$  where the arg max is unique.  $\square$

It is also interesting to note that increasing the predictive mean of a single alternative usually, but not universally, encourages KG to measure it. Thus, having a large predictive mean is similar, but not identical, to having a large unnormalized influence. We formalize this in the following proposition.

PROPOSITION 4.3. *If KG prefers alternative  $i$  in state  $(\mu, \beta)$ , then it also prefers the same alternative  $i$  in state  $(\mu + a e_i, \beta)$  for all positive real numbers  $a$  such that  $\mu_i + a \leq \max_x \mu_x$ , i.e., for  $0 \leq a \leq -\mu_i + \max_x \mu_x$ .*

We leave the proof until Appendix A.

**5. Asymptotic optimality.** In this section we show that the KG policy is asymptotically optimal in the limit as the number of measurements  $N$  grows large. This means that, given the opportunity to measure infinitely often, KG will discover which alternative is best. In some sense, this is a convergence result because it shows that the policy's estimate of which alternative is best will converge to the alternative that is truly best.

The KG policy is not alone in possessing this property. Indeed, the following well-known policies are all asymptotically optimal: the equal-allocation policy which distributes its measurements in a round-robin fashion equally among the alternatives; the uniform exploration policy which randomly chooses its measurements with equal



probability across the alternatives; and the Boltzmann exploration policy discussed in section 8 which randomly chooses its measurements according to exponentially weighted probabilities.

These policies differ from KG in that they explore for exploration's sake and for the long-term benefit it provides, while KG is purely myopic. Moreover, we argue that KG's asymptotic optimality is notable exactly because the policy is entirely myopic, maximizing its single-period expected reward without regard for the long-term. This is not generally the case with myopic policies for other problems. That a myopic policy is also optimal in the long-term shows that this ranking and selection problem has a special structure, and it foreshadows what is further suggested by our numerical experiments: that this myopic policy, KG, performs quite well in many cases which are neither myopic nor asymptotic.

In addition, one policy, interval estimation, performs very well in our numerical experiments but is not asymptotically optimal as in some cases it "sticks," measuring one alternative only and obtaining its true value perfectly without learning about the others [19]. Indeed, one can construct cases in which this policy's performance is arbitrarily bad compared to any asymptotically optimal policy. Although a policy's asymptotic optimality is not evidence of quality by itself, its absence should raise concern among those who might use a policy lacking it. Finally, a natural question is whether other policies, such as those in the OCBA family and those proposed in [12], are asymptotically optimal. This question is currently open as these other policies are more complex and require more care during analysis than does KG. Nevertheless, we believe that the proof techniques applied here may be extended to show that many other Bayesian look-ahead policies are also asymptotically optimal.

To show that KG is asymptotically optimal, we begin by showing in Proposition 5.1 that the asymptotic value of a policy is well defined and bounded above by the value  $\mathbb{E} \max_x Y_x$  of learning every alternative exactly. Then we show in Proposition 5.2 that this value is achieved by any stationary policy that measures every alternative infinitely often. Thus, any stationary policy that samples every alternative infinitely often is asymptotically optimal. Finally, we show in Theorem 5.1 that KG is asymptotically optimal. The proof centers on the notion that, as the number of times an alternative is measured increases, the variance of the value of that alternative shrinks toward 0. Eventually, that variance will be so low that KG will prefer to measure another alternative. This argument is used to show that KG samples every alternative infinitely often and thus is asymptotically optimal.

Since we will be varying the number  $N$  of measurements allowed, we use the notation  $V^0(\cdot; N)$  to denote the value function at time 0 when the problem's terminal time is  $N$ . We then define the *asymptotic value function*  $V(\cdot; \infty)$  by the limit  $V(s; \infty) := \lim_{N \rightarrow \infty} V^0(s; N)$  for  $s \in \mathbb{S}$ . Similarly, we denote the *asymptotic value function for stationary policy*  $\pi$  by  $V^\pi(\cdot; \infty)$  and define it by  $V^\pi(s; \infty) := \lim_{N \rightarrow \infty} V^{\pi,0}(s; N)$  for  $s \in \mathbb{S}$ . Proposition 5.1 shows that both limits exist.

If  $V^\pi(s; \infty)$  is equal to  $V(s; \infty)$  for every  $s \in \mathbb{S}$ , then  $\pi$  is said to be *asymptotically optimal*. In particular, if a stationary policy  $\pi$  achieves the upper bound  $U(\cdot)$  on  $V(\cdot; \infty)$  shown in Proposition 5.1, then  $\pi$  must be asymptotically optimal. We will use this later to show that KG is asymptotically optimal. The proof of Proposition 5.1 is found in Appendix A.

**PROPOSITION 5.1.** *Let  $s \in \mathbb{S}$ . Then the limit  $V(s; \infty)$  exists and is bounded above by*

$$(20) \quad U(s) := \mathbb{E} \left[ \max_x Y_x \mid S^0 = s \right] < \infty,$$

where we recall that  $\{Y_x\}_{x \in \{1, \dots, M\}}$  are independent and  $Y_x \sim \mathcal{N}(\mu_x^0, (\beta_x^0)^{-1})$ . Furthermore,  $V^\pi(s; \infty)$  exists and is finite for every stationary policy  $\pi$ .

For any finite terminal time  $N$  we define the random variable  $\eta_x^N$  as the number of times that alternative  $x$  is measured up to but not including the terminal time  $N$ . We also define  $\eta_x^\infty$  as the limit of the  $\eta_x^N$ ; namely,

$$\eta_x^N := \sum_{k=1}^N 1_{\{x^k=x\}} \quad \text{and} \quad \eta_x^\infty := \lim_{N \rightarrow \infty} \eta_x^N.$$

The limit  $\eta_x^\infty$  exists because  $\eta_x^N$  is nondecreasing in  $N$  a.s. Note that we allow the limit  $\eta_x^\infty$  to be infinite.

Proposition 5.2 formalizes the idea that if we measure every alternative infinitely often, then we eventually learn the true value of every alternative. This implies asymptotic optimality. We then use Proposition 5.2 in the proof of Theorem 5.1 to show that KG is asymptotically optimal. The proofs for both Theorem 5.1 and Proposition 5.2 are found in Appendix A.

**PROPOSITION 5.2.** *If  $\pi$  is a stationary policy under which  $\eta_x^\infty = \infty$  a.s. for every  $x$ , then  $\pi$  is asymptotically optimal.*

**THEOREM 5.1.** *The KG policy is asymptotically optimal and has value  $U(S^0)$ .*

**6. Bound on suboptimality.** We have shown that KG is optimal when  $N = 1$  and in the limit as  $N \rightarrow \infty$ . In this section we address the range of  $N$  between these extremes by bounding KG's suboptimality in this region. This bound will be tight for small  $N$  and will grow as  $N$  increases.

We begin with a theorem that implies our bound as a corollary. This theorem shows that there is a limit on how much we may learn through any single measurement.

**THEOREM 6.1.** *Let  $s = (\mu, \beta) \in \mathbb{S}$  and  $c = (2\pi)^{-1/2} \max_x \tilde{\sigma}(\beta_x)$ . Then*

$$V^n(s) \leq V^{N-1}(s) + c(N - n - 1).$$

The proof is found in Appendix A. We combine this result with Theorem 3.1 to bound KG's suboptimality. Here,  $V^{KG,n}(s)$  is the value of the KG policy at time  $n$  when  $S^n = s$ .

**COROLLARY 6.1.** *Let  $s = (\mu, \beta) \in \mathbb{S}$  and  $c = (2\pi)^{-1/2} \max_x \tilde{\sigma}(\beta_x)$ . Then*

$$V^n(s) - V^{KG,n}(s) \leq c(N - n - 1).$$

*Proof.* By Remark 4.1, we have  $V^{N-1}(s) = V^{KG,N-1}(s)$ . From Theorem 3.1 we have  $V^{KG,N-1}(s) \leq V^{KG,n}(s)$ . Substituting the inequality  $V^{N-1}(s) \leq V^{KG,n}(s)$  into Theorem 6.1 shows the corollary.  $\square$

**7. Optimality for finite horizon special cases.** We saw in Remark 4.1 that KG is optimal when  $N = 1$ . We will show that KG is optimal in two other special cases: first, when there are only two alternatives to measure; second, when the measurements are free from noise,  $(\sigma^\varepsilon)^2 = 0$ , and when the parameters of the time 0 prior can be ordered by  $\mu_1^0 \geq \mu_2^0 \geq \dots \geq \mu_M^0$  and  $\sigma_{11}^0 \geq \sigma_{22}^0 \geq \dots \geq \sigma_{MM}^0$ . Before showing optimality under these conditions, we first define and discuss a property called the KG persistence property. This property is useful because it provides a sufficient condition for optimality.

**7.1. Persistence of the knowledge-gradient policy.** Proofs of the optimality of the KG policy in these special cases is based on the KG persistence property. A problem setting is said to have the KG persistence property if, operating the problem under some policy other than KG, an alternative preferred by KG will remain preferred until the alternative is measured. Below, in Theorem 7.1, we show that if a problem setting has the KG persistence property, then KG is optimal in that problem setting. Before stating this theorem, we formally define the KG persistence property and an associated term, “covering of the future.”

**DEFINITION 7.1.** *A sequence of subsets of  $\mathbb{S}$ ,  $\{\mathbb{S}^n\}_{n=k}^N$ , is called a covering of the future from  $k$  if  $T(s, x, Z^{n+1}) \in \mathbb{S}^{n+1}$  a.s. for every  $s \in \mathbb{S}^n$ ,  $x \in \{1, \dots, M\}$ , and  $n \in \{k, \dots, N-1\}$ .*

**DEFINITION 7.2.** *We say that the KG persistence property holds on a covering  $\{\mathbb{S}^n\}_{n=k}^N$  of the future from  $k$  if  $X^{KG}(T(s, x, Z^{n+1})) = X^{KG}(s)$  a.s. for every  $s \in \mathbb{S}^n$ ,  $x \neq X^{KG}(s)$ , and  $n \in \{k, \dots, N-1\}$ .*

This KG persistence property gives us a sufficient condition for the optimality of the KG policy, as stated in the following theorem.

**THEOREM 7.1.** *If the KG persistence property holds on a covering  $\{\mathbb{S}^n\}_{n=k}^N$  of the future from  $k$  for some  $k \in \{0 \dots N-1\}$ , then  $V^{KG,k}(s) = V^k(s)$  for every  $s \in \mathbb{S}^k$ .*

We leave the proof until Appendix A, but we give a sketch here. Consider a time  $n < N-1$  and the alternative that KG prefers. If the problem setting has the KG persistence property, then, even if we do not measure that alternative now, KG will continue to prefer it until we reach the final measurement  $N-1$ . At this measurement, KG is optimal by construction and so it is now provably optimal to measure this persistent alternative. Thus, there exists an optimal policy that measures the persistent alternative a.s., and by the temporal symmetry in the model, there exists an optimal policy that measures the persistent alternative immediately at time  $n$ . This argument is used with induction to show that there exists an optimal policy making the same measurements as KG.

**7.2. Optimality for two alternatives.** We use the KG persistence principle to show that KG is optimal when there are exactly two alternatives to consider, i.e.,  $M = 2$ . In this case we will see that the optimal policy is one that, at each decision point, measures the alternative with the largest variance. This policy is actually deterministic, and it was shown in [15] that this policy is optimal among the class of deterministic policies. Theorem 7.2 extends this result to show that this same policy is also optimal among the class of fully sequential policies. It is not generally true that the best deterministic policy is also as good as or better than every sequential policy, but Theorem 7.2 shows that this is exactly the case for this particular problem.

We will see that the policy of measuring the alternative with the largest variance is optimal because knowing the correct implementation decision is the same as knowing the true sign of  $Y_1 - Y_2$ . Each measurement measures only one of  $Y_1$  or  $Y_2$ , and an equal reduction in variance for  $Y_1$  or  $Y_2$  contributes equally to the overall reduction in variance of  $Y_1 - Y_2$ , regardless of which expected value is bigger. Thus, the best way to learn about the difference between points  $Y_1 - Y_2$  is to measure that point about which the least is known.

To show that KG is optimal when  $M = 2$ , we need to show that KG persistence holds when  $M = 2$  and then refer to Theorem 7.1.

**LEMMA 7.1.** *If  $M = 2$ , then  $X^{KG}(s) \in \arg \min_x \beta_x$  for each  $s = (\mu, \beta) \in \mathbb{S}$  with ties broken by choosing the alternative with the smaller index.*

*Proof.* By (19) from Theorem 4.1, it is enough to show equality between the sets

$\arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(s))$  and  $\arg \min_x \beta_x$ . When  $M = 2$ ,  $\zeta_x(s) = -|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)$ , so  $\arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(s)) = \arg \max_x \tilde{\sigma}(\beta_x) f(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x))$ . The function  $\tilde{\sigma}$  is strictly decreasing by Remark 3.1. This fact will be used on its own, and it also implies that  $-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)$  is a decreasing function of  $\beta_x$ . The function  $f$  is nondecreasing by Lemma 4.1, so the function  $\beta_x \mapsto f(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x))$  is the composition of a nondecreasing function with a nonincreasing function and is thus itself nonincreasing. Thus, the function  $\beta_x \mapsto \tilde{\sigma}(\beta_x) f(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x))$  is the product of a strictly decreasing function with a nonincreasing function and is thus itself strictly decreasing. This implies that  $\arg \max_x \tilde{\sigma}(\beta_x) f(-|\mu_1 - \mu_2|/\tilde{\sigma}(\beta_x)) = \arg \min_x \beta_x$ .  $\square$

**THEOREM 7.2.** *If  $M = 2$ , then KG is optimal.*

*Proof.* Let  $\mathbb{S}^n = \mathbb{S}$  for all  $n$ , and note that  $\{\mathbb{S}^n\}_{n=0}^N$  is a covering of the future from 0. We will show that the KG persistence property holds on  $\{\mathbb{S}^n\}_{n=0}^N$ .

Let  $n \in \{0, \dots, N-1\}$  and  $s = (\mu, \beta) \in \mathbb{S}$ . First consider the case when  $\beta_1 \leq \beta_2$ . By Lemma 7.1,  $X^{KG}(s) = 1$ . The precision component of  $T(s, 2, Z^{n+1})$  is  $(\beta_1, \beta_2 + \beta^\epsilon)$ . Since  $\beta_1 \leq \beta_2 \leq \beta_2 + \beta^\epsilon$  and by Lemma 7.1,  $X^{KG}(T(s, 2, Z^{n+1})) = 1$  a.s.

Now consider the case when  $\beta_1 > \beta_2$ . By Lemma 7.1,  $X^{KG}(s) = 2$ . The precision component of  $T(s, 1, Z^{n+1})$  is  $(\beta_1 + \beta^\epsilon, \beta_2)$ . Since  $\beta_1 + \beta^\epsilon \geq \beta_1 > \beta_2$  and by Lemma 7.1,  $X^{KG}(T(s, 1, Z^{n+1})) = 2$  a.s.

In both cases,  $x \neq X^{KG}(s)$  implies  $X^{KG}(T(s, x, Z^{n+1})) = X^{KG}(s)$  a.s., so KG persistence holds. Then, by Theorem 7.1,  $V^{KG,0}(s) = V^0(s)$  for every  $s \in \mathbb{S}$ , and KG is optimal.  $\square$

This theorem is founded on the intuition that the policy that learns the most is also the one that changes our beliefs the most. This has a comparison in other measurement problems—for example, the problem in which we have a quadratic function with known second derivative and we measure the first derivative to find the maximum of the function. In this case the optimal policy is also the one that maximizes the variance of the change in our final belief with respect to our current belief. In both cases we measure the change between our current and final beliefs by taking the variance. In other problems the variance is likely not the right measure of change, but the same intuition would apply with some other measure of change.

**7.3. Optimality when the state space is ordered.** The KG policy is also optimal when there is no measurement noise, i.e.,  $(\sigma^\epsilon)^2 = 0$ , and when the components of  $S^0$  may be ordered in such a way that we have  $\mu_1^0 \geq \dots \geq \mu_M^0$  together with  $\beta_1^0 \leq \dots \leq \beta_M^0$ . In other words, the optimality result requires that we may order the alternatives with increasing means while simultaneously ordering them with increasing variances. With the assumption of no measurement noise, the problem is interesting only if the number of alternatives  $M$  is larger than the measurement budget  $N$ .

We present this optimality result formally in the theorem below, but first, as these conditions are particularly restrictive, we motivate them with an example. Consider a problem in marketing research in which we have a collection of potential advertising campaigns, some of which are more ambitious than others. The predictive distributions for the value obtained from the ambitious campaigns have larger mean but larger variance as well. We may test a few of these campaigns in test markets before committing to one of them. We will assume that the number of test markets allowed is smaller than the number of potential campaigns. If we are willing to make two additional assumptions—that loss is linear and that test markets give us perfect knowledge of the campaign's true value—then the example meets the conditions of the theorem. These additional assumptions would not be met perfectly satisfied in reality, but it is not too unreasonable to imagine situations in which loss would be

approximately linear, and in which the knowledge obtained from a test market would be large enough that one would not wish to perform a second test market. With this marketing application as an illustrative example, we expect that this sort of ordering of means and variances may also occur in financial applications, or wherever greater expected reward brings greater risk along with it.

**THEOREM 7.3.** *If  $(\sigma^\varepsilon)^2 = 0$  and  $s = (\mu, \beta) \in \mathbb{S}$  is such that the implication*

$$(\beta_i \neq \infty \text{ and } \beta_j \neq \infty \text{ and } \beta_i < \beta_j) \implies \mu_i \geq \mu_j$$

*holds for every  $i, j \in \{1, \dots, M\}$ , then  $V^0(s) = V^{KG,0}(s)$ .*

The full proof can be found in Appendix A, but the essential idea is that when this ordering holds, the tension between exploration and exploitation is gone, and KG will simply choose that alternative with the largest variance. This is because the alternative with the largest variance is also the alternative with the largest mean among those which are not yet perfectly known. This ordering by variances is persistent, as it was in the  $M = 2$  case. Thus, the KG persistence property holds and KG is optimal.

**8. Computational experiments.** We compared KG against other sampling policies using Monte Carlo simulation on 100 randomly generated problems and found that it performs competitively. In particular, KG performed best when measured by average performance across all the problems, and the margin by which it outperformed the best competing policies in favorable cases was significantly larger than the margin by which it was outperformed in unfavorable cases. Its comparative performance was particularly good when the measurement budget was not much larger than the number of alternatives to measure, and we would argue that performing well in these cases is particularly important.

The space of problems is parameterized by a number of measurements  $N$ , a number of alternatives  $M$ , an initial precision  $\beta^0 \in (0, \infty]^M$ , an initial mean  $\mu^0 \in \mathbb{R}^M$ , and a measurement noise  $(\sigma^\varepsilon)^2 \in [0, \infty)$ . We chose a collection of 100 problems randomly generated within this space according to the following distribution:  $M$  was integer-valued between 2 and 100.  $N$  was chosen by first choosing  $M$  and then choosing a ratio  $N/M$  uniformly from the set  $\{1, 3, 10\}$ . Each  $\mu_x$  was uniformly distributed in the interval  $[-1, 1]$ , and each  $\beta_x$  was independently chosen as 1 with probability .9 and 1000 with probability .1. The noise variance  $(\sigma^\varepsilon)^2$  was set to 1 in all cases.

For each problem, we performed simulations in which true function values were generated independently according to the prior. Rather than collecting the value obtained by the policy in each simulation, we collected the opportunity cost realized, where the opportunity cost is the difference in true value between the best option and the option chosen by the policy. The difference in expected opportunity cost is the same as the difference in policy value, but samples of opportunity cost have less error, and this allowed us to obtain accurate estimates with fewer simulations. We ran  $10^5$  simulations for each policy.

We compared KG against seven other policies: the OCBA for linear loss of [18], the LL(S) policy of [12], the interval estimation (IE) policy of [19], Boltzmann exploration (see, e.g., [28]), equal allocation, and exploitation. Several of these policies required choosing one or more parameters, which we did by simulating several choices on all 100 problems and taking the parameters whose resulting opportunity cost was smallest when summed over all 100 problems. We briefly describe each policy and its tuning.

- **OCBA.** This policy has three parameters: the number of alternatives to allocate to in each stage,  $m$ ; the number of measurements to allocate to each

alternative in the first stage,  $n_0$ ; and the number of measurements per chosen alternative to allocate in each stage,  $\tau$ . We set  $n_0$  to 0 because our prior is informative and thus may be thought of as already providing the results of a first stage. To calibrate  $m$  and  $\tau$ , we ran initial experiments with 5000 samples each with settings of  $m = 1, \tau \in \{1, 2, 5, 10\}$ , and also with  $\tau = 1, m \in \{2, 5, 10\}$ . We found that  $m = 1, \tau = 1$  performed best.

- *LL(S) for known variance.* The LL(S) policy allows normal measurement errors with *unknown* variance and uses a normal-gamma prior for the unknown mean and measurement precision. We adapted this policy to the known-variance case by taking the limit as the gamma prior on the precision becomes a point mass at the known variance. Details can be found in Appendix B. The policy has two parameters,  $n_0$  and  $\tau$ . We set  $n_0$  to 0 as we did with OCBA. We tested the values 1, 2, 3, 4, 5, 10 for  $\tau$  on our collection of 100 problems with 5000 samples for each problem and found that  $\tau = 1$  worked best for every problem. This is the value we used in comparison with KG.
- *Interval estimation.* IE is parameterized by  $z_{\alpha/2}$ . As [19] suggests that values of 2, 2.5, or 3 often work best for  $z_{\alpha/2}$ , we tested values between 2 and 4 in increments of .1 and found that  $z_{\alpha/2} = 3.1$  worked best. Although we found IE worked very well when properly tuned, we also found it to be very sensitive to the choice of tuning parameter.
- *Boltzmann exploration.* Boltzmann exploration chooses its measurements by  $\mathbb{P}\{x^n = x \mid \mathcal{F}^n\} = \frac{\exp(\mu_x^n/T^n)}{\sum_{x'=1}^M \exp(\mu_{x'}^n/T^n)}$ , where the policy is parameterized by a decreasing sequence of “temperature” coefficients  $(T^n)_{n=0}^{N-1}$ . We tuned this temperature sequence within the set of exponentially decreasing sequences defined by  $T^{n+1} = \gamma T^n$  for some constant  $\gamma \in (0, 1]$ . The set of all such sequences is parameterized by  $\gamma$  and  $T^N$ . We tested  $\gamma \in \{.1, .5, .8, .9, 1\}$  with  $T^N \in \{.1, 1, 10\}$  and found that  $\gamma = 1$  performed best. We then tested the set of possible  $T^N$  between .1 and 10 with  $\gamma$  fixed to 1 and found that  $T^N = .55$  performed best.
- *Equal allocation.* The equal-allocation policy is  $x^n \in \arg \min_x \beta_x^n$ , since we think of the prior as providing the results of some previous first-stage measurements, and we interpret  $\beta_x^n/\beta^e$  as the number of measurements of alternative  $x$  taken by time  $n$ . It requires no tuning.
- *Exploitation.* The exploitation policy is  $x^n \in \arg \max_x \mu_x$ . It requires no tuning.

The work required to tune other policies highlights one practical advantage of KG policy: it requires no tuning.

**8.1. Results.** On each of the 100 randomly generated problems, we took  $10^5$  samples of opportunity cost from every policy. The distribution of opportunity cost is not normal, as it is positive a.s. and often equal to 0. We averaged groups of 500 samples to obtain approximately normal samples from which we estimated expected opportunity cost as well as standard errors on these estimates. The difference in value between KG and any other policy on any particular problem was then estimated as the difference in sampled opportunity costs, with standard error equal to the square root of the sum of the squared standard errors. The resulting standard errors of the difference, reporting maximum and averaged values across the 100 problems, were .0018 and .0007 for IE; .0018 and .0007 for OCBA; .0019 and .0007 for LL(S); .0020 and .0009 for Boltzmann exploration; .0024 and .0013 for equal allocation; and .0026 and .0021 for exploitation.

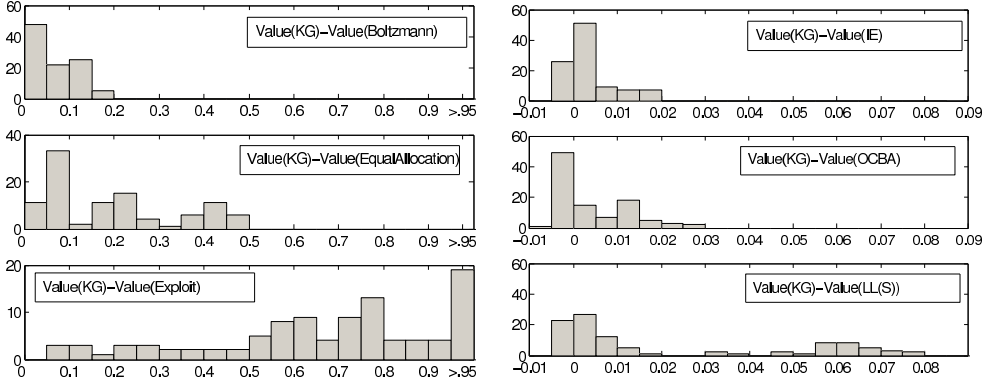


FIG. 1. Histogram of the sampled difference in value for competing policies aggregated across the 100 randomly generated problems.

We show in Figure 1 the sample estimates of  $V^{KG} - V^\pi$  aggregated across the randomly generated problems for each of the competing policies  $\pi$ . Bars to the right of 0 indicate that KG outperformed the plotted policy on those problems, and bars to the left indicate the converse. Note that the scale of the histograms in the right-hand plots is much smaller than in the left-hand plots. The histograms show that Boltzmann exploration, equal allocation, and exploitation policies were all outperformed by KG in every problem setting tested, while IE, OCBA for linear loss, and the LL(S) policy performed relatively better. Each of these three better competing policies performed better than KG on some problems and were outperformed on others; however, the tail to the right of 0 is larger than to the left. This indicates that the amount by which KG outperformed the competing policies was significantly larger than the amount by which it was outperformed.

We note a seeming discrepancy between our numerical work and that of Chick, Branke, and Schmidt [10], who tested a variance-unknown version of the KG policy called  $LL_1$ . They found that  $LL_1$  performed well in small-sample settings, but poorly elsewhere. In contrast, we found that KG, a very similar policy, performed quite well overall. We believe that the difference lies in the stopping rule used. We simply stopped our sampling policies after a fixed horizon  $N$ , but [10] drew many of its conclusions from experiments using the expected opportunity cost Bonferonni (EOC Bonf) stopping rule introduced in [2]. In experiments not pictured here we found that KG also performed poorly with EOC Bonf stopping, but much better when it was stopped using a stopping rule that we introduce now.

This new rule stops as soon as the expected myopic value of the next measurement, as determined by  $Q^{N-1}(s, x) - \max_{x'} \mu_{x'} = \tilde{\sigma}(\beta_x) f(\zeta_x(s))$ , drops below a threshold  $c$ . That is, the number of measurements  $N$  to take under this rule is defined by  $N = \inf\{n \geq 0 : \tilde{\sigma}(\beta_x^n) f(\zeta_x(S^n)) < c\}$ . The threshold  $c$  should be interpreted as the cost of one measurement. Since the expected marginal value of each subsequent measurement decreases on average, it is reasonable to stop measuring as soon as the marginal expected value of the next measurement drops below its cost. Replacing EOC Bonf with this new stopping rule may improve the performance of the KG sampling policy enough to make it competitive with LL(S) and other commonly used policies in an adaptive stopping setting. Our initial experiments suggest that this may be the case, but space limitations prevent a thorough discussion of the experimental issues.

**9. Conclusion.** The KG measurement policy, as first proposed in [16] and as analyzed here, has several attractive features. Under the assumption of independent normally distributed priors with normal sampling errors of common known variance, we showed that the policy is optimal in both extremes of the number of measurements allowed ( $N = 1$  and  $N \rightarrow \infty$ ), as well as in other special cases, and has bounded suboptimality in the remaining cases. We showed numerically that it performs competitively with, or significantly better than, several other sequential measurement policies in a broad class of problem settings. In addition, KG is simple in concept, easy to implement, fast to compute, and requires no tuning. This simplicity may make it an attractive alternative to its more complex but similarly performing cousins, the OCBA and the LL(S) policy.

One important limitation of the version of the policy discussed herein is its assumption of common known variance, which often fails to be met in practice. To lift this assumption, it is possible to place a normal-gamma prior on the unknown means and variances, as was done in [12], and recompute the optimal single-step look ahead policy. Indeed, if we begin with a noninformative normal-gamma prior for the true mean  $Y_x$  and unknown sampling variance  $\beta_x^\epsilon$  of alternative  $x$ , and after sampling have vectors of statistics  $(\mu, \hat{\sigma}^2, n)$  where  $(\mu_x, \hat{\sigma}_x^2, n_x)$  indicate the sample mean, sample variance, and number of samples taken for alternative  $x$ , then a calculation similar to that of Theorem 4.1 reveals that the corresponding KG policy is  $\arg \max_x \tilde{\sigma}_x f_{n_x-1}(\zeta_x)$ , where we must redefine  $\tilde{\sigma}_x := \sqrt{\hat{\sigma}^2/n_x(n_x+1)}$ , leave  $\zeta_x$  defined as before, and define  $f_n(z) := \frac{\nu+z^2}{\nu-1} \varphi_\nu(z) + z\Phi_\nu(z)$ , where  $\varphi_\nu$  and  $\Phi_\nu$  are, respectively, the probability density function and cumulative density function of the student- $t$  distribution with  $\nu$  degrees of freedom. This provides a version of KG for the unknown-variance case. This was derived earlier and independently in [10], and is discussed there in much greater detail, together with a numerical analysis of its properties.

Additionally, the KG policy as described herein has used a fixed number of samples instead of an adaptive stopping rule, while [2] has shown that such rules generally improve the efficiency of budgeted ranking and selection policies. Nevertheless, as implied briefly in section 8 and as discussed in [10], one can certainly use an adaptive stopping rule with the KG sampling policy. Future work is needed to assess the quality of such adaptively stopped policies, and to determine which stopping rules are best to use with KG, but this is by no means an insurmountable obstacle.

Other limitations would seem to present more difficulty. The use of common random numbers has proved immensely beneficial for simulation-based ranking and selection. References [11] and [14] discuss Bayesian ranking and selection policies taking advantage of common random numbers, as does [21] for the frequentist formulation, and it may be possible to extend the KG approach along these lines as well. Indeed, KG's benefits may be overshadowed by its inability to leverage common random numbers in simulation-based ranking and selection unless this extension can be made. In addition, KG assumes the alternatives have a common measurement cost, while in practice it may be more expensive or time consuming to measure some alternatives than others. It may be possible to lift this restriction by dividing the benefit of measurement by the cost so as to obtain a normalized quantity for comparison (a benefit per unit cost), but it may also be that the OCBA approach is more appropriate in such instances.

Despite these limitations, KG has great potential for application. As demonstrated here, it should be considered a reasonable alternative to other measurement policies for those applications that meet its assumptions of a fixed sampling budget



and normally distributed errors with common known variance.

### Appendix A. Proofs.

**Proof of Proposition 3.2.** We proceed by induction on  $n$ . For  $n = N - 1$  and  $s = (\mu, \beta)$  we have

$$\begin{aligned} Q^{N-1}(s, x) &= \mathbb{E} [V^N(T(s, x, Z^N))] = \mathbb{E} \left[ (\mu_x + \tilde{\sigma}(\beta_x) Z^N) \vee \max_{x' \neq x} \mu_{x'} \right] \\ &\geq \mu_x \vee \max_{x' \neq x} \mu_{x'} = V^N(s), \end{aligned}$$

where the inequality is justified by Jensen's inequality and the convexity of the max operator. Now we prove the induction step. For  $0 \leq n < N$ ,

$$\begin{aligned} Q^n(s, x) &= \mathbb{E} [V^{n+1}(T(s, x, Z^{n+1}))] = \mathbb{E} \left[ \max_{x' \in \{1, \dots, M\}} Q^{n+1}(T(s, x, Z^{n+1}), x') \right] \\ &\geq \max_{x' \in \{1, \dots, M\}} \mathbb{E} [Q^{n+1}(T(s, x, Z^{n+1}), x')] \\ (21) \quad &= \max_{x' \in \{1, \dots, M\}} \mathbb{E} [V^{n+2}(T(T(s, x, Z^{n+1}), x'), Z^{n+2})]. \end{aligned}$$

In this equation both decisions  $x$  and  $x'$  are fixed, so the state to which we arrive when we measure  $x$  first and  $x'$  second,  $T(T(s, x, Z^{n+1}), x', Z^{n+2})$ , is equal in distribution to the state to which we arrive when we measure  $x'$  first and  $x$  second,  $T(T(s, x', Z^{n+2}), x, Z^{n+1})$ . This allows us to exchange the time-order of the decisions  $x$  and  $x'$  in (21) to write

$$\begin{aligned} Q^n(s, x) &\geq \max_{x' \in \{1, \dots, M\}} \mathbb{E} [V^{n+2}(T(T(s, x', Z^{n+2}), x, Z^{n+1}))] \\ &= \max_{x' \in \{1, \dots, M\}} \mathbb{E} [\mathbb{E} [V^{n+2}(T(T(s, x', Z^{n+2}), x, Z^{n+1})) \mid Z^{n+2}]] \\ &= \max_{x' \in \{1, \dots, M\}} \mathbb{E} [Q^{n+1}(T(s, x', Z^{n+2}), x)]. \end{aligned}$$

Then the induction hypothesis tells us that

$$Q^{n+1}(T(s, x', Z^{n+2}), x) \geq V^{n+2}(T(s, x', Z^{n+2})) \text{ a.s.,}$$

allowing us to write

$$Q^n(s, x) \geq \max_{x' \in \{1, \dots, M\}} \mathbb{E} [V^{n+2}(T(s, x', Z^{n+2}))] = \max_{x' \in \{1, \dots, M\}} Q^{n+1}(s, x') = V^{n+1}(s).$$

**Proof of Theorem 3.1.** We proceed by induction on  $n$ . Consider the base case, which is  $n = N - 1$ . Fix  $s = (\mu, \beta) \in \mathbb{S}$ . Then  $V^N(s) = \max_x \mu_x$  is convex in its arguments, so we can employ Jensen's inequality to write

$$\begin{aligned} V^{\pi, N-1}(s) &= \mathbb{E} [V^{\pi, N}(T(s, X^\pi(s), Z^N))] \geq V^{\pi, N}(\mathbb{E} [T(s, X^\pi(s), Z^N)]) \\ &= V^{\pi, N}(\mu, \beta + \beta^\epsilon e_{X^\pi(s)}) = V^{\pi, N}(\mu, \beta) = V^{\pi, N}(s). \end{aligned}$$

Now consider the induction step. For  $n < N - 1$ ,

$$V^{\pi, n}(s) = \mathbb{E} [V^{\pi, n+1}(T(s, X^\pi(s), Z^{n+1}))] \geq \mathbb{E} [V^{\pi, n+2}(T(s, X^\pi(s), Z^{n+1}))]$$

by the induction hypothesis. Then, by the definition of  $V^{\pi, n+1}$  in terms of  $V^{\pi, n+2}$  from (10), we have  $V^{\pi, n}(s) \geq V^{\pi, n+1}(s)$ .

**Proof of Theorem 4.1.** By (15), computing  $X^{KG}(s)$  reduces to computing  $Q^{N-1}(s, x)$  for each  $x \in \{1, \dots, M\}$ . By definition (11) we have, for a generic state  $s$  and standard normal random variable  $Z$ ,

$$(22) \quad Q^{N-1}(s, x) = \mathbb{E} [V^N(T(s, x, Z))] = \mathbb{E} \left[ (\mu_x + \tilde{\sigma}(\beta_x)Z) \vee \max_{x' \neq x} \mu_{x'} \right].$$

This expectation is the expectation of the maximum of a constant and a normal random variable, for which we have an analytical expression from [13]. Let  $a \in \mathbb{R}$  be an arbitrary constant and  $W \sim \mathcal{N}(b, c^2)$  an arbitrary normal random variable. Then [13] tells us that

$$(23) \quad \mathbb{E} [W \vee a] = a\Phi\left(\frac{a-b}{c}\right) + b\Phi\left(\frac{b-a}{c}\right) + c\varphi\left(\frac{a-b}{c}\right),$$

which can be rewritten as

$$\begin{aligned} \mathbb{E} [W \vee a] &= a\Phi\left(\frac{a-b}{c}\right) + b\left(1 - \Phi\left(\frac{a-b}{c}\right)\right) + c\varphi\left(\frac{a-b}{c}\right) \\ &= b + (a-b)\Phi\left(\frac{a-b}{c}\right) + c\varphi\left(\frac{a-b}{c}\right) \\ &= b + c\left[\left(\frac{a-b}{c}\right)\Phi\left(\frac{a-b}{c}\right) + \varphi\left(\frac{a-b}{c}\right)\right]. \end{aligned}$$

Fix  $x$  and consider two cases. First, consider the case that  $\mu_x > \max_{x'} \mu_{x'}$ . This is the case in which we measure an alternative that is uniquely the best according to the prior. Then  $\mu_x - \max_{x' \neq x} \mu_{x'}$  is positive and  $(\max_{x' \neq x} \mu_{x'} - \mu_x)/\tilde{\sigma}(\beta_x) = \zeta_x(s)$ . Substitute  $\zeta_x(s)$  for  $(a-b)/c$  and write (22) as

$$Q^{N-1}(s, x) = \mu_x + \tilde{\sigma}(\beta_x) [\zeta_x(s)\Phi(\zeta_x(s)) + \varphi(\zeta_x(s))] = \mu_x + \tilde{\sigma}(\beta_x)f(\zeta_x(s)),$$

which can be rewritten in our case using  $\mu_x = \max_{x'} \mu_{x'}$  as

$$(24) \quad Q^{N-1}(s, x) = \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x)f(\zeta_x(s)).$$

Now consider the case that  $\mu_x \leq \max_{x'} \mu_{x'}$ . We rewrite (23) again using the substitution  $\Phi(-z) = 1 - \Phi(z)$  and also using the symmetric property of the normal probability density function,  $\varphi(-z) = \varphi(z)$ , as

$$\mathbb{E} [Z \vee a] = a + c\left[\left(\frac{b-a}{c}\right)\Phi\left(\frac{b-a}{c}\right) + \varphi\left(\frac{b-a}{c}\right)\right].$$

In the case we are considering,  $\mu_x - \max_{x' \neq x} \mu_{x'} \leq 0$  and  $(\mu_x - \max_{x' \neq x} \mu_{x'})/\tilde{\sigma}(\beta_x) = \zeta_x(s)$ . Substitute  $\zeta_x(s)$  for  $(b-a)/c$  and write (22) as

$$\begin{aligned} Q^{N-1}(s, x) &= \max_{x' \neq x} \mu_{x'} + \tilde{\sigma}(\beta_x) [\zeta_x(s)\Phi(\zeta_x(s)) + \varphi(\zeta_x(s))] \\ &= \max_{x' \neq x} \mu_{x'} + \tilde{\sigma}(\beta_x)f(\zeta_x(s)), \end{aligned}$$

which can be rewritten in our case using  $\max_{x' \neq x} \mu_{x'} = \max_{x'} \mu_{x'}$  as

$$(25) \quad Q^{N-1}(s, x) = \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)).$$

In both cases the expression for  $Q^{N-1}(s, x)$  agrees with (18), and we use this expression to rewrite (15) as

$$X^{KG}(s) \in \arg \max_x \max_{x'} \mu_{x'} + \tilde{\sigma}(\beta_x) f(\zeta_x(s)) = \arg \max_{x \in \{1, \dots, M\}} \tilde{\sigma}(\beta_x) f(\zeta_x(s)),$$

since  $\max_{x'} \mu_{x'}$  does not depend on  $x$ .

**Proof of Proposition 4.3.** By Theorem 4.1, KG prefers the alternative with the largest value of  $\tilde{\sigma}(\beta_x) f(\zeta_x(S))$ . Fix  $S = (\mu, \beta)$ , and let  $a$  be as in the statement of Proposition 4.3. Let  $i$  be the alternative preferred by KG, so

$$(26) \quad i = \arg \max_{x \in \{1, \dots, M\}} \tilde{\sigma}(\beta_x) f(\zeta_x(S)),$$

where we recall that we are breaking ties by choosing the smallest index. Note that the theorem's condition on  $a$  trivializes the case when  $\mu_i = \max_x \mu_x$  because here the range of  $a$  contains only the value 0, for which the theorem is obviously true. Thus, without loss of generality we may assume  $\mu_i < \max_x \mu_x$ , and let  $j \in \arg \max_x \mu_x$ . Then  $j \neq i$ .

Let  $S' = (\mu + ae_i, \beta)$ . We will first show for all alternatives  $x \neq i$  that

$$(27) \quad \tilde{\sigma}(\beta_i) f(\zeta_i(S')) \geq \tilde{\sigma}(\beta_x) f(\zeta_x(S')).$$

This will show that  $i \in \arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S'))$ . We will then show that the implication

$$(28) \quad \tilde{\sigma}(\beta_x) f(\zeta_x(S)) < \tilde{\sigma}(\beta_i) f(\zeta_i(S)) \implies \tilde{\sigma}(\beta_x) f(\zeta_x(S')) < \tilde{\sigma}(\beta_i) f(\zeta_i(S'))$$

holds for all  $x \neq i$ . This will suffice to show the proposition because if we choose any  $x' \notin \arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S))$ , (26) will imply  $\tilde{\sigma}(\beta_{x'}) f(\zeta_{x'}(S)) < \tilde{\sigma}(\beta_i) f(\zeta_i(S))$ . The implication (28) will then imply that  $\tilde{\sigma}(\beta_{x'}) f(\zeta_{x'}(S')) < \tilde{\sigma}(\beta_i) f(\zeta_i(S'))$  and, moreover, that  $x' \notin \arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S'))$ . Taking the contrapositive of the statement

$$x' \notin \arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S)) \implies x' \notin \arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S'))$$

reveals that

$$x' \in \arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S')) \implies x' \in \arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S)).$$

By this argument, (28) implies that  $\arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S')) \subseteq \arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S))$ . Therefore  $i$  is the element of  $\arg \max_x \tilde{\sigma}(\beta_x) f(\zeta_x(S))$  with the smallest index, and thus  $i$  is the alternative that KG prefers in state  $S'$ .

We will show (27) and (28) by treating three cases separately, noting in general that  $\zeta_i(\mu, \beta) \leq \zeta_i(\mu + ae_i, \beta)$ . The first case is when  $x \neq i, j$ . Then

$$\zeta_x(S') = \zeta_x(\mu + ae_i, \beta) = \zeta_x(\mu, \beta) = \zeta_x(S).$$

Thus, (27) is true because

$$\tilde{\sigma}(\beta_i)f(\zeta_i(S')) \geq \tilde{\sigma}(\beta_i)f(\zeta_i(S)) \geq \tilde{\sigma}(\beta_x)f(\zeta_x(S)) = \tilde{\sigma}(\beta_x)f(\zeta_x(S')),$$

and (28) is true because if  $\tilde{\sigma}(\beta_x)f(\zeta_x(S)) < \tilde{\sigma}(\beta_i)f(\zeta_i(S))$ , then

$$\tilde{\sigma}(\beta_x)f(\zeta_x(S')) = \tilde{\sigma}(\beta_x)f(\zeta_x(S)) < \tilde{\sigma}(\beta_i)f(\zeta_i(S)) \leq \tilde{\sigma}(\beta_i)f(\zeta_i(S')).$$

The second case is when  $x = j$  and  $\mu_i + a < \max_{x' \neq j} \mu_{x'}$ . Then again  $\zeta_j(S') = \zeta_j(S)$  because  $j \neq i$ , and both (27) and (28) hold by the same reasoning as in the first case.

The third case is when  $x = j$  and  $\mu_i + a \geq \max_{x' \neq j} \mu_{x'}$ . Then we have  $\zeta_j(\mu + ae_i, \beta) = \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_j)}$ . For  $x = j$ , KG's preference of alternative  $i$  implies that  $\beta_i \leq \beta_j$ . Otherwise, by Remark 4.2 and because  $\mu_j \geq \mu_i$ , KG would prefer alternative  $j$ . This shows that

$$\zeta_i(\mu + ae_i, \beta) = \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_i)} \geq \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_j)} = \zeta_j(\mu + ae_i, \beta).$$

This shows (27). To show (28), assume the antecedent of condition (28). Since  $|\mu_j - \max_{x' \neq j} \mu_{x'}| \leq |\mu_j - \mu_i|$  and  $\tilde{\sigma}(\beta_j)f(\zeta_j(S)) < \tilde{\sigma}(\beta_i)f(\zeta_i(S))$ , it must be that  $\tilde{\sigma}(\beta_j) < \tilde{\sigma}(\beta_i)$  since otherwise  $j$  would have been KG's choice in state  $S$ . Thus,

$$\zeta_i(\mu + ae_i, \beta) = \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_i)} > \frac{-|\mu_i + a - \mu_j|}{\tilde{\sigma}(\beta_j)} = \zeta_j(\mu + ae_i, \beta).$$

**Proof of Proposition 5.1.** We will show that  $V^0(S^0; N)$  is a nondecreasing function of  $N$  bounded from above by  $U(S^0)$ , which will imply that the limit  $V(S^0; \infty)$  exists and is bounded as claimed. To show that  $V^0(S^0; N)$  is nondecreasing in  $N$ , note that  $V^0(S^0; N - 1) = V^1(S^0; N)$ , and thus

$$V^0(S^0; N) - V^0(S^0; N - 1) = V^0(S^0; N) - V^1(S^0; N).$$

This difference is positive by Corollary 3.2.

Now we show that  $V^0(S^0; N) \leq U(S^0)$ . For every  $N \geq 1$  and policy  $\pi$ ,

$$\begin{aligned} \mathbb{E}^\pi \left[ \max_x \mu_x^N \right] &= \mathbb{E}^\pi \left[ \max_x \mathbb{E}_N^\pi [Y_x] \right] \leq \mathbb{E}^\pi \left[ \mathbb{E}_N^\pi \left[ \max_x Y_x \right] \right] \\ &= \mathbb{E}^\pi \left[ \max_x Y_x \right] = \mathbb{E} \left[ \max_x Y_x \right]. \end{aligned}$$

This value is independent of  $\pi$  and is equal to  $U(S^0)$ . Thus

$$V^0(S^0; N) := \sup_\pi \mathbb{E}^\pi \left[ \max_x \mu_x^N \right] \leq U(S^0)$$

for every  $N \geq 1$ . Taking the limit as  $N \rightarrow \infty$  shows  $V(S^0; \infty) \leq U(S^0)$ .

Finally, we show that the limit  $V^\pi(S^0; \infty)$  exists and is finite for every stationary policy  $\pi$ . Fix a stationary policy  $\pi$ . Then Theorem 3.1 implies that  $V^{\pi, 0}(S^0; N)$  is nondecreasing in  $N$ , and  $V^{\pi, 0}(S^0; N)$  is bounded by  $V^0(S^0; N)$ , which is itself uniformly bounded in  $N$  by  $U(S^0)$ . Then  $V^\pi(S^0; \infty)$  is the limit of a nondecreasing bounded sequence. Hence, it exists.

**Proof of Proposition 5.2.** We assumed in the formal model in section 3.1 that our measurement-noise variance  $(\sigma^\varepsilon)^2$  is finite. This implies via the strong law of large numbers that the sequence of posterior predictive means  $\mu_x^N$  converges as  $\lim_{N \rightarrow \infty} \mu_x^N = Y_x$  a.s. for each  $x = 1, \dots, M$ . Thus  $\lim_{N \rightarrow \infty} \max_x \mu_x^N$  exists a.s. and in probability. We will show next that the sequence  $(\max_x \mu_x^N)_{N \geq 1}$  is uniformly integrable, and then convergence in probability together with uniform integrability implies convergence in  $L^1$  (see, e.g., [20, Theorem 3.12]). Convergence in  $L^1$  of  $\max_x \mu_x^N$  as  $N \rightarrow \infty$  implies

$$V^\pi(S^0; \infty) = \lim_{N \rightarrow \infty} \mathbb{E}^\pi \left[ \max_x \mu_x^N \right] = \mathbb{E}^\pi \left[ \lim_{N \rightarrow \infty} \max_x \mu_x^N \right] = \mathbb{E}^\pi \left[ \max_x Y_x \right] = U(S^0).$$

Proposition 5.1 showed that  $U(S^0) \geq V(S^0; \infty)$ , so  $V^\pi(S^0; \infty) = V(S^0; \infty)$  and  $\pi$  must be asymptotically optimal.

To complete the proof we must show uniform integrability of the sequence  $(\max_x \mu_x^N)_{N \geq 1}$ . For every fixed  $K$  we have

$$\begin{aligned} \mathbb{E} \left[ \max_x \mu_x^N \mid 1_{\{\max_x \mu_x^N \geq K\}} \right] &\leq \mathbb{E} \left[ \max_x |\mu_x^N| \mid 1_{\{\max_x |\mu_x^N| \geq K\}} \right] \\ &= \mathbb{E} \left[ \max_x |\mathbb{E}_N[Y_x]| \mid 1_{\{\max_x |\mathbb{E}_N[Y_x]| \geq K\}} \right] \leq \mathbb{E} \left[ \max_x \mathbb{E}_N[|Y_x|] \mid 1_{\{\max_x \mathbb{E}_N[|Y_x|] \geq K\}} \right] \\ &\leq \mathbb{E} \left[ \mathbb{E}_N \left[ \max_x |Y_x| \right] \mid 1_{\{\mathbb{E}_N[\max_x |Y_x|] \geq K\}} \right] = \mathbb{E} \left[ \mathbb{E}_N \left[ \max_x |Y_x| \mid 1_{\{\mathbb{E}_N[\max_x |Y_x|] \geq K\}} \right] \right] \\ &= \mathbb{E} \left[ \max_x |Y_x| \mid 1_{\{\mathbb{E}_N[\max_x |Y_x|] \geq K\}} \right]. \end{aligned}$$

We assumed in the formal model in section 3.1 that  $\max_x |Y_x|$  was integrable. This implies via Markov's inequality that

$$\mathbb{P} \left\{ \mathbb{E}_N \left[ \max_x |Y_x| \right] \geq K \right\} \leq \frac{\mathbb{E} [\mathbb{E}_N [\max_x |Y_x|]]}{K} = \frac{\mathbb{E} [\max_x |Y_x|]}{K}.$$

This is bounded uniformly in  $N$ , and the bound goes to zero as  $K \rightarrow \infty$ .

**Proof of Theorem 5.1.** First note that KG is stationary. We will show that  $\lim_{N \rightarrow \infty} \eta_x^N = \infty$  a.s. for all  $x$  under KG, and then Proposition 5.2 will complete the proof.

First we show that, for each  $x$ ,  $\{\mu_x^n\}_{n=0}^\infty$  is a uniformly integrable martingale with respect to the filtration  $\mathcal{F}$  and hence converges.  $\mu_x^n$  is defined by  $\mu_x^n := \mathbb{E}[Y_x \mid \mathcal{F}^n]$  and thus is  $\mathcal{F}^n$ -measurable and, by the tower property of conditional expectation, satisfies the martingale identity.  $Y_x$  is a normal random variable with finite variance. Thus,  $Y_x \in L^2 \subset L^1$ , and by the Doob uniform integrability lemma [20, Lemma 5.5], the collection of conditional expectations  $\{\mu_x^n\}_n$  is uniformly integrable (and hence each  $\mu_x^n$  is integrable). Thus,  $\{\mu_x^n\}_n$  is a uniformly integrable martingale and hence converges a.s. to an integrable random variable  $\mu_x^\infty$ . In addition,  $\lim_{n \rightarrow \infty} \beta_x^n \stackrel{a.s.}{=} \beta_x^0 + \beta^\varepsilon \eta_x^\infty$  for each  $x$ .

By the computation performed in Theorem 4.1, the Q-factors for each alternative  $x$  are continuous functions of their arguments  $(\mu, \beta)$ , and, hence,

$$\lim_{n \rightarrow \infty} Q^{N-1}(S^n; x) \stackrel{a.s.}{=} \max_{x'} \mu_{x'}^\infty + \tilde{\sigma}(\beta_x^\infty) f \left( \frac{\mu_x^\infty - \max_{x'' \neq x} \mu_{x''}^\infty}{\tilde{\sigma}(\beta_x^\infty)} \right).$$

Define  $\Omega_0$  to be the almost sure event on which this convergence holds, and define the event  $\mathcal{H}_x$  to be  $\mathcal{H}_x := \{\omega : \eta_x^\infty(\omega) < \infty\}$ . Then,

$$(29) \quad \lim_{n \rightarrow \infty} Q^{N-1}(S^n(\omega); x) > \max_{x'} \mu_{x'}^\infty(\omega) \quad \text{for all } \omega \in \mathcal{H}_x \cap \Omega_0,$$

$$(30) \quad \lim_{n \rightarrow \infty} Q^{N-1}(S^n(\omega); x) = \max_{x'} \mu_{x'}^\infty(\omega) \quad \text{for all } \omega \in \mathcal{H}_x^c \cap \Omega_0.$$

Let  $A$  be any subset of  $\{1, \dots, M\}$ , and define the event  $\mathcal{H}_A$  to be  $\mathcal{H}_A := (\cap_{x \in A} \mathcal{H}_x) \cap (\cap_{x \in A^c} \mathcal{H}_x^c)$ . We will show that, if  $A \neq \emptyset$ , then  $\mathbb{P}(\mathcal{H}_A) = 0$ . This will prove the theorem because  $\Omega = \cup_{A \subseteq \{1, \dots, M\}} \mathcal{H}_A$ , so if we know that  $A \neq \emptyset \implies \mathbb{P}(\mathcal{H}_A) = 0$ , then  $1 = \mathbb{P}(\mathcal{H}_\emptyset) = \mathbb{P}\{\lim_{n \rightarrow \infty} \eta_x^n = \infty \text{ for all } x\}$ .

Fix  $A$  nonempty and suppose for contradiction that  $\mathcal{H}_A \cap \Omega_0$  is nonempty so that we may choose  $\omega \in \mathcal{H}_A \cap \Omega_0$  to be an element of this set. By (29) and (30), for all  $x \in A$  and all  $y \in A^c$ ,

$$\lim_{n \rightarrow \infty} Q^{N-1}(S^n(\omega); x) > \lim_{n \rightarrow \infty} Q^{N-1}(S^n(\omega); y),$$

and there exists a finite number  $K_{xy}$  such that, for all  $n > K_{xy}$ ,

$$Q^{N-1}(S^n(\omega); x) > Q^{N-1}(S^n(\omega); y).$$

Let  $K := \max_{x \in A, y \in A^c} K_{xy}$  if  $A^c$  is nonempty, and let  $K := 1$  if  $A^c$  is empty. Then  $K$  is finite and for all  $n > K$  and all  $x \in A$  and  $y \in A^c$ ,

$$Q^{N-1}(S^n(\omega); x) > Q^{N-1}(S^n(\omega); y).$$

Therefore, KG distributes all measurements  $n > K$  only to alternatives in the set  $A$ , and  $\sum_{x \in A} \eta_x^\infty(\omega) = \infty$ . This is a contradiction because  $x \in A$  implies  $\omega \in \mathcal{H}_x$ , which implies  $\eta_x^\infty(\omega) < \infty$ .

Thus,  $\mathbb{P}(\mathcal{H}_\emptyset \cap \Omega_0) = 0$ , and since  $\mathbb{P}(\Omega_0) = 1$ ,  $\mathbb{P}(\mathcal{H}_\emptyset) = 0$ .

**Proof of Theorem 6.1.** Note that  $\varphi(0) = (2\pi)^{-1/2}$ , where  $\varphi$  is the normal probability density function. We will use this throughout. We induct backward over  $n$ . First, when  $n = N - 1$ , the theorem is trivially true with equality. Now, under the assumption that the theorem is true for some  $n + 1$ ,

$$\begin{aligned} V^n(s) &= \max_x \mathbb{E} [V^{n+1}(T(s, x, Z^{n+1}))] \\ &\leq \max_x \mathbb{E} \left[ V^{N-1}(T(s, x, Z^{n+1})) + \varphi(0)(N - n - 2) \max_x \tilde{\sigma}(\beta_{x'} + \beta^\epsilon 1_{\{x=x'\}}) \right]. \end{aligned}$$

Then, since  $\tilde{\sigma}$  is a decreasing function and  $\beta_{x'}^n \leq \beta_{x'}^n + \beta^\epsilon 1_{\{x=x'\}}$ ,

$$V^n(s) \leq \max_x \mathbb{E} \left[ V^{N-1}(T(s, x, Z^{n+1})) + \varphi(0)(N - n - 2) \max_{x'} \tilde{\sigma}(\beta_{x'}) \right].$$

Since the last term is a constant and does not depend on  $x$ , we may move it outside the maximum and expectation operators, giving

$$(31) \quad V^n(s) \leq \max_x \mathbb{E} [V^{N-1}(T(s, x, Z^{n+1}))] + \varphi(0)(N - n - 2) \max_{x'} \tilde{\sigma}(\beta_{x'}).$$

We will rewrite the first term on the right-hand side of this inequality as a maximum over a set of Q-factors using the definition of  $V^{N-1}$  in terms of  $Q^{N-1}$ , but before

making this substitution, let us bound  $Q^{N-1}$ . We rewrite the expression (24) for  $Q^{N-1}$  as  $Q^{N-1}(s, x') = \max_{x''} \mu_{x''} + \tilde{\sigma}(\beta_{x'}) f(\zeta_{x'}) = V^N(s) + \tilde{\sigma}(\beta_{x'}) f(\zeta_{x'})$ . Lemma 4.1 tells us that  $f$  is nondecreasing, so  $\zeta_{x'} \leq 0$  implies that  $f(\zeta_{x'}) \leq f(0) = \varphi(0)$ . Thus,

$$Q^{N-1}(s, x') \leq V^N(s) + \varphi(0) \tilde{\sigma}(\beta_{x'}).$$

Using this and the definition of the value function in terms of the Q-factors from (10) and (11), we have

$$\begin{aligned} V^{N-1}(T(s, x, Z^{n+1})) &= \max_{x'} Q^{N-1}(T(s, x, Z^{n+1}), x') \\ &\leq \max_{x'} V^N(T(s, x, Z^{n+1})) + \varphi(0) \tilde{\sigma}(\beta_{x'} + \beta^\epsilon 1_{\{x=x'\}}) \\ &= V^N(T(s, x, Z^{n+1})) + \varphi(0) \max_{x'} \tilde{\sigma}(\beta_{x'} + \beta^\epsilon 1_{\{x=x'\}}) \\ &\leq V^N(T(s, x, Z^{n+1})) + \varphi(0) \max_{x'} \tilde{\sigma}(\beta_{x'}). \end{aligned}$$

Combining this bound with (31) and moving the  $\tilde{\sigma}(\beta_x)$  outside the maximization and expectation operators, we obtain

$$\begin{aligned} V^n(s) &\leq \max_x \mathbb{E} \left[ V^N(T(s, x, Z^{n+1})) + \varphi(0) \max_{x'} \tilde{\sigma}(\beta_{x'}^n) \right] + \varphi(0)(N - n - 2) \max_{x'} \tilde{\sigma}(\beta_{x'}) \\ &= \max_x \mathbb{E} \left[ V^N(T(s, x, Z^{n+1})) \right] + \varphi(0)(N - n - 1) \max_{x'} \tilde{\sigma}(\beta_{x'}) \\ &= V^{N-1}(s) + \varphi(0)(N - n - 1) \max_{x'} \tilde{\sigma}(\beta_{x'}), \end{aligned}$$

where in the last step we used the definition of  $V^N$  in terms of  $V^{N-1}$  from (10).

**Proof of Theorem 7.1.** The proof is by induction backward on  $k$ . The theorem holds for the base case,  $k = N - 1$ , by Remark 4.1. Now let  $k < N - 1$ . Let  $\pi^*$  be an optimal policy, with decision function  $X^{*k}$  at time  $k$ . Let  $s = (\mu, \beta) \in \mathbb{S}^k$ . Then

$$(32) \quad V^k(s) = \mathbb{E} [V^{k+1}(T(s, X^{*k}(s), Z^{k+1}))] = \mathbb{E} [V^{KG, k+1}(T(s, X^{*k}(s), Z^{k+1}))]$$

by the induction hypothesis, since  $\{\mathbb{S}^n\}_{n=k+1}^N$  is a covering of the future from  $k + 1$  on which KG persistence holds, and  $T(s, X^{*k}(s), Z^{k+1}) \in \mathbb{S}^{k+1}$  a.s.

Consider two cases. In the first case, suppose  $X^{*k}(s) = X^{KG}(s)$ . By (32),

$$V^k(s) = \mathbb{E} [V^{KG, k+1}(T(s, X^{KG}(s), Z^{k+1}))] = V^{KG, k}(s).$$

In the second case, suppose  $X^{*k}(s) \neq X^{KG}(s)$ . Then, abbreviating the random state at time  $k + 1$  under the optimal policy by  $S^{k+1} = T(s, X^{*k}(s), Z^{k+1})$ ,

$$\begin{aligned} V^k(s) &= \mathbb{E} [V^{KG, k+2}(T(S^{k+1}, X^{KG}(S^{k+1}), Z^{k+2}))] \\ (33) \quad &= \mathbb{E} [V^{KG, k+2}(T(S^{k+1}, X^{KG}(s), Z^{k+2}))], \end{aligned}$$

since  $X^{KG}(s) = X^{KG}(S^{k+1})$  a.s. by the KG persistence property. Let  $S^{k+2} = T(S^{k+1}, X^{KG}(s), Z^{k+2})$ . Then  $V^k(s) = \mathbb{E} [V^{KG, k+2}(S^{k+2})]$ .

Note that  $S^{k+2}$  is the state to which we arrive when we measure  $X^{*k}(s)$  at time  $k$  and  $X^{KG}(s)$  at time  $k + 1$ . Let  $E_x = e_x(e_x)^T$  be a matrix of all zeros except for

a single 1 at row  $x$ , column  $x$ , and let  $\stackrel{d}{=}$  denote equality in distribution. Then the definition (8) of the transition function  $T$  and  $X^{KG}(s) \neq X^{*,k}(s)$  imply

$$\begin{aligned} S^{k+2} &= T(S^{k+1}, X^{KG}(s), Z^{k+2}) \\ &= T(T(s, X^{*,k}(s), Z^{k+1}), X^{KG}(s), Z^{k+2}) \\ &= \mu + \tilde{\sigma}(\beta_{X^{KG}(s)})Z^{k+1} + \tilde{\sigma}(\beta_{X^{*,k}(s)})Z^{k+2} + \beta^\epsilon E_{X^{KG}(s)} + \beta^\epsilon E_{X^{*,k}(s)} \\ &\stackrel{d}{=} \mu + \tilde{\sigma}(\beta_{X^{KG}(s)})Z^{k+2} + \tilde{\sigma}(\beta_{X^{*,k}(s)})Z^{k+1} + \beta^\epsilon E_{X^{KG}(s)} + \beta^\epsilon E_{X^{*,k}(s)} \\ &= T(T(s, X^{KG}(s), Z^{k+1}), X^{*,k}(s), Z^{k+2}). \end{aligned}$$

Thus, we have that  $V^k(s) = \mathbb{E}[V^{KG,k+2}(S^{k+2})]$  equals

$$\mathbb{E}[V^{KG,k+2}(T(T(s, X^{KG}(s), Z^{k+1}), X^{*,k}(s), Z^{k+2}))].$$

This quantity is the value of making decisions  $X^{KG}(s)$  at time  $k$ ,  $X^{*,k}(s)$  at time  $k+1$ , and then following KG afterward. This value must be less than the value of making the same decision  $X^{KG}(s)$  at time  $k$  and following the optimal policy afterward. Thus,  $V^k(s) \leq \mathbb{E}[V^{k+1}(T(s, X^{KG}(s), Z^{k+1}))]$ . Now,  $T(s, X^{KG}(s), Z^{n+1}) \in \mathbb{S}^{n+1}$  a.s., so by the induction hypothesis we may replace the optimal value function with the KG value function when operating on this state. This allows us to write

$$V^k(s) \leq \mathbb{E}[V^{KG,k+1}(T(s, X^{KG}(s), Z^{k+1}))] = V^{KG,k}(s).$$

Finally,  $V^k(s) \geq V^{KG,k}(s)$  implies  $V^k(s) = V^{KG,k}(s)$ .

**Proof of Theorem 7.3.** For  $n \in \{0, \dots, N-1\}$ , define  $\mathbb{S}^n$  to be the set of all  $s = (\mu, \beta) \in \mathbb{S}$  satisfying

$$(34) \quad (\beta_i \neq \infty \text{ and } \beta_j \neq \infty \text{ and } \beta_i < \beta_j) \implies \mu_i \geq \mu_j$$

for all  $i, j \in \{1, \dots, M\}$ . Note that the sets  $\mathbb{S}^n$  are identical for all  $n$ . We will show that  $\{\mathbb{S}^n\}$  is a covering of the future from 0.

Let  $n \in \{0, \dots, N-2\}$ ,  $x \in \{1, \dots, M\}$ ,  $s \in \mathbb{S}^n$ , and  $S^n = s$  a.s. Consider  $S^{n+1} := T(S^n, x, Z^{n+1})$ . Let  $i, j \in \{1, \dots, M\}$  meet the conditions of the implication (34) for  $S^{n+1}$ , so  $\beta_i^{n+1} \neq \infty$  and  $\beta_j^{n+1} \neq \infty$  and  $\beta_i^{n+1} < \beta_j^{n+1}$ . We will show that  $\mu_i^n \geq \mu_j^n$ , which will show that  $S^{n+1}$  meets condition (34) and is in  $\mathbb{S}^{n+1}$ .

First,  $\beta^n \leq \beta^{n+1}$  componentwise implies that  $\beta_i^n \neq \infty$  and  $\beta_j^n \neq \infty$ . Also, since  $(\sigma^\epsilon)^2 = 0$ ,  $\beta_x^{n+1} = \infty$ , which implies that  $x \neq i, j$ , and the measurement between  $S^n$  and  $S^{n+1}$  alters neither the  $i$  component nor the  $j$  component. Thus,  $\beta_i^n = \beta_i^{n+1} < \beta_j^{n+1} = \beta_j^n$ . This shows that  $i, j$  meet the conditions of the implication (34) for  $S^n$  as well as  $S^{n+1}$ . Thus, since  $S^n \in \mathbb{S}^n$ ,  $\mu_i^n \geq \mu_j^n$ . Then, again because  $x \neq i, j$  implies that the means of the  $i$  and  $j$  components did not change from time  $n$  to  $n+1$ ,  $\mu_i^{n+1} \geq \mu_j^{n+1}$ , showing that  $S^{n+1}$  meets the condition (34), and  $S^{n+1} \in \mathbb{S}^{n+1}$ . Thus,  $\{\mathbb{S}^n\}$  is a covering of the future from 0.

Now we will show that KG is persistent on  $\{\mathbb{S}^n\}$ . Let  $s \in \mathbb{S}^n$  and  $S^n = s$  a.s. Condition (34) together with Remarks 4.2 and 4.3 implies  $X^{KG}(S^n) \in \arg \min_x \beta_x^n$ , with ties broken by the smallest index. Let  $x \neq X^{KG}(S^n)$ . We showed that  $S^{n+1} := T(S^n, x, Z^{n+1}) \in \mathbb{S}^{n+1}$  a.s. Thus, again by condition (34) and Remarks 4.2 and 4.3,



$X^{KG}(S^{n+1}) \in \arg \min_{x'} \beta_{x'}^{n+1}$ . We use the state transition function for the case with  $(\sigma^\epsilon) = 0$ ,  $\beta_{x'}^{n+1} = \beta_{x'}^n + \infty 1_{\{x'=x\}}$ , and we consider two cases.

In the first case suppose  $\beta_{x'}^n < \infty$  for some  $x' \neq x$ . Then, since  $\beta_x^{n+1} = \infty$ , we have  $\beta_{x'}^{n+1} = \beta_{x'}^n < \beta_x^{n+1}$ . Thus, we may drop  $x$  from the argmin set as in

$$\arg \min_{x'} \beta_{x'}^{n+1} = \arg \min_{x' \neq x} \beta_{x'}^{n+1} = \arg \min_{x' \neq x} \beta_{x'}^n.$$

$X^{KG}(S^n)$  is the element of this set with the smallest index, and since  $X^{KG}(S^{n+1})$  is also defined to be the element of this set with the smallest index,  $X^{KG}(S^{n+1}) = X^{KG}(S^n)$ .

In the second case suppose  $\beta_{x'}^n = \infty$  for all  $x' \neq x$ . Then, by  $X^{KG} \in \arg \min_{x'} \beta_{x'}^n$ , and since  $X^{KG}(S^n) \neq x$ , we also have that  $\beta_x^n = \infty$ . The state transition rule for  $\beta$  implies that  $\beta_{x'}^{n+1} = \infty$  for all  $x'$ . Thus,  $\arg \min_{x'} \beta_{x'}^n = \{1, \dots, M\} = \arg \min_{x'} \beta_{x'}^{n+1}$ , and since the tie-breaking rule is fixed to choose the element with the smallest index,  $X^{KG}(S^{n+1}) = X^{KG}(S^n)$ .

In both cases KG is persistent on  $\{S^n\}$ , and Theorem 7.1 shows that  $V^{KG,0}(s) = V^0(s)$  for all  $s \in \mathbb{S}^0$ .

**Appendix B. Known variance LL(S) policy.** The LL(S) policy was developed for normal measurement errors with *unknown* variance and uses a normal-gamma prior for the unknown mean and measurement precision. To adapt it to the known-variance case, we take both the shape and rate parameters in the gamma prior on the measurement precision to infinity while keeping their ratio fixed to the known measurement precision  $\beta^\epsilon$ ; we obtain a prior in which the measurement precision is known perfectly and the alternative's true value is still normally distributed. Taking this limit in the allocation given by [12, Corollary 1] provides the following policy. The steps below describe how the policy allocates  $\tau$  measurements for the stage beginning at a generic time  $n$ , and should be repeated a total of  $N/\tau$  times beginning at time 0 and finishing at time  $N$ . We use the notation  $[i]$  to indicate the alternative whose  $\mu^n$  component is  $i$ th largest. That is,  $\mu_{[M]}^n \geq \dots \geq \mu_{[1]}^n$ .

- (i) For each alternative calculate  $n_i = \beta_i^n / \beta^\epsilon$ , which may be interpreted as the effective number of times alternative  $i$  has been sampled.
- (ii) Initialize  $\mathcal{S}$ , the set of alternatives under consideration for measurement in the current stage, to  $\mathcal{S} = \{1, \dots, M\}$ .
- (iii) For each  $i \in \mathcal{S} \setminus \{[M]\}$  set  $\lambda_{i,M}$  as follows. If  $[M] \notin \mathcal{S}$ , set  $\lambda_{i,M} = \beta_{[i]}^n$ . If  $[M] \in \mathcal{S}$ , set  $\lambda_{i,M} = ((\beta_{[M]}^n)^{-1} + (\beta_{[i]}^n)^{-1})^{-1}$ .
- (iv) Calculate a tentative number of samples  $r_{[i]}$  to take from alternative  $[i]$ ,

$$r_{[i]} = \frac{\tau + \sum_{j \in \mathcal{S}} n_j}{\sum_{j \in \mathcal{S}} \sqrt{\gamma_j / \gamma_{[i]}}} - n_{[i]},$$

where

$$\gamma_{[i]} = \begin{cases} \sqrt{\lambda_{i,M}} \phi \left( \sqrt{\lambda_{i,M}} (\mu_{[M]}^n - \mu_{[i]}^n) \right) & \text{if } [i] \neq [M], \\ \sum_{[j] \in \mathcal{S} \setminus \{[M]\}} \gamma_{[j]} & \text{if } [i] = [M]. \end{cases}$$

- (v) For each  $[i] \in \mathcal{S}$  with  $r_{[i]} < 0$ , remove  $[i]$  from  $\mathcal{S}$  and set  $r_{[i]} = 0$ . If any  $[i]$  was removed, then return to step (iii).
- (vi) Round the  $r_{[i]}$  to integer values so that they still sum to  $\tau$ .
- (vii) Run  $r_{[i]}$  additional samples for each alternative  $[i]$ .

## REFERENCES

- [1] R. BECHHOFFER, T. SANTNER, AND D. GOLDSMAN, *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*, John Wiley & Sons, New York, 1995.
- [2] J. BRANKE, S. CHICK, AND C. SCHMIDT, *New developments in ranking and selection: An empirical comparison of the three main approaches*, in Proceedings of the 2005 Winter Simulation Conference, M. Kuhl, N. Steiger, F. Armstrong, and J. Joines, eds., IEEE, Piscataway, NJ, 2005, pp. 708–717.
- [3] H. CHANG, M. FU, J. HU, AND S. MARCUS, *Simulation-Based Algorithms for Markov Decision Processes*, Springer, Berlin, 2007.
- [4] C. CHEN, *An effective approach to smartly allocate computing budget for discrete event simulation*, in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp. 2598–2603.
- [5] C. CHEN, L. DAI, AND H. CHEN, *A gradient approach for smartly allocating computing budget for discrete event simulation*, in Proceedings of the 28th Winter Simulation Conference, IEEE Computer Society, Washington, DC, 1996, pp. 398–405.
- [6] C. CHEN, K. DONOHUE, E. YÜCESAN, AND J. LIN, *Optimal computing budget allocation for Monte Carlo simulation with application to product design*, Simul. Model. Practice Theory, 11 (2003), pp. 57–74.
- [7] C. CHEN, J. LIN, E. YÜCESAN, AND S. CHICK, *Simulation budget allocation for further enhancing the efficiency of ordinal optimization*, Discrete Event Dyn. Syst., 10 (2000), pp. 251–270.
- [8] H. CHEN, C. CHEN, AND E. YÜCESAN, *Computing efforts allocation for ordinal optimization and discrete event simulation*, IEEE Trans. Automat. Control, 45 (2000), pp. 960–964.
- [9] H. CHEN, L. DAI, C. CHEN, AND E. YÜCESAN, *New development of optimal computing budget allocation for discrete event simulation*, in Proceedings of the 29th Winter Simulation Conference, IEEE Computer Society, Washington, DC, 1997, pp. 334–341.
- [10] S. CHICK, J. BRANKE, AND C. SCHMIDT, *New myopic sequential sampling procedures*, INFORMS J. Computing, submitted.
- [11] S. CHICK AND K. INOUE, *New procedures to select the best simulated system using common random numbers*, Manage. Sci., 47 (2001), pp. 1133–1149.
- [12] S. CHICK AND K. INOUE, *New two-stage and sequential procedures for selecting the best simulated system*, Oper. Res., 49 (2001), pp. 732–743.
- [13] C. CLARK, *The greatest of a finite set of random variables*, Oper. Res., 9 (1961), pp. 145–163.
- [14] M. C. FU, J.-Q. HU, C.-H. CHEN, AND X. XIONG, *Simulation allocation for determining the best design in the presence of correlated sampling*, INFORMS J. Comput., 19 (2007), pp. 101–111.
- [15] S. GUPTA AND K. MIESCKE, *Bayesian look ahead one stage sampling allocations for selecting the largest normal mean*, Statist. Papers, 35 (1994), pp. 169–177.
- [16] S. GUPTA AND K. MIESCKE, *Bayesian look ahead one-stage sampling allocations for selection of the best population*, J. Statist. Plann. Inference, 54 (1996), pp. 229–244.
- [17] M. HARTMANN, *An improvement on Paulson's procedure for selecting the population with the largest mean from k normal populations with a common unknown variance*, Sequential Anal., 10 (1991), pp. 1–16.
- [18] D. HE, S. CHICK, AND C. CHEN, *Opportunity cost and OCBA selection procedures in ordinal optimization for a fixed number of alternative systems*, IEEE Trans. Systems Man Cybernetics, Part C: Applications and Reviews, 37 (2007), pp. 951–961.
- [19] L. P. KAEHLING, *Learning in Embedded Systems*, MIT Press, Cambridge, MA, 1993.
- [20] O. KALLENBERG, *Foundations of Modern Probability*, Springer, New York, 1997.
- [21] S. KIM AND B. NELSON, *Selecting the best system*, in Simulation, Handbooks Oper. Res. Management Sci. 13, North-Holland, Amsterdam, 2006, pp. 501–534.
- [22] S. KIM AND B. NELSON, *On the asymptotic validity of fully sequential selection procedures for steady-state simulation*, Oper. Res., 54 (2006), pp. 475–488.
- [23] S.-H. KIM AND B. L. NELSON, *A fully sequential procedure for indifference-zone selection in simulation*, ACM Trans. Model. Comput. Simul., 11 (2001), pp. 251–273.
- [24] B. NELSON, J. SWANN, D. GOLDSMAN, AND W. SONG, *Simple procedures for selecting the best simulated system when the number of alternatives is large*, Oper. Res., 49 (2001), pp. 950–963.
- [25] E. PAULSON, *A sequential procedure for selecting the population with the largest mean from k normal populations*, Ann. Math. Statist., 35 (1964), pp. 174–180.

- [26] E. PAULSON, *Sequential procedures for selecting the best one of  $K$  Koopman-Darmois populations*, Sequential Anal., 13 (1994), pp. 207–220.
- [27] Y. RINOTT, *On two-stage selection procedures and related probability-inequalities*, Comm. Statist. A—Theory Methods, 7 (1978), pp. 799–811.
- [28] S. SINGH, T. JAAKKOLA, M. LITTMAN, AND C. SZEPESVARI, *Convergence results for single-step on-policy reinforcement-learning algorithms*, Mach. Learn., 38 (2000), pp. 287–308.

## KERNEL DENSITY ESTIMATION AND GOODNESS-OF-FIT TEST IN ADAPTIVE TRACKING\*

BERNARD BERCU<sup>†</sup> AND BRUNO PORTIER<sup>‡</sup>

**Abstract.** We investigate the asymptotic properties of a recursive kernel density estimator of the driven noise of multivariate ARMAX models in adaptive tracking. We provide an almost sure pointwise and uniform strong law of large numbers as well as a pointwise and multivariate central limit theorem. We also carry out a goodness-of-fit test together with some simulation experiments.

**Key words.** adaptive control, kernel density estimation, goodness-of-fit test

**AMS subject classifications.** 93C40, 62G07, 62G10

**DOI.** 10.1137/070694739

**1. Introduction.** Since the pioneer work of Aström and Wittenmark [1], a wide range of literature is available on parametric estimation and adaptive tracking for linear regression models [4], [5], [6] [9], [13], [14], [15], [16]. However, only a few references may be found on nonparametric estimation in adaptive tracking [20], [21], [22], [25]. Our goal is to investigate the asymptotic properties of a kernel density estimator associated with the driven noise of a linear regression in adaptive tracking and to carry out a goodness-of-fit test. Consider the multivariate ARMAX model of order  $(p, q, r)$  given, for all  $n \geq 0$ , by

$$(1.1) \quad A(R)X_n = B(R)U_n + C(R)\varepsilon_n,$$

where  $X_n, U_n$ , and  $\varepsilon_n$  are the  $d$ -dimensional system output, input, and driven noise, respectively. Denote by  $R$  the shift-back operator and set

$$\begin{aligned} A(R) &= I_d - A_1R - \cdots - A_pR^p, \\ B(R) &= B_1R + \cdots + B_qR^q, \\ C(R) &= I_d + C_1R + \cdots + C_rR^r, \end{aligned}$$

where  $A_i, B_j$ , and  $C_k$  are unknown matrices and  $I_d$  is the identity matrix of order  $d$ . For the sake of simplicity, we shall assume that the high frequency gain matrix  $B_1$  is known with  $B_1 = I_d$ . Hence, the unknown parameter of the model is given by

$$\theta^t = (A_1, \dots, A_p, B_2, \dots, B_q, C_1, \dots, C_r).$$

Relation (1.1) can be rewritten as

$$(1.2) \quad X_{n+1} = \theta^t \Psi_n + U_n + \varepsilon_{n+1},$$

---

\*Received by the editors June 18, 2007; accepted for publication (in revised form) April 28, 2008; published electronically September 8, 2008.

<http://www.siam.org/journals/sicon/47-5/69473.html>

<sup>†</sup>Institut de Mathématiques de Bordeaux, Université Bordeaux 1, UMR 5251 and INRIA Bordeaux Sud-Ouest, Team CQFD, 351 cours de la libération, 33405 Talence cedex, France (Bernard.Bercu@math.u-bordeaux1.fr).

<sup>‡</sup>Département de Génie Mathématiques, Laboratoire de Mathématique, INSA de Rouen, LMI-EA 3226, place Emile Blondel, BP 08, 76131 Mont-Saint-Aignan cedex, France (Bruno.Portier@insa-rouen.fr).

where

$$\Psi_n^t = (X_n^t, \dots, X_{n-p+1}^t, U_{n-1}^t, \dots, U_{n-q+1}^t, \varepsilon_n^t, \dots, \varepsilon_{n-r+1}^t).$$

The most common way for estimating  $\theta$  is to make use of the extended least-squares (ELS) algorithm given, for all  $n \geq 0$ , by

$$\begin{aligned}\hat{\theta}_{n+1} &= \hat{\theta}_n + S_n^{-1} \Phi_n (X_{n+1} - U_n - \hat{\theta}_n^t \Phi_n)^t, \\ \hat{\varepsilon}_{n+1} &= X_{n+1} - U_n - \hat{\theta}_n^t \Phi_n, \\ \Phi_n^t &= (X_n^t, \dots, X_{n-p+1}^t, U_{n-1}^t, \dots, U_{n-q+1}^t, \hat{\varepsilon}_n^t, \dots, \hat{\varepsilon}_{n-r+1}^t),\end{aligned}$$

where the initial value  $\hat{\theta}_0$  may be arbitrarily chosen. Moreover,

$$S_n = \sum_{i=0}^n \Phi_i \Phi_i^t + S,$$

where  $S$  is a positive definite and deterministic matrix introduced in order to avoid useless invertibility assumption. The crucial role played by the control  $U_n$  is to regulate the dynamic of the process  $(X_n)$  by forcing  $X_n$  to track step-by-step a bounded predictable reference trajectory  $x_n^*$ . Via the certainty equivalence principle [1], the adaptive tracking control  $U_n$  is given, for all  $n \geq 0$ , by

$$(1.3) \quad U_n = x_{n+1}^* - \hat{\theta}_n^t \Phi_n.$$

By substituting (1.3) into (1.2), we obtain the closed-loop system

$$(1.4) \quad X_{n+1} - x_{n+1}^* = \pi_n + \varepsilon_{n+1},$$

where

$$\pi_n = \theta^t \Psi_n - \hat{\theta}_n^t \Phi_n$$

is the prediction error at time  $n$ . In the following, we shall assume that the driven noise  $(\varepsilon_n)$  is a sequence of centered independent and identically distributed random vectors with positive definite covariance matrix  $\Gamma$  and unknown probability density function denoted by  $f$ .

The purpose of this paper is to study the asymptotic properties of a kernel density estimator (KDE) of  $f$ . Since the pioneer works of Parzen [18] and Rosenblatt [23], the asymptotic properties of such a kernel estimator have been widely investigated in the context of independent and identically distributed random variables as well as for mixing random variables. We refer the reader to [10], [11], [24] for some excellent books on density estimation for stationary processes. Although the stability of ARMAX models in adaptive tracking has been deeply investigated in the literature [9], [12], one can realize that kernel density estimation results are not available in adaptive tracking.

Let us now define our KDE of  $f$  associated with model (1.2). When the sequence  $(\varepsilon_n)$  is observable, the traditional Parzen–Rosenblatt KDE of  $f$  is given, for all  $x \in \mathbb{R}^d$  and  $n \geq 1$ , by

$$f_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^n K\left(\frac{\varepsilon_i - x}{h_n}\right),$$

where the kernel  $K$  is a chosen density function and the bandwidth  $(h_n)$  is a sequence of positive real numbers decreasing to zero. In our situation, the sequence  $(\varepsilon_n)$  is of course unobservable. However, when the tracking objective is fulfilled, the prediction error  $\pi_n$  is as close as possible to zero. Consequently, via (1.4), we can choose  $X_n - x_n^*$  as a predictor of  $\varepsilon_n$ . Moreover, since we are in an adaptive tracking framework, it is more suitable to make use of a recursive kernel density estimator (RKDE) of  $f$  given, for all  $x \in \mathbb{R}^d$  and  $n \geq 1$ , by

$$(1.5) \quad \hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i^d} K\left(\frac{X_i - x_i^* - x}{h_i}\right).$$

Our purpose is first to show that  $\hat{f}_n$  behaves pretty well as a RKDE of  $f$  in adaptive tracking and second to carry out a goodness-of-fit test for  $f$  based on  $\hat{f}_n$ . Such a goodness-of-fit test is very popular in time series, in particular, for testing the normality hypothesis. For independent and identically distributed samples, we can mention the well-known and very popular Kolmogorov–Smirnov and Cramér–Von Mises statistical tests based on the empirical distribution function as well as the Bickel and Rosenblatt test [7] based on a KDE. Recently, for stationary autoregressive processes, several authors have proposed goodness-of-fit tests based on KDE [2], [17]. However, to the best of our knowledge, no work is concerned with asymptotic properties of KDE in adaptive tracking.

The paper is organized as follows. Section 2 is devoted to the asymptotic behavior of  $\hat{f}_n$ . We establish the almost sure pointwise and uniform convergence of  $\hat{f}_n$  to  $f$  as well as a pointwise law of iterated logarithm (LIL) and a pointwise multivariate central limit theorem (CLT). Section 3 is concerned with the goodness-of-fit test for  $f$ . Finally, some simulation experiments are given in section 4. All technical proofs are postponed in appendices.

**2. Main results.** In the following, we shall assume that the kernel  $K$  is a non-negative function, bounded with compact support, such that

$$\int_{\mathbb{R}^d} K(t) dt = 1 \quad \text{and} \quad \int_{\mathbb{R}^d} K^2(t) dt = \tau^2.$$

For example, for some  $s > 0$  and some known positive constants  $a_s, b_s, c_s$ , one can make use of the uniform kernel on the sphere of  $\mathbb{R}^d$  with radius  $s$ ,  $K(t) = a_s \mathbb{I}_{(\|t\| \leq s)}$ , the Epanechnikov kernel with scaling factor  $s$ ,  $K(t) = b_s (1 - \|t\|^2/s^2) \mathbb{I}_{(\|t\| \leq s)}$ , and the Gaussian kernel with truncation level  $s$ ,  $K(t) = c_s \exp(-\|t\|^2/2) \mathbb{I}_{(\|t\| \leq s)}$ .

Moreover, we shall assume that the bandwidth  $(h_n)$  is a sequence of positive real numbers, decreasing to zero, such that  $nh_n^d$  tends to infinity and

$$\sum_{i=1}^n h_i = O(nh_n).$$

This mild condition, due to the recursive form of  $\hat{f}_n$ , is clearly not restrictive. For example, one can choose  $h_n = n^{-\alpha}$  with  $\alpha \in ]0, 1/d[$ .

Furthermore, we shall also make use of the classical assumptions of causality and passivity as well as the traditional smoothness hypothesis on the probability density function  $f$ .

**Causality** [A1]. For all  $z \in \mathbb{C}$  with  $|z| \leq 1$ ,  $\det(z^{-1}B(z)) \neq 0$ .

**Passivity** [A2]. For all  $z \in \mathbb{C}$  with  $|z| = 1$ ,  $\det(C(z)) \neq 0$  and  $C^{-1}(z) > \frac{1}{2}I_d$ .

**Density** [A3]. The function  $f$  is positive and differentiable with bounded gradient.

We shall now propose several asymptotic results for the RKDE  $\hat{f}_n$  of  $f$ , the first one dealing with the almost sure convergence properties of  $\hat{f}_n$ .

**THEOREM 2.1.** Assume that [A1] to [A3] hold and suppose that  $(\varepsilon_n)$  has finite moment of order  $a > 2$ . In addition, assume that  $nh_n^d$  tends to infinity faster than  $(\log n)^2$ . Then, for any  $x \in \mathbb{R}^d$ ,  $\hat{f}_n(x)$  converges a.s. to  $f(x)$ . As soon as the bandwidth  $(h_n)$  satisfies  $\max(nh_n^{d+2}, n^b h_n^d) = o(\log \log n)$  for some  $b \in ]2/a, 1[$ , we also have

$$(2.1) \quad \limsup_{n \rightarrow \infty} \left( \frac{nh_n^d}{2\tau^2 \|f\|_\infty \log \log n} \right)^{1/2} \left| \hat{f}_n(x) - f(x) \right| \leq 1 \quad \text{a.s.}$$

Moreover, assume that the kernel  $K$  is Lipschitz and that the bandwidth  $(h_n)$  is given by  $h_n = n^{-\alpha}$  with  $\alpha \in ]0, 1/d[$ . Then,  $\hat{f}_n$  converges a.s. to  $f$ , uniformly on all compact sets of  $\mathbb{R}^d$  and, for any  $\beta \in ](1+c)/2, 1[$  with  $c = \max(b, \alpha d)$ ,

$$(2.2) \quad \sup_{x \in \mathbb{R}^d} \left| \hat{f}_n(x) - f(x) \right| = O(n^{-\alpha}) + o(n^{\beta-1}) \quad \text{a.s.}$$

*Proof.* The proof is given in Appendix A.  $\square$

**Remark 1.** The bandwidth condition associated with the almost sure pointwise convergence is clearly not restrictive and it is satisfied when  $h_n = n^{-\alpha}$  with  $\alpha \in ]0, 1/d[$ . In this particular case, the bandwidth condition required for the LIL is obviously satisfied as soon as  $\alpha \in ]\delta, 1/d[$  with  $\delta = \max(1/(d+2), b/d)$ .

**Remark 2.** In the particular case of controlled autoregressive process

$$(2.3) \quad X_{n+1} = A_1 X_n + \cdots + A_p X_{n-p+1} + U_n + \varepsilon_{n+1},$$

the assumptions [A1] and [A2] are clearly useless and the associated prediction errors sequence  $(\pi_n)$  satisfies (see, e.g., [5])

$$(2.4) \quad \sum_{i=0}^n \|\pi_i\|^2 = O(\log n) \quad \text{a.s.}$$

Thanks to this sharp result on the sequence  $(\pi_n)$ , we only have to assume that  $\max(nh_n^{d+2}, h_n^d \log n) = o(\log \log n)$  for the LIL. This bandwidth condition is immediately satisfied when  $h_n = n^{-\alpha}$  with  $\alpha \in ]1/(d+2), 1/d[$ . Moreover, for the uniform convergence, it is only necessary to assume that  $\beta \in ](1+\alpha d)/2, 1[$ . All of the above is also true for the scalar nonlinear controlled autoregressive process

$$(2.5) \quad X_{n+1} = \theta \varphi(X_n, \dots, X_{n-p+1}) + U_n + \varepsilon_{n+1}$$

under suitable moment assumption on  $(\varepsilon_n)$  and as soon as the function  $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}$  does not increase to infinity faster than a polynomial of degree  $< 4$  [6]. We are again able to deduce such results because the associated prediction errors sequence  $(\pi_n)$  satisfies (2.4). Finally, Theorem 2.1 holds for the RKDE associated with nonlinear controlled autoregressive processes as soon as the associated prediction errors sequence  $(\pi_n)$  satisfies a stability property such as (2.4).

Our second result is a pointwise and a multivariate CLT for  $\hat{f}_n$ .

THEOREM 2.2. Assume that [A1] to [A3] hold and suppose that  $(\varepsilon_n)$  has finite moment of order  $a > 2$ . Moreover, assume that the bandwidth  $(h_n)$  satisfies  $\max(nh_n^{d+2}, n^b h_n^d) = o(1)$  for some  $b \in ]2/a, 1[$ , together with

$$(2.6) \quad \lim_{n \rightarrow \infty} \frac{h_n^d}{n} \sum_{i=1}^n h_i^{-d} = \ell_h$$

for some finite constant  $\ell_h > 0$ . Then, for any  $x \in \mathbb{R}^d$ , we have the pointwise CLT

$$(2.7) \quad G_n(x) = \sqrt{n h_n^d} \left( \hat{f}_n(x) - f(x) \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau^2 \ell_h f(x)) = G(x).$$

In addition, for  $N$  distinct points  $x_1, \dots, x_N$  of  $\mathbb{R}^d$ , we also have

$$(2.8) \quad (G_n(x_1), \dots, G_n(x_N)) \xrightarrow{\mathcal{L}} (G(x_1), \dots, G(x_N)),$$

where  $G(x_1), \dots, G(x_N)$  are independent Gaussian random variables.

*Proof.* The proof is given in Appendix B.  $\square$

*Remark 3.* Convergence (2.7) is identical to the one obtained by Duflo [12] for stationary processes. Besides, it is worthless to require the bandwidth condition (2.6) for the nonrecursive KDE of  $f$ , and  $\ell_h$  has to be replaced by 1 in (2.7). Finally, if  $h_n = n^{-\alpha}$ , it is necessary to assume that  $\alpha \in ]\delta, 1/d[$  with  $\delta = \max(1/(d+2), b/d)$  and we obviously have  $\ell_h = (1 + \alpha d)^{-1}$ . In addition, for the controlled autoregressive processes given by (2.3) or (2.5), we only have to assume that  $\alpha \in ]1/(d+2), 1/d[$ .

*Remark 4.* When the density function  $f$  belongs to  $C^2(\mathbb{R}^d)$  with a bounded second derivative and for symmetric kernel  $K$ , we can relax the bandwidth condition by  $\max(nh_n^{d+4}, n^b h_n^d) = o(1)$ .

**3. Application to a goodness-of-fit test.** We shall now propose a statistical test associated with the probability density function  $f$  based on the convergence results of section 2. We wish to test

$$\mathcal{H}_0 : \langle\langle f = f_0 \rangle\rangle \quad \text{vs} \quad \mathcal{H}_1 : \langle\langle f \neq f_0 \rangle\rangle$$

where  $f_0$  is a given probability density function. It is well known that such a goodness-of-fit test is very important and it has been widely investigated in time series analysis since the pioneer works of Kolmogorov–Smirnov and Cramér–Von Mises. Indeed, many statistical procedures require the assumption of normality for the driven white noise (see, e.g., [3] or [8]). Consequently, a goodness-of-fit test for the white noise density is of particular interest. However, no such a statistical test is available in the adaptive tracking framework, although several situations require the normality assumption on the driven white noise. Our purpose is to provide a goodness-of-fit test for  $f$  based on the RKDE  $\hat{f}_n$ . Such an approach has been already used by Bickel and Rosenblatt [7]. Indeed, for the independent and identically distributed sample, they proposed a statistical test based on the integrated quadratic deviation between the true density and a KDE of  $f$ . This approach has been extended to the scalar autoregressive framework by Lee and Na [17] and more recently by Bachmann and Dette [2]. However, due to some technical reasons, it seems impossible to extend this approach to our adaptive tracking context. Therefore, we propose a new strategy and we carry out a goodness-of-fit test for  $f$  based on the multivariate CLT for  $\hat{f}_n$  together with the LIL. Our statistical test consists of a suitably normalized sum of



the quadratic deviation between the true density and the RKDE  $\widehat{f}_n$  evaluated on  $N$  distinct points of  $\mathbb{R}^d$ . More precisely, it is defined by

$$T_n(N) = \frac{1}{\tau^2 \ell_h} \sum_{j=1}^N \frac{(\widehat{f}_n(x_j) - f_0(x_j))^2}{\widehat{f}_n(x_j)},$$

where  $x_1, \dots, x_N$  are  $N$  distinct points of  $\mathbb{R}^d$ . We shall make use of

$$\sigma^2 = \frac{1}{\tau^2 \ell_h} \sum_{j=1}^N \frac{(f(x_j) - f_0(x_j))^2}{f(x_j)} \quad \text{and} \quad \lambda^2 = \frac{1}{\tau^2 \ell_h} \sum_{j=1}^N \frac{(f^2(x_j) - f_0^2(x_j))^2}{f^3(x_j)}.$$

**THEOREM 3.1.** *Assume that [A1] to [A3] hold and suppose that  $(\varepsilon_n)$  has finite moment of order  $a > 2$ . Moreover, assume that the bandwidth  $(h_n)$  shares the same assumptions as in Theorem 2.2 and is such that  $nh_n^d$  goes to infinity faster than  $(\log n)^2$ . Then, under  $\mathcal{H}_0$ ,*

$$(3.1) \quad nh_n^d T_n(N) \xrightarrow{\mathcal{L}} \chi^2(N).$$

Moreover, under  $\mathcal{H}_1$  and if one can find  $x \in \{x_1, x_2, \dots, x_N\}$  such that  $f(x) \neq f_0(x)$ , then  $T_n(N)$  converges a.s. towards  $\sigma^2$ . In addition, we also have

$$(3.2) \quad \sqrt{nh_n^d} (T_n(N) - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda^2).$$

*Remark 5.* According to these asymptotic results, it is possible to construct a goodness-of-fit test associated with  $f$ . On the one hand, under the null hypothesis  $\mathcal{H}_0$ , we can approximate for  $n$  large enough the distribution of  $nh_n^d T_n(N)$  by a  $\chi^2(N)$  one. On the other hand, under the alternative hypothesis  $\mathcal{H}_1$ , if  $\sigma^2$  is positive,  $nh_n^d T_n(N)$  goes a.s. to infinity, which guarantees that the asymptotic power of our test is equal to 1. From a practical point of view, the null hypothesis  $\mathcal{H}_0$  will be rejected at level  $\delta$  whenever  $nh_n^d T_n(N) > a_\delta$  where  $a_\delta$  stands for the  $(1 - \delta)$  quantile of the  $\chi^2(N)$  distribution. Finally, one can observe that the weak convergence (3.2) allows us to evaluate the probability of the type II error of our test.

*Remark 6.* It is also possible to make use of the test statistic  $Z_n(N)$  defined by

$$Z_n(N) = \frac{1}{\tau^2 \ell_h} \sum_{j=1}^N \frac{(\widehat{f}_n(x_j) - f_0(x_j))^2}{f_0(x_j)}.$$

In that case, Theorem 3.1 holds with

$$\sigma^2 = \frac{1}{\tau^2 \ell_h} \sum_{j=1}^N \frac{(f(x_j) - f_0(x_j))^2}{f_0(x_j)} \quad \text{and} \quad \lambda^2 = \frac{4}{\tau^2 \ell_h} \sum_{j=1}^N \frac{(f(x_j) - f_0(x_j))^2 f(x_j)}{f_0^2(x_j)}.$$

This statistical test should improve the empirical level under  $\mathcal{H}_0$ , but it should certainly degrade the empirical power under  $\mathcal{H}_1$ . Nevertheless, it is easier to compute than  $T_n(N)$  because it allows one to avoid the division by  $\widehat{f}_n(x_j)$ , which can be equal to zero due to the use of a compactly supported kernel.

*Proof.* The proof is straightforward by use of Theorem 2.1 together with Theorem 2.2. As a matter of fact, we have the decomposition

$$(3.3) \quad T_n(N) - \sigma^2 = A_n + B_n,$$

where

$$A_n = \frac{1}{\tau^2 \ell_h} \sum_{j=1}^N \frac{\left( \widehat{f}_n(x_j) - f(x_j) \right)^2}{\widehat{f}_n(x_j)},$$

$$B_n = \frac{1}{\tau^2 \ell_h} \sum_{j=1}^N \frac{\left( \widehat{f}_n(x_j) - f(x_j) \right)}{\widehat{f}_n(x_j)} \frac{(f^2(x_j) - f_0^2(x_j))}{f(x_j)}.$$

We can deduce from (2.8) and the pointwise almost sure convergence of  $\widehat{f}_n$  to  $f$  that

$$(3.4) \quad \sqrt{\frac{n h_n^d}{\tau^2 \ell_h}} \left( \frac{\widehat{f}_n(x_1) - f(x_1)}{\sqrt{\widehat{f}_n(x_1)}}, \dots, \frac{\widehat{f}_n(x_N) - f(x_N)}{\sqrt{\widehat{f}_n(x_N)}} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_N),$$

where  $I_N$  stands for the identity matrix of order  $N$ . Hence, it immediately follows from (3.4) that

$$(3.5) \quad n h_n^d A_n \xrightarrow{\mathcal{L}} \chi^2(N).$$

Consequently, we clearly obtain (3.1) from (3.3) together with (3.5) since, under the null hypothesis  $\mathcal{H}_0$ ,  $\sigma^2$  and  $B_n$  vanish. Under the alternative hypothesis  $\mathcal{H}_1$ , it is straightforward to see that  $T_n(N)$  converges a.s. towards  $\sigma^2$  via the almost sure pointwise convergence of  $\widehat{f}_n$  to  $f$ . Only convergence (3.2) remains to be proven. On the one hand, by the pointwise LIL, we infer that

$$|A_n| = O\left(\frac{\log \log n}{n h_n^d}\right) \quad \text{a.s.},$$

which implies that

$$(3.6) \quad \sqrt{n h_n^d} A_n = o(1) \quad \text{a.s.}$$

as  $n h_n^d$  goes to infinity faster than  $(\log n)^2$ . On the other hand, we can deduce from (3.4) that

$$(3.7) \quad \sqrt{n h_n^d} B_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda^2).$$

Finally, convergence (3.2) immediately follows from the conjunction of (3.3), (3.6), and (3.7), which completes the proof of Theorem 3.1.  $\square$

**4. Simulation experiments.** In this section, we investigate the finite sample properties of our statistical test under both hypotheses  $\mathcal{H}_0$  and  $\mathcal{H}_1$  without some bootstrap procedure as is usual in this context of nonparametric tests. Since it has never been experimented, we shall not restrict ourselves to models of form (1.2), but

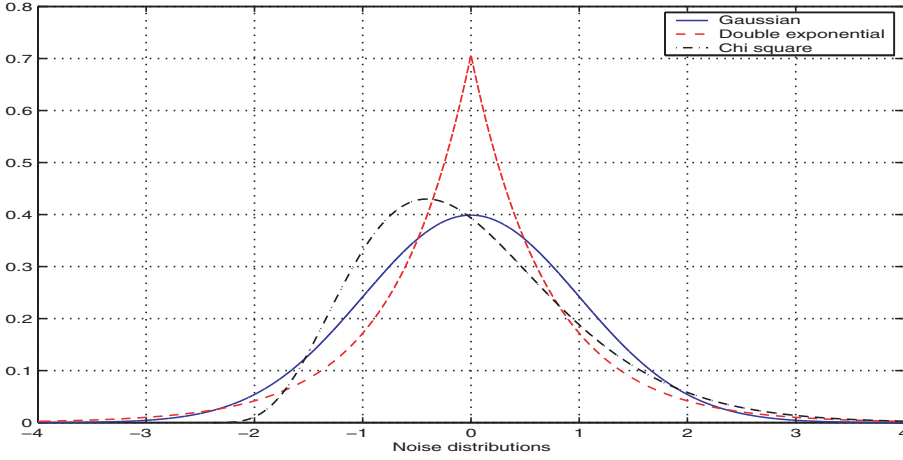


FIG. 4.1.

we will also consider some closely related stationary models. Our goal is to show that our statistical test behaves pretty well in many different situations. The different models that we will study are given as follows.

$$(WN) \quad X_n = \varepsilon_n,$$

$$(AR) \quad X_{n+1} = \theta X_n + \varepsilon_{n+1},$$

$$(ARX) \quad X_{n+1} = \theta X_n + U_n + \varepsilon_{n+1},$$

$$(NARX) \quad X_{n+1} = \theta X_n^2 + U_n + \varepsilon_{n+1},$$

where  $(\varepsilon_n)$  is a sequence of centered independent and identically distributed random variables with probability density function  $f$ . We choose  $\theta = 7/10$ ,  $\theta = 2$ , and  $\theta = 1/2$  for the AR, ARX, and NARX models, respectively. We consider three choices of noise distributions, given in Figure 4.1, that we combine two-by-two in order to study the performances of our statistical test under both  $\mathcal{H}_0$  and  $\mathcal{H}_1$ . The first one is the standard normal distribution

$$f_0(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

The second one is the normalized double exponential distribution

$$f_1(x) = \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}|x|\right).$$

The last one is the standardized chi-square distribution with twelve degrees of freedom

$$f_2(x) = \frac{9}{5}(x + \sqrt{6})^5 \exp\left(-\sqrt{6}(x + \sqrt{6})\right) \mathbb{1}_{(x \geq -\sqrt{6})}.$$

For AR, ARX, and NARX models, we estimate the unknown parameter  $\theta$  by use of the standard least-squares estimator  $\hat{\theta}_n$ . For the AR model, the probability density function  $f$  is estimated using the RKDE given by (1.5) where  $X_n - x_n^*$  is replaced

by  $X_n - \hat{\theta}_n X_{n-1}$ . For ARX and NARX models, the adaptive control  $U_n$  is given by  $U_n = -\hat{\theta}_n X_n$  and  $U_n = -\hat{\theta}_n X_n^2$ , respectively.

For each model and each test of  $\mathcal{H}_0$  against  $\mathcal{H}_1$ , we base our estimations on 800 independent realizations of sample sizes  $n = 200, 500$ , and  $1000$ . We are interested in the empirical level under  $\mathcal{H}_0$  to be compared with the theoretical level equal to 5% and the empirical power under  $\mathcal{H}_1$ , as well as the closeness between the simulated distribution of our statistical test and the corresponding theoretical distribution. The implementation of our statistic test  $T_n(N)$  requires the choice of design points together with the specification of a bandwidth and a kernel for the RKDE  $\hat{f}_n$ . The RKDE  $\hat{f}_n$  is constructed by use of the Epanechnikov kernel

$$K(t) = \frac{3}{4} (1 - t^2) \mathbb{I}_{(|t| \leq 1)}$$

and the bandwidth  $h_n = n^{-1/3}$ . For the denominator of  $T_n(N)$ , we use the Gaussian kernel and the usual bandwidth  $h_n = n^{-1/5}$ . Via this choice, we avoid a possible division by zero and we provide a smoother version for the estimation of  $f$ . Finally, for ARX and NARX models, we use a short learning period of  $\tau = 100$  time steps. This learning period allows us to forget the transitory phase.

For the choice of  $N$  and the points  $x_1, \dots, x_N$ , we use the design points selection rule proposed by Poggi and Portier and fully described in [19]. More precisely, we proceed as follows. Starting from an estimate of the distribution of the driven noise, we choose  $N$  equidistant points  $x_1, \dots, x_N$  so that the density at those points is not too small and in such a way that they are sufficiently distant to ensure sufficient accuracy in the use of the multivariate CLT. Typically, we choose points  $x_1, \dots, x_N$  such that the distance between two neighboring points is  $4n^{-1/3}$ . This last condition allows us to make sure that the independence property in the multivariate CLT, which holds asymptotically, remains true for small to moderate sample sizes. We take  $N = 8, 13$ , and  $22$  equidistant points for sample sizes  $n = 200, 500$ , and  $1000$ , respectively. It should be noted that only a few number of points is needed to make a decision.

In the sequel, the abbreviations  $\mathcal{G}f_0$ ,  $\mathcal{G}f_1$ , and  $\mathcal{G}f_2$  mean that the driven noise ( $\varepsilon_n$ ) is generated with the normal  $f_0$  distribution, the double exponential  $f_1$  distribution, and the chi-square  $f_2$  distribution, respectively, while  $\mathcal{H}f_0$ ,  $\mathcal{H}f_1$ , and  $\mathcal{H}f_2$  mean that we are testing the assumptions  $\mathcal{H}_0 : \langle\langle f = f_0 \rangle\rangle$ ,  $\mathcal{H}_0 : \langle\langle f = f_1 \rangle\rangle$ , and  $\mathcal{H}_0 : \langle\langle f = f_2 \rangle\rangle$ , respectively. Finally, as we have chosen a test level  $\alpha = 5\%$  and we have generated 800 trials, the Kolmogorov–Smirnov fitting statistic in *italic* has to be compared with the critical value 0.048.

We shall now comment on the test results contained in Tables 4.1–4.4. First of all, one can verify that our statistical test behaves pretty well under  $\mathcal{H}_0$ . Indeed, for each model and each noise distribution, the empirical level is close to the 5% theoretical value level as one can realize with the values in **bold**. In addition, the simulated distribution of  $n^{2/3}T_n(N)$  is close to the  $\chi^2(N)$  distribution as one can observe with the values in *italic* of the Kolmogorov–Smirnov fitting statistic to be compared with the critical value at 5% equal to 0.048. Next, one can verify that the empirical power increases with the sample size, from 20% to 40% for  $n = 200$ , to 96% to 100% for  $n = 1000$ ; it is more difficult to decide between  $f_0$  and  $f_2$  than between  $f_1$  and  $f_2$ , which is the easier situation. Finally, if one superimposes the four tables, one can observe that the results for the different models are almost the same. In conclusion, our statistical test behaves pretty well for small to moderate sample sizes and for a large class of models.

TABLE 4.1

**WN model.** Results under  $\mathcal{H}_0$  and  $\mathcal{H}_1$  with test level 5%. Empirical level in bold and percentage of correct decisions.

	$n = 200, N = 8$			$n = 500, N = 13$			$n = 1000, N = 22$		
	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$
$\mathcal{G}f_0$	<b>4.2%</b> 0.035	35.7%	26.2%	<b>5.3%</b> 0.029	84.1%	70%	<b>5.2%</b> 0.024	99.8%	98.6%
$\mathcal{G}f_1$	49%	<b>5.3%</b> 0.047	74.1%	91.2%	<b>5.1%</b> 0.041	99.3%	100%	<b>4.2%</b> 0.030	100%
$\mathcal{G}f_2$	19.2%	53.5%	<b>4.2%</b> 0.047	60%	97.3%	<b>4.7%</b> 0.031	96.7%	100%	<b>4.5%</b> 0.009

TABLE 4.2

**AR model.** Results under  $\mathcal{H}_0$  and  $\mathcal{H}_1$  with test level 5%. Empirical level in bold and percentage of correct decisions.

	$n = 200, N = 8$			$n = 500, N = 13$			$n = 1000, N = 22$		
	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$
$\mathcal{G}f_0$	<b>4.5%</b> 0.045	31.2%	25.6%	<b>4.8%</b> 0.014	82%	65%	<b>3.7%</b> 0.023	99.8%	98.8%
$\mathcal{G}f_1$	49.7%	<b>5.7%</b> 0.032	73.1%	90.5%	<b>5%</b> 0.014	99.1%	100%	<b>4.8%</b> 0.019	100%
$\mathcal{G}f_2$	19.3%	54.6%	<b>3.7%</b> 0.045	62%	96.6%	<b>3.5%</b> 0.022	96.6%	100%	<b>3.8%</b> 0.013

TABLE 4.3

**ARX model.** Results under  $\mathcal{H}_0$  and  $\mathcal{H}_1$  with test level 5% and learning period  $\tau = 100$ . Empirical level in bold and percentage of correct decisions.

	$n = 200, N = 8$			$n = 500, N = 13$			$n = 1000, N = 22$		
	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$
$\mathcal{G}f_0$	<b>3.8%</b> 0.042	35.7%	28%	<b>4.2%</b> 0.029	81.5%	66%	<b>3.7%</b> 0.018	99.7%	98.2%
$\mathcal{G}f_1$	45.8%	<b>5.5%</b> 0.053	71.5%	87.5%	<b>4.7%</b> 0.021	99.3%	100%	<b>5%</b> 0.022	100%
$\mathcal{G}f_2$	21.2%	54.5%	<b>3.2%</b> 0.029	62%	95.6%	<b>2.5%</b> 0.040	96.7%	100%	<b>5.1%</b> 0.029

TABLE 4.4

**NARX model.** Results under  $\mathcal{H}_0$  and  $\mathcal{H}_1$  with test level 5% and learning period  $\tau = 100$ . Empirical level in bold and percentage of correct decisions.

	$n = 200, N = 8$			$n = 500, N = 13$			$n = 1000, N = 22$		
	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$	$\mathcal{H}f_0$	$\mathcal{H}f_1$	$\mathcal{H}f_2$
$\mathcal{G}f_0$	<b>3%</b> 0.037	37.1%	28.5%	<b>4.8%</b> 0.029	83.5%	68.2%	<b>4.3%</b> 0.037	99.5%	98.6%
$\mathcal{G}f_1$	44.6%	<b>5.2%</b> 0.021	72%	89.8%	<b>4.5%</b> 0.022	99.2%	100%	<b>5.1%</b> 0.017	100%
$\mathcal{G}f_2$	19.8%	58.3%	<b>3.7%</b> 0.021	63.2%	95.5%	<b>4.7%</b> 0.05	97.2%	100%	<b>5%</b> 0.039

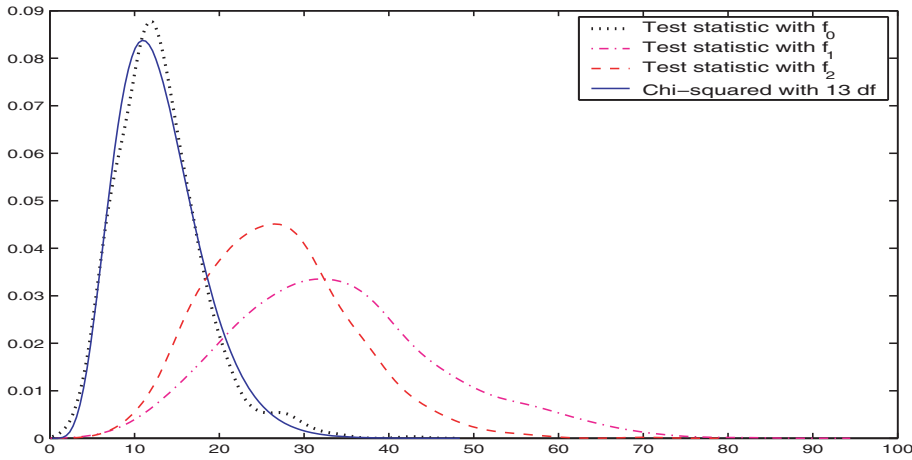


FIG. 4.2.

Figure 4.2 illustrates the empirical level and power of our test for the NARX model. We base our estimation on 800 trials of sample size  $n = 500$  with  $N = 13$  equidistant points. The driven noise  $(\varepsilon_n)$  is generated with the normal distribution  $f_0$ , and we are successively testing the assumptions  $\mathcal{H}f_0$ ,  $\mathcal{H}f_1$ , and  $\mathcal{H}f_2$ . On the one hand, when we test the hypothesis  $\mathcal{H}f_0$ , we can observe that the distribution of our statistical test  $n^{2/3}T_n(N)$  is superimposed with the  $\chi^2(N)$  one. It clearly illustrates the good approximation of the distribution of  $n^{2/3}T_n(N)$  by a  $\chi^2(N)$  one under  $\mathcal{H}f_0$  for moderate sample size. On the other hand, when we test the hypothesis  $\mathcal{H}f_1$  as well as  $\mathcal{H}f_2$ , we can effectively see that the distribution of our statistical test  $n^{2/3}T_n(N)$  is totally different from the  $\chi^2(N)$  one. Consequently, the power of separation of our statistical test is clearly significant.

**Appendix A.** This appendix is devoted to the proof of Theorem 2.1. In order to prove the asymptotic properties of our RKDE  $\widehat{f}_n$  of  $f$ , we are led to introduce the martingale  $(M_n)$  associated with the sequence  $(\widehat{f}_n)$ . To be more precise, we infer from (1.5) that for all  $x \in \mathbb{R}^d$  and  $n \geq 1$ ,

$$(A.1) \quad n\left(\widehat{f}_n(x) - f(x)\right) = M_n(x) + R_n(x)$$

with

$$(A.2) \quad M_n(x) = \sum_{i=1}^n \left( K_i(X_i - x_i^* - x) - \mathbb{E}[K_i(X_i - x_i^* - x) | \mathcal{F}_{i-1}] \right),$$

$$(A.3) \quad R_n(x) = \sum_{i=1}^n \mathbb{E}[K_i(X_i - x_i^* - x) | \mathcal{F}_{i-1}] - n f(x),$$

where, for all  $y \in \mathbb{R}^d$ ,  $K_n(y) = h_n^{-d}K(h_n^{-1}y)$  and  $\mathcal{F}_n$  denotes the  $\sigma$ -algebra of the events occuring up to time  $n$ . The almost sure properties of  $(M_n)$  are given by the two following lemmas.

LEMMA A.1. Assume that  $nh_n^d$  tends to infinity faster than  $(\log n)^2$ . Then, for any  $x \in \mathbb{R}^d$ , we have  $M_n(x) = o(n)$  a.s. More precisely,

$$(A.4) \quad \limsup_{n \rightarrow \infty} \frac{|M_n(x)|}{\sqrt{2\tau^2 \|f\|_\infty n h_n^{-d} \log \log n}} \leq 1 \quad a.s.$$

*Proof.* For any  $x \in \mathbb{R}^d$ ,  $(M_n(x))$  is a square integrable real martingale. In addition, its increasing process  $(\langle M(x) \rangle_n)$  satisfies  $\langle M(x) \rangle_n = O(nh_n^{-d})$ . As a matter of fact, for all  $x \in \mathbb{R}^d$ ,

$$\langle M(x) \rangle_n = \sum_{i=1}^n \mathbb{E} [K_i^2 (X_i - x_i^* - x) | \mathcal{F}_{i-1}] - \sum_{i=1}^n (\mathbb{E} [K_i (X_i - x_i^* - x) | \mathcal{F}_{i-1}])^2.$$

Consequently, we deduce from (1.4) that for all  $x \in \mathbb{R}^d$

$$(A.5) \quad \langle M(x) \rangle_n \leq \sum_{i=1}^n h_i^{-2d} \int_{\mathbb{R}^d} K^2(h_i^{-1}(\pi_{i-1} + s - x)) f(s) ds.$$

Via the change of variables  $t = h_i^{-1}(\pi_{i-1} + s - x)$  into (A.5), we find that

$$\langle M(x) \rangle_n \leq \sum_{i=1}^n h_i^{-d} \int_{\mathbb{R}^d} K^2(t) f(h_i t + x - \pi_{i-1}) dt \leq \tau^2 \|f\|_\infty \sum_{i=1}^n h_i^{-d}.$$

Therefore, as  $(h_n)$  is decreasing,  $\langle M(x) \rangle_n = O(nh_n^{-d})$ . Hence, it follows from the strong law of large numbers for martingales (see, e.g., [12], Theorem 1.3.15, p. 20) that for all  $\gamma > 0$ ,

$$|M_n(x)|^2 = o(nh_n^{-d}(\log n)^{1+\gamma}) \quad a.s.,$$

which ensures that  $M_n(x) = o(n)$  a.s. since  $nh_n^d$  tends to infinity faster than  $(\log n)^2$ . Furthermore, for any  $x \in \mathbb{R}^d$ ,  $|M_n(x) - M_{n-1}(x)| \leq 2h_n^{-d} \|K\|_\infty$  which clearly implies that

$$|M_n(x) - M_{n-1}(x)| \leq C_n \sqrt{\frac{nh_n^{-d}}{\log \log n}},$$

where  $(C_n)$  is a deterministic sequence which tends to zero. Finally, we immediately obtain (A.4) from the upper bound in the law of iterated logarithm for martingales (see, e.g., [12], Theorem 6.4.24, p. 209).  $\square$

LEMMA A.2. Assume that the kernel  $K$  is Lipschitz and that the bandwidth  $(h_n)$  is given by  $h_n = n^{-\alpha}$  with  $\alpha \in ]0, 1/d[$ . Then, for any constants  $A > 0$  and  $\gamma > 0$ , we have the expanded uniform strong law

$$(A.6) \quad \sup_{\|x\| \leq An^\gamma} |M_n(x)| = o(n^\beta) \quad a.s.,$$

where  $\beta \in ](1 + \alpha d)/2, 1[$ .

*Proof.* Result (A.6) follows from the expanded uniform strong law for martingales given by Theorem 6.4.34, p. 220 of [12]. First of all, for all  $x \in \mathbb{R}^d$ , set  $\Delta M_n(x) = M_n(x) - M_{n-1}(x)$ . We already saw in the proof of Lemma A.1 that there exists two

positive constants  $a, b$  such that, for all  $n \geq 1$ ,  $\langle M(0) \rangle_n \leq an^{1+\alpha d}$  and  $|\Delta M_n(0)| \leq bn^{\alpha d}$ . In addition, since the kernel  $K$  is bounded and Lipschitz, for all  $\delta \in ]0, 1[$ , one can find some positive constant  $C_\delta$  such that, for any  $x, y \in \mathbb{R}^d$

$$(A.7) \quad |K(x) - K(y)| \leq C_\delta \|x - y\|^\delta.$$

Hence, for any  $x, y \in \mathbb{R}^d$ , we can derive that

$$|\Delta M_n(x) - \Delta M_n(y)| \leq 2 C_\delta \|x - y\|^\delta n^{\alpha(d+\delta)}.$$

Furthermore, similarly to (A.5), we have for any  $x, y \in \mathbb{R}^d$

$$\langle M(x) - M(y) \rangle_n \leq \sum_{i=1}^n i^{2\alpha d} \int_{\mathbb{R}^d} \left( K(i^\alpha(\pi_{i-1} + s - x)) - K(i^\alpha(\pi_{i-1} + s - y)) \right)^2 f(s) ds,$$

which, by the change of variables  $t = i^\alpha(\pi_{i-1} + s - x)$ , leads to

$$(A.8) \quad \langle M(x) - M(y) \rangle_n \leq \|f\|_\infty \sum_{i=1}^n i^{\alpha d} \int_{\mathbb{R}^d} (K(t) - K(t + i^\alpha(x - y)))^2 dt.$$

In addition, as  $K$  is a density function, it follows from (A.7) that

$$\int_{\mathbb{R}^d} (K(t) - K(t + i^\alpha(x - y)))^2 dt \leq 2 C_{2\delta} \|x - y\|^{2\delta} i^{2\alpha\delta}.$$

Therefore, we deduce from (A.8) that for any  $x, y \in \mathbb{R}^d$

$$\langle M(x) - M(y) \rangle_n \leq 2 C_{2\delta} \|x - y\|^{2\delta} n^{1+\alpha d+2\alpha\delta}.$$

Since the power  $\delta$  can be chosen as small as one wishes, all four conditions of Theorem 6.4.34 of [12] are fulfilled which leads to Lemma A.2.  $\square$

*Proof of Theorem 2.1.* In order to prove Theorem 2.1, it remains to study the almost sure asymptotic behavior of the remainder  $R_n(x)$  in (A.1). It follows from (A.3) that

$$\begin{aligned} R_n(x) &= \sum_{i=1}^n h_i^{-d} \int_{\mathbb{R}^d} K(h_i^{-1}(\pi_{i-1} + s - x)) f(s) ds - n f(x), \\ &= \sum_{i=1}^n \int_{\mathbb{R}^d} K(t) (f(h_i t + x - \pi_{i-1}) - f(x)) dt, \end{aligned}$$

via the change of variables  $t = h_i^{-1}(\pi_{i-1} + s - x)$ . As the density function  $f$  is differentiable with a bounded gradient, we obtain by a Taylor expansion that

$$\sup_{x \in \mathbb{R}^d} |R_n(x)| = O\left(\sum_{i=1}^n h_i\right) + O\left(\sum_{i=1}^n \|\pi_{i-1}\|\right) \quad \text{a.s.}$$

Moreover, since  $(\varepsilon_n)$  has a finite moment of order  $a > 2$ , we deduce from [A1] and [A2] together with Theorem 1 of [13] that

$$(A.9) \quad \sum_{i=1}^n \|\pi_{i-1}\|^2 = O(n^b) \quad \text{a.s.}$$



for all  $b \in ]2/a, 1[$ . Hence, it follows from (A.9) together with the Cauchy–Schwarz inequality that

$$(A.10) \quad \sup_{x \in \mathbb{R}^d} |R_n(x)| = O(nh_n) + O(\sqrt{n^{1+b}}) \quad \text{a.s.}$$

Consequently,  $R_n(x) = o(n)$  a.s., which ensures that  $\widehat{f}_n(x)$  converges a.s. to  $f(x)$ . Moreover, we obtain (2.1) from the conjunction of Lemma A.1 and result (A.10). The uniform, almost sure convergence on  $\mathbb{R}^d$  still remains to be proven. Hereafter, we take  $h_n = n^{-\alpha}$  with  $\alpha \in ]0, 1/d[$ . On the one hand, we find from Lemma A.2 with  $A = 2$  and  $\gamma = 1/2$ , that

$$(A.11) \quad \sup_{\|x\| \leq 2\sqrt{n}} |M_n(x)| = o(n^\beta) \quad \text{a.s.},$$

where  $\beta \in ](1+\alpha d)/2, 1[$ . From now on, we choose  $\beta \in ](1+c)/2, 1[$  with  $c = \max(b, \alpha d)$ . Since  $\beta > (1+b)/2$ , it implies that  $n^{1+b} = o(n^{2\beta})$ . Hence, it follows from the conjunction of (A.10) and (A.11) that

$$(A.12) \quad \sup_{\|x\| \leq 2\sqrt{n}} \left| \widehat{f}_n(x) - f(x) \right| = O(n^{-\alpha}) + o(n^{\beta-1}) \quad \text{a.s.}$$

On the other hand, we claim that

$$(A.13) \quad \sup_{\|x\| > 2\sqrt{n}} \left| \widehat{f}_n(x) - f(x) \right| = O\left(\frac{1}{n}\right) \quad \text{a.s.}$$

As a matter of fact, since  $(\varepsilon_n)$  has a finite moment of order  $a > 2$ , we infer from Lemma 2 of [13] that  $\|X_n\|^2 = O(n^b)$  a.s. for some  $b \in ]2/a, 1[$ , which implies that

$$\sup_{i \leq n} \|X_i - x_i^*\|^2 = o(n) \quad \text{a.s.}$$

Hence, for  $n$  large enough,  $\|X_i - x_i^*\| < \sqrt{n}$  a.s., which ensures that, for  $x$  such that  $\|x\| > 2\sqrt{n}$ ,  $\|X_i - x_i^* - x\| > \sqrt{n}$  a.s. Therefore, since  $K$  is compactly supported, it clearly leads to

$$(A.14) \quad \sup_{\|x\| > 2\sqrt{n}} \left| n\widehat{f}_n(x) \right| = \sup_{\|x\| > 2\sqrt{n}} \left| \sum_{i=1}^n K_i(X_i - x_i^* - x) \right| = O(1) \quad \text{a.s.}$$

In addition, since  $(\varepsilon_n)$  has a finite moment of order  $a > 2$  and  $f$  is positive, it follows that  $f(x) = O(\|x\|^{-3})$  for large values of  $x$ , leading to

$$(A.15) \quad \sup_{\|x\| > 2\sqrt{n}} f(x) = O\left(\frac{1}{n}\right).$$

Consequently, we obtain (A.13) from (A.14) and (A.15). Finally, we deduce (2.2) from (A.12) and (A.13), which completes the proof of Theorem 2.1.  $\square$

**Appendix B.** This appendix is concerned with the proof of Theorem 2.2. We first propose a CLT for the martingale  $(M_n)$ .

LEMMA B.1. Assume that [A1] to [A3] hold and suppose that  $(\varepsilon_n)$  has a finite moment of order  $a > 2$ . Moreover, assume that the bandwidth  $(h_n)$  shares the same assumptions as in Theorem 2.2. Then, for any  $x \in \mathbb{R}^d$ ,

$$(B.1) \quad \frac{M_n(x)}{\sqrt{nh_n^{-d}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau^2 \ell_h f(x)).$$

*Proof.* In order to prove Lemma B.1, it is necessary to study the asymptotic behavior of the increasing process  $\langle M(x) \rangle_n$  properly normalized. For all  $i \geq 1$  and  $x \in \mathbb{R}^d$ , we have

$$\begin{aligned} \mathbb{E}[K_i(X_i - x_i^* - x) | \mathcal{F}_{i-1}] &= h_i^{-d} \int_{\mathbb{R}^d} K(h_i^{-1}(\pi_{i-1} + s - x)) f(s) \, ds, \\ &= \int_{\mathbb{R}^d} K(t) f(h_i t + x - \pi_{i-1}) \, dt \leq \|f\|_\infty, \end{aligned}$$

which implies that

$$(B.2) \quad \sum_{i=1}^n (\mathbb{E}[K_i(X_i - x_i^* - x) | \mathcal{F}_{i-1}])^2 = O(n) \quad \text{a.s.}$$

Moreover, we also have

$$\begin{aligned} \mathbb{E}[K_i^2(X_i - x_i^* - x) | \mathcal{F}_{i-1}] &= h_i^{-2d} \int_{\mathbb{R}^d} K^2(h_i^{-1}(\pi_{i-1} + s - x)) f(s) \, ds \\ &= h_i^{-d} \int_{\mathbb{R}^d} K^2(t) f(h_i t + x - \pi_{i-1}) \, dt. \end{aligned}$$

Consequently, we obtain the decomposition

$$\sum_{i=1}^n \mathbb{E}[K_i^2(X_i - x_i^* - x) | \mathcal{F}_{i-1}] = A_n + \tau^2 B_n + \tau^2 f(x) C_n,$$

where

$$\begin{aligned} A_n &= \sum_{i=1}^n h_i^{-d} \int_{\mathbb{R}^d} K^2(t) (f(h_i t + x - \pi_{i-1}) - f(x - \pi_{i-1})) \, dt, \\ B_n &= \sum_{i=1}^n h_i^{-d} (f(x - \pi_{i-1}) - f(x)), \\ C_n &= \sum_{i=1}^n h_i^{-d}. \end{aligned}$$

As the gradient of  $f$  is bounded, we clearly have  $|A_n| = O(nh_n^{1-d})$  a.s. and

$$|B_n| = O\left(\sum_{i=1}^n h_i^{-d} \|\pi_{i-1}\|\right) \quad \text{a.s.}$$

Hence, it follows from (A.9) that

$$|B_n| = O\left(h_n^{-d} \sqrt{n^{1+b}}\right) \quad \text{a.s.}$$

for all  $b \in ]2/a, 1[$ . Furthermore, we immediately get from (2.6) that  $n^{-1}h_n^d C_n$  converges to  $\ell_h$  as  $n$  goes to infinity. Putting together those three contributions, we find that

$$(B.3) \quad \lim_{n \rightarrow \infty} \frac{h_n^d}{n} \langle M(x) \rangle_n = \tau^2 \ell_h f(x) \quad \text{a.s.}$$

In order to make use of the CLT for martingales (see, e.g., [12], Corollary 2.1.10, p. 46), it remains to check that Lindeberg's condition is satisfied. For all  $a > 0$  and  $x \in \mathbb{R}^d$ , let

$$\Lambda_n(a, x) = \frac{h_n^d}{n} \sum_{i=1}^n \mathbb{E} \left[ |\Delta M_i(x)|^2 \mathbb{I}_{(|\Delta M_i(x)| \geq a \sqrt{nh_n^{-d}})} | \mathcal{F}_{i-1} \right].$$

We already saw that for all  $i \leq n$ ,  $|\Delta M_i(x)| \leq 2 h_n^{-d} \|K\|_\infty$ . Hence, we clearly have for all  $i \leq n$

$$\mathbb{I}_{(|\Delta M_i(x)| \geq a \sqrt{nh_n^{-d}})} \leq \mathbb{I}_{(2\|K\|_\infty \geq a \sqrt{nh_n^{-d}})}.$$

Consequently, we find that for all  $a > 0$  and  $x \in \mathbb{R}^d$ ,

$$\begin{aligned} \Lambda_n(a, x) &\leq \frac{h_n^d}{n} \mathbb{I}_{(2\|K\|_\infty \geq a \sqrt{nh_n^{-d}})} \sum_{i=1}^n \mathbb{E} \left[ |\Delta M_i(x)|^2 | \mathcal{F}_{i-1} \right], \\ &\leq \frac{h_n^d}{n} \mathbb{I}_{(2\|K\|_\infty \geq a \sqrt{nh_n^{-d}})} \tau^2 \|f\|_\infty \sum_{i=1}^n h_i^{-d}, \\ &\leq \tau^2 \|f\|_\infty \mathbb{I}_{(2\|K\|_\infty \geq a \sqrt{nh_n^{-d}})}. \end{aligned}$$

Therefore, as  $nh_n^d$  tends to infinity, we can deduce that, for all  $a > 0$  and  $x \in \mathbb{R}^d$ ,  $\Lambda_n(a, x)$  tends to zero a.s. Finally, Lindeberg's condition is satisfied, which achieves the proof of Lemma B.1.  $\square$

*Proof of Theorem 2.2.* We are now in position to prove Theorem 2.2. It follows from (A.1) that for any  $x \in \mathbb{R}^d$

$$(B.4) \quad \sqrt{nh_n^d} (\hat{f}_n(x) - f(x)) = \frac{M_n(x) + R_n(x)}{\sqrt{nh_n^{-d}}}.$$

Consequently, (2.7) immediately follows from (A.10) together with (B.1) and (B.4) as soon as  $\max(nh_n^{d+2}, n^b h_n^d) = o(1)$ . The multivariate CLT remains to be proven. Taking the previous results into account, it is enough to prove that for two distinct points  $x, y \in \mathbb{R}^d$ , the random vector

$$\frac{1}{\sqrt{nh_n^{-d}}} \begin{pmatrix} M_n(x) \\ M_n(y) \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} G(x) \\ G(y) \end{pmatrix},$$

where  $G(x)$  and  $G(y)$  are two independent Gaussian random variables. We can easily show this convergence by remarking that for two distinct points  $x, y \in \mathbb{R}^d$

$$(B.5) \quad \lim_{n \rightarrow \infty} \frac{h_n^d}{n} \sum_{i=1}^n \mathbb{E} [\Delta M_i(x) \Delta M_i(y) | \mathcal{F}_{i-1}] = 0 \quad \text{a.s.}$$

Indeed, for all  $i \geq 1$ , we have

$$\begin{aligned}\mathbb{E}[\Delta M_i(x)\Delta M_i(y)|\mathcal{F}_{i-1}] &\leq \mathbb{E}[K_i(X_i - x_i - x)K_i(X_i - x_i - y)|\mathcal{F}_{i-1}], \\ &\leq \mathbb{E}[K_i(\pi_{i-1} + \varepsilon_i - x)K_i(\pi_{i-1} + \varepsilon_i - y)|\mathcal{F}_{i-1}],\end{aligned}$$

which implies that

$$\mathbb{E}[\Delta M_i(x)\Delta M_i(y)|\mathcal{F}_{i-1}] \leq h_i^{-d} \int_{\mathbb{R}^d} K(t)K(t + h_i^{-1}(x - y))f(h_it + x - \pi_{i-1})dt.$$

Therefore, as the gradient of  $f$  is bounded, we obtain from (A.9) that

$$\sum_{i=1}^n \mathbb{E}[\Delta M_i(x)\Delta M_i(y)|\mathcal{F}_{i-1}] \leq H_n(x, y) + O(nh_n^{1-d}) + O(h_n^{-d}\sqrt{n^{1+b}}) \quad \text{a.s.}$$

for all  $b \in ]2/a, 1[$ , where

$$H_n(x, y) = \sum_{i=1}^n h_i^{-d} f(x) \int_{\mathbb{R}^d} K(t)K(t + h_i^{-1}(x - y))dt.$$

However, using the fact that  $K$  is compactly supported, we can deduce that for  $i$  large enough, the integral at the right-hand side of  $H_n(x, y)$  is zero. Finally, we obtain that convergence (B.5) is satisfied, which completes the proof of Theorem 2.2.  $\square$

*Remark 7.* Result (B.5) ensures the asymptotic independence of the random variables  $G_n(x_1), \dots, G_n(x_N)$  in the multivariate CLT. Since the kernel  $K$  is compactly supported, for finite values of  $n$ , the left-hand side of (B.5) can be very small if we choose two points  $x$  and  $y$  sufficiently distant. This last point clarifies the design points selection rule described in section 4.

## REFERENCES

- [1] K. J. ASTRÖM AND B. WITTENMARK, *Adaptive Control*, 2nd ed., Addison-Wesley, New York, 1995.
- [2] D. BACHMANN AND H. DETTE, *A note on the Bickel-Rosenblatt test in autoregressive time series*, Statist. Probab. Lett., 74 (2005), pp. 221–234.
- [3] M. BASSEVILLE AND I. V. NIKIFOROV, *Detection of Abrupt Changes—Theory and Application*, Prentice Hall, Englewood Cliffs, NJ, 1993.
- [4] B. BERCU, *Weighted estimation and tracking for ARMAX models*, SIAM J. Control Optim., 33 (1995), pp. 89–106.
- [5] B. BERCU, *Central limit theorem and law of iterated logarithm for least squares algorithms in adaptive tracking*, SIAM J. Control Optim., 36 (1998), pp. 910–928.
- [6] B. BERCU AND B. PORTIER, *Adaptive control of parametric nonlinear autoregressive models via a new martingale approach*, IEEE Trans. Automat. Control, 47 (2002), pp. 1524–1528.
- [7] P. BICKEL AND P. ROSENBLATT, *On some global measures of the deviation of density function estimators*, Ann. Statist., 1 (1973), pp. 1071–1095.
- [8] P. J. BROCKWELL AND R. A. DAVIES, *Time Series: Theory and Methods*, 2nd ed., Series in Statistics, Springer, New York, 1991.
- [9] H. F. CHEN AND L. GUO, *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston, 1991.
- [10] L. DEVROYE, *A Course in Density Estimation*, Birkhäuser, Boston, 1987.
- [11] L. DEVROYE AND G. LUGOSI, *Combinatorial Methods in Density Estimation*, Springer-Verlag, New York, 2001.
- [12] M. DUFLO, *Random Iterative Models*, Springer-Verlag, Berlin, 1997.

- [13] L. GUO AND H. F. CHEN, *The Aström Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers*, IEEE Trans. Automat. Control, 36 (1991), pp. 802–812.
- [14] L. GUO, *Further results on least squares based adaptive minimum variance control*, SIAM J. Control Optim., 32 (1994), pp. 187–212.
- [15] L. GUO, *Self-convergence of weighted least squares with applications to stochastic adaptive control*, IEEE Trans. Automat. Control, 41 (1996), pp. 79–89.
- [16] T. L. LAI AND C. Z. WEI, *Extended least squares and their applications to adaptive control and prediction in linear systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 898–906.
- [17] S. LEE AND S. NA, *On the Bickel-Rosenblatt test for first-order autoregressive models*, Statist. Probab. Lett., 56 (2002), pp. 23–35.
- [18] E. PARZEN, *On estimation of a probability density function and mode*, Ann. Math. Stat., 33 (1962), pp. 1065–1076.
- [19] J. M. POGGI AND B. PORTIER, *A test of linearity for functional autoregressive models*, J. Time Ser. Anal., 18 (1997), pp. 615–640.
- [20] J.-M. POGGI AND B. PORTIER, *Nonlinear adaptive tracking using kernel estimators: estimation and test for linearity*, SIAM J. Control Optim., 39 (2000), pp. 707–727.
- [21] B. PORTIER, *Adaptive control of discrete-time nonlinear systems combining nonparametric and parametric estimators*, Commun. Inf. Syst., 2 (2002), pp. 69–90.
- [22] B. PORTIER AND A. OULIDI, *Nonparametric estimation and adaptive control of functional autoregressive models*, SIAM J. Control Optim., 39 (2000), pp. 411–432.
- [23] M. ROSENBLATT, *Remarks on some nonparametric estimates of a density function*, Ann. Math. Stat., 27 (1956), pp. 832–837.
- [24] B. W. SILVERMAN, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, New York, 1986.
- [25] L.-L. XIE AND L. GUO, *How much uncertainty can be dealt with by feedback?*, IEEE Trans. Automat. Control, 45 (2000), pp. 2203–2217.

## IMPORTANT MOMENTS IN SYSTEMS AND CONTROL\*

CHRISTOPHER I. BYRNES<sup>†</sup> AND ANDERS LINDQUIST<sup>‡</sup>

**Abstract.** The moment problem matured from its various special forms in the late 19th and early 20th centuries to a general class of problems that continues to exert profound influence on the development of analysis and its applications to a wide variety of fields. In particular, the theory of systems and control is no exception, where the applications have historically been to circuit theory, optimal control, robust control, signal processing, spectral estimation, stochastic realization theory, and the use of the moments of a probability density. Many of these applications are also still works in progress. In this paper, we consider the generalized moment problem, expressed in terms of a basis of a finite-dimensional subspace  $\mathfrak{P}$  of the Banach space  $C[a, b]$  and a “positive” sequence  $c$ , but with a new wrinkle inspired by the applications to systems and control. We seek to parameterize solutions which are positive “rational” measures in a suitably generalized sense. Our parameterization is given in terms of smooth objects. In particular, the desired solution space arises naturally as a manifold which can be shown to be diffeomorphic to a Euclidean space and which is the domain of some canonically defined functions. The analysis of these functions, and related maps, yields interesting corollaries for the moment problem and its applications, which we compare to those in the recent literature and which play a crucial role in part of our proof.

**Key words.** moment problems, interpolation, rational positive measures

**AMS subject classifications.** 30E05, 44A60, 93B28

**DOI.** 10.1137/070693941

**1. Introduction.** Given a sequence of complex numbers,  $(c_0, c_1, \dots, c_n)$ , and a basis,  $(\alpha_0, \alpha_1, \dots, \alpha_n)$ , of a (finite-dimensional) subspace  $\mathfrak{P}$  of the Banach space  $C[a, b]$  of complex-valued continuous functions defined on the real interval  $[a, b]$ , the generalized moment problem [21] is to find a positive measure  $d\mu$  such that

$$(1.1) \quad \int_a^b \alpha_k(t) d\mu(t) = c_k, \quad k = 0, 1, \dots, n.$$

This problem is a beautiful generalization of several important classical moment problems, including the power moment problem, the trigonometric moment problem, and the moment problem arising in Nevanlinna–Pick interpolation. There are, of course, necessary conditions stemming from the positivity of  $d\mu$  and whether a particular  $\alpha_k$  is real-valued or not; these will be summarized in section 2.

Among the pioneers in the use of power moments, where  $\alpha_k(t) = t^k$ , we should mention Chebyshev and his students, particularly Markov and Lyapunov, who used them in connection with the classical central limit theorem in the 19th century. On a finite interval this problem is usually called the Hausdorff moment problem and was solved by Hausdorff for an infinite sequence of moments in 1921. The power moment problem for an infinite sequence of moments on an infinite interval is known

---

\*Received by the editors June 7, 2007; accepted for publication (in revised form) May 18, 2008; published electronically September 8, 2008. This research was supported in part by grants from AFOSR, Swedish Research Council, Swedish Foundation of Strategic Research, and the Göran Gustafsson Foundation.

<http://www.siam.org/journals/sicon/47-5/69394.html>

<sup>†</sup>Department of Electrical and Systems Engineering, Washington University, St. Louis, MO 63130 (chrisbyrnes@seas.wustl.edu).

<sup>‡</sup>Department of Mathematics, Royal Institute of Technology, 100 44 Stockholm, Sweden (alq@math.kth.se).

as the Hamburger moment problem, while on the semi-infinite interval this is called the Stieltjes moment problem. We refer to [21], especially pages 166–171 and the references therein, for a more detailed historical and technical treatment.

*Remark 1.1.* In classical treatments of the power moment problems [21] it is typical to take  $\mathfrak{P}$  to be the real subspace  $\text{span}_{\mathbb{R}}\{\alpha_0, \dots, \alpha_n\}$ . While in this case the role of the functions  $\alpha_k$  are clear and familiar to any student of probability, it is reasonable to ask why we need  $\mathfrak{P}$ . One of many good reasons for this is that  $\mathfrak{P}$  is a natural space of “test functions” with which to develop necessary and sufficient conditions on the candidate moments for the solvability of the moment equations. For example, if  $p(t) = p_0 + p_1 t + \dots + p_n t^n > 0$  for all  $t \in [a, b]$ , then solvability of the moment equations for a positive measure  $d\mu$  implies that

$$(1.2) \quad \sum_{i=0}^n p_i c_i = \int_a^b p(t) d\mu > 0.$$

This has been refined, in a neat way, to give necessary and sufficient conditions for the solvability of the generalized moment problem (see [21] and the discussion in section 2).

In the trigonometric moment problem, where  $\alpha_k(t) = e^{ikt}$  defined on  $[-\pi, \pi]$ , the constants  $c_k$  are, of course, the first  $n + 1$  Fourier coefficients of  $d\mu$ . The corresponding moment problem was classically considered by Carathéodory in potential theory, where the moment conditions place a constraint on the boundary value data for Laplace’s equation on the unit disc. Through subsequent classical work by Schur, Toeplitz, Nevanlinna, Pick, and many others, this has been influential in the development of modern analysis (see, e.g., [16]). Applications of the trigonometric moment problem to systems and control also have a long and fruitful history, including the rational covariance extension problem originally posed by Kalman [18] and later observed to be related to the trigonometric moment problem in [11]. However, to be applicable to problems in spectral estimation and stochastic realization theory there are system theoretic constraints that must be added to the trigonometric moment problem, relating to the rationality of, and the degree of, a solution. These challenges were noticed early on [18, 20, 13], and the ultimate breakthroughs relied (and still do rely) on the nontrivial use of topology, nonlinear convex optimization, or a combination thereof (see [13, 9] and the SIGEST paper [4] and references therein).

*Remark 1.2.* In this setting, the classical theory was developed for a complex subspace  $\mathfrak{P}$  of “test functions” and follows, mutatis mutandis, the real case. Explicitly, in order to develop the corresponding necessary conditions it is necessary to take those  $p \in \mathfrak{P}$  for which the trigonometric polynomial  $P := \text{Re}(p)$  is positive on  $[-\pi, \pi]$ . The complex-valued enhancement of condition (1.2) is then

$$(1.3) \quad \text{Re} \left( \sum_{i=0}^n p_i c_i \right) = \frac{1}{2} \sum_{i=0}^n (\bar{p}_i c_i + p_i \bar{c}_i) = \int_{-\pi}^{\pi} P d\mu > 0.$$

One of the many gems in this classical literature is the use [21, p. 65] of the Riesz–Fejér theorem to evaluate the quadratic form on the right-hand side of (1.3), where  $P > 0$ , as

$$\sum_{i,j=0}^n c_{i-j} z_i \bar{z}_j = \bar{z}^T T_n z > 0,$$

where  $z = (z_0, \dots, z_n) \in \mathbb{C}^n \setminus 0$  and  $T_n$  is the standard Toeplitz form fashioned out of the moment sequence  $c = (c_i)$ . For a general moment problem, the form on the left-hand side of (1.3) is classically denoted by  $\langle c, p \rangle$ .

*Remark 1.3.* In both the power and the trigonometric moment problems we were led to consider the “polynomials”  $P = \operatorname{Re}(p)$  for  $p \in \mathfrak{P}$ . For this reason, the functions  $P := \operatorname{Re}(p)$  for  $p \in \mathfrak{P}$  in an arbitrary generalized moment problem are referred to as “polynomials” for  $\mathfrak{P}$ . Following this precedent, we shall refer to the ratio  $P/Q$  with  $p, q \in \mathfrak{P}$  as “rational functions” for  $\mathfrak{P}$ .

In the Nevanlinna–Pick interpolation problem for distinct interpolation points  $z_0, z_1, \dots, z_n$ , the basis functions are given by

$$\alpha_k(t) = \frac{1}{2\pi} \frac{e^{it} + z_k}{e^{it} - z_k}, \quad k = 0, 1, \dots, n,$$

which coincide on  $[-\pi, \pi]$ , modulo an additive constant, with Cauchy kernels. Higher order kernels can of course be used for multiple points. As for the case of trigonometric polynomials, it turns out that it is more helpful to identify the interval with the unit circle and, in this case, to think of  $\mathfrak{P}$  in terms of Hardy spaces. This has also led to profound developments in several complex variables and in operator theory as well as in the applications of mathematics to circuit theory [10, 17] and to robust control [25, 19, 14, 12, 22]. For this problem as well, the applications to systems and control impose additional constraints to the classical moment problem whose treatment still requires nonlinear methods drawn from geometry, topology, and/or optimization [17, 14, 2, 4].

*Remark 1.4.* For the classical Nevanlinna–Pick interpolation problem, using the Riesz–Fejér theorem, the quadratic form (1.3) can also be evaluated, with some work [21, pp. 67–69], as the value of the celebrated Pick form. Moreover, it turns out that  $\mathfrak{P}$  is a finite-dimensional coinvariant subspace of  $H^2$  so that the elements of  $\mathfrak{P}$  are rational functions  $\sigma/\tau$ , where  $\tau$  is fixed, and the “polynomials” are the real parts of elements in  $\mathfrak{P}$ . This of course implies that the “rational functions” are rational in the usual sense.

The generalized moment problem is about measures, and combining these two concepts leads us to the following definition.

DEFINITION 1.5. *Any measure of the form*

$$(1.4) \quad d\mu = \frac{P(t)}{Q(t)} dt,$$

where  $P$  and  $Q$  are positive polynomials for  $\mathfrak{P}$ , is a rational positive measure.

PROBLEM 1.6. Given a sequence of complex numbers  $c_0, c_1, \dots, c_n$  and a subspace  $\mathfrak{P}$ , the generalized moment problem for rational measures is to parameterize all positive rational measures  $\frac{P(t)}{Q(t)} dt$  such that

$$(1.5) \quad \int_a^b \alpha_k(t) \frac{P(t)}{Q(t)} dt = c_k, \quad k = 0, 1, \dots, n.$$

The problem itself is motivated by classical applications and examples, in both finite and infinite dimensions, and also reflects the importance of rational functions in systems and control. In this paper we give a concise description of all solutions of this generalized moment problem for a broad class of subspaces  $\mathfrak{P}$ .



**2. The main result.** In order to state our result, we first need to compute the dimension of  $\mathfrak{P}$  as a real vector space, taking into account the cases where a basis element is real, purely imaginary, or neither. In order for the moment equations to hold it is necessary that  $c_k$  be real whenever  $\alpha_k$  is real. Moreover, a purely imaginary moment condition can always be reduced to a real one, and henceforth we shall assume that  $\alpha_0, \dots, \alpha_{r-1}$  are real functions and  $\alpha_r, \dots, \alpha_n$  are complex-valued functions whose real and imaginary parts, taken together with  $\alpha_0, \dots, \alpha_{r-1}$ , are linearly independent over  $\mathbb{R}$ . In particular, we may regard  $\mathfrak{P}$  as a real vector space of dimension  $2n - r + 2$ . Since we have chosen a fixed basis, we may regard each

$$(2.1) \quad p := \sum_{k=0}^n p_k \alpha_k \in \mathfrak{P}$$

also as an  $(n+1)$ -tuple of points  $(p_0, p_1, \dots, p_n)$ , where  $p_0, p_1, \dots, p_{r-1}$  are real and  $p_r, p_{r+1}, \dots, p_n$  are complex. Moreover,  $p$  is determined by its real part  $P := \operatorname{Re}(p)$ , a notation we shall keep throughout. Next we define the subset  $\mathfrak{P}_+$  of those elements  $p \in \mathfrak{P}$  such that  $P > 0$ . We shall assume that  $\mathfrak{P}_+$  is nonempty and is therefore an open, convex set having dimension  $2n - r + 2$ .

The rational measures we seek as solutions have the property that  $d\mu(E) > 0$  for every Borel measurable subset  $E \subset [a, b]$  having nonzero measure. For this reason, we will seek a necessary condition expressible in terms of the slightly larger space  $\overline{\mathfrak{P}}_+ \setminus \{0\}$  of “test” functions. That is, we define  $\mathfrak{C}_+$  as the set of sequences  $c = (c_0, c_1, \dots, c_n)$  such that

$$(2.2) \quad \langle c, p \rangle := \operatorname{Re} \left\{ \sum_{k=0}^n p_k c_k \right\} = \int_a^b P \, d\mu > 0$$

for all  $p \in \overline{\mathfrak{P}}_+ \setminus \{0\}$ . We will call such a sequence *positive*. In particular,  $\mathfrak{C}_+$  is also a nonempty, open convex subset of  $\mathbb{R}^{2n-r+2}$  of dimension  $2n - r + 2$ .

*Remark 2.1.* Since we are seeking a solution to the moment problem for a smaller class of positive measure than in the classical treatment, our necessary conditions are, not surprisingly, stronger than the classical conditions. In particular, a sequence  $c$  which we call positive is referred to as *strictly positive* in [21]. A sequence  $c$  is positive *in the classical sense* if it satisfies  $\langle c, p \rangle \geq 0$  for all  $p \in \overline{\mathfrak{P}}_+$ , a condition that does not quite capture the case of positive rational measures.

We shall now fix  $c \in \mathfrak{C}_+$  and consider the set  $\mathfrak{M}_c$  of all pairs of polynomials  $(p, q)$  for which the rational measure  $P(t)/Q(t)dt$  solves Problem 1.6 for the positive sequence  $c$ . There is a natural parameterization of  $\mathfrak{M}_c$  as a submanifold of the product space  $\mathfrak{P}_+ \times \mathfrak{P}_+$ , and, as a subset of a product space,  $\mathfrak{M}_c$  comes with two mappings:

$$\pi_1 : \mathfrak{M}_c \rightarrow \mathfrak{P}_+ \quad \text{and} \quad \pi_2 : \mathfrak{M}_c \rightarrow \mathfrak{P}_+,$$

where  $\pi_1$  and  $\pi_2$  are the restrictions to  $\mathfrak{M}_c$  of the two mappings

$$\operatorname{proj}_1 : \mathfrak{P}_+ \times \mathfrak{P}_+ \rightarrow \mathfrak{P}_+ \quad \text{and} \quad \operatorname{proj}_2 : \mathfrak{P}_+ \times \mathfrak{P}_+ \rightarrow \mathfrak{P}_+,$$

defined by  $\operatorname{proj}_1(p, q) = p$  and  $\operatorname{proj}_2(p, q) = q$ .

**THEOREM 2.2.** *Suppose that  $\mathfrak{P}$  consists of Lipschitz continuous functions. Then, for each  $c \in \mathfrak{C}_+$ ,  $\mathfrak{M}_c$  is a smooth submanifold of  $\mathfrak{P}_+ \times \mathfrak{P}_+$ , diffeomorphic to  $\mathbb{R}^{2n-r+2}$ . Moreover, each of the maps  $\pi_1, \pi_2$  is a diffeomorphism of  $\mathfrak{M}_c$  onto its image, which is an open submanifold of  $\mathfrak{P}_+$ . Finally,  $\pi_1 : \mathfrak{M}_c \rightarrow \mathfrak{P}_+$  is surjective.*

*Remark 2.3.* To the best of our knowledge, all instances of the generalized moment problem that arise in systems and control involve subspaces of  $C[a, b]$  consisting of Lipschitz continuous functions. Moreover, this class of subspaces has been of considerable classical interest. For example, an important class of spaces considered in the classical literature on the generalized moment problem [21] consists of those spaces  $\mathfrak{P}$  spanned by a Chebyshev system (or T-system), which are characterized by a bound on the number of zeros for any nonzero polynomial in  $\mathfrak{P}$ . These spaces arise in important applications of the generalized moment problem, e.g., the power moment problem and the trigonometric moment problem of odd order, and have remarkable approximation properties in the Banach space  $C[a, b]$ . We remark that [21] contains a neat application, generalizing Feldbaum's theorem on the number of switchings, of Chebyshev systems to the time-optimal control of scalar-input linear control systems. For our present purposes, we recall the classical result that, if  $\mathfrak{P}$  is spanned by a Chebyshev system and contains a constant function, then, after a reparameterization,  $\mathfrak{P}$  consists of Lipschitz continuous functions [21, p. 37].

*Remark 2.4.* We have remarked that the finite-dimensional Nevanlinna–Pick problem can be recast in a Hardy space setting, where the space  $\mathfrak{P}$  is a coinvariant subspace (defined, in fact, by a finite Blaschke product) in  $H^2(\mathbb{D})$ . In a seminal paper [23], Sarason developed a vast generalization of this problem to one involving liftings of a partial isometry  $T$ , defined on an arbitrary coinvariant subspace, which commute there with the restriction of the shift operator. Among many other results, Sarason showed that, under general conditions, the lift of  $T$  has an  $H^\infty$  symbol which is *rational* with respect to the coinvariant subspace. The corresponding problem for  $T$  being a strict contraction was studied in [3], where optimization methods were used to show that the lifting of such a  $T$  always had such a generalized rational symbol. Moreover, it was shown that this symbol is completely parameterized by its numerator in parallel with the conclusion in Theorem 2.2 that  $\pi_1$  is a bijection. In this light, it is interesting to enquire whether a general version of Problem 1.6 can be formulated, and solved, in a meaningful infinite-dimensional setting.

The formulations of Definition 1.5 and Problem 1.6 for generalized rational measures and of Theorem 2.2 are new and have some appeal both for the intrinsic simplicity of the formulation and as a unification of a variety of specific applications and more general results on the moment problem. There are of course antecedents in the literature to some parts of the theorem and its corollaries. We shall review these results as a conclusion to our outline of the proof in section 3.

**3. An outline of the proof.** The proof of our main result can be reduced to several steps. The first part involves establishing some smoothness results for  $\mathfrak{M}_c$  and the maps  $\pi_1$  and  $\pi_2$ . This, of course, depends upon the ambient spaces and their properties, as investigated in section 4. In Proposition 4.1, we establish the required smoothness and prove that each of the maps  $\pi_1$  and  $\pi_2$  is a local diffeomorphism, whenever  $\mathfrak{M}_c$  is nonempty.

The final steps in the proof are to demonstrate that  $\mathfrak{M}_c$  is nonempty for each positive sequence  $c$ , that  $\pi_1$  is a bijection, and that  $\pi_2$  is an injection. For suppose that  $\mathfrak{M}_c$  is nonempty. By the inverse function theorem, the image of each  $\pi_i$  is an open subset  $U_i$  of  $\mathfrak{P}_+$ . Therefore, to say that  $\pi_1$  is also a bijection is to say that it has an inverse defined on  $U_1 = \mathfrak{P}_+$ , which from the inverse function theorem must also be differentiable. That is, the map

$$\pi_1 : \mathfrak{M}_c \rightarrow \mathfrak{P}_+$$

is a diffeomorphism. Similarly, to say that  $\pi_2$  is an injection is to say that

$$\pi_2 : \mathfrak{M}_c \rightarrow U_2$$

is a diffeomorphism. Taken together, these steps conclude the proof. The proofs of the last three steps are, however, not just set-theoretic.

For example, the analysis of the map  $\pi_2$  boils down to the analysis of a linear map between closed convex sets. If  $\mathfrak{P}$  contains the constant functions, we may, for example, choose  $q = 1$  which leads to a new constrained problem, the generalized moment problem for positive *polynomial measures*. In section 5, after proving in Lemma 5.1 that  $\pi_2$  is injective, we analyze its image using the auxiliary problem for polynomial measures. In particular, we deduce Proposition 5.5 which asserts that  $\pi_2$  fails to be surjective for general  $c$  in dimension greater than one, under the auxiliary hypothesis that the zero set of any  $p \in \mathfrak{P}$  has measure zero.

In contrast, the analysis of  $\pi_1$  is nonlinear and the result is nicer. To say that, for every  $c$  and for each  $p$ , there exists a unique  $q$  is to say that, for each fixed  $p$ , and any  $c$ , there exists a unique  $q$  so that the corresponding rational measure solves the moment problem for  $c$ . As for the map  $\pi_2$ , this results in a related constrained moment problem, which defines a smooth mapping  $F^p : \mathfrak{P}_+ \rightarrow \mathfrak{C}_+$ . We show that  $\pi_1$  is a diffeomorphism if and only if  $F^p$  is a diffeomorphism. The local smoothness results obtained in section 4 imply that  $F^p$  is a local diffeomorphism, so that  $F^p(\mathfrak{P}_+) \subset \mathfrak{C}_+$  is open. In section 6, we prove Lemma 6.3, which asserts that  $F^p$  is proper whenever  $\mathfrak{P}$  consists of Lipschitz continuous functions. In particular,  $F^p(\mathfrak{P}_+) = \mathfrak{C}_+$ , and an application of Hadamard's global inverse function theorem shows that  $F^p$  is a diffeomorphism. This has several important and interesting consequences, including Corollary 6.4, which asserts that  $\mathfrak{M}_c$  is nonempty, for each positive sequence  $c$ . In addition, we observe in Corollary 6.7 that  $\pi_1$  is surjective and hence a bijection, thereby concluding the proof of Theorem 2.2.

*Remark 3.1.* The proof of Theorem 2.2 both touches upon and gives new proofs of certain results in the literature on generalized moment problems with a degree, or a complexity, constraint. Some of these were developed in some specific applications to problems arising in systems and control and, later, in a more general setting. For example, in the SIGEST paper [4], we surveyed the trigonometric moment problem and its manifestation in our work, and the work of Georgiou, on the covariance extension problem. In [4], we also reviewed our joint work with Georgiou [2] on the Nevanlinna–Pick moment problem. In both of these problems, a specialized version of Theorem 6.5 emerged. It is fair to say that, at the time, everybody interested in this circle of problems recognized that this kind of result capped off the brilliant introduction of topological methods into these problems by Georgiou [13, 14]. Motivated by the similarities between these problems and by their common role as classical instances of the generalized moment problem, we concluded [4] with a sketch of a unified approach to both applications in the form of a constrained generalized moment problem, as treated in section 6. The resulting formulation stated a version of Theorem 6.5 for arbitrary subspaces  $\mathfrak{P}$  and referred, as did the more recent paper [7], to the unpublished report [6] for more details and proofs. However, it is also fair to say that, at the time, both the formulation of the general problem in terms of (generalized) rational measures and Theorem 2.2 remained unanticipated.

The basic technical lemma in [6] has been generalized here as Lemma 6.3 and is proved in the case when  $\mathfrak{P}$  consists of Lipschitz continuous functions. This is unlikely to be the most general form of the technical lemma but, in light of Remark 2.3, could

be the most interesting form for finite-dimensional subspaces  $\mathfrak{P}$ . A brief overview of this result and the hypotheses under which versions of Theorem 6.5 have been established can be described as follows.

- The proof of the corresponding results in [6] required that the subspace  $\mathfrak{P}$  consists of functions of class  $C^2$ .
- Georgiou [15] developed an innovative approach to the generalized moment problem with complexity constraints based on a one-parameter embedding argument, similar to the path-lifting proof of the Banach–Mazur theorem in [1]. Using this method, Georgiou was able to prove an analogue of Theorem 6.5 for subspaces  $\mathfrak{P}$  consisting of functions of class  $C^1$ .
- In [8], an alternative approach to this constrained moment problem was developed from a detailed analysis of an underlying variational problem, proving in particular that all minimizers arise as interior points. The proof holds under a condition concerning certain divergent integrals that is valid whenever  $\mathfrak{P}$  consists of Lipschitz continuous functions.

In contrast, the approach followed here avoids the use of variational methods and relies instead on nonlinear analysis, such as Hadamard’s global inverse function theorem, to give a streamlined yet self-contained proof of existence and uniqueness results for a class of constrained moment problems en route to our ultimate goal, Theorem 2.2.

**4. Some basic results on smoothness.** We now turn to the smoothness of  $\mathfrak{M}_c$  and the maps  $\pi_1$  and  $\pi_2$ . The map

$$M : \mathfrak{P}_+ \times \mathfrak{P}_+ \rightarrow \mathfrak{C}_+,$$

defined via

$$M(p, q) = \int_a^b \begin{pmatrix} \alpha_1(t) \\ \alpha_2(t) \\ \vdots \\ \alpha_n(t) \end{pmatrix} \frac{P(t)}{Q(t)} dt,$$

has  $\mathfrak{M}_c$  as its level set  $M^{-1}(c)$ .

For simplicity, we view  $\mathfrak{P}$  and  $\mathfrak{C}$  as real vector spaces, so that  $\mathfrak{P}$  is spanned by the real basis  $(\alpha_i)$ , where we have replaced a complex-valued  $(\alpha_k)$  by its real and imaginary parts. The Jacobian,  $\text{Jac}(M)_{(p_0, q_0)}$ , of  $M$  at a point  $(p_0, q_0)$  takes the form

$$(4.1) \quad \text{Jac}(M) = (\partial M / \partial p, \partial M / \partial q) = (M_p, M_q),$$

where  $M_p$  is the square matrix whose  $(i, j)$ th entry is

$$(4.2) \quad (M_p)_{(i, j)} = \int_a^b \alpha_i(t) \alpha_j(t) \frac{1}{Q(t)} dt$$

and where  $M_q$  is defined by

$$(4.3) \quad (M_q)_{(i, j)} = - \int_a^b \alpha_i(t) \alpha_j(t) \frac{P(t)}{Q^2(t)} dt,$$

each being evaluated at the point  $(p_0, q_0)$ . Thus,  $M_p$  (or  $-M_q$ ) is the Gramian matrix of the real basis  $(\alpha_i)$  with respect to the positive definite inner product defined by

$Q(t)^{-1}dt$  (or  $P(t)/Q^2(t)dt$ ) on  $C[a, b]$ . Therefore,  $\text{Jac}(M)$  has rank  $2n - r + 2$  at each point  $(p, q)$  so that, by the implicit function theorem, we obtain the following result.

**PROPOSITION 4.1.** *For each  $c \in \mathfrak{C}_+$ ,  $\mathfrak{M}_c$  is either empty or a submanifold of  $\mathfrak{P}_+ \times \mathfrak{P}_+$  of real dimension  $2n - r + 2$ .*

As restrictions of a smooth map to a smooth submanifold of the product, both  $\pi_1$  and  $\pi_2$  are smooth maps from  $\mathfrak{M}_c$  to  $\mathfrak{P}_+$ . Suppose that  $M(p_0, q_0) = c$  so that, in particular,  $\mathfrak{M}_c$  is nonempty. The tangent space  $T_{(p_0, q_0)}(\mathfrak{M}_c)$  to  $\mathfrak{M}_c$  at  $(p_0, q_0)$  is given by the kernel of  $\text{Jac}(M)_{(p_0, q_0)}$ . By inspection, we have

$$(4.4) \quad \ker \text{Jac}(M)_{(p_0, q_0)} = \left\{ \begin{bmatrix} M_p^{-1}x \\ -M_q^{-1}x \end{bmatrix} : x \in \mathbb{R}^{2n-r+2} \right\}.$$

We wish to show that

$$\text{rank } \text{Jac}(\pi_1)_{(p_0, q_0)} = 2n - r + 2.$$

This will occur if and only if

$$\dim \ker \text{Jac}(\pi_1)_{(p_0, q_0)} = 0,$$

which, since  $\pi_1 = \text{proj}_1|_{\mathfrak{M}_c}$ , is equivalent to the condition that the subspace

$$(4.5) \quad \ker \text{Jac}(\text{proj}_1)_{(p_0, q_0)} \cap \ker \text{Jac}(M)_{(p_0, q_0)}$$

is trivial. Now, since

$$\ker \text{Jac}(\text{proj}_1)_{(p_0, q_0)} = \left\{ \begin{bmatrix} 0 \\ y \end{bmatrix} : y \in \mathbb{R}^{2n-r+2} \right\},$$

the intersection (4.5) is parametrized by solutions to the equation  $M_p^{-1}x = 0$ . Since this implies  $x = 0$ , it follows that the intersection (4.5) is the trivial subspace  $\{0\}$ .

In particular, the Jacobian of  $\pi_1$  at  $(p_0, q_0)$  is nonsingular. A similar argument shows that the Jacobian of  $\pi_2$  at  $(p_0, q_0)$  is nonsingular, and, therefore, the final result in this section then follows from the inverse function theorem.

**PROPOSITION 4.2.** *Whenever  $\mathfrak{M}_c$  is nonempty, each of the maps  $\pi_1$  and  $\pi_2$  is a local diffeomorphism.*

**5. Injectivity of  $\pi_2$  and the generalized moment problem for polynomial measures.** Since the map

$$L_+ : \mathfrak{P}_+ \rightarrow \mathfrak{C}_+, \quad p \mapsto \int_b^a \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \frac{P}{Q} dt, \quad k = 0, 1, \dots, n,$$

is linear in  $p$  for a fixed  $q \in \mathfrak{P}_+$ , the inverse image  $\pi_2^{-1}(q)$  in  $\mathfrak{M}_c$  is convex for each fixed  $q \in \mathfrak{P}_+$ . If  $\pi_2^{-1}(q)$  is nonempty, then Proposition 4.2 implies that it consists of a single point.

**LEMMA 5.1.** *The map  $\pi_2$  is an injection.*

The question of whether  $\pi_2^{-1}$  is nonempty is more interesting. To this end, we shall now keep  $q$  fixed and vary  $p$ . It follows from the above that the corresponding

map  $L_+ : \mathfrak{P}_+ \rightarrow \mathfrak{C}_+$  from a positive polynomial  $p$  to a positive sequence  $c$  is a convex injection. To say that  $L_+(p) = c$  is to say that  $(p, q) \in \mathfrak{M}_c$ , so we are interested in the image of  $L_+$ . This is also of independent interest. For example, if  $\mathfrak{P}$  contains the constant functions, choosing  $q = 1$  leads to a special case of Problem 1.6.

PROBLEM 5.2. Given a sequence of complex numbers  $c_0, c_1, \dots, c_n$  and a subspace  $\mathfrak{P}$ , the generalized moment problem for polynomial measures is to parameterize all positive polynomial measures  $P(t)dt$  such that

$$(5.1) \quad \int_a^b \alpha_k(t) P(t) dt = c_k, \quad k = 0, 1, \dots, n.$$

More generally, to say that  $L_+$  is surjective for a fixed  $q \in \mathfrak{P}_+$  is equivalent to asserting that  $L_+ : \partial\mathfrak{P}_+ \rightarrow \partial\mathfrak{C}_+$ . This is trivially true for  $\dim(\mathfrak{P}) = 1$ . In order to analyze the image of  $L_+$ , we shall assume an auxiliary hypothesis, introduced in [8] in a similar context.

HYPOTHESIS 5.3. The zero set of any  $p \in \mathfrak{P}$  has Lebesgue measure zero.

Remark 5.4. Every  $\mathfrak{P}$  spanned by a Chebyshev system (or T-system) satisfies Hypothesis 5.3. In particular, this applies to the power moment problem. The spaces  $\mathfrak{P}$  corresponding to the trigonometric moment problem and the Nevanlinna–Pick interpolation problem satisfy Hypothesis 5.3. More generally, finite-dimensional spaces of analytic functions always satisfy Hypothesis 5.3.

PROPOSITION 5.5. *If Hypothesis 5.3 holds, then the convex injection  $L_+$  fails to be surjective in dimensions greater than one.*

Proof. Following [21] consider the curve

$$U(t) = \begin{pmatrix} u_0(t) \\ u_1(t) \\ \vdots \\ u_n(t) \end{pmatrix}, \quad a \leq t \leq b,$$

and the subset  $U = \{U(t) : t \in [a, b]\} \subset \mathbb{C}^{n+1}$ . Let  $K(U)$  denote its closed convex conic hull. Clearly, the dual cone satisfies  $K(U)^\top = \overline{\mathfrak{P}}_+$ , from which we have the following theorem.

THEOREM 5.6 (see [21]).  $K(U) = \overline{\mathfrak{C}}_+$ .

Therefore to say that  $L_+(\overline{\mathfrak{P}}_+) = \overline{\mathfrak{C}}_+$  is to say that  $U \subset L_+(\overline{\mathfrak{P}}_+)$ . In particular, for each  $t_0 \in [a, b]$  there exists  $p_{t_0} \in \overline{\mathfrak{P}}_+$  such that

$$(5.2) \quad \int_a^b r(t) \frac{P_{t_0}(t)}{Q(t)} dt = r(t_0)$$

for  $r \in \mathfrak{P}$ . Clearly,  $P_{t_0}(t_0) = \int_a^b \frac{P_{t_0}^2(t)}{Q(t)} dt \geq 0$  with equality only if  $P_{t_0} = 0$ . On the other hand,  $P_{t_0} = 0$  is impossible since there exists  $r \in \mathfrak{P}_+$  with  $R(t) > 0$ , which would contradict (5.2).

Now suppose  $r \in \overline{\mathfrak{P}}_+$ . We claim that either  $R$  never vanishes or  $r = 0$ . Indeed, if  $R(t_0) = 0$  for some  $t_0 \in [a, b]$ , then

$$(5.3) \quad \int_a^b R(t) \frac{P_{t_0}(t)}{Q(t)} dt = R(t_0) = 0,$$

and therefore  $R(t)P_{t_0}(t) = 0$  for all  $t \in [a, b]$ . Since  $P_{t_0}(t_0) \neq 0$ , there exists  $\epsilon > 0$  such that  $R(t) = 0$  for  $t \in [t_0 - \epsilon, t_0 + \epsilon]$ , contrary to Hypothesis 5.3.

Since any positive  $R$  satisfies  $R \in \mathfrak{P}_+$ , this implies that  $\partial\mathfrak{P}_+ = \{0\}$  and hence  $n = 0$  and  $\dim(\mathfrak{P}) = 1$ .  $\square$

In particular, this shows that the generalized moment problem for polynomial measures is unsolvable for a set of positive sequences having positive measure, whenever  $\mathfrak{P}$  has dimension at least two.

**6. Hadamard's theorem and bijectivity of  $\pi_1$ .** As in the previous section, we shall begin with an analysis of the “fiber”  $\pi_1^{-1}(p)$  of the map  $\pi_1$  over a fixed  $p$  in  $\mathfrak{P}_+$ , and, as before, for an arbitrary  $c \in \mathfrak{C}_+$ , this leads to a related constrained moment problem, defined as follows.

Consider the function

$$(6.1) \quad F^p : \mathfrak{P}_+ \rightarrow \mathfrak{C}_+,$$

defined componentwise via

$$F_k^p(q) = \int_a^b \alpha_k(t) \frac{P(t)}{Q(t)} dt,$$

and a given positive sequence  $c = (c_0, \dots, c_n)$ .

**PROPOSITION 6.1.** *If  $\mathfrak{P}$  consists of Lipschitz continuous functions, the map (6.1) is a surjective local diffeomorphism.*

*Proof.* We want to show that the image of the map (6.1) is both open and closed in  $\mathfrak{C}_+$ . We begin with a lemma ensuring that  $F^p(\mathfrak{P}_+)$  is open.

**LEMMA 6.2.** *For each  $q \in \mathfrak{P}_+$ ,  $\det \text{Jac}(F^p)|_q \neq 0$ .*

*Proof.* This follows immediately from the fact that  $\text{Jac}(F^p)|_q = M_q$ , where  $M_q$  is the invertible matrix defined in (4.3).  $\square$

In particular, by the inverse function theorem,  $F^p$  is a local diffeomorphism and by the implicit function theorem,  $F^p(\mathfrak{P}_+) \subset \mathfrak{C}_+$  is open. Since  $\mathfrak{C}_+$  is connected, the proposition will follow provided that  $F^p(\mathfrak{P}_+)$  is also closed in  $\mathfrak{C}_+$ .

**LEMMA 6.3.** *If  $\mathfrak{P}$  consists of Lipschitz continuous functions, the map (6.1) is proper.*

*Proof.* We show that, for any compact set  $K$  in  $\mathfrak{C}_+$ ,  $(F^p)^{-1}(K)$  is bounded. To see this, suppose  $(c_j)$  is a sequence in  $\mathfrak{C}_+$  converging to  $c \in \mathfrak{C}_+$  such that  $F^p(q_j) = c_j$  for some  $q_j \in \mathfrak{P}_+$ . We claim that the sequence  $M_j := \|q_j\|$  is bounded in any norm on the vector space  $\mathfrak{P}_+$ . Setting  $r_j := q_j/M_j$ , we first observe

$$F_k^p(q_j) = M_j \int_a^b \alpha_k \frac{P}{R_j} dt.$$

In particular,

$$(6.2) \quad \lim_{j \rightarrow \infty} M_j \int_a^b \frac{P^2}{R_j} dt = \lim_{j \rightarrow \infty} \langle c_j, p \rangle = \langle c, p \rangle > 0.$$

Since the sequence  $(P^2/R_j)$  is bounded away from zero, it follows that the sequence  $(M_j)$  is bounded. Therefore, the preimage of a convergent sequence in  $K$  has a cluster point in the closure of  $\mathfrak{P}_+$ . If  $q$  lies on the boundary of  $\mathfrak{P}_+$ , then  $Q$  is a nonnegative function in  $\mathfrak{P}$  having a zero  $t_0 \in [a, b]$ . Since  $Q$  is Lipschitz continuous at  $t_0$ , by

definition, there exist an  $\varepsilon > 0$  and an  $M > 0$  such that  $Q(t) \leq M|t - t_0|$  whenever  $|t - t_0| < \varepsilon$  and  $t \in [a, b]$ . In particular, if  $t_0 \in (a, b)$ ,

$$\int_a^b \frac{P^2}{Q} dt \geq \frac{1}{M} \int_{t_0-\varepsilon}^{t_0+\varepsilon} \frac{P^2}{|t - t_0|} dt = +\infty,$$

contrary to the assumption. If  $t_0 = a$  or  $t_0 = b$ , a similar estimate holds. Hence,  $q \in \mathfrak{P}_+$ , establishing that  $F^p$  is proper.  $\square$

Since the image of a proper map is closed, this concludes the proof of the proposition.  $\square$

COROLLARY 6.4. *If  $\mathfrak{P}$  consists of Lipschitz continuous functions, then  $\mathfrak{M}_c \neq \emptyset$  for each  $c \in \mathfrak{C}_+$ .*

We have shown that (6.1) is a proper, local diffeomorphism onto the convex set  $\mathfrak{C}_+$ . Since any open convex subset of  $\mathbb{R}^n$  is itself diffeomorphic to  $\mathbb{R}^n$  (see, e.g., [5, p. 771]), (6.1) is a diffeomorphism by Hadamard’s theorem. We record this important fact as follows.

THEOREM 6.5. *If  $\mathfrak{P}$  consists of Lipschitz continuous functions, the mapping*

$$F^p : \mathfrak{P}_+ \rightarrow \mathfrak{C}_+$$

*is a diffeomorphism.*

Remark 6.6. An alternative proof of this result was derived in [8] using convex optimization methods.

COROLLARY 6.7. *If  $\mathfrak{P}$  consists of Lipschitz continuous functions, for each  $c \in \mathfrak{C}_+$  the restriction*

$$\pi_1 : \mathfrak{M}_c \rightarrow \mathfrak{P}_+$$

*of the first projection is bijective. That is, for every positive sequence  $c$  and every choice of  $p$  in  $\mathfrak{P}_+$ , there is a unique  $q$  such that  $(p, q)$  lies in  $\mathfrak{M}_c$ .*

The conclusion of Corollary 6.7 defines, for each fixed  $c \in \mathfrak{C}_+$ , a map

$$g^c : \mathfrak{P}_+ \rightarrow \mathfrak{P}_+,$$

where  $g^c(p)$  is the unique  $q$  such that  $(p, q) \in \mathfrak{M}_c$ . This map was also studied in [8]. In more explicit terms,  $q$  is the unique function in  $\mathfrak{P}_+$  such that  $Q$  is the denominator in the rational measure with numerator  $P$  solving the moment equations

$$\int_a^b \alpha_k(t) \frac{P(t)}{Q(t)} dt = c_k, \quad k = 0, 1, \dots, n,$$

for  $c$ . Moreover, in the language of Theorem 2.2, we see that

$$g^c = \pi_2 \circ \pi_1^{-1}.$$

We summarize these observations in the following result.

COROLLARY 6.8 (see [8]). *The mapping*

$$g^c : \mathfrak{P}_+ \rightarrow \mathfrak{P}_+$$

*is a diffeomorphism onto its image.*



REFERENCES

- [1] M. S. BERGER, *Nonlinearity and Functional Analysis*, Academic Press, New York, 1977.
- [2] C. I. BYRNES, T. T. GEORGIU, AND A. LINDQUIST, *A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint*, IEEE Trans. Automat. Control, 46 (2001), pp. 822–839.
- [3] C. I. BYRNES, T. T. GEORGIU, A. LINDQUIST, AND A. MEGRETSKI, *Generalized interpolation in  $H^\infty$  with a complexity constraint*, Trans. Amer. Math. Soc., 358 (2006), pp. 965–987.
- [4] C. I. BYRNES, S. V. GUSEV, AND A. LINDQUIST, *From finite covariance windows to modeling filters: A convex optimization approach*, SIAM Rev., 43 (2001), pp. 645–675.
- [5] C. I. BYRNES AND A. LINDQUIST, *On the duality between filtering and Nevanlinna-Pick interpolation*, SIAM J. Control Optim., 39 (2000), pp. 757–775.
- [6] C. I. BYRNES AND A. LINDQUIST, *Interior Point Solutions of Variational Problems and Global Inverse Function Theorems*, Technical report TRITA/MAT-01-OS13, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, 2001.
- [7] C. I. BYRNES AND A. LINDQUIST, *A convex optimization approach to generalized moment problems*, in Control and Modeling of Complex Systems: Cybernetics in the 21st Century, K. Hashimoto, Y. Oishi, and Y. Yamamoto, eds., Birkhäuser Boston, Boston, 2003, pp. 3–21.
- [8] C. I. BYRNES AND A. LINDQUIST, *The generalized moment problem with complexity constraint*, Integral Equations Operator Theory, 56 (2006), pp. 163–180.
- [9] C. I. BYRNES, A. LINDQUIST, S. V. GUSEV, AND A. V. MATVEEV, *A complete parameterization of all positive rational extensions of a covariance sequence*, IEEE Trans. Automat. Control, 40 (1995), pp. 1841–1857.
- [10] PH. DELSARTE, Y. GENIN, AND Y. KAMP, *On the role of the Nevanlinna-Pick problem in circuit and system theory*, Internat. J. Circuit Theory Appl., 9 (1981), pp. 177–187.
- [11] PH. DELSARTE, Y. GENIN, Y. KAMP, AND P. VAN DOOREN, *Speech modelling and the trigonometric moment problem*, Philips J. Res., 37 (1982), pp. 277–292.
- [12] J. C. DOYLE, B. A. FRANCIS, AND A. R. TANNENBAUM, *Feedback Control Theory*, Macmillan, New York, 1992.
- [13] T. T. GEORGIU, *Realization of power spectra from partial covariances*, IEEE Trans. Acoust. Speech Signal Process., 35 (1987), pp. 438–449.
- [14] T. T. GEORGIU, *A topological approach to Nevanlinna-Pick interpolation*, SIAM J. Math. Anal., 18 (1987), pp. 1248–1260.
- [15] T. T. GEORGIU, *Solution of the general moment problem via a one-parameter imbedding*, IEEE Trans. Automat. Control, 50 (2005), pp. 811–826.
- [16] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, Los Angeles, 1958.
- [17] J. W. HELTON, *Non-Euclidean functional analysis and electronics*, Bull. Amer. Math. Soc. (N.S.), 7 (1982), pp. 1–64.
- [18] R. E. KALMAN, *Realization of covariance sequences*, in Proceedings of the Toeplitz Memorial Conference, Tel Aviv, Israel, 1981, Oper. Theory Adv. Appl. 4, Birkhäuser, Basel, Boston, 1982, pp. 331–342.
- [19] H. KIMURA, *Robust stabilizability for a class of transfer functions*, IEEE Trans. Automat. Control, 29 (1984), pp. 788–793.
- [20] H. KIMURA, *Positive partial realization of covariance sequences*, in Modelling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 499–513.
- [21] M. G. KREIN AND A. A. NUDELMAN, *The Markov Moment Problem and Extremal Problems*, AMS, Providence, RI, 1977.
- [22] R. NAGAMUNE, *Closed-loop shaping based on Nevanlinna-Pick interpolation with degree constraint*, IEEE Trans. Automat. Control, 49 (2004), pp. 300–305.
- [23] D. SARASON, *Generalized interpolation in  $H^\infty$* , Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.
- [24] M. SPIVAK, *Calculus on Manifolds*, W. A. Benjamin, New York, 1965.
- [25] A. R. TANNENBAUM, *Feedback stabilization of linear dynamical plants with uncertainty in the gain factor*, Internat. J. Control, 32 (1980), pp. 1–16.

## EFFICIENT ON-LINE COMPUTATION OF CONSTRAINED OPTIMAL CONTROL\*

MATO BAOTIĆ<sup>†</sup>, FRANCESCO BORRELLI<sup>‡</sup>, ALBERTO BEMPORAD<sup>§</sup>, AND  
MANFRED MORARI<sup>¶</sup>

**Abstract.** We consider constrained finite-time optimal control problems for discrete-time linear time-invariant systems with constraints on inputs and outputs based on linear and quadratic performance indices. The solution to such problems is a time-varying piecewise affine (PWA) state-feedback law and can be computed by means of multiparametric programming. By exploiting the properties of the value function and the piecewise affine optimal control law of the constrained finite-time optimal control (CFTOC), we propose two new algorithms that avoid storing the polyhedral regions. The new algorithms significantly reduce the on-line storage demands and computational complexity during evaluation of the PWA feedback control law resulting from the CFTOC.

**Key words.** constrained finite time optimal control, multiparametric programming, piecewise affine function evaluation

**AMS subject classifications.** 49N05, 93C05, 93C55, 90C05, 90C20, 90C31

**DOI.** 10.1137/060659314

**1. Introduction.** Recently, in [4, 3], the authors have shown how to compute the solution to the constrained finite-time optimal control (CFTOC) problem as a piecewise affine (PWA) state-feedback law. Such a law is computed off-line by using a multiparametric programming solver [4, 7, 13], which divides the state space into polyhedral regions, and for each region determines the linear gain and offset which produces the optimal control action.

This method reveals its effectiveness when a receding horizon control (RHC) strategy is used [14, 15]. RHC requires solving at each sampling time an open-loop CFTOC problem. The optimal command signal is applied to the process only during the sampling interval that follows. At the next time step a new optimal control problem based on new measurements of the state is solved over a shifted horizon. Having a precomputed solution as an explicit PWA function of the state vector reduces the on-line computation of the RHC control law to a function evaluation, thus avoiding the on-line solution of a quadratic or linear program.

The only drawback of such a PWA feedback control law is that the number of polyhedral regions could grow dramatically with the number of constraints in the optimal control problem. In this paper we focus on efficient on-line methods for the evaluation of such a PWA control law. The simplest algorithm would require (i) the storage of the list of polyhedral regions and of the corresponding affine control laws

---

\*Received by the editors May 8, 2006; accepted for publication (in revised form) May 1, 2008; published electronically September 8, 2008. All of the authors were attending the Automatic Control Laboratory, ETH Zurich, Switzerland, when initial research reported in this paper was carried out.  
<http://www.siam.org/journals/sicon/47-5/65931.html>

<sup>†</sup>Corresponding author. Automatic Control Laboratory, ETH Zentrum - ETL, Physikstrasse 3, CH-8092 Zürich, Switzerland and Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia (mato.baotic@fer.hr).

<sup>‡</sup>Department of Mechanical Engineering, University of California, 5132 Etcheverry Hall, Berkeley, CA 94720-1740 (fborrelli@me.berkeley.edu).

<sup>§</sup>Dip. Ingegneria dell'Informazione, University of Siena, I-53100 Siena, Italy (bemporad@unisi.it).

<sup>¶</sup>Automatic Control Laboratory, ETH Zentrum - ETL, Physikstrasse 3, CH-8092 Zürich, Switzerland (morari@control.ee.ethz.ch).

and (ii) a sequential search through the list of polyhedra for the  $i$ th polyhedron that contains the current state in order to implement the  $i$ th control law. By exploiting the properties of the value function and the optimal control law, for CFTOC problems based on linear programming (LP) and quadratic programming (QP), we propose two new algorithms that avoid storing the polyhedral regions. The new algorithms significantly reduce the on-line storage demands and computational complexity during evaluation of the explicit solution of the CFTOC problem.

The same problem has been recently approached in a different manner in [23]. The algorithm proposed there—with the controller gains of the PWA control law organized on a balanced search tree—is less efficient in terms of memory requirements, but has a logarithmic average computation complexity. At the expense of the optimality of the solution a similar computational complexity can be achieved with an approximate point location algorithm described in [12].

Several papers also propose the use of fast solvers for the on-line solution of constrained predictive control problems (cf. [16, 10]); these algorithms pursue the same goal as this paper, namely, to reduce the on-line computational burden in RHC, but from a different perspective. They use fast LP or QP solvers (in place of a general-purpose solver) tailored to the special dynamic structure of the underlying optimal control problem [16]. Note that a proper comparison of the proposed algorithms with “fast” on-line LP and QP solvers requires the simultaneous analysis of several issues such as speed of computation, storage demand, and real time code verifiability. This is an involved study and as such is outside the scope of this paper.

The paper is organized as follows. For discrete-time linear time-invariant systems the basics of CFTOC problems and of RHC are summarized in section 2. In section 3, for LP-based and QP-based optimal control we present two new algorithms to evaluate on-line explicit optimal control laws and compare their complexity in terms of time and storage against the simplest algorithm mentioned above. Finally, in section 4 an example is given that confirms the efficiency of the new methods.

**2. CFTOC, RHC, and their state-feedback PWA solution.** Throughout this paper (lower and upper case) italic letters denote scalars, vectors, and matrices (e.g.,  $A, a, \dots$ ), while upper case calligraphic letters denote sets (e.g.,  $\mathcal{A}, \mathcal{B}, \dots$ ). For a matrix (vector)  $A$ ,  $A'$  denotes its transpose, while  $A_{(i)}$  denotes the  $i$ th row (element),  $\mathbb{R}$  is the set of real numbers, and  $\mathbb{N}$  is the set of positive integer numbers.

**2.1. CFTOC problem formulation.** Consider the discrete-time linear time-invariant system

$$(1) \quad x(t+1) = Ax(t) + Bu(t)$$

subject to the constraints

$$(2) \quad E^x x(t) + E^u u(t) \leq E$$

at all time instants  $t \geq 0$ .

In (1)–(2),  $n_x \in \mathbb{N}$ ,  $n_u \in \mathbb{N}$ , and  $n_E \in \mathbb{N}$  are the number of states, inputs, and constraints; respectively,  $x(t) \in \mathbb{R}^{n_x}$  is the state vector,  $u(t) \in \mathbb{R}^{n_u}$  is the input vector,  $A \in \mathbb{R}^{n_x \times n_x}$ ,  $B \in \mathbb{R}^{n_x \times n_u}$ ,  $E^x \in \mathbb{R}^{n_E \times n_x}$ ,  $E^u \in \mathbb{R}^{n_E \times n_u}$ ,  $E \in \mathbb{R}^{n_E}$ , the pair  $(A, B)$  is stabilizable, and the vector inequality (2) is considered elementwise.

Let  $x_0 = x(0)$  be the initial state and consider the constrained finite-time optimal

control problem

$$(3) \quad \begin{aligned} J^*(x_0) &:= \min_U J(x_0, U) \\ \text{s.t.} \quad &\begin{cases} x_{k+1} = Ax_k + Bu_k, \\ E^x x_k + E^u u_k \leq E, \quad k = 0, \dots, N-1, \end{cases} \end{aligned}$$

where  $N \in \mathbb{N}$  is the horizon length,  $U := [u'_0, \dots, u'_{N-1}]' \in \mathbb{R}^{n_u N}$  is the optimization vector,  $x_i$  denotes the state at time  $i$  if the initial state is  $x_0$  and the control sequence  $\{u_0, \dots, u_{i-1}\}$  is applied to the system (1),  $J^* : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$  is the value function, and the cost function  $J : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u N} \rightarrow \mathbb{R}$  is given either as a piecewise linear function (i.e., sum of  $l_1$  or  $l_\infty$  norms)

$$(4a) \quad J(x_0, U) = \|Q^{x_N} x_N\|_\ell + \sum_{k=0}^{N-1} \|Q^x x_k\|_\ell + \|Q^u u_k\|_\ell, \quad \ell \in \{1, \infty\},$$

or as a quadratic function

$$(4b) \quad J(x_0, U) = x'_N Q^{x_N} x_N + \sum_{k=0}^{N-1} x'_k Q^x x_k + u'_k Q^u u_k.$$

In the following, we will assume that  $Q^x$ ,  $Q^u$ ,  $Q^{x_N}$  are full column rank matrices when the cost function (4a) is used, and that  $Q^x = (Q^x)' \succeq 0$ ,  $Q^u = (Q^u)' \succ 0$ ,  $Q^{x_N} \succeq 0$ , when the cost function (4b) is used, where  $Q \succ 0$  denotes positive definiteness (resp.,  $Q \succeq 0$  positive semidefiniteness).

The optimization problem (3) can be translated into a linear program (LP) when the piecewise linear cost function (4a) is used [3] or into a quadratic program (QP) when the quadratic cost function (4b) is used [4]. We denote by  $U^* = [(u_0^*)', \dots, (u_{N-1}^*)']'$  one of the possible optimizers of problem (3)–(4). An optimizer  $U^*$  can be computed by solving an LP or a QP once  $x_0$  is fixed or can be computed explicitly for all  $x_0$  within a given range of values as explained in subsections 2.3 and 2.4.

**2.2. RHC strategy.** Consider the problem of regulating the discrete-time linear time-invariant system (1) to the origin while fulfilling the constraints (2). The solution  $U^*$  to CFTOC problem (3)–(4) is an open-loop optimal control trajectory over a finite horizon. A receding horizon control strategy employs it to obtain a feedback control law in the following way: Assume that a full measurement of the state  $x(t)$  is available at the current time  $t \geq 0$ . Then, the CFTOC problem (3)–(4) is solved at each time  $t$  for  $x_0 = x(t)$ , and

$$(5) \quad u(t) = u_0^*$$

is applied as an input to system (1). For a detailed discussion on RHC strategy, see, e.g., [21, 9, 18, 4, 3, 14].

**2.3. Solution of CFTOC, linear cost case.** Consider the problem (3) with the piecewise linear cost function (4a) and  $\ell = \infty$ . Using a standard transformation [3], we introduce the vector  $v := [u'_0, \dots, u'_{N-1}, \varepsilon_1^x, \dots, \varepsilon_N^x, \varepsilon_0^u, \dots, \varepsilon_{N-1}^u]' \in \mathbb{R}^{n_v}$ ,  $n_v :=$

$(n_u + 2)N$ ,  $\varepsilon_k^x \geq \|Q^x x_k\|_\infty$ ,  $\varepsilon_N^x \geq \|Q^{x_N} x_N\|_\infty$ ,  $\varepsilon_k^u \geq \|Q^u u_k\|_\infty$ , and substitute  $x_k = A^k x_0 + \sum_{j=0}^{k-1} A^j B u_{k-1-j}$  in (3)–(4), which can be rewritten as the linear program<sup>1</sup>

$$(6) \quad \begin{aligned} J^*(x) &:= \min_v \quad c^T v \\ \text{s.t.} \quad &L^v v \leq L + L^x x, \end{aligned}$$

where  $x = x_0$ , and matrices  $c \in \mathbb{R}^{n_v}$ ,  $L^v \in \mathbb{R}^{n_L \times n_v}$ ,  $L^x \in \mathbb{R}^{n_L \times n_x}$ ,  $L \in \mathbb{R}^{n_L}$  are easily obtained from  $Q^x$ ,  $Q^u$ ,  $Q^{x_N}$  and (3)–(4), as explained in [3]. For a given  $x$  we denote with  $\mathcal{V}^*(x)$  the set of optimizers for the problem (6). Note that, in general,  $\mathcal{V}^*(x)$  is a set valued function, i.e.,  $\mathcal{V}^* : \mathbb{R}^{n_x} \rightarrow 2^{\mathbb{R}^{n_v}}$ .

Because the problem depends on  $x$  the implementation of RHC can be performed in two different ways: solve the LP (6) on-line at each time step for a given  $x$  or solve (6) *off-line* for all  $x$  within a given range of values, i.e., by considering (6) as a *multiparametric linear program* (mp-LP) [11].

Solving an mp-LP means computing the value function  $J^*(x) : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$  and one (out of possibly many) optimizer function  $v^*(x) : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_v}$  for all possible vectors  $x$  in a given set  $\mathcal{X}$ . The solution to mp-LP problems can be approached simply by exploiting the properties of the primal and dual optimality conditions as proposed in [7, 11].

In [11] the following results about the properties of the solution are proved.

**THEOREM 1.** *Consider the mp-LP (6). Then the set of feasible parameters  $\mathcal{X}_f$  is convex. The value function  $J^* : \mathcal{X}_f \rightarrow \mathbb{R}$  is convex and piecewise affine. There always exists a continuous and PWA selection of an optimizer function  $v^* : \mathcal{X}_f \rightarrow \mathbb{R}^{n_v}$ . In particular, if the optimizer  $\mathcal{V}^*(x)$  is unique for all  $x \in \mathcal{X}_f$ , then  $v^*(x) = \mathcal{V}^*(x)$ .*

Once the multiparametric problem (6) has been solved off-line for a polyhedral set  $\mathcal{X} \subseteq \mathbb{R}^{n_x}$  of states, the explicit solution  $v^*(x)$  of CFTOC problem (6) is available as a PWA function of  $x$ , and the receding horizon controller (3)–(5) is also available explicitly, as the optimal input  $u(t)$  consists simply of  $n_u$  components of  $v^*(x(t))$ ,

$$(7) \quad u(t) = [I_{n_u} \ 0 \ \cdots \ 0] v^*(x(t)).$$

**COROLLARY 1.** *The RHC (7), defined by the optimization problem (3), (4a), and (5), is a continuous and piecewise affine function,  $u : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_u}$ , and has the form*

$$(8) \quad u(x) = F_i x + G_i \quad \forall x \in \mathcal{P}_i, \quad i = 1, \dots, N_{\mathcal{P}},$$

where  $F_i \in \mathbb{R}^{n_u \times n_x}$ ,  $G_i \in \mathbb{R}^{n_u}$ ,  $\{\mathcal{P}_i\}_{i=1}^{N_{\mathcal{P}}}$  is a polyhedral partition of  $\mathcal{X}_f$  (i.e.,  $\cup_i \mathcal{P}_i = \mathcal{X}_f$ , and  $\mathcal{P}_i$  and  $\mathcal{P}_j$  have disjoint interiors  $\forall i \neq j$ ), with  $\mathcal{P}_i = \{x \in \mathbb{R}^{n_x} \mid P_i^x x \leq P_i\}$ ,  $P_i^x \in \mathbb{R}^{p_i \times n_x}$ ,  $P_i \in \mathbb{R}^{p_i}$ , and  $p_i$  is the number of halfspaces defining polyhedron  $\mathcal{P}_i$ ,  $i = 1, \dots, N_{\mathcal{P}}$ .

In the rest of this paper we will assume, without loss of generality, that  $\mathcal{P}_i$  in (8) and  $i = 1, \dots, N_{\mathcal{P}}$ , are full dimensional sets in  $\mathbb{R}^{n_x}$  corresponding to the so-called critical regions of the optimization problem (6) (see [7] for more details). We will denote with  $N_{\mathcal{H}}$  the total number of halfspaces defining the polyhedral partition of  $\mathcal{X}_f$ ,

$$(9) \quad N_{\mathcal{H}} := \sum_{i=1}^{N_{\mathcal{P}}} p_i.$$

<sup>1</sup>The same holds for  $\ell = 1$  with a different optimization vector [3].

*Remark 2.1.* Typically the total number of halfspaces defining polyhedral partition of feasible set  $\mathcal{X}_f$  is much bigger than the number of polyhedral regions in it, i.e.,  $N_{\mathcal{H}} \gg N_{\mathcal{P}}$ . The reasoning is the following. Assume, as is the case in practical applications, that all  $\mathcal{P}_i$  are bounded. Since the smallest number of halfspaces defining a bounded polyhedron in  $\mathbb{R}^{n_x}$  is  $n_x + 1$  (achieved by a simplex), we have  $N_{\mathcal{H}} \geq (n_x + 1)N_{\mathcal{P}}$ .

**2.4. Solution of CFTOC, quadratic cost case.** Consider the problem (3) with the quadratic cost function (4b). By substituting  $x_k = A^k x_0 + \sum_{j=0}^{k-1} A^j B u_{k-1-j}$  in (3)–(4), this can be rewritten as the quadratic program

$$(10) \quad \begin{aligned} J^*(x) = \frac{1}{2} x' Y x + \min_U \quad & \frac{1}{2} U' H U + x' F U \\ \text{s.t.} \quad & M^U U \leq M + M^x x, \end{aligned}$$

where  $x = x_0$ , the column vector  $U := [u'_0, \dots, u'_{N-1}]' \in \mathbb{R}^{n_U}$ ,  $n_U := n_u N$ , is the optimization vector,  $H = H' \succ 0$ , and  $H, F, Y, M^U, M^x, M$  are easily obtained from  $Q^x, Q^u, Q^{x_N}$  and (3)–(4) (see [4] for details).

As in the linear cost case, because the problem depends on  $x$ , the implementation of RHC can be performed by either solving the QP (10) on-line or, as shown in [4, 22], by solving problem (10) *off-line* for all  $x$  within a given range of values, i.e., by considering (10) as a *multiparametric quadratic program* (mp-QP).

Once the multiparametric problem (10) is solved off-line, i.e., the solution  $U^*(x)$  of the CFTOC problem (10) is found, the state-feedback PWA RHC law is simply

$$(11) \quad u(t) = [I_{n_u} \ 0 \ \cdots \ 0] U^*(x(t)).$$

In [4] the authors give a self-contained proof of the following properties of the solution.

**THEOREM 2.** *Consider the multiparametric quadratic program (10), and let  $H \succ 0$ . Then the set of feasible parameters  $\mathcal{X}_f$  is convex, the optimizer  $U^* : \mathcal{X}_f \rightarrow \mathbb{R}^s$  is continuous and piecewise affine, and the value function  $J^* : \mathcal{X}_f \rightarrow \mathbb{R}$  is continuous, convex, and piecewise quadratic.*

The proof of the properties listed in Theorem 2 can be found in [5, “Maximum Theorem” on page 116]. It also follows from [1, Theorem 3.2.1(I) and Theorem 3.3.3].

**COROLLARY 2.** *The RHC control law (11), defined by the optimization problem (3), (4b), and (5), is continuous and piecewise affine and has the form (8).*

The optimization problem (10), where  $\mathcal{X}_f$  is a lower dimensional set, can be dealt with in the same way as in the linear cost case (see [7] for details). Hence, in the following we will assume, without loss of generality, that  $\mathcal{P}_i$ ,  $i = 1, \dots, N_{\mathcal{P}}$ , are full dimensional sets in  $\mathbb{R}^{n_x}$  corresponding to the so-called critical regions of the optimization problem (10); cf. [4].

Corollaries 1 and 2 state that by using a multiparametric solver the computation of RHC action becomes a simple PWA function evaluation. In the next section we propose a method to efficiently evaluate such a PWA function without storing the polyhedral regions  $\mathcal{P}_i$ ,  $i = 1, \dots, N_{\mathcal{P}}$ .

**3. Efficient on-line algorithms.** The on-line implementation of the control law (8) is executed simply according to the following algorithm.

## ALGORITHM 1.

1. Measure the current state  $x(t)$
2. Search for the  $i$ th polyhedron that contains  $x(t)$ ,  $(P_i^x x(t) \leq P_i)$
3. Implement the  $i$ th control law  $(u(t) = F_i x(t) + G_i)$

In Algorithm 1, step 2 is critical and is the only step whose efficiency can be improved. A simple implementation of step 2 would consist of searching for the polyhedral region that contains the state  $x(t)$  as in the following algorithm.

## ALGORITHM 2.

1.  $i = 0$ , notfound=TRUE
2. WHILE  $i \leq N_{\mathcal{P}}$  AND notfound
  - 2.1.  $j = 0$ , feasible=TRUE
  - 2.2. WHILE  $j \leq p_i$  AND feasible
    - 2.2.1. IF  $(P_i^x)_{(j)} x(t) > (P_i)_{(j)}$  THEN feasible=FALSE
  - 2.3. END
  - 2.4. IF feasible THEN notfound=FALSE
3. END

Recalling the expression (9) for  $N_{\mathcal{H}}$  (the total number of halfspaces defining the polyhedral partition of the feasible set  $\mathcal{X}_f$ ), it is easy to see that Algorithm 2 requires  $(n_x + 1)N_{\mathcal{H}}$  real numbers to store all polyhedra  $\mathcal{P}_i$ , and in the worst case (when the state is contained in the last region of the list) Algorithm 2 will give a solution after  $n_x N_{\mathcal{H}}$  multiplications,  $(n_x - 1)N_{\mathcal{H}}$  sums, and  $N_{\mathcal{H}}$  comparisons.

*Remark 3.1.* In the algorithms presented in the following sections we implicitly assume that  $x(t)$  belongs to the feasible set  $\mathcal{X}_f$ . If this (reasonable) assumption does not hold, then we should include a set of *boundaries* of feasible parameter space  $\mathcal{X}_f$ , and we should (before using any of proposed algorithms) first check if the point  $x(t)$  is inside the boundaries of  $\mathcal{X}_f$ . Note that such a step is not needed for Algorithm 2 since there we automatically detect if the point  $x(t)$  is outside of the feasible set  $\mathcal{X}_f$ .

By using the properties of the value function, we will show how Algorithm 2 can be replaced by more efficient algorithms that have less computational complexity and *avoid storing the polyhedral regions*  $\mathcal{P}_i$ ,  $i = 1, \dots, N_{\mathcal{P}}$ , therefore reducing the storage demand significantly.

In the following we will distinguish between optimal control based on LP and optimal control based on QP.

**3.1. Efficient implementation, linear cost case.** From Theorem 1, the value function  $J^*(x)$  corresponding to the solution of the CFTOC problem (3) with the piecewise linear cost (4a) is convex and PWA:

$$(12) \quad J^*(x) = T'_i x + V_i \quad \forall x \in \mathcal{P}_i, \quad i = 1, \dots, N_{\mathcal{P}},$$

where  $T_i \in \mathbb{R}^{n_x}$ ,  $V_i \in \mathbb{R}$ .

By exploiting the convexity of the value function the storage of the polyhedral regions  $\mathcal{P}_i$  can be avoided. From the equivalence of the representations of PWA convex functions (cf. [17], [8, page 80]) the function  $J^*(x)$  in (12) can be represented alternatively as

$$(13) \quad J^*(x) = \max \{T'_i x + V_i\}_{i=1}^{N_{\mathcal{P}}} \quad \text{for } x \in \mathcal{X}_f = \cup_{i=1}^{N_{\mathcal{P}}} \mathcal{P}_i.$$

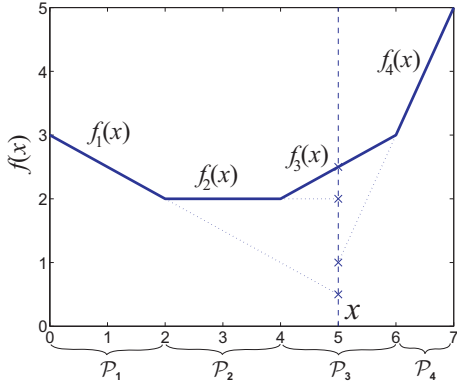


FIG. 1. Example for Algorithm 3 in one dimension: For a given point  $x \in \mathcal{P}_3$  ( $x = 5$ ), we have  $f_3(x) = \max(f_1(x), f_2(x), f_3(x), f_4(x))$ .

TABLE 1  
Complexity comparison of Algorithms 2 and 3.

	Algorithm 2	Algorithm 3
Storage demand (real numbers)	$(n_x + 1)N_{\mathcal{H}}$	$(n_x + 1)N_{\mathcal{P}}$
Number of flops (worst case)	$2n_x N_{\mathcal{H}}$	$2n_x N_{\mathcal{P}}$

From (12) and (13), the polyhedral region  $\mathcal{P}_j$  containing  $x$  can be identified simply by searching for the maximum number in the list  $\{T'_i x + V_i\}_{i=1}^{N_{\mathcal{P}}}$ ,

(14) 
$$x \in \mathcal{P}_j \Leftrightarrow T'_j x + V_j = \max \{T'_i x + V_i\}_{i=1}^{N_{\mathcal{P}}}.$$

Therefore, instead of searching for the polyhedron  $j$  that contains the point  $x$  via Algorithm 2, we can just store the value function and identify region  $j$  by searching for the maximum in the list of numbers composed of the single affine function  $T'_i x + V_i$  evaluated at  $x$ .

ALGORITHM 3.

- 1. Compute the list  $\mathcal{T} = \{t_i := T'_i x + V_i\}_{i=1}^{N_{\mathcal{P}}}$
- 2. Find  $i$  such that  $t_i = \max_{t_j \in \mathcal{T}} t_j$

For illustration, see the example in Figure 1, where we have  $N_{\mathcal{P}} = 4$ ,  $f_1(x) = -0.5x + 3$ ,  $f_2(x) = 2$ ,  $f_3(x) = 0.5x$ , and  $f_4(x) = 2x - 9$ .

Algorithm 3 requires the storage of  $(n_x + 1)N_{\mathcal{P}}$  real numbers and will give a solution after  $n_x N_{\mathcal{P}}$  multiplications,  $(n_x - 1)N_{\mathcal{P}}$  sums, and  $N_{\mathcal{P}} - 1$  comparisons. In Table 1 we compare the complexity of Algorithm 3 against Algorithm 2 in terms of storage demand and number of flops.

*Remark 3.2.* Algorithm 3 will outperform Algorithm 2 since typically the total number of halfspaces defining polyhedral partition of feasible set  $\mathcal{X}_f$  is much larger than the number of polyhedral regions, i.e.,  $N_{\mathcal{H}} \gg N_{\mathcal{P}}$  (see Remark 2.1).

*Remark 3.3.* Note that Algorithm 3 cannot be applied if the solution to (6) contains dual degenerate regions, i.e., two regions  $\mathcal{P}_i$  and  $\mathcal{P}_j$  that have the same cost expressions but different control expressions (i.e.,  $[T'_i \ V_i] = [T'_j \ V_j]$ ,  $[F'_i \ G_i] \neq [F'_j \ G_j]$ ). If dual degeneracy occurs one can use Algorithm 4 combined with the procedure



described in section 3.2.2. An alternative approach consists of modifying Algorithm 3 in order to be able to discern between dual degenerate regions (e.g., by means of Algorithm 2).

**3.2. Efficient implementation, quadratic cost case.** Consider the explicit solution of CFTOC problem (3) with the quadratic cost (4b). Theorem 2 states that the value function  $J^*(x)$  is convex and piecewise quadratic and the simple Algorithm 3 described in the previous subsection cannot be used here. Instead, a modified approach is described below.

We will first establish the following general result: given a general polyhedral partition of the state space, we can locate where the state lies (i.e., in which polyhedron) by using a search procedure based on the information provided by an “appropriate” PWA continuous function defined over the same polyhedral partition. We will refer to such an “appropriate” PWA function as a *PWA descriptor function*. In the following, first we outline the properties of the PWA descriptor function and then we describe the search procedure itself. In later subsections we will finally show how the gradient of the value function (under certain regularity conditions) and the optimizer (always) can be used for the construction of PWA descriptor functions.

**DEFINITION 1.** *Two polyhedra  $\mathcal{P}_i, \mathcal{P}_j$  of  $\mathbb{R}^{n_x}$  are called neighboring polyhedra if their interiors are disjoint and  $\mathcal{P}_i \cap \mathcal{P}_j$  is  $(n_x - 1)$ -dimensional (i.e., is a common facet).*

Let  $\{\mathcal{P}_i\}_{i=1}^{N_P}$  be the polyhedral partition obtained by solving the mp-QP (10). For each polyhedra  $\mathcal{P}_i$  we denote with  $\mathcal{C}_i$  the list of all its neighbors,

$$(15) \quad \mathcal{C}_i := \{j \mid \mathcal{P}_j \text{ is a neighbor of } \mathcal{P}_i, j = 1, \dots, N_P, j \neq i\}, \quad i = 1, \dots, N_P.$$

In the following, we will assume that every facet is shared by only two neighboring polyhedral regions. This so-called *facet-to-facet* property is almost always satisfied by the solution of the mp-QP (10); cf. [19]. In such a case the list  $\mathcal{C}_i$  has at most  $p_i$  elements ( $p_i$  is the number of halfspaces defining polyhedron  $\mathcal{P}_i$ ) since some of the boundaries of  $\mathcal{P}_i$  may be outer boundaries of the polyhedral partition  $\{\mathcal{P}_i\}_{i=1}^{N_P}$  and they would not introduce element in the list  $\mathcal{C}_i$ . We give the following definition of a PWA descriptor function.

**DEFINITION 2 (PWA descriptor function).** *A scalar continuous real-valued PWA function  $f : \mathcal{X}_f \rightarrow \mathbb{R}$*

$$(16) \quad f(x) = f_i(x) := A'_i x + B_i \quad \text{if } x \in \mathcal{P}_i,$$

*with  $A_i \in \mathbb{R}^{n_x}$ ,  $B_i \in \mathbb{R}$ , is called a descriptor function if*

$$(17) \quad A_i \neq A_j \quad \forall j \in \mathcal{C}_i, \quad i = 1, \dots, N_P,$$

*where  $\cup_i \mathcal{P}_i = \mathcal{X}_f \subset \mathbb{R}^{n_x}$ , and  $\mathcal{C}_i$  is the list of neighbors of  $\mathcal{P}_i$ .*

In the following, we will show that the PWA descriptor function defined above has all of the properties we need to be able to locate in which polyhedron the state  $x$  lies, because the sign of  $f_i(x) - f_j(x)$  changes only when the point  $x$  crosses the separating hyperplane between  $\mathcal{P}_i$  and  $\mathcal{P}_j$ . Thus for all  $x$  in  $\mathcal{P}_i$ , the difference  $f_i(x) - f_j(x)$  has the same sign.

**DEFINITION 3 (ordering function).** *Let  $f(x)$  be a PWA descriptor function on the polyhedral partition  $\{\mathcal{P}_i\}_{i=1}^{N_P}$ . We define ordering function  $O_i(x)$  as*

$$(18) \quad O_i(x) := [O_{i,j}(x)]_{j \in \mathcal{C}_i},$$

where

$$(19) \quad O_{i,j}(x) := \begin{cases} +1 & \text{if } f_i(x) \geq f_j(x), \\ -1 & \text{if } f_i(x) < f_j(x), \end{cases}$$

with  $i \in \{1, \dots, N_{\mathcal{P}}\}$ ,  $j \in \mathcal{C}_i$ . Note that, for simplicity, we will assume that the order in which the elements of  $\mathcal{C}_i$  are used when creating  $O_i(x)$  in (18) is uniquely defined. Namely, we use sorted (e.g., in an increasing order) list  $\mathcal{C}_i$ .

**THEOREM 3.** Let  $f(x)$  be a PWA descriptor function on the polyhedral partition  $\{\mathcal{P}_i\}_{i=1}^{N_{\mathcal{P}}}$ . Let  $\xi_i \in \mathbb{R}^{n_x}$  be any point in the interior of  $\mathcal{P}_i$ , and define

$$(20) \quad S_{i,j} := O_{i,j}(\xi_i),$$

$$(21) \quad S_i := O_i(\xi_i),$$

with  $i = 1, \dots, N_{\mathcal{P}}$ ,  $j \in \mathcal{C}_i$ . Then the following holds:

$$(22) \quad x \in \text{int}(\mathcal{P}_i) \Leftrightarrow O_{i,j}(x) = S_{i,j} \quad \forall j \in \mathcal{C}_i \Leftrightarrow O_i(x) = S_i.$$

*Proof.* Let  $\mathcal{F} = \mathcal{P}_i \cap \mathcal{P}_j$  be the common facet of two neighboring polyhedra  $\mathcal{P}_i$  and  $\mathcal{P}_j$ . Define the linear function

$$(23) \quad g_{i,j}(x) = f_i(x) - f_j(x).$$

From the continuity of descriptor function  $f(x)$ , it follows that  $g_{i,j}(x) = 0 \quad \forall x \in \mathcal{F}$ . As  $\mathcal{P}_i$  and  $\mathcal{P}_j$  are disjoint convex polyhedra and  $A_i \neq A_j$ , it follows that  $g_{i,j}(x) = 0$  is a separating hyperplane between  $\mathcal{P}_i$  and  $\mathcal{P}_j$ . We have the following.

(i) “ $\Rightarrow$ ” part. Since  $g_{i,j}(x) = 0$  is a separating hyperplane between  $\mathcal{P}_i$  and  $\mathcal{P}_j$  it follows that  $g_{i,j}(x)$  does not change sign for all  $x \in \text{int}(\mathcal{P}_i)$ . Hence, we have  $O_{i,j}(x) = S_{i,j}$ .

(ii) “ $\Leftarrow$ ” part by contradiction. Assume that  $O_{i,j}(\bar{x}) = S_{i,j}$  while  $\bar{x} \notin \mathcal{P}_i$ . It is easy to see that  $\forall \bar{x} \notin \mathcal{P}_i$ ,  $\exists j \in \mathcal{C}_i$  such that  $(P_i^x)_{(j)}\bar{x} > (P_i)_{(j)}$ . This, however, implies that  $g_{i,j}(\bar{x})$  has a different sign compared to  $g_{i,j}(\xi)$ ,  $\xi \in \text{int}(\mathcal{P}_i)$ . Hence we have  $O_{i,j}(\bar{x}) \neq S_{i,j}$ , a contradiction.  $\square$

Theorem 3 states that the ordering function  $O_i(x)$  and the vector  $S_i$  uniquely characterize  $\mathcal{P}_i$ . Therefore, to check on-line if the polyhedral region  $\mathcal{P}_i$  contains the state  $x$  it is sufficient to compute the binary vector  $O_i(x)$  and compare it with  $S_i$ . Vectors  $S_i$  are calculated off-line for  $i = 1, \dots, N_{\mathcal{P}}$ , by comparing the values of  $f_i(x)$  and  $f_j(x) \quad \forall j \in \mathcal{C}_i$ , in a point  $\xi$  belonging to  $\text{int}(\mathcal{P}_i)$ , for instance, the Chebyshev center of  $\mathcal{P}_i$ .

In Figure 2 a one dimensional example illustrates the procedure with  $N_{\mathcal{P}} = 4$  regions and for  $f_1(x) = x$ ,  $f_2(x) = 2$ ,  $f_3(x) = x - 3$ ,  $f_4(x) = -\frac{1}{3}x + \frac{19}{3}$ . The list of neighboring regions  $\mathcal{C}_i$  and the vector  $S_i$  can be constructed by simply looking at the figure:  $\mathcal{C}_1 = \{2\}$ ,  $\mathcal{C}_2 = \{1, 3\}$ ,  $\mathcal{C}_3 = \{2, 4\}$ ,  $\mathcal{C}_4 = \{3\}$ ,  $S_1 = -1$ ,  $S_2 = [-1 \ 1]$ ,  $S_3 = [1 \ -1]$ ,  $S_4 = -1$ . The point  $x = 4$  is in region 2 and we have  $O_2(x) = [-1 \ 1] = S_2$ , while  $O_3(x) = [-1 \ -1] \neq S_3$ ,  $O_1(x) = 1 \neq S_1$ ,  $O_4(x) = 1 \neq S_4$ . The failure of a match  $O_i(x) = S_i$  provides information on good search direction(s). The solution can be found by searching in the direction where a constraint is violated, i.e., we should check the neighboring region  $\mathcal{P}_j$  for which  $O_{i,j}(x) \neq S_{i,j}$ .

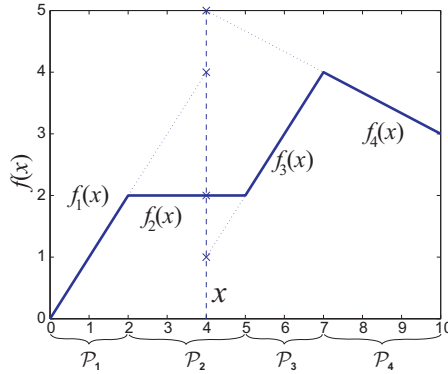


FIG. 2. Example for Algorithm 4 in one dimension: For a given point  $x \in \mathcal{P}_2$  ( $x = 4$ ) we have  $O_2(x) = [-1 \ 1] = S_2$ , while  $O_1(x) = 1 \neq S_1 = -1$ ,  $O_3(x) = [-1 \ -1] \neq S_3 = [1 \ -1]$ ,  $O_4(x) = 1 \neq S_4 = -1$ .

The overall procedure is composed of two parts:

1. (*off-line*). Construction of the scalar continuous real-valued PWA function  $f(\cdot)$  in (16) satisfying (17), and computation of the list of neighbors  $\mathcal{C}_i$  and the vector  $S_i$ ,
2. (*on-line*). Search for the  $i$ th polyhedron that contains  $x(t)$  by the execution of the following algorithm.

ALGORITHM 4.

1.  $\mathcal{I} = \{1, \dots, N_{\mathcal{P}}\}$
2.  $i \leftarrow \mathcal{I}$
3.  $\mathcal{I} = \mathcal{I} \setminus \{i\}$ ,  $\mathcal{C} = \mathcal{C}_i$
4. WHILE  $\mathcal{C} \neq \emptyset$ 
  - 4.1.  $j \leftarrow \mathcal{C}$ ,  $\mathcal{C} = \mathcal{C} \setminus \{j\}$
  - 4.2. Compute  $O_{i,j}(x)$
  - 4.3. IF  $O_{i,j}(x) \neq S_{i,j}$ 
    - 4.3.1. IF  $j \notin \mathcal{I}$  THEN GOTO step 2. ELSE  $i = j$  and GOTO step 3.
  - 4.4. END
5. END

Algorithm 4 does not require the storage of the polyhedra  $\mathcal{P}_i$ , but only the storage of one affine function  $f_i(x)$  per polyhedron, i.e.,  $N_{\mathcal{P}}(n_x + 1)$  real numbers and the list of neighbors  $\mathcal{C}_i$  which demands (at most)  $N_{\mathcal{H}}$  integers. Algorithm 4 in the worst case terminates after  $N_{\mathcal{P}}n_x$  multiplications,  $N_{\mathcal{P}}(n_x - 1)$  sums, and  $N_{\mathcal{H}}$  comparisons.

In Table 2 we compare the complexity of Algorithm 4 against the standard Algorithm 2 in terms of storage demand and the number of flops.

*Remark 3.4.* Note that the computation of  $O_i(x)$  in Algorithm 4 requires the evaluation of  $p_i$  linear functions, but the overall computation never exceeds  $N_{\mathcal{P}}$  linear function evaluations. Consequently, Algorithm 4 will outperform Algorithm 2, since typically  $N_{\mathcal{H}} \gg N_{\mathcal{P}}$ .

Now that we have shown how to locate the polyhedron in which the state lies by using a PWA descriptor function, we need a procedure for the construction of such a function.

TABLE 2  
Complexity comparison of Algorithms 2 and 4.

	Algorithm 2	Algorithm 4
Storage demand (real numbers)	$(n_x + 1)N_{\mathcal{H}}$	$(n_x + 1)N_{\mathcal{P}}$
Number of flops (worst case)	$2n_x N_{\mathcal{H}}$	$(2n_x - 1)N_{\mathcal{P}} + N_{\mathcal{H}}$

The image of the descriptor function is the set of real numbers  $\mathbb{R}$ . In the following, we will show how a descriptor function can be generated from a vector valued function  $m : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^s$ . This general result will be used in the next subsections.

DEFINITION 4 (vector valued PWA descriptor function). *A continuous vector valued PWA function  $m : \mathcal{X}_f \rightarrow \mathbb{R}^s$ ,*

$$(24) \quad m(x) = \bar{A}_i x + \bar{B}_i \quad \text{if } x \in \mathcal{P}_i,$$

*is called a vector valued PWA descriptor function if*

$$(25) \quad \bar{A}_i \neq \bar{A}_j \quad \forall j \in \mathcal{C}_i, \quad i = 1, \dots, N_{\mathcal{P}},$$

where  $\cup_i \mathcal{P}_i = \mathcal{X}_f \subset \mathbb{R}^{n_x}$ ,  $\bar{A}_i \in \mathbb{R}^{s \times n_x}$ ,  $\bar{B}_i \in \mathbb{R}^s$ ,  $s \in \mathbb{N}$ ,  $s \geq 2$ , and  $\mathcal{C}_i$  is the list of neighbors of  $\mathcal{P}_i$ .

THEOREM 4. *Let  $m : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^s$  be a vector valued PWA descriptor function defined over a polyhedral partition  $\{\mathcal{P}_i\}_{i=1}^{N_{\mathcal{P}}}$ . Then  $\exists w \in \mathbb{R}^s$  such that  $f(x) := w'm(x)$  is a PWA descriptor function defined over the same polyhedral partition.*

*Proof.* Let  $\mathcal{N}_{i,j}$  be the null-space of  $(\bar{A}_i - \bar{A}_j)'$ . Since, by definition,  $\bar{A}_i - \bar{A}_j \neq \mathbf{0}$ , it follows that  $\mathcal{N}_{i,j}$  is not full dimensional, i.e.,  $\mathcal{N}_{i,j} \subseteq \mathbb{R}^{s-1}$ . Consequently, it is always possible to find a vector  $w \in \mathbb{R}^s$  such that  $w'(\bar{A}_i - \bar{A}_j) \neq \mathbf{0}$  holds for all  $i = 1, \dots, N_{\mathcal{P}}$  and  $\forall j \in \mathcal{C}_i$ . Clearly,  $f(x) = w'm(x)$  is then a valid PWA descriptor function.  $\square$

As shown in the proof of Theorem 4, once we have vector valued PWA descriptor function, practically any randomly chosen vector  $w \in \mathbb{R}^s$  is likely to be satisfactory for the construction of a PWA descriptor function. From a numerical point of view, however, we would like to obtain  $w$ , which is as far away as possible from the null-spaces  $\mathcal{N}_{i,j}$ . We show one algorithm for finding such a vector  $w$ .

For a given vector valued PWA descriptor function we form a set of vectors  $a_k \in \mathbb{R}^s$ ,  $\|a_k\| = 1$ ,  $k = 1, \dots, N_a$ , by taking and normalizing one (and only one) nonzero column from each matrix  $(\bar{A}_i - \bar{A}_j) \forall j \in \mathcal{C}_i$ ,  $i = 1, \dots, N_{\mathcal{P}}$ . Here  $N_a := \sum_i |\mathcal{C}_i|/2 \leq N_{\mathcal{H}}$  and  $|\mathcal{C}_i|$  denotes cardinality of set  $\mathcal{C}_i$ . The vector  $w \in \mathbb{R}^s$  satisfying the set of equations  $w'a_k \neq 0$ ,  $k = 1, \dots, N_a$ , can then be constructed by using the following algorithm.<sup>2</sup>

ALGORITHM 5.

1.  $w \leftarrow [1, \dots, 1]'$ ,  $R \leftarrow 1$
2. WHILE  $k \leq N_a$ 
  - 2.1.  $d \leftarrow w'a_k$
  - 2.2. IF  $0 \leq d \leq R$  THEN  $w \leftarrow w + \frac{1}{2}(R - d)a_k$ ,  $R \leftarrow \frac{1}{2}(R + d)$
  - 2.3. IF  $-R \leq d < 0$  THEN  $w \leftarrow w - \frac{1}{2}(R + d)a_k$ ,  $R \leftarrow \frac{1}{2}(R - d)$
3. END

Essentially, Algorithm 5 constructs a sequence of balls  $\mathcal{B} = \{x \mid x = w + r, \|r\|_2 \leq R\}$ . As depicted in Figure 3, we start with the initial ball of radius  $R = 1$ ,

<sup>2</sup>Index  $k$  goes to  $N_a := \sum_i |\mathcal{C}_i|/2$  since the term  $(\bar{A}_j - \bar{A}_i)$  is the same as  $(\bar{A}_i - \bar{A}_j)$ , and thus there is no need to consider it twice.



DEFINITION 5 (nondegenerate QP). *We say that the QP (10) is nondegenerate if, for each  $x \in \mathcal{X}_f$ , the rows of  $M_{(\mathcal{A}^*(x))}^U$  are linearly independent.*

LEMMA 1. *Suppose that the QP problem (10) is nondegenerate. Consider the value function  $J^*(x)$  in (26), and let  $\mathcal{P}_i, \mathcal{P}_j$  be two neighboring polyhedra corresponding to the set of active constraints  $\mathcal{A}_i$  and  $\mathcal{A}_j$ , respectively. Then*

$$(28) \quad Q_i - Q_j \preceq 0 \quad \text{or} \quad Q_i - Q_j \succeq 0 \quad \text{and} \quad Q_i \neq Q_j$$

and

$$(29) \quad Q_i - Q_j \preceq 0 \quad \text{iff} \quad \mathcal{A}_i \subset \mathcal{A}_j.$$

*Proof.* Let  $\mathcal{P}_i$  and  $\mathcal{P}_j$  be two neighboring polyhedra, and let  $\mathcal{A}_i$  and  $\mathcal{A}_j$  be the corresponding sets of active constraints at the optimum of QP (10). Let  $\mathcal{A}_i \subset \mathcal{A}_j$ . We want to prove that the difference between the quadratic terms of  $q_i(x)$  and  $q_j(x)$  is negative semidefinite, i.e.,  $Q_i - Q_j \preceq 0$  and  $Q_i \neq Q_j$ .

Without loss of generality, we can assume that  $\mathcal{A}_i = \emptyset$ . If this is not the case, then a simple substitution of variables based on the set of active constraints  $M_{(\mathcal{A}_i)}^U U = M_{(\mathcal{A}_i)} + M_{(\mathcal{A}_i)}^x x$  transforms problem (10) into a QP in a lower dimensional space.

With the substitution  $z = U + H^{-1}F'x$ , problem (10) can be translated into the following:

$$(30) \quad \begin{aligned} J_z^*(x) &= \min_z \frac{1}{2} z' H z \\ \text{s.t. } Gz &\leq W + Sx, \end{aligned}$$

where  $G := M^U$ ,  $W := M$ ,  $S := M^x + M^U H^{-1}F'$ , and  $J_z^*(x) = J^*(x) - \frac{1}{2}x'(Y - FH^{-1}F')x$ . For the unconstrained case we have  $z^* = 0$  and  $J_z^*(x) = 0$ . Consequently,

$$(31) \quad q_i(x) = \frac{1}{2}x'(Y - FH^{-1}F')x.$$

For the constrained case, as shown in [4], from the set of active constraints  $G_{(\mathcal{A}_j)}z = W_{(\mathcal{A}_j)} + S_{(\mathcal{A}_j)}x$  and the Karush–Kuhn–Tucker (KKT) conditions we obtain

$$(32) \quad z = H^{-1}G'_{(\mathcal{A}_j)}\Gamma^{-1}(W_{(\mathcal{A}_j)} + S_{(\mathcal{A}_j)}x),$$

$$(33) \quad \lambda_{(\mathcal{A}_j)} = -\Gamma^{-1}(W_{(\mathcal{A}_j)} + S_{(\mathcal{A}_j)}x),$$

where  $\Gamma = G_{(\mathcal{A}_j)}H^{-1}G'_{(\mathcal{A}_j)}$ ,  $\Gamma = \Gamma' \succ 0$  as the rows of  $G_{(\mathcal{A}_j)}$  are linearly independent, and  $\lambda_{(\mathcal{A}_j)}$  are the Lagrange multipliers of the active constraints  $\lambda_{(\mathcal{A}_j)} \geq 0$ . The corresponding value function is

$$(34) \quad \begin{aligned} q_j(x) &= \frac{1}{2}x'(Y - FH^{-1}F' + S'_{(\mathcal{A}_j)}\Gamma^{-1}S_{(\mathcal{A}_j)})x \\ &\quad + W'_{(\mathcal{A}_j)}\Gamma^{-1}S_{(\mathcal{A}_j)}x + \frac{1}{2}W'_{(\mathcal{A}_j)}\Gamma^{-1}W_{(\mathcal{A}_j)}. \end{aligned}$$

The difference of the quadratic terms of  $q_i(x)$  and  $q_j(x)$  gives

$$(35) \quad Q_i - Q_j = -\frac{1}{2}S'_{(\mathcal{A}_j)}\Gamma^{-1}S_{(\mathcal{A}_j)} \preceq 0.$$

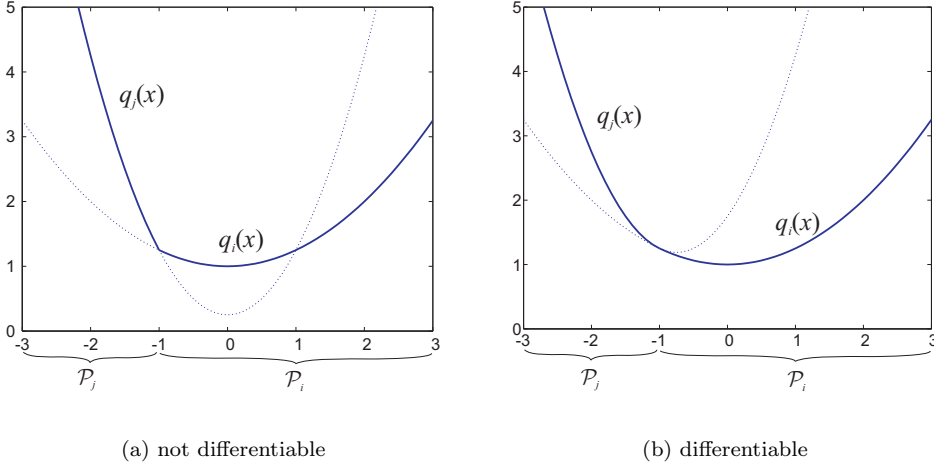


FIG. 4. Two piecewise quadratic convex functions.

What is left to prove is that  $Q_i \neq Q_j$ . We will prove it by showing that  $Q_i = Q_j$  if and only if  $\mathcal{P}_i = \mathcal{P}_j$ . For this purpose we recall [4] that the polyhedron  $\mathcal{P}_j$ , where the set of active constraints  $\mathcal{A}_j$  is constant, is defined as

$$(36) \quad \mathcal{P}_j = \left\{ x \mid GH^{-1}G'_{(\mathcal{A}_j)}\Gamma^{-1}(W_{(\mathcal{A}_j)} + S_{(\mathcal{A}_j)}x) \leq W + Sx, -\Gamma^{-1}(W_{(\mathcal{A}_j)} + S_{(\mathcal{A}_j)}x) \geq 0 \right\}.$$

From (35) we conclude that  $Q_i = Q_j$  if and only if  $S_{(\mathcal{A}_j)} = 0$ . The continuity of  $J_z^*(x)$  implies that  $q_i(x) - q_j(x) = 0$  on the common facet of  $\mathcal{P}_i$  and  $\mathcal{P}_j$ . Therefore, by comparing (31) and (34) we see that  $S_{(\mathcal{A}_j)} = 0$  implies  $W_{(\mathcal{A}_j)} = 0$ . Finally, for  $S_{(\mathcal{A}_j)} = 0$  and  $W_{(\mathcal{A}_j)} = 0$ , from (36) it follows that  $\mathcal{P}_j = \mathcal{P}_i := \{x \mid 0 \leq W + Sx\}$ .  $\square$

The following property of CPWQ functions was proved in [20].

**THEOREM 6.** Consider the value function  $J^*(x)$  in (26) satisfying (28) and its quadratic expression  $q_i(x)$  and  $q_j(x)$  on two neighboring polyhedra  $\mathcal{P}_i, \mathcal{P}_j$ . Then

$$(37) \quad q_i(x) = q_j(x) + (a'x - b)(\gamma a'x - \bar{b}),$$

where  $\gamma \in \mathbb{R}/\{0\}$ ,  $a \in \mathbb{R}^{n_x}$ ,  $b \in \mathbb{R}$ ,  $\bar{b} \in \mathbb{R}$ .

Equation (37) states that the functions  $q_i(x)$  and  $q_j(x)$  in two neighboring regions  $\mathcal{P}_i, \mathcal{P}_j$  of a CPWQ function satisfying (28) either intersect on two parallel hyperplanes,  $a'x - b$  and  $\gamma a'x - \bar{b}$  if  $\bar{b} \neq \gamma b$  (see Figure 4(a)), or are tangent in one hyperplane,  $a'x - b$  if  $\bar{b} = \gamma b$  (see Figure 4(b)). We will prove next that if the QP problem (10) is nondegenerate, then  $J^*(x)$  is a  $C^{(1)}$  function by showing that the case depicted in Figure 4(a) is not consistent with Lemma 1. In fact, Figure 4(a) depicts case  $Q^i - Q^j \preceq 0$ , that implies  $\mathcal{A}_i \subset \mathcal{A}_j$  by Lemma 1. However,  $q_j(0) < q_i(0)$  and from the definition of  $q_i$  and  $q_j$  this contradicts the fact that  $\mathcal{A}_i \subset \mathcal{A}_j$ .

**THEOREM 7.** If the QP problem (10) is nondegenerate, then the value function  $J^*(x)$  in (26) is  $C^{(1)}$ .

*Proof.* To show that  $J^*$  is  $C^{(1)}$  we need to prove that the cost functions  $q_i(x)$  and  $q_j(x)$  of any two neighboring regions  $\mathcal{P}_i$  and  $\mathcal{P}_j$  have the same gradient on the

common boundary. It is straightforward to see from (37) that this is indeed the case for  $\bar{b} = \gamma b$ . We will prove by contradiction that  $\bar{b} = \gamma b$ . Suppose there exist two neighboring polyhedra  $\mathcal{P}_i$  and  $\mathcal{P}_j$  such that  $\bar{b} \neq \gamma b$ . Without loss of generality, assume that (i)  $Q_i - Q_j \preceq 0$  and (ii)  $\mathcal{P}_i$  is in the halfspace  $a'x \leq b$  defined by the common boundary. Let  $\mathcal{F}_{ij}$  be the common facet between  $\mathcal{P}_i$  and  $\mathcal{P}_j$  and  $\text{relint}(\mathcal{F}_{ij})$  its relative interior.

From both (i) and (37), either  $\gamma < 0$  or  $\gamma = 0$  if  $Q_i - Q_j = 0$ . Take  $x_0 \in \text{relint}(\mathcal{F}_{ij})$ . For sufficiently small  $\varepsilon \geq 0$ , the point  $x := x_0 - a\varepsilon$  belongs to  $\mathcal{P}_i$ .

Let  $J^*(\varepsilon) := J^*(x_0 - a\varepsilon)$ ,  $q_i(\varepsilon) := q_i(x_0 - a\varepsilon)$ , and consider that

$$(38) \quad q_j(\varepsilon) = q_j(\varepsilon) + (a'a\varepsilon)(\gamma a'a\varepsilon + (\bar{b} - \gamma b)).$$

From convexity of  $J^*(\varepsilon)$ ,  $J^{*-}(\varepsilon) \leq J^{*+}(\varepsilon)$ , where  $J^{*-}(\varepsilon)$  and  $J^{*+}(\varepsilon)$  are the left and right derivatives of  $J^*(\varepsilon)$  with respect to  $\varepsilon$ . This implies  $q'_j(\varepsilon) \leq q'_i(\varepsilon)$ , where  $q'_j(\varepsilon)$  and  $q'_i(\varepsilon)$  are the derivatives of  $q_j(\varepsilon)$  and  $q_i(\varepsilon)$ , respectively. Condition  $q'_j(\varepsilon) \leq q'_i(\varepsilon)$  is true if and only if  $-(\bar{b} - \gamma b) \leq 2\gamma(a'a)\varepsilon$ , which implies  $-(\bar{b} - \gamma b) < 0$  since  $\gamma < 0$  and  $\varepsilon > 0$ .

From (38)  $q_j(\varepsilon) < q_i(\varepsilon) \forall \varepsilon \in (0, \frac{-(\bar{b}-\gamma b)}{\gamma a'a})$ .

Thus there exists  $x \in \mathcal{P}_i$  with  $q_j(x) < q_i(x)$ . This is a contradiction since from Theorem 5,  $\mathcal{A}_i \subset \mathcal{A}_j$ .  $\square$

Note that, in case of degeneracy, the value function  $J^*(x)$  in (26) may not be  $C^{(1)}$ ; counterexamples are given in [6].

In Theorem 7 we have proven that the value function is  $C^{(1)}$  for the nondegenerate QP (10). Now we want to show that the gradient of  $J^*(x)$  is a vector valued PWA descriptor function.

**THEOREM 8.** *Consider the value function  $J^*(x)$  in (26), and assume that the CFTOC problem (3) leads to a nondegenerate QP (10). Then the gradient  $m(x) := \nabla J^*(x)$  is a vector valued PWA descriptor function.*

*Proof.* From Theorem 7 we see that  $m(x)$  is a continuous vector valued PWA function, while from (26) we get

$$(39) \quad m(x) := \nabla J^*(x) = 2Q_i x + T_i \quad \text{if } x \in \mathcal{P}_i, \quad i = 1, \dots, N_P.$$

Since from Lemma 1 we know that  $Q_i \neq Q_j$  for all neighboring polyhedra, it follows that  $m(x)$  satisfies all conditions for a vector valued PWA descriptor function.  $\square$

Combining results of Theorems 8 and 4, it follows that by using Algorithm 5 we can construct a PWA descriptor function from the gradient of the value function  $J^*(x)$ .

*Remark 3.5.* Note that some CFTOC problems may lead to a degenerate QP (10). Identifying the classes of control problems which are guaranteed to be nondegenerate is currently an unsolved problem and the topic of ongoing research. An example of a control problem where the nondegeneracy fails is when, for some initial point, the optimal control is saturated on the whole time horizon (this gives as many active constraints as the dimension of  $U^*$ ) and the state hits the state constraint at least once (so there are more active constraints than the dimension of  $U^*$ ). However, degeneracy does not necessarily imply the loss of continuous differentiability of the value function, which thus might have no effect on our procedure. Next, we provide a simple example of a nondegenerate CFTOC which, by adding a terminal constraint, becomes degenerate and whose cost is not continuously differentiable.



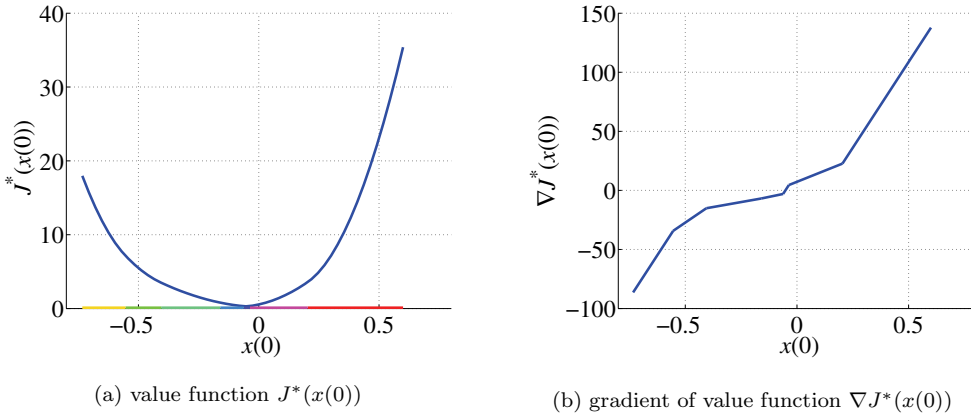


FIG. 5. Solution to the CFTOC problem (42).

**Example.** Consider the system

$$(40) \quad x_{k+1} = -1.5x_k + u_k,$$

where  $x_k \in \mathbb{R}$  and  $u_k \in \mathbb{R}$  are state and input at time  $k$ , respectively, subject to the following constraints  $\forall k \geq 0$ :

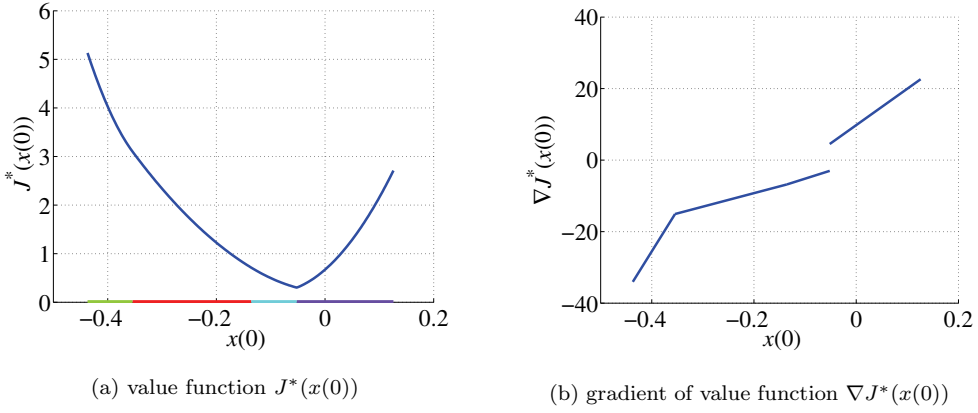
$$(41) \quad -1 \leq x_k \leq 1, \quad -0.5 \leq u_k \leq -0.1.$$

Consider the CFTOC problem (3) for system (40)–(41) with quadratic cost (4a), weighting matrices  $Q^x = 0.1$ ,  $Q^u = 10$ ,  $Q^{x_N} = 0$ , and horizon  $N = 3$ , i.e.,

$$(42) \quad \begin{aligned} J^*(x(0)) = \min_{u_0, u_1, u_2} \quad & \sum_{k=0}^2 0.1x_k^2 + 10u_k^2 \\ \text{s.t.} \quad & x_{k+1} = -1.5x_k + u_k, \quad k = 0, \dots, 2, \\ & -1 \leq x_k \leq 1, \quad k = 0, \dots, 2, \\ & -0.5 \leq u_k \leq -0.1, \quad k = 0, \dots, 2, \\ & x_0 = x(0). \end{aligned}$$

Then the corresponding QP (10) is nondegenerate and  $J^*(x(0))$  is continuously differentiable everywhere in its domain (see Figure 5). However, if the additional constraint  $x_3 = 0$  is added to the CFTOC problem (42), then the corresponding QP (10) becomes degenerate and  $J^*(x(0))$  is not continuously differentiable at  $x(0) = -0.0518$  (see Figure 6).

**3.2.2. Generating a PWA descriptor function from the optimizer.** A clear drawback of the procedure described in the previous subsection is the requirement that the QP (10) be nondegenerate. Luckily, there is another way to construct a vector valued PWA descriptor function  $m(x)$  that always works and does not require any additional checking. This second procedure emerges naturally if we look at the properties of the optimizer  $U^*(x)$  corresponding to the state feedback solution of the CFTOC problem (3). From Theorem 2, the optimizer  $U^*(x)$  is continuous in  $x$  and

FIG. 6. Solution to the CFTOC problem (42) with additional constraint  $x_3 = 0$ .

PWA:

$$(43) \quad U^*(x) = l_i(x) := \bar{F}_i x + \bar{G}_i \quad \text{if } x \in \mathcal{P}_i, \quad i = 1, \dots, N_{\mathcal{P}},$$

where  $\bar{F}_i \in \mathbb{R}^{s \times n_x}$ ,  $\bar{G}_i \in \mathbb{R}^s$ , and  $\mathcal{P}_i$ ,  $i = 1, \dots, N_{\mathcal{P}}$ , are the full dimensional critical regions of the optimization problem (10); cf. [4].

All we need to show now is the following lemma.

LEMMA 2. Consider the CFTOC problem (3) and corresponding QP (10). Let  $\mathcal{P}_i$ ,  $\mathcal{P}_j$  be two (full dimensional) neighboring polyhedra of the state feedback solution (43). Then  $\bar{F}_i \neq \bar{F}_j$ .

*Proof.* We prove Lemma 2 by contradiction. Suppose that the optimizer is the same for both polyhedra, i.e.,  $[\bar{F}_i \ \bar{G}_i] = [\bar{F}_j \ \bar{G}_j]$ . Let  $\mathcal{A}_i$  and  $\mathcal{A}_j$  be the sets of active constraints for  $\mathcal{P}_i$  and  $\mathcal{P}_j$ . From (27) we see that for an active constraint in  $\mathcal{P}_i$ ,  $v \in \mathcal{A}_i$ , the following holds:

$$(44) \quad M_{(v)}^U(\bar{F}_i x + \bar{G}_i) = M_{(v)} + M_{(v)}^x x \quad \forall x \in \mathcal{P}_i, \forall v \in \mathcal{A}_i,$$

because (44) must hold  $\forall x \in \mathcal{P}_i$ , where  $\mathcal{P}_i$  is a full dimensional polyhedron in  $\mathbb{R}^{n_x}$ . Since  $M_{(v)}^U \bar{F}_i - M_{(v)}^x \in \mathbb{R}^{n_x}$  the above statement is satisfied if and only if

$$(45) \quad M_{(v)}^U \bar{F}_i - M_{(v)}^x = \mathbf{0}', \quad M_{(v)}^U \bar{G}_i - M_{(v)} = 0 \quad \forall v \in \mathcal{A}_i.$$

By assumption  $[\bar{F}_i \ \bar{G}_i] = [\bar{F}_j \ \bar{G}_j]$ , hence the expression (45) implies that

$$(46) \quad M_{(v)}^U(\bar{F}_j x + \bar{G}_j) = M_{(v)} + M_{(v)}^x x \quad \forall x, \forall v \in \mathcal{A}_i.$$

Clearly, (46) holds  $\forall x \in \mathcal{P}_j$ , and we conclude that all active constraints of  $\mathcal{P}_i$  are also active in  $\mathcal{P}_j$ . The same reasoning holds for  $\mathcal{P}_j$  and  $\forall v \in \mathcal{A}_j$ , thus implying that  $\mathcal{A}_i = \mathcal{A}_j$ . However, the same sets of active constraints and the same optimal control law in QP (10) can generate only one region, thus implying that  $\mathcal{P}_i = \mathcal{P}_j$ , which contradicts the assumption of Lemma 2. Therefore we have  $[\bar{F}_i \ \bar{G}_i] \neq [\bar{F}_j \ \bar{G}_j]$ . Finally, observing that for the two neighboring polyhedra, due to the continuity of  $U^*(x)$ ,  $\bar{F}_i = \bar{F}_j$  implies  $\bar{G}_i = \bar{G}_j$ , for which we prove that  $\bar{F}_i \neq \bar{F}_j$ .  $\square$

From Lemma 2 and Theorem 4 it follows that an appropriate PWA descriptor function  $f(x)$  can be calculated from the optimizer  $U^*(x)$  by using Algorithm 5.

*Remark 3.6.* Note that even if we are implementing RHC strategy, the construction of the PWA descriptor function is based on the full optimization vector  $U^*(x)$  and the corresponding matrices  $\bar{F}_i$  and  $\bar{G}_i$ .

*Remark 3.7.* In some cases the use of the optimal control profile  $U^*(x)$  for the construction of a descriptor function  $f(x)$  can be extremely simple. If there is a row  $r$ ,  $r \leq n_u$  ( $n_u$  is the dimension of  $u$ ), for which  $(\bar{F}_i)_{(r)} \neq (\bar{F}_j)_{(r)} \forall i = 1, \dots, N_P, \forall j \in \mathcal{C}_i$ , it is enough to set  $A_i' = (\bar{F}_i)_{(r)}$  and  $B^i = (\bar{G}_i)_{(r)}$ , where  $(\bar{F}_i)_{(r)}$  and  $(\bar{G}_i)_{(r)}$  denote the  $r$ th row of the matrices  $\bar{F}_i$  and  $\bar{G}_i$ , respectively. In this way we *avoid* the storage of the descriptor function, since it is equal to one component of the control law, which is stored anyway.

*Remark 3.8.* A natural question is whether this approach can be readily extended to the control of PWA systems. In general, the answer is no. However, we point out that whenever a PWA descriptor function can be found one can use Algorithm 4. For instance, the optimal control of a PWA system with piecewise linear cost results in a PWA control over polyhedral partition of the feasible state space (cf. [2]). If this control law satisfies conditions of the vector valued PWA descriptor function (see Definition 4), then we can use Algorithm 4. Note that the value function (12) in the linear cost case is also a descriptor function.

All described algorithms can be easily implemented and combined with available tools for deriving optimal controllers for linear systems (e.g., multiparametric toolbox [13] under MATLAB).

**4. Example.** As an example, we compare the performance of Algorithms 2, 3, and 4 on a CFTOC problem for the discrete-time system

$$(47) \quad \begin{cases} x(t+1) = \begin{bmatrix} 4 & -1.5 & 0.5 & -0.25 \\ 4 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0.5 \\ 0 \\ 0 \\ 0 \end{bmatrix} u(t), \\ y(t) = [0.08333 \quad 0.2292 \quad 0.1146 \quad 0.02083] x(t) \end{cases}$$

resulting from the linear system  $y = \frac{1}{s^4}u$ , sampled at  $T_s = 1$ , and subject to the input and output constraints

$$(48) \quad -1 \leq u(t) \leq 1, \quad -10 \leq y(t) \leq 10.$$

**4.1. CFTOC, linear cost case.** To regulate (47)–(48), we design a receding horizon controller based on the optimization problem (3) with the piecewise linear cost function (4a),  $\ell = \infty$ ,  $N = 2$ ,  $Q = \text{diag}\{5, 10, 10, 10\}$ ,  $R = 0.8$ ,  $P = 0$ . The PWA solution of the mp-LP problem was computed in 240 s on a Pentium III 900 MHz machine running MATLAB 6.0. The corresponding polyhedral partition of the state-space consists of  $N_P = 136$  regions with  $N_H = 1138$  halfspaces. In Table 3 we report the comparison between the complexity of Algorithms 2 and 3 for this example.

The average on-line RHC computation for a set of 1000 random points in the state space is 2259 flops (Algorithm 2) and 1088 flops (Algorithm 3).

**4.2. CFTOC, quadratic cost case.** To regulate (47)–(48), we design a receding horizon controller based on the optimization problem (3) with the quadratic cost

TABLE 3  
Complexity comparison of Algorithms 2 and 3 for the example in section 4.1.

	Algorithm 2	Algorithm 3
Storage demand (real numbers)	5690	680
Number of flops (worst case)	9104	1088

TABLE 4  
Complexity comparison of Algorithms 2 and 4 for the example in section 4.2.

	Algorithm 2	Algorithm 4
Storage demand (real numbers)	9740	1065
Number of flops (worst case)	15584	3439

function (4b),  $N = 7$ ,  $Q = I$ ,  $R = 0.01$ ,  $P = 0$ . The PWA solution of the mp-QP problem was computed in 560 s on a Pentium III 900 MHz machine running MATLAB 6.0. The corresponding polyhedral partition of the state space consists of  $N_{\mathcal{P}} = 213$  regions with  $N_{\mathcal{H}} = 1948$  halfspaces. For this example the choice of  $w = [1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1]'$  is satisfactory to obtain a descriptor function from the optimizer. In Table 4 we report the comparison between the complexity of Algorithms 2 and 4 for this example.

The average on-line RHC computation for a set of 1000 random points in the state space is 2114 flops (Algorithm 2) and 175 flops (Algorithm 4).

**5. Conclusion.** By exploiting properties of the value function and the optimal solution to the CFTOC problem, we presented two algorithms that significantly improved the efficiency of the on-line calculation of the control action (LP-based and QP-based) in terms of storage demand and computational complexity. The following improvements are then achieved:

1. There is no need to store the polyhedral partition of the state space.
2. In the worst case, the optimal control law is computed after the evaluation of one linear function per polyhedron.

**Acknowledgments.** The authors would like to thank the anonymous reviewers whose careful scrutiny and many useful suggestions considerably improved the quality of this manuscript.

## REFERENCES

- [1] B. BANK, J. GUDDAT, D. KLATTE, B. KUMMER, AND K. TAMMER, *Non-Linear Parametric Optimization*, Akademie-Verlag, Berlin, 1982.
- [2] M. BAOTIĆ, F. J. CHRISTOPHERSEN, AND M. MORARI, *Constrained optimal control of hybrid systems with a linear performance index*, IEEE Trans. Automat. Control, 51 (2006), pp. 1903–1919.
- [3] A. BEMPORAD, F. BORRELLI, AND M. MORARI, *Model predictive control based on linear programming—The explicit solution*, IEEE Trans. Automat. Control, 47 (2002), pp. 1974–1985.
- [4] A. BEMPORAD, M. MORARI, V. DUA, AND E. N. PISTIKOPOULOS, *The explicit linear quadratic regulator for constrained systems*, Automatica J. IFAC, 38 (2002), pp. 3–20.
- [5] C. BERGE, *Topological Spaces*, Dover Publications, Inc., Mineola, NY, 1997.
- [6] A. B. BERKELAAR, K. ROOS, AND T. TERLAKY, *The optimal set and optimal partition approach to linear and quadratic programming*, in Advances in Sensitivity Analysis and Parametric Programming, T. Gal and H. J. Greenberg, eds., Internat. Ser. Oper. Res. Management Sci. 6, Kluwer Academic Publishers, Boston, 1997.
- [7] F. BORRELLI, A. BEMPORAD, AND M. MORARI, *Geometric algorithm for multiparametric linear programming*, J. Optim. Theory Appl., 118 (2003), pp. 515–540.

- [8] S. BOYD AND L. VANDENBERGHE, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
- [9] D. CHMIELEWSKI AND V. MANOUSIOUTHAKIS, *On constrained infinite-time linear quadratic optimal control*, Systems Control Lett., 29 (1996), pp. 121–129.
- [10] H. J. FERREAU, H. G. BOCK, AND M. DIEHL, *An online active set strategy to overcome the limitations of explicit MPC*, Internat. J. Robust Nonlinear Control, 18 (2008), pp. 816–830.
- [11] T. GAL, *Postoptimal Analyses, Parametric Programming, and Related Topics*, 2nd ed., Walter de Gruyter, Berlin, 1995.
- [12] C. JONES, P. GRIEDER, AND S. RAKOVIĆ, *A logarithmic-time solution to the point location problem for parametric linear programming*, Automatica J. IFAC, 42 (2006), pp. 2215–2218.
- [13] M. KVASNICA, P. GRIEDER, M. BAOTIĆ, AND M. MORARI, *Multi-Parametric Toolbox (MPT)*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2993, Springer-Verlag, Philadelphia, 2004, pp. 448–462.
- [14] J. M. MACIEJOWSKI, *Predictive Control with Constraints*, Prentice-Hall, Harlow, UK, 2002.
- [15] D. Q. MAYNE, J. B. RAWLINGS, C. V. RAO, AND P. O. M. SCOKAERT, *Constrained model predictive control: Stability and optimality*, Automatica J. IFAC, 36 (2000), pp. 789–814.
- [16] J. A. MENDEZ, B. KOUVARITAKIS, AND J. A. ROSSITER, *State-space approach to interpolation in MPC*, Internat. J. Robust Nonlinear Control, 10 (2000), pp. 27–38.
- [17] M. SCHECHTER, *Polyhedral functions and multiparametric linear programming*, J. Optim. Theory Appl., 53 (1987), pp. 269–280.
- [18] P. O. M. SCOKAERT AND J. B. RAWLINGS, *Constrained linear quadratic regulation*, IEEE Trans. Automat. Control, 43 (1998), pp. 1163–1169.
- [19] J. SPJØTVOLD, E. C. KERRIGAN, C. N. JONES, P. TØNDEL, AND T. A. JOHANSEN, *On the facet-to-facet property of solutions to convex parametric quadratic programs*, Automatica J. IFAC, 42 (2006), pp. 2209–2214.
- [20] J. SUN, *On the structure of convex piecewise quadratic functions*, J. Optim. Theory Appl., 72 (1992), pp. 499–510.
- [21] M. SZNAIER AND M. J. DAMBORG, *Suboptimal control of linear systems with state and control inequality constraints*, in Proceedings of the 26th IEEE Conference on Decision & Control, Los Angeles, CA, 1987, pp. 761–762.
- [22] P. TØNDEL, T. A. JOHANSEN, AND A. BEMPORAD, *An algorithm for multiparametric quadratic programming and explicit MPC solutions*, Automatica J. IFAC, 39 (2003), pp. 489–497.
- [23] P. TØNDEL, T. A. JOHANSEN, AND A. BEMPORAD, *Evaluation of piecewise affine control via binary search tree*, Automatica J. IFAC, 39 (2003), pp. 945–950.

## PARTIALLY OBSERVED INVENTORY SYSTEMS: THE CASE OF RAIN CHECKS\*

ALAIN BENSOUSSAN<sup>†</sup>, METIN ÇAKANYILDIRIM<sup>‡</sup>, J. ADOLFO MINJÁREZ-SOSA<sup>§</sup>,  
SURESH P. SETHI<sup>¶</sup>, AND RUIXIA SHI<sup>‡</sup>

**Abstract.** In many inventory control contexts, inventory levels are only partially (i.e., not fully) observed. This may be due to nonobservation of demand, spoilage, misplacement, or theft of inventory. We study a discrete-time periodic-review inventory system where the unmet demand is backordered. When the inventory level is nonnegative, the inventory manager does not know the exact inventory level. Otherwise, inventory shortages occur, and the inventory manager issues rain checks to customers. The shortages are fully observed via the rain checks. The inventory manager determines the order quantity based on the partial information on the inventory level. The objective is to minimize the expected total discounted cost over an infinite horizon. The dynamic programming formulation of this problem has an infinite dimensional state space. We use the methodology of the unnormalized probability to establish the existence of an optimal feedback policy when the periodic cost has linear growth. Moreover, uniqueness and continuity of the solution to dynamic programming equations are proved when the discount factor is sufficiently small.

**Key words.** stochastic inventory problem, partial observations, the Zakai equation, rain check

**AMS subject classifications.** 90B05, 93E20, 93C41, 90C39

**DOI.** 10.1137/070688663

**1. Introduction.** Most inventory models assume that the inventory levels at any given time are fully observed. Under this assumption, some of the most celebrated results, such as the optimality of the base stock policy [1], have been obtained. However, in reality, the inventory levels are often only partially observed. In such cases, most of the well-known inventory policies are not even admissible, let alone optimal.

There are a number of reasons for partial observability of inventory levels. We list only some, as they have been discussed in detail in Bensoussan, Çakanyildirim, and Sethi [4] (referred to as BCS hereafter) and references therein. Demand may be incorrectly observed or observed with some delay. Inventory may be misplaced or stolen. Some products such as groceries and drugs deteriorate over time and are no longer fit for sale after a period of time. In addition, the saleable quantities received may be different from those ordered because of uncertain yields.

Even though partial observations in inventory systems are very common, there has not been much research activity in this area. The main reason may be the mathematical difficulty. While a finite dimensional space suffices to accommodate the system state in the full observation case, an infinite dimensional state space is required in the partial observation setting. More specifically, the inventory level at

---

\*Received by the editors April 18, 2007; accepted for publication (in revised form) April 20, 2008; published electronically September 17, 2008.

<http://www.siam.org/journals/sicon/47-5/68866.html>

<sup>†</sup>International Center for Decision and Risk Analysis, School of Management, SM 30, University of Texas at Dallas, P.O. Box 830688, Richardson, TX 75083-0688 (alain.bensoussan@utdallas.edu).

<sup>‡</sup>School of Management, SM 30, University of Texas at Dallas, P.O. Box 830688, Richardson, TX 75083-0688 (metin@utdallas.edu, ruixia.shi@utdallas.edu).

<sup>§</sup>Departamento de Matemáticas, Universidad de Sonora, Rosales s/n, Centro, 83000 Hermosillo, Sonora, Mexico (aminjare@gauss.mat.uson.mx).

<sup>¶</sup>Center for Intelligent Supply Networks, School of Management, SM 30, University of Texas at Dallas, P.O. Box 830688, Richardson, TX 75083-0688 (sethi@utdallas.edu).

a given time is no longer a system state in  $\mathbb{R}^n$ ; it must now be represented by its conditional probability given the partial observations available at that time. Thus, the analysis takes place in the space of probability distributions.

When there is no physical inventory, i.e., the inventory is zero or negative, then none of the following would happen: transaction errors, misplaced inventories, spoilage, or theft. Most companies pay utmost attention to an item when its inventory reaches zero. In these companies, employees walk around the shelves to see if the item is stocked out. This process is implemented at the office supply store Staples and is called “zero-balance walk” in [9] and [12]. BCS study this situation, in which the positive inventory is represented by a conditional distribution, no backordering is allowed, and an empty shelf is noticed when the inventory reaches zero. BCS were the first to introduce a methodology to handle the difficulty of dealing with partially observed discrete-time inventory systems.

This paper extends BCS to the case when backordering is allowed, i.e., when sales are not lost. It is very common in reality that a store issues rain checks to a customer when there is a stockout. A rain check is an assurance to a customer that a sold out item can be purchased later at the same price. Usually, the inventory manager (IM) monitors the issuing of the rain checks. Thus, in the inventory systems with rain checks, the negative inventory level is fully observed by the IM. It is the purpose of this paper to formulate and analyze the partially observed inventory system with rain checks. However, compared to the work of BCS, where the inventory level is observed when it is zero, the full observation of negative inventory in the present paper makes the analysis much more difficult. Indeed, in the model of BCS, the optimality equation is represented by a pair formed by a constant and a function of the conditional distribution of the inventory level. The constant corresponds to the case of zero inventory (observable term), while the function represents the optimality equation with positive inventory (nonobservable term). However, in the present model, instead of a constant, the observable term via rain checks is represented by a function of the backordered amount. Therefore, the present study must consider suitable functional spaces to define appropriate dynamic programming operators, which necessitates a more complicated mathematical analysis.

It is worth emphasizing that even though the study of the rain check model is more involved than that of the zero-balance walk model, we are able to develop an alternative approach that allows us to improve the results in BCS. They show the existence of a solution to the dynamic programming equations for the problem by obtaining a decreasing sequence of solutions from a value iteration scheme that converge to the value function. Furthermore, the existence of an optimal ordering policy in BCS is obtained under the condition of bounded ordering quantities or with a sufficiently small discount factor. We, on the other hand, do not require these conditions to get the existence of an optimal feedback ordering policy. This is accomplished by first proving the convergence of an increasing sequence of solutions, obtained from a value iteration algorithm, to the value function, and then using a selection theorem to establish the existence of an optimal feedback policy. Furthermore, for a sufficiently small discount factor, and using a decreasing approximation scheme as in BCS, we show in addition that the value function is continuous and is the unique solution of the optimality equation. Moreover, the value iteration procedure developed in this paper is a valid approximation scheme for any discount factor, whereas the BCS scheme is restricted to a sufficiently small discount factor.

The mathematical treatment for our problem is similar to those applying to the standard partially observed stochastic control problems. These consist of introducing

a filtering process for the partially observed variable, and then the partially observed problem is transformed into an equivalent fully observed problem where the filter becomes the new system state. In our case, the filtering process is defined by the conditional density of the inventory level given the observed history. The filters are obtained recursively in Theorem 1. Besides the difficulties associated with the infinite dimensional state space of the new problem, the equation defining the filtering process is highly nonlinear, which leads to a nontrivial problem. However, by introducing what is known as the unnormalized probability, we transform this equation to a linear one. Therefore, our approach finally consists of using an unnormalized filtering process, which allows us to easily prove our main results.

The rest of this paper is organized as follows. In the next section we describe our model and obtain the evolution equation for the inventory distribution, which is linearized by introducing the unnormalized probability. In section 3, we formulate the problem and provide the dynamic programming equations for both normalized and unnormalized probabilities. We present our results in section 4 and prove them in section 5.

**2. Model development.** We study an infinite-horizon, single-item, discrete-time periodic-review inventory system in which the inventory level is partially observed due to unobservable demand. We assume that the unmet demand is backordered. Because of the rain checks, the IM knows exactly the backordered quantity. But when the inventory is nonnegative, he or she knows only the probability distribution of the inventory level. Thus, the IM only partially observes the inventory level.

We divide the timeline into intervals of equal length. Each interval is called a period. In practice, a period can be one week, one month, etc. The sequence of events in any given period  $t$  is as follows: at the beginning of the period (before the demand occurs), the IM observes if the inventory level is negative or nonnegative. If the inventory level is nonnegative, he or she has the probability distribution of the inventory level based on prior observations. Otherwise, the IM knows exactly the backordered quantity. Then, the IM determines how much to order, and the order is delivered immediately. Next the demand occurs, and the IM's order can be used to meet the demand of period  $t$ . But the demand is not observed by the IM unless the inventory level falls below zero. In each period, the IM incurs inventory related costs.

We let  $I_t$  denote the inventory level at the beginning of period  $t$ ,  $q_t$  denote the order quantity determined by the IM at that time taking values in  $\mathbb{R}^+ := [0, \infty)$ , and  $D_t$  denote the random demand occurring in period  $t$ . We assume that  $D_t$ ,  $t = 1, 2, 3, \dots$ , are nonnegative, independently and identically distributed random variables, and let the generic demand be denoted by  $D \geq 0$  with the density and distribution functions  $f$  and  $F$ , respectively. We assume that  $E(D) < \infty$ . Unsatisfied demand is fully backordered, so that the evolution of inventory is given by

$$(1) \quad I_{t+1} = I_t + q_t - D_t \quad \text{for } t \geq 1.$$

$I_t$  can be either negative or nonnegative. If  $I_t$  is negative, the demand in period  $t - 1$  is not fully met, and the negative part of  $I_t$  represents the backordered amount. Let  $z_t$  denote the backordered amount in period  $t$ ; then  $z_t = (I_t)^-$ . If  $I_t$  is nonnegative, obviously it is the ending inventory carried over from period  $t - 1$ . But the IM cannot tell exactly how much it is. What he or she can do, however, is to evolve the conditional distribution of the inventory level.



In our model, for period  $t \geq 1$ , the IM can only observe

$$(2) \quad z_t = (I_t)^- \quad \text{for } t \geq 1,$$

where  $(x)^- = \max\{0, -x\}$ . If  $z_t > 0$ , it is the backordered quantity. If  $z_t = 0$ , the inventory level at the beginning of period  $t$  is nonnegative, and a distribution of the inventory level can be derived based on prior observations. The IM determines the order quantity  $q_t$  for each period  $t$ . The order quantity  $q_t$  is adapted to the sigma field  $\mathcal{Z}_t := \sigma(\{z_j : 1 \leq j \leq t\})$ . Thus,  $\mathcal{Z}_t$  is the history available to the IM in period  $t$ . By an ordering policy (or simply a policy), we mean a sequence  $\tilde{q} = \{q_t\}$  of  $\mathbb{R}^+$ -valued random variables such that  $q_t$  is  $\mathcal{Z}_t$ -measurable for each  $t \geq 1$ . Let  $\Gamma$  be the set of such policies.

When the demand is met entirely in period  $t - 1$ , inventory holding costs are incurred on the remaining inventory carried from period  $t - 1$  to period  $t$ .<sup>1</sup> Otherwise, there are backorder costs. Then, given the cost function  $c(I_t, q_t)$  depending only on the inventory level  $I_t$  and the order size  $q_t$  in period  $t$ ,  $t = 1, 2, 3, \dots$ , and given the policy  $\tilde{q} \in \Gamma$ , the total expected discounted cost can be written as

$$J(\zeta, \pi, \tilde{q}) := \mathbb{E} \sum_{t=1}^{\infty} \alpha^{t-1} c(I_t, q_t),$$

where  $0 < \alpha < 1$  is the discount factor. The pair  $(\zeta, \pi(\cdot))$  is the inventory state at the beginning of period 1. When  $\zeta > 0$ ,  $\zeta$  is the backordered quantity at the beginning of period 1. If  $\zeta = 0$ , the inventory level  $I_1$  is nonnegative and  $\pi \in \Pi$  is the probability density of  $I_1$ , where  $\Pi$  is the set of density functions such that  $\int x\pi(x)dx < \infty$ . Observe that  $f \in \Pi$ .

We want to find a policy  $\tilde{q} \in \Gamma$  that minimizes  $J(\zeta, \pi, \tilde{q})$ . When  $\zeta > 0$ ,  $\pi$  is of no consequence and  $J(\zeta, \pi, \tilde{q})$  does not depend on  $\pi$ , which will be noted explicitly below.

**2.1. Evolution of state probabilities.** At the beginning of the first period, a prior probability density  $\pi \in \Pi$  of nonnegative inventory is given. From the second period on, the IM needs to derive the probability distribution function for nonnegative inventory to use in his or her decision process. In what follows, we show how these distribution functions evolve over time.

It is convenient to introduce the indicator random variables

$$(3) \quad \mathbb{I}_{z_t=0} = \mathbb{I}_{I_t^- = 0} = \mathbb{I}_{I_t \geq 0} \quad \text{for } t \geq 1.$$

When there is no backorder, i.e.,  $z_t = I_t^- = 0$  or  $I_t \geq 0$ ,  $\mathbb{I}_{z_t=0} = 1$ . Otherwise,  $\mathbb{I}_{z_t=0} = 0$ . The indicator random variable  $\mathbb{I}_{z_t=0}$  is a discrete-time Markov chain with the state space  $\{0, 1\}$ : 1 means there is no backorder, and 0 means there is.

Let  $\pi_t(\cdot)$  be the conditional density of  $I_t$  given  $\mathcal{Z}_{t-1}$  and  $I_t \geq 0$ . That is,

$$P(I_t \leq x | \mathcal{Z}_{t-1}, I_t \geq 0) = \int_0^x \pi_t(y) dy.$$

<sup>1</sup>See Remark 2.3 in [8] for inventory holding cost accounting based on ending or beginning inventory levels.

For any bounded real test function  $\varphi(\cdot)$ , the conditional Bayes theorem gives

$$(4) \quad \int_0^\infty \varphi(x) \pi_t(x) dx = E[\varphi(I_t) | \mathcal{Z}_{t-1}, I_t \geq 0] = \frac{E[\varphi(I_t) \mathbb{I}_{I_t \geq 0} | \mathcal{Z}_{t-1}]}{E[\mathbb{I}_{I_t \geq 0} | \mathcal{Z}_{t-1}]} \\ = \frac{E[\varphi(I_t) \mathbb{I}_{I_t \geq 0} | \mathcal{Z}_{t-1}]}{P(I_t \geq 0 | \mathcal{Z}_{t-1})}.$$

In order to obtain a recursive expression for  $\pi_t$  in terms of  $\pi_{t-1}$ , we express  $E(\varphi(I_t) | \mathcal{Z}_t)$  in terms of the conditional density  $\pi_t$  in the following lemma.

LEMMA 1.

$$(5) \quad E(\varphi(I_t) | \mathcal{Z}_t) = \mathbb{I}_{z_t > 0} \varphi(-z_t) + \mathbb{I}_{z_t = 0} \int_0^\infty \varphi(\eta) \pi_t(\eta) d\eta.$$

From this lemma we can derive the density function  $\pi_t$  as stated in the following theorem.

THEOREM 1. *For  $t \geq 2$ , the conditional density  $\pi_t$  can be expressed recursively as follows:*

$$(6) \quad \pi_t(x) = \mathbb{I}_{z_{t-1} > 0} \left\{ \frac{f(-z_{t-1} + q_{t-1} - x) \mathbb{I}_{-z_{t-1} + q_{t-1} - x \geq 0}}{F(q_{t-1} - z_{t-1})} \right\} \\ + \mathbb{I}_{z_{t-1} = 0} \left\{ \frac{\int_{(x-q_{t-1})^+}^\infty f(y + q_{t-1} - x) \pi_{t-1}(y) dy}{\int_0^\infty F(y + q_{t-1}) \pi_{t-1}(y) dy} \right\}.$$

For  $t = 1$ ,  $\pi_1(x) = \pi(x)$ .

Thus,  $\pi_t$  evolves according to a highly nonlinear equation which corresponds to the Kushner equation [11] in our inventory context.

We can use the following method to linearize (6). We define the sequence of functions  $\{p_t\}$  by the recursive linear equation

$$(7) \quad p_t(x) = \mathbb{I}_{z_{t-1} > 0} f(-z_{t-1} + q_{t-1} - x) \mathbb{I}_{-z_{t-1} + q_{t-1} - x \geq 0} \\ + \mathbb{I}_{z_{t-1} = 0} \int_{(x-q_{t-1})^+}^\infty f(y + q_{t-1} - x) p_{t-1}(y) dy, \\ p_1(x) = \pi(x),$$

which corresponds to the Zakai equation obtained in the context of systems with diffusions in [2] and [13]. Also,

$$\lambda_t = \int_0^\infty p_t(x) dx.$$

Then, we have  $\lambda_1 = 1$  and, for  $t \geq 2$  from (7),

$$(8) \quad \lambda_t = \mathbb{I}_{z_{t-1} > 0} F(-z_{t-1} + q_{t-1}) + \mathbb{I}_{z_{t-1} = 0} \int_0^\infty F(q_{t-1} + y) p_{t-1}(y) dy.$$

Moreover,

$$(9) \quad p_t(x) = \lambda_t \pi_t(x).$$

Clearly, (9) holds for  $t = 1$ . Assuming (9) to hold for any  $t$ , we proceed to  $t + 1$  by multiplying (6) side-by-side by (8) to obtain

$$\begin{aligned} \lambda_{t+1}\pi_{t+1}(x) &= \mathbb{I}_{z_t > 0} f(-z_t + q_t - x) \mathbb{I}_{-z_t + q_t - x \geq 0} \\ &+ \mathbb{I}_{z_t = 0} \frac{\int_{(x-q_t)^+}^{\infty} f(y + q_t - x) \pi_t(y) dy \int_0^{\infty} F(q_t + y) p_t(y) dy}{\int_0^{\infty} F(y + q_t) \pi_t(y) dy}. \end{aligned}$$

By multiplying the numerator and the denominator of the second term on the right-hand side by  $\lambda_t$ , we establish (9) for  $t + 1$ .

On account of the weighting factor  $\lambda_t$  in (9),  $p_t(x)$  can be viewed as an unnormalized probability; see [10] or [13]. We can easily get the normalized probability  $\pi_t(x)$  by

$$\pi_t(x) = \frac{p_t(x)}{\int_0^{\infty} p_t(x) dx}.$$

To write (6) and (7) in the operator form, we define the spaces

$$\mathcal{H} := \left\{ p \in L^1(\mathbb{R}^+) : \int_0^{\infty} x |p(x)| dx < \infty \right\} \quad \text{and} \quad \mathcal{H}^+ := \{ p \in \mathcal{H} | p(x) \geq 0, x \in \mathbb{R}^+ \},$$

where  $L^1(\mathbb{R}^+)$  is the space of integrable functions whose domain is the set of nonnegative real numbers. Note that  $\Pi \subseteq \mathcal{H}^+$ . Observe that  $\mathcal{H}^+$  is not a subspace of  $\mathcal{H}$ , for it does not include  $-p$  for some  $p$ . In the remainder, we identify  $\mathcal{H}^+$  as the set of all unnormalized probabilities with the norm

$$\|p\| := \int_0^{\infty} |p(x)| dx + \int_0^{\infty} x |p(x)| dx.$$

Since  $\Pi \subseteq \mathcal{H}^+$ , the norm applies to the normalized probability  $\pi$  to say that the inventory level must have a finite mean. For a sequence  $\{p_n\}$  in  $\mathcal{H}^+$  and  $p \in \mathcal{H}^+$ , “ $p_n \rightarrow p$ ” means  $\|p_n - p\| \rightarrow 0$  as  $n \rightarrow \infty$ , which is equivalent to

$$(10) \quad \int_0^{\infty} |p_n(x) - p(x)| dx \rightarrow 0 \quad \text{and} \quad \int_0^{\infty} x |p_n(x) - p(x)| dx \rightarrow 0.$$

Let  $\mathbb{L}$  be the space of functions  $\phi$  with linear growth, i.e.,

$$\mathbb{L} := \left\{ \phi : \sup_{x \geq 0} \frac{|\phi(x)|}{1+x} < \infty \right\}$$

with the norm

$$\|\phi\|_{\mathbb{L}} := \sup_{x \geq 0} \frac{|\phi(x)|}{1+x}.$$

Furthermore, we define the product

$$\langle p, \phi \rangle := \int_0^{\infty} p(x) \phi(x) dx \quad \text{for } p \in \mathcal{H}, \phi \in \mathbb{L}.$$

Similarly, we define the space  $\mathbb{L}^+ := \{\phi \in \mathbb{L} : \phi(x) \geq 0, x \in \mathbb{R}^+\}$ . We assume that, for every fixed  $q_t$ , the one-period cost  $c(I_t, q_t)$  is in  $\mathbb{L}^+$ .

For any order quantity  $q$  and the inventory level  $y$  before the receipt of the order, we let  $w = y + q$  be the inventory level after the receipt. We also let  $\mathcal{L}(\mathcal{H}, \mathcal{H})$  denote the space of bounded linear maps from  $\mathcal{H}$  to  $\mathcal{H}$ . Then, for  $y \geq 0$ , we can define the linear operator  $\varrho(q) \in \mathcal{L}(\mathcal{H}, \mathcal{H})$  as

$$\varrho(q)p(x) := \int_{(x-q)^+}^{\infty} f(y+q-x)p(y)dy.$$

For  $y < 0$ , we define

$$\varrho_0(w)(x) := \begin{cases} f(w-x)\mathbb{I}_{x \leq w} & \text{if } w \geq 0, \\ 0 & \text{if } w < 0. \end{cases}$$

After defining the operators above, we can define the corresponding nonlinear operator  $\theta(q, \cdot)$  and the function  $\theta_0(w)$ , respectively, as

$$\begin{aligned} \theta(q, p) &:= \frac{\varrho(q)p}{\int_0^{\infty} F(y+q)p(y)dy}, \\ \theta_0(w) &:= \frac{\varrho_0(w)}{F(w)}. \end{aligned}$$

From the definition of  $\theta(q, p)$ , we can see that the operator  $\theta$  is homogenous of degree 0 in  $p$  and is well defined if  $\langle \varrho(q)p, 1 \rangle > 0$ ; i.e.,

$$(11) \quad \langle \varrho(q)p, 1 \rangle = \int_0^{\infty} p(y)F(y+q)dy \neq 0.$$

With these operators, we can write (6) and (7) in the operator form

$$(12) \quad \pi_t = \mathbb{I}_{z_{t-1} > 0} \theta_0(q_{t-1} - z_{t-1}) + \mathbb{I}_{z_{t-1} = 0} \theta(q_{t-1}, \pi_{t-1}),$$

$$(13) \quad p_t = \mathbb{I}_{z_{t-1} > 0} \varrho_0(q_{t-1} - z_{t-1}) + \mathbb{I}_{z_{t-1} = 0} \varrho(q_{t-1})p_{t-1},$$

with the initial conditions  $\pi_1 = p_1 = \pi$ . Once again, we emphasize that (13) is a linear equation, while (12) is not.

For the linear operator  $\varrho(q)$ , the norm is defined as

$$\|\varrho(q)\|_{\mathcal{L}} := \sup_{p \in \mathcal{H}} \frac{\|\varrho(q)p\|}{\|p\|},$$

where we adopt the convention that  $0/0 = 0$ , here and throughout. For this norm, we have the following lemma.

LEMMA 2.  $\|\varrho(q)p\| \leq \|p\| + q \int_0^{\infty} |p(y)| dy$  and  $\|\varrho(q)\|_{\mathcal{L}} \leq 1 + q$ .

From this lemma we can see that, when  $q = 0$ , the operator  $\varrho(q)p$  is a contraction mapping. Additional properties of operators  $\varrho$  and  $\theta$  are given in BCS.

**3. Dynamic programming formulation.** We assume that  $c(I, q)$  is a continuous function of linear growth in  $I$  for every fixed  $q$ ; i.e.,  $c(\cdot, q) \in \mathbb{L}^+$ . A cost function with linear growth satisfies the following inequality:

$$(14) \quad \text{Linear growth assumption: } c(y, q) \leq c_0 + c_1 q + h y^+ + s y^-, \quad y \in \mathfrak{R},$$

where  $c_1$ ,  $h$ , and  $s$  are positive constants. The constant  $c_0$  can be interpreted as the maximum expected backorder cost that can be incurred in a period if the beginning inventory level is zero. Indeed, we set  $c_0 = c(0, 0)$ .  $c_0$  will be bounded by the cost of backordering  $E(D)$  units. If the unit order cost is  $c$ , obviously the ordering cost  $cq$  constitutes a lower bound on  $c(I, q)$ . Under the linear growth assumption, we have the following lemma.

LEMMA 3. *If  $q_n \rightarrow q$  in the Euclidean norm and  $p_n \rightarrow p$  in the norm  $\|\cdot\|$  as  $n \rightarrow \infty$ , then*

$$\int_0^\infty c(y, q_n) p_n(y) dy \rightarrow \int_0^\infty c(y, q) p(y) dy.$$

From this lemma, we can see that the function

$$(15) \quad (q, p) \rightarrow \int_0^\infty c(y, q) p(y) dy$$

is continuous.

An example of a single-period cost with linear growth is  $c(I, q) = cq + hI^+ + sE[(D - I - q)^+]$ , which is common in the inventory literature [7]. The cost parameters  $c$ ,  $h$ , and  $s$  can be interpreted as the unit ordering cost, the unit holding cost for each unit of inventory carried to this period from the previous period, and the penalty cost for each backordered unit. Since  $c(I, q) = cq + hI^+ + sE[(D - I - q)^+] \leq cq + hI^+ + sE[(D - I)^+] \leq cq + hI^+ + sE(D) + sI^- = c_0 + cq + hI^+ + sI^-$ , it holds that  $c(I, q)$  is of linear growth in  $I$  for a fixed  $q$ .

The total cost function can be written as

$$(16) \quad \begin{aligned} J(\zeta, \pi, \tilde{q}) &= \sum_{t=1}^{\infty} \alpha^{t-1} E[E[c(I_t, q_t) | \mathcal{Z}_t]] \\ &= \sum_{t=1}^{\infty} \alpha^{t-1} E\{\mathbb{I}_{z_t > 0} c(-z_t, q_t) + \mathbb{I}_{z_t = 0} \langle c(I_t, q_t), \pi_t \rangle\}, \end{aligned}$$

where  $\pi_t$  is the solution of (6) and  $z_1 = \zeta \in \mathbb{R}^+$  is the initial condition. When  $\zeta = 0$ , the inventory level at the beginning of the first period is nonnegative and the given density function is  $\pi_1 = \pi$ . When  $\zeta > 0$ , it is the observed backordered quantity at the beginning of period one.

The value function is defined as

$$(17) \quad V(\zeta, \pi) := \inf_{\tilde{q} \in \Gamma} J(\zeta, \pi, \tilde{q}).$$

By looking one period ahead from period one, the value function can be written as

$$(18) \quad V(\zeta, \pi) = \inf_q \left\{ \mathbb{I}_{\zeta > 0} c(-\zeta, q) + \mathbb{I}_{\zeta = 0} \int_0^\infty c(y, q) \pi(y) dy + \alpha E[V(z_2, \pi_2) | \zeta, \pi] \right\},$$

which is the inventory-related cost in period one plus the cost-to-go.

By (1) and (2), we can express

$$z_2 = (I_1 + q - D_1)^- = \mathbb{I}_{\zeta > 0} (-\zeta + q - D_1)^- + \mathbb{I}_{\zeta = 0} (I_1 + q - D_1)^-.$$

Then, the cost-to-go can be written as

$$(19) \quad \begin{aligned} \mathbb{E}[V(z_2, \pi_2)|\zeta, \pi] &= \mathbb{I}_{\zeta>0} \mathbb{E}[V((- \zeta + q - D)^-, \theta_0(q - \zeta))] \\ &\quad + \mathbb{I}_{\zeta=0} \mathbb{E}[V((I_1 + q - D)^-, \theta(q, \pi))]. \end{aligned}$$

By conditioning on the event  $[q - \zeta < D]$ , we can express the first term on the right-hand side of (19) as

$$(20) \quad \begin{aligned} \mathbb{E}[V((- \zeta + q - D)^-, \theta_0(q - \zeta))] &= V(0, \theta_0(q - \zeta))F(q - \zeta) \\ &\quad + \int_{q-\zeta}^{\infty} V(\eta + \zeta - q, \theta_0(q - \zeta))f(\eta)d\eta. \end{aligned}$$

For a fixed  $I_1$ , we obtain

$$(21) \quad \begin{aligned} \mathbb{E}[V((I_1 + q - D)^-, \theta(q, \pi))|I_1] &= V(0, \theta(q, \pi))F(I_1 + q) \\ &\quad + \int_{q+I_1}^{\infty} V(\eta - I_1 - q, \theta(q, \pi))f(\eta)d\eta. \end{aligned}$$

Using (21), we can express the second term on the right-hand side of (19) explicitly as

$$\begin{aligned} &\mathbb{I}_{\zeta=0} \mathbb{E}[V((I_1 + q - D)^-, \theta(q, \pi))] \\ &= \mathbb{I}_{\zeta=0} V(0, \theta(q, \pi)) \int_0^{\infty} F(y + q)\pi(y)dy \\ &\quad + \mathbb{I}_{\zeta=0} \int_0^{\infty} \pi(y) \left[ \int_{q+y}^{\infty} V(\eta - y - q, \theta(q, \pi))f(\eta)d\eta \right] dy \\ &= \mathbb{I}_{\zeta=0} V(0, \theta(q, \pi)) \int_0^{\infty} F(y + q)\pi(y)dy \\ &\quad + \mathbb{I}_{\zeta=0} \int_0^{\infty} \pi(y) \int_0^{\infty} V(\eta, \theta(q, \pi))f(q + y + \eta)d\eta dy. \end{aligned}$$

From (20) and the above equation, we obtain

$$\begin{aligned} &\mathbb{E}[V(z_2, \pi_2)|\zeta, \pi] \\ &= \mathbb{I}_{\zeta>0} \left\{ V(0, \theta_0(q - \zeta))F(q - \zeta) + \int_0^{\infty} V(\eta, \theta_0(q - \zeta))f(\eta + q - \zeta)d\eta \right\} \\ &\quad + \mathbb{I}_{\zeta=0} V(0, \theta(q, \pi)) \int_0^{\infty} F(y + q)\pi(y)dy \\ &\quad + \mathbb{I}_{\zeta=0} \int_0^{\infty} \pi(y) \int_0^{\infty} V(\eta, \theta(q, \pi))f(q + y + \eta)d\eta dy. \end{aligned}$$

From the above expression, we can see that, when  $\zeta > 0$ , we have

$$\begin{aligned} V(\zeta, \pi) &= \inf_q \left\{ c(-\zeta, q) + \alpha V(0, \theta_0(q - \zeta))F(q - \zeta) \right. \\ &\quad \left. + \alpha \int_0^{\infty} V(\eta, \theta_0(q - \zeta))f(\eta + q - \zeta)d\eta \right\}. \end{aligned}$$

Since  $V(\zeta, \pi)$  is independent of  $\pi$  when  $\zeta > 0$ , we can let  $v(\zeta) := V(\zeta, \pi)$ .

On the other hand, when  $\zeta = 0$ ,

$$\begin{aligned} V(0, \pi) = \inf_q \bigg\{ & \int_0^\infty c(y, q) \pi(y) dy \\ & + \alpha V(0, \theta(q, \pi)) \int_0^\infty F(y + q) \pi(y) dy \\ & + \alpha \int_0^\infty \pi(y) \int_0^\infty V(\eta, \theta(q, \pi)) f(q + y + \eta) d\eta dy \bigg\}. \end{aligned}$$

If we denote  $W(\pi) := V(0, \pi)$ , then together with  $v(\zeta)$  we have the system

$$\begin{aligned} v(\zeta) = \inf_q \bigg\{ & c(-\zeta, q) + \alpha W(\theta_0(q - \zeta)) F(q - \zeta) \\ & + \alpha \int_0^\infty v(\eta) f(\eta + q - \zeta) d\eta \bigg\} \quad \text{for } \zeta > 0, \end{aligned} \quad (22)$$

$$\begin{aligned} W(\pi) = \inf_q \bigg\{ & \int_0^\infty c(y, q) \pi(y) dy \\ & + \alpha W(\theta(q, \pi)) \int_0^\infty F(y + q) \pi(y) dy \\ & + \alpha \int_0^\infty \pi(y) \int_0^\infty v(\eta) f(q + y + \eta) d\eta dy \bigg\}. \end{aligned} \quad (23)$$

Observe that

$$V(\zeta, \pi) = \mathbb{I}_{\zeta > 0} v(\zeta) + \mathbb{I}_{\zeta = 0} W(\pi), \quad (\zeta, \pi) \in \mathbb{R}^+ \times \Pi. \quad (24)$$

Since  $\pi$  evolves according to the nonlinear operator  $\theta$ , a direct study of the system (22)–(23) is not easy. If we can use  $p \in \mathcal{H}^+$  as the system state instead of  $\pi$ , then the analysis becomes much easier, because  $p$  evolves with the linear operator  $\varrho$ . To make ideas concrete, we define a new value function  $Y(\cdot)$  as follows:

$$Y(p) := W\left(\frac{p}{\lambda}\right) \lambda, \quad \lambda := \int_0^\infty p(x) dx. \quad (25)$$

Then, from (23) we have

$$\begin{aligned}
 Y(p) &= \lambda \inf_q \left\{ \int_0^\infty c(y, q)(p(y)/\lambda) dy \right. \\
 &\quad + \alpha W(\theta(q, p/\lambda)) \int_0^\infty F(y + q)(p(y)/\lambda) dy \\
 &\quad \left. + \alpha \int_0^\infty p(y)/\lambda \int_0^\infty v(\eta) f(q + y + \eta) d\eta dy \right\} \\
 &= \inf_q \left\{ \int_0^\infty c(y, q)p(y) dy + \alpha \left[ W(\theta(q, p)) \int_0^\infty F(y + q)p(y) dy \right. \right. \\
 &\quad \left. \left. + \int_0^\infty p(y) \int_0^\infty v(\eta) f(q + y + \eta) d\eta dy \right] \right\}.
 \end{aligned}$$

The above equation follows from the fact that the operator  $\theta$  is homogenous of degree 0. Next, we simplify the term  $W(\theta(q, \pi))$  and  $W(\theta_0(q - \zeta))$  on the right-hand side of the above equation and of (22), respectively. From the definition of  $\theta$  and (11), we get

$$\begin{aligned}
 Y(\varrho(q)p) &= \left\{ \int_0^\infty \varrho(q)p(x) dx \right\} \left\{ W \left( \frac{\varrho(q)p}{\langle \varrho(q)p, 1 \rangle} \right) \right\} \\
 &\stackrel{(11)}{=} \left\{ \int_0^\infty F(y + q)p(y) dy \right\} \{ W(\theta(q, p)) \}.
 \end{aligned}$$

By the definition of  $\theta_0$ , we get

$$Y(\varrho_0(q - \zeta)) = W \left( \frac{\varrho_0(q - \zeta)}{F(q - \zeta)} \right) F(q - \zeta) = W(\theta_0(q - \zeta)) F(q - \zeta).$$

Using these results in the expression of  $Y(p)$ , we obtain the following new system of equations:

$$(26) \quad v(\zeta) = \inf_q \left\{ c(-\zeta, q) + \alpha \int_0^\infty v(\eta) f(\eta + q - \zeta) d\eta + \alpha Y(\varrho_0(q - \zeta)) \right\}, \quad \zeta \in \mathfrak{R}^+,$$

$$\begin{aligned}
 Y(p) &= \inf_q \left\{ \int_0^\infty c(y, q)p(y) dy + \alpha Y(\varrho(q)p) \right. \\
 (27) \quad &\quad \left. + \alpha \int_0^\infty p(y) \int_0^\infty v(\eta) f(y + q + \eta) d\eta dy \right\}, \quad p \in \mathcal{H}^+.
 \end{aligned}$$

As in (24), observe that the corresponding unnormalized value function  $Z : \mathfrak{R}^+ \times \mathcal{H}^+ \rightarrow \mathfrak{R}$  takes the form

$$(28) \quad Z(\zeta, p) = \mathbb{I}_{\zeta > 0} v(\zeta) + \mathbb{I}_{\zeta = 0} Y(p), \quad (\zeta, p) \in \mathfrak{R}^+ \times \mathcal{H}^+.$$



Moreover,

$$(29) \quad Y(\mu p) = \mu Y(p) \quad \text{for every } \mu > 0.$$

Thus,  $Y(0) = 0$ .

Compared with the value functions in BCS, the left-hand side of (26) is an unknown number  $v$ , but in the present paper it is a function  $v(\cdot)$ . This makes the analysis of the system more difficult. In what follows, we will show the details of the methodology we use to prove the existence of an optimal control. Important costs and spaces are summarized in Table 1. Sequences of cost functions and some of the spaces will be formally defined soon. In this paper, we interpret decreasing and increasing sequences in a nonstrict way.

TABLE 1  
*Frequently used notation.*

Definition	Normalized system		Unnormalized system	
	Function	Space	Function	Space
Probability	$\pi$	$\Pi$	$p$	$\mathcal{H}^+$
Value function—backorder case	$v$	$\mathbb{L}^+$	$v$	$\mathbb{L}^+$
Value function—no backorder case	$W$	$\mathcal{B}^\Pi$	$Y$	$\mathcal{B}^+$
Value function—combined case	$V$	$\tilde{\mathcal{G}}$	$Z$	$\mathcal{G}$
Increasing value iteration sequence	$\{V_n\}$	$\tilde{\mathcal{G}}$	$\{Z_n\}$	$\mathcal{G}$
Decreasing value iteration sequence	$\{V^n\}$	$\tilde{\mathcal{G}}$	$\{Z^n\}$	$\mathcal{G}$

**4. Main results.** When the cost functions are bounded, we can apply commonplace arguments of the standard theory (see, e.g., [6]) to obtain a contraction property for an appropriate operator. Then by using Banach's fixed point theorem, we prove directly the existence and uniqueness of the solution to the optimality equation as well as the value iteration algorithm. However, consideration of the unbounded one-period costs as in (14) brings some challenges. For instance, the nice contraction property does not work for the discounted cost criterion, and so we need to consider ceiling functions for the optimal cost. These functions, in turn, will define the appropriate functional space where the solution to the system (22)–(23) is unique and is the value function. Moreover, besides the result of existence of the optimal ordering policy, ceiling functions allow us to prove that the value iteration algorithm converges to the value function.

Before proving our main results, we give some preliminaries.

**4.1. Preliminaries.** We define the functional space

$$\mathcal{B} := \left\{ \phi(p) : \mathcal{H} \rightarrow \mathbb{R} : \sup_{p \in \mathcal{H}} \frac{|\phi(p)|}{\|p\|} < \infty \right\}$$

with the norm

$$\|\phi\|_{\mathcal{B}} := \sup_{p \in \mathcal{H}} \frac{|\phi(p)|}{\|p\|}.$$

From the definition, we can see that  $\mathcal{B}$  is a Banach space. For any  $\phi \in \mathcal{B}$ , we must have  $\phi(0) = 0$ .

We define the space

$$\mathcal{B}^+ := \left\{ \phi : \mathcal{H}^+ \rightarrow \mathfrak{R}^+ : \sup_{p \in \mathcal{H}^+} \frac{\phi(p)}{\|p\|} < \infty \right\} \subseteq \mathcal{B}$$

and denote by  $\mathcal{B}^\Pi$  the subset of  $\mathcal{B}^+$  restricted to  $\Pi$ . That is,

$$\mathcal{B}^\Pi := \left\{ \phi : \Pi \rightarrow \mathfrak{R}^+ : \sup_{\pi \in \Pi} \frac{\phi(\pi)}{\|\pi\|} < \infty \right\}.$$

Similar to  $\mathbb{L}^+$ ,  $\mathcal{B}^+$  is also used to accommodate the dynamic programming cost.

We consider  $v^0 \in \mathbb{L}^+$  and  $Y^0 \in \mathcal{B}^+$  defined as

$$(30) \quad v^0(\zeta) := \frac{\alpha s E(D)}{(1-\alpha)^2} + \frac{c_0 + s\zeta}{1-\alpha}, \quad \zeta \in \mathfrak{R}^+.$$

$$(31) \quad Y^0(p) := \frac{a_0}{1-\alpha} \|p\|, \quad p \in \mathcal{H}^+,$$

where

$$a_0 = \max \left\{ h, \frac{\alpha s E(D)}{(1-\alpha)^2} + \frac{c_0 + sE(D)}{1-\alpha} \right\}.$$

For any  $\phi_1 \in \mathbb{L}^+$  and  $\phi_2 \in \mathcal{B}^+$  such that  $\phi_1(\cdot) \leq v^0(\cdot)$  and  $\phi_2(\cdot) \leq Y^0(\cdot)$ , we define the set

$$\mathcal{G} = \left\{ \phi : \mathfrak{R}^+ \times \mathcal{H}^+ \rightarrow \mathfrak{R} \mid \phi(\zeta, p) = \mathbb{I}_{\zeta > 0} \phi_1(\zeta) + \mathbb{I}_{\zeta = 0} \phi_2(p) \right.$$

$$\left. \text{and } \phi(\zeta, p) \leq \mathbb{I}_{\zeta > 0} v^0(\zeta) + \mathbb{I}_{\zeta = 0} Y^0(p) \right\}.$$

Thus the domain of the members of the set  $\mathcal{G}$  is  $\mathcal{H}^+ \times \mathfrak{R}^+$ , where  $\mathcal{H}^+$  is the space of unnormalized probabilities. If  $\phi_1(\zeta)$  and  $\phi_2(p)$  are solutions of the system (26)–(27), then they correspond to  $v(\zeta)$  and  $Y(p)$ , respectively. As a result,  $\phi(\zeta, p)$  corresponds to  $Z(\zeta, p)$ . The counterpart of the normalized probabilities  $\Pi \times \mathfrak{R}^+$  is denoted by  $\tilde{\mathcal{G}}$  and defined for functions  $\tilde{\phi}_1 \in \mathbb{L}^+$  and  $\tilde{\phi}_2 \in \mathcal{B}^\Pi$  such that  $\tilde{\phi}_1(\cdot) \leq v^0(\cdot)$  and  $\tilde{\phi}_2(\cdot) \leq Y^0(\cdot)$ :

$$\tilde{\mathcal{G}} = \left\{ \tilde{\phi} : \mathfrak{R}^+ \times \Pi \rightarrow \mathfrak{R} \mid \tilde{\phi}(\zeta, \pi) = \mathbb{I}_{\zeta > 0} \tilde{\phi}_1(\zeta) + \mathbb{I}_{\zeta = 0} \tilde{\phi}_2(\pi) \right.$$

$$\left. \text{and } \tilde{\phi}(\zeta, \pi) \leq \mathbb{I}_{\zeta > 0} v^0(\zeta) + \mathbb{I}_{\zeta = 0} Y^0(\pi) \right\}.$$

If  $\tilde{\phi}_1$  and  $\tilde{\phi}_2$  are solutions of the system (22)–(23), then  $\tilde{\phi}_1$  corresponds to  $v(\zeta)$  and  $\tilde{\phi}_2$  corresponds to  $W(\pi)$ . As a result  $\tilde{\phi}(\zeta, \pi) = V(\zeta, \pi)$ . In the next subsection, we will show that the value functions  $Z$  and  $V$  are in the sets  $\mathcal{G}$  and  $\tilde{\mathcal{G}}$ , respectively.

For  $\phi \in \mathcal{G}$ , we define the operators

$$T_q^{(1)}(\phi_1, \phi_2) = c(-\zeta, q) + \alpha \int_0^\infty \phi_1(\eta) f(\eta + q - \zeta) d\eta + \alpha \phi_2(\varrho_0(q - \zeta)),$$

$$\begin{aligned} T_q^{(2)}(\phi_1, \phi_2) &= \int_0^\infty c(y, q) p(y) dy \\ &\quad + \alpha \int_0^\infty p(y) \int_0^\infty \phi_1(\eta) f(y + q + \eta) d\eta dy + \alpha \phi_2(\varrho(q)p). \end{aligned}$$

We can see that if  $\phi_1$  and  $\phi_2$  are solutions of (26)–(27), then  $T_q^{(1)}(\phi_1, \phi_2)$  and  $T_q^{(2)}(\phi_1, \phi_2)$  represent the cost when the ordering quantity is  $q$  and when there are backorders and no backorders, respectively. By combining these two cases, we define the operator  $T_q$  as follows:

$$T_q\phi(\zeta, p) := \mathbb{I}_{\zeta>0}T_q^{(1)}(\phi_1, \phi_2) + \mathbb{I}_{\zeta=0}T_q^{(2)}(\phi_1, \phi_2).$$

Let

$$(32) \quad \begin{aligned} T^{(1)}(\phi_1, \phi_2) &:= \inf_{q \geq 0} T_q^{(1)}(\phi_1, \phi_2), \\ T^{(2)}(\phi_1, \phi_2) &:= \inf_{q \geq 0} T_q^{(2)}(\phi_1, \phi_2), \\ T\phi(\zeta, p) &:= \inf_{q \geq 0} T_q\phi(\zeta, p). \end{aligned}$$

The operators  $T^1$ ,  $T^2$ , and  $T$  represent the corresponding optimal costs. After defining the above operators, we have the following lemma.

LEMMA 4. *When  $\zeta \geq 0$ , for any  $\tilde{q} \in \Gamma$ ,  $J(\zeta, \pi, \tilde{q}) \leq v^0(\zeta)$ . Moreover, the operator  $T$  maps  $\mathcal{G}$  into itself, and, for  $\phi_1 \leq v^0$  and  $\phi_2 \leq Y^0$ ,  $T^{(1)}(\phi_1, \phi_2) \leq v^0(\zeta)$  and  $T^{(2)}(\phi_1, \phi_2) \leq Y^0(p)$ .*

From this lemma, we can see that  $v^0$  and  $Y^0$  are ceiling functions for the optimal cost. The solution  $(v, Y)$  of the system (26)–(27) satisfies  $v \leq v^0$  and  $Y \leq Y^0$ .

**4.2. Existence of a solution of the Bellman equation and an optimal feedback policy.** We define a value iteration procedure as follows. Let  $\{v_n\}$  and  $\{Y_n\}$  be sequences of functions defined by  $v_1 = Y_1 = 0$ , and, for  $n \geq 1$ ,

$$(33) \quad v_{n+1} = T^{(1)}(v_n, Y_n), \quad Y_{n+1} = T^{(2)}(v_n, Y_n).$$

From the definitions of operators  $T^{(1)}$  and  $T^{(2)}$ ,  $\{v_n\}$  and  $\{Y_n\}$  are increasing sequences. This can be established by induction. Clearly,  $v_1 \leq v_2$  and  $Y_1 \leq Y_2$ . Let us assume that  $v_n \leq v_{n+1}$  and  $Y_n \leq Y_{n+1}$ . Then,

$$\begin{aligned} v_{n+1}(\zeta) &= T^{(1)}(v_n, Y_n) \\ &= \inf_{q \geq 0} \left\{ c(-\zeta, q) + \alpha \int_0^\infty v_n(\eta) f(\eta + q - \zeta) d\eta + \alpha Y_n(\varrho_0(q - \zeta)) \right\} \\ &\leq \inf_{q \geq 0} \left\{ c(-\zeta, q) + \alpha \int_0^\infty v_{n+1}(\eta) f(\eta + q - \zeta) d\eta + \alpha Y_{n+1}(\varrho_0(q - \zeta)) \right\} \\ &= T^{(1)}(v_{n+1}, Y_{n+1}) = v_{n+2}(\zeta), \end{aligned}$$

and similarly we prove that  $Y_{n+1}(p) \leq Y_{n+2}(p)$ .

Let

$$(34) \quad Z_n(\zeta, p) := \mathbb{I}_{\zeta>0}v_n(\zeta) + \mathbb{I}_{\zeta=0}Y_n(p), \quad (\zeta, p) \in \mathbb{R}^+ \times \mathcal{H}^+.$$

Observe that

$$(35) \quad Z_{n+1}(\zeta, p) = \mathbb{I}_{\zeta>0}T^{(1)}(v_n(\zeta), Y_n(p)) + \mathbb{I}_{\zeta=0}T^{(2)}(v_n(\zeta), Y_n(p)) = TZ_n(\zeta, p).$$

Then we have the following theorem.

THEOREM 2. (a) *There exist lower semicontinuous (l.s.c.) functions  $\bar{Y} \leq Y^0$  and  $\bar{v} \leq v^0$  such that  $Z_n \nearrow \bar{Z}$  and  $\bar{Z} = T\bar{Z}$ , where*

$$\bar{Z}(\zeta, p) := \mathbb{I}_{\zeta > 0} \bar{v}(\zeta) + \mathbb{I}_{\zeta = 0} \bar{Y}(p), \quad (\zeta, p) \in \mathbb{R}^+ \times \mathcal{H}^+.$$

(b) *For each  $(\zeta, p) \in \mathbb{R}^+ \times \mathcal{H}^+$ , there exists a measurable function  $g_{\bar{Z}} : \mathbb{R}^+ \times \mathcal{H}^+ \rightarrow \mathbb{R}^+$  such that  $T\bar{Z}(\zeta, p) = T_{g_{\bar{Z}}(\zeta, p)} \bar{Z}(\zeta, p)$ .*

From this theorem we can see that  $\bar{v} \in \mathcal{H}^+$  and  $\bar{Y} \in \mathcal{B}^+$ . So far we have analyzed the unnormalized system (26)–(27). We show that there exists a solution of this system. Next, we will show that there is a solution of system (22)–(23). This result immediately follows from Theorem 2. Let  $\{(W_n, v_n)\}$  be the normalized value iteration functions corresponding to  $\{(Y_n, v_n)\}$ . That is,  $W_n$  is a function on  $\Pi$  satisfying (see (25))

$$(36) \quad Y_n(p) = W_n \left( \frac{p}{\int p(x) dx} \right) \int_0^\infty p(x) dx.$$

Let  $V_n(\zeta, \pi) := \mathbb{I}_{\zeta > 0} v_n(\zeta) + \mathbb{I}_{\zeta = 0} W_n(\pi)$ ,  $(\zeta, \pi) \in \mathbb{R}^+ \times \Pi$ . Then, as  $v_1 = W_1 = 0$ , we have  $V_1 = 0$ . Moreover,

$$V_{n+1}(\zeta, \pi) = \tilde{T}V_n(\zeta, \pi) := \inf_{q \geq 0} \tilde{T}_q V_n(\zeta, \pi), \quad (\zeta, \pi) \in \mathbb{R}^+ \times \Pi,$$

where, for a function  $\tilde{\phi} \in \tilde{\mathcal{G}}$ ,

$$\tilde{T}_q \tilde{\phi}(\zeta, \pi) := \mathbb{I}_{\zeta > 0} \tilde{T}_q^{(1)}(\tilde{\phi}_1(\zeta), \tilde{\phi}_2(\pi)) + \mathbb{I}_{\zeta = 0} \tilde{T}_q^{(2)}(\tilde{\phi}_1(\zeta), \tilde{\phi}_2(\pi))$$

and

$$\begin{aligned} \tilde{T}_q^{(1)}(\tilde{\phi}_1, \tilde{\phi}_2) &= c(-\zeta, q) + \alpha \tilde{\phi}_2(\theta_0(q - \zeta)) F(q - \zeta) + \alpha \int_0^\infty \tilde{\phi}_1(\eta) f(\eta + q - \zeta) d\eta, \\ \tilde{T}_q^{(2)}(\tilde{\phi}_1, \tilde{\phi}_2) &= \int_0^\infty c(y, q) \pi(y) dy + \alpha \left[ \tilde{\phi}_2(\theta(q, \pi)) \int_0^\infty F(y + q) \pi(y) dy \right. \\ &\quad \left. + \int_0^\infty \pi(y) \int_0^\infty \tilde{\phi}_1(\eta) f(q + y + \eta) d\eta dy \right]. \end{aligned}$$

The above two equations correspond to (22) and (23), which yield the value functions for the normalized probabilities.

We know that  $Y_n(\pi) = W_n(\pi)$  for all  $\pi \in \Pi$  from (36). From (29), Theorem 2(a) yields the existence of an l.s.c. function  $\bar{W}(\pi) \leq Y^0(\pi)$  such that  $W_n \nearrow \bar{W}$ . After defining  $\bar{V}(\zeta, \pi) := \mathbb{I}_{\zeta > 0} \bar{v}(\zeta) + \mathbb{I}_{\zeta = 0} \bar{W}(\pi)$ ,  $(\zeta, \pi) \in \mathbb{R}^+ \times \Pi$ , we have

$$(37) \quad V_n \nearrow \bar{V} \quad \text{as } n \rightarrow \infty,$$

and  $\bar{V}(\zeta, \pi) = \tilde{T}\bar{V}(\zeta, \pi)$  for  $(\zeta, \pi) \in \mathbb{R}^+ \times \Pi$ . Hence, similarly to (18), for each  $(\zeta, \pi) \in \mathbb{R}^+ \times \Pi$ ,

$$(38) \quad \bar{V}(\zeta, \pi) = \inf_q \left\{ \mathbb{I}_{\zeta > 0} c(-\zeta, q) + \mathbb{I}_{\zeta = 0} \int_0^\infty c(y, q) \pi(y) dy + \alpha E[\bar{V}(z_2, \pi_2) | \zeta, \pi] \right\}.$$

Furthermore, from Theorem 2(b), there exists a map  $\bar{g} : \mathbb{R}^+ \times \Pi \rightarrow \mathbb{R}^+$  such that

$$(39) \quad \bar{V}(\zeta, \pi) = \tilde{T}\bar{V}(\zeta, \pi) = \tilde{T}_{\bar{g}(\zeta, \pi)} \bar{V}(\zeta, \pi), \quad (\zeta, \pi) \in \mathbb{R}^+ \times \Pi.$$

So we can see that there is a solution of the system (22)–(23). In the next theorem, we establish that this solution is actually the value function.

**THEOREM 3.** (a) *For each  $(\zeta, \pi) \in \mathfrak{R}^+ \times \Pi$ , we have  $\bar{V}(\zeta, \pi) = V(\zeta, \pi)$ , where  $V$  is the optimal value function defined in (17). Hence,  $\bar{Z}(\zeta, p) = Z(\zeta, p)$  for each  $(\zeta, p) \in \mathfrak{R}^+ \times \mathcal{H}^+$ .*

(b) *The functions  $\bar{V}$  and  $\bar{Z}$  are the minimal solutions in  $\tilde{\mathcal{G}}$  and  $\mathcal{G}$  of the optimality equations (24) and (28), respectively.*

(c) *There exists an optimal feedback policy  $\tilde{q}^* \in \Gamma$  for the partially observed inventory problem. That is,*

$$V(\zeta, \pi) := \inf_{\tilde{q} \in \Gamma} J(\zeta, \pi, \tilde{q}) = J(\zeta, \pi, \tilde{q}^*) \quad \forall (\zeta, \pi) \in \mathfrak{R}^+ \times \Pi.$$

In this subsection, we have shown that, starting from  $v^0$  and  $Y^0$ , sequences  $\{v_n\}$  and  $\{Y_n\}$  are increasing and converge to  $\bar{v}$  and  $\bar{Y}$ , respectively, and that the pair  $(\bar{v}, \bar{Y})$  is a solution of the system (26)–(27). Since  $Y_n(\pi) = W_n(\pi)$ , sequences  $v_n$  and  $W_n$  converge to  $\bar{v}$  and  $\bar{W}$ , respectively. As a result the pair  $(\bar{v}, \bar{W})$  is a solution of (22)–(23). Theorem 3 states that  $\bar{V}(\zeta, \pi)$  is the value function defined in (17). In the next subsection, we will show that the solution of the system (26)–(27) is unique. As a result, the solution of the system (22)–(23) is also unique.

#### 4.3. Uniqueness and continuity of the solution of the Bellman equation.

We consider the unnormalized value iteration procedure (33)–(35), but starting with the functions  $v^0$  and  $Y^0$ . That is, for  $(\zeta, p) \in \mathfrak{R}^+ \times \mathcal{H}^+$ ,

$$(40) \quad \begin{aligned} v^1(\zeta) &= v^0(\zeta), & Y^1(p) &= Y^0(p), \\ v^{n+1}(\zeta) &= T^{(1)}(v^n(\zeta), Y^n(p)), & Y^{n+1}(p) &= T^{(2)}(v^n(\zeta), Y^n(p)). \end{aligned}$$

Using induction arguments, it is easy to see that  $\{v^n\}$  and  $\{Y^n\}$  are nonnegative decreasing sequences. From the previous subsection we have  $v^2 \leq v^1$  and  $Y^2 \leq Y^1$ . To establish the decreasing property for  $n \geq 1$ , we assume that it holds for  $n = k$ , that is,  $v^k \leq v^{k-1}$  and  $Y^k \leq Y^{k-1}$ , and then establish it for  $n = k+1$ . It follows that

$$\begin{aligned} v^{k+1}(\zeta) &= T^{(1)}(v^k, Y^k) \\ &= \inf_{q \geq 0} \left\{ c(-\zeta, q) + \alpha \int_0^\infty v^k(\eta) f(\eta + q - \zeta) d\eta + \alpha Y^k(\varrho_0(q - \zeta)) \right\} \\ &\leq \inf_{q \geq 0} \left\{ c(-\zeta, q) + \alpha \int_0^\infty v^{k-1}(\eta) f(\eta + q - \zeta) d\eta + \alpha Y^{k-1}(\varrho_0(q - \zeta)) \right\} \\ &= T^{(1)}(v^{k-1}, Y^{k-1}) = v^k(\zeta), \end{aligned}$$

and similarly we prove that  $Y^{k+1}(p) \leq Y^k(p)$ .

Since  $\{v^n\}$  and  $\{Y^n\}$  are nonnegative decreasing sequences and both have lower bounds 0, there exist functions  $\underline{v} \leq v^0$  and  $\underline{Y} \leq Y^0$  such that

$$(41) \quad v^n \searrow \underline{v} \quad \text{and} \quad Y^n \searrow \underline{Y} \quad \text{as} \quad n \rightarrow \infty.$$

In other words, if we denote

$$\begin{aligned} Z^n(\zeta, p) &= \mathbb{I}_{\zeta > 0} v^n(\zeta) + \mathbb{I}_{\zeta = 0} Y^n(p), \\ \underline{Z}(\zeta, p) &= \mathbb{I}_{\zeta > 0} \underline{v}(\zeta) + \mathbb{I}_{\zeta = 0} \underline{Y}(p), \quad (\zeta, p) \in \mathfrak{R}^+ \times \mathcal{H}^+, \end{aligned}$$

we have

$$(42) \quad Z^{n+1} = TZ^n \text{ and } \underline{Z} \in \mathcal{G}, \quad Z^n \searrow \underline{Z}, \quad \text{as } n \rightarrow \infty.$$

Since  $Z^n$  is a decreasing sequence, we have  $Z^n \geq \underline{Z}$  for a finite  $n$ . Applying  $T$  infinitely many times on both sides, we obtain

$$(43) \quad \underline{Z} \geq T\underline{Z}.$$

Moreover, by applying the monotone convergence theorem together with (40) and (41), we can prove that for all  $q \in \mathbb{R}^+$ ,  $\underline{v}(\zeta) \leq T_q^{(1)}(\underline{v}(\zeta), \underline{Y}(p))$  and  $\underline{Y}(p) \leq T_q^{(2)}(\underline{v}(\zeta), \underline{Y}(p))$ , which implies  $\underline{Z}(\zeta, p) \leq T\underline{Z}(\zeta, p)$ . This inequality combined with (43) yields

$$(44) \quad \underline{Z}(\zeta, p) = T\underline{Z}(\zeta, p).$$

Now, because  $Z \in \mathcal{G}$  (see (28) and Theorem 3(b)), we have  $v(\zeta) \leq v^0(\zeta)$  and  $Y(p) \leq Y^0(p)$ . Then, as  $Z(\zeta, p)$  is a fixed point of the operator  $T$ , we obtain  $Z(\zeta, p) \leq Z^n(\zeta, p)$  for all  $n$ , which implies from (42) that

$$(45) \quad Z(\zeta, p) \leq \underline{Z}(\zeta, p) \quad \forall (\zeta, p) \in \mathbb{R}^+ \times \mathcal{H}^+.$$

On the other hand, let  $W^n : \Pi \rightarrow \mathbb{R}$  be the corresponding normalized value iteration function defined in the same way as (36), and define  $V^n(\zeta, \pi) = \mathbb{I}_{\zeta > 0} v^n(\zeta) + \mathbb{I}_{\zeta = 0} W^n(\pi)$ . Observe that  $V^{n+1} = \tilde{T}V^n$ . Then, there exists a function  $\underline{V} \in \tilde{\mathcal{G}}$  such that  $V^n \searrow \underline{V}$ . Furthermore, from (44) and (45), we can prove that, for each  $(\zeta, \pi) \in \mathbb{R}^+ \times \Pi$ ,

$$(46) \quad \underline{V}(\zeta, \pi) = \tilde{T}\underline{V}(\zeta, \pi)$$

and

$$(47) \quad V(\zeta, \pi) \leq \underline{V}(\zeta, \pi).$$

**THEOREM 4.** Suppose  $\alpha(1 + \frac{a_0}{c(1-\alpha)}) < 1$ . Then we have the following.

(a) For each  $(\zeta, \pi) \in \mathbb{R}^+ \times \Pi$ ,  $V(\zeta, \pi) = \underline{V}(\zeta, \pi)$ . Moreover,  $\underline{V}$  is the maximal solution in  $\tilde{\mathcal{G}}$  of the optimality equation (24). Hence,  $Z(\zeta, p) = \underline{Z}(\zeta, p)$  for each  $(\zeta, p) \in \mathbb{R}^+ \times \mathcal{H}^+$ , and  $\underline{Z}$  is the maximal solution of (28) in  $\mathcal{G}$ . Moreover, the solutions of (24) and (28) are unique.

(b) The value functions  $V$  and  $Z$  are continuous on  $\mathbb{R}^+ \times \Pi$  and  $\mathbb{R}^+ \times \mathcal{H}^+$ , respectively.

**4.4. Conclusions.** We have studied an infinite-horizon single-stage periodic-review inventory control system with backorders. In this system, the inventory level is partially observed when it is nonnegative. Partial inventory observations eventually lead to a dynamic program in the space of probability distributions. This dynamic program is highly nonlinear. We apply the methodology of unnormalized probability to linearize the dynamic programming equations. This linearization facilitates the analysis.

As we discussed in the introduction, partial observations in inventory systems are very common. But there is not much research done on this topic. This motivated us to analyze this problem with rain checks. In Theorems 2 and 3, we prove the convergence of a value iteration algorithm to the optimal value function, as well as

the existence of an optimal feedback ordering policy. In addition, the value function is the minimal solution of the optimality equation. Furthermore, from Theorem 4, for a sufficiently small discount factor, the solution of the optimality equation is unique and continuous, and it is the value function.

We conclude the paper by describing an extension of the model and discussing some computational aspects related to the rain check model.

Sometimes, when the customer comes to the store during a stock out, he or she does not ask for a rain check, leaves the store, and comes back later. Some stores do not have a practice of offering rain checks. In the absence of rain checks, the IM cannot fully observe the backordered quantity. This setting gives rise to an interesting model where even less information is available to IM than in the current model. That is, in this new context, the IM cannot fully observe the negative inventory, and the available information to make his or her decisions in each period would be the sign of the inventory. The authors currently are studying this extension [3].

Although the unnormalized probability simplifies the analysis, it does not lead to a dynamic program in finite dimensional spaces. Finiteness of these spaces, which accommodate the conditional distribution  $p$  of the inventory level, can be imposed to develop approximations. The challenge here is to find an appropriate finite family of functions to represent  $p$ . This is a computational issue worth exploring. We provide some numerical techniques in [5] to solve the dynamic programming equation.

## 5. Proofs.

### 5.1. Preliminary results.

LEMMA 5. *The function*

$$(48) \quad (q, p) \rightarrow \int p(y) \int \phi_1(\eta) f(q + y + \eta) d\eta dy$$

is continuous for every bounded function  $\phi_1 : \mathbb{R}^+ \rightarrow \mathbb{R}$ .

*Proof.* Let  $\phi_1$  be a bounded function and  $\{(q_n, p_n)\}$  be a sequence in  $\mathbb{R}^+ \times \mathcal{H}^+$  converging to  $(q, p) \in \mathbb{R}^+ \times \mathcal{H}^+$ . Then, by adding and subtracting the term  $\int \int \phi_1(\eta) f(q_n + y + \eta) p(y) d\eta dy$ , we have

$$\begin{aligned} & \left| \int \int \phi_1(\eta) f(q_n + y + \eta) p_n(y) d\eta dy - \int \int \phi_1(\eta) f(q + y + \eta) p(y) d\eta dy \right| \\ & \leq \int \int \phi_1(\eta) f(q_n + y + \eta) |p_n(y) - p(y)| d\eta dy \\ & \quad + \int \int \phi_1(\eta) p(y) |f(q_n + y + \eta) - f(q + y + \eta)| d\eta dy, \end{aligned}$$

which, by the dominated convergence theorem, converges to zero as  $n \rightarrow \infty$ . This proves the continuity of the function defined in (48).  $\square$

The next lemma is a key result to prove that the operators  $T_q$  and  $T$  map continuous functions into continuous functions (see Remark 1(b) and Lemma 7). Thus, in each iteration of the monotone approximations introduced in section 4, we get continuous functions when we use both increasing and decreasing processes.

LEMMA 6. *For each nonnegative measurable function  $\phi_1$  on  $\mathbb{R}^+$ , the function*

$$(49) \quad (q, p) \rightarrow \int p(y) \int \phi_1(\eta) f(q + y + \eta) d\eta dy$$

is continuous.

*Proof.* Let  $\{(q_n, p_n)\}$  be a sequence in  $\mathbb{R}^+ \times \mathcal{H}^+$  converging to  $(q, p) \in \mathbb{R}^+ \times \mathcal{H}^+$ , and let  $\phi_1$  be a measurable function on  $\mathbb{R}^+$ . Then, there exists a sequence  $\{\phi_1^k\}$  of bounded functions on  $\mathbb{R}^+$  such that  $\phi_1^k \nearrow \phi_1$ , as  $k \rightarrow \infty$ . Therefore from Lemma 5, for each  $k \in \mathbb{N}$ ,

$$\liminf_{n \rightarrow \infty} \int \int \phi_1^k(\eta) f(q_n + y + \eta) p_n(y) d\eta dy = \int \int \phi_1^k(\eta) f(q + y + \eta) p(y) d\eta dy.$$

Thus,

$$\begin{aligned} \liminf_{n \rightarrow \infty} \int \int \phi_1(\eta) f(q_n + y + \eta) p_n(y) d\eta dy \\ \geq \liminf_{n \rightarrow \infty} \int \int \phi_1^k(\eta) f(q_n + y + \eta) p_n(y) d\eta dy \\ = \int \int \phi_1^k(\eta) f(q + y + \eta) p(y) d\eta dy. \end{aligned}$$

Letting  $k \rightarrow \infty$  and using Fatou's lemma, we obtain

$$\liminf_{n \rightarrow \infty} \int \int \phi_1(\eta) f(q_n + y + \eta) p_n(y) d\eta dy \geq \int \int \phi_1(\eta) f(q + y + \eta) p(y) d\eta dy.$$

That is, the function (49) is l.s.c. Now, applying the same arguments to the function  $-\phi_1$ , we have that  $-\int \int \phi_1(\eta) f(q + y + \eta) p(y) d\eta dy$  is l.s.c. in  $(q, p)$ . Hence, the function (49) is upper semicontinuous (u.s.c.), which yields the desired result.  $\square$

*Remark 1.* (a) Applying arguments similar to those in the proofs of Lemmas 5 and 6, we can show the continuity of the functions

$$(x, q, p) \rightarrow \int f(y + q - x) p(y) dy, \quad (x, q, p) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathcal{H}^+,$$

and

$$(x, q) \rightarrow \int \phi_1(\eta) f(q - x + \eta) d\eta, \quad (x, q) \in \mathbb{R}^+ \times \mathbb{R}^+,$$

for each measurable function  $\phi_1$  on  $\mathbb{R}^+$ .

(b) Let  $\phi \in \mathcal{G}$  be a continuous function on  $\mathbb{R}^+ \times \mathcal{H}^+$ . That is,  $\phi$  is a function of the form  $\phi(\zeta, p) = \mathbb{I}_{\zeta > 0} \phi_1(\zeta) + \mathbb{I}_{\zeta = 0} \phi_2(p)$ , where  $\phi_1$  and  $\phi_2$  are continuous functions on  $\mathbb{R}^+$  and  $\mathcal{H}^+$ , respectively. Then, in view of (15), Lemmas 5 and 6, and Remark 1(a), we have that  $T_q \phi(\zeta, p)$  is continuous in  $(\zeta, q, p) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathcal{H}^+$ .

*Remark 2.* (a) Observe that by defining the operator  $T$  for each  $\phi \in \mathcal{G}$ , we can obtain bounds for the range of  $q$  for which  $T_q^{(1)}(\phi_1, \phi_2) \leq v^0(\zeta)$  and  $T_q^{(2)}(\phi_1, \phi_2) \leq Y^0(p)$ ,  $(\zeta, p) \in \mathbb{R}^+ \times \mathcal{H}^+$ . Hence, if we denote by  $q_-$  and  $q_+$  the ordering quantity under the negative and nonnegative initial inventory levels, respectively, then from Lemma 4 we have

$$cq_- \leq T_q^{(1)}(\phi_1(\zeta), \phi_2(p)) \leq v^0(\zeta) = \frac{\alpha s E(D)}{(1 - \alpha)^2} + \frac{c_0 + s\zeta}{1 - \alpha}.$$

Hence,

$$(50) \quad q_- \leq \frac{\alpha s E(D)}{(1 - \alpha)^2 c} + \frac{c_0 + s\zeta}{(1 - \alpha)c}.$$



Similarly, we have

$$cq_+ \int_0^\infty p(y)dy \leq T_q^{(2)}(\phi_1(\zeta), \phi_2(p)) \leq Y^0(p) = \frac{a_0}{1-\alpha} \|p\|,$$

which implies

$$(51) \quad q_+ \leq \frac{a_0}{c(1-\alpha) \int_0^\infty p(x)dx} \|p\|.$$

Then from (50) and (51), for each  $(\zeta, p) \in \mathfrak{R}^+ \times \mathcal{H}^+$ ,  $q$  must belong to  $\mathcal{Q}^*(\zeta, p) := \mathcal{Q}_1^*(\zeta) \cup \mathcal{Q}_2^*(p)$ , where

$$\mathcal{Q}_1^*(\zeta) := \left\{ q \in \mathfrak{R}^+ : q \leq \frac{\alpha s E(D)}{(1-\alpha)^2 c} + \frac{c_0 + s\zeta}{(1-\alpha)c} \right\}$$

and

$$\mathcal{Q}_2^*(p) := \left\{ q \in \mathfrak{R}^+ : q \leq \frac{a_0}{c(1-\alpha) \int_0^\infty p(x)dx} \|p\| \right\}.$$

Thus, for a fixed  $(\zeta, p) \in \mathfrak{R}^+ \times \mathcal{H}^+$ ,  $q$  remains bounded.

(b) Taking into account the latter, since for each  $\phi \in \mathcal{G}$  the map  $(\zeta, p, q) \rightarrow T_q \phi(\zeta, p)$  is continuous, there exists a measurable function  $g_\phi: \mathfrak{R}^+ \times \mathcal{H}^+ \rightarrow \mathfrak{R}^+$  that attains the minimum in (32). That is,

$$(52) \quad T\phi(\zeta, p) = T_{g_\phi} \phi(\zeta, p) \quad \forall (\zeta, p) \in \mathfrak{R}^+ \times \mathcal{H}^+.$$

LEMMA 7. For each continuous function  $\phi \in \mathcal{G}$ ,  $T\phi(\zeta, p)$  is continuous in  $(\zeta, p) \in \mathfrak{R}^+ \times \mathcal{H}^+$ .

*Proof.* Let  $\{(\zeta_k, p_k)\}$  be a sequence in  $\mathfrak{R}^+ \times \mathcal{H}^+$  such that  $(\zeta_k, p_k) \rightarrow (\zeta, p) \in \mathfrak{R}^+ \times \mathcal{H}^+$ ,  $p \neq 0$ , and  $q_k = g_\phi(\zeta_k, p_k) \in \mathcal{Q}^*(\zeta_k, p_k)$  satisfies (see Remark 2(b))

$$(53) \quad T\phi(\zeta_k, p_k) = T_{q_k} \phi(\zeta_k, p_k).$$

Clearly  $q_k$  remains in a compact set. Then, we can extract a subsequence  $\{(\zeta_{k_l}, p_{k_l}, q_{k_l})\}$  of  $\{(\zeta_k, p_k, q_k)\}$  such that  $(\zeta_{k_l}, p_{k_l}, q_{k_l}) \rightarrow (\zeta, p, q')$  for some  $q' \in \mathcal{Q}^*(\zeta, p)$ . Now from the continuity of  $T_q \phi(\zeta, p)$  (see Remark 1(b)), we have

$$\lim_{l \rightarrow \infty} T_{q_{k_l}} \phi(\zeta_{k_l}, p_{k_l}) = T_{q'} \phi(\zeta, p).$$

Hence, from (53) and (32),

$$(54) \quad \liminf_{k \rightarrow \infty} T\phi(\zeta_k, p_k) = T_{q'} \phi(\zeta, p) \geq T\phi(\zeta, p).$$

On the other hand, we have

$$T\phi(\zeta_k, p_k) \leq T_q \phi(\zeta_k, p_k) \quad \forall q \in \mathfrak{R}^+.$$

Then, again from the continuity of  $T_q \phi(\zeta, p)$ ,

$$\limsup_{k \rightarrow \infty} T\phi(\zeta_k, p_k) \leq T_q \phi(\zeta, p) \quad \forall q \in \mathfrak{R}^+,$$

which yields

$$(55) \quad \limsup_{k \rightarrow \infty} T\phi(\zeta_k, p_k) \leq T\phi(\zeta, p).$$

Therefore, by combining (54) and (55), we obtain

$$\lim_{k \rightarrow \infty} T\phi(\zeta_k, p_k) = T\phi(\zeta, p). \quad \square$$

*Remark 3.* (a) If  $\phi \in \mathcal{G}$  is only l.s.c., we can follow arguments similar to those in the proof of Lemma 7 (see also Remark 1(b)) to show that the functions  $T_q\phi(\zeta, p)$  and  $T\phi(\zeta, p)$  are l.s.c. in  $(\zeta, p, q)$  and  $(\zeta, p)$ , respectively. In addition, the selection theorem ensures the existence of a measurable function  $g_\phi: \mathfrak{R}^+ \times \mathcal{H}^+ \rightarrow \mathfrak{R}^+$  satisfying (52).

(b) It is worth noting that if  $\phi_1: \mathfrak{R}^+ \rightarrow \mathfrak{R}^+$  and  $\phi_2: \mathcal{H}^+ \rightarrow \mathfrak{R}^+$  are continuous, then  $T^{(1)}(\phi_1(\zeta), \phi_2(p))$  and  $T^{(2)}(\phi_1(\zeta), \phi_2(p))$  are continuous (see (64) and (65) for definitions of  $T^{(1)}$  and  $T^{(2)}$ ).

**5.2. Proof of Lemma 1.** We can express the left-hand side of (5) as

$$(56) \quad \mathbb{E}(\varphi(I_t)|\mathcal{Z}_t) = \mathbb{E}[\varphi(I_t)(\mathbb{I}_{z_t>0} + \mathbb{I}_{z_t=0})|\mathcal{Z}_t] = \varphi(-z_t)\mathbb{I}_{z_t>0} + \mathbb{E}[\varphi(I_t)\mathbb{I}_{z_t=0}|\mathcal{Z}_t].$$

From the last term of (56), we have

$$\begin{aligned} \mathbb{E}(\varphi(I_t)\mathbb{I}_{z_t=0}|\mathcal{Z}_t) &= \mathbb{I}_{z_t=0}\mathbb{E}(\varphi(I_t)|\mathcal{Z}_t) \\ &= \mathbb{I}_{z_t=0}\psi(z_1, \dots, z_{t-1}, z_t) \\ (57) \quad &= \mathbb{I}_{z_t=0}\psi(z_1, \dots, z_{t-1}, 0), \end{aligned}$$

where the first equality follows from the fact that  $\mathbb{I}_{z_t=0}$  is  $\mathcal{Z}_t$  measurable. The second equality introduces a measurable function  $\psi$  to express  $\mathbb{E}(\varphi(I_t)|\mathcal{Z}_t)$  as  $\psi(z_1, \dots, z_{t-1}, z_t)$ . If we set  $z_t = 0$  into the second equality, we can obtain the last equality.

Taking conditional expectation of (57) with respect to  $\mathcal{Z}_{t-1}$  and observing that  $\mathcal{Z}_{t-1} \subseteq \mathcal{Z}_t$  and  $\psi(z_1, \dots, z_{t-1}, 0)$  is  $\mathcal{Z}_{t-1}$ -measurable, we obtain

$$\begin{aligned} \mathbb{E}[\varphi(I_t)\mathbb{I}_{z_t=0}|\mathcal{Z}_{t-1}] &= \psi(z_1, \dots, z_{t-1}, 0)\mathbb{E}[\mathbb{I}_{I_t \geq 0}|\mathcal{Z}_{t-1}] \\ &= \psi(z_1, \dots, z_{t-1}, 0)\mathbb{P}(I_t \geq 0|\mathcal{Z}_{t-1}) \end{aligned}$$

or

$$(58) \quad \psi(z_1, \dots, z_{t-1}, 0) = \frac{\mathbb{E}(\varphi(I_t)\mathbb{I}_{I_t \geq 0}|\mathcal{Z}_{t-1})}{\mathbb{P}(I_t \geq 0|\mathcal{Z}_{t-1})}.$$

We insert (58) into (57) and then the result into (56) to get

$$\mathbb{E}(\varphi(I_t)|\mathcal{Z}_t) = \mathbb{I}_{z_t>0}\varphi(-z_t) + \mathbb{I}_{z_t=0} \frac{\mathbb{E}(\varphi(I_t)\mathbb{I}_{z_t=0}|\mathcal{Z}_{t-1})}{\mathbb{P}(I_t \geq 0|\mathcal{Z}_{t-1})}.$$

By using (3) and (4), we have

$$\mathbb{E}(\varphi(I_t)|\mathcal{Z}_t) = \mathbb{I}_{z_t>0}\varphi(-z_t) + \mathbb{I}_{z_t=0} \int_0^\infty \varphi(\eta)\pi_t(\eta)d\eta. \quad \square$$

**5.3. Proof of Theorem 1.** We start to prove this theorem from the right-hand side of (4). Take the numerator and obtain

$$\begin{aligned}
 & \mathbb{E}(\varphi(I_t) \mathbb{I}_{z_t=0} | \mathcal{Z}_{t-1}) \\
 &= \mathbb{E}(\varphi(I_{t-1} + q_{t-1} - D_{t-1}) \mathbb{I}_{I_{t-1}+q_{t-1}-D_{t-1} \geq 0} | \mathcal{Z}_{t-1}) \\
 &= \mathbb{E} \left( \mathbb{E}(\varphi(I_{t-1} + q_{t-1} - D_{t-1}) \mathbb{I}_{I_{t-1}+q_{t-1}-D_{t-1} \geq 0} | \mathcal{Z}_{t-1}, I_{t-1}) | \mathcal{Z}_{t-1} \right) \\
 &\quad \text{because } \mathcal{Z}_{t-1} = \sigma(\{z_1, \dots, z_{t-1}\}) \subseteq \sigma(\{z_1, \dots, z_{t-1}, I_{t-1}\}) \\
 &= \mathbb{E} \left( \int_0^\infty \varphi(I_{t-1} + q_{t-1} - y) \mathbb{I}_{I_{t-1}+q_{t-1}-y \geq 0} f(y) dy | \mathcal{Z}_{t-1} \right) \\
 &\quad \text{set } x := I_{t-1} + q_{t-1} - y \\
 &= \mathbb{E} \left( \int_0^\infty \varphi(x) f(I_{t-1} + q_{t-1} - x) \mathbb{I}_{I_{t-1}+q_{t-1}-x \geq 0} dx | \mathcal{Z}_{t-1} \right) \\
 (59) \quad &= \int_0^\infty \varphi(x) \mathbb{E} (f(I_{t-1} + q_{t-1} - x) \mathbb{I}_{I_{t-1}+q_{t-1}-x \geq 0} | \mathcal{Z}_{t-1}) dx.
 \end{aligned}$$

Letting the time index in the first equality of (5) be  $t-1$  instead of  $t$  and replacing  $\varphi(I_{t-1})$  with  $f(I_{t-1} + q_{t-1} - x) \mathbb{I}_{I_{t-1}+q_{t-1}-x \geq 0}$ , we obtain

$$\begin{aligned}
 & \mathbb{E}(f(I_{t-1} + q_{t-1} - x) \mathbb{I}_{I_{t-1}+q_{t-1}-x \geq 0} | \mathcal{Z}_{t-1}) \\
 &= \mathbb{I}_{z_{t-1} > 0} f(-z_{t-1} + q_{t-1} - x) \mathbb{I}_{-z_{t-1}+q_{t-1}-x \geq 0} \\
 (60) \quad &+ \mathbb{I}_{z_{t-1}=0} \int_{(x-q_{t-1})^+}^\infty f(\eta + q_{t-1} - x) \pi_{t-1}(\eta) d\eta.
 \end{aligned}$$

By inserting (60) into (59), we get

$$\begin{aligned}
 \mathbb{E}(\varphi(I_t) \mathbb{I}_{z_t=0} | \mathcal{Z}_{t-1}) &= \mathbb{I}_{z_{t-1} > 0} \int_0^\infty \varphi(x) f(-z_{t-1} + q_{t-1} - x) \mathbb{I}_{-z_{t-1}+q_{t-1}-x \geq 0} dx \\
 &+ \mathbb{I}_{z_{t-1}=0} \int_0^\infty \varphi(x) \left( \int_{(x-q_{t-1})^+}^\infty f(\eta + q_{t-1} - x) \pi_{t-1}(\eta) d\eta \right) dx.
 \end{aligned}$$

Similarly, we can express explicitly the denominator of the right-hand side of (4) as

$$\begin{aligned}
 \mathbb{P}(I_t \geq 0 | \mathcal{Z}_{t-1}) &= \mathbb{E}(\mathbb{I}_{I_{t-1}+q_{t-1}-D_{t-1} \geq 0} | \mathcal{Z}_{t-1}) \\
 &= \mathbb{E}\{\mathbb{E}(\mathbb{I}_{I_{t-1}+q_{t-1}-D_{t-1} \geq 0} | \mathcal{Z}_{t-1}, I_{t-1}) | \mathcal{Z}_{t-1}\} \\
 &= \mathbb{E}\{F(I_{t-1} + q_{t-1}) | \mathcal{Z}_{t-1}\} \\
 &= \mathbb{I}_{z_{t-1} > 0} F(-z_{t-1} + q_{t-1}) + \mathbb{I}_{z_{t-1}=0} \int_0^\infty F(y + q_{t-1}) \pi_{t-1}(y) dy.
 \end{aligned}$$

Because we have already obtained explicit expressions of the numerator and denominator of the right-hand side of (4), we can express the fraction, say the right-hand side of (4), as follows:

$$\begin{aligned} \frac{\mathbb{E}(\varphi(I_t)\mathbb{I}_{Z_t=0}|\mathcal{Z}_{t-1})}{\mathbb{P}(I_t \geq 0|\mathcal{Z}_{t-1})} &= \mathbb{I}_{z_{t-1}>0} \frac{\int_0^\infty \varphi(x)f(-z_{t-1}+q_{t-1}-x)\mathbb{I}_{-z_{t-1}+q_{t-1}-x \geq 0}dx}{F(-z_{t-1}+q_{t-1})} \\ &\quad + \mathbb{I}_{z_{t-1}=0} \frac{\int_0^\infty \varphi(x) \left( \int_{(x-q_{t-1})^+}^\infty f(\eta+q_{t-1}-x)\pi_{t-1}(\eta)d\eta \right) dx}{\int_0^\infty F(y+q_{t-1})\pi_{t-1}(y)dy}. \end{aligned}$$

Then, we have

$$\begin{aligned} \int_0^\infty \varphi(\eta)\pi_t(\eta)d\eta &= \mathbb{I}_{z_{t-1}>0} \frac{\int_0^\infty \varphi(x)f(-z_{t-1}+q_{t-1}-x)\mathbb{I}_{-z_{t-1}+q_{t-1}-x \geq 0}dx}{F(-z_{t-1}+q_{t-1})} \\ &\quad + \mathbb{I}_{z_{t-1}=0} \frac{\int_0^\infty \varphi(x) \left( \int_{(x-q_{t-1})^+}^\infty f(\eta+q_{t-1}-x)\pi_{t-1}(\eta)d\eta \right) dx}{\int_0^\infty F(y+q_{t-1})\pi_{t-1}(y)dy}. \end{aligned}$$

Since the above equation is satisfied for all test functions  $\varphi(x)$ , we have

$$\begin{aligned} \pi_t(x) &= \mathbb{I}_{z_{t-1}>0} \left\{ \frac{f(-z_{t-1}+q_{t-1}-x)\mathbb{I}_{-z_{t-1}+q_{t-1}-x \geq 0}}{F(q_{t-1}-z_{t-1})} \right\} \\ &\quad + \mathbb{I}_{z_{t-1}=0} \left\{ \frac{\int_{(x-q_{t-1})^+}^\infty f(y+q_{t-1}-x)\pi_{t-1}(y)dy}{\int_0^\infty F(y+q_{t-1})\pi_{t-1}(y)dy} \right\}. \end{aligned}$$

This equality is exactly (6), which completes the proof.  $\square$

**5.4. Proof of Lemma 2.** By changing the order of integration, we obtain

$$\begin{aligned} \int_0^\infty |\varrho(q)p(x)|dx &\leq \int_0^\infty \int_{(x-q)^+}^\infty f(y+q-x)|p(y)|dydx \\ &= \int_0^\infty \int_0^{y+q} f(y+q-x)|p(y)|dx dy \\ &= \int_0^\infty |p(y)| \int_0^{y+q} f(y+q-x)dx dy \\ (61) \qquad &\leq \int_0^\infty |p(y)|F(y+q)dy. \end{aligned}$$

Using similar operations, we see that

$$\begin{aligned} \int_0^\infty x|\varrho(q)p(x)|dx &\leq \int_0^\infty \int_{(x-q)^+}^\infty xf(y+q-x)|p(y)|dydx \\ &= \int_0^\infty \int_0^{y+q} xf(y+q-x)|p(y)|dx dy \\ &= \int_0^\infty |p(y)| \int_0^{y+q} (y+q-z)f(z)dz dy \\ (62) \qquad &\leq \int_0^\infty (y+q)|p(y)|F(y+q)dy. \end{aligned}$$

Using inequalities (61) and (62), we get

$$\begin{aligned}
 \|\varrho(q)p\| &= \int_0^\infty |\varrho(q)p(x)|dx + \int_0^\infty x|\varrho(q)p(x)|dx \\
 &\stackrel{(61,62)}{\leq} \int_0^\infty |p(y)|F(y+q)dy + \int_0^\infty |p(y)|(y+q)F(y+q)dy \\
 (63) \quad &\leq \|p\| + q \int_0^\infty |p(y)|dy.
 \end{aligned}$$

From (63) we can easily obtain  $\|\varrho(q)\|_{\mathcal{L}} \leq 1 + q$ .  $\square$

**5.5. Proof of Lemma 3.** Since  $c(y, q) \geq 0$ ,  $p_n \in \mathcal{H}^+$ , and  $p \in \mathcal{H}^+$ , we have

$$\begin{aligned}
 &\lim_{n \rightarrow \infty} \left| \int_0^\infty c(y, q_n)p_n(y)dy - \int_0^\infty c(y, q)p(y)dy \right| \\
 &= \lim_{n \rightarrow \infty} \left| \int_0^\infty c(y, q_n)p_n(y)dy - \int_0^\infty c(y, q)p_n(y)dy \right. \\
 &\quad \left. + \int_0^\infty c(y, q)p_n(y)dy - \int_0^\infty c(y, q)p(y)dy \right| \\
 &\leq \lim_{n \rightarrow \infty} \left| \int_0^\infty c(y, q_n)p_n(y)dy - \int_0^\infty c(y, q)p_n(y)dy \right| \\
 &\quad + \lim_{n \rightarrow \infty} \left| \int_0^\infty c(y, q)p_n(y)dy - \int_0^\infty c(y, q)p(y)dy \right| \\
 &= \lim_{n \rightarrow \infty} \left| \int_0^\infty [c(y, q_n) - c(y, q)]p_n(y)dy \right| + \lim_{n \rightarrow \infty} \int_0^\infty c(y, q)|p_n(y) - p(y)|dy \\
 &\leq \lim_{n \rightarrow \infty} \int_0^\infty |c(y, q_n) - c(y, q)|p_n(y)dy \\
 &\quad + \lim_{n \rightarrow \infty} \int_0^\infty [c_0 + c_1q + hy]|p_n(y) - p(y)|dy = 0,
 \end{aligned}$$

where both limits approach zero, respectively, on account of continuity of  $c(y, q)$  in  $q$  and (10).  $\square$

**5.6. Proof of Lemma 4.** The function  $v^0$  is an upper bound for the discounted cost of backordering. Indeed, let  $\tilde{q}_0 = \{0, 0, \dots\} \in \Gamma$  and  $\tilde{q} \in \Gamma$  be an arbitrary policy.

Then,

$$\begin{aligned}
 J(\zeta, \pi, \tilde{q}) &\leq J(\zeta, \pi, \tilde{q}_0) = \sum_{t=1}^{\infty} \alpha^{t-1} \mathbb{E}\{\mathbb{I}_{z_t > 0} c(-z_t, q_t)\} \leq \sum_{t=1}^{\infty} \alpha^{t-1} \mathbb{E}\{\mathbb{I}_{z_t > 0} (c_0 + sz_t)\} \\
 &\leq (c_0 + s\zeta) + \alpha \left( c_0 + s(\zeta + E(D)) \right) + \alpha^2 \left( c_0 + s(\zeta + 2E(D)) \right) \\
 &\quad + \alpha^3 \left( c_0 + s(\zeta + 3E(D)) \right) + \cdots \\
 &= c_0 + s\zeta + sE(D) + (c_0 + s\zeta)(\alpha + \alpha^2 + \alpha^3 + \cdots) \\
 &\quad + sE(D)(2\alpha + 3\alpha^2 + 4\alpha^3 + 5\alpha^4 + \cdots) \\
 &= \frac{\alpha s E(D)}{(1-\alpha)^2} + \frac{c_0 + s\zeta}{1-\alpha} = v^0(\zeta), \quad \zeta \in \mathbb{R}^+,
 \end{aligned}$$

where we use the upper bound of the cost in (14) to obtain the second inequality.

For  $\phi_1 \leq v^0$  and  $\phi_2 \leq Y^0$ , since  $Y^0(0) = 0$ ,

$$\begin{aligned}
 T^{(1)}(\phi_1, \phi_2) &= \inf_q T_q^{(1)}(\phi_1, \phi_2) \leq c(-\zeta, 0) + \alpha \int_0^\infty v^0(\eta) f(\eta - \zeta) d\eta + \alpha Y^0(0) \\
 &= c(-\zeta, 0) + \alpha \int_0^\infty v^0(\eta) f(\eta - \zeta) d\eta \\
 &= c_0 + s\zeta + \alpha \int_0^\infty \left[ \frac{\alpha s E(D)}{(1-\alpha)^2} + \frac{c_0 + s\eta}{1-\alpha} \right] f(\eta - \zeta) d\eta.
 \end{aligned}$$

We know that  $\int_0^\infty \eta f(\eta - \zeta) d\eta \leq E(D) + \zeta$ , for  $\zeta \in \mathbb{R}^+$ . Then,

$$\begin{aligned}
 T^{(1)}(\phi_1, \phi_2) &\leq c_0 + s\zeta + \frac{\alpha^2 s E(D)}{(1-\alpha)^2} + \frac{\alpha c_0}{1-\alpha} + \alpha(E(D) + \zeta) \frac{s}{1-\alpha} \\
 (64) \qquad &= \frac{\alpha^2 s E(D)}{(1-\alpha)^2} + \frac{c_0 + s\zeta}{1-\alpha} \leq v^0(\zeta).
 \end{aligned}$$

Thus,  $T^{(1)}(\phi_1, \phi_2) \in \mathbb{L}^+$ .

For  $\phi_1 \leq v^0$  and  $\phi_2 \leq Y^0$ ,

$$\begin{aligned}
 T^{(2)}(\phi_1, \phi_2) &= \inf_q T_q^{(2)}(\phi_1, \phi_2) \leq T_0^{(2)}(\phi_1, \phi_2) \leq T_0^{(2)}(v^0, Y^0) \\
 &= \int_0^\infty c(y, 0)p(y)dy + \alpha \int_0^\infty p(y) \int_0^\infty v^0(\eta)f(y+\eta)d\eta dy + \alpha Y^0(\varrho(0)p) \\
 &\leq \int_0^\infty (c_0 + hy)p(y)dy \\
 &\quad + \alpha \int_0^\infty p(y) \int_0^\infty \left[ \frac{\alpha s E(D)}{(1-\alpha)^2} + \frac{c_0 + s\eta}{1-\alpha} \right] f(y+\eta)d\eta dy \\
 &\quad + \alpha \frac{a_0}{1-\alpha} \|\varrho(0)p\| \\
 &\leq \left\{ c_0 + \frac{\alpha s E(D)}{(1-\alpha)^2} + \frac{\alpha c_0 + s E(D)}{1-\alpha} \right\} \int_0^\infty p(y)dy \\
 &\quad + h \int_0^\infty yp(y)dy + \alpha \frac{a_0}{1-\alpha} \|p\| \\
 (65) \quad &\leq a_0 \|p\| + \alpha \frac{a_0}{1-\alpha} \|p\| = \frac{a_0}{1-\alpha} \|p\| = Y^0(p).
 \end{aligned}$$

Obviously,  $T^{(2)}(\phi_1, \phi_2) \in \mathcal{B}^+$ . Therefore, the operator  $T$  maps  $\mathcal{G}$  into itself. For  $\phi_1 \leq v^0$  and  $\phi_2 \leq Y^0$ ,  $T^{(1)}(\phi_1, \phi_2) \leq v^0(\zeta)$ ,  $T^{(2)}(\phi_1, \phi_2) \leq Y^0(p)$ .  $\square$

**5.7. Proof of Theorem 2.** From Remark 3(a), part (b) of the theorem is a consequence of part (a).

To prove part (a), observe that, since  $Y_1 = v_1 = 0$ , we have  $v_1 \leq v^0$  and  $Y_1 \leq Y^0$ . Since  $\{v_n\}$  and  $\{Y_n\}$  are increasing sequences of continuous functions such that  $v_n \leq v^0$  and  $Y_n \leq Y^0$  for all  $n \geq 1$ , there are l.s.c. functions  $\bar{v} \leq v^0$  and  $\bar{Y} \leq Y^0$  such that

$$v_n \nearrow \bar{v} \quad \text{and} \quad Y_n \nearrow \bar{Y} \quad \text{as } n \rightarrow \infty.$$

Hence, we have that, for each  $(\zeta, p) \in \mathcal{R}^+ \times \mathcal{H}^+$ ,

$$(66) \quad Z_n(\zeta, p) \nearrow \bar{Z}(\zeta, p)$$

and  $\bar{Z} \in \mathcal{G}$ . Now, using the fact that the operators  $T^{(1)}$  and  $T^{(2)}$  are monotone, we have  $Z_n = TZ_{n-1} \leq T\bar{Z}$ . Thus, from (66),

$$(67) \quad \bar{Z}(\zeta, p) \leq T\bar{Z}(\zeta, p), \quad (\zeta, p) \in \mathcal{R}^+ \times \mathcal{H}^+.$$

To obtain the reverse inequality, let  $\bar{q}_n = \bar{g}_{Z_n}(\zeta, p)$  be such that  $Z_{n+1}(\zeta, p) = TZ_n(\zeta, p) = T_{\bar{q}_n}Z_n(\zeta, p)$ . Thus, observe that for any  $N$ ,

$$Z_{n+1}(\zeta, p) \geq T_{\bar{q}_n}Z_N(\zeta, p) \quad \forall n \geq N.$$

Then from (66),

$$(68) \quad \bar{Z}(\zeta, p) \geq T_{\bar{q}_n}Z_N(\zeta, p).$$

In addition, we can extract a subsequence  $\{\bar{q}_{n_k}\}$  of  $\{\bar{q}_n\}$  such that  $\bar{q}_{n_k} \rightarrow \bar{q} \in \mathcal{Q}^*(\zeta, p)$  as  $k \rightarrow \infty$ . Then, by the continuity of the function  $q \rightarrow T_qZ_N(\zeta, p)$ , we have that  $T_{\bar{q}_{n_k}}Z_N(\zeta, p) \rightarrow T_{\bar{q}}Z_N(\zeta, p)$  as  $k \rightarrow \infty$ . Therefore, from (68), we have  $\bar{Z}(\zeta, p) \geq T_{\bar{q}}Z_N(\zeta, p)$ . Letting  $N \rightarrow \infty$ , we obtain  $\bar{Z}(\zeta, p) \geq T_{\bar{q}}\bar{Z}(\zeta, p) \geq T\bar{Z}(\zeta, p)$ , which, combined with (67), yields  $\bar{Z}(\zeta, p) = T\bar{Z}(\zeta, p)$  for each  $(\zeta, p) \in \mathcal{R}^+ \times \mathcal{H}^+$ .  $\square$

**5.8. Proof of Theorem 3.** (a) Because  $V_1 = 0$ , we have  $V_1 \leq V$ . Then, by applying induction arguments, we can prove that  $V_n(\zeta, \pi) \leq V(\zeta, \pi)$  for all  $n$  and  $(\zeta, \pi) \in \mathfrak{R}^+ \times \Pi$ . Therefore (see (37)), since  $V_n \nearrow \bar{V}$  as  $n \rightarrow \infty$ ,

$$(69) \quad \bar{V}(\zeta, \pi) \leq V(\zeta, \pi) \quad \forall (\zeta, \pi) \in \mathfrak{R}^+ \times \Pi.$$

To prove the reverse inequality, let  $\hat{q}_t = \hat{g}(z_t, \pi_t)$  be the map satisfying (see (38) and (39))

$$(70) \quad \bar{V}(z_t, \pi_t) = \mathbb{E} [\mathbb{I}_{z_t > 0} c(-z_t, \hat{q}_t) + \mathbb{I}_{z_t = 0} \langle c(I_t, \hat{q}_t), \pi_t \rangle | \mathcal{Z}_t] + \alpha \mathbb{E} [\bar{V}(z_{t+1}, \pi_{t+1}) | \mathcal{Z}_t].$$

Hence,

$$\begin{aligned} & \mathbb{E} [\alpha^{t-1} \bar{V}(z_t, \pi_t)] - \mathbb{E} [\alpha^t \bar{V}(z_{t+1}, \pi_{t+1})] \\ &= \alpha^{t-1} \mathbb{E} [\mathbb{I}_{z_t > 0} c(-z_t, \hat{q}_t) + \mathbb{I}_{z_t = 0} \langle c(I_t, \hat{q}_t), \pi_t \rangle]. \end{aligned}$$

Summing up for  $t = 1, 2, \dots, M$  yields

$$\begin{aligned} \bar{V}(\zeta, \pi) &= \sum_{t=1}^M \alpha^{t-1} \mathbb{E} [\mathbb{I}_{z_t > 0} c(-z_t, \hat{q}_t) + \mathbb{I}_{z_t = 0} \langle c(I_t, \hat{q}_t), \pi_t \rangle] + \alpha^M \mathbb{E} [\bar{V}(z_{M+1}, \pi_{M+1})] \\ (71) \quad &\geq \sum_{t=1}^M \alpha^{t-1} \mathbb{E} [\mathbb{I}_{z_t > 0} c(-z_t, \hat{q}_t) + \mathbb{I}_{z_t = 0} \langle c(I_t, \hat{q}_t), \pi_t \rangle]. \end{aligned}$$

Letting  $M \rightarrow \infty$  and denoting  $\hat{q} = \{\hat{g}, \hat{g}, \dots\}$ , from (16) and (17), we get

$$(72) \quad \bar{V}(\zeta, \pi) \geq J(\zeta, \pi, \hat{q}) \geq V(\zeta, \pi) \quad \forall (\zeta, \pi) \in \mathfrak{R}^+ \times \Pi,$$

which, from (69), proves part (a).

(b) Let  $\tilde{\phi} \in \tilde{\mathcal{G}}$  be an arbitrary function such that  $\tilde{\phi}(\zeta, \pi) = \tilde{T}\tilde{\phi}(\zeta, \pi)$ . That is,  $\tilde{\phi}$  is of the form  $\tilde{\phi}(\zeta, \pi) = \mathbb{I}_{\zeta > 0} \tilde{\phi}_1(\zeta) + \mathbb{I}_{\zeta = 0} \tilde{\phi}_2(\pi)$ , and  $\tilde{\phi}_1$  and  $\tilde{\phi}_2$  satisfy the system (22)–(23). Then, applying the arguments in the proof of part (a) with  $\tilde{\phi}$  instead of  $\bar{V}$  (see (72)), we conclude that  $\tilde{\phi} \geq V$ . Since  $V = \bar{V}$ , it follows that  $\bar{V} \leq \tilde{\phi}$ . Hence,  $\bar{V}$  is minimal in  $\tilde{\mathcal{G}}$ . In addition, the corresponding unnormalized value function  $\bar{Z}$  is minimal in  $\mathcal{G}$ .

(c) Let  $\tilde{q}^* = \{g^*, g^*, \dots\} \in \Gamma$  be the policy determined by the map  $g^* : \mathfrak{R}^+ \times \Pi \rightarrow \mathcal{Q}$ . By denoting  $q_t^* = g^*(z_t, \pi_t)$  (see (38) and (39)), we write

$$V(z_t, \pi_t) = \mathbb{E} [\mathbb{I}_{z_t > 0} c(-z_t, q_t^*) + \mathbb{I}_{z_t = 0} \langle c(I_t, q_t^*), \pi_t \rangle | \mathcal{Z}_t] + \alpha \mathbb{E} [V(z_{t+1}, \pi_{t+1}) | \mathcal{Z}_t].$$

Then the first inequality in (72) implies

$$V(\zeta, \pi) \geq J(\zeta, \pi, \tilde{q}^*) \quad \forall (\zeta, \pi) \in \mathfrak{R}^+ \times \Pi.$$

Therefore, from (17),  $\tilde{q}^*$  is optimal.  $\square$



**5.9. Proof of Theorem 4.** (a) Let  $\tilde{q} \in \Gamma$  be an arbitrary policy and  $\{q_1, q_2, \dots\}$  be the decisions corresponding to application of  $\tilde{q}$ . Then, from (46), we can proceed as in (70) and (71) to obtain

(73)

$$\underline{V}(\zeta, \pi) = \sum_{t=1}^M \alpha^{t-1} E [\mathbb{I}_{z_t > 0} c(-z_t, \hat{q}_t) + \mathbb{I}_{z_t = 0} \langle c(I_t, \hat{q}_t), \pi_t \rangle] + \alpha^M E [\underline{V}(z_{M+1}, \pi_{M+1})].$$

Now, as  $\underline{V} \in \tilde{\mathcal{G}}$ , it follows that

$$(74) \quad \underline{V}(z_{M+1}, \pi_{M+1}) \leq \mathbb{I}_{z_{M+1} > 0} v^0(z_{M+1}) + \mathbb{I}_{z_{M+1} = 0} Y^0(\pi_{M+1}).$$

On the other hand, observe that from (2)  $z_t > 0$  if and only if  $I_t < 0$ , for each  $t \geq 1$ . Then, if  $z_{M+1} > 0$ , we have  $z_{M+1} = -I_{M+1} \leq D_M - I_M$ . Hence, if  $z_t > 0$  for all  $t$ , iterating this inequality we have

$$-E[I_{M+1}] \leq (M-1)E(D) - \zeta,$$

where  $I_1 = \zeta$ . Therefore, from (30),

$$\begin{aligned} & \alpha^N E [\mathbb{I}_{z_{M+1} > 0} v^0(z_{M+1})] \\ &= \alpha^M E \left[ \mathbb{I}_{z_{M+1} > 0} \left[ \frac{\alpha s E(D)}{(1-\alpha)^2} + \frac{c_0 + s z_{M+1}}{1-\alpha} \right] \right] \\ (75) \quad & \leq \alpha^M E \left[ \frac{\alpha s E(D)}{(1-\alpha)^2} + \frac{c_0 + s(M-1)E(D) - s\zeta}{1-\alpha} \right] \rightarrow 0 \quad \text{as } M \rightarrow \infty. \end{aligned}$$

On the other hand, from (31),

$$\begin{aligned} Y^0(\pi_{M+1}) &= \frac{a_0}{1-\alpha} \|\pi_{M+1}\| = \frac{a_0}{1-\alpha} \left( 1 + \int_0^\infty x \pi_{M+1}(x) dx \right) \\ &= \frac{a_0}{1-\alpha} (1 + E[I_{M+1} | \mathcal{Z}_M]). \end{aligned}$$

Then,

$$(76) \quad E[Y^0(\pi_{M+1})] = \frac{a_0}{1-\alpha} + \frac{a_0}{1-\alpha} E[I_{M+1}].$$

If  $z_{M+1} = 0$ , we have, from (51),

$$\begin{aligned} I_{M+1} &= I_M + q_M - D_M \leq I_M + q_M \leq I_M + \frac{a_0}{c(1-\alpha)} \|\pi_M\| \\ &= I_M + \frac{a_0}{c(1-\alpha)} + \frac{a_0}{c(1-\alpha)} E[I_M | \mathcal{Z}_{M-1}]. \end{aligned}$$

Hence,

$$E[I_{M+1}] \leq \left( 1 + \frac{a_0}{c(1-\alpha)} \right) E[I_M] + \frac{a_0}{c(1-\alpha)}.$$

Therefore, if  $z_t = 0$  for all  $t$ , iteration of this inequality yields

$$E[I_{M+1}] \leq \left(1 + \frac{a_0}{c(1-\alpha)}\right)^M (E(I_1) + 1) - 1.$$

Thus, from (76), as  $M \rightarrow \infty$ ,

$$(77) \quad \alpha^M E[\mathbb{I}_{z_{M+1}=0} Y^0(\pi_{M+1})] \leq \alpha^M \left[ \frac{a_0}{1-\alpha} (1 + E(I_1)) \left(1 + \frac{a_0}{c(1-\alpha)}\right)^M \right] \rightarrow 0.$$

Combining (74), (75), and (77), we have

$$\alpha^M [\underline{V}(z_{M+1}, \pi_{M+1})] \rightarrow 0 \text{ as } M \rightarrow \infty.$$

Letting  $M \rightarrow \infty$  in (73), we get  $\underline{V}(\zeta, \pi) \leq J(\zeta, \pi, \tilde{q})$ , and, as  $\tilde{q} \in \Gamma$  was arbitrary, we have

$$(78) \quad \underline{V}(\zeta, \pi) \leq V(\zeta, \pi).$$

This combined with (47) yields  $V(\zeta, \pi) = \underline{V}(\zeta, \pi)$ .

Finally, let  $\tilde{\phi} \in \tilde{\mathcal{G}}$  be an arbitrary function such that  $\tilde{\phi}(\zeta, \pi) = \tilde{T}\tilde{\phi}(\zeta, \pi)$ . Then, by using  $\tilde{\phi}$  instead of  $\underline{V}$ , we obtain  $V \geq \tilde{\phi}$  (see (78)). Since  $\underline{V} = V$ ,  $\underline{V} \geq \tilde{\phi}$ . Then,  $\underline{V}$  is maximal in  $\tilde{\mathcal{G}}$ . Similarly,  $\underline{Z}$  is maximal in  $\mathcal{G}$ . Furthermore, in view of the results in Theorem 3, we can conclude that the solutions of (24) and (28) are unique.

(b) According to part (a), we see that the value iteration functions  $Z^n$  converge decreasingly to the value function  $Z$ ; that is,

$$(79) \quad Z^n \searrow Z \text{ as } n \rightarrow \infty.$$

On the other hand, since  $v^1(\zeta) = v^0(\zeta)$  and  $Y^1(p) = Y^0(p)$  are continuous functions, we have from Lemma 7 that  $Z^n$  is continuous on  $\mathfrak{R}^+ \times \mathcal{H}^+$  for all  $n \geq 1$ . Therefore, from (79), we can ensure that the value function  $Z$  is u.s.c. Hence, from the lower semicontinuity of  $Z$  given in Theorems 2 and 3, we conclude that  $Z$  is continuous on  $\mathfrak{R}^+ \times \mathcal{H}^+$ . This result also yields the continuity of  $V$  on  $\mathfrak{R}^+ \times \Pi$ . This completes the proof.  $\square$

## REFERENCES

- [1] K. J. ARROW, S. KARLIN, AND H. SCARF, *Studies in the Mathematical Theory of Inventory and Production*, Stanford University Press, Stanford, CA, 1958.
- [2] A. BENSOUSSAN, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.
- [3] A. BENSOUSSAN, M. ÇAKANYILDIRIM, J. MINJÁREZ-SOSA, S. P. SETHI, AND R. SHI, *An Incomplete Information Inventory Model with Presence of Inventories or Backorders as Only Observations*, working paper, School of Management, University of Texas at Dallas, Richardson, TX, 2007.
- [4] A. BENSOUSSAN, M. ÇAKANYILDIRIM, AND S. P. SETHI, *Partially observed inventory systems: The case of zero-balance walk*, SIAM J. Control Optim., 46 (2007), pp. 176–209.
- [5] A. BENSOUSSAN, M. ÇAKANYILDIRIM, S. P. SETHI, AND R. SHI, *Computation of Approximate Optimal Policies in an Inventory Model with Rain Checks*, working paper, School of Management, University of Texas at Dallas, Richardson, TX, 2007.
- [6] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
- [7] D. BEYER, F. CHENG, S. P. SETHI, AND M. I. TAKSAR, *Markovian Demand Inventory Models*, Springer, New York, 2008.

- [8] D. BEYER AND S. P. SETHI, *Average cost optimality in inventory models with Markovian demands*, J. Optim. Theory Appl., 92 (1997), pp. 497–526.
- [9] M. L. FISHER, A. RAMAN, AND A. S. MCCLELLAND, *Rocket science retailing is almost here: Are you ready?*, Harvard Business Rev., 78 (2000), pp. 115–124.
- [10] C. J. GEYER, *On the convergence of Monte Carlo maximum likelihood calculations*, J. Roy. Statist. Soc. Ser. B, 56 (1994), pp. 261–274.
- [11] H. J. KUSHNER, *Dynamic equations for nonlinear filtering*, J. Differential Equations, 3 (1967), pp. 179–190.
- [12] A. RAMAN, N. DEHORATIUS, AND Z. TON, *Execution: The missing link in retail operations*, California Management Rev., 43 (2001), pp. 136–152.
- [13] M. ZAKAI, *On the optimal filtering of diffusion processes*, Z. Wahrsch. Verw. Gebiete, 11 (1969), pp. 230–243.

# STABILITY OF SOLUTIONS FOR SOME CLASSES OF NONLINEAR DAMPED WAVE EQUATIONS\*

GENNI FRAGNELLI<sup>†</sup> AND DIMITRI MUGNAI<sup>‡</sup>

**Abstract.** We consider two classes of semilinear wave equations with nonnegative damping which may be of type “on–off” or integrally positive. In both cases we give a sufficient condition for the asymptotic stability of the solutions. In the case of integrally positive damping we show that such a condition is also necessary.

**Key words.** damped nonlinear wave equations, integrally positive damping, on–off damping

**AMS subject classifications.** 35L70, 93D20, 35B35

**DOI.** 10.1137/070689735

**1. Introduction.** We are concerned with some classes of nonlinear abstract damped wave equations, whose prototype is the usual wave equation in a bounded domain  $\Omega \subset \mathbb{R}^N$ ,  $N \geq 1$ ,

$$(W) \quad u_{tt} = \Delta u - h(t)u_t + f(u),$$

and its associated Cauchy problem,

$$(1.1) \quad \begin{cases} u_{tt} = \Delta u - h(t)u_t + f(u) & \text{in } (0, +\infty) \times \Omega, \\ u(t, x) = 0 & \text{in } (0, +\infty) \times \partial\Omega, \\ u(0, x) = u_0(x), \quad u_t(0, x) = u_1(x) & x \in \Omega, \end{cases}$$

though we can handle equations in a more general Banach setting of the form

$$(E) \quad u'' + B(t)u' + Au = f(u),$$

with  $B$  and  $A$  suitably given (see section 2 for the precise setting).

This problem has been already investigated by many authors in the case of ordinary differential equations or systems of ordinary differential equations (i.e., when  $u$  depends only on  $t$ ) when  $f$  is linear (for example, see [5], [8]) and also when  $f$  is nonlinear (see [16], [18], [19]).

In the case of hyperbolic partial differential equations like (1.1), the problem has been studied when  $f$  is linear (see [4], [12]), when  $f$  is nonlinear but with linear growth (see [22]), and when  $f$  is nonlinear with superlinear growth (see [14] in the case of constant damping; see [17] for more general cases).

Concerning the damping  $h$ , the following different assumptions are alternatively made: on–off [7], increasing [1], bounded (in many senses) [4], [20], integrally positive

---

\*Received by the editors April 27, 2007; accepted for publication (in revised form) May 5, 2008; published electronically September 17, 2008. This work was started while the authors were visiting Prof. László Hatvani at the Bolyai Institute of the University of Szeged. The first author acknowledges financial support from GNAMPA. The second author acknowledges financial support from CNR in the framework of the bilateral project CNR MTA 132.07.

<http://www.siam.org/journals/sicon/47-5/68973.html>

<sup>†</sup>Dipartimento di Ingegneria dell'Informazione, Università di Siena, Via Roma 56, 53100 Siena, Italy (fragnelli@dii.unisi.it).

<sup>‡</sup>Dipartimento di Matematica e Informatica, Università di Perugia, Via Vanvitelli 1, 06123 Perugia, Italy (mugnai@dipmat.unipg.it). This author's research was supported by the M.I.U.R. project *Metodi Variazionali ed Equazioni Differenziali Nonlineari*.

[22], etc. In particular, on–off dampers are suitable for describing a wide variety of communication network models (circuits which can be switched on or off), as well as systems where a control depending on time is necessary.

Very interesting results in the special case of  $f \equiv 0$  and damping of type on–off can be found in [7], and also when the term  $a(t)u_t$  in (1.1) is replaced by  $a(t)g(u_t)$ , where  $g$  is a nonlinear function with linear growth (see also [13]). For the case  $f \equiv 0$  we also mention a logarithmic decay estimate proved in [3] when the term  $a(t)u_t$  in (1.1) is replaced by  $(1+t)^\theta a(x)g(u_t)$ , with  $a$  bounded and strictly positive on a subdomain of  $\Omega$  and  $g$  possibly having superlinear growth at infinity.

In [22] Zhang shows that, if  $f$  has linear growth and  $h$  is integrally positive (see Definition 3.1 below), then any solution  $u(u_0, u_1)$  of (2.1) converges to 0 in the norm  $\|\nabla u\|_{L^2} + \|u_t\|_{L^2}$  if and only if

$$(1.2) \quad \int_0^\infty e^{-H(t)} \int_0^t e^{H(s)} ds dt = +\infty,$$

where  $H(t) = \int_0^t h(s)ds$ . Actually, the proof contains some gaps, in particular, in proving that the  $L^2$ -norm of  $u'$  converges to 0 as  $t \rightarrow \infty$ . However, we recover that result in section 3.

In this paper we show that when  $h$  is integrally positive, condition (1.2) is sufficient to prove global stability for problem (1.1), and also when  $f$  is superlinear and satisfies a sign condition. More precisely, in Theorem 3.1 we prove that condition (1.2) is sufficient under the assumption

$$uf(u) \leq 0 \quad \text{for every } u \in \mathbb{R},$$

which was already assumed, for example, in [17] and [20]. Note that such a condition is trivially verified when  $f(u) = -|u|^{p-1}u$ ,  $p \geq 1$ , which is our prototype, and which was studied, for example, in [14], where a global existence result is proved for  $h = \text{constant}$ . Under a natural nonsupercritical growth condition on  $f$ , we also show that (1.2) is also a necessary condition for global stability to hold (see Theorem 3.2). We remark that the requirement  $uf(u) \leq 0$  is a bit stronger than  $sf(s) < \lambda_1 s^2$ ,  $s \neq 0$ , which is essentially assumed in [22], and also in [9] where the damping is concentrated in a subset of  $\omega$  and Neumann-type conditions are assumed on the boundary (as in [2])—as usual, here  $\lambda_1$  denotes the first eigenvalue of  $-\Delta$  on  $H_0^1(\Omega)$ .

For damping of type on–off, in [7] the following case is considered: Let  $(a_n, b_n)_n$  be a sequence of open disjoint intervals of  $(0, \infty)$  such that  $a_n \rightarrow \infty$  and suppose there exists  $M_n \geq m_n > 0$  such that

$$(1.3) \quad m_n \leq h(t) \leq M_n \quad \forall t \in (a_n, b_n).$$

If  $f \equiv 0$ , then any solution  $u(u_0, u_1)$  of (2.1) converges to 0 in the norm  $\|\nabla u\|_{L^2} + \|u_t\|_{L^2}$  if

$$(1.4) \quad \sum_{n=1}^\infty m_n(b_n - a_n) \min \left( (b_n - a_n)^2, \frac{1}{1 + m_n M_n} \right) = \infty.$$

In this result the fact that  $f \equiv 0$  is essential in the proof of stability. In the nonlinear case under consideration, with the assumption

$$sf(s) - \int_0^s f(\sigma) d\sigma \leq 0 \quad \forall s \in \mathbb{R},$$

we show that (1.4), which was essentially already introduced in [16] for systems of ordinary differential equations, is again sufficient for the stability (see Theorem 4.2). As for the case  $f \equiv 0$  in [7], we still don't know if (1.4) is also necessary for stability to hold.

However, in the case of integrally positive damping, we give a complete characterization of stability for signed nonsupercritical nonlinearities.

**2. The abstract setting.** We will use an abstract setting which is a bit less general than the one in [7], but more natural for our purposes. Let us consider a second order evolution problem of the form

$$(2.1) \quad \begin{cases} u'' + B(t)u' + Au = f(u), & t > 0, \\ u(0) = u_0 \in V, \ u'(0) = u_1 \in H. \end{cases}$$

Here  $H$  denotes a real Hilbert space with scalar product  $\langle \cdot, \cdot \rangle_H$  and norm  $\| \cdot \|_H$ ,  $A : D(A) \subset H \rightarrow H$  is a linear self-adjoint coercive operator on  $H$  with dense domain, and  $V = D(A^{1/2})$  with norm  $\|v\|_V = \|A^{1/2}v\|_H$  is such that

$$(2.2) \quad \begin{aligned} &V \hookrightarrow H \equiv H' \hookleftarrow V' \quad \text{with dense embeddings, and} \\ &\exists \lambda_1 > 0 : \|v\|_H^2 \leq \frac{1}{\lambda_1} \|v\|_V^2 = \frac{1}{\lambda_1} \|A^{1/2}v\|_H^2 \quad \text{for any } v \in V. \end{aligned}$$

Concerning the time-dependent operator  $B$ , in section 3 we assume it is actually a nonnegative function which can be 0 in a set of measure 0 (see Definition 3.1 below), while for the results of section 4 we let it be a more particular “positive” nonlinear operator (see below for the precise assumptions), for which we assume that  $B \in L^\infty(0, \infty; \text{Lip}(H, H'))$ .

Finally, on the nonlinearity  $f$  we assume alternatively

$$(2.3) \quad sf(s) \leq 0 \quad \forall s \in \mathbb{R},$$

which implies

$$F(s) := \int_0^s f(\sigma) d\sigma \leq 0 \quad \forall s \in \mathbb{R}$$

or

$$(2.4) \quad sf(s) - F(s) \leq 0 \quad \forall s \in \mathbb{R}.$$

*Remark 2.1.* In both cases,  $f(u) = -|u|^{p-1}u$ ,  $p \geq 1$ , is the prototype function.

We also remark that the sign assumptions on  $f$  look quite reasonable and hard to relax. Indeed, it is well known that solutions of  $u_{tt} + a(x, t)u_t - \Delta u = |u|^{p-1}u$  in  $\Omega$ ,  $a(x, t) \geq 0$  and  $p > 1$  might blow up in finite time (see, for example, [10] or [11]).

By *solution* of (2.1), we mean a function  $u$  such that for any  $T > 0$  there holds  $u \in L^2(0, T; V) \cap H^1(0, T; H) \cap H^2(0, T; V')$  with  $\langle Bu', u' \rangle_H \in L^2(0, T)$  and

$$Au \in L^2(0, T; V'), \quad Bu' \in L^2(0, T; V'), \quad f(u) \in L^2(0, T; H),$$

with  $u(0) = u_0$ ,  $u'(0) = u_1$ , and such that

$$u'' + Bu' + Au = f(u) \quad \text{in } L^2(0, T; V').$$

*Remark 2.2.* In the case of problem (1.1) the condition  $f(u) \in L^2(0, T; H)$  is automatically satisfied when  $H = L^2(\Omega)$ ,  $V = H_0^1(\Omega)$ , and  $f(u) = -|u|^{p-1}u$ ,  $p \geq 1$  if  $N = 1, 2$  or  $1 \leq p \leq N/(N-2)$  if  $N \geq 3$ .

For any solution  $u$  of problem (2.1) we denote by  $E_u$ , or simply by  $E$ , if there is no need to specify  $u$ , the energy associated to such a solution,

$$(2.5) \quad E(t) = \frac{1}{2} \|u'(t)\|_H^2 + \frac{1}{2} \|u(t)\|_V^2 - \mathcal{F}(u(t)),$$

where  $\mathcal{F}(u)$  is the real-valued functional such that

$$\mathcal{F}(0) = 0 \text{ and } \mathcal{F}'(u)(\phi) = \langle f(u), \phi \rangle_{V', V}.$$

Of course in the case of problem (1.1) we have  $H = L^2(\Omega)$ ,  $V = H_0^1(\Omega)$ , and

$$\mathcal{F}(u) = \int_{\Omega} F(u) dx.$$

The following result, proved in [7] when  $f \equiv 0$ , still holds true in the nonlinear case thanks to the assumption  $f(u) \in L^2(0, T; H)$ . The proof is an adaptation to the one given therein and is thus omitted.

LEMMA 2.1. *For any solution  $u$  of (2.1) we have that*

- $u \in C([0, T]; V) \cap C^1([0, T]; H)$ ;
- *the associated energy  $E_u$  is locally absolutely continuous on  $[0, \infty)$  and*

$$(2.6) \quad E'_u(t) = -\langle B(t)u'(t), u'(t) \rangle_H \quad \text{a.e. in } [0, \infty).$$

In our setting we will also need the following obvious corollary.

LEMMA 2.2. *If in addition  $\langle B(t)w, w \rangle_H \geq 0$  for a.e.  $t \geq 0$  and for every  $w \in H$ , then for any solution  $u$  of (2.1) we have that  $E_u$  is nonincreasing.*

*Remark 2.3.* In the case of problem (1.1) equality (2.6) reads

$$E'(t) = - \int_{\Omega} h(t) u_t^2 dx,$$

which is nonpositive if  $h \geq 0$  a.e. in  $[0, \infty)$ .

**3. The integrally positive case.** Let us start with the following definitions.

DEFINITION 3.1. *A function  $h : [0, +\infty) \rightarrow [0, +\infty)$  is said to be integrally positive if for every  $a > 0$  there exists  $\delta > 0$  such that*

$$\int_t^{t+a} h(s) ds \geq \delta \quad \forall t > 0.$$

*Remark 3.1.* We underline the fact that according to this definition, the function  $h$  may vanish somewhere, but not on any interval.

DEFINITION 3.2. *Solutions of (E) are said to be uniformly bounded in  $D(A) \times D(A^{1/2})$  if for any  $B_1 > 0$  there exists  $B_2 > 0$  such that*

$$\text{if } (u_0, u_1) \in D(A) \times D(A^{1/2}), \quad \|Au_0\|_H + \|A^{1/2}u_1\|_H \leq B_1, \text{ and } t_0 \in [0, \infty),$$

*and if  $u(t, t_0, u_0, u_1)$  denotes the solution of (E) such that  $u(t_0) = u_0$  and  $u'(t_0) = u_1$ , then  $\forall t \geq t_0$ ,*

- $f(u(t, t_0, u_0, u_1)) \in H$  and

$$\bullet \|Au(t, t_0, u_0, u_1)\|_H + \|A^{1/2}u'(t, t_0, u_0, u_1)\|_H + \|f(u(t, t_0, u_0, u_1))\|_H \leq B_2.$$

Let us remark that such a definition is a natural modification of the one introduced in [22], due to the presence of the requirement on  $\|f(u)\|_H$ . Moreover, we also underline the fact that  $B_2$  is independent of the initial time  $t_0$ .

For the following result we concentrate on (1.1), where  $A = -\Delta$ ,  $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$ ,  $D(A^{1/2}) = H_0^1(\Omega)$ , and  $H = L^2(\Omega)$ , and we recall that  $\|\Delta u\|_2$  is a norm on  $H^2(\Omega) \cap H_0^1(\Omega)$  which is equivalent to the usual one (for example, see [15]), where we have set  $\|\cdot\|_2 = \|\cdot\|_{L^2(\Omega)}$ .

Having in mind the prototype  $f(s) = -|s|^{p-1}s$ ,  $p \geq 1$ , we also make the following natural assumption:

If  $u$  solves (2.1),  $\forall T > 0 \exists C = C(u, T) > 0$  and  $q = q(u, T) > 0$  such that

$$(3.1) \quad \left| \int_{\Omega} F(u(t)) dx \right| \leq C \|A^{1/2}u(t)\|_2^q \quad \forall t \geq T.$$

Of course, in the model case  $f(s) = -|s|^{p-1}s$ ,  $p \geq 1$ , and  $p \leq N/(N-2)$  if  $N \geq 3$  described in Remark 2.2, we have also  $p+1 \leq (2N-2)/(N-2)$  if  $N \geq 3$ ; thus the Sobolev inequality can be applied, and we can always take  $q = p+1$  and  $C$  depending only on the measure of  $\Omega$  and the Sobolev constant.

Now we can state our first fundamental result.

**THEOREM 3.1.** *Assume (1.2), (2.3), and (3.1). If  $h$  is integrally positive and solutions of (W) are uniformly bounded in  $H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega)$ , then every solution  $u$  of (1.1) satisfies*

$$(3.2) \quad \|A^{1/2}u(t)\|_2 + \|u'(t)\|_2 \rightarrow 0 \quad \text{as } t \rightarrow +\infty.$$

*Proof.* Take  $(u_0, u_1) \in D(A) \times D(A^{1/2})$  and denote by  $u$  the associated solution of (1.1). By Remark 2.3 there exists  $E_{\infty} \geq 0$  such that

$$(3.3) \quad \begin{aligned} & \lim_{t \rightarrow +\infty} E(t) \\ &= \lim_{t \rightarrow +\infty} \left( \frac{1}{2} \|u'(t)\|_2^2 + \frac{1}{2} \|A^{1/2}u(t)\|_2^2 - \int_{\Omega} F(u) dx \right) = E_{\infty}. \end{aligned}$$

We want to show that  $E_{\infty} = 0$ , so let us assume by contradiction that  $E_{\infty} > 0$ .

By (2.3)  $F(s) = \int_0^s f(\tau) d\tau \leq 0$  for any  $s \in \mathbb{R}$ , so that by (3.3) there exists  $L \in [0, 2E_{\infty}]$  such that

$$(3.4) \quad \limsup_{t \rightarrow +\infty} \|u'(t)\|_2^2 = L.$$

Assume by contradiction that  $L > 0$ . We must distinguish several cases.

*Case 1.*  $\|u'(t)\|_2^2 \equiv L \forall t > 0$ . Then, by (2.6) and Remark 2.3 we get

$$(3.5) \quad \begin{aligned} 0 < E_{\infty} &= E(0) + \int_0^{\infty} E'(\tau) d\tau = E(0) - \int_0^{\infty} h(\tau) \|u'(\tau)\|_2^2 d\tau \\ &= E(0) - L \int_0^{\infty} h(\tau) d\tau. \end{aligned}$$

Since  $h$  is integrally positive, there exists  $\delta > 0$  such that

$$(3.6) \quad \int_n^{n+1} h(\tau) d\tau \geq \delta \quad \forall n \in \mathbb{N}.$$



Therefore, (3.5) and (3.6) imply

$$0 < E(0) - L \sum_{n=1}^{\infty} \delta = -\infty,$$

and a contradiction arises.

*Case 2.*  $\|u'(t)\|_2^2 \neq L$ . Set  $\liminf_{t \rightarrow +\infty} \|u'(t)\|_2^2 = \ell \in [0, L]$ , and assume first that  $\ell < L$ . Then, since  $u \in C^1([0, T]; H)$  by Lemma 2.1, there exist two sequences  $(s_n)_n$  and  $(t_n)_n$  such that

1.  $0 < s_n < t_n < s_{n+1} \quad \forall n \in \mathbb{N}$ ;
2.  $s_n \rightarrow +\infty$  as  $n \rightarrow \infty$ ;
3.  $\frac{L+\ell}{2} = \|u'(s_n)\|_2^2 < \|u'(t_n)\|_2^2 = \frac{3L+\ell}{4} \quad \forall n \in \mathbb{N}$ ;
4.  $\frac{L+\ell}{2} \leq \|u'(t)\|_2^2 \leq \frac{3L+\ell}{4} \quad \forall t \in (s_n, t_n)$ .

Since solutions of (1.1) are uniformly bounded, there exists  $M > 0$ , depending on  $\|Au_0\|_2$  and  $\|A^{1/2}u_1\|_2$ , such that

$$\begin{aligned} \frac{d}{dt} \|u'(t)\|_2^2 &= 2\langle u'(t), u''(t) \rangle \\ &= -2\langle u'(t), Au(t) \rangle - 2h(t)\|u'(t)\|_2^2 + 2\langle u'(t), f(u) \rangle \\ &\leq 2\|u'(t)\|_2 \|Au(t)\|_2 + 2\|u'(t)\|_2 \|f(u)\|_2 \leq M. \end{aligned}$$

Therefore

$$\frac{L-\ell}{4} = \frac{3L+\ell}{4} - \frac{L+\ell}{2} = \int_{s_n}^{t_n} \frac{d}{dt} \|u'(t)\|_2^2 dt \leq M(t_n - s_n),$$

so that

$$(3.7) \quad t_n - s_n \geq \frac{L-\ell}{4M} \quad \forall n \in \mathbb{N}.$$

In this way, by (2.6) and (3.7) we get

$$\begin{aligned} (3.8) \quad 0 < E_{\infty} &= E(0) + \int_0^{\infty} E'(\tau) d\tau \leq E(0) - \int_{\cup_n (s_n, t_n)} h(\tau) \|u'(\tau)\|_2^2 d\tau \\ &\leq E(0) - \frac{L+\ell}{2} \int_{\cup_n (s_n, s_n + \frac{L-\ell}{4M})} h(\tau) d\tau. \end{aligned}$$

Since  $h$  is integrally positive, there exists  $\delta > 0$  such that

$$\int_{(s_n, s_n + \frac{L-\ell}{4M})} h(\tau) d\tau \geq \delta \quad \forall n \in \mathbb{N},$$

so that (3.8) again gives  $0 < -\infty$ .

*Case 3.*  $u'(t)$  has limit. Now assume that there exists  $\lim_{t \rightarrow +\infty} \|u'(t)\|_2^2 = L$ . Then there exists  $M > 0$  such that

$$(3.9) \quad \|u'(t)\|_2^2 \geq \frac{L}{2} \quad \forall t \geq M.$$

In this way, by (2.6), Remark 2.3, and (3.9) we get

$$\begin{aligned}
 0 < E_\infty &= E(0) + \int_0^\infty E'(\tau) d\tau \leq C_M - \int_M^\infty h(\tau) \|u'(\tau)\|_2^2 d\tau \\
 (3.10) \qquad &\leq C_M - \frac{L}{2} \int_M^\infty h(\tau) d\tau,
 \end{aligned}$$

where  $C_M$  is a constant depending on  $E(0)$  and  $M$ . Since  $h$  is integrally positive, as in (3.6), there exists  $\delta > 0$  such that

$$\int_n^{n+1} h(\tau) d\tau \geq \delta \quad \forall n \in \mathbb{N},$$

so that (3.10) gives  $0 < -\infty$ , again a contradiction.

Thus we can conclude that  $L = 0$ , i.e.,

$$(3.11) \qquad \lim_{t \rightarrow +\infty} \|u'(t)\|_2 = 0.$$

As a consequence, (3.3) implies

$$\lim_{t \rightarrow \infty} \left( \frac{1}{2} \|A^{1/2} u(t)\|_2^2 - \int_\Omega F(u(t)) dx \right) = E_\infty > 0.$$

Then there exists  $T > 0$  such that for any  $t \geq T$ , one has

$$\frac{1}{2} \|A^{1/2} u(t)\|_2^2 - \int_\Omega F(u(t)) dx \geq \frac{E_\infty}{2}.$$

By (3.1), there exists  $\gamma = \gamma(u) > 0$  such that

$$(3.12) \qquad \|A^{1/2} u(t)\|_2 \geq \gamma \quad \forall t > T.$$

Now set

$$(3.13) \qquad v(t) = \langle u(t), u'(t) \rangle = \frac{1}{2} \frac{d}{dt} \|u(t)\|_2^2,$$

so that

$$\begin{aligned}
 v'(t) &= \|u'(t)\|^2 + \langle u(t), -h(t)u'(t) - Au + f(u) \rangle_{L^2} \\
 &= \|u'(t)\|^2 - h(t)v(t) - \|A^{1/2}u\|_2^2 + \langle u, f(u) \rangle_{L^2} \\
 &\leq \|u'(t)\|^2 - h(t)v(t) - \|A^{1/2}u\|_2^2
 \end{aligned}$$

by (2.3). Finally, (3.11) and (3.12) imply that there exist  $T_0 > T$  and  $\delta > 0$  such that

$$v'(t) \leq -\delta - h(t)v(t) \quad \forall t \geq T_0,$$

that is,

$$\frac{d}{dt} \left( v(t) e^{H(t)} \right) \leq -\delta e^{H(t)} \quad \forall t \geq T_0,$$

where  $H(t) = \int_0^t h(\tau) d\tau$ . Integrating between  $T_0$  and  $t$  gives

$$v(t) \leq v(T_0)e^{-H(t)+H(T_0)} - \delta e^{-H(t)} \int_{T_0}^t e^{H(\tau)} d\tau.$$

Integrating again between  $T_0$  and  $t$ , by (3.13) we get

$$\begin{aligned} \frac{1}{2}\|u(t)\|_2^2 &\leq \frac{1}{2}\|u(T_0)\|_2^2 \\ &+ v(T_0)e^{H(T_0)} \int_{T_0}^t e^{-H(s)} ds - \delta \int_{T_0}^t e^{-H(s)} \int_{T_0}^s e^{H(\tau)} d\tau ds. \end{aligned}$$

Letting  $t \rightarrow \infty$ , by (1.2) the right-hand side of the previous inequality goes to  $-\infty$ , and a contradiction arises. In fact, by (3.6) we get  $H(t) = \int_0^t h(\tau) d\tau > \delta[t] \geq \delta(t-1)$ , where  $[t]$  denotes the integer part of  $t$  (i.e., the greatest integer not bigger than  $t$ ), and then

$$\int_{T_0}^t e^{-H(\tau)} d\tau \leq \int_{T_0}^t e^{-\delta(\tau-1)} d\tau \leq \frac{e^{-\delta(T_0-1)}}{\delta}.$$

Thus  $E_\infty = 0$ , and since  $F(u) \leq 0$  for any  $u$ , (3.3) implies that  $\|u'(t)\|_2^2 + \|A^{1/2}u(t)\|_2^2 \rightarrow 0$  as  $t \rightarrow \infty$  and (3.2) clearly follows.  $\square$

*Remark 3.1.* 1. In proving the previous result, actually what is really needed is that

$$\text{if } (u_0, u_1) \in D(A) \times H \text{ and } \|Au_0\|_H + \|u_1\|_H \leq B_1,$$

then  $\forall t \geq t_0$   $f(u(t, t_0, u_0, u_1)) \in H$  and

$$\|Au(t, t_0, u_0, u_1)\|_H + \|u'(t, t_0, u_0, u_1)\|_H + \|f(u(t, t_0, u_0, u_1))\|_H \leq B_2.$$

However, we preferred to maintain the definition proposed in [22], since it is natural to deal with solutions whose time derivative is still in  $H_0^1(\Omega)$ , as it happens when it is possible to apply a regularity result.

2. Moreover, the proof above extends immediately to the abstract case, and this is the reason why we maintained the abstract formulation, writing, for example,  $\|A^{1/2}u\|_2$  in place of  $\|Du\|_2$ .

*Remark 3.2.* In proving the analogue of Theorem 3.1 in [22] for a sublinear  $f$ , Zhang didn't take into account the different possibilities about the limit  $L$  defined in (3.4). However, adapting our proof to any function  $f$  with sublinear growth and such that  $sf(s) < \lambda_1 s^2$ ,  $s \neq 0$ , like in [22], we can recover the stability result quoted therein.

As in [22], we prove that condition (1.2) is also necessary for asymptotic stability to hold, even without the assumption that  $h$  is integrally positive and without the sign assumption on  $f$ , though we need  $f$  to be nonsupercritical, in the usual sense. Moreover, we can even require a weaker a priori bound condition.

**DEFINITION 3.1.** *Solutions of (E) are said weakly uniformly bounded in  $D(A^{1/2}) \times H$  if for any  $B_1 > 0$  there exists  $B_2 > 0$  such that*

$$\text{if } (u_0, u_1) \in D(A^{1/2}) \times H, \|A^{1/2}u_0\|_H + \|u_1\|_H \leq B_1, \text{ and } t_0 \in [0, \infty),$$

then  $\forall t \geq t_0$ ,

- $f(u(t, t_0, u_0, u_1)) \in H$  and
- $\|A^{1/2}u(t, t_0, u_0, u_1)\|_H + \|u'(t, t_0, u_0, u_1)\|_H + \|f(u(t, t_0, u_0, u_1))\|_H \leq B_2$ , where  $u(t, t_0, u_0, u_1)$  denotes the solution of (E) such that  $u(t_0) = u_0$  and  $u'(t_0) = u_1$ .

**THEOREM 3.2.** Suppose that (2.3) holds and that solutions of (W) are weakly uniformly bounded in  $D(A^{1/2}) \times L^2(\Omega)$ . In addition, assume that

$$(3.14) \quad \exists a, b \geq 0, p \in \begin{cases} [1, \infty) & \text{if } N = 1, 2, \\ [1, \frac{N}{N-2}] & \text{if } N \geq 3, \end{cases} \text{ such that } |f(s)| \leq a + b|s|^p \quad \forall s \in \mathbb{R}.$$

If every solution of (1.1) satisfies (3.2), then (1.2) holds.

*Proof.* Since all solutions of (W) are weakly uniformly bounded in  $D(A^{1/2}) \times L^2(\Omega)$ , for any  $D > 0$  there exists  $M > 0$  such that, for any  $u_0, u_1 \in D(A^{1/2}) \times L^2(\Omega)$  with  $\|A^{1/2}u_0\|_2^2 + \|u_1\|_2^2 \leq D$ , for any  $t_0 \geq 0$  there holds in particular

$$(3.15) \quad \|A^{1/2}u(t, t_0, u_0, u_1)\|_2 \leq M \quad \forall t > 0.$$

Moreover, by the Hölder and Sobolev inequalities, there exist  $S_1, S_{p+1} > 0$  such that for any  $u \in H_0^1(\Omega)$  there holds

$$(3.16) \quad \int_{\Omega} |u| dx \leq S_1 \|A^{1/2}u\|_2 \quad \text{and} \quad \int_{\Omega} |u|^{p+1} dx \leq S_{p+1} \|A^{1/2}u\|_2^{p+1}.$$

In fact, note that if  $N \geq 3$ , then  $p+1 \leq (2N-2)/(N-2)$  and Sobolev's theorem can be applied.

Assume by contradiction that

$$\int_0^\infty e^{-H(t)} \int_0^t e^{H(s)} ds dt < \infty.$$

Then, for any  $\gamma > 0$  there exists  $t_0$  such that

$$(3.17) \quad \int_{t_0}^\infty e^{-H(t)} \int_{t_0}^t e^{H(s)} ds dt < \frac{D}{8\gamma}.$$

Now take  $\phi \in H_0^1(\Omega)$  such that  $\|A^{1/2}\phi\|_2^2 = D/2$ ,  $\|\phi\|_2^2 = D/2$ , and consider the solution  $u$  of (W) such that  $u(t_0) = \phi$  and  $u_t(t_0) = 0$ , so that  $(\phi, 0)$  guarantees  $\|A^{1/2}u(t_0)\|_2^2 + \|u_t(t_0)\|_2^2 \leq D$ . Finally, for  $t \geq t_0$ , set  $w(t) = \langle u(t), u'(t) \rangle_{L^2} = \frac{1}{2}(\|u(t)\|_2^2)'$ . Differentiating, we get

$$\begin{aligned} w'(t) &= \|u'(t)\|_2^2 + \langle u(t), u''(t) \rangle_{L^2} \\ &= \|u'(t)\|_2^2 + \langle u(t), -Au \rangle - h(t)w(t) + \langle u(t), f(u) \rangle_{L^2} \\ (3.18) \quad &= \|u'(t)\|_2^2 - \|A^{1/2}u(t)\|_2^2 - h(t)w(t) + \langle u(t), f(u) \rangle_{L^2} \\ &\geq -\|A^{1/2}u(t)\|_2^2 - h(t)w(t) - \int_{\Omega} (a|u(t)| + b|u(t)|^{p+1}) dt. \end{aligned}$$

By (3.16) we get

$$(3.19) \quad w'(t) \geq -\|A^{1/2}u(t)\|_2^2 - h(t)w(t) - aS_1 \|A^{1/2}u\|_2 - bS_{p+1} \|A^{1/2}u\|_2^{p+1}.$$

By the Young inequality we can find  $\eta > 0$  such that (3.19) gives

$$w'(t) \geq -\eta(1 + \|A^{1/2}u(t)\|_2^{p+1}) - h(t)w(t) - bS_{p+1}\|A^{1/2}u\|_2^{p+1},$$

and by (3.15),

$$w'(t) \geq -(\eta + \eta M^{p+1} + bS_{p+1}M^{p+1}) - h(t)w(t).$$

Setting  $\gamma = \eta + \eta M^{p+1} + bS_{p+1}M^{p+1} > 0$  (which is independent of  $t_0$ ), we have  $w'(t) + h(t)w(t) \geq -\gamma$ , i.e.,

$$(3.20) \quad (e^{H(t)}w(t))' \geq -\gamma.$$

Integrating (3.20) twice between  $t_0$  and  $t$  gives

$$\frac{1}{2}\|u(t)\|_2^2 \geq \frac{1}{2}\|u(t_0)\|_2^2 - \gamma \int_{t_0}^t e^{-H(s)} \int_{t_0}^s e^{H(\tau)} d\tau ds.$$

Finally, by (3.17), we get

$$\frac{1}{2}\|u(t)\|_2^2 \geq \frac{D}{8}.$$

By Poincaré's inequality we obtain

$$\|A^{1/2}u(t)\|_2^2 \geq \frac{\lambda_1 D}{4},$$

so that (3.2) cannot hold.  $\square$

*Remark 3.3.* 1. Of course a uniform bound implies a weak uniform bound by Poincaré's inequality. However, we preferred to present Theorem 3.2 under the more general assumption of a weak uniform bound on the set of solutions.

2. Again, in the proof of Theorem 3.2 we maintained the abstract form  $A$  for  $-\Delta$ , since an analogous version for the abstract problem (2.1) can be provided at once, where  $\lambda_1$  is now given by (2.2).

Summing up, in view of Theorems 3.2 and 3.1, and recalling that (3.14) implies (3.1) when  $a = 0$ , we can conclude with the following result.

**THEOREM 3.3.** *Let  $\Omega$  be a bounded domain of  $\mathbb{R}^N$ ,  $N \geq 1$ . Assume (2.3), (3.1), and (3.14); moreover, suppose that  $h$  is integrally positive and that solutions of (W) are uniformly bounded in  $H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega)$ . Then every solution of (1.1) verifies*

$$\|Du(t)\|_2 + \|u'(t)\|_2 \rightarrow 0 \quad \text{as } t \rightarrow +\infty$$

if and only if (1.2) holds.

**4. The on-off case.** As in section 2, we set  $V = D(A^{1/2}) \subset H$ , and by (2.2) there exists  $\lambda_1 > 0$  such that

$$\|v\|_H^2 \leq \frac{1}{\lambda_1} \|v\|_V^2 = \frac{1}{\lambda_1} \|A^{1/2}v\|_H^2 \quad \text{for any } v \in V.$$

We remark that in the standard case of problem (1.1) we have  $H = L^2(\Omega)$ ,  $V = H_0^1(\Omega)$  and  $\lambda_1$  is the first eigenvalue of  $-\Delta$  on  $H_0^1(\Omega)$ , the inequality above being the usual Poincaré's inequality.

The inequality expressed in (2.4) can be replaced in the abstract case by the condition that

$$(4.1) \quad \langle f(u), u \rangle_{V',V} - \mathcal{F}(u) \leq 0 \quad \forall u \in V,$$

though it would suffice to hold only for those  $u$ 's which solve (2.1). Of course this condition, in effect, reduces to (2.4) in the model case of problem (1.1).

THEOREM 4.1. *Fix  $T > 0$  and assume there exist  $M, m > 0$  such that*

$$(4.2) \quad m\|v\|_H^2 \leq \langle B(t)v, v \rangle_H \quad \forall t \in [0, T], \forall v \in H,$$

$$(4.3) \quad \|B(t)v\|_H^2 \leq M\langle B(t)v, v \rangle_H \quad \forall t \in [0, T], \forall v \in H.$$

Moreover, suppose that (2.2) and (4.1) hold. Then for every  $(u_0, u_1) \in V \times H$  the solution  $u$  of problem (2.1) satisfies

$$(4.4) \quad E_u(T) \leq \frac{1}{1 + \frac{T^3}{30} \frac{1}{\frac{4}{\lambda_1 m} + \frac{3T^2}{32m} + \frac{MT^2}{16\lambda_1}}} E_u(0).$$

*Proof.* As in [6] and [7], for any  $t \in [0, T]$  we set  $\theta(t) = t^2(T - t)^2$ , so that  $\theta'(t) = 2t(T - t)(T - 2t)$  and

$$(4.5) \quad |\theta'(t)| \leq 2T\theta^{1/2}(t) \quad \forall t \in [0, T],$$

$$(4.6) \quad \max_{t \in [0, T]} \theta(t) = \frac{T^4}{16},$$

and

$$(4.7) \quad \int_0^T \theta(t) dt = \frac{T^5}{30}.$$

Therefore (2.6) gives

$$(4.8) \quad E(0) - E(T) = \int_0^T \langle Bu', u' \rangle_H dt \geq 0.$$

Multiplying (2.1) by  $\theta u$  gives

$$\int_0^T \theta \left\{ \langle u'' + Bu', u \rangle_{V',V} + \|A^{1/2}u\|_H^2 - \langle f(u), u \rangle_{V',V} \right\} dt = 0.$$

Therefore, integrating by parts,

$$\int_0^T \theta \|A^{1/2}u\|_H^2 dt = \int_0^T \left\{ \langle (\theta u)', u' \rangle_H - \theta \langle Bu', u \rangle_H + \theta \langle f(u), u \rangle_{V',V} \right\} dt,$$

that is,

$$\int_0^T \theta \|A^{1/2}u\|_H^2 dt = \int_0^T \left\{ \theta \|u'\|_H^2 + \theta' \langle u, u' \rangle_H - \theta \langle Bu', u \rangle_H + \theta \langle f(u), u \rangle_{V',V} \right\} dt.$$

Thus for any  $\varepsilon, \eta > 0$  we get

$$\begin{aligned} \int_0^T \theta \|A^{1/2}u\|_H^2 dt &\leq \int_0^T \left\{ \varepsilon \theta \|u\|_H^2 + \frac{1}{4\varepsilon} \theta \|Bu'\|_H^2 \right. \\ &\quad \left. + \eta \theta'^2 \|u\|_H^2 + \frac{1}{4\eta} \|u'\|_H^2 + \theta \|u'\|_H^2 + \theta \langle f(u), u \rangle_{V',V} \right\} dt. \end{aligned}$$

By (4.5), (4.6), (4.3), and (2.2) we get

$$\begin{aligned} (4.9) \quad \int_0^T \theta \|A^{1/2}u\|_H^2 dt &\leq \int_0^T \left\{ \frac{\varepsilon}{\lambda_1} \theta \|A^{1/2}u\|_H^2 + \frac{MT^4}{64\varepsilon} \langle Bu', u' \rangle_H \right. \\ &\quad \left. + \frac{4\eta T^2}{\lambda_1} \theta \|A^{1/2}u\|_H^2 + \frac{1}{4\eta} \|u'\|_H^2 + \frac{T^4}{16} \|u'\|_H^2 + \theta \langle f(u), u \rangle_{V',V} \right\} dt. \end{aligned}$$

Let us choose  $\varepsilon$  and  $\eta$  so that

$$\frac{4\eta T^2}{\lambda_1} = \frac{\varepsilon}{\lambda_1} = \frac{1}{4}.$$

Then (4.9) reads

$$\begin{aligned} \int_0^T \theta \|A^{1/2}u\|_H^2 dt &\leq \int_0^T \left\{ \frac{1}{4} \theta \|A^{1/2}u\|_H^2 + \frac{MT^4}{16\lambda_1} \langle Bu', u' \rangle_H + \theta \langle f(u), u \rangle_{V',V} \right. \\ &\quad \left. + \frac{1}{4} \theta \|A^{1/2}u\|_H^2 + \frac{4T^2}{\lambda_1} \|u'\|_H^2 + \frac{T^4}{16} \|u'\|_H^2 \right\} dt, \end{aligned}$$

so that

$$\begin{aligned} &\frac{1}{2} \int_0^T \theta \|A^{1/2}u\|_H^2 dt \\ &\leq \int_0^T \left\{ C_T \|u'\|_H^2 + \frac{MT^4}{16\lambda_1} \langle Bu', u' \rangle_H + \theta \langle f(u), u \rangle_{V',V} \right\} dt, \end{aligned}$$

where  $C_T = \frac{4T^2}{\lambda_1} + \frac{T^4}{16}$ . By (4.8) this means

$$\begin{aligned} (4.10) \quad &\frac{1}{2} \int_0^T \theta \|A^{1/2}u\|_H^2 dt \\ &\leq \int_0^T \left\{ C_T \|u'\|_H^2 + \theta \langle f(u), u \rangle_{V',V} \right\} dt + \frac{MT^4}{16\lambda_1} (E(0) - E(T)). \end{aligned}$$

By (2.5),

$$\frac{1}{2} \int_0^T \theta \|A^{1/2}u\|_H^2 dt = \int_0^T \theta E(t) dt - \frac{1}{2} \int_0^T \theta \|u'\|_H^2 dt + \int_0^T \theta \mathcal{F}(u) dt,$$

and by Lemma 2.2,

$$\begin{aligned} (4.11) \quad &\frac{1}{2} \int_0^T \theta \|A^{1/2}u\|_H^2 dt \\ &\geq E(T) \int_0^T \theta(t) dt - \frac{1}{2} \int_0^T \theta \|u'\|_H^2 dt + \int_0^T \theta \mathcal{F}(u) dt. \end{aligned}$$

Therefore (4.7), (4.10), (4.6), and (4.11) give

$$\begin{aligned} \frac{T^5}{30} E(T) &\leq \int_0^T \theta[\langle f(u), u \rangle_{V', V} - \mathcal{F}(u)] dt \\ &+ \left( C_T + \frac{T^4}{32} \right) \int_0^T \|u'\|_H^2 dt + \frac{MT^4}{16\lambda_1} (E(0) - E(T)), \end{aligned}$$

which implies, together with (4.1),

$$\left( \frac{T^5}{30} + \frac{MT^4}{16\lambda_1} \right) E(T) \leq \left( C_T + \frac{T^4}{32} \right) \int_0^T \|u'\|_H^2 dt + \frac{MT^4}{16\lambda_1} E(0).$$

By (4.2) we get

$$\left( \frac{T^5}{30} + \frac{MT^4}{16\lambda_1} \right) E(T) \leq \frac{1}{m} \left( C_T + \frac{T^4}{32} \right) \int_0^T \langle Bu', u' \rangle_H dt + \frac{MT^4}{16\lambda_1} E(0).$$

By (2.6) this implies

$$\left( \frac{T^5}{30} + \frac{MT^4}{16\lambda_1} \right) E(T) \leq \frac{1}{m} \left( C_T + \frac{T^4}{32} \right) (E(0) - E(T)) + \frac{MT^4}{16\lambda_1} E(0),$$

and so

$$\left( \frac{T^5}{30} + \frac{C_T}{m} + \frac{T^4}{32m} + \frac{MT^4}{16\lambda_1} \right) E(T) \leq \left( \frac{C_T}{m} + \frac{T^4}{32m} + \frac{MT^4}{16\lambda_1} \right) E(0)$$

and (4.4) follows.  $\square$

*Remark 4.1.* Theorem 4.1 can be generalized to any interval  $[a, b]$ , obtaining

$$E_u(b) \leq \frac{1}{1 + \frac{T^3}{30} \frac{1}{\frac{4}{\lambda_1 m} + \frac{3T^2}{32m} + \frac{MT^2}{16\lambda_1}}} E_u(a)$$

with  $T = b - a$ . Indeed, in  $[a, b]$  take  $\theta(t) = (t - a)^2(b - t)^2$ ; then  $\theta'(t) = 2\sqrt{\theta}(b - 2t + a)$ , so that (4.5), (4.6), (4.7) are replaced by  $|\theta'(t)| \leq 2(b - a)\theta^{1/2}(t)$ ,  $\max_{[a, b]} \theta = \frac{(b-a)^4}{16}$ , and

$$\int_a^b \theta(t) dt = \int_0^T x^2(T - x)^2 dx = \frac{(b - a)^5}{30},$$

respectively, which are the same formal estimates, since  $b - a = T$ . Analogously, (4.8) is replaced by

$$E(a) - E(b) = \int_a^b \langle Bu', u' \rangle_H dt \geq 0.$$

Now multiply (2.1) by  $\theta u$  and integrate in  $[a, b]$ ; performing the same estimates as in the proof of Theorem 4.1, we get the desired result.

Theorem 4.1 is the essential tool for the following stability result, which can be proved by extending the method of Smith (see [21]) as already done in [7].



**THEOREM 4.2.** *Let  $(a_n, b_n)_n$  be a sequence of disjoint open intervals in  $(0, +\infty)$  with  $a_n \rightarrow +\infty$  and assume that (1.3), (1.4), (2.2), and (4.1) hold. Then for every  $(u_0, u_1) \in D(A^{1/2}) \times H$  the solution  $u$  of problem (2.1) is such that  $E_u(t) \rightarrow 0$  as  $t \rightarrow +\infty$ .*

*Proof.* The proof is now similar to the proof of [7, Theorem 3.2], since the main tool in this proof is an inequality of the type of (4.4) proved in Theorem 4.1. We sketch it for completeness.

Apply Theorem 4.1 in the form of Remark 4.1 in the interval  $(a_n, b_n)$  instead of the interval  $(0, T)$ , obtaining

$$(4.12) \quad E_u(b_n) \leq \frac{1}{1 + \frac{T_n^3}{30} \frac{1}{\frac{4}{\lambda_1 m_n} + \frac{3T_n^2}{32m_n} + \frac{M_n T_n^2}{16\lambda_1}}} E_u(a_n),$$

where we have set  $T_n = b_n - a_n$ .

Defining

$$k_n := \frac{m_n T_n^3}{128 + 3\lambda_1 T_n^2 + 2m_n M_n T_n^2} \quad \text{and} \quad c = \frac{16\lambda_1}{15},$$

(4.12) can be rewritten as

$$E_u(b_n) \leq \frac{1}{1 + ck_n} E_u(a_n).$$

Since  $E$  is nonincreasing, by iteration we get that for any  $n \in \mathbb{N}$ ,

$$E(a_{n+1}) \leq E(b_n) \leq \frac{1}{1 + ck_n} E_u(a_n) \leq E(a_0) \prod_{j=0}^n \frac{1}{1 + ck_j} \leq E(0) \prod_{j=0}^n \frac{1}{1 + ck_j}.$$

Since  $E$  is nonincreasing, Theorem 4.2 will be proved if we show that  $E(a_{n+1}) \rightarrow 0$  as  $n \rightarrow \infty$ ; therefore, let us show that

$$\prod_{j=0}^{\infty} \frac{1}{1 + ck_j} = 0, \quad \text{or equivalently,} \quad \sum_{j=0}^{\infty} \log \frac{1}{1 + ck_j} = -\infty.$$

This condition obviously holds if  $k_j \not\rightarrow 0$  as  $j \rightarrow \infty$ , while if  $k_j \rightarrow 0$ , it means that

$$\sum_{j=1}^{\infty} k_j = +\infty.$$

This last condition is equivalent to (1.4); indeed, if  $T_j^2(3\lambda_1 + 2m_j M_j) := T_j^2 c_j \geq 128$ , then  $k_j \geq \frac{1}{2c_j}$ , while if  $T_j^2 c_j \leq 128$ , then  $k_j \geq \frac{T_j^2}{2 \times 128}$ , concluding that

$$k_j \geq \frac{1}{2} m_j (b_j - a_j) \min \left( \frac{(b_j - a_j)^2}{128}, \frac{1}{\lambda_1 + 2m_j M_j} \right).$$

On the other hand,

$$k_j \leq m_j (b_j - a_j) \min \left( \frac{(b_j - a_j)^2}{128}, \frac{1}{\lambda_1 + 2m_j M_j} \right),$$

and the claim follows.  $\square$

### 5. Some concrete applications.

**5.1. The integrally positive case.** Consider again problem (1.1), where  $\Omega$  is a bounded domain of  $\mathbb{R}^N$ ,  $N \geq 1$ . First, let us briefly show that the set of solutions of (W) is weakly uniformly bounded in  $H_0^1(\Omega) \times L^2(\Omega)$  under the subcritical growth assumption on  $f$  of Theorem 3.2, even without any sign condition on  $f$ . Indeed, by Lemma 2.2, if  $u$  solves (E) with  $u(t_0) = u_0$  and  $u'(t_0) = u_1$  for some  $t_0 \geq 0$ , we get for  $t \geq t_0$ ,

$$E(t) \leq E(t_0) = \frac{1}{2}\|u_1\|_2^2 + \frac{1}{2}\|u_0\|_{H_0^1(\Omega)}^2 - \int_{\Omega} F(u_0) \, dx.$$

But

$$\int_{\Omega} |F(u_0)| \, dx \leq a \int_{\Omega} |u_0| \, dx + \frac{b}{p+1} \int_{\Omega} |u_0|^{p+1} \, dx,$$

and by (3.16) there exists  $S > 0$ ,

$$\int_{\Omega} |F(u_0)| \, dx \leq S \left( \|u_0\|_{H_0^1(\Omega)} + \|u_0\|_{H_0^1(\Omega)}^{p+1} \right).$$

Therefore, if  $\|u_0\|_{H_0^1(\Omega)} + \|u_1\|_2 \leq B_1$ , then

$$E(t_0) \leq \frac{B_1^2}{2} + S(B_1 + B_1^p) := B_2.$$

Hence for every  $t \geq t_0$ ,

$$\begin{aligned} \|u(t, t_0, u_0, u_1)\|_{H_0^1(\Omega)}^2 + \|u'(t, t_0, u_0, u_1)\|_2^2 \\ = 2E(t) + 2 \int_{\Omega} F(u) \, dx \leq 2E(t_0) = 2B_2; \end{aligned}$$

that is, Definition 3.1 is verified, as claimed, whatever  $t_0$  is.

Finally, we recall that the request that the set of solutions of (1.1) is uniformly bounded in  $H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega)$  is not so strange. Several examples are considered in [22], and we refer to those cases therein for a sublinear  $f$ , simply recalling the following.

*Example 5.1* (see [22, Example 3.1]). Assume that  $f(s) = \alpha s$  for some constant  $\alpha < \lambda_1$ , where  $\lambda_1$  is the first eigenvalue of  $-\Delta$  on  $H_0^1(\Omega)$ . Then the set of solutions of (1.1) is uniformly bounded in  $H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega)$ .

Finally, we show that the set of solutions of (1.1) is uniformly bounded in  $H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega)$  also in more general cases. This result appears in [22] for  $f$  having linear growth, under the additional assumptions that  $h$  is bounded above and below by strictly positive constants (though there is a mistake in the final step on page 198). We take the latter assumption, but we let  $f$  have superlinear growth. However, let us note that the growth condition we give on  $f$  immediately implies condition (3.1), which is therefore useless from now on.

**LEMMA 5.1.** *Suppose that  $N = 1$ , and that  $f$  is an absolutely continuous function satisfying (2.3) such that*

$$\exists b_1 > 0, p \in [1, \infty) : \quad |f'(s)| \leq b_1 |s|^{p-1} \quad \forall s \in \mathbb{R}.$$

Assume that there exist two positive constants  $\alpha < \beta$  such that  $\alpha \leq h(t) \leq \beta$  for any  $t \geq 0$ . Finally, assume that  $\forall P > 0$  there exists  $Q > 0$  such that  $\|Du\|_2 \leq P$  implies  $\|Df(u)\|_2 \leq Q\|Du\|_2$ . Then the set of solutions of (W) is uniformly bounded in  $H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega)$ .

*Proof.* First, let us note that the condition on  $f'$  implies that  $|f(s)| \leq b_2|s|^p$  for any  $s$ . Now take  $B_1 > 0$  and  $(u_0, u_1) \in H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega)$  such that  $\|\Delta u_0\|_2 + \|Du_1\|_2 \leq B_1$ ; take  $t_0 \geq 0$ , and observe that the growth condition on  $f$  and the Sobolev inequality immediately ensure that  $f(u) \in L^2(\Omega)$  for any  $t \geq t_0$ .

By Lemma 2.2 and Remark 2.3 we get, setting  $u(t) = u(t, t_0, u_0, u_1)$ ,

$$\begin{aligned} \|u'(t)\|_2^2 + \|Du(t)\|_2^2 &\leq 2E(t_0) + 2 \int_{\Omega} F(u(t)) dx \\ &\leq 2E(t_0) = \|u_1\|_2^2 + \|Du_0\|_2^2 - 2 \int_{\Omega} F(u_0) dx. \end{aligned}$$

By the Poincaré and Sobolev inequalities and the growth condition on  $f$ , we get

$$\|u_1\|_2^2 + \|Du_0\|_2^2 - 2 \int_{\Omega} F(u_0) dx \leq C \left( \|Du_1\|_2^2 + \|\Delta u_0\|_2^2 + \|u\|_1 + \|u\|_{p+1}^{p+1} \right).$$

Applying again the Poincaré and Sobolev inequalities, since  $\|\Delta u_0\|_2 + \|Du_1\|_2 \leq B_1$ , we get the existence of a constant  $B > 0$  such that

$$(5.1) \quad \|u'(t)\|_2^2 + \|Du(t)\|_2^2 \leq B \quad \forall t \geq t_0.$$

Now set

$$V(t) = \int_{\Omega} u(t)u'(t) dx.$$

Proceeding as in [22], we can find

$$V'(t) = \int_{\Omega} (u'(t)^2 + u''(t)u(t)) dx,$$

and using the equation in (1.1),

$$V'(t) = \int_{\Omega} u'(t)^2 dx - \int_{\Omega} |Du(t)|^2 dx - h(t) \int_{\Omega} u'(t)u(t) dx + \int_{\Omega} f(u(t))u(t) dx.$$

By (2.3) and the Hölder, Young, and Poincaré inequalities,

$$V'(t) \leq \left(1 + \frac{\beta^2}{2\varepsilon}\right) \int_{\Omega} u'(t)^2 dx + \left(\frac{\varepsilon}{2\lambda_1} - 1\right) \int_{\Omega} |Du(t)|^2 dx$$

for any  $\varepsilon > 0$ . Therefore

$$V(t) \leq V(t_0) + \left(1 + \frac{\beta^2}{2\varepsilon}\right) \int_{t_0}^t \|u'(\tau)\|_2^2 d\tau + \left(\frac{\varepsilon}{2\lambda_1} - 1\right) \int_{t_0}^t \|Du(\tau)\|_2^2 d\tau.$$

Choosing  $\varepsilon < 2\lambda_1$  we get the existence of a positive constant  $C$  such that

$$\int_{t_0}^t \|Du(\tau)\|_2^2 d\tau \leq C \left( V(t_0) + \int_{t_0}^t \|u'(\tau)\|_2^2 d\tau \right) - V(t) \quad \forall t \geq t_0.$$

But

$$|V(t)| \leq \|u'(t)\|_2 \|u(t)\|_2 \leq \frac{1}{\sqrt{\lambda_1}} \|u'(t)\|_2 \|Du(t)\|_2 \leq \frac{B}{\sqrt{\lambda_1}} \quad \forall t \geq t_0$$

by (5.1), so that

$$(5.2) \quad \int_{t_0}^t \|Du(\tau)\|_2^2 d\tau \leq C_1 \left( 1 + \int_{t_0}^t \|u'(\tau)\|_2^2 d\tau \right) \quad \forall t \geq t_0$$

for some positive constant  $C_1$ .

Using again Lemma 2.2 and Remark 2.3, we have

$$E(t) = E(t_0) - \int_{t_0}^t h(\tau) \|u'(\tau)\|_2^2 d\tau,$$

so that

$$\int_{t_0}^t \|u'(\tau)\|_2^2 d\tau \leq \frac{E(t_0)}{\alpha} \leq C_2 \quad \forall t \geq t_0$$

for some positive constant  $C_2$ . In this way (5.2) implies the existence of  $C_3 > 0$  such that

$$(5.3) \quad \int_{t_0}^t \|Du(\tau)\|_2^2 d\tau \leq C_3 \quad \forall t \geq t_0.$$

Finally, introduce

$$W(t) = \frac{1}{2} \int_{\Omega} |Du'(t)|^2 dx + \frac{1}{2} \int_{\Omega} |\Delta u(t)|^2 dx.$$

As in [22], we find

$$W(t) = W(t_0) - \int_{t_0}^t h(\tau) \|Du'(\tau)\|_2^2 d\tau + \int_{t_0}^t \int_{\Omega} Du'(\tau) \cdot Df(u(\tau)) dx d\tau,$$

so that by Cauchy's inequality,

$$W(t) \leq W(t_0) - \int_{t_0}^t h(\tau) \|Du'(\tau)\|_2^2 d\tau + \int_{t_0}^t \|Du'(\tau)\|_2 \|Df(u(\tau))\|_2 d\tau.$$

By Young's inequality, for any  $\varepsilon > 0$ ,

$$\begin{aligned} W(t) \leq & W(t_0) - \int_{t_0}^t h(\tau) \|Du'(\tau)\|_2^2 d\tau \\ & + \varepsilon \int_{t_0}^t \|Du'(\tau)\|_2^2 d\tau + \frac{1}{2\varepsilon} \int_{t_0}^t \|Df(u(\tau))\|_2^2 d\tau. \end{aligned}$$

By assumption on  $h$ , we finally get

$$W(t) \leq W(t_0) + (\varepsilon - \alpha) \int_{t_0}^t h(\tau) \|Du'(\tau)\|_2^2 d\tau + \frac{1}{2\varepsilon} \int_{t_0}^t \|Df(u(\tau))\|_2^2 d\tau.$$

Choosing  $\varepsilon < \alpha$  and recalling that  $\|\Delta u_0\|_2 + \|Du_1\|_2 \leq B_1$ , we get

$$(5.4) \quad W(t) \leq C + C \int_{t_0}^t \|Df(u(\tau))\|_2 d\tau \quad \forall t \geq t_0$$

for some positive constant  $C$ .

By assumption, taking  $P := \sqrt{B}$  in (5.1), from (5.4) we get

$$(5.5) \quad W(t) \leq C + CQ \int_{t_0}^t \|Du(\tau)\|_2^p d\tau \quad \forall t \geq t_0.$$

By (5.3) we find  $D > 0$  such that  $W(t) \leq D$  for any  $t$ ; i.e., the set of solutions is uniformly bounded in  $H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega)$ , as claimed.  $\square$

As a final application, Lemma 5.1 gives the following.

**PROPOSITION 5.1.** *Suppose that  $N = 1$ , and that  $f$  is an absolutely continuous function satisfying (2.3) such that*

$$\exists b_1 > 0, p \in [1, \infty) : \quad |f'(s)| \leq b_1 |s|^{p-1} \quad \forall s \in \mathbb{R}.$$

*Finally, assume that there exist two positive constants  $\alpha < \beta$  such that  $\alpha \leq h(t) \leq \beta$  for any  $t \geq 0$ . Then the set of solutions of (W) is uniformly bounded in  $H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega)$ .*

*Proof.* The only thing to prove is that for any  $P > 0$  there exists  $Q > 0$  such that  $\|Du\|_2 \leq P$  implies  $\|Df(u)\|_2 \leq Q$ . But this is just an application of Sobolev's and Poincaré's inequalities. Indeed, if  $\|Du\|_2 \leq P$ , then

$$\begin{aligned} \|Df(u)\|_2 &= \|f'(u)Du\|_2 \leq b_1 \| |u|^{p-1} Du \|_2 \\ &\leq C \|Du\|_2^p \leq CP^{p-1} \|Du\|_2 = Q \|Du\|_2, \end{aligned}$$

as soon as  $Q = CP^{p-1}$ . Now apply Lemma 5.1.  $\square$

Summing up, Theorems 3.1 and 3.2 and Proposition 5.1 imply the following final result.

**PROPOSITION 5.2.** *Assume  $N = 1$  and suppose that  $f$  is an absolutely continuous function satisfying (2.3) and such that*

$$\exists b_1 > 0, p \in [1, \infty) : \quad |f'(s)| \leq b_1 |s|^{p-1} \quad \forall s \in \mathbb{R}.$$

*Moreover, assume that there exist two positive constants  $\alpha < \beta$  such that  $\alpha \leq h(t) \leq \beta$  for any  $t \geq 0$ . Then every solution of (1.1) satisfies*

$$\|u\|_{H_0^1(\Omega)} + \|u'\|_{L^2(\Omega)} \rightarrow 0 \quad \text{as } t \rightarrow +\infty$$

*if and only if (1.2) holds.*

We end up with the following.

**Remark 5.1.** If  $h(t) \geq \alpha > 0 \forall t \geq 0$ , as in the previous cases, it is easily seen that  $h$  is integrally positive, and if  $h(t) \equiv \alpha > 0$ , then also (1.2) holds. A more interesting example is given by the periodic function  $h(t) = |\sin t|$ , which also satisfies (1.2); indeed, denoting again the integer part of a real number  $x$  by  $[x]$ , we have

$$2\frac{t}{\pi} - 2 \leq 2\left[\frac{t}{\pi}\right] \leq H(t) = \int_0^{\left[\frac{t}{\pi}\right]\pi} |\sin s| ds + \int_{\left[\frac{t}{\pi}\right]\pi}^t |\sin s| ds \leq 2\left[\frac{t}{\pi}\right] + 2 \leq 2\frac{t}{\pi} + 2,$$

so that  $\int_0^t e^{H(s)} ds \geq \frac{\pi}{2}(e^{2t/\pi-2} - e^{-2})$  and

$$\int_0^\tau e^{-H(t)} \int_0^t e^{H(s)} ds dt \geq \frac{\pi}{2} \int_0^\tau (e^{-4} - e^{-2t/\pi-4}) dt \rightarrow \infty$$

as  $\tau \rightarrow \infty$ , so that (1.2) holds.

**5.2. The on-off case.** A particular case of the abstract problem considered in section 4 is the following nonlinear wave system in a bounded domain  $\Omega$  of  $\mathbb{R}^N$ ,  $N \geq 1$ :

$$(\widetilde{W}) \quad \begin{cases} u_{tt} = \Delta u - h(t)g(u_t) + f(u) & \text{in } (0, +\infty) \times \Omega, \\ u(t, x) = 0 & \text{in } (0, +\infty) \times \partial\Omega, \\ u(0, x) = u_0(x), \quad u_t(0, x) = u_1(x), & x \in \Omega, \end{cases}$$

where the following assumptions are made:

$$(A) \quad \begin{cases} g : \mathbb{R} \longrightarrow \mathbb{R} \text{ is a } C^1 \text{ function with } g(0) = 0, \\ \exists B \geq A > 0 \text{ such that } 0 < A \leq g'(v) \leq B \quad \forall v \in \mathbb{R}, \\ f \text{ satisfies (2.4),} \end{cases}$$

while  $u_0 \in H_0^1(\Omega)$  and  $u_1 \in L^2(\Omega)$ .

In this case Theorem 4.1 can be easily generalized as follows.

**THEOREM 5.1.** *Fix  $T > 0$  and assume there exist  $M, m > 0$  such that*

$$(5.6) \quad 0 < m \leq h(t) \leq M \quad \forall t \in [0, T].$$

*Suppose (A) holds. Then for every  $(u_0, u_1) \in H_0^1(\Omega) \times L^2(\Omega)$  the solution  $u$  of problem  $(\widetilde{W})$  satisfies*

$$E(T) \leq \frac{1}{1 + \frac{T^3}{30} \frac{1}{\frac{4}{\lambda_1} + \frac{3T^2}{32m} + \frac{MT^2}{16\lambda_1}}} E(0).$$

In the same way as Theorem 4.2 is implied by Theorem 4.1, Theorem 5.1 immediately gives the following fundamental application.

**THEOREM 5.2.** *Let  $(a_n, b_n)_n$  be a sequence of disjoint open intervals in  $(0, +\infty)$  with  $a_n \rightarrow +\infty$  and assume that (1.3), (1.4), and (A) hold. Then for every  $(u_0, u_1) \in H_0^1(\Omega) \times L^2(\Omega)$  the solution  $u$  of problem  $(\widetilde{W})$  is such that  $E_u(t) \rightarrow 0$  as  $t \rightarrow +\infty$ .*

Of course the abstract setting we gave in section 2 lets us deal with higher order problems. Indeed, consider the problem

$$(H) \quad \begin{cases} u_{tt} = \Delta^{2m} u - h(t)g(u_t) + f(u) & \text{in } (0, +\infty) \times \Omega, \\ Cu(t, x) = 0 \in \mathbb{R}^{2m} & \text{in } (0, +\infty) \times \partial\Omega, \\ u(0, x) = u_0(x) \in D(\Delta^m), \quad u_t(0, x) = u_1(x), & x \in \Omega, \end{cases}$$

where  $m \in \mathbb{N}$ ,  $g$  is as before, and  $C$  is a boundary operator. For example, if  $m = 1$  and  $Cu = (u, \partial u / \partial \nu)$ ,  $\nu$  being the unit outward normal to  $\partial\Omega$ , we have  $D(\Delta^m) = H_0^2(\Omega)$ , while, in the case  $Cu = (u, \Delta u)$  we have  $D(\Delta^m) = H_0^1(\Omega) \cap H^2(\Omega)$ . Other generalizations are easy to do. For this case we have the following.

**THEOREM 5.3.** *Let  $(a_n, b_n)_n$  be a sequence of disjoint open intervals in  $(0, +\infty)$  with  $a_n \rightarrow +\infty$  and assume that (1.3), (1.4), and (2.4) hold. Then for every  $(u_0, u_1) \in D(\Delta^m) \times L^2(\Omega)$  the solution  $u$  of problem (H) is such that  $E_u(t) \rightarrow 0$  as  $t \rightarrow +\infty$ .*

**Acknowledgments.** The authors wish to thank Prof. Hatvani for some interesting discussions on this subject and for bringing [22] to their attention.

## REFERENCES

- [1] F. ALABAU-BOUSSOIRA, *A general formula for decay rates of nonlinear dissipative systems*, C. R. Math. Acad. Sci. Paris, 338 (2004), pp. 35–40.
- [2] F. ALABAU-BOUSSOIRA, *Convexity and weighted integral inequalities for energy decay rates of nonlinear dissipative hyperbolic systems*, Appl. Math. Optim., 51 (2005), pp. 61–105.
- [3] M. BELLASSOUED, *Decay of solutions of the wave equation with arbitrary localized nonlinear damping*, J. Differential Equations, 211 (2005), pp. 303–332.
- [4] S. COX AND E. ZUAZUA, *The rate at which energy decays in a damped string*, Comm. Partial Differential Equations, 19 (1994), pp. 213–243.
- [5] Á. ELBERT, *Stability of some difference equations*, in Advances in Difference Equations, (Veszprém, 1995), Gordon and Breach, Amsterdam, 1997, pp. 165–187.
- [6] A. HARAUX, *On a completion problem in the theory of distributed control of wave equations*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. X (Paris, 1987–1988), Pitman Res. Notes Math. Ser. 220, Pitman, Boston, 1991, pp. 241–271.
- [7] A. HARAUX, P. MARTINEZ, AND J. VANCOSTENOBLE, *Asymptotic stability for intermittently controlled second-order evolution equations*, SIAM J. Control Optim., 43 (2005), pp. 2089–2108.
- [8] L. HATVANI AND T. KRISZTIN, *Necessary and sufficient conditions for intermittent stabilization of linear oscillators by large damping*, Differential Integral Equations, 10 (1997), pp. 265–272.
- [9] I. LASIECKA AND D. TOUNDYKOV, *Energy decay rates for the semilinear wave equation with nonlinear localized damping and source terms*, Nonlinear Anal., 64 (2006), pp. 1757–1797.
- [10] H. A. LEVINE, *Some additional remarks on the nonexistence of global solutions to nonlinear wave equations*, SIAM J. Math. Anal., 5 (1974), pp. 138–146.
- [11] H. A. LEVINE, S. R. PARK, AND J. SERRIN, *Global existence and global nonexistence of solutions of the Cauchy problem for a nonlinearly damped wave equation*, J. Math. Anal. Appl., 228 (1998), pp. 181–205.
- [12] J. LÓPEZ-GÓMEZ, *On the linear damped wave equation*, J. Differential Equations, 134 (1997), pp. 26–45.
- [13] P. MARTINEZ, *Precise decay rate estimates for time-dependent dissipative systems*, Israel J. Math., 119 (2000), pp. 291–324.
- [14] S. A. MESSAOUDI, *Energy decay of solutions of a semilinear wave equation*, Int. J. Appl. Math., 2 (2000), pp. 1037–1048.
- [15] D. MUGNAI, *On a “reversed” variational inequality*, Topol. Methods Nonlinear Anal., 17 (2001), pp. 321–357.
- [16] P. PUCCI AND J. SERRIN, *Asymptotic stability for intermittently controlled nonlinear oscillators*, SIAM J. Math. Anal., 25 (1994), pp. 815–835.
- [17] P. PUCCI AND J. SERRIN, *Asymptotic stability for nonautonomous dissipative wave systems*, Comm. Pure Appl. Math., 49 (1996), pp. 177–216.
- [18] P. PUCCI AND J. SERRIN, *Precise damping conditions for global asymptotic stability for nonlinear second order systems*, Acta Math., 170 (1993), pp. 275–307.
- [19] P. PUCCI AND J. SERRIN, *Precise damping conditions for global asymptotic stability for nonlinear second order systems. II*, J. Differential Equations, 113 (1994), pp. 505–534.
- [20] P. PUCCI AND J. SERRIN, *Stability for abstract evolution equations*, in Partial Differential Equations and Applications, Lecture Notes in Pure and Appl. Math. 177, Dekker, New York, 1996, pp. 279–288.
- [21] R. A. SMITH, *Asymptotic stability of  $x'' + a(t)x' + x = 0$* , Quart. J. Math. Oxford Ser. (2), 12 (1961), pp. 123–126.
- [22] B. ZHANG, *Asymptotic behavior of solutions of a nonlinear damped wave equation*, Differential Equations Dynam. Systems, 2 (1994), pp. 173–204.

## STOCHASTIC DYNAMIC OPTIMIZATION WITH DISCOUNTED STOCHASTIC DOMINANCE CONSTRAINTS\*

DARINKA DENTCHEVA<sup>†</sup> AND ANDRZEJ RUSZCZYŃSKI<sup>‡</sup>

**Abstract.** We introduce a stochastic dynamic optimization problem, where risk aversion is expressed by a stochastic ordering constraint. The constraint requires that a random reward sequence depending on our decisions dominates a given benchmark random sequence. The dominance is defined by discounting both processes with a family of discount sequences, and by applying a univariate order. We describe the generator of this order. We develop necessary and sufficient conditions of optimality for convex stochastic control problems with the new ordering constraint, and we derive an equivalent control problem featuring implied utility functions. Furthermore, we prove the existence of an optimal random discount sequence such that the solution of the risk averse problem is also a solution of an expected value problem with this discount. Finally, we derive a version of the maximum principle for the problem with discounted dominance constraints.

**Key words.** stochastic control, stochastic programming, stochastic orders, risk, utility, maximum principle

**AMS subject classifications.** Primary, 90C15, 90C48, 93E20; Secondary, 46N10, 60E15, 90C34

**DOI.** 10.1137/070679569

**1. Introduction.** In our earlier publications [5, 6], we have introduced and analyzed the following optimization model with stochastic dominance constraints:

$$\begin{aligned} (1.1) \quad & \max \mathbb{E}[H(z)] \\ (1.2) \quad & \text{s.t. } G(z) \succeq_{(2)} Y, \\ (1.3) \quad & z \in Z_0. \end{aligned}$$

In this problem  $Z_0$  is a convex closed subset of a Banach space  $\mathcal{Z}$ , and  $G$  and  $H$  are continuous operators from  $\mathcal{Z}$  to the space of integrable random variables  $\mathcal{L}_1(\Omega, \mathcal{F}, P)$ . The random variable  $Y$  plays the role of a benchmark outcome. For example, one may set  $Y = G(\bar{z})$ , where  $\bar{z} \in Z_0$  is some reasonable value of the decision vector, which is currently employed in the system.

The relation  $\succeq_{(2)}$  used in (1.2) is the *stochastic dominance* relation of the second order. It is defined as follows: A random variable  $X$  *dominates* another random variable  $Y$  in the second order if

$$(1.4) \quad \mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$$

for every concave nondecreasing function  $u(\cdot)$ , for which these expected values are finite. The relation  $\succeq_{(2)}$  is also called an *increasing concave order* (see [24, 33] and

---

\*Received by the editors January 8, 2007; accepted for publication (in revised form) May 17, 2008; published electronically September 17, 2008. This research was supported by the NSF awards DMS-0603728 and DMS-0604060.

<http://www.siam.org/journals/sicon/47-5/67956.html>

<sup>†</sup>Department of Mathematical Sciences, Stevens Institute of Technology, Castle Point on Hudson, Hoboken, NJ 07030 (darinka.dentcheva@stevens.edu).

<sup>‡</sup>Department of Management Science and Information Systems, Rutgers University, 94 Rockefeller Rd., Piscataway, NJ 08854 (rusz@business.rutgers.edu).



the references therein). This relation is used in economics for comparing random outcomes.

Model (1.1)–(1.3) is a convenient way to express risk-aversion in a stochastic optimization problem. It has recently been applied to electricity market models in [2] and to financial optimization in [7]. Also, the idea of a stochastic ordering constraint has been recently borrowed by [10].

Our objective in this paper is to extend our model (1.1)–(1.3) to a dynamic setting, with  $G(z)$  representing a random sequence, rather than a scalar random variable. We are interested in modeling risk aversion in a stochastic control problem for a discrete-time linear dynamic system governed by the equations

$$s_{t+1} = A_t s_t + B_t v_t + e_t, \quad t = 1, \dots, T.$$

Here  $s_t$  denotes the state vector at time  $t$  and  $v_t$  denotes the control vector. The vectors  $e_t$  and the matrices  $A_t$  and  $B_t$  are random. The initial state  $s_1$  is given.

Assume that on the probability space  $(\Omega, \mathcal{F}, P)$  we have a filtration  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \subset \mathcal{F}_{T+1}$ , with  $\mathcal{F}_1 = \{\emptyset, \Omega\}$  and  $\mathcal{F}_{T+1} = \mathcal{F}$ . The  $\sigma$ -field  $\mathcal{F}_t$  is generated by the information available at time  $t$ , when control  $v_t$  is chosen. We assume that  $e_t \in \mathcal{L}_p^{n_s}(\Omega, \mathcal{F}_{t+1}, P)$ ,  $v_t \in \mathcal{L}_p^{n_v}(\Omega, \mathcal{F}_t, P)$ ,  $s_t \in \mathcal{L}_p^{n_s}(\Omega, \mathcal{F}_t, P)$ , with some  $p \in [1, \infty)$ . The matrices  $A_t$  and  $B_t$  are elements of spaces of matrices of appropriate dimensions, which are measurable with respect to  $\mathcal{F}_t$  and essentially bounded. The standard symbol  $\mathcal{L}_p^m(\Omega, \mathcal{F}, P)$  denotes the space of all  $\mathcal{F}$ -integrable mappings  $X : \Omega \rightarrow \mathbb{R}^m$ , for which  $\mathbb{E}\|X\|^p < \infty$ . If the values are taken in  $\mathbb{R}$ , then the superscript  $m$  is omitted.

Specific conditions impose additional constraints on our actions:  $v_t \in V_t$ ,  $P$ -a.s., where each  $V_t$  is a convex closed set in  $\mathbb{R}^{n_v}$ .

Assume that the random outcomes  $X_t$ , representing the performance measures of the system at  $t = 1, \dots, T+1$ , are scalar and given by

$$(1.5) \quad \begin{aligned} X_t(\omega) &= g_t(s_t(\omega), v_t(\omega)), \quad \text{for } t = 1, \dots, T, \\ X_{T+1}(\omega) &= g_{T+1}(s_{T+1}(\omega)), \quad \omega \in \Omega. \end{aligned}$$

The functions  $g_t : \mathbb{R}^{n_s} \times \mathbb{R}^{n_v} \rightarrow \mathbb{R}$  are concave and satisfy the growth condition

$$|g_t(s, v)| \leq C_1 + C_2(\|s\|^p + \|v\|^p),$$

with some constants  $C_1$  and  $C_2$ . If  $p = 1$ , then this condition (for concave functions) amounts to global Lipschitz continuity.

We adopt the convention that larger values of  $X_t$  are preferred; for example,  $X_t$  may represent profits at time  $t$ .

Relations (1.5) define mappings  $G_t : \mathcal{L}_p^{n_s}(\Omega, \mathcal{F}_t, P) \times \mathcal{L}_p^{n_v}(\Omega, \mathcal{F}_t, P) \rightarrow \mathcal{L}_1(\Omega, \mathcal{F}_t, P)$ ,  $t = 1, \dots, T$ , and  $G_{T+1} : \mathcal{L}_p^{n_s}(\Omega, \mathcal{F}_{T+1}, P) \rightarrow \mathcal{L}_1(\Omega, \mathcal{F}_{T+1}, P)$ , in the following way:

$$\begin{aligned} [G_t(s_t, v_t)](\omega) &= g_t(s_t(\omega), v_t(\omega)), \quad t = 1, \dots, T, \\ [G_{T+1}(s_{T+1})](\omega) &= g_{T+1}(s_{T+1}(\omega)), \quad \omega \in \Omega. \end{aligned}$$

We write  $G(s, v) = (G_1(s_1, v_1), \dots, G_T(s_T, v_T), G_{T+1}(s_{T+1}))$ . Here  $s = (s_1, \dots, s_{T+1})$  and  $v = (v_1, \dots, v_T)$  represent the variables of our problem.

In the simplest formulation of the problem the objective functional is just the expected value; the problem reads

$$(1.6) \quad \begin{aligned} & \max \sum_{t=1}^T \mathbb{E} G_t(s_t, v_t) + \mathbb{E} G_{T+1}(s_{T+1}) \\ & \text{s.t. } s_{t+1} = A_t s_t + B_t v_t + e_t, \quad t = 1, \dots, T, \\ & \quad v_t \in V_t \text{ a.s., } \quad t = 1, \dots, T. \end{aligned}$$

Owing to the growth condition on  $g_t(\cdot, \cdot)$ , the objective functional is finite and continuous everywhere.

Our goal is to model risk aversion in this problem by using stochastic orders. To this end we compare the multivariate distribution of the rewards  $(X_1, X_2, \dots, X_{T+1})$  with the distribution of some benchmark outcomes  $(Y_1, Y_2, \dots, Y_{T+1})$ . We shall add to the problem formulation an appropriate stochastic ordering constraint

$$(1.7) \quad (X_1, X_2, \dots, X_{T+1}) \succeq (Y_1, Y_2, \dots, Y_{T+1}).$$

In section 2 we specify the multivariate stochastic ordering constraint (1.7) as a discounted stochastic dominance and construct a generator of this order. In section 3 we introduce and analyze a control problem with the discounted stochastic dominance constraint and develop necessary and sufficient conditions of optimality by deriving an equivalent formulation featuring implied utility functions. These results refine and generalize to the dynamic case the optimality conditions obtained in [5, 6, 8]. In section 4, we derive the existence of random discount factors in the necessary conditions of optimality. Finally, in section 5 we prove a version of the maximum principle for our model.

Let us introduce some notation used throughout this paper. The space of continuous functions on a compact set  $D \subset \mathbb{R}^n$  is denoted  $\mathcal{C}(D)$ . The space of regular countably additive measures on a compact set  $D \subset \mathbb{R}^n$  having finite variation is denoted  $\mathcal{M}(D)$ ; its subset of nonnegative measures is denoted by  $\mathcal{M}_+(D)$ . For a Banach space  $\mathcal{Z}$  we denote its topological dual by  $\mathcal{Z}^*$ . We denote the space of the random outcomes by  $\mathcal{X} = \mathcal{L}_1(\Omega, \mathcal{F}_1, P) \times \dots \times \mathcal{L}_1(\Omega, \mathcal{F}_{T+1}, P)$ . Its dual is  $\mathcal{X}^* = \mathcal{L}_\infty(\Omega, \mathcal{F}_1, P) \times \dots \times \mathcal{L}_\infty(\Omega, \mathcal{F}_{T+1}, P)$ .

**2. Stochastic dominance for random reward sequences.** The notion of stochastic ordering for scalar random variables (or *stochastic dominance of first order*) has been introduced in statistics in [20, 17] and further applied and developed in economics [26, 11, 12]. It is defined as follows. For a random variable  $X$  we consider its distribution function,  $F(X; \eta) = P[X \leq \eta]$ ,  $\eta \in \mathbb{R}$ . We say that a random variable  $X$  *dominates in the first order* a random variable  $Y$  if

$$(2.1) \quad F(X; \eta) \leq F(Y; \eta) \quad \text{for all } \eta \in \mathbb{R}.$$

We denote this relation by  $X \succeq_{(1)} Y$ . We refer the reader to [21, 24, 33, 35] for a detailed treatment of stochastic ordering.

Consider a scalar random variable  $X \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$  and define the function

$$(2.2) \quad F_2(X; \eta) = \int_{-\infty}^{\eta} F(X; \alpha) d\alpha \quad \text{for } \eta \in \mathbb{R}.$$

As an integral of a nondecreasing function, it is a convex function of  $\eta$ .

DEFINITION 2.1. A random variable  $X \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$  dominates in the second order another random variable  $Y \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$  if

$$(2.3) \quad F_2(X; \eta) \leq F_2(Y; \eta) \quad \text{for all } \eta \in \mathbb{R}.$$

We denote relation (2.3) by  $X \succeq_{(2)} Y$ . In a similar way, we can define higher order dominance relations (see [24]).

The stochastic dominance relation  $\succeq_{(2)}$  introduces a preorder among integrable random variables (see, e.g., [21] and the references therein). Partial orders appear in abstract optimization problems when the values of the objective operator are elements of a topological vector space (see, e.g., [19]). It is usually assumed that the partial order is generated by a convex cone. The stochastic dominance relations in  $\mathcal{L}_1(\Omega, \mathcal{F}, P)$  are not generated by cones in this space therefore we cannot follow the theory in the spirit of [19].

Changing the order of integration in (2.2), we get (see, e.g., [25])

$$(2.4) \quad F_2(X; \eta) = \mathbb{E}[(\eta - X)_+].$$

Therefore, an equivalent representation of the second order stochastic dominance relation is

$$(2.5) \quad \mathbb{E}[(\eta - X)_+] \leq \mathbb{E}[(\eta - Y)_+] \quad \text{for all } \eta \in \mathbb{R}.$$

Let us consider the set  $\mathcal{U}$  of concave nondecreasing functions  $u : \mathbb{R} \rightarrow \mathbb{R}$  satisfying the following conditions:

$$(2.6) \quad \begin{aligned} \lim_{t \rightarrow -\infty} u(t)/t &< \infty, \\ \lim_{t \rightarrow \infty} u(t) &= 0. \end{aligned}$$

With every  $u \in \mathcal{U}$  we can associate a measure  $\nu$  on  $\mathbb{R}$  as follows:  $\nu([\tau, \infty)) = u'_-(\tau)$ , where  $u'_-$  is the left derivative of  $u$ . We can represent  $u(t)$  as follows:

$$(2.7) \quad \begin{aligned} u(t) &= - \int_t^\infty u'_-(\tau) d\tau = - \int_t^\infty \nu([\tau, \infty)) d\tau \\ &= - \int_t^\infty \int_\tau^\infty \nu(d\eta) d\tau = - \int_t^\infty \int_t^\eta d\tau \nu(d\eta) \\ &= - \int_t^\infty (\eta - t) \nu(d\eta) = - \int_{-\infty}^\infty \max(0, \eta - t) \nu(d\eta). \end{aligned}$$

By (2.6), for every random variable  $X \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$  and for every  $u \in \mathcal{U}$  the quantity

$$\mathbb{E}[u(X)] = \int u(X(\omega)) P(d\omega)$$

is well defined and finite. The following fact is well known in the theory of stochastic dominance.

PROPOSITION 2.2. For each  $X, Y \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$  the relation  $X \succeq_{(2)} Y$  is equivalent to

$$(2.8) \quad \mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)] \quad \text{for all } u \in \mathcal{U}.$$

Consider random vectors  $(X_1, \dots, X_{T+1})$  and  $(Y_1, \dots, Y_{T+1})$  in  $\mathcal{L}_1^{T+1}(\Omega, \mathcal{F}, P)$ . The simplest way to define a stochastic ordering relation  $X \succeq_{(2)}^{\text{sep}} Y$  between these vectors, is to require the stochastic dominance relation for each coordinate

$$(2.9) \quad X_t \succeq_{(2)} Y_t, \quad t = 1, \dots, T+1.$$

The analysis in our earlier paper [6] includes this case. This approach, however, ignores the temporal structure and the dependency between the coordinates of the vector  $(X_1, \dots, X_{T+1})$ .

Therefore, we are taking a different approach, by considering discounted sums of the rewards,  $\sum_{t=1}^{T+1} \varrho_t X_t$ , and the corresponding discounted sums of the benchmark  $\sum_{t=1}^{T+1} \varrho_t Y_t$ . The sequence of discount factors  $\{\varrho_t\}$  is assumed to belong to a compact set  $D$ , where

$$D \subseteq \{\varrho \in \mathbb{R}^{T+1} : 1 \geq \varrho_1 \geq \varrho_2 \geq \dots \geq \varrho_{T+1} \geq 0\}.$$

DEFINITION 2.3. A random sequence  $(X_1, \dots, X_{T+1}) \in \mathcal{X}$  dominates a random sequence  $(Y_1, \dots, Y_{T+1}) \in \mathcal{X}$  in the discounted second order, if for all  $\varrho \in D$  the relation

$$(2.10) \quad \sum_{t=1}^{T+1} \varrho_t X_t \succeq_{(2)} \sum_{t=1}^{T+1} \varrho_t Y_t$$

is satisfied.

We denote this relation by  $X \succeq_{(2)}^D Y$ .

The example below shows that the discounted order  $\succeq_{(2)}^D$  neither implies nor is implied by the coordinate order (2.9).

Example 1. Consider the special case with finite set  $D$ , consisting of  $T+1$  elements  $\varrho^k$ ,  $k = 1, \dots, T+1$ , with

$$(2.11) \quad \varrho_t^k = \begin{cases} 1 & \text{if } t \leq k, \\ 0 & \text{if } t > k. \end{cases}$$

Then the relation  $X \succeq_{(2)}^D Y$  is equivalent to the following system of dominance relations:

$$(2.12) \quad \sum_{t=1}^k X_t \succeq_{(2)} \sum_{t=1}^k Y_t, \quad k = 1, \dots, T+1.$$

Consider now three random sequences:  $X$ ,  $Y$ , and  $Z$ , and let  $\varepsilon > 0$ . The sequence  $Y$  has i.i.d. components  $Y_t$ ,  $t = 1, \dots, T$  with expected value  $\mu$  and variance  $\sigma^2$ , the sequence  $X$  has components  $X_t = Y_1 + \varepsilon$ ,  $t = 1, \dots, T$ , where  $\varepsilon > 0$ , and the sequence  $Z$  has components  $Z_1 = Y_1 + \varepsilon$ ,  $Z_2 = Y_2 - \varepsilon$ ,  $Z_t = Y_t$ ,  $t = 3, \dots, T$ .

It is obvious that  $X_t \succeq_{(2)} Y_t$ ,  $t = 1, \dots, T$ . However,  $X \not\succeq_{(2)}^D Y$ , if  $F_2(Y_1; \mu - \varepsilon) > 0$ , and  $T$  is large enough. Indeed, for every  $k$  we have to verify relation (2.12). As

the second order stochastic dominance relation is preserved under multiplication by positive constants, we need to compare

$$Y_1 + \varepsilon \quad \text{and} \quad \frac{1}{k} \sum_{t=1}^k Y_t.$$

At  $\eta = \mu$  we have

$$F_2 \left( \frac{1}{k} \sum_{t=1}^k X_t; \mu \right) = \mathbb{E} \left[ (\mu - Y_1 - \varepsilon)_+ \right] = F_2(Y_1; \mu - \varepsilon) > 0.$$

On the other hand, by the central limit theorem

$$\lim_{k \rightarrow \infty} F_2 \left( \frac{1}{k} \sum_{t=1}^k Y_t; \mu \right) = \mathbb{E} \left[ \left( \mu - \frac{1}{k} \sum_{t=1}^k Y_t \right)_+ \right] = 0.$$

Therefore, for  $T$  and  $k$  large enough, we shall have

$$\sum_{t=1}^k X_t \not\prec_{(2)} \sum_{t=1}^k Y_t.$$

Consider now the sequences  $Z$  and  $Y$ . Clearly,  $Z_2 \not\prec_{(2)} Y_2$ . However, directly from (2.12) we see that  $Z \succeq_{(2)}^D Y$ .

Suppose, for simplicity, that all  $Y_t$  are normal, and let the sequence  $W$  have components  $W_1 = \mu + \alpha(Y_1 - \mu)$ ,  $W_2 = \mu + \beta(Y_2 - \mu)$ ,  $W_t = Y_t$ ,  $t = 3, \dots, T$ , where  $\alpha \in (0, 1)$ ,  $\beta \in (1, \sqrt{2 - \alpha^2})$ . Again, it is easy to see that  $W_2 \not\prec_{(2)} Y_2$ , since  $W_2$  and  $Y_2$  have equal expected values, but  $W_2$  has a larger variance than  $Y_2$ . Still, we can verify (2.12) under the specified assumptions, and establish that  $W \succeq_{(2)}^D Y$ .

In fact, the relations  $Z \succeq_{(2)}^D Y$  and  $W \succeq_{(2)}^D Y$  hold true for every set  $D$  such that  $q_1$  is bounded from below. This concludes the example.

A very special case occurs when  $D = \{(1, 1, \dots, 1)\}$ . In this case we are interested in the total terminal wealth and require that it dominates (in the second order) some benchmark wealth distribution. The information about the whole process  $Y$  is irrelevant in this case, only the distribution of the sum  $Y_1 + \dots + Y_{T+1}$  matters.

The question of orderings of linear or positive combinations of random vectors was investigated in several earlier publications [23, 22, 32]. One approach to defining a multivariate stochastic order is to specify its generator. This approach is adopted in [23], where further conditions are specified under which linear combinations of the vector components are related via a univariate order. In [23], the concept of a directional convex order for random vectors is introduced, by specifying that the generator contains all directionally convex functions. The directional convex order of two vectors implies a convex order of nonnegative linear combinations of the components of the vectors. In a similar way one can also derive a directional concave *nondecreasing* order, which will imply stochastic dominance of nonnegative linear combinations. Our approach is different: we define the order by requiring stochastic dominance of nonnegative combinations and then we derive its generator.

Example 1 shows that the coordinate order  $X_i \succeq_{(2)} Y_i$ ,  $i = 1, \dots, n$ , does not carry over to the multivariate order  $X \succeq_{(2)}^D Y$ . The directional convex order is implied by the coordinate order if both vectors have a common copula of a certain type

(see [23]). A similar implication holds true for our order  $\succeq_{(2)}^D$  if we know that all random reward vectors occurring in our optimization problem have a common copula (of this type) with the benchmark. In this case, we would be able to replace the multivariate dominance constraint by univariate constraints for the marginals. This would allow solving the problem by the techniques of [5, 6]. Unfortunately, imposing a constraint on a copula in an optimization problem appears to be difficult. Without that, constraints on the marginals are insufficient.

Inverse versions of multivariate orders (based on generalizations of the Lorenz curve) were analyzed in [1, 15, 16].

For brevity we write

$$\langle \varrho, X \rangle = \sum_{t=1}^{T+1} \varrho_t X_t.$$

By virtue of Proposition 2.2 we obtain the following relation:

$$(2.13) \quad \mathbb{E}[u(\langle \varrho, X \rangle)] \geq \mathbb{E}[u(\langle \varrho, Y \rangle)] \quad \text{for all } u \in \mathcal{U} \text{ and all } \varrho \in D.$$

Using (2.7), we conclude that for every nonnegative measure  $\nu$  on  $\mathbb{R}$  and for every  $\varrho \in D$

$$\int_{\Omega} \int_{\mathbb{R}} \max(0, \eta - \langle \varrho, X(\omega) \rangle) \nu(d\eta) P(d\omega) \leq \int_{\Omega} \int_{\mathbb{R}} \max(0, \eta - \langle \varrho, Y(\omega) \rangle) \nu(d\eta) P(d\omega).$$

For every  $\lambda \in \mathcal{M}_+(\mathbb{R} \times D)$  we define a concave nondecreasing function  $\varphi_{\lambda} : \mathbb{R}^{T+1} \rightarrow \mathbb{R}$  as follows:

$$\varphi_{\lambda}(x) = - \int_{\mathbb{R} \times D} \max(0, \eta - \langle \varrho, x \rangle) \lambda(d\eta, d\varrho).$$

We now show that the class of functions

$$\Phi = \{\varphi_{\lambda} : \lambda \in \mathcal{M}_+(\mathbb{R} \times D)\}$$

is a generator of the order  $\succeq_{(2)}^D$ .

PROPOSITION 2.4. *For each  $X, Y \in \mathcal{X}$  the relation  $X \succeq_{(2)}^D Y$  is equivalent to*

$$(2.14) \quad \mathbb{E}[\varphi(X)] \geq \mathbb{E}[\varphi(Y)] \quad \text{for all } \varphi \in \Phi.$$

*Proof.* Assume  $X \succeq_{(2)}^D Y$ . Take any  $\lambda \in \mathcal{M}_+(\mathbb{R} \times D)$ . We shall show inequality (2.14) for  $\varphi_{\lambda}$ . Let  $\lambda_{\varrho}$  be the conditional measure of  $\lambda$  on  $\mathbb{R}$ , for  $\varrho \in D$  (see, e.g., [9, Theorem 10.2.2]). Denote by  $\mu$  the marginal measure of  $\lambda$  on  $D$ . The integral over  $\mathbb{R} \times D$  can be written as an iterated integral (see, e.g., [9, Theorem 10.21])

$$\varphi_{\lambda}(x) = - \int_D \int_{\mathbb{R}} \max(0, \eta - \langle \varrho, x \rangle) \lambda_{\varrho}(d\eta) \mu(d\varrho).$$

From the definition of  $\succeq_{(2)}^D$  and (2.13) it follows that

$$\int_{\Omega} \int_{\mathbb{R}} \max(0, \eta - \langle \varrho, X(\omega) \rangle) \lambda_{\varrho}(d\eta) P(d\omega) \leq \int_{\Omega} \int_{\mathbb{R}} \max(0, \eta - \langle \varrho, Y(\omega) \rangle) \lambda_{\varrho}(d\eta) P(d\omega).$$

Integrating with the measure  $\mu$  and changing the order of integration, we obtain

$$\begin{aligned}\mathbb{E}[\varphi_\lambda(X)] &= - \int_{\Omega} \int_D \int_{\mathbb{R}} \max(0, \eta - \langle \varrho, X(\omega) \rangle) \lambda_{\varrho}(d\eta) \mu(d\varrho) P(d\omega) \\ &\geq - \int_{\Omega} \int_D \int_{\mathbb{R}} \max(0, \eta - \langle \varrho, Y(\omega) \rangle) \lambda_{\varrho}(d\eta) \mu(d\varrho) P(d\omega) = \mathbb{E}[\varphi_\lambda(Y)],\end{aligned}$$

as required. To prove the converse, we observe that we can choose measures  $\lambda$  such that their marginal measures  $\mu$  on  $D$  are atomic. The last displayed inequality becomes equivalent to the inequality in (2.13). The latter is equivalent to the definition of the order  $\succeq_{(2)}^D$ .  $\square$

It is clear from our analysis that an order can be defined using any compact set  $D \subset \mathbb{R}_+^{T+1}$ , and the generator of this order will have the form specified in Proposition 2.4. However, discounting of future rewards is a standard approach in stochastic control and in the practice of finance, and this is the reason for requiring the monotonicity of the sequences  $\varrho$  included in  $D$ .

Another way to characterize the relation  $\succeq_{(2)}^D$  is to use the integrated distribution functions. From (2.3) we obtain the equivalent characterization:

$$(2.15) \quad F_2(\langle \varrho, X \rangle; \eta) \leq F_2(\langle \varrho, Y \rangle; \eta) \quad \text{for all } \varrho \in D \text{ and all } \eta \in \mathbb{R}.$$

**3. Optimality conditions. The implied utility functions.** We introduce the following stochastic dynamic optimization problem with discounted dominance constraints:

$$\begin{aligned}(3.1) \quad & \max \sum_{t=1}^T \mathbb{E}G_t(s_t, v_t) + \mathbb{E}G_{T+1}(s_{T+1}) \\ & \text{s.t. } s_{t+1} = A_t s_t + B_t v_t + e_t, \quad t = 1, \dots, T, \\ & (G_1(s_1, v_1), \dots, G_T(s_T, v_T), G_{T+1}(s_{T+1})) \succeq_{(2)}^D (Y_1, \dots, Y_T, Y_{T+1}) \\ & v_t \in V_t \text{ a.s., } \quad t = 1, \dots, T.\end{aligned}$$

Define the space of controls  $(v_1, \dots, v_T)$  by  $\mathcal{V} = \mathcal{L}_p^{n_v}(\Omega, \mathcal{F}_1, P) \times \dots \times \mathcal{L}_p^{n_v}(\Omega, \mathcal{F}_T, P)$ . The space of state trajectories  $(s_2, \dots, s_{T+1})$  is denoted by  $\mathcal{S} = \mathcal{L}_p^{n_s}(\Omega, \mathcal{F}_2, P) \times \dots \times \mathcal{L}_p^{n_s}(\Omega, \mathcal{F}_{T+1}, P)$ .

The relation  $X \succeq_{(2)}^D Y$  can be equivalently formulated as (2.15). For technical reasons, which will become apparent later, we restrict the range of  $\eta \in \mathbb{R}$ , for which we impose (2.15), to an interval  $[a, b]$ . This slightly changes the generator of the relation, which will be discussed in due course. We relax problem (3.1) to the following:

$$\begin{aligned}(3.2) \quad & \max \sum_{t=1}^T \mathbb{E}G_t(s_t, v_t) + \mathbb{E}G_{T+1}(s_{T+1}) \\ (3.3) \quad & \text{s.t. } s_{t+1} = A_t s_t + B_t v_t + e_t, \quad t = 1, \dots, T, \\ (3.4) \quad & F_2(\langle \varrho, G(s, v) \rangle; \eta) \leq F_2(\langle \varrho, Y \rangle; \eta) \text{ for all } \varrho \in D \text{ and all } \eta \in [a, b], \\ (3.5) \quad & v_t \in V_t \text{ a.s., } \quad t = 1, \dots, T.\end{aligned}$$

If all  $G_t(s_t, v_t)$  have uniformly bounded distributions, (3.4) is equivalent to  $X \succeq_{(2)}^D Y$  for appropriately chosen  $a$  and  $b$ . However, if the distributions are not uniformly bounded, (3.4) is a relaxation of the relation  $X \succeq_{(2)}^D Y$ . This means that every feasible solution of problem (3.1) is also feasible for (3.2)–(3.5), but not necessarily the other way around. Even with this relaxation the problem remains very difficult. In the very special case when  $a = b$ , problem (3.2)–(3.5) is still much more difficult than an expected value problem. Observe that if  $D = \{(1, \dots, 1)\}$ , constraint (3.4) is a conditional value at risk constraint for the cumulative wealth.

We define the set  $\mathcal{U}([a, b])$  of functions  $u(\cdot)$  satisfying the following conditions:

$u(\cdot)$  is concave and nondecreasing;

$u(t) = 0$  for all  $t \geq b$ ;

$u(t) = u(a) + \gamma(t - a)$  for all  $t \leq a$ , where  $\gamma > 0$ .

It is evident that  $\mathcal{U}([a, b])$  is a convex cone. Moreover, the subgradients of each function  $u \in \mathcal{U}([a, b])$  are bounded for all  $t \in \mathbb{R}$ .

Define the class of functions

$$\Phi([a, b], D) = \{\varphi_\lambda : \lambda \in \mathcal{M}_+([a, b] \times D)\},$$

where

$$\varphi_\lambda(x) = - \int_{[a, b] \times D} \max(0, \eta - \langle \varrho, x \rangle) \lambda(d\eta, d\varrho).$$

By exactly the same argument as Proposition 2.4 we obtain the following observation.

**PROPOSITION 3.1.** *The set  $\Phi([a, b], D)$  is a generator of the order (3.4).*

We introduce the functional  $L : \mathcal{S} \times \mathcal{V} \times \Phi([a, b], D) \rightarrow \mathbb{R}$ , which plays the role of a partial Lagrangian associated with problem (3.2)–(3.5):

$$\begin{aligned} L(s, v, \varphi) = \mathbb{E} & \left[ \sum_{t=1}^T G_t(s_t, v_t) + G_{T+1}(s_{T+1}) \right. \\ & \left. + \left( \varphi(G_1(s_1, v_1), \dots, G_T(s_T, v_T), G_{T+1}(s_{T+1})) - \varphi(Y_1, \dots, Y_T, Y_{T+1}) \right) \right]. \end{aligned}$$

For a fixed  $\varphi(\cdot)$ , we use  $L$  as an objective functional in an auxiliary control problem:

$$(3.6) \quad \max \mathbb{E} \left[ \sum_{t=1}^T G_t(s_t, v_t) + G_{T+1}(s_{T+1}) \right. \\ \left. + \left( \varphi(G_1(s_1, v_1), \dots, G_T(s_T, v_T), G_{T+1}(s_{T+1})) - \varphi(Y_1, \dots, Y_T, Y_{T+1}) \right) \right]$$

$$(3.7) \quad \text{s.t. } s_{t+1} = A_t s_t + B_t v_t + e_t, \quad t = 1, \dots, T,$$

$$(3.8) \quad v_t \in V_t \text{ a.s., } \quad t = 1, \dots, T.$$

Let  $Z_0$  denote the convex set of  $(s, v)$  satisfying conditions (3.7)–(3.8). The following property plays the role of a constraint qualification condition.



DEFINITION 3.2. *Problem (3.2)–(3.5) satisfies the uniform dominance condition if there exists a pair  $(\tilde{s}, \tilde{v}) \in Z_0$  such that*

$$\inf_{(\eta, \varrho) \in [a, b] \times D} \left\{ F_2(\langle \varrho, Y \rangle; \eta) - F_2(\langle \varrho, G(\tilde{s}, \tilde{v}) \rangle; \eta) \right\} > 0.$$

This condition is the reason for considering constraints (3.3) in the finite interval  $[a, b]$ . A constraint qualification of Slater type cannot be satisfied for  $\eta \in \mathbb{R}$  since for any random variable  $X$  the function  $F_2(X; \eta)$  converges to zero, whenever  $\eta \rightarrow -\infty$ .

*Example 2.* Consider any compact set  $D$  such that  $\varrho_1$  is uniformly bounded from below by a positive number for  $\varrho \in D$ . Consider the sequence  $Y$ , where all  $Y_t$  are independent and have normal distribution with mean  $\mu$  and variance  $\sigma^2 > 0$ . Let the sequence  $W = G(\tilde{s}, \tilde{v})$  have components  $W_1 = \mu + \alpha(Y_1 - \mu)$ ,  $W_2 = \mu + \beta(Y_2 - \mu)$ ,  $W_t = Y_t$ ,  $t = 3, \dots, T$ , where  $\alpha \in (0, 1)$ ,  $\beta \in (1, \sqrt{2 - \alpha^2})$ . We considered this pair in Example 1 and know that  $W \succeq_{(2)}^D Y$ . For every bounded interval  $[a, b]$  the random variable  $W$  satisfies the uniform dominance condition, because for all  $\varrho \in D$

$$\mathbb{E}\langle \varrho, W \rangle = \mathbb{E}\langle \varrho, Y \rangle \quad \text{and} \quad \text{Var}\langle \varrho, W \rangle < \text{Var}\langle \varrho, Y \rangle$$

and all random variables involved are normal. Observe that strict dominance would not hold if the interval  $[a, b]$  was replaced by an infinite interval. This concludes the example.

THEOREM 3.3. *Assume that the uniform dominance condition is satisfied. If  $(\hat{s}, \hat{v})$  is an optimal solution of (3.2)–(3.5), then there exist  $\hat{\varphi} \in \Phi([a, b], D)$  such that  $(\hat{s}, \hat{v})$  is an optimal solution of problem (3.6)–(3.8) with  $\varphi = \hat{\varphi}$ , and*

$$(3.9) \quad \mathbb{E}[\hat{\varphi}(G(\hat{s}, \hat{v}))] = \mathbb{E}[\hat{\varphi}(Y)].$$

*Conversely, if for some function  $\hat{\varphi} \in \Phi([a, b], D)$  an optimal solution  $(\hat{s}, \hat{v})$  of (3.6)–(3.8) satisfies (3.4) and (3.9), then  $(\hat{s}, \hat{v})$  is an optimal solution of (3.2)–(3.5).*

*Proof.* Let us rewrite (3.2)–(3.5) in the general form

$$\max \sum_{t=1}^T \mathbb{E}G_t(s_t, v_t) + \mathbb{E}G_{T+1}(s_{T+1})$$

$$\text{s.t. } \Gamma(s, v) \in K,$$

$$(s, v) \in Z_0,$$

where  $\Gamma : \mathcal{S} \times \mathcal{V} \rightarrow \mathcal{C}([a, b] \times D)$  is a continuous operator defined as

$$[\Gamma(s, v)](\eta, \varrho) = F_2(\langle \varrho, Y \rangle; \eta) - F_2(\langle \varrho, G(s, v) \rangle; \eta), \quad \eta \in [a, b], \quad \varrho \in D.$$

The set  $K$  is the cone of nonnegative functions in  $\mathcal{C}([a, b] \times D)$ . Observe that for every  $\varrho \in D$  the function  $(s, v) \rightarrow \eta - \langle \varrho, G(s, v) \rangle$  is convex for almost all  $\omega \in \Omega$ , and the function  $x \rightarrow (x)_+$  is convex and nondecreasing. Therefore, the composition

$$F_2(\langle \varrho, G(s, v) \rangle; \eta) = \mathbb{E}[(\eta - \langle \varrho, G(s, v) \rangle)_+]$$

is a convex function of  $(s, v)$ . It follows that the operator  $\Gamma$  is concave with respect to the cone  $K$ , that is, for any  $(s^1, v^1), (s^2, v^2)$  in  $Z_0$  and all  $\alpha \in [0, 1]$ ,

$$\Gamma(\alpha s^1 + (1 - \alpha)s^2, \alpha v^1 + (1 - \alpha)v^2) - [\alpha \Gamma(s^1, v^1) + (1 - \alpha)\Gamma(s^2, v^2)] \in K.$$

By the Riesz representation theorem, the dual space to  $\mathcal{C}([a, b] \times D)$  is the space  $\mathcal{M}([a, b] \times D)$ . We introduce the Lagrangian  $\Lambda : \mathcal{S} \times \mathcal{V} \times \mathcal{M}_+([a, b] \times D) \rightarrow \mathbb{R}$ ,

$$(3.10) \quad \Lambda(s, v, \lambda) = \sum_{t=1}^T \mathbb{E} G_t(s_t, v_t) + \mathbb{E} G_{T+1}(s_{T+1}) + \int_{[a, b] \times D} [\Gamma(s, v)](\eta, \varrho) \lambda(d\eta, d\varrho).$$

Let us observe that the uniform dominance condition implies that the following generalized Slater condition is satisfied. There exists a point  $(\tilde{s}, \tilde{v}) \in Z_0$  such that

$$\Gamma(\tilde{s}, \tilde{v}) \in \text{int } K.$$

Moreover,  $(\tilde{s}, \tilde{v}) \in Z_0$ . By [3, Proposition 2.106], this is equivalent to the regularity condition

$$0 \in \text{int}[\Gamma(Z_0) - K].$$

Therefore we can use the necessary conditions of optimality in Banach spaces (see, e.g., [3, Theorem 4]). We conclude that there exists a measure  $\hat{\lambda} \in \mathcal{M}_+([a, b] \times D)$  such that

$$(3.11) \quad \Lambda(\hat{s}, \hat{v}, \hat{\lambda}) = \max_{(s, v) \in Z_0} \Lambda(s, v, \hat{\lambda})$$

and

$$(3.12) \quad \int_{[a, b] \times D} [F_2(\langle \varrho, Y \rangle; \eta) - F_2(\langle \varrho, G(\hat{s}, \hat{v}) \rangle; \eta)] \hat{\lambda}(d\eta, d\varrho) = 0.$$

We shall transform these conditions to the postulated form. Using the representation of  $F_2$  as expected shortfall and changing the order of integration, we obtain

$$\begin{aligned} \int_{[a, b] \times D} F_2(\langle \varrho, Y \rangle; \eta) \hat{\lambda}(d\eta, d\varrho) &= \int_{[a, b] \times D} \int_{\Omega} \max(0, \eta - \langle \varrho, Y(\omega) \rangle) P(d\omega) \hat{\lambda}(d\eta, d\varrho) \\ &= \int_{\Omega} \int_{[a, b] \times D} \max(0, \eta - \langle \varrho, Y(\omega) \rangle) \hat{\lambda}(d\eta, d\varrho) P(d\omega) = -\mathbb{E} \hat{\varphi}(Y). \end{aligned}$$

We have set

$$(3.13) \quad \hat{\varphi}(x) = \varphi_{\hat{\lambda}}(x) = - \int_{[a, b] \times D} \max(0, \eta - \langle \varrho, x \rangle) \hat{\lambda}(d\eta, d\varrho).$$

Clearly,  $\hat{\varphi} \in \Phi([a, b] \times D)$ . Thus, condition (3.11) implies the optimality in problem (3.6)–(3.8), and condition (3.12) implies (3.9).

Let us now prove the converse. If  $\hat{\varphi} \in \Phi([a, b] \times D)$ , then there exist a measure  $\hat{\lambda} \in \mathcal{M}_+([a, b] \times D)$  such that

$$\hat{\varphi}(x) = - \int_{[a, b] \times D} \max(0, \eta - \langle \varrho, x \rangle) \hat{\lambda}(d\eta, d\varrho)$$

and therefore

$$\mathbb{E} \hat{\varphi}(Y) = - \int_{[a, b] \times D} F_2(\langle \varrho, Y \rangle; \eta) \hat{\lambda}(d\eta, d\varrho).$$

Thus, the maximizer  $(\hat{s}, \hat{v})$  of problem (3.6)–(3.8) is also the maximizer of  $\Lambda(s, v, \hat{\lambda})$ . It follows from sufficient conditions of optimality (see, e.g., [3, Proposition 3.3]) that if  $(\hat{s}, \hat{v})$  satisfies (3.3) and (3.9), then it is optimal for (3.2)–(3.5).  $\square$

The main result of this section is an equivalent formulation with an implied utility function  $\hat{\varphi}(\cdot)$  in (3.6). In general, it cannot be decomposed into terms corresponding to successive time periods.

We return to the special case of a finite set  $D$  defined by (2.11). The integral with respect to  $\hat{\lambda}$  can be written as an iterated integral with respect to its marginal measure  $\mu$  in  $D$  and the conditional measures  $\lambda_{\varrho}$  on  $[a, b]$ . Denoting  $\mu_t = \mu(\{\varrho^t\})$ , we obtain

$$\hat{\varphi}(x) = - \sum_{t=1}^{T+1} \mu_t \int_a^b \max(0, \eta - (x_1 + \cdots + x_t)) \lambda_{\varrho^t}(d\eta) = \sum_{t=1}^{T+1} u_t(x_1 + \cdots + x_t),$$

where  $u_t(s) = \mu_t \int_a^b \max(0, \eta - s) \lambda_{\varrho^t}(d\eta)$  is an element of  $\mathcal{U}([a, b])$ . It follows that the resulting optimal control problem has the form

$$\begin{aligned} \max \quad & \sum_{t=1}^T \mathbb{E}(G_t(s_t, v_t) + u_t(G_1(s_1, v_1) + \cdots + G_t(s_t, v_t))) \\ & + \mathbb{E}G_{T+1}(s_{T+1}) + \mathbb{E}u_{T+1}(G_1(s_1, v_1) + \cdots + G_T(s_T, v_T) + G_{T+1}(s_{T+1})) \\ \text{s.t.} \quad & s_{t+1} = A_t s_t + B_t v_t + e_t, \quad t = 1, \dots, T, \\ & v_t \in V_t \text{ a.s.,} \quad t = 1, \dots, T. \end{aligned}$$

We observe that the objective functional (1.6) is modified by adding to each term an expected utility of the cumulative reward up to time  $t$ . If we add a state variable  $W_t$  representing the cumulative reward up to time  $t$ , then the utility becomes decomposable.

**4. The implied random discount.** We can now apply techniques of convex optimization to analyze the auxiliary control problem (3.6)–(3.8). For related expected value models (without the implied utility function  $\varphi(\cdot)$ ), see [14, 30, 31].

Our goal in this section is to demonstrate the existence of random discount factors in (1.6) such that the optimal solution of (3.1) is also optimal for the discounted expected value problem.

After skipping the term depending on  $Y$ , problem (3.6)–(3.8) can be compactly written as follows:

$$(4.1) \quad \max_{(s,v) \in Z_0} \mathbb{E} \left[ \sum_{t=1}^T G_t(s_t, v_t) + G_{T+1}(s_{T+1}) + \varphi(G_1(s_1, v_1), \dots, G_T(s_T, v_T), G_{T+1}(s_{T+1})) \right].$$

**THEOREM 4.1.** *Assume that the uniform dominance condition is satisfied. If  $(\hat{s}, \hat{v})$  is an optimal solution of (3.2)–(3.5), then there exist  $\xi_t \in \mathcal{L}_{\infty}(\Omega, \mathcal{F}_t, P)$ ,  $t = 1, \dots, T+1$ , with*

$$(4.2) \quad \xi_1 \geq \xi_2 \geq \cdots \geq \xi_T \geq \xi_{T+1} \geq 0, \quad \text{a.s.,}$$

such that  $(\hat{s}, \hat{v})$  is an optimal solution of the control problem

$$\begin{aligned}
 (4.3) \quad & \max \sum_{t=1}^T \mathbb{E}(1 + \xi_t)G_t(s_t, v_t) + \mathbb{E}(1 + \xi_{T+1})G_{T+1}(s_{T+1}) \\
 & \text{s.t. } s_{t+1} = A_t s_t + B_t v_t + e_t, \quad t = 1, \dots, T, \\
 & v_t \in V_t \text{ a.s., } \quad t = 1, \dots, T.
 \end{aligned}$$

*Proof.* By virtue of Theorem 3.3, the optimal solution of problem (3.2)–(3.5) is also a solution of problem (4.1) where  $Z_0$  is the set of  $(s, v)$  satisfying conditions (3.7)–(3.8). As the function  $\varphi(\cdot)$  is nondecreasing, problem (4.1) is equivalent to

$$\begin{aligned}
 (4.4) \quad & \max \mathbb{E} \left[ \sum_{t=1}^{T+1} X_t + \varphi(X_1, \dots, X_T, X_{T+1}) \right] \\
 & \text{s.t. } G_t(s_t, v_t) \geq X_t, \quad t = 1, \dots, T, \\
 & G_{T+1}(s_{T+1}) \geq X_{T+1}, \\
 & (s, v) \in Z_0, \\
 & X_t \in \mathcal{L}_1(\Omega, \mathcal{F}_t, P), \quad t = 1, \dots, T+1.
 \end{aligned}$$

Indeed, if  $(\hat{s}, \hat{v})$  is an optimal solution of (4.1), then the triple  $(\hat{s}, \hat{v}, \hat{X})$ , with  $\hat{X}_t = G_t(\hat{s}_t, \hat{v}_t)$ ,  $t = 1, \dots, T$  and with  $\hat{X}_{T+1} = G_{T+1}(\hat{s}_{T+1})$ , is an optimal solution of (4.4). Conversely, if a triple  $(\bar{s}, \bar{v}, \bar{X})$  is an optimal solution of (4.4), then also the triple  $(\bar{s}, \bar{v}, \hat{X})$ , with  $\hat{X}_t = G_t(\bar{s}_t, \bar{v}_t)$ ,  $t = 1, \dots, T$ , and with  $\hat{X}_{T+1} = G_{T+1}(\bar{s}_{T+1})$ , is an optimal solution of this problem. Then it is evident that the pair  $(\bar{s}, \bar{v})$  solves (4.1).

Let us define the set  $C$  of the triples  $(s, v, X) \in \mathcal{S} \times \mathcal{V} \times \mathcal{X}$  satisfying the constraints of problem (4.4). The set  $C$  is convex. The objective functional of this problem is concave and continuous. Therefore, we can apply optimality conditions to the problem

$$\begin{aligned}
 (4.5) \quad & \max \mathbb{E} \left[ \sum_{t=1}^{T+1} X_t + \varphi(X_1, \dots, X_T, X_{T+1}) \right] \\
 & \text{s.t. } (s, v, X) \in C.
 \end{aligned}$$

To this end, we need to subdifferentiate the objective functional. By the definition of the class  $\Phi([a, b] \times D)$ , the nonlinear part of the objective functional can be rewritten as follows:

$$I(X) = \mathbb{E}\varphi(X_1, \dots, X_T, X_{T+1}) = - \int_{\Omega} \int_{[a, b] \times D} \max(0, \eta - \langle \varrho, X(\omega) \rangle) \lambda(d\eta, d\varrho) P(d\omega)$$

for some measure  $\lambda \in \mathcal{M}_+([a, b] \times D)$ . This shows that  $I(\cdot)$  is concave and continuous on  $\mathcal{X}$ . The subdifferential of  $I(\cdot)$  can be calculated by the theory of convex integral functionals [27, 28, 29, 4]. According to [4, Theorem VII–7],

$$\begin{aligned}
 \partial I(X) &= \partial \int_{\Omega} \varphi(X_1(\omega), \dots, X_T(\omega), X_{T+1}(\omega)) P(d\omega) \\
 &= \left\{ \xi \in \mathcal{X}^* : \xi(\omega) \in \partial \varphi(X_1(\omega), \dots, X_T(\omega), X_{T+1}(\omega)), \text{ a.s.} \right\}.
 \end{aligned}$$

It remains to calculate  $\partial\varphi(\cdot)$ . The subdifferential of the integrand  $x \rightarrow \max(0, \eta - \langle \varrho, x \rangle)$  is given by the following multifunction:  $M : \mathbb{R}^{T+1} \times [a, b] \times D \rightrightarrows \mathbb{R}^{T+1}$ ,

$$M(x, \eta, \varrho) = \begin{cases} \{-\varrho\} & \text{if } \langle \varrho, x \rangle < \eta, \\ \text{conv}\{-\varrho, 0\} & \text{if } \langle \varrho, x \rangle = \eta, \\ \{0\} & \text{if } \langle \varrho, x \rangle > \eta. \end{cases}$$

By Strassen's theorem (see [34] and [18, Theorem 1.1]),

$$\partial\varphi(x) = -\partial \int_{[a,b] \times D} \max(0, \eta - \langle \varrho, x \rangle) \lambda(d\eta, d\varrho) = - \int_{[a,b] \times D} M(x, \eta, \varrho) \lambda(d\eta, d\varrho).$$

The last integral is understood as the collection of integrals of all measurable selections of  $M$ .

Now we can formulate the optimality conditions for problem (4.5). If  $(\hat{s}, \hat{v}, \hat{X})$  is an optimal solution of this problem, then we can assume that  $\hat{X}_t = G_t(\hat{s}_t, \hat{v}_t)$ ,  $t = 1, \dots, T$  and  $\hat{X}_{T+1} = G_{T+1}(\hat{s}_{T+1})$ , as already argued. The necessary and sufficient condition of optimality reads: there exists a subgradient

$$(4.6) \quad (\xi_1, \dots, \xi_{T+1}) \in \partial\hat{\varphi}(\hat{X}_1, \dots, \hat{X}_T, \hat{X}_{T+1})$$

such that the triple  $(\hat{s}, \hat{v}, \hat{X})$  is also a solution of the problem

$$\begin{aligned} \max \mathbb{E} \sum_{t=1}^{T+1} (1 + \xi_t) X_t \\ \text{s.t. } (s, v, X) \in C. \end{aligned}$$

Observe that the coordinates of every element of  $-M(x, \eta, \varrho)$  are nonincreasing, because the coordinates of  $\varrho$  are nonincreasing. Therefore the coordinates of every element of  $\partial\varphi(x)$  are nonincreasing. Thus, for almost every  $\omega \in \Omega$ , the coordinates of  $\xi(\omega)$  are nonincreasing and nonnegative.  $\square$

It follows from Theorem 4.1 that a random discount sequence  $(1 + \xi_t)$ ,  $t = 1, \dots, T$ , applied to the rewards in the expected value formulation yields the optimal solution of the dominance-constrained problem. The monotonicity of the random discount factors is inherited from the monotonicity of the sequences  $\varrho \in D$ . However, none of the deterministic discount sequences  $\varrho \in D$  (or their convex combinations) can be substituted for  $\xi$  or for its realizations, as can be seen from formulae (3.13) and (4.6).

**5. The maximum principle.** In this section we consider the discounted control problem (4.3) with the aim of deriving a counterpart of the maximum principle.

As the functionals  $\mathbb{E}G_t(\cdot, \cdot)$  are continuous, they are subdifferentiable. By virtue of Strassen's theorem (see [34] and [18, Theorem 1.1]), every subgradient  $(\sigma_t^s, \sigma_t^v) \in \partial\mathbb{E}G_t(s_t, v_t)$  has the form  $(\sigma_t^s(\omega), \sigma_t^v(\omega)) \in \partial g_t(\hat{s}_t(\omega), \hat{v}_t(\omega))$ . We apply necessary and sufficient conditions of optimality as formulated in [13, Theorem 5]. At the optimal solution  $(\hat{s}, \hat{v})$  there exist the following:

- dual variables  $\hat{y}_t \in \mathcal{L}_q^{n_s}(\Omega, \mathcal{F}_{t+1}, P)$ ,  $t = 1, \dots, T$ , with  $1/p + 1/q = 1$ , and
- subgradients  $(\sigma_t^s, \sigma_t^v) \in \mathcal{L}_q^{n_s}(\Omega, \mathcal{F}_t, P) \times \mathcal{L}_q^{n_v}(\Omega, \mathcal{F}_t, P)$ ,  $t = 1, \dots, T$ , and  $\sigma_{T+1}^s \in \mathcal{L}_q^{n_s}(\Omega, \mathcal{F}_{T+1}, P)$  such that for all  $t$  and  $P$ -almost all  $\omega \in \Omega$

$$\begin{aligned} (5.1) \quad & (\sigma_t^s(\omega), \sigma_t^v(\omega)) \in \partial g_t(\hat{s}_t(\omega), \hat{v}_t(\omega)), \\ & \sigma_{T+1}^s(\omega) \in \partial g_{T+1}(\hat{s}_{T+1}(\omega)), \end{aligned}$$

and the pair  $(\hat{s}, \hat{v})$  is also a solution of the problem

$$\begin{aligned} \max \mathbb{E} \sum_{t=1}^{T+1} (1 + \xi_t) \langle \sigma_t^s, s_t \rangle + \mathbb{E} \sum_{t=1}^T (1 + \xi_t) \langle \sigma_t^v, v_t \rangle + \mathbb{E} \sum_{t=1}^T \langle \hat{y}_t, A_t s_t + B_t v_t - s_{t+1} \rangle \\ \text{s.t. } v_t \in V_t \text{ a.s., } t = 1, \dots, T. \end{aligned}$$

Observe that  $\mathbb{E} \langle \hat{y}_t, A_t s_t + B_t v_t - s_{t+1} \rangle = \mathbb{E} \langle \mathbb{E}[\hat{y}_t | \mathcal{F}_t], A_t s_t + B_t v_t \rangle - \mathbb{E} \langle \hat{y}_t, s_{t+1} \rangle$ . For the last optimization problem to have a solution, it is necessary that the dual variables  $\hat{y}$  satisfy the adjoint equations

$$\begin{aligned} (5.2) \quad y_T &= (1 + \xi_{T+1}) \sigma_{T+1}^s, \\ y_{t-1} &= A'_t \mathbb{E}[y_t | \mathcal{F}_t] + (1 + \xi_t) \sigma_t^s, \quad t = T, \dots, 2. \end{aligned}$$

Here  $A'_t$  is the transpose of  $A_t$ . Assuming that they hold true for  $\hat{y}$ , we obtain the problem

$$\begin{aligned} \max \sum_{t=1}^T \mathbb{E} \langle (1 + \xi_t) \sigma_t^v + B'_t \hat{y}_t, v_t \rangle \\ \text{s.t. } v_t \in V_t \text{ a.s., } t = 1, \dots, T. \end{aligned}$$

It follows that for  $t = 1, \dots, T$  and for  $P$ -almost all  $\omega \in \Omega$ , the optimal control  $\hat{v}_t(\omega)$  is a solution of the deterministic problem

$$(5.3) \quad \max_{v_t(\omega) \in V_t} \langle (1 + \xi_t(\omega)) \sigma_t^v(\omega) + B'_t(\omega) \mathbb{E}[y_t | \mathcal{F}_t](\omega), v_t(\omega) \rangle.$$

We summarize these considerations in the following statement.

**THEOREM 5.1.** *Assume that the uniform dominance condition is satisfied. If  $(\hat{s}, \hat{v})$  is an optimal solution of (3.2)–(3.5), then there exist discount factors  $\xi_t \in \mathcal{L}_\infty(\Omega, \mathcal{F}_t, P)$ ,  $t = 1, \dots, T + 1$ , subgradients  $(\sigma_t^s, \sigma_t^v) \in \mathcal{L}_q^{n_s}(\Omega, \mathcal{F}_t, P) \times \mathcal{L}_q^{n_v}(\Omega, \mathcal{F}_t, P)$  satisfying (5.1), and dual variables  $\hat{y}_t \in \mathcal{L}_q^{n_s}(\Omega, \mathcal{F}_{t+1}, P)$ ,  $t = 1, \dots, T$ , such that the adjoint equations (5.2) are satisfied and for all  $t = 1, \dots, T$  and for almost all  $\omega \in \Omega$  the control  $\hat{v}_t(\omega)$  is a solution of (5.3).*

Observing that in (5.3) the vector  $\sigma_t^v(\omega)$  is a subgradient of  $g(\hat{s}_t(\omega), \cdot)$  at  $\hat{v}_t(\omega)$ , we can reformulate the maximum principle by using the Pontryagin function

$$P(s_t(\omega), v_t(\omega), y_t(\omega), \xi_t(\omega)) = (1 + \xi_t(\omega))g(s_t(\omega), v_t(\omega)) + \langle \mathbb{E}[y_t | \mathcal{F}_t](\omega), B_t(\omega)v_t(\omega) \rangle.$$

It follows that for all  $t = 1, \dots, T$  and for almost all  $\omega \in \Omega$ , the control  $\hat{v}_t(\omega)$  is a solution of the problem

$$(5.4) \quad \max_{v_t(\omega) \in V_t} P(\hat{s}_t(\omega), v_t(\omega), y_t(\omega), \xi_t(\omega)).$$

It is important to stress that this theorem provides necessary conditions only. Even if the operators  $G$  are linear, the risk-averse control problem (3.1) is nonlinear, because the ordering constraint results in nonlinear utility functions  $\varphi(\cdot)$  in the necessary and sufficient conditions of Theorem 3.3 of this paper.

**Acknowledgment.** The authors are also grateful to two anonymous referees for their insightful comments which helped to improve the presentation of this paper.

## REFERENCES

- [1] B. C. ARNOLD, *Majorization and The Lorenz Order: A Brief Introduction*, Lecture Notes in Statist. 43, Springer-Verlag, Berlin, 1987.
- [2] D. BERLEANT, M. DANCRE, J.-P. ARGAUD, AND G. SHEBLE, *Electric company portfolio optimization under interval stochastic dominance constraints*, in Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, Pittsburgh, PA, 2005, pp. 416–422.
- [3] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
- [4] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, 1977.
- [5] D. DENTCHEVA AND A. RUSZCZYŃSKI, *Optimization with stochastic dominance constraints*, SIAM J. Optim., 14 (2003), pp. 548–566.
- [6] D. DENTCHEVA AND A. RUSZCZYŃSKI, *Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints*, Math. Program., 99 (2004), pp. 329–350.
- [7] D. DENTCHEVA AND A. RUSZCZYŃSKI, *Portfolio optimization with stochastic dominance constraints*, J. Banking and Finance, 30 (2006), pp. 433–451.
- [8] D. DENTCHEVA AND A. RUSZCZYŃSKI, *Optimization with multivariate stochastic dominance constraints*, Math. Program., 117 (2009), pp. 111–127.
- [9] R. M. DUDLEY, *Real Analysis and Probability*, Cambridge University Press, Cambridge, UK, 2002.
- [10] N. EL KAROUI AND A. MEZIOU, *Constrained optimization with respect to stochastic dominance: Applications to portfolio insurance*, Math. Finance, 16 (2006), pp. 103–117.
- [11] P. C. FISHBURN, *Utility Theory for Decision Making*, John Wiley and Sons, New York, 1970.
- [12] J. HADAR AND W. RUSSELL, *Rules for ordering uncertain prospects*, Amer. Econ. Rev., 59 (1969), pp. 25–34.
- [13] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, Nauka, Moscow, 1974 (in Russian) (English transl., North-Holland, Amsterdam, 1979).
- [14] A. J. KING, *Duality and martingales: A stochastic programming perspective on contingent claims*, Math. Program., 91 (2002), pp. 543–562.
- [15] G. KOSHEVOY AND K. MOSLER, *The Lorenz zonoid of a multivariate distribution*, J. Amer. Statist. Assoc., 91 (1996), pp. 873–882.
- [16] G. KOSHEVOY AND K. MOSLER, *Lift zonoids, random convex hulls and the variability of random vectors*, Bernoulli, 4 (1998), pp. 377–399.
- [17] E. LEHMANN, *Ordered families of distributions*, Ann. Math. Statist., 26 (1955), pp. 399–419.
- [18] V. L. LEVIN, *Convex Analysis in Spaces of Measurable Functions and Its Applications in Economics*, Nauka, Moscow, 1985 (in Russian).
- [19] D. T. LUC, *Theory of vector optimization*, in Lecture Notes in Econ. and Math. Systems, Springer-Verlag, Berlin, 1989.
- [20] H. B. MANN AND D. R. WHITNEY, *On a test of whether one of two random variables is stochastically larger than the other*, Ann. Math. Statistics, 18 (1947), pp. 50–60.
- [21] K. MOSLER AND M. SCARSINI, EDS., *Stochastic Orders and Decision Under Risk*, Institute of Mathematical Statistics, Hayward, CA, 1991, pp. 261–281.
- [22] P. MULIERE AND M. SCARSINI, *Multivariate decisions with unknown price vector*, Econom. Lett., 29 (1999), pp. 13–19.
- [23] A. MÜLLER AND M. SCARSINI, *Stochastic comparison of random vectors with a common copula*, Math. Oper. Res., 26 (2001), pp. 723–740.
- [24] A. MÜLLER AND D. STOYAN, *Comparison Methods for Stochastic Models and Risks*, John Wiley & Sons, Chichester, 2002.
- [25] W. OGRYCZAK AND A. RUSZCZYŃSKI, *From stochastic dominance to mean-risk models: Semideviations as risk measures*, European J. Oper. Res., 116 (1999), pp. 33–50.
- [26] J. P. QUIRK AND R. SAPOSNIK, *Admissibility and measurable utility functions*, Rev. Econ. Stud., 29 (1962), pp. 140–146.
- [27] R. T. ROCKAFELLAR, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.
- [28] R. T. ROCKAFELLAR, *Integrals which are convex functionals II*, Pacific J. Math., 39 (1971), pp. 439–469.

- [29] R. T. ROCKAFELLAR, *Convex integral functionals and duality*, in Contributions to Nonlinear Functional Analysis, Academic Press, New York, 1971, pp. 215–239.
- [30] R. T. ROCKAFELLAR, *Duality and optimality in multistage stochastic programming*, Ann. Oper. Res., 85 (1999), pp. 1–19.
- [31] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Generalized linear-quadratic problems of deterministic and stochastic optimal control in discrete time*, SIAM J. Control Optim., 28 (1990), pp. 810–822.
- [32] M. SCARSINI AND M. SHAKED, *Some conditions for stochastic equality*, Naval Res. Logist., 37 (1990), pp. 617–625.
- [33] M. SHAKED AND J. G. SHANTHIKUMAR, *Stochastic Orders and Their Applications*, Academic Press, Boston 1994.
- [34] V. STRASSEN, *The existence of probability measures with given marginals*, Ann. Math. Statist., 38 (1965), pp. 423–439.
- [35] R. SZEKLI, *Stochastic Ordering and Dependence in Applied Probability*, Lecture Notes in Stat. 97, Springer-Verlag, New York, 1995.



## NUMERICAL VERIFICATION OF OPTIMALITY CONDITIONS\*

ARND RÖSCH<sup>†</sup> AND DANIEL WACHSMUTH<sup>‡</sup>

**Abstract.** A class of optimal control problems for a semilinear elliptic partial differential equation with control constraints is considered. It is well known that sufficient second-order conditions ensure the stability of optimal solutions, and the convergence of numerical methods. Otherwise, such conditions are very difficult to verify (analytically or numerically). We will propose a new approach as follows: Starting with a numerical solution for a fixed mesh we will show the existence of a local minimizer of the continuous problem. Moreover, we will prove that this minimizer satisfies the sufficient second-order conditions.

**Key words.** optimal control, sufficient optimality condition, semilinear elliptic equation, numerical verification, error estimates

**AMS subject classifications.** Primary, 49K20; Secondary, 49M25, 65N15

**DOI.** 10.1137/060663714

**1. Introduction.** In this paper, we consider the optimal control problem (P) of minimizing  $J(y, u)$  given by

$$(1.1) \quad J(y, u) := \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\nu}{2} \|u\|_{L^2(\Omega)}^2$$

subject to the semilinear boundary value problem

$$(1.2) \quad \begin{aligned} (Ay)(x) + f(y(x)) &= u(x) && \text{in } \Omega \\ y &= 0 && \text{on } \Gamma \end{aligned}$$

and to the control constraints

$$(1.3) \quad a \leq u(x) \leq b \text{ a.e. in } \Omega.$$

In this setting,  $A$  is a uniformly bounded elliptic differential operator and  $\Omega$  is a bounded domain of  $\mathbb{R}^N$ ,  $N = 2, 3$ , with boundary  $\Gamma$ . Moreover,  $\nu$  is a fixed positive number. Precise assumptions on and definitions of the quantities introduced above are formulated at the end of this section.

Sufficient second-order optimality conditions are well-established mathematical tools. The development of such conditions for control constrained problems governed by partial differential equations started with the papers of Goldberg and Tröltzsch [7, 8]. Two new trends narrowed the gap between necessary and sufficient optimality conditions a few years later. Dontchev et al. [6] introduced strongly active sets in the theory of sufficient second-order conditions. The idea is that no coercivity is needed on subspaces where the first-order conditions are sufficient. The equivalence of the coercivity condition with a positivity condition for the second derivative of the Lagrangian for a class of semilinear elliptic problems was proved by Bonnans [4].

---

\*Received by the editors June 26, 2006; accepted for publication (in revised form) April 30, 2008; published electronically September 19, 2008.

<http://www.siam.org/journals/sicon/47-5/66371.html>

<sup>†</sup>Universität Duisburg-Essen, Fachbereich Mathematik, Forsthausweg 2, D-47057 Duisburg, Germany (arnd.roesch@uni-due.de).

<sup>‡</sup>Institut für Mathematik, Technische Universität Berlin, Str. des 17. Juni 136, D-10623 Berlin, Germany (wachsmut@math.tu-berlin.de).

Lipschitz stability results for optimal control problems governed by partial differential equations (PDEs) can be proved using second-order optimality conditions; see for instance the papers of Malanowski and Tröltzsch [11, 12] and Tröltzsch [16]. Moreover, it is possible to show locally quadratic convergence of the SQP-method; see Arada, Raymond, and Tröltzsch [3]. These conditions are necessary to derive error estimates for discretized optimal control problems; see Arada, Casas, and Tröltzsch [2] and Casas, Mateos, and Tröltzsch [5].

For all of these results, it is essential to assume that the *unknown* solution satisfies a sufficient optimality condition. To the best of our knowledge, there are only four references available in which it is shown that such a condition is fulfilled: the articles of Arada, Raymond, and Tröltzsch [3] and Casas, Mateos, and Tröltzsch [5]; the book chapter of Mittelman and Tröltzsch [13]; and the monograph of Tröltzsch [17]. However, [17] essentially used the fact that the solution of the considered problems is *known*. If that is not the case and the optimization problem is not convex, then there is no known way up to now to check whether the sufficient optimality condition is satisfied or not. Consequently, the validity of results based thereon is not clear. Even if the numerical method behaves well for a finite sequence of meshes, it is not clear whether the numerical solution is close to a stationary point of the original problem or not.

The numerical verification of a second-order optimality condition is the main concern of this paper. Our starting point is the following very realistic situation: A numerical solution of a discretized optimal control problem is given with information on the mesh size and the discretization error. Under certain conditions, we will show that a local minimum of the continuous problem exists in a neighborhood of the numerical solution.

The paper is organized as follows. Section 2 is devoted to optimality conditions of the continuous problem. In section 3 we sketch our strategy. The state equation, discretization, and objective functional are estimated in the following sections 4–6. The main theorem on verification of optimality conditions is stated and proved in section 7. We will discuss all assumptions and the results through a numerical example in section 8.

**Assumptions.** First, we want to specify the assumptions on the various ingredients of the considered optimal control problem.

(A1)  $\Omega \subset \mathbb{R}^N$ ,  $N \in \{2, 3\}$ , is a bounded domain that is either convex and polygonal or of the class  $C^{1,1}$ . The parameter  $\nu$  is assumed to be positive. The control bounds  $a, b$  are real numbers satisfying  $a < b$ .

We denote by the  $H^1$ -norm  $\|\cdot\|_{H^1(\Omega)}$  as follows:  $\|y\|_{H^1(\Omega)}^2 = \|y\|_{L^2(\Omega)}^2 + \|\nabla y\|_{L^2(\Omega)}^2$ . We will denote the imbedding constants from  $H_0^1(\Omega)$  to  $L^p(\Omega)$  by  $I_p$ , i.e.,

$$\|y\|_{L^p(\Omega)} \leq I_p \|y\|_{H^1(\Omega)} \quad \forall y \in H_0^1(\Omega).$$

$A$  is a uniformly elliptic differential operator defined by

$$(Ay)(x) = - \sum_{i,j=1}^N \frac{\partial}{\partial x_i} \left( a_{ij}(x) \frac{\partial}{\partial x_j} y(x) \right) + c_0(x)y(x)$$

with functions  $a_{ij}$  that belong to  $C^{0,1}(\bar{\Omega})$ , satisfying the condition  $a_{ij}(x) = a_{ji}(x)$  and

$$\delta_0 \|y\|_{H^1(\Omega)}^2 \leq \langle Ay, y \rangle_{H^{-1}, H^1}, \quad \langle Ay_1, y_2 \rangle_{H^{-1}, H^1} \leq \delta_1 \|y_1\|_{H^1(\Omega)} \|y_2\|_{H^1(\Omega)} \quad \forall y \in H_0^1(\Omega).$$

Let us denote by  $a(\cdot, \cdot)$  as follows the bilinear form induced by  $A$ :

$$a(u, v) = \langle Au, v \rangle_{H^{-1}, H^1}.$$

**(A2)** The function  $f = f(y) : \Omega \rightarrow \mathbb{R}$  is of class  $C^2$  with  $f(0) = 0$ . It satisfies the following conditions on boundedness and Lipschitz-continuity: For all  $\tilde{M} > 0$  there are constants  $c_f, c_{f'}, c_{f''} > 0$  such that

$$\begin{aligned} |f(y_1) - f(y_2)| &\leq c_f |y_1 - y_2|, \\ |f'(y_1) - f'(y_2)| &\leq c_{f'} |y_1 - y_2|, \\ |f''(y_1) - f''(y_2)| &\leq c_{f''} |y_1 - y_2| \end{aligned}$$

hold for all  $|y_i| \leq \tilde{M}$ ,  $i = 1, 2$ . Moreover, we require  $f'(y) \geq 0$  in  $\mathbb{R}$ .

**2. Optimality conditions for the continuous problem.** A function  $y$  is called a weak solution of the semilinear elliptic equation (see section 4) if it satisfies

$$(2.1) \quad a(y, v) + (f(y), v) = (u, v) \quad \forall v \in V = H^1(\Omega) \cap C(\bar{\Omega}).$$

Solvability as well as regularity results for the weak formulation of the state equation are stated in section 4.

The set of admissible controls  $U_{ad} \subset U := L^2(\Omega)$  is defined by

$$U_{ad} = \{u \in L^2(\Omega) : a \leq u(x) \leq b \text{ a.e. in } \Omega\}.$$

We define the Lagrange functional by

$$(2.2) \quad L(y, u, p) = J(y, u) - a(y, p) - (f(y), p) + (u, p).$$

Then the necessary first-order optimality conditions are given by

$$(2.3) \quad L_y(\bar{y}, \bar{u}, \bar{p})(v) = 0 \quad \text{for } v \in V,$$

$$(2.4) \quad L_u(\bar{y}, \bar{u}, \bar{p})(u - \bar{u}) \geq 0 \quad \text{for } u \in U_{ad}.$$

Every admissible point  $(u, y)$  is called stationary if there exists an adjoint  $p$  such that the optimality system (2.3)–(2.4) is satisfied. Equation (2.3) is equivalent to the adjoint equation defined by

$$(2.5) \quad a(v, p) + (f'(\bar{y})p, v) = (y - y_d, v) \quad \forall v \in V = H^1(\Omega) \cap C(\bar{\Omega}).$$

Existence and regularity results for the adjoint equation can be found in section 4.

A sufficient second-order condition, henceforth called (SSC), is given by

$$(2.6) \quad L''_{(y,u)}(\bar{y}, \bar{u}, \bar{p})(y, u) \geq \delta \|u\|_{L^2(\Omega)}^2$$

for all  $(y, u) \in V \times U$  satisfying the linearized equation

$$(2.7) \quad a(y, v) + (f'(\bar{y})y, v) = (u, v) \quad \forall v \in V.$$

Here, the second derivative of the Lagrangian is given by

$$(2.8) \quad L''_{(y,u)}(\bar{y}, \bar{u}, \bar{p})(y, u) = \|y\|_{L^2(\Omega)}^2 + \nu \|u\|_{L^2(\Omega)}^2 + (f''(\bar{y})y^2, \bar{p}).$$

Note that checking (SSC) requires the knowledge of  $(\bar{y}, \bar{u}, \bar{p})$ , which typically cannot be assumed. Hence, it is impossible to show that (SSC) holds in advance.

**3. The discretized problem.** Typically, control and state are discretized (for instance, by finite elements) to obtain numerical approximations  $(\bar{y}_h, \bar{u}_h, \bar{p}_h)$  of solutions of the continuous problem. Here, the following questions arise:

- Is the numerical solution close to a stationary point of the continuous (i.e., undiscretized) problem?
- Is the numerical solution close to a local minimizer of the continuous problem?

In this article, we will derive conditions for the solution of the discretized problem  $(\bar{y}_h, \bar{u}_h, \bar{p}_h)$  that ensure the existence of a local minimizer of (P) in a specified neighborhood of the numerical solution  $(\bar{y}_h, \bar{u}_h, \bar{p}_h)$ . Under additional assumptions, we can even show that this unknown minimizer fulfills (SSC).

We will formulate general assumptions for the discretization of the optimal control problem. The assumption fits into finite element discretizations of the elliptic state equation (2.1), which is replaced by

$$(3.1) \quad a(y_h, v_h) + (f(y_h), v_h) = (u, v_h) \quad \forall v_h \in V_h,$$

where  $V_h \subset V$  is a finite dimensional space. The set of admissible controls is given by

$$(3.2) \quad U_{ad}^h := \{u_h \in U_h : a \leq u_h(x) \leq b\},$$

where  $U_h$  is the space of functions that are piecewise constant over the elements of a triangulation of  $\Omega$ . We refer to section 5 for the precise formulation of the assumptions on the discretization.

The corresponding discretized adjoint equation is given by

$$(3.3) \quad a(v_h, p_h) + (f'(\bar{y}_h)p_h, v_h) = (y - y_d, v_h) \quad \forall v_h \in V_h.$$

The discretized optimal control problem is given by the following: Minimize  $J(y_h, u_h)$  subject to the discretized equation (3.1) and to the control constraint  $u \in U_{ad}^h$ .

In what follows, we assume that  $(\bar{y}_h, \bar{u}_h, \bar{p}_h)$  solves that discretized problem without numerical errors. However, it would be sufficient for all results that the numerical error is smaller than the discretization error. There are different types of numerical errors in a practical computation, e.g., rounding errors, iterative solvers (Newton), quadrature rules for the evaluation of the right-hand sides, and so on. In our computations we chose the numerical methods in such a way that these types of numerical errors are essentially smaller than the discretization error.

Now, let us sketch our strategy. We *assume* a condition similar to (2.6), namely, that

$$(3.4) \quad L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y, u) \geq \delta \|u\|_{L^2(\Omega)}^2$$

holds for all  $(y, u) \in V \times U$  fulfilling the linearized equation

$$(3.5) \quad a(y, v) + (f'(\bar{y}_h)y, v) = (u, v) \quad \forall v \in V.$$

So we assume the coercivity for the discretized solution instead of the continuous one. Note that this condition is required for all  $u \in U$  (and not only in  $U^h$ ). The inequality (3.4) means

$$(3.6) \quad \|y\|_{L^2(\Omega)}^2 + \nu \|u\|_{L^2(\Omega)}^2 + (f''(\bar{y}_h)y^2, \bar{p}_h) \geq \delta \|u\|_{L^2(\Omega)}^2.$$

Since  $\bar{y}_h, \bar{p}_h$  are known, there is a chance to verify this condition; i.e., it would suffice, for instance, that

$$(3.7) \quad f''(\bar{y}_h(x))\bar{p}_h(x) > -1$$

hold on  $\Omega$ . Moreover, if  $(\bar{y}, \bar{u}, \bar{p})$  satisfies (SSC), then such an assumption will be true for sufficiently fine discretizations.

The main difference between the conditions (3.4)–(3.5) and (SSC) is that the point  $(\bar{y}_h, \bar{u}_h, \bar{p}_h)$ , where the second-derivative of  $L$  is evaluated in (3.4), is known. Hence, there is the possibility to check that condition. Indeed, inequality (3.7), which is easy to verify, is sufficient for the fulfillment of (3.4). As discussed above, it is much more difficult to prove (SSC) for the unknown solution  $\bar{u}$ . However, one cannot guarantee that (SSC) or (3.4) is fulfilled if, for instance, good convergence of numerical algorithms is observed, since both are indeed sufficient conditions.

One could of course try to check the inequality (2.6) above for all admissible controls  $u \in U_{ad}$  with associated state  $y$  and adjoint  $p$ . Then this inequality must hold for all  $(u, y, p)$  instead of  $(\bar{u}, \bar{y}, \bar{p})$ . However, this is the global convexity of the original problem. It is clear that in the convex case all theoretical results are valid.

Let us continue with the sketch of our strategy. The pair  $(\bar{u}_h, \bar{y}_h)$  is not admissible for the continuous problem. Therefore, we introduce the auxiliary state  $y^h$  as the solution of the elliptic equation with right-hand side  $\bar{u}_h$ ,

$$(3.8) \quad a(y^h, v) + (f(y^h), v) = (\bar{u}_h, v) \quad \forall v \in V.$$

Under the assumption that  $\delta$  in (3.4) is large enough, we will show the existence of a radius  $r > 0$  such that

$$J(y, u) - J(y^h, \bar{u}_h) > 0 \quad \text{if } \|u - \bar{u}_h\|_{L^2(\Omega)} = r.$$

This result will be the key for our argumentation; i.e., since  $(y^h, \bar{u}_h)$  is an admissible pair for the continuous problem, we obtain the existence of a local minimizer  $\hat{u}$  of (P) in the neighborhood  $\{u \in U_{ad} : \|u - \bar{u}_h\|_{L^2(\Omega)} < r\}$  of  $\bar{u}_h$ ; see section 6. Under additional assumptions, this local minimizer is unique and fulfills the sufficient condition (SSC). That is, we get the fulfillment of (SSC) as an a posteriori result and not as an a priori assumption.

In order to get a computable bound for the radius  $r$ , all estimations have to be carried out carefully, and all constants have to be known. A representative collection of those constants includes

- embedding constants  $H_0^1(\Omega) \hookrightarrow L^p(\Omega)$ ,
- interpolation constants for the finite elements,
- global bounds on the state, and
- norm of the solution operators of the PDEs.

In the following sections, we will estimate several ingredients of the optimal control problem and its discretization. The solution operator of the semilinear elliptic equation is studied in section 4. The finite element discretization and associated error estimates for the state equations can be found in section 5. The difference between the objective functionals  $J(y, u) - J(y^h, \bar{u}_h)$  is investigated and estimated in section 6. The existence of a local minimizer  $\hat{u}$  of (P) in the neighborhood of  $\bar{u}_h$  is proved in section 7. Also the proof that  $\hat{u}$  satisfies (SSC) is contained in that section.

#### 4. State equation. Now, let us study briefly the state equation.

**THEOREM 4.1.** *The semilinear state equation admits a unique solution  $y \in V \cap H^2(\Omega)$  for all  $u \in L^2$  and satisfies the estimates*

$$(4.1) \quad \|y\|_{H^2(\Omega)} \leq c_S \|u\|_{L^2(\Omega)},$$

$$(4.2) \quad \|y\|_{L^\infty(\Omega)} \leq c_{SL^\infty} \|u\|_{L^2(\Omega)}.$$

Since  $U_{ad}$  is bounded, we have in addition

$$(4.3) \quad \|y\|_{L^\infty(\Omega)} < M$$

for all solutions  $y$  associated with admissible controls  $u$ .

For the proof we refer to Grisvard [9].

Thanks to Theorem 4.1, we can interpret all assumptions of **(A2)**, which were defined on bounded sets, as global estimates as follows

**(A2)'** The function  $f = f(y) : \Omega \rightarrow \mathbb{R}$  is of class  $C^2$  with  $f(0) = 0$ . It satisfies the following conditions on boundedness and Lipschitz-continuity: There are constants  $c_f, c_{f'}, c_{f''} > 0$  such that

$$\begin{aligned} |f(y_1) - f(y_2)| &\leq c_f |y_1 - y_2|, \\ |f'(y_1) - f'(y_2)| &\leq c_{f'} |y_1 - y_2|, \\ |f''(y_1) - f''(y_2)| &\leq c_{f''} |y_1 - y_2| \end{aligned}$$

hold for all  $y_1, y_2 \in \mathbb{R}$ .

This assumption implies the boundedness of the derivatives of  $f$ :  $|f'(y)| \leq c_f$  and  $|f''(y)| \leq c_{f'}$  for all  $y \in \mathbb{R}$ .

The Lipschitz-continuity of the control-to-state mapping is an immediate consequence of the monotonicity of  $f$ .

LEMMA 4.2. *Let  $y_1$  and  $y_2$  be the solutions of (2.1) for controls  $u_1$  and  $u_2$ , respectively. Then the estimate*

$$(4.4) \quad \|y_1 - y_2\|_{L^2(\Omega)} \leq \|y_1 - y_2\|_{H^1(\Omega)} \leq c_L \|u_1 - u_2\|_{L^2(\Omega)}$$

is valid with  $c_L = \frac{I_2}{\delta_0}$ .

A similar estimate is available for solutions of the linearized system

$$(4.5) \quad a(y, v) + (f'(\bar{y})y, v) = (u, v) \quad \forall v \in V = H^1(\Omega) \cap C(\bar{\Omega}).$$

Here,  $\bar{y}$  is a given function of  $V$ .

COROLLARY 4.3. *The linearized state equation (4.5) admits a unique solution  $y$  for each  $u \in U$ , and it holds that*

$$\|y_1 - y_2\|_{L^2(\Omega)} \leq \|y_1 - y_2\|_{H^1(\Omega)} \leq c_L \|u_1 - u_2\|_{L^2(\Omega)}$$

with the same constant  $c_L$  as in (4.4) above.

We conclude this section with the state of an existence result for the adjoint equation (2.5).

LEMMA 4.4. *The adjoint equation (2.5) admits a unique solution. Moreover, the estimate*

$$(4.6) \quad \|p\|_{H^2(\Omega)} \leq c_p \|y - y_d\|_{L^2(\Omega)}$$

is valid.

For the proof we refer to Grisvard [9].

**5. Finite element discretization.** The state  $V$  is approximated by a finite dimensional subspace  $V_h \subset V$ . Here, we impose the following requirement.

**(A3)** The mesh parameter  $h$  is assumed to be smaller than 1. We assume the existence of an operator  $i_h : V \rightarrow V_h$  with the following properties:

$$(5.1) \quad \|y - i_h y\|_{L^2(\Omega)} \leq c_1 h^2 \|y\|_{H^2(\Omega)},$$

$$(5.2) \quad \|y - i_h y\|_{H^1(\Omega)} \leq c_2 h \|y\|_{H^2(\Omega)}.$$

This assumption is a standard property of conforming finite element discretizations.

The control is discretized piecewise constant on a mesh  $\mathcal{T}_h$  containing open sets  $T$  (finite elements),

$$U_h := \{q_h \in L^2(\Omega) : q_h|_T = \text{const for all } T \in \mathcal{T}_h\}.$$

We require the following for the mesh  $\mathcal{T}_h$ .

(A4) The diameter of the largest element of  $\mathcal{T}_h$  is bounded by  $h$ . Moreover,

$$\bigcup_{T \in \mathcal{T}_h} \bar{T} = \bar{\Omega}, \quad T_i \cap T_j = \emptyset \quad \forall T_i, T_j \in \mathcal{T}_h, i \neq j.$$

Now, we give an error estimate for the semilinear equation.

LEMMA 5.1. *Let  $y$  and  $y_h$  be the unique solutions of the semilinear equations (2.1) and (3.1), respectively. Then, the following error estimates are valid:*

$$(5.3) \quad \|y - y_h\|_{L^2(\Omega)} \leq c_{L^2} h^2 \|y\|_{H^2(\Omega)},$$

$$(5.4) \quad \|y - y_h\|_{H^1(\Omega)} \leq c_{H^1} h \|y\|_{H^2(\Omega)}$$

with  $c_{L^2} = (\delta_1 c_2 + c_f c_1) c_{H^1} c_M$  and  $c_{H^1} = (\delta_1 c_2 + c_f c_1) / \delta_0$ .

*Proof.* Using the ellipticity assumption (A1), the monotonicity of  $f$  in (A2)', and Galerkin orthogonality, the  $H^1$ -estimate follows by standard arguments. For the  $L^2$ -estimate, we introduce an auxiliary state  $g$  as the solution of

$$a(g, v) + (f_y^{y_h} g, v) = (e, v) \quad \forall v \in V$$

with  $e = \frac{y - y_h}{\|y - y_h\|_{L^2(\Omega)}}$ . The expression  $f_y^{y_h} g$  is defined by

$$f_y^{y_h} g = \int_0^1 f'(y_h + s(y - y_h)) g \, ds.$$

Note that  $\|e\|_{L^2(\Omega)} = 1$ . Since  $f'$  is globally bounded, we have

$$\|g\|_{H^2(\Omega)} \leq c_M$$

with some constant  $c_M > 0$ . Using these definitions, we find

$$\begin{aligned} \|y - y_h\|_{L^2(\Omega)} &= (e, y - y_h) \\ &= a(g, y - y_h) + (f_y^{y_h} g, y - y_h) \\ &= a(g, y - y_h) + (g, f_y^{y_h}(y - y_h)) \\ &= a(g, y - y_h) + (g, f(y) - f(y_h)). \end{aligned}$$

Testing the difference of the weak formulation (2.1) and the discrete equation (3.1) by  $v_h = i_h g$ , we obtain

$$\begin{aligned} \|y - y_h\|_{L^2(\Omega)} &= a(g - i_h g, y - y_h) + (g - i_h g, f(y) - f(y_h)) \\ &\leq \delta_1 \|g - i_h g\|_{H^1(\Omega)} \|y - y_h\|_{H^1(\Omega)} + c_f \|g - i_h g\|_{L^2(\Omega)} \|y - y_h\|_{L^2(\Omega)} \\ &\leq \delta_1 c_2 c_{H^1} h^2 \|g\|_{H^2(\Omega)} \|y\|_{H^2(\Omega)} + c_f c_1 h^2 \|g\|_{H^2(\Omega)} \|y - y_h\|_{L^2(\Omega)} \\ &\leq (\delta_1 c_2 c_{H^1} c_M h^2 + c_f c_1 c_{H^1} c_M h^3) \|y\|_{H^2(\Omega)}. \end{aligned}$$

Now, for  $h < 1$  we have  $c_{L^2} = (\delta_1 c_2 + c_f c_1) c_{H^1} c_M$ .  $\square$

Applying these results, we can give error estimates for the auxiliary function  $y^h$  introduced in (3.8).

COROLLARY 5.2. *It holds for the difference  $y^h - \bar{y}_h$  that*

$$(5.5) \quad \|y^h - \bar{y}_h\|_{L^2(\Omega)} \leq c_{L^2} c_S h^2 \|\bar{u}_h\|_{L^2(\Omega)},$$

$$(5.6) \quad \|y^h - \bar{y}_h\|_{H^1(\Omega)} \leq c_{H^1} c_S h \|\bar{u}_h\|_{L^2(\Omega)}.$$

*Proof.* The claim follows directly from Lemma 5.1 and Theorem 4.1.  $\square$

**6. Estimation of the objective functional; existence of a local minimizer.** Throughout the following sections, we assume the fulfillment of **(A1)**, **(A3)**, and **(A4)**, which are assumptions concerning the state equation and the discretization.

Let us fix a mesh according to assumptions **(A3)** and **(A4)**. Let  $(\bar{y}_h, \bar{u}_h, \bar{p}_h)$  be the solution of the discretized problem for this particular mesh. Hence, the discretized state equation (3.1), the discretized adjoint equation (3.3), and the variational inequality

$$(\nu \bar{u}_h + \bar{p}_h, u_h - \bar{u}_h) \geq 0 \quad \forall u_h \in U_{ad}^h$$

are satisfied. Furthermore, we impose the following conditions on the discrete solution  $(\bar{y}_h, \bar{u}_h, \bar{p}_h)$ .

**(A5)** We assume that the coercivity condition (3.4)–(3.5) holds at  $(\bar{y}_h, \bar{u}_h, \bar{p}_h)$  with some  $\delta > 0$ . We require the regularities  $\bar{y}_h \in L^\infty(\Omega)$  and  $\bar{p}_h \in W^{1,\infty}(\Omega)$ .

The latter assumption allows us to use the modified assumption **(A2)'** on the bounded set  $\{y \in L^\infty(\Omega) : \|y\|_\infty \leq \max(M, \|\bar{y}_h\|_\infty)\}$ . Here,  $M$  is the global bound provided by Theorem 4.1.

We are only interested in discretized solutions for a fixed mesh. Hence, the regularity assumptions  $\bar{y}_h \in L^\infty(\Omega)$  and  $\bar{p}_h \in W^{1,\infty}(\Omega)$  are in fact assumptions on the finite element space. We do not require that the norms  $\|\bar{y}_h\|_{L^\infty(\Omega)}$  and  $\|\bar{p}_h\|_{W^{1,\infty}(\Omega)}$  be bounded independently of  $h$ .

Now, we will investigate the behavior of the objective functional in the neighborhood of  $\bar{u}_h$ . The following lemma expresses  $J(y, u) - J(y^h, \bar{u}_h)$  as the sum of several addends, which will be estimated in what follows.

LEMMA 6.1. *Let  $u$  be an admissible control and  $y$  the associated state (solution of (2.1)). Then we can represent the difference of the objective values as*

$$(6.1) \quad J(y, u) - J(y^h, \bar{u}_h) = L(\bar{y}_h, \bar{u}_h, \bar{p}_h) - L(y^h, \bar{u}_h, \bar{p}_h)$$

$$(6.2) \quad + L_y(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h)$$

$$(6.3) \quad + L_u(\bar{y}_h, \bar{u}_h, \bar{p}_h)(u - \bar{u}_h)$$

$$(6.4) \quad + \frac{1}{2} L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h, u - \bar{u}_h)^2$$

$$(6.5) \quad + r_2,$$

where  $r_2$  denotes a second-order remainder term.

Now, we will estimate the terms (6.1)–(6.5).

LEMMA 6.2. *The discretization error in the Lagrange function can be estimated by*

$$(6.6) \quad |L(y^h, \bar{u}_h, \bar{p}_h) - L(\bar{y}_h, \bar{u}_h, \bar{p}_h)| < c_j h^2 \|\bar{u}_h\|_{L^2(\Omega)}$$



with  $c_j = c_{L^2} c_S (\|\bar{y}_h - y_d\|_{L^2(\Omega)} + \frac{1}{2} c_{L^2} c_S \|\bar{u}_h\|_{L^2(\Omega)})$ .

*Proof.* Here, we find for the objective

$$\begin{aligned} |J(y^h, \bar{u}_h) - J(\bar{y}_h, \bar{u}_h)| &= \left| \frac{1}{2} \|y^h - y_d\|_{L^2(\Omega)}^2 - \frac{1}{2} \|\bar{y}_h - y_d\|_{L^2(\Omega)}^2 \right| \\ &= \left| \frac{1}{2} \|y^h - \bar{y}_h\|_{L^2(\Omega)}^2 + (\bar{y}_h - y_d, y^h - \bar{y}_h) \right| \\ &\leq c_{L^2} c_S h^2 \|\bar{u}_h\|_{L^2(\Omega)} \left( \|\bar{y}_h - y_d\|_{L^2(\Omega)} + \frac{1}{2} c_{L^2} c_S h^2 \|\bar{u}_h\|_{L^2(\Omega)} \right). \end{aligned}$$

It remains to estimate the term associated with the semilinear equation in the difference  $L(y^h, \bar{u}_h, \bar{p}_h) - L(\bar{y}_h, \bar{u}_h, \bar{p}_h)$ . The semilinear equation is satisfied for the pair  $(y^h, \bar{u}_h)$ . Consequently, this term vanishes. The pair  $(\bar{y}_h, \bar{u}_h)$  fulfills only the equation in  $V_h$ . However, the test function  $\bar{p}_h$  belongs to  $V_h$ . Hence, this term vanishes, too. Consequently, the assertion is true with  $c_j = c_{L^2} c_S (\|\bar{y}_h - y_d\|_{L^2(\Omega)} + \frac{1}{2} c_{L^2} c_S \|\bar{u}_h\|_{L^2(\Omega)})$ .  $\square$

LEMMA 6.3. *The error in the adjoint equation can be estimated by*

$$(6.7) \quad |L_y(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h)| \leq c_y h^2 \|\bar{y}_h - y_d\|_{L^2(\Omega)} \|y\|_{H^2(\Omega)} \quad \forall y \in V$$

with  $c_y = c_p(c_{H^1} \delta_1 + c_{L^2} c_f)$ .

*Proof.* We start with

$$L_y(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h) = (\bar{y}_h - y_d, y - \bar{y}_h) - a(\bar{p}_h, y - \bar{y}_h) - (f'(\bar{y}_h) \bar{p}_h, y - \bar{y}_h).$$

Moreover, we define a function  $p^h$  as a solution of

$$(6.8) \quad a(p^h, v) + (f'(\bar{y}_h) p^h, v) = (\bar{y}_h - y_d, v) \quad \forall v \in V.$$

We obtain for  $v = y - \bar{y}_h$ ,

$$(\bar{y}_h - y_d, y - \bar{y}_h) = a(p^h, y - \bar{y}_h) + (f'(\bar{y}_h) p^h, y - \bar{y}_h).$$

Therefore, we can continue using Galerkin orthogonality as follows:

$$\begin{aligned} L_y(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h) &= a(p^h - \bar{p}_h, y - \bar{y}_h) + (f'(\bar{y}_h)(p^h - \bar{p}_h), y - \bar{y}_h) \\ &= a(p^h - \bar{p}_h, y - i_h y) + (f'(\bar{y}_h)(p^h - \bar{p}_h), y - i_h y) \\ &\quad + a(p^h - \bar{p}_h, i_h y - \bar{y}_h) + (f'(\bar{y}_h)(p^h - \bar{p}_h), i_h y - \bar{y}_h) \\ &= a(p^h - \bar{p}_h, y - i_h y) + (f'(\bar{y}_h)(p^h - \bar{p}_h), y - i_h y). \end{aligned}$$

Consequently, we obtain

$$\begin{aligned} |L_y(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h)| &\leq \delta_1 \|p^h - \bar{p}_h\|_{H^1(\Omega)} \|y - i_h y\|_{H^1(\Omega)} \\ &\quad + c_f \|p^h - \bar{p}_h\|_{L^2(\Omega)} \|y - i_h y\|_{L^2(\Omega)}. \end{aligned}$$

The proving technique of Lemma 5.1 delivers

$$(6.9) \quad \|p^h - \bar{p}_h\|_{H^1(\Omega)} \leq c_{H^1} h \|p^h\|_{H^2(\Omega)} \quad \text{and} \quad \|p^h - \bar{p}_h\|_{L^2(\Omega)} \leq c_{L^2} h^2 \|p^h\|_{H^2(\Omega)}.$$

Moreover, we can apply the inequalities (5.1) and (5.2) for the interpolation error of  $y$ . Finally, we get

$$|L_y(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h)| \leq (c_{H^1} \delta_1 + c_{L^2} c_f h^2) h^2 \|p^h\|_{H^2(\Omega)} \|y\|_{H^2(\Omega)}.$$

Of course, the norm  $\|p^h\|_{H^2(\Omega)}$  can be estimated by means of Lemma 4.4,

$$(6.10) \quad \|p^h\|_{H^2(\Omega)} \leq c_p \|\bar{y}_h - y_d\|_{L^2(\Omega)},$$

and (6.7) is obtained with  $c_y = c_p(c_{H^1}\delta_1 + c_{L^2}c_f)$  if  $h \leq 1$ .  $\square$

LEMMA 6.4. *The error in the optimality condition can be estimated by*

$$(6.11) \quad L_u(\bar{y}_h, \bar{u}_h, \bar{p}_h)(u - \bar{u}_h) \geq -c_u h \|u - \bar{u}_h\|_{L^2(\Omega)} \|\bar{p}_h\|_{W^{1,\infty}(\Omega)} \quad \forall u \in U_{ad}$$

with  $c_u = |T_h^i|^{1/2}$ , where  $T_h^i$  is the set of elements, where  $\bar{p}_h + \nu \bar{u}_h$  changes the sign.

*Proof.* The optimality condition for  $\bar{u}_h$  is given by

$$(6.12) \quad (\bar{p}_h + \nu \bar{u}_h, u_h - \bar{u}_h) \geq 0 \quad \forall u_h \in U_{ad}^h.$$

This implies  $u_h = a$  on all elements  $T_h$  where  $\bar{p}_h + \nu \bar{u}_h$  is a.e. positive. Analogously  $u_h = b$  holds on all elements  $T_h$  where  $\bar{p}_h + \nu \bar{u}_h$  is a.e. negative. For the set  $T_h^a$  of all such elements we find for arbitrary  $u \in U_{ad}$

$$(6.13) \quad (\bar{p}_h + \nu \bar{u}_h, u - \bar{u}_h)_{L^2(T_h^a)} \geq 0 \quad \forall u \in U_{ad}.$$

It remains to estimate the error on the set  $T_h^i$  of elements where the expression  $\bar{p}_h + \nu \bar{u}_h$  changes the sign. Since  $\bar{u}_h$  is constant on each element, we find

$$|\bar{p}_h + \nu \bar{u}_h| \leq h \|\bar{p}_h\|_{W^{1,\infty}(\Omega)} \quad \text{on } T_h^i.$$

From this, we conclude

$$(6.14) \quad (\bar{p}_h + \nu \bar{u}_h, u - \bar{u}_h)_{L^2(T_h^i)} \geq -|T_h^i|^{1/2} h \|u - \bar{u}_h\|_{L^2(T_h^i)} \|\bar{p}_h\|_{W^{1,\infty}(\Omega)} \quad \forall u \in U_{ad}.$$

Combining (6.13) and (6.14), (6.11) is obtained with  $c_u = |T_h^i|^{1/2}$ .  $\square$

LEMMA 6.5. *Let  $u \in U_{ad}$  be given together with the associated solution  $y$  of the semilinear state equation (2.1). Then it holds with  $r = \|u - \bar{u}_h\|_{L^2(\Omega)}$  that*

$$(6.15) \quad L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h, u - \bar{u}_h)^2 \geq \delta r^2 - d_1 h r^2 - d_2 h^2 r - d_3 r^3 - d_4 h^4$$

with constants  $d_i$  specified below; see (6.21).

*Proof.* At first, let us define  $d$  as the solution of the linearized equation

$$a(d, v) + (f'(\bar{y}_h)d, v) = (u - \bar{u}_h, v) \quad \forall v \in V.$$

Observe that  $d$  can be used as a test function in the coercivity condition (3.4), whereas  $y - \bar{y}_h$  would not be suitable there. Furthermore, we have by Corollary 4.3 the estimate

$$(6.16) \quad \|d\|_{L^2(\Omega)} \leq c_L \|u - \bar{u}_h\|_{L^2(\Omega)}.$$

Now, we rewrite the left-hand side in (6.15) as

$$(6.17) \quad \begin{aligned} L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h, u - \bar{u}_h)^2 \\ = L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(d, u - \bar{u}_h)^2 + L_{yy}(\bar{y}_h, \bar{u}_h, \bar{p}_h)[y - \bar{y}_h - d, y - \bar{y}_h + d]. \end{aligned}$$

The first addend gives the coercivity  $\geq \delta r^2$  using the sufficient condition (3.4). The second one can be estimated using (2.8) by

$$(6.18) \quad \begin{aligned} |L_{yy}(\bar{y}_h, \bar{u}_h, \bar{p}_h)[y - \bar{y}_h - d, y - \bar{y}_h + d]| \\ \leq (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) \|y - \bar{y}_h - d\|_{L^2(\Omega)} \|y - \bar{y}_h + d\|_{L^2(\Omega)}. \end{aligned}$$

Using (4.4), (5.5), and (6.16), we find

$$\begin{aligned}
 \|y - \bar{y}_h + d\|_{L^2(\Omega)} &\leq \|y - y^h\|_{L^2(\Omega)} + \|y^h - \bar{y}_h\|_{L^2(\Omega)} + \|d\|_{L^2(\Omega)} \\
 (6.19) \quad &\leq c_L \|u - \bar{u}_h\|_{L^2(\Omega)} + c_{L^2} c_S h^2 \|\bar{u}_h\|_{L^2(\Omega)} + c_L \|u - \bar{u}_h\|_{L^2(\Omega)} \\
 &= 2c_L r + c_{L^2} c_S \|\bar{u}_h\|_{L^2(\Omega)} h^2.
 \end{aligned}$$

We could treat  $y - \bar{y}_h - d$  in the same way. However, we need a sharper result. To this end, we use the splitting  $y - \bar{y}_h - d = (y - y^h - d) + (y^h - \bar{y}_h)$ .

The first function  $y - y^h - d =: y_1$  is the weak solution of

$$a(y_1, v) + (f'(\bar{y}_h)y_1, v) = -(f(y) - f(y^h) - f'(\bar{y}_h)(y - y^h), v) \quad \forall v \in V.$$

We transform the right-hand side into

$$\begin{aligned}
 f(y) - f(y^h) - f'(\bar{y}_h)(y - y^h) &= (f'(y^h) - f'(\bar{y}_h))(y - y^h) \\
 &\quad + \int_0^1 [f'(y^h + s(y - y^h)) - f'(y^h)](y - y^h) ds.
 \end{aligned}$$

Its  $L^2$ -norm is estimated by

$$\begin{aligned}
 \|f(y) - f(y^h) - f'(\bar{y}_h)(y - y^h)\|_{L^2(\Omega)} \\
 \leq c_{f'} \left( \|y^h - \bar{y}_h\|_{L^4(\Omega)} \|y - y^h\|_{L^4(\Omega)} + \frac{1}{2} \|y - y^h\|_{L^4(\Omega)}^2 \right).
 \end{aligned}$$

Hence, we can estimate

$$\begin{aligned}
 \|y_1\|_{L^2(\Omega)} &= \|y - d - y^h\|_{L^2(\Omega)} \leq c_L \|f(y) - f(y^h) - f'(\bar{y}_h)(y - y^h)\|_{L^2(\Omega)} \\
 &\leq c_L c_{f'} \left( \|y^h - \bar{y}_h\|_{L^4(\Omega)} \|y - y^h\|_{L^4(\Omega)} + \frac{1}{2} \|y - y^h\|_{L^4(\Omega)}^2 \right).
 \end{aligned}$$

Analogously to (5.5) we get an  $L^4$ -error estimate for  $y - y^h$  by Lemma 5.1,

$$\|y^h - \bar{y}_h\|_{L^4(\Omega)} \leq I_4 c_{H^1} h c_S \|\bar{u}_h\|_{L^2(\Omega)}.$$

Applying Corollary 4.3, we find

$$\|y - y^h\|_{L^4(\Omega)} \leq I_4 \|y - y^h\|_{H^1(\Omega)} \leq I_4 c_L r.$$

Altogether, we derived the estimate

$$\|y - d - y^h\|_{L^2(\Omega)} \leq c_L^2 c_{f'} I_4^2 \left( c_{H^1} c_S \|\bar{u}_h\|_{L^2(\Omega)} h r + \frac{1}{2} c_L r^2 \right),$$

which yields

$$\begin{aligned}
 (6.20) \quad \|y - \bar{y}_h - d\|_{L^2(\Omega)} &\leq \|y - y^h - d\|_{L^2(\Omega)} + \|y^h - \bar{y}_h\|_{L^2(\Omega)} \\
 &\leq c_L^2 c_{f'} I_4^2 \left( c_{H^1} c_S \|\bar{u}_h\|_{L^2(\Omega)} h r + \frac{1}{2} c_L r^2 \right) + c_{L^2} c_S h^2 \|\bar{u}_h\|_{L^2(\Omega)}.
 \end{aligned}$$

Now we can proceed with the estimation of  $L_{yy}$  already started in (6.18) using the inequalities (6.19) and (6.20),

$$\begin{aligned}
 & |L_{yy}(\bar{y}_h, \bar{u}_h, \bar{p}_h)[y - \bar{y}_h - d, y - \bar{y}_h + d]| \\
 & \leq (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) \|y - \bar{y}_h + d\|_{L^2(\Omega)} \|y - \bar{y}_h - d\|_{L^2(\Omega)} \\
 & \leq (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) (2c_L r + c_{L^2} c_S \|\bar{u}_h\|_{L^2(\Omega)} h^2) \\
 & \quad \cdot \left( c_L^2 c_{f'} I_4^2 \left( c_{H^1} c_S \|\bar{u}_h\|_{L^2(\Omega)} h r + \frac{1}{2} c_L r^2 \right) + c_{L^2} c_S h^2 \|\bar{u}_h\|_{L^2(\Omega)} \right) \\
 & \leq d_1 h r^2 + d_2 h^2 r + d_3 r^3 + d_4 h^4
 \end{aligned}$$

with constants defined by

$$\begin{aligned}
 (6.21) \quad & d_0 = (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}), \\
 & d_1 = d_0 \cdot c_L^2 c_{f'} I_4^2 \cdot \left( c_{H^1} c_S \|\bar{u}_h\|_{L^2(\Omega)} \cdot 2c_L + \frac{1}{2} c_L c_{L^2} c_S \|\bar{u}_h\|_{L^2(\Omega)} \right), \\
 & d_2 = d_0 \cdot (c_L^2 c_{f'} I_4^2 \cdot c_{H^1} c_S \|\bar{u}_h\|_{L^2(\Omega)} \cdot c_{L^2} c_S \|\bar{u}_h\|_{L^2(\Omega)} + 2c_L \cdot c_{L^2} c_S \|\bar{u}_h\|_{L^2(\Omega)}), \\
 & d_3 = d_0 \cdot c_L^2 c_{f'} I_4^2 \cdot c_L^2, \\
 & d_4 = d_0 \cdot (c_{L^2} c_S \|\bar{u}_h\|_{L^2(\Omega)})^2.
 \end{aligned}$$

Here, we used again  $h < 1$ . Finally, we obtain for the second derivative

$$L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h, u - \bar{u}_h)^2 \geq \delta r^2 - d_1 h r^2 - d_2 h^2 r - d_3 r^3 - d_4 h^4,$$

and the claim is proved.  $\square$

LEMMA 6.6. *Assume that (2.6) holds. Then, the estimate*

$$(6.22) \quad J(y, u) - J(y^h, \bar{u}_h) \geq \delta r^2 - a_1 h^2 - a_2 h r - a_3 h^2 r - a_4 r^3 - a_5 h r^2$$

is valid for all  $(y, u)$  satisfying (2.1) and  $\|u - \bar{u}_h\|_{L^2(\Omega)} = r$ . The constants  $a_i$  are computed in the course of the proof; see (6.26).

*Proof.* Since  $(y, u)$  and  $(y^h, \bar{u}_h)$  fulfill (2.1), we find

$$\begin{aligned}
 J(y, u) - J(y^h, \bar{u}_h) &= L(y, u, \bar{p}_h) - L(y^h, \bar{u}_h, \bar{p}_h) \\
 &= L(y, u, \bar{p}_h) - L(\bar{y}_h, \bar{u}_h, \bar{p}_h) \\
 (6.23) \quad &+ L(\bar{y}_h, \bar{u}_h, \bar{p}_h) - L(y^h, \bar{u}_h, \bar{p}_h).
 \end{aligned}$$

The second difference was already estimated in Lemma 6.2. We will now focus on the first difference,

$$\begin{aligned}
 L(y, u, \bar{p}_h) - L(\bar{y}_h, \bar{u}_h, \bar{p}_h) &= L_y(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h) + L_u(\bar{y}_h, \bar{u}_h, \bar{p}_h)(u - \bar{u}_h) \\
 (6.24) \quad &+ \frac{1}{2} L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y - \bar{y}_h, u - \bar{u}_h)^2 + r_2.
 \end{aligned}$$

Note that the quadratic part of the objective is approximated exactly. Only the nonlinear function  $f$  causes the remainder part  $r_2$ . We estimated all addends in Lemmas 6.3, 6.4, and 6.5. It remains to investigate the remainder term. Here, we find

$$\begin{aligned}
 (6.25) \quad |r_2| &= \left| \int_{\Omega} \int_0^1 \int_0^s (f''(\bar{y}_h + t(y - \bar{y}_h)) - f''(\bar{y}_h)) (y - \bar{y}_h)^2 dt ds \bar{p}_h dx \right| \\
 &\leq \frac{1}{6} c_{f''} \|y - \bar{y}_h\|_{L^3(\Omega)}^3 \|\bar{p}_h\|_{L^\infty(\Omega)}.
 \end{aligned}$$

We estimate the right-hand side of (6.25) using Lemma 4.2 and Corollary 5.2 as

$$\begin{aligned} \|y - \bar{y}_h\|_{L^3(\Omega)} &\leq \|y - y^h\|_{L^3(\Omega)} + \|y^h - \bar{y}_h\|_{L^3(\Omega)} \\ &\leq c_L I_3 \|u - \bar{u}_h\|_{L^2(\Omega)} + c_{H^1} c_S I_3 h \|\bar{u}_h\|_{L^2(\Omega)}. \end{aligned}$$

Using  $\|u - \bar{u}_h\|_{L^2(\Omega)} = r$  and  $(a + b)^3 \leq 4(a^3 + b^3)$  for positive  $a, b$ , we obtain

$$|r_2| \leq \frac{2}{3} c_{f''} \|\bar{p}_h\|_{L^\infty(\Omega)} (c_L^3 I_3^3 r^3 + (c_{H^1} c_S I_3 \|\bar{u}_h\|_{L^2(\Omega)})^3 h^3) =: c_{r1} r^3 + c_{r2} h^3.$$

By means of Lemmas 6.2, 6.3, 6.4, and 6.5 and (6.25) we get

$$\begin{aligned} J(y, u) - J(y^h, \bar{u}_h) &\geq -c_j \|\bar{u}_h\|_{L^2(\Omega)} h^2 - c_y \|\bar{y}_h - y_d\|_{L^2(\Omega)} \|y\|_{H^2(\Omega)} h^2 \\ &\quad - c_u \|\bar{p}_h\|_{W^{1,\infty}(\Omega)} h r + \delta r^2 - d_1 h r^2 - d_2 h^2 r - d_3 r^3 - d_4 h^4 - c_{r1} r^3 - c_{r2} h^3. \end{aligned}$$

Here, it remains to estimate  $\|y\|_{H^2(\Omega)}$ :

$$\|y\|_{H^2(\Omega)} \leq c_S \|u\|_{L^2(\Omega)} \leq c_S (\|\bar{u}_h\|_{L^2(\Omega)} + r).$$

Consequently, we obtain

$$\begin{aligned} J(y, u) - J(y^h, \bar{u}_h) &\geq \delta r^2 - (c_y \|\bar{y}_h - y_d\|_{L^2(\Omega)} c_S \|\bar{u}_h\|_{L^2(\Omega)} + c_{r2} + d_4 h^2) h^2 \\ &\quad - (c_u \|\bar{p}_h\|_{W^{1,\infty}(\Omega)}) h r - (c_y \|\bar{y}_h - y_d\|_{L^2(\Omega)} c_S + d_2) h^2 r - (d_3 + c_{r1}) r^3 - d_1 h r^2. \end{aligned}$$

Setting

$$\begin{aligned} a_1 &:= c_j \|\bar{u}_h\|_{L^2(\Omega)} + c_y \|\bar{y}_h - y_d\|_{L^2(\Omega)} c_S \|\bar{u}_h\|_{L^2(\Omega)} + c_{r2} + d_4, \\ a_2 &:= c_u \|\bar{p}_h\|_{W^{1,\infty}(\Omega)}, \\ (6.26) \quad a_3 &:= c_y \|\bar{y}_h - y_d\|_{L^2(\Omega)} c_S + d_2, \\ a_4 &:= d_3 + c_{r1}, \\ a_5 &:= d_1, \end{aligned}$$

the assertion is obtained.  $\square$

Now, let us fix a mesh-size  $h$  with associated solution  $\bar{u}_h$ . Suppose that the polynomial given by (6.22) is positive for some  $r > 0$ . Then we have that the value  $J(y, u)$  is greater than the  $J(y^h, \bar{u}_h)$  for all  $u$  having  $L^2$ -distance  $r$  to  $\bar{u}_h$ . Hence, there exists a local minimum of the optimal control problem (1.1)–(1.3) inside the neighborhood  $\{u : \|u - \bar{u}_h\|_{L^2(\Omega)} < r\}$  of  $\bar{u}_h$ .

**COROLLARY 6.7.** *Assume that there exists a positive value  $r$  such that*

$$-a_4 r^3 + (\delta - a_5 h) r^2 - (a_2 h + a_3 h^2) r - a_1 h^2 > 0.$$

*Then*

$$(6.27) \quad J(y, u) - J(y^h, \bar{u}_h) > 0$$

*holds for all  $(y, u)$  satisfying (2.1) and  $\|u - \bar{u}_h\|_{L^2(\Omega)} = r$ .*

A sufficient condition that the assumption of the previous Corollary 6.7 is fulfilled is given as the last result in this section.

COROLLARY 6.8. *Let us assume that  $\sigma := \delta - a_5 h > 0$  holds. Let us suppose further that there exists  $r_+ > 0$  that fulfills*

$$(6.28) \quad r_+ > \max \left\{ \frac{3(a_2 h + a_3 h^2)}{\sigma}, \sqrt{\frac{3a_1 h^2}{\sigma}} \right\},$$

and if  $a_4 > 0$  additionally, then

$$(6.29) \quad r_+ < \frac{\sigma}{3a_4}.$$

Then the prerequisite of Corollary 6.7 is satisfied.

*Proof.* The first condition, (6.28), gives

$$\frac{2}{3}(\delta - a_5 h)r_+^2 = \frac{2}{3}\sigma r_+^2 > (a_2 h + a_3 h^2)r_+ + a_1 h^2 \geq 0.$$

If  $a_4$  is not zero, then the second one implies  $\frac{1}{3}\sigma r_+^2 > a_4 r_+^3 > 0$ . Now, the polynomial  $-a_4 r^3 + (\delta - a_5 h)r^2 - (a_2 h + a_3 h^2)r - a_1 h^2$  admits a positive value for  $r_+$ . Thus, it satisfies the assumptions of the previous Corollary 6.7.  $\square$

**Remark 6.9.** The assumptions of Corollary 6.7 directly link the discretization parameter  $h$  to the coercivity factor  $\delta$ . The discretization has to be fine enough; i.e.,  $h$  has to be small enough, such that  $\delta - a_5 h > 0$  can be fulfilled. Hence, if the sufficient condition (2.6) holds for the solution of the original, continuous problem, then for sufficiently small  $h$  the assumptions of Corollary 6.7 will be satisfied.

**7. Verification of optimality conditions.** In the following,  $r$  denotes a fixed radius fulfilling the assumptions of Corollary 6.7. We are now going to prove that in an  $r$ -neighborhood of  $\bar{u}_h$  there exists a solution of the optimality system connected with the continuous problem (1.1)–(1.3).

**THEOREM 7.1.** *There exist at least one control  $\hat{u}$  with associated state  $\hat{y}$  and adjoint state  $\hat{p}$  in an  $r$ -neighborhood of  $\bar{u}_h$  fulfilling the first-order necessary optimality conditions (2.3).*

*Proof.* We investigate the optimal control problem for

$$U_{ad}^r := U_{ad} \cap \{u \in U : \|u - \bar{u}_h\|_{L^2(\Omega)} \leq r\}.$$

This set is weakly compact. Hence, there exists at least one solution  $\hat{u}$  of the modified problem. Since  $\bar{u}_h$  is feasible and

$$J(y, u) - J(y^h, \bar{u}_h) > 0$$

is satisfied for all controls  $u$  with  $\|u - \bar{u}_h\|_{L^2(\Omega)} = r$ , we have  $\|\hat{u} - \bar{u}_h\|_{L^2(\Omega)} < r$ . Consequently, the local minimizer  $\hat{u}$  of the modified problem is also a local minimizer for the original problem. In particular,  $\hat{u}$  has to fulfill the first-order necessary optimality conditions.  $\square$

As a consequence, we can give an error estimate for the associated state  $\hat{y}$  and adjoint state  $\hat{p}$  in terms of  $r$  and  $h$ .

**LEMMA 7.2.** *Let  $(\hat{y}, \hat{u}, \hat{p})$  fulfill the first-order necessary optimality conditions. Then one can estimate the distance to the discrete solution by*

$$(7.1) \quad \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} \leq c_{y1} r + c_{y2} h^2$$

and

$$(7.2) \quad \|\hat{p} - \bar{p}_h\|_{L^2(\Omega)} \leq c_{p1} r + c_{p2} h^2,$$

with constants  $c_{yi}$  and  $c_{pi}$  independent of  $h$ ,  $r$ , and  $(\hat{y}, \hat{u}, \hat{p})$ :

$$\begin{aligned} c_{y1} &= c_L, \quad c_{y2} = c_{L^2} c_S \|\bar{u}_h\|_{L^2(\Omega)}, \quad c_{p1} = c_L (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) c_{y1}, \\ c_{p2} &= c_L (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) c_{y2} + c_p c_{L^2} \|\bar{y}_h - y_d\|_{L^2(\Omega)}. \end{aligned}$$

*Proof.* The difference  $\hat{y} - \bar{y}_h$  can be treated using the auxiliary function  $y^h$  defined in (3.8),

$$\begin{aligned} \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} &\leq \|\hat{y} - y^h\|_{L^2(\Omega)} + \|y^h - \bar{y}_h\|_{L^2(\Omega)} \\ &\leq c_L r + c_{L^2} c_S h^2 \|\bar{u}_h\|_{L^2(\Omega)}. \end{aligned}$$

Here, we applied the estimate (5.5). The claim (7.1) follows with  $c_{y1} = c_L$  and  $c_{y2} = c_{L^2} c_S \|\bar{u}_h\|_{L^2(\Omega)}$ .

For the estimation of the adjoint states, recall the definition of  $p^h$  in (6.8) and the corresponding estimate (6.9),

$$\|p^h - \bar{p}_h\|_{L^2(\Omega)} \leq c_{L^2} h^2 \|p^h\|_{H^2(\Omega)} \leq c_p c_{L^2} h^2 \|\bar{y}_h - y_d\|_{L^2(\Omega)}.$$

Now, we introduce the splitting

$$\|\hat{p} - \bar{p}_h\|_{L^2(\Omega)} \leq \|\hat{p} - p^h\|_{L^2(\Omega)} + \|p^h - \bar{p}_h\|_{L^2(\Omega)},$$

and it remains to investigate  $d := \hat{p} - p^h$ . It is a solution of

$$a(d, v) + (f'(\hat{y})d, v) = (\hat{y} - \bar{y}_h, v) - ((f'(\hat{y}) - f'(\bar{y}_h))\bar{p}_h, v) \quad \forall v \in V.$$

Hence, we obtain

$$\begin{aligned} \|\hat{p} - p^h\|_{L^2(\Omega)} &\leq c_L (\|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} + \|(f'(\hat{y}) - f'(\bar{y}_h))\bar{p}_h\|_{L^2(\Omega)}) \\ &\leq c_L (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)}. \end{aligned}$$

The estimate (7.2) is satisfied with  $c_{p1} = c_L (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) c_{y1}$  and  $c_{p2} = c_L (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) c_{y2} + c_p c_{L^2} \|\bar{y}_h - y_d\|_{L^2(\Omega)}$ .  $\square$

With the same technique, we get an  $H^1$ -error estimate for the state and adjoint state.

**COROLLARY 7.3.** *Under the assumptions of the previous lemma, we have*

$$\|\hat{y} - \bar{y}_h\|_{H^1(\Omega)} \leq c_L r + c_{H^1} c_S \|\bar{u}_h\|_{L^2(\Omega)} h$$

and

$$\|\hat{p} - \bar{p}_h\|_{H^1(\Omega)} \leq c_L (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) \|\hat{y} - \bar{y}_h\|_{H^1(\Omega)} + c_p c_{H^1} \|\bar{y}_h - y_d\|_{L^2(\Omega)} h.$$

Next, we will introduce additional assumptions on  $r$  that guarantee that the control  $\hat{u}$  fulfills a second-order sufficient optimality condition.

**THEOREM 7.4.** *Under the assumption*

$$r < \frac{\delta - c_h h^2}{c_r}$$

every control  $\hat{u}$  in an  $r$ -neighborhood of  $\bar{u}_h$  fulfills, together with its associated state  $\hat{y}$  and adjoint  $\hat{p}$ , the coercivity

$$(7.3) \quad L''_{(y,u)}(\hat{y}, \hat{u}, \hat{p})(y, u)^2 \geq \delta' \|u\|_{L^2(\Omega)}^2$$

for all  $y$  given as a solution of

$$(7.4) \quad a(y, v) + (f'(\hat{y})y, v) = (u, v) \quad \forall v \in V.$$

Moreover, there exists only one local minimum in this neighborhood.

The constants  $c_r$  and  $c_h$  will be determined in the course of the proof.

*Proof.* Let  $\hat{u}$  be a stationary point with associated state  $\hat{y}$  and adjoint state  $\hat{p}$ . Further, let  $u \in U$  be an arbitrary control with associated solutions  $y$  and  $y^h$  of the linearized equations (7.4) and (3.5), respectively. Then the pair  $(y, u)$  is suitable in (7.3), whereas  $(y^h, u)$  can be utilized as test functions in (3.4).

At first, we have to estimate the difference  $d := y - y^h$  as well as the sum  $y + y^h$  in terms of  $u$  and  $h$ . The difference fulfills the equation

$$0 = a(d, v) + (f'(\hat{y})y, v) - (f'(\bar{y}_h)y^h, v) = a(d, v) + (f'(\hat{y})d, v) + ((f'(\hat{y}) - f'(\bar{y}_h))y^h, v)$$

for all  $v \in V$ . Testing with  $d$  itself, we obtain

$$\|d\|_{H^1(\Omega)} \leq \frac{I_4}{\delta_0} \|(f'(\hat{y}) - f'(\bar{y}_h))y^h\|_{L^{4/3}(\Omega)}.$$

Hence, we can estimate  $d$  as a solution of a linearized equation with right-hand side  $-(f'(\hat{y}) - f'(\bar{y}_h))y^h$ , which is estimated by

$$\begin{aligned} \|(f'(\hat{y}) - f'(\bar{y}_h))y^h\|_{L^{4/3}(\Omega)} &\leq c_{f'} \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} \|y^h\|_{L^4(\Omega)} \\ &\leq c_{f'} \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} c_L I_4 \|u\|_{L^2(\Omega)}. \end{aligned}$$

Collecting all these inequalities, we find for  $d = y - y^h$

$$\begin{aligned} \|y - y^h\|_{L^2(\Omega)} &\leq c_L I_4 \|(f'(\hat{y}) - f'(\bar{y}_h))y^h\|_{L^2(\Omega)} \\ &\leq c_L \delta_0^{-1} I_4^2 c_{f'} \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)}. \end{aligned}$$

The sum  $y + y^h$  is estimated by

$$\|y + y^h\|_{L^2(\Omega)} \leq 2c_L \|u\|_{L^2(\Omega)}.$$

Now, we can investigate the second derivative of the Lagrangian. We start with the decomposition

$$\begin{aligned} L''_{(y,u)}(\hat{y}, \hat{u}, \hat{p})(y, u)^2 &= L''_{(y,u)}(\hat{y}, \hat{u}, \hat{p})(y, u)^2 - L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y, u)^2 \\ &\quad + L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y, u)^2 - L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y^h, u)^2 \\ &\quad + L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y^h, u)^2. \end{aligned}$$

The last addend gives us the desired coercivity by (3.4). So we have to estimate the



two differences in this equation. The first one yields

$$\begin{aligned}
 & \left| L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y, u)^2 - L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y^h, u)^2 \right| \\
 &= \left| (y - y^h, y + y^h) + \int_{\Omega} f''(\bar{y}_h) \bar{p}_h (y - y^h)(y + y^h) \right| \\
 &\leq (1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) \|y - y^h\|_{L^2(\Omega)} \|y + y^h\|_{L^2(\Omega)} \\
 &\leq \underbrace{(1 + c_{f'} \|\bar{p}_h\|_{L^\infty(\Omega)}) 2c_L^2 \delta_0^{-1} I_4^2 c_{f'}}_{C_1} \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)}^2 \\
 &\leq C_1 \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)}^2.
 \end{aligned}$$

Let us proceed with the second difference,

$$\begin{aligned}
 & \left| L''_{(y,u)}(\hat{y}, \hat{u}, \hat{p})(y, u)^2 - L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y, u)^2 \right| \\
 &= \left| \int_{\Omega} (f''(\hat{y})\hat{p} - f''(\bar{y}_h)\bar{p}_h) y^2 dx \right| \leq \|f''(\hat{y})\hat{p} - f''(\bar{y}_h)\bar{p}_h\|_{L^2(\Omega)} \|y\|_{L^4(\Omega)}^2 \\
 &\leq c_L^2 I_4^2 \|f''(\hat{y})\hat{p} - f''(\bar{y}_h)\bar{p}_h\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)}^2.
 \end{aligned}$$

Using Lipschitz estimates, we obtain for the right-hand side,

$$\begin{aligned}
 \|f''(\hat{y})\hat{p} - f''(\bar{y}_h)\bar{p}_h\|_{L^2(\Omega)} &\leq \|f''(\hat{y})\hat{p} - f''(\hat{y})\bar{p}_h\|_{L^2(\Omega)} + \|f''(\hat{y})\bar{p}_h - f''(\bar{y}_h)\bar{p}_h\|_{L^2(\Omega)} \\
 &\leq c_{f'} \|\hat{p} - \bar{p}_h\|_{L^2(\Omega)} + c_{f''} \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} \|\bar{p}_h\|_{L^\infty(\Omega)}.
 \end{aligned}$$

Hence, we arrive at

$$\begin{aligned}
 & \left| L''_{(y,u)}(\hat{y}, \hat{u}, \hat{p})(y, u)^2 - L''_{(y,u)}(\bar{y}_h, \bar{u}_h, \bar{p}_h)(y, u)^2 \right| \\
 &\leq c_L^2 I_4^2 (c_{f'} \|\hat{p} - \bar{p}_h\|_{L^2(\Omega)} + c_{f''} \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} \|\bar{p}_h\|_{L^\infty(\Omega)}) \|u\|_{L^2(\Omega)}^2.
 \end{aligned}$$

So, we find

$$\begin{aligned}
 L''_{(y,u)}(\hat{y}, \hat{u}, \hat{p})(y, u)^2 &\geq \left\{ \delta - (c_L^2 I_4^2 c_{f'}) \|\hat{p} - \bar{p}_h\|_{L^2(\Omega)} \right. \\
 &\quad \left. - (C_1 + c_L^2 I_4^2 c_{f''} \|\bar{p}_h\|_{L^\infty(\Omega)}) \|\hat{y} - \bar{y}_h\|_{L^2(\Omega)} \right\} \|u\|_{L^2(\Omega)}^2.
 \end{aligned}$$

Here, we can apply the estimates of  $\|\hat{p} - \bar{p}_h\|_{L^2(\Omega)}$  and  $\|\hat{y} - \bar{y}_h\|_{L^2(\Omega)}$  of Lemma 7.2. We get consequently,

$$L''_{(y,u)}(\hat{y}, \hat{u}, \hat{p})(y, u)^2 \geq \{\delta - c_r r - c_h h^2\} \|u\|_{L^2(\Omega)}^2$$

with

$$c_r = C_1 c_{y_1} + c_L^2 I_4^2 (c_{f'} c_{p1} + c_{f''} \|\bar{p}_h\|_{L^\infty(\Omega)} c_{y1})$$

and

$$c_h = C_1 c_{y_2} + c_L^2 I_4^2 (c_{f'} c_{p2} + c_{f''} \|\bar{p}_h\|_{L^\infty(\Omega)} c_{y2}),$$

and the claim is proved.  $\square$

Finally, the sufficient condition provided by the previous theorem gives local optimality of  $\hat{u}$ . Moreover, the method of proof yields an estimate of the neighborhood of  $\hat{u}$ , where  $\hat{u}$  is locally optimal.

**THEOREM 7.5.** *Let the assumptions of Theorem 7.4 be satisfied. Then it holds with some  $\hat{\delta} > 0$  that*

$$J(y, u) - J(\hat{y}, \hat{u}) \geq \hat{\delta} \|u - \hat{u}\|_{L^2(\Omega)}^2$$

for all admissible  $u \in U_{ad}$  with  $\|u - \hat{u}\|_{L^2(\Omega)} < r'$  for some sufficiently small  $r'$ .

*Proof.* By assumption,  $(\hat{y}, \hat{u}, \hat{p})$  fulfills the necessary optimality conditions together with the coercivity relation (7.3). Let  $(y, u)$  be another admissible pair. We have

$$J(\hat{y}, \hat{u}) = L(\hat{y}, \hat{u}, \hat{p}) \quad \text{and} \quad J(y, u) = L(y, u, \hat{p}),$$

since  $(\hat{y}, \hat{u})$  and  $(y, u)$  are admissible. Taylor expansion of the Lagrange function yields

$$\begin{aligned} J(y, u) - J(\hat{y}, \hat{u}) &= L(y, u, \hat{p}) - L(\hat{y}, \hat{u}, \hat{p}) \\ (7.5) \quad &= L_y(\hat{y}, \hat{u}, \hat{p})(y - \hat{y}) + L_u(\hat{y}, \hat{u}, \hat{p})(u - \hat{u}) \\ &\quad + \frac{1}{2} L''(\hat{y}, \hat{u}, \hat{p})[(y - \hat{y}, u - \hat{u})]^2 + r_2. \end{aligned}$$

Since the necessary conditions (2.3) are satisfied at  $\hat{y}, \hat{u}$  with adjoint state  $\hat{p}$ , the first term vanishes. The second addend is nonnegative due to the variational inequality. The remainder term  $r_2$  satisfies (cf. (6.25))

$$(7.6) \quad |r_2| \leq \frac{1}{6} c_{f''} \|y - \hat{y}\|_{L^4(\Omega)}^3 \|\hat{p}\|_{L^4(\Omega)} \leq \frac{1}{6} c_{f''} c_L^3 I_4^4 \|u - \hat{u}\|_{L^2(\Omega)}^3 \|\hat{p}\|_{H^1(\Omega)}.$$

The pair  $(y - \hat{y}, u - \hat{u})$  is not suitable as a test function in (7.3), since  $y - \hat{y}$  is not the solution of a linearized equation. Let us introduce an auxiliary state  $d$  as the weak solution of

$$a(d, v) + (f'(\hat{y})d, v) = (u - \hat{u}, v) \quad \forall v \in V.$$

When we use  $d$  instead of  $y - \hat{y}$ , we make the small error  $r_1 := (y - \hat{y}) - d$ , which is itself the weak solution of

$$a(r_1, v) + (f_y^{\hat{y}} r_1, v) = ((f_y^{\hat{y}} - f'(\hat{y}))d, v) \quad \forall v \in V$$

with  $f_y^{\hat{y}} = \int_0^1 f'(\hat{y} + s(y - \hat{y})) ds$ . Since  $f'$  and  $f_y^{\hat{y}}$  are nonnegative, we get the estimate

$$\begin{aligned} (7.7) \quad \|r_1\|_{H^1(\Omega)} &\leq c_L \|(f_y^{\hat{y}} - f'(\hat{y}))d\|_{L^2(\Omega)} \leq \frac{1}{2} c_L c_{f'} \|y - \hat{y}\|_{L^4(\Omega)} \|d\|_{L^4(\Omega)} \\ &\leq \frac{1}{2} c_L^3 c_{f'} I_4^2 \|u - \hat{u}\|_{L^2(\Omega)}. \end{aligned}$$

Substituting  $y - \hat{y}$  by  $d + r_1$ , we obtain

$$\begin{aligned} \frac{1}{2} L_{yy}(\hat{y}, \hat{u}, \hat{p})[y - \hat{y}]^2 &= \frac{1}{2} L_{yy}(\hat{y}, \hat{u}, \hat{p})[d]^2 + L_{yy}(\hat{y}, \hat{u}, \hat{p})[d, r_1] + \frac{1}{2} L_{yy}(\hat{y}, \hat{u}, \hat{p})[r_1]^2 \\ &= \frac{1}{2} L_{yy}(\hat{y}, \hat{u}, \hat{p})[d]^2 + \frac{1}{2} \tilde{r}_2. \end{aligned}$$

The remainder term is given by

$$\tilde{r}_2 = (r_1, 2d + r_1) + \int_{\Omega} f''(\hat{y}) \hat{p} r_1 (2d + r_1) dx$$

and can be estimated by

$$(7.8) \quad \begin{aligned} |\tilde{r}_2| &\leq (1 + c_{f'} I_4^3 \|\hat{p}\|_{H^1(\Omega)}) \|r_1\|_{H^1(\Omega)} (\|d\|_{H^1(\Omega)} + \|r_1\|_{H^1(\Omega)}) \\ &\leq \underbrace{\frac{1}{2} (1 + c_{f'} I_4^3 \|\hat{p}\|_{H^1(\Omega)}) (2c_L \|u - \hat{u}\|_{L^2(\Omega)} + \|r_1\|_{H^1(\Omega)}) c_L^3 c_{f'} I_4^2}_{r_0} \|u - \hat{u}\|_{L^2(\Omega)}^2 \end{aligned}$$

with  $r_0 \rightarrow 0$  as  $\|u - \hat{u}\|_{L^2(\Omega)} \rightarrow 0$ .

So far, we achieved the following estimate for the difference of the objective values:

$$J(y, u) - J(\hat{y}, \hat{u}) \geq \frac{1}{2} L''(\hat{y}, \hat{u}, \hat{p}) [d, u]^2 - |r_2| - r_0 \|u - \hat{u}\|_{L^2(\Omega)}^2.$$

In the next step, we apply the coercivity (7.3) given by Theorem 7.4. Furthermore, we utilize the estimates of  $r_0$ ,  $r_1$ , and  $r_2$  in (7.6), (7.7), and (7.8), respectively. To shorten the estimates, let us assume  $\|u - \hat{u}\|_{L^2(\Omega)} \leq R$ . We obtain

$$(7.9) \quad \begin{aligned} J(y, u) - J(\hat{y}, \hat{u}) &\geq \left\{ \frac{\delta'}{2} - \frac{1}{2} c_L^3 c_{f'} I_4^2 (1 + c_{f'} I_4^3 \|\hat{p}\|_{H^1(\Omega)}) \left( 2c_L R + \frac{1}{2} c_L^3 c_{f'} I_4^2 R^2 \right) \right. \\ &\quad \left. - \frac{1}{6} c_{f''} I_4^4 R^3 \|\hat{p}\|_{H^1(\Omega)} \right\} \|u - \hat{u}\|_{L^2(\Omega)}^2. \end{aligned}$$

It remains to estimate  $\|\hat{p}\|_{H^1(\Omega)}$ . Using the splitting

$$\|\hat{p}\|_{H^1(\Omega)} \leq \|\hat{p} - \bar{p}_h\|_{H^1(\Omega)} + \|\bar{p}_h\|_{H^1(\Omega)},$$

Corollary 7.3 provides us with a computable estimate of that norm.

For  $R < R_0$  small enough, the factor in braces in (7.9) is greater than zero. This implies quadratic growth of the objective functional in the neighborhood  $\{u \in U_{ad} : \|u - \hat{u}\|_{L^2(\Omega)} \leq R_0\}$ .  $\square$

**8. Example.** In this section, we will apply our results to consider the optimal control problem of minimizing  $J(y, u)$  given by

$$(8.1) \quad J(y, u) := \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\nu}{2} \|u\|_{L^2(\Omega)}^2$$

subject to the semilinear boundary value problem

$$(8.2) \quad \begin{aligned} -(\Delta y)(x) + y^3(x) &= u(x) && \text{in } \Omega \\ y &= 0 && \text{on } \Gamma \end{aligned}$$

and to the control constraints

$$(8.3) \quad -0.1 \leq u(x) \leq 0.5 \text{ a.e. in } \Omega.$$

The domain  $\Omega$  is the unit square  $\Omega = (0, 1)^2$ , its boundary denoted by  $\Gamma$ . The parameter  $\nu$  was chosen as  $\nu = 0.1$ . The desired state  $y_d$  is given by

$$(8.4) \quad y_d = 8 \sin \pi x_1 \sin \pi x_2 - 4.$$

We find that the set of admissible controls is bounded in  $L^2(\Omega)$  by

$$\|u\|_{L^2(\Omega)} \leq 0.5 =: M_U \quad \forall u \in U_{ad}.$$

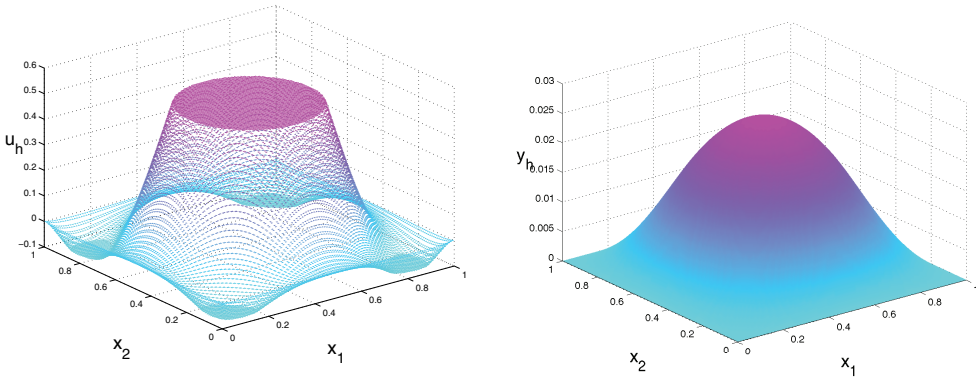


FIG. 8.1. *Discrete solution: control  $\bar{u}_h$  and state  $\bar{y}_h$ .*

**8.1. The solution.** At first, let us show the computed solutions for a fixed discretization. The state and adjoint state were discretized using piecewise linear and continuous functions on a regular triangulation of the domain. The control was discretized according to **(A3)** by piecewise constant functions. See also section 8.3.4 below for the discretization details. The grid consists of 20,000 triangles with 10,201 nodes. The discretization parameter was  $h = \sqrt{2}/100 = 0.01414\dots$ . Here and in what follows “ $\dots$ ” means truncation of floating point numbers after four leading digits. For the approximate evaluation of integrals we used a six-point quadrature of fourth order; see [15, Table 4.1]. A plot of the computed control  $\bar{u}_h$  and state  $\bar{y}_h$  can be found in Figure 8.1.

Now let us report the norms of the solution, which are needed for the following computations:

$$\begin{aligned} \|y_h\|_{L^\infty(\Omega)} &= 0.02752\dots, \quad \|y_h\|_{H^1(\Omega)} = 0.05638\dots, \quad \|y_h - y_d\|_{L^2(\Omega)} = 2.457\dots, \\ \|u_h\|_{L^2(\Omega)} &= 0.2856\dots, \\ \|p_h\|_{L^\infty(\Omega)} &= 0.1092\dots, \quad \|p_h\|_{H^1(\Omega)} = 0.2434\dots, \quad \|p_h\|_{W^{1,\infty}(\Omega)} = 0.3541\dots \end{aligned}$$

The set of elements  $T_h^i$ , where the control constraint is inactive (see Lemma 6.4), has measure  $|T_h^i| = 0.7220\dots$ .

**8.2. Check of the condition.** Now, let us report whether the condition (3.4) on the discrete solution is satisfied. For the computed solution it holds that

$$f''(\bar{y}_h(x))\bar{p}_h(x) \geq -0.01803\dots$$

Hence, the condition (3.7) and

$$\|y\|_{L^2(\Omega)}^2 + \nu\|u\|_{L^2(\Omega)}^2 + (f''(\bar{y}_h)y^2, \bar{p}_h) \geq \delta\|u\|_{L^2(\Omega)}^2$$

are satisfied with  $\delta = \nu = 0.1$ .

**8.3. Computation of the constants.** In the following short sections, we will explain how all those constants involved in the proofs are computed.

**8.3.1. Imbedding constants.** At first, we will compute the imbedding constants  $I_p$  of the imbeddings  $H_0^1(\Omega) \hookrightarrow L^p(\Omega)$ . We obtain, with the use of the eigenfunction of  $-\Delta$ , the constant of the imbedding in  $L^2(\Omega)$  as

$$I_2 = \frac{1}{\pi + 1} = 0.2415\dots$$

For the imbedding constants  $I_4$  and  $I_6$  we have the following. Because of the inequalities  $\|y\|_{L^4(\Omega)}^4 \leq \frac{1}{2}\|y\|_{L^2(\Omega)}^2\|\nabla y\|_{L^2(\Omega)}^2$  and  $\|y\|_{L^6(\Omega)}^6 \leq \frac{9}{8}\|y\|_{L^4(\Omega)}^4\|\nabla y\|_{L^2(\Omega)}^2$  (see [14]), the imbedding constants  $I_4$  and  $I_6$  can be computed as

$$I_4 = \left(\frac{1}{2}I_2^2\right)^{1/4} = 2^{-1/4}I_2^{1/2} = 0.4132\dots, \quad I_6 = \left(\frac{9}{8}I_4^4\right)^{1/6} = 0.5658\dots$$

Now, it remains to compute  $I_3$  for the imbedding in  $L^3(\Omega)$ . Using the interpolation inequality  $\|y\|_{L^3(\Omega)} \leq \|y\|_{L^2(\Omega)}^{1/3}\|y\|_{L^4(\Omega)}^{2/3}$ , we find

$$I_3 = I_2^{1/3}I_4^{2/3} = 0.3454\dots$$

The interpolation between  $L^2(\Omega)$  and  $L^6(\Omega)$  would give a larger imbedding constant of  $\tilde{I}_3 = I_2^{1/2}I_6^{1/2} = 0.3696\dots$  in our case.

**8.3.2. Solution mapping  $u \mapsto y$ .** At first, we investigate the bilinear form  $a$ . We have  $|a(y_1, y_2)| \leq \|y_1\|_{H^1(\Omega)}\|y_2\|_{H^1(\Omega)}$ , which gives  $\delta_1 = 1$ . Furthermore it holds that

$$a(y, y) = \|\nabla y\|_{L^2(\Omega)}^2 = \|y\|_{H^1(\Omega)}^2 - \|y\|_{L^2(\Omega)}^2 \geq (1 - I_2^2)\|y\|_{H^1(\Omega)}^2,$$

and we obtain  $\delta_0 = 1 - I_2^2 = 0.9417\dots$

The Lipschitz constant of the solution mapping now is given by Lemma 4.2 as

$$c_L = \frac{I_2}{\delta_0} = 0.2564\dots$$

In the following, we will apply the identity  $|y|_{H^2(\Omega)} = \|\Delta y\|_{L^2(\Omega)}$ , which can easily be proved by Fourier expansion. Now, let us estimate the constant  $c_S$  of Theorem 4.1. Since  $f(0) = 0$  holds, we find for the nonlinear equation,

$$\|y\|_{H^1(\Omega)} \leq c_L\|u\|_{L^2(\Omega)}.$$

Hence, we can derive

$$\begin{aligned} \|\Delta y\|_{L^2(\Omega)} &\leq \|f(y)\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)} \\ &\leq \|y\|_{L^6(\Omega)}^3 + \|u\|_{L^2(\Omega)} \leq (I_6^3 c_L^3 \|u\|_{L^2(\Omega)}^2 + 1)\|u\|_{L^2(\Omega)} \\ &\leq (I_6^3 c_L^3 M_U^2 + 1)\|u\|_{L^2(\Omega)}, \end{aligned}$$

which gives finally

$$\|y\|_{H^2(\Omega)} \leq \sqrt{c_L^2 + (I_6^3 c_L^3 M_U^2 + 1)^2}\|u\|_{L^2(\Omega)}$$

and the value of the constant  $c_S = \sqrt{c_L^2 + (I_6^3 c_L^3 M_U^2 + 1)^2} = 1.033\dots$

**8.3.3. Global estimates of the nonlinearity.** Furthermore, we need an  $L^\infty$ -bound of the solutions to the state equation. Here, we use Stampacchia's result; see [10, 17].

**COROLLARY 8.1.** *Let  $y$  be the solution of the nonlinear state equation (2.1) for a given right-hand side  $u \in L^2(\Omega)$ . Then it holds that  $y \in L^\infty(\Omega)$  with*

$$\|y\|_{L^\infty(\Omega)} \leq 4 \frac{I_6^2}{\sqrt{\delta_0}} |\Omega|^{1/6} \|u\|_{L^2(\Omega)}.$$

*Proof.* Let  $k$  be a real number. Define

$$v(x) = \begin{cases} y(x) - k & \text{if } y(x) \geq k, \\ 0 & \text{if } |y(x)| < k, \\ y(x) + k & \text{if } y(x) \leq -k, \end{cases} \quad \Omega(k) = \{x : |y(x)| \geq k\}.$$

Now, we test (2.1) by  $v$ ,

$$a(y, v) + (f(y), v) = (u, v).$$

Because of  $f(0) = 0$  and the monotonicity of  $f$ , it holds that  $(f(y), v) \geq 0$ . The right-hand side we estimate by

$$\begin{aligned} |(u, v)| &\leq \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq \|u\|_{L^2(\Omega)} \|v\|_{L^6(\Omega)} |\Omega(k)|^{1/3} \\ (8.5) \quad &\leq \|u\|_{L^2(\Omega)} |\Omega(k)|^{1/3} I_6 \|v\|_{H^1(\Omega)} \leq \frac{\delta_0}{2} \|v\|_{H^1(\Omega)}^2 + \frac{|\Omega(k)|^{2/3} I_6^2}{2\delta_0} \|u\|_{L^2(\Omega)}^2. \end{aligned}$$

Further, it holds that

$$|a(y, v)| \geq a(v, v) \geq \delta_0 \|v\|_{H^1(\Omega)}^2.$$

Now, we proceed with

$$(8.6) \quad \|v\|_{H^1(\Omega)}^2 \geq \frac{1}{I_6^2} \|v\|_{L^6(\Omega)}^2 = \frac{1}{I_6^2} \left( \int_{\Omega(k)} (|y| - k)^6 dx \right)^{1/3}.$$

Since  $\Omega(h) \subset \Omega(k)$  for  $h > k$ , we find

$$\begin{aligned} \left( \int_{\Omega(k)} (|y| - k)^6 dx \right)^{1/3} &\geq \left( \int_{\Omega(h)} (|y| - k)^6 dx \right)^{1/3} \geq \left( \int_{\Omega(h)} (h - k)^6 dx \right)^{1/3} \\ &= (h - k)^2 |\Omega(h)|^{1/3}. \end{aligned}$$

Altogether, we obtain

$$\frac{\delta_0}{I_6^2} (h - k)^2 |\Omega(h)|^{1/3} \leq \frac{I_6^2}{\delta_0} |\Omega(k)|^{2/3} \|u\|_{L^2(\Omega)}^2,$$

or, equivalently,

$$|\Omega(h)| \leq \left( \frac{I_6^4 \|u\|_{L^2(\Omega)}^2}{\delta_0^2} \right)^3 \frac{1}{(h - k)^6} |\Omega(k)|^2.$$

Then by a result of Stampacchia it holds that  $|\Omega(d)| = 0$  for

$$d = 4 \frac{I_6^2 \|u\|_{L^2(\Omega)}}{\delta_0} |\Omega|^{1/6}.$$

This implies

$$\|y\|_{L^\infty(\Omega)} \leq d = 4 \frac{I_6^2}{\delta_0} |\Omega|^{1/6} \|u\|_{L^2(\Omega)},$$

and the claim is proved. Due to the assumptions on  $f$  this estimate is valid for every suitable choice of the semilinearity.  $\square$

In our case, the constant  $c_{SL^\infty} = 4 \frac{I_6^2}{\delta_0} |\Omega|^{1/6}$  (see Theorem 4.1) has the value  $c_{SL^\infty} = 1.3596 \dots$ . Choices of  $L^p$ -spaces other than  $L^6$  in the estimates (8.5) and (8.6) yield other values of  $c_{SL^\infty}$ . Using the imbedding constants computed in this section,  $p = 6$  gave the smallest constant over all possible combinations from  $\{3, 4, 6\}$ .

Now, we can compute the global bound of the  $L^\infty$ -norms of all possible states by

$$M = c_{SL^\infty} M_U = 0.6798 \dots$$

Then the Lipschitz estimates of  $f$  holds with the following constants:

$$\begin{aligned} c_f &= 3M^2 = 1.386 \dots, \\ c_{f'} &= 6M = 4.079 \dots, \\ c_{f''} &= 6. \end{aligned}$$

Now, let us check whether the condition (3.6) holds uniformly for all admissible controls  $u$  with associated states  $y$  and adjoints  $p$ . To do so, we have to compute bounds of the  $L^\infty$ -norm of  $f''(y)p$ . Straightforward estimation gives

$$\begin{aligned} \|f''(y)p\|_{L^\infty(\Omega)} &\leq c_{f'} \|p\|_{L^\infty(\Omega)} \leq c_{f'} c_{SL^\infty} \|y - y_d\|_{L^2(\Omega)} \\ &\leq c_{f'} c_{SL^\infty} (I_2 c_L M_U + \|y_d\|_{L^2(\Omega)}) \leq 13.82 \dots \end{aligned}$$

Since this bound is larger than one, we cannot prove a priori that condition (3.7) holds for all admissible controls.

**8.3.4. Finite elements constants.** The discrete space  $V_h$  is chosen to be the set of continuous functions that are linear on the triangles of the triangulation  $T$ , e.g.,

$$V_h = \{v \in C(\bar{\Omega}) : v|_{T_h} \in P_1(T_h) \quad \forall T_h \in T, v|_\Gamma = 0\}.$$

The unit square was triangulated uniformly by orthogonal and congruent triangles.

It is known that the interpolation operator fulfills the requirements (5.1) and (5.2) of **(A3)**. However, we could not find any estimate of the associated constant  $c_1$  in the literature. So we decided to define the operator  $i_h$  as follows.

For  $y \in H_2(\Omega) \cap H_0^1(\Omega)$ , we set  $y_h := i_h y$  to be the solution of

$$a(y_h, v) = (-\Delta y, v) \quad \forall v \in V_h.$$

Then, the following upper bounds for the values of  $c_1$  and  $c_2$  can be found in [1, Theorem 5]:

$$c_1 = 0.2381 \dots, \quad c_2 = 0.4888 \dots$$

Before we can derive the constants of Lemma 5.1, we have to compute the constant  $c_M$  that appears in the proof of that lemma. The norm of the auxiliary function  $g$  defined there can be estimated by  $\|g\|_{H^1(\Omega)} \leq c_L$ . The  $L^2$ -norm of its Laplacian is

$$\|\Delta g\|_{L^2(\Omega)} \leq \|e\|_{L^2(\Omega)} + \|f_y^{y_h} g\|_{L^2(\Omega)} \leq 1 + \|g\|_{L^2(\Omega)} c_f \leq 1 + c_f c_L.$$

Hence, we obtain  $c_M = \sqrt{c_L^2 + (1 + c_f c_L)^2} = 1.504 \dots$

Now, we can compute the desired constants as

$$\begin{aligned} c_{L^2} &= (\delta_1 c_2 + c_f c_1) c_{H^1} c_M = 0.9806 \dots, \\ c_{H^1} &= (\delta_1 c_2 + c_f c_1) / \delta_0 = 0.8688 \dots \end{aligned}$$

**8.3.5. Auxiliary constants.** Now, let us investigate the auxiliary adjoint state  $p^h$  defined in (6.8). Here we want to compute the constant  $c_p$  as used in (6.10). By Corollary 4.3, we find

$$\|p^h\|_{H^1(\Omega)} \leq c_L \|\bar{y}_h - y_d\|_{L^2(\Omega)},$$

and with

$$\begin{aligned} \|\Delta p^h\|_{L^2(\Omega)} &= \|f'(\bar{y}_h)p^h\|_{L^2(\Omega)} + \|\bar{y}_h - y_d\|_{L^2(\Omega)} \\ &\leq 3I_2 \|\bar{y}_h\|_{L^\infty(\Omega)}^2 \|p^h\|_{H^1(\Omega)} + \|\bar{y}_h - y_d\|_{L^2(\Omega)} \\ &\leq (3I_2 c_L \|\bar{y}_h\|_{L^\infty(\Omega)}^2 + 1) \|\bar{y}_h - y_d\|_{L^2(\Omega)} \end{aligned}$$

we find  $c_p = \sqrt{c_L^2 + (3I_2 c_L \|\bar{y}_h\|_{L^\infty(\Omega)}^2 + 1)^2} = 1.032\dots$

**8.4. Verification results.** Finally, we report the estimates of the objective functional. The coefficients in the polynomial (6.22) were computed to

$$a_1 = 2.503\dots, a_2 = 0.3009\dots, a_3 = 5.843\dots, a_4 = 0.00337\dots, a_5 = 0.00786\dots,$$

which leads to the estimate

$$J(y, u) - J(y^h, \bar{u}_h) \geq -0.00337r^3 + 0.09989r^2 - 0.00542r - 0.00050.$$

This polynomial admits positive values for  $r \in [0.1033, 29.60]$ . That implies the existence of a local minimizer  $\hat{u}$  of  $J$  within the set

$$B_h = \{u \in U_{ad} : \|u - \bar{u}_h\|_{L^2(\Omega)} < 0.1033\}.$$

Furthermore, it follows that the global minimizer of  $J$  belongs to that neighborhood, since the upper bound 29.60 of the positivity interval is much larger than the diameter of the set of admissible controls.

Regarding the check of the sufficient condition—coercivity of the second-derivative of the Lagrangian—we achieved the following. For every  $r < 5.978$ , Theorem 7.4 gives coercivity of  $L''_{(y,u)}$  with a positive  $\delta'$ . Finally, Theorem 7.5 yields quadratic growth of the cost functional in an  $L^2$ -neighborhood of  $\hat{u}$  with radius  $r = 0.8543$ . Hence, it follows that the local minimizer  $\hat{u}$  of  $J$ , whose existence is proved above, is unique in the specified neighborhood  $B_h$ . Moreover, we know already that the global minimizer belongs to that neighborhood. Consequently, the function  $\hat{u}$  is the *unique globally optimal control* of our original problem (P). That is, the optimal control problem (8.1)–(8.4) is uniquely solvable.

Let us emphasize, that we proved a posteriori that the sufficient optimality condition holds at the still unknown control  $\hat{u}$ . Also we computed an approximation  $\bar{u}_h$  of that control with a computable error bound  $\|\hat{u} - \bar{u}_h\|_{L^2(\Omega)} < 0.1033$ .

**8.5. Dependence on  $h$ .** We computed the solution of the discrete problem for a sequence of discretizations. Then we performed the calculation of the different radii as above.

In Table 8.1,  $r_+$  denotes the smallest positive number such that the polynomial in the estimate in the estimation of the cost functional admits a positive value. That is,  $r_+$  is an upper bound for the approximation error in the controls. Our computable error bound decreases linearly with  $h$ , which is the expected convergence order for piecewise constant trial functions for the control [2].



TABLE 8.1

h	$r_+$
0.01414	0.1033
0.00707	0.04947
0.00354	0.02421
0.00177	0.01198

## REFERENCES

- [1] T. APEL AND M. DOBROWOLSKI, *Anisotropic interpolation with applications to the finite element method*, Computing, 47 (1992), pp. 277–293.
- [2] N. ARADA, E. CASAS, AND F. TRÖLTZSCH, *Error estimates for a semilinear elliptic control problem*, Comput. Optim. Appl., 23 (2002), pp. 201–229.
- [3] N. ARADA, J.-P. RAYMOND, AND F. TRÖLTZSCH, *On an augmented Lagrangian SQP method for a class of optimal control problems in Banach spaces*, Comput. Optim. Appl., 22 (2002), pp. 369–398.
- [4] J. F. BONNANS, *Second-order analysis for control constrained optimal control problems of semilinear elliptic equations*, Appl. Math. Optim., 38 (1998), pp. 303–325.
- [5] E. CASAS, M. MATEOS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl., 31 (2005), pp. 193–219.
- [6] A. L. DONTCHEV, W. W. HAGER, A. B. POORE, AND B. YANG, *Optimality, stability, and convergence in optimal control*, Appl. Math. Optim., 31 (1995), pp. 297–326.
- [7] H. GOLDBERG AND F. TRÖLTZSCH, *Second order optimality conditions for a class of control problems governed by nonlinear integral equations with application to parabolic boundary control*, Optimization, 20 (1989), pp. 687–698.
- [8] H. GOLDBERG AND F. TRÖLTZSCH, *Second-order sufficient optimality conditions for a class of nonlinear parabolic boundary control problems*, SIAM J. Control Optim., 31 (1993), pp. 1007–1025.
- [9] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.
- [10] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980; reprinted as Classics Appl. Math. 31, SIAM, Philadelphia, 2000.
- [11] K. MALANOWSKI AND F. TRÖLTZSCH, *Lipschitz stability of solutions to parametric optimal control problems for parabolic equations*, Z. Anal. Anwendungen, 18 (1999), pp. 469–489.
- [12] K. MALANOWSKI AND F. TRÖLTZSCH, *Lipschitz stability of solutions to parametric optimal control for elliptic equations*, Control Cybernet., 29 (2000), pp. 237–256.
- [13] H. D. MITTELMANN AND F. TRÖLTZSCH, *Sufficient optimality in a parabolic control problem*, in Trends in Industrial and Applied Mathematics, A. H. Siddiqi and M. Kocvara, eds., Kluwer, Dordrecht, 2002, pp. 305–316.
- [14] M. PLUM AND C. WIENERS, *New solutions of the Gelfand problem*, J. Math. Anal. Appl., 269 (2002), pp. 588–606.
- [15] G. STRANG AND G. J. FIX, *An Analysis of the Finite Element Method*, Prentice–Hall Series in Automatic Computation, Prentice–Hall, Englewood Cliffs, NJ, 1973.
- [16] F. TRÖLTZSCH, *Lipschitz stability of solutions of linear-quadratic parabolic control problems with respect to perturbations*, Dynam. Contin. Discrete Impuls. Systems, 7 (2000), pp. 289–306.
- [17] F. TRÖLTZSCH, *Optimale Steuerung partieller Differentialgleichungen*, Vieweg, Wiesbaden, 2005.

## FAST SWITCHING ANALYSIS OF LINEAR SWITCHED SYSTEMS USING EXPONENTIAL SPLITTING\*

M. PORFIRI<sup>†</sup>, D. G. ROBERSON<sup>‡</sup>, AND D. J. STILWELL<sup>‡</sup>

**Abstract.** Stability of periodic switched linear systems with fast switching patterns is studied by combining the method of averaging with exponential splitting. This approach yields less conservative bounds on stabilizing fast switching rates than can be obtained with prior approaches. Such bounds can be useful for analysis and design of switching control laws. The method is also generalized to arbitrary (including nonperiodic nonswitched) time-varying systems. In particular, the effects of time-varying state perturbations on the stability of linear periodic switched systems are analyzed.

**Key words.** averaging method, exponential splitting, fast switching, stability, switched systems, linear systems

**AMS subject classifications.** 34C29, 93C05, 93D20

**DOI.** 10.1137/060665750

**1. Introduction.** A switched system is a dynamical system which consists of a family of time-invariant subsystems and a policy that oversees the switching among them. For example,

$$\dot{x} = A_{\rho(t)}x,$$

where  $\rho(t)$  is a piecewise constant switching signal that selects from a family of matrix-valued coefficients  $\Theta = \{A_1, A_2, \dots, A_M\}$ , represents such a system. This class of dynamical systems finds extensive application in various areas of engineering practice such as power electronics, network communication, hybrid control, traffic flow, biosystem modeling, etc. (see, e.g., [16] and references therein). Switched systems have been studied for several decades in the systems and control literature (see, e.g., [8]).

Switching among the possible system configurations may be orchestrated according to different protocols, as discussed in [16] and [19]. Switching events may occur as a function of the system state, or at given instants in time independent of the system state. In the latter case, when one or more of the subsystems (the elements of  $\Theta$ ) are stable, dwell time approaches, where the switching signal selects stable subsystems for relatively long time intervals, are effective in ensuring system stability. However, when no subsystems are stable, or when system constraints prevent long intervals between switching instants, alternate strategies must be considered. Periodic switching is another well-studied strategy (see, e.g., [20]) that can be useful when such restrictions exist.

Periodic switching is a state-independent switching scheme that utilizes a periodic rule for selecting subsystems. That is, there is a  $T > 0$  such that  $\rho(t) = \rho(t + T)$  for

---

\*Received by the editors July 21, 2006; accepted for publication (in revised form) May 22, 2008; published electronically September 19, 2008. This work was supported by the National Science Foundation via grants IIS-0238092 and CMMI-0745753, and by the Office of Naval Research via grants N000140310444, N000140510780, and N000140510516.

<http://www.siam.org/journals/sicon/47-5/66575.html>

<sup>†</sup>Mechanical and Aerospace Engineering Department, Polytechnic University, Brooklyn, NY 11201 (mporfiri@poly.edu).

<sup>‡</sup>The Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 (grayr@vt.edu, stilwell@vt.edu).

all  $t \geq 0$ . The stability of the resulting periodic linear system may be analyzed using a number of different techniques. The spectral properties and contractibility of the state transition matrix over a switching period may be used to characterize system stability. Such an approach involves computation and multiplication of matrix exponentials, which may present computational challenges (see, e.g., [10]), especially as the system dimension and the number of subsystems grow. Standard Floquet arguments may also be used to study the periodic system, as in [6]. The main drawback of this approach is its inherent dependence on matrix logarithms, which limits qualitative system analysis and control design. In [4] sufficient conditions for uniform asymptotic stability are stated in terms of convex combinations of the matrix measures of the elements of  $\Theta$ . In [13] the conditions of [4] are generalized by including matrix condition numbers in the sufficient conditions. In [17] it is shown that the stability of a switched system may be assessed by studying an auxiliary time-invariant system whose state coefficient matrix is a certain time average of the original system. In particular, it is shown that the stability of the average system is inherited by the switched system whenever the switching signal is sufficiently fast, or equivalently when the switching period  $T$  is sufficiently small (fast switching). In [20] the average system is used to design a stabilizing feedback law for periodic switched systems. More specifically, it is shown that if the average system is stabilizable and detectable, then for any feedback matrix that stabilizes the average system there exists a minimum switching rate that guarantees stability of the periodic system.

Computing the slowest allowable switching rate (or the maximum switching period) is an important factor to consider when analyzing stability and designing feedback control laws. In [17] the estimate of the maximum switching period is based on a perturbation analysis of the system spectral properties. The estimate may also be established by specializing to switched linear systems the method of averaging proposed in [7] for general linear time-varying systems (see, e.g., [11]). The approach of [7] relies on a decomposition of the system transition matrix into a part due to the time average of the system dynamics and a perturbation that is on the order of  $T^2$ , where  $T$  is the switching period. For the time-varying system  $\dot{x} = \mathbf{A}(t)x$ , the state transition matrix satisfies

$$\Phi(s + \tau, s) = e^{\bar{\mathbf{A}}_\tau(s)\tau} + \mathbf{E}_\tau(s),$$

where

$$\bar{\mathbf{A}}_\tau(s) = \frac{1}{\tau} \int_s^{s+\tau} \mathbf{A}(\sigma) d\sigma$$

and

$$\|\mathbf{E}_\tau(s)\| \leq \alpha^2 \tau^2 e^{\alpha\tau}$$

if  $\|\mathbf{A}(t)\| \leq \alpha$  for all  $t$ . This decomposition is well known and appears in several textbooks (see, e.g., [1] and [12, Exercise 4.25]). Henceforth, this approach to estimating the maximum switching period is referred to as the standard approach, and the estimate itself as the standard estimate. The estimate of [17] is difficult to compute numerically, as it involves quantifying the sensitivity of matrices to perturbations. The standard estimate is easily applicable and requires the knowledge of elementary quantities. Nevertheless, it generally leads to conservative results that may be even a few orders of magnitude less than the exact value computed using Floquet theory.

The present work is focused on characterizing the slowest stabilizing switching rate through a computationally tractable expression. Loosely speaking, improved bounds are due to directly expressing the effects of the commutation relationship of the elements of  $\Theta$ . The general idea stems from combining the standard approach with well-developed results from numerical algebra (see, e.g., [14, 10]). By exploiting the concept of matrix exponential splitting, which is widely used in numerical analysis of partial differential equations (see, e.g., [9]), a novel switching rate estimate is obtained. The estimate bounds the slowest switching rate that guarantees uniform asymptotic stability of switched systems characterized by asymptotically stable averages. The new estimate is compared to the standard estimate and its improvement is shown. These results are also extended to a more general linear class of dynamical systems, and new results on the method of averaging are established. In particular, the effects of time-varying state perturbations on the stability of linear periodic switched systems are analyzed.

The paper is organized as follows. Section 2 contains the general system description and a discussion of fundamental matrix measure and exponential splitting concepts. Section 3 addresses our principle contribution, an improved estimate for the maximum switching period for stability under fast switching. Section 4 addresses the stability analysis of more general dynamical systems and presents stability conditions for perturbed switched systems. Section 5 contains concluding remarks.

**2. System description.** We consider the homogeneous linear time-varying system

$$(2.1) \quad \dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x},$$

where  $t \in \mathbb{R}^+$  indicates the time variable,  $\mathbf{x} \in \mathbb{R}^n$  is the system state, and  $\mathbf{A}(\cdot) : \mathbb{R}^+ \rightarrow \mathbb{R}^{n \times n}$  is a bounded and right continuous matrix function. By a right continuous function in  $\mathbb{R}^+$  we mean a function whose value at each point  $t$  in  $\mathbb{R}^+$  equals its right-hand limit at  $t$ . We further assume that the number of discontinuities of  $\mathbf{A}$  is finite over any finite interval  $[s, s + \tau]$ , where  $s, \tau \in \mathbb{R}^+$ . Thus we do not consider systems which exhibit Zeno behavior or chattering, as described in [8] and [16].

The homogeneous systems that we consider include switched systems for which  $\mathbf{A}(t) = \mathbf{A}_{\rho(t)}$  with  $\rho(t)$  being a switching signal that selects from among a family of constant matrix functions. In section 3 this class of switched systems is considered, with the additional conditions that  $\rho(t)$  is periodic and selects from among a finite set of matrix functions  $\Theta = \{A_1, A_2, \dots, A_M\}$ . The switched systems considered in section 4 are not subject to the additional conditions, so that  $\rho(t)$  may be nonperiodic and  $\Theta$  may be countably infinite. In section 4 we also consider more general linear time-varying systems, where  $\mathbf{A}(t)$  may be nonconstant between points of discontinuity. For existence and uniqueness of these types of problems one may refer to [3] and [18].

**2.1. Matrix measure.** Let  $\|\cdot\|$  denote a norm in  $\mathbb{R}^n$  and the corresponding induced norm in  $\mathbb{R}^{n \times n}$ . The one-sided directional derivative of  $\|\cdot\|$  at the identity matrix  $\mathbf{I} \in \mathbb{R}^{n \times n}$  in the direction  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is called the matrix measure of  $\mathbf{A}$  and is denoted by  $\mu(\mathbf{A})$ . That is,

$$(2.2) \quad \mu(\mathbf{A}) = \lim_{h \rightarrow 0^+} \frac{\|\mathbf{I} + h\mathbf{A}\| - 1}{h}.$$

Basic properties of the matrix measure for  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$  and  $c \in \mathbb{R}$  are

$$\begin{aligned}
 (2.3a) \quad & \mu(\mathbf{I}) = 1, \quad \mu(-\mathbf{I}) = -1, \quad \mu(\mathbf{0}) = 0, \\
 (2.3b) \quad & \mu(c\mathbf{A}) = |c|\mu(\operatorname{sgn}(c)\mathbf{A}), \quad \mu(\mathbf{A} + c\mathbf{I}) = \mu(\mathbf{A}) + c, \\
 (2.3c) \quad & \mu(\mathbf{A} + \mathbf{B}) \leq \mu(\mathbf{A}) + \mu(\mathbf{B}), \\
 (2.3d) \quad & \|e^{\mathbf{A}}\| \leq e^{\mu(\mathbf{A})}, \\
 (2.3e) \quad & \max_{i=1, \dots, n} [\operatorname{Re}(\lambda_i(\mathbf{A}))] \leq \mu(\mathbf{A}) \leq \|\mathbf{A}\|,
 \end{aligned}$$

where  $\lambda_i(\mathbf{A})$  indicates the  $i$ th eigenvalue of  $\mathbf{A}$ . These properties hold for the matrix measure defined in any induced matrix norm. Derivation of the relationships (2.3) may be found in [15]. An additional property of the matrix measure (see, e.g., [7]) places upper and lower bounds on the transition matrix norm of the system (2.1) according to the relationship

$$(2.4) \quad e^{-\int_{\tau}^t \mu(-\mathbf{A}(\vartheta))d\vartheta} \leq \|\Phi(t, \tau)\| \leq e^{\int_{\tau}^t \mu(\mathbf{A}(\vartheta))d\vartheta}$$

for all  $t, \tau \in \mathbb{R}^+$  and  $t \geq \tau$ .

The computation of the matrix measure is generally very involved, but there are special cases where reduced effort is needed (see, e.g., [15, 2]). In what follows, we primarily make use of the vector  $\mathbf{P}$ -norm and the corresponding matrix induced norm and matrix measure, defined as

$$\begin{aligned}
 (2.5) \quad & \|\mathbf{x}\|_P = \sqrt{\mathbf{x}^T \mathbf{P} \mathbf{x}}, \quad \|\mathbf{A}\|_P = \max_{i=1, \dots, n} \sigma_i(\mathbf{P}^{1/2} \mathbf{A} \mathbf{P}^{-1/2}), \\
 & \mu_P(\mathbf{A}) = \frac{1}{2} \max_{i=1, \dots, n} \lambda_i(\mathbf{P}^{1/2} \mathbf{A} \mathbf{P}^{-1/2} + \mathbf{P}^{-1/2} \mathbf{A}^T \mathbf{P}^{1/2}),
 \end{aligned}$$

where  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $\sigma_i(\mathbf{A})$  indicates the  $i$ th singular value of  $\mathbf{A}$ , and  $\mathbf{P}$  is any symmetric positive definite matrix in  $\mathbb{R}^{n \times n}$ . In particular, when  $\mathbf{P} = \mathbf{I}$  the matrix measure corresponding to the Euclidean norm is obtained.

**2.2. Exponential splitting.** Exponential splitting methods are numerical tools for computing matrix exponentials (see, e.g., [9]). They involve decomposing a matrix into the sum of matrices, whose exponentials can be easily computed, and approximating the original matrix exponential in terms of the exponentials of the simpler matrices. That is, traditional splitting methods approximate the matrix exponential  $e^{(\mathbf{A}+\mathbf{B})h}$  by the product  $e^{\mathbf{A}h}e^{\mathbf{B}h}$ , where  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$  and  $h \in \mathbb{R}^+$ , and where  $e^{\mathbf{A}h}$  and  $e^{\mathbf{B}h}$  are relatively easy to compute.

We apply splitting methods in the opposite direction. That is, we approximate the product of exponentials  $e^{\mathbf{A}h}e^{\mathbf{B}h}$  by the single exponential  $e^{(\mathbf{A}+\mathbf{B})h}$ . The motivation for this approach is the recognition that the product represents the state transition matrix for a switched system, and the single exponential represents the transition matrix for a corresponding average system. We consider the problem of estimating the global error in the first order exponential splitting by defining

$$\mathbf{E}(h) = e^{\mathbf{A}h}e^{\mathbf{B}h} - e^{(\mathbf{A}+\mathbf{B})h}.$$

PROPOSITION 2.1. *Given  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$  and  $h \in \mathbb{R}^+$ ,*

$$\|\mathbf{E}(h)\| \leq \frac{1}{2} h^2 \|[\mathbf{A}, \mathbf{B}]\| e^{h(\mu(\mathbf{A}) + \mu(\mathbf{B}))},$$

where

$$[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$$

is the commutator of  $\mathbf{A}$  and  $\mathbf{B}$ .

*Proof.* Part of the following arguments are borrowed from [14]. The main difference in those results lies in the estimate itself, where we make use of additional properties of the matrix measure. By differentiating  $\mathbf{E}(h)$ , we obtain

$$\frac{d}{dh} \mathbf{E}(h) = (\mathbf{A} + \mathbf{B})\mathbf{E}(h) + [e^{\mathbf{A}h}, \mathbf{B}] e^{\mathbf{B}h}.$$

The above ordinary differential equation (ODE) with homogeneous initial conditions may be solved to obtain

$$(2.6) \quad \mathbf{E}(h) = \int_0^h e^{(h-\vartheta)(\mathbf{A}+\mathbf{B})} \mathbf{S}(\vartheta) e^{\vartheta \mathbf{B}} d\vartheta,$$

where

$$\mathbf{S}(\vartheta) = [e^{\mathbf{A}\vartheta}, \mathbf{B}].$$

It is clear that  $\mathbf{S}(0) = 0$ . By differentiating  $\mathbf{S}(\vartheta)$ , we obtain

$$\frac{d}{d\vartheta} \mathbf{S}(\vartheta) = \mathbf{AS}(\vartheta) + [\mathbf{A}, \mathbf{B}] e^{\mathbf{A}\vartheta}.$$

The above ODE may be solved, yielding

$$(2.7) \quad \mathbf{S}(\vartheta) = \int_0^\vartheta e^{(\vartheta-p)\mathbf{A}} [\mathbf{A}, \mathbf{B}] e^{p\mathbf{A}} dp.$$

Substituting (2.7) into (2.6), we obtain

$$\mathbf{E}(h) = \int_0^h e^{(h-\vartheta)(\mathbf{A}+\mathbf{B})} \left( \int_0^\vartheta e^{(\vartheta-p)\mathbf{A}} [\mathbf{A}, \mathbf{B}] e^{p\mathbf{A}} dp \right) e^{\vartheta \mathbf{B}} d\vartheta.$$

Applying standard matrix norm inequalities, then using property (2.3d) followed by (2.3b), we obtain

$$\begin{aligned} \|\mathbf{E}(h)\| &\leq \int_0^h \left\| e^{(h-\vartheta)(\mathbf{A}+\mathbf{B})} \right\| \int_0^\vartheta \left\| e^{(\vartheta-p)\mathbf{A}} \right\| \left\| [\mathbf{A}, \mathbf{B}] \right\| \left\| e^{p\mathbf{A}} \right\| dp \left\| e^{\vartheta \mathbf{B}} \right\| d\vartheta \\ &\leq \left\| [\mathbf{A}, \mathbf{B}] \right\| \int_0^h e^{\mu((h-\vartheta)\mathbf{A}) + \mu((h-\vartheta)\mathbf{B})} e^{\mu(\vartheta \mathbf{B})} \int_0^\vartheta e^{\mu((\vartheta-p)\mathbf{A})} e^{\mu(p\mathbf{A})} dp d\vartheta \\ &= \left\| [\mathbf{A}, \mathbf{B}] \right\| \int_0^h e^{(h-\vartheta)\mu(\mathbf{A})} e^{(h-\vartheta)\mu(\mathbf{B})} e^{\vartheta \mu(\mathbf{B})} \int_0^\vartheta e^{(\vartheta-p)\mu(\mathbf{A})} e^{p\mu(\mathbf{A})} dp d\vartheta \\ &= \left\| [\mathbf{A}, \mathbf{B}] \right\| \int_0^h e^{h\mu(\mathbf{A})} e^{h\mu(\mathbf{B})} \int_0^\vartheta dp d\vartheta \\ &= \frac{1}{2} h^2 \left\| [\mathbf{A}, \mathbf{B}] \right\| e^{h(\mu(\mathbf{A}) + \mu(\mathbf{B}))}. \quad \square \end{aligned}$$

The problem of estimating the global error in the first order exponential splitting has also been studied in [5], using a Lie algebraic framework. The resulting error bound in Proposition 2 of [5] involves the determination of the limit of a series of nested commutators, arising from the Baker–Campbell–Hausdorff formula, which can be potentially difficult to compute.

**3. Stability analysis of periodic switching systems.** The present section is devoted to the analysis of homogeneous time-varying systems of the form (2.1), where  $\mathbf{A}$  is a  $T$ -periodic piecewise constant bounded matrix function. At any point in time,  $\mathbf{A}(t)$  takes values from the set  $\Theta = \{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_M\}$  according to the policy

$$(3.1) \quad \mathbf{A}(t) = \mathbf{A}_m \text{ for } t \in \left[ kT + T \sum_{i=0}^{m-1} \delta_i, kT + T \sum_{i=0}^m \delta_i \right),$$

where  $\delta_i$  is the duty cycle of the subsystem corresponding to the matrix  $\mathbf{A}_i$  with the convention  $\delta_0 = 0$ , so that  $0 \leq \delta_i < 1$  and  $\sum_{i=1}^M \delta_i = 1$ , and where  $k \in \mathbb{N}$ . Our goal is to determine an estimate for the maximum switching period  $T$  for which (2.1) is uniformly asymptotically stable.

For the periodic switched linear system (3.1), the transition matrix  $\Phi$  over any period  $[kT, (k+1)T)$  is called a monodromy. Due to periodicity, the monodromy over the interval for any  $k$  equals the monodromy over the interval for  $k = 0$ , so that

$$(3.2) \quad \begin{aligned} \Phi((k+1)T, kT) &= \Phi(T, 0) = \Phi\left(T, T \sum_{i=1}^{M-1} \delta_i\right) \cdots \Phi(T(\delta_1 + \delta_2), T\delta_1) \Phi(T\delta_1, 0) \\ &= e^{T\mathbf{A}_M \delta_M} \cdots e^{T\mathbf{A}_2 \delta_2} e^{T\mathbf{A}_1 \delta_1}. \end{aligned}$$

In what follows we show a sufficient condition for uniform asymptotic stability based on the transition matrix contractibility.

PROPOSITION 3.1. *If for some norm*

$$(3.3) \quad \sum_{i=1}^M \delta_i \mu(\mathbf{A}_i) < 0,$$

*then the system is uniformly asymptotically stable for every positive  $T$ . Otherwise, if the average system matrix*

$$(3.4) \quad \bar{\mathbf{A}}_T = \sum_{i=1}^M \delta_i \mathbf{A}_i$$

*is Hurwitz, then there exists a  $T^* > 0$  such that when  $T < T^*$  the system is uniformly asymptotically stable.  $T^*$  satisfies*

$$(3.5) \quad e^{T^* \mu_P(\bar{\mathbf{A}}_T)} + \frac{1}{2} (T^*)^2 \Gamma_P(\mathbf{A}) \exp\left(T^* \sum_{i=1}^M \delta_i \mu_P(\mathbf{A}_i)\right) = 1,$$

*where*

$$(3.6) \quad \Gamma_P(\mathbf{A}) = \sum_{i=1}^{M-1} \sum_{j=i+1}^M \delta_i \delta_j \|\mathbf{A}_i, \mathbf{A}_j\|_P.$$

*The norm is constructed as in (2.5), with  $\mathbf{P}$  being the unique symmetric positive definite solution of*

$$(3.7) \quad \bar{\mathbf{A}}_T^T \mathbf{P} + \mathbf{P} \bar{\mathbf{A}}_T + \mathbf{R} = 0,$$

*where  $\mathbf{R} = \mathbf{R}^T > 0$  is arbitrary.*

The proof of Proposition 3.1 is given at the end of the section, after a preliminary proposition on the application of exponential splitting to the monodromy matrix. We note that although Proposition 3.1 provides sufficient stability conditions, the condition of Hurwitz  $\bar{\mathbf{A}}_T$  is actually necessary for stability for small  $T$ . Indeed, if  $\bar{\mathbf{A}}_T$  is not Hurwitz, there exists a  $T^* > 0$  such that the system is unstable for all  $T < T^*$ , as Theorem 2.1 in [7] indicates. Constructing the norm as in (3.7) guarantees that the measure of  $\bar{\mathbf{A}}_T$  is negative, given (2.5) and the fact that  $\bar{\mathbf{A}}_T$  is Hurwitz. If an arbitrary norm is chosen, the corresponding measure of  $\bar{\mathbf{A}}_T$  may be positive, even if  $\bar{\mathbf{A}}_T$  is Hurwitz. The commutators appearing in (3.6) reflect the similarity of the eigenvectors of the subsystem state coefficient matrices, a set of relationships not considered in the standard approach. If all of the matrices  $\mathbf{A}_1, \dots, \mathbf{A}_M$  commute and  $\bar{\mathbf{A}}_T$  is Hurwitz, the switching system is uniformly asymptotically stable. Indeed, the commutator in (3.5) vanishes for every  $i, j$ , and the monodromy matrix is a contraction for any  $T$ . If all of the matrices  $\mathbf{A}_1, \dots, \mathbf{A}_M$  are symmetric and negative definite, then  $\bar{\mathbf{A}}_T$  is Hurwitz and the switched system is uniformly asymptotically stable. Indeed, the matrix measure induced by the Euclidean norm satisfies (3.3).

**PROPOSITION 3.2.** *For the switched periodic linear system defined by (3.1), the difference  $\mathbf{E}_T$  between the monodromy matrix over the period  $[0, T)$  and the transition matrix of the average system over the same period, defined by*

$$(3.8) \quad \Phi(T, 0) = \mathbf{E}_T + e^{T\bar{\mathbf{A}}_T},$$

*satisfies the norm bound*

$$(3.9) \quad \|\mathbf{E}_T\| \leq \frac{1}{2} T^2 \Gamma(\mathbf{A}) \exp \left( T \sum_{i=1}^M \delta_i \mu(\mathbf{A}_i) \right),$$

*where*

$$(3.10) \quad \Gamma(\mathbf{A}) = \sum_{i=1}^{M-1} \sum_{j=i+1}^M \delta_i \delta_j \|[\mathbf{A}_i, \mathbf{A}_j]\|.$$

*Proof.* We rewrite the global error in (3.8) as

$$\begin{aligned} \mathbf{E}_T = & \sum_{i=0}^{M-2} e^{T\mathbf{A}_{M+1}\delta_{M+1}} e^{T\mathbf{A}_M\delta_M} \dots e^{T\mathbf{A}_{M+1-i}\delta_{M+1-i}} \\ & \times \left( e^{T\delta_{M-i}\mathbf{A}_{M-i}} e^{T\sum_{j=i+1}^{M-1} \delta_{M-j}\mathbf{A}_{M-j}} - e^{T\sum_{j=i}^{M-1} \delta_{M-j}\mathbf{A}_{M-j}} \right), \end{aligned}$$

where  $\mathbf{A}_{M+1} = \mathbf{I}$  and  $\delta_{M+1} = 0$ . Applying Proposition 2.1 for each bracketed term, the Schwarz inequality, the triangle inequality, and the measure properties (2.3b),



(2.3c), and (2.3d), we obtain

$$\begin{aligned}
 \|\mathbf{E}_T\| &\leq \sum_{i=0}^{M-2} \|e^{T\mathbf{A}_{M+1}\delta_{M+1}}\| \dots \|e^{T\mathbf{A}_{M+1-i}\delta_{M+1-i}}\| \\
 &\quad \times \left\| e^{T\mathbf{A}_{M-i}} e^{T\sum_{j=i+1}^{M-1} \delta_{M-j}\mathbf{A}_{M-j}} - e^{T\sum_{j=i}^{M-1} \delta_{M-j}\mathbf{A}_{M-j}} \right\| \\
 &\leq \frac{1}{2} T^2 \sum_{i=0}^{M-2} \left\| \left[ \delta_{M-i}\mathbf{A}_{M-i}, \sum_{j=i+1}^{M-1} \delta_{M-j}\mathbf{A}_{M-j} \right] \right\| \\
 &\quad \times e^{T(\sum_{j=0}^i \delta_{M-j}\mu(\mathbf{A}_{M-j}) + \mu(\sum_{j=i+1}^{M-1} \delta_{M-j}\mathbf{A}_{M-j}))} \\
 &\leq \frac{1}{2} T^2 \sum_{i=0}^{M-2} \sum_{j=i+1}^{M-1} \delta_{M-i}\delta_{M-j} \|\mathbf{A}_{M-i}, \mathbf{A}_{M-j}\| e^{T\sum_{l=0}^{M-1} \delta_{M-l}\mu(\mathbf{A}_{M-l})}. \quad \square
 \end{aligned}$$

*Proof of Proposition 3.1.* To demonstrate uniform asymptotic stability, we show that the monodromy matrix  $\Phi(T, 0)$  is a contraction. This approach relies on the boundedness of the transition matrix over any interval whose duration is less than one switching period, due to the fact that the set  $\Theta$  of state coefficient matrices contains a finite number of elements. If the monodromy is a contraction,  $\|\Phi(t, 0)\|$  can be made arbitrarily small by choosing  $t$  sufficiently large, establishing uniform asymptotic stability (see, e.g., [12, Exercise 6.15]).

To establish the first sufficient condition, we use (3.2) to obtain an expression for the monodromy matrix norm in the norm for which (3.3) holds. Applying the Schwarz inequality and matrix measure properties (2.3b) and (2.3d) yields

$$\|\Phi(T, 0)\| \leq e^{T\delta_M\mu(\mathbf{A}_M)} \dots e^{T\delta_2\mu(\mathbf{A}_2)} e^{T\delta_1\mu(\mathbf{A}_1)} = e^{T\sum_{i=1}^M \delta_i\mu(\mathbf{A}_i)} < 1.$$

This condition appeared in [4] for the Euclidean norm.

To establish the second sufficient condition, we use (3.8) to obtain a matrix measure norm expression. Using the triangle inequality together with (2.3d) yields

$$\|\Phi(T, 0)\| \leq \|\mathbf{E}_T\| + e^{T\mu(\bar{\mathbf{A}}_T)}.$$

Thus, considering Proposition 3.2, we look for switching periods satisfying

$$(3.11) \quad \frac{1}{2} T^2 \Gamma(\mathbf{A}) \exp\left(T \sum_{i=1}^M \delta_i \mu(\mathbf{A}_i)\right) + e^{T\mu(\bar{\mathbf{A}}_T)} < 1.$$

We specialize to the norm defined in (2.5), where  $\mathbf{P}$  is the solution of (3.7). Because the matrix  $\mathbf{P}$ -norm is induced from the vector  $\mathbf{P}$ -norm, properties (2.3) hold for the matrix measure derived from this norm. The measure of the average matrix is

$$\begin{aligned}
 \mu_P(\bar{\mathbf{A}}_T) &= \frac{1}{2} \max_{i=1, \dots, n} \lambda_i \left( \mathbf{P}^{1/2} \bar{\mathbf{A}}_T \mathbf{P}^{-1/2} + \mathbf{P}^{-1/2} \bar{\mathbf{A}}_T^T \mathbf{P}^{1/2} \right) \\
 &= \frac{1}{2} \max_{i=1, \dots, n} \lambda_i \left( \mathbf{P}^{-1/2} (\bar{\mathbf{A}}_T^T \mathbf{P} + \mathbf{P} \bar{\mathbf{A}}_T) \mathbf{P}^{-1/2} \right) \\
 &= \frac{1}{2} \max_{i=1, \dots, n} \lambda_i \left( -\mathbf{P}^{-1/2} \mathbf{R} \mathbf{P}^{-1/2} \right).
 \end{aligned}$$

Because both  $\mathbf{P}$  and  $\mathbf{R}$  are positive definite, it follows that  $\mu_P(\bar{\mathbf{A}}_T) < 0$ . For the chosen norm, then, the second term of the LHS of (3.11) is a decreasing function of

$T$  which attains the value 1 at  $T = 0$ . On the other hand, the first term is a function that increases with  $T$ , since (3.3) does not hold by hypothesis. However, the first term vanishes at  $T = 0$ , and, moreover, its derivative with respect to  $T$  is 0 at  $T = 0$ . Observing that the LHS of (3.11) is continuous in  $T$ , it can be characterized as a function that equals 1 at  $T = 0$ , decreases for small  $T$ , and increases with larger  $T$ . Hence there is a time  $T^*$  such that (3.5) holds. Because the LHS of (3.11) provides an upper bound on  $\|\Phi(T, 0)\|$ , it follows that when  $T < T^*$ , the monodromy matrix is a contraction with respect to  $\|\cdot\|_P$ .  $\square$

In [13], condition (3.3) for the Euclidean norm is generalized by applying a modified version of property (2.3d), where a similarity transformation is exploited and the norm of the matrix exponential is bounded by using the condition number of the transformation and the measure of the transformed matrix. In [13], it is shown that this condition may yield equivalent, less conservative, or more conservative results than (3.3), depending on the problem at hand and on the used similarity transformation.

Since they are defined in terms of a matrix norm, the matrix measure and commutator norm quantities appearing in (3.5) and (3.10) do not grow with matrix dimension. Therefore, the estimate of the slowest stabilizing switching period provided by (3.5) does not degrade with system dimension. It should be noted that because (3.10) accounts for all possible commutation relationships between subsystem matrices, it tends to make the estimate of (3.5) more conservative as the number of subsystems grows. Nevertheless, it is always less conservative than the standard estimate represented by the RHS of (3.12). The following proposition formalizes this result.

**PROPOSITION 3.3.** *For the switched system in (3.1), it is the case that*

$$(3.12) \quad \frac{1}{2} T^2 \Gamma(\mathbf{A}) \exp \left( T \sum_{i=1}^M \delta_i \mu(\mathbf{A}_i) \right) \leq T^2 \alpha^2 e^{T\alpha},$$

where  $\Gamma(\mathbf{A})$  is defined in (3.10) and

$$\alpha = \operatorname{esssup}_{t \in [0, T]} \|\mathbf{A}(t)\| = \max_{i=1, \dots, M} \|\mathbf{A}_i\|.$$

*Proof.* By noticing that for any  $i, j \in \{1, \dots, M\}$ ,

$$\|[\mathbf{A}_i, \mathbf{A}_j]\| \leq 2\|\mathbf{A}_i \mathbf{A}_j\| \leq 2\alpha^2,$$

and that

$$\sum_{i=1}^{M-1} \sum_{j=i+1}^M \delta_i \delta_j \leq 1,$$

a trivial application of the matrix measure properties (2.3c) and (2.3e) leads to the claim (3.12).  $\square$

The standard estimate shows severe limitations since it discards the properties of the commutators, and, moreover, never yields decreasing global estimates. Loosely speaking, the standard estimate ignores the relations between the eigenvectors of the state matrices that are accounted for by the commutator. This means that even when (3.3) is satisfied, or when the matrices all commute and have a Hurwitz average, it still provides an upper bound for the maximum switching period.

As a numerical example, consider the two-dimensional periodic switched system from [19], whose state matrices are

$$\mathbf{A}_1 = \begin{bmatrix} -4 & -5 \\ 7 & 7 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 3 & 5 \\ -7 & -13 \end{bmatrix},$$

with the normalized time durations  $\delta_1 = 0.6, \delta_2 = 0.4$ . It is easy to check that (i)  $\mathbf{A}_1$  and  $\mathbf{A}_2$  are not Hurwitz; (ii)  $\bar{\mathbf{A}}_T$  is Hurwitz; and (iii) condition (3.3) for Euclidean norms is not satisfied. We estimate the maximum time that guarantees the stability of the switched system with three different methods: (i) solving (3.5) as elucidated by Proposition 3.1, with the matrix measure defined by the norm computed with  $\mathbf{R} = \mathbf{I}$ ; (ii) solving (3.5) upon using the error bound (3.12) as done in [11]; and (iii) determining directly the maximum switching period that guarantees the contraction property of the monodromy matrix with respect to the norm suggested in Proposition 3.1. These estimates are compared with the exact solution obtained by requiring that the characteristic multipliers (see, e.g., [12]) be less than one. In Table 1, we report the results of the computations. We emphasize that both the standard and the present approaches approximate the estimate obtained from the contraction condition. Nevertheless, the present approach provides a better estimation, as predicted by the theoretical analysis and illustrated by the numerical example.

TABLE 1

*Different estimates of the maximum switching period for uniform asymptotic stability and exact solution.*

	Switching period
Present work	$1.2 \times 10^{-1}$
Standard estimate	$4.3 \times 10^{-3}$
Contraction condition	$3.2 \times 10^{-1}$
Exact solution	$9.6 \times 10^{-1}$

Although the expression (3.5) appears complicated, it actually provides a convenient computational approach to estimating the critical switching period  $T^*$ . Indeed,  $\Gamma(\mathbf{A})$  defined in (3.10) is independent of  $T^*$ , as are the matrix measures of the individual subsystems and the average system. In contrast, the exact solution from Floquet theory requires expensive matrix logarithm computations that must be repeated for each candidate  $T^*$ . Any algorithm using Floquet theory should consider small time steps beginning at zero to account for the oscillatory behavior of Floquet exponents.

**4. Stability analysis of general linear time-varying systems.** In the present section we focus on the generalization of the method of averaging to arbitrary time-varying linear systems. The state matrix in (2.1) may be continuous, in which case the theory developed in this section provides a useful bound on the error incurred by estimating the transition matrix of the time-varying system over some time interval by the transition matrix of the average system over the same time interval. The state matrix may have points of discontinuity, however. In contrast to the switching systems studied in section 3, the state matrix  $\mathbf{A}(t)$  may vary between the discontinuities, and the system is not constrained to be periodic.

We begin by specifying a technique for estimating the transition matrix of (2.1) over an interval. The method involves dividing the interval into a large number of subintervals and approximating the state matrix over each subinterval by its value at the beginning of the subinterval. The transition matrix of the time-varying system (2.1) is then approximated by the transition matrix of the resulting switched system.

PROPOSITION 4.1. *Suppose that  $\mathbf{A}(t)$  in (2.1) is right continuous, with a finite number  $D$  of discontinuities on the interval  $[s, s + \tau]$  for any  $s, \tau \in \mathbb{R}^+$ . Consider a set of  $N + 1$  time instants  $\{t_i\}$  in  $[s, s + \tau]$ , where  $s = t_0 < t_1 < \cdots < t_N = s + \tau$ , consisting of the union of the points of discontinuity of  $\mathbf{A}(t)$  and the points  $s + k\frac{\tau}{M}$ , where  $k = 0, \dots, M$ , so that  $M \leq N \leq M + D$ . Define  $\mathbf{A}_i^\# = \mathbf{A}(t_{i-1})$  for all  $i = 1, \dots, N$ . Let  $\mathbf{A}^\#(t)$  be the corresponding switched system state matrix, whose pointwise in time value is selected from the set  $\Theta^\# = \{\mathbf{A}_1^\#, \dots, \mathbf{A}_N^\#\}$ . Let  $\Phi(t, s)$  be the transition matrix associated with  $\mathbf{A}(t)$ , and let  $\Phi^\#(t, s)$  be the transition matrix associated with  $\mathbf{A}^\#(t)$ . Then  $\Phi^\#(t, s) \rightarrow \Phi(t, s)$  uniformly on  $[s, s + \tau]$  as  $M \rightarrow \infty$ , that is, as  $N \rightarrow \infty$ .*

*Proof.* Let  $\mathbf{R}(t, s) = \Phi^\#(t, s) - \Phi(t, s)$ , and differentiate to obtain

$$(4.1) \quad \begin{aligned} \dot{\mathbf{R}}(t, s) &= \mathbf{A}^\#(t)\Phi^\#(t, s) - \mathbf{A}(t)\Phi(t, s) \\ &= \mathbf{A}(t)\mathbf{R}(t, s) + (\mathbf{A}^\#(t) - \mathbf{A}(t))\Phi^\#(t, s). \end{aligned}$$

Since  $\mathbf{R}(t, s) = 0$  at  $t = s$ , the solution to the differential equation (4.1) is

$$\mathbf{R}(t, s) = \int_s^t \Phi(t, \xi)(\mathbf{A}^\#(\xi) - \mathbf{A}(\xi))\Phi^\#(\xi, s)d\xi.$$

The error may be bounded as

$$\|\mathbf{R}(t, s)\| \leq \int_s^t \|\Phi(t, \xi)\| \|\mathbf{A}^\#(\xi) - \mathbf{A}(\xi)\| \|\Phi^\#(\xi, s)\| d\xi \leq \alpha\beta\gamma\tau,$$

where

$$\alpha = \operatorname{esssup}_{s, t \in [s, s + \tau]} \|\Phi(t, s)\|, \quad \beta = \operatorname{esssup}_{t \in [s, s + \tau]} \|\mathbf{A}^\#(t) - \mathbf{A}(t)\|, \quad \gamma = \operatorname{esssup}_{s, t \in [s, s + \tau]} \|\Phi^\#(t, s)\|.$$

Let

$$\zeta = \operatorname{esssup}_{t \in [s, s + \tau]} e^{\|\mathbf{A}(t)\|\tau}.$$

Because  $\mathbf{A}(t)$  has a finite number of discontinuities,  $\|\mathbf{A}(t)\|$  is bounded on the compact interval  $[s, s + \tau]$ . Using (2.3d), (2.3e), and (2.4), it follows that both  $\alpha$  and  $\gamma$  are bounded by  $\zeta$ , a quantity that is not influenced by switching behavior. As  $N \rightarrow \infty$ , the maximum time interval between switching instants vanishes, and right continuity of  $\mathbf{A}(t)$  ensures that  $\beta \rightarrow 0$ . It follows that  $\|\mathbf{R}(t, s)\| \rightarrow 0$ , so that the error between the transitions matrices  $\Phi^\#(t, s)$  and  $\Phi(t, s)$ , at every point  $t \in [s, s + \tau]$ , is bounded by a common vanishing quantity. Therefore  $\Phi^\#(t, s) \rightarrow \Phi(t, s)$  uniformly on  $[s, s + \tau]$  as  $N \rightarrow \infty$ .  $\square$

This result is now used to bound the difference between the time-varying system transition matrix and the corresponding average system transition matrix over a time interval.

PROPOSITION 4.2. *Suppose that  $\mathbf{A}(t)$  in (2.1) is right continuous and bounded, with a finite number of discontinuities on the interval  $[s, s + \tau]$  for any  $s, \tau \in \mathbb{R}^+$ . The transition matrix  $\Phi(s + \tau, s)$  of (2.1) over the interval  $[s + \tau, s]$  is given by*

$$(4.2) \quad \Phi(s + \tau, s) = \bar{\Phi}_\tau(s + \tau, s) + \mathbf{E}_\tau(s),$$

where  $\bar{\Phi}_\tau(s + \tau, s)$  is the transition matrix of the time-invariant average system represented by the sample average of  $\mathbf{A}(t)$  on the interval  $[s, s + \tau]$ . That is, the transition

matrix of

$$(4.3) \quad \bar{\mathbf{A}}_\tau(s) = \frac{1}{\tau} \int_s^{s+\tau} \mathbf{A}(\vartheta) d\vartheta$$

and  $\mathbf{E}_\tau(s)$  satisfies

$$(4.4) \quad \|\mathbf{E}_\tau(s)\| \leq \frac{1}{2} \int_s^{s+\tau} \int_\vartheta^{s+\tau} \|[\mathbf{A}(\vartheta), \mathbf{A}(\xi)]\| d\xi d\vartheta \exp\left(\int_s^{s+\tau} \mu(\mathbf{A}(\xi)) d\xi\right).$$

*Proof.* Define the  $N+1$  time instants  $\{t_i\}$  in  $[s, s+\tau]$ , the switching system state matrix  $\mathbf{A}^\#(t)$ , and the transition matrix  $\bar{\Phi}^\#(t, s)$  as in Proposition 4.1. Define the average  $\bar{\mathbf{A}}_\tau^\#(s)$  of the  $\mathbf{A}_i^\#$  matrices over the interval  $[s, s+\tau]$  as

$$\bar{\mathbf{A}}_\tau^\#(s) = \sum_{i=1}^N \delta_i \mathbf{A}_i^\#, \quad \delta_i = \frac{t_i - t_{i-1}}{\tau}.$$

Define  $\bar{\Phi}_\tau^\#(s+\tau, s)$  as the transition matrix over the interval  $[s, s+\tau]$  of the time-invariant system whose state matrix is  $\bar{\mathbf{A}}_\tau^\#(s)$ , and define the error  $\mathbf{E}_\tau^\#(s) = \Phi^\#(s+\tau, s) - \bar{\Phi}_\tau^\#(s+\tau, s)$ . From (3.9),  $\mathbf{E}_\tau^\#(s)$  is bounded by

$$(4.5) \quad \|\mathbf{E}_\tau^\#(s)\| \leq \frac{1}{2} \tau^2 \Gamma(\mathbf{A}^\#) \exp\left(\tau \sum_{i=1}^N \delta_i \mu(\mathbf{A}_i^\#)\right),$$

where  $\Gamma$  is defined as in (3.10). According to Proposition 4.1,  $\Phi^\#(s+\tau, s) \rightarrow \Phi(s+\tau, s)$  as  $N \rightarrow \infty$ . Similarly,  $\bar{\Phi}_\tau^\#(s+\tau, s) \rightarrow \bar{\Phi}_\tau(s+\tau, s)$  as  $N \rightarrow \infty$ . It follows that  $\mathbf{E}_\tau^\#(s) \rightarrow \mathbf{E}_\tau(s)$  as  $N \rightarrow \infty$ . The summations on the RHS of (4.5), including the double summation appearing in  $\Gamma(\mathbf{A}^\#)$ , are performed over the set of time instants  $\{t_i\}$  on the interval  $[s, s+\tau]$ , with the size of the steps between time instants determined by the  $\delta_i$ . Because of the way the  $N+1$  time instants are defined, the maximum step size approaches zero as  $N \rightarrow \infty$ . Thus the summations are consistent with the definition of Riemann sums (see, e.g., [3]), and they converge to integrals as  $N \rightarrow \infty$ . It follows that the RHS of (4.5) converges to the RHS of (4.4), and therefore that condition (4.4) holds.  $\square$

When the matrix  $\mathbf{A}$  is constant, the remainder in (4.2) vanishes. In this case, the present estimate yields the exact result since the commutator is zero. This is in sharp contrast with the standard estimate, which disregards the influence of the commutator, resulting in a generally looser bound depending only on the supremum of the time-varying matrix norm.

As a sample application of Proposition 4.2 (in the following proposition) we consider perturbed periodic switched systems.

**PROPOSITION 4.3.** *Consider the system (2.1) with*

$$(4.6) \quad \mathbf{A}(t) = \mathbf{F}(t) + \varepsilon \mathbf{B}(t),$$

where  $\mathbf{F}$  is a  $T$ -periodic switching matrix of the form (3.1),  $\mathbf{B}$  is a bounded right continuous matrix function which has at most a finite number of discontinuities within any switching period, and  $\varepsilon$  scales the perturbation magnitude. Assume that  $\mathbf{B}$  has zero average over any switching period. If for some norm

$$\sum_{i=1}^M \delta_i \mu(\mathbf{F}_i) + \varepsilon \operatorname{esssup}_{t \in \mathbb{R}^+} \mu(\mathbf{B}(t)) < 0,$$

then the system (4.6) is uniformly asymptotically stable for every positive  $T$ . Otherwise, if the average matrix

$$\bar{\mathbf{A}}_T = \sum_{i=1}^M \delta_i \mathbf{F}_i$$

of (4.6) is Hurwitz, then there exists a  $T^* > 0$  such that when  $T < T^*$  the system is uniformly asymptotically stable.  $T^*$  satisfies

$$e^{T^* \mu_P(\bar{\mathbf{A}}_T)} + \frac{1}{2} (T^*)^2 (\Gamma_P(\mathbf{F}) + 2\beta_1\varepsilon + \beta_2\varepsilon^2) \exp\left(T^* \left(\sum_{l=1}^M \delta_l \mu_P(\mathbf{F}_l) + \varepsilon\beta_3\right)\right) = 1,$$

where  $\Gamma_P$  is defined as in (3.6), the norm is constructed according to (3.7), and

$$\begin{aligned} \beta_1 &= \operatorname{esssup}_{t \in \mathbb{R}^+, i=1, \dots, M} \|[\mathbf{F}_i, \mathbf{B}(t)]\|_P, & \beta_2 &= \operatorname{esssup}_{t, \xi \in \mathbb{R}^+} \|[\mathbf{B}(t), \mathbf{B}(\xi)]\|_P, \\ \beta_3 &= \operatorname{esssup}_{t \in \mathbb{R}^+} \mu_P(\mathbf{B}(t)). \end{aligned}$$

*Proof.* We look for switching periods where the transition matrix of the system (4.6) from  $kT$  to  $(k+1)T$  is a contraction for every  $k \in \mathbb{Z}^+$ . To establish the first sufficient condition, we use matrix measure properties (2.4), (2.3c), and (2.3b) to obtain a bound on the transition matrix between consecutive switching events

$$\begin{aligned} \left\| \Phi\left(kT + \sum_{j=0}^i \delta_j T, kT + \sum_{j=0}^{i-1} \delta_j T\right) \right\| &\leq \exp\left(\delta_i T \operatorname{esssup}_{t \in \mathbb{R}^+} \mu(\mathbf{A}(t))\right) \\ &\leq \exp\left(\delta_i T \mu(\mathbf{F}_i) + \delta_i T \varepsilon \operatorname{esssup}_{t \in \mathbb{R}^+} \mu(\mathbf{B}(t))\right), \end{aligned}$$

where  $\delta_0 = 0$  by convention. Using the Schwarz inequality then yields

$$\|\Phi((k+1)T, kT)\| \leq \exp\left(T \sum_{i=1}^M \delta_i \mu(\mathbf{F}_i) + T \varepsilon \operatorname{esssup}_{t \in \mathbb{R}^+} \mu(\mathbf{B}(t))\right) < 1.$$

To establish the second sufficient condition, note that because  $\mathbf{B}$  has zero average over any switching period, Proposition 4.2 can be applied to the transition matrix of (4.6) to yield

$$(4.7) \quad \Phi((k+1)T, kT) = \exp(\bar{\mathbf{A}}_T T) + \mathbf{E}_T(kT),$$

where  $\|\mathbf{E}_T(kT)\|$  is bounded according to (4.4). We specialize to the norm constructed in (2.5), where  $\mathbf{P}$  is the unique symmetric positive definite solution of (3.7), and where  $\mathbf{R} = \mathbf{R}^T > 0$  is arbitrary. We look for switching periods satisfying

$$(4.8) \quad \frac{1}{2} \int_{kT}^{(k+1)T} \int_{\vartheta}^{(k+1)T} \|[\mathbf{A}(\vartheta), \mathbf{A}(\xi)]\|_P d\xi d\vartheta \exp\left(\int_{kT}^{(k+1)T} \mu_P(\mathbf{A}(\xi)) d\xi\right) + e^{T \mu_P(\bar{\mathbf{A}}_T)} < 1.$$

By applying the triangle inequality and accounting for the boundedness of  $\mathbf{B}$ , the double integral in (4.8) may be further bounded by

$$\int_{kT}^{(k+1)T} \int_{\vartheta}^{(k+1)T} \|[\mathbf{A}(\vartheta), \mathbf{A}(\xi)]\|_P d\xi d\vartheta \leq \Gamma_P(\mathbf{F}) + \beta_1\varepsilon + \beta_2\varepsilon^2.$$

By applying property (2.3c) to the exponent in (4.4), by applying the norm triangle inequality to (4.7), and by using property (2.3d), the claim follows.  $\square$

In the general case, when the state matrix is not of the form (4.6), the averaging method applies to systems of type (2.1) with state matrix  $\varpi\mathbf{A}(t)$ , where  $\varpi > 0$ . The averaging method addresses the effects of  $\varpi$  on the system stability. The small parameter  $\varpi$  has the effect of rescaling the system (2.1) to the fast time  $t/\varpi$ . Indeed, by a change of variable, the system may be rewritten as

$$\dot{\mathbf{x}}(t) = \mathbf{A}\left(\frac{t}{\varpi}\right)\mathbf{x}(t).$$

**PROPOSITION 4.4.** *Under the hypotheses of Proposition 4.2 and assuming that  $\mathbf{A}$  is bounded, if there is a period  $T$  such that*

$$(4.9) \quad \mu(\bar{\mathbf{A}}_T(kT)) < -\beta$$

*for all  $k \in \mathbb{N}$  and for some positive  $\beta$ , then there exists an  $\eta > 0$  such that the system (2.1) with state matrix  $\varpi\mathbf{A}(t)$ , where  $\varpi > 0$ , is uniformly asymptotically stable for  $\varpi T < \eta$ .*

*Proof.* From Proposition 4.2 the transition matrix of the system (2.1) with state matrix  $\varpi\mathbf{A}(t)$  from  $kT$  to  $(k+1)T$  is

$$(4.10) \quad \Phi((k+1)T, kT) = \exp(\varpi\bar{\mathbf{A}}_T(kT)T) + \mathbf{E}_T(kT),$$

where  $\|\mathbf{E}_T(kT)\|$  is bounded according to (4.4). That is,

$$\begin{aligned} \|\mathbf{E}_T(kT)\| &\leq \frac{1}{2} \varpi^2 \int_{kT}^{(k+1)T} \int_{\vartheta}^{(k+1)T} \|[\mathbf{A}(\vartheta), \mathbf{A}(\xi)]\| d\sigma d\vartheta \\ &\quad \times \exp\left(\varpi \int_{kT}^{(k+1)T} \mu(\mathbf{A}(\xi)) d\sigma\right). \end{aligned}$$

Since  $\mathbf{A}$  is bounded, the RHS of the previous inequality may be further bounded by

$$g(\varpi T) = \frac{1}{2} \varpi^2 T^2 \varepsilon \exp(\varpi T \alpha),$$

where

$$(4.11) \quad \alpha = \operatorname{esssup}_{t \in \mathbb{R}^+} \mu(\mathbf{A}(t)), \quad \varepsilon = \operatorname{esssup}_{t, \xi \in \mathbb{R}^+} \|[\mathbf{A}(t), \mathbf{A}(\xi)]\|.$$

The function  $g(\varpi T)$  is zero at  $\varpi = 0$ , as is its first derivative with respect to  $\varpi$ . By computing the norm of both sides of (4.10), by applying the norm triangle inequality and property (2.3d), and by accounting for (4.9), we obtain

$$\|\Phi((k+1)T, kT)\| \leq \exp(-\varpi T \beta) + g(\varpi T).$$

The condition (4.9) is analogous to (3.3) in Proposition 3.1 and ensures that the matrix measure is uniformly negative for the more general system considered here. Thus by

selecting  $\varpi$  sufficiently small,  $\Phi$  defines a contraction between any two consecutive instants  $kT, (k+1)T$ , and the system is asymptotically stable.  $\square$

**5. Conclusions.** A novel framework for analyzing periodic linear switched systems has been presented. By combining results on exponential splitting methods with the method of averaging, we have provided an improved understanding of the switching mechanism. We have elucidated the importance of commuting relations among the different subsystems constituting the switched system on stability properties. A new estimate of the maximum switching period that guarantees that the switched system is uniformly asymptotically stable when the average system matrix is Hurwitz has been established. It has been shown that this estimate is less conservative than the standard estimate. Moreover, our estimate is consistent with sufficient stability conditions in the literature, such as in [4]. The estimate effectiveness has been validated through a sample problem from [19] and compared to other estimates and to the exact solution from Floquet theory. The effect of time-varying perturbations on the system stability has been discussed and a modified estimate of the switching rate accounting for additive perturbations has been developed. Moreover, the method has been generalized to arbitrary time-varying linear systems, and a sufficient condition for uniform asymptotic stability is stated in terms of some properties of sampled averages.

**Acknowledgments.** The authors thank the anonymous reviewers for their helpful comments, and particularly for a correction in the proof of Proposition 3.1.

#### REFERENCES

- [1] R. W. BROCKETT, *Finite Dimensional Linear Systems*, Wiley, New York, 1970.
- [2] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.
- [3] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960.
- [4] J. EZZINE AND A. H. HADDAD, *Controllability and observability of hybrid systems*, Int. J. Control, 49 (1989), pp. 2045–2055.
- [5] J. EZZINE AND A. H. HADDAD, *Error bounds in the averaging of hybrid systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 1188–1192.
- [6] C. GÖKÇEK, *Stability analysis of periodically switched linear systems using Floquet theory*, Math. Probl. Eng., 1 (2004), pp. 1–10.
- [7] R. L. KOSUT, B. D. O. ANDERSON, AND I. M. Y. MAREELS, *Stability theory for adaptive systems: Method of averaging and persistency of excitation*, IEEE Trans. Automat. Control, 32 (1987), pp. 26–34.
- [8] D. LIBERZON, *Switching in Systems and Control*, Birkhäuser Boston, Boston, MA, 2003.
- [9] R. I. McLACHLAN, G. QUISPEL, AND W. REINOUT, *Splitting methods*, Acta Numer., 11 (2002), pp. 341–434.
- [10] C. MOLER AND C. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49.
- [11] D. G. ROBERSON AND D. J. STILWELL, *Control of an autonomous underwater vehicle platoon with a switched communication network*, in Proceedings of the IEEE American Control Conference, Portland, OR, 2005, pp. 4333–4338.
- [12] W. J. RUGH, *Linear System Theory*, 2nd ed., Prentice-Hall, London, 1995.
- [13] C. SCHWARTZ AND A. H. HADDAD, *Stability criteria for linear periodic switched systems*, in Proceedings of the IEEE American Control Conference, Denver, CO, 2003, pp. 3215–3219.
- [14] Q. SHENG, *Global error estimates for exponential splitting*, IMA J. Numer. Anal., 14 (1993), pp. 27–56.
- [15] T. STRÖM, *On logarithmic norms*, SIAM J. Numer. Anal., 12 (1975), pp. 741–753.
- [16] Z. SUN AND S. S. GE, *Switched Linear Systems: Control and Design*, Springer, London, 2005.
- [17] J. TOKARZEWSKI, *Stability of periodically switched linear systems and the switching frequency*, Int. J. Systems Sci., 18 (1987), pp. 697–726.



- [18] I. I. VRABIE, *Differential Equations: An Introduction to Basic Concepts, Results and Applications*, World Scientific, Singapore, 2004.
- [19] M. WICKS, P. PELETIES, AND R. DECARLO, *Switched controller synthesis for the quadratic stabilisation of a pair of unstable linear systems*, Eur. J. Control, 4 (1998), pp. 140–147.
- [20] G. XIE AND L. WANG, *Periodical stabilization of switched linear systems*, J. Comput. Appl. Math., 181 (2005), pp. 176–187.

## A CONSTRUCTIVE APPROACH TO A CLASS OF ERGODIC HJB EQUATIONS WITH UNBOUNDED AND NONSMOOTH COST\*

PATRICK CATTIAUX<sup>†</sup>, PAOLO DAI PRA<sup>‡</sup>, AND SYLVIE ROELLY<sup>§</sup>

**Abstract.** We consider a class of ergodic Hamilton–Jacobi–Bellman (HJB) equations related to long-time asymptotics of nonsmooth multiplicative functional of diffusion processes. Under suitable ergodicity assumptions on the underlying diffusion, we show existence of these asymptotics and that they solve the related HJB equation in the viscosity sense.

**Key words.** long-time asymptotics, cluster expansion, viscosity solution, HJB equation

**AMS subject classifications.** 47D07, 49L25, 60H10, 60J35, 60J60, 93E20

**DOI.** 10.1137/070698634

**1. Introduction.** Let  $(x_t)_{t \geq 0}$  be a continuous-time, homogeneous Markov process with infinitesimal generator  $L$ . To fix ideas, assume  $x_t$  is  $\mathbb{R}^d$ -valued. Given a function  $c : \mathbb{R}^d \rightarrow \mathbb{R}$  and  $\gamma > 0$ , we are interested in obtaining long-time asymptotics of the functional

$$S(T, x) := \log E_x \left[ \exp \left( \gamma \int_0^T c(x_t) dt \right) \right],$$

where  $E_x$  is the expectation conditioned to  $x_0 = x$ . Let  $\varphi(T, x) = e^{S(T, x)}$ . At least at the formal level,  $\varphi$  is a solution of the equation

$$\partial_t \varphi(t, x) = L\varphi(t, x) + \gamma c(x)\varphi(t, x).$$

If a Perron–Frobenius-type theorem holds for the operator  $L + \gamma c$ , then for  $T$  large  $\varphi(T, x)$  gets close to  $e^{\lambda T} v(x)$ , where  $\lambda$  is the largest eigenvalue of  $L + \gamma c$ , and  $v$  is the corresponding strictly positive eigenfunction. In other words, setting  $V(x) := \log v(x)$ , we obtain

$$S(T, x) = \lambda T + V(x) + o(T),$$

i.e.,

$$(1.1) \quad \lambda = \lim_{T \rightarrow +\infty} \frac{1}{T} \log E_x \left[ \exp \left( \gamma \int_0^T c(x_t) dt \right) \right]$$

and

$$(1.2) \quad V(x) = \lim_{T \rightarrow +\infty} \left\{ \log E_x \left[ \exp \left( \gamma \int_0^T c(x_t) dt \right) \right] - \lambda T \right\}.$$

---

\*Received by the editors July 30, 2007; accepted for publication (in revised form) June 5, 2008; published electronically October 13, 2008. This research was supported by the Deutsch-Französische Universität through the doctoral program CDFA 01-06.

<http://www.siam.org/journals/sicon/47-5/69863.html>

<sup>†</sup>Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse cedex 09, France (cattiaux@math.univ-toulouse.fr).

<sup>‡</sup>Matematica Pura e Applicata, Università di Padova, 31522 Padova, Italy (daipra@math.unipd.it).

<sup>§</sup>Institut für Mathematik, Universität Potsdam, 14469 Potsdam, Germany (roelly@math.uni-potsdam.de).

Note also that the pair  $(\lambda, V)$  is a solution of the nonlinear equation

$$(1.3) \quad \lambda = e^{-V} L(e^V) + \gamma c.$$

The actual proof of the existence of the limits (1.1) and (1.2) is, in general, not simple, and various assumptions are required. If the empirical measures

$$\mathcal{L}_t := \frac{1}{t} \int_0^t \delta_{x_s} ds$$

of the Markov process obey a large deviation principle with rate function  $i(\mu)$  (which is known under fairly general conditions), and  $c(\cdot)$  is measurable and bounded (but suitable growth conditions on  $c(\cdot)$  may suffice), then the limit (1.1) exists, and

$$(1.4) \quad \lambda = \sup_{\mu} \left[ \int c d\mu - i(\mu) \right],$$

where in (1.3)  $\mu$  varies over probability measures on  $\mathbb{R}^d$ . The existence of the limit (1.2), i.e., the second-order asymptotics of  $S(T, x)$ , is a harder problem to solve. For processes taking values in a compact space, where things are simpler, we refer to [7, section 4]. In this paper we consider  $\mathbb{R}^d$ -valued diffusions of the form

$$(1.5) \quad dx_t = b(x_t)dt + dB_t,$$

where  $(B_t)_{t \geq 0}$  is a standard Brownian motion and  $b$  is a regular drift function. Thus the associated infinitesimal generator is  $L = \frac{1}{2}\Delta + b(x) \cdot \nabla$ . The first results in this context date back to [6] and [14], where conditions are given for the existence of the solution of (1.3), which takes the form of the Hamilton–Jacobi–Bellman (HJB) equation

$$(1.6) \quad \begin{aligned} \lambda &= \frac{1}{2}\Delta V(x) + \max_{u \in \mathbb{R}^d} \left[ (b(x) + u) \cdot \nabla V(x) + \gamma c(x) - \frac{1}{2}|u|^2 \right] \\ &= \frac{1}{2}\Delta V(x) + b(x) \cdot \nabla V(x) + \frac{1}{2}|\nabla V(x)|^2 + \gamma c(x). \end{aligned}$$

In [6] it is also shown that, under sufficient ergodicity of  $(x_t)_{t \geq 0}$  and if  $c(\cdot)$  is bounded and sufficiently smooth, then (1.3) has a solution, which need not be the unique one, for which (1.1) and (1.2) hold. More recent results, which require boundedness from above of  $c(\cdot)$  but not smoothness, can be found in [5, Appendix B] or [20, Proposition 6.4].

The case of  $c(\cdot)$  unbounded has been recently dealt with in [11] and [10].

In [11] the authors deal with a discrete-time process; it is plausible that many of their proofs can be adapted to continuous time. Their approach is based on a rather sophisticated spectral theory (see also [12]). The translation of their results into our context would allow to prove the existence of the limits (1.1) and (1.2) for any measurable  $c(\cdot)$  whose growth at infinity is *strictly less* than quadratic. The assumptions are related to contractivity of the transition operator. In term of continuous-time diffusions, this corresponds to existence of *spectral gap* for the infinitesimal generator of the diffusion, in the space  $L^2(m)$ , where  $m$  is the invariant measure (see assumption A5 below).

In [10] the authors allow quadratic growth for  $c(\cdot)$ , but require differentiability. Using PDE methods they show, under reasonable conditions on  $b(\cdot)$ , that (1.3) indeed has multiple solutions, even after identifying solutions that differ by a constant. It is shown in [10] that there exists  $\lambda \in \mathbb{R}$  such that the equation

$$\mu = \frac{1}{2}\Delta V + b \cdot \nabla V + \frac{1}{2}|\nabla V|^2 + \gamma c$$

has a (smooth) solution if and only if  $\mu \geq \lambda$ . Moreover, for  $\mu = \lambda$ , this solution is unique up to additive constant. Kaise and Sheu also indicate that this  $\lambda$  should be as in (1.1). They do not address the possibility of interpreting *one* solution  $V$  as in (1.2).

The main objective of this paper is to propose a totally different approach to the above problems. On one hand we tackle (1.1) and (1.2), for the diffusion (1.5), directly, without relying on properties of (1.3). This makes it easy to avoid any regularity condition on  $c(\cdot)$ . On the other hand, unlike in [11], we allow  $c(\cdot)$  to have quadratic growth. We remark that quadratic growth of  $c(\cdot)$  makes a Gaussian concentration property (see assumption A3 below) a natural assumption. This property is by no means implied by existence of spectral gap for the generator (which corresponds to our assumption A5). Our assumptions A1–A6 are discussed in detail in section 3.

The method we propose is based on *cluster expansion*, a well-known method in statistical mechanics and combinatorics. Besides the technical advantage of allowing quadratic growth without requiring regularity, our approach has, we believe, other positive aspects as follows:

1. It is considerably simpler than both PDE and spectral methods. Moreover, in principle it allows us to obtain explicit estimates on the limits (1.1) and (1.2) in terms of various parameters related to the drift  $b(\cdot)$ .
2. It is a very robust method which can be adapted to various modifications of the problem considered here. For example, in (1.1) and (1.2) the integral

$$\int_0^T c(x_t) dt$$

could be replaced by

$$\int_{[0,T]} c(x_t) d\mu(t),$$

where  $\mu$  could be of the following forms:

- i.  $\mu$  is a  $\sigma$ -finite periodic measure, for instance,  $\mu(dt) = \sum_{k \geq 0} \delta_{k\Delta}(dt)$  for some  $\Delta > 0$ . In this last case the cost acts only at discrete time.
- ii.  $\mu$  is a random measure, independent of  $(B_t)_{t \geq 0}$ , translation invariant, and sufficiently ergodic in law. For instance, we could take  $\mu(dt) = \sum_n \delta_{\tau_n}(dt)$ , where  $(\tau_n)_{n \geq 0}$  are the points of a Poisson process.

Moreover, jump processes, rather than diffusions, should also be treatable.

We also remark that, although in this paper we consider diffusions whose diffusion coefficient is the identity matrix, the uniformly elliptic case could be dealt with after minor modifications. It is worth noticing that the whole content of section 2 is based on assumptions A1–A6 below, which do not refer to any specific form of the Markov process. The fact that the process is a diffusion plays a role in sections 3 and 4.

In the case when  $c(\cdot)$  has quadratic growth,  $S(t, x)$  could possibly explode in finite time, unless  $\gamma$  is sufficiently small. At the present stage, our results hold for  $\gamma$  in some

interval  $[0, \bar{\gamma}]$ , which is certainly *not optimal*. Note, however, that one could get an explicit expression for  $\bar{\gamma}$  (carefully following the proofs) as a function of the constants  $c_b$  and  $K_b$  appearing in conditions (DC) and (CC) of section 3.

The paper is organized as follows. In section 2 we prove existence of the limits (1.1) and (1.2) under some general conditions (A1–A6 below) on the diffusion process. In section 3 we give explicit sufficient conditions on the drift  $b$  for A1–A6 to hold. In section 4 we show that  $V$  and  $\lambda$  given by (1.2) and (1.1), respectively, are linked to (1.6); more precisely, we show that  $V$  is a viscosity solution of (1.6).

**2. Existence of the limits  $(\lambda, V)$ .** We begin by stating our assumptions on the  $\mathbb{R}^d$ -valued diffusion

$$(2.1) \quad dx_t = b(x_t)dt + dB_t.$$

**A1.** Equation (2.1) has, for every deterministic initial condition, a unique strong solution.

**A2.** There is a  $C > 0$  such that

$$|c(x)| \leq C(|x|^2 + 1), \quad x \in \mathbb{R}^d.$$

**A3.** The process  $(x_t)_t$  which is a solution of (2.1) has a unique invariant probability measure  $m(dx)$  such that, for some  $\beta > 0$ ,

$$\int e^{\beta|x|^2} m(dx) < +\infty.$$

**A4.** The transition probability of the process  $(x_t)_t$  admits a density  $p_t(x, y)$  with respect to the measure  $m$ . Furthermore, there exist  $K > 0$ ,  $p > 2$ , and  $t_0 > 0$  such that

$$\sup_{t \geq t_0} \|p_t(\cdot, \cdot)\|_{\mathbb{L}^p(m \otimes m)} \leq K.$$

**A5.** Let  $P_t$  be the semigroup associated with the process  $(x_t)_t$ . It extends as a continuous semigroup on  $\mathbb{L}^2(m)$  and satisfies

$$\forall f \in \mathbb{L}^2(m), \quad \lim_{t \rightarrow +\infty} \|P_t f - \int f dm\|_{\mathbb{L}^2(m)} = 0.$$

**A6.** For all  $a > 0$  and all  $x$ , there exists  $\beta_{a,x} > 0$  such that

$$E_x \left[ e^{\beta_{a,x} \int_0^a |x_s|^2 ds} \right] < +\infty.$$

We shall say that A6 is uniformly satisfied if for all  $a > 0$  there exist  $\beta_a > 0$  and a locally bounded function  $h_a$  such that for all  $x$ ,

$$E_x \left[ e^{\beta_a \int_0^a |x_s|^2 ds} \right] \leq h_a(x).$$

Section 3 will be devoted to giving sufficient conditions for these hypotheses to hold.

*Remark 1.* Assumption A4 implies that the semigroup  $P_t$  is continuous from  $\mathbb{L}^2(m)$  into  $\mathbb{L}^p(m)$ ,  $p > 2$ , for  $t \geq t_0$ . Hence, according to the Gross hypercontractivity theorem (see, e.g., [1]),  $m$  satisfies a defective logarithmic Sobolev inequality. If  $m$  is absolutely continuous with respect to the Lebesgue measure,  $m(dx) = e^{-V} dx$ , and

$V$  is locally bounded, a result by Röckner and Wang says that  $m$  satisfies a so-called weak Poincaré inequality; hence, thanks to a result by Aida,  $m$  will satisfy a tight log-Sobolev inequality (for all these results see the book of Wang [19]). In particular  $m$  will satisfy both a spectral gap inequality, so that Assumption A5 is satisfied, and a Gaussian concentration inequality implying A3.

**THEOREM 1.** *Under A1–A6 there is  $\bar{\gamma} > 0$  such that for every  $\gamma < \bar{\gamma}$  the limits (1.1) and (1.2) exist.*

*Furthermore, if A6 is uniformly satisfied, the convergence in (1.2) is uniform on compact sets.*

*Proof of Theorem 1.* We begin by showing that the limits (1.1) and (1.2) exist along suitable sequences.

**PROPOSITION 1.** *Under A1–A6, for every time-step  $a > 0$  large enough, there exists  $\gamma(a)$  such that for all  $\gamma < \gamma(a)$  and all  $x \in \mathbb{R}^d$ , the limits*

$$(2.2) \quad \lambda_a = \lim_{n \rightarrow +\infty} \frac{1}{an} \log E_x \left[ \exp \left( \gamma \int_0^{an} c(x_t) dt \right) \right]$$

and

$$(2.3) \quad V_a(x) = \lim_{n \rightarrow +\infty} \left\{ \log E_x \left[ \exp \left( \gamma \int_0^{an} c(x_t) dt \right) \right] - \lambda_a an \right\}$$

exist.

*Proof of Proposition 1.* The proof is done via a cluster expansion technique. The convergence of the expansion requires us to choose  $\gamma$  small enough and the time-step  $a$  large enough.

Define

$$(2.4) \quad \psi_\gamma(t, x, y) := \log E_{xy} \left[ \exp \left( \gamma \int_0^t c(x_s) ds \right) \right],$$

where  $E_{xy}$  denotes the expectation under the law of the bridge of  $(x_s)_{0 \leq s \leq t}$  between  $x$  and  $y$ , and consider a time-step  $a > 0$ . Then

$$(2.5) \quad e^{S(an, x)} = E_x \left[ \exp \left( \sum_{k=0}^{n-1} \psi_\gamma(a, x_{ka}, x_{(k+1)a}) \right) \right] = \mathbb{E} \left[ \exp \left( \sum_{k=0}^{n-1} \phi_\gamma(a, \xi_k, \xi_{k+1}) \right) \right],$$

where  $\mathbb{E}$  is the expectation with respect to a probability  $\mathbb{P}$ , and  $\xi_0 = x$ ,  $\xi_1, \dots, \xi_n$  are random variables that, under  $\mathbb{P}$ , are independent and identically distributed (i.i.d.) with law  $m(dx)$ , and

$$\phi_\gamma(a, x, y) = \psi_\gamma(a, x, y) + \log p_a(x, y).$$

A cluster in this context is a subset of  $\mathbb{Z}^+ := \{0, 1, 2, \dots\}$  of the form  $\{k, k+1, \dots, k+l\}$ . We say that two clusters are separated if there is an integer which is strictly larger than all elements of one cluster and strictly smaller than all elements of the other. We denote by  $\mathcal{C}$  the set of all clusters, while  $\mathcal{C}_n$  denotes the set of clusters

contained in  $\{0, 1, \dots, n-1\}$ . The usual cluster expansion procedure yields

$$\begin{aligned} \exp \left( \sum_{k=0}^{n-1} \phi_\gamma(a, \xi_k, \xi_{k+1}) \right) &= \prod_{k=0}^{n-1} \left[ \left( e^{\phi_\gamma(a, \xi_k, \xi_{k+1})} - 1 \right) + 1 \right] \\ &= \sum_{\tau \subseteq \{0, 1, \dots, n-1\}} \prod_{k \in \tau} \left( e^{\phi_\gamma(a, \xi_k, \xi_{k+1})} - 1 \right) = \sum_{p \geq 0} \frac{1}{p!} \sum_{\substack{\tau_1, \dots, \tau_p \in \mathcal{C}_n \\ \text{separated}}} \prod_{i=1}^p \prod_{k \in \tau_i} \left( e^{\phi_\gamma(a, \xi_k, \xi_{k+1})} - 1 \right) \\ &= \sum_{p \geq 0} \frac{1}{p!} \sum_{\substack{\tau_1, \dots, \tau_p \in \mathcal{C}_n \\ \text{separated}}} q_{\tau_1} q_{\tau_2} \cdots q_{\tau_p}, \end{aligned}$$

where

$$q_{\tau_i} := \prod_{k \in \tau_i} \left( e^{\phi_\gamma(a, \xi_k, \xi_{k+1})} - 1 \right),$$

and we have used the fact that any subset of  $\{0, \dots, n-1\}$  is a union of  $p$  separated clusters for some  $p \geq 0$ , and these clusters can be rearranged in  $p!$  ways. The key fact is that if  $\tau_i$  and  $\tau_j$  are separated clusters, then  $q_{\tau_i}$  and  $q_{\tau_j}$  are independent. Thus, by (2.5),

$$e^{S(an, x)} = \sum_{p \geq 0} \frac{1}{p!} \sum_{\substack{\tau_1, \dots, \tau_p \in \mathcal{C}_n \\ \text{separated}}} \mathbb{E}(q_{\tau_1}) \mathbb{E}(q_{\tau_2}) \cdots \mathbb{E}(q_{\tau_p}).$$

The logarithm of the above expression can be rewritten as

$$S(an, x) = \sum_{\tau \in \mathcal{C}_n, \tau \neq \emptyset} \sum_{p \geq 0} \sum_{\substack{\tau_1, \dots, \tau_p \in \mathcal{C}_n \\ \tau_1 \cup \dots \cup \tau_p = \tau}} a_p(\tau_1, \tau_2, \dots, \tau_p) \mathbb{E}(q_{\tau_1}) \mathbb{E}(q_{\tau_2}) \cdots \mathbb{E}(q_{\tau_p}) =: \sum_{\tau \in \mathcal{C}_n, \tau \neq \emptyset} \Gamma_\tau,$$

where the coefficients  $a_p(\tau_1, \dots, \tau_p)$  come from the Taylor expansion of the logarithm (see [13, page 492]). Now note that  $\Gamma_\tau$  depends on  $x$  if and only if  $0 \in \tau$ , i.e.,  $\tau = \{0, 1, \dots, m\}$  for some  $m$ . In what follows, we write  $\Gamma_m$  in place of  $\Gamma_{\{0, 1, \dots, m\}}$ . Thus

$$(2.6) \quad \sum_{\tau \in \mathcal{C}_n, \tau \neq \emptyset} \Gamma_\tau = \sum_{m=0}^{n-1} \Gamma_m + \sum_{i=1}^{n-1} \sum_{\substack{\tau \in \mathcal{C}_n, \tau \not\ni 0 \\ i \in \tau}} \frac{1}{|\tau|} \Gamma_\tau = \sum_{m=0}^{n-1} \Gamma_m + (n-1) \sum_{\substack{\tau \in \mathcal{C}_n, \tau \not\ni 0 \\ 1 \in \tau}} \frac{1}{|\tau|} \Gamma_\tau,$$

where we used the fact that, for  $0 \notin \tau$ ,  $\Gamma_\tau$  is invariant by translation and permutation of  $\tau$ . Thus, at a formal level, the limits (2.2) and (2.3) should be given by

$$(2.7) \quad \lambda_a = \frac{1}{a} \sum_{\substack{\tau \in \mathcal{C}, \tau \not\ni 0 \\ 1 \in \tau}} \frac{1}{|\tau|} \Gamma_\tau,$$

$$(2.8) \quad V_a(x) = \sum_{m=0}^{+\infty} \Gamma_m - a \lambda_a.$$

As usual (see, e.g., [4]), the convergence of the above sums will follow from the following strong cluster estimates:

$$(2.9) \quad \exists \rho < 1 \quad \forall \tau \in \mathcal{C} \text{ with } 0 \notin \tau, \quad |\mathbb{E}(q_\tau)| \leq \rho^{|\tau|}$$

and

$$(2.10) \quad \forall \tau \in \mathcal{C} \text{ with } \tau \ni 0, \quad |\mathbb{E}(q_\tau)| \leq C(x)\rho^{|\tau|},$$

where  $C(\cdot)$  is a locally bounded function of  $x$ .

Thus, we have only to prove the estimates (2.9) and (2.10) for  $\gamma$  sufficiently small. We begin by proving (2.9). By the generalized H  lder inequality in [15, Lemma 5.2], we have

$$(2.11) \quad |\mathbb{E}(q_\tau)| = \left| \mathbb{E} \left[ \prod_{k \in \tau} \left( e^{\phi_\gamma(a, \xi_k, \xi_{k+1})} - 1 \right) \right] \right| \leq \prod_{k \in \tau} \mathbb{E}^{1/2} \left[ \left( e^{\phi_\gamma(a, \xi_k, \xi_{k+1})} - 1 \right)^2 \right] = \rho^{|\tau|},$$

where

$$\rho := \mathbb{E}^{1/2} \left[ \left( e^{\phi_\gamma(a, \xi_1, \xi_2)} - 1 \right)^2 \right].$$

We now show that  $\rho$  can be made strictly less than 1 by choosing  $a$  sufficiently large and  $\gamma$  small enough:

$$\begin{aligned} \rho^2 &= \mathbb{E} \left( \left( e^{\psi_\gamma(a, \xi_1, \xi_2)} p_a(\xi_1, \xi_2) - 1 \right)^2 \right) \\ &= \int_{\mathbb{R}^{2d}} \left[ E_{xy} \left( e^{\gamma \int_0^a c(x_s) ds} \right) p_a(x, y) - 1 \right]^2 m(dx) m(dy) \\ &= \int_{\mathbb{R}^{2d}} \left[ E_{xy} \left( e^{\gamma \int_0^a c(x_s) ds} - 1 \right) p_a(x, y) + (p_a(x, y) - 1) \right]^2 m(dx) m(dy) \\ &\leq 2 \int_{\mathbb{R}^{2d}} E_{xy}^2 \left( e^{\gamma \int_0^a c(x_s) ds} - 1 \right) p_a^2(x, y) m(dx) m(dy) \\ &\quad + 2 \int_{\mathbb{R}^{2d}} (p_a(x, y) - 1)^2 m(dx) m(dy) \\ &=: 2I_1(a, \gamma) + 2I_2(a). \end{aligned}$$

We first analyze  $I_1(a, \gamma)$ . For any  $\varepsilon \in ]0, 1[$ , by the H  lder inequality,

$$\begin{aligned} I_1(a, \gamma) &= \left[ \int_{\mathbb{R}^{2d}} E_{xy}^{2/\varepsilon} \left( e^{\gamma \int_0^a c(x_s) ds} - 1 \right) p_a(x, y) m(dx) m(dy) \right]^\varepsilon \\ &\quad \cdot \left[ \int_{\mathbb{R}^{2d}} p_a^{\frac{2-\varepsilon}{1-\varepsilon}}(x, y) m(dx) m(dy) \right]^{1-\varepsilon} \\ &\leq E_m^\varepsilon \left( |e^{\gamma \int_0^a c(x_s) ds} - 1|^{2/\varepsilon} \right) \| p_a \|_{\mathbb{L}^{\frac{2-\varepsilon}{1-\varepsilon}}(m \otimes m)}^{2-\varepsilon}, \end{aligned}$$

where  $E_m$  denotes the expectation under the law of  $(x_t)_t$  with initial measure  $m$ . Thanks to assumption A4, for  $a$  large enough and  $\varepsilon$  small enough (such that  $\frac{2-\varepsilon}{1-\varepsilon} \leq p$ ),

$\| p_a \|_{\mathbb{L}^{\frac{2-\varepsilon}{1-\varepsilon}}(m \otimes m)}^{2-\varepsilon} < +\infty$ . To control  $J(a, \gamma) := E_m(|e^{\gamma \int_0^a c(x_s) ds} - 1|^{2/\varepsilon})$ , we represent  $e^{\gamma \int_0^a c(x_s) ds} - 1$  as  $\gamma \int_0^a c(x_s) ds \int_0^1 e^{u\gamma \int_0^a c(x_s) ds} du$  and obtain

$$\begin{aligned} J(a, \gamma) &= \gamma^{2/\varepsilon} E_m \left( \left| \int_0^a c(x_s) ds \right|^{2/\varepsilon} \left( \int_0^1 e^{u\gamma \int_0^a c(x_s) ds} du \right)^{2/\varepsilon} \right) \\ &\leq \gamma^{2/\varepsilon} E_m^{1/2} \left( \left| \int_0^a c(x_s) ds \right|^{4/\varepsilon} \right) E_m^{1/2} \left( e^{\frac{4\gamma}{\varepsilon} \int_0^a c(x_s) ds} \right). \end{aligned}$$



By Jensen's inequality,

$$\begin{aligned} J(a, \gamma) &\leq \gamma^{2/\varepsilon} a^{2/\varepsilon - 1/2} E_m^{1/2} \left( \int_0^a |c(x_s)|^{4/\varepsilon} ds \right) \left[ \frac{1}{a} \int_0^a E_m \left( e^{\frac{4\gamma}{\varepsilon} ac(x_s)} \right) ds \right]^{1/2} \\ &\leq (\gamma a)^{2/\varepsilon} \left[ \int |c(x)|^{4/\varepsilon} m(dx) \right]^{1/2} \left[ \int e^{\frac{4\gamma a}{\varepsilon} c(x)} m(dx) \right]^{1/2}. \end{aligned}$$

The first integral term on the right-hand side in the above inequality is finite due to assumptions A2 and A3. For the same reason, if  $\gamma a < \frac{\varepsilon}{4C}\beta$ , the last integral term of the right-hand side is finite. Then, for  $\gamma a$  small enough,  $J(a, \gamma)$  and  $I_1(a, \gamma)$  are as small as we want.

We now prove that  $I_2(a)$  goes to 0 as  $a \rightarrow +\infty$ .

LEMMA 1. *If A4 and A5 are satisfied,  $\lim_{a \rightarrow +\infty} \int_{\mathbb{R}^{2d}} (p_a(x, y) - 1)^2 m(dx) m(dy) = 0$ .*

*Proof.* By assumption A5, the semigroup  $P_t$  is a contraction on  $\mathbb{L}^2(m)$ , and

$$\lim_{t \rightarrow +\infty} \int \left| P_t f(x) - \left( \int f dm \right) \right|^2 m(dx) = 0.$$

Notice also that, for  $a > b > 0$  and  $m$  almost all  $y$ ,

$$p_a(x, y) = P_{a-b} p_b(\cdot, y)(x)$$

in  $\mathbb{L}^2(m)$  for all rational times  $a$  and  $b$  and, by invariance of  $m$ ,  $\int p_b(x, y) m(dx) = 1$ . Consider now an increasing sequence  $(a_n)_{n \geq 0}$  such that  $a_n \rightarrow +\infty$ . We have to show that, for any such sequence,

$$(2.12) \quad \lim_n \int_{\mathbb{R}^{2d}} (p_{a_n}(x, y) - 1)^2 m(dx) m(dy) = 0.$$

It is not restrictive to assume  $a_1 > t_0$ , where  $t_0$  is the constant in assumption A4. For ( $m$  almost all) fixed  $y$ ,

$$\int (p_{a_n}(x, y) - 1)^2 m(dx) = \int (P_{a_n - a_1} p_{a_1}(\cdot, y)(x) - 1)^2 m(dx) \rightarrow 0$$

by assumption A5. But thanks to assumption A4, the sequence

$$y \mapsto \int (p_{a_n}(x, y) - 1)^2 m(dx)$$

is uniformly integrable, which implies (2.12) by the Vitali convergence theorem.  $\square$

We can now conclude that for  $\gamma a$  small enough and for  $a$  large enough, the cluster estimate  $\rho$  is smaller than 1, which completes the proof of (2.9).

For the proof of (2.10), we proceed in the same way, just observing that the first factor in the right-hand side of (2.11) is now dependent on  $x$ . The additional term to control is  $E_x \left[ e^{q\gamma \int_0^a c(x_s) ds} \right]$  for some large  $q > 1$ . This can be done using A2 and A6. Thus, we completed the proof of Proposition 1.

To complete the proof of Theorem 1, we shall show why the limits (2.2) and (2.3) do not depend on the time-step  $a$ , yielding the limits (1.1) and (1.2). So we choose once and for all some convenient  $a$  and consider the corresponding set of convenient  $\gamma$ 's, yielding for each  $\gamma$  a  $\lambda$  obtained thanks to Proposition 1. For large  $T$  we choose  $n$  such that  $a(n-1) \leq T < an$ .

Notice that

$$S^-(an, x) \leq S(T, x) \leq S^+(an, x),$$

where

$$e^{S^-(an, x)} := E_x \left[ e^{\gamma \int_0^{a(n-1)} c(x_s) ds} e^{-\gamma \int_{a(n-1)}^{an} c^-(x_s) ds} \right]$$

and

$$e^{S^+(an, x)} := E_x \left[ e^{\gamma \int_0^{a(n-1)} c(x_s) ds} e^{\gamma \int_{a(n-1)}^{an} c^+(x_s) ds} \right].$$

Both  $S^-(an, x)$  and  $S^+(an, x)$  can be calculated using the same cluster expansion, except that now we have to replace  $\psi_\gamma(a, x_{a(n-1)}, x_{an})$  with a function  $\psi^-$  (resp.,  $\psi^+$ ) obtained by replacing  $c$  with  $-c^-$  (resp.,  $c^+$ ). We thus obtain a similar decomposition  $S^-(an, x) = \sum_{T \in \mathcal{C}_n, T \neq \emptyset} \Gamma_T^-$  with  $\Gamma_T^- = \Gamma_T$  if  $n-1 \notin T$  and with  $\Gamma_T^-$  obviously modified if  $n-1 \in T$ . In particular, in the decomposition (2.6) we see that, in the first sum, the only modified term is  $\Gamma_{n-1}$ . But since  $-c^-$  also satisfies A2, estimates similar to those in (2.9) and (2.10) hold true for both  $S^-$  and  $S^+$ , whose difference goes to 0 as  $n$  goes to infinity. This yields the desired result.

### 3. Some properties of diffusion processes and their invariant measures.

In this section we provide explicit conditions on the drift  $b(\cdot)$  for assumptions A1 and A3–A6 to hold for the diffusion process in (2.1). Our main result, Theorem 2 below, will be stated in terms of the following two drift conditions:

$$(3.1) \quad \text{Condition (DC)} \quad \exists c_b > 0 \text{ and } \exists R \geq 0 \text{ s.t. for } |x| \geq R, \quad b(x) \cdot x \leq -c_b |x|^2.$$

The second condition is usually called a ‘‘curvature condition.’’ Assume  $b \in C^1$ , and for  $\xi \in \mathbb{R}^d$  recall the notation

$$(3.2) \quad \langle \nabla_\xi b(x), \xi \rangle = \sum_{i,j} \xi_i \partial_i b_j(x) \xi_j.$$

The curvature condition is then as follows:

$$(3.3) \quad \text{Condition (CC)} \quad \exists K_b \in \mathbb{R} \text{ s.t. } \forall x \text{ and all } \xi, \quad \langle \nabla_\xi b(x), \xi \rangle \leq K_b |\xi|^2.$$

In what follows,  $L$  denotes the generator  $b(x) \cdot \nabla + \frac{1}{2} \Delta$ , and  $P_t$  denotes the associated semigroup.

**THEOREM 2.** *Let  $b \in C^1$ . If (DC) holds, then Assumptions A1 and A3 are satisfied, and assumption A6 is uniformly satisfied.*

*If (DC) and (CC) are both satisfied with constants  $c_b, K_b$  such that  $c_b > 2K_b$ , then assumptions A4 and A5 are satisfied.*

**Remark 2.** The condition  $c_b > 2K_b$  is a mild condition. Indeed, if we replace (DC) by a stronger (but more symmetric) condition, namely,

$$(b(x) - b(y)) \cdot (x - y) \leq -c_b |x - y|^2$$

(if  $b = -\nabla V$ , this is a convexity assumption), then we may choose  $K_b < 0$ .

It is also worth noticing that if we reinforce (DC), assuming the condition

$$\lim_{|x| \rightarrow \infty} b(x) \cdot \frac{x}{|x|^2} = -\infty,$$

then we may always choose  $c_b > 2K_b$  (if  $K_b$  is finite, of course). In this situation it can be shown (see [19, Corollary 5.7.7]) that the semigroup  $P_t$  is superbounded.

*Proof of Theorem 2.* We divide the proof into several steps.

*Step 1. A1 holds under (DC).* Since  $b \in C^1$ , it is locally Lipschitz; thus existence and strong uniqueness are ensured up to the explosion time, starting from any  $x$ . Define now  $\psi(x) = 1 + |x|^2$ . By condition (DC) it is easy to check that  $L\psi \leq C\psi$  for some  $C > 0$ . By applying Ito's rule to  $\psi(x_t)$  up to the exit time of the level sets of  $\psi$  (in the same spirit as in [16, Théorème 2.2.19]), it follows that the explosion time is a.s. infinite.

*Step 2. A3 holds under (DC).* Existence and uniqueness of the invariant measure  $m$  under (DC) follow, for instance, from [17, Theorem 7.4.21]. We prove here that, if (DC) holds, then for all  $\beta < c_b$ ,  $\int e^{\beta|y|^2} m(dy) < +\infty$ . Since (DC) holds, there exists  $D > 0$  such that  $b(x) \cdot x \leq -c_b|x|^2 + D$  for all  $x$ ; we set  $c = c_b$  in the proof of this step.

Let  $g_n$  be a smooth nondecreasing concave function defined on  $\mathbb{R}^+$  such that  $g_n(u) = u$  if  $u \leq n-1$  and  $g_n(u) = n$  if  $u \geq n$  (such a function exists). Let  $f_n(x) = \exp(\beta g_n(|x|^2))$  for  $\beta < c$ .

Then  $\nabla f_n(x) = 2\beta f_n(x)g'_n(|x|^2)x$  and

$$\Delta f_n(x) = 2\beta f_n(x) (2g''_n(|x|^2)|x|^2 + 2\beta(g'_n)^2(|x|^2)|x|^2 + dg'_n(|x|^2))$$

so that

$$\begin{aligned} Lf_n(x) &= \beta f_n(x) ((2g''_n(|x|^2)|x|^2 + dg'_n(|x|^2)) + 2g'_n(|x|^2)(\beta g'_n(|x|^2)|x|^2 + b(x) \cdot x)) \\ &\leq \beta f_n(x)(d + 2D - 2(c - \beta)|x|^2) \\ &\leq \beta(d + 2D)e^{\beta \frac{d+2D}{c-\beta}} - \beta(d + 2D) f_n(x) \end{aligned}$$

since

$$d + 2D - 2(c - \beta)|x|^2 \leq -(c - \beta)|x|^2 \leq -(d + 2D)$$

for  $|x|^2 \geq \frac{d+2D}{c-\beta}$ .

In short, there exist  $c_1$  and  $c_2$  positive constants such that for all  $n$ ,  $Lf_n \leq c_1 - c_2 f_n$ .

Define  $h_n(s) = E_x[e^{\beta g_n(|x_s|^2)}]$ . Ito's formula yields

$$h_n(t) \leq h_n(0) + c_1 t - c_2 \int_0^t h_n(s) ds,$$

and hence by applying Gronwall's lemma, we obtain

$$(3.4) \quad E_x \left[ e^{\beta g_n(|x_t|^2)} \right] \leq \frac{c_1}{c_2} + e^{-c_2 t} e^{\beta g_n(|x|^2)}.$$

Integrating (3.4) with respect to the invariant measure  $m$  yields

$$(1 - e^{-c_2 t}) \int e^{\beta g_n(|y|^2)} m(dy) \leq \frac{c_1}{c_2}.$$

We may thus choose  $t$  large enough for  $e^{-c_2 t} \leq 1/2$  and then use the monotone convergence theorem with  $n \rightarrow +\infty$  in order to obtain  $\int e^{\beta|y|^2} m(dy) < +\infty$  for  $\beta < c_b$ .

*Step 3.* A6 is uniformly satisfied under (DC). Using Ito's formula up to the exit time  $T_M$  of the ball of center 0 and radius  $M$ , we have

$$E_x \left[ e^{\theta |x_{t \wedge T_M}|^2} \right] = e^{\theta |x|^2} + E_x \left[ \int_0^{t \wedge T_M} (2\theta b(x_s) \cdot x_s + d\theta + 2\theta^2 |x_s|^2) e^{\theta |x_s|^2} ds \right].$$

In particular, if condition (DC) holds with  $c_b > \theta$ , the integrand in the right-hand side is nonpositive for large values of  $|x_s|$ , and hence we can let  $M$  go to infinity in order to show that there exists some constant  $\kappa$  (depending on (DC) and  $\theta$ ) such that

$$E_x \left[ e^{\theta |x_t|^2} \right] < e^{\theta |x|^2} + \kappa t.$$

Accordingly, using the Jensen inequality, we obtain

$$\begin{aligned} E_x \left[ e^{\int_0^a \beta |x_s|^2 ds} \right] &= E_x \left[ e^{\frac{1}{a} \int_0^a a \beta |x_s|^2 ds} \right] \\ &\leq \frac{1}{a} E_x \left[ \int_0^a e^{a \beta |x_s|^2} ds \right] < +\infty \end{aligned}$$

as soon as  $a\beta < c_b$ . The proof is completed.

*Step 4.* A4 holds if (DC) and (CC) are both satisfied and if  $c_b > 2K_b$ . Since  $b \in C^1$ , Malliavin calculus shows that the law of  $x_t$  is absolutely continuous w.r.t. the Lebesgue measure for all initial conditions  $x$  and all  $t > 0$ . Hence  $m$  is also absolutely continuous w.r.t. the Lebesgue measure, and it can be shown that  $dm/dy$  is a.e. positive. Thus, the existence of  $p_t(x, y)$  follows. The proof of the integrability condition stated in A4 relies on a beautiful Harnack inequality derived by Wang (see [19, Theorem 2.5.2]),

$$(3.5) \quad (P_t f(x))^\alpha \leq P_t f^\alpha(y) \exp \left( \frac{\alpha}{2(\alpha-1)} K_b (1 - e^{-2K_b t})^{-1} |x - y|^2 \right),$$

holding for  $t > 0$ ,  $\alpha > 1$ , all  $(x, y)$ , and all nonnegative continuous and bounded  $f$ , with the convention  $K_b(1 - e^{-2K_b t})^{-1} = 1/2t$  if  $K_b \leq 0$  (see also [1, Lemma 7.5.4] if  $\alpha = 2$ ).

Using (3.5), we show that for all  $p > 2$ ,  $p_t(\cdot, \cdot) \in \mathbb{L}^p(m \otimes m)$  for all  $t$  such that

$$c_b > \frac{K_b p(p-1)}{1 - e^{-2K_b t}}.$$

In particular, if  $K_b \leq 0$ , then for all  $p > 2$  there exists  $t_p$  such that  $p_t(\cdot, \cdot) \in \mathbb{L}^p(m \otimes m)$  for  $t \geq t_p$ , while for  $K_b > 0$  such a  $t_p$  exists provided  $c_b > K_b p(p-1)$ . We shall first derive an upper bound for the density.

Let  $\alpha > 1$ ,  $D_t := \{x \in \mathbb{R}^d, |x| \leq \gamma(t)\}$  for some increasing function  $\gamma$  going to  $\infty$ , and let  $f$  be nonnegative and bounded. Integrating the Harnack inequality for  $P_t$  with respect to  $m(dy)$  on  $D_t$  and denoting

$$\kappa(t) = \frac{\alpha}{2(\alpha-1)} K_b (1 - e^{-2K_b t})^{-1},$$

we get

$$\begin{aligned} ((P_t f)(x))^\alpha &\leq \int_{D_t} (P_t f^\alpha)(y) e^{\kappa(t)|x-y|^2} m(dy)/m(D_t) \\ &\leq \int f^\alpha(y) (P_t^*(\mathbf{1}_{D_t}(\cdot) e^{\kappa(t)|x-\cdot|^2}))(y) m(dy)/m(D_t) \\ &\leq e^{2\kappa(t)(|x|^2 + \gamma^2(t))} \int f^\alpha(y) m(dy)/m(D_t) \end{aligned}$$

since

$$\| \mathbb{1}_{D_t}(\cdot) e^{\kappa(t)|x-\cdot|^2} \|_\infty \leq e^{2\kappa(t)(|x|^2 + \gamma^2(t))}.$$

If

$$\theta(t, x) = \left( e^{2\kappa(t)\gamma^2(t)} / m(D_t) \right) e^{2\kappa(t)|x|^2},$$

we thus have

$$(3.6) \quad ((P_t f)(x))^\alpha \leq \theta(t) \int f^\alpha(y) m(dy).$$

Applying the previous inequality with a continuous approximation of  $f_N(z) = p_t^\beta(x, z) \mathbb{1}_{\{p_t(x, z) \leq N\}}$  and then taking limits, we have

$$\left( \int p_t^{1+\beta}(x, z) \mathbb{1}_{\{p_t(x, z) \leq N\}} m(dz) \right)^\alpha \leq \theta(t, x) \int p_t^{\alpha\beta}(x, y) \mathbb{1}_{\{p_t(x, y) \leq N\}} m(dy);$$

i.e., by letting  $N$  go to  $\infty$  and choosing  $1 + \beta = \alpha\beta$  and hence  $\beta = 1/(\alpha - 1)$  we obtain

$$(3.7) \quad \int p_t^{\frac{\alpha}{\alpha-1}}(x, y) m(dy) \leq \theta^{1/(\alpha-1)}(t, x).$$

By Step 2, the right-hand side in (3.7) is in  $\mathbb{L}^1(m)$  provided

$$(3.8) \quad c_b > \frac{2\kappa(t)}{\alpha - 1} = \frac{\alpha K_b}{(\alpha - 1)^2(1 - e^{-2K_b t})}.$$

In particular if  $p > 2$ , define  $1 < \alpha = p/(p - 1) < 2$ . Hence  $p_t(\cdot, \cdot) \in \mathbb{L}^p(m \otimes m)$  provided

$$c_b > \frac{K_b p(p - 1)}{1 - e^{-2K_b t}}.$$

*Step 5.* A5 holds if (DC) and (CC) are both satisfied, and  $c_b > K_b$ . (Note that the condition needed here is weaker than  $c_b > 2K_b$ .) It is well known that the  $L^2(m)$ -contractivity stated in assumption A5 is implied by *hypercontractivity* of  $P_t$ , which means that for all  $1 < p < q < +\infty$  there exists  $t_{p,q}$  such that for  $t \geq t_{p,q}$ ,  $P_t$  is a bounded operator from  $\mathbb{L}^p(m)$  into  $\mathbb{L}^q(m)$  with norm equal to 1. The fact that  $c_b > K_b$  implies hypercontractivity of  $P_t$  is shown in [19, Theorem 5.7.3, Corollary 5.7.2, and Theorem 5.7.1].

**4. The limiting function as viscosity solution.** We consider the function

$$(4.1) \quad \varphi(t, x) := E_x \left[ \exp \left( \gamma \int_0^t c(x_s) ds \right) \right].$$

We have shown that (under some assumptions we shall assume to be in force below), for  $\gamma$  sufficiently small, the limits

$$(4.2) \quad \lambda := \lim_t \frac{1}{t} \log \varphi(t, x)$$

and

$$(4.3) \quad V(x) := \lim_t [\log \varphi(t, x) - \lambda t]$$

exist uniformly over compact sets. We want to show that  $V$  is a *viscosity solution* of the HJB equation (1.6) or, equivalently, that  $v(x) := e^{V(x)}$  is a viscosity solution of the linear equation

$$(4.4) \quad - \left[ \frac{1}{2} \Delta v + b \cdot \nabla v + \gamma c v \right] + \lambda v = 0.$$

We first prove that  $\varphi(T - t, x)$  is a continuous viscosity solution of a suitable evolution equation. Then by using (4.3) we show that (4.4) holds.

This problem has been dealt with in [9] in a much more general setting. However, the assumptions given in [9] are not satisfied here, due to the unboundedness of  $c$ . Thus, some modifications of their proof are needed.

**PROPOSITION 2.** *Assume that conditions A1–A5 are satisfied and that condition (DC) is satisfied, so that condition A6 is uniformly satisfied (in particular, the strong Feller property holds). Moreover, let  $\bar{\gamma}$  be as in Theorem 1, and assume  $\gamma < \bar{\gamma}$  (hence the limits (4.2) and (4.3) exist). Then  $v(\cdot)$  is continuous and is a viscosity solution of (4.4).*

*Proof. Step 1. Continuity of  $\varphi(t, x)$ .* We first establish continuity in  $x$ .

First note that, according to the proofs in section 2, for  $\gamma < \bar{\gamma}$ , one can find some  $\delta \in ]1, \frac{\bar{\gamma}}{\gamma}[$  and some function  $h_\gamma(t, x)$ , which is bounded on compact sets such that

$$(4.5) \quad E_x \left[ \exp \left( \gamma \delta \int_0^t |c(x_s)| ds \right) \right] \leq h_\gamma(t, x)$$

for all  $t > 0$  and  $x \in \mathbb{R}^d$ .

Note that, for  $0 < \epsilon < t$ ,

$$(4.6) \quad \begin{aligned} & |\varphi(t, x) - \varphi(t, y)| \\ &= \left| E_x \left[ \varphi(t - \epsilon, x_\epsilon) \exp \left( \gamma \int_0^\epsilon c(x_s) ds \right) \right] \right. \\ &\quad \left. - E_y \left[ \varphi(t - \epsilon, x_\epsilon) \exp \left( \gamma \int_0^\epsilon c(x_s) ds \right) \right] \right| \\ &\leq E_x \left[ \left| e^{\gamma \int_0^\epsilon c(x_s) ds} - 1 \right| \varphi(t - \epsilon, x_\epsilon) \right] + E_y \left[ \left| e^{\gamma \int_0^\epsilon c(x_s) ds} - 1 \right| \varphi(t - \epsilon, x_\epsilon) \right] \\ &\quad + |E_x[\varphi(t - \epsilon, x_\epsilon)] - E_y[\varphi(t - \epsilon, x_\epsilon)]|. \end{aligned}$$

We begin by estimating the first term in the right-hand side of (4.6). By the Hölder inequality,

$$(4.7) \quad \begin{aligned} & E_x \left[ \left| e^{\gamma \int_0^\epsilon c(x_s) ds} - 1 \right| \varphi(t - \epsilon, x_\epsilon) \right] \\ &\leq \left\{ E_x \left[ \left| e^{\gamma \int_0^\epsilon c(x_s) ds} - 1 \right|^p \right] \right\}^{1/p} \left\{ E_x \left[ \exp \left( \gamma \delta \int_0^t |c(x_s)| ds \right) \right] \right\}^{1/\delta}, \end{aligned}$$

where  $p = \frac{\delta}{\delta-1}$ . Our goal is to show that the left-hand side of (4.7) goes to 0 as  $\epsilon \rightarrow 0$ , uniformly in  $x$  varying in a compact set. By (4.5), the second factor in the right-hand side of (4.7) is locally bounded. Thus, it is enough to show that

$$E_x \left[ \left| e^{\gamma \int_0^\epsilon c(x_s) ds} - 1 \right|^p \right]$$

goes to zero uniformly in compact sets. By the inequality  $|e^x - 1| \leq |x|e^{|x|}$ , the Cauchy-Schwarz inequality, and Jensen's inequality,

$$\begin{aligned} E_x \left[ \left| e^{\gamma \int_0^\epsilon c(x_s) ds} - 1 \right|^p \right]^2 &\leq \gamma^{2p} E_x \left[ \left( \int_0^\epsilon c(x_s) ds \right)^p e^{\gamma p \int_0^\epsilon |c(x_s)| ds} \right]^2 \\ &\leq \gamma^{2p} E_x \left[ \left( \int_0^\epsilon c(x_s) ds \right)^{2p} \right] E_x \left[ e^{2\gamma p \int_0^\epsilon |c(x_s)| ds} \right] \\ &\leq \epsilon^{2p-1} \gamma^{2p} \int_0^\epsilon E_x [|c(x_s)|^{2p}] ds E_x \left[ e^{2\gamma p \int_0^\epsilon |c(x_s)| ds} \right] \\ (4.8) \quad &\leq \epsilon^{2p-2} \gamma^{2p} \int_0^\epsilon E_x [|c(x_s)|^{2p}] ds \int_0^\epsilon E_x \left[ e^{2\gamma p \epsilon |c(x_s)|} \right] ds. \end{aligned}$$

Since  $p > 1$ , it is enough to show that the two integrals in (4.8) are locally bounded. This follows easily from the assumption that  $c(\cdot)$  has quadratic growth (see A2, where the constant  $C$  is defined), and from the proof of the first part of Theorem 2, as soon as  $2\gamma p C \epsilon < c_b$ , that holds true for  $\epsilon$  small enough. Indeed we get some exponential integrability which is strong enough to control both terms.

It remains to deal with the last term in (4.6). It is enough to show that, for given  $\epsilon > 0$ , the map

$$(4.9) \quad x \mapsto E_x[\varphi(t - \epsilon, x_\epsilon)]$$

is continuous in  $x$ . For this purpose, we realize all diffusion starting from any  $x \in \mathbb{R}^d$  in the same probability space. We denote by  $X_t(x)$  the diffusion starting from  $x$ , and denote by  $E$  the expectation in this probability space. Thus

$$E_x[\varphi(t - \epsilon, x_\epsilon)] = E[\varphi(t - \epsilon, X_\epsilon(x))].$$

By (4.5),

$$E[\varphi^\delta(t - \epsilon, X_\epsilon(x))]$$

is locally bounded in  $x$ . This implies that, for any ball  $B$ , the family of random variables

$$(\varphi(t - \epsilon, X_\epsilon(x)))_{x \in B}$$

is uniformly integrable. Thus, letting  $\varphi_M(x) := \varphi(t - \epsilon, x) \mathbf{1}_{[0, M]}(|\varphi(t - \epsilon, x)|)$  ( $\mathbf{1}_A$  is the indicator function of the set  $A$ ), we have that

- for every  $M > 0$ ,  $E[\varphi_M(X_\epsilon(x))]$  is continuous in  $x$  by the strong Feller property; and
- $E[\varphi(t - \epsilon, X_\epsilon(x))] - E[\varphi_M(X_\epsilon(x))]$  goes to zero as  $M \rightarrow +\infty$  uniformly in  $x \in B$  for any ball  $B$ .

From these two statements, continuity of (4.9) follows.

To get joint continuity in  $(t, x)$  just observe that, by the integrability condition (4.5), we can differentiate in  $t$   $\varphi(t, x)$  and show that this derivative is locally bounded. Thus  $\varphi(t, x)$  is locally Lipschitz in  $t$ , locally uniformly in  $x$ . This, together with continuity in  $x$ , implies joint continuity.

*Step 2. Viscosity solution of the parabolic equation.* In what follows we introduce the upper-semicontinuous (resp., lower-semicontinuous) extension  $c^*$  (resp.,  $c_*$ ) of  $c(\cdot)$ :

$$c^*(x) := \limsup_{y \rightarrow x} c(y), \quad c_*(x) := \liminf_{y \rightarrow x} c(y).$$

Moreover, let  $v_T(t, x) := \varphi(T - t, x)$ . We now show that  $v_T$  is a viscosity solution (in  $[0, T]$ ) of the parabolic equation

$$(4.10) \quad - \left( \partial_t v_T + b \cdot \nabla v_T + \frac{1}{2} \Delta v_T + \gamma c v_T \right) = 0.$$

Since  $v_T$  is continuous, this amounts to showing that the following two properties hold true:

- i. (supersolution property). Let  $(t, x) \in [0, T) \times \mathbb{R}^d$  and let  $\psi : [0, T) \times \mathbb{R}^d \rightarrow \mathbb{R}$  be a smooth function such that  $\psi(t, x) = v_T(t, x)$  and  $v_T - \psi$  has a local maximum at  $(t, x)$  (there may be no such function). Then

$$- \left( \partial_t \psi(t, x) + b(x) \cdot \nabla \psi(t, x) + \frac{1}{2} \Delta \psi(t, x) + \gamma c^*(x) v_T(t, x) \right) \leq 0.$$

- ii. (subsolution property). Let  $(t, x) \in [0, T) \times \mathbb{R}^d$  and let  $\psi : [0, T) \times \mathbb{R}^d \rightarrow \mathbb{R}$  be a smooth function such that  $\psi(t, x) = v_T(t, x)$  and  $v_T - \psi$  has a local minimum at  $(t, x)$ . Then

$$- \left( \partial_t \psi(t, x) + b(x) \cdot \nabla \psi(t, x) + \frac{1}{2} \Delta \psi(t, x) + \gamma c_*(x) v_T(t, x) \right) \geq 0.$$

$v_T - \psi$  has a *strict* local extreme in  $(t, x)$ . Indeed, if  $v_T - \psi$  has a local extreme at  $(t, x)$  and  $\tilde{\psi}(s, y) := \psi(s, y) \pm [(s - t)^2 + |x - y|^4]$  (where the sign depends on whether we are dealing with a maximum or a minimum), then  $v_T - \tilde{\psi}$  has a *strict* local extreme in  $(t, x)$ , and  $\psi$  and  $\tilde{\psi}$  have the same first space and time derivatives and second space derivatives at  $(t, x)$ . We now observe the following identities:

$$\begin{aligned} \varphi(t, x) &= 1 - \int_0^t \frac{d}{ds} E_x \left[ \exp \left( \int_s^t \gamma c(x_\tau) d\tau \right) \right] ds \\ &= 1 + \gamma \int_0^t E_x \left[ c(x_s) \exp \left( \gamma \int_s^t c(x_\tau) d\tau \right) \right] ds \\ &= 1 + \gamma \int_0^t E_x [c(x_s) \varphi(t - s, x_s)] ds, \end{aligned}$$

where all steps are justified by (4.5). It follows that, for  $\epsilon > 0$ ,

$$\varphi(t, x) - E_x[\varphi(t - \epsilon, x_\epsilon)] = \gamma E_x \left[ \int_0^\epsilon c(x_s) \varphi(t - s, x_s) ds \right].$$



By a change  $t \mapsto T - t$  of the time variable, we get

$$(4.11) \quad v_T(t, x) - E_x[v_T(t + \epsilon, x_\epsilon)] = \gamma E_x \left[ \int_0^\epsilon c(x_s) v_T(t + s, x_s) ds \right].$$

Now we use (4.11) to prove that  $v_T$  has the subsolution property. The supersolution property is proved in the same way. Note that both properties are local, so it is not restrictive to assume the test functions  $\psi$  to have compact support.

So let  $\psi$  be a smooth function with compact support such that  $\psi(t, x) = v_T(t, x)$  and  $v_T - \psi$  has a local minimum at  $(t, x)$ . We first claim that

$$(4.12) \quad \limsup_{\epsilon \rightarrow 0} \frac{v_T(t, x) - E_x[v_T(t + \epsilon, x_\epsilon)]}{\epsilon} \leq \limsup_{\epsilon \rightarrow 0} \frac{\psi(t, x) - E_x[\psi(t + \epsilon, x_\epsilon)]}{\epsilon}.$$

This is done by a simple localization. Let  $\rho > 0$  be such that  $v_T(s, y) \geq \psi(s, y)$  for  $(s, y) \in [t - \rho, t + \rho] \times B(x, \rho)$ . Then, for  $|\epsilon| < \rho$ ,

$$\begin{aligned} & \frac{v_T(t, x) - E_x[v_T(t + \epsilon, x_\epsilon)]}{\epsilon} \\ &= E_x \left[ \frac{v_T(t, x) - v_T(t + \epsilon, x_\epsilon)}{\epsilon} \mathbb{1}_{|x_\epsilon - x| \leq \rho} \right] + E_x \left[ \frac{v_T(t, x) - v_T(t + \epsilon, x_\epsilon)}{\epsilon} \mathbb{1}_{|x_\epsilon - x| > \rho} \right] \\ &\leq E_x \left[ \frac{\psi(t, x) - \psi(t + \epsilon, x_\epsilon)}{\epsilon} \mathbb{1}_{|x_\epsilon - x| \leq \rho} \right] + E_x \left[ \frac{v_T(t, x) - v_T(t + \epsilon, x_\epsilon)}{\epsilon} \mathbb{1}_{|x_\epsilon - x| > \rho} \right] \\ &= E_x \left[ \frac{\psi(t, x) - \psi(t + \epsilon, x_\epsilon)}{\epsilon} \right] \\ &\quad - E_x \left[ \frac{\psi(t, x) - \psi(t + \epsilon, x_\epsilon)}{\epsilon} \mathbb{1}_{|x_\epsilon - x| > \rho} \right] + E_x \left[ \frac{v_T(t, x) - v_T(t + \epsilon, x_\epsilon)}{\epsilon} \mathbb{1}_{|x_\epsilon - x| > \rho} \right]. \end{aligned}$$

Thus, in order to obtain (4.12), it is enough to show that the last two terms go to zero as  $\epsilon \rightarrow 0$ . We deal only with the last (the others being easier),

$$\begin{aligned} \left| E_x \left[ \frac{v_T(t, x) - v_T(t + \epsilon, x_\epsilon)}{\epsilon} \mathbb{1}_{|x_\epsilon - x| > \rho} \right] \right| &\leq \frac{2}{\epsilon} E_x \left[ e^{\gamma \int_0^T |c(x_s)| ds} \mathbb{1}_{|x_\epsilon - x| > \rho} \right] \\ &\leq \frac{2}{\epsilon} E_x \left[ e^{\gamma \delta \int_0^T |c(x_s)| ds} \right] E_x(\mathbb{1}_{|x_\epsilon - x| > \rho})^{1 - \frac{1}{\delta}}, \end{aligned}$$

which goes to zero as  $\epsilon \rightarrow 0$  since, by small time estimates (see, e.g., [18]),  $E_x(\mathbb{1}_{|x_\epsilon - x| > \rho}) = o(\epsilon)$ . This establishes (4.12). On the other hand, by Ito's rule,

$$(4.13) \quad \lim_{\epsilon \rightarrow 0} \frac{\psi(t, x) - E_x[\psi(t + \epsilon, x_\epsilon)]}{\epsilon} = - \left( \partial_t \psi(t, x) + b(t, x) \cdot \nabla \psi(t, x) + \frac{1}{2} \Delta \psi(t, x) \right).$$

Putting together (4.11), (4.12), and (4.13), the subsolution property follows from

$$(4.14) \quad \liminf_{\epsilon \rightarrow 0} \frac{1}{\epsilon} E_x \left[ \int_0^\epsilon c(x_s) v_T(t + s, x_s) ds \right] \geq c_*(x) v_T(t, x),$$

where the above convergence is again controlled by small time estimates and the fact that  $v_T$  is continuous.

*Step 3. Conclusion.* Letting  $\tilde{v}_T(t, x) := v_T(t, x)e^{-\lambda(T-t)}$ , it is easily checked that  $\tilde{v}_T$  is a viscosity solution of

$$(4.15) \quad - \left( \partial_t \tilde{v}_T + b \cdot \nabla \tilde{v}_T + \frac{1}{2} \Delta \tilde{v}_T + \gamma c v_T \right) + \lambda \tilde{v} = 0.$$

Moreover,  $\tilde{v}_T(t, x) \rightarrow v(x)$  as  $T \rightarrow +\infty$  uniformly on compact sets. In particular,  $v$  is continuous. We now sketch a standard argument to show that  $v$  is a viscosity solution of (4.4).

Let  $x \in \mathbb{R}^d$ , and let  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$  be a smooth function such that  $v(x) = \psi(x)$  and  $v - \psi$  has a local minimum at  $x$ . Fix  $t > 0$ , and define  $\tilde{\psi}(s, y) := \psi(y) - |y - x|^4 - (s - t)^2$ . Note that  $v - \tilde{\psi}$  has a *strict* local minimum at  $(t, x)$ , and

$$(4.16) \quad \partial_t \tilde{\psi}(t, x) + b(t, x) \cdot \nabla \tilde{\psi}(t, x) + \frac{1}{2} \Delta \tilde{\psi}(t, x) = b(x) \cdot \nabla \psi(x) + \frac{1}{2} \Delta \psi(x).$$

A simple exercise in uniform convergence shows that there is a sequence  $(t_n, x_n) \rightarrow (t, x)$  as  $n \rightarrow +\infty$  such that  $\tilde{v}_n - \tilde{\psi}$  has a local minimum at  $(t_n, x_n)$ . Therefore, since  $\tilde{v}$  is a viscosity solution of (4.15),

$$(4.17) \quad - \left( \partial_t \tilde{\psi}(t_n, x_n) + b(x_n) \cdot \nabla \tilde{\psi}(t_n, x_n) + \frac{1}{2} \Delta \tilde{\psi}(t_n, x_n) + \gamma c_*(x_n) \tilde{v}_n(x_n) \right) + \lambda \tilde{v}_n(x_n) \geq 0.$$

Letting  $n \rightarrow +\infty$  and using (4.16) and lower-semicontinuity of  $c_*$ , we obtain the subsolution property for (4.4). The supersolution property is obtained in the same way.  $\square$

**Acknowledgment.** The authors thank an anonymous referee for bringing to their attention the reference [12].

## REFERENCES

- [1] C. ANÉ, S. BLACHÈRE, D. CHAFAI, P. FOUGÈRES, I. GENTIL, F. MALRIEU, C. ROBERTO, AND G. SCHEFFER, *Sur les inégalités de Sobolev logarithmiques*, Panoramas et Synthèses 10, Société Mathématique de France, Paris, 2000.
- [2] D. BAKRY, M. LEDOUX, AND Z. QIAN, *Logarithmic Sobolev Inequalities, Poincaré Inequalities and Heat Kernel Bounds*, unpublished manuscript, available online at <http://www.lsp.upstlse.fr/Fp/Qian/BLQ.pdf> (1997).
- [3] V. I. BOGACHEV, N. V. KRYLOV, AND M. RÖCKNER, *On regularity of transition probabilities and invariant measures of singular diffusions under minimal conditions*, Comm. Partial Differential Equations, 26 (2001), pp. 2037–2080.
- [4] P. DAI PRA AND S. RÆLLY, *An existence result for infinite-dimensional Brownian diffusions with non-regular and non-Markovian drift*, Markov Process. Related Fields, 10 (2004), pp. 113–136.
- [5] J. FENG AND T. G. KURTZ, *Large Deviations for Stochastic Processes*, Math. Surveys Monogr. 131, American Mathematical Society, Providence, RI, 2006.
- [6] W. H. FLEMING AND W. M. MCENEANEY, *Risk-sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.
- [7] W. FLEMING, S. J. SHEU, AND M. SONER, *On the existence of the dominant eigenfunction and its application to the large deviation properties of an ergodic Markov process*, Stochastics, 22 (1987), pp. 187–199.
- [8] H. FÖLLMER, *Time reversal on Wiener space*, in Stochastic Processes—Mathematics and Physics, Lecture Notes in Math. 1158, Springer, 1986, pp. 119–129.
- [9] A. GULISASHVILI AND J. A. VAN CASTEREN, *Non-autonomous Kato Classes and Feynman-Kac Propagators*, World Scientific, Hackensack, NJ, 2006.
- [10] H. KAISE AND S. J. SHEU, *On the structure of solutions of ergodic type Bellman equation related to risk-sensitive control*, Ann. Probab., 34 (2006), pp. 284–320.
- [11] I. KONTOYIANNIS AND S. P. MEYN, *Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes*, Electron. J. Probab., 10 (2005), pp. 61–123.
- [12] I. KONTOYIANNIS AND S. P. MEYN, *Spectral theory and limit theorems for geometrically ergodic Markov processes*, Ann. Appl. Probab., 13 (2003), pp. 304–362.
- [13] R. KOTECKÝ AND D. PREISS, *Cluster expansions for abstract polymer models*, Comm. Math. Phys., 103 (1986), pp. 491–498.

- [14] W. M. McENEANEY AND K. ITO, *Infinite time-horizon risk-sensitive systems with quadratic growth*, in Proceedings of the 36th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 1997, pp. 3413–3418.
- [15] R. A. MINLOS, A. VERBEURE, AND V. ZAGREBNOV, *A quantum crystal model in the light-mass limit: Gibbs states*, Rev. Math. Phys., 12 (2000), pp. 981–1032.
- [16] G. ROYER, *Une initiation aux inégalités de Sobolev logarithmiques*, Cours Spécialisés 5, Société Mathématique de France, Paris, 1999.
- [17] D. W. STROOCK, *Probability Theory: An Analytic View*, Cambridge University Press, Cambridge, UK, 1999.
- [18] S. R. S. VARADHAN, *Diffusion processes in a small time interval*, Comm. Pure Appl. Math., 20 (1967), pp. 659–685.
- [19] F. Y. WANG, *Functional Inequalities, Markov Processes, and Spectral Theory*, Science Press, Beijing, 2004.
- [20] L. WU, *Uniformly integrable operators and large deviations for Markov processes*, J. Funct. Anal., 172 (2000), pp. 301–376.

## DYNAMIC PROGRAMMING PRINCIPLE FOR ONE KIND OF STOCHASTIC RECURSIVE OPTIMAL CONTROL PROBLEM AND HAMILTON–JACOBI–BELLMAN EQUATION\*

ZHEN WU<sup>†</sup> AND ZHIYONG YU<sup>‡</sup>

**Abstract.** In this paper, we study one kind of stochastic recursive optimal control problem with the obstacle constraint for the cost functional described by the solution of a reflected backward stochastic differential equation. We give the dynamic programming principle for this kind of optimal control problem and show that the value function is the unique viscosity solution of the obstacle problem for the corresponding Hamilton–Jacobi–Bellman equation.

**Key words.** reflected backward stochastic differential equation, recursive optimal control problem, dynamic programming principle, Hamilton–Jacobi–Bellman equation, viscosity solution

**AMS subject classifications.** 93E20, 60H10, 35K15

**DOI.** 10.1137/060671917

**1. Introduction.** The nonlinear backward stochastic differential equation (BSDE) was been introduced by Pardoux and Peng [12]. Independently, Duffie and Epstein [6] introduced BSDE from economic background. In [6] they presented a stochastic differential recursive utility which is an extension of the standard additive utility with the instantaneous utility depending not only on the instantaneous consumption rate but also on the future utility. Actually, it corresponds to the solution of a particular BSDE whose generator does not depend on the variable  $z$ . From a mathematical point of view the results in [12] are more general. Then El Karoui, Peng, and Quenez [11] gave some important properties of BSDEs such as comparison theorem and applications in mathematical finance and optimal control theory. And also in the same paper, the authors gave the formulation of recursive utilities and their properties from the BSDE point of view. The recursive optimal control problem is presented as a kind of optimal control problem whose cost functional is described by the solution of BSDE. In 1992, Peng [13] got the Bellman’s dynamic programming principle for this kind of problem and proved that the value function is a viscosity solution of one kind of quasi-linear second order partial differential equation (PDE) which is the well known Hamilton–Jacobi–Bellman equation. Later in 1997, he virtually generalized these results to a much more general situation, under Markovian and even non-Markovian framework (see [14, Chapter 2]). In the Chinese version, Peng used the backward semigroup property of BSDE to prove Bellman’s dynamic programming principle for the recursive optimal problem introduced by a BSDE under Markovian and non-Markovian framework. He also proved that the value function is a viscosity solution of a generalized Hamilton–Jacobi–Bellman equation.

---

\*Received by the editors October 10, 2006; accepted for publication (in revised form) May 13, 2008; published electronically October 13, 2008. This work is supported by the National Natural Science Foundation (10671112), the National Basic Research Program of China (973 Program, No. 2007CB814901 and No. 2007CB814904), the Natural Science Foundation of Shandong Province (Z2006A01), and the Chinese New Century Young Teachers Program.

<http://www.siam.org/journals/sicon/47-5/67191.html>

<sup>†</sup>School of Mathematics and System Science, Shandong University, Jinan 250100, China (wuzhen@sdu.edu.cn).

<sup>‡</sup>Corresponding author. School of Mathematics and System Science, and School of Economics, Shandong University, Jinan 250100, China (yuzhiyong@sdu.edu.cn).

In 1997, El Karoui et al. [9] studied the reflected BSDE with one continuous barrier. The solution of the reflected BSDE is forced to stay above a given continuous stochastic process called the barrier. For this purpose they introduced an increasing process to push the solution upwards in a kind of minimal way. Using two different methods, optimal stopping problem with fixed point principle and penalization method, they got the existence and uniqueness of the solution for this kind of reflected BSDE. The solution of reflected BSDE provides a probabilistic representation for the unique viscosity solution of an obstacle problem for a nonlinear parabolic partial differential equation within the Markovian framework. This kind of reflected BSDE also has applications in the financial market. We know from El Karoui et al. [11] that the pricing of the European contingent claim can be formulated in terms of BSDEs. Since pricing an American option can be formulated as an optimal stopping problem, El Karoui, Pardoux, and Quenez [10], showed that the price of an American option corresponds to the solution of a reflected BSDE, even in an imperfect market.

Cvitanic and Karatzas [5] extended the result to reflected BSDE with upper and lower barriers. Hamadène and Lepeltier [7] used reflected BSDE to solve a mixed optimal stochastic control problem. In this kind of problem, the controller has two actions, one is control and the other is stopping his control strategy in order to maximize his payoff. And then Hamadène, Lepeltier, and Wu [8] generalized the results for infinite horizon reflected BSDEs with one or two barriers and also applied those results to the mixed control and mixed game problem.

In our paper, we study one kind of recursive optimal control problem with the obstacle constraint for the cost functional; i.e., the cost functional of the control system is described by the solution of a reflected BSDE with one lower barrier. This kind of recursive optimal control problem has some practical meaning. For example, in the financial market, an investor requires his recursive utility function to be bigger than a certain function of his wealth. For this purpose, an increasing process is introduced to push the cost functional upward; we require this push to be minimum. From the results in [9] and [7], we know that, in many cases, this problem can be regarded as a mixed optimal stochastic control problem. We also will show that the pricing problem of the American option when loan interest is higher than deposit interest can be reformulated to this kind of recursive optimal control problem.

One of our interesting problems is that if the dynamic programming principle still holds for this recursive optimal control problem. Using some properties of the reflected BSDE and analysis techniques we give a positive answer for this question. Then we show that, if the problem is formulated within a Markovian framework, the value function is the unique viscosity solution of the obstacle problem for a nonlinear parabolic PDE which is called the Hamilton–Jacobi–Bellman (HJB) equation.

Our paper generalizes the dynamic programming principle of the recursive control problem in [13] and [14] to the obstacle constraint case and has the following advantages and improvements. First, in Peng [13] and [14], the recursive cost function does not have the obstacle constraint. In our paper, the optimization problem has obstacle constraints for the cost function which is described by the solution of reflected BSDEs. Such a problem has more financial applications and can formulate the pricing problem of the American option in a complete or incomplete market. Many ideas and methods of proof in our paper for dynamic programming principle come from [13] and [14]. However, we generalize the results to obstacle constraint case whose coefficients just satisfy Lipschitz condition, so our results are more general. Second, the method of proof in our paper is mainly based on elementary mathematics analysis technique and the properties of reflected BSDEs. We improved two properties of

reflected BSDEs in [9] (see Propositions 2.1 and 2.2) which play important roles for the continuation of the value function about  $t$  (see Proposition 3.12, also our proof is more clear). Finally, the proof of existence for the viscosity solution of the obstacle problem for the corresponding HJB equations in our paper is completely different from the one in Peng [13] and [14]. Our result is the generalization of the one in Karoui et al. [9] for the obstacle problem of nonlinear parabolic PDEs. We need to consider changing the order between  $\lim$  and  $\sup$  as according to Peng, [13] and [14], there lacks the uniqueness proof for the viscosity solution. In our paper, we apply the method introduced by Barles, Buckdahn, and Pardoux [1] to give the uniqueness proof for the viscosity solution of corresponding HJB equations.

The paper is organized as follows. In section 2, we present some preliminary results about reflected BSDE which play an important role to study the dynamic programming principle of the optimal control problem. In section 3, we formulate the recursive optimal control problem with the obstacle constraint for the cost functional and prove that the dynamic programming principle holds. In section 4, we prove that the value function of the control problem is the unique viscosity solution of the obstacle problem for the corresponding HJB equation.

**2. Preliminary results of the reflected BSDE.** In this section, we give some preliminary results of the reflected BSDE which are useful for the dynamic programming principle for the recursive optimal control problem with the obstacle constraint for the cost functional.

Let  $\{W_t, 0 \leq t \leq T\}$  be a  $d$ -dimensional standard Brownian motion defined on a probability space  $(\Omega, \mathcal{F}, P)$ . Let  $\{\mathcal{F}_t, 0 \leq t \leq T\}$  be the natural filtration of  $\{W_t\}$ , where  $\mathcal{F}_0$  contains all  $P$ -null sets of  $\mathcal{F}$ , and, let  $\mathcal{P}$  be the  $\sigma$ -algebra of predictable subsets of  $\Omega \times [0, T]$ .

Let us introduce some notations

$$L^2 = \left\{ \xi \text{ is an } \mathcal{F}_T\text{-measurable random variable s.t. } \mathbb{E}(|\xi|^2) < +\infty \right\},$$

$$H^2 = \left\{ \{\varphi_t, 0 \leq t \leq T\} \text{ is an adapted process s.t. } \mathbb{E} \int_0^T |\varphi_t|^2 dt < +\infty \right\},$$

$$S^2 = \left\{ \{\varphi_t, 0 \leq t \leq T\} \text{ is an adapted process s.t. } \mathbb{E} \left( \sup_{0 \leq t \leq T} |\varphi_t|^2 \right) < +\infty \right\},$$

and the following reflected BSDE with one barrier:

$$(2.1) \quad Y_t = \xi + \int_t^T g(s, Y_s, Z_s) ds + K_T - K_t - \int_t^T Z_s dW_s, \quad 0 \leq t \leq T.$$

Here  $\xi \in L^2$ ,  $g$  is a map from  $\Omega \times [0, T] \times \mathbb{R} \times \mathbb{R}^d$  into  $\mathbb{R}$  satisfying

- (i) for all  $(y, z) \in \mathbb{R} \times \mathbb{R}^d$ ,  $g(\cdot, y, z) \in H^2$ ,
- (ii) for some  $L > 0$  and all  $y, y' \in \mathbb{R}$ ,  $z, z' \in \mathbb{R}^d$ , a.s.

$$|g(t, y, z) - g(t, y', z')| \leq L(|y - y'| + |z - z'|),$$

$\{S_t, 0 \leq t \leq T\}$  called an “obstacle” is a continuous adapted real-valued process satisfying

- (iii)  $\mathbb{E}(\sup_{0 \leq t \leq T} |S_t|^2) < +\infty$ .

Then from Theorem 5.2 in [9], there exists a unique solution  $\{(Y_t, Z_t, K_t), 0 \leq t \leq T\}$  taking values in  $\mathbb{R}$ ,  $\mathbb{R}^d$  and  $\mathbb{R}_+$ , respectively, and satisfying

(iv)  $Y \in S^2$ ,  $Z \in H^2$ , and  $K_T \in L^2$ ;

(v)  $Y_t \geq S_t$ ,  $0 \leq t \leq T$ ;

(vi)  $\{K_t\}$  is adapted, continuous and increasing,  $K_0 = 0$ , and  $\int_0^T (Y_t - S_t) dK_t = 0$ .

Now we give two more accurate estimates of the solutions than that of Propositions 3.5 and 3.6 in [9]. They are necessary for proof of the dynamic programming principle of our optimal control problem and play an important role for the continuation properties of value function  $u(t, x)$  about  $t$  and  $x$ . The analogous generalization results for the nonreflected case can be seen in [2]. Since some proof technique is derived from [2], we omit it.

PROPOSITION 2.1. *Let  $\{(Y_t, Z_t, K_t), 0 \leq t \leq T\}$  be the solution of the above reflected BSDE  $(\xi, g, S)$ , then there exists a constant  $C$  such that*

$$\begin{aligned} & \mathbb{E}^{\mathcal{F}_t} \left\{ \sup_{t \leq s \leq T} Y_s^2 + \int_t^T |Z_s|^2 + |K_T - K_t|^2 \right\} \\ & \leq C \mathbb{E}^{\mathcal{F}_t} \left\{ \xi^2 + \left( \int_t^T g(s, 0, 0) ds \right)^2 + \sup_{t \leq s \leq T} S_s^2 \right\}. \end{aligned}$$

PROPOSITION 2.2. *Let  $(\xi, g, S)$  and  $(\xi', g', S')$  be two triplets satisfying the above assumptions. Suppose  $(Y, Z, K)$  is the solution of the reflected BSDE  $(\xi, g, S)$  and  $(Y', Z', K')$  is the solution of the reflected BSDE  $(\xi', g', S')$ . Define*

$$\Delta \xi = \xi - \xi', \quad \Delta g = g - g', \quad \Delta S = S - S';$$

$$\Delta Y = Y - Y', \quad \Delta Z = Z - Z', \quad \Delta K = K - K'.$$

Then there exists a constant  $C$  such that

$$\begin{aligned} & \mathbb{E}^{\mathcal{F}_t} \left\{ \sup_{t \leq s \leq T} |\Delta Y_s|^2 + \int_t^T |\Delta Z_s|^2 ds + |\Delta K_T - \Delta K_t|^2 \right\} \\ & \leq C \mathbb{E}^{\mathcal{F}_t} \left\{ |\Delta \xi|^2 + \left( \int_t^T |\Delta g(s, Y_s, Z_s)| ds \right)^2 \right\} + C \left( \mathbb{E}^{\mathcal{F}_t} \left\{ \sup_{t \leq s \leq T} |\Delta S_s|^2 \right\} \right)^{1/2} \Psi_{t,T}^{1/2}, \end{aligned}$$

where

$$\begin{aligned} \Psi_{t,T} = \mathbb{E}^{\mathcal{F}_t} \left\{ & |\xi|^2 + \left( \int_t^T |g(s, 0, 0)| ds \right)^2 + \sup_{t \leq s \leq T} |S_s|^2 \right. \\ & \left. + |\xi'|^2 + \left( \int_t^T |g'(s, 0, 0)| ds \right)^2 + \sup_{t \leq s \leq T} |S'_s|^2 \right\}. \end{aligned}$$

**3. Formulation of the problem and the dynamic programming principle.** In this section, we first formulate one kind of stochastic recursive optimal control problem with the obstacle constraint for the cost functional, and then we prove that the dynamic programming principle still holds for this kind of optimization problem.

We introduce the admissible control set  $\mathcal{U}$  defined by

$$\mathcal{U} := \{v(\cdot) \in H^2 \mid v(\cdot) \text{ take value in } U \subset \mathbb{R}^k\}.$$

An element of  $\mathcal{U}$  is called an admissible control. Here  $U$  is a compact subset of  $\mathbb{R}^k$ ; however, this restriction is often satisfied in practical applications.

For a given admissible control, we consider the following control system:

$$(3.1) \quad \begin{cases} dX_s^{t,\zeta;v} &= b(s, X_s^{t,\zeta;v}, v_s)ds + \sigma(s, X_s^{t,\zeta;v}, v_s)dW_s, & s \in [t, T], \\ X_t^{t,\zeta;v} &= \zeta, \end{cases}$$

where  $t \geq 0$  is regarded as the initial time, and  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$  as the initial state. The mappings

$$b : [0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}^n, \quad \sigma : [0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}^{n \times d}$$

satisfy the following conditions:

(H3.1)  $b$  and  $\sigma$  are continuous in  $t$ ,

(H3.2) for some  $L > 0$ , and all  $x, x' \in \mathbb{R}^n$ ,  $v, v' \in U$ , a.s.

$$|b(t, x, v) - b(t, x', v')| + |\sigma(t, x, v) - \sigma(t, x', v')| \leq L(|x - x'| + |v - v'|).$$

Obviously, under the above assumptions, for any  $v(\cdot) \in \mathcal{U}$ , the control system (3.1) has a unique strong solution  $\{X_s^{t,\zeta;v}, 0 \leq t \leq s \leq T\}$ , and we also have the following estimates:

PROPOSITION 3.1. For all  $t \in [0, T]$ ,  $\zeta, \zeta' \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ ,  $v(\cdot), v'(\cdot) \in \mathcal{U}$ ,

$$(3.2) \quad \mathbb{E}^{\mathcal{F}_t} \left\{ \sup_{t \leq s \leq T} |X_s^{t,\zeta;v}|^2 \right\} \leq C(1 + |\zeta|^2),$$

$$(3.3) \quad \mathbb{E}^{\mathcal{F}_t} \left\{ \sup_{t \leq s \leq T} |X_s^{t,\zeta;v} - X_s^{t,\zeta';v'}|^2 \right\} \leq C|\zeta - \zeta'|^2 + C\mathbb{E}^{\mathcal{F}_t} \left\{ \int_t^T |v_s - v'_s|^2 ds \right\},$$

where the constant  $C$  depends on  $L, T$ , and the compact set  $U$ .

PROPOSITION 3.2. For all  $t \in [0, T]$ ,  $x \in \mathbb{R}^n$ ,  $v(\cdot) \in \mathcal{U}$ ,  $\delta \in [0, T - t]$ ,

$$(3.4) \quad \mathbb{E} \left\{ \sup_{t \leq s \leq t+\delta} |X_s^{t,x;v} - x|^2 \right\} \leq C\delta,$$

where the constant  $C$  depends on  $x, L$ , and the compact set  $U$ .

Now for any given admissible control  $v(\cdot) \in \mathcal{U}$ , we consider the following reflected BSDE:

$$(3.5) \quad \begin{aligned} Y_s^{t,\zeta;v} &= \Phi(X_T^{t,\zeta;v}) + \int_s^T g(r, X_r^{t,\zeta;v}, Y_r^{t,\zeta;v}, Z_r^{t,\zeta;v}, v_r)dr \\ &\quad + K_T^{t,\zeta;v} - K_s^{t,\zeta;v} - \int_s^T Z_r^{t,\zeta;v} dW_r, \quad t \leq s \leq T, \end{aligned}$$



where

$$\Phi = \Phi(x) : \mathbb{R}^n \rightarrow \mathbb{R}, \quad h = h(t, x) : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R},$$

$$g = g(t, x, y, z, v) : [0, T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \times U \rightarrow \mathbb{R}$$

satisfy the following conditions:

(H3.3)  $g$  and  $h$  are continuous in  $t$ ;

(H3.4) for some  $L > 0$ , and all  $x, x' \in \mathbb{R}^n$ ,  $y, y' \in \mathbb{R}$ ,  $z, z' \in \mathbb{R}^d$ ,  $v, v' \in U$ , a.s.

$$\begin{aligned} & |g(t, x, y, z, v) - g(t, x', y', z', v')| + |\Phi(x) - \Phi(x')| + |h(t, x) - h(t, x')| \\ & \leq L(|x - x'| + |y - y'| + |z - z'| + |v - v'|). \end{aligned}$$

Then from Theorem 5.2 in [9], there exists a triple  $(Y^{t,\zeta;v}, Z^{t,\zeta;v}, K^{t,\zeta;v})$ , which is the unique solution of the reflected BSDE (3.5), satisfying

- (i)  $Y^{t,\zeta;v} \in S^2$ ,  $Z^{t,\zeta;v} \in H^2$  and  $K_T^{t,\zeta;v} \in L^2$ ;
- (ii)  $Y_s^{t,\zeta;v} \geq h(s, X_s^{t,\zeta;v})$ ,  $t \leq s \leq T$ ;
- (iii)  $\{K_s^{t,\zeta;v}\}$  is increasing and continuous,  $K_t^{t,\zeta;v} = 0$ , and

$$\int_t^T (Y_s^{t,\zeta;v} - h(s, X_s^{t,\zeta;v})) dK_s^{t,\zeta;v} = 0.$$

Moreover, we get the following estimates for the solution of (3.5) from Propositions 2.1 and 2.2.

PROPOSITION 3.3.

$$(3.6) \quad \mathbb{E}^{\mathcal{F}_t} \left\{ \sup_{t \leq s \leq T} |Y_s^{t,\zeta;v}|^2 + \int_t^T |Z_s^{t,\zeta;v}|^2 ds + |K_T^{t,\zeta;v}|^2 \right\} \leq C(1 + |\zeta|^2).$$

PROPOSITION 3.4.

$$\begin{aligned} & \mathbb{E}^{\mathcal{F}_t} \left\{ \sup_{t \leq s \leq T} |Y_s^{t,\zeta;v} - Y_s^{t,\zeta';v'}|^2 + \int_t^T |Z_s^{t,\zeta;v} - Z_s^{t,\zeta';v'}|^2 ds \right. \\ & \quad \left. + |K_T^{t,\zeta;v} - K_T^{t,\zeta';v'}|^2 \right\} \\ (3.7) \quad & \leq C|\zeta - \zeta'|^2 + C\mathbb{E}^{\mathcal{F}_t} \left\{ \int_t^T |v_s - v'_s|^2 ds \right\} \\ & \quad + C(1 + |\zeta| + |\zeta'|) \left( |\zeta - \zeta'|^2 + \mathbb{E}^{\mathcal{F}_t} \left\{ \int_t^T |v_s - v'_s|^2 ds \right\} \right)^{1/2}. \end{aligned}$$

Given a control process  $v(\cdot) \in \mathcal{U}$ , we introduce the associated cost functional

$$(3.8) \quad J(t, x; v(\cdot)) := Y_s^{t,x;v} \Big|_{s=t}, \quad (t, x) \in [0, T] \times \mathbb{R}^n,$$

and we define the value function of the stochastic optimal control problem

$$(3.9) \quad u(t, x) := \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} J(t, x; v(\cdot)), \quad (t, x) \in [0, T] \times \mathbb{R}^n.$$

*Remark 3.5.* This is one kind of stochastic recursive optimal control problem with the obstacle constraint for the cost functional  $Y_s^{t,x;v} \geq h(s, X_s^{t,x;v})$ ,  $t \leq s \leq T$ . In the financial market, if  $X_s^{t,x;v}$  represents the wealth of the investor and  $Y_s^{t,x;v}$  represents the recursive utility cost functional, then the constraint means that the investor requires his cost functional value to be bigger than a given function of his wealth at any time. For this, an additional increasing process is introduced to push  $Y_s^{t,x;v}$  upwards in a kind of minimal way. In addition, for an investment policy the investor may choose the terminal time (as a stopping time) up to which he maximizes his wealth. We precise this point as follows.

For any  $(t, x) \in [0, T] \times \mathbb{R}^n$  and a given admissible control  $v(\cdot) \in \mathcal{U}$ , let  $(Y^{t,x;v}, Z^{t,x;v}, K^{t,x;v})$  be the solution of the reflected BSDE (3.5). From Proposition 2.3 in [9], we know that, for each  $s \in [t, T]$ ,

$$Y_s^{t,x;v} = \operatorname{esssup}_{\tau \in \mathcal{T}_s} \mathbb{E}^{\mathcal{F}_s} \left\{ \int_s^\tau g(r, X_r^{t,x;v}, Y_r^{t,x;v}, Z_r^{t,x;v}, v_r) dr \right. \\ \left. + h(\tau, X_\tau^{t,x;v}) \mathbf{1}_{\{\tau < T\}} + \Phi(X_T^{t,x;v}) \mathbf{1}_{\{\tau = T\}} \right\},$$

where  $\mathcal{T}$  is the set of all stopping times dominated by  $T$ , and  $\mathcal{T}_s = \{\tau \in \mathcal{T}; s \leq \tau \leq T\}$ . Furthermore, the stopping time

$$(3.10) \quad D_s^{t,x;v} = \inf\{s \leq r \leq T; Y_r^{t,x;v} = h(r, X_r^{t,x;v})\}$$

is optimal.

Here  $g$  is a function of  $(s, X, Y, Z, v)$  satisfying assumptions (H3.3) and (H3.4); we will give some special examples to illustrate some applications for this recursive optimization problem.

*Example 1.* If  $g$  is independent on  $(y, z)$ , then our problem is a mixture of an optimal stopping time problem and a “classical” optimal stochastic control problem. The value function is

$$u(t, x) := \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} Y_t^{t,x;v} = \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} \operatorname{esssup}_{\tau \in \mathcal{T}_t} \mathbb{E}^{\mathcal{F}_t} \left\{ \int_t^\tau g(r, X_r^{t,x;v}, v_r) dr \right. \\ \left. + h(\tau, X_\tau^{t,x;v}) \mathbf{1}_{\{\tau < T\}} + \Phi(X_T^{t,x;v}) \mathbf{1}_{\{\tau = T\}} \right\}.$$

Furthermore, the stopping time  $D_t^{t,x;v}$  defined by (3.10) is optimal. The details about this mixed optimal control problem can be found in [7] and [8].

*Example 2.* Let  $(\beta_t, \gamma_t)$  be a bounded  $(\mathbb{R}, \mathbb{R}^d)$ -valued predictable continuous vector process, and  $\varphi(t, x, v)$  be a  $\mathbb{R}$ -valued continuous function which is Lipschitz continuous in  $(x, v)$  and, for each  $(x, v)$ ,  $\varphi(\cdot, x, v)$  belongs to  $H^2$ . Let

$$g(t, x, y, z, v) = \varphi(t, x, v) + \beta_t y + \langle \gamma, z \rangle.$$

Let  $\{\Gamma_{t,s}, t \leq s \leq T\}$  be the adjoint process satisfying the linear SDE:

$$d\Gamma_{t,s} = \Gamma_{t,s}[\beta_s ds + \langle \gamma_s, dW_s \rangle], \quad \Gamma_{t,t} = 1.$$

Then the value function is

$$u(t, x) := \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} Y_t^{t, x; v} = \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} \operatorname{esssup}_{\tau \in \mathcal{T}_t} \mathbb{E}^{\mathcal{F}_t} \left\{ \int_t^\tau \Gamma_{t, r} \varphi(r, X_r^{t, x; v}, v_r) dr \right. \\ \left. + \Gamma_{t, \tau} h(\tau, X_\tau^{t, x; v}) \mathbf{1}_{\{\tau < T\}} + \Gamma_{t, T} \Phi(X_T^{t, x; v}) \mathbf{1}_{\{\tau = T\}} \right\}.$$

Furthermore, the stopping time  $D_t^{t, x; v}$  defined by (3.10) is optimal. This is also a mixed optimal control problem.

*Example 3.* This involves hedging the American option when loan interest is higher than deposit interest.

In [10], El Karoui, Pardoux, and Quenez proved that the price of an American option corresponds to the solution of a reflected BSDE. Now we will show that under some requirements in a financial market, such as when loan interest is higher than deposit interest, the price of an American option corresponds to the value function of a recursive optimal control problem with the obstacle constraints described by the solution of reflected BSDE.

We take the same notations as those in section 3 of [11] and in section 5 of [10], all variables are 1-dimensional for simplification. We suppose that the investor is allowed to borrow money at time  $t$  at an interest rate  $R_t > r_t$ , where  $r_t$  the bond rate. Then, from the results in [11] and [10], the wealth of the investor satisfies

$$-dX_t = b(t, X_t, Z_t)dt - Z_t dW_t, \quad 0 \leq t \leq T,$$

$$b(t, X_t, Z_t) := - \left[ r_t X_t + \theta_t Z_t - (R_t - r_t) \left( X_t - \frac{Z_t}{\sigma_t} \right)^- \right],$$

where  $Z_t := \sigma_t \pi_t$ ,  $\theta_t := \sigma_t^{-1}(b_t - r_t)$ . Also,  $b_t$  represents the instantaneous expected return rate in stock;  $\sigma_t$  implies that the instantaneous volatility of the stock is invertible;  $b_t$ ,  $r_t$ ,  $R_t$ , and  $\sigma_t$  are all deterministic bounded functions; and  $\sigma_t^{-1}$  is also bounded.

We consider the problem of pricing an American contingent claim at each time  $t$ , which consists of the selection of a stopping time  $\tau \in \mathcal{T}_t$  and a payoff  $S_\tau$  on exercise if  $\tau < T$  and  $\xi$  if  $\tau = T$ . Here  $\{S_t, 0 \leq t \leq T\}$  satisfies assumptions (H3.3)–(H3.4). Set

$$\tilde{S}_s = \xi \mathbf{1}_{\{\tau = T\}} + S_s \mathbf{1}_{\{\tau < T\}}, \quad 0 \leq s \leq T,$$

then the price of American contingent claim  $(\tilde{S}_s, 0 \leq s \leq T)$  at time  $t$  is given by

$$X_t = \operatorname{esssup}_{\tau \in \mathcal{T}_t} X_t(\tau, \tilde{S}_\tau),$$

where  $X_t(\tau, \tilde{S}_\tau)$ , noted by  $X_t^\tau$ , satisfies BSDE:

$$\begin{cases} -dX_s^\tau = b(s, X_s^\tau, Z_s^\tau)ds - Z_s^\tau dW_s, & 0 \leq s \leq \tau, \\ X_\tau^\tau = \tilde{S}_\tau. \end{cases}$$

For each  $(\omega, t)$ ,  $b(t, x, z)$  is a convex function of  $(x, z)$ . It follows from Proposition 3.6 and (3.22) in [10] (pp. 40–41) that we have

$$X_t^\tau = \operatorname{esssup}_{r_t \leq \beta_t \leq R_t} X_t^{\beta, \tau},$$

where  $X_t^{\beta, \tau}$  satisfies

$$\begin{cases} -dX_s^{\beta, \tau} = b^\beta(s, X_s^{\beta, \tau}, Z_s^{\beta, \tau})ds - Z_s^{\beta, \tau}dW_s, & 0 \leq s \leq \tau, \\ X_\tau^{\beta, \tau} = \tilde{S}_\tau, \end{cases}$$

$$b^\beta(t, X_t, Z_t) := -\beta_t X_t - \left[ \theta_t + \frac{r_t - \beta_t}{\sigma_t} \right] Z_t,$$

and  $\beta_t$  is a bounded  $\mathbb{R}$ -valued adapted process which can be regarded as an interest rate process in finance. So,

$$\begin{aligned} X_t &:= \operatorname{esssup}_{\tau \in \mathcal{T}_t} X_t(\tau, \tilde{S}_\tau) \\ &= \operatorname{esssup}_{\tau \in \mathcal{T}_t} \operatorname{esssup}_{r_t \leq \beta_t \leq R_t} X_t^{\beta, \tau} = \operatorname{esssup}_{r_t \leq \beta_t \leq R_t} \operatorname{esssup}_{\tau \in \mathcal{T}_t} X_t^{\beta, \tau} \\ &= \operatorname{esssup}_{r_t \leq \beta_t \leq R_t} X_t^\beta, \end{aligned}$$

where  $X_t^\beta := \operatorname{esssup}_{\tau \in \mathcal{T}_t} X_t^{\beta, \tau}$ . Then from Proposition 5.1 in [10], there exist  $Z_s^\beta \in H^2$  and  $K_s^\beta$ , which is an increasing adapted continuous process with  $K_0 = 0$ , such that  $(X_s^\beta, Z_s^\beta, K_s^\beta)$  satisfies the following reflected BSDE:

$$\begin{cases} -dX_s^\beta = b^\beta(s, X_s^\beta, Z_s^\beta)ds + dK_s^\beta - Z_s^\beta dW_s, \\ X_T^\beta = \xi, \quad 0 \leq s \leq T, \end{cases}$$

with  $X_t^\beta \geq S_t$ ,  $0 \leq t \leq T$ , and  $\int_0^T (X_t^\beta - S_t) dK_t^\beta = 0$ . Here process  $K_t^\beta$  may be interpreted as a cumulative consumption process. Furthermore, the stopping time

$$D_t^\beta = \inf\{t \leq s \leq T; X_s^\beta = S_s\}$$

is optimal. Then we formulate the pricing problem of an American option to the stochastic recursive optimal control problem with the obstacle constraint which is studied in our paper.

The above example provides the practical background for our optimal control problem (3.9). Now we continue to prove that the celebrated dynamic programming principle still holds for this optimization problem. Some proof ideas come from the proof of the dynamic programming principle for recursive problem given by Peng in the Chinese version [14]. However, we generalize his conclusion to the obstacle constraint case and improve proof method which becomes more readable.

For each  $t > 0$ , we denote by  $\{\mathcal{F}_s^t, t \leq s \leq T\}$  the natural filtration of the Brownian motion  $\{W_s - W_t, t \leq s \leq T\}$ , augmented by the P-null sets of  $\mathcal{F}$ . Also, we introduce the following subspaces of  $\mathcal{U}$ :

$$\begin{aligned} \mathcal{U}^t &:= \left\{ v(\cdot) \in \mathcal{U} \mid v(s) \text{ is } \{\mathcal{F}_s^t\} \text{ progressively measurable } \forall t \leq s \leq T \right\}, \\ \bar{\mathcal{U}}^t &:= \left\{ v_s = \sum_{j=1}^N v_s^j 1_{A_j} \mid v_s^j \in \mathcal{U}^t, \{A_j\}_{j=1}^N \text{ is a partition of } (\Omega, \mathcal{F}_t) \right\}. \end{aligned}$$

Now we will prove the following proposition.

PROPOSITION 3.6. *Under the assumptions (H3.1)–(H3.4), the value function  $u(t, x)$  defined in (3.9) is a deterministic function.*

*Proof.* First, we will prove

$$(3.11) \quad \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} J(t, x; v(\cdot)) = \operatorname{esssup}_{v(\cdot) \in \bar{\mathcal{U}}^t} J(t, x; v(\cdot)).$$

Obviously, from the fact that  $\bar{\mathcal{U}}^t$  is a subset of  $\mathcal{U}$ , we get

$$\operatorname{esssup}_{v(\cdot) \in \mathcal{U}} J(t, x; v(\cdot)) \geq \operatorname{esssup}_{v(\cdot) \in \bar{\mathcal{U}}^t} J(t, x; v(\cdot)).$$

We need to consider the inverse inequality. For any  $v(\cdot), \tilde{v}(\cdot) \in \mathcal{U}$ , from Proposition 3.4 we know

$$\mathbb{E} \left\{ \left| Y_t^{t,x;v} - Y_t^{t,x;\tilde{v}} \right|^2 \right\} \leq C \mathbb{E} \int_t^T |v_s - \tilde{v}_s|^2 ds + C(1 + |x|) \left( \mathbb{E} \int_t^T |v_s - \tilde{v}_s|^2 ds \right)^{1/2}.$$

Note that  $\bar{\mathcal{U}}^t$  is dense in  $\mathcal{U}$ , then, for each  $v(\cdot) \in \mathcal{U}$ , there exists a sequence  $\{v_n(\cdot)\}_{n=1}^\infty \in \bar{\mathcal{U}}^t$  such that

$$\lim_{n \rightarrow \infty} \mathbb{E} \left\{ \left| Y_t^{t,x;v_n} - Y_t^{t,x;v} \right|^2 \right\} = 0.$$

So there exists a subsequence, we denote without loss of generality  $\{v_n(\cdot)\}_{n=1}^\infty$ , such that

$$\lim_{n \rightarrow \infty} Y_t^{t,x;v_n} = Y_t^{t,x;v} \quad a.s..$$

From the definition in (3.8),

$$\lim_{n \rightarrow \infty} J(t, x; v_n(\cdot)) = J(t, x; v(\cdot)) \quad a.s..$$

By the arbitrariness of  $v(\cdot)$  and the definition of essential supremum, we get

$$\operatorname{esssup}_{v(\cdot) \in \bar{\mathcal{U}}^t} J(t, x; v(\cdot)) \geq \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} J(t, x; v(\cdot)).$$

Then we obtain (3.11).

Second, we want to prove

$$(3.12) \quad \operatorname{esssup}_{v(\cdot) \in \bar{\mathcal{U}}^t} J(t, x; v(\cdot)) = \operatorname{esssup}_{v(\cdot) \in \mathcal{U}^t} J(t, x; v(\cdot)).$$

Obviously,

$$\operatorname{esssup}_{v(\cdot) \in \bar{\mathcal{U}}^t} J(t, x; v(\cdot)) \geq \operatorname{esssup}_{v(\cdot) \in \mathcal{U}^t} J(t, x; v(\cdot)).$$

In order to get the inverse inequality, we need Lemma 3.7. The main idea of the lemma is to consider the partition of probability space, which is first introduced by Theorem 4.7 in [14]. The proof of Lemma 3.7 is essentially the same as that in [14], hence we omit it.

LEMMA 3.7.

$$\begin{aligned} X^{t,x;\sum_{j=1}^N v^j 1_{A_j}} &= \sum_{j=1}^N 1_{A_j} X^{t,x;v^j}, & Y^{t,x;\sum_{j=1}^N v^j 1_{A_j}} &= \sum_{j=1}^N 1_{A_j} Y^{t,x;v^j}, \\ Z^{t,x;\sum_{j=1}^N v^j 1_{A_j}} &= \sum_{j=1}^N 1_{A_j} Z^{t,x;v^j}, & K^{t,x;\sum_{j=1}^N v^j 1_{A_j}} &= \sum_{j=1}^N 1_{A_j} K^{t,x;v^j}. \end{aligned}$$

For all  $v(\cdot) \in \bar{\mathcal{U}}^t$ , we have

$$J(t, x; v(\cdot)) = J\left(t, x; \sum_{j=1}^N v^j(\cdot) 1_{A_j}\right) = \sum_{j=1}^N 1_{A_j} J(t, x; v^j(\cdot)).$$

Note that  $v^j(\cdot)$  ( $j = 1, 2, \dots, N$ ) are  $\{\mathcal{F}_s^t\}$  progressively measurable, then  $J(t, x; v^j(\cdot))$  ( $j = 1, 2, \dots, N$ ) are deterministic. Without loss of generality, we assume that

$$J(t, x; v^1(\cdot)) \geq J(t, x; v^j(\cdot)) \quad \forall j = 2, 3, \dots, N.$$

So

$$J(t, x; v(\cdot)) \leq J(t, x; v^1(\cdot)) \leq \operatorname{esssup}_{v(\cdot) \in \mathcal{U}^t} J(t, x; v(\cdot)).$$

Since  $v(\cdot)$  is arbitrary, we get

$$\operatorname{esssup}_{v(\cdot) \in \bar{\mathcal{U}}^t} J(t, x; v(\cdot)) \leq \operatorname{esssup}_{v(\cdot) \in \mathcal{U}^t} J(t, x; v(\cdot)),$$

then obtain (3.12).

However, when  $v(\cdot) \in \mathcal{U}^t$ , the cost functional  $J(t, x; v(\cdot))$  is deterministic; hence

$$u(t, x) = \sup_{v(\cdot) \in \mathcal{U}^t} J(t, x; v(\cdot))$$

is deterministic and the proof is completed.  $\square$

Next we will discuss the continuity of value function  $u(t, x)$  with respect to  $x$ . We have the following estimates:

LEMMA 3.8. *For each  $t \in [0, T]$ ,  $x$  and  $x' \in \mathbb{R}^n$ , we have*

- (i)  $|u(t, x) - u(t, x')|^2 \leq C|x - x'|^2 + C(1 + |x| + |x'|)|x - x'|$ ,
- (ii)  $|u(t, x)| \leq C(1 + |x|)$ .

*Proof.* In order to prove this lemma, Propositions 2.1 and 2.2 as well as Propositions 3.3 and 3.4, play important roles.

From Propositions 3.3 and 3.4, for each admissible control  $v(\cdot) \in \mathcal{U}$ , we have

$$(3.13) \quad |J(t, x; v(\cdot))| \leq C(1 + |x|),$$

$$(3.14) \quad |J(t, x; v(\cdot)) - J(t, x'; v(\cdot))|^2 \leq C|x - x'|^2 + C(1 + |x| + |x'|)|x - x'|.$$

On the other hand, for each  $\varepsilon > 0$ , there exist  $v(\cdot)$  and  $v'(\cdot) \in \mathcal{U}$  such that

$$J(t, x; v'(\cdot)) \leq u(t, x) \leq J(t, x; v(\cdot)) + \varepsilon, \quad J(t, x'; v(\cdot)) \leq u(t, x') \leq J(t, x'; v'(\cdot)) + \varepsilon.$$

Then from the estimate (3.13) of  $J$ , we get

$$-C(1 + |x|) \leq J(t, x; v'(\cdot)) \leq u(t, x) \leq J(t, x; v(\cdot)) + \varepsilon \leq C(1 + |x|) + \varepsilon.$$

From the arbitrariness of  $\varepsilon$ , we obtain (ii). Similarly,

$$J(t, x; v'(\cdot)) - J(t, x'; v'(\cdot)) - \varepsilon \leq u(t, x) - u(t, x') \leq J(t, x; v(\cdot)) - J(t, x'; v(\cdot)) + \varepsilon,$$

$$\begin{aligned} & |u(t, x) - u(t, x')| \\ & \leq \max \left\{ |J(t, x; v(\cdot)) - J(t, x'; v(\cdot))|, |J(t, x; v'(\cdot)) - J(t, x'; v'(\cdot))| \right\} + \varepsilon, \\ & |u(t, x) - u(t, x')|^2 \\ & \leq 2 \max \left\{ |J(t, x; v(\cdot)) - J(t, x'; v(\cdot))|^2, |J(t, x; v'(\cdot)) - J(t, x'; v'(\cdot))|^2 \right\} + 2\varepsilon^2 \\ & \leq 2C|x - x'|^2 + 2C(1 + |x| + |x'|)|x - x'| + 2\varepsilon^2. \end{aligned}$$

Then we can obtain (i).  $\square$

We also have

LEMMA 3.9. *For all  $t \in [0, T]$ , for all  $v(\cdot) \in \mathcal{U}$ , and for all  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , we have*

$$J(t, \zeta; v(\cdot)) = Y_t^{t, \zeta; v}.$$

*Proof.* We first study a simple case.  $\zeta$  is in the following form:  $\zeta = \sum_{i=1}^N 1_{A_i} x_i$ , where  $\{A_i\}_{i=1}^N$  is a finite partition of  $(\Omega, \mathcal{F}_t)$ , and  $x_i \in \mathbb{R}^n$ , for  $1 \leq i \leq N$ . The similar argument in Lemma 3.7 leads to

$$Y_s^{t, \zeta; v} = Y_s^{t, \sum_{i=1}^N 1_{A_i} x_i; v} = \sum_{i=1}^N 1_{A_i} Y_s^{t, x_i; v}, \quad s \in [t, T].$$

From the definition of cost functional (3.8), we deduce that

$$Y_t^{t, \zeta; v} = \sum_{i=1}^N 1_{A_i} Y_t^{t, x_i; v} = \sum_{i=1}^N 1_{A_i} J(t, x_i; v(\cdot)) = J\left(t, \sum_{i=1}^N 1_{A_i} x_i; v(\cdot)\right) = J(t, \zeta; v(\cdot)).$$

Therefore, for simple functions, we get the desired result.

Given a general  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , we can choose a sequence of simple functions  $\{\zeta_i\}$  which converges to  $\zeta$  in  $L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ . Consequently, from Proposition 3.4 and

estimate (3.14), we have

$$\begin{aligned}
& \mathbb{E} \left\{ \left| Y_t^{t, \zeta; v} - Y_t^{t, \zeta_i; v} \right|^2 \right\} \\
& \leq \mathbb{E} \left\{ C |\zeta - \zeta_i|^2 + C(1 + |\zeta| + |\zeta_i|) |\zeta - \zeta_i| \right\} \\
& \leq C \mathbb{E} \left\{ |\zeta - \zeta_i|^2 \right\} + C \left( \mathbb{E} \left\{ (1 + |\zeta| + |\zeta_i|)^2 \right\} \right)^{1/2} \left( \mathbb{E} \left\{ |\zeta - \zeta_i|^2 \right\} \right)^{1/2} \\
& \rightarrow 0 \quad \text{as } i \rightarrow \infty, \\
& \mathbb{E} \left\{ |J(t, \zeta; v(\cdot)) - J(t, \zeta_i; v(\cdot))|^2 \right\} \\
& \leq \mathbb{E} \left\{ C |\zeta - \zeta_i|^2 + C(1 + |\zeta| + |\zeta_i|) |\zeta - \zeta_i| \right\} \\
& \leq C \mathbb{E} \left\{ |\zeta - \zeta_i|^2 \right\} + C \left( \mathbb{E} \left\{ (1 + |\zeta| + |\zeta_i|)^2 \right\} \right)^{1/2} \left( \mathbb{E} \left\{ |\zeta - \zeta_i|^2 \right\} \right)^{1/2} \\
& \rightarrow 0 \quad \text{as } i \rightarrow \infty.
\end{aligned}$$

With the help of  $Y_t^{t, \zeta_i; v} = J(t, \zeta_i; v(\cdot))$ , the proof is completed.  $\square$

For the value function of our recursive optimal control problem, we have the following lemma.

LEMMA 3.10. *Fixed  $t \in [0, T)$  and  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , for each  $v(\cdot) \in \mathcal{U}$ , we have*

$$(3.15) \quad u(t, \zeta) \geq Y_t^{t, \zeta; v}.$$

*On the other hand, for each  $\varepsilon > 0$ , there exists an admissible control  $v(\cdot) \in \mathcal{U}$  such that*

$$(3.16) \quad u(t, \zeta) \leq Y_t^{t, \zeta; v} + \varepsilon, \quad a.s..$$

*Proof.* We first prove (3.15). When  $\zeta$  is a simple function:  $\zeta = \sum_{i=1}^N 1_{A_i} x_i$ , for all  $v(\cdot) \in \mathcal{U}$ , we have

$$Y_t^{t, \zeta; v} = Y_t^{t, \sum_{i=1}^N 1_{A_i} x_i; v} = \sum_{i=1}^N 1_{A_i} Y_t^{t, x_i; v} \leq \sum_{i=1}^N 1_{A_i} u(t, x_i) = u(t, \zeta).$$

If  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , we can choose a sequence of simple functions  $\{\zeta_i\}$  which converges to  $\zeta$  in  $L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ . Consequently, similar to Lemma 3.9, we have

$$\mathbb{E} \left\{ \left| Y_t^{t, \zeta; v} - Y_t^{t, \zeta_i; v} \right|^2 \right\} \rightarrow 0; \quad \mathbb{E} \left\{ |u(t, \zeta) - u(t, \zeta_i)|^2 \right\} \rightarrow 0.$$

Then there exists a subsequence, without loss of generality we use same notation, such that

$$\lim_{i \rightarrow \infty} Y_t^{t, \zeta_i; v} = Y_t^{t, \zeta; v}, \quad a.s., \quad \lim_{i \rightarrow \infty} u(t, \zeta_i) = u(t, \zeta), \quad a.s..$$

Here  $Y_t^{t, \zeta_i; v} \leq u(t, \zeta_i)$ ,  $i = 1, 2, \dots$ , so  $Y_t^{t, \zeta; v} \leq u(t, \zeta)$ .

We now turn to (3.16). We first consider the case that  $\zeta$  is a bounded random variable:  $\zeta \in L^\infty(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ . We suppose that  $|\zeta| \leq M$ , and construct a simple



random variable  $\eta \in L^\infty(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ ,  $\eta = \sum_{i=1}^N 1_{A_i} x_i$  such that

- (i)  $|\eta| \leq |\zeta|$ ,
- (ii)  $|\eta - \zeta| \leq \min \left\{ \frac{\varepsilon}{6\sqrt{C}}, \frac{\varepsilon^2}{36C(1+2M)} \right\}$ .

For any  $v(\cdot) \in \mathcal{U}$ , we have

$$\left| Y_t^{t, \zeta; v} - Y_t^{t, \eta; v} \right| \leq \frac{\varepsilon}{3}, \quad |u(t, \zeta) - u(t, \eta)| \leq \frac{\varepsilon}{3}.$$

Then, for each  $x_i$ , we can choose an  $\{\mathcal{F}_s^t\}$ -adapted admissible control  $v^i(\cdot)$  such that

$$u(t, x_i) \leq Y_t^{t, x_i; v^i} + \frac{\varepsilon}{3}.$$

We denote  $v(\cdot) := \sum_{i=1}^N 1_{A_i} v^i(\cdot)$ , then

$$\begin{aligned} Y_t^{t, \zeta; v} &\geq - \left| Y_t^{t, \zeta; v} - Y_t^{t, \eta; v} \right| + Y_t^{t, \eta; v} \geq -\frac{\varepsilon}{3} + \sum_{i=1}^N 1_{A_i} Y_t^{t, x_i; v^i} \\ &\geq -\frac{\varepsilon}{3} + \sum_{i=1}^N 1_{A_i} \left( u(t, x_i) - \frac{\varepsilon}{3} \right) = -\frac{2}{3}\varepsilon + u(t, \eta) \\ &\geq -\varepsilon + u(t, \zeta). \end{aligned}$$

Therefore, for  $\zeta \in L^\infty(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , we have the desired result (3.16).

Given a general  $\zeta \in L^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ , we note that  $\zeta$  has the following form:

$$\zeta = \sum_{i=1}^{\infty} 1_{A_i} \zeta_i,$$

where  $\{A_i\}_{i=1}^{\infty}$  is a partition of  $(\Omega, \mathcal{F}_t)$ ,  $x_i \in \mathbb{R}^n$  ( $i = 1, 2, \dots$ ),  $|\zeta_i| \leq i$ , and  $\zeta_i \in L^\infty(\Omega, \mathcal{F}_t, P; \mathbb{R}^n)$ . So, for every  $\zeta_i$ , there exists  $v^i(\cdot) \in \mathcal{U}$ , such that

$$u(t, \zeta_i) \leq Y_t^{t, \zeta_i; v^i} + \varepsilon.$$

We denote  $v(\cdot) = \sum_{i=1}^{\infty} 1_{A_i} v^i(\cdot)$  and get

$$\begin{aligned} u(t, \zeta) &= u\left(t, \sum_{i=1}^{\infty} 1_{A_i} \zeta_i\right) = \sum_{i=1}^{\infty} 1_{A_i} u(t, \zeta_i) \leq \sum_{i=1}^{\infty} 1_{A_i} (Y_t^{t, \zeta_i; v^i} + \varepsilon) \\ &= \sum_{i=1}^{\infty} 1_{A_i} Y_t^{t, \zeta_i; v^i} + \varepsilon = Y_t^{t, \zeta; v} + \varepsilon. \end{aligned}$$

The proof is now completed.  $\square$

Now we start to discuss the (generalized) dynamic programming principle for our recursive optimal control problem (3.9).

First, we introduce a family of (backward) semigroups which is original from Peng's idea in [14].

Given the initial condition  $(t, x)$ , an admissible control  $v(\cdot) \in \mathcal{U}$ , a positive number  $\delta \leq T - t$ , and a real-valued random variable  $\eta \in L^2(\Omega, \mathcal{F}_{t+\delta}, P; \mathbb{R})$ , we denote

$$G_{t, t+\delta}^{t, x; v}[\eta] := Y_t,$$

where  $(Y_s, Z_s, K_s)_{t \leq s \leq t+\delta}$  is the solution of the following reflected BSDE with time horizon  $t + \delta$ :

$$\begin{aligned} Y_s &= \eta + \int_s^{t+\delta} g(r, X_r^{t,x;v}, Y_r, Z_r, v_r) dr + K_{t+\delta} - K_s \\ &\quad - \int_s^{t+\delta} Z_r dW_r, \quad t \leq s \leq t + \delta, \end{aligned}$$

satisfying

- (i)  $Y \in S^2$ ,  $Z \in H^2$  and  $K_{t+\delta} \in L^2$ ,
- (ii)  $Y_s \geq h(s, X_s^{t,x;v})$ ,  $t \leq s \leq t + \delta$ ,
- (iii)  $\{K_s\}$  is increasing and continuous,  $K_t = 0$ ,  $\int_t^{t+\delta} (Y_s - h(s, X_s^{t,x;v})) dK_s = 0$ .

Obviously,

$$G_{t,T}^{t,x;v}[\Phi(X_T^{t,x;v})] = G_{t,t+\delta}^{t,x;v}[Y_{t+\delta}^{t,x;v}].$$

Then our (generalized) dynamic programming principle holds.

**THEOREM 3.11.** *Under the assumptions (H3.1)–(H3.4), the value function  $u(t, x)$  obeys the following dynamic programming principle: For each  $0 < \delta \leq T - t$ ,*

$$(3.17) \quad u(t, x) = \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} G_{t,t+\delta}^{t,x;v}[u(t + \delta, X_{t+\delta}^{t,x;v})].$$

*Proof.* We have

$$\begin{aligned} u(t, x) &= \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} G_{t,T}^{t,x;v}[\Phi(X_T^{t,x;v})] = \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} G_{t,t+\delta}^{t,x;v}[Y_{t+\delta}^{t,x;v}] \\ &= \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} G_{t,t+\delta}^{t,x;v} \left[ Y_{t+\delta}^{t+\delta, X_{t+\delta}^{t,x;v};v} \right]. \end{aligned}$$

From Lemma 3.10 and the comparison theorem of reflected BSDE (Theorem 4.1 in [9]),

$$u(t, x) \leq \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} G_{t,t+\delta}^{t,x;v}[u(t + \delta, X_{t+\delta}^{t,x;v})].$$

On the other hand, from Lemma 3.10, for every  $\varepsilon > 0$ , we can find an admissible control  $\bar{v}(\cdot) \in \mathcal{U}$  such that

$$u(t + \delta, X_{t+\delta}^{t,x;v}) \leq Y_{t+\delta}^{t+\delta, X_{t+\delta}^{t,x;v};\bar{v}} + \varepsilon.$$

For each  $v(\cdot) \in \mathcal{U}$ , we denote  $\tilde{v}(s) = 1_{\{s \leq t+\delta\}} v(s) + 1_{\{s > t+\delta\}} \bar{v}(s)$ . From the above inequality and the comparison theorem, we get

$$Y_{t+\delta}^{t+\delta, X_{t+\delta}^{t,x;\tilde{v}};\tilde{v}} \geq u(t + \delta, X_{t+\delta}^{t,x;\tilde{v}}) - \varepsilon, \quad u(t, x) \geq \operatorname{esssup}_{\tilde{v}(\cdot) \in \mathcal{U}_{\bar{v}}} G_{t,t+\delta}^{t,x;\tilde{v}}[u(t + \delta, X_{t+\delta}^{t,x;\tilde{v}}) - \varepsilon],$$

and here

$$\mathcal{U}_{\bar{v}} := \{\tilde{v}(\cdot) \in \mathcal{U}; \tilde{v}(s) = 1_{\{s \leq t+\delta\}} v(s) + 1_{\{s > t+\delta\}} \bar{v}(s) \text{ for some } v(\cdot) \in \mathcal{U}\}.$$

By Proposition 2.2, there exists a positive constant  $C_0$  such that

$$u(t, x) \geq \operatorname{esssup}_{\tilde{v}(\cdot) \in \mathcal{U}_{\tilde{v}}} G_{t, t+\delta}^{t, x; \tilde{v}}[u(t + \delta, X_{t+\delta}^{t, x; \tilde{v}})] - C_0 \varepsilon.$$

Therefore, letting  $\varepsilon \downarrow 0$ , we obtain

$$u(t, x) \geq \operatorname{esssup}_{\tilde{v}(\cdot) \in \mathcal{U}_{\tilde{v}}} G_{t, t+\delta}^{t, x; \tilde{v}}[u(t + \delta, X_{t+\delta}^{t, x; \tilde{v}})].$$

Because  $\tilde{v}(\cdot)$  acts only on  $[t, t + \delta]$  for  $G_{t, t+\delta}^{t, x; \tilde{v}}$ , from the definition of  $\tilde{v}(\cdot)$  and the arbitrariness of  $v(\cdot) \in \mathcal{U}$ , we know that the above inequality can be written as

$$u(t, x) \geq \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} G_{t, t+\delta}^{t, x; v}[u(t + \delta, X_{t+\delta}^{t, x; v})],$$

which is our desired conclusion.  $\square$

At the end of this section, we devote ourselves to the continuity of  $u(t, x)$  with respect to  $t$ .

**PROPOSITION 3.12.** *Under the Assumption (H3.1)–(H3.4), the value function  $u(t, x)$  defined by (3.9) is continuous in  $t$ .*

*Proof.* We define  $Y_s^{t, x; v}$  for all  $s \in [0, T]$  by choosing  $Y_s^{t, x; v} \equiv Y_t^{t, x; v}$  for  $0 \leq s \leq t$ . And we define the “obstacle”

$$S_s^{t, x; v} = \begin{cases} h(s, X_s^{t, x; v}), & t \leq s \leq T, \\ h(t, x), & 0 \leq s \leq t. \end{cases}$$

For fixed  $x \in \mathbb{R}^n$  for all  $0 \leq t_1 \leq t_2 \leq T$ , we analyze the difference between  $u(t_1, x)$  and  $u(t_2, x)$ . For all  $\varepsilon > 0$ , there exist  $v_1(\cdot) \in \mathcal{U}$ ,  $v_2(\cdot) \in \mathcal{U}$ , such that

$$Y_{t_1}^{t_1, x; v_2} \leq u(t_1, x) \leq Y_{t_1}^{t_1, x; v_1} + \varepsilon, \quad Y_{t_2}^{t_2, x; v_1} \leq u(t_2, x) \leq Y_{t_2}^{t_2, x; v_2} + \varepsilon.$$

Then

$$Y_{t_1}^{t_1, x; v_2} - Y_{t_2}^{t_2, x; v_2} - \varepsilon \leq u(t_1, x) - u(t_2, x) \leq Y_{t_1}^{t_1, x; v_1} - Y_{t_2}^{t_2, x; v_1} + \varepsilon,$$

$$|u(t_1, x) - u(t_2, x)| \leq \max \{ |Y_{t_1}^{t_1, x; v_1} - Y_{t_2}^{t_2, x; v_1}|, |Y_{t_1}^{t_1, x; v_2} - Y_{t_2}^{t_2, x; v_2}| \} + \varepsilon.$$

Here we only estimate  $|Y_{t_1}^{t_1, x; v_1} - Y_{t_2}^{t_2, x; v_1}|$  since it is same as estimating  $|Y_{t_1}^{t_1, x; v_2} - Y_{t_2}^{t_2, x; v_2}|$ . From Proposition 2.2, we have

$$\begin{aligned} & |Y_{t_1}^{t_1, x; v_1} - Y_{t_2}^{t_2, x; v_1}|^2 = |Y_0^{t_1, x; v_1} - Y_0^{t_2, x; v_1}|^2 \\ & \leq \mathbb{E} \left\{ \sup_{0 \leq s \leq T} |Y_s^{t_1, x; v_1} - Y_s^{t_2, x; v_1}|^2 \right\} \\ & \leq C \mathbb{E} \{ |\Phi(X_T^{t_1, x; v_1}) - \Phi(X_T^{t_2, x; v_1})|^2 \} \\ (3.18) \quad & + C \mathbb{E} \left\{ \left( \int_0^T |1_{[t_1, T]} g(s, X_s^{t_1, x; v_1}, Y_s^{t_1, x; v_1}, Z_s^{t_1, x; v_1}, v_1(s)) \right. \right. \\ & \quad \left. \left. - 1_{[t_2, T]} g(s, X_s^{t_2, x; v_1}, Y_s^{t_1, x; v_1}, Z_s^{t_1, x; v_1}, v_1(s)) \right| ds \right)^2 \right\} \\ & + C \Psi_{0, T}^{1/2} \left( \mathbb{E} \left\{ \sup_{0 \leq s \leq T} |S_s^{t_1, x; v_1} - S_s^{t_2, x; v_1}|^2 \right\} \right)^{1/2}, \end{aligned}$$

where

$$\begin{aligned} \Psi_{0,T} = \mathbb{E} & \left\{ |\Phi(X_T^{t_1,x;v_1})|^2 + \left( \int_{t_1}^T |g(s, X_s^{t_1,x;v_1}, 0, 0, v_1(s))| ds \right)^2 \right. \\ & + \sup_{t_1 \leq s \leq T} |h(s, X_s^{t_1,x;v_1})|^2 + |\Phi(X_T^{t_2,x;v_1})|^2 \\ & \left. + \left( \int_{t_2}^T |g(s, X_s^{t_2,x;v_1}, 0, 0, v_1(s))| ds \right)^2 + \sup_{t_2 \leq s \leq T} |h(s, X_s^{t_2,x;v_1})|^2 \right\}. \end{aligned}$$

Now we consider the items for the right-hand side of inequality (3.18).

For the first item: From the Lipschitz condition and Propositions 3.1 and 3.2, we get

$$I \leq C \mathbb{E} \{ |X_T^{t_1,x;v_1} - X_T^{t_2,x;v_1}|^2 \} \leq C \mathbb{E} \{ |X_{t_1}^{t_2,x;v_1} - x|^2 \} \leq C(t_2 - t_1).$$

For the second item: By the Lipschitz condition,  $(a+b)^2 \leq a^2/2 + b^2/2$ , and Propositions 3.1, 3.2, and 3.3, we get

$$II \leq C(t_2 - t_1).$$

For the third item: By the same arguments we get

$$\Psi_{0,T} \leq C.$$

Next, we know that

$$\begin{aligned} |S_s^{t_1,x;v_1} - S_s^{t_2,x;v_1}|^2 &= |h(s, X_s^{t_1,x;v_1}) - h(s, X_s^{t_2,x;v_1})|^2 \\ &\leq C |X_s^{t_1,x;v_1} - X_s^{t_2,x;v_1}|^2, & s \in [t_2, T], \\ |S_s^{t_1,x;v_1} - S_s^{t_2,x;v_1}|^2 &= |h(s, X_s^{t_1,x;v_1}) - h(t_2, x)|^2 \\ &\leq C |X_s^{t_1,x;v_1} - x|^2 + 2|h(s, x) - h(t_2, x)|^2, & s \in [t_1, t_2], \\ |S_s^{t_1,x;v_1} - S_s^{t_2,x;v_1}|^2 &= |h(t_1, x) - h(t_2, x)|^2, & s \in [0, t_1]. \end{aligned}$$

So we have

$$\begin{aligned} & \mathbb{E} \left\{ \sup_{0 \leq s \leq T} |S_s^{t_1,x;v_1} - S_s^{t_2,x;v_1}|^2 \right\} \\ & \leq \mathbb{E} \left\{ \left( \sup_{0 \leq s \leq t_1} + \sup_{t_1 \leq s \leq t_2} + \sup_{t_2 \leq s \leq T} \right) |S_s^{t_1,x;v_1} - S_s^{t_2,x;v_1}|^2 \right\} \\ & \leq C(t_2 - t_1) + |h(t_1, x) - h(t_2, x)|^2 + 2 \sup_{t_1 \leq s \leq t_2} |h(s, x) - h(t_2, x)|^2 \\ & \leq C(t_2 - t_1) + 3 \sup_{t_1 \leq s \leq t_2} |h(s, x) - h(t_2, x)|^2. \end{aligned}$$

From the previous analysis, we get

$$\begin{aligned} |Y_{t_1}^{t_1,x;v_1} - Y_{t_2}^{t_2,x;v_1}|^2 &\leq C(t_2 - t_1) + 3 \sup_{t_1 \leq s \leq t_2} |h(s, x) - h(t_2, x)|^2, \\ |Y_{t_1}^{t_1,x;v_1} - Y_{t_2}^{t_2,x;v_1}| &\leq C(t_2 - t_1)^{1/2} + 3 \sup_{t_1 \leq s \leq t_2} |h(s, x) - h(t_2, x)|. \end{aligned}$$

With the same arguments applying to  $|Y_{t_1}^{t_1,x;v_2} - Y_{t_2}^{t_2,x;v_2}|^2$ , we have

$$|u(t_1, x) - u(t_2, x)| \leq C(t_2 - t_1)^{1/2} + 3 \sup_{t_1 \leq s \leq t_2} |h(s, x) - h(t_2, x)| + \varepsilon.$$

Because of the arbitrariness of  $\varepsilon$ , we get

$$|u(t_1, x) - u(t_2, x)| \leq C(t_2 - t_1)^{1/2} + 3 \sup_{t_1 \leq s \leq t_2} |h(s, x) - h(t_2, x)|.$$

From the continuity of  $h(t, x)$  with respect to  $t$ , we obtain the continuity of  $u(t, x)$  with respect to  $t$ . The proof is now completed.  $\square$

*Remark 3.13.* Lemma 3.8(i) and Proposition 3.12 imply the joint continuity of  $u$  in  $(t, x)$ , which is required in the next section.

#### 4. Viscosity solution of an obstacle problem for the HJB equation.

In this section, we consider the relation between the value function of the above recursive optimal control problem and the following obstacle problem for nonlinear second-order parabolic PDE, which is called the Hamilton–Jacobi–Bellman equation:

$$(4.1) \quad \begin{cases} \min \left( u(t, x) - h(t, x), \right. \\ \quad \left. -\frac{\partial u}{\partial t}(t, x) - \sup_{v \in U} \{ \mathcal{L}(t, x, v)u(t, x) \right. \\ \quad \left. + g(t, x, u(t, x), \nabla u(t, x)\sigma(t, x, v), v) \} \right) = 0, \\ u(T, x) = \Phi(x), \end{cases}$$

where  $\mathcal{L}$  is a family of second order linear partial differential operators,

$$\mathcal{L}(t, x, v)\varphi = \frac{1}{2} \text{Tr}((\sigma\sigma^T)(t, x, v)D^2\varphi) + \langle b(t, x, v), D\varphi \rangle.$$

Here the function  $b, \sigma, g, \Phi, h$  satisfies (H3.1) and (H3.4), respectively.

We want to prove that the value function  $u(t, x)$  introduced by (3.9) is the unique viscosity solution of the obstacle problem for HJB equation obstacle problem (4.1). We first recall the definition of a viscosity solution for (4.1) from [4] and [1]. Below,  $S^n$  will denote the set of  $n \times n$  symmetric matrices and  $\Theta_{s,t} := [s, t] \times \mathbb{R}^n$ .

**DEFINITION 4.1.** Let  $u(t, x) \in C(\Theta_{0,T})$  and  $(t, x) \in [0, T] \times \mathbb{R}^n$ . We denote by  $\mathcal{P}_{\Theta_{0,T}}^{2,+}u(t, x)$ , the parabolic superjet of  $u$  at  $(t, x)$ , the set of triples  $(p, q, X) \in \mathbb{R} \times \mathbb{R}^n \times S^n$  which satisfies

$$u(s, y) \leq u(t, x) + p(s - t) + \langle q, y - x \rangle + \frac{1}{2} \langle X(y - x), y - x \rangle + o(|s - t| + |y - x|^2).$$

Similarly, we denote by  $\mathcal{P}_{\Theta_{0,T}}^{2,-}u(t, x)$ , the parabolic subjet of  $u$  at  $(t, x)$ , the set of triples  $(p, q, X) \in \mathbb{R} \times \mathbb{R}^n \times S^n$  which satisfies

$$u(s, y) \geq u(t, x) + p(s - t) + \langle q, y - x \rangle + \frac{1}{2} \langle X(y - x), y - x \rangle + o(|s - t| + |y - x|^2).$$

*Example 4.* Suppose that  $\varphi \in C^{1,2}(\Theta_{0,T})$ . If  $u - \varphi$  attains local maximum at  $(t, x)$ , then

$$\left( \frac{\partial \varphi}{\partial t}(t, x), \nabla \varphi(t, x), D^2 \varphi(t, x) \right) \in \mathcal{P}_{\Theta_{0,T}}^{2,+} u(t, x).$$

If  $u - \varphi$  reaches a local minimum at  $(t, x)$ , then

$$\left( \frac{\partial \varphi}{\partial t}(t, x), \nabla \varphi(t, x), D^2 \varphi(t, x) \right) \in \mathcal{P}_{\Theta_{0,T}}^{2,-} u(t, x).$$

DEFINITION 4.2.

- (a) We call  $u(t, x) \in C([0, T] \times \mathbb{R}^n)$  a viscosity subsolution of (4.1) if  $u(T, x) \leq \Phi(x)$ ,  $x \in \mathbb{R}^n$ , and at any point  $(t, x) \in \Theta_{0,T}$ , for any  $(p, q, X) \in \mathcal{P}_{\Theta_{0,T}}^{2,+} u(t, x)$ , we have

$$\min \left( u(t, x) - h(t, x), \right. \\ \left. -p - \sup_{v \in U} \left\{ \frac{1}{2} Tr(aX) + \langle b, q \rangle + g(t, x, u(t, x), q\sigma(t, x, v), v) \right\} \right) \leq 0.$$

In other words, at any point  $(t, x)$  where  $u(t, x) > h(t, x)$ , we have

$$-p - \sup_{v \in U} \left\{ \frac{1}{2} Tr(aX) + \langle b, q \rangle + g(t, x, u(t, x), q\sigma(t, x, v), v) \right\} \leq 0.$$

- (b) We call  $u(t, x) \in C([0, T] \times \mathbb{R}^n)$  a viscosity supersolution of (4.1) if  $u(T, x) \geq \Phi(x)$ ,  $x \in \mathbb{R}^n$ , and at any point  $(t, x) \in \Theta_{0,T}$ , for any  $(p, q, X) \in \mathcal{P}_{\Theta_{0,T}}^{2,-} u(t, x)$ , we have

$$\min \left( u(t, x) - h(t, x), \right. \\ \left. -p - \sup_{v \in U} \left\{ \frac{1}{2} Tr(aX) + \langle b, q \rangle + g(t, x, u(t, x), q\sigma(t, x, v), v) \right\} \right) \geq 0.$$

In other words, at each point, we have both  $u(t, x) \geq h(t, x)$  and

$$-p - \sup_{v \in U} \left\{ \frac{1}{2} Tr(aX) + \langle b, q \rangle + g(t, x, u(t, x), q\sigma(t, x, v), v) \right\} \geq 0.$$

- (c)  $u(t, x) \in C([0, T] \times \mathbb{R}^n)$  is said to be a viscosity solution of (4.1) if it is both a viscosity sub- and supersolution.

We are going to use the approximation of the reflected BSDE by penalization, which was studied in section 6 of [9]. For each  $(t, x) \in [0, T] \times \mathbb{R}^n$ ,  $n \in \mathbb{N}$ , let  $\{( {}^n Y_s^{t,x;v}, {}^n Z_s^{t,x;v} ), t \leq s \leq T \}$  be the solution of the BSDEs

$$\begin{aligned} {}^n Y_s^{t,x;v} &= \Phi(X_T^{t,x;v}) + \int_s^T g(r, X_r^{t,x;v}, {}^n Y_r^{t,x;v}, {}^n Z_r^{t,x;v}, v_r) dr \\ &\quad + n \int_s^T ( {}^n Y_r^{t,x;v} - h(r, X_r^{t,x;v}) )^- dr - \int_s^T {}^n Z_r^{t,x;v} dW_r, \quad t \leq s \leq T. \end{aligned}$$

We define

$$(4.2) \quad J_n(t, x; v(\cdot)) := {}^n Y_t^{t, x; v}, \quad v(\cdot) \in \mathcal{U}, \quad 0 \leq t \leq T, \quad x \in \mathbb{R}^n;$$

$$(4.3) \quad u_n(t, x) := \operatorname{esssup}_{v(\cdot) \in \mathcal{U}} J_n(t, x; v(\cdot)), \quad 0 \leq t \leq T, \quad x \in \mathbb{R}^n.$$

It is known from [13] and [14] that  $u_n(t, x)$  defined in (4.3) is the viscosity solution of the PDE

$$\begin{cases} -\frac{\partial u_n}{\partial t}(t, x) - \sup_{v \in U} \{ \mathcal{L}(t, x, v) u_n(t, x) + g_n(t, x, u_n(t, x), \nabla u_n(t, x) \sigma(t, x, v), v) \} = 0, \\ u_n(T, x) = \Phi(x), \end{cases}$$

where

$$g_n(t, x, r, p\sigma(t, x, v), v) = g(t, x, r, p\sigma(t, x, v), v) + n(r - h(t, x))^-.$$

Then

LEMMA 4.3.  $u_n(t, x) \uparrow u(t, x)$ ,  $0 \leq t \leq T$ ,  $x \in \mathbb{R}^n$ .

*Proof.* From the result of section 6 in [9], for each  $0 \leq t \leq T$ ,  $x \in \mathbb{R}^n$ ,

$$J_n(t, x; v(\cdot)) \uparrow J(t, x; v(\cdot)) \quad \text{as } n \rightarrow \infty.$$

From the monotonic property of  $J_n$  and the definition of  $u_n$  in (4.3), we get the monotonic property of  $u_n$ . Next, we will prove that  $u_n$  converges in  $n$ .

For each  $0 \leq t \leq T$ ,  $x \in \mathbb{R}^n$  for all  $\varepsilon > 0$ , there exists  $v(\cdot) \in \mathcal{U}$  such that

$$u(t, x) < Y_t^{t, x; v} + \varepsilon,$$

then

$$0 \leq u(t, x) - u_n(t, x) \leq Y_t^{t, x; v} - {}^n Y_t^{t, x; v} + \varepsilon.$$

Because  ${}^n Y_t^{t, x; v} \uparrow Y_t^{t, x; v}$ , a.s., we take limit on both sides,

$$0 \leq \limsup_{n \rightarrow \infty} (u(t, x) - u_n(t, x)) \leq \varepsilon.$$

From the arbitrariness of  $\varepsilon$ , we get the desired result.  $\square$

Remark 4.4. Since  $u_n$  and  $u$  are continuous, it follows from Dini's theorem that the convergence in the lemma is uniform on compact sets.

THEOREM 4.5.  $u$  defined by (3.9) is a viscosity solution of HJB equation (4.1).

*Proof.* We first prove that  $u$  is a subsolution of (4.1). Let  $(t, x)$  be a point where  $u(t, x) > h(t, x)$ , and let  $(p, q, X) \in \mathcal{P}_{\Theta_{0,T}}^{2,+} u(t, x)$ . From Lemma 6.1 in [4], there exists sequences

$$n_j \rightarrow +\infty, \quad (t_j, x_j) \rightarrow (t, x), \quad (p_j, q_j, X_j) \in \mathcal{P}_{\Theta_{0,T}}^{2,+} u_{n_j}(t_j, x_j),$$

such that

$$(p_j, q_j, X_j) \rightarrow (p, q, X).$$

But for any  $j$ ,

$$\begin{aligned} -p_j - \sup_{v \in U} \left\{ \frac{1}{2} \text{Tr}(aX_j) + \langle b, q_j \rangle + g(t_j, x_j, u_{n_j}(t_j, x_j), q_j \sigma(t_j, x_j, v), v) \right. \\ \left. + n_j(u_{n_j}(t_j, x_j) - h(t_j, x_j))^- \right\} \leq 0. \end{aligned}$$

From the assumption that  $u(t, x) > h(t, x)$  and the uniform convergence of  $u_n$ , it follows that for  $j$  large enough  $u_{n_j}(t_j, x_j) > h(t_j, x_j)$ , hence

$$-p_j - \sup_{v \in U} \left\{ \frac{1}{2} \text{Tr}(aX_j) + \langle b, q_j \rangle + g(t_j, x_j, u_{n_j}(t_j, x_j), q_j \sigma(t_j, x_j, v), v) \right\} \leq 0.$$

Let us admit for a moment the following lemma.

LEMMA 4.6.

$$\begin{aligned} \lim_{j \rightarrow \infty} \sup_{v \in U} \left\{ \frac{1}{2} \text{Tr}(aX_j) + \langle b, q_j \rangle + g(t_j, x_j, u_{n_j}(t_j, x_j), q_j \sigma(t_j, x_j, v), v) \right\} \\ = \sup_{v \in U} \lim_{j \rightarrow \infty} \left\{ \frac{1}{2} \text{Tr}(aX_j) + \langle b, q_j \rangle + g(t_j, x_j, u_{n_j}(t_j, x_j), q_j \sigma(t_j, x_j, v), v) \right\}. \end{aligned}$$

Letting  $j \rightarrow \infty$  in the above inequality yields

$$-p - \sup_{v \in U} \left\{ \frac{1}{2} \text{Tr}(aX) + \langle b, q \rangle + g(t, x, u(t, x), q \sigma(t, x, v), v) \right\} \leq 0,$$

then we get that  $u$  is a subsolution of (4.1).

We will obtain the result by proving that  $u$  is a supersolution of (4.1). Let  $(t, x)$  be an arbitrary point in  $\Theta_{0,T}$ , and  $(p, q, X) \in \mathcal{P}_{\Theta_{0,T}}^{2,-} u(t, x)$ . We already know that  $u(t, x) \geq h(t, x)$ . By the same arguments as above, there exist sequences

$$n_j \rightarrow +\infty, \quad (t_j, x_j) \rightarrow (t, x), \quad (p_j, q_j, X_j) \in \mathcal{P}_{\Theta_{0,T}}^{2,-} u_{n_j}(t_j, x_j),$$

such that

$$(p_j, q_j, X_j) \rightarrow (p, q, X).$$

While for any  $j$ , we have

$$\begin{aligned} -p_j - \sup_{v \in U} \left\{ \frac{1}{2} \text{Tr}(aX_j) + \langle b, q_j \rangle + g(t_j, x_j, u_{n_j}(t_j, x_j), q_j \sigma(t_j, x_j, v), v) \right. \\ \left. + n_j(u_{n_j}(t_j, x_j) - h(t_j, x_j))^- \right\} \geq 0. \end{aligned}$$

Because  $n_j(u_{n_j}(t_j, x_j) - h(t_j, x_j))^- \geq 0$ , we get

$$-p_j - \sup_{v \in U} \left\{ \frac{1}{2} \text{Tr}(aX_j) + \langle b, q_j \rangle + g(t_j, x_j, u_{n_j}(t_j, x_j), q_j \sigma(t_j, x_j, v), v) \right\} \geq 0.$$

As  $j \rightarrow \infty$ , we conclude that

$$-p - \sup_{v \in U} \left\{ \frac{1}{2} \text{Tr}(aX) + \langle b, q \rangle + g(t, x, u(t, x), q \sigma(t, x, v), v) \right\} \geq 0. \quad \square$$



To complete the proof, we need the proof of Lemma 4.6.

*Proof of Lemma 4.6.* For convenience, we denote

$$f_j(v) = \frac{1}{2}Tr(a, X_j) + \langle b, q_j \rangle + g(t_j, x_j, u_{n_j}(t_j, x_j), q_j \sigma(t_j, x_j, v), v).$$

First, for all  $v \in U$ ,

$$f_j(v) \leq \sup_{v \in U} f_j(v), \quad \lim_{j \rightarrow \infty} f_j(v) \leq \liminf_{j \rightarrow \infty} \sup_{v \in U} f_j(v),$$

so

$$(4.4) \quad \sup_{v \in U} \lim_{j \rightarrow \infty} f_j(v) \leq \liminf_{j \rightarrow \infty} \sup_{v \in U} f_j(v).$$

Second, we consider a subsequence  $\{j_k\}_{k=1}^{\infty}$  such that

$$\lim_{j_k \rightarrow \infty} \sup_{v \in U} f_{j_k}(v) = \limsup_{j \rightarrow \infty} \sup_{v \in U} f_j(v).$$

For all  $\varepsilon > 0$ , for all  $j_k$ ,  $\exists v_{j_k} \in U$  such that

$$\sup_{v \in U} f_{j_k}(v) \leq f_{j_k}(v_{j_k}) + \varepsilon.$$

Because  $U$  is compact, there also exists a convergent subsequence denoted by  $\{v_{j_k}\}_{k=1}^{\infty}$ , the limit is denoted by  $v_0$ . We consider the difference between  $f_{j_k}(v_{j_k})$  and  $f_{j_k}(v_0)$ : From the Lipschitz condition we get

$$|f_{j_k}(v_{j_k}) - f_{j_k}(v_0)| \leq C|v_{j_k} - v_0|^2 + C|v_{j_k} - v_0|,$$

where  $C$  only depends on the Lipschitz constant. It follows that for  $j_k$  large enough

$$|f_{j_k}(v_{j_k}) - f_{j_k}(v_0)| \leq \varepsilon.$$

Then

$$\sup_{v \in U} f_{j_k}(v) \leq f_{j_k}(v_0) + 2\varepsilon,$$

$$\limsup_{j \rightarrow \infty} \sup_{v \in U} f_j(v) = \lim_{j_k \rightarrow \infty} \sup_{v \in U} f_{j_k}(v) \leq \lim_{j_k \rightarrow \infty} f_{j_k}(v_0) + 2\varepsilon = \lim_{j \rightarrow \infty} f_j(v_0) + 2\varepsilon,$$

$$\limsup_{j \rightarrow \infty} \sup_{v \in U} f_j(v) \leq \sup_{v \in U} \lim_{j \rightarrow \infty} f_j(v_0) + 2\varepsilon.$$

Since  $\varepsilon$  is arbitrary,

$$(4.5) \quad \limsup_{j \rightarrow \infty} \sup_{v \in U} f_j(v) \leq \sup_{v \in U} \lim_{j \rightarrow \infty} f_j(v_0).$$

Combine (4.4) and (4.5), the result follows.  $\square$

To establish a uniqueness result for viscosity solution of (4.1), we use some techniques and methods from [1]. These techniques and methods can also be found in [3] for the uniqueness for viscosity solutions of Hamilton–Jacobi–Bellman–Isaacs equations related to stochastic differential games.

LEMMA 4.7. *Let  $u_1 \in C([0, T] \times \mathbb{R}^n)$  be a viscosity subsolution and  $u_2 \in C([0, T] \times \mathbb{R}^n)$  be a viscosity supersolution of (4.1). Then the function  $w := u_1 - u_2$  is a viscosity subsolution of the system*

$$(4.6) \quad \begin{cases} \min \left( w(t, x) \right. \\ \left. - \frac{\partial w}{\partial t}(t, x) - \sup_{v \in U} \{ \mathcal{L}(t, x, v)w(t, x) + L|w| + L|\nabla w \sigma(t, x, v)| \} \right) = 0, \\ w(T, x) = 0. \end{cases}$$

The proof is similar to Lemma 3.7 in [1], hence we omit it.

Now we are going to construct one suitable smooth supersolution for (4.6).

LEMMA 4.8. *For any  $A > 0$ , there exists  $C_1 > 0$  such that the function*

$$\chi(t, x) = \exp \{ (C_1(T - t) + A)\psi(x) \},$$

where

$$\psi(x) = \left[ \log \left( (|x|^2 + 1)^{\frac{1}{2}} \right) + 1 \right]^2$$

satisfies

$$\min \left( \chi(t, x), -\frac{\partial \chi}{\partial t}(t, x) - \sup_{v \in U} \{ \mathcal{L}(t, x, v)\chi(t, x) + L\chi(t, x) + L|\nabla \chi \sigma(t, x, v)| \} \right) > 0$$

in  $[t_1, T] \times \mathbb{R}^n$ , where  $t_1 = T - (A/C_1)$ .

*Proof.* Obviously, the function  $\chi$  defined in Lemma 4.8 satisfies  $\chi(t, x) > 0$ , for each  $(t, x) \in [0, T] \times \mathbb{R}^n$ . We consider estimates on the first and second order derivatives of  $\psi$ :

$$|D\psi(x)| \leq \frac{2[\psi(x)]^{\frac{1}{2}}}{(|x|^2 + 1)^{\frac{1}{2}}} \quad \text{and} \quad |D^2\psi(x)| \leq \frac{C \left( 1 + [\psi(x)]^{\frac{1}{2}} \right)}{|x|^2 + 1} \quad \text{in } \mathbb{R}^n.$$

These estimates imply that, if  $t \in [t_1, T]$ ,

$$|D\chi(t, x)| \leq C\chi(t, x) \frac{[\psi(x)]^{\frac{1}{2}}}{(|x|^2 + 1)^{\frac{1}{2}}}, \quad |D^2\chi(t, x)| \leq C\chi(t, x) \frac{\psi(x)}{|x|^2 + 1},$$

where the constant  $C$  only depend on  $A$ . Then we get

$$(4.7) \quad \begin{aligned} & \frac{\partial \chi}{\partial t}(t, x) + \sup_{v \in U} \{ \mathcal{L}(t, x, v)\chi(t, x) + L\chi(t, x) + L|\nabla \chi \sigma(t, x, v)| \} \\ &= \frac{\partial \chi}{\partial t}(t, x) + \sup_{v \in U} \left\{ \frac{1}{2} \text{Tr}((\sigma \sigma^T) D^2 \chi) + \langle b, D\chi \rangle + L\chi(t, x) + L|\nabla \chi \sigma(t, x, v)| \right\} \\ &\leq -C_1\chi(t, x)\psi(x) + \sup_{v \in U} \left\{ \frac{1}{2} \frac{|\sigma(t, x, v)|^2}{|x|^2 + 1} C\chi(t, x)\psi(x) \right. \\ &\quad \left. + \frac{|b(t, x, v)|}{(|x|^2 + 1)^{\frac{1}{2}}} C\chi(t, x)[\psi(x)]^{\frac{1}{2}} + L\chi(t, x) + L \frac{|\sigma(t, x, v)|}{(|x|^2 + 1)^{\frac{1}{2}}} C\chi(t, x)[\psi(x)]^{\frac{1}{2}} \right\}. \end{aligned}$$

Because  $b$  and  $\sigma$  are linear growth in  $x$ ,  $[\psi(x)]^{\frac{1}{2}} \leq \psi(x)$ , and  $1 \leq \psi(x)$ , the inequality (4.7) satisfies

$$\begin{aligned} (4.7) &< -C_1\chi(t, x)\psi(x) + \frac{1}{2}C\chi(t, x)\psi(x) + C\chi(t, x)\psi(x) \\ &\quad + L\chi(t, x)\psi(x) + LC\chi(t, x)\psi(x) \\ &= -(C_1 - \frac{1}{2}C - C - L - LC)\chi(t, x)\psi(x). \end{aligned}$$

It is clear that when  $C_1$  is large enough the right-hand side of the above inequality is negative and the proof is completed.  $\square$

With these preparations we can prove the uniqueness result for the viscosity solution of (4.1).

**THEOREM 4.9.** *Assume that  $b$ ,  $\sigma$ ,  $g$ ,  $\Phi$ , and  $h$  satisfy (H3.1) and (H3.4), respectively. Then there exists at most one viscosity solution of HJB equation (4.1) in the class of continuous functions which grow at most polynomially at infinity.*

*Proof.* Let  $u_1, u_2 \in C([0, T] \times \mathbb{R}^n)$  be two viscosity solutions of HJB equation (4.1). We define  $w := u_1 - u_2$ , then we have

$$\lim_{|x| \rightarrow \infty} w(t, x) e^{-A[\log((|x|^2+1)^{\frac{1}{2}})]^2} = 0$$

uniformly for  $t \in [0, T]$ , for some  $A > 0$ . This implies, in particular, that  $w(t, x) - \alpha\chi(t, x)$  is bounded from above in  $[t_1, T] \times \mathbb{R}^n$  for any  $\alpha > 0$  and that

$$M := \max_{[t_1, T] \times \mathbb{R}^n} (w - \alpha\chi)(t, x) e^{-L(T-t)}$$

is achieved at some point  $(t_0, x_0) \in [t_1, T] \times \mathbb{R}^n$  (depend on  $\alpha$ ). Here  $t_1$  is defined as in Lemma 4.8. Then we have two cases.

The first case:  $w(t_0, x_0) \leq 0$ .

Therefore we have

$$u_1(t, x) - u_2(t, x) \leq \alpha\chi(t, x), \quad (t, x) \in [t_1, T] \times \mathbb{R}^n.$$

Letting  $\alpha$  tend to zero, we obtain

$$(4.8) \quad u_1(t, x) \leq u_2(t, x), \quad (t, x) \in [t_1, T] \times \mathbb{R}^n.$$

The second case:  $w(t_0, x_0) > 0$ .

By the maximum point property, we deduce that

$$w(t, x) - \alpha\chi(t, x) \leq (w(t_0, x_0) - \alpha\chi(t_0, x_0))e^{-L(t-t_0)}, \quad (t, x) \in [t_1, T] \times \mathbb{R}^n,$$

and this inequality implies that the function  $w - \varphi$  has a global maximum point at  $(t_0, x_0)$  where

$$\varphi(t, x) = \alpha\chi(t, x) + (w(t_0, x_0) - \alpha\chi(t_0, x_0))e^{-L(t-t_0)}.$$

Since  $w$  is a viscosity subsolution of (4.6) from Lemma 4.7 and  $\varphi(t_0, x_0) = w(t_0, x_0) > 0$ , if  $t_0 \in [t_1, T]$ , we notice that

$$((\partial/\partial t)\varphi(t_0, x_0), \nabla\varphi(t_0, x_0), D^2\varphi(t_0, x_0)) \in \mathcal{P}_{\Theta_{t_1, T}}^{2,+} u(t_0, x_0),$$

then

$$\begin{aligned} & -\frac{\partial \varphi}{\partial t}(t_0, x_0) - \sup_{v \in U} \left\{ \frac{1}{2} \text{Tr}((\sigma \sigma^T)(t_0, x_0, v) D^2 \varphi(t_0, x_0)) + \langle b(t_0, x_0, v), D\varphi(t_0, x_0) \rangle \right. \\ & \left. + L\varphi(t_0, x_0) + L|\nabla \varphi(t_0, x_0) \sigma(t_0, x_0, v)| \right\} \leq 0. \end{aligned}$$

By the definition of  $\varphi$ , we rewrite the above inequality as

$$\begin{aligned} & \alpha \left[ -\frac{\partial \chi}{\partial t}(t_0, x_0) - \sup_{v \in U} \left\{ \frac{1}{2} \text{Tr}((\sigma \sigma^T)(t_0, x_0, v) D^2 \chi(t_0, x_0)) + \langle b(t_0, x_0, v), D\chi(t_0, x_0) \rangle \right. \right. \\ & \left. \left. + L\chi(t_0, x_0) + L|\nabla \chi(t_0, x_0) \sigma(t_0, x_0, v)| \right\} \right] \leq 0. \end{aligned}$$

This is a contradiction with Lemma 4.8. Therefore  $t_0 = T$ , which is a contradiction with the fact that  $w(t, x)$  is a viscosity subsolution of (4.6) (see Lemma 4.7). Then the second case does not happen.

If we change  $w(t, x) = u_1 - u_2$  for  $w'(t, x) = u_2 - u_1$ , then the same arguments lead to

$$(4.9) \quad u_2(t, x) \leq u_1(t, x), \quad (t, x) \in [t_1, T] \times \mathbb{R}^n.$$

Combining (4.8) with (4.9), we have

$$u_1(t, x) = u_2(t, x), \quad (t, x) \in [t_1, T] \times \mathbb{R}^n.$$

Applying successively the same argument on the intervals  $[t_2, t_1]$  where  $t_2 = (t_1 - A/C_1)^+$  and then, if  $t_2 > 0$  on  $[t_3, t_2]$  where  $t_3 = (t_2 - A/C_1)^+$ , etc. We finally obtain that

$$u_1(t, x) = u_2(t, x), \quad (t, x) \in [0, T] \times \mathbb{R}^n.$$

The proof is completed.  $\square$

**Acknowledgments.** The authors express their gratitude to Prof. Shige Peng for his elicitation and inspiring idea in recursive stochastic dynamic programming principle. The authors also thank Dr. Juan Li and Dr. Mingyu Xu for their constructive suggestions. In particular, the authors would like to thank the associate editor and two anonymous referees for their careful reading and helpful comments and suggestions which greatly improved this paper.

#### REFERENCES

- [1] G. BARLES, R. BUCKDAHN, AND E. PARDOUX, *Backward stochastic differential equations and integral-partial differential equations*, Stochastics Stochastics Rep., 60 (1997), pp. 57–83.
- [2] P. BRIAND, F. COQUET, Y. HU, J. MÉMIN, AND S. PENG, *A converse comparison theorem for BSDEs and related properties of g-expectation*, Electron. Comm. Probab., 5 (2000), pp. 101–117.
- [3] R. BUCKDAHN AND J. LI, *Stochastic differential games and viscosity solutions of Hamilton-Jacobi-Bellman-Isaacs equations*, SIAM J. Control Optim., 47 (2008), pp. 444–475.
- [4] M. G. CRANDALL, H. ISHII, AND P.-L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

- [5] J. CVITANIC AND I. KARATZAS, *Backward stochastic differential equations with reflection and Dynkin games*, Ann. Probab., 24 (1996), pp. 2024–2056.
- [6] D. DUFFIE AND L. EPSTEIN, *Stochastic differential utility*, Econometrica, 60 (1992), pp. 353–394.
- [7] S. HAMADÈNE AND J. P. LEPELTIER, *Reflected backward stochastic differential equations and mixed game problem*, Stochastic Process. Appl., 85 (2000), pp. 177–188.
- [8] S. HAMADÈNE, J. P. LEPELTIER, AND Z. WU, *Infinite horizon reflected backward stochastic differential equations and applications in mixed control and game problems*, Probab. Math. Statist., 19 (1999), pp. 211–234.
- [9] N. EL KAROUI, C. KAPOUDJIAN, E. PARDOUX, S. PENG, AND M. C. QUENEZ, *Reflected solutions, of backward SDE's, and related obstacle problems for PDE's*, Ann. Probab., 25 (1997), pp. 702–737.
- [10] N. EL KAROUI, E. PARDOUX, AND M. C. QUENEZ, *Reflected backward SDEs and American options*, in Numerical Methods in Finance, L. C. G. Rogers and D. Talay, eds., Cambridge University Press, Cambridge, UK, 1997, pp. 215–231.
- [11] N. EL KAROUI, S. PENG, AND M. C. QUENEZ, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.
- [12] E. PARDOUX AND S. PENG, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.
- [13] S. PENG, *A generalized dynamic programming principle and Hamilton-Jacobi-Bellmen equation*, Stochastics Stochastics Rep., 38 (1992), pp. 119–134.
- [14] S. PENG, *Backward stochastic differential equations—stochastic optimization theory and viscosity solutions of HJB equations*, in Topics on Stochastic Analysis, J. Yan, S. Peng, S. Fang, and L. Wu, eds., Science Press, Beijing, 1997, pp. 85–138 (in Chinese).

## ROBUST STABILITY OF POLYTOPIC SYSTEMS VIA AFFINE PARAMETER-DEPENDENT LYAPUNOV FUNCTIONS\*

GUANG-HONG YANG<sup>†</sup> AND JIUXIANG DONG<sup>‡</sup>

**Abstract.** This paper studies robust stability of linear systems with polytopic uncertainty. New necessary and sufficient conditions for the existence of an affine parameter-dependent Lyapunov function assuring the Hurwitz or the Schur stability of a polytopic system are presented. These conditions are composed of a family of linear matrix inequality conditions of increasing precision. At each step, a set of linear matrix inequalities provides sufficient conditions for the existence of the affine parameter-dependent Lyapunov function, and necessity is asymptotically attained. Compared with the existing results in the literature, it is shown that the new stability conditions provide less conservative tests at each step. Numerical examples are given to illustrate the effectiveness of the new results.

**Key words.** linear systems, polytopic uncertainty, Hurwitz stability, Schur stability, linear matrix inequalities (LMIs), parameter-dependent Lyapunov functions

**AMS subject classifications.** 93C05, 34D20, 93D09, 93D20

**DOI.** 10.1137/060668948

**1. Introduction.** Robust stability analysis of systems with uncertainties is one of the most fundamental problems in system and control theory [1]. In particular, the Lyapunov approach is popular in investigating the Hurwitz and Schur stabilities of linear systems with polytopic uncertainty. In past years, the Hurwitz and the Schur stabilities of linear systems with polytopic uncertainty has received a great deal of attention, and some significant results have been obtained; see [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] and the references therein. For obtaining less conservative evaluations of stability domains, higher degree parameter-dependent Lyapunov functions have been exploited. Sufficient conditions based on homogeneous polynomially parameter-dependent Lyapunov functions  $v(x) = x^T P(\alpha)x$  with  $P(\alpha)$  of arbitrary degree on  $\alpha$  are given in [16], and a family of linear matrix inequality (LMI) conditions of increasing precision for the Hurwitz stability is presented in [17]. The paper [18] provides a systematic procedure for generating sufficient LMI conditions for assuring robust stability based on homogeneous polynomially parameter-dependent Lyapunov matrix functions of arbitrary degree on the uncertain parameters. Moreover, in [19],

---

\*Received by the editors September 3, 2006; accepted for publication (in revised form) May 28, 2008; published electronically October 13, 2008.

This work was supported in part by Program for New Century Excellent Talents in University (NCET-04-0283), the Funds for Creative Research Groups of China (60521003), Program for Changjiang Scholars and Innovative Research Team in University (IRT0421), the State Key Program of National Natural Science of China (grant 60534010), the Funds of National Science of China (grant 60674021), the 111 Project (B08015), and the Funds of Ph.D. program of MOE, China (grant 20060145019).

<http://www.siam.org/journals/sicon/47-5/66894.html>

<sup>†</sup>Corresponding author. College of Information Science and Engineering, Northeastern University, Shenyang, 110004, People's Republic of China, and Key Laboratory of Integrated Automation of Process Industry (Ministry of Education), Northeastern University, Shenyang, 110004, People's Republic of China (yangguanghong@ise.neu.edu.cn, yang-guanghong@163.com).

<sup>‡</sup>College of Information Science and Engineering, Northeastern University, Shenyang, 110004, People's Republic of China, and Key Laboratory of Integrated Automation of Process Industry (Ministry of Education), Northeastern University, Shenyang, 110004, People's Republic of China (dongjiuxiang@ise.neu.edu.cn, dong\_jiuxiang@163.com).

a piecewise Lyapunov function, instead of a single Lyapunov function, method is also exploited to reduce conservatism. In particular, based on Pólya's theorem [21], necessary and sufficient conditions for the existence of an affine parameter-dependent Lyapunov function assuring the Hurwitz or the Schur stability of a polytopic system are given in [20], and a systematic procedure for constructing a family of LMI conditions of increasing precision is developed, where the necessity is asymptotically attained.

This paper will continue to study the Hurwitz and Schur stabilities of polytopic systems. New necessary and sufficient conditions for the existence of an affine parameter-dependent Lyapunov function assuring the Hurwitz or Schur stability of a polytopic system are derived, which are composed of a family of linear matrix inequality conditions of increasing precision. At each step, a set of LMIs provides sufficient conditions for the existence of the affine parameter-dependent Lyapunov function, and necessity is asymptotically attained. Compared with the results in [20], the new stability conditions provide less conservative tests at each step and allow faster convergence to the necessary conditions for the existence of an affine parameter-dependent Lyapunov function in terms of steps. However, it should be pointed out that the faster convergence is not in terms of time, since the numerical complexity of the new proposed conditions is much greater than that in [20].

The paper is organized as follows. In the next section, system description and some preliminaries are given. In section 3, new necessary and sufficient conditions for the existence of an affine parameter-dependent Lyapunov function assuring the Hurwitz or Schur stability of a polytopic system are presented. Section 4 presents two examples to illustrate the effectiveness of the proposed methods. Finally, section 5 concludes the paper.

*Notation.* For a square matrix  $E$ ,  $\text{He}(E)$  is defined as  $\text{He}(E) = E + E^T$ .  $\mathbb{R}$  represents the set of real numbers,  $\mathbb{N}$  denotes the natural numbers set  $\{0, 1, 2, \dots\}$ , and  $p!$  denotes factorial, i.e.,  $p! = p(p-1)(p-2) \cdots (2)(1)$  for  $p \in \mathbb{N}$  with  $0! = 1$ .

**2. System description and preliminaries.** Consider a linear system described by

$$\dot{x}(t) = A(\alpha)x(t)$$

for the continuous-time case, and by

$$x(k+1) = A(\alpha)x(k)$$

for the discrete-time case, where  $A(\alpha)$  belongs to a polytope of real square matrices defined by

$$(1) \quad \mathcal{A} = \left\{ A(\alpha) : A(\alpha) = \sum_{i=1}^N \alpha_i A_i; \alpha \in \Delta_N \right\},$$

$$\Delta_N = \left\{ \alpha \in \mathbb{R}^N : \sum_{i=1}^N \alpha_i = 1; \alpha_i \geq 0 \right\},$$

where  $N$  is the number of vertices and  $\Delta_N$  stands for unit simplex.

In this paper, the Hurwitz and Schur stabilities problem of  $\mathcal{A}$  will be studied by considering an affine parameter-dependent Lyapunov matrix given by

$$(2) \quad P(\alpha) = \sum_{i=1}^N \alpha_i P_i, \quad \alpha \in \Delta_N,$$

with  $P_i = P_i^T > 0$ ,  $i = 1, \dots, N$ . Through the use of the Lyapunov function  $V(x) = x^T P(\alpha)x$ , the Hurwitz stability of  $\mathcal{A}$  can be assured if and only if the following condition holds:

$$(3) \quad \Gamma(\alpha) = A^T(\alpha)P(\alpha) + P(\alpha)A(\alpha) < 0;$$

the Schur stability of  $\mathcal{A}$  can be assured if and only if the following condition holds:

$$A^T(\alpha)P(\alpha)A(\alpha) - P(\alpha) < 0.$$

By  $\sum_{i=1}^N \alpha_i = 1$ , (3) can be rewritten as

$$(4) \quad \Gamma(\alpha) = \left( \sum_{i=1}^N \alpha_i \right)^d (A^T(\alpha)P(\alpha) + P(\alpha)A(\alpha)) < 0,$$

where  $d \in \mathbb{N}$ . By writing (4) (with  $\alpha_i$  as a variable) as a sum of terms in an extended form and using the known Pólya's theorem [21], two computationally verifiable necessary and sufficient conditions, which guarantee that (3) holds, are given in [20, Lemmas 1 and 2].

The purpose of this paper is to derive new necessary and sufficient conditions, provide less conservative tests at each step, and allow faster convergence to exact robust stability bounds for the existence results of an affine parameter-dependent Lyapunov function.

The following definitions are needed, which are consistent with those in [20]. Define  $\mathcal{K}(d)$  as the set of  $N$ -tuples obtained as all possible combinations of  $k_1, k_2, \dots, k_N$ ,  $k_i \in \mathbb{N}$ ,  $i = 1, \dots, N$ , such that  $k_1 + k_2 + \dots + k_N = d$ .  $\mathcal{K}_\ell(d)$  is the  $\ell$ th  $N$ -tuple of  $\mathcal{K}(d)$  which is lexically ordered  $\ell = 1, \dots, J(d)$ . For a fixed  $N$ , the number of elements in  $\mathcal{K}(d)$  is given by  $J(d) = (N + d - 1)! / (d!(N - 1)!)$ , and the associated standard multinomial coefficients are  $\mathcal{C}^\ell(d) = d! / (k_1!k_2! \cdots k_N!)$ ,  $k_1k_2 \cdots k_N = \mathcal{K}_\ell(d)$ ,  $\ell = 1, \dots, J(d)$ . As an example consider the case when  $N = 3$  and  $d = 2$ , which yields  $J(2) = 6$ ,  $\mathcal{K}(2) = \{002, 011, 020, 101, 110, 200\}$ , and the coefficients  $\mathcal{C}^\ell(d) = \{1, 2, 1, 2, 2, 1\}$ .

Consider the modified multinomial coefficients

$$\begin{aligned} \mathcal{C}_i^\ell(d, a) &= \begin{cases} \frac{d!}{k_1! \cdots (k_i - a)! \cdots k_N!} & \text{if } k_i - a \in \mathbb{N}, \\ 0 & \text{otherwise,} \end{cases} \\ \mathcal{C}_{ij}^\ell(d, a, b) &= \begin{cases} \frac{d!}{k_1! \cdots (k_i - a)! \cdots (k_j - b)! \cdots k_N!} & \text{if } \begin{cases} k_i - a \in \mathbb{N}, \\ k_j - b \in \mathbb{N}, \end{cases} \\ 0 & \text{otherwise,} \end{cases} \\ \mathcal{C}_{ij\ell}^\ell(d, a, b, c) &= \begin{cases} \frac{d!}{k_1! \cdots (k_i - a)! \cdots (k_j - b)! \cdots (k_\ell - c)! \cdots k_N!} & \text{if } \begin{cases} k_i - a \in \mathbb{N}, \\ k_j - b \in \mathbb{N}, \\ k_\ell - c \in \mathbb{N}, \end{cases} \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

all of which depend on  $k_1k_2 \cdots k_N = \mathcal{K}_\ell(d)$ ,  $\ell = 1, \dots, J(d)$ .

LEMMA 1 (see [20]). (i) *An affine parameter-dependent Lyapunov matrix  $P(\alpha)$  given by (2) assures the Hurwitz stability of  $\mathcal{A}$  if and only if there exist symmetric positive definite matrices  $P_i$ ,  $i = 1, \dots, N$ , and a sufficiently large  $d$ , such that the*



following LMIs hold:

$$(5) \quad T_\ell = \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) T_{H_i} + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) T_{H_{ij}} < 0, \\ k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d+2) \quad \text{for } \ell = 1, \dots, J(d+2),$$

where

$$T_{H_i} = A_i^T P_i + P_i A_i, \\ T_{H_{ij}} = A_i^T P_j + P_j A_i + A_j^T P_i + P_i A_j.$$

(ii) An affine parameter-dependent Lyapunov matrix  $P(\alpha)$  given by (2) assures the Schur stability of  $\mathcal{A}$  if and only if there exist symmetric positive definite matrices  $P_i$ ,  $i = 1, \dots, N$ , and a sufficiently large  $d$ , such that the following LMIs hold:

$$D_\ell = \sum_{i=1}^N \mathcal{C}_i^\ell(d, 3) T_{S_i} + \sum_{i=1}^N \sum_{j \neq i; j=1}^N \mathcal{C}_{ij}^\ell(d, 2, 1) T_{S_{ij}} + \sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \sum_{k=j+1}^N \mathcal{C}_{ijk}^\ell(d, 1, 1, 1) T_{S_{ijk}} < 0, \\ k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d+3) \quad \text{for } \ell = 1, \dots, J(d+3),$$

where

$$T_{S_i} = A_i^T P_i A_i - P_i, \\ T_{S_{ij}} = A_i^T P_i A_j + A_j^T P_i A_i + A_i^T P_j A_i - 2P_i - P_j, \\ T_{S_{ijk}} = A_j^T P_i A_k + A_k^T P_i A_j + A_i^T P_j A_k + A_k^T P_j A_i + A_i^T P_k A_j + A_j^T P_k A_i \\ - 2(P_i + P_j + P_k).$$

LEMMA 2 (see [20]). An affine parameter-dependent Lyapunov matrix  $P(\alpha)$  given by (2) assures the stability of  $\mathcal{A}$  if and only if there exist symmetric positive definite matrices  $P_i$ ,  $i = 1, \dots, N$ , matrices  $X_i \in \mathbb{R}^{2n \times 2n}$ ,  $i = 1, \dots, N$ , and a sufficiently large  $d$  such that the following LMIs hold:

$$(6) \quad \Lambda_\ell = \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) T_{F_i} + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) T_{F_{ij}} < 0, \\ k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d+2) \quad \text{for } \ell = 1, \dots, J(d+2),$$

where

$$(7) \quad T_{F_i} = Q_i + X_i B_i + B_i^T X_i^T, \\ T_{F_{ij}} = Q_i + Q_j + X_j B_i + B_i^T X_j^T + X_i B_j + B_j^T X_i^T, \\ B_i = \begin{bmatrix} A_i & -I \end{bmatrix},$$

and  $Q_i$ ,  $i = 1, \dots, N$ , are, respectively, given, for the Hurwitz and Schur cases, by

$$(8) \quad Q_{H_i} = \begin{bmatrix} 0 & P_i \\ P_i & 0 \end{bmatrix}, \quad Q_{S_i} = \begin{bmatrix} -P_i & 0 \\ 0 & P_i \end{bmatrix}.$$

It should be pointed out that the condition of Lemma 2 is less conservative than that of Lemma 1 for each  $d$ .

In particular, the following equality from [20] will be used in the next section:

$$\begin{aligned}
 (9) \quad & (\alpha_1 + \cdots + \alpha_N)^d (A^T(\alpha)P(\alpha) + P(\alpha)A(\alpha)) \\
 &= (\alpha_1 + \cdots + \alpha_N)^d \left( \sum_{i=1}^N \alpha_i^2 T_{H_i} + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \alpha_i \alpha_j T_{H_{ij}} \right) \\
 &= \sum_{\ell=1}^{J(d+2)} \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} T_\ell,
 \end{aligned}$$

where  $P(\alpha)$  is the same as in (2);  $T_{H_i}$ ,  $T_{H_{ij}}$ , and  $T_\ell$  are the same as in (5).

Moreover, the following two technical lemmas will be needed in the later development.

LEMMA 3. (i)

$$\begin{aligned}
 (10) \quad & \sum_{\ell=1}^{J(d+2)} \sum_{1 \leq i \leq N} (\alpha_1^{k_1} \cdots \alpha_i^{k_i-2} \cdots \alpha_N^{k_N}) (\alpha_i^2) J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N} \\
 &= \sum_{\ell=1}^{J(d)} (\alpha_1^{k_1} \cdots \alpha_i^{k_i} \cdots \alpha_N^{k_N}) \sum_{1 \leq i \leq N} (\alpha_i^2) J_{ii}^{k_1 \cdots k_i \cdots k_N},
 \end{aligned}$$

where

$$(11) \quad J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N} = 0 \quad \text{for } k_i - 2 < 0.$$

(ii)

$$\begin{aligned}
 & \sum_{\ell=1}^{J(d+2)} \operatorname{He} \left( \sum_{1 \leq i < j \leq N} (\alpha_1^{k_1} \cdots \alpha_i^{k_i-1} \cdots \alpha_j^{k_j-1} \cdots \alpha_N^{k_N}) (\alpha_i \alpha_j) J_{ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} \right) \\
 &= \sum_{\ell=1}^{J(d)} (\alpha_1^{k_1} \cdots \alpha_i^{k_i} \cdots \alpha_j^{k_j} \cdots \alpha_N^{k_N}) \sum_{1 \leq i < j \leq N} (\alpha_i \alpha_j) \operatorname{He}(J_{ij}^{k_1 \cdots k_i \cdots k_j \cdots k_N}),
 \end{aligned}$$

where

$$J_{ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} = 0 \quad \text{for } k_i - 1 < 0 \quad \text{or } k_j - 1 < 0.$$

*Proof.* (i) Consider an arbitrary term of the left-hand side of equality (10) given by

$$(12) \quad (\alpha_1^{k_1} \cdots \alpha_i^{k_i-2} \cdots \alpha_N^{k_N}) (\alpha_i^2) J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N},$$

where  $k_1 + \cdots + k_N = d + 2$ .

If  $k_i - 2 < 0$ , then from (11), the term (12) is zero.

If  $k_i - 2 \geq 0$ , then  $k_1 \cdots (k_i - 2) \cdots k_N \in \mathcal{H}(d)$ , which implies that (12) also is one term of the right-hand side of equality (10).

On the other hand, consider an arbitrary term of the right-hand side of equality (10) given by

$$(13) \quad (\alpha_1^{k_1} \cdots \alpha_i^{k_i} \cdots \alpha_N^{k_N}) (\alpha_i^2) J_{ii}^{k_1 \cdots k_i \cdots k_N},$$

where  $k_1 + \dots + k_N = d$ . Since  $k_1 \dots k_i \dots k_N \in \mathcal{K}(d)$ ,  $k_1 \dots (k_i + 2) \dots k_N \in \mathcal{K}(d + 2)$ . Equation (13) can be rewritten as

$$(\alpha_1^{k_1} \dots \alpha_i^{k_i+2-2} \dots \alpha_N^{k_N})(\alpha_i^2)J_{ii}^{k_1 \dots (k_i+2-2) \dots k_N},$$

which implies that (13) also is one term of the left-hand side of equality (10).

Moreover, all terms are different from each other in the left-hand (right-hand) side of equality (10), so both sides of equality (10) are the same. Thus, it follows that equality (10) holds.

(ii) The proof is similar to that for (i), and thus is omitted here.  $\square$

LEMMA 4. *If*

$$(14) \quad \begin{aligned} J^{k_1 k_2 \dots k_N} &= \begin{bmatrix} J_{ij}^{k_1 k_2 \dots k_N} \end{bmatrix}_{N \times N} > 0, \\ k_1 k_2 \dots k_N &= \mathcal{K}_\ell(d) \quad \text{for } \ell = 1, \dots, J(d), \end{aligned}$$

then for arbitrary  $\epsilon > 0$ , the following inequalities hold:

$$(15) \quad \begin{aligned} J_b^{k_1 k_2 \dots k_N} &= \begin{bmatrix} J_{bij}^{k_1 k_2 \dots k_N} \end{bmatrix}_{N \times N} > 0, \\ k_1 k_2 \dots k_N &= \mathcal{K}_\ell(d) \quad \text{for } \ell = 1, \dots, J(d), \end{aligned}$$

where

$$\begin{aligned} J_{bii}^{k_1 k_2 \dots k_N} &= \begin{bmatrix} J_{ii}^{k_1 k_2 \dots k_N} + \epsilon I & 0 \\ 0 & \epsilon I \end{bmatrix}, \quad 1 \leq i \leq N, \\ J_{bij}^{k_1 k_2 \dots k_N} &= \begin{bmatrix} J_{ij}^{k_1 k_2 \dots k_N} & 0 \\ 0 & 0 \end{bmatrix}, \quad 1 \leq i < j \leq N. \end{aligned}$$

*Proof.* From condition (14), we have

$$(16) \quad \begin{aligned} J_a^{k_1 k_2 \dots k_N} &= \begin{bmatrix} J_{aij}^{k_1 k_2 \dots k_N} \end{bmatrix}_{N \times N} \geq 0, \\ k_1 k_2 \dots k_N &= \mathcal{K}_\ell(d) \quad \text{for } \ell = 1, \dots, J(d), \end{aligned}$$

where

$$J_{aij}^{k_1 k_2 \dots k_N} = \begin{bmatrix} J_{ij}^{k_1 k_2 \dots k_N} & 0 \\ 0 & 0 \end{bmatrix}, \quad 1 \leq i \leq j \leq N;$$

then from (16), for arbitrary  $\epsilon > 0$ , it follows that

$$(17) \quad \begin{aligned} J_a^{k_1 k_2 \dots k_N} + \text{diag}[\epsilon I \dots \epsilon I] &= \begin{bmatrix} J_{aij}^{k_1 k_2 \dots k_N} \end{bmatrix}_{N \times N} + \text{diag}[\epsilon I \dots \epsilon I] > 0, \\ k_1 k_2 \dots k_N &= \mathcal{K}_\ell(d) \quad \text{for } \ell = 1, \dots, J(d). \end{aligned}$$

Choose

$$\begin{aligned} J_{bii}^{k_1 k_2 \dots k_N} &= J_{a ii}^{k_1 k_2 \dots k_N} + \epsilon I, \quad 1 \leq i \leq N, \\ J_{bij}^{k_1 k_2 \dots k_N} &= J_{a ij}^{k_1 k_2 \dots k_N}, \quad 1 \leq i < j \leq N. \end{aligned}$$

By (17), it follows that (15) holds. Thus, the proof is complete.  $\square$

**3. Main results.** In the following, new necessary and sufficient conditions are given for  $\mathcal{A}$  to be Hurwitz or Schur stable via an affine parameter-dependent Lyapunov matrix  $P(\alpha)$  given by (2).

**THEOREM 5.** (i) *An affine parameter-dependent Lyapunov matrix  $P(\alpha)$  given by (2) assures the Hurwitz stability of  $\mathcal{A}$  if and only if there exist a sufficiently large  $d$ , symmetric positive matrices  $P_i$ ,  $i = 1, \dots, N$ ,  $J_{ii}^{k_1 k_2 \dots k_N}$  and matrices  $J_{ij}^{k_1 k_2 \dots k_N} = (J_{ji}^{k_1 k_2 \dots k_N})^T$ ,  $k_1 k_2 \dots k_N = \mathcal{K}_\ell(d)$  for  $\ell = 1, \dots, J(d)$  such that the following LMIs hold:*

$$(18) \quad T_\ell + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \dots (k_i-2) \dots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} \right) < 0,$$

$$k_1 k_2 \dots k_N = \mathcal{K}_\ell(d+2) \quad \text{for } \ell = 1, \dots, J(d+2),$$

$$(19) \quad J^{k_1 k_2 \dots k_N} = \begin{bmatrix} J_{ij}^{k_1 k_2 \dots k_N} \end{bmatrix}_{N \times N} > 0,$$

$$k_1 k_2 \dots k_N = \mathcal{K}_\ell(d) \quad \text{for } \ell = 1, \dots, J(d),$$

where

$$J_{ii}^{k_1 \dots (k_i-2) \dots k_N} = 0 \quad \text{for } k_i - 2 < 0,$$

$$J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} = 0 \quad \text{for } k_i - 1 < 0 \quad \text{or } k_j - 1 < 0,$$

and  $T_\ell$ ,  $\ell = 1, \dots, J(d+2)$  are the same as in (5).

(ii) *An affine parameter-dependent Lyapunov matrix  $P(\alpha)$  given by (2) assures the Schur stability of  $\mathcal{A}$  if and only if there exist a sufficiently large  $d$ , symmetric positive matrices  $P_i$ ,  $i = 1, \dots, N$ ,  $J_{ii}^{k_1 k_2 \dots k_N}$  and matrices  $J_{ij}^{k_1 k_2 \dots k_N} = (J_{ji}^{k_1 k_2 \dots k_N})^T$ ,  $k_1 k_2 \dots k_N = \mathcal{K}_\ell(d+1)$  for  $\ell = 1, \dots, J(d+1)$  such that the following LMIs hold:*

$$D_\ell + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \dots (k_i-2) \dots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} \right) < 0,$$

$$k_1 k_2 \dots k_N = \mathcal{K}_\ell(d+3) \quad \text{for } \ell = 1, \dots, J(d+3),$$

$$J^{k_1 k_2 \dots k_N} = \begin{bmatrix} J_{ij}^{k_1 k_2 \dots k_N} \end{bmatrix}_{N \times N} > 0,$$

$$k_1 k_2 \dots k_N = \mathcal{K}_\ell(d+1) \quad \text{for } \ell = 1, \dots, J(d+1),$$

where

$$J_{ii}^{k_1 \dots (k_i-2) \dots k_N} = 0 \quad \text{for } k_i - 2 < 0,$$

$$J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} = 0 \quad \text{for } k_i - 1 < 0 \quad \text{or } k_j - 1 < 0,$$

and  $D_\ell$ ,  $\ell = 1, \dots, J(d+3)$  are the same as in (5).

*Proof.* (i) *Sufficiency.* By pre- and postmultiplying (19) by

$$\begin{bmatrix} \alpha_1 I & \alpha_2 I & \dots & \alpha_N I \end{bmatrix}$$

and its transpose, we then have

$$(20) \quad \sum_{1 \leq i \leq N} \alpha_i^2 J_{ii}^{k_1 k_2 \dots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} \alpha_i \alpha_j J_{ij}^{k_1 k_2 \dots k_N} \right) > 0,$$

$$k_1 k_2 \dots k_N = \mathcal{K}_\ell(d), \quad \ell = 1, \dots, J(d).$$

Multiplying (18) by  $\alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} = \mathcal{K}_\ell(d+2)$ , it follows that

$$\begin{aligned}
 (21) \quad & \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} \left( T_\ell + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N} \right. \\
 & \quad \left. + \operatorname{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} \right) \right) \\
 & = \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} T_\ell + \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} \sum_{1 \leq i \leq N} J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N} \\
 & \quad + \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} \operatorname{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} \right) \\
 & = \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} T_\ell + \sum_{1 \leq i \leq N} (\alpha_1^{k_1} \cdots \alpha_i^{k_i-2} \cdots \alpha_N^{k_N}) (\alpha_i^2) J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N} \\
 & \quad + \operatorname{He} \left( \sum_{1 \leq i < j \leq N} (\alpha_1^{k_1} \cdots \alpha_i^{k_i-1} \cdots \alpha_j^{k_j-1} \cdots \alpha_N^{k_N}) (\alpha_i \alpha_j) J_{ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} \right) < 0, \\
 & \quad k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d+2), \quad \ell = 1, \dots, J(d+2).
 \end{aligned}$$

By summing (21) from  $\ell = 1$  to  $J(d+2)$ , we then can obtain

$$\begin{aligned}
 (22) \quad & \sum_{\ell=1}^{J(d+2)} \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} T_\ell + \sum_{\ell=1}^{J(d+2)} \sum_{1 \leq i \leq N} (\alpha_1^{k_1} \cdots \alpha_i^{k_i-2} \cdots \alpha_N^{k_N}) (\alpha_i^2) J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N} \\
 & + \sum_{\ell=1}^{J(d+2)} \operatorname{He} \left( \sum_{1 \leq i < j \leq N} (\alpha_1^{k_1} \cdots \alpha_i^{k_i-1} \cdots \alpha_j^{k_j-1} \cdots \alpha_N^{k_N}) (\alpha_i \alpha_j) J_{ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} \right) < 0.
 \end{aligned}$$

Applying Lemma 3 to (22), it follows that

$$\begin{aligned}
 & \sum_{\ell=1}^{J(d+2)} \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} T_\ell + \sum_{\ell=1}^{J(d+2)} \sum_{1 \leq i \leq N} (\alpha_1^{k_1} \cdots \alpha_i^{k_i-2} \cdots \alpha_N^{k_N}) (\alpha_i^2) J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N} \\
 & + \sum_{\ell=1}^{J(d+2)} \operatorname{He} \left( \sum_{1 \leq i < j \leq N} (\alpha_1^{k_1} \cdots \alpha_i^{k_i-1} \cdots \alpha_j^{k_j-1} \cdots \alpha_N^{k_N}) (\alpha_i \alpha_j) J_{ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} \right) \\
 & = \sum_{\ell=1}^{J(d+2)} \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} T_\ell + \sum_{\ell=1}^{J(d)} (\alpha_1^{k_1} \cdots \alpha_i^{k_i} \cdots \alpha_N^{k_N}) \sum_{1 \leq i \leq N} (\alpha_i^2) J_{ii}^{k_1 \cdots k_i \cdots k_N} \\
 & + \sum_{\ell=1}^{J(d)} (\alpha_1^{k_1} \cdots \alpha_i^{k_i} \cdots \alpha_j^{k_j} \cdots \alpha_N^{k_N}) \sum_{1 \leq i < j \leq N} (\alpha_i \alpha_j) \operatorname{He} (J_{ij}^{k_1 \cdots k_i \cdots k_j \cdots k_N})
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{\ell=1}^{J(d+2)} \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} T_\ell + \sum_{\ell=1}^{J(d)} (\alpha_1^{k_1} \cdots \alpha_i^{k_i} \cdots \alpha_N^{k_N}) \\
&\quad \times \left( \sum_{1 \leq i \leq N} (\alpha_i^2) J_{ii}^{k_1 \cdots k_i \cdots k_N} + \sum_{1 \leq i < j \leq N} (\alpha_i \alpha_j) \text{He} (J_{ij}^{k_1 \cdots k_i \cdots k_j \cdots k_N}) \right) < 0.
\end{aligned}$$

Combining the above inequality with (20), we have

$$(23) \quad \sum_{\ell=1}^{J(d+2)} \alpha_1^{k_1} \alpha_2^{k_2} \cdots \alpha_N^{k_N} T_\ell < 0.$$

From (9) and (23), it follows that

$$\begin{aligned}
&(\alpha_1 + \cdots + \alpha_N)^d (A(\alpha)^T P(\alpha) + P(\alpha) A(\alpha)) \\
&= A(\alpha)^T P(\alpha) + P(\alpha) A(\alpha) < 0,
\end{aligned}$$

which implies that  $\mathcal{A}$  is stable. Thus, the proof of the sufficiency is complete.

*Necessity.* If  $\mathcal{A}$  is stable via an affine parameter-dependent Lyapunov matrix  $P(\alpha)$  given by (2), then by Lemma 1, there exist symmetric positive definite matrices  $P_i$ ,  $i = 1, \dots, N$ , and a sufficiently large  $d$  such that (5) holds for  $\ell = 1, \dots, J(d+2)$ . Therefore, there exists a scalar  $\epsilon > 0$  such that

$$\begin{aligned}
(24) \quad &\sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) T_{H_i} + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) T_{H_{ij}} + \epsilon I \\
&= T_\ell + \epsilon I < 0, \\
&k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d+2), \quad \ell = 1, \dots, J(d+2).
\end{aligned}$$

Choose

$$\begin{aligned}
(25) \quad &J_{ii}^{k_1 k_2 \cdots k_N} = \frac{1}{N} \epsilon I, \quad 1 \leq i \leq N, \quad k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d), \quad \ell = 1, \dots, J(d), \\
&J_{ij}^{k_1 k_2 \cdots k_N} = 0, \quad 1 \leq i < j \leq N, \quad k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d), \quad \ell = 1, \dots, J(d).
\end{aligned}$$

Then

$$\begin{aligned}
(26) \quad &T_\ell + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} \right) \\
&\leq T_\ell + N \frac{1}{N} \epsilon I, \quad k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d+2), \quad \ell = 1, \dots, J(d+2).
\end{aligned}$$

From (24) and (26), we then have

$$\begin{aligned}
&T_\ell + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} \right) < 0, \\
&k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d+2), \quad \ell = 1, \dots, J(d+2);
\end{aligned}$$

i.e., (18) holds. Moreover, by (25), it follows that (19) holds. Thus, the conditions (18) and (19) are satisfied.

(ii) The proof is similar to that for (i), and thus is omitted here.  $\square$

*Remark 1.* Theorem 5 presents a necessary and sufficient condition for an affine parameter-dependent Lyapunov matrix  $P(\alpha)$  given by (2) to assure the stability of  $\mathcal{A}$  in terms of solutions of a set of LMIs. For each fixed  $d$ , the condition given by Theorem 5 is sufficient to assure the stability of  $\mathcal{A}$ , which is less conservative than that given by Lemma 1 (see [20]). In fact, if the condition given by Lemma 1 holds for a given  $d$ , then the condition given by Theorem 5 holds for a sufficiently small  $\epsilon > 0$  and  $J^{k_1 k_2 \dots k_N} = \epsilon I$ . Thus, compared with Lemma 1 (see [20]), the new stability condition provide a less conservative test for each  $d$ , which allows faster convergence to the necessary conditions for the existence of an affine parameter-dependent Lyapunov function in terms of  $d$ . However, it should be pointed out that the faster convergence is not on the time, since the numerical complexity of the condition given by Theorem 5 is much greater than that of Lemma 1 (see [20]); see Table 1.

The following theorem presents the relationship between the conditions of Theorem 5 for  $d = p$  and  $d = p + 1$ .

**THEOREM 6.** *If the condition of Theorem 5 holds for  $d = p$ , then the condition of Theorem 5 also holds for  $d = p + 1$ .*

*Proof.* If  $d = p$ , the condition of Theorem 5 holds; i.e., there exist symmetric positive matrices  $P_i$ ,  $i = 1, \dots, N$ ,  $J_{ii}^{k_1 k_2 \dots k_N}$  and matrices  $J_{ij}^{k_1 k_2 \dots k_N} = (J_{ji}^{k_1 k_2 \dots k_N})^T$ ,  $k_1 k_2 \dots k_N \in \mathcal{K}_\ell(p)$  for  $\ell = 1, \dots, J(p)$  satisfying (18) and (19).

By  $\sum_{i=1}^N \alpha_i = 1$ , we have

$$\begin{aligned} & (\alpha_1 + \dots + \alpha_N)^{p+1} (A^T(\alpha)P(\alpha) + P(\alpha)A(\alpha)) \\ &= (\alpha_1 + \dots + \alpha_N)^p (A^T(\alpha)P(\alpha) + P(\alpha)A(\alpha)). \end{aligned}$$

Combining the above equality with (9), it follows that

$$\sum_{\ell=1}^{J(p+3)} \alpha_1^{k_1} \alpha_2^{k_2} \dots \alpha_N^{k_N} T_{k_1 \dots k_N} = (\alpha_1 + \dots + \alpha_N) \sum_{\ell=1}^{J(p+2)} \alpha_1^{k_1} \alpha_2^{k_2} \dots \alpha_N^{k_N} T_{k_1 \dots k_N}$$

which further implies that, for  $k_1 k_2 \dots k_N \in \mathcal{K}(p+3)$ ,

$$T_{k_1 \dots k_N} = \sum_{r=1, k_r \geq 1}^N T_{k_1 \dots (k_r-1) \dots k_N}.$$

For  $k_1 k_2 \dots k_N \in \mathcal{K}(p+1)$ , choose

$$\begin{aligned} J_{ii}^{k_1 \dots k_N} &= \sum_{r=1, k_r \geq 1}^N J_{ii}^{k_1 \dots (k_r-1) \dots k_N}, \\ J_{ij}^{k_1 \dots k_N} &= \sum_{r=1, k_r \geq 1}^N J_{ij}^{k_1 \dots (k_r-1) \dots k_N}. \end{aligned}$$

Then, for  $k_1 k_2 \dots k_N \in \mathcal{K}(p+3)$ ,

$$\begin{aligned} (27) \quad & T_{k_1 \dots k_N} + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \dots (k_i-2) \dots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} \right) \\ &= \sum_{r=1, k_r \geq 1}^N \left( T_{k_1 \dots (k_r-1) \dots k_N} + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \dots (k_i-2) \dots (k_r-1) \dots k_N} \right. \\ & \quad \left. + \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots (k_r-1) \dots k_N} \right). \end{aligned}$$

From (18) with  $d = p$ , it follows that (27) is less than zero, which implies that for  $d = p + 1$ , (18) holds. Moreover, for  $k_1 k_2 \cdots k_N \in \mathcal{K}(p + 1)$ ,

$$\begin{aligned}
 (28) \quad & J^{k_1 k_2 \cdots k_N} \\
 &= \left[ J_{ij}^{k_1 k_2 \cdots k_N} \right]_{N \times N} \\
 &= \left[ \sum_{i=1, k_r \geq 1}^N J_{ij}^{k_1 \cdots (k_r-1) \cdots k_N} \right]_{N \times N} \\
 &= \sum_{i=1, k_r \geq 1}^N \left[ J_{ij}^{k_1 \cdots (k_r-1) \cdots k_N} \right]_{N \times N}.
 \end{aligned}$$

From (19) with  $d = p$ , it follows that (28) is greater than zero, which further implies that for  $d = p + 1$ , (19) holds. Thus, the condition of Theorem 5 for  $d = p + 1$  holds.  $\square$

*Remark 2.* Theorem 6 shows that the family of LMI conditions given by Theorem 5 is of increasing precision.

For a better understanding and comparison of the LMI conditions of Lemma 1 and Theorem 5, the concise LMIs are presented in the following for a polytope  $\mathcal{A}$  with  $N = 2$  vertices. For this case,

$$T_{H_1} = A_1^T P_1 + P_1 A_1; \quad T_{H_2} = A_2^T P_2 + P_2 A_2; \quad T_{H_{12}} = A_1^T P_2 + P_2 A_1 + A_2^T P_1 + P_1 A_2.$$

Then, for  $d = 0$ , we have  $J(2) = 3$ ,  $\mathcal{K}(0) = \{00\}$ ,  $\mathcal{K}(2) = \{02, 11, 20\}$ . The LMIs of Lemma 1 are

$$\begin{aligned}
 T_{H_1} &< 0, \\
 T_{H_2} &< 0, \\
 T_{H_{12}} &< 0.
 \end{aligned}$$

The LMIs of Theorem 5 are

$$\begin{aligned}
 T_{H_1} + J_{11}^{00} &< 0, \\
 T_{H_2} + J_{22}^{00} &< 0, \\
 T_{H_{12}} + \mathbf{He}(J_{12}^{00}) &< 0, \\
 \begin{bmatrix} J_{11}^{00} & J_{12}^{00} \\ (J_{12}^{00})^T & J_{22}^{00} \end{bmatrix} &> 0.
 \end{aligned}$$

For  $d = 1$ ,  $J(3) = 4$ ,  $\mathcal{K}(1) = \{01, 10\}$ ,  $\mathcal{K}(3) = \{03, 12, 21, 30\}$ . The LMIs of Lemma 1 are

$$\begin{aligned}
 T_{H_1} &< 0, \\
 T_{H_2} &< 0, \\
 T_{H_1} + T_{H_{12}} &< 0, \\
 T_{H_2} + T_{H_{12}} &< 0.
 \end{aligned}$$

The LMIs of Theorem 5 are

$$\begin{aligned}
 T_{H_1} + J_{11}^{10} &< 0, \\
 T_{H_2} + J_{22}^{01} &< 0, \\
 T_{H_1} + T_{H_{12}} + J_{11}^{01} + \mathbf{He}(J_{12}^{10}) &< 0,
 \end{aligned}$$



$$\begin{aligned} T_{H_2} + T_{H_{12}} + J_{22}^{10} + \text{He}(J_{12}^{01}) &< 0, \\ \begin{bmatrix} J_{11}^{10} & J_{12}^{10} \\ (J_{12}^{10})^T & J_{22}^{10} \end{bmatrix} &> 0, \\ \begin{bmatrix} J_{11}^{01} & J_{12}^{01} \\ (J_{12}^{01})^T & J_{22}^{01} \end{bmatrix} &> 0. \end{aligned}$$

For  $d = 2$ , we have  $J(4) = 5$ ,  $\mathcal{K}(2) = \{02, 11, 20\}$ ,  $\mathcal{K}(4) = \{04, 13, 22, 31, 40\}$ . The LMIs of Lemma 1 are

$$\begin{aligned} T_{H_1} &< 0, \\ T_{H_2} &< 0, \\ 2T_{H_1} + T_{H_{12}} &< 0, \\ T_{H_1} + T_{H_2} + 2T_{H_{12}} &< 0, \\ 2T_{H_2} + T_{H_{12}} &< 0. \end{aligned}$$

The LMIs of Theorem 5 are

$$\begin{aligned} T_{H_1} + J_{11}^{20} &< 0, \\ T_{H_2} + J_{22}^{02} &< 0, \\ 2T_{H_1} + T_{H_{12}} + J_{11}^{11} + \text{He}(J_{12}^{20}) &< 0, \\ T_{H_1} + T_{H_2} + 2T_{H_{12}} + J_{11}^{02} + J_{22}^{20} + \text{He}(J_{12}^{11}) &< 0, \\ 2T_{H_2} + T_{H_{12}} + J_{22}^{11} + \text{He}(J_{12}^{02}) &< 0, \\ \begin{bmatrix} J_{11}^{20} & J_{12}^{20} \\ (J_{12}^{20})^T & J_{22}^{20} \end{bmatrix} &> 0, \\ \begin{bmatrix} J_{11}^{11} & J_{12}^{11} \\ (J_{12}^{11})^T & J_{22}^{11} \end{bmatrix} &> 0, \\ \begin{bmatrix} J_{11}^{02} & J_{12}^{02} \\ (J_{12}^{02})^T & J_{22}^{02} \end{bmatrix} &> 0. \end{aligned}$$

**THEOREM 7.** *An affine parameter-dependent Lyapunov matrix  $P(\alpha)$  given by (2) assures the stability of  $\mathcal{A}$  if and only if there exist a sufficiently large  $d$ , symmetric positive definite matrices  $P_i$ ,  $i = 1, \dots, N$ ,  $J_{ii}^{k_1 k_2 \dots k_N}$  and matrices  $X_i \in \mathbb{R}^{2n \times 2n}$ ,  $i = 1, \dots, N$ ,  $J_{ij}^{k_1 k_2 \dots k_N} = (J_{ji}^{k_1 k_2 \dots k_N})^T$ ,  $k_1 k_2 \dots k_N = \mathcal{K}_\ell(d)$ ,  $\ell = 1, \dots, J(d)$ , such that the following LMIs hold:*

$$\begin{aligned} (29) \quad \Lambda_\ell + \sum_{1 \leq i \leq N} J_{bii}^{k_1 \dots (k_i-2) \dots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{bij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} \right) &< 0, \\ k_1 k_2 \dots k_N = \mathcal{K}_\ell(d+2), \quad \ell = 1, \dots, J(d+2), \\ J_b^{k_1 k_2 \dots k_N} = \begin{bmatrix} J_{bij}^{k_1 k_2 \dots k_N} \end{bmatrix}_{N \times N} &> 0, \\ (30) \quad k_1 k_2 \dots k_N = \mathcal{K}_\ell(d), \quad \ell = 1, \dots, J(d), \end{aligned}$$

where

$$\begin{aligned} J_{bii}^{k_1 \dots (k_i-2) \dots k_N} &= 0 \quad \text{for} \quad k_i - 2 < 0, \\ J_{bij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} &= 0 \quad \text{for} \quad k_i - 1 < 0 \quad \text{or} \quad k_j - 1 < 0, \end{aligned}$$

and  $\Lambda_\ell$ ,  $\ell = 1, \dots, J(d+2)$ , are the same as in Lemma 2.

*Proof.* The proof is similar to that for Theorem 5, and omitted here.  $\square$

*Remark 3.* For each fixed  $d$ , the condition given by Theorem 7 is sufficient to assure the stability of  $\mathcal{A}$ , which is less conservative than that given by Lemma 2 (see [20]), and allows faster convergence to the necessary conditions for the existence of an affine parameter-dependent Lyapunov function in terms of  $d$ . Moreover, similarly to Theorem 6, it can be shown that the family of LMI conditions given by Theorem 7 is of increasing precision.

For the relationship between Theorems 5 and 7, we have the following theorem.

**THEOREM 8.** *For a fixed  $d$ , if condition (i) (resp., (ii)) of Theorem 5 holds, then the condition of Theorem 7 for assuring the Hurwitz (resp., Schur) stability of  $\mathcal{A}$  holds.*

*Proof.* First, consider the Hurwitz stability of  $\mathcal{A}$ .

For a fixed  $d$ , if condition (i) of Theorem 5 holds, then there exist symmetric positive matrices  $P_i$ ,  $i = 1, \dots, N$ ,  $J_{ii}^{k_1 k_2 \dots k_N}$  and matrices  $J_{ij}^{k_1 k_2 \dots k_N} = (J_{ji}^{k_1 k_2 \dots k_N})^T$ ,  $k_1 k_2 \dots k_N = \mathcal{K}_\ell(d)$  for  $\ell = 1, \dots, J(d)$  such that (18) and (19) hold. Equation (18) can be rewritten as follows:

$$\begin{aligned} & \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) \text{He}(A_i^T P_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \text{He}(A_i^T P_j + A_j^T P_i) \\ & + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \dots (k_i-2) \dots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} \right) < 0, \\ & k_1 k_2 \dots k_N = \mathcal{K}_\ell(d+2) \quad \text{for } \ell = 1, \dots, J(d+2). \end{aligned}$$

Then there exists a symmetric positive matrix  $G$  such that

$$\begin{aligned} (31) \quad & \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) \text{He}(A_i^T P_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \text{He}(A_i^T P_j + A_j^T P_i) \\ & + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \dots (k_i-2) \dots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} \right) \\ & + \eta^{-1} \left( \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) A_i + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) (A_i + A_j) \right)^T G \\ & \times \left( \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) A_i + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) (A_i + A_j) \right) < 0, \\ & k_1 k_2 \dots k_N = \mathcal{K}_\ell(d+2) \quad \text{for } \ell = 1, \dots, J(d+2), \end{aligned}$$

where

$$\eta = 2 \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) + 4 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1).$$

Applying the Schur complement to (31), we then have

$$\begin{aligned} (32) \quad & \begin{bmatrix} R_{11}^{k_1 k_2 \dots k_N} & * \\ R_{21}^{k_1 k_2 \dots k_N} & R_{22}^{k_1 k_2 \dots k_N} \end{bmatrix} < 0, \quad k_1 k_2 \dots k_N = \mathcal{K}_\ell(d+2) \\ & \text{for } \ell = 1, \dots, J(d+2), \end{aligned}$$

where

$$\begin{aligned}
 R_{11}^{k_1 k_2 \dots k_N} &= \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) \text{He}(A_i^T P_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \text{He}(A_i^T P_j + A_j^T P_i) \\
 &\quad + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \dots (k_i-2) \dots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} \right), \\
 R_{21}^{k_1 k_2 \dots k_N} &= G \left( \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) A_i + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) (A_i + A_j) \right), \\
 R_{22}^{k_1 k_2 \dots k_N} &= -\eta G = - \left( 2 \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) + 4 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \right) G.
 \end{aligned}$$

Let

$$(33) \quad X_{i1} = P_i, \quad X_{i2} = G;$$

then  $R_{11}^{k_1 k_2 \dots k_N}$ ,  $R_{21}^{k_1 k_2 \dots k_N}$ , and  $R_{22}^{k_1 k_2 \dots k_N}$  can be rewritten as follows:

$$\begin{aligned}
 R_{11}^{k_1 k_2 \dots k_N} &= \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) \text{He}(A_i^T X_{i1}) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \text{He}(A_i^T X_{j1} + A_j^T X_{i1}) \\
 &\quad + \sum_{1 \leq i \leq N} J_{ii}^{k_1 \dots (k_i-2) \dots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} \right), \\
 R_{21}^{k_1 k_2 \dots k_N} &= \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) X_{i2} A_i + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) (X_{j2} A_i + X_{i2} A_j), \\
 R_{22}^{k_1 k_2 \dots k_N} &= \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) \text{He}(-X_{i2}) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \text{He}(-X_{i2} - X_{j2}).
 \end{aligned}$$

By (32), it follows that

$$\begin{aligned}
 (34) \quad &\left[ \begin{array}{c} \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) \text{He}(A_i^T X_{i1}) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \text{He}(A_i^T X_{j1} + A_j^T X_{i1}) \\ \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) X_{i2} A_i + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) (X_{j2} A_i + X_{i2} A_j) \\ \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) \text{He}(-X_{i2}) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \text{He}(-X_{i2} - X_{j2}) \\ + \left[ \begin{array}{c} \sum_{1 \leq i \leq N} J_{ii}^{k_1 \dots (k_i-2) \dots k_N} + \text{He} \left( \sum_{1 \leq i < j \leq N} J_{ij}^{k_1 \dots (k_i-1) \dots (k_j-1) \dots k_N} \right) \\ 0 \end{array} \right] \end{array} \right] \begin{array}{c} * \\ 0 \\ 0 \end{array} \\
 &= \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) \left[ \begin{array}{cc} \text{He}(A_i^T X_{i1}) & * \\ X_{i2} A_i & -\text{He}(X_{i2}) \end{array} \right]
 \end{aligned}$$

$$\begin{aligned}
& + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \begin{bmatrix} \text{He}(A_i^T X_{j1} + A_j^T X_{i1}) & * \\ X_{j2} A_i + X_{i2} A_j & -\text{He}(X_{i2} + X_{j2}) \end{bmatrix} \\
& + \sum_{1 \leq i \leq N} \begin{bmatrix} J_{ii}^{k_1 \cdots (k_i-2) \cdots k_N} & 0 \\ 0 & 0 \end{bmatrix} \\
& + \sum_{1 \leq i < j \leq N} \text{He} \left( \begin{bmatrix} J_{ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} & 0 \\ 0 & 0 \end{bmatrix} \right) \\
& = \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) \begin{bmatrix} \text{He}(A_i^T X_{i1}) & * \\ X_{i2} A_i & -\text{He}(X_{i2}) \end{bmatrix} \\
& + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \begin{bmatrix} \text{He}(A_i^T X_{j1} + A_j^T X_{i1}) & * \\ X_{j2} A_i + X_{i2} A_j & -\text{He}(X_{i2} + X_{j2}) \end{bmatrix} \\
& + \sum_{1 \leq i \leq N} J_{a ii}^{k_1 \cdots (k_i-2) \cdots k_N} + \sum_{1 \leq i < j \leq N} \text{He}(J_{a ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N}) < 0, \\
& k_1 k_2 \cdots k_N = \mathcal{C}_\ell(d+2) \quad \text{for } \ell = 1, \dots, J(d+2),
\end{aligned}$$

where

$$J_{a ij}^{k_1 k_2 \cdots k_N} = \begin{bmatrix} J_{ij}^{k_1 k_2 \cdots k_N} & 0 \\ 0 & 0 \end{bmatrix}, \quad 1 \leq i \leq j \leq N.$$

From (33) and (34), we then can obtain

$$\begin{aligned}
(35) \quad & \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) \begin{bmatrix} \text{He}(A_i^T X_{i1}) & * \\ X_{i2} A_i + P_i - X_{i1}^T & -\text{He}(X_{i2}) \end{bmatrix} \\
& + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) \\
& \quad \times \begin{bmatrix} \text{He}(A_i^T X_{j1} + A_j^T X_{i1}) & * \\ X_{j2} A_i + X_{i2} A_j + P_i + P_j - X_{i1}^T - X_{j1}^T & -\text{He}(X_{i2} + X_{j2}) \end{bmatrix} \\
& + \sum_{1 \leq i \leq N} J_{a ii}^{k_1 \cdots (k_i-2) \cdots k_N} + \sum_{1 \leq i < j \leq N} \text{He}(J_{a ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N}) < 0, \\
& k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d+2) \quad \text{for } \ell = 1, \dots, J(d+2).
\end{aligned}$$

Let

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix}.$$

By (35), we have

$$\begin{aligned}
(36) \quad & \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2) (Q_i + X_i B_i + B_i^T X_i^T) \\
& + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1) (Q_i + Q_j + X_j B_i + B_i^T X_j^T + X_i B_j + B_j^T X_i^T)
\end{aligned}$$

$$\begin{aligned}
 & + \sum_{1 \leq i \leq N} J_{a ii}^{k_1 \cdots (k_i-2) \cdots k_N} + \sum_{1 \leq i < j \leq N} \operatorname{He}(J_{a ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N}) < 0, \\
 & k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d+2) \quad \text{for } \ell = 1, \dots, J(d+2),
 \end{aligned}$$

where  $B_i$  are the same as in (7), and  $Q_i$  are the same as in (8).

From (36), there exists a sufficiently small  $\epsilon > 0$  such that

$$\begin{aligned}
 (37) \quad & \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2)(Q_i + X_i B_i + B_i^T X_i^T) \\
 & + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1)(Q_i + Q_j + X_j B_i + B_i^T X_j^T + X_i B_j + B_j^T X_i^T) \\
 & + \sum_{1 \leq i \leq N} J_{a ii}^{k_1 \cdots (k_i-2) \cdots k_N} + \sum_{1 \leq i < j \leq N} \operatorname{He}(J_{a ij}^{k_1 \cdots (k_i-2) \cdots k_N}) + N\epsilon I \\
 = & \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2)(Q_i + X_i B_i + B_i^T X_i^T) \\
 & + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1)(Q_i + Q_j + X_j B_i + B_i^T X_j^T + X_i B_j + B_j^T X_i^T) \\
 & + \sum_{1 \leq i \leq N} (J_{a ii}^{k_1 \cdots (k_i-2) \cdots k_N} + \epsilon I) + \sum_{1 \leq i < j \leq N} \operatorname{He}(J_{a ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N}) < 0, \\
 & k_1 k_2 \cdots k_N = \mathcal{K}_\ell(d+2) \quad \text{for } \ell = 1, \dots, J(d+2).
 \end{aligned}$$

For  $k_1 k_2 \cdots k_N \in \mathcal{K}(d+2)$ , let

$$\begin{aligned}
 (38) \quad & J_{b ii}^{k_1 \cdots (k_i-2) \cdots k_N} = \begin{cases} J_{a ii}^{k_1 \cdots (k_i-2) \cdots k_N} + \epsilon I & \text{if } k_i - 2 \geq 0, \\ 0 & \text{otherwise,} \end{cases} \quad 1 \leq i \leq N, \\
 & J_{b ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N} = J_{a ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N}, \quad 1 \leq i < j \leq N.
 \end{aligned}$$

Then

$$\sum_{1 \leq i \leq N} J_{b ii}^{k_1 \cdots (k_i-2) \cdots k_N} \leq \sum_{1 \leq i \leq N} (J_{a ii}^{k_1 \cdots (k_i-2) \cdots k_N} + \epsilon I).$$

By (37), it follows that

$$\begin{aligned}
 (39) \quad & \sum_{i=1}^N \mathcal{C}_i^\ell(d, 2)(Q_i + X_i B_i + B_i^T X_i^T) \\
 & + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathcal{C}_{ij}^\ell(d, 1, 1)(Q_i + Q_j + X_j B_i + B_i^T X_j^T + X_i B_j + B_j^T X_i^T) \\
 & + \sum_{1 \leq i \leq N} J_{b ii}^{k_1 \cdots (k_i-2) \cdots k_N} + \sum_{1 \leq i < j \leq N} \operatorname{He}(J_{b ij}^{k_1 \cdots (k_i-1) \cdots (k_j-1) \cdots k_N}) < 0;
 \end{aligned}$$

i.e., (29) holds. Moreover, from Lemma 4, (19), and (38), it follows that (30) holds. Thus, the condition of Theorem 7 is satisfied for assuring the Hurwitz stability of  $\mathcal{A}$ . For the case of Schur stability, the proof is similar, and thus is omitted here.  $\square$

TABLE 1  
Numbers of LMIs and decision variables.

Methods	Number of LMIs	Number of decision variables
Lemma 1 (i) [20]	$J(d+2)$	$nN(n+1)/2$
Lemma 1 (ii) [20]	$J(d+3)$	$nN(n+1)/2$
Theorem 5 (i)	$J(d+2)+J(d)$	$nN(n+1)/2+nN(nN+1)J(d)/2$
Theorem 5 (ii)	$J(d+3)+J(d+1)$	$nN(n+1)/2+nN(nN+1)J(d+1)/2$
Lemma 2 [20]	$J(d+2)$	$nN(n+1)/2+4n^2N$
Theorem 7	$J(d+2)+J(d)$	$nN(n+1)/2+4n^2N+nN(2nN+1)J(d)$

*Remark 4.* Theorem 8 shows that for a fixed  $d$ , the condition given by Theorem 7 is less conservative than that given by Theorem 5, which allows faster convergence to the necessary conditions for the existence of an affine parameter-dependent Lyapunov function in terms of  $d$ . But, more computational time for the stability test given by Theorem 7 is required for each  $d$ , since more decision variables are involved in the condition given by Theorem 7.

*Remark 5.* Regarding computational complexity, the number of decision variables and the number of LMIs are used as measures of computational burden. Table 1 summarizes the numbers of decision variables and LMIs for each method with a fixed  $d$ . From the table, it is easy to see that Theorems 5 and 7 require more decision variables and LMIs than Lemmas 1 and 2, respectively, which implies that the methods given by Theorems 5 and 7 are of more computational burden for a given  $d$ .

**4. Example.** In this section, two numerical examples are given to illustrate the effectiveness of the robust stability criteria given by Theorems 5 and 7.

*Example 1.* Consider an uncertain matrix  $\mathcal{A}$  described by (1) with  $N = 2$ ,

$$A_1 = \begin{bmatrix} -0.0996 & 0.9846 & -0.4496 \\ -0.9045 & -0.0387 & 0.9657 + \delta \\ 0.6933 & -0.7612 & -0.4179 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -0.6327 & -0.3142 & 0.8716 \\ 0.6263 & -0.6932 & 0.9598 \\ -0.0090 & -0.9367 & 0.3422 \end{bmatrix}.$$

It is concerned with the robust Hurwitz stability for  $|\delta| \leq \gamma$ . Applying Lemmas 1 and 2 and Theorems 5 and 7 to this example, the stability bounds and cpu times are obtained as shown in Tables 2–6.

TABLE 2  
Stability bounds and cpu times for Example 1 by Lemma 1.

$d$	0	1	2	3	4	5	
Stability bound	infeasible	infeasible	infeasible	infeasible	0.0019	0.0043	
Cpu time (sec)	0.0313	0.3438	0.4844	0.7031	0.1701	0.1875	
$d$	6	7	8	9	10	11	12
Stability bound	0.0064	0.0070	0.0077	0.0079	0.0082	0.00827	0.00833
Cpu time (sec)	0.2656	0.3438	0.3440	0.3594	0.5313	0.6250	0.5469
$d$	13	14	15	16	17	18	
Stability bound	0.00836	0.00839	0.00840	0.00841	0.008420	0.008427	
Cpu time (sec)	0.5938	0.6375	0.7188	0.7500	0.8750	0.8906	
	$d$		19	20	21		
	Stability bound		0.008429	0.00843	0.00843		
	Cpu time (sec)		0.9375	0.9688	1.0469		

TABLE 3  
*Stability bounds and cpu times for Example 1 by Lemma 2.*

$d$	0	1
Stability bound	0.00843	0.00843
Cpu time (sec)	0.2498	0.3281

TABLE 4  
*Stability bounds and cpu times for Example 1 by Theorem 5.*

$d$	0	1
Stability bound	0.00843	0.00843
Cpu time (sec)	0.1875	0.4291

TABLE 5  
*Stability bounds and cpu times for Example 1 by Theorem 7.*

$d$	0	1
Stability bound	0.00843	0.00843
Cpu time (sec)	1.4219	5.7813

TABLE 6  
*The obtained stability bounds and the total cpu time for Example 1.*

Methods	Lemma 1	Lemma 2	Theorem 5	Theorem 7
Stability bound	0.00843	0.00843	0.00843	0.00843
Total cpu time (sec)	12.3551	0.5779	0.6166	7.8750

From the above computational results, it can be seen that Theorems 5 and 7 cannot provide more conservative results than Lemmas 1 and 2 for the same  $d$ . However, Theorems 5 and 7 require more computational time for each  $d$  than Lemmas 1 and 2, respectively. Table 6 shows that Lemma 2 performs best for the example since it has the lowest consumed cpu time for obtaining the stability bound of 0.00843.

*Example 2.* Consider an uncertain matrix  $\mathcal{A}$  described by (1) with  $N = 3$ ,

$$A_1 = \begin{bmatrix} -0.646 & -0.875 & -0.997 \\ -0.732 & -0.993 & -0.126 \\ 0.707 & 0.289 & -0.3 + \delta \end{bmatrix}, \quad A_2 = \begin{bmatrix} -0.419 & 0.896 & -0.854 \\ -0.198 & -0.417 & 0.592 \\ 0.574 & 0.113 & -0.970 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} -0.789 & -0.533 & 0.353 \\ -0.469 & -0.390 & 0.676 \\ -0.970 & -0.914 & 0.053 \end{bmatrix}.$$

It also is concerned with the robust Hurwitz stability for  $|\delta| \leq \gamma$ . Applying Lemmas 1 and 2 and Theorems 5 and 7 to this example, the stability bounds and cpu times are obtained, as shown in Tables 7–11.

From the computational results, it can be seen that Theorems 5 and 7 can provide less conservative results than Lemmas 1 and 2 for the same  $d$ , respectively. For this example, Theorem 7 performs best for the example, since the total cpu times consumed by using Lemmas 1 and 2 and Theorem 5 for obtaining the smaller stability bounds have exceeded the total consumed cpu time of 656.3907 seconds by using Theorem 7 for obtaining the stability bound of 0.27025.

These two examples show that, in general, Lemma 1 (resp., Lemma 2) will usually perform better than Theorem 5 (resp., Theorem 7) for obtaining a stability bound if it can be achieved by using Lemma 1 (resp., Lemma 2) for a smaller  $d$ . But, if the stability bound is achieved by using Lemma 1 (resp., Lemma 2) for a much larger  $d$  than that used by Theorem 5 (resp., Theorem 7), then Theorem 5 (resp., Theorem 7) will be preferred, because the programming for a larger  $d$  also is costly in terms of time.

TABLE 7  
*Stability bounds and cpu times for Example 2 by Lemma 1.*

$d$	0	1	2	3	4	5	6
Stability bound	0.2444	0.2482	0.2504	0.2520	0.2547	0.2571	0.2584
Cpu time (sec)	0.0781	0.2031	0.6719	1.5625	1.6406	2.6563	3.6875
$d$	7	8	9	10	11	12	
Stability bound	0.2594	0.2608	0.2620	0.2626	0.2630	0.2636	
Cpu time (sec)	5.3594	7.1719	8.5781	10.9531	13.9219	18.2813	
$d$	13	14	15	16	17	18	
Stability bound	0.2641	0.2645	0.2649	0.2653	0.2657	0.2658	
Cpu time (sec)	23.4622	30.0269	34.7135	39.2688	45.7103	52.6619	
$d$	19	20	21	22	23	24	
Stability bound	0.2660	0.2662	0.2663	0.2665	0.2667	0.2668	
Cpu time (sec)	59.5203	61.6933	79.8125	100.0139	125.9238	152.0312	
$d$	25						
Stability bound	0.26698						
Cpu time (sec)	178.8125						

TABLE 8  
*Stability bounds and cpu times for Example 2 by Lemma 2.*

$d$	0	1	2	3	4	5	6
Stability bound	0.2647	0.2657	0.2665	0.2669	0.2672	0.2673	0.2676
Cpu time (sec)	0.8750	1.6563	2.6094	4.0938	7.0625	9.4219	13.5156
$d$	7	8	9	10	11	12	
Stability bound	0.26793	0.26815	0.26824	0.26831	0.26845	0.26854	
Cpu time (sec)	26.1406	42.5156	53.2969	68.3594	85.3906	98.7500	
$d$	13	14	15	16	17	18	
Stability bound	0.268663	0.26871	0.26885	0.26891	0.26897	0.26900	
Cpu time (sec)	109.2193	121.6217	136.7125	151.3281	169.9214	188.3438	

TABLE 9  
*Stability bounds and cpu times for Example 2 by Theorem 5.*

$d$	0	1	2	3	4	5
Stability bound	0.26983	0.26992	0.26999	0.270049	0.27009	0.27011
Cpu time (sec)	0.5938	2.7969	34.0156	73.1563	221.4137	520.6406

TABLE 10  
*Stability bounds and cpu times for Example 2 by Theorem 7.*

$d$	0	1
Stability bound	0.27025	0.27025
Cpu time (sec)	92.4219	563.9688

TABLE 11  
*The obtained stability bounds and the total cpu times for Example 2.*

Methods	Lemma 1	Lemma 2	Theorem 5	Theorem 7
Stability bound	0.26698	0.26900	0.27011	0.27025
Total cpu time (sec)	1058.4168	1290.8344	852.6169	656.3907



**5. Conclusion.** In this paper, new necessary and sufficient conditions for the existence of an affine parameter-dependent Lyapunov function assuring the Hurwitz or Schur stability of a polytopic system are given, which are composed of a family of LMI conditions of increasing precision. At each step, a set of LMIs provides sufficient conditions for the existence of the affine parameter-dependent Lyapunov function, and necessity is asymptotically attained. Compared with the results in [20], the new stability conditions provide less conservative tests at each step and allow faster convergence to the necessary conditions for the existence of an affine parameter-dependent Lyapunov function in terms of steps. But, it should be emphasized that the faster convergence is not in terms of time, since the numerical complexity of the new proposed conditions is much greater than that in [20]. The numerical examples have shown the effectiveness of the new proposed stability criteria.

## REFERENCES

- [1] S. BOYD, L. EL GHAOU, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.
- [2] L. WANG, *Robust stability of a class of polynomial families under nonlinearly correlated perturbations*, Syst. Control Lett., 30 (1997), pp. 25–30.
- [3] V. L. KHARITONOV, *Robust stability of nested polynomial families*, Automatica, 32 (1996), pp. 365–367.
- [4] M. C. DE OLIVEIRA, J. BERNUSSOU, AND J. C. GEROMEL, *A new discrete-time robust stability condition*, Syst. Control Lett., 37 (1999), pp. 261–265.
- [5] M. W. MCCONLEY, M. A. DAHLEH, AND E. FERON, *Polytopic control Lyapunov functions for robust stabilization of a class of nonlinear systems*, Syst. Control Lett., 34 (1998), pp. 77–83.
- [6] U. SHAKED, *Improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty*, IEEE Trans. Automat. Control, 46 (2001), pp. 652–656.
- [7] Y.-Y. CAO AND Z. LIN, *A descriptor system approach to robust stability analysis and controller synthesis*, IEEE Trans. Automat. Control, 49 (2004), pp. 2081–2084.
- [8] C. LIN, Q.-G. WANG, AND T. H. LEE, *A less conservative robust stability test for linear uncertain time-delay systems*, IEEE Trans. Automat. Control, 51 (2006), pp. 87–91.
- [9] D. C. W. RAMOS AND P. L. D. PERES, *An LMI condition for the robust stability of uncertain continuous-time linear systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 675–678.
- [10] P. GAHNET, A. NEMIROVSKI, A. J. LAUB, AND M. CHILALI, *LMI Control Toolbox User's Guide*, The Math Works Inc., Natick, MA, 1995.
- [11] J. F. STURM, *Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11/12 (1999), pp. 625–653. Also available online at [http://www.optimization-online.org/DB\\_HTML/2001/10/395.html](http://www.optimization-online.org/DB_HTML/2001/10/395.html).
- [12] J. GEROMEL, M. C. DE OLIVEIRA, AND L. HSU, *LMI characterization of structural and robust stability*, Linear Algebra Appl., 285 (1998), pp. 69–80.
- [13] S. ZHOU, J. LAM, AND G. FENG, *New characterization of positive realness and control of a class of uncertain polytopic discrete-time systems*, Syst. Control Lett., 54 (2005), pp. 417–427.
- [14] V. J. S. LEITE AND P. L. D. PERES, *An improved LMI condition for robust D-stability of uncertain polytopic systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 500–504.
- [15] V. J. S. LEITE, R. C. L. F. OLIVEIRA, R. J. DE OLIVEIRA, AND R. L. D. PERES, *D-stability of polytopes of polynomial matrices: Characterization through LMIs*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas, 2004, pp. 863–868.
- [16] G. CHESI, A. GARULLI, A. TESI, AND A. VICINO, *Robust stability of polytopic systems via polynomially parameter-dependent Lyapunov functions*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 4670–4675.
- [17] P.-A. BLIMAN, *A convex approach to robust stability for linear systems with uncertain scalar parameters*, SIAM J. Control Optim., 42 (2004), pp. 2016–2042.
- [18] R. C. L. F. OLIVEIRA AND P. L. D. PERES, *LMI conditions for robust stability analysis based on polynomially parameter-dependent Lyapunov functions*, Syst. Control Lett., 55 (2006), pp. 52–61.

- [19] V. J. S. LEITE AND P. L. D. PERES, *Robust control through piecewise Lyapunov functions for discrete time-varying uncertain systems*, Int. J. Control, 77 (2004), pp. 230–238.
- [20] R. C. L. F. OLIVEIRA AND P. L. D. PERES, *Stability of polytopes of matrices via affine parameter-dependent Lyapunov functions: Asymptotically exact LMI conditions*, Linear Algebra Appl., 405 (2005), pp. 209–228.
- [21] G. H. HARDY, J. E. LITTLEWOOD, AND G. PÓLYA, *Inequalities*, 2nd ed., Cambridge University Press, Cambridge, UK, 1952.

## MESH INDEPENDENCE OF KLEINMAN–NEWTON ITERATIONS FOR RICCATI EQUATIONS IN HILBERT SPACE\*

J. A. BURNS<sup>†</sup>, E. W. SACHS<sup>‡</sup>, AND L. ZIETSMAN<sup>†</sup>

**Abstract.** In this paper we consider the convergence of the infinite dimensional version of the Kleinman–Newton algorithm for solving the algebraic Riccati operator equation associated with the linear quadratic regulator problem in a Hilbert space. We establish mesh independence for this algorithm and apply the result to systems governed by delay equations. Numerical examples are presented to illustrate the results.

**Key words.** Riccati, Kleinman–Newton, mesh independence

**AMS subject classifications.** 49, 65, 93

**DOI.** 10.1137/060653962

**1. Introduction.** The problem of constructing numerical schemes for optimization-based design and control of infinite dimensional systems leads to technical and practical issues that are not present if one is interested only in simulation. For example, if one uses finite elements or the method of lines to simulate a system of partial differential equations (PDEs), then the resulting finite dimensional approximate system is often very large and can have millions of state variables. The corresponding approximating Riccati equations are immense, and special numerical techniques are required to solve such equations. Many of these large-scale Riccati solvers are based on iterative algorithms (see [9], [10], and [30]) and take advantage of the mathematical structure of the approximating system (symmetry, sparseness, etc.).

There are two basic issues that need to be addressed in developing practical numerical approximations for control. First, it is essential that the approximation scheme leads to finite dimensional approximating Riccati equations that converge (under mesh refinement) to the solution of the infinite dimensional Riccati equation. This is a well-studied problem (see [7], [14], [26], [33], and [43]). It is now well known that to obtain norm convergence for the Riccati equation, the approximation scheme must satisfy some form of convergence, dual convergence, and uniform preservation of stabilizability and detectability under mesh refinement (see [7] and [33]). These concepts will be made more precise in section 7.1. The important point here is that many “standard” convergent approximation schemes do not satisfy all the conditions necessary for norm convergence of the Riccati operators (see [16]). If this issue is ignored when one develops an approximation scheme for control design and optimization, then the resulting numerical algorithm can fail to produce accurate and useful results (see the numerical examples in section 9). In this paper we show that these properties are also key ingredients in establishing mesh independence of Newton-type algorithms.

---

\*Received by the editors March 9, 2006; accepted for publication (in revised form) June 2, 2008; published electronically October 22, 2008. This research was supported in part by the Air Force Office of Scientific Research under grant F49620-03-1-0243 and by the DARPA Special Projects Office.

<http://www.siam.org/journals/sicon/47-5/65396.html>

<sup>†</sup>Interdisciplinary Center for Applied Mathematics, Virginia Tech, Blacksburg, VA 24061 (jaburns@vt.edu, lzietsma@vt.edu).

<sup>‡</sup>Fachbereich IV, Abteilung Mathematik, Universität Trier, 54286 Trier, Germany, and Interdisciplinary Center for Applied Mathematics, Virginia Tech, Blacksburg, VA 24061 (sachs@uni-trier.de, esachs@vt.edu).

The second important issue is concerned with the development of an effective algorithm for the numerical solution of the (large-scale) finite dimensional Riccati equations that arise once the problem has been discretized. During the past five years considerable attention has been devoted to the problem of developing accurate and fast numerical methods for control of large-scale systems. In 2004, the first issue of the *IEEE Control Systems Magazine* (Volume 24, Issue 1) was devoted to this topic. Much of the motivation for this emphasis comes from the fact that such systems often arise as discretizations of control problems with PDEs as the governing system. The observation that these large-scale finite dimensional Riccati equations come from discretizations of PDE control systems makes it possible to exploit special algorithms such as multigrid techniques (see [42] and [45]) and parallel iterative solvers (see [30]). Considerable progress has been made at this level by Benner and Saak (see [9] and [10]). Also, Grasedyck, Hackbusch, and Khoromskij (see [30] and the references therein) have developed impressive computational algorithms for Riccati and Lyapunov equations that arise in these cases. Many of these large-scale Riccati solvers are based on iterative algorithms.

It is impossible to address all the potential problems in constructing approximation schemes for optimal control of infinite dimensional systems in a single paper, so we limit our discussion to the well-studied linear quadratic optimal control problem and show how specific approximation assumptions are needed to address convergence and efficiency of an algorithm. In particular, we focus on convergence and mesh independence of the Kleinman–Newton algorithm for solving the operator Riccati equation defined by the linear quadratic regulator (LQR) problem. This problem is simple enough to allow for a rather complete analysis of convergence and mesh independence and yet complex enough to illustrate how both convergence and mesh independence might fail for perfectly good “standard” (convergent) numerical approximations.

**2. A short review of the mesh independence principle.** There are two basic aspects of the mesh independence principle (MIP) for Newton-type methods (see [1] and [2]). Roughly speaking, the MIP may be broken down into convergence under mesh refinement of the Newton iteration counts on a given mesh.

Let  $\mathcal{F} : D(\mathcal{F}) \subseteq E \longrightarrow E$  be a nonlinear operator on an infinite dimensional Hilbert space  $E$ , and consider the equation

$$(2.1) \quad \mathcal{F}(x) = 0.$$

Let  $E^N \subseteq E$  be a sequence of finite dimensional approximating spaces, and consider the sequence of discretized equations

$$(2.2) \quad \mathcal{F}^N(x^N) = 0,$$

where  $\mathcal{F}^N : D(\mathcal{F}^N) \subseteq E^N \longrightarrow E^N$ . In this paper we are interested in the problem of solving the Riccati operator equation associated with LQR feedback control of systems governed by delay and PDEs. In this setting, (2.1) is an infinite dimensional Riccati equation defined by a PDE control system, and (2.2) is an approximating Riccati equation obtained by some type of finite element or finite difference scheme applied to the PDE system. Here  $N$  is related to the size of the mesh used to define the discretized equations on a grid. Assume that (2.1) and (2.2) have unique solutions  $x_\infty \in D(\mathcal{F})$  and  $x_\infty^N \in D(\mathcal{F}^N)$ , respectively. We say that the *approximation scheme converges* if

$$(2.3) \quad \lim_{N \rightarrow +\infty} \|x_\infty^N - P^N x_\infty\|_{E^N} = 0,$$

where  $P^N : E \rightarrow E^N$  is the orthogonal projection of  $E$  onto  $E^N$ .

Now assume that one applies a Newton-type algorithm to the infinite dimensional problem (2.1) and that the same algorithm is also applied to the corresponding finite dimensional approximate problem (2.2). For the moment assume both schemes produce quadratically convergent iterations  $x_k$  and  $x_k^N$ ,  $k = 1, 2, \dots$ . For a given  $\varepsilon > 0$ ,  $x_0 \in D(\mathcal{F})$  and  $x_0^N \in D(\mathcal{F}^N)$  define the numbers  $M(\varepsilon, x_0)$  and  $M^N(\varepsilon, x_0^N)$  by

$$M(\varepsilon, x_0) \triangleq \inf\{k : \|x_k - x_\infty\| < \varepsilon\} \quad \text{and} \quad M^N(\varepsilon, x_0^N) \triangleq \inf\{k : \|x_k^N - x_\infty^N\|_{E^N} < \varepsilon\},$$

respectively. Here,  $x_0$  and  $x_0^N$  are the starting values for the iterations. The (strong) MIP (see Theorem 2.1 in [2]) takes the form

$$(2.4) \quad M(\varepsilon, x_0) = M^N(\varepsilon, P^N x_0) + \tau(N),$$

where  $\tau(N) \rightarrow 0$  as  $N \rightarrow +\infty$ . Also, assume there are constants  $c$  and  $c^N$  such that

$$(2.5) \quad \|x_{k+1} - x_\infty\| \leq c \|x_k - x_\infty\|^2$$

and

$$(2.6) \quad \|x_{k+1}^N - x_\infty^N\|_{E^N} \leq c^N \|x_k^N - x_\infty^N\|_{E^N}^2,$$

respectively. Let  $\hat{c}$  and  $\hat{c}^N$  be the minimal values of  $c$  and  $c^N$  that satisfy (2.5) and (2.6), where  $x_0^N = P^N x_0$ . As noted in [2], since  $P^N : E \rightarrow E^N$  is the orthogonal projection of  $E$  onto  $E^N$ , in some cases one can show that another form of the strong MIP is given by

$$(2.7) \quad \hat{c}^N = \hat{c} + \gamma(N),$$

where  $\gamma(N) \rightarrow 0$  as  $N \rightarrow +\infty$ . The basic idea behind these strong versions of mesh independence is that the number of iterations required to achieve a given error tolerance is independent of the mesh size and asymptotically converges to the number of infinite dimensional iterations (theoretically) required to attain the same tolerance. A weaker form of the MIP would require only that, if one has the estimates (2.5) and (2.6), then

$$(2.8) \quad c^N = c + \delta(N),$$

where  $\delta(N) \rightarrow 0$  as  $N \rightarrow +\infty$ . Although the constants  $c$  and  $c^N$  are not the minimal values, it follows that the number of iterations required to solve the discretized equations  $\mathcal{F}^N(x^N) = 0$  is essentially independent of the mesh size.

**3. Mesh independence for the infinite dimensional Riccati equation.** In this paper we focus on the case where the nonlinear function  $\mathcal{F} = \mathcal{F}(\Pi)$  is defined by an infinite dimensional Riccati operator equation of the form

$$(3.1) \quad \mathcal{F}(\Pi) = A^* \Pi + \Pi A - \Pi B B^* \Pi + C^* C = 0,$$

where  $A$  generates a strongly continuous semigroup on a Hilbert space  $H$ . Here  $\mathcal{F} : D(\mathcal{F}) \subseteq E \rightarrow E$ , where  $E$  is the space of bounded linear operators on  $H$ .

*Remark 3.1.* It is important to note that in most applications the operator  $A$  is unbounded, and even if the  $B$  and  $C$  operators are bounded, the nonlinear operator

$\mathcal{F}$  will not be continuous on its domain. Therefore,  $\mathcal{F}$  will not have a Lipschitz continuous Fréchet derivative, and the analysis used in [1] and [2] is not directly applicable. In particular, convergence proofs for the infinite dimensional Newton algorithm that depend on the existence of the Fréchet derivative cannot be used in this setting. As noted by Damm and Hinrichsen in [23], the existence of the Fréchet derivative can be relaxed if one works in ordered Banach spaces. Indeed, they provide a general convergence result under the blanket assumptions that  $E$  is ordered by a closed, solid, regular convex cone and that  $\mathcal{F}$  is continuous on its domain (see page 50 in [23]). Moreover, even when using the ordered space approach, we see that Fréchet differentiability was needed to obtain quadratic convergence of the Newton method (see page 56 in [23]). However, for the delay systems below (and other PDE control systems) these assumptions do not hold.

In the finite dimensional case, the “natural” space of operators is the set  $E = \mathcal{H}^n$  of  $n \times n$  Hermitian matrices with (Frobenius) trace norm. As noted in [23], if one uses the cone  $\mathcal{C} = \mathcal{H}_+^n = \{\Pi \in \mathcal{H}^n : \Pi \geq 0\}$ , then  $\mathcal{C}$  satisfies the blanket assumptions above. In an infinite dimensional setting, verifying these assumptions is nontrivial or impossible, depending on the choice of  $E$ . One might be tempted to use the infinite dimensional analogue and set  $E = \mathcal{H}$  to be the set of all trace class operators on the Hilbert space  $H$ . If the solution  $\Pi$  to (3.1) is not of trace class (see Example 1), then this is not a reasonable choice for  $E$ . Even if the solution is of trace class, one might still need to work in a larger space to develop practical approximation schemes for numerical solutions. In this setting, if  $E = \mathcal{L}(H, H)$  is the space of bounded linear operators on  $H$  and one sets  $\mathcal{C} = \mathcal{H}_+ = \{\Pi \in \mathcal{H} : \Pi \geq 0\}$  to be the cone of nonnegative definite trace operators, then  $\mathcal{C}$  is not solid. Hence, a direct application of the results in [23] is not possible.

In the case when the nonlinear equation (3.1) is a Riccati equation defining an LQR controller, it is possible to extend the finite dimensional proof of Mehrmann in [40] to a rather general class of infinite dimensional problems. We take this approach and present a complete convergence proof for the infinite dimensional Kleinman–Newton algorithm in the space  $E = \mathcal{L}(H, H)$ . Although this proof is similar in spirit to the results in [40], there are some technical details that require attention. Moreover, this approach provides explicit bounds and estimates that we later use to establish mesh independence. This is another reason we do not use the approach in [23] based on ordered spaces. The following example illustrates that even if  $A$ ,  $B$ , and  $C$  are bounded, the solution to the operator Riccati equation (3.1) does not have to be of trace class. Later we shall use this example to illustrate the importance of the compactness assumptions.

*Example 1.* Let  $H = \mathbb{R} \times L^2$ , and define the operators  $A$ ,  $B$ , and  $C$  on  $H$  by

$$A = \begin{bmatrix} -1 & 0 \\ 0 & -I \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \end{bmatrix}^T, \quad \text{and} \quad C = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & \sqrt{2}I \end{bmatrix},$$

respectively. Here  $I$  is the identity on  $L^2$ . By direct computation it follows that

$$\Pi = \begin{bmatrix} 1 & 0 \\ 0 & I \end{bmatrix}$$

is the solution to the Riccati equation (3.1). Since the identity operator is not compact,  $\Pi$  is not of trace class. On the other hand, for this simple example,  $\mathcal{F}$  is continuously Fréchet differentiable, and convergence of the infinite dimensional Kleinman–Newton algorithm follows directly from [1] and [2].

It is well known that under very mild conditions on the LQR problem, and assuming that the discretization scheme preserves the basic control system properties, both (2.1) and (2.2) have unique solutions in the set of nonnegative self-adjoint bounded linear operators. The issues to be resolved are as follows:

- (i) What conditions must be placed on the discretization scheme to guarantee that the solutions to the approximating equations (2.2) converge *in norm* to the solution of the infinite dimensional problem (2.1)? In some sense this is a classic numerical analysis problem. However, for the Riccati equation (3.1), the conditions for norm convergence are nontrivial and the best results are stated in terms of control system properties.
- (ii) Does the infinite dimensional Kleinman–Newton algorithm converge quadratically when applied to the infinite dimensional operator Riccati equation? As noted above, there are several approaches to this question. For applications to delay and PDE systems, our approach offers explicit bounds which are helpful in establishing mesh independence, and this approach does not require continuity of  $\mathcal{F}$ .
- (iii) Finally, what conditions must be placed on the discretization scheme to guarantee that the Kleinman–Newton algorithm satisfies MIP estimates of the form (2.4), (2.7), or (2.8)?

In this paper we focus on these issues. First, we give a brief review of what is known about convergence of discretization schemes for the infinite dimensional LQR control problem.

**4. A summary of approximation results for LQR control.** Most of the numerical schemes for approximating systems governed by PDEs developed during the past 50 years focused on methods that provided convergent and efficient simulations. However, the LQR problem is an optimal control problem on an infinite time interval, and it is possible to clearly identify two additional requirements that need to be placed on an approximation scheme to ensure convergence of the control design. Moreover, we shall show that these requirements also play a role in determining mesh independence of the Kleinman–Newton algorithm. In particular, dual convergence and preservation of exponential stability (POES) play central roles in both convergence and mesh independence. The POES condition was first introduced by Banks and Kunisch in [7] as a technical assumption needed to establish strong convergence of the Riccati operators for parabolic PDE control problems. This condition is equivalent to the uniform stabilizability defined in Assumption 7.3 below. In some cases one can relax the POES assumption and still obtain strong (and even norm) convergence of the Riccati operators. For example, the spline scheme developed for delay systems by Kappel and Salamon in [36] produced convergent Riccati operators even though POES was not satisfied (see [35] and [37]). Kappel and Salamon replaced the POES assumption with a uniform output and input-output stability condition and proved strong convergence of the Riccati operators. Ito [33], [34] used the version given in Assumption 7.3 to establish norm convergence. We will say more about this when we discuss the numerical results below.

In 1969 Sasai and Shimemura [47] was among the first researchers to recognize the importance of dual convergence for infinite dimensional LQR problems (also see [46] and [48]). Gibson (see [27], [28], [29]) established a general framework for developing approximation schemes for LQR problems and applied his results to control systems governed by delay and hyperbolic PDEs. If one is interested only in weak convergence of the functional gains, then dual convergence may not be essential (see [24] and

[25]). However, as observed in [8] and [13], weak convergence may not be sufficient for practical design, and, as shown in [16], not all standard schemes yield dual convergent algorithms. In particular, the finite element scheme developed in [5] is not dual convergent and does not produce strongly convergent functional gains.

At this point, there is no general method that can address the issue of dual convergence. However, for delay systems, two approaches have emerged. The first method is based on constructing numerical schemes that are dual convergent. The excellent survey by Kappel [35] focuses on this approach. A second approach is based on constructing separate numerical schemes for the forward problem and the dual problem. This is the approach carried out for delay equations by Germani, Manes, and Pepe in the paper [26]. It is important to note that extending any of these methods to other types of PDE-based systems is not a trivial exercise.

The key point here is that in order to develop numerical schemes for control of infinite dimensional systems, one must first ensure that certain control system properties are preserved under the approximation. Once this issue is resolved, it is important to consider the problem of numerically solving the finite dimensional problem. In particular, it is possible to construct several numerical schemes that preserve the required control system properties (stabilizability, detectability, etc.), but the resulting finite dimensional control problems may differ dramatically in conditioning and computational complexity.

In this paper we focus on an iterative method for solving the Riccati equations associated with LQR problems. We show that the infinite dimensional Kleinman–Newton iterations converge to the Riccati operator for the infinite dimensional problem, and we investigate mesh independence. Although the basic ideas used in the proof are similar to those found in papers on the finite dimensional problem, there are certain estimates that provide insight into general connections between preservation of control system properties, convergence, and mesh independence. These results provide a framework that can be employed to analyze specific finite dimensional approximations. We close with an application of these results to delay systems and a discussion of the averaging schemes found in [3] and [16] and the spline/finite element schemes given in [4], [5], and [8].

As noted above, the MIP does not make sense unless one has *norm convergence* of the discretized Riccati operators. It is known (see [16]) that the spline/finite element scheme for delay systems given in [5] does not produce norm convergent Riccati operators. This problem can also be seen in the much simpler Example 2 given in section 7.1 below.

**5. Problem setting and basics.** We consider the following LQR problem in an abstract Hilbert space setting. Let  $U$ ,  $H$ , and  $Y$  be Hilbert spaces over the reals. If  $Z$  and  $W$  are any two Hilbert spaces, then we denote by  $\mathcal{L}(Z, W)$  the linear space of linear bounded operators from  $Z$  into  $W$ . In the special case where  $W = Z$ , we set  $\mathcal{L}(Z) = \mathcal{L}(Z, Z)$ . The system equation is given in state space form by

$$(5.1) \quad \dot{z}(t) = Az(t) + Bu(t), \quad t \geq 0, \quad z(0) = z_0 \in H,$$

where  $A$  generates a strongly continuous semigroup on  $H$  and  $B \in \mathcal{L}(U, H)$ .

The assumption that  $B \in \mathcal{L}(U, H)$  implies we are considering only bounded input operators. Let  $C \in \mathcal{L}(H, Y)$ , and define the quadratic cost function  $J(u)$  by

$$(5.2) \quad J(u) = \int_0^\infty (\|Cz(s)\|^2 + \|u(s)\|^2) ds,$$



where  $z(s)$  is the solution to (5.1) for a given control  $u \in L^2(0, \infty; U)$ . The LQR control problem is to minimize the quadratic cost  $J(u)$  over all controls  $u \in L^2(0, \infty; U)$ .

It is well known (see [11], [12]) that under certain assumptions, the optimal control is given by state feedback  $u_{opt} = -Kz(t)$ , where

$$(5.3) \quad K = B^*X,$$

and  $X \in \mathcal{L}(H)$  is a solution of an abstract algebraic Riccati operator equation of the form

$$(5.4) \quad A^*X + XA - XBB^*X + C^*C = 0.$$

In order to formulate an abstract Newton method for solving this nonlinear operator equation in  $\mathcal{L}(H)$ , we first have to specify the appropriate mappings. Let  $\Sigma(H) = \{\Pi \in \mathcal{L}(H) : \Pi \text{ self-adjoint}\}$  be the space of self-adjoint bounded linear operators on  $H$ , and let  $\Sigma^+(H) = \{\Pi \in \Sigma(H) : (\Pi x, x) \geq 0 \text{ for all } x \in H\}$  denote the subspace of nonnegative operators in  $\Sigma(H)$ .

Since  $A$  is an unbounded operator in (5.4), we define a map  $\mathcal{A}$  which can be formally written as  $\mathcal{A}(\Pi) = A^*\Pi + \Pi A$  and can be defined rigorously as in [12, page 151]. In particular, for a given  $\Pi \in \Sigma(H)$ , set

$$\phi_\Pi(x, y) = (\Pi x, Ay) + (Ax, \Pi y), \quad x, y \in D(A),$$

and define

$$D(\mathcal{A}) = \{\Pi \in \Sigma(H) : \phi_\Pi \text{ can be extended to a continuous sesquilinear operator on } H \times H\}.$$

This unique extension of  $\phi_\Pi$  as a continuous sesquilinear form on  $H \times H$  will also be denoted by  $\phi_\Pi$ . For each  $\Pi \in D(\mathcal{A})$ , one can define a linear operator, denoted by  $\mathcal{A}(\Pi) \in \Sigma(H)$ , by the identity

$$(5.5) \quad (\mathcal{A}(\Pi)x, y) = \phi_\Pi(x, y) = (\Pi x, Ay) + (Ax, \Pi y), \quad x, y \in D(A), \quad \Pi \in D(\mathcal{A}).$$

Therefore, we have defined a linear operator  $\mathcal{A} : D(\mathcal{A}) \subset \Sigma(H) \rightarrow \Sigma(H)$ , and in [12, page 152], it is shown that for  $\Pi \in D(\mathcal{A})$  and  $x \in D(A)$  one has  $\Pi x \in D(A^*)$  and

$$(5.6) \quad \mathcal{A}(\Pi)x = A^*\Pi x + \Pi Ax.$$

The previous notation allows us to precisely define a solution of the abstract Riccati operator equation.

**DEFINITION 5.1.** *The bounded linear operator  $X$  is called a strict solution of the Riccati equation (5.4) if  $X \in D(\mathcal{A})$  and*

$$(5.7) \quad \mathcal{A}(X) - XBB^*X + C^*C = 0.$$

*The bounded linear operator  $X$  is called a weak solution of the Riccati equation (5.4) if  $X \in D(\mathcal{A})$  and*

$$(5.8) \quad (Xx, Ay) + (Ax, Xy) - (B^*Xx, B^*Xy) + (Cx, Cy) = 0, \quad x, y \in D(A).$$

It is shown on page 262 in [12] that for  $X \in \Sigma^+(H)$  a strict solution is equivalent to a weak solution of (5.4). Although we might use the notation of the operator equation, in this paper we deal with weak solutions. The existence and uniqueness of

solutions to the Riccati operator equation are not necessarily guaranteed. We follow the definitions and notation in [12].

DEFINITION 5.2. (i) *The system  $(A, B)$  is called stabilizable if there exists a bounded linear operator  $K : H \rightarrow U$  such that  $(A - BK)$  generates an exponentially stable  $C_0$ -semigroup on  $H$ .*

(ii) *The pair  $(A, C)$  is called detectable if there exists a bounded linear operator  $F : Y \rightarrow X$  such that  $(A + FC)$  generates an exponentially stable  $C_0$ -semigroup on  $H$ .*

Since  $B$  and  $C$  are bounded, the following theorem follows from [11, Part III, Prop. 2.3, Prop. 3.2, and Cor. 4.2].

THEOREM 5.3. *If  $(A, B)$  is stabilizable, then there exists a minimal solution  $X_{\min} \in \Sigma^+(H)$  of the Riccati equation (5.8). If, in addition,  $(A, C)$  is detectable, then  $A - BB^*X_{\min}$  generates an exponentially stable semigroup and  $X_{\min}$  is the unique solution of the Riccati equation (5.8) in  $\Sigma^+(H)$ .*

In the Newton iteration, we will see that the Newton steps are defined by the solutions of a generalized Lyapunov equation. Therefore, we recall a few facts about Lyapunov equations in Hilbert spaces. The following result is found on pages 19–28 in [11].

THEOREM 5.4. *Let  $S(\cdot)$  denote a strongly continuous semigroup on a Hilbert space  $H$  with an infinitesimal generator  $A$ . Then the following statements are equivalent.*

(i) *The semigroup  $S(\cdot)$  is exponentially stable; i.e., there exist  $\omega > 0$  and  $M \geq 1$  such that*

$$(5.9) \quad \|S(t)x\| \leq Me^{-\omega t}\|x\| \quad \text{for all } x \in H, \quad t \geq 0.$$

(ii) *There exists a positive  $P \in \Sigma^+(H)$  such that*

$$(5.10) \quad (Px, Ay) + (Ax, Py) + (x, y) = 0, \quad x, y \in D(A).$$

When applying Newton's method, we obtain Lyapunov equations that are more general where the identity term  $(x, y)$  is replaced by a more general term  $(x, Qy)$  with possible nonnegativity or positivity properties. However, from the representation formula for solutions of Lyapunov equations, one can establish the following result (see [22, page 252]).

THEOREM 5.5. *Let  $S(\cdot)$  denote a strongly continuous semigroup on a Hilbert space  $H$  with an infinitesimal generator  $A$ . If  $S(\cdot)$  is exponentially stable and  $Q \in \Sigma(H)$ , then there exists a unique solution  $X \in \Sigma(H)$  of*

$$(5.11) \quad (Xx, Ay) + (Ax, Xy) + (x, Qy) = 0, \quad x, y \in D(A).$$

Moreover,  $X$  has the representation

$$(5.12) \quad X = \int_0^\infty S^*(t)QS(t)dt,$$

and if  $Q \in \Sigma^+(H)$ , then  $X \in \Sigma^+(H)$ .

**6. The Kleinman–Newton method in Hilbert space.** In this section we define the Kleinman–Newton algorithm and establish convergence. Throughout this section we make the following assumption.

ASSUMPTION 6.1. *Let  $A$  be the infinitesimal generator of a semigroup on  $H$ ,  $B \in \mathcal{L}(U, H)$ , and  $C \in \mathcal{L}(H, Y)$ . We assume that*

- (i) the system  $(A, B)$  is stabilizable and the pair  $(A, C)$  is detectable, and
- (ii) an operator  $X_0 \in \Sigma^+(H)$  is given such that  $A - BB^*X_0$  generates an exponentially stable semigroup on  $H$ .

We solve the abstract algebraic equation (5.7) by Newton's method. We seek a bounded linear operator  $X \in D(\mathcal{A})$ , which provides a solution to the Riccati equation  $\mathcal{F}(X) = 0$  where the nonlinear mapping  $\mathcal{F} : D(\mathcal{A}) \subset \Sigma(H) \rightarrow \Sigma(H)$  is defined by

$$(6.1) \quad \mathcal{F}(X) = \mathcal{A}(X) - XBB^*X + C^*C.$$

In particular, we look for weak solutions to the equation  $\mathcal{F}(X) = 0$  as defined by (5.8).

Formally applying Newton's method to  $\mathcal{F}(X) = 0$  leads to the iteration

$$(6.2) \quad \begin{aligned} &\mathcal{A}(X_{k+1} - X_k) - X_kBB^*(X_{k+1} - X_k) - (X_{k+1} - X_k)BB^*X_k \\ &+ \mathcal{A}(X_k) - X_kBB^*X_k + C^*C = 0, \end{aligned}$$

or equivalently,

$$(6.3) \quad \mathcal{A}(X_{k+1}) - X_kBB^*X_{k+1} - X_{k+1}BB^*X_k + X_kBB^*X_k + C^*C = 0.$$

After rearranging some terms, it follows that the weak formulation of this scheme has the form

$$(6.4) \quad \begin{aligned} &(X_{k+1}x, (A - BB^*X_k)y) + ((A - BB^*X_k)x, X_{k+1}y) \\ &= -(B^*X_kx, B^*X_ky) - (Cx, Cy), \quad x, y \in D(A). \end{aligned}$$

In what follows we will establish the following convergence theorems for Newton's method. We split up the statements into three parts according to the tools being used in the proof. We follow the structure of the proof for the finite dimensional case given in [40] (see pages 91–94).

**THEOREM 6.2.** *If Assumption 6.1 holds, then*

- (i) the Newton iteration (6.4) is well defined and has a unique solution  $X_k \in \Sigma^+(H)$ ,  $k = 1, 2, \dots$ , and
- (ii) the closed-loop operators  $A - BB^*X_k$ ,  $k = 1, 2, \dots$ , generate exponentially stable semigroups  $S_k(t)$ .

*Proof.* Let us assume, by induction, that  $X_k \in \Sigma^+(H)$  is such that  $A - BB^*X_k$  generates an exponentially stable semigroup  $S_k(t)$ . In particular, there exist  $M_k \geq 1$  and  $\omega_k > 0$  such that

$$(6.5) \quad \|S_k(t)\| \leq M_k e^{-\omega_k t}.$$

Theorem 5.5 applied to equation (6.4) yields the existence of the next iterate,  $X_{k+1} \in \Sigma^+(H)$ . By adding and subtracting terms involving  $X_{k+1}$  and reordering, (6.4) can now be rewritten as

$$(6.6) \quad \begin{aligned} &(X_{k+1}x, (A - BB^*X_{k+1})y) + ((A - BB^*X_{k+1})x, X_{k+1}y) \\ &= -(B^*X_{k+1}x, B^*X_{k+1}y) - (Cx, Cy) \\ &\quad - (B^*(X_{k+1} - X_k)x, B^*(X_{k+1} - X_k)y), \quad x, y \in D(A). \end{aligned}$$

From (6.6) the operator  $X_{k+1}$  can be viewed as a solution  $X_{k+1} \in \Sigma^+(H)$  of a Lyapunov equation of the form (5.11) with the infinitesimal generator  $A - BB^*X_{k+1}$  of a semigroup denoted by  $S_{k+1}(t)$ . Since  $X_{k+1}$  exists, we define

$$V(t, z) := (X_{k+1}S_{k+1}(t)z, S_{k+1}(t)z), \quad z \in D(A - BB^*X_{k+1}).$$

It follows from (6.6) that  $\frac{dV}{dt}(t, z) = -\Phi(t)$ , where

$$\Phi(t) = \|B^*X_{k+1}S_{k+1}(t)z\|^2 + \|CS_{k+1}(t)z\|^2 + \|B^*(X_{k+1} - X_k)S_{k+1}(t)z\|^2.$$

Since  $X_{k+1} \in \Sigma^+(H)$ , an integration of the previous equation yields

$$\int_0^t \Phi(s) ds = V(0, z) - V(t, z) \leq V(0, z)$$

for all  $t > 0$  and  $z \in D(A - BB^*X_{k+1})$ . The domain  $D(A - BB^*X_{k+1})$  is dense in  $H$ , which implies that for all  $z \in H$  there is a constant  $c_z$  such that

$$(6.7) \quad \int_0^\infty \|B^*(X_{k+1} - X_k)S_{k+1}(t)z\|^2 dt \leq \int_0^\infty \Phi(t) dt \leq c_z.$$

Set  $A_k = A - BB^*X_k$  so that  $A_{k+1} = A - BB^*X_{k+1} = A_k + BB^*(X_k - X_{k+1})$ . Observe that  $z_{k+1}(t) = S_{k+1}(t)z$  is the solution of

$$\dot{z}_{k+1}(t) = A_{k+1}z_{k+1}(t) = A_k z_{k+1}(t) + BB^*(X_k - X_{k+1})z_{k+1}(t), \quad z_{k+1}(0) = z,$$

and is given by

$$z_{k+1}(t) = S_k(t)z + \int_0^t S_k(t-s)BB^*(X_k - X_{k+1})z_{k+1}(s)ds.$$

Using the assumption that  $S_k(t)$  is exponentially stable so that (6.5) holds, we have

$$\begin{aligned} \|z_{k+1}(t)\| &\leq \|S_k(t)z\| + \int_0^t \|S_k(t-s)\| \|BB^*(X_k - X_{k+1})z_{k+1}(s)\| ds \\ &\leq M_k e^{-\omega_k t} \|z\| + M_k \|B\| \int_0^t e^{-\omega_k(t-s)} \|B^*(X_k - X_{k+1})\| \|z_{k+1}(s)\| ds. \end{aligned}$$

Gronwall's inequality, together with (6.7), yields

$$\int_0^\infty \|S_{k+1}(t)z\|^2 dt = \int_0^\infty \|z_{k+1}(t)\|^2 dt \leq c_z, \quad z \in D(A - BB^*X_{k+1}).$$

By density this holds for all  $z \in H$ , and by Theorem 5.1.2 in [22] we obtain that  $S_{k+1}$  is an exponentially stable semigroup. This concludes the induction step and the proof.  $\square$

**THEOREM 6.3.** *If Assumption 6.1 holds, then*

- (i) *the sequence  $X_k$  converges in  $\Sigma^+(H)$  and  $\lim_{k \rightarrow \infty} X_k = X_\infty \in \Sigma^+(H)$ . Moreover,  $X_\infty$  is a weak solution to  $\mathcal{F}(X_\infty) = 0$ ;*
- (ii) *the closed-loop operator  $A - BB^*X_\infty$  generates an exponentially stable semigroup  $S_\infty(t)$  satisfying*

$$(6.8) \quad \|S_\infty(t)\| \leq M e^{-\omega t}$$

*for constants  $M \geq 1$  and  $\omega > 0$ ;*

- (iii) *the Newton iterates satisfy  $0 \leq X_\infty \leq \cdots \leq X_{k+1} \leq X_k \leq \cdots \leq X_1$ .*

*Proof.* If we increase the index in (6.4) by one, we obtain

$$\begin{aligned} (X_{k+2}x, (A - BB^*X_{k+1}y)) + ((A - BB^*X_{k+1})x, X_{k+2}y) \\ = -(B^*X_{k+1}x, B^*X_{k+1}y) - (Cx, Cy), \quad x, y \in D(A). \end{aligned}$$

Subtracting this from (6.6) yields

$$(6.9) \quad \begin{aligned} & ((X_{k+1} - X_{k+2})x, (A - BB^*X_{k+1}y)) + ((A - BB^*X_{k+1})x, (X_{k+1} - X_{k+2})y) \\ & = -(B^*(X_{k+1} - X_k)x, (B^*(X_{k+1} - X_k)y)), \quad x, y \in D(A). \end{aligned}$$

We can infer from Theorem 5.5 that  $X_{k+1} - X_{k+2} \geq 0$ ,  $k = 0, 1, 2, \dots$

Therefore  $X_k \in \Sigma^+(H)$  is a sequence of operators which is decreasing and bounded from below by 0. By [39, page 282], there exists an  $X_\infty \in \Sigma^+(H)$  with

$$\lim_{k \rightarrow +\infty} X_k x = X_\infty x \quad \text{for all } x \in H.$$

Passing to the limit in (6.4), we deduce that  $X_\infty$  satisfies the Riccati equation (5.8) in its weak form, and hence  $\mathcal{F}(X_\infty) = 0$ . Theorem 5.3 implies that the solution  $X_\infty$  is unique, and by Assumption 6.1,  $A - BB^*X_\infty$  generates an exponentially stable semigroup.  $\square$

THEOREM 6.4. *If Assumption 6.1 holds, then for all  $k = 0, 1, 2, \dots$ ,*

$$\|X_{k+1} - X_\infty\| \leq c\|X_k - X_\infty\|^2,$$

where

$$c = \int_0^\infty \|S_\infty^*(t)\| \|BB^*\| \|S_\infty(t)\| dt \leq \frac{M^2}{2\omega} \|BB^*\|$$

and the constants  $M \geq 1$  and  $\omega > 0$  are given by (6.8).

*Proof.* By (6.3), the limit operator  $X_\infty$  satisfies the equation

$$(6.10) \quad \mathcal{A}(X_\infty) - X_\infty BB^*X_\infty + C^*C = 0.$$

To shorten notation in the rest of the proof, all the equations are to be understood in the weak sense. Equation (6.10) can be rewritten as

$$(6.11) \quad \begin{aligned} & (A - BB^*X_{k+1})^*X_\infty + X_\infty(A - BB^*X_{k+1}) \\ & = -X_\infty BB^*X_\infty - C^*C + (X_\infty - X_{k+1})BB^*X_\infty + X_\infty BB^*(X_\infty - X_{k+1}). \end{aligned}$$

If we subtract (6.6) from (6.11), we obtain

$$(6.12) \quad \begin{aligned} & (A - BB^*X_{k+1})^*(X_\infty - X_{k+1}) + (X_\infty - X_{k+1})(A - BB^*X_{k+1}) \\ & = X_\infty BB^*X_\infty - X_{k+1} BB^*X_\infty - X_\infty BB^*X_{k+1} + X_{k+1} BB^*X_{k+1} \\ & \quad + (X_{k+1} - X_k)BB^*(X_{k+1} - X_k) \\ & = (X_\infty - X_{k+1})BB^*(X_\infty - X_{k+1}) + (X_{k+1} - X_k)BB^*(X_{k+1} - X_k). \end{aligned}$$

This implies that

$$(6.13) \quad \begin{aligned} & (A - BB^*X_\infty)^*(X_\infty - X_{k+1}) + (X_\infty - X_{k+1})(A - BB^*X_\infty) \\ & = -(X_\infty - X_{k+1})BB^*(X_\infty - X_{k+1}) + (X_{k+1} - X_k)BB^*(X_{k+1} - X_k). \end{aligned}$$

Note that  $\Delta = X_\infty - X_{k+1}$  is the solution to the Lyapunov equation

$$(A_\infty)^* \Delta + \Delta(A_\infty) = -\hat{Q},$$

where

$$\hat{Q} = (X_\infty - X_{k+1})BB^*(X_\infty - X_{k+1}) - (X_{k+1} - X_k)BB^*(X_{k+1} - X_k)$$

and  $A_\infty = A - BB^*X_\infty$  generates the exponentially stable semigroup  $S_\infty(t)$ . It follows from (5.12) in Theorem 5.5 above that the representation formula in Corollary 4.2 of [28] can be used to derive the following:

(6.14)

$$\begin{aligned} 0 \leq X_{k+1} - X_\infty &= \int_0^\infty S_\infty^*(t) \{ -(X_\infty - X_{k+1})BB^*(X_\infty - X_{k+1}) \\ &\quad + (X_{k+1} - X_k)BB^*(X_{k+1} - X_k) \} S_\infty(t) dt \\ &\leq \int_0^\infty S_\infty^*(t) ((X_{k+1} - X_k)BB^*(X_{k+1} - X_k)) S_\infty(t) dt. \end{aligned}$$

Taking norms and using the fact that, for self-adjoint operators,  $\|S\| = \sup_{\|x\| \leq 1} (x, Sx)$ , we obtain

(6.15)

$$\|X_{k+1} - X_\infty\| \leq \|X_{k+1} - X_k\|^2 \int_0^\infty \|S_\infty^*(t)\| \|BB^*\| \|S_\infty(t)\| dt = c \|X_{k+1} - X_k\|^2,$$

where

(6.16)

$$c = \int_0^\infty \|S_\infty^*(t)\| \|BB^*\| \|S_\infty(t)\| dt = \|BB^*\| \int_0^\infty \|S_\infty^*(t)\| \|S_\infty(t)\| dt \leq \frac{M^2}{2\omega} \|BB^*\|$$

follows from (6.8). Since all the operators are self-adjoint, we have

$$0 \leq X_k - X_{k+1} \leq X_k - X_\infty \quad \Rightarrow \quad \|X_k - X_{k+1}\| \leq \|X_k - X_\infty\|,$$

which implies the quadratic rate of convergence

$$\|X_{k+1} - X_\infty\| \leq c \|X_k - X_\infty\|^2. \quad \square$$

**7. Approximation and mesh independence results.** In this section we focus on the problem of developing numerical schemes that yield convergent and mesh-independent approximations of the infinite dimensional Riccati equation

$$\mathcal{F}(X) = \mathcal{A}(X) - XBB^*X + C^*C = 0,$$

where  $BB^*, CC^* \in \mathcal{L}(H)$  and  $\mathcal{A} : D(\mathcal{A}) \subset \Sigma(H) \rightarrow \Sigma(H)$  is defined as in section 5 by  $\mathcal{A}(\Pi)x = A^*\Pi x + \Pi Ax$ . Although it is possible to work in a very abstract setting, we use the approximation setup found in Ito's paper [33]. The resulting framework is general enough to handle a large class of problems and helps us keep the technical discussion to a minimum.

We consider a sequence of approximating problems defined by  $(H^N, A^N, B^N, C^N)$ , where  $H^N \subset H$  is a sequence of finite dimensional subspaces of  $H$ , and  $A^N \in$

$\mathcal{L}(H^N, H^N)$ ,  $B^N \in \mathcal{L}(U, H^N)$ , and  $C^N \in \mathcal{L}(H^N, Y)$  are bounded linear operators. Let  $P^N : H \rightarrow H^N$  denote the orthogonal projection of  $H$  onto  $H^N$  satisfying  $\|P^N\| \leq 1$ , and as  $N \rightarrow \infty$  we have  $\|P^N x - x\| \rightarrow 0$  for all  $x \in H$ . Note that if  $T^N$  is any bounded linear operator on  $H^N$ , i.e.,  $T^N \in \mathcal{L}(H^N, H^N)$ , then  $T^N P^N$  belongs to  $\mathcal{L}(H, H^N)$  and the operator norms satisfy

$$\|T^N\|_{\mathcal{L}(H^N, H^N)} = \|T^N P^N\|_{\mathcal{L}(H, H^N)}.$$

Therefore, we can use the notation  $\|T^N\| = \|T^N P^N\|$  without referring to the specific spaces.

Define the finite dimensional approximations for  $\mathcal{A}$ ,

$$\mathcal{A}^N : \Sigma(H^N) \rightarrow \Sigma(H^N), \quad N = 1, 2, \dots,$$

by

$$\mathcal{A}^N(\Pi^N) = [A^N]^* \Pi^N + \Pi^N A^N.$$

The resulting approximating Riccati equation becomes

$$(7.1) \quad \mathcal{F}^N(X^N) = \mathcal{A}^N(X^N) - X^N B^N (B^N)^* X^N + (C^N)^* C^N = 0.$$

In this section we distinguish between two types of sequences. Let  $X_k \in \mathcal{L}(H)$  denote the iterates of the Newton method for the infinite dimensional Riccati equation  $\mathcal{F}(X) = 0$ . Likewise,  $X_k^N \in \mathcal{L}(H^N)$  denotes the iterates of the Newton method for the discretized Riccati equation  $\mathcal{F}^N(X^N) = 0$ .

We first review the conditions on the approximating scheme  $(H^N, A^N, B^N, C^N)$  which are sufficient to guarantee that the approximating Riccati equation (7.1) admits a unique nonnegative solution  $X_\infty^N$ , and  $X_\infty^N P^N$  converges to the unique nonnegative solution  $X_\infty$  of the operator Riccati equation (6.1). These results can be found in Ito's paper [33]. We then focus on the issue of mesh independence for the Kleinman-Newton algorithm and present convergence rates.

**7.1. Convergence of approximating Riccati operators.** In order to discuss convergence of the finite dimensional approximating Riccati operators, we need to assume that the numerical scheme preserves the basic stabilizability and detectability conditions needed to guarantee that the LQR problem is well-posed. It is important to note that even standard numerical schemes may not preserve these important control system properties. However, for the delay systems considered below it is known that all of the schemes discussed in Kappel's survey [35] satisfy these conditions (see [15], [18], [19], and [28]). Moreover, as we see below, although these conditions are sufficient for the finite dimensional Newton iterates to converge, they do not guarantee that the limit of the Newton iterates  $X_\infty^N$  converges to  $X_\infty$  as  $N \rightarrow +\infty$ . We shall need additional properties on the approximating sequence  $(H^N, A^N, B^N, C^N)$ . We break these assumptions into three distinct hypotheses concerning the convergence of the operators, convergence of the adjoint operators, and preservation of uniform stabilizability/detectability under the approximation.

**ASSUMPTION 7.1 (convergence).** Assume that there is an  $N_s$  such that for all  $N > N_s$  the following conditions hold:

(C-i) For each  $x \in H$ ,  $S^N(t)P^N x \rightarrow S(t)x$  and the convergence is uniform in  $t$  on bounded subintervals of  $[0, +\infty)$ .

(C-ii) For each  $u \in U$ ,  $B^N u \rightarrow Bu$ , and for each  $x \in H$ ,  $C^N P^N x \rightarrow Cx$ .

ASSUMPTION 7.2 (dual convergence). Assume that there is an  $N_s$  such that for all  $N > N_s$  the following conditions hold:

(C\*-i) For each  $x \in H$ ,  $[S^N(t)]^* P^N x \longrightarrow S^*(t)x$  and the convergence is uniform in  $t$  on bounded subintervals of  $[0, +\infty)$ .

(C\*-ii) For each  $x \in H$ ,  $[B^N]^* P^N x \longrightarrow B^* x$ , and for each  $y \in Y$ ,  $[C^N]^* y \longrightarrow C^* y$ .

ASSUMPTION 7.3 (uniformly stabilizable and detectable). Assume that there is an  $N_s$  such that for all  $N > N_s$  the following conditions hold:

(US) The family of pairs  $(A^N, B^N)$  is uniformly stabilizable; i.e., there exist a sequence of operators  $K^N \in \mathcal{L}(H^N, U)$  and positive constants  $M_1 \geq 1$ ,  $\omega_1 > 0$  such that  $\sup \|K^N\| < +\infty$  and the semigroups  $e^{(A^N - B^N K^N)t}$  generated by closed-loop operators  $A^N - B^N K^N$  satisfy

$$\|e^{(A^N - B^N K^N)t} P^N\| \leq M_1 e^{-\omega_1 t}, \quad t \geq 0.$$

(UD) The family of pairs  $(A^N, C^N)$  is uniformly detectable; i.e., there exist a sequence of operators  $G^N \in \mathcal{L}(Y, H^N)$  and positive constants  $M_2 \geq 1$ ,  $\omega_2 > 0$  such that  $\sup \|G^N\| < +\infty$  and the semigroups  $e^{(A^N - G^N C^N)t}$  generated by closed-loop operators  $A^N - G^N C^N$  satisfy

$$\|e^{(A^N - G^N C^N)t} P^N\| \leq M_2 e^{-\omega_2 t}, \quad t \geq 0.$$

The following results may be found in [33, Theorems 2.1 and 2.2].

THEOREM 7.4. If Assumptions 7.1, 7.2, and 7.3 hold, then for all  $N > N_s$  the Riccati equation (7.1) admits a unique nonnegative solution  $X_\infty^N$ ,  $\sup \|X_\infty^N\| < +\infty$ , and there exist positive constants  $M_3 \geq 1$ ,  $\omega_3 > 0$  (independent of  $N$ ) such that the closed-loop semigroups  $S_\infty^N(t)$  generated by operators  $(A^N - B^N [B^N]^* X_\infty^N)$  satisfy

$$(7.2) \quad \|S_\infty^N(t)\| = \|S_\infty^N(t) P^N\| \leq M_3 e^{-\omega_3 t}, \quad t \geq 0.$$

THEOREM 7.5. If  $(A, B)$  is stabilizable,  $(A, C)$  is detectable, and Assumptions 7.1, 7.2, and 7.3 hold, then the unique nonnegative solutions  $X_\infty^N$  to the Riccati equation (7.1) converge strongly to  $X_\infty$ . Moreover, the closed-loop semigroups  $S_\infty^N(t)$  converge strongly to the closed-loop semigroup  $S_\infty(t)$  and

$$(7.3) \quad \|S_\infty(t)\| \leq M_3 e^{-\omega_3 t}, \quad t \geq 0.$$

Note that the previous results yield only strong convergence of the solutions of the Riccati equations. However, if  $B$  is bounded with finite dimensional range, then strong convergence of  $X_\infty^N$  to  $X_\infty$  implies norm convergence of the feedback gain operators. In particular, if  $\text{rank}(B) < +\infty$ , then

$$(7.4) \quad \lim_{N \rightarrow +\infty} \|K^N - K\| = 0,$$

where  $K^N = [B^N]^* X_\infty^N$  (see Theorems 6.2 and 6.8 in [28]). Moreover, under certain compactness assumptions on  $B$  and  $C$ , one can establish norm convergence of the Riccati operators. There are a number of results along this line (see [24], [28], [33], [34], [43], and [44]) and some make use of the smoothing property of the semigroup (e.g., analytic, differentiable). The following theorem is not the most general result, but it is directly applicable to delay and parabolic PDE control systems. Also, this result can be used to obtain rates of convergence for the approximating Riccati operators. The proof follows directly from Ito's paper [33, pp. 158–160].



**THEOREM 7.6.** *Suppose Assumptions 7.1, 7.2, and 7.3 hold,  $B$  and  $C$  are compact, and  $X_\infty x \in D(A^*)$  for all  $x \in H$ . If  $B^N = P^N B$  and  $C^N = C P^N$ , then  $(A, B)$  is stabilizable,  $(A, C)$  is detectable, and there exists a constant  $\hat{\beta} > 0$  such that*

$$(7.5) \quad \|X_\infty^N - P^N X_\infty P^N\| \leq \Delta^N,$$

where  $\Delta^N$  is given by

$$(7.6) \quad \Delta^N \triangleq \hat{\beta} \{ \|(A^* - [A^N]^* P^N) X_\infty\| + \|B\| \|(B^* - [B^N]^* P^N) X_\infty\| \}.$$

If, in addition,  $\lim_{N \rightarrow +\infty} \|(A^* - [A^N]^* P^N) X_\infty\| = 0$ , then

$$\lim_{N \rightarrow +\infty} \|X_\infty^N - P^N X_\infty P^N\| = 0.$$

*Remark.* The assumptions of convergence, dual convergence, and uniform preservation of stability and detectability in Ito's result are sufficient, but it is not yet clear if they are necessary for operator norm convergence (see [13], [16], and [26]). However, most approximations that yield operator norm convergence satisfy these or even stronger assumptions (see [14]). We note that, especially for nonnormal problems, it may not be easy to check these conditions for a specific approximation scheme. For example, it is not known which numerical algorithms used in computational fluid dynamics are dual convergent when applied to nonnormal control systems typical in this area (see [15] and [20]).

At first glance the compactness assumptions on the  $B$  and  $C$  operators seem rather strong. This assumption certainly excludes some idealized boundary control problems. On the other hand, if one includes actuator or sensor dynamics at the boundary (a reasonable assumption in many boundary control problems), then the resulting  $B$  and  $C$  operators are often compact in practical problems. Moreover, the simple example below provides some insight into the importance of this assumption and perhaps a way out of this technical difficulty.

*Example 2.* Consider Example 1 above and let  $H^N = \mathbb{R} \times \mathbb{R}^N$ , and define  $P^N : H \rightarrow H^N$  to be the natural projection onto  $H^N$ . If  $A^N = P^N A$ ,  $B^N = P^N B$ , and  $C^N = C P^N$ , then all the conditions in the previous theorem are satisfied except that  $C$  is not compact. The solution to the finite dimensional Riccati equation

$$\mathcal{F}^N(X) \triangleq [A^N]^* X^N + X^N [A^N] - X^N [B^N] [B^N]^* X^N + Q^N = 0$$

is  $X^N = \begin{bmatrix} 1 & 0 \\ 0 & I^N \end{bmatrix}$ , where  $I^N$  is the identity on  $\mathbb{R}^N$ . Clearly,  $X^N$  does not converge to  $X$  in the uniform operator norm since  $I$  is not compact. It is interesting to note that the feedback gain operators

$$(7.7) \quad K^N = [B^N]^* X^N = \begin{bmatrix} 1 & 0 \end{bmatrix} = K$$

converge uniformly. This situation occurs in many problems and can be exploited to address mesh independence issues. We shall discuss this issue in a future paper. We turn now to the issue of mesh independence.

**7.2. Mesh independence of the Kleinman-Newton algorithm.** We turn now to the application of the Kleinman-Newton algorithm to the finite dimensional Riccati equation

$$(7.8) \quad \mathcal{F}^N(X^N) = \mathcal{A}^N(X^N) - X^N B^N (B^N)^* X^N + (C^N)^* C^N = 0.$$

We solve the abstract algebraic equation (7.8) by Newton's method. In particular, we seek a bounded linear operator  $X^N \in D(\mathcal{A}^N)$ , which provides a solution to the Riccati equation  $\mathcal{F}^N(X^N) = 0$ , where the nonlinear mapping  $\mathcal{F}^N : D(\mathcal{A}^N) \subset \Sigma(H^N) \rightarrow \Sigma(H^N)$  is as defined by (7.1) above. Just as for the infinite dimensional case, applying Newton's method to  $\mathcal{F}^N(X^N) = 0$  leads to the iteration

$$(7.9) \quad \begin{aligned} \mathcal{A}^N(X_{k+1}^N - X_k^N) - X_k^N B^N [B^N]^* (X_{k+1}^N - X_k^N) - (X_{k+1}^N - X_k^N) B^N [B^N]^* X_k^N \\ + \mathcal{A}^N(X_k^N) - X_k^N B^N [B^N]^* X_k^N + [C^N]^* C^N = 0, \end{aligned}$$

or equivalently,

$$(7.10) \quad \mathcal{A}^N(X_{k+1}^N) - X_k^N B^N [B^N]^* X_{k+1}^N - X_{k+1}^N B^N [B^N]^* X_k^N + X_k^N B^N [B^N]^* X_k^N + [C^N]^* C^N = 0.$$

We assume that the approximation scheme preserves the basic Assumption 6.1 for all  $N$  sufficiently large. This ensures that the Newton iterations in the finite dimensional spaces converge monotonically as in Theorem 6.3. In particular, we shall use the following hypothesis.

**ASSUMPTION 7.7.** *Let  $S^N(t)$  be the semigroup generated by  $A^N$  on  $H^N$ ,  $B^N \in \mathcal{L}(U, H^N)$ , and  $C^N \in \mathcal{L}(H^N, Y)$ . Assume that there is an  $N_s$  such that for all  $N > N_s$  the following conditions hold:*

- (N-i) *The system  $(A^N, B^N)$  is stabilizable and the pair  $(A^N, C^N)$  is detectable.*
- (N-ii) *An operator  $X_0^N \in \Sigma^+(H^N)$  is given such that  $A^N - B^N(B^N)^* X_0^N$  generates an exponentially stable semigroup on  $H^N$ .*

The following results are the finite dimensional versions of Theorems 6.3 and 6.4 above. The proofs are almost identical.

**THEOREM 7.8.** *If Assumption 7.7 holds, then for all  $N > N_s$ ,*

- (i) *the sequence  $X_k^N$  converges in  $\Sigma^+(H^N)$  and  $\lim_{k \rightarrow +\infty} X_k^N = X_\infty^N \in \Sigma^+(H^N)$ . Moreover,  $X_\infty^N$  is a solution to  $\mathcal{F}^N(X_\infty^N) = 0$ ;*
- (ii) *the closed-loop operator  $A^N - B^N[B^N]^* X_\infty^N$  generates an exponentially stable semigroup  $S_\infty^N(t)$  satisfying*

$$(7.11) \quad \|S_\infty^N(t)\| \leq M_N e^{-\omega_N t}$$

*for constants  $M_N \geq 1$  and  $\omega_N > 0$ ;*

- (iii) *the Newton iterates satisfy  $0 \leq X_\infty^N \leq \dots \leq X_{k+1}^N \leq X_k^N \leq \dots \leq X_1^N$ .*

**THEOREM 7.9.** *If Assumption 7.7 holds, then for all  $N > N_s$  and  $k = 0, 1, 2, \dots$ ,*

$$\|X_{k+1}^N - X_\infty^N\| \leq c^N \|X_k^N - X_\infty^N\|^2,$$

where

$$(7.12) \quad \begin{aligned} c^N &= \int_0^\infty \| [S_\infty^N(t)]^* \| \| B^N [B^N]^* \| \| S_\infty^N(t) \| \, dt \\ &= \| B^N [B^N]^* \| \int_0^\infty \| [S_\infty^N(t)]^* \| \| S_\infty^N(t) \| \, dt \leq (M_N^2 / 2\omega_N) \| B^N [B^N]^* \| \end{aligned}$$

and the constants  $M_N \geq 1$  and  $\omega_N > 0$  are as given by (7.11) above.

The constant  $c^N$  in Theorem 7.9 is not necessarily the minimal value  $\hat{c}^N$ . However, it is possible that there exists an  $\alpha$ , independent of  $N$ , such that the finite dimensional iterates  $X_k^N$  satisfy

$$(7.13) \quad \|X_{k+1}^N - X_\infty^N\| \leq \alpha \|X_k^N - X_\infty^N\|^2.$$

Clearly, the bound

$$c^N \leq (M_N^2/2\omega_N)\|B^N[B^N]^*\|$$

is not tight. As we shall see below, there are convergent approximation schemes satisfying Assumption 7.1 and a fixed  $\alpha$  satisfying (7.13) with

$$c^N \leq \alpha < \lim_{N \rightarrow +\infty} (M_N^2/2\omega_N)\|B^N[B^N]^*\| = +\infty.$$

Therefore, the rate of convergence dictated by the constant  $\alpha$  in (7.13) provides some level of mesh independence for the finite dimensional problems. However, it is important to note that the previous estimates do not imply that the approximating Riccati operators  $X_\infty^N$  converge in norm to  $X_\infty$ . Hence, even the existence of a constant  $\alpha$  for which (7.13) holds does not provide true mesh independence.

Although the previous theorem is well known, the explicit value of the constant  $c^N$  in (7.12) provides some insight into those approximation properties that might be important in establishing an MIP. There are several factors that influence this constant, but clearly the choice of the approximation scheme  $(H^N, A^N, B^N, C^N)$  plays a fundamental role in determining  $c^N$  and its value as the mesh is refined. We shall illustrate this dependency with the numerical examples below. However, using Ito's theorem, Theorem 7.4 above, we have the following mesh independence result.

**THEOREM 7.10.** *If Assumptions 7.1, 7.2, and 7.3 hold and  $B$  is compact, then there exist  $\alpha^N$  and  $\alpha$  such that*

$$(7.14) \quad \|X_{k+1}^N - X_\infty^N\| \leq \alpha^N \|X_k^N - X_\infty^N\|^2,$$

$$(7.15) \quad \|X_{k+1} - X_\infty\| \leq \alpha \|X_k - X_\infty\|^2,$$

and  $\alpha^N = \alpha + \delta(N)$ , where  $\delta(N) \rightarrow 0$  as  $N \rightarrow +\infty$ .

*Proof.* First note that Assumption 7.3 implies that Assumption 7.7 holds so that the Kleinman-Newton iterates  $X_k^N$  exist and Theorem 7.9 is valid. From Assumption 7.1 we have convergence  $B^N u \rightarrow Bu$  for  $u \in U$ , and Assumption 7.2 yields the dual convergence  $[B^N]^* P^N x \rightarrow B^* x$  for  $x \in H$ . Since  $B$  is compact, it follows from Theorem 3.2 in [21] that  $\|B^N - B\| \rightarrow 0$ , and  $\|[B^N]^* P^N - B^*\| \rightarrow 0$  so that  $\|B^N[B^N]^* P^N - BB^*\| \rightarrow 0$ . Let

$$(7.16) \quad \beta(N) = \|B^N[B^N]^* P^N\| - \|BB^*\|.$$

Theorem 7.4 yields the existence of positive constants  $M_3 \geq 1$ ,  $\omega_3 > 0$  (independent of  $N$ ) such that the closed-loop semigroups  $S_\infty^N(t)$  generated by the operators  $(A^N - B^N[B^N]^* X_\infty^N)$  satisfy

$$\|S_\infty^N(t)^*\| = \|S_\infty^N(t)\| \leq M_3 e^{-\omega_3 t}, \quad t \geq 0,$$

and

$$\|S_\infty^*(t)\| = \|S_\infty(t)\| \leq M_3 e^{-\omega_3 t}, \quad t \geq 0.$$

Hence, the estimate for  $c^N$  in (7.12) is bounded by

$$c^N = \int_0^\infty \|S_\infty^N(t)^*\| \|B^N[B^N]^*\| \|S_\infty^N(t)\| dt \leq \frac{M_3^2}{2\omega_3} \|B^N[B^N]^* P^N\|.$$

Let

$$\alpha^N = \frac{M_3^2}{2\omega_3} \|B^N [B^N]^* P^N\| \geq c^N \text{ and } \alpha = \frac{M_3^2}{2\omega_3} \|BB^*\| \geq c,$$

where  $c$  is given by (6.16). It follows that  $\alpha^N$  and  $\alpha$  satisfy (7.14) and (7.15), respectively. Moreover,

$$\alpha^N = \frac{M_3^2}{2\omega_3} \|B^N [B^N]^* P^N\| = \frac{M_3^2}{2\omega_3} \{\|BB^*\| + \beta(N)\} = \alpha + \frac{M_3^2}{2\omega_3} \beta(N) = \alpha + \delta(N),$$

where

$$\delta(N) = \frac{M_3^2}{2\omega_3} \beta(N) \longrightarrow 0 \text{ as } N \longrightarrow +\infty,$$

and this completes the proof.  $\square$

All that one can imply from Theorem 7.10 is that the finite dimensional iterates  $X_k^N$  satisfy

$$\|X_{k+1}^N - X_\infty^N\| \leq (\alpha + \delta(N)) \|X_k^N - X_\infty^N\|^2.$$

If, in addition, one has norm convergence  $\lim_{N \rightarrow +\infty} \|P^N X_\infty P^N - X_\infty^N\| = 0$ , then the inequality

$$\begin{aligned} \|X_{k+1}^N - P^N X_\infty P^N\| &= \|X_{k+1}^N - X_\infty^N + X_\infty^N - P^N X_\infty P^N\| \\ &\leq \|X_{k+1}^N - X_\infty^N\| + \|X_\infty^N - P^N X_\infty P^N\| \\ &\leq (\alpha + \delta(N)) \|X_k^N - X_\infty^N\|^2 + \|X_\infty^N - P^N X_\infty P^N\| \end{aligned}$$

provides a useful overall convergence rate of

$$(7.17) \quad \|X_{k+1}^N - P^N X_\infty P^N\| \leq (\alpha + \delta(N)) \|X_k^N - X_\infty^N\|^2 + \|X_\infty^N - P^N X_\infty P^N\|$$

in terms of the Newton iterates and the finite dimensional approximations. Applying Theorems 7.5 and 7.6 above yields the following mesh independence result.

**THEOREM 7.11.** *Suppose Assumptions 7.1, 7.2, and 7.3 hold,  $B$  and  $C$  are compact, and  $X_\infty x \in D(A^*)$  for all  $x \in H$ . If  $B^N = P^N B$  and  $C^N = C P^N$ , then there exist  $\delta(N) \longrightarrow 0$  as  $N \longrightarrow +\infty$  and  $\hat{\beta}$  such that*

$$(7.18) \quad \|X_{k+1}^N - P^N X_\infty P^N\| \leq (\alpha + \delta(N)) \|X_k^N - X_\infty^N\|^2 + \Delta^N,$$

where  $\alpha$  is as given by (7.15) in Theorem 7.10 and  $\Delta^N = \hat{\beta} \{ \|(A^* - [A^N]^* P^N) X_\infty\| + \|B\| \|(B^* - [B^N]^* P^N) X_\infty\| \}$ . If, in addition, for some  $p > 0$  we have

$$\Delta^N = O(1/N^p),$$

then the MIP holds with a rate of  $O(1/N^p)$ .

Observe that the rate determined by

$$\Delta^N = \hat{\beta} \{ \|(A^* - [A^N]^* P^N) X_\infty\| + \|B\| \|(B^* - [B^N]^* P^N) X_\infty\| \}$$

depends on the order of the approximating scheme  $(H^N, A^N, B^N, C^N)$ . In particular, the rate of convergence for the adjoint approximations  $[A^N]^*$  plays a key role. For

the delay systems discussed below, it follows that  $B^* = [B^N]^*$  for all  $N > 1$  so that the rate is essentially determined by how well one can approximate  $A^*$ , i.e., if one can obtain an estimate of the form

$$\|(A^* - [A^N]^* P^N)X_\infty\| = O(1/N^p).$$

We shall apply this estimate to the delay systems considered in the next section. In general, the convergence results depend on the regularity of the semigroups and the type of approximations. Obtaining these rates depends on each individual problem. Ito considered both delay systems and parabolic PDE control systems. He established convergence rates for the standard finite element scheme applied to the parabolic PDE problem. He also gave convergence rates for the two schemes we will discuss below involving delay systems (see [33] and [34]). In both cases he made heavy use of the regularity of the semigroups generated by  $A$ .

**8. Control of delay systems.** In this section we consider the LQR problem for delay differential equations. In particular, the system is defined by

$$(8.1) \quad \dot{x}(t) = A_0 x(t) + A_1 x(t-r) + B_0 u(t), \quad t > 0,$$

with initial data

$$(8.2) \quad x(0) = \eta, \quad x(s) = \varphi(s), \quad -r < s < 0,$$

where  $\eta \in \mathbb{R}^n$  and  $\varphi(\cdot) \in L^2(-r, 0; \mathbb{R}^n)$ . Here,  $A_0$  and  $A_1$  are  $n \times n$  constant real matrices and  $B_0$  is an  $n \times m$  matrix.

Let  $C_0 = [C_0]^T \geq 0$  be a symmetric real-valued matrix and define the cost function

$$(8.3) \quad J(u) = \int_0^{+\infty} \{(C_0 x(s))^T C_0 x(s) + \|u(s)\|^2\} ds.$$

The corresponding LQR problem is to minimize the quadratic cost (8.3) over all controls  $u \in L^2(0, +\infty; \mathbb{R}^m)$ .

In order to present this problem in an infinite dimensional setting we use the Hilbert space  $H = \mathbb{R}^n \times L^2(-r, 0; \mathbb{R}^n)$ . Define the operator  $A$  with domain

$$D(A) = \{(\eta, \phi(\cdot)) \in H : \phi(\cdot) \in H^1(-r, 0; \mathbb{R}^n) \quad \text{and} \quad \phi(0) = \eta\}$$

by

$$A \begin{bmatrix} \eta \\ \phi(\cdot) \end{bmatrix} = \begin{bmatrix} A_0 \eta + A_1 \varphi(-r) \\ \phi'(\cdot) \end{bmatrix}.$$

Also, let  $B : \mathbb{R}^m \rightarrow H$  be defined by

$$Bu = \begin{bmatrix} B_0 u \\ 0 \end{bmatrix},$$

and let  $C : H \rightarrow \mathbb{R}^m$  be given by

$$Cz = C \begin{bmatrix} \eta \\ \phi(\cdot) \end{bmatrix} = C_0 \eta.$$

It is well known (see [3]) that  $A$  generates a  $C_0$ -semigroup  $S(t) : H \rightarrow H$ ,  $t \geq 0$ , such that

$$(8.4) \quad S(t)(\eta, \phi(\cdot)) = (x(t), x_t(\cdot)) \in H$$

for all  $(\eta, \phi(\cdot)) \in H$ , where  $x(t)$  is the solution to (8.1)–(8.2) and  $x_t(s) = x(t+s)$  for all  $-r < s < 0$ . Moreover, the delay system is equivalent to the infinite dimensional system in  $H$  defined by

$$(8.5) \quad \dot{z}(t) = Az(t) + Bu(t), \quad t > 0,$$

with initial data

$$(8.6) \quad z(0) = \begin{bmatrix} \eta \\ \phi(\cdot) \end{bmatrix},$$

and the LQR cost function has the form

$$J(u) = \int_0^{+\infty} \{\|Cz(s)\|^2 + \|u(s)\|^2\} ds.$$

Finally, the Hilbert adjoint  $A^*$  is defined on the domain

$$D(A^*) = \{(\xi, \psi(\cdot)) \in H : \psi(\cdot) \in H^1(-r, 0; \mathbb{R}^n), \quad \psi(-r) = A_1^T \xi\}$$

by

$$A^* \begin{bmatrix} \xi \\ \psi(\cdot) \end{bmatrix} = \begin{bmatrix} A_0^T \xi + \psi(0) \\ -\psi'(\cdot) \end{bmatrix}.$$

Observe that the linear operator  $A$  is not normal; this can cause problems when approximating the LQR control problem (see [8], [13], [16], [15], and [35]).

**8.1. Approximations of the delay system.** We consider two different numerical schemes for approximating the LQR control problem for the delay system. Since the  $B$  and  $C$  operators act only on the finite dimensional part of the state, the main issue is how to approximate  $A$ . We focus on a finite volume method known as the “AVE” scheme in [4] and a conforming finite element scheme first described by Banks and Kappel in [5], and hence we do not give the details here. The key difference between these two schemes is how they approximate the initial condition  $\phi$  in (8.1). The “AVE” scheme uses an averaging technique and characteristic functions, while the Banks–Kappel scheme uses a continuous finite element technique. Although both schemes are convergent, only the “AVE” scheme is dual convergent, and hence produces convergent approximations of the operator

$$\mathcal{A}(\Pi) = A^* \Pi + A \Pi, \quad \Pi \in D(\mathcal{A}).$$

The papers by Rosen (see [43], [44], and [45]) provide considerable insight in this problem.

**The AVE/finite volume scheme.** For each  $N > 1$ , create a partition on  $[-r, 0]$  by defining  $\tau_j^N = -jr/N$ , where  $j = 0, \dots, N$ . On  $[-r, 0]$ , define  $\chi_j^N(\cdot)$  to be the characteristic function on  $[\tau_j^N, \tau_{j-1}^N)$  for  $j = 2, \dots, N$ , and define  $\chi_1^N(\cdot)$  to be the

characteristic function on  $[\tau_1^N, \tau_0^N]$ . Define the finite dimensional subspace  $H_{AVE}^N$  of  $H$  by

$$(8.7) \quad H_{AVE}^N \equiv \left\{ (\eta, \phi^N(\cdot)) \in H : \phi^N(s) = \sum_{j=1}^N v_j^N \chi_j^N(s), v_j^N \in \mathbb{R}^n \right\}.$$

The projection  $P^N$  of  $H$  into  $H_{AVE}^N$  is defined by

$$P^N(\eta, \phi(\cdot)) = \left( \phi_0^N, \sum_{j=1}^N \phi_j^N \chi_j^N(\cdot) \right),$$

where

$$\phi_0^N \equiv \eta \quad \text{and for } j = 1, \dots, N, \quad \phi_j^N \equiv \frac{N}{r} \int_{\tau_j^N}^{\tau_{j-1}^N} \phi(s) ds.$$

To approximate the operator  $A$ , we first define  $L^N : H_{AVE}^N \rightarrow \mathbb{R}^n$  and  $D^N : H_{AVE}^N \rightarrow L^2(-r, 0; \mathbb{R}^n)$  by

$$L^N \left( \eta, \sum_{j=1}^N v_j^N \chi_j^N(\cdot) \right) = A_0 \eta + A_1 v_N^N$$

and

$$D^N \left( \eta, \sum_{j=1}^N v_j^N \chi_j^N(\cdot) \right) = \frac{N}{r} \sum_{j=1}^N \{v_{j-1}^N - v_j^N\} \chi_j^N(\cdot),$$

respectively, where  $v_0^N = \eta$ . The AVE approximation  $A_{AVE}^N : H_{AVE}^N \rightarrow H_{AVE}^N \subseteq H$  is given by

$$(8.8) \quad A_{AVE}^N(\eta, \psi) \equiv (L^N(\eta, \psi), D^N(\eta, \psi)).$$

In order to complete the approximation scheme, we define

$$(8.9) \quad B_{AVE}^N = P^N B \quad \text{and} \quad C_{AVE}^N = C P^N,$$

and this yields the AVE approximation scheme  $(H_{AVE}^N, A_{AVE}^N, B_{AVE}^N, C_{AVE}^N)$ .

Observe that

$$B_{AVE}^N u = P^N \begin{bmatrix} B_0 u \\ 0 \end{bmatrix} = \begin{bmatrix} B_0 u \\ 0 \end{bmatrix} = Bu$$

and

$$C_{AVE}^N \begin{bmatrix} \eta \\ \phi(\cdot) \end{bmatrix} = C P^N \begin{bmatrix} \eta \\ \phi(\cdot) \end{bmatrix} = C \begin{bmatrix} \eta \\ \phi^N(\cdot) \end{bmatrix} = C_0 \eta = C \begin{bmatrix} \eta \\ \phi(\cdot) \end{bmatrix}.$$

Hence, the operators  $B$  and  $C$  are compact and satisfy the conditions in Theorem 7.6. Norm convergence of the input and output operators is trivial for this approximation.

Since  $\|B^N[B^N]^*P^N\| = \|BB^*\|$  for all  $N \geq 1$ , it follows that  $\beta(N)$  defined in (7.16) satisfies

$$\beta(N) = \|B^N[B^N]^*P^N\| - \|BB^*\| = 0,$$

and hence  $\delta(N) = \frac{M_3^2}{2\omega_3}\beta(N) = 0$  for all  $N \geq 1$ . Moreover, the AVE scheme satisfies all the assumptions in Theorem 7.6 above, and the following convergence and mesh independence result holds (see pages 164–166 in [33]).

**THEOREM 8.1.** *The AVE approximation scheme  $(H_{AVE}^N, A_{AVE}^N, B_{AVE}^N, C_{AVE}^N)$  satisfies all the assumptions in Theorem 7.6. There exist constants  $\hat{M}$  and  $\alpha$  independent of  $N$  such that*

$$\|X_\infty^N - P^N X_\infty P^N\| \leq \frac{\hat{M}}{\sqrt{N}},$$

and the finite dimensional Kleinman–Newton iterates satisfy

$$\|X_{k+1}^N - P^N X_\infty P^N\| \leq \alpha \|X_k^N - X_\infty^N\|^2 + \frac{\hat{M}}{\sqrt{N}}.$$

Note that the overall convergence rate for the AVE scheme is  $O(1/\sqrt{N})$ . In order to improve this rate, several “high order” spline-based schemes were proposed. The first of these schemes was developed by Banks and Kappel in [5]. Because this spline scheme failed to produce strongly convergent Riccati operators, several modifications were developed to overcome this issue. A nice summary of these schemes and their properties can be found in Kappel’s survey paper [35]. We briefly describe the scheme below.

**The Banks–Kappel (BK) spline scheme.** We now describe the “BK” finite element spline-based scheme first proposed by Banks and Kappel in [5]. For each  $N > 1$ , create a partition on  $[-r, 0]$  by defining  $\tau_j^N = -jr/N$ , where  $j = 0, \dots, N$ . For ease of notation we set  $\tau_{N+1}^N = -r$  and  $\tau_{-1}^N = 0$ . On  $[-r, 0]$ , define the standard linear B-splines by

$$B_j^N(s) = \begin{cases} \frac{N}{r}(s - \tau_{j+1}^N), & s \in [\tau_{j+1}^N, \tau_j^N], \\ \frac{N}{r}(\tau_{j-1}^N - s), & s \in [\tau_j^N, \tau_{j-1}^N], \\ 0 & \text{otherwise.} \end{cases}$$

Define the finite dimensional subspace  $H_{BK}^N$  of  $H$  by

$$(8.10) \quad H_{BK}^N \equiv \left\{ (\phi^N(0), \phi^N(\cdot)) \in H : \phi^N(s) = \sum_{j=0}^N v_j^N B_j^N(s), v_j^N \in \mathbb{R}^n \right\}.$$

Let  $P^N$  denote the orthogonal projection of  $H$  into  $H_{BK}^N$  and note that since  $H_{BK}^N \subseteq D(A) \subseteq H$ , the range of  $P^N$  is contained in the domain of  $A$ . Therefore, we define the spline approximation  $A_{BK}^N : H_{BK}^N \rightarrow H_{BK}^N \subseteq H$  by

$$(8.11) \quad A_{BK}^N = P^N A = P^N A P^N.$$

In order to complete the approximation scheme, we define

$$(8.12) \quad B_{BK}^N = P^N B \quad \text{and} \quad C_{BK} = C P^N,$$



and this yields the BK spline approximation scheme  $(H_{BK}^N, A_{BK}^N, B_{BK}^N, C_{BK}^N)$ .

**Note.** The BK spline scheme satisfies Assumption 7.1 and hence yields a convergent numerical scheme in the sense that, for a given initial condition and input function, the approximations of the forward problem converge on finite time intervals. However, unlike the AVE scheme above, the BK spline scheme fails to satisfy the dual convergence assumption, Assumption 7.2 (see [16]), and the uniformly stabilizable and detectable assumption, Assumption 7.3 (see [17] and [18]). The BK spline scheme does satisfy the basic assumption, Assumption 7.7, so that the finite dimensional Kleinman–Newton algorithm converges quadratically with constant  $c^N$  possibly depending on  $N$ . However, the approximating Riccati operators do not converge strongly so, in particular, norm convergence fails.

In the next section we present numerical results based on these two schemes. The numerical results will confirm (as Theorem 8.1 implies) that the AVE scheme is mesh independent and the approximating Riccati operators converge. However, the numerical results below also show that the BK spline scheme is not mesh independent, although there is a bound on  $c^N$ .

**9. Numerical results.** In this section we illustrate the importance of Assumptions 7.1, 7.2, and 7.3 in obtaining strong convergence (norm convergence) of feedback gain operators as well as strong mesh independence of the Kleinman–Newton iterations.

For this discussion we use the two schemes discussed in section 8.1. The AVE scheme satisfies all the assumptions of Theorems 7.6 and 7.10. Therefore, both forms of strong mesh independence, (2.4) and (2.7), are satisfied and the approximate Riccati operators converge in norm to  $X_\infty$ . This is not the case for the BK scheme since it fails to satisfy Assumptions 7.2 and 7.3. In the numerical approximations below,  $X_\infty$  is taken as the (converged) fine grid solution of the Riccati equation using the AVE scheme.

In this section we use the following notation: Let  $\hat{c}_{AVE}^N$  and  $\hat{c}_{BK}^N$ , respectively, denote the values  $\hat{c}^N$  if the AVE and BK schemes are used for the approximations. Also, let  $\hat{M}_{AVE}^N(\varepsilon, x_0^N)$  and  $\hat{M}_{BK}^N(\varepsilon, x_0^N)$  denote the values  $\hat{M}^N(\varepsilon, x_0^N)$  for the AVE and BK schemes.

Mesh independence implies that a finite dimensional process behaves asymptotically the same as the underlying infinite dimensional process. Thus, in order to compare the behavior of the approximation schemes, it is necessary that the starting operators in the approximation spaces are the projections of the starting operator in the infinite dimensional space onto the respective approximation spaces. To accomplish this,  $X_{0,AVE}^N$  and  $X_{0,BK}^N$  are expressed in terms of multiples of the identity operator in the respective approximation spaces. For the other obvious choice, the zero operator, the convergence was too fast to make observations about quadratic convergence or mesh independence. Since the mass matrices,  $\text{MASS}_{AVE}$  and  $\text{MASS}_{BK}$ , are the projections of  $I_{\mathbb{R} \times L_2(0,1)}$  onto the approximation spaces  $H_{AVE}^N$  and  $H_{BK}^N$ , respectively, the starting matrices will be multiples of the appropriate mass matrices.

All computations in this section have been performed on a PowerPC G5, 2.7GHz, using MATLAB version 7.0.0. All Lyapunov equations are solved by implementing the MATLAB Lyapunov solver which uses the SLICOT routines SB03MD and SG03AD.

**Numerical Example 1.** The results presented here are typical for all the runs on a one-dimensional delay equation,

$$\dot{x}(t) = x(t) + x(t-1) + u(t),$$

with cost function

$$J(u(\cdot)) = \int_0^{+\infty} \{10^4[x(t)]^2 + [u(t)]^2\} dt.$$

The starting operator,  $X_0$ , equals 100 times the identity in  $H = \mathbb{R} \times L_2(0, 1)$ ; thus  $X_0 = 100I_{\mathbb{R} \times L_2(0,1)}$ . The finite dimensional approximations for  $X_0$  using the AVE and BK schemes result in  $X_{0,AVE}^N = 100\text{MASS}_{AVE}$  and  $X_{0,BK}^N = 100\text{MASS}_{BK}$ , respectively. The tolerance is set to be  $\|X_k^N - X_\infty^N\| < 10^{-8} = \varepsilon$ .

Since the AVE scheme satisfies the criteria for both forms of strong mesh independence, (2.4) and (2.7), we expect mesh-independent behavior of the quantities in Table 9.1. Indeed, we notice that  $\hat{c}_{AVE}^N \rightarrow 10^{-2}$  and  $\hat{M}_{AVE}^N \rightarrow 3$ , confirming the theoretical results.

For the BK scheme this observation cannot be made from the numerical results. This is in line with the fact that the BK scheme fails to satisfy Assumptions 7.2 and 7.3. Note that mesh independence would imply that  $\hat{c}_{BK}^N \rightarrow \hat{c} \approx 10^{-2}$  and  $\hat{M}_{BK}^N \rightarrow M(10^{-8}, 100I) \approx 3$  based on the results from the AVE scheme. The results for the BK scheme show that  $\hat{c}_{BK}^{1024} \approx 1.6 \times 10^2 \gg \hat{c}$  and  $\hat{M}_{BK}^N \geq 5 > M(10^{-8}, 100I)$ .

A further comparison of the two approximation schemes includes the actual CPU-time per Newton iteration that was used by the MATLAB process. This was computed using the MATLAB function `cputime` as well as `tic` and `toc`. The resulting times were identical. Both schemes use roughly the same amount of CPU-time per iteration. For example, for  $N = 256$ , the size of the problem is 257. The AVE scheme uses on average 10.9s of CPU-time per iteration, and the BK scheme uses 10.6s. Consequently, the number of iterations that the two schemes use is a direct measure of the total computational time. In general, the BK scheme needs more iterations than the AVE scheme to obtain a specific accuracy, and in some cases significantly more.

**Numerical Example 2.** We present typical results for the two-dimensional delay equation,

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -1.6 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 & 0 \\ -1 & -1 \end{bmatrix} x(t-1) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t),$$

with cost function

$$J(u(\cdot)) = \int_0^{+\infty} \left\{ \left( \begin{bmatrix} 10 & 0 \\ 0 & 0 \end{bmatrix} x(t) \right)^2 + [u(t)]^2 \right\} dt.$$

The starting operator,  $X_0$ , equals twice the identity in  $H = \mathbb{R}^2 \times L_2(0, 1; \mathbb{R}^2)$ ; thus  $X_0 = 2I_{\mathbb{R}^2 \times L_2(0,1;\mathbb{R}^2)}$ . For the AVE and BK schemes, the starting matrices are  $X_{0,AVE}^N = 2\text{MASS}_{AVE}$  and  $X_{0,BK}^N = 2\text{MASS}_{BK}$ , respectively. As before,  $\varepsilon$  is taken to be  $10^{-8}$ ,  $\|X_k^N - X_\infty^N\| < 10^{-8}$ .

The results presented in Table 9.2 are similar to the results observed in Example 1. The AVE scheme yields strong mesh independence, and the Riccati operators converge strongly, while this is not true for the BK scheme. In particular, the optimal feedback law defined by (5.3) has the form

$$Kz(t) = k_0z(t) + \int_{-1}^0 k_1(s)z(t+s)ds + \int_{-1}^0 k_2(s)z(t+s)ds,$$

where  $k_i(s)$ ,  $i = 1, 2$ , are called the functional gains. Figures 9.1 and 9.2 illustrate that the AVE scheme (the solid line) yields strong convergence of the gains, while the BK scheme (the dashed line) does not.

TABLE 9.1

$N$	$\hat{c}_{AVE}^N$	$\hat{M}_{AVE}^N$	$\hat{c}_{BK}^N$	$\hat{M}_{BK}^N$
8	$1.01497 \times 10^{-2}$	3	$3.98797 \times 10^6$	14
16	$6.12638 \times 10^{-3}$	3	$3.31429 \times 10^6$	12
32	$6.18756 \times 10^{-3}$	3	$1.81757 \times 10^6$	10
64	$6.57817 \times 10^{-3}$	3	$5.04459 \times 10^5$	8
128	$7.11713 \times 10^{-3}$	3	$3.04030 \times 10^4$	7
256	$7.78503 \times 10^{-3}$	3	$7.82120 \times 10^4$	7
512	$8.55008 \times 10^{-3}$	3	$6.02653 \times 10^3$	6
1024	$9.32057 \times 10^{-3}$	3	$1.59829 \times 10^2$	5

TABLE 9.2

$N$	$\hat{c}_{AVE}^N$	$\hat{M}_{AVE}^N$	$\hat{c}_{BK}^N$	$\hat{M}_{BK}^N$
8	$4.77751 \times 10^{-2}$	8	$1.28997 \times 10^6$	16
16	$4.94581 \times 10^{-2}$	8	$8.69896 \times 10^5$	14
32	$4.98913 \times 10^{-2}$	8	$4.56992 \times 10^4$	12
64	$5.04197 \times 10^{-2}$	9	$3.59473 \times 10^3$	11
128	$5.07046 \times 10^{-2}$	9	$9.31569 \times 10^3$	11
256	$5.08547 \times 10^{-2}$	9	$4.18416 \times 10^1$	10
512	$5.09322 \times 10^{-2}$	9	$2.38082 \times 10^1$	10

As in Numerical Example 1, the two schemes use roughly the same CPU-time per Newton iteration. For  $N = 256$ , the problem size is 514, and the average CPU-time per iteration is 102s for the AVE scheme and 105s for the BK scheme.

We note that Figures 9.1 and 9.2 verify the theoretical results in this paper as well as those established in the earlier paper by Burns, Ito, and Propst [16]. In particular, in [16] it was proved that the BK scheme does not produce approximating Riccati operators that converge in norm. Hence the oscillations seen in these figures, which are indicative of weak convergence, are the best one can expect. However, the AVE scheme is norm convergent and this is also illustrated in Figures 9.1 and 9.2.

We close this section with an example that illustrates the need for infinite dimensional feedback. As noted earlier, the convergence theory in Damm and Hinrichsen [23] is easily applied to a wide variety of finite dimensional (matrix) Riccati equations. However, applying this method to infinite dimensional Riccati equations is not straightforward. A special feedback problem for a delay system was used to illustrate their results. They considered the problem of stabilizing a delay system with finite dimensional feedback only, which leads to a finite dimensional Riccati type (matrix) equation. In particular, the control system considered in [23] is given by the delay differential equation

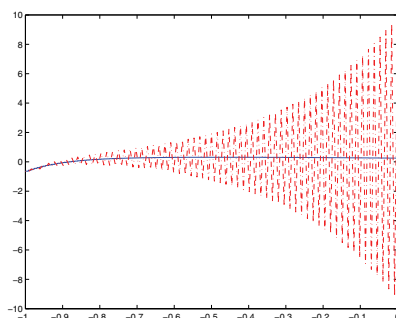
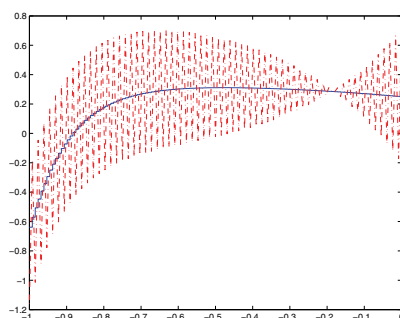
$$\dot{x}(t) = A_0x(t) + A_1x(t-r) + Bu(t).$$

The problem (see problem 5 on page 58 in [23]) is to find a finite dimensional feedback law of the form

$$u(t) = -Kx(t)$$

so that the closed-loop system

$$\dot{x}(t) = [A_0 - BK]x(t) + A_1x(t-r)$$

FIG. 9.1. *Numerical Example 2: Functional gain  $k_1$ .*FIG. 9.2. *Numerical Example 2: Functional gain  $k_2$ .*

is stable for all delays  $r > 0$ . This leads to a matrix Riccati equation for the matrix  $K$ .

If one considers the retarded delay equation,

$$(9.1) \quad \dot{x}(t) = \alpha x(t) + \beta x(t-r) + \gamma \int_{-r}^0 x(t+s)ds,$$

then one will know (see Corollary 2.8 in [32]) that (9.1) is stable independent of delay if and only if

$$\alpha < 0, \quad \gamma < 0, \quad 0 < -\gamma \leq \frac{\alpha^2 - \beta^2}{2}.$$

If  $\gamma > 0$ , then (9.1) is not stable for all  $r > 0$ . If one starts with the control system

$$(9.2) \quad \dot{x}(t) = x(t) + 2x(t-r) + \int_{-r}^0 x(t+s)ds + u(t)$$

and uses only current state feedback

$$u(t) = -kx(t),$$

then the closed-loop system has the form

$$(9.3) \quad \dot{x}(t) = [1 - k]x(t) + 2x(t - r) + \int_{-r}^0 x(t + s)ds,$$

and this system is never stable independent of delay since  $\gamma = +1$ . On the other hand, the complete state (infinite dimensional) feedback law,

$$u(t) = -k_0x(t) + (k_1 - 1) \int_{-r}^0 x(t + s)ds,$$

leads to the closed loop system

$$(9.4) \quad \dot{x}(t) = [1 - k_0]x(t) + 2x(t - r) + k_1 \int_{-r}^0 x(t + s)ds,$$

which is stable independent of delay if and only if

$$1 < k_0, \quad k_1 < 0, \quad 0 < -k_1 \leq \frac{[1 - k_0]^2 - 4}{2}.$$

If we set  $k_0 = 4$  and  $k_1 = 5/4 < 5/2$ , then (9.4) is stable independent of delay. In particular, the control system (9.2) is stable independent of delay, and the MIP holds if we apply the AVE scheme to this problem. Numerical results on mesh independence and convergence for this problem are almost identical to the previous two numerical examples and will not be presented here.

**10. Conclusions.** The theoretical results above provide precise conditions on approximation schemes needed to guarantee an MIP. The numerical results are interesting for two reasons. First, they provide numerical support for the mesh independence of the AVE scheme. Also, since the BK scheme does not generate norm convergent Riccati solutions, it is certainly not a mesh-independent scheme. However, the numerical results alone might be used to *incorrectly* justify some type of mesh independence.

There are many PDE control problems in which the linearization is not normal. For example, in channel flow control, when one linearizes about a nonzero equilibrium, the resulting  $A$  operator is highly nonnormal. Thus dual convergence is extremely important. Moreover, we have tested the theoretical results above on self-adjoint parabolic PDE control systems such as the ones considered by Banks and Kunisch [7]. Since dual convergence is not an issue and in [7] POES is established for this class of problems, our results imply mesh independence for standard finite element schemes. We have also applied the theory to some non-self-adjoint PDE problems. These PDE results, along with numerical examples, will appear in a forthcoming paper.

We have established a mesh independence result for the infinite dimensional version of the Kleinman-Newton algorithm for solving the algebraic Riccati operator equation associated with the LQR problem in a Hilbert space. We applied the results to systems governed by delay equations and presented numerical examples to illustrate the ideas. The results provide insight into the type of approximation schemes that lead to mesh independence. In particular, we showed that it is sufficient that the approximation be convergent, dual convergent, and uniformly stabilizable and detectable. As noted by Kappel in [35], it is possible to obtain (at least strong) convergence of the approximating Riccati operators without POES. This leaves open the

question of whether or not it is possible to achieve mesh independence without preserving stabilizability and detectability *uniformly* under approximation. However, it is important to note again that mesh independence alone does not imply convergence. We are currently looking into this issue and other issues concerning the numerical conditioning of the finite dimensional approximating Riccati equations.

**Acknowledgments.** The authors wish to thank the referees for their feedback and many helpful suggestions.

#### REFERENCES

- [1] E. L. ALLGOWER, K. BÖHMER, F. A. POTRA, AND W. C. RHEINOLDT, *A mesh-independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.
- [2] E. L. ALLGOWER AND K. BÖHMER, *Application of the mesh independence principle to mesh refinement strategies*, SIAM J. Numer. Anal., 24 (1987), pp. 1335–1351.
- [3] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: Numerical methods based on averaging approximations*, SIAM J. Control Optim., 16 (1978), pp. 169–208.
- [4] H. T. BANKS, J. A. BURNS, AND E. M. CLIFF, *Parameter estimation and identification for systems with delays*, SIAM J. Control Optim., 19 (1981), pp. 791–828.
- [5] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.
- [6] H. T. BANKS AND K. KUNISCH, *An approximation theory for nonlinear partial differential equations with applications to identification and control*, SIAM J. Control Optim., 20 (1982), pp. 815–849.
- [7] H. T. BANKS AND K. KUNISCH, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684–698.
- [8] H. T. BANKS, I. G. ROSEN, AND K. ITO, *A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 830–855.
- [9] P. BENNER, *Solving large-scale control problems*, IEEE Control Systems Magazine, 24 (2004), pp. 44–59.
- [10] P. BENNER AND J. SAAK, *Linear-Quadratic Regulator Design for Optimal Cooling of Steel Profiles*, Preprint SFB393/05, TUChemnitz, Chemnitz, Germany, 2005.
- [11] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite Dimensional Systems*, Vol. I, Birkhäuser, Boston, Basel, Berlin, 1992.
- [12] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite Dimensional Systems*, Vol. II, Birkhäuser, Boston, Basel, Berlin, 1993.
- [13] J. BORGAARD, J. A. BURNS, E. VUGRIN, AND L. ZIETSMAN, *On strong convergence of feedback operators for non-normal distributed parameter systems*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas, 2004, pp. 1526–1531.
- [14] P. BUBÁK, C. V. M. VAN DER MEE, AND A. C. M. RAN, *Approximation of solutions of Riccati equations*, SIAM J. Control Optim., 44 (2005), pp. 1419–1435.
- [15] J. A. BURNS, *Nonlinear distributed parameter control systems with non-normal linearizations: Applications and approximations*, in Research Directions in Distributed Parameter Systems, R. C. Smith and M. A. Demetriou, eds., SIAM, Philadelphia, 2003, pp. 17–53.
- [16] J. A. BURNS, K. ITO, AND G. PROPST, *On nonconvergence of adjoint semigroups for control systems with delays*, SIAM J. Control Optim., 26 (1988), pp. 1442–1454.
- [17] J. A. BURNS AND G. H. PEICHL, *Preservation of controllability under approximation and controllability radii for hereditary systems*, Differential Integral Equations, 2 (1989), pp. 439–452.
- [18] J. A. BURNS AND G. H. PEICHL, *A note on the asymptotic behavior of controllability radii for a scalar hereditary system*, in Proceedings of the 4th International Conference on Control of Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser, Basel, 1989, pp. 27–39.
- [19] J. A. BURNS AND G. H. PEICHL, *Control system radii and robustness under approximation*, in Robust Optimization-Directed Design, A. Kurdila, P. Pardalos, and M. Zabrankin, eds., Springer, New York, 2006, pp. 25–61.

- [20] J. A. BURNS AND J. R. SINGLER, *Modeling transition: New scenarios, system sensitivity, and feedback control*, in Transition and Turbulence Control, Lecture Notes Ser. 8, Institute for Mathematical Sciences, National University of Singapore, M. Gad-el-Hak and H. M. Tsai, eds., World Scientific Publishers, Singapore, 2006, pp. 1–37.
- [21] F. CHATELIN, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1983.
- [22] R. F. CURTAIN AND H. J. ZWART, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer, New York, Berlin, Heidelberg, 1995.
- [23] T. DAMM AND D. HINRICHSSEN, *Newton's method for concave operators with resolvent positive derivatives in ordered Banach spaces*, Linear Algebra Appl., 363 (2003), pp. 43–64.
- [24] A. DE SANTIS, A. GERMANI, AND L. JETTO, *Approximation of the algebraic Riccati equation in the Hilbert space of Hilbert-Schmidt operators*, SIAM J. Control Optim., 31 (1993), pp. 847–874.
- [25] A. GERMANI, L. JETTO, AND M. PICCIONI, *Galerkin approximation for optimal linear filtering of infinite-dimensional linear systems*, SIAM J. Control Optim., 26 (1988), pp. 1287–1305.
- [26] A. GERMANI, C. MANES, AND P. PEPE, *A twofold spline approximation for finite horizon LQG control of hereditary systems*, SIAM J. Control Optim., 39 (2000), pp. 1233–1295.
- [27] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537–565.
- [28] J. S. GIBSON, *Linear-quadratic optimal control of hereditary differential systems: Infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95–139.
- [29] J. S. GIBSON AND A. ADAMIAN, *Approximation theory for linear-quadratic-Gaussian optimal control of flexible structures*, SIAM J. Control Optim., 29 (1991), pp. 1–37.
- [30] L. GRASEDYCK, W. HACKBUSCH, AND B. N. KHOROMSKIJ, *Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices*, Computing, 70 (2003), pp. 121–165.
- [31] N. J. HIGHAM, M. KONSTANTINOV, V. MEHRMANN, AND P. PETKOV, *The sensitivity of computational control problems*, IEEE Control Systems Magazine, 24 (2004), pp. 28–43.
- [32] C. HUANG AND S. VANDEWALLE, *An analysis of delay-dependent stability for ordinary and partial differential equations with fixed and distributed delays*, SIAM J. Sci. Comput., 25 (2004), pp. 1608–1632.
- [33] K. ITO, *Strong convergence and convergence rates of approximating solutions for algebraic Riccati equations in Hilbert spaces*, in Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Springer, New York, 1987, pp. 151–166.
- [34] K. ITO, *Finite-dimensional compensators for infinite-dimensional systems via Galerkin-type approximations*, SIAM J. Control Optim., 28 (1990), pp. 1251–1269.
- [35] F. KAPPEL, *Approximations of LQR-problems for delay systems: A survey*, in Computation and Control II, K. Bowers and J. Lund, eds., Birkhäuser, Cambridge, 1991, pp. 187–224.
- [36] F. KAPPEL AND D. SALAMON, *Spline approximations for retarded systems and the Riccati equation*, SIAM J. Control Optim., 25 (1987), pp. 1082–1117.
- [37] F. KAPPEL AND D. SALAMON, *On the stability properties of spline approximations for retarded systems*, SIAM J. Control Optim., 27 (1989), pp. 407–431.
- [38] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1976.
- [39] L. A. LUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Gordon and Breach, New York, 1961.
- [40] V. L. MEHRMANN, *The Autonomous Linear Quadratic Control Problem*, Springer, Berlin, Heidelberg, 1991.
- [41] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
- [42] T. PENZL, *A Multi-Grid Method for Generalized Lyapunov Equations*, Preprint SFB393/97, TUChemnitz, Chemnitz, Germany, 1997.
- [43] I. G. ROSEN, *On Hilbert-Schmidt norm convergence of Galerkin approximations for operator Riccati equations*, in Control and Estimation of Distributed Parameter Systems, Internat. Ser. Numer. Math. 91, Birkhäuser, Basel, 1989, pp. 335–349.
- [44] I. G. ROSEN, *Convergence of Galerkin approximations for operator Riccati equations—A nonlinear evolution equation approach*, J. Math. Anal. Appl., 155 (1991), pp. 226–248.
- [45] I. G. ROSEN AND C. WANG, *A multilevel technique for the approximate solution of operator Lyapunov and Riccati equations*, SIAM J. Numer. Anal., 32 (1995), pp. 514–541.
- [46] H. SASAI, *Approximation of optimal control problems governed by nonlinear evolution equations*, Internat. J. Control, 28 (1978), pp. 313–324.

- [47] H. SASAI AND E. SHIMEMURA, *On convergence of approximating solutions of a class of linear optimal control problems of distributed parameter systems*, Bul. Inst. Politehn. Iași (N.S.), 15 (1969), pp. 39–47.
- [48] H. SASAI AND E. SHIMEMURA, *On the convergence of approximating solutions for linear distributed parameter optimal control problems*, SIAM J. Control Optim., 9 (1971), pp. 263–273.



## PROBABILISTIC ROBUSTNESS ANALYSIS—RISKS, COMPLEXITY, AND ALGORITHMS\*

XINJIA CHEN<sup>†</sup>, KEMIN ZHOU<sup>†</sup>, AND JORGE ARAVENA<sup>†</sup>

**Abstract.** It is becoming increasingly apparent that probabilistic approaches can overcome conservatism and computational complexity of the classical worst-case deterministic framework and may lead to designs that are actually safer. In this paper we argue that a comprehensive probabilistic robustness analysis requires a detailed evaluation of the robustness function, and we show that such an evaluation can be performed with essentially any desired accuracy and confidence using algorithms with complexity that is linear in the dimension of the uncertainty space. Moreover, we show that the average memory requirements of such algorithms are absolutely bounded and well within the capabilities of today's computers. In addition to efficiency, our approach permits control over statistical sampling error and the error due to discretization of the uncertainty radius. For a specific level of tolerance of the discretization error, our techniques provide an efficiency improvement upon conventional methods which is inversely proportional to the accuracy level; i.e., our algorithms get better as the demands for accuracy increase.

**Key words.** robustness analysis, risk analysis, randomized algorithms, uncertain system, computational complexity

**AMS subject classifications.** 93D09, 93D15, 68W20, 68W40

**DOI.** 10.1137/060668407

**1. Introduction.** In recent years, a number of researchers have proposed probabilistic control methods for overcoming the computational complexity and conservatism of the deterministic worst-case robust control framework (see, e.g., [1, 2, 3, 4, 5, 6, 7, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22] and the references therein).

The philosophy of probabilistic control theory is to sacrifice cases of extreme uncertainty. Such a paradigm has led to the concept of *confidence degradation function* (originated by Barmish, Lagoa, and Tempo [2]), which has been demonstrated to be extremely powerful for the robustness analysis of uncertain systems. Such a function,  $\mathcal{P}(\cdot)$ , is defined as  $\mathcal{P}(r) = \inf_{0 < \rho \leq r} \mathbb{P}(\rho)$  with

$$\mathbb{P}(\rho) = \text{vol}\{X \in \mathcal{B}_\rho \mid \text{the robustness requirement is guaranteed for } X\} / \text{vol}\{\mathcal{B}_\rho\},$$

where the volume function  $\text{vol}\{\cdot\}$  is the Lebesgue measure, and  $\mathcal{B}_\rho$  denotes the uncertainty bounding set with radius  $\rho$ . Interestingly, it was discovered in [2] that such a function is not necessarily monotone decreasing in the uncertainty radius. In view of this fact, and for the purpose of avoiding confusion with the concept of *confidence band* used in the evaluation of the accuracy of the estimate of  $\mathbb{P}(r)$ , the confidence degradation function is referred to as the *robustness function* in this paper. Accordingly, a graph representation of the robustness function is called the *robustness curve*. It can be seen that the robustness function is a natural extension of the concept of robustness margin. From the robustness curve, one can determine the *probabilistic robustness margin* [2] and estimate the deterministic robustness margin.

---

\*Received by the editors August 25, 2006; accepted for publication (in revised form) June 8, 2008; published electronically October 22, 2008. This research was supported in part by grants from NASA (NCC5-573), LEQSF (NASA/LEQSF(2001-04)-01), the NNSFC Young Investigator Award for Overseas Collaborative Research (60328304), and the NNSFC (10377004).

<http://www.siam.org/journals/sicon/47-5/66840.html>

<sup>†</sup>Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803 (chan@ece.lsu.edu, kemin@ece.lsu.edu, aravena@ece.lsu.edu).

In addition to overcoming the NP hard complexity and conservatism of deterministic robustness analysis methods, the robustness function can address very complex problems which are intractable by deterministic worst-case methods. Moreover, the probability that the robustness requirement is guaranteed can be inferred from the robustness function, while the deterministic margin loses the connection with such a probability. Based on the assumption that the density function of uncertainty is radially symmetric and nonincreasing with respect to the norm of uncertainty, it has been shown in [2] that the probability that the robustness requirement is guaranteed is no less than  $\mathcal{P}(r) = \inf_{\rho \in (0, r]} \mathbb{P}(\rho)$  when the uncertainty is included in a bounding set with radius  $r$ . The underlying assumption is supported by modeling and manufacturing considerations that the uncertainty is unstructured so that all directions are equally likely and that small perturbations from the nominal model are more likely than large perturbations. Since  $\mathbb{P}(\cdot)$  is not monotonically decreasing [2], the lower bound of the probability depends on  $\mathbb{P}(\rho)$  for all  $\rho \in (0, r]$ . It is not clear whether it is feasible to estimate  $\mathcal{P}(r)$ , since the estimation of  $\mathbb{P}(\rho)$  for every  $\rho$  relies on intensive Monte Carlo simulation and  $\mathbb{P}(\rho)$  needs to be estimated for numerous values of  $\rho$ . For such a probabilistic method to overcome the NP hard of worst-case methods, it is necessary to show that the complexity for estimating  $\mathcal{P}(r)$  for a given  $r$  is polynomial in terms of computer running *time* and memory *space*. In this paper, we demonstrate that the complexity in terms of space and time is surprisingly low and is *linear in the uncertainty dimension and the logarithm of the relative width of the range of uncertainty radius*.

In the next section we argue that both the deterministic robustness margin and its risk-adjusted version—the probabilistic robustness margin—have inherent limitations. We note that, as compared to robustness margins, the robustness function with respect to an uncertainty radius varying over a wide range can provide more insight into the system performance. In order to construct the robustness function for a wide range of uncertainty radii, the conventional method independently estimates  $\mathbb{P}(r_i)$  for each grid point of uncertainty. If there are  $m$  grid points and  $N$  is the sample size for each radius, then the total number of simulations is  $Nm$ . In section 3, we use the sample reuse principle and demonstrate that the robustness curve for an arbitrarily wide range of uncertainty radii can be accurately constructed with surprisingly low complexity. Clearly, the number of grid points,  $m$ , must tend to infinity as the tolerance tends to zero. However, we show that with our algorithms, the *equivalent number of grid points* (ENGP),  $m_{eq}$ , is strictly bounded from above in the sense that in order to guarantee the same level of accuracy for the estimation of the robustness function, the required average computational effort is the same as that of a conventional grid with  $m_{eq}$  points. Moreover, we show that the average memory requirement is also absolutely bounded and is well within the reach of modern computers.

The remainder of the paper is organized as follows. Section 2 provides an example illustrating the pitfalls of the deterministic robustness margin and the probabilistic robustness margin. Section 4 discusses the control of estimation error of the robustness function and the required complexity. Section 5 investigates the difficulties of the conventional data structure. Section 6 describes our new algorithms, analyzes the complexity of data processing and memory space, and introduces the concept of confidence band. The proposed randomized algorithms are applied to control problems in section 7. Section 8 is the conclusion. The proofs of all the theorems are included in the appendices.

Throughout this paper, all probabilistic statements are associated with the same probability space,  $(\Omega, \mathcal{F}, \text{Pr})$ , where the uncertainty is defined as a multivariate

random variable. The notations  $\Omega$ ,  $\mathcal{F}$ , and  $\Pr$  denote, respectively, the sample space,  $\sigma$ -algebra, and probability measure.

**2. The risk of robustness margins.** In this section we make the case for the need to have a robustness function in order to properly estimate how well a control system tolerates uncertainties. Conventional robust control approaches the issue with a “worst-case” philosophy. In this regard, it has been demonstrated (see Chen, Aravena and Zhou [5]) that it is not uncommon for a probabilistic controller to be significantly less risky than a deterministic worst-case control. The reasons are the “uncertainty in modeling uncertainties” and the fact that the worst-case design cannot, in some instances, be “all encompassing.” Therefore, the worst-case approach has an associated risk that usually is overlooked, while the probabilistic approach acknowledges the risk and manages it.

From manufacturing and modeling considerations, it is reasonable to assume that the density of the distribution of uncertainty decreases with an increasing uncertainty norm. Such an assumption leads to the worst-case property of uniform distribution in robustness analysis [2]. However, the decay rate of density is generally unknown to the designer. Therefore, for a given uncertainty radius  $r$ , one does not have adequate knowledge about the probability that the uncertainty is included in set  $\mathcal{B}_r$ . It is important to note that the system robustness depends critically on the distribution of the uncertainty norm.

Attempts to improve the analysis have led to the definitions of a *deterministic robustness margin* and a *probabilistic robustness margin*. Both are numbers that purportedly allow the user to estimate the tolerance to uncertainties. We contend that both can be misleading—and essentially for the same reason. To demonstrate this viewpoint, we consider a feedback system, shown in Figure 1.

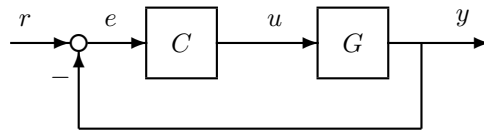


FIG. 1. Standard feedback configuration.

The transfer function of the plant is  $G(s) = \frac{q}{s-p}$ , where  $p$  and  $q$  are uncertain parameters. The uncertainty bounding set with radius  $r > 0$  is

$$\mathcal{B}_r = \{(x, y) : |x - q_0| \leq r, \quad |y - p_0| \leq r\}, \quad p_0 < 0, \quad q_0 > 0.$$

Consider two controllers  $C_A = \frac{K_A}{s+\sigma}$ ,  $\sigma > 0$ , and  $C_B = K_B$  such that

$$1 < K_B < \frac{K_A}{\sigma}, \quad \frac{K_A q_0 - \sigma p_0}{K_A + \sigma} < \sigma - p_0.$$

Suppose that the robustness requirement is stability. It can be shown that the robustness function for controller  $A$  is

$$\mathcal{P}^A(r) = \begin{cases} 1 & \text{for } 0 < r < \rho_A, \\ \frac{1}{2} - \frac{K_A \left( r + \frac{\sigma}{K_A} - q_0 \right)^2}{8\sigma r^2} - \frac{p_0 - \beta}{2r} & \text{for } \rho_A \leq r \leq \rho_A^*, \\ \frac{1}{2} - \frac{(r + \beta - p_0) \left( r + \frac{\sigma(\beta + p_0 - r)}{2K_A} - q_0 \right)}{4r^2} - \frac{(p_0 - \beta)}{2r} & \text{for } r > \rho_A^*, \end{cases}$$

where

$$\rho_A = \frac{K_A q_0 - \sigma p_0}{K_A + \sigma}$$

is the deterministic robustness margin,  $\beta = \min(\sigma, p_0 + r)$ , and  $\rho_A^* = \frac{K_A q_0 - \sigma p_0}{K_A - \sigma}$ . It can be shown that the robustness function for controller  $B$  is given by

$$\mathcal{P}^B(r) = \begin{cases} 1 & \text{for } 0 < r < \rho_B, \\ 1 - \frac{K_B \left(r + \frac{p_0+r}{K_B} - q_0\right)^2}{8r^2} & \text{for } \rho_B \leq r \leq \rho_B^*, \\ \frac{1}{2} - \frac{\frac{p_0}{K_B} - q_0}{2r} & \text{for } r > \rho_B^*, \end{cases}$$

where

$$\rho_B = \frac{K_B q_0 - p_0}{K_B + 1}$$

is the deterministic robustness margin and  $\rho_B^* = \frac{K_B q_0 - p_0}{K_B - 1}$ .

We consider an example with  $p_0 = -10$ ,  $q_0 = 50$ ,  $\sigma = 40$ ,  $K_A = 100\sigma$ , and  $K_B = 10$ . The corresponding robustness functions are displayed in Figure 2. We obtained deterministic margins  $\rho_A = 49.6040$  and  $\rho_B = 46.3636$ . Since  $\rho_A > \rho_B$ , a comparison based on the deterministic margin simply suggests that controller  $A$  is more robust than controller  $B$ . Quite to the contrary, a judgment based on the robustness curves indicates that controller  $B$  may be more robust. The risk of the probabilistic robustness margin can also be illustrated by this example.

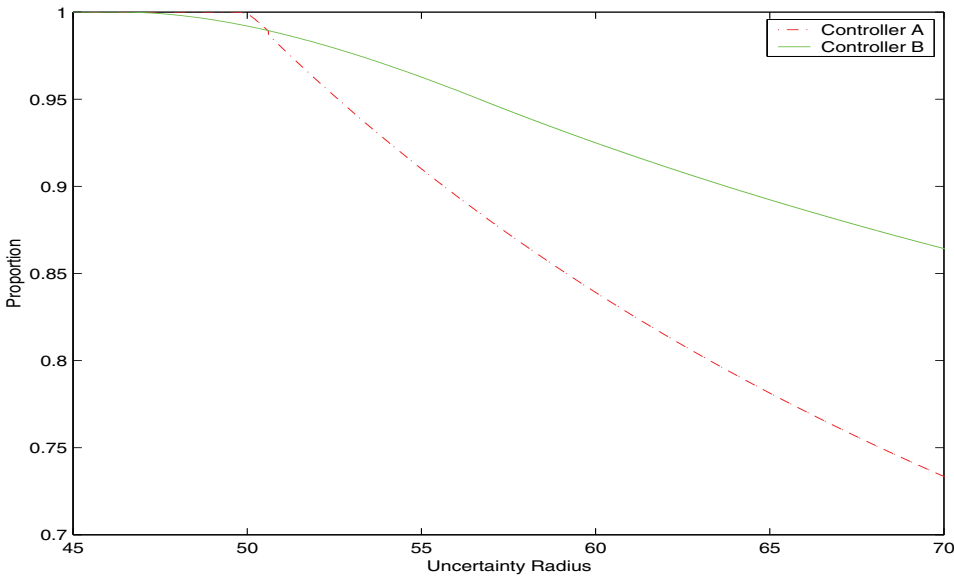


FIG. 2. Comparison of controller alternatives.

Robust analysis should be able to help a designer reliably determine which controller design is more robust. However, it appears that the concepts of robustness margin fail to meet such fundamental needs of control engineering. On the other

hand, the robustness curve serves the purpose of giving the designer complete information on how well a control system tolerates uncertainties.

From the previous discussion, it can be seen that there are two crucial factors to be considered in order to make a reliable judgment about the system robustness:

- (i) How fast the robustness curve rolls off.
- (ii) The dependency of coverage probability of uncertainty bounding set  $\mathcal{B}_r$  on the radius  $r$ . Here the coverage probability refers to the probability that the uncertainty is included in the bounding set.

The second factor can be difficult since a designer generally lacks knowledge of the coverage probability corresponding to a bounding set of fixed radius. To overcome such a difficulty, the only choice is to construct the robustness curve for a wide range of uncertainty radius. The construction of the robustness curve may be seen as a computationally challenging task since the probability of guaranteeing a robustness requirement needs to be estimated for many values of uncertainty radius. However, as we demonstrate in the next section, by using the sample reuse principle one can construct the robustness curve for virtually the entire scope of uncertainty range  $(0, \infty)$  with absolutely bounded average computational complexity, regardless of the size of the grid. For example, we shall show that for an uncertainty range as large as  $(10^{-10}, 10^{10})$ , on average, one needs less than 50 times memory and computational resources than those needed to evaluate the uncertainty range  $(1, e)$  with the same resolution.

**3. Equivalent number of grid points.** Throughout this paper, we assume that the uncertainty sets are homogeneous star-shaped (see, e.g., [2]). That is, the uncertainty bounding set with radius  $r$  is  $\mathcal{B}_r = \{rX \mid X \in \mathcal{B}_1\}$ , where  $\mathcal{B}_1$  denotes the uncertainty bounding set such that  $cX \in \mathcal{B}_1$  for any  $X \in \mathcal{B}_1$  and any  $c \in [0, 1]$ . Clearly, most of the commonly used uncertainty bounding sets such as the  $l_p$  balls and spectral norm balls are homogeneous star-shaped.

We shall consider the problem of constructing the robustness curve for the arbitrary robustness requirement under the assumption of uncertainty sets. Conventionally, the robustness curve for a range of uncertainty radii  $[\frac{a}{\lambda}, a]$  with  $a > 0$ ,  $\lambda > 1$  is constructed by choosing a set of grid points  $\frac{a}{\lambda} = r_1 < r_2 < \dots < r_m = a$  and, for every grid point, performing  $N$  independent and identically distributed (i.i.d.) Monte Carlo simulations. Hence, the total number of simulations is a deterministic constant  $mN$ . To reduce computational complexity, we shall make use of the following intuitive concept:

*Let  $X$  be an observation of a random variable with uniform distribution over  $\mathcal{B}_\rho \supseteq \mathcal{B}_r$  such that  $X \in \mathcal{B}_r$ . Then  $X$  can also be viewed as an observation of a random variable with uniform distribution over  $\mathcal{B}_r$ .*

In order to apply this concept, it is necessary to perform the simulation in a backward direction so that appropriate evaluations of the robust requirement for larger uncertainty sets can be saved for later use in simulations on smaller uncertainty sets [6]. The sample reuse principle allows a single simulation to be used for multiple radii. Thus, the actual total number of simulations is significantly reduced if the dimension of the uncertainty is not too high. In order to quantify this reduction, we introduce the *equivalent number of grid points* (ENGP),  $m_{\text{eq}}$ , defined as

$$m_{\text{eq}} = \frac{\text{expected total number of simulations}}{N}.$$

In our approach, the number of simulations required at uncertainty radius  $r_i$ , denoted by  $\mathbf{n}_i$  for  $i = 1, \dots, m$ , is a random number. The total number of simulations

can be represented by the random variable  $\mathbf{n} = \sum_{i=1}^m \mathbf{n}_i$ . The expected value of the total number of simulations is  $\mathbb{E}[\mathbf{n}] = \sum_{i=1}^m \mathbb{E}[\mathbf{n}_i]$ , where  $\mathbb{E}[X]$  denotes the expectation of random variable  $X$ . Hence, we can formally define

$$m_{\text{eq}} = \frac{\mathbb{E}[\mathbf{n}]}{N}.$$

Due to sample reuse, we can achieve a substantial reduction of simulations, i.e.,  $\mathbb{E}[\mathbf{n}] \ll mN$ . To quantify the reduction of the computational effort, we have introduced the notion of *sample reuse factor* [6], which is defined as

$$(3.1) \quad \mathcal{F}_{\text{reuse}} = \frac{mN}{\mathbb{E}[\mathbf{n}]} = \frac{m}{m_{\text{eq}}}.$$

In our approach,  $N$  i.i.d. simulation results are collected for each grid point. Hence, the accuracy of estimation is the same as that of the conventional method. However, the average number of simulations in our approach is  $\mathbb{E}[\mathbf{n}]$ , which is equivalent to the complexity of  $m_{\text{eq}}$  grid points in the conventional scheme. As a direct consequence of Theorem 1 of [6], we have that, for any discretization scheme,  $m_{\text{eq}}$  is independent of the sample size  $N$ . Moreover, we have the following general result.

**THEOREM 3.1.** *Let  $d$  be the dimension of uncertainty parameter space. Then, for an arbitrary gridding scheme, the equivalent number of grid points based on the principle of sample reuse is strictly bounded from above by  $1 + d \ln \lambda$ , i.e.,*

$$m_{\text{eq}} < 1 + d \ln \lambda.$$

See Appendix A for a proof. As mentioned at the end of the introduction, all probabilistic statements in this paper refer to the same probability space  $(\Omega, \mathcal{F}, \text{Pr})$  used to define the uncertainty variable. To make this possible, one can define  $m \times N$  mutually independent random variables  $X_{ij}$ ,  $i = 1, \dots, m$ ;  $j = 1, \dots, N$ , on the same probability space  $(\Omega, \mathcal{F}, \text{Pr})$  such that for each  $i$ ,  $X_{ij}$ ,  $j = 1, \dots, N$ , are i.i.d. random variables uniformly distributed over the bounding set with radius  $r_i$ . Each simulation corresponds to obtaining an observation for  $X_{ij}$  and evaluating the performance of the system for that observation. Without sample reuse, one needs to obtain  $m \times N$  observations corresponding to these  $m \times N$  random variables. The idea of sample reuse is to reuse observations and the corresponding evaluation results of system performance. Clearly, the number of simulations  $\mathbf{n}_i$ , associated with uncertainty radius  $r_i$ , is a random variable which can be defined as a function of random variables  $X_{ij}$ ,  $i = 1, \dots, m$ ;  $j = 1, \dots, N$ . Needless to say, the definition of such a function is complex. It can be shown that other variables can be implicitly defined on the same probability space  $(\Omega, \mathcal{F}, \text{Pr})$ .

By an “arbitrary” discretization scheme, we mean two things: (i) the number of grid points can be arbitrarily large, and (ii) the grid points can be distributed arbitrarily over the specified range of uncertainty radius.

A fundamental question of robust control is whether randomized algorithms have polynomial complexity. In light of the fact that the cost of each simulation depends on problem cases, the computational complexity is usually measured in terms of the number of simulations. This theorem reveals the following important facts:

- (a) The complexity is linear in the dimension of the uncertainty space.
- (b) The complexity depends linearly on the logarithm of the “relative” width,  $\lambda$ , of the interval of uncertainty radii. This proves that our algorithms are capable of estimating the robustness function for a wide range of uncertainty.

- (c) Our algorithms can arbitrarily reduce the grid error while keeping the complexity strictly below a constant bound.

In order to illustrate these points, Figure 3 displays the variation of  $m_{\text{eq}}$  for various dimensions of the uncertainty space and for values of  $\lambda$  up to  $\lambda = 10^{20}$  corresponding to the uncertainty range  $(10^{-10}, 10^{10})$  (which may be deemed a good approximation to  $(0, \infty)$ ). As can be seen from Figure 3, even for dimensions as high as  $d = 1024$  the equivalent number of grid points,  $m_{\text{eq}}$ , is very reasonable.

It should be noted that the sample reuse technique does not work well for high dimension  $d$ . The reason is that, for very large  $d$ , the samples tend to concentrate on the “surface” of the bounding set, and thus the efficiency of sample reuse is diminished.

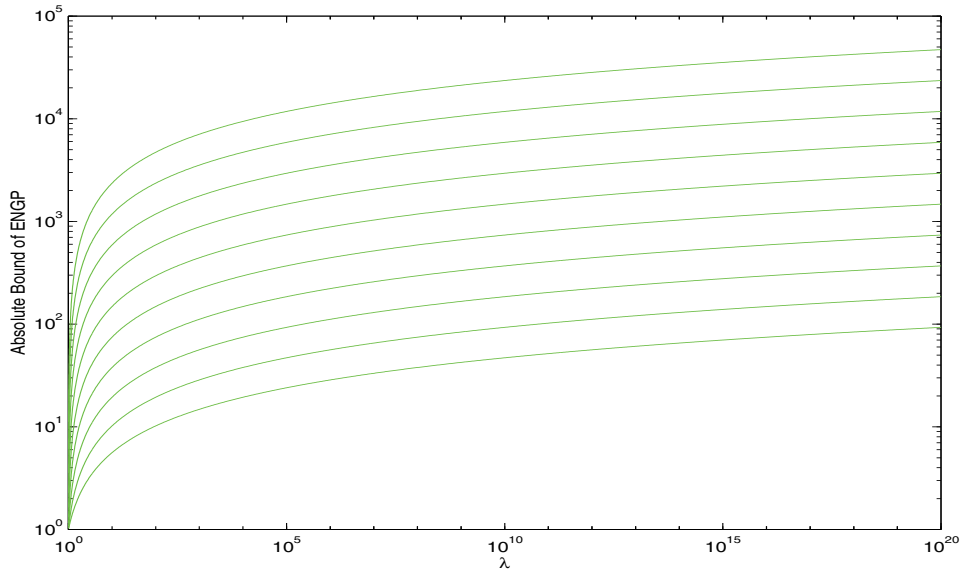


FIG. 3. Absolute bounds for  $m_{\text{eq}}$  (ENGP) ( $d = 2^i$ ,  $i = 1, \dots, 10$ ).

**4. Error control.** In addition to efficiency, another important issue in any numerical approach is error control. This point has been emphasized in many control engineering problems. For instance, when computing the  $H_\infty$  norm of a system, a lower bound and an upper bound are obtained, and it is required that the gap between them be less than a prescribed tolerance. A similar situation arises in the computation of the structured singular value ( $\mu$ ).

For the specific case of the estimation of the robustness function, there are two sources of error: (i) the statistical sampling error due to the finiteness of the sample size,  $N$  (sample size error), and (ii) the discretization error due to the finite number of points in any partition. Control of the sample size error has been well studied and emphasized. Existing techniques include the Chernoff bounds [9], binomial confidence interval [7, 10], etc. However, we claim that control of the discretization error is not sufficiently emphasized. In fact, one can argue that *controlling the sample size error can be meaningless if the discretization error is not controlled*. This will be the case, for example, for those situations where a risk at the level of a small  $\varepsilon$  (e.g.,  $\varepsilon = 0.001$ ) may be significant or unacceptable. *How can any estimation be useful if the discretization error is not ensured to be less than the tolerance  $\varepsilon$ ?*

In this section, we first introduce an interpolation result necessary for analyzing error control methods. Afterward, we discuss two different schemes which ensure a discretization error less than a given  $\epsilon \in (0, 1)$ . The first is a uniform partition whereby the uncertainty radius interval  $[\frac{a}{\lambda}, a]$  is partitioned by  $m$  points,

$$(4.1) \quad r_i = a - \frac{(m-i)(\lambda-1)}{(m-1)\lambda}a, \quad i = 1, \dots, m.$$

In the second scheme we consider a geometric-type partition of the form

$$(4.2) \quad r_i = a \left( \frac{1}{\lambda} \right)^{\frac{m-i}{m-1}}, \quad i = 1, \dots, m.$$

For any partition of the uncertainty radius interval, we have the following linear interpolation results.

**THEOREM 4.1.** *Given an arbitrary partition of the uncertainty radius interval  $[\frac{a}{\lambda}, a]$  with  $\frac{a}{\lambda} = r_1 < r_2 < \dots < r_m = a$ , define*

$$\mathbb{P}^*(r) = \frac{(r - r_i) \mathbb{P}(r_{i+1}) + (r_{i+1} - r) \mathbb{P}(r_i)}{r_{i+1} - r_i},$$

$$g(r) = (r_{i+1} - r) \left( \frac{r}{r_i} \right)^{-d} + (r - r_i) \left( \frac{r_{i+1}}{r} \right)^{-d}.$$

Then, for all  $r \in [r_i, r_{i+1}]$ ,

$$|\mathbb{P}(r) - \mathbb{P}^*(r)| \leq 1 - \frac{g(r_*)}{r_{i+1} - r_i} \leq \frac{d}{2r_i}(r_{i+1} - r_i),$$

where  $r_* \in (r_i, r_{i+1})$  is the unique solution of equation

$$\left( \frac{r_{i+1}}{r} \right)^{-d} \left[ 1 + \left( 1 - \frac{r_i}{r} \right) d \right] - \left( \frac{r}{r_i} \right)^{-d} \left[ 1 + \left( \frac{r_{i+1}}{r} - 1 \right) d \right] = 0$$

with respect to  $r$ , which can be solved by a bisection search.

See Appendix B for a proof. As mentioned before, these interpolation results will be used in the construction of a tight confidence band for the robustness function.

*Remark 1.* To guarantee a prescribed tolerance  $\epsilon \in (0, 1)$ , the number of grid points must be larger than a certain number. It has been shown by Barmish, Lagoa, and Tempo [2] that if

$$(4.3) \quad m \geq 1 + \frac{2(\lambda-1)d}{\epsilon},$$

then  $|\mathbb{P}(r) - \mathbb{P}(r_i)| < \epsilon$  for all  $r \in [r_i, r_{i+1}]$  for  $i = 1, \dots, m-1$ . This bound shows that, for fixed error  $\epsilon$ , the complexity is polynomial. From another perspective, it also shows that the number of grid points and the computational complexity tend to infinity as the tolerance tends to zero. For example, the robustness analysis problem for complex uncertainty of size  $30 \times 30$  over an interval of uncertainty with  $\lambda = 10$  requires  $m \geq 3,240,000,001$  in order to guarantee  $\epsilon \leq 10^{-5}$ . The bound, however, does not account for the sample reuse principle. Using our approach, the equivalent number of grid points for this case is bounded from above by  $1 + 1800 \times \ln(10)$ .



The following result is our extension of the result by Barmish, Lagoa, and Tempo cited above, and quantifies the advantage of using linear interpolation.

THEOREM 4.2. *Let*

$$(4.4) \quad m = 2 + \left\lfloor \frac{(\lambda - 1)d}{2\epsilon} \right\rfloor,$$

where  $\lfloor \cdot \rfloor$  denotes the floor function. Then, for a uniform gridding scheme,

$$|\mathbb{P}(r) - \mathbb{P}^*(r)| < \epsilon \quad \forall r \in [r_i, r_{i+1}]$$

for  $i = 1, \dots, m - 1$ . Moreover, the equivalent number of grid points is

$$m_{\text{eq}}(\epsilon) = m - \sum_{i=1}^{m-1} \left( 1 - \frac{1}{\frac{m-1}{\lambda-1} + i} \right)^d.$$

See Appendix C for a proof.

*Remark 2.* We point out that when using linear interpolation, the number of grid points given by (4.4) is approximately  $\frac{1}{4}$  of the bound given by (4.3).

We now analyze a discretization scheme whereby the partition of the uncertainty interval under study is defined by a geometric series.

THEOREM 4.3. *For a geometric discretization scheme with*

$$(4.5) \quad m = 2 + \left\lfloor \frac{\ln \lambda}{\ln \left( 1 + \frac{2\epsilon}{d} \right)} \right\rfloor$$

and

$$r_i = a \left( \frac{1}{\lambda} \right)^{\frac{m-i}{m-1}}$$

for  $i = 1, \dots, m$ , the following statements hold true:

(I)

$$|\mathbb{P}(r) - \mathbb{P}^*(r)| < \epsilon \quad \forall r \in [r_i, r_{i+1}], \quad i = 1, \dots, m - 1.$$

(II)

$$m_{\text{eq}}(\epsilon) = 1 + \left( 1 + \left\lfloor \frac{\ln \lambda}{\ln \left( 1 + \frac{2\epsilon}{d} \right)} \right\rfloor \right) \left[ 1 - \left( \frac{1}{\lambda} \right)^{1 + \frac{d}{\left\lfloor \frac{\ln \lambda}{\ln \left( 1 + \frac{2\epsilon}{d} \right)} \right\rfloor}} \right].$$

(III)

$$\mathcal{F}_{\text{reuse}} > \frac{1}{2\epsilon} \left( 1 - \frac{1}{1 + d \ln \lambda} \right).$$

See Appendix D for a proof.

*Remark 3.* Since  $1 + d \ln \lambda \gg 1$  in many situations, the sample reuse factor for the geometric discretization scheme may be written in a more elegant form. That is,

$$\mathcal{F}_{\text{reuse}} \approx \frac{1}{2\epsilon},$$

which is inversely proportional to the tolerance of the discretization error. For example, to ensure that the discretization error is less than  $10^{-4}$ , which is a rather weak requirement for many applications, our algorithm reduces the computational effort by a factor of 5,000 when compared to a conventional approach.

The two discretization schemes considered here, and others, have bounded complexity, but the distributions of the total number of simulations are different. Hence it is reasonable to ask if there is a “best discretization.” Our results indicate that the geometric scheme is generally more efficient, as shown by the comparison of grid points in Figure 4 and the comparison of ENGP in Figure 5.

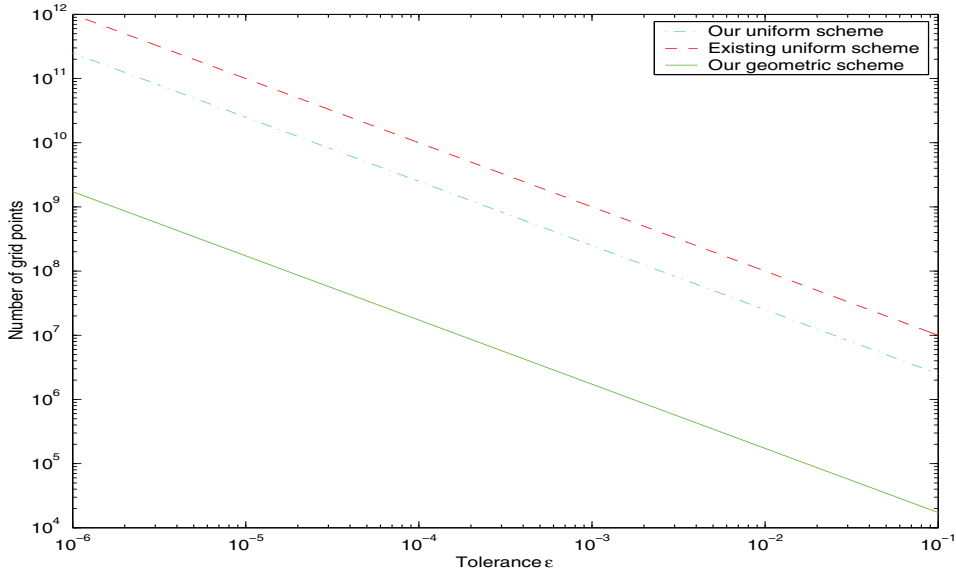


FIG. 4. Comparison of the number of grid points ( $\lambda = 10^3$ ,  $d = 500$ ).

**5. The difficulties of conventional data structure.** Our previous sample reuse algorithm [6] uses the same data structure as that of the conventional algorithm. That is, the data structure for implementing the algorithms is basically a matrix of fixed size. In such a data structure, for each grid point  $r_i$ , there is a record  $(k_i, n_i)$ , where  $k_i$  represents the number of cases guaranteeing (or violating) the robustness requirement among  $n_i$  simulations. In the course of our experiment, the number  $n_i$  increases from 0 to sample size  $N$ . In the following two subsections, we demonstrate that the conventional data structure is not suitable for controlling the error due to finite gridding.

**5.1. The issue of data processing.** Clearly, the total number of records is exactly the number of grid points  $m$ . For the conventional method, to accomplish  $N$  simulations for each grid point, the total number of times that we update the data record is  $Nm$ . As illustrated in section 4, to control the error due to finite gridding requires an extremely large number,  $m$ , of grid points even for the moderate requirement of  $\epsilon$ . Therefore,  $Nm$  is usually a *very large number*. It can be shown that *if the sample reuse algorithm employs the same data structure as that of the conventional method, then, for any gridding scheme with  $m$  grid points, the total number of times that we update the data record is also  $Nm$ .* This is true because, for every time a

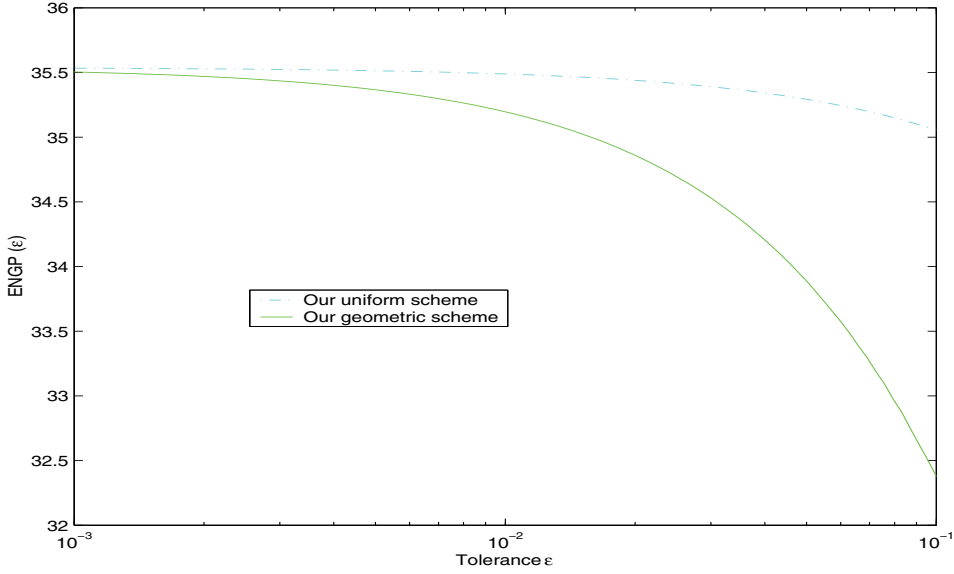


FIG. 5. Comparison of  $m_{\text{eq}}$  ( $\lambda = 10^3$ ,  $d = 5$ ,  $1 + d \ln \lambda = 35.5388$ ).

record  $(k_i, n_i)$  is updated, the number  $n_i$  can be increased only by 1, and the number  $n_i$  must be  $N$  when the experiment is completed. To see that data processing with the conventional data structure is a severe challenge, one can consider the example discussed in Remark 1 of section 4. With  $m \geq 3,240,000,001$  and normal sample size  $10^4 < N < 10^6$ , it can be seen that  $Nm$  will be in the range of  $3 \times 10^{13}$  to  $3 \times 10^{15}$ . This is an enormous burden for today's computing technology. For a modern computer with 1.9 GHz CPU and 256 M bytes RAM, it takes about 20 seconds to execute  $10^7$  times the command  $n_i \leftarrow n_i + 1$  written in the MATLAB language. It can be reasonably inferred that updating the data record  $3 \times 10^{13}$  times will take about  $20 \times 10^{-7} \times 3 \times 10^{13}$  seconds (i.e., about 700 days).

**5.2. The issue of memory space.** For the conventional data structure, the total number of records is  $m$ . To execute the sample reuse algorithm or the conventional one with such a data structure, each record must occupy some physical addresses. Such addresses are necessary for storing and visualizing the outcome of simulations. Of course, to obtain the outcome simulations may require a much higher amount of computer internal memory to execute the algorithm. Since  $m$  is usually a very large number, the consumption of memory used to store and visualize the output of simulation can be enormous. To see this, consider again the example discussed in Remark 1 of section 4. Since a floating point number occupies 2 bytes, storing a tuple of the form  $(k_i, n_i)$  needs 4 bytes. For  $m \geq 3,240,000,001$ , the data record will consume  $4 \times 3,240,000,001 \approx 13 \times 10^9$  bytes (i.e., about 13 giga bytes) of RAM. Such a requirement, just for visualizing the outcome of the simulations, is a challenging task even for modern computers.

**6. New techniques of sample reuse.** In the previous section, we have shown that any algorithm using the conventional data structure suffers from the problems of the complexity of data processing and memory space. This is because the sample size  $N$  is usually very large and the number,  $m$ , of points in the partition of uncertainty

radius approaches infinity as the tolerance,  $\epsilon$ , approaches zero (see Theorem 4.2). In this section, we shall demonstrate that, by introducing a dynamic data structure and a new sample reuse algorithm, the average requirement of memory and the computational effort devoted to data processing are absolutely bounded, independent of the tolerance, and well within the power of modern computers.

**6.1. Data structure.** In order to address the memory issue and minimize the effort devoted to data processing, an appropriate data structure is critical. The key idea is to make use of the observation that, *for a set of consecutive grid points with identical records of simulation results, it suffices to store the information of the smallest and the largest grid points.* To illustrate our techniques, we enumerate, in chronological order of generation, the samples generated from various uncertainty bounding sets as  $X_1, X_2, \dots$ . When samples  $X_1, X_2, \dots, X_j$  have been generated, *the state of the experiment is completely represented by functions  $\mathbf{s}(i, j)$  and  $\mathbf{v}(i, j)$* , where

$$\mathbf{s}(i, j) = \sum_{k=1}^{l(i, j)} Y_i^k, \quad \mathbf{v}(i, j) = \sum_{k=1}^{l(i, j)} Z_i^k$$

with

$$(6.1) \quad Y_i^k = \begin{cases} 1 & \text{if } X_k \in \mathcal{B}_{r_i}, \\ 0 & \text{otherwise,} \end{cases}$$

$$(6.2) \quad l(i, j) = \max \left\{ \ell : 1 \leq \ell \leq j, \sum_{k=1}^{\ell} Y_i^k \leq N \right\},$$

and

$$(6.3) \quad Z_i^k = \begin{cases} 1 & \text{if } X_k \in \mathcal{B}_{r_i} \text{ and the robustness requirement is violated for } X_k, \\ 0 & \text{otherwise} \end{cases}$$

for  $i = 1, \dots, m$  and  $k = 1, \dots, j$ . The definitions of  $Y_i^k$ ,  $Z_i^k$  given in (6.1) and (6.3) are based on the principle of sample reuse in that a sample of uncertainty originally generated from larger bounding sets can be reused for smaller bounding sets.

The reason we introduce variable  $l(i, j)$  by (6.2) is that, for grid point  $r_i$ , once  $N$  equivalent simulations are available, the subsequent simulations can be ignored. By the principle of sample reuse,  $\mathbf{s}(i, j)$  and  $\mathbf{v}(i, j)$  are, respectively, the numbers of samples and violations (of the robustness requirement) for an uncertainty bounding set with radius  $r_i$ , collected from various bounding sets when samples  $X_1, X_2, \dots, X_j$  have been generated. When the experiment is completed, we have  $\mathbf{n}$  samples  $X_1, X_2, \dots, X_{\mathbf{n}}$  and

$$\mathbf{s}(i, \mathbf{n}) = N, \quad \mathbb{P}(r_i) = 1 - \frac{\mathbf{v}(i, \mathbf{n})}{N}, \quad i = 1, \dots, m.$$

For each value of  $j$  in the course of the simulation experiment, it suffices to record  $\mathbf{s}(i, j)$  and  $\mathbf{v}(i, j)$  for  $i = 1, \dots, m$  in order to save the outcome of simulation. Since the number of grid points,  $m$ , can be huge, this direct method of storing simulation results requires too much memory and data processing effort. To address this problem, we need to make use of the observation that  $\mathbf{s}(i, j)$  is piecewise constant with respect to

$i$  in the sense that  $\mathbf{s}(i, j)$  is like a discrete stair function of  $i$ . Thus, to record  $\mathbf{s}(i, j)$  for  $i = 1, \dots, m$ , we need only record the values of  $i$  at which  $\mathbf{s}(i, j)$  changes and the different values of  $\mathbf{s}(i, j)$ . This method also applies to  $\mathbf{v}(i, j)$ . To implement such a method, we shall introduce two matrices  $S^j$  and  $V^j$  as follows.

Since  $\mathbf{s}(i, j)$  is piecewise constant with respect to  $i$ , there exists a matrix  $S^j$  such that, for  $i = 1, \dots, m$ ,

$$(6.4) \quad \mathbf{s}(i, j) = \begin{cases} [S^j]_{\ell, 2} & \text{for } [S^j]_{\ell, 1} \leq i < [S^j]_{\ell+1, 1} \quad \text{with } 1 \leq \ell \leq \kappa - 1, \\ [S^j]_{\kappa, 2} & \text{for } [S^j]_{\kappa, 1} \leq i \leq m, \end{cases}$$

where  $\kappa$  is the number of rows of  $S^j$  and  $[A]_{i,j}$  denotes the element of matrix  $A$  in the  $i$ th row and the  $j$ th column. According to (6.4), for any positive integer  $i$  no greater than  $m$ , we can determine  $\mathbf{s}(i, j)$  from matrix  $S^j$  based on two cases. In the case that  $[S^j]_{\kappa, 1} \leq i \leq m$ , we can set  $\mathbf{s}(i, j) = [S^j]_{\kappa, 2}$ . In the case that there exists an integer  $\ell$  such that  $1 \leq \ell \leq \kappa - 1$  and that  $[S^j]_{\ell, 1} \leq i < [S^j]_{\ell+1, 1}$ , we can set  $\mathbf{s}(i, j) = [S^j]_{\ell, 2}$ . Roughly speaking, the first column of matrix  $S^j$  records the indexes of grid points for which the accumulated numbers of samples are jumping to different values. The second column of matrix  $S^j$  records the corresponding accumulated numbers of samples.

Similarly,  $\mathbf{v}(i, j)$  is piecewise constant with respect to  $i$ , and there exists a matrix  $V^j$  such that, for  $i = 1, \dots, m$ ,

$$(6.5) \quad \mathbf{v}(i, j) = \begin{cases} [V^j]_{\ell, 2} & \text{for } [V^j]_{\ell, 1} \leq i < [V^j]_{\ell+1, 1} \quad \text{with } 1 \leq \ell \leq \tau - 1, \\ [V^j]_{\tau, 2} & \text{for } [V^j]_{\tau, 1} \leq i \leq m, \end{cases}$$

where  $\tau$  is the number of rows of  $V^j$ . According to (6.5), for any positive integer  $i$  no greater than  $m$ , we can determine  $\mathbf{v}(i, j)$  from matrix  $V^j$  based on two cases. In the case that  $[V^j]_{\tau, 1} \leq i \leq m$ , we can set  $\mathbf{v}(i, j) = [V^j]_{\tau, 2}$ . In the case that there exists an integer  $\ell$  such that  $1 \leq \ell \leq \tau - 1$  and that  $[V^j]_{\ell, 1} \leq i < [V^j]_{\ell+1, 1}$ , we can set  $\mathbf{v}(i, j) = [V^j]_{\ell, 2}$ . Loosely speaking, the first column of matrix  $V^j$  records the indexes of grid points for which the accumulated numbers of violations are jumping to different values. The second column of matrix  $V^j$  records the corresponding accumulated numbers of violations.

In this paper, matrices  $S^j$  and  $V^j$  are, respectively, referred to as the *matrix of sample sizes* and the *matrix of violations*. At any stage in which samples  $X_1, \dots, X_j$  have been generated, the status of the experiment is completely characterized by matrices  $S^j, V^j$ . Both matrices are of two columns but of varying number of rows in the course of our experiment. Obviously, the numbers of rows of these matrices can be substantially less than the number of grid points  $m$ .

To save memory and data processing effort, we shall take advantage of the piecewise constant property of the accumulated numbers of samples and violations by constructing matrices  $S^j$  and  $V^j$  when we have generated uncertainty samples  $X_1, \dots, X_j$ . As can be seen in what follows, such matrices can be constructed recursively. Once we have  $S^j$  and  $V^j$ , we can generate sample  $X_{j+1}$  and update  $S^j, V^j$  as  $S^{j+1}, V^{j+1}$  in accordance with (6.4) and (6.5).

**6.2. Sample reuse algorithm.** In this section, we shall present our sample reuse algorithms as follows.

**Initialization** We initialize the matrices of sample sizes and violations by the following steps:

◇ Generate sample  $X_1$  uniformly from uncertainty set with radius  $r_m$ .

◇ Compute  $j$  such that  $X_1 \in \mathcal{B}_{r_i}$  for  $j \leq i \leq m$  and  $X_1 \notin \mathcal{B}_{r_i}$  for  $1 \leq i \leq j-1$ .

◇ Let  $S^1 = \begin{bmatrix} 1 & 1 \end{bmatrix}$  if  $j = 1$  and  $S^1 = \begin{bmatrix} 1 & 0 \\ j & 1 \end{bmatrix}$  if  $j > 1$ .

◇ Let  $V^1 = \begin{bmatrix} 1 & 0 \end{bmatrix}$  if the robustness requirement is satisfied for  $X_1$ , and let  $V^1 = S^1$  if the robustness requirement is violated for  $X_1$ .

**Sample generation** If  $[S^j]_{\kappa,1} < N$  then generate sample  $X_{j+1}$  uniformly from uncertainty set with radius  $r_m$ , otherwise generate sample  $X_{j+1}$  uniformly from uncertainty set with radius  $[S^j]_{\kappa-1,1}$ .

**Updating matrices** Update  $S^j$  as  $S^{j+1}$  by the method described in section 6.2.1. If the robustness requirement is satisfied for  $X_{j+1}$  then let  $V^{j+1} = V^j$ , otherwise update  $V^j$  as  $V^{j+1}$  by the method described in section 6.2.2.

**Stopping criterion** The sampling process is terminated if  $S^j$  has only one row and  $[S^j]_{1,2} = N$ .

In the above, we describe the basic steps of our sample reuse algorithm. The main purpose of the initialization is to build the starting matrices  $S^1$  and  $V^1$  based on the first sample generated from the largest bounding set. The first sample must be drawn from the bounding set of the largest radius  $r_m$ . The second step of initiation is to determine the index  $j$  of the smallest bounding set which includes the sample  $X_1$ . The third step of initiation is to determine the starting matrix of sample sizes. If  $j = 1$ , then the sample  $X_1$  can be used by all bounding sets, and thus  $\mathbf{s}(i, 1) = 1$  for  $i = 1, \dots, m$ . It follows that we should let  $S^1 = \begin{bmatrix} 1 & 1 \end{bmatrix}$  in accordance with (6.4). If  $j > 1$ , then  $\mathbf{s}(i, 1) = 1$  for  $i = j, \dots, m$  and  $\mathbf{s}(i, 1) = 0$  for  $i = 1, \dots, j-1$ . Consequently, we should let  $S^1 = \begin{bmatrix} 1 & 0 \\ j & 1 \end{bmatrix}$  according to (6.4). The fourth step of initiation is to determine the starting matrix of violations. If the robustness requirement is satisfied for  $X_1$ , then  $\mathbf{v}(i, 1) = 0$  for  $i = 1, \dots, m$ , and thus we should let  $V^1 = \begin{bmatrix} 1 & 0 \end{bmatrix}$  be consistent with (6.5). On the other hand, if the robustness requirement is violated for  $X_1$ , then the matrix of violations is the same as the matrix of sample sizes.

In the step of sample generation, a new sample is generated from the currently largest bounding set for which the number of collected samples has not reached  $N$ . The new sample is used to update the data record based on (6.4) and (6.5) in the step of updating matrices. Finally, all bounding sets will have enough collected samples, and thus the simulation should be stopped when  $S^j$  has one row and  $[S^j]_{1,2} = N$ .

**6.2.1. Sample sizes tracking.** In this section, we describe how to update the matrix of sample sizes. The key idea is to ensure condition (6.4). Let  $\kappa$  be the number of rows of  $S^j$ . We proceed as follows.

**Step (1)** Compute an index  $j^*$  such that  $X_{j+1} \in \mathcal{B}_{r_i}$  for  $j^* \leq i \leq m$  and  $X_{j+1} \notin \mathcal{B}_{r_i}$  for  $1 \leq i \leq j^* - 1$  (note that explicit formulas for computing  $j^*$  are available when using uniform or geometric grid scheme).

The purpose of this step is to find the index of the smallest bounding set which includes the new sample  $X_{j+1}$ . The bounding sets with index no less than  $j^*$  can use the new sample.

**Step (2)** Modify  $S^j$  as a temporary matrix  $\widehat{S}^{j+1}$  based on the following three cases.

Case (1).  $[S^j]_{\ell^*,1} < j^* < [S^j]_{\ell^*+1,1}$  for some  $\ell^* \in \{1, \dots, \kappa-1\}$ ;

Case (2).  $j^* = [S^j]_{\ell^*,1}$  for some  $\ell^* \in \{1, \dots, \kappa\}$ ;

Case (3).  $j^* > [S^j]_{\kappa,1}$ .

Case (1) corresponds to the situation wherein the new sample  $X_{j+1}$  falls between two nested bounding sets with radius  $r_p$  and  $r_q$  such that  $p = [S^j]_{\ell^*,1}$ ,  $q = [S^j]_{\ell^*+1,1}$  and wherein bounding set  $\mathcal{B}_{r_i}$  has the same number of collected samples for all  $i$  satisfying  $p \leq i < q$ . Therefore, a new row

needs to be inserted into the matrix of sample sizes. Moreover, the number of collected samples needs to be increased for bounding set  $\mathcal{B}_{r_i}$  with  $i \geq j^*$ . Specifically, this can be accomplished as follows:

In Case (1), define  $\widehat{S}^{j+1}$  as a  $(\kappa + 1) \times 2$  matrix such that

$$\begin{aligned} [\widehat{S}^{j+1}]_{\ell,1} &= [S^j]_{\ell,1}, & [\widehat{S}^{j+1}]_{\ell,2} &= [S^j]_{\ell,2}, & \ell &= 1, \dots, \ell^*, \\ [\widehat{S}^{j+1}]_{\ell^*+1,1} &= j^*, & [\widehat{S}^{j+1}]_{\ell^*+1,2} &= 1 + [S^j]_{\ell^*,2}, \\ [\widehat{S}^{j+1}]_{\ell+1,1} &= [S^j]_{\ell,1}, & [\widehat{S}^{j+1}]_{\ell+1,2} &= 1 + [S^j]_{\ell,2}, & \ell &= \ell^* + 1, \dots, \kappa. \end{aligned}$$

Case (2) corresponds to the situation wherein the new sample  $X_{j+1}$  falls between two nested bounding sets with radius  $r_{p-1}$  and  $r_p$  such that  $p = [S^j]_{\ell^*,1}$ . The new sample can be used by bounding set  $\mathcal{B}_{r_i}$  with  $i \geq p$ . No new row needs to be inserted into the matrix of sample sizes. However, by the principle of sample reuse, the number of collected samples needs to be increased for bounding set  $\mathcal{B}_{r_i}$  with  $i \geq p$ . Specifically, this can be accomplished as follows:

In Case (2), define  $\widehat{S}^{j+1}$  as a  $\kappa \times 2$  matrix such that

$$\begin{aligned} [\widehat{S}^{j+1}]_{\ell,1} &= [S^j]_{\ell,1}, & [\widehat{S}^{j+1}]_{\ell,2} &= [S^j]_{\ell,2}, & \ell &= 1, \dots, \ell^* - 1, \\ [\widehat{S}^{j+1}]_{\ell,1} &= [S^j]_{\ell,1}, & [\widehat{S}^{j+1}]_{\ell,2} &= 1 + [S^j]_{\ell,2}, & \ell &= \ell^*, \dots, \kappa. \end{aligned}$$

Case (3) corresponds to the situation wherein the new sample  $X_{j+1}$  falls between two nested bounding sets with radius  $r_p$  and  $r_m$  such that  $p = [S^j]_{\kappa,1}$ . The new sample can be used by bounding set  $\mathcal{B}_{r_i}$  with  $p \leq i \leq m$ . A new row needs to be added to the matrix of sample sizes. Moreover, by the principle of sample reuse, the number of collected sample sizes needs to be increased for bounding set  $\mathcal{B}_{r_i}$  with  $i \geq p$ . Specifically, this can be accomplished as follows:

In Case (3), define  $\widehat{S}^{j+1}$  as a  $(\kappa + 1) \times 2$  matrix such that

$$\begin{aligned} [\widehat{S}^{j+1}]_{\ell,1} &= [S^j]_{\ell,1}, & [\widehat{S}^{j+1}]_{\ell,2} &= [S^j]_{\ell,2}, & \ell &= 1, \dots, \kappa, \\ [\widehat{S}^{j+1}]_{\kappa+1,1} &= j^*, & [\widehat{S}^{j+1}]_{\kappa+1,2} &= 1 + [S^j]_{\kappa,2}. \end{aligned}$$

**Step (3)** Let  $\widehat{\kappa}$  denote the number of rows of  $\widehat{S}^{j+1}$ . If  $[\widehat{S}^{j+1}]_{\widehat{\kappa},2} < N$ , then let  $S^{j+1} = \widehat{S}^{j+1}$ ; otherwise find index  $\ell_*$  by a bisection search such that  $[\widehat{S}^j]_{\ell_*,1,2} < N \leq [\widehat{S}^j]_{\ell_*,2}$  and define  $S^{j+1}$  as an  $\ell_* \times 2$  matrix such that

$$\begin{aligned} [S^{j+1}]_{\ell,1} &= [\widehat{S}^{j+1}]_{\ell,1}, & [S^{j+1}]_{\ell,2} &= [\widehat{S}^{j+1}]_{\ell,2}, & \ell &= 1, \dots, \ell_* - 1, \\ [S^{j+1}]_{\ell_*,1} &= [\widehat{S}^{j+1}]_{\ell_*,1}, & [S^{j+1}]_{\ell_*,2} &= [\widehat{S}^{j+1}]_{\ell_*,2}. \end{aligned}$$

The purpose of this step is to make sure that no more than  $N$  samples are used for each bounding set. In this step, the records for those bounding sets having enough collected samples have been merged.

**6.2.2. Violations tracking.** In this section, we describe how to update the matrix of violations in the case that the robustness requirement is violated for  $X_{j+1}$ . The key idea is to ensure condition (6.5). Let  $\kappa$  be the number of rows of  $S^j$ . Let  $\tau$  be the number of rows of  $V^j$ . Let  $j^*$  be the index obtained in the process of updating  $S^j$  such that  $X_{j+1} \in \mathcal{B}_{r_i}$  for  $j^* \leq i \leq m$  and  $X_{j+1} \notin \mathcal{B}_{r_i}$  for  $1 \leq i \leq j^* - 1$ . We proceed as follows.

**Step (i)** Identify the *maximal* number  $\iota$  such that the experiment for uncertainty radius  $r_p$  with  $p = [V^j]_{\iota,1}$  has not been completed by the following method.

◇ If  $[S^j]_{\kappa,2} < N$ , then let  $\iota = \kappa$ ; otherwise find  $\iota$  by a bisection search such that  $[V^j]_{\iota,1} < [S^j]_{\kappa,1}$ ,  $[V^j]_{\iota+1,1} \geq [S^j]_{\kappa,1}$ .

Here “the experiment for a bounding set is completed” means that the number of collected samples reaches  $N$ . Obviously, if  $[S^j]_{\kappa,2} < N$ , the experiment has not completed for all bounding sets. If  $[S^j]_{\kappa,2} = N$ , then the experiment is completed for bounding set  $\mathcal{B}_{r_i}$  with  $i \geq [S^j]_{\kappa,1}$ . The maximal number  $\iota$  satisfies  $[V^j]_{\iota,1} < [S^j]_{\kappa,1}$ ,  $[V^j]_{\iota+1,1} \geq [S^j]_{\kappa,1}$ .

**Step (ii)** Modify  $V^j$  as a temporary matrix  $\widehat{V}^j$  based on the following two cases.

Case (a).  $[S^j]_{\kappa,2} < N$  or  $[S^j]_{\kappa,2} = N$ ,  $[V^j]_{\iota+1,1} = [S^j]_{\kappa,1}$ .

Case (b).  $[S^j]_{\kappa,2} = N$  and the index  $\iota$  guarantees  $[V^j]_{\iota+1,1} > [S^j]_{\kappa,1}$ .

In Case (a), we define  $\widehat{V}^j = V^j$ . In Case (b), we define  $\widehat{V}^j$  as a  $(\tau + 1) \times 2$  matrix such that

$$\begin{aligned} [\widehat{V}^j]_{\ell,1} &= [V^j]_{\ell,1}, & [\widehat{V}^j]_{\ell,2} &= [V^j]_{\ell,2}, & \ell &= 1, \dots, \iota, \\ [\widehat{V}^j]_{\iota+1,1} &= [S^j]_{\kappa,1}, & [\widehat{V}^j]_{\iota+1,2} &= [V^j]_{\iota,2}, \\ [\widehat{V}^j]_{\ell+1,1} &= [V^j]_{\ell,1}, & [\widehat{V}^j]_{\ell+1,2} &= [V^j]_{\ell,2}, & \ell &= \iota + 1, \dots, \tau. \end{aligned}$$

This step prepares us for updating the matrix of violations in Step (iii). It is ensured that, in both Cases (a) and (b), the rows of matrix  $\widehat{V}^j$  of index greater than  $\iota$  correspond to the completed experiments. Hence, as will be seen in Step (iii), such rows of matrix  $\widehat{V}^j$  need not to be updated.

**Step (iii)** Obtain  $V^{j+1}$  by modifying  $\widehat{V}^j$  based on the following three cases.

Case (i).  $[\widehat{V}^j]_{\ell^*,1} < j^* < [\widehat{V}^j]_{\ell^*+1,1}$  for some  $\ell^* \in \{1, \dots, \iota - 1\}$ .

Case (ii).  $j^* = [\widehat{V}^j]_{\ell^*,1}$  for some  $\ell^* \in \{1, \dots, \iota\}$ .

Case (iii).  $j^* > [\widehat{V}^j]_{\iota,1}$ .

Case (i) corresponds to the situation wherein the new sample  $X_{j+1}$  falls between two nested bounding sets with radius  $r_p$  and  $r_q$  such that  $p = [\widehat{V}^j]_{\ell^*,1}$ ,  $q = [\widehat{V}^j]_{\ell^*+1,1}$  and wherein bounding set  $\mathcal{B}_{r_i}$  has the same number of collected violations for all  $i$  satisfying  $p \leq i < q$ . Therefore, a new row needs to be inserted into the matrix of violations. Moreover, the number of collected violations needs to be increased for bounding set  $\mathcal{B}_{r_i}$  with  $i \geq j^*$ . Recall that the rows of  $\widehat{V}^j$  of index greater than  $\iota$  correspond to the complete experiments; those rows should be passed without change to  $V^{j+1}$ . Specifically, this can be accomplished as follows:

Let  $\widehat{\tau}$  be the number of rows of  $\widehat{V}^j$ . In Case (i), define  $V^{j+1}$  as a  $(\widehat{\tau} + 1) \times 2$  matrix such that

$$\begin{aligned} [V^{j+1}]_{\ell,1} &= [\widehat{V}^j]_{\ell,1}, & [V^{j+1}]_{\ell,2} &= [\widehat{V}^j]_{\ell,2}, & \ell &= 1, \dots, \ell^*, \\ [V^{j+1}]_{\ell^*+1,1} &= j^*, & [V^{j+1}]_{\ell^*+1,2} &= 1 + [\widehat{V}^j]_{\ell^*,2}, \\ [V^{j+1}]_{\ell+1,1} &= [\widehat{V}^j]_{\ell,1}, & [V^{j+1}]_{\ell+1,2} &= 1 + [\widehat{V}^j]_{\ell,2}, & \ell &= \ell^* + 1, \dots, \iota, \\ [V^{j+1}]_{\ell+1,1} &= [\widehat{V}^j]_{\ell,1}, & [V^{j+1}]_{\ell+1,2} &= [\widehat{V}^j]_{\ell,2}, & \ell &= \iota + 1, \dots, \widehat{\tau}. \end{aligned}$$

Case (ii) corresponds to the situation wherein the new sample  $X_{j+1}$  falls between two nested bounding sets with radius  $r_{p-1}$  and  $r_p$  such that  $p = [\widehat{V}^j]_{\ell^*,1}$ . The new sample can be used by bounding set  $\mathcal{B}_{r_i}$  with  $i \geq p$ . No new row needs to be inserted into the matrix of violations. However, by the principle of sample reuse, the number of collected violations needs



to be increased for bounding set  $\mathcal{B}_{r_i}$  with  $i \geq p$ . The rows of  $\widehat{V}^j$  of index greater than  $\iota$  should be directly passed to  $V^{j+1}$ . Specifically, this can be accomplished as follows:

In Case (ii), define  $V^{j+1}$  as a  $\widehat{\tau} \times 2$  matrix such that

$$\begin{aligned} [V^{j+1}]_{\ell,1} &= [\widehat{V}^j]_{\ell,1}, & [V^{j+1}]_{\ell,2} &= [\widehat{V}^j]_{\ell,2}, & \ell &= 1, \dots, \ell^* - 1, \\ [V^{j+1}]_{\ell,1} &= [\widehat{V}^j]_{\ell,1}, & [V^{j+1}]_{\ell,2} &= 1 + [\widehat{V}^j]_{\ell,2}, & \ell &= \ell^*, \dots, \iota, \\ [V^{j+1}]_{\ell,1} &= [\widehat{V}^j]_{\ell,1}, & [V^{j+1}]_{\ell,2} &= [\widehat{V}^j]_{\ell,2}, & \ell &= \iota + 1, \dots, \widehat{\tau}. \end{aligned}$$

Case (iii) corresponds to the situation wherein the new sample  $X_{j+1}$  falls between two nested bounding sets with radius  $r_p$  and  $r_m$  such that  $p = [\widehat{V}^j]_{\iota,1}$ . The new sample can be used by bounding set  $\mathcal{B}_{r_i}$  with  $j^* \leq i \leq m$  if the experiment has not been completed. A new row needs to be added to the matrix of violations. Moreover, by the principle of sample reuse, the number of collected violations needs to be increased for bounding set  $\mathcal{B}_{r_i}$  with  $i \geq j^*$  if the experiment has not been completed. Specifically, this can be accomplished as follows:

In Case (iii), define  $V^{j+1}$  as a  $(\widehat{\tau} + 1) \times 2$  matrix such that

$$\begin{aligned} [V^{j+1}]_{\ell,1} &= [\widehat{V}^j]_{\ell,1}, & [V^{j+1}]_{\ell,2} &= [\widehat{V}^j]_{\ell,2}, & \ell &= 1, \dots, \iota, \\ [V^{j+1}]_{\iota+1,1} &= j^*, & [V^{j+1}]_{\iota+1,2} &= 1 + [\widehat{V}^j]_{\iota,2}, \\ [V^{j+1}]_{\ell+1,1} &= [\widehat{V}^j]_{\ell,1}, & [V^{j+1}]_{\ell+1,2} &= [\widehat{V}^j]_{\ell,2}, & \ell &= \iota + 1, \dots, \widehat{\tau}. \end{aligned}$$

**6.3. Complexity of data processing and memory.** It can be seen that the memory requirement and the computation due to data processing are determined by the sizes of matrices  $S^j$  and  $V^j$ . To quantify the complexity, we have the following results.

THEOREM 6.1.

For any  $j$ , the following statements hold true:

- (I) The number of rows of matrix  $S^j$  is no more than  $N$ .
- (II) The expected number of rows of matrix  $V^j$  is no greater than

$$(6.6) \quad 1 + N \left[ P_e(a) + 2d \int_{\frac{a}{\hbar}}^a \frac{P_e(x)}{x} dx \right] \leq 1 + N P_e(a) (1 + 2d \ln \hbar)$$

where  $P_e(x) = 1 - \min_{y \in [\frac{a}{\lambda}, x]} \mathbb{P}(y)$  for all  $x \in [\frac{a}{\lambda}, a]$  and

$$\hbar = \max \left( \min \left( \lambda, \frac{a}{\rho_0} \right), 1 \right)$$

with  $\rho_0 = \sup\{r \mid \mathbb{P}(r) = 1\}$ .

See Appendix E for a proof. We now revisit the robustness analysis problem discussed in Remark 1 of section 4 from the perspective of memory complexity. Assume that each data record  $(k_i, n_i)$  (i.e., each row of  $V^j$ ) occupies 4 bytes of computer internal memory (RAM). As illustrated in section 5.2, when using the conventional data structure, it takes 13G (gigabytes) of RAM to save the data and visualize the results. On the other hand, in our new algorithm, if the smallest proportion is

$p_* = \min_{r \in [\frac{a}{\lambda}, a]} \mathbb{P}(r) > 0.999$  and  $\hbar < \frac{3}{2}$ , the RAM requirement will be equivalent to

$$\begin{aligned} & 4 \times [1 + (1 - p_*) (1 + 2d \ln \hbar) N] \\ &= 4 \times \left[ 1 + (1 - 0.999) \times \left( 1 + 2 \times 1800 \times \ln \frac{3}{2} \right) \times 10^6 \right] \\ &< 6.2 \times 10^6 \text{ bytes} \approx 6.2 \text{ M bytes.} \end{aligned}$$

It can be seen that such a memory requirement is extremely low as compared to that of the conventional method. Theorem 6.1 also reveals that the complexity of data processing is very low.

**6.4. Confidence band.** To be useful, any numerical techniques should provide a method for error assessment. Monte Carlo simulation is no exception. The following results allow us to construct a confidence band for the robustness curve. Such a post-experimental statistical inference can remedy the conservatism of an a priori choice of sample size  $N$  based on the Chernoff bound. In order to overcome the computational complexity of the Clopper–Pearson confidence interval [10], we have developed new methods to facilitate the construction of the confidence band.

**THEOREM 6.2.** *Let  $\delta \in (0, 1)$ . Let  $\mathcal{L}(k) = \frac{k}{N} + \frac{3}{4} \frac{1 - \frac{2k}{N} - \sqrt{1 + 4\theta k(1 - \frac{k}{N})}}{1 + \theta N}$  and  $\mathcal{U}(k) = \frac{k}{N} + \frac{3}{4} \frac{1 - \frac{2k}{N} + \sqrt{1 + 4\theta k(1 - \frac{k}{N})}}{1 + \theta N}$  with  $\theta = \frac{9}{8 \ln \frac{2}{\delta}}$ . Let  $\zeta = \frac{r - r_i}{r_{i+1} - r_i}$ . Let*

$$K_i = N - \mathbf{v}(i, \mathbf{n}), \quad i = 1, \dots, m.$$

*Let  $\varsigma = 1 - \frac{g(r_*)}{r_{i+1} - r_i}$ . Define  $\bar{\mathbb{P}}(r) = \zeta \mathcal{U}(K_i) + (1 - \zeta) \mathcal{U}(K_{i+1}) + \varsigma$  and  $\underline{\mathbb{P}}(r) = \zeta \mathcal{L}(K_i) + (1 - \zeta) \mathcal{L}(K_{i+1}) - \varsigma$ . Then*

$$\Pr\{\underline{\mathbb{P}}(r) < \mathbb{P}(r) < \bar{\mathbb{P}}(r) \ \forall r \in [r_i, r_{i+1}]\} > 1 - \delta.$$

See Appendix F for a proof. The family of intervals  $[\underline{\mathbb{P}}(r), \bar{\mathbb{P}}(r)]$ ,  $r \in [a/\lambda, a]$ , is referred to as the *confidence band*. It is important to note that the confidence band can be efficiently constructed by making use of the piecewise constant property of  $\mathbf{v}(i, \mathbf{n})$ . It can be shown that the computational complexity of constructing the confidence band is also absolutely bounded.

**7. Examples.** In this section, we shall illustrate by examples the application of our techniques in control of uncertain systems. For a robustness analysis problem, one can define an indicator function  $\mathbb{I}(\cdot)$  such that

$$\mathbb{I}(\Delta) = \begin{cases} 1 & \text{if the robustness requirement is satisfied for } \Delta, \\ 0 & \text{otherwise} \end{cases}$$

for any uncertainty instance  $\Delta$ . Let  $\Delta_1, \dots, \Delta_N$  be  $N$  i.i.d. samples uniformly distributed over an uncertainty bounding set  $\mathcal{B}_r$ . Then,  $\mathbb{I}(\Delta_i)$ ,  $i = 1, \dots, N$ , are i.i.d. Bernoulli random variables with a success probability  $\mathbb{P}(r)$ . A minimum variance unbiased estimator of  $\mathbb{P}(r)$  is taken as

$$\hat{\mathbb{P}}(r) = \frac{\sum_{i=1}^N \mathbb{I}(\Delta_i)}{N}.$$

The Chernoff bound [9] asserts that, for any  $\varepsilon$ ,  $\delta \in (0, 1)$ ,

$$\Pr\left\{\left|\hat{\mathbb{P}}(r) - \mathbb{P}(r)\right| < \varepsilon\right\} > 1 - \delta$$

provided that

$$(7.1) \quad N > \frac{\ln \frac{2}{\delta}}{2\epsilon^2}.$$

Since the explicit bound (7.1) is independent of the quantity  $\mathbb{P}(r)$  that we want to estimate, we can use it as the guideline for the choice of sample size  $N$ .

Now we first consider the robust stability problem studied in [12] by a deterministic approach. The system considered in [12] is represented by Figure 6. The compensator is  $C(s) = \frac{s+2}{s+10}$  and the plant is  $P(s) = \frac{800(1+0.1\delta_1)}{s(s+4+0.2\delta_2)(s+6+0.3\delta_3)}$  with parametric uncertainty  $\Delta = [\delta_1, \delta_2, \delta_3]$ . For  $\epsilon = \delta = 0.01$ , we determined the sample size  $N$  from the Chernoff bound (7.1) as  $N = 26492$ . In order to make the discretization error  $\epsilon$  be less than  $10^{-4}$  with a geometrical gridding over uncertainty radius interval [3, 8], we obtained the number of grid point as  $m = 14714$  by formula (4.5). In this problem case, the evaluation of system performance for sample  $\Delta_i$  is actually the evaluation of the indicator function

$$\mathbb{I}(\Delta_i) = \begin{cases} 1 & \text{if the closed-loop system is stable for } \Delta_i, \\ 0 & \text{otherwise.} \end{cases}$$

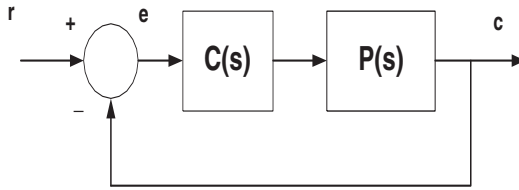


FIG. 6. *Uncertain system.*

The deterministic robustness margin is found to be 3.44 by a branch and bound technique (see page 163 of [12]). As shown in Figure 7, we have applied our new sample reuse algorithms to obtain an estimation of the function  $\mathbb{P}(r)$  and corresponding confidence band with confidence coefficient 99% (based on Theorem 6) for assessing the accuracy of the estimation. Owing to the power of the new method of sample reuse, the ENGP is less than 4 and the memory requirement is negligible. In addition to the low computational complexity, the robustness curve obtained by the new algorithms provides more insight into the system robustness than the deterministic robustness margin.

Next, we consider a robust performance problem involving time-domain specifications for the same system shown by Figure 6. The robustness requirement is that the rise time and settling time should be no more than 0.25 and 3.5 seconds, respectively, and the overshoot should be no more than 70% under the condition that the closed-loop system is stable. For this robust performance problem, the corresponding indicator function becomes

$$\mathbb{I}(\Delta_i) = \begin{cases} 1 & \text{if the closed-loop system is stable} \\ & \text{and the time specification is satisfied for } \Delta_i, \\ 0 & \text{otherwise.} \end{cases}$$

The estimation of  $\mathbb{P}(r)$  together with its corresponding confidence band are shown in Figure 8, where the sample size  $N$  is taken as 26,492 and the number of grid

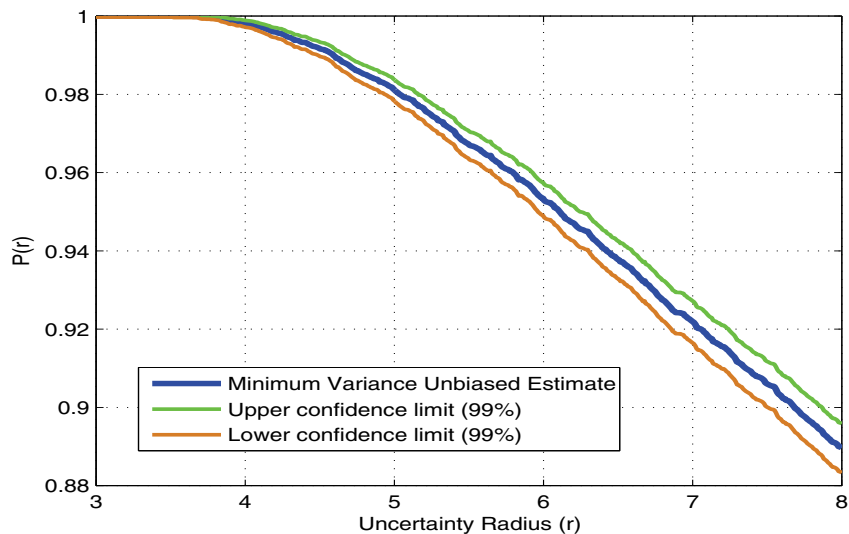


FIG. 7. Robustness analysis with stability requirement.

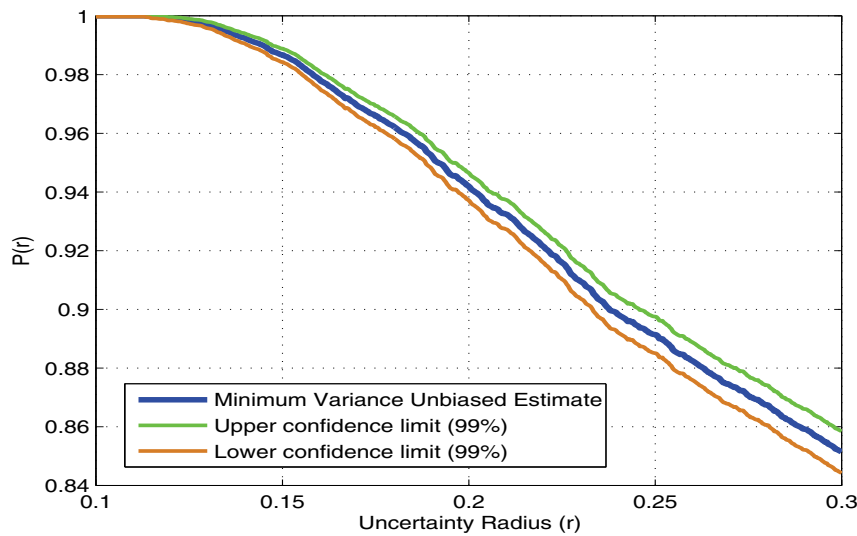


FIG. 8. Robustness analysis with time specifications.

points is taken by (4.5) as 16,481 to ensure the discretization error less than  $10^{-4}$  for a geometric gridding over uncertainty radius interval  $[0.1, 0.3]$ . It is well known that this type of problem is, in general, intractable by the deterministic approach. However, the new sample reuse algorithms can provide an efficient and insightful solution.

**8. Conclusion.** It is possible to make a case for the statement that the probabilistic robustness analysis is essentially the study of the robustness function,

especially its probabilistic implications, efficient evaluation, and computational complexity. We have addressed these issues in this paper. In particular, we have developed randomized algorithms which offer more insights for system robustness. We rigorously show that, in both aspects of computer running time and memory requirement, the complexity of such randomized algorithms is not only linear in the dimension of uncertainty space, but also surprisingly low. While the complexity of the conventional method grows linearly with the number of grid points, and the error due to interpolation is not well controlled, our techniques completely resolve such issues. In short, our method guarantees accuracy and efficiency.

**Appendix A. Proof of Theorem 3.1.** We first establish a basic inequality that will be used to prove the theorem.

LEMMA A.1. *For any  $x > 1$ ,*

$$\frac{1}{x} + \ln x > 1.$$

*Proof.* Let

$$f(x) = \frac{1}{x} + \ln x.$$

Then  $f(1) = 1$  and

$$\frac{d f(x)}{d x} = \frac{x-1}{x^2} > 0 \quad \forall x > 1.$$

It follows that  $f(x) > 1$  for all  $x > 1$ .  $\square$

Now we are in the position to prove Theorem 3.1. Observing that

$$\left(\frac{r_m}{r_1}\right)^d = \prod_{i=1}^{m-1} \left(\frac{r_{i+1}}{r_i}\right)^d,$$

we have

$$\ln \left(\frac{r_m}{r_1}\right)^d = \sum_{i=1}^{m-1} \ln \left(\frac{r_{i+1}}{r_i}\right)^d.$$

Therefore,

$$\sum_{i=1}^{m-1} \left(\frac{r_i}{r_{i+1}}\right)^d + \ln \left(\frac{r_m}{r_1}\right)^d = \sum_{i=1}^{m-1} \left[ \frac{1}{\left(\frac{r_{i+1}}{r_i}\right)^d} + \ln \left(\frac{r_{i+1}}{r_i}\right)^d \right].$$

Since  $\left(\frac{r_{i+1}}{r_i}\right)^d > 1$ ,  $i = 1, \dots, m-1$ , it follows from Lemma A.1 that

$$\frac{1}{\left(\frac{r_{i+1}}{r_i}\right)^d} + \ln \left(\frac{r_{i+1}}{r_i}\right)^d > 1, \quad i = 1, \dots, m-1.$$

Hence,

$$\sum_{i=1}^{m-1} \left(\frac{r_i}{r_{i+1}}\right)^d + \ln \left(\frac{r_m}{r_1}\right)^d > m-1,$$

or equivalently,

$$m - \sum_{i=1}^{m-1} \left( \frac{r_i}{r_{i+1}} \right)^d < 1 + \ln \left( \frac{r_m}{r_1} \right)^d = 1 + d \ln \lambda.$$

Finally, by Theorem 1 of [6] and the definition of  $m_{\text{eq}}$ , we have

$$m_{\text{eq}} = m - \sum_{i=1}^{m-1} \left( \frac{r_i}{r_{i+1}} \right)^d < 1 + d \ln \lambda.$$

**Appendix B. Proof of Theorem 4.1.** To prove the theorem, we need some preliminary results. It is derived in [2] that  $\left| \frac{d\mathbb{P}(r)}{dr} \right| < \frac{2d}{r}$  when  $\mathbb{P}(\cdot)$  is differentiable. The following lemma indicates that the bound on the rate of variation of  $\mathbb{P}(\cdot)$  can be much tighter.

LEMMA B.1. *For the arbitrary robustness requirement,*

$$|\mathbb{P}(r + \Delta r) - \mathbb{P}(r)| \leq 1 - \left( 1 + \frac{\Delta r}{r} \right)^{-d} < \frac{d}{r} \Delta r$$

for any  $r > 0$  and any  $\Delta r > 0$ .

*Proof.* Let  $\mathcal{Q}_r \subseteq \mathcal{B}_r$  be the set such that the robustness requirement is satisfied. Let

$$I_1 = \frac{\text{vol}(\mathcal{Q}_{r+\Delta r})}{\text{vol}(\mathcal{B}_{r+\Delta r})} - \frac{\text{vol}(\mathcal{Q}_r)}{\text{vol}(\mathcal{B}_{r+\Delta r})}, \quad I_2 = \frac{\text{vol}(\mathcal{Q}_r)}{\text{vol}(\mathcal{B}_{r+\Delta r})} - \frac{\text{vol}(\mathcal{Q}_r)}{\text{vol}(\mathcal{B}_r)}.$$

Let “ $\setminus$ ” denote the operation of set minus. Observing that  $\mathcal{Q}_{r+\Delta r} \setminus \mathcal{Q}_r \subseteq \mathcal{B}_{r+\Delta r} \setminus \mathcal{B}_r$ , we have  $\text{vol}(\mathcal{Q}_{r+\Delta r}) - \text{vol}(\mathcal{Q}_r) \leq \text{vol}(\mathcal{B}_{r+\Delta r}) - \text{vol}(\mathcal{B}_r)$ . Using this fact and the identity  $\text{vol}(\mathcal{B}_r) = r^d \text{vol}(\mathcal{B}_1)$ , we have

$$0 \leq I_1 \leq \frac{\text{vol}(\mathcal{B}_{r+\Delta r}) - \text{vol}(\mathcal{B}_r)}{\text{vol}(\mathcal{B}_{r+\Delta r})} = 1 - \left( 1 + \frac{\Delta r}{r} \right)^{-d}$$

and

$$- \left[ 1 - \left( 1 + \frac{\Delta r}{r} \right)^{-d} \right] \leq - \frac{\text{vol}(\mathcal{Q}_r)}{\text{vol}(\mathcal{B}_r)} \frac{\text{vol}(\mathcal{B}_{r+\Delta r}) - \text{vol}(\mathcal{B}_r)}{\text{vol}(\mathcal{B}_{r+\Delta r})} = I_2.$$

Therefore,  $|\mathbb{P}(r + \Delta r) - \mathbb{P}(r)| = |I_1 + I_2| \leq 1 - \left( 1 + \frac{\Delta r}{r} \right)^{-d} < \frac{d}{r} \Delta r$ , where the last inequality follows from inequality  $1 - \left( 1 + \frac{x}{r} \right)^{-d} < \frac{dx}{r}$  for all  $x > 0$ . To prove this inequality, we can define function  $h(x) = 1 - \left( 1 + \frac{x}{r} \right)^{-d} - \frac{dx}{r}$ ,  $x > 0$ , and check that  $h(0) = 0$  and  $\frac{\partial h(x)}{\partial x} = \frac{d}{r} \left[ \left( 1 + \frac{x}{r} \right)^{-(d+1)} - 1 \right] < 0$  for all  $x > 0$ .  $\square$

We are now in the position to prove the theorem. It can be shown that

$$\begin{aligned} |\mathbb{P}(r) - \mathbb{P}^*(r)| &= \left| \frac{(r_{i+1} - r)[\mathbb{P}(r) - \mathbb{P}(r_i)] + (r - r_i)[\mathbb{P}(r) - \mathbb{P}(r_{i+1})]}{r_{i+1} - r_i} \right| \\ (B.1) \quad &\leq \frac{(r_{i+1} - r)|\mathbb{P}(r) - \mathbb{P}(r_i)| + (r - r_i)|\mathbb{P}(r) - \mathbb{P}(r_{i+1})|}{r_{i+1} - r_i}. \end{aligned}$$

By Lemma B.1 and inequality (B.1), we have

$$\begin{aligned} |\mathbb{P}(r) - \mathbb{P}^*(r)| &\leq \frac{(r_{i+1} - r) \left[ 1 - \left( \frac{r}{r_i} \right)^{-d} \right] + (r - r_i) \left[ 1 - \left( \frac{r_{i+1}}{r} \right)^{-d} \right]}{r_{i+1} - r_i} \\ &= 1 - \frac{g(r)}{r_{i+1} - r_i}. \end{aligned}$$

Note that

$$\frac{\partial g(r)}{\partial r} = \Phi(r) - \Psi(r),$$

where

$$\Phi(r) = \left( \frac{r_{i+1}}{r} \right)^{-d} \left[ 1 + \left( 1 - \frac{r_i}{r} \right) d \right], \quad \Psi(r) = \left( \frac{r}{r_i} \right)^{-d} \left[ 1 + \left( \frac{r_{i+1}}{r} - 1 \right) d \right].$$

It can be verified that

$$\Phi(r_i) = \left( \frac{r_{i+1}}{r_i} \right)^{-d} < 1, \quad \Psi(r_i) = 1 + \left( \frac{r_{i+1}}{r_i} - 1 \right) d > 1, \quad \frac{\partial g(r)}{\partial r} \Big|_{r=r_i} < 0,$$

$$\Phi(r_{i+1}) = 1 + \left( 1 - \frac{r_i}{r_{i+1}} \right) d > 1, \quad \Psi(r_{i+1}) = \left( \frac{r_{i+1}}{r_i} \right)^{-d} < 1, \quad \frac{\partial g(r)}{\partial r} \Big|_{r=r_{i+1}} > 0.$$

It can be checked that  $\Phi(r)$  is a monotone increasing function of  $r$  and that  $\Psi(r)$  is a monotone decreasing function of  $r$ . Hence,  $\frac{\partial g(r)}{\partial r}$  is a monotone increasing function of  $r$ . Moreover, there exists a unique number  $r_\star \in (r_i, r_{i+1})$  such that  $\frac{\partial g(r)}{\partial r} \Big|_{r=r_\star} = 0$ , i.e.,  $\Phi(r_\star) = \Psi(r_\star)$ . Furthermore,  $g(r)$  is a convex function of  $r$ . Consequently,

$$\min_{r \in [r_i, r_{i+1}]} g(r) = g(r_\star),$$

and we have shown

$$|\mathbb{P}(r) - \mathbb{P}^*(r)| \leq 1 - \frac{g(r_\star)}{r_{i+1} - r_i} \quad \forall r \in [r_i, r_{i+1}].$$

Since  $\frac{\partial g(r)}{\partial r}$  is a monotone increasing function of  $r$ , we can compute  $r_\star$  by a bisection search over interval  $(r_i, r_{i+1})$ .

By Lemma B.1 and inequality (B.1), we have

$$\begin{aligned} |\mathbb{P}(r) - \mathbb{P}^*(r)| &\leq \frac{(r_{i+1} - r)(r - r_i) \frac{d}{r_i} + (r - r_i)(r_{i+1} - r) \frac{d}{r}}{r_{i+1} - r_i} \\ &\leq \frac{(r_{i+1} - r)(r - r_i) \frac{d}{r_i} + (r - r_i)(r_{i+1} - r) \frac{d}{r_i}}{r_{i+1} - r_i} \\ &\leq \frac{2d}{r_i(r_{i+1} - r_i)} \max_{r \in [r_i, r_{i+1}]} (r_{i+1} - r)(r - r_i) \\ &= \frac{2d}{r_i(r_{i+1} - r_i)} \frac{(r_{i+1} - r_i)^2}{4} \\ &= \frac{d(r_{i+1} - r_i)}{2r_i}. \end{aligned}$$

**Appendix C. Proof of Theorem 4.2.** By Theorem 4.1,  $|\mathbb{P}(r) - \mathbb{P}^*(r)| \leq \frac{d(r_{i+1}-r_i)}{2r_i}$  for all  $r \in [r_i, r_{i+1}]$ . Thus, it suffices to show  $\frac{d(r_{i+1}-r_i)}{2r_i} < \epsilon$ , i.e.,

$$(C.1) \quad \frac{r_{i+1}}{r_i} < 1 + \frac{2\epsilon}{d}.$$

By definition (4.1), for  $i = 1, \dots, m-1$ ,

$$\begin{aligned} \frac{r_{i+1}}{r_i} &= \frac{a - \frac{(m-i-1)(\lambda-1)}{(m-1)\lambda}a}{a - \frac{(m-i)(\lambda-1)}{(m-1)\lambda}a} \\ &= 1 + \frac{\lambda-1}{m-1 + (\lambda-1)(i-1)} \\ &\leq 1 + \frac{\lambda-1}{m-1}. \end{aligned}$$

By virtue of (C.1), to guarantee that the gridding error is less than  $\epsilon$ , it suffices to ensure  $1 + \frac{\lambda-1}{m-1} < 1 + \frac{2\epsilon}{d}$ , i.e.,  $m > 1 + \frac{d(\lambda-1)}{2\epsilon}$ . Hence, it suffices to have

$$m \geq 2 + \left\lfloor \frac{(\lambda-1)d}{2\epsilon} \right\rfloor.$$

It can be verified that

$$\frac{r_i}{r_{i+1}} = 1 - \frac{1}{\frac{m-1}{\lambda-1} + i}, \quad i = 1, \dots, m-1.$$

By Theorem 1 of [6], the sample reuse factor is given by

$$\begin{aligned} \mathcal{F}_{\text{reuse}} &= \frac{m}{m - \sum_{i=1}^{m-1} \left( \frac{r_i}{r_{i+1}} \right)^d} \\ &= \frac{m}{m - \sum_{i=1}^{m-1} \left( 1 - \frac{1}{\frac{m-1}{\lambda-1} + i} \right)^d}. \end{aligned}$$

Therefore,

$$m_{\text{eq}}(\epsilon) = \frac{m}{\mathcal{F}_{\text{reuse}}} = m - \sum_{i=1}^{m-1} \left( 1 - \frac{1}{\frac{m-1}{\lambda-1} + i} \right)^d.$$

**Appendix D. Proof of Theorem 4.3.** By virtue of (4.2), we have  $\frac{r_{i+1}}{r_i} = \lambda^{\frac{1}{m-1}}$ . Hence, by (C.1), it suffices to show  $\lambda^{\frac{1}{m-1}} < 1 + \frac{2\epsilon}{d}$ , which can be reduced to  $m > 1 + \frac{\ln \lambda}{\ln(1 + \frac{2\epsilon}{d})}$ . This inequality is equivalent to  $m \geq 2 + \left\lfloor \frac{\ln \lambda}{\ln(1 + \frac{2\epsilon}{d})} \right\rfloor$ . By equation (3.1) and Theorem 1 of [6], we have  $\mathbb{E}[\mathbf{n}] = \left[ m - (m-1) \left( \frac{1}{\lambda} \right)^{\frac{d}{m-1}} \right] N$  and hence obtain  $m_{\text{eq}}(\epsilon)$ . Note that

$$\mathcal{F}_{\text{reuse}} = \frac{m}{m_{\text{eq}}(\epsilon)} > \frac{m}{1 + d \ln \lambda} = \frac{2 + \left\lfloor \frac{\ln \lambda}{\ln(1 + \frac{2\epsilon}{d})} \right\rfloor}{1 + d \ln \lambda} > \frac{\frac{\ln \lambda}{\ln(1 + \frac{2\epsilon}{d})}}{1 + d \ln \lambda}.$$



Making use of the inequality  $\ln(1+x) < x$  for all  $x > 0$ , we have  $\ln\left(1 + \frac{2\epsilon}{d}\right) < \frac{2\epsilon}{d}$ . Therefore,

$$\begin{aligned}\mathcal{F}_{\text{reuse}} &> \frac{\frac{\ln \lambda}{\frac{2\epsilon}{d}}}{1 + d \ln \lambda} \\ &= \frac{1}{2\epsilon} \left(1 - \frac{1}{1 + d \ln \lambda}\right).\end{aligned}$$

### Appendix E. Proof of Theorem 6.1.

**Proof of statement (I).** Obviously,  $[S^j]_{1,2} \geq 1$ ,  $[S^j]_{\kappa,2} \leq N$ . From the rules of sampling, we can perform induction with respect to  $j$  and have  $[S^j]_{\ell+1,2} - [S^j]_{\ell,2} \geq 1$ ,  $\ell = 1, \dots, \kappa - 1$ . Observing that

$$\begin{aligned}[S^j]_{\kappa,2} &= [S^j]_{1,2} + \sum_{\ell=1}^{\kappa-1} ([S^j]_{\ell+1,2} - [S^j]_{\ell,2}) \\ &\geq 1 + \sum_{\ell=1}^{\kappa-1} ([S^j]_{\ell+1,2} - [S^j]_{\ell,2}) \\ &\geq 1 + \kappa - 1 \\ &= \kappa,\end{aligned}$$

we have  $\kappa \leq [S^j]_{\kappa,2} \leq N$ .

**Proof of statement (II).** We need some preliminary results.

LEMMA E.1. *Let  $1 \leq i \leq m - 1$ . Then*

$$\left| \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] - \frac{d(r_{i+1} - r_i)}{r_{i+1}} \right| \leq \frac{d(d-1)}{2} \left( \frac{r_{i+1} - r_i}{r_{i+1}} \right)^2.$$

*Proof.* Note that

$$\begin{aligned}\left| \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] - \frac{d(r_{i+1} - r_i)}{r_{i+1}} \right| &= \left| \left( \frac{r_i}{r_{i+1}} \right)^d - \left( 1 - \frac{d(r_{i+1} - r_i)}{r_{i+1}} \right) \right| \\ &= \left| (1-t)^d - (1-dt) \right|,\end{aligned}$$

where  $t = \frac{r_{i+1} - r_i}{r_{i+1}}$ . It can be checked that  $\left| (1-t)^d - (1-dt) \right| = \frac{d(d-1)}{2} t^2$  for  $d = 1, 2$ . For  $d \geq 3$ , by Taylor's expansion formula, there exists  $\xi \in (0, t)$  such that

$$(1-t)^d = 1 - dt + \frac{d(d-1)}{2} (1-\xi)^{d-2} t^2.$$

Observing that  $0 < t < 1$  since  $0 < r_i < r_{i+1}$ , we hence have  $0 < (1-\xi)^{d-2} < 1$  and  $\left| (1-t)^d - (1-dt) \right| < \frac{d(d-1)}{2} t^2$  for  $d \geq 3$ . Therefore, for any  $d \geq 1$ ,

$$\begin{aligned}\left| \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] - \frac{d(r_{i+1} - r_i)}{r_{i+1}} \right| \\ \leq \frac{d(d-1)}{2} t^2 = \frac{d(d-1)}{2} \left( \frac{r_{i+1} - r_i}{r_{i+1}} \right)^2. \quad \square\end{aligned}$$

LEMMA E.2. Define the maximum gap between grid points as

$$\varpi = \max_{1 \leq i \leq m-1} (r_{i+1} - r_i).$$

Then,

$$\sum_{i=1}^{m-1} \left( \frac{r_{i+1} - r_i}{r_{i+1}} \right)^2 < \frac{\lambda(\lambda-1)\varpi}{a}.$$

*Proof.* Note that

$$\sum_{i=1}^{m-1} \left( \frac{r_{i+1} - r_i}{r_{i+1}} \right)^2 \leq \varpi \sum_{i=1}^{m-1} \frac{r_{i+1} - r_i}{r_{i+1}^2} < \varpi \sum_{i=1}^{m-1} \frac{r_{i+1} - r_i}{\left(\frac{a}{\lambda}\right)^2}.$$

By successive cancellation,

$$\sum_{i=1}^{m-1} (r_{i+1} - r_i) = r_m - r_1 = a - \frac{a}{\lambda}.$$

Hence,

$$\sum_{i=1}^{m-1} \left( \frac{r_{i+1} - r_i}{r_{i+1}} \right)^2 < \varpi \sum_{i=1}^{m-1} \frac{r_{i+1} - r_i}{\left(\frac{a}{\lambda}\right)^2} = \varpi \frac{a - \frac{a}{\lambda}}{\left(\frac{a}{\lambda}\right)^2} = \frac{\lambda(\lambda-1)\varpi}{a}. \quad \square$$

LEMMA E.3. The expected number of rows of the matrix of violations  $V^{\mathbf{n}}$  is no greater than  $1 + NP_e(a) + 2N \sum_{j=1}^{m-1} P_e(r_j) \left[ 1 - \left(\frac{r_j}{r_{j+1}}\right)^d \right]$ .

*Proof.* Let  $X_1^j, \dots, X_{\mathbf{n}_j}^j$  be the samples generated from the uncertainty set with radius  $r_j$ . For  $i = 1, \dots, m$ , define the random variable  $Y_i^j$  such that  $Y_i^j = X_i^j$  if the robustness requirement is violated for  $X_i^j$  and  $Y_i^j = 0$  otherwise. By the principle of sample reuse, the value of  $\mathbf{n}_j$  depends only on the samples generated from uncertainty sets with radius  $r_k$ ,  $j+1 \leq k \leq m$ . Consequently, event  $\{\mathbf{n}_j = \nu\}$  is independent of event  $\{Y_i^j = 1\}$  and  $\Pr\{Y_i^j = 1, \mathbf{n}_j = \nu\} = \Pr\{Y_i^j = 1\} \Pr\{\mathbf{n}_j = \nu\}$ . By the definitions of  $Y_i^j$  and  $P_e(\cdot)$ , we have  $\Pr\{Y_i^j = 1\} = 1 - \mathbb{P}(r_j) \leq P_e(r_j)$ . Therefore,

$$\begin{aligned} \mathbb{E} \left[ \sum_{i=1}^{\mathbf{n}_j} Y_i^j \right] &= \sum_{\nu=1}^N \sum_{i=1}^{\nu} \Pr\{Y_i^j = 1, \mathbf{n}_j = \nu\} \\ &= \sum_{\nu=1}^N \sum_{i=1}^{\nu} \Pr\{Y_i^j = 1\} \Pr\{\mathbf{n}_j = \nu\} \\ &\leq \sum_{\nu=1}^N \nu P_e(r_j) \Pr\{\mathbf{n}_j = \nu\} \\ &= P_e(r_j) \mathbb{E}[\mathbf{n}_j] \end{aligned}$$

for  $j = 1, \dots, m$ . We now consider  $V^{\mathbf{n}}$  with  $\mathbf{n} = \sum_{i=1}^m \mathbf{n}_i$ . By the mechanism of the sample reuse algorithms, for  $j = 1, \dots, m-1$ , every new sample from the uncertainty set with radius  $r_j$  at most creates  $2Y_i^j$ ,  $i = 1, \dots, \mathbf{n}_j$ , new rows for the matrix of violations (see section 6.2.2). Note that  $X_1$  creates at most  $1 + Y_1^m$  rows

for  $V^1$ . Every new sample from the uncertainty set with radius  $r_m$  creates at most  $Y_i^m$ ,  $i = 2, \dots, \mathbf{n}_m$ , new rows for the matrix of violations. Hence,

$$\begin{aligned} & \mathbb{E} [\text{the number of rows of matrix } V^{\mathbf{n}}] \\ & \leq 1 + \mathbb{E} \left[ \sum_{i=1}^{\mathbf{n}_m} Y_i^m \right] + 2 \mathbb{E} \left[ \sum_{j=1}^{m-1} \sum_{i=1}^{\mathbf{n}_j} Y_i^j \right] \\ & \leq 1 + P_e(r_m) \mathbb{E}[\mathbf{n}_m] + 2 \sum_{j=1}^{m-1} P_e(r_j) \mathbb{E}[\mathbf{n}_j]. \end{aligned}$$

By Lemma 6 of [6], we have

$$(E.1) \quad \mathbb{E}[\mathbf{n}_j] = \left[ 1 - \left( \frac{r_j}{r_{j+1}} \right)^d \right] N, \quad j = 1, \dots, m-1.$$

By (E.1) and using the fact that  $\mathbb{E}[\mathbf{n}_m] = N$ ,  $P_e(r_m) = P_e(a)$ , we have

$$\begin{aligned} & \mathbb{E} [\text{the number of rows of matrix } V^{\mathbf{n}}] \\ & \leq 1 + NP_e(a) + 2N \sum_{j=1}^{m-1} P_e(r_j) \left[ 1 - \left( \frac{r_j}{r_{j+1}} \right)^d \right]. \quad \square \end{aligned}$$

LEMMA E.4. *For any grid scheme,*

$$\begin{aligned} & \left| \sum_{i=1}^{m-1} P_e(r_i) \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] - d \sum_{i=1}^{m-1} \frac{P_e(r_{i+1})(r_{i+1} - r_i)}{r_{i+1}} \right| \\ & < \frac{d(d-1)\lambda(\lambda-1)\varpi}{2a} + \frac{\lambda d}{a} \sum_{i=1}^{m-1} P_e(r_{i+1})(r_{i+1} - r_i) - \frac{\lambda d}{a} \sum_{i=1}^{m-1} P_e(r_i)(r_{i+1} - r_i). \end{aligned}$$

*Proof.* Note that

$$\begin{aligned} & \left| \sum_{i=1}^{m-1} P_e(r_i) \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] - d \sum_{i=1}^{m-1} \frac{P_e(r_{i+1})(r_{i+1} - r_i)}{r_{i+1}} \right| \\ & \leq \sum_{i=1}^{m-1} \left| P_e(r_i) \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] - d \frac{P_e(r_i)(r_{i+1} - r_i)}{r_{i+1}} \right| \\ & \quad + \sum_{i=1}^{m-1} \left| d \frac{P_e(r_i)(r_{i+1} - r_i)}{r_{i+1}} - d \frac{P_e(r_{i+1})(r_{i+1} - r_i)}{r_{i+1}} \right| \\ & < \sum_{i=1}^{m-1} \left| \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] - \frac{d(r_{i+1} - r_i)}{r_{i+1}} \right| \\ & \quad + \frac{\lambda d}{a} \sum_{i=1}^{m-1} [P_e(r_{i+1})(r_{i+1} - r_i) - P_e(r_i)(r_{i+1} - r_i)], \end{aligned}$$

where the last inequality follows from the facts that  $0 \leq P_e(r_i) \leq P_e(r_{i+1}) \leq 1$  and

$r_{i+1} > \frac{a}{\lambda}$ . Making use of Lemmas E.1 and E.2, we have

$$\begin{aligned}
 & \left| \sum_{i=1}^{m-1} P_e(r_i) \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] - d \sum_{i=1}^{m-1} \frac{P_e(r_{i+1})(r_{i+1} - r_i)}{r_{i+1}} \right| \\
 & \leq \frac{d(d-1)}{2} \sum_{i=1}^{m-1} \left( \frac{r_{i+1} - r_i}{r_{i+1}} \right)^2 \\
 & \quad + \frac{\lambda d}{a} \sum_{i=1}^{m-1} P_e(r_{i+1})(r_{i+1} - r_i) - \frac{\lambda d}{a} \sum_{i=1}^{m-1} P_e(r_i)(r_{i+1} - r_i) \\
 & \leq \frac{d(d-1)\lambda(\lambda-1)\varpi}{2a} + \frac{\lambda d}{a} \sum_{i=1}^{m-1} P_e(r_{i+1})(r_{i+1} - r_i) \\
 & \quad - \frac{\lambda d}{a} \sum_{i=1}^{m-1} P_e(r_i)(r_{i+1} - r_i). \quad \square
 \end{aligned}$$

LEMMA E.5. For a set of grid points  $\mathcal{G} = \{r_\ell \mid 1 \leq \ell \leq m\}$  with  $\frac{a}{\lambda} = r_1 < r_2 < \dots < r_m = a$ , define function  $\aleph(\cdot)$  such that

$$\aleph(\mathcal{G}) = \sum_{\ell=1}^{m-1} P_e(r_\ell) \left[ 1 - \left( \frac{r_\ell}{r_{\ell+1}} \right)^d \right].$$

Then, for any two sets of grid points  $\mathcal{G}_1$  and  $\mathcal{G}_2$  such that  $\mathcal{G}_1 \subset \mathcal{G}_2$ ,

$$\aleph(\mathcal{G}_1) \leq \aleph(\mathcal{G}_2).$$

*Proof.* Consider two sequences of grid points  $\mathcal{G}_1 = \{r_\ell \mid 1 \leq \ell \leq m\}$  and  $\mathcal{G}_2 = \{\hat{r}_\ell \mid 1 \leq \ell \leq m+1\}$  such that

$$\frac{a}{\lambda} = r_1 < r_2 < \dots < r_m = a, \quad \frac{a}{\lambda} = \hat{r}_1 < \hat{r}_2 < \dots < \hat{r}_{m+1} = a,$$

and such that  $\mathcal{G}_2$  is obtained from  $\mathcal{G}_1$  by adding a grid point  $\hat{r}_{i+1}$  to interval  $(r_i, r_{i+1})$  where  $1 \leq i \leq m-1$ , i.e.,  $\hat{r}_j = r_j$ ,  $j = 1, \dots, i$  and  $\hat{r}_{j+1} = r_j$ ,  $j = i+1, \dots, m$ . By the definition of function  $\aleph(\cdot)$ , we have

$$\begin{aligned}
 & \aleph(\mathcal{G}_2) - \aleph(\mathcal{G}_1) \\
 & = \sum_{\tau=1}^m P_e(\hat{r}_\tau) \left[ 1 - \left( \frac{\hat{r}_\tau}{\hat{r}_{\tau+1}} \right)^d \right] - \sum_{\tau=1}^{m-1} P_e(r_\tau) \left[ 1 - \left( \frac{r_\tau}{r_{\tau+1}} \right)^d \right] \\
 & = P_e(r_i) \left[ 1 - \left( \frac{r_i}{\hat{r}_{i+1}} \right)^d \right] + P_e(\hat{r}_{i+1}) \left[ 1 - \left( \frac{\hat{r}_{i+1}}{r_{i+1}} \right)^d \right] - P_e(r_i) \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right].
 \end{aligned}$$

By virtue of the fact that  $P_e(\hat{r}_{i+1}) \geq P_e(r_i)$ , we have

$$\begin{aligned}
 & \aleph(\mathcal{G}_2) - \aleph(\mathcal{G}_1) \\
 & \geq P_e(r_i) \left[ 1 - \left( \frac{r_i}{\hat{r}_{i+1}} \right)^d \right] + P_e(r_i) \left[ 1 - \left( \frac{\hat{r}_{i+1}}{r_{i+1}} \right)^d \right] - P_e(r_i) \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] \\
 & = P_e(r_i) \left[ 1 - \left( \frac{r_i}{\hat{r}_{i+1}} \right)^d - \left( \frac{\hat{r}_{i+1}}{r_{i+1}} \right)^d + \left( \frac{r_i}{r_{i+1}} \right)^d \right].
 \end{aligned}$$

Recalling that  $r_i < \widehat{r}_{i+1} < r_{i+1}$ , we have

$$\begin{aligned} & 1 - \left( \frac{r_i}{\widehat{r}_{i+1}} \right)^d - \left( \frac{\widehat{r}_{i+1}}{r_{i+1}} \right)^d + \left( \frac{r_i}{r_{i+1}} \right)^d \\ &= \frac{\widehat{r}_{i+1}^d - r_i^d}{\widehat{r}_{i+1}^d} - \frac{\widehat{r}_{i+1}^d - r_i^d}{r_{i+1}^d} \\ &= \frac{(\widehat{r}_{i+1}^d - r_i^d)(r_{i+1}^d - \widehat{r}_{i+1}^d)}{\widehat{r}_{i+1}^d r_{i+1}^d} \\ &> 0. \end{aligned}$$

It follows that  $\aleph(\mathcal{G}_2) - \aleph(\mathcal{G}_1) \geq 0$ .  $\square$

We are now in the position to prove statement (II) of the theorem. For any set of grid points, we can reduce the maximal gap between grid points by adding grid points. Every new grid point is placed at the middle of one of the previous intervals which possesses the *largest* width in order to ensure that, as more grid points are added, the maximal gap of grid points tends to zero. In this process, we create a series of nested sets of grid points  $\mathcal{G}_k$ ,  $k = 1, 2, \dots, \infty$ , such that  $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{G}_3 \subset \dots$ . Note that

$$\begin{aligned} & \left| \sum_{i=1}^{m-1} P_e(r_i) \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] - d \int_{\frac{a}{\lambda}}^a \frac{P_e(x)}{x} dx \right| \\ & \leq \left| \sum_{i=1}^{m-1} P_e(r_i) \left[ 1 - \left( \frac{r_i}{r_{i+1}} \right)^d \right] - d \sum_{i=1}^{m-1} \frac{P_e(r_{i+1})(r_{i+1} - r_i)}{r_{i+1}} \right| \\ & \quad + \left| d \sum_{i=1}^{m-1} \frac{P_e(r_{i+1})(r_{i+1} - r_i)}{r_{i+1}} - d \int_{\frac{a}{\lambda}}^a \frac{P_e(x)}{x} dx \right| \\ (E.2) \quad & \leq \frac{d(d-1)\lambda(\lambda-1)\varpi}{2a} + \frac{\lambda d}{a} \sum_{i=1}^{m-1} P_e(r_{i+1})(r_{i+1} - r_i) \\ & \quad - \frac{\lambda d}{a} \sum_{i=1}^{m-1} P_e(r_i)(r_{i+1} - r_i) \\ & \quad + d \left| \sum_{i=1}^{m-1} \frac{P_e(r_{i+1})(r_{i+1} - r_i)}{r_{i+1}} - \int_{\frac{a}{\lambda}}^a \frac{P_e(x)}{x} dx \right|, \end{aligned}$$

where inequality (E.2) follows from Lemma E.4. By Lemma B.1,  $\mathbb{P}(\cdot)$  is a continuous function with respect to  $r$ . Consequently,  $\frac{P_e(x)}{x}$  is Riemann integrable over interval  $[\frac{a}{\lambda}, a]$  and

$$\lim_{\varpi \rightarrow 0} \sum_{i=1}^{m-1} \frac{P_e(r_{i+1})(r_{i+1} - r_i)}{r_{i+1}} = \int_{\frac{a}{\lambda}}^a \frac{P_e(x)}{x} dx.$$

Moreover, since  $P_e(x)$  is Riemann integrable, we have

$$\lim_{\varpi \rightarrow 0} \sum_{i=1}^{m-1} P_e(r_{i+1})(r_{i+1} - r_i) = \lim_{\varpi \rightarrow 0} \sum_{i=1}^{m-1} P_e(r_i)(r_{i+1} - r_i) = \int_{\frac{a}{\lambda}}^a P_e(x) dx.$$

Hence, the right-hand side of inequality (E.2) can be made arbitrarily small by successively cutting the gap between grid points in half with new grid points. This proves

that

$$\lim_{k \rightarrow \infty} \aleph(\mathcal{G}_k) = d \int_{\frac{a}{\lambda}}^a \frac{P_e(x)}{x} dx.$$

On the other hand, by Lemma E.5, we have  $\aleph(\mathcal{G}_1) \leq \aleph(\mathcal{G}_2) \leq \aleph(\mathcal{G}_3) \leq \dots$ . Combining the convergency and the monotone property of sequence  $\{\aleph(\mathcal{G}_k)\}_{k=1}^\infty$ , we can conclude that  $\aleph(\mathcal{G}) \leq d \int_{\frac{a}{\lambda}}^a \frac{P_e(x)}{x} dx$  for any set of grid points  $\mathcal{G}$ . By Lemma E.3, the expected number of rows of the matrix of violations  $V^n$  is no greater than

$$\begin{aligned} 1 + NP_e(a) + 2N\aleph(\mathcal{G}) &\leq 1 + NP_e(a) + 2Nd \int_{\frac{a}{\lambda}}^a \frac{P_e(x)}{x} dx \\ &= 1 + NP_e(a) + 2Nd \int_{\frac{a}{h}}^a \frac{P_e(x)}{x} dx \end{aligned}$$

for any  $\mathcal{G}$ . Such a bound applies to any  $V^j$  because the number of rows of  $V^j$  is nondecreasing with respect to  $j$ . Finally, the inequality of (6.6) can be proved by making use of the observation that  $P_e(x) \leq P_e(a)$  for all  $x \in [\frac{a}{h}, a]$ .

**Appendix F. Proof of Theorem 6.2.** We need the following lemma, which has recently been obtained in [8].

LEMMA F.1. *Let  $X_i$ ,  $i = 1, \dots, N$ , be i.i.d. Bernoulli random variables such that  $\Pr\{X_i = 1\} = 1 - \Pr\{X_i = 0\} = \mathbb{P}_X > 0$ . Let  $K = \frac{\sum_{i=1}^N X_i}{N}$ . Then*

$$\Pr\{\mathcal{L}(K) < \mathbb{P}_X < \mathcal{U}(K)\} > 1 - \delta.$$

Applying Lemma F.1, we have  $\Pr\{\mathcal{L}(K_{i+1}) < \mathbb{P}(r_{i+1}) < \mathcal{U}(K_{i+1})\} > 1 - \frac{\delta}{2}$  and  $\Pr\{\mathcal{L}(K_i) < \mathbb{P}(r_i) < \mathcal{U}(K_i)\} > 1 - \frac{\delta}{2}$ . Hence, by Bonferroni's inequality [11],

$$\Pr\{\mathcal{L}(K_{i+1}) < \mathbb{P}(r_{i+1}) < \mathcal{U}(K_{i+1}), \mathcal{L}(K_i) < \mathbb{P}(r_i) < \mathcal{U}(K_i)\} > 1 - \delta.$$

By the definitions of  $\mathbb{P}^*(r)$ ,  $\bar{\mathbb{P}}(r)$ , and  $\underline{\mathbb{P}}(r)$ , we have that event  $\{\mathcal{L}(K_{i+1}) < \mathbb{P}(r_{i+1}) < \mathcal{U}(K_{i+1}), \mathcal{L}(K_i) < \mathbb{P}(r_i) < \mathcal{U}(K_i)\}$  implies event  $\{\underline{\mathbb{P}}(r) + \varsigma < \mathbb{P}^*(r) < \bar{\mathbb{P}}(r) - \varsigma$  for all  $r \in [r_i, r_{i+1}]\}$ . Hence,  $\Pr\{\underline{\mathbb{P}}(r) + \varsigma < \mathbb{P}^*(r) < \bar{\mathbb{P}}(r) - \varsigma$  for all  $r \in [r_i, r_{i+1}]\} > 1 - \delta$ . By Theorem 4.1 and the gridding scheme,  $\Pr\{|\mathbb{P}^*(r) - \mathbb{P}(r)| < \varsigma$  for all  $r \in [r_i, r_{i+1}]\} = 1$ . Applying Bonferroni's inequality, we have

$$(F.1) \quad \Pr\{\underline{\mathbb{P}}(r) + \varsigma < \mathbb{P}^*(r) < \bar{\mathbb{P}}(r) - \varsigma, |\mathbb{P}^*(r) - \mathbb{P}(r)| < \varsigma \forall r \in [r_i, r_{i+1}]\} > 1 - \delta.$$

Finally, the theorem is proved by observing that the left-hand side of inequality (F.1) is no greater than  $\Pr\{\underline{\mathbb{P}}(r) < \mathbb{P}(r) < \bar{\mathbb{P}}(r)$  for all  $r \in [r_i, r_{i+1}]\}$ .

## REFERENCES

- [1] E. W. BAI, R. TEMPO, AND M. FU, *Worst-case properties of the uniform distribution and randomized algorithms for robustness analysis*, Math. Control Signals Systems, 11 (1998), pp. 183–196.
- [2] B. R. BARMISH, C. M. LAGOA, AND R. TEMPO, *Radially truncated uniform distributions for probabilistic robustness of control systems*, in Proceedings of the American Control Conference, Albuquerque, NM, 1997, pp. 853–857.
- [3] B. R. BARMISH AND C. M. LAGOA, *The uniform distribution: A rigorous justification for its use in robustness analysis*, Math. Control Signals Systems, 10 (1997), pp. 203–222.

- [4] B. R. BARMISH AND P. S. SHCHERBAKOV, *On avoiding vertexization of robustness problems: The approximate feasibility concept*, IEEE Trans. Automat. Control, 42 (2002), pp. 819–824.
- [5] X. CHEN, J. ARAVENA, AND K. ZHOU, *Risk analysis in robust control—making the case for probabilistic robust control*, in Proceedings of the American Control Conference, Portland, OR, 2005, pp. 1533–1538.
- [6] X. CHEN, K. ZHOU, AND J. L. ARAVENA, *Fast construction of robustness degradation function*, SIAM J. Control Optim., 42 (2004), pp. 1960–1971.
- [7] X. CHEN, K. ZHOU, AND J. ARAVENA, *Fast universal algorithms for robustness analysis*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 1926–1931.
- [8] X. CHEN, K. ZHOU, AND J. ARAVENA, *Explicit formula for constructing binomial confidence interval with guaranteed coverage probability*, Comm. Statist. A—Theory Methods, 37 (2008), pp. 1173–1180.
- [9] H. CHERNOFF, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, Ann. Math. Statist., 23 (1952), pp. 493–507.
- [10] C. J. CLOPPER AND E. S. PEARSON, *The use of confidence or fiducial limits illustrated in the case of the binomial*, Biometrika, 26 (1934), pp. 404–413.
- [11] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Wiley, New York, 1968.
- [12] R. R. DE GASTON AND M. G. SAFONOV, *Exact calculation of the multiloop stability margin*, IEEE Trans. Automat. Control, 33 (1988), pp. 156–171.
- [13] P. F. HOKAYEM AND C. T. ABDALLAH, *Quasi-Monte Carlo methods in robust control design*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 2435–2440.
- [14] S. KANEV, B. DE SCHUTTER, AND M. VERHAEGEN, *An ellipsoid algorithm for probabilistic robust controller design*, Systems Control Lett., 49 (2003), pp. 365–375.
- [15] V. KOLTCHINSKII, C. T. ABDALLAH, M. ARIOLA, P. DORATO, AND D. PANCHENKO, *Improved sample complexity estimates for statistical learning control of uncertain systems*, IEEE Trans. Automat. Control, 46 (2000), pp. 2383–2388.
- [16] C. M. LAGOA, *Probabilistic enhancement of classic robustness margins: A class of nonsymmetric distributions*, in Proceedings of the IEEE American Control Conference, Chicago, IL, 2000, pp. 3802–3806.
- [17] C. M. LAGOA, X. LI, M. C. MAZZARO, AND M. SZNAIER, *Sampling random transfer functions*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 2429–2434.
- [18] C. MARRISON AND R. F. STENGEL, *Robust control system design using random search and genetic algorithms*, IEEE Trans. Automat. Control, 42 (1997), pp. 835–839.
- [19] L. R. RAY AND R. F. STENGEL, *A Monte Carlo approach to the analysis of control systems robustness*, Automatica, 3 (1993), pp. 229–236.
- [20] S. R. ROSS AND B. R. BARMISH, *Distributionally robust gain analysis for systems containing complexity*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 5020–5025.
- [21] R. F. STENGEL AND L. R. RAY, *Stochastic robustness of linear time-invariant systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 82–87.
- [22] Q. WANG AND R. F. STENGEL, *Robust control of nonlinear systems with parametric uncertainty*, Automatica, 38 (2002), pp. 1591–1599.

## ROBUSTNESS OF $\lambda$ -TRACKING IN THE GAP METRIC\*

ACHIM ILCHMANN<sup>†</sup> AND MARKUS MUELLER<sup>†</sup>

**Abstract.** For  $m$ -input,  $m$ -output, finite-dimensional, linear systems satisfying the classical assumptions of adaptive control (i.e., (i) minimum phase, (ii) relative degree one, and (iii) “positive” high-frequency gain), it is well known that the adaptive  $\lambda$ -tracker “ $u = -k e$ ,  $k = \max\{0, |e| - \lambda\}|e|$ ” achieves  $\lambda$ -tracking of the tracking error  $e$  if applied to such a system: all states of the closed-loop system are bounded, and  $|e|$  is ultimately bounded by  $\lambda$ , where  $\lambda > 0$  is prespecified and may be arbitrarily small. Invoking the conceptual framework of the nonlinear gap metric, we show that the  $\lambda$ -tracker is robust. In the present setup this means in particular that the  $\lambda$ -tracker copes with bounded input and output disturbances, and, more importantly, it may even be applied to a system not satisfying any of the classical conditions (i)–(iii) as long as the initial conditions and the disturbances are “small” and the system is “close” (in terms of “small” gap) to a system satisfying (i)–(iii).

**Key words.** adaptive control,  $\lambda$ -tracking, gap metric, robust stability

**AMS subject classifications.** 93D21, 93D09, 93C40, 93D25

**DOI.** 10.1137/07070142X

### Nomenclature

$\mathbb{C}_+, \mathbb{C}_-$	$= \{s \in \mathbb{C} \mid \operatorname{Re} s > 0\}, \{s \in \mathbb{C} \mid \operatorname{Re} s < 0\}$ , respectively
$A > 0$	if and only if $x^T A x > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$ , where $A = A^T \in \mathbb{R}^{n \times n}$
$ x $	$= \sqrt{x^T x}$ , the Euclidean norm of $x \in \mathbb{R}^n$
$ A $	$= \max \{ Ax  \mid x \in \mathbb{R}^m,  x  = 1\}$ , the induced matrix norm for $A \in \mathbb{R}^{n \times m}$
$\ v\ _{\mathcal{V}}$	the norm of $v \in \mathcal{V}$ for any normed vector space $\mathcal{V}$
$L^p(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell)$	the space of $p$ -integrable functions $y : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell$ , $1 \leq p < \infty$ with norm
$\ y\ _{L^p(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell)}$	$= \left(\int_0^\infty  y(t) ^p dt\right)^{\frac{1}{p}}$
$L^p_{\text{loc}}(I \rightarrow \mathbb{R}^\ell)$	the space of locally $p$ -integrable functions $y : I \rightarrow \mathbb{R}^\ell$ , with $\int_K  y(t) ^p dt < \infty$ for all compact $K \subset I$ , where $1 \leq p < \infty$ and $I \subset \mathbb{R}_{\geq 0}$ is an interval
$L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell)$	the space of essentially bounded functions $y : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell$ with norm
$\ y\ _{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell)}$	$= \operatorname{ess\,sup}_{t \geq 0}  y(t) $
$L^\infty_{\text{loc}}(I \rightarrow \mathbb{R}^\ell)$	the space of locally bounded functions $y : I \rightarrow \mathbb{R}^\ell$ , with $\operatorname{ess\,sup}_{t \in K}  y(t)  < \infty$ for all compact $K \subset I$ , where $I \subset \mathbb{R}_{\geq 0}$ is an interval

\*Received by the editors August 28, 2007; accepted for publication (in revised form) June 16, 2008; published electronically October 22, 2008.

<http://www.siam.org/journals/sicon/47-5/70142.html>

<sup>†</sup>Institute of Mathematics, Technical University Ilmenau, Weimarer Straße 25, 98693 Ilmenau, Germany (achim.ilchmann@tu-ilmenau.de, markus.mueller@tu-ilmenau.de).



$W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell)$  the Sobolev space of absolutely continuous functions  $y : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell$  with  $y, \dot{y} \in L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell)$  and norm

$$\|y\|_{W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell)} = \|y\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell)} + \|\dot{y}\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^\ell)}$$

$W_{\text{loc}}^{1,\infty}(I \rightarrow \mathbb{R}^\ell)$  the space of absolutely continuous functions  $y : I \rightarrow \mathbb{R}^\ell$ , with  $y, \dot{y} \in L_{\text{loc}}^\infty(I \rightarrow \mathbb{R}^\ell)$ , where  $I \subset \mathbb{R}_{\geq 0}$  is an interval

$$\text{dist}(e, [-\lambda, \lambda]) = \max\{0, |e| - \lambda\} \text{ for } e \in \mathbb{R}^m \text{ and } \lambda > 0$$

$$d_\lambda(e) = \max\{0, |e| - \lambda\} \text{ for } e \in \mathbb{R}^m \text{ and } \lambda > 0$$

**1. Introduction.** In this paper we show robustness of  $\lambda$ -stabilization and  $\lambda$ -tracking (i.e., stabilization and tracking with a final accuracy of prespecified  $\lambda > 0$ ) for linear  $n$ -dimensional,  $m$ -input,  $m$ -output systems of the form

$$(1.1) \quad \begin{cases} \dot{x}(t) = Ax(t) + Bu_1(t), & x(0) = x^0, \\ y_1(t) = Cx(t), \end{cases}$$

where  $A \in \mathbb{R}^{n \times n}$ ,  $B, C^T \in \mathbb{R}^{n \times m}$ ,  $x^0 \in \mathbb{R}^n$ , subject to additive input/output disturbances  $u_0, y_0$ , respectively,

$$(1.2) \quad u_0 = u_1 + u_2, \quad y_0 = y_1 + y_2,$$

as depicted in Figure 1, where the *plant*  $P$  maps the interior input signal  $u_1$  to the interior output signal  $y_1$  and the *controller*  $C$  maps the interior output signal  $y_2$  to the interior input signal  $u_2$ . In our setup,  $P$  will always be a linear initial value problem of the form (1.1) and the controller  $C$  will be a dynamical system, specified in due course. In case of zero disturbances  $u_0 \equiv y_0 \equiv 0$ , it is well known that (1.1) can be stabilized by proportional high-gain ( $k \gg 0$ ) output feedback

$$(1.3) \quad u_2(t) = -k y_2(t),$$

provided (1.1) is *minimum phase*, i.e.,

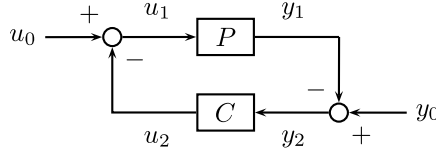
$$\forall s \in \overline{\mathbb{C}}_+ : \det \begin{bmatrix} sI - A & B \\ C & 0 \end{bmatrix} \neq 0,$$

and its transfer function  $C(sI - A)^{-1}B$  has *strict relative degree* one with “positive” high-frequency gain, i.e.,  $CB + (CB)^T > 0$ . This system class is denoted, for  $n, m \in \mathbb{N}$  with  $n \geq m$ , as

$$\widetilde{\mathcal{M}}_{n,m} := \left\{ (A, B, C) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m} \times \mathbb{R}^{m \times n} \left| \begin{array}{l} CB + (CB)^T > 0 \\ \forall s \in \overline{\mathbb{C}}_+ : \det \begin{bmatrix} sI_n - A & B \\ C & 0 \end{bmatrix} \neq 0 \end{array} \right. \right\}.$$

Note that the state space dimension  $n \in \mathbb{N}$  need not be known but only the input/output dimension  $m \in \mathbb{N}$ , and, most importantly, only structural assumptions are required, but the system entries are completely unknown. A sufficiently high-gain  $k$  in (1.3) can be found adaptively. More precisely, any system  $(A, B, C) \in \widetilde{\mathcal{M}}_n$  can be stabilized adaptively, in the presence of  $L^2$  input/output disturbances, by the controller (ubiquitous in the adaptive control literature)

$$(1.4) \quad \begin{cases} \dot{k}(t) = |y_2(t)|^2, & k(0) = k^0 \in \mathbb{R}, \\ u_2(t) = -k(t) y_2(t) \end{cases}$$

FIG. 1. The closed-loop system  $[P, C]$ .

in the sense that all states of the closed-loop system (1.1), (1.2), (1.4) are bounded and  $\lim_{t \rightarrow \infty} y_1(t) = 0$ . This approach has been introduced by the seminal work of [14, 15, 17]; see also the survey [6].

The surprising property of the controller (1.4) is not only its simplicity but also its robustness: it is also applicable in the presence of additive  $L^2$  input/output disturbances, and it may stabilize systems (1.1) not belonging to  $\widetilde{\mathcal{M}}_{n,m}$  but sufficiently “close”—in terms of the *gap metric* defined in section 3—to some  $(A, B, C)$  in normal form and belonging to  $\widetilde{\mathcal{M}}_{n,m}$ . This has been proved in [4].

However, the controller (1.4) has the shortcomings that, if tracking is the control objective, it needs to be combined with an internal model (thus becoming more involved) and, more importantly, fails for stabilizing nonlinear systems or in the presence of additive arbitrarily small input or output  $L^\infty$ -disturbances. To overcome these shortcomings, the so-called  $\lambda$ -tracker

$$(1.5) \quad \begin{cases} \dot{k}(t) = \text{dist}(y_2(t), [-\lambda, \lambda]) \cdot |y_2(t)|, & k(0) = k^0, \\ u_2(t) = -k(t)y_2(t) \end{cases}$$

for  $\lambda > 0$  and  $k^0 \in \mathbb{R}$  has been introduced by [9]. The application of the  $\lambda$ -tracker (1.5) to any system (1.1) belonging to  $\widetilde{\mathcal{M}}_{n,m}$ , via (1.2), satisfies, in the presence of arbitrary input/output disturbances  $u_0, y_0$  which are bounded with essentially bounded derivative, arbitrary initial conditions  $x^0 \in \mathbb{R}^n$ ,  $k^0 \in \mathbb{R}$  and any arbitrarily small design parameter  $\lambda > 0$ , the control objectives of  $\lambda$ -tracking:

- all signals and their derivatives of the closed-loop system (1.1), (1.2), (1.5) are bounded;
- $\limsup_{t \rightarrow \infty} \text{dist}(y_2(t), [-\lambda, \lambda]) = 0$ .

This result has been generalized to nonlinear and infinite-dimensional systems [10] and applied, to name but a few, to regulate biogas tower reactors [8], chemical reactors [12], and insulin delivery for diabetic patients [2] by preserving the simplicity of the control strategy. Note also that it is a tracking result without invoking an internal model: set  $y_0(\cdot) \equiv y_{\text{ref}}(\cdot)$  as a prespecified reference signal.

The purpose of the present paper is to show robustness properties of the  $\lambda$ -tracker in terms of the gap metric. For example, we consider

$$(1.6) \quad \begin{cases} \dot{x} = \tilde{A}x + \tilde{b}u_1, & x(0) = \tilde{x}^0, \\ y_1 = \tilde{c}x, \end{cases}$$

with  $\tilde{x}^0 \in \mathbb{R}^3$  and where, for  $\alpha, N, M > 0$ ,

$$\tilde{A} := \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \alpha NM & -NM + \alpha N + \alpha M & \alpha - N - M \end{bmatrix}, \quad \tilde{b} := \begin{bmatrix} 0 \\ 0 \\ N \end{bmatrix}, \quad \tilde{c} := [M, -1, 0].$$

This system does not belong to the class  $\widetilde{\mathcal{M}}_{3,1}$ ; its transfer function  $\frac{N(M-s)}{(s-\alpha)(s+N)(s+M)}$  does not satisfy any of the classical structural assumptions in adaptive control:

$$(1.7) \quad \left\{ \begin{array}{l} \bullet \text{ it is not minimum phase;} \\ \bullet \text{ it has relative degree two;} \\ \bullet \text{ and its high-frequency gain } -N < 0 \text{ has the "wrong" sign.} \end{array} \right.$$

However, defining, for  $n, m \in \mathbb{N}$  with  $n \geq m$ , the system class

$$\mathcal{P}_{n,m}$$

$$:= \{ (A, B, C) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m} \times \mathbb{R}^{m \times n} \mid (A, B, C) \text{ is stabilizable and detectable} \},$$

(1.6) belongs to  $\mathcal{P}_{3,1}$ , and we will show in subsection 3.5 and Example 4.6 that (1.6) is close (in terms of the gap metric) to a system belonging to  $\widetilde{\mathcal{M}}_{1,1}$  for  $N, M$  sufficiently large and  $\tilde{x}^0$  sufficiently small, and thus (1.5) applied to (1.6) achieves  $\lambda$ -tracking.

Instead of systems  $(A, B, C) \in \widetilde{\mathcal{M}}_{n,m}$  we restrict our attention to systems in Byrnes–Isidori normal form; see, for example, [13, section 4]. That is, for each  $(A, B, C) \in \widetilde{\mathcal{M}}_{n,m}$  the matrix

$$T = [B(CB)^{-1}, V], \text{ where } V \in \mathbb{R}^{n \times (n-m)} \text{ with } \text{rk } V = n - m \text{ and } \text{im } V = \ker C,$$

converts (1.1) via the coordinate transformation  $\begin{pmatrix} y_1 \\ z \end{pmatrix} = T^{-1}x$  into

$$(1.8) \quad \left\{ \begin{array}{lll} \dot{y}_1 & = & A_1 y_1 + A_2 z + CB u_1, & y_1(0) & = & y_1^0 \in \mathbb{R}, \\ \dot{z} & = & A_3 y_1 + A_4 z, & z(0) & = & z^0 \in \mathbb{R}^{n-m}, \end{array} \right. \quad \begin{pmatrix} y_1^0 \\ z^0 \end{pmatrix} = T^{-1}x^0,$$

where

$$\begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} := T^{-1}AT, \quad \begin{bmatrix} B_1 \\ 0_{(n-m) \times m} \end{bmatrix} := \begin{bmatrix} CB \\ 0 \end{bmatrix} = T^{-1}B, \quad [I_m \quad 0_{m \times (n-m)}] = CT.$$

By the minimum-phase property,  $A_4$  has spectrum in the open left half complex plane  $\mathbb{C}_-$ . Therefore, we introduce, for  $n, m \in \mathbb{N}$  with  $n \geq m$ , the system class

$$\mathcal{M}_{n,m}$$

$$:= \left\{ (A, B, C) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m} \times \mathbb{R}^{m \times n} \left| \begin{array}{l} A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \\ C = [I_m \quad 0], \quad B_1, A_1 \in \mathbb{R}^{m \times m}, \\ \text{spec}(A_4) \subset \mathbb{C}_-, \quad B_1 + B_1^T > 0. \end{array} \right. \right\}.$$

We will study properties of the closed-loop system generated by the application of the  $\lambda$ -tracker (1.5) to systems (1.1) of class  $\mathcal{M}_{n,m}$  or of class  $\mathcal{P}_{n,m}$  in the presence of disturbances  $(u_0, y_0) \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m) \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$  satisfying the interconnection equations (1.2). The closed-loop system (1.8), (1.5), (1.2) is depicted in Figure 2.

The paper is organized as follows. In section 2 we show that  $\lambda$ -tracking is possible for all linear systems (1.1) belonging to class  $\mathcal{M}_{n,m}$  in the presence of  $W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow$

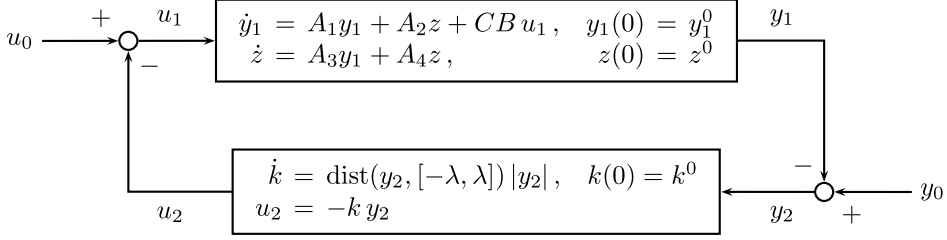


FIG. 2. The adaptive closed-loop system.

$\mathbb{R}^m$ ) input/output disturbances; see Figure 2. In section 3, we collect the basics of the framework of gap metric and graph topology from [5, 3, 4] necessary for our setup. Section 4 contains our main result, i.e., robustness of  $\lambda$ -tracking. We show that if the initial conditions, the input/output disturbances, and the gap between a nominal system belonging to class  $\mathcal{P}_{q,m}$  and a system belonging to class  $\mathcal{M}_{n,m}$  (for  $m, q, n \in \mathbb{N}$  with  $q, n \geq m$ ) are sufficiently small, then the controller (1.5) achieves  $\lambda$ -tracking for the nominal system.

**2.  $\lambda$ -tracking.** In this section we show that the control strategy given by (1.5) applied to any linear system of class  $\mathcal{M}_{n,m}$  achieves  $\lambda$ -tracking in the presence of  $W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$  input/output disturbances; see Figure 2. Set, for  $n, m \in \mathbb{N}$  with  $n \geq m$ ,

$$\mathcal{D}_{n,m} := \mathcal{M}_{n,m} \times (\mathbb{R}^m \times \mathbb{R}^{n-m} \times \mathbb{R}) \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m) \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m).$$

**PROPOSITION 2.1.** *Let  $m, n \in \mathbb{N}$  with  $n \geq m$  and  $\lambda > 0$ . Then there exists a continuous map  $\nu: \mathcal{D}_{n,m} \rightarrow \mathbb{R}_{\geq 0}$  such that, for all  $d = ([\begin{smallmatrix} A_1 & A_2 \\ A_3 & A_4 \end{smallmatrix}], B, C, (y_1^0, z^0, k^0), u_0, y_0) \in \mathcal{D}_{n,m}$ , the associated closed-loop initial value problem (1.8), (1.2), (1.5) satisfies*

$$(2.1) \quad \|(u_2, y_2, z, k)\|_{W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{m+n+1})} \leq \nu(d)$$

and

$$(2.2) \quad \limsup_{t \rightarrow \infty} |y_2(t)| \leq \lambda.$$

The result that  $\lambda$ -tracking works for the class of systems  $\mathcal{M}_n$  goes back to [9], and input disturbances are also considered in [7]. However, to prove the robustness of the  $\lambda$ -tracker in section 4, the existence of a continuous function  $\nu(\cdot)$  satisfying (2.1) is crucial. Therefore, we had to find a new proof showing (2.1) which easily shows (2.2).

*Proof of Proposition 2.1.* Let  $d = ([\begin{smallmatrix} A_1 & A_2 \\ A_3 & A_4 \end{smallmatrix}], B, C, (y_1^0, z^0, k^0), u_0, y_0) \in \mathcal{D}_{n,m}$  and set, for notational convenience,

$$\begin{aligned} h(\cdot) &:= \dot{y}_0(\cdot) - A_1 y_0(\cdot) - C B u_0(\cdot), \\ e(\cdot) &:= y_2(\cdot). \end{aligned}$$

The closed-loop initial value problem (1.8), (1.2), (1.5) is then given by

$$(2.3) \quad \begin{cases} \dot{e} = A_1 e - A_2 z - k C B e + h, & e(0) = e^0 := y_0(0) - y_1^0, \\ \dot{z} = -A_3 e + A_4 z + A_3 y_0, & z(0) = z^0, \\ \dot{k} = d_\lambda(e) |e|, & k(0) = k^0, \end{cases}$$

where  $d_\lambda$  is defined in the Nomenclature. We divide the proof into 10 steps.

*Step 1.* Since the right-hand side of (2.3) is continuous and locally Lipschitz, it follows from the theory of ordinary differential equations that (2.3) has a solution

$$(e, z, k): [0, \omega) \rightarrow \mathbb{R}^{n-m} \times \mathbb{R}^m \times \mathbb{R}_{\geq 0}$$

on a maximal interval of existence  $[0, \omega)$  for some  $\omega \in (0, \infty]$ . This solution is unique.

*Step 2.* We define some constants that are used in the following steps of the proof.

Since  $\text{spec}(A_4) \subset \mathbb{C}_-$  we have

$$(2.4) \quad \exists M_1, \mu > 0 \quad \forall t \geq 0 : |\exp(A_4 t)| \leq M_1 \exp(-\mu t).$$

Set

$$\sigma_1 := \min \text{spec} (CB + (CB)^T) / 2,$$

$$M_2 := M_1 + M_1 |A_3| (\|y_0\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)} + \lambda + \mu) / \mu,$$

$$M_3 := M_2 (1 + |z^0|) / \lambda + M_2 (1 + 1/\mu),$$

$$M_4 := |A_1| + |A_2| + \|h\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)} / \lambda,$$

$$M_5 := |k^0| + 2(M_4 + M_3 M_4 + 1) / \sigma_1,$$

$$M_6 := M_5 + |k^0| + |e^0|^2 / 2,$$

$$M_7 := (d_\lambda(e^0)^2 + 2(M_6 + |k^0|) [\sigma_1 (M_6 + |k^0|) / 2 + M_4 + M_3 M_4])^{\frac{1}{2}} + \lambda,$$

$$M_8 := M_2 (1 + |z^0| + M_7 / \mu).$$

*Step 3.* We estimate the  $z$ -dynamics in the form

$$(2.5) \quad \forall t \in [0, \omega) : \int_0^t d_\lambda(e(\tau)) |z(\tau)| d\tau \leq M_3 [k(t) - k^0].$$

Applying variation of constants to the second equation in (2.3) and invoking (2.4) gives, for all  $t \in [0, \omega)$ ,

$$\begin{aligned} |z(t)| &\leq M_1 e^{-\mu t} |z^0| + \int_0^t M_1 e^{-\mu(t-\tau)} |A_3| (|e(\tau)| + |y_0(\tau)|) d\tau \\ &\leq M_1 e^{-\mu t} |z^0| + M_1 |A_3| \int_0^t e^{-\mu(t-\tau)} (d_\lambda(e(\tau)) + \|y_0\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)} + \lambda) d\tau \\ &\leq M_1 e^{-\mu t} |z^0| + M_1 |A_3| \int_0^t e^{-\mu(t-\tau)} \|y_0\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)} d\tau \\ &\quad + M_1 |A_3| \int_0^t e^{-\mu(t-\tau)} \lambda d\tau + M_1 |A_3| \int_0^t e^{-\mu(t-\tau)} d_\lambda(e(\tau)) d\tau \\ &\leq M_1 |z^0| + \frac{M_1 |A_3|}{\mu} (\|y_0\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)} + \lambda) + M_1 |A_3| \int_0^t e^{-\mu(t-\tau)} d_\lambda(e(\tau)) d\tau \\ (2.6) \quad &\leq M_2 \left[ 1 + |z^0| + \int_0^t e^{-\mu(t-\tau)} d_\lambda(e(\tau)) d\tau \right]. \end{aligned}$$

Let

$$\forall t \in [0, \infty) \quad \forall \varphi \in L^2_{\text{loc}}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}) : (L * \varphi)(t) := \int_0^t e^{-\mu(t-\tau)} \varphi(\tau) \, d\tau.$$

Invoking the well-known inequality (see, for example, [16, p. 298])

$$\forall t \geq 0 : \|L * \varphi\|_{L^2([0,t] \rightarrow \mathbb{R})} \leq \|e^{-\mu \cdot}\|_{L^1(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})} \|\varphi\|_{L^2([0,t] \rightarrow \mathbb{R})} = \frac{1}{\mu} \|\varphi\|_{L^2([0,t] \rightarrow \mathbb{R})}$$

and the fact that

$$\forall e \in \mathbb{R} : d_\lambda(e)^2 \leq d_\lambda(e) |e|$$

yields, by (2.3), (2.6), and the Cauchy–Schwarz inequality, for all  $t \in [0, \omega)$ ,

$$\begin{aligned} & \int_0^t d_\lambda(e(\tau)) |z(\tau)| \, d\tau \\ & \leq M_2 \int_0^t d_\lambda(e(\tau)) [1 + |z^0| + (L * d_\lambda(e))(\tau)] \, d\tau \\ & \leq M_2 [1 + |z^0|] \frac{1}{\lambda} \int_0^t d_\lambda(e(\tau)) |e(\tau)| \, d\tau \\ & \quad + M_2 \left[ \|d_\lambda(e)\|_{L^2([0,t] \rightarrow \mathbb{R})}^2 + \|L * d_\lambda(e)\|_{L^2([0,t] \rightarrow \mathbb{R})}^2 \right] \\ & \leq M_2 [1 + |z^0|] \frac{1}{\lambda} \int_0^t d_\lambda(e(\tau)) |e(\tau)| \, d\tau + M_2 \left( 1 + \frac{1}{\mu} \right) \int_0^t d_\lambda(e(\tau))^2 \, d\tau. \end{aligned}$$

This proves (2.5).

*Step 4.* We estimate the  $e$ -dynamics in the form

$$(2.7) \quad \forall t \in [0, \omega) :$$

$$\frac{1}{2} d_\lambda(e(t))^2 \leq \frac{1}{2} d_\lambda(e^0)^2 - (k(t) - k^0) \left[ \frac{\sigma_1}{2} (k(t) + k^0) - M_4 - M_3 M_4 \right].$$

By (2.3) and Step 2 we have, omitting the argument  $t$ ,

$$\begin{aligned} \frac{d}{dt} \left( \frac{1}{2} d_\lambda(e(t))^2 \right) &= d_\lambda(e) |e|^{-1} e^T \dot{e} \\ &= d_\lambda(e) |e|^{-1} e^T [A_1 e - A_2 z - k C B e + h] \\ &\leq d_\lambda(e) |e| |A_1| + d_\lambda(e) |z| |A_2| + d_\lambda(e) \|h\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)} \\ &\quad - k d_\lambda(e) |e|^{-1} e^T \left( \frac{1}{2} (C B + (C B)^T) e \right) \\ &\leq -(k \sigma_1 - |A_1|) d_\lambda(e) |e| + |A_2| d_\lambda(e) |z| + d_\lambda(e) \|h\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)} \\ &\leq -(k \sigma_1 - |A_1|) d_\lambda(e) |e| + |A_2| d_\lambda(e) |z| + d_\lambda(e) \frac{|e|}{\lambda} \|h\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)} \\ &\leq -(k \sigma_1 - M_4) d_\lambda(e) |e| + M_4 d_\lambda(e) |z|, \end{aligned}$$

and hence, by integration and invoking (2.5), we arrive at

$$\forall t \in [0, \omega) : \frac{1}{2}d_\lambda(e(t))^2 \leq \frac{1}{2}d_\lambda(e^0)^2 - \int_0^t (k(\tau)\sigma_1 - M_4)\dot{k}(\tau) d\tau + M_3M_4[k(t) - k^0],$$

which yields (2.7).

*Step 5.* We show boundedness of  $k$  in the form

$$(2.8) \quad \forall t \in [0, \omega) : k(t) \leq M_6.$$

Suppose there exists  $T \in [0, \omega)$  such that  $k(T) = M_5$ ; otherwise (2.8) is obvious. Then, by monotonicity of  $k$ , it follows from (2.7) that, for all  $t \in [T, \omega)$ ,

$$\begin{aligned} 0 &\leq \frac{1}{2}d_\lambda(e(t))^2 \leq \frac{1}{2}d_\lambda(e^0)^2 - \frac{\sigma_1}{2}(k(t) - k^0) \left[ k(t) + k^0 - \frac{2}{\sigma_1}(M_4 + M_3M_4) \right] \\ &\leq \frac{1}{2}d_\lambda(e^0)^2 - \frac{\sigma_1}{2}(k(t) - k^0) \left[ M_5 + k^0 - \frac{2}{\sigma_1}(M_4 + M_3M_4) \right] \\ &= \frac{1}{2}d_\lambda(e^0)^2 - \frac{\sigma_1}{2}(k(t) - k^0) \left[ |k^0| + k^0 + \frac{2}{\sigma_1} \right] \\ &\leq \frac{1}{2}d_\lambda(e^0)^2 - (k(t) - k^0), \end{aligned}$$

and thus

$$\forall t \in [T, \omega) : k(t) - k^0 \leq \frac{1}{2}d_\lambda(e^0)^2 \leq \frac{1}{2}|e^0|^2$$

and

$$\forall t \in [0, T) : k(t) - k^0 \leq M_5 - k^0,$$

whence (2.8).

*Step 6.* We show boundedness of  $e$  in the form

$$(2.9) \quad \forall t \in [0, \omega) : |e(t)| \leq M_7.$$

An application of (2.8) to (2.7) gives, for all  $t \in [0, \omega)$ ,

$$\begin{aligned} |e(t)| &\leq d_\lambda(e(t)) + \lambda \\ &\leq \left( d_\lambda(e^0)^2 - 2(k(t) - k^0) \left[ \frac{\sigma_1}{2}(k(t) + k^0) - M_4 - M_3M_4 \right] \right)^{\frac{1}{2}} + \lambda \\ &\leq \left( d_\lambda(e^0)^2 + 2(M_6 + |k^0|) \left[ \frac{\sigma_1}{2}(M_6 + |k^0|) + M_4 + M_3M_4 \right] \right)^{\frac{1}{2}} + \lambda. \end{aligned}$$

Note that the argument of the root in the second line is nonnegative; see Step 5. Now (2.9) follows from Step 2.

*Step 7.* Boundedness of  $z$  in the form

$$(2.10) \quad \forall t \in [0, \omega) : |z(t)| \leq M_2 \left[ 1 + |z^0| + \int_0^t e^{-\mu(t-\tau)} M_7 d\tau \right] \leq M_8$$

follows from applying (2.9) to (2.6).

*Step 8.* We show  $\omega = \infty$ .

Since  $\omega$  was chosen maximal, (2.8)–(2.10) yield  $\omega = \infty$ .

*Step 9.* We show (2.1).

It follows from Steps 5–8 that  $(u_2, y_2, z, k)$  is uniformly bounded in terms of  $d = ([\begin{smallmatrix} A_1 & A_2 \\ A_3 & A_4 \end{smallmatrix}], B, C, (y_1^0, z^0, k^0), u_0, y_0)$ . Moreover, applying Steps 5–8 again and invoking (2.3) yields uniform boundedness of  $(\dot{u}_2, \dot{y}_2, \dot{z}, \dot{k})$  in terms of  $d$ . Now the existence of a continuous function  $\nu: \mathcal{D}_{n,m} \rightarrow \mathbb{R}_{\geq 0}$  such that (2.1) holds is straightforward by invoking the constants from Step 2.

*Step 10.* We show (2.2).

Since  $k \in L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$  by (2.1) it follows from  $\|d_\lambda(y_2) |y_2|\|_{L^1([0,t] \rightarrow \mathbb{R})} = k(t) - k^0$  that  $d_\lambda(y_2) |y_2| \in L^1(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$ .

Since  $y_2 \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$  there exists  $M > 0$  such that  $\text{ess sup}_{t \geq 0} |(\dot{y}_2)_i(t)| < M$  for all  $i \in \{1, \dots, m\}$ , which gives

$$\forall s \geq 0 \, \forall i \in \{1, \dots, m\} \, \forall t \in [0, s] \, \exists \tau_i \in (t, s) : (\dot{y}_2)_i(\tau_i) = \frac{(y_2)_i(s) - (y_2)_i(t)}{s - t} < M,$$

and so

$$\forall i \in \{1, \dots, m\} \, \forall t \in [0, s] : |(y_2)_i(s) - (y_2)_i(t)| < M(s - t).$$

For  $\delta = \frac{\varepsilon}{M}$  we arrive at

$$\forall i \in \{1, \dots, m\} \, \forall \varepsilon > 0 \, \exists \delta > 0 \, \forall t, s \in \mathbb{R}_{\geq 0} \text{ with } |t - s| < \delta : |(y_2)_i(t) - (y_2)_i(s)| < \varepsilon;$$

i.e.,  $y_2$  is uniformly continuous. Boundedness and uniform continuity of  $y_2$  and the continuity of  $e \mapsto d_\lambda(e) |e|$  give uniform continuity of  $t \mapsto d_\lambda(y_2(t)) |y_2(t)|$ . So Barb alat’s lemma (see [1]) gives

$$\lim_{t \rightarrow \infty} d_\lambda(y_2(t)) |y_2(t)| = 0,$$

which yields (2.2) and completes the proof.  $\square$

**3. The concept of gap metric.** The material in this section is based on [5, section II], [4, section 2], and [3, section 2] and contains the fundamental results necessary for proving robustness in section 4.

**3.1. Terminology.** Let  $\mathcal{X}$  be a nonempty set and, for  $0 < \omega \leq \infty$ , let  $\mathcal{S}_\omega$  denote the set of locally integrable maps  $[0, \omega) \rightarrow \mathcal{X}$ . For simplicity, we write  $\mathcal{S} := \mathcal{S}_\infty$ . For  $0 < \tau < \omega \leq \infty$ ,  $T_\tau : \mathcal{S}_\omega \rightarrow \mathcal{S}$  denotes the operator given by

$$T_\tau v := \begin{cases} v(t), & t \in [0, \tau), \\ 0, & t \in [\tau, \infty). \end{cases}$$

With  $\mathcal{V} \subset \mathcal{S}$  we associate spaces as follows:

$$\mathcal{V}_e = \{v \in \mathcal{S} \mid \forall \tau > 0 : T_\tau v \in \mathcal{V}\}, \quad \text{the extended space;}$$

$$\mathcal{V}_\omega = \{v \in \mathcal{S}_\omega \mid \forall \tau \in (0, \omega) : T_\tau v \in \mathcal{V}\}, \quad 0 < \omega \leq \infty;$$

$$\mathcal{V}_a = \bigcup_{\omega \in (0, \infty]} \mathcal{V}_\omega, \quad \text{the ambient space.}$$



The ambient space  $\mathcal{V}_a$  is a subset of the union of all  $\mathcal{S}_\omega$ ,  $\omega \in (0, \infty]$ . Thus, for  $v \in \mathcal{V}_a$  the domain of  $v$ , i.e., the set of values in  $[0, \infty)$  where  $v$  is defined, denoted by  $\text{dom}(v)$ , is not obvious. If  $v, w \in \mathcal{V}_a$  with  $v|_I = w|_I$  on  $I = \text{dom}(v) \cap \text{dom}(w)$ , then we write  $v = w$ . For  $(u, y) \in \mathcal{V}_a \times \mathcal{V}_a$ , the domains of  $u$  and  $y$  may be different; we adopt the convention

$$\text{dom}(u, y) := \text{dom}(u) \cap \text{dom}(y).$$

We say  $\mathcal{V} \subset \mathcal{S}$  is a *signal space* if and only if it is a vector space and has the property that  $\sup_{\tau \geq 0} \|T_\tau v\|_{\mathcal{V}} < \infty$  implies  $v \in \mathcal{V}$ . In our applications,  $\mathcal{V}$  will frequently be the normed signal space  $W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ , in which case  $\mathcal{V}_e = W_{\text{loc}}^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ ,  $\mathcal{V}_\omega = W_{\text{loc}}^{1,\infty}([0, \omega) \rightarrow \mathbb{R}^m)$  for  $\omega \in (0, \infty]$ , and  $\mathcal{V}_a = \cup_{0 < \omega \leq \infty} W_{\text{loc}}^{1,\infty}([0, \omega) \rightarrow \mathbb{R}^m)$ . It is important to note that  $\mathcal{V}_\omega \supsetneq W^{1,\infty}([0, \omega) \rightarrow \mathbb{R}^m)$ .

For a normed signal space  $\mathcal{U}$  and the Euclidean space  $\mathbb{R}^l$ ,  $l \in \mathbb{N}$ , we will also consider subsets of  $\mathcal{V} = \mathbb{R}^l \times \mathcal{U}$ , which, on identifying each  $\theta \in \mathbb{R}^l$  with the constant signal  $t \mapsto \theta$ , can be thought of as a normed signal space with the norm given by  $\|(\theta, x)\|_{\mathcal{V}} = \sqrt{|\theta|^2 + \|x\|_{\mathcal{U}}^2}$ .

**3.2. Well posedness.** A mapping  $Q: \mathcal{X}_1 \rightarrow \mathcal{X}_2$  between signal spaces is said to be *causal* if and only if, for all  $\tau > 0$ ,  $x, y \in \mathcal{X}_1$ ,  $T_\tau x = T_\tau y$  implies  $T_\tau Qx = T_\tau Qy$ . Let  $\mathcal{U}$  and  $\mathcal{Y}$  be normed signal spaces and let  $P: \mathcal{U}_a \rightarrow \mathcal{Y}_a$  and  $C: \mathcal{Y}_a \rightarrow \mathcal{U}_a$  be causal mappings representing a plant and controller, respectively. Our central concern is the system of equations

$$(3.1) \quad [P, C] : \quad y_1 = Pu_1, \quad u_2 = Cy_2, \quad u_0 = u_1 + u_2, \quad y_0 = y_1 + y_2$$

corresponding to the closed-loop feedback configuration as depicted in Figure 1; see section 1. By a solution of (3.1) we mean the following. For  $w_0 = (u_0, y_0) \in \mathcal{W} := \mathcal{U} \times \mathcal{Y}$ , a pair  $(w_1, w_2) = ((u_1, y_1), (u_2, y_2)) \in \mathcal{W}_a \times \mathcal{W}_a$ ,  $\mathcal{W}_a := \mathcal{U}_a \times \mathcal{Y}_a$ , is a *solution* of (3.1) if and only if (3.1) holds on  $\text{dom}(w_1, w_2)$ . The (possibly empty) set of all solutions is denoted by

$$\mathcal{X}_{w_0} := \{(w_1, w_2) \in \mathcal{W}_a \times \mathcal{W}_a \mid (w_1, w_2) \text{ solves (3.1)}\}.$$

The closed-loop system  $[P, C]$ , given by (3.1), is said to have

- the *existence property* if and only if  $\mathcal{X}_{w_0} \neq \emptyset$ ,
- the *uniqueness property* if and only if

$$\begin{aligned} \forall w_0 \in \mathcal{W} : \quad & \left[ (\hat{w}_1, \hat{w}_2), (\tilde{w}_1, \tilde{w}_2) \in \mathcal{X}_{w_0} \right. \\ & \left. \implies (\hat{w}_1, \hat{w}_2) = (\tilde{w}_1, \tilde{w}_2) \quad \text{on} \quad \text{dom}(\hat{w}_1, \hat{w}_2) \cap \text{dom}(\tilde{w}_1, \tilde{w}_2) \right]. \end{aligned}$$

Assume that  $[P, C]$  has the existence and uniqueness properties. For each  $w_0 \in \mathcal{W}$ , define  $\omega_{w_0}$ ,  $0 < \omega_{w_0} \leq \infty$ , by the property

$$[0, \omega_{w_0}) := \bigcup_{(\hat{w}_1, \hat{w}_2) \in \mathcal{X}_{w_0}} \text{dom}(\hat{w}_1, \hat{w}_2)$$

and define  $(w_1, w_2) \in \mathcal{W}_a \times \mathcal{W}_a$ , with  $\text{dom}(w_1, w_2) = [0, \omega_{w_0})$ , by the property  $(w_1, w_2)|_{[0, t)} \in \mathcal{X}_{w_0}$  for all  $t \in [0, \omega_{w_0})$ . This construction induces the operator

$$H_{P,C} : \mathcal{W} \rightarrow \mathcal{W}_a \times \mathcal{W}_a, \quad w_0 \mapsto (w_1, w_2).$$

The closed-loop system  $[P, C]$ , given by (3.1), is said to be

- *locally well posed* if and only if it has the existence and uniqueness properties and the operator  $H_{P,C}: \mathcal{W} \rightarrow \mathcal{W}_a \times \mathcal{W}_e$ ,  $w_0 \mapsto (w_1, w_2)$ , is causal,
- *globally well posed* if and only if it is locally well posed and  $H_{P,C}(\mathcal{W}) \subset \mathcal{W}_e \times \mathcal{W}_e$ ,
- $\mathcal{W}$ -*stable* if and only if it is locally well posed and  $H_{P,C}(\mathcal{W}) \subset \mathcal{W} \times \mathcal{W}$ ,
- *regularly well posed* if and only if it is locally well posed and

$$(3.2) \quad \forall w_0 \in \mathcal{W} : [\omega_{w_0} < \infty \implies T_{\omega_{w_0}} H_{P,C}(w_0) \notin \mathcal{W} \times \mathcal{W}].$$

If  $[P, C]$  is globally well posed, then for each  $w_0 \in \mathcal{W}$  the solution  $H_{P,C}(w_0)$  exists on the half-line  $\mathbb{R}_{\geq 0}$ . Regular well posedness means that if the closed-loop system has a finite escape time  $\omega_{w_0} > 0$  for some disturbance  $w_0 \in \mathcal{W}$ , then at least one of the components  $u_1$ ,  $u_2$  or  $y_1$ ,  $y_2$  is not a restriction to  $[0, \omega_{w_0})$  of a function in  $\mathcal{U}$  or  $\mathcal{Y}$ , respectively. If  $[P, C]$  is regularly well posed and satisfies

$$\forall w_0 \in \mathcal{W} : [\omega_{w_0} < \infty \implies T_{\omega_{w_0}} H_{P,C}(w_0) \in \mathcal{W} \times \mathcal{W}],$$

there does not exist a solution of  $[P, C]$  with a finite escape time, and therefore  $[P, C]$  is globally well posed. However, global well posedness does not guarantee that each solution belongs to  $\mathcal{W} \times \mathcal{W}$ ; the latter is ensured by  $\mathcal{W}$ -stability of  $[P, C]$ . Note also that neither regular nor global well posedness implies the other.

**3.3. Graphs and gain-function stability.** In our investigation of robustness of stability properties of a closed-loop system, the concept of graphs and gain-function stability will play a central role. Corresponding to a plant operator  $P$  (respectively, the controller operator  $C$ ) is a subset of  $\mathcal{W}$ , called the *graph* of the plant  $\mathcal{G}_P$  (respectively, the controller  $\mathcal{G}_C$ ), defined as

$$\mathcal{G}_P = \left\{ \begin{pmatrix} u \\ Pu \end{pmatrix} \middle| u \in \mathcal{U}, Pu \in \mathcal{Y} \right\} \subset \mathcal{W}, \quad \mathcal{G}_C = \left\{ \begin{pmatrix} Cy \\ y \end{pmatrix} \middle| Cy \in \mathcal{U}, y \in \mathcal{Y} \right\} \subset \mathcal{W}.$$

Note that we identify  $\mathcal{G}_P \ni \begin{pmatrix} u \\ Pu \end{pmatrix} = (u, Pu) \in \mathcal{W}$ , and analogously for  $\mathcal{G}_C$ .

A causal operator  $F: \mathcal{X} \rightarrow \mathcal{V}_a$ , where  $\mathcal{X}, \mathcal{V}$  are subsets of normed signal spaces, is said to be *gain-function stable* if and only if  $F(\mathcal{X}) \subset \mathcal{V}$  and the following nonlinear so-called *gain function* is well defined:

$$(3.3) \quad g[F]: (r_0, \infty) \rightarrow \mathbb{R}_{\geq 0},$$

$$r \mapsto g[F](r) = \sup \left\{ \|T_\tau Fx\|_{\mathcal{V}} \mid x \in \mathcal{X}, \|T_\tau x\|_{\mathcal{X}} \in (r_0, r], \tau > 0 \right\},$$

where  $r_0 := \inf_{x \in \mathcal{X}} \|x\|_{\mathcal{X}} < \infty$ . Observe that  $\|T_\tau Fx\|_{\mathcal{V}} \leq g[F](\|T_\tau x\|_{\mathcal{X}})$ . A closed-loop system  $[P, C]$  is said to be *gain-function stable* if and only if it is globally well posed and  $H_{P,C}: \mathcal{W} \rightarrow \mathcal{W}_e \times \mathcal{W}_e$  is gain-function stable.

Note the following facts:

- global well posedness of  $[P, C]$  implies that  $\text{im } H_{P,C} \subset \mathcal{W}_e \times \mathcal{W}_e$ ;
- gain-function stability of  $[P, C]$  implies  $\mathcal{W}$ -stability of  $[P, C]$ ;
- if  $[P, C]$  is  $\mathcal{W}$ -stable, then  $H_{P,C}: \mathcal{W} \rightarrow \mathcal{G}_P \times \mathcal{G}_C$  is a bijective operator with inverse  $H_{P,C}^{-1}: (w_1, w_2) \mapsto w_1 + w_2$ .

To see (iii), note that  $H_{P,C}(\mathcal{W}) \subset \mathcal{W} \times \mathcal{W}$  implies that  $H_{P,C}(\mathcal{W}) \subset \mathcal{G}_P \times \mathcal{G}_C$ , and since for any  $w_1 \in \mathcal{G}_P \subset \mathcal{W}$ ,  $w_2 \in \mathcal{G}_C \subset \mathcal{W}$  we have  $w_1 + w_2 \in \mathcal{W}$ , it follows that  $H_{P,C}(\mathcal{W}) \supset \mathcal{G}_P \times \mathcal{G}_C$ . Therefore, we can think of a gain-function stable  $H_{P,C}$  as a surjective operator  $H_{P,C}: \mathcal{W} \rightarrow \mathcal{G}_P \times \mathcal{G}_C$ . The inverse of  $H_{P,C}: \mathcal{W} \rightarrow \mathcal{G}_P \times \mathcal{G}_C$  is obviously  $H_{P,C}^{-1}: (w_1, w_2) \mapsto w_1 + w_2$ .

Finally, with a closed-loop system  $[P, C]$ , we associate the following two parallel projection operators:  $\Pi_{P//C}: \mathcal{W} \rightarrow \mathcal{W}_a$ ,  $w_0 \mapsto w_1$ , and  $\Pi_{C//P}: \mathcal{W} \rightarrow \mathcal{W}_a$ ,  $w_0 \mapsto w_2$ . Clearly,  $H_{P,C} = (\Pi_{P//C}, \Pi_{C//P})$  and  $\Pi_{P//C} + \Pi_{C//P} = I$ . Therefore, gain-function stability of one of the operators  $\Pi_{P//C}$  and  $\Pi_{C//P}$  implies the gain-function stability of the other, and so gain-function stability of either operator implies gain-function stability of the closed-loop system  $[P, C]$ .

**3.4. The nonlinear gap.** The essence of the paper is a study of robust stability in a specific adaptive control context. Robust stability is the property that the stability properties of a globally well-posed closed-loop system  $[P, C]$  persists under “sufficiently small” perturbations of the plant. In other words, robust stability is the property that  $[P_1, C]$  inherits the stability properties of  $[P, C]$  when the plant  $P$  is replaced by any plant  $P_1$  sufficiently “close” to  $P$ . In the context of this paper, plants  $P$  and  $P_1$  are deemed to be close if and only if their respective graphs are *close* in the gap sense of [5]. The nonlinear gap is defined as follows:

Let, for signal spaces  $\mathcal{U}$  and  $\mathcal{Y}$ ,

$$\Gamma := \{P: \mathcal{U}_a \rightarrow \mathcal{Y}_a \mid P \text{ is causal}\}$$

and, for  $P_1, P_2 \in \Gamma$ , define the (possibly empty) set

$$\mathcal{O}_{P_1, P_2} := \{\Phi: \mathcal{G}_{P_1} \rightarrow \mathcal{G}_{P_2} \mid \Phi \text{ is causal, surjective, and } \Phi(0) = 0\}.$$

The *directed nonlinear gap* is given by

$$\vec{\delta}: \Gamma \times \Gamma \rightarrow [0, \infty],$$

$$(P_1, P_2) \mapsto \vec{\delta}(P_1, P_2) := \inf_{\Phi \in \mathcal{O}_{P_1, P_2}} \sup_{x \in \mathcal{G}_{P_1} \setminus \{0\}, \tau > 0} \left( \frac{\|T_\tau(\Phi - I)(x)\|_{\mathcal{U} \times \mathcal{Y}}}{\|T_\tau x\|_{\mathcal{U} \times \mathcal{Y}}} \right),$$

with the convention that  $\vec{\delta}(P_1, P_2) := \infty$  if  $\mathcal{O}_{P_1, P_2} = \emptyset$ , and the *nonlinear gap*  $\delta$  is

$$\delta: \Gamma \times \Gamma \rightarrow [0, \infty], \quad (P_1, P_2) \mapsto \delta(P_1, P_2) := \max\{\vec{\delta}(P_1, P_2), \vec{\delta}(P_2, P_1)\}.$$

**3.5. Example.** In this subsection we illustrate the previous graph and gap concepts by two operators  $P_\alpha, P_{N,M,\alpha}$  induced by state space systems

$$(3.4) \quad \begin{cases} P_\alpha & : & \dot{x} = \alpha x + u_1, & x(0) = x^0, \\ & & y_1 = x, \end{cases}$$

$$(3.5) \quad \begin{cases} P_{N,M,\alpha} & : & \dot{x} = \tilde{A}x + \tilde{b}u_1, & x(0) = \tilde{x}^0, \\ & & y_1 = \tilde{c}x \end{cases}$$

for  $\alpha > 0$ ,  $x^0 \in \mathbb{R}$ , and  $(\tilde{A}, \tilde{b}, \tilde{c})$  as in (1.6), with  $\tilde{x}^0 \in \mathbb{R}^3$ . Throughout this example assume that  $x^0 = 0$ ,  $\tilde{x}^0 = 0$ . The second purpose of this example is to show that  $P_\alpha$  is close to  $P_{N,M,\alpha}$  in the sense that

$$(3.6) \quad \limsup_{M \rightarrow \infty} \vec{\delta}(P_\alpha, P_{2M,M,\alpha}) = 0.$$

First, recall that  $(\widetilde{A}, \widetilde{b}, \widetilde{c}) \in \mathcal{P}_{3,1} \setminus \widetilde{\mathcal{M}}_{3,1}$  and  $(\alpha, 1, 1) \in \mathcal{M}_{1,1}$ .

Second, recall that the graphs of  $P_\alpha$  and  $P_{N,M,\alpha}$  are given, respectively, by

$$\mathcal{G}_{P_\alpha} = \left\{ \begin{pmatrix} u_1 \\ y_1 \end{pmatrix} \middle| u_1, y_1 \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}), \, y_1 \text{ solves (3.4)} \right\},$$
$$\mathcal{G}_{P_{N,M,\alpha}} = \left\{ \begin{pmatrix} u_1 \\ y_1 \end{pmatrix} \middle| u_1, y_1 \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}), \, y_1 \text{ solves (3.5)} \right\}.$$

To determine an upper bound for the *gap* between  $P_\alpha$  and  $P_{N,M,\alpha}$ , consider the bijective mapping  $\Phi$  from graph  $\mathcal{G}_{P_\alpha}$  to graph  $\mathcal{G}_{P_{N,M,\alpha}}$  given by

$$\Phi : \mathcal{G}_{P_\alpha} \rightarrow \mathcal{G}_{P_{N,M,\alpha}}, \quad \begin{pmatrix} u \\ \int_0^\cdot e^{\alpha(\cdot-s)} u(s) \, ds \end{pmatrix} \mapsto \begin{pmatrix} u \\ \widetilde{c} \int_0^\cdot e^{\widetilde{A}(\cdot-s)} \widetilde{b} u(s) \, ds \end{pmatrix}.$$

By the definition of the nonlinear gap (see section 3.4), we obtain

$$\vec{\delta}(P_\alpha, P_{N,M,\alpha}) \leq \sup_{w \in \mathcal{G}_{P_\alpha} \setminus \{0\}} \frac{\|(\Phi - I)(w)\|_{\mathcal{W}}}{\|w\|_{\mathcal{W}}},$$

where  $\mathcal{W} := W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}) \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$  and, for  $w = (u, y) \in \mathcal{W}$ , the norm is defined by

$$\|(u, y)\|_{\mathcal{W}} := \|u\|_{W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})} + \|y\|_{W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})}.$$

To estimate

$$|(\Phi - I)(w)(t)| \quad \text{for } w := \begin{pmatrix} u \\ \int_0^\cdot e^{\alpha(\cdot-s)} u(s) \, ds \end{pmatrix} \in \mathcal{G}_{P_\alpha}$$

we calculate that the output  $y_1$  of (3.5) is given, for all  $t \geq 0$ , by

$$\begin{aligned} y_1(t) &= \widetilde{c} \int_0^t e^{\widetilde{A}(t-s)} \widetilde{b} u_1(s) \, ds = \int_0^t \frac{N(M - \alpha)}{(\alpha + N)(\alpha + M)} e^{\alpha(t-s)} u_1(s) \, ds \\ &\quad + \int_0^t \frac{N(N + M)}{(N - M)(\alpha + N)} e^{-N(t-s)} u_1(s) \, ds \\ &\quad + \int_0^t \frac{-2NM}{(N - M)(\alpha + M)} e^{-M(t-s)} u_1(s) \, ds, \end{aligned}$$

and thus, for all  $t \geq 0$ ,

$$\begin{aligned}
 |(\Phi - I)(w)(t)| &\leq \left| \left( \frac{N(M-\alpha)}{(\alpha+N)(\alpha+M)} - 1 \right) \int_0^t e^{\alpha(t-s)} u(s) \, ds \right| \\
 &\quad + \left| \frac{N(N+M)}{(N-M)(\alpha+N)} \int_0^t e^{-N(t-s)} u(s) \, ds \right| \\
 &\quad + \left| \frac{-2NM}{(N-M)(\alpha+M)} \int_0^t e^{-M(t-s)} u(s) \, ds \right| \\
 &\leq \left| \frac{N(M-\alpha)}{(\alpha+N)(\alpha+M)} - 1 \right| \left| \int_0^t e^{\alpha(t-s)} u(s) \, ds \right| \\
 &\quad + \left( \left| \frac{N(N+M)}{(N-M)(\alpha+N)} \int_0^t e^{-N(t-s)} \, ds \right| \right. \\
 &\quad \left. + \left| \frac{-2NM}{(N-M)(\alpha+M)} \int_0^t e^{-M(t-s)} \, ds \right| \right) \|u\|_{L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})} \\
 &\leq \left| \frac{N(M-\alpha)}{(\alpha+N)(\alpha+M)} - 1 \right| \left\| \int_0^\cdot e^{\alpha(\cdot-s)} u(s) \, ds \right\|_{W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})} \\
 &\quad + \left( \left| \frac{N+M}{(N-M)(\alpha+N)} \right| + \left| \frac{2N}{(N-M)(\alpha+M)} \right| \right) \|u\|_{W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})}.
 \end{aligned}$$

Hence

$$\vec{\delta}(P_\alpha, P_{N,M,\alpha}) \leq \left| \frac{N(M-\alpha)}{(\alpha+N)(\alpha+M)} - 1 \right| + \left| \frac{N+M}{(N-M)(\alpha+N)} \right| + \left| \frac{2N}{(N-M)(\alpha+M)} \right|,$$

which yields (3.6).

#### 4. Robustness of the $\lambda$ -tracker.

**4.1. Well posedness of the closed-loop system.** For  $m, n \in \mathbb{N}$  with  $n \geq m$ , consider  $\mathcal{P}_{n,m}$  as a subspace of the Euclidean space  $\mathbb{R}^{n^2+2mn}$  by identifying a plant  $\theta = (A, B, C)$  with a vector  $\theta$  consisting of the elements of the plant matrices, ordered lexicographically. With normed signal spaces  $\mathcal{U}$  and  $\mathcal{Y}$  and  $(\theta, x^0) \in \mathcal{P}_{n,m} \times \mathbb{R}^n$ , where  $x^0 \in \mathbb{R}^n$  is the initial value of a linear system (1.1), we associate the causal plant operator

$$(4.1) \quad \tilde{P}(\theta, x^0) : \mathcal{U}_a \rightarrow \mathcal{Y}_a, \quad u_1 \mapsto \tilde{P}(\theta, x^0)(u_1) := y_1,$$

where, for  $u_1 \in \mathcal{U}_a$  with  $\text{dom}(u_1) = [0, \omega)$ , we have  $y_1 = cx$ ,  $x$  being the unique solution of (1.1) on  $[0, \omega)$ . Note that  $\tilde{P}$  is a map from  $\cup_{n \geq m} (\mathcal{P}_{n,m} \times \mathbb{R}^n)$  to the space of maps  $\mathcal{U}_a \rightarrow \mathcal{Y}_a$ . Consider, for  $\lambda > 0$ , the adaptive strategy (1.5) and associate the causal control operator, parameterized by  $\lambda$  and the initial value  $k^0 \in \mathbb{R}$ , i.e.,

$$(4.2) \quad \tilde{C}(\lambda, k^0) : \mathcal{Y}_a \rightarrow \mathcal{U}_a, \quad y_2 \mapsto \tilde{C}(\lambda, k^0)(y_2) := u_2.$$

Note that  $\tilde{C}$  is a map from  $\mathbb{R}_{>0} \times \mathbb{R}$  to the space of causal maps  $\mathcal{Y}_a \rightarrow \mathcal{U}_a$ .

In this subsection we show that, for  $\mathcal{U} = \mathcal{Y} = W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ , the closed-loop system  $[\tilde{P}(\theta, x^0), \tilde{C}(\lambda, k^0)]$  of any plant of the form (1.1) (with associated operator  $\tilde{P}(\theta, x^0)$ ) and adaptive controller (1.5) (with associated operator  $\tilde{C}(\lambda, k^0)$ ), where  $(\theta, x^0) \in \mathcal{P}_{n,m} \times \mathbb{R}^n$  and  $(\lambda, k^0) \in \mathbb{R}_{>0} \times \mathbb{R}$ , is regularly well posed. Furthermore, we show that, for  $\theta \in \mathcal{M}_{n,m}$ , the closed-loop system  $[\tilde{P}(\theta, x^0), \tilde{C}(\lambda, k^0)]$  is globally well posed and  $(\mathcal{U} \times \mathcal{Y})$ -stable.

**PROPOSITION 4.1.** *Let  $m, n \in \mathbb{N}$  with  $n \geq m$ ,  $\lambda > 0$ ,  $(\theta, x^0, k^0) \in \mathcal{M}_{n,m} \times \mathbb{R}^n \times \mathbb{R}$ , and  $u_0, y_0 \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ . Then, for plant operator  $\tilde{P}(\theta, x^0)$  and control operator  $\tilde{C}(\lambda, k^0)$ , given by (4.1) and (4.2), respectively, the closed-loop initial value problem  $[\tilde{P}(\theta, x^0), \tilde{C}(\lambda, k^0)]$ , given by (1.8), (1.2), (1.5), is globally well posed and  $(W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m) \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m))$ -stable.*

*Proof.* The proposition is a direct consequence of Proposition 2.1.  $\square$

Note that, for  $(A, B, C) \in \mathcal{P}_{n,m}$ ,  $x^0 \in \mathbb{R}^n$ ,  $\lambda > 0$ , and  $k^0 \in \mathbb{R}$ , the closed-loop initial value problem (1.1), (1.2), (1.5) may be written as

$$(4.3) \quad \begin{cases} \dot{x}(t) &= Ax(t) + B[u_0(t) - u_2(t)], & x(0) = x^0 \in \mathbb{R}^n, \\ \dot{k}(t) &= d_\lambda(y_2(t)) |y_2(t)|, & k(0) = k^0 \in \mathbb{R}, \\ y_2(t) &= y_0(t) - Cx(t), \\ u_2(t) &= -k(t)y_2(t), \end{cases}$$

where  $d_\lambda$  is defined in the Nomenclature.

**PROPOSITION 4.2.** *Let  $m, n \in \mathbb{N}$  with  $n \geq m$ ,  $\lambda > 0$ ,  $(\theta, x^0, k^0) \in \mathcal{P}_{n,m} \times \mathbb{R}^n \times \mathbb{R}$ , and  $u_0, y_0 \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ . Then, for plant operator  $\tilde{P}(\theta, x^0)$  and control operator  $\tilde{C}(\lambda, k^0)$ , given by (4.1) and (4.2), respectively, the closed-loop initial value problem  $[\tilde{P}(\theta, x^0), \tilde{C}(\lambda, k^0)]$ , given by (4.3), has the following properties:*

- (i) *there exists a unique maximal solution  $(x, k) : [0, \omega) \rightarrow \mathbb{R}^n \times \mathbb{R}$  for some  $\omega \in (0, \infty]$ ;*
- (ii) *if  $k \in W^{1,\infty}([0, \omega) \rightarrow \mathbb{R})$ , then  $\omega = \infty$ ;*
- (iii) *if  $y_2 \in W^{1,\infty}([0, \omega) \rightarrow \mathbb{R}^m)$ , then  $\omega = \infty$ ;*
- (iv)  *$[\tilde{P}(\theta, x^0), \tilde{C}(\lambda, k^0)]$  is regularly well posed.*

*Proof.* (i) Since the right-hand side of (4.3) is continuous and locally Lipschitz, the statement follows from the theory of ordinary differential equations.

(ii) Suppose  $k \in W^{1,\infty}([0, \omega) \rightarrow \mathbb{R})$  and, for contradiction,  $\omega < \infty$ . Since  $d_\lambda(y_2)^2 \leq d_\lambda(y_2) |y_2| = \dot{k} \in L^\infty([0, \omega) \rightarrow \mathbb{R}_{\geq 0})$ , we have  $d_\lambda(y_2) \in L^\infty([0, \omega) \rightarrow \mathbb{R}_{\geq 0})$  and  $d_\lambda(y_2) + \lambda \in L^\infty([0, \omega) \rightarrow \mathbb{R}_{\geq 0})$ . Thus  $y_2 \in L^\infty([0, \omega) \rightarrow \mathbb{R}^m)$ .

Since  $k \in L^\infty([0, \omega) \rightarrow \mathbb{R})$ , variation of constants applied to (4.3) yields the existence of constants  $c_0, c_1 > 0$  such that

$$(4.4) \quad \forall t \in [0, \omega) : |x(t)| \leq c_0 \left( e^{c_1 \omega} + \int_0^\omega e^{c_1(\omega-s)} (|u_0(s)| + |y_2(s)|) ds \right).$$

Since  $y_2 \in L^\infty([0, \omega) \rightarrow \mathbb{R}^m)$  and  $u_0 \in L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ , it follows from the convolution in (4.4) that the right-hand side of (4.4) is bounded on  $[0, \omega)$ , which contradicts the maximality of the solution  $x$ . Hence  $\omega = \infty$ .

(iii) Suppose  $y_2 \in W^{1,\infty}([0, \omega) \rightarrow \mathbb{R}^m)$  and, for contradiction,  $\omega < \infty$ . Then  $\dot{k} = d_\lambda(y_2) |y_2| \in L^\infty([0, \omega) \rightarrow \mathbb{R})$ , and, combined with

$$\forall t \in [0, \omega) :$$

$$k(t) - k^0 = \int_0^t d_\lambda(y_2(s)) |y_2(s)| \, ds \leq \int_0^t \|y_2\|_{L^\infty([0, \omega) \rightarrow \mathbb{R}^m)}^2 \, ds = \omega \|y_2\|_{L^\infty([0, \omega) \rightarrow \mathbb{R}^m)}^2,$$

we arrive at  $k \in W^{1,\infty}([0, \omega) \rightarrow \mathbb{R})$ . Now (ii) yields that  $\omega = \infty$ . This is a contradiction, and so  $\omega = \infty$ .

(iv) Let  $\mathcal{W} = W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m) \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ . By (i), the closed-loop  $[\tilde{P}(\theta, x^0), \tilde{C}(\lambda, k^0)]$  is locally well posed. To prove that  $[\tilde{P}(\theta, x^0), \tilde{C}(\lambda, k^0)]$  is regularly well posed, it suffices to show that (3.2) holds. For arbitrary  $w_0 = (u_0, y_0) \in \mathcal{W}$  consider  $(w_1, w_2) = H_{\tilde{P}(\theta, x^0), \tilde{C}(\lambda, k^0)}(w_0)$ , where  $\text{dom}(w_1, w_2) = [0, \omega)$  is maximal. Suppose, contrary to the right-hand side of (3.2), that  $T_\omega(w_1, w_2) \in \mathcal{W} \times \mathcal{W}$ . Then  $y_2 \in W^{1,\infty}([0, \omega) \rightarrow \mathbb{R}^m)$ , which, in view of (iii), yields  $\omega = \infty$ , i.e., the contrary of the left-hand side of (3.2). Hence the closed-loop system is regularly well posed.  $\square$

**4.2. Robustness.** In Propositions 4.1 and 4.2 we have established that, for  $(\theta, x^0, k^0) \in \mathcal{M}_{n,m} \times \mathbb{R}^n \times \mathbb{R}$  and  $m, n \in \mathbb{N}$  with  $n \geq m$ ,  $\lambda > 0$ ,  $u_0, y_0 \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ , the closed-loop system  $[\tilde{P}(\theta, x^0), \tilde{C}(\lambda, k^0)]$  is globally well posed and has certain stability properties. Furthermore, in Proposition 2.1  $\lambda$ -tracking is shown for linear systems belonging to class  $\mathcal{M}_{n,m}$ .

The purpose of this subsection is to determine conditions under which these properties are maintained when the plant  $\tilde{P}(\theta, x^0)$  is perturbed to a plant  $\tilde{P}(\tilde{\theta}, \tilde{x}^0)$ , where  $(\tilde{\theta}, \tilde{x}^0) \in \mathcal{P}_{q,m} \times \mathbb{R}^q$  for some  $q \in \mathbb{N}$ , in particular when  $\tilde{\theta} \notin \mathcal{M}_{q,m}$ . The main result, Theorem 4.5, shows that stability properties and  $\lambda$ -tracking persist if (a) the plants  $\tilde{P}(\tilde{\theta}, 0)$  and  $\tilde{P}(\theta, 0)$  are sufficiently close (in the gap sense) and (b) the initial data  $\tilde{x}^0$  and disturbance  $w_0 = (u_0, y_0)$  are sufficiently small.

To establish gap margin results, we will need to construct the augmented plant and controller operators as in [4]. Note that  $0 \notin \mathcal{M}_{n,m}$ . Define  $\tilde{\mathcal{U}} := \mathbb{R}^{n^2+2n} \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$  and let  $\tilde{\mathcal{W}} := \tilde{\mathcal{U}} \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ , which can be considered as signal spaces by identifying  $\theta \in \mathbb{R}^{n^2+2mn}$  with the constant function  $t \mapsto \theta$  and endowing  $\tilde{\mathcal{U}}$  with the norm  $\|(\theta, u)\|_{\tilde{\mathcal{U}}} := \sqrt{|\theta|^2 + \|u\|_{W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)}^2}$ . For given  $\tilde{P}(\theta, 0)$  as in (4.1), we define the (augmented) plant operator as

$$(4.5) \quad P : \tilde{\mathcal{U}}_a \rightarrow W_a^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m), \quad (\theta, u_1) = \tilde{u}_1 \mapsto y_1 = P(\tilde{u}_1) := \tilde{P}(\theta, 0)(u_1).$$

Fix  $\lambda > 0$  and  $k^0 \in \mathbb{R}$  and define, for  $\tilde{C}(\lambda, k^0)$  as in (4.2), the (augmented) controller operator as

$$(4.6) \quad C : W_a^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m) \rightarrow \tilde{\mathcal{U}}_a, \quad y_2 \mapsto \tilde{u}_2 = C(y_2) := \left(0, \tilde{C}(\lambda, k^0)(y_2)\right) = (0, u_2).$$

For each nonempty  $\Omega \subset \mathcal{M}_{n,m}$ , define

$$(4.7) \quad \mathcal{W}^\Omega := (\Omega \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)) \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m) \quad \text{and} \quad H_{P,C}^\Omega := H_{P,C}|_{\mathcal{W}^\Omega}.$$

It follows from Proposition 4.1 that  $H_{P,C}^\Omega : \mathcal{W}^\Omega \rightarrow \widetilde{\mathcal{W}} \times \widetilde{\mathcal{W}}$  is a causal operator for any  $\Omega \subset \mathcal{M}_{n,m}$ . In Proposition 4.3 we show gain-function stability of  $H_{P,C}^\Omega$ . This is a supposition of Theorem 5.2 in [3], the latter being used to show Proposition 4.4 and thus the main result, Theorem 4.5.

**PROPOSITION 4.3.** *Let  $m, n \in \mathbb{N}$  with  $n \geq m$ ,  $k^0 \in \mathbb{R}$ , and  $\lambda > 0$  and assume  $\Omega \subset \mathcal{M}_{n,m}$  is closed. Then, for the closed-loop system  $[P, C]$  given by (3.1), (4.5), and (4.6), the operator  $H_{P,C}^\Omega$  given by (4.7) is gain-function stable.*

*Proof.* Note that  $((\theta, u_1), y_1) = ((\theta, u_0), y_0) - ((0, u_2), y_2)$ . For  $\nu : \mathcal{D}_{n,m} \rightarrow \mathbb{R}_{\geq 0}$  as in Proposition 2.1 and  $\mathcal{W}^\Omega$  given by (4.7), we have

$$\forall ((\theta, u_0), y_0) \in \mathcal{W}^\Omega :$$

$$\begin{aligned} \|H_{P,C}^\Omega((\theta, u_0), y_0)\|_{\widetilde{\mathcal{W}} \times \widetilde{\mathcal{W}}} &= \|((\theta, u_1), y_1), ((0, u_2), y_2))\|_{\widetilde{\mathcal{W}} \times \widetilde{\mathcal{W}}} \\ &\leq \|((\theta, u_0), y_0)\|_{\widetilde{\mathcal{W}}} + 2\|((0, u_2), y_2)\|_{\widetilde{\mathcal{W}}} \\ &\leq \|(u_0, y_0)\|_{\mathcal{W}} + |\theta| + 2\nu(\theta, (0, k^0), u_0, y_0), \end{aligned}$$

and so, for  $r_0 := \inf_{w \in \mathcal{W}^\Omega} \|w\|_{\widetilde{\mathcal{W}}}$  and  $r \in (r_0, \infty)$ , closedness of  $\Omega$  yields

$$\begin{aligned} &g[H_{P,C}^\Omega](r) \\ &:= \sup \left\{ \|(u_0, y_0)\|_{\mathcal{W}} + |\theta| + 2\nu(\theta, (0, k^0), u_0, y_0) \mid \begin{array}{l} (\theta, u_0, y_0) \in \mathcal{W}^\Omega, \\ \|(\theta, u_0, y_0)\|_{\widetilde{\mathcal{W}}} \leq r \end{array} \right\} < \infty. \end{aligned}$$

Thus, a gain function for  $H_{P,C}^\Omega$  exists, and the proof is complete.  $\square$

The following proposition establishes  $(W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m) \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m))$ -stability of the closed-loop system  $[\tilde{P}(\tilde{\theta}, \tilde{x}^0), \tilde{C}(\lambda, k^0)]$  for a system  $\tilde{\theta}$  belonging to the system class  $\mathcal{P}_{q,m}$  if, for a system  $\theta$  belonging to  $\mathcal{M}_{n,m}$ , the gap between  $\tilde{P}(\tilde{\theta}, 0)$  and  $\tilde{P}(\theta, 0)$ , the initial value  $\tilde{x}^0 \in \mathbb{R}^q$  and the input/output disturbances  $w_0 = (u_0, y_0)$  are sufficiently small. The proof is based on results from [3].

**PROPOSITION 4.4.** *Let  $m, n, q \in \mathbb{N}$  with  $n, q \geq m$ ,  $\mathcal{U} = \mathcal{Y} = W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ ,  $\mathcal{W} = \mathcal{U} \times \mathcal{Y}$ , and  $\theta \in \mathcal{M}_{n,m}$ . For  $(\tilde{\theta}, \tilde{x}^0, k^0) \in \mathcal{P}_{q,m} \times \mathbb{R}^q \times \mathbb{R}$  and  $\lambda > 0$ , consider  $\tilde{P}(\tilde{\theta}, \tilde{x}^0) : \mathcal{U}_a \rightarrow \mathcal{Y}_a$  and  $\tilde{C}(\lambda, k^0) : \mathcal{Y}_a \rightarrow \mathcal{U}_a$  defined by (4.1) and (4.2), respectively. Then there exist a continuous function  $\eta : (0, \infty) \rightarrow (0, \infty)$  and a function  $\psi : \mathcal{P}_{q,m} \rightarrow (0, \infty)$  such that the following holds:*

$$(4.8) \quad \forall (\tilde{\theta}, \tilde{x}^0, w_0, r) \in \mathcal{P}_{q,m} \times \mathbb{R}^q \times \mathcal{W} \times (0, \infty) :$$

$$\left. \begin{array}{l} \psi(\tilde{\theta})|\tilde{x}^0| + \|w_0\|_{\mathcal{W}} \leq r, \\ \tilde{\delta}(\tilde{P}(\tilde{\theta}, 0), \tilde{P}(\tilde{\theta}, 0)) \leq \eta(r) \end{array} \right\} \implies H_{\tilde{P}(\tilde{\theta}, \tilde{x}^0), \tilde{C}(\lambda, k^0)}(w_0) \in \mathcal{W} \times \mathcal{W}.$$

*Proof.* We need to show how the gain-function stability of the augmented closed loop  $[P, C]$ , given by (4.5) and (4.6), yields the robustness property (4.8) for the unaugmented closed-loop system  $[\tilde{P}(\tilde{\theta}, \tilde{x}^0), \tilde{C}(\lambda, k^0)]$ .

By Proposition 4.2 the closed-loop system  $[\tilde{P}(\tilde{\theta}, \tilde{x}^0), \tilde{C}(\lambda, k^0)]$  is regularly well posed for all  $\tilde{\theta} \in \mathcal{P}_{q,m}$ . Consider the augmented operators defined by (4.5) and (4.6),



i.e.,

$$\begin{aligned} P: \mathcal{P}_{n,m} \times \mathcal{U}_a &\rightarrow \mathcal{Y}_a, & (\tilde{\theta}, u_1) &\mapsto P(\tilde{\theta}, u_1) = \tilde{P}(\tilde{\theta}, 0)(u_1), \\ C: \mathcal{Y}_a &\rightarrow \mathcal{P}_{n,m} \times \mathcal{U}_a, & y_2 &\mapsto C(y_2) = (0, \tilde{C}(\lambda, k^0)(y_2)). \end{aligned}$$

For  $\theta \in \mathcal{M}_{n,m}$  set  $\Omega = \{\theta\}$ . By Proposition 4.3,  $H_{\tilde{P},C}^\Omega = H_{P,C}|_{\mathcal{W}^\Omega}$ , given by (4.7), is gain-function stable. By, for example, the proof of Theorem 4.D in [18],  $T_\tau \Pi_{\tilde{P}(\theta,0)/\tilde{C}(\lambda,k^0)}$  is continuous for all  $\tau > 0$ , and so  $T_\tau \Pi_{P/C}|_{\mathcal{W}^\Omega}$  is continuous for all  $\tau > 0$ .

Then [3, Theorem 5.2] gives the existence of a continuous function  $\mu: (0, \infty) \times \Omega \rightarrow (0, \infty)$  such that

$$\forall (\theta, \tilde{\theta}, w_0, r) \in \Omega \times \mathcal{P}_{q,m} \times \mathcal{W} \times (0, \infty) :$$

$$\left[ \|w_0\|_{\mathcal{W}} \leq r \wedge \delta \left( \tilde{P}(\theta, 0), \tilde{P}(\tilde{\theta}, 0) \right) \leq \mu(r, \theta) \right] \implies H_{\tilde{P}(\tilde{\theta},0), \tilde{C}(\lambda,k^0)}(w_0) \in \mathcal{W} \times \mathcal{W}.$$

Note that the proof of [3, Theorem 5.2] holds also for the signal space  $W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ , although it is proved in [3] for  $\mathcal{U} = \mathcal{Y} = L^p(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ ,  $1 \leq p \leq \infty$ .

To prove (4.8) we will use [3, Theorem 5.3]. The statement of [3, Theorem 5.3] has been proved for  $\mathcal{U} = \mathcal{Y} = L^p(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ ,  $1 \leq p \leq \infty$ . The statement holds also for  $\mathcal{U} = \mathcal{Y} = W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ . To see this, invoke the fact that for any Hurwitz matrix  $M \in \mathbb{R}^{n \times n}$  it is  $(t \mapsto \exp(Mt))$ ,  $(t \mapsto \frac{d}{dt} \exp(Mt)) \in L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{n \times n})$ . Now the statement of [3, Theorem 5.3] for  $\mathcal{U} = \mathcal{Y} = W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$  yields the existence of a continuous function  $\mu: (0, \infty) \times \Omega \rightarrow (0, \infty)$  and a function  $\psi: \mathcal{P}_{q,m} \rightarrow (0, \infty)$  such that

$$(4.9) \quad \forall (\tilde{\theta}, \theta, \tilde{x}^0, w_0, r) \in \mathcal{P}_{q,m} \times \mathcal{M}_{n,m} \times \mathbb{R}^q \times \mathcal{W} \times (0, \infty) :$$

$$\left. \begin{aligned} &\psi(\tilde{\theta})|\tilde{x}^0| + \|w_0\|_{\mathcal{W}} \leq r, \\ &\delta \left( \tilde{P}(\theta, 0), \tilde{P}(\tilde{\theta}, 0) \right) \leq \mu(r, \theta) \end{aligned} \right\} \implies H_{\tilde{P}(\tilde{\theta},\tilde{x}^0), \tilde{C}(\lambda,k^0)}(w_0) \in \mathcal{W} \times \mathcal{W}.$$

Finally, statement (4.8) follows on setting  $\eta(\cdot) = \mu(\cdot, \theta)$ .  $\square$

Note that [3, Theorem 5.3] requires stabilizability of system  $\tilde{\theta} \in \mathcal{P}_{q,m}$ .

Finally, we are in the position to state and prove the main result of the present paper. Loosely speaking, we show that the  $\lambda$ -tracker also works for systems  $(\tilde{A}, \tilde{B}, \tilde{C}) \in \mathcal{P}_{q,m}$  which are not necessarily minimum phase, may have higher relative degree, and may have negative high-frequency gain. However,  $(\tilde{A}, \tilde{B}, \tilde{C})$  has to be sufficiently close—in the terms of the gap metric—to a system  $(A, B, C) \in \tilde{\mathcal{M}}_{n,m}$ , and the initial value  $\tilde{x}^0 \in \mathbb{R}^q$  for  $(\tilde{A}, \tilde{B}, \tilde{C})$  and the input/output disturbances  $(u_0, y_0)$  have to be sufficiently small.

**THEOREM 4.5.** *Let  $m, n, q \in \mathbb{N}$  with  $n, q \geq m$ ,  $\mathcal{U} = \mathcal{Y} = W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ ,  $\mathcal{W} = \mathcal{U} \times \mathcal{Y}$ ,  $k^0 \in \mathbb{R}$ ,  $\lambda > 0$ , and  $\theta \in \mathcal{M}_{n,m}$ . For  $(\tilde{\theta}, \tilde{x}^0) \in \mathcal{P}_{q,m} \times \mathbb{R}^q$  consider the associated operators  $\tilde{P}(\tilde{\theta}, \tilde{x}^0): \mathcal{U}_a \rightarrow \mathcal{Y}_a$  and  $\tilde{C}(\lambda, k^0): \mathcal{Y}_a \rightarrow \mathcal{U}_a$  defined by (4.1) and (4.2), respectively, and the closed-loop initial value problem (1.1), (1.2), (1.5). Then there exist a continuous function  $\eta: (0, \infty) \rightarrow (0, \infty)$  and a function  $\psi: \mathcal{P}_{q,m} \rightarrow (0, \infty)$*

such that the following holds:

$$(4.10) \quad \forall \left( \tilde{\theta}, \tilde{x}^0, w_0, r \right) \in \mathcal{P}_{q,m} \times \mathbb{R}^q \times \mathcal{W} \times (0, \infty) :$$

$$\left. \begin{array}{l} \psi(\tilde{\theta})|\tilde{x}^0| + \|w_0\|_{\mathcal{W}} \leq r, \\ \vec{\delta} \left( \tilde{P}(\tilde{\theta}, 0), \tilde{P}(\tilde{\theta}, 0) \right) \leq \eta(r) \end{array} \right\} \implies \begin{cases} \limsup_{t \rightarrow \infty} |y_2(t)| \leq \lambda, \\ k \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}), \\ x \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^q), \end{cases}$$

where  $(x, k)$  and  $y_2$  satisfy (4.3).

*Proof. Step 1.* We show that

$$(4.11) \quad ((u_1, y_1), (u_2, y_2)) = H_{\tilde{P}(\tilde{\theta}, \tilde{x}^0), \tilde{C}(\lambda, k^0)}(w_0) \in \mathcal{W} \times \mathcal{W}.$$

Choose functions  $\eta: (0, \infty) \rightarrow (0, \infty)$  and  $\psi: \mathcal{P}_{q,m} \rightarrow (0, \infty)$  from Proposition 4.4. Let

$$\left( \tilde{\theta}, \tilde{x}^0, w_0, r \right) \in \mathcal{P}_{q,m} \times \mathbb{R}^q \times \mathcal{W} \times (0, \infty) :$$

$$\psi(\tilde{\theta})|\tilde{x}^0| + \|w_0\|_{\mathcal{W}} \leq r \wedge \vec{\delta} \left( \tilde{P}(\tilde{\theta}, 0), \tilde{P}(\tilde{\theta}, 0) \right) \leq \eta(r).$$

Then Proposition 4.4 gives (4.11).

*Step 2.* By Proposition 4.2 it follows that (4.3) has a unique solution

$$(x, k): [0, \omega) \rightarrow \mathbb{R}^q \times \mathbb{R}$$

on a maximal interval of existence  $[0, \omega)$  for some  $\omega \in (0, \infty]$ . Proposition 4.2(iii) yields  $\omega = \infty$ .

*Step 3.* We show that  $\dot{k} \in L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$ . Suppose, for contradiction, that  $\dot{k} \notin L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$ ; i.e., there exists a sequence  $(t_i) \in (\mathbb{R}_{\geq 0})^\mathbb{N}$  with  $t_i > t_{i+1}$  and  $\lim_{i \rightarrow \infty} \dot{k}(t_i) = \infty$ . Then

$$\lim_{i \rightarrow \infty} d_\lambda(y_2(t_i)) |y_2(t_i)| = \infty$$

and thus

$$\lim_{i \rightarrow \infty} |y_2(t_i)| = \infty,$$

a contradiction to Step 1.

*Step 4.* We show that  $k \in L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$ . Suppose, for contradiction, that  $k \notin L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$ ; i.e.,  $\lim_{t \rightarrow \infty} k(t) = \infty$ . Since  $u_2 \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ , the fourth equation in (4.3) yields  $\lim_{t \rightarrow \infty} y_2(t) = 0$ , and thus

$$\exists T > 0 \quad \forall t \geq T : \dot{k}(t) = d_\lambda(y_2(t)) |y_2(t)| = 0,$$

which contradicts the assumption on  $k$ .

*Step 5.* By Steps 3 and 4 we obtain  $k \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$ .

*Step 6.* By Proposition 4.4 we have in particular  $y_2, \dot{y}_2 \in L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$ . Similar to Step 10 of the proof of Proposition 2.1, we may establish that  $y_2$  is uniformly continuous.

*Step 7.* By Step 6 and continuity of  $e \mapsto d_\lambda(e)|e|$  we obtain that  $t \mapsto d_\lambda(y_2(t)) |y_2(t)|$  is uniformly continuous. Hence, in view of  $\dot{k} = d_\lambda(y_2) |y_2| \in L^1(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$

$\mathbb{R}$ ), which is equivalent to  $k \in L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$ , and Barbălat's lemma (see [1]),  $\lim_{t \rightarrow \infty} d_\lambda(y_2(t))|y_2(t)| = 0$  holds. This gives  $\limsup_{t \rightarrow \infty} |y_2(t)| \leq \lambda$ .

*Step 8.* It remains to show that  $x \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^q)$ . Let  $(\tilde{A}, \tilde{B}, \tilde{C}) \in \mathcal{P}_{q,m}$  associated with (1.1). Detectability of  $(\tilde{A}, \tilde{B}, \tilde{C})$  yields the existence of  $F \in \mathbb{R}^{q \times m}$  such that  $\text{spec}(\tilde{A} + F\tilde{C}) \subset \mathbb{C}_-$ . Setting  $g := -[F + k\tilde{B}](y_0 - y_2) + \tilde{B}u_0 + \tilde{B}ky_0$  gives

$$(4.12) \quad \dot{x} = [\tilde{A} - k\tilde{B}\tilde{C}]x + \tilde{B}u_0 + \tilde{B}ky_0 = [\tilde{A} + F\tilde{C}]x + g.$$

By Proposition 4.4 and Step 5 we have  $y_2 \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$  and  $k \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$ , and since  $w_0 = (u_0, y_0) \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m) \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m)$  it follows that  $g \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^q)$ . Hence, by (4.12) we obtain  $x \in L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^q)$ . The first equation in (4.3) then gives  $\dot{x} \in L^\infty(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^q)$ , which shows that  $x \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^q)$ , and the proof is complete.  $\square$

*Example 4.6.* Finally, we revisit example (1.6).

In subsection 3.5 we have already shown that for zero initial conditions the gap between the system  $(\tilde{A}, \tilde{b}, \tilde{c}) \in \mathcal{P}_{3,1} \setminus \mathcal{M}_{3,1}$  and  $(\alpha, 1, 1) \in \mathcal{M}_{1,1}$  tends to zero as  $N = 2M$  and  $M$  tends to infinity; see (3.6). Now, in view of Theorem 4.5, there exist a continuous function  $\eta: (0, \infty) \rightarrow (0, \infty)$  and a function  $\psi: \mathcal{P}_{3,1} \rightarrow (0, \infty)$  such that

$$\forall (\tilde{x}^0, w_0, r) \in \mathbb{R}^3 \times \mathcal{W} \times (0, \infty) :$$

$$\left. \begin{aligned} &\psi((\tilde{A}, \tilde{b}, \tilde{c}))|\tilde{x}^0| + \|w_0\|_{\mathcal{W}} \leq r, \\ &\delta\left(\tilde{P}_1((\alpha, 1, 1), 0), \tilde{P}_2((\tilde{A}, \tilde{b}, \tilde{c}), 0)\right) \leq \eta(r) \end{aligned} \right\} \implies \begin{cases} k \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}), \\ \limsup_{t \rightarrow \infty} |y_0(t) - y_1(t)| \leq \lambda, \\ x \in W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^3), \end{cases}$$

where  $\mathcal{W} = W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R}) \times W^{1,\infty}(\mathbb{R}_{\geq 0} \rightarrow \mathbb{R})$ .

This means in particular that  $\lambda$ -tracking is achieved by the adaptive control strategy (1.5) applied to system (1.6) despite the fact that it has unstable zero dynamics, has relative degree two, and has negative high-frequency gain. The only restrictions are that the zero is “far” in the right half complex plane, the initial condition  $\tilde{x}^0$  is “small,” and the  $W^{1,\infty}$  input/output disturbances  $u_0$  and  $y_0$  are “small,” too.

**5. Conclusions.** We have shown the robustness of the  $\lambda$ -tracker (1.5) for a class of linear systems close in the gap metric to minimum phase systems with strict relative degree one; moreover, the  $\lambda$ -tracker copes with certain bounded input/output disturbances. Although this result may be worth knowing, the shortcoming of the control strategy (1.5) is that the gain  $k(\cdot)$  is a monotone nondecreasing function which converges to a potentially “too high” gain, thus amplifying noise. Recently, a simple time-varying proportional output feedback law called “funnel control” has been introduced to achieve “practical” tracking with prespecified transient behavior [11]; this control law is applicable even to nonlinear minimum phase systems, and the gain is no longer monotone but may decrease again. The conceptual results of the present paper may indicate how to use the gap metric framework to show the robustness of the funnel controller.

## REFERENCES

- [1] I. BARBĂLAT, *Systèmes d'équations différentielles d'oscillations nonlinéaires*, Rev. Math. Pures Appl., 4 (1959), pp. 267–270.

- [2] E. BULLINGER, C. W. FREI, T. J. SIEBER, A. H. GLATTFELDER, F. ALLGÖWER, AND A. M. ZBINDEN, *Adaptive  $\lambda$ -tracking in anesthesia*, in Proceedings of the 4th IFAC Symposium on Modelling and Control in Biomedical Systems, E. Carson and E. Salzsieder, eds., Pergamon, Oxford, UK, 2000, pp. 181–186.
- [3] M. FRENCH, *Adaptive control and robustness in the gap metric*, IEEE Trans. Automat. Control, 53 (2008), pp. 461–478.
- [4] M. FRENCH, A. ILCHMANN, AND E. P. RYAN, *Robustness in the graph topology of a common adaptive controller*, SIAM J. Control Optim., 45 (2006), pp. 1736–1757.
- [5] T. T. GEORGIOU AND M. C. SMITH, *Robustness analysis of nonlinear feedback systems: An input-output approach*, IEEE Trans. Automat. Control, 42 (1997), pp. 1200–1221.
- [6] A. ILCHMANN, *Non-identifier-based adaptive control of dynamical systems: A survey*, IMA J. Math. Control Inform., 8 (1991), pp. 321–366.
- [7] A. ILCHMANN, *Adaptive  $\lambda$ -tracking for polynomial minimum phase systems*, Dynam. Stability Systems, 13 (1998), pp. 341–371.
- [8] A. ILCHMANN AND M. PAHL, *Adaptive multivariable pH regulation of a biogas tower reactor*, Eur. J. Control, 4 (1998), pp. 116–131.
- [9] A. ILCHMANN AND E. P. RYAN, *Universal  $\lambda$ -tracking for nonlinearly-perturbed systems in the presence of noise*, Automatica J. IFAC, 30 (1994), pp. 337–346.
- [10] A. ILCHMANN, E. P. RYAN, AND C. J. SANGWIN, *Systems of controlled functional differential equations and adaptive tracking*, SIAM J. Control Optim., 40 (2002), pp. 1746–1764.
- [11] A. ILCHMANN, E. P. RYAN, AND P. TOWNSEND, *Tracking with prescribed transient behavior for nonlinear systems of known relative degree*, SIAM J. Control Optim., 46 (2007), pp. 210–230.
- [12] A. ILCHMANN, M. THUTO, AND S. TOWNLEY, *Input constrained adaptive tracking with applications to exothermic chemical reaction models*, SIAM J. Control Optim., 43 (2004), pp. 154–173.
- [13] A. ISIDORI, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, Berlin, 1995.
- [14] I. M. Y. MAREELS, *A simple self-tuning controller for stably invertible systems*, Systems Control Lett., 4 (1984), pp. 5–16.
- [15] A. S. MORSE, *Recent problems in parameter adaptive control*, in Outils et Modèles Mathématiques pour l'Automatique, l'Analyse de Systèmes et le Traitement du Signal, I. D. Landau, ed., Birkhäuser Verlag, Paris, 1983, pp. 733–740.
- [16] M. VIDYASAGAR, *Nonlinear Systems Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
- [17] J. C. WILLEMS AND C. I. BYRNES, *Global adaptive stabilization in the absence of information on the sign of the high frequency gain*, in Analysis and Optimization of Systems, Lecture Notes in Control and Inform. Sci. 62, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, 1984, pp. 49–57.
- [18] E. ZEIDLER, *Nonlinear Functional Analysis and its Applications I: Fixed Point Theorems*, Springer-Verlag, New York, 1986.

## A SECTOR CONDITION FOR LIMIT CYCLE ROBUSTNESS\*

ULF T. JÖNSSON†

**Abstract.** Robustness of periodic oscillations in autonomous feedback systems is considered for systems with separable nonlinearities. Local quadratic separation of the nonlinear dynamics from the linear part of the dynamics is used to characterize a set of systems that exhibit periodic oscillation in a bounded frequency and amplitude range. The quadratic constraint is generated as a time-periodic sector condition that characterizes the nonlinearity around a nominal periodic solution. The main analysis condition is formulated as an operator inequality involving the nominal dynamics and the sector constraint. This is an infinite dimensional robustness test that must be truncated to be verified numerically. We discuss two possible ways of performing the analysis.

**Key words.** limit cycles, uncertain system, robustness

**AMS subject classifications.** 93D09, 37C27, 49N20

**DOI.** 10.1137/070701108

**1. Introduction.** We consider the existence and robustness of periodic solutions of the feedback system

$$(1.1) \quad \begin{aligned} y &= \Phi(w), \\ w &= Hy, \end{aligned}$$

where  $H$  is a bounded, linear time-invariant operator and  $\Phi$  is a nonlinear operator. The case of our interest is when  $H$  is only approximately known. We will assume that there exists a nominal model and that the size of the uncertainty in this model is characterized by one parameter. This is not restrictive and allows us to consider the uncertainty class in the structured singular value framework [15, 3, 17].

The question to be addressed is if a periodic solution to the nominal system implies the existence of a periodic solution to every system in the uncertainty class. Bounds on the possible location of such periodic solutions will also be obtained. The result thus provides a quantitative measure on whether a simplified model of an oscillator predicts the real limit cycle solution with acceptable accuracy. This is a fundamental problem in modeling and analysis of systems with limit cycle solutions, which has application in areas such as systems biology [19] and oscillatory control of mechanical systems [6].

The results are obtained using local quadratic constraints to separate the nonlinear from the linear part of the dynamics. The quadratic constraints are chosen to be time periodic in order to explore as much as possible the structure of the nonlinear part of the dynamics. The verification of the resulting quadratic inequality on the linear dynamics is performed using relaxation techniques from the robust control field. The techniques used and the problem under investigation are related to classical 1970's works on harmonic balance and describing functions; see, e.g., [13, 11, 12, 2, 16]. The

---

\*Received by the editors August 24, 2007; accepted for publication (in revised form) May 29, 2008; published electronically November 5, 2008. A brief version of this paper appeared in *Proceedings of the 46th IEEE Conference on Decision and Control*, New Orleans, LA, 2007. This work was supported by the Swedish Research Council.

<http://www.siam.org/journals/sicon/47-6/70110.html>

†Department of Mathematics, Division of Optimization and Systems Theory, Royal Institute of Technology, SE 100 44 Stockholm, Sweden (ulfj@math.kth.se).

focus of the present paper is on the robustness of a given limit cycle to perturbations in the dynamics, while the focus of the 1970's research was on predicting the existence of a limit cycle in a given system. However, it turns out that both of these questions can be addressed simultaneously in all the above works. Various sector criteria are also used to derive the results in [13, 11, 2], but two major differences compared to our approach are that the sector bounds are either time invariant or based on the describing function in [13, 11, 2] and that the separation between the nonlinear and linear parts of the dynamics is more explicit in our paper. Moreover, the verification of the sector bounds in this paper is based on techniques that were not fully developed in the 1970's. Quadratic constraints for predicting oscillations have also previously been used in [21, 14]. More recent works using control theoretic methods to address fundamental robustness issues for limit cycle systems can be found in [5, 18, 9]. The technique employed in this paper and the obtained results are significantly different from these works. Our main contribution is to show how time-varying sector bounds can be used in the analysis of limit cycles.

The paper is organized as follows. In section 2 the system model and properties of the limit cycle are discussed. Uncertainty is introduced to the system model in section 3, where our basic robustness result is also derived. This result provides the foundation for all subsequent results of the paper. We prove that our proposed method always works, provided that the predefined uncertainty descriptions are sufficiently small, and we show that the verification of the robustness conditions involves the computation of a robustness margin for the nominal system. In section 4 we show that the main conditions after truncation can be verified using linear matrix inequalities. Finally, we apply the ideas of the paper to an example involving a Van der Pol oscillator connected to a mass and spring system.

**Notation.** We let  $\mathbf{L}_2(1)$  denote the space of square integrable 1-periodic functions and  $C(1)$  the set of continuous 1-periodic functions equipped with the norm  $\|v\|_{C(1)} = \max_{t \in [0,1]} |v(t)|$ , where  $|\cdot|$  denotes the Euclidean norm. The functions in  $\mathbf{L}_2(1)$  and  $C(1)$  are assumed to take values in  $\mathbf{R}^m$ , but we suppress this information from the notation.

We make extensive use of the topological inclusion  $C(1) \hookrightarrow \mathbf{L}_2(1)$ , which follows since  $\|v\|_{\mathbf{L}_2(1)} \leq \|v\|_{C(1)}$ . We will almost always denote inner products, norms, and induced norms on  $\mathbf{L}_2(1)$  without the suffix, e.g.,  $\|\cdot\| := \|\cdot\|_{\mathbf{L}_2(1)}$  and  $\|\cdot\| := \|\cdot\|_{\mathbf{L}_2(1) \rightarrow \mathbf{L}_2(1)}$ . At several places we consider the space  $C(1) \times \mathbf{R}$  with the norm  $\|(y, T)\|_{C(1) \times \mathbf{R}} = (\|y\|_{C(1)}^2 + |T|^2)^{1/2}$  and similarly for  $\mathbf{L}_2(1) \times \mathbf{R}$ . We define

$$l_2(\mathbf{C}) = \left\{ \{\hat{y}[k]\}_{k=-\infty}^{\infty} : \hat{y}[k] \in \mathbf{C}^m; \hat{y}[-k] = \overline{\hat{y}[k]}; \sum_{k=-\infty}^{\infty} |\hat{y}[k]|^2 < \infty \right\}$$

and let  $\mathcal{F} : \mathbf{L}_2(1) \rightarrow l_2$  denote the Fourier transform defined by  $\hat{y}[k] = (\mathcal{F}y)[k] = \int_0^1 y(t) e^{-j2\pi kt} dt$ . We use that the Fourier transform is an isometric isomorphism between these two spaces, and we frequently use the Parseval relation that relates the inner product of the two spaces:

$$\langle y_1, y_2 \rangle = \int_0^1 y_1(t)^T y_2(t) dt = \sum_{k=-\infty}^{\infty} \hat{y}_1[k]^* \hat{y}_2[k].$$

Given a periodic matrix function  $\beta \in C(1)$  we define the corresponding frequency

domain representation using the (bi-infinite) block Toeplitz matrix

$$(1.2) \quad \text{Toep}[\beta] = \begin{bmatrix} \ddots & & & & \\ & \hat{\beta}[0] & \hat{\beta}[1] & \hat{\beta}[2] & \\ \dots & \hat{\beta}[-1] & \hat{\beta}[0] & \hat{\beta}[1] & \dots \\ & \hat{\beta}[-2] & \hat{\beta}[-1] & \hat{\beta}[0] & \\ & & & \ddots & \ddots \end{bmatrix},$$

which defines the frequency domain representation of multiplication, i.e.,  $\mathcal{F}\beta v = \text{Toep}[\beta]\hat{v}$  for any  $v \in \mathbf{L}_2(1)$ .

Let  $\Psi : \mathbf{L}_2(1) \rightarrow \mathbf{L}_2(1)$  be a bounded self-adjoint linear operator and  $\hat{\Psi}$  its corresponding frequency domain representation ( $\hat{\cdot}$  notation always refers to a frequency domain representation). Using  $\Psi$ , a continuous quadratic form  $\sigma_\Psi : \mathbf{L}_2(1) \rightarrow \mathbf{R}$  can be defined as

$$\sigma_\Psi(v) = \langle v, \Psi v \rangle_{\mathbf{L}_2(1)} = \left\langle \hat{v}, \hat{\Psi} \hat{v} \right\rangle_{l_2},$$

where the explicit dependence of  $\sigma$  on  $\Psi$  normally will be suppressed from the notation. One possibility is to use multiplication operators defined by  $(\Psi v)(t) = \psi(t)v(t)$ , where  $\psi \in C(1)$  and then  $\hat{\Psi} = \text{Toep}[\psi]$ . We define

$$\lambda_{\max}(\Psi) = \sup_{v \in \mathbf{L}_2(1)} \frac{\langle v, \Psi v \rangle_{\mathbf{L}_2(1)}}{\|v\|_{\mathbf{L}_2(1)}^2} \quad \text{and} \quad \lambda_{\min}(\Psi) = \inf_{v \in \mathbf{L}_2(1)} \frac{\langle v, \Psi v \rangle_{\mathbf{L}_2(1)}}{\|v\|_{\mathbf{L}_2(1)}^2}.$$

The operator  $\Psi$  is called positive definite if  $\lambda_{\min}(\Psi) > 0$ . The notation  $\lambda_{\min}$  and  $\lambda_{\max}$  is used to indicate that the above quantities are the maximum and minimum spectral values. If  $\Psi$  is positive definite, we often use that  $\langle v, \Psi v \rangle = \|\Psi^{1/2}v\|^2$ , where  $\Psi^{1/2}$  is the unique positive definite square root of  $\Psi$  [22].

We let  $P_N$  denote the orthogonal projection onto the  $2N + 1$  dimensional space spanned by the Fourier bases  $\{e^{j\omega k}\}_{-N}^N$  and  $Q_N = I - P_N$ .

If  $X \subset \mathbf{X}$  is a proper subspace, then the inclusion map  $I_X : X \rightarrow \mathbf{X}$  is defined by  $I_X x = x$  for all  $x \in X$ .

Finally,  $\text{cl } D$  denotes the closure of the set  $D$ , and  $\partial D$  denotes its boundary.

**2. Limit cycle model.** In this section we define some preliminaries that are adopted from [8]. Here a *limit cycle* will denote a periodic solution of (1.1). One of the main difficulties in limit cycle analysis is that both the trajectory and the period time may change as the dynamics are perturbed. In order to simplify the use of standard techniques from robust control we normalize the period time to one. By doing this the actual period time becomes a parameter in the linear dynamics. To understand how this works, let the nonlinear operator  $\Phi : \mathbf{L}_\infty(-\infty, \infty) \rightarrow \mathbf{L}_\infty(-\infty, \infty)$  in (1.1) be defined by the time-domain relation

$$(2.1) \quad (\Phi(w))(t) = \varphi(w(t)),$$

where  $\varphi : \mathbf{R}^m \rightarrow \mathbf{R}^m$  will be assumed to be  $C^1$ . We further assume that the linear part of the dynamics is defined by a strictly proper transfer function  $H(s)$  with corresponding convolution kernel  $h$ . If under these conditions there exists a  $T_0$  periodic solution  $y$  to (1.1), then the following equivalent closed loop equation holds:

$$y(t) = \varphi \left( \int_{-\infty}^{\infty} h(t - \tau)y(\tau)d\tau \right),$$

where  $y(t + T_0) = y(t)$  for all  $t$ . We may now normalize the period time and define  $y_0(t) = y(tT_0)$ , which is a 1-periodic function. Then the system equation can be rewritten as

$$y_0(t) = \varphi \left( \int_{-\infty}^{\infty} T_0 h(T_0(t - \tau)) y_0(\tau) d\tau \right),$$

which in operator notation is  $y_0 = \varphi(H[T_0]y_0)$ , where  $H[T_0](s) = H(s/T_0)$ . The conclusion is that we can always normalize the period time of  $y$  to be 1 by making the true period time a parameter in the system dynamics. Hence, searching for a limit cycle corresponds to searching for a pair  $z = (y, T)$  of a 1-periodic function and a period time. The above transformation simplifies the search for a limit cycle and has previously been used in various equivalent forms; see, e.g., [13].

The discussion above motivates us to consider the system

$$(2.2) \quad y = \Phi(H[T]y),$$

where  $H[T](s) = H(s/T)$  and  $\Phi : \mathbf{L}_{\infty}(-\infty, \infty) \rightarrow \mathbf{L}_{\infty}(-\infty, \infty)$  is a memoryless nonlinear operator defined by the time-domain relation in (2.1). We assume that the convolution operators corresponding to the transfer functions  $H(s)$ ,  $sH(s)$ , and  $sH'(s)$  are bounded, e.g. (here  $\mathcal{L}$  denotes the two-sided Laplace transform)

$$\mathcal{L}^{-1}\{sH(s)\} = \dot{h}_c(t) + \sum_{k=0}^{\infty} \dot{h}_k \delta(t - T_k),$$

where  $\dot{h}_c(\cdot) \in \mathbf{L}_1(-\infty, \infty)$  and  $\sum_{k=0}^{\infty} |\dot{h}_k| < \infty$ , and  $T_0 < T_1 < T_2 \dots$ . We have the following time-domain representation when  $H$  acts on functions in  $\mathbf{L}_2(1)$  (and  $C(1)$ ):

$$(2.3) \quad \begin{aligned} y(t) &= \int_{-\infty}^{\infty} h(t - \tau) v(\tau) d\tau \\ &= \int_0^1 \tilde{h}(t, \tau) v(\tau) d\tau, \end{aligned}$$

where  $\tilde{h} : [0, 1] \times [0, 1] \rightarrow \mathbf{R}^{m \times m}$  is defined by

$$\tilde{h}(t, \tau) = \sum_{k=-\infty}^{\infty} h(t + k - \tau)$$

and it satisfies the periodicity property  $\tilde{h}(1, \tau) = \tilde{h}(0, \tau)$ , for  $\tau \in [0, 1]$ , which ensures that  $y(t)$  is 1-periodic for any  $v \in \mathbf{L}_2(1)$ . The assumptions imply that any periodic solution of the system equation will be absolutely continuous. We will embed this solution into the larger space  $\mathbf{L}_2(1)$  while the output of the linear part of the dynamics is embedded into  $C(1)$ . The use of these two different topologies will in the next section allow us to locally characterize the nonlinear part of the dynamics using quadratic forms. This leads to a resulting robustness criterion that can be formulated as a frequency domain inequality on the linear dynamics. Finally, we note that the linear part of the dynamics is not restricted to be causal, although this is usually the case in applications.

We will next define the return difference operator for the system in (2.2). Its derivative is a measure of the local robustness of the limit cycle which will be used



in our results. Let  $z = (y, T) \in \mathbf{L}_2(1) \times \mathbf{R}$  and define the return difference  $F : \mathbf{L}_2(1) \times \mathbf{R} \rightarrow \mathbf{L}_2(1)$  by

$$(2.4) \quad F(z) = y - \Phi(H[T]y).$$

We note that any solution to  $F(z) = 0$  with  $T > 0$  corresponds to a limit cycle. By using (2.3) it follows that the operator  $F$  has the following equivalent time-domain representations:

$$\begin{aligned} (F(z))(t) &= y(t) - \varphi \left( \int_{-\infty}^{\infty} Th(T(t-\tau))y(\tau)d\tau \right) \\ &= y(t) - \varphi \left( \int_0^1 T\tilde{h}(Tt, T\tau)y(\tau)d\tau \right). \end{aligned}$$

It is now easy to see that the limit cycle at best is unique modulo time translations. Indeed, let  $z_0 = (y_0, T_0)$  be a nontrivial 1-periodic solution of (2.2) and let  $S_d : C(1) \times \mathbf{R} \rightarrow C(1) \times \mathbf{R}$  be the time translation operator defined by  $S_d(y(t), T) = (y(t-d), T)$ . Then using the first representation of  $F$  above

$$\begin{aligned} (F(S_d z_0))(t) &= y_0(t-d) - \varphi \left( \int_{-\infty}^{\infty} T_0 h(T_0(t-\tau))y_0(\tau-d)d\tau \right) \\ &= y_0(t-d) - \varphi \left( \int_{-\infty}^{\infty} T_0 h(T_0(t-d-\tau))y_0(\tau)d\tau \right) \\ (2.5) \quad &= (F(z_0))(t-d) = 0, \end{aligned}$$

which holds for any  $d \in \mathbf{R}$ . This shows that the time translated 1-periodic solution is still a valid periodic solution of (2.2). The limit cycle is thus not unique but belongs to the manifold  $\mathcal{Z}_0 = \{S_d z_0 : d \in [0, 1]\}$ , which, since  $(\mathcal{F}y(t-d))[k] = e^{j2\pi kd}\hat{y}[k]$ , equivalently can be represented in the frequency domain as

$$(2.6) \quad \mathcal{Z}_0 = \{(\hat{y}, \omega_0) : \hat{y}[k] = e^{j2\pi kd}\hat{y}_0[k]; d \in [0, 1]\},$$

where  $\omega_0 = 2\pi/T_0$ .

In order to proceed we need an expression for the derivative of the return difference.

**PROPOSITION 2.1.** *The operator  $F : \mathbf{L}_2(1) \times \mathbf{R} \rightarrow \mathbf{L}_2(1)$  defined in (2.4) has the Fréchet derivative  $F'(z_0) \in \mathcal{L}(\mathbf{L}_2(1) \times \mathbf{R}, \mathbf{L}_2(1))$  with block structure*

$$(2.7) \quad F'(z_0) = \begin{bmatrix} F'_y(z_0) & F'_T(z_0) \end{bmatrix} = \begin{bmatrix} I - L(z_0) & K(z_0) \end{bmatrix},$$

where  $L(z_0) : \mathbf{L}_2(1) \rightarrow \mathbf{L}_2(1)$ ,  $K(z_0) : \mathbf{R} \rightarrow \mathbf{L}_2(1)$  are defined as

$$(2.8) \quad \begin{aligned} L(z_0) &= \Phi'(H[T_0]y_0) \circ H[T_0], \\ K(z_0) &= -\Phi'(H[T_0]y_0) \circ D_T H[T_0]y_0 \end{aligned}$$

and where  $(\Phi'(w_0))(t) = \varphi'(w_0(t))$  and

$$(D_T H[T_0])(s) = -\frac{s}{T_0^2} H' \left( \frac{s}{T_0} \right).$$

*Proof.* We first show that  $\Phi' : C(1) \rightarrow \mathcal{L}(C(1), \mathbf{L}_2(1))$  can be defined by  $(\Phi'(w)h)(t) = \varphi'(w(t))h(t)$ . Indeed, since  $\varphi$  is  $C^1$  we have

$$\begin{aligned} \lim_{\|h\|_{C(1)} \rightarrow 0} \frac{\|\Phi(w+h) - \Phi(w) - \Phi'(w)h\|_{\mathbf{L}_2(1)}}{\|h\|_{C(1)}} \\ \leq \lim_{\|h\|_{C(1)} \rightarrow 0} \frac{\|\Phi(w+h) - \Phi(w) - \Phi'(w)h\|_{C(1)}}{\|h\|_{C(1)}} = 0, \end{aligned}$$

which proves differentiability. Due to the assumptions on  $H$  one can show that

$$\|H\|_{\mathbf{L}_2(1) \rightarrow C(1)}^2 \leq \sup_{t \in [0,1]} \int_0^1 |\tilde{h}(t, \tau)|^2 d\tau < \infty$$

and  $D_T H[T_0]y_0 \in C(1)$ . It follows by the composite mapping theorem that  $F = I - \Phi \circ H$  has a well-defined Fréchet derivative on the form given in the theorem statement.  $\square$

Differentiation of (2.5) with respect to  $d$  at  $d = 0$  gives

$$\begin{aligned} \dot{y}_0(t) &= \varphi'((H[T_0]y_0)(t)) \int_{-\infty}^{\infty} T_0 h(T_0(t - \tau)) \dot{y}_0(\tau) d\tau \\ &= (L(z_0)\dot{y}_0)(t). \end{aligned}$$

This shows that  $(\dot{y}_0, 0) \in \text{Ker } F'(z_0)$ , where  $F'(z_0)$  is the Fréchet derivative.

A right inverse of  $F'(z_0)$  is a bounded linear operator  $F'^{\dagger} \in \mathcal{L}(\mathbf{L}_2(1), \mathbf{L}_2(1) \times \mathbf{R})$  satisfying  $F'(z_0)F'^{\dagger} = I$ . The next result shows that  $\text{Ker } F'(z_0) = \text{span}\{(\dot{y}_0, 0)\}$  is a necessary and sufficient condition for the existence of a right inverse.

**PROPOSITION 2.2.** *There exists a bounded right inverse  $F'(z_0)^{\dagger}$  if and only if  $\text{Ker } F'(z_0) = \text{span}\{(\dot{y}_0, 0)\}$ . Moreover, the periodic solution  $z_0 \in \mathbf{L}_2(1) \times \mathbf{R}$  of (2.2) is isolated if  $F'(z_0)$  has a bounded right inverse.*

*Proof.* Since  $L(z_0)$  and  $K(z_0)$  are compact it follows that  $F'(z_0)$  is a Fredholm operator with index  $\text{Ind}(F'(z_0)) = 1$ . It thus follows that  $F'(z_0)$  is surjective if and only if  $\dim \text{Ker } F'(z_0) = 1$ , which by the Banach isomorphism theorem implies that there exists a bounded linear operator  $F'(z_0)^{\dagger} \in \mathcal{L}(\mathbf{L}_2(1), \mathbf{L}_2(1) \times \mathbf{R})$  such that  $F'(z_0)F'(z_0)^{\dagger} = I$ .

The last statement is related to the submersion theorem; see Theorem 3.5.4 in [1].  $\square$

A limit cycle solution of (2) belongs to the manifold  $\mathcal{Z}_0$  in (2.6), and it is thus only unique modulo time translation. By fixing the phase a priori we get a unique solution which simplifies the analysis. For this reason we introduce the space

$$(2.9) \quad \mathbf{Y} = \{y \in \mathbf{L}_2(1) : \langle y, v \rangle = 0\},$$

where the choice of  $v$  is either of the following:

- (a) A common choice is to let  $\langle y, v \rangle = \text{re} \{e_1^T \hat{y}[1]\} = 0$ ,<sup>1</sup> where  $e_1 = [1 \ 0 \ \dots \ 0]^T$  is the first unit vector.
- (b) The choice  $v = \dot{y}_0$  implies that  $\mathbf{Y}$  is the set of periodic functions in “phase” with the nominal solution  $y_0$ .

<sup>1</sup>Here we assume that  $e_1^T \hat{y}_0[1] \neq 0$  so that the constraint defines useful phase information. This assumption remains throughout the paper.

We let  $\mathbf{Z} = \mathbf{Y} \times \mathbf{R}$  and restrict  $F$  in (2.4) to  $\mathbf{Z}$ . By considering the system on  $\mathbf{Z}$  implies that we fix the phase and ideally consider one particular point on the limit cycle manifold. For the choice of  $v$  in (a) above, this is evident since it implies that only the phase  $d = 0$  is valid in (2.6). In (b) a similar conclusion holds (at least) locally. To see this we notice that  $f(d) = \langle S_d y_0, \dot{y}_0 \rangle$  satisfies  $f(0) = 0$  and  $f'(0) = \|\dot{y}_0\|^2 \neq 0$ . For both cases (a) and (b) we may use the following simplified version of Proposition 2.2 for the Fréchet derivative of the restricted return difference

**COROLLARY 2.3.** *Let  $F : \mathbf{Z} \rightarrow \mathbf{L}_2(1)$  be defined as in (2.4). The derivative  $F'(z_0)$  is defined as in (2.7) with  $L(z_0) : \mathbf{Y} \rightarrow \mathbf{L}_2(1)$ ,  $K(z_0) : \mathbf{R} \rightarrow \mathbf{L}_2(1)$  defined as in (2.8). It follows that  $F'(z_0)$  has a bounded inverse if and only if  $\text{Ker } F'(z_0) = 0$ .*

*Proof.* The proof follows easily from Proposition 2.2. Indeed, the restriction of  $F$  in (2.4) to the subspace  $\mathbf{Z}$  can be formulated as  $F|_{\mathbf{Z}} = F \circ (I_{\mathbf{Y}}, I_{\mathbf{R}})$ , where  $I_{\mathbf{Y}}$  is the inclusion of  $\mathbf{Y}$  into  $\mathbf{L}_2(1)$  defined by  $I_{\mathbf{Y}}y = y$  for all  $y \in \mathbf{Y}$ , and  $I_{\mathbf{R}}$  is the identity on  $\mathbf{R}$ . Therefore we need only show that  $\text{Ker } F'(z_0) = \text{span}\{(\dot{y}_0, 0)\} \Leftrightarrow \text{Ker } F'|_{\mathbf{Z}}(z_0) = \{(0, 0)\}$ . This is obvious if  $v = \dot{y}_0$  in (2.9) because then  $\text{span}\{(\dot{y}_0, 0)\} \cap \mathbf{Y} \times \mathbf{R} = \{0\}$ . For the second case of our concern, we have  $\mathbf{Y} = \{y \in \mathbf{L}_2(1) : e_1^T \text{Re } \hat{y}[1] = 0\}$ . Since  $\hat{y}_0[1] = j\omega_0 \hat{y}_0[1] \in \mathbf{R}$  and is nonzero by assumption we again have  $\text{span}\{(\dot{y}_0, 0)\} \cap \mathbf{Y} \times \mathbf{R} = \{0\}$ . Hence, the proof of Corollary 2.3 follows from Proposition 2.2.  $\square$

**3. Robustness of limit cycle.** In this section we derive an existence and robustness result for limit cycles. The conditions are formulated as integral quadratic constraints on the linear and nonlinear parts of the system. We consider the set of systems

$$(3.1) \quad y = \Phi(H[T, \theta])y), \quad \theta \in [0, 1],$$

where the following hold:

- (A1)  $\Phi : C(1) \rightarrow \mathbf{L}_2(1)$  is a nonlinear operator, which in time domain is defined by  $(\Phi(w))(t) = \varphi(w(t))$ , where  $\varphi : \mathbf{R}^m \rightarrow \mathbf{R}^m$  is a  $C^1$  function.
- (A2)  $H[T, \theta](s) = H(s/T, \theta)$  is for each  $\theta \in [0, 1]$  a strictly proper transfer function as defined in section 2. We assume that the dependence on the parameter  $\theta$  is Lipschitz continuous, i.e., for  $\theta_1, \theta_2 \in [0, 1]$

$$\|H[T, \theta_1] - H[T, \theta_2]\|_{\mathbf{L}_2(1) \rightarrow C(1)} \leq c|\theta_1 - \theta_2|,$$

where  $c$  is a positive constant. This bound is assumed to be uniform over all period times in the set  $\mathcal{T}$  defined in (3.4) below.

The system (3.1) is called *nominal* when  $\theta = 0$ , and we assume that the nominal system has a solution. Our goal is to prove that there remains a limit cycle solution in a neighborhood of the nominal solution for all values of the perturbation parameter  $\theta \in [0, 1]$ .

We will exploit that the linear part of the system maps to continuous signals, where the sup-norm topology can be used to define neighborhoods of the nominal solution. It is then reasonably easy to describe how the nonlinear operator maps this neighborhood to the output. We will do this using “local” integral quadratic constraints.

We know from above that the limit cycle is only unique modulo time translation. We will therefore fix the phase in order to get unique solutions and easier analysis. This is done by restricting the analysis to the space  $\mathbf{Y}$  defined in (2.9). We next summarize some further assumptions and definitions.

- (A3) We assume that the nominal system has a solution  $z_0 = (y_0, T_0) \in \mathbf{Z}$ , i.e.,  $y_0 = \Phi(H[T_0, 0]y_0)$ , where  $\mathbf{Z} = \mathbf{Y} \times \mathbf{R}$  with norm  $\|z\| = (\|y\|^2 + T^2)^{1/2}$  and where  $\mathbf{Y}$  is defined in (2.9).
- (D1) Let  $\mathcal{Z} \subset \mathbf{Z}$  be a bounded, convex, and open set defined as

$$(3.2) \quad \mathcal{Z} = \{z = (y, T) \in \mathbf{Z} : \|\Psi_Z^{1/2}(z - z_0)\| < 1\},$$

where  $\Psi_Z$  is a positive definite operator on  $\mathbf{L}_2(1) \times \mathbf{R}$  with block structure

$$(3.3) \quad \Psi_Z = \begin{bmatrix} \Psi_Y & 0 \\ 0 & \psi_T \end{bmatrix}.$$

The constraint in (3.2) can equivalently be written as  $\sigma_Z(z - z_0) < 0$ , where  $\sigma_Z : \mathbf{L}_2(1) \times \mathbf{R} \rightarrow \mathbf{R}^+$  is a continuous quadratic form defined as

$$\sigma_Z(z) = \langle z, \Psi_Z z \rangle - 1, \quad \text{where} \quad \langle z, \Psi_Z z \rangle = \langle y, \Psi_Y y \rangle + \psi_T T^2.$$

The above definition implies that the period time belongs to

$$(3.4) \quad \mathcal{T} = (T_{\min}, T_{\max}),$$

where  $T_{\min} = T_0 - \psi_T^{-1/2}$  and  $T_{\max} = T_0 + \psi_T^{-1/2}$ . We assume  $T_{\min} > 0$ , which ensures that  $H[T, \theta]$  is continuous for all  $T \in \mathcal{T}$ .

- (D2) Let  $\mathcal{W} \subset C(1)$  be a bounded, closed, and convex set that contains  $w_0 = H[T_0]y_0$  in its interior. The restriction of the nonlinear operator to  $\mathcal{W}$ ,  $\Phi|_{\mathcal{W}}(\cdot) : \mathcal{W} \rightarrow \mathbf{L}_2(1)$  will be Lipschitz continuous due to (A1), and we denote the Lipschitz constant  $L_\Phi$ .

We use that the operator in the right-hand side of (3.1) is continuous and compact under (A1)–(A3), (D1)–(D2).

**LEMMA 3.1.** *The operator  $\Psi : \mathbf{Z} \times [0, 1] \rightarrow \mathbf{L}_2(1)$  defined by  $\Psi(z, \theta) = \Phi(H[T, \theta]y)$  is compact on  $\text{cl } \mathcal{Z}$  for each  $\theta \in [0, 1]$ . Moreover,  $F : \text{cl } \mathcal{Z} \times [0, 1] \rightarrow \mathbf{L}_2(1)$  defined as  $F(z, \theta) = z - \Psi(z, \theta)$  is a homotopy of compact transformations on  $\text{cl } \mathcal{Z}$ . This means that for given  $\epsilon > 0$  there exists  $\delta > 0$  such that*

$$\|\Psi(z, \theta_1) - \Psi(z, \theta_2)\| \leq \epsilon \quad \forall z \in \text{cl } \mathcal{Z}, \quad |\theta_1 - \theta_2| \leq \delta.$$

*Proof.*  $\Psi = \Phi \circ H$  is compact since it is a composition of a compact operator  $H$  with a continuous operator with bounded range  $\Phi$  (by (D2)). The continuity with respect to  $\theta$  follows from (D2). Indeed, due to the Lipschitz continuity of  $\Phi$  there exists  $L_\Phi$  such that

$$\begin{aligned} \|\Phi(H[T, \theta_2]y) - \Phi(H[T, \theta_1]y)\| &\leq L_\Phi \|H[T, \theta_2] - H[T, \theta_1]\|_{C(1)} \|y\| \\ &\leq c L_\Phi \sup_{y \in \text{cl } \mathcal{Y}} \|y\| |\theta_2 - \theta_1|, \end{aligned}$$

where  $\mathcal{Y} = \{y \in \mathbf{Y} : \|\Psi_Y^{1/2}(y - y_0)\| < 1\}$  and  $c$  was defined in (A2).  $\square$

We use topological degree theory to derive a robustness result for the system in (3.1). The degree provides a modulo two count of the number of solutions to an operator equation in a given set. Let  $F = I - \Psi$ , where  $\Psi : Z \rightarrow Z$  is a compact operator on a Banach space  $Z$  with a countable basis. In our application  $F$  will depend on a parameter as in Lemma 3.1, and we will derive conditions under which

the existence of a solution to  $F(z_0, 0) = 0$  implies the existence of a solution to  $F(z, 1) = 0$ . The proof relies on an invariance property of the topological degree, which in our case will be verified using time-periodic sector conditions.

To define the degree  $d(F, D, 0)$  of  $F$  at 0 with respect to a bounded open set  $D$  we let  $P_n$  denote a projection on the finite dimensional space spanned by the first  $n$  basis elements,  $F_n = P_n \circ F \circ P_n$ , and  $D_n = P_n D$ . If  $F$  is  $C^1$ , then we define the degree as

$$d(F, D, 0) = \lim_{n \rightarrow \infty} d(F_n, D_n, 0),$$

$$\text{where } d(F_n, D_n, 0) = \sum_{z \in F_n^{-1}(0) \cap D_n} \text{sign}(\det(F'_n(z))).$$

Here it is assumed that  $0 \notin F_n(\partial D_n)$  and  $\det(F'_n(z)) \neq 0$  for all  $z \in F_n^{-1}(0) \cap D_n$ . Otherwise let  $F_\epsilon = I - \Psi_\epsilon$ , where  $\Psi_\epsilon$  is a compact  $C^1$  approximation such that  $\|F - F_\epsilon\| \leq \epsilon$  and define

$$d(F, D, 0) = \lim_{\epsilon \rightarrow 0} d(F_\epsilon, D, 0).$$

We refer the reader to [4, 10, 13] for treatments of degree theory and its application to periodic systems analysis. The next result is derived using degree theory and provides the foundation for the subsequent results of the paper.

**THEOREM 3.2.** *Consider the set of systems (3.1) under the assumptions (A1)–(A3) and (D1)–(D2). Let  $F : \text{cl } \mathcal{Z} \times [0, 1] \rightarrow \mathbf{L}_2(1)$  be defined as*

$$(3.5) \quad F(z, \theta) = y - \Phi(H[T, \theta]y).$$

*Suppose that*

- (i)  $F'_z(z_0, 0) : \mathbf{Z} \rightarrow \mathbf{L}_2(1)$  has a bounded inverse;
- (ii)  $H[T, \theta]y \subset \mathcal{W}$  for all  $(y, T) \in \text{cl } \mathcal{Z}$  and  $\theta \in [0, 1]$ ;

*and suppose that there exists a continuous quadratic form  $\sigma : \mathbf{L}_2(1) \times \mathbf{L}_2(1) \rightarrow \mathbf{R}$  such that*

- (iii)  $\sigma(w - w_0, \Phi(w) - \Phi(w_0)) \geq 0$  for all  $w \in \mathcal{W}$ ;
- (iv) *there exists  $\epsilon > 0$  such that*
  - (a)  $\sigma(H[T, \theta]y - H[T_0, 0]y_0, y - y_0) \leq -\epsilon$  for all  $(y, T) \in \partial \mathcal{Z}$ ,  $\theta \in [0, 1]$ ;
  - (b)  $\sigma(H[T, 0]y - H[T_0, 0]y_0, y - y_0) \leq -\epsilon \|z - z_0\|^2$  for all  $z = (y, T) \in \mathcal{Z}$ .

*Then for all  $\theta \in [0, 1]$  there exists  $(y_\theta, T_\theta) \in \mathcal{Z}$  such that  $y_\theta = \Phi(H[T_\theta, \theta]y_\theta)$ .*

**Remark 1.** It is generally not evident that (iv)(a) would imply (iv)(b). We will develop computational tests that provide sufficient conditions that simultaneously verify both (iv)(a) and (iv)(b).

**Proof.** By assumption (A3),  $F(z_0, 0) = 0$ . We will use topological degree theory to show that for all  $\theta \in [0, 1]$  there exists  $z_\theta \in \mathcal{Z}$  such that  $F(z_\theta, \theta) = 0$ . In order to make the problem considered here fit the basic case in [10] we introduce  $\tilde{F} : \text{cl } \mathcal{Z} \times [0, 1] \rightarrow \mathbf{Z}$  defined as

$$\tilde{F}(z, \theta) = (P_Y F(z, \theta), \langle F(z, \theta), v \rangle),$$

where  $P_Y : \mathbf{L}_2(1) \rightarrow \mathbf{Y}$  is the projection onto  $\mathbf{Y}$ . To prove the result we will show that

- (a)  $\deg(\tilde{F}(\cdot, 0), \mathcal{Z}, 0) \neq 0$ ;
- (b) (ii)–(iv) implies that  $0 \notin \tilde{F}(\partial \mathcal{Z}, \theta)$  for all  $\theta \in [0, 1]$ .

These two properties together with Lemma 3.1 imply by the invariance of the degree under homotopy that  $\deg(\tilde{F}(\cdot, \theta), \mathcal{Z}, \theta) = \text{const} \neq 0$  for  $\theta \in [0, 1]$ . This in turn implies the existence of  $(y_\theta, T_\theta) \in \mathcal{Z}$  such that  $\tilde{F}(z, \theta) = 0$  for all  $\theta \in [0, 1]$ . See [10] for a proof of these facts.

We can use the representation  $\tilde{F} = \Pi \circ F$ , where  $\Pi : \mathbf{L}_2(1) \rightarrow \mathbf{Z}$  defined by  $\Pi y = (P_Y y, \langle y, v \rangle)$  is an invertible linear operator. Therefore

$$\deg(\tilde{F}(\cdot, \theta), \mathcal{Z}, 0) \neq 0 \quad \Leftrightarrow \quad \deg(F(\cdot, \theta), \mathcal{Z}, 0) \neq 0,$$

and it is thus sufficient to prove properties (a) and (b) for  $F$  in order to conclude that there exists  $(y_\theta, T_\theta) \in \mathcal{Z}$  such that  $F(z, \theta) = 0$ , i.e.,  $y = \Phi(H[T, \theta])y$ , for all  $\theta \in [0, 1]$ .

We will first use (ii)–(iv) to show that  $\|F(z, \theta)\| > 0$  for all  $z \in \partial\mathcal{Z}$  and  $\theta \in [0, 1]$ . This proves that  $0 \notin F(\partial\mathcal{Z}, \theta)$  for all  $\theta \in [0, 1]$ . We consider solutions to the equation system

$$(3.6) \quad \begin{aligned} y &= \Phi(w), \\ w &= H[T, \theta]y. \end{aligned}$$

If  $z = (y, T) \in \partial\mathcal{Z}$ , then by (ii),  $w = H[T, \theta]y \in \mathcal{W}$ . We can therefore use (iii)–(iv) to derive quantitative relations for solutions  $(y, T, w)$  of (3.6) in  $\partial\mathcal{Z} \times \mathcal{W}$ .

We have by (ii) and (iv)(a)  $(F(z, \theta) = y - \Phi(w))$  and  $(F(z_0, 0) = y_0 - \Phi(w_0))$

$$\begin{aligned} -\epsilon &\geq \sigma(H[T, \theta]y - H[T_0, 0]y_0, y - y_0) \\ &= \sigma(w - w_0, y - y_0 - (\Phi(w) - \Phi(w_0)) + (\Phi(w) - \Phi(w_0))) \\ &\geq \sigma(w - w_0, \Phi(w) - \Phi(w_0)) - \bar{c}_1 \|F(z, \theta) - F(z_0, 0)\|^2 \\ &\quad - \bar{c}_2 (\|w - w_0\| + \|\Phi(w) - \Phi(w_0)\|) \cdot \|F(z, \theta) - F(z_0, 0)\| \\ &\geq -\bar{c}_1 \|F(z, \theta) - F(z_0, 0)\|^2 - \bar{c}_3 \|w - w_0\|_{C(1)} \cdot \|F(z, \theta) - F(z_0, 0)\|, \end{aligned}$$

where  $\bar{c}_1, \bar{c}_2$  are nonnegative constants determined by the quadratic form and

$$\bar{c}_3 = \bar{c}_2 + \bar{c}_2 L_\Phi.$$

We used that  $\|\Phi(w) - \Phi(w_0)\|_{\mathbf{L}_2(1)} \leq L_\Phi \|w - w_0\|_{\mathbf{L}_2(1)} \leq L_\Phi \|w - w_0\|_{C(1)}$  to obtain the last inequality. This gives the inequality

$$\|F(z, \theta) - F(z_0, 0)\| \geq -\frac{\bar{c}_3 \|w - w_0\|_{C(1)}}{2\bar{c}_1} + \sqrt{\frac{\epsilon}{\bar{c}_1} + \frac{(\bar{c}_3 \|w - w_0\|_{C(1)})^2}{4\bar{c}_1^2}}.$$

This in turn shows that  $\|F(z, \theta) - F(z_0, 0)\| = \|F(z, \theta)\| > 0$  for  $z \in \partial\mathcal{Z}$ ,  $\theta \in [0, 1]$ . If  $\bar{c}_1 = 0$ , then we have

$$\|F(z, \theta) - F(z_0, 0)\| \geq \frac{\epsilon}{\bar{c}_3 \|w - w_0\|_{C(1)}},$$

and the same conclusion holds.

An analogous derivation using (iv)(b) shows that

$$\|F(z, 0) - F(z_0, 0)\| \geq \delta(\|z - z_0\|) \quad \forall z \in \text{cl } \mathcal{Z},$$

where  $\delta(\cdot)$  is a positive definite function. For example, if  $\bar{c}_1 \neq 0$ , then we can take

$$\delta(\|z - z_0\|) = \begin{cases} \left( \sqrt{\frac{\epsilon}{\bar{c}_1} + \frac{\bar{c}_2^2 \|w - w_0\|_{C(1)}^2}{4\bar{c}_1^2}} - \frac{\bar{c}_2}{2\bar{c}_1} \frac{\|w - w_0\|_{C(1)}}{\|z - z_0\|} \right) \|z - z_0\|, & z \neq z_0, \\ 0, & z = z_0. \end{cases}$$

Using that  $F(z_0, 0) = 0$  implies that  $\|F(z, 0)\| > 0$  when  $z \neq z_0$  in  $\text{cl } \mathcal{Z}$  and hence that  $z_0$  is a unique periodic solution of the nominal system in  $\mathcal{Z}$ . Therefore, since  $F'_z(z_0, 0)$  has a bounded inverse,

$$\deg(F(\cdot, 0), \mathcal{Z}, 0) = \lim_{N \rightarrow \infty} \text{sign}(\det(F'_N(z_0, 0))) \neq 0,$$

where  $F_N(\cdot, 0) = P_N F(P_N \cdot, 0)$  denotes the truncation of  $F$  to the finite dimensional space  $P_N \mathbf{L}_2(1)$ .

We have shown that conditions (a) and (b) above hold, and this implies that the conclusion of the theorem is valid.  $\square$

In the next sections we discuss the choice of quadratic forms and the verification of conditions (ii), (iii), and (iv) of Theorem 3.2.

**3.1. Quadratic forms.** The constraint (iii) in Theorem 3.2 typically originates from a pointwise sector condition on the form

$$(3.7) \quad (\beta(t)(w(t) - w_0(t)) - ((\Phi(w))(t) - (\Phi(w_0))(t)))^T \\ \times ((\Phi(w))(t) - (\Phi(w_0))(t) - \alpha(t)(w(t) - w_0(t))) \geq 0 \quad \forall t \in [0, 1],$$

where  $\beta, \alpha \in C(1)$  (note that  $\beta, \alpha$  may be matrix valued functions). This corresponds to a quadratic inequality on the form  $\sigma(w - w_0, \Phi(w) - \Phi(w_0)) \geq 0$ , where

$$(3.8) \quad \sigma(w - w_0, \Phi(w) - \Phi(w_0)) = \left\| \frac{\beta - \alpha}{2}(w - w_0) \right\|^2 - \left\| \Phi(w) - \Phi(w_0) - \frac{\beta + \alpha}{2}(w - w_0) \right\|^2,$$

which may be equivalently formulated as

$$(3.9) \quad \sigma(w, y) = \left\langle \begin{bmatrix} w \\ y \end{bmatrix}, \Psi \begin{bmatrix} w \\ y \end{bmatrix} \right\rangle,$$

where  $\Psi : \mathbf{L}_2(1) \times \mathbf{L}_2(1) \rightarrow \mathbf{L}_2(1) \times \mathbf{L}_2(1)$  has the block representation

$$(3.10) \quad \Psi = \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{12}^* & -I \end{bmatrix}$$

with  $\Psi_{11} = -\frac{1}{2}(\beta^T \alpha + \alpha^T \beta)$  and  $\Psi_{12} = \frac{1}{2}(\beta + \alpha)$ . We primarily use the formulation in (3.9)–(3.10) in the remaining sections of the paper.

**3.2. A proof of the method.** Next we show that it always will be possible to verify robustness using Theorem 3.2, provided that  $\mathcal{Z}$  is small, and  $H[T, \theta]$  and  $H[T_0, 0]$  are close enough. For simplicity, we consider only the case with one scalar nonlinearity.

**PROPOSITION 3.3.** *Let  $\mathcal{W} \subset C(1)$  and suppose  $(\Phi(w))(t) = \varphi(w(t))$  satisfies the sector condition (3.7) for all  $w \in \mathcal{W}$  when*

$$\beta(t) = \varphi'(w_0(t)) + \epsilon^{1/2} \quad \text{and} \quad \alpha(t) = \varphi'(w_0(t)) - \epsilon^{1/2}.$$

*Let  $F'(z_0, 0) : \mathbf{Z} \rightarrow \mathbf{L}_2(1)$  be defined as*

$$F'(z_0, 0) = \begin{bmatrix} I - \Phi'(w_0)H[T_0, 0] & -\Phi'(w_0)D_T H[T_0, 0]y_0 \end{bmatrix}.$$

If  $F'(z_0, 0)$  has a bounded inverse and if  $\|H[T, \theta] - H[T, 0]\|$  is small enough, then condition (iv) in Theorem 3.2 can be verified by making  $\epsilon$  and  $\mathcal{Z}$  sufficiently small.

*Remark 2.* The remaining conditions of Theorem 3.2 will be satisfied under weak assumptions. For example, if in (A1) it is assumed that  $\varphi \in C^2$ , then it is possible to take  $\epsilon^{1/2} = O(\|z - z_0\|^2)$ , where  $O(\cdot)$  is the ordo notation. From (3.12) below, it follows that (iv) is satisfied if  $\mathcal{Z}$  is sufficiently small (provided that  $\|H[T, \theta] - H[T, 0]\|$  is small enough). In addition, it can also easily be verified that then (ii) and (iii) can be satisfied if  $\mathcal{W}$  and  $\mathcal{Z}$  are chosen appropriately small.

*Proof.* Using (3.7) with  $\alpha$  and  $\beta$  as in the statement of the proposition it follows that (iv)(a) gives the condition<sup>2</sup>

$$(3.11) \quad \|y - y_0 - \Phi'(w_0)H[T, \theta](y - y_0) - \Phi'(w_0)(H[T, \theta] - H[T_0, 0])y_0\|^2 \geq \epsilon(1 + \|\Delta_1(z, \theta)\|^2)$$

for all  $(y, T) \in \partial\mathcal{Z}$ , where  $\Delta_1(z, \theta) := H[T, \theta]y - H[T_0, 0]y_0$ . The expression within the left-hand side norm can be rewritten as  $F'(z_0, 0)(z - z_0) + \Delta_2$ , where

$$\begin{aligned} \Delta_2(z, \theta) &= -\Phi'(w_0)(H[T, \theta] - H[T_0, 0])(y - y_0) \\ &\quad - \Phi'(w_0)(H[T, \theta] - H[T_0, 0] - D_T H[T_0, 0](T - T_0))y_0. \end{aligned}$$

By using that  $F'(z_0, 0)$  has a bounded inverse we can bound the norm on the left-hand side of (3.11) as

$$\begin{aligned} \|F'(z_0)(z - z_0) + \Delta_2\| &\geq \|F'(z_0)(z - z_0)\| - \|\Delta_2\| \\ &\geq \|F'(z_0)^{-1}\|^{-1}\|z - z_0\| - \|\Delta_2\|. \end{aligned}$$

Hence, condition (iv)(a) in Theorem 3.2 is implied by the condition

$$(3.12) \quad \|F'(z_0)^{-1}\|^{-1} \geq \sup_{z \in \partial\mathcal{Z}} \frac{1}{\|z - z_0\|} \left( \epsilon^{1/2} \sqrt{1 + \|\Delta_1(z, \theta)\|^2} + \|\Delta_2(z, \theta)\| \right).$$

If the perturbation of  $H$  is sufficiently small, then there exist  $c_1(\mathcal{Z}), c_2(\mathcal{Z}) > 0$  such that (since  $\|\Delta_2(z, 0)\| = O(\|z - z_0\|^2)$  and  $\|\Delta_1(z, 0)\| = O(\|z - z_0\|)$ , where  $O(\cdot)$  is the ordo notation)

$$\begin{aligned} \|\Delta_2(z, \theta)\| &\leq c_1(\mathcal{Z})\|z - z_0\|^2, \\ \|\Delta_1(z, \theta)\| &\leq c_2(\mathcal{Z})\|z - z_0\| \end{aligned}$$

for all  $z \in \partial\mathcal{Z}$ . It follows that the right-hand side of (3.12) can be made sufficiently small for the inequality to be satisfied by making  $\epsilon$  and  $\mathcal{Z}$  small enough.

For condition (iv)(b) we let  $\theta = 0$ . The same arguments as above gives the condition

$$\|F'(z_0)^{-1}\|^{-1} \geq \sup_{z \in \mathcal{Z}, z \neq z_0} \frac{1}{\|z - z_0\|} \left( \epsilon^{1/2} \sqrt{\|z - z_0\|^2 + \|\Delta_1(z, 0)\|^2} + \|\Delta_2(z, 0)\| \right),$$

which by the above arguments holds if  $\epsilon$  and  $\mathcal{Z}$  are sufficiently small.  $\square$

<sup>2</sup>We use that the quadratic form is real valued, and hence  $\langle w, y \rangle = \langle y, w \rangle$ .



**3.3. Verification of the quadratic constraints in (iv).** Here we discuss how condition (iv) in Theorem 3.2 can be formulated as a condition on a stability margin that can be computed from an operator inequality that involves the nominal dynamics, the sector description of the nonlinearity, and the set description in (D1). Let us start to derive a description of  $H[T, \theta]y - H[T_0, 0]y_0$  that is more tractable for our analysis. It can be rewritten as

$$(3.13) \quad H[T, \theta]y - H[T_0, 0]y_0 = \tilde{H}(z - z_0) + \tilde{w}(z, \theta),$$

where  $\tilde{w}(z, \theta) = \tilde{\Delta}[T, \theta](z - z_0) + w_{\Delta}[T, \theta]$  and

$$(3.14) \quad \tilde{H} = \begin{bmatrix} H[T_0, 0] & D_T H[T_0, 0]y_0 \end{bmatrix},$$

$$(3.15) \quad \tilde{\Delta}[T, \theta] = \begin{bmatrix} \Delta[T, \theta] & \delta[T] \end{bmatrix},$$

$$(3.16) \quad w_{\Delta}[T, \theta] = (H[T, \theta] - H[T, 0])y_0$$

and where

$$(3.17) \quad \begin{aligned} \Delta[T, \theta] &= H[T, \theta] - H[T_0, 0], \\ \delta[T] &= \begin{cases} \frac{1}{T - T_0} (H[T, 0] - H[T_0, 0] - D_T H[T_0, 0](T - T_0))y_0, & T \neq T_0, \\ 0, & T = T_0. \end{cases} \end{aligned}$$

We assume that we have the following constraint for the nonlinearity:

$$(3.18) \quad \sigma(w - w_0, \Phi(w) - \Phi(w_0)) \geq 0 \quad \forall w \in \mathcal{W},$$

where the quadratic form is defined by the self-adjoint operator  $\Psi : \mathbf{L}_2(1) \times \mathbf{L}_2(1) \rightarrow \mathbf{L}_2(1) \times \mathbf{L}_2(1)$ . Then we have the following result.

**THEOREM 3.4.** *Suppose there exists  $\tau > 0$  such that*

$$(3.19) \quad I_{\mathbf{Z}}^*(\Pi^* \Psi \Pi + \tau \Psi_Z) I_{\mathbf{Z}} \leq 0,$$

where  $\Psi_Z$  is defined in (3.3),

$$\Pi = \begin{bmatrix} H[T_0, 0] & D_T H[T_0, 0]y_0 \\ I & 0 \end{bmatrix},$$

and  $I_{\mathbf{Z}} : \mathbf{Z} \rightarrow \mathbf{L}_2(1) \times \mathbf{R}$  is the inclusion of  $\mathbf{Z}$  into  $\mathbf{L}_2(1) \times \mathbf{R}$ . Then condition (iv) in Theorem 3.2 holds if

$$(3.20) \quad \tau \geq \epsilon + 2 \sup_{\theta \in [0, 1]} \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta]),$$

where  $\mathcal{T}$  is defined in (3.4),

$$\Upsilon_1[T, \theta] = \Omega_1[T, \theta] + \Omega_1[T, \theta]^*,$$

and

$$\Omega_1[T, \theta] = \begin{bmatrix} \tilde{\Delta}[T, \theta] \Psi_Z^{-1/2} & w_\Delta[T, \theta] \\ \begin{bmatrix} 0 & 0 \end{bmatrix} & 0 \end{bmatrix}^* \Psi \begin{bmatrix} (\tilde{H} + \frac{1}{2} \tilde{\Delta}[T, \theta]) \Psi_Z^{-1/2} & \frac{1}{2} w_\Delta[T, \theta] \\ \begin{bmatrix} \Psi_Y^{-1/2} & 0 \end{bmatrix} & 0 \end{bmatrix}.$$

*Remark 3.* We have  $I_Z = \begin{bmatrix} I_Y & 0 \\ 0 & 1 \end{bmatrix}$ , where  $I_Y$  is the inclusion of  $\mathbf{Y}$  in  $\mathbf{L}_2(1)$ . The inequality in (3.19) cannot be verified unless the phase is fixed using the inclusion map. There would otherwise be a zero eigenvalue in the first term of (3.19) that forces  $\tau$  to be zero.

*Remark 4.* The parameter  $\tau$  in (3.19) is a measure on how much uncertainty (size of  $\mathcal{Z}$  and perturbation of  $H$ ) can be tolerated by the nominal limit cycle solution. It can be interpreted as a stability margin, which should be as large as possible for a given set description in (D1).

*Remark 5.* The operator inequality (3.19) can be truncated and verified as a linear matrix inequality in the frequency domain. We discuss this in more detail in the next section, where we also show how relaxation techniques from robust control can be used to integrate (3.19) and (3.20) into a parameterized operator inequality of form similar to (3.19).

*Proof.* Let us first consider (iv)(a). We have  $(\bar{z} = \begin{bmatrix} z \\ 1 \end{bmatrix}) \in \mathbf{Z} \times \mathbf{R}$  and likewise for  $\bar{z}_0$

$$\begin{aligned} \sigma(H[T, \theta]y - H[T_0, 0]y_0, y - y_0) &= \left\langle \begin{bmatrix} \tilde{H} + \tilde{\Delta}[T, \theta] & w_\Delta \\ \begin{bmatrix} I & 0 \end{bmatrix} & 0 \end{bmatrix} (\bar{z} - \bar{z}_0), \Psi \begin{bmatrix} \tilde{H} + \tilde{\Delta}[T, \theta] & w_\Delta \\ \begin{bmatrix} I & 0 \end{bmatrix} & 0 \end{bmatrix} (\bar{z} - \bar{z}_0) \right\rangle \\ &= \langle z - z_0, \Pi^* \Psi \Pi (z - z_0) \rangle + \langle \bar{z} - \bar{z}_0, \tilde{\Upsilon}_1[T, \theta] (\bar{z} - \bar{z}_0) \rangle \\ &= \sigma_{\Pi^* \Psi \Pi}(z - z_0) + \sigma_{\tilde{\Upsilon}_1}(\bar{z} - \bar{z}_0), \end{aligned}$$

where  $\tilde{\Upsilon}_1 = \tilde{\Psi}_Z^{1/2} \Upsilon_1 \tilde{\Psi}_Z^{1/2}$  and where  $\tilde{\Psi}_Z = \text{diag}(\Psi_Z, 1)$ .

By using this representation of the quadratic form, we get the following implications:

$$\begin{aligned} \sup_{z \in \partial \mathcal{Z}} \sigma(H[T, \theta]y - H[T_0, 0]y_0, y - y_0) &= \sup_{z \in \partial \mathcal{Z}} \sigma_{\Pi^* \Psi \Pi}(z - z_0) + \sigma_{\tilde{\Upsilon}_1}(\bar{z} - \bar{z}_0) \\ &\leq \sup_{z \in \partial \mathcal{Z}} \sigma_{\Pi^* \Psi \Pi}(z - z_0) + 2 \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta]), \end{aligned}$$

where we used that the second term is bounded by

$$\begin{aligned} \sup_{z \in \partial \mathcal{Z}} \sigma_{\tilde{\Upsilon}_1}(\bar{z} - \bar{z}_0) &= \sup_{z \in \partial \mathcal{Z}} \langle \bar{z} - \bar{z}_0, \tilde{\Upsilon}_1(\bar{z} - \bar{z}_0) \rangle \\ &= \sup_{\|\tilde{\Psi}_Z^{1/2}(\bar{z} - \bar{z}_0)\|^2=1} \langle \tilde{\Psi}_Z^{1/2}(\bar{z} - \bar{z}_0), \Upsilon_1 \tilde{\Psi}_Z^{1/2}(\bar{z} - \bar{z}_0) \rangle \\ &\leq \sup_{\|\tilde{\Psi}_Z^{1/2}(\bar{z} - \bar{z}_0)\|^2=2} \langle \tilde{\Psi}_Z^{1/2}(\bar{z} - \bar{z}_0), \Upsilon_1 \tilde{\Psi}_Z^{1/2}(\bar{z} - \bar{z}_0) \rangle \\ &\leq 2 \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta]). \end{aligned}$$

By Lagrange relaxation (S-procedure relaxation) we can further simplify:<sup>3</sup>

$$\begin{aligned}
 & \sup_{z \in \partial \mathcal{Z}} \sigma_{\Pi^* \Psi \Pi}(z - z_0) + 2 \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta]) \\
 &= \inf_{\tau \in \mathbf{R}} \sup_{z \in \mathbf{Z}} \sigma_{\Pi^* \Psi \Pi}(z - z_0) + \tau \sigma_Z(z - z_0) + 2 \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta]) \\
 &= \inf_{\tau \in \mathbf{R}} \sup_{z \in \mathbf{Z}} \langle z - z_0, (\Pi^* \Psi \Pi + \tau \Psi_Z)(z - z_0) \rangle + 2 \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta]) - \tau \\
 &\leq 2 \sup_{\theta \in [0, 1]} \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta]) - \tau \leq -\epsilon,
 \end{aligned}$$

where in the second last inequality we used (3.19) and the last inequality follows from (3.20). This implies (iv)(a).

We next consider condition (iv)(b). For this we need to introduce  $\Upsilon_2[T] = \Omega_2[T] + \Omega_2[T]^*$ , where

$$\Omega_2[T] = \begin{bmatrix} \tilde{\Delta}[T, 0] \Psi_Z^{-1/2} \\ 0 \end{bmatrix}^* \Psi \begin{bmatrix} \tilde{H} + \frac{1}{2} \tilde{\Delta}[T, 0] \Psi_Z^{-1/2} \\ [\Psi_Y^{-1/2} \quad 0] \end{bmatrix}.$$

We have

$$\begin{aligned}
 & \sigma(H[T, 0]y - H[T_0, 0]y_0, y - y_0) \\
 &= \left\langle \begin{bmatrix} \tilde{H} + \tilde{\Delta}[T, 0] \\ I \quad 0 \end{bmatrix} (z - z_0), \Psi \begin{bmatrix} \tilde{H} + \tilde{\Delta}[T, 0] \\ I \quad 0 \end{bmatrix} (z - z_0) \right\rangle \\
 &= \langle z - z_0, (\Pi^* \Psi \Pi + \tilde{\Upsilon}_2[T])(z - z_0) \rangle \\
 &= \sigma_{\Pi^* \Psi \Pi}(z - z_0) + \sigma_{\tilde{\Upsilon}_2}(z - z_0),
 \end{aligned}$$

where  $\tilde{\Upsilon}_2 = \Psi_Z^{1/2} \Upsilon_2 \Psi_Z^{1/2}$ .

Now let  $\mathcal{Z}(\eta) = \{z \in \mathbf{Z} : \sigma_Z(z - z_0, \eta) = 0\}$ , where  $\sigma_Z(z, \eta) = \langle z, \Psi_Z z \rangle - \eta$  and  $\eta \in [0, 1]$ . By using this representation of the quadratic form, we get the following inequality:

$$\begin{aligned}
 \sup_{z \in \mathcal{Z}(\eta)} \sigma(H[T, 0]y - H[T_0, 0]y_0, y - y_0) &= \sup_{z \in \mathcal{Z}(\eta)} \sigma_{\Pi^* \Psi \Pi}(z - z_0) + \sigma_{\tilde{\Upsilon}_2}(z - z_0) \\
 &\leq \sup_{z \in \mathcal{Z}(\eta)} \sigma_{\Pi^* \Psi \Pi}(z - z_0) + \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_2[T])\eta,
 \end{aligned}$$

where we used that the last term can be bounded by

$$\begin{aligned}
 \sup_{z \in \mathcal{Z}(\eta)} \sigma_{\tilde{\Upsilon}_2}(z - z_0) &= \sup_{\|\Psi_Z^{1/2}(z - z_0)\|^2 = \eta} \langle \Psi_Z^{1/2}(z - z_0), \Upsilon_2 \Psi_Z^{1/2}(z - z_0) \rangle \\
 &= \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_2[T])\eta.
 \end{aligned}$$

<sup>3</sup>The first implication is an equivalence since this is a relaxation of one quadratic form defined over a real space [20].

Next we use the S-procedure lossless condition in [20], which gives

$$\begin{aligned}
 & \sup_{z \in \mathcal{Z}(\eta)} \sigma_{\Pi^* \Psi \Pi}(z - z_0) + \lambda_{\max}(\Upsilon_2[T])\eta \\
 &= \inf_{\tau \in \mathbf{R}} \sup_{z \in \mathbf{Z}} \sigma_{\Pi^* \Psi \Pi}(z - z_0) + \tau \sigma_Z(z - z_0, \eta) + \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_2[T])\eta \\
 &= \inf_{\tau \in \mathbf{R}} \sup_{z \in \mathbf{Z}} \langle z - z_0, (\Pi^* \Psi \Pi + \tau \Psi_Z)(z - z_0) \rangle + \sup_{T \in \mathcal{T}} (\lambda_{\max}(\Upsilon_2[T]) - \tau)\eta \\
 &\leq \left( \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_2[T]) - \tau \right) \eta \leq -\epsilon \eta \leq -\epsilon \lambda_{\min}(\Psi_Z) \|z - z_0\|^2,
 \end{aligned}$$

where we used (3.19) in the second inequality. In the third inequality we use (3.20) and that  $\sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_2[T]) = \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, 0]) \leq 2 \sup_{\theta \in [0, 1]} \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta])$  (because  $\sup_{\theta} \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta]) \geq 0$ ). Finally, the last inequality follows since

$$\eta = \langle z - z_0, \Psi_Z(z - z_0) \rangle \geq \lambda_{\min}(\Psi_Z) \|z - z_0\|^2 \quad \forall z \in \mathcal{Z}(\eta).$$

This implies (iv)(b).  $\square$

**3.4. The set inclusion.** Estimates of  $\mathcal{W}$  can be obtained in several ways. It is generally useful to have bounds in terms of the  $C(1)$ -topology, and this is achieved using the next result.

**PROPOSITION 3.5.** *Suppose that  $\mathcal{Z}$  is characterized as in (D1). We have*

$$H[T, \theta]y \in \mathcal{W} = \{w \in C(1) : |w(t) - w_0(t)| \leq \nu(t)\}$$

for all  $(y, T) \in \text{cl } \mathcal{Z}$  and  $\theta \in [0, 1]$ , where

$$\nu(t) = \|\mathbf{1}(e^{j2\pi t}) \hat{H} \hat{\Psi}_Z^{-1/2}\| + \sup_{\theta \in [0, 1], T \in \mathcal{T}} \{\|\tilde{\Delta}[T, \theta] \Psi_Z^{-1/2}\|_{\mathbf{L}_2(1) \times \mathbf{R} \rightarrow C(1)} + \|w_{\Delta}[T, \theta]\|_{C(1)}\}$$

and where  $\hat{H}$  is the frequency domain representation of the operator defined in (3.14),  $\tilde{\Delta}$  and  $w_{\Delta}$  are defined in (3.15)–(3.16),  $\mathcal{T}$  is defined in (3.4), and

$$(3.21) \quad \mathbf{1}(e^{j2\pi t}) = \begin{bmatrix} \dots & e^{j4\pi t} I & e^{j2\pi t} I & I & e^{-j2\pi t} I & e^{-j4\pi t} I & \dots \end{bmatrix}.$$

If  $\hat{\Psi}_Z$  is diagonal, then we can use

$$(3.22) \quad \|\tilde{\Delta}[T, \theta] \Psi_Z^{-1/2}\|_{\mathbf{L}_2(1) \times \mathbf{R} \rightarrow C(1)} \leq \sqrt{\sum_{k=-\infty}^{\infty} |\Delta(j2\pi k/T, \theta) \hat{\Psi}_Y^{-1/2}(k, k)|^2 + \|\delta[T] \psi_T^{-1/2}\|_{C(1)}^2}.$$

*Proof.* We use that  $w(t) = \mathbf{1}(e^{j2\pi t})\hat{w}$ , and therefore we have

$$\begin{aligned}
 |w(t) - w_0(t)| &= |\mathbf{1}(e^{j2\pi t})(\hat{H}[T, \theta]\hat{y} - \hat{H}[T_0, 0]\hat{y}_0)| \\
 &= \left| \mathbf{1}(e^{j2\pi t}) \hat{H}(\hat{z} - \hat{z}_0) + \mathbf{1}(e^{j2\pi t}) \hat{\Delta}[T, \theta](\hat{z} - \hat{z}_0) + \mathbf{1}(e^{j2\pi t}) \hat{w}_{\Delta} \right| \\
 &\leq \|\mathbf{1}(e^{j2\pi t}) \hat{H} \hat{\Psi}_Z^{-1/2}\| \cdot \|\hat{\Psi}_Z^{1/2}(\hat{z} - \hat{z}_0)\| \\
 &\quad + \|\tilde{\Delta}[T, \theta] \Psi_Z^{-1/2}\|_{\mathbf{L}_2(1) \times \mathbf{R} \rightarrow C(1)} \|\Psi_Z^{1/2}(z - z_0)\| + |\mathbf{1}(e^{j2\pi t}) \hat{w}_{\Delta}| \\
 &\leq \|\mathbf{1}(e^{j2\pi t}) \hat{H} \hat{\Psi}_Z^{-1/2}\| + \|\tilde{\Delta}[T, \theta] \Psi_Z^{-1/2}\|_{\mathbf{L}_2(1) \times \mathbf{R} \rightarrow C(1)} + \|w_{\Delta}\|_{C(1)},
 \end{aligned}$$

where we used that  $\|\Psi_Z^{1/2}(z - z_0)\| \leq 1$  by the definition of  $\mathcal{Z}$  in (3.2).  $\square$

**4. Algorithm for robustness analysis.** Here we will assemble the results of the previous sections into an algorithm that allows us to verify the robustness of limit cycles. From Theorems 3.2 and 3.4 we get the following algorithm:

- (R1) Verify that  $F'_z(z_0, 0)$  has a bounded inverse.
- (R2) Verify that  $H[T, \theta]y \in \mathcal{W}$  for all  $(y, T) \in \text{cl } \mathcal{Z}$ ,  $\theta \in [0, 1]$ .
- (R3) Determine the sector bound (iii) in Theorem 3.2 with  $\sigma$  on the form (3.9)–(3.10).
- (R4a) Solve  $\max_{\tau > 0}$  such that  $I_Z^*(\Pi^* \Psi \Pi + \tau \Psi_Z)I_Z \leq 0$ .
- (R4b) Verify that  $\tau > 2 \sup_{\theta \in [0, 1]} \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta])$ .

Given (RA1)–(RA4) we can conclude that for each  $\theta \in [0, 1]$  there exists a solution  $(y_\theta, T_\theta) \in \mathcal{Z}$  of  $y = \Phi(H[T, \theta]y)$ .

We will next discuss how the steps of the algorithm can be verified numerically. For the first step we may use Proposition 4 in [8]. For the second step we use Proposition 3.5. An application will be illustrated in our example in the next section. The third step is simply to determine  $\alpha, \beta \in C(1)$  such that the time-varying sector bound in (3.7) holds. The last two steps are more involved and will be discussed in detail in the next two subsections. Using (R4a)–(R4b) may lead to unnecessary conservatism, and as an alternative we will also show that (R3)–(R4) can be replaced by the following:

- (R4') Use quadratic relaxation of the uncertainties to verify condition (iv) in Theorem 3.2.

**4.1. Verifying the operator inequality in (R4a).** The operator inequality in step (R4) is infinite dimensional and must be truncated in order to be verified. This is possible due to the compactness of  $H$  and since the sector condition in (3.8) can be defined by functions  $\alpha$  and  $\beta$  of limited frequency content. The inequality has a frequency domain representation, and for the computation we truncate the high frequency contribution to obtain a finite dimensional inequality. The Schur complement formula together with estimates of the high frequency contributions can be used to derive conditions under which the finite dimensional formula gives a lower bound on  $\tau$  in step (R4a).

We next define a high versus low frequency decomposition of each operator according to the following definitions and notation, where  $I_N$  and  $I_{\bar{N}}$  are the identity operators on  $P_N \mathbf{L}_2(1)$  and  $Q_N \mathbf{L}_2(1)$ , respectively, and where we assume the phase is fixed according to

$$(4.1) \quad \mathbf{Y} = \{y \in \mathbf{L}_2(1) : \text{re } e_1^T \hat{y}[1] = 0\},$$

where  $e_1$  is the first unit vector.

- (F1) For the inclusion map  $I_Z : Q_N \mathbf{L}_2(1) \oplus \mathbf{Y}_N \times \mathbf{R} \rightarrow Q_N \mathbf{L}_2(1) \oplus P_N \mathbf{L}_2(1) \times \mathbf{R}$  we use the block structure

$$(4.2) \quad I_Z = \begin{bmatrix} I_{\bar{N}} & 0 \\ 0 & I_{Z_N} \end{bmatrix}, \quad \text{where} \quad I_{Z_N} = \begin{bmatrix} I_{Y_N} & 0 \\ 0 & 1 \end{bmatrix},$$

where  $I_{Y_N}$  is the inclusion map from  $\mathbf{Y}_N = P_N \mathbf{Y}$  into  $P_N \mathbf{L}_2(1)$ .

- (F2) The operator  $\Psi_Y$  defined in (D1) has the low versus high frequency block decomposition  $\Psi_Y : Q_N \mathbf{L}_2(1) \oplus P_N \mathbf{L}_2(1) \rightarrow Q_N \mathbf{L}_2(1) \oplus P_N \mathbf{L}_2(1)$ ,

$$(4.3) \quad \Psi_Y = \begin{bmatrix} \psi_{Y_{\bar{N}}} I_{\bar{N}} & 0 \\ 0 & \Psi_{Y_N} \end{bmatrix},$$

and we let

$$(4.4) \quad \Psi_{Z_N} = \begin{bmatrix} \Psi_{Y_N} & 0 \\ 0 & \psi_T \end{bmatrix}.$$

(F3) We use the following block decompositions:

$$\begin{aligned} H[T_0, 0] &= \begin{bmatrix} H_{\bar{N}}[T_0, 0] & 0 \\ 0 & H_N[T_0, 0] \end{bmatrix}, \\ D_T H[T_0, 0] &= \begin{bmatrix} D_T H_{\bar{N}}[T_0, 0] & 0 \\ 0 & D_T H_N[T_0, 0] \end{bmatrix}, \\ \Pi_N &= \begin{bmatrix} H_N[T_0, 0] & D_T H_N[T_0, 0] y_{0,N} \\ I_N & 0 \end{bmatrix}, \end{aligned}$$

where  $H_N = I_{P_N \mathbf{L}_2(1)}^* H I_{P_N \mathbf{L}_2(1)}$  and  $H_{\bar{N}} = I_{Q_N \mathbf{L}_2(1)}^* H I_{Q_N \mathbf{L}_2(1)}$ , and similarly for  $D_T H_{\bar{N}}[T_0, 0]$  and  $D_T H_N[T_0, 0]$ . Finally,  $y_{0,N} = P_N y_0$ .

(F4) We let the blocks of  $\Psi$  in (3.10) have the high versus low frequency decompositions

$$\Psi_{11} = \begin{bmatrix} \Psi_{11, \bar{N}\bar{N}} & \Psi_{11, \bar{N}N} \\ \Psi_{11, N\bar{N}} & \Psi_{11, NN} \end{bmatrix}, \quad \Psi_{12} = \begin{bmatrix} \Psi_{12, \bar{N}\bar{N}} & \Psi_{12, \bar{N}N} \\ \Psi_{12, N\bar{N}} & \Psi_{12, NN} \end{bmatrix},$$

where

$$\begin{aligned} \Psi_{11, NN} &= I_{P_N \mathbf{L}_2(1)}^* \Psi_{11} I_{P_N \mathbf{L}_2(1)}, \\ \Psi_{11, N\bar{N}} &= I_{P_N \mathbf{L}_2(1)}^* \Psi_{11} I_{Q_N \mathbf{L}_2(1)}, \\ \Psi_{11, \bar{N}N} &= I_{Q_N \mathbf{L}_2(1)}^* \Psi_{11} I_{P_N \mathbf{L}_2(1)}, \\ \Psi_{11, \bar{N}\bar{N}} &= I_{Q_N \mathbf{L}_2(1)}^* \Psi_{11} I_{Q_N \mathbf{L}_2(1)}, \end{aligned}$$

and analogously for the blocks in defining the decomposition of  $\Psi_{12}$ . Finally, let  $\Psi_N = I_{P_N \mathbf{L}_2(1)}^* \Psi I_{P_N \mathbf{L}_2(1)}$ .

It is straightforward to see that

$$(4.5) \quad I_{\mathbf{Z}}^* (\Pi^* \Psi \Pi + \tau \Psi_Z) I_{\mathbf{Z}} = \begin{bmatrix} \Delta_{11} + (\tau \psi_{Y_{\bar{N}}} - 1) I_{\bar{N}} & \Delta_{12} \\ \Delta_{12}^* & I_{\mathbf{Z}_N}^* (\Pi_N^* \Psi_N \Pi_N + \tau \Psi_{Z_N}) I_{\mathbf{Z}_N} + \Delta_{22} \end{bmatrix},$$

where

$$\begin{aligned} \Delta_{11} &= H_{\bar{N}}^* \Psi_{11, \bar{N}\bar{N}} H_{\bar{N}} + H_{\bar{N}}^* \Psi_{12, \bar{N}\bar{N}} + \Psi_{12, \bar{N}\bar{N}}^* H_{\bar{N}}, \\ \Delta_{12} &= \left[ (H_{\bar{N}}^* \Psi_{11, \bar{N}N} + \Psi_{12, \bar{N}N}^*) H_N I_{\mathbf{Y}_N} + H_{\bar{N}}^* \Psi_{12, \bar{N}N} I_{\mathbf{Y}_N} \quad (H_{\bar{N}}^* \Psi_{11, \bar{N}} + \Psi_{12, \bar{N}}^*) D_T H y_0 \right], \\ \Delta_{22} &= \begin{bmatrix} 0 & (H_N^* \Psi_{11, N\bar{N}} + \Psi_{12, N\bar{N}}^*) D_T H_{\bar{N}} y_{0, \bar{N}} \\ (\cdot)^* & y_0^* D_T H[T_0, 0]^* \Psi_{11} D_T H[T_0, 0] y_0 - y_{0,N}^* D_T H_N[T_0, 0]^* \Psi_{11, NN} D_T H_N[T_0, 0] y_{0,N} \end{bmatrix} \end{aligned}$$

and where in  $\Delta_{12}$  we used the notation

$$\Psi_{11, \bar{N}} = \begin{bmatrix} \Psi_{11, \bar{N}\bar{N}} & \Psi_{11, \bar{N}N} \end{bmatrix}, \quad \Psi_{12, \bar{N}}^* = \begin{bmatrix} \Psi_{12, \bar{N}\bar{N}}^* & \Psi_{12, \bar{N}N}^* \end{bmatrix}.$$

Using the above definitions we can state the following result.

PROPOSITION 4.1. *Suppose there exist  $\epsilon > 0$  and  $\rho \in (0, 1 - \|\Delta_{11}\|)$  such that*

- (i)  $\tau = (1 - \rho - \|\Delta_{11}\|)/\psi_{Y_N}$ ;
- (ii)  $\epsilon > \|\Delta_{22}\| + \|\Delta_{12}\|^2/\rho$ ;
- (iii)  $I_{\mathbf{Z}_N}^*(\Pi_N^* \Psi_N \Pi_N + \tau \Psi_{\mathbf{Z}_N}) I_{\mathbf{Z}_N} \leq -\epsilon$ .

Then  $I_{\mathbf{Z}}^*(\Pi^* \Psi \Pi + \tau \Psi_{\mathbf{Z}}) I_{\mathbf{Z}} \leq 0$ .

*Proof.* The upper left block of (4.5) satisfies

$$\Delta_{11} + (\tau \psi_{Y_N} - 1) I_N \leq -\rho I_N$$

by (i). We can therefore apply the Schur complements formula and conclude that  $I_{\mathbf{Z}}^*(\Pi^* \Psi \Pi + \tau \Psi_{\mathbf{Z}}) I_{\mathbf{Z}} < 0$  since

$$\begin{aligned} I_{\mathbf{Z}_N}^*(\Pi_N^* \Psi_N \Pi_N + \tau \Psi_{\mathbf{Z}_N}) I_{\mathbf{Z}_N} + \Delta_{22} - \Delta_{12}^* (\Delta_{11} + (\tau \psi_{Y_N} - 1) I_N)^{-1} \Delta_{12} \\ \leq -\epsilon + \|\Delta_{22}\| + \|\Delta_{12}\|^2/\rho < 0. \quad \square \end{aligned}$$

The inequality in (iii) can be verified in the frequency domain as the finite dimensional linear matrix inequality

$$(4.6) \quad \hat{I}_{\mathbf{Z}_N}^* (\hat{\Pi}_N^* \hat{\Psi}_N \hat{\Pi}_N + \tau \hat{\Psi}_{\mathbf{Z}_N}) \hat{I}_{\mathbf{Z}_N} \leq -\epsilon \hat{I}_N,$$

where  $\hat{I}_N$  is a unit matrix of dimension  $2N + 1$ ,

$$\hat{\Pi}_N = \begin{bmatrix} \hat{H}_N[T_0, 0] & D_T \hat{H}_N[T_0, 0] \hat{y}_{0,N} \\ \hat{I}_N & 0 \end{bmatrix},$$

where  $\hat{y}_{0,N}$  is the Fourier transform of  $y_0$  truncated at its lowest  $N$  frequencies,

$$(4.7) \quad \hat{H}_N[T, 0] = \text{diag}(H(j2k\pi N/T, 0) : k = N, \dots, -N),$$

and analogously for  $D_T \hat{H}_N[T_0, 0]$ . Further,  $\hat{\Psi}_{\mathbf{Z}_N}$  is the frequency domain representation of  $\Psi_{\mathbf{Z}_N}$  in (4.4),  $\hat{I}_{\mathbf{Z}_N}$  is the frequency domain representation of the inclusion map  $I_{\mathbf{Z}_N}$  defined in (4.2), and, finally,  $\hat{\Psi}_N$  is the  $2N + 1 \times 2N + 1$  block Toeplitz representation of  $\Psi_N$ . It has the form

$$\hat{\Psi}_N = \begin{bmatrix} -P_N \left( \frac{1}{2} (\mathcal{B}^* \mathcal{A} + \mathcal{A}^* \mathcal{B}) \right) P_N & \frac{1}{2} P_N (\mathcal{A} + \mathcal{B}) P_N \\ \frac{1}{2} P_N (\mathcal{A} + \mathcal{B})^* P_N & -I_N \end{bmatrix},$$

where  $\mathcal{B} = \text{Toep}[\beta]$ ,  $\mathcal{A} = \text{Toep}[\alpha]$  are (bi-infinite) block Toeplitz matrices defined as in (1.2).

The maximal value of the parameter  $\tau$  for which (4.6) is satisfied can be computed as

$$\tau = -\lambda_{\min}((\hat{I}_{\mathbf{Z}_N}^* \hat{\Psi}_{\mathbf{Z}_N} \hat{I}_{\mathbf{Z}_N})^{-1/2} (\hat{I}_{\mathbf{Z}_N}^* \hat{\Pi}_N^* \hat{\Psi}_N \hat{\Pi}_N \hat{I}_{\mathbf{Z}_N} + \epsilon \hat{I}_N) (\hat{I}_{\mathbf{Z}_N}^* \hat{\Psi}_{\mathbf{Z}_N} \hat{I}_{\mathbf{Z}_N})^{-1/2}).$$

Note that all norms  $\|\Delta_{k,l}\|$ ,  $k, l = 1, 2$ , in Proposition 4.1 can be made arbitrarily small by choosing  $N$  large enough compared to the frequency content in  $\Psi$  and the roll-off decay of  $H$ . Hence, if (iii) is satisfied, then the remaining inequalities in Proposition 4.1 will be satisfied if  $N$  is large enough.

**4.2. Verifying the perturbation bound.** A straightforward calculation shows that

$$\Upsilon_1[T, \theta] = \begin{bmatrix} \Upsilon_{11}[T, \theta] & \Upsilon_{12}[T, \theta] & \Upsilon_{13}[T, \theta] \\ \Upsilon_{12}^*[T, \theta] & \Upsilon_{22}[T, \theta] & \Upsilon_{23}[T, \theta] \\ \Upsilon_{13}^*[T, \theta] & \Upsilon_{23}^*[T, \theta] & \Upsilon_{33}[T, \theta] \end{bmatrix},$$

where (here we suppress the dependence on  $T$  and  $\theta$ )

$$\begin{aligned} \Upsilon_{11} &= \Psi_Y^{-1/2}(\Delta^* \Psi_{11} \Delta + \text{Sym}(\Delta^* \Psi_{11} H) + \text{Sym}(\Delta^* \Psi_{12})) \Psi_Y^{-1/2}, \\ \Upsilon_{12} &= \Psi_Y^{-1/2}(\Delta^* \Psi_{11}(D_T H + \delta) + (H^* \Psi_{11} + \Psi_{12}^*) \delta) \psi_T^{-1/2}, \\ \Upsilon_{13} &= \Psi_Y^{-1/2}((\Delta^* + H^*) \Psi_{11} + \Psi_{12}^*) w_\Delta, \\ \Upsilon_{22} &= \text{Sym} \left( \psi_T^{-1/2} \delta^* \Psi_{11} \left( D_T H + \frac{1}{2} \delta \right) \psi_T^{-1/2} \right), \\ \Upsilon_{23} &= \psi_T^{-1/2}(\delta + D_T H)^* \Psi_{11} w_\Delta, \\ \Upsilon_{33} &= w_\Delta^* \Psi_{11} w_\Delta, \end{aligned}$$

where  $\text{Sym}(H) = H + H^*$  and  $\Delta, \delta$  are defined in (3.17). We will use the following result.

PROPOSITION 4.2.

$$\sup_{\theta \in [0,1]} \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_1[T, \theta]) \leq \lambda_{\max} \left( \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{12} & \gamma_{22} & \gamma_{23} \\ \gamma_{13} & \gamma_{23} & \gamma_{33} \end{bmatrix} \right),$$

where

$$\begin{aligned} \gamma_{kk} &= \sup_{\theta \in [0,1]} \sup_{T \in \mathcal{T}} \lambda_{\max}(\Upsilon_{kl}[T, \theta]), \quad k = 1, 2, 3, \\ \gamma_{kl} &= \sup_{\theta \in [0,1]} \sup_{T \in \mathcal{T}} \|\Upsilon_{kl}[T, \theta]\|, \quad k \neq l. \end{aligned}$$

*Proof.* Note that  $\Upsilon_1 : \mathbf{L}_2(1) \times \mathbf{R}^2 \rightarrow \mathbf{L}_2(1) \times \mathbf{R}^2$ . An arbitrary unit length vector  $v \in \mathbf{L}_2(1) \times \mathbf{R}^2$  can be decomposed as  $v = \sum_{k=1}^3 \alpha_k \bar{v}_k$ , where  $\bar{v}_1 = (v_1, 0, 0)$  with  $v_1 \in \mathbf{L}_2(1)$  and  $\|v_1\| = 1$ , and  $\bar{v}_2 = (0, v_2, 0)$ ,  $\bar{v}_3 = (0, 0, v_3)$  with  $v_2, v_3 \in [-1, 1]$ , and the real valued scalars  $\alpha_k$  satisfy  $\sum_{k=1}^3 \alpha_k^2 = 1$ . It follows that

$$\begin{aligned} \lambda_{\max}(\Upsilon_1) &= \sup_{\|v\|_{\mathbf{L}_2(1) \times \mathbf{R}^2} = 1} \langle v, \Upsilon_1 v \rangle \\ &= \sup_{|\alpha|=1} \sup_{\|v_1\|=1} \sup_{v_2, v_3 \in [-1, 1]} \sum_{k=1}^3 \sum_{l=1}^3 \alpha_k \alpha_l \langle v_k, \Upsilon_{kl} v_l \rangle \\ &\leq \sup_{|\alpha|=1} \alpha_k \alpha_l \gamma_{kl} \\ &= \lambda_{\max}(\Gamma), \end{aligned}$$

where  $\Gamma = [\gamma_{kl}]_{k,l=1}^3$ . This concludes the proof.  $\square$

The coefficients needed in Proposition 4.2 can easily be estimated. To avoid unnecessary conservatism, it is important to explore that  $\Psi_{11}$  is usually not positive definite, and simple minded norm estimates should therefore be avoided.



**4.3. Relaxation of uncertainties using quadratic forms.** Here we will discuss how condition (iv) in Theorem 3.2 can be verified using relaxation techniques from robust control. This provides an alternative means of verifying steps (R4a)–(R4b) in the algorithm for robustness analysis, which often is less conservative but computationally more demanding.

We assume that the nonlinearity satisfies the quadratic form (3.18), where  $\Psi$  is defined as in (F4). By using the representation in (3.14)–(3.16), we obtain the following equivalent representation of (3.13) as a linear fractional transformation [23]:

$$(4.8) \quad \begin{bmatrix} \delta w \\ v \end{bmatrix} = \underbrace{\begin{bmatrix} \tilde{H} & II_{\Delta} & I_W \\ I & 0 & 0 \end{bmatrix}}_G \begin{bmatrix} \delta z \\ u \\ \check{w}_{\Delta} \end{bmatrix},$$

$$(4.9) \quad u = \underbrace{\begin{bmatrix} I_{\Delta}^{-1} \Delta[T, \theta] & 0 \\ 0 & I_{\delta}^{-1} \delta[T] \end{bmatrix}}_{\check{\Delta}} v,$$

$$(4.10) \quad II_{\Delta} = \begin{bmatrix} I_{\Delta} & I_{\delta} \end{bmatrix},$$

where  $\check{w}_{\Delta} = I_W^{-1} w_{\Delta}$ ,  $\delta z = z - z_0$ , and  $\delta w = H[T, \theta]y - H[T_0, 0]y_0$ . The operators  $I_{\Delta}$ ,  $I_{\delta}$ , and  $I_W$  are introduced in order to create roll off so that the same computational ideas as in Proposition 4.1 can be used. In order to apply Proposition 4.1 for truncation of the operator inequality in (ii)(a) in Proposition 4.3 below we may assume that the operators act through multiplication in the frequency domain according to  $\hat{u}[k] = \hat{I}_{\Delta}[k]\hat{v}[k]$ , where

$$(4.11) \quad \hat{I}_{\Delta}[k] = \begin{cases} I, & |k| \leq N, \\ \nu, & |k| > N, \end{cases}$$

and where  $N$  is a large integer and  $\nu$  is a small number.  $I_{\delta}$  and  $I_W$  may be defined analogously. For the discussion up to the next proposition including its proof one may assume that these operators are equal to the identity.

We will use quadratic constraints to characterize the uncertainties in the system according to the following definitions:

1. The uncertain operator  $\check{\Delta}[T, \theta]$  defined in (4.9) is assumed to satisfy the quadratic inequality

$$(4.12) \quad \sigma_{\Delta}(v, \check{\Delta}[T, \theta]v, \lambda) \geq 0 \quad \forall \lambda \in \Lambda, \theta \in [0, 1], T \in \mathcal{T},$$

where  $\Lambda$  is a convex cone of parameters. The parameterization can be used to characterize the structure of  $\check{\Delta}$  as in  $\mu$ -analysis [15, 3, 17]. We assume that the quadratic form has the representation

$$(4.13) \quad \sigma_{\Delta}(v, u, \lambda) = \left\langle \begin{bmatrix} v \\ u \end{bmatrix}, \Psi_{\Delta}(\lambda) \begin{bmatrix} v \\ u \end{bmatrix} \right\rangle, \quad \Psi_{\Delta} = \begin{bmatrix} \Psi_{\Delta, (11)} & \Psi_{\Delta, (12)} \\ \Psi_{\Delta, (12)}^* & \Psi_{\Delta, (22)} \end{bmatrix},$$

where  $\Psi_{\Delta}$  is assumed to be linear in  $\lambda$ .

2. The uncertain signal  $\check{w}_{\Delta} = I_W^{-1} w_{\Delta}$  satisfies the quadratic inequality

$$(4.14) \quad \sigma_{W_{\Delta}}(\check{w}_{\Delta}) = 1 - \langle \check{w}_{\Delta}, \Psi_{W_{\Delta}} \check{w}_{\Delta} \rangle \geq 0,$$

where the positive definite and self-adjoint operator  $\Psi_{W_\Delta}$  can be constructed as a diagonal operator with components that estimate the size of the frequency components of  $w_\Delta$ .

By using the standard S-procedure relaxation technique we obtain the following condition that can replace (iv) in Theorem 3.2.

PROPOSITION 4.3. *Condition (iv) in Theorem 3.2 is satisfied if there exist  $\tau_1, \tau_2 \geq 0$  and  $\lambda \in \Lambda$  such that*

- (a)  $I_Z^*(\check{\Pi}^* \Psi \check{\Pi} + \check{\Psi}_Z(\tau_1, \tau_2, \lambda))I_Z \leq 0$ ,
- (b)  $\tau_1 > \tau_2$ ,

where

$$\begin{aligned} \check{\Pi} &= \begin{bmatrix} \check{H} \\ I & 0 \end{bmatrix} := \begin{bmatrix} H[T_0, 0] & [D_T H[T_0, 0]y_0 & II_\Delta & I_W] \\ I & \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \end{bmatrix}, \\ \check{\Psi}_Z(\tau_1, \tau_2, \lambda) &= \begin{bmatrix} \Psi_{\Delta,11}(\lambda) + \tau_1 \Psi_Z & \Psi_{\Delta,12}(\lambda) & 0 \\ \Psi_{\Delta,12}^*(\lambda) & \Psi_{\Delta,22}(\lambda) & 0 \\ 0 & 0 & -\tau_2 \Psi_{W_\Delta} \end{bmatrix}, \\ I_Z &= \begin{bmatrix} I_Z & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}, \end{aligned}$$

where  $\Psi$  is defined in (F4),  $\Psi_\Delta$  is defined in (4.13), and  $II_\Delta$  is defined in (4.10).

*Remark 6.* The same computational ideas as presented in Proposition 4.1 can be applied to verify (a). The parameters  $N$  and  $\nu$  in (4.11) must be chosen sufficiently large and small, respectively, so that the norms  $\|\Delta_{k,l}\|$  are small enough.

*Proof.* To verify (iv)(a) in Theorem 3.2 we consider the optimization problem  $(\delta z = (\delta y, \delta T) = (y - y_0, T - T_0))$ :

$$\begin{aligned} &\sup_{\delta z, u, \check{w}_\Delta} \sigma(\delta w, \delta y) \\ &\text{subject to } \begin{bmatrix} \delta w \\ v \end{bmatrix} = G \begin{bmatrix} \delta z \\ u \\ \check{w}_\Delta \end{bmatrix} \quad \text{such that} \quad \begin{cases} \sigma_\Delta(v, u, \lambda) \geq 0 & \forall \lambda \in \Lambda, \\ \sigma_Z(\delta z) = 0, \\ \sigma_{W_\Delta}(\check{w}_\Delta) \geq 0 \end{cases} \\ (4.15) \quad &\leq \inf_{\tau_1 \in \mathbf{R}, \tau_2 \geq 0, \lambda \in \Lambda} \sup_{\delta z, u, w_\Delta} \sigma(\tilde{H}\delta z + II_\Delta u + I_W w_\Delta, \delta y) + \sigma_\Delta(\delta z, u, \lambda) \\ &\quad + \tau_1 \sigma_Z(\delta z) + \tau_2 \sigma_{W_\Delta}(\check{w}_\Delta) \\ &= \inf_{\tau_1 \in \mathbf{R}, \tau_2 \geq 0, \lambda \in \Lambda} \sup_{\delta z, u, w_\Delta} \xi^* I_Z^* (\check{\Pi}^* \Psi \check{\Pi} + \check{\Psi}_Z(\tau_1, \tau_2, \lambda)) I_Z \xi - \tau_1 + \tau_2 \\ &\leq -\tau_1 + \tau_2 < 0, \end{aligned}$$

where  $\xi = [\delta z^* \quad u^* \quad \check{w}_\Delta^*]^*$  and  $G$  was defined in (4.8). The first inequality follows by S-procedure relaxation, the operator representation obtained after the equality follows by the definition of  $\check{\Pi}$  and  $\check{\Psi}_Z$ , and, finally, the last two inequalities follow by conditions (a)–(b) in the statement of the proposition.

Next we show that condition (iv)(b) in Theorem 3.2 follows from (a)–(b). We

need to introduce the operators

$$G_2 = \begin{bmatrix} \tilde{H} & II_\Delta \\ I & 0 \end{bmatrix},$$

$$\check{\Pi}_2 = \begin{bmatrix} H[T_0, 0] & [D_T H[T_0, 0]y_0 & II_\Delta] \\ I & \begin{bmatrix} 0 & 0 \end{bmatrix} \end{bmatrix},$$

$$\check{\Psi}_{\check{Z}_2}(\tau, \lambda) = \begin{bmatrix} \Psi_{\Delta, 11}(\lambda) + \tau \Psi_Z & \Psi_{\Delta, 12}(\lambda) \\ \Psi_{\Delta, 12}^*(\lambda) & \Psi_{\Delta, 22}(\lambda) \end{bmatrix}, \quad I_{\check{Z}_2} = \begin{bmatrix} I_Z & 0 \\ 0 & I \end{bmatrix}$$

and let  $\mathcal{Z}(\eta) = \{z \in \mathbf{Z} : \sigma_Z(z - z_0, \eta) = 0\}$ , where  $\sigma_Z(z, \eta) = \langle z, \Psi_Z z \rangle - \eta$  and  $\eta \in [0, 1]$ . Again we consider S-procedure relaxation ( $\eta \in [0, 1]$ ):

$$\begin{aligned} & \sup_{z \in \mathcal{Z}(\eta)} \sigma(H[T, 0]y - H[T_0, 0]y_0, y - y_0) \\ & \leq \sup_z \sigma(\delta w, \delta y, \lambda) \\ & \quad \text{subject to } \begin{bmatrix} \delta w \\ v \end{bmatrix} = G_2 \begin{bmatrix} \delta z \\ u \end{bmatrix} \quad \text{such that} \quad \begin{cases} \sigma_\Delta(v, u, \lambda) \geq 0 & \forall \lambda \in \Lambda \\ \sigma_Z(\delta z, \eta) = 0 \end{cases} \\ & \leq \inf_{\tau, \lambda \in \Lambda} \sup_{\delta z, u} \sigma(\tilde{H}\delta z + II_\Delta u, \delta y, \lambda) + \sigma_\Delta(\delta z, u, \lambda) + \tau \sigma_Z(\delta z, \eta) \\ & = \inf_{\tau, \lambda \in \Lambda} \sup_{\delta z, u} \xi^* I_{\check{Z}_2}^* (\check{\Pi}_2^* \Psi_1 \check{\Pi}_2 + \check{\Psi}_{\check{Z}_2}(\tau, \lambda)) I_{\check{Z}_2} \xi - \tau \eta \\ & \leq -\tau \eta, \end{aligned}$$

where  $\xi = [\delta z^* \quad u^*]^*$ . Here the last inequality follows because condition (a) implies that  $I_{\check{Z}_2}^* (\check{\Pi}_2^* \Psi_1 \check{\Pi}_2 + \check{\Psi}_{\check{Z}_2}(\tau, \lambda)) I_{\check{Z}_2} \leq 0$ . In order to continue the proof we notice that the definition of  $\mathcal{Z}(\eta)$  implies that

$$\|\Psi_Z^{1/2}(z - z_0)\|^2 = \eta,$$

which by the above inequality implies that

$$\sigma(H[T, 0]y - H[T_0, 0]y_0, y - y_0) \leq -\tau \eta \leq -L \|z - z_0\|^2,$$

where  $L = \tau \lambda_{\min}(\Psi_Z) > 0$  since  $\tau > 0$  and  $\Psi_Z > 0$ . This proves condition (iv)(b).  $\square$

**5. Example.** Let us consider the Van der Pol oscillator

$$\ddot{w}(t) + m(w(t)^2 - 1)\dot{w}(t) + w(t) = \theta \Delta(\dot{w}).$$

To represent this system on the form (3.1) we introduce the new coordinates

$$\begin{aligned} x_1 &= -\dot{w} - m(w^3/3 - w), \\ x_2 &= w. \end{aligned}$$

The transformed system can be shown to have the form (3.1) with  $\varphi(w) = -mw^3/3 + (2 + m)w$  and

$$H(s, \theta) = (I - \theta H_0(s) \Delta(s))^{-1} H_0(s),$$

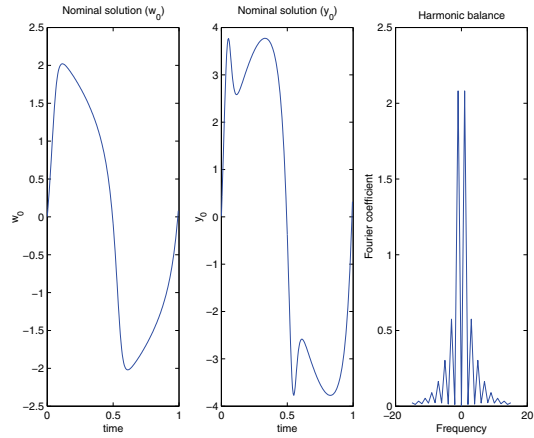


FIG. 5.1. The nominal limit cycle solution. The left-hand figure shows  $w_0(t)$ , and the middle figure shows  $y_0(t) = \varphi(w_0(t))$ . The right-hand figure shows the spectrum of the nominal solution  $|\hat{y}_0[k]|$ .

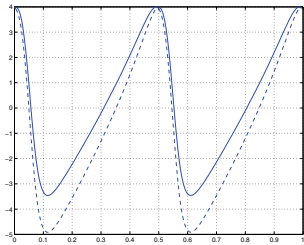


FIG. 5.2. The functions  $\alpha(t)$  and  $\beta(t)$  that define the low frequency sector condition.

where  $H_0(s) = \frac{s}{s^2+2s+1}$  and where  $\Delta$  is a stable transfer function with  $|\Delta(j\omega)| \leq \mu$  for all  $\omega \in \mathbf{R}$ . The nominal Van der Pol oscillator has a periodic solution with  $T_0 = 7.7$  and the corresponding 1-periodic trajectories in Figure 5.1.

We will use Theorem 3.2 to derive a bound on  $\mu$  for which the perturbed Van der Pol oscillator has a periodic solution in a neighborhood of  $(y_0, T_0)$  defined as in (D1) with  $\Psi_Y$  the diagonal operator,

$$\hat{\Psi}_Y[k, k] = \begin{cases} \psi_{Y_N} k, & |k| \leq 7, \\ \psi_{Y_{\bar{N}}}, & |k| > N, \end{cases}$$

where  $\psi_{Y_N} = 45$ , and  $\psi_{Y_{\bar{N}}} = 45$ . We let  $\psi_T = 45$ , which implies that  $\mathcal{T} = (7.55, 7.85)$ .

We first consider the case when  $\mu = 0.0082$  and verify (i)–(iv) in Theorem 3.2.

- (R1) It can be verified numerically that condition (i) in Theorem 3.2 is satisfied.
- (R2) We use Proposition 3.5. Some detailed information on the computation can be found in the appendix. We obtained the bound  $\|w - w_0\|_{C(1)} \leq 0.17$ .
- (R3) For  $\mathcal{W} = \{w \in C(1) : \|w - w_0\|_{C(1)} \leq 0.17\}$  we use a quadratic constraint on the form (3.9)–(3.10) with  $\alpha$  and  $\beta$  as in Figure 5.2.
- (R4) We use Proposition 4.3. We let the uncertainties be characterized by the

quadratic forms

$$\sigma_{\Delta}(v, u, \lambda) = \hat{v}^* \begin{bmatrix} \Lambda_1 I & 0 \\ 0 & \lambda_2 \end{bmatrix} \hat{v} - \hat{u}^* \begin{bmatrix} \hat{\Psi}_{\Delta}(\lambda_1) & 0 \\ 0 & \lambda_2 \gamma_{\delta} I \end{bmatrix} \hat{u},$$

where  $\lambda_2 \geq 0$  and  $\Lambda_1 = \text{diag}(\dots \lambda_{12}, \lambda_{1,1}, \lambda_{1,0}, \lambda_{1,-1}, \lambda_{1,-2}, \dots)$ , where  $\lambda_{1,k} = \lambda_{1,-k} \geq 0$  for all  $k$ . We use  $\hat{\Psi}_{\Delta} = \text{diag}(\lambda_{1k} \gamma_{\Delta}[k])_{k=-\infty}^{\infty}$ , with

$$\begin{aligned} \gamma_{\Delta}[k]^{-1/2} &= \sup_{T \in \mathcal{T}} \sup_{\theta \in [0,1]} |H(j2\pi k/T, \theta) - H_0(j2\pi k/T_0)| \\ &\leq \sup_{T \in \mathcal{T}} \frac{|H_0(j2\pi k/T) - H_0(j2\pi k/T_0)| + \mu |H_0(j2\pi k/T) H_0(j2\pi k/T_0)|}{1 - \mu |H_0(j2\pi k/T)|}, \end{aligned}$$

and we let

$$\begin{aligned} (5.1) \quad \gamma_{\delta}^{-1} &= \sup_{T \in \mathcal{T}} \|\delta[T]\|^2 \\ &= \sup_{T \in \mathcal{T}} \sum_{k=-\infty}^{\infty} \frac{|(H_0(j2\pi k/T) - H_0(j2\pi k/T_0)) - D_T H_0(j2\pi k/T_0)(T - T_0)) \hat{y}_0[k]|^2}{(T - T_0)^2}. \end{aligned}$$

For  $w_{\Delta}$  we use the quadratic constraint in (4.14) with  $\hat{\Psi}_{W_{\Delta}} = \gamma_w I_{2N+1}$ , where

$$\gamma_w^{-1} = \sup_{T \in \mathcal{T}} \sum_{k=-\infty}^{\infty} \frac{\mu^2 |H_0(j2\pi k/T)^2 \hat{y}_0[k]|^2}{(1 - \mu |H(j2\pi k/T)|)^2}.$$

The condition of Proposition 4.3 can be verified numerically as follows: With  $\tau_1 = 2.01 \cdot 10^{-4}$ ,  $\tau_2 = 2.00 \cdot 10^{-4}$ ,  $\lambda_{1,k} = 4.16 \cdot 10^{-4}$  for all  $k$ , and  $\lambda_2 = 0.0034$  we have

- (a)  $I_{\check{Z}}^*(\check{\Pi}^* \Psi \check{\Pi} + \check{\Psi}_{\check{Z}}(\tau_1, \tau_2, \lambda)) I_{\check{Z}} \leq 0$ , where we used the quadratic forms introduced above;
- (b)  $\tau_1 > \tau_2$ .

Further results can be found in Table 5.1, where the column for  $\|w - w_0\|_{C(1)}$  is the estimate the maximal perturbation of the trajectory obtained in (R2) and  $|T - T_0| = 1/\sqrt{\psi_T}$  is the maximum perturbation of the period time in the set  $\mathcal{Z}$ . The theoretical estimates compare fairly well with simulation results where we used

$$(5.2) \quad \Delta(s) = \mu_{sim} \omega_0^2 / (s^2 + 2\zeta \omega_0 s + \omega_0^2)$$

with  $\zeta = 0.2$  and  $\omega_0 = 0.83$ , which corresponds to  $\|\Delta\|_{\infty} \approx 0.01$  when  $\mu_{sim} = 0.004$ . Further simulations in Table 5.2 show that the period time is very sensitive to perturbations.

The case with  $\Delta$  as in (5.2) is a model of a mass and spring system excited by a Van der Pol oscillator; see Figure 5.3. The oscillator adapts its frequency to the resonance frequency of the mechanical system, and that is the reason why our analysis indicates large sensitivity to model uncertainty. The analysis agrees very well with the simulation for small perturbations. The upper bounds obtained from the simulations are only about 20% larger than the computed bounds.

TABLE 5.1

The table shows the performance of the algorithm for three different sizes of the perturbation. The last column indicates how large the perturbation is in relation to the  $\mathbf{H}_\infty$ -norm of the nominal dynamics.

$\Psi_{Y_N}$	$\psi_{Y_N}$	$\psi_T$	$\mu$	$\ w - w_0\ _{C(1)}$	$ T - T_0 $	$\frac{\mu}{\ H_0\ _{\mathbf{H}_\infty}}$
45	700	45	0.0082	0.17	0.15	1.6%
45	700	20	0.0095	0.19	0.22	1.9%
35	600	15	0.0099	0.20	0.26	2.0%

TABLE 5.2

The table shows the perturbation of the trajectory and the period time obtained by simulating the Van der Pol oscillator with the perturbation in (5.2). The last column indicates that the bounds on  $\mu$  obtained in Table 5.1 predict the perturbation of the period time well.

$\mu_{sim}$	$\ \Delta\ _\infty$	$\ w - w_0\ _{C(1)}$	$ T - T_0 $	$\frac{\mu}{\ \Delta\ _\infty}$
0.040	0.0102	0.010	0.15	0.80
0.046	0.0117	0.012	0.22	0.81
0.049	0.0125	0.012	0.28	0.79

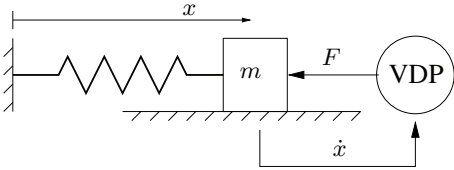


FIG. 5.3. Van der Pol oscillator interconnected with a mass and spring system.

**6. Discussion.** The numerical example indicates that the method has the potential to accurately predict the effect of perturbation. However, as the perturbation becomes larger the uncertainty in model, period time, and trajectory accumulate, and it becomes difficult to make the estimates accurate. The reason is partly due to the set inclusion test (ii) in Theorem 3.2, which is not developed to its full potential. Another reason is that the coefficients defining the quadratic constraints could be improved using structured singular value calculations.

In our results we have used a neighborhood  $\mathcal{Z}$  defining the limit cycle but also a neighborhood  $\mathcal{W}$  defining the trajectory at the input of the nonlinearity. We let  $\mathcal{Z}$  be defined by  $\mathbf{L}_2$ -topology, while  $\mathcal{W}$  is defined in the uniform topology in order to allow the incremental sector condition for the nonlinearity to be defined in time domain. It should also be mentioned that the set inclusion condition (the linear part of the dynamics maps the set  $\mathcal{Z}$  into  $\mathcal{W}$ ) could be integrated with the sector condition. We believe there are both advantages and disadvantages in doing so.

In a previous work [7] we presented ideas related to the present paper. However, there we used a different assumption on the neighborhood  $\mathcal{Z}$  of the nominal solution, which appears to be harder to use than the one defined in the present paper. In fact, the boundary of the neighborhood was interpreted wrongly in Theorem 1 of [7], which led to mistakes in the computations.

**7. Concluding remarks.** We have derived a new robustness result for limit cycles of systems with separable nonlinearities. In our main results we use a time-varying sector condition to characterize the nonlinearity around an a priori given

neighborhood of the nominal solution. By using relaxation results from robust control we obtain linear matrix inequalities defined by the sector condition, which, if satisfied, proves the existence of a limit cycle within the predefined neighborhood of the nominal solution.

**Appendix.** For the first term in the expression for  $\nu$  we exploit that  $H[T_0, 0] : \mathbf{L}_2(1) \rightarrow \mathbf{L}_2(1)$  are compact operators. This implies that we can truncate the expression at some sufficiently high frequency and obtain the bound

$$\nu_1(t) = \sigma_{\max} \left( \mathbf{1}_N(e^{j2\pi t}) \begin{bmatrix} \hat{H}_N[T_0, 0] & D_T \hat{H}_N[T_0, 0] \end{bmatrix} \hat{\Psi}_{Z_N}^{-1/2} \right),$$

where  $\hat{H}_N$  is defined as in (4.7), analogously for  $D_T \hat{H}_N[T_0, 0]$ , and

$$\mathbf{1}_N(e^{j2\pi t}) = \begin{bmatrix} e^{j2\pi Nt} I & \dots & e^{j2\pi Nt} I & I & e^{-j2\pi Nt} I & \dots & e^{-j2\pi Nt} I \end{bmatrix}.$$

Let us next consider the third term  $\nu_3 = \sup_{\theta \in [0,1]} \sup_{T \in \mathcal{T}} \|w_\Delta[T, \theta]\|_{C(1)}$ . We use that

$$w_\Delta[T, \theta] = (H[T, \theta] - H[T_0, 0])y_0 = \theta H_0[T] \Delta[T] (I - \theta H_0[T] \Delta[T])^{-1} H_0[T] y_0,$$

which suggests the norm bound

$$\|w_\Delta[T, \theta]\|_{C(1)} \leq \|H_0[T]\|_{\mathbf{L}_2(1) \rightarrow C(1)} \|\Delta[T] (I - \theta H_0[T] \Delta[T])^{-1} H_0[T] y_0\|_{\mathbf{L}_2(1)},$$

where

$$\begin{aligned} \|H_0[T]\|_{\mathbf{L}_2(1) \rightarrow C(1)} &= \sqrt{\sum_{k=-\infty}^{\infty} |H_0(j2\pi k/T)|^2}, \\ \|\Delta[T] (I - \theta H_0[T] \Delta[T])^{-1} H_0[T] y_0\|_{\mathbf{L}_2(1)} &\leq \mu \|H_0[T] y_0\|_{\mathbf{L}_2(1)} / (1 - \mu \|H_0[T]\|), \end{aligned}$$

where  $\|H_0[T]\| = \sup_k |H_0[T]|$ . We can once again exploit compactness of  $H_0$  and truncate the expressions at some sufficiently high frequency.

For the second term we may use that  $\hat{\Psi}_Z$  is chosen diagonal and therefore use the simplified expression given in (3.22) in Proposition 3.5. We use the bound

$$\begin{aligned} &|\Delta(j2\pi k/T, \theta) \hat{\Psi}_Y(k, k)^{-1/2}| \\ &\leq \frac{|H_0(j2\pi k/T) - H_0(j2\pi k/T_0)| + \mu |H_0(j2\pi k/T) H_0(j2\pi k/T_0)|}{(1 - \mu |H_0(j2\pi k/T)|) |\hat{\Psi}_Y(k, k)|^{1/2}} \end{aligned}$$

for the first term, and for the second term we use the bound

$$\|\delta[T] \psi_T^{-1/2}\|_{C(1)} \leq \|H_0[T_0]\|_{\mathbf{L}_2(1) \rightarrow C(1)} \|\check{\delta}[T]\|_{\mathbf{L}_2(1)},$$

where

$$\check{\delta}[T] = H_0[T_0]^{-1} \delta[T] \psi_T^{-1/2},$$

which can be estimated in a similar way as  $\|\delta[T]\|_{\mathbf{L}_2(1)}$  in (5.2).

## REFERENCES

- [1] R. ABRAHAM, J. E. MARSDEN, AND T. RATIU, *Manifolds, Tensor Analysis, and Applications*, Springer-Verlag, New York, 1988.
- [2] A. R. BERGEN, L. O. CHUA, A. I. MEES, AND E. W. SZETO, *Error bounds for general describing function problems*, IEEE Trans. Circuits Syst., 29 (1982), pp. 345–354.
- [3] M. K. H. FAN, A. L. TITS, AND J. C. DOYLE, *Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics*, IEEE Trans. Automat. Control, 36 (1991), pp. 25–38.
- [4] M. FARKAS, *Periodic Motions*, Springer-Verlag, New York, 1994.
- [5] T. T. GEORGIOU AND M. C. SMITH, *Robustness of a relaxation oscillator*, Internat. J. Robust Nonlinear Control, 10 (2000), pp. 1005–1024.
- [6] T. IWASAKI AND B. LIU, *Feedback control with central pattern generator for decentralized coordination of prototype mechanical rectifier*, in Proceedings of the IEEE American Control Conference, Boston, MA, 2004, pp. 3059–3064.
- [7] U. JÖNSSON, *A sector condition for local robustness of limit cycles*, in Proceedings of the IEEE American Control Conference, Minneapolis, MN, 2006, pp. 5014–5019.
- [8] U. JÖNSSON AND A. MEGRETSKI, *Limit cycle analysis using a system right inverse*, in Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 2005.
- [9] U. T. JÖNSSON AND A. MEGRETSKI, *A small-gain theory for limit cycles of systems on Luré form*, SIAM J. Control Optim., 44 (2005), pp. 909–938.
- [10] N. G. LLOYD, *Degree Theory*, Cambridge University Press, London, UK, 1978.
- [11] A. MEES AND A. BERGEN, *Describing functions revisited*, IEEE Trans. Automat. Control, 20 (1975), pp. 473–478.
- [12] A. I. MEES, *The describing function matrix*, IMA J. Appl. Math., 10 (1972), pp. 49–67.
- [13] A. I. MEES, *Dynamics of Feedback Systems*, Wiley, London, UK, 1981.
- [14] A. MEGRETSKI, *Frequency criteria for the existence of periodic solutions of systems with integral relations*, Differ. Uravn., 26 (1999), pp. 594–599.
- [15] A. PACKARD AND J. C. DOYLE, *The complex structured singular value*, Automatica, 29 (1993), pp. 71–109.
- [16] M. G. SAFONOV, *Describing Function Analysis of Nonlinear Periodic Systems*, Master's thesis, Department of Electrical Engineering, MIT, Cambridge, MA, 1971.
- [17] M. G. SAFONOV, *Stability margins of diagonally perturbed multivariable feedback systems*, IEE Proc., 129 (1982), pp. 251–256.
- [18] R. SEPULCHRE AND G.-B. STAN, *Feedback mechanisms for global oscillations in Lure systems*, Systems Control Lett., 54 (2005), pp. 809–818.
- [19] J. STELLING, E. D. GILLES, AND F. J. DOYLE, *Robustness properties of circadian clock architectures*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 13210–13215.
- [20] V. A. YAKUBOVICH, *S-procedure in nonlinear control theory*, Vestnik Leningrad Univ., 1 (1971), pp. 62–77 (in Russian); Vestnik Leningrad Univ., 4 (1977), pp. 73–93 (in English).
- [21] V. A. YAKUBOVICH, *Frequency domain criteria for oscillations in nonlinear systems with one stationary nonlinear component*, Sibirsk. Mat. Zh., 14 (1973), pp. 1100–1129.
- [22] N. YOUNG, *An Introduction to Hilbert Space*, Cambridge Math. Textbooks, Cambridge, UK, 1988.
- [23] K. ZHOU, *Essentials of Robust Control*, Prentice-Hall, Englewood Cliffs, NJ, 1998.



## OPTIMAL CONTROL FOR NONCONVEX RATE-INDEPENDENT EVOLUTION PROCESSES\*

FILIP RINDLER†

**Abstract.** Energetic solutions to rate-independent systems allow for an effective modeling of many physical systems displaying hysteretic effects, e.g., phase transformations in shape-memory alloys, elastoplasticity, and ferroelectricity. For some engineering applications, optimal control of such systems is desirable. We establish existence results for these continuous-time optimal control problems by a combination of the direct method with  $\Gamma$ -convergence arguments. Applicability to the common situation of a controlled external loading is demonstrated and a concrete partial differential inclusion as well as an academic model of the foreign exchange market including trading costs are investigated.

**Key words.** optimal control, optimization, rate-independent system, differential inclusion, variational inequality, energetic solution

**AMS subject classifications.** 49J24, 49J40, 49J45, 74C05, 74H20

**DOI.** 10.1137/080718711

**1. Introduction.** Many nonlinear physical systems possess the property of rate-independence, which means that their solution behavior does not depend on the speed of the forcing, but solely on its *path*. Examples from the material sciences include such diverse phenomena as phase transformation in shape-memory alloys [12, 20], elastoplasticity [1, 8, 14], and ferroelectricity (electrostriction) [9, 21]; see also the unifying approach of the theory of generalized standard materials in [16]. Sometimes, these systems are also referred to as “quasistatic,” because evolution takes place on a much longer time-scale than the system relaxation time. In engineering applications [24, 27], the need arises to (optimally) control such systems, for example in systems involving so-called “smart actuators.” Further, optimization with respect to material parameters could be desirable. Contrary to the situation for parabolic problems, cf., e.g., [22, 28], only very few investigations seem to have been carried out in the rate-independent or hysteretic case; cf. [9, 11].

The aim of this work is to show existence for optimal control problems involving rate-independent systems, especially in the situation where an external loading (e.g., a force or an electromagnetic field) is controlled. To illustrate the results, we also consider two concrete applications: We look at the optimal control problem for an evolutionary partial differential inclusion and we consider a simplistic model of controlling currency trading conditions on the foreign exchange market, taking into account transaction costs and (feedback) effects on the real values of the currencies caused by (large) transactions.

We base the optimal control results on the concept of energetic solutions; cf. [6, 13, 17, 18, 19] and the survey [15]. In the finite-dimensional (spatially discretized) setting, optimal control problems for such systems have recently been investigated as “mathematical programs with evolutionary equilibrium constraints” (MPEECs) in [9, 11] (with applications to delamination and micromagnetics). To the best of the

---

\*Received by the editors March 17, 2008; accepted for publication (in revised form) June 27, 2008; published electronically November 5, 2008.

<http://www.siam.org/journals/sicon/47-6/71871.html>

†Technische Universität Berlin, Institut für Mathematik, Sekr. MA 5-2, Straße des 17. Juni 136, D-10623 Berlin, Germany (rindler@math.tu-berlin.de).

author's knowledge, however, no results for the infinite-dimensional case are available at present.

When dealing with the problems previously described, one has to overcome the difficulty that, in general, solutions are not guaranteed to be unique and no continuous dependence of the solution on the data can be assumed (this holds only if the problems are uniformly convex); cf. [2]. In general, this disallows the use of the standard implicit programming approach, where the occurrence of the solution corresponding to a control is expressed through a continuous solution operator and thus is eliminated from the minimization problem; cf. [23] for a recent monograph. In certain cases, though, uniqueness and stability of solutions are guaranteed and the implicit programming approach is feasible; cf. [9, 11]. We choose a different path here and work with joint minimizing sequences of controls and corresponding solutions and exploit upper semicontinuity properties of the (set-valued) solution operators.

This paper is organized as follows: After describing the general setup in section 2, we show solvability of the optimal control problem in a fairly general setting in section 3. Then, in section 4, we demonstrate the application of the existence result to the common situation in which the control parameter is an external loading (e.g., an external force) and conclude with the aforementioned concrete applications in section 5. The appendix presents some terminology on the Sobolev–Bochner spaces  $W^{1,p}(0, T; \mathcal{X}; \mathcal{Y})$  and, for the convenience of the reader, a direct proof of a well-known compactness result.

**2. Setup.** To explain the setup in a special case, assume for the moment that we are given two Banach spaces  $\mathcal{Z}, \mathcal{U}$ , where  $\mathcal{Z}$  is the *state space* of the system, and  $\mathcal{U}$  is the *admissible control space*. Rate-independent systems can be modeled as *evolutionary doubly-nonlinear (sub)differential inclusions* [3, 25], here including a *control parameter*  $u \in \mathcal{U}$ ,

$$(DI) \quad 0 \in \partial_y \mathcal{R}(y(t), \dot{y}(t); u) + D_y \mathcal{E}(t, y(t); u) \quad (\text{in } \mathcal{Z}^*)$$

for almost every  $t \in [0, T]$ . By writing out the definition of the subdifferential, this is equivalent to the *evolutionary quasivariational inequality*

$$(VI) \quad \langle D_y \mathcal{E}(t, y(t); u), z - \dot{y}(t) \rangle + \mathcal{R}(y(t), z; u) - \mathcal{R}(y(t), \dot{y}(t); u) \geq 0$$

for all  $z \in \mathcal{Z}$  and almost all  $t \in [0, T]$ . In both cases,  $\mathcal{E}: [0, T] \times \mathcal{Z} \times \mathcal{U} \rightarrow \mathbb{R}_\infty = \mathbb{R} \cup \{+\infty\}$  is a possibly nonconvex, Gâteaux-differentiable *energy-storage functional*. The *dissipation potential*  $\mathcal{R}: \mathcal{Z} \times \mathcal{Z} \times \mathcal{U} \rightarrow \mathbb{R}_\infty$  is such that for all  $u \in \mathcal{U}$ ,  $\mathcal{R}(z, \cdot; u): \mathcal{Z} \rightarrow \mathbb{R}_\infty$  is convex and 1-homogeneous, i.e.,  $\mathcal{R}(z, \lambda \dot{z}; u) = \lambda \mathcal{R}(z, \dot{z}; u)$  for all  $(z, \dot{z}, u) \in \mathcal{Z} \times \mathcal{Z} \times \mathcal{U}$ . This implies 0-homogeneity of  $\partial \mathcal{R}(\dot{z}, \cdot; u)$ , i.e., the mathematical manifestation of rate-independence. Additionally, we impose an *initial condition*  $y(0) = y_0 \in \mathcal{Z}$  (in fact, not all initial values are admissible, but this will be made precise later). The optimal control problem then consists of minimizing a given *cost functional*  $\mathcal{J}(y, u)$  over all admissible controls  $u$  and corresponding solutions  $y$ .

We deliberately omitted specifying the space in which we look for solutions to (DI) or (VI), but it should be noted that these two formulations are only sensible if  $y$  is at least weakly differentiable in time. Rate-independent systems, however, do not always yield solutions that are so regular and we must switch to a different notion of solution. This is accomplished by the framework of energetic solutions [6, 13, 17, 18, 19]. Consider the basic setup as presented in the survey [15] and assume the *state space*  $\mathcal{Z}$  of the system to be a Hausdorff topological space. We

further need a Hausdorff *control space*  $\mathcal{U}$ , which contains all admissible controls (in fact, optimization with respect to material parameters is also possible, but we stick to the term “control” for clarity). In the following, all topological notions (in particular *convergence* and *compactness*) are to be understood in a sequential sense. Also note that we do *not* identify functions on the interval  $[0, T]$  (into some space) that are equal almost everywhere, because modification on a null set may change whether a given function is a solution (to the energetic formulation introduced below).

In the terminology of [15], we will always be working in the case of a “reduced” energy functional, which means that  $\mathcal{Z}$  might be part of a larger state space  $\mathcal{Q} = \mathcal{F} \times \mathcal{Z}$ , where  $\mathcal{F}$  and  $\mathcal{Z}$  contain the conservative and dissipative parts of a state, respectively, and the energy functional  $\mathcal{E}$  is given through  $\mathcal{E}(t, z) = \min\{\mathcal{E}_0(t, \varphi, z) : \varphi \in \mathcal{F}\}$  for  $t \in [0, T]$ ,  $z \in \mathcal{Z}$ , with some energy functional  $\mathcal{E}_0: [0, T] \times \mathcal{F} \times \mathcal{Z} \rightarrow \mathbb{R}_\infty$ . If the minimizer of  $\mathcal{E}_0(t, \varphi, z)$  is unique for fixed  $t, z$ , then one can reconstruct the conservative part from the dissipative part. Usually, for the subsequent assumptions to hold, one must also require continuous dependence of this unique minimizer on  $t$  and  $z$  (which holds in many interesting applications); see [15, section 3.4] for details. Sometimes (as in the concrete partial differential inclusion in section 5.1), the distinction between the conservative and dissipative parts of the state space is not existent and we can directly work with the setup as presented below.

The system properties are modeled by an energy-storage functional  $\mathcal{E}: [0, T] \times \mathcal{Z} \times \mathcal{U} \rightarrow \mathbb{R}_\infty := \mathbb{R} \cup \{+\infty\}$  (now not necessarily Gâteaux-differentiable anymore) and a *dissipation distance*  $\mathcal{D}: \mathcal{Z} \times \mathcal{Z} \times \mathcal{U} \rightarrow [0, \infty]$ . While  $\mathcal{E}(t, z; u)$  is the potential energy of the state  $z \in \mathcal{Z}$  at time  $t \in [0, T]$  and with control  $u \in \mathcal{U}$ , the dissipation distance  $\mathcal{D}(z_1, z_2; u)$  measures the dissipated energy when the state is changed from  $z_1 \in \mathcal{Z}$  to  $z_2 \in \mathcal{Z}$  (in the situation of (DI), (VI):  $\mathcal{D}(z_1, z_2; u) := \inf\{\int_0^1 \mathcal{R}(v(\tau), \dot{v}(\tau); u) \, d\tau : v \in W^{1,1}(0, 1; \mathcal{Z}), v(0) = z_1, v(1) = z_2\}$  or simply  $\mathcal{D}(z_1, z_2; u) = \mathcal{R}(z_2 - z_1; u)$  if  $D_y \mathcal{R} \equiv 0$ ; cf. sections 3.6 and 4.1 of [15]). Therefore, we require  $\mathcal{D}(\cdot, \cdot; u)$  to be a *quasimetric*, i.e., we assume the triangle inequality and the positivity  $\mathcal{D}(z_1, z_2; u) = 0$  if and only if  $z_1 = z_2$  (see (A3)). With our physical interpretation in mind,  $\mathcal{D}(\cdot, \cdot; u)$  does not need to be symmetric (consider, for example, crack formation in brittle materials [4, 5] or delamination [10]). Our control on the system is expressed through the dependence of  $\mathcal{E}$  and  $\mathcal{D}$  on the control parameter  $u \in \mathcal{U}$ . We tacitly assume  $\mathcal{E}(t, \cdot; u)$  to be proper, i.e., not identically  $+\infty$  for all  $t \in [0, T]$  and  $u \in \mathcal{U}$ .

For a process  $y: [0, T] \rightarrow \mathcal{Z}$ , the *total dissipation*  $\text{Diss}_{\mathcal{D}}(y; [r, s]; u)$  of  $y$  in the subinterval  $[r, s] \subseteq [0, T]$  is the total (pointwise)  $\mathcal{D}$ -variation of  $y$ , i.e.,

$$\text{Diss}_{\mathcal{D}}(y; [r, s]; u) := \sup \left\{ \sum_{j=1}^N \mathcal{D}(y(\tau_{j-1}), y(\tau_j); u) : r = \tau_0 < \dots < \tau_N = s \right\}.$$

The space  $\text{BV}_{\mathcal{D}}([0, T]; \mathcal{Z}; \mathcal{U})$  of *functions with bounded dissipation* consists of all processes  $y: [0, T] \rightarrow \mathcal{Z}$  such that  $\text{Diss}_{\mathcal{D}}(y; [0, T]; u) < \infty$  for all  $u \in \mathcal{U}$ . Since the spaces  $\mathcal{Z}$  and  $\mathcal{U}$  will always be clear from the context, we abbreviate  $\text{BV}_{\mathcal{D}}([0, T]; \mathcal{Z}; \mathcal{U})$  to  $\text{BV}_{\mathcal{D}}([0, T])$ . We say that a sequence  $(y_k)_k \subseteq \text{BV}_{\mathcal{D}}([0, T])$  converges in  $\text{BV}_{\mathcal{D}}([0, T])$  to  $y \in \text{BV}_{\mathcal{D}}([0, T])$  if  $\sup_{k \in \mathbb{N}} \text{Diss}_{\mathcal{D}}(y_k; [0, T]; u) < \infty$  for all  $u \in \mathcal{U}$  and  $y_k(t) \rightarrow y(t)$  for all  $t \in [0, T]$  (pointwise convergence). The generalized Helly selection principle [13, Theorem 3.2] is the fundamental compactness result in  $\text{BV}_{\mathcal{D}}([0, T])$  and says that from each sequence with uniformly bounded dissipation we can select a subsequence converging in  $\text{BV}_{\mathcal{D}}([0, T])$ , i.e., pointwise.

The *set of stable states* at time  $t \in [0, T]$  is defined to be

$$\mathcal{S}(t; u) := \left\{ z \in \mathcal{Z} : \mathcal{E}(t, z; u) < \infty \text{ and } \mathcal{E}(t, z; u) \leq \mathcal{E}(t, \hat{z}; u) + \mathcal{D}(z, \hat{z}; u) \right. \\ \left. \text{for all } \hat{z} \in \mathcal{Z} \right\}.$$

The “stability” expresses itself through the fact that the system must leave these states only when the exterior conditions change (i.e., if  $\mathcal{E}$  and hence the stable states change in the course of time).

Given a control  $u \in \mathcal{U}$  and an initial value  $y_0 \in \mathcal{S}(0; u)$ , let

$$\text{Sol}: \mathcal{Z} \times \mathcal{U} \rightrightarrows \text{BV}_{\mathcal{D}}([0, T]), \quad (y_0, u) \mapsto \text{Sol}(y_0, u) \subseteq \text{BV}_{\mathcal{D}}([0, T])$$

denote the set-valued solution operator that associates with  $(y_0, u)$  all *energetic solutions* to the (initial-value) evolution problem associated with  $\mathcal{E}(\cdot, \cdot; u)$  and  $\mathcal{D}(\cdot, \cdot; u)$ , i.e., all processes  $y \in \text{BV}_{\mathcal{D}}([0, T])$  that satisfy the *stability condition*

$$(S) \quad y(t) \in \mathcal{S}(t; u)$$

and the *energy balance*

$$(E) \quad \mathcal{E}(t, y(t); u) + \text{Diss}(y; [0, t]; u) = \mathcal{E}(0, y(0); u) + \int_0^t \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau$$

for all  $t \in [0, T]$  (this includes the requirement that  $\tau \mapsto \partial_t \mathcal{E}(\tau, y(\tau); u) \in L^1(0, T)$ ), as well as the *initial condition*  $y(0) = y_0$ .

If the assumptions stated after the subdifferential inclusion (DI) and variational inequality (VI) are satisfied,  $\mathcal{E}(t, \cdot; u)$  is *convex*, and  $y \in W^{1,1}(0, T; \mathcal{Z})$ , then the energetic formulation is equivalent to (DI) and (VI); cf. [15, 19]. In comparison, however, the integrated formulation of energetic solutions has the advantage that the solution  $y$  need not be differentiable, and neither a linear structure on  $\mathcal{Z}$  nor Gâteaux-differentiability of the functional  $\mathcal{E}$  needs to be assumed. Further, the dissipation distance  $\mathcal{D}$  need not be given through a dissipation potential  $\mathcal{R}$  and hence more general types of dissipation can be treated.

The optimal control problem for a rate-independent system consists of minimizing a given *cost functional*  $\mathcal{J}: \text{BV}_{\mathcal{D}}([0, T]) \times \mathcal{U} \rightarrow \mathbb{R}_{\infty}$ , depending on the control and the corresponding solution. Thus, our objective is the following continuous-time *optimal control problem*:

$$(OC) \quad \begin{cases} \text{For given initial value } y_0 \in \mathcal{Z}, \text{ find an } \textit{optimal control } u \in \mathcal{U} \\ \text{and a corresponding } \textit{optimal solution } y \in \text{Sol}(y_0, u), \text{ satisfying} \\ y \in \text{Argmin}\{ \mathcal{J}(\hat{y}, \hat{u}) : \hat{y} \in \text{Sol}(y_0, \hat{u}), \hat{u} \in \mathcal{U} \}. \end{cases}$$

As for noncontrolled systems, the initial value needs to fulfill a certain compatibility property; we will come back to that later; see (CC).

Later, we will also incorporate side conditions on the solution  $y$  and the control  $u$  in the form  $y(t) \in R_{\text{sol}} \subseteq \mathcal{Z}$  for all  $t \in [0, T]$  and  $u \in R_{\text{ctrl}} \subseteq \mathcal{U}$  with  $R_{\text{sol}}, R_{\text{ctrl}}$  closed; cf. Remark 3.7.

**3. Existence of optimal controls and processes.** In this section, we establish solvability of the optimal control problem (OC) stated in the previous section. After some auxiliary lemmas we prove a generalization of the standard existence theorem

in the theory of rate-independent systems; cf. [13, 15, 17]. In particular, an improved version of the lower energy estimate, Proposition 3.2, allows us to dispense with the uniform continuity assumption on the power  $\partial_t \mathcal{E}$ . As eventually we are interested in optimal control, this is an important generalization. Also, we employ different convergence conditions (in particular (C4)) than in the cited works in order to make the theory applicable to the scenarios we have in mind.

**3.1. Assumptions.** Most of the following assumptions are by now standard in the theory of energetic solutions to rate-independent systems:

*Uniform control of power:*

- (A1) There exist  $c_0^E \geq 0$ ,  $c_1^E > 0$  such that for all  $u \in \mathcal{U}$  and  $z \in \mathcal{Z}$  with  $\mathcal{E}(s, z; u) < \infty$  for some  $s \in [0, T]$ , it holds that
- (i)  $\mathcal{E}(\cdot, z; u) \in W^{1,\infty}(0, T)$  and
  - (ii)  $|\partial_t \mathcal{E}(t, z; u)| \leq c_1^E(\mathcal{E}(t, z; u) + c_0^E)$  for all  $t \in [0, T]$ .

*Uniform coercivity:*

- (A2) For all  $t \in [0, T]$ ,  $u \in \mathcal{U}$ , and  $E \in \mathbb{R}$ , it holds that
- $$\bigcup_{u \in \mathcal{U}} \{z \in \mathcal{Z} : \mathcal{E}(t, z; u) \leq E\} \text{ is relatively compact.}$$

*Quasimetric:*

- (A3) For all  $u \in \mathcal{U}$  and all  $z_1, z_2, z_3 \in \mathcal{Z}$ , it holds that
- (i)  $\mathcal{D}(z_1, z_2; u) = 0$  if and only if  $z_1 = z_2$  (*positivity*) and
  - (ii)  $\mathcal{D}(z_1, z_3; u) \leq \mathcal{D}(z_1, z_2; u) + \mathcal{D}(z_2, z_3; u)$  (*triangle inequality*).

Note that (ii) from (A1) implicitly contains the lower bound  $\mathcal{E}(\cdot, z; u) \geq -c_0^E$ . By the Gronwall inequality, we also get

$$(3.1) \quad \mathcal{E}(t, z; u) + c_0^E \leq (\mathcal{E}(s, z; u) + c_0^E) e^{c_1^E |t-s|}$$

for all  $z \in \mathcal{Z}$ ,  $u \in \mathcal{U}$ , and  $t, s \in [0, T]$ ; cf. [15, section 3.1].

We further need certain continuity properties of  $\mathcal{E}$  and  $\mathcal{D}$ :

*Lower semicontinuity of energy-storage functional:*

- (C1) For all  $t \in [0, T]$ ,  $u^k \rightarrow u$  in  $\mathcal{U}$  and  $z^k \rightarrow z$  in  $\mathcal{Z}$  with  $\sup_{k \in \mathbb{N}} \mathcal{E}(t, z^k; u^k) < \infty$ , it holds that
- $$\mathcal{E}(t, z; u) \leq \liminf_{k \rightarrow \infty} \mathcal{E}(t, z^k; u^k).$$

*Lower semicontinuity of dissipation distance:*

- (C2) For all  $u^k \rightarrow u$  in  $\mathcal{U}$  and  $z^k \rightarrow z$ ,  $\tilde{z}^k \rightarrow \tilde{z}$  in  $\mathcal{Z}$  with  $\sup_{k \in \mathbb{N}, t \in [0, T]} (\mathcal{E}(t, z^k; u^k) + \mathcal{E}(t, \tilde{z}^k; u^k)) < \infty$ , it holds that
- $$\mathcal{D}(z, \tilde{z}; u) \leq \liminf_{k \rightarrow \infty} \mathcal{D}(z^k, \tilde{z}^k; u^k).$$

*Upper semicontinuity of stability sets:*

- (C3) For all  $t_k \rightarrow t$  in  $[0, T]$ ,  $u^k \rightarrow u$  in  $\mathcal{U}$  and  $z^k \rightarrow z$  in  $\mathcal{Z}$  satisfying  $z^k \in \mathcal{S}(t_k; u^k)$  and  $\sup_{k \in \mathbb{N}} \mathcal{E}(t_k, z^k; u^k) < \infty$ , it holds that  $z \in \mathcal{S}(t, u)$ .

*Convergence of power in mean:*

For all  $u^k \rightarrow u$  in  $\mathcal{U}$  and  $y^k \rightarrow y$  in  $\text{BV}_{\mathcal{D}}([0, T])$  with

$$(C4) \quad \sup_{k \in \mathbb{N}, t \in [0, T]} \mathcal{E}(t, y^k(t); u^k) < \infty \text{ and } \tau \mapsto \partial_t \mathcal{E}(\tau, y^k(\tau); u^k) \in L^1(0, T),$$

it holds that  $\tau \mapsto \partial_t \mathcal{E}(\tau, y(\tau); u) \in L^1(0, T)$  and

$$\int_0^s \partial_t \mathcal{E}(\tau, y^k(\tau); u^k) \, d\tau \longrightarrow \int_0^s \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau$$

for all  $s \in [0, T]$ .

*Convergence of suitable power interpolants:*

For all  $u \in \mathcal{U}$ ,  $y \in \text{BV}_{\mathcal{D}}([0, T])$  and  $s \in [0, T]$  such that

$$(C5) \quad \tau \mapsto \partial_t \mathcal{E}(\tau, y(\tau); u) \in L^1(0, T), \text{ there exists a sequence of partitions}$$

$$(\Pi_k = (r = \tau_0^k, \dots, \tau_{N(k)}^k = s))_k \text{ of } [0, s] \text{ with } \|\Pi_k\| \rightarrow 0 \text{ and}$$

$$\sum_{j=1}^{N(k)} \int_{\tau_{j-1}^k}^{\tau_j^k} \partial_t \mathcal{E}(\xi, y(\tau_j^k); u) \, d\xi \longrightarrow \int_0^s \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau.$$

Note that by (3.1) in (C2) it suffices to require  $\sup_{k \in \mathbb{N}} (\mathcal{E}(t, z^k; u^k) + \mathcal{E}(t, \tilde{z}^k; u^k)) < \infty$  for *some*  $t \in [0, T]$ . Further, it is easy to see that conditions (A2), (A3), and (C2) together imply the following positivity property of the dissipation distance:

$$(3.2) \quad \begin{aligned} &\text{For all } u \in \mathcal{U}, (z^k)_k \subseteq \mathcal{Z} \text{ with } \sup_{k \in \mathbb{N}, t \in [0, T]} \mathcal{E}(t, z^k; u) < \infty \text{ and} \\ &\min \{ \mathcal{D}(z^k, z; u), \mathcal{D}(z, z^k; u) \} \rightarrow 0 \text{ for some } z \in \mathcal{Z}, \text{ it holds that} \\ &z^k \rightarrow z \text{ in } \mathcal{Z}. \end{aligned}$$

The conditional upper semicontinuity (C3) of the stability sets can be established, for example, by strengthening the assumptions (C1) and (C2) on  $\mathcal{E}$  and  $\mathcal{D}$  (cf. [17] for a fine hierarchy of increasingly stronger conditions):

*$\Gamma$ -continuity of energy-storage functional:*

For all  $t \in [0, T]$ ,  $u^k \rightarrow u$  in  $\mathcal{U}$ , it holds that

$$(C1') \quad \begin{aligned} (i) \quad &\text{For all } z^k \rightarrow z \text{ in } \mathcal{Z} \text{ with } \sup_{k \in \mathbb{N}} \mathcal{E}(t, z^k; u^k) < \infty, \text{ the} \\ &\text{lim inf-inequality } \mathcal{E}(t, z; u) \leq \liminf_{k \rightarrow \infty} \mathcal{E}(t, z^k; u^k) \text{ holds.} \\ (ii) \quad &\text{For all } z \in \mathcal{Z} \text{ with } \mathcal{E}(t, z; u) < \infty, \text{ there exists a recovery} \\ &\text{sequence } z^k \rightarrow z \text{ in } \mathcal{Z}, \text{ i.e., } \mathcal{E}(t, z; u) = \lim_{k \rightarrow \infty} \mathcal{E}(t, z^k; u^k). \end{aligned}$$

*Continuity of dissipation distance:*

For all  $u^k \rightarrow u$  in  $\mathcal{U}$  and  $z^k \rightarrow z, \tilde{z}^k \rightarrow \tilde{z}$  in  $\mathcal{Z}$  with

$$(C2') \quad \sup_{k \in \mathbb{N}, t \in [0, T]} (\mathcal{E}(t, z^k; u^k) + \mathcal{E}(t, \tilde{z}^k; u^k)) < \infty, \text{ it holds that}$$

$$\mathcal{D}(z^k, \tilde{z}^k; u^k) \rightarrow \mathcal{D}(z, \tilde{z}; u).$$

Notice that (C1') means that  $\mathcal{E}(t, \cdot; u) = \Gamma\text{-}\lim_k \mathcal{E}(t, \cdot; u^k)$  on sets of bounded energy. Obviously, the last two requirements are stronger than (C1) and (C2), respectively. As mentioned above, (C1') and (C2') together also imply (C3) as observed in Proposition 2.2 of [17]. For the convenience of the reader and because in the cited work even more general statements are made, we here show an adapted version.

LEMMA 3.1. *Conditions (C1') and (C2') imply (C3).*

*Proof.* First, note that by (A1) and (3.1), for all  $\tilde{z} \in \mathcal{Z}$  and  $u \in \mathcal{U}$  such that  $\mathcal{E}(s, \tilde{z}; u) < E$  for some  $s \in [0, T]$ , there exists a Lipschitz constant  $L = L(E) > 0$  such that

$$(3.3) \quad |\mathcal{E}(t, \tilde{z}; u) - \mathcal{E}(s, \tilde{z}; u)| \leq L |t - s|$$

for all  $s, t \in [0, T]$  (in fact,  $L(E) = c_1^E(E^* + c_0^E)e^{c_1^E T}$ ).

For a concise notation, define

$$\mathcal{H}(t, z, \hat{z}; u) := \mathcal{E}(t, \hat{z}; u) + \mathcal{D}(z, \hat{z}; u) - \mathcal{E}(t, z; u)$$

for  $u \in \mathcal{U}$ ,  $z, \hat{z} \in \mathcal{Z}$ . Note that  $z \in \mathcal{S}(t; u)$  if and only if  $\mathcal{H}(t, z, \hat{z}; u) \geq 0$  for all  $\hat{z} \in \mathcal{Z}$  with  $\mathcal{E}(t, \hat{z}; u) < \infty$ .

Now, let  $u^k \rightarrow u$  in  $\mathcal{U}$ ,  $(t_k, z^k) \rightarrow (t, z)$  in  $[0, T] \times \mathcal{Z}$  and  $z^k \in \mathcal{S}(t_k; u^k)$  with  $E^* := \sup_{k \in \mathbb{N}} \mathcal{E}(t_k, z^k; u^k) < \infty$ . To show  $\mathcal{H}(t, z, \hat{z}; u) \geq 0$  for all  $\hat{z} \in \mathcal{Z}$  with  $\mathcal{E}(t, \hat{z}; u) < \infty$ , let  $\hat{z}^k \rightarrow \hat{z}$  in  $\mathcal{Z}$  be a recovery sequence for  $\hat{z}$  with respect to the  $\Gamma$ -converging sequence  $(\mathcal{E}(t, \cdot; u^k))_k$ , which exists by (C1'). In particular, also employing the Lipschitz-continuity (3.3), we have

$$\limsup_{k \rightarrow \infty} \mathcal{E}(t_k, \hat{z}^k; u^k) \leq \limsup_{k \rightarrow \infty} \mathcal{E}(t, \hat{z}^k; u^k) + \lim_{k \rightarrow \infty} L_1 |t - t_k| \leq \mathcal{E}(t, \hat{z}; u),$$

where  $L_1$  is the Lipschitz constant associated to the energy bound on the recovery sequence. From the lim inf-inequality of  $\mathcal{E}(\cdot, z^k; u^k)$ , we further get

$$\limsup_{k \rightarrow \infty} -\mathcal{E}(t_k, z^k; u^k) \leq -\liminf_{k \rightarrow \infty} \mathcal{E}(t, z^k; u^k) + \lim_{k \rightarrow \infty} L_2 |t - t_k| \leq -\mathcal{E}(t, z; u)$$

with  $L_2$  corresponding to the energy bound  $E^*$ .

Combining the previous two estimates and using the continuous convergence  $\mathcal{D}(z^k, \hat{z}^k; u^k) \rightarrow \mathcal{D}(z, \hat{z}; u)$  of the dissipation distance (C2'), we arrive at

$$\mathcal{H}(t, z, \hat{z}; u) \geq \limsup_{k \rightarrow \infty} \mathcal{H}(t_k, z^k, \hat{z}^k; u^k) \geq 0,$$

where the last inequality holds, because  $z^k \in \mathcal{S}(t_k; u^k)$ .  $\square$

Finally, we require the control space  $\mathcal{U}$  and the cost functional  $\mathcal{J}$  to satisfy the following standard requirements:

(J1) *Compactness of  $\mathcal{U}$ :*  
 $\mathcal{U}$  is compact.

*Lower semicontinuity of  $\mathcal{J}$ :*  
 For all  $u^k \rightarrow u$  in  $\mathcal{U}$  and  $y^k \rightarrow y$  in  $\text{BV}_{\mathcal{D}}([0, T])$  with  $y^k \in \text{Sol}(y_0, u^k)$ ,  
 (J2)  $y \in \text{Sol}(y_0, u)$  and  $\sup_{k \in \mathbb{N}, t \in [0, T]} \mathcal{E}(t, z^k; u^k) < \infty$ , it holds that  

$$\mathcal{J}(y; u) \leq \liminf_{k \rightarrow \infty} \mathcal{J}(y^k; u^k).$$

Condition (J1) might seem restrictive, but can be omitted if additional coerciveness conditions with respect to  $u$  are imposed on  $\mathcal{J}$ . As many cost functionals, however, do not even depend on  $u$ , here we give priority to the situation with compact  $\mathcal{U}$ .

**3.2. Existence of solutions for a fixed control.** To construct solutions, we will rely on a *discrete-time incremental problem*. For this, fix for each  $k \in \mathbb{N}$  a partition

$$\Pi_k = (0 = t_0^k, t_1^k, \dots, t_{N(k)}^k = T),$$

of the interval  $[0, T]$ , where the *fineness*  $\|\Pi_k\| := \max\{(t_j^k - t_{j-1}^k) : j = 1, \dots, N(k)\}$  goes to zero as  $k \rightarrow \infty$ . For a control  $u \in \mathcal{U}$  and a stable initial value  $y_0 \in \mathcal{S}(0; u)$ , we then want to find solutions  $y^k = (y_0^k, \dots, y_N^k) = (y^k(t_0), \dots, y^k(t_N)) \in \mathcal{Z}^{\Pi_k}$  to

$$(IP_k) \quad \begin{cases} \text{Given } y_0^k = y_0 \in \mathcal{S}(0; u), \text{ inductively find } y_j^k \in \mathcal{Z} \text{ such that} \\ y_j^k \in \operatorname{Argmin}\{\mathcal{E}(t_j^k, \hat{z}; u) + \mathcal{D}(y_{j-1}^k, \hat{z}; u) : \hat{z} \in \mathcal{Z}\} \text{ for } j = 1, \dots, N. \end{cases}$$

The next proposition is a generalized version of Proposition 5.7 in [15] (see also [6, 20]). In this version, we do not need uniform continuity of the power  $\partial_t \mathcal{E}(\cdot, z; u)$  for fixed  $z \in \mathcal{Z}$  (and  $u \in \mathcal{U}$ ) anymore. This generalization is necessary in order to allow for nonsmooth controls (which can cause the time-evolution of the functionals to be nonsmooth).

**PROPOSITION 3.2.** *Assume (A1) and (C5). Let  $u \in \mathcal{U}$ , and let  $y \in \operatorname{BV}_{\mathcal{D}}([0, T])$  be a stable process, i.e.,  $y(t) \in \mathcal{S}(t; u)$  for all  $t \in [0, T]$ . Further, let  $\tau \mapsto \partial_t \mathcal{E}(\tau, y(\tau); u) \in L^1(0, T)$ . Then, for all  $t \in [0, T]$  the process  $y$  satisfies the lower energy estimate*

$$\mathcal{E}(t, y(t); u) + \operatorname{Diss}_{\mathcal{D}}(y; [0, t]; u) \geq \mathcal{E}(0, y(0); u) + \int_0^t \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau.$$

*Proof.* Let  $0 = \tau_0 < \tau_1 < \dots < \tau_N = t$  be any partition of  $[0, t]$ . The stability of  $y(\tau_{j-1})$  tested with  $y(\tau_j)$  gives

$$\begin{aligned} \mathcal{E}(\tau_{j-1}, y(\tau_{j-1}); u) &\leq \mathcal{E}(\tau_{j-1}, y(\tau_j); u) + \mathcal{D}(y(\tau_{j-1}), y(\tau_j); u) \\ &= \mathcal{E}(\tau_j, y(\tau_j); u) - \int_{\tau_{j-1}}^{\tau_j} \partial_t \mathcal{E}(\xi, y(\tau_j); u) \, d\xi \\ &\quad + \mathcal{D}(y(\tau_{j-1}), y(\tau_j); u). \end{aligned}$$

Summing this over  $j = 1, \dots, N$  leads to

$$\begin{aligned} \mathcal{E}(t, y(t); u) + \operatorname{Diss}(y; [0, t]; u) &\geq \mathcal{E}(t, y(t); u) + \sum_{j=1}^N \mathcal{D}(y(\tau_{j-1}), y(\tau_j); u) \\ &\geq \mathcal{E}(0, y(0); u) + \sum_{j=1}^N \int_{\tau_{j-1}}^{\tau_j} \partial_t \mathcal{E}(\xi, y(\tau_j); u) \, d\xi. \end{aligned}$$

By (C5), the last sum converges to  $\int_0^t \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau$  for a suitable sequence of partitions.  $\square$

We now have all of the necessary ingredients to prove the existence of a solution to (S) and (E) for some arbitrary, but fixed, control. In other words, we show that  $\operatorname{Sol}(y_0; u)$  is nonempty for all  $u \in \mathcal{U}$  and  $y_0 \in \mathcal{S}(0; u)$ .

**THEOREM 3.3.** *Assume (A1)–(A3), (C1)–(C5), and let  $u \in \mathcal{U}$ . Then, for all initial values  $y_0 \in \mathcal{S}(0; u)$ , there exists a solution process  $y \in \operatorname{BV}_{\mathcal{D}}([0, T])$ , i.e.,  $y \in \operatorname{Sol}(y_0; u)$ , for which it holds that*

- (i)  *$y$  is the pointwise limit of a subsequence (not renumbered) of the piecewise-constant, right-continuous interpolants  $\bar{y}^k$  of solutions  $y^k$  to  $(IP_k)$ ,*



- (ii)  $\mathcal{E}(t, \bar{y}^k(t); u) \rightarrow \mathcal{E}(t, y(t); u)$  for all  $t \in [0, T]$ ,
- (iii)  $\text{Diss}_{\mathcal{D}}(\bar{y}^k; [0, t]; u) \rightarrow \text{Diss}_{\mathcal{D}}(y; [0, t]; u)$  for all  $t \in [0, T]$ .

*Proof. Step 1: Incremental problems.* For each  $k \in \mathbb{N}$ , by assumptions (C1) and (C2) (which imply lower semicontinuity of  $\mathcal{E}(t_j^k, \cdot; u) + \mathcal{D}(y_{j-1}^k, \cdot; u)$ ) the direct method of the Calculus of Variations, applied inductively for  $j = 1, \dots, N(k)$ , shows existence of a discrete-time solution  $y^k = (y_0^k, y_1^k, \dots, y_{N(k)}^k)$  of  $(\text{IP}_k)$ .

*Step 2: A priori estimates.* For brevity of notation, set  $e_j^k := \mathcal{E}(t_j^k, y_j^k; u)$  and  $\delta_j^k := \mathcal{D}(y_{j-1}^k, y_j^k; u)$ . Since  $y_j^k$  is a minimizer of  $\mathcal{E}(t_j^k, \cdot; u) + \mathcal{D}(y_{j-1}^k, \cdot; u)$ , we have for all  $\hat{z} \in \mathcal{Z}$  and  $j = 1, \dots, N(k)$  that

$$e_j^k \leq \mathcal{E}(t_j^k, \hat{z}; u) + \mathcal{D}(y_{j-1}^k, \hat{z}; u) - \delta_j^k \leq \mathcal{E}(t_j^k, \hat{z}; u) + \mathcal{D}(y_j^k, \hat{z}; u),$$

i.e.,  $y_j^k \in \mathcal{S}(t_j^k; u)$ . For  $j = 0$  we assumed  $y_0^k = y_0 \in \mathcal{S}(0; u)$ .

Similarly, testing the minimality of  $y_j^k$  with  $y_{j-1}^k$  and using the growth estimate (A1) and its consequence (3.1) gives

$$(3.4) \quad e_j^k + \delta_j^k \leq \mathcal{E}(t_j^k, y_{j-1}^k; u) = e_{j-1}^k + \int_{t_{j-1}^k}^{t_j^k} \partial_t \mathcal{E}(\tau, y_{j-1}^k; u) \, d\tau$$

$$(3.5) \quad \begin{aligned} &\leq e_{j-1}^k + \int_{t_{j-1}^k}^{t_j^k} c_1^E (e_{j-1}^k + c_0^E) e^{c_1^E (\tau - t_{j-1}^k)} \, d\tau \\ &= e_{j-1}^k + (e_{j-1}^k + c_0^E) (e^{c_1^E (t_j^k - t_{j-1}^k)} - 1). \end{aligned}$$

The last estimate implies  $e_j^k + \delta_j^k + c_0^E \leq (e_{j-1}^k + c_0^E) e^{c_1^E (t_j^k - t_{j-1}^k)}$  and, by repeated application,

$$(3.6) \quad e_j^k + \delta_j^k + c_0^E \leq (e_0^k + c_0^E) e^{c_1^E t_j^k}.$$

Combining this with a similar estimate as in (3.5), for the right-continuous, piecewise-constant interpolant  $\bar{y}^k$  of  $y^k$ , we deduce the uniform energy bound (recall  $\delta_j^k, c_0^E \geq 0$ )

$$(3.7) \quad \bar{e}^k(t) := \mathcal{E}(t, \bar{y}^k(t); u) \leq (\mathcal{E}(0, y_0; u) + c_0^E) e^{c_1^E t} =: E^* e^{c_1^E t}$$

for all  $k \in \mathbb{N}$ ,  $t \in [0, T]$ . By (A1), we get

$$(3.8) \quad \bar{P}^k(t) := \partial_t \mathcal{E}(t, \bar{y}^k(t); u) \leq c_1^E (E^* e^{c_1^E t} + c_0^E)$$

for all  $k \in \mathbb{N}$ ,  $t \in [0, T]$ .

Rearranging (3.5), summing up, and using (3.6) gives

$$\begin{aligned} \sum_{j=1}^{N(k)} \delta_j^k &\leq e_0^k - e_{N(k)}^k + \sum_{j=1}^{N(k)} (e_{j-1}^k + c_0^E) (e^{c_1^E (t_j^k - t_{j-1}^k)} - 1) \\ &\leq (e_0^k + c_0^E) - (e_{N(k)}^k + c_0^E) + (e_0^k + c_0^E) \sum_{j=1}^{N(k)} (e^{c_1^E t_j^k} - e^{c_1^E t_{j-1}^k}) \\ &\leq E^* e^{c_1^E T} \end{aligned}$$

since  $e_{N(k)}^k \geq -c_0^E$ , which is contained implicitly in point (ii) of (A1). It follows that the dissipation of  $\bar{y}^k$  stays uniformly bounded, i.e.,

$$(3.9) \quad \text{Diss}(\bar{y}^k; [0, T]; u) = \sum_{j=1}^{N(k)} \delta_j^k \leq E^* e^{c_1^E T}.$$

*Step 3: Selection of subsequences.* The  $k$ -independent bounds (3.7), (3.9) together with the assumptions (A2), (A3), (C2), and (3.2), allow us to invoke the generalized Helly selection principle, Theorem 3.2 in [13], in order to get subsequences (not renumbered) and limit functions  $y \in \text{BV}_{\mathcal{D}}([0, T])$  and  $\delta^\infty: [0, T] \rightarrow \mathbb{R}$  such that for all  $t, r, s \in [0, T]$  with  $r < s$ , it holds that

$$(3.10) \quad \begin{aligned} \bar{y}^k(t) &\rightarrow y(t), & \text{Diss}(\bar{y}^k; [0, t]; u) &\rightarrow \delta^\infty(t), & \text{and} \\ \text{Diss}(y; [r, s]; u) &\leq \delta^\infty(s) - \delta^\infty(r), & \delta^\infty(0) &= 0. \end{aligned}$$

The convergence of the power in mean (C4) (with  $u^k := u$ , and notice that  $\bar{P}^k$  is measurable by (A1) and in  $L^1(0, T)$  by (3.8)) now implies  $\tau \mapsto \partial_t \mathcal{E}(\tau, y(\tau); u) \in L^1(0, T)$  and

$$(3.11) \quad \int_0^t \bar{P}^k(\tau) \, d\tau \longrightarrow \int_0^t \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau$$

for all  $t \in [0, T]$ .

*Step 4: Stability of the limit process.* Fix  $t \in [0, T]$ , and let  $m(t, k) \in \{1, \dots, N(k)\}$  be the largest integer such that  $t_{m(t, k)}^k \leq t$ . If  $t > 0$ , then also  $m(t, k) > 0$  for  $k$  large enough and because  $\bar{y}^k(t) = \bar{y}^k(t_{m(t, k)}^k) \in \mathcal{S}(t_{m(t, k)}^k; u)$  (this was shown in Step 2), (3.10), and assumption (C3) (with  $u^k := u$ ) imply that  $y(t) \in \mathcal{S}(t; u)$ . For the initial value, we already know  $y(0) = y_0 \in \mathcal{S}(0; u)$  by hypothesis. Hence, the limit process satisfies (S).

*Step 5: Energy estimates in the limit.* By the lower semicontinuity assumption (C1) on the energy functional we get

$$(3.12) \quad \mathcal{E}(t, y(t); u) \leq \liminf_{k \rightarrow \infty} \bar{e}^k(t) =: e_\infty(t) \leq e^\infty(t) := \limsup_{k \rightarrow \infty} \bar{e}^k(t).$$

Let  $m(t, k)$  be defined as above. We sum (3.4) for  $j = 1, \dots, m(t, k)$  and get, also employing the uniform bound (3.8) on  $\bar{P}^k = \partial_t \mathcal{E}(\cdot, \bar{y}^k(\cdot); u)$ ,

$$\begin{aligned} \bar{e}^k(t) + \text{Diss}(\bar{y}^k; [0, t]; u) &\leq \bar{e}_{m(t, k)}^k + \sum_{j=1}^{m(t, k)} \delta_j^k + c \|\Pi_k\| \\ &\leq \bar{e}^k(0) + \int_0^t \bar{P}^k(\tau) \, d\tau + c \|\Pi_k\| \end{aligned}$$

for all  $t \in [0, T]$ . Passing to the limit in this inequality and using (3.10), (3.12), and the convergence of the power in mean (3.11) as well as  $\bar{e}^k(0) = \mathcal{E}(0, y_0; u)$  and  $\delta^\infty(0) = 0$ , we get

$$(3.13) \quad \begin{aligned} \mathcal{E}(t, y(t); u) + \text{Diss}(y; [0, t]; u) &\leq e_\infty(t) + \delta^\infty(t) \leq e^\infty(t) + \delta^\infty(t) \\ &\leq \mathcal{E}(0, y_0; u) + \int_0^t \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau \end{aligned}$$

for all  $t \in [0, T]$ , i.e., the upper half of (E).

The lower energy estimate follows directly from Proposition 3.2 (cf. Step 3). Hence, (3.13) is an equality and we have shown (E) as well as the convergence  $\bar{e}^k(t) \rightarrow e_\infty(t) = e^\infty(t)$ . Then,  $\mathcal{E}(t, y(t); u) \leq e_\infty(t)$  and  $\text{Diss}_{\mathcal{D}}(y; [0, t]; u) \leq \delta^\infty(t)$  (by (C1) and (3.10), respectively) imply (ii) and (iii) from the statement of the theorem.  $\square$

**3.3. Existence of optimal controls.** After having established the existence of a solution, we can now move to the principal aim of this work, i.e., the existence proof for optimal controls and corresponding solutions. We combine the direct method and an upper semicontinuity property of the solution operator  $u^k \mapsto \text{DSol}(y_0; u^k)$  with respect to the convergence  $u^k \rightarrow u$  in  $\mathcal{U}$  (a variant of the “ $\Gamma$ -stability” first discovered in Theorem 3.1 of [17]).

**THEOREM 3.4.** *Assume (A1)–(A3), (C1)–(C5), and (J1)–(J2). Further, let the initial value  $y_0 \in \mathcal{Z}$  satisfy the compatibility condition*

$$(CC) \quad \begin{aligned} (i) \quad & y_0 \in \bigcap_{u \in \mathcal{U}} \mathcal{S}(0; u) \\ (ii) \quad & \mathcal{E}(0, y_0; u^k) \rightarrow \mathcal{E}(0, y_0; u) \quad \text{for all } u^k \rightarrow u \text{ in } \mathcal{U}. \end{aligned}$$

*Then (OC) has at least one solution.*

**REMARK 3.5.** *Obviously, if  $\mathcal{E}$  satisfies the continuity property that  $\mathcal{E}(t, z; u^k) \rightarrow \mathcal{E}(t, z; u)$  for all  $u^k \rightarrow u$  in  $\mathcal{U}$  and all  $t \in [0, T]$ ,  $z \in \mathcal{Z}$  fixed, then the compatibility condition (CC) reduces to point (i) alone.*

*Proof.* Let  $(y^k, y^k) \in \mathcal{U} \times \text{BV}_{\mathcal{D}}([0, T])$  with  $y^k \in \text{Sol}(y_0; u^k)$  be a minimizing sequence for  $\mathcal{J}$ , i.e.,

$$\mathcal{J}(y^k; u^k) \rightarrow \inf \{ \mathcal{J}(\hat{y}, \hat{u}) : \hat{y} \in \text{Sol}(y_0, \hat{u}), \hat{u} \in \mathcal{U} \}.$$

By (J1), we can select a converging subsequence (not relabelled)  $u^k \rightarrow u$  in  $\mathcal{U}$ . We show that there exists  $y \in \text{BV}_{\mathcal{D}}([0, T])$  with  $y \in \text{Sol}(y_0; u)$  such that  $y^k(t) \rightarrow y(t)$  in  $\mathcal{Z}$  for all  $t \in [0, T]$ . By an inspection of the proof of Theorem 3.3, we see that in (3.7) and (3.9), the value of  $E^*$  can be chosen independently of  $k$  by point (ii) of (CC). Thus, as in the previous theorem, invoking the generalized Helly selection principle (this time the  $\Gamma$ -convergence version in Theorem A.1 of [17]) we conclude that there exists  $y: [0, T] \rightarrow \mathcal{Z}$  such that  $y^k(t) \rightarrow y(t)$  for all  $t \in [0, T]$ . Assumption (C4) gives

$$(3.14) \quad \int_0^t \partial_t \mathcal{E}(\tau, y^k(\tau); u^k) \, d\tau \longrightarrow \int_0^t \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau$$

for all  $t \in [0, T]$ .

By the uniform boundedness of  $\mathcal{E}(0, y_0; u^k)$  (cf. point (ii) of (CC)) and (A1), from Gronwall’s inequality it follows that  $\mathcal{E}(t, y^k(t); u^k)$  is uniformly bounded [15, section 3.1]. Thus, the stability (S) of  $y$  follows immediately from the upper semicontinuity of the stability sets (C3) (note that we have pointwise convergence of the processes  $y^k$  to  $y$  and each  $y^k$  is stable).

The upper part of the energy balance can be derived using the energy balances  $(E_k)$  for  $y^k$  together with assumptions (C1), (C2), and (ii) of the compatibility

condition (CC) as follows:

$$\begin{aligned} & \mathcal{E}(t, y(t); u) + \text{Diss}_{\mathcal{D}}([0, t]; y(t); u) \\ & \leq \liminf_{k \rightarrow \infty} (\mathcal{E}(t, y^k(t); u^k) + \text{Diss}_{\mathcal{D}}([0, t]; y^k(t); u^k)) \\ & = \liminf_{k \rightarrow \infty} \left( \mathcal{E}(0, y_0; u^k) + \int_0^t \partial_t \mathcal{E}(\tau, y^k(\tau); u^k) \, d\tau \right) \\ & = \mathcal{E}(0, y_0; u) + \int_0^t \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau, \end{aligned}$$

where we have also employed (3.14) and the fact that by (C2)

$$\text{Diss}_{\mathcal{D}}(y; [0, t]; u) \leq \liminf_{k \rightarrow \infty} \text{Diss}_{\mathcal{D}}(y^k; [0, t]; u^k).$$

The lower half of the energy balance follows again from Proposition 3.2.

Hence, we have found  $(u, y) = \lim_{k \rightarrow \infty} (u^k, y^k)$  with  $y \in \text{Sol}(y; u)$ , which, by the choice of  $(u^k, y^k)$  and the lower semicontinuity (J2), must be a minimizer for  $\mathcal{J}$ .  $\square$

**REMARK 3.6.** *If the initial value  $y_0$  does not satisfy (CC) (which might be difficult to check), then the solution  $y$  need not be stable at time  $t = 0$ . In this case, however, a solution immediately jumps to a stable state as soon as  $t > 0$  (see Step 2 in the proof of Theorem 3.3, where stability is shown for all points of the discrete solution away from the initial value). This “self-stabilization” is a very useful feature in the framework of energetic solutions. Of course, (CC) is not necessary for initial stability of an optimal control-solution pair  $(u, y)$ . If further knowledge on the possible optimal controls is available, it suffices to require initial stability for such controls.*

**REMARK 3.7.** *Certain side-conditions can be incorporated into the optimal control problem as well: Constraints on the control may be enforced by setting*

$$\tilde{\mathcal{J}}(y; u) := \begin{cases} \mathcal{J}(y; u) & \text{if } u \in R_{\text{ctrl}}, \\ +\infty & \text{otherwise,} \end{cases}$$

where  $R_{\text{ctrl}} \subseteq \mathcal{U}$  is closed. Closed and time-independent constraints on the solution may be directly incorporated into the functional  $\mathcal{E}$  in a similar manner. In both cases, all assumptions from section 3.1 are easily verified. Note that time-independent constraints cannot be treated since if  $\mathcal{E}(s, z; u) < \infty$  for one  $s \in [0, T]$ , then already  $\mathcal{E}(s, z; u) < \infty$  for all  $s \in [0, T]$  by (3.1).

**4. Optimal control of external loadings.** In many applications,  $\mathcal{Z}$  is a Banach space equipped with its weak topology, and controls act on the process in the form of an external loading (e.g., a force). Technically, this means that  $\mathcal{U}$  is a space of functions with values in (a subset of) the dual space  $\mathcal{Z}^*$  and the energy-storage functional depends on  $u$  only through the duality pairing  $\langle u(t), y(t) \rangle$  (we will specify later in which space the duality product is taken). In this section, we apply and adapt the previously developed results to this special situation and also give conditions on the space  $\mathcal{U}$  and on the functional  $\mathcal{J}$  such that (J1) and (J2) are fulfilled.

**4.1. Control space.** To gain the necessary compactness for the general theory, on several instants we need not only the space  $\mathcal{Z}$  itself, but also a larger *pivot space*  $\mathcal{P}$  such that the compact embeddings

$$(4.1) \quad \mathcal{Z} \xhookrightarrow{c} \mathcal{P} \quad \text{and} \quad \mathcal{P}^* \xhookrightarrow{c} \mathcal{Z}^*$$

hold. Of course, by Schauder's theorem, only the first compact embedding needs to be verified. Note that if, additionally,  $\mathcal{P}$  is a Hilbert space, then we can identify  $\mathcal{P}$  and  $\mathcal{P}^*$  and get the Gelfand triple  $\mathcal{Z} \hookrightarrow \mathcal{P} \cong \mathcal{P}^* \hookrightarrow \mathcal{Z}^*$ . Here, however, we do not need the additional Hilbert space structure. Further, we require  $\mathcal{Z}$  and  $\mathcal{P}$  to be reflexive and separable.

EXAMPLE 4.1. *The most prominent example for spaces  $\mathcal{Z}, \mathcal{P}$  as in (4.1) is given through the Gelfand triple  $H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$ , i.e.,  $\mathcal{P} = \mathcal{P}^* = L^2(\Omega)$ ,  $\mathcal{Z} = H_0^1(\Omega)$  (equipped with the norm  $\|z\|_{H_0^1(\Omega)} := \|\nabla z\|_{L^2(\Omega)}$ ), where  $\Omega \subseteq \mathbb{R}^d$  is a bounded Lipschitz domain.*

The control space  $\mathcal{U}$  is rooted on the framework of Sobolev–Bochner spaces. Some results from this theory are collected in the appendix. The space  $\mathcal{U}$  will always be a subset of the space  $W^{1,\infty}(0, T; \mathcal{P}^*)$  equipped with its weak\* topology. Further, we require (also cf. the remark after (J2))

*Boundedness of  $\mathcal{U}$ :*

(U) There exists a constant  $M_{\mathcal{U}} \geq 0$  such that for all  $u \in \mathcal{U}$

$$\|u\|_{W^{1,\infty}(0,T;\mathcal{P}^*)} \leq M_{\mathcal{U}}.$$

In many applications (see the choice of the energy functional in the next section), the last condition means that the power  $\tau \mapsto \partial_t \mathcal{E}(\tau, z; u) = \langle \dot{u}(\tau), z \rangle_{\mathcal{P}^* \times \mathcal{P}}$  ( $z \in \mathcal{Z}$ ,  $u \in \mathcal{U}$ ) stays  $L^\infty$ -bounded on bounded subsets of the state space  $\mathcal{Z}$ .

EXAMPLE 4.2. *The simplest choice for  $\mathcal{U}$  is*

$$\mathcal{U} := B_{W^{1,\infty}(0,T;\mathcal{P}^*)}(M_{\mathcal{U}}),$$

*the ball in  $W^{1,\infty}(0, T; \mathcal{P}^*)$  with radius  $M_{\mathcal{U}} > 0$ . This bounded set is weakly\* sequentially compact; also cf. the appendix.*

**4.2. Energy-storage functional and dissipation distance.** We assume the energy-storage functional to have the form of a sum of potential energy and external loading, i.e.,

$$(4.2) \quad \mathcal{E}(t, z; u) = \mathcal{W}(z) + \langle u(t), z \rangle_{\mathcal{P}^* \times \mathcal{P}}$$

with  $\mathcal{W}: \mathcal{Z} \rightarrow \mathbb{R}_\infty$ . We only treat time-independent  $\mathcal{W}$ , but cf. Remark 4.5.

To fulfill all of the requirements on  $\mathcal{E}$  in our special case, we require

*Superlinear growth of  $\mathcal{W}$ :*

(W1) There exist  $\mu > 0$ ,  $\kappa \geq 0$ ,  $s > 1$  such that  $\mathcal{W}(z) \geq \mu \|z\|_{\mathcal{Z}}^s - \kappa$  for all  $z \in \mathcal{Z}$ .

*Lower semicontinuity of  $\mathcal{W}$ :*

(W2)  $\mathcal{W}$  is in weakly sequentially lower semicontinuous in  $\mathcal{Z}$ .

LEMMA 4.3. *Let  $\mathcal{U} \subseteq W^{1,\infty}(0, T; \mathcal{P}^*)$  fulfill (U), let  $\mathcal{E}$  be defined through (4.2), and assume (W1) and (W2). Then, (A1), (A2), (C1'), (C4), and (C5) hold.*

*Proof.* Point (i) of assumption (A1) is clear. Let  $\gamma$  be the operator norm of the continuous embedding  $\mathcal{Z} \hookrightarrow \mathcal{P}$ , i.e.,  $\|z\|_{\mathcal{P}} \leq \gamma \|z\|_{\mathcal{Z}}$  for all  $z \in \mathcal{Z}$ . Choose  $K$  be so large that  $\mu\gamma^{-1}K^{s-1} - M_{\mathcal{U}} > 0$ . Using (W1), (4.2), we get for all  $t \in [0, T]$ ,  $z \in \mathcal{Z}$  with  $\|z\|_{\mathcal{Z}} \geq K$  and  $u \in \mathcal{U}$

$$\mathcal{E}(t, z; u) \geq \mathcal{W}(z) - M_{\mathcal{U}} \|z\|_{\mathcal{P}} \geq \mu \|z\|_{\mathcal{Z}}^s - \kappa - M_{\mathcal{U}} \|z\|_{\mathcal{P}} \geq (\mu\gamma^{-1}K^{s-1} - M_{\mathcal{U}}) \|z\|_{\mathcal{P}} - \kappa.$$

Then also

$$|\partial_t \mathcal{E}(t, z; u)| = |\langle \dot{u}(t), z \rangle_{\mathcal{P}^* \times \mathcal{P}}| \leq M_{\mathcal{U}} \|z\|_{\mathcal{P}} \leq \frac{M_{\mathcal{U}}}{\mu \gamma^{-1} K^{s-1} - M_{\mathcal{U}}} (\mathcal{E}(t, z; u) + \kappa).$$

For  $\|z\|_{\mathcal{Z}} \leq K$  we have

$$|\partial_t \mathcal{E}(t, z; u)| \leq \gamma K M_{\mathcal{U}} \quad \text{and} \quad \mathcal{E}(t, z; u) \geq -\kappa - \gamma K M_{\mathcal{U}}.$$

Hence, the uniform control of the power, point (ii) of (A1), follows with  $c_1^E := M_{\mathcal{U}}/(\mu \gamma^{-1} K^{s-1} - M_{\mathcal{U}}) > 0$  and  $c_0^E := \kappa + \gamma K M_{\mathcal{U}} + \gamma K(\mu \gamma^{-1} K^{s-1} - M_{\mathcal{U}}) > 0$ .

By (W1), for  $\|z\|_{\mathcal{Z}} \geq K$  we also have

$$\mathcal{E}(t, z; u) \geq \mathcal{W}(z) - M_{\mathcal{U}} \|z\|_{\mathcal{P}} \geq \mu \|z\|_{\mathcal{Z}}^s - \kappa - \gamma M_{\mathcal{U}} \|z\|_{\mathcal{Z}} \geq (\mu K^{s-1} - \gamma M_{\mathcal{U}}) \|z\|_{\mathcal{Z}} - \kappa.$$

Thus, for all  $E \in \mathbb{R}$ , the set  $\bigcup_{u \in \mathcal{U}} \{z \in \mathcal{Z} : \mathcal{E}(t, z; u) \leq E\}$  is contained in the weakly sequentially compact  $\mathcal{Z}$ -ball with radius  $\max\{K, (E + \kappa)/(\mu K^{s-1} - \gamma M_{\mathcal{U}})\} < \infty$  and (A2) follows.

To show (C4), let  $u^k \xrightarrow{*} u$  in  $\mathcal{U}$  and  $y^k(t) \rightharpoonup y(t)$  in  $\mathcal{Z}$  for all  $t \in [0, T]$ , and assume  $\mathcal{E}(t, y^k(t), u^k) \leq E^*$  for all  $t \in [0, T]$ ,  $k \in \mathbb{N}$ , and some fixed  $E^* \in \mathbb{R}$ . Because  $y$  is the pointwise weak limit of Bochner-measurable functions  $y^k$ , it is itself weakly measurable (for each fixed  $z^* \in \mathcal{Z}^*$ , the real functions  $\tau \mapsto \langle z^*, y^k(\tau) \rangle$  are measurable and converge pointwise to the measurable function  $\tau \mapsto \langle z^*, y(\tau) \rangle$ ). Hence, by the Pettis theorem (see, e.g., [29, p. 131]),  $y$  is Bochner-measurable. Then, the measurability of  $\tau \mapsto \partial_t \mathcal{E}(\tau, y(\tau); u)$  follows from the fact that  $\partial_t \mathcal{E}(\cdot, \cdot; u)$  is a Carathéodory mapping.

As shown above, all sublevels of  $\mathcal{E}(t, \cdot; u)$  are uniformly (in  $t, u$ ) norm-bounded in  $\mathcal{P}$ , and hence the processes  $y^k$  and  $y$  are also uniformly (in  $t, k$ ) bounded in  $\mathcal{P}$ . Because  $\mathcal{Z} \xhookrightarrow{c} \mathcal{P}$ , the pointwise weak convergence in  $\mathcal{Z}$  of  $y^k(t)$  to  $y(t)$  implies  $y^k(t) \rightarrow y(t)$  (strongly) in  $\mathcal{P}$  for all  $t \in [0, T]$ . By Lebesgue's dominated convergence theorem, pointwise-a.e. (strong) convergence in  $\mathcal{P}$  implies (strong) convergence of  $y^k \rightarrow y$  in  $L^1(0, T; \mathcal{P})$  and  $\tau \mapsto \partial_t \mathcal{E}(\tau, y(\tau); u) \in L^1(0, T)$ . Because  $L^1(0, T; \mathcal{P})^* \cong L^\infty(0, T; \mathcal{P}^*)$  (see the appendix), we have for all  $t \in [0, T]$

$$\begin{aligned} \int_0^t \partial_t \mathcal{E}(\tau, y^k(\tau); u^k) \, d\tau &= \int_0^t \langle \dot{u}^k(\tau), y^k(\tau) \rangle_{\mathcal{P}^* \times \mathcal{P}} \, d\tau \\ &= \langle \dot{u}^k, y^k \rangle_{L^\infty(0, t; \mathcal{P}^*) \times L^1(0, t; \mathcal{P})} \\ &\longrightarrow \langle \dot{u}, y \rangle_{L^\infty(0, t; \mathcal{P}^*) \times L^1(0, t; \mathcal{P})} = \int_0^t \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau, \end{aligned}$$

because of the (weak  $\times$  strong)-continuity of the duality pairing. This shows (C4).

From Proposition A.2 and the (weak  $\times$  strong)-continuity of the duality pairing, for *every* fixed  $t \in [0, T]$ , the continuity of  $(u, z) \mapsto \langle u(t), z \rangle_{\mathcal{P}^* \times \mathcal{P}}$  with respect to sequences converging in  $(W^{1, \infty}(0, T; \mathcal{P}^*), \text{weak}^*) \times (\mathcal{Z}, \text{weak})$  follows. Hence,  $\mathcal{E}$  is the sum of the lower semicontinuous function  $\mathcal{W}$  and the continuous perturbation  $\langle u(t), z \rangle_{\mathcal{P}^* \times \mathcal{P}}$ ; therefore (C1') holds (take constant recovery sequences  $z^k := z$ ).

The convergence of suitable interpolants of the power (C5) can be seen by observing

$$\begin{aligned} &\left| \sum_{j=1}^N \int_{\tau_{j-1}}^{\tau_j} \partial_t \mathcal{E}(\xi, y(\tau_j); u) \, d\xi - \int_0^s \partial_t \mathcal{E}(\tau, y(\tau); u) \, d\tau \right| \\ &\leq M_{\mathcal{U}} \sum_{j=1}^N \int_{\tau_{j-1}}^{\tau_j} \|y(\tau_j) - y(\tau)\|_{\mathcal{P}} \, d\tau \end{aligned}$$

for any partition  $\Pi = (0 = \tau_0, \dots, \tau_N = s)$  and selecting a suitable sequence of partitions such that the last integrand vanishes by Lemma 4.12 from [4].  $\square$

REMARK 4.4. *Of course, the same proof also works in the case  $s = 1$  (linear growth) if we additionally require  $\mu\gamma^{-1} - M_{\mathcal{U}} > 0$ .*

REMARK 4.5. *Also time-dependent  $\mathcal{W} = \mathcal{W}(t, z)$  can be treated if (W1) is satisfied uniformly in  $t$  and if  $\mathcal{W}$  fulfills assumptions (A1), (C4), and (C5).*

For the dissipation distance  $\mathcal{D}: \mathcal{Z} \times \mathcal{Z} \times \mathcal{U} \rightarrow [0, \infty]$ , we require (A3) to hold. For reasons of simplicity, we also assume the property (C2'). Of course, more general situations are imaginable, in which only the weaker property (C2) is satisfied, but requiring (C2'), we immediately have (C3) as observed in Lemma 3.1.

REMARK 4.6. *If  $\mathcal{Z} \xhookrightarrow{c} \mathcal{P} \hookrightarrow \mathcal{V}$ , we can always choose the  $\mathcal{V}$ -norm (not depending on  $u \in \mathcal{U}$ ) as dissipation distance, i.e.,*

$$\mathcal{D}(z_1, z_2; u) := \|z_2 - z_1\|_{\mathcal{V}} \quad \text{for } z_1, z_2 \in \mathcal{Z}.$$

*Then, assumption (A3) holds by the properties of the norm. The validity of (C2') follows, because  $\|\cdot\|_{\mathcal{V}}$  is weaker than the weak topology in  $\mathcal{Z}$  since  $\mathcal{Z} \xhookrightarrow{c} \mathcal{V}$ .*

EXAMPLE 4.7. *Like in Example 4.1, consider the spaces  $\mathcal{Z} = H_0^1(\Omega)$ ,  $\mathcal{P} = L^2(\Omega)$ , and let  $\mathcal{D}(z_1, z_2; u) := \|z_2 - z_1\|_{L^1(\Omega)}$ . By the previous remark, this dissipation distance is admissible.*

**4.3. Initial value.** Concerning the compatibility condition (CC) on the initial value, Remark 3.5 applies by Proposition A.2 and we only have to show point (i) in (CC), i.e.,

$$y_0 \in \bigcap_{u \in \mathcal{U}} \mathcal{S}(0; u).$$

If  $\mathcal{D}(\cdot, \cdot; u) = \mathcal{D}(\cdot, \cdot)$  does not depend on the control  $u$ , then in our situation of a stored-energy functional  $\mathcal{E}(t, z; u) = \mathcal{W}(z) + \langle u(t), z \rangle_{\mathcal{P}^* \times \mathcal{P}}$ , this condition translates into

$$\mathcal{W}(y_0) + \langle u(0), y_0 \rangle_{\mathcal{P}^* \times \mathcal{P}} \leq \mathcal{W}(\hat{z}) + \langle u(0), \hat{z} \rangle_{\mathcal{P}^* \times \mathcal{P}} + \mathcal{D}(y_0, \hat{z})$$

for all  $\hat{z} \in \mathcal{Z}$ ,  $u \in \mathcal{U}$ . Adding the zero  $\mathcal{D}(y_0, y_0)$  on the left-hand side and rearranging, we see that this is equivalent to the variational inequality

$$\mathcal{W}(y_0) + \mathcal{D}(y_0, y_0) - \langle u(0), \hat{z} - y_0 \rangle_{\mathcal{P}^* \times \mathcal{P}} \leq \mathcal{W}(\hat{z}) + \mathcal{D}(y_0, \hat{z})$$

and (if  $z \mapsto \mathcal{W}(z) + \mathcal{D}(y_0, z)$  is convex, or if we extend the definition of subdifferential to nonconvex functions) to the differential inclusion

$$-u(0) \in \partial(\mathcal{W}(\cdot) + \mathcal{D}(y_0, \cdot))[y_0] \quad (\text{in } \mathcal{P}^*) \quad \text{for all } u \in \mathcal{U}.$$

If we have an additional condition on the initial value of  $u$ , e.g.,  $u(0) = 0$ , then this last condition can be statically checked before the optimization is started. In case of zero initial values  $u(0) = 0$  for the controls, the above condition translates to  $y_0 \in \mathcal{S}_{\mathcal{W}}(0)$ , where  $\mathcal{S}_{\mathcal{W}}(0)$  denotes the stability set at time  $t = 0$  for the functional  $\mathcal{W}$ .

**4.4. Cost functional.** Which functionals are now lower semicontinuous in our concrete setting?

EXAMPLE 4.8. One simple example of a cost functional  $\mathcal{J}$  satisfying (J2) is given through

$$\mathcal{J}(y; u) := \|y(T) - v\|_{\mathcal{Z}} \quad (\text{or } \|y(T) - v\|_{\mathcal{P}}),$$

where  $v \in \mathcal{Z}$  is some fixed element. This functional represents a prescribed terminal datum and seeks to find a solution minimizing the distance, measured in the  $\mathcal{Z}(\mathcal{P})$ -norm, to this fixed element. Condition (J2) holds by the weak lower semicontinuity of norms.

The following proposition is more general.

PROPOSITION 4.9. Let  $g: [0, T] \times \mathcal{P} \times \mathcal{U} \rightarrow \mathbb{R}_{\infty}$  be such that

- (i)  $g(\cdot, z; u): [0, T] \rightarrow \mathbb{R}_{\infty}$  is Lebesgue-measurable for fixed  $z \in \mathcal{P}$  and  $u \in \mathcal{U}$ .
- (ii)  $|g(t, z_1; u) - g(t, z_2; u)| \leq \omega_1(\|z_1 - z_2\|_{\mathcal{P}})$  for all  $z_1, z_2 \in \mathcal{Z}$ ,  $t \in [0, T]$ , and  $u \in \mathcal{U}$ , where  $\omega_1: [0, \infty) \rightarrow [0, \infty)$  is a modulus of continuity, i.e., increasing and  $\omega_1(0) = 0$  ( $\omega_1$  not depending on  $u$ ).
- (iii)  $g(t, z; \cdot): \mathcal{U} \rightarrow \mathbb{R}_{\infty}$  for fixed  $t \in [0, T]$  and  $z \in \mathcal{Z}$ , is continuous with respect to sequences converging weakly\* in  $\mathcal{U}$ .
- (iv)  $|g(t, z; u)| \leq h(t)(1 + \omega_2(\max\{\|z\|_{\mathcal{P}}, \|u\|_{\mathcal{U}}\}))$  for all  $t \in [0, T]$ , where the function  $\omega_2: [0, \infty) \rightarrow [0, \infty)$  is bounded on bounded sets and  $h \in L^1(0, T)$ .

Then the functional

$$\mathcal{J}(y; u) := \int_0^T g(t, y(t); u) \, dt$$

fulfills (J2).

*Proof.* We commence with a pointwise lower semicontinuity property of the integrand. For this, let  $z^k \rightarrow z$  strongly in  $\mathcal{P}$  and  $u^k \xrightarrow{*} u$  in  $\mathcal{U}$ . Then, for all  $t \in [0, T]$ , by (ii)

$$\begin{aligned} g(t, z; u) &= g(t, z; u) - g(t, z; u^k) + g(t, z; u^k) - g(t, z^k; u^k) + g(t, z^k; u^k) \\ &\leq (g(t, z; u) - g(t, z; u^k)) + \omega_1(\|z - z^k\|_{\mathcal{P}}) + g(t, z^k; u^k). \end{aligned}$$

By (iii),  $g(t, z; u) \leq \liminf_{k \rightarrow \infty} g(t, z^k; u^k)$ , and  $g(t, \cdot; \cdot): \mathcal{P} \times \mathcal{U} \rightarrow \mathbb{R}_{\infty}$  is (strong  $\times$  weak\*)-sequentially lower semicontinuous.

Now the integrand is the composition of the Carathéodory function  $g$  with a Bochner-measurable function  $y$  and hence it is measurable itself (this can be shown in a standard way using approximations of  $y$  by simple functions and exploiting assumptions (i), (ii)).

We now show the lower semicontinuity property (J2). For this, let  $u^k \xrightarrow{*} u$  in  $\mathcal{U}$ ,  $y^k \rightarrow y$  in  $BV_{\mathcal{D}}([0, T])$ , in particular  $y^k(t) \rightarrow y(t)$  strongly in  $\mathcal{P}$  for all  $t \in [0, T]$ , and  $E^* := \sup_{k \in \mathbb{N}, t \in [0, T]} \mathcal{E}(t, y^k(t); u^k) < \infty$ . By (A2), established in Lemma 4.3, the set  $\bigcup_{u \in \mathcal{U}} \{z \in \mathcal{Z} : \mathcal{E}(t, z; u) \leq E^*\}$  is relatively weakly compact. Thus,  $\|y^k(t)\|_{\mathcal{P}}$  stays uniformly (in  $t$  and  $k$ ) bounded. As  $u^k \xrightarrow{*} u$ , this also holds for  $\|u^k\|_{\mathcal{U}}$ , and (iv) yields an integrable majorant to  $|g(t, y^k(t); u^k)|$ . Thus, using Fatou's lemma, we infer from the pointwise lower semicontinuity of the integrand  $g$ ,

$$\mathcal{J}(y; u) \leq \liminf_{k \rightarrow \infty} \mathcal{J}(y^k; u^k),$$

i.e., (J2).  $\square$



Note that the modulus of continuity  $\omega_1$  from (ii) may depend on a bound on  $\|z_1\|_{\mathcal{P}}, \|z_2\|_{\mathcal{P}}$  as well, because the uniform energy bound required in (J2) implies as before that for any sequence of  $y^k$  occurring in (J2) we can find a uniform (in  $t, k$ ) bound on  $\|y^k(t)\|_{\mathcal{P}}$ .

An important example for such a cost functional  $\mathcal{J}$  is the following.

EXAMPLE 4.10. For  $g(t, z; u) := \|z - w(t)\|_{\mathcal{P}}^2$  with  $w \in L^\infty(0, T; \mathcal{P})$ , the last proposition and the small remark preceding this example show that

$$\mathcal{J}(y; u) := \int_0^T \|y(t) - w(t)\|_{\mathcal{P}}^2 dt = \|y - w\|_{L^2(0, T; \mathcal{P})}^2$$

is an admissible cost functional. We can now optimize the control as to make the system process follow this prescribed movement  $w$  as closely as possible in the  $L^2$ -sense.

**5. Applications.** This section presents two concrete applications of the abstract theory developed so far.

**5.1. Optimal control of a partial differential inclusion.** One possible application of the theory presented so far is the task to (optimally) control solutions of the initial-boundary value problem for the partial differential inclusion [3, 13]

$$(5.1) \quad \begin{cases} 0 \in \kappa \operatorname{Sign}(\dot{y}) - \operatorname{div}[a \nabla y] + D_y F(\cdot, y) - u & \text{in } [0, T] \times \Omega, \\ y = 0 & \text{on } [0, T] \times \partial\Omega, \quad \text{and} \quad y(0, \cdot) = y_0 & \text{in } \Omega. \end{cases}$$

Here,  $\Omega$  is a bounded Lipschitz domain in  $\mathbb{R}^d$ ,  $a \in L^\infty(\Omega)$  is a function bounded from below by  $2\mu > 0$ , and  $\kappa \in L^\infty(\Omega)$  is a function bounded from above and below by positive constants. The function  $F: \Omega \times \mathbb{R} \rightarrow [0, \infty]$  is assumed bounded, continuous, and differentiable in the second argument. As usual,  $\dot{y}$  and  $\nabla y$  denote the temporal derivative and the spatial gradient, respectively. We want to control  $u: [0, T] \times \Omega \rightarrow \mathbb{R}$  under the constraints

$$(5.2) \quad u(0, \cdot) = 0, \quad \varphi_*(t, x) \leq u(t, x) \leq \varphi^*(t, x), \quad \psi_*(t, x) \leq \partial_t u(t, x) \leq \psi^*(t, x)$$

for all  $t \in [0, T]$  and a.e.  $x \in \Omega$ , where  $\varphi_*, \varphi^*, \psi_*, \psi^* \in L^\infty(0, T; L^2(\Omega))$  (note that we do not identify functions which are equal a.e. in time), and the pointwise condition

$$(5.3) \quad \gamma_*(x) \leq y(t, x) \leq \gamma^*(x) \quad \text{for all } t \in [0, T] \text{ and a.e. } x \in \Omega,$$

where  $\gamma_*, \gamma^*: \Omega \rightarrow \mathbb{R} \cup \{-\infty, +\infty\}$ . Note that if  $\gamma_* \equiv -\infty$ ,  $\gamma^* \equiv +\infty$ , then we have not imposed any pointwise condition on  $y$  at all.

Concerning the spaces, we choose  $\mathcal{Z} := H_0^1(\Omega)$ ,  $\mathcal{P} := L^2(\Omega)$  (see Example 4.1) and

$$\mathcal{U} := \{u \in W^{1,\infty}(0, T; L^2(\Omega)) : u \text{ satisfies condition (5.2)}\}.$$

For any  $p \in (1, \infty)$ , consider  $\mathcal{U}$  as a subset of  $W^{1,p}(0, T; L^2(\Omega))$ , and observe that in this space,  $\mathcal{U}$  is weakly closed (since it is strongly closed and convex). Hence, because weak\* convergence in the space  $W^{1,\infty}(0, T; L^2(\Omega))$  implies weak convergence in the space  $W^{1,p}(0, T; L^2(\Omega))$ , the control space  $\mathcal{U}$  is weakly\* closed in  $W^{1,\infty}(0, T; L^2(\Omega))$  (the boundedness allows us to work with sequences). The pointwise restriction (5.3) is weakly closed in  $H_0^1(\Omega)$ , because it only affects the functions itself (and not their gradients) and for these we can assume pointwise a.e. convergence.

The energetic formulation of the partial differential inclusion (5.1) is based on the following functionals:

$$\begin{aligned}\mathcal{E}(t, z; u) &:= \mathcal{W}(z) - \int_{\Omega} u(t, x) z(x) \, dx, \\ \mathcal{W}(z) &:= \int_{\Omega} \frac{a(x)}{2} |\nabla z(x)|^2 + F(x, z(x)) \, dx, \\ \mathcal{D}(z_1, z_2) &:= \int_{\Omega} \kappa(x) |z_2(x) - z_1(x)| \, dx.\end{aligned}$$

For the initial value, we require

$$y_0 \in \mathcal{S}_{\mathcal{W}}(0) = \operatorname{Argmin}\{ \mathcal{W}(\hat{z}) + \mathcal{D}(y_0, \hat{z}) : \hat{z} \in H_0^1(\Omega) \}.$$

While this might still be difficult to verify (note that  $y_0$  also occurs under the integral), by section 4.3, this is exactly the compatibility condition (CC) in our concrete setting. The system, however, self-corrects an inadmissible initial value if necessary, cf. Remark 3.6.

As the cost functional, we choose

$$(5.4) \quad \mathcal{J}(y) := \int_0^T \int_{\Omega} |y(t, x) - w(t, x)|^2 \, dx \, dt,$$

i.e., we want to make the solution follow the function  $w \in L^\infty(0, T; L^2(\Omega))$  (cf. Example 4.10).

By the positivity of  $a(\cdot)$ , it is clear that (W1) holds with  $s = 2$ ,  $\mu > 0$  (the lower bound for  $a(\cdot)$  was  $2\mu$ ), and  $\kappa = 0$ . Condition (W2) follows from the fact that sublevels of  $\mathcal{W}$  must be closed by the weak lower semicontinuity of norms and the term with  $F$  is weakly continuous by Lebesgue's dominated convergence theorem. Hence, Lemma 4.3 implies (A1), (A2), (C1'), (C4), and (C5). From the assumptions on  $\kappa(\cdot)$ ,  $\mathcal{D}(\cdot, 0)$  is an equivalent norm to the norm of  $L^1(\Omega)$  and Remark 4.6 (see also Example 4.7) yields the validity of assumptions (A3) and (C2'). Lemma 3.1 gives (C3).

Finally, our choice for  $\mathcal{J}$  is admissible by Example 4.10. Thus, Theorem 3.4 (also cf. Remark 3.7) shows solvability of the optimal control problem, meaning that there exists an optimal control  $u \in W^{1,\infty}(0, T; L^2(\Omega))$  and a corresponding solution process  $y \in \operatorname{BV}_{\mathcal{D}}([0, T]; L^1(\Omega)) \cap L^\infty(0, T; H_0^1(\Omega))$  fulfilling (5.1)–(5.3), and minimizing the cost functional  $\mathcal{J}$  as defined in (5.4) over all such constrained choices for the control and solution.

**5.2. Stackelberg games and the foreign exchange market.** The optimal control problem (OC) can be viewed as an asymmetric (evolutionary) Stackelberg game, which we can employ to look at a simplistic proof-of-concept model of currency trading.

Set  $\mathcal{Z} := \{ z \in \mathbb{R}^n : z \geq 0 \text{ and } |z|_1 = 1 \}$ , where  $|\cdot|_1$  denotes the 1-norm. We think of  $z \in \mathcal{Z}$  as representing the fractions of a trader's total amount of money, which are invested in one of  $n$  different currencies

Define the *real-value functional*

$$\tilde{\mathcal{E}}(t, z; u) := (r(t) - f(z) + u(t)) \cdot z \quad \text{and} \quad \mathcal{D}(z, \tilde{z}) := \frac{\gamma}{2} |\tilde{z} - z|_1 \quad \text{for } z, \tilde{z} \in \mathbb{R}^n$$

with a (in reality estimated or predicted) *rating function*  $r \in W^{1,\infty}(0, T; \mathbb{R}^n)$  for the currencies ( $r(t)_j > 1$ , the  $j$ th component of  $r(t)$ , means that the currency is undervalued and  $r(t)_j < 1$  means that it is overvalued) and a Lipschitz-continuous *feedback function*  $f \in C(\mathbb{R}^n; \mathbb{R}^n)$  simulating the effect of large buys on the currency's real value (with respect to purchasing power, "the law of one price"). As a simple choice, one can take  $f(z) = cz$  with  $c > 0$  (not too large). Then, buying more currency will make the real value of the currency decrease linearly (adjustment of exchange rates). We assume that  $f$  satisfies the coercivity inequality  $f(z) \cdot z \geq \mu |z|_1^s - \kappa$  for some fixed  $\mu > 0$ ,  $\kappa \geq 0$ ,  $s > 1$ . The control  $u \in \mathcal{U} := \{u \in W^{1,\infty}(0, T; \mathbb{R}^n) : \|u\|_{W^{1,\infty}(0, T; \mathbb{R}^n)} \leq M_{\mathcal{U}}\}$ ,  $M_{\mathcal{U}} > 0$ , represents the external command we have over the market (for example, a central bank can raise or lower interests). Finally, with  $\gamma > 0$  expressing the fractional trading costs (assumed uniform for all currencies),  $\mathcal{D}$  models the total costs when currencies are exchanged.

At each point in time, the  $j$ th component of the vector  $r(t) - f(y(t)) + u(t)$  represents the total value factor of the  $j$ th currency. The currency traders will now want to maximize the value of their money, but taking into account the trading costs. If  $z_0 \in \mathcal{Z}$  is the current state and we are at time  $t \in (0, T]$ , then the traders will always seek to maximize  $\hat{\mathcal{E}}(t, \hat{z}; u) - \mathcal{D}(z_0, \hat{z})$  over all  $\hat{z} \in \mathcal{Z}$ . Setting  $\mathcal{E}(t, z; u) := -\hat{\mathcal{E}}(t, z; u)$  and switching from a maximization to a minimization problem, we arrive at the discrete-time incremental problem (IP). Hence, the energetic formulation (S) and (E) with the functionals  $\mathcal{E}$  and  $\mathcal{D}$  is the continuous-time system that models the described behavior.

We want to exercise our control over the market to reach some objective, for example keeping the real currency values as close to 1 as possible (in the Euclidean sense). The latter goal can be modeled by taking

$$\mathcal{J}(y, u) := \int_0^T |r(t) - f(y(t)) + u(t) - (1, \dots, 1)^T|_2^2 \, dt.$$

It should be noted that if several solutions  $y$  exist for a fixed control  $u$ , then here we "choose" the best one as to minimize our functional. This optimistic view might not be realistic in certain applications. However, if unique solvability of the rate-independent system is guaranteed or if we have another mechanism of forcing the "right" solution, then this problem does not occur. Other cases might lead to saddle point problems and are not considered here.

Just as before, the tools of section 4 are applicable (with  $\mathcal{Z} = \mathcal{P} = \mathbb{R}^n$  and the  $|\cdot|_1$ -norm) and Theorem 3.4 shows solvability of the optimal control problem.

**Appendix. Sobolev–Bochner spaces.** This appendix presents some facts about Sobolev–Bochner spaces. In particular, for the convenience of the reader we prove a rather well-known compactness result in a somewhat more direct fashion than in [26]. For a general reference on these spaces, see, e.g., [25, Chapter 7] and [7, 26].

For a separable, reflexive Banach space  $\mathcal{X}$ , denote by  $L^p(0, T; \mathcal{X})$ ,  $p \in [1, \infty]$ , the space of  $p$ -Bochner-integrable functions with values in  $\mathcal{X}$ . It is well known that for  $p \in [1, \infty)$ , the space  $L^p(0, T; \mathcal{X})$  is separable and  $L^p(0, T; \mathcal{X})^* \cong L^{p'}(0, T; \mathcal{X}^*)$ , where  $1/p + 1/p' = 1$ ; cf. [7, section IV.1].

For a separable, reflexive Banach space  $\mathcal{X}$ , and a separable Banach space  $\mathcal{Y}$  with  $\mathcal{X} \hookrightarrow \mathcal{Y}$ , the *Sobolev–Bochner space*  $W^{1,p}(0, T; \mathcal{X}; \mathcal{Y})$ ,  $p \in [1, \infty]$  is defined as

$$W^{1,p}(0, T; \mathcal{X}; \mathcal{Y}) := \{u \in L^p(0, T; \mathcal{X}) : \dot{u} \in L^p(0, T; \mathcal{Y})\},$$

where  $\dot{u}$  is the generalized (or distributional) time-derivative, i.e.,

$$\int_0^T u(t) \dot{\varphi}(t) \, dt = - \int_0^T \dot{u}(t) \varphi(t) \, dt \quad (\text{in } \mathcal{Y}) \quad \text{for all } \varphi \in C_c^\infty(0, T).$$

The space  $W^{1,p}(0, T; \mathcal{X}; \mathcal{Y})$  is a Banach space with the norm

$$\|u\|_{W^{1,p}(0, T; \mathcal{X}; \mathcal{Y})} := \begin{cases} \|u\|_{L^p(0, T; \mathcal{X})} + \|\dot{u}\|_{L^p(0, T; \mathcal{Y})} & \text{if } p \in [1, \infty), \\ \max \{ \|u\|_{L^\infty(0, T; \mathcal{X})}, \|\dot{u}\|_{L^\infty(0, T; \mathcal{Y})} \} & \text{if } p = \infty. \end{cases}$$

If  $\mathcal{X} = \mathcal{Y}$ , we write  $W^{1,p}(0, T; \mathcal{X})$  for  $W^{1,p}(0, T; \mathcal{X}; \mathcal{X})$ . If there is another Banach space  $\mathcal{Y}$  with  $\mathcal{X} \hookrightarrow \mathcal{Y}$ , it is obvious that  $W^{1,p}(0, T; \mathcal{X}) \hookrightarrow W^{1,p}(0, T; \mathcal{X}; \mathcal{Y})$ .

For  $p \in (1, \infty)$  and  $\mathcal{X}, \mathcal{Y}$  reflexive, by identifying  $W^{1,p}(0, T; \mathcal{X}; \mathcal{Y})$  with a closed subspace of the product  $L^p(0, T; \mathcal{X}) \times L^p(0, T; \mathcal{Y})$  of reflexive Banach spaces [7, Bemerkung IV.1.11], one easily sees that  $W^{1,p}(0, T; \mathcal{X}; \mathcal{Y})$  is reflexive. Weak convergence  $u_n \rightharpoonup u$  of a sequence  $(u_n)_n \subseteq W^{1,p}(0, T; \mathcal{X}; \mathcal{Y})$ ,  $p \in (1, \infty)$ , means that  $u_n \rightharpoonup u$  in  $L^p(0, T; \mathcal{X})$  and  $\dot{u}_n \rightharpoonup \dot{u}$  in  $L^p(0, T; \mathcal{Y})$ . For  $p = \infty$ ,  $u_n \overset{*}{\rightharpoonup} u$  in  $W^{1,\infty}(0, T; \mathcal{X}; \mathcal{Y})$  means  $u_n \overset{*}{\rightharpoonup} u$  in  $L^\infty(0, T; \mathcal{X})$  and  $\dot{u}_n \overset{*}{\rightharpoonup} \dot{u}$  in  $L^\infty(0, T; \mathcal{Y})$ .

For  $p' \in (1, \infty]$ , norm-bounded sets in  $W^{1,p'}(0, T; \mathcal{X}^*)$  are weakly (weakly\* if  $p' = \infty$ ) relatively compact. More precisely, let  $(u_n)_n \subseteq W^{1,p'}(0, T; \mathcal{X}^*)$  be bounded. Then, since  $L^p(0, T; \mathcal{X})$  is the separable predual to  $L^{p'}(0, T; \mathcal{X}^*)$ , where  $1/p + 1/p' = 1$ , there exists a subsequence (not relabelled) such that  $u_n \rightharpoonup u$  ( $u_n \overset{*}{\rightharpoonup} u$ ) and  $\dot{u} \rightharpoonup v$  ( $\dot{u} \overset{*}{\rightharpoonup} v$ ) in  $L^{p'}(0, T; \mathcal{X}^*)$  and an integration by parts shows  $v = \dot{u}$ .

The following lemma shows how we can deduce uniform (strong) convergence from  $L^p$ -convergence if the involved functions are uniformly equicontinuous.

**LEMMA A.1.** *Let  $\mathcal{X}$  be a Banach space, and let  $(u_n)_n \subseteq C([0, T]; \mathcal{X})$ ,  $n \in \mathbb{N} \cup \{\infty\}$ , be a sequence of uniformly equicontinuous functions, i.e., for all  $\varepsilon > 0$  we can find  $\delta > 0$  such that*

$$\|u_n(s) - u_n(r)\|_{\mathcal{X}} < \varepsilon \quad \text{for all } n \in \mathbb{N}_\infty \text{ and } r, s \in [0, T] \text{ with } |s - r| < \delta.$$

*Then, for all  $p \in [1, \infty]$ ,  $u_n \rightarrow u_\infty$  in  $L^p(0, T; \mathcal{X})$  implies  $u_n \rightarrow u_\infty$  in  $C([0, T]; \mathcal{X})$ .*

*Proof.* The case  $p = \infty$  is trivial; therefore let  $p \in [1, \infty)$ . By contradiction, assume that there exists a (sub)sequence  $(t_n)_n \subseteq [0, T]$  such that  $\|u_\infty(t_n) - u_n(t_n)\|_{\mathcal{X}} \geq 3\varepsilon > 0$  for all  $n \in \mathbb{N}$ . As  $[0, T]$  is compact, we can assume without loss of generality that  $t_n \rightarrow t \in [0, T]$ . By equicontinuity, choose  $\delta \in (0, T)$  such that

$$\|u_n(s) - u_n(r)\|_{\mathcal{X}} \leq \varepsilon$$

for all  $n \in \mathbb{N}_\infty$  and all  $r, s \in [0, T]$  with  $|s - r| < \delta$ . For  $n$  so large that  $|t - t_n| < \delta/2$  and all  $s \in [0, T]$  with  $|s - t| < \delta/2$ , in particular  $|s - t_n| < \delta$ , it holds that

$$\begin{aligned} \|u_n(s) - u_\infty(s)\|_{\mathcal{X}} &\geq \left\| (u_n(s) - u_n(t_n)) - (u_\infty(s) - u_\infty(t_n)) \right\|_{\mathcal{X}} \\ &\quad - \|u_\infty(t_n) - u_n(t_n)\|_{\mathcal{X}} \geq \varepsilon. \end{aligned}$$

Hence,

$$\begin{aligned} \|u_n - u_\infty\|_{L^p(0, T; \mathcal{X})}^p &= \int_0^T \|u_n(\tau) - u_\infty(\tau)\|_{\mathcal{X}}^p \, d\tau \\ &\geq \int_{\max\{t-\delta/2, 0\}}^{\min\{t+\delta/2, T\}} \|u_n(\tau) - u_\infty(\tau)\|_{\mathcal{X}}^p \, d\tau \geq \frac{\delta \varepsilon^p}{2}, \end{aligned}$$

contradicting  $\|u_n - u_\infty\|_{L^p(0,T;\mathcal{X})} \rightarrow 0$ .  $\square$

Applying this lemma, we can show the compact embedding  $W^{1,p}(0,T;\mathcal{X};\mathcal{Y}) \xhookrightarrow{c} C([0,T];\mathcal{X})$  if  $\mathcal{X} \xhookrightarrow{c} \mathcal{Y}$ .

PROPOSITION A.2. *If  $\mathcal{X} \xhookrightarrow{c} \mathcal{Y}$ , then it holds that  $W^{1,p}(0,T;\mathcal{X};\mathcal{Y}) \xhookrightarrow{c} C([0,T];\mathcal{Y})$  for all  $p \in (1, \infty]$ .*

*Proof.* It is a well-known fact that  $W^{1,p}(0,T;\mathcal{X};\mathcal{Y}) \hookrightarrow C([0,T];\mathcal{Y})$ ; see, e.g., [25, Lemma 7.1]. The continuous representative of  $v \in W^{1,p}(0,T;\mathcal{X};\mathcal{Y})$  is given through

$$\tilde{v}(t) = \tilde{v}(0) + \int_0^t v'(\tau) \, d\tau \quad (\in \mathcal{Y})$$

(note that  $v' \in L^1(0,T;\mathcal{Y})$  by the Hölder inequality, and that  $\tilde{v}(0)$  has a sense with respect to the continuous representative). We therefore have

$$(A.1) \quad \|\tilde{v}(t) - \tilde{v}(s)\|_{\mathcal{Y}} \leq \int_s^t \|v'(\tau)\|_{\mathcal{Y}} \, d\tau \leq (t-s)^{(p-1)/p} \|v'\|_{L^p(0,T;\mathcal{Y})},$$

by the Hölder inequality.

Since the space  $C([0,T];\mathcal{Y})$  is metric, it suffices to work with sequences to show the compact embedding. Therefore let  $(v_n)_n$  be a bounded sequence in  $W^{1,p}(0,T;\mathcal{X};\mathcal{Y})$ . The Aubin–Lions lemma [25, Lemma 7.7] applied to the inclusions  $\mathcal{X} \xhookrightarrow{c} \mathcal{Y} \hookrightarrow \mathcal{Y}$  shows  $W^{1,p}(0,T;\mathcal{X};\mathcal{Y}) \xhookrightarrow{c} L^p(0,T;\mathcal{Y})$ ; hence, we can select a subsequence (not relabelled) such that  $v_n \rightarrow v$  (strongly) in  $L^1(0,T;\mathcal{X})$ . The last estimate (A.1) shows that the sequence  $(v_n)_n$  and its limit  $v$  are uniformly equicontinuous. Now, Lemma A.1 implies  $v_n \rightarrow v$  (strongly) in  $C([0,T];\mathcal{Y})$ . Hence, the compact embedding  $W^{1,p}(0,T;\mathcal{X};\mathcal{Y}) \xhookrightarrow{c} C([0,T];\mathcal{Y})$  holds.  $\square$

In the situation of the preceding proposition, it is clear that  $W^{1,p}(0,T;\mathcal{X}) \hookrightarrow W^{1,p}(0,T;\mathcal{X};\mathcal{Y}) \xhookrightarrow{c} C([0,T];\mathcal{Y})$ , and hence the result also holds for  $W^{1,p}(0,T;\mathcal{X})$ .

**Acknowledgments.** I thank Alexander Mielke, Jiří Outrata, Martin Kružík, and Tomáš Roubíček for fruitful discussions and for reading preliminary versions of this work. Also, I wish to thank the anonymous referees for their comments, which led to the improvement of this paper.

## REFERENCES

- [1] H.-D. ALBER, *Materials with Memory*, Lecture Notes in Math. 1682, Springer-Verlag, Berlin, 1998.
- [2] M. BROKATE, P. KREJČÍ, AND H. SCHNABEL, *On uniqueness in evolution quasivariational inequalities*, J. Convex Anal., 11 (2004), pp. 111–130.
- [3] P. COLLI AND A. VISINTIN, *On a class of doubly nonlinear evolution equations*, Comm. Partial Differential Equations, 15 (1990), pp. 737–756.
- [4] G. DAL MASO, G. A. FRANCFORT, AND R. TOADER, *Quasistatic crack growth in nonlinear elasticity*, Arch. Ration. Mech. Anal., 176 (2005), pp. 165–225.
- [5] G. FRANCFORT AND J.-J. MARIGO, *Revisiting brittle fracture as an energy minimization problem*, J. Mech. Phys. Solids, 46 (1998), pp. 1319–1342.
- [6] G. FRANCFORT AND A. MIELKE, *Existence results for a class of rate-independent material models with nonconvex elastic energies*, J. Reine Angew. Math., 595 (2006), pp. 55–91.
- [7] H. GAJEWSKI, K. GRÖGER, AND K. ZACHARIAS, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Akademie-Verlag, Berlin, 1974.
- [8] W. HAN AND B. REDDY, *Plasticity. Mathematical Theory and Numerical Analysis*, Interdiscip. Appl. Math. 9, Springer-Verlag, New York, 1999.

- [9] M. KOČVARA, M. KRUŽÍK, AND J. OUTRATA, *On the control of an evolutionary equilibrium in micromagnetics*, in Optimization with Multivalued Mappings, S. Dempe and V. Kalashnikov, eds., Springer, New York, 2006, pp. 143–168.
- [10] M. KOČVARA, A. MIELKE, AND T. ROUBÍČEK, *A rate-independent approach to the delamination problem*, Math. Mech. Solids, 11 (2006), pp. 423–447.
- [11] M. KOČVARA AND J. OUTRATA, *On the modeling and control of delamination processes*, in Control and Boundary Analysis, J. Cagnol and J.-P. Zolesion, eds., Lecture Notes in Pure and Appl. Math. 240, Chapman and Hall, Boca Raton, FL, 2004, pp. 169–187.
- [12] M. KRUŽÍK, A. MIELKE, AND T. ROUBÍČEK, *Modelling of microstructure and its evolution in shape-memory-alloy single-crystals, in particular in CuAlNi*, Meccanica, 40 (2005), pp. 389–418.
- [13] A. MAINIK AND A. MIELKE, *Existence results for energetic models for rate-independent systems*, Calc. Var. Partial Differential Equations, 22 (2005), pp. 73–99.
- [14] A. MIELKE, *Energetic formulation of multiplicative elasto-plasticity using dissipation distances*, Contin. Mech. Thermodyn., 15 (2003), pp. 351–382.
- [15] A. MIELKE, *Evolution of rate-independent systems*, in Handb. Differ. Equa., Evolutionary Equations, Vol. 2, C. Dafermos and E. Feireisl, eds., Elsevier/North-Holland, Amsterdam, 2005, pp. 461–559.
- [16] A. MIELKE, *A mathematical framework for generalized standard materials in the rate-independent case*, in Multifield Problems in Solid and Fluid Mechanics, Lecture Notes in Appl. Computat. Mech. 28, R. Helmig, A. Mielke, and B. I. Wohlmuth, eds., Springer-Verlag, Berlin, 2006, pp. 351–379.
- [17] A. MIELKE, T. ROUBÍČEK, AND U. STEFANELLI,  *$\Gamma$ -limits and relaxations for rate-independent evolutionary problems*, Calc. Var. Partial Differential Equations, 31 (2008), pp. 387–416.
- [18] A. MIELKE AND F. THEIL, *A mathematical model for rate-independent phase transformations with hysteresis*, in Proceedings of the Workshop on Models of Continuum Mechanics in Analysis and Engineering, H.-D. Alber, R. Baean, and R. Farwig, eds., Shaker-Verlag, Herzogenrath, Germany, 1999, pp. 117–129.
- [19] A. MIELKE AND F. THEIL, *On rate-independent hysteresis models*, NoDEA Nonlinear Differential Equations Appl., 11 (2004), pp. 151–189.
- [20] A. MIELKE, F. THEIL, AND V. LEVITAS, *A variational formulation of rate-independent phase transformations using an extremum principle*, Arch. Ration. Mech. Anal., 162 (2002), pp. 137–177.
- [21] A. MIELKE AND A. TIMOFTE, *An energetic material model for time-dependent ferroelectric behavior: Existence and uniqueness*, Math. Methods Appl. Sci., 29 (2006), pp. 1393–1410.
- [22] P. NEITTAANMÄKI AND D. TIBA, *Optimal Control of Nonlinear Parabolic Systems*, Pure Appl. Math. 179, Marcel Dekker, New York, 1994.
- [23] J. OUTRATA, M. KOČVARA, AND J. ZOWE, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, Nonconvex Optim. Appl. 28, Kluwer, Dordrecht, 1998.
- [24] A. REIMERS AND E. D. TORRE, *Fast Preisach-based magnetization model and fast inverse hysteresis model*, IEEE Trans. Mag., 34 (1998), pp. 3857–3866.
- [25] T. ROUBÍČEK, *Nonlinear Partial Differential Equations with Applications*, Internat. Ser. Numer. Math. 153, Birkhäuser Verlag, Basel, 2005.
- [26] J. SIMON, *Compact sets in the space  $L^p(0, T; B)$* , Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
- [27] X. TAN, J. BARAS, AND P. KRISHNAPRASAD, *Control of hysteresis in smart actuators with application to micro-positioning*, Systems Control Lett., 54 (2005), pp. 483–492.
- [28] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [29] K. YOSIDA, *Functional Analysis*, 6th ed., Grundlehren math. Wiss. 123, Springer-Verlag, Berlin, 1980.

## THE MODIFIED OPTIMAL $\mathcal{H}_\infty$ CONTROL PROBLEM FOR DESCRIPTOR SYSTEMS\*

PHILIP LOSSE<sup>†</sup>, VOLKER MEHRMANN<sup>‡</sup>, LISA KATRIN POPPE<sup>‡</sup>, AND TIMO REIS<sup>‡</sup>

**Abstract.** The  $\mathcal{H}_\infty$  control problem is studied for linear constant coefficient descriptor systems. Necessary and sufficient optimality conditions are derived for systems of arbitrary index. These conditions are formulated in terms of deflating subspaces of even matrix pencils containing only the parameters of the original system. It is shown that this approach leads to a more numerically robust and efficient method in computing the optimal value  $\gamma$  in contrast to other methods such as the widely used Riccati- and linear matrix inequality (LMI)-based approaches. The results are illustrated by a numerical example.

**Key words.** descriptor system,  $\mathcal{H}_\infty$  control, algebraic Riccati equation, even matrix pencil,  $\gamma$ -iteration, deflating subspace

**AMS subject classifications.** 34A09, 93B40, 93B36, 65F15, 65L80, 93B52, 93C05

**DOI.** 10.1137/070710093

**1. Introduction.** Solving the optimal infinite-horizon output (or measurement) feedback  $\mathcal{H}_\infty$  control problem is one of the central tasks in robust control; see, e.g., [16, 19, 28, 39, 40]. For standard state space systems, where the dynamics of the system is modeled by a linear constant coefficient ordinary differential equation (ODE), the analysis of this problem is well studied [10] and numerical methods have been developed and integrated in control software packages such as [5, 20, 1, 29]. These methods work well for a wide range of problems in computing close to optimal (sub-optimal) controllers, but the exact computation of the optimal value  $\gamma$  in  $\mathcal{H}_\infty$  control is considered a challenge [8]. In order to avoid some of the numerical difficulties that arise when approaching the optimum, in [3, 4] several improvements in the previously known methods were presented. These are based on the solution of structured eigenvalue problems with structured methods.

In this paper we study the more general case that the dynamics is constrained, i.e., described by a *differential-algebraic equation (DAE)* or *descriptor system*. Descriptor systems arise in the control of constrained mechanical systems (see, e.g., [12, 30, 35, 36, 37]), in electrical circuit simulation (see, e.g., [18, 17]), and in particular in heterogeneous systems, where different models are coupled [27].

Robust control for descriptor systems has been studied in [31, 32, 33] using linear matrix inequalities (LMIs) and in [38] via generalized Riccati equations and  $J$ -spectral factorization. In contrast to these approaches, we extend the analysis and the robust numerical methods that were derived via deflating subspaces in [3, 4]. We discuss

---

\*Received by the editors December 4, 2007; accepted for publication (in revised form) July 21, 2008; published electronically November 5, 2008.

<http://www.siam.org/journals/sicon/47-6/71009.html>

<sup>†</sup>Fakultät für Mathematik, TU Chemnitz, Reichenhainer Straße 41, D-09126 Chemnitz, Germany (philip.losse@mathematik.tu-chemnitz.de). The research of this author was supported by *Deutsche Forschungsgemeinschaft* through the project BE-2174/6-1,2.

<sup>‡</sup>Institut für Mathematik, TU Berlin, Straße des 17. Juni 136, D-10623 Berlin, Germany (mehrman@math.tu-berlin.de, poppe@math.tu-berlin.de, reis@math.tu-berlin.de). The second and third authors were partially supported by *Deutsche Forschungsgemeinschaft* through the project ME 790/16-1.

descriptor systems of the form

$$(1.1) \quad \mathbf{P} : \begin{aligned} E\dot{x}(t) &= Ax(t) + B_1w(t) + B_2u(t), & x(t_0) &= x^0, \\ z(t) &= C_1x(t) + D_{11}w(t) + D_{12}u(t), \\ y(t) &= C_2x(t) + D_{21}w(t) + D_{22}u(t), \end{aligned}$$

where  $E, A \in \mathbb{R}^{n,n}$ ,  $B_i \in \mathbb{R}^{n,m_i}$ ,  $C_i \in \mathbb{R}^{p_i,n}$ , and  $D_{ij} \in \mathbb{R}^{p_i,m_j}$  for  $i, j = 1, 2$ . (Here, by  $\mathbb{R}^{k,l}$  we denote the set of real  $k \times l$  matrices.)

In this system,  $x(t) \in \mathbb{R}^n$  is the state vector,  $u(t) \in \mathbb{R}^{m_2}$  is the control input vector, and  $w(t) \in \mathbb{R}^{m_1}$  is an exogenous input that may include noise, linearization errors, and unmodeled dynamics. The vector  $y(t) \in \mathbb{R}^{p_2}$  contains measured outputs, while  $z(t) \in \mathbb{R}^{p_1}$  is a regulated output or an estimation error. Our approach can also be extended to rectangular systems, and systems in behavior formulation, using a remodeling, as was suggested in [23, 22] (see also [24]), but here we study only the formulation in (1.1).

The optimal  $\mathcal{H}_\infty$  control problem is typically formulated in a frequency domain. For this we need the following notation. The space  $\mathcal{H}_\infty^{p,m}$  consists of all  $\mathbb{C}^{p,m}$ -valued functions that are analytic and bounded in the complex half-plane

$$\mathbb{C}^+ = \{s \in \mathbb{C} : \operatorname{Re}(s) > 0\}.$$

For  $F \in \mathcal{H}_\infty^{p,m}$  the  $\mathcal{H}_\infty$ -norm is given by

$$\|F\|_\infty = \sup_{s \in \mathbb{C}^+} \sigma_{\max}(F(s)),$$

where  $\sigma_{\max}(F(s))$  denotes the maximal singular value of the matrix  $F(s)$ .

In robust control,  $\|F\|_\infty$  is used as a measure of the worst-case influence of the disturbances  $w$  on the output  $z$ , where in this case  $F$  is the transfer function mapping noise or disturbance inputs to error signals [40].

Solving the optimal  $\mathcal{H}_\infty$  control problem is the task of designing a dynamic controller, as presented in Figure 1.1, that minimizes (or at least approximately minimizes) this measure.

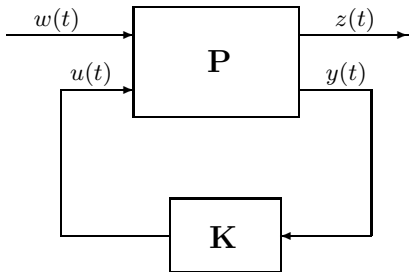


FIG. 1.1. Interconnection with controller.

Put more rigorously, the optimal  $\mathcal{H}_\infty$  control problem is the following.

DEFINITION 1.1 (the optimal  $\mathcal{H}_\infty$  control problem). *For the descriptor system (1.1), determine a controller (dynamic compensator)*

$$(1.2) \quad \mathbf{K} : \begin{aligned} \hat{E}\dot{\hat{x}}(t) &= \hat{A}\hat{x}(t) + \hat{B}y(t), \\ u(t) &= \hat{C}\hat{x}(t) + \hat{D}y(t) \end{aligned}$$



with  $\hat{E}, \hat{A} \in \mathbb{R}^{N,N}$ ,  $\hat{B} \in \mathbb{R}^{N,p_2}$ ,  $\hat{C} \in \mathbb{R}^{m_2,N}$ ,  $\hat{D} \in \mathbb{R}^{m_2,p_2}$ , and transfer function  $K(s) = \hat{C}(s\hat{E} - \hat{A})^{-1}\hat{B} + \hat{D}$  such that the closed-loop system resulting from the combination of (1.1) and (1.2), that is,

$$\begin{aligned} E\dot{x}(t) &= (A + B_2\hat{D}Z_1C_2)x(t) + (B_2Z_2\hat{C})\hat{x}(t) + (B_1 + B_2\hat{D}Z_1D_{21})w(t), \\ (1.3) \quad \hat{E}\dot{\hat{x}}(t) &= \hat{B}Z_1C_2x(t) + (\hat{A} + \hat{B}Z_1D_{22}\hat{C})\hat{x}(t) + \hat{B}Z_1D_{21}w(t), \\ z(t) &= (C_1 + D_{12}Z_2\hat{D}C_2)x(t) + D_{12}Z_2\hat{C}\hat{x}(t) + (D_{11} + D_{12}\hat{D}Z_1D_{21})w(t) \end{aligned}$$

with  $Z_1 = (I_{p_2} - D_{22}\hat{D})^{-1}$  and  $Z_2 = (I_{m_2} - \hat{D}D_{22})^{-1}$ , has the following properties:

- (1) System (1.3) is internally stable; that is, the solution  $\begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix}$  of the system with  $w \equiv 0$  is asymptotically stable, i.e.,  $\lim_{t \rightarrow \infty} \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} = 0$ .
- (2) The closed-loop transfer function  $T_{zw}(s)$  from  $w$  to  $z$  satisfies  $T_{zw} \in \mathcal{H}_\infty^{p_1, m_1}$  and is minimized in the  $\mathcal{H}_\infty$ -norm.

In principle, there is no restriction on the dimension  $N$  of the auxiliary state  $\hat{x}$  in (1.2), although smaller dimensions  $N$  are preferred for practical implementation and computation.

As in the case of the optimal  $\mathcal{H}_\infty$  control problems for ordinary state space systems, it is also necessary to study two closely related optimization problems, the modified optimal  $\mathcal{H}_\infty$  control problem and the suboptimal  $\mathcal{H}_\infty$  control problem.

**DEFINITION 1.2** (the modified optimal  $\mathcal{H}_\infty$  control problem). *For the descriptor system (1.1) let  $\Gamma$  be the set of positive real numbers  $\gamma$  for which there exists an internally stabilizing dynamic controller of the form (1.2) so that the transfer function  $T_{zw}(s)$  of the closed-loop system (1.3) satisfies  $T_{zw} \in \mathcal{H}_\infty^{p_1, m_1}$  with  $\|T_{zw}\|_\infty < \gamma$ . Determine  $\gamma_{mo} = \inf \Gamma$ . If no internally stabilizing dynamic controller exists, we set  $\Gamma = \emptyset$  and  $\gamma_{mo} = \infty$ .*

Note that it is possible that there is no internally stabilizing dynamic controller with the property  $\|T_{zw}\|_\infty = \gamma_{mo}$ . In this case one solves the suboptimal  $\mathcal{H}_\infty$  control problem.

**DEFINITION 1.3** (the suboptimal  $\mathcal{H}_\infty$  control problem). *For the descriptor system (1.1) and  $\gamma \in \Gamma$  with  $\gamma > \gamma_{mo}$ , determine an internally stabilizing dynamic controller of the form (1.2) such that the closed-loop transfer function satisfies  $T_{zw} \in \mathcal{H}_\infty^{p_1, m_1}$  with  $\|T_{zw}\|_\infty < \gamma$ . We call such a controller a  $\gamma$ -suboptimal controller or simply a suboptimal controller.*

The outline of the paper is as follows. In the forthcoming section we present the notation and some definitions that are used throughout the paper. Section 3 contains the main result of the paper and states conditions for the existence of an appropriate controller in terms of deflating subspaces of matrix pencils. The proof is given in three parts. First, we briefly discuss the standard state space case. The results are then generalized to descriptor systems of index 1 and, thereafter, to systems with arbitrary index. In section 4 we illustrate the presented theory by means of a numerical example.

**2. Preliminaries.** In this section we introduce some notation and definitions. For symmetric matrices  $A$  and  $B$ , by  $A \geq B$  and  $A > B$  we denote that  $A - B$  is positive semidefinite and positive definite, respectively. The spectral radius of a matrix  $A \in \mathbb{R}^{n,n}$  is denoted by  $\rho(A)$ . The set of complex numbers with positive real part is denoted by  $\mathbb{C}^+$  and the set of positive real numbers by  $\mathbb{R}^+$ .

Let  $\lambda E - A$  be a matrix pencil with  $E, A \in \mathbb{R}^{n,n}$ . Then  $\lambda E - A$  is called *regular* if  $\det(\lambda E - A) \neq 0$  for some  $\lambda \in \mathbb{C}$ .

A pencil  $P(\lambda) = \lambda E - A$  is called *even* if  $P(-\lambda)^T = P(\lambda)$ , i.e., if  $E = -E^T$  and  $A = A^T$ .

For regular pencils, *generalized eigenvalues* are the pairs  $(\alpha, \beta) \in \mathbb{C}^2 \setminus \{(0, 0)\}$  for which  $\det(\alpha E - \beta A) = 0$ . If  $\beta \neq 0$ , then the pair represents the finite eigenvalue  $\lambda = \alpha/\beta$ . If  $\beta = 0$ , then  $(\alpha, \beta)$  represent the eigenvalue infinity. In the following we use the notation with  $\lambda$ .

The solution and many properties of the *free descriptor system* (with  $u, w = 0$ ) can be characterized in terms of the *Weierstrass canonical form* (WCF).

**THEOREM 2.1** (see [13]). *For a regular matrix pencil  $\lambda E - A$ , there exist matrices  $W_f, V_f \in \mathbb{R}^{n_f, n_f}$ ,  $W_\infty, V_\infty \in \mathbb{R}^{n_\infty, n_\infty}$  with the property that  $W = \begin{bmatrix} W_f & W_\infty \end{bmatrix}$ ,  $V = \begin{bmatrix} V_f & V_\infty \end{bmatrix}$  are square and invertible, with*

$$(2.1a) \quad W^T E V = \begin{bmatrix} W_f^T \\ W_\infty^T \end{bmatrix} E \begin{bmatrix} V_f & V_\infty \end{bmatrix} = \begin{bmatrix} I_{n_f} & 0 \\ 0 & N \end{bmatrix},$$

$$(2.1b) \quad W^T A V = \begin{bmatrix} W_f^T \\ W_\infty^T \end{bmatrix} A \begin{bmatrix} V_f & V_\infty \end{bmatrix} = \begin{bmatrix} A_f & 0 \\ 0 & I_{n_\infty} \end{bmatrix},$$

$A_f \in \mathbb{R}^{n_f, n_f}$  is in real Jordan canonical form, and  $N \in \mathbb{R}^{n_\infty, n_\infty}$  is a nilpotent matrix, also in Jordan canonical form. We call  $n_f, n_\infty$  the number of finite or infinite eigenvalues, respectively.

The index of nilpotency of the nilpotent matrix  $N$  in (2.1a) is called the *index* of the system, and if  $E$  is nonsingular, then the pencil is said to have *index zero*.

**DEFINITION 2.2.** *A subspace  $\mathcal{L} \subset \mathbb{R}^n$  is called a deflating subspace for the pencil  $\lambda E - A$  if for a matrix  $X_{\mathcal{L}} \in \mathbb{R}^{n, k}$  with full column rank and  $\text{im } X_{\mathcal{L}} = \mathcal{L}$  there exist matrices  $Y_{\mathcal{L}} \in \mathbb{R}^{n, k}$ ,  $R_{\mathcal{L}} \in \mathbb{R}^{k, k}$ , and  $U_{\mathcal{L}} \in \mathbb{R}^{k, k}$  such that*

$$(2.2) \quad EX_{\mathcal{L}} = Y_{\mathcal{L}}R_{\mathcal{L}}, \quad AX_{\mathcal{L}} = Y_{\mathcal{L}}U_{\mathcal{L}}.$$

*A deflating subspace  $\mathcal{L}$  of  $\lambda E - A$  is called stable (semistable) if all finite eigenvalues of  $\lambda R_{\mathcal{L}} - U_{\mathcal{L}}$  are in the open (closed) left half-plane.*

*Let  $\mathcal{J} = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}$ , where  $I_n$  is the  $n \times n$  identity matrix. A subspace  $\mathcal{L} \subset \mathbb{R}^{2n}$  is called isotropic if  $x^T \mathcal{J} y = 0$  for all  $x, y \in \mathcal{L}$ . An isotropic subspace with  $\dim \mathcal{L} = n$  is called Lagrangian.*

In the notation of (2.1a)–(2.1b) with

$$(2.3) \quad B_{i,f} = W_f^T B_i, \quad B_{i,\infty} = W_\infty^T B_i, \quad C_{i,f} = C_i V_f, \quad C_{i,\infty} = C_i V_\infty, \quad i = 1, 2,$$

classical solutions of (1.2) take the form  $x(t) = V_f x_f(t) + V_\infty x_\infty(t)$ , where  $x_f$  and  $x_\infty$  satisfy

$$(2.4a) \quad \dot{x}_f(t) = A_f x_f(t) + B_{1,f} w(t) + B_{2,f} u(t),$$

$$(2.4b) \quad N \dot{x}_\infty(t) = x_\infty(t) + B_{1,\infty} w(t) + B_{2,\infty} u(t).$$

If the pencil  $\lambda E - A$  has index  $\nu$ , then this system has the explicit solution

$$(2.5a) \quad x_f(t) = e^{A_f(t-t_0)} x_f(t_0) + \int_{t_0}^t e^{A_f(t-\tau)} (B_{1,f} w(\tau) + B_{2,f} u(\tau)) d\tau,$$

$$(2.5b) \quad x_\infty(t) = - \sum_{i=0}^{\nu-1} \frac{d^i}{dt^i} N^i (B_{1,\infty} w(t) + B_{2,\infty} u(t)).$$

In contrast to standard state space systems, this shows that the initial condition  $x_\infty(t_0)$  is restricted by (2.5b). Moreover, if  $\nu > 1$ , then the solution will depend on derivatives of the input  $u$  and the disturbance  $w$ .

Note further that for the closed-loop system (1.3) to be *internally stable*, the controller has to be designed so that both  $x_f$  and  $x_\infty$  are asymptotically stable. While for the finite part this can be guaranteed if the spectrum of the matrix  $A_f$  lies in the open left half-plane, for the infinite part this has to be explicitly achieved by the construction of the controller.

As in the case of standard state space systems, certain conditions will be needed to guarantee the existence of optimal  $\mathcal{H}_\infty$  controls. First, these are stabilizability and detectability conditions, which for descriptor systems are the following; see [6, 9].

DEFINITION 2.3. Let  $E, A \in \mathbb{R}^{n,n}$ ,  $B \in \mathbb{R}^{n,m}$ , and  $C \in \mathbb{R}^{p,n}$ . Further, let  $T_\infty, S_\infty$  be matrices with  $\text{im } T_\infty = \ker E^T$  and  $\text{im } S_\infty = \ker E$ .

- (i) The triple  $(E, A, B)$  is called *finite dynamics stabilizable* if  $\text{rank}[\lambda E - A, B] = n$  for all  $\lambda \in \mathbb{C}^+$ ;
- (ii)  $(E, A, B)$  is *impulse controllable* if  $\text{rank}[E, AS_\infty, B] = n$ ;
- (iii)  $(E, A, B)$  is *strongly stabilizable* if it is both *finite dynamics stabilizable* and *impulse controllable*;
- (iv) the triple  $(E, A, C)$  is *finite dynamics detectable* if  $\text{rank}[\lambda E^T - A^T, C^T] = n$  for all  $\lambda \in \mathbb{C}^+$ ;
- (v)  $(E, A, C)$  is *impulse observable* if  $\text{rank}[E^T, A^T T_\infty, C^T] = n$ ;
- (vi)  $(\lambda E - A, C)$  is *strongly detectable* if it is both *finite dynamics detectable* and *impulse observable*.

After introducing our notation and giving some preliminary results, in the next section we derive the theoretical basis for the optimal  $\mathcal{H}_\infty$  control problem for descriptor systems.

**3. The modified optimal  $\mathcal{H}_\infty$  control problem.** In this section we approach the problem of determining  $\gamma_{mo}$  for a given system (1.1). As in the case of standard state space systems (see [16, 19, 28, 40]), we need several assumptions on the system matrices. In the following we set  $r = \text{rank } E$ .

*Assumptions.*

- (A1) The triple  $(E, A, B_2)$  is strongly stabilizable and the triple  $(E, A, C_2)$  is strongly detectable; see Definition 2.3.
- (A2)  $\text{rank} \begin{bmatrix} A - i\omega E & B_2 \\ C_1 & D_{12} \end{bmatrix} = n + m_2$  for all  $\omega \in \mathbb{R}$ .
- (A3)  $\text{rank} \begin{bmatrix} A - i\omega E & B_1 \\ C_2 & D_{21} \end{bmatrix} = n + p_2$  for all  $\omega \in \mathbb{R}$ .
- (A4) For matrices  $T_\infty, S_\infty \in \mathbb{R}^{n,n-r}$  with  $\text{im } S_\infty = \ker E$  and  $\text{im } T_\infty = \ker E^T$ , the following hold:

$$\begin{aligned} \text{rank} \begin{bmatrix} T_\infty^T AS_\infty & T_\infty^T B_2 \\ C_1 S_\infty & D_{12} \end{bmatrix} &= n + m_2 - r, \\ \text{rank} \begin{bmatrix} T_\infty^T AS_\infty & T_\infty^T B_1 \\ C_2 S_\infty & D_{21} \end{bmatrix} &= n + p_1 - r. \end{aligned}$$

It is well known for standard state space systems that assumption (A1) is essential for the existence of a controller that internally stabilizes the system. We will see that a similar result holds for the descriptor case. Assumptions (A2) and (A3) correspond to the typical claim that the system does not have transmission zeros on the imaginary axis. This is assumed in many works about  $\mathcal{H}_\infty$  control of standard state space

systems, since eigenvalues on the imaginary axis of the Hamiltonian matrices that are used in the computation of an optimal controller usually lead to problems in the computation of a semistable subspace; see [26, 34].

Further typical assumptions in the  $\mathcal{H}_\infty$  control of standard state space systems are that  $D_{12}, D_{21}^T$  have full column rank; see [19, 28, 40]. The conditions in (A4) reduce to these rank conditions if  $E$  is invertible.

For the construction of optimal or suboptimal controllers, we will make use of the following two even matrix pencils, which generalize the pencils constructed in [3, 4]. Let

$$(3.1) \quad \lambda N_H + M_H(\gamma) = \left[ \begin{array}{cc|ccc} 0 & -\lambda E^T - A^T & 0 & 0 & -C_1^T \\ \lambda E - A & 0 & -B_1 & -B_2 & 0 \\ \hline 0 & -B_1^T & -\gamma^2 I_{m_1} & 0 & -D_{11}^T \\ 0 & -B_2^T & 0 & 0 & -D_{12}^T \\ -C_1 & 0 & -D_{11} & -D_{12} & -I_{p_1} \end{array} \right]$$

and

$$(3.2) \quad \lambda N_J + M_J(\gamma) = \left[ \begin{array}{cc|ccc} 0 & -\lambda E - A & 0 & 0 & -B_1 \\ \lambda E^T - A^T & 0 & -C_1^T & -C_2^T & 0 \\ \hline 0 & -C_1 & -\gamma^2 I_{p_1} & 0 & -D_{11} \\ 0 & -C_2 & 0 & 0 & -D_{21} \\ -B_1^T & 0 & -D_{11}^T & -D_{21}^T & -I_{m_1} \end{array} \right].$$

Our approach is based on considering deflating subspaces of the matrix pencils (3.1) and (3.2), where the subspaces are spanned by the columns of the matrices  $X_H$  and  $X_J$  that are partitioned conformably with the pencils, i.e.,

$$(3.3) \quad X_H(\gamma) = \begin{bmatrix} X_{H,1}(\gamma) \\ X_{H,2}(\gamma) \\ X_{H,3}(\gamma) \\ X_{H,4}(\gamma) \\ X_{H,5}(\gamma) \end{bmatrix}, \quad X_J(\gamma) = \begin{bmatrix} X_{J,1}(\gamma) \\ X_{J,2}(\gamma) \\ X_{J,3}(\gamma) \\ X_{J,4}(\gamma) \\ X_{J,5}(\gamma) \end{bmatrix},$$

with

$$\begin{aligned} X_{H,1}(\gamma), X_{H,2}(\gamma), X_{J,1}(\gamma), X_{J,2}(\gamma) &\in \mathbb{R}^{n,r}, & X_{H,4}(\gamma) &\in \mathbb{R}^{m_2,r}, \\ X_{J,4}(\gamma) &\in \mathbb{R}^{p_2,r}, & X_{H,3}(\gamma), X_{J,5}(\gamma) &\in \mathbb{R}^{m_1,r}, & X_{H,5}(\gamma), X_{J,3}(\gamma) &\in \mathbb{R}^{p_1,r}. \end{aligned}$$

We extend the results in [3, 4] to general descriptor systems and use deflating subspaces of the even pencils (3.1) and (3.2) to characterize the elements of the set  $\Gamma$  in Definition 1.1. For this we introduce the following conditions which will be necessary for the existence of a controller with the desired properties associated with a parameter  $\gamma \in \Gamma$ .

- (C1) The index of both pencils (3.1) and (3.2) is at most one.
- (C2) There exists a matrix  $X_H(\gamma)$  as in (3.3) such that
  - (C2.a) the space  $\text{im } X_H(\gamma)$  is a semistable deflating subspace of  $\lambda N_H + M_H(\gamma)$  and  $\text{im} \begin{bmatrix} E X_{H,1} \\ X_{H,2} \end{bmatrix}$  is an  $r$ -dimensional isotropic subspace of  $\mathbb{R}^{2n}$ ;
  - (C2.b)  $\text{rank } E X_{H,1}(\gamma) = r$ .

(C3) There exists a matrix  $X_J(\gamma)$  as in (3.3) such that

(C3.a) the space  $\text{im } X_J(\gamma)$  is a semistable deflating subspace of  $\lambda N_J + M_J(\gamma)$  and  $\text{im} \begin{bmatrix} E^T X_{J,1} \\ X_{J,2} \end{bmatrix}$  is an  $r$ -dimensional isotropic subspace of  $\mathbb{R}^{2n}$ ;

(C3.b)  $\text{rank } E^T X_{J,1}(\gamma) = r$ .

Based on these conditions on the pencils, we introduce the following sets.

DEFINITION 3.1. Consider system (1.1) and the associated even pencils  $\lambda N_H + M_H(\gamma)$  in (3.1) and  $\lambda N_J + M_J(\gamma)$  in (3.2). Define the sets

$$\Gamma_H = \{\gamma \in \mathbb{R}^+ \mid \text{the index of } \lambda N_H + M_H(\gamma) \text{ is greater than one}\},$$

$$\Gamma_J = \{\gamma \in \mathbb{R}^+ \mid \text{the index of } \lambda N_J + M_J(\gamma) \text{ is greater than one}\},$$

and set  $\hat{\gamma}_H = \sup \Gamma_H$ ,  $\hat{\gamma}_J = \sup \Gamma_J$ , and  $\hat{\gamma} = \max\{\hat{\gamma}_H, \hat{\gamma}_J\}$ .

Note that, in general, the sets  $\Gamma_H$  and  $\Gamma_J$  may be all of  $\mathbb{R}^+$ , but as we will show later it follows from assumptions (A1)–(A4) that  $\hat{\gamma}_H$  and  $\hat{\gamma}_J$ , and therefore also  $\hat{\gamma}$  are finite. If  $\gamma > \hat{\gamma}$  then, since both  $\lambda N_H + M_H(\gamma)$  and  $\lambda N_J + M_J(\gamma)$  have index at most one, it follows that these pencils have  $2r$  finite eigenvalues, where  $r = \text{rank } E$ . Due to the fact that the pencils are even, and thus the eigenvalues occur in pairs  $\lambda, -\lambda$  (see [25]), it follows that there exist at least  $r$  eigenvalues in the closed left half-complex plane and at most  $r$  eigenvalues in the open left half-plane.

The next group of sets is related to conditions (C2.a) and (C2.b).

DEFINITION 3.2. Consider (1.1) and the associated even pencils  $\lambda N_H + M_H(\gamma)$  in (3.1) and  $\lambda N_J + M_J(\gamma)$  in (3.2). Define the sets

$$\Gamma_H^L = \{\gamma \geq \hat{\gamma} \mid \text{the pencil } \lambda N_H + M_H(\gamma) \text{ satisfies (C2.a)}\},$$

$$\Gamma_J^L = \{\gamma \geq \hat{\gamma} \mid \text{the pencil } \lambda N_J + M_J(\gamma) \text{ satisfies (C3.a)}\}, \quad \Gamma^L = \Gamma_J^L \cap \Gamma_H^L,$$

$$\Gamma_H^R = \{\gamma \geq \hat{\gamma} \mid \text{the pencil } \lambda N_H + M_H(\gamma) \text{ satisfies (C2)}\},$$

$$\Gamma_J^R = \{\gamma \geq \hat{\gamma} \mid \text{the pencil } \lambda N_J + M_J(\gamma) \text{ satisfies (C3)}\}, \quad \Gamma^R = \Gamma_J^R \cap \Gamma_H^R,$$

and set

$$\begin{aligned} \hat{\gamma}_H^L &= \inf \Gamma_H^L, & \hat{\gamma}_J^L &= \inf \Gamma_J^L, & \hat{\gamma}^L &= \inf \Gamma^L, \\ \hat{\gamma}_H^R &= \inf \Gamma_H^R, & \hat{\gamma}_J^R &= \inf \Gamma_J^R, & \hat{\gamma}^R &= \inf \Gamma^R. \end{aligned}$$

For the numerical method to compute the optimal or suboptimal  $\mathcal{H}_\infty$ , it will turn out to be essential to figure out these extremal values. Finally, we discuss the situation of finite purely imaginary eigenvalues.

DEFINITION 3.3. Consider (1.1) and the associated even pencils  $\lambda N_H + M_H(\gamma)$  in (3.1) and  $\lambda N_J + M_J(\gamma)$  in (3.2). Define the sets

$$\Gamma_H^I = \left\{ \gamma \geq \hat{\gamma} \mid \begin{array}{l} \text{the pencil } \lambda N_H + M_H(\gamma) \text{ has at least one} \\ \text{finite eigenvalue on the imaginary axis} \end{array} \right\},$$

$$\Gamma_J^I = \left\{ \gamma \geq \hat{\gamma} \mid \begin{array}{l} \text{the pencil } \lambda N_J + M_J(\gamma) \text{ has at least one} \\ \text{finite eigenvalue on the imaginary axis} \end{array} \right\}, \quad \Gamma^I = \Gamma_J^I \cap \Gamma_H^I,$$

and set

$$\hat{\gamma}_H^I = \inf \Gamma_H^I, \quad \hat{\gamma}_J^I = \inf \Gamma_J^I, \quad \hat{\gamma}^I = \inf \Gamma^I.$$

In the case where  $\Gamma_H^I$ ,  $\Gamma_J^I$ , or  $\Gamma^I$  is empty, we set  $\hat{\gamma}_H^I = \infty$ ,  $\hat{\gamma}_J^I = \infty$ , or  $\hat{\gamma}^I = \infty$ , respectively.

As in the classical  $\mathcal{H}_\infty$  control problem for state space systems (see [4]), we also need some further rank conditions, which are characterized in the following theorem that is proved in full generality in subsection 3.3.

**THEOREM 3.4.** *Consider a system of the form (1.1) satisfying assumptions (A1)–(A4). Let  $X_H(\gamma)$  and  $X_J(\gamma)$  be deflating subspace matrices of the form (3.3) that satisfy conditions (C2) and (C3), respectively. Then there exist parameters  $\hat{\gamma}_H^k \geq \hat{\gamma}_H^L$ ,  $\hat{\gamma}_J^k \geq \hat{\gamma}_J^L$ , and  $\hat{k}_H, \hat{k}_J \in \mathbb{N}$  with the property that for all  $\gamma_{H,1}, \gamma_{H,2} > \hat{\gamma}_H^k$ ,  $\gamma_{J,1}, \gamma_{J,2} > \hat{\gamma}_J^k$ , the rank conditions hold:*

$$(3.4) \quad \begin{aligned} \text{rank } E^T X_{H,2}(\gamma_{H,1}) &= \text{rank } E^T X_{H,2}(\gamma_{H,2}) = \hat{k}_H, \\ \text{rank } EX_{J,2}(\gamma_{J,1}) &= \text{rank } EX_{J,2}(\gamma_{J,2}) = \hat{k}_J. \end{aligned}$$

This rank property will be the basis for the formulation of a further condition on the pencils in (3.1) and (3.2) and on the blocks of the deflating subspace matrices  $X_H(\gamma) \in \mathbb{R}^{2n+m_1+m_2+p_1, r}$ ,  $X_J(\gamma) \in \mathbb{R}^{2n+p_1+p_2+m_1, r}$  satisfying (C2) (resp., (C3)).

(C4) The matrix

$$(3.5) \quad \mathcal{Y}(\gamma) = \begin{bmatrix} -\gamma X_{H,2}^T(\gamma) E X_{H,1}(\gamma) & X_{H,2}^T(\gamma) E X_{J,2}(\gamma) \\ X_{J,2}^T(\gamma) E^T X_{H,2}(\gamma) & -\gamma X_{J,2}^T(\gamma) E^T X_{J,1}(\gamma) \end{bmatrix}$$

is symmetric and positive semidefinite and satisfies  $\text{rank } \mathcal{Y}(\gamma) = \hat{k}_H + \hat{k}_J$ . Since  $X_H(\gamma)$  and  $X_J(\gamma)$  are unique up to a multiplication from the right with invertible matrices,  $\mathcal{Y}(\gamma)$  is unique up to a block-diagonal congruence transformation. Therefore, the value  $\text{rank } \mathcal{Y}(\gamma)$  is well defined.

Note that if we consider  $\mathcal{Y}(\gamma)$  in the standard case  $E = I_n$ , then it slightly differs from the matrix  $\mathcal{Y}(\gamma)$  used in [4]. This is due to the fact that the pencils (3.1) and (3.2) are expressed in a slightly different form in the generalization to descriptor systems.

Condition (C4) then leads to another set that has to be considered.

**DEFINITION 3.5.** *Consider a system of the form (1.1) that satisfies assumptions (A1)–(A4). Then we define*

$$\Gamma^\rho = \left\{ \gamma \geq \hat{\gamma} \mid \begin{array}{l} \text{the matrix } \mathcal{Y}(\gamma) \text{ is positive semidefinite} \\ \text{with } \text{rank } \mathcal{Y}(\gamma) = \hat{k}_H + \hat{k}_J \end{array} \right\}$$

and we set  $\hat{\gamma}^\rho := \inf \Gamma^\rho$ .

In this section we have introduced several assumptions and conditions as well as sets of  $\gamma$ -parameters that will be used in the next section to derive conditions for the optimal and suboptimal  $\gamma$ -parameters.

We proceed in three steps, first recalling the standard state space case in subsection 3.1, then considering the index one case in subsection 3.2, and finally considering the general case in subsection 3.3.

**3.1. The standard state space case.** In the first step, we briefly review the results from [10, 4] for the standard state space, that is,  $E = I_n$ . The relation between the values introduced in Definitions 3.1–3.3 is given by the following proposition.

**PROPOSITION 3.6** (see [4]). *Consider a system of the form (1.1) with  $E = I_n$ . Then the following inequality holds:*

$$(3.6) \quad 0 \leq \hat{\gamma} \leq \hat{\gamma}^L \leq \hat{\gamma}^R.$$

If  $\hat{\gamma}^I < \infty$ , then  $\hat{\gamma}^I = \hat{\gamma}^L > \hat{\gamma}$ . If  $\hat{\gamma}^\rho$  exists, then  $\hat{\gamma}^\rho \geq \hat{\gamma}^R$ .

Furthermore, it was shown in [4] that Theorem 3.4 holds if  $E = I_n$ . Therefore, (C4) represents a well-defined condition, and we can present the main result for the modified optimal  $\mathcal{H}_\infty$  control problem of standard systems.

**PROPOSITION 3.7** (see [4]). *Consider system (1.1) with  $E = I_n$  and the even pencils  $\lambda N_H + M_H(\gamma)$  and  $\lambda N_J + M_J(\gamma)$  as in (3.1) and (3.2), respectively. Suppose that assumptions (A1)–(A4) hold.*

*Then there exists an internally stabilizing controller such that the transfer function from  $w$  to  $z$  satisfies  $T_{zw} \in \mathcal{H}_\infty^{p_1, m_1}$  with  $\|T_{zw}\|_\infty < \gamma$  if and only if  $\gamma$  is such that conditions (C1)–(C4) hold.*

*Furthermore, the set of  $\gamma$  satisfying conditions (C1)–(C4) is nonempty.*

**3.2. The index one case.** To extend Proposition 3.7 to the case where the index of  $\lambda E - A$  is  $\nu = 1$ , we will make use of the WCF from Theorem 2.1. Transforming system (1.1) and using the notation introduced in (2.3), the explicit solution (2.5b) reduces to  $x_\infty(t) = -B_{1,\infty}w(t) - B_{2,\infty}u(t)$ . Inserting this into the transformed out equations, we obtain the standard state space system (often called the slow or finite dynamics subsystem)

$$(3.7) \quad \begin{aligned} \dot{x}_f(t) &= A_f x_f(t) + B_{1,f}w(t) + B_{2,f}u(t), \\ z(t) &= C_{1,f}x_f(t) + (D_{11} - C_{1,\infty}B_{1,\infty})w(t) + (D_{12} - C_{1,\infty}B_{2,\infty})u(t), \\ y(t) &= C_{2,f}x_f(t) + (D_{21} - C_{2,\infty}B_{1,\infty})w(t) + (D_{22} - C_{2,\infty}B_{2,\infty})u(t). \end{aligned}$$

**LEMMA 3.8.** *Consider system (1.1) and suppose that the index of  $\lambda E - A$  is at most one. Then for  $i \in \{1, 2, 3, 4\}$ , system (1.1) satisfies (Ai) if and only if the slow subsystem (3.7) satisfies (Ai).*

*Proof.* Any system of index at most one is both impulse controllable and observable (see [9, 26]) and, furthermore, finite dynamics stabilizability (detectability) is equivalent to stabilizability (detectability) of the slow subsystem obtained from the WCF. Then from Theorem 2.1 (see also [9, 26]), the equivalence of the corresponding conditions (A1) is immediate.

The equivalence for the corresponding conditions (A2) is obtained by using the transformation matrices to WCF, since

$$\begin{aligned} & \begin{bmatrix} W_f^T & 0 \\ -C_{1,\infty}W_\infty^T & I_{p_1} \\ W_\infty^T & 0 \end{bmatrix} \begin{bmatrix} A - i\omega E & B_2 \\ C_1 & D_{12} \end{bmatrix} \begin{bmatrix} V_f & -V_\infty B_{2,\infty} & V_\infty \\ 0 & I_{m_2} & 0 \end{bmatrix} \\ &= \begin{bmatrix} A_f - i\omega I_{n_f} & B_{2,f} & 0 \\ C_{1,f} & D_{12} - C_{2,\infty}B_{2,\infty} & 0 \\ 0 & 0 & I_{n_\infty} \end{bmatrix}. \end{aligned}$$

The proof for the equivalence of the corresponding conditions (A3) is analogous.

We now consider condition (A4). By definition, the columns of the matrices  $T_\infty, S_\infty$  span the left and right nullspace of  $E$ . Thus there exist invertible matrices  $M_l, M_r \in \mathbb{R}^{n_\infty, n_\infty}$  such that  $W_\infty = T_\infty M_l$ ,  $V_\infty = S_\infty M_r$ . The assertion then

follows from

$$\begin{aligned} & \begin{bmatrix} M_l^T & 0 \\ -C_{1,\infty}M_l^T & I_{p_1} \end{bmatrix} \begin{bmatrix} T_\infty^T A S_\infty & T_\infty^T B_2 \\ C_1 S_\infty & D_{12} \end{bmatrix} \begin{bmatrix} M_r & -M_r B_{2,\infty} \\ 0 & I_{m_2} \end{bmatrix} \\ &= \begin{bmatrix} I_{n_\infty} & 0 \\ 0 & D_{12} - C_{1,\infty} B_{2,\infty} \end{bmatrix}, \\ & \begin{bmatrix} M_l^T & 0 \\ -C_{2,\infty}M_l^T & I_{p_2} \end{bmatrix} \begin{bmatrix} T_\infty^T A S_\infty & T_\infty^T B_1 \\ C_2 S_\infty & D_{21} \end{bmatrix} \begin{bmatrix} M_r & -M_r B_{1,\infty} \\ 0 & I_{m_1} \end{bmatrix} \\ &= \begin{bmatrix} I_{n_\infty} & 0 \\ 0 & D_{21} - C_{2,\infty} B_{1,\infty} \end{bmatrix}. \quad \square \end{aligned}$$

After proving the equivalence of the conditions (Ai), we now show that the  $\Gamma$  sets and  $\gamma$  parameters introduced in Definitions 3.1–3.3 and 3.5 are those of the slow subsystem. We denote by  $\lambda N_{H,st} + M_{H,st}(\gamma)$  and  $\lambda N_{J,st} + M_{J,st}(\gamma)$  the even pencils (3.1) and (3.2) constructed from the data of system (3.7).

LEMMA 3.9. *Consider the system (1.1) and assume that the index of  $\lambda E - A$  is at most one. Let  $\lambda N_H + M_H(\gamma)$  and  $\lambda N_J + M_J(\gamma)$  be the even pencils constructed from the data of (1.1), and let  $\lambda N_{H,st} + M_{H,st}(\gamma)$ ,  $\lambda N_{J,st} + M_{J,st}(\gamma)$  be the corresponding pencils constructed from the data of (3.7).*

*Let  $\Gamma_H, \Gamma_J, \Gamma_H^L, \Gamma_J^L, \Gamma_H^R, \Gamma_J^R, \Gamma_H^I, \Gamma_J^I$  be the sets introduced in Definitions 3.1–3.3 and 3.5, and let  $\mathcal{Y}(\gamma)$  be the matrix introduced in (3.5).*

*Analogously, let  $\Gamma_{H,st}, \Gamma_{J,st}, \Gamma_{H,st}^L, \Gamma_{J,st}^L, \Gamma_{H,st}^R, \Gamma_{J,st}^R, \Gamma_{H,st}^I, \Gamma_{J,st}^I$ , and  $\mathcal{Y}_{st}(\gamma)$  be correspondingly defined for the slow subsystem (3.7). Then,*

$$\begin{aligned} \Gamma_{J,st} &= \Gamma_H, & \Gamma_{H,st}^L &= \Gamma_H^L, & \Gamma_{H,st}^R &= \Gamma_H^R, & \Gamma_{H,st}^I &= \Gamma_H^I, \\ \Gamma_{J,st} &= \Gamma_J, & \Gamma_{J,st}^L &= \Gamma_J^L, & \Gamma_{J,st}^R &= \Gamma_J^R, & \Gamma_{J,st}^I &= \Gamma_J^I, & \text{rank } \mathcal{Y}(\gamma) &= \text{rank } \mathcal{Y}_{st}(\gamma). \end{aligned}$$

*Proof.* First, we consider the pencil  $\lambda N_H + M_H(\gamma)$  and introduce the transformation matrix

$$(3.8) \quad P_H = \begin{bmatrix} V_f^T & 0 & 0 & 0 & 0 \\ 0 & W_f^T & 0 & 0 & 0 \\ B_{1,\infty}^T V_\infty^T & 0 & I_{m_1} & 0 & 0 \\ B_{2,\infty}^T V_\infty^T & 0 & 0 & I_{m_2} & 0 \\ 0 & -C_{1,\infty} W_\infty^T & 0 & 0 & I_{p_2} \\ V_\infty^T & 0 & 0 & 0 & 0 \\ 0 & W_\infty^T & 0 & 0 & 0 \end{bmatrix}^T.$$

We obtain that

$$(3.9) \quad \lambda P_H^T N_H P_H + P_H^T M_H(\gamma) P_H = \begin{bmatrix} \lambda N_{H,st} + M_{H,st}(\gamma) & 0 & 0 \\ 0 & I_{n_\infty} & 0 \\ 0 & 0 & I_{n_\infty} \end{bmatrix}.$$

This directly implies  $\Gamma_{H,st} = \Gamma_H$  and  $\Gamma_{H,st}^I = \Gamma_H^I$ . Analogously, we can show that  $\Gamma_{J,st} = \Gamma_J$  and  $\Gamma_{J,st}^I = \Gamma_J^I$ . Furthermore, it can be concluded from (3.9) that the columns of a matrix

$$X_{H,st} = [X_{H,st,1}^T \quad X_{H,st,2}^T \quad X_{H,st,3}^T \quad X_{H,st,4}^T \quad X_{H,st,5}^T]^T$$



partitioned conformably to the block structure of  $\lambda N_{H,st} + M_{H,st}(\gamma)$  span a semistable deflating subspace if and only if the columns of

$$(3.10) \quad \begin{bmatrix} X_{H,1} \\ X_{H,2} \\ X_{H,3} \\ X_{H,4} \\ X_{H,5} \end{bmatrix} = \begin{bmatrix} V_f X_{H,st,1} + V_\infty B_{1,\infty} X_{H,st,3} + V_\infty B_{2,\infty} X_{H,st,4} \\ W_f X_{H,st,2} - W_\infty C_{1,\infty}^T X_{H,st,5} \\ X_{H,st,3} \\ X_{H,st,4} \\ X_{H,st,5} \end{bmatrix}$$

span the semistable deflating subspace of  $\lambda N_H + M_H(\gamma)$ .

An analogous relation can be derived for the relation between spanning matrices of deflating subspaces of  $\lambda N_J + M_J(\gamma)$  and  $\lambda N_{J,st} + M_{J,st}(\gamma)$ . Using the fact that  $EV_\infty = 0$ ,  $W_\infty^T E = 0$  and  $W_f^T EV_f = I_{n_f}$ , it follows that

$$(3.11a) \quad \text{rank } EX_{H,1}(\gamma) = \text{rank } X_{H,st,1}(\gamma),$$

$$(3.11b) \quad \text{rank } E^T X_{J,1}(\gamma) = \text{rank } X_{J,st,1}(\gamma),$$

$$(3.11c) \quad \text{rank } E^T X_{H,2}(\gamma) = \text{rank } X_{H,st,2}(\gamma),$$

$$(3.11d) \quad \text{rank } EX_{J,2}(\gamma) = \text{rank } X_{J,st,2}(\gamma),$$

$$(3.11e) \quad X_{H,2}(\gamma)^T EX_{H,1}(\gamma) = X_{H,st,2}(\gamma)^T X_{H,st,1}(\gamma),$$

$$(3.11f) \quad X_{H,2}(\gamma)^T EX_{J,2}(\gamma) = X_{H,st,2}(\gamma)^T X_{J,st,2}(\gamma),$$

$$(3.11g) \quad X_{J,2}(\gamma)^T E^T X_{J,1}(\gamma) = X_{J,st,2}(\gamma)^T X_{J,st,1}(\gamma).$$

Equations (3.11e) and (3.11g) and the relations between the stable deflating subspaces of  $\lambda N_H + M_H(\gamma)$ ,  $\lambda N_{H,st} + M_{H,st}(\gamma)$  and  $\lambda N_J + M_J(\gamma)$ ,  $\lambda N_{J,st} + M_{J,st}(\gamma)$ , respectively, imply that  $\Gamma_{H,st}^L = \Gamma_H^L$  and  $\Gamma_{J,st}^L = \Gamma_J^L$ . Additionally, from (3.11a) and (3.11b), we obtain  $\Gamma_{H,st}^R = \Gamma_H^R$  and  $\Gamma_{J,st}^R = \Gamma_J^R$ .

By using (3.11e)–(3.11g) we then obtain that the matrices  $\mathcal{Y}(\gamma)$  and  $\mathcal{Y}_{st}(\gamma)$  coincide; in particular, we have  $\text{rank } \mathcal{Y}(\gamma) = \text{rank } \mathcal{Y}_{st}(\gamma)$ .  $\square$

An immediate consequence is that Proposition 3.6 holds for systems of index at most one. Furthermore, from (3.11c) and (3.11b) and the corresponding fact for standard systems, we can conclude that Theorem 3.4 holds for systems of index at most one.

With these preparations we can formulate the following extension of Proposition 3.7 for systems of index at most one.

**PROPOSITION 3.10.** *Consider system (1.1) such that the index of the pencil  $\lambda E - A$  is at most one, and the even pencils  $\lambda N_H + M_H(\gamma)$  and  $\lambda N_J + M_J(\gamma)$  as in (3.1) and (3.2), respectively. Suppose that assumptions (A1)–(A4) hold.*

*Then there exists an internally stabilizing controller such that the transfer function from  $w$  to  $z$  satisfies  $T_{zw} \in \mathcal{H}_\infty^{p_1, m_1}$  with  $\|T_{zw}\|_\infty < \gamma$  if and only if  $\gamma$  is such that conditions (C1)–(C4) hold.*

*Furthermore, the set of  $\gamma$  satisfying conditions (C1)–(C4) is nonempty.*

*Proof.* The closed-loop transfer function  $T_{zw}(s)$  of the system (3.7) with a controller of the form (1.2) is equal to the closed-loop transfer function of the system (1.1) with the same controller.

Since (1.1) is strongly stabilizable (strongly detectable), if and only if system (3.7) is stabilizable (detectable), a controller that internally stabilizes (3.7) also stabilizes the finite dynamics of (1.1).

Therefore, the existence of a controller with desired properties for (1.1) is equivalent to the existence of such a controller for (3.7). Since by Lemma 3.8 the validity

of assumptions (A1)–(A4) for (3.7) is equivalent to those of (1.1) and, furthermore, also by Lemma 3.9 the corresponding conditions (C1)–(C4) of these two systems are equivalent, the assertion follows.  $\square$

We have seen so far that the standard state space case and the index one case follow after some simple transformation. In the next subsection we now study the general case.

**3.3. The general case.** In this section we formulate the results for the modified optimal  $\mathcal{H}_\infty$  control problem for descriptor systems of arbitrary index. A key tool in the proof will be an a priori static output feedback  $u(t) = Ky(t) + \bar{u}(t)$  resulting in a system

$$(3.12) \quad \begin{aligned} E\dot{x}(t) &= (A + B_2KC_2)x(t) + (B_1 + B_2KD_{21})w(t) + B_2\bar{u}(t), & x(t_0) &= x^0, \\ z(t) &= (C_1 + D_{12}KC_2)x(t) + (D_{11} + D_{12}KD_{21})w(t) + D_{12}\bar{u}(t), \\ y(t) &= C_2x(t) + D_{21}w(t). \end{aligned}$$

The feedback matrix  $K$  will be constructed in a way so that system (3.12) has index one. Then we are able to apply the results of the previous section. If (1.2) is a controller for (3.12), then a controller for system (1.1) is given by

$$(3.13) \quad \begin{aligned} \hat{E}\dot{\hat{x}}(t) &= \hat{A}\hat{x}(t) + \hat{B}y(t), \\ u(t) &= \hat{C}\hat{x}(t) + (\hat{D} - K)y(t). \end{aligned}$$

To proceed, we need the following results about the existence of a static output feedback  $K$  that leads to a system of index at most one.

LEMMA 3.11 (see [7, 9]). *Consider matrices  $C \in \mathbb{R}^{p,n}$  and  $B \in \mathbb{R}^{n,m}$  and a regular matrix pencil  $\lambda E - A$ . Then there exists  $K \in \mathbb{R}^{p,m}$  such that the pencil  $\lambda E - (A + BKC)$  is regular and has index at most one if and only if the triple  $(E, A, B)$  is impulse controllable and the triple  $(E, A, C)$  is impulse observable; see Definition 2.3.*

To make use of this result, we show that a static output feedback does not change assumptions (A1)–(A4).

LEMMA 3.12. *Consider system (1.1) and let  $K \in \mathbb{R}^{m_2,p_2}$  such that the pencil  $\lambda E - (A + B_2KC_2)$  is regular. Then for every  $i \in \{1, 2, 3, 4\}$ , system (1.1) satisfies (Ai) if and only if the system (3.12) satisfies (Ai).*

*Proof.* The invariance of strong stabilizability and strong detectability under output feedback is trivial. The proof for the equivalence of the corresponding assumptions (A2) follows from the identity

$$\begin{bmatrix} A - i\omega E & B_2 \\ C_1 & D_{12} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ KC_2 & I_{m_2} \end{bmatrix} = \begin{bmatrix} A + B_2KC_2 - i\omega E & B_2 \\ C_1 + D_{12}KC_2 & D_{12} \end{bmatrix},$$

while the equivalence statement for (A3) can be shown analogously. The fact that (3.12) satisfies (A4) if and only if (1.1) satisfies (A4) is a consequence of

$$\begin{aligned} \begin{bmatrix} T_\infty^T AS_\infty & T_\infty^T B_2 \\ C_1 S_\infty & D_{12} \end{bmatrix} \begin{bmatrix} I_{n-r} & 0 \\ KC_2 S_\infty & I_{m_2} \end{bmatrix} &= \begin{bmatrix} T_\infty^T (A + B_2KC_2) S_\infty & T_\infty^T B_2 \\ (C_1 + D_{12}KC_2) S_\infty & D_{12} \end{bmatrix}, \\ \begin{bmatrix} I_{n-r} & T_\infty^T B_2 K \\ 0 & I_{p_2} \end{bmatrix} \begin{bmatrix} T_\infty^T AS_\infty & T_\infty^T B_1 \\ C_2 S_\infty & D_{21} \end{bmatrix} &= \begin{bmatrix} T_\infty^T (A + B_2KC_2) S_\infty & T_\infty^T (B_1 + B_2KD_{21}) \\ C_2 S_\infty & D_{21} \end{bmatrix}. \quad \square \end{aligned}$$

In the following lemma we show that the sets introduced in Definitions 3.1–3.3 and 3.5 are invariant under output feedback as well.

LEMMA 3.13. Consider the system (1.1) and let  $K \in \mathbb{R}^{m_2, p_2}$  be such that the pencil  $\lambda E - (A + BKC)$  is regular. Let  $\Gamma_H, \Gamma_J, \Gamma_H^L, \Gamma_J^L, \Gamma_H^R, \Gamma_J^R, \Gamma_H^I, \Gamma_J^I$  be the sets introduced in Definitions 3.1–3.3 and 3.5, and let  $\mathcal{Y}(\gamma)$  be the matrix introduced in (3.5). Furthermore, let  $\Gamma_{H,K}, \Gamma_{J,K}, \Gamma_{H,K}^L, \Gamma_{J,K}^L, \Gamma_{H,K}^R, \Gamma_{J,K}^R, \Gamma_{H,K}^I, \Gamma_{J,K}^I$ , and  $\mathcal{Y}_K(\gamma)$  be the corresponding quantities for the system (3.12). Then,

$$\begin{aligned} \Gamma_{J,K} &= \Gamma_H, & \Gamma_{H,K}^L &= \Gamma_H^L, & \Gamma_{H,K}^R &= \Gamma_H^R, & \Gamma_{H,K}^I &= \Gamma_H^I, \\ \Gamma_{J,K} &= \Gamma_J, & \Gamma_{J,K}^L &= \Gamma_J^L, & \Gamma_{J,K}^R &= \Gamma_J^R, & \Gamma_{J,K}^I &= \Gamma_J^I, & \text{rank } \mathcal{Y}(\gamma) &= \text{rank } \mathcal{Y}_K(\gamma). \end{aligned}$$

*Proof.* Let  $\lambda N_{H,K} + M_{H,K}(\gamma)$  be the even pencil associated with the system (3.12). Then, with the transformation matrices

$$T_{H,K} = \begin{bmatrix} I_n & 0 & 0 & 0 & 0 \\ 0 & I_n & 0 & 0 & 0 \\ 0 & 0 & I_{m_1} & 0 & 0 \\ KC_2 & 0 & KD_{21} & I_{m_2} & 0 \\ 0 & 0 & 0 & 0 & I_{p_1} \end{bmatrix}, \quad T_{J,K} = \begin{bmatrix} I_n & 0 & 0 & 0 & 0 \\ 0 & I_n & 0 & 0 & 0 \\ 0 & 0 & I_{p_1} & 0 & 0 \\ K^T B_2^T & 0 & K^T D_{12}^T & I_{p_2} & 0 \\ 0 & 0 & 0 & 0 & I_{m_1} \end{bmatrix},$$

we have the identities

$$\begin{aligned} \lambda T_{H,K}^T N_H T_{H,K} + T_{H,K}^T M_H(\gamma) T_{H,K} &= \lambda N_{H,K} + M_{H,K}(\gamma), \\ \lambda T_{J,K}^T N_J T_{J,K} + T_{J,K}^T M_J(\gamma) T_{J,K} &= \lambda N_{J,K} + M_{J,K}(\gamma). \end{aligned}$$

Thus, we have that the pencils  $\lambda N_H + M_H(\gamma)$  and  $\lambda N_{H,K} + M_{H,K}(\gamma)$  have the same index and eigenvalues. Similarly, this holds for  $\lambda N_J + M_J(\gamma)$  and  $\lambda N_{J,K} + M_{J,K}(\gamma)$ . Therefore we have  $\Gamma_{H,K} = \Gamma_H, \Gamma_{J,K} = \Gamma_J, \Gamma_{H,K}^L = \Gamma_H^L, \Gamma_{J,K}^L = \Gamma_J^L$ . The relations  $\Gamma_{H,K}^R, \Gamma_{J,K}^R, \Gamma_{H,K}^I, \Gamma_{J,K}^I$  follow from the facts that

$$\begin{aligned} \text{im} \begin{bmatrix} X_{H,1}^T & X_{H,2}^T & X_{H,3}^T & X_{H,4}^T & X_{H,5}^T \end{bmatrix}^T, \\ \text{im} \begin{bmatrix} X_{J,1}^T & X_{J,2}^T & X_{J,3}^T & X_{J,4}^T & X_{J,5}^T \end{bmatrix}^T \end{aligned}$$

are semistable deflating subspaces of  $\lambda N_{H,K} + M_{H,K}(\gamma)$  and  $\lambda N_{J,K} + M_{J,K}(\gamma)$ , respectively, if and only if

$$(3.14) \quad \begin{aligned} \text{im} \begin{bmatrix} X_{H,1}^T & X_{H,2}^T & X_{H,3}^T & (X_{H,4} + KC_2 X_{H,1} + KD_{21} X_{H,3})^T & X_{H,5}^T \end{bmatrix}^T, \\ \text{im} \begin{bmatrix} X_{J,1}^T & X_{J,2}^T & X_{J,3}^T & (X_{J,4} + K^T B_2^T X_{J,1} + K^T D_{12}^T X_{J,3})^T & X_{J,5}^T \end{bmatrix}^T \end{aligned}$$

are semistable deflating subspace of  $\lambda N_H + M_H(\gamma)$  and  $\lambda N_J + M_J(\gamma)$ . From (3.14), we further obtain that  $\mathcal{Y}(\gamma) = \mathcal{Y}_K(\gamma)$ , and thus their ranks coincide.  $\square$

With these auxiliary results, we are now in a position to prove Theorem 3.4.

*Proof of Theorem 3.4.* First we apply an a priori feedback  $K \in \mathbb{R}^{m_2, p_2}$  to (1.1) such that the resulting system (3.12) has index at most one. Then we know from (3.14) that the corresponding matrices  $X_{H,1}, X_{H,2}, X_{J,1}, X_{J,2}$  of (1.1) and (3.12) are equal. Since Theorem 3.4 holds for systems of index one, the assertion follows.  $\square$

Lemma 3.13 also implies that the assertion of Proposition 3.6 still holds for the general case; i.e., for (1.1), the inequality  $0 \leq \hat{\gamma} \leq \hat{\gamma}^L \leq \hat{\gamma}^R$  is valid. In the case where  $\hat{\gamma}^I < \infty$ , we have  $\hat{\gamma}^I = \hat{\gamma}^L > \hat{\gamma}$ , and if  $\hat{\gamma}^\rho$  exists, then  $\hat{\gamma}^\rho \geq \hat{\gamma}^R$ .

With the described framework, we can now formulate the main result for the modified  $\mathcal{H}_\infty$  control problem for descriptor systems.

**THEOREM 3.14.** *Consider system (1.1) and the even pencils  $\lambda N_H + M_H(\gamma)$  and  $\lambda N_J + M_J(\gamma)$  as in (3.1) and (3.2), respectively. Suppose that assumptions (A1)–(A4) hold.*

*Then there exists an internally stabilizing controller such that the transfer function from  $w$  to  $z$  satisfies  $T_{zw} \in \mathcal{H}_\infty^{p_1, m_1}$  with  $\|T_{zw}\|_\infty < \gamma$  if and only if  $\gamma$  is such that the conditions (C1)–(C4) hold.*

*Furthermore, the set of  $\gamma$  satisfying conditions (C1)–(C4) is nonempty.*

*Proof.* Due to Lemma 3.11, there exists a matrix  $K \in \mathbb{R}^{m_2, p_2}$  such that system (3.12) has index at most one. Lemma 3.12 implies that (3.12) satisfies (A1)–(A4) as well. Furthermore, by Lemma 3.13, the validity of conditions (C1)–(C4) for system (1.1) is equivalent to the respective conditions for system (3.12).

Proposition 3.10 then implies that conditions (C1)–(C4) for (3.12) are fulfilled if and only if there exists a desired controller for (3.12).

Since an application of the controller (1.2) to (3.12) yields the same closed-loop system as that given by (3.13) with controller (1.1), the desired result follows immediately.  $\square$

**THEOREM 3.15.** *Consider system (1.1) and suppose that assumptions (A1)–(A4) hold. Then the set  $\Gamma^\rho$  is nonempty, and optimal  $\gamma$  for the modified optimal  $\mathcal{H}_\infty$  control problem is given by*

$$(3.15) \quad \gamma_{mo} = \hat{\gamma}^\rho.$$

*Proof.* Let  $\Gamma$  be the set of  $\gamma > 0$  for which an internally stabilizing controller exists such that the transfer function from  $w$  to  $z$  satisfies  $\|T_{zw}\|_\infty < \gamma$ .

We know from Theorem 3.14 that  $\Gamma$  is nonempty, and for some  $\gamma > 0$ , we have  $\gamma \in \Gamma$  if and only if conditions (C1)–(C4) are fulfilled. By the definition of  $\Gamma_H$ ,  $\Gamma_J$ ,  $\Gamma^R$ , and  $\Gamma^\rho$ , the existence of a controller with desired properties is therefore equivalent to

$$(3.16) \quad \gamma \in \Gamma_H \cap \Gamma_J, \quad \gamma \in \Gamma^R, \quad \gamma \in \Gamma^\rho.$$

Especially, we have that  $\Gamma^\rho$  is nonempty. By the definition of  $\hat{\gamma}$ ,  $\hat{\gamma}^R$ , and  $\hat{\gamma}^\rho$ , condition (3.16) is the same as

$$(3.17) \quad \gamma > \hat{\gamma}, \quad \gamma > \hat{\gamma}^R, \quad \gamma \in \hat{\gamma}^\rho.$$

Hence,  $\gamma \in \Gamma$  is equivalent to

$$(3.18) \quad \gamma > \max\{\hat{\gamma}, \hat{\gamma}^R, \hat{\gamma}^\rho\}.$$

However, since by Lemma 3.13 we have that Proposition 3.6 still holds for arbitrary descriptor systems, the equation  $\hat{\gamma}^\rho = \max\{\hat{\gamma}, \hat{\gamma}^R, \hat{\gamma}^\rho\}$  holds. Thus we have that  $\hat{\gamma}_{mo} = \inf \Gamma = \hat{\gamma}^\rho$ .  $\square$

**4. Numerical example.** To illustrate the functionality of our approach, consider the following example from [38] which is also discussed in [31]. The descriptor system is given by (1.1) with  $D_{21} = 1$ ,  $D_{11} = D_{22} = 0$ , and

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix},$$

$$C_1 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad C_2 = [1 \ 0 \ 1], \quad D_{12} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

This system is of index two, and the associated pencils  $\lambda N_H + M_H(\gamma)$  and  $\lambda N_J + M_J(\gamma)$  have index one for  $\gamma \neq 0$ . The goal is to find the minimum value  $\gamma$  that satisfies conditions (C1)–(C4). Using our experimental code for the computation of stable deflating subspaces of structured matrix pencils as in [2] and a bisection to determine the optimal value for  $\gamma$ , we computed  $\gamma_{opt}$  given by  $\gamma^p = 0.7678$ , which is smaller than the suboptimal value obtained in [38, 31].

**5. Conclusion.** In this paper we have developed conditions for optimal and suboptimal  $\mathcal{H}_\infty$  control for descriptor systems of arbitrary index. We have expressed criteria for the existence of an internally stabilizing controller in terms of even pencils containing only parameters of the original system. With these criteria the  $\gamma$ -iteration can be performed in a numerically stable way, especially if a structure-preserving method is used for the computation of the deflating subspaces. The computation of a controller for a given  $\gamma$  can be performed by first following the steps in the proofs and then using controller formulas for standard systems [14, 15, 11]. However, this approach does not lead to controller formulas stated in terms of the original system variables. In [21] such formulas are given for systems with nonsingular system matrix  $E$ ; the general case is currently under investigation.

## REFERENCES

- [1] G. BALAS, R. CHIANG, A. PACKARD, AND M. SAFONOV, *Robust Control Toolbox. For Use with MATLAB. User's Guide, Version 3*, The MathWorks, Inc., Natick, MA, 2005.
- [2] P. BENNER, R. BYERS, V. MEHRMANN, AND H. XU, *Numerical computation of deflating subspaces of skew-Hamiltonian/Hamiltonian pencils*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 165–190.
- [3] P. BENNER, R. BYERS, V. MEHRMANN, AND H. XU, *A robust numerical method for optimal  $\mathcal{H}_\infty$  control*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Atlantis, Paradise Island, Bahamas, 2004, pp. 424–425.
- [4] P. BENNER, R. BYERS, V. MEHRMANN, AND H. XU, *A robust numerical method for the  $\gamma$ -iteration in  $\mathcal{H}_\infty$  control*, Linear Algebra Appl., 425 (2007), pp. 548–570.
- [5] P. BENNER, V. MEHRMANN, V. SIMA, S. VAN HUFFEL, AND A. VARGA, *SLICOT — a subroutine library in systems and control theory*, in Applied and Computational Control, Signals, and Circuits, Vol. 1, B. N. Datta, ed., Birkhäuser Boston, Boston, MA, 1999, pp. 499–546.
- [6] A. BUNSE-GERSTNER, R. BYERS, V. MEHRMANN, AND N. K. NICHOLS, *Feedback design for regularizing descriptor systems*, Linear Algebra Appl., 299 (1999), pp. 119–151.
- [7] A. BUNSE-GERSTNER, V. MEHRMANN, AND N. K. NICHOLS, *Regularization of descriptor systems by output feedback*, IEEE Trans. Automat. Control, 39 (1994), pp. 1742–1748.
- [8] B. M. CHEN, *Exact computation of optimal value in  $\mathcal{H}_\infty$  control*, in Unsolved Problems in Mathematical Systems and Control Theory, V. D. Blondel and A. Megretski, eds., Princeton University Press, Princeton, NJ, 2004, pp. 271–275. Available online from <http://pup.princeton.edu/titles/7790.html>.
- [9] L. DAI, *Singular Control Systems*, Lecture Notes in Control and Inform. Sci. 118, Springer-Verlag, Berlin, Heidelberg, 1989.
- [10] J. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State-space solutions to standard  $\mathcal{H}_2$  and  $\mathcal{H}_\infty$  control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.
- [11] J. DOYLE AND K. GLOVER, *State-space formulae for all stabilizing controllers that satisfy an  $\mathcal{H}_\infty$  norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.
- [12] E. EICH-SOELLNER AND C. FÜHRER, *Numerical Methods in Multibody Dynamics*, Teubner, Stuttgart, 1998.
- [13] F. R. GANTMACHER, *Theory of Matrices*, Vol. 2, Chelsea, New York, 1959.
- [14] K. GLOVER, D. J. N. LIMBEER, J. C. DOYLE, E. M. KASENALLY, AND M. G. SAFONOV, *A characterization of all solutions to the four block general distance problem*, SIAM J. Control Optim., 29 (1991), pp. 283–324.

- [15] K. C. GOH AND M. G. SAFONOV,  $\mathcal{H}_\infty$  control: Inverse free formulae for  $D_{11} \neq 0$  and eliminating pole-zero cancellation via interpolation, in Proceedings of the 32nd IEEE Conference on Decision and Control, San Antonio, TX, 1993, pp. 1152–1157.
- [16] M. GREEN AND D. J. N. LIMBEER, *Linear Robust Control*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
- [17] M. GÜNTHER AND U. FELDMANN, *CAD-based electric-circuit modeling in industry. II. Impact of circuit configurations and parameters*, Surveys Math. Indust., 8 (1999), pp. 131–157.
- [18] M. GÜNTHER AND U. FELDMANN, *CAD-based electric-circuit modeling in industry. I. Mathematical structure and index of network equations*, Surveys Math. Indust., 8 (1999), pp. 97–129.
- [19] D.-W. GU, P. HR. PETKOV, AND M. M. KONSTANTINOV, *Direct Formulae for the  $\mathcal{H}_\infty$  Sub-Optimal Central Controller*, NICONET Report 1998–7, The Working Group on Software (WGS), Leuven-Heverlee, Belgium, 1998. Available online at <http://www.slicot.org/index.php?site=reports>.
- [20] D.-W. GU, P. HR. PETKOV, AND M. M. KONSTANTINOV,  *$\mathcal{H}_\infty$  and  $\mathcal{H}_2$  Optimization Toolbox in SLICOT*, SLICOT Working Note 1999–12, The Working Group on Software (WGS), Leuven-Heverlee, Belgium, 1999. Available online at <http://www.slicot.org/index.php?site=reports>.
- [21] A. KARTHIKEYAN AND M. G. SAFONOV, *Simplified matrix pencil all-solutions  $H_\infty$  controller formulae*, SICE Journal of Control Measurement and System Integration, 1 (2008), pp. 137–142.
- [22] P. KUNKEL, V. MEHRMANN, AND W. RATH, *Analysis and numerical solution of control problems in descriptor form*, Math. Control Signals Systems, 14 (2001), pp. 29–61.
- [23] P. KUNKEL AND V. MEHRMANN, *Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems*, Math. Control Signals Systems, 14 (2001), pp. 233–256.
- [24] P. KUNKEL AND V. MEHRMANN, *Differential-Algebraic Equations. Analysis and Numerical Solution*, EMS Publishing House, Zürich, Switzerland, 2006.
- [25] D. S. MACKEY, N. MACKEY, C. MEHL, AND V. MEHRMANN, *Structured polynomial eigenvalue problems: Good vibrations from good linearizations*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1029–1051.
- [26] V. MEHRMANN, *The Autonomous Linear Quadratic Control Problem. Theory and Numerical Solution*, Lecture Notes in Control and Inform. Sci. 163, Springer-Verlag, Heidelberg, 1991.
- [27] M. OTTER, H. ELMQVIST, AND S. E. MATTSO, *Multi-domain modeling with Modelica*, in Handbook of Dynamic System Modeling, Paul Fishwick, ed., Chapman and Hall/CRC Press, Boca Raton, FL, 2006, pp. 36.1–36.27.
- [28] I. R. PETERSEN, V. A. UGRINOVSKII, AND A. V. SAVKIN, *Robust Control Design Using  $\mathcal{H}_\infty$  Methods*, Springer-Verlag, London, UK, 2000.
- [29] P. HR. PETKOV, D.-W. GU, AND M. M. KONSTANTINOV, *Fortran 77 Routines for  $\mathcal{H}_\infty$  and  $\mathcal{H}_2$  Design of Continuous-Time Linear Control Systems*, NICONET Report 1998–8, The Working Group on Software (WGS), Leuven-Heverlee, Belgium, 1998. Available online at <http://www.slicot.org/index.php?site=reports>.
- [30] P. C. RABIER AND W. C. RHEINOLDT, *Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint*, SIAM, Philadelphia, 2000.
- [31] A. REHM AND F. ALLGÖWER,  *$\mathcal{H}_\infty$  Control of Differential-Algebraic-Equation Systems*, Technical report, Institut für Systemdynamik, Universität Stuttgart, Stuttgart, 1998.
- [32] A. REHM AND F. ALLGÖWER, *An LMI approach towards  $\mathcal{H}_\infty$  control of descriptor systems*, in IFAC International Symposium on Advanced Control of Chemical Processes (ADCHEM 2000), Vol. 1, Pisa, 2000, pp. 57–62.
- [33] A. REHM AND F. ALLGÖWER,  *$\mathcal{H}_\infty$  control of differential algebraic equation systems: The linearizing change of variables approach revisited*, in Proceedings of the IEEE American Control Conference, Vol. 4, Arlington, VA, 2001, pp. 2948–2952.
- [34] W. SCHIEHLEN, *Multibody Systems Handbook*, Springer-Verlag, Heidelberg, 1990.
- [35] T. SCHMIDT AND M. HOU, *Rollringgetriebe*. Internal Report, Sicherheitstechnische Regelungen und Meßtechnik, Bergische Universität, GH Wuppertal, Wuppertal, Germany, 1992.
- [36] R. SCHÜPPHAUS AND P. C. MÜLLER, *Control analysis and synthesis of linear mechanical descriptor systems*, in Advanced Multibody System Dynamics, W. Schiehlen, ed., Kluwer Academic Publishers, Stuttgart, 1990, pp. 463–468.
- [37] A. STEINBRECHER, *Numerical Solution of Quasi-Linear Differential-Algebraic Equations and Industrial Simulation of Multibody Systems*, Ph.D. thesis, Institut für Mathematik, Technische Universität Berlin, Berlin, 2005.

- [38] K. TAKABA, N. MORIHIRA, AND T. KATAYAMA,  *$\mathcal{H}_\infty$ -control for descriptor systems — a  $J$ -spectral factorization approach*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Vol. 3, IEEE Press, Piscataway, NJ, 1994, pp. 2251–2256.
- [39] H. L. TRENTelman, A. A. STOORVOGEL, AND M. HAUTUS, *Control Theory for Linear Systems*, Springer-Verlag, London, UK, 2001.
- [40] K. ZHOU, J. C. DOYLE, AND K. GLOVER, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1995.

## THE EXISTENCE OF STRONGLY MDS CONVOLUTIONAL CODES\*

RYAN HUTCHINSON†

**Abstract.** It is known that maximum distance separable and maximum distance profile convolutional codes exist over large enough finite fields of any characteristic for all parameters  $(n, k, \delta)$ . It has been conjectured that the same is true for convolutional codes that are strongly maximum distance separable. Using methods from linear systems theory, we resolve this conjecture by showing that, over a large enough finite field of any characteristic, codes which are simultaneously maximum distance profile and strongly maximum distance separable exist for all parameters  $(n, k, \delta)$ .

**Key words.** MDS codes, convolutional codes, column distances, linear systems, minimal partial realization problem

**AMS subject classification.** 94B10

**DOI.** 10.1137/050638977

**1. Introduction.** In recent literature on convolutional codes, several new classes of codes with optimal distance properties have been introduced. These classes of codes are known as maximum distance separable (MDS) codes, maximum distance profile (MDP) codes, and strongly MDS (sMDS) codes. MDS codes are characterized by the property that they have the maximum possible free distance for a given choice of code parameters. sMDS codes are a subclass of MDS codes having the property that this maximum possible free distance is attained at the earliest possible encoding step. MDP codes are characterized by the property that their column distances grow at the maximum possible rate for a given choice of code parameters.

In [18], it is shown that MDS convolutional codes exist for all parameters  $(n, k, \delta)$  over sufficiently large finite fields; in [11], a similar result is obtained for codes having the MDP property. In [8], sMDS convolutional codes are introduced and studied, and they are shown to exist for parameters  $(n, k, \delta)$  satisfying  $(n - k) \mid \delta$ . In addition, it is conjectured that convolutional codes possessing the MDP and sMDS properties together exist for all  $(n, k, \delta)$ . In this work, we show that this conjecture is correct. The approach used is systems-theoretic in nature; to obtain the proof, we make use of the well-known interpretation of a convolutional code as an input-state-output linear system as well as results from partial realization theory.

The structure of this paper is as follows. In section 2, we review relevant ideas from the theory of convolutional codes. We recall as well a connection between convolutional codes and input-state-output linear systems that we will use to obtain our results. In section 3, we use a linear systems representation of convolutional codes to give a characterization of the sMDS property. In section 4, we use this characterization to show the existence, for all parameters  $(n, k, \delta)$ , of codes possessing both the MDP and sMDS properties.

**2. Convolutional codes and linear systems.** In this section, we recall some facts about convolutional codes and their connection with linear systems. Throughout

---

\*Received by the editors August 25, 2005; accepted for publication (in revised form) June 17, 2008; published electronically November 19, 2008. This work was supported in part by NSF grants DMS-00-72383 and CCR-02-05310.

<http://www.siam.org/journals/sicon/47-6/63897.html>

†Department of Mathematics and Computer Science, Bemidji State University, Bemidji, MN 56601 (rhutchinson@bemidjistate.edu).



this paper,  $0$  will be understood to be the zero matrix or vector of the appropriate size. Let  $k$  and  $n$  be positive integers with  $k < n$ ,  $p$  a prime number,  $\mathbb{K}$  the algebraic closure of the prime field  $\mathbb{F}_p$ , and  $\mathbb{F}$  a finite subfield of  $\mathbb{K}$ .

DEFINITION 2.1. A convolutional code  $\mathcal{C}$  of rate  $k/n$  is a rank- $k$  direct summand of  $\mathbb{F}[s]^n$ .

$\mathcal{C}$  is a free  $\mathbb{F}[s]$ -module and may thus be viewed as the column space of a full-rank matrix  $G(s) \in \mathbb{F}[s]^{n \times k}$ , called a *generator matrix* for  $\mathcal{C}$ . Two full-rank  $n \times k$  matrices  $G_1(s)$  and  $G_2(s)$  generate the same code if and only if there exists a unimodular matrix  $U(s) \in \mathbb{F}[s]^{k \times k}$  such that  $G_1(s) = G_2(s)U(s)$ .

When convenient, we will (at times with a slight abuse of notation) make use of the fact that  $\mathbb{F}[s]^n$  and  $\mathbb{F}^n[s]$  are isomorphic  $\mathbb{F}[s]$ -modules and think of codewords as elements of  $\mathbb{F}^n[s]$ . For example, the columns of a generator matrix  $G(s)$  may be thought of as polynomials with coefficients in  $\mathbb{F}^n$ ; we refer to the degrees of these polynomials as the *column degrees* of  $G(s)$  and denote the degree of the  $j$ th column by  $\delta_j$ . The *high-order coefficient matrix* of  $G(s)$ ,  $G_\infty$ , is the matrix whose  $j$ th column is the column coefficient of  $s^{\delta_j}$  in the  $j$ th column of  $G(s)$ . In general,  $G_\infty$  need not have full rank. It is always possible, though, to find a unimodular matrix  $U(s) \in \mathbb{F}[s]^{k \times k}$  such that  $G(s)U(s)$  has a full-rank high-order coefficient matrix (see [7]). If  $G_\infty$  has full rank, then  $G(s)$  is called a *minimal generator matrix*.

An important invariant of a convolutional code is its degree, defined as follows.

DEFINITION 2.2. The degree  $\delta$  of a convolutional code  $\mathcal{C}$  is the maximal degree of a (polynomial) determinant of a  $k \times k$  submatrix of a generator matrix of  $\mathcal{C}$ .

This definition makes sense, as multiplication by a unimodular matrix preserves the degrees of such determinants. We note that, if  $G(s)$  is a minimal generator matrix of  $\mathcal{C}$  with column degrees  $\delta_1, \dots, \delta_k$ , then  $\delta = \sum_{j=1}^k \delta_j$ . A code of rate  $k/n$  and degree  $\delta$  will be referred to as an  $(n, k, \delta)$ -code.

We turn next to notions of distance. We first recall the definition of Hamming weight.

DEFINITION 2.3. Let  $v \in \mathbb{F}^n$  and  $v(s) := \sum_{t=0}^d v_t s^t \in \mathbb{F}^n[s]$ . The Hamming weight of  $v$ ,  $\text{wt}(v)$ , is the number of nonzero components of  $v$ . The Hamming weight of  $v(s)$  is  $\text{wt}(v(s)) := \sum_{t=0}^d \text{wt}(v_t)$ .

For the purpose of error control coding, it is important that the minimum weight among the codewords of a code be as large as possible. This leads to the concept of free distance.

DEFINITION 2.4. The free distance of a convolutional code  $\mathcal{C}$  is

$$d_{\text{free}}(\mathcal{C}) := \min\{\text{wt}(v(s)) \mid v(s) \in \mathcal{C} \setminus \{0\}\}.$$

Column distances also play an important part in what follows. They measure the minimum possible distance between truncated codewords.

DEFINITION 2.5. Let  $\mathcal{C}$  be a convolutional code. For  $j \in \mathbb{N}_0$ , the  $j$ th column distance of  $\mathcal{C}$  is

$$d_j^{\mathcal{C}} := \min \left\{ \sum_{t=0}^j \text{wt}(v_t) \mid v(s) \in \mathcal{C} \text{ and } v_0 \neq 0_n \right\},$$

where  $v_j = 0_n$  if  $j > \deg v(s)$ .

The following result gives upper bounds for the column distances and the free distance of a convolutional code.

PROPOSITION 2.6. *Let  $\mathcal{C}$  be an  $(n, k, \delta)$ -code.*

1. *For every  $j \in \mathbb{N}_0$ ,*

$$d_j^c(\mathcal{C}) \leq (n - k)(j + 1) + 1.$$

*If  $d_j^c(\mathcal{C}) = (n - k)(j + 1) + 1$  for some  $j$ , then  $d_i^c(\mathcal{C}) = (n - k)(i + 1) + 1$  if  $i \in \{0, \dots, j\}$ .*

- 2.

$$d_{free}(\mathcal{C}) \leq (n - k) \left( \left\lfloor \frac{\delta}{k} \right\rfloor + 1 \right) + \delta + 1.$$

Statement 1 is proved in [8], and statement 2 is proved in [18]. The bound in 2 is called the *generalized Singleton bound*.

Set  $L := \lfloor \frac{\delta}{k} \rfloor + \lfloor \frac{\delta}{n-k} \rfloor$  and  $M := \lfloor \frac{\delta}{k} \rfloor + \lceil \frac{\delta}{n-k} \rceil$ . We are now ready to define the code properties of interest in this work.

DEFINITION 2.7. *Let  $\mathcal{C}$  be an  $(n, k, \delta)$ -code. Then,*

1.  *$\mathcal{C}$  is called a maximum distance profile (MDP) code if*

$$d_L^c(\mathcal{C}) = (n - k)(L + 1) + 1.$$

2.  *$\mathcal{C}$  is called a maximum distance separable (MDS) code if*

$$d_{free}(\mathcal{C}) = (n - k) \left( \left\lfloor \frac{\delta}{k} \right\rfloor + 1 \right) + \delta + 1.$$

3.  *$\mathcal{C}$  is called a strongly MDS (sMDS) code if*

$$d_M^c(\mathcal{C}) = (n - k) \left( \left\lfloor \frac{\delta}{k} \right\rfloor + 1 \right) + \delta + 1.$$

Using the fact that no column distance of  $\mathcal{C}$  can exceed the generalized Singleton bound, one can show that  $L$  is the largest possible value of  $j$  for which  $d_j^c(\mathcal{C})$  can attain the upper bound in statement 1 of Proposition 2.6. If  $\mathcal{C}$  is an MDP code, then, by Proposition 2.6,  $d_i^c(\mathcal{C})$  attains this upper bound when  $i \in \{0, \dots, L\}$ . Thus, statement 1 says that the column distances of an MDP code are maximal until it is no longer possible. Similarly, one can show that, if  $j < M$ , then  $d_j^c(\mathcal{C}) < (n - k)(\lfloor \frac{\delta}{k} \rfloor + 1) + \delta + 1$ . Thus, statement 3 says that, for an sMDS code, the sequence  $\{d_j^c(\mathcal{C})\}_{j \geq 0}$  attains the generalized Singleton bound at the smallest possible value of  $j$ .

In the second part of this section, we introduce a connection between convolutional codes and linear systems that we will use to obtain our results. Background information for this discussion and applications of ideas from systems theory to the construction of convolutional codes may be found, for example, in [2, 3, 17, 19, 20].

Let  $A \in \mathbb{K}^{\delta \times \delta}$ ,  $B \in \mathbb{K}^{\delta \times k}$ ,  $C \in \mathbb{K}^{(n-k) \times \delta}$ , and  $D \in \mathbb{K}^{(n-k) \times k}$ . Note that, since the number of entries in the matrices  $(A, B, C, D)$  is finite, these matrices are actually defined over a finite subfield  $\mathbb{F}$  of  $\mathbb{K}$ . The matrices  $(A, B, C, D)$  describe a time-invariant linear system through the equations

$$\begin{aligned} (2.1) \quad & x_{t+1} = Ax_t + Bu_t, \\ & y_t = Cx_t + Du_t, \\ & x_0 = 0, \end{aligned}$$

where  $x_t \in \mathbb{F}^\delta$ ,  $u_t \in \mathbb{F}^k$ , and  $y_t \in \mathbb{F}^{n-k}$  are called the *state vector*, *input vector*, and *output vector* at time  $t$ , respectively. The matrix quadruple  $(A, B, C, D)$  is called a *realization* for the system. We recall the following well-known definition.

DEFINITION 2.8.  $(A, B)$  is called a reachable pair if

$$\text{rank} \left( \begin{bmatrix} B & AB & \cdots & A^{\delta-2}B & A^{\delta-1}B \end{bmatrix} \right) = \delta.$$

$(A, C)$  is called an observable pair if

$$\text{rank} \left( \begin{bmatrix} C^T & (CA)^T & \cdots & (CA^{\delta-2})^T & (CA^{\delta-1})^T \end{bmatrix}^T \right) = \delta.$$

If  $(A, B)$  is a reachable pair and  $(A, C)$  is an observable pair, then  $(A, B, C, D)$  is called a *minimal realization*. In this case,  $\delta$  is called the *McMillan degree* of the system. We denote by  $S_{k,n}^\delta$  the set of minimal realizations of systems over  $\mathbb{K}$  having input vectors of size  $k$ , output vectors of size  $n - k$ , and McMillan degree  $\delta$ .

Let  $\{x_t\}_{t \geq 0}$  be a sequence of vectors in  $\mathbb{F}^\delta$  and  $\{(y_t)\}_{t \geq 0}$ , where  $y_t \in \mathbb{F}^{n-k}$  and  $u_t \in \mathbb{F}^k$ , a sequence of vectors in  $\mathbb{F}^n$  having the following properties:

1. Equations (2.1) are satisfied for all  $t \in \mathbb{N}_0$ .
2. There exists a  $d \in \mathbb{N}_0$  such that  $x_{d+1} = 0$  and  $u_t = 0$  for  $t \geq d + 1$ .

These properties guarantee that the sequence  $\{(y_t)\}_{t \geq 0}$  has finite weight. We refer to the truncated sequence  $\{(y_t)\}_{t=0}^d$  as a *finite-weight sequence* for  $(A, B, C, D)$ . The following remarks connect finite-weight sequences and codewords.

Let  $(A, B, C, D) \in S_{k,n}^\delta$ . The corresponding transfer function is  $T(s) := C(sI - A)^{-1}B + D$ . Let  $Q^{-1}(s)P(s)$  be a left coprime factorization of  $T(s)$ , and set  $H(s) := [-Q(s) \ P(s)]$ . Set

$$y(s) := y_0s^d + y_1s^{d-1} + \cdots + y_d \in \mathbb{F}^{n-k}[s]$$

and

$$u(s) := u_0s^d + u_1s^{d-1} + \cdots + u_d \in \mathbb{F}^k[s],$$

and use their coefficients to form the vector sequence  $\{(y_t)\}_{t=0}^d$ . We then have the following equivalent conditions; see [3, 19, 20] for more details.

1. The set  $\{(y_t)\}_{t=0}^d$  of vectors is a finite-weight sequence for  $(A, B, C, D)$ .
- 2.

$$\left[ \begin{array}{c|cccccc} 0 & \cdots & 0 & A^d B & A^{d-1} B & \cdots & AB & B \\ \hline & & & D & 0 & \cdots & \cdots & 0 \\ & & & CB & D & \ddots & & \vdots \\ & & & CAB & CB & \ddots & \ddots & \vdots \\ & & & \vdots & \vdots & \ddots & \ddots & 0 \\ & & & CA^{d-1} B & CA^{d-2} B & \cdots & CB & D \end{array} \right] \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_d \\ \hline u_0 \\ u_1 \\ \vdots \\ u_d \end{bmatrix} = 0.$$

3. There exists a “state vector polynomial”

$$x(s) = x_0s^d + x_1s^{d-1} + \cdots + x_d \in \mathbb{F}^\delta[s]$$

such that

$$\begin{bmatrix} sI - A & 0 & -B \\ -C & I_{n-k} & -D \end{bmatrix} \begin{bmatrix} x(s) \\ y(s) \\ u(s) \end{bmatrix} = 0.$$

4.

$$H(s) \begin{bmatrix} y(s) \\ u(s) \end{bmatrix} = [-Q(s)P(s)] \begin{bmatrix} y(s) \\ u(s) \end{bmatrix} = 0.$$

5.

$$y(s) = T(s)u(s).$$

Further, the right  $\mathbb{F}[s]$ -kernel of  $H(s)$  is an  $(n, k, \delta)$ -code  $\mathcal{C}$ .

The code  $\mathcal{C}$  is not quite suitable for our purposes. This is due to the fact that the finite-weight sequence

$$\begin{pmatrix} y_0 \\ u_0 \end{pmatrix}, \begin{pmatrix} y_1 \\ u_1 \end{pmatrix}, \dots, \begin{pmatrix} y_{d-1} \\ u_{d-1} \end{pmatrix}, \begin{pmatrix} y_d \\ u_d \end{pmatrix}$$

corresponds with the codeword

$$\begin{pmatrix} y_d \\ u_d \end{pmatrix} + \begin{pmatrix} y_{d-1} \\ u_{d-1} \end{pmatrix}s + \dots + \begin{pmatrix} y_1 \\ u_1 \end{pmatrix}s^{d-1} + \begin{pmatrix} y_0 \\ u_0 \end{pmatrix}s^d \in \mathcal{C}.$$

Working in the systems setting, we will show that there is a realization  $(A, B, C, D) \in S_{k,n}^\delta$  for which any finite-weight sequence

$$\begin{pmatrix} y_0 \\ u_0 \end{pmatrix}, \begin{pmatrix} y_1 \\ u_1 \end{pmatrix}, \dots, \begin{pmatrix} y_{d-1} \\ u_{d-1} \end{pmatrix}, \begin{pmatrix} y_d \\ u_d \end{pmatrix}$$

(with  $u_0 \neq 0$ ) formed using (2.1) has the properties that

$$\sum_{t=0}^L \text{wt} \left( \begin{pmatrix} y_t \\ u_t \end{pmatrix} \right) = (L+1)(n-k) + 1$$

and

$$\sum_{t=0}^M \text{wt} \left( \begin{pmatrix} y_t \\ u_t \end{pmatrix} \right) \geq (n-k) \left( \left\lfloor \frac{\delta}{k} \right\rfloor + 1 \right) + \delta + 1.$$

Due to the order reversal noted above, it will not necessarily be true that  $d_L^c(\mathcal{C}) = (L+1)(n-k) + 1$ . The next result shows how to overcome this problem.

**PROPOSITION 2.9.** *Let  $\mathcal{C}$  be an  $(n, k, \delta)$ -code with minimal generator matrix  $G(s)$ . Let  $\overline{G(s)}$  be the matrix obtained by replacing each entry  $p_{ij}(s)$  of  $G(s)$  by  $\overline{p_{ij}(s)} := s^{\delta_j} p_{ij}(s^{-1})$ , where  $\delta_j$  is the  $j$ th column degree of  $G(s)$ . Then,  $\overline{G(s)}$  is a minimal generator matrix of an  $(n, k, \delta)$ -code  $\overline{\mathcal{C}}$ , and*

$$\begin{pmatrix} y_0 \\ u_0 \end{pmatrix} + \begin{pmatrix} y_1 \\ u_1 \end{pmatrix}s + \dots + \begin{pmatrix} y_{d-1} \\ u_{d-1} \end{pmatrix}s^{d-1} + \begin{pmatrix} y_d \\ u_d \end{pmatrix}s^d \in \overline{\mathcal{C}}$$

*if and only if*

$$\begin{pmatrix} y_d \\ u_d \end{pmatrix} + \begin{pmatrix} y_{d-1} \\ u_{d-1} \end{pmatrix}s + \dots + \begin{pmatrix} y_1 \\ u_1 \end{pmatrix}s^{d-1} + \begin{pmatrix} y_0 \\ u_0 \end{pmatrix}s^d \in \mathcal{C}.$$

*Proof.* First,  $\overline{G(s)}$  has rank  $k$ , since a  $k \times k$  minor of  $\overline{G(s)}$  is zero if and only if the corresponding minor of  $G(s)$  is.

Next, let  $G_0$  denote the  $n \times k$  matrix whose  $ij$ th entry is  $p_{ij}(0)$  and  $\overline{G}_0$  the  $n \times k$  matrix whose  $ij$ th entry is  $\overline{p_{ij}(0)}$ . Because  $\mathcal{C}$  is a summand of  $\mathbb{F}[s]^n$ ,  $G_0$  has full rank, so that each column of  $G_0$  has at least one nonzero entry. This means that  $G_0 = \overline{G}_\infty$ ,  $\overline{G}_\infty$  has full rank, corresponding columns of  $G(s)$  and  $\overline{G(s)}$  have the same column degrees, and  $\overline{\overline{G(s)}} = G(s)$ . From the definition of  $\overline{G(s)}$ , we have that  $G_\infty = \overline{G}_0$ ; since  $G(s)$  is minimal, it follows that  $\overline{G}_0$  also has full rank.

Suppose  $p(s) \in \mathbb{F}[s]$  has degree  $d$  and is a common divisor of the  $k \times k$  minors of  $\overline{G(s)}$ . Then,  $p(0) \neq 0$ , since  $\overline{G}_0$  has full rank, so that  $s^d p(s^{-1})$  has degree  $d$ . Since  $\overline{\overline{G(s)}} = G(s)$ ,  $s^d p(s^{-1})$  is a common divisor of the  $k \times k$  minors of  $G(s)$ . As  $\mathcal{C}$  is a summand of  $\mathbb{F}[s]^n$ , it follows that  $d = 0$ , so that the only common divisors of the  $k \times k$  minors of  $\overline{G(s)}$  are the nonzero elements of  $\mathbb{F}$ . Thus, the column space of  $\overline{G(s)}$  is a summand of  $\mathbb{F}[s]^n$ , which means that it is a rate  $k/n$  convolutional code  $\overline{\mathcal{C}}$ . It follows from the remarks in the preceding paragraph that  $\overline{G(s)}$  is a minimal generator matrix of  $\overline{\mathcal{C}}$  and that  $\overline{\mathcal{C}}$  has degree  $\delta$ .

Consider the vector polynomials

$$v(s) := v_d + v_{d-1}s + \cdots + v_1s^{d-1} + v_0s^d$$

and

$$\overline{v(s)} := v_0 + v_1s + \cdots + v_{d-1}s^{d-1} + v_ds^d$$

in  $\mathbb{F}^n[s]$ , and note that  $\overline{\overline{v(s)}} = s^d v(s^{-1})$ . Thinking of  $v(s)$  and  $\overline{v(s)}$  as column vectors in  $\mathbb{F}[s]^n$ , we observe that a  $(k+1) \times (k+1)$  minor of  $\begin{bmatrix} G(s) & | & v(s) \end{bmatrix}$  is zero if and only if the corresponding minor of  $\begin{bmatrix} \overline{G(s)} & | & \overline{v(s)} \end{bmatrix}$  is. Since  $\mathcal{C}$  and  $\overline{\mathcal{C}}$  are summands of  $\mathbb{F}[s]^n$ , this means that  $v(s) \in \mathcal{C}$  if and only if  $\overline{v(s)} \in \overline{\mathcal{C}}$ .  $\square$

This result, together with the remarks preceding it, shows that

$$\begin{pmatrix} y_0 \\ u_0 \end{pmatrix}, \begin{pmatrix} y_1 \\ u_1 \end{pmatrix}, \dots, \begin{pmatrix} y_{d-1} \\ u_{d-1} \end{pmatrix}, \begin{pmatrix} y_d \\ u_d \end{pmatrix}$$

is a finite-weight sequence for  $(A, B, C, D)$  if and only if

$$\begin{pmatrix} y_0 \\ u_0 \end{pmatrix} + \begin{pmatrix} y_1 \\ u_1 \end{pmatrix}s + \cdots + \begin{pmatrix} y_{d-1} \\ u_{d-1} \end{pmatrix}s^{d-1} + \begin{pmatrix} y_d \\ u_d \end{pmatrix}s^d \in \overline{\mathcal{C}}.$$

$\overline{\mathcal{C}}$ , then, will have the property that  $d_M^c(\overline{\mathcal{C}}) = (n-k)(\lfloor \frac{\delta}{k} \rfloor + 1) + \delta + 1$ . For the rest of the paper, we will refer to the code  $\overline{\mathcal{C}}$  as the code represented by the matrices  $(A, B, C, D)$ .

**3. Trivial rank deficiency and the sMDS property.** In this section, we give conditions on the entries of the matrices in a realization  $(A, B, C, D) \in S_{k,n}^\delta$  guaranteeing that the convolutional code these matrices represent has both the MDP and sMDS properties. For  $(A, B, C, D) \in S_{k,n}^\delta$  and  $j \in \mathbb{N}_0$ , we form the matrices

$$(3.1) \quad \mathcal{T}_j := \begin{bmatrix} D & 0 & \cdots & \cdots & 0 \\ CB & D & \ddots & & \vdots \\ CAB & CB & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ CA^{j-1}B & CA^{j-2}B & \cdots & CB & D \end{bmatrix}.$$

NOTATION 3.1. Let  $l_1, l_2 \in \mathbb{N}$  satisfy  $1 \leq l_1 \leq (j+1)(n-k)$  and  $1 \leq l_2 \leq (j+1)k$ . Let  $1 \leq i_1 < \dots < i_{l_1} \leq (j+1)(n-k)$  and  $1 \leq j_1 < \dots < j_{l_2} \leq (j+1)k$  be two sequences of integers. We denote by  $(\mathcal{T}_j)_{j_1, \dots, j_{l_2}}^{i_1, \dots, i_{l_1}}$  the  $l_1 \times l_2$  submatrix obtained from  $\mathcal{T}_j$  by intersecting rows  $i_1, \dots, i_{l_1}$  and columns  $j_1, \dots, j_{l_2}$ .

Notice that, if  $(\mathcal{T}_j)_{j_1}^{i_1} \neq 0$ , then  $j_1 \leq \lceil \frac{i_1}{n-k} \rceil k$ .

In what follows, the notion of trivial rank deficiency plays an important role. To define trivial rank deficiency, we think of replacing the entries of the block matrices in  $\mathcal{T}_j$  with the indeterminates of the polynomial ring  $R := \mathbb{K}[x_1, x_2, \dots, x_{(j+1)(n-k)k}]$ . Specifically, we replace the entry  $(s, t)$  of the matrix  $D$  with the indeterminate  $x_{(s-1)k+t}$  and the entry  $(s, t)$  of the matrix  $CA^iB$  with the indeterminate  $x_{(i+1)(n-k)k+(s-1)k+t}$ . The zero entries above the block diagonal remain zero.

DEFINITION 3.2. Let  $c$  be an integer with  $0 \leq c \leq n-k-1$ , and let  $l$  be an integer satisfying  $1 \leq l \leq \min\{(j+1)(n-k)-c, (j+1)k\}$ . A square submatrix of  $\mathcal{T}_j$  is said to be trivially rank deficient if the determinant of this submatrix is zero when it is viewed as a matrix over  $R$  in the manner described above. A submatrix  $(\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+c}}$  of  $\mathcal{T}_j$  is called trivially rank deficient if all  $\binom{l+c}{l}$   $l \times l$  submatrices of  $(\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+c}}$  are trivially rank deficient.

To say that  $(\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+c}}$  is trivially rank deficient is to say that  $(\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+c}}$  has less than full rank regardless of how elements of  $\mathbb{K}$  are substituted for the indeterminates of  $R$ . The next lemma shows how to determine if a given submatrix is trivially rank deficient.

LEMMA 3.3. Let  $l$  be an integer with  $1 \leq l \leq \min\{(j+1)(n-k)-c, (j+1)k\}$ , and let  $(\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+c}}$  be an  $(l+c) \times l$  submatrix of  $\mathcal{T}_j$ . Then, the following are equivalent:

1.  $(\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+c}}$  is trivially rank deficient.
2.  $(\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{i_{1+c}, i_{2+c}, \dots, i_{l+c}}$  is trivially rank deficient.
3. The inequality

$$j_t > \left\lceil \frac{i_{t+c}}{n-k} \right\rceil k$$

holds for some  $t \in \{1, \dots, l\}$ .

*Proof.* For notational convenience, we set  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}} := (\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{i_{1+c}, i_{2+c}, \dots, i_{l+c}}$  and  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}} := (\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+c}}$ .

$1 \implies 2$ : Suppose  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is trivially rank deficient. Then, by definition, all  $\binom{l+c}{l}$   $l \times l$  submatrices of  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  are trivially rank deficient. In particular,  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is trivially rank deficient.

$2 \implies 3$ : Suppose that  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is trivially rank deficient. We first use induction on  $l$  to prove that  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is lower block triangular and has a 0 on its diagonal. If  $l = 1$  and  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is trivially rank deficient, then the claim is trivially true. Suppose  $l$  satisfies  $2 \leq l \leq \min\{(j+1)(n-k)-c, (j+1)k\}$ , that the induction hypothesis is satisfied for  $1, 2, \dots, l-1$ , and that  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is trivially rank deficient. If  $(\mathcal{T}_j)_{j_1}^{i_{l+c}} = 0$ , then every entry in  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is 0, and thus all diagonal entries are 0. If  $(\mathcal{T}_j)_{j_1}^{i_{l+c}} \neq 0$ , then let  $x_\ell$  be the indeterminate corresponding with  $(\mathcal{T}_j)_{j_1}^{i_{l+c}}$  when  $\mathcal{T}_j$  is viewed over  $R$  in the manner described before Definition 3.2. Notice that, when  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is viewed in this way, the indeterminate  $x_\ell$  appears exactly once. Since  $x_\ell$  is transcendental over  $\mathbb{K}(x_1, \dots, x_{\ell-1})$ , doing a cofactor expansion along the first column of  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  (still

viewing  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  over  $R$ ) shows that the  $(l-1) \times (l-1)$  submatrix  $(\mathcal{T}_j)_{j_2, j_3, \dots, j_l}^{i_1+c, i_2+c, \dots, i_{l-1}+c}$  is trivially rank deficient. By the induction hypothesis,  $(\mathcal{T}_j)_{j_2, j_3, \dots, j_l}^{i_1+c, i_2+c, \dots, i_{l-1}+c}$  is lower block triangular and has a 0 on its diagonal. It follows that there is an integer  $h$  satisfying  $1 \leq h \leq l-1$  such that  $(\mathcal{T}_j)_{j_{h+1}}^{i_{h+c}} = 0$ . This, in turn, means that  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is lower block triangular. Because we assumed that  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is trivially rank deficient, it follows that at least one of  $(\mathcal{T}_j)_{j_1, j_2, \dots, j_h}^{i_1+c, i_2+c, \dots, i_h+c}$  and  $(\mathcal{T}_j)_{j_{h+1}, j_{h+2}, \dots, j_l}^{i_{h+1}+c, i_{h+2}+c, \dots, i_l+c}$  is trivially rank deficient. By the induction hypothesis, at least one of these submatrices is lower block triangular and has a 0 on its diagonal. As the diagonals of these submatrices lie on the diagonal of  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$ , the claim follows.

Next, we note that the diagonal entries of  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  are the entries  $(\mathcal{T}_j)_{j_1}^{i_1+c}, (\mathcal{T}_j)_{j_2}^{i_2+c}, \dots, (\mathcal{T}_j)_{j_l}^{i_l+c}$ . From the structure of  $\mathcal{T}_j$ , it is clear that, when  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is viewed over  $R$ , a diagonal entry  $(\mathcal{T}_j)_{j_t}^{i_t+c}$  is 0 if and only if

$$j_t > \left\lceil \frac{i_t+c}{n-k} \right\rceil k.$$

It follows that

$$j_t > \left\lceil \frac{i_t+c}{n-k} \right\rceil k$$

for some  $t \in \{1, \dots, l\}$ .

$3 \implies 1$ : If

$$j_t > \left\lceil \frac{i_t+c}{n-k} \right\rceil k$$

for some  $t \in \{1, \dots, l\}$ , then  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  has a 0 on its diagonal and is lower block triangular.  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is therefore trivially rank deficient. Let  $(\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{w_1, w_2, \dots, w_l}$  be a submatrix of  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$ . Since  $w_t \leq i_t+c$ ,

$$j_t > \left\lceil \frac{w_t}{n-k} \right\rceil k$$

holds as well. As before, it follows that  $(\mathcal{T}_j)_{j_1, j_2, \dots, j_l}^{w_1, w_2, \dots, w_l}$  is trivially rank deficient. Consequently, all  $\binom{l+c}{l}$   $l \times l$  submatrices of  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  are trivially rank deficient, so that  $(\mathcal{T}_j)_{\bar{j}}^{\bar{i}}$  is trivially rank deficient.  $\square$

We next characterize the MDP and sMDS properties in terms of trivial rank deficiency. We denote by  $r$  the difference of the generalized Singleton bound and the upper bound for the  $L$ th column distance:

$$\begin{aligned} r &:= (n-k) \left( \left\lfloor \frac{\delta}{k} \right\rfloor + 1 \right) + \delta + 1 - \left( \left\lfloor \frac{\delta}{k} \right\rfloor + \left\lfloor \frac{\delta}{n-k} \right\rfloor + 1 \right) (n-k) - 1 \\ &= \delta - \left\lfloor \frac{\delta}{n-k} \right\rfloor (n-k). \end{aligned}$$

Note that  $r$  is the remainder of  $\delta$  on division by  $n-k$ . If  $r = 0$ , then  $L = M$ , and a code is MDP if and only if it is sMDS. This case was considered in [8] and [11], so we will assume that  $r \in \{1, \dots, n-k-1\}$ . In this situation,  $M = L + 1$ .

**THEOREM 3.4.** *Let  $(A, B, C, D) \in S_{k,n}^\delta$  and  $\mathcal{C}$  be the  $(n, k, \delta)$ -code represented by  $(A, B, C, D)$ . Then,  $\mathcal{C}$  is an MDP code if and only if every square submatrix of  $\mathcal{T}_L$*

that is not trivially rank deficient has full rank.  $\mathcal{C}$  is an sMDS code if and only if, for every integer  $l$  satisfying  $1 \leq l \leq \min\{(M+1)(n-k) - (n-k-r), (M+1)k\}$ , every submatrix  $(\mathcal{T}_M)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+n-k-r}}$  that is not trivially rank deficient has full rank.

*Proof.* The first statement is [11, Corollary 2.5]. Next we consider the second statement.

$\Leftarrow$ : Suppose that

$$v := \begin{bmatrix} y_0^T & y_1^T & \cdots & y_M^T & | & u_0^T & u_1^T & \cdots & u_M^T \end{bmatrix}^T$$

is formed from the first  $M+1$  vectors of a finite-weight sequence for  $(A, B, C, D)$  with  $u_0 \neq 0$ , so that the matrix equation

$$\begin{bmatrix} -I_{(M+1)(n-k)} & \begin{bmatrix} D & 0 & \cdots & \cdots & 0 \\ CB & D & \ddots & & \vdots \\ CAB & CB & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ CA^{M-1}B & CA^{M-2}B & \cdots & CB & D \end{bmatrix} \end{bmatrix} v = 0$$

is satisfied, and denote the weight of

$$u := \begin{bmatrix} u_0^T & u_1^T & \cdots & u_M^T \end{bmatrix}^T$$

by  $w$ . For  $t \in \{1, \dots, w\}$ , let  $j_t$  denote the position of the  $t$ th nonzero entry in  $u$ , and let  $\bar{u}$  denote the vector obtained from  $u$  by deleting all of the zero entries. Suppose that  $(\mathcal{T}_M)_{\tilde{j}}^{\tilde{i}} := (\mathcal{T}_M)_{j_1, j_2, \dots, j_w}^{i_1, i_2, \dots, i_{w+n-k-r}}$  is a submatrix of  $\mathcal{T}_M$  such that

$$(3.2) \quad (\mathcal{T}_M)_{\tilde{j}}^{\tilde{i}} \bar{u} = 0.$$

Since  $u_0 \neq 0$ , we have that

$$j_1 \leq k \leq \left\lceil \frac{i_1}{n-k} \right\rceil k \leq \left\lceil \frac{i_{1+n-k-r}}{n-k} \right\rceil k.$$

By Lemma 3.3,  $(\mathcal{T}_M)_{j_1}^{i_1, i_2, \dots, i_{1+n-k-r}}$  is not trivially rank deficient, so that it has full rank. This means that at least one of its entries is nonzero, and, since (3.2) holds, it follows that

$$j_2 \leq \left\lceil \frac{i_{1+n-k-r}}{n-k} \right\rceil k \leq \left\lceil \frac{i_{2+n-k-r}}{n-k} \right\rceil k.$$

By Lemma 3.3,  $(\mathcal{T}_M)_{j_1, j_2}^{i_1, i_2, \dots, i_{2+n-k-r}}$  is not trivially rank deficient, so that it has full rank. Consequently, at least one  $2 \times 2$  minor of  $(\mathcal{T}_M)_{j_1, j_2}^{i_1, i_2, \dots, i_{2+n-k-r}}$  is nonzero. Again, since (3.2) holds, it follows that

$$j_3 \leq \left\lceil \frac{i_{2+n-k-r}}{n-k} \right\rceil k \leq \left\lceil \frac{i_{3+n-k-r}}{n-k} \right\rceil k.$$

Continuing, we see that, for  $t \in \{1, \dots, w\}$ ,

$$j_t \leq \left\lceil \frac{i_{t+n-k-r}}{n-k} \right\rceil k.$$



A final application of Lemma 3.3 gives that  $(\mathcal{T}_M)_{\tilde{j}}^{\tilde{i}}$  is not trivially rank deficient. By hypothesis, it must have full rank, which contradicts the hypothesis that  $(\mathcal{T}_M)_{\tilde{j}}^{\tilde{i}} \bar{u} = 0$ . Consequently, at most  $w+n-k-r-1$  rows of  $(\mathcal{T}_M)_{\tilde{j}}^{\tilde{i}}$  are in the left kernel of  $\bar{u}$ . It follows that  $v$  has weight at least  $w + ((M+1)(n-k) - (w+n-k-r-1)) = M(n-k) + 1 + r = (L+1)(n-k) + 1 + r$ , which means that  $d_M^c(\mathcal{C}) \geq (L+1)(n-k) + 1 + r$ . Recalling Proposition 2.6 and the definition of  $r$ , we conclude that  $d_M^c(\mathcal{C}) = (L+1)(n-k) + 1 + r$ , so that  $\mathcal{C}$  is sMDS.

$\Rightarrow$ : We prove the contrapositive. Suppose that the matrix (3.1) has a  $(w+n-k-r) \times w$  submatrix  $(\mathcal{T}_M)_{\tilde{j}}^{\tilde{i}} := (\mathcal{T}_M)_{j_1, j_2, \dots, j_w}^{i_1, i_2, \dots, i_{w+n-k-r}}$  that is not trivially rank deficient and that has less than full rank. There then exists a vector  $(\mathcal{T}_M)_{\tilde{j}}^{\tilde{i}} \bar{u} \neq 0$  of weight  $w' \leq w$  such that  $\bar{u} = 0$ . Let

$$u := \begin{bmatrix} u_0^T & u_1^T & \cdots & u_M^T \end{bmatrix}^T \in \mathbb{F}^{Mk}$$

be the vector in which the  $j_t$ th entry is the  $t$ th entry of  $\bar{u}$  and all other entries are zero; because of the block Toeplitz structure of  $\mathcal{T}_M$ , we may assume that  $u_0 \neq 0$ . Using (2.1), we form the vector

$$v := \begin{bmatrix} y_0^T & y_1^T & \cdots & y_M^T & | & u_0^T & u_1^T & \cdots & u_M^T \end{bmatrix}^T.$$

Because  $[-I_{(M+1)(n-k)} \mid \mathcal{T}_M]v = 0$ , the weight of  $v$  is at most  $w' + (M+1)(n-k) - (w+n-k-r) \leq (M+1)(n-k) - (n-k-r) = (L+1)(n-k) + r < (L+1)(n-k) + r + 1$ . We may choose additional information vectors  $u_{M+1}, \dots, u_d$  so that  $x_d = 0$  (see, for example, [2]); in other words, it is possible to extend  $v$  to a finite-weight sequence for  $(A, B, C, D)$  with weight less than the generalized Singleton bound. Thus,  $d_M^c(\mathcal{C}) < (L+1)(n-k) + r + 1$ , so that  $\mathcal{C}$  is not an sMDS code.  $\square$

Theorem 3.4 gives polynomial conditions on the entries of a realization  $(A, B, C, D) \in S_{k,n}^\delta$  that may be used to determine whether or not the convolutional code these matrices represent has the MDP and sMDS properties. In the next section, we use this information to show that we can find a realization  $(A, B, C, D) \in S_{k,n}^\delta$  representing an  $(n, k, \delta)$ -code that has the MDP and sMDS properties.

**4. Proof of the existence of sMDS convolutional codes.** Recall that we defined the block matrices making up  $\mathcal{T}_M$  in terms of matrices  $(A, B, C, D)$ . In this section, we will work in the opposite direction. Let  $\{F_0, F_1, \dots, F_j\}$  be a sequence of matrices in  $\mathbb{K}^{(n-k) \times k}$ . Slightly abusing notation, we set

$$(4.1) \quad \mathcal{T}_j := \begin{bmatrix} F_0 & 0 & \cdots & 0 \\ F_1 & F_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ F_j & F_{j-1} & \cdots & F_0 \end{bmatrix}.$$

The plan is to show the existence of a sequence  $\{F_0, F_1, \dots, F_M\}$  of matrices in  $\mathbb{K}^{(n-k) \times k}$  such that

1.  $\mathcal{T}_M$  has the property that, for all integers  $l$  with  $1 \leq l \leq \min\{(M+1)(n-k) - (n-k-r), (M+1)k\}$ , every submatrix  $(\mathcal{T}_M)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+n-k-r}}$  that is not trivially rank deficient has full rank;
2. there is a minimal partial realization  $(A, B, C, D) \in S_{k,n}^\delta$  of this matrix sequence (this means that  $D = F_0$  and  $CA^{i-1}B = F_i$  for  $1 \leq i \leq M$ ).

The matrices  $(A, B, C, D)$  will represent the desired code. We begin with the following lemma.

LEMMA 4.1. *There exists a sequence  $\{F_0, F_1, \dots, F_L\}$  of matrices in  $\mathbb{K}^{(n-k) \times k}$  such that every square submatrix of  $\mathcal{T}_L$  that is not trivially rank deficient has full rank.*

*Proof.* Note that we may think of such a matrix sequence  $\{F_0, F_1, \dots, F_L\}$  as a point in  $\mathbb{K}^{(L+1)(n-k)k}$ . To begin, think of the matrix (4.1) with  $j = L$  as being defined over the polynomial ring  $\mathbb{K}[x_1, x_2, \dots, x_{(L+1)(n-k)k}]$ , the entries corresponding with the indeterminates of this ring in a manner analogous to that in the previous section. When viewed in this way, the determinant of a square submatrix of  $\mathcal{T}_L$  that is not trivially rank deficient is a nonzero polynomial in  $\mathbb{K}[x_1, x_2, \dots, x_{(L+1)(n-k)k}]$ , and there is a finite number of such polynomials. The solution sets of these polynomials make up a proper algebraic subset of  $\mathbb{K}^{(L+1)(n-k)k}$ , the complement of which is a nonempty Zariski open set. Choose  $\{F_0, \dots, F_L\}$  to be a point in this open set.  $\square$

To determine the degree of a minimal partial realization of a matrix sequence  $\{F_0, F_1, \dots, F_M\}$ , we consider the matrices

$$\mathcal{F}_{x,y} := \begin{bmatrix} F_1 & F_2 & \cdots & F_y \\ F_2 & F_3 & \cdots & F_{y+1} \\ \vdots & \vdots & & \vdots \\ F_x & F_{x+1} & \cdots & F_{x+y-1} \end{bmatrix}.$$

In [22, Lemma 3], it is shown that the degree of a minimal partial realization of  $\{F_0, F_1, \dots, F_M\}$  is given by the expression

$$(4.2) \quad \sum_{x=1}^M \text{rank } \mathcal{F}_{x, M+1-x} - \sum_{x=1}^{M-1} \text{rank } \mathcal{F}_{x, M-x}.$$

The next results show that, starting with a matrix sequence  $\{F_0, F_1, \dots, F_L\}$  as described in Lemma 4.1, we can find a matrix  $F_M$  so that the expression (4.2) evaluates to  $\delta$ .

LEMMA 4.2. *Let  $\{F_0, \dots, F_M\}$  be a sequence of matrices in  $\mathbb{K}^{(n-k) \times k}$  such that every square submatrix of  $\mathcal{T}_L$  that is not trivially rank deficient has full rank. Then,*

1. *for  $x \in \{1, \dots, M-1\}$ ,  $\text{rank } \mathcal{F}_{x, M-x} = \min\{x(n-k), (M-x)k\}$ .*
2. *If  $\text{rank } \mathcal{F}_{x, M+1-x} < \min\{x(n-k), (M+1-x)k\}$ , then  $x = \lceil M \frac{k}{n} \rceil$ . If  $x \in \{1, \dots, M\} \setminus \{\lceil M \frac{k}{n} \rceil\}$ , then  $\text{rank } \mathcal{F}_{x, M+1-x} = \min\{x(n-k), (M+1-x)k\}$ .*
3. *Set  $\bar{x} := \lceil M \frac{k}{n} \rceil$ . The expression (4.2) reduces to  $\text{rank } \mathcal{F}_{\bar{x}, M+1-\bar{x}}$ .*

*Proof.* To verify the first claim, observe that  $\mathcal{F}_{x, M-x}$  differs by a column permutation from a submatrix of  $\mathcal{T}_L$  that has full rank.

For the second claim, suppose first that  $x(n-k) \leq (M+1-x)k$ . The hypothesis is then that  $\text{rank } \mathcal{F}_{x, M+1-x} < x(n-k)$ . If  $x < M$ , it follows from 1 that  $\text{rank } \mathcal{F}_{x, M-x} = \min\{x(n-k), (M-x)k\}$ , which means that  $x(n-k) > (M-x)k$ . Together, this gives

$$(M-x)k < x(n-k) \leq (M+1-x)k$$

(note that the first inequality also holds if  $x = M$ ). This can be rewritten as

$$M \frac{k}{n} < x \leq (M+1) \frac{k}{n}.$$

If we suppose instead that  $(M+1-x)k \leq x(n-k)$ , similar reasoning leads to

$$(M+1)\frac{k}{n} \leq x < M\frac{k}{n} + 1.$$

In all, we have

$$M\frac{k}{n} < x < M\frac{k}{n} + 1.$$

Since  $x$  is an integer,  $x = \lceil M\frac{k}{n} \rceil$ . The second statement follows immediately.

The third claim follows directly from the first two, since  $x < \bar{x} \implies x(n-k) < (M-x)k$  and  $x > \bar{x} \implies x(n-k) > (M+1-x)k$ .  $\square$

**THEOREM 4.3.** *Let  $\{F_0, \dots, F_L\}$  be a sequence of matrices in  $\mathbb{K}^{(n-k) \times k}$  such that every square submatrix of  $\mathcal{T}_L$  that is not trivially rank deficient has full rank. Then, one can find a matrix  $F_M \in \mathbb{K}^{(n-k) \times k}$  such that*

1. *the matrix*

$$\mathcal{F}_{\bar{x}, M+1-\bar{x}} = \begin{bmatrix} F_1 & F_2 & \cdots & F_{M+1-\bar{x}} \\ F_2 & F_3 & \cdots & F_{M+2-\bar{x}} \\ \vdots & \vdots & \ddots & \vdots \\ F_{\bar{x}} & F_{\bar{x}+1} & \cdots & F_M \end{bmatrix}$$

*has rank  $\delta$ ;*

2. *the matrix  $\mathcal{T}_M$  has the property that, for every integer  $l$  with  $1 \leq l \leq \min\{(M+1)(n-k)-(n-k-r), (M+1)k\}$ , every submatrix  $(\mathcal{T}_M)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+n-k-r}}$  that is not trivially rank deficient has full rank.*

*Proof.* We may write

$$\delta = \left\lfloor \frac{\delta}{n-k} \right\rfloor (n-k) + r = \left\lfloor \frac{\delta}{k} \right\rfloor k + r',$$

where  $1 \leq r < n-k$  and  $0 \leq r' < k$ . Since

$$M = L+1 = \left\lfloor \frac{\delta}{n-k} \right\rfloor + \left\lfloor \frac{\delta}{k} \right\rfloor + 1,$$

we see that

$$\begin{aligned} \frac{Mk}{n} &= \left\lfloor \frac{\delta}{n-k} \right\rfloor \frac{k}{n} + \left\lfloor \frac{\delta}{k} \right\rfloor \frac{k}{n} + \frac{k}{n} = \left\lfloor \frac{\delta}{n-k} \right\rfloor - \left\lfloor \frac{\delta}{n-k} \right\rfloor \frac{n-k}{n} + \left\lfloor \frac{\delta}{k} \right\rfloor \frac{k}{n} + \frac{k}{n} \\ &= \left\lfloor \frac{\delta}{n-k} \right\rfloor - \frac{\delta-r}{n} + \frac{\delta-r'}{n} + \frac{k}{n} = \left\lfloor \frac{\delta}{n-k} \right\rfloor + \frac{k-r'+r}{n}. \end{aligned}$$

Since  $1 < k-r'+r < n$ , we have  $\lfloor \frac{\delta}{n-k} \rfloor = \bar{x}-1$ , so that  $\delta = (\bar{x}-1)(n-k) + r$  and

$$\frac{Mk}{n} = \bar{x}-1 + \frac{k-r'+r}{n}.$$

Multiplying both sides by  $n$  and subtracting  $\bar{x}k$  from both sides, we get

$$(M-\bar{x})k = (\bar{x}-1)(n-k) + r - r' = \delta - r',$$

from which it follows that  $(M-\bar{x})k \leq \delta$ . Since  $r' < k$ , it also follows that  $\delta < (M+1-\bar{x})k$ .

We next want to see that we may find a matrix  $F_M$  as described in the statement of the theorem. We first consider the top  $r$  rows of  $F_M$ . Using the same reasoning as in the proof of Lemma 4.1, we may find elements of  $\mathbb{K}$  to form these top  $r$  rows so that all square submatrices of the top  $M(n-k) + r$  rows of  $\mathcal{T}_M$  that are not trivially rank deficient have full rank. In particular, all square submatrices of the top  $\delta$  rows of  $\mathcal{F}_{\bar{x}, M+1-\bar{x}}$  have full rank. Denote the  $r \times k$  matrix consisting of these  $r$  rows by  $F'_M$ . Since  $\delta < (M+1-\bar{x})k$ ,  $\text{rank } \mathcal{F}_{\bar{x}, M+1-\bar{x}} \geq \delta$  will hold regardless of how the entries of the bottom  $n-k-r$  rows of  $F_M$  are chosen. To find entries for these rows so that  $\text{rank } \mathcal{F}_{\bar{x}, M+1-\bar{x}} = \delta$ , consider the top  $\delta$  rows of  $\mathcal{F}_{\bar{x}, M-\bar{x}}$ . Since  $\delta \geq (M-\bar{x})k$ , we may choose  $M-\bar{x}$  of these  $\delta$  rows to form an  $(M-\bar{x})k \times (M-\bar{x})k$  submatrix that necessarily has full rank. This means that the last  $n-k-r$  rows of  $\mathcal{F}_{\bar{x}, M-\bar{x}}$  may each be expressed as a linear combination of the rows of our chosen submatrix. Consequently, we may take the last  $n-k-r$  rows of  $F_M$  to be the corresponding linear combinations of the rows of

$$\begin{bmatrix} F_{M+1-\bar{x}} \\ F_{M+2-\bar{x}} \\ \vdots \\ F'_M \end{bmatrix}$$

extending the rows of our chosen submatrix. With this, we have found an  $F_M$  so that  $\text{rank } \mathcal{F}_{\bar{x}, M+1-\bar{x}} = \delta$ .

Suppose finally that  $(\mathcal{T}_M)_{\bar{j}}^{\bar{i}} := (\mathcal{T}_M)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_{l+n-k-r}}$  is a submatrix of  $\mathcal{T}_M$  that is not trivially rank deficient and does not have full rank. Then, in particular,  $(\mathcal{T}_M)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_l}$  does not have full rank. Since  $(\mathcal{T}_M)_{j_1, j_2, \dots, j_l}^{i_1, i_2, \dots, i_l}$  is contained in the top  $M(n-k) + r$  rows of  $\mathcal{T}_M$ , it must be trivially rank deficient. By Lemma 3.3, there exists a smallest integer  $t \in \{1, \dots, l\}$  such that

$$j_t > \left\lceil \frac{i_t}{n-k} \right\rceil k.$$

Since  $(\mathcal{T}_M)_{\bar{j}}^{\bar{i}}$  is not trivially rank deficient, it also follows from Lemma 3.3 that

$$j_\tau \leq \left\lceil \frac{i_\tau + n - k - r}{n - k} \right\rceil k \quad \forall \tau \in \{1, \dots, l\},$$

so that  $(\mathcal{T}_M)_{\bar{j}}^{\bar{i}} := (\mathcal{T}_M)_{j_t, j_{t+1}, \dots, j_l}^{i_t + n - k - r, i_{t+1} + n - k - r, \dots, i_{l+n-k-r}}$  is not trivially rank deficient. Since  $j_t > k$ ,  $(\mathcal{T}_M)_{\bar{j}}^{\bar{i}}$  must be a submatrix of  $\mathcal{T}_L$ , which means that  $(\mathcal{T}_M)_{\bar{j}}^{\bar{i}}$  has full rank. Recalling how  $t$  was chosen, we see that  $(\mathcal{T}_M)_{\bar{j}}^{\bar{i}}$  has full rank. This is a contradiction. We conclude that if a submatrix  $(\mathcal{T}_M)_{\bar{j}}^{\bar{i}}$  is not trivially rank deficient, then it has full rank.  $\square$

**COROLLARY 4.4.** *Let  $\{F_0, \dots, F_M\}$  be as in Theorem 4.3. Then, the expression (4.2) evaluates to  $\delta$ .*

We are now ready to finish our existence proof.

**THEOREM 4.5.** *An MDP and sMDS  $(n, k, \delta)$ -code exists over a sufficiently large finite field of characteristic  $p$ .*

*Proof.* By Lemma 4.1 and Theorem 4.3, we can find a sequence  $\{F_0, \dots, F_M\}$  of matrices in  $\mathbb{K}^{(n-k) \times k}$  such that

1. every square submatrix of  $\mathcal{T}_L$  that is not trivially rank deficient has full rank;

2. every  $(l + n - k - r) \times l$  submatrix of  $\mathcal{T}_M$  that is not trivially rank deficient has full rank;
3. the minimum possible degree of a partial realization of  $\{F_0, \dots, F_M\}$  is  $\delta$ .

Since there are a finite number of entries in the matrices  $\{F_0, \dots, F_M\}$ , the entries all belong to some finite subfield  $\mathbb{F}$  of  $\mathbb{K}$ . From [22, Theorem 1], there is a minimal realization  $(A, B, C, D) \in S_{k,n}^\delta$  of the sequence  $\{F_0, \dots, F_M\}$  with entries in  $\mathbb{F}$ . By Theorem 3.4, the  $(n, k, \delta)$ -code represented by  $(A, B, C, D)$  is both MDP and sMDS.  $\square$

With this, we have shown that the conjecture in [8] that codes having both the MDP and sMDS properties exist for all parameters  $(n, k, \delta)$  is correct. It is still an open problem as to how one may construct matrices of the form (3.1) leading to codes with these properties, and this must be left for future research.

**Acknowledgments.** The author wishes to thank Joachim Rosenthal, Heide Gluesing-Luerssen, and José Ignacio Iglesias Curto for helpful comments during the preparation of this paper. He also wishes to thank the anonymous referees for their careful readings and detailed comments.

#### REFERENCES

- [1] A. C. ANTOLAS, *On recursiveness and related topics in linear systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 1121–1135.
- [2] P. J. ANTSAKLIS AND A. N. MICHEL, *Linear Systems*, Birkhäuser, Boston, 2006.
- [3] D. F. DELCHAMPS, *State Space and Input-Output Linear Systems*. Springer-Verlag, New York, 1988.
- [4] P. L. FALB, *Methods of Algebraic Geometry in Control Theory: Part I*, Birkhäuser, Boston, 1990.
- [5] P. L. FALB, *Methods of Algebraic Geometry in Control Theory: Part II*, Birkhäuser, Boston, 1999.
- [6] M. FLIESS, *On the structure of linear recurrent error-control codes*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 703–713.
- [7] G. D. FORNEY, JR., *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, SIAM J. Control, 13 (1975), pp. 493–520.
- [8] H. GLUESING-LUERSSEN, J. ROSENTHAL, AND R. SMARANDACHE, *Strongly MDS convolutional codes*, IEEE Trans. Inform. Theory, 52 (2006), pp. 584–598.
- [9] H. GLUESING-LUERSSEN AND W. SCHMALE, *On cyclic convolutional codes*, Acta Appl. Math., 82 (2004), pp. 183–237.
- [10] C. N. HADJICOSTIS, *Non-concurrent error detection and correction in fault-tolerant linear finite-state machines*, IEEE Trans. Automat. Control, 48 (2003), pp. 2133–2140.
- [11] R. HUTCHINSON, J. ROSENTHAL, AND R. SMARANDACHE, *Maximum distance profile convolutional codes*, Systems Control Lett., 54 (2004), pp. 53–63.
- [12] R. JOHANNESSON AND K. ZIGANGIROV, *Distances and distance bounds for convolutional codes—an overview*, in Topics in Coding Theory. In Honour of L. H. Zetterberg, Lecture Notes in Control and Inform. Sci. 128, Springer-Verlag, Berlin, 1989, pp. 109–136.
- [13] R. JOHANNESSON AND K. SH. ZIGANGIROV, *Fundamentals of Convolutional Coding*, IEEE Press, New York, 1999.
- [14] B. LANGFELD, *Minimal Cyclic Convolutional Codes*, Diploma thesis, University of Oldenburg, Oldenburg, Germany, 2003.
- [15] V. LOMADZE, *Finite-dimensional time-invariant linear dynamical systems: Algebraic theory*, Acta Appl. Math., 19 (1990), pp. 149–201.
- [16] R. J. MCELIECE, *The algebraic theory of convolutional codes*, in Handbook of Coding Theory, Vol. 1, Elsevier, Amsterdam, 1998, pp. 1065–1138.
- [17] J. ROSENTHAL, J. M. SCHUMACHER, AND E. V. YORK, *On behaviors and convolutional codes*, IEEE Trans. Inform. Theory, 42 (1996), pp. 1881–1891.
- [18] J. ROSENTHAL AND R. SMARANDACHE, *Maximum distance separable convolutional codes*, Appl. Algebra Engrg. Comm. Comput., 10 (1999), pp. 15–32.
- [19] J. ROSENTHAL AND E. V. YORK, *BCH convolutional codes*, IEEE Trans. Inform. Theory, 45 (1999), pp. 1833–1844.

- [20] J. ROSENTHAL, *Connections between linear systems and convolutional codes*, in Codes, Systems and Graphical Models, IMA Vol. Math. Appl. 123, B. Marcus and J. Rosenthal, eds., Springer-Verlag, New York, 2001, pp. 39–66.
- [21] R. SMARANDACHE, H. GLUESING-LUERSSEN, AND J. ROSENTHAL, *Constructions for MDS-convolutional codes*, IEEE Trans. Inform. Theory, 47 (2001), pp. 2045–2049.
- [22] A. J. TETHER, *Construction of minimal linear state-variable models from finite input-output data*, IEEE Trans. Automat. Control, 15 (1970), pp. 427–436.

## MINIMAL TIME SEQUENTIAL BATCH REACTORS WITH BOUNDED AND IMPULSE CONTROLS FOR ONE OR MORE SPECIES\*

P. GAJARDO<sup>†</sup>, H. RAMÍREZ C.<sup>‡</sup>, AND A. RAPAPORT<sup>§</sup>

**Abstract.** We consider the optimal control problem of feeding in minimal time a tank where several species compete for a single resource, with the objective being to reach a given level of the resource. We allow controls to be bounded measurable functions of time plus possible impulses. For the one-species case, we show that the immediate one-impulse strategy (filling the whole reactor with one single impulse at the initial time) is optimal when the growth function is monotonic. For nonmonotonic growth functions with one maximum, we show that a particular singular arc strategy (precisely defined in section 3) is optimal. These results extend and improve former ones obtained for the class of measurable controls only. For the two-species case with monotonic growth functions, we give conditions under which the immediate one-impulse strategy is optimal. We also give optimality conditions for the singular arc strategy (at a level that depends on the initial condition) to be optimal. The possibility for the immediate one-impulse strategy to be nonoptimal while both growth functions are monotonic is a surprising result and is illustrated with the help of numerical simulations.

**Key words.** minimal time problem, chemostat, Hamilton–Jacobi–Bellman equation, Pontryagin’s minimum principle, impulse control

**AMS subject classifications.** 49J15, 49N25

**DOI.** 10.1137/070695204

**1. Introduction.** Sequential batch reactors (SBR) are often used in biotechnological industries, notably in waste-water treatment. Typically, a tank is filled with activated sludge or biological microorganisms capable of degrading some undesirable substrate. The method then consists of a sequence of cycles composed of three phases:

- Phase 1: Filling the reactor with water to be treated,
- Phase 2: Waiting for the concentration of the undesirable substrate to decrease until a given (low) concentration,
- Phase 3: Emptying the *clean* water from the reactor, leaving the sludge inside.

The time necessary to achieve such cycles can be substantially long and can have an economic impact on the overall process. Manipulating the input flow during the filling phase clearly has an influence on the total duration of the cycle (more precisely, the duration of Phases 1 and 2, the duration of Phase 3 being fixed). But the nonlinear kinetics of the biological reactions do not always make easy the determination of the input flow strategy that minimizes the total time obvious.

Very similar problems (optimizing the production of biomass at a fixed terminal time) have already been tackled with the help of optimal control theory [1, 11, 10],

---

\*Received by the editors June 22, 2007; accepted for publication (in revised form) July 6, 2008; published electronically November 19, 2008. This research was partially supported by the INRIA-CONICYT program.

<http://www.siam.org/journals/sicon/47-6/69520.html>

<sup>†</sup>Departamento de Matemática, Universidad Técnica Federico Santa María, Avda. España 1680 Casilla 110-V, Valparaíso, Chile (pedro.gajardo@usm.cl). Partially supported by FONDECYT project 1080173 and Fondo Basal, Centro de Modelamiento Matemático, Universidad de Chile.

<sup>‡</sup>Departamento de Ingeniería Matemática and Center for Mathematical Modelling (CNRS UMI 2807), Universidad de Chile, Casilla 170/3 Santiago, Chile (hramirez@dim.uchile.cl). Partially supported by FONDECYT project 1070297.

<sup>§</sup>INRA-INRIA project MERE. UMR Analyse des Systèmes et Biométrie, 2 Place Viala, 34060 Montpellier Cedex 2, France (rapaport@ensam.inra.fr).

which has led to computational methods [30, 12]. For models with one biological species, a solution of the minimal time problem has been proposed by Moreno in [20] for monotonic as well as nonmonotonic kinetics. It has been proved that, for monotonic growth functions, such as the Monod law (see [29]), the optimal solution consists of a *most rapid approach* strategy, namely, filling the tank up to its maximum capacity as fast as possible and then waiting. For nonmonotonic growths with one maximum, such as the Haldane law (see [29]), a *singular arc* strategy which consists of maintaining the resource level that maximizes the growth function for most of the time, have been proved to be optimal. The optimality proofs are based on a technique due to Miele [13] using Green's theorem. More precisely, proofs rely on a reformulation of the problem in a planar one.

In the present work, we consider minimal time problems where more than one species can compete for the same substrate. For these cases, the problem cannot be reformulated into a planar one, and the technique mentioned above does not apply. Nevertheless, we are interested in characterizing biological systems for which the *most rapid approach* strategy is again optimal. We are also interested in identifying conditions for which a *singular arc* strategy could be optimal.

We shall allow adding unbounded or impulsive controls to the usual measurable bounded controls. The practical motivation for such a consideration comes from the fact that a bounded measurable control can be incorporated into a device that tunes the speed of a pump over a certain range, while an unbounded control can be assimilated to an instantaneous dilution of a positive volume, as in [9]. A similar optimal control problem for fed-batch processes has been studied in [32] but for a fixed terminal time and a final cost. A characterization of minimal time functions with impulse controls and state constraints has been proposed in [7, 26]. In [7], some restrictive conditions are considered on the jumps that do not apply to the present problem. In [26], the minimal time function is characterized but as a function of a bound on the total variation allowed on the unbounded control. For related results concerning the regularity of the value function for minimal time problems, see [24] and the references therein.

For our problem with a scalar control, we use a smooth time parameterization in the spirit of [33] and [34], which differs from more general approaches that use discontinuous time transformation (see, for instance, [3, 8, 14, 15, 16, 18, 19] or [35]). The possibility of immediately reaching the target with a single jump has also led us to extend the definition of the singular arc strategy to the framework of impulse controls.

Even though the main contribution of this paper is the analysis of the two species case, it is worth noting that the former results of Moreno [20] for the one-species case without impulse controls did not consider the parametric configuration  $s^* < s_{out}$  (the notation will be defined in section 3). This case leads to more complicated optimal trajectories, as we shall show. Furthermore, we provide an explicit expression for the value function for any parametric configuration.

The paper is organized as follows. In the next section, we state the minimal time problem with impulse control and give an equivalent formulation with measurable controls. In section 3, we define the *one-impulse* and *singular arc* strategies. Section 4 characterizes the cost of the *one-impulse* strategy, which plays an important role in the following sections. Section 5 gives the Hamilton–Jacobi formulation of the problem and states optimality results for the strategies presented in section 3. The use of the minimum principle is presented in section 6. Finally, applications to the one- and two-species cases are given in sections 7 and 8, respectively.



**2. Formulation of the problem.** The dynamics of an SBR with several species can be described by the following set of ordinary differential equations (see [29]):

$$(2.1) \quad \begin{cases} \dot{x}_i = \mu_i(s)x_i - \frac{F}{v}x_i, & x_i(t_0) = y_i \quad (i = 1 \cdots n), \\ \dot{s} = -\sum_{j=1}^n \mu_j(s)x_j + \frac{F}{v}(s_{in} - s), & s(t_0) = z, \\ \dot{v} = F, & v(t_0) = w, \end{cases}$$

where  $x_i$ ,  $s$ , and  $v$  stand, respectively, for the concentration of the  $i$ th species, the concentration of the substrate, and the current volume of water present in the tank. The parameter  $s_{in} > 0$  is a constant which represents the substrate concentration in the input flow. The growth functions  $\mu_i(\cdot)$  are nonnegative smooth functions such that  $\mu_i(0) = 0$ , and the input flow  $F$  is a nonnegative control variable.

Given a (desirable) substrate concentration  $s_{out} \in ]0, s_{in}[$  and a volume (of the reactor)  $v_{max} > 0$ , consider the domain  $\mathcal{D} = (\mathbb{R}_+^n \setminus \{0\}) \times ]0, s_{in}] \times ]0, v_{max}[$  and the target  $\mathcal{T} = \mathbb{R}_+^n \times ]0, s_{out}] \times \{v_{max}\}$ . From any initial condition  $\xi = (y, z, w)$  in  $\mathcal{D}$  at time  $t_0$ , the objective is to reach  $\mathcal{T}$  in minimal time. Let us write  $V(\cdot)$  the value function of the problem

$$(2.2) \quad V(\xi) = \inf_{F(\cdot)} \{t - t_0 \mid s^{t_0, \xi, F}(t) \leq s_{out}, v^{t_0, \xi, F}(t) = v_{max}\},$$

where  $s^{t_0, \xi, F}(\cdot)$ ,  $v^{t_0, \xi, F}(\cdot)$  denote solutions of (2.1), with initial condition  $\xi \in \mathcal{D}$  at time  $t_0$  and control  $F(\cdot)$ .

We allow here  $F(\cdot)$  to be a nonnegative measurable function plus possible positive impulses. The question of the proper treatment of optimal control problems with unbounded or *impulse* controls has already been studied in the literature (see [5, 6, 8, 14, 15, 17, 18, 19, 21, 22, 23, 26, 27, 35]). Instead of an ordinary control  $F(\cdot)$ , we consider a measure  $dF(\cdot)$  that we decompose into a sum of a measure absolutely continuous with respect to the Lebesgue measure  $u(t)dt$  and a singular or *impulsive* part  $d\sigma$  (see [33, 34]):

$$(2.3) \quad dF(t) = u(t)dt + d\sigma.$$

Here,  $u(\cdot)$  is a measurable nonnegative control that we impose to be bounded from above by  $u_{max}$ , because it corresponds to the use of a *pump* device. At time  $t$ , the nonnegative *impulse*  $d\sigma$  corresponds to an (instantaneous) addition of volume from  $v^-(t)$  to  $v^+(t)$ . When  $d\sigma$  is nonnull, it implies that the concentrations  $x_i$  and  $s$  jump as follows:

$$\begin{cases} x_i^+(t) = x_i^-(t) \frac{v^-(t)}{v^+(t)}, \\ s^+(t) = s^-(t) \frac{v^-(t)}{v^+(t)} + s_{in} \left(1 - \frac{v^-(t)}{v^+(t)}\right). \end{cases}$$

Notice that such a jump is equivalent to integrate the dynamics

$$(2.4) \quad \frac{dx_i}{d\tau} = -\frac{u}{v}x_i, \quad \frac{ds}{d\tau} = \frac{u}{v}(s_{in} - s), \quad \frac{dv}{d\tau} = u,$$

from  $\tau^-$  to  $\tau^+$ , with any regular nonnegative control  $u(\cdot)$  bounded from above by  $u_{max}$ , provided that the integral constraint is fulfilled:

$$(2.5) \quad \int_{\tau^-}^{\tau^+} u(\tau)d\tau = v^+(t) - v^-(t).$$

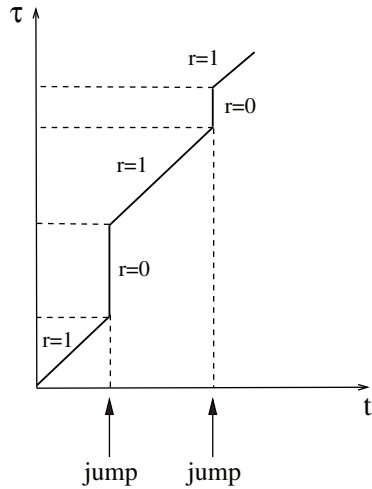


FIG. 2.1. Time parameterization.

Consider then a time parameterization  $\tau \geq t_0$  such that  $dt = r(\tau)d\tau$  (see Figure 2.1), where

$$r(\tau) = \begin{cases} 1 & \text{when } dF \text{ is absolutely continuous (a.c.) at } t(\tau), \\ 0 & \text{otherwise.} \end{cases}$$

Then, dynamics (2.1) with  $dF$  regular, and dynamics (2.4) with nonnull  $d\sigma$  can be gathered into the system

$$(2.6) \quad \begin{cases} \frac{dx_i}{d\tau} = r\mu_i(s)x_i - \frac{u}{v}x_i & (i = 1 \cdots n), \\ \frac{ds}{d\tau} = -r \sum_{j=1}^n \mu_j(s)x_j + \frac{u}{v}(s_{in} - s), \\ \frac{dv}{d\tau} = u, \end{cases}$$

where the controls  $u(\cdot)$  and  $r(\cdot)$  are sought among measurable functions w.r.t.  $\tau$ , taking values in  $[0, u_{\max}]$  and  $\{0, 1\}$ , respectively. Notice that, in this formulation,  $u(\cdot)$  plays both the role of an ordinary control when  $r = 1$  and the control of the amplitude of the jump (2.5) when  $r = 0$ , with the same single constraint  $u \in [0, u_{\max}]$ .

*Remark 1.* We could have considered two distinct controls, as in [18], for instance, if we write the system (2.1) as

$$\dot{X}(t) = f(X(t)) + F(t)g(X(t)),$$

where  $X = (x_i, s, v)$ ,  $i = 1, \dots, n$  and functions  $f$  and  $g$  are suitably chosen to be compatible with the dynamics (2.1), we could consider

$$\dot{X}(t) = f(X(t)) + (u_1(t) + u_2(t))g(X(t)),$$

with  $u_1$  a measurable nonnegative bounded (by  $u_{\max}$ ) control and  $u_2$  is an unbounded nonnegative control.

In this framework, the time reparametrization  $dt = r d\tau$ , with  $r \in [0, 1]$ , leads to the dynamics

$$(2.7) \quad \frac{dX}{d\tau} = r(\tau)[f(X(\tau)) + u_1(\tau)g(X(\tau))] + (1 - r(\tau))w_2(\tau)g(X(\tau)),$$

with  $u_1$  and  $w_2$  nonnegative bounded controls. One can also require  $w_2$  to be bounded by the same bounds  $u_{\max}$  as  $u_1$ , where  $u_1(\tau)r(\tau)d\tau = u_1(t)dt$  and  $(1 - r(\tau))w_2(\tau)d\tau = du_2(t)$ . The dynamics (2.7) is equivalent to

$$\frac{dX}{d\tau} = r(\tau)f(X(\tau)) + u(\tau)g(X(\tau)),$$

where  $u = ru_1 + (1 - r)w_2$  belongs to  $[0, u_{\max}]$ .

*Remark 2.* Since one can always take  $r = 0$  and  $u = 0$  on an arbitrarily large  $\tau$ -interval without modifying the total time  $\int_{t_0}^{\tau} r(\theta)d\theta$ , the minimal time problem has no unique solution. Hence, without loss of generality, we will be only interested in controls that never take null values simultaneously, that is, satisfying  $r(\tau) \neq 0$  or  $u(\tau) \neq 0$  for all time  $\tau$ .

Let us define the set of admissible controls by

$$(2.8) \quad \mathcal{C} = \{(u, r) : [0, +\infty) \mapsto [0, u_{\max}] \times \{0, 1\} \setminus \{(0, 0)\} \text{ Lebesgue measurable}\},$$

and let us write now  $V(\cdot)$  the value function of the reformulated problem (2.6)

$$(2.9) \quad V(\xi) = \inf_{(u, r)(\cdot) \in \mathcal{C}} \left\{ \int_{t_0}^{\tau} r(\theta)d\theta \mid s^{t_0, \xi, u, r}(\tau) \leq s_{out}, v^{t_0, \xi, u, r}(\tau) = v_{\max} \right\},$$

where  $s^{t_0, \xi, u, r}(\cdot)$ ,  $v^{t_0, \xi, u, r}(\cdot)$  denote solutions of (2.6), with initial condition  $\xi \in \mathcal{D}$  at time  $t_0$  and controls  $u(\cdot)$  and  $r(\cdot)$ .

*Remark 3.* Any trajectory of the dynamics (2.6) with initial condition  $\xi = (y, z, w) \in \mathcal{D}$  lies in the region defined by

$$(2.10) \quad \rho(\xi) = v \left( \sum_{j=1}^n x_j + s - s_{in} \right) = w \left( \sum_{j=1}^n y_j + z - s_{in} \right).$$

By using the above fact, one can write the variable  $s$  in terms of the other variables as follows:

$$(2.11) \quad s = \frac{\rho(\xi)}{v} - \sum_{j=1}^n x_j + s_{in}.$$

This is a key step in the approach used in Moreno [20] that reformulates the problem with one species in a planar one. However, since it does not simplify our results, we shall work with all of the variables. In the proof of Proposition 7.4 only, equality (2.11) will be used.

**3. The one-impulse and singular arc strategies.** From an initial state  $\xi = (y, z, w) \in \mathcal{D}$  at time  $t_0$ , we define the *immediate one impulse* strategy (that we shall denote *IOI strategy* in the following), which consists in making the following:

1. An impulse of volume  $v_{\max} - w$  at  $t_0$ . This can be achieved by  $r(\tau) = 0$ ,  $u(\tau) = u_{\max}$ , for  $\tau \in [t_0, t_0 + (v_{\max} - w)/u_{\max}]$ .
2. A null control (no feeding) until the concentration  $s(\tau)$  reaches  $s_{out}$ .

For convenience, we shall denote by  $\tilde{y}(\xi)$  and  $\tilde{z}(\xi)$  the concentrations obtained with an impulse of volume  $v_{\max} - w$  from a state  $\xi = (y, z, w) \in \mathcal{D}$ :

$$(3.1) \quad \tilde{y}(\xi) = y \frac{w}{v_{\max}}, \quad \tilde{z}(\xi) = z \frac{w}{v_{\max}} + s_{in} \left( 1 - \frac{w}{v_{\max}} \right).$$

Notice that, for the particular case  $\tilde{z}(\xi) \leq s_{out}$ , the first step only is used.

A second strategy considered in this paper is defined as follows. Consider a time  $t_0$ , a state  $\xi = (y, z, w) \in \mathcal{D}$ , a level substrate  $\bar{s}$  in  $]0, s_{in}[$ , and define the quantity

$$(3.2) \quad s^\dagger(\bar{s}, w) = s_{in} - (s_{in} - \max(\bar{s}, s_{out})) \frac{v_{\max}}{w}.$$

The *singular arc* strategy on the level  $\bar{s}$ , denoted by  $SA(\bar{s})$ , consists of the following steps.

1. *First step*:

- a. If  $z > s^\dagger(\bar{s}, w)$  and  $z < \bar{s}$ , make an impulse of volume  $w(\bar{s} - z)/(s_{in} - \bar{s})$  at  $t_0$ . This can be achieved by  $r(\tau) = 0$  and  $u(\tau) = u_{\max}$ , for  $\tau \in [t_0, \bar{t}]$ , where  $\bar{t} = t_0 + w(\bar{s} - z)/u_{\max}(s_{in} - \bar{s})$  (then  $s$  and  $v$  jump to  $\bar{s}$  and  $\bar{v} = w(s_{in} - z)/(s_{in} - \bar{s}) \leq v_{\max}$ , respectively).
- b. If  $z \geq \bar{s}$  and  $z > s^\dagger(\bar{s}, w)$ , apply a null control (no feeding) until the concentration  $s(\cdot)$  reaches the value  $\max(\bar{s}, s^\dagger(\bar{s}, w))$ , i.e.,  $r(\tau) = 1$ ,  $u(\tau) = 0$  for  $\tau \in [t_0, \bar{t}]$ , where  $\bar{t}$  is such that  $s^{t_0, y, z}(\bar{t}) = \max(\bar{s}, s^\dagger(\bar{s}, w))$  and  $s^{t_0, y, z}(\cdot)$  is the solution of the free dynamics

$$(3.3) \quad \begin{cases} \frac{dx_i}{d\tau} = \mu_i(s)x_i, & x_i(t_0) = y_i \quad (i = 1 \cdots n), \\ \frac{ds}{d\tau} = -\sum_{j=1}^n \mu_j(s)x_j, & s(t_0) = z. \end{cases}$$

- c. If  $z \leq s^\dagger(\bar{s}, w)$ , make an impulse of volume  $v_{\max} - w$  and go to the third step.

2. *Second step*:

- a. If the current state  $s$  is equal to  $\bar{s}$ , make a *singular arc*<sup>1</sup> by taking  $r(\tau) = 1$  and a suitable control  $u(\cdot)$  ensuring  $s(\tau) = \bar{s}$  for any  $\tau \in (\bar{t}, \tilde{T}]$ , where  $\tilde{T}$  is such that  $v(\tilde{T}^+) = v^\dagger(\bar{s})$  and the volume  $v^\dagger(\bar{s})$  is defined as follows:

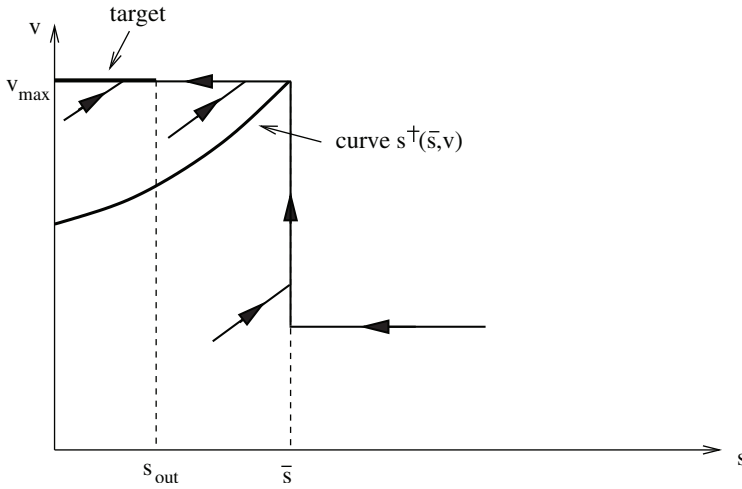
$$(3.4) \quad v^\dagger(\bar{s}) = v_{\max} \min \left( 1, \frac{s_{in} - s_{out}}{s_{in} - \bar{s}} \right).$$

If  $v^\dagger(\bar{s}) < v_{\max}$  (or, equivalently,  $\bar{s} < s_{out}$ ), then make an impulse of volume  $v_{\max} - v^\dagger(\bar{s})$ , and the process is finished. Otherwise, go to the third step.

- b. If the current state  $s$  is equal to  $s^\dagger(\bar{s}, w)$ , make an impulse of volume  $v_{\max} - w$ , and the process is finished.

3. *Third step*: Apply  $r = 1$  and a null control  $u$  until the concentration  $s(\cdot)$  reaches  $s_{out}$ .

<sup>1</sup>See [4, Part III Chapter 2] for a formal definition.


 FIG. 3.1. The  $SA(\bar{s})$  synthesis when  $\bar{s} > s_{out}$ .

Notice that  $z \leq s^\dagger(\bar{s}, w)$  implies that before reaching the substrate level  $\bar{s}$  with an impulse of volume, one reaches the volume  $v_{\max}$ . When  $z > s^\dagger(\bar{s}, w)$  and  $s^\dagger(\bar{s}, w) > \bar{s}$ , the variable  $s$  reaches the value  $s^\dagger(\bar{s}, w)$  before  $\bar{s}$ , and then an impulse drives directly to the target.

Observe also that, in order to apply the singular arc strategy on  $\bar{s}$ , imposing  $ds/d\tau = 0$ , the following constraint on the control  $u$  must be satisfied:

$$\frac{v}{(s_{in} - \bar{s})} \sum_{j=1}^n \mu_j(\bar{s}) x_j = u \leq u_{\max}.$$

Since the maximum level of substrate on which one can apply a singular arc, starting from  $\xi \in \mathcal{D}$ , is given by  $\tilde{z}(\xi)$  defined in (3.1), a sufficient condition, on the initial condition  $\xi$ , in order to guarantee the above inequality is to have

$$(3.5) \quad M \left( \frac{\rho(\xi)}{s_{in} - \tilde{z}(\xi)} + v_{\max} \right) \leq u_{\max},$$

where  $\rho(\xi)$  is defined by (2.10) and

$$M = \max_{\substack{s \in [0, s_{in}] \\ j=1, \dots, n}} \mu_j(s).$$

Indeed, from the definition of  $\rho(\xi)$ , one has

$$\begin{aligned} \frac{v}{(s_{in} - \bar{s})} \sum_{j=1}^n \mu_j(\bar{s}) x_j &\leq M \frac{v}{(s_{in} - \bar{s})} \sum_{j=1}^n x_j = M \left( \frac{\rho(\xi)}{s_{in} - \bar{s}} + v \right) \\ &\leq M \left( \frac{\rho(\xi)}{s_{in} - \tilde{z}(\xi)} + v_{\max} \right). \end{aligned}$$

The synthesis of the  $SA(\bar{s})$  strategy is depicted on Figures 3.1 and 3.2, depending on the position of  $\bar{s}$  relatively to  $s_{out}$ .

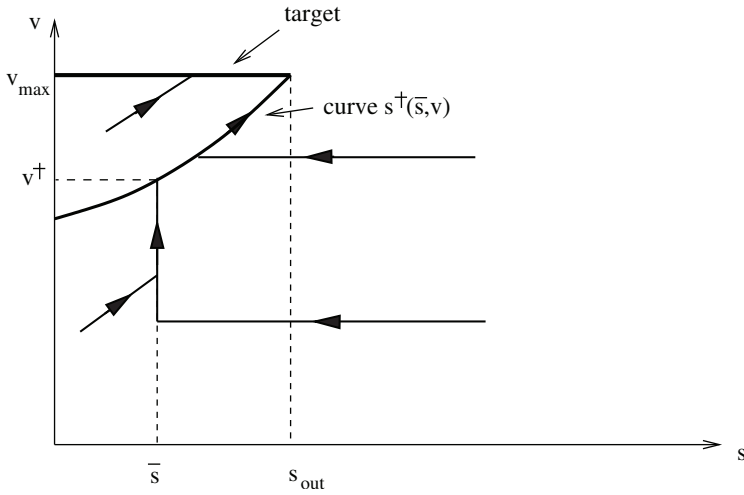


FIG. 3.2. The  $SA(\bar{s})$  synthesis when  $\bar{s} < s_{out}$ .

*Remark 4.* An impulse of volume  $\delta w$  at time  $\tau$  can be achieved by any control law  $u(\cdot)$  such that there exists  $\delta\tau > 0$  satisfying

$$\int_{\tau}^{\tau+\delta\tau} u(\theta)d\theta = \delta w,$$

with  $r(\theta) = 0$  for  $\theta \in [\tau, \tau + \delta\tau]$ . For the sake of simplicity, we shall systematically take  $u(\theta) = u_{\max}$  for  $\theta \in [\tau, \tau + \delta w/u_{\max}]$ .

**4. The cost of the one-impulse strategies.** We consider a family of functions  $\varphi_c(\cdot)$  defined on  $(\mathbb{R}_+^n \setminus \{0\}) \times \mathbb{R}_+$  and parameterized by  $c > 0$ :

$$(4.1) \quad \varphi_c(y, z) = \inf \left\{ t - t_0 \mid s^{t_0, y, z}(t) \leq c \right\},$$

where  $s^{t_0, y, z}(\cdot)$  is the solution of the free dynamics (3.3). A standard analysis of minimal time problems shows that  $\varphi_c(\cdot)$  are Lipschitz-continuous functions and solutions, in the viscosity sense, of the partial differential equation (see, for instance, [2])

$$(4.2) \quad \sum_{j=1}^n (\partial_{y_j} \varphi_c(y, z) - \partial_z \varphi_c(y, z)) \mu_j(z) y_j + 1 = 0$$

on the domain  $(\mathbb{R}_+^n \setminus \{0\}) \times (c, +\infty)$ , with boundary conditions

$$(4.3) \quad \varphi_c(\cdot, z) = 0 \quad \forall z \in (0, c].$$

The time cost of the IOI strategy can then be simply written in terms of the above functions as follows:

$$(4.4) \quad T_{IOI}(\xi) = \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)),$$

where  $\tilde{y}(\xi)$  and  $\tilde{z}(\xi)$  are given by (3.1).

*Remark 5.* Observe that the suboptimal IOI strategy has a finite time cost  $T_{IOI}(\xi)$  for a any initial condition  $\xi$  in the domain  $\mathcal{D}$ . Consequently, the optimal value  $V(\xi)$  is finite for any  $\xi$  in  $\mathcal{D}$ .

**5. The Hamilton–Jacobi characterization.** Let us define the *Hamiltonian* as the mapping  $H : \mathbb{R}_+^n \times [0, s_{in}] \times [0, v_{\max}] \times [0, u_{\max}] \times [0, 1] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  given by

$$(5.1) \quad H(x, s, v, u, r, p, k, q) = r + qu + k \left( \frac{u}{v} (s_{in} - s) - r \sum_{j=1}^n x_j \mu_j(s) \right) + \sum_{j=1}^n p_j x_j \left( r \mu_j(s) - \frac{u}{v} \right).$$

The Hamilton–Jacobi–Bellman equation associated to minimal time problem with dynamics (2.6) is

$$(5.2) \quad \min_{u \in [0, u_{\max}]} \min_{r \in \{0, 1\}} H(y, z, w, u, r, \partial_y V(\xi), \partial_z V(\xi), \partial_w V(\xi)) = 0,$$

or equivalently

$$(5.3) \quad \min \left( 0, \sum_{j=1}^n (\partial_{y_j} V(\xi) - \partial_z V(\xi)) \mu_j(z) y_j + 1 \right) + \frac{u_{\max}}{w} \min \left( 0, - \sum_{j=1}^n \partial_{y_j} V(\xi) y_j + \partial_z V(\xi) (s_{in} - z) + \partial_w V(\xi) w \right) = 0,$$

for any  $\xi \in \mathcal{D}$ , with the boundary condition

$$(5.4) \quad V(\xi) = 0 \quad \forall \xi \in \mathcal{T}.$$

*Remark 6.* Notice that the control variable  $r(\cdot)$  does not take values in a convex set. This does not a priori guarantee the existence of an admissible optimal trajectory in  $\mathcal{C}$  defined in (2.8). In the following, we prove the existence of optimal trajectories by exhibiting particular strategies for which  $r(\cdot)$  takes values 0 or 1.

It is straightforward to check that (5.3) is equivalent to a system of two partial differential inequalities:

$$(5.5) \quad \Delta_r V(\xi) = \sum_{j=1}^n (\partial_{y_j} V(\xi) - \partial_z V(\xi)) \mu_j(z) y_j + 1 \geq 0,$$

$$(5.6) \quad \Delta_u V(\xi) = - \sum_{j=1}^n \partial_{y_j} V(\xi) y_j + \partial_z V(\xi) (s_{in} - z) + \partial_w V(\xi) w \geq 0,$$

independently of the upper bound  $u_{\max}$ .

*Remark 7.* Note that the situation when  $\Delta_r V$  and  $\Delta_u V$  are both strictly positive corresponds to controls  $u = 0$  and  $r = 0$ . Since we consider only controls in  $\mathcal{C}$  (defined in (2.8)), we obtain that inequalities (5.5)–(5.6) are equivalent to

$$\min\{\Delta_r V(\xi), \Delta_u V(\xi)\} = 0.$$

As it is well known from the theory of first order Hamilton–Jacobi partial differential equation (p.d.e.) [2], the differentiability of the value function is not guaranteed. Furthermore, in the present case, the uniqueness of the solution of system (5.4)–(5.5)–(5.6) among smooth functions is not guaranteed either (one can easily check that  $V \equiv 0$  is always a solution). Nevertheless, one has the following result, providing sufficient conditions for the existence of an optimal trajectory in  $\mathcal{C}$  and the smoothness of the value function.

PROPOSITION 5.1. *If there exist*

- (a) *a nonnegative continuous function  $V(\cdot)$  that fulfills the boundary condition (5.4) and such that at any  $\xi \in \mathcal{D}$ , with  $V(\xi) > 0$ ,  $V(\cdot)$  is  $C^1$  and fulfills the partial differential inequalities (5.5) and (5.6);*
- (b) *two maps  $\xi \mapsto u^*(\xi) \geq 0$ ,  $\xi \mapsto r^*(\xi) \in \{0, 1\}$ , with*

$$H(\xi, u^*(\xi), r^*(\xi), \nabla V(\xi)) = 0, \quad \forall \xi \in \mathcal{D} \text{ such that (s.t.) } V(\xi) > 0,$$

$$r^*(\xi) = 0, \quad \forall \xi \in \mathcal{D} \text{ s.t. } V(\xi) = 0,$$

*and such that the closed-loop dynamics*

$$(5.7) \quad \begin{cases} \frac{dx_i}{d\tau} = r^*(X(\tau))\mu_i(s)x_i - \frac{u^*(X(\tau))}{v}x_i & (i = 1 \cdots n), \\ \frac{ds}{d\tau} = -r^*(X(\tau)) \sum_{j=1}^n \mu_j(s)x_j + \frac{u^*(X(\tau))}{v}(s_{in} - s), \\ \frac{dv}{d\tau} = u^*(X(\tau)), \end{cases}$$

*admits an absolutely continuous solution  $X(\cdot) = (x(\cdot), s(\cdot), v(\cdot))$  that reaches the target in finite time, for any initial condition  $X(t_0) = \xi \in \mathcal{D}$ , then  $V(\xi)$  is the value function (2.9) at any  $\xi \in \mathcal{D}$  such that the solution of (5.7) fulfills  $u^*(X(\tau)) \leq u_{\max}$  for any  $\tau \geq t_0$  such that  $X(\theta) \notin \mathcal{T}$  whatever is  $\theta \in [t_0, \tau)$ .*

*Proof.* Fix an initial condition  $\xi \in \mathcal{D}$  at time  $t_0$ , and consider admissible controls  $(r(\cdot), u(\cdot)) \in \mathcal{C}$  such that the trajectory  $X(\cdot) = (x(\cdot), s(\cdot), v(\cdot))$  solution of system (2.6) reaches the target in finite time, say, at time  $\tau_c$ . Define then the function

$$\mathcal{V}(\tau) = V(X(\tau)),$$

and consider the set  $\mathcal{N} = \{\tau \in [t_0, \tau_c] \mid \mathcal{V}(\tau) > 0\}$ . Clearly,  $\mathcal{V}(\cdot)$  is absolutely continuous on  $\mathcal{N}$  and one has

$$\begin{aligned} \mathcal{V}(\tau_c) - \mathcal{V}(t_0) &= \int_{\mathcal{N}} \mathcal{V}'(\tau) d\tau \\ &= \int_{\mathcal{N}} H(x(\tau), s(\tau), v(\tau), u(\tau), r(\tau), \partial_y V(X(\tau)), \partial_z V(X(\tau)), \partial_w V(X(\tau)) - r(\tau) d\tau. \end{aligned}$$

From (5.2), one deduces the inequality

$$\mathcal{V}(\tau_c) - \mathcal{V}(t_0) \geq - \int_{\mathcal{N}} r(\tau) d\tau,$$

and with the boundary condition (5.4),

$$V(\xi) = \mathcal{V}(t_0) \leq \int_{\mathcal{N}} r(\tau) d\tau \leq \int_{t_0}^{\tau_c} r(\tau) d\tau.$$

This last inequality being valid for any admissible controls  $(u(\cdot), r(\cdot))$ , one deduces that

$$\inf_{u(\cdot), r(\cdot)} \left\{ \int_{t_0}^{\tau} r(\theta) d\theta \mid s^{t_0, \xi, u, r}(\tau) \leq s_{out}, v^{t_0, \xi, u, r}(\tau) = v_{\max} \right\} \geq V(\xi).$$



Consider now the trajectory  $X^*(\cdot)$  solution of (5.7) that reaches the target at time  $\tau_c^*$ . The function  $\mathcal{V}^*(\cdot) = V(X^*(\cdot))$  and the set  $\mathcal{N}^* = \{\tau \in [t_0, \tau_c^*] \mid \mathcal{V}^*(\tau) > 0\}$  verify that

$$\mathcal{V}^*(\tau_c^*) = \mathcal{V}^*(t_0) - \int_{\mathcal{N}^*} r^*(\tau) d\tau = \mathcal{V}^*(t_0) - \int_{t_0}^{\tau_c^*} r^*(\tau) d\tau.$$

We finally obtain

$$\begin{aligned} V(\xi) &= \mathcal{V}^*(t_0) = \int_{t_0}^{\tau_c^*} r^*(\tau) d\tau \\ &\geq \inf_{u(\cdot), r(\cdot)} \left\{ \int_{t_0}^{\tau} r(\theta) d\theta \mid s^{t_0, \xi, u, r}(\tau) \leq s_{out}, v^{t_0, \xi, u, r}(\tau) = v_{max} \right\}, \end{aligned}$$

which proves that  $V(\cdot)$  is the value function.  $\square$

*Remark 8.* For a function  $V(\cdot)$  that fulfills condition (a) of Proposition 5.1 independently of  $u_{max}$ , the existence of a pair of admissible feedbacks  $u^*(\cdot), r^*(\cdot)$  that leads to the target is related to the value of  $u_{max}$ .

On the other hand, the above result establish that there exists at most one function  $V$  (the value function of our problem) satisfying the conditions of Proposition 5.1. Nevertheless, one cannot conclude the uniqueness of the optimal trajectory.

*Remark 9.* One can easily check that the function  $V \equiv 0$  is a  $C^1$  solution of the Hamilton–Jacobi–Bellman equation (5.3) that fulfills the boundary condition (5.4). But (5.2) imposes to have  $r \equiv 0$ . Clearly, such controls do not allow one to reach the target, and the conditions of Proposition 5.1 are not fulfilled.

For technicalities, we shall assume in the following that functions  $\varphi_c(\cdot)$  defined in (4.1) possess some regularity.

*Assumption A0.* For any  $c > 0$ , the function  $\varphi_c(\cdot)$  is  $C^1$  on  $(\mathbb{R}_+^n \setminus \{0\}) \times (c, +\infty)$ .

**LEMMA 5.2.** *Under Assumption A0, at any  $\xi \in \mathcal{D}$  such that  $T_{IOI}(\xi) > 0$ , one has*

$$(5.8) \quad \Delta_r T_{IOI}(\xi) = \sum_{j=1}^n (\partial_{y_j} \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) - \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi))) \tilde{y}_j (\mu_j(z) - \mu_j(\tilde{z}(\xi))).$$

*Proof.* At  $\xi = (y, z, w) \in \mathcal{D}$  such that  $T_{IOI}(\xi) > 0$ , Assumption A0 guarantees that  $T_{IOI}(\cdot)$  is  $C^1$ . Let us write its partial derivatives as follows:

$$\begin{cases} \partial_{y_j} T_{IOI}(\xi) = \frac{w}{v_{max}} \partial_{y_j} \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) & (j = 1 \cdots n), \\ \partial_z T_{IOI}(\xi) = \frac{w}{v_{max}} \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)), \\ \partial_w T_{IOI}(\xi) = \sum_j \frac{y_j}{v_{max}} \partial_{y_j} \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) - \frac{s_{in} - z}{v_{max}} \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)). \end{cases}$$

Then, one has

$$(5.9) \quad \Delta_r T_{IOI}(\xi) = \sum_j \left( \partial_{y_j} \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) - \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \right) \mu_j(z) \tilde{y}_j(\xi) + 1.$$

Equation (4.2) with  $c = s_{out}$  at  $(\tilde{y}(\xi), \tilde{z}(\xi))$  provides the equality

$$(5.10) \quad \sum_j \left( \partial_{y_j} \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) - \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \right) \mu_j(\tilde{z}(\xi)) \tilde{y}_j(\xi) = -1.$$

Combining (5.9) and (5.10) gives, finally,

$$\begin{aligned} \Delta_r T_{IOI}(\xi) &= \sum_j \left( \partial_{y_j} \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \right. \\ &\quad \left. - \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \right) \tilde{y}_j(\xi) (\mu_j(z) - \mu_j(\tilde{z}(\xi))). \quad \square \end{aligned}$$

We then obtain the following result concerning the optimality of the IOI strategy for any initial condition.

PROPOSITION 5.3. *Under Assumption A0, the IOI strategy is optimal for any  $\xi \in \mathcal{D}$  if and only if*

$$(5.11) \quad \Delta_r T_{IOI}(\xi) \geq 0 \quad \forall \xi \in \mathcal{D} \text{ s.t. } T_{IOI}(\xi) > 0.$$

*Proof.* We proceed to show that the function  $T_{IOI}(\cdot)$  fulfills conditions of Proposition 5.1.

If  $\xi \in \mathcal{T}$ , one has  $T_{IOI}(\xi) = 0$ , thus boundary condition (5.4) is fulfilled. At  $\xi \in \mathcal{D}$  such that  $T_{IOI}(\xi) > 0$ ,  $T_{IOI}(\cdot)$  is  $C^1$  under assumption A0.

Notice that condition (5.11) is exactly the first partial differential inequality (5.5). The verification of the second partial differential inequality (5.6) is easy:

$$\begin{aligned} \Delta_u T_{IOI}(\xi) &= - \sum_j \partial_{y_j} \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \tilde{y}_j + \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \frac{w}{v_{\max}} (s_{in} - z) \\ &\quad + \sum_j \partial_{y_j} \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \tilde{y}_j - \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \frac{w}{v_{\max}} (s_{in} - z) = 0. \end{aligned}$$

So, condition (a) of Proposition 5.1 is fulfilled.

Finally, the IOI strategy, as defined in section 3, straightforwardly fulfills condition (b) of Proposition 5.1.  $\square$

Let us now consider the function

$$\psi(\xi, c) = \varphi_c(y, z) + T_{IOI}(x_c, c, w), \quad \xi \in \mathcal{D}, \quad c \in (0, z),$$

where  $x_c = x^{t_0, y, z}(t_c)$  such that  $s^{t_0, y, z}(t_c) = c$ , with  $(x^{t_0, y, z}(\cdot), s^{t_0, y, z}(\cdot))$  solution of the free dynamics (3.3). Concerning the optimality of the IOI strategy, the study of the function  $\psi(\cdot)$  allows us to show that condition (5.11) is also necessary for a given initial condition  $\xi \in \mathcal{D}$  such that  $T_{IOI}(\xi) > 0$ . For this purpose, the next technical lemma will be useful.

LEMMA 5.4. *Under Assumption A0, one has*

$$\partial_c \psi(\xi, z) = - \frac{\Delta_r T_{IOI}(\xi)}{n}, \quad \xi \in \mathcal{D} \text{ s.t. } T_{IOI}(\xi) > 0.$$

$$\sum_{j=1} \mu_j(z) y_j$$

*Proof.* From (3.3), one has

$$\frac{\partial x_i(t_c)}{\partial c} = -\frac{\mu_i(c)x_i(t_c)}{\sum_{j=1}^n \mu_j(c)x_j(t_c)}, \quad \partial_c \varphi_c(y, z) = -\frac{1}{\sum_{j=1}^n \mu_j(c)x_j(t_c)}.$$

Then, one can write

$$\begin{aligned} \partial_c \psi(\xi, z) &= [\partial_c \varphi_c(y, z)]_{c=z} \\ &+ \left[ \sum_{j=1}^n \partial_{y_j} T_{IOI}(x(t_c), c, w) \frac{\partial x_j(t_c)}{\partial c} + \partial_z T_{IOI}(x(t_c), c, w) \right]_{c=z} \\ &= -\frac{1}{\sum_{j=1}^n \mu_j(z)y_j} - \frac{\frac{w}{v_{\max}} \sum_{j=1}^n \partial_{y_j} \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \mu_j(z)y_j}{\sum_{j=1}^n \mu_j(z)y_j} \\ &+ \frac{w}{v_{\max}} \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \\ &+ \frac{1 + \sum_{j=1}^n (\partial_{y_j} \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) - \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi))) \mu_j(z) \tilde{y}_j(\xi)}{\sum_{j=1}^n \mu_j(z)y_j}. \end{aligned}$$

Using the property (4.2) for  $c = s_{out}$  at  $(\tilde{y}(\xi), \tilde{z}(\xi))$ , and the expression (5.8) given by Lemma 5.2, finally gives

$$\partial_c \psi(\xi, c) = -\frac{\Delta_r T_{IOI}(\xi)}{\sum_{j=1}^n \mu_j(z)y_j}. \quad \square$$

Proposition 5.3 states that if (5.11) is not satisfied, then there exists an initial condition  $\xi \in \mathcal{D}$  for which the IOI strategy is not optimal. The following proposition characterizes some initial conditions for when this occurs.

**PROPOSITION 5.5.** *At states  $\xi \in \mathcal{D}$  such that  $T_{IOI}(\xi) > 0$  and  $\Delta_r T_{IOI}(\xi) < 0$ , the IOI strategy cannot be optimal.*

*Proof.* When  $T_{IOI}(\xi) > 0$  and  $\Delta_r T_{IOI}(\xi) < 0$  at  $\xi \in \mathcal{D}$ , Lemma 5.4 gives the existence of  $c^* < z$  such that  $\psi(\xi, c^*) < \psi(\xi, z) = T_{IOI}(\xi)$ . Consequently, there is a strategy (consisting in applying a null control until the time  $t_{c^*}$  such that  $s(t_{c^*}) = c^*$  and then the IOI strategy) which has a better cost than the IOI strategy.  $\square$

**6. Derivation from the minimum principle.** In this section we apply the Pontryagin's minimum principle (PMP) (see [4, 25]) to the minimal time problem with dynamics (2.6).

The PMP states that when  $(x, s, v, u, r)(\cdot)$  is a solution of the minimal time problem associated to the system (2.6), then there exists an  $n$ -dimensional *multiplier*  $p(\cdot)$

and scalar multipliers  $q(\cdot)$  and  $k(\cdot)$  such that

$$(6.1) \quad \begin{cases} \frac{dp_i}{d\tau} = p_i u/v - r(p_i - k)\mu_i(s), & p_i(T) = 0, \\ \frac{dk}{d\tau} = -r \sum_{j=1}^n (p_j - k)x_j \mu'_j(s) + ku/v, & k(T) = 1, \\ \frac{dq}{d\tau} = \frac{u}{v^2} \left( k(s_{in} - s) - \sum_{j=1}^n p_j x_j \right), \end{cases}$$

where  $T$  is the optimal terminal time. In addition, the Hamiltonian (defined in (5.1))

$$(u, r) \longrightarrow H(x(\tau), s(\tau), v(\tau), u, r, p(\tau), k(\tau), q(\tau))$$

is minimized in  $u(\tau)$  and  $r(\tau)$ , at any  $\tau \in [t_0, T]$ .

Define the auxiliary variables  $\tilde{p}_i = p_i - k$ , who play an important role in what follows.

First, we observe that the dynamics of  $\tilde{p} = (\tilde{p}_1, \dots, \tilde{p}_n)$  can be written as follows:

$$(6.2) \quad \frac{d\tilde{p}}{d\tau} = A(\tau)\tilde{p}, \quad \tilde{p}_i(T) = -1,$$

where  $A(\tau)$  is an  $n \times n$  time dependent matrix. Consequently, one has  $\tilde{p}(\tau) \neq 0$  for any  $\tau \in [t_0, T]$ .

On another hand, one has

$$(6.3) \quad \operatorname{argmin}_{u,r} H(x, s, v, u, r, p, k, q) = \operatorname{argmin}_{u,r} u\phi_u(x, s, v, p, k, q) + r\phi_r(x, s, p, k),$$

where

$$(6.4) \quad \begin{cases} \phi_u(x, s, v, p, k, q) = q + \frac{k}{v}(s_{in} - s) - \frac{1}{v} \sum_{j=1}^n p_j x_j, \\ \phi_r(x, s, p, k) = 1 + \sum_{j=1}^n (p_j - k)\mu_j(s)x_j = 1 + \sum_{j=1}^n \tilde{p}_j \mu_j(s)x_j. \end{cases}$$

If we derive with respect to the fictitious time  $\tau$ , we obtain

$$(6.5) \quad \begin{cases} \frac{d\phi_u}{d\tau} = -r \frac{(s_{in} - s)}{v} \langle \tilde{p}, m \rangle, \\ \frac{d\phi_r}{d\tau} = u \frac{(s_{in} - s)}{v} \langle \tilde{p}, m \rangle, \end{cases}$$

where  $m = m(\tau)$  is given by

$$(6.6) \quad m(\tau) = \begin{pmatrix} \mu'_1(s(\tau))x_1(\tau) \\ \vdots \\ \mu'_n(s(\tau))x_n(\tau) \end{pmatrix}$$

and  $\langle \cdot, \cdot \rangle$  is the Euclidean inner product in  $\mathbb{R}^n$ . Finally, it is straightforward to check that (see (6.2))

$$(6.7) \quad \frac{d}{d\tau} \langle \tilde{p}, m \rangle = \left\langle \tilde{p}, A^\top m + \frac{dm}{d\tau} \right\rangle.$$

The next result links the optimal value function  $V(\cdot)$  to Pontryagin's multipliers  $(p, q, k)(\cdot)$ .

**PROPOSITION 6.1.** *Let  $V(\cdot)$  be the optimal value function defined in (2.2). Consider an initial vector  $\xi = (y, z, w) \in \mathcal{D}$  and denote by  $(p, q, k)(\cdot)$  the corresponding Pontryagin's multipliers. If the value function  $V$  is  $C^1$  at  $\xi$ , it holds that*

$$(6.8) \quad \Delta_r V(\xi) = \phi_r(x, s, p, k) \quad \text{and} \quad \Delta_u V(\xi) = v\phi_u(x, s, p, k, q),$$

where  $\Delta_r V(\xi)$  and  $\Delta_u V(\xi)$  are defined by (5.5) and (5.6), respectively.

*Proof.* See Theorem 12.5.1 in [31].  $\square$

We end this section by introducing our second assumption and a lemma whose proof is direct.

**Assumption A1.** The functions  $\mu_i(\cdot)$  are nondecreasing.

**LEMMA 6.2.** *Under Assumption A1, the following assertions hold:*

- i. *The matrix  $A(\tau)$  has nonnegative off-diagonal terms, i.e., the dynamical system (6.2) is cooperative (see [28]);*
- ii. *The vector  $m(\cdot)$ , defined in (6.6), lies in  $\mathbb{R}_+^n$ .*

**7. The one-species case.** For this case, it is straightforward to check that, for any  $c > 0$ , the partial differential equation (4.2) with boundary condition (4.3) admits the (unique) nonnegative continuous solution that is  $C^1$  on  $(\mathbb{R}_+ \setminus \{0\}) \times (c, +\infty)$ , given by the expression

$$(7.1) \quad \varphi_c(y, z) = \begin{cases} \int_c^z \frac{d\zeta}{\mu(\zeta)(y + z - \zeta)} & \text{for } z > c, \\ 0 & \text{for } z \leq c. \end{cases}$$

Hence, Assumption A0 is fulfilled. We give a technical lemma that will be useful in the following.

**LEMMA 7.1.** *Let  $z > c > 0$ . If  $\mu(\cdot)$  is nonincreasing on  $[c, z]$ , then the following inequalities are satisfied:*

$$(7.2) \quad \partial_z \varphi_c(y, z) \geq \frac{1}{\mu(z)(y + z - c)},$$

$$(7.3) \quad \partial_z \varphi_c(y, z) \leq \frac{1}{\mu(c)(y + z - c)} + \frac{1}{\mu(z)y} - \frac{1}{\mu(c)y}.$$

*Proof.* Notice first that the partial derivative of function  $\varphi_c$  given in (7.1) verifies

$$\partial_z \varphi_c(y, z) = \frac{1}{\mu(z)y} - \int_c^z \frac{d\zeta}{\mu(\zeta)(y + z - \zeta)^2}.$$

Since  $\mu(\cdot)$  is nonincreasing, one can write

$$\partial_z \varphi_c(y, z) \geq \frac{1}{\mu(z)y} - \int_c^z \frac{d\zeta}{\mu(z)(y + z - \zeta)^2} = \frac{1}{\mu(z)(y + z - c)},$$

and

$$\begin{aligned} \partial_z \varphi_c(y, z) &\leq \frac{1}{\mu(z)y} - \int_c^z \frac{d\zeta}{\mu(c)(y + z - \zeta)^2} \\ &= \frac{1}{\mu(z)y} + \frac{1}{\mu(c)(y + z - c)} - \frac{1}{\mu(c)y}. \quad \square \end{aligned}$$

### 7.1. Increasing growth functions.

PROPOSITION 7.2. *Under Assumption A1, the IOI strategy is optimal for any initial condition  $\xi$  in  $\mathcal{D}$ , and the value function is*

$$(7.4) \quad V(\xi) = \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)),$$

where  $\varphi_{s_{out}}(\cdot)$  is given by (7.1) and  $(\tilde{y}(\xi), \tilde{z}(\xi))$  by (3.1).

*Proof.* For  $(\tilde{y}(\xi), \tilde{z}(\xi))$  such that  $\tilde{z}(\xi) > s_{out}$ , the map  $\xi \mapsto \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi))$  is  $C^1$  at  $\xi$ , and from (4.2), one has

$$(\partial_y \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) - \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)))\mu(\tilde{z}(\xi))\tilde{y}(\xi) = -1.$$

Then, condition (5.11) of Proposition 5.3 simply becomes  $\mu(\tilde{z}(\xi)) - \mu(z) \geq 0$ , which is fulfilled when  $\mu(\cdot)$  is nondecreasing and  $\tilde{z} \geq z$ .  $\square$

*Remark 10.* This proposition extends to impulse controls a result obtained by Moreno [20] for measurable control (with a different technique based on Green's theorem). It states that, for a monotonic growth functions  $\mu$ , the *one bang* control

$$F = \begin{cases} F_{\max} & \text{if } v < v_{\max}, \\ 0 & \text{if } v = v_{\max}, \end{cases}$$

is optimal.

It is clear that there is no uniqueness of optimal control laws (see Remarks 2 and 4). Eventually, we could have another control  $(u(\cdot), r(\cdot))$  that implies a strategy different to the IOI one, with the same value function defined in (7.4). The following proposition shows that the IOI strategy is, in fact, the unique admissible optimal one. This result is obtained using PMP.

PROPOSITION 7.3. *Under Assumption A1, one has that, for any initial condition  $\xi$  in  $\mathcal{D}$ , the IOI strategy is the unique optimal control law.*

*Proof.* From the PMP, there exist multipliers  $p$  and  $k$  solutions of system (6.1), which, in this case ( $n = 1$ ), satisfies  $\langle \tilde{p}, m \rangle = \tilde{p}\mu'(s)x < 0$  for  $m$  defined by (6.6) and  $\tilde{p} = p - k$ . Then, from (6.5), one has that  $\frac{d\phi_r}{d\tau} \leq 0$  and  $\frac{d\phi_u}{d\tau} \geq 0$  along all of the optimal trajectories. Also, relations (6.8), (5.5), and (5.6) imply that  $\phi_u \geq 0$  and  $\phi_r \geq 0$ .

Since the admissible controls  $(u, r)$  are in  $\mathcal{C}$  defined by (2.8) (see Remark 2) and the states of the system (2.6) must reach the target  $\mathcal{T}$ , the only possibilities for  $\phi_u$  and  $\phi_r$  are as follows:

- i.  $\phi_u = 0$  at the beginning and then  $\phi_u > 0$ ;
- ii.  $\phi_r > 0$  at the beginning and then  $\phi_r = 0$ .

Indeed, the admissible control set  $\mathcal{C}$  allows us to consider only configurations such that  $\phi_u \cdot \phi_r = 0$ , and  $u \neq 0$  or  $r \neq 0$ . This, together with (6.5), discards the choices  $\phi_u > 0$  or  $\phi_r = 0$ , at the beginning.

On the other hand, (6.5) implies also that  $u \neq 0$  until the time when  $\phi_u$  switches from  $\phi_u = 0$  to  $\phi_u > 0$ . This time coincides with the time when  $v = v_{\max}$  and, consequently, also with the time when  $\phi_r$  switches from  $\phi_r > 0$  to  $\phi_r = 0$  (because  $u$  becomes null at such time). Therefore, one has the following:

- i.  $u = u_{\max}$  until  $v$  reaches  $v_{\max}$  and then  $u = 0$ ;
- ii.  $r = 0$  until  $v = v_{\max}$  and then  $u = 0$  and  $r = 1$  until the concentration  $s$  reaches  $s_{out}$ .

This proves the optimality of the IOI strategy, and, by construction, we have uniqueness.  $\square$

**7.2. Nonmonotonic growth functions with one maximum.** In this section we consider a continuously differentiable growth function  $\mu(\cdot)$ , which is nonmonotonic and attains a unique isolated maximum point  $s^* \in (0, s_{in})$ . More precisely, this growth function satisfies  $\mu'(s) > 0$  for all  $s \in [0, s^*)$ ,  $\mu'(s) < 0$  for all  $s > s^*$ , and  $\mu'(s^*) = 0$ .

One instance of such functions is typically the Haldane law, given by the expression

$$\mu(s) = \frac{\bar{\mu}s}{K + s + s^2/R},$$

where  $K$  is the affinity constant and  $R$  is the inhibition constant. This kind of growth function occurs in bioprocesses where the substrate is a toxic substance and, for big concentrations, inhibits the activity of the microorganisms [29].

The following proposition solves our minimal time problem for this type of growth function. This solution has been previously obtained in [20] for the class of measurable and bounded controls, and under the assumption  $s^* > s_{out}$ . Furthermore, we give the expression of the value function  $V(\cdot)$ . For convenience, we define the number

$$s^\diamond = \max(s^*, s_{out}).$$

**PROPOSITION 7.4.** *For any initial condition  $\xi = (y, z, w) \in \mathcal{D}$  that satisfies (3.5), the  $SA(s^*)$  strategy (defined in section 3) is optimal. Furthermore, the value function at  $\xi$  is given by the expression*

$$(7.5) \quad V(\xi) = \begin{cases} \varphi_{s^*}(y, z) + \gamma(\xi) + \varphi_{s_{out}}(\tilde{y}(\xi) + \tilde{z}(\xi) - s^\diamond, s^\diamond), & \text{for } z > s^* \geq s^\dagger(s^*, w), \\ \gamma\left(y \frac{s_{in} - s^*}{s_{in} - z}, s^*, w \frac{s_{in} - z}{s_{in} - s^*}\right) + \varphi_{s_{out}}(\tilde{y}(\xi) + \tilde{z}(\xi) - s^\diamond, s^\diamond), & \text{for } z \leq s^* \text{ and } z > s^\dagger(s^*, w), \\ \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)), & \text{for } z \leq s^\dagger(s^*, w), \\ \varphi_{s^\dagger(s^*, w)}(y, z) & \text{for } z > s^\dagger(s^*, w) > s^*, \end{cases}$$

where  $\varphi(\cdot)$  is given by (7.1),  $\tilde{y}(\cdot)$  and  $\tilde{z}(\cdot)$  by (3.1),  $s^\dagger(\cdot)$  by (3.2), and

$$(7.6) \quad \gamma(\xi) = \frac{1}{\mu(s^*)} \log \left( \frac{w(y + z - s_{in}) + v_{\max}(s_{in} - s^\diamond)}{w(y + z - s^*)} \right).$$

*Proof.* First, observe that if the initial condition  $\xi$  is in the target, i.e.,  $z \leq s_{out}$  and  $w = v_{\max}$ , then  $s^\dagger(s^*, w) = s^\diamond \geq s_{out} \geq z$  and  $\tilde{z}(\xi) = z$  and, therefore it corresponds to the third case in the definition of  $V$  obtaining  $\varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) = 0$ , and hence the boundary condition (5.4) is satisfied. We prove now that  $V(\cdot)$  is  $C^1$  and fulfills the Hamilton–Jacobi inequalities (5.5) and (5.6), concluding the result from Proposition 5.1 because one can easily check that  $V(\xi)$  is the cost associated to the  $SA(s^*)$  strategy from initial condition  $\xi$ .

For  $z > s^*$  and  $s^* \geq s^\dagger(s^*, w)$ ,  $V(\cdot)$  is  $C^1$  and its partial derivatives are

$$\begin{aligned}
 (7.7) \quad \partial_y V(\xi) &= \partial_y \varphi_{s^*}(y, z) + \frac{1}{\mu(s^*)} \left( \frac{w}{w(y+z-s_{in}) + (s_{in} - s^\diamond)v_{\max}} - \frac{1}{y+z-s^*} \right) \\
 &\quad + \partial_y \varphi_{s_{out}}(\tilde{y}(\xi) + \tilde{z}(\xi) - s^\diamond, s^\diamond) \frac{w}{v_{\max}}, \\
 \partial_z V(\xi) &= \partial_z \varphi_{s^*}(y, z) + \frac{1}{\mu(s^*)} \left( \frac{w}{w(y+z-s_{in}) + (s_{in} - s^\diamond)v_{\max}} - \frac{1}{y+z-s^*} \right) \\
 &\quad + \partial_y \varphi_{s_{out}}(\tilde{y}(\xi) + \tilde{z}(\xi) - s^\diamond, s^\diamond) \frac{w}{v_{\max}}, \\
 \partial_w V(\xi) &= \frac{1}{\mu(s^*)} \left( \frac{y+z-s_{in}}{w(y+z-s_{in}) + (s_{in} - s^\diamond)v_{\max}} - \frac{1}{w} \right) \\
 &\quad + \partial_y \varphi_{s_{out}}(\tilde{y}(\xi) + \tilde{z}(\xi) - s^\diamond, s^\diamond) \frac{y+z-s_{in}}{v_{\max}}.
 \end{aligned}$$

Then, one has straightforwardly

$$\begin{aligned}
 \Delta_r V(\xi) &= (\partial_y \varphi_{s^*}(y, z) - \partial_z \varphi_{s^*}(y, z)) \mu(z)y + 1, \\
 \Delta_u V(\xi) &= -\partial_y \varphi_{s^*}(y, z)y + \partial_z \varphi_{s^*}(y, z)(s_{in} - z) + \frac{1}{\mu(s^*)} \left( \frac{y+z-s_{in}}{y+z-s^*} - 1 \right).
 \end{aligned}$$

Using the property (4.2) fulfilled by the function  $\varphi_{s^*}$ , one obtains

$$\begin{aligned}
 \Delta_r V(\xi) &= 0, \\
 \Delta_u V(\xi) &= (s_{in} - s^*) \left( \partial_z \varphi_{s^*}(y, z) - \frac{1}{\mu(s^*)(y+z-s^*)} \right) \\
 &\quad + (y+z-s^*) \left( \frac{1}{\mu(z)(y+z-s^*)} - \partial_z \varphi_{s^*}(y, z) \right).
 \end{aligned}$$

Consequently, inequality (5.5) is fulfilled. For the second inequality (5.6), we distinguish two cases:

i.  $y+z-s_{in} \leq 0$ . Since  $\mu(z) \leq \mu(s^*)$ , one can write the inequality

$$\Delta_u V(\xi) \geq (y+z-s_{in}) \left( \frac{1}{\mu(s^*)(y+z-s^*)} - \partial_z \varphi_{s^*}(y, z) \right),$$

and with inequality (7.2) given by Lemma 7.1 with  $c = s^*$ , one deduces that

$$\Delta_u V(\xi) \geq \frac{y+z-s_{in}}{y+z-s^*} \left( \frac{1}{\mu(s^*)} - \frac{1}{\mu(z)} \right) \geq 0.$$

ii.  $y+z-s_{in} > 0$ . With inequality (7.3) given by Lemma 7.1 with  $c = s^*$ , one can write

$$\begin{aligned}
 \Delta_u V(\xi) &= (s_{in} - y - z) \partial_z \varphi_{s^*}(y, z) + \frac{1}{\mu(z)} - \frac{s_{in} - s^*}{\mu(s^*)(y+z-s^*)} \\
 &\geq (s_{in} - y - z) \left( \frac{1}{\mu(z)y} + \frac{1}{\mu(s^*)(y+z-s^*)} - \frac{1}{\mu(s^*)y} \right) + \frac{1}{\mu(z)} \\
 &\quad - \frac{1}{\mu(s^*)} - \frac{s_{in} - y - z}{\mu(s^*)(y+z-s^*)} \\
 &= \frac{s_{in} - z}{y} \left( \frac{1}{\mu(z)} - \frac{1}{\mu(s^*)} \right) \geq 0.
 \end{aligned}$$



For  $z < s^*$  and  $z > s^\dagger(s^*, w)$ ,  $V(\cdot)$  is  $C^1$  and its partial derivatives are

$$(7.8) \quad \begin{aligned} \partial_y V(\xi) &= \frac{1}{\mu(s^*)} \left( \frac{w}{w(y+z-s_{in}) + (s_{in} - s^\diamond)v_{\max}} - \frac{1}{y} \right) \\ &\quad + \partial_y \varphi_{s_{out}}(\tilde{y}(\xi) + \tilde{z}(\xi) - s^\diamond, s^\diamond) \frac{w}{v_{\max}}, \\ \partial_z V(\xi) &= \frac{1}{\mu(s^*)} \frac{w}{w(y+z-s_{in}) + (s_{in} - s^\diamond)v_{\max}} \\ &\quad + \partial_y \varphi_{s_{out}}(\tilde{y}(\xi) + \tilde{z}(\xi) - s^\diamond, s^\diamond) \frac{w}{v_{\max}}, \\ \partial_w V(\xi) &= \frac{1}{\mu(s^*)} \left( \frac{y+z-s_{in}}{w(y+z-s_{in}) + (s_{in} - s^\diamond)v_{\max}} - \frac{1}{w} \right) \\ &\quad + \partial_y \varphi_{s_{out}}(\tilde{y}(\xi) + \tilde{z}(\xi) - s^\diamond, s^\diamond) \frac{y+z-s_{in}}{v_{\max}}. \end{aligned}$$

One has clearly, from (7.1),

$$\partial_y \varphi_{s^*}(y, s^*) = 0 \quad \text{and} \quad \partial_z \varphi_{s^*}(y, s^*) = \frac{1}{\mu(s^*)y},$$

which implies, from expressions (7.7) and (7.8),

$$\lim_{z \rightarrow s^*+} \nabla V(\xi) = \lim_{z \rightarrow s^*-} \nabla V(\xi).$$

Consequently,  $V(\cdot)$  is  $C^1$  at points  $\xi$  such that  $z = s^*$ . From the expressions (7.8), one can straightforward check the following equalities:

$$\Delta_r V(\xi) = 1 - \frac{\mu(s^*)}{\mu(z)} \geq 0 \quad \text{and} \quad \Delta_u V(\xi) = 0.$$

Consider now points  $\xi$  such that  $z < s^\dagger(s^*, w)$  and  $V(\xi) > 0$ . At such points,  $V(\cdot)$  is  $C^1$  and its partial derivatives are

$$\begin{cases} \partial_y V(\xi) = \partial_y \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \frac{w}{v_{\max}}, \\ \partial_z V(\xi) = \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \frac{w}{v_{\max}}, \\ \partial_w V(\xi) = \partial_y \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \frac{y}{v_{\max}} - \partial_z \varphi_{s_{out}}(\tilde{y}(\xi), \tilde{z}(\xi)) \frac{s_{in} - z}{v_{\max}}. \end{cases}$$

Notice that when  $z = s^\dagger(s^*, w)$ , with  $s^\dagger(s^*, w) > s^*$ , one has  $V(\xi) = 0$ . One can easily check that  $V(\cdot)$  is also  $C^1$  at points  $\xi$  such that  $z = s^\dagger(s^*, w)$  and  $V(\xi) > 0$ , because, at such points, one has  $\tilde{z}(\xi) = s^\diamond = s^*$ , and furthermore, function  $\varphi_{s_{out}}(\cdot)$  fulfills the property (from (4.2))

$$\partial_z \varphi_{s_{out}}(\tilde{y}(\xi), s^*) = \partial_y \varphi_{s_{out}}(\tilde{y}(\xi), s^*) + \frac{1}{\mu(s^*)\tilde{y}(\xi)}.$$

Notice that one has also  $V(\xi) = T_{IOI}(\xi)$ . Since  $\Delta_u T_{IOI}(\xi) = 0$  (see the proof of Proposition 5.3), the function  $V(\cdot)$  is a solution of the Hamilton–Jacobi equation

(5.3) at  $\xi$  when the single condition  $\Delta_r T_{IOI}(\xi) \geq 0$  is fulfilled. Here, this condition simply becomes  $\mu(\tilde{z}(\xi)) - \mu(z) = \mu(s^*) - \mu(z) \geq 0$ , which is fulfilled because  $s^*$  is the maximum of the function  $\mu$ .

Finally, we consider situations for which  $z > s^\dagger(s^*, w) > s^*$ . This case occurs only when  $s^* < s_{out}$ . At such points  $\xi = (y, z, w)$ , the function  $V(\cdot)$  is  $C^1$  and its partial derivatives are

$$\begin{cases} \partial_y V(\xi) = \partial_y \varphi_{s^\dagger(s^*, w)}(y, z), \\ \partial_z V(\xi) = \partial_z \varphi_{s^\dagger(s^*, w)}(y, z), \\ \partial_w V(\xi) = -\frac{(s_{in} - s_{out})v_{\max}}{\mu(s^\dagger(s^*, w))(y + z - s^\dagger(s^*, w))w^2}. \end{cases}$$

One can easily check that the first inequality  $\Delta_r V(\xi) \geq 0$  is fulfilled. For the second one, let us write

$$\begin{aligned} (7.9) \quad \Delta_u V(\xi) &= -\partial_y \varphi_{s^\dagger(s^*, w)}(y, z)y + \partial_z \varphi_{s^\dagger(s^*, w)}(y, z)(s_{in} - z) \\ &\quad - \frac{(s_{in} - s_{out})v_{\max}}{\mu(s^\dagger(s^*, w))(y + z - s^\dagger(s^*, w))w^2} \\ &= \frac{1}{\mu(z)} + \partial_z \varphi_{s^\dagger(s^*, w)}(y, z)(s_{in} - z - y) \\ &\quad - \frac{(s_{in} - s_{out})v_{\max}}{\mu(s^\dagger(s^*, w))(y + z - s^\dagger(s^*, w))w^2} \end{aligned}$$

from the expression (4.2). We distinguish now two cases.

- i.  $s_{in} - z - y \geq 0$ . Expressions (7.9) and (7.2) given by Lemma 7.1 with  $c = s^\dagger(s^*, w)$  give together

$$\Delta_u V(\xi) \geq \frac{s_{in} - s^\dagger(s^*, w)}{y + z - s^\dagger(s^*, w)} \left( \frac{1}{\mu(z)} - \frac{1}{\mu(s^\dagger(s^*, w))} \right) \geq 0.$$

- ii.  $s_{in} - z - y < 0$ . Gathering expressions (7.9) and (7.3) given by Lemma 7.1 with  $c = s^\dagger(s^*, w)$  leads to the following inequality:

$$\Delta_u V(\xi) \geq \frac{s_{in} - z}{y} \left( \frac{1}{\mu(z)} - \frac{1}{\mu(s^\dagger(s^*, w))} \right) \geq 0. \quad \square$$

*Remark 11.* For initial conditions  $\xi = (y, z, w)$ , with  $z \geq s^* > s^\dagger(s^*, w)$ , the value of  $\gamma(\xi)$ , defined by (7.6), represents the time spent on the singular arc  $s = s^*$  from volume  $w$  up to volume  $v^\dagger(\bar{s})$  defined in (3.4). Indeed, at time  $t_1 = \varphi_{s^*}(y, z)$ , one has  $s(t_1) = s^*$ ,  $v(t_1) = w$ , and  $x(t_1) = y + z - s^*$  (see the invariant  $\rho(\xi)$  defined in (2.10)). Then, the suitable control in order to keep  $s = s^*$  is obtained from the equation  $\frac{ds}{d\tau} = 0$ , that is,

$$u = u_s(v) = \frac{\mu(s^*)}{s_{in} - s^*} \rho(\xi) + v\mu(s^*).$$

This implies that the  $v(\cdot)$  is the solution of the ordinary differential equation

$$\frac{dv}{d\tau} = u_s(v) = \frac{\mu(s^*)}{s_{in} - s^*} w(y + z - s_{in}) + \mu(s^*)v$$

up to time  $t_2$  such that  $v(t_2) = v^\dagger$ , that can be solved analytically, leading to  $t_2 - t_1 = \gamma(\xi)$ .

*Remark 12.* Proposition 7.4 extends to impulse controls the result obtained by Moreno [20] for measurable controls, based on a Green's theorem argumentation.

Moreover, notice that when  $s^* < s_{out}$ , the  $SA(s^*)$  strategy imposes a final impulse before reaching the target. Such situations were not considered in [20].

*Remark 13.* Proposition 7.4 establishes that the time associated to the singular arc strategy is the optimal value function, but we could have that there exists another optimal control different to this strategy. Nevertheless, one can prove uniqueness (analogously as in Proposition 7.3) using the PMP. We will not develop this approach here because it is very similar to the one used in Proposition 7.3 and in the proof of Theorem 8.2 below.

**8. The two-species case.** We first consider functions  $\mu_i(\cdot)$  that are  $C^2$  and such that  $\mu'_1/\mu'_2$  is a strictly monotonic function. Without loss of generality, we assume that  $\mu'_1/\mu'_2$  is strictly decreasing, which is equivalent to the following condition.

*Assumption A2.*  $\mu'_1(s)\mu''_2(s) > \mu''_1(s)\mu'_2(s)$  for any  $s \in (0, s_{in}]$ .

*Remark 14.* Assumption A2 is fulfilled for nonproportional Monod growth functions. That is, for growth functions

$$\mu_i(s) = \frac{\mu_{\max,i}s}{K_i + s},$$

Assumption A2 holds when  $K_1 < K_2$ .

**LEMMA 8.1.** *Under Assumption A2, a singular arc<sup>2</sup>  $I$  is characterized by  $\frac{ds}{d\tau} = 0$  on  $I$ .*

*Proof.* Consider  $p$  and  $k$  solutions of PMP-system (6.1) and  $m$  defined in (6.6). Define the auxiliary variable  $\tilde{p} = p - k$ . Recall that the property  $\tilde{p}(\tau) \neq 0$ , for any time  $\tau$ , follows from (6.2).

If  $I$  is a singular arc, it follows that the first derivatives of  $\phi_u$  and  $\phi_r$  are null on  $I$ . Since controls  $u$  and  $r$  are not simultaneously null, it is equivalent to write  $\langle \tilde{p}, m \rangle = 0$  on  $I$  (via (6.5)). Differentiating this last equation w.r.t.  $\tau$  and using expression (6.7), it holds that

$$(8.1) \quad \langle \tilde{p}, m \rangle = 0 \quad \text{and} \quad \left\langle \tilde{p}, A^\top m + \frac{dm}{d\tau} \right\rangle = 0 \quad \text{on } I.$$

Since  $\tilde{p}$  is always nonnull and has dimension 2, equalities (8.1) are satisfied if  $m$  and  $A^\top m + \frac{dm}{d\tau}$  are linearly dependent on  $I$ . We easily verify that, under A2, this is equivalent to  $\frac{ds}{d\tau} = 0$  on  $I$ . Indeed, a simple computation leads to

$$\left\| m \times \left( A^\top m + \frac{dm}{d\tau} \right) \right\| = \left| \frac{ds}{d\tau} \right| x_1 x_2 (\mu'_1(s)\mu''_2(s) - \mu''_1(s)\mu'_2(s)).$$

Vectors  $A^\top m + \frac{dm}{d\tau}$  are linearly dependent exactly when their cross product (in  $\mathbb{R}^3$ ) is null, which holds if and only if  $\frac{ds}{d\tau} = 0$ . We have thus proved that if  $I$  is a singular arc, then  $\frac{ds}{d\tau} = 0$  on  $I$ .

<sup>2</sup>See [4, Part III Chapter 2] for an exact definition. In our case, a singular arc consists of an open interval of time  $I$ , where  $\phi_u = \phi_r = 0$ , and then no information on controls  $u$  and  $r$  can be obtained directly from (6.3).

Reciprocally, suppose that  $\frac{ds}{d\tau} = 0$  on  $I$ . The above arguments imply that  $m$  and  $A^\top m + \frac{dm}{d\tau}$  are linearly dependent on  $I$ , obtaining from (6.7) that  $\frac{d\langle \tilde{p}, m \rangle}{d\tau} = \lambda(\tau) \langle \tilde{p}, m \rangle$  on  $I$ , for a real-valued continuous map  $\lambda : I \rightarrow \mathbb{R}$ .

This is equivalent to saying that  $\langle \tilde{p}, m \rangle = Ce^{\int_0^\tau \lambda(t) dt}$  on  $I$  for a real constant  $C$  and a real-valued continuous map  $\lambda : I \rightarrow \mathbb{R}$ . Moreover, by replacing  $\frac{ds}{d\tau} = 0$  in (2.6), it necessarily requires strictly positive controls  $u$  and  $r$ .

Suppose first that  $C = 0$ . Then (6.5) implies that  $\phi_u$  and  $\phi_r$  are both constant on  $I$ . Since  $u$  and  $r$  are strictly positive, this holds only if  $\phi_u = \phi_r = 0$  on  $I$ , i.e.,  $I$  is a singular arc.

Now, suppose that  $C \neq 0$ . One obtains from (6.5) that  $\phi_u$  and  $\phi_r$  are both strictly monotonic on  $I$ . This, together with inequalities (5.5)–(5.6), implies that  $\phi_u$  and  $\phi_r$  are both strictly positive on  $I$ . But, since  $u$  and  $r$  are also strictly positive, it contradicts (6.3). We thus conclude that  $C = 0$ , and hence  $I$  is a singular arc.  $\square$

Consider now the next assumption on growth functions  $\mu_i(\cdot)$ .

*Assumption A3.* For any  $s_1, s_2 \in [s_{out}, s_{in}]$ , one has

$$(8.2) \quad s_2 \geq s_1 \Rightarrow \mu_2(s_2)\mu_1(s_1) \geq \mu_1(s_2)\mu_2(s_1).$$

*Remark 15.* Assumption A3 is fulfilled for growth functions constant or linear on  $[s_{out}, s_{in}]$ . For Monod functions

$$\mu_i(s) = \frac{\mu_{\max, i} s}{K_i + s},$$

condition (8.2) is exactly fulfilled when  $K_1 \leq K_2$ .

**THEOREM 8.2.** *Assume that Assumptions A0–A1–A2–A3 are fulfilled. Then, for any initial condition in  $\mathcal{D}$  that satisfies (3.5), the optimal solution of the minimal time problem associated to dynamics (2.6) consists in either the IOI strategy or the SA( $s^*$ ) strategy for some  $s^* \in ]s_{out}, s_{in}[$ .*

To prove this theorem, we need the three following technical lemmas.

**LEMMA 8.3.** *Under Assumptions A1 and A3, one has*

$$(8.3) \quad \tilde{p}_1 \geq 0 \text{ and } \phi_r = 0 \Rightarrow \langle \tilde{p}, m \rangle \leq 0.$$

*Proof of Lemma 8.3.* Observe that Assumption A3 implies that  $\mu'_1\mu_2 - \mu_1\mu'_2 \leq 0$ . If  $\phi_r = 0$ , then

$$\tilde{p}_2 = \frac{-(1 + \tilde{p}_1\mu_1x_1)}{\mu_2x_2},$$

and then

$$\langle \tilde{p}, m \rangle = \frac{\tilde{p}_1}{\mu_2}(\mu'_1\mu_2 - \mu_1\mu'_2)x_1 - \frac{\mu'_2}{\mu_2},$$

which proves the desired result.  $\square$

**LEMMA 8.4.** *Under Assumptions A1 and A2, for  $\tilde{p}(\tau) \in E(\tau) = \{\tilde{p} = (\tilde{p}_1, \tilde{p}_2) \mid \langle \tilde{p}, m(\tau) \rangle = 0\}$ , one has*

$$(8.4) \quad \operatorname{sgn} \left( \frac{d}{d\tau} \langle \tilde{p}, m \rangle \right) = -\operatorname{sgn} \left( \frac{ds}{d\tau} \tilde{p}_1 \right).$$

*Proof of Lemma 8.4.* Let  $\tilde{p}(\tau) = (\tilde{p}_1, \tilde{p}_2)$  be in  $E(\tau)$ , that is,  $\tilde{p}_2 = -\tilde{p}_1\mu'_1(s)x_1/\mu'_2(s)x_2$ . It is straightforward to check that the property

$$\frac{d}{d\tau}\langle\tilde{p}, m\rangle = \left\langle A^\top m + \frac{dm}{d\tau}, \tilde{p} \right\rangle = \frac{ds}{d\tau}\tilde{p}_1 \left[ \mu''_1 - \frac{\mu'_1\mu''_2}{\mu'_2} \right] x_1$$

is fulfilled, from which (8.4) is deduced (recalling A2).  $\square$

LEMMA 8.5. *If, for some interval of time  $[\tau_-, \tau_+]$ , one has  $\phi_r = \phi_u = 0$ , then*

- (a) *if  $\tilde{p}_1 > 0$ ,  $\phi_r$  and  $\phi_u$  remain equal to zero for all  $\tau \geq \tau_+$ ;*
- (b) *if  $\tilde{p}_1 < 0$ , either  $\langle\tilde{p}, m\rangle \geq 0$  for all  $\tau \geq \tau_+$  or  $\langle\tilde{p}, m\rangle \leq 0$  for all  $\tau \geq \tau_+$ .*

*Proof of Lemma 8.5.* Notice that if we have  $\phi_r = \phi_u = 0$  in some interval of time, necessarily, by (6.5),  $\langle\tilde{p}, m\rangle = 0$  on this interval. Since  $(\tilde{p}_1, \tilde{p}_2) \neq (0, 0)$  for all  $\tau$  and the vector  $m$  lies in  $]0, +\infty[ \times ]0, +\infty[$  (under Assumption 1), we deduce that  $\tilde{p}_1 \neq 0$ , and therefore it does not change its sign in this interval.

Suppose now that  $\bar{\tau} = \sup\{\tau \mid \langle\tilde{p}, m\rangle = 0\} < +\infty$ . If

$$(8.5) \quad \exists \bar{\delta} > 0 \quad \text{such that} \quad \forall \delta \in ]0, \bar{\delta}], \quad \text{one has} \quad \langle\tilde{p}(\bar{\tau} + \delta), m(\bar{\tau} + \delta)\rangle > 0,$$

since  $\phi_r$  and  $\phi_u$  have to be nonnegative, necessarily, according to (6.5), the control  $r$  must be zero,  $u = u_{\max}$ , and therefore  $\frac{ds}{d\tau} > 0$  until  $\langle\tilde{p}, m\rangle$  changes its sign.

- (a) If  $\tilde{p}_1 > 0$  on  $[\tau_-, \tau_+]$ , let us prove that  $\langle\tilde{p}, m\rangle$  does not become positive. If it occurs, we have (8.5), and, in such a case,  $\frac{ds}{d\tau} > 0$ . Nevertheless, by (8.4), we obtain  $\frac{d}{d\tau}\langle\tilde{p}, m\rangle \leq 0$  at  $\bar{\tau}$ , which contradicts (8.5).

With similar arguments, one can prove that  $\langle\tilde{p}, m\rangle$  does not become negative, and hence  $\bar{\tau} = +\infty$ .

- (b) For the case  $\tilde{p}_1 < 0$ , if we have (8.5), then  $\frac{ds}{d\tau} > 0$  until  $\langle\tilde{p}, m\rangle$  changes its sign. This change will never happen, because, from (8.4), one has  $\frac{d}{d\tau}\langle\tilde{p}, m\rangle \geq 0$  on the set  $E(\tau) = \{\tilde{p} = (\tilde{p}_1, \tilde{p}_2) \mid \langle\tilde{p}, m(\tau)\rangle = 0\}$ , and therefore,  $\langle\tilde{p}, m\rangle$  remains nonnegative.

Analogous arguments allow us to prove that if there exists  $\bar{\delta} > 0$  such that for all  $\delta \in ]0, \bar{\delta}]$  one has  $\langle\tilde{p}(\bar{\tau} + \delta), m(\bar{\tau} + \delta)\rangle < 0$ , then  $\langle\tilde{p}, m\rangle$  remains nonpositive.  $\square$

As a corollary of Lemma 8.5, one has that if the optimal strategy includes a singular arc ( $\phi_r = \phi_u = 0$  during an interval), then it must occur with  $\tilde{p}_1 < 0$ . Indeed, if  $\tilde{p}_1 > 0$  and  $\phi_r = \phi_u = 0$  during an interval, the last equalities will remain for every larger time. This situation is not allowed because  $\tilde{p}$  must be equal to  $(-1, -1)$  at the final time.

On the other hand, if the optimal strategy includes a singular arc (with  $\tilde{p}_1 < 0$ ), after this process, necessarily,  $\langle\tilde{p}, m\rangle$  must be negative. In fact, if  $\langle\tilde{p}, m\rangle$  is positive, it does not change its sign, and then  $\tilde{p}$  will never be equal to  $(-1, -1)$ .

Thus, if a singular arc occurs, the volume at the end of this process must be equal to  $v_{\max}$  because, as  $\langle\tilde{p}, m\rangle$  will remain nonpositive and  $\phi_r$  and  $\phi_u$  have to be nonnegative, the control  $u$  is equal to zero and  $r = 1$  for the rest of time, and then the process of filling the tank has necessarily finished.

As a last consequence of Lemma 8.5, we obtain that in the case of a singular arc, which is equivalent to keep the level of substrate  $s$  constant (see Lemma 8.1), this level has to be greater than  $s_{out}$ . Indeed, if it is not the case, we have that all of the processes finish at the end of the singular arc because  $v = v_{\max}$  and  $s_{out}$  is greater than the current substrate level, and hence the target has been reached. This cannot occur, because, at the end, the vector  $\tilde{p}$  must be equal to  $(-1, -1)$ .

*Proof of Theorem 8.2.* As in the proof of Proposition 7.3, the positivity of  $\phi_u$  and  $\phi_r$  plays a crucial role (cf. (6.8), (5.5), and (5.6)).

Moreover, the considered admissible control set  $\mathcal{C}$  (see Remark 2) tell us that only optimal controls such that  $u \neq 0$  or  $r \neq 0$  are considered. And, therefore,  $\phi_u$  and  $\phi_r$  cannot be both strictly positive.

Recall that, under Assumption A1, the matrix  $A$  of system (6.2) is cooperative and the vector  $m$ , defined by (6.6), is always in  $]0, +\infty[ \times ]0, +\infty[$ . Then, since  $\tilde{p} = (p_1 - k, p_2 - k)$  is equal to  $(-1, -1)$  at the final time  $T$ , we deduce that  $\tilde{p} \notin \mathbb{R}_+^2$  at any time  $\tau$ , and moreover, once  $\tilde{p}$  reaches the negative octant  $\mathbb{R}_-^2$ , it remains there until time  $T$ . Thus, thanks to Lemma 8.4, the study of the sign of  $\tilde{p}_1$  will be a key issue in the proof. Indeed, the arguments above imply that either  $\tilde{p}_1$  remains always negative, or it is positive at initial time and then it becomes negative until the final time  $T$ .

Hence, our proof splits into the following two cases.

*Case 1:*  $\tilde{p}_1(t_0) > 0$ . Let us first discard the following case:

(a)  $v < v_{\max}$ ,  $\tilde{p}_1 > 0$ ,  $\langle \tilde{p}, m \rangle < 0$ , and  $\phi_r = 0$ .

In this situation, one has necessarily  $u = 0$  and  $r = 1$  in order to keep  $\phi_r$  nonnegative. Thus,  $s$  decreases,  $\phi_u$  increases, and there exists a time  $\tau$  such that  $\langle \tilde{p}, m \rangle = 0$  at  $\tau$  and  $\langle \tilde{p}, m \rangle > 0$  for larger time (because  $\frac{d}{d\tau} \langle \tilde{p}, m \rangle > 0$ , cf. (8.4)). Since  $\phi_r = 0$ , we obtain a contradiction with Lemma 8.3.

Thus, if  $v < v_{\max}$ ,  $\tilde{p}_1 > 0$ , and  $\langle \tilde{p}, m \rangle < 0$ , then  $\phi_r > 0$  and  $\phi_u = 0$ . In such a case, one has  $r = 0$  and  $u = u_{\max}$ , which implies that  $s$  increases and  $\phi_r$  decreases until a time such that  $\phi_r = 0$  (in order to reach the target). Since a singular arc is not possible with  $\tilde{p}_1 > 0$  (see Lemma 8.5), we discard the case  $\langle \tilde{p}, m \rangle = \frac{d}{d\tau} \langle \tilde{p}, m \rangle = 0$ . On the other hand, as the case (a) above is not possible and  $\frac{d}{d\tau} \langle \tilde{p}, m \rangle \leq 0$  (on the set  $E(\tau)$ ), the equality  $v = v_{\max}$  has to be fulfilled when  $\phi_r = 0$ . For larger times, since  $\phi_r$  must be nonnegative, one has  $u = 0$ ,  $r = 1$ , and then the obtained trajectory is exactly synthesized by the IOI strategy.

If  $v < v_{\max}$ ,  $\tilde{p}_1 > 0$ , and  $\langle \tilde{p}, m \rangle > 0$ , from Lemma 8.3, the unique possibility is to have  $\phi_r > 0$ , and consequently  $\phi_u = 0$ . In such a case, one has  $r = 0$  and  $u = u_{\max}$ . Hence,  $s$  and  $\phi_r$  increase. Then, there exists necessarily a time such that  $\langle \tilde{p}, m \rangle = 0$ , in order to reach the target. Note that, for larger time,  $\langle \tilde{p}, m \rangle < 0$  holds due to  $\frac{d}{d\tau} \langle \tilde{p}, m \rangle < 0$ . In this situation,  $\phi_r$  decreases until  $\phi_r = 0$ , and from Lemma 8.3, it has to coincide with the time at which  $v = v_{\max}$ . After, since  $\phi_r$  must be nonnegative, one has  $u = 0$  and  $r = 1$ . The obtained trajectory is again synthesized by the IOI strategy.

Hence, we have proved that when  $\tilde{p}_1(t_0)$  is positive, the IOI strategy is optimal.

*Case 2:*  $\tilde{p}_1(t_0) < 0$ . Recall, from the above discussion, that  $\tilde{p}_1$  remains always negative. We now proceed to discard the following cases:

(b)  $v < v_{\max}$ ,  $\tilde{p}_1 < 0$ ,  $\phi_u > 0$ , and  $\langle \tilde{p}, m \rangle < 0$ .

This case implies that  $u = 0$  and  $r = 1$ . Thus,  $s$  decreases, and, by (6.5),  $\phi_u$  increases. This together with (8.4) imply that the sign of  $\frac{d}{d\tau} \langle \tilde{p}, m \rangle$  is negative (on the set  $E(\tau)$  defined in Lemma 8.4). Consequently,  $\langle \tilde{p}, m \rangle$  remains always negative, and  $\phi_u$  always increases. Then the target cannot be reached because the tank is never fulfilled.

(c)  $v < v_{\max}$ ,  $\tilde{p}_1 < 0$ ,  $\phi_u = 0$ , and  $\langle \tilde{p}, m \rangle > 0$ .

In this case, one necessarily obtains  $r = 0$  and  $u = u_{\max}$  (in order to keep  $\phi_u$  nonnegative). Hence,  $s$  and  $\phi_r$  increase (see (6.5)). This together with (8.4) imply that the sign of  $\frac{d}{d\tau} \langle \tilde{p}, m \rangle$  is positive (on  $E(\tau)$ ). Consequently,  $\langle \tilde{p}, m \rangle$  remains always positive, which is a contradiction with  $\tilde{p} = (-1, -1)$  at the final time.

Hence, if  $v < v_{\max}$ ,  $\tilde{p}_1 < 0$ , and  $\langle \tilde{p}, m \rangle > 0$  necessarily  $\phi_u > 0$  and  $\phi_r = 0$ , then one has  $u = 0$  and  $r = 1$  implying that  $s$  and  $\phi_u$  decrease until  $\langle \tilde{p}, m \rangle = 0$ . Equation (8.4) allows us to say that the sign of  $\frac{d}{d\tau} \langle \tilde{p}, m \rangle$  is nonpositive, and hence  $\langle \tilde{p}, m \rangle \leq 0$  for larger times. If at time such that  $\langle \tilde{p}, m \rangle = 0$ , we have  $\phi_u > 0$  immediately after one has  $\langle \tilde{p}, m \rangle < 0$ . Then  $u = 0$  (in order to keep  $\phi_r$  nonnegative), and consequently  $\phi_u$  remains always positive, and the tank will not be fulfilled, which discard this case. Thus, necessarily  $\varphi_u$  becomes zero when  $\langle \tilde{p}, m \rangle = 0$ . If  $\langle \tilde{p}, m \rangle$  remains equal to zero for an interval of time, this corresponds to a singular arc. If not, that is,  $\langle \tilde{p}, m \rangle < 0$  immediately after, this situation implies that  $u = 0$  and  $r = 1$  onwards, which is impossible because the tank will never be filled.

Finally, we study the remaining case  $v < v_{\max}$ ,  $\tilde{p}_1 < 0$ ,  $\phi_u = 0$ , and  $\langle \tilde{p}, m \rangle < 0$ . One has necessarily  $r = 0$  and  $u = u_{\max}$ , in order to keep  $\phi_u$  nonnegative. Therefore,  $s$  increases and  $\phi_r$  decreases (see (6.5)) until a time  $\tau^*$  when one of the following three cases occur:

- case  $\phi_r > 0$  and  $\langle \tilde{p}, m \rangle = 0$ . This implies that  $u = u_{\max}$  and  $r = 0$ , and consequently  $s$  increases. This together with (8.4) implies that the sign of  $\frac{d}{d\tau} \langle \tilde{p}, m \rangle$  is positive. Consequently,  $\phi_r$  will always remain positive, which cannot allow one to reach the target. This case is thus discarded.
- case  $\phi_r = 0$  and  $\langle \tilde{p}, m \rangle < 0$ . This implies that  $u = 0$  and  $r = 1$ , and therefore (by (6.5))  $\phi_u$  becomes positive. Since (b) of Case 2 above has been discarded, it necessarily follows that  $v$  reaches  $v_{\max}$  at the same time  $\tau^*$ . The optimal trajectory is synthesized by the IOI strategy.
- case  $\phi_r = 0$  and  $\langle \tilde{p}, m \rangle = 0$ . Due to equality  $\phi_u = 0$  holding at the same time  $\tau^*$ , this configuration corresponds to a singular arc.

Thus, we have proved that if  $(u, r)(\cdot) \in \mathcal{C}$  is optimal, then it corresponds to an IOI strategy or to the singular arc strategy for a level  $s^* \geq s_{out}$ . We finish concluding that a singular arc cannot be applied on a substrate level greater than  $s_{in}$  because the domain  $\mathcal{D} = (\mathbb{R}_+^n \setminus \{0\}) \times ]0, s_{in}] \times ]0, v_{\max}[$  is invariant.  $\square$

*Remark 16.* The value of  $s^*$  depends on the initial condition and cannot be, in general, explicitly determined as in the case with one nonmonotonic species (see Proposition 7.4).

We give now an example that shows that an  $SA(\cdot)$  strategy can be better than an IOI strategy.

*Example 1.* We consider functions  $\mu_1(\cdot)$ ,  $\mu_2(\cdot)$  that fulfill Assumptions A1, A2, and A3, but not A4:

$$\mu_1(s) = s^2, \quad \mu_2(s) = 5\sqrt{s}$$

for the values  $s_{out} = 0.1$  and  $s_{in} = 5$  (see Figure 8.1).

We compare the strategies IOI and  $SA(s^*)$ , where  $s^*$  minimizes the cost of the  $SA(\bar{s})$  strategy for  $\bar{s} \in (s_{out}, s_{in})$ . For  $v_{\max} = 10$  and initial conditions with  $y_1 = 1$ ,  $z = 3$ , and  $w = 1$ , we have computed numerically  $s^*$  for different values of  $y_2$ . Results are reported in Table 8.1.

This example shows that, in the presence of a small population of a species more efficient for small substrate concentrations, the singular arc strategy may be better than the IOI one.

We focus now on sufficient conditions for which the IOI strategy is always optimal. We first consider functions  $\mu_i(\cdot)$  such that their graphs do not cross away from 0 (without loss of generality, one can assume that  $\mu_2$  is above  $\mu_1$ ).

*Assumption A4.*  $\mu_2(s) \geq \mu_1(s) \quad \forall s \in (0, s_{in}]$ .

Then, the functions  $\varphi_c(\cdot)$  possess the following properties.

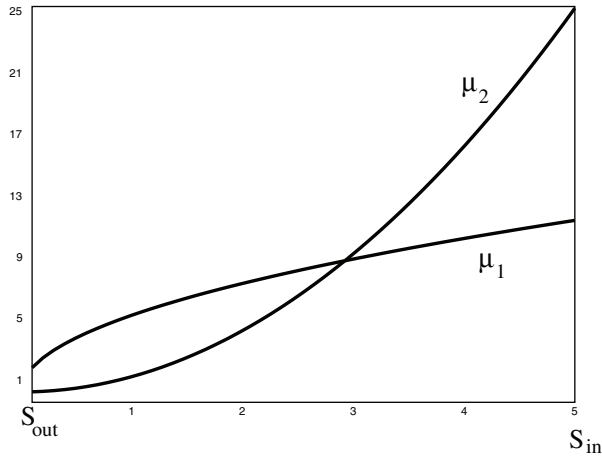


FIG. 8.1. Graphs of the two growth functions.

TABLE 8.1

$y_2$	$T(IOI)$	$s^*$	$T(SA(s^*))$	Gain
0	2.320765	not reached	—	—
$10^{-4}$	2.126756	1.678	1.973976	7%
$10^{-3}$	1.700494	1.97	1.515415	11%
$10^{-2}$	1.186231	2.46	1.101530	7%
0.05	0.867009	3.05	0.842255	3%
0.1	0.746446	3.34	0.739046	1%
0.5	0.522361	not reached	—	—

LEMMA 8.6. Under Assumptions A0, A1, and A4, for any  $c \in (0, s_{in})$ ,  $(y, z) \in (\mathbb{R}^2_+ \setminus \{0\}) \times (c, s_{in}]$ , one has

- i.  $\partial_z \varphi_c(y, z) \geq \partial_{y_2} \varphi_c(y, z)$ ,
- ii.  $\partial_{y_1} \varphi_c(y, z) \geq \partial_{y_2} \varphi_c(y, z)$ ,
- iii.  $\partial_{y_2} \varphi_c(y, z) \leq 0$ .

*Proof.* Notice that the dynamics (3.3) possesses the property that  $M = x_1 + x_2 + s$  is constant. Then, fix a value  $M$ , and consider the reduced system

(8.6) 
$$\begin{cases} \frac{dx_1}{d\tau} = f_1(x_1, s) = \mu_1(s)x_1, \\ \frac{ds}{d\tau} = f_2(x_1, s, M) = -\mu_1(s)x_1 - \mu_2(s)(M - s - x_1). \end{cases}$$

The Jacobian matrix  $\mathcal{J}_M$  of this two-dimensional vector field has the structure

$$\mathcal{J}_M(x_1, s) = \begin{pmatrix} \star & \mu_1'(s)x_1 \\ \mu_2(s) - \mu_1(s) & \star \end{pmatrix},$$

which has nonnegative off-diagonal terms for any fixed  $M$ . So, the dynamical system (8.6) is cooperative (see [28]).

- i. Consider two sets of initial conditions for system (3.3):

$$(y_1^-, y_2^-, z^-) = (y_1, y_2 + \delta, z) \quad \text{and} \quad (y_1^+, y_2^+, z^+) = (y_1, y_2, z + \delta),$$



where  $\delta$  is a positive (small) number. We check that these two initial conditions have the same invariant  $x_1 + x_2 + s = M + \delta$ , and, from the cooperative property of system (8.6) with  $M + \delta$ , one has  $s^+(t) \geq s^-(t)$  for any  $t \geq 0$ . Thus  $\varphi_c(y_1, y_2, z + \delta) \geq \varphi_c(y_1, y_2 + \delta, z)$ , or equivalently

$$\frac{\varphi_c(y_1, y_2, z + \delta) - \varphi_c(y_1, y_2, z)}{\delta} \geq \frac{\varphi_c(y_1, y_2 + \delta, z) - \varphi_c(y_1, y_2, z)}{\delta},$$

and letting  $\delta$  take arbitrary small values, we obtain

$$\partial_z \varphi_c(y, z) \geq \partial_{y_2} \varphi_c(y, z).$$

ii. Similarly, we consider the sets of initial conditions

$$(y_1^-, y_2^-, z^-) = (y_1, y_2 + \delta, z) \quad \text{and} \quad (y_1^+, y_2^+, z^+) = (y_1 + \delta, y_2, z)$$

and obtain, when  $\delta$  tends toward zero,

$$\partial_{y_1} \varphi_c(y, z) \geq \partial_{y_2} \varphi_c(y, z).$$

iii. We consider the sets of initial conditions

$$(y_1^-, y_2^-, z^-) = (y_1, y_2 + \delta, z) \quad \text{and} \quad (y_1^+, y_2^+, z^+) = (y_1, y_2, z),$$

with nonnegative  $\delta$ . The first initial condition leads to the invariant  $x_1 + x_2 + s = M + \delta$ , while the second one has  $x_1 + x_2 + s = M$  as an invariant. Dynamics (8.6) is such that  $f_2(x_1, s, M + \delta) \leq f_2(x_1, s, M)$ , and, by the cooperative property, we conclude that  $\varphi_c(y_1, y_2 + \delta, z) \leq \varphi_c(y_1, y_2, z)$  or equivalently

$$\partial_{y_2} \varphi_c(y, z) \leq 0. \quad \square$$

*Remark 17.* From expression (3.1), one can notice that inequality

$$(8.7) \quad \tilde{z} \geq z$$

is always fulfilled. Thus, under assumptions A0, A1, and A4, if the function  $\varphi_{s_{out}}(\cdot)$  is such that

$$(8.8) \quad \partial_z \varphi_{s_{out}}(y, z) \geq \partial_{y_1} \varphi_{s_{out}}(y, z), \quad \forall (y, z) \in (\mathbb{R}_+^2 \setminus \{0\}) \times (s_{out}, s_{in}],$$

then, along with point i. of Lemma 8.6, inequality (8.7), and Lemma 5.2, we deduce immediately that condition (5.11) of Proposition 5.3 is fulfilled. Unfortunately, condition (8.8) is rarely met, even for simple growth rates. In Figure 8.2, we plot iso-values of the function

$$\psi_c(y) = \varphi_c(y, M - y_1 - y_2)$$

(computed numerically) for  $\mu_1(s) = s$ ,  $\mu_2(s) = 5s$ ,  $M = 10$ , and  $c = 1$ . We see that  $\partial_{y_2} \psi_c(\cdot)$  is everywhere nonpositive, but the sign of  $\partial_{y_1} \psi_c(\cdot)$  can change. Notice that the partial derivatives of  $\psi_c(\cdot)$  are linked to the partial derivatives of  $\varphi_c(\cdot)$  as follows:

$$\partial_{y_j} \psi_c(y) = \partial_{y_j} \varphi_c(y, z) - \partial_z \varphi_c(y, z), \quad (j = 1, 2),$$

with  $z = M - y_1 - y_2$ , and we conclude that condition (8.8) is not fulfilled.

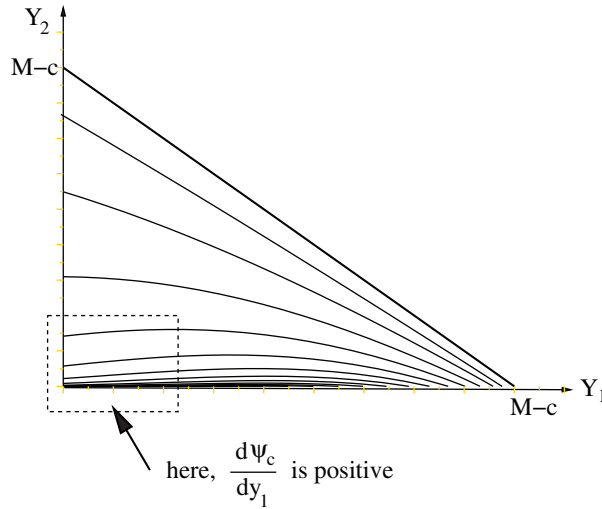


FIG. 8.2. Iso-values of the function  $\psi_c(\cdot)$ .

The following proposition imposes conditions on growth functions  $\mu_i(\cdot)$  that guarantee the optimality of the IOI strategy.

**PROPOSITION 8.7.** *Under Assumptions A0, A3, and A4, the IOI strategy is optimal for any initial condition in  $\mathcal{D}$ .*

*Proof.* Consider  $\xi \in \mathcal{D}$  such that  $T_{IOI}(\xi) > 0$ . Let us write  $\Delta_r T_{IOI}(\xi)$  given in (5.8) as follows:

$$\Delta_r T_{IOI}(\xi) = \sum_{j=1}^2 \left( \partial_{y_j} \varphi_{s_{out}}(\tilde{y}, \tilde{z}) - \partial_z \varphi_{s_{out}}(\tilde{y}, \tilde{z}) \right) \mu_j(\tilde{z}) \tilde{y}_j \frac{\mu_j(z) - \mu_j(\tilde{z})}{\mu_j(\tilde{z})}.$$

Recall that  $\tilde{z} \geq z$  (8.7) and  $\mu_i(\cdot)$  are nondecreasing (Assumption A1). Then, by Assumption A3, one has

$$\frac{\mu_2(z) - \mu_2(\tilde{z})}{\mu_2(\tilde{z})} \leq \frac{\mu_1(z) - \mu_1(\tilde{z})}{\mu_1(\tilde{z})} \leq 0.$$

Lemma 8.6 gives the inequality

$$\partial_{y_2} \varphi_{s_{out}}(\tilde{y}, \tilde{z}) - \partial_z \varphi_{s_{out}}(\tilde{y}, \tilde{z}) \leq 0,$$

and consequently, from equation (4.2), we obtain

$$\begin{aligned} \Delta_r T_{IOI}(\xi) &\geq \sum_{j=1}^2 \left( \partial_{y_j} \varphi_{s_{out}}(\tilde{y}, \tilde{z}) - \partial_z \varphi_{s_{out}}(\tilde{y}, \tilde{z}) \right) \mu_j(\tilde{z}) \tilde{y}_j \frac{\mu_1(z) - \mu_1(\tilde{z})}{\mu_1(\tilde{z})} \\ &= - \frac{\mu_1(z) - \mu_1(\tilde{z})}{\mu_1(\tilde{z})} \geq 0 \end{aligned}$$

and conclude by Proposition 5.3.  $\square$

**9. Conclusion.** In this work, we have analyzed the minimal time problem for fed-batch reactors with several species, for which impulse controls are allowed. We

have shown that even when all of the growth functions are monotonic, the most rapid approach strategy is not necessarily optimal. In certain situations, it is better to follow a singular arc instead of applying an impulse, a departure from the optimal strategy in the one-species case. We believe that this result holds important implications for biotechnological applications.

**Acknowledgments.** The authors thank the INRIA-CONICYT program for its support. The authors are also grateful to anonymous referees for their relevant suggestions.

## REFERENCES

- [1] G. D'ANS, D. GOTTLIEB, AND P. KOKOTOVIC, *Optimal control of bacterial growth*, Automatica, 8 (1972), pp. 729–736.
- [2] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Cambridge, MA, 1997.
- [3] D. BEROVIC AND R. VINTER, *The application of dynamic programming to optimal inventory control*, IEEE Trans. Automat. Control, 49 (2004), pp. 676–685.
- [4] F. BONNANS AND P. ROUCHON, *Commande et optimisation de systèmes dynamiques*, Éditions de l'École Polytechnique, Montreal, 2005.
- [5] A. BRESSAN AND F. RAMPAZZO, *Impulsive control systems with commutative vector fields*, J. Optim. Theory Appl., 71 (1991), pp. 67–83.
- [6] A. BRESSAN AND F. RAMPAZZO, *Impulsive control systems without commutativity assumptions*, J. Optim. Theory Appl., 81 (1994), pp. 435–457.
- [7] E. CRÜCK, *Problèmes de cible sous contraintes d'état pour des systèmes non linéaires avec sauts d'état*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 403–408.
- [8] V. A. DYKHTA AND O. N. SAMSONYUK, *Optimal Impulse Control with Applications* (in Russian), Fizmatlit, Moscow, 2000.
- [9] C. GAO, K. LI, E. FENG, AND Z. XIU, *Nonlinear impulsive system of fed-batch culture in fermentative production and its properties*, Chaos Solitons Fractals, 28 (2006), pp. 271–277.
- [10] J. HONG, *Optimal substrate feeding policy for fed batch fermentation with substrate and product inhibition kinetics*, Biotechnol. Bioengng., 28 (1986), pp. 1421–1431.
- [11] R. L. IRVINE AND L. H. KETCHUM, *Sequencing batch reactors for biological wastewater treatment*, Crit. Rev. Environ. Control, 18 (1989), pp. 255–294.
- [12] H. C. LIM, Y. J. TAYEB, J. M. MODAK, AND P. BONTE, *Computational algorithms for optimal feed rates for a class of fed-batch fermentation: Numerical results for penicillin and cell production*, Biotechnol. Bioengng., 28 (1986), pp. 1408–1420.
- [13] A. MIELE, *Extremization of linear integrals by Green's Theorem*, Optimization Techniques, Academic Press, New York, 1962, pp. 69–98.
- [14] B. M. MILLER, *Conditions for optimality in generalized control problems. I. Necessary conditions for optimality*, Automat. Remote Control, 53 (1992), pp. 362–370.
- [15] B. M. MILLER, *Conditions for optimality in generalized control problems. II. Sufficient conditions for optimality*, Automat. Remote Control, 53 (1992), pp. 505–513.
- [16] B. M. MILLER AND E. J. RUBINOVICH, *Optimal impulse control problem with constrained number of impulses*, Math. Comput. Simulation, 34 (1992), pp. 23–49.
- [17] B. MILLER, *The generalized solutions of ordinary differential equations in the impulse control problems*, J. Math. Systems Estim. Control, 6 (1994), pp. 415–435.
- [18] B. MILLER, *The generalized solutions of nonlinear optimization problems with impulse control*, SIAM J. Control Optim., 34 (1996), pp. 1420–1440.
- [19] B. M. MILLER AND E. Y. RUBINOVICH, *Impulsive Control in Continuous and Discrete-Continuous Systems*, Kluwer/Plenum Publishers, New York, 2003.
- [20] J. MORENO, *Optimal control of bioreactors for the wastewater treatment*, Optimal Control Appl. Methods, 20 (1999), pp. 145–164.
- [21] M. MOTTA AND F. RAMPAZZO, *Space-time trajectories of nonlinear systems driven by ordinary and impulsive controls*, Differ. Integral Equ., 8 (1995), pp. 269–288.
- [22] M. MOTTA AND F. RAMPAZZO, *Nonlinear systems with unbounded controls and state constraints: A problem of proper extension*, NoDEA Nonlinear Differential Equations Appl., 3 (1996), pp. 191–216.

- [23] M. MOTTA AND F. RAMPAZZO, *Dynamic programming for nonlinear systems driven by ordinary and impulsive controls*, SIAM J. Control Optim., 34 (1996), pp. 199–225.
- [24] M. MOTTA AND C. SARTORI, *Minimum time with bounded energy, minimum energy with bounded time*, SIAM J. Control Optim., 42 (2003), pp. 789–809.
- [25] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Translated from Russian by K. N. Trirkoff; L. W. Neustadt, ed., Interscience Publishers, John Wiley & Sons, Inc., New York, London, 1962.
- [26] F. RAMPAZZO AND C. SARTORI, *The minimum time function with unbounded controls*, J. Math. Systems Estim. Control, 8 (1998), p. 34.
- [27] R. W. RISHEL, *An extended Pontryagin principle for control systems whose control laws contain measures*, J. Soc. Ind. Appl. Math. Ser. A Control, 3 (1965), pp. 191–205.
- [28] H. L. SMITH, *Monotone Dynamical Systems*. AMS, Providence, RI, 1995.
- [29] H. L. SMITH AND P. WALTMAN, *The Theory of the Chemostat*, Cambridge University Press, London, 1995.
- [30] R. G. TSONEVA, T. D. PATARINSKA, AND I. P. POPCHEV, *Augmented Lagrange decomposition method for optimal control calculation of batch fermentation processes*, Bioprocess Engng., 18 (1998), pp. 143–153.
- [31] R. VINTER, *Optimal Control*, Birkhäuser, Cambridge, MA, 2000.
- [32] H. WANG, E. FENG, AND Z. XIU, *Optimality condition of the nonlinear impulsive system in fed-batch fermentation*, Nonlinear Anal., 68 (2008), pp. 12–23.
- [33] P. WOLENSKI AND S. ZABIC, *A differential solution concept for impulsive systems*, Dyn. Contin. Discrete Impuls. Syst. Ser. A, Math. Anal., 13 (2006), pp. 199–210.
- [34] P. WOLENSKI AND S. ZABIC, *A sampling method and approximation results for impulsive systems*, SIAM J. Control Optim., 46 (2007), pp. 983–998.
- [35] S. T. ZAVALISHCHIN AND A. N. SESEKIN, *Dynamic impulse systems: Theory and applications*, Math. Appl. 394. Kluwer Academic Publishers Group, Dordrecht, 1997.

# UNIFORM BOUNDARY CONTROLLABILITY OF A SEMIDISCRETE 1-D WAVE EQUATION WITH VANISHING VISCOSITY\*

SORIN MICU†

**Abstract.** This article deals with the approximation of the boundary control of the linear one-dimensional wave equation. It is known that the high frequency spurious oscillations that the classical methods of finite difference and finite element introduce lead to nonuniform controllability properties (see [J. A. Infante and E. Zuazua, *M2AN Math. Model. Numer. Anal.*, 33 (1999), pp. 407–438]. A space-discrete scheme with an added numerical vanishing viscous term is introduced and analyzed. The extra numerical damping filters out the high numerical frequencies and ensures the convergence of the sequence of discrete controls to a control of the continuous conservative wave equation when the mesh size tends to zero.

**Key words.** control, wave equation, semidiscrete approximation, numerical viscosity

**AMS subject classifications.** 93B40, 65M06, 93B05

**DOI.** 10.1137/070696933

**1. Introduction.** The following boundary exact controllability property for the one-dimensional (1-D) linear wave equation is known to hold: given  $T \geq 2$  and  $(u^0, u^1) \in L^2(0, 1) \times H^{-1}(0, 1)$  there exists a control function  $v \in L^2(0, T)$  such that the solution of the wave equation

$$(1.1) \quad \begin{cases} u'' - u_{xx} = 0 & \text{for } x \in (0, 1), \ t > 0, \\ u(t, 0) = 0, \quad u(t, 1) = v(t) & \text{for } t > 0, \\ u(0, x) = u^0(x), \quad u'(0, x) = u^1(x) & \text{for } x \in (0, 1) \end{cases}$$

satisfies

$$(1.2) \quad u(T, \cdot) = u'(T, \cdot) = 0.$$

By  $'$  we denote the time derivative.

For the study of this controllability problem the moments theory has been successfully used (see, for instance, [1, 25]). Also, the Hilbert uniqueness method (HUM) (see [16]) has provided a different and general way to study this and similar multidimensional problems.

In past years there was an increasing interest in the numerical approximations of the controls. For instance, HUM was used in [8, 10, 11] to deduce numerical algorithms with finite differences in the context of the two dimensional (2-D) wave equation. In these references a bad behavior of the approximate controls was observed. Let us briefly explain this phenomenon.

Generally speaking, the control's approximation strategy consists in discretizing the continuous problem (1.1), finding a control for each discrete problem, and finally making the mesh size  $h$  tend to zero. What one normally expects is to get a control of (1.1). However, as mentioned before, this is not necessarily true. Indeed, negative

---

\*Received by the editors July 11, 2007; accepted for publication (in revised form) July 9, 2008; published electronically November 19, 2008.

<http://www.siam.org/journals/sicon/47-6/69693.html>

†Facultatea de Matematica si Informatica, Universitatea din Craiova, Craiova DOLJ 200585, Romania (sd\_micu@yahoo.com). This research was partially supported by grant MTM2005-00714 of MEC (Spain) and grant CEEEX-05-D11-36/2005 (Romania).

numerical results may be obtained as a consequence of the spurious high frequency oscillations that do not exist at the continuous level but that are generated by any semidiscrete dynamics. Precisely the controls of the highest modes may have large norms which are not uniformly bounded in  $h$ . Moreover, a dispersion phenomenon appears, and the velocity of propagation of these high frequency numerical waves may converge to zero when the mesh size  $h$  tends to zero. Hence, neither is the control time uniformly bounded. All of these phenomena occur in the semidiscrete models obtained by finite differences or by the classical finite element method (see [13, 30] for a detailed analysis of the 1-D case and [29] for the 2-D case, in the context of the dual observability problem). In any of these models, the controllability property is *not uniform* as the discretization parameter  $h$  goes to zero. As a consequence there are initial data of the wave equation (even regular ones) for which the corresponding sequence of discrete controls will diverge in the  $L^2$ -norm.

Note that the spurious oscillations correspond to the high frequencies of the discrete model, and therefore they weakly converge to zero. Consequently, their existence is compatible with the convergence of the numerical scheme. However, when we are dealing with the exact controllability problem, the role of the high frequencies becomes much more important.

Since the main problem is the existence of the spurious high frequencies generated by the discretization process, the idea of eliminating or reducing them in one way or another arises naturally. All of the numerical experiments using a Tychonoff regularization technique [10, 11], a bigrid algorithm [8, 11] or a mixed finite element approximation [9] are based on this idea. How to do that in an optimal and general way and how to show mathematically the uniform controllability results are less clear. Nevertheless, in past years many theoretical results were obtained. In [18] the high frequency modes of the discrete initial data are filtered out in an appropriate manner to ensure the existence of a uniformly bounded sequence of discrete controls. In [2] a mixed finite element method is analyzed and an explicit sequence of discrete controls which tends to the HUM control of the limit wave equation (1.1) is constructed. The analysis of a bigrid method is presented in [21], where uniform results are also proved.

This paper considers a different method to achieve the uniform controllability. The idea is to introduce in the discrete equation a numerical viscous term which vanishes when the mesh size  $h$  tends to zero. The dissipation has the role to damp out the bad spurious high frequencies responsible for the large norm controls, eventually ensuring the uniform controllability of the system. The method has the advantage of being simple, general, and quite natural. Note that the amount of dissipation introduced in the discrete system is very important. If the dissipation is uniformly bounded in  $h$  (i.e., all of the eigenvalues belong to a vertical strip), it cannot compensate for the effect of the controls' growth, and the uniform result does not hold. On the other hand, the damping term should vanish in the limit and therefore cannot be enforced too much. We have treated a particular case in which we were able to prove the uniform controllability result. However, to give the optimal amount of dissipation capable of ensuring the best convergence rate and at the same time the uniformity needs further investigation (see Remark 1).

The proof of the uniform controllability result is based on the moments theory. More precisely, we construct a biorthogonal sequence in  $L^2(-\frac{T}{2}, \frac{T}{2})$  to the family of complex exponentials  $\Lambda = (e^{-\bar{\lambda}_n(h)t})_n$ , where  $\lambda_n(h)$  are the eigenvalues of the corresponding discrete operator. This is done via the Fourier transform of some entire functions of exponential type. A careful analysis of the behavior of these entire functions on the real axis gives estimates for the norm of the biorthogonal sequence.

These will allow us to show the uniform boundedness on  $h$  of the corresponding sequence of discrete controls for a large class of initial data.

Estimates for biorthogonal sequences to families of exponential functions may be found, for instance, in [6, 7] for the case of the heat equation, [5] for the case of a multidimensional wave equation, or [12] for the case of a dissipative beam equation. However, our family of exponentials has several different characteristics which make the study more difficult, such as the dependence on the discretization step  $h$  and the complexity of the exponents, which are contained neither in a sector of the real axis nor on a vertical strip of the imaginary axis as in the previous works. The Fourier transform of the biorthogonal elements has two qualitatively different behaviors on the real axis, depending on if the value of  $|x|$  is smaller than  $\frac{1}{h}$  or not. This gives interesting and new type estimates reflecting the behavior of the real parts of our exponents, which are of order  $h$  for the low frequencies but become of order  $\frac{1}{h}$  for the highest ones.

In [28] the controllability of a hyperbolic-parabolic coupled system is studied. The corresponding spectrum is a union of two families  $\{\lambda_l^l\}_{l \geq 1} \cup \{\lambda_k^h\}_{k \geq 1}$ . The first one is purely real and behaves like  $-l^2\pi^2$ , and the second one is complex and looks like  $-\frac{1}{\sqrt{2k\pi}} + k\pi i$ . Taken separately, each family has good controllability properties, and the main problem is to join them. Although there are similarities with our case (the spectrum is contained neither in a sector of the real axis nor on a vertical strip of the imaginary axis), we cannot use the technique from [28] since the high part of our spectrum is far away from any known controllable family of eigenfunctions.

The numerical viscosity technique was used in many different contexts (see, for instance, [4, 17] in the case of hyperbolic conservation laws or [22, 23, 26, 27] in the case of the semidiscrete dissipative plate or wave equation). In [22, 26, 27] an additional vanishing viscosity term is introduced in the interior of the domain to make the initially dissipative system uniformly stable. Consequently, the uniform controllability property holds if a vanishing control is added in the interior of the domain. However, our aim is more ambitious and consists of showing that no additional control is needed to ensure the convergence of this scheme. From the numerical point of view this is an advantage since the extra storage and computation for an additional control are avoided.

In [3] a singular limit problem from a parabolic to a hyperbolic equation is studied from the controllability point of view. Although the context and the approach are different, there exists at least one similarity to our case: the parabolic character vanishes in the limit when a controlled hyperbolic system is obtained.

It is known that a structural damping of the form  $-\varepsilon u_{txx}$  makes the continuous wave equation not even spectrally controllable (see [24]). This is due to the accumulation of the spectrum on  $-\varepsilon$ . In our semidiscrete case the viscosity parameter  $\varepsilon$  and the mesh size  $h$  are related and tend to zero simultaneously. The resulting system is consistent with the conservative wave equation, and therefore, unlike in [24], the uniform controllability property does hold.

The article is organized in the following way. The discrete control problem is presented in section 2, and an analysis of the energy decay rate is given in section 3. The control problem is transformed into an equivalent problem of moments in section 4. The main result is contained in section 5, where a biorthogonal sequence is constructed and evaluated. The technical but fundamental proof of the estimate on the real axis of the entire functions whose Fourier transforms give the biorthogonal sequence is given in the appendix at the end of the paper. In section 6 the existence of a bounded sequence of discrete controls is proved, and it is shown that any weak

limit is a control of the continuous wave equation. Section 7 presents some numerical experiments to support the theoretical results obtained in the paper.

**2. The discrete problem.** In this paper we study a finite-difference space discretization of (1.1). In order to do this, let us consider  $N \in \mathbb{N}^*$ , a step  $h = \frac{1}{N+1}$ , and an equidistant mesh of the interval  $(0, 1)$ ,  $0 = x_0 < x_1 < \dots < x_N < x_{N+1} = 1$ , with  $x_j = jh$ ,  $0 \leq j \leq N+1$ . The central finite-difference approximation of the space derivatives leads to the following semidiscretization of (1.1):

$$(2.1) \quad \begin{cases} u_j''(t) - \frac{u_{j+1}(t) + u_{j-1}(t) - 2u_j(t)}{h^2} = 0 & \text{for } 1 \leq j \leq N, \ t > 0, \\ u_0(t) = 0, \ u_{N+1}(t) = v_h(t) & \text{for } t > 0, \\ u_j(0) = u_j^0(x), \ u_j' = u_j^1(x) & \text{for } 1 \leq j \leq N. \end{cases}$$

System (2.1) consists of  $N$  linear differential equations with  $N$  unknowns  $u_1, u_2, \dots, u_N$ . Roughly speaking,  $u_j(t)$  approximates  $u(t, x_j)$ , the solution of (1.1), provided that  $(u_j^0, u_j^1)_{0 \leq j \leq N+1}$  is an approximation for the initial datum in (1.1). In fact, we shall choose

$$(2.2) \quad u_j^0 = u^0(jh), \quad u_j^1 = u^1(jh), \quad 0 \leq j \leq N+1.$$

Our aim is to study the following controllability property for (2.1): *given  $T > 2$  and  $(u_j^0, u_j^1)_{1 \leq j \leq N} \in \mathbb{C}^{2N}$ , there exists a control function  $v_h \in H^1(0, T)$  such that the corresponding solution  $(u_j, u_j')_{1 \leq j \leq N}$  of (2.1) satisfies*

$$(2.3) \quad u_j(T) = u_j'(T) = 0, \quad 1 \leq j \leq N.$$

If this holds for any  $(u_j^0, u_j^1)_{1 \leq j \leq N} \in \mathbb{C}^{2N}$ , we say that (2.1) is *exactly controllable in time  $T$* .

It is not difficult to see that the controllability problem we have just addressed has a positive answer and a sequence of discrete controls  $(v_h)_{h>0}$  may be easily found. Considerably more difficult is to show that the sequence  $(v_h)_{h>0}$  converges in some sense to a control  $v$  of the continuous wave equation (1.1), corresponding to the initial datum  $(u^0, u^1)$ , which verifies (2.2). In fact, due to the spurious high frequencies introduced by the discretization, the numerical scheme (2.1) gives an unbounded sequence of controls  $(v_h)_{h>0}$  even when very regular initial data  $(u^0, u^1)$  are considered (see [13, 18]).

In order to deal with the spurious high frequencies, we add a dissipative term, vanishing when the mesh size tends to zero. More precisely, we consider instead of (2.1) the following scheme:

$$(2.4) \quad \begin{cases} u_j''(t) - \frac{u_{j+1}(t) + u_{j-1}(t) - 2u_j(t)}{h^2} - \varepsilon \frac{u_{j+1}'(t) + u_{j-1}'(t) - 2u_j'(t)}{h^2} = 0, & t > 0, \\ u_0(t) = 0, \ u_{N+1}(t) = v_h(t), & t > 0, \\ u_j(0) = u_j^0(x), \ u_j' = u_j^1(x), & 1 \leq j \leq N, \end{cases}$$

and we address the same controllability problem as before.

The parameter  $\varepsilon$ , which multiplies the viscous term  $\frac{u_{j+1}'(t) + u_{j-1}'(t) - 2u_j'(t)}{h^2}$ , depends on the step size  $h$  and tends to zero as  $h \rightarrow 0$ :

$$(2.5) \quad \lim_{h \rightarrow 0} \varepsilon(h) = 0.$$



Hence, in (2.4), the term  $\varepsilon \frac{u'_{j+1}(t) + u'_{j-1}(t) - 2u'_j(t)}{h^2}$  represents a vanishing numerical viscosity that will eventually ensure the boundedness of the sequence  $(v_h)_{h>0}$ .

*Remark 1.* A similar damping term, with  $\varepsilon = h^2$ , is used in [27] in order to ensure an exponentially uniform decay rate of the discrete energy, when a dissipative term is already present on the boundary. In this case, the extra damping constitutes a bounded perturbation sufficient to restore the uniform decay rate. The spectrum of the corresponding operator is shifted to the left, but it remains in a vertical strip. However, this is not sufficient to ensure the uniform boundary controllability property. Therefore, we shall reinforce the damping by considering  $\varepsilon = h$ . This selection verifies (2.5) and, at the same time, ensures the amount of dissipation needed for the uniform control of the high frequencies. Let us finally mention that we do not know the optimal choice for  $\varepsilon(h)$ , ensuring both the uniformity in  $h$  and the best convergence rate. For instance, any  $\varepsilon(h) = h^\alpha$ , with  $\alpha \in (1, 2)$ , could still produce uniform controllability results with a better convergence rate. The numerical experiments confirm this, but the theoretical study is more difficult and needs further investigation. In [23] such dissipative terms were used to achieve uniform stability of the approximating models.

Let us first write (2.4) in a vectorial form, which is easier to deal with. We define the following matrix from  $\mathcal{M}_{N \times N}(\mathbb{R})$ :

$$A_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 \end{pmatrix}.$$

If we denote the unknown of (2.4) by  $U_h(t) = (u_1(t), u_2(t), \dots, u_N(t))^T$ , system (2.4) may be written in the following equivalent vectorial form:

$$(2.6) \quad \begin{cases} U_h''(t) + A_h U_h(t) + \varepsilon A_h U_h'(t) = F_h(t), & \text{for } t > 0, \\ U_h(0) = U_h^0, \quad U_h'(0) = U_h^1, \end{cases}$$

$U_h^0 = (u_j^0)_{1 \leq j \leq N}$  and  $U_h^1 = (u_j^1)_{1 \leq j \leq N}$  being the initial data of (2.4). The vector  $F_h$  is given by

$$F_h(t) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \frac{1}{h^2} (v_h(t) + \varepsilon v_h'(t)) \end{pmatrix}.$$

In (2.6) we have taken into account that  $u_{N+1}(t) = v_h(t)$  and  $u_0(t) = 0$  for all  $t > 0$ .

Before studying the decay properties of the solutions of (2.6), let us define in  $\mathbb{C}^N$  the canonical inner product

$$(2.7) \quad (f, g) = h \sum_{k=1}^N f_k \overline{g_k},$$

where  $f = (f_k)_{1 \leq k \leq N}$  and  $g = (g_k)_{1 \leq k \leq N}$  belong to  $\mathbb{C}^N$ .

Also, we consider in  $\mathbb{C}^{2N}$  the inner product defined by

$$(2.8) \quad (f, g)_1 = h \left[ \sum_{k=1}^{N-1} \frac{f_{k+1} - f_k}{h} \frac{\bar{g}_{k+1} - \bar{g}_k}{h} + \frac{1}{h^2} (f_1 \bar{g}_1 + f_N \bar{g}_N) \right] + h \sum_{k=N+1}^{2N} f_k \bar{g}_k,$$

where  $f = (f_k)_{1 \leq k \leq 2N}$  and  $g = (g_k)_{1 \leq k \leq 2N}$  are two vectors from  $\mathbb{C}^{2N}$ . The corresponding norm will be denoted by  $\|\cdot\|_1$ .

*Remark 2.* The following equivalent form of the inner product (2.8) justifies its definition and its usefulness for our problem:

$$(2.9) \quad (f, g)_1 = (A_h f^1, g^1) + (f^2, g^2),$$

where  $f^1 = (f_k)_{1 \leq k \leq N}$ ,  $f^2 = (f_k)_{N+1 \leq k \leq 2N}$ ,  $g^1 = (g_k)_{1 \leq k \leq N}$ ,  $g^2 = (g_k)_{N+1 \leq k \leq 2N}$  and  $f = (f_1, f_2)$ ,  $g = (g_1, g_2)$ .

The following discrete duality product will be needed in the study of the control problem:

$$(2.10) \quad \langle (f^1, f^2), (g^1, g^2) \rangle_D = - (f^1, g^2) + (f^2 + \varepsilon A_h f^1, g^1),$$

where  $f = (f^1, f^2)$  and  $g = (g^1, g^2)$  are two vectors from  $\mathbb{C}^{2N}$  as in Remark 2.

**3. Energy estimates.** In this section we briefly study system (2.4) in the uncontrolled case, i.e., when  $v_h = 0$ , and show its dissipative character. Here (2.4) is a homogeneous linear system of  $N$  differential equations of order two and has a unique solution  $U \in \mathcal{C}^\omega([0, \infty), \mathbb{R}^N)$  (the set of analytic functions defined in  $[0, \infty)$  and with values in  $\mathbb{R}^N$ ). The energy of (2.4) is defined by

$$(3.1) \quad E_h(t) = \frac{h}{2} \sum_{j=0}^N \left[ |u'_j(t)|^2 + \left| \frac{u_{j+1}(t) - u_j(t)}{h} \right|^2 \right]$$

and represents a discretization of the continuous energy corresponding to (1.1)

$$(3.2) \quad E(t) = \frac{1}{2} \int_0^1 [|u'(t)|^2 + |u_x(t)|^2] dx.$$

*Remark 3.* The energy may be expressed in terms of the inner product (2.7):

$$(3.3) \quad E_h(t) = \frac{1}{2} [(U'_h(t), U'_h(t)) + (U_h(t), A_h U_h(t))].$$

It is easy to show that system (2.4) is dissipative. More precisely, we have the following proposition.

**PROPOSITION 3.1.** *If  $v_h = 0$  in (2.4) and  $E_h(t)$  is the energy function (3.1), then*

$$(3.4) \quad \frac{dE_h}{dt}(t) = -\varepsilon (A_h U'_h(t), U'_h(t)) = -\varepsilon h \sum_{j=0}^N \left| \frac{u'_{j+1}(t) - u'_j(t)}{h} \right|^2.$$

*Proof.* Equality (3.4) follows easily, multiplying (2.6) by  $U'_h$ . Indeed,

$$\begin{aligned} 0 &= (U''_h + A_h U_h + \varepsilon A_h U'_h, U'_h) \\ &= \frac{1}{2} [(U'_h, U'_h) + (A_h U_h, U_h)]' + \varepsilon (A_h U'_h, U'_h) \\ &= \frac{dE_h}{dt}(t) + \varepsilon (A_h U'_h, U'_h), \end{aligned}$$

and the proof finishes.  $\square$

Proposition 3.1 indicates that the energy of (2.4) decreases with  $t$ . In the next theorem we shall give an estimate for the decay rate of the energy. Let us first consider the Fourier decomposition of the solutions of (2.4) by using the eigenvectors of the self-adjoint operator  $A_h$ . It is well known (see, for instance, [15]) that the eigenvalues of  $A_h$  are  $(\nu_j^2)_{1 \leq j \leq N}$ , where

$$(3.5) \quad \nu_j = \frac{2}{h} \sin\left(\frac{j\pi h}{2}\right), \quad 1 \leq j \leq N,$$

and the corresponding eigenvectors are given by

$$(3.6) \quad \varphi^j = \sqrt{2}(\sin(j\pi hk))_{1 \leq k \leq N} \in \mathbb{R}^N, \quad 1 \leq j \leq N.$$

The following property holds.

LEMMA 3.2. *The set of vectors  $(\varphi^j)_{1 \leq j \leq N}$  forms an orthonormal basis in  $\mathbb{C}^N$  with respect to the inner product defined by (2.7).*

*Proof.* We have

$$(\varphi^l, \varphi^j) = 2h \sum_{k=1}^N \sin(j\pi hk) \sin(l\pi hk) = h \sum_{k=0}^N (\cos((l-j)\pi hk) - \cos((l+j)\pi hk)).$$

But, for  $q \in \mathbb{Z}$ ,

$$\sum_{k=0}^N \cos(q\pi hk) = \begin{cases} N+1 & \text{if } q = 0, \\ \frac{1-(-1)^q}{2} & \text{if } q \in \mathbb{Z}^*. \end{cases}$$

It follows that  $(\varphi^l, \varphi^j) = \delta_{lj}$ , and the proof finishes.  $\square$

Let us expand the initial datum  $(U_h^0, U_h^1)$  of (2.6) as follows:

$$(3.7) \quad U_h^0 = \sum_{j=1}^N a_{jh}^0 \varphi^j \quad \text{and} \quad U_h^1 = \sum_{j=1}^N a_{jh}^1 \varphi^j.$$

The corresponding solution  $U_h(t)$  is given by

$$(3.8) \quad U_h(t) = \sum_{j=1}^N a_{jh}(t) \varphi^j,$$

where the coefficients  $a_{jh}(t)$  can be computed explicitly. Indeed, we have the following lemma.

LEMMA 3.3. *If the initial datum  $(U_h^0, U_h^1)$  of system (2.6) are given by (3.7), then the corresponding solution of (2.6), with  $v_h = 0$ , is given by (3.8), where*

$$(3.9) \quad a_{jh}(t) = \frac{a_{jh}^1 - \lambda_j^- a_{jh}^0}{\lambda_j^+ - \lambda_j^-} e^{\lambda_j^+ t} - \frac{a_{jh}^1 - \lambda_j^+ a_{jh}^0}{\lambda_j^+ - \lambda_j^-} e^{\lambda_j^- t}$$

and

$$(3.10) \quad \lambda_j^\pm = \frac{1}{2} \left( -\varepsilon \nu_j^2 \pm \sqrt{\varepsilon^2 \nu_j^4 - 4\nu_j^2} \right), \quad j = 1, 2, \dots, N.$$

*Proof.* The proof is elementary, and we omit it.  $\square$

*Remark 4.* The values  $(\lambda_j^\pm)_{1 \leq j \leq N}$  from (3.10) are the eigenvalues of the operator

$$\mathcal{A} = \begin{pmatrix} 0 & -I \\ A_h & \varepsilon A_h \end{pmatrix}$$

corresponding to (2.6). Note that they are complex numbers with a negative real part. We shall write

$$\lambda_j = \frac{1}{2} \left( -\varepsilon \nu_j^2 + \operatorname{sgn}(j) \sqrt{\varepsilon^2 \nu_j^4 - 4 \nu_j^2} \right), \quad 1 \leq |j| \leq N.$$

Let us remark that, in the case  $\varepsilon = h$ ,  $\lambda_j$  has the simpler form

$$(3.11) \quad \lambda_j = i \frac{2}{h} \sin \left( \frac{j\pi h}{2} \right) \left( \cos \left( \frac{j\pi h}{2} \right) + i \sin \left( \frac{j\pi h}{2} \right) \right).$$

We pass now to evaluate the decay rate of the energy corresponding to (2.6).

From now on we shall suppose that  $\varepsilon = h$ . This is the case we shall analyze from the controllability point of view later on in the paper.

**THEOREM 3.4.** *Let  $\varepsilon = h$  and  $v_h = 0$  in (2.4). There exist two positive constants  $C$  and  $\omega$ , not depending on  $h$ , such that*

$$(3.12) \quad E_h(t) \leq \frac{C}{h^2} e^{-\omega h t} E_h(0) \quad \forall t > 0.$$

*Proof.* From (3.3) and Lemmas 3.2 and 3.3 we have that

$$E_h(t) = \frac{1}{2} [(U'(t), U'(t)) + (A_h U(t), U(t))] = \frac{1}{2} \left[ \sum_{j=1}^N (|a'_{jh}(t)|^2 + \nu_j^2 |a_{jh}(t)|^2) \right].$$

From (3.9) we obtain that, for  $1 \leq j \leq N$ ,

$$\begin{aligned} |a'_{jh}(t)|^2 + \nu_j^2 |a_{jh}(t)|^2 &\leq 8e^{-\varepsilon \nu_j^2 t} \frac{|\lambda_j^+|^2 + \nu_j^2}{|\lambda_j^+ - \lambda_j^-|^2} [|a_{jh}^1|^2 + |\lambda_j^+|^2 |a_{jh}^0|^2] \\ &= 16e^{-\varepsilon \nu_j^2 t} \frac{\nu_j^2}{|\lambda_j^+ - \lambda_j^-|^2} [|a_{jh}^1|^2 + \nu_j^2 |a_{jh}^0|^2]. \end{aligned}$$

Hence,

$$(3.13) \quad E_h(t) \leq 8 \sum_{j=1}^N \left[ e^{-\varepsilon \nu_j^2 t} \frac{\nu_j^2}{|\lambda_j^+ - \lambda_j^-|^2} (|a_{jh}^1|^2 + \nu_j^2 |a_{jh}^0|^2) \right].$$

It follows that

$$E_h(t) \leq 8 \max_{1 \leq j \leq N} \left\{ \frac{\nu_j^2}{|\lambda_j^+ - \lambda_j^-|^2} \right\} e^{-\omega h t} \sum_{j=1}^N [|a_{jh}^1|^2 + \nu_j^2 |a_{jh}^0|^2] \leq \frac{1}{4h^2} e^{-\omega h t} E_h(0),$$

with

$$0 < \omega \leq 4 \leq \frac{4}{h^2} \sin^2 \frac{\pi h}{2} = \frac{1}{h} \min_{1 \leq j \leq N} \{\varepsilon \nu_j^2\}.$$

The proof is finished.  $\square$

*Remark 5.* Note that the decay rate of the energy is not uniform when  $h$  tends to zero. Nevertheless, from (3.13), it follows that

$$(3.14) \quad E_h(t) \leq \sum_{j=1}^N e^{-h\nu_j^2 t} \frac{2}{\cos\left(\frac{j\pi h}{2}\right)} (|a_{jh}^1|^2 + \nu_j^2 |a_{jh}^0|^2).$$

Since  $h\nu_j^2$  increases with  $j$ , it follows that the high frequencies are sensibly more dissipated than the lower ones. This is precisely the mechanism we shall take advantage of in the control problem.

**4. The problem of moments.** We return now to the controllability problem for (2.6). Let us recall that (2.6) is exactly controllable in time  $T$  if, for any  $(U_h^0, U_h^1) \in \mathbb{C}^{2N}$ , there exists a control function  $v_h \in H^1(0, T)$  such that the corresponding solution  $(U_h, U_h')$  of (2.6) satisfies  $U_h(T) = U_h'(T) = 0$ .

First, we deduce a variational characterization of the controllability property. Let  $(\phi_h, \phi_h')$  be the solution of the following backward homogeneous system:

$$(4.1) \quad \begin{cases} \phi_h''(t) + A_h \phi_h(t) - \varepsilon A_h \phi_h'(t) = 0 & \text{for } t \in (0, T), \\ \phi_h(T) = \phi_h^0, \quad \phi_h'(T) = \phi_h^1, \end{cases}$$

where  $(\phi^0, \phi^1) \in \mathbb{C}^{2N}$  are given.

Multiplying (4.1) by the solution  $U_h$  of (2.6) and integrating in time, we obtain

$$\begin{aligned} 0 &= \int_0^T (U_h, \phi_h'' + A_h \phi_h - \varepsilon A_h \phi_h') dt \\ &= [-(U_h' + \varepsilon A_h U_h, \phi_h) + (U_h, \phi_h')]_0^T + \int_0^T (U_h'' + A_h U_h + \varepsilon A_h U_h', \phi_h) dt \\ &= [-(U_h' + \varepsilon A_h U_h, \phi_h) + (U_h, \phi_h')]_0^T + \int_0^T (F_h, \phi_h) dt. \end{aligned}$$

Hence,

$$(4.2) \quad \langle (U_h(t), U_h'(t)), (\phi_h(t), \phi_h'(t)) \rangle_D \Big|_0^T = \int_0^T \frac{1}{h} (v_h(t) + \varepsilon v_h'(t)) \overline{\phi_N(t)} dt.$$

For any  $f_h \in L^2(0, T)$ , let  $v_h \in H^1(0, T)$  be a solution of

$$(4.3) \quad \varepsilon v_h' + v_h = f_h, \quad t \in (0, T).$$

From (4.2) we obtain the following variational characterization of the controllability property.

**LEMMA 4.1.** *Given  $T > 0$ , system (2.4) is exactly controllable in time  $T$  if and only if, for any  $(U_h^0, U_h^1) \in \mathbb{C}^{2N}$ , there exists  $f_h \in L^2(0, T)$  such that, for any  $(\phi_h^0, \phi_h^1) \in \mathbb{C}^{2N}$ ,*

$$(4.4) \quad \int_0^T f_h(t) \frac{\overline{\phi_N(t)}}{h} dt = - \langle (U_h^0, U_h^1), (\phi_h(0), \phi_h'(0)) \rangle_D,$$

$(\phi_h, \phi_h')$  being the corresponding solution of (4.1). A control  $v_h$  for (2.4) is any solution of (4.3).

In order to write (4.4) as an equivalent problem of moments, we use the Fourier expansion of the solutions of (4.1) as we have done for (2.6). Let us decompose the initial datum  $(\phi^0, \phi^1)$  of (4.1) as

$$(4.5) \quad \phi_h^0 = \sum_{j=1}^N b_{jh}^0 \varphi^j \text{ and } \phi_h^1 = \sum_{j=1}^N b_{jh}^1 \varphi^j.$$

The corresponding solution  $\phi_h$  of (4.1) has the form

$$(4.6) \quad \phi_h(t) = \sum_{j=1}^N b_{jh}(t) \varphi^j,$$

where the coefficients  $b_{jh}(t)$  are given by

$$(4.7) \quad b_{jh}(t) = \frac{b_{jh}^1 - \mu_j^- b_{jh}^0}{\mu_j^+ - \mu_j^-} e^{\mu_j^+(t-T)} + \frac{-b_{jh}^1 + \mu_j^+ b_{jh}^0}{\mu_j^+ - \mu_j^-} e^{\mu_j^-(t-T)},$$

and  $\mu_j^\pm = \frac{1}{2} \left( \varepsilon \nu_j^2 \pm \sqrt{\varepsilon^2 \nu_j^4 - 4\nu_j^2} \right)$  are the roots of the characteristic equation

$$(4.8) \quad \mu^2 - \varepsilon \nu_j^2 \mu + \nu_j^2 = 0, \quad 1 \leq j \leq N.$$

In the sequel we shall write

$$\mu_n = \frac{1}{2} \left( \varepsilon \nu_n^2 + \operatorname{sgn}(n) \sqrt{\varepsilon^2 \nu_n^4 - 4\nu_n^2} \right) = \begin{cases} \mu_n^+ & \text{if } n > 0, \\ \mu_n^- & \text{if } n < 0, \end{cases}$$

and these are the eigenvalues of the adjoint problem (4.1). We have the following new characterization of the controllability property in terms of a problem of moments.

**THEOREM 4.2.** *Let  $T > 0$ . System (2.6) is exactly controllable in time  $T$  if and only if, for any  $(U_h^0, U_h^1) \in \mathbb{C}^{2N}$  of form (3.7), there exists  $f_h \in L^2(0, T)$  such that*

$$(4.9) \quad \int_0^T f_h(t) e^{\mu_n t} dt = \frac{(-1)^n h}{\sqrt{2} \sin(|n|\pi h)} \left( \frac{\nu_n^2}{\mu_n} a_{|n|h}^0 + a_{|n|h}^1 \right), \quad 1 \leq |n| \leq N.$$

*Proof.* The proof is a direct consequence of Lemma 4.1. Indeed, it is sufficient to verify (4.4) for  $(\phi_h^0, \phi_h^1) = (\varphi^{|n|}, \mu_n \varphi^{|n|})$ ,  $1 \leq |n| \leq N$ . By taking into account (4.6)–(4.7), we deduce that in this case  $\phi = e^{(t-T)\mu_n} \varphi^{|n|}$  and

$$\phi_N(t) = (-1)^{n+1} \sqrt{2} \sin(|n|\pi h) e^{(t-T)\mu_n}.$$

On the other hand,

$$\begin{aligned} \langle (U_h^0, U_h^1), (\phi_h(0), \phi_h'(0)) \rangle_D &= \left\langle (U_h^0, U_h^1), e^{-\mu_n T} \left( \varphi^{|n|}, \mu_n \varphi^{|n|} \right) \right\rangle_D \\ &= e^{-\bar{\mu}_n T} \left( -\bar{\mu}_n \sum_{1 \leq m \leq N} a_{mh}^0 \left( \varphi^m, \varphi^{|n|} \right) \right. \\ &\quad \left. + \sum_{1 \leq m \leq N} (a_{mh}^1 + \varepsilon \nu_m^2 a_{mh}^0) \left( \varphi^m, \varphi^{|n|} \right) \right) \\ &= e^{-\bar{\mu}_n T} \left( a_{|n|h}^0 (-\bar{\mu}_n + \varepsilon \nu_n^2) + a_{|n|h}^1 \right) \\ &= e^{-\bar{\mu}_n T} \left( \frac{\nu_n^2}{\bar{\mu}_n} a_{|n|h}^0 + a_{|n|h}^1 \right). \end{aligned}$$

From Lemma 4.1 it follows immediately that (4.9) holds.  $\square$

*Remark 6.* According to Theorem 4.2 a control  $v_h$  for (2.6) is obtained by solving the following problem of moments: find  $g_h \in L^2(-\frac{T}{2}, \frac{T}{2})$  such that

$$(4.10) \quad \int_{-\frac{T}{2}}^{\frac{T}{2}} g_h(t) e^{\mu_n t} dt = \beta_{nh} e^{-\mu_n \frac{T}{2}}, \quad 1 \leq |n| \leq N,$$

where  $g_h(s - \frac{T}{2}) = f_h(s)$  and  $\beta_{nh} = \frac{(-1)^{n+1}h}{\sqrt{2} \sin(|n|\pi h)} (\frac{\nu_n^2}{\mu_n} a_{|n|h}^0 + a_{|n|h}^1)$ .

**5. Biorthogonal sequences.** Let us consider the sequence  $(\mu_n)_{\substack{|n| \leq N \\ n \neq 0}}$  of the eigenvalues of the matrix operator

$$\mathcal{A}^* = \begin{pmatrix} 0 & -I \\ A_h & -\varepsilon A_h \end{pmatrix}$$

corresponding to the adjoint problem (4.1). Recall that we have considered  $\varepsilon = h$  and consequently

$$(5.1) \quad \mu_n = i \frac{2}{h} \sin\left(\frac{n\pi h}{2}\right) \left( \cos\left(\frac{n\pi h}{2}\right) - i \sin\left(\frac{n\pi h}{2}\right) \right).$$

To solve the problem of moments (4.10), we construct an explicit biorthogonal sequence  $(\Theta_m)_{\substack{|m| \leq N \\ m \neq 0}}$  to the family of complex exponentials  $(e^{\mu_n t})_{\substack{|n| \leq N \\ n \neq 0}}$  in  $L^2(-\frac{T}{2}, \frac{T}{2})$  and we estimate the norm of the elements of this biorthogonal sequence.

We recall that  $(\Theta_m)_{\substack{|m| \leq N \\ m \neq 0}}$  is a biorthogonal sequence to  $(e^{\mu_n t})_{\substack{|n| \leq N \\ n \neq 0}}$  in  $L^2(-\frac{T}{2}, \frac{T}{2})$  if (see [1] and [31])

$$(5.2) \quad \int_{-\frac{T}{2}}^{\frac{T}{2}} \Theta_m(t) e^{\mu_n t} dt = \delta_{mn} \quad \forall m, n = \pm 1, \pm 2, \dots, \pm N.$$

*Remark 7.* If  $(\Theta_m)_{1 \leq |m| \leq N}$  is a biorthogonal sequence in  $L^2(-\frac{T}{2}, \frac{T}{2})$  to the family  $(e^{\mu_n t})_{1 \leq |n| \leq N}$ , then

$$(5.3) \quad g_h(t) = \sum_{\substack{|n| \leq N \\ n \neq 0}} \beta_{nh} e^{-\mu_n \frac{T}{2}} \Theta_n$$

is a solution of the problem of moments (4.10). Note that in (5.3) we have a finite sum, but the number of terms tends to infinity as  $h$  goes to zero. We also have

$$(5.4) \quad \|g_h\|_{L^2} \leq \sum_{\substack{|n| \leq N \\ n \neq 0}} |\beta_{nh}| e^{-\Re(\mu_n) \frac{T}{2}} \|\Theta_n\|_{L^2},$$

and an estimate for the norm of  $g_h$  is obtained from the estimate of the norm of the biorthogonal sequence. From (5.4) we see that the  $L^2$ -norm of  $g_h$  may be uniformly bounded in  $h$  even if  $\|\Theta_n\|_{L^2}$  is large, provided that the negative exponentials  $e^{-\Re(\mu_n) \frac{T}{2}}$  (due to the dissipation term we have introduced) are small enough. Our aim is to show that, for  $T$  sufficiently large,  $\|\Theta_n\|_{L^2} e^{-\Re(\mu_n) \frac{T}{2}}$  is sufficiently small to ensure the uniform boundedness of the sum.

Since  $(e^{\mu_n t})_{\substack{|n| \leq N \\ n \neq 0}}$  is a finite family of exponential functions, it follows immediately that it has infinitely many biorthogonal families in  $L^2(-\frac{T}{2}, \frac{T}{2})$ . However, we are interested not only in the existence but also on the dependence of these biorthogonal families on  $N$ . Our aim is to construct an explicit biorthogonal and to evaluate the norm of its elements. We shall do that in several steps:

1. We construct an entire function of exponential type, the product  $\Upsilon_m$ , with the property that  $\Upsilon_m(-i\mu_n) = \delta_{mn}$  (Lemma 5.1).
2. We evaluate  $\Upsilon_m$  on the real axis (Lemma 5.2).
3. We construct an entire function, the multiplier  $G$ , of exponential type such that  $\Upsilon_m G$  is bounded on the real axis (Lemma 5.3).
4. The Fourier transform of the entire function  $\Upsilon_m(z)G(z)(\frac{\sin(z+i\mu_m)}{z+i\mu_m})^2$  gives the element  $\Theta_m$  of a biorthogonal sequence. Moreover, from the Plancherel theorem, an estimate for the norm of  $\Theta_m$  is obtained too (Theorem 5.4).

This method was used in several controllability problems for the heat [6] or the wave equations [5]. However, here we deal with two parameters  $h$  and  $m$  and complex exponents  $\mu_n$ . The most difficult part consists in the estimate at step 2. This estimate will have different forms, according to the relation existing between  $x$  and  $h$ .

**5.1. The product.** Let us first define, for each  $m$  such that  $1 \leq |m| \leq N$ , the following product function:

$$(5.5) \quad \Upsilon_m(z) = \prod_{\substack{1 \leq |n| \leq N \\ n \neq m}} \left( \frac{1 + \frac{z}{i\mu_n}}{1 - \frac{\mu_m}{\mu_n}} \right) \prod_{1 \leq |n| \leq N} \exp\left(-\frac{z + i\mu_m}{i\mu_n}\right).$$

LEMMA 5.1. *The function  $\Upsilon_m$  has the following properties:*

- (i)  $\Upsilon_m(-i\mu_n) = \delta_{nm}$ ,  $1 \leq |n| \leq N$ ;
- (ii) *There exists a constant  $B_1$  independent of  $h$  and  $m$  such that  $\Upsilon_m$  is an entire function of exponential type at most  $B_1$ ; i.e., there exists a constant  $0 < A_m < 1$  such that*

$$(5.6) \quad |\Upsilon_m(z)| \leq A_m \exp(B_1|z|) \quad \forall z \in \mathbb{C}.$$

*Proof.* The first property is evident. Let us pass directly to show (5.6). Since  $\mu_{-n} = \overline{\mu_n}$ , we have that, for any  $z \in \mathbb{C}$ ,

$$\prod_{1 \leq |n| \leq N} \exp\left(\frac{z}{\mu_n}\right) = \exp\left(z \sum_{1 \leq |n| \leq N} \frac{1}{\mu_n}\right) = \exp\left(z \sum_{1 \leq n \leq N} \frac{2\Re(\mu_n)}{|\mu_n|^2}\right) = \exp(Nh z)$$

and therefore

$$(5.7) \quad \left| \exp\left(-\frac{z}{i\mu_m}\right) \prod_{1 \leq |n| \leq N} \exp\left(-\frac{\mu_m}{\mu_n}\right) \right| \leq \exp(-Nh \Re(\mu_m)) \exp\left(\frac{|z|}{2}\right).$$

On the other hand, for any  $z \neq -i\mu_n$ ,

$$\begin{aligned} & \left| \prod_{\substack{1 \leq |n| \leq N \\ n \neq m}} \left(1 + \frac{z}{i\mu_n}\right) \exp\left(-\frac{z}{i\mu_n}\right) \right| = \exp\left(\sum_{\substack{1 \leq |n| \leq N \\ n \neq m}} \ln \left| \left(1 + \frac{z}{i\mu_n}\right) \exp\left(-\frac{z}{i\mu_n}\right) \right| \right) \\ &= \exp\left(\sum_{\substack{1 \leq |n| \leq [|z|] \\ n \neq m}} \ln \left| 1 + \frac{z}{i\mu_n} \right| - \sum_{\substack{1 \leq |n| \leq [|z|] \\ n \neq m}} \Re\left(\frac{z}{i\mu_n}\right) + \sum_{\substack{[|z|]+1 \leq |n| \leq N \\ n \neq m}} \ln \left| \left(1 + \frac{z}{i\mu_n}\right) \exp\left(-\frac{z}{i\mu_n}\right) \right| \right). \end{aligned}$$



Since  $|\mu_n| = \frac{2}{h} \sin(\frac{|n|\pi h}{2}) \geq 2|n|$  and  $|\frac{z}{\mu_n}| < \frac{1}{2}$  if  $[|z|] + 1 \leq |n|$ , we deduce that

$$\begin{aligned} \sum_{\substack{1 \leq |n| \leq [|z|] \\ n \neq m}} \ln \left| 1 + \frac{z}{i\mu_n} \right| &\leq 2 \sum_{1 \leq n \leq [|z|]} \ln \left( 1 + \frac{|z|}{2n} \right) \leq 2 \int_0^{|z|} \ln \left( 1 + \frac{|z|}{2s} \right) ds \\ &= |z| (3 \ln 3 - 2 \ln 2) < 2|z|, \\ \sum_{\substack{[|z|]+1 \leq |n| \leq N \\ n \neq m}} \ln \left| \left( 1 + \frac{z}{i\mu_n} \right) \exp \left( -\frac{z}{i\mu_n} \right) \right| &\leq \sum_{\substack{[|z|]+1 \leq |n| \leq N \\ n \neq m}} \left| \frac{z}{i\mu_n} \right|^2 \leq \sum_{[|z|]+1 \leq |n| \leq N} \frac{|z|^2}{4n^2} \\ &\leq \frac{|z|}{2}, \exp \left( -\sum_{\substack{1 \leq |n| \leq [|z|] \\ n \neq m}} \Re \left( \frac{z}{i\mu_n} \right) \right) \\ &\leq \left| \exp \left( iz \sum_{1 \leq |n| \leq [|z|]} \frac{2\Re(\mu_n)}{|\mu_n|^2} \right) \right| \exp \left( \frac{|z|}{|\mu_m|} \right) \\ &\leq \exp \left( Nh|z| + \frac{|z|}{2} \right). \end{aligned}$$

It follows that

$$(5.8) \quad \left| \prod_{\substack{1 \leq |n| \leq N \\ n \neq m}} \left( 1 + \frac{z}{i\mu_n} \right) \exp \left( -\frac{z}{i\mu_n} \right) \right| \leq \exp(4|z|).$$

Finally, we evaluate

$$\begin{aligned} \prod_{\substack{1 \leq |n| \leq N \\ n \neq m}} \left| \frac{\mu_n}{\mu_n - \mu_m} \right|^2 &= \prod_{\substack{1 \leq |n| \leq N \\ n \neq m}} \frac{\sin^2 \left( \frac{n\pi h}{2} \right)}{\sin^2 \left( \frac{(n-m)\pi h}{2} \right)} \\ &= \frac{\prod_{k=N-|m|+1}^N \sin^2 \left( \frac{k\pi h}{2} \right)}{\prod_{k=N+1}^{N+|m|} \sin^2 \left( \frac{k\pi h}{2} \right)} = \frac{\prod_{k=1}^{|m|} \cos^2 \left( \frac{k\pi h}{2} \right)}{\prod_{k=0}^{|m|-1} \cos^2 \left( \frac{k\pi h}{2} \right)} = \cos^2 \left( \frac{m\pi h}{2} \right). \end{aligned}$$

Hence,

$$(5.9) \quad \prod_{\substack{1 \leq |n| \leq N \\ n \neq m}} \left| \frac{\mu_n}{\mu_n - \mu_m} \right| = \cos \left( \frac{m\pi h}{2} \right).$$

From (5.7), (5.8), and (5.9) we obtain that

$$|\Upsilon_m(z)| \leq \cos \left( \frac{m\pi h}{2} \right) \exp(-Nh \Re(\mu_m)) \exp \left( \frac{9}{2}|z| \right),$$

and (5.6) follows if  $B_1 > \frac{9}{2}$  and  $A_m = \cos \left( \frac{m\pi h}{2} \right) \exp(-Nh \Re(\mu_m))$ .  $\square$

**5.2. Estimate of the product on the real axis.** A key point in the construction and evaluation of the biorthogonal sequence is the following result concerning the behavior of  $\Upsilon_m$  on the real axis.

LEMMA 5.2. *The following estimate holds for the function  $\Upsilon_m$  on the real axis:*

$$(5.10) \quad |\Upsilon_m(x)| \leq \begin{cases} C_1 \cos\left(\frac{m\pi h}{2}\right) \left| \frac{x - i\mu_m}{i\mu_m} \right| \exp\left(\omega_1 \sqrt{\frac{|x|}{h}}\right), & |x| \geq \frac{1}{h}, \\ C_1 \cos\left(\frac{m\pi h}{2}\right) \left| \frac{x - i\mu_m}{i\mu_m} \right| \exp(\omega_1 h |x|^2), & |x| \leq \frac{1}{h}, \end{cases}$$

where  $\omega_1$  and  $C_1$  are two positive constants independent of  $h$ .

*Proof.* The proof is technical, and it will be given in the appendix.  $\square$

*Remark 8.* An estimate of the product function on the real axis is always important in these type of problems. For instance, in [6], where the heat equation is considered and consequently only real exponents are used, the product has a behavior like  $\exp(\omega\sqrt{x})$ . On the other hand, if purely imaginary exponents are considered, as in the wave equation, it is easy to see that the corresponding product is bounded. Estimate (5.10) combines these two behaviors. Our product is like the wave equation if  $hx$  is small and like the heat equation if  $hx$  is sufficiently large.

**5.3. The multiplier.** The aim of this section is to construct an entire function with a sufficient decay on the real axis to compensate for the growth of the product  $\Upsilon_m$  evaluated in Lemma 5.2. We adapt an idea of Ingham [14], used several times in the study of completeness problems for exponential functions.

LEMMA 5.3. *Let  $\varepsilon > 0$  and  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  be the function defined by*

$$(5.11) \quad \varphi(x) = \begin{cases} \varepsilon x^2, & |x| \leq \frac{1}{\varepsilon}, \\ \sqrt{\frac{|x|}{\varepsilon}}, & |x| > \frac{1}{\varepsilon}. \end{cases}$$

*There exists an entire function  $G_\varepsilon$  of exponential type such that*

$$(5.12) \quad |G_\varepsilon(x)| \leq C_2 \exp(-\varphi(x)) \quad \forall x \in \mathbb{R},$$

$$(5.13) \quad |G_\varepsilon(-i\mu_m)| \geq \exp(-\omega_2 \Re(\mu_m)), \quad 1 \leq |m| \leq N,$$

where  $C_2$  and  $\omega_2$  are two positive constants, independent of  $N$  and  $\varepsilon$ .

*Proof.* Let  $(\rho_n)_{n \geq 1}$  be the nonincreasing sequence defined by

$$(5.14) \quad \rho_n = \begin{cases} e\varepsilon, & n \leq \frac{1}{\varepsilon}, \\ e\sqrt{\frac{1}{\varepsilon n^3}}, & n > \frac{1}{\varepsilon}. \end{cases}$$

Note that  $\rho_n = e \frac{\varphi(n)}{n^2}$  and

$$\begin{aligned} \sum_{n \geq 1} \rho_n &= e \sum_{n=1}^{\left[\frac{1}{\varepsilon}\right]} \varepsilon + e \sum_{n=\left[\frac{1}{\varepsilon}\right]+1}^{\infty} \sqrt{\frac{1}{\varepsilon n^3}} \\ &\leq e + \int_{\left[\frac{1}{\varepsilon}\right]}^{\infty} \sqrt{\frac{1}{\varepsilon s^3}} ds = e + \frac{2e}{\sqrt{\varepsilon}} \frac{1}{\sqrt{\left[\frac{1}{\varepsilon}\right]}} \\ &\leq e + \frac{2e}{\sqrt{1-\varepsilon}} := l < \infty. \end{aligned}$$

Now, we define the function

$$H(z) = \prod_{n \geq 1} \frac{\sin(\rho_n z)}{\rho_n z}.$$

Since

$$\left| \frac{\sin(\rho_n z)}{\rho_n z} \right| = \left| \sum_{k \geq 0} (-1)^k \frac{(\rho_n z)^{2k}}{(2k+1)!} \right| \leq \sum_{k \geq 0} \frac{|\rho_n z|^{2k}}{(2k)!} \leq \exp(\rho_n |z|),$$

we have that

$$|H(z)| = \prod_{n \geq 1} \left| \frac{\sin(\rho_n z)}{\rho_n z} \right| \leq \exp \left( |z| \sum_{n \geq 1} \rho_n \right)$$

and  $H$  is an entire function of exponential type less than  $l$ .

We pass to evaluate  $H(x)$  by considering the following two cases:

- $|x| > \frac{1}{\varepsilon}$ . We consider  $\nu = \lceil \sqrt{\frac{|x|}{\varepsilon}} \rceil \geq [\frac{1}{\varepsilon}]$ , and we have that

$$\begin{aligned} |H(x)| &\leq \prod_{n=1}^{\nu} \frac{|\sin(\rho_n x)|}{|\rho_n x|} \leq \prod_{n=1}^{\nu} \frac{1}{\rho_n |x|} \leq \left( \frac{1}{\rho_{\nu} |x|} \right)^{\nu} \\ &= \left( \frac{\sqrt{\varepsilon \nu^3}}{e |x|} \right)^{\nu} \leq e^{-\nu} \leq e \exp \left( -\sqrt{\frac{|x|}{\varepsilon}} \right). \end{aligned}$$

- $|x| \leq \frac{1}{\varepsilon}$ . We consider  $\nu = [\frac{1}{\varepsilon}]$ , and we have that

$$|H(x)| \leq \prod_{n=1}^{\nu} \frac{|\sin(\rho_n x)|}{|\rho_n x|} = \left( \frac{|\sin(e \varepsilon x)|}{|e \varepsilon x|} \right)^{\nu}.$$

Since  $e \varepsilon |x| \leq e$  and  $\sin(t) \leq t - \frac{\sin(e)}{6e} t^3$  for all  $t \in [0, e]$ , it follows that

$$\begin{aligned} H(x) &\leq \left( 1 - \frac{\sin(e)}{6e} (e \varepsilon |x|)^2 \right)^{\nu} = \exp \left( \nu \ln \left( 1 - \frac{\sin(e)}{6e} (e \varepsilon |x|)^2 \right) \right) \\ &\leq \exp \left( -\nu \frac{\sin(e)}{6e} (e \varepsilon |x|)^2 \right) \leq \exp \left( \frac{e \sin(e)}{6} \right) \exp \left( -\frac{1}{\varepsilon} \frac{\sin(e)}{6e} (e \varepsilon |x|)^2 \right) \\ &\leq e \exp \left( -\frac{e \sin(e)}{6} \varepsilon |x|^2 \right) \leq e \exp \left( -\frac{1}{6} \varepsilon |x|^2 \right). \end{aligned}$$

It follows that the function  $G_{\varepsilon}(z) = (H(z))^6$  verifies (5.12), with  $C_2 = e^6$ .

We prove that (5.13) holds too. We have that

$$|H(-i \mu_m)| = \prod_{n=1}^{\infty} \left| \frac{\sin(i \rho_n \mu_m)}{i \rho_n \mu_m} \right| = \prod_{\rho_n |\mu_m| \leq 1} \left| \frac{\sin(i \rho_n \mu_m)}{i \rho_n \mu_m} \right| \prod_{\rho_n |\mu_m| > 1} \left| \frac{\sin(i \rho_n \mu_m)}{i \rho_n \mu_m} \right|.$$

If  $\rho_n |\mu_m| \leq 1$ ,

$$|\sin(i \rho_n \mu_m)| \geq \sin(\rho_n |\mu_m|) \geq \rho_n |\mu_m| - \frac{(\rho_n |\mu_m|)^3}{6}$$

and consequently

$$\begin{aligned} \prod_{\rho_n |\mu_m| \leq 1} \left| \frac{\sin(i \rho_n \mu_m)}{i \rho_n \mu_m} \right| &= \exp \left( \sum_{\rho_n |\mu_m| \leq 1} \ln \left| \frac{\sin(i \rho_n \mu_m)}{i \rho_n \mu_m} \right| \right) \\ &\geq \exp \left( \sum_{\rho_n |\mu_m| \leq 1} \ln \left( 1 - \frac{(\rho_n |\mu_m|)^2}{6} \right) \right) \\ &\geq \exp \left( -\frac{|\mu_m|^2}{6} \sum_{n \geq 1} \rho_n^2 \right). \end{aligned}$$

Since

$$\sum_{n \geq 1} \rho_n^2 = \sum_{n \leq [\frac{1}{\varepsilon}]} e^2 \varepsilon^2 + \sum_{n > [\frac{1}{\varepsilon}]} e^2 \frac{1}{\varepsilon n^3} \leq e^2 \varepsilon + \frac{e^2}{\varepsilon} \int_{[\frac{1}{\varepsilon}]}^{\infty} \frac{ds}{s^3} \leq 4e^2 \varepsilon,$$

it follows that

$$(5.15) \quad \prod_{\rho_n |\mu_m| \leq 1} \left| \frac{\sin(i \rho_n \mu_m)}{i \rho_n \mu_m} \right| \geq \exp \left( -\frac{2e^2}{3} \Re(\mu_m) \right).$$

If  $\rho_n |\mu_m| > 1$ , we have that  $|\mu_m| > \frac{1}{\rho_n} \geq \frac{1}{e\varepsilon}$  and

$$\Re(\rho_n \mu_m) = \rho_n \Re(\mu_m) = \rho_n \varepsilon |\mu_m|^2 \geq \varepsilon |\mu_m| \geq \frac{1}{e}.$$

It follows that

$$\begin{aligned} \prod_{\rho_n |\mu_m| > 1} \left| \frac{\sin(i \rho_n \mu_m)}{i \rho_n \mu_m} \right| &= \prod_{\rho_n |\mu_m| > 1} \left| \frac{e^{\rho_n \mu_m} - e^{-\rho_n \mu_m}}{2\rho_n \mu_m} \right| \\ &\geq \prod_{\rho_n |\mu_m| > 1} \frac{e^{\rho_n \Re(\mu_m)} - e^{-\rho_n \Re(\mu_m)}}{2\rho_n |\mu_m|} \geq \prod_{\rho_n |\mu_m| > 1} \frac{2\rho_n \Re(\mu_m)}{2\rho_n |\mu_m|} \\ &= \prod_{\rho_n |\mu_m| > 1} \varepsilon |\mu_m| \geq \prod_{\rho_n |\mu_m| > 1} \frac{1}{e} \\ &= \exp \left( -\left| \left\{ n \geq 1 : \rho_n \geq \frac{1}{|\mu_m|} \right\} \right| \right). \end{aligned}$$

Since  $|\mu_m| > \frac{1}{e\varepsilon}$  we have that

$$\left| \left\{ n \geq 1 : \rho_n \geq \frac{1}{|\mu_m|} \right\} \right| = \left\lceil \sqrt[3]{\frac{e^2}{\varepsilon} |\mu_m|^2} \right\rceil \leq e^2 \varepsilon |\mu_m|^2 = e^2 \Re(\mu_m)$$

and therefore

$$(5.16) \quad \prod_{\rho_n |\mu_m| > 1} \left| \frac{\sin(i \rho_n \mu_m)}{i \rho_n \mu_m} \right| \geq \exp(-e^2 \Re(\mu_m)).$$

From (5.15) and (5.16) it follows that inequality (5.13) holds with  $\omega_2 = 10e^2$ , and the proof ends.  $\square$

**5.4. Biorthogonal function estimates.** We are now ready to prove the desired result on the biorthogonal sequence.

**THEOREM 5.4.** *For any  $T > 0$  sufficiently large but independent of  $h$ , there exists a sequence  $(\Theta_m)_{\substack{|m| \leq N \\ m \neq 0}}$ , biorthogonal in  $L^2(-\frac{T}{2}, \frac{T}{2})$  to the family  $(e^{\mu_n t})_{\substack{|n| \leq N \\ n \neq 0}}$ , such that*

$$(5.17) \quad \|\Theta_m\|_{L^2(-\frac{T}{2}, \frac{T}{2})} \leq M \cos\left(\frac{m\pi h}{2}\right) \exp(\omega \Re(\mu_m)), \quad 1 \leq |m| \leq N,$$

where  $M$  and  $\omega$  are positive constants, independent of  $m$  and  $N$ .

*Remark 9.* Theorem 5.4 provides a biorthogonal set for any  $T > 0$ . However, for estimate (5.17) we need a time  $T$  sufficiently large (but independent of the discretized problem). The value  $T = 2(B_1 + 74\pi^4 e + 2)$ , for any  $B_1 > \frac{9}{2}$ , is obtained in the proof. We may improve it by finer estimates, but we are still far from  $T = 2$ , which is probably the optimal value.

*Proof.* We define

$$(5.18) \quad \Xi_m(z) = \Upsilon_m(z) \left[ \frac{G_h(z)}{G_h(-i\mu_m)} \right]^{\omega_1} \left( \frac{\sin(z + i\mu_m)}{z + i\mu_m} \right)^2,$$

where  $\Upsilon_m$  is given by (5.5) and  $G_h$  is the function constructed in Lemma 5.3, with  $\varepsilon = h$ . We have that

- $\Xi_m(-i\mu_n) = \delta_{nm}$ ,  $1 \leq |n|, |m| \leq N$ ;
- $\Xi_m$  is an entire function of exponential type  $B = B_1 + 6l\omega_1 + 2$ , independent of  $N$ ;
- by using the properties of  $G_h$  from Lemma 5.3 and the estimates of  $\Upsilon_m$  from Lemma 5.2, we obtain that

$$\begin{aligned} \int_{-\infty}^{\infty} |\Xi_m(x)|^2 dx &\leq \frac{C_1^2 C_2^{2\omega_1} \cos^2\left(\frac{m\pi h}{2}\right)}{|G(-i\mu_m)|^{2\omega_1}} \int_{-\infty}^{\infty} \left| \frac{x - i\mu_m}{i\mu_m} \right|^2 \left| \frac{\sin(x + i\mu_m)}{x + i\mu_m} \right|^4 dx \\ &\leq \frac{2C_1^2 C_2^{2\omega_1} \cos^2\left(\frac{m\pi h}{2}\right)}{|G(-i\mu_m)|^{2\omega_1}} \left( \frac{e^{2\Re(\mu_m)}}{|\mu_m|^2} \int_{-\infty}^{\infty} \left| \frac{\sin(t + i\Re(\mu_m))}{t + i\Re(\mu_m)} \right|^2 dt \right. \\ &\quad \left. + 4 \int_{-\infty}^{\infty} \left| \frac{\sin(t + i\Re(\mu_m))}{t + i\Re(\mu_m)} \right|^4 dt \right) \\ &\leq \frac{8C_1^2 C_2^{2\omega_1} \cos^2\left(\frac{m\pi h}{2}\right)}{|G(-i\mu_m)|^{2\omega_1}} e^{2(1+\pi)\Re(\mu_m)} \int_{-\infty}^{\infty} \left| \frac{\sin(t)}{t} \right|^2 dt \\ &\leq 8\pi C_1^2 C_2^{2\omega_1} \cos^2\left(\frac{m\pi h}{2}\right) e^{(2+2\pi+2\omega_1\omega_2)\Re(\mu_m)}. \end{aligned}$$

We introduce now the Fourier transform of  $\Xi_m$

$$(5.19) \quad \Theta_m(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Xi_m(x) e^{-xz} dx,$$

and we note that  $\{\Theta_m\}_{\substack{|m| \leq N \\ m \neq 0}}$  is the biorthogonal sequence we are looking for.

Indeed, from the properties of  $\Xi_m$ , by using the Paley–Wiener theorem, it follows that  $\Theta_m(t)$  has compact support in  $[-B, B]$ , it belongs to  $L^2(-B, B)$ , and

$$\int_{-B}^B \Theta_m(t) e^{\mu_n t} dt = \Xi_m(-i\mu_n) = \delta_{nm}, \quad 1 \leq |n| \leq N.$$

It follows that  $(\Theta_m)_{\substack{|m| \leq N \\ m \neq 0}}$  is a biorthogonal sequence to  $\{e^{\mu_n t}\}_{\substack{|n| \leq N \\ n \neq 0}}$  in  $L^2(-B, B)$ . Moreover, from Plancherel's theorem we have

$$\sqrt{2\pi} \|\Theta_m\|_{L^2(-B, B)} = \|\Xi_m\|_{L^2(-\infty, \infty)} \leq C_1 C_2^{\omega_1} \sqrt{8\pi} \cos\left(\frac{m\pi h}{2}\right) e^{(1+\pi+\omega_1\omega_2)\Re(\mu_m)},$$

and the proof is finished.  $\square$

**6. Convergence results.** The aim of this section is to show that a control for the continuous system (1.1) may be obtained as a limit of controls of the corresponding semidiscrete problems (2.6). The control time  $T$  will be considered sufficiently large (but independent of the discretized problem) such that estimate (5.17) from Theorem 5.4 holds.

Let  $(u^0, u^1) \in L^2(0, 1) \times H^{-1}(0, 1)$  be the initial datum of (1.1). We consider that  $(u^0, u^1)$  has a Fourier decomposition

$$(6.1) \quad (u^0, u^1) = \sum_{n \geq 1} (a_n^0, a_n^1) \sqrt{2} \sin(n\pi x).$$

Since  $\sqrt{2} \sin(n\pi x)$  is orthonormal in  $L^2(0, 1)$ , it follows that  $(u^0, u^1) \in L^2(0, 1) \times H^{-1}(0, 1)$  if and only if

$$(6.2) \quad \sum_{n \geq 1} \left( |a_n^0|^2 + \frac{1}{|n|^2 \pi^2} |a_n^1|^2 \right) < \infty.$$

Now, let  $(U_h^0, U_h^1)_{h>0}$  be a sequence of discretizations of  $(u^0, u^1)$  given by (3.7). Assume that  $(a_{nh}^0, a_{nh}^1)_{n \in \mathbb{Z}^*}$ , the Fourier coefficients of the discrete initial data, verify

$$(6.3) \quad (a_{nh}^0)_n \rightharpoonup (a_n^0)_n, \quad \left( \frac{a_{nh}^1}{\lambda_n} \right)_n \rightharpoonup \left( \frac{a_n^1}{n\pi i} \right)_n \text{ when } h \rightarrow 0 \text{ in } \ell^1.$$

*Remark 10.* The usual discretization by points

$$(6.4) \quad (U_h^0, U_h^1) = \left( (u^0(jh))_{1 \leq j \leq N}, (u^1(jh))_{1 \leq j \leq N} \right)$$

leads to a convergence property of the Fourier coefficients sequence that depends on the regularity of  $(u^0, u^1)$ . Indeed, it is not difficult to prove the following:

(i) If  $u^0$  and  $u^1$  are piecewise continuous functions in  $[0, 1]$ , then

$$(6.5) \quad (a_{nh}^0)_n \rightarrow (a_n^0)_n, \quad \left( \frac{a_{nh}^1}{\lambda_n} \right)_n \rightarrow \left( \frac{a_n^1}{n\pi i} \right)_n \text{ when } h \rightarrow 0 \text{ in } \ell^1.$$

(ii) If  $u^0$  and  $u^1$  are one-time derivable with a continuous derivative in  $[0, 1]$ , then

$$(6.6) \quad (a_{nh}^0)_n \rightarrow (a_n^0)_n, \quad \left( \frac{a_{nh}^1}{\lambda_n} \right)_n \rightarrow \left( \frac{a_n^1}{n\pi i} \right)_n \text{ when } h \rightarrow 0 \text{ in } \ell^1.$$

We prove now the existence of a bounded sequence of controls for the semidiscrete problem.

**THEOREM 6.1.** *Let us suppose that the initial data of (1.1) are such that*

$$(6.7) \quad \sum_{n \geq 1} \left( |a_n^0| + \frac{1}{n\pi} |a_n^1| \right) < \infty.$$

There exists a control  $v_h$  of the semidiscrete problem (2.4), with  $\varepsilon = h$ , such that the sequence  $(v_h)_{h>0}$  is bounded in  $L^2(0, T)$ .

*Proof.* Let  $(\Theta_m)_{\substack{|m| \leq N \\ m \neq 0}}$  be the biorthogonal sequence in  $L^2(-\frac{T}{2}, \frac{T}{2})$  to  $\{e^{i\lambda_n t}\}_{\substack{|n| \leq N \\ n \neq 0}}$  constructed in Theorem 5.4. From Remarks 6 and 7, we obtain that

$$(6.8) \quad f_h(t) = \sum_{1 \leq |n| \leq N} \frac{(-1)^{n+1}h}{\sqrt{2} \sin(|n|\pi h)} \left( \frac{\nu_n^2}{\mu_n} a_{|n|h}^0 + a_{|n|h}^1 \right) e^{-\mu_n \frac{T}{2}} \Theta_n \left( t - \frac{T}{2} \right)$$

is a solution of (4.10). Moreover,

$$(6.9) \quad \|f_h\|_{L^2(0,T)} \leq \sum_{1 \leq |n| \leq N} \frac{h \exp\left(-\Re(\mu_n) \frac{T}{2}\right)}{\sqrt{2} |\sin(n\pi h)|} \left( |\nu_n| |a_{|n|h}^0| + |a_{|n|h}^1| \right) \|\Theta_n\|_{L^2(-\frac{T}{2}, \frac{T}{2})}.$$

From the estimates for the norm of  $\Theta_n$  given by Theorem 5.4, it follows that for any  $T > 2\omega$ ,

$$\|f_h\|_{L^2(0,T)} \leq M \sum_{1 \leq |n| \leq N} \left( a_{|n|h}^0 + \frac{1}{|\nu_n|} a_{|n|h}^1 \right).$$

It follows that any sequence of controls  $(v_h)_{h>0}$  given by (4.3) is uniformly bounded in  $L^2(0, T)$ , and the proof ends.  $\square$

*Remark 11.* Theorem 6.1 shows that the initial data which verify (6.7) can be uniformly controlled with the scheme (2.4). The method we have used does not allow us to prove the optimal result for the initial data in  $L^2 \times H^{-1}$  which verifies (6.2). This is a general limitation of the biorthogonal technique already mentioned in [6].

Since the sequence of controls  $(v_h)_h$  given by Theorem 6.1 is bounded in  $L^2(0, T)$ , there exists a subsequence, denoted in the same way, and  $v \in L^2(0, T)$  such that  $v_h \rightharpoonup v$  in  $L^2(0, T)$  when  $h \rightarrow 0$ . In the next theorem we show that  $v$  is a control for the corresponding continuous problem.

**THEOREM 6.2.** *If  $v \in L^2(0, T)$  is a weak limit of the bounded sequence  $(v_h)_h$  given by Theorem 6.1, then  $v$  is a control for the continuous problem (1.1).*

*Proof.* Like in the case of the semidiscrete problem, it is easy to prove that  $v \in L^2(0, T)$  is a control for (1.1) if and only if the equality

$$(6.10) \quad \int_0^T v(t) \overline{\varphi}_x(t, 1) dt = \langle u^1, \overline{\varphi}(0) \rangle_{H^{-1}, H_0^1} - \int_0^1 u^0 \overline{\varphi}'(0)$$

holds for any  $(\varphi^0, \varphi^1) \in H_0^1(0, 1) \times L^2(0, 1)$  and for  $\varphi$  the solution of the adjoint equation

$$(6.11) \quad \begin{cases} \varphi'' - \varphi_{xx} = 0 & \text{for } x \in (0, 1), \ t > 0, \\ \varphi(t, 0) = \varphi(t, 1) = 0 & \text{for } t > 0, \\ \varphi(T, x) = \varphi^0(x), \ \varphi'(T, x) = \varphi^1(x) & \text{for } x \in (0, 1). \end{cases}$$

Now, the Fourier decomposition allows us to show that  $v$  is a control for (1.1) if and only if

$$(6.12) \quad \int_0^T v(t) e^{-in\pi t} dt = \frac{(-1)^n}{\sqrt{2}} \left( -a_{|n|}^0 i + \frac{a_{|n|}^1}{n\pi} \right) \quad \forall n \neq 0.$$

Note that this is the moment problem for the continuous system (1.1), similar to (4.9) from Theorem 4.2. Let  $v$  be a weak limit in  $L^2(0, T)$  of the sequence  $(v_h)_h$ . It follows that  $v$  is a weak limit of the sequence  $(f_h)_h$  too.

Note that, for each  $n \in \mathbb{Z}^*$ ,

$$e^{\mu_n t} \rightarrow e^{-in\pi t} \text{ in } L^2(0, T)$$

and

$$\frac{h}{\sin(|n|\pi h)} \left( \frac{\nu_n^2}{\mu_n} a_{|n|h}^0 + a_{|n|h}^1 \right) \rightarrow a_n^0 i + \frac{a_n^1}{n\pi}$$

when  $h$  tends to zero.

By passing to the limit in (4.9), it follows that  $v$  satisfies (6.12). Hence, the limit  $v$  is a control for the problem (1.1), and the proof finishes.  $\square$

*Remark 12.* The weak convergence of the controlled discrete solutions to the controlled continuous solution of (1.1) may be proved too. Moreover, if the discrete initial data converge stronger to the continuous ones, the sequence of discrete HUM controls converges strongly to the continuous HUM control. Both proofs are rather technical and very similar to Theorems 3.2 and 3.3 from [2], and we omit them.

**7. Numerical results.** In this section we present a numerical experiment based on the scheme with an added viscosity term. More precisely, we approximate the HUM control for (1.1) (the control of minimal  $L^2$ -norm) denoted by  $\hat{v}$  by using (2.4) as the discretization of (1.1).

The algorithm we have used to compute the approximate controls is inspired by the one proposed by Glowinski, Li, and Lions [10] (see also [8, 11]), and it is based on a conjugate gradient implementation of the HUM method. To this end, we use the approximations  $(U_h^0, U_h^1)$  of the initial data by taking the values of  $u^0$  and  $u^1$  at the nodes. Then, we minimize the following functional

$$(7.1) \quad J(\phi_h^0, \phi_h^1) = \frac{1}{2} \int_0^T \left( \frac{\phi_N(t) - \varepsilon \phi'_N(t)}{h} \right)^2 dt + \langle (U_h^0, U_h^1), (\phi_h(0), \phi'_h(0)) \rangle_D$$

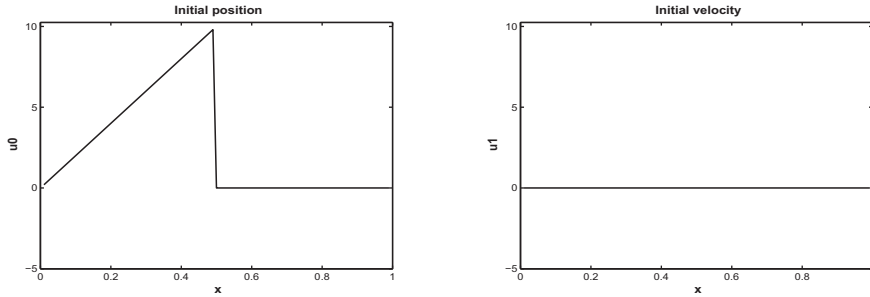
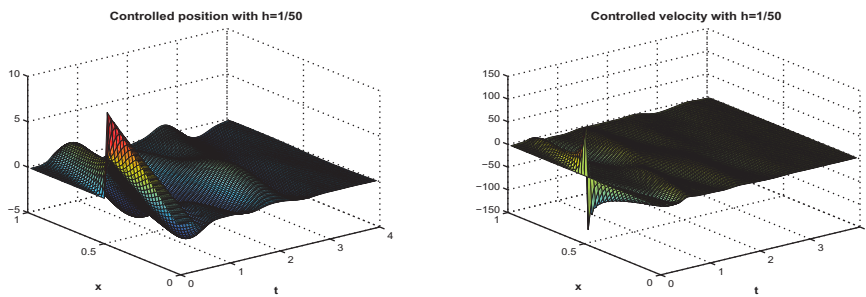
over all  $(\phi^0, \phi^1) \in \mathbb{C}^{2N}$ ,  $\phi_N$  being the last component of  $\phi$ , the solution of the adjoint system (4.1).

The minimizer  $(\hat{\phi}_h^0, \hat{\phi}_h^1)$  of  $J$  provides the control  $\hat{v}_h$  of the discrete system (2.4) with minimal  $L^2$ -norm. Since in Theorem 6.1 we have proved the existence of a bounded sequence of discrete controls, it follows that  $(\hat{v}_h)_{h>0}$  is bounded too. Its (unique) weak limit is the HUM control of (1.1) denoted by  $\hat{v}$ .

In the algorithm several wave equations have to be solved. To do that, we also need a time discretization. Except for these numerical experiments, our article does not deal with the fully discrete problem, and the convergence remains to be done. However, the uniform results we have obtained for the semidiscrete case suggest that this is a good scheme to be discretized in time. The full discrete problem is analyzed, for instance, in [19, 20].

Let  $\Delta t$  be the time step and  $l = \frac{\Delta t}{h}$  be the Courant number. We recall that the classical central difference scheme for the constant coefficients 1-D wave equation, with Courant number  $l$  equal to one, provides the exact solution at the nodes. Therefore, in this very particular situation, the 1-D version of the conjugate gradient algorithm described in [8, 11] gives a very accurate approximation of the control for any  $h > 0$ . We use this special situation to compute “exact” controls that allow us to compare the results obtained with our method. However, this scheme fails to converge if  $l \neq 1$ . Note that, in view of the possible generalizations to other types of equations



FIG. 1. *Initial data  $(u^0, u^1)$  to be controlled.*FIG. 2. *The controlled position and velocity with  $h = \frac{1}{50}$ .*

or multidimensional domains, it is of utmost importance to find robust algorithms which are not sensitive on the various discretization parameters. As the numerical results illustrate, the convergence of our method does not depend on the election of the Courant number  $l = \Delta t/h$ , which is related to the theoretical uniform result we have obtained.

**Numerical example.** In this example we consider a singular situation with discontinuous initial data  $(u^0, u^1)$ . We take

$$u^0(x) = \begin{cases} 20x & \text{if } x \in (0, 1/2), \\ 0 & \text{if } x \in (1/2, 1), \end{cases} \quad u^1(x) = 0.$$

In Figure 1 we present a picture of these data. Note that  $u^0 \in L^2(0, 1) \setminus H^1(0, 1)$ .

Since we are not looking for the optimal control time, we take  $T = 4$ . This is an arbitrary choice. The value of  $T$  for which we have proved the convergence is much larger, and it is given in Remark 9. It may be improved by finer estimates, but, in this case, our option was for simplicity.

The algorithm based on the finite difference scheme (2.1) gives, when  $l := \Delta t/h < 1$ , unsatisfactory approximations of the HUM control corresponding to  $(u^0, u^1)$ . In the following experiments we have taken  $l = 7/8$  and we have used (2.4). Figure 2 shows the controlled solution (position and velocity) when  $h = 1/50$ . Figure 3 illustrates the convergence of the discrete controls (dashed line) as  $h$  is decreased. The results are compared with those obtained by the finite difference scheme, with  $\Delta t = h$  (solid line). We may conclude that, even in this singular situation, the viscosity method (2.4) provides satisfactory approximations of the HUM control.

The first line of results in Table 7.1 shows that the error in the  $L^2$ -norm decreases with  $h$  but at a slow rate when  $\varepsilon = h$ . It seems that this may be improved by choosing

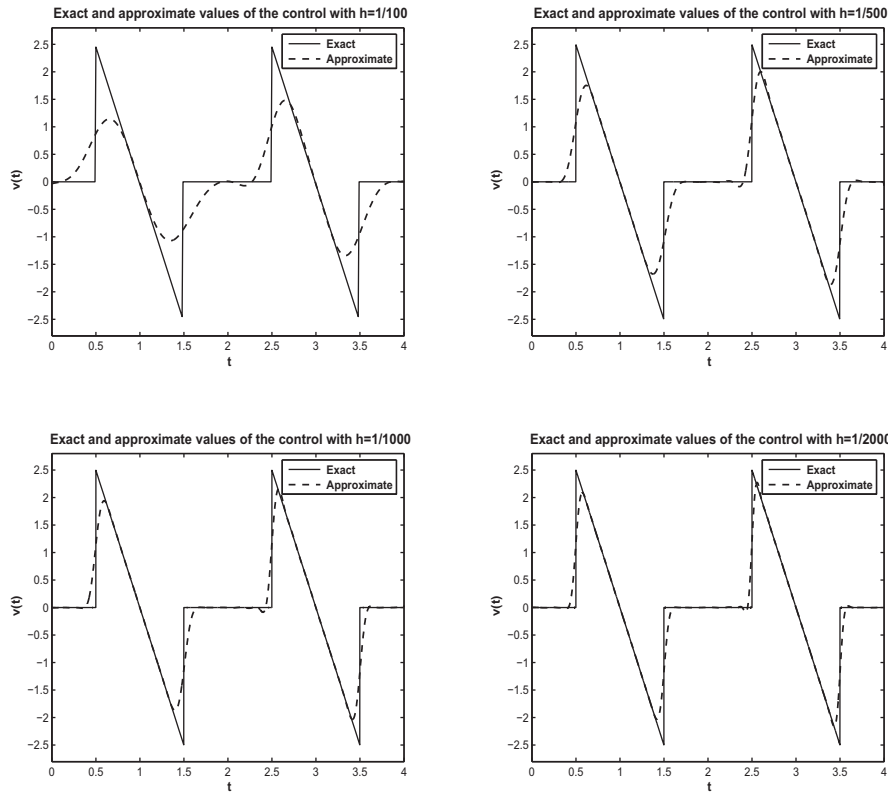


FIG. 3. Approximations of the control with  $h = \frac{1}{100}, \frac{1}{500}, \frac{1}{1000}$ , and  $\frac{1}{2000}$ .

TABLE 7.1  
Numerical results for  $\|v_h\|_{L^2}$  obtained with  $\Delta t = 7/8h$  and different values of the parameters  $\varepsilon$  and  $h$ . The exact result is  $\|v\|_{L^2} = 2.0106$ .

h	1/100	1/500	1/1000
$\ v_h\ _{L^2}$ with $\varepsilon = h$	1.4656	1.8013	1.8750
$\ v_h\ _{L^2}$ with $\varepsilon = h^{1.5}$	1.8495	1.9877	2.0101
$\ v_h\ _{L^2}$ with $\varepsilon = h^{1.7}$	1.9117	2.0100	2.0242
$\ v_h\ _{L^2}$ with $\varepsilon = h^{1.9}$	1.9540	2.0225	2.0316

$\varepsilon = h^\alpha$ , with  $\alpha \in (1, 2)$  in (2.4). Numerical simulations with different values of the parameter  $\varepsilon$  are presented in Figure 4 and Table 7.1. We note better convergence rates with larger values of  $\alpha$ , with an optimum close to 1.7. However, the theoretical analysis of this problem should be based on the error estimates for the control, which, to our knowledge, have not been yet obtained, regardless of the convergent discrete scheme.

8. Appendix. Proof of Lemma 5.2. We have that

$$\Upsilon_m(x) = \left(1 + \frac{x}{i\mu_{-m}}\right) \prod_{\substack{1 \leq |n| \leq N \\ n \neq m}} \frac{\mu_n}{\mu_m - \mu_n} \prod_{\substack{1 \leq |n| \leq N \\ n \neq |m|}} \left(1 + \frac{x}{i\mu_n}\right) \prod_{1 \leq |n| \leq N} \exp\left(-\frac{x + i\mu_m}{i\mu_n}\right).$$

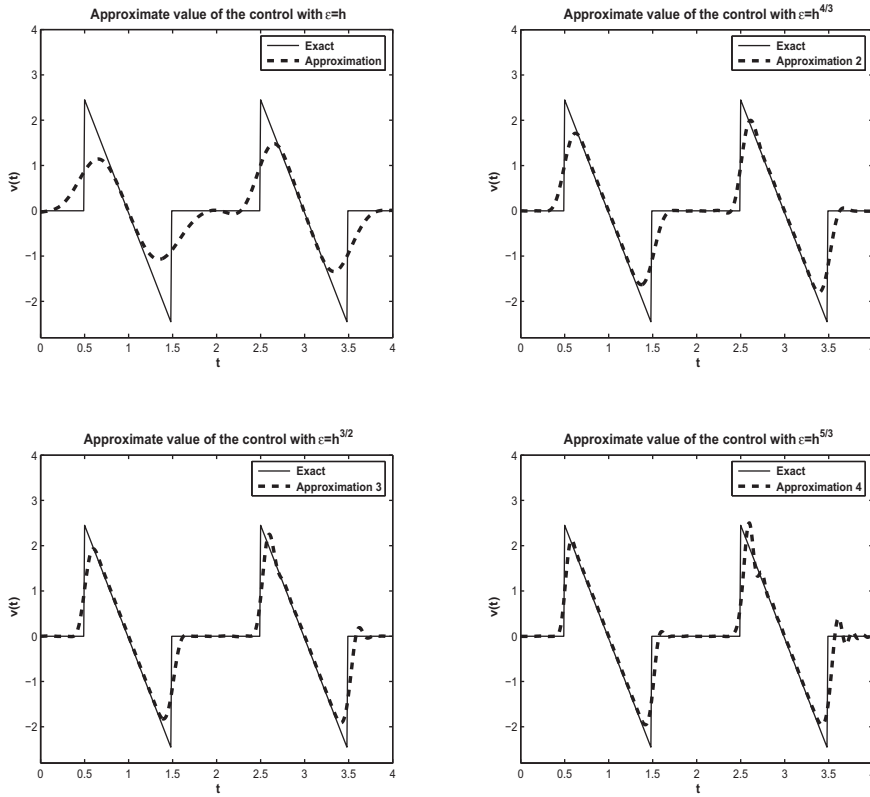


FIG. 4. Approximations of the control with different values of the parameter  $\varepsilon$  when  $h = \frac{1}{100}$ .

The first product in  $\Upsilon_m$  is evaluated in (5.9), and we have that

$$\left| \prod_{\substack{1 \leq |n| \leq N \\ n \neq m}} \frac{\mu_n}{\mu_m - \mu_n} \right| = \cos \left( \frac{m\pi h}{2} \right) \leq 1.$$

For the third product in  $\Upsilon_m$  we have that

$$\begin{aligned} \left| \prod_{1 \leq |n| \leq N} \exp \left( -\frac{x + i\mu_m}{i\mu_n} \right) \right| &= \prod_{1 \leq |n| \leq N} \left| \exp \left( \frac{ix}{\mu_n} \right) \right| \prod_{1 \leq |n| \leq N} \left| \exp \left( -\frac{\mu_m}{\mu_n} \right) \right| \\ &= \exp(-Nh \Re(\mu_m)) < 1. \end{aligned}$$

We pass to evaluate the second product in  $\Upsilon_m$ . We denote

$$P(x) = \prod_{\substack{1 \leq |n| \leq N \\ n \neq |m|}} \left| 1 + \frac{x}{i\mu_n} \right|^2 = \prod_{\substack{1 \leq n \leq N \\ n \neq |m|}} \frac{\left( x^2 - \frac{4}{h^2} \sin^2 \left( \frac{n\pi h}{2} \right) \right)^2 + \frac{16x^2}{h^2} \sin^4 \left( \frac{n\pi h}{2} \right)}{\frac{16}{h^4} \sin^4 \left( \frac{n\pi h}{2} \right)},$$

and we consider the cases  $|x| \leq \pi$ ,  $|x| > \frac{1}{2h}$ , and  $\pi < |x| \leq \frac{1}{2h}$ .

Case 1 ( $|x| \leq \pi$ ). In this case we have that

$$\begin{aligned} P(x) &\leq \prod_{\substack{1 \leq n \leq N \\ n \neq |m|}} \frac{\left(\pi^2 + \frac{4}{h^2} \sin^2\left(\frac{n\pi h}{2}\right)\right)^2 + \frac{16\pi^2}{h^2} \sin^4\left(\frac{n\pi h}{2}\right)}{\frac{16}{h^4} \sin^4\left(\frac{n\pi h}{2}\right)} \\ &\leq \prod_{\substack{1 \leq n \leq N \\ n \neq |m|}} \left( \left(1 + \frac{\pi^2}{\frac{4}{h^2} \sin^2\left(\frac{n\pi h}{2}\right)}\right)^2 + \pi^2 h^2 \right) \leq \prod_{\substack{1 \leq n \leq N \\ n \neq |m|}} \left( \left(1 + \frac{\pi^2}{4n^2}\right)^2 + \pi^2 h^2 \right) \\ &\leq \prod_{\substack{1 \leq n \leq N \\ n \neq |m|}} \left(1 + \pi h + \frac{\pi^2}{4n^2}\right)^2 \leq \exp \left( 2 \sum_{\substack{1 \leq n \leq N \\ n \neq |m|}} \left(\pi h + \frac{\pi^2}{4n^2}\right) \right). \end{aligned}$$

Hence,

$$(8.1) \quad P(x) \leq \exp(2\pi + \pi^2) \quad \forall |x| \leq \pi.$$

Case 2 ( $|x| > \frac{1}{2h}$ ). We have that

$$\begin{aligned} P(x) &\leq \prod_{\substack{1 \leq n \leq N \\ n \neq |m|}} \frac{x^4 + \frac{16}{h^4} \sin^4\left(\frac{n\pi h}{2}\right) + \frac{16x^2}{h^2} \sin^4\left(\frac{n\pi h}{2}\right)}{\frac{16}{h^4} \sin^4\left(\frac{n\pi h}{2}\right)} \leq \prod_{\substack{1 \leq n \leq N \\ n \neq |m|}} \frac{x^4 + \frac{16}{h^4} + \frac{16x^2}{h^2}}{\frac{16}{h^4} \sin^4\left(\frac{n\pi h}{2}\right)} \\ &\leq \prod_{\substack{1 \leq n \leq N \\ n \neq |m|}} \frac{x^4 (1 + 256 + 64)}{\frac{16}{h^4} \sin^4\left(\frac{n\pi h}{2}\right)} \leq \prod_{\substack{1 \leq n \leq N \\ n \neq |m|}} \frac{256x^4}{n^4} = \exp \left( 4 \sum_{\substack{1 \leq n \leq N \\ n \neq |m|}} \ln \left( \frac{4|x|}{n} \right) \right) \\ &\leq \exp \left( 4 \int_0^N \ln \left( \frac{4|x|}{s} \right) ds \right) = \exp \left( 4N \ln \left( \frac{4|x|}{N} \right) + 4N \right) \leq \exp \left( 8N \sqrt{\frac{4|x|}{N}} \right), \end{aligned}$$

where we have used that  $\frac{4|x|}{N} > 2$  and  $\ln(t) + 1 \leq 2\sqrt{t}$  for any  $t > 1$ . It follows that

$$(8.2) \quad P(x) \leq \exp \left( 16 \sqrt{\frac{|x|}{h}} \right), \quad \forall |x| > \frac{1}{2h}.$$

Case 3 ( $\pi \leq |x| \leq \frac{1}{2h}$ ). The analysis of this case is much more difficult. First of all let us note that, in view of the particular form of  $P(x)$ , it is sufficient to consider  $x > 0$ . Moreover, there exists a unique  $x_h \in (1, N)$  such that

$$(8.3) \quad x = \frac{2}{h} \sin \left( \frac{x_h \pi h}{2} \right),$$

and let  $p^* \in [1, N] \cap \mathbb{N}$  be the nearest integer to  $x_h$ . We have that

$$(8.4) \quad 2x_h \leq x \leq \pi x_h,$$

$$(8.5) \quad p^* - \frac{1}{2} \leq x_h \leq p^* + \frac{1}{2},$$

$$(8.6) \quad \frac{1}{2}x_h \leq p^* \leq x_h + \frac{1}{2} \leq \frac{x+1}{2} \leq \frac{1}{4h} + \frac{1}{2}.$$

We write  $P(x) = P_1(x)P_2(x)P_3(x)$ , where

$$P_1(x) = \frac{\left(x^2 - \frac{4}{h^2} \sin^2\left(\frac{p^*\pi h}{2}\right)\right)^2 + \frac{16x^2}{h^2} \sin^4\left(\frac{p^*\pi h}{2}\right)}{\frac{16}{h^4} \sin^4\left(\frac{p^*\pi h}{2}\right)},$$

$$P_2(x) = \prod_{\substack{1 \leq n \leq N \\ n \neq p^*, |m|}} \frac{\left(x^2 - \frac{4}{h^2} \sin^2\left(\frac{n\pi h}{2}\right)\right)^2 + \frac{16x^2}{h^2} \sin^4\left(\frac{n\pi h}{2}\right)}{\left(x^2 - \frac{4}{h^2} \sin^2\left(\frac{n\pi h}{2}\right)\right)^2},$$

$$P_3(x) = \prod_{\substack{1 \leq n \leq N \\ n \neq p^*, |m|}} \frac{\left(x^2 - \frac{4}{h^2} \sin^2\left(\frac{n\pi h}{2}\right)\right)^2}{\frac{16}{h^4} \sin^4\left(\frac{n\pi h}{2}\right)}.$$

We shall evaluate  $P_1$ ,  $P_2$ , and  $P_3$  successively. First of all note that

$$\left(x^2 - \frac{4}{h^2} \sin^2\left(\frac{n\pi h}{2}\right)\right)^2 = \frac{16}{h^4} \sin^2\left(\frac{(x_h - n)\pi h}{2}\right) \sin^2\left(\frac{(x_h + n)\pi h}{2}\right),$$

and since  $(x_h + n)h \leq \frac{3}{2}$  and  $\sin^2\left(\frac{(x_h + n)\pi h}{2}\right) \geq \frac{2(x_h + n)^2 h^2}{9}$  we obtain that

$$(8.7) \quad \frac{32(x_h + n)^2(x_h - n)^2}{9} \leq \left(x^2 - \frac{4}{h^2} \sin^2\left(\frac{n\pi h}{2}\right)\right)^2 \leq \pi^4(x_h + n)^2(x_h - n)^2.$$

We have that

$$P_1(x) \leq \frac{\pi^4|x_h - p^*|^2|x_h + p^*|^2}{16(p^*)^4} + x^2 h^2 \leq \frac{9\pi^4}{16(p^*)^2}|x_h - p^*|^2 + x^2 h^2$$

from where we deduce that

$$(8.8) \quad P_1(x) \leq \frac{9\pi^6}{16x^2} + x^2 h^2.$$

Now, we evaluate  $P_2$ :

$$\begin{aligned} P_2(x) &\leq \prod_{\substack{1 \leq n \leq N \\ n \neq p^*, |m|}} \left(1 + \frac{\frac{16x^2}{h^2} \sin^4\left(\frac{n\pi h}{2}\right)}{\frac{32}{9}(x_h + n)^2(x_h - n)^2}\right) \leq \prod_{\substack{1 \leq n \leq N \\ n \neq p^*, |m|}} \left(1 + \frac{9\pi^4 x^2 h^2 n^4}{32(x_h + n)^2(x_h - n)^2}\right) \\ &\leq \prod_{\substack{1 \leq n \leq N \\ n \neq p^*, |m|}} \left(1 + \frac{9\pi^4 x^2 h^2 n^2}{32(x_h - n)^2}\right) = \prod_{\substack{1 \leq n \leq N \\ n \neq p^*, |m|}} \left(1 + \frac{r n^2}{(x_h - n)^2}\right), \end{aligned}$$

where  $r = \frac{9\pi^4 x^2 h^2}{32}$ . We have that

$$\begin{aligned}
 \prod_{\substack{1 \leq n \leq N \\ n \neq p^*, |m|}} \left( 1 + \frac{r n^2}{(x_h - n)^2} \right) &\leq \prod_{\substack{1 \leq n \leq N \\ n \neq p^*}} \left( 1 + \frac{r n^2}{(x_h - n)^2} \right) \\
 &= \prod_{\substack{1 \leq n \leq N \\ n \notin \{p^*, p^* - 1, p^* + 1\}}} \left( 1 + \frac{r n^2}{(x_h - n)^2} \right) \left( 1 + \frac{r (p^* - 1)^2}{(x_h - p^* + 1)^2} \right) \\
 &\quad \left( 1 + \frac{r (p^* + 1)^2}{(x_h - p^* - 1)^2} \right) \\
 &\leq (1 + 4r(p^* + 1)^2)^2 \prod_{\substack{1 \leq n \leq N \\ n \notin \{p^*, p^* - 1, p^* + 1\}}} \left( 1 + \frac{r n^2}{(x_h - n)^2} \right) \\
 &\leq \exp(3\pi^2 h x^2) \exp \left( \int_0^{N+1} \ln \left( 1 + \frac{r s^2}{(x_h - s)^2} \right) ds \right).
 \end{aligned}$$

Now,

$$\begin{aligned}
 \int_0^{N+1} \ln \left( 1 + \frac{r s^2}{(x_h - s)^2} \right) ds &= \left( \int_0^{x_h} + \int_{x_h}^{N+1} \right) \ln \left( 1 + \frac{r s^2}{(x_h - s)^2} \right) ds \\
 &= (N + 1 - x_h) \ln \left( 1 + \frac{r (N + 1)^2}{(N + 1 - x_h)^2} \right) \\
 &\quad + \left( \int_0^{x_h} + \int_{x_h}^{N+1} \right) \frac{2rsx_h ds}{rs^2 + (x_h - s)^2}.
 \end{aligned}$$

Moreover,

$$\begin{aligned}
 (N + 1 - x_h) \ln \left( 1 + \frac{r (N + 1)^2}{(N + 1 - x_h)^2} \right) &\leq \frac{r (N + 1)^2}{N + 1 - x_h} \leq \frac{4r}{3} (N + 1) \\
 &= \frac{3\pi^4}{8} h x^2, \int_{\frac{x_h}{1+hx_h}}^{\frac{x_h}{1-hx_h}} \frac{2rsx_h ds}{rs^2 + (s - x_h)^2} \\
 &\leq \int_{\frac{x_h}{1+hx_h}}^{\frac{x_h}{1-hx_h}} \frac{2rsx_h ds}{rs^2} = 2x_h \ln \left( \frac{1 + hx_h}{1 - hx_h} \right) \\
 &\leq 4h(x_h)^2, \left( \int_0^{\frac{x_h}{1+hx_h}} + \int_{\frac{x_h}{1-hx_h}}^{N+1} \right) \frac{2rsx_h ds}{rs^2 + (s - x_h)^2} \\
 &\leq \int_0^{\frac{x_h}{1+hx_h}} \frac{2r(x_h)^2 ds}{(s - x_h)^2} \\
 &\quad + \int_{\frac{x_h}{1-hx_h}}^{N+1} \left( \frac{2rx_h ds}{(s - x_h)} + \frac{2r(x_h)^2 ds}{(s - x_h)^2} \right) \\
 &= \frac{2r}{h} + 4rx_h \ln \left( \frac{1 - hx_h}{hx_h} \right) \\
 &\quad + 2r \frac{(1 - hx_h)^2 - h^2(x_h)^2}{h(1 - hx_h)} \\
 &\leq \frac{2r}{h} + 4\frac{r}{h} + 2\frac{r}{h} \leq \frac{9\pi^4}{4} h x^2.
 \end{aligned}$$

It follows that

$$(8.9) \quad P_2(x) \leq \exp\left(\frac{3\pi^4}{8}hx^2 + \frac{4}{3}hx^2 + \frac{9\pi^4}{4}hx^2\right) \leq \exp(3\pi^4hx^2).$$

Finally, we evaluate  $P_3$ . First, note that if  $m \neq p^*$ ,

$$\left(\frac{\frac{4}{h^2}\sin^2\left(\frac{m\pi h}{2}\right)}{x^2 - \frac{4}{h^2}\sin^2\left(\frac{m\pi h}{2}\right)}\right)^2 \leq \frac{9m^4\pi^4}{32(x_h + m)^2(x_h - m)^2} \leq \frac{9\pi^4 m^2}{32(x_h - m)^2} \leq \frac{9\pi^4 x^2}{8}.$$

Therefore,

$$P_3(x) \leq \frac{9\pi^4 x^2}{8} \prod_{\substack{1 \leq n \leq N \\ n \neq p^*}} \frac{\sin^2\left(\frac{(x_h - n)\pi h}{2}\right) \sin^2\left(\frac{(x_h + n)\pi h}{2}\right)}{\sin^4\left(\frac{n\pi h}{2}\right)}.$$

We have that  $x_h = p^* + \alpha$ , with  $\alpha \in [-\frac{1}{2}, \frac{1}{2}]$ , and we evaluate

$$\begin{aligned} E(x) &= \prod_{\substack{1 \leq n \leq N \\ n \neq p^*}} \frac{\sin^2\left(\frac{(p^* + \alpha - n)\pi h}{2}\right) \sin^2\left(\frac{(p^* + \alpha + n)\pi h}{2}\right)}{\sin^4\left(\frac{n\pi h}{2}\right)} \\ &= \frac{\prod_{\substack{1 \leq k \leq 2p^* - 1 \\ k \neq p^*}} \sin^2\left(\frac{(k + \alpha)\pi h}{2}\right) \prod_{k=1}^{N-p^*} \sin^2\left(\frac{(k - \alpha)\pi h}{2}\right) \prod_{k=2p^*+1}^{N+p^*} \sin^2\left(\frac{(k + \alpha)\pi h}{2}\right)}{\prod_{\substack{1 \leq k \leq N \\ k \neq p^*}} \sin^4\left(\frac{k\pi h}{2}\right)} \\ &= \frac{\sin^4\left(\frac{p^*\pi h}{2}\right)}{\sin^2\left(\frac{(2p^* + \alpha)\pi h}{2}\right) \sin^2\left(\frac{(p^* + \alpha)\pi h}{2}\right)} \\ &\quad \times \prod_{k=1}^N \frac{\sin^2\left(\frac{(k + \alpha)\pi h}{2}\right) \sin^2\left(\frac{(k - \alpha)\pi h}{2}\right)}{\sin^4\left(\frac{k\pi h}{2}\right)} \prod_{k=N-p^*+1}^N \frac{\sin^2\left(\frac{(k + p^* + \alpha)\pi h}{2}\right)}{\sin^2\left(\frac{(k - \alpha)\pi h}{2}\right)}. \end{aligned}$$

By taking into account that  $\sin(a)\sin(b) \leq \sin^2\left(\frac{a+b}{2}\right)$  and  $2p^* + 1 < \frac{1}{2h}$ , we deduce that

$$E(x) \leq \frac{\sin^2\left(\frac{p^*\pi h}{2}\right)}{\sin^2\left(\frac{(p^* + \alpha)\pi h}{2}\right)} \prod_{k=N-p^*+1}^N \frac{\sin^2\left(\frac{(k + p^* + \alpha)\pi h}{2}\right)}{\sin^2\left(\frac{(k - \alpha)\pi h}{2}\right)}.$$

Since  $0 \leq \frac{p^* + \alpha}{2}\pi h \leq \pi$  and the function  $h(x) = \frac{\sin x}{x}$  is decreasing on  $[0, \pi]$ , it follows that

$$\frac{\sin\left(\frac{(k + p^* + \alpha)\pi h}{2}\right)}{\sin\left(\frac{(k - \alpha)\pi h}{2}\right)} \leq \frac{k + p^* + \alpha}{k - \alpha}$$

and thus

$$\begin{aligned}
 E(x) &\leq 2 \prod_{k=N-p^*+1}^N \left( \frac{k+p^*+\alpha}{k-\alpha} \right)^2 = 2 \exp \left( 2 \sum_{k=N-p^*+1}^N \ln \left( 1 + \frac{p^*+2\alpha}{k-\alpha} \right) \right) \\
 &\leq 2 \exp \left( 2 \int_{N-p^*}^N \ln \left( 1 + \frac{p^*+2\alpha}{s-\alpha} \right) ds \right) \leq 2 \exp \left( 2(p^*+2\alpha) \int_{N-p^*}^N \frac{ds}{s-\alpha} \right) \\
 &= 2 \exp \left( 2(p^*+2\alpha) \ln \left( \frac{N-\alpha}{N-p^*-\alpha} \right) \right) \leq 2 \exp \left( 2(p^*+2\alpha) \frac{p^*}{N-p^*-\alpha} \right) \\
 &\leq 2 \exp \left( 4(p^*)^2 \frac{1}{N - \frac{N+1}{4} - 1} \right) \leq 2 \exp \left( 16(p^*)^2 \frac{1}{3N-5} \right).
 \end{aligned}$$

Hence,

$$(8.10) \quad P_3(x) \leq \frac{9\pi^4 x^2}{4} \exp(16hx^2).$$

Finally from (8.8), (8.9), and (8.10) we obtain that

$$P(x) \leq \frac{9\pi^4 x^2}{4} \left( \frac{9\pi^6}{16x^2} + x^2 h^2 \right) \exp(3\pi^4 hx^2 + 16hx^2) \leq \frac{81\pi^{10}}{32} \exp(4\pi^4 hx^2).$$

The proof is complete, with  $C_1 = \frac{81\pi^{10}}{32}$  and  $\omega_1 = 4\pi^4$ .  $\square$

**Acknowledgments.** The author is grateful to Enrique Zuazua for several discussions and suggestions related to this work and for encouraging him to study this problem. Also, the author thanks Carlos Castro for invaluable help concerning the program and numerical experiments.

## REFERENCES

- [1] S. A. AVDONIN AND S. A. IVANOV, *Families of Exponentials. The Method of Moments in Controllability Problems for Distributed Parameter Systems*, Cambridge University Press, New York, 1995.
- [2] C. CASTRO AND S. MICU, *Boundary controllability of a linear semi-discrete 1-D wave equation derived from a mixed finite element method*, Numer. Math., 102 (2006), pp. 413–462.
- [3] J. M. CORON AND S. GUERRERO, *Singular optimal control: A linear 1-D parabolic-hyperbolic example*, Asymptot. Anal., 44 (2005), pp. 237–257.
- [4] R. J. DiPERNA, *Convergence of approximate solutions to conservation laws*, Arch. Ration. Mech. Anal., 82 (1983), pp. 27–70.
- [5] H. O. FATTORINI, *Estimates for sequences biorthogonal to certain complex exponentials and boundary control of the wave equation*, in New Trends in Systems Analysis, Lecture Notes in Control and Inform. Sci., Springer, Berlin, 1977, pp. 111–124.
- [6] H. O. FATTORINI AND D. L. RUSSELL, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Ration. Mech. Anal., 4 (1971), pp. 272–292.
- [7] H. O. FATTORINI AND D. L. RUSSELL, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Q. J. Appl. Math., 32 (1974), pp. 45–69.
- [8] R. GLOWINSKI, *Ensuring well-posedness by analogy; Stokes problem and boundary control for the wave equation*, J. Comput. Phys., 103 (1992), pp. 189–221.
- [9] R. GLOWINSKI, W. KINTON, AND M. F. WHEELER, *A mixed finite element formulation for the boundary controllability of the wave equation*, Internat. J. Numer. Methods Engrg., 27 (1989), pp. 623–636.
- [10] R. GLOWINSKI, C. H. LI, AND J.-L. LIONS, *A numerical approach to the exact boundary controllability of the wave equation (I). Dirichlet controls: Description of the numerical methods*, Japan J. Appl. Math., 7 (1990), pp. 1–76.



- [11] R. GLOWINSKI AND J.-L. LIONS, *Exact and approximate controllability for distributed parameter systems*, Acta Numer., (1996), pp. 159–333.
- [12] S. HANSEN, *Bounds on functions biorthogonal to sets of complex exponentials; control of damped elastic systems*, J. Math. Anal. Appl., 158 (1991), pp. 487–508.
- [13] J. A. INFANTE AND E. ZUAZUA, *Boundary observability for the space semi-discretization of the 1-D wave equation*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 407–438.
- [14] A. E. INGHAM, *A note on Fourier transform*, J. London Math. Soc., 9 (1934), pp. 29–32.
- [15] E. ISAAKSON AND H. B. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1996.
- [16] J.-L. LIONS, *Contrôlabilité exacte perturbations et stabilisation de systèmes distribués*, Tome 1, Masson, Paris, 1988.
- [17] A. MAJDA AND S. OSHER, *Numerical viscosity and the entropy condition*, Comm. Pure Appl. Math., 32 (1979), pp. 797–838.
- [18] S. MICU, *Uniform boundary controllability of a semi-discrete 1-D wave equation*, Numer. Math., 91 (2002), pp. 723–768.
- [19] A. MÜNCH, *A uniformly controllable and implicate scheme for the 1-D wave equation*, ESAIM M2AN Math. Model. Numer. Anal., 39 (2005), pp. 377–418.
- [20] M. NEGREANU AND E. ZUAZUA, *Discrete Ingham inequalities and applications*, SIAM J. Numer. Anal., 44 (2006), pp. 412–448.
- [21] M. NEGREANU AND E. ZUAZUA, *Uniform boundary controllability of a discrete 1-D wave equation*, Systems Control Lett., 48 (2003), pp. 261–280.
- [22] K. RAMDANI, T. TAKAHASHI, AND M. TUCSNAK, *Internal stabilization of the plate equation in a square: The continuous and the semi-discretized problems*, J. Math. Pures Appl., 85 (2006), pp. 17–37.
- [23] K. RAMDANI, T. TAKAHASHI, AND M. TUCSNAK, *Uniformly exponentially stable approximations for a class of second order evolution equations - Application to LQR problems*, ESAIM Control Optim. Calc. Var., 13 (2007), pp. 503–527.
- [24] L. ROSIER AND P. ROUCHON, *On the controllability of a wave equation with structural damping*, Int. J. Tomogr. Stat., 5 (2007), pp. 79–84.
- [25] D. L. RUSSELL, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
- [26] L. R. TCHEUGOUÉ TÉBOU AND E. ZUAZUA, *Uniform exponential long time decay for the space semi-discretization of a locally damped wave equation via an artificial numerical viscosity*, Numer. Math., 95 (2003), pp. 563–598.
- [27] L. R. TCHEUGOUÉ TÉBOU AND E. ZUAZUA, *Uniform boundary stabilization of the finite difference space discretization of the 1 - d wave equation*, Adv. Comput. Math., 26 (2007), pp. 337–365.
- [28] X. ZHANG AND E. ZUAZUA, *Polynomial decay and control of a 1-d hyperbolic-parabolic coupled system*, J. Differential Equations, 204 (2004), pp. 380–438.
- [29] E. ZUAZUA, *Boundary observability for the finite difference space semi-discretizations of the 2-D wave equation in the square*, J. Math. Pures Appl., 78 (1999), pp. 523–563.
- [30] E. ZUAZUA, *Propagation, observation, and control of waves approximated by finite difference methods*, SIAM Rev., 47 (2005), pp. 197–243.
- [31] R. M. YOUNG, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

## ON THE LINEAR-EXPONENTIAL FILTERING PROBLEM FOR GENERAL GAUSSIAN PROCESSES\*

M. L. KLEPTSZYNA<sup>†</sup>, A. LE BRETON<sup>‡</sup>, AND M. VIOT<sup>‡</sup>

**Abstract.** The explicit solution of the filtering problem with exponential criteria for a general Gaussian signal is obtained through an approach which is based on a conditional Cameron–Martin-type formula. This key formula is derived for conditional expectations of exponentials of some quadratic functionals of a general continuous Gaussian process. The formula involves conditional expectations and conditional covariances in some auxiliary optimal risk-neutral filtering problem which is used in the proof. Closed form equations of the Itô–Volterra- and Riccati–Volterra-types for these ingredients are provided. Particular cases for which the results can be further elaborated are investigated.

**Key words.** Gaussian process, optimal filtering, filtering error, Riccati–Volterra equation, risk-sensitive filtering, exponential criteria

**AMS subject classifications.** Primary, 60G15; Secondary, 60G44, 62M20

**DOI.** 10.1137/070705908

**1. Introduction.** The linear exponential Gaussian (LEG) filtering problem, i.e., with an exponential of integral performance index (see definition (2) below), and the so-called risk-sensitive (RS) filtering problem (see [3] and the statement (64) below) have been given a great deal of interest over the last decades. Numerous results have been already reported in specific models, specially around Markov models, but, as far as we know, without exhibiting the relationship between these two problems. See, e.g., Whittle [24], [25], [26], Speyer, Fan, and Banavar [23], Elliott et al. [2], [6], [7], Dey and Moore [4], [5], and Bensoussan and van Schuppen [1] for contributions on this subject and related LEG and RS control problems. Therein the notion of the “information state” has been introduced without any clear probabilistic meaning for auxiliary processes which are involved, even in the Gauss–Markov case. Moreover, the method proposed in [2] does not work in a non-Markovian situation. On the other hand, the general solution for the optimal risk-neutral linear filtering problem and a Cameron–Martin-type formula for a general Gaussian process has been obtained in [10], [11], and [17]. It seems natural to use the approach proposed in [10] and [11] to derive the solution of the LEG and RS filtering problems for a general Gaussian process, to refine their link and also to give a probabilistic interpretation for the ingredients of the information state. For details, see sections 5 and 6, where in particular we prove that the LEG and RS filtering problems have the same solution and we propose an example to show that, in a bit more general setting, two similar problems may have different solutions.

In the present paper we deal with a signal process  $X = (X_t, t \geq 0)$ , which is an arbitrary  $n$ -dimensional continuous Gaussian process, and an observation process

---

\*Received by the editors October 20, 2007; accepted for publication (in revised form) July 8, 2008; published electronically November 19, 2008.

<http://www.siam.org/journals/sicon/47-6/70590.html>

<sup>†</sup>Corresponding author. Laboratoire de Statistique et Processus, Université du Maine, Av. Olivier Messiaen, 72085 Le Mans, Cedex 9, France (Marina.Kleptsyna@univ-lemans.fr).

<sup>‡</sup>Laboratoire Jean Kuntzmann, Université J. Fourier, BP 53, 38041 Grenoble, Cedex 9, France (Alain.Le-Breton@imag.fr, acf.viot@orange.fr).

$Y = (Y_t, t \geq 0)$  in  $\mathbb{R}^d$  governed by the linear equation

$$(1) \quad Y_t = \int_0^t A(s)X_s ds + \tilde{B}_t, \quad t \geq 0.$$

Here the deterministic function  $A = (A(s), s \geq 0)$  is continuous with values in the set of  $d \times n$  matrices, and  $\tilde{B} = (\tilde{B}_t, t \geq 0)$  denotes a  $d$ -dimensional standard Brownian motion independent of  $X$ . Clearly the pair  $(X, Y)$  is Gaussian but, in the general setting, neither Markovian nor a semimartingale.

Suppose that only  $Y$  is observed and one wishes to minimize with respect to  $h \in \mathcal{H}$  and  $g \in \mathcal{G}$  the following quantity:

$$(2) \quad L_T(h, g, \mu) = \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} (X_T - g)' M (X_T - g) + \frac{\mu}{2} \int_0^T (X_s - h(s))' Q_s (X_s - h(s)) ds \right\} \right],$$

where  $h = (h(s), 0 \leq s \leq T) \in \mathcal{H}$  means that  $h$  is a  $(\mathcal{Y}_s)$ -adapted process and  $g \in \mathcal{G}$  means that  $g$  is a  $\mathcal{Y}_T$ -measurable variable. Here nonnegative definite symmetric deterministic matrices  $M$  and  $Q_s, 0 \leq s \leq T$ , are given, and  $(\mathcal{Y}_s)$  is the natural filtration of  $Y$ , i.e.,  $\mathcal{Y}_s = \sigma(\{Y_u, 0 \leq u \leq s\}), 0 \leq s \leq T$ . Note that, according to the sign of the real parameter  $\mu$ , there are two different cases for this LEG filtering problem (the terminology is taken from the LEG optimal control problem):

- $\mu < 0$ , called the *risk-preferring* filtering problem,
- $\mu > 0$ , called the *risk-averse* filtering problem.

It is well-known (see, e.g., [23] for the discrete time Markov case setting) that the solution to this problem is not the conditional expectation of  $X_t$  given the  $\sigma$ -field  $\mathcal{Y}_t$ . Our first aim is to show that the solution can be completely explicit: the characteristics of the optimal solution are obtained as the solution of a closed form system of Volterra-type equations which actually reduce to the equations known for the RS setting when the signal process  $X$  is Gauss–Markov (see, e.g., [20]). Our second aim is to give the probabilistic interpretation of this optimal solution in terms of an auxiliary optimal *risk-neutral* filtering problem. Actually, we extend the filtering approach initiated in [11] for one-dimensional (1D) processes to obtain a conditional Cameron–Martin-type formula for the *conditional Laplace transform* of a quadratic functional of the involved process. Namely, we give an explicit representation for the random variable

$$(3) \quad \mathcal{L}_T = \mathbb{E} \left[ \exp \left\{ \frac{\mu}{2} (X_T - g)' M (X_T - g) + \frac{\mu}{2} \int_0^T (X_s - h(s))' Q_s (X_s - h(s)) ds \right\} \middle/ \mathcal{Y}_T \right].$$

The paper is organized as follows. In section 2 we derive the solution of the LEG filtering problem in the case  $M = 0$ ; explicit Itô–Volterra-type equations, involving the covariance function of the filtered process, are obtained. In particular, in section 2.1, an appropriate auxiliary risk-neutral filtering problem is matched to that of deriving the key Cameron–Martin-type formula for the *conditional Laplace transform* of a quadratic functional of the involved process. The solution of this auxiliary filtering

problem is discussed in section 2.2. In section 3 we investigate some specific cases where the results can be further elaborated. In particular, more general observation models where the observation noise is not a Brownian motion or it is not independent of the signal are considered. Section 4 is devoted to the case  $M$  nonnegative definite. Of course, the results of section 2 are particular cases of those of section 4, but we prefer to keep both sections because a really simple independent proof can be given when  $M = 0$ . In section 5 we discuss the relation between LEG and RS filtering problems. Finally, section 6 is devoted to the interpretation for the ingredients of the information state.

**2. Solution of the LEG filtering problem: The case  $M = 0$ .** It what follows all random variables and processes are defined on a given stochastic basis  $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$  satisfying the usual conditions, and processes are  $(\mathcal{F}_t)$ -adapted. We consider a  $\mathbb{R}^n$ -valued continuous Gaussian process  $X = (X_t, t \geq 0)$  with mean function  $m = (m_t, t \geq 0)$  and covariance function  $\Gamma = (\Gamma(t, s), t \geq 0, s \geq 0)$ , i.e.,

$$\mathbb{E}X_t = m_t, \quad \mathbb{E}(X_t - m_t)(X_s - m_s)' = \Gamma(t, s), \quad t \geq 0, s \geq 0.$$

Let us introduce the following condition  $(C_\mu)$ .

$(C_\mu)$  the Riccati–Volterra equation

$$(4) \quad \bar{\gamma}(t, s) = \Gamma(t, s) - \int_0^s \bar{\gamma}(t, r)[A_r' A_r - \mu Q_r] \bar{\gamma}'(s, r) dr$$

has a unique and bounded solution on  $\{(t, s) : 0 \leq s \leq t \leq T\}$  such that for  $0 \leq t \leq T$  the matrix  $\bar{\gamma}(t, t)$  is nonnegative definite.

Notice that for all  $\mu$  *negative* the condition  $(C_\mu)$  is satisfied (cf. Theorem 2 below), and if  $\mu$  is *positive*, the condition  $(C_\mu)$  is satisfied for  $\mu$  sufficiently small, for example, those such that for any  $t \leq T$ ,  $A_t' A_t - \mu Q_t$  is nonnegative definite.

The first result of the present paper is the following theorem.

**THEOREM 1.** *Suppose that the condition  $(C_\mu)$  is satisfied. Let  $\bar{h} = (\bar{h}(t), 0 \leq t \leq T)$  be the unique solution of the Itô–Volterra equation*

$$(5) \quad \bar{h}(t) = m_t + \int_0^t \bar{\gamma}(t, s) A_s' [dY_s - A_s \bar{h}(s) ds],$$

where  $\bar{\gamma}$  is the unique solution of (4).

Then  $\bar{h} = (\bar{h}(t), 0 \leq t \leq T)$  is the solution of the LEG filtering problem, i.e.,

$$(6) \quad \bar{h} = \arg \min_{h \in \mathcal{H}} \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} \int_0^T (X_s - h(s))' Q_s (X_s - h(s)) ds \right\} \right].$$

Moreover, the corresponding optimal risk is given by

$$\mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} \int_0^T (X_s - \bar{h}(s))' Q_s (X_s - \bar{h}(s)) ds \right\} \right] = \mu \exp \left\{ \frac{\mu}{2} \int_0^T \text{tr}(\bar{\gamma}(s, s) Q_s) ds \right\}.$$

Theorem 1 is a direct consequence of the results of section 2.1. Its proof will be given at the end of section 2.1.

*Remark 1.*

- (i) When  $\mu$  is *negative*, functions  $\bar{h}$  and  $\bar{\gamma}$  can be interpreted in terms of an appropriate *risk-neutral* filtering problem (see the proofs of Proposition 1 and Theorem 1), but when  $\mu$  is *positive*, there is no such connection anymore.

- (ii) Let us note that in the Markovian case, (4) and (5) reduce to those given in Elliott, Aggoun, and Moore [7] for the solution of the risk-sensitive filtering problem. See also sections 3 and 5 for more details.
- (iii) It is worth emphasizing that, taking  $\mu = \mathbf{0}$  in (4), one gets through (5) the solution  $\bar{h}$  of the *risk-neutral* filtering problem of the signal  $X$ , given the observation  $Y$ , i.e.,  $\bar{h}(t) = \mathbb{E}(X_t/\mathcal{Y}_t)$  (see, e.g., [10]).

**2.1. Conditional version of a Cameron–Martin formula: The case  $M = \mathbf{0}$ .** In [10] and [11] it was proved that for any  $\mathbb{R}^n$ -valued continuous Gaussian process  $X = (X_t, t \geq 0)$ , with mean function  $m = (m_t, t \geq 0)$  and covariance function  $\Gamma = (\Gamma(t, s), t \geq 0, s \geq 0)$ , the following equality holds for any  $T > 0$  and  $\mu$  negative :

$$(7) \quad \mathbb{E} \exp \left\{ \frac{\mu}{2} \int_0^T X'_s Q_s X_s ds \right\} = \exp \left\{ \frac{\mu}{2} \int_0^T [z'(s) Q_s z(s) + \text{tr}(\gamma(s, s) Q_s)] ds \right\},$$

where  $\gamma = (\gamma(t, s), 0 \leq s \leq t \leq T)$  is the unique solution of the Riccati–Volterra equation

$$(8) \quad \gamma(t, s) = \Gamma(t, s) + \int_0^s \gamma(t, u) \mu Q_u \gamma'(s, u) du, \quad 0 \leq s \leq t \leq T,$$

such that  $\gamma(t, t)$  is nonnegative definite and  $z = (z_s, 0 \leq s \leq T)$  is the unique solution of the integral equation

$$(9) \quad z_s = m_s + \int_0^s \gamma(s, u) \mu Q_u z_u du.$$

Moreover, in [10] and [11] the link between the computation of the Laplace transform (7) and the resolution of an appropriate *risk-neutral* filtering problem was exhibited, and hence the probabilistic interpretation of functions  $z$  and  $\gamma$  was established. More precisely, new processes  $\bar{Y} = (\bar{Y}_t, t \geq 0)$  and  $\bar{\xi} = (\bar{\xi}_t, t \geq 0)$  were defined by

$$(10) \quad d\bar{Y}_t = (-\mu Q_t)^{\frac{1}{2}} X_t dt + d\bar{B}_t, \quad t \geq 0,$$

and

$$(11) \quad d\bar{\xi}_t = X'_t (-\mu Q_t)^{\frac{1}{2}} d\bar{Y}_t, \quad t \geq 0,$$

where  $\bar{B} = (\bar{B}_t; t \in [0, T])$  is a Brownian motion, independent of  $X$ . It was established that

- $\gamma(t, t)$  is nothing but the covariance of the filtering error of  $X$ , given  $\bar{Y}$ , i.e.,

$$(12) \quad \gamma(t, t) = \gamma_{XX}(t) = \mathbb{E}[X_t - \bar{\pi}_t(X)][X_t - \bar{\pi}_t(X)]',$$

and

- $z_t = \bar{\pi}_t(X) - \gamma_{X\bar{\xi}}(t)$ , with

$$(13) \quad \gamma_{X\bar{\xi}}(t) = \mathbb{E}[(X_t - \bar{\pi}_t(X))(\bar{\xi}_t - \bar{\pi}_t(\bar{\xi})) / \bar{\mathcal{Y}}_t],$$

where for any process  $\eta = (\eta_t; t \in [0, T])$  such that  $\mathbb{E}|\eta_t| < +\infty$ , the notation  $\bar{\pi}_t(\eta)$  is used for the conditional expectation of  $\eta_t$ , given the *auxiliary*  $\sigma$ -field  $\bar{\mathcal{Y}}_t = \sigma(\{\bar{Y}_s, 0 \leq s \leq t\})$ ,

$$\bar{\pi}_t(\eta) = \mathbb{E}(\eta_t / \bar{\mathcal{Y}}_t).$$

*Remark 2.* Let us mention that, since the pair  $(X, \xi)$  is only conditionally Gaussian given  $\bar{\mathcal{Y}}_t$ , the conditional covariance  $\gamma_{X\xi}(t)$  is random as well as  $\bar{\pi}_t(X)$ . But the difference  $z_t = \bar{\pi}_t(X) - \gamma_{X\xi}(t)$ , which is the solution of (9), is actually deterministic.

*Remark 3.* If  $\mu$  is *positive*, but sufficiently small in order that (8) has a unique and bounded solution on  $\{(t, s) : 0 \leq s \leq t \leq T\}$ , due to analytical properties of involved functions with respect to  $\mu$ , equality (7) is still valid. But it is worth emphasizing that there is no connection anymore between functions  $z$  and  $\bar{\gamma}$  and a risk-neutral filtering problem.

In this section we are interested in a conditional version of (7). We assume that the observation process  $Y$  satisfies (1). We fix some arbitrary  $(\mathcal{Y}_t)$ -adapted  $\mathbb{R}^n$ -valued continuous process  $h = (h(t), t \geq 0)$  and some continuous deterministic function  $Q_t$  with values in the set of nonnegative definite symmetric  $n \times n$  matrices. We denote  $G_t$  by

$$(14) \quad G_t = \exp \left\{ \frac{\mu}{2} \int_0^t (X_s - h(s))' Q_s (X_s - h(s)) ds \right\}$$

or, in the differential form,  $dG_t = \frac{\mu}{2} G_t (X_t - h(t))' Q_t (X_t - h(t)) dt$ . We denote by  $\mathbb{J}_T$  the conditional expectation of  $G_T$ , given the  $\sigma$ -field  $\mathcal{Y}_T$ :

$$(15) \quad \mathbb{J}_T = \mathbb{E} \left[ \exp \left\{ \frac{\mu}{2} \int_0^T (X_s - h(s))' Q_s (X_s - h(s)) ds \right\} \middle/ \mathcal{Y}_T \right] = \pi_T(G),$$

where for any process  $\eta = (\eta_t, t \in [0, T])$  such that  $\mathbb{E}|\eta_t| < +\infty$ , the notation  $\pi_t(\eta)$  is used for the conditional expectation of  $\eta_t$ , given the  $\sigma$ -field  $\mathcal{Y}_t = \sigma(\{Y_s, 0 \leq s \leq t\})$ ,

$$\pi_t(\eta) = \mathbb{E}(\eta_t / \mathcal{Y}_t).$$

Then we can state the following proposition.

**PROPOSITION 1.** *Suppose that the condition  $(C_\mu)$  is satisfied. Let  $z^h = (z_s^h, 0 \leq s \leq T)$  be the unique solution of the Itô–Volterra equation*

$$(16) \quad z_t^h = m_t + \int_0^t \bar{\gamma}(t, s) \mu Q_s [z_s^h - h(s)] ds + \int_0^t \bar{\gamma}(t, s) A'_s [dY_s - A_s z_s^h ds],$$

where  $\bar{\gamma}$  is the unique solution of (4).

Then the following equality holds:

$$(17) \quad \mathbb{J}_T = \exp \left\{ \frac{\mu}{2} \int_0^T (z_s^h - h(s))' Q_s (z_s^h - h(s)) ds + \frac{\mu}{2} \int_0^T \text{tr}(\bar{\gamma}(s, s) Q_s) ds \right. \\ \left. + \int_0^T (z_s^h - \pi_s(X))' A'_s d\nu_s - \frac{1}{2} \int_0^T \|A_s (z_s^h - \pi_s(X))\|^2 ds \right\},$$

where  $(\nu_t, t \geq 0)$  is the innovation process associated to  $Y$ , i.e.,

$$(18) \quad d\nu_t = dY_t - A_t \pi_t(X) dt, \quad \nu_0 = 0.$$

*Remark 4.* For  $\mu$  *negative* the probabilistic interpretation of functions  $z_t^h$  and  $\bar{\gamma}$  will be clarified below in terms of the solution of an auxiliary *risk-neutral* filtering problem.

*Remark 5.* Let us note that if  $A_t \equiv 0$  and  $h_t \equiv 0$ , then the conditional expectation is nothing else but the ordinary expectation, and the result of Proposition 1 reduces to that of [10] and [11] (see (7)–(9) above).

*Proof of Proposition 1.* Actually, for our goals it is sufficient to work with  $\mu < 0$  since the result will be valid for sufficiently small  $\mu > 0$ , because of the analytical properties of the involved functions. To simplify the notations we work with  $\mu = -1$ ; then for the general situation it is sufficient to replace  $Q$  by  $-\mu Q$ .

In this proof we follow directly the idea of [11] for the ordinary Cameron–Martin formula for general Gaussian processes. Let us introduce the auxiliary pair of processes  $\bar{Y}_t = (Y_t^1, Y_t^2)$  such that

$$(19) \quad \begin{cases} dY_t^1 = dY_t, & Y_0^1 = 0, \\ dY_t^2 = Q_t^{\frac{1}{2}}(X_t - h(t))dt + d\bar{B}_t, & Y_0^2 = 0, \end{cases}$$

where  $\bar{B}_t$  is a standard Brownian motion, independent of  $(X, \tilde{B})$ . We emphasize that the first component  $Y^1$  is exactly our observation process.

Below, for any process  $\eta = (\eta_t; t \in [0, T])$  such that  $\mathbb{E}|\eta_t| < +\infty$ , again the notation  $\bar{\pi}_t(\eta)$  is used for the conditional expectation of  $\eta_t$ , given the *new auxiliary*  $\sigma$ -field  $\bar{\mathcal{Y}}_t = \sigma(\{\bar{Y}_s, 0 \leq s \leq t\})$ ,

$$\bar{\pi}_t(\eta) = \mathbb{E}(\eta_t / \bar{\mathcal{Y}}_t) .$$

Let also  $\xi_t$  be defined by

$$(20) \quad d\xi_t = (X_t - h(t))' Q_t^{\frac{1}{2}} dY_t^2, \quad \xi_0 = 0.$$

We see that the conditional distribution of  $(X_t, \xi_t)$  with respect to  $\bar{\mathcal{Y}}_t$  is Gaussian with the conditional expectation  $(\bar{\pi}_t(X), \bar{\pi}_t(\xi))$  and the conditional covariance  $\begin{pmatrix} \bar{\gamma}_{XX}(t) & \bar{\gamma}_{X\xi}(t) \\ \bar{\gamma}_{X\xi}(t)' & \bar{\gamma}_{\xi\xi}(t) \end{pmatrix}$ , where

$$(21) \quad \begin{aligned} \bar{\gamma}_{XX}(t) &= \mathbb{E}[(X_t - \bar{\pi}_t(X))(X_t - \bar{\pi}_t(X))' / \bar{\mathcal{Y}}_t] , \\ \bar{\gamma}_{X\xi}(t) &= \mathbb{E}[(X_t - \bar{\pi}_t(X))(\xi_t - \bar{\pi}_t(\xi)) / \bar{\mathcal{Y}}_t] , \end{aligned}$$

and

$$(22) \quad \bar{\gamma}_{\xi\xi}(t) = \mathbb{E}[(\xi_t - \bar{\pi}_t(\xi))(\xi_t - \bar{\pi}_t(\xi)) / \bar{\mathcal{Y}}_t] .$$

Of course, since the pair  $(X, \bar{Y})$  is Gaussian, the conditional covariance  $\bar{\gamma}_{XX}(t)$  is deterministic and is nothing but the covariance of the filtering error, i.e.,

$$(23) \quad \bar{\gamma}_{XX}(t) = \mathbb{E}[(X_t - \bar{\pi}_t(X))(X_t - \bar{\pi}_t(X))'] .$$

Now, we turn to the proof of equality (17) with  $\mu = -1$ . The general filtering theorem in [18] gives the following representation (where  $G$  is defined by (14)):

$$(24) \quad d\pi_t(G) = -\frac{1}{2}\pi_t[G(X - h)'Q(X - h)]dt + [\pi_t'(GX) - \pi_t(G)\pi_t'(X)]A_t'd\nu_t,$$

with the innovation process  $\nu_t$ . We emphasize that here  $\pi$  means again the conditional expectation with respect to a  $\sigma$ -algebra generated by the observation process

$Y$ . Representation (24) can be rewritten in the following equivalent integral form:

$$(25) \quad \pi_t(G) = \exp \left\{ -\frac{1}{2} \int_0^t \alpha_2(s) ds + \int_0^t [\alpha_1(s) - \pi_s(X)]' A'_s d\nu_s \right. \\ \left. - \frac{1}{2} \int_0^t \|A_s(\alpha_1(s) - \pi_s(X))\|^2 ds \right\},$$

with

$$\alpha_1(t) = \frac{\pi_t(XG)}{\pi_t(G)}$$

and

$$\alpha_2(t) = \frac{\pi_t(G(X-h)'Q(X-h))}{\pi_t(G)}.$$

Now, for a fixed  $t \geq 0$ , we define the random variable  $\zeta_t$  by

$$\zeta_t = \int_0^t (X_s - h(s))' Q_s^{\frac{1}{2}} d\bar{B}_s + \frac{1}{2} \int_0^t (X_s - h(s))' Q_s (X_s - h(s)) ds.$$

Since the pair  $(X, Y)$  is independent of  $\bar{B}$  it is easy to check that  $\mathbb{E}e^{-\zeta_t} = 1$ , and so we can define the new probability  $\tilde{\mathbb{P}}_t$  such that

$$(26) \quad \rho_t = \frac{d\tilde{\mathbb{P}}_t}{d\mathbb{P}} = e^{-\zeta_t}.$$

The Girsanov theorem tells us that under  $\tilde{\mathbb{P}}_t$  the distribution of the system  $((X_s, \bar{Y}_s), 0 \leq s \leq t)$  (where  $\bar{Y}$  is given by (19)) is the same as that of  $((X_s, Y_s^1, \bar{B}_s), 0 \leq s \leq t)$  under  $\mathbb{P}$ . Therefore, denoting by  $\tilde{\mathbb{E}}_t$  a conditional expectation computed with respect to  $\tilde{\mathbb{P}}_t$ , we obtain

$$\alpha_1(t) = \frac{\tilde{\mathbb{E}}_t(X_t G_t / \mathcal{Y}_t)}{\tilde{\mathbb{E}}_t(G_t / \mathcal{Y}_t)}$$

and

$$\alpha_2(t) = \frac{\tilde{\mathbb{E}}_t(G_t(X_t - h(t))' Q_t(X_t - h(t)) / \mathcal{Y}_t)}{\tilde{\mathbb{E}}_t(G_t / \mathcal{Y}_t)}.$$

Since in particular under  $\tilde{\mathbb{P}}_t$ ,  $X$  and  $Y^1$  are independent of  $\bar{B}$ , the above conditional expectations can be replaced by conditional expectations given  $\bar{\mathcal{Y}}_t$  under  $\tilde{\mathbb{P}}_t$  so that

$$\alpha_1(t) = \frac{\tilde{\mathbb{E}}_t(X_t G_t / \bar{\mathcal{Y}}_t)}{\tilde{\mathbb{E}}_t(G_t / \bar{\mathcal{Y}}_t)}$$

and

$$\alpha_2(t) = \frac{\tilde{\mathbb{E}}_t(G_t(X_t - h(t))' Q_t(X_t - h(t)) / \bar{\mathcal{Y}}_t)}{\tilde{\mathbb{E}}_t(G_t / \bar{\mathcal{Y}}_t)}.$$



But from the well-known Bayes formula of the filtering theory, these equalities can be rewritten as

$$\alpha_1(t) = \frac{\mathbb{E}(X_t G_t \rho_t / \bar{\mathcal{Y}}_t)}{\mathbb{E}(\rho_t / \bar{\mathcal{Y}}_t)} \frac{\mathbb{E}(\rho_t / \bar{\mathcal{Y}}_t)}{\mathbb{E}(G_t \rho_t / \bar{\mathcal{Y}}_t)}$$

and

$$\alpha_2(t) = \frac{\mathbb{E}(G_t \rho_t (X_t - h(t))' Q_t (X_t - h(t)) / \bar{\mathcal{Y}}_t)}{\mathbb{E}(\rho_t / \bar{\mathcal{Y}}_t)} \frac{\mathbb{E}(\rho_t / \bar{\mathcal{Y}}_t)}{\mathbb{E}(G_t \rho_t / \bar{\mathcal{Y}}_t)}.$$

Hence, since from definitions (19) and (20) we have  $e^{-\xi_t} = G_t \rho_t$ , it follows that

$$\alpha_1(t) = \frac{\mathbb{E}(X_t e^{-\xi_t} / \bar{\mathcal{Y}}_t)}{\mathbb{E}(e^{-\xi_t} / \bar{\mathcal{Y}}_t)}$$

and

$$\alpha_2(t) = \frac{\mathbb{E}(e^{-\xi_t} (X_t - h(t))' Q_t (X_t - h(t)) / \bar{\mathcal{Y}}_t)}{\mathbb{E}(e^{-\xi_t} / \bar{\mathcal{Y}}_t)}.$$

Now, we observe that (under  $\mathbb{P}$ ) the conditional distribution of  $X$ , given the  $\sigma$ -field  $\mathcal{Y}_t$ , is Gaussian, and moreover from (20) for any  $t \geq 0$ , given  $\bar{\mathcal{Y}}_t$ , the variable  $\xi_t$  is a linear functional of  $X$ . Consequently, the conditional distribution of  $(X_t, \xi_t)$ , given the  $\bar{\mathcal{Y}}_t$ , is also Gaussian. The classical properties of Gaussian vectors give the following equalities:

$$\alpha_1(t) = \frac{\mathbb{E}(X_t e^{-\xi_t} / \bar{\mathcal{Y}}_t)}{\mathbb{E}(e^{-\xi_t} / \bar{\mathcal{Y}}_t)} = \bar{\pi}_t(X) - \bar{\gamma}_{X\xi}(t)$$

and

$$\begin{aligned} \alpha_2(t) &= \frac{\mathbb{E}((X_t - h(t))' Q_t (X_t - h(t)) e^{-\xi_t} / \bar{\mathcal{Y}}_t)}{\mathbb{E}(e^{-\xi_t} / \bar{\mathcal{Y}}_t)} \\ &= [\bar{\pi}_t(X) - \bar{\gamma}_{X\xi}(t) - h(t)]' Q_t [\bar{\pi}_t(X) - \bar{\gamma}_{X\xi}(t) - h(t)] + \text{tr}(\bar{\gamma}_{XX}(t) Q_t). \end{aligned}$$

Inserting this into (25) gives immediately (17), with  $z_t^h = \bar{\pi}_t(X) - \bar{\gamma}_{X\xi}(t)$  and  $\bar{\gamma}(t, t) = \bar{\gamma}_{XX}(t)$ . To complete the proof of Proposition 1, the equations for  $z^h$  and  $\bar{\gamma}$  will be derived in section 2.2.  $\square$

*Remark 6.*

- (i) Actually, the probabilistic meaning of  $z_t^h$  and  $\bar{\gamma}(t, s)$  in terms of a risk-neutral filtering problem is valid only when  $\mu$  is negative. It is worth mentioning that  $\bar{\pi}_t(X)$  and  $\bar{\gamma}_{X\xi}(t)$  are only  $\bar{\mathcal{Y}}_t$ -measurable variables but that the difference  $z_t^h = \bar{\pi}_t(X) - \bar{\gamma}_{X\xi}(t)$  is actually a  $\mathcal{Y}_t$ -measurable variable.
- (ii) Of course, the proof above is still valid if we suppose only that the conditional distribution of the signal  $X$ , given the observation  $Y$ , is Gaussian. Hence, in this case, (17) holds but with modified equations for the ingredients  $z_t^h$  and  $\bar{\gamma}(t, s)$ .

*Proof of Theorem 1.* The statement of Theorem 1 is the direct consequence of Proposition 1. Indeed, we claim that the following chain of inequalities holds for any

$h \in \mathcal{H}$ :

$$\begin{aligned}
 & \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} \int_0^T (X_s - h(s))' Q_s (X_s - h(s)) ds \right\} \right] \\
 &= \mathbb{E} \left[ \mu \mathbb{E} \left[ \exp \left\{ \frac{\mu}{2} \int_0^T (X_s - h(s))' Q_s (X_s - h(s)) ds \right\} \middle| \mathcal{Y}_T \right] \right] \\
 &\stackrel{(a)}{\geq} \exp \left\{ \frac{\mu}{2} \int_0^T \text{tr}(\bar{\gamma}(s, s) Q_s) ds \right\} \\
 &\quad \times \mu \mathbb{E} \left[ \exp \left\{ \int_0^T (z_s^h - \pi_s(X))' A'_s d\nu_s - \frac{1}{2} \int_0^T \|A_s (z_s^h - \pi_s(X))\|^2 ds \right\} \right] \\
 &\stackrel{(b)}{\geq} \mu \exp \left\{ \frac{\mu}{2} \int_0^T \text{tr}(\bar{\gamma}(s, s) Q_s) ds \right\}.
 \end{aligned}$$

Of course, under condition  $(C_\mu)$ , since the term in the last line is finite, it is sufficient to consider the case

$$(27) \quad \mathbb{E} \left[ \exp \left\{ \frac{\mu}{2} \int_0^T (X_s - h(s))' Q_s (X_s - h(s)) ds \right\} \right] < \infty,$$

which gives the first equality. Inequality (a) follows directly from Proposition 1. For inequality (b) we consider separately the cases  $\mu < 0$  and  $\mu > 0$ .

For  $\mu < 0$ , inequality (b) is due to the submartingale property:

$$\mathbb{E} \left[ \exp \left\{ \int_0^T (z_s^h - \pi_s(X))' A'_s d\nu_s - \frac{1}{2} \int_0^T \|A_s (z_s^h - \pi_s(X))\|^2 ds \right\} \right] \leq 1.$$

For  $\mu > 0$ , (16) and conditions  $(C_\mu)$  and (27) give that the property  $\sup_{0 \leq s \leq T} \mathbb{E} [\exp \{\varepsilon \|z_s^h - \pi_s(X)\|^2\}] < \infty$  holds for some  $\varepsilon > 0$ , and hence, due to Theorem 6.1 in [18],

$$\mathbb{E} \left[ \exp \left\{ \int_0^T (z_s^h - \pi_s(X))' A'_s d\nu_s - \frac{1}{2} \int_0^T \|A_s (z_s^h - \pi_s(X))\|^2 ds \right\} \right] = 1.$$

Now, to obtain the lower bound we must take  $\bar{h} = z^{\bar{h}}$ , where  $z^h$  is the solution of (16), which means that  $\bar{h}$  is the unique solution of (5). Finally, for  $\bar{h}$  the lower bound is attained because the system  $(X, Y, \bar{h})$  is Gaussian, and hence inequalities (a) and (b) are actually equalities (see Example 3a in section 6.2 of [18]).  $\square$

**2.2. Auxiliary filtering problem.** Here we deal with the filtering problem of the signal  $(X, \xi)$ ,  $\xi$  defined in (20) from the observation of  $\bar{Y}$  defined in (19) (see [9, Chapter 10] and [10] for a similar setting).

The following statement provides equations for the conditional mean and the variance of the filtering error. Let us observe that, contrary to the Markov case, to update these quantities one needs to have access to the entire past observation record.

**THEOREM 2.** *The conditional mean  $\bar{\pi}_t(X)$  and the variance of the filtering error  $\bar{\gamma}_{XX}(t)$  are given by the equations*

$$(28) \quad \bar{\pi}_t(X) = m_t + \int_0^t \bar{\gamma}(t, s) \left[ A'_s d\bar{\nu}_s^1 + Q_s^{\frac{1}{2}} d\bar{\nu}_s^2 \right], \quad t \geq 0,$$

$$(29) \quad \bar{\gamma}_{xx}(t) = \bar{\gamma}(t, t), \quad t \geq 0,$$

where  $\bar{\gamma}$  is the unique solution of (4), with  $\mu = -1$ , such that  $\bar{\gamma}(t, t)$ ,  $t \geq 0$ , is a nonnegative definite symmetric matrix and  $\bar{\nu}_t = (\bar{\nu}_t^1, \bar{\nu}_t^2)$  is the innovation process for the pair  $(Y^1, Y^2)$ :

$$(30) \quad d\bar{\nu}_t^1 = dY_t^1 - A_t \bar{\pi}_t(X) dt, \quad d\bar{\nu}_t^2 = dY_t^2 - Q_t^{\frac{1}{2}} [\bar{\pi}_t(X) - h(t)] dt.$$

*Proof.* The difficulty is that in general  $X$  is not a semimartingale. To go over it in order to be able to apply the well-known filtering theory for semimartingales (see, e.g., [18, 19]), for a fixed  $t \geq 0$ , we introduce the process  $X^t = (X_s^t, 0 \leq s \leq t)$  as follows:

$$X_s^t = \mathbb{E}[X_t / \sigma(\{X_r, 0 \leq r \leq s\})], \quad 0 \leq s \leq t.$$

Clearly, by its definition, the process  $X^t$  is a continuous martingale (with mean  $m_t$ ) and  $X_t^t = X_t$ . Moreover, the pair  $(X, X^t)$  is Gaussian and independent of  $(B, \tilde{B})$ , and so the distribution of  $(X, X^t, Y)$  is still Gaussian. In particular, the conditional covariance  $\bar{\gamma}(t, s) = \mathbb{E}[(X_s^t - \bar{\pi}_s(X^t))(X_s - \bar{\pi}_s(X))' / \bar{\mathcal{Y}}_s]$  is deterministic. Hence, setting

$$\delta_X(s) = X_s - \bar{\pi}_s(X), \quad \delta_X^t(s) = X_s^t - \bar{\pi}_s(X^t), \quad 0 \leq s \leq t,$$

we can write

$$(31) \quad \bar{\gamma}(t, s) = \mathbb{E} \delta_X^t(s) \delta_X(s)', \quad 0 \leq s \leq t.$$

Of course, due to  $X_t^t = X_t$ , which also implies  $\delta_X^t(t) = \delta_X(t)$ , for  $s = t$  equality (31) reduces to (29).

Applying the fundamental filtering theorem to the pair  $(X^t, \bar{Y})$  of semimartingales, we obtain immediately that

$$(32) \quad \bar{\pi}_s(X^t) = m_t + \int_0^s \bar{\gamma}(t, r) \left[ A_r' d\bar{\nu}_r^1 + Q_r^{\frac{1}{2}} d\bar{\nu}_r^2 \right], \quad 0 \leq s \leq t.$$

Hence, again because  $X_t^t = X_t$  and due to definition (30) of  $\bar{\nu}$ , for  $s = t$  (32) reduces to (28).

Therefore, to complete the proof of the first part of the theorem, we need only to show that the function  $\bar{\gamma}$  which has just been defined in (31) is nothing but the unique solution of (4) such that  $\bar{\gamma}(t, t)$ ,  $t \geq 0$ , is a nonnegative symmetric matrix. From (32), using (1) and (30), we can write

$$(33) \quad \begin{aligned} \delta_X^t(s) = (X_s^t - m_t) &- \int_0^s \bar{\gamma}(t, r) A_r' [dB_r + A_r \delta_X(r) dr] \\ &- \int_0^s \bar{\gamma}(t, r) Q_r^{1/2} [d\bar{B}_r + Q_r^{1/2} \delta_X(r) dr]. \end{aligned}$$

Then, fixing  $0 \leq s \leq t$ , we may apply the Itô formula to obtain the semimartingale

decomposition of the process  $(\delta_X^t(u)\delta_X^s(u)', 0 \leq u \leq s)$ :

$$\begin{aligned}
 \delta_X^t(u)\delta_X^s(u)' &= \int_0^u \delta_X^t(r) \{dX_r^s - \bar{\gamma}(s, r)A_r'[dB_r + A_r\delta_X(r)dr]\}' \\
 &\quad + \int_0^u \delta_X^t(r) \left\{dX_r^s - \bar{\gamma}(s, r)Q_r^{\frac{1}{2}} \left[d\bar{B}_r + Q_r^{\frac{1}{2}}\delta_X(r)dr\right]\right\}' \\
 (34) \quad &\quad + \int_0^u \{dX_r^t - \bar{\gamma}(t, r)A_r'[dB_r + A_r\delta_X(r)dr]\}(\delta_X^s(r))' \\
 &\quad + \int_0^u \left\{dX_r^t - \bar{\gamma}(t, r)Q_r^{\frac{1}{2}} \left[d\bar{B}_r + Q_r^{\frac{1}{2}}\delta_X(r)dr\right]\right\}(\delta_X^s(r))' \\
 &\quad + \langle X^t - m_t, X^s - m_s \rangle_u + \int_0^u \bar{\gamma}(t, r)[Q_r + A_r'A_r]\bar{\gamma}(s, r)'dr.
 \end{aligned}$$

Let us point out here that, due to the Gaussian property of the pair  $(X^t, X^s)$  of martingales, the bracket  $\langle X^t - m_t, X^s - m_s \rangle_u$  is given by

$$\langle X^t - m_t, X^s - m_s \rangle_u = \mathbb{E}(X_u^t - m_t)(X_u^s - m_s)'$$

(see, e.g., Theorem 4.9.5 in [19]), and in particular for  $u = s$  one gets

$$\langle X^t - m_t, X^s - m_s \rangle_s = \Gamma(t, s).$$

Then let us put  $u = s$  in (34) and compute the expectations of both sides, taking into account the martingale property of  $X^t, X^s$ , and  $(B, \bar{B})$  and definition (31). It is easy to check that this shows that  $\bar{\gamma}$  defined in (31), which of course is such that  $\bar{\gamma}(t, t)$ ,  $t \geq 0$ , satisfies (4), with  $\mu = -1$ . The uniqueness of such a solution of (4) is proved below in Lemma 2, which completes the proof of the theorem.  $\square$

Now, we identify the function  $\bar{\pi}_s(X) - \bar{\gamma}_{X\xi}(s)$  appearing in the proof of Proposition 1.

LEMMA 1. *Let  $\bar{\pi}_t(X)$  and  $\bar{\gamma}_{X\xi}(s)$  be the conditional mean and the conditional covariance given by (28) and (21), respectively. Then the quantity*

$$(35) \quad z_s^h = \bar{\pi}_s(X) - \bar{\gamma}_{X\xi}(s), \quad s \geq 0,$$

*satisfies (16), with  $\mu = -1$ .*

*Proof.* Using the complementary notation

$$\delta_\xi(s) = \xi_s - \bar{\pi}_s(\xi), \quad 0 \leq s \leq t,$$

we define

$$\tilde{\gamma}(t, s) = \mathbb{E}(\delta_X^t(s)\delta_\xi(s)/\bar{\mathcal{Y}}_s), \quad 0 \leq s \leq t.$$

Of course, because  $X_t^t = X_t$ , the quantity that we want to identify is nothing but  $\bar{\gamma}_{X\xi}(t) = \tilde{\gamma}(t, t)$ . From (1) and (20) the process  $\xi$  is a semimartingale with the decomposition

$$(36) \quad \xi_t = \int_0^t (X_s - h(s))'Q_s(X_s - h(s))ds + \int_0^t (X_s - h(s))'Q_s^{\frac{1}{2}}d\bar{B}_s.$$

Hence, the fundamental filtering theorem gives

$$\begin{aligned}
 \bar{\pi}_t(\xi) &= \int_0^t \bar{\pi}_s((X_s - h(s))'Q_s(X_s - h(s)))ds \\
 (37) \quad &\quad + \int_0^t [\bar{\pi}_s((X_s - h(s)))' + \bar{\gamma}'_{X\xi}(s)]Q_s^{\frac{1}{2}}d\bar{B}_s^2 + \int_0^t \bar{\gamma}_{X\xi}(s)A_s'd\bar{B}_s^1.
 \end{aligned}$$

From the two previous equations, using (30), it follows that for  $0 \leq s \leq t$

$$\begin{aligned}
 \delta_\xi(t) = & \int_0^t ((X_s - h(s))' Q_s ((X_s - h(s)) - \bar{\pi}_s((X - h)' Q(X - h))) ds \\
 (38) \quad & - \int_0^t (\delta_X'(s) Q_s \bar{\pi}_s(X - h) + \bar{\gamma}'_{X\xi}(s) [Q_s + A'_s A_s] \delta_X(s)) ds \\
 & + \int_0^t (\delta_X(s) - \bar{\gamma}_{X\xi}(s))' Q_s^{\frac{1}{2}} d\bar{B}_s - \int_0^t \bar{\gamma}'_{X\xi}(s) A'_s dB_s.
 \end{aligned}$$

Using (33) and (38), applying the Itô formula, it is easy to determine the semimartingale decomposition of the process  $(\delta_X^t(s) \delta_\xi(s), 0 \leq s \leq t)$ . Then, again applying the fundamental filtering theorem and using the conditional Gaussian properties of  $(X_s, X_s^t, \xi_s)$ , given the  $\sigma$ -field  $\bar{\mathcal{Y}}_s$ , the following equation is readily obtained:

$$\begin{aligned}
 (39) \quad \bar{\gamma}(t, s) = & \int_0^s \bar{\gamma}(t, r) Q_r^{\frac{1}{2}} d\bar{\nu}_r^2 - \int_0^s \bar{\gamma}(t, r) [Q_r + A'_r A_r] \bar{\gamma}_{X\xi}(r) dr \\
 & + \int_0^s [\bar{\gamma}(t, r) Q_r \bar{\pi}_r(X_r - h(r))] dr.
 \end{aligned}$$

Putting  $s = t$  we can write the difference (28) and (39) which leads to

$$\begin{aligned}
 \bar{\pi}_t(X) - \bar{\gamma}_{X\xi}(t) = & m_t - \int_0^t \bar{\gamma}(t, r) [Q_r + A'_r A_r] [\bar{\pi}_r(X) - \bar{\gamma}_{X\xi}(r)] dr \\
 & + \int_0^t \bar{\gamma}(t, r) Q_r h(r) dr + \int_0^t \bar{\gamma}(t, r) A'_r dY_r.
 \end{aligned}$$

This means exactly that  $z_r = \bar{\pi}_r(X) - \bar{\gamma}_{X\xi}(r)$  satisfies (16). The uniqueness of the solution of this equation is proved in Lemma 2, which achieves the proof of the present lemma.  $\square$

Actually, as a direct consequence of (38) we have the following corollary, which will be useful in section 4.

**COROLLARY 1.** *The auxiliary conditional expectation  $\bar{\pi}_t(\xi)$  and the auxiliary filtering error  $\bar{\gamma}_{\xi\xi}(t)$  satisfy the following equations:*

$$\begin{aligned}
 (40) \quad \bar{\pi}_t(\xi) = & \int_0^t \text{tr}(\bar{\gamma}_{XX}(s) Q(s)) ds + \int_0^t (\bar{\pi}_s(X) - h(s))' Q_s (\bar{\pi}_s(X) - h(s)) ds \\
 & + \int_0^t [(\bar{\pi}_s(X) - h(s))' + \bar{\gamma}'_{X\xi}(s)] Q_s^{\frac{1}{2}} d\bar{\nu}_s^2 + \int_0^t \bar{\gamma}'_{X\xi}(s) A'_s d\bar{\nu}_s^1,
 \end{aligned}$$

$$\begin{aligned}
 (41) \quad \bar{\gamma}_{\xi\xi}(t) = & \int_0^t \text{tr}(\bar{\gamma}_{XX}(s) Q(s)) ds + 2 \int_0^t \bar{\gamma}_{X\xi}(s) Q_s^{\frac{1}{2}} d\bar{\nu}_s^2 \\
 & - \int_0^t \bar{\gamma}'_{X\xi}(s) [Q(s) + A'(s) A(s)] \bar{\gamma}_{X\xi}(s) ds \\
 & + 2 \int_0^t (\bar{\pi}_s(X) - h(s))' Q(s) \bar{\gamma}_{X\xi}(s) ds.
 \end{aligned}$$

Finally, to complete the proof of Theorem 2, we derive uniqueness concerning (4) and (16).

**LEMMA 2.** *For arbitrary negative and sufficiently small positive  $\mu$ , (4) and (16) have unique continuous solutions  $\bar{\gamma}$  and  $z^h$  such that  $\bar{\gamma}(t, t)$ ,  $0 \leq t \leq T$ , is a nonnegative definite symmetric matrix.*

*Proof.* The proof for the 1-D case, when  $\bar{\gamma}(t, t) \geq 0$ ,  $t \geq 0$  and  $\mu < 0$ , has been done in [11], and the general case  $n \geq 1$  can be treated by the same way, using the simple fact that for a symmetric nonnegative definite matrix  $\gamma$  we have  $\lambda' \gamma \lambda \geq 0$  for any  $\lambda \in \mathbb{R}^n$ . Also, the proof proposed in [11] is valid for sufficiently small positive  $\mu$ , for example, those such that for any  $t \leq T$ ,  $A_t' A_t - \mu Q_t$  is nonnegative.  $\square$

**3. Particular cases and applications.** Here we deal with some specific cases where the results can be further elaborated. Strictly speaking, only the models which are investigated in subsections 3.1 and 3.3 are particular cases since they involve an observation noise  $\tilde{B}$ , which is both a Brownian motion and independent of the signal  $X$ . For these two examples, we can apply directly Theorem 1, and moreover the special structure of the covariances allows us to simplify the answer. Another method is also proposed. Actually, a direct consequence of Proposition 1 is that only equations for the variance of the filtering error  $\bar{\gamma}_{XX}(t)$  and for the difference  $z_t^h = \bar{\pi}_t(X) - \bar{\gamma}_{X\xi}(t)$  must be written. Here the conditional terms correspond to the auxiliary filtering problem of  $(X, \xi)$ , given the observations  $\tilde{Y}$  defined by (19). Equations for these ingredients will be first obtained for  $\mu = -1$  by means of a direct application of the general filtering theorem (see section 8.1 in [18]) or an approach to filtering proposed in [17] for Itô–Volterra-type models. Then, for any  $\mu$  such that the condition  $(C_\mu)$  is satisfied, it is sufficient to replace  $Q$  by  $-\mu Q$  in the equations. But it is worth emphasizing again that the replacement  $Q$  by  $-\mu Q$  means that the equations obtained from the auxiliary filtering problem cease to have a filtering interpretation. Finally, to get the solution of the LEG filtering problem defined by (6), we just have to take  $\bar{h}(t) = z_t^h$  for  $t \in [0, T]$ .

In the models under study in the other subsections, either  $X$  is not independent of  $\tilde{B}$  (cf. subsection 3.2) or  $\tilde{B}$  is not a Brownian motion (cf. subsections 3.4 and 3.5), and hence Theorem 1 is not directly applicable. Each of these cases is handled by means of an appropriate transformation of the considered original model into a standard scheme, where the Brownian and independence properties are satisfied, to which Theorem 1 can be applied. To this end, either a simple extension of the state space is used or the approach proposed in [14], [15], and [13] for a fractional Brownian noise is applied. Actually, more general settings should be tractable through a similar approach (see also Mandal and Mandrekar [21] for complementary relevant material), but this is kept for possible further investigation.

In what follows of this section we will use the notation  $(B, \tilde{B}) = ((B_t, \tilde{B}_t), t \geq 0)$  for a  $(n + d)$ -dimensional standard Brownian motion.

**3.1. Ornstein–Uhlenbeck-type signal observed through a linear channel with an independent Brownian noise.** At first we deal with the standard Gauss–Markov case where the  $\mathbb{R}^n$ -valued process  $X$  is governed by the stochastic differential equation

$$(42) \quad dX_t = a(t)X_t dt + \sigma_1(t)dB_t, \quad t \geq 0, \quad X_0 = 0,$$

where  $a = (a(t), t \geq 0)$  and  $\sigma_1 = (\sigma_1(t), t \geq 0)$  are  $n \times n$  matrix-valued continuous functions and  $B = (B_t, t \geq 0)$  is a standard Brownian motion in  $\mathbb{R}^n$ . Here, due to (42), denoting by  $\Pi_s$  the solution of the differential equation  $\dot{\Pi}_s = a(s)\Pi_s$ ,  $s \geq 0$ , with the initial condition  $\Pi_0 = I_n$  ( $n \times n$  identity matrix), we have

$$\Gamma(t, s) = \Pi_t \Pi_s^{-1} \Gamma(s, s), \quad 0 \leq s \leq t,$$

where  $\Gamma(s, s)$  solves the Lyapunov differential equation

$$\frac{d}{ds}\Gamma(s, s) = a(s)\Gamma(s, s) + \Gamma(s, s)a'(s) + \sigma_1(s)\sigma_1'(s), \quad s \geq 0, \quad \Gamma(0, 0) = 0.$$

Denote by  $\bar{\gamma}_{xx}$  the unique nonnegative definite solution of the Riccati differential equation

$$(43) \quad \begin{aligned} \dot{\bar{\gamma}}_{xx}(s) &= \sigma_1(s)\sigma_1'(s) + a(s)\bar{\gamma}_{xx}(s) + \bar{\gamma}_{xx}(s)a'(s) \\ &\quad - \bar{\gamma}_{xx}(s)[- \mu Q_s + A'(s)A(s)]\bar{\gamma}_{xx}(s), \quad 0 \leq s \leq T, \end{aligned}$$

with initial condition  $\bar{\gamma}_{xx}(0) = 0$ . From the classical filtering theory it is well-known that (for  $\mu < 0$ )  $\bar{\gamma}_{xx}(s)$  is nothing but the covariance of the filtering error of the signal  $X$  given by the auxiliary observation  $\bar{Y}$ , defined by (19). Then, it is readily seen that the function  $\bar{\gamma}(t, s)$ , where  $\bar{\gamma}(t, s) = \Pi_t \Pi_s^{-1} \bar{\gamma}_{xx}(s)$ , is the solution of (4) and that moreover (5) for the solution  $\bar{h}$  of the LEG filtering problem (6) can be reduced to the following one:

$$(44) \quad d\bar{h}_t = a(t)\bar{h}_t dt + \bar{\gamma}_{xx}(t)A'(t) [dY_t - A(t)\bar{h}_t dt].$$

Actually, (44) can also be obtained directly from the general filtering theory (for  $\mu = -1$  and replacing  $Q$  by  $-\mu Q$ ). For arbitrary  $h$  the general filtering theorem gives the following equation for  $z^h$ :

$$(45) \quad dz_t^h = a(t)z_t^h dt + \bar{\gamma}_{xx}(t)\mu Q_t [z_t^h - h_t] dt + \bar{\gamma}_{xx}(t)A'(t) [dY_t - A(t)z_t^h dt],$$

where  $\bar{\gamma}_{xx}$  is the solution of (43). Hence, again the solution  $\bar{h} = z^{\bar{h}}$  of the LEG filtering problem (6) is given by (44).

Let us emphasize that these equations are nothing but those given in Elliott, Aggoun, and Moore [7] for the solution of the RS filtering problem. Below, in section 5, it will be explained why the LEG and RS filtering problems have the same solution.

**3.2. Correlated signal-observation noises.** Actually, it is a slight generalization of the previous setting. We suppose that the  $\mathbb{R}^n$ -valued signal process  $X$  is governed by the stochastic differential equation

$$dX_t = a_t X_t dt + \sigma_1(t)dB_t + \sigma_2(t)d\tilde{B}_t, \quad X_0 = 0,$$

and the observation process  $Y$  is governed by (1). Applying the general filtering theorem (for arbitrary  $h$ ) we obtain

$$(46) \quad \begin{cases} dz_t^h = a_t z_t^h dt + \bar{\gamma}_{xx}(t)\mu Q_t [z_t^h - h(t)]dt \\ \quad \quad \quad + [\sigma_2(t) + \bar{\gamma}_{xx}(t)A'(t)][dY_t - A(t)z_t^h dt], \\ \dot{\bar{\gamma}}_{xx}(t) = [a(t) - \sigma_2(t)A_t]\bar{\gamma}_{xx}(t) + \bar{\gamma}_{xx}(t)[a(t) - \sigma_2(t)A(t)]' \\ \quad \quad \quad - \bar{\gamma}_{xx}(t)[- \mu Q_t + A'(t)A(t)]\bar{\gamma}_{xx}(t) + \sigma_1(t)\sigma_1'(t). \end{cases}$$

The solution  $\bar{h}$  of the LEG filtering problem (6) is given by

$$(47) \quad d\bar{h}_t = a(t)\bar{h}_t dt + [\sigma_2(t) + \bar{\gamma}_{xx}(t)A'(t)] [dY_t - A(t)\bar{h}_t dt].$$

**3.3. Itô–Volterra-type signal observed through a linear channel with Brownian noise.** Here we deal with the case where the  $\mathbb{R}^n$ -valued process  $X$  is governed by the stochastic integral equation

$$(48) \quad X_t = \int_0^t a(t, s) X_s ds + \int_0^t K(t, s) dB_s, \quad t \geq 0,$$

where  $a$  is a continuous function and the kernel  $K$  is such that the stochastic integral is well-defined and continuous with respect to  $t$ . Following the approach proposed in [17] we can introduce the auxiliary process  $X_t^\tau$ :

$$(49) \quad X_t^\tau = \int_0^t a(\tau, s) X_s ds + \int_0^t K(\tau, s) dB_s, \quad 0 \leq t \leq \tau,$$

which is a Gaussian semimartingale (for fixed  $\tau$ ). We see that since  $X_t^\tau = X_t$  the covariance function of the signal  $X$  can be represented in the form

$$\Gamma(t, s) = \int_s^t a(t, r) \Gamma(r, s) dr + G(t, s), \quad 0 \leq s \leq t,$$

where  $G(\tau, s) = \mathbb{E} X_s^\tau X_s'$  is the unique solution of the equation

$$G(t, s) = \int_0^s [a(t, r) G(s, r) + G(t, r) a'(s, r)] + \int_0^s K(t, r) K'(s, r) dr.$$

Let us denote by  $\gamma(t, s)$  the unique solution of the integral equation

$$(50) \quad \begin{aligned} \gamma(t, s) = & \int_0^s [a(t, r) \gamma(s, r) + \gamma(t, r) a'(s, r)] dr \\ & - \int_0^s \gamma(t, r) [-\mu Q_r + A_r' A_r] \gamma(s, r) dr + \int_0^s K(t, r) K'(s, r) dr. \end{aligned}$$

It follows from [17] that (for  $\mu < 0$ )  $\gamma(\tau, s)$  is nothing but the covariance of the filtering error of the signal  $X^\tau$  given the auxiliary observations  $\bar{Y}$  defined by (19):  $\gamma(\tau, s) = \mathbb{E}(X_s^\tau - \bar{\pi}_s(X^\tau))(X_s - \bar{\pi}_s(X))'$ . It can be shown that  $\bar{\gamma}(t, s)$ , where

$$(51) \quad \bar{\gamma}(t, s) = \int_s^t a(t, r) \bar{\gamma}(r, s) dr + \gamma(t, s), \quad 0 \leq s \leq t,$$

is the solution of (4). Now, using (51), we can reduce (5) for the solution  $\bar{h}$  of the LEG filtering problem (6) to the following one:

$$(52) \quad \bar{h}_t = \int_0^t a(t, s) \bar{h}_s ds + \int_0^t \gamma(t, s) A'(s) [dY_s - A(s) \bar{h}_s ds].$$

Actually, using an approach proposed in [17], (52) can also be obtained directly from the general filtering theory (for  $\mu = -1$  and replacing  $Q$  by  $-\mu Q$ ) which gives for arbitrary  $h$  the following equation for  $z^h$ :

$$(53) \quad \begin{aligned} z_t^h = & \int_0^t a(t, s) z_s^h ds + \int_0^t \gamma(t, s) \mu Q_s [z_s^h - h_s] ds \\ & + \int_0^t \gamma(t, s) A'(s) [dY_s - A(s) z_s^h ds], \end{aligned}$$

where  $\gamma$  is the solution of (50). Hence, again the solution  $\bar{h} = z^{\bar{h}}$  of the LEG filtering problem (6) is given by (52).



**3.4. Ornstein–Uhlenbeck-type signal observed through linear channel with Ornstein–Uhlenbeck-type noises.** Here we deal with the system

$$\begin{cases} dX_t = a(t)X_t dt + dG_t, \\ dY_t = A(t)X_t dt + d\tilde{G}_t, \end{cases}$$

where  $\tilde{G}_t$  and  $G_t$  are two independent Ornstein–Uhlenbeck processes, which means that

$$\begin{cases} dG_t = \beta_t G_t dt + dB_t, \\ d\tilde{G}_t = \tilde{\beta}_t \tilde{G}_t dt + d\tilde{B}_t. \end{cases}$$

To write the solution of the LEG filtering problem (6) for this model, it is sufficient to introduce the extension of the state space and to apply the result formulated in section 3.2. Indeed, our observation model can be written in the form

$$\begin{cases} d\tilde{X}_t = \tilde{a}_t \tilde{X}_t dt + \tilde{\sigma}_1(t) dB_t + \tilde{\sigma}_2(t) d\tilde{B}_t, \\ dY_t = \tilde{A}_t \tilde{X}_t dt + d\tilde{B}_t, \end{cases}$$

with  $\tilde{X}'_t = (X_t \quad \tilde{G}_t \quad G_t)$ ,  $\tilde{\sigma}'_1 = (1 \quad 0 \quad 1)$ ,  $\tilde{\sigma}'_2 = (0 \quad 1 \quad 0)$ ,

$$\tilde{a}_t = \begin{pmatrix} a(t) & 0 & \beta_t \\ 0 & \tilde{\beta}_t & 0 \\ 0 & 0 & \beta_t \end{pmatrix}, \text{ and } \tilde{A}_t = \begin{pmatrix} A(t) & \tilde{\beta}_t & 0 \end{pmatrix}.$$

Also, we can rewrite the quantity (2) as

$$(54) \quad L_T(h, \mu) = \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} \int_0^T \left( \tilde{X}_s - \tilde{h}(s) \right)' \tilde{Q}_s \left( \tilde{X}_s - \tilde{h}(s) \right) ds \right\} \right],$$

with  $\tilde{h}'_t = (h_t \quad h_t^2 \quad h_t^3)$  and

$$\tilde{Q}_t = \begin{pmatrix} Q_t & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Thus the solution of the LEG filtering problem  $\bar{h}$  is the first component of the solution of a LEG filtering problem which can be obtained from the result of section 3.2 (see (46)–(47) with appropriate changes).

**3.5. Fractional Ornstein–Uhlenbeck-type signal observed through linear channel with fractional Brownian noise.** In this section we deal with the observation model

$$\begin{cases} dX_t = aX_t dt + dB_t^H, \quad X_0 = 0, \\ dY_t = AX_t dt + d\tilde{B}_t^H, \quad Y_0 = 0, \end{cases}$$

with constant coefficients  $a$  and  $A$ . The case of varying coefficients can be treated by the same way, using the approach proposed in [13] and [15].

Here  $B_t^H$  and  $\tilde{B}_t^H$  are two independent *normalized fractional Brownian motions* on  $[0, T]$  with the same Hurst parameter  $H \in [1/2, 1)$ . By the definition,  $V^H =$

$(V_t^H, t \in [0, T])$  is a *normalized fractional Brownian motion* on  $[0, T]$  with Hurst parameter  $H$  means that  $V^H$  is a Gaussian process with continuous paths such that  $V_0^H = 0$ ,  $\mathbb{E}V_t^H = 0$ , and

$$(55) \quad \mathbb{E}V_s^H V_t^H = \frac{1}{2} \left[ s^{2H} + t^{2H} - |s - t|^{2H} \right], \quad 0 \leq s, t \leq T.$$

Of course, the fractional Brownian motion reduces to the standard Brownian motion when  $H = 1/2$ , and for  $H \neq 1/2$ , the fractional Brownian motion is outside the world of semimartingales. Nevertheless, it is proved in [12]–[13] that the initial observation model  $(X, Y)$  can be analyzed as a part of the Gaussian signal-observation model  $((X, \zeta)', Z^0)$ , where the filtration generated by  $Z^0$  coincides with  $(\mathcal{Y}_t)$ :

$$(56) \quad \begin{cases} X_t = \int_0^t a X_s ds + \int_0^t K_H(t, s) dM_s^H, \\ \zeta_t = \int_0^t \frac{a}{2} A_H(s) \zeta_s ds + \int_0^t b_H(s) dM_s^H, \\ Z_t^0 = \int_0^t \frac{A}{2} \sqrt{\lambda_H} s^{H-\frac{1}{2}} b'_{1-H}(s) \zeta_s ds + \tilde{B}_t, \end{cases}$$

where the  $2 \times 2$  matrix  $A_H(s)$  and the 2-D vector  $b_H(s)$  are given by

$$A_H(s) = \begin{pmatrix} 1 & s^{1-2H} \\ s^{2H-1} & 1 \end{pmatrix}, \quad b_H(s) = (1 \quad s^{1-2H}).$$

Here for  $H \in (1/2, 1)$

$$(57) \quad K_H(t, s) = H(2H - 1) \int_s^t r^{H-\frac{1}{2}} (r - s)^{H-\frac{3}{2}} dr, \quad 0 \leq s \leq t,$$

and for  $H = 1/2$  the convention  $K_{1/2} \equiv 1$  is used. The process  $M^H$  is a Gaussian martingale, called in [22] the *fundamental martingale*, whose variance function  $\langle M^H \rangle$  is  $\langle M^H \rangle_t = \lambda_H^{-1} t^{2-2H}$ , with

$$\lambda_H = \frac{2H\Gamma(3-2H)\Gamma(H+1/2)}{\Gamma(3/2-H)}.$$

Now, we are in a particular situation of section 3.3. Indeed, it is sufficient to introduce the new matrices

$$\begin{aligned} \tilde{a}(t, s) &= \begin{pmatrix} a & 0 \\ 0 & \frac{a}{2} A_H(s) \end{pmatrix}, \\ \tilde{A}_t &= \begin{pmatrix} 0 & \frac{A}{2} \sqrt{\lambda_H} t^{H-\frac{1}{2}} b'_{1-H}(t) \end{pmatrix}, \\ \tilde{K}'(t, s) &= (\sqrt{\lambda_H} s^{\frac{1}{2}-H} K_H(t, s) \quad \sqrt{\lambda_H} s^{\frac{1}{2}-H} b'_H(s)). \end{aligned}$$

As it was done in section 3.4 we can rewrite the quantity (2) as (54) with  $\tilde{X}' = (X \quad \zeta')$ ,  $\tilde{h}'_t = (h_t \quad (h_t^2)')$ , and

$$\tilde{Q}_t = \begin{pmatrix} Q_t & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Thus the solution of the LEG filtering problem  $\bar{h}$  is the first component of the solution of a LEG filtering problem which can be obtained from the result of section 3.3 (see (50)–(52) with appropriate changes).

Let us note that the case  $H < \frac{1}{2}$  can be treated in the same way, using Corollary 5.2 from [8].

**4. Solution of the LEG filtering problem: The case  $M$  nonnegative definite.** In this section we work with a more general setting of the LEG filtering problem than in section 2. Let us formulate the condition  $(C_\mu^*)$ .

$(C_\mu^*)$  the Riccati–Volterra equation (4) has a unique and bounded solution on  $\{(t, s) : 0 \leq s \leq t \leq T\}$  such that for  $0 \leq t \leq T$  the matrix  $\bar{\gamma}(t, t)$  is nonnegative definite and moreover

$$\det(Id - \mu M \bar{\gamma}(T, T)) > 0.$$

Of course, if  $M = 0$ , then condition  $(C_\mu^*)$  reduces to condition  $(C_\mu)$ , and again for all  $\mu$  negative  $(C_\mu^*)$  is satisfied, and if  $\mu$  is positive,  $(C_\mu^*)$  is satisfied for  $\mu$  sufficiently small.

We state the following extension of Theorem 1.

**THEOREM 3.** *Suppose that the condition  $(C_\mu^*)$  is satisfied. Let  $\bar{h} = (\bar{h}(t), 0 \leq t \leq T)$  be the unique solution of the Itô–Volterra equation (5). Then the pair  $(\bar{h}, \bar{g})$ , where  $\bar{g} = \bar{h}(T)$ , is the solution of the LEG filtering problem, i.e.,*

$$(58) \quad \begin{aligned} (\bar{h}, \bar{g}) = \arg \min_{(h, g) \in \mathcal{H} \times \mathcal{G}} \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} (X_T - g)' M (X_T - g) \right. \right. \\ \left. \left. + \frac{\mu}{2} \int_0^T (X_s - h(s))' Q_s (X_s - h(s)) ds \right\} \right]. \end{aligned}$$

Moreover, the corresponding optimal risk is given by

$$\begin{aligned} \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} (X_T - \bar{g})' M (X_T - \bar{g}) + \frac{\mu}{2} \int_0^T (X_s - \bar{h}(s))' Q_s (X_s - \bar{h}(s)) ds \right\} \right] \\ = \mu [\det(Id - \mu M \bar{\gamma}(T, T))]^{-\frac{1}{2}} \exp \left\{ \frac{\mu}{2} \int_0^T \text{tr}(\bar{\gamma}(s, s) Q_s) ds \right\}. \end{aligned}$$

Theorem 3 is a direct consequence of the results of section 4.1. Its proof will be given in section 4.1.

**4.1. Conditional version of Cameron–Martin formula: The case  $M$  nonnegative definite.** In this section we give the generalization of representation (17) including the final conditions. Indeed, in this section we are interested in the explicit representation

$$(59) \quad \mathcal{J}_T = \mathbb{E} \left[ \exp \left\{ \frac{\mu}{2} (X_T - g)' M (X_T - g) \right. \right. \\ \left. \left. + \frac{\mu}{2} \int_0^T (X_s - h(s))' Q_s (X_s - h(s)) ds \right\} \middle/ \mathcal{Y}_T \right],$$

with any  $(\mathcal{Y}_s)$ -adapted process  $h(s)$ ,  $\mathcal{Y}_T$ -measurable variable  $g$ , and symmetric deterministic nonnegative definite matrices  $M$  and  $Q$ . Here the observation process  $Y$  is defined by (1).

PROPOSITION 2. Suppose that the condition  $(C_\mu^*)$  is satisfied. Let  $z^h = (z_s^h, 0 \leq s \leq T)$  be the unique solution of the Itô-Volterra equation (16). Then the following equality holds:

$$\begin{aligned} \mathcal{J}_T = & [\det(Id - \mu M \bar{\gamma}(T, T))]^{-\frac{1}{2}} \exp \left\{ \frac{1}{2} (z_T^h - g)' [Id - \mu M \bar{\gamma}(T, T)]^{-1} \mu M (z_T^h - g) \right. \\ & + \frac{\mu}{2} \int_0^T (z_s^h - h(s))' Q_s (z_s^h - h(s)) ds + \frac{\mu}{2} \int_0^T \text{tr}(\bar{\gamma}(s, s) Q_s) ds \Big\} \\ & \times \exp \left\{ \int_0^T (z_s^h - \pi_s(X))' A'_s d\nu_s - \frac{1}{2} \int_0^T \|A(s) (z_s^h - \pi_s(X))\|^2 ds \right\}. \end{aligned}$$

*Proof.* The proof is based on the idea developed in the proof of Proposition 1. Again, we work with  $\mu = -1$ , and then we can replace  $Q$  and  $M$  by  $-\mu Q$  and  $-\mu M$ , respectively. For the proof we shall note only that the classical Bayes formula gives that

$$(60) \quad \mathcal{J}_T = \frac{\mathbb{E} [\exp(-\frac{1}{2}(X_T - g)' M (X_T - g) - \xi_T) / \bar{\mathcal{Y}}_T]}{\mathbb{E} [\rho_T / \bar{\mathcal{Y}}_T]} = \frac{\varphi_1(T)}{\bar{\pi}_T(\rho)},$$

where  $\rho_t$  and  $\xi_t$  are defined by (26) and (20), respectively. But the conditional Gaussian properties of the pair  $(X, \xi)$ , given  $\bar{\mathcal{Y}}_T$ , gives the following:

$$\begin{aligned} \ln \varphi_1(T) = & -\frac{1}{2} \ln[\det(Id + M \bar{\gamma}_{XX}(T))] \\ (61) \quad & -\frac{1}{2} (z_T^h - g)' [(Id + M \bar{\gamma}_{XX}(T))]^{-1} M (z_T^h - g) \\ & - \bar{\pi}_T(\xi) + \frac{1}{2} \bar{\gamma}_{\xi\xi}(T) \end{aligned}$$

(see, for example, [18, Lemma 11.6]).

Since

$$d\rho_t = -\rho_t (X_t - h(t))' Q_t^{\frac{1}{2}} d\bar{B}_t,$$

thanks to the general filtering theorem we can write

$$(62) \quad d\bar{\pi}_t(\rho) = \bar{\pi}_t(\rho) \left\{ -\bar{\pi}'_t(X - h) Q_t^{\frac{1}{2}} d\bar{\nu}_t^2 + \left[ \frac{\bar{\pi}'_t(\rho X)}{\bar{\pi}_t(\rho)} - \bar{\pi}'_t(X) \right] A'_t d\bar{\nu}_t^1 \right\}.$$

The same arguments that we have used for the proof of Proposition 1 give the equality

$$\frac{\bar{\pi}_t(\rho X)}{\bar{\pi}_t(\rho)} = \pi_t(X).$$

Hence,

$$d\bar{\pi}_t(\rho) = \bar{\pi}_t(\rho) \left\{ -\bar{\pi}'_t(X - h) Q_t^{\frac{1}{2}} d\bar{\nu}_t^2 + [\pi'_t(X) - \bar{\pi}'_t(X)] A'_t d\bar{\nu}_t^1 \right\}$$

or equivalently

$$\begin{aligned} (63) \quad \bar{\pi}_T(\rho) = & \exp \left\{ -\int_0^T \bar{\pi}'_s(X - h) Q_s^{\frac{1}{2}} d\bar{\nu}_s^2 + \int_0^T [\pi'_s(X) - \bar{\pi}'_s(X)] A'_s d\bar{\nu}_s^1 \right. \\ & \left. - \frac{1}{2} \int_0^T \|A_s [\pi_s(X) - \bar{\pi}_s(X)]\|^2 ds - \frac{1}{2} \int_0^T \|Q_s^{\frac{1}{2}} [\bar{\pi}_s(X - h)]\|^2 ds \right\}. \end{aligned}$$

Now, it follows from (61) that to prove the statement of the proposition it is sufficient to write the expression for

$$\Psi_T = \frac{\exp(-\bar{\pi}_T(\xi) + \frac{1}{2}\bar{\gamma}_{\xi\xi}(T))}{\bar{\pi}_T(\rho)}.$$

The equalities (40), (41), and (63) imply

$$\begin{aligned} \ln(\Psi_t) = & -\frac{1}{2} \int_0^T \operatorname{tr}(\bar{\gamma}(s, s)Q_s)ds - \frac{1}{2} \int_0^T \bar{\pi}_s((X-h))'Q_s\bar{\pi}_s((X-h))ds \\ & + \int_0^T (\bar{\pi}_s(X) - h(s))'Q_s\bar{\gamma}_{X\xi}(s)ds - \frac{1}{2} \int_0^t \bar{\gamma}'_{X\xi}(s)Q_s\bar{\gamma}_{X\xi}(s)ds \\ & + \int_0^T [\pi_s(X) - \bar{\pi}_s(X) - \bar{\gamma}_{X\xi}(s)]'A'(s)d\bar{\nu}_s^1 \\ & - \frac{1}{2} \int_0^T \bar{\gamma}'_{X\xi}(s)A'(s)A(s)\bar{\gamma}_{X\xi}(s)ds + \frac{1}{2} \int_0^T \|A_s[\pi_s(X) - \bar{\pi}_s(X)]\|^2 ds. \end{aligned}$$

Replacing  $d\bar{\nu}_t^1$  by  $d\bar{\nu}_t^1 = d\nu_t + A(t)[\pi_t(X) - \bar{\pi}_t(X)]dt$ , we obtain that

$$\begin{aligned} \ln(\Psi_t) = & -\frac{1}{2} \int_0^T \operatorname{tr}(\bar{\gamma}(s, s)Q_s)ds - \frac{1}{2} \int_0^T (\bar{\pi}_s(X-h) - \bar{\gamma}_{X\xi}(s))'Q_s(\bar{\pi}_s(X-h) - \bar{\gamma}_{X\xi}(s))ds \\ & + \int_0^T [\pi_s(X) - \bar{\pi}_s(X) - \bar{\gamma}_{X\xi}(s)]'A'(s)d\nu_s - \frac{1}{2} \int_0^T \|A(s)(\pi_s(X) - \bar{\pi}_s(X) - \bar{\gamma}_{X\xi}(s))\|^2 ds, \end{aligned}$$

and it gives the statement of the proposition.  $\square$

*Proof of Theorem 3.* We follow the same lines as in the proof of Theorem 1, starting here from Proposition 2. Again, we claim that the following chain of inequalities holds for any  $(g, h) \in \mathcal{G} \times \mathcal{H}$ :

$$\begin{aligned} & \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} (X_T - g)'M(X_T - g) + \frac{\mu}{2} \int_0^T (X_s - h(s))'Q_s(X_s - h(s))ds \right\} \right] \\ & \stackrel{(a)}{\geq} [\det(Id - \mu M\bar{\gamma}(T, T))]^{-\frac{1}{2}} \exp \left\{ \frac{\mu}{2} \int_0^T \operatorname{tr}(\bar{\gamma}(s, s)Q_s)ds \right\} \\ & \times \mathbb{E} \mu \left[ \exp \left\{ \int_0^T (z_s^h - \pi_s(X))'A'_s d\nu_s - \frac{1}{2} \int_0^T \|A(s)(z_s^h - \pi_s(X))\|^2 ds \right\} \right] \\ & \stackrel{(b)}{\geq} \mu [\det(Id - \mu M\bar{\gamma}(T, T))]^{-\frac{1}{2}} \exp \left\{ \frac{\mu}{2} \int_0^T \operatorname{tr}(\bar{\gamma}(s, s)Q_s)ds \right\}. \end{aligned}$$

Inequality (a) follows directly from Proposition 2, and inequality (b) is valid due to the same arguments as those we used in the proof of Theorem 1. Here to attain the lower bound we must take  $\bar{h} = z^{\bar{h}}$  and  $\bar{g} = z^{\bar{h}}_T$ , where  $z^h$  is the solution of (16), which means that  $\bar{h}$  is the unique solution of (5). Again, for  $(\bar{g}, \bar{h})$  the lower bound is attained because the system  $(X, Y, \bar{h})$  is Gaussian, and hence inequalities (a) and (b) are actually equalities.

**5. LEG and RS filtering problems.** Let  $\bar{h} = (\bar{h}(s), 0 \leq s \leq T)$  be the solution of the LEG filtering problem (6) on  $[0, T]$  given by (5). For any fixed  $t \leq T$ , let us

denote  $\hat{g}_t$  by

$$\hat{g}_t = \arg \min_{g \in \mathcal{Y}_t} \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} (X_t - g)' Q_t (X_t - g) + \frac{\mu}{2} \int_0^t (X_s - \bar{h}(s))' Q_s (X_s - \bar{h}(s)) ds \right\} / \mathcal{Y}_t \right],$$

where  $g \in \mathcal{Y}_t$  means that  $g$  is a  $\mathcal{Y}_t$ -measurable variable. It follows directly from Proposition 2 that, provided that  $\det(Id - \mu Q_t \bar{\gamma}(t, t)) > 0$ , the equality  $\hat{g}_t = z_t^{\bar{h}}$  holds. Since it was noted in the proof of Theorem 3 that  $\bar{h}(t) = z_t^{\bar{h}}$ , hence we have also  $\hat{g}_t = \bar{h}(t)$ . It means that for  $t \in [0, T]$  the solution  $\bar{h}$  of the LEG filtering problem satisfies the following recursive equation:

$$(64) \quad \bar{h}(t) = \arg \min_{g \in \mathcal{Y}_t} \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} (X_t - g)' Q_t (X_t - g) + \frac{\mu}{2} \int_0^t (X_s - \bar{h}(s))' Q_s (X_s - \bar{h}(s)) ds \right\} / \mathcal{Y}_t \right].$$

Indeed, in the literature, the recursion (64) is the basic *definition* of the so-called RS filtering problem which was introduced in [6]. Therefore, we have also proved the following statement where  $(C_\mu^{**})$  denotes the condition.

$(C_\mu^{**})$  the Riccati–Volterra equation (4) has a unique and bounded solution on  $\{(t, s) : 0 \leq s \leq t \leq T\}$  such that for  $0 \leq t \leq T$  the matrix  $\bar{\gamma}(t, t)$  is nonnegative definite and moreover

$$\det(Id - \mu Q_t \bar{\gamma}(t, t)) > 0.$$

**THEOREM 4.** *Assume that the condition  $(C_\mu^{**})$  is satisfied. Let  $\bar{h} = (\bar{h}(t), 0 \leq t \leq T)$  be the unique solution of the Itô–Volterra equation (5), i.e.,  $\bar{h}$  is the solution of the LEG filtering problem (6). Then  $\bar{h}$  is the solution of the RS filtering problem (64).*

Actually, we did not find in the literature any trace of the discussion about the relation between the LEG filtering problem (6) and the RS filtering problem (64), even in a Gauss–Markov case. As a complement to our observation that these two problems have the same solution, we propose an example to show that in a bit more general setting, two similar problems may have different solutions.

For given positive symmetric deterministic matrices  $\Lambda_s, 0 \leq s \leq T$ , let us set  $\Phi_t(h) = (X_t' h(t)') \Lambda_t \begin{pmatrix} X_t \\ h(t) \end{pmatrix}$ . We can define  $\bar{h} \in \mathcal{H}$  as a solution of a *LEG-type filtering problem*:

$$(65) \quad \bar{h} = \arg \min_{h \in \mathcal{H}} \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} \int_0^T \Phi_s(h) ds \right\} \right],$$

where  $h = (h(s), 0 \leq s \leq T) \in \mathcal{H}$  means that  $h$  is a  $(\mathcal{Y}_s)$ -adapted process.

We can also define  $\hat{h}$  as the solution of the following recursive equation (*RS-type filtering problem*):

$$(66) \quad \hat{h}(t) = \arg \min_{g \in \mathcal{Y}_t} \mathbb{E} \left[ \mu \exp \left\{ \frac{\mu}{2} \Phi_t(g) + \frac{\mu}{2} \int_0^t \Phi_s(\hat{h}) ds \right\} / \mathcal{Y}_t \right],$$

where  $g \in \mathcal{Y}_t$  means that  $g$  is a  $\mathcal{Y}_t$ -measurable variable.

The question which we discuss now is the following: Does the equality  $\bar{h} = \hat{h}$  hold?

As we have just proved, the answer is positive for singular matrices  $\Lambda$ , namely, when  $\Lambda_{11} = \Lambda_{22} = -\Lambda_{12} = Q$ . But in the general situation the answer may be negative. Without technical details, which can be found in [16], we propose the following example:  $\Lambda = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ ,  $A = 1$ ,  $\mu = -1$ , and  $X_t = B_t$ . Even in this Markov case,  $\hat{h} \neq \bar{h}$ . More explicitly, one gets

$$\bar{h}(t) = \int_0^t \bar{H}(T, t, s) dY_s,$$

where

$$\begin{aligned} \bar{H}(T, t, s) &= \frac{\sinh(\sqrt{3}s) \cosh(T-t)}{\sqrt{\alpha_t \alpha_s}}, \\ \alpha_t &= \frac{\sqrt{3}+1}{2} \cosh(T + (\sqrt{3}-1)t) + \frac{\sqrt{3}-1}{2} \cosh(T - (\sqrt{3}+1)t). \end{aligned}$$

Clearly,  $\bar{H}$  depends on  $T$ , and by the definition of recursion (66),  $\hat{h}(t)$  does not depend on  $T$ . More precisely, one obtains

$$\hat{h}(t) = \int_0^t \hat{H}(t, s) dY_s,$$

where

$$\hat{H}(t, s) = \frac{\sinh(\sqrt{3}s) \sqrt[3]{\cosh(\sqrt{3}t)}}{\sqrt{3}(\cosh(\sqrt{3}s))^{\frac{2}{3}}},$$

and so  $\hat{h}$  and  $\bar{h}$  are different.

**6. Information state, interpretation.** In this section we discuss the probabilistic interpretation of the ingredients of the “information state” which was introduced in the context of RS filtering and LEG control problems. By the definition, the information state contains all of the information needed to describe the solution of the concerned optimization problem. In particular, it takes into account the cost function but not only estimates of the signal, and it should give the total information about the model states available in the measurement.

*Risk-sensitive filtering.* In the context of the RS filtering problem the definition of the information state can be found, for example, in [7]. It is the density  $\lambda_t$ , with respect to the Lebesgue measure, of the nonnormalized random measure  $\omega_t$ :

$$(67) \quad \omega_t(dx) = \mathbb{E} \left[ \mathbb{I}(X_t \in dx) \exp \left\{ \frac{\mu}{2} \int_0^t (X_s - h(s))' Q_s (X_s - h(s)) ds \right\} \middle/ \mathcal{Y}_t \right],$$

where  $h \in \mathcal{H}$  and the observation  $Y$  is defined by (1).

In a classical Gauss–Markov setting, an explicit representation of  $\lambda_t$  can be obtained as the solution of some stochastic partial differential equation (see, e.g., [3]).

We claim that for a general Gaussian signal  $X$  the density  $\lambda_t$  satisfies the following equality:

$$(68) \quad \begin{aligned} \lambda_t(x) = & [(2\pi)^n \det(\bar{\gamma}(t, t))]^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x - z_t^h)' \bar{\gamma}^{-1}(t, t) (x - z_t^h) \right. \\ & + \frac{\mu}{2} \int_0^t (z_s^h - h(s))' Q_s (z_s^h - h(s)) ds + \frac{\mu}{2} \int_0^t \text{tr}(\bar{\gamma}(s, s) Q_s) ds \Big\} \\ & \times \exp \left\{ \int_0^t (z_s^h - \pi_s(X))' A'(s) d\nu_s - \frac{1}{2} \int_0^t \|A(s) (z_s^h - \pi_s(X))\|^2 ds \right\}, \end{aligned}$$

where  $\bar{\gamma}$  and  $z^h$  are the solutions of (4) and (16), respectively.

Indeed, to prove (68) we can replace the equality (60) by the following:

$$\omega_t(dx) = \frac{\mathbb{E} [\mathbb{I}(X_t \in dx) \exp(-\xi_t) / \bar{\mathcal{Y}}_t]}{\mathbb{E} [\rho_t / \bar{\mathcal{Y}}_t]},$$

where  $\rho_t$  and  $\xi_t$  are defined by (26) and (20), respectively. Again, conditionally Gaussian properties of the pair  $(X, \xi)$  imply that

$$(69) \quad \begin{aligned} & \mathbb{E} [\mathbb{I}(X_t \in dx) \exp \{-\xi_t\} / \bar{\mathcal{Y}}_t] = [(2\pi)^n \det(\bar{\gamma}(t, t))]^{-\frac{1}{2}} \\ & \times \exp \left\{ -\frac{1}{2} (x - z_t^h)' \bar{\gamma}^{-1}(t, t) (x - z_t^h) - \bar{\pi}_t(\xi) + \frac{1}{2} \bar{\gamma}_{\xi\xi}(t) \right\} dx. \end{aligned}$$

Now, it is sufficient to replace (61) in the proof of Proposition 2 by (69).

It is worth emphasizing that (for negative  $\mu$ ) now we know the probabilistic interpretation of the involved pair  $(z^h, \bar{\gamma})$ . Actually, we have proved that  $z^h$  is the difference  $\bar{\pi}_t(X) - \bar{\gamma}_{X\xi}(t)$  and  $\bar{\gamma}$  is nothing but the covariance of the filtering error of  $X$  in view of auxiliary observations  $\bar{Y}$ .

Of course, after a simple integration of  $\lambda_t$ , (68) gives Propositions 1 and 2 and therefore the solution of the LEG and RS filtering problems.

*Linear exponential Gaussian control.* In the context of the LEG control problem for a partially observed process, the information state is also defined (see, e.g., [7]) as the density  $\lambda_t$ , with respect to the Lebesgue measure, of the nonnormalized random measure  $\omega_t$ :

$$(70) \quad \omega_t(dx) = \mathbb{E} \left[ \mathbb{I}(X_t \in dx) \exp \left\{ \frac{\mu}{2} \int_0^t X_s' Q_s X_s ds \right\} / \mathcal{Y}_t \right],$$

where  $X$  is the controlled state governed by the equation

$$(71) \quad dX_t = a(t)X_t dt + b(t)u_t dt + \sigma_1(t)dB_t, \quad t \geq 0, \quad X_0 = 0,$$



driven by a standard Brownian motion  $B = (B_t, t \geq 0)$  in  $\mathbb{R}^n$  and  $u \in \mathcal{H}$  corresponding to the available observation  $Y$  defined by (1).

By the same way that we have just explained, for the conditionally Gaussian pair  $(X, Y)$ , one can check that the density  $\lambda_t$  satisfies the following equality:

$$(72) \quad \begin{aligned} \lambda_t(x) = & [(2\pi)^n \det(\bar{\gamma}(t, t))]^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x - z_t)' \bar{\gamma}^{-1}(t, t) (x - z_t) \right. \\ & + \frac{\mu}{2} \int_0^t z_s' Q_s z_s ds + \frac{\mu}{2} \int_0^t \text{tr}(\bar{\gamma}(s, s) Q_s) ds \Big\} \\ & \times \exp \left\{ \int_0^t (z_s - \pi_s(X))' A'(s) d\nu_s - \frac{1}{2} \int_0^t \|A(s)(z_s - \pi_s(X))\|^2 ds \right\}, \end{aligned}$$

where  $\bar{\gamma}$  is the solution of (43) and  $z$  is the solution of the equation

$$(73) \quad dz_t = a(t)z_t dt + b(t)u_t dt + \bar{\gamma}(t)\mu Q_t z_t dt + \bar{\gamma}(t)A'(t)[dY_t - A(t)z_t dt].$$

Actually, it is the equation for the difference  $z = \bar{\pi}_t(X) - \bar{\gamma}_{X\xi}(t)$ , where the conditional expectations are taken with respect to the auxiliary observation process  $\bar{Y}$  defined by (19), with  $h = 0$ .

Equality (72) gives the possibility to rewrite the cost function in terms of the completely observable process  $z$ , namely,

$$\begin{aligned} & \mathbb{E} \left[ \exp \left\{ \frac{\mu}{2} \int_0^T X_s' Q_s X_s ds \right\} \right] \\ &= \mathbb{E} \left\{ \mathbb{E} \left[ \exp \left\{ \frac{\mu}{2} \int_0^T X_s' Q_s X_s ds \right\} \middle/ \mathcal{Y}_T \right] \right\} \\ &= \mathbb{E} \left[ \exp \left\{ \frac{\mu}{2} \int_0^T z_s' Q_s z_s ds + \frac{\mu}{2} \int_0^T \text{tr}(\bar{\gamma}(s) Q_s) ds + \int_0^T (z_s - \pi_s(X))' A'(s) d\nu_s \right. \right. \\ &\quad \left. \left. - \frac{1}{2} \int_0^T \|A(s)(z_s - \pi_s(X))\|^2 ds \right\} \right] \\ &= \exp \left\{ \frac{\mu}{2} \int_0^T \text{tr}(\bar{\gamma}(s) Q_s) ds \right\} \tilde{\mathbb{E}} \exp \left\{ \frac{\mu}{2} \int_0^T z_s' Q_s z_s ds \right\}, \end{aligned}$$

where  $\tilde{\mathbb{E}}$  stands for an expectation with respect to the new measure  $\tilde{\mathbb{P}}$  such that

$$\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} = \exp \left\{ \int_0^T (z_s - \pi_s(X))' A'(s) d\nu_s - \frac{1}{2} \int_0^T \|A(s)(z_s - \pi_s(X))\|^2 ds \right\}.$$

With respect to this new measure the solution of (73) can be represented as

$$(74) \quad dz_t = a(t)z_t dt + b(t)u_t dt + \bar{\gamma}(t)\mu Q_s z_t dt + \bar{\gamma}(t)A'(t)d\tilde{\nu}_t,$$

with the new Brownian motion  $\tilde{\nu}$ :

$$(75) \quad d\tilde{\nu}_t = d\nu_t - A(t)(z_t - \pi_t(X)) dt.$$

Thus, the new process  $z$  plays the role of the completely observed controlled state (see [1]).

Now, we emphasize that the probabilistic interpretation of the information state  $z$  used in [1] is nothing but  $z = \bar{\pi}_t(X) - \bar{\gamma}_{X\xi}(t)$ , where the conditional expectations are taken with respect to the auxiliary observation process  $\bar{Y}$  defined by (19), with  $h = 0$ . Also,  $\bar{\gamma}$  is the conditional covariance of  $X$ .

**Acknowledgments.** We are grateful to the reviewers for their valuable suggestions and comments, which helped to improve the presentation of this paper.

#### REFERENCES

- [1] A. BENSOUSSAN AND J.H. VAN SCHUPPEN, *Optimal control of partially observable stochastic systems with an exponential-of-integral performance index*, SIAM J. Control Optim., 23 (1985), pp. 599–613.
- [2] C.D. CHARALAMBUS, S. DEY, AND R.J. ELLIOTT, *New finite-dimensional risk sensitive filters: Small-noise limits*, IEEE Trans. Automat. Control, 43 (1997), pp. 1424–1429.
- [3] I. COLLINGS, M. JAMES, AND J.B. MOORE, *An information state approach to risk-sensitive tracking problems*, J. Math. Systems Estim. Control, 6 (1996), pp. 1–24.
- [4] S. DEY AND J.B. MOORE, *Risk sensitive filtering and smoothing for hidden Markov models*, Systems Control Lett., 25 (1995), pp. 361–366.
- [5] S. DEY AND J.B. MOORE, *Risk-sensitive filtering and smoothing via reference probability methods*, IEEE Trans. Automat. Control, 42 (1997), pp. 1587–1591.
- [6] S. DEY, R.J. ELLIOTT, AND J.B. MOORE, *Finite dimensional risk-sensitive estimation for continuous time nonlinear systems*, in Proceedings of the European Control Conference, Brussels, 1997.
- [7] R.J. ELLIOTT, L. AGGOUN, AND J.B. MOORE, *Hidden Markov Models: Estimation and Control*, Springer, Berlin, 1994.
- [8] C. JOST, *Transformation formulas for fractional Brownian motion*, Stochastic Process. Appl., 116 (2006), pp. 1341–1357.
- [9] G. KALLIANPUR, *Stochastic Filtering Theory*, Springer, New York, 1980.
- [10] M.L. KLEPTSINA AND A. LE BRETON, *Optimal linear filtering of general multidimensional Gaussian processes - Application to Laplace transforms of quadratic functionals*, J. Appl. Math. Stoch. Anal., 14 (2001), pp. 215–226.
- [11] M.L. KLEPTSINA AND A. LE BRETON, *A Cameron-Martin type formula for general Gaussian processes - A filtering approach*, Stoch. Stoch. Rep., 72 (2002), pp. 229–250.
- [12] M.L. KLEPTSINA AND A. LE BRETON, *Statistical analysis of the fractional Ornstein-Uhlenbeck type process*, Stat. Inference Stoch. Process, 5 (2002), pp. 229–248.
- [13] M.L. KLEPTSINA AND A. LE BRETON, *Extension of the Kalman-Bucy filter to elementary linear systems with fractional Brownian noises*, Stat. Inference Stoch. Process., 5 (2002), pp. 249–271.
- [14] M.L. KLEPTSINA, A. LE BRETON, AND M.-C. ROUBAUD, *An elementary approach to filtering in systems with fractional Brownian observation noise*, in Proceedings of the 7th Vilnius Conference, Prob. Theory and Math. Stat., TEV, Vilnius, Lithuania, B. Grigelionis et al., eds., VSP, Zeist, The Netherlands, 1999, pp. 373–392.
- [15] M.L. KLEPTSINA, A. LE BRETON, AND M.-C. ROUBAUD, *General approach to filtering with fractional Brownian noises - Application to linear systems*, Stoch. Stoch. Rep., 71 (2000), pp. 119–140.
- [16] M.L. KLEPTSINA, A. LE BRETON, AND M. VIOT, *About the Relationship Between Linear Exponential Gaussian and Risk-Sensitive Filtering Problems*, <http://www.univ-lemans.fr/sciences/statist/download/Kleptsyna/Manut.pdf> (2008).
- [17] M.L. KLEPTSINA AND A.YU. VERETENNIKOV, *Linear filtering and properties of conditional laws of Itô-Volterra equations*, Statistics and Control of Stochastic Processes, Springer, Berlin, 1985, pp. 179–196.
- [18] R.S. LIPTSER AND A.N. SHIRYAEV, *Statistics of Random Processes I - General Theory*, Springer, New York, 1977.
- [19] R.S. LIPTSER AND A.N. SHIRYAEV, *Theory of Martingales*, Kluwer Academic, Dordrecht, 1989.
- [20] J.B. MOORE, R.J. ELLIOTT, AND S. DEY, *Risk-sensitive generalization of minimum variance estimation and control*, IFAC Symposium on Nonlinear Control Systems Design, Lake Tahoe, CA, 1995, pp. 465–470.
- [21] P.K. MANDAL AND V. MANDREKAR, *A Bayes formula for Gaussian noise processes and its applications*, SIAM J. Control Optim., 39 (2000), pp. 852–871.

- [22] I. NORROS, E. VALKEILA, AND J. VIRTAMO, *An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions*, Bernoulli, 5 (1999), pp. 571–587.
- [23] J.L. SPEYER, C. FAN, AND R.N. BANAVAR, *Optimal stochastic estimation with exponential criteria*, in Proceedings of the 31st IEEE Conference on Decision and Control, IEEE, Tucson, AZ, 1992, Vol. 2, pp. 2293–2298.
- [24] P. WHITTLE, *Optimization Over Time: Dynamic Programming and Stochastic Control*, Wiley Probability Math. Stat., 11, John Wiley, New York, 1982.
- [25] P. WHITTLE, *Risk-Sensitive Optimal Control*, John Wiley, New York, 1990.
- [26] P. WHITTLE, *A risk-sensitive maximum principle: The case imperfect state observations*, IEEE Trans. Automat. Control, 36 (1991), pp. 793–801.

## POLICY ITERATION AND THE MAX-PLUS FINITE ELEMENT METHOD\*

D. MCCAFFREY†

**Abstract.** We set out a policy iteration algorithm for the solution of a nonconvex nonlinear-affine finite horizon differential game. This involves the use of a max-plus finite element method (FEM) to solve a convex optimal control problem in the value determination step of the algorithm, where this convex problem is the result of fixing the policy of the second, or outer, player. A quadratic program is then solved in the policy improvement step to improve the feedback policy employed by the outer player. We show that the algorithm converges in a finite number of steps and that the approximation error at convergence is of order  $\sqrt{\Delta t} + \Delta x(\Delta t)^{-1}$ .

**Key words.** max-plus algebra, policy iteration, nonconvex differential game, finite element method, Hamilton–Jacobi–Isaacs equation,  $H_\infty$  control

**AMS subject classifications.** Primary, 49L20, 91A23; Secondary, 65M60, 06A15, 12K10

**DOI.** 10.1137/070705465

**1. Introduction.** We consider the finite horizon differential game

$$(1) \quad v(x, T) = \inf_{a(\cdot)} \sup_{b(\cdot)} \int_0^T \left\{ \frac{1}{2}x(s)^2 + \frac{1}{2}a(s)^2 - \frac{\gamma^2}{2}b(s)^2 \right\} ds + \phi(x(T))$$

over trajectories  $(x(\cdot), a(\cdot), b(\cdot))$  satisfying

$$(2) \quad \dot{x}(s) = f(x(s)) + g(x(s))a(s) + h(x(s))b(s), \quad x(0) = x,$$

where

$$x(s) \in X \subseteq \mathbb{R}^n, \quad a(s) \in U \subseteq \mathbb{R}^m, \quad b(s) \in W \subseteq \mathbb{R}^r.$$

This problem arises, for example, as the differential game formulation of a well-known class of nonlinear affine  $H_\infty$  control problems. The reader can consult [11] for an exposition of the differential game approach to general nonlinear  $H_\infty$  control, and [12, 13, 8] for background on the specific nonlinear affine subclass of  $H_\infty$  control problems considered here. It is sufficient to note that the first (or inner) player, with input  $b(\cdot)$ , is referred to as the disturbance, and the second (or outer) player, who responds with policy  $a(\cdot)$ , is referred to as the control. In a strict formulation of the problem, the second player actually responds with strategies, i.e., nonanticipating functionals from the space of disturbance inputs to the space of control inputs, but for our purposes here, where we apply very coarse discretization to the space of control inputs, the above notation will be sufficient, provided the reader bears in mind the simplifications we are making. Also note that the value  $\gamma$  appearing in the cost function in (1) is given as an external parameter.

---

\*Received by the editors October 16, 2007; accepted for publication (in revised form) April 30, 2008; published electronically December 5, 2008. A preliminary version of this paper was presented at the International Workshop on Idempotent and Tropical Mathematics and Problems of Mathematical Physics, Independent University of Moscow, Moscow, August 25–30, 2007.

<http://www.siam.org/journals/sicon/47-6/70546.html>

†Department of Automatic Control and Systems Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK (david@mccaffrey275.fsnet.co.uk).

The problem is usually considered over an infinite horizon, where the objective of the second player is to maintain the  $L_2$  system output  $\int \frac{1}{2}x(s)^2 + \frac{1}{2}a(s)^2 ds$  less than or equal to  $\gamma^2$  times the  $L_2$  system input  $\int \frac{1}{2}b(s)^2 ds$ , i.e., to achieve an attenuation level  $\leq \gamma^2$ . Under suitable regularity assumptions, usually expressed in terms of the local controllability and observability at the origin of the system (2), it is known that the infinite horizon problem is solvable, in the sense that the value function is finite, on some nonlocal neighborhood of the origin for  $\gamma$  above some minimum level. So in particular it is known (see, for instance, [11]) that the value function  $v(x, t)$  for the finite horizon problem is a (possibly nonsmooth) solution to the Hamilton–Jacobi–Isaacs equation

$$(3) \quad H(x, \partial v / \partial x) = \partial v / \partial t$$

with initial condition  $v(x, 0) = \phi(x)$  for  $(x, t) \in X \times (0, T]$ , where the Hamiltonian is defined as

$$H(x, p) = \min_a \max_b \left\{ p(f(x) + g(x)a + h(x)b) + \frac{1}{2}x^2 + \frac{1}{2}a^2 - \frac{\gamma^2}{2}b^2 \right\}.$$

Note that this Hamiltonian is nonconvex in  $p$ .

Suppose we choose some feedback function  $\hat{a}(x)$  and, on any solution trajectory  $(x(\cdot), a(\cdot), b(\cdot))$ , define the control input  $a(s) = \hat{a}(x(s))$  for all  $s$ . We can then define  $f_{\hat{a}}(x, b) = f(x) + g(x)\hat{a}(x) + h(x)b$  and  $l_{\hat{a}}(x, b) = \frac{1}{2}x^2 + \frac{1}{2}\hat{a}(x)^2 - \frac{\gamma^2}{2}b^2$ , and consider the finite horizon optimal control problem

$$(4) \quad v_{\hat{a}}(x, T) = \sup_{b(\cdot)} \int_0^T l_{\hat{a}}(x(s), b(s)) ds + \phi(x(T))$$

over trajectories  $(x(\cdot), b(\cdot))$  satisfying

$$\dot{x}(s) = f_{\hat{a}}(x(s), b(s)), \quad x(0) = x.$$

In this case, the value function  $v_{\hat{a}}(x, t)$  satisfies the Hamilton–Jacobi equation

$$(5) \quad H_{\hat{a}}(x, \partial v_{\hat{a}} / \partial x) = \partial v_{\hat{a}} / \partial t$$

with initial condition  $v_{\hat{a}}(x, 0) = \phi(x)$  for  $(x, t) \in X \times (0, T]$ , where the Hamiltonian is defined as  $H_{\hat{a}}(x, p) = \max_b \{ p f_{\hat{a}}(x, b) + l_{\hat{a}}(x, b) \}$ . Note that this Hamiltonian is convex in  $p$  for all  $x$ .

A max-plus analogue of the finite element method (FEM) is set out in [1] for the numerical computation of the value function  $v_{\hat{a}}$  solving convex optimal control problem (4). This involves a max-plus variational formulation in which successive time steps of the semigroup action are approximated by projection onto two idempotent semimodules, the first spanned by a set of test functions and the second by a set of finite element basis functions. The errors associated with these projections are estimated in [1] to be of order  $\sqrt{\Delta t} + \Delta x(\Delta t)^{-1}$ , where  $\Delta t$  is the time discretization step and  $\Delta x$  the space discretization step.

In this paper, we set out a policy iteration algorithm for the solution of the nonconvex differential game (1). This involves the use of the max-plus FEM to solve (5) for a given fixed control feedback  $a(x)$  in the value determination step of the algorithm, and then the use of a quadratic program to improve the control feedback in the policy improvement step. This quadratic program is given explicitly and very naturally

in terms of the coefficients of the max-plus basis functions appearing in the value determination step. We show that the algorithm converges in a finite number of steps and that the approximation error arising from projection of the converged solution is also of order  $\sqrt{\Delta t} + \Delta x(\Delta t)^{-1}$ . We leave open the question of how to formulate the policy iteration algorithm for the steady state solution to the corresponding infinite horizon differential game, although we make some comments in this direction at the end of the paper.

**2. Policy iteration.** In order to provide context and notation for what follows, we briefly set out here, in abstract terms, a policy iteration algorithm for games. This goes back to [6]. Here we have adapted it to a finite horizon context. In section 4, we will develop a concrete implementation of this algorithm adapted to the max-plus FEM.

Let  $S^t$  denote the evolution semigroup of the PDE (3). This associates to any function  $\phi$  the function  $v^t = v(\cdot, t)$ , where  $v$  is the value function of the differential game (1). Similarly, let  $S_{\hat{a}}^t$  denote the evolution semigroup of the PDE (5) for some fixed feedback function  $\hat{a}(\cdot)$ . This associates to any function  $\phi$  the function  $v_{\hat{a}}^t = v_{\hat{a}}(\cdot, t)$ , where  $v_{\hat{a}}$  is the value function of the optimal control problem (4). Maslov [7] observed that the semigroup  $S_{\hat{a}}^t$  is max-plus linear.

Now let  $p$  denote the cycle index within the policy iteration algorithm, and let  $q \in \{0, \dots, N-1\}$  denote the time step index, so that the full time horizon  $T$  is divided into  $N$  equal steps of length  $\tau$ , i.e.,  $T = N\tau$ , with the  $q$ th step running from  $q\tau$  to  $(q+1)\tau$ .

Suppose, on the  $p$ th cycle of policy iteration, we are given some fixed time dependent control policy  $\hat{a}_p = (\hat{a}_p^0(x), \dots, \hat{a}_p^{N-1}(x))$ . Note that, for each  $x \in X$ , this policy is constant on time step  $q\tau$  to  $(q+1)\tau$  for each  $q$ . Then, for a given  $q$ , we can consider the evolution semigroup  $S_{\hat{a}_p^q}^\tau$  of the PDE (5) with  $\hat{a} = \hat{a}_p^q$ . By application of the time iteration

$$v_p^{(q+1)\tau}(\cdot) = S_{\hat{a}_p^q}^\tau(v_p^{q\tau}(\cdot))$$

for  $q \in \{0, \dots, N-1\}$ , and defining  $v_p^0(\cdot) = \phi(\cdot)$ , we can then compute a time-discretization  $\{v_p^\tau(\cdot), \dots, v_p^{N\tau}(\cdot)\}$  of the value function  $v_{\hat{a}_p}^t$  of the optimal control problem (4) with time-dependent control policy  $\hat{a}_p$ . This is the value determination step of the algorithm, and it is max-plus linear.

The policy improvement step then involves computing the value function, which is obtained after “one step” of the differential game (1) if we start from function  $v_{\hat{a}_p}^t$ . This is achieved by computing

$$(6) \quad u_{p+1}^{(q+1)\tau}(\cdot) = S^\tau(v_p^{q\tau}(\cdot))$$

for  $q \in \{0, \dots, N-1\}$ . The policy improvement is then obtained by choosing a new policy  $\hat{a}_{p+1} = (\hat{a}_{p+1}^0(x), \dots, \hat{a}_{p+1}^{N-1}(x))$  which minimizes the right-hand side of (6) for each  $q \in \{0, \dots, N-1\}$ . This computation (6) is max-plus nonlinear. We will see below that it can be expressed as a quadratic program in the coefficients of a max-plus basis expansion of  $v_p^{q\tau}$ .

The algorithm is initiated by making some sensible choice of initial fixed time-dependent control policy  $\hat{a}_0 = (\hat{a}_0^0(x), \dots, \hat{a}_0^{N-1}(x))$ . On each new cycle  $p+1$  for  $p \geq 0$ , we start with  $v_{p+1}^0(\cdot) = v_p^0(\cdot) = \phi(\cdot)$ . Assume, with a view to induction on  $q$ , that  $v_{p+1}^{q\tau}(\cdot) \leq v_p^{q\tau}(\cdot)$  for  $q \geq 0$ . Then by the order preserving character of  $S_{\hat{a}}^t$ , we have

$$v_{p+1}^{(q+1)\tau} = S_{\hat{a}_{p+1}^q}^\tau(v_{p+1}^{q\tau}) \leq S_{\hat{a}_{p+1}^q}^\tau(v_p^{q\tau}).$$

Then since  $\hat{a}_{p+1}^q$  is the argmin in (6), we have

$$(7) \quad S_{\hat{a}_{p+1}^q}^\tau(v_p^{q\tau}) = S^\tau(v_p^{q\tau}) \leq S_{\hat{a}_p^q}^\tau(v_p^{q\tau}) = v_p^{(q+1)\tau}.$$

It then follows by induction on  $q$  that  $v_{p+1}^{q\tau}(\cdot) \leq v_p^{q\tau}(\cdot)$  for all  $q \in \{0, \dots, N\}$ ; i.e., the policy improvement step (6) does indeed lead to a lower value function of the optimal control problem (4) on the next iteration of value determination.

In general, it is necessary to impose conditions to guarantee that this process converges to a time discretization  $\{v^\tau(\cdot), \dots, v^{N\tau}(\cdot)\}$  of the value function  $v^t$  of the differential game (1). Typical conditions are that the sequence  $\{v_p^{q\tau}(\cdot)\}$  is uniformly bounded below and strictly monotone in  $p$ , which in turn usually follows by suitably restricting the class of control policies over which the minimization in (6) is performed.

In our case, given the particular structure of the differential game (1) and the specific discretizations which we apply within the context of the max-plus FEM, we show that the sequence of policy improvements actually terminates after  $N$  steps; i.e., the first iteration of policy improvement produces the optimal policy for the outer player on the first time step, the second iteration of policy improvement produces the optimal outer player policy on the second time step, and so on. This means that in practice the algorithm can be implemented without needing to cycle through policy iterations. Rather, it can be implemented like a value iteration with one pass through each of the  $N$  steps of the time discretization in turn, so that starting with initial data  $v_0^0(\cdot) = \phi(\cdot)$  and an initial choice of control policy  $\hat{a}_0^0(\cdot)$  at time zero, the value determination and policy improvement steps are applied to the one step game to compute the value function  $v_1^1(\cdot)$  and optimal outer player policy  $\hat{a}_1^1(\cdot)$  after one time step. These are then used as starting values for value determination and policy improvement applied again to the one step game to compute the value function  $v_2^2(\cdot)$  and optimal outer player policy  $\hat{a}_2^2(\cdot)$  after two time steps, and so on. However, in order to prove that this process is well defined, we consider below the full iteration of policy over the whole time horizon and show that the diagonal path stabilizes within this scheme. Also, the policy iteration framework allows for a potential generalization of the algorithm to deal with the infinite horizon problem, which we outline in the final section of the paper.

**3. The max-plus FEM.** We now briefly review the max-plus FEM set out in [1] for the numerical computation of the value function  $v_{\hat{a}}$  solving the optimal control problem (4). The various results on idempotent semirings, semimodules, linear maps, residuation, and projectors can be found in, for instance, [2, 3, 4].

Let  $\mathbb{R}_{\max}$  denote the idempotent semiring obtained from  $\mathbb{R}$ , with its usual order  $\leq$ , by defining idempotent addition as  $a \oplus b := \max(a, b)$  and multiplication as  $ab := a + b$ . Then let  $\bar{\mathbb{R}}_{\max} := \mathbb{R}_{\max} \cup \{+\infty\}$ , with the convention that  $-\infty$  is absorbing for the multiplication. This has the property of being complete as an idempotent semiring, which means that any subset has a least upper bound and the left and right multiplications  $x \mapsto ax$  and  $x \mapsto xa$  are residuated. This in turn means that the sets  $\{x \in \bar{\mathbb{R}}_{\max} : ax \leq b\}$  and  $\{x \in \bar{\mathbb{R}}_{\max} : xa \leq b\}$  both have maximal elements.

For a set  $X$ , we consider the set  $\bar{\mathbb{R}}_{\max}^X$  of  $\bar{\mathbb{R}}_{\max}$  valued functions on  $X$ . This is a semimodule over  $\bar{\mathbb{R}}_{\max}$  with respect to componentwise addition  $(u, v) \mapsto u \oplus v$ , defined by  $(u \oplus v)(x) = u(x) \oplus v(x)$ , and componentwise scalar multiplication  $(\lambda, u) \mapsto u\lambda$ , defined by  $(u\lambda)(x) = u(x)\lambda$ , where  $u, v \in \bar{\mathbb{R}}_{\max}^X$ ,  $\lambda \in \bar{\mathbb{R}}_{\max}$ , and  $x \in X$ . Note that the natural order on  $\bar{\mathbb{R}}_{\max}^X$  arising from the idempotent addition, i.e., the order defined by  $u \leq v \iff u \oplus v = v$ , corresponds to the componentwise partial order  $u \leq v \iff$

$u(x) \leq v(x)$  for all  $x \in X$ . Then  $\bar{\mathbb{R}}_{\max}^X$  has the property of being complete as a semi-module, which means again that any subset has a least upper bound (with respect to the natural order) and the right and left multiplications  $R_\lambda : \bar{\mathbb{R}}_{\max}^X \rightarrow \bar{\mathbb{R}}_{\max}^X : u \mapsto u\lambda$  and  $L_u : \bar{\mathbb{R}}_{\max} \rightarrow \bar{\mathbb{R}}_{\max}^X : \lambda \mapsto u\lambda$  are residuated. This in turn means that the sets  $\{u \in \bar{\mathbb{R}}_{\max}^X : u\lambda \leq v\}$  and  $\{\lambda \in \bar{\mathbb{R}}_{\max} : u\lambda \leq v\}$  both have maximal elements.

Now consider a map  $f : S \rightarrow T$  between partially ordered sets  $S$  and  $T$ . Suppose that  $f$  is monotone, i.e.,  $s \leq s' \Rightarrow f(s) \leq f(s')$ . Then  $f$  is said to be residuated if and only if for all  $t \in T$ , the set  $\{s \in S : f(s) \leq t\}$  has a maximal element. The residual map  $f^\# : T \rightarrow S$  is then defined as

$$f^\#(t) = \max\{s \in S : f(s) \leq t\}.$$

It follows that the left multiplication map  $L_u : \bar{\mathbb{R}}_{\max} \rightarrow \bar{\mathbb{R}}_{\max}^X$  is residuated. The action of the residual map  $L_u^\#$  on a given  $v \in \bar{\mathbb{R}}_{\max}^X$  is then defined as

$$L_u^\#(v) = \max\{\lambda \in \bar{\mathbb{R}}_{\max} : u\lambda \leq v\}.$$

Now let  $X$  and  $Y$  be sets and consider an operator  $A : \bar{\mathbb{R}}_{\max}^Y \rightarrow \bar{\mathbb{R}}_{\max}^X$  from  $\bar{\mathbb{R}}_{\max}$  valued functions on  $Y$  to  $\bar{\mathbb{R}}_{\max}$  valued functions on  $X$ . Such an operator is called linear if, for all  $u_1, u_2 \in \bar{\mathbb{R}}_{\max}^Y$ , and  $\lambda_1, \lambda_2 \in \bar{\mathbb{R}}_{\max}$ ,  $A(u_1\lambda_1 \oplus u_2\lambda_2) = A(u_1)\lambda_1 \oplus A(u_2)\lambda_2$ . Given some  $\bar{\mathbb{R}}_{\max}$  valued function  $a \in \bar{\mathbb{R}}_{\max}^{X \times Y}$  on  $X \times Y$ , we are then interested in the linear operator  $A : \bar{\mathbb{R}}_{\max}^Y \rightarrow \bar{\mathbb{R}}_{\max}^X$  with kernel  $a$  which maps any function  $u \in \bar{\mathbb{R}}_{\max}^Y$  to the function  $Au \in \bar{\mathbb{R}}_{\max}^X$  defined, in terms of the normal arithmetic operations on  $\mathbb{R}$ , by

$$(8) \quad Au(x) = \sup_{y \in Y} \{a(x, y) + u(y)\}.$$

Then, as shown in [3], this kernel operator  $A$  is residuated; i.e., for any  $v \in \bar{\mathbb{R}}_{\max}^X$ , the set  $\{u \in \bar{\mathbb{R}}_{\max}^Y : Au \leq v\}$  has a maximal element. The residual map  $A^\# : \bar{\mathbb{R}}_{\max}^X \rightarrow \bar{\mathbb{R}}_{\max}^Y$  then takes any  $v \in \bar{\mathbb{R}}_{\max}^X$  to this maximal element in  $\bar{\mathbb{R}}_{\max}^Y$  defined, again in terms of the normal arithmetic operations on  $\mathbb{R}$ , as the function

$$(9) \quad (A^\#v)(y) = \inf_{x \in X} \{-a(x, y) + v(x)\}.$$

The next notion to be introduced, for a kernel operator  $B : \bar{\mathbb{R}}_{\max}^Y \rightarrow \bar{\mathbb{R}}_{\max}^X$ , is that of projection on the image  $\text{im}B$  of  $B$ . The projector is denoted  $P_{\text{im}B}$  and is a map  $\bar{\mathbb{R}}_{\max}^X \rightarrow \bar{\mathbb{R}}_{\max}^X$  defined for all  $v \in \bar{\mathbb{R}}_{\max}^X$  by

$$P_{\text{im}B}(v) = \max\{w \in \text{im}B : w \leq v\}.$$

As shown in [4], this projector on the subsemimodule  $\text{im}B$  can be expressed as a composition

$$P_{\text{im}B} = B \circ B^\#$$

of  $B$  and its residual  $B^\#$ . If  $b(x, y)$  denotes the kernel of  $B$ , then this formula can be expressed, in the normal arithmetic of  $\mathbb{R}$ , as

$$(10) \quad B \circ B^\#(v)(x) = \sup_{y \in Y} \left( b(x, y) + \inf_{\xi \in X} (-b(\xi, y) + v(\xi)) \right).$$



Given a kernel operator  $C : \bar{\mathbb{R}}_{\max}^X \rightarrow \bar{\mathbb{R}}_{\max}^Z$  with kernel  $c(z, x)$ , we can consider the transposed operator  $C^* : \bar{\mathbb{R}}_{\max}^Z \rightarrow \bar{\mathbb{R}}_{\max}^X$  with kernel  $c^*(x, z) = c(z, x)$ . We can then define a dual projector on the  $\bar{\mathbb{R}}_{\min}$ -subsemimodule  $-\text{im}C^*$  in terms of

$$P^{-\text{im}C^*}(v) = \min\{w \in -\text{im}C^* : w \geq v\}$$

for all  $v \in \bar{\mathbb{R}}_{\max}^X$ . Then, as above, this projector can be expressed as a composition

$$P^{-\text{im}C^*} = C^\# \circ C,$$

which, in the normal arithmetic of  $\mathbb{R}$ , has the form

$$(11) \quad C^\# \circ C(v)(x) = \inf_{z \in Z} \left( -c(z, x) + \sup_{\xi \in X} (c(z, \xi) + v(\xi)) \right).$$

In greater generality than given here, it is proved in Theorem 1 of [1] that, with  $\Pi_B^C := P_{\text{im}B} \circ P^{-\text{im}C^*}$ , then for all  $v \in \bar{\mathbb{R}}_{\max}^X$ ,

$$(12) \quad \Pi_B^C(v) = \max\{w \in \text{im}B : Cw \leq Cv\}.$$

Now we can define the max-plus FEM for approximating the value function  $v_a^t = v_a(\cdot, t)$  for the optimal control problem (4). Let  $Y = \{1, \dots, I\}$ ,  $X = \mathbb{R}^n$ , and  $Z = \{1, \dots, J\}$ . Consider a family  $\{w_1, \dots, w_I\}$  of finite element functions  $w_i : X \rightarrow \bar{\mathbb{R}}_{\max}$ , and a family  $\{z_1, \dots, z_J\}$  of test functions  $z_j : X \rightarrow \bar{\mathbb{R}}_{\max}$ . The vectors  $\lambda = (\lambda_i)_{i=1, \dots, I} \in \bar{\mathbb{R}}_{\max}^I$  and  $\mu = (\mu_j)_{j=1, \dots, J} \in \bar{\mathbb{R}}_{\max}^J$  can be considered as  $\bar{\mathbb{R}}_{\max}$  valued functions on  $Y$  and  $Z$ , respectively. So, as above in (8), we can define max-plus kernel operators  $W : \bar{\mathbb{R}}_{\max}^Y \rightarrow \bar{\mathbb{R}}_{\max}^X$  and  $Z^* : \bar{\mathbb{R}}_{\max}^Z \rightarrow \bar{\mathbb{R}}_{\max}^X$  with kernels  $W = \text{col}(w_i)_{1 \leq i \leq I}$  and  $Z^* = \text{col}(z_j)_{1 \leq j \leq J}$ . The action of  $W$ , which plays the role of operator  $B$  above, is

$$W\lambda(x) = \sup_{i \in Y} \{w_i(x) + \lambda_i\},$$

while  $Z^*$  gives rise to the transposed operator  $Z : \bar{\mathbb{R}}_{\max}^X \rightarrow \bar{\mathbb{R}}_{\max}^Z$  which plays the role of operator  $C$  above, and acts as

$$(Zv)_j = \sup_{x \in X} \{z_j(x) + v(x)\} = \langle z_j | v \rangle,$$

where  $\langle \cdot | \cdot \rangle$  denotes the max-plus scalar product. Then from (10) and (11), we can give the specific form of the corresponding two projectors,

$$(13) \quad P_{\text{im}W}(v)(x) = \sup_{i \in Y} \left( w_i(x) + \inf_{\xi \in X} (-w_i(\xi) + v(\xi)) \right),$$

$$(14) \quad P^{-\text{im}Z^*}(v)(x) = \inf_{j \in Z} \left( -z_j(x) + \sup_{\xi \in X} (z_j(\xi) + v(\xi)) \right).$$

To start the algorithm, we approximate the initial data  $v_a^0 = \phi$  with the maximal element  $\leq v_a^0$  in the space  $\text{im}W$  spanned by the finite element functions. The approximation of  $v_a^0$  is denoted with a subscript  $h$  and takes the form

$$v_{ah}^0(x) = (W\lambda^0)(x) = \sup_{i \in Y} (w_i(x) + \lambda_i^0),$$

where the coefficients  $\lambda_i^0$  are determined from the residuation of  $W$  given in formula (9) as

$$(15) \quad \lambda_i^0 = \inf_{x \in X} (-w_i(x) + \phi(x)).$$

As an induction assumption, suppose that at time step  $q\tau$  we have a vector of coefficients  $\lambda_i^{q\tau}$  giving an approximation

$$v_{\hat{a}h}^{q\tau}(x) = \sup_{i \in Y} (w_i(x) + \lambda_i^{q\tau})$$

of  $v_{\hat{a}}^{q\tau}$  by the maximal element  $\leq v_{\hat{a}}^{q\tau}$  in the space  $\text{im}W$ . Then the approximation  $v_{\hat{a}h}^{(q+1)\tau}$  of  $v_{\hat{a}}^{(q+1)\tau}$  at the next time step can be calculated as

$$v_{\hat{a}h}^{(q+1)\tau}(\cdot) = P_{\text{im}W} \circ P^{-\text{im}Z^*} \circ S_{\hat{a}}^\tau \circ v_{\hat{a}h}^{q\tau}(\cdot)$$

The coefficients of this approximation are given, from (13) and (14), by

$$\lambda_i^{(q+1)\tau} = \inf_{\xi \in X} \left( -w_i(\xi) + \inf_{j \in Z} \left( -z_j(\xi) + \sup_{\eta \in X} (z_j(\eta) + S_{\hat{a}}^\tau \circ v_{\hat{a}h}^{q\tau}(\eta)) \right) \right).$$

It follows from (12) that  $v_{\hat{a}h}^{(q+1)\tau}$  is the maximal element in the space  $\text{im}W$  spanned by the finite element functions which satisfies  $Zv_{\hat{a}h}^{(q+1)\tau} \leq ZS_{\hat{a}}^\tau v_{\hat{a}h}^{q\tau}$ , i.e., the maximal element satisfying  $\langle z_j | v_{\hat{a}h}^{(q+1)\tau} \rangle \leq \langle z_j | S_{\hat{a}}^\tau v_{\hat{a}h}^{q\tau} \rangle$  for each test function  $z_j$ . So  $v_{\hat{a}h}^{(q+1)\tau}$  is the maximal solution to a max-plus variational formulation of the semigroup equation.

If (see section 3.3 of [1]) we further approximate the semigroup action  $S_{\hat{a}}^\tau v_{\hat{a}h}^{q\tau}$  by

$$(16) \quad \left( \tilde{S}_{\hat{a}}^\tau v_{\hat{a}h}^{q\tau} \right)(x) = \sup_{i \in Y} (w_i(x) + \lambda_i^{q\tau} + \tau H_{\hat{a}}(x, \partial w_i / \partial x)),$$

then  $\lambda_i^{(q+1)\tau}$  can be written explicitly as

$$(17) \quad \lambda_i^{(q+1)\tau} = \inf_{\xi \in X} \left( -w_i(\xi) + \inf_{j \in Z} \left( -z_j(\xi) + \sup_{\eta \in X} \left( z_j(\eta) + \sup_{k \in Y} (w_k(\eta) + \lambda_k^{q\tau} + \tau H_{\hat{a}}(\eta, \partial w_k / \partial x|_{\eta})) \right) \right) \right).$$

It is shown in [1] that the error  $\|v_{\hat{a}h}^T - v_{\hat{a}}^T\|_\infty$  on this approximation of the value function at time  $T$  is less than the sum of the approximation errors

$$\|P_{\text{im}W}(v_{\hat{a}}^{q\tau}) - v_{\hat{a}}^{q\tau}\|_\infty \quad \text{and} \quad \|P^{-\text{im}Z^*}(v_{\hat{a}}^{q\tau}) - v_{\hat{a}}^{q\tau}\|_\infty$$

for  $q = 0, \dots, N$ , arising from the projections on  $\text{im}W$  and  $-\text{im}Z^*$ , plus a term  $\sup_{i \in Y} \|\tilde{S}_{\hat{a}}^\tau w_i - S_{\hat{a}}^\tau w_i\|_\infty$  arising from the approximation of the semigroup action on the finite elements  $w_i$ .

In particular, if we choose two sets  $(\hat{x}_i)_{i \in Y}$  and  $(\hat{x}_j)_{j \in Z}$  of discretization points, and take the finite element functions to be  $w_i(x) = -\frac{c}{2} \|x - \hat{x}_i\|_2^2$  for some fixed constant  $c$ , and test functions to be  $z_j(x) = -\alpha \|x - \hat{x}_j\|_1$  for some fixed constant  $\alpha$ , then by Theorem 22 of [1] the error  $\|v_{\hat{a}h}^T - v_{\hat{a}}^T\|_\infty = O(\tau + \Delta x(\tau)^{-1})$ , where  $\Delta x$  is the maximal radius of the cells of the two Voronoi tessellations centered on the points  $(\hat{x}_i)_{i \in Y}$  and  $(\hat{x}_j)_{j \in Z}$ , respectively. See [10] for details on Voronoi tessellations. Below we will consider two cases for the test functions  $z_j$ , the first being the notationally simpler, limiting case where  $\alpha \rightarrow \infty$ , and the second being the more general case of  $\alpha < \infty$ .

#### 4. Policy iteration with max-plus FEM in the value determination step.

Recall from section 2 that  $p$  denotes the cycle index within the policy iteration algorithm, and  $q \in \{0, \dots, N-1\}$  denotes the time step index. Recall also that we restrict consideration of time-dependent feedback control policies  $a(x, t)$  to those in the form of sequences of  $N$  constant-in-time policy components  $(a^0(x), \dots, a^{N-1}(x))$ . In this section, we further restrict our choice of the individual policy components to functions  $a^q(\cdot)$  chosen from the set  $A = \{a(\cdot) : X \rightarrow U\}$  of functions which are locally constant with respect to  $x$  on cells of the Voronoi tessellation  $V_Z$  centered on the origins  $(\hat{x}_j)_{j \in Z}$  of the test functions  $z_j$ .

So suppose, as an induction hypothesis, that on iteration  $p$ , we have a set of constant vectors  $\{a_p^{qj}\}$  for  $q \in \{0, \dots, N-1\}$  and  $j \in Z$ , where each  $a_p^{qj} \in U \subseteq \mathbb{R}^m$ . These give rise to a fixed policy  $a_p$  which, for a given  $q$ , takes the form  $a_p^q(x) = a_p^{q\mu(x)}$ , where  $\mu(x) \in Z$  is the index of the cell of the Voronoi tessellation  $V_Z$  containing  $x$ . Note that the process can be initiated for  $p = 0$  by choosing some set of fixed vectors  $a^j$  (say, for all  $j$ , set them equal to the zero vector) such that  $a_0^q(x) = a^j$  for all  $q \in \{0, \dots, N-1\}$  and for all  $x \in \text{cell } j$  of  $V_Z$ , where cell  $j$  is the one centered on the origin  $\hat{x}_j$  of test function  $z_j$ .

**4.1. Value determination step.** The max-plus FEM outlined above can be applied to approximate the value function  $v_{a_p}^t$  solving the optimal control problem (4) with fixed strategy  $a_p$ . The coefficients of the expansion of this approximation, with respect to the finite elements  $w_i$ , are obtained as follows. For  $q = 0$  and  $i \in Y$ , the coefficients  $\lambda_{pi}^0 = \lambda_i^0$  are defined in (15) above. Then, using (17), for  $q \in \{0, \dots, N-1\}$  and  $i \in Y$  we get

$$\lambda_{pi}^{(q+1)\tau} = \inf_{\xi \in X} \left( -w_i(\xi) + \inf_{j \in Z} \left( -z_j(\xi) + \sup_{\eta \in X} \left( z_j(\eta) + \sup_{k \in Y} \left( w_k(\eta) + \lambda_{pk}^{q\tau} + \tau H_{a_p^q}(\eta, \partial w_k / \partial x|_{\eta}) \right) \right) \right) \right).$$

Note that in the Hamiltonian  $H_{a_p^q}$  we apply the policy  $a_p^q(\eta) = a_p^{q\mu(\eta)}$ , where  $\mu(\eta) \in Z$  is the index of the cell of the Voronoi tessellation  $V_Z$  containing  $\eta$ . The above can be rearranged to give

$$\begin{aligned} \lambda_{pi}^{(q+1)\tau} &= \inf_{j \in Z} \left( \inf_{\xi \in X} (-w_i(\xi) - z_j(\xi)) + \sup_{\eta \in X} \left( z_j(\eta) + \sup_{k \in Y} \left( w_k(\eta) + \lambda_{pk}^{q\tau} + \tau H_{a_p^q}(\eta, \partial w_k / \partial x|_{\eta}) \right) \right) \right) \\ &= \inf_{j \in Z} \left( -\langle w_i | z_j \rangle + \sup_{k \in Y} \left( \lambda_{pk}^{q\tau} + \sup_{\eta \in X} (z_j(\eta) + w_k(\eta) + \tau H_{a_p^q}(\eta, \partial w_k / \partial x|_{\eta})) \right) \right). \end{aligned}$$

For a given policy  $a$ , let

$$(18) \quad T_{jka} = \sup_{\eta \in X} (z_j(\eta) + w_k(\eta) + \tau H_a(\eta, \partial w_k / \partial x|_{\eta})).$$

In the normal max-plus FEM, the  $T_{jka}$  terms can be calculated offline. This would be difficult in the application of max-plus FEM to policy iteration, since we don't know

the policies  $a$  in advance. The relevant  $a$  for each  $p$  iteration is known at the start of that iteration and so, in principle, the next set of  $T_{jka}$  terms for a given  $a$  could be calculated at the start of that iteration. However, this would be slow. An alternative is to approximate the  $T_{jka}$  online by

$$(19) \quad \tilde{T}_{jka} = \langle z_j | w_k \rangle + \tau H_a \left( \eta_{jk}^{opt}, \partial w_k / \partial x |_{\eta_{jk}^{opt}} \right),$$

where  $\eta_{jk}^{opt} = \arg \sup \langle z_j | w_k \rangle = \arg \sup (z_j(\eta) + w_k(\eta))$ . Note that this approximation  $\tilde{T}$  is presented in [1], where it is shown in Theorem 22 that the resulting error estimate on the max-plus FEM deteriorates to  $\|v_{ah}^T - v_a^T\|_\infty = O(\sqrt{\tau} + \Delta x(\tau)^{-1})$ . So, finally, the coefficients of the expansion of the approximation to the value function  $v_{a_p}^t$  for fixed strategy  $a_p$  are given by

$$(20) \quad \lambda_{pi}^{(q+1)\tau} = \inf_{j \in Z} \left( -\langle w_i | z_j \rangle + \sup_{k \in Y} \left( \lambda_{pk}^{q\tau} + \tilde{T}_{jka_p^q} \right) \right).$$

**4.2. Policy improvement step.** For each  $i$  and  $q$ , there exists  $\bar{j}(iq) \in Z$ , which achieves the inf in (20), so that

$$(21) \quad \lambda_{pi}^{(q+1)\tau} = -\langle w_i | z_{\bar{j}} \rangle + \sup_{k \in Y} \left( \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}ka_p^q} \right).$$

For each  $k$ , the Hamiltonian within  $\tilde{T}_{\bar{j}ka_p^q}$  is evaluated at  $\eta_{jk}^{opt}$ , and so the strategy  $a_p^q$  applied in the Hamiltonian term takes the value  $a_p^{q\mu(\bar{j}k)}$ , where  $\mu(\bar{j}k) \in Z$  is the index of the cell of the Voronoi tessellation  $V_Z$  containing  $\eta_{jk}^{opt}$ .

Now suppose, as our first simpler case, that the constant term  $\alpha \rightarrow \infty$  in the test functions  $z_j(x) = -\alpha \|x - \hat{x}_j\|_1$ . The test functions are therefore defined as

$$(22) \quad z_j = \begin{cases} 0 & \text{at } x = \hat{x}_j, \\ -\infty & \text{otherwise} \end{cases}$$

for each  $j \in Z$ .

Then we have  $\eta_{jk}^{opt} = \hat{x}_{\bar{j}}$  for all  $k \in Y$ , and the index  $\mu(\bar{j}k) \in Z$  of the cell of the Voronoi tessellation  $V_Z$  containing  $\hat{x}_{\bar{j}}$  can be denoted simply  $\mu(\bar{j})$ . We note that  $\mu(\bar{j})$  can of course be represented simply as  $\bar{j}$ . However, it is useful for developing the later discussion of general test functions if we denote the cell index in the current discussion by  $\mu(\bar{j})$ .

It follows that  $a_p^q(\hat{x}_{\bar{j}}) = a_p^{q\mu(\bar{j})}$  is the policy value applied in the Hamiltonian term in  $\tilde{T}_{\bar{j}ka_p^q}$  for all  $k$ . So every term  $\tilde{T}_{\bar{j}ka_p^q}$  uses the same policy value  $a_p^{q\mu(\bar{j})}$  for all  $k \in Y$  within the  $\sup_{k \in Y}$  operation in (21).

Now let  $\bar{k}(iq) = \arg \sup_{k \in Y}$  in (21), so that

$$\lambda_{pi}^{(q+1)\tau} = -\langle w_i | z_{\bar{j}} \rangle + \lambda_{p\bar{k}}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a_p^q}.$$

Then we can improve the policy  $a_p^q$  in cell  $\mu(\bar{j})$  of  $V_Z$  by taking

$$(23) \quad \min_{a \in U} \tilde{T}_{\bar{j}\bar{k}a}$$

subject to

$$(24) \quad \lambda_{p\bar{k}}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a} \geq \lambda_{p\bar{k}}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a_p^q}$$

for all  $k \in Y$ , i.e., subject to

$$\lambda_{p\bar{k}}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a} \geq \sup_{k \neq \bar{k} \in Y} \left( \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}ka} \right).$$

This optimization is feasible since the current policy value  $a_p^{q\mu(\bar{j})}$  satisfies

$$\lambda_{p\bar{k}}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a_p^{q\mu(\bar{j})}} \geq \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}ka_p^{q\mu(\bar{j})}}$$

for all  $k \in Y$ . Furthermore, the function  $\tilde{T}_{\bar{j}\bar{k}a}$  is convex in  $a$ . In fact, the Hamiltonian within  $\tilde{T}_{\bar{j}\bar{k}a}$  is evaluated at fixed  $x = \hat{x}_{\bar{j}}$  and  $p = \partial w_{\bar{k}} / \partial x|_{\hat{x}_{\bar{j}}}$  and so is a positive definite quadratic form in  $a$ . Similarly, expanding the constraint (24) for each  $k$ , as is done below in section 4.3, shows that this is just a set of linear constraints in  $a$ . So the optimization (23) is a positive definite quadratic program and has a unique global minimum.

**THEOREM 4.1.** *For each  $q \in \{0, \dots, N-1\}$  and  $i \in Y$ , there is a  $\bar{j}(iq)$  and a  $\bar{k}(iq)$  defined as above. For this  $\bar{j}$  and  $\bar{k}$ , let  $\bar{a} = \arg \min_{a \in U} \tilde{T}_{\bar{j}\bar{k}a}$  subject to the constraint (24). For all  $x \in \text{cell}$  with index  $\mu(\bar{j})$  of the Voronoi tessellation  $V_Z$ , take new policy*

$$a_{p+1}^q(x) = a_{p+1}^{q\mu(\bar{j})} := \bar{a}.$$

*Note that for each  $q$ , there may be some remaining cells of  $V_Z$  whose indices  $\neq \mu(\bar{j}(iq))$  for any  $i \in Y$ . In these cells we leave the policy at time step  $q$  unchanged; i.e., if  $\mu^*$  is the index of such a cell, then for all  $x \in \text{cell}$  with index  $\mu^*$ ,*

$$a_{p+1}^q(x) = a_p^{q\mu^*}.$$

*Then the resulting new policy  $a_{p+1} = \{a_{p+1}^{qj}\}$  is an improvement on the old one  $a_p = \{a_p^{qj}\}$  in the sense that the corresponding functions  $v_{a_{(p+1)h}}^{q\tau}$  and  $v_{a_ph}^{q\tau}$ , i.e., the approximations to the value functions which solve the optimal control problem (4) with fixed policies  $a_{p+1}$  and  $a_p$ , respectively, satisfy*

$$(25) \quad v_{a_{(p+1)h}}^{q\tau} \leq v_{a_ph}^{q\tau}$$

*for all  $q \in \{0, \dots, N\}$ . Furthermore, this sequence of policy improvements terminates after at most  $N$  steps, so that the sequence  $\{v_{a_{Nh}}^{\tau}(\cdot), \dots, v_{a_{Nh}}^{N\tau}(\cdot)\}$  constitutes a time-discretized finite element approximation to the value function  $v^t$  of the differential game (1).*

*Proof.* First we deal with uniqueness. If, for a given  $q$ , there are two  $i$  giving rise to the same  $\bar{j}(iq)$ , then these both result in the same policy improvement  $a_{p+1}^q(x) = \bar{a}$  in cell  $\mu(\bar{j})$  since the term  $\tilde{T}_{\bar{j}\bar{k}a_p^q}$  in (21) does not depend on  $i$ . Similarly, if there are  $i_1$  and  $i_2$  giving rise to  $\bar{j}_1 = (i_1q) \neq \bar{j}(i_2q) = \bar{j}_2$ , then the corresponding cells  $\mu(\bar{j}_1)$  and  $\mu(\bar{j}_2)$  of  $V_Z$  are different, and so the optimizations (23) and (24), applied separately to  $\bar{j}_1$  and  $\bar{j}_2$ , give rise to policy improvements in different cells of  $V_Z$ .

Next, we show that the above quadratic program does lead to a policy improvement. Suppose, with a view to induction on  $q$ , that  $\lambda_{(p+1)i}^{q\tau} \leq \lambda_{pi}^{q\tau}$  for all  $i$ . Then

$$\begin{aligned}
 \lambda_{pi}^{(q+1)\tau} &= -\langle w_i | z_{\bar{j}} \rangle + \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a_p^q} \\
 &= -\langle w_i | z_{\bar{j}} \rangle + \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a_p^{q\mu(\bar{j})}} \\
 &\geq -\langle w_i | z_{\bar{j}} \rangle + \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}\bar{a}} \\
 &= -\langle w_i | z_{\bar{j}} \rangle + \sup_{k \in Y} \left( \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}\bar{a}} \right) \\
 &\geq -\langle w_i | z_{\bar{j}} \rangle + \sup_{k \in Y} \left( \lambda_{(p+1)k}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}\bar{a}} \right) \\
 &= -\langle w_i | z_{\bar{j}} \rangle + \sup_{k \in Y} \left( \lambda_{(p+1)k}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a_{p+1}^q} \right)
 \end{aligned}$$

since every term  $\tilde{T}_{\bar{j}\bar{k}a_{p+1}^q}$  uses the same policy value in the same cell  $\mu(\bar{j})$ . So

$$(26) \quad \lambda_{pi}^{(q+1)\tau} \geq \inf_{\bar{j}} \left( -\langle w_i | z_{\bar{j}} \rangle + \sup_{k \in Y} \left( \lambda_{(p+1)k}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a_{p+1}^q} \right) \right) = \lambda_{(p+1)i}^{(q+1)\tau}.$$

Since this holds for all  $i$ , it then follows that for any  $x \in X$ ,

$$\sup_i \left( \lambda_{(p+1)i}^{(q+1)\tau} + w_i(x) \right) \leq \sup_i \left( \lambda_{pi}^{(q+1)\tau} + w_i(x) \right),$$

i.e.,  $v_{a_{(p+1)h}}^{(q+1)\tau}(x) \leq v_{a_ph}^{(q+1)\tau}(x)$ . The proof follows by induction, after noting that from (15) the coefficients at the initial time step  $q = 0$  satisfy  $\lambda_{(p+1)i}^0 = \lambda_{pi}^0 = \lambda_i^0$ .

Lastly, we deal with termination in  $N$  steps. Again this is done by induction on  $q$ . To start with, note that the coefficients  $\lambda_i^0$  at the zeroth time step are independent of all choices of control policy and so constitute a terminal point for the policy iteration algorithm at step  $q = 0$  of the time discretization, i.e.,  $\lambda_{(p+1)i}^0 = \lambda_{pi}^0 = \lambda_i^0$  for all  $p \geq 0$ .

Suppose as an induction hypothesis that the coefficients  $\lambda_{pi}^{q\tau}$  for the  $q$ th time discretization step stabilize after  $p = q$  iterations of the above policy improvement process, i.e.,  $\lambda_{(p+1)i}^{q\tau} = \lambda_{pi}^{q\tau}$  for all  $p \geq q$ . Then consider what happens to  $\lambda_{(p+1)i}^{(q+1)\tau}$  for a given  $i \in Y$  during the  $(p+1)$ th policy improvement iteration. First, we identify the active  $\bar{j}(iq) \in Z$  and  $\bar{k}(iq) \in Y$ , which achieve the inf and sup in (20) and (21), respectively. Then we perform the minimization (23) subject to constraints (24) with these values of  $\bar{j}$  and  $\bar{k}$ . Now, as noted above, this is a positive definite quadratic program in  $a$  and so has a unique global minimum  $\bar{a}$ , which represents the policy to be applied at time step  $q$  in cell  $\mu(\bar{j})$  on iteration  $(p+1)$ . Furthermore, the  $\lambda_{pk}^{q\tau}$  terms appearing in (24) are all assumed by induction to be constant for all  $p \geq q$ , and all the other terms appearing in (23) and (24) are fixed by our choice of  $\bar{j}$  and  $\bar{k}$ . So the improved policy  $\bar{a}$ , computed on iteration  $(p+1)$  and to be applied at time step  $q$  in cell  $\mu(\bar{j})$ , is unique and will not change if recomputed for the same time step  $q$  and index  $\bar{j}$  at any later iteration  $(p+r) > (p+1)$  of the policy improvement cycle.

Now examination of (26) shows that, for each  $i$ , two things can happen on the  $(p+1)$ th policy improvement iteration during the improvement from  $\lambda_{pi}^{(q+1)\tau}$  to  $\lambda_{(p+1)i}^{(q+1)\tau}$ . First, the active  $\bar{j}(iq)$  which achieves the inf can remain the same. In this case, the above described unique improvement  $\bar{a}$  to the policy applied at time step

$q$  in cell  $\mu(\bar{j})$  continues to be the policy value applied in the Hamiltonian term appearing within  $\tilde{T}_{\bar{j}\bar{k}a_{p+1}^q}$  in the evaluation of  $\lambda_{(p+1)i}^{(q+1)\tau}$ . This then remains unchanged by subsequent policy improvement iterations.

Alternatively, the active  $\bar{j}(iq)$  which achieves the inf can change in (26) between iteration  $p$  and  $(p+1)$ . Let  $\bar{j}_p$  and  $\bar{j}_{p+1}$  denote the active  $\bar{j}$  on iterations  $p$  and  $(p+1)$ , respectively. Then on iteration  $p$  we have

$$\lambda_{pi}^{(q+1)\tau} = -\left\langle w_i | z_{\bar{j}_p} \right\rangle + \sup_{k \in Y} \left( \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}_p k a_p^q} \right),$$

while on iteration  $(p+1)$  we have

$$\begin{aligned} \lambda_{(p+1)i}^{(q+1)\tau} &= -\left\langle w_i | z_{\bar{j}_{p+1}} \right\rangle + \sup_{k \in Y} \left( \lambda_{(p+1)k}^{q\tau} + \tilde{T}_{\bar{j}_{p+1} k a_{p+1}^q} \right) \\ &= -\left\langle w_i | z_{\bar{j}_{p+1}} \right\rangle + \sup_{k \in Y} \left( \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}_{p+1} k a_{p+1}^q} \right) \end{aligned}$$

by our induction hypothesis. So for the active  $\bar{j}$  to change, the term

$$\sup_{k \in Y} \left( \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}_{p+1} k a_{p+1}^q} \right)$$

must have decreased during the  $(p+1)$ th policy improvement iteration. In other words,  $\bar{j}_{p+1} = \bar{j}(i^*q)$  must be the value of  $j$  which achieves the inf in (20) on iteration  $p$  for some other  $i^* \neq i$ . The optimization in (23) and (24) must therefore also be evaluated for this  $\bar{j}(i^*q)$ , on the same iteration  $(p+1)$  of policy improvement, to produce a unique improvement (denote it as  $\bar{a}^*$ ) applied at time step  $q$  in cell  $\mu(\bar{j}(i^*q))$ . The improvement  $\bar{a}^*$  in this particular cell is then applied in the Hamiltonian term appearing within  $\tilde{T}_{\bar{j}_{p+1}\bar{k}a_{p+1}^q}$  in the evaluation of  $\lambda_{(p+1)i}^{(q+1)\tau}$ . By the same argument as above, this policy value  $\bar{a}^*$ , applied at time step  $q$  in cell  $\mu(\bar{j}(i^*q))$ , then remains unchanged by subsequent policy improvement iterations.

It follows that in both cases  $\lambda_{(p+2)i}^{(q+1)\tau} = \lambda_{(p+1)i}^{(q+1)\tau}$  for  $p \geq q$ , i.e., for  $p+1 \geq q+1$ , and the induction step is proved. Since there is a total of  $N$  steps in the time discretization, the policy improvement algorithm converges after  $N$  steps.  $\square$

It follows from the above proof that, as mentioned at the end of section 2, the algorithm in practice can be implemented without needing to cycle through policy iterations, but rather can be implemented like a value iteration, with the above defined value determination and policy improvement steps applied to each step of the time discretization in turn.

**4.3. Quadratic program optimization.** The Hamiltonian appearing in (5) has the form

$$\begin{aligned} H_a(x, p) &= \max_b \{ p f_a(x, b) + l_a(x, b) \} \\ &= p(f + ga) + \frac{1}{2}x^2 + \frac{1}{2}a^2 + \frac{1}{2\gamma^2}phh^T p, \end{aligned}$$

and for  $z_j$  given by (22),  $\tilde{T}_{\bar{j}\bar{k}a}$  has the specific form

$$\tilde{T}_{\bar{j}\bar{k}a} = w_{\bar{k}}(\hat{x}_{\bar{j}}) + \tau H_a(\hat{x}_{\bar{j}}, \partial w_{\bar{k}} / \partial x |_{\hat{x}_{\bar{j}}}).$$

The policy improvement optimization set out in (23) and (24) can then be formulated as

$$\min_{a \in U} \tau H_a(\hat{x}_{\bar{j}}, \partial w_{\bar{k}} / \partial x |_{\hat{x}_{\bar{j}}})$$

subject to

$$\tau H_a(\hat{x}_{\bar{j}}, \partial w_{\bar{k}} / \partial x|_{\hat{x}_{\bar{j}}}) \geq \lambda_{pk}^{q\tau} - \lambda_{p\bar{k}}^{q\tau} + w_k(\hat{x}_{\bar{j}}) - w_{\bar{k}}(\hat{x}_{\bar{j}}) + \tau H_a(\hat{x}_{\bar{j}}, \partial w_k / \partial x|_{\hat{x}_{\bar{j}}})$$

for all  $k \in Y$ . This can be simplified into the quadratic program

$$\min_{a \in U} \left( \frac{\partial w_{\bar{k}}}{\partial x} ga + \frac{1}{2} a^2 \right)$$

subject to

$$\begin{aligned} \left( \frac{\partial w_{\bar{k}}}{\partial x} - \frac{\partial w_k}{\partial x} \right) ga &\geq \frac{1}{\tau} (w_k - w_{\bar{k}}) + \frac{1}{\tau} (\lambda_{pk}^{q\tau} - \lambda_{p\bar{k}}^{q\tau}) \\ &+ \frac{1}{2\gamma^2} \left( \frac{\partial w_k}{\partial x} - \frac{\partial w_{\bar{k}}}{\partial x} \right)^T h h^T \left( \frac{\partial w_k}{\partial x} + \frac{\partial w_{\bar{k}}}{\partial x} \right) \\ &+ \left( \frac{\partial w_k}{\partial x} - \frac{\partial w_{\bar{k}}}{\partial x} \right) f \end{aligned}$$

for all  $k \in Y$ , and evaluated at  $\hat{x}_{\bar{j}}$ . It can thus be seen that the constraints (24) are just a set of linear constraints on  $a$ .

It is interesting to examine the two cases under which the unique minimum of (23) and (24) is obtained. The first case is an unconstrained minimum and occurs at

$$a = -g^T(\hat{x}_{\bar{j}}) \frac{\partial w_{\bar{k}}}{\partial x} \Big|_{\hat{x}_{\bar{j}}} = c g^T(\hat{x}_{\bar{j}}) (\hat{x}_{\bar{j}} - \hat{x}_{\bar{k}}).$$

Now the max-plus FEM, with test functions given by (22), only evaluates the control feedback at the points  $\hat{x}_{\bar{j}}$ . So for  $x \in \text{cell } \mu(\bar{j})$  of  $V_Z$ , we can consider this unconstrained minimum to be an approximation to the feedback

$$(27) \quad \hat{a}(x) = -g^T(x) \frac{\partial w_{\bar{k}}}{\partial x} \Big|_x = c g^T(x) (x - \hat{x}_{\bar{k}}).$$

Furthermore, as in the proof of the above theorem, if this unconstrained minimum represents the feedback to be applied in cell  $\mu(\bar{j})$  at time step  $q$  and is obtained on iteration  $(p+1)$  of the policy improvement cycle, then we can assume that the value function  $v_{a_{ph}}^{q\tau}$  at time step  $q$  has already converged on iteration  $p$  to the finite element approximation to the value function for the differential game (1). In other words, on iteration  $(p+1)$ , the initial condition for the  $(q+1)$ th time step is the final approximation, obtained on iteration  $p$ , to the optimal value of the differential game at time step  $q$ . If we examine (16) and (17), we can see that this final approximation term, evaluated at  $\eta_{\bar{j}\bar{k}}^{opt} = \hat{x}_{\bar{j}}$ , is given by

$$(28) \quad v_{a_{ph}}^{q\tau}(\hat{x}_{\bar{j}}) = w_{\bar{k}}(\hat{x}_{\bar{j}}) + \lambda_{p\bar{k}}^{q\tau},$$

where  $\lambda_{p\bar{k}}^{q\tau}$  is the final value of the  $\bar{k}$ th coefficient of the value function at time step  $q$ . Now, by the Pontryagin maximum principle, along an extremal of the Hamilton–Jacobi–Isaacs equation (3),  $a(\cdot)$  should satisfy (27). Within the cell  $\mu(\bar{j})$ , the initial condition for the Cauchy problem (3) is given (approximately) by the smooth function (28), and so for sufficiently small time step  $\tau$ , this extremal will be unique, and therefore optimal. Hence, the unconstrained minimum represents an approximation



to the classical optimal feedback solution to the problem (1) in the case where there is a unique extremal over time step  $\tau$ , as identified, for instance, in [12, 13].

The second case is a constrained minimum and occurs when

$$(29) \quad \lambda_{p\bar{k}}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a} = \lambda_{pk}^{q\tau} + \tilde{T}_{\bar{j}ka}$$

for some  $k \in Y$  with  $k \neq \bar{k}$ . To understand what is happening here, we note that the term  $\tilde{T}_{\bar{j}ka}$  defined in (19) is an approximation to the term  $T_{\bar{j}ka}$  defined in (18), which in turn, if we go back to (16) and (17), can be seen to be the  $\bar{j}$ th coefficient of the projection on the space of test functions of the approximation of the semigroup

$$\tilde{S}_a^\tau w_k$$

at time step  $q+1$ , acting on the basis function  $w_k$  at time step  $q$ . If we combine this with (28), then we can see that the left-hand side of (29) is the approximate evaluation at  $\eta_{\bar{j}\bar{k}}^{opt} = \hat{x}_{\bar{j}}$  of the semigroup  $\tilde{S}_a^\tau$  acting on the initial condition  $w_{\bar{k}} + \lambda_{p\bar{k}}^{q\tau}$ , while the right-hand side is evaluation at  $\eta_{\bar{j}k}^{opt} = \hat{x}_{\bar{j}}$  of the same semigroup acting on the initial condition  $w_k + \lambda_{pk}^{q\tau}$ . So the left-hand side of (29) corresponds to the action of the semigroup on the particular component of the basis expansion of  $v_{a_p h}^{q\tau}$  which turns out to be active in cell  $\mu(\bar{j})$  at time step  $q+1$ , namely, the component corresponding to the  $\bar{k}$  which achieves the sup in (21). As the policy  $a$  applied during the  $(q+1)$ th step is reduced in the  $(p+1)$ th cycle of policy iteration, the constraint in (29) becomes active for some  $k \neq \bar{k}$  when we can no longer uniquely identify the source of the active component in cell  $\mu(\bar{j})$  at time step  $q+1$ , i.e., when there are two extremals through  $\hat{x}_{\bar{j}}$ , one corresponding to the semigroup action in time  $\tau$  starting from initial data  $w_{\bar{k}} + \lambda_{p\bar{k}}^{q\tau}$  and the other corresponding to the same semigroup action over the same time period but starting from initial data  $w_k + \lambda_{pk}^{q\tau}$ .

**4.4. General case  $z_j$  test functions.** Recall from above that we have

$$\tilde{T}_{jka} = \langle z_j | w_k \rangle + \tau H_a \left( \eta_{jk}^{opt}, \partial w_k / \partial x |_{\eta_{jk}^{opt}} \right)$$

and

$$\eta_{jk}^{opt} = \arg \sup \langle z_j | w_k \rangle,$$

and assume now that for each  $j \in Z$ ,  $z_j(x) = -\alpha \|x - \hat{x}_j\|_1$  for some constant  $\alpha < \infty$ .

In order to obtain the above theorem in this more general case, we need only one condition, namely, that for a given  $j$ , the set of points  $\eta_{jk}^{opt} : k \in Y$  must lie in the same cell of the Voronoi tessellation  $V_Z$  as the origin  $\hat{x}_j$  of the test function  $z_j$ . We denoted the index of this cell above as  $\mu(\bar{j})$ . This can clearly be arranged by a suitable choice of the two grids of origins  $\hat{x}_k$  and  $\hat{x}_j$  of the basis and test functions, respectively, since this choice is made in advance of the execution of the algorithm. We give here a simple sufficient condition which guarantees slightly more than this, namely, that for a given  $j$ ,  $\eta_{jk}^{opt} = \hat{x}_j$  for all  $k \in Y$ .

**LEMMA 4.2.** *Suppose that the fixed constant  $\alpha$  which appears in each test function  $z_j$  satisfies  $\frac{c}{2} \text{diam} X \leq \alpha$ , where  $c$  is the fixed constant which appears in each basis function  $w_k$  and  $X$  is the region of state space within which the problem is defined. Then for a given  $k \in Y$  and  $j \in Z$ , the function  $w_k(x) + z_j(x) = -\frac{c}{2} \|x - \hat{x}_k\|_2^2 - \alpha \|x - \hat{x}_j\|_1$  has a global maximum at  $\hat{x}_j$ .*

*Proof.*

$$\begin{aligned}
 -\frac{c}{2} \|\hat{x}_j - \hat{x}_k\|_2^2 - \alpha \|\hat{x}_j - \hat{x}_j\|_1 &= -\frac{c}{2} \|\hat{x}_j - \hat{x}_k\|_2^2 \\
 &\geq -\frac{c}{2} \|x - \hat{x}_k\|_2^2 - \frac{c}{2} \|x - \hat{x}_j\|_2^2 \\
 &\geq -\frac{c}{2} \|x - \hat{x}_k\|_2^2 - \frac{c}{2} \text{diam} X \|x - \hat{x}_j\|_1 \\
 &\geq -\frac{c}{2} \|x - \hat{x}_k\|_2^2 - \alpha \|x - \hat{x}_j\|_1. \quad \square
 \end{aligned}$$

Now we can repeat the argument of section 4.2. For each  $i$  and  $q$ , there exists  $\bar{j}(iq) \in Z$  which achieves the inf in (20). Then for each  $k$ , the Hamiltonian within the  $\tilde{T}_{\bar{j}ka_p^q}$  term is evaluated at  $\eta_{\bar{j}k}^{opt}$ , and by our assumption above, the grids of origins have been chosen so as to ensure that  $\eta_{\bar{j}k}^{opt} \in \text{cell } \mu(\bar{j})$  of  $V_Z$  for each  $k \in Y$ . It follows that the strategy  $a_p^q$  applied in the Hamiltonian term within  $\tilde{T}_{\bar{j}ka_p^q}$  takes the value  $a_p^{q\mu(\bar{j})}$ . Then let  $\bar{k}(iq) = \arg \sup_{k \in Y}$  in (21), so that

$$\lambda_{pi}^{(q+1)\tau} = -\langle w_i | z_{\bar{j}} \rangle + \lambda_{p\bar{k}}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a_p^q}.$$

Then we can improve the policy  $a_p^q$  in cell  $\bar{\mu}(\bar{j})$  of  $V_Z$  by taking

$$\min_{a \in U} \tilde{T}_{\bar{j}\bar{k}a}$$

subject to

$$\lambda_{p\bar{k}}^{q\tau} + \tilde{T}_{\bar{j}\bar{k}a} \geq \lambda_{pk}^{q\tau} + \tilde{T}_{jka}$$

for all  $k \in Y$ . As above, this optimization is a positive definite quadratic program and so has a unique minimum. So Theorem 4.1 still holds for this more general set of test functions; i.e., the above minimization gives a strict policy improvement on each iteration and terminates after at most  $N$  steps.

**5. Error estimation.** Recall that we are denoting by  $v^t$  the value function of the differential game (1). The projected policy iteration algorithm outlined in the previous section converges in a finite number of steps to a time-discretized finite element approximation to  $v^t$ , which we denote by  $v_h^t$ . Suppose that the final optimal policy for the projected policy iteration algorithm is  $a_1$ , and that the true optimal policy, in the same class of policies, for the exact problem is  $a_2$ . Then we can consider  $v_h^t$  to be the value function of the optimal control problem obtained by fixing  $a_1$  as the policy of the outer player in the projected version of the game, i.e.,  $v_h^t = v_{a_1h}^t$ . Similarly, we can consider  $v^t$  to be the value function of the optimal control problem obtained by fixing  $a_2$  as the policy of the outer player in the exact version of the game, i.e.,  $v^t = v_{a_2}^t$ . The semigroup  $S_{a_2}$  associated with this exact optimal control problem is then max-plus linear. Then we need to estimate

$$\|S_{a_1h} - S_{a_2}\|_\infty \leq \|S_{a_1h} - S_{a_2h}\|_\infty + \|S_{a_2h} - S_{a_2}\|_\infty,$$

where  $S_{a,h} = P_{\text{im}W} \circ P^{-\text{im}Z^*} \circ S_a$  is the projection of the semigroup action  $S_a$  with policy  $a$ . The second term on the right-hand side is the FEM error already estimated in Theorem 22 of [1], namely,  $O(\sqrt{\tau} + \Delta x(\tau)^{-1})$ , where the constant in the big- $O$  term

is expressed in terms of  $\text{diam}X$ , with the constants  $c$  and  $\alpha$  appearing in the basis and test functions, and Lipschitz constants and bounds for the two functions  $f_{a_2}$  and  $l_{a_2}$  appearing in (4). So to estimate the total error, it suffices to prove the following.

PROPOSITION 5.1.  $\|S_{a_1h} - S_{a_2h}\|_\infty = 0$ .

*Proof.* If  $a_1$  is also the final optimal policy for the exact algorithm, then  $a_1 = a_2$  and we are done.

Suppose  $a_1$  is not the optimal policy for the exact problem. Then since  $a_2$  has to be an improvement on  $a_1$ , we have  $S_{a_2} \leq S_{a_1}$  with respect to the natural order on  $\bar{\mathbb{R}}_{\max}^X$ . It follows from the lemma below that  $S_{a_2h} \leq S_{a_1h}$ . But since  $a_1$  is the final optimal policy for the projected algorithm, we must also have  $S_{a_1h} \leq S_{a_2h}$ . Hence  $S_{a_1h} = S_{a_2h}$ , and the result is proved.  $\square$

LEMMA 5.2. *For arbitrary residuated (kernel) operators  $B : \bar{\mathbb{R}}_{\max}^Y \rightarrow \bar{\mathbb{R}}_{\max}^X$  and  $C : \bar{\mathbb{R}}_{\max}^X \rightarrow \bar{\mathbb{R}}_{\max}^Z$  between complete semimodules of  $\bar{\mathbb{R}}_{\max}$  valued functions on sets  $Y$ ,  $X$ , and  $Z$ , let  $v_h = \Pi_B^C(v) = P_{\text{im}B} \circ P^{-\text{im}C^*} \circ v$ . If functions  $v_1 \geq v_2$  in the natural order on  $\bar{\mathbb{R}}_{\max}^X$ , then  $v_{1,h} \geq v_{2,h}$ .*

*Proof.* Recall from above that

$$\begin{aligned} P_{\text{im}B}(v) &= \max\{w \in \text{im}B : w \leq v\}, \\ P^{-\text{im}C^*}(v) &= \min\{w \in -\text{im}C^* : w \geq v\}. \end{aligned}$$

So, for  $v_1 \geq v_2$ , it follows that

$$(30) \quad P^{-\text{im}C^*}(v_1) \geq P^{-\text{im}C^*}(v_2).$$

Now

$$\begin{aligned} P_{\text{im}B} \circ P^{-\text{im}C^*}(v_1) &= \max\{w \in \text{im}B : w \leq P^{-\text{im}C^*}(v_1)\}, \\ P_{\text{im}B} \circ P^{-\text{im}C^*}(v_2) &= \max\{w \in \text{im}B : w \leq P^{-\text{im}C^*}(v_2)\}. \end{aligned}$$

By (30), any  $w$  in the second set must be  $\leq P^{-\text{im}C^*}(v_1)$ . Hence, any such  $w$  must be  $\leq \max\{w \leq P^{-\text{im}C^*}(v_1)\} = P_{\text{im}B} \circ P^{-\text{im}C^*}(v_1)$ , and so  $\max\{\text{any such } w\}$  must be  $\leq P_{\text{im}B} \circ P^{-\text{im}C^*}(v_1)$ , i.e.,  $P_{\text{im}B} \circ P^{-\text{im}C^*}(v_2) \leq P_{\text{im}B} \circ P^{-\text{im}C^*}(v_1)$  as required.  $\square$

**6. Conclusion.** In this paper, we have set out a policy iteration algorithm for the solution of a nonconvex nonlinear-affine finite horizon differential game. This involves the use of a max-plus finite element method (FEM) to solve a convex optimal control problem in the value determination step of the algorithm, where this convex problem is the result of fixing the policy of the second, or outer, player. A quadratic program is then solved in the policy improvement step in order to improve the control feedback. We have shown that the algorithm converges in a finite number of steps and that the approximation error at convergence is of order  $\sqrt{\tau} + \Delta x(\tau)^{-1}$ .

We make some brief remarks here on the question of how to formulate the policy iteration algorithm for the steady state solution to the corresponding infinite horizon differential game. The results of [9] give conditions under which the infinite horizon convex nonlinear-affine optimal control problem, arising from fixing the policy of the outer player, has a unique steady state solution. Furthermore, this reference gives an algorithm, which converges in a finite number of steps, for the computation of a max-plus approximation to this steady state solution. As remarked in section 4 of [1], this approximation can be considered as the limit of the above max-plus FEM in the case where the operator  $Z$  is the identity, i.e., where the space of test functions is the set of all functions. So if the infinite horizon differential game can be formulated in such a way that the conditions of [9] are satisfied by the corresponding optimal

control problem for all relevant choices of fixed outer player policy, then we have a guaranteed unique fixed point for each such problem, i.e., a unique solution to the max-plus linear problem

$$v_p = S_{\hat{a}_p}^\tau(v_p)$$

obtained by setting the outer player policy to  $\hat{a}_p$ . Furthermore we have an approximation to  $v_p$  in the form of a max-plus expansion relative to the basis of finite element functions  $\{w_i\}$ . As in section 2 above, we then compute the one step iteration

$$u_{p+1} = S^\tau(v_p)$$

of the nonconvex game and select a new (stationary) policy  $\hat{a}_{p+1}$ , which minimizes the right-hand side of the iteration. If we approximate  $v_p$  by its max-plus basis expansion, and use the max-plus FEM to compute the evolution of the coefficients  $\lambda_i$  of the max-plus basis expansion under the action of the semigroup with a given outer player policy, then we can use the policy improvement method of section 4.2 to determine the minimizing policy via a quadratic program. We can then compute the fixed point of the new convex optimal control problem

$$v_{p+1} = S_{\hat{a}_{p+1}}^\tau(v_{p+1})$$

obtained by setting the outer player policy to  $\hat{a}_{p+1}$ , again using the algorithm of [9]. Since

$$v_p = S_{\hat{a}_p}^\tau(v_p) \geq S^\tau(v_p) = u_{p+1} = S_{\hat{a}_{p+1}}^\tau(v_p)$$

it follows from Lemma 24 of [5] that  $v_p$  is  $\geq$  the fixed point of  $S_{\hat{a}_{p+1}}^\tau$ , i.e.,  $v_p \geq v_{p+1}$ . This descending sequence is bounded below by the fixed point of the nonconvex game, the existence of which on a nonlocal neighborhood of the origin is given by the results of [8] for the particular nonlinear affine form of game considered here, or more generally, by the results of [11].

Finally, by an argument similar to that described in the second to last paragraph of section 2, we could then hope to prove that the policy iteration on the infinite horizon differential game converges, subject to the following two pieces of work being successfully completed. The first, as mentioned above, is to formulate conditions on the infinite horizon differential game to ensure that the results of [9] can be applied to deduce unique fixed points for the sequence of infinite horizon convex optimal control problems. The second is to work out the details of combining the max-plus approximation of [9], for solving the eigenvector equation  $v_p = S_{\hat{a}_p}^\tau(v_p)$ , with the max-plus FEM approximation to the one step iteration  $u_{p+1} = S^\tau(v_p)$  of the nonconvex game and the policy improvement quadratic program for determining a new minimizing policy. This is left to a future paper.

**Acknowledgment.** I would like to thank Stephane Gaubert for useful conversations during the development of this paper.

#### REFERENCES

- [1] M. AKIAN, S. GAUBERT, AND A. LAKHOUA, *The max-plus finite element method for solving deterministic optimal control problems: Basic properties and convergence analysis*, SIAM J. Control Optim., 47 (2008), pp. 817–848.

- [2] M. AKIAN, S. GAUBERT, AND V. KOLOKOLTSOV, *Set coverings and invertibility of functional Galois connections*, in Idempotent Mathematics and Mathematical Physics, V. P. Maslov and G. L. Litvinov, eds., Contemp. Math. 377, American Mathematical Society, Providence, RI, 2005, pp. 19–51.
- [3] F. BACCELLI, G. COHEN, G. J. OLSDER, AND J.-P. QUADRAT, *Synchronization and Linearity: An Algebra for Discrete Events Systems*, John Wiley & Sons, New York, 1992.
- [4] G. COHEN, S. GAUBERT, AND J.-P. QUADRAT, *Duality and separation theorem in idempotent semimodules*, Linear Algebra Appl., 379 (2004), pp. 395–422.
- [5] S. GAUBERT AND J. GUNAWARDENA, *A non-linear hierarchy for discrete event dynamical systems*, in Proceedings of the Fourth Workshop on Discrete Event Systems (WODES'98), Cagliari, Italy, 1998. Available online at <http://amadeus.inria.fr/gaubert/papers.html>.
- [6] A. J. HOFFMAN AND R. M. KARP, *On nonterminating stochastic games*, Management Sci., 12 (1966), pp. 359–370.
- [7] V. P. MASLOV, *On a new principle of superposition for optimisation problems*, Russ. Math. Surveys, 42 (1987), pp. 43–54.
- [8] D. MCCAFFREY, *Geometric existence theory for the control-affine  $H_\infty$  problem*, J. Math. Anal. Appl., 324 (2006), pp. 682–695.
- [9] W. M. MCENEANEY, *Max-plus eigenvector representations for solution of nonlinear  $H_\infty$  problems: Basic concepts*, IEEE Trans. Automat. Control, 48 (2003), pp. 1150–1163.
- [10] J.-R. SACK AND J. URRUTIA, *Handbook of Computational Geometry*, North-Holland, Amsterdam, 2000.
- [11] P. SORAVIA,  *$H_\infty$  control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.
- [12] A. J. VAN DER SCHAFT, *On a state space approach to nonlinear  $H_\infty$  control*, Systems Control Lett., 16 (1991), pp. 1–8.
- [13] A. J. VAN DER SCHAFT,  *$L_2$  gain analysis of nonlinear systems and nonlinear state feedback  $H_\infty$  control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.

## DISSIPATIVITY OF UNCONTROLLABLE SYSTEMS, STORAGE FUNCTIONS, AND LYAPUNOV FUNCTIONS\*

DEBASATTAM PAL<sup>†</sup> AND MADHU N. BELUR<sup>†</sup>

**Abstract.** Dissipative systems have played an important role in the analysis and synthesis of dynamical systems. The commonly used definition of dissipativity often requires an assumption on the controllability of the system. In this paper we use a definition of dissipativity that is slightly different (and less often used in the literature) to study a linear, time-invariant, possibly uncontrollable dynamical system. We provide a necessary and sufficient condition for an uncontrollable system to be strictly dissipative with respect to a supply rate under the assumption that the uncontrollable poles are not “mixed”; i.e., no pair of uncontrollable poles is symmetric about the imaginary axis. This condition is known to be related to the solvability of a Lyapunov equation; we link Lyapunov functions for autonomous systems to storage functions of an uncontrollable system. The set of storage functions for a controllable system has been shown to be a convex bounded polytope in the literature. We show that for an uncontrollable system the set of storage functions is unbounded, and that the unboundedness arises precisely due to the set of Lyapunov functions for an autonomous linear system being unbounded. Further, we show that stabilizability of a system results in this unbounded set becoming bounded from below. Positivity of storage functions is known to be very important for stability considerations because the maximum stored energy that can be drawn out is bounded when the storage function is positive. In this paper we establish the link between stabilizability of an uncontrollable system and existence of positive definite storage functions. In most of the results in this paper, we assume that no pair of the uncontrollable poles of the system is symmetric about the imaginary axis; we explore the extent of necessity of this assumption and also prove some results for the case of single output systems regarding this necessity.

**Key words.** dissipativity, uncontrollability, storage functions, behaviors, algebraic Riccati equation, Hamiltonian matrix, Lyapunov equation

**AMS subject classifications.** 93D05, 15A63, 93B05, 93B07, 15A18, 15A03

**DOI.** 10.1137/070699019

**1. Introduction.** Dissipativity of dynamical systems helps in the analysis and design of control systems. Dissipativity theory allows problems like LQR, circle criterion, Popov criterion, passivity synthesis,  $\mathcal{H}_\infty$  control, and Riccati inequalities to be analyzed under a common framework. An important assumption in some of these developments is that of controllability of the dynamical system. In this paper we study dissipativity of general linear time-invariant systems, possibly uncontrollable.

Uncontrollable systems arise naturally in the process of modeling dynamical systems. The inability to shape one or more system variables in an arbitrary desired fashion is frequently encountered in systems. For example, loss of controllability could happen to otherwise controllable systems when certain system parameters satisfy relevant equations arising in controllability check methods: see a simple electrical circuit below in section 5.1. Uncontrollability could also arise generically due to *structural* inabilities to influence one or more system variables. (See [14, 12] and references therein about structural controllability studies.)

In the context of synthesis of a dynamical system, one sometimes has to settle for an uncontrollable realization of a given transfer function: the case of nonminimality of

---

\*Received by the editors August 1, 2007; accepted for publication (in revised form) July 1, 2008; published electronically December 5, 2008. The research was supported in part by SERC division, Department of Science and Technology, India.

<http://www.siam.org/journals/sicon/47-6/69901.html>

<sup>†</sup>The authors are with the Department of Electrical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India (debpal@ee.iitb.ac.in, belur@ee.iitb.ac.in).

the transformerless synthesis of a positive real transfer function is well known. This issue concerns the synthesis of a positive real transfer matrix using only resistors, capacitors, and inductors. The currently known methods (the Bott–Duffin method [3] and its variants) bring us naturally to systems that are dissipative, but are not controllable. See [6] for a recent overview about this classical problem.

Dissipativity of a system is about the absence of any source of energy within the system, and hence all interactions with the environment have to satisfy the condition that the “net energy” is directed inwards. This is made precise below in Definition 3.1. Such a property is intrinsic to the system and therefore should be independent of the question of controllability. For example, a passive electrical network made out of passive circuit elements must continue to be dissipative even if it loses controllability. In this paper we consider a general linear time-invariant system and work on a theory of dissipativity free from any controllability assumption. Our work is based on the signature characteristic of a dissipative system to store energy, i.e., existence of a storage function. An important issue that immediately arises is whether to include unobservable variables to describe this storage of energy (see [33, 5]). Our main result sorts out this issue: for the case of strict dissipativity, we show that a storage function depending only on the manifest variables suffices, and no unobservable variables are necessary (see Remark 5.1).

The present theory of dissipative systems is well-developed primarily for controllable systems because it is possible there to define dissipativity without taking recourse to the existence of a storage function. This is done using an integral inequality involving only the compactly supported trajectories allowed by the system. This definition turns out to be inadequate for a general, possibly uncontrollable, linear behavior. In order to overcome this inadequacy, there has been prior work of taking existence of storage functions satisfying a dissipation inequality as a definition of dissipativity; see [28, 5], for example. In this paper we further develop using this definition, prove results regarding existence, and relate it to the situation of controllability. The principal finding is that a certain condition on the uncontrollable poles, which we call as the *unmixing* condition, plays a key role. If no pair of the uncontrollable poles of the system is symmetric with respect to the imaginary axis, then the noncontrollability poses no hindrance to strict dissipativity, i.e., the strict dissipativities of the behavior and its controllable part are equivalent (Theorem 3.4). This result is utilized to show useful identities about positive storage functions and unboundedness of the set of storage functions for the case of uncontrollability.

The paper is structured as follows. The rest of this section has a few words about the notation we follow. Section 2 contains some preliminaries we require regarding behavioral theory. The next section (section 3) has some definitions that we need in order to state the main result of this paper and for the proofs. In this section we also present the main result: a necessary and sufficient condition for a general linear time-invariant system to be strictly dissipative with respect to a supply rate that depends on the manifest variables, under the assumption that the set of uncontrollable poles satisfies the unmixing condition. Interestingly, this unmixing condition on the uncontrollable poles is reminiscent of the solvability condition of Lyapunov equations: this is elaborated in sections 3 and 8. This paper utilizes the wealth of existing literature on Hamiltonian matrices and Riccati equations; section 4 has results about relations among Riccati equations, the Hamiltonian matrix, and dissipativity. A proof of the main result follows in section 5, together with some auxiliary results. In section 6 we present a necessary and sufficient condition for existence of positive storage functions: here we relate stabilizability to positive storage functions. In section 7 we present

some insight on the nature, namely, unboundedness and convexity, of the set of all storage functions of an uncontrollable dissipative behavior. (The set of storage functions is known to be *bounded* in the case of controllability.) Section 8 explores into the extent of necessity of the unmixing property that we have assumed throughout this paper. In this section we show an interesting result about rank one symmetric matrices and the solvability of the Lyapunov equation. When one or more pairs of the uncontrollable poles have symmetry about the imaginary axis, it turns out that the solvabilities of a certain Riccati equation and the corresponding Riccati *inequality* differ significantly from the situation in the controllable case. We conclude the paper in section 9 following which is an appendix containing some proofs and peripheral results needed for the proofs.

The notation we follow is standard.  $\mathbb{R}$  and  $\mathbb{C}$  stand for the fields of real and complex numbers. The ring of polynomials in  $\xi$  with real coefficients is denoted by  $\mathbb{R}[\xi]$ .  $\mathbb{R}^{p \times w}[\xi]$  stands for the set of  $p \times w$  matrices with entries from  $\mathbb{R}[\xi]$ . In the context of quadratic differential forms, we require polynomials in two indeterminates:  $\zeta$  and  $\eta$ . The set of such polynomials with real coefficients is denoted by  $\mathbb{R}[\zeta, \eta]$ , and the set of  $w \times w$  matrices with entries from  $\mathbb{R}[\zeta, \eta]$  by  $\mathbb{R}^{w \times w}[\zeta, \eta]$ .  $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^w)$  denotes the space of all infinitely often differentiable functions from  $\mathbb{R}$  to  $\mathbb{R}^w$ , and  $\mathcal{D}(\mathbb{R}, \mathbb{R}^w)$  denotes its subspace of all compactly supported trajectories. We use  $\bullet$  when it is unnecessary to specify a dimension. For example,  $R \in \mathbb{R}^{\bullet \times w}$  means  $R$  is a real matrix with  $w$  columns. When dealing with many variables, in order to keep track of the dimensions, we use the same letter as a generic variable  $w$ , but in typewriter font  $\mathfrak{w}$ , to denote the number of components; for example,  $w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathfrak{w}})$ . In the context of stability, we require certain regions of the complex plane  $\mathbb{C}$ . The open left and right half complex planes are denoted by  $\mathbb{C}^-$  and  $\mathbb{C}^+$ , respectively. To improve readability within text, we use  $\text{col}(\cdot, \cdot)$  to stack its arguments into a column, i.e.,  $\text{col}(w_1, w_2) = [w_1^T \ w_2^T]^T$ .

**2. Behaviors, QDFs, and state representations.** A linear differential behavior  $\mathfrak{B}$  is defined to be the subspace of  $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathfrak{w}})$  consisting of the solutions to a set of ordinary linear differential equations with constant coefficients; i.e.,

$$\mathfrak{B} := \left\{ w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathfrak{w}}) \mid R \left( \frac{d}{dt} \right) w = 0 \right\},$$

where  $R(\xi)$  is a polynomial matrix having  $\mathfrak{w}$  number of columns:  $R \in \mathbb{R}^{\bullet \times \mathfrak{w}}[\xi]$ . We shall denote the set of linear differential behaviors with  $\mathfrak{w}$  number of variables by  $\mathfrak{L}^{\mathfrak{w}}$ . The linear differential behavior  $\mathfrak{B} \in \mathfrak{L}^{\mathfrak{w}}$  can also be written as  $\mathfrak{B} = \ker R(\frac{d}{dt})$ . That is why this representation is called a *kernel representation* of  $\mathfrak{B}$ . We call  $w$  the *manifest variable*; these are the variables of interest. In this paper,  $w$  is the variable through which the system exchanges energy with the environment. It turns out that we can assume, without loss of generality, that  $R(\xi)$  is of full row rank (see [17]); in this paper, a kernel representation matrix  $R(\xi)$  is assumed to be of full row rank. For a behavior  $\mathfrak{B} = \ker R(\frac{d}{dt})$ , the row rank of  $R(\xi)$  gives the output cardinality (number of outputs in the system). Though the variables  $w$  can often be partitioned into inputs and outputs in more than one way, the output cardinality remains the same:  $\text{rank } R$ . Further, the cardinality does not depend on the  $R$  used to define it, but depends only on  $\mathfrak{B}$ . In this sense, the output cardinality is an integer invariant of  $\mathfrak{B}$  and we denote it by  $\mathfrak{p}(\mathfrak{B})$ . The number of inputs to the system, the input cardinality, is another integer invariant of  $\mathfrak{B}$ . This integer is denoted by  $\mathfrak{m}(\mathfrak{B})$  and is calculated using  $\mathfrak{m}(\mathfrak{B}) = \mathfrak{w} - \mathfrak{p}(\mathfrak{B})$ , where  $\mathfrak{w}$  is the number of components in the manifest variable  $w$ .



A concept of central importance for this paper is that of controllability. A behavior  $\mathfrak{B} \in \mathfrak{L}^w$  is said to be *controllable* if for every  $w', w'' \in \mathfrak{B}$ , there exists a  $w \in \mathfrak{B}$  and a  $\tau > 0$  such that

$$\begin{aligned} w(t) &= w'(t) & \text{for all } t \leq 0, \\ &= w''(t) & \text{for all } t \geq \tau. \end{aligned}$$

We denote the set of all controllable behaviors with  $w$  variables as  $\mathfrak{L}_{\text{cont}}^w$ . A behavior  $\mathfrak{B} = \ker R(\frac{d}{dt})$  is controllable if and only if  $R(\lambda)$  does not lose rank for any  $\lambda \in \mathbb{C}$ . An important characterization of controllable behaviors is that they also admit *image representations*. It was shown in [31] that  $\mathfrak{B}$  is controllable if and only if it can be represented as

$$\mathfrak{B} := \left\{ w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^w) \mid \exists \ell \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^m) \text{ such that } w = M \left( \frac{d}{dt} \right) \ell \right\},$$

for some polynomial matrix  $M \in \mathbb{R}^{w \times m}[\xi]$ . This representation of  $\mathfrak{B} \in \mathfrak{L}_{\text{cont}}^w$  is called an image representation. It turns out that for an image representation, without loss of generality, one can assume  $M(\xi)$  to have the property that  $M(\lambda)$  has full column rank for every  $\lambda \in \mathbb{C}$ . We call such an  $M$  satisfying this property a *right-prime* polynomial matrix.  $M$  being right-prime means that we are able to deduce the  $\ell$  trajectory corresponding to a  $w$  trajectory satisfying the equation  $w = M(\frac{d}{dt})\ell$ . Hence, such an  $M$  is also said to induce an *observable* image representation.

In the context of uncontrollable systems, we use the key notion of *uncontrollable poles* and *uncontrollable characteristic polynomial*. Suppose  $\mathfrak{B} = \ker R(\frac{d}{dt})$  and suppose  $\mathfrak{B}$  is *not* controllable. Then there exist one or more complex numbers  $\lambda$  such that  $R(\lambda)$  loses rank. These complex numbers, together with multiplicities,<sup>1</sup> are defined as uncontrollable poles in the definition below. Uncontrollable poles are the roots of a monic polynomial called the uncontrollable characteristic polynomial. See [32] for details.

**DEFINITION 2.1.** Let  $R \in \mathbb{R}^{p \times w}[\xi]$  have full row rank and suppose  $R(\frac{d}{dt})w = 0$  is a kernel representation for  $\mathfrak{B}$ . Consider a factorization of  $R$  into  $R(\xi) = F(\xi)R_{\text{cont}}(\xi)$  such that  $R_{\text{cont}} \in \mathbb{R}^{p \times w}[\xi]$ ,  $R_{\text{cont}}(\lambda)$  has full row rank for every complex number  $\lambda$ , and  $\det F$  is a monic polynomial. The uncontrollable characteristic polynomial of  $\mathfrak{B}$ , denoted by  $\chi_{\text{un}}(\mathfrak{B})$ , is defined as  $\det F$ . The set of uncontrollable poles is defined as roots of  $\chi_{\text{un}}$ , and is denoted by  $\Lambda_{\text{un}}(\mathfrak{B})$ .

If the behavior  $\mathfrak{B}$  is clear from the context, we write just  $\chi_{\text{un}}$  and  $\Lambda_{\text{un}}$ . Notice that if  $\mathfrak{B}$  is controllable, then  $\chi_{\text{un}} = 1$ . When a behavior is not controllable, we often require the *controllable part* of  $\mathfrak{B}$ . This is the largest controllable behavior contained in  $\mathfrak{B}$ ; the controllable part of  $\mathfrak{B}$  is denoted by  $\mathfrak{B}_{\text{cont}}$ . Consider the above definition in which  $R$  has been factorized as described to obtain  $R_{\text{cont}}$ . A kernel representation for  $\mathfrak{B}_{\text{cont}}$  is induced by  $R_{\text{cont}}$ . For a detailed exposition on behaviors, controllability, and uncontrollable characteristic polynomial, we refer the reader to [17, 32].

This paper deals with dissipativity and in this context we deal with quadratic forms in the system variables and a finite number of their derivatives. It turns out to be very natural to associate two variable polynomial matrices to such quadratic forms. Consider a two variable polynomial matrix  $\Phi(\zeta, \eta) := \sum_{i,k} \Phi_{ik} \zeta^i \eta^k \in \mathbb{R}^{w \times w}[\zeta, \eta]$ , where  $\Phi_{ik} \in \mathbb{R}^{w \times w}$ . A Quadratic Differential Form (QDF)  $Q_\Phi$  induced by  $\Phi(\zeta, \eta)$  is a

<sup>1</sup>See Remark 4.2 below.

map  $Q_\Phi : \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^w) \rightarrow \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R})$  defined by

$$Q_\Phi(w) := \sum_{i,k} \left( \frac{d^i w}{dt^i} \right)^T \Phi_{ik} \left( \frac{d^k w}{dt^k} \right).$$

When dealing with quadratic forms in  $w$  and its derivatives, we can assume without loss of generality that  $\Phi(\zeta, \eta) = \Phi^T(\eta, \zeta)$ . We call such  $\Phi(\zeta, \eta)$  a symmetric two-variable polynomial matrix, and we denote the set of all such symmetric two-variable polynomial matrices by  $\mathbb{R}_s^{w \times w}[\zeta, \eta]$ . A quadratic form induced by a real symmetric constant matrix  $S \in \mathbb{R}_s^{w \times w}$  is a special QDF and we shall often need this in this paper. For a given  $\Phi \in \mathbb{R}_s^{w \times w}[\zeta, \eta]$ , we often require the one variable polynomial matrix  $\Phi(-\xi, \xi)$ : we shall denote this by  $\partial\Phi(\xi)$ . Due to the symmetry of  $\Phi(\zeta, \eta)$  the one variable polynomial matrix  $\partial\Phi(\xi)$  is *para-Hermitian*, i.e.,  $\partial\Phi(-\xi) = \partial\Phi^T(\xi)$ . Notice that this property makes  $\partial\Phi(j\omega)$  Hermitian for all  $\omega \in \mathbb{R}$ . Throughout this paper, ample use is made of the well-developed theory of QDFs; only the essential results of which are reviewed here. See [28] for a thorough and complete treatment on QDFs.

The notion of state is central to this paper due to the claim that, for uncontrollable systems, also the storage of energy is possible due to memory elements in the system. The state variable  $x$  is an auxiliary variable that relates to the memory of the system. Consider a behavior  $\mathfrak{B} \in \mathfrak{L}^w$  with manifest variables  $w$ . A variable  $x$  of the system is called a state variable if it satisfies the system equations together with  $w$ , and has the concatenation property. More precisely, if  $(w', x')$  and  $(w'', x'')$  are two smooth trajectories allowed by the system, and  $x'(0) = x''(0)$ , then the new trajectory  $(w, x)$  formed by concatenating  $(w', x')$  and  $(w'', x'')$  at  $t = 0$ , i.e.,

$$\begin{aligned} (w, x)(t) &= (w', x')(t) && \text{for all } t \leq 0 \\ &= (w'', x'')(t) && \text{for all } t > 0, \end{aligned}$$

also satisfies the system equations in a distributional sense. A formal treatment on this is contained in [19], where it was proved that a variable  $x$  is a state variable for  $\mathfrak{B}$  if and only if the behavior satisfies

$$\mathfrak{B} = \left\{ w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^w) \mid \exists x \in \mathfrak{C}^\infty \text{ such that } E \frac{d}{dt} x + Fx + Gw = 0 \right\}$$

for suitable real constant matrices  $E, F$  and  $G$ . The above first order representation is called a state representation. Such a representation is said to be minimal if it has the least number of state variables among all state representations describing the behavior. The dimension of the state space for a minimal state representation is defined to be the *McMillan degree* of the behavior, and is denoted by  $\mathbf{n}(\mathfrak{B})$ . It was shown in [19] that a set of state variables can be obtained through the manifest variables by a *state map*,  $X \in \mathbb{R}^{\bullet \times w}[\xi]$ , which gives the state variables by  $x = X(\frac{d}{dt})w$ . Among all state maps, if  $X$  has the minimum number of rows, then it is called a minimal state map; in this case, the number of rows equals  $\mathbf{n}(\mathfrak{B})$ . If  $\mathfrak{B} \in \mathfrak{L}^w$  has an *input/output* partition  $w = \text{col}(w_1, w_2)$  where  $w_1$  is input and  $w_2$  is output such that the transfer function from  $w_1$  to  $w_2$  is *proper*, then  $\mathfrak{B}$  admits a minimal state representation that is more special and well known: an *input/state/output* (i/s/o) representation

$$\frac{d}{dt}x = Ax + Bw_1, \quad w_2 = Cx + Dw_1,$$

with  $(C, A)$  observable. Controllability of the behavior  $\mathfrak{B}$  and that of the pair  $(A, B)$  are related as shown in the following well-known result. We write  $\Lambda_{\text{un}}^{(A, B)}$  for the set of

uncontrollable eigenvalues of the  $(A, B)$  pair (counted with multiplicities). Proposition 2.2 shows when a behavior allows a state map that gives rise to an i/s/o representation with observability properties (see [31, 19]). We state this as a proposition for easy reference later in this paper.

**PROPOSITION 2.2.** *Let behavior  $\mathfrak{B} \in \mathfrak{L}^w$  be given by  $\mathfrak{B} = \ker R(\frac{d}{dt})$ , where  $R(\xi)$  has full row rank. Let  $(w_1, w_2)$  be an input/output partition of  $w$  such that the resulting transfer function matrix is proper. Suppose  $n$  is the McMillan degree of  $\mathfrak{B}$  and  $X \in \mathbb{R}^{n \times w}[\xi]$  gives a minimal state map. Then there exist  $A \in \mathbb{R}^{n \times n}$ ,  $B \in \mathbb{R}^{n \times m}$ ,  $C \in \mathbb{R}^{p \times n}$ , and  $D \in \mathbb{R}^{p \times m}$ , such that  $\frac{d}{dt}x = Ax + Bw_1$ ,  $w_2 = Cx + Dw_1$  with  $(C, A)$  observable, and  $x = X(\frac{d}{dt})w$ . Moreover,  $\mathfrak{B}$  is controllable if and only if the above i/s/o representation is state-controllable. If  $\mathfrak{B}$  is uncontrollable, then  $\Lambda_{\text{un}}(\mathfrak{B})$ , the set of uncontrollable poles of  $\mathfrak{B}$ , is equal to  $\Lambda_{\text{un}}^{(A, B)}$  the set of uncontrollable eigenvalues of  $A$  (counted with multiplicities).*

**3. Dissipative systems: Definition and main result.** Dissipative systems are those that have no source of energy within, and hence any energy stored within the system has to have been supplied from its environment. This intuitive physical concept was made concrete in [30, 28] using the *dissipation inequality*: the rate of increase of stored energy is at most the power supplied to the system. In this paper, the power supplied and the stored energy are both QDFs in the manifest variables  $w$  of the system. (See Remark 5.1 below regarding storage function's dependence on just the manifest variables.) In this paper we use the following definition of dissipativity; its relation to other definitions is discussed below.

**DEFINITION 3.1.** *A linear differential behavior  $\mathfrak{B} \in \mathfrak{L}^w$  is said to be dissipative with respect to supply rate  $S \in \mathbb{R}_s^{w \times w}$  if there exists a quadratic differential form  $Q_\Psi(w)$  such that*

$$(3.1) \quad \frac{d}{dt}Q_\Psi(w) \leq Q_S(w) \text{ for all } w \in \mathfrak{B}.$$

*The quadratic differential form  $Q_\Psi$  is called a storage function for  $\mathfrak{B}$  with respect to the supply rate  $S$ .*

The inequality (3.1) above is called the dissipation inequality. In some control problems like in LQR and the suboptimal  $\mathcal{H}_\infty$  control, a stricter notion of dissipativity plays a key role. In this paper we shall deal primarily with strict dissipativity, although many of our results are valid for just dissipativity also (see Remark 5.7 below). We define strict dissipativity as follows.

**DEFINITION 3.2.** *A linear differential behavior  $\mathfrak{B} \in \mathfrak{L}^w$  is said to be strictly dissipative with respect to  $S \in \mathbb{R}_s^{w \times w}$  if there exists an  $\epsilon > 0$  and a storage function  $Q_\Psi(w)$  such that*

$$\frac{d}{dt}Q_\Psi(w) \leq Q_S(w) - \epsilon|w|^2 \quad \text{for all } w \in \mathfrak{B}.$$

Because the above definitions require the existence of a hitherto unknown storage function, it has been common to use an equivalent statement for the definition of (strict) dissipativity when dealing with *controllable* systems. The following result from [28] shows the equivalence.

**PROPOSITION 3.3.** *Let  $\mathfrak{B} \in \mathfrak{L}_{\text{cont}}^w$  and  $S \in \mathbb{R}_s^{w \times w}$  be nonsingular. Then the following statements are equivalent.*

1. *There exists a storage function  $Q_\Psi(w)$  such that  $\frac{d}{dt}Q_\Psi(w) \leq Q_S(w) - \epsilon|w|^2$  for all  $w \in \mathfrak{B}$ .*

2. For all  $w \in \mathfrak{B} \cap \mathfrak{D}$  the integral inequality  $\int_{\mathbb{R}} Q_S(w) dt \geq \epsilon \int_{\mathbb{R}} |w|^2 dt$  is satisfied.

The above proposition shows that the existence of a storage function satisfying the dissipation inequality is equivalent to saying that the total energy transferred into the system is strictly positive whenever we start the system from rest and bring the system back to rest. Statement 2 was used as the definition of strict dissipativity in [28]. With  $\epsilon = 0$ , we get the definition of nonstrict dissipativity given in [28, 29]. It is important to note here that the second statement above holds over only compactly supported trajectories in  $\mathfrak{B}$ , while the first holds for all  $w \in \mathfrak{B}$ . Controllability of  $\mathfrak{B}$  is crucial for the compactly supported trajectories in  $\mathfrak{B}$  to be representative enough of the whole behavior for the above equivalence to hold (see [16]). Definitions using an integral over a finite interval (which means energy supplied for a finite period of time) ends up having initial and final storage function values in the defining inequality. Use of compactly supported trajectories makes the integrals over the whole of  $\mathbb{R}$  well-defined and also makes the defining integral inequality free from the initial and final storage function values. However, for an uncontrollable behavior, Statement 2 of Proposition 3.3 puts no restrictions on the trajectories in the behavior which are outside the controllable part (see [16]), and hence this cannot be used as a definition of dissipativity.

We define signature of a real symmetric nonsingular matrix  $S$ , denoted by  $\sigma(S)$  as the pair of integers  $\sigma(S) = (\sigma_-(S), \sigma_+(S))$ , where  $\sigma_-(S)$  and  $\sigma_+(S)$  are the number of negative and positive eigenvalues of  $S$ , respectively. In this paper we shall deal only with the case when the positive signature  $\sigma_+(S)$  equals the input cardinality  $\mathfrak{m}(\mathfrak{B})$  of the behavior  $\mathfrak{B}$ . Dissipativity with respect to this matrix is required for the  $\mathcal{H}_\infty$  norm of a corresponding transfer matrix to be at most one. Further, as shown in [29, Proposition 2, Part I],  $\sigma_+(S) = \mathfrak{m}(\mathfrak{B})$  means that the behavior has as high an input cardinality as  $S$ -dissipativity allows; we call this condition the *maximum input cardinality condition*.

We are now ready to state one of the main results of this paper. The following theorem tells that if a certain unmixing condition is satisfied for the uncontrollable poles, then the controllable part of a behavior being strictly dissipative is equivalent to the existence of a storage function for the whole behavior's strict dissipativity. Recall from Definition 2.1 that the uncontrollable characteristic polynomial  $\chi_{\text{un}}$  of  $\mathfrak{B}$  is the monic polynomial whose roots (with suitable multiplicities) are those complex numbers where  $R(\xi)$  loses rank. Theorem 3.4 below states that if the uncontrollable poles are such that no pair of the uncontrollable poles is symmetric with respect to the imaginary axis, then noncontrollability of  $\mathfrak{B}$  poses no hindrance to strict dissipativity of  $\mathfrak{B}$ ; i.e., strict dissipativities of  $\mathfrak{B}$  and  $\mathfrak{B}_{\text{cont}}$  are equivalent.

**THEOREM 3.4.** *Consider a linear differential behavior  $\mathfrak{B} \in \mathfrak{L}^w$  and a nonsingular  $S \in \mathbb{R}_s^{w \times w}$  with the input cardinality of  $\mathfrak{B}$  equal to the positive signature of  $S$ :  $\mathfrak{m}(\mathfrak{B}) = \sigma_+(S)$ . Assume that the uncontrollable characteristic polynomial of  $\mathfrak{B}$ ,  $\chi_{\text{un}}$ , is such that  $\chi_{\text{un}}(\xi)$  and  $\chi_{\text{un}}(-\xi)$  are coprime. Then,  $\mathfrak{B}$  is strictly  $S$ -dissipative if and only if its controllable part  $\mathfrak{B}_{\text{cont}}$  is strictly  $S$ -dissipative.*

We call the condition of coprimeness of  $\chi_{\text{un}}(\xi)$  and  $\chi_{\text{un}}(-\xi)$  the *unmixing condition*. In the context of autonomous systems, it is well known (see [34], for example) that the unmixing condition is a necessary and sufficient condition for the existence of a unique solution to the Lyapunov equation. In section 8 we explore the extent of necessity of this condition. For some autonomous behaviors (and hence some uncontrollable behaviors), we show that the unmixing condition is not necessary (section 8).

Throughout this paper, we shall assume  $S$  has the following form:

$$(3.2) \quad \Sigma := \begin{bmatrix} I_m & 0 \\ 0 & -I_p \end{bmatrix}.$$

Lemma 3.5 below shows that, for dissipativity considerations, taking  $\Sigma$  as in (3.2) is without any loss of generality. Dissipativity with respect to a different constant matrix  $S$  can be easily treated by modifying the behavior suitably, as shown in the following lemma.

**LEMMA 3.5.** *Consider  $\mathfrak{B} \in \mathfrak{L}^w$  and a real symmetric nonsingular matrix  $S \in \mathbb{R}^{w \times w}$ . Let  $S = T^T \Sigma T$ , with  $T \in \mathbb{R}^{w \times w}$  nonsingular, be a symmetric factorization of  $S$ . Define  $\tilde{\mathfrak{B}} := T\mathfrak{B}$ . Then  $\mathfrak{B}$  is  $S$ -dissipative if and only if  $\tilde{\mathfrak{B}}$  is  $\Sigma$ -dissipative.*

Before we continue with other preliminaries, results, and proofs, we list the assumptions we make in the rest of this paper that are without loss of any generality. While we do write these standing assumptions explicitly in some results, they are sometimes skipped for brevity. For a kernel representation, the polynomial matrix  $R(\xi)$  is assumed to be full row rank, and the matrix  $M(\xi)$  in an image representation is assumed to be right-prime. Further, when we start with an i/s/o representation of a behavior, we assume this to be minimal. In the context of dissipativity,  $\Sigma$  is used for the supply rate. We assume  $\Sigma$  to be symmetric and nonsingular.

**4. Dissipativity, Riccati equation, and Hamiltonian matrix.** In this section we first briefly review existing literature about how the dissipation inequality gives us a well-studied Linear Matrix Inequality (LMI). We bring out the connection between a certain para-Hermitian polynomial matrix related to the behavior and an associated Hamiltonian matrix. See [8, 13, 25] for related results.

For the case that the input cardinality of the behavior is equal to the positive signature of  $\Sigma$ , a necessary condition for dissipativity of  $\mathfrak{B} \in \mathfrak{L}_{\text{cont}}^w$  is that a partition of  $w = (w_1, w_2)$  corresponding to the matrix  $\Sigma$  (see (3.2)) results in an input/output partition for  $\mathfrak{B}$  such that  $w_1$  is input and  $w_2$  is output (see [28, Remark 5.11]). Note that the above partition means  $Q_\Sigma(w) = |w_1|^2 - |w_2|^2$ . Further, due to the dissipativity, the transfer function from  $w_1$  to  $w_2$  turns out to be *proper*. This implies that  $\mathfrak{B}$  allows an i/s/o representation as

$$(4.1) \quad \frac{d}{dt}x = Ax + Bw_1, \quad w_2 = Cx + Dw_1,$$

with  $(C, A)$  observable (see Proposition 2.2). The link among dissipativity of a controllable behavior, storage functions, and LMIs is the subject of [30, 21, 4]; we state this result below for easy reference.

**PROPOSITION 4.1.**  *$\mathfrak{B} \in \mathfrak{L}_{\text{cont}}^w$  is  $\Sigma$ -dissipative if and only if there exists a solution  $K = K^T \in \mathbb{R}^{n \times n}$  for the following LMI*

$$(4.2) \quad \begin{bmatrix} (C^T C + A^T K + K A) & (K B + C^T D) \\ (B^T K + D^T C) & -(I_m - D^T D) \end{bmatrix} \leq 0,$$

where  $\frac{d}{dt}x = Ax + Bw_1$  and  $w_2 = Cx + Dw_1$  is a state-controllable and state-observable i/s/o representation of  $\mathfrak{B}$  with the input/output partition induced by the block matrix  $\Sigma$ .

In the above proposition, the memoryless state function  $x^T K x$  acts as a storage function for the controllable behavior  $\mathfrak{B}$ . See Corollary 5.6 below for our result regarding uncontrollable behaviors. The above LMI is the well-known *bounded-real*

LMI. Assume  $(I_m - D^T D) > 0$  (some implications of this assumption will be clarified in the next section). The Schur complement of  $(I_m - D^T D)$  in inequality (4.2) gives the Algebraic Riccati Inequality (ARI)<sup>2</sup>

$$(4.3) \quad \begin{aligned} & \left( A + B(I_m - D^T D)^{-1} D^T C \right)^T K + K \left( A + B(I_m - D^T D)^{-1} D^T C \right) \\ & + C^T (I_p - DD^T)^{-1} C + KB(I_m - D^T D)^{-1} B^T K \leq 0. \end{aligned}$$

Note that  $K$  satisfies the ARI if and only if  $K$  satisfies the above LMI.

The corresponding equation is the Algebraic Riccati Equation (ARE), and we use properties of this equation for various results in this paper. Interestingly, the solution to the ARE can be found from certain  $n$ -dimensional invariant subspaces of a  $2n \times 2n$  matrix known as the *Hamiltonian* matrix. This paper uses properties of the Hamiltonian matrix to relate to dissipativity. The procedure of constructing a solution to the ARE from an  $n$ -dimensional eigenspace of the Hamiltonian matrix comes in several texts, for example, [7, 10]. For easy reference we present this result as Proposition 4.4 below. We first define *Lambda-sets* of the roots of an even polynomial  $p(\xi)$  having no roots on the imaginary axis, for it will be of importance in the sequel.

*Remark 4.2.* As a convention in this paper, a set of roots of a polynomial (or that of eigenvalues of a real constant matrix) has every element appearing as many number of times as its multiplicity (algebraic multiplicity in case of eigenvalues), and therefore equality of such sets means equality with the multiplicities counted. This helps avoid writing certain polynomials are equal after ensuring monicity.

The definition of a Lambda-set plays an important role in the partition of a set of complex numbers which are symmetric with respect to the imaginary axis. This notion is similar to that of an  $S$ -set [20].  $\bar{\Lambda}$  below denotes the set of complex conjugates of the elements in  $\Lambda$ .

**DEFINITION 4.3.** Let  $p(\xi)$  be a nonzero even polynomial in  $\xi$  with no roots on the imaginary axis. A set of complex numbers  $\Lambda \subset \text{roots}(p(\xi))$  is said to be a Lambda-set of roots  $(p(\xi))$  if it satisfies the following properties:

1.  $\Lambda = \bar{\Lambda}$ ,
2.  $\Lambda \cap (-\Lambda) = \emptyset$ , and
3.  $\Lambda \cup (-\Lambda) = \text{roots}(p(\xi))$  (counted with multiplicity).

The disjointness condition (condition 2) requires that  $p$  has no roots on the imaginary axis. We called this the unmixing condition in the remark following Theorem 3.4. Proposition 4.4 below is well known; it relates ARE solutions to the *Hamiltonian* matrix  $H$ , defined below.

**PROPOSITION 4.4.** Consider the ARE:  $A^T K + KA + C^T C + KBB^T K = 0$  in the unknown real symmetric matrix  $K = K^T \in \mathbb{R}^{n \times n}$ . Corresponding to this ARE, construct the Hamiltonian matrix,  $H := \begin{bmatrix} A & BB^T \\ -C^T C & -A^T \end{bmatrix}$ . Assume  $H$  does not have eigenvalues on the imaginary axis and let  $\Lambda$  be a Lambda-set of  $\text{spec}(H)$ . Suppose the  $n$ -dimensional  $H$ -invariant subspace corresponding to  $\Lambda$  is given by

$$(4.4) \quad \mathcal{X}_\Lambda(H) := \text{im} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

where  $X_1, X_2 \in \mathbb{R}^{n \times n}$ . A real symmetric solution  $K$  to the ARE satisfying  $\text{spec}(A + BB^T K) = \Lambda$  exists if and only if  $X_1$  is nonsingular.

<sup>2</sup>We used the fact that positive definiteness of  $I_m - D^T D$  and  $I_p - DD^T$  are equivalent.

If  $X_1$  as defined above is nonsingular, then  $K := X_2 X_1^{-1}$  is a solution to the ARE. Thus the solvability of the ARE through suitable  $n$ -dimensional invariant subspaces of the Hamiltonian matrix gives a condition for the existence of a storage function (state function) that satisfies the dissipation inequality. Note that the above result is independent of any controllability assumption. For controllable behaviors that have an i/s/o representation given by (4.1) such that  $(A, B)$  is controllable and  $(C, A)$  is observable, it turns out that the eigenvalues of the Hamiltonian matrix are exactly equal to the roots of the determinant of the corresponding para-Hermitian matrix  $\partial\Phi(\xi)$  coming from the image representation matrix of the behavior and  $\Sigma$ . We state this result as a lemma below. The result is quite expected, and we prove it (in Appendix A) for the sake of completeness, and since we shall extend it to the case of uncontrollability in Theorem 5.4.

**LEMMA 4.5.** *Let  $\mathfrak{B} \in \mathfrak{L}_{\text{cont}}^w$  have an i/s/o representation,  $\frac{d}{dt}x = Ax + Bw_1$ ,  $w_2 = Cx + Dw_1$  with  $(A, B)$  controllable and  $(C, A)$  observable and  $\Sigma := \begin{bmatrix} I_m & 0 \\ 0 & -I_p \end{bmatrix}$ . Assume  $(I_m - D^T D) > 0$ . Define the real  $2n \times 2n$  Hamiltonian matrix*

$$H := \begin{bmatrix} A + B(I_m - D^T D)^{-1} D^T C & B(I_m - D^T D)^{-1} B^T \\ -C^T(I_p - DD^T)^{-1} C & -(A + B(I_m - D^T D)^{-1} D^T C)^T \end{bmatrix}.$$

*Suppose  $w = M(\frac{d}{dt})\ell$ ;  $\ell \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^m)$  is an observable image representation of  $\mathfrak{B}$  and consider  $\partial\Phi(\xi) := M^T(-\xi)\Sigma M(\xi)$ . Then, the Hamiltonian matrix eigenvalues are same as the zeros of  $\partial\Phi(\xi)$ , counted with multiplicities, i.e.,  $\text{spec}(H) = \text{roots}(\det \partial\Phi(\xi))$ .*

*Proof.* See Appendix A.  $\square$

The para-Hermitian matrix  $\partial\Phi(\xi)$  comes from the image representation of the behavior and  $\Sigma$ , whereas the Hamiltonian matrix is formed from the i/s/o representation with the i/o partition induced by  $\Sigma$ . Lemma 4.5 above nicely brings out a relation between these two matrices and helps in establishing a system theoretic meaning to the Hamiltonian matrix (see [15]). We shall make use of this lemma in proving our main result: Theorem 3.4.

For the special case that  $(A, B)$  is controllable, the result in Proposition 4.4 can be further extended. It has been shown in [7, 10] that if  $(A, B)$  is controllable and if  $H$  has no eigenvalues on the imaginary axis, then  $H$  gives a solution to the ARE. We state this result as a proposition below. We shall make use of this result within some proofs in the sequel to infer about the existence of a solution to an ARE coming from a strict dissipation inequality

**PROPOSITION 4.6.** *Consider the Hamiltonian matrix given by  $H := \begin{bmatrix} A & BB^T \\ -C^T C & -A^T \end{bmatrix}$ . If  $(A, B)$  is controllable and  $H$  has no roots on the imaginary axis, then there exists a real symmetric solution to the ARE:  $A^T K + KA + C^T C + KBB^T K = 0$ .*

**5. Dissipativity of uncontrollable behaviors.** In this section we prove Theorem 3.4 using the results presented in the last section and some more presented here. We first consider an example of a simple electrical circuit as shown in Figure 5.1. Under the condition  $R_1 C \neq L/R_2$ , the port variables (manifest variables)  $(v, i)$  satisfy the following differential equation:

$$\left[ \left( LC \frac{d^2}{dt^2} + (R_1 + R_2)C \frac{d}{dt} + 1 \right) - \left( R_1 LC \frac{d^2}{dt^2} + (R_1 R_2 C + L) \frac{d}{dt} + R_2 \right) \right] \begin{bmatrix} v \\ i \end{bmatrix} = 0.$$

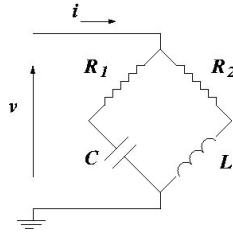


FIG. 5.1. An LCR circuit.

For the case that  $R_1 C = L/R_2$ , and  $R_1 = R_2$ , the system becomes uncontrollable. The corresponding kernel representation is

$$\left[ \left( R_2 C \frac{d}{dt} + 1 \right) - \left( L \frac{d}{dt} + R_2 \right) \right] \begin{bmatrix} v \\ i \end{bmatrix} = 0.$$

If the voltage across the capacitor  $v_C$  and current through the inductor  $i_L$  are considered as internal system variables, then we can write the following dissipation inequality:

$$\frac{d}{dt} (C v_C^2 + L i_L^2) \leqslant \begin{bmatrix} v & i \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} v \\ i \end{bmatrix}.$$

However, it turns out that the latent variables  $(v_C, i_L)$  are not observable from  $(v, i)$ , and so the storage function in the left-hand side of above inequality cannot be written in terms of a QDF in just the manifest variables (see Remark 5.1 below). We ask the question: is it possible to find a storage function in terms of the manifest variables, or do we have to have, for some cases, storage functions in terms of “hidden” variables only (variables that are unobservable from the manifest variables are also said to be hidden)? Our main result Theorem 3.4 addresses this issue under the unmixing assumption, and gives a necessary and sufficient condition for the existence of a storage function in terms of manifest variables. Thus Theorem 3.4 rules out the necessity of hidden variables to construct storage functions.

For the case of the above example, as derived in [33],  $q(v - R_1 i)^2$  with any  $q > 0$  is a storage function, i.e.,

$$\frac{d}{dt} q(v - R_1 i)^2 \leqslant v i,$$

which is a dissipation inequality in just the manifest variables. The fact that this storage function has no apparent interpretation as physical energy is discussed in Remark 5.1 below. Further,  $q > 0$  makes the set of storage functions unbounded for this case: in section 7 we shall prove the unboundedness for general uncontrollable behaviors.

*Remark 5.1.* The question of whether to allow unobservable variables into the storage function has been an issue in [33, 5]. We call a storage function *observable* if it can be expressed as a function of the manifest variables  $w$  and its derivatives. In this paper a storage function is observable by definition. In certain physical systems, like the electrical circuit above, one is able to construct a storage function from the configuration of the individual elements within the system. However, as noted in above references, the situation that the internal system variables may not be observable from



the manifest variables (for example, in the above circuit,  $(v, i)$ , through which energy is exchanged with the environment) raises the issue of whether to allow a storage function to depend on unobservable system variables also. An important contribution of this paper is that we have resolved this issue at least for the case of strict dissipativity. Under the unmixing and the maximum input cardinality conditions, we have obtained an observable storage function for strictly dissipative behaviors. However, it may turn out for some cases, like in the circuit above, that the observable storage function we obtain has no physical energy interpretation. In general for a network consisting of an interconnection of passive elements, which are either lossless elements with memory and strictly dissipative elements without memory (see [30, p. 336]), the sum of energies stored in the individual passive elements gives a natural storage function. The fact that this property is not presently captured in the definition is perhaps causing the lack of physical energy interpretation of the observable storage function.

The following example shows that we have improved one of the main results (Theorem 2) in [5] regarding observable storage functions. Consider the uncontrollable behavior,  $\mathfrak{B} \in \mathfrak{L}^2$ , given by the kernel representation  $R(\frac{d}{dt})w = 0$ , where  $R(\xi) = [2(\xi^2 + 3\xi + 2) - (2\xi^2 + 3\xi + 1)]$ . The QDF induced by  $X^T(\zeta)KX(\eta)$  with

$$X(\xi) := \begin{bmatrix} (2\xi + 6) & -(2\xi + 3) \\ 1 & -1 \end{bmatrix} \quad \text{and} \quad K := \frac{1}{4} \begin{bmatrix} 0.118 & -0.014 \\ -0.014 & 0.472 \end{bmatrix}$$

serves as an observable storage function with respect to the supply rate  $S = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ .

Before the proof of Theorem 3.4 we state and prove the following theorem, which is an important result in its own right and will also be of importance for proving Theorem 3.4. We show that for a controllable behavior, though the definition of strict dissipativity is existential in  $\epsilon$ , it is equivalent to a pair of conditions that are verifiable without  $\epsilon$ . The second condition ensures that  $\partial\Phi(\xi)$  has no zeros on the imaginary axis while the first condition, loosely speaking, rules out the existence of zeros of  $\partial\Phi(\xi)$  at infinity.

**THEOREM 5.2.** *Consider  $\mathfrak{B} \in \mathfrak{L}_{\text{cont}}^v$  that has an observable image representation  $\mathfrak{B} = \text{im } M(\frac{d}{dt})$ . Define  $\partial\Phi(\xi) = M^T(-\xi)\Sigma M(\xi)$ . Let  $\mathbf{n}$  be the McMillan degree of  $\mathfrak{B}$ . Then  $\mathfrak{B}$  is strictly dissipative with respect to  $\Sigma$  if and only if the following are satisfied:*

1.  $\deg(\det \partial\Phi(\xi)) = 2\mathbf{n}$ ,
2.  $\partial\Phi(j\omega) > 0$  for all  $\omega \in \mathbb{R}$ .

In order to prove the above theorem, we use the following lemma, whose proof is in the appendix. Notice that in both, the theorem above and the lemma below, condition 2 does *not* imply condition 1. Strictness of the dissipativity (i.e., existence of an  $\epsilon > 0$ ) plays a role in this implication.

**LEMMA 5.3.** *Let a controllable behavior  $\mathfrak{B} = \text{im } M(\frac{d}{dt})$  have i/s/o representation  $\frac{d}{dt}x = Ax + Bw_1$  and  $w_2 = Cx + Dw_1$ . Suppose  $\mathbf{n}$  is the McMillan degree of  $\mathfrak{B}$ . Define  $\partial\Phi(\xi) = M^T(-\xi)\Sigma M(\xi)$ . Then  $(I_{\mathbf{m}} - D^TD) > 0$  if the following conditions are satisfied:*

1.  $\deg(\det \partial\Phi(\xi)) = 2\mathbf{n}$ ,
2.  $\partial\Phi(j\omega) > 0$  for all  $\omega \in \mathbb{R}$ .

*Proof.* See Appendix A.  $\square$

*Proof of Theorem 5.2. (If)* Assuming both conditions 1 and 2 are true, we shall show that the behavior is strictly  $\Sigma$ -dissipative. It follows from [28, Proposition 5.2] that the second condition,  $\partial\Phi(j\omega) > 0$  for all real  $\omega$ , implies that  $\mathfrak{B}$  is guaranteed to be dissipative with respect to  $\Sigma$ . Therefore, it follows from Proposition 4.1 that  $\mathfrak{B}$

allows an i/s/o representation (4.1) and there exists  $K = K^T \in \mathbb{R}^{n \times n}$  such that the following LMI is satisfied:

$$\begin{bmatrix} (C^T C + A^T K + K A) & (K B + C^T D) \\ (B^T K + D^T C) & -(I_m - D^T D) \end{bmatrix} \leq 0.$$

Using Lemma 5.3, it follows that conditions 1 and 2 of Theorem 5.2 imply  $(I_m - D^T D) > 0$ . Therefore, the LMI has rank-minimizing solutions coming from the following ARE:

$$\begin{aligned} & \left( A + B (I_m - D^T D)^{-1} D^T C \right)^T K + K \left( A + B (I_m - D^T D)^{-1} D^T C \right) \\ & + C^T (I_p - D D^T)^{-1} C + K B (I_m - D^T D)^{-1} B^T K = 0. \end{aligned}$$

So from Proposition 4.4 there exists a Hamiltonian matrix  $H$  given below corresponding to the above ARE such that its solutions come from the  $n$ -dimensional (generalized) eigenspaces of

$$H := \begin{bmatrix} A + B (I_m - D^T D)^{-1} D^T C & B (I_m - D^T D)^{-1} B^T \\ -C^T (I_p - D D^T)^{-1} C & - \left( A + B (I_m - D^T D)^{-1} D^T C \right)^T \end{bmatrix}.$$

It follows from Lemma 4.5 that  $\text{spec}(H) = \text{roots}(\det \partial \Phi(\xi))$ . Since  $\partial \Phi(j\omega) > 0$  for all  $\omega \in \mathbb{R}$  roots  $(\det \partial \Phi(\xi)) \cap j\mathbb{R} = \emptyset$ . Therefore, due to Lemma 4.5,  $H$  does not have any purely imaginary eigenvalues. So from the continuity of eigenvalues (see [11]) there exists  $\epsilon \in \mathbb{R}$  small enough such that the following matrix

$$H_\epsilon = H - \epsilon \begin{bmatrix} 0 & 0 \\ C^T C & 0 \end{bmatrix}$$

also has no eigenvalues on the imaginary axis. Note that from Proposition A.1 in the appendix,  $(A, B)$  controllable implies so is  $[(A + B(I_m - D^T D)^{-1} D^T C), B(I_m - D^T D)^{-1/2}]$ .

Hence it follows from Proposition 4.6 that  $H_\epsilon$  gives a solution  $K$  to the corresponding ARE

$$\begin{aligned} & \left( A + B (I_m - D^T D)^{-1} D^T C \right)^T K + K \left( A + B (I_m - D^T D)^{-1} D^T C \right) \\ & + C^T (I_p - D D^T)^{-1} C + \epsilon C^T C + K B (I_m - D^T D)^{-1} B^T K = 0, \end{aligned}$$

which implies, from Proposition 4.1, that  $\mathfrak{B}$  is dissipative with respect to  $\begin{bmatrix} I_m & 0 \\ 0 & -(1 + \epsilon)I_p \end{bmatrix}$  for some  $\epsilon > 0$ .

Utilizing the Lemma A.2 in the appendix, we conclude that  $\mathfrak{B}$  being dissipative with respect to  $\begin{bmatrix} I_m & 0 \\ 0 & -(1 + \epsilon)I_p \end{bmatrix}$  implies  $\mathfrak{B}$  is strictly  $\Sigma$ -dissipative, and this completes the “if” part of Theorem 5.2.

**(Only if)** First we show that  $\mathfrak{B}$  being strictly  $\Sigma$ -dissipative implies condition 2 holds. Then we shall further show that assuming condition 2 holds,  $\mathfrak{B}$  being strictly  $\Sigma$ -dissipative implies that condition 1 holds. Assume that  $\mathfrak{B}$  is strictly  $\Sigma$ -dissipative. Definition 3.2 implies there exists  $\epsilon > 0$  such that  $\mathfrak{B}$  is dissipative with respect to  $\Sigma - \epsilon I_m$ . It then follows from Proposition 3.3 and [28, Proposition 5.2] that

$$(5.1) \quad \partial \Phi(j\omega) \geq \epsilon M^T(-j\omega) M(j\omega).$$

Note that  $M^T(-j\omega)M(j\omega) \geq 0$  for all real  $\omega$ . Since  $M(\xi)$  is right-prime, it follows from Lemma A.5 that  $M^T(-j\omega)M(j\omega) > 0$  for all  $\omega \in \mathbb{R}$ . This implies, from inequality (5.1) that  $\partial\Phi(j\omega) > 0$  for all  $\omega \in \mathbb{R}$ . This shows that if  $\mathfrak{B}$  is strictly  $\Sigma$ -dissipative, then condition 2 holds.

Now we show that  $\mathfrak{B}$  being strictly  $\Sigma$ -dissipative together with  $\partial\Phi(j\omega) > 0$  for all  $\omega \in \mathbb{R}$  implies that  $\deg(\det \partial\Phi(\xi)) = 2n$ . Consider a spectral factorization of  $\partial\Phi(\xi) = N^T(-\xi)N(\xi)$  (existence of  $N \in \mathbb{R}^{n \times n}[\xi]$  is guaranteed due to the inequality  $\partial\Phi(j\omega) > 0$ ; see [18]). Again, since  $M(\xi)$  is right-prime,  $M^T(-j\omega)M(j\omega) > 0$  for all  $\omega \in \mathbb{R}$  (Lemma A.5). So  $M^T(-\xi)M(\xi)$  also allows a spectral factorization

$$M^T(-\xi)M(\xi) = D^T(-\xi)D(\xi); \quad D \in \mathbb{R}^{n \times n}(\xi).$$

By Lemma A.5 in the appendix,  $M(\xi)$  being a right-prime polynomial matrix implies that  $\deg(\det(M^T(-\xi)M(\xi))) = 2n$ . Therefore  $\deg(\det D(\xi)) = n$ . Since  $\mathfrak{B}$  is strictly  $\Sigma$ -dissipative, inequality (5.1) holds, which we can now rewrite in terms of the spectral factors of  $\partial\Phi(\xi)$  and  $M^T(-\xi)M(\xi)$  as

$$\begin{aligned} N^T(-j\omega)N(j\omega) &\geq \epsilon D^T(-j\omega)D(j\omega) \quad \forall \omega \in \mathbb{R} \\ (5.2) \quad &\Rightarrow [N(-j\omega)D^{-1}(-j\omega)]^T [N(j\omega)D^{-1}(j\omega)] \geq \epsilon I_m \quad \forall \omega \in \mathbb{R}. \end{aligned}$$

The above inequality (5.2) implies that there exists a real  $\epsilon > 0$  such that the minimum singular value of the rational function matrix  $N(j\omega)D^{-1}(j\omega)$  is at least  $\epsilon$  for all real  $\omega$ . This further implies that  $\det(N(\xi)D^{-1}(\xi))$  is a biproper rational function, which means  $\deg(\det N(\xi)) = n$  and hence  $\deg(\det \partial\Phi(\xi)) = 2n$ .  $\square$

The following result will be important in order to prove the main result: Theorem 3.4. It is an extension of Lemma 4.5, where we saw that for a controllable behavior  $\mathfrak{B}$  the set of eigenvalues of the Hamiltonian matrix is equal to the set of roots of the determinant of the para-Hermitian matrix  $\partial\Phi(\xi)$ . For the case of uncontrollability we show that every uncontrollable pole  $\lambda$  of  $\mathfrak{B}$ , together with  $-\lambda$ , is an eigenvalue of  $H$ , in addition to those coming from the controllable part of  $\mathfrak{B}$  like in Lemma 4.5.

**THEOREM 5.4.** *Let  $\mathfrak{B} \in \mathfrak{L}^w$  and let  $\chi_{\text{un}}$  be its uncontrollable characteristic polynomial. Assume  $\mathfrak{B}$  has an observable i/s/o representation  $\frac{d}{dt}x = Ax + Bw_1$ ,  $w_2 = Cx + Dw_1$ , and suppose  $I_m - D^T D$  is invertible. Let  $\mathfrak{B}_{\text{cont}} = \text{im } M(\frac{d}{dt})$ . Define  $\partial\Phi(\xi) = M^T(-\xi)\Sigma M(\xi)$ . Construct the Hamiltonian matrix*

$$H := \begin{bmatrix} A + B(I_m - D^T D)^{-1} D^T C & B(I_m - D^T D)^{-1} B^T \\ -C^T(I_p - DD^T)^{-1} C & -\left(A + B(I_m - D^T D)^{-1} D^T C\right)^T \end{bmatrix},$$

*Then, the Hamiltonian matrix eigenvalues, the zeros of  $\partial\Phi(\xi)$ , and the uncontrollable poles of  $\mathfrak{B}$  are related by:  $\text{spec}(H) = \text{roots}[\det \partial\Phi(\xi) \chi_{\text{un}}(\xi) \chi_{\text{un}}(-\xi)]$ , counted with multiplicities.*

*Proof.* See Appendix A.  $\square$

The next result is one of the main results of this paper and it is pivotal for proving Theorem 3.4. It brings out an important property about certain  $n$ -dimensional invariant subspaces of the Hamiltonian matrix. We already know from Proposition 4.4 that statement 2 in the theorem below is equivalent to existence of a solution to the ARE. In this sense, Theorem 5.5 below is an important extension to Proposition 4.6. The theorem shows that a given Lambda set results in a solution to the ARE if and only if this Lambda set contains the uncontrollable poles of the system. The proof

comes following a line of argument similar to the one used in proving [10, Theorem 7.2].

**THEOREM 5.5.** *Consider the Hamiltonian matrix,  $H = \begin{bmatrix} A & BB^T \\ -C^T C & -A^T \end{bmatrix}$ . Define  $\Lambda_{\text{un}}^{(A,B)}$  to be the set of uncontrollable eigenvalues of  $(A, B)$  pair and let  $\Lambda$  be a Lambda-set of  $\text{spec}(H)$ . Then the following are equivalent.*

1.  $\Lambda \supseteq \Lambda_{\text{un}}^{(A,B)}$ .
2. The  $n$ -dimensional invariant subspace of  $H$  corresponding to  $\Lambda$  is complementary to  $\text{im} \begin{bmatrix} 0_n \\ I_n \end{bmatrix}$ .

*Proof. (1  $\Rightarrow$  2)* Denote by  $\mathcal{X}_\Lambda(H)$  the invariant subspace of  $H$  corresponding to a Lambda-set  $\Lambda$  of  $\text{spec}(H)$ . Since  $\Lambda$  is a Lambda-set of  $\text{spec}(H)$   $\Lambda \cap (-\Lambda) = \emptyset$  and  $\Lambda \cup (-\Lambda) = \text{spec}(H)$ , and therefore  $\dim(\mathcal{X}_\Lambda(H)) = n$ . Let  $\mathcal{X}_\Lambda(H)$  be given by

$$(5.3) \quad \mathcal{X}_\Lambda(H) = \text{im} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

where  $X_1, X_2 \in \mathbb{R}^{n \times n}$ . Since  $\mathcal{X}_\Lambda(H)$  is an invariant subspace corresponding to  $\Lambda$ , we have the following equality

$$(5.4) \quad H \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} H_\Lambda,$$

where  $H_\Lambda \in \mathbb{R}^{n \times n}$  is such that  $\text{spec}(H_\Lambda) = \Lambda$ .

In order to prove that  $\mathcal{X}_\Lambda(H)$  is complementary to  $\text{im} \begin{bmatrix} 0_n \\ I_n \end{bmatrix}$ , we have to show that  $\ker X_1 = \{0\}$ . Assume to the contrary that  $\ker X_1$  is nontrivial; we shall show that this will lead to a contradiction to  $\Lambda \supseteq \Lambda_{\text{un}}^{(A,B)}$ .

We may assume without loss of generality that  $\mathcal{X}_\Lambda(H)$  is a generalized (right) eigenspace of  $H$  with respect to  $\Lambda$ . Using Lemma A.4 from the appendix we get,  $\text{im} \begin{bmatrix} X_2 \\ -X_1 \end{bmatrix}$  is a generalized left-eigenspace of  $H$  corresponding to  $-\Lambda$ . Since  $\Lambda$  is a Lambda-set,  $\Lambda \cap (-\Lambda) = \emptyset$ , so the two generalized eigenspaces are orthogonal to each other, i.e.,

$$\begin{bmatrix} X_2^T & -X_1^T \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = 0.$$

Because  $\mathcal{X}_\Lambda(H)$  is  $H$ -invariant, the last equation leads to

$$(5.5) \quad \begin{aligned} & \begin{bmatrix} X_2^T & -X_1^T \end{bmatrix} \begin{bmatrix} A & BB^T \\ -C^T C & -A^T \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = 0 \\ \Rightarrow & X_2^T A X_1 + X_1^T A^T X_2 + X_1^T C^T C X_1 + X_2^T B B^T X_2 = 0. \end{aligned}$$

Let  $x \in \ker X_1$ . Pre- and postmultiplying (5.5) by  $x^T$  and  $x$ , respectively, we get  $x^T X_2^T B B^T X_2 x = 0$ , which implies that  $B^T X_2 x = 0$ . Consider (5.4), which gives  $A X_1 + B B^T X_2 = X_1 H_\Lambda$ . After postmultiplying by  $x$  we get

$$\begin{aligned} A X_1 x + B B^T X_2 x &= X_1 H_\Lambda x \\ \Rightarrow X_1 H_\Lambda x &= 0 \quad \Rightarrow H_\Lambda x \in \ker X_1. \end{aligned}$$

This implies  $\ker X_1$  is  $H_\Lambda$ -invariant. Therefore, there exists  $v \neq 0$  an eigenvector of  $H_\Lambda$  such that  $X_1 v = 0$ . Let the eigenvalue corresponding to  $v$  be  $\lambda$ . Since  $\text{spec}(H_\Lambda) = \Lambda$ ,  $\lambda \in \Lambda$ . Now, from (5.4) we can write

$$(5.6) \quad -C^T C X_1 - A^T X_2 = X_2 H_\Lambda.$$

Postmultiplying (5.6) by  $v$  we get

$$\begin{aligned} -C^T C X_1 v - A^T X_2 v &= X_2 H_\Lambda v \\ \Rightarrow -A^T X_2 v &= \lambda X_2 v. \end{aligned}$$

But this means  $X_2 v$  is a left eigenvector of  $A$  with eigenvalue  $-\lambda$ . Moreover  $-\lambda \notin \Lambda$  (because  $\Lambda \cap (-\Lambda) = \emptyset$ ) and  $B^T X_2 v = 0$ . This together means that  $-\lambda$  is an uncontrollable eigenvalue of  $A$ , which is a contradiction to  $\Lambda \supseteq \Lambda_{\text{un}}^{(A,B)}$ . Thus  $\ker X_1$  cannot be nontrivial, and therefore  $\mathcal{X}_\Lambda$  is complementary to  $\text{im} \begin{bmatrix} 0_n \\ I_n \end{bmatrix}$ .

**(2  $\Rightarrow$  1)** We assume that  $\mathcal{X}_\Lambda(H)$ , the generalized eigenspace of  $H$  corresponding to  $\Lambda$  (a Lambda-set of  $\text{spec}(H)$ ), is complementary to  $\text{im} \begin{bmatrix} 0_n \\ I_n \end{bmatrix}$ , and we show that  $\Lambda \supseteq \Lambda_{\text{un}}^{(A,B)}$ . Suppose  $\Lambda \not\supseteq \Lambda_{\text{un}}^{(A,B)}$ . Then there exists  $\lambda \in \Lambda_{\text{un}}^{(A,B)}$  but  $\lambda \notin \Lambda$ . Since  $\Lambda$  is a Lambda-set of  $\text{spec}(H)$ ,  $\Lambda \cup (-\Lambda) = \text{spec}(H)$ . Also  $\Lambda_{\text{un}}^{(A,B)} \subset \text{spec}(H)$  (see Theorem 5.4). These two facts together imply that  $-\lambda \in \Lambda$ . Now  $\lambda \in \Lambda_{\text{un}}^{(A,B)}$  implies that there exists a nonzero vector  $v \in \mathbb{C}^n$  such that  $v^T A = \lambda v^T$  and  $v^T B = 0$ . Therefore the  $2n$  vector constructed as  $w = \text{col}(0_n, v)$  satisfies  $Hw = -\lambda w$ ; i.e.,  $w$  is an eigenvector of  $H$  with eigenvalue  $-\lambda$ . Since  $-\lambda \in \Lambda$ , the last  $w$  being an eigenvector of  $H$  with eigenvalue  $-\lambda$  means  $\mathcal{X}_\Lambda$  is *not* complementary to  $\text{im} \begin{bmatrix} 0_n \\ I_n \end{bmatrix}$  because  $w$  has the upper  $n$  entries zero. Thus  $\Lambda \not\supseteq \Lambda_{\text{un}}^{(A,B)}$  leads to a contradiction to statement 2. This proves “2  $\Rightarrow$  1.”  $\square$

With the above results we are now in a position to prove our main result, Theorem 3.4. The “if” part requires construction of a storage function for the uncontrollable behavior  $\mathfrak{B}$  to show strict dissipativity. We use the Hamiltonian matrix properties proved above, combined with some perturbation arguments, to show the Riccati equality and inequality have solutions, and then construct a storage function for proving the strict dissipativity. The unmixing property plays a key role in this construction.

*Proof of Theorem 3.4. (Only if)* We assume that  $\mathfrak{B} \in \mathfrak{L}^w$  is strictly  $\Sigma$ -dissipative, and we show that this implies  $\mathfrak{B}_{\text{cont}}$  is strictly  $\Sigma$ -dissipative. Since  $\mathfrak{B}$  is strictly  $\Sigma$ -dissipative, according to Definition 3.2 there exists  $\Psi(\zeta, \eta) \in \mathbb{R}_s^{w \times w}[\zeta, \eta]$  and a real number  $\epsilon > 0$  such that

$$\frac{d}{dt} Q_\Psi(w) \leq Q_\Sigma(w) - \epsilon |w|^2 \quad \text{for all } w \in \mathfrak{B}.$$

Integrating both sides of the above inequality, considering only those trajectories in  $\mathfrak{B} \cap \mathfrak{D}$  we get

$$\int_{-\infty}^{\infty} Q_\Sigma(w) dt \geq \epsilon \int_{-\infty}^{\infty} |w|^2 dt \quad \text{for all } w \in \mathfrak{B} \cap \mathfrak{D}.$$

Since  $\mathfrak{B} \cap \mathfrak{D} = \mathfrak{B}_{\text{cont}} \cap \mathfrak{D}$  (see [16]), the above inequality implies that  $\mathfrak{B}_{\text{cont}}$  is strictly  $\Sigma$ -dissipative.

**(If)** We assume that the controllable part  $\mathfrak{B}_{\text{cont}}$  is strictly  $\Sigma$ -dissipative and that the unmixing condition on the uncontrollable poles holds. We show that  $\mathfrak{B}$  too is strictly  $\Sigma$ -dissipative. A necessary condition for  $\mathfrak{B}_{\text{cont}}$  to be strictly  $\Sigma$ -dissipative is that the transfer function for  $\mathfrak{B}_{\text{cont}}$  with the i/o partition induced by  $\Sigma$  is proper. Therefore, from Proposition 2.2 above,  $\mathfrak{B}$  has a state map  $X(\xi)$  and an i/s/o representation with  $x = X(\frac{d}{dt})w$  such that  $\frac{d}{dt}x = \tilde{A}x + \tilde{B}w_1$ ,  $w_2 = \tilde{C}x + \tilde{D}w_1$  with  $(\tilde{C}, \tilde{A})$  pair observable and  $(\tilde{A}, \tilde{B})$  pair uncontrollable with  $\Lambda_{\text{un}}^{(\tilde{A}, \tilde{B})} = \Lambda_{\text{un}}(\mathfrak{B})$ , counting mul-

tiplicities. Let  $\mathfrak{B}_{\text{cont}}$  have an image representation

$$\mathfrak{B}_{\text{cont}} = \text{im} \begin{bmatrix} W_1 \left( \frac{d}{dt} \right) \\ W_2 \left( \frac{d}{dt} \right) \end{bmatrix}; \quad W_1 \in \mathbb{R}^{m \times m}[\xi], W_2 \in \mathbb{R}^{p \times m}[\xi].$$

Since  $\mathfrak{B}_{\text{cont}}$  is strictly  $\Sigma$ -dissipative, from Theorem 5.2,

$$\begin{aligned} \deg \left( \det \left[ W_1^T(-\xi)W_1(\xi) - W_2^T(-\xi)W_2(\xi) \right] \right) &= 2 \times n(\mathfrak{B}_{\text{cont}}) \text{ and} \\ W_1^T(-j\omega)W_1(j\omega) - W_2^T(-j\omega)W_2(j\omega) &> 0 \quad \forall \omega \in \mathbb{R}. \end{aligned}$$

Again, from Lemma 5.3, these two facts together imply that  $I_m - \tilde{D}^T \tilde{D} > 0$ . This implies that there exists a Hamiltonian matrix given by

$$H = \begin{bmatrix} \tilde{A} + \tilde{B} \left( I_m - \tilde{D}^T \tilde{D} \right)^{-1} \tilde{D}^T \tilde{C} & \tilde{B} \left( I_m - \tilde{D}^T \tilde{D} \right)^{-1} \tilde{B}^T \\ -\tilde{C}^T \left( I_p - \tilde{D} \tilde{D}^T \right)^{-1} \tilde{C} & - \left( \tilde{A} + \tilde{B} \left( I_m - \tilde{D}^T \tilde{D} \right)^{-1} \tilde{D}^T \tilde{C} \right)^T \end{bmatrix}.$$

Define  $A := \tilde{A} + \tilde{B}(I_m - \tilde{D}^T \tilde{D})^{-1} \tilde{D}^T \tilde{C}$ ,  $B := \tilde{B}(I_m - \tilde{D}^T \tilde{D})^{-\frac{1}{2}}$ , and  $C := (I_p - \tilde{D} \tilde{D}^T)^{-\frac{1}{2}} \tilde{C}$ . Using Theorem 5.4 we get  $\text{spec}(H) = \text{roots} [\det \partial \Phi(\xi) \chi_{\text{un}}(\xi) \chi_{\text{un}}(-\xi)]$ , where  $\partial \Phi(\xi) := W_1^T(-\xi)W_1(\xi) - W_2^T(-\xi)W_2(\xi)$ . Since  $\mathfrak{B}_{\text{cont}}$  is strictly  $\Sigma$ -dissipative, Theorem 5.2 implies that  $\text{roots}(\det \partial \Phi(\xi)) \cap j\mathbb{R} = \emptyset$ . By assumption  $\chi_{\text{un}}(\xi)$  and  $\chi_{\text{un}}(-\xi)$  are coprime, i.e.,  $\Lambda_{\text{un}} \cap (-\Lambda_{\text{un}}) = \emptyset$ . Thus  $H$  has no eigenvalues on the imaginary axis and so  $\text{spec}(H)$  allows a Lambda-set  $\Lambda$ . Moreover, we can construct  $\Lambda$  in such a way that  $\Lambda_{\text{un}} \subseteq \Lambda$ . From Proposition 2.2 it follows that the set of uncontrollable eigenvalues of  $(A, B)$  pair is exactly equal to  $\Lambda_{\text{un}}$ . Hence from Theorem 5.5 a generalized eigenspace of  $H$  corresponding to  $\Lambda$ ,  $\mathcal{X}_\Lambda := \text{im} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ , with  $X_1, X_2 \in \mathbb{R}^{n \times n}$ , is complementary to  $\text{im} \begin{bmatrix} 0_n \\ I_n \end{bmatrix}$ , which implies  $X_1$  is nonsingular allowing us to define  $K = X_2 X_1^{-1} \in \mathbb{R}^{n \times n}$ . Again since  $\Lambda \cap (-\Lambda) = \emptyset$ , applying Lemma A.4 we get

$$\begin{bmatrix} X_2^T & -X_1^T \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = 0, \text{ which implies that } (X_2 X_1^{-1})^T = X_2 X_1^{-1}.$$

Thus  $K = K^T$  is a symmetric solution to the ARE:  $A^T K + K A + C^T C + K B B^T K = 0$ .

In order to complete the proof we have to show that there exists a storage function for the strict dissipation inequality of Definition 3.2. For this we make use of Lemma A.2 of Appendix A to infer that the strict dissipation inequality is equivalent to the following LMI

$$(5.7) \quad \begin{bmatrix} - \left( \tilde{C}^T \tilde{C} + \tilde{A}^T K + K \tilde{A} \right) & - \left( K \tilde{B} + \tilde{C}^T \tilde{D} \right) \\ - \left( \tilde{B}^T K + \tilde{D}^T \tilde{C} \right) & s I_m - \tilde{D}^T \tilde{D} \end{bmatrix} - \epsilon \begin{bmatrix} \tilde{C}^T \tilde{C} & 0 \\ 0 & 0 \end{bmatrix} \geq 0,$$

for some  $\epsilon > 0$ . The corresponding Hamiltonian matrix turns out to be

$$\begin{aligned} H_\epsilon &:= \begin{bmatrix} \tilde{A} + \tilde{B}(I_m - \tilde{D}^T \tilde{D})^{-1} \tilde{D}^T \tilde{C} & \tilde{B}(I_m - \tilde{D}^T \tilde{D})^{-1} \tilde{B}^T \\ -\tilde{C}^T (I_p - \tilde{D} \tilde{D}^T)^{-1} \tilde{C} & -(\tilde{A} + \tilde{B}(I_m - \tilde{D}^T \tilde{D})^{-1} \tilde{D}^T \tilde{C})^T \end{bmatrix} - \epsilon \begin{bmatrix} 0 & 0 \\ \tilde{C}^T \tilde{C} & 0 \end{bmatrix} \\ &= H - \epsilon \begin{bmatrix} 0 & 0 \\ \tilde{C}^T \tilde{C} & 0 \end{bmatrix}. \end{aligned}$$

Observe that the kind of perturbation that takes  $H$  to  $H_\epsilon$  is such that  $\text{spec}(H_\epsilon) \supset \Lambda_{\text{un}}$ . Further, since the perturbation is analytic, according to [11], there exists an  $\epsilon_1 > 0$  small enough such that the following property of  $H$  holds for  $H_{\epsilon_1}$  also:

$$\text{spec}(H_{\epsilon_1}) \cap j\mathbb{R} = \phi.$$

Therefore,  $\text{spec}(H_{\epsilon_1})$  also allows a Lambda-set  $\Lambda_{\epsilon_1}$  such that  $\Lambda_{\epsilon_1} \supseteq \Lambda_{\text{un}}$ . It follows from Theorem 5.5 that there exists  $K_{\epsilon_1} = K_{\epsilon_1}^T \in \mathbb{R}^{n \times n}$ , which is a rank-minimizing solution to the LMI (5.7) with  $\epsilon = \epsilon_1$ .

The observable i/s/o representation is obtained from the manifest variables through a state map as  $x = X(\frac{d}{dt})w$ . Thus,  $X(\xi)$  and  $K_{\epsilon_1}$  give  $\Psi(\zeta, \eta) := X^T(\zeta)K_{\epsilon_1}X(\eta)$  that satisfies

$$\frac{d}{dt}Q_\Psi(w) = \frac{d}{dt} \left[ \left( X \left( \frac{d}{dt} \right) w \right)^T K_{\epsilon_1} X \left( \frac{d}{dt} \right) w \right] \leq Q_\Sigma(w) - \epsilon_1 |w|^2 \quad \forall w \in \mathfrak{B},$$

which from Definition 3.2 means that  $\mathfrak{B}$  is strictly  $\Sigma$ -dissipative.  $\square$

It is evident from the above proof that the storage function we construct to show strict dissipativity is, in fact, a memoryless quadratic function of the state of the system. More concretely, under the assumption that the uncontrollable poles are unmixed, such a storage function, which is a memoryless state function, exists if the behavior is strictly dissipative. We state this important consequence as a corollary below.

**COROLLARY 5.6.** *Suppose  $\mathfrak{B} \in \mathfrak{L}^w$  has uncontrollable poles satisfying  $\Lambda_{\text{un}} \cap (-\Lambda_{\text{un}}) = \phi$ , and let  $X \in \mathbb{R}^{n \times w}[\xi]$  give a minimal state map for  $\mathfrak{B}$ . Consider a nonsingular  $\Sigma \in \mathbb{R}_s^{w \times w}$  with  $\mathfrak{m}(\mathfrak{B}) = \sigma_+(\Sigma)$ . Then, the following are equivalent.*

1.  $\mathfrak{B}$  is strictly  $\Sigma$ -dissipative.
2. There exists a  $K \in \mathbb{R}^{n \times n}$  and  $\epsilon > 0$  such that  $\frac{d}{dt}[(X(\frac{d}{dt})w)^T K X(\frac{d}{dt})w] \leq w^T(\Sigma - \epsilon I)w$  for all  $w \in \mathfrak{B}$ .

The above important and intuitive result that storage of energy requires memory elements, namely states, was proved formally for the controllable case in [27]. Note that for the special case that the behavior is controllable we have provided a new and alternative proof of the main result of [27], by showing that for a strictly dissipative behavior there exists a storage function which is a state function.

**Remark 5.7.** In this paper we have worked with strict dissipativity as defined in Definition 3.2. One of the main reasons to invoke strictness is that it guarantees nonsingularity of  $I_m - D^T D$  and hence the existence of the Hamiltonian matrix and the Riccati equation. It also rules out the possibility of the Hamiltonian matrix having purely imaginary eigenvalues, and thus enables us to use Lambda-set arguments. It remains to explore which of the above results are true for the case of nonstrict dissipativity.

**Remark 5.8.** The question of solvability of the positive-real LMI without imposing system theoretic assumptions like controllability or observability has been dealt with in [9]. However, a very restrictive assumption made there is that the whole set of eigenvalues of the system matrix  $A$  satisfies the unmixing property, i.e.,  $\text{spec}(A) \cap \text{spec}(-A) = \phi$ . According to our main result (Theorem 3.4) this assumption is not necessary. It is sufficient that only the uncontrollable poles satisfy the unmixing property. We shall see later in section 8 the extent of necessity of this unmixing property. The following example shows how the positive-real LMI is solvable when some elements of  $\text{spec}(A)$  have symmetry with respect to the imaginary axis and the system is uncontrollable.

*Example 5.9.* Consider an i/s/o system with the following  $A, B, C, D$  matrices:

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad D = 1.$$

Observe that  $\text{spec}(A) = \{1, -1\}$ , which is symmetric with respect to the imaginary axis. Here  $\Lambda_{\text{un}} = \{-1\}$ , and the other eigenvalue ( $= 1$ ) is controllable. An equivalent kernel representation of the manifest behavior is given by

$$\left[ \left( \frac{d^2}{dt^2} - \frac{d}{dt} - 2 \right) \quad - \left( \frac{d^2}{dt^2} - 1 \right) \right] \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0.$$

We ask the question: is this i/s/o system  $S := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$  dissipative or equivalently, is there a real symmetric solution  $K = K^T \in \mathbb{R}^{2 \times 2}$  for the following LMI

$$\begin{bmatrix} -A^T K - K A & C^T - K B \\ C - B^T K & D + D^T \end{bmatrix} \geq 0?$$

Obviously,  $\Lambda_{\text{un}} = \{-1\}$  satisfies the unmixing property, and one can check that the controllable part  $\mathfrak{B}_{\text{cont}} = \ker \begin{bmatrix} \frac{d}{dt} - 2 & \frac{d}{dt} + 1 \end{bmatrix}$  is strictly  $S$ -dissipative, which from Theorem 3.4 implies that  $\mathfrak{B}$  is strictly  $S$ -dissipative. This can be verified by checking that the following real symmetric matrix induces a storage function that satisfies the dissipation inequality

$$K = \begin{bmatrix} -0.957 & -1.457 \\ -1.457 & -1.957 \end{bmatrix},$$

and therefore solves the LMI.

**6. Positive storage functions and stabilizability.** In this section we establish an important link between stabilizability of systems and positive definiteness of storage functions of strictly dissipative systems. The importance of this link lies in the fact that the energy stored in physical systems is a nonnegative quantity and dissipative physical systems satisfy an additional property that, if the system was initially discharged, then the net energy supplied into the system *upto any time instant* is nonnegative; this is called *half-line dissipativity*. We review these concepts (from [28]) below and prove similar results for uncontrollable systems in this section.

For this paper, we need half-line dissipativity for only the negative half of the real line:  $\mathbb{R}_-$ . A controllable behavior  $\mathfrak{B} \in \mathfrak{L}_{\text{cont}}^w$  is said to be  $\Sigma$ -dissipative on  $\mathbb{R}_-$  if  $\int_{-\infty}^0 Q_{\Sigma}(w)dt \geq 0$  for all  $w \in \mathfrak{B} \cap \mathfrak{D}$ . (Due to time invariance of  $\mathfrak{B}$ , it is enough to integrate up to 0.) Half-line dissipativity is related to (semi-)definiteness of the storage function. A storage function  $Q_{\Psi}$  is called nonnegative if  $Q_{\Psi}(w)(t) \geq 0$  for all  $t \in \mathbb{R}$  and  $w \in \mathfrak{B}$ . For controllable behaviors, it was shown in [28] that existence of a nonnegative storage function is equivalent to dissipativity of  $\mathfrak{B}$  on  $\mathbb{R}_-$ . The importance of nonnegative storage functions is due to such functions being bounded from below (namely, by zero), because of which we expect that when the supply of energy is stopped, then the trajectories cannot become unbounded. This link to stability was made precise and proved in [29, Proposition 1, Part I].

We saw in Corollary 5.6 that, for a dissipative behavior  $\mathfrak{B}$  with a minimal state map  $X \in \mathbb{R}^{n \times w}[\xi]$ , a storage function  $Q_{\Psi}$  is associated to a symmetric matrix  $K \in \mathbb{R}^{n \times n}$  such that  $Q_{\Psi}(w) = (X(\frac{d}{dt})w)^T K X(\frac{d}{dt})w$ . Hence  $Q_{\Psi}$  is nonnegative if and only



if  $K \geq 0$  (see [28]). In the context of strict dissipativity, we define a *positive definite* storage function. A storage function  $Q_\Psi$  is called positive definite if  $K > 0$ .

The following result is one of the main results of this paper. It relates existence of positive definite storage functions to stability of the autonomous part of the uncontrollable dissipative behavior. A behavior with a stable autonomous part is nothing but a *stabilizable* behavior. A behavior  $\mathfrak{B} \in \mathfrak{L}^w$  is called stabilizable if for every  $w \in \mathfrak{B}$ , there exists a  $w' \in \mathfrak{B}$  such that  $w(t) = w'(t)$  for  $t \leq 0$  and  $w'(t) \rightarrow 0$  as  $t \rightarrow \infty$ . A behavior is stabilizable if and only if  $\Lambda_{\text{un}} \subset \mathbb{C}^-$  (see [32]).

**THEOREM 6.1.** *Let a linear differential behavior  $\mathfrak{B} \in \mathfrak{L}^w$  be strictly  $\Sigma$ -dissipative with  $\mathfrak{m}(\mathfrak{B}) = \sigma_+(\Sigma)$ . Then there exists a positive definite storage function if and only if the following are satisfied:*

1. *there exists  $\epsilon > 0$  such that  $\int_{\mathbb{R}_-} Q_\Sigma(w) dt \geq \int_{\mathbb{R}_-} \epsilon |w|^2 dt$  for all  $w \in \mathfrak{B} \cap \mathfrak{D}$  and*
2.  *$\Lambda_{\text{un}} \subset \mathbb{C}^-$ .*

The first condition is clearly a necessary condition for existence of a positive definite storage function; namely, the controllable part has to be strictly dissipative on  $\mathbb{R}_-$ . The second condition is also necessary because of the notion that the storage function behaves like a Lyapunov function for an autonomous system, and as is well known, a positive Lyapunov function exists if and only if the autonomous system is asymptotically stable. The fact that these two conditions are together *sufficient* for the existence of a positive definite storage function for the whole behavior is one of the main contributions of this paper. Also notice that  $\Lambda_{\text{un}} \subset \mathbb{C}^-$  is a very special case of the unmixing condition. Thus the uncontrollability of the stabilizable behavior poses no hindrance to existence of a storage function for strict dissipativity as long as the controllable/autonomous parts allow storage/Lyapunov functions individually. As noted above, this is the principal finding of this paper.

*Proof. (If)* In this part of the proof we show that the existence of  $\epsilon > 0$  such that  $\int_{\mathbb{R}_-} Q_\Sigma(w) dt \geq \int_{\mathbb{R}_-} \epsilon |w|^2 dt$  for all  $w \in \mathfrak{B} \cap \mathfrak{D}$  and  $\Lambda_{\text{un}} \subset \mathbb{C}^-$  together imply that  $\mathfrak{B}$  has a positive definite storage function. First, let the controllable part be given by an observable image representation

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} W_1(\frac{d}{dt}) \\ W_2(\frac{d}{dt}) \end{bmatrix} \ell; \quad \ell \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^m),$$

where  $W_1 \in \mathbb{R}^{m \times m}[\xi]$  and  $W_2 \in \mathbb{R}^{p \times m}[\xi]$ . Since  $\mathfrak{B} \cap \mathfrak{D} = \mathfrak{B}_{\text{cont}} \cap \mathfrak{D}$ ,  $\int_{\mathbb{R}_-} Q_\Sigma(w) dt \geq 0$  for all  $w \in \mathfrak{B} \cap \mathfrak{D}$  implies that the QDF induced by the two-variable polynomial matrix  $\Phi(\zeta, \eta) := [W_1^T(\zeta) \ W_2^T(\zeta)] \Sigma [W_1^T(\eta) \ W_2^T(\eta)]^T$  is strictly *half-line* positive, that is  $\int_{\mathbb{R}_-} Q_\Phi(\ell) dt > 0$  for all nonzero  $\ell \in \mathfrak{D}(\mathbb{R}, \mathbb{R}^m)$ . This, according to Theorem 6.4 of [28], implies that  $W_1(\xi)$  is Hurwitz. Since  $\mathfrak{B}$  is strictly  $\Sigma$ -dissipative the transfer function from  $w_1$  to  $w_2$  is proper, and so from Proposition A.3,  $\mathfrak{B}$  has a state observable i/s/o representation in Kalman decomposed form as

$$A = \begin{bmatrix} A_c & A_{cp} \\ 0 & A_u \end{bmatrix}, B = \begin{bmatrix} B_c \\ 0 \end{bmatrix}, C = [C_c \ C_u],$$

where  $\frac{d}{dt}x_c = A_c x_c + B_c w_1, w_2 = C_c x + D w_1$  gives a state controllable and state observable i/s/o representation of the controllable part  $\mathfrak{B}_{\text{cont}}$ . This implies that roots  $(\det W_1(\xi)) = \text{spec}(A_c)$ , and, therefore,  $W_1(\xi)$  being Hurwitz implies that so is  $A_c$ . Once again, from Proposition A.3, the set of uncontrollable poles  $\Lambda_{\text{un}} = \text{spec}(A_u)$ . So  $\Lambda_{\text{un}} \subset \mathbb{C}^-$  implies  $A_u$  is also Hurwitz. Hence  $A$  is Hurwitz. Because  $\mathfrak{B}$  is strictly

$\Sigma$ -dissipative, the following ARI has a solution

$$(6.1) \quad A^T K + K A + C^T C + (D^T C + B^T K)^T (I_m - D^T D)^{-1} (D^T C + B^T K) \leq 0.$$

Since  $(D^T C + B^T K)^T (I_m - D^T D)^{-1} (D^T C + B^T K) \geq 0$ ,  $(C, A)$  being observable implies that  $[C^T C + (D^T C + B^T K)^T (I_m - D^T D)^{-1} (D^T C + B^T K), A]$  too is observable (see [34]). Hence the above inequality (6.1), when treated as a Lyapunov equation  $A^T K + K A + Q \leq 0$ , where  $Q := [C^T C + (D^T C + B^T K)^T (I_m - D^T D)^{-1} (D^T C + B^T K)]$ , has  $(Q, A)$  observable and  $A$  Hurwitz, which means that all the solutions  $K$  are positive definite (see [34]). Thus a positive definite solution to (6.1) induces a positive definite storage function and this completes the proof of the “if” part.

**(Only if)** First we shall show that  $\mathfrak{B}$  being strictly dissipative with respect to  $\Sigma$  with a positive definite storage function implies that there exists  $\epsilon > 0$  such that  $\int_{\mathbb{R}_-} Q_\Sigma(w) dt \geq \int_{\mathbb{R}_-} \epsilon |w|^2 dt$  for all  $w \in \mathfrak{B} \cap \mathfrak{D}$ . Let  $Q_\Psi$  be a positive definite storage function that satisfies the strict dissipation inequality

$$(6.2) \quad \frac{d}{dt}(Q_\Psi(w)) \leq Q_\Sigma(w) - \epsilon |w|^2, \text{ for all } w \in \mathfrak{B}.$$

Considering only those trajectories in  $\mathfrak{B} \cap \mathfrak{D}$  and integrating over  $\mathbb{R}_-$  we get

$$\int_{\mathbb{R}_-} Q_\Sigma(w) dt - \int_{\mathbb{R}_-} \epsilon |w|^2 dt \geq Q_\Psi(w)(0) \Rightarrow \int_{\mathbb{R}_-} Q_\Sigma(w) dt \geq \int_{\mathbb{R}_-} \epsilon |w|^2 dt,$$

where the last implication uses  $Q_\Psi(w) \geq 0$ .

Next we show that  $\mathfrak{B}$  being strictly dissipative with a positive storage function implies that  $\Lambda_{\text{un}} \subset \mathbb{C}^-$ . Observe that  $\mathfrak{B}$  being strictly  $\Sigma$ -dissipative implies that the partition  $w = (w_1, w_2)$  induced by  $\Sigma$  results in  $\mathfrak{B}_{\text{cont}}$  having a *proper* transfer function from  $w_1$  to  $w_2$ . Hence it follows from Proposition 2.2 that  $\mathfrak{B}$  allows an i/s/o representation  $\frac{d}{dt}x = Ax + Bw_1$ ,  $w_2 = Cx + Dw_1$ , where  $(A, B)$  is state uncontrollable and  $(C, A)$  is state observable. In order to show that  $\Lambda_{\text{un}} \subset \mathbb{C}^-$ , we shall show, in fact, that  $A$  is Hurwitz. We show this implication by contradiction, i.e., if  $\lambda \notin \mathbb{C}^-$  is an eigenvalue of  $A$ , then there does not exist *any* nonnegative definite storage functions. Let  $\lambda \in \text{spec}(A)$  and  $x_0 \neq 0$ , the corresponding eigenvector of  $A$ . The following  $w$  is an element of  $\mathfrak{B}$

$$(6.3) \quad \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ Cx_0 e^{\lambda t} \end{bmatrix}.$$

The behavior  $\mathfrak{B}$  being strictly  $\Sigma$ -dissipative implies that there exists a storage function  $Q_\Psi(w)$  that satisfies the following strict dissipation inequality:

$$(6.4) \quad \frac{d}{dt}(Q_\Psi(w)) \leq |w_1|^2 - |w_2|^2 - \epsilon |w|^2 \text{ for all } w \in \mathfrak{B}.$$

Putting  $w$  as in (6.3) we get, by direct differentiation of the left hand-side,

$$(6.5) \quad 2\Re(\lambda)Q_\Psi(w) \leq -x_0^T C^T C x_0 e^{2\Re(\lambda)t} - \epsilon |w|^2.$$

Consider the case when  $\lambda \in \mathbb{C}^+$ . The right-hand side of the above inequality is negative definite, which implies that  $Q_\Psi(w) \not\geq 0$  for this  $w \in \mathfrak{B}$ . This proves the contradiction for the case  $\Re(\lambda) > 0$ . Next we consider the case when  $\lambda \in j\mathbb{R}$ . In that case  $\Re(\lambda) = 0$ , which makes inequality (6.5) impossible. Thus we have shown that

$\lambda \notin \mathbb{C}^-$  eliminates nonnegativity of *any* storage function. Hence in order to have a positive definite storage function it is necessary that  $\Lambda_{\text{un}} \subset \mathbb{C}^-$ . This completes the proof that existence of a nonnegative storage function implies  $\Lambda_{\text{un}} \subset \mathbb{C}^-$ .  $\square$

Within the above proof, we have, in fact, shown that *every* storage function for the behavior is positive definite. This has been shown for the controllable case in [28, Theorem 6.4]. Intuitively, storage functions being positive is closer to their interpretation as energy-like functions. Also, the meaning of dissipativity that there is no source of energy in the system appeals to both positive definite storage functions and the stabilizability of the system. In the following section we explore other properties of the set of storage functions, like (un)boundedness of this set.

**7. Set of all storage functions for an uncontrollable system.** Another important topic of interest is the set of all storage functions of a dissipative behavior. For LQR/LQG theory and  $\mathcal{H}_\infty$  control, certain extremum storage functions give stabilizing controllers. In this section we show that the set of storage functions is unbounded for uncontrollable dissipative systems and that for stabilizable systems, this set is bounded from below.

We have seen before how the solutions to an ARI give storage functions as state functions. Thus to further explore the set of all storage functions we shall look into the set of solutions of a related ARI. It has been shown (in [27, 28], for example) that with respect to a given state space representation i.e., with respect to a given state map  $x = X(\frac{d}{dt})w$ , there is a one to one correspondence between storage functions and solutions to the ARI. Moreover, it is known that the set of storage functions for a controllable dissipative behavior is a bounded convex polytope with its vertices given by the storage functions coming from so-called spectral factorizations of  $\partial\Phi(\xi)$ . These storage functions correspond to the algebraic Riccati *equality* solutions. However, this set of solutions to the ARI turns out to lose the boundedness property when the behavior loses controllability. This constitutes the theorem below, one of the main results of this paper.

**THEOREM 7.1.** *Let  $\mathfrak{B} \in \mathfrak{L}^w$  be uncontrollable, and suppose the set of its uncontrollable poles  $\Lambda_{\text{un}}$  satisfies the unmixing property, i.e.,  $\Lambda_{\text{un}} \cap (-\Lambda_{\text{un}}) = \emptyset$ . Further, let  $\mathfrak{B}$  be strictly  $\Sigma$ -dissipative. Then the set of all storage functions is an unbounded convex set.*

*Proof.* By Proposition 2.2,  $\mathfrak{B}$  allows an i/s/o representation as  $\frac{d}{dt}x = \tilde{A}x + \tilde{B}w_1$ ,  $w_2 = \tilde{C}x + \tilde{D}w_1$ . Since  $\mathfrak{B}$  is strictly  $\Sigma$ -dissipative,  $(I_m - \tilde{D}^T\tilde{D}) > 0$  and all the storage functions come from real symmetric solutions of the following ARI

$$A^TK + KA + C^TC + KBB^TK \leq 0,$$

where  $A := \tilde{A} + \tilde{B}(I_m - \tilde{D}^T\tilde{D})^{-1}\tilde{D}^T\tilde{C}$ ,  $B := \tilde{B}(I_m - \tilde{D}^T\tilde{D})^{-\frac{1}{2}}$ , and  $C := (I_p - \tilde{D}\tilde{D}^T)^{-\frac{1}{2}}\tilde{C}$ . That the set of solutions to this ARI is convex is well known (see [4], for example). In order to show the unboundedness of the solution set, we once again make use of the Kalman decomposition result (Proposition A.3): there exists a similarity transformation that results in  $A, B, C$  matrices in the following forms:  $A = \begin{bmatrix} A_c & A_{cp} \\ 0 & A_u \end{bmatrix}$ ,  $B = \begin{bmatrix} B_c \\ 0 \end{bmatrix}$ , and  $C = [C_c \quad C_u]$ . In order to show that the set of ARI solutions is unbounded, we shall show that there exists a nonzero  $P \in \mathbb{R}_s^{n \times n}$  such that if  $K \in \mathbb{R}_s^{n \times n}$  is a solution to the ARI, then for all  $\lambda > 0$ , the new real symmetric matrix defined by  $\tilde{K} := K + \lambda P$  is also a solution of the ARI. Define  $P$

$$P := \begin{bmatrix} 0 & 0 \\ 0 & P_1 \end{bmatrix},$$

where  $P_1 \in \mathbb{R}_s^{n_u \times n_u}$  satisfies the following Lyapunov inequality

$$(7.1) \quad A_u^T P_1 + P_1 A_u \leq 0.$$

Since the set of uncontrollable poles satisfy the unmixing property, that is,  $\Lambda_{\text{un}} \cap (-\Lambda_{\text{un}}) = \emptyset$ , the Lyapunov inequality is guaranteed to have a nonzero solution. Consider the following expression in  $\tilde{K}$ :

$$\begin{aligned} & A^T \tilde{K} + \tilde{K} A + C^T C + \tilde{K} B B^T \tilde{K} \\ &= A^T K + K A + C^T C + K B B^T K + \lambda (A^T P + P A) + \lambda^2 P B B^T P \\ & \quad + \lambda (K B B^T P + P B B^T K). \end{aligned}$$

Using the Kalman decomposed form of  $A, B$ , and  $C$  above and the structure of  $P$ , the above expression simplifies to

$$(A^T K + K A + C^T C + K B B^T K) + \lambda \begin{bmatrix} 0 & 0 \\ 0 & A_u^T P_1 + P_1 A_u \end{bmatrix}.$$

From the fact that  $K$  is a solution to the ARI, and that  $P_1$  satisfies the Lyapunov inequality (7.1) the above expression is negative semidefinite for all  $\lambda > 0$ . Thus  $A^T \tilde{K} + \tilde{K} A + C^T C + \tilde{K} B B^T \tilde{K} \leq 0$ , and  $\tilde{K}$  is also a solution of the ARI for all  $\lambda > 0$ . This proves that the set of solutions to the ARI is unbounded.  $\square$

Notice that within the above proof we used that, if a solution  $K$  to the Riccati inequality exists, then a solution  $P_1$  to the Lyapunov inequality (7.1) can be added to  $K$  giving solutions  $\tilde{K}$  of the ARI. Though we have demonstrated the existence of solutions to the ARI primarily under the unmixing condition on the uncontrollable poles, this method shows that whenever the Riccati inequality admits solutions, uncontrollability forces the set of storage functions to be unbounded.

The following simple example shows a pictorial representation of the set of all storage functions for a controllable behavior.

*Example 7.2.* Consider behavior  $\mathfrak{B}$  having an i/s/o representation with  $A = \begin{bmatrix} 0 & -2 \\ 1 & -3 \end{bmatrix}$ ,  $B = \begin{bmatrix} 10 \\ 1 \end{bmatrix}$ ,  $C = [0 \quad 1]$  and supply rate  $S = \begin{bmatrix} 100 & 0 \\ 0 & -1 \end{bmatrix}$ . Let all the symmetric solutions to the corresponding ARI be of the form  $K = \begin{bmatrix} k_1 & k_2 \\ k_2 & k_3 \end{bmatrix}$ . The figure below shows the set of all ARI solutions. Clearly, the set is a bounded convex polytope.

*Remark 7.3.* Figure 7.1 corresponds to a controllable behavior, and hence the set is bounded. As the behavior becomes uncontrollable, all Lambda-sets no longer admit Riccati equation solutions (see Theorem 5.5). However, the set of storage functions is now unbounded along certain directions specified by the Lyapunov equation corresponding to the autonomous part of the behavior (as shown in Theorem 7.1). For some specific examples, the transition to uncontrollability causes the Riccati equation solutions corresponding to inadmissible Lambda-sets to move to infinity along the direction of the Lyapunov equation solution. Further, loosely speaking, when restricted to the controllable part, the storage functions corresponding to the Riccati equation solutions are unaffected by translation along this direction. It remains to formulate these observations concretely and prove them.

A very interesting fact about this unbounded set of all storage functions comes up for the case when the behavior is uncontrollable but stabilizable; i.e., the set of uncontrollable poles  $\Lambda_{\text{un}}$  is contained in the open left half of the complex plane (see the previous section for the definition and related results about stabilizability). We

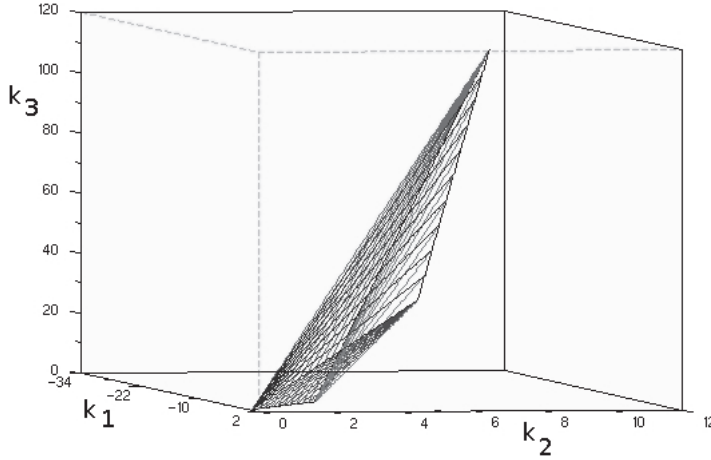


FIG. 7.1. The set of all storage functions.

show below that for stabilizability, the set of storage functions, though an unbounded set, is bounded from below. In other words, there exists a storage function  $Q_{\Psi_-}$  such that every storage function  $Q_{\Psi}$  satisfies  $Q_{\Psi}(w) - Q_{\Psi_-}(w) \geq 0$  for all  $w \in \mathfrak{B}$ . We state this result as a theorem below.

**THEOREM 7.4.** *Let  $\mathfrak{B} \in \mathfrak{L}^w$  be an uncontrollable, strictly  $\Sigma$ -dissipative behavior. Also assume that the set of uncontrollable poles  $\Lambda_{\text{un}} \subset \mathbb{C}^-$ . Then the set of all storage functions is bounded from below; i.e., there exists a storage function  $Q_{\Psi_-}(w)$  for  $\mathfrak{B}$  such that for each storage function  $Q_{\Psi}(w)$  for  $\mathfrak{B}$ ,  $Q_{\Psi_-}(w) \leq Q_{\Psi}(w)$  for all  $w \in \mathfrak{B}$ .*

Note the analogy of this result with that for controllable behaviors, where the set of storage functions is bounded and has a maximum and a minimum element (see [28, Theorem 5.7]). While we have shown unboundedness for the case of uncontrollability, stabilizability ensures the existence of the minimum element in this unbounded set.

*Proof.* To prove the above result, we use Corollary 5.6 and look into the solutions to the corresponding ARI. Let the behavior  $\mathfrak{B}$  have an i/s/o representation:  $\frac{d}{dt}x = \tilde{A}x + \tilde{B}w_1$ ,  $w_2 = \tilde{C}x + \tilde{D}w_1$ . Then the set of all storage functions comes from the solution set of the following ARI

$$A^T K + K A + C^T C + K B B^T K \leq 0,$$

where  $A := \tilde{A} + \tilde{B}(I_m - \tilde{D}^T \tilde{D})^{-1} \tilde{D}^T \tilde{C}$ ,  $B := \tilde{B}(I_m - \tilde{D}^T \tilde{D})^{-\frac{1}{2}}$ , and  $C := (I_p - \tilde{D} \tilde{D}^T)^{-\frac{1}{2}} \tilde{C}$ . Construct the corresponding Hamiltonian matrix

$$H := \begin{bmatrix} A & B B^T \\ -C^T C & -A^T \end{bmatrix}.$$

By assumption,  $\mathfrak{B}$  is strictly  $\Sigma$ -dissipative, which implies  $\text{spec}(H) \cap j\mathbb{R} = \emptyset$ . Also from Theorem 5.4,  $\text{spec}(H) \supset \Lambda_{\text{un}}$  and by assumption  $\Lambda_{\text{un}} \subset \mathbb{C}^-$ . These facts together imply that there exists a Lambda-set (say  $\Lambda$ ) of  $\text{spec}(H)$  such that  $\Lambda \supset \Lambda_{\text{un}}$  and  $\Lambda \subset \mathbb{C}^-$ . By Theorem 5.5, since  $\Lambda \supseteq \Lambda_{\text{un}}$ , there exists a real symmetric solution,  $K_-$  to the ARE such that  $\text{spec}(A + B B^T K_-) = \Lambda$ . We shall show that this particular solution to the ARE serves as a “minimum” storage function. To show this consider any real symmetric solution  $K$  to the ARI

$$(7.2) \quad A^T K + K A + C^T C + K B B^T K \leq 0.$$

Also rewrite the ARE in  $K_-$  as follows

$$(7.3) \quad (A + BB^T K_-)^T K_- + K_- (A + BB^T K_-) + C^T C - K_- BB^T K_- = 0.$$

Subtracting (7.3) from (7.2), and further adding and subtracting the terms  $K_- BB^T K_-$  and  $K BB^T K_-$  we get

$$(7.4) \quad (A + BB^T K_-)^T (K - K_-) + (K - K_-) (A + BB^T K_-) + (K - K_-) BB^T (K - K_-) \leq 0.$$

Observe that the above inequality closely resembles the Lyapunov inequality. In order to conclude that  $(K - K_-) \geq 0$ , it is enough to notice that  $\text{spec}(A + BB^T K_-) = \Lambda \subset \mathbb{C}^-$ , and the constant-like term  $(K - K_-) BB^T (K - K_-)$  is nonnegative. This proves our claim that the set of all storage functions is bounded from below.  $\square$

We saw in the previous section that, for a strictly dissipative and stabilizable behavior  $\mathfrak{B}$ , dissipativity on  $\mathbb{R}_-$  of  $\mathfrak{B}_{\text{cont}}$  assures the existence of positive definite storage functions. Combining this result with the one above, we infer that the lower bound of the set of storage functions is, in fact, positive (see discussions following Theorem 6.1). This formalizes the intuition that such a system is devoid of any energy sources within it, and hence the maximum extractable energy<sup>3</sup> from *any* given state is bounded.

Using a very similar argument as in the above proof, one can show that if the behavior is antistabilizable, meaning all the uncontrollable poles are unstable, i.e.,  $\Lambda_{\text{un}} \subset \mathbb{C}^+$ , then the set of storage functions is bounded *from above*.

**8. Necessity of the unmixing of uncontrollable poles.** As seen in Theorem 3.4, the unmixing property of the uncontrollable poles makes strict dissipativity of the controllable part equivalent to that of the whole behavior. In this section we shall see to what extent the unmixing property is necessary. As mentioned in the introduction, the unmixing property serves as a sufficient condition for solvability of a Lyapunov equation and the corresponding Lyapunov operator becomes singular when this condition is not satisfied. We show in this section that the Lyapunov operator is onto if and only if there exists an observable rank one symmetric matrix in its image. This interesting result is utilized for exploring the extent of the unmixing condition for strict dissipativity. Theorem 8.3 below shows that for a system with single output, unmixing is necessary for existence of nonzero lossless trajectories satisfying a dissipation *equality* (see Remark 8.4 below).

When the Lyapunov operator is singular, then nonsymmetric solutions can exist even when the constant term is symmetric. This general solvability condition can be obtained from a certain eigenspace of a Hamiltonian matrix. The fact that a Lyapunov equation is a special case of an ARE with the quadratic term being zero motivates the following result. We state this as a lemma below since it will be needed later in this section. Related results on Lyapunov equation solvability can be found in [24].

**LEMMA 8.1.**  $K \in \mathbb{R}^{n \times n}$  is a solution (not necessarily symmetric) to the following Lyapunov equation  $A^T K + K A + C^T C = 0$  if and only if  $\text{im} \begin{bmatrix} I_n \\ K \end{bmatrix}$  is an invariant space of the Hamiltonian matrix defined as

$$H = \begin{bmatrix} A & 0 \\ -C^T C & -A^T \end{bmatrix}.$$

<sup>3</sup>This has been called *available storage* in [28].

*Proof. (If)* We first assume that  $\mathcal{X}_A := \text{im} \begin{bmatrix} I_n \\ K \end{bmatrix}$  is an  $n$ -dimensional invariant subspace of  $H$ . We shall show that this implies  $K$  satisfies the Lyapunov equation  $A^T K + K A + C^T C = 0$ . Clearly  $\begin{bmatrix} K & -I_n \end{bmatrix} \begin{bmatrix} I_n \\ K \end{bmatrix} = 0$ . Since  $\mathcal{X}_A$  is  $H$ -invariant the last equation can be written as  $\begin{bmatrix} K & -I_n \end{bmatrix} \begin{bmatrix} A & 0 \\ -C^T C & -A^T \end{bmatrix} \begin{bmatrix} I_n \\ K \end{bmatrix} = 0$ . Hence  $A^T K + K A + C^T C = 0$ .

**(Only if)** In this part we shall show that  $K$  being a solution to the Lyapunov equation  $A^T K + K A + C^T C = 0$  implies that  $\text{im} \begin{bmatrix} I_n \\ K \end{bmatrix}$  is an  $n$ -dimensional invariant subspace of  $H$ . Assuming  $K$  satisfies the Lyapunov equation, we can write the following equality:

$$\begin{bmatrix} A & 0 \\ -C^T C & -A^T \end{bmatrix} \begin{bmatrix} I_n \\ K \end{bmatrix} = \begin{bmatrix} I_n \\ K \end{bmatrix} A.$$

This implies that  $\text{im} \begin{bmatrix} I_n \\ K \end{bmatrix}$  is an eigenspace of  $H$  corresponding to  $\text{spec}(A)$ .  $\square$

It follows from the lemma above that there is only one set of eigenvalues of the Hamiltonian matrix, which gives a solution to the Lyapunov equation: namely, the eigenvalues of  $A$ . Unlike the ARE, for the Lyapunov equation there is no choice for the set of eigenvalues of the Hamiltonian matrix. Another important fact that follows is that the set of eigenvalues, i.e.,  $\text{spec}(A)$  is no longer a Lambda-set when there is mixing, i.e.,  $\text{spec}(A) \cap \text{spec}(-A) \neq \emptyset$ . Owing to this particular fact, the eigenspace of the Hamiltonian matrix which gives a solution is no longer guaranteed to be perpendicular to the left-eigenspace corresponding to the rest of the eigenvalues of  $H$ . However, from Lemma A.4 if  $\text{im} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$  is a (generalized) right-eigenspace of dimension  $n$  corresponding  $\text{spec}(A)$ , then  $\text{im} \begin{bmatrix} X_2 \\ -X_1 \end{bmatrix}$  is an  $n$ -dimensional (generalized) left-eigenspace corresponding to  $\text{spec}(-A)$ . Since the right and left eigenspaces are no longer guaranteed to be perpendicular, we do not necessarily have  $X_2 X_1^{-1} = X_1^{-T} X_2^T$ . This implies that when the unmixing condition is not satisfied there can be nonsymmetric and nonunique solutions to the Lyapunov equation. However, the existence of a nonsymmetric solution guarantees existence of a symmetric solution. If  $K$  is a solution to the Lyapunov equation, then so are  $K^T$  and  $(K + K^T)/2$ . With this simple observation we now give a necessary and sufficient condition for the existence of solution to a Lyapunov equation for the special case that the constant term is of rank one. Interestingly for this case when the constant term is rank 1 the unmixing condition becomes necessary!

**THEOREM 8.2.** *Consider the Lyapunov equation  $A^T K + K A + C^T C = 0$  with  $(C, A)$  pair observable. Assume  $\text{rank}(C^T C) = 1$ . Then there exists a solution  $K$  to the Lyapunov equation if and only if  $\text{spec}(A) \cap \text{spec}(-A) = \emptyset$ .*

*Proof. (If)* This implication is well known (see [26], for example). In fact, uniqueness and symmetry are guaranteed.

**(Only if)** Suppose  $A$  has a mixed spectrum, that is,  $\text{spec}(A) \cap \text{spec}(-A) \neq \emptyset$ . We shall show that this implies that the Lyapunov equation  $A^T K + K A + C^T C = 0$  has no solution  $K$ . In Theorem 8.1 it has been shown that the eigenspace of the Hamiltonian matrix corresponding to only  $\text{spec}(A)$  can give a solution to the Lyapunov equation. Hence to prove that there does not exist any solution  $K$ , it is sufficient to show that any eigenvector of the Hamiltonian matrix

$$H := \begin{bmatrix} A & 0 \\ -C^T C & -A^T \end{bmatrix},$$

corresponding to an eigenvalue  $\lambda \in \text{spec}(A) \cap \text{spec}(-A)$  shall have all upper  $n$  elements zero. For that, let  $v := \text{col}(v_1, v_2) \in \mathbb{C}^{2n}$ ,  $v \neq 0$  be an eigenvector of  $H$  corresponding to  $\lambda \in \text{spec}(A) \cap \text{spec}(-A^T)$ . It then follows that

$$\begin{bmatrix} A & 0 \\ -C^T C & -A^T \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \end{bmatrix},$$

which means  $Av_1 = \lambda v_1$  and  $[\lambda I_n - (-A^T)]v_2 = -C^T C v_1$ . Suppose  $v_1 \neq 0$ , then the above implies  $v_1$  is an eigenvector of  $A$ . Since  $(C, A)$  is observable  $Cv_1 \neq 0$ . Further,  $\text{rank}(C^T C) = 1$ ,  $Cv_1 \neq 0$  together imply  $[\lambda I_n - (-A^T)]v_2 = -C^T C v_1$  implies<sup>4</sup>

$$(8.1) \quad \text{im } C^T \subseteq \text{im } [\lambda I_n - (-A^T)].$$

This gives  $\ker C \supseteq \ker [\lambda I_n - (-A)]$ . Since  $\lambda \in \text{spec}(A) \cap \text{spec}(-A)$ ,  $-\lambda \in \text{spec}(A)$ . But the inclusion  $\ker C \supseteq \ker [\lambda I_n - (-A)]$  implies  $\begin{bmatrix} -\lambda I_n & -A \\ C & \end{bmatrix}$  loses rank, which means  $-\lambda$  is unobservable. This contradicts the observability of  $(C, A)$ . Hence  $v_1 = 0$ , and therefore, the eigenspace of  $H$  corresponding to  $\text{spec}(A)$  cannot have the form  $\text{im } \begin{bmatrix} I_n \\ K \end{bmatrix}$ . Thus, using Lemma 8.2, the Lyapunov equation is not solvable.  $\square$

It is well known that the unmixing condition is equivalent to existence and uniqueness of solution to the Lyapunov equation. In other words, the unmixing condition is equivalent to the image of the Lyapunov operator containing *all* symmetric matrices. The above theorem shows that unmixing is necessary and sufficient for the image to contain a symmetric matrix of rank one (satisfying observability conditions). However, the corresponding Lyapunov inequality can have solutions when the equation is not solvable; see Remark 8.4 below.

We have seen in the previous sections how dissipativity is related to the solvability of certain ARE/ARI. For an uncontrollable system the corresponding ARE behaves like a Lyapunov equation on certain subspaces of the state space. Following this observation we shall now utilize Theorem 8.2 to show how the unmixing becomes necessary for the solvability of the ARE. We shall make use of the fact that  $C^T C$  being rank one means that the system has only one output.

**THEOREM 8.3.** *Consider  $\mathfrak{B} \in \mathfrak{L}^w$  having a single output, i.e.,  $p(\mathfrak{B}) = 1$ . Suppose  $R \in \mathbb{R}^{1 \times w}[\xi]$ , with  $R \neq 0$ , gives a kernel representation for  $\mathfrak{B}$ . Define  $\Lambda_{\text{un}} = \{\lambda \in \mathbb{C} \mid R(\lambda) = 0\}$  and let  $(A, B, C, D)$  give a minimal i/s/o representation for  $\mathfrak{B}$ , with  $(C, A)$  observable. Then, the following are equivalent.*

1.  $I_m - D^T D > 0$  and there exists  $K$  satisfying

$$(8.2) \quad A^T K + K A + C^T C + (K B + C^T D)(I_m - D^T D)^{-1}(B^T K + D^T C) = 0,$$

with  $\text{spec}(A + B(I_m - D^T D)^{-1}(B^T K + D^T C)) \cap j\mathbb{R} = \emptyset$ .

2.  $\Lambda_{\text{un}} \cap (-\Lambda_{\text{un}}) = \emptyset$  and  $\mathfrak{B}_{\text{cont}}$  is strictly  $\Sigma$ -dissipative.

*Proof.* **(2  $\Rightarrow$  1)** From Lemma 5.3 strict dissipativity of  $\mathfrak{B}_{\text{cont}}$  implies  $I_m - D^T D > 0$ , so the following Hamiltonian matrix exists

$$(8.3) \quad H := \begin{bmatrix} A + B(I_m - D^T D)^{-1} D^T C & B(I_m - D^T D)^{-1} B^T \\ -C^T (I_p - D D^T)^{-1} C & -\left(A + B(I_m - D^T D)^{-1} D^T C\right)^T \end{bmatrix}.$$

<sup>4</sup>Existence of a solution  $v_2$  for just one  $v_1 \neq 0$  implies the inclusion (8.1) because of the rank condition.



Let the observable image representation of  $\mathfrak{B}_{\text{cont}}$  induce the para-Hermitian matrix  $\partial\Phi(\xi)$ . Then by Theorem 5.4 we get  $\text{spec}(H) = \text{roots} [\det \partial\Phi(\xi) \chi_{\text{un}}(\xi) \chi_{\text{un}}(-\xi)]$ . Once again strict  $\Sigma$ -dissipativity of  $\mathfrak{B}_{\text{cont}}$  implies  $\partial\Phi(j\omega) > 0$  for all  $\omega \in \mathbb{R}$  (see Theorem 5.2) meaning  $\text{roots} (\det \partial\Phi(\xi))$  has no roots on the imaginary axis. This together with the assumption  $\Lambda_{\text{un}} \cap (-\Lambda_{\text{un}}) = \emptyset$  implies that  $\text{spec}(H) \cap j\mathbb{R} = \emptyset$  and so it allows a Lambda-set  $\Lambda \supseteq \Lambda_{\text{un}}$ . It then follows from Theorem 5.5 that there exists  $K \in \mathbb{R}^{n \times n}$  such that  $\text{im} \begin{bmatrix} I_n \\ K \end{bmatrix}$  is the  $n$ -dimensional invariant subspace of  $H$  corresponding to  $\Lambda$ . This shows that  $K$  solves the ARE. Also this solution  $K$  has the property  $\text{spec}(A + B(I_m - D^T D)^{-1}(B^T K + D^T C)) = \Lambda$  (see Proposition 4.4) which means  $\text{spec}(A + B(I_m - D^T D)^{-1}(B^T K + D^T C)) \cap j\mathbb{R} = \emptyset$  because  $\Lambda \subset \text{spec}(H)$  and  $\text{spec}(H) \cap j\mathbb{R} = \emptyset$ .

**(1  $\Rightarrow$  2)** We shall first show that statement 1 implies that  $\mathfrak{B}_{\text{cont}}$  is strictly  $\Sigma$ -dissipative.  $I_m - D^T D$  is assumed to be positive definite, and therefore the Hamiltonian matrix of equation (8.3) exists. Again, Theorem 5.4 implies  $\text{spec}(H) = \text{roots} [\det \partial\Phi(\xi) \chi_{\text{un}}(\xi) \chi_{\text{un}}(-\xi)]$ . Counting multiplicities both sides we get  $\deg (\det \partial\Phi(\xi)) = 2n(\mathfrak{B}_{\text{cont}})$ . We claim that  $H$  has no eigenvalues on the imaginary axis, because otherwise any solution to the ARE, if it exists, would result in  $\text{spec}(A + B(I_m - D^T D)^{-1}(B^T K + D^T C)) \cap j\mathbb{R} \neq \emptyset$  which contradicts statement 1. Thus  $\partial\Phi(\xi)$  also has no zeros on the imaginary axis, which implies  $\partial\Phi(j\omega) > 0$  for all  $\omega \in \mathbb{R}$ . This positive definiteness together with  $\deg (\det \partial\Phi(\xi)) = 2n_c$  implies that  $\mathfrak{B}_{\text{cont}}$  is strictly  $\Sigma$ -dissipative (see Theorem 5.2).

In order to complete the proof it remains to show that the ARE (8.2) having a solution implies that  $\Lambda_{\text{un}}$  is unmixed. Since  $(C, A)$  is observable  $K$  is nonsingular (see [7, Lemma 3]). We have assumed that  $\mathfrak{B}$  is uncontrollable; therefore,  $(A, B)$  pair is also uncontrollable and the set of uncontrollable poles  $\Lambda_{\text{un}}$  is exactly equal to the set of uncontrollable eigenvalues of  $A$  (this follows from Proposition 2.2). So if  $\Lambda_{\text{un}}$  has cardinality  $n_u$  (counting multiplicities), then there exists a full column rank matrix  $T_u \in \mathbb{R}^{n \times n_u}$  obtained from the generalized eigenvectors of  $A^T$  such that  $A^T T_u = T_u A_u$  and  $T_u^T B = 0$ , where  $A_u \in \mathbb{R}^{n_u \times n_u}$  is in Jordan form with  $\text{spec}(A_u) = \Lambda_{\text{un}}$ . Also, since  $K$  is nonsingular, there exists  $T \in \mathbb{R}^{n \times n_u}$  full column rank, such that  $KT = T_u$ . Then pre- and postmultiplying the ARE (8.2) by  $T^T$  and  $T$ , respectively, and making use of the fact that  $A^T T_u = T_u A_u$  and  $T_u^T B = 0$  we get

$$T^T K T A_u + A_u^T T^T K T + T^T C^T (I_p - D D^T)^{-1} C T = 0.$$

Define  $K_u := T^T K T \in \mathbb{R}^{n_u \times n_u}$  and  $C_u := (I_p - D D^T)^{-\frac{1}{2}} C T$ . The above Lyapunov equation can be rewritten as  $A_u^T K_u + K_u A_u + C_u^T C_u = 0$ .

Now we show that the new positive semidefinite matrix  $C_u^T C_u$  is also rank one. In order to prove this we shall use a contradiction argument. Since  $T$  is full column rank,  $\text{rank } C_u^T C_u \leq 1$ . Assume  $\text{rank } C_u^T C_u = 0$ . This implies  $C T = 0$  because  $\mathfrak{B}$  being strictly  $\Sigma$ -dissipative implies  $(I_p - D D^T)^{-1} > 0$  (see footnote 2 in section 4). Postmultiplying the ARE (8.2) by  $T$  and utilizing the fact that  $B^T K T = 0$  we get  $A^T K T + K A T + C^T (I_p - D D^T)^{-1} C T = 0$ . Since  $C T = 0$  and  $A^T T_u = T_u A_u$  the last equation gives

$$(8.4) \quad K T A_u + K A T = 0.$$

Equation (8.4) along with the fact that  $K$  is nonsingular implies that  $\text{im } T$  is  $A$ -invariant. This means  $\text{im } T$  is a nontrivial  $A$ -invariant subspace contained in  $\ker C$ , which is not possible due to observability of  $(C, A)$ . Hence  $C T \neq 0$  and so

$\text{rank}(C_u^T C_u) = 1$ . Therefore, from Theorem 8.2 it follows that  $\text{spec}(A_u) = \Lambda_{\text{un}}$  satisfies the unmixing property.  $\square$

*Remark 8.4.* Statement 1 above would have been equivalent to strict dissipativity of  $\mathfrak{B}$  if  $\mathfrak{B}$  were controllable, but this is not the case in general. More precisely, under controllability assumptions on  $(A, B)$ , the ARE not having a solution implies that the ARI also does not have a solution (see [22, 23]). However, this turns out to be untrue for an uncontrollable behavior. For example, consider the special case of an autonomous system with  $A = \text{diag}([1 \ -1])$  and  $C = [1 \ 1]$ . For this case the Lyapunov equation is not solvable, though the corresponding inequality admits a solution. More generally the ARE having a solution means that there exist nonzero lossless trajectories in the behavior (see [1]). For a controllable dissipative behavior, there are always nonzero lossless trajectories. The last theorem makes it clear that for an uncontrollable system with a single output, unmixing of uncontrollable poles is necessary for the existence of nontrivial lossless trajectories. We do not digress more into the notion of losslessness because it is not central to the subject of this paper. A thorough understanding of the necessity of the unmixing condition requires further investigation.

That the unmixing is not necessary in general for more than one output is quite expected. The following example gives one such simple instance.

*Example 8.5.* Consider the autonomous system with i/s/o representation  $\frac{d}{dt}x = Ax$ ,  $y = Cx$ , where

$$A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

Observe that  $A$  has “mixed” eigenvalues, i.e.,  $\Lambda_{\text{un}} \cap (-\Lambda_{\text{un}}) \neq \emptyset$ .  $\Sigma$ -dissipativity of such an autonomous system together with  $\sigma_+(\Sigma) = \mathfrak{m}(\mathfrak{B})$  is equivalent to existence of a real symmetric solution to the following Lyapunov inequality:  $A^T K + K A + C^T C \leq 0$ . Notice that  $K = \begin{bmatrix} 2+b & a \\ a & -2-c \end{bmatrix}$  with  $a, b, c \in \mathbb{R}$  and  $b, c \geq 0$  gives a solution to the above Lyapunov inequality. This example shows that the unmixing condition of uncontrollable poles is not necessary for the system to be dissipative.

**9. Concluding remarks.** In this paper we studied dissipativity for a general, possibly uncontrollable, LTI system. Our starting point was a more appropriate, though less often used, definition of dissipativity in terms of a differential inequality called the dissipation inequality. With this definition we brought out an equivalence between the dissipativities of a behavior and its controllable part, under the important unmixing condition (Theorem 3.4). For the case of strict dissipativity, Theorem 3.4 also settles the issue of whether to allow unobservable variables in the defining dissipation inequality: the theorem rules out the requirement of unobservable variables. The important intuitive idea that storage of energy should take place through the state variables comes as a natural consequence of Theorem 3.4. We stated this result as a corollary.

Next we looked into the set of all storage functions for a strictly  $\Sigma$ -dissipative system. It is well known that this set is a bounded convex polyhedron for a controllable system. We showed that for an uncontrollable system the set loses its boundedness property. Further, this set becomes bounded from below if the system is stabilizable. If in addition the controllable part is strictly  $\Sigma$ -dissipative on  $\mathbb{R}_-$ , then we showed that this lower bound on the set is positive. We used this result to formalize the physical notion of stored energy being finite in a dissipative system that has no source of energy within: it is *not* possible to extract an indefinite amount of energy from a stabilizable

system whose controllable part is strictly dissipative on  $\mathbb{R}_-$ . In this paper, though we utilized results about Hamiltonian matrices and Riccati equations, our main results pertain to the system directly, without an *a priori* input/output partition of the system variables. This is an important feature and advantage of the behavioral approach.

The unmixing condition plays a crucial role in most of the main results of this paper. In order to address the necessity of the unmixing condition we made use of the important observation that for an uncontrollable behavior a storage function acts like a Lyapunov function over certain trajectories, and we showed that unmixing is not necessary in general for existence of a Lyapunov function and therefore for dissipativity. However, an interesting situation arises when the system has only one output. In Theorem 8.2 we showed that under suitable observability conditions a singular Lyapunov operator cannot have a rank one symmetric matrix in its image. This helped us to bring out the fact that the unmixing is necessary for an uncontrollable behavior to have nonzero lossless trajectories in it. The extent of unmixing condition for a more general situation remains to be investigated.

In this paper we have dealt only with the maximum input cardinality case, i.e., the case when the number of inputs is equal to the positive signature of the supply rate function  $\Sigma$ . A study of the general case can also be utilized for dissipativity synthesis problems for uncontrollable systems.

#### Appendix A. Auxiliary results and proofs.

The following standard result from state space theory [34] is needed in the proof of Lemma 4.5; we state it for easy reference.

**PROPOSITION A.1.** *Consider the following state space representation of a dynamical system.  $\frac{d}{dt}x = Ax + Bw_1$ ,  $w_2 = Cx + Dw_1$ . Let  $F_c \in \mathbb{R}^{m \times n}$ ,  $F_o \in \mathbb{R}^{n \times p}$  and  $G_c \in \mathbb{R}^{m \times m}$ ,  $G_o \in \mathbb{R}^{p \times p}$  with  $G_c, G_o$  nonsingular. Then,*

- $(A + BF_c, BG_c)$  is controllable if and only if  $(A, B)$  is controllable,
- $(G_o C, A + F_o C)$  is observable if and only if  $(C, A)$  is observable.

*Proof of Lemma 4.5.* As seen earlier, the supply function matrix  $\Sigma$  induces an input/output partition  $w = (w_1, w_2)$  where  $w_1$  is input and  $w_2$  is output. To prove this theorem we shall consider two cases, first the case when  $\mathfrak{B}$  has a *strictly* proper transfer function from  $w_1$  to  $w_2$  and secondly the case when  $\mathfrak{B}$  has a proper transfer function (not necessarily strictly proper) with the same input/output partition. For the second case we shall show that there exists a simple transformation that changes it to the first case and thus the proof is complete.

**Case 1 ( $\mathfrak{B}$  has strictly proper transfer function from  $w_1$  to  $w_2$ ) :**  $\mathfrak{B}$  allows an i/s/o representation with  $D = 0$  as follows

$$(A.1) \quad \frac{d}{dt}x = Ax + Bw_1, \quad w_2 = Cx.$$

$\Sigma$  is given as  $\begin{bmatrix} I_n & 0 \\ 0 & -I_p \end{bmatrix}$ , so the corresponding Hamiltonian matrix is  $H = \begin{bmatrix} A & BB^T \\ -C^T C & -A^T \end{bmatrix}$ . Consider the observable image representation of  $\mathfrak{B}$ :  $w = M(\frac{d}{dt})\ell$ , and partition  $M$  corresponding to  $w = (w_1, w_2)$  to get

$$(A.2) \quad \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} W_1(\frac{d}{dt}) \\ W_2(\frac{d}{dt}) \end{bmatrix} \ell.$$

Without loss of generality we can assume that  $\det(W_1(\xi))$  is a monic polynomial. Now, the transfer function for  $\mathfrak{B}$  is given by  $G(\xi) := W_2(\xi)W_1^{-1}(\xi) = C(\xi I_n - A)^{-1}B$ .

The characteristic polynomial of  $H$  is given by  $\chi(H) = \det(\xi I_{2n} - H)$ . Using Schur complement this can be written as

$$\begin{aligned}\chi(H) &= \det(\xi I_n - A) \det[(\xi I_n + A^T) + C^T C(\xi I_n - A)^{-1} B B^T] \\ &= \det(\xi I_n - A) \det(\xi I_n + A^T) \det\left[I_n + (\xi I_n + A^T)^{-1} C^T C(\xi I_n - A)^{-1} B B^T\right].\end{aligned}$$

Now using the identity,  $\det(I + PQ) = \det(I + QP)$  we get

$$\begin{aligned}\chi(H) &= \det(\xi I_n - A) \det(\xi I_n + A^T) \det\left[I_m + B^T (\xi I_n + A^T)^{-1} C^T C(\xi I_n - A)^{-1} B\right] \\ &= \det(\xi I_n - A) \det(\xi I_n + A^T) \det[I_m - G^T(-\xi)G(\xi)] \\ &= \det(\xi I_n - A) \det(\xi I_n + A^T) \det\left[W_1^{-T}(-\xi)(W_1^T(-\xi)W_1(\xi) \right. \\ &\quad \left. - W_2^T(-\xi)W_2(\xi))W_1^{-1}(\xi)\right] \\ &= \left(\frac{\det(\xi I_n - A) \det(\xi I_n + A^T)}{\det W_1^T(-\xi) \det W_1(\xi)}\right) \det[W_1^T(-\xi)W_1(\xi) - W_2^T(-\xi)W_2(\xi)].\end{aligned}$$

Since  $(A, B)$  is controllable and  $(C, A)$  is observable, and due to observability of the image representation (A.2),  $\det W_1(\xi) = \det(\xi I_n - A)$ ; therefore, the above equation simplifies to

$$\begin{aligned}\chi(H) &= -\det[W_1^T(-\xi)W_1(\xi) - W_2^T(-\xi)W_2(\xi)] \\ &= -\det \partial \Phi(\xi)\end{aligned}$$

This proves  $\text{spec}(H) = \text{roots}(\det \partial \Phi(\xi))$  and hence the lemma for the strictly proper transfer matrix case.

**Case 2 ( $\mathfrak{B}$  has proper transfer function from  $w_1$  to  $w_2$ ):**  $\mathfrak{B}$  allows the following i/s/o representation,  $\frac{d}{dt}x = Ax + Bw_1$ ,  $w_2 = Cx + Dw_1$  with  $(A, B)$  controllable and  $(C, A)$  observable. Since  $(I_m - D^T D) > 0$ , the corresponding Hamiltonian matrix is given by

$$H := \begin{bmatrix} A + B(I_m - D^T D)^{-1} D^T C & B(I_m - D^T D)^{-1} B^T \\ -C^T(I_p - DD^T)^{-1} C & -(A + B(I_m - D^T D)^{-1} D^T C)^T \end{bmatrix}.$$

Because  $D$  is nonzero, the arguments in case 1 do not hold. As mentioned earlier, we show that there exists a transformation on the manifest variables that changes this situation to that in case 1. Consider the quadratic form  $Q_\Sigma(w)$

$$\begin{aligned}Q_\Sigma(w) &= |w_1|^2 - |w_2|^2 = w_1^T w_1 - x^T C^T C x - x^T C^T D w_1 - w_1^T D^T C x - w_1^T D^T D w_1 \\ &= \begin{bmatrix} w_1 \\ x \end{bmatrix}^T \Sigma_1 \begin{bmatrix} w_1 \\ x \end{bmatrix},\end{aligned}$$

where  $\Sigma_1$  is defined as  $\Sigma_1 := \begin{bmatrix} I_n - D^T D & -D^T C \\ -C^T D & -C^T C \end{bmatrix}$ . Notice that  $\Sigma_1$  can be factorized as

$$(A.3) \quad \begin{bmatrix} (I_m - D^T D)^{\frac{1}{2}} & -(I_m - D^T D)^{-\frac{1}{2}} D^T C \\ 0 & I_n \end{bmatrix}^T \tilde{\Sigma} \begin{bmatrix} (I_m - D^T D)^{\frac{1}{2}} & -(I_m - D^T D)^{-\frac{1}{2}} D^T C \\ 0 & I_n \end{bmatrix},$$

where  $\tilde{\Sigma} := \begin{bmatrix} I_m & 0 \\ 0 & -C^T(I_p - DD^T)^{-1}C \end{bmatrix}$ . Define  $T := \begin{bmatrix} (I_m - D^TD)^{\frac{1}{2}} & -(I_m - D^TD)^{-\frac{1}{2}}D^TC \\ 0 & I_n \end{bmatrix}$ . Further define  $\begin{bmatrix} \tilde{w}_1 \\ \tilde{x} \end{bmatrix} := T \begin{bmatrix} w_1 \\ x \end{bmatrix}$ . Then from (A.3) we have

$$(A.4) \quad Q_{\Sigma}(w) = \begin{bmatrix} \tilde{w}_1 \\ \tilde{x} \end{bmatrix}^T \tilde{\Sigma} \begin{bmatrix} \tilde{w}_1 \\ \tilde{x} \end{bmatrix}.$$

Now, since  $(w_1, x)$  satisfies the state equation,  $(\tilde{w}_1, \tilde{x})$  must satisfy the following:

$$\begin{aligned} & [-B \quad \frac{d}{dt}I_n - A] T^{-1} \begin{bmatrix} \tilde{w}_1 \\ \tilde{x} \end{bmatrix} = 0 \\ \Rightarrow & [-B \quad \frac{d}{dt}I_n - A] \begin{bmatrix} (I_m - D^TD)^{-\frac{1}{2}} & (I_m - D^TD)^{-1}D^TC \\ 0 & I_n \end{bmatrix} \begin{bmatrix} \tilde{w}_1 \\ \tilde{x} \end{bmatrix} = 0 \\ \Rightarrow & [-B(I_m - D^TD)^{-\frac{1}{2}} \quad \frac{d}{dt}I_n - A - B(I_m - D^TD)^{-1}D^TC] \begin{bmatrix} \tilde{w}_1 \\ \tilde{x} \end{bmatrix} = 0. \end{aligned}$$

Therefore,  $(\tilde{w}_1, \tilde{x})$  satisfies the state equation,  $\frac{d}{dt}\tilde{x} = \tilde{A}\tilde{x} + \tilde{B}\tilde{w}_1$ , where  $\tilde{A} := A + B(I_m - D^TD)^{-1}D^TC$ ,  $\tilde{B} := B(I_m - D^TD)^{-\frac{1}{2}}$ . By denoting  $\tilde{C} := (I_p - DD^T)^{-\frac{1}{2}}C$ , we can define a new behavior  $\tilde{\mathfrak{B}}$  that has an i/s/o representation given by  $\frac{d}{dt}\tilde{x} = \tilde{A}\tilde{x} + \tilde{B}\tilde{w}_1$ ,  $\tilde{w}_2 = \tilde{C}\tilde{x}$ . Now, from (A.4) we have  $\tilde{w} := \text{col}(\tilde{w}_1, \tilde{w}_2)$  satisfies

$$(A.5) \quad Q_{\Sigma}(\tilde{w}) = Q_{\Sigma}(w).$$

In order to infer about the controllability of the new behavior  $\tilde{\mathfrak{B}}$ , we make use of Proposition A.1 above; since  $(A, B)$  is controllable, so is  $(\tilde{A}, \tilde{B})$ , and similarly, since  $(C, A)$  is observable, so is  $(\tilde{C}, \tilde{A})$ . Therefore,  $\tilde{\mathfrak{B}}$  allows an observable image representation

$$\tilde{\mathfrak{B}} = \text{im} \begin{bmatrix} \tilde{W}_1(\frac{d}{dt}) \\ \tilde{W}_2(\frac{d}{dt}) \end{bmatrix}.$$

From (A.5) we have

$$\partial\tilde{\Phi}(\xi) := \tilde{W}_1^T(-\xi)\tilde{W}_1(\xi) - \tilde{W}_2^T(-\xi)\tilde{W}_2(\xi) = \partial\Phi(\xi).$$

Also,  $\tilde{\mathfrak{B}}$  gives the Hamiltonian matrix,  $\tilde{H} = \begin{bmatrix} \tilde{A} & \tilde{B}\tilde{B}^T \\ -\tilde{C}^T\tilde{C} & -\tilde{A}^T \end{bmatrix} = H$ . Thus for every  $\mathfrak{B} \in \mathfrak{L}_{\text{cont}}^w$  with an input/output partition such that the transfer function is proper, there exists  $\tilde{\mathfrak{B}} \in \mathfrak{L}_{\text{cont}}^w$  with a corresponding i/o partition that gives a strictly proper transfer function, such that  $\partial\Phi(\xi)$  matrices coming from  $\mathfrak{B}$  and  $\tilde{\mathfrak{B}}$  are the same and so are the corresponding Hamiltonian matrices. Therefore, from case 1 it follows that

$$\text{spec}(\tilde{H}) = \text{roots}(\det \partial\tilde{\Phi}(\xi)) \Rightarrow \text{spec}(H) = \text{roots}(\det \partial\Phi(\xi)).$$

This completes the proof of Lemma 4.5.  $\square$

*Proof of Lemma 5.3.* Let the image representation matrix have a partition as

$$M(\xi) = \begin{bmatrix} W_1(\xi) \\ W_2(\xi) \end{bmatrix}; \quad W_1 \in \mathbb{R}^{m \times m}[\xi], W_2 \in \mathbb{R}^{p \times m}[\xi].$$

Since the image representation is observable,  $W_1(\xi)$  and  $W_2(\xi)$  are relatively right-prime. So the transfer function from  $w_1$  to  $w_2$  is given by  $G(\xi) = W_2(\xi)W_1^{-1}(\xi)$ . Also, since this i/o partition allows an i/s/o representation with state dimension  $\mathbf{n}$ ,  $G(\xi)$  is proper and  $\deg(\det W_1(\xi)) = \mathbf{n}$ . From the i/s/o representation we have the transfer function as  $G(\xi) = C(\xi I_{\mathbf{n}} - A)^{-1}B + D$ . Now the second condition can be written as

$$\begin{aligned} \partial\Phi(j\omega) &> 0 \quad \forall \omega \in \mathbb{R} \\ \Rightarrow W_1^T(-j\omega)W_1(j\omega) - W_2^T(-j\omega)W_2(j\omega) &> 0 \quad \forall \omega \in \mathbb{R}. \end{aligned}$$

Since  $\deg(\det \partial\Phi(\xi)) = 2\mathbf{n}$  and  $\deg(\det W_1(\xi)) = \mathbf{n}$ , we infer the biproperness of the rational function  $\det[W_1^{-T}(-j\omega)[W_1^T(-j\omega)W_1(j\omega) - W_2^T(-j\omega)W_2(j\omega)]W_1^{-1}(j\omega)]$ . Using the biproperness, we can see that

$$\lim_{\omega \rightarrow \infty} (\det[W_1^{-T}(-j\omega)[W_1^T(-j\omega)W_1(j\omega) - W_2^T(-j\omega)W_2(j\omega)]W_1^{-1}(j\omega)]) \neq 0.$$

This fact together with  $W_1^T(-j\omega)W_1(j\omega) - W_2^T(-j\omega)W_2(j\omega) > 0$  for all  $\omega \in \mathbb{R}$  implies that

$$\lim_{\omega \rightarrow \infty} W_1^{-T}(-j\omega)[W_1^T(-j\omega)W_1(j\omega) - W_2^T(-j\omega)W_2(j\omega)]W_1^{-1}(j\omega) > 0.$$

The expression within the above limit is nothing but  $I_{\mathbf{m}} - G^T(-j\omega)G(j\omega)$ , the limit of which is  $I_{\mathbf{m}} - D^TD$ . This proves  $(I_{\mathbf{m}} - D^TD) > 0$ , and completes the proof of Lemma 5.3.  $\square$

The following lemma related to the strict dissipation inequality was used in the proof of Theorem 5.2 above. The proof is straightforward and so omitted (see [2]).

LEMMA A.2. *Let  $\mathfrak{B} \in \mathfrak{L}_{\text{cont}}^{\mathbf{w}}$ . Then  $\mathfrak{B}$  is strict  $\Sigma$ -dissipative if and only if  $\mathfrak{B}$  is dissipative with respect to  $\Sigma_1 := \begin{bmatrix} I_{\mathbf{m}} & 0 \\ 0 & -(1 + \epsilon_1)I_{\mathbf{p}} \end{bmatrix}$  for some  $\epsilon_1 > 0$ .*

The following proposition is regarding the standard Kalman decomposition (see [34]); we require it in the proof of Theorem 5.4, which follows after this proposition.

PROPOSITION A.3. *Let  $\mathfrak{B}$  be uncontrollable with an i/s/o representation  $\frac{d}{dt}x = Ax + Bw_1$ ,  $w_2 = Cx + Dw_1$ , where  $w = (w_1, w_2)$ . Then there exists a nonsingular matrix  $T \in \mathbb{R}^{\mathbf{n} \times \mathbf{n}}$  such that  $T^{-1}AT = \begin{bmatrix} A_c & A_{cp} \\ 0 & A_u \end{bmatrix}$ ,  $TB = \begin{bmatrix} B_c \\ 0 \end{bmatrix}$ , and  $CT^{-1} = [C_c \quad C_u]$ . Further,*

$$(A.6) \quad \frac{d}{dt}x_c = A_c x_c + B_c w_1, \quad w_2 = C_c x_c + D w_1$$

*gives an i/s/o representation for  $\mathfrak{B}_{\text{cont}}$ .*

*Proof of Theorem 5.4.* The last proposition enables us to consider the following i/s/o representation without loss of generality:

$$\begin{aligned} \frac{d}{dt} \begin{bmatrix} x_c \\ x_u \end{bmatrix} &= \begin{bmatrix} \tilde{A}_c & \tilde{A}_{cp} \\ 0 & \tilde{A}_u \end{bmatrix} \begin{bmatrix} x_c \\ x_u \end{bmatrix} + \begin{bmatrix} \tilde{B}_c \\ 0 \end{bmatrix} w_1 \\ w_2 &= \begin{bmatrix} \tilde{C}_c & \tilde{C}_u \end{bmatrix} \begin{bmatrix} x_c \\ x_u \end{bmatrix} + \tilde{D} w_1. \end{aligned}$$

Then the corresponding Hamiltonian matrix gets the following form

$$H = \begin{bmatrix} A_c & A_{cp} & B_c B_c^T & 0 \\ 0 & A_u & 0 & 0 \\ -C_c^T C_c & -C_c^T C_u & -A_c^T & 0 \\ -C_u^T C_c & -C_u^T C_u & -A_{cp}^T & -A_u^T \end{bmatrix},$$

where

$$\begin{aligned} A_c &= \tilde{A}_c + \tilde{B}_c \left( I_m - \tilde{D}^T \tilde{D}^T \right)^{-1} \tilde{D}^T \tilde{C}_c, & B_c &= \tilde{B}_c \left( I_m - \tilde{D}^T \tilde{D} \right)^{-\frac{1}{2}}, \\ A_{cp} &= \tilde{A}_{cp} + \tilde{B}_c \left( I_m - \tilde{D}^T \tilde{D}^T \right)^{-1} \tilde{D}^T \tilde{C}_u, & C_c &= \left( I_p - \tilde{D} \tilde{D}^T \right)^{-\frac{1}{2}} \tilde{C}_c, \\ A_u &= \tilde{A}_u, & C_u &= \left( I_p - \tilde{D} \tilde{D}^T \right)^{-\frac{1}{2}} \tilde{C}_u. \end{aligned}$$

Once again from Proposition A.3,  $\frac{d}{dt}x_c = \tilde{A}_c x_c + \tilde{B}_c w_1$ ,  $w_2 = \tilde{C}_c x_c + \tilde{D} w_1$  is an i/s/o representation of  $\mathfrak{B}_{\text{cont}}$ . So the corresponding Hamiltonian matrix for  $\mathfrak{B}_{\text{cont}}$  is given by  $H_c := \begin{bmatrix} \tilde{A}_c & \tilde{B}_c \tilde{C}_c^T \\ -\tilde{C}_c^T \tilde{C}_c & -\tilde{A}_c^T \end{bmatrix}$ . Now from Proposition 2.2 we can assume the i/s/o representation for  $\mathfrak{B}$  to be state observable, which implies that the controllable part is state observable; i.e., the  $(\tilde{C}_c, \tilde{A}_c)$  pair is observable. Again, since  $(\tilde{C}_c, \tilde{A}_c)$  is observable and  $\mathfrak{B}_{\text{cont}}$  is controllable, the i/s/o representation of  $\mathfrak{B}_{\text{cont}}$  is also state controllable; that is, the  $(\tilde{A}_c, \tilde{B}_c)$  pair is controllable. Therefore from Lemma 4.5, we get  $\text{spec}(H_c) = \text{roots}(\det \partial \Phi(\xi))$ . Again, from the Kalman-decomposed i/s/o representation of  $\mathfrak{B}$ , we get

$$\Lambda_{\text{un}} = \text{spec}(\tilde{A}_u) = \text{spec}(A_u).$$

Thus to prove Theorem 5.4, all we need to show is  $\det(\xi I_{2n} - H) = \det(\xi I - H_c) \det(\xi I - A_u) \det(\xi I + A_u)$ .

To show the above equality we shall find the determinant of the polynomial matrix  $(\xi I_{2n} - H)$  applying some elementary transformations on it as shown below. Let  $\mathbf{n}_c$  be the number of controllable eigenvalues and  $\mathbf{n}_u$  be that of uncontrollable eigenvalues.

$$\begin{aligned} \det(\xi I_{2n} - H) &= \det \left( \begin{bmatrix} I_{\mathbf{n}_c} & 0 & 0 & 0 \\ 0 & 0 & I_{\mathbf{n}_c} & 0 \\ 0 & I_{\mathbf{n}_u} & 0 & 0 \\ 0 & 0 & 0 & I_{\mathbf{n}_u} \end{bmatrix} (\xi I_{2n} - H) \begin{bmatrix} I_{\mathbf{n}_c} & 0 & 0 & 0 \\ 0 & 0 & I_{\mathbf{n}_u} & 0 \\ 0 & I_{\mathbf{n}_c} & 0 & 0 \\ 0 & 0 & 0 & I_{\mathbf{n}_u} \end{bmatrix} \right) \\ &= \det \begin{bmatrix} \xi I_{\mathbf{n}_c} - A_c & -B_c B_c^T & -A_{cp} & 0 \\ C_c^T C_c & \xi I_{\mathbf{n}_c} + A_c^T & C_c^T C_u & 0 \\ 0 & 0 & \xi I_{\mathbf{n}_u} - A_u & 0 \\ C_u^T C_c & A_{cp}^T & C_u^T C_u & \xi I_{\mathbf{n}_u} + A_u^T \end{bmatrix}. \end{aligned}$$

Now, exploiting the block structure of the above matrix, we can find out its determinant as

$$\begin{aligned} &\det \begin{bmatrix} \xi I_{\mathbf{n}_c} - A_c & -B_c B_c^T & -A_{cp} & 0 \\ C_c^T C_c & \xi I_{\mathbf{n}_c} + A_c^T & C_c^T C_u & 0 \\ 0 & 0 & \xi I_{\mathbf{n}_u} - A_u & 0 \\ C_u^T C_c & A_{cp}^T & C_u^T C_u & \xi I_{\mathbf{n}_u} + A_u^T \end{bmatrix} \\ &= \det(\xi I_{\mathbf{n}_u} + A_u^T) \det(\xi I_{\mathbf{n}_u} - A_u) \det \begin{bmatrix} \xi I_{\mathbf{n}_c} - A_c & -B_c B_c^T \\ C_c^T C_c & \xi I_{\mathbf{n}_c} + A_c^T \end{bmatrix} \\ &= \det(\xi I_{\mathbf{n}_u} + A_u^T) \det(\xi I_{\mathbf{n}_u} - A_u) \det(\xi I_{\mathbf{n}_c} - H_c) \\ (A.7) \quad &\Rightarrow \text{spec}(H) = \text{roots}[\chi_{\text{un}}(-\xi) \chi_{\text{un}}(\xi) \partial \Phi(\xi)], \end{aligned}$$

where the last equality follows from Lemma 4.5.  $\square$

The following well-known result about left and right eigenspaces of the Hamiltonian matrix was used in the proof of Theorem 5.5. This lemma can be proved by straightforward verification.

LEMMA A.4. *Let  $\Lambda$  be a set of eigenvalues of the Hamiltonian matrix  $H$ . If  $\text{im} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$  is the generalized right eigenspace of  $H$  corresponding to  $\Lambda$  with  $X_1, X_2 \in \mathbb{R}^{n \times \bullet}$ , then  $\text{im} \begin{bmatrix} X_2 \\ -X_1 \end{bmatrix}$  is the generalized left eigenspace of  $H$  corresponding to  $-\Lambda$ .*

The following lemma is used in the proof of 5.2. The result says that right-primeness is equivalent to the absence of any zeros in the complex plane for the corresponding Hermitian matrix. Statement 3 below makes a similar statement, but at the point  $\infty$ .

LEMMA A.5. *Consider  $M \in \mathbb{R}^{w \times n}[\xi]$  with  $M$  having full column rank. Suppose the behavior having the image representation  $w = M(\frac{d}{dt})\ell$  has McMillan degree  $n$ . Then, the following are equivalent.*

1.  $M(\xi)$  is right-prime,
2.  $M^T(\bar{\lambda})M(\lambda) > 0$  for all  $\lambda \in \mathbb{C}$ , and
3.  $\deg [\det (M^T(-\xi)M(\xi))] = 2n$ .

*Proof.* We shall prove the following chain of implications:  $1 \Rightarrow 2$ ,  $2 \Rightarrow 1$ ,  $1 \Rightarrow 3$  and  $3 \Rightarrow 1$ .

(1  $\Rightarrow$  2) We assume right-primeness of  $M$  and show that  $M^T(\bar{\lambda})M(\lambda)$  is positive definite for all  $\lambda \in \mathbb{C}$ . That  $M^T(\bar{\lambda})M(\lambda) \geq 0$  for all  $\lambda \in \mathbb{C}$  is obvious. Due to the right-primeness,  $M(\lambda)$  has full column rank for all  $\lambda \in \mathbb{C}$ , and this implies the required positive definiteness for all  $\lambda$ .

(2  $\Rightarrow$  1)  $M^T(\bar{\lambda})M(\lambda) > 0$  for all  $\lambda \in \mathbb{C}$  means that  $M(\lambda)$  is injective for all  $\lambda$  proving its full column rank property for all  $\lambda$ , and hence right-primeness.

(1  $\Rightarrow$  3)  $M(\xi)$  can be partitioned (after a permutation of rows, if needed) into  $\text{col}(W_1(\xi), W_2(\xi))$  such that  $W_2W_1^{-1}$  is a proper rational matrix, and  $W_1(\xi)$  has determinant of degree  $n$ . Suppose  $G(\xi) := W_2(\xi)W_1(\xi)^{-1}$ . (See [19] for McMillan degree's relation to an *observable* image representation, and observability of the image representation is equivalent to right-primeness of the matrix  $M$ .) Consider  $\det (M^T(-\xi)M(\xi))$  which equals  $\det (W_1^T(-\xi)W_1(\xi) + W_2^T(-\xi)W_2(\xi))$

$$\begin{aligned} &= \det (W_1(-\xi))\det (W_1(\xi))\det \left( I + (W_2(-\xi)W_1(-\xi)^{-1})^T (W_2(\xi)W_1(\xi)^{-1}) \right) \\ &= \det (W_1(-\xi))\det (W_1(\xi))\det (I + G^T(-\xi)G(\xi)). \end{aligned}$$

In order to determine the degree of  $\det (M^T(-\xi)M(\xi))$ , we let  $\xi \rightarrow \infty$  to get rid of the strictly proper part within the last term above:  $\lim_{\xi \rightarrow \infty} \det (I + G(-\xi)^T G(\xi)) = a$  (say). Notice that  $0 < a < \infty$ ;  $a < \infty$  because of the properness of  $W_2W_1^{-1} = G$ , while due to the positive definiteness of  $(I + \lim_{\xi \rightarrow \infty} G^T(-\xi)G(\xi))$  we obtain that its determinant  $a > 0$ . Thus the degree of  $\det (M(-\xi)^T M(\xi))$  is twice the degree of  $\det W_1$ , and is thus  $2n$ .

(3  $\Rightarrow$  1) In order to prove this, we assume  $M(\xi)$  is not right-prime and arrive at the required contradiction. Non-right-primeness of  $M$  means that  $M$  can be factored into  $M = \bar{M}F$  such that  $\bar{M}$  is right-prime, and  $F$  has a nonzero and non-constant polynomial as its determinant. This implies that  $w = \bar{M}(\frac{d}{dt})\ell$  is an observable kernel representation for  $\mathfrak{B}$  and hence  $\bar{M}$  can be partitioned (after possibly a permutation of rows) into  $\bar{M} = \text{col}(W_1, W_2)$  such that degree of  $\det W_1 = n$ . Notice that  $\det (M^T(-\xi)M(\xi)) = \det F(-\xi) \det F(\xi) \det (\bar{M}(-\xi)^T \bar{M}(\xi))$ . We now use the proof of (1  $\Rightarrow$  3) above and that  $\bar{M}$  is right-prime to conclude that the degree of  $\det (\bar{M}(-\xi)^T \bar{M}(\xi))$  is  $2n$ . Hence degree of  $\det M^T(-\xi)M(\xi)$  equals



$2n + 2(\deg(\det F))$ . Since  $F$  has determinant a nonzero, nonconstant polynomial, we obtain  $\det M^T(-\xi)M(\xi) > 2n$ , thus obtaining the required contradiction. This proves  $(3 \Rightarrow 1)$  and also the lemma.  $\square$

**Acknowledgments.** During the course of this work, we had several useful discussions with Dr. Harish K. Pillai; we thank him for his insightful input. We also thank Prof. Jan C. Willems for his constructive suggestions.

## REFERENCES

- [1] M. N. BELUR, H. K. PILLAI, AND H. L. TRENTelman, *Dissipative systems synthesis, a linear algebraic approach*, Linear Algebra Appl., 425 (2007), pp. 739–756.
- [2] M. N. BELUR AND H. L. TRENTelman, *The strict dissipativity synthesis problem and the rank of the coupling qdf*, Systems Control Lett., 51 (2004), pp. 247–258.
- [3] R. BOTT AND R. J. DUFFIN, *Impedance synthesis without transformers*, J. Appl. Phys., 20 (1949), p. 816.
- [4] S. BOYD, L. E. GHAOUI, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.
- [5] M. K. ÇAMLIBEL, J. C. WILLEMS, AND M. N. BELUR, *On the dissipativity of uncontrollable systems*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Hawaii, December 2003.
- [6] M. Z. CHEN AND M. C. SMITH, *Electrical and mechanical passive network synthesis*, in Recent Advances in Learning and Control, V. D. Blondel, S. P. Boyd, and H. Kimura, eds., Springer-Verlag, New York, 2008, Ch. 3, pp. 35–50.
- [7] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State-space solutions to standard  $\mathcal{H}_2$  and  $\mathcal{H}_\infty$  control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.
- [8] L. E. FAIBUSOVICH, *Matrix Riccati inequality: Existence of solutions*, Sys. Control Lett., 9 (1987), pp. 59–64.
- [9] A. FERRANTE AND L. PANDOLFI, *On the solvability of the positive real lemma equations*, Sys. Control Lett., 47 (2002), pp. 211–219.
- [10] B. A. FRANCIS, *A Course in  $\mathcal{H}_\infty$  Control Theory*, Springer-Verlag, New York, 1987.
- [11] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1984.
- [12] C. LIN, *Structural controllability*, IEEE Trans. Automat. Control, 19 (1974), pp. 201–208.
- [13] G. MEINSMA, *J-spectral factorization and equalizing vectors*, Sys. Control Lett., 25 (1995), pp. 243–249.
- [14] K. MUROTA, *Systems Analysis by Graphs and Matroids: Structural Solvability and Controllability*, Springer-Verlag, Berlin, 1987.
- [15] D. PAL, *Optimal control: Problems at optimality*, Master’s thesis, IIT Bombay, 2007.
- [16] H. K. PILLAI AND S. SHANKAR, *A behavioural approach to control of distributed systems*, SIAM J. Control Optim., 37 (1998), pp. 388–408.
- [17] J. W. POLDERMAN AND J. C. WILLEMS, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Springer-Verlag, New York, 1998.
- [18] A. C. M. RAN AND L. RODMAN, *Factorization of matrix polynomials with symmetries*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 845–864.
- [19] P. RAPISARDA AND J. C. WILLEMS, *State maps for linear systems*, SIAM J. Control Optim., 35 (1997), pp. 1053–1091.
- [20] P. RAPISARDA, *Linear Differential Systems*, Ph.D. thesis, University of Groningen, The Netherlands, 1998.
- [21] C. W. SCHERER, P. GAHINET, AND M. CHILALI, *Multi-objective output-feedback control via lmi optimization*, IEEE Trans. Automat. Control, 42 (1997), pp. 896–911.
- [22] C. W. SCHERER, *The Riccati inequality and state-space  $\mathcal{H}_\infty$ -optimal control*, Ph.D. thesis, University of Würzburg, Germany, 1990.
- [23] C. W. SCHERER, *The state-feedback  $\mathcal{H}_\infty$ -problem at optimality*, Automat., 30 (1994), pp. 293–305.
- [24] C. W. SCHERER, *A complete algebraic solvability test for the nonstrict lyapunov inequality*, Sys. Control Lett., 25 (1995), pp. 327–335.
- [25] M. SEBEK AND H. KWAKERNAAK, *J-spectral factorization*, in Proceedings of the 30th IEEE Conference on Decision and Control, Brighton, UK, December 1991, pp. 1278–1283.
- [26] J. SYLVESTER, *Sur l’équation en matrices  $px = xq$* , Comptes Rendus, 99 (1884), pp. 67–71.

- [27] H. L. TRENTELMAN AND J. C. WILLEMS, *Every storage function is a state function*, Systems Control Lett., 32 (1997), pp. 249–259.
- [28] J. C. WILLEMS AND H. L. TRENTELMAN, *On quadratic differential forms*, SIAM J. Control Optim., 36 (1998), pp. 1703–1749.
- [29] J. C. WILLEMS AND H. L. TRENTELMAN, *Synthesis of dissipative systems using quadratic differential forms: Parts I and II*, IEEE Trans. Automat. Control, 47 (2002), pp. 53–86.
- [30] J. C. WILLEMS, *Dissipative dynamical systems - Part I: General theory, Part II: Linear systems with quadratic supply rates*, Archive for Rational Mechanics and Analysis, 45 (1972), pp. 321–351, 352–393.
- [31] J. C. WILLEMS, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.
- [32] J. C. WILLEMS, *On interconnections, control and feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 326–339.
- [33] J. C. WILLEMS, *Dissipative dynamical systems*, Eur. J. Control, 13 (2007), pp. 134–151.
- [34] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.

## A BEHAVIORAL APPROACH TO PLAY IN MECHANICAL NETWORKS\*

FRANK SCHEIBE<sup>†</sup> AND MALCOLM C. SMITH<sup>†</sup>

**Abstract.** This paper shows that the treatment of play or backlash as an input-output operator in mechanical networks leads to solutions which are unsatisfactory from a physical point of view. This contrasts with a simple behavioral definition of ideal play. With this definition, play cannot be treated in isolation as an input-output relation. As a result, methods of nonlinear feedback systems are not readily applicable to establish well-posedness of solutions of mechanical networks incorporating this play element. By means of simple network examples, this paper explores the issues involved in establishing well-posedness of mechanical networks incorporating springs, dampers, masses, and inerters together with the behavioral model of ideal play. Connections with the approach of Nordin, Galic, and Gutman are analyzed and a model implementation given.

**Key words.** mechanical networks, backlash, behavior, well-posedness, inerter

**AMS subject classifications.** 70K99, 70Q05, 93C05

**DOI.** 10.1137/070704605

**1. Introduction.** This paper is concerned with the mathematical modeling of mechanical networks made up of standard linear elements and a single nonlinear element, play. There are two main themes in the paper. The first is the question of how well the mathematical models match the behavior of actual physical devices. The second is a purely mathematical question of the well-posedness of interconnections of the linear dynamical elements with play. Our approach will exhibit a close connection between these themes.

A number of different models for play have been proposed in the literature. Figure 1 shows two different definitions, the dead-zone model and the hysteresis model. These have been justified by the expected behavior of a clearance in series with a spring and damper, respectively; see [3, p. 122] and [7, p. 68]. The hysteresis model is commonly used as a basis for a formal mathematical approach to play [11, 1]. Both definitions aim to describe an apparently well-defined phenomena and give rise to two different mathematical descriptions. This raises the question of which model, or indeed if either, is more satisfactory? The hysteresis model will be considered in detail in section 2 in its formal mathematical definition as the “play operator.” We will argue that this definition leads to behavior in mechanical networks which appears unrealistic from a physical point of view. A similar point can be made with respect to the dead-zone model (cf. [13]).

In section 3 we propose a formal definition of play, in which only the sign of the force at the extremes of relative displacement is specified, and the force is zero between the extremes. This model can be thought of as a behavioral model of play in the sense of Willems [20], since it does not admit an input-output definition in isolation. This appears to be the simplest possible definition which avoids the objection raised in section 2. We note that this definition does not seek to model the contact mechanics which might be relevant at engagement and disengagement.

---

\*Received by the editors October 5, 2007; accepted for publication (in revised form) July 21, 2008; published electronically December 5, 2008.

<http://www.siam.org/journals/sicon/47-6/70460.html>

<sup>†</sup>Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK (frank.scheibe@cantab.net, mcs@eng.cam.ac.uk).

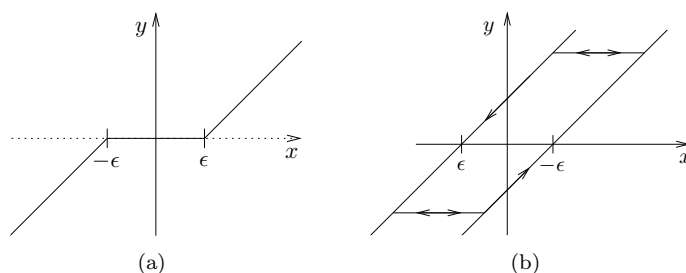


FIG. 1. (a) Graph of dead-zone play model. (b) Graph of hysteresis play model.

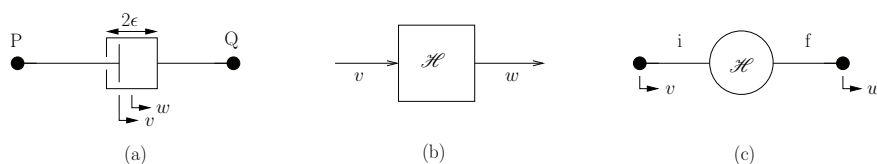


FIG. 2. The play operator. (a) Physical representation. (b) Input-output modeling symbol. (c) Terminal modeling symbol.

Section 3 goes on to examine the question of well-posedness of mechanical networks incorporating this play element. So far it has not been possible to develop a general framework for well-posedness with this play element, such as the linear complementarity framework for a class of hybrid systems developed in [18, 12, 9, 2]. Instead, we will examine a number of typical cases from first principles. We will encounter situations in which not only unique solutions are obtained, but also networks which admit multiple solutions. In the latter case we are able to illuminate this nonuniqueness by energy considerations at transitions. A further perspective is provided by the incorporation of compliance and buffer networks and the study of their limiting behavior. These networks are strongly suggestive of models proposed for impact mechanics.

Section 4 considers a semi-ideal model for play which consists of a parallel spring-damper buffer in series with our behavioral model of ideal play. We point out that the definition coincides almost always with a model of Nordin, Galic, and Gutman [13]. The latter model is expressed as a dynamical system in input-output form. A numerical implementation of this semi-ideal play is described.

The paper is structured as follows. Section 2 is a critique of the play operator. Section 3 describes the behavioral definition of play and analyzes the well-posedness of a simple network. Sections 3.3–3.6 examine the well-posedness of a network with play in series with an inerter. Section 4 analyzes a semi-ideal play model and its relation to a model of Nordin, Galic, and Gutman and presents a numerical implementation. Concluding remarks are given in section 5.

**2. The play operator: A critique.** A standard treatment of play between mechanical elements makes use of the play operator (hysteron). The formal definition of the operator is illustrated in Figure 2(a). The position of the piston ( $v$ ) is considered to be the input, and the position of the cylinder ( $w$ ) is considered to be the output (follower). The behavior of the play operator is then characterized by the graph of Figure 1(b). Within the play region  $|v - w| < \epsilon$  the follower  $Q$  remains stationary. Otherwise, it follows  $P$  with an offset of  $\pm\epsilon$ . The input-output behavior of this model

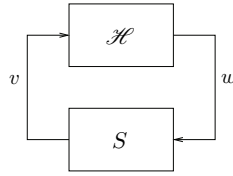


FIG. 3. Input-output representation of play in feedback arrangement.

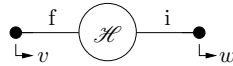


FIG. 4. Terminal modeling symbol with input and follower reversed.

can be defined formally for piecewise monotonic (continuous) inputs (see [1, pp. 24–25] and [11, pp. 6–8]). The model can then be extended to inputs which are continuous functions by a limiting argument (see [1, p. 42] and [11, pp. 14–15]). This defines the “play operator.” Two alternative modeling symbols are used to represent the play operator, depending on the need for an input-output or a terminal representation (see Figure 2(b),(c)).

When the play element is connected to a network with dynamical elements, the well-posedness of the dynamical equations can be analyzed as shown in Figure 3.

**PROPOSITION 2.1.** *Let  $S$  be an linear time-invariant dynamical system with transfer function  $G(s)$ . If  $G(s)$  is proper and  $|G(\infty)| < 1$ , then the dynamical system of Figure 3 is well-posed.*

*Proof.* This relies on the notion of instantaneous gain in a feedback loop and makes use of the contraction mapping theorem in Banach spaces. General results of this type can be found in [19, sec. 4.3.3] and [4, p. 48].  $\square$

The behavior of the play operator as shown in Figure 1(b) can also be expressed as a condition on three hybrid states as follows:

G1 (engagement–extension):  $w = v + \epsilon$ ,  $\dot{v} = \dot{w} \leq 0$ .

G2 (engagement–compression):  $w = v - \epsilon$ ,  $\dot{v} = \dot{w} \geq 0$ .

G3 (disengagement):  $|v - w| < \epsilon$ ,  $\dot{w} = 0$ .

If the terminals in Figure 2(a) are reversed so that  $Q$  is the input and  $P$  is the follower as represented in Figure 4, then the three hybrid states become the following:

H1 (engagement–extension):  $v = w - \epsilon$ ,  $\dot{v} = \dot{w} \geq 0$ .

H2 (engagement–compression):  $v = w + \epsilon$ ,  $\dot{v} = \dot{w} \leq 0$ .

H3 (disengagement):  $|v - w| < \epsilon$ ,  $\dot{v} = 0$ .

**2.1. A network example.** This section analyzes the behavior of the mechanical network illustrated in Figure 5. The dynamical equations are

$$(2.1a) \quad m_2 \ddot{y} = k(z - y) + c(\dot{z} - \dot{y}),$$

$$(2.1b) \quad m_1 \ddot{z} = -k(z - y) - c(\dot{z} - \dot{y}).$$

Eliminating  $z$ , we can then calculate the transfer function

$$\frac{\hat{y}(s)}{\hat{u}(s)} = -\frac{m_1 cs}{m_1 m_2 s^2 + m_2 cs + k(m_1 + m_2)},$$

which is strictly proper. Hence, from Proposition 2.1 the network in Figure 5 is well-posed. Next, the dynamical equations in each of the states H1–H3 are evaluated explicitly, assuming the constants to be  $m_1 = m_2 = 1$  kg,  $c = 2$  Nsm<sup>−1</sup>, and  $k = 5/2$  Nm<sup>−1</sup>.

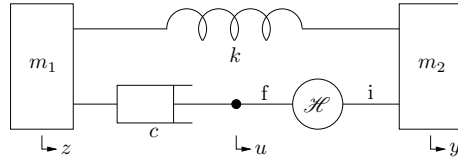


FIG. 5. Damped harmonic oscillator network.

H1 (engagement–extension). This corresponds to  $u = y - \epsilon$  and  $\dot{u} = \dot{y} \geq 0$ . The dynamical equations can be written in the form  $\dot{x}(t) = A_{H1}x(t)$  with  $x = [y, \dot{y}, z, \dot{z}]^T$ . Choosing the initial conditions

$$(2.2) \quad x(0) = [A, B - 2A, -A, -B + 2A]^T + [C, D, C, D]^T,$$

we can calculate  $x(t) = e^{A_{H1}t}x(0)$  to find

$$(2.3a) \quad y(t) - z(t) = 2e^{-2t}(A \cos(t) + B \sin(t)),$$

$$(2.3b) \quad y(t) + z(t) = 2(C + Dt).$$

H2 (engagement–compression). This corresponds to  $u = y + \epsilon$  and  $\dot{u} = \dot{y} \leq 0$ . The equations for  $x(t)$ ,  $y(t)$ , and  $z(t)$  are identical to those in case H1.

H3 (disengagement). This corresponds to  $|y - u| < \epsilon$  and  $\dot{u} = 0$ . The dynamical equations can be written in the form  $\dot{x}(t) = A_{H3}x(t)$  with  $x = [y, \dot{y}, z, \dot{z}]^T$ . In the following, only the solution in the special case  $x(0) = [A, 0, -A, 0]$  is needed. By explicit computation we can find  $x(t) = e^{A_{H3}t}x(0)$ , which gives

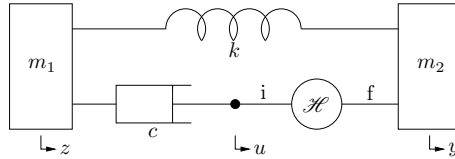
$$(2.4) \quad y(t) = -z(t) = 1/2e^{-t}A(2 \cos(2t) + \sin(2t)).$$

**2.1.1. The force through the play element.** Here, a particular solution for the network of section 2.1 is calculated, which involves a transition from state H1 to state H3. Consider an initial condition of the form (2.2) in which  $A = C = D = 0$  and  $B > 0$ . From (2.3a) and (2.3b) we find that  $y(t) = e^{-2t}B \sin(t)$ . It follows that a transition from state H1 to state H3 must occur at the first time  $t_1 = \arctan(1/2)$  for which  $\dot{y}(t_1) = 0$ , since otherwise  $\dot{y}$  becomes negative in violation of the conditions for state H1. The corresponding displacement at disengagement is  $y(t_1) = B \exp(-2 \arctan(1/2))/\sqrt{5}$ . For simplicity we choose  $B$  so that  $y(t_1) = 1$ , and from (2.4) we find that for  $t > t_1$ ,  $y(t - t_1) = e^{-(t-t_1)}(2 \cos(2(t - t_1)) + \sin(2(t - t_1)))/2$ . It is straightforward to see that this solution is consistent with a transition to state H3, namely,  $|y(t) - u(t)| < \epsilon$  for  $t - t_1$  sufficiently small, since  $u(t) = 1 - \epsilon$  for  $t > t_1$  in state H3. The solution can be continued forward in time until the next transition occurs to state H2 when  $y(t_2) = 1 - 2\epsilon$ . For  $t > t_1$  and in state H3, the force exerted by the damper is equal to

$$(2.5) \quad cz(t - t_1) = 5e^{-(t-t_1)} \sin(2(t - t_1))/2.$$

This is also equal to the force through the play element.

Equation (2.5) highlights our first observation about the play operator in mechanical networks that appears unsatisfactory; that is, during the disengagement state H3 the force transmitted by the play element is not necessarily zero.

FIG. 6. Figure 5 with reversed terminals of  $\mathcal{H}$ .

**2.1.2. Dependence on inertial frame.** Next, the particular solution of section 2.1.1 is considered, but with a steady “drift” term added. Setting  $A = C = 0$  we find that  $y(t) = e^{-2t}B \sin(t) + Dt$ . This, and the corresponding solution for  $z(t)$ , differs only from that in section 2.1.1 by the addition of a constant velocity  $D$ . As before, a transition from state H1 to H3 occurs when  $\dot{y}(t) = 0$ . But it is easy to see that  $\dot{y}(t) > 0$  for all  $t \geq 0$  if  $D$  is sufficiently large. In such a case there will be no transition to state H3. Even if a transition occurs, the transition time will be dependent on  $D$ . Evidently, the behavior of the system fails to be invariant to a simple translation of the inertial frame. This is the second property of the play operator that appears unsatisfactory from a physical point of view. It is also curious that the output of the play operator remains stationary during disengagement.

For this solution we now consider the force through the play element, which is the same as the force through the damper, which equals

$$c(\dot{z} - \dot{y}) = -4e^{-2t}B(\cos(t) - 2\sin(t)).$$

Clearly the sign of this force oscillates. This again appears unsatisfactory from a physical point of view, since we would expect the force acting on the play element to be positive in the extension state.

**2.1.3. Reversal of terminals of the play operator.** The network in Figure 5 is considered next, but with the terminals of the play operator reversed (see Figure 6), and with the same parameters as in section 2.1. The dynamical equations are again given by (2.1), and on eliminating  $z$  we find that

$$\frac{\hat{u}(s)}{\hat{y}(s)} = -\frac{m_1 m_2 s^2 + m_2 cs + k(m_1 + m_2)}{m_1 cs}.$$

Since this transfer function is nonproper, it is not possible to use Proposition 2.1 and give a general statement on well-posedness. However, we show below that solutions can be computed in specific cases. The dynamical equations in each of the states G1–G3 are as follows:

G1 (engagement–extension). This corresponds to  $y = u + \epsilon$ ,  $\dot{u} = \dot{y} \leq 0$ . The equations of  $u(t)$ ,  $y(t)$ , and  $z(t)$  are identical to those in case H1 in section 2.1.

G2 (engagement–compression). This corresponds to  $y = u - \epsilon$ ,  $\dot{u} = \dot{y} \geq 0$ . The equations for  $u(t)$ ,  $y(t)$ , and  $z(t)$  are identical to those in case H2 in section 2.1.

G3 (disengagement). This corresponds to  $|y - u| < \epsilon$ ,  $\dot{y} = 0$ . In this state  $y(t) \equiv y^*$  (say) and the dynamical equations become

$$(2.6a) \quad 0 = k(z - y^*) + c(\dot{z} - \dot{u}),$$

$$(2.6b) \quad m_1 \ddot{z} = -k(z - y^*) - c(\dot{z} - \dot{u}).$$

Adding both equations gives  $0 = m_1 \ddot{z}$ , which implies  $z(t) = E_1 + E_2 t$ . From (2.6a) we can find an expression of  $\dot{u}(t)$ , which leads to

$$(2.7) \quad u(t) = E_3 + (E_2 + (k/c)(E_1 - y^*))t + kt^2 E_2 / (2c),$$

where  $E_1$ ,  $E_2$ , and  $E_3$  are constants.

Now consider an initial condition of the form (2.2) in which  $A = C = D = 0$  and  $B > 0$  for the network of section 2.1.3, and a transition from G2 to G3. Again from (2.3a) and (2.3b) we find that a transition to G3 occurs when  $t_1 = \arctan(1/2)$ . Once again  $B$  is chosen so that  $y(t_1) = 1$  (which is the same value as obtained in section 2.1.1). In order that a solution for the dynamical equations in (2.1) exists through  $t_1$ , it is necessary that  $y(t)$ ,  $z(t)$ ,  $\dot{y}(t)$ , and  $\dot{z}(t)$  are continuous at  $t = t_1$ . Therefore, a solution to the equations in the G3 state is sought with the following initial conditions:  $y(t_1) = -z(t_1) = 1$ ,  $\dot{y}(t_1) = \dot{z}(t_1) = 0$ . It follows that  $y(t) \equiv 1$ , while the G3 state persists. Since  $\ddot{y}(t) + \ddot{z}(t) = 0$  we must also have  $z(t) \equiv -1$ , while the G3 state persists. From (2.7) we find that  $u(t) = 1 + \epsilon - 5t/2$ . Since  $y - u = -\epsilon + 5t/2$ , the solution is consistent with a transition to G3 at  $t = t_1$ .

It is interesting to make a comparison between this solution and the one in section 2.1.1 without the terminals reversed. Dynamically, the two solutions are identical for  $0 < t < t_1$ , except for the fact that the play element is in compression rather than extension. During disengagement the two solutions differ. With the choice of terminals used in this section, the solution during disengagement appears curious in that the play operator results in both masses being stationary, while the input terminal of the play operator moves with constant velocity.

**2.2. Summary of critique.** For a simple mechanical network incorporating the play operator in series with a damper, we have identified several properties of the network behavior which appear unsatisfactory from a physical point of view. These are summarized as follows:

1. During disengagement the force through the play element is not necessarily zero.
2. The solutions of the network equations depend on the choice of inertial frame, namely, the addition of a constant velocity to all states may change switching times or eliminate them altogether.
3. During engagement the force through the play element is not restricted in sign.
4. The behavior of the network is not invariant to a switch of terminals of the play operator.

A similar critique can be given for the dead-zone model of play. (See [13], where some similar points are made for the dead-zone model.)

**3. The ideal play.** This section proposes a definition of ideal play which does not suffer from the criticism identified in section 2. Since the ideal play does not admit an input-output graph, mathematical properties like well-posedness and the exclusion of limit points of switching are arrived at by analyzing individual transition scenarios.

**3.1. A behavioral definition of the ideal play.** Consider a physical representation of play as shown in Figure 7(a) where  $z_1, z_2$  are the terminal positions and  $F$  is the equal and opposite force applied at the terminals. The ideal play is defined to be completely characterized by the following three states:

- I1 (engagement–extension):  $z_2 - z_1 = \epsilon$ ,  $F \leq 0$ .
- I2 (engagement–compression):  $z_2 - z_1 = -\epsilon$ ,  $F \geq 0$ .
- I3 (disengagement):  $|z_2 - z_1| < \epsilon$ ,  $F = 0$ .

Note that the definition is invariant to terminal reversal (i.e., the transformation  $z_1 \rightarrow -z_2$ ,  $z_2 \rightarrow -z_1$ ,  $F \rightarrow F$ ). Also, objections 1 and 3 in section 2.2 no longer apply. Finally, we note that this definition allows the mechanical network to maintain invariance to the choice of inertial frame, since the three states depend only on the



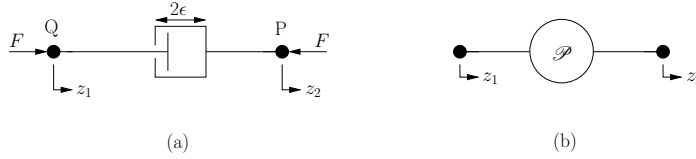


FIG. 7. (a) *Physical representation of ideal play.* (b) *Terminal modeling symbol for ideal play.*

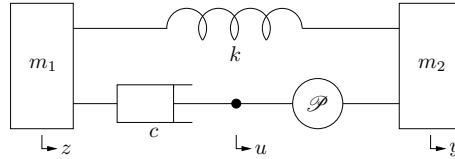


FIG. 8. *Damped harmonic oscillator network with ideal play.*

difference between  $z_1$  and  $z_2$ . We will use the network symbol shown in Figure 7(b) to represent the ideal play.

It is important to remark that the ideal play cannot be represented as a function or input-output operator since  $F$  does not determine  $z_2 - z_1$  uniquely—nor does  $z_2 - z_1$  determine  $F$  uniquely. Fundamentally, the ideal play will need to be connected to other elements in order that the complete network has a unique solution for given initial conditions.

The issue of well-posedness for networks incorporating the ideal play will be explored in the rest of this section. A number of typical cases will be examined from first principles. So far it has not been possible to derive general conditions for well-posedness by, for example, exploiting general approaches to hybrid or switched systems, e.g., the linear complementarity approach of [18, 12, 9, 2]. The use of such methods to consider well-posedness with ideal play is a topic for future research.

**3.2. A network incorporating ideal play.** Consider the network in Figure 8, which is the same as the networks considered in Figures 5 and 6 but with the new ideal model of play. The dynamical equations are

$$(3.1a) \quad m_2 \ddot{y} = k(z - y) + F,$$

$$(3.1b) \quad m_1 \ddot{z} = -k(z - y) - F,$$

$$(3.1c) \quad F = c(\dot{z} - \dot{u}).$$

We will now write down the form of the solutions in each state.

I1 (engagement–extension). This corresponds to  $y - u = \epsilon$  and  $F \leq 0$ . The dynamical equations for  $x = [y, \dot{y}, z, \dot{z}]$  are identical to those in state H1 in section 2.1.

I2 (engagement–compression). This corresponds to  $y - u = -\epsilon$  and  $F \geq 0$ . The dynamical equations are identical to those in state H2 in section 2.1.

I3 (disengagement). This corresponds to  $|y - u| < \epsilon$  and  $F = 0$ . It can be seen directly from (3.1) that

$$(3.2a) \quad y(t) - z(t) = A \cos(\omega_3 t) + B \sin(\omega_3 t),$$

$$(3.2b) \quad m_2 y(t) + m_1 z(t) = C + Dt,$$

where  $\omega_3 = \sqrt{\frac{k(m_1 + m_2)}{m_1 m_2}}$ .

We immediately see that the solution of the network of Figure 8 differs from the networks of both Figure 5 and Figure 6 as is seen from the purely oscillatory form of the dynamical equations in the disengagement region.

**3.2.1. A special well-posedness argument.** We now establish a well-posedness property of the network in Figure 8. The approach taken is to examine in detail all possible transition scenarios and to show that there is a well-defined transition in all cases. The proposition describes only the transitions between I1 and I3—the transitions between I2 and I3 are analogous. The terminology “just before  $t_0$ ” means formally “for  $t < t_0$  and  $|t_0 - t|$  sufficiently small,” and a similar meaning holds for “just after  $t_0$ ”.

**PROPOSITION 3.1.** *Consider the network of Figure 8 in which the displacements  $y(t)$ ,  $z(t)$ ,  $u(t)$  are assumed to be continuous.*

(a) *For any solutions of the network,  $\dot{y}(t)$  and  $\dot{z}(t)$  are always continuous, including transitions between states; however,  $\dot{u}(t)$  need not be continuous at transitions.*

(b) *Suppose the system is in state I1 just before  $t_0$  and that  $F(t) \uparrow 0$  as  $t \uparrow t_0$ . Then there is a unique continuation of the system into either state I1 or I3 after  $t_0$ .*

(c) *Suppose the system is in I3 just before  $t_0$  and  $y - u \uparrow \epsilon$ . If  $\dot{y}(t_0) - \dot{u}(t_0^-) > 0$ , then there is a unique continuation of the system into state I1 after  $t_0$ . If  $\dot{y}(t_0) - \dot{u}(t_0^-) = 0$ , then there is a unique continuation of the system into either I1 or I3 after  $t_0$ .*

(d) *Given any initial condition at time  $t = 0$ , the solutions of the network exist,  $y(t) - z(t)$  is uniformly bounded on  $[0, \infty)$ , and there are no limit points of switching between the states I1, I2, I3.*

*Proof.* (a) Suppose that there is a discontinuity in either  $\dot{y}$  or  $\dot{z}$ . By direct observation of the differential equations (3.1a)–(3.1b) this implies a  $\delta$ -function in  $\ddot{y}$  or  $\ddot{z}$ , which means that  $F$  must provide an impulsive force. From (3.1c) this can happen only if the displacement  $u(t)$  is discontinuous, which is excluded. Therefore,  $\dot{y}(t)$ ,  $\dot{z}(t)$  are always continuous, including transitions between states, whereas  $\dot{u}(t)$  need not be continuous at transitions.

(b) Suppose the network is in state I1 just before time  $t_0$  and that  $\dot{z}(t) - \dot{u}(t) \uparrow 0$  as  $t \uparrow t_0$ . Since  $y(t) - u(t) = \epsilon$  just before  $t_0$ , we have the following conditions:  $y(t_0) - u(t_0) = \epsilon$ ,  $\dot{y}(t_0) = \dot{u}(t_0^-) = \dot{z}(t_0)$ . If the system remains in state I1, then  $\dot{u}(t_0^+) = \dot{y}(t_0)$ , whereas if there is a transition to I3, then  $\dot{u}(t_0^+) = \dot{z}(t_0)$ . Thus, in either case  $\dot{u}(t_0^-) = \dot{u}(t_0^+)$ . Let us therefore consider an “initial” condition  $x(t_0) = [Y, D, Z, D]^T$ . If the network were to continue in the state I1, then we could calculate that

$$(3.3) \quad \dot{z} - \dot{u} = \dot{z} - \dot{y} = -4(Z - Y) \frac{\omega_3}{\omega_1} \exp\left(-\frac{c(m_1 + m_2)}{2m_1 m_2}(t - t_0)\right) \sin(\omega_1(t - t_0)),$$

where  $\omega_1 = \sqrt{\frac{(m_1 + m_2)(4km_1 m_2 - c^2(m_1 + m_2))}{4m_1^2 m_2^2}}$ , which is consistent with  $F \leq 0$  if and only if  $Z - Y \geq 0$ . If the network were to continue in state I3, we could calculate that  $\dot{y} - \dot{u} = \dot{y} - \dot{z} = \omega_3(Z - Y) \sin(\omega_3(t - t_0))$ , which is consistent with  $y - u < \epsilon$  after  $t_0$  if and only if  $Z - Y < 0$ . Thus, there is always an unambiguous continuation into I1 or I3.

(c) Suppose the network is in state I3 just before time  $t_0$  and that  $y(t) - u(t) \uparrow \epsilon$  as  $t \uparrow t_0$ . It follows that  $\dot{y}(t_0) - \dot{u}(t_0^-) \geq 0$  since  $y - u$  tends to  $\epsilon$  from below. Let us suppose that  $\dot{y}(t_0) - \dot{u}(t_0^-) > 0$ . The network cannot remain in state I3 after time  $t_0$  since the condition  $|y(t) - u(t)| < \epsilon$  would be violated. (Note that if the network remains in I3,  $\dot{z}(t_0^-) - \dot{u}(t_0^-) = \dot{z}(t_0^+) - \dot{u}(t_0^+) = 0$ , so that  $\dot{u}(t_0^+) = \dot{u}(t_0^-)$ .) For a transition to I1 we need the condition  $\dot{u}(t_0^+) = \dot{y}(t_0)$ , whereas  $\dot{u}(t_0^-) = \dot{z}(t_0)$ , which suggests a discontinuity in  $\dot{u}$  at  $t = t_0$ . For a valid transition to I1 it is necessary that  $F \leq 0$  after  $t_0$ ,

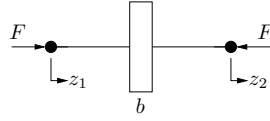


FIG. 9. (Ideal) inerter modeling symbol.

which must hold since  $\dot{z}(t_0) - \dot{u}(t_0^+) = \dot{u}(t_0^-) - \dot{y}(t_0) < 0$ . The only remaining case is when  $\dot{y}(t_0) - \dot{u}(t_0^-) = 0$ . Since  $\dot{u}(t_0^-) = \dot{z}(t_0)$  we expect  $\dot{u}(t)$  to be continuous at  $t = t_0$  irrespective of any transition to I1 or I3. This case therefore reduces to the precise situation analyzed in (b).

(d) Within I1 or I2,

$$(3.4) \quad (\ddot{y} - \ddot{z}) + (c/m)(\dot{y} - \dot{z}) + (k/m)(y - z) = 0,$$

where  $m = m_1 m_2 / (m_1 + m_2)$ , and within I3,

$$(3.5) \quad (\ddot{y} - \ddot{z}) + (k/m)(y - z) = 0.$$

We can check that  $V = (y - z)^2 + (m/k)(\dot{y} - \dot{z})^2$  is a common Lyapunov function for (3.4) and (3.5). Hence,

$$(3.6) \quad |y(t) - z(t)| \leq \sqrt{(y(0) - z(0))^2 + (m/k)(\dot{y}(0) - \dot{z}(0))^2}$$

for all  $t \geq 0$ , independent of switching between states.

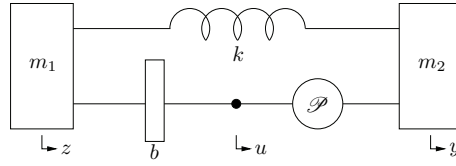
Suppose now that a switching occurs at  $t_0$  as in (b). We will show that the “dwell time” in state I3 is bounded from below. If there is a unique continuation of the system into state I1, then there is nothing to prove. So let us assume that there is a unique continuation in I3. As in (b) we can find that  $Y - Z > 0$  and  $y - u = \epsilon + (Y - Z)(\cos(\omega_3(t - t_0)) - 1)$ . There are two possibilities. If  $Y - Z \leq \epsilon$ , then the system returns to I1 after one oscillation cycle with  $t_1 - t_0 = 2\pi/\omega_3$ . If  $Y - Z > \epsilon$ , there is a transition to I2 with

$$(3.7) \quad t_1 - t_0 = \arccos\left(1 - \frac{2\epsilon}{Y - Z}\right) / \omega_3.$$

Since  $Y - Z$  is bounded above by the right-hand side of (3.6), the dwell time given by (3.7) is bounded from below.

Consequently, it is concluded that there can be no limit point of switching times since the sequence of states occupied by the system must alternate between I3 and either I1 or I2. This in turn means that we can find solutions for  $y(t)$ ,  $z(t)$ ,  $u(t)$  for all  $t \geq 0$ .  $\square$

**3.3. The inerter.** In [16] an ideal modeling element termed the *inerter* was introduced with the following definition. The (ideal) inerter is defined to be a two-terminal mechanical element with the property that the equal and opposite force applied at the terminals is equal to the relative acceleration between them, i.e.,  $F = b(\ddot{z}_1 - \ddot{z}_2)$  in the notation of Figure 9, where  $b$  is the constant of proportionality in kilograms. For the purpose of modeling mechanical networks we will assume that the ideal inerter has zero mass (which is similar to the assumption that ideal springs and dampers have zero mass). Mechanical realizations of inerters have been described in [17, 15].

FIG. 10. *Harmonic oscillator network with an inerter and ideal play.*

**3.4. A network incorporating an inerter with play.** Consider the network in Figure 10, which differs from Figure 8 only in that an inerter replaces the damper. The dynamical equations are

$$(3.8a) \quad m_2 \ddot{y} = k(z - y) + F,$$

$$(3.8b) \quad m_1 \ddot{z} = -k(z - y) - F,$$

$$(3.8c) \quad F = b(\ddot{z} - \ddot{u}).$$

We will now write down the form of the solutions in each state.

J1 (engagement–extension). This corresponds to  $y - u = \epsilon$  and  $F \leq 0$ . The dynamical equations have solutions given by

$$\begin{aligned} y(t) - z(t) &= A_1 \cos(\omega_1 t) + B_1 \sin(\omega_1 t), \\ m_2 y(t) + m_1 z(t) &= C_1 + D_1 t. \end{aligned}$$

where  $\omega_1 = \sqrt{\frac{k(m_1 + m_2)}{m_1 m_2 + b(m_1 + m_2)}}$ .

J2 (engagement–compression). This corresponds to  $y - u = -\epsilon$  and  $F \geq 0$ . The dynamical equations are identical to state J1.

J3 (disengagement). This corresponds to  $|y - u| < \epsilon$  and  $F = 0$ . The dynamical equations have solutions given by

$$\begin{aligned} (3.9) \quad y(t) - z(t) &= A_3 \cos(\omega_3 t) + B_3 \sin(\omega_3 t), \\ m_2 y(t) + m_1 z(t) &= C_3 + D_3 t. \end{aligned}$$

where  $\omega_3 = \sqrt{\frac{k(m_1 + m_2)}{m_1 m_2}}$ .

**3.4.1. Well-posedness and the need for impulsive forces.** We will now study the well-posedness of the network in Figure 10. We will show that there is a qualitative difference in the behavior compared to the network of Figure 8. In the first place, the dynamical equations (3.8) do not imply directly that  $\dot{y}(t)$  and  $\dot{z}(t)$  are continuous (as was the case in Proposition 3.1(a)). This leaves open the possibility of impulsive forces being generated at transitions between states. Physical intuition might suggest the absence of impulsive forces for transitions from engagement to disengagement (Remark 1). With such an assumption we will show that there are uniquely defined transitions in such cases (Proposition 3.2(a)). For transitions from disengagement to engagement it is not obvious that impulsive forces can be dispensed with. In Proposition 3.2(b) we will show that impulsive forces are needed for there to be a well-defined transition. We will also see that the dynamical equations do not define the strength of the impulse uniquely. Thus, multiple solutions of the dynamical equations exist and well-posedness is lost. Once again, the proposition is stated only for transitions between J1 and J3.

PROPOSITION 3.2. *Consider the network of Figure 10 in which the displacements  $y(t)$ ,  $z(t)$ , and  $u(t)$  are assumed to be continuous.*

(a) *Suppose the system is in state J1 just before  $t_0$  and  $F = b(\ddot{z} - \ddot{u}) \uparrow 0$  (strictly) as  $t \uparrow t_0$ . Then the system undergoes a well-defined transition to J3 at  $t_0$  under the assumption that no impulsive forces are generated.*

(b) *Suppose the system is in state J3 just before  $t_0$  and  $y(t) - u(t) \uparrow \epsilon$  as  $t \uparrow t_0$ . Then  $\dot{y}(t_0^-) - \dot{u}(t_0^-) =: \alpha \geq 0$ . A well-defined transition to J1 occurs under the following two conditions:*

(i) *An impulsive force  $P\delta(t - t_0)$  occurs with  $P = P_0 := -(m_1 m_2 b)/(m_1 m_2 + b(m_1 + m_2))\alpha$  in the inerter;*

(ii)  *$\ddot{y}(t_0^-) - \ddot{u}(t_0^-) > 0$ .*

*A well-defined transition to J3 occurs under the following condition:*

(iii) *An impulsive force  $P\delta(t - t_0)$  in the inerter occurs with  $P < P_0$ .*

*In all cases of well-defined transitions  $P$  is negative, which is consistent with the engagement condition being extensive. There is no solution of the system equations assuming an impulsive force with  $P > P_0$ .*

*Proof.* (a) Suppose the network is in state J1 just before time  $t_0$  and that  $\ddot{z}(t) - \ddot{u}(t) \uparrow 0$  as  $t \uparrow t_0$ . Since  $y(t) - u(t) = \epsilon$  just before  $t_0$ , we have the following conditions:  $y(t_0) - u(t_0) = \epsilon$ ,  $\dot{y}(t_0^-) - \dot{u}(t_0^-) = 0$ ,  $\ddot{z}(t_0^-) = \ddot{u}(t_0^-) = \ddot{y}(t_0^-)$ . Within J1 we can check from (3.8a) and (3.8b) that

$$(3.10) \quad z - y = -\frac{m_1 m_2 + b(m_1 + m_2)}{k(m_1 + m_2)} (\ddot{z} - \ddot{y}).$$

This implies  $z(t_0) = y(t_0)$  (but there is no obvious relationship between  $\dot{z}(t_0^-)$  and  $\dot{y}(t_0^-)$ ). This means  $y(t) - z(t) = B_1 \sin(\omega_1(t - t_0))$  just before  $t = t_0$ , and we deduce from (3.10) that  $B_1 \geq 0$  in order to respect the sign constraint on  $F$  in J1.

By assumption, no impulsive forces are generated at  $t_0$ , so  $\dot{z}(t_0^+) = \dot{z}(t_0^-)$ ,  $\dot{y}(t_0^+) = \dot{y}(t_0^-)$ , and  $\dot{u}(t_0^+) = \dot{u}(t_0^-)$ . We conclude that  $\dot{y}(t_0) - \dot{z}(t_0) = B_1 \geq 0$ . Furthermore, there cannot be a continuation of the network after  $t_0$  in state J1 unless  $B_1 = 0$ .

Let us consider whether a continuation in state J3 is always possible. From (3.9) we find that  $y(t) - z(t) = (B_1/\omega_3) \sin(\omega_3(t - t_0))$ . Further, we have  $u(t) = z(t) - \epsilon + B_1(t - t_0)$  after  $t_0$  since  $\ddot{u} = \ddot{z}$  in J3,  $u(t_0) - z(t_0) = -\epsilon$ , and  $\dot{u}(t_0) - \dot{z}(t_0) = \dot{y}(t_0) - \dot{z}(t_0) = B_1$ . Therefore,

$$\begin{aligned} y(t) - u(t) &= y(t) - z(t) + \epsilon - B_1(t - t_0) \\ &= \epsilon + B_1 (\omega_3^{-1} \sin(\omega_3(t - t_0)) - (t - t_0)), \end{aligned}$$

which we can readily check to be decreasing after  $t_0$ , which is consistent with a transition to J3.

(b) We now consider a possible transition from J3 to J1. Suppose the network is in state J3 and  $y(t) - u(t) \uparrow \epsilon$  as  $t \uparrow t_0$  and  $\ddot{z} = \ddot{u}$  just before time  $t_0$ . Again we must have  $\dot{y}(t_0^-) - \dot{u}(t_0^-) \geq 0$  since  $y(t) - u(t) \uparrow \epsilon$ . Let us consider the case  $\dot{y}(t_0^-) - \dot{u}(t_0^-) =: \alpha > 0$ . If there is a transition from J3 to J1, it must be true that  $\dot{y}(t_0^+) - \dot{u}(t_0^+) = 0$ . Thus, there must be discontinuities in velocities. This in turn implies that there must be an impulsive force in the inerter. Let us consider an impulse of the form  $P\delta(t - t_0)$  in  $F$ . (On physical grounds we would expect  $P \leq 0$  since J1 is extensive.) From (3.8c) we observe that

$$(3.11a) \quad (\dot{z}(t_0^+) - \dot{u}(t_0^+)) - (\dot{z}(t_0^-) - \dot{u}(t_0^-)) = P/b,$$

while (3.8a) and (3.8b) imply

$$(3.11b) \quad \dot{y}(t_0^+) - \dot{y}(t_0^-) = P/m_2,$$

$$(3.11c) \quad \dot{z}(t_0^+) - \dot{z}(t_0^-) = -P/m_1.$$

We therefore find that

$$0 = \dot{y}(t_0^+) - \dot{u}(t_0^+) = P/m_1 + P/m_2 + P/b + \alpha,$$

which means that  $P = P_0$  and indeed we have  $P < 0$ . For a valid transition to J1 we must have  $F \leq 0$  for  $t > t_0$ . If the state remains in J1, we must have  $\ddot{z}(t) - \ddot{y}(t) = A_1 \cos(\omega_1(t - t_0)) + B_1 \sin(\omega_1(t - t_0))$  and  $\ddot{u}(t) = \ddot{y}(t)$ . Therefore,  $F \leq 0$  after  $t_0$  requires  $A_1 < 0$ . Note that  $A_1 = y(t_0) - z(t_0) = -\omega_3^{-2}(\ddot{y}(t_0^-) - \ddot{u}(t_0^-))$ . This establishes the required conditions (i) and (ii) for a well-defined transition to J1.

Let us now consider the possibility that an impulsive force at  $t_0$  allows a continuation for the system in state J3 after  $t_0$ . We can check that if  $P < P_0$ , then  $\dot{y}(t_0^+) - \dot{u}(t_0^+) < 0$ , and that we have a valid solution for the state J3 after  $t_0$ . If  $P > P_0$ , there is no valid solution.  $\square$

*Remark 1.* In Proposition 3.2(a) the analysis was restricted to the case where no impulse occurs at disengagement. In the next paragraph we will show that if an impulse does occur at disengagement, the system experiences an overall increase in energy. Thus, it seems reasonable to exclude such behavior on the grounds that the system is passive.

Assume (3.11) applies. Without loss of generality we can restrict our attention to trajectories for which  $m_2 y + m_1 z \equiv 0$  (cf. (3.8)), so we may take

$$(3.12) \quad [y, \dot{y}, z, \dot{z}](t_0^-) = [A, B, -(m_2/m_1)A, -(m_2/m_1)B].$$

Suppose the system transfers from J1 to J3 at  $t_0$ , as in Proposition 3.2(a). Then  $\dot{y}(t_0^+) = -(m_1/m_2)\dot{z}(t_0^+) = P/m_2 + B$  and  $\dot{u}(t_0^+) = B - P(m_1^{-1} + b^{-1})$  from (3.11). Let  $E = (m_1 \dot{z}^2 + m_2 \dot{y}^2 + b(\dot{z} - \dot{u})^2)/2$  denote the kinetic energy of the system. Then we can verify that

$$\begin{aligned} E(t_0^-) &= \frac{1}{2} m_2 B^2 \left( 1 + b \left( \frac{1}{m_2} + \frac{1}{m_1} \right) \right) \left( 1 + \frac{m_2}{m_1} \right), \\ E(t_0^+) &= E(t_0^-) + \frac{1}{2} P^2 \left( \frac{1}{m_1} + \frac{1}{m_2} + \frac{1}{b} \right). \end{aligned}$$

We observe that for any  $P \neq 0$ ,  $E(t_0^+) > E(t_0^-)$ .

*Remark 2.* It is interesting to calculate the change in kinetic energy of the system when an impulse occurs at engagement as in Proposition 3.2(b). In the next paragraph we will show that energy is dissipated, providing  $2P_0 < P \leq P_0$  (recall that  $P_0 < 0$ ). We will see that the maximum energy is dissipated when  $P = P_0$  and energy is conserved when  $P = 2P_0$ . If  $P < 2P_0$ , then the impulse leads to an increase in kinetic energy.

To see this, consider (3.11) with (3.12) and  $\dot{y}(t_0^-) - \dot{u}(t_0^-) := \alpha \geq 0$ . Then we can verify that

$$\begin{aligned} \dot{z}(t_0^+) &= -(m_2/m_1)\dot{y}(t_0^+) = -P/m_1 - (m_2/m_1)B, \\ \dot{z}(t_0^-) - \dot{u}(t_0^-) &= \alpha - B(m_2/m_1 + 1), \\ \dot{z}(t_0^+) - \dot{u}(t_0^+) &= \alpha - B(m_2/m_1 + 1) + P/b. \end{aligned}$$

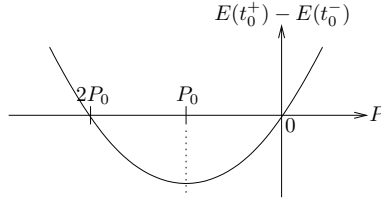


FIG. 11. Change in kinetic energy due to an impulse of strength  $P$  at  $t = t_0$ . From Proposition 3.2, solutions of the system equations exist only if  $P \leq P_0$ . Energy is dissipated when  $2P_0 < P \leq P_0$  and energy increases when  $P < 2P_0$ .

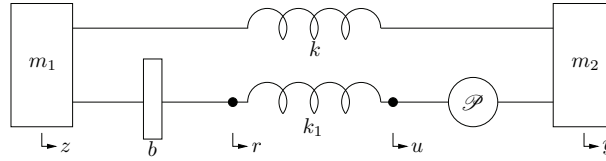


FIG. 12. Network with an inerter, a compliance spring, and ideal play.

As in Remark 1 we can compute the kinetic energy at  $t_0^-$  and  $t_0^+$  to find that

$$E(t_0^-) = \frac{1}{2} (m_2 B)^2 (m_1^{-1} + m_2^{-1}) + \frac{1}{2} b \left( \alpha - m_2 B (m_1^{-1} + m_2^{-1}) \right)^2,$$

$$E(t_0^+) = \frac{1}{2} (P + m_2 B)^2 (m_1^{-1} + m_2^{-1}) + \frac{1}{2} b \left( \alpha - m_2 B (m_1^{-1} + m_2^{-1}) + \frac{P}{b} \right)^2.$$

The required conclusions follow from these expressions. In particular we can show that  $E(t_0^+)$  has a minimum at  $P = P_0$ . This is illustrated in Figure 11.

**3.5. Compliance in series with an inerter and play.** Consider the network in Figure 12 which differs from Figure 10 by the insertion of a spring in series with the inerter and play. The dynamical equations are

$$(3.13a) \quad m_2 \ddot{y} = k(z - y) + F$$

$$(3.13b) \quad m_1 \ddot{z} = -k(z - y) - F,$$

$$(3.13c) \quad F = b(\ddot{z} - \ddot{r}),$$

$$(3.13d) \quad F = k_1(r - u).$$

We now write down the form of the solutions in each state.

K1 (engagement–extension). This corresponds to  $u = y - \epsilon$ ,  $F \leq 0$ . We can solve for  $\ddot{r}$  and reduce the dynamical equations to

$$(3.14a) \quad \ddot{y} = m_2^{-1} (k(z - y) + k_1(r - u)),$$

$$(3.14b) \quad \ddot{z} = -m_1^{-1} (k(z - y) + k_1(r - u)),$$

$$(3.14c) \quad \ddot{r} = -m_1^{-1} (k(z - y) + k_1(r - u) (1 + m_1/b)).$$

These can be written in state-space form as follows:  $\dot{x}(t) = A_{K1}x(t) + B_{K1}\epsilon$  with  $x = [y, \dot{y}, z, \dot{z}, r, \dot{r}]^T$ . In section 3.5.2 we will solve the equations symbolically by means of the expression

$$(3.15) \quad \mathcal{L}^{-1}\{X(s)\} = \mathcal{L}^{-1}\{\Psi(s)\}x(0) + \mathcal{L}^{-1}\{s^{-1}\Psi(s)\}B_{K1}\epsilon,$$

where the resolvent of  $A_{K1}$ ,  $\Psi(s) = (sI - A_{K1})^{-1}$ , is obtained using Maple.

K2 (engagement–compression). This corresponds to  $u = y + \epsilon$ ,  $F \geq 0$ . The dynamical equations are identical to state K1 but with a sign change in  $\epsilon$ .

K3 (disengagement). This corresponds to  $|y - u| < \epsilon$  and  $F = 0$ . The dynamical equations for  $y$  and  $z$  are identical to case J3 and we have  $\ddot{z}(t) = \ddot{r}(t)$  and  $r(t) = u(t)$ .

**3.5.1. A special well-posedness argument.** This section shows that there is a qualitative difference between the behaviors of the networks of Figures 10 and 12. In the first place, the inclusion of the series spring prevents impulsive forces from being generated. Second, the modeling equations specify a unique transition from disengagement to engagement.

**PROPOSITION 3.3.** *Consider the network of Figure 12 in which the displacements  $y(t)$ ,  $z(t)$ ,  $u(t)$ ,  $r(t)$  are assumed to be continuous.*

(a) *For any solutions of the network,  $\dot{y}(t)$ ,  $\dot{z}(t)$ ,  $\dot{r}(t)$ ,  $\ddot{y}(t)$ ,  $\ddot{z}(t)$ , and  $\ddot{r}(t)$  are continuous including transitions between states. However,  $\dot{u}(t)$  is not necessarily continuous at transitions.*

(b) *Suppose the system is in state K1 just before  $t_0$  and  $F \uparrow 0$  (strictly), i.e.,  $\dot{r}(t_0^-) - \dot{u}(t_0^-) > 0$ . Then there is a well-defined transition to state K3.*

(c) *Suppose the system is in state K3 just before  $t_0$  and  $y - u \uparrow \epsilon$  (strictly), i.e.,  $\dot{y}(t_0^-) - \dot{u}(t_0^-) > 0$ . Then there is a well-defined transition to state K1.*

*Proof.* (a) Substitution of (3.13d) into (3.13a) and (3.13b) shows that  $\ddot{y}$ ,  $\ddot{z}$  are continuous, and equating (3.13c) and (3.13d) shows that  $\ddot{r}$  is continuous. Hence,  $\dot{y}$ ,  $\dot{z}$ , and  $\dot{r}$  are continuous, but it is not necessarily true that  $\dot{u}$  is continuous.

(b) Since

$$(3.16) \quad \dot{r}(t_0^-) - \dot{u}(t_0^-) = \dot{r}(t_0^-) - \dot{y}(t_0^-) > 0,$$

a continuation in state K1 is not possible, as this would violate the condition that  $F \leq 0$ . A well-defined transition to K3 occurs if

$$(3.17) \quad \dot{y}(t_0^+) - \dot{u}(t_0^+) < 0.$$

By definition of state K3 we must have  $\dot{r}(t_0^+) = \dot{u}(t_0^+)$ . From (a),  $\dot{y}(t_0^+) = \dot{y}(t_0^-)$  and  $\dot{r}(t_0^+) = \dot{r}(t_0^-)$ . Therefore,  $\dot{y}(t_0^+) - \dot{u}(t_0^+) = \dot{y}(t_0^-) - \dot{r}(t_0^-)$ . Hence, (3.17) follows from (3.16). Incidentally, we have also shown that  $\dot{u}$  is discontinuous at  $t_0$  since  $\dot{u}(t_0^+) - \dot{u}(t_0^-) = \dot{r}(t_0^-) - \dot{y}(t_0^-)$ . This follows since  $y = u + \epsilon$  just before  $t_0$  so that  $\dot{u}(t_0^-) = \dot{y}(t_0^-)$ .

(c) Since

$$(3.18) \quad \dot{y}(t_0^-) - \dot{u}(t_0^-) > 0,$$

a continuation in state K3 is not possible, as this would violate the condition  $|y - u| < \epsilon$ . For a well-defined transition to K1 we need  $F \leq 0$  after  $t_0$ . For this to happen it would be sufficient to have

$$(3.19) \quad \dot{r}(t_0^+) - \dot{u}(t_0^+) < 0.$$

By definition of state K3 we must have  $\dot{r}(t_0^+) = \dot{u}(t_0^+)$ . Therefore, from (a)  $\dot{y}(t_0^-) - \dot{u}(t_0^-) = \dot{y}(t_0^+) - \dot{r}(t_0^+)$ . Hence, (3.19) follows from (3.18). As in (b) we have shown that  $\dot{u}$  is discontinuous at  $t_0$ .  $\square$

*Remark 3.* In the following we extend Proposition 3.3(b) to the case  $\dot{r}(t_0^-) - \dot{u}(t_0^-) = 0$  when the system is in state K1 and  $F \uparrow 0$ . For a transition to K3 we require  $y - u < \epsilon$  for  $t > t_0$ . We now derive a condition for this to hold. For a transition to K3,  $\dot{r}(t_0^+) - \dot{u}(t_0^+) = 0$ . Since  $\dot{r}$  is continuous (see Proposition 3.3(a)),  $\dot{u}$



is continuous at  $t_0$ . This gives  $\dot{y}(t_0^-) - \dot{u}(t_0^-) = \dot{y}(t_0^+) - \dot{u}(t_0^+) = 0$ . Hence it is necessary to consider  $\ddot{y}(t_0^+) - \ddot{u}(t_0^+)$ . From Proposition 3.3(a) and  $F \equiv 0$  for  $t > t_0$  it follows that  $\ddot{r}(t_0^-) = \ddot{r}(t_0^+) = \ddot{u}(t_0^+)$ . Also, without loss of generality, we can restrict our attention to trajectories for which  $m_2 y + m_1 z \equiv 0$  so that  $y(t_0^-) = A$ ,  $z(t_0^-) = -(m_2/m_1)A$ , and  $r(t_0^-) = A - \epsilon$ . We can calculate

$$(3.20) \quad \ddot{y}(t_0^+) - \ddot{u}(t_0^+) = \ddot{y}(t_0^-) - \ddot{r}(t_0^-) = -m_2 k (m_1^{-1} + m_2^{-1})^2 A,$$

where the right-hand side follows from (3.14a) and (3.14c). Thus, from (3.20) there is a transition to K3 if  $A > 0$ . (Note that the sign of  $A$  is the opposite to the sign of the force in the spring  $k$  since  $z(t_0^-) - y(t_0^-) = -(1 + m_2/m_1)A$ .)

Now let us consider the possibility that there is no transition to K3, but the system remains in K1 (again with  $F \uparrow 0$  and  $\dot{r}(t_0^-) - \dot{u}(t_0^-) = 0$ ). For this to occur it is necessary that  $F < 0$  after  $t_0$ , i.e.,  $\ddot{r} - \ddot{u} < 0$  after  $t_0$ . Note that  $\ddot{u}$  is continuous at  $t_0$  since  $\ddot{y}$  is continuous and  $\ddot{y} = \ddot{u}$  in K1. Therefore,

$$(3.21) \quad \ddot{r}(t_0^+) - \ddot{u}(t_0^+) = \ddot{r}(t_0^-) - \ddot{u}(t_0^-) = -(\ddot{y}(t_0^-) - \ddot{r}(t_0^-)).$$

Comparing (3.21) with (3.20) we see that the system remains in state K1 if  $A < 0$ .

To conclude, even in the pathological case that  $F \uparrow 0$  from state K1 and  $\dot{r}(t_0^-) - \dot{u}(t_0^-) = 0$ , we have a well-defined transition to K3 or K1, depending on the sign of the force in the spring  $k$ . The case where  $A = 0$  requires consideration of  $\ddot{y}(t_0^+) - \ddot{u}(t_0^+)$  and  $\ddot{r}(t_0^+) - \ddot{u}(t_0^+)$  to determine whether there is a transition to K3 or whether the system remains in K1. The details are cheerfully left to the reader.<sup>1</sup>

Similarly, we can extend Proposition 3.3(c) to include the case  $\dot{y}(t_0^-) - \dot{u}(t_0^-) = 0$  when the system is in state K3 and  $y - u \uparrow \epsilon$ . We will now show that the conditions are identical to the ones laid out above. For a transition back to K3,  $\ddot{r}(t_0^-) = \ddot{u}(t_0^-)$  and  $\ddot{r}(t_0^+) = \ddot{u}(t_0^+)$  and the case is identical to the one analyzed in (3.20), namely, the situation occurs if  $A > 0$ . For a transition to K1,  $F \leq 0$  for  $t > t_0$  and we require again  $r - u < 0$  for  $t > t_0$ . Since  $\dot{y}(t_0^+) = \dot{u}(t_0^+)$ ,  $\dot{u}$  is continuous at  $t_0$ , and we also have  $\dot{y}(t_0) = \dot{r}(t_0)$ . Therefore, for a transition to K1, we need  $\ddot{r}(t_0^+) - \ddot{u}(t_0^+) < 0$ . We can check that

$$(3.22) \quad \begin{aligned} \ddot{r}(t_0^+) - \ddot{u}(t_0^+) &= \ddot{r}(t_0^+) - \ddot{y}(t_0^+) = \ddot{r}(t_0^-) - \ddot{y}(t_0^-) = -(\ddot{y}(t_0^-) - \ddot{z}(t_0^-)) \\ &= -k (m_1^{-1} + m_2^{-1}) (z(t_0^-) - y(t_0^-)) = m_2 k (m_1^{-1} + m_2^{-1})^2 A, \end{aligned}$$

where the right-hand side follows from (3.13a) and (3.13b) with  $F \equiv 0$  in state K3. Equation (3.22) is the same condition as in (3.21) and (3.20). Thus, there is a well-defined transition to K1 if  $A < 0$ .

**3.5.2. Convergence to a solution of section 3.4.1.** The networks in Figures 10 and 12 are similar and differ only by the spring  $k_1$ . In this section we investigate the convergence behavior of the latter network when  $k_1 \rightarrow \infty$ . We will focus on the transition from disengagement to engagement, where impulsive forces were needed for the network in Figure 10. We will show that the “dwell time” in the engagement state tends to zero as  $k_1 \rightarrow \infty$ . Furthermore, the force through the play element in the engagement state approaches an impulse in the limit as  $k_1 \rightarrow \infty$ . Finally, the strength of the impulse is uniquely determined and implies conservation of energy through the

<sup>1</sup>We are pleased to cite Professor Harry Dym as the originator of this useful sentence [5, p. 454].

impulsive impact. For simplicity we carry out the analysis with numerical values for the parameters  $m_1$ ,  $m_2$ ,  $b$ , and  $k$ .

First, consider the solution in case K1. The determinant of  $sI - A_{K1}$  is given by

$$\left( s^4 + \frac{m_1 m_2 k_1 + b(k + k_1)(m_1 + m_2)}{m_1 m_2 b} s^2 + \frac{k k_1 (m_1 + m_2)}{m_1 m_2 b} \right) s^2.$$

With  $m_1 = m_2 = 1$  kg,  $b = 1/2$  kg, and  $k = 1$  Nm<sup>-1</sup>, two roots are at the origin, and the roots of the quadratic in  $s^2$  are

$$-\left( 2k_1 + 1 \pm \sqrt{4k_1^2 + 1} \right) = -k_1 \left( 2 + \frac{1}{k_1} \pm 2 \left( 1 + \frac{1}{8k_1^2} + O(k_1^{-4}) \right) \right).$$

Thus, the natural frequencies are given by  $s = \pm j\omega_1, \pm j\omega_2$ , where

$$\begin{aligned} \omega_1 &= 1 + O(k_1^{-1}), \\ \omega_2 &= \sqrt{4k_1 + 1} + O(k_1^{-3/2}). \end{aligned}$$

From (3.13a) and (3.13b) we see that  $m_2 \ddot{y} + m_1 \ddot{z} = 0$ . Without loss of generality we can restrict our attention to trajectories for which  $m_2 y(t) + m_1 z(t) \equiv 0$ . We therefore consider a transition from K3 to K1, with the initial condition for the dynamic equations in state K1 equal to

$$(3.23) \quad x_0 = [A, B, -(m_2/m_1)A, -(m_2/m_1)B, A - \epsilon, B - K]^T.$$

We then solve (3.15) to find that there is a single dominant term in the force through the inerter given by  $-bK\omega_2 \sin(\omega_2 t)/2$ . We observe that this dominant term is negative for  $0 < t < t_1 := \pi/\omega_2$ . This suggests that the system will remain in state K1 until time  $t_1$  (approximately) at which there is a transition to K3. Qualitatively, this behavior is approaching one where there is an impulsive force at engagement with an immediate transition back to state K3. This is similar to behavior that was identified in section 3.4.1. Further note that

$$(3.24) \quad -\int_0^{\pi/\omega_2} bK\omega_2 \sin(\omega_2 t)/2 = -K/2,$$

which would be the expected strength of the impulse.

It is also interesting to find that the dominant term in  $\dot{y}(t) - \dot{r}(t)$  is equal to  $K \cos(\omega_2 t)$  for large  $k_1$ . This shows that  $\dot{y}(t_1) - \dot{r}(t_1) = -K$ . Note that  $\dot{u}(t_1^+) = \dot{r}(t_1)$  since there is a transition back to K3 at  $t_1$ . Therefore, we obtain

$$(3.25) \quad \dot{y}(t_0^+) - \dot{u}(t_0^+) = -K$$

in the limit as  $k_1 \rightarrow \infty$ . Using the method of section 3.4.1, (3.25) together with the assumption that  $\dot{y}(t_0^-) - \dot{u}(t_0^-) = K$  requires

$$(3.26) \quad P = -2K \frac{m_1 m_2 b}{m_1 b + m_2 b + m_1 m_2},$$

which gives the same value as (3.24). It is also interesting to note that this limiting solution is one in which energy is conserved through the impulsive impact, as discussed in Remark 2.

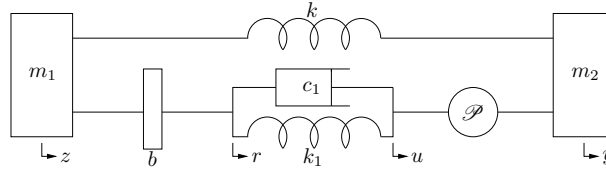


FIG. 13. Network with an inerter, a parallel spring-damper buffer, and ideal play.

**3.6. Buffer in series with an inerter and play.** Consider the network in Figure 13, which differs from Figure 10 by the insertion of a buffer consisting of a parallel spring-damper in series with the inerter and play. The dynamical equations are

$$(3.27a) \quad m_2 \ddot{y} = k(z - y) + F$$

$$(3.27b) \quad m_1 \ddot{z} = -k(z - y) - F,$$

$$(3.27c) \quad F = b(\ddot{z} - \ddot{r}),$$

$$(3.27d) \quad F = k_1(r - u) + c_1(\dot{r} - \dot{u}).$$

We now write down the form of the solutions in each state.

L1 (engagement–extension). This corresponds to  $u = y - \epsilon$ ,  $F \leq 0$ . We can solve for  $\ddot{r}$  and reduce the dynamical equations to

$$(3.28a) \quad \ddot{y} = m_2^{-1}(k(z - y) + k_1(r - u) + c_1(\dot{r} - \dot{u})),$$

$$(3.28b) \quad \ddot{z} = -m_1^{-1}(k(z - y) + k_1(r - u) + c_1(\dot{r} - \dot{u})),$$

$$(3.28c) \quad \ddot{r} = -m_1^{-1}(k(z - y) + (k_1(r - u) + c_1(\dot{r} - \dot{u}))(1 + m_1/b)).$$

These can be written in state-space form with  $x = [y, \dot{y}, z, \dot{z}, r, \dot{r}]^T$ . In section 3.6.2 we will solve these equations symbolically.

L2 (engagement–compression). This corresponds to  $u = y + \epsilon$ ,  $F \geq 0$ . The dynamical equations are identical to state L1 but with a sign change in  $\epsilon$ .

L3 (disengagement). This corresponds to  $|y - u| < \epsilon$  and  $F = 0$ . The dynamical equations for  $y$  and  $z$  are identical to case J3, and we have  $\ddot{z}(t) = \ddot{r}(t)$ .

**3.6.1. A special well-posedness argument.** This section shows that there is a qualitative difference between the behaviors of the networks of Figures 10 and 13. In the first place, the inclusion of the parallel spring-damper prevents impulsive forces from being generated. Second, the modeling equations specify unique transitions between system states.

**PROPOSITION 3.4.** *Consider the network of Figure 13 in which the displacements  $y(t)$ ,  $z(t)$ ,  $u(t)$ ,  $r(t)$  are assumed to be continuous.*

(a) *For any solutions of the network,  $\dot{y}(t)$ ,  $\dot{z}(t)$ , and  $\dot{r}(t)$  are continuous, including transitions between states, and no impulsive forces may occur. Furthermore,  $\dot{u}(t)$  is continuous for a transition from engagement to disengagement as also are  $\ddot{y}$ ,  $\ddot{z}$ , and  $\ddot{r}$ . On the other hand,  $\dot{u}$  may be discontinuous for a transition from state L3 to engagement.*

(b) *Suppose the system is in state L1 just before  $t_0$  and  $F \uparrow 0$  (strictly), i.e.,  $\dot{F}(t_0^-) > 0$ . Then there is a well-defined transition to state L3.*

(c) *Suppose the system is in state L3 just before  $t_0$  and  $y - u \uparrow \epsilon$  (strictly), i.e.,  $\dot{y}(t_0^-) - \dot{u}(t_0^-) > 0$ . Then there is a well-defined transition to state L1.*

*Proof.* (a) Substitution of (3.27d) into (3.27a) and (3.27b) shows that  $\dot{y}$ ,  $\dot{z}$  are continuous; otherwise impulsive behavior occurs, which would violate the continuity of  $r$  and  $u$ . Equating (3.27c) and (3.27d) shows that  $\dot{r}$  is continuous.

For a transition from state L1 to L3, we expect the force through the buffer to be continuous, i.e.,  $F(t_0^-) = F(t_0^+)$ . Since  $r$ ,  $u$ ,  $\dot{r}$  are continuous,  $\dot{u}$  has to be continuous through the transition from (3.27d). Consequently, (3.27a), (3.27b), and (3.27c) show that also  $\ddot{y}$ ,  $\ddot{z}$ ,  $\ddot{r}$  are continuous.

For a transition from state L3 to L1,  $F(t_0^-) = 0$  and  $F(t_0^+) \leq 0$ . Since  $r$ ,  $u$ , and  $\dot{r}$  are continuous,  $\dot{u}$  will be discontinuous if  $F(t_0^+) < 0$  from (3.27d).

(b) A continuation in state L1 would require that  $F(t) \leq 0$  for  $t > t_0$ , which is not possible since  $F(t_0) = 0$  and  $\dot{F}(t_0^-) > 0$  (solutions not leaving L1 will have  $\dot{F}$  continuous).

A well-defined transition to L3 requires that  $y(t) - u(t) < \epsilon$  for  $t$  just after  $t_0$ . We will now verify that this holds. Let us assume that a transition to L3 occurs. We first observe that  $F(t) = 0$  for  $t$  just after  $t_0$  and  $F(t_0^-) = 0$  so that  $F(t)$  is continuous at  $t_0$ . (There cannot be an impulse in  $F(t)$  at  $t_0$  since this would violate continuity of displacements using (3.27d).) Since  $r$ ,  $u$ , and  $\dot{r}$  are continuous at  $t_0$ , then  $\dot{u}$  is continuous at  $t_0$  from (3.27d). Therefore,  $\dot{y}(t_0^+) - \dot{u}(t_0^+) = \dot{y}(t_0^-) - \dot{u}(t_0^-) = 0$ . Hence, we need to consider  $\ddot{y}(t_0^+) - \ddot{u}(t_0^+)$ . By assumption,  $F \equiv 0$  after  $t_0$ , which means that

$$\dot{F}(t_0^+) = k_1 (\dot{r}(t_0^+) - \dot{u}(t_0^+)) + c_1 (\ddot{r}(t_0^+) - \ddot{u}(t_0^+)) = 0.$$

Now observe that

$$\begin{aligned} \ddot{y}(t_0^+) - \ddot{u}(t_0^+) &= \ddot{y}(t_0^-) - (k_1/c_1 (\dot{r}(t_0^-) - \dot{u}(t_0^-)) + \ddot{r}(t_0^-)) \\ &= - (k_1 (\dot{r}(t_0^-) - \dot{u}(t_0^-)) + c_1 (\ddot{r}(t_0^-) - \ddot{u}(t_0^-))) / c_1 \\ (3.29) \quad &= -\dot{F}(t_0^-) / c_1 < 0, \end{aligned}$$

using the facts that  $\dot{r}$ ,  $\dot{u}$ ,  $\ddot{r}$  are continuous and that  $\ddot{y}(t_0^-) - \ddot{u}(t_0^-) = 0$ . Equation (3.29) shows that  $y(t) - u(t) < \epsilon$  after  $t_0$ .

(c) A continuation in L3 would require  $y - u < \epsilon$ , which is not possible since  $\dot{y}(t_0^-) - \dot{u}(t_0^-) > 0$  (solutions not leaving L3 will have  $\dot{y} - \dot{u}$  continuous). A well-defined transition to L1 requires that  $F \leq 0$  after  $t_0$ . We will now verify that this holds. In L1,  $y(t) \equiv u(t) + \epsilon$  so that  $\dot{y}(t_0^+) = \dot{u}(t_0^+)$ . Hence

$$\begin{aligned} F(t_0^+) &= k_1 (r(t_0^+) - u(t_0^+)) + c_1 (\dot{r}(t_0^+) - \dot{u}(t_0^+)) \\ &= k_1 (r(t_0^-) - u(t_0^-)) + c_1 (\dot{r}(t_0^-) - \dot{u}(t_0^-)) + c_1 (\dot{u}(t_0^-) - \dot{u}(t_0^+)) \\ &= c_1 (\dot{u}(t_0^-) - \dot{y}(t_0^-)) < 0. \quad \square \end{aligned}$$

**3.6.2. Convergence to solutions of section 3.4.1.** The networks in Figures 10 and 13 are similar and differ only by the spring  $k_1$  and damper  $c_1$ . In this section we investigate the convergence behavior of the network of Figure 13 when  $c_1 = a\sqrt{k_1}$  and  $k_1 \rightarrow \infty$ . This choice of  $c_1$  renders imaginary all (four) nonzero roots of the characteristic equation corresponding to state L1 in the interval  $a \in [0, 1)$ . We will focus on the transition from disengagement to engagement where impulsive forces were needed for the network in Figure 10. We will show that the “dwell time” in the engagement state tends to zero as  $k_1 \rightarrow \infty$ . Furthermore, the force through the play element in the engagement state approaches an impulse in the limit as  $k_1 \rightarrow \infty$ . Finally, the strength of the impulse is uniquely determined and covers a range of impulse

strengths  $P$  as a function of  $a$  as determined in Remark 2. For simplicity we carry out the analysis with numerical values for the parameters  $m_1$ ,  $m_2$ ,  $b$ , and  $k$ .

Consider first the solution in case L1. The determinant of  $sI - A_{L1}$  is given by

$$s^2 \left( s^4 + c_1 \frac{m_1 m_2 + b(m_1 + m_2)}{m_1 m_2 b} s^3 + \frac{m_1 m_2 k_1 + b(m_1 + m_2)(k + k_1)}{m_1 m_2 b} s^2 + k \frac{m_1 + m_2}{m_1 m_2 b} (c_1 s + k_1) \right).$$

Two roots are at the origin, and with  $m_1 = m_2 = 1$  kg,  $b = 1/2$  kg, and  $k = 1$  Nm<sup>-1</sup> the roots of the quartic are

$$\begin{aligned} s_{1,2} &= \pm j + O\left(k_1^{-1/2}\right), \\ s_{3,4} &= -2a\sqrt{k_1} \pm j2\sqrt{k_1}\sqrt{1-a^2} + O\left(k_1^{-1/2}\right). \end{aligned}$$

From (3.27a) and (3.27b) we see that  $m_2 \ddot{y} + m_1 \ddot{z} = 0$ . Without loss of generality we can restrict our attention to trajectories for which  $m_2 y + m_1 z \equiv 0$ . We therefore consider a transition from L3 to L1 with the initial condition for the dynamic equations in state L1 as in (3.23). We then solve the state-space equations for L1 to find the dominant term in the force through the inerter given by

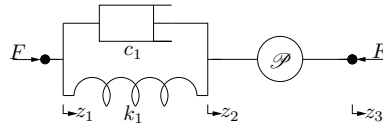
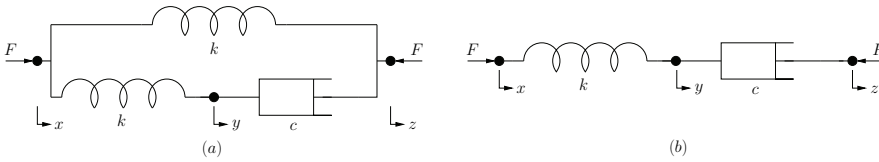
$$b(\ddot{z} - \ddot{r})_{\text{Dom}} = -b \left( \frac{1-2a^2}{\sqrt{1-a^2}} \sin(\omega_2 t) + 2a \cos(\omega_2 t) \right) \sqrt{k_1} K \exp\left(-2a\sqrt{k_1} t\right),$$

where  $\omega_2 = 2\sqrt{k_1}\sqrt{1-a^2}$  (subscript Dom means dominant term). In order that  $F \leq 0$  in state L1, it is necessary that  $K > 0$ . We calculate that this dominant term is negative for  $0 < t < t_1$ . Time  $t_1$  marks the first zero crossing of  $(1-2a^2)/\sqrt{1-a^2} \sin(\omega_2 t_1) + 2a \cos(\omega_2 t_1) = 0$ . This suggests that the system remains in state L1 until time  $t_1$  (approximately) at which there is a transition to L3. Qualitatively, this behavior is approaching one where there is an impulsive force at engagement with an immediate transition back to state L3. This is similar to behavior that was identified in section 3.4.1. We can further calculate the expected strength of the impulse as follows:

$$\begin{aligned} (3.30) \quad P(a) &:= \int_0^{t_1} b(\ddot{z} - \ddot{r})_{\text{Dom}} dt \\ &= -\frac{bK}{2} e^{-2a\sqrt{k_1} t} \left[ 2a \left( \sqrt{1-a^2} \sin(\omega_2 t) - a \cos(\omega_2 t) \right) \right. \\ &\quad \left. - \frac{1-2a^2}{\sqrt{1-a^2}} \left( \sqrt{1-a^2} \cos(\omega_2 t) + a \sin(\omega_2 t) \right) \right] \Big|_0^{t_1}. \end{aligned}$$

The function is strictly monotonic with  $P(0) = -K/2$  and  $\lim_{a \rightarrow 1} P(a) = -(1 + e^{-2})K/4$ .

In summary, we have determined a range of unique solutions for network Figure 13, depending on the amount of damping  $c_1$ , which approximate an impulse in the limit as  $k_1 \rightarrow \infty$ . The buffer network provides impulse strengths in the range  $2P_0 < P < (1 + e^{-2})P_0$  for the chosen numerical values. This captures some, but not all, dissipative solutions illustrated in Figure 11. In particular, the coalescing case where  $P = P_0$  is not captured.

FIG. 14. *Semi-ideal play model.*FIG. 15. *Surface compliance models. (a) Standard linear solid model. (b) Maxwell fluid.*

**4. Semi-ideal play.** We define a semi-ideal play model as having an ideal play element in series with a parallel spring-damper network (Figure 14). The parallel spring-damper buffer was used in section 3.6 for the study of well-posedness, and it is also related to the Kelvin–Voigt model of solids [6, p. 7] to account for the elasticity and friction of bodies in contact. We mention that other buffer networks have also been proposed, e.g., the standard linear and the Maxwell models [8, p. 24], [10, p. 184] shown in Figure 15.

Consider the displacement across the semi-ideal play element to be  $z_d = z_1 - z_3$ , the displacement of the parallel spring-damper buffer network to be  $z_n = z_1 - z_2$ , and the displacement of the ideal play element to be  $z_p = z_2 - z_3$ , with corresponding velocities. Then we can write the following equation for the force:

$$(4.1) \quad F(t) = k_1 (z_d(t) - z_p(t)) + c_1 (\dot{z}_d(t) - \dot{z}_p(t)).$$

Three disjoint cases follow from the description of the ideal play element introduced in section 3.1:

$$(4.2a) \quad (\text{engagement-extension}): F \leq 0, \quad z_p(t) = -\epsilon,$$

$$(4.2b) \quad (\text{disengagement}): |z_p(t)| < \epsilon, \quad F = 0,$$

$$(4.2c) \quad (\text{engagement-compression}): F \geq 0, \quad z_p(t) = \epsilon.$$

**4.1. Connections to an approach of Nordin, Galic, and Gutman.** In [13, sec. 2.1], [14, sec. 2.2] a rotational backlash model is proposed consisting of a compliant shaft in series with a clearance gap. In translational form this can be represented by a network of the type shown in Figure 14, namely, a parallel spring-damper in series with some play element. The corresponding mathematical model given in [13] is as follows:

$$(4.3) \quad \dot{z}_p = \begin{cases} (a) & \max(0, \dot{z}_d + (k_1/c_1)(z_d - z_p)) & \text{if } z_p = -\epsilon, \\ (b) & \dot{z}_d + (k_1/c_1)(z_d - z_p) & \text{if } |z_p| < \epsilon, \\ (c) & \min(0, \dot{z}_d + (k_1/c_1)(z_d - z_p)) & \text{if } z_p = \epsilon \end{cases}$$

with output

$$(4.4) \quad F = \begin{cases} 0 & \text{if } |z_p| < \epsilon, \\ k_1 (z_d - z_p) + c_1 \dot{z}_d & \text{if } |z_p| = \epsilon. \end{cases}$$

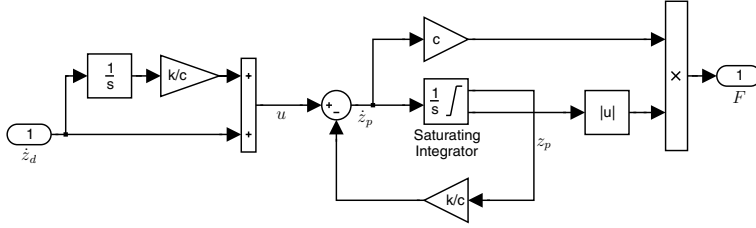


FIG. 16. Semi-ideal play model implementation using Simulink.

At first sight it is not clear what definition of play is being used; however, we will now show that it is practically identical with our behavioral definition.

**PROPOSITION 4.1.** *Let  $(t_1, t_2)$  be an open interval in which (4.1) and (4.2a) hold; then (4.3a) and (4.4) hold over the same interval. The converse statement also holds. The corresponding equivalences for (b) and (c) also hold.*

*Proof.* (4.1), (4.2a)  $\Rightarrow$  (4.3a), (4.4). Since  $z_p(t) = -\epsilon$  on  $(t_1, t_2)$ ,  $\dot{z}_p(t) = 0$  on  $(t_1, t_2)$ . Hence, from (4.1),  $F = k_1(z_d(t) - z_p(t)) + c_1\dot{z}_d(t)$ , which gives (4.4), and we can also write  $\dot{z}_p(t) = \max(0, \dot{z}_d(t) + (k_1/c_1)(z_d(t) - z_p(t)))$  since  $F \leq 0$ , which shows (4.3a).

(4.3a), (4.4)  $\Rightarrow$  (4.1), (4.2a). Since  $z_p(t) = -\epsilon$  on  $(t_1, t_2)$ ,  $\dot{z}_p(t) = 0$  on  $(t_1, t_2)$ . Hence, from (4.3a),  $\dot{z}_d(t) + (k_1/c_1)(z_d(t) - z_p(t)) \leq 0$  on  $(t_1, t_2)$ , and hence (4.1) and (4.2a) hold on  $(t_1, t_2)$ .

(4.1), (4.2b)  $\Leftrightarrow$  (4.3b), (4.4). These are trivially identical.

(4.1), (4.2c)  $\Leftrightarrow$  (4.3c), (4.4). This is similar to (4.1), (4.2a)  $\Leftrightarrow$  (4.3a), (4.4).  $\square$

**Remark 4.** For network interconnections where switching between engagement and disengagement happens at isolated times, the solutions of (4.1)–(4.2) and (4.3)–(4.4) coincide almost everywhere. For solutions in which displacements are continuous, displacements must agree at all times, with the possibility that velocities disagree at isolated times.

**Remark 5.** The definition of semi-ideal play as an input-output model suggests the possibility of using results from feedback systems to establish well-posedness of networks with play, as in Proposition 2.1. We point out that this may be more difficult than with the play operator since the model may give discontinuous outputs for continuous inputs.

**4.2. A network implementation of semi-ideal play.** This section presents a numerical implementation of the play model (4.3)–(4.4) which appears to work well in both ordinary and pathological cases. To determine the state of operation it is necessary to observe the relative displacement across the ideal play element  $z_p(t)$ , which can be calculated by integration of the relative velocity  $\dot{z}_p = \dot{z}_d - \dot{z}_n$ . Integration followed by saturation introduces integrator wind-up. To avoid such complications, we consider a saturating integrator,

$$z_p(t) = \int_0^t \zeta \dot{z}_p(t) dt \quad \text{with} \quad \zeta = \begin{cases} 0 & \text{if } |z_p(t)| = \epsilon, \\ 1 & \text{if } |z_p(t)| < \epsilon, \end{cases}$$

which is readily available in MATLAB Simulink. The complete model is shown in Figure 16 with input  $\dot{z}_d$ , output  $F$ , and  $u = (k_1/c_1)z_d + \dot{z}_d$ .

The saturating integrator produces a saturation signal, which outputs one of the three states  $(-1, 0, 1)$  corresponding to  $(z_p(t) = -\epsilon, -\epsilon < z_p(t) < \epsilon, z_p(t) = \epsilon)$ . This signal would suffice to switch the force output correctly in a continuous time system.

However, the sample time in a digital implementation delays the signal  $z_p(t)$ , the output of the integrator, by a time step to the corresponding force signal  $F$ . Thus,  $F$  can change sign during a transition. This single step instant violates the model description in (4.3)–(4.4). It is recommended to reduce the integrator sample time to much below the time frame of any principle dynamics to reduce this effect.

Figure 17 shows numerical simulations obtained with the implementation of Figure

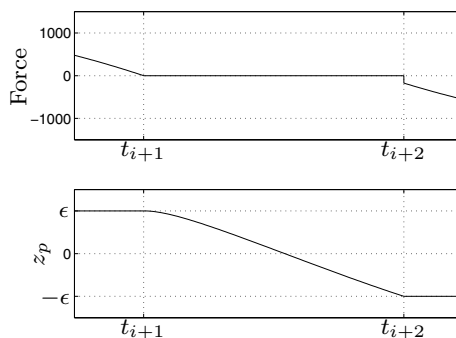
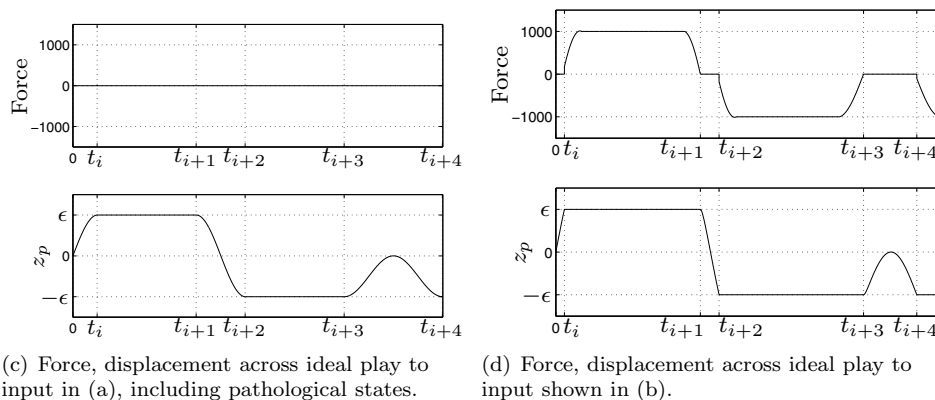
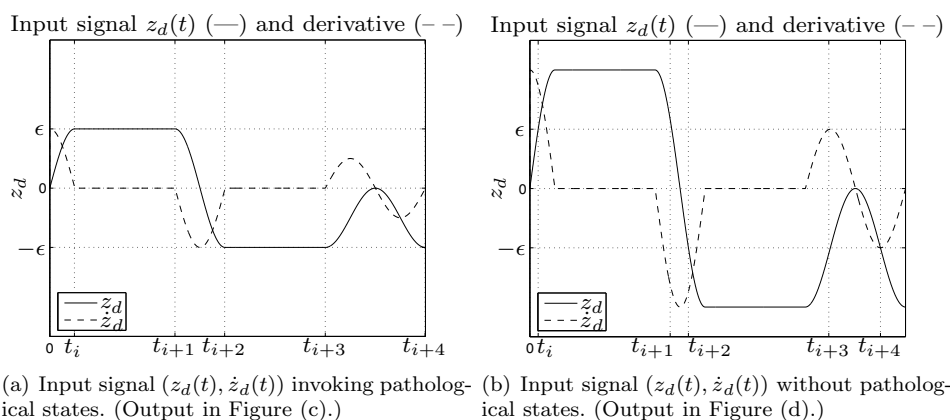


FIG. 17. Input and output signals to network in Figure 16. Figures (a) and (c), and (b) and (d), are signal pairs under the same conditions. The parameters used are  $\epsilon = 0.1$ ,  $k = 10 \text{ kNm}^{-1}$ ,  $c = 1 \text{ kNsm}^{-1}$ .



16. Figures 17(a), (c) consider a very special situation in which  $\dot{z}_d$  is chosen so that engagement and disengagement “only just” occur with the force remaining at zero throughout. Figures 17(b), (d), (e) show ordinary situations of engagement and disengagement in which the force is continuous at disengagement, but not at engagement.

**5. Conclusions.** This paper has shown that the treatment of play as an input-output operator in mechanical networks leads to unsatisfactory solutions from a physical point of view (see section 2.2 for a summary). In contrast, a behavioral definition of play does not suffer from these objections and appears more reasonable from a physical point of view. Well-posedness has been analyzed for several simple networks by considering in detail the transitions between engagement and disengagement states. Networks incorporating inerters were also considered, and it was shown that impulsive forces may sometimes be generated. In such cases multiple solutions may exist with different amounts of instantaneous energy loss. It was shown that the incorporation of a buffer network (either a spring or a parallel spring-damper) recovers well-posedness, and in the limit as the stiffness of the buffer tends to infinity, a range of impulse strengths is obtained. It was seen that the use of such buffer networks is related to the modeling of impact in contact mechanics. Finally, connections with the work of Nordin, Galic, and Gutman were studied. It was pointed out that their input-output model was essentially identical with the ideal play (behavioral definition) in series with a parallel spring-damper. A MATLAB Simulink implementation of the model was presented.

## REFERENCES

- [1] M. BROKATE AND J. SPREKELS, *Hysteresis and Phase Transitions*, Appl. Math. Sci. 121, Springer, New York, 1996.
- [2] M.K. ÇAMLİBEL, W.P.M.H. HEEMELS, A.J. VAN DER SCHAFT, AND J.M. SCHUMACHER, *Switched networks and complementarity*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 50 (2003), pp. 1036–1046.
- [3] R.L. COSGRIFF, *Nonlinear Control Systems*, McGraw-Hill, New York, 1958.
- [4] C.A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.
- [5] P. DEWILDE AND H. DYM, *Schur recursions, error formulas, and convergence of rational estimators for stationary stochastic sequences*, IEEE Trans. Inform. Theory, 27 (1981), pp. 446–461.
- [6] W. FLÜGGE, *Viscoelasticity*, Blaisdell Publishing Company, Waltham, MA, 1967; 2nd revised ed., Springer-Verlag, New York, 1975.
- [7] A. GELB AND W.E. VANDER VELDE, *Multiple-input describing function and nonlinear system design*, McGraw-Hill, New York, 1968.
- [8] W. GOLDSMITH, *Impact*, Edward Arnold Ltd., London, 1960.
- [9] W.P.M.H. HEEMELS, M.K. ÇAMLİBEL, AND J.M. SCHUMACHER, *On the dynamic analysis of piecewise-linear networks*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 49 (2002), pp. 315–327.
- [10] K.L. JOHNSON, *Contact Mechanics*, Cambridge University Press, Cambridge, UK, 1985.
- [11] M.A. KRASNOSEL'SKII AND A.V. POKROVSKII, *Systems with Hysteresis*, Springer-Verlag, Berlin, 1989 (in English); Nauka, Moscow, 1983 (in Russian).
- [12] Y.J. LOOTSMA, A.J. VAN DER SCHAFT, AND M.K. ÇAMLİBEL, *Uniqueness of solutions of linear relay systems*, Automatica, 35 (1999), pp. 467–478.
- [13] M. NORDIN, J. GALIC, AND P.-O. GUTMAN, *New models for backlash and gear play*, Internat. J. Adapt. Control Signal Process., 11 (1997), pp. 49–63.
- [14] M. NORDIN AND P.-O. GUTMAN, *Controlling mechanical systems with backlash—a survey*, Automatica, 38 (2002), pp. 1633–1649.
- [15] C. PAPAGEORGIOU AND M.C. SMITH, *Positive real synthesis using matrix inequalities for mechanical networks: Application to vehicle suspension*, IEEE Trans. Control Syst. Tech., 14 (2006), pp. 423–435.

- [16] M.C. SMITH, *Synthesis of mechanical networks: The inerter*, IEEE Trans. Automat. Control, 47 (2002), pp. 1648–1662.
- [17] M.C. SMITH AND F.-C. WANG, *Performance benefits in passive vehicle suspensions employing inerters*, Vehicle Syst. Dyn., 42 (2004), pp. 235–257.
- [18] A.J. VAN DER SCHAFT AND J.M. SCHUMACHER, *Complementarity modeling of hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 483–490.
- [19] J.C. WILLEMS, *The Analysis of Feedback Systems*, MIT Press, Cambridge, MA, 1971.
- [20] J.C. WILLEMS, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

## OPTIMAL CONSUMPTION IN A GROWTH MODEL WITH THE COBB–DOUGLAS PRODUCTION FUNCTION\*

HIROAKI MORIMOTO<sup>†</sup> AND XUN YU ZHOU<sup>‡</sup>

**Abstract.** An optimal consumption problem is studied in a growth model for the Cobb–Douglas production function in a finite horizon. The problem is transferred into a stochastic Ramsey problem so as to reduce the dimension of the state space. The corresponding state equation is a stochastic differential equation with inherently non-Lipschitz coefficients, whose unique solvability is established. The unique existence of the classical solution of the Hamilton–Jacobi–Bellman equation associated with the original problem is proved, and a synthesis of the optimal consumption policy is presented in the feedback form.

**Key words.** economic growth, Cobb–Douglas production function, Ramsey problem, Hamilton–Jacobi–Bellman equation, viscosity solutions

**AMS subject classifications.** 49L20, 49L25, 91B62

**DOI.** 10.1137/070709153

**1. Introduction.** We deal with the economic growth model originated by Merton [7] for the Cobb–Douglas production function in the finite horizon. We define the following quantities:

$T$  = finite horizon;  
 $y_t$  = labor supply at time  $t \in [0, T]$ ;  
 $z_t$  = capital stock at time  $t \in [0, T]$ ;  
 $\nu$  = the constant rate of depreciation,  $\nu \geq 0$ ;  
 $c_t z_t$  = consumption rate at time  $t \in [0, T]$ ,  $0 \leq c(t) \leq 1$ ;  
 $c_t z_t / y_t$  = the totality of consumption rate per person;  
 $F(z, y)$  = the Cobb–Douglas production function  $z^\alpha y^{1-\alpha}$ ,  $0 < \alpha < 1$ , producing the commodity for the capital stock  $z > 0$  and the labor force  $y > 0$ ;  
 $n, \sigma$  = nonzero constant coefficients;  
 $U(c)$  = the utility function for the consumption rate  $c \geq 0$ .

We assume that the labor supply  $y_t$  and the capital stock  $z_t$  are governed by the stochastic differential equation (SDE)

$$(1.1) \quad dy_t = ny_t dt + \sigma y_t dB_t, \quad y_0 = y > 0,$$

$$(1.2) \quad \dot{z}_t = F(z_t, y_t) - \nu z_t - c_t z_t, \quad 0 < t \leq T, \quad z_0 = z > 0,$$

---

\*Received by the editors November 24, 2007; accepted for publication (in revised form) August 7, 2008; published electronically December 5, 2008.

<http://www.siam.org/journals/sicon/47-6/70915.html>

<sup>†</sup>Department of Mathematics, Graduate School of Science and Engineering, Ehime University, Matsuyama 790-0826, Japan (morimoto@mserv.sci.ehime-u.ac.jp).

<sup>‡</sup>Mathematical Institute and Nomura Centre for Mathematical Finance, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, UK (zhouxy@maths.ox.ac.uk) and Department of System Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong. This author is supported in part by a start-up fund at the University of Oxford and by RGC Earmarked grants CUHK418605 and CUHK418606.

on a complete probability space  $(\Omega, \mathcal{F}, P)$  carrying a standard Brownian motion  $\{B_t\}$ . Let  $c = \{c_t\}$  be a consumption policy per capita such that

$$(1.3) \quad c_t \text{ is progressively measurable with respect to the filtration } \mathcal{F}_t = \sigma(B_s, s \leq t), \\ 0 \leq c_t \leq 1, \quad 0 \leq t \leq T,$$

and we denote by  $\mathcal{A}$  the class of all consumption policies  $\{c_t\}$  per capita.

The purpose of this paper is to present a synthesis of optimal consumption policy  $c^*$  so as to maximize the expected utilities

$$(1.4) \quad J(c) = E \left[ \int_0^T U(c_t z_t / y_t) dt \right]$$

per person with finite horizon  $T$  over the class  $\mathcal{A}$ . The Hamilton–Jacobi–Bellman (HJB) equation associated with this problem is given by

$$(1.5) \quad V_t + \frac{1}{2} \sigma^2 y^2 V_{yy} + n y V_y + \{F(z, y) - \nu z\} V_z + \max_{0 \leq c \leq 1} \{U(c z / y) - c z V_z\} = 0, \quad 0 \leq t < T, \\ V(T, z, y) = 0, \quad z > 0, \quad y > 0,$$

where the subscripts denote the partial derivatives, and the utility function  $U(c)$  is assumed to have the following properties:

$$(1.6) \quad U \in C[0, \infty) \cap C^2(0, \infty), \quad U''(c) < 0 \text{ for } c > 0, \\ U'(\infty) = U(0+) = 0, \quad U'(0+) = U(\infty) = \infty.$$

The last two conditions constitute what is known as the Inada condition. Its economic interpretation is that, while the utility is very small (respectively, very large) for a very small (respectively, very large) consumption rate, the marginal utility diminishes as the consumption rate becomes extremely large.

Under (1.6), by the uniform continuity of  $U$  near 0, we have that

$$(1.7) \quad \forall \varepsilon > 0, \exists C_\varepsilon > 0 : |U(c) - U(\bar{c})| \leq C_\varepsilon |c - \bar{c}| + \varepsilon \quad \text{for } c, \bar{c} > 0.$$

The technical difficulty in solving the problem lies in the fact that the HJB equation (1.5) is a parabolic PDE with two spatial variables  $y$  and  $z$ . The main approach to be employed is to reduce the dimension by turning the problem into a so-called Ramsey problem [7]. Through an analysis of the Ramsey problem, together with the viscosity solution technique, we are able to show that (1.5) admits a smooth solution  $V$ , and the optimal consumption policy  $c^*$  can be represented in a feedback form. A major technical hurdle to overcome is to prove the existence and uniqueness of solutions to the state equation of the Ramsey problem, whose drift coefficient is inherently non-Lipschitz. Moreover, we need to estimate the Hölder order of the solution in time. It should be noted that the stochastic Ramsey problem is analytically studied in [5], but in the *infinite* time horizon. The resulting HJB equation is an elliptic PDE, which is very different from the parabolic PDE dealt with in the present paper. We also refer to [6] for the growth model with the constant-returns-to-scale production function replacing  $F(z, y)$  of (1.2).

This paper is organized as follows. In sections 2 and 3, we reduce (1.5) to the two-dimensional HJB equation associated with the stochastic Ramsey problem, and we show the existence of viscosity solutions of the HJB equation. Sections 4 and 5, respectively, are devoted to the  $C^2$ -regularity and the concavity of the viscosity solution. In section 6, we give a synthesis of the optimal consumption policy.

**2. The stochastic Ramsey problem.** We consider the HJB equation (1.5) and seek the solution  $V(t, z, y)$  of (1.5) of the form

$$(2.1) \quad V(t, z, y) = v(t, x), \quad x = z/y.$$

Clearly,

$$(2.2) \quad yV_y = -xv_x, \quad yV_z = v_x, \quad y^2V_{yy} = x^2v_{xx} + 2xv_x.$$

Then, by (1.5),  $v(t, x)$  solves the HJB equation

$$(2.3) \quad \begin{aligned} v_t(t, x) + \frac{1}{2}\sigma^2 x^2 v_{xx}(t, x) + (x^\alpha - \mu x)v_x(t, x) + \tilde{U}(x, v_x(t, x)) &= 0, \quad 0 \leq t < T, \\ v(T, x) &= 0, \quad x > 0, \end{aligned}$$

where  $\mu = n + \nu - \sigma^2$  and

$$(2.4) \quad \tilde{U}(x, p) = \max_{0 \leq c \leq 1} \{U(cx) - cpx\}, \quad p \in \mathbf{R}.$$

We observe that (2.3) is the HJB equation associated with the stochastic Ramsey problem so as to maximize

$$(2.5) \quad \bar{J}(c) = E \left[ \int_0^T U(c_t R_t) dt \right]$$

over the class  $\mathcal{A}$ , subject to

$$(2.6) \quad \begin{aligned} dR_t &= (R_t^\alpha - \mu R_t - c_t R_t)dt - \sigma R_t dB_t, \quad 0 < t \leq T, \\ R_0 &= x > 0. \end{aligned}$$

The above SDE does not satisfy the Lipschitz condition, as normally required for existence and uniqueness. Moreover, we need to estimate the dependence, if any, of the solution on the time and the initial state. We solve these problems by an ad hoc technique.

**PROPOSITION 2.1.** *For each  $c \in \mathcal{A}$ , there exists a unique positive solution  $\{R_t\} = \{R_t^x\}$  of (2.6), which satisfies*

$$(2.7) \quad E \left[ \sup_{0 \leq t \leq T} R_t^2 \right] \leq C(1 + x^2),$$

$$(2.8) \quad E[|R_r - R_s|] \leq C(1 + x)|r - s|^{1/2}, \quad 0 \leq s \leq r \leq T,$$

$$(2.9) \quad E[|R_s^x - R_s^y|] \leq C|x - y|^{1-\alpha}(1 + x^\alpha + y^\alpha), \quad x, y > 0, \quad 0 \leq s \leq T,$$

where the constant  $C > 0$  depends only on  $\alpha, T, \mu, \sigma$ .

*Proof.* By Itô's formula,

$$\begin{aligned} dR_t^{1-\alpha} &= (1 - \alpha)R_t^{-\alpha}\{(R_t^\alpha - \mu R_t - c_t R_t)dt - \sigma R_t dB_t\} \\ &\quad + \frac{1}{2}\sigma^2(1 - \alpha)(-\alpha)R_t^{-\alpha-1}R_t^2 dt. \end{aligned}$$

Hence, setting  $x_t = R_t^{1-\alpha}$ , we have

$$\begin{aligned}
 (2.10) \quad dx_t &= (1-\alpha) \left\{ 1 - \left( \mu + c_t + \frac{1}{2} \sigma^2 \alpha \right) x_t \right\} dt - (1-\alpha) \sigma x_t dB_t \\
 &= (1-\alpha) \left\{ 1 - \left( c_t + \frac{1}{2} \sigma^2 \alpha \right) x_t \right\} dt - x_t (1-\alpha) (\mu dt + \sigma dB_t), \quad x_0 = x^{1-\alpha}.
 \end{aligned}$$

By linearity, (2.11) admits a unique solution  $\{x_t\}$ . Also, we apply the comparison theorem to (2.11) and

$$d\bar{x}_t = (1-\alpha) \left\{ - \left( \mu + c_t + \frac{1}{2} \sigma^2 \alpha \right) \bar{x}_t \right\} dt - (1-\alpha) \sigma \bar{x}_t dB_t, \quad \bar{x}_0 = x_0.$$

Then

$$\begin{aligned}
 (2.11) \quad x_t &\geq \bar{x}_t \\
 &= x_0 \exp \left\{ (1-\alpha) \left( -\mu t - \int_0^t c_s ds - \frac{1}{2} \sigma^2 \alpha t \right) \right. \\
 &\quad \left. - (1-\alpha) \sigma B_t - \frac{1}{2} (1-\alpha)^2 \sigma^2 t \right\} > 0.
 \end{aligned}$$

Thus, we obtain a positive solution  $\{R_t\}$  of (2.6). Let  $\{\hat{x}_t\}$  be the solution of

$$d\hat{x}_t = -\hat{x}_t (1-\alpha) (\mu dt + \sigma dB_t), \quad \hat{x}_0 = x_0.$$

Setting  $H_t = x_t / \hat{x}_t$  and  $\bar{\alpha} = \sigma^2 (1-\alpha) \alpha / 2$ , we have

$$\begin{aligned}
 dH_t &= (1-\alpha) \left\{ \frac{1}{\hat{x}_t} - \left( c_t + \frac{1}{2} \sigma^2 \alpha \right) H_t \right\} dt \\
 &\leq \left( \frac{1-\alpha}{\hat{x}_t} - \bar{\alpha} H_t \right) dt, \quad H_0 = 1.
 \end{aligned}$$

Therefore

$$(2.12) \quad x_t \leq \hat{x}_t e^{-\bar{\alpha} t} \left\{ 1 + (1-\alpha) \int_0^t \frac{e^{\bar{\alpha} s}}{\hat{x}_s} ds \right\},$$

which yields (2.7).

Now, let  $\beta = 1/(1-\alpha) > 1$  and  $M_t = \exp\{-(1-\alpha)\sigma B_t - (1/2)(1-\alpha)^2\sigma^2 t\}$ . By (2.12) and Doob's maximal inequality, we have

$$\begin{aligned}
 (2.13) \quad E \left[ \sup_{0 \leq t \leq T} x_t^\beta \right] &\leq C \left( 1 + x_0^\beta E \left[ \sup_{0 \leq t \leq T} M_t^\beta \right] \right) \\
 &\leq C \left( 1 + x_0^\beta \left( \frac{\beta}{\beta-1} \right)^\beta E[M_T^\beta] \right) \\
 &\leq C'(1+x),
 \end{aligned}$$

where the constant  $C' > 0$  depends only on  $\alpha, T, \mu, \sigma$ . Hence, by (2.11), (2.13), and the moment inequality for martingales, we get

$$\begin{aligned} E[|x_r - x_s|^\beta] &\leq 2^\beta \left( E \left[ \left| \int_s^r (1 - \alpha) \left\{ 1 - \left( \mu + c_t + \frac{1}{2} \sigma^2 \alpha \right) x_t \right\} dt \right|^\beta \right] \right. \\ &\quad \left. + E \left[ \left| \int_s^r (1 - \alpha) \sigma x_t dB_t \right|^\beta \right] \right) \\ &\leq C \left( E \left[ \left( \int_s^r (1 + x_t^2) dt \right)^{\beta/2} \right] |r - s|^{\beta/2} + E \left[ \left( \int_s^r x_t^2 dt \right)^{\beta/2} \right] \right) \\ &\leq C'(1 + x) |r - s|^{\beta/2}, \quad 0 \leq s \leq r \leq T. \end{aligned}$$

Since

$$|x^\beta - y^\beta| = \left| \int_y^x \beta t^{\beta-1} dt \right| \leq \beta |x - y| (|x|^{\beta-1} + |y|^{\beta-1}), \quad x, y \geq 0,$$

we observe by Hölder's inequality that

$$\begin{aligned} E[|R_r - R_s|] &= E[|x_r^\beta - x_s^\beta|] \\ &\leq \beta (E[|x_r - x_s|^\beta])^{1/\beta} (E[(|x_r|^{\beta-1} + |x_s|^{\beta-1})^{\beta/(\beta-1)}])^{1-1/\beta} \\ &\leq \beta (C'(1 + x) |r - s|^{\beta/2})^{1/\beta} \left( E \left[ 2^{\beta/(\beta-1)} \sup_{0 \leq t \leq T} |x_t|^\beta \right] \right)^{1-1/\beta} \\ &\leq C(1 + x) |r - s|^{1/2}, \end{aligned}$$

which implies (2.8).

Next, we set  $r_t = (R_t^y)^{1-\alpha}$ . Then, by (2.11),

$$d(x_t - r_t) = (1 - \alpha) \left( -\mu - c_t - \frac{1}{2} \sigma^2 \alpha \right) (x_t - r_t) dt - (1 - \alpha) \sigma (x_t - r_t) dB_t,$$

or equivalently,

$$\begin{aligned} x_s - r_s &= (x_0 - r_0) \exp \left\{ (1 - \alpha) \left( -\mu t - \int_0^s c_t dt - \frac{1}{2} \sigma^2 \alpha s \right) \right. \\ &\quad \left. - (1 - \alpha) \sigma B_s - \frac{1}{2} (1 - \alpha)^2 \sigma^2 s \right\}. \end{aligned}$$

Hence

$$E[|x_s - r_s|^\beta] \leq C |x_0 - r_0|^\beta.$$

By Hölder's inequality, we deduce

$$\begin{aligned} E[|R_s^x - R_s^y|] &= E[|x_s^\beta - r_s^\beta|] \\ &\leq \beta (E[|x_s - r_s|^\beta])^{1/\beta} (E[(|x_s|^{\beta-1} + |r_s|^{\beta-1})^{\beta/(\beta-1)}])^{1-1/\beta} \\ &\leq C |x_0 - r_0| (1 + x^\alpha + y^\alpha), \end{aligned}$$

which implies (2.9).

**3. Viscosity solutions.** We study the viscosity solution  $v$  of the HJB equation (2.3), i.e.,

$$(3.1) \quad v_t + \frac{1}{2}\sigma^2 x^2 v_{xx} + (x^\alpha - \mu x)v_x + \tilde{U}(x, v_x) = 0 \quad \text{in } Q := [0, T] \times (0, \infty),$$

$$(3.2) \quad v(T, x) = 0, \quad x > 0.$$

DEFINITION 3.1. *Let  $v \in C([0, T] \times (0, \infty))$  satisfy (3.2). Then  $v$  is called a viscosity solution of (3.1) if the following assertions are satisfied:*

$$a + \frac{1}{2}\sigma^2 x^2 X + (x^\alpha - \mu x)\lambda + \tilde{U}(x, \lambda) \geq 0 \quad \forall (a, \lambda, X) \in \mathcal{P}^{2,+}v(s, x), \quad \forall (s, x) \in Q,$$

$$a + \frac{1}{2}\sigma^2 x^2 X + (x^\alpha - \mu x)\lambda + \tilde{U}(x, \lambda) \leq 0 \quad \forall (a, \lambda, X) \in \mathcal{P}^{2,-}v(s, x), \quad \forall (s, x) \in Q,$$

where  $\mathcal{P}^{2,+}$  and  $\mathcal{P}^{2,-}$  are the second parabolic superdifferentials and subdifferentials [1] defined by

$$\begin{aligned} \mathcal{P}^{2,+}v(s, x) &= \left\{ (a, \lambda, X) \in \mathbf{R}^3 : \right. \\ &\quad \left. \limsup_{(t,y) \in Q \rightarrow (s,x)} \frac{v(t, y) - v(s, x) - a(t-s) - \lambda(y-x) - \frac{1}{2}X(y-x)^2}{|t-s| + |y-x|^2} \leq 0 \right\}, \\ \mathcal{P}^{2,-}v(s, x) &= \left\{ (a, \lambda, X) \in \mathbf{R}^3 : \right. \\ &\quad \left. \liminf_{(t,y) \in Q \rightarrow (s,x)} \frac{v(t, y) - v(s, x) - a(t-s) - \lambda(y-x) - \frac{1}{2}X(y-x)^2}{|t-s| + |y-x|^2} \geq 0 \right\}. \end{aligned}$$

Define

$$(3.3) \quad v(s, x) = \sup_{c \in \mathcal{A}} E \left[ \int_s^T U(c_t X_t) dt \right],$$

where  $\{X_t\}$  is the solution of (2.6) for  $t \in (s, T]$  with  $X_s = x$ , that is,

$$(3.4) \quad dX_t = (X_t^\alpha - \mu X_t - c_t X_t)dt - \sigma X_t dB_t, \quad s < t \leq T, \quad X_s = x > 0,$$

and the supremum is taken over all systems  $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}; \{B_t\}, \{c_t\})$ . We choose  $b_1 > 0$  such that  $x^\alpha - \mu x \leq b_1$ . Taking sufficiently large  $b_0 > b_1$ , we observe that  $\zeta(t, x) := e^{T-t}(x + b_0)$  fulfills

$$\begin{aligned} \zeta_t + \frac{1}{2}\sigma^2 x^2 \zeta_{xx} + (x^\alpha - \mu x)\zeta_x + \tilde{U}(x, \zeta_x) &\leq e^{T-t}\{-b_0 + (x^\alpha - \mu x)\} + \tilde{U}(x, e^{T-t}) \\ (3.5) \quad &\leq e^{T-t}(-b_0 + b_1) + U \circ (U')^{-1}(e^{T-t}) \\ &\leq -b_0 + b_1 + U \circ (U')^{-1}(1) < 0, \quad (t, x) \in [0, T] \times (0, \infty). \end{aligned}$$

LEMMA 3.2. *We assume (1.6). Then the following assertions are valid:*

$$(3.6) \quad 0 \leq v(s, x) \leq \zeta(s, x).$$



For any  $\varepsilon > 0$ , there exists  $C_\varepsilon > 0$  such that

$$(3.7) \quad \begin{aligned} |v(s, x) - v(r, y)| &\leq C_\varepsilon \{|s - r|^{1/2}(1 + x + y) + |x - y|\} + \varepsilon(1 + x + y), \\ x, y > 0, \quad 0 \leq r \leq s \leq T. \end{aligned}$$

*Proof.* By Itô's formula and (3.5), we have

$$(3.8) \quad \begin{aligned} 0 &\leq \zeta(T, X_T) \\ &= \zeta(s, x) + \int_s^T \left\{ \zeta_t(t, X_t) + [(X_t)^\alpha - \mu X_t - c_t X_t] \zeta_x(t, X_t) \right. \\ &\quad \left. + \frac{1}{2} \sigma^2 X_t^2 \zeta_{xx}(t, X_t) \right\} dt - \int_s^T \sigma X_t \zeta_x(t, X_t) dB_t \\ &\leq \zeta(s, x) - \int_s^T U(c_t X_t) dt - \int_s^T \sigma X_t e^{T-t} dB_t \quad a.s. \end{aligned}$$

By (2.7), we note that  $\{\int_s^t \sigma X_r e^{-r} dB_r\}$  is a martingale. Therefore, we deduce (3.6). Now, by (3.3), we have

$$(3.9) \quad \begin{aligned} |v(s, x) - v(r, y)| &\leq \sup_{c \in \mathcal{A}} E \left[ \left| \int_s^T U(c_t X_t) dt - \int_r^T U(c_t Y_t) dt \right| \right] \\ &\leq \sup_{c \in \mathcal{A}} E \left[ \int_s^T |U(c_t X_t) - U(c_t Y_t)| dt \right] + \sup_{c \in \mathcal{A}} E \left[ \int_r^s U(c_t Y_t) dt \right] \\ &\equiv J_1 + J_2, \end{aligned}$$

where  $\{Y_t\}$  denotes the solution of (3.4) with  $Y_r = y$ . By (2.9) and Young's inequality, choosing a suitable constant  $\delta > 0$  for any  $\varepsilon' > 0$ , we note that

$$\begin{aligned} E[|X_t - Y_t|] &\leq C \left\{ \frac{1 - \alpha}{\delta} (|x - y|^{1-\alpha})^{1/(1-\alpha)} + \alpha \delta (1 + x^\alpha + y^\alpha)^{1/\alpha} \right\} \\ &= C_{\varepsilon'} |x - y| + \varepsilon' (1 + x + y). \end{aligned}$$

Also, by (2.7),

$$E[X_t] \leq C(1 + x).$$

Hence, by (1.7),

$$\begin{aligned} J_1 &\leq \sup_{c \in \mathcal{A}} E \left[ \int_0^T \{C_\varepsilon |X_t - Y_t| + \varepsilon\} dt \right] \\ &\leq C_\varepsilon T \{C_{\varepsilon'} |x - y| + \varepsilon' (1 + x + y)\} + \varepsilon T. \end{aligned}$$

By the same calculation as (3.8), taking into account (2.7) and (2.8), we get

$$\begin{aligned} J_2 &\leq E[\zeta(r, Y_r) - \zeta(s, Y_s)] \\ &\leq E[|\zeta(r, Y_r) - \zeta(r, Y_s)|] + E[|\zeta(r, Y_s) - \zeta(s, Y_s)|] \\ &\leq e^T \{E[|Y_s - Y_r|] + E[|s - r| |Y_s + b_0|]\} \\ &\leq C |s - r|^{1/2} (1 + y). \end{aligned}$$

Therefore, we deduce (3.7).

**THEOREM 3.3.** *We assume (1.6). Then the value function  $v$  of (3.3) is a viscosity solution of (3.1).*

*Proof.* By Lemma 3.2, we see that  $v \in C([0, T] \times (0, \infty))$ , and by (3.3),  $v(T, x) = 0$ . According to [2], the viscosity property of  $v$  follows from the dynamic programming principle for  $v$ ; that is,

$$(3.10) \quad v(s, x) = \sup_{c \in \mathcal{A}} E \left[ \int_s^\tau U(c_t X_t) dt + v(\tau, X_\tau) \right] \quad \forall (s, x) \in [0, T] \times (0, \infty)$$

for any  $\tau \in [s, T]$ , where the supremum is taken over all systems  $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}; \{B_t\}, \{c_t\})$ .

Let  $\bar{v}$  be the right-hand side of (3.10) and set  $J_{(s,x)}(c) = E[\int_s^T U(c_t X_t) dt]$ . For each  $c = \{c_t\} \in \mathcal{A}$ , let  $\tilde{X}_t = X_{t+\tau}$  and  $\tilde{B}_t = B_{t+\tau} - B_\tau$ . Then we have

$$d\tilde{X}_t = (\tilde{X}_t^\alpha - \mu \tilde{X}_t - \tilde{c}_t \tilde{X}_t) dt - \sigma \tilde{X}_t d\tilde{B}_t, \quad t \in (0, T - \tau], \quad \tilde{X}_0 = X_\tau,$$

where  $\tilde{c} = \{\tilde{c}_t\}$  is the shifted process of  $c$  by  $\tau$ , i.e.,  $\tilde{c}_t = c_{t+\tau}$ . By (3.4), we see that

$$E \left[ \int_\tau^T U(c_t X_t) dt | \mathcal{F}_\tau \right] = E \left[ \int_0^{T-\tau} U(\tilde{c}_t \tilde{X}_t) dt | \mathcal{F}_\tau \right] = J_{(\tau, X_\tau)}(c) \quad \text{a.s.}$$

with respect to the conditional probability measure  $P(\cdot | \mathcal{F}_\tau)$ . Hence

$$\begin{aligned} J_{(s,x)}(c) &= E \left[ \int_s^\tau U(c_t X_t) dt + \int_\tau^T U(c_t X_t) dt \right] \\ &\leq E \left[ \int_s^\tau U(c_t X_t) dt + v(\tau, X_\tau) \right]. \end{aligned}$$

Taking the supremum, we deduce  $v \leq \bar{v}$ .

Conversely, let  $\{S_j : j = 1, \dots, n+1\}$  be a sequence of disjoint subsets of  $(0, \infty)$  such that

$$\text{diam}(S_j) < \delta, \quad \bigcup_{j=1}^n S_j = (0, R), \quad \text{and} \quad S_{n+1} = [R, \infty)$$

for  $\delta, R > 0$  chosen later. For any  $\varepsilon > 0$ , we take  $x_j \in S_j$  and  $c^{(j)} \in \mathcal{A}$  such that

$$(3.11) \quad v(\tau, x_j) - \varepsilon \leq J_{(\tau, x_j)}(c^{(j)}), \quad j = 1, \dots, n+1.$$

By the same argument as (3.7), we note that

$$|J_{(\tau, x)}(c) - J_{(\tau, y)}(c)| + |v(\tau, x) - v(\tau, y)| \leq C_\varepsilon |x - y| + \frac{\varepsilon}{4}(1 + x + y), \quad x, y > 0, \quad c \in \mathcal{A},$$

for some constant  $C_\varepsilon > 0$ . We choose  $0 < \delta < 1$  such that  $C_\varepsilon \delta < \varepsilon/2$ . Then, we have

$$|J_{(\tau, x)}(c^{(j)}) - J_{(\tau, y)}(c^{(j)})| + |v(\tau, x) - v(\tau, y)| \leq \varepsilon(1 + x), \quad x, y \in S_j,$$

from which

$$J_{(\tau, X_\tau)}(c^{(j)}) \geq J_{(\tau, x_j)}(c^{(j)}) - \varepsilon(1 + X_\tau) \quad \text{if} \quad X_\tau \in S_j, \quad j = 1, \dots, n.$$

Hence

$$\begin{aligned}
 J_{(\tau, X_\tau)}(c^{(j)}) &= J_{(\tau, X_\tau)}(c^{(j)}) - J_{(\tau, x_j)}(c^{(j)}) + J_{(\tau, x_j)}(c^{(j)}) \\
 (3.12) \quad &\geq -\varepsilon(1 + X_\tau) + v(\tau, x_j) - \varepsilon \\
 &\geq -2\varepsilon(1 + X_\tau) + v(\tau, X_\tau) - \varepsilon \quad \text{if } X_\tau \in S_j, \quad j = 1, \dots, n.
 \end{aligned}$$

By definition, we find  $c \in \mathcal{A}$  such that

$$\bar{v}(s, x) - \varepsilon \leq E \left[ \int_s^\tau U(c_t X_t) dt + v(\tau, X_\tau) \right].$$

As in the proof of Theorem IV-1.1 of [3], we can take  $c, c^{(j)}$  on the same probability space. Define

$$c_t^\tau = c_t 1_{\{t < \tau\}} + c_t^{(j)} 1_{\{\tau \leq t \leq T\}} \quad \text{if } X_\tau \in S_j, \quad j = 1, \dots, n+1.$$

It is easy to see that  $\{c_t^\tau\}$  belongs to  $\mathcal{A}$ . Let  $\{X_t^\tau\}$  be the solution of

$$dX_t^\tau = [(X_t^\tau)^\alpha - \mu X_t^\tau - c_t^\tau X_t^\tau] dt - \sigma X_t^\tau dB_t, \quad s < t \leq T, \quad X_s^\tau = x > 0.$$

Clearly,  $X_t^\tau = X_t$  a.s. if  $s \leq t < \tau$ . Further, for each  $j = 1, \dots, n+1$ , we have on  $\{X_\tau \in S_j\}$ ,

$$X_r^\tau = X_\tau + \int_\tau^r [(X_t^\tau)^\alpha - \mu X_t^\tau - c_t^\tau X_t^\tau] dt - \int_\tau^r \sigma X_t^\tau dB_t, \quad \tau < r \leq T \quad \text{a.s.}$$

Hence  $X_t^\tau = X_t^{(j)}$  for all  $t \in [\tau, T]$  a.s. on  $\{X_\tau \in S_j\}$ , where  $\{X_t^{(j)}\}$  denotes the solution of

$$dX_t^{(j)} = [(X_t^{(j)})^\alpha - \mu X_t^{(j)} - c_t^{(j)} X_t^{(j)}] dt - \sigma X_t^{(j)} dB_t, \quad \tau < t \leq T, \quad X_\tau^{(j)} = X_\tau.$$

Thus, we get

$$\begin{aligned}
 J_{(\tau, X_\tau)}(c^\tau) &= E \left[ \int_\tau^T U(c_t^\tau X_t^\tau) dt | \mathcal{F}_\tau \right] \\
 (3.13) \quad &= E \left[ \int_\tau^T U(c_t^{(j)} X_t^{(j)}) dt | \mathcal{F}_\tau \right] \\
 &= J_{(\tau, X_\tau)}(c^{(j)}) \quad \text{a.s. on } \{X_\tau \in S_j\}.
 \end{aligned}$$

Next, taking into account (3.6) and (2.7), we choose  $R > 0$  such that

$$\begin{aligned}
 \sup_{c \in \mathcal{A}} E[v(\tau, X_\tau) 1_{\{X_\tau \geq R\}}] &\leq \sup_{c \in \mathcal{A}} e^T E[(X_\tau + b_0) 1_{\{X_\tau \geq R\}}] \\
 (3.14) \quad &\leq \sup_{c \in \mathcal{A}} \frac{e^T}{R} E[X_\tau^2 + b_0 X_\tau] \\
 &\leq \frac{e^T}{R} \{(1 + b_0)C(1 + x^2) + b_0\} < \varepsilon.
 \end{aligned}$$

By (3.11)–(3.14) and (2.7), we have

$$\begin{aligned}
 E \left[ \int_{\tau}^T U(c_t^{\tau} X_t^{\tau}) dt \right] &= E \left[ E \left[ \int_{\tau}^T U(c_t^{\tau} X_t^{\tau}) dt | \mathcal{F}_{\tau} \right] \right] \\
 &= E[J_{(\tau, X_{\tau})}(c^{\tau})] \\
 &= E \left[ \sum_{j=1}^{n+1} J_{(\tau, X_{\tau})}(c^{(j)}) 1_{\{X_{\tau} \in S_j\}} \right] \\
 &\geq E \left[ \sum_{j=1}^n \{v(\tau, X_{\tau}) - 3\varepsilon(1 + X_{\tau})\} 1_{\{X_{\tau} \in S_j\}} \right] \\
 &\geq E[\{v(\tau, X_{\tau}) - v(\tau, X_{\tau}) 1_{\{X_{\tau} \geq R\}}\}] - 3\varepsilon E[1 + X_{\tau}] \\
 &\geq E[v(\tau, X_{\tau})] - \varepsilon - 3\varepsilon\{2 + C(1 + x^2)\}.
 \end{aligned}$$

Thus

$$\begin{aligned}
 v(s, x) &\geq E \left[ \int_s^{\tau} U(c_t^{\tau} X_t^{\tau}) dt + \int_{\tau}^T U(c_t^{\tau} X_t^{\tau}) dt \right] \\
 &\geq E \left[ \int_s^{\tau} U(c_t X_t) dt + v(\tau, X_{\tau}) \right] - 4\varepsilon\{2 + C(1 + x^2)\} \\
 &\geq \bar{v}(s, x) - \varepsilon - 4\varepsilon\{2 + C(1 + x^2)\}.
 \end{aligned}$$

Therefore, letting  $\varepsilon \rightarrow 0$ , we obtain  $\bar{v} \leq v$ . The proof is complete.

**4. Classical solutions.** In this section, we study the classical solutions of the HJB equation (3.1) with the terminal condition (3.2). First, for any interval  $[\xi_1, \xi_2]$  with  $\xi_1 > 0$ , we consider the parabolic equation

$$(4.1) \quad u_t + \frac{1}{2}\sigma^2 x^2 u_{xx} + (x^{\alpha} - \mu x)u_x + \tilde{U}(x, u_x) = 0, \quad 0 \leq t < T, \quad \xi_1 < x < \xi_2,$$

with the (parabolic) boundary condition

$$(4.2) \quad u(T, x) = v(T, x) = 0, \quad x \in [\xi_1, \xi_2],$$

$$(4.3) \quad u(t, \xi_1) = v(t, \xi_1), \quad u(t, \xi_2) = v(t, \xi_2), \quad t \in [0, T).$$

**THEOREM 4.1.** *Let  $u_i \in C([0, T] \times [\xi_1, \xi_2])$ ,  $i = 1, 2$ , be two viscosity solutions of (4.1)–(4.3). Then, under (1.6), we have  $u_1 = u_2$ .*

*Proof.* It is sufficient to show that  $u_1 \leq u_2$ . Suppose there exists  $(t_0, x_0) \in [0, T) \times (\xi_1, \xi_2)$  such that  $u_1(t_0, x_0) - u_2(t_0, x_0) > 0$ . Then we find  $\eta > 0$  such that

$$\varrho := \sup_{(t, x) \in (0, T) \times (\xi_1, \xi_2)} \left\{ u_1(t, x) - u_2(t, x) - 2\eta \frac{1}{t} \right\} > 0.$$

By boundedness, we have

$$(4.4) \quad u_1(t, x) - u_2(t, x) - 2\eta \frac{1}{t} \rightarrow -\infty \quad \text{uniformly in } x \text{ as } t \downarrow 0.$$

Thus, by (4.2) and (4.3), there exists  $(\bar{t}, \bar{x}) \in (0, T) \times (\xi_1, \xi_2)$  such that

$$\varrho = u_1(\bar{t}, \bar{x}) - u_2(\bar{t}, \bar{x}) - 2\eta \frac{1}{\bar{t}}.$$

Define

$$\Psi_k(t, x, y) = u_1(t, x) - u_2(t, y) - \frac{k}{2}|x - y|^2 - 2\eta \frac{1}{t}$$

for  $k > 0$ . By (4.2) and (4.4), there exists  $(t_k, x_k, y_k) \in (0, T) \times [\xi_1, \xi_2]^2$  such that

$$(4.5) \quad \Psi_k(t_k, x_k, y_k) = \sup \Psi_k(t, x, y) \geq \Psi_k(\bar{t}, \bar{x}, \bar{x}) = \varrho,$$

from which

$$\frac{k}{2}|x_k - y_k|^2 < u_1(t_k, x_k) - u_2(t_k, y_k) - 2\eta \frac{1}{t_k}.$$

Thus

$$(4.6) \quad |x_k - y_k| \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

By the definition of  $(t_k, x_k, y_k)$ , we have

$$\Psi_k(t_k, x_k, y_k) \geq \Psi_k(t_k, x_k, x_k).$$

Hence, by uniform continuity,

$$(4.7) \quad \frac{k}{2}|x_k - y_k|^2 \leq u_2(t_k, x_k) - u_2(t_k, y_k) \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

By (4.6), (4.5), and (4.3), and by extracting a subsequence, we have

$$t_k \rightarrow \tilde{t} \in (0, T), \quad (x_k, y_k) \rightarrow (\tilde{x}, \tilde{x}) \in (\xi_1, \xi_2)^2 \quad \text{as } k \rightarrow \infty.$$

Now, we may consider that  $(t_k, x_k, y_k) \in (0, T) \times (\xi_1, \xi_2)^2$ . Applying Ishii's lemma [1, Thm. 8.3] to

$$\Psi_k(t, x, y) = w_1(t, x) - w_2(t, y) - \frac{k}{2}|x - y|^2,$$

we obtain  $a, b \in \mathbf{R}$  and  $X, Y \in \mathbf{R}$  such that

$$(4.8) \quad \begin{aligned} (a, k(x_k - y_k), X) &\in \bar{\mathcal{P}}^{2,+} w_1(t_k, x_k), \\ (b, k(x_k - y_k), Y) &\in \bar{\mathcal{P}}^{2,-} w_2(t_k, y_k), \\ a - b = 0, \quad \begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} &\leq 3k \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \end{aligned}$$

where  $w_1(t, x) = u_1(t, x) - \eta/t$  and  $w_2(t, y) = u_2(t, y) + \eta/t$ . From the definition of  $\mathcal{P}^{2,+} u_1(t_k, x_k)$ ,  $\mathcal{P}^{2,-} u_2(t_k, y_k)$ , it follows that

$$\mathcal{P}^{2,+}u_1(t, x) = \left\{ (\hat{a}, \lambda, \hat{X}) + \eta \left( \frac{-1}{t^2}, 0, 0 \right) : (\hat{a}, \lambda, \hat{X}) \in \mathcal{P}^{2,+}w_1(t, x) \right\},$$

$$\mathcal{P}^{2,-}u_2(t, x) = \left\{ (\hat{a}, \lambda, \hat{X}) - \eta \left( \frac{-1}{t^2}, 0, 0 \right) : (\hat{a}, \lambda, \hat{X}) \in \mathcal{P}^{2,-}w_2(t, x) \right\}.$$

Hence

$$(\bar{a}, \lambda_1, \bar{X}) := (a, k(x_k - y_k), X) + \eta \left( \frac{-1}{t_k^2}, 0, 0 \right) \in \bar{\mathcal{P}}^{2,+}u_1(t_k, x_k),$$

$$(\bar{b}, \lambda_2, \bar{Y}) := (b, k(x_k - y_k), Y) - \eta \left( \frac{-1}{t_k^2}, 0, 0 \right) \in \bar{\mathcal{P}}^{2,-}u_2(t_k, y_k).$$

By Definition 3.1,

$$\begin{aligned} \bar{a} + \frac{1}{2}\sigma^2 x_k^2 \bar{X} + (x_k^\alpha - \mu x_k)\lambda_1 + \tilde{U}(x_k, \lambda_1) &\geq 0, \\ \bar{b} + \frac{1}{2}\sigma^2 y_k^2 \bar{Y} + (y_k^\alpha - \mu y_k)\lambda_2 + \tilde{U}(y_k, \lambda_2) &\leq 0. \end{aligned}$$

Putting these inequalities together, we get

$$\begin{aligned} 2\eta \frac{1}{t_k^2} &\leq \frac{1}{2}\sigma^2 (x_k^2 \bar{X} - y_k^2 \bar{Y}) + \{(x_k^\alpha - \mu x_k)\lambda_1 - (y_k^\alpha - \mu y_k)\lambda_2\} \\ &\quad + |\tilde{U}(x_k, \lambda_1) - \tilde{U}(y_k, \lambda_2)| \\ &\equiv I_1 + I_2 + I_3. \end{aligned}$$

By (4.8) and (4.7), it is clear that

$$I_1 = \frac{\sigma^2}{2}(x_k^2 X - y_k^2 Y) \leq \frac{\sigma^2}{2}3k|x_k - y_k|^2 \quad \rightarrow \quad 0 \quad \text{as } k \rightarrow \infty.$$

Since  $x^\alpha$  is Lipschitz on  $[\xi_1, \xi_2]$ , we see by (4.7) that

$$I_2 = k\{(x_k^\alpha - y_k^\alpha) - \mu(x_k - y_k)\}(x_k - y_k) \quad \rightarrow \quad 0 \quad \text{as } k \rightarrow \infty.$$

By (1.7), (4.6), and (4.7), we have

$$\begin{aligned} I_3 &\leq \max_{0 \leq c \leq 1} |U(cx_k) - U(cy_k)| + |x_k \lambda_1 - y_k \lambda_2| \\ &\leq C_\varepsilon |x_k - y_k| + \varepsilon + k|x_k - y_k|^2 \\ &\rightarrow 0 \quad \text{as } k \rightarrow \infty \quad \text{and } \varepsilon \rightarrow 0. \end{aligned}$$

Consequently, we deduce

$$2\eta \frac{1}{T^2} \leq 2\eta \frac{1}{t^2} \leq 0,$$

which is a contradiction.

**THEOREM 4.2.** *We assume (1.6). Then there exists a solution  $v \in C^{1,2}([0, T] \times (0, \infty)) \cap C([0, T] \times (0, \infty))$  of (3.1), (3.2).*

*Proof.* By (1.6), we have

$$U(0) \leq U'(c)(-c) + U(c) \quad \forall c > 0.$$

Then for any  $\xi_1 > 0$ ,

$$C_0 := \sup_{0 < c \leq \xi_1} cU'(c) \leq U(\xi_1) < \infty.$$

Hence, for  $x_1, x_2 \in [\xi_1, \xi_2]$  where  $\xi_2 > \xi_1$ ,

$$|U(cx_1) - U(cx_2)| \leq cU'(c\xi_1)|x_1 - x_2| \leq \frac{C_0}{\xi_1}|x_1 - x_2|, \quad 0 \leq c \leq 1.$$

Thus, for  $p_1, p_2 \in \mathbf{R}$ ,

$$\begin{aligned} |\tilde{U}(x_1, p_1) - \tilde{U}(x_2, p_2)| &\leq \max_{0 \leq c \leq 1} |U(cx_1) - U(cx_2)| + |x_1 p_1 - x_2 p_2| \\ &\leq \left( \frac{C_0}{\xi_1} + |p_1| \right) |x_1 - x_2| + \xi_2 |p_1 - p_2|. \end{aligned}$$

According to [4], by uniform ellipticity, there exists a unique solution  $u \in C([0, T] \times [\xi_1, \xi_2]) \cap C^{1,2}([0, T] \times (\xi_1, \xi_2))$  of (4.1)–(4.3). Clearly,  $v$  is a viscosity solution of (4.1)–(4.3). By Theorem 4.1, we have  $u = v$  and  $v$  is smooth. Since  $\xi_1, \xi_2$  are arbitrary, we obtain the assertion.

**COROLLARY 4.3.** *We make the assumption of Theorem 4.2. Then there exists a solution  $V \in C^{1,2}([0, T] \times (0, \infty)^2)$  of (1.5).*

*Proof.* The proof follows from Theorem 4.2 and (2.1).

**5. Concavity.** In this section, we study the concavity of the solution  $v$  to (3.1), (3.2).

**THEOREM 5.1.** *We assume (1.6). Then  $v(s, x)$  is concave in  $x \in (0, \infty)$  for each  $s \in [0, T]$ . In addition, we have*

$$(5.1) \quad v_x(s, x) > 0 \quad \text{for } x > 0.$$

*Proof.* Let  $x_i > 0, i = 1, 2$ , and  $0 \leq \theta \leq 1$ . For any  $\varepsilon > 0$ , there exists  $c^{(i)} \in \mathcal{A}$  such that

$$v(s, x_i) - \varepsilon < E \left[ \int_s^T U(c_t^{(i)} X_t^{(i)}) dt \right],$$

where  $\{X_t^{(i)}\}$  denotes the solution of (3.4) corresponding to  $c^{(i)}$  with  $X_s^{(i)} = x_i$  on the same probability space. We set

$$\bar{c}_t = \frac{\theta c_t^{(1)} X_t^{(1)} + (1 - \theta) c_t^{(2)} X_t^{(2)}}{\theta X_t^{(1)} + (1 - \theta) X_t^{(2)}},$$

which belongs to  $\mathcal{A}$ . Define  $\{\bar{X}_t\}$  and  $\{\tilde{X}_t\}$  by

$$\begin{aligned} d\bar{X}_t &= [(\bar{X}_t)^\alpha - \mu \bar{X}_t - \bar{c}_t \bar{X}_t] dt - \sigma \bar{X}_t dB_t, \quad s < t \leq T, \quad \bar{X}_s = \theta x_1 + (1 - \theta) x_2, \\ \tilde{X}_t &= \theta X_t^{(1)} + (1 - \theta) X_t^{(2)}. \end{aligned}$$

By concavity

$$\tilde{X}_r \leq \theta x_1 + (1 - \theta)x_2 + \int_s^r [(\tilde{X}_t)^\alpha - \mu \tilde{X}_t - \bar{c}_t \tilde{X}_t] dt - \int_s^r \sigma \tilde{X}_t dB_t \quad \text{a.s.}$$

By the comparison theorem, we have

$$\tilde{X}_t \leq \bar{X}_t, \quad t \in (s, T] \quad \text{a.s.}$$

Thus,

$$\begin{aligned} v(s, \theta x_1 + (1 - \theta)x_2) &\geq E \left[ \int_s^T U(\bar{c}_t \bar{X}_t) dt \right] \geq E \left[ \int_s^T U(\bar{c}_t \tilde{X}_t) dt \right] \\ &= E \left[ \int_s^T U(\theta c_t^{(1)} X_t^{(1)} + (1 - \theta) c_t^{(2)} X_t^{(2)}) dt \right] \\ &\geq \theta E \left[ \int_s^T U(c_t^{(1)} X_t^{(1)}) dt \right] + (1 - \theta) E \left[ \int_s^T U(c_t^{(2)} X_t^{(2)}) dt \right] \\ &> \theta v(s, x_1) + (1 - \theta)v(s, x_2) - \varepsilon. \end{aligned}$$

Therefore, letting  $\varepsilon \rightarrow 0$ , we obtain the concavity of  $v$ .

To prove (5.1), by Theorem 4.2, we note that  $v$  is smooth. By nonnegativity and concavity, we see that

$$v_x(s, x) \geq 0, \quad x > 0,$$

for every  $s \in [0, T)$ . Suppose that  $v_x(s, x_0) = 0$  for some  $x_0 > 0$ . Then,  $v_x(s, x) = 0$  for all  $x \geq x_0$ . Hence  $v(s, x)$  can be written as  $v(s, x) = h(s)$  for  $x \geq x_0$ . By (3.1), we have

$$U(x) = -v_t(s, x) = -h'(s), \quad x \geq x_0.$$

This is contrary to (1.6). Therefore we obtain (5.1).

**6. Optimal policies.** We give a synthesis of the optimal policy  $c^* = \{c_t^*\}$  for the optimization problem (1.4) subject to (1.1)–(1.3). We consider the SDE

$$(6.1) \quad dX_t^* = [(X_t^*)^\alpha - \mu X_t^* - \gamma(t, X_t^*) X_t^*] dt - \sigma X_t^* dB_t, \quad 0 < t \leq T, \quad X_0^* = x > 0,$$

where  $\gamma(t, x) = I(x, v_x(t, x))$  and  $I(x, p)$  denotes the maximizer of (2.4) for  $x, p > 0$ , i.e.,

$$(6.2) \quad I(x, p) = \begin{cases} (U')^{-1}(p)/x & \text{if } U'(x) \leq p, \\ 1 & \text{otherwise.} \end{cases}$$

LEMMA 6.1. *Under (1.6), there exists a unique positive solution  $\{X_t^*\}$  of (6.1).*

*Proof.* By (5.1), we notice that  $\gamma(t, x)$  is well defined. Let  $\{N_t\}$  be the solution of (2.6) corresponding to  $c_t = 0$ . Define the probability measure  $\hat{P}$  on  $(\Omega, \mathcal{F}_T, P)$  by

$$d\hat{P}/dP = \exp \left\{ \int_0^T \gamma(s, N_s)/\sigma dB_s - \frac{1}{2} \int_0^T (\gamma(s, N_s)/\sigma)^2 ds \right\}.$$



By the very definition (6.2) we have  $0 \leq \gamma(t, x) \leq 1$ ; so Girsanov's theorem yields that

$$\hat{B}_t := B_t - \int_0^t \gamma(s, N_s)/\sigma \, ds \quad \text{is a Brownian motion on } (\Omega, \mathcal{F}_T, \hat{P}).$$

Hence,

$$dN_t = [(N_t)^\alpha - \mu N_t - \gamma(t, N_t)N_t]dt - \sigma N_t d\hat{B}_t \quad \text{under } \hat{P}.$$

Thus, (6.1) admits a positive weak solution.

Now, by (6.2), we have

$$\gamma(t, x)x = \min\{(U')^{-1} \circ v_x(t, x), x\}.$$

Also, by (1.6) and concavity,

$$\frac{\partial}{\partial x}(U')^{-1} \circ v_x(t, x) = \frac{v_{xx}(t, x)}{U'' \circ (U')^{-1} \circ v_x(t, x)} \geq 0.$$

Thus,  $\gamma(t, x)x$  is nondecreasing on  $(0, \infty)$  for each  $t$ . We rewrite (6.1) in the form of (2.11). Then, we see that the pathwise uniqueness holds for (6.1). Therefore, by the Yamada–Watanabe theorem [3], we deduce that (6.1) admits a unique strong solution  $\{X_t^*\}$ .

**THEOREM 6.2.** *We assume (1.6). Then the optimal consumption policy  $\{c_t^*\}$  is given by the feedback form*

$$(6.3) \quad c_t^* = c^*(t, z_t^*, y_t),$$

where  $c^*(t, z, y) = I(z/y, yV_z(t, z, y))$  and  $\{z_t^*\}$  is the unique solution of

$$(6.4) \quad \dot{z}_t^* = F(z_t^*, y_t) - \nu z_t^* - c_t^* z_t^*, \quad 0 < t \leq T, \quad z_0^* = z > 0.$$

*Proof.* We set  $X_t^* = z_t^*/y_t$ . By Itô's formula and (2.2), we see that  $X_t^*$  solves (6.1). Therefore, by Lemma 6.1, there exists a unique positive solution  $\{z_t^*\}$  of (6.4).

By Theorems 4.2 and 5.1, we note that

$$0 < v_x(t, x)x \leq v(t, x) - v(t, 0+) \leq v(t, x), \quad x > 0.$$

Hence, by (3.6) and (2.7),

$$\begin{aligned} E \left[ \int_0^T \{v_x(s, X_s^*)X_s^*\}^2 ds \right] &\leq E \left[ \int_0^T \{v(s, X_s^*)\}^2 ds \right] \\ &\leq E \left[ \int_0^T \zeta(s, X_s^*)^2 ds \right] < \infty. \end{aligned}$$

By (2.2), this yields that  $\{\int_0^t \sigma y_s V_y(s, z_s^*, y_s) dB_s\}$  is a martingale. By (1.5) and (6.2),  $c^*$  satisfies

$$V_t + \frac{1}{2}\sigma^2 y^2 V_{yy} + nyV_y + \{F(z, y) - \nu z\}V_z + \{U(c^*z/y) - c^*zV_z\} = 0.$$

Applying Itô's formula to (1.1) and (6.4), we get

$$E[V(T, z_T^*, y_T)] = V(0, z, y) - E \left[ \int_0^T U(c_t^* z_t^*/y_t) dt \right],$$

which implies

$$E \left[ \int_0^T U(c_t^* z_t^*/y_t) dt \right] = V(0, z, y).$$

By the same calculation as above, we obtain

$$E \left[ \int_0^T U(c_t z_t/y_t) dt \right] \leq V(0, z, y)$$

for any  $c \in \mathcal{A}$ . The proof is complete.

*Remark 6.3.* From the proof of Theorem 6.2, it follows that

$$\sup_{c \in \mathcal{A}} E \left[ \int_s^T U(c_t z_t/y_t) dt \right] = V(s, z, y).$$

Thus, under (1.6), we see that the smooth solution  $V$  of the HJB equation (1.5) is unique. Furthermore, let  $u$  be the solution of (4.1) on the entire domain  $[0, T] \times (0, \infty)$  with  $u(T, x) = 0$ ,  $x > 0$ . Setting  $x = z/y$  and  $V(t, z, y) = u(t, z/y)$  for  $z, y > 0$ , by (2.2), we have that  $V$  satisfies (1.5). Therefore, we obtain the uniqueness of  $u$ .

**Acknowledgments.** The authors wish to thank the two anonymous referees for their comments which led to an improved version of this paper.

#### REFERENCES

- [1] M. G. CRANDALL, H. ISHII, AND P. L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.
- [2] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
- [3] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.
- [4] O. A. LADYŽENSKAYA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasi-Linear Equations of Parabolic Type*, Transl. Math. Monogr., 23, AMS, Providence, RI, 1968.
- [5] C. LIU AND H. MORIMOTO, *Optimal consumption for the stochastic Ramsey problem with non-Lipschitz coefficients*, submitted.
- [6] H. MORIMOTO, *Optimal consumption models in economic growth*, J. Math. Anal. Appl., 337 (2008), pp. 480–492.
- [7] R. C. MERTON, *An asymptotic theory of growth under uncertainty*, Rev. Econ. Studies, 42 (1975), pp. 375–393.

## ANALYTICAL PARAMETERIZATION OF ROTORS AND PROOF OF A GOLDBERG CONJECTURE BY OPTIMAL CONTROL THEORY\*

TÉRENCE BAYEN†

**Abstract.** Curves which can be rotated freely in an  $n$ -gon (that is, a regular polygon with  $n$  sides) so that they always remain in contact with every side of the  $n$ -gon are called rotors. Using optimal control theory, we prove that the rotor with minimal area consists of a finite union of arcs of circles. Moreover, the radii of these arcs are exactly the distances of the diagonals of the  $n$ -gon from the parallel sides. Finally, using the extension of Noether's theorem to optimal control (as performed in [D. F. M. Torres, *WSEAS Trans. Math.*, 3 (2004), pp. 620–624]), we show that a minimizer is necessarily a regular rotor, which proves a conjecture formulated in 1957 by Goldberg (see [M. Goldberg, *Amer. Math. Monthly*, 64 (1957), pp. 71–78]).

**Key words.** shape optimization, convexity, constant width bodies, rotors, support function, optimal control, Pontryagin maximum principle, switching point, bang control, Noether theorem

**AMS subject classifications.** 49J15, 49Q10, 78M50, 80M50

**DOI.** 10.1137/070705325

**1. Introduction.** In this paper, we investigate properties of *rotors*, that is, convex curves that can be freely rotated inside a regular polygon  $P_n$  with  $n$  sides,  $n \geq 3$ , while remaining in contact with every side of  $P_n$ . When  $n = 4$ ,  $P_4$  is a square of side  $\alpha$ , and a rotor of  $P_4$  is called a curve of constant width  $\alpha$  or an orbiform. When  $n = 3$ ,  $P_3$  is an equilateral triangle, and a rotor of  $P_3$  is called a  $\Delta$ -curve. There are infinitely many such curves besides the circle (see section 2).

Orbiforms have been studied geometrically since the 19th century (see [5], [22], [24], [27], [34]). In particular, Reuleaux's name is attached to those orbiforms obtained by intersecting a finite number of discs of equal radii  $\alpha$ . The Reuleaux triangle is the most famous of these orbiforms: it consists of the intersection of three circles of radius one and whose centers are on the vertices of an equilateral triangle of side one. Orbiforms have many interesting properties and applications in mechanics (see [5], [6], [7], [23], [24], [25], [34]). For example, Reuleaux triangles are used in boring square holes, and they are also part of the Wankel engine used by Japan's Mazda cars. Nowadays, the study of rotors is potentially interesting in mechanics for the design of engines or propellers, for example, in the Navy.

An interesting shape optimization problem consists in determining the convex body maximizing or minimizing the area in the class of rotors. It is easy to show that the disc always has maximal area in this class. This is a consequence of the isoperimetric inequality, as all rotors have the same perimeter (see Barbier's theorem in section 2.3). The question of finding a rotor of least area is more difficult. First, notice that the problem of minimizing the area is well posed, as rotors are convex bodies (see section 2.2). This question has been solved for  $n = 4$  (that is, in the case of orbiforms) by Blaschke using the mixed-volume (see [5]) and Lebesgue (see [22]). They show that the Reuleaux triangle has the least area in the class of constant width bodies of  $\mathbb{R}^2$ . Fujiwara has given the first analytic proof of this result (see [11], [12]).

---

\*Received by the editors October 15, 2007; accepted for publication (in revised form) April 30, 2008; published electronically January 7, 2009.

<http://www.siam.org/journals/sicon/47-6/70532.html>

†Laboratoire des Signaux et Systèmes, Ecole Supérieure d'Electricité, 91192 Gif-sur-Yvette, France (terence.bayen@lss.supelec.fr).

More recently, Harrell gave a modern proof using minimization under constraints (see [18]). The study of these problems in  $\mathbb{R}^2$  is useful for extensions in  $\mathbb{R}^3$  and in the domain of spectral analysis. For example, the problem of finding a constant width body of minimal volume in  $\mathbb{R}^3$  has recently been investigated (see [4], [20]). The optimization of eigenvalues with respect to the domain  $\Omega$  is also an intense field of research (see [19] for an overview of many spectral problems involving convexity). These questions require a careful study of dimension 2.

The  $\Delta$ -curves have many similar geometrical properties to the orbiforms (see [7], [34]). Fujiwara gave an analytic proof in [11] that, among all  $\Delta$ -curves inscribed in an equilateral triangle of side one, the one of minimal area is the  $\Delta$ -biangle or lens. It consists of two circular arcs of radius  $\frac{\sqrt{3}}{2}$  and of length  $\frac{\pi}{3}$ . This result was also established by Blaschke and later by Weissbach (see [33]).

Whereas the cases  $n = 3$  and  $n = 4$  have been investigated, the question of finding the rotor of least area for  $n \geq 5$  is open. Standard geometrical proofs cannot be applied in this case (see [13]). In [15] and [16], Goldberg constructs a family of “trammel” rotors in a regular polygon,  $(O_n^{ln\pm 1})_{l \in \mathbb{N}^*}$ , that have  $2(ln \pm 1)$  symmetries, and he conjectured in [15] that the minimizer is a rotor called  $O_n^{n-1}$  obtained for  $l = 1$ . The boundary of a rotor  $O_n^{ln\pm 1}$  consists of a finite union of arcs of circles of different radii  $r_i$  and of equal sectors (see section 2.6). The values  $r_i$  are exactly the distances of the diagonals of the  $n$ -gon from the parallel sides. In this class,  $O_n^{n-1}$  has the minimum number of arcs. An analytic description of these regular rotors is given in [10] by Focke. In 1975, Klötzler made an analytic study of the minimization problem using optimal control theory (see [2], [3], [21]). He showed in [21] that a minimizer consists of a union of arcs of circles of radii  $r_i$ , but he failed to prove that a minimizer is in the class  $(O_n^{ln\pm 1})_{l \in \mathbb{N}^*}$ . His idea consists in reformulating the initial minimization problem into an optimal control problem by choosing the radius of curvature as the control variable. Unfortunately, he seems to prove that the regular rotors  $O_n^{ln\pm 1}$  are local minimizers of the area in the subclass  $\mathcal{R}_n^{ln\pm 1}$  of rotors having the same number of arcs and the same radii of curvature. This result contradicts the one of Firey (see [9]) in the case  $n = 4$ : the author shows that regular Reuleaux polygons with  $N$  sides,  $N \geq 5$ , maximize the area in the class of Reuleaux polygons with the same number of sides. Moreover, in [2], the author performs only convex perturbations of a regular rotor  $O_n^{ln\pm 1}$ . This kind of perturbation increases the area by the concavity of the functional (the Brunn–Minkowski theorem; see [8]). The main difficulty is to consider nonconvex perturbations of those rotors which are not obtained by a strictly convex combination of two rotors.

The aim of the paper is to prove the following theorem conjectured by Goldberg in 1957 (see [15]).

**THEOREM 1.1.** *Among all rotors of a regular polygon  $P_n$  ( $n \geq 3$ ), the one of minimal area is the regular rotor  $O_n^{n-1}$ .*

In section 2, we give an analytic parameterization of a rotor using the support function of a convex body (see [6] or [28] for an overview of the properties of the support function). In section 3, we formulate the minimization problem into an optimal control problem which is similar to the one obtained by Klötzler (see [21]). Indeed, the convexity constraints enable us to choose the radius of curvature of the boundary of a rotor as the control variable. Thanks to this new parameterization, the initial shape optimization is well posed. By the Pontryagin maximum principle (PMP), we show that the extremal trajectories are “bang-bang,” and we determine the corresponding number of switching points. We thus restrict the class of extremal trajectories step by step. Whereas the computation of the extremal trajectories performed by Klötzler is

incomplete (he does not show that the switching points of an extremal trajectory are equidistant), we prove, in section 4, Theorem 1.1 by using an extension of Noether's theorem to optimal control theory provided in [29]. We compute conserved quantities along an extremal trajectory, and thus we can characterize the switching points of an extremal (see section 4). This shows that the rotors corresponding to the extremal trajectories belong to the class  $(O_n^{ln\pm 1})_{l \in \mathbb{N}^*}$ . We then conclude the proof of Goldberg's conjecture by Proposition 2.11. Note that by this proposition, there is no need to examine the optimality of extremal trajectories.

## 2. Construction of a rotor.

**2.1. Support function of a convex body.** A body or a domain in  $\mathbb{R}^N$ ,  $N \geq 2$ , is a nonempty compact connected subset of  $\mathbb{R}^N$ . Let  $K$  be a convex body. The support function of  $K$  is defined as the map  $h_K : \mathbb{R}^N \setminus \{0\} \rightarrow \mathbb{R}$  with

$$h_K(\nu) := \max_{x \in K} x \cdot \nu, \quad \nu \in \mathbb{R}^N \setminus \{0\}.$$

The support function is clearly homogeneous of degree 1. A convex body is uniquely determined by its support function (see [6, p. 29] or [20]). Let  $K$  be a convex body of nonempty interior and assume that the origin is inside  $K$ . Recall that, for a convex body, a hyperplane  $H$  is a hyperplane of support for  $K$  if there exists  $x \in K \cap H$  such that  $K$  is included in one of the half-spaces defined by  $H$ . If  $\nu$  belongs to  $\mathcal{S}^{N-1}$ ,  $h_K(\nu)$  can be interpreted as the distance from the origin to the support hyperplane of  $K$  with normal vector  $\nu$  (see Figure 1). The support function is nonnegative if and only if the origin is inside  $K$ . The next proposition characterizes the degree of regularity of the support function (see [6, p. 28] or [28]).

**PROPOSITION 2.1.** *Let  $K$  be a convex body of  $\mathbb{R}^N$  and  $h_K$  its support function. Then  $h_K$  is of class  $C^1$  if and only if  $K$  is strictly convex.*

From now on, we consider convex bodies in dimension 2. The support function of a convex body  $K$  of  $\mathbb{R}^2$  will be denoted by  $p_K(\theta) := h_K(e^{i\theta})$ ,  $\theta \in \mathbb{R}$ , or  $p(\theta)$  to simplify. The function  $p_K$  is  $2\pi$ -periodic. If  $K$  is a convex body, we denote by  $\partial K$  its boundary. Given  $(z_1, z_2) \in \mathbb{C}^2$ , their scalar product in  $\mathbb{R}^2$  will be written indifferently  $\Re(\overline{z_1} z_2)$  or  $z_1 \cdot z_2$ .

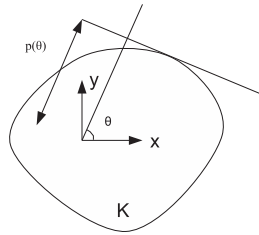


FIG. 1. The support function of a convex body  $K$  is the distance  $p(\theta)$  between the tangent to  $K$  orthogonal to  $(\cos(\theta), \sin(\theta))$  and the origin.

**PROPOSITION 2.2.** *Let  $K$  be a strictly convex body and  $p$  its support function. We assume that the boundary of  $K$ ,  $\partial K$ , is Lipschitz. Then  $\partial K$  can be described by the equations*

$$(2.1) \quad \begin{cases} x(\theta) = p(\theta) \cos(\theta) - \dot{p}(\theta) \sin(\theta), \\ y(\theta) = p(\theta) \sin(\theta) + \dot{p}(\theta) \cos(\theta), \end{cases}$$

where  $\theta \in \mathbb{R}$ .

*Proof.* Let  $\theta$  be in  $[0, 2\pi]$  and  $u_\theta$  be the vector of coordinates  $(\cos(\theta), \sin(\theta))$ . The support function  $p(\theta)$  is defined by

$$p(\theta) := \max_{x \in K} x \cdot u_\theta,$$

and  $p$  is of class  $C^1$  by strict convexity. As  $K$  is compact, the maximum is reached at some point of coordinates  $(x(\theta), y(\theta))$ , and we have

$$(2.2) \quad x(\theta) \cos(\theta) + y(\theta) \sin(\theta) = p(\theta).$$

As the boundary of  $K$  is Lipschitz, the functions  $(x, y)$  are differentiable a.e. (Rademacher's theorem). Moreover, the vector  $u_\theta$  is orthogonal to the support line given by  $X \cos(\theta) + Y \sin(\theta) = 0$ ; hence, we must have

$$(\dot{x}(\theta), \dot{y}(\theta)) \cdot \vec{u}_\theta = 0.$$

By derivation of (2.2), we get

$$-x(\theta) \sin(\theta) + y(\theta) \cos(\theta) = \dot{p}(\theta),$$

which gives (2.1).  $\square$

Equation (2.1) can be rewritten as  $z(\theta) := x(\theta) + iy(\theta) = (p(\theta) + i\dot{p}(\theta))e^{i\theta}$ .

In the following, the space  $C^{1,1}$  denotes the set of maps  $p : \mathbb{R} \rightarrow \mathbb{R}$ , of class  $C^1$ , and such that  $\dot{p}$  is locally Lipschitz.

**PROPOSITION 2.3.** *Let  $K$  be a convex body and  $p$  its support function. We assume that  $p$  is of class  $C^{1,1}$ . Then the radius of curvature  $p + \ddot{p}$  of the boundary  $\partial K$  exists a.e., and, for a.e.  $\theta \in \mathbb{R}$ ,*

$$(2.3) \quad p(\theta) + \ddot{p}(\theta) \geq 0.$$

*Proof.* As  $p$  is of class  $C^{1,1}$ , the functions  $(x(\theta), y(\theta))$  are differentiable a.e., and by standard formulas, the radius of curvature  $f$  of  $\partial K$  is given by  $f = p + \ddot{p}$ . As the body  $K$  is convex,  $f$  must be nonnegative, and consequently we have  $f(\theta) = p(\theta) + \ddot{p}(\theta) \geq 0$  for a.e.  $\theta \in \mathbb{R}$ .  $\square$

If  $K$  is a convex body of support function  $p$  and if  $p$  is of class  $C^{1,1}$ , the tangent vector to  $\partial K$  is given by

$$\dot{z}(\theta) = i(p(\theta) + \ddot{p}(\theta))e^{i\theta}.$$

When  $p + \ddot{p} = 0$  on a set  $A$  of positive measure, then we have  $\dot{z} = 0$ . Geometrically speaking, this means that the boundary  $\partial K$  has a corner: for  $\theta \in A$ , the point  $z(\theta)$  is stationary. For a given function  $f \in L^\infty(\mathbb{R}, \mathbb{R})$  and  $2\pi$ -periodic, we denote by

$$c_1(f) = \frac{1}{2\pi} \int_0^{2\pi} f(\theta) e^{i\theta} d\theta$$

the first Fourier coefficient of  $f$ .

PROPOSITION 2.4. *Let  $f \in L^\infty(\mathbb{R}, \mathbb{R})$  be a  $2\pi$ -periodic function. Then any function  $p$  that satisfies  $f = p + \ddot{p}$  is of class  $C^{1,1}$ , and  $p$  is  $2\pi$ -periodic if and only if  $c_1(f) = 0$ .*

*Proof.* Let  $f \in L^\infty(\mathbb{R}, \mathbb{R})$  be a  $2\pi$ -periodic function. A function  $p$  satisfies  $f = p + \ddot{p}$  if and only if there exists  $(a, b) \in \mathbb{R}^2$  such that, for all  $\theta \in \mathbb{R}$ ,

$$(2.4) \quad p(\theta) = \int_0^\theta f(t) \sin(\theta - t) dt + a \cos(\theta) + b \sin(\theta).$$

By (2.4), any function  $p$  that satisfies  $p + \ddot{p} = f$  is of class  $C^{1,1}$ . Moreover, any such function  $p$  is of class  $C^{1,1}$  and is  $2\pi$ -periodic if and only if its restriction on  $[0, 2\pi]$  satisfies  $p(0) = p(2\pi)$ ,  $\dot{p}(0) = \dot{p}(2\pi)$ . But we have

$$\int_0^{2\pi} (p(\theta) + \ddot{p}(\theta)) e^{i\theta} d\theta = \dot{p}(2\pi) - \dot{p}(0) - i(p(2\pi) - p(0)).$$

Hence, any function  $p$  satisfying (2.4) is  $2\pi$ -periodic if and only if  $p(2\pi) = p(0)$  and  $\dot{p}(2\pi) = \dot{p}(0)$ , that is, if and only if  $c_1(f) = 0$ .  $\square$

If we deal with  $f = p + \ddot{p}$  instead of  $p$ , we get an additional condition  $c_1(f) = 0$  which says that the boundary  $\partial K$  given by (2.1) is closed. The next theorem is a consequence of the two previous propositions.

THEOREM 2.1. (i) *Let  $K$  be a strictly convex body of  $\mathbb{R}^2$  and  $p$  its support function. If  $p$  is of class  $C^{1,1}$ , then  $p + \ddot{p} \geq 0$ .*

(ii) *Conversely, let  $f \in L^\infty(\mathbb{R}, \mathbb{R})$  be a  $2\pi$ -periodic function such that  $f \geq 0$  and  $c_1(f) = 0$ . If  $p$  is a function satisfying  $f = p + \ddot{p}$ , then  $p$  is of class  $C^{1,1}$ , is  $2\pi$ -periodic (in the sense of  $C^{1,1}$  maps), and is the support function of a strictly convex body.*

Let  $K$  be a strictly convex body. We denote by  $p$  its support function of class  $C^1$  and by  $\mathcal{A}(p)$  its area. By Stokes's formula and by (2.1), we have

$$(2.5) \quad \mathcal{A}(p) = \frac{1}{2} \int_0^{2\pi} (p^2(\theta) - \dot{p}^2(\theta)) d\theta.$$

By integrating by parts, the area becomes

$$(2.6) \quad \mathcal{A}(p) = \frac{1}{2} \int_0^{2\pi} p(\theta) (p(\theta) + \ddot{p}(\theta)) d\theta,$$

which has a sense because  $p + \ddot{p}$  is a positive Radon measure, and (2.6) can be interpreted as the product of a positive Radon measure and a continuous function. In the next section, we show that the support function of a rotor is of class  $C^{1,1}$ , and (2.6) is clearly defined in that case.

**2.2. Construction of a rotor by its support function.** In this section, we recall classical definitions and properties of rotors (see [6], [17], [34]). Let  $K$  be a convex domain and  $P$  be a convex polygon.  $P$  will be called a *tangential polygon* of  $K$  and  $K$  an *osculating domain* in  $P$  if  $K \subset P$  and every side of  $P$  has a nonempty intersection with  $K$  (see [17]). We say that a polygon  $P$  is *equiangular* if all of its interior angles at the vertices are equal. We say that a convex polygon  $P$  is an  $n$ -gon if it is a regular polygon with  $n$  sides,  $n \geq 3$ .

DEFINITION 2.1. *A convex domain  $K$  will be called a rotor in a polygon  $Q$  if, for every rotation  $\rho$ , there exists a translation vector  $p_\rho$  such that  $\rho K + p_\rho$  is an osculating domain in  $K$ .*

In the following, we assume that  $Q$  is a regular polygon with  $n \geq 3$  sides; that is, we consider only rotors of a regular polygon. Hence,  $K$  is a rotor in a regular  $n$ -gon  $Q$  if and only if all tangential equiangular  $n$ -gons are regular and have equal perimeters. A rotor of an  $n$ -gon  $P_n$  has the property to rotate inside  $P_n$  while remaining in contact with all sides of  $P_n$ . The disc is the most simple example of a rotor. A rotor is a strictly convex domain (see [17], [34]). Consequently, the support function of a rotor is of class  $C^1$ .

Let  $r$  be the radius of the inscribed circle of the  $n$ -gon  $P_n$  and let  $\delta := \frac{2\pi}{n}$ . We give in the following theorem an analytic description of a rotor which will be used in the rest of the paper.

THEOREM 2.2. (i) *Let  $K$  be a rotor and  $p$  its support function. Then  $p$  satisfies*

$$(2.7) \quad p(\theta) - 2\cos(\delta)p(\theta + \delta) + p(\theta + 2\delta) = 4r\sin^2\left(\frac{\delta}{2}\right) \quad \forall \theta \in [0, 2\pi].$$

Moreover,  $p$  is of class  $C^{1,1}$  and satisfies (2.3).

(ii) *Conversely, let  $p$  be a  $2\pi$ -periodic function of class  $C^{1,1}$ . Assume that  $p$  satisfies (2.3) and (2.7). Then  $p$  is the support function of a rotor  $K$ .*

The characterization of a rotor by (2.7) is well known (see [7], [10], [21]), but we show in particular that the support function of a rotor is actually of class  $C^{1,1}$ . Before doing the proof of the theorem, we set some notation:

$$(2.8) \quad S_n(p) := p(\theta) - 2\cos(\delta)p(\theta + \delta) + p(\theta + 2\delta)$$

and

$$(2.9) \quad C_n := 4r\sin^2\left(\frac{\delta}{2}\right).$$

*Proof of (i).* We refer the reader to Chapter 8 of [34] for the following geometric property. By definition of a rotor, the tangents to  $\partial K$  at each contact point are the sides of the  $n$ -gon. Hence, the perpendiculars to these paths at their contact points meet in a point which is the instantaneous center of rotation of the body. A simple computation yields (2.7). We now prove that  $p$  is of class  $C^{1,1}$ . First, we have

$$(2.10) \quad \sum_{0 \leq k \leq n-1} p(\theta + k\delta) = nr \quad \forall \theta \in \mathbb{R}.$$

Indeed, by writing (2.7) at points  $\theta, \theta + \delta, \dots, \theta + (n-1)\delta$  and adding all of these equalities, we get (2.10). As  $K$  is strictly convex, its support function  $p$  is of class  $C^1$ . We now show that  $p$  satisfies the inequality

$$(2.11) \quad (\dot{p}(\theta') - \dot{p}(\theta))\sin(\theta - \theta') \leq (p(\theta) + p(\theta'))(1 - \cos(\theta - \theta')) \quad \forall (\theta, \theta') \in [0, 2\pi].$$

By definition of the support function, we have, for all  $(\theta, \theta') \in [0, 2\pi]$ ,

$$(x(\theta'), y(\theta')) \cdot (\cos(\theta), \sin(\theta)) \leq p(\theta).$$

Taking into account (2.1), we get

$$\dot{p}(\theta')\sin(\theta - \theta') \leq p(\theta) - p(\theta')\cos(\theta' - \theta).$$

If we permute  $\theta$  and  $\theta'$ , we obtain

$$\dot{p}(\theta)\sin(\theta' - \theta) \leq p(\theta') - p(\theta)\cos(\theta' - \theta).$$



Adding the last two inequalities yields (2.11). We now write (2.11) at the points  $\theta + k\delta$  and  $\theta' + k\delta$ ,  $0 \leq k \leq n-1$ . We get, for all  $(\theta, \theta') \in [0, 2\pi]$  and  $0 \leq k \leq n-1$ ,

(2.12)

$$(\dot{p}(\theta' + k\delta) - \dot{p}(\theta + k\delta)) \sin(\theta - \theta') \leq (p(\theta + k\delta) + p(\theta' + k\delta))(1 - \cos(\theta - \theta')).$$

By (2.10), we obtain, for all  $(\theta, \theta') \in [0, 2\pi]$ ,

$$(2.13) \quad \sum_{1 \leq k \leq n-1} p(\theta + k\delta) = nr - p(\theta)$$

and

$$(2.14) \quad \sum_{1 \leq k \leq n-1} \dot{p}(\theta + k\delta) = -\dot{p}(\theta).$$

Combining (2.12), (2.13), and (2.14), we obtain

$$(-\dot{p}(\theta') + \dot{p}(\theta)) \sin(\theta - \theta') \leq (2nr - p(\theta) - p(\theta'))(1 - \cos(\theta - \theta')).$$

Therefore, by (2.11) and the previous inequality, we get, for all  $(\theta, \theta') \in [0, 2\pi]$ ,

$$|(\dot{p}(\theta') - \dot{p}(\theta)) \sin(\theta - \theta')| \leq 2nr \sin^2 \left( \frac{\theta - \theta'}{2} \right).$$

Consequently,  $\dot{p}$  satisfies the inequality

$$|\dot{p}(\theta') - \dot{p}(\theta)| \leq 2nr \left| \tan \left( \frac{\theta - \theta'}{2} \right) \right|$$

for all  $(\theta, \theta') \in [0, 2\pi]$  such that  $|\theta - \theta'| \notin \{0, \pi, 2\pi\}$ . This inequality proves that  $\dot{p}$  is Lipschitz, and thus  $p$  is of class  $C^{1,1}$ . As  $K$  is convex and  $p$  is of class  $C^{1,1}$ , it satisfies (2.3). This concludes the proof of (i).

*Proof of (ii).* Let us assume that conditions (2.3) and (2.7) are satisfied. As  $p$  is of class  $C^{1,1}$ , is  $2\pi$ -periodic, and satisfies (2.3), it is the support function of a strictly convex body  $K$ . A straightforward computation using (2.7) shows that an osculating polygon to  $K$  is equiangular; consequently,  $K$  is a rotor.  $\square$

An example of a function  $p$  satisfying (2.7) is given by

$$(2.15) \quad p(\theta) = 1 + \frac{1}{1 - (ln - 1)^2} \cos((ln - 1)\theta),$$

where  $l \in \mathbb{N}^*$  (see Figure 2). A simple computation shows that we have  $S_n(p) = C_n$  with  $r = 1$ . Moreover, we easily have  $p(\theta) + \ddot{p}(\theta) = 1 + \cos((ln - 1)\theta) \geq 0$  for all  $\theta \in \mathbb{R}$ . Hence,  $p$  is the support function of a rotor  $K$  in an  $n$ -gon. The boundary of  $K$  is of class  $C^\infty$  because  $p$  is of class  $C^\infty$ .

In the following, we denote by  $E$  the set of the functions  $p \in C^{1,1}(\mathbb{R})$  that are  $2\pi$ -periodic and that satisfy (2.3) and (2.7). The problem of finding a rotor of minimal area is now equivalent to the optimization problem

$$(2.16) \quad \min_{p \in E} \mathcal{A}(p).$$

The existence of a minimizer for problem (2.16) easily follows from standard compactness arguments (see [32], [34]).

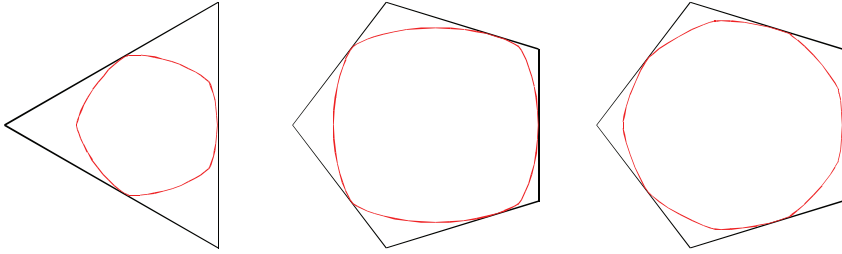


FIG. 2. Example of rotors whose support function is given by (2.15) for  $n = 3$ ,  $l = 2$  and  $n = 5$ ,  $l = 1, 2$ .

**2.3. Basic properties of rotors.** This section is devoted to well-known results about rotors which can be found in the case  $n = 3$  or  $n = 4$  in [5], [7], and [34]. Let us first recall Barbier’s theorem, which is a simple consequence of (2.7).

**THEOREM 2.3.** *Let  $r$  be the radius of the inscribed circle in  $P_n$ . Then the perimeter of every rotor  $\mathcal{R}$  of  $P_n$  is equal to  $2\pi r$ .*

*Proof.* Let  $\mathcal{R}$  be a rotor and  $p$  be its support function. The perimeter  $L$  of  $\mathcal{R}$  is given by the integral of the radius of curvature:

$$L = \int_0^{2\pi} (p(\theta) + \ddot{p}(\theta))d\theta,$$

which is well defined, as  $p$  is of class  $C^{1,1}$ . As  $\dot{p}$  is  $2\pi$ -periodic, the perimeter becomes  $L = \int_0^{2\pi} p(\theta)d\theta$ . Now integrating (2.7) on the interval  $[0, 2\pi]$  and using the  $2\pi$ -periodicity of  $p$ , we get  $L = 2\pi r$ .  $\square$

**PROPOSITION 2.5.** *Among all rotors of a regular polygon  $P_n$ , the one of maximal area is the disc of radius  $r$ .*

*Proof.* By the isoperimetric inequality, the body of maximal area among all closed curves having the same perimeter is the disc, and the disc is a rotor of  $P_n$ .  $\square$

When  $n = 4$ , a rotor is called a constant width body.

**DEFINITION 2.2.** *The width of a convex curve in a given direction is the distance between a pair of supporting lines of the curve perpendicular to this direction. If the width is constant in every direction, the curve is a curve of constant width.*

Equivalently, a constant width body has the property to rotate inside a square while remaining tangent to the four sides of the square. The relation (2.7) can be simplified in the case  $n = 4$ , which corresponds to the constant width bodies. The support function of  $K$  in this case satisfies

$$(2.17) \quad p(\theta) + p(\theta + \pi) = 2r \quad \forall \theta \in \mathbb{R},$$

which is exactly saying that any pair of parallel support lines to  $K$  is separated by the distance  $2r$  (see [14]).

**2.4. Formulation of the constraints on the interval  $[0, 2\delta]$ .** In this section, we derive consequences of (2.7) which will be useful in formulating the optimal control problem associated with the minimization problem. Let us define the reals  $s_k$  and  $t_k$  for  $k = 0, \dots, n - 1$  by

$$(2.18) \quad s_k := \frac{\sin(k\delta)}{\sin(\delta)}, \quad t_k := 2 \frac{\sin(\frac{k\delta}{2}) \sin(\frac{(k-1)\delta}{2})}{\cos(\frac{\delta}{2})} r.$$

LEMMA 2.1. *Let  $p$  be a  $2\pi$ -periodic map in  $C^{1,1}(\mathbb{R})$  satisfying (2.7). Then we have*

$$(2.19) \quad p(\theta + k\delta) = s_k p(\theta + \delta) - s_{k-1} p(\theta) + t_k \quad \forall \theta \in [0, 2\pi].$$

*Proof.* Let  $\theta \in [0, 2\pi]$  and  $v_k := p(\theta + k\delta)$ . We have by (2.7)

$$(2.20) \quad v_k - 2\cos(\delta)v_{k+1} + v_{k+2} = 4r\sin^2\left(\frac{\delta}{2}\right).$$

We solve this linear recurrent sequence and get

$$v_k = a\omega^k + \overline{a}\overline{\omega}^k + r,$$

where  $\omega := e^{i\delta}$  and  $v_0 = p(\theta)$ ,  $v_1 = p(\theta + \delta)$ . This gives (2.19).  $\square$

COROLLARY 2.1. *If  $n$  is even, a rotor  $K$  in an  $n$ -gon is a constant width body.*

*Proof.* Let  $K$  be a rotor and  $p$  be its support function which satisfies (2.7). We assume that  $n = 2m$ ,  $m \in \mathbb{N}^*$ . Using (2.19) with  $k = m$ , we get  $s_m = 0$ ,  $s_{m-1} = 1$ , and  $t_m = 2r$ . Consequently,  $p$  satisfies

$$p(\theta + m\delta) = -p(\theta) + 2r,$$

which is exactly saying that  $K$  is of constant width as  $m\delta = \pi$ .  $\square$

We now reformulate the area of a rotor on the interval  $[0, 2\delta]$ . Let  $r$  be the radius of the inscribed circle to the  $n$ -gon and  $P \in C^{1,1}(\mathbb{R}, \mathbb{R})$ ,  $F \in L^\infty(\mathbb{R}, \mathbb{R})$  be the maps defined by

$$(2.21) \quad \begin{cases} P(\theta) := p(\theta) - r, \\ F(\theta) := p(\theta) + \ddot{p}(\theta) - r = P(\theta) + \ddot{P}(\theta). \end{cases}$$

LEMMA 2.2. *Let  $p$  be the support function of a rotor and  $f$  its radius of curvature. The area of a rotor is given by*

$$\mathcal{A}(p) = \frac{n}{4\sin^2(\frac{\delta}{2})} \tilde{\mathcal{A}}(P) + \pi r^2,$$

where

$$(2.22) \quad \begin{aligned} \tilde{\mathcal{A}}(P) = \int_0^\delta & \left( P(\theta)F(\theta) + P(\theta + \delta)F(\theta + \delta) \right. \\ & \left. - \cos(\delta)(F(\theta)P(\theta + \delta) + F(\theta + \delta)P(\theta)) \right) d\theta. \end{aligned}$$

*Proof.* We have by (2.6)

$$\begin{aligned} \mathcal{A}(f) &= \frac{1}{2} \int_0^{2\pi} p(\theta)f(\theta)d\theta = \frac{1}{2} \sum_{0 \leq k \leq n-1} \int_{k\delta}^{(k+1)\delta} p(\theta)f(\theta)d\theta \\ &= \frac{1}{2} \sum_{0 \leq k \leq n-1} \int_0^\delta p(\theta + k\delta)f(\theta + k\delta)d\theta. \end{aligned}$$

Replacing  $p(\theta + k\delta)$  and  $f(\theta + k\delta)$  using (2.19), we get the result by the equalities

$$\sum_{0 \leq k \leq n-1} s_k^2 = \sum_{0 \leq k \leq n-1} s_{k-1}^2 = \frac{2}{2 \sin^2(\delta)}$$

and

$$\sum_{0 \leq k \leq n-1} s_k t_k = -\frac{n}{4 \cos^2(\frac{\delta}{2})}, \quad \sum_{0 \leq k \leq n-1} s_k s_{k-1} = \frac{n \cos(\delta)}{2 \sin^2(\delta)}. \quad \square$$

Note that in the special case of sets of constant width ( $n = 4$ ), one finds the usual functional (see [14]):

$$(2.23) \qquad \mathcal{A}(p) = \pi r^2 - \int_0^\pi p(\theta)(1 - f(\theta))d\theta,$$

which can be easily obtained by (2.6) and (2.17).

**2.5. Simplification of the functional.** Before going into details for solving the minimization problem (2.16), we diagonalize the functional (2.22) (see [21] for the same parameterization). In particular, we establish the equivalence between the parameterization of a rotor by its support function and the new parameterization. The following parameterization will be useful in defining an optimal control problem equivalent to (2.16). We set

$$\gamma := \cos(\delta), \quad \sigma := \sin(\delta), \quad \omega^{\frac{1}{2}} := e^{\frac{i\delta}{2}}, \quad \omega^{-\frac{1}{2}} := e^{-\frac{i\delta}{2}};$$

that is, we denote by  $\omega^{\frac{1}{2}}$  and  $\omega^{-\frac{1}{2}}$  a square root of  $\omega$  and  $\overline{\omega}$ .

Recall that given a rotor  $K$  of support function  $p$ , the functions  $P$  and  $F$  are defined by (2.21), and by (2.8) and (2.9) we have  $S_n(f) = C_n$  if and only if  $S_n(F) = 0$ . We now define the functions  $W \in C^{1,1}(\mathbb{R}, \mathbb{C})$  and  $Z \in L^\infty(\mathbb{R}, \mathbb{C})$  by

$$(2.24) \qquad \begin{cases} W(\theta) := P(\theta) - \overline{\omega}P(\theta + \delta), \\ Z(\theta) := F(\theta) - \overline{\omega}F(\theta + \delta), \end{cases}$$

where  $\theta \in \mathbb{R}$ , so that

$$(2.25) \qquad W + \check{W} = Z.$$

The functions  $W$  and  $Z$  can be interpreted as the *complex support function* and the *complex radius of curvature* associated with a rotor. We denote by  $X_1, X_3, U, V$  the real and imaginary parts of  $W$  and  $Z$ :

$$\begin{cases} W = X_1 + iX_3, \\ Z = U + iV, \end{cases}$$

so that we have

$$(2.26) \qquad \begin{cases} X_1(\theta) = P(\theta) - \gamma P(\theta + \delta), \\ X_3(\theta) = \sigma P(\theta + \delta), \\ U(\theta) = F(\theta) - \gamma F(\theta + \delta), \\ V(\theta) = \sigma F(\theta + \delta). \end{cases}$$

We have, equivalently,

$$(2.27) \quad \begin{cases} P(\theta) = X_1(\theta) + \frac{\gamma}{\sigma} X_3(\theta), \\ P(\theta + \delta) = \frac{1}{\sigma} X_3(\theta), \\ F(\theta) = U(\theta) + \frac{\gamma}{\sigma} V(\theta), \\ F(\theta + \delta) = \frac{1}{\sigma} V(\theta + \delta). \end{cases}$$

PROPOSITION 2.6. *The functions  $W$  and  $Z$  satisfy the relations*

$$(2.28) \quad \begin{cases} W(\theta + \delta) = \overline{\omega} W(\theta) \quad \forall \theta \in \mathbb{R}, \\ Z(\theta + \delta) = \overline{\omega} Z(\theta) \quad \text{a.e. } \theta \in \mathbb{R}. \end{cases}$$

*Proof.* Let  $p$  be the support function of a rotor. We have by (2.7)  $S_n(p) = C_n$ , where  $C_n$  is given by (2.9). Thus,  $S_n(P) = 0$ , that is,

$$(2.29) \quad \forall \theta \in \mathbb{R}, \quad P(\theta) - 2\gamma P(\theta + \delta) + P(\theta + 2\delta) = 0.$$

Eliminating  $P(\theta + 2\delta)$  in the equation above, we get

$$\forall \theta \in \mathbb{R}, \quad W(\theta + \delta) = P(\theta + \delta) - \overline{\omega}(2\gamma P(\theta + \delta) - P(\theta)),$$

which gives  $W(\theta + \delta) = \overline{\omega} W(\theta)$  for all  $\theta \in \mathbb{R}$ . By derivation of the previous equation, we get  $Z(\theta + \delta) = \overline{\omega} Z(\theta)$  for all  $\theta \in \mathbb{R}$ .  $\square$

In the following,  $\mathcal{P}_n$  denotes the regular polygon whose center is the origin and whose vertices are the points of coordinates  $(r^* \omega^k e^{i\alpha})_{0 \leq k \leq n-1}$ , where  $r^* := 2r \sin(\frac{\delta}{2})$  and  $\alpha := -\frac{\pi}{2} - \frac{\delta}{2}$ .

PROPOSITION 2.7. *Let  $K$  be a rotor,  $p$  its support function, and  $f = p + \ddot{p}$  its radius of curvature. We denote by  $Z$  its complex radius of curvature. Then we have  $f \geq 0$  if and only if  $Z(\theta) \in \mathcal{P}_n$  for a.e.  $\theta \in [0, \delta]$ .*

*Proof.* Let us consider for  $0 \leq k \leq n-1$  the map defined by

$$u_k(x, y) = s_k y - s_{k-1} x + t_k.$$

By Lemma 2.1, we have, for  $\theta \in [0, \delta]$  and for  $0 \leq k \leq n-1$ ,

$$f(\theta + k\delta) = u_k(f(\theta), f(\theta + \delta)).$$

Therefore, we have, for  $\theta \in [0, \delta]$ ,

$$\begin{aligned} f \geq 0 &\iff u_k(f(\theta), f(\theta + \delta)) \geq 0, \quad k = 0, \dots, n-1 \\ &\iff s_k(f(\theta + \delta) - r) - s_{k-1}(f(\theta) - r) + t_k + r(s_k - s_{k-1}) \geq 0 \\ &\iff \sin(k\delta)F(\theta + \delta) - \sin((k-1)\delta)F(\theta) \geq -\sigma r \\ &\iff \Im(\sin(k\delta)Z(\theta) - \sin((k-1)\delta)Z(\theta - \delta)) \geq -\sigma^2 r \\ &\iff \Im(\sin(k\delta)Z(\theta) - \sin((k-1)\delta)\omega Z(\theta)) \geq -\sigma^2 r \\ &\iff \Im(\omega^{k-1}Z(\theta)) \geq -\sigma r. \end{aligned}$$

Let  $z = x + iy$  be a complex number,  $D_k$  the hyperplane of equation  $\Im(\omega^{k-1}z) = -\sigma r$ , and  $H_k$  the half-plane defined for  $z \in \mathbb{C}$  by  $\Im(\omega^{k-1}z) \geq -\sigma r$ . We easily have that  $z \in D_{k+1}$  if and only if  $\omega z \in D_k$ . Hence, for  $\theta \in [0, \delta]$ ,  $Z(\theta)$  satisfies  $\Im(\omega^{k-1}Z(\theta)) \geq -\sigma r$ ,  $0 \leq k \leq n-1$ , if and only if  $Z(\theta)$  belongs to the intersection of the half-spaces  $H_k$ . This intersection is nonempty, as 0 belongs to  $H_k$  for all  $0 \leq k \leq n-1$  and is convex as all  $H_k$  are convex; hence it is a nonempty convex polygon. Moreover, a simple computation yields that the vertices of  $\mathcal{P}_n$  are given by the intersection  $D_k \cap D_{k+1}$  and are of coordinates  $-2ir \sin(\frac{\delta}{2})e^{i(k-\frac{1}{2})\delta}$  for  $0 \leq k \leq n-1$ .  $\square$

It is convenient to work with  $\mathcal{P}_n$  because we will see in the next section that the optimal control takes its values at the vertices of  $\mathcal{P}_n$  (the extremal points of  $\mathcal{P}_n$ ).

PROPOSITION 2.8. *Let  $p$  be the support function of a rotor  $K$ . Then the area of  $K$  is given by*

$$(2.30) \quad \mathcal{A}(p) = \pi r^2 + \frac{n}{4\sigma^2} \int_0^\delta U X_1 + V X_3 = \pi r^2 + \frac{n}{4\sigma^2} \int_0^\delta \Re(Z\overline{W}).$$

*Proof.* The area of the rotor  $K$  described by  $p \in E$  is given by (2.22). Replacing  $P(\theta)$ ,  $P(\theta + \delta)$ ,  $F(\theta)$ , and  $F(\theta + \delta)$  by  $W(\theta)$ ,  $W(\theta + \delta)$ ,  $Z(\theta)$ , and  $Z(\theta + \delta)$ , we get (2.30) by using (2.28).  $\square$

Notice the similarity between (2.6) and (2.30).

DEFINITION 2.3. *Let  $\Gamma$  be the set of the complex functions  $W$  in  $C^{1,1}([0, \delta])$  that satisfy*

$$(2.31) \quad \begin{cases} W(\delta) = \overline{\omega}W(0), \\ \dot{W}(\delta) = \overline{\omega}\dot{W}(0) \end{cases}$$

*and such that the function  $Z = W + \ddot{W}$  takes its values in the polygon  $\mathcal{P}_n$ .*

DEFINITION 2.4. *We denote by  $\mathcal{Z}$  the set of the complex valued functions  $Z \in L^\infty(\mathbb{R}, \mathbb{C})$  satisfying*

$$Z(\theta + \delta) = \overline{\omega}Z(\theta) \quad \forall \theta \in \mathbb{R}$$

*and*

$$Z(\theta) \in \mathcal{P}_n \quad \forall \theta \in \mathbb{R}.$$

We can now prove the equivalence between the parameterization of a rotor  $K$  by its support function  $p$  and its complex support function  $W$ .

THEOREM 2.4. (i) *Let  $W = X_1 + iX_3$  be a function in  $\Gamma$ . Let us define the function  $\tilde{p}$  on  $[0, 2\delta]$  by  $\tilde{p} = P + r$ , where  $P$  is given by (2.27). Then, if we extend  $\tilde{p}$  on the interval  $[0, 2\pi]$  by (2.19) and if we denote by  $p$  this extension, then  $p$  is the support function of a rotor.*

(ii) *Conversely, if  $p$  is the support function of a rotor  $K$  and  $P := p - r$ , then the function  $W|_{[0, \delta]}$  defined by (2.24) belongs to  $\Gamma$ .*

*Proof of (i).* First, let us take  $W = X_1 + iX_3 \in \Gamma$ . We have by (2.31)

$$(2.32) \quad \begin{cases} \frac{1}{\sigma}X_3(0) = X_1(\delta) + \frac{\gamma}{\sigma}X_3(\delta), \\ \sigma X_1(0) - \gamma X_3(0) = -X_3(\delta) \end{cases}$$

and

$$(2.33) \quad \begin{cases} \frac{1}{\sigma}\dot{X}_3(0) = \dot{X}_1(\delta) + \frac{\gamma}{\sigma}\dot{X}_3(\delta), \\ \sigma\dot{X}_1(0) - \gamma\dot{X}_3(0) = -\dot{X}_3(\delta). \end{cases}$$

We now define a function  $P$  on the interval  $[0, 2\delta]$  by

$$P(\theta) = X_1(\theta) + \frac{\gamma}{\sigma} X_3(\theta), \quad P(\theta + \delta) = \frac{1}{\sigma} X_3(\theta)$$

for  $\theta \in [0, \delta]$ . By (2.32), we have

$$P(\delta^-) = P(\delta^+),$$

and by (2.33) we have

$$\dot{P}(\delta^-) = \dot{P}(\delta^+).$$

Consequently, the function  $P$  is of class  $C^1$  on  $[0, 2\delta]$ . By (2.32) we also get

$$S_n(P)(0) = 0,$$

and by (2.33) we get

$$S_n(\dot{P})(0) = 0.$$

Hence, the functions  $P$  and  $\dot{P}$  satisfy  $S_n(P) = 0$  and  $S_n(\dot{P}) = 0$  for  $\theta = 0$ . If we extend  $p = P + r$  to the interval  $[0, 2\pi]$  by (2.19) and to  $\mathbb{R}$  by  $2\pi$ -periodicity, it satisfies, by construction,  $S_n(p) = C_n$ . We also have  $p(0) = p(2\pi)$  and  $\dot{p}(0) = \dot{p}(2\pi)$  by (2.19) so that the function  $p$  is of class  $C^1$ . Finally, we have  $p + \ddot{p} \geq 0$  because  $Z \in \mathcal{P}_n$ . We conclude that  $p$  is the support function of a rotor.

*Proof of (ii).* Let us now consider the support function  $p$  of a rotor. We define a function  $W$  by (2.24). First, the condition (2.3) satisfied by  $p$  implies that  $Z = W + \ddot{W}$  takes its value in  $\mathcal{P}_n$ . Let us show that  $W$  satisfies (2.31). By (2.26), we have

$$\frac{1}{\sigma} X_3(0) = X_1(\delta) + \frac{\gamma}{\sigma} X_3(\delta),$$

and by using  $S_n(P)(0) = 0$ , we get

$$\sigma X_1(0) - \gamma X_3(0) = -X_3(\delta).$$

These two real conditions imply  $W(\delta) = \overline{\omega}W(0)$ . By using (2.27) and the equality  $S_n(\dot{P})(0) = 0$ , we get  $\dot{W}(\delta) = \overline{\omega}\dot{W}(0)$ . Hence,  $W$  belongs to  $\Gamma$ .  $\square$

*Remark 2.1.* Let us make two remarks. First, any function  $W \in \Gamma$  such that  $Z = W + \ddot{W}$  satisfies, by (2.31), the condition

$$(2.34) \quad \int_0^\delta Z(\theta) e^{i\theta} d\theta = 0.$$

Second, (2.30) remains unchanged if we replace  $W$  by  $W e^{i\alpha}$  and  $Z$  by  $Z e^{i\alpha}$ , where  $\alpha \in \mathbb{R}$ .

From now on, we will mainly deal with the set  $\Gamma$  instead of the set  $E$ , as there is a one-to-one correspondence between these two sets. For  $W \in \Gamma$  such that  $W = X_1 + iX_3$  and  $Z = W + \ddot{W} = U + iV$ , we denote by  $J(W)$  the functional

$$(2.35) \quad J(W) = \int_0^\delta U X_1 + V X_3 = \int_0^\delta \Re(Z \overline{W})$$

and by  $\mathcal{A}(W)$  the area of a rotor. An integration by parts shows that we have

$$J(W) = \int_0^\delta Z \overline{W} = \int_0^\delta |W|^2 - |\dot{W}|^2,$$

and as  $J(W) \in \mathbb{R}$ , we have

$$\int_0^\delta \Im(Z \overline{W}) = 0.$$

The area of a rotor becomes

$$\mathcal{A}(W) = \pi r^2 + \frac{n}{4\sigma^2} J(W).$$

The initial problem, finding the rotor of least area (problem (2.16)), is now equivalent to

$$(2.36) \quad \min_{W \in \Gamma} J(W).$$

In sections 3 and 4, we will solve problem (2.36) using the optimal control theory.

**2.6. Fourier series of regular rotors.** Before going further into the analysis of (2.36), we describe by Fourier series the two families of regular rotors  $O_n^{ln \pm 1}$  introduced in section 1. An analogous description is given by Focke (see [10]), but here we use the new parameterization  $(W, Z)$ , which simplifies the computations.

We consider the subset  $J \subset \mathbb{Z}$  defined for  $n \geq 3$  by

$$J = (n\mathbb{Z} + 1) \cup (n\mathbb{Z} - 1) \setminus \{\pm 1\}$$

and let  $p$  be the support function of a rotor. Then  $p$  is given by

$$(2.37) \quad p(\theta) = r + c_1 e^{i\theta} + c_{-1} e^{-i\theta} + \sum_{j \in J} c_j e^{ij\theta},$$

where  $c_j$  are the Fourier coefficients of  $p$ . In the case of constant width bodies, the support function becomes

$$p(\theta) = r + c_1 e^{i\theta} + c_{-1} e^{-i\theta} + \sum_{l \in \mathbb{Z}^*} \left( c_{4l-1} e^{i(4l-1)\theta} + c_{4l+1} e^{i(4l+1)\theta} \right).$$

By the Parseval equality, the area of a rotor  $K$  becomes

$$(2.38) \quad \mathcal{A}(p) = \pi \left( r^2 - \sum_{j \in J} \frac{|c_j|^2}{j^2 - 1} \right).$$

Let  $m \in \mathbb{N}^*$ ,  $\varepsilon = \pm 1$ ,  $L = mn - \varepsilon$ ,  $\tau = \frac{\delta}{L}$ , and  $s = L - 1$ . We can easily check that the complex function defined by

$$(2.39) \quad Z(\theta) = \sum_{0 \leq j \leq s} \omega^{\varepsilon j} \mathbb{1}_{[j\tau, (j+1)\tau[}$$

is an element of  $\mathcal{Z}$ . We will define the regular rotors by (2.39).



DEFINITION 2.5. We call regular rotor any element  $W$  of  $\Gamma$  such that  $W + \ddot{W}$  is of the form (2.39). The first series consists of the rotors obtained for  $\varepsilon = 1$ , and the second series is obtained for  $\varepsilon = -1$ .

The integer  $L = s + 1$  denotes the number of intervals of the subdivision  $[0, \delta]$ . We now consider the set

$$J_\varepsilon = \left\{ k \in \mathbb{Z}, \quad k \equiv \varepsilon[n] \right\}.$$

PROPOSITION 2.9. The Fourier series of a regular rotor is given by

$$(2.40) \quad Z(\theta) = \frac{n}{\pi} e^{-\frac{i\varepsilon\delta}{2}} \sin\left(\frac{\varepsilon\delta}{2}\right) \sum_{k \in J_\varepsilon} \frac{e^{ikL\theta}}{k}.$$

*Proof.* The function  $\theta \mapsto e^{i\theta} Z(\theta)$  is  $\delta$ -periodic, as we have  $Z(\theta + \delta) = \overline{\omega} Z(\theta)$ . Thus, one has, for a.e.  $\theta \in \mathbb{R}$ ,

$$Z(\theta) e^{i\theta} = \sum_{k \in \mathbb{Z}} c_k e^{ikn\theta},$$

where the Fourier coefficients are given by

$$c_k = \frac{n}{2\pi} \int_0^\delta e^{-i(kn-1)\theta} Z(\theta) d\theta.$$

Using (2.39), we get, for  $k \in \mathbb{Z}$ ,

$$c_k = \frac{i}{kn-1} (e^{-i(kn-1)\tau} - 1) \sum_{0 \leq j \leq s} \omega^{\varepsilon j} e^{-i(kn-1)j\tau}.$$

The previous sum can be easily computed, and we get  $c_0 = 0$  and

$$c_k \neq 0 \iff \omega^\varepsilon e^{-i(kn-1)\tau} = 1,$$

because  $\tau = \frac{\delta}{L}$ . For  $\varepsilon = 1$ , one has

$$\omega^\varepsilon e^{-i(kn-1)\tau} = 1 \iff \exists j \in \mathbb{Z}, \quad kn-1 = (jn+1)L.$$

For  $\varepsilon = +1$ , we finally obtain

$$c_k = \frac{n}{\pi(jn+1)} e^{-i\frac{\delta}{2}} \sin\left(\frac{\delta}{2}\right).$$

For  $\varepsilon = -1$ , a similar computation yields

$$c_k = -\frac{n}{\pi(jn-1)} e^{i\frac{\delta}{2}} \sin\left(\frac{\delta}{2}\right).$$

This gives (2.40).  $\square$

The Fourier series of  $Z$  can also be written as

$$Z(\theta) = \frac{n}{\pi} e^{-i\varepsilon\frac{\delta}{2}} \sin\left(\frac{\varepsilon\delta}{2}\right) \sum_{j \in \mathbb{Z}} \frac{e^{i((mnj-\varepsilon j+\varepsilon m)n-1)\theta}}{jn+\varepsilon}.$$

The first series of rotors obtained for  $\varepsilon = +1$  will be called  $O_n^{mn-1}$ , and the second series obtained for  $\varepsilon = -1$  will be called  $O_n^{mn+1}$  (see [10], [21]). For  $n = 4$ , the two families  $O_4^{4m-1}$  and  $O_4^{4m+1}$  describe the odd Reuleaux polygons (see [9]). A Reuleaux polygon consists of the intersection of  $N$  circles of radii 1 ( $N$  is odd) and whose centers are the vertices of an  $N$ -gon of side 1. An analogous geometrical description of  $O_n^{ln\pm 1}$  can be found in [16].

PROPOSITION 2.10. *Let  $K$  be a rotor and  $Z$  its complex radius of curvature. If  $Z$  is given by (2.39), then the area of  $K$  becomes*

$$(2.41) \quad \mathcal{A}(K) = \pi r^2 - \frac{r^2 n^2}{2\pi} \tan^2 \left( \frac{\delta}{2} \right) \sum_{j \in \mathbb{Z}} \frac{1}{(jn+1)^2 ((mn-\varepsilon)^2 (jn+1)^2 - 1)}.$$

*Proof.* By (2.30), we have

$$\mathcal{A}(K) = \pi r^2 + \frac{n}{4\sigma^2} \int_0^\delta \overline{Z}(\theta) W(\theta) d\theta,$$

where  $W$  is in  $\Gamma$  and satisfies  $W + \ddot{W} = Z$ . By (2.40), the function  $W$  is given by

$$W(\theta) = -\frac{n}{\pi} e^{-i\varepsilon \frac{\delta}{2}} \sum_{k \in J_\varepsilon} \frac{e^{ikL\theta}}{k(k^2 L^2 - 1)}.$$

Applying the Parseval equality yields (2.41).  $\square$

The following proposition has been proved in [10]. It will be useful for proving Goldberg's conjecture (see section 4). We give a short proof using the expression of the area of a rotor given by (2.41).

PROPOSITION 2.11. *In the class of the regular rotors  $O_n^{mn\pm 1}$ , the one of minimal area is  $O_n^{n-1}$  obtained for  $m = 1$  and  $\varepsilon = +1$ . Its Fourier series is given by*

$$(2.42) \quad Z(\theta) = \frac{n}{\pi} e^{-i\frac{\delta}{2}} \sin \left( \frac{\delta}{2} \right) \sum_{j \in \mathbb{Z}} \frac{e^{i(((n-1)j+1)n-1)\theta}}{jn+1}.$$

*Proof.* The area of a rotor  $K$  described by  $Z \in \mathcal{Z}$  is an increasing function of  $m \in \mathbb{N}^*$  by (2.41). Thus the minimum in the class of regular rotors is obtained for  $m = 1$ . The minimum between  $O_n^{n-1}$  and  $O_n^{n+1}$  is clearly  $O_n^{n-1}$ .  $\square$

It is easy to see that  $O_n^{n-1}$  is invariant with respect to the action of the dihedral group of order  $2(n-1)$ ,  $D_{n-1}$ . For example, the Reuleaux triangle is invariant with respect to the group  $D_3$  and the  $\Delta$ -biangle is invariant with respect to the group  $D_2$ . Anyway, it seems difficult to prove that a minimizer of problem (2.36) has these symmetries.

### 3. The minimization problem as an optimal control problem.

**3.1. First consequences of the PMP.** In the case of the sets of constant width ( $n = 4$ ), one can deal with one control on the interval  $[0, \pi]$  because the functional to minimize is given by (2.23) (see [14]). The optimal control problem in the general case ( $n \geq 3$ ) requires a sharper analysis here because we have to deal with a control  $(U, V) \in \mathbb{R}^2$  on  $[0, \delta]$  as  $\gamma \neq 0$ .

Let us consider the polygon  $\mathcal{P}'_n$  which corresponds to the initial polygon  $\mathcal{P}_n$  by a homotheticity of ratio  $\lambda = \frac{1}{2 \sin(\frac{\delta}{2})}$  and a rotation of angle  $\alpha = \frac{\pi}{2} + \frac{\delta}{2}$ . Hence, the vertices of the polygon  $\mathcal{P}'_n$  are the points of coordinates  $(\omega^j)_{0 \leq j \leq n-1}$ . We consider

the differential system (harmonic oscillator) on the interval  $[0, \delta]$  described by the equations

$$(3.1) \quad \begin{cases} \dot{X}_1 = X_2, \\ \dot{X}_2 = -X_1 + U, \\ \dot{X}_3 = X_4, \\ \dot{X}_4 = -X_3 + V, \end{cases}$$

where the control  $(U, V)$  takes its values within the polygon  $\mathcal{P}'_n$ . As the vector  $(X_1, X_3)$  satisfies the boundary conditions given by (2.31), the PMP will lead to transversality conditions. Notice that the initial and final states are not fixed, but they are linked by (2.31).

By the linearity of (3.1), the problem (2.36) is clearly equivalent to minimizing (2.30), where  $(X_1, X_2, X_3, X_4)$  satisfies (2.31) and (3.1) and the control  $(U, V)$  takes its values within the polygon  $\mathcal{P}'_n$ . We have thus reformulated the initial shape optimization problem into an optimal control problem:

$$(3.2) \quad \min \left\{ \int_0^\delta U X_1 + V X_3, (U, V) \in \mathcal{P}'_n, (X_1, X_2, X_3, X_4) \text{ satisfies (2.31) and (3.1)} \right\}.$$

DEFINITION 3.1. We denote by  $X = (X_1, X_2, X_3, X_4) \in \mathbb{R}^4$  the state variable and  $q = (q_1, q_2, q_3, q_4) \in \mathbb{R}^4$  the dual variable. The Hamiltonian of the system  $H := H(X, q, U, V, p_0)$  is given by

$$(3.3) \quad H = q_1 X_2 + q_2 (-X_1 + U) + q_3 X_4 + q_4 (-X_3 + V) + p_0 (U X_1 + V X_3),$$

where  $p_0 \in \mathbb{R}$ .

We first prove the existence of an optimal control of (3.2).

THEOREM 3.1. There exists an optimal control for problem (3.2).

*Proof.* There exists an admissible trajectory of (3.2) corresponding to  $Z = 0$ ; hence, the set of admissible trajectories is nonempty. The existence of an optimal control will follow from an application of Filippov's theorem (see [1] or [31, p. 98]). First, we check that the trajectories are uniformly bounded. Indeed, the set of admissible controls is compact, and by linearity of (3.1), we obtain a uniform bound by Gronwall's lemma. Second, given  $(X_1, X_2, X_3, X_4) \in \mathbb{R}^4$ , the set defined by

$$\left\{ (X_1 U + X_3 V, X_2, -X_1 + U, X_4, -X_3 + V), (U, V) \in \mathcal{P}'_n \right\}$$

is clearly convex. By Filippov's theorem (see [31]), we get the result.  $\square$

By the PMP, there exists a map  $X : [0, \delta] \rightarrow \mathbb{R}^4$  absolutely continuous, a map  $q : [0, \delta] \rightarrow \mathbb{R}^4$  absolutely continuous, a constant  $p_0 \leq 0$ , and an optimal control  $Z(\theta) = (U(\theta), V(\theta))$  satisfying the equations

$$(3.4a) \quad \dot{X} = \frac{\partial H}{\partial q},$$

$$(3.4b) \quad \dot{q} = -\frac{\partial H}{\partial X},$$

and

$$(3.5) \quad \max_{(\tilde{U}, \tilde{V}) \in \mathcal{P}'_n} H(X(\theta), q(\theta), \tilde{U}, \tilde{V}, p_0) = H(X(\theta), q(\theta), U(\theta), V(\theta), p_0).$$

Moreover, the pair  $(p_0, q)$  is nontrivial, and  $q$  satisfies transversality conditions that we will make explicit in the paragraph below.

DEFINITION 3.2. *We will call an extremal trajectory a quadruplet  $(X, q, p_0, Z)$  satisfying (3.4a), (3.4b), (3.5) and such that the pair  $(X, q)$  is absolutely continuous on  $[0, \delta]$ ,  $p_0 \leq 0$ , and  $(p_0, q)$  is nonzero. The control  $Z = (U, V)$  corresponding to an extremal trajectory will be called an extremal control.*

As the system is autonomous, the Hamiltonian of the system is conserved along the extremal trajectories of the system. By (3.4b), the variable  $q$  satisfies the dual system:

$$(3.6) \quad \begin{cases} \dot{q}_1 = q_2 - p_0 U, \\ \dot{q}_2 = -q_1, \\ \dot{q}_3 = q_4 - p_0 V, \\ \dot{q}_4 = -q_3. \end{cases}$$

The system (3.1) can also be written as

$$(3.7) \quad \ddot{W} + W = Z,$$

where

$$W = X_1 + iX_3, \quad Z = U + iV,$$

and from now on, for convenience, we will mainly deal with complex variables. We write the dual variable  $q = (q_1, q_2, q_3, q_4)$  in the following way:

$$(3.8) \quad \Pi = q_2 + iq_4,$$

so that we have

$$(3.9) \quad \dot{\Pi} = -q_1 - iq_3.$$

We get from (3.6)

$$(3.10) \quad \ddot{\Pi} + \Pi = p_0 Z.$$

It follows that  $W$  and  $\Pi$  are of class  $C^{1,1}$  on the interval  $[0, \delta]$ , as the control  $Z$  is bounded. Let us now compute the transversality conditions by using the variables  $(W, \Pi)$ . The vector of  $\mathbb{C}^4$ ,

$$(W(0), \dot{W}(0), W(\delta), \dot{W}(\delta)),$$

takes its values in the subspace  $M$  of  $\mathbb{C}^4$  defined by

$$M := \{(A, B, \bar{\omega}A, \bar{\omega}B), (A, B) \in \mathbb{C}^2\}.$$

The orthogonal of  $M$  in  $\mathbb{C}^4$  (with respect to the canonical scalar product in  $\mathbb{C}^4$ ) is simply

$$M^\perp = \{(A', B', -\bar{\omega}A', -\bar{\omega}B'), (A', B') \in \mathbb{C}^2\}.$$

By the PMP, the vector  $(-q(0), q(\delta)) = (-\Pi(0), -\dot{\Pi}(0), \Pi(\delta), \dot{\Pi}(\delta))$  is in  $M^\perp$  (see [26], [31] for transversality conditions in the periodic case). Hence, we have  $\Pi(\delta) = \overline{\omega}\Pi(0)$  and  $\dot{\Pi}(\delta) = \overline{\omega}\dot{\Pi}(0)$ ; consequently,  $\Pi$  satisfies (2.31), that is, the same boundary conditions as  $W$ . Note that the Hamiltonian can be expressed as follows:

$$(3.11) \quad H = -\Re(W\overline{\Pi}) - \Re(\dot{W}\overline{\dot{\Pi}}) + \Re((p_0W + \Pi)\overline{Z}).$$

By (3.8) and (3.9), the scalar product in  $\mathbb{C}^2$  between  $W$  and  $\Pi$  is given by

$$(3.12) \quad \langle W, \Pi \rangle := \sum_{1 \leq i \leq 4} q_i X_i = -\Re(W\overline{\Pi}) + \Re(\dot{W}\overline{\dot{\Pi}}).$$

We now simplify the system (3.4a)–(3.4b) by expressing the dual variable  $\Pi$  as a function of the state variable  $W$ . This corresponds to a reduction of the number of degrees of freedom of the system (3.4a)–(3.4b).

LEMMA 3.1. *Let  $W$  be an extremal trajectory of the system and  $\Pi = q_2 + iq_4$  its dual variable. Then there exists  $A \in \mathbb{C}$  such that the function  $\Pi - p_0W$  is of the form*

$$\Pi(\theta) - p_0W(\theta) = Ae^{-i\theta}, \quad \theta \in [0, \delta].$$

*Proof.* We have by (3.7) and (3.10)

$$\ddot{\Pi} + \Pi = p_0(U + iV) = p_0Z = p_0(\ddot{W} + W),$$

and, consequently, the function  $y = \Pi - p_0W$  satisfies  $\ddot{y} + y = 0$ . There exist two constants  $(A, B) \in \mathbb{C}^2$  such that, for all  $\theta \in [0, \delta]$ , we have

$$(3.13) \quad \Pi(\theta) - p_0W(\theta) = Ae^{-i\theta} + Be^{i\theta}.$$

Let us prove that  $B = 0$ . For  $\theta = 0$  and  $\theta = \delta$ , we get

$$\Pi(0) - p_0W(0) = A + B, \quad \Pi(\delta) - p_0W(\delta) = A\overline{\omega} + B\omega.$$

But, as  $(W, \Pi)$  belong to  $\Gamma$ , we have by the transversality conditions

$$\Pi(\delta) - p_0W(\delta) = \overline{\omega}\Pi(0) - p_0\overline{\omega}W(0) = A\overline{\omega} + B\overline{\omega}.$$

Thus, we conclude that  $B = 0$ .  $\square$

We now show that an extremal trajectory is not abnormal.

LEMMA 3.2. *Let  $(X, q, p_0, Z)$  be an extremal trajectory. Then the constant  $p_0$  is strictly negative.*

*Proof.* Let us assume that  $p_0 = 0$ . As the point  $(0, 0)$  belongs to  $\mathcal{P}'_n$ , we get by the PMP the following: for almost  $\theta \in [0, \delta]$ ,

$$q_2(\theta)U(\theta) + q_4(\theta)V(\theta) \geq 0.$$

Consequently,

$$\int_0^\delta (q_2(\theta)U(\theta) + q_4(\theta)V(\theta))d\theta \geq 0.$$

But, we have

$$\int_0^\delta (q_2(\theta)U(\theta) + q_4(\theta)V(\theta))dt = \int_0^\delta \operatorname{Re}(\overline{\Pi}(\theta)Z(\theta))d\theta,$$

and by the previous lemma and (2.34), we have

$$\int_0^\delta \bar{\Pi} Z = \int_0^\delta \bar{A} e^{i\theta} Z(\theta) d\theta = 0.$$

Hence, the function  $\Re(\Pi Z)$  must be zero on the interval  $[0, \delta]$ . If  $\Pi$  is not zero, then the extremal control associated with this trajectory is orthogonal to  $\Pi$ . This contradicts (3.5) by choosing a control  $\tilde{Z} \in \mathcal{P}'_n$  such that  $\Re(\Pi \tilde{Z}) > 0$ . Hence,  $\Pi$  must be 0 everywhere. This is not possible because by the PMP, the pair  $(\Pi, p_0)$  is not zero.  $\square$

In the following, we take  $p_0 = -1$  for any extremal trajectory of the system. Let  $(W, \Pi, Z)$  be an extremal trajectory defined by  $\frac{\partial H}{\partial U} = \frac{\partial H}{\partial V} = 0$ ; that is, we have  $\Pi = W$ . As  $p_0 = -1$ , we get by Lemma 3.1

$$W(\theta) = \frac{A}{2} e^{-i\theta}, \quad \theta \in [0, \delta].$$

Such an extremal trajectory represents the disc which maximizes the area, and this case can be excluded.

LEMMA 3.3. *Let  $W$  be an extremal trajectory of the system and  $\Pi$  its dual variable. Then there exists an extremal trajectory of the system,  $W_1$ , with dual variable  $\Pi_1$ , such that*

$$\Pi_1 = -W_1$$

*and such that the functional of both extremals is identical.*

*Proof.* For  $\lambda \in \mathbb{C}$ , we consider the functions  $(W_1, \Pi_1)$  defined on  $[0, \delta]$  by

$$\begin{cases} W_1(\theta) = W(\theta) + \lambda e^{-i\theta}, \\ \Pi_1(\theta) = \Pi(\theta) + \lambda e^{-i\theta}. \end{cases}$$

We have

$$\ddot{W}_1 + W_1 = Z, \quad \ddot{\Pi}_1 + \Pi_1 = -Z.$$

We can easily check that  $W_1$  and  $\Pi_1$  satisfy (2.31). By (2.34), the functional remains unchanged:

$$\int_0^\delta \Re(Z \bar{W}_1) = \int_0^\delta \Re(Z \bar{W}).$$

Hence,  $(W_1, \Pi_1)$  is also an optimal trajectory. Recall that the Hamiltonian along this trajectory is defined by

$$H_1 = -\Re(W_1 \bar{\Pi}_1) - \Re(W'_1 \bar{\Pi}'_1) + \Re((\Pi_1 - W_1) \bar{Z}).$$

Using Lemma 3.1, we have  $\Pi = -W + A e^{-i\theta}$ , and by a computation, we get

$$H_1 = H + 2\Re(A \bar{\lambda}) + 2|\lambda|^2,$$

where  $H$  is given by (3.11). This shows that the PMP (3.5) gives the same extremal control for  $(W, \Pi)$  and for  $(W_1, \Pi_1)$ , as both Hamiltonian are equal up to a constant. Finally, we have

$$\Pi_1 + W_1 = (A + 2\lambda) e^{-i\theta},$$

and by taking  $\lambda$  such that  $A = -2\lambda$ , we get the lemma.  $\square$

From now on, we consider extremal solutions  $(W, Z)$  of the system such that the dual variable  $\Pi$  satisfies  $\Pi = -W$  (by Lemma 3.3). To simplify, we will say that  $W$  is an extremal trajectory of the system if  $\Pi = -W$  and if it satisfies the PMP. The Hamiltonian of the system is constant along such an extremal and can be written using (3.11):

$$(3.14) \quad H = |W|^2 + |\dot{W}|^2 - 2\Re(W \cdot \bar{Z}) = |W - Z|^2 + |\dot{W}|^2 - |Z|^2.$$

*Remark 3.1.* By (3.14), and by using (3.5), we get  $H \geq 0$  along an extremal trajectory.

**3.2. Computation of the extremal control.** We now examine in more detail the consequences of the PMP to describe the extremal trajectories. Let us recall the definition of a switching point.

**DEFINITION 3.3.** Let  $Z = (U, V)$  be an extremal control of problem (3.2). A point  $\tau \in ]0, \delta[$  is called a switching point if, for every  $\varepsilon > 0$  such that  $[\tau - \varepsilon, \tau + \varepsilon] \subset ]0, \delta[$ , the control  $Z$  is nonconstant on  $[\tau - \varepsilon, \tau + \varepsilon]$ .

To restrict the class of extremal trajectories, we prove step by step the following:

- An extremal is bang-bang, and the associated control takes its values on the vertices of  $\mathcal{P}'_n$  (Lemma 3.4).
- An extremal control takes its values regularly on the vertices of  $\mathcal{P}'_n$  (Theorem 3.2).
- The number of switching points of an extremal control is finite (Theorem 3.3).
- The number of switching points of an extremal control is prescribed (Theorem 3.4).
- The distance between two consecutive switching points is constant (Proposition 4.1).

We first prove two lemmas which will be useful in proving Theorems 3.2 and 3.3.

**LEMMA 3.4.** Let  $W$  be an extremal trajectory of the system. Then the extremal control takes its values on the vertices of  $\mathcal{P}'_n$ .

*Proof.* First, we show that the extremal control takes its values on the vertices of  $\mathcal{P}'_n$ . By (3.5) and (3.14), the extremal control is a solution of the maximization problem

$$(3.15) \quad \max_{z \in \mathcal{P}'_n} \phi(z),$$

where  $\phi$  is defined on  $\mathcal{P}'_n$  by  $\phi(z) := -2\Re(\bar{z}W(\theta))$  and  $\theta \in [0, \delta]$  is fixed. Let  $z_0$  be a point where the maximum in (3.15) is obtained.

If  $W(\theta) = 0$ , then the maximum in (3.15) can be taken arbitrarily in  $\mathcal{P}'_n$  and, in particular, on a vertex of  $\mathcal{P}'_n$ . Let us now assume that  $W(\theta) \neq 0$ . The maximum of  $\phi$  is necessarily on the boundary of  $\mathcal{P}'_n$  because  $\nabla\phi(z_0) \neq 0$ . Hence,  $z_0$  is of the form  $z_0 = t_0\omega^j + (1 - t_0)\omega^{j+1}$ , where  $t_0 \in [0, 1]$  and  $0 \leq j \leq n - 1$ . If  $W(\theta)$  is orthogonal to  $\omega^{j+1} - \omega_j$ , then we can take  $z_0 = \omega^j$  or  $z_0 = \omega^{j+1}$ . If this is not the case, let us define the function  $\psi$  on  $[0, 1]$  by

$$\psi(t) = -2\Re((t\bar{\omega}^j + (1 - t)\bar{\omega}^{j+1})W(\theta)).$$

As we have  $\dot{\psi}(t_0) \neq 0$ , the maximum in (3.15) cannot be reached at  $t_0$ . Hence, the maximum in (3.15) is reached on a vertex of  $\mathcal{P}'_n$ , and this proves the lemma.  $\square$

LEMMA 3.5. *Let  $W$  be an extremal trajectory of the system and  $\tau_j$ ,  $j \in \mathbb{N}$ , a switching point of the extremal control  $Z$  such that  $Z(\tau_j^-) = \omega^{k_j}$  and  $Z(\tau_j^+) = \omega^{k_{j+1}}$  with  $(k_j, k_{j+1}) \in \mathbb{N}^2$ . Then there exists  $t_j \in \mathbb{R}$  such that*

$$(3.16) \quad W(\tau_j) = t_j \omega^{\frac{k_j + k_{j+1}}{2}}.$$

*Proof.* The Hamiltonian is constant along an extremal trajectory, and the functions  $\theta \mapsto |W(\theta)|^2$  and  $\theta \mapsto |W'(\theta)|^2$  are continuous. Hence, the function  $\theta \mapsto \Re(W(\theta)\overline{Z(\theta)})$  is continuous, and at a switching point  $\tau_j$ , we get

$$\Re(W(\tau_j)\overline{\omega^{k_j}}) = \Re(W(\tau_j)\overline{\omega^{k_{j+1}}}).$$

Geometrically speaking, the vector  $W(\tau_j)$  is orthogonal to the segment  $[\omega^{k_j}, \omega^{k_{j+1}}]$ ; hence it takes the form given by (3.16).  $\square$

By Lemma 3.4, an extremal trajectory is bang-bang: the extremal control associated with this trajectory takes the extremal values of the convex polygon  $\mathcal{P}'_n$ . We now show that the extremal control goes all over the vertices  $\omega^j$  clockwise or counterclockwise.

THEOREM 3.2. *Let  $W$  be an extremal trajectory of the system. There exists  $\varepsilon \in \{\pm 1\}$  such that if  $\tau_j$  is a switching point with  $Z(\tau_j^-) = \omega^{k_j}$  and  $Z(\tau_j^+) = \omega^{k_{j+1}}$ , then*

$$k_{j+1} - k_j = \varepsilon.$$

*Proof.* By Lemma 3.5, we have at a switching point  $\tau_j$

$$W(\tau_j) = t_j \omega^{\frac{k_j + k_{j+1}}{2}},$$

where  $t_j \in \mathbb{R}$ . Geometrically speaking, the vector  $W(\tau_j)$  is parallel to the median of the segment  $[\omega^{k_j}, \omega^{k_{j+1}}]$ , which is a side or a diagonal of the polygon  $\mathcal{P}'_n$ . The line  $\Delta$  directed by  $W(\tau_j)$  contains 0, 1, or 2 vertices of  $\mathcal{P}'_n$ .

First, assume that  $\Delta$  does not contain any vertex of  $\mathcal{P}'_n$ . If  $|k_j - k_{j+1}| \neq 1$ , there exists another vertex  $\omega^s := (U_s, V_s)$  of  $\mathcal{P}'_n$ , which is different from  $\omega^j$  and  $\omega^{j+1}$ , and such that

$$-2\Re(\overline{W(\tau_j)}\omega^s) > -2\Re(\overline{W(\tau_j)}\omega^j)$$

or

$$-2\Re(\overline{W(\tau_j)}\omega^s) > -2\Re(\overline{W(\tau_j)}\omega^{j+1}).$$

This means that the scalar product between  $W(\tau_j)$  and  $\omega^s$  is less than the scalar product between  $W(\tau_j)$  and  $\omega^{k_j}$  or  $\omega^{k_{j+1}}$ . Assume, for example, that the first inequality is satisfied by  $\omega^s$ . We obtain by (3.14)

$$H(W(\tau_j), \Pi(\tau_j), U_s, V_s, p_0) > H(W(\tau_j), \Pi(\tau_j), U(\tau_j^-), V(\tau_j^-), p_0).$$

This contradicts (3.5), that is, the maximality of the Hamiltonian along an extremal.

Now assume that  $\Delta$  contains only one vertex of  $\mathcal{P}'_n$  (in this case  $n$  is necessarily even) and  $|k_j - k_{j+1}| \neq 1$ . The segment  $[\omega^{k_j}, \omega^{k_{j+1}}]$  is parallel to a side  $[\omega^r, \omega^{r+1}]$  of  $\mathcal{P}'_n$ . Let us call  $\omega^l$  the vertex of  $\mathcal{P}'_n$  opposite to  $[\omega^r, \omega^{r+1}]$ . As in the previous case, we get a contradiction in (3.5). Indeed, one has

$$H(W(\tau_j), \Pi(\tau_j), U_s, V_s, p_0) > H(W(\tau_j), \Pi(\tau_j), U(\tau_j^-), V(\tau_j^-), p_0),$$

with  $s$  equal to  $r$ ,  $r + 1$ , or  $l$  and with  $\omega^s := (U_s, V_s)$ .



If  $\Delta$  contains two vertices  $\omega^s$  and  $\omega^l$  of  $\mathcal{P}_n$  and if  $|k_j - k_{j+1}| \neq 1$ , we get a similar contradiction in (3.5) by considering the vertex  $\omega^s$  or  $\omega^l$ .

We have thus proved that  $|k_{j+1} - k_j| = 1$  for any switching point  $\tau_j$ . To conclude the proof of the theorem, we have to show that the extremal control does not contain a subsequence of the form  $\{\omega^p, \omega^{p+1}, \omega^p, \dots\}$ , where  $p \in \mathbb{N}$ . Let us assume that an extremal control  $Z$  takes the form

$$Z(\theta) = \mathbb{1}_{[\tau_1, \tau_2[} + \omega \mathbb{1}_{[\tau_2, \tau_3[} + \mathbb{1}_{[\tau_3, \tau_4[} + \tilde{Z}(\theta), \quad \theta \in [0, \delta],$$

where  $\tau_1 < \tau_2 < \tau_3 < \tau_4$  and  $(\tau_2, \tau_3)$  are two consecutive switching points, and  $\tilde{Z}$  is the restriction of  $Z$  on  $[0, \delta] \setminus [\tau_1, \tau_4]$ :

$$\tilde{Z} = Z|_{[0, \delta] \setminus [\tau_1, \tau_4]}.$$

It is always possible to consider this case by multiplying  $Z$  by  $\overline{\omega}^p$ , since it does not change the extremality of  $(W, Z)$ . As  $Z$  is switching from 1 to  $\omega$  for  $\theta = \tau_2$ , we have by Lemma 3.5  $W(\tau_2) = t_2 \omega^{\frac{1}{2}}$ ,  $t_2 \in \mathbb{R}$ . Notice that by (3.14), we necessarily have  $t_1 < 0$ . Indeed, by the maximality condition, the value of the Hamiltonian on the extremal is greater than the value of the Hamiltonian obtained with  $(\tilde{U}, \tilde{V}) = (0, 0)$ . At the switching point  $\tau_3$ , we similarly have  $W(\tau_3) = t_3 \omega^{\frac{1}{2}}$ , where  $t_3 < 0$ . Hence, the vectors  $W(\tau_2)$  and  $W(\tau_3)$  are parallel. For  $\theta \in [\tau_2, \tau_3]$ , the function  $\theta \mapsto W(\theta)$  describes an arc of an ellipse whose center is the point  $\omega$ . Indeed, by (3.7), we have

$$W(\theta) = \omega + A_2 e^{i\theta} + B_2 e^{-i\theta}, \quad (A_2, B_2) \in \mathbb{C}^2.$$

Hence, the vectors  $W(\tau_2)$  and  $W(\tau_3)$  are equal or opposite because the line directed by  $W(\tau_2)$  crosses the ellipse in at most two points. But, as we have  $W(\tau_2) \cdot W(\tau_3) = t_2 t_3 > 0$ , we must have

$$W(\tau_2) = W(\tau_3).$$

This condition will bring a contradiction. Let  $\mathcal{E}$  be the ellipse of center  $\omega$  on which the function  $W$  takes its values for  $\theta \in [\tau_1, \tau_2]$ .

*First case.*  $\mathcal{E}$  is not degenerated. The function  $W$  satisfies  $W(\tau_2) = W(\tau_3)$ . As  $W$  is of class  $C^1$ , it must go all over the ellipse, and this is possible only if  $\tau_2 = \tau_1 + 2k\pi$ ,  $k \in \mathbb{N}^*$ . As  $(\tau_2, \tau_3)$  belong to the interval  $[0, \delta]$ , we get a contradiction.

*Second case.*  $\mathcal{E}$  is a segment which contains  $W(\tau_2)$  and  $\omega$ . For  $\theta \in [\tau_2, \tau_3]$ ,  $W(\theta)$  takes its values within this segment. For  $\theta \in [\tau_1, \tau_2]$ , the function  $\theta \mapsto W(\theta)$  takes its values within an ellipse  $\mathcal{E}'$  whose center is the point  $(1, 0)$ . By Lemma 3.5,  $W$  satisfies, for  $\theta = \tau_2$ ,  $W(\tau_2) = t_2 \omega^{\frac{1}{2}}$ . Hence, the function  $W$  cannot be of class  $C^1$  at the point  $\theta = \tau_2$ , since  $W(\theta)$  is parallel to  $W(\tau_2)$  for  $\theta \in [\tau_2, \tau_3]$ . We thus get a contradiction.

We have thus proved that for any switching point  $\tau_j$ ,  $k_{j+1} - k_j = \varepsilon$ , where  $\varepsilon = \pm 1$  is fixed by the rotation of  $Z$  clockwise or counterclockwise. This concludes the proof of the theorem.  $\square$

We now show that an extremal control switches a finite number of times on the interval  $[0, \delta]$ .

**THEOREM 3.3.** *Let  $W$  be an extremal trajectory of the system. Then there exists a subdivision  $(\tau_j)_{0 \leq j \leq r}$  of  $[0, \delta]$  with  $r \in \mathbb{N}^*$  such that  $\tau_0 = 0$  and  $\tau_{r+1} = \delta$  and such that on each  $[\tau_j, \tau_{j+1}[$  the extremal control  $(U, V)$  satisfies  $Z = \omega^{\varepsilon j + h}$ , where  $h \in \mathbb{N}$ ,  $\varepsilon = \pm 1$ .*

*Proof.* Let us prove that the number of switching points is finite on the interval  $[0, \delta]$ . Assume that there exists a sequence  $(\tau_j)$  of switching points in  $[0, \delta]$  that converges to a point  $\tau \in [0, \delta]$ . We will show that

$$(3.17) \quad W(\tau) = 0, \quad \dot{W}(\tau) = 0.$$

Assume that  $Z$  rotates clockwise, that is,  $\varepsilon = \pm 1$ . We have by Lemma 3.5

$$W(\tau_j) = t_j \omega^{j+\frac{1}{2}}.$$

As  $W$  is of class  $C^{1,1}$  on  $[0, \pi]$ , the sequence  $(t_j)$  is bounded. Consequently (up to a subsequence), we can assume that the sequence  $(t_j)$  converges to a real  $t \in \mathbb{R}$ . Assume that  $t \neq 0$ ; then there exists  $j_0 \in \mathbb{N}$  such that, for  $j \geq j_0$ , we have  $t_j \neq 0$ . Hence,  $\frac{W(\tau_j)}{W(\tau_{j+1})}$  converges to 1 and

$$\frac{W(\tau_j)}{W(\tau_{j+1})} = \frac{t_j}{t_{j+1}} \bar{\omega},$$

which converges to  $\bar{\omega}$ . Thus  $t = 0$  and  $W(\tau) = 0$ . Again, we get a contradiction if we assume that  $\dot{W}(\tau) \neq 0$ . This shows (3.17). The Hamiltonian  $H$  along this extremal is 0. By (3.5) and by (3.14), the value of  $H$  is greater than the value of  $H$  for  $(\tilde{U}, \tilde{V}) = (0, 0)$ . It follows that  $W \equiv 0$  and  $Z \equiv 0$ . This extremal represents the disc, which is not a minimizer. An extremal trajectory then has a finite number of switching points. Finally, if we consider  $\omega^h$ ,  $h \in \mathbb{N}$ , the initial value of the control, and  $\varepsilon = \pm 1$ , the rotation clockwise or counterclockwise of the control, then we get the theorem.  $\square$

We now compute the exact number of switching points of an extremal. We prove the following result.

**THEOREM 3.4.** *Let  $W$  be an extremal trajectory and  $Z$  the extremal control. Then we have*

$$(3.18) \quad Z = \sum_{0 \leq j \leq s} \omega^{\varepsilon j + h} \mathbb{1}_{[\tau_j, \tau_{j+1}[},$$

where  $\varepsilon \in \{\pm 1\}$ ,  $h \in \mathbb{N}$ , and  $\tau_0 = 0 < \tau_1 < \dots < \tau_s < \tau_{s+1} = \delta$ . Moreover, the number  $L$  of switching points of  $Z$  in the interval  $[0, \delta]$  is given by

$$(3.19) \quad L = s + 1 = ln - \varepsilon, \quad l \in \mathbb{N}^*.$$

*Proof.* By Theorem 3.3, an extremal control  $Z$  takes the values  $(\omega^{\varepsilon j + h})_{0 \leq j \leq n-1}$  with  $h \in \mathbb{N}$  and  $\varepsilon = \pm 1$  on a finite subdivision of  $[0, \delta]$  denoted by  $(\tau_j)_{0 \leq j \leq s+1}$  with  $\tau_0 = 0$  and  $\tau_{s+1} = \delta$ . Without loss of generality, we can assume that  $\varepsilon = +1$ . If  $Z = \omega^h$  for  $\theta = 0^+$ , by performing a rotation of the control, that is, by changing  $Z$  into  $Z\bar{\omega}^h$ , we can always assume that  $Z(0^+) = 1$ . By extending the function  $W$  to  $\mathbb{R}$  by the relation  $W(\theta + \delta) = \bar{\omega}W(\theta)$  (recall that  $W$  is in  $\Gamma$ ), we can assume that 0 is a switching point. The function  $Z$  is now given by

$$Z = \sum_{0 \leq j \leq s} \omega^j \mathbb{1}_{[\tau_j, \tau_{j+1}[},$$

with  $\tau_0 = 0 < \tau_1 < \dots < \tau_s < \tau_{s+1} = \delta$ . As  $Z$  is in  $\mathcal{Z}$ , we must have  $Z(\delta^+) = \bar{\omega}Z(0^+) = \bar{\omega}$ . On the interval  $[\tau_s, \delta]$ , we have  $Z = \omega^s$ . Consequently, the point  $\delta$  is

a switching point, and we must have  $\omega^{s+1} = \bar{\omega}$ . Thus,  $s+1$  is of the form  $s+1 = -1+ln$ ,  $l \in \mathbb{N}^*$ . The number of switching points in the interval  $[0, \delta]$  is  $s+1$ , as  $\delta$  is not considered as a switching point of this interval. We have proved the theorem in the case where  $\varepsilon = +1$ . When the control  $Z$  satisfies  $Z = \bar{\omega}^j$ , the proof is the same, and we must have  $\bar{\omega}^{s+1} = \bar{\omega}$ . Consequently,  $s$  is given by  $s = ln$ ,  $l \in \mathbb{N}^*$ . In this case the number of switching points is  $s+1 = ln+1$ . This ends the proof of the theorem.  $\square$

In the case of regular rotors  $O_n^{ln \pm 1}$ , the switching points are of the form  $j\tau$ ,  $j = 1, \dots, s = ln \pm 1 - 1$  with  $\tau = \frac{\delta}{s+1}$ , and the associated control is given by (2.39). In the next section, we show that the distance between two consecutive switching points  $\tau_j$  and  $\tau_{j+1}$  of an extremal is constant. This will prove that a minimizer is necessarily a regular rotor.

An extremal  $(W, Z)$  given by (3.18) satisfies on each interval  $[\tau_j, \tau_{j+1}]$

$$(3.20) \quad W(\theta) = A_j e^{i\theta} + B_j e^{-i\theta} + \omega^{\varepsilon j + h}.$$

A simple computation using (3.14) shows that the Hamiltonian along this trajectory is

$$(3.21) \quad H = 2|A_j|^2 + 2|B_j|^2 - 1 \quad \forall 0 \leq j \leq s,$$

and, as  $H$  is constant, we have

$$|A_j|^2 + |B_j|^2 \equiv cst \quad \forall 0 \leq j \leq s.$$

**4. Conserved quantities along the extremal trajectories.** In this section we prove by an extension of Noether's theorem in optimal control theory that the angular momentum is conserved along an extremal trajectory. Combining the two conserved quantities (Hamiltonian and angular momentum) we will show that extremal trajectories describe regular rotors. We use the results of Torres (see [29], [30]) in order to prove the conservation of the angular momentum.

**4.1. Conservation of the angular momentum.** Let  $M$  be the function defined on the interval  $[0, \delta]$  by

$$M(\theta) = \Im((\overline{W}(\theta) - \overline{Z}(\theta))\dot{W}(\theta)), \quad \theta \in [0, \delta],$$

where  $(W(\theta), Z(\theta))$  is an admissible trajectory of problem (3.2). This quantity is usually called the *angular momentum* in mechanics (cross product between the position and the velocity). If  $(W(\theta), Z(\theta))$  is an extremal trajectory of (3.2) given by (3.18), we have, for  $0 \leq j \leq s$ , and  $\theta \in [\tau_j, \tau_{j+1}[$ ,

$$M(\theta) = \Im((\overline{W}(\theta) - \overline{\omega}^{\varepsilon j + h})\dot{W}(\theta)).$$

By differentiating, we get

$$\dot{M}(\theta) = 0 \quad \forall \theta \in [\tau_j, \tau_{j+1}[.$$

This proves that the function  $M(\theta)$  is piecewise constant on each  $[\tau_j, \tau_{j+1}]$ . We now show a stronger result.

**THEOREM 4.1.** *Along an extremal trajectory of (3.2), the quantity  $M(\theta)$  is constant:*

$$\forall \theta \in [0, \delta], \quad \dot{M}(\theta) = 0.$$

*Proof.* Let us consider the  $C^1$  transformation  $h^\alpha : \mathbb{C} \times \mathbb{C} \rightarrow \mathbb{C}$ ,  $\alpha \in \mathbb{R}_+$ , defined by

$$(4.1) \quad h^\alpha(W, Z) = e^{i\alpha}(W - Z) + Z.$$

Geometrically speaking,  $h^\alpha(W, Z)$  is the image of  $W - Z$  by the rotation of angle  $\alpha$  and whose center is  $Z$ . For any  $(W, Z) \in \mathbb{C}^2$ , we have  $h^0(W, Z) = W$ . Now, given an extremal trajectory  $(W(\theta), Z(\theta))$  of (3.2), we denote by  $W^\alpha$  the image of  $(W(\theta), Z(\theta))$  by  $h^\alpha$ . We then have on  $[0, \delta]$

$$\ddot{W}^\alpha + W^\alpha = Z.$$

Consequently,  $W^\alpha$  satisfies the same equation as  $W$ , and the extremal control associated with  $W^\alpha$  is  $Z$ . Let  $L : \mathbb{C} \times \mathbb{C}$  be the  $C^1$  map defined by

$$L(W, Z) = \Re(W\overline{Z}).$$

If  $(W(\theta), Z(\theta))$  is an extremal trajectory, we have

$$L(W^\alpha, Z) = \cos(\alpha)L(W, Z) - \sin(\alpha)\Im(W\overline{Z}) + 1 - \cos(\alpha).$$

Considering the  $C^1$  map  $F : \mathbb{C} \times \mathbb{C} \times \mathbb{R}_+ \rightarrow \mathbb{R}$  defined by

$$F(W, \dot{W}, \alpha) = -\sin(\alpha)\Im(W\overline{\dot{W}}),$$

we then have along an extremal trajectory  $(W(\theta), Z(\theta))$

$$L(W^\alpha(\theta), Z(\theta)) = \cos(\alpha)L(W(\theta), Z(\theta)) + \frac{d}{d\theta}F(W(\theta), \dot{W}(\theta), \alpha) + 1 - \cos(\alpha) \quad \forall \theta \in [0, \delta].$$

By (3.12), the scalar product between the state variable  $W$  and the dual variable  $\Pi$  is

$$\langle W, \Pi \rangle = -\Re(W\overline{\Pi}) + \Re(\dot{W}\overline{\Pi}).$$

Now we are in position to derive consequences of the invariance theorem (see [29]). Let  $(W(\theta), Z(\theta))$  be an extremal trajectory of (3.2),  $H$  the Hamiltonian along this trajectory, and  $\Pi(\theta)$  the dual variable. We then have

$$(4.2) \quad p_0 \frac{\partial F(W(\theta), Z(\theta), \alpha)}{\partial \alpha} \Big|_{\alpha=0} + \left\langle \frac{\partial W^\alpha(\theta)}{\partial \alpha} \Big|_{\alpha=0}, \Pi(\theta) \right\rangle - H \equiv cst$$

for all  $\theta \in [0, \delta]$ . But, we have

$$p_0 \frac{\partial F(W(\theta), Z(\theta), \alpha)}{\partial \alpha} \Big|_{\alpha=0} = -\Im(\overline{W}(\theta)\dot{W}(\theta)) \quad \forall \theta \in [0, \delta],$$

and by Lemma 3.3, we can take  $\Pi = -W$  so that

$$\left\langle \frac{\partial W^\alpha(\theta)}{\partial \alpha} \Big|_{\alpha=0}, \Pi(\theta) \right\rangle = \Im(\overline{\dot{W}}(\theta)Z(\theta)) + 2\Im(\dot{W}(\theta)\overline{W}(\theta)) \quad \forall \theta \in [0, \delta].$$

As the Hamiltonian is constant along an extremal trajectory, we get by (4.2)

$$\Im((\overline{W} - \overline{Z})\dot{W}) \equiv cst.$$

This ends the proof of the theorem.  $\square$

#### 4.2. Conserved quantities and equidistance of the switching points.

Thanks to the two conserved quantities along a Pontryagin extremal, we are now in position to prove the equidistance of the switching points. We first show the following proposition.

**PROPOSITION 4.1.** *For an extremal trajectory given by (3.18), we have for  $0 \leq j \leq s$ ,  $\tau_{j+1} - \tau_j = \tau_1 - \tau_0$ .*

*Proof.* A simple computation shows that for an extremal given by (3.18) we have on each  $[\tau_j, \tau_{j+1}[$ ,  $0 \leq j \leq s$ ,

$$(4.3) \quad M(\theta) = |A_j|^2 - |B_j|^2, \quad \theta \in [\tau_j, \tau_{j+1}[.$$

Thus, by (3.21) and Theorem 4.1, we get, for  $0 \leq j \leq s$ ,

$$(4.4) \quad \begin{cases} |A_j| = |A_0|, \\ |B_j| = |B_0|. \end{cases}$$

Since  $W$  is of class  $C^1$  at each switching point  $\tau_j$ , the coefficients  $A_j$  and  $B_j$ ,  $1 \leq j \leq s$ , are given by

$$(4.5) \quad \begin{cases} A_j = A_{j-1} + \frac{1}{2}(\omega^{j-1} - \omega^j)e^{-i\tau_j}, \\ B_j = B_{j-1} + \frac{1}{2}(\omega^{j-1} - \omega^j)e^{i\tau_j}. \end{cases}$$

Combining (4.4) and (4.5), we get

$$(4.6) \quad \Re(A_j \overline{A_{j-1}}) \equiv cst, \quad 1 \leq j \leq s.$$

Geometrically speaking, the complex  $(A_j)_{0 \leq j \leq s}$  lie on a circle whose center is the origin and whose radius is  $|A_0|$ , and  $A_{j+1}$  is the image of  $A_j$  by a rotation of a fixed angle by (4.6). In terms of the switching point  $(\tau_j)_{1 \leq j \leq s}$ , the phase between  $A_{j+1} - A_j$  and  $A_j - A_{j-1}$  is  $\delta - (\tau_{j+1} - \tau_j)$  by (4.5). But using (4.4) and (4.6), the phase between these two complex numbers is the same as the phase between  $A_j$  and  $A_{j-1}$ . By (4.6), the phase between  $A_{j+1} - A_j$  and  $A_j - A_{j-1}$  is constant, which ends the proof of the proposition.  $\square$

**COROLLARY 4.1.** *Let  $W$  be an extremal trajectory of the system and  $Z$  the extremal control. Then the corresponding rotor is in the class  $(O_n^{ln\pm 1})_{l \in \mathbb{N}^*}$ , and the extremal control  $Z$  is given by (2.39).*

*Proof.* This is a consequence of the previous proposition, as two consecutive switching points of an extremal are equidistant. The corresponding rotor given by (3.18) satisfies  $\tau_{j+1} - \tau_j \equiv cst$ , and it is necessarily an element of  $(O_n^{ln\pm 1})_{l \in \mathbb{N}^*}$ .  $\square$

By Proposition 2.11, the rotor of minimal area in the class  $O_n^{ln\pm 1}$  is  $O_n^{n-1}$  (with the least number of arcs). As the rotor of minimal area necessarily belongs to this class (by the PMP), it is  $O_n^{n-1}$ . By (2.39) the optimal control  $Z_{min}$  corresponding to  $O_n^{n-1}$  is obtained for  $s+1 = n-1$  and is given by

$$(4.7) \quad Z_{min} = \sum_{0 \leq j \leq n-2} \omega^j \mathbb{1}_{[j \frac{\delta}{n-1}, (j+1) \frac{\delta}{n-1}[}.$$

This proves Goldberg's conjecture (Theorem 1.1). Note that there is no necessity of verifying the optimality of the extremal trajectories corresponding to  $(O_n^{ln\pm 1})_{l \in \mathbb{N}^*} \setminus \{O_n^{n-1}\}$ , as  $O_n^{n-1}$  is of minimal area in this class.

TABLE 1  
Values of the  $r_j$  for  $n = 3, 4, 5, 6$ .

$n$	$r_0$	$r_1$	$r_2$	$r_3$	$r_4$	$r_5$
3	0	$3r$	0			
4	0	$2r$	$2r$	0		
5	0	$r_1$	$r_2$	$r_1$	0	
6	0	$r$	$2r$	$2r$	$r$	0

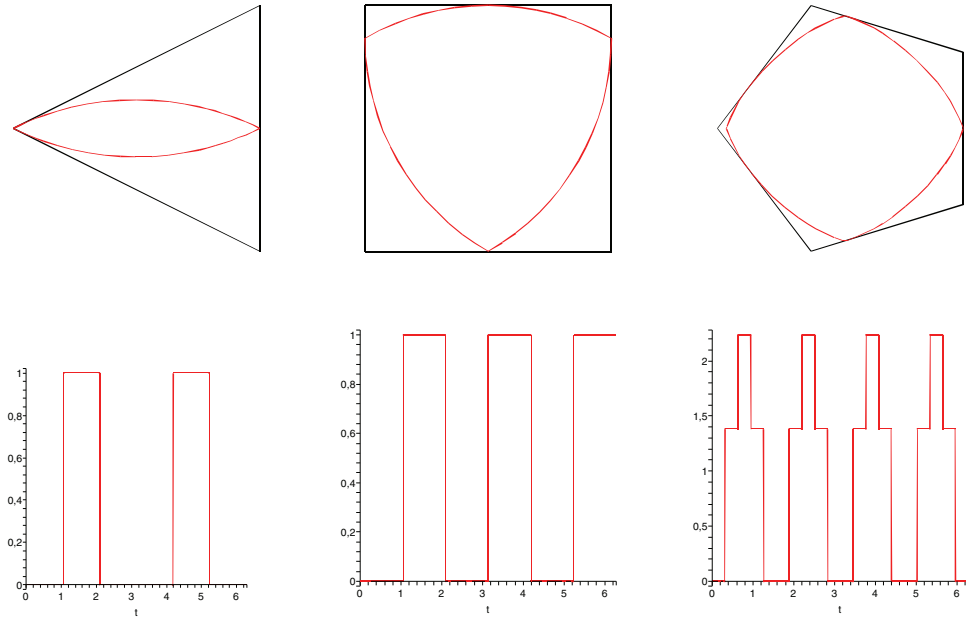


FIG. 3. The minimizers in the cases  $n = 3$  ( $\Delta$ -biangle),  $n = 4$  (Reuleaux triangle), and  $n = 5$  ( $O_5^4$ ) and their respective radii of curvature on the interval  $[0, 2\pi]$ .

Geometrically speaking, if we come back to the initial parameterization of a rotor by its support function  $p$ , the rotor  $O_n^{n-1}$  is the union of arcs of circles of radii  $r_j$ :

$$\begin{aligned} r_j &= \frac{r}{\cos\left(\frac{\delta}{2}\right)} \left( \cos\left(\frac{\delta}{2}\right) - \cos\left(\left(j + \frac{1}{2}\right)\delta\right) \right) \\ &= \frac{r}{\cos\left(\frac{\delta}{2}\right)} \Re(\omega^{1/2} - \omega^{j+1/2}), \quad j = 0, \dots, n-1. \end{aligned}$$

These values of the radii of curvature are precisely equal to the distances of the diagonals of the  $n$ -gon from the parallel sides (see [15], [16]), and the sectors are all equal to  $\frac{2\pi}{n(n-1)}$ , as the switching points are equidistant. In Table 1, we give the different values of the radii  $r_j$  for  $n = 3, 4, 5, 6$ . For  $n = 5$ , there are two different radii  $r_1 < r_2$ , and  $r$  denotes the radius of the inscribed circle. If  $n$  is even, there are exactly  $\frac{n-2}{2}$  values of the  $r_j$ , and if  $n$  is odd, there are exactly  $\frac{n-1}{2}$  values of the  $r_j$ . By (2.39), the radius of curvature of the boundary of  $O_n^{n-1}$  is  $\frac{2\pi}{n-1}$ -periodic (see [10]). We have represented in Figure 3 the minimizers of the area for  $n = 3$ ,  $n = 4$ , and  $n = 5$  and their respective radii of curvature.

**Acknowledgments.** The author is grateful to to E. Trélat for some helpful advice and to J. P. François for some helpful discussions.

## REFERENCES

- [1] A. AGRACHEV AND Y. L. SACHKOV, *Control Theory from the Geometric Viewpoint*, Springer-Verlag, Berlin, 2004.
- [2] J. A. ANDREJEWA AND R. KLÖTZLER, *Zur analytischen Lösung geometrischer Optimierungsaufgaben mittels Dualität bei Steuerungsproblemen I*, Z. Angew. Math. Mech., 64 (1984), pp. 35–44.
- [3] J. A. ANDREJEWA AND R. KLÖTZLER, *Zur analytischen Lösung geometrischer Optimierungsaufgaben mittels Dualität bei Steuerungsproblemen II*, Z. Angew. Math. Mech., 64 (1984), pp. 147–153.
- [4] T. BAYEN, T. LACHAND-ROBERT, AND E. OUDET, *Analytic parametrization of three-dimensional bodies of constant width*, Arch. Ration. Mech. Anal., 186 (2007), pp. 225–249.
- [5] W. BLASCHKE, *Konvexe Bereiche gegebener konstanter Breite und kleinsten Inhalts*, Math. Ann., 76 (1915), pp. 504–513.
- [6] T. BONNESEN AND W. FENCHEL, *Theory of Convex Bodies*, BCS Associates, Moscow, ID, 1987.
- [7] G. D. CHAKERIAN AND H. GROEMER, *Convex bodies of constant width*, in Convexity and Its Applications, Birkhäuser, Basel, 1983, pp. 49–96.
- [8] B. DACOROGNA, *Introduction to the Calculus of Variations*, Imperial College Press, London, 2004.
- [9] W. J. FIREY, *Isoperimetric ratios of Reuleaux polygons*, Pacific J. Math., 10 (1960), pp. 823–829.
- [10] J. FOCKE, *Symmetrische  $n$ -Orbiformen kleinsten Inhalts*, Acta Math. Acad. Sci. Hungar., 20 (1969), pp. 39–68.
- [11] M. FUJIWARA, *Analytical proof of Blaschke's theorem on the curve of constant breadth with minimum area*, Proc. Tokyo Imp. Acad. Japan, 3 (1927), pp. 307–309.
- [12] M. FUJIWARA, *Analytical proof of Blaschke's theorem on the curve of constant breadth with minimum area*, II, Proc. Tokyo Imp. Acad. Japan, 7 (1931), pp. 300–302.
- [13] M. FUJIWARA AND S. KAKEYA, *On some problems of maxima and minima for the curve of constant breadth and the in-resolvable curve of the equilateral triangle*, Tôhoku Math. J., 11 (1917), pp. 92–110.
- [14] M. GHANDEHARI, *An optimal control formulation of the Blaschke-Lebesgue theorem*, J. Math. Anal. Appl., 200 (1996), pp. 322–331.
- [15] M. GOLDBERG, *Trammel rotors in regular polygons*, Amer. Math. Monthly, 64 (1957), pp. 71–78.
- [16] M. GOLDBERG, *Rotors in polygons and polyhedra*, Math. Comput., 14 (1960), pp. 229–239.
- [17] P. M. GRUBER AND J. M. WILLS, EDS., *Handbook of Convex Geometry*, Vols. A and B, North-Holland, Amsterdam, 1993.
- [18] E. HARRELL, *A direct proof of a theorem of Blaschke and Lebesgue*, J. Geom. Anal., 12 (2002), pp. 81–88.
- [19] A. HENROT, *Extremum Problems for Eigenvalues of Elliptic Operators*, Birkhäuser, Basel, 2006.
- [20] R. HOWARD, *Convex bodies of constant width and constant brightness*, Adv. Math., 204 (2006), pp. 241–261.
- [21] R. KLÖTZLER, *Beweis einer Vermutung über  $n$ -Orbiformen kleinsten Inhalts*, Z. Angew. Math. Mech., 55 (1975), pp. 557–570.
- [22] H. LEBESGUE, *Sur quelques questions de minimum, relatives aux courbes orbiformes, et sur leurs rapports avec le calcul des variations*, J. Math. Pures Appl. (8), 4 (1921), pp. 67–96.
- [23] E. MEISSNER, *Über die Anwendung von Fourierreihen auf einige Aufgaben der Geometrie und Kinematik*, Vierteljahresschr. Naturfor. Ges. Zürich, 54 (1909), pp. 309–329.
- [24] E. MEISSNER, *Über Punktmengen konstanter Breite*, Vierteljahresschr. Naturfor. Ges. Zürich, 56 (1911), pp. 42–50.
- [25] E. MEISSNER, *Drei Gipsmodelle von Flächen konstanter Breite*, Z. Math. Phys., 60 (1912), pp. 92–94.
- [26] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, John Wiley and Sons, New York, London, 1962.
- [27] F. REULEAUX, *The Kinematics of Machinery: Outline of a Theory of Machines*, Macmillan, London, 1876.
- [28] R. SCHNEIDER, *Convex Bodies: The Brunn-Minkowski Theory*, Cambridge University Press, Cambridge, UK, 1993.

- [29] D. F. M. TORRES, *Conserved quantities along the Pontryagin extremals of quasi-invariant optimal control problems*, in Proceedings of the 10th Mediterranean Conference on Control and Automation, 2002.
- [30] D. F. M. TORRES, *On the Noether invariance principle for constrained optimal control problems*, WSEAS Trans. Math., 3 (2004), pp. 620–624.
- [31] E. TRÉLAT, *Contrôle Optimal*, Vuibert, Paris, 2005.
- [32] F. A. VALENTINE, *Convex Sets*, McGraw–Hill, New York, Toronto, London, 1964.
- [33] B. WEISSBACH, *Rotoren im regulären Dreieck*, Publ. Math. Debrecen, 19 (1972), pp. 21–27.
- [34] I. M. YAGLOM AND V. G. BOLTYANSKII, *Convex Figures*, Holt, Rinehart and Winston, New York, 1961.



## COMPOSITE SYSTEMS WITH UNCERTAIN COUPLINGS OF FIXED STRUCTURE: SCALED RICCATI EQUATIONS AND THE PROBLEM OF QUADRATIC STABILITY\*

DIEDERICH HINRICHSSEN<sup>†</sup> AND ANTHONY J. PRITCHARD<sup>‡</sup>

*Tony Pritchard, my friend and close collaborator of many years, died suddenly from a heart attack on the 12th of August 2007 while on holidays with his family. Some weeks before his death, we had agreed that after a final revision he would submit this paper to SICON as the corresponding author.—Diederich Hinrichsen.*

**Abstract.** We consider large scale systems consisting of a finite number of separate uncertain subsystems which interact via uncertain couplings of arbitrarily prescribed structure. The couplings are viewed as structured perturbations of the block-diagonal system representing the collection of the separate nominal subsystems (the “nominal system”). We define spectral value sets and stability radii for these time-invariant structured perturbations and derive formulas for their computation. Scaled Riccati equations are introduced to obtain explicit formulas for the stability radii with respect to time-varying and possibly nonlinear perturbations of the given structure. From these we derive necessary and sufficient conditions under which the stability radii with respect to time-invariant and time-varying perturbations are equal. These results are obtained by constructing joint quadratic Liapunov functions of optimal robustness. With their help we prove necessary and sufficient conditions for quadratic stability and sufficient conditions for the validity of a generalized Aizerman conjecture.

**Key words.** interconnected system, Riccati equation, structured perturbation, spectral value set, stability radius, time-varying perturbation, quadratic stability, Aizerman conjecture

**AMS subject classifications.** 15A18, 93C05, 93C73, 93D09

**DOI.** 10.1137/070707919

**1. Introduction.** Composite systems play a role in many different areas of application. They arise naturally where large scale systems are composed of a finite number of separate interacting subsystems. Examples of such systems are traffic, biochemical [2], and power networks [6]; compartmental models in physiology and ecology [15]; and systems of cooperative robotic vehicles [10]; see also [3], [20]. Often in these composite systems the interconnection structure is fixed, but the magnitudes of the couplings between the subsystems are uncertain and may even change in time or be dependent on current states. The *interconnection structure* specifies for each subsystem  $\Sigma_i$  the set of subsystems  $\Sigma_j$  which can directly influence it. A structure of this type may be described by a graph, and tools from graph theory have been used by several authors to analyze such systems; see, e.g., [8], [22]. Due to economic costs, reliability, and availability of information flows, etc. there will also, in general, be structural constraints imposed on the control of composite systems. For instance, the subsystems may be separated geographically and for each one only local measurements may be available for feedback control. This leads to problems of *decentralized control* which have been studied extensively in the literature; see, e.g., [3], [21]. As a result of such structural constraints closed loop interconnected systems will also, in general, not be fully coupled but have a fixed interconnection structure.

---

\*Received by the editors November 10, 2007; accepted for publication (in revised form) June 19, 2008; published electronically January 7, 2009.

<http://www.siam.org/journals/sicon/47-6/70791.html>

<sup>†</sup>Zentrum für Technomathematik, Universität Bremen, 28334 Bremen, Germany (dh@math.uni-bremen.de).

<sup>‡</sup>Former address: Institute of Mathematics, University of Warwick, Coventry, UK.

In this paper we do not deal with control aspects, but our results may provide an aid to the design of decentralized feedback controls. We analyze stability properties of composite systems consisting of (possibly uncertain) subsystems interacting via uncertain couplings of a fixed structure. The couplings are viewed as structured perturbations of the block-diagonal system representing the collection of the disconnected nominal subsystems. This block-diagonal system will be regarded as the “nominal system”. We assume that the individual nominal subsystems are stable or have been stabilized (e.g., via decentralized control). Then a basic question is whether the overall system remains stable provided the uncertainties (possibly time-varying) are bounded in norm by some given realistic uncertainty level.

There is extensive literature on stability properties of composite systems. Various tools such as input-output methods [24], [3], passivity methods, scalar and vector Liapunov functions [17], [20], and (dynamic) graphs [22] have been used to obtain sufficient stability criteria. Our aim is to develop *nonconservative* results on the stability of uncertain composite systems with norm bounded couplings and an arbitrary fixed interconnection structure. If the interconnection structure is a priori known, then the application of robustness estimates for unstructured perturbations will, in general, lead to conservative results. On the other hand, it is well known from  $\mu$ -analysis that tight robustness margins for structured perturbations are not easily obtained.

In a first step we determine the *spectral value sets* [12] associated with uncertain composite systems of the above type, i.e., the set of eigenvalues of all the composite systems which are obtained from the block-diagonal nominal one by adding perturbations which preserve the prescribed interaction structure and are bounded in an appropriate norm by a given uncertainty level  $\delta > 0$ . We then derive computable formula for the *stability radius* [12] of the nominal system with respect to these (complex time-invariant) perturbations. Similar results, but with respect to different perturbation norms, have been obtained in the recent paper [16], and it has been shown there that for the special case of *one-dimensional* subsystems these results are closely related to classical spectral inclusion theorems of linear algebra due to Gershgorin, Brauer, and Brualdi. In fact, they can be used to show that the inclusion regions of Brauer and Brualdi (see [14], [23]) are tight for the corresponding perturbation structures.

In [16] *constant* variations of the coupling parameters were considered. However, time-varying couplings may often occur in practice. The notion of *connective stability* was introduced by Siljak [20] to determine conditions under which a composite system remains stable despite time-varying perturbations whereby subsystems are disconnected and again connected during the evolution of the system. It may also occur that some of the interconnected subsystems are nonlinear or time-varying, but for simplicity they have been modelled approximately by time-invariant ones. Then the “real” subsystems can be viewed as nonlinear or time-varying perturbations of the time-invariant nominal subsystems. In this paper we will focus on the establishment of nonconservative stability results for uncertain composite systems with such *time-varying and/or state-dependent* uncertainties, both in the couplings and in the individual subsystems.

In general, the determination of the stability radius with respect to time-varying structured perturbations is a difficult problem [7], [26]. But if no perturbation structure is prescribed and arbitrary complex matrix perturbations are allowed (“complex full block case”), then the situation is quite simple. It is known that in the complex full block case the stability radii with respect to time-invariant and time-varying perturbations coincide [12, section 5.6]. This result (which does *not* carry over to *real*

perturbations) has been obtained in [11] by characterizing the complex stability radius via a parametrized Riccati equation. By means of the Riccati equation a quadratic Liapunov function of maximal robustness can be constructed. The construction of such Liapunov functions for *structured* perturbations has been stated as an open problem in [16, Remark 3.2].

In this paper we will show that for uncertain block-diagonal systems with irreducible perturbation structures quadratic Liapunov functions of maximal robustness can be constructed by applying a scaling technique known from  $\mu$ -analysis. Via scaling, new parametrized algebraic Riccati equations are obtained whose Hermitian solutions provide us with quadratic Liapunov functions of approximately optimal robustness (optimality is achieved if the perturbation structure is irreducible). By means of these quadratic Liapunov functions one can derive computable formulas for the stability radius of uncertain block-diagonal systems with respect to *time-varying* perturbations of the prescribed structure. In general, the stability radii with respect to time-varying and time-invariant perturbations are different. We will derive necessary and sufficient conditions for them to be equal.

The robust stability problem with respect to time-varying perturbations is closely related to the more general stability problem for linear differential inclusions of the form  $\dot{x}(t) \in Ax(t)$ , where  $A$  is a compact set of  $n \times n$  matrices. Molchanov and Pyatnitskij [18] have shown that such a differential inclusion is asymptotically stable, if and only if there exists a joint convex positive homogeneous strict Liapunov function for all the time-invariant systems  $\dot{x} = Ax$ ,  $A \in A$ . This result shows a fundamental difference between robustness issues for time-varying and for time-invariant perturbations. In general, the stability of a set of time-invariant systems  $\dot{x} = Ax$ ,  $A \in A$ , does not guarantee the existence of a joint Liapunov function for all these systems. In contrast, if all the *time-varying* systems  $\dot{x}(t) = A(t)x(t)$ ,  $A(\cdot) : \mathbb{R}_+ \rightarrow A$  measurable, are uniformly asymptotically stable then, by Molchanov and Pyatnitskij's theorem, such a joint Liapunov function always exists.

A more specific problem is to characterize those sets of time-invariant linear systems for which a joint *quadratic* Liapunov function can be found. This is the problem of *quadratic stability*; see [5]. It is well known that for uncertain systems with full block perturbations, quadratic stability is equivalent to asymptotic stability; see [11]. Moreover, it is known that this equivalence does not, in general, hold for *structured perturbations* [19]. To the best of our knowledge, there are no general quadratic stability criteria available for structured perturbations besides reformulations of the problem in terms of linear matrix inequalities; see [1]. In this paper we consider block-diagonal systems under bounded perturbations of arbitrarily prescribed structure and derive explicit necessary and sufficient criteria for the quadratic stability of these systems. To check these criteria one only has to determine the spectral radius of a certain matrix; see Theorem 7.7.

The organization of this paper is as follows. In the next section we introduce the concepts of *spectral value set* and *stability radius* and for the special case of *full block* perturbations recall some characterizations of them via transfer functions and parametrized algebraic Riccati equations. The section concludes with the proof of a stability theorem for nonlinear time-varying full block perturbations. After fixing, in section 3, the terminology and notations for the analysis of composite systems, computable formulae for the corresponding spectral value sets and stability radii are presented in section 4. To derive these formulae, we apply the Perron–Frobenius theory of nonnegative matrices. We do not employ tools from  $\mu$ -analysis, but briefly express our results in terms of  $\mu$ -values. In section 5, we introduce scaled parametrized alge-

braic Riccati equations and show how they can be used to construct joint quadratic Liapunov functions of approximately optimal robustness (optimal robustness if the underlying perturbation structure is irreducible). The scaled Riccati equations introduced in section 5 will be the key tool for proving the main results of this paper. In section 6, we derive a computable formula for the stability radius of a block-diagonal system with respect to *time-varying* parameter perturbations. As a consequence we obtain necessary and sufficient conditions under which the stability radii with respect to time-invariant and time-varying perturbations of the prescribed structure coincide. In section 7, we show that this equality holds if the subsystems are either one-dimensional or positive. We then conclude this paper by applying our results to two classical robustness topics, the problems of *quadratic* and of *absolute* stability.

**2. Preliminaries.** In this section we introduce some basic concepts and fix the notation. The symbols  $\mathbb{N}, \mathbb{R}, \mathbb{R}_+, \mathbb{C}$  denote the sets of positive integers, real numbers, nonnegative real numbers, and complex numbers, respectively. For any  $N \in \mathbb{N}$  we set  $\underline{N} := \{1, 2, \dots, N\}$ . By  $\mathbb{K}^{n \times m}$  we denote the set of  $n$  by  $m$  matrices with entries in  $\mathbb{K}$ , where  $\mathbb{K} = \mathbb{R}$  or  $\mathbb{C}$ . If  $A = (a_{ij}) \in \mathbb{C}^{n \times m}$ , then we define  $|A| := (|a_{ij}|)$ , and for real matrices  $A = (a_{ij}), B = (b_{ij}) \in \mathbb{R}^{n \times m}$  we write  $A \leq B$  if  $a_{ij} \leq b_{ij}$  for all  $i \in \underline{n}, j \in \underline{m}$ . If  $A$  is square, then  $\sigma(A)$ ,  $\rho(A) = \mathbb{C} \setminus \sigma(A)$ , and  $\alpha(A) = \max\{\operatorname{Re} \lambda; \lambda \in \sigma(A)\}$  denote its *spectrum*, *resolvent set*, and *spectral abscissa*, respectively. By  $\mathbb{L}_{n,l,q}$  we denote the set of triples of matrices  $(A, B, C)$  with  $A \in \mathbb{C}^{n \times n}, B \in \mathbb{C}^{n \times l}, C \in \mathbb{C}^{q \times n}$ ,  $n, l, q \in \mathbb{N}$ . The open left half-plane is denoted by  $\mathbb{C}_- = \{s \in \mathbb{C}; \operatorname{Re} s < 0\}$ , and  $A \in \mathbb{C}^{n \times n}$  is called *Hurwitz stable* if  $\sigma(A) \subset \mathbb{C}_-$ . We use the conventions

$$(1) \quad 0^{-1} = \infty, \quad \infty^{-1} = 0, \quad \inf \emptyset = \infty.$$

In the following definitions we suppose that  $(A, B, C) \in \mathbb{L}_{n,l,q}$  and  $\Delta \subset \mathbb{K}^{l \times q}$  is a  $\mathbb{K}$ -linear subspace provided with a norm  $\|\cdot\|_\Delta$ . For a more detailed account of the definitions and some results presented in this section, see [12]. We consider perturbations of the form

$$(2) \quad A \rightsquigarrow A_\Delta = A + B\Delta C, \quad \Delta \in \Delta.$$

DEFINITION 2.1. The  $\mu$ -value of a matrix  $M \in \mathbb{C}^{q \times l}$  (with respect to the normed perturbation space  $(\Delta, \|\cdot\|_\Delta)$ ) is defined by

$$(3) \quad \mu_\Delta(M) := [\inf\{\|\Delta\|_\Delta; \Delta \in \Delta, \det(I_l - \Delta M)\}]^{-1}.$$

DEFINITION 2.2. Given a system  $(A, B, C) \in \mathbb{L}_{n,l,q}$  and the perturbation space  $(\Delta, \|\cdot\|_\Delta)$ , the spectral value set of  $A$  of level  $\delta > 0$ , with respect to perturbations of the form (2), is the following subset of the complex plane:

$$(4) \quad \sigma_\Delta(A, B, C; \delta) := \bigcup_{\Delta \in \Delta, \|\Delta\|_\Delta < \delta} \sigma(A + B\Delta C).$$

DEFINITION 2.3. Given a system  $(A, B, C) \in \mathbb{L}_{n,l,q}$  and the perturbation space  $(\Delta, \|\cdot\|_\Delta)$ , the stability radius of  $A$  with respect to perturbations of the form (2) is defined by

$$(5) \quad r_\Delta(A, B, C) = \inf\{\|\Delta\|_\Delta; \Delta \in \Delta, \sigma(A + B\Delta C) \not\subset \mathbb{C}_-\}.$$

It is easily seen that the infimum in (5) is in fact a minimum if  $r_{\Delta}(A, B, C)$  is finite. In this case the stability radius is the norm of a smallest perturbation in  $\Delta$  which destabilizes  $A$ .  $r_{\Delta}(A, B, C) = 0$  if and only if  $\sigma(A) \not\subset \mathbb{C}_-$ .  $r_{\Delta}(A, B, C) = \infty$  if and only if  $\sigma(A + B\Delta C) \subset \mathbb{C}_-$  for all  $\Delta \in \Delta$ .

Spectral value sets and stability radii can be characterized via  $\mu$ -values as follows; see [12].

PROPOSITION 2.4. *Let  $(A, B, C) \in \mathbb{L}_{n,l,q}$  be a given system and  $G(s) = C(sI_n - A)^{-1}B$  the associated transfer function. Then*

$$(6) \quad \sigma_{\Delta}(A, B, C; \delta) = \sigma(A) \cup \{s \in \rho(A); \mu_{\Delta}(G(s)) > \delta^{-1}\}, \quad \delta > 0;$$

$$(7) \quad r_{\Delta}(A, B, C) = \left( \sup_{\omega \in \mathbb{R}} \mu_{\Delta}(G(i\omega)) \right)^{-1} \quad \text{if } \sigma(A) \subset \mathbb{C}_-.$$

Specializing to *block-diagonal* and *full block* perturbations, further results are known if the perturbation norm  $\|\cdot\|_{\Delta}$  is an operator norm. Let  $\mathbb{C}^l, \mathbb{C}^q$  be endowed with arbitrary norms and  $\mathbb{C}^{l \times q}, \mathbb{C}^{q \times l}$  with the induced operator norms  $\|\cdot\|_{\mathcal{L}(\mathbb{C}^q, \mathbb{C}^l)}$  and  $\|\cdot\|_{\mathcal{L}(\mathbb{C}^l, \mathbb{C}^q)}$ . In the case where  $\mathbb{C}^l, \mathbb{C}^q$  are provided with 2-norms, we write  $\|\cdot\|_2$  for vector norms and the corresponding operator norms are denoted by  $\|\cdot\|_{2,2}$ . Suppose  $\Delta \subset \mathbb{C}^{l \times q}$  is the vector space of block-diagonal perturbations of the following form provided with the corresponding operator norm:

$$(8) \quad \Delta = \{\text{diag}(\Delta_1, \dots, \Delta_N); \Delta_i \in \mathbb{C}^{l_i \times q_i}, i \in \underline{N}\}, \quad \|\cdot\|_{\Delta} = \|\cdot\|_{\mathcal{L}(\mathbb{C}^q, \mathbb{C}^l)},$$

where  $N \geq 1$ ,  $l_i \geq 1$ ,  $q_i \geq 1$ ,  $i \in \underline{N}$  are given such that  $\sum_{i=1}^N l_i = l$ ,  $\sum_{i=1}^N q_i = q$ . Then estimates of the associated  $\mu$ -values, spectral value sets, and stability radii can be obtained by the following scaling method. For any scaling vector  $\gamma = (\gamma_1, \dots, \gamma_N)$ , where  $\gamma_i > 0$ ,  $i \in \underline{N}$ , we set  $L_{\gamma} = \text{diag}(\gamma_1 I_{l_1}, \dots, \gamma_N I_{l_N})$  and  $R_{\gamma} = \text{diag}(\gamma_1 I_{q_1}, \dots, \gamma_N I_{q_N})$ . Then  $L_{\gamma} \Delta R_{\gamma}^{-1} = \Delta$  for all  $\Delta \in \Delta$  and this fact implies (see [12, section 4.4]) that

$$(9) \quad \mu_{\Delta}(G) \leq \|R_{\gamma} G L_{\gamma}^{-1}\|_{\mathcal{L}(\mathbb{C}^l, \mathbb{C}^q)}, \quad G \in \mathbb{C}^{q \times l}.$$

As a consequence we have

$$(10) \quad \sigma_{\Delta}(A, B, C; \delta) \subset \sigma(A) \cup \{s \in \rho(A); \|R_{\gamma} G(s) L_{\gamma}^{-1}\|_{\mathcal{L}(\mathbb{C}^l, \mathbb{C}^q)} > \delta^{-1}\};$$

$$(11) \quad r_{\Delta}(A, B, C) \geq \left( \sup_{\omega \in \mathbb{R}} \|R_{\gamma} G(i\omega) L_{\gamma}^{-1}\|_{\mathcal{L}(\mathbb{C}^l, \mathbb{C}^q)} \right)^{-1} \quad \text{if } \sigma(A) \subset \mathbb{C}_-.$$

In the *full block case* where  $\Delta = \mathbb{K}^{l \times q}$  precise formulae are obtained without any scaling. For this case the spectral value sets and stability radii are denoted by  $\sigma_{\mathbb{K}}(A, B, C; \delta)$  and  $r_{\mathbb{K}}(A, B, C)$ , respectively, and are called the *complex* and *real* spectral value sets and stability radii according to whether  $\mathbb{K} = \mathbb{C}$  or  $\mathbb{K} = \mathbb{R}$ . For the complex case ( $\Delta = \mathbb{C}^{l \times q}$ ,  $\|\cdot\|_{\Delta} = \|\cdot\|_{\mathcal{L}(\mathbb{C}^q, \mathbb{C}^l)}$ ), one has  $\mu_{\mathbb{C}^{l \times q}}(G(s)) = \|G(s)\|_{\mathcal{L}(\mathbb{C}^l, \mathbb{C}^q)}$ , and therefore Proposition 2.4 implies

$$(12) \quad \sigma_{\mathbb{C}}(A, B, C; \delta) = \sigma(A) \cup \{s \in \rho(A); \|G(s)\|_{\mathcal{L}(\mathbb{C}^l, \mathbb{C}^q)} > \delta^{-1}\};$$

$$(13) \quad r_{\mathbb{C}}(A, B, C) = \left( \max_{\omega \in \mathbb{R}} \|G(i\omega)\|_{\mathcal{L}(\mathbb{C}^l, \mathbb{C}^q)} \right)^{-1} \quad \text{if } \sigma(A) \subset \mathbb{C}_-.$$

The formula for the real stability radius  $r_{\mathbb{R}}(A, B, C)$  (where  $A, B, C$  are supposed to be *real*) is more complicated; see [12, section 5.3.3].

In the following we will assume that both  $\mathbb{C}^l$  and  $\mathbb{C}^q$  are provided with the 2-norm  $\|\cdot\|_2$  so that the norms  $\|\cdot\|_{\mathcal{L}(\mathbb{C}^q, \mathbb{C}^l)}, \|\cdot\|_{\mathcal{L}(\mathbb{C}^l, \mathbb{C}^q)}$  are both spectral norms.<sup>1</sup> Then

$$(14) \quad r_{\mathbb{C}}(A, B, C) = \left( \max_{\omega \in \mathbb{R}} \|G(i\omega)\|_{2,2} \right)^{-1} = \|G(\cdot)\|_{H^\infty}^{-1} \quad \text{if } \sigma(A) \subset \mathbb{C}_-.$$

Throughout the rest of this paper we will reserve the notation  $r_{\mathbb{C}}(A, B, C)$  for the complex stability radius with respect to the spectral norm on  $\mathbb{C}^{l \times q}$ . This stability radius can be characterized in terms of the following parametrized algebraic Riccati equation:

$$(15) \quad PA + A^*P + \rho^2 C^*C + PBB^*P = 0,$$

where  $\rho \geq 0$ . We denote the real vector space of all the Hermitian matrices in  $\mathbb{C}^{n \times n}$  by  $\mathcal{H}_n(\mathbb{C})$  and the usual order relation on  $\mathcal{H}_n(\mathbb{C})$  by  $\preceq$ . The following results have been proved in [11].<sup>2</sup>

**THEOREM 2.5.** *Suppose that  $(A, B, C) \in \mathbb{L}_{n,l,q}$  is a given system with  $\sigma(A) \subset \mathbb{C}_-$ , and  $G(s) = C(sI_n - A)^{-1}B$  is the associated transfer matrix. Then (15) has a Hermitian solution if and only if  $\rho \leq r_{\mathbb{C}} := r_{\mathbb{C}}(A, B, C)$ . Moreover, the following statements are equivalent:*

- (i) *There exists a (unique) solution  $P_\rho \in \mathcal{H}_n(\mathbb{C})$  of (15) such that  $\sigma(A + BB^*P_\rho) \subset \mathbb{C}_-$ .*
- (ii)  *$\rho < r_{\mathbb{C}}$ .*

*If  $\rho = r_{\mathbb{C}}$ , then (15) has a unique solution  $P_{r_{\mathbb{C}}} \in \mathcal{H}_n(\mathbb{C})$  satisfying  $\sigma(A + BB^*P_{r_{\mathbb{C}}}) \subset \mathbb{C}_-$ . For each  $\rho \leq r_{\mathbb{C}}$ ,  $P_\rho$  is the smallest Hermitian solution of (15). If  $P$  is any Hermitian solution of (15), then  $P \succeq 0$ .  $P$  is positive definite if  $(A, C)$  is observable. If  $P$  is positive definite, then  $V_\rho(x) = \langle x, Px \rangle$  is a joint quadratic Liapunov function for all perturbed systems*

$$(16) \quad \dot{x} = A_\Delta x = (A + B\Delta C)x, \quad \Delta \in \mathbb{C}^{l \times q}, \quad \|\Delta\|_{2,2} \leq \rho.$$

Since there is no joint Liapunov function for all perturbed systems  $\dot{x} = A_\Delta x$  with  $\|\Delta\|_{2,2} < \rho$  if  $\rho > r_{\mathbb{C}}(A, B, C)$ , we may call the quadratic Liapunov function  $V_{r_{\mathbb{C}}}(x)$  (in case  $P \succ 0$ ) one of *maximal robustness* for the perturbation space  $\Delta = \mathbb{C}^{l \times q}$  endowed with the spectral norm.

We conclude this section with a theorem which extends Proposition 5.2 in [11] and will be useful for the treatment of time-varying nonlinear perturbations in section 6. Since the proof in [11] works only for time-invariant perturbations, we will give a full proof.

Let  $(A, B, C) \in \mathbb{L}_{n,l,q}$  be a given system with  $\sigma(A) \subset \mathbb{C}_-$ . We suppose that  $\Omega$  is an open neighborhood of 0 in  $\mathbb{C}^n$  and consider nonlinear time-varying perturbations of  $\dot{x} = Ax$  of “output feedback form”,

$$(17) \quad \dot{x} = Ax + B\Delta(x, t)y, \quad y = Cx, \quad \text{where } \Delta(\cdot, \cdot) \in \Delta_{nt}(\Omega).$$

Here  $\Delta_{nt}(\Omega)$  is the vector space of all *bounded* matrix functions  $\Delta(\cdot, \cdot) : \Omega \times \mathbb{R}_+ \rightarrow \mathbb{C}^{l \times q}$  with the *Carathéodory properties*; i.e.,  $\Delta(x, \cdot) : \mathbb{R}_+ \rightarrow \mathbb{C}^{l \times q}$  is measurable for each

<sup>1</sup>For any Hermitian matrix  $H$  we denote by  $\lambda_{\max}(H)$  the maximal eigenvalue of  $H$ . For any matrix  $M \in \mathbb{C}^{h \times k}$ ,  $h, k \in \mathbb{N}$ , we denote by  $\|M\|_{2,2} = [\lambda_{\max}(M^*M)]^{1/2} = [\lambda_{\max}(MM^*)]^{1/2}$  the spectral norm of  $M$ .

<sup>2</sup>Note that  $P$  is a solution of (15) if and only if  $-P$  satisfies (ARE) $_\rho$  in [11].

$x \in \Omega$ ,  $\Delta(\cdot, t) : \Omega \rightarrow \mathbb{C}^{l \times q}$  is continuous for each  $t \in \mathbb{R}_+$ , and for each compact product set  $K \times I \subset \Omega \times \mathbb{R}_+$  there exists an integrable  $k(\cdot) : I \rightarrow \mathbb{R}_+$  such that

$$\|\Delta(x, t)Cx - \Delta(\hat{x}, t)C\hat{x}\|_2 \leq k(t)\|x - \hat{x}\|_2, \quad (x, t), (\hat{x}, t) \in K \times I.$$

The norm on  $\Delta_{nt}(\Omega)$  is taken to be

$$(18) \quad \|\Delta(\cdot, \cdot)\| = \sup_{x \in \Omega, t \geq 0} \|\Delta(x, t)\|_{2,2}, \quad \Delta(\cdot, \cdot) \in \Delta_{nt}(\Omega).$$

By Carathéodory's theorem for every  $(t_0, x^0) \in \mathbb{R}_+ \times \Omega$ , there exists a unique solution  $x(t) = x(t; t_0, x^0)$  of (17) with  $x(t_0) = x^0$  on some maximal semiopen interval  $[t_0, t_+(t_0, x^0))$ , where  $t_+(t_0, x^0) > t_0$ ; see [12, Theorem 2.1.14]. In the following theorem we will see that  $t_+(t_0, x^0) = \infty$  if  $\|\Delta(\cdot, \cdot)\| \leq r_{\mathbb{C}}(A, B, C)$  and  $x^0$  is sufficiently close to the equilibrium state  $\bar{x} = 0$ . Recall that a quadratic function  $V(x) = \langle x, Px \rangle$  is said to be a *Liapunov function* (respectively, *strict Liapunov function*) for the nonlinear time-varying system  $\dot{x} = Ax + B\Delta(x, t)Cx$  at the origin if  $P \succ 0$  and  $\dot{V}(x, t) \leq 0$  (respectively,  $\sup_{t \geq t_0} \dot{V}(x, t) < 0$  for  $x \neq 0$ ) on some neighborhood of the origin. The existence of a quadratic Liapunov function ensures that the origin is uniformly stable, whereas the existence of a strict quadratic Liapunov function implies uniform asymptotic stability; see [12, Theorem 3.2.17]. Unfortunately, these criteria are not applicable in the present situation where  $P$  is obtained as a solution of the algebraic Riccati equation (15), since the corresponding  $V(\cdot)$  is, in general, neither a strict Liapunov function for the system (17) nor need it be positive definite. Nevertheless, we will see in the following proof that  $V(\cdot)$  can be used to establish asymptotic stability. For simplicity, we call the system (17) *uniformly (asymptotically) stable* if  $\bar{x} = 0$  is a uniformly (asymptotically) stable equilibrium position of the system (17).

**THEOREM 2.6.** *Suppose  $(A, B, C) \in \mathbb{L}_{n,l,q}$  is a given system  $\sigma(A) \subset \mathbb{C}_-$ , and  $\Omega$  is an open neighborhood of 0 in  $\mathbb{C}^n$ . Then we have the following.*

- (i) *The nonlinear system (17) is asymptotically stable for all  $\Delta \in \Delta_{nt}(\Omega)$  satisfying  $\|\Delta(\cdot, \cdot)\| < r_{\mathbb{C}}(A, B, C)$ . Moreover, when this condition is satisfied we have  $x(t; t_0, x^0) \rightarrow 0$  as  $t \rightarrow \infty$  for all  $(t_0, x^0) \in \mathbb{R}_+ \times \Omega$  for which  $t_+(t_0, x^0) = \infty$ .*
- (ii) *Suppose  $\rho \leq r_{\mathbb{C}}(A, B, C)$ , and let  $P$  be a Hermitian solution of the algebraic Riccati equation (15). Then  $P \succeq 0$ . If  $\delta > 0$  satisfies  $D_\delta = \{x; x \in \mathbb{C}^n \text{ and } \langle x, Px \rangle < \delta\} \subset \Omega$ , then  $D_\delta$  is a joint domain of attraction of the equilibrium point  $\bar{x} = 0$  for all the systems (17) with  $\Delta \in \Delta_{nt}(\Omega)$ ,  $\|\Delta(\cdot, \cdot)\| < \rho$ .*
- (iii) *Suppose  $\rho \leq r_{\mathbb{C}}(A, B, C)$  and  $P$  is a Hermitian solution of (15). If  $r < \rho$ , then there exists a constant  $k > 0$  such that the derivative of the quadratic function  $V_\rho(x) = \langle x, Px \rangle$  along trajectories of (17) satisfies  $\dot{V}_\rho(x) \leq -k\|Cx\|^2$ ,  $x \in \Omega$ , for all  $\Delta \in \Delta_{nt}(\Omega)$ ,  $\|\Delta(\cdot, \cdot)\| \leq r$ . If  $(A, C)$  is observable,  $V_\rho(x)$  is a joint Liapunov function at  $\bar{x} = 0$  for all perturbed systems (87) with  $\Delta \in \Delta_{nt}(\Omega)$  satisfying  $\|\Delta(\cdot, \cdot)\| \leq \rho$ . In particular, (17) is uniformly stable if  $\Delta \in \Delta_{nt}(\Omega)$  and  $\|\Delta(\cdot, \cdot)\| \leq r_{\mathbb{C}}(A, B, C)$ .*

*Proof.* We first prove the third statement. Let  $\rho \leq r_{\mathbb{C}}(A, B, C)$ . Since  $A$  is Hurwitz stable, every Hermitian solution  $P$  of (15) is positive semidefinite. Multiplying (15) from the right by  $x \in \ker P$  and from the left by  $x^*$ , we see that  $\ker P \subset \ker C$ . Let  $x(t) = x(t; t_0, x^0)$ ,  $(t_0, x^0) \in \mathbb{R}_+ \times \Omega$ , be an arbitrary solution of (17). The derivative

of the time-invariant quadratic function  $V_\rho(x) = \langle x, Px \rangle$  along this solution is

$$\begin{aligned}
 \dot{V}_\rho(x(t)) &= [\langle (Ax(t) + B\Delta(x(t), t)Cx(t)), Px(t) \rangle \\
 &\quad + \langle x(t), P(Ax(t) + B\Delta(x(t), t)Cx(t)) \rangle] \\
 (19) \quad &= \langle (PA + A^*P)x(t), x(t) \rangle + 2\operatorname{Re} \langle B\Delta(x(t), t)Cx(t), Px(t) \rangle \\
 &= -\rho^2 \|y(t)\|_2^2 - \|B^*Px(t)\|_2^2 + 2\operatorname{Re} \langle \Delta(x(t), t)y(t), B^*Px(t) \rangle \\
 &= -\|\Delta(x(t), t)y(t) - B^*Px(t)\|_2^2 - [\rho^2 \|y(t)\|_2^2 - \|\Delta(x(t), t)y(t)\|_2^2] \\
 &\leq -[\rho^2 \|y(t)\|_2^2 - \|\Delta(x(t), t)y(t)\|_2^2], \quad t \in [t_0, t_+(t_0, x^0)],
 \end{aligned}$$

where  $y(t) = Cx(t)$ . Setting  $k = \rho^2 - r^2$  we obtain  $\dot{V}_\rho(x) \leq -k\|Cx\|^2$ ,  $x \in \Omega$ , for all  $\Delta(\cdot, \cdot) \in \Delta_{nt}(\Omega)$ ,  $\|\Delta(\cdot, \cdot)\| \leq r$ . If  $(A, C)$  is observable, then  $P \succ 0$  and so  $V_\rho$  is a time-invariant Liapunov function on  $\Omega$  for all perturbed systems (17) with  $\Delta \in \Delta_{nt}(\Omega)$  satisfying  $\|\Delta(\cdot, \cdot)\| \leq \rho$ . In particular, setting  $\rho = r_{\mathbb{C}}(A, B, C)$ , the system (17) is uniformly stable for all  $\Delta \in \Delta_{nt}(\Omega)$ ,  $\|\Delta(\cdot, \cdot)\| \leq r_{\mathbb{C}}(A, B, C)$  (see [12, Theorem 3.2.17]). This proves (iii).

We now abandon the observability assumption and prove (i). Suppose that  $\Delta \in \Delta_{nt}(\Omega)$  and  $\|\Delta(\cdot, \cdot)\| < \rho$ . Then

$$(20) \quad \dot{V}_\rho(x(t)) \leq 0, \quad t \in [t_0, t_+(t_0, x^0)], \quad \text{and} \quad \dot{V}_\rho(x(t)) = 0 \Rightarrow y(t) = Cx(t) = 0.$$

Consider the decomposition

$$x(t) = x_1(t) + x_2(t), \quad x_1(t) \in (\ker P)^\perp, \quad x_2(t) \in \ker P, \quad t \in [t_0, t_+(t_0, x^0)].$$

Since there exists  $\alpha > 0$  such that  $\langle x, Px \rangle = \langle x_1, Px_1 \rangle \geq \alpha \|x_1\|_2^2$  for all  $x \in \mathbb{C}^n$ , we have

$$(21) \quad \alpha \|x_1(t)\|_2^2 \leq \langle x_1(t), Px_1(t) \rangle \leq \langle x_1(t_0), Px_1(t_0) \rangle \leq \|P\|_{2,2} \|x_1(t_0)\|_2^2$$

for all  $(t_0, x^0) \in \mathbb{R}_+ \times \Omega$  and  $t \in [t_0, t_+(t_0, x^0)]$ . But  $\ker P \subset \ker C$ , so it follows that there exists a constant  $c > 0$ , independent of  $(t_0, x^0) \in \mathbb{R}_+ \times \Omega$ , such that

$$(22) \quad \|y(t)\|_2 = \|Cx(t)\|_2 = \|Cx_1(t)\|_2 \leq c \|x_1(t_0)\|_2, \quad t \in [t_0, t_+(t_0, x^0)].$$

Since  $\|e^{At}\| \leq Me^{-\omega t}$  with suitable  $M > 0$ ,  $\omega > 0$ , we have for all  $(t_0, x^0) \in \mathbb{R}_+ \times \Omega$

$$\begin{aligned}
 (23) \quad \|x(t)\|_2 &= \|e^{A(t-t_0)}x^0 + \int_{t_0}^t e^{A(t-s)}B\Delta(x(s), s)y(s)ds\|_2 \\
 &\leq Me^{-\omega(t-t_0)}\|x^0\|_2 + (M/\omega)\|B\|_{2,2}\|\Delta(\cdot, \cdot)\| \sup_{t \in [t_0, t_+(t_0, x^0)]} \|y(t)\|_2 \\
 &\leq Me^{-\omega(t-t_0)}\|x^0\|_2 + (M/\omega)\|B\|_{2,2}\|\Delta(\cdot, \cdot)\| c \|x_1(t_0)\|_2, \quad t \in [t_0, t_+(t_0, x^0)].
 \end{aligned}$$

Let  $\varepsilon > 0$  and  $B(0, \varepsilon) = \{x \in \mathbb{C}^n; \|x\|_2 < \varepsilon\}$  be such that the closed ball  $\overline{B(0, \varepsilon)}$  is contained in  $\Omega$ . Then by (23) there exists  $\delta' > 0$  so that

$$x^0 \in B(0, \delta'), \quad t_0 \in \mathbb{R}_+, \quad t \in [t_0, t_+(t_0, x^0)] \implies x(t) = x(t; t_0, x^0) \in B(0, \varepsilon).$$

Hence  $t_+(t_0, x^0) = \infty$  for  $(t_0, x^0) \in \mathbb{R}_+ \times B(0, \delta')$  and the system (17) is uniformly stable.

Now assume that  $(t_0, x^0) \in \mathbb{R}_+ \times \Omega$ ,  $t_+(t_0, x^0) = \infty$  and  $y(t) = Cx(t, t_0, x^0)$  does not tend to zero as  $t \rightarrow \infty$ . Then there exist  $\varepsilon_1 > 0$  and a sequence  $t_k \rightarrow \infty$  such that



$\|y(t_k)\|_2 > 2\varepsilon_1$  for  $k \in \mathbb{N}$ . Now  $\|x(t)\|_2$ ,  $t \geq t_0$  is bounded by (23), and so via (17) we see that  $\|\dot{x}(t)\|_2$ ,  $t \geq t_0$ , is also bounded. Hence there exists  $\eta > 0$  such that

$$\|y(t)\|_2 > \varepsilon_1 \quad \text{for } t \in [t_k, t_k + \eta], \quad k \in \mathbb{N}.$$

Let  $\gamma := \rho^2 - \|\Delta(\cdot, \cdot)\|^2 > 0$ . Then (19) implies for all  $k \in \mathbb{N}$

$$\begin{aligned} & V_\rho(x(t_k + \eta)) - V_\rho(x(t_k)) \\ &= \int_{t_k}^{t_k + \eta} \dot{V}_\rho(x(t)) dt \leq -\int_{t_k}^{t_k + \eta} (\rho^2 - \|\Delta(x(t), t)\|^2) \|y(t)\|_2^2 dt \leq -\gamma \eta \varepsilon_1^2. \end{aligned}$$

Since  $V_\rho(x(t))$  is not increasing in  $t$ , this contradicts the fact that  $V_\rho(x(t)) \geq 0$ ,  $t \geq t_0$ . Hence  $\lim_{t \rightarrow \infty} \|y(t)\|_2 = 0$ . Replacing  $t_0$  by a sufficiently large  $t'_0$ ,  $x^0$  by  $x(t'_0)$ , and  $t_+(t_0, x^0)$  by  $\infty$ , the first inequality in (23) becomes

$$\|x(t)\|_2 \leq M e^{-\omega(t-t'_0)} \|x(t'_0)\|_2 + (M/\omega) \|B\|_{2,2} \|\Delta(\cdot, \cdot)\| \sup_{t \in [t'_0, \infty)} \|y(t)\|_2, \quad t \in [t'_0, \infty).$$

We conclude that  $x(t) \rightarrow 0$  as  $t \rightarrow \infty$ . Since  $t_+(t_0, x^0) = \infty$  for all  $x^0 \in B(0, \delta')$ , setting  $\rho = r_{\mathbb{C}}(A, B, C)$  completes the proof of (i).

Now suppose that  $D_\delta = \{x \in \mathbb{C}^n; \langle x, Px \rangle < \delta\} \subset \Omega$  for some  $\delta > 0$  and  $\Delta \in \Delta_{nt}(\Omega)$ ,  $\|\Delta(\cdot, \cdot)\| < \rho$ . Let  $x^0 \in D_\delta$  and choose  $\delta_1 > 0$  such that  $\langle x^0, Px^0 \rangle < \delta_1 < \delta$ ; then  $\overline{D_{\delta_1}} \subset \Omega$  and  $\overline{D_{\delta_1}}$  is invariant for (17). Hence  $t_+(t_0, x^0) = \infty$ , and so, by (i),  $x(t) \rightarrow 0$  as  $t \rightarrow \infty$ . This proves (ii).  $\square$

*Remark 2.7.* (i) If  $P \succ 0$ , then there will always exist a  $\delta$  such that  $D_\delta \subset \Omega$ . However, this need not be the case if  $\ker P \neq \{0\}$  (e.g., for bounded  $\Omega$ ).

(ii) Suppose that  $\Delta \in \Delta_{nt}(\Omega)$  and there exist  $k > 0$  and a positive semidefinite  $P \in \mathcal{H}_n(\mathbb{C})$  with  $\ker P \subset \ker C$  such that the quadratic function  $V(x) = \langle x, Px \rangle$  satisfies  $\dot{V}(x) \leq -k\|Cx\|^2$  along the solutions of (17). Then by similar arguments as in the previous proof one can show that  $\bar{x} = 0$  is an asymptotically stable equilibrium point of (17).

**3. Composite systems.** Let us introduce some additional notation. In the following,  $\mathbf{q}, \mathbf{l}$  are finite  $N$ -tuples  $\mathbf{l} = (l_1, \dots, l_N)$ ,  $\mathbf{q} = (q_1, \dots, q_N)$ ,  $l_j, q_j, N \in \mathbb{N}$ , and we set  $q = |\mathbf{q}| = \sum_{i=1}^m q_i$ ,  $l = |\mathbf{l}| = \sum_{j=1}^m l_j$ . We denote by  $\mathbb{K}^{l \times \mathbf{q}}$  the set of  $N \times N$  block matrices

$$(24) \quad [\Delta_{ij}] = [\Delta_{ij}]_{i,j \in \underline{N}} = \begin{bmatrix} \Delta_{11} & \cdots & \Delta_{1N} \\ \vdots & & \vdots \\ \Delta_{N1} & \cdots & \Delta_{NN} \end{bmatrix}, \quad \Delta_{ij} \in \mathbb{K}^{l_i \times q_j} \text{ for } (i, j) \in \underline{N} \times \underline{N}.$$

For any positive integer  $k$  we denote by  $\mathbb{K}^{k \times \mathbf{q}}$  the set of all block rows  $[X_1, \dots, X_N]$ ,  $X_j \in \mathbb{K}^{k \times q_j}$ ,  $j \in \underline{N}$ , and by  $\mathbb{K}^{1 \times k}$  the set of block columns  $[Y_1^\top, \dots, Y_N^\top]^\top$ ,  $Y_i \in \mathbb{K}^{l_i \times k}$ ,  $i \in \underline{N}$ . The partitioned vectors in  $\mathbb{K}^{\mathbf{q}} := \mathbb{K}^{\mathbf{q} \times 1}$  are denoted by  $x = (x_i)$ , where  $x_i \in \mathbb{K}^{q_i}$  for  $i \in \underline{N}$ . The block-diagonal matrix with blocks  $\Delta_j \in \mathbb{K}^{l_j \times q_j}$ ,  $j \in \underline{N}$ , is denoted by

$$\Delta := \oplus_{j=1}^N \Delta_j = \text{diag}(\Delta_1, \dots, \Delta_N) = \begin{bmatrix} \Delta_1 & & & 0 \\ & \Delta_2 & & \\ & & \ddots & \\ 0 & & & \Delta_N \end{bmatrix} \in \mathbb{K}^{1 \times \mathbf{q}}.$$

Suppose that  $E \in \mathbb{R}^{N \times N}$  is a given nonnegative matrix with entries  $e_{ij} \geq 0$ , and let  $\mathcal{I} = \{(i, j) \in \underline{N} \times \underline{N}; e_{ij} > 0\}$ . For each  $i \in \underline{N}$  let  $\mathcal{I}_i = \{j \in \underline{N}; e_{ij} > 0\}$  denote the

set of positions of the positive entries of  $E$  in row  $i$ . We say that  $\Delta = (\Delta_{ij}) \in \mathbb{C}^{1 \times \mathbf{q}}$  is of structure  $E$  if  $e_{ij} = 0$  implies  $\Delta_{ij} = 0$ . Let  $\Delta_E \subset \mathbb{C}^{1 \times \mathbf{q}}$  be the vector space of all the block matrices  $\Delta \in \mathbb{C}^{1 \times \mathbf{q}}$  of structure  $E$ .

To describe the perturbations  $A \rightsquigarrow A_\Delta$  in a concise way we will make use of the *Hadamard product* of matrices. Given  $X = (x_{ij}), Y = (y_{ij}) \in \mathbb{C}^{h \times k}$ , where  $h, k$  are positive integers, the *Hadamard product* of  $X$  and  $Y$ , denoted by  $X \circ Y$ , is defined by  $X \circ Y = (x_{ij}y_{ij}) \in \mathbb{C}^{h \times k}$ . For  $k \in \mathbb{Z}$  the  $k$ th Hadamard power of  $X$  is defined by  $X^{\circ k} = (x_{ij}^{\circ k})$ , where  $x_{ij}^{\circ k} = x_{ij}^k$  if  $x_{ij} \neq 0$  and  $x_{ij}^{\circ k} = 0$  if  $x_{ij} = 0$ . Given a matrix  $X = (x_{ij}) \in \mathbb{C}^{N \times N}$  and a block matrix  $Y = [Y_{jk}]_{j,k \in \underline{N}} \in \mathbb{C}^{1 \times \mathbf{q}}$ , the *Hadamard block product* of  $X$  and  $Y$  is by definition the block matrix  $X \circ Y := [x_{ij}Y_{ij}]_{i,j \in \underline{N}}$ . Note that for every  $\Delta = (\Delta_{ij}) \in \mathbb{C}^{1 \times \mathbf{q}}$  we have  $E \circ \Delta = [e_{ij}\Delta_{ij}]_{i,j \in \underline{N}} \in \Delta_E$ .

Given  $(A_i, B_i, C_i) \in \mathbb{L}_{n_i, l_i, q_i}$ ,  $i \in \underline{N}$ , the object of this paper is to study the variation of the spectrum of the block-diagonal matrix  $A = \oplus_{i=1}^N A_i \in \mathbb{C}^{n \times n}$  under perturbations of the form

$$(25) \quad A \rightsquigarrow A_\Delta := A + B(E \circ \Delta)C, \quad \Delta \in \Delta_E,$$

where  $\mathbf{n} = (n_1, \dots, n_N)$  is given,  $n := |\mathbf{n}|$ , and  $B, C$  are the block-diagonal matrices  $B = \oplus_{i=1}^N B_i \in \mathbb{C}^{n \times 1}$ ,  $C = \oplus_{i=1}^N C_i \in \mathbb{C}^{q \times n}$ . The nonnegative matrix  $E$  has a double role. On the one hand, it defines the *structure of the admissible perturbations*, i.e., the perturbation set  $\Delta_E$ . On the other hand, the positive entries  $e_{ij}$  of  $E$  provide *weights* for the blocks  $\Delta_{ij}$ . Note that since these weights cannot, in general, be absorbed by the matrices  $B_i$  and/or  $C_j$ , they provide an additional scaling flexibility. The scaled blocks  $e_{ij}\Delta_{ij}$  represent uncertain couplings between the subsystems described by the triplets  $(A_i, B_i, C_i) \in \mathbb{L}_{n_i, l_i, q_i}$ ,  $i \in \underline{N}$ . In fact, consider the system

$$(26) \quad \Sigma : \quad \dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t),$$

which is the direct sum of the  $N$  subsystems

$$(27) \quad \Sigma_i : \quad \dot{x}_i(t) = A_i x_i(t) + B_i u_i(t), \quad y_i(t) = C_i x_i(t), \quad i \in \underline{N}.$$

The transfer matrix of  $\Sigma$  is the direct sum of the transfer matrices of the subsystems

$$(28) \quad G(s) = C(sI - A)^{-1}B = \oplus_{i=1}^m G_i(s), \quad G_i(s) := C_i(sI_{n_i} - A_i)^{-1}B_i, \quad i \in \underline{N}.$$

Introducing the couplings

$$(29) \quad u_i(t) = \sum_{j \in \mathcal{I}_i} e_{ij} \Delta_{ij} y_j(t), \quad i \in \underline{N},$$

one obtains the coupled subsystem equations

$$(30) \quad \dot{x}_i(t) = A_i x_i(t) + B_i \sum_{j \in \mathcal{I}_i} e_{ij} \Delta_{ij} C_j x_j(t), \quad i \in \underline{N},$$

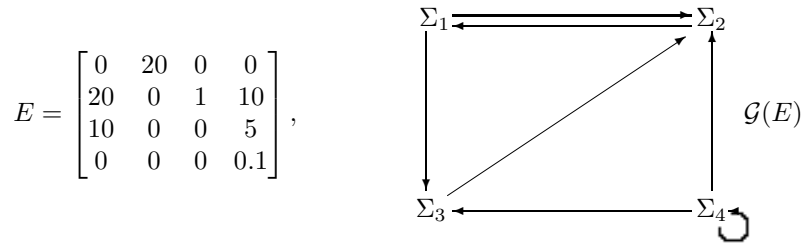
which together describe the composite system

$$(31) \quad \Sigma_\Delta : \quad \begin{bmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_N \end{bmatrix} = (A + B(E \circ \Delta)C) \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = A_\Delta \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}.$$

Thus the perturbed system  $\Sigma_\Delta$  with system matrix  $A_\Delta$  can be viewed as the composite system obtained by interconnecting the subsystems  $\Sigma_i$  via the uncertain couplings (29) determined by the unknown perturbation blocks  $\Delta_{ij}$ . The unperturbed (“nominal”) system  $\Sigma_0 : \dot{x} = Ax$  obtained by setting  $\Delta = 0$  is simply the direct sum of the subsystems  $\dot{x}_i = A_i x_i$ .

The interconnection structure of the composite system, determined by the matrix  $E$ , is best illustrated by drawing the associated graph  $\mathcal{G}(E)$  [13]. The node set of this directed graph is  $\underline{N}$  or alternatively, in the present context of interconnected systems, the set of subsystems  $\{\Sigma_i; i \in \underline{N}\}$ . The set of directed arcs is given by  $\mathcal{I} = \{(i, j); e_{ij} > 0\}$ , where the pair  $(i, j)$  denotes the arc from the node  $\Sigma_j$  to the node  $\Sigma_i$ .<sup>3</sup>

*Example 3.1.* Consider the following structure matrix  $E$  and the associated graph  $\mathcal{G}(E)$ .



The set of directed arcs of  $\mathcal{G}(E)$  is  $\mathcal{I} = \{(1, 2), (2, 1), (2, 3), (2, 4), (3, 1), (3, 4), (4, 4)\}$ . The diagonal entries of  $E$  portray a perturbation structure where the first three subsystems are unperturbed whereas the fourth one is subjected to perturbations of small weight. The off-diagonal entries model a situation where perturbation blocks of similar size will cause a strong interaction between the first two subsystems, a medium influence of the fourth on the second subsystem, a medium influence of the first on the third subsystem, a lesser influence of the fourth on the third subsystem, and a comparatively small influence of the third on the second subsystem.

**4. Stability radii and spectral value sets.** In this section we derive computable formulas for the spectral value sets and stability radii of the block-diagonal system  $\dot{x} = Ax$  with respect to structured perturbations of the form (25). We continue to use the set-up of the previous section and begin by extending the perturbation space from  $\Delta_E$  to  $\Delta = \mathbb{C}^{1 \times q} \supset \Delta_E$ . Let  $E^0$  be the  $N \times N$  matrix obtained from  $E$  by normalizing all of its nonzero entries to 1; i.e.,  $E^0 = (e_{ij}^0)$  with  $e_{ij}^0 = 1$  if  $e_{ij} > 0$  and  $e_{ij}^0 = 0$  if  $e_{ij} = 0$ . Clearly,  $E = E \circ E^0$  and we have for all  $\Delta \in \mathbb{C}^{1 \times q}$

$$(32) \quad \Delta_0 := E^0 \circ \Delta \in \Delta_E \quad \text{and} \quad E \circ \Delta = E \circ \Delta_0.$$

The block matrix  $\Delta_0 = [e_{ij}^0 \Delta_{ij}]$  is obtained from  $\Delta$  by replacing all of the blocks  $\Delta_{ij}$  in  $\Delta$  for which  $e_{ij} = 0$  by zero blocks of the same dimensions. Because of (32) the set of perturbed matrices  $A_\Delta$  is not extended if we extend the perturbation space from  $\Delta_E$  to  $\Delta = \mathbb{C}^{1 \times q}$ :

$$\{A_\Delta; \Delta \in \Delta_E\} = \{A_\Delta; \Delta \in \mathbb{C}^{1 \times q}\}.$$

<sup>3</sup>Note that this is the reverse of the standard notation in graph theory. We have used our notation in order to be in harmony with the system theoretic interpretation.

Since this leads to a substantial simplification in the notation we will henceforth allow  $\Delta$  to vary in  $\mathbb{C}^{1 \times \mathbf{q}}$  and consider perturbations of the form

$$(33) \quad A \rightsquigarrow A_\Delta := A + B(E \circ \Delta)C = A + B(E \circ \Delta_0)C, \quad \Delta \in \mathbf{\Delta} = \mathbb{C}^{1 \times \mathbf{q}}.$$

Let us introduce a norm on the extended perturbation space  $\mathbf{\Delta}$ . If  $\Delta_i = [\Delta_{i1}, \dots, \Delta_{iN}]$  is the  $i$ th block row of  $\Delta = [\Delta_{jk}] \in \mathbb{C}^{1 \times \mathbf{q}}$ ,  $i \in \underline{N}$ , then

$$(34) \quad \|\Delta_i\|_{2,2}^2 = \lambda_{\max}(\Delta_i \Delta_i^*) = \lambda_{\max} \left( \sum_{j \in \underline{N}} \Delta_{ij} \Delta_{ij}^* \right).$$

On the space of perturbation matrices  $\mathbf{\Delta} = \mathbb{C}^{1 \times \mathbf{q}}$  we introduce the mixed norm  $\|\cdot\|_{\mathbf{\Delta}}$  defined by

$$(35) \quad \|\Delta\|_{\mathbf{\Delta}} = \max_{i \in \underline{N}} \|\Delta_i\|_{2,2} = \max_{i \in \underline{N}} \left[ \lambda_{\max} \left( \sum_{j \in \underline{N}} \Delta_{ij} \Delta_{ij}^* \right) \right]^{1/2}, \quad \Delta = (\Delta_{ij}) \in \mathbf{\Delta}.$$

$\|\Delta\|_{\mathbf{\Delta}}$  is the operator norm of  $\Delta$  as a linear map from  $\mathbb{C}^{\mathbf{q}}$  provided with the 2-norm to  $\mathbb{C}^1$  provided with the  $(2|\infty)$ -Hölder norm  $\|\cdot\|_{2|\infty}$  defined by

$$\|u\|_{2|\infty} = \max_{i \in \underline{N}} \|u_i\|_2, \quad u = (u_i) \in \mathbb{C}^1, \quad u_i \in \mathbb{C}^{l_i}, \quad i \in \underline{N}.$$

The spectral value set (Definition 2.2) of the block-diagonal matrix  $A$  at uncertainty level  $\delta$  (with respect to the perturbations of the form (33)) is denoted by  $\sigma_{\mathbf{\Delta}}(A, B, C, E; \delta)$  and the corresponding stability radius (Definition 2.3) by  $r_{\mathbf{\Delta}}(A, B, C, E)$ . Since for every  $\Delta \in \mathbf{\Delta}$  we have  $\Delta = E^0 \circ \Delta \in \mathbf{\Delta}_E$  and  $\|\Delta_0\|_{\mathbf{\Delta}} \leq \|\Delta\|_{\mathbf{\Delta}}$ , we get from (32) and (33)

$$(36) \quad \begin{aligned} \sigma_{\mathbf{\Delta}}(A, B, C, E; \delta) &= \bigcup_{\Delta \in \mathbf{\Delta}, \|\Delta\|_{\mathbf{\Delta}} < \delta} \sigma(A_\Delta) = \bigcup_{\Delta \in \mathbf{\Delta}_E, \|\Delta\|_{\mathbf{\Delta}} < \delta} \sigma(A_\Delta) = \sigma_{\mathbf{\Delta}_E}(A, B, C, E; \delta), \\ r_{\mathbf{\Delta}}(A, B, C, E) &= \inf\{\|\Delta\|_{\mathbf{\Delta}}; \Delta \in \mathbf{\Delta}, \sigma(A_\Delta) \not\subset \mathbb{C}_-\} = r_{\mathbf{\Delta}_E}(A, B, C, E). \end{aligned}$$

Let  $\Pi_N$  be the group of permutations of the set  $\underline{N}$  and  $\pi \in \Pi_N$ . For every matrix  $X = (x_{ij}) \in \mathbb{C}^{N \times N}$  and every block matrix  $Y = [Y_{ij}]_{i,j \in \underline{N}} \in \mathbb{C}^{1 \times \mathbf{q}}$  we define the matrix (respectively, block matrix) obtained by simultaneous permutation  $\pi$  of its rows and columns (respectively, block rows and block columns) as follows:

$$(37) \quad \pi(X) = (x_{\pi(i)\pi(j)})_{i,j \in \underline{N}} \quad \text{and} \quad \pi(Y) = [Y_{\pi(i)\pi(j)}]_{i,j \in \underline{N}} \in \mathbb{C}^{\pi(1) \times \pi(\mathbf{q})},$$

where  $\pi(1) = (l_{\pi(1)}, \dots, l_{\pi(N)})$  and  $\pi(\mathbf{q}) = (q_{\pi(1)}, \dots, q_{\pi(N)})$ . Then

$$(38) \quad \pi(X) \circ \pi(Y) = \pi(X \circ Y), \quad X \in \mathbb{C}^{N \times N}, \quad Y \in \mathbb{C}^{1 \times \mathbf{q}}, \quad \pi \in \Pi_N.$$

In order to determine  $\sigma_{\mathbf{\Delta}}(A, B, C, E; \delta)$  and  $r_{\mathbf{\Delta}}(A, B, C, E)$  we will sometimes make the assumption that the structure matrix  $E$  is irreducible.  $E$  is said to be *irreducible* if it is not possible to find a perturbation  $\pi \in \Pi_N$  such that  $\pi(E) = \begin{bmatrix} E^{11} & 0 \\ E^{21} & E^{22} \end{bmatrix}$ , with  $E^{11} \in \mathbb{R}^{N_1 \times N_1}$ ,  $E^{22} \in \mathbb{R}^{(N-N_1) \times (N-N_1)}$  for some  $N_1 \in [1, N]$ . The irreducibility of  $E$  has a nice graph theoretical interpretation. Let  $\mathcal{G}(E)$  be the directed graph

corresponding to the matrix  $E$ ; see section 3. Then the matrix  $E$  is irreducible if and only if  $\mathcal{G}(E)$  is strongly connected; see [4, section 2.2]. If  $E$  is reducible, it is known (see [4, section 2.3]) that  $E$  can be reduced by a simultaneous row and column permutation  $\pi \in \Pi_N$  to a block-triangular form

$$(39) \quad \pi(E) = \begin{bmatrix} E_{11} & 0 & \cdots & 0 \\ E_{21} & E_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ E_{s1} & E_{s2} & \cdots & E_{ss} \end{bmatrix},$$

where each diagonal block  $E_{hh} \in \mathbb{R}^{N_h \times N_h}$ ,  $h \in \underline{s}$  is square and is either irreducible or a  $1 \times 1$  null matrix. Applying the permutation  $\pi$  to  $A + B(E \circ \Delta)C$  we obtain

$$(40) \quad \pi(A + B(E \circ \Delta)C) = \pi(A) + \pi(B)(\pi(E) \circ \pi(\Delta))\pi(C),$$

where

$$(41) \quad \begin{aligned} \pi(A) &= \oplus_{h=1}^s A_\pi^h, \quad \pi(B) = \oplus_{h=1}^s B_\pi^h, \quad \pi(C) = \oplus_{h=1}^s C_\pi^h, \\ \pi(\Delta) &= (\Delta_\pi^{hk})_{h,k \in \underline{s}} \in \mathbb{C}^{\pi(1) \times \pi(\mathbf{q})}. \end{aligned}$$

The superblocks  $A_\pi^h$ ,  $B_\pi^h$ ,  $C_\pi^h$ ,  $h \in \underline{s}$ , are block-diagonal matrices with  $N_h$  diagonal blocks on the diagonal and, for any  $\Delta \in \Delta_E$ ,  $\pi(\Delta)$  is a lower block-triangular matrix with the superblocks  $\Delta_\pi^{hk}$  consisting of  $N_h \times N_k$  blocks  $\Delta_{ij}$  of  $\Delta$ ,  $h, k \in \underline{s}$ . Clearly, the blocks on the diagonals of  $\pi(A)$ ,  $\pi(B)$ ,  $\pi(C)$  are a permutation of the diagonal blocks of  $A$ ,  $B$ ,  $C$ , respectively. Moreover, if  $\Delta \in \Delta_E$ , the superblocks  $\Delta_\pi^{hk}$  are of the structure  $E_{hk}$  in the sense that if an entry  $E_{hk}(i, j)$  of  $E_{hk}$  is zero, then the block  $\Delta_\pi^{hk}(i, j) = 0$ .

*Example 4.1.* Suppose we are given four subsystems  $(A_i, B_i, C_i) \in L_{n_i, l_i, q_i}$ ,  $i = 1, \dots, 4$ , and the interconnection matrix  $E \in \mathbb{R}^{4 \times 4}$  is as in Example 3.1. The graph  $\mathcal{G}(E)$  has two strongly connected components with node sets  $\{\Sigma_1, \Sigma_2, \Sigma_3\}$  and  $\{\Sigma_4\}$ . Choose  $\pi \in \Pi_4$  to be the permutation which maps 1, 2, 3, 4 to 4, 1, 2, 3, respectively. Then the permuted matrix  $\pi(E)$  is of the form (39) with  $s = 2$ :

$$\begin{aligned} \pi(E) &= \left[ \begin{array}{c|ccc} 0.1 & 0 & 0 & 0 \\ \hline 0 & 0 & 20 & 0 \\ 10 & 20 & 0 & 1 \\ 5 & 10 & 0 & 0 \end{array} \right] = \begin{bmatrix} E_{11} & 0 \\ E_{21} & E_{22} \end{bmatrix}, \\ E_{11} &= [0.1], \quad E_{21} = \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}, \quad E_{22} = \begin{bmatrix} 0 & 20 & 0 \\ 20 & 0 & 1 \\ 10 & 0 & 0 \end{bmatrix}. \end{aligned}$$

The block-diagonal matrices  $\pi(A)$ ,  $\pi(B)$ ,  $\pi(C)$  each consist of two superblocks. For instance,  $\pi(A) = A_\pi^1 \oplus A_\pi^2$ , where  $A_\pi^1 = A_4$  and  $A_\pi^2 = A_1 \oplus A_2 \oplus A_3$ . The perturbation matrices  $\Delta \in \Delta_E$ , respectively,  $\pi(\Delta) \in \Delta_E^\pi = \pi(\Delta_E)$ , are of the following form:

$$\begin{aligned} \Delta &= \begin{bmatrix} 0 & \Delta_{12} & 0 & 0 \\ \Delta_{21} & 0 & \Delta_{23} & \Delta_{24} \\ \Delta_{31} & 0 & 0 & \Delta_{34} \\ 0 & 0 & 0 & \Delta_{44} \end{bmatrix}, \\ \pi(\Delta) &= \left[ \begin{array}{c|ccc} \Delta_{44} & 0 & 0 & 0 \\ \hline 0 & 0 & \Delta_{12} & 0 \\ \Delta_{24} & \Delta_{21} & 0 & \Delta_{23} \\ \hline \Delta_{34} & \Delta_{31} & 0 & 0 \end{array} \right] = \left[ \begin{array}{c|ccc} \Delta_\pi^{11} & 0 & & \\ \hline \Delta_\pi^{21} & & \Delta_\pi^{22} & \end{array} \right]. \end{aligned}$$

The spectrum and the perturbation norm remain invariant under the simultaneous permutation of block rows and block columns:

$$\begin{aligned}\sigma(A + B(E \circ \Delta)C) &= \sigma(\pi(A) + \pi(B)(\pi(E) \circ \pi(\Delta))\pi(C)), \\ \|\Delta\|_{\Delta} &= \max_{i \in \underline{N}} \left[ \lambda_{\max} \left( \sum_{j \in \underline{N}} \Delta_{ij} \Delta_{ij}^* \right) \right]^{1/2} \\ &= \max_{i \in \underline{N}} \left[ \lambda_{\max} \left( \sum_{j \in \underline{N}} \Delta_{\pi(i)\pi(j)} \Delta_{\pi(i)\pi(j)}^* \right) \right]^{1/2} = \|\pi(\Delta)\|_{\pi(\Delta)}.\end{aligned}$$

Hence

$$(42) \quad \begin{aligned}\sigma_{\Delta}(A, B, C, E; \delta) &= \sigma_{\pi(\Delta)}(\pi(A), \pi(B), \pi(C), \pi(E); \delta), \\ r_{\Delta}(A, B, C, E) &= r_{\pi(\Delta)}(\pi(A), \pi(B), \pi(C), \pi(E)),\end{aligned}$$

and we may therefore assume, wherever convenient, that  $E$  has the same block-triangular structure as  $\pi(E)$  in (39). Because of the block-triangular structure of  $\pi(A) + \pi(B)(\pi(E) \circ \pi(\Delta))\pi(C)$  the proof of the following lemma is straightforward.

LEMMA 4.2. Suppose that  $E$  is reduced to lower block-triangular form as in (39) by a permutation  $\pi \in \Pi_N$ , and  $\pi(A), \pi(B), \pi(C)$  and  $\pi(\Delta)$ ,  $\Delta \in \Delta$  are decomposed as in (41). If, for any pair  $(h, k) \in \underline{s} \times \underline{s}$ , we provide the space of superblocks  $\Delta_{\pi}^{hk} = \{\Delta_{\pi}^{hk}; \Delta \in \Delta\}$  with the norm

$$(43) \quad \|\Delta_{\pi}^{hk}\|_{\Delta_{\pi}^{hk}} = \max_{i \in \underline{N}_h} \left[ \lambda_{\max} \left( \sum_{j \in \underline{N}_k} \Delta_{ij}^{hk} (\Delta_{ij}^{hk})^* \right) \right]^{1/2}, \quad \Delta_{\pi}^{hk} = (\Delta_{ij}^{hk})_{i \in \underline{N}_h, j \in \underline{N}_k},$$

then

$$(44) \quad \sigma_{\Delta}(A, B, C, E; \delta) = \bigcup_{h \in \underline{s}} \sigma_{\Delta_{\pi}^{hh}}(A_{\pi}^h, B_{\pi}^h, C_{\pi}^h, E_{hh}; \delta), \quad \delta > 0,$$

and

$$(45) \quad r_{\Delta}(A, B, C, E) = \min_{h \in \underline{s}} r_{\Delta_{\pi}^{hh}}(A_{\pi}^h, B_{\pi}^h, C_{\pi}^h, E_{hh}).$$

Remark 4.3. If  $E_{hh}$  is a  $1 \times 1$  null matrix, then  $A_{\pi}^h = [A_i]$  for some  $i \in \underline{N}$ . It follows that the eigenvalues of  $A_i$  are fixed eigenvalues of all perturbed matrices  $A + B(E \circ \Delta)C$ ,  $\Delta \in \Delta$ , and  $r_{\Delta_{\pi}^{hh}}(A_{\pi}^h, B_{\pi}^h, C_{\pi}^h, E_{hh}) = \infty$  if  $A$  is asymptotically stable. Hence Lemma 4.2 shows that the analysis of the spectral perturbation problem under consideration can be reduced to the case where  $E$  is irreducible.

We illustrate Lemma 4.2 by applying it to the data of Example 4.1.

Example 4.4. Using the notation of Example 4.1 we obtain from (44) and (45) that

$$\sigma_{\Delta}(A, B, C, E; \delta) = \sigma_{\Delta_{\pi}^{11}}(A_{\pi}^1, B_{\pi}^1, C_{\pi}^1, E_{11}; \delta) \cup \sigma_{\Delta_{\pi}^{22}}(A_{\pi}^2, B_{\pi}^2, C_{\pi}^2, E_{22}; \delta),$$

$$r_{\Delta}(A, B, C, E) = \min\{r_{\Delta_{\pi}^{11}}(A_{\pi}^1, B_{\pi}^1, C_{\pi}^1, E_{11}), r_{\Delta_{\pi}^{22}}(A_{\pi}^2, B_{\pi}^2, C_{\pi}^2, E_{22})\},$$

where

$$A_\pi^1 = A_4, \quad A_\pi^2 = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{bmatrix}, \quad E_{11} = [0.1], \quad E_{22} = \begin{bmatrix} 0 & 20 & 0 \\ 20 & 0 & 1 \\ 10 & 0 & 0 \end{bmatrix}$$

and the components  $B_\pi^i, C_\pi^i$ ,  $i = 1, 2$ , of the block-diagonal matrices  $B, C$  have a structure analogous to those of  $A$ . The perturbations of  $A_\pi^1$  are of the form

$$A_\pi^1 = A_4 \rightsquigarrow A_\pi^1 + 0.1B_\pi^1\Delta_\pi^{11}C_\pi^1 = A_4 + 0.1B_4\Delta_{44}C_4,$$

whereas the perturbations of  $A_\pi^2$  are of the form

$$A_\pi^2 = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{bmatrix} \rightsquigarrow A_\pi^2 + B_\pi^2(E_{22} \circ \Delta_\pi^{22})C_\pi^2 = \begin{bmatrix} A_1 & 20\Delta_{12} & 0 \\ 20\Delta_{21} & A_2 & \Delta_{23} \\ 10\Delta_{11} & 0 & A_3 \end{bmatrix}.$$

Note that by the previous formulas the perturbation blocks  $\Delta_{24}, \Delta_{34}$  have no influence on  $\sigma_\Delta(A, B, C, E; \delta)$  and  $r_\Delta(A, B, C, E)$ . The spectrum of  $A_\Delta$  remains invariant if these blocks in the perturbation matrix  $\Delta$  are changed. In order to determine the stability radius  $r_\Delta(A, B, C, E)$  of  $A$  with respect to perturbations of the form (33) we will make use of the Perron–Frobenius theory of nonnegative matrices  $M \in \mathbb{R}_+^{N \times N}$ ; see [4], [13, Chapter 8].

LEMMA 4.5. Suppose  $M, M' \in \mathbb{R}_+^{N \times N}$ ,  $\alpha, \beta \in \mathbb{R}$ . Then we have the following.

- (i)  $\varrho(M) \in \sigma(M)$ , and there exists a nonnegative eigenvector  $z$  of  $M$  corresponding to the eigenvalue  $\varrho(M)$  (Perron vector). If  $M$  is irreducible, then the Perron vector is uniquely determined modulo multiplication by a positive scalar and all of its coordinates are positive.
- (ii) If there exists  $z \in \mathbb{R}_+^n$ ,  $z \neq 0$  such that  $Mz \geq \alpha z$ , then  $\varrho(M) \geq \alpha$ .
- (iii) If there exists  $z > 0$  such that  $Mz \leq \beta z$ , then  $\varrho(M) \leq \beta$ .
- (iv) If  $M \leq M'$ , then  $\varrho(M) \leq \varrho(M')$ . If  $M$  is irreducible and  $M \leq M'$ ,  $M \neq M'$ , then  $\varrho(M) < \varrho(M')$ .
- (v) If  $(M_k)$  is a sequence in  $\mathbb{R}_+^{N \times N}$  converging to  $M$ , then  $\varrho(M_k) \rightarrow \varrho(M)$  as  $k \rightarrow \infty$ .

*Proof.* (i)–(iv) follow from the Perron–Frobenius theory of nonnegative matrices; see [13, Theorem 8.3.1] and [13, Theorem 8.4.4] for (i), [13, Theorem 8.3.2] for (ii), [13, Corollary 8.1.29] for (iii), [13, Corollary 8.1.19] and [4, Corollary 1.3.29] for (iv).

(v) follows from the continuous dependence of the spectrum  $\sigma(M)$  on the matrix  $M$ .  $\square$

Remark 4.6. Let  $E$  be a nonnegative  $N \times N$  matrix,  $\mathcal{G}(E)$  the associated directed graph, and  $\mathcal{Z}_0(E)$  the set of cycles of this digraph. Then

$$\varrho(E) = 0 \quad \Leftrightarrow \quad E^N = 0 \quad \Leftrightarrow \quad \mathcal{Z}_0(E) = \emptyset.$$

The following theorem is the main result of this section and the key tool for determining  $\sigma_\Delta(A, B, C, E; \delta)$  and  $r_\Delta(A, B, C, E)$ .

THEOREM 4.7. Suppose  $E \in \mathbb{R}_+^{N \times N}$  and  $(A_i, B_i, C_i) \in \mathbb{L}_{n_i, l_i, q_i}$ ,  $i \in \underline{N}$ , are given. If  $A = \oplus_{i=1}^N A_i$  is nonsingular,  $B = \oplus_{i=1}^N B_i$ ,  $C = \oplus_{i=1}^N C_i$ , and  $\Delta = \mathbb{C}^{1 \times \mathbf{q}}$  is provided with the norm (35), then the “distance of  $A$  from singularity” with respect to perturbations of the form (33) is given by

(46)

$$d_\Delta(A, B, C, E) := \inf\{\|\Delta\|_\Delta; \Delta \in \Delta \text{ and } \det(A + B(E \circ \Delta)C) = 0\} = 1/\sqrt{\varrho(D^2 E \circ^2)},$$

where  $D = \text{diag}(\|C_1 A_1^{-1} B_1\|_{2,2}, \dots, \|C_N A_N^{-1} B_N\|_{2,2})$ . If additionally  $A, B, C$  are real, then also

$$(47) \quad \inf\{\|\Delta\|_{\Delta}; \Delta \in \mathbf{\Delta} \cap \mathbb{R}^{1 \times \mathbf{q}} \text{ and } \det(A + B(E \circ \Delta)C) = 0\} = 1/\sqrt{\varrho(D^2 E^{\circ 2})}.$$

*Proof.* Suppose that  $A$  is nonsingular and  $(A + B(E \circ \Delta)C)x = 0$  for some  $\Delta = (\Delta_{ij}) \in \mathbf{\Delta}$ ,  $x = (x_i) \in \mathbb{C}^{\mathbf{n}}$ ,  $x_i \in \mathbb{C}^{n_i}$ ,  $x \neq 0$ . Then

$$(48) \quad -A_i x_i = B_i \sum_{j \in \underline{N}} e_{ij} \Delta_{ij} C_j x_j, \quad i \in \underline{N}.$$

Setting  $y_i = C_i x_i$  and  $z_i = \|y_i\|_2^2$ ,  $i \in \underline{N}$ , we obtain by (48) and (35)

$$y_i = -C_i A_i^{-1} B_i \sum_{j \in \underline{N}} e_{ij} \Delta_{ij} y_j \quad \text{and} \quad z_i \leq \|C_i A_i^{-1} B_i\|_{2,2}^2 \|\Delta\|_{\Delta}^2 \sum_{j \in \underline{N}} e_{ij}^2 z_j, \quad i \in \underline{N}.$$

The  $y_i$  cannot all be zero since otherwise  $Ax = 0$  by (48) and  $x \neq 0$  would imply that  $A$  is singular. Hence  $z = (z_i) \in \mathbb{R}_+^{\mathbf{N}}$  satisfies  $z \leq \|\Delta\|_{\Delta}^2 D^2 E^{\circ 2} z$ ,  $z \neq 0$ . By Lemma 4.5 this implies  $\|\Delta\|_{\Delta}^{-2} \leq \varrho(D^2 E^{\circ 2})$  and so the inequality  $\geq$  in (46). We also see that there cannot exist a  $\Delta \in \mathbf{\Delta}$  such that  $\det(A + B(E \circ \Delta)C) = 0$  if  $\varrho(D^2 E^{\circ 2}) = 0$ . So equality holds in (46) if  $\varrho(D^2 E^{\circ 2}) = 0$  (making use of the conventions (1)).

To prove the converse inequality  $\leq$  in (46) we may therefore assume  $\varrho(D^2 E^{\circ 2}) > 0$ . We construct a perturbation matrix  $\Delta \in \mathbf{\Delta}_E$  satisfying  $\|\Delta\|_{\Delta}^2 = 1/\varrho(D^2 E^{\circ 2})$  and a vector  $x \in \mathbb{C}^{\mathbf{n}}$ ,  $x \neq 0$ , such that  $(A + B(E \circ \Delta)C)x = 0$ . By the Perron–Frobenius theory there exists  $z = (z_1, \dots, z_N)^{\top} \geq 0$ ,  $z \neq 0$ , such that  $D^2 E^{\circ 2} z = \varrho(D^2 E^{\circ 2}) z$ ; that is

$$(49) \quad \|C_i A_i^{-1} B_i\|_{2,2}^2 \sum_{j \in \underline{N}} e_{ij}^2 z_j = \varrho(D^2 E^{\circ 2}) z_i, \quad i \in \underline{N}.$$

If  $\|C_i A_i^{-1} B_i\|_{2,2} = 0$ , then necessarily  $z_i = 0$  by (49). Hence there exists, for every  $i \in \underline{N}$ , a vector  $u_i \in \mathbb{C}^{q_i}$  such that  $y_i := C_i A_i^{-1} B_i u_i$  satisfies

$$(50) \quad \|y_i\|_2^2 = \|C_i A_i^{-1} B_i\|_{2,2}^2 \|u_i\|_2^2 \quad \text{and} \quad \|y_i\|_2^2 = z_i.$$

Here we can choose  $u_i = 0$  if  $z_i = 0$ . Setting  $x_i = A_i^{-1} B_i u_i$  for  $i \in \underline{N}$  we obtain that  $x_i = 0$  whenever  $z_i = 0$ . Now define, for  $i \in \underline{N}$ ,

$$(51) \quad \Delta_{ij} = \begin{cases} -\frac{e_{ij} u_i y_j^*}{\sum_{k \in \underline{N}} e_{ik}^2 \|y_k\|_2^2} & \text{if } j \in \mathcal{I}_i \text{ and } \sum_{k \in \underline{N}} e_{ik}^2 \|y_k\|_2^2 \neq 0, \\ 0 & \text{if } j \notin \mathcal{I}_i \text{ or } \sum_{k \in \underline{N}} e_{ik}^2 \|y_k\|_2^2 = 0. \end{cases}$$

If  $\sum_{k \in \underline{N}} e_{ik}^2 \|y_k\|_2^2 = \sum_{k \in \underline{N}} e_{ik}^2 z_k = 0$  for some  $i \in \underline{N}$ , then  $\Delta_{ij} = 0$  for all  $j \in \underline{N}$  by (51), and  $z_i = 0$  by (49) so that  $x_i = 0$  and

$$A_i x_i + B_i \sum_{j \in \underline{N}} e_{ij} \Delta_{ij} C_j x_j = B_i \sum_{j \in \underline{N}} e_{ij} \Delta_{ij} y_j = 0.$$

On the other hand, if  $\sum_{k \in \underline{N}} e_{ik}^2 \|y_k\|_2^2 \neq 0$ , then by (51)

$$\begin{aligned} A_i x_i + B_i \sum_{j \in \underline{N}} e_{ij} \Delta_{ij} C_j x_j &= A_i x_i - B_i u_i \left( \sum_{j \in \underline{N}} e_{ij}^2 \|y_j\|_2^2 \right) / \left( \sum_{j \in \underline{N}} e_{ij}^2 \|y_j\|_2^2 \right) \\ &= A_i x_i - B_i u_i = 0. \end{aligned}$$



We conclude that  $(A + B(E \circ \Delta)C)x = 0$ , where  $x := (x_i) \in \mathbb{C}^n$ ,  $x \neq 0$ . It remains to show that  $\|\Delta\|_{\Delta} = 1/\sqrt{\varrho(D^2 E^{\circ 2})}$ . Now, for every  $i \in \underline{N}$  such that  $\sum_{j \in \underline{N}} e_{ij}^2 \|y_j\|_2^2 \neq 0$ , we have

$$\sum_{j=1}^N \Delta_{ij} \Delta_{ij}^* = \sum_{j \in \underline{N}} \frac{e_{ij}^2 u_i u_i^* \|y_j\|_2^2}{(\sum_{j \in \underline{N}} e_{ij}^2 \|y_j\|_2^2)^2} = \frac{u_i u_i^*}{\sum_{j \in \underline{N}} e_{ij}^2 \|y_j\|_2^2},$$

and so by (50) and (49)

$$\begin{aligned} \lambda_{\max} \left( \sum_{j \in \underline{N}} \Delta_{ij} \Delta_{ij}^* \right) &= \frac{\lambda_{\max}(u_i u_i^*)}{\sum_{j \in \underline{N}} e_{ij}^2 \|y_j\|_2^2} \\ &= \frac{\|u_i\|_2^2}{\sum_{j \in \underline{N}} e_{ij}^2 z_j} = \begin{cases} \frac{z_i / \|C_i A_i^{-1} B_i\|_{2,2}^2}{\sum_{j \in \underline{N}} e_{ij}^2 z_j} = 1/\varrho(D^2 E^{\circ 2}), & z_i \neq 0, \\ 0, & z_i = 0. \end{cases} \end{aligned}$$

Since  $z \neq 0$  there exists  $i \in \underline{N}$  such that  $z_i \neq 0$  and hence by (49)  $\sum_{j \in \underline{N}} e_{ij}^2 \|y_j\|_2^2 \neq 0$ . It follows that  $\|\Delta\|_{\Delta}^2 = \max_{i \in \underline{N}} \lambda_{\max}(\sum_{j \in \underline{N}} \Delta_{ij} \Delta_{ij}^*) = 1/\varrho(D^2 E^{\circ 2})$ , and this concludes the proof of (46).

If  $A, B, C$  are real, the  $u_i$ ,  $i \in \underline{N}$ , can be chosen to be real and then the perturbation matrix  $\Delta$  constructed above is real, and (47) follows from the previous proof.  $\square$

*Remark 4.8.* If  $\varrho(D^2 E^{\circ 2}) = 0$ , then  $A + B(E \circ \Delta)C$  is nonsingular for all  $\Delta \in \Delta_E$ . If  $\varrho(D^2 E^{\circ 2}) > 0$ , then the above proof shows how to construct a perturbation  $\Delta \in \Delta_E$  (respectively,  $\Delta \in \Delta_E \cap \mathbb{R}^{1 \times \mathbf{q}}$ ) of minimum norm  $\|\Delta\|_{\Delta}$  such that  $A + B(E \circ \Delta)C$  is singular. In this case the “inf” can be replaced by “min” in (46).

**COROLLARY 4.9.** Suppose  $E \in \mathbb{R}_+^{N \times N}$  and  $(A_i, B_i, C_i) \in \mathbb{L}_{n_i, l_i, q_i}$ ,  $i \in \underline{N}$ , are given. If  $A = \oplus_{i=1}^N A_i$ ,  $B = \oplus_{i=1}^N B_i$ ,  $C = \oplus_{i=1}^N C_i$ , and  $\Delta = \mathbb{C}^{1 \times \mathbf{q}}$  is provided with the norm (35), then we have the following.

- (i) For every  $\delta > 0$  the spectral value set of  $A$  of level  $\delta$  with respect to perturbations of the form (33) is

$$\begin{aligned} (52) \quad \sigma_{\Delta}(A, B, C, E; \delta) &= \bigcup_{\Delta \in \Delta, \|\Delta\|_{\Delta} < \delta} \sigma(A + B(E \circ \Delta)C) \\ &= \sigma(A) \cup \{s \in \rho(A); \varrho(D(s)^2 E^{\circ 2}) > \delta^{-2}\}, \end{aligned}$$

where  $D(s) = \text{diag}(\|G_1(s)\|_{2,2}, \dots, \|G_N(s)\|_{2,2})$ ,  $G_i(s) = C_i(sI_{n_i} - A_i)^{-1} B_i$ ,  $i \in \underline{N}$ .

- (ii) If  $A$  is Hurwitz stable, then its stability radius with respect to perturbations of the form (33) is

$$(53) \quad r_{\Delta}(A, B, C, E) = \left[ \max_{\omega \in \mathbb{R}} \varrho(D(i\omega)^2 E^{\circ 2}) \right]^{-1/2}.$$

*Proof.* (i) By definition we have  $\sigma(A) \subset \sigma_{\Delta}(A, B, C, E; \delta)$ . Now suppose  $s \in \rho(A)$ . Since  $s \in \sigma(A + B(E \circ \Delta)C)$  holds if and only if  $\det((sI_n - A) - B(E \circ \Delta)C) = 0$ , we obtain from (46) that  $s \in \sigma_{\Delta}(A, B, C, E; \delta)$  if and only if  $1/\sqrt{\varrho(D(s)^2 E^{\circ 2})} < \delta$ . This proves (52).

(ii) By the definition of  $r_{\Delta}(A, B, C, E)$ , the continuity of the spectrum, and by (52), we have for every  $\delta \geq 0$

$$\begin{aligned} \delta > r_{\Delta}(A, B, C, E) &\Leftrightarrow \sigma_{\Delta}(A, B, C, E; \delta) \cap i\mathbb{R} \neq \emptyset \\ &\Leftrightarrow \exists \omega \in \mathbb{R} : \varrho(D(i\omega)^2 E^{\circ 2}) > 1/\delta^2. \end{aligned}$$

Observing that the function  $\omega \mapsto \varrho(D(i\omega)^2 E^{\circ 2})$  admits a maximum on  $\mathbb{R}$  since  $\lim_{|\omega| \rightarrow \infty} \varrho(D(i\omega)^2 E^{\circ 2}) = 0$ , this proves (53).  $\square$

*Remark 4.10.* Suppose  $A, B, C$  are real and  $\omega \mapsto \varrho(D(i\omega)^2 E^{\circ 2})$  admits its maximum on  $\mathbb{R}_+$  at  $\omega = 0$ . Then  $D(0) = D$  and it follows from Theorem 4.7 that the stability radii of  $A$  with respect to complex and with respect to real perturbations of structure  $E$  are equal:

$$r_{\Delta}(A, B, C, E) = r_{\Delta_{\mathbb{R}}}(A, B, C, E) = [\varrho(D(0)^2 E^{\circ 2})]^{-1/2}.$$

As an illustration of Corollary 4.9 we consider an example where  $E$  is of “cyclic” structure.

*Example 4.11.* Suppose that  $(A_i, B_i, C_i) \in \mathbb{L}_{n_i, l_i, q_i}$ ,  $i \in \underline{N}$ ,  $G_i(s) = C_i(sI_{n_i} - A_i)^{-1}B_i$ , and

$$(54) \quad E = \begin{bmatrix} 0 & e_{12} & 0 & \cdots & 0 \\ 0 & 0 & e_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_{N-1, N} & 0 \\ e_{N, 1} & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad D(s) = \text{diag}(\|G_1(s)\|_{2,2}, \dots, \|G_N(s)\|_{2,2}).$$

Then

$$\begin{aligned} \varrho(D(s)^2 E^{\circ 2}) &= \varrho \left( \begin{bmatrix} 0 & e_{12}^2 \|G_1(s)\|_{2,2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_{N-1, N}^2 \|G_{N-1}(s)\|_{2,2}^2 \\ e_{N, 1}^2 \|G_N(s)\|_2^2 & 0 & \cdots & 0 \end{bmatrix} \right) \\ &= \sqrt[N]{e_{12}^2 \cdots e_{N-1, N}^2 e_{N, 1}^2 \|G_1(s)\|_{2,2}^2 \cdots \|G_N(s)\|_{2,2}^2}, \end{aligned}$$

and we obtain from (53) that

$$\begin{aligned} r_{\Delta}(A, B, C, E) &= \left[ \max_{\omega \in \mathbb{R}} \varrho(D(i\omega)^2 E^{\circ 2}) \right]^{-1/2} \\ &= \left[ \max_{\omega \in \mathbb{R}} \sqrt[N]{e_{12} \cdots e_{N-1, N} e_{N, 1} \|G_1(i\omega)\|_{2,2} \cdots \|G_N(i\omega)\|_{2,2}} \right]^{-1}. \end{aligned}$$

In particular, if we have, e.g.,

$$N = 2, \quad A = \begin{bmatrix} -1 + i & 0 \\ 0 & -2 + i \end{bmatrix}, \quad B = C = I_2, \quad E = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

then  $r_{\Delta} = r_{\Delta}(A, I_2, I_2, E)$  is given by

$$\begin{aligned} r_{\Delta} &= \left[ \max_{\omega \in \mathbb{R}} \sqrt{|1 + (\omega - 1)i|^{-1} \cdot |2 + (\omega - 1)i|^{-1}} \right]^{-1} \\ &= \left[ \min_{\omega \in \mathbb{R}} [1 + (\omega - 1)^2][4 + (\omega - 1)^2] \right]^{1/4} = \sqrt{2}. \end{aligned}$$

Although the perturbations (33) are structured, our approach does not make use of  $\mu$ -values. In order to explain how our results are related to  $\mu$ -analysis, we describe the perturbed system  $\dot{x} = A_{\Delta}x$  equivalently by an equation of the form

$$\dot{x} = Ax + B\tilde{\Delta}\tilde{C}x,$$

where  $\tilde{C}$  is a suitable matrix and  $\tilde{\Delta}$  is a block-diagonal counterpart of  $\Delta$  with spectral norm  $\|\tilde{\Delta}\|_{2,2} = \|\Delta\|_{\Delta}$ . Given  $\Delta = (\Delta_{ij}) \in \mathbf{\Delta}$ , we define  $\tilde{\Delta}$  by

$$(55) \quad \tilde{\Delta} = \text{diag}(\Delta^1, \dots, \Delta^N) \in \mathbb{C}^{1 \times Nq}, \quad \Delta^i = [\Delta_{i1}, \Delta_{i2}, \dots, \Delta_{iN}] \in \mathbb{C}^{l_i \times q}.$$

Note that the spectral norm of  $\tilde{\Delta}$  satisfies

$$(56) \quad \|\tilde{\Delta}\|_{2,2}^2 = \max_{i \in \underline{N}} \|\Delta^i\|_{2,2}^2 = \|\Delta\|_{\Delta}^2, \quad \Delta = (\Delta_{ij}) \in \mathbf{\Delta}.$$

Now define  $\tilde{C} \in \mathbb{C}^{Nq \times n}$  by

$$(57) \quad \tilde{C} = \begin{bmatrix} \tilde{C}^1 \\ \tilde{C}^2 \\ \vdots \\ \tilde{C}^N \end{bmatrix}, \quad \tilde{C}^i = \text{diag}(e_{i1}C_1, \dots, e_{iN}C_N) \in \mathbb{C}^{q \times n}.$$

Note that each block row of  $\tilde{C}$  contains at most one nonzero block and hence the nonzero blocks of any two different block columns of  $\tilde{C}$  occur at different positions. As a consequence  $\tilde{C}^* \tilde{C}$  is block-diagonal,

$$(58) \quad \tilde{C}^* \tilde{C} = \sum_{i \in \underline{N}} \tilde{C}^{i*} \tilde{C}^i = \text{diag} \left( \left( \sum_{i \in \underline{N}} e_{i1}^2 \right) C_1^* C_1, \dots, \left( \sum_{i \in \underline{N}} e_{iN}^2 \right) C_N^* C_N \right) \in \mathbb{C}^{n \times n}.$$

Moreover, we have

$$(59) \quad \Delta^i \tilde{C}^i = [e_{i1} \Delta_{i1} C_1, \dots, e_{iN} \Delta_{iN} C_N], \quad \tilde{\Delta} \tilde{C} = \begin{bmatrix} \Delta^1 \tilde{C}^1 \\ \vdots \\ \Delta^N \tilde{C}^N \end{bmatrix} = (E \circ \Delta) C,$$

and hence

$$(60) \quad A_{\Delta} = A + B(E \circ \Delta)C = A + B\tilde{\Delta}\tilde{C}, \quad \Delta \in \mathbf{\Delta}.$$

Defining  $\tilde{\mathbf{\Delta}} = \{\tilde{\Delta}; \Delta \in \mathbf{\Delta}\}$ , it follows from (60) and (56) that

$$\inf\{\|\Delta\|_{\Delta}; \Delta \in \mathbf{\Delta} \text{ and } \det(A_{\Delta}) = 0\} = \inf\{\|\tilde{\Delta}\|_{2,2}; \tilde{\Delta} \in \tilde{\mathbf{\Delta}} \text{ and } \det(A + B\tilde{\Delta}\tilde{C}) = 0\}.$$

On the other hand, since  $\det(I + UV) = \det(I + VU)$  for  $U \in \mathbb{C}^{h \times k}$ ,  $V \in \mathbb{C}^{k \times h}$ ,  $h, k \in \mathbb{N}$ , we have the equivalence

$$\det(A + B\tilde{\Delta}\tilde{C}) = 0 \Leftrightarrow \det(I + \tilde{\Delta}\tilde{C}A^{-1}B) = 0.$$

If we endow  $\tilde{\mathbf{\Delta}}$  with the spectral norm, we therefore obtain

$$(61) \quad d_{\mathbf{\Delta}}(A, B, C, E) = \inf\{\|\Delta\|_{\Delta}; \Delta \in \mathbf{\Delta} \text{ and } \det(A_{\Delta}) = 0\} = 1/\mu_{\tilde{\mathbf{\Delta}}}(\tilde{G}),$$

where  $\tilde{G} := \tilde{C}A^{-1}B$ .

*Remark 4.12.* Especially for sparse matrices  $E$  one can reduce the dimensions of  $\tilde{\Delta}$  and  $\tilde{C}$  a lot by eliminating in the rows of  $\Delta_i$  (see (55)) all of the blocks  $\Delta_{ij}$  for which  $e_{ij} = 0$  and removing in  $\tilde{C}^i$  (see (57)) all of the block rows for which  $e_{ij} = 0$ . However, this would considerably complicate the notation and not provide any computational advantage since the blown-up matrices  $\tilde{\Delta}$  and  $\tilde{C}$  will not be used in calculations.

The characterization (61) expresses the distance  $d_{\Delta}(A, B, C, E)$  in terms of the  $\mu$ -value of the blown-up matrix  $\tilde{G} := \tilde{C}A^{-1}B$  with respect to the set  $\tilde{\Delta}$  of blown-up block-diagonal perturbations  $\tilde{\Delta}$ . The question arises if it is possible to characterize  $d_{\Delta}(A, B, C, E)$  in terms of the  $\mu$ -value of the more natural “transfer function at  $s = 0$ ,”  $G = CA^{-1}B$ , and the given perturbation set  $\Delta_E$ . Since  $\Delta_E$  only reflects the structure of  $E$  but not the magnitude of its nonzero entries, such a characterization will only be possible, if the missing information is included in the set-up, e.g., via a suitable norm on  $\Delta_E$  depending on  $E$ . The following corollary shows how this can be done.

**COROLLARY 4.13.** *Suppose the conditions of Theorem 4.7 and let  $\Delta_E$  be provided with the norm*

$$(62) \quad \|\Delta\|_{\Delta_E} = \|E^{\circ-1} \circ \Delta\|_{\Delta}.$$

If  $G = CA^{-1}B$ , then  $\mu_{\Delta_E}(G) = \sqrt{\varrho(D^2E^{\circ 2})}$  and

$$(63) \quad \inf\{\|\Delta\|_{\Delta}; \Delta \in \Delta \text{ and } \det(A + B(E \circ \Delta)C) = 0\} = 1/\mu_{\Delta_E}(G).$$

*Proof.* For every  $\Delta \in \Delta_E$  define  $\Delta_E = E \circ \Delta \in \Delta_E$ . Then  $\|\Delta_E\|_{\Delta_E} = \|\Delta\|_{\Delta}$  for all  $\Delta \in \Delta_E$  and  $\Delta \mapsto \Delta_E$  is a norm preserving isomorphism from the normed space  $(\Delta_E, \|\cdot\|_{\Delta})$  onto  $(\Delta_E, \|\cdot\|_{\Delta_E})$ . Moreover, we obtain from (58) that for every  $\Delta \in \Delta$

$$(64) \quad \begin{aligned} \det(A + B(E \circ \Delta)C) = 0 &\Leftrightarrow \det(I + A^{-1}B(E \circ \Delta)C) = 0 \\ &\Leftrightarrow \det(I + \Delta_E G) = \det(I + (E \circ \Delta)CA^{-1}B) = 0. \end{aligned}$$

Now by definition  $\mu_{\Delta_E}(G)^{-1}$  equals

$$\inf\{\|\Delta_E\|_{\Delta_E}; \Delta_E \in \Delta_E, \det(I + \Delta_E G) = 0\} = \inf\{\|\Delta\|_{\Delta}; \Delta \in \Delta_E, \det(I + \Delta G) = 0\}.$$

Hence (63) follows from (64), and  $\mu_{\Delta_E}(G) = \sqrt{\varrho(D^2E^{\circ 2})}$  is a consequence of Theorem 4.7.  $\square$

*Remark 4.14.* There are many different possibilities for representing the perturbed matrices  $A_{\Delta} = A + B(E \circ \Delta)C$ ,  $\Delta \in \Delta$ , in the form  $A_{\Delta} = A + \tilde{B}\tilde{\Delta}\tilde{C}$ ,  $\tilde{\Delta} \in \tilde{\Delta}$ , where  $\tilde{\Delta}$  is a vector space of block-diagonal matrices. Equation (60) is just one of these representations, with  $\tilde{B} = B$ ,  $\tilde{C}$  defined by (57), and  $\tilde{\Delta}$  is the set of all block-diagonal perturbations  $\tilde{\Delta}$  of the form (55). Another representation of  $A_{\Delta}$  with block-diagonal perturbations  $\tilde{\Delta}$  is  $A_{\Delta} = A + \tilde{B}\tilde{\Delta}\tilde{C}$ , where

$$(65) \quad \tilde{B} = \oplus_{i \in \underline{N}} [B_{i1}, \dots, B_{iN}], \quad B_{ij} = B_i; \quad \tilde{\Delta} = \oplus_{i \in \underline{N}} \oplus_{j \in \underline{N}} \Delta_{ij}; \quad \tilde{C} = \begin{bmatrix} \oplus_{j \in \underline{N}} e_{1j} C_j \\ \vdots \\ \oplus_{j \in \underline{N}} e_{Nj} C_j \end{bmatrix}.$$

In order to relate this perturbation structure to the Riccati equation associated with the system  $(\tilde{A}, \tilde{B}, \tilde{C})$  via Theorem 2.5, we must provide  $\Delta$  with a norm  $\|\cdot\|_{\Delta}$  such

that  $\|\Delta\|_{\Delta} = \|\tilde{\Delta}\|_{2,2}$  for all  $\Delta \in \Delta_E$ . If we choose  $\tilde{B}, \tilde{C}, \tilde{\Delta}$  as in (65), then a suitable norm on  $\Delta$  would be  $\|\Delta\|_{\Delta} := \max_{i,j \in \underline{N}} \|\Delta_{ij}\|_{2,2}$ . If instead the weighted norm

$$\|\Delta\|_{\Delta_E} := \max_{(i,j) \in \mathcal{I}} e_{ij}^{-1} \|\Delta_{ij}\|_{2,2}$$

is introduced on  $\Delta_E$ , then

$$\|E \circ \Delta\|_{\Delta_E} = \|\Delta\|_{\Delta} = \|\tilde{\Delta}\|_{2,2}, \quad \Delta \in \Delta_E,$$

and one can prove that, with respect to the norms  $\|\cdot\|_{\Delta}$  and  $\|\cdot\|_{\Delta_E}$  just defined,

$$\inf\{\|\Delta\|_{\Delta}; \Delta \in \Delta \text{ and } \det(A + B(E \circ \Delta)C) = 0\} = 1/\mu_{\Delta_E}(G).$$

In [16, Theorem 4.2] the following formula for the  $\mu$ -value of any block-diagonal matrix  $G = \text{diag}(G_1, \dots, G_N) \in \mathbb{C}^{q \times l}$  (with respect to the norm  $\|\cdot\|_{\Delta_E}$  on  $\Delta_E$ ) was derived:

$$(66) \quad \mu_{\Delta_E}(G) = \varrho(ED), \quad \text{where} \quad D = \text{diag}(\|G_1\|_{2,2}, \dots, \|G_N\|_{2,2}).$$

An analysis of the corresponding Riccati equations and quadratic stability problem is more complicated and for lack of space cannot be dealt with in this article.

**5. Riccati equation.** We continue to use the set-up of section 3. There are various parametrized algebraic Riccati equations which can be used to construct joint quadratic Liapunov functions for perturbed systems of the form  $\dot{x} = A_{\Delta}x$  (31). If we neglected the structure matrix  $E$  and considered the full block case where  $A_{\Delta} = A + B\Delta C$ ,  $\Delta \in \mathbb{C}^{l \times q}$ , then the associated Riccati equation would be

$$(67) \quad PA + A^*P + \rho^2 C^*C + PBB^*P = 0.$$

If  $\sigma(A) \subset \mathbb{C}_-$ , then by Theorem 2.5 this Riccati equation has a Hermitian solution if and only if

$$(68) \quad \rho \leq \left[ \max_{\omega \in \mathbb{R}} \|G(i\omega)\|_{2,2} \right]^{-1} = \left[ \max_{\omega \in \mathbb{R}} \max_{i \in \underline{N}} \|G_i(i\omega)\|_{2,2} \right]^{-1} = \min_{i \in \underline{N}} g_i^{-1}, \quad g_i := \max_{\omega \in \mathbb{R}} \|G_i(i\omega)\|_{2,2},$$

where  $G(s)$  is the transfer function of the block-diagonal system  $(A, B, C)$ ; see (28).

If we take the perturbation structure into account, then different Riccati equations can be considered, depending on the representation of the perturbed system matrix  $A_{\Delta} = A + B(E \circ \Delta)C$  in the form  $A_{\Delta} = A + \tilde{B}\tilde{\Delta}\tilde{C}$ . In this section we will focus on the block-diagonal representation (60) where the matrices  $\tilde{B} = B, \tilde{C}$  and  $\tilde{\Delta}$  are given by (57) and (55). The parametrized algebraic Riccati equation associated with the system  $(A, B, \tilde{C})$  is given by

$$(69) \quad PA + A^*P + \rho^2 \tilde{C}^* \tilde{C} + PBB^*P = 0,$$

$$\tilde{C}^* \tilde{C} = \text{diag} \left( \left( \sum_{i \in \underline{N}} e_{i1}^2 \right) C_1^* C_1, \dots, \left( \sum_{i \in \underline{N}} e_{iN}^2 \right) C_N^* C_N \right).$$

We cannot expect to be able to characterize  $r_{\Delta}(A, B, C, E)$  by the solvability of this Riccati equation. In fact, by Theorem 2.5, (69) has a Hermitian solution if and only if

$$(70) \quad \rho \leq r_{\mathbb{C}}(A, B, \tilde{C}) = \inf\{\|\tilde{\Delta}\|_{2,2}; \tilde{\Delta} \in \mathbb{C}^{l \times Nq} \text{ and } \sigma(A + B\tilde{\Delta}\tilde{C}) \not\subset \mathbb{C}_-\} = \left[ \max_{\omega \in \mathbb{R}} \|\tilde{G}(i\omega)\|_{2,2} \right]^{-1},$$

where  $\tilde{G} = \tilde{C}(sI - A)^{-1}B$ . Thus (69) is related to a much larger class of perturbations than the block-diagonal perturbations of the form (55). We will therefore expect that, in general,  $r_{\mathbb{C}}(A, B, \tilde{C}) < r_{\Delta}(A, B, C, E)$ .

In order to get a better lower bound of  $r_{\Delta}(A, B, C, E)$  we use the scaling technique described in section 2. For any scaling vector  $\gamma = (\gamma_1, \dots, \gamma_N)$ , where  $\gamma_i > 0$ ,  $i \in \underline{N}$ , we set  $R_{\gamma} = \text{diag}(\gamma_1 I_{\mathbf{q}}, \dots, \gamma_N I_{\mathbf{q}})$  and  $L_{\gamma} = \text{diag}(\gamma_1 I_{l_1}, \dots, \gamma_N I_{l_N})$ . Then since  $\tilde{\Delta}$  is block-diagonal of the form (55) we have  $\tilde{\Delta} = L_{\gamma}^{-1} \tilde{\Delta} R_{\gamma}$ , and hence setting  $B_{\gamma} = B L_{\gamma}^{-1}$ ,  $\tilde{C}_{\gamma} = R_{\gamma} \tilde{C}$ , we obtain from (60) that

$$(71) \quad B(E \circ \Delta)C = B \tilde{\Delta} \tilde{C} = B_{\gamma} \tilde{\Delta} \tilde{C}_{\gamma}.$$

$\tilde{C}_{\gamma}$  has the same structure as  $\tilde{C}$  but with  $e_{ij}$  replaced by  $\gamma_i e_{ij}$ . More precisely,

$$(72) \quad \begin{aligned} B_{\gamma} &= \text{diag}(\gamma_1^{-1} B_1, \dots, \gamma_N^{-1} B_N), \\ \tilde{C}_{\gamma} &= \begin{bmatrix} \tilde{C}_{\gamma}^1 \\ \vdots \\ \tilde{C}_{\gamma}^N \end{bmatrix}, \quad \tilde{C}_{\gamma}^i = \text{diag}(\gamma_i e_{i1} C_1, \dots, \gamma_i e_{iN} C_N). \end{aligned}$$

The Riccati equation associated with the triple  $(A, B_{\gamma}, \tilde{C}_{\gamma})$  is

$$(73) \quad PA + A^*P + \rho^2 \tilde{C}_{\gamma}^* \tilde{C}_{\gamma} + PB_{\gamma} B_{\gamma}^* P = 0.$$

**THEOREM 5.1.** *Suppose that  $E \in \mathbb{R}_+^{N \times N}$ ,  $(A_i, B_i, C_i) \in \mathbb{L}_{n_i, l_i, q_i}$ ,  $i \in \underline{N}$ , and a scaling vector  $\gamma = (\gamma_1, \dots, \gamma_N) > 0$  are given; then  $A = \oplus_{i=1}^N A_i$ ,  $\sigma(A) \subset \mathbb{C}_-$ ,  $B_{\gamma}$  and  $\tilde{C}_{\gamma}$  are defined by (72), and  $\Delta$  is provided with the norm (35). Then we have the following:*

- (i) *The algebraic Riccati equation (73) has a Hermitian solution if and only if*

$$(74) \quad \rho^2 \leq \rho_{\gamma}^2 := \left[ \max_{j \in \underline{N}} g_j^2 \sum_{i=1}^N \frac{\gamma_i^2 e_{ij}^2}{\gamma_j^2} \right]^{-1} = \min_{j \in \underline{N}} \frac{\gamma_j^2}{g_j^2 \sum_{i=1}^N \gamma_i^2 e_{ij}^2},$$

where  $g_i = \max_{\omega \in \mathbb{R}} \|G_i(i\omega)\|_{2,2}$  and  $G_i = C_i(sI - A_i)^{-1}B_i$ ,  $i \in \underline{N}$ .

- (ii) *If  $\rho \leq \rho_{\gamma}$ , then the smallest solution  $P(\rho, \gamma)$  of (73) (satisfying  $\sigma(A + B_{\gamma} B_{\gamma}^* P) \subset \overline{\mathbb{C}_-}$ ) is of the form  $P(\rho, \gamma) = \oplus_{j \in \underline{N}} P_j(\rho, \gamma)$ , where  $P_j(\rho, \gamma)$ ,  $j \in \underline{N}$ , is the smallest solution of the algebraic Riccati equation*

$$(75) \quad P_j A_j + A_j^* P_j + \rho^2 \left( \sum_{i=1}^N \gamma_i^2 e_{ij}^2 \right) C_j^* C_j + \gamma_j^{-2} P_j B_j B_j^* P_j = 0.$$

- (iii)  $\sigma(A + B(E \circ \Delta)C) \subset \mathbb{C}_-$  for all  $\Delta \in \Delta$  with  $\|\Delta\|_{\Delta} < \rho_{\gamma}$ .  
 (iv) *Suppose  $E$  does not have a zero column,  $\rho \leq \rho_{\gamma}$ , and  $P$  is a Hermitian solution of (73). If  $r < \rho$ , then there exists a constant  $k > 0$  such that the derivative of the quadratic function  $V_{\rho}(x) = \langle x, Px \rangle$  along trajectories of  $\dot{x} = A_{\Delta} x$  (31) satisfies  $\dot{V}_{\rho}(x) \leq -k \|Cx\|^2$ ,  $x \in \mathbb{C}^n$ , for all  $\Delta \in \Delta$ ,  $\|\Delta\|_{\Delta} \leq r$ . If the pairs  $(A_j, C_j)$ ,  $j \in \underline{N}$ , are observable, then the pair  $(A, \tilde{C}_{\gamma})$  is observable and  $V_{\rho}(x) = \langle x, Px \rangle$  is a joint Liapunov function for all perturbed systems  $\dot{x} = (A + B(E \circ \Delta)C)x$  with  $\Delta \in \Delta$ ,  $\|\Delta\|_{\Delta} \leq \rho$ .*

*Proof.* (i) It follows from (58) and (72) that  $\tilde{C}_\gamma^* \tilde{C}_\gamma$  is the block-diagonal matrix

$$(76) \quad \tilde{C}_\gamma^* \tilde{C}_\gamma = \text{diag} \left( \left( \sum_{i=1}^N \gamma_i^2 e_{i1}^2 \right) C_1^* C_1, \dots, \left( \sum_{i=1}^N \gamma_i^2 e_{iN}^2 \right) C_N^* C_N \right).$$

So if  $G_\gamma(s)$  is the transfer function of the system  $(A, B_\gamma, \tilde{C}_\gamma)$  and  $\Theta = \int_0^\infty e^{A^*t} \tilde{C}_\gamma^* \tilde{C}_\gamma e^{At} dt$  is the observability Gramian of the pair  $(A, \tilde{C}_\gamma)$ , then

$$(77) \quad \begin{aligned} G_\gamma(s)^* G_\gamma(s) \\ = \text{diag} \left( \left( \sum_{i=1}^N \gamma_i^2 e_{i1}^2 / \gamma_1^2 \right) G_1(s)^* G_1(s), \dots, \left( \sum_{i=1}^N \gamma_i^2 e_{iN}^2 / \gamma_N^2 \right) G_N(s)^* G_N(s) \right), \end{aligned}$$

$$(78) \quad \Theta = \text{diag} \left( \left( \sum_{i=1}^N \gamma_i^2 e_{i1}^2 \right) \Theta_1, \dots, \left( \sum_{i=1}^N \gamma_i^2 e_{iN}^2 \right) \Theta_N \right), \quad \Theta_j = \int_0^\infty e^{A_j^*t} C_j^* C_j e^{A_j t} dt.$$

By (77) we have  $\|G_\gamma(i\omega)\|_{2,2}^2 = \max_{j \in \underline{N}} \left( \sum_{i=1}^N \gamma_i^2 e_{ij}^2 / \gamma_j^2 \right) \|G_j(i\omega)\|_{2,2}^2$  and hence

$$(79) \quad \max_{\omega \in \mathbb{R}} \|G_\gamma(i\omega)\|_{2,2}^2 = \max_{j \in \underline{N}} \left[ \left( \sum_{i=1}^N \gamma_i^2 e_{ij}^2 / \gamma_j^2 \right) \max_{\omega \in \mathbb{R}} \|G_j(i\omega)\|_{2,2}^2 \right] = \max_{j \in \underline{N}} \left( g_j^2 \sum_{i=1}^N \gamma_i^2 e_{ij}^2 / \gamma_j^2 \right).$$

Therefore (i) follows from Theorem 2.5.

(ii) If  $\rho \leq \rho_\gamma$ , then  $\rho^2 \leq \left[ \left( \sum_{i=1}^N \gamma_i^2 e_{ij}^2 / \gamma_j^2 \right) \max_{\omega \in \mathbb{R}} \|G_j(i\omega)\|_{2,2}^2 \right]^{-1}$  for  $j \in \underline{N}$  and it follows from Theorem 2.5 that every Riccati equation (75),  $j \in \underline{N}$  has a smallest Hermitian solution  $P_j = P_j(\rho, \gamma)$  satisfying  $\sigma(A_j + \gamma_j^{-2} B_j B_j^* P_j) \subset \overline{\mathbb{C}_-}$ . Because of the block-diagonal structure of  $A$ ,  $\tilde{C}_\gamma^* \tilde{C}_\gamma$ , and  $B_\gamma B_\gamma^*$ , the block-diagonal Hermitian matrix  $P := \oplus_{j \in \underline{N}} P_j(\rho, \gamma)$  is a solution of (73) satisfying  $\sigma(A + B_\gamma B_\gamma^* P) \subset \overline{\mathbb{C}_-}$ . Since there exists only one such solution of (73) (which is the smallest one), (ii) follows.

(iii) Since  $\rho_\gamma = [\max_{\omega \in \mathbb{R}} \|G_\gamma(i\omega)\|_{2,2}]^{-1} = r_{\mathbb{C}}(A, B_\gamma, \tilde{C}_\gamma)$  by (79), (14),  $\|\Delta\|_{\Delta} = \|\tilde{\Delta}\|_{2,2}$ , by (56), and  $\sigma(A + B(E \circ \Delta)C) = \sigma(A + B_\gamma \tilde{\Delta} \tilde{C}_\gamma)$  by (71), (iii) follows from the definition of  $r_{\mathbb{C}}(A, B_\gamma, \tilde{C}_\gamma)$ .

(iv) If the pairs  $(A_j, C_j)$ ,  $j \in \underline{N}$ , are observable and  $E$  does not have a zero column, then  $\Theta_j \succ 0$  and it follows from (78) that  $\Theta \succ 0$ ; i.e.,  $(A, \tilde{C}_\gamma)$  is observable. This proves the observability assertion in (iv). As  $\rho_\gamma = r_{\mathbb{C}}(A, B_\gamma, \tilde{C}_\gamma)$ , the remaining statements in (iv) follow by applying Theorem 2.6(iii) to  $(A, B_\gamma, \tilde{C}_\gamma)$  and observing that by (76) there exist  $\alpha > 0$ ,  $\beta > 0$  such that  $\alpha C^* C \preceq \tilde{C}_\gamma^* \tilde{C}_\gamma \preceq \beta C^* C$  (since by assumption  $\forall j \in \underline{N} \exists i \in \underline{N} : e_{ij} \neq 0$ ).  $\square$

$\rho_\gamma$  is completely determined by the scaling vector  $\gamma > 0$ , the  $H_\infty$ -norms  $g_i$  of the transfer functions  $G_i(s)$  of the subsystems  $\Sigma_i$ , and by the structure matrix  $E$ . To indicate these dependencies we write  $\rho_\gamma = \rho_\gamma(D_g, E)$ , where  $D_g = \text{diag}(g_1, \dots, g_N)$ . Equation (74) implies that  $\rho_\gamma = \infty$  if and only if  $g_j^2 \sum_{i=1}^N \gamma_i^2 e_{ij}^2 = 0$  (i.e.,  $g_j = 0$  or the  $j$ th column of  $E$  is zero) for all  $j \in \underline{N}$ .

It follows from the third statement in Theorem 5.1 and the definition of the stability radius  $r_{\Delta} = r_{\Delta}(A, B, C, E)$  that  $\rho_\gamma \leq r_{\Delta}$ . But, in general, we will have  $\rho_\gamma < r_{\Delta}$ . It is an interesting question whether  $\hat{\rho} := \sup_{\gamma \in (0, \infty)^N} \rho_\gamma$  satisfies  $\hat{\rho} = r_{\Delta}$ .

and whether there exists a scaling vector  $\hat{\gamma} \in (0, \infty)^N$  such that  $\rho_{\hat{\gamma}} = \hat{\rho}$ . We will see that under certain conditions there exists an optimizing  $\gamma$ , but even in this case one has, in general,  $\hat{\rho} \neq r_{\Delta}$ .

First, we derive a formula for  $\hat{\rho}$ . Since the  $\rho_{\gamma}$  only depend upon the nonnegative matrix  $E$  and the diagonal matrix  $D_g$ , we formulate the result for these data.

**THEOREM 5.2.** *Suppose  $E \in \mathbb{R}_+^{N \times N}$ ,  $g_1, \dots, g_N \geq 0$  are given and, for any scaling vector  $\gamma = (\gamma_1, \dots, \gamma_N) > 0$ , the number  $\rho_{\gamma}$  is defined by (74). Then  $\hat{\rho} = \sup_{\gamma \in (0, \infty)^N} \rho_{\gamma}$  is determined by*

$$(80) \quad \hat{\rho}^2 = \varrho(E^{\circ 2} D_g^2)^{-1} = \varrho(D_g^2 E^{\circ 2})^{-1}, \text{ where } D_g = \text{diag}(g_1, \dots, g_N).$$

For the proof of this theorem we will make use of the following lemma.

**LEMMA 5.3.** *For every nonnegative matrix  $M = (m_{ij}) \in \mathbb{R}_+^{N \times N}$ ,*

$$(81) \quad \varrho(M) = \inf_{y > 0} \max_{j \in \underline{N}} \frac{(y^{\top} M)_j}{y_j} = \inf_{y > 0} \max_{j \in \underline{N}} \frac{1}{y_j} \sum_{i \in \underline{N}} y_i m_{ij}.$$

Moreover, the following conditions are equivalent:

- (i)  $\varrho(M) = \min_{y > 0} \max_{j \in \underline{N}} \frac{(y^{\top} M)_j}{y_j}$ .
- (ii) *There exists a positive row vector  $z > 0$  such that  $\max_{j \in \underline{N}} \frac{1}{z_j} \sum_{i \in \underline{N}} z_i m_{ij} = \varrho(M)$ .*
- (iii) *There exists a positive row vector  $z > 0$  such that  $zM \leq \varrho(M)z$ .*
- (iv) *Let  $\pi \in \Pi_N$  be a permutation such that  $\pi(M) = (M_{hk})_{h,k \in \underline{s}}$  is of block-triangular structure where each diagonal block  $M_{hh} \in \mathbb{R}_+^{\nu_h \times \nu_h}$ ,  $h \in \underline{s}$ , is square and is either irreducible or a  $1 \times 1$  null matrix. Then, for every  $k \in \underline{s}$ ,*

$$(82) \quad \varrho(M_{kk}) = \varrho(M) \Rightarrow M_{ik} = 0 \text{ for all } i = k + 1, \dots, s.$$

In particular, conditions (i)–(iii) hold if  $M$  is irreducible.

*Proof.* Let  $(M_k)$  be a decreasing sequence of positive matrices  $M_k > 0$  converging towards  $M$ . By [13, Corollary 8.1.31]

$$\varrho(M_k) = \min_{y > 0} \max_{j \in \underline{N}} \frac{(y^{\top} M_k)_j}{y_j}, \quad k \in \mathbb{N}.$$

Hence it follows by continuity and monotonicity of the spectral radius on  $\mathbb{R}_+^{N \times N}$  (see Lemma 4.5) that

$$\begin{aligned} \varrho(M) &= \inf_{k \in \mathbb{N}} \varrho(M_k) = \inf_{k \in \mathbb{N}} \min_{y > 0} \max_{j \in \underline{N}} \frac{(y^{\top} M_k)_j}{y_j} \\ &= \inf_{y > 0} \inf_{k \in \mathbb{N}} \max_{j \in \underline{N}} \frac{(y^{\top} M_k)_j}{y_j} = \inf_{y > 0} \max_{j \in \underline{N}} \frac{(y^{\top} M)_j}{y_j}. \end{aligned}$$

This proves (81). Let us now show the equivalence of the conditions (i)–(iv). (i)  $\Leftrightarrow$  (ii) follows directly from (81). The equivalence of (ii) and (iii) follows because  $zM \leq \varrho(M)z$  and  $z > 0$  imply  $\max_{j \in \underline{N}} (zM)_j / z_j = \varrho(M)$ , again by (81).

Since conditions (i)–(iii) hold for  $M$  if and only if they hold for any permutation  $\pi(M)$ , we may assume that  $M = (M_{ij})$  is already of the block-triangular form described in (iv).



(iii)  $\Rightarrow$  (iv). Suppose that  $z = (z_1, \dots, z_N) > 0$  satisfies  $zM \leq \varrho(M)z$  and  $\varrho(M_{kk}) = \varrho(M)$  for some  $k \in \underline{s}$ . If we partition  $z$  in a compatible way with  $M$ ,  $z = (z^1, \dots, z^s)$ ,  $z^i \in \mathbb{R}_+^{1 \times \nu_i}$ , then  $zM \leq \varrho(M)z$  implies

$$z^k M_{kk} \leq \sum_{i=k}^s z^i M_{ik} \leq \varrho(M)z^k = \varrho(M_{kk})z^k.$$

On the other hand, it follows from (81) applied to  $M_{kk}$  that  $z^k M_{kk} \geq \varrho(M_{kk})z^k$ . Therefore  $z^k M_{kk} = \sum_{i=k}^s z^i M_{ik}$  and so  $\sum_{i=k+1}^s z^i M_{ik} = 0$  which implies  $M_{ik} = 0$  for  $i = k+1, \dots, s$ .

(iv)  $\Rightarrow$  (iii) is proved by induction on  $s$ . If  $s = 1$ , then  $M$  is irreducible or  $N = 1$  and  $M = [0]$ . In the first case, (iii) holds by Lemma 4.5(i). In the second case,  $\varrho(M) = 0$  and (iii) holds for any  $z = (z_1) > 0$ . Now suppose the implication (iv)  $\Rightarrow$  (iii) has been proved for  $s = 1, \dots, k-1$  for some  $k \geq 2$  and  $M$  is of the form (39) with  $s = k$ . Assume that (iv) holds.  $M$  can be written as a triangular  $2 \times 2$  block matrix

$$(83) \quad M = \begin{bmatrix} M^{11} & 0_{\nu_1 \times (N-\nu_1)} \\ M^{21} & M^{22} \end{bmatrix}, \quad M^{11} = M_{11},$$

$$M^{21} = \begin{bmatrix} M_{21} \\ \vdots \\ M_{k1} \end{bmatrix}, \quad M^{22} = \begin{bmatrix} M_{22} & 0 & \cdots & 0 \\ M_{32} & M_{33} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ M_{k2} & M_{k3} & \cdots & M_{kk} \end{bmatrix}.$$

Let  $\varrho = \varrho(M)$ ,  $\varrho_1 = \varrho(M^{11})$ ,  $\varrho_2 = \varrho(M^{22})$ ,  $I_1 = \{1, \dots, \nu_1\}$ ,  $I_2 = \{\nu_1 + 1, \dots, N\}$ . Then  $\varrho = \max\{\varrho_1, \varrho_2\}$ . First, suppose  $\varrho_1 = \varrho$ . Then  $M^{11}$  is either irreducible or  $M^{11} = [0]$ ,  $\nu_1 = 1$ , and  $\varrho_1 = 0$ . In either case there exists a left Perron vector  $z^1 = (z_1, \dots, z_{\nu_1}) > 0$  such that  $z^1 M^{11} = \varrho_1 z^1$ . If  $\varrho_2 = \varrho$ , then  $M^{22}$  satisfies condition (iv) with  $M$  replaced by  $M^{22}$ . Therefore there exists, by assumption of induction, a row vector  $z^2 = (z_{\nu_1+1}, \dots, z_N) > 0$  such that  $z^2 M^{22} \leq \varrho_2 z^2$ . Since  $M_{i1} = 0$  for  $i = 2, \dots, k$  it follows that  $z = (z^1, z^2) = (z_1, \dots, z_N)$  satisfies  $zM \leq \varrho z$ . If  $\varrho_2 < \varrho$ , then we may apply (81) to conclude that there exists  $z^2 = (z_{\nu_1+1}, \dots, z_N) > 0$  such that  $z^2 M^{22} < \varrho z^2$ . Again we obtain  $zM \leq \varrho z$  for  $z = (z^1, z^2)$ .

Finally, suppose that  $\varrho_1 < \varrho$ . As above there exists a row vector  $z^1 = (z_1, \dots, z_{\nu_1}) > 0$  such that  $z^1 M^{11} = \varrho_1 z^1$ . On the other hand, we necessarily have  $\varrho_2 = \varrho$ . Hence  $M^{22}$  satisfies condition (iv) with  $M$  replaced by  $M^{22}$ . Therefore there exists, by assumption of induction, a row vector  $z^2 = (z_{\nu_1+1}, \dots, z_N) > 0$  such that  $z^2 M^{22} \leq \varrho_2 z^2$ . For any  $\alpha > 0$  define  $z(\alpha) = (z^1, \alpha z^2) = (z_1, \dots, z_{\nu_1}, \alpha z_{\nu_1+1}, \dots, \alpha z_N) > 0$ . Since  $\varrho_1 < \varrho$  we obtain  $z(\alpha)M \leq \varrho z(\alpha)$  if  $\alpha > 0$  is chosen sufficiently small. We conclude that in both cases,  $\varrho_1 = \varrho$  and  $\varrho_1 < \varrho$ , there exists a row vector  $z > 0$  such that  $zM \leq \varrho(M)z$ . This proves (iv)  $\Rightarrow$  (iii).  $\square$

*Proof of Theorem 5.2.* First, note that the second equality in (80) follows from the general fact that  $\varrho(ED) = \varrho(DE)$  for any pair of square matrices  $D, E$ . If  $\gamma > 0$  is any scaling vector, then we have by (74)

$$(84) \quad \rho_\gamma(D_g, E)^{-2} = \max_{j \in \underline{N}} \frac{1}{\gamma_j^2} \sum_{i=1}^N \gamma_i^2 e_{ij}^2 g_j^2 = \max_{j \in \underline{N}} \frac{(\gamma^2 E^{\circ 2} D_g^2)_j}{\gamma_j^2}.$$

Hence, applying (81) to  $M = E^{\circ 2} D_g^2$  with  $y = \gamma^2$  we obtain

$$\begin{aligned}\hat{\rho}^{-2} &= \left[ \sup_{\gamma \in (0, \infty)^N} \rho_\gamma(D_g, E) \right]^{-2} \\ &= \inf_{\gamma \in (0, \infty)^N} \rho_\gamma(D_g, E)^{-2} = \inf_{\gamma \in (0, \infty)^N} \max_{j \in \underline{N}} \frac{(\gamma^2 E^{\circ 2} D_g^2)_j}{\gamma_j^2} = \varrho(E^{\circ 2} D_g^2),\end{aligned}$$

and this proves (80).  $\square$

The following remark indicates that generically  $\hat{\rho} < r_\Delta(A, B, C, E)$ .

*Remark 5.4.* It follows from Corollary 4.9 and Theorem 5.2 that

$$(85) \quad r_\Delta(A, B, C, E) = \hat{\rho} \quad \Leftrightarrow \quad \max_{\omega \in \mathbb{R}} \varrho(E^{\circ 2} D(\omega)^2) = \varrho(E^{\circ 2} D_g^2).$$

Now suppose that  $E$  is irreducible and no  $G_i(s)$  vanishes identically. Then  $\hat{\rho}^2 = r_\Delta(A, B, C, E)$  if and only if there exists a joint maximum  $\omega_0 \in \mathbb{R}$  of the  $N$  functions  $\omega \mapsto \|G_i(\omega)\|_{2,2}$ , i.e., such that  $\|G_i(\omega_0)\|_{2,2} = \max_{\omega \in \mathbb{R}} \|G_i(\omega)\|_{2,2}$  for all  $i \in \underline{N}$ . In fact, if this condition is not satisfied, then  $0 \leq D(\omega) \leq D_g$ ,  $D(\omega) \neq D_g$  and so  $\varrho(D(\omega)^2 E^{\circ 2}) = \varrho(E^{\circ 2} D(\omega)^2) < \varrho(E^{\circ 2} D_g^2)$  for all  $\omega \in \mathbb{R}$  by [4, Corollary 2.1.5] since  $E^{\circ 2} D_g^2$  is irreducible. Conversely, if the condition is satisfied, then the right-hand equality in (85) follows directly from  $\|G_i(\omega_0)\|_{2,2} = g_i$ ,  $i \in \underline{N}$ .

Now suppose that  $E$  is reducible. Then  $\hat{\rho} = r_\Delta(A, B, C, E)$  if and only if there exist a strongly connected component of  $\mathcal{G}_E$  with node set  $J \subset \underline{N}$  and a joint maximum  $\omega_0 \in \mathbb{R}$  such that  $\|G_i(\omega_0)\|_{2,2} = \max_{j \in \underline{N}} \max_{\omega \in \mathbb{R}} \|G_j(\omega)\|_{2,2}$  for all  $i \in J$ .

We will now investigate under which conditions there exists an *optimal* scaling vector, i.e.,  $\gamma > 0$  such that  $\rho_\gamma = \hat{\rho}$ .

**THEOREM 5.5.** *Suppose  $E \in \mathbb{R}_+^{N \times N}$ ,  $g_1, \dots, g_N \geq 0$  are given,  $D_g = \text{diag}(g_1, \dots, g_N)$ , and, for any scaling vector  $\gamma > 0$ , the number  $\rho_\gamma = \rho_\gamma(D_g, E)$  is defined by (74) and  $\hat{\rho}(D_g, E) = \sup_{\gamma \in (0, \infty)^N} \rho_\gamma$ . Then the following conditions are equivalent for every scaling vector  $\hat{\gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_N) > 0$ :*

- (i)  $\hat{\gamma}$  is optimal; i.e.,  $\rho_{\hat{\gamma}}(D_g, E) = \hat{\rho}(D_g, E)$ .
- (ii)  $\hat{\gamma}^2 = (\hat{\gamma}_1^2, \dots, \hat{\gamma}_N^2)$  satisfies  $\max_{j \in \underline{N}} \frac{(\hat{\gamma}^2 E^{\circ 2} D_g^2)_j}{\hat{\gamma}_j^2} = \varrho(E^{\circ 2} D_g^2)$ .

*In particular, there exists an optimal scaling vector if and only if  $M = E^{\circ 2} D_g^2$  satisfies one of the equivalent conditions (i)–(iii) of Lemma 5.3.*

*Proof.* By (74) and (80) condition (i) holds if and only if  $\hat{\gamma}^2$  satisfies

$$\max_{j \in \underline{N}} \frac{1}{\hat{\gamma}_j^2} \sum_{i=1}^N \hat{\gamma}_i^2 e_{ij}^2 g_j^2 = \rho_{\hat{\gamma}}(D_g, E)^{-2} = \hat{\rho}(D_g, E)^{-2} = \varrho(E^{\circ 2} D_g^2).$$

This proves (i)  $\Leftrightarrow$  (ii). Setting  $z = \hat{\gamma}^2$ , condition (ii) is identical with condition (ii) of Lemma 5.3 for  $M = E^{\circ 2} D_g^2$  and this yields the last statement of the theorem.  $\square$

As a consequence of Theorem 5.5 there always exists an optimal scaling vector if  $E^{\circ 2} D_g^2$  is irreducible.

**6. Nonlinear and/or time-varying perturbations.** Throughout this section we suppose the following:

$$(86) \quad \begin{aligned} &(A_i, B_i, C_i) \in \mathbb{L}_{n_i, l_i, q_i}, \quad \sigma(A_i) \subset \mathbb{C}_-, \quad G_i(s) = C_i(sI - A_i)^{-1} B_i, \quad i \in \underline{N}, \\ &A = \oplus_{i=1}^N A_i, \quad B = \oplus_{i=1}^N B_i, \quad C = \oplus_{i=1}^N C_i, \quad D(s) = \oplus_{i=1}^N \|G_i(s)\|_{2,2}, \\ &g_i := \max_{\omega \in \mathbb{R}} \|G_i(\omega)\|_{2,2}, \quad i \in \underline{N}, \quad D_g = \text{diag}(g_1, \dots, g_N), \quad E \in \mathbb{R}_+^{N \times N}. \end{aligned}$$

Let  $\Omega$  be an open neighborhood of 0 in  $\mathbb{C}^n$ , and consider time-varying nonlinear perturbations of  $\dot{x} = Ax$  of the form

$$(87) \quad \dot{x} = Ax + B(E \circ \Delta(x, t))y, \quad y = Cx,$$

where  $\Delta(\cdot, \cdot) \in \mathbf{\Delta}_{nt}(\Omega)$ . Here  $\mathbf{\Delta}_{nt}(\Omega)$  is the vector space of all bounded block matrix valued functions  $\Delta(\cdot, \cdot) = (\Delta_{ij}(\cdot, \cdot)) : \Omega \times \mathbb{R}_+ \rightarrow \mathbb{C}^{l \times q}$  with the Carathéodory properties (see section 2) provided with the norm

$$(88) \quad \|\Delta(\cdot, \cdot)\|_{\mathbf{\Delta}_{nt}} = \sup_{x \in \Omega, t \geq 0} \|\Delta(x, t)\|_{\mathbf{\Delta}}, \quad \Delta(\cdot, \cdot) = (\Delta_{ij}(\cdot, \cdot)) \in \mathbf{\Delta}_{nt}(\Omega).$$

Note that for every  $\Delta(\cdot, \cdot) \in \mathbf{\Delta}_{nt}(\Omega)$ ,  $\Delta^E(\cdot, \cdot) = E \circ \Delta(\cdot, \cdot) \in \mathbf{\Delta}_{nt}(\Omega)$  is of structure  $E$ ; i.e.,

$$\forall (x, t) \in \Omega \times \mathbb{R}_+ : \quad e_{ij} = 0 \quad \Rightarrow \quad \Delta_{ij}^E(x, t) = 0.$$

By Carathéodory's theorem, for every  $\Delta(\cdot, \cdot) \in \mathbf{\Delta}_{nt}(\Omega)$ ,  $(t_0, x^0) \in \mathbb{R}_+ \times \Omega$ , there exists a unique solution  $x_{\Delta}(t) = x_{\Delta}(t; t_0, x^0)$  of (87) with  $x_{\Delta}(t_0) = x^0$  on some maximal semiopen interval  $[t_0, t_+^{\Delta}(t_0, x^0))$ , where  $t_+^{\Delta}(t_0, x^0) > t_0$ ; see [12, Theorem 2.1.14]. In the following theorem we will see that  $t_+^{\Delta}(t_0, x^0) = \infty$  if  $\|\Delta(\cdot, \cdot)\|_{\mathbf{\Delta}_{nt}} < \hat{\rho}(D_g, E)$  and  $x^0$  is sufficiently close to the equilibrium state  $\bar{x} = 0$ . For simplicity, we call the nonlinear system (87) *uniformly (asymptotically) stable* if  $\bar{x} = 0$  is a uniformly (asymptotically) stable equilibrium position of the system (87).

**THEOREM 6.1.** *Suppose (86). Then we have the following.*

- (i) *The nonlinear system (87) is asymptotically stable for all  $\Delta \in \mathbf{\Delta}_{nt}(\Omega)$  satisfying*

$$(89) \quad \|\Delta(\cdot, \cdot)\|_{\mathbf{\Delta}_{nt}} < \hat{\rho}(D_g, E) = \sup_{\gamma > 0} \rho_{\gamma}(D_g, E).$$

*Moreover, if (89) holds, then every trajectory  $x_{\Delta}(t) = x_{\Delta}(t; t_0, x^0)$ ,  $(t_0, x^0) \in \mathbb{R}_+ \times \Omega$  of (87) with an infinite life span  $[t_0, \infty)$  tends to 0 as  $t \rightarrow \infty$ .*

- (ii) *Suppose  $\rho \leq \rho_{\gamma}(D_g, E)$  for some scaling vector  $\gamma > 0$  and  $P$  is a Hermitian solution of the associated algebraic Riccati equation (73). If  $\delta > 0$  is such that  $D_{\delta} = \{x \in \mathbb{C}^n; \langle x, Px \rangle < \delta\} \subset \Omega$ , then  $D_{\delta}$  is a joint domain of attraction of the equilibrium point  $\bar{x} = 0$  for all the systems (87) with  $\Delta \in \mathbf{\Delta}_{nt}(\Omega)$ ,  $\|\Delta(\cdot, \cdot)\|_{\mathbf{\Delta}_{nt}} < \rho$ .*
- (iii) *Suppose  $E$  does not have a zero column,  $\rho \leq \rho_{\gamma}(D_g, E)$  for some scaling vector  $\gamma > 0$ , and  $P$  is a Hermitian solution of (73). If  $r < \rho$ , then there exists a constant  $k > 0$  such that the derivative of the quadratic function  $V_{\rho}(x) = \langle x, Px \rangle$  along trajectories of (87) satisfies  $\dot{V}_{\rho}(x) \leq -k\|Cx\|^2$ ,  $x \in \Omega$ , for all  $\Delta \in \mathbf{\Delta}_{nt}(\Omega)$ ,  $\|\Delta(\cdot, \cdot)\|_{\mathbf{\Delta}_{nt}} \leq r$ . If the pairs  $(A_j, C_j)$ ,  $j \in \underline{N}$ , are observable, then  $V_{\rho}(x)$  is a joint Liapunov function at  $\bar{x} = 0$  for all perturbed systems (87) with  $\Delta \in \mathbf{\Delta}_{nt}(\Omega)$ ,  $\|\Delta(\cdot, \cdot)\|_{\mathbf{\Delta}_{nt}} \leq \rho$ .*

*Proof.* For any  $\Delta(\cdot, \cdot) \in \mathbf{\Delta}_{nt}(\Omega)$  define  $\tilde{\Delta}(\cdot, \cdot) : \Omega \times \mathbb{R}_+ \rightarrow \mathbb{C}^{l \times Nq}$  by

$$\tilde{\Delta}(x, t) = \text{diag}(\Delta^1(x, t), \dots, \Delta^N(x, t)), \quad (x, t) \in \Omega \times \mathbb{R}_+,$$

where  $\Delta^i(x, t) \in \mathbb{C}^{l_i \times q}$ ,  $i \in \underline{N}$ , is the  $i$ th block row of  $\Delta(x, t)$ . We provide the vector space  $\tilde{\mathbf{\Delta}}_{nt}(\Omega)$  of all these  $\tilde{\Delta}(\cdot, \cdot)$  with the norm (see (56))

$$(90) \quad \begin{aligned} \|\tilde{\Delta}(\cdot, \cdot)\|_{\tilde{\mathbf{\Delta}}_{nt}} &:= \sup_{x \in \Omega, t \geq 0} \|\tilde{\Delta}(x, t)\|_{2,2} \\ &= \sup_{x \in \Omega, t \geq 0} \|\Delta(x, t)\|_{\mathbf{\Delta}} = \|\Delta(\cdot, \cdot)\|_{\mathbf{\Delta}_{nt}}, \quad \Delta(\cdot, \cdot) \in \mathbf{\Delta}_{nt}(\Omega). \end{aligned}$$

The map  $\Delta(\cdot, \cdot) \mapsto \tilde{\Delta}(\cdot, \cdot)$  is an (isometric) isomorphism from  $\Delta_{nt}(\Omega)$  onto  $\tilde{\Delta}_{nt}(\Omega)$ . By (59) we have  $B(E \circ \Delta(x, t))C = B\tilde{\Delta}(x, t)\tilde{C}$ , where  $\tilde{C}$  is as in (57). Since  $\tilde{\Delta}(x, t)$  is block-diagonal, we may apply scaling to obtain from (71) that, for all  $\gamma = (\gamma_1, \dots, \gamma_N) > 0$ ,

$$(91) \quad B(E \circ \Delta(x, t))C = B\tilde{\Delta}(x, t)\tilde{C} = B_\gamma\tilde{\Delta}(x, t)\tilde{C}_\gamma, \quad (x, t) \in \Omega \times \mathbb{R}_+, \quad \Delta(\cdot, \cdot) \in \Delta_{nt}(\Omega),$$

where  $B_\gamma, \tilde{C}_\gamma$  are as in (72). Hence the uncertain time-varying nonlinear system (87) with  $\Delta(\cdot, \cdot) \in \Delta_{nt}(\Omega)$  can equivalently be described by

$$(92) \quad \dot{x} = Ax + B_\gamma\tilde{\Delta}(x, t)\tilde{C}_\gamma x,$$

where  $\tilde{\Delta}(\cdot, \cdot) \in \tilde{\Delta}_{nt}(\Omega)$  is the associated block-diagonal perturbation.

The parametrized Riccati equation (73) is of the form (15) with  $(A, B, C)$  replaced by  $(A, B_\gamma, \tilde{C}_\gamma)$ . Hence, by Theorem 2.5, there exist Hermitian solutions  $P$  of (73) if

$$\rho \leq \rho_\gamma = \left[ \max_{\omega \in \mathbb{R}} \|G_\gamma(i\omega)\|_{2,2} \right]^{-1} = r_{\mathbb{C}}(A, B_\gamma, \tilde{C}_\gamma), \quad \text{where } G_\gamma(s) = \tilde{C}_\gamma(sI_n - A)^{-1}B_\gamma.$$

By (91) the perturbed system (87) is of the form (17) with  $(A, B, C)$  replaced by  $(A, B_\gamma, \tilde{C}_\gamma)$  and  $\Delta(\cdot, \cdot) \in \Delta_{nt}(\Omega)$  replaced by  $\tilde{\Delta}(\cdot, \cdot) \in \tilde{\Delta}_{nt}(\Omega)$ . By Theorem 5.1  $(A, \tilde{C}_\gamma)$  is observable if the pairs  $(A_j, C_j)$ ,  $j \in \underline{N}$ , are observable and  $E$  does not have a zero column. Hence (ii) and (iii) follow by application of Theorem 2.6 to  $(A, B_\gamma, \tilde{C}_\gamma)$  making use of (91) and the fact that there exist  $\alpha > 0, \beta > 0$  such that  $\alpha C^*C \preceq \tilde{C}_\gamma^* \tilde{C}_\gamma \preceq \beta C^*C$ . Finally (i) follows from Theorem 2.6(i) since for every  $\Delta \in \Delta_{nt}(\Omega)$  satisfying  $\|\Delta(\cdot, \cdot)\|_{\Delta_{nt}} < \hat{\rho}(D_g, E)$  there exists  $\gamma > 0$  such that  $\|\tilde{\Delta}(\cdot, \cdot)\|_{\tilde{\Delta}_{nt}} = \|\Delta(\cdot, \cdot)\|_{\Delta_{nt}} < \rho_\gamma(D_g, E) = r_{\mathbb{C}}(A, B_\gamma, \tilde{C}_\gamma)$ .  $\square$

We now examine whether or not (89) is a tight robustness estimate. In order to do this we introduce a stability radius with respect to time-varying linear and nonlinear perturbations. Consider the following time-varying linear system:

$$(93) \quad \dot{x}(t) = Ax(t) + B(E \circ \Delta(t))Cx(t),$$

where  $\Delta(\cdot) \in \Delta_{tv}$  and  $\Delta_{tv}$  is the vector space of all bounded measurable block matrix valued functions  $\Delta(\cdot): \mathbb{R}_+ \rightarrow \mathbb{C}^{l \times q}$ . We provide  $\Delta_{tv}$  with the  $tv$ -norm

$$(94) \quad \|\Delta(\cdot)\|_{\Delta_{tv}} = \sup_{t \geq 0} \|\Delta(t)\|_{\Delta}, \quad \Delta \in \Delta_{tv}.$$

Similarly, we denote by  $\Delta_n$  the vector space of all perturbations  $\Delta \in \Delta_n := \Delta_{nt}(\mathbb{C}^n)$  which are independent of time, i.e.,  $\Delta(x, t) = \Delta(x)$ , and provide it with the norm induced from  $\Delta_{nt}$ :

$$(95) \quad \|\Delta(\cdot)\|_n = \sup_{x \in \mathbb{C}^n} \|\Delta(x)\|_{\Delta}, \quad \Delta \in \Delta_n.$$

Note that, with the obvious embeddings  $\Delta \subset \Delta_{tv} \subset \Delta_{nt}$ , the norm  $\|\cdot\|_{\Delta_{tv}}$  is the restriction of the norm  $\|\cdot\|_{\Delta_{nt}}$  to  $\Delta_{tv}$  and the norm  $\|\cdot\|_{\Delta}$  is the restriction of the norm  $\|\cdot\|_{\Delta_{tv}}$  to  $\Delta$ .

DEFINITION 6.2. *Given  $A \in \mathbb{C}^{n \times n}$  the stability radius of  $A$  with respect to complex time-varying linear perturbations  $\Delta(\cdot) \in \Delta_{tv}$  is defined by*

$$r_{\Delta_{tv}}(A, B, C, E) = \inf\{\|\Delta(\cdot)\|_{\Delta_{tv}}; \Delta(\cdot) \in \Delta_{tv} \text{ and (93) is not asymptotically stable}\}.$$

The stability radius of  $A$  with respect to complex nonlinear (respectively, nonlinear time-varying) perturbations,  $r_{\Delta_n}(A, B, C, E)$  and  $r_{\Delta_{nt}}(A, B, C, E)$ , is defined analogously.

In the full block case we have (see [12, section 5.6])

$$r_{\Delta_{nt}}(A, B, C) = r_{\Delta_n}(A, B, C) = r_{\Delta_{tv}}(A, B, C) = r_{\Delta}(A, B, C).$$

We will see that for structured perturbations the last equality does, in general, not hold.

THEOREM 6.3. *Suppose the standing assumption of this section. Then*

$$(96) \quad r_{\Delta_{nt}}(A, B, C, E) = r_{\Delta_{tv}}(A, B, C, E) = \varrho(E^{\circ 2} D_g^2)^{-1/2} = \hat{\rho}(D_g, E).$$

*Proof.* It follows from the definitions and the isometric embeddings  $\Delta_{tv} \subset \Delta_{nt}$  that  $r_{\Delta_{nt}}(A, B, C, E) \leq r_{\Delta_{tv}}(A, B, C, E)$ . On the other hand, Theorem 6.1 (with  $\Omega = \mathbb{C}^n$ ) and Theorem 5.2 imply that  $r_{\Delta_{nt}}(A, B, C, E) \geq \hat{\rho}(D_g, E) = \sup_{\gamma > 0} \rho_{\gamma}(D_g, E) = \varrho(E^{\circ 2} D_g^2)^{-1/2}$ . To prove (96) it therefore suffices to show  $r_{\Delta_{tv}}(A, B, C, E) \leq \varrho(E^{\circ 2} D_g^2)^{-1/2}$ . Suppose

$$(97) \quad \omega_i \in \mathbb{R} \text{ is such that } \|C_i(\omega_i I_{n_i} - A_i)^{-1} B_i\|_{2,2} = g_i, \quad i \in \underline{N}.$$

Replacing  $A_i$  with  $A_i - \omega_i I_{n_i}$ ,  $i \in \underline{N}$ , in Theorem 4.7 we see that  $D$  is replaced by  $D_g = \text{diag}(g_1, \dots, g_N)$ . So just as in the proof of Theorem 4.7 one can find  $\Delta = (\Delta_{ij}) \in \Delta_E$  with  $\|\Delta\|_{\Delta} = \varrho(E^{\circ 2} D_g^2)^{-1/2}$  such that

$$(98) \quad \det(A_{\omega} + B(E \circ \Delta)C) = 0, \quad A_{\omega} = \text{diag}(A_1 - \omega_1 I_{n_1}, \dots, A_N - \omega_N I_{n_N}).$$

So there exists a nonzero  $x = (x_i)_{i \in \underline{N}} \in \mathbb{C}^n$

$$(99) \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \text{ such that } \begin{bmatrix} \omega_1 x_1 \\ \vdots \\ \omega_N x_N \end{bmatrix} = (A + B(E \circ \Delta)C) \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}.$$

Let

$$(100) \quad \Delta(t) = (\Delta_{ij}(t)), \quad \Delta_{ij}(t) = \Delta_{ij} e^{i(\omega_i - \omega_j)t}, \quad x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_N(t) \end{bmatrix} = \begin{bmatrix} e^{i\omega_1 t} x_1 \\ \vdots \\ e^{i\omega_N t} x_N \end{bmatrix}.$$

Then for each  $i \in \underline{N}$ ,

$$\begin{aligned} \dot{x}_i(t) &= \omega_i e^{i\omega_i t} x_i = e^{i\omega_i t} \left[ A_i x_i + B_i \sum_{j=1}^N e_{ij} \Delta_{ij} C_j x_j \right] \\ &= A_i x_i(t) + B_i \sum_{j=1}^N e_{ij} \Delta_{ij} e^{i(\omega_i - \omega_j)t} C_j e^{i\omega_j t} x_j = A_i x_i(t) + B_i \sum_{j=1}^N e_{ij} \Delta_{ij}(t) C_j x_j(t). \end{aligned}$$

Hence  $x(t)$  satisfies  $\dot{x}(t) = Ax(t) + B(E \circ \Delta(t))Cx(t)$  and  $\|\Delta(\cdot)\|_{\Delta_{tv}} = \|\Delta\|_{\Delta} = \varrho(E^{\circ 2} D_g^2)^{-1/2}$ . Since  $x(t)$  does not tend to 0 as  $t \rightarrow 0$ , it follows that  $\dot{x}(t) = Ax(t) + B(E \circ \Delta(t))Cx(t)$  is not asymptotically stable and so  $r_{\Delta_{tv}}(A) \leq \varrho(E^{\circ 2} D_g^2)^{-1/2}$ . This completes the proof.  $\square$

*Remark 6.4.* The previous theorem shows that the robustness bound (89) is tight. Moreover, we have seen in the proof that there exists a minimum norm perturbation in  $\Delta_{tv} \subset \Delta_{nt}$  which destroys asymptotic stability; that is, there is a  $\Delta(\cdot) \in \Delta_{tv}$  of tv-norm  $r_{\Delta_{tv}}(A, B, C, E) = r_{\Delta_{nt}}(A, B, C, E)$  for which the system (93) is not asymptotically stable.

Now suppose that there is a scaling vector  $\hat{\gamma} > 0$  satisfying  $\rho_{\hat{\gamma}} = \hat{\rho} = \hat{\rho}(D_g, E)$  and  $\hat{P} \succ 0$  is a Hermitian solution of the algebraic Riccati equation (73) with  $\rho = \hat{\rho}$ . Then  $V_{\hat{\rho}}(x) = \langle x, \hat{P}x \rangle$  is a joint Liapunov function of *maximal robustness* for the uncertain system (87):  $V_{\hat{\rho}}(x)$  is a Liapunov function for every system (87) with  $\Delta \in \Delta_{nt}$ ,  $\|\Delta(\cdot, \cdot)\|_{\Delta_{nt}} \leq \hat{\rho} = \varrho(E^{\circ 2} D_g^2)^{-1/2}$ , and one can prove for every  $\rho > \varrho(E^{\circ 2} D_g^2)^{-1/2}$  that there does not exist a Liapunov function for all the systems (87) with  $\Delta \in \Delta_{tv}$ ,  $\|\Delta(\cdot, \cdot)\|_{\Delta_{tv}} \leq \rho$ . The proof proceeds similarly to the proof of Theorem 6.3 showing that there exist a time-varying perturbation  $\Delta \in \Delta_{tv}$  with  $\varrho(E^{\circ 2} D_g^2)^{-1/2} < \|\Delta(\cdot, \cdot)\|_{\Delta_{tv}} < \rho$  and a solution  $x(t)$  of (93) such that  $\lim_{t \rightarrow \infty} \|x(t)\|_2 = \infty$ .

**COROLLARY 6.5.** *Suppose (86) and consider the following statements:*

- (i)  $r_{\Delta_{tv}}(A, B, C, E) = r_{\Delta}(A, B, C, E)$ .
- (ii)  $r_{\Delta_{nt}}(A, B, C, E) = r_{\Delta_n}(A, B, C, E) = r_{\Delta_{tv}}(A, B, C, E) = r_{\Delta}(A, B, C, E)$ .
- (iii)  $\max_{\omega \in \mathbb{R}} \varrho(E^{\circ 2} D(\omega)^2) = \varrho(E^{\circ 2} D_g^2)$ .
- (iv) *There exists a joint maximum  $\omega_0 \in \mathbb{R}$  of the  $N$  functions  $\omega \mapsto \|G_i(\omega)\|_{2,2}$ ,  $i \in \underline{N}$ .*

Then (i)  $\Leftrightarrow$  (ii)  $\Leftrightarrow$  (iii)  $\Leftarrow$  (iv). If  $E$  is irreducible and  $g_i > 0$  for all  $i \in \underline{N}$ , then the four statements are all equivalent.

*Proof.* The equivalence (i)  $\Leftrightarrow$  (ii) follows from Theorem 6.3 because by definition  $r_{\Delta_{nt}}(A, B, C, E) \leq r_{\Delta_n}(A, B, C, E) \leq r_{\Delta}(A, B, C, E)$ . The equivalence (ii)  $\Leftrightarrow$  (iii) follows from Theorem 6.3 and (53).

(iv)  $\Rightarrow$  (i). Suppose that  $\omega_0 \in \mathbb{R}$  is a joint maximum of the  $N$  functions  $\omega \mapsto \|G_i(\omega)\|_{2,2}$ ; i.e.,  $\|G_i(\omega_0)\|_{2,2} = g_i$ ,  $i \in \underline{N}$ . Then Theorem 6.3 and Lemma 4.5(iv) imply

$$r_{\Delta_{tv}}(A, B, C, E) = \varrho(E^{\circ 2} D_g^2)^{-1/2} = \varrho(E^{\circ 2} D(\omega_0)^2)^{-1/2} = \left[ \max_{\omega \in \mathbb{R}} \varrho(E^{\circ 2} D(\omega)^2) \right]^{-1/2}$$

and hence (i) by (53).

(i)  $\Rightarrow$  (iv). Now suppose that  $E$  is irreducible,  $g_i > 0$  for all  $i \in \underline{N}$ , and there does not exist a joint maximum of the functions  $\omega \mapsto \|G_i(\omega)\|_{2,2}$ ,  $i \in \underline{N}$ . Let  $\omega_0 \in \mathbb{R}$  be a maximum of  $\omega \mapsto \varrho(E^{\circ 2} D(\omega)^2)$ . Then  $D(\omega_0) \leq D_g$  and  $D(\omega_0) \neq D_g$ . Since  $E$  and hence  $E^{\circ 2} D_g^2$  are irreducible, it follows from [4, Corollary 2.1.5] that

$$r_{\Delta}(A, B, C, E)^{-2} = \varrho(E^{\circ 2} D(\omega_0)^2) < \varrho(E^{\circ 2} D_g^2) = r_{\Delta_{tv}}(A, B, C, E)^{-2}.$$

This concludes the proof.  $\square$

*Remark 6.6.* (i) It is noteworthy that by the last statement in the previous corollary the equality  $r_{\Delta_{tv}}(A, B, C, E) = r_{\Delta}(A, B, C, E)$  depends only on the  $N$  isolated subsystems  $(A_i, B_i, C_i)$ ,  $i \in \underline{N}$ , and not on their specific interconnection, provided that  $E$  is irreducible.

(ii) Suppose  $A, B, C$  are real and each function  $\omega \mapsto \|G_i(\omega)\|_{2,2}$ ,  $i \in \underline{N}$ , admits its maximum on  $\mathbb{R}_+$  at  $\omega = 0$ . Then  $\omega \mapsto \varrho(D(\omega)^2 E^{\circ 2})$  admits its maximum on  $\mathbb{R}_+$  at  $\omega = 0$ , and we conclude from Corollary 6.5, Theorem 6.3, and Remark 4.10 that

$$\begin{aligned} r_{\Delta_{nt}}(A, B, C, E) &= r_{\Delta_{tv}}(A, B, C, E) \\ &= r_{\Delta}(A, B, C, E) = r_{\Delta_{\mathbb{R}}}(A, B, C, E) = \varrho(E^{\circ 2} D_g^2)^{-1/2}. \end{aligned}$$

In particular, the stability radii of  $A$  with respect to *time-invariant* and *time-varying real* perturbations of structure  $E$  are equal.

In the following extended example we determine the stability radii of a system composed of two interacting oscillators with respect to time-invariant and time-varying complex perturbations. We also determine the stability radius with respect to time-invariant real perturbations of the same structure.

*Example 6.7.* Consider a composite system consisting of two harmonic oscillators

$$\Sigma_i : \quad \dot{x}_i(t) = \begin{bmatrix} 0 & 1 \\ -\nu_i^2 & -2\xi_i\nu_i \end{bmatrix} x_i(t) + \begin{bmatrix} 0 \\ b_i \end{bmatrix} u_i(t), \quad y_i(t) = \begin{bmatrix} c_i & 0 \end{bmatrix} x_i(t), \quad i = 1, 2,$$

interconnected via  $u_1 = \delta_{12}y_2$ ,  $u_2 = \delta_{21}y_1$ . We assume  $\nu_i, \xi_i > 0$ ,  $b_i, c_i \in \mathbb{R} \setminus \{0\}$ ,  $\delta_{12}, \delta_{21} \in \mathbb{C}$ . The coupled system is of the form (31) with  $N = 2$ ,  $\mathbf{l} = \mathbf{q} = (1, 1)$  and matrices

$$(101) \quad A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\nu_1^2 & -2\xi_1\nu_1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\nu_2^2 & -2\xi_2\nu_2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ b_1 & 0 \\ 0 & 0 \\ 0 & b_2 \end{bmatrix}, \quad C = \begin{bmatrix} c_1 & 0 & 0 & 0 \\ 0 & 0 & c_2 & 0 \end{bmatrix}, \quad E = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The associated perturbation space  $\Delta_E$  is given by  $\Delta_E = \left\{ \begin{bmatrix} 0 & \delta_{12} \\ \delta_{21} & 0 \end{bmatrix}; \delta_{12}, \delta_{21} \in \mathbb{C} \right\}$  and provided with the norm  $\|\Delta\| := \|\Delta\|_{\Delta} = \max\{|\delta_{12}|, |\delta_{21}|\}$ ,  $\Delta \in \Delta_E$ . The transfer functions of the two subsystems  $\Sigma_i$  are  $G_i(s) = c_i b_i / (s^2 + 2\xi_i \nu_i s + \nu_i^2)$  and so

$$G_i(i\omega) = c_i b_i / (\nu_i^2 - \omega^2 + 2\xi_i \nu_i i\omega),$$

$$g_i = \max_{\omega \in \mathbb{R}} |G_i(i\omega)| = |c_i b_i| / \min_{\omega \in \mathbb{R}} [(\nu_i^2 - \omega^2)^2 + 4\xi_i^2 \nu_i^2 \omega^2]^{1/2}.$$

We conclude that  $r_{\Delta_{tv}} = r_{\Delta_{tv}}(A, B, C, E)$  and  $r_{\Delta} = r_{\Delta}(A, B, C, E)$  are determined by

$$r_{\Delta_{tv}} = \varrho(E^{\circ 2} D_g^2)^{-1/2} = \varrho \left( \begin{bmatrix} 0 & g_2^2 \\ g_1^2 & 0 \end{bmatrix} \right)^{-1/2} = \frac{1}{\sqrt{g_1 g_2}},$$

$$r_{\Delta} = \frac{1}{\max_{\omega \in \mathbb{R}} \sqrt{|G_1(i\omega)| |G_2(i\omega)|}}.$$

A simple calculation gives

$$|G_i(i\omega)| = \frac{|c_i b_i|}{\sqrt{(\nu_i^2 - \omega^2)^2 + 4\xi_i^2 \nu_i^2 \omega^2}}, \quad g_i = \begin{cases} |c_i b_i| / \nu_i^2 & \text{if } 1 \leq 2\xi_i^2, \\ \frac{|c_i b_i|}{2\nu_i^2 \xi_i \sqrt{1 - \xi_i^2}} & \text{if } 1 > 2\xi_i^2, \end{cases} \quad i = 1, 2.$$

Since  $E$  is irreducible, the equality  $r_{\Delta} = r_{\Delta_{tv}}$  holds if and only if there exists a joint minimum of the two even functions  $f_i : \omega \mapsto (\nu_i^2 - \omega^2)^2 + 4\xi_i^2 \nu_i^2 \omega^2$ ,  $i = 1, 2$ , on  $\mathbb{R}$ ; see Corollary 6.5. An easy calculation shows that  $f_i$  has a unique (local and global) minimum at  $\omega_0 = 0$  if  $1 \leq 2\xi_i^2$ , and has exactly two minima at  $\omega_{0i} = \pm \nu_i \sqrt{1 - 2\xi_i^2}$  if  $1 > 2\xi_i^2$ ,  $i = 1, 2$ . Hence  $r_{\Delta} = r_{\Delta_{tv}}$  holds if and only if either  $1 \leq 2\xi_1^2$  and  $1 \leq 2\xi_2^2$  or  $\nu_1^2(1 - 2\xi_1^2) = \nu_2^2(1 - 2\xi_2^2) > 0$ .

The frequencies  $\omega_{0i}$  maximize the amplitude (gain) responses of the two subsystems, so it is not surprising that they play critical roles in the stability analysis: We

have seen in the proof of Theorem 6.3 that periodic perturbations of these frequencies can be constructed which lead to nondecaying oscillations. The critical values  $\xi_i = \sqrt{1/2}$ ,  $i = 1, 2$ , are those values of the damping for which engineers regard the subsystems  $\Sigma_i$  as having only one significant overshoot.

We now consider *real* perturbations of the same structure; i.e.,  $\delta_{12}, \delta_{21} \in \mathbb{R}$ . Then

$$A + B(E \circ \Delta)C = A + B\Delta C = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\nu_1^2 & -2\xi_1\nu_1 & b_1\delta_{12}c_2 & 0 \\ 0 & 0 & 0 & 1 \\ b_2\delta_{21}c_1 & 0 & -\nu_2^2 & -2\xi_2\nu_2 \end{bmatrix},$$

$$\Delta = \begin{bmatrix} 0 & \delta_{12} \\ \delta_{21} & 0 \end{bmatrix} \in \mathbf{\Delta}_E \cap \mathbb{R}^{2 \times 2}.$$

$\Delta$  is marginally destabilizing if there exists  $\omega \in \mathbb{R}$  such that  $\det(\omega I - A - B\Delta C) = 0$ ; i.e.,

$$(102) \quad [(\nu_1^2 - \omega^2) + 2\xi_1\nu_1\omega][(\nu_2^2 - \omega^2) + 2\xi_2\nu_2\omega] = b_1b_2c_1c_2\delta_{12}\delta_{21}.$$

But this can only be the case if

$$2\xi_1\nu_1\omega(\nu_2^2 - \omega^2) + 2\xi_2\nu_2\omega(\nu_1^2 - \omega^2) = 0,$$

i.e., if  $\omega = 0$  or  $\omega^2(\xi_1\nu_1 + \xi_2\nu_2) = \nu_1\nu_2(\xi_1\nu_2 + \xi_2\nu_1)$ . We denote this latter value of  $\omega^2$  by  $\omega_c^2$ . In order to minimize  $\|\Delta\| = \max\{|\delta_{12}|, |\delta_{21}|\}$  in (102) we must choose  $\delta_{12} = \pm\delta_{21}$ . Hence if  $\omega = 0$ , then we have  $|b_1b_2c_1c_2|\|\Delta\|^2 = \nu_1^2\nu_2^2$  and if  $\omega^2 = \omega_c^2$ , we have

$$|b_1b_2c_1c_2|\|\Delta\|^2 = -(\nu_1^2 - \omega_c^2)(\nu_2^2 - \omega_c^2) + 4\xi_1\xi_2\nu_1\nu_2\omega_c^2.$$

It is easy to see that

$$(\xi_1\nu_1 + \xi_2\nu_2)(\nu_1^2 - \omega_c^2) = \xi_1\nu_1(\nu_1^2 - \nu_2^2), \quad (\xi_1\nu_1 + \xi_2\nu_2)(\nu_2^2 - \omega_c^2) = \xi_2\nu_2(\nu_2^2 - \nu_1^2).$$

Hence

$$(\xi_1\nu_1 + \xi_2\nu_2)^2|b_1b_2c_1c_2|\|\Delta\|^2 = \xi_1\xi_2\nu_1\nu_2[(\nu_2^2 - \nu_1^2)^2 + 4\nu_1\nu_2(\xi_1\nu_1 + \xi_2\nu_2)(\xi_1\nu_2 + \xi_2\nu_1)].$$

We conclude that

$$|b_1b_2c_1c_2|r_{\mathbf{\Delta}_{\mathbb{R}}}^2 = \min \left\{ \nu_1^2\nu_2^2, \frac{\xi_1\xi_2\nu_1\nu_2[(\nu_2^2 - \nu_1^2)^2 + 4\nu_1\nu_2(\xi_1\nu_1 + \xi_2\nu_2)(\xi_1\nu_2 + \xi_2\nu_1)]}{(\xi_1\nu_1 + \xi_2\nu_2)^2} \right\}.$$

As an example we consider two identical oscillators so that  $\xi_1 = \xi_2 = \xi$ ,  $\nu_1 = \nu_2 = \nu$  and for simplicity we choose  $b_1 = b_2 = c_1 = c_2 = 1$ . Then  $r_{\mathbf{\Delta}_{\mathbb{R}}} = \min\{\nu^2, 2\xi\nu^2\}$ . So  $r_{\mathbf{\Delta}_{\mathbb{R}}} = \nu^2$  if  $\xi \geq 1/2$  and  $r_{\mathbf{\Delta}_{\mathbb{R}}} = 2\xi\nu^2$  if  $\xi < 1/2$ . This is to be compared with the above results for complex perturbations that  $r_{\mathbf{\Delta}} = r_{\mathbf{\Delta}_{tv}} = \nu^2$  if  $2\xi^2 \geq 1$  and  $r_{\mathbf{\Delta}} = r_{\mathbf{\Delta}_{tv}} = 2\xi\nu^2\sqrt{1 - \xi^2}$  if  $2\xi^2 < 1$ . So we see that  $r_{\mathbf{\Delta}} < r_{\mathbf{\Delta}_{\mathbb{R}}}$  if  $\xi < 1/\sqrt{2}$ .

**7. Special cases and concluding remarks.** In this final section we illustrate our results by applying them to special classes of systems. In particular, we will prove that if the subsystems  $\Sigma_j$  are all real and one-dimensional or if they are all positive, then there is a joint maximum at  $\omega = 0$  of the  $N$  functions  $\omega \mapsto \|G_i(\omega)\|_{2,2}$ ,  $i \in \underline{N}$ , and consequently the stability radii with respect to time-varying and time-invariant parameter perturbations are equal. In a second part of this section we will discuss the relationship between our results and two classical problems of robust stability connected with the existence of joint Liapunov functions for a given set of time-invariant linear systems.



**One-dimensional subsystems.** Suppose that each subsystem  $\Sigma_i$  is one-dimensional; i.e.,  $l_i = n_i = q_i = 1$  and  $(A_i, B_i, C_i) = (a_i, b_i, c_i) \in \mathbb{L}_{1,1,1}$ ,  $i \in \underline{N}$ . Since  $B(E \circ \Delta)C = (b_i e_{ij} \delta_{ij} c_j) = (b_i e_{ij} c_j \delta_{ij})$  for  $\Delta = (\delta_{ij}) \in \mathbb{C}^{N \times N}$ , we can incorporate the diagonal entries of  $B = \text{diag}(b_1, \dots, b_N)$ ,  $C = \text{diag}(c_1, \dots, c_N)$  into  $E \in \mathbb{R}_+^{N \times N}$  and suppose  $B = C = I_N$  without restriction of generality. Hence let

$$(103) \quad A = \text{diag}(a_1, \dots, a_N) \in \mathbb{C}^{N \times N}, \quad B = I_N, \quad C = I_N, \quad E \in \mathbb{R}_+^{N \times N}.$$

We consider perturbations of the form

$$(104) \quad A \rightsquigarrow A_\Delta = A + E \circ \Delta, \quad \Delta \in \mathbf{\Delta} = \mathbb{C}^{N \times N}.$$

Note that  $E \circ \Delta \in \mathbf{\Delta}_E = \{(\delta_{ij}) \in \mathbb{C}^{N \times N}; \delta_{ij} = 0 \text{ if } e_{ij} = 0\}$  for every  $\Delta \in \mathbf{\Delta}$ .  $\mathbf{\Delta}$  is provided with the norm

$$\|\Delta\| = \max_{i \in \underline{N}} \left[ \sum_{j \in \underline{N}} |\delta_{ij}|^2 \right]^{1/2}, \quad \Delta = (\delta_{ij}) \in \mathbf{\Delta}.$$

We write  $\sigma_\Delta(A, E; \delta)$  for  $\sigma_\Delta(A, I_N, I_N, E; \delta)$ ,  $r_\Delta(A, E)$  for  $r_\Delta(A, I_N, I_N, E)$  and similarly for the other stability radii,  $r_{\Delta_{tv}}$ ,  $r_{\Delta_n}$ , and  $r_{\Delta_{nt}}$ . If  $\text{Re } a_i \neq 0$ ,  $i \in \underline{N}$ , then the transfer functions  $G_i(s)$ , the scalars  $g_i$  defined in (68), and the diagonal matrix  $D_g = \text{diag}(g_1, \dots, g_N)$  are given by

$$(105) \quad G_i(s) = (s - a_i)^{-1}, \quad g_i = \max_{\omega \in \mathbb{R}} |G_i(i\omega)| = |\text{Re } a_i|^{-1}, \quad i \in \underline{N}, \quad D_g = |\text{Re } A|^{-1}.$$

**THEOREM 7.1.** *Suppose (103). Then we have the following.*

- (i) *For every  $\delta > 0$  the spectral value set of  $A$  at level  $\delta > 0$  with respect to perturbations of the form (104) is*

$$(106) \quad \begin{aligned} \sigma_\Delta(A, E; \delta) &= \bigcup_{\Delta \in \mathbf{\Delta}, \|\Delta\| < \delta} \sigma(A + E \circ \Delta) \\ &= \sigma(A) \cup \{\lambda \in \rho(A); \varrho(|\lambda I_N - A|^{-2} E^{\circ 2}) > \delta^{-2}\}. \end{aligned}$$

- (ii) *If  $A$  is Hurwitz stable, then its stability radius with respect to perturbations of the form (104) is*

$$(107) \quad r_\Delta(A, E) = \left[ \max_{\omega \in \mathbb{R}} \varrho(|i\omega I_N - A|^{-2} E^{\circ 2}) \right]^{-1/2}.$$

*If, additionally,  $A$  is real, then*

$$(108) \quad \begin{aligned} r_{\Delta_{nt}}(A, E) &= r_{\Delta_n}(A, E) = r_{\Delta_{tv}}(A, E) \\ &= r_\Delta(A, E) = r_{\Delta_{\mathbb{R}}}(A, E) = \varrho(|A|^{-2} E^{\circ 2})^{-1/2}. \end{aligned}$$

- (iii) *If  $A$  is Hurwitz stable, then its stability radii with respect to time-varying linear or nonlinear perturbations of structure  $E$  are given by*

$$(109) \quad r_{\Delta_{nt}}(A, E) = r_{\Delta_{tv}}(A, E) = \varrho(|\text{Re } A|^{-2} E^{\circ 2})^{-1/2} = r_\Delta(\text{Re } A, E).$$

*Proof.* Part (i) and (107) follow from Corollary 4.9 since

$$D(s)^2 = [\text{diag}(|(s - a_1)^{-1}|, \dots, |(s - a_N)^{-1}|)]^2 = |(sI_N - A)^{-2}|, \quad s \in \rho(A).$$

If, additionally,  $A$  is real, then the functions  $\omega \mapsto |G_i(i\omega)| = |i\omega - a_i|^{-1}$  have a joint maximum at  $\omega = 0$  and  $D_g = |A|$ . Hence (108) follows directly from Remark 6.6(ii).

(iii) The first two equalities of (109) follow from Theorem 6.3 because of (105). The last equality follows from (108).  $\square$

**Remark 7.2.** Let  $A$  be any complex diagonal Hurwitz matrix and suppose that  $E$  is irreducible. The functions  $f_i : \omega \mapsto |G_i(i\omega)| = |i\omega - a_i|^{-1}$  have a unique maximum at  $\omega_i = \text{Im } a_i$ . Hence it follows from Corollary 6.5 that  $r_{\Delta_{tv}}(A, E) = r_{\Delta}(A, E)$  if and only if all the diagonal elements  $a_i$  of  $A$  have the same imaginary part,  $i \in \underline{N}$ . The “if” holds without the assumption that  $E$  is irreducible.

**Positive subsystems.** Positive systems occur frequently in the modelling of real processes where the state coordinates represent variables which cannot take negative values. They are used to model phenomena in such diverse fields as economics, population dynamics, and biology; see [9]. We will now show that positive block-diagonal systems have properties similar to those of the real scalar diagonal ones. A system  $(A, B, C) \in \mathbb{L}_{n,l,q}$  is called *positive* if  $B \geq 0$ ,  $C \geq 0$  and  $A$  is a Metzler matrix; i.e.,  $A + rI_n \geq 0$  for some  $r > 0$ . We use the same notation as in section 6.

**THEOREM 7.3.** Suppose (86) and that the subsystems  $(A_i, B_i, C_i) \in \mathbb{L}_{n_i, l_i, q_i}$  are positive. If  $\mathbf{n} = (n_1, \dots, n_N)$  and  $\Delta_{\mathbb{R}} = \Delta \cap \mathbb{R}^{\mathbf{n} \times \mathbf{n}}$ , then

$$\begin{aligned} r_{\Delta_{nt}}(A, B, C, E) &= r_{\Delta_{tv}}(A, B, C, E) \\ &= r_{\Delta}(A, B, C, E) = r_{\Delta_{\mathbb{R}}}(A, B, C, E) = \varrho(E \circ^2 D_g^2)^{-1/2}, \end{aligned}$$

where  $D_g = \text{diag}(\|C_1 A_1^{-1} B_1\|_{2,2}, \dots, \|C_N A_N^{-1} B_N\|_{2,2})$ .

*Proof.* Since  $(A_i, B_i, C_i)$  is positive, we have  $g_i = \max_{\omega \in \mathbb{R}} \|G_i(i\omega)\|_{2,2} = \|C_i A_i^{-1} B_i\|_{2,2}$  for  $i \in \underline{N}$ ; see [12, Example 5.3.22]. Thus the  $N$  functions  $\omega \mapsto \|G_i(i\omega)\|_{2,2}$ ,  $i \in \underline{N}$ , have a joint maximum at  $\omega_0 = 0$ . Hence the result follows from Remark 6.6(ii).  $\square$

**Quadratic and absolute stability.** We conclude this paper by applying our results to two classical robustness issues, the problems of *quadratic* and of *absolute* stability. As in the previous section we consider systems with structured norm-bounded uncertainties of the following kind:

$$(110) \quad \dot{x} = Ax + B(E \circ \Delta)y, \quad y = Cx, \quad \Delta \in \Delta(r),$$

$$(111) \quad \dot{x} = Ax + B(E \circ \Delta(t))y, \quad y = Cx, \quad \Delta \in \Delta_{tv}(r),$$

$$(112) \quad \dot{x} = Ax + B(E \circ \Delta(x))y, \quad y = Cx, \quad \Delta \in \Delta_n(r),$$

$$(113) \quad \dot{x} = Ax + B(E \circ \Delta(x, t))y, \quad y = Cx, \quad \Delta \in \Delta_{nt}(r),$$

where the data  $(A, B, C, E)$  are given as in (86),  $\Delta = \mathbb{C}^{1 \times \mathbf{q}}$ ,  $\Delta_{tv}$ ,  $\Delta_n$ ,  $\Delta_{nt}$  are defined as in the previous section, and for any uncertainty level  $r > 0$ ,  $\Delta$  varies within the following sets of norm bounded perturbations:

$$\begin{aligned} \Delta(r) &= \{\Delta \in \Delta; \|\Delta\| \leq r\}, \quad \Delta_n(r) = \{\Delta \in \Delta_n; \|\Delta\|_n \leq r\}, \\ \Delta_{tv}(r) &= \{\Delta \in \Delta_{tv}; \|\Delta\|_{tv} \leq r\}, \quad \Delta_{nt}(r) = \{\Delta \in \Delta_{nt}; \|\Delta\|_{nt} \leq r\}. \end{aligned}$$

The following definition specifies different concepts of robust stability.

**DEFINITION 7.4.** Let  $r > 0$ . The uncertain system  $(A, B, C, E)$  is said to be

- (i) asymptotically stable at level  $r$  if all the systems (110) are asymptotically stable,
- (ii) tv-stable at level  $r$  if all the systems (111) are asymptotically stable,
- (iii) absolutely stable at level  $r$  if all the systems (112) are asymptotically stable,
- (iv) nt-stable at level  $r$  if all the systems (113) are asymptotically stable.

For any uncertainty level  $r > 0$  the following implications are either trivial or easily proved:

$$(114) \quad \text{nt-stable} \Rightarrow \text{tv-stable} \Rightarrow \text{absolutely stable} \Rightarrow \text{asymptotically stable}.$$

In what follows we will discuss the relationship of these concepts with the notion of *quadratic stability*. Here we use a modification of the usual definition; compare [5].

**DEFINITION 7.5.** We say that the uncertain system  $(A, B, C, E)$  is quadratically stable at level  $r$  if there is a quadratic function  $V(x) = \langle x, Px \rangle$ , with a positive definite Hermitian matrix  $P$ , such that the derivative of  $V$  along the trajectories of (110)

$$(115) \quad \dot{V}(x) = \langle Px, Ax + B(E \circ \Delta)y \rangle + \langle Ax + B(E \circ \Delta)y, Px \rangle, \quad y = Cx$$

satisfies  $\dot{V}(x) \leq -k\|y\|^2$  for some  $k > 0$  and all  $\Delta \in \mathbf{\Delta}(r)$ .

The usual definition of quadratic stability requires that  $V(x)$  is a *strict* Liapunov function; i.e.,  $\dot{V}(x)$  is negative definite. Then it is well known that quadratic stability implies all the four properties of robust stability defined in the preceding definition. However, the quadratic Liapunov functions constructed from Riccati equations of the form (15) are usually not strict but only satisfy  $\dot{V}(x) \leq -k\|y\|^2$ . So as a direct consequence of Liapunov theory, we are only able to conclude stability but not asymptotic stability. However, a similar argument, as in the proof of Theorem 2.6, can be used to show that the condition  $\dot{V}(x) \leq -k\|y\|^2$  in fact suffices to prove asymptotic stability of the uncertain systems (110). This motivates our definition of a slightly weaker concept of quadratic stability.

In the following we will apply our previous results to the following two problems.

**Problem of quadratic stability.** Under what conditions does any one of the above properties of robust stability imply quadratic stability?

**Aizerman problem.** Under what conditions does the asymptotic stability of the uncertain system  $(A, B, C, E)$  at level  $r > 0$  imply that it is absolutely stable at level  $r$ ? (If this implication holds for every  $r > 0$ , we say that the *structured complex version of the generalized Aizerman conjecture* holds true for  $(A, B, C, E)$ .)

**Remark 7.6.** Note that the original Aizerman conjecture was stated for the *real* single input single output case where  $(A, B, C) \in \mathbb{L}_{n,1,1}$  is real and  $\mathbf{\Delta} = \mathbb{R}$ . In this real case the Aizerman conjecture is not true: there are counterexamples of  $(A, B, C) \in \mathbb{L}_{n,1,1}$  for which every matrix in the set  $\{A + B\Delta C; \Delta \in \mathbb{R}, |\Delta| \leq 1\}$  is asymptotically stable but  $(A, B, C)$  is not absolutely stable at level 1; see [25, section 7.3].

Assuming observability it is well known that in the full block case all of the above concepts of robust stability are equivalent. Note that the full block case is a very special case of the situation considered in this paper, namely  $N = 1$ ,  $\mathbf{n} = (n)$ ,  $\mathbf{l} = (l)$ ,  $\mathbf{q} = (q)$ ,  $(A, B, C) = (A_1, B_1, C_1) \in \mathbb{L}_{n,l,q}$ . In this special case we have that an uncertain *observable* system  $(A, B, C)$  is asymptotically stable at level  $r > 0$  if and only if it is quadratically stable at level  $r > 0$ ; see [12, section 5.6]. Moreover, the complex version of the Aizerman conjecture is true in the full block case [12, Theorem 5.6.22].

For structured perturbations these results, in general, no longer hold. Counterexamples have been given in [19]. Besides reformulations of the property in terms of linear matrix inequalities or  $\mu$ -analysis there are apparently no general necessary and sufficient criteria available for quadratic stability of systems with structured uncertainties; compare [1]. In our framework where the nominal system is block-diagonal and the perturbations  $E \circ \Delta \in \Delta_E$  have an arbitrarily prescribed zero structure (defined by  $E$ ), we can solve the problem of quadratic stability and establish its precise relationship with the other notions of robust stability. Moreover, we obtain computable tests for deciding whether a given uncertain system  $(A, B, C, E)$  is quadratically stable at level  $r$  or not. Finally, we obtain systematic procedures for the construction of counterexamples of uncertain systems which are asymptotically stable at a level  $r > 0$  but neither tv-stable nor quadratically stable at this level.

**THEOREM 7.7.** *Suppose (86). Then for any  $r > 0$  we have the following.*

- (i) *The uncertain system  $(A, B, C, E)$  at level  $r$  is tv-stable if and only if  $r < \varrho(E^{\circ 2} D_g^2)^{-1/2}$ . Moreover, it is tv-stable if and only if it is nt-stable.*
- (ii) *The asymptotic stability of the uncertain system  $(A, B, C, E)$  at level  $r$  implies the tv-stability at level  $r$  if and only if  $\max_{\omega \in \mathbb{R}} \varrho(D(i\omega)^2 E^{\circ 2}) = \varrho(E^{\circ 2} D_g^2)$ . In this case the structured complex version of the generalized Aizerman conjecture is valid for  $(A, B, C, E)$ .*
- (iii) *Suppose the pairs  $(A_j, C_j)$ ,  $j \in \underline{N}$ , are observable and  $E$  does not have a zero column. Then the uncertain system  $(A, B, C, E)$  is quadratically stable at level  $r$  if and only if it is tv-stable at level  $r$  (or, equivalently,  $r < \varrho(E^{\circ 2} D_g^2)$ ).*

*Proof.* In the proof of the three statements we implicitly make use of the fact that the uncertain system  $(A, B, C, E)$  at level  $r$  is tv-stable (respectively, nt-stable; respectively, asymptotically stable) if and only if  $r < r_{\Delta_{tv}}(A, B, C, E)$  (respectively,  $r < r_{\Delta_{nt}}(A, B, C, E)$ ; respectively,  $r < r_{\Delta}(A, B, C, E)$ ). This follows from the definitions of  $r_{\Delta_{nt}}$ ,  $r_{\Delta_{tv}}$ ,  $r_{\Delta}$  and the fact that there is a  $\Delta(\cdot) \in \Delta_{tv} \subset \Delta_{nt}$  with  $\|\Delta(\cdot)\|_{\Delta_{tv}} = r_{\Delta_{tv}} = r_{\Delta_{nt}}$  (respectively,  $\Delta \in \Delta$  with  $\|\Delta\|_{\Delta} = r_{\Delta}$ ) such that the corresponding perturbed system  $\dot{x} = Ax + B(E \circ \Delta(t))Cx$  (respectively,  $\dot{x} = Ax + B(E \circ \Delta)Cx$ ) is not asymptotically stable; see Remark 6.4.

(i) This follows directly from (96).

(ii) The asymptotic stability of the uncertain system  $(A, B, C, E)$  at level  $r$  implies the tv-stability at level  $r$  if and only if  $r_{\Delta_{tv}}(A, B, C, E) = r_{\Delta}(A, B, C, E)$ . Hence the first statement in (ii) follows from Corollary 6.5. Furthermore, if  $\max_{\omega \in \mathbb{R}} \varrho(D(i\omega)^2 E^{\circ 2}) = \varrho(E^{\circ 2} D_g^2)$ , then  $r_{\Delta_{nt}}(A, B, C, E) = r_{\Delta}(A, B, C, E)$  by Corollary 6.5 and therefore the structured complex version of the generalized Aizerman conjecture holds true for  $(A, B, C, E)$ .

(iii) Suppose that the uncertain system  $(A, B, C, E)$  at level  $r$  is tv-stable. Then  $r < r_{\Delta_{tv}}(A, B, C, E) = \hat{\rho}(D_g, E)$  by Theorem 6.3 and (80) and there exists a scaling vector  $\gamma > 0$  such that  $r < \rho_{\gamma}(D_g, E)$ . Let  $P$  be any Hermitian solution of the scaled algebraic Riccati equation (73) with  $\rho = \rho_{\gamma}(D_g, E)$ . By Theorem 6.1(iii)  $V(x) = \langle x, Px \rangle$  is a quadratic Liapunov function for (110) and there exists  $k > 0$  such that along the trajectories of (110) we have  $\dot{V}(x) \leq -k\|y\|^2$ ,  $y = Cx$  for all  $\Delta \in \Delta(r)$ . Hence the uncertain system (110) is quadratically stable. Conversely, if the uncertain system  $(A, B, C, E)$  is quadratically stable at level  $r$  and  $V(x) = \langle x, Px \rangle$  is the corresponding Liapunov function (see Definition 7.5), then for any solution  $x(\cdot) \neq 0$  of (111) on  $[t_0, \infty)$  we have  $V(x(t)) > 0$  and  $\dot{V}(x(t)) \leq -k\|Cx(t)\|^2$ ,  $t \geq t_0$  for some  $k > 0$ . So the origin is uniformly stable and just as in the last part of the proof of Theorem 2.6 we have  $x(t) \rightarrow 0$  as  $t \rightarrow \infty$ . Hence  $r < r_{\Delta_{tv}}$ , and this completes

the proof of (iv).  $\square$

Applying this theorem to the special cases described above, we see that the complex version of the Aizerman conjecture is valid if the subsystems  $(A_j, B_j, C_j)$ ,  $j \in \underline{N}$ , are either one-dimensional or positive. Moreover, the properties of asymptotic and quadratic stability are equivalent in these cases provided that the subsystems are observable and  $E$  does not have a zero column.

To illustrate Theorem 7.7 we return to Example 6.7 and construct a joint Liapunov function for the system consisting of two harmonic oscillators with uncertain couplings of norm  $< r_{\Delta_{tv}}(A, B, C, E)$ .

*Example 7.8.* As in Example 6.7 let

$$A_i = \begin{bmatrix} 0 & 1 \\ -\nu_i^2 & -2\xi_i\nu_i \end{bmatrix}, \quad B_i = \begin{bmatrix} 0 \\ b_i \end{bmatrix}, \quad C_i = [c_i \quad 0], \quad i = 1, 2; \quad E = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

$$\Delta_E = \left\{ \begin{bmatrix} 0 & \delta_{12} \\ \delta_{21} & 0 \end{bmatrix}; \delta_{12}, \delta_{21} \in \mathbb{C} \right\},$$

and consider the (four-dimensional) uncertain interconnected system described by

$$(116) \quad \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = [A + B(E \circ \Delta)C] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} A_1 & B_1\delta_{12}C_2 \\ B_2\delta_{21}C_1 & A_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \Delta \in \Delta_E;$$

see (101). For simplicity of presentation we assume that  $\nu_i = 1$ ,  $i = 1, 2$ . With  $\gamma = (\gamma_1, \gamma_2) > 0$  and  $\rho \leq \rho_\gamma$  the Riccati equations (75) take the form

$$(117) \quad P_1 A_1 + A_1^* P_1 + \rho^2 \gamma_2^2 C_1^* C_1 + \gamma_1^{-2} P_1 B_1 B_1^* P_1 = 0,$$

$$(118) \quad P_2 A_2 + A_2^* P_2 + \rho^2 \gamma_1^2 C_2^* C_2 + \gamma_2^{-2} P_2 B_2 B_2^* P_2 = 0.$$

In Example 6.7 we have shown that

$$r_{\Delta_{tv}} = \varrho \left( \begin{bmatrix} 0 & g_2^2 \\ g_1^2 & 0 \end{bmatrix} \right)^{-1/2} = \frac{1}{\sqrt{g_1 g_2}},$$

$$\text{where } g_i = \begin{cases} |c_i b_i| & \text{if } 1 \leq 2\xi_i^2, \\ \frac{|c_i b_i|}{2\xi_i \sqrt{1 - \xi_i^2}} & \text{if } 1 > 2\xi_i^2, \end{cases} \quad i = 1, 2.$$

By Theorem 5.2 we have  $\hat{\rho} = (g_1 g_2)^{-1/2}$ . To find an optimal scaling vector  $\hat{\gamma} > 0$  we determine a positive Perron vector  $\hat{\gamma}^2$  of  $E^{\circ 2} D_g^2$ :

$$E^{\circ 2} D_g^2 \begin{bmatrix} \hat{\gamma}_1^2 \\ \hat{\gamma}_2^2 \end{bmatrix} = \begin{bmatrix} 0 & g_2^2 \\ g_1^2 & 0 \end{bmatrix} \begin{bmatrix} \hat{\gamma}_1^2 \\ \hat{\gamma}_2^2 \end{bmatrix} = g_1 g_2 \begin{bmatrix} \hat{\gamma}_1^2 \\ \hat{\gamma}_2^2 \end{bmatrix}, \quad \text{i.e., } \hat{\gamma}_1^2 g_2 = \hat{\gamma}_2^2 g_1.$$

Choosing  $\hat{\gamma}_1 = \sqrt{g_1}$ ,  $\hat{\gamma}_2 = \sqrt{g_2}$ , we obtain  $\rho_\gamma = \hat{\rho}$  by Theorem 5.5. Then if  $P_1 = \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix}$ , the Riccati equation for the first subsystem with  $\rho = \hat{\rho}$  and  $\gamma = \hat{\gamma}$  takes the form

$$\begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & -2\xi_1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 1 & -2\xi_1 \end{bmatrix} \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix} + g_1^{-1} \begin{bmatrix} c_1^2 & 0 \\ 0 & 0 \end{bmatrix} \\ + g_1^{-1} \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & b_1^2 \end{bmatrix} \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix} = 0.$$

A similar equation holds for  $P_2$  with  $\xi_1$  replaced by  $\xi_2$  and  $g_1$  replaced by  $g_2$ . Note that although the Riccati equations (117) and (118) are coupled through the parameters  $\rho$  and  $\gamma$ , they are decoupled at the optimal parameter values. The Riccati equation for  $P_1$  can be decomposed to

$$\begin{aligned} -2p_2 + g_1^{-1}c_1^2 + g_1^{-1}b_1^2p_2^2 &= 0, \\ p_1 - 2\xi_1p_2 - p_3 + g_1^{-1}b_1^2p_2p_3 &= 0, \\ 2p_2 - 4\xi_1p_3 + g_1^{-1}b_1^2p_3^2 &= 0. \end{aligned}$$

By solving the first quadratic equation for  $p_2$ , then the third quadratic equation for  $p_3$ , and finally the middle one for  $p_1$ , one obtains the following positive definite solution of (117) (with  $\rho = \hat{\rho}$  and  $\gamma = \hat{\gamma}$ ):

$$\begin{aligned} P_1 &= \frac{|c_1|}{|b_1|} \begin{bmatrix} 2\xi_1 & 1 \\ 1 & 2\xi_1 + \sqrt{4\xi_1^2 - 2} \end{bmatrix} \quad \text{if } 1 \leq 2\xi_1^2 \quad \text{and} \\ P_1 &= \frac{|c_1|}{|b_1|\sqrt{1 - \xi_1^2}} \begin{bmatrix} 1 & \xi_1 \\ \xi_1 & 1 \end{bmatrix} \quad \text{if } 1 > 2\xi_1^2. \end{aligned}$$

Replacing  $\xi_1$  by  $\xi_2$  and  $g_1$  by  $g_2$  one obtains an analogous formula for a solution  $P_2 \succ 0$  of (117) (with  $\rho = \hat{\rho}$  and  $\gamma = \hat{\gamma}$ ). By Theorem 5.1,  $V(x) = \langle x_1, P_1x_1 \rangle + \langle x_2, P_2x_2 \rangle$  is a joint Liapunov function for all the perturbed systems (116) with

$$\Delta = \begin{bmatrix} 0 & \delta_{12} \\ \delta_{21} & 0 \end{bmatrix} \in \mathbf{\Delta}_E, \quad \|\Delta\| = \max\{|\delta_{12}|, |\delta_{21}|\} \leq \rho_{\hat{\gamma}} = (g_1g_2)^{-1/2} = r_{\mathbf{\Delta}_{tv}}.$$

By Remark 6.4,  $V(\cdot)$  is a quadratic Liapunov function of maximal robustness for the uncertain system (116).

## REFERENCES

- [1] F. AMATO, *Robust Control of Linear Systems Subject to Uncertain Time-Varying Perturbations*, Lecture Notes in Control and Inform. Sci. 325, Springer-Verlag, Berlin, 2006.
- [2] M. ARCAK AND E. D. SONTAG, *A passivity-based stability criterion for a class of interconnected systems and applications to biochemical reaction networks*, IEEE Trans. Automat. Control, 38 (2007), pp. 799–803.
- [3] M. ATHANS, ED., *Special Issue on Large-Scale Systems and Decentralized Control*, IEEE Trans. Automat. Control, AC-23 (1978).
- [4] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Classics Appl. Math. 9, SIAM, Philadelphia, 1994.
- [5] S. BOYD, L. EL GHAOU, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.
- [6] R. CASTELLANOS, A. R. MESSINA, AND H. SARMIENT, *Robust stability analysis of large power systems using structured singular value theory*, Journal of Electrical Power and Energy Systems, 27 (2005), pp. 389–397.
- [7] F. COLONIUS AND W. KLIEMANN, *The Dynamics of Control*, Birkhäuser Boston, Boston, 2000.
- [8] E. J. DAVISON, *Connectability and structural controllability of composite systems*, Automatica, 13 (1977), pp. 109–123.
- [9] L. FARINA AND S. RINALDI, *Positive Linear Systems*, John Wiley & Sons, New York, 2000.
- [10] J. FEDDEMA, C. LEWIS, AND D. SCHOENWALD, *Decentralized control of cooperative robotic vehicles: Theory and application*, IEEE Trans. Robotics Automat., 18 (2002), pp. 852–864.
- [11] D. HINRICHSSEN AND A. J. PRITCHARD, *Stability radius for structured perturbations and the algebraic Riccati equation*, Systems Control Lett., 8 (1986), pp. 105–113.
- [12] D. HINRICHSSEN AND A. J. PRITCHARD, *Mathematical Systems Theory I: Modelling, State Space Analysis, Stability and Robustness*, Texts Appl. Math. 48, Springer-Verlag, Berlin, 2005.

- [13] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990.
- [14] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
- [15] J. A. JACQUEZ AND C. P. SIMON, *Qualitative theory of compartmental systems*, SIAM Rev., 35 (1993), pp. 43–79.
- [16] M. KAROW, D. HINRICHSSEN, AND A. J. PRITCHARD, *Interconnected systems with uncertain couplings: Explicit formulae for  $\mu$ -values, spectral value sets, and stability radii*, SIAM J. Control Optim., 45 (2006), pp. 856–884.
- [17] A. N. MICHEL AND R. K. MILLER, *Qualitative Analysis of Large Scale Dynamical Systems*, Academic Press, New York, 1977.
- [18] A. P. MOLCHANOV AND Y. E. S. PYATNITSKY, *Criteria of asymptotic stability of differential and difference inclusions encountered in control theory*, Systems Control Lett., 13 (1989), pp. 59–64.
- [19] M. A. ROTEA, M. CORLESS, D. DA, AND I. R. PETERSEN, *Systems with structured uncertainty: Relations between quadratic and robust stability*, IEEE Trans. Automat. Control, 38 (1993), pp. 799–803.
- [20] D. D. ŠILJAK, *Large Scale Dynamic Systems. Stability and Structure*, North-Holland Series in System Science and Engineering, North-Holland, New York, 1979.
- [21] D. D. ŠILJAK, *Decentralized Control of Complex Systems*, Math. Sci. Engrg. 184, Academic Press, Boston, 1991.
- [22] D. D. ŠILJAK, *Dynamic graphs*, Nonlinear Anal. Hybrid Syst., 2 (2008), pp. 544–567.
- [23] R. S. VARGA, *Geršgorin and His Circles*, Springer Ser. Comput. Math. 36, Springer-Verlag, Berlin, 2004.
- [24] M. VIDYASAGAR, *Input-Output Analysis of Large-Scale Interconnected Systems. Decomposition, Well-posedness, and Stability*, Lecture Notes in Control and Inform. Sci. 29, Springer-Verlag, Berlin, 1981.
- [25] J. C. WILLEMS, *The Analysis of Feedback Systems*, Research Monograph 62, MIT Press, Cambridge, MA, 1971.
- [26] F. WIRTH, *Asymptotic behavior of the value functions of discrete-time discounted optimal control*, J. Optim. Theory Appl., 110 (2001), pp. 183–210.

## A CLASS OF SINGULAR CONTROL PROBLEMS AND THE SMOOTH FIT PRINCIPLE\*

XIN GUO<sup>†</sup> AND PASCAL TOMECEK<sup>‡</sup>

**Abstract.** This paper analyzes a class of singular control problems for which value functions are not necessarily smooth. Necessary and sufficient conditions for the well-known smooth fit principle, along with the regularity of the value functions, are given. Explicit solutions for the optimal policy and for the value functions are provided. In particular, when payoff functions satisfy the usual Inada conditions, the boundaries between action and continuation regions are smooth and strictly monotonic, as postulated and exploited in the existing literature (see [A. K. Dixit and R. S. Pindyck, *Investment under Uncertainty*, Princeton University Press, Princeton, NJ, 1994]; [M. H. A. Davis et al., *Adv. in Appl. Probab.*, 19 (1987), pp. 156–176]; [T. Ø. Kobila, *Stochastics Stochastics Rep.*, 43 (1993), pp. 29–63]; [A. B. Abel and J. C. Eberly, *J. Econom. Dynam. Control*, 21 (1997), pp. 831–852]; [A. Øksendal, *Finance Stoch.*, 4 (2000), pp. 223–250]; [J. A. Scheinkman and T. Zariphopoulou, *J. Econom. Theory*, 96 (2001), pp. 180–207]; [A. Merhi and M. Zervos, *SIAM J. Control Optim.*, 46 (2007), pp. 839–876]; and [L. H. Alvarez, *A General Theory of Optimal Capacity Accumulation under Price Uncertainty and Costly Reversibility*, Working Paper, Helsinki Center of Economic Research, Helsinki, Finland, 2006]). Illustrative examples for both smooth and nonsmooth cases are discussed to emphasize the pitfall of solving singular control problems with a priori smoothness assumptions.

**Key words.** smooth fit, singular control, switching control, reversible investment

**AMS subject classifications.** 49L20, 91B32, 91B38, 91B70

**DOI.** 10.1137/070685336

**1. Introduction.** Consider the following problem in reversible investment/capacity planning that arises naturally in resource extraction and power generation: Facing the risk of market uncertainty, a company extracts a resource (such as oil or gas) and chooses the capacity level in response to the random fluctuation of market price for the resources, subject to some capacity constraints and the associated cost for capacity expansion and contraction. The goal of the company is to maximize its long-term profit, subject to these constraints and the rate of resource extraction.

This kind of capacity planning with price uncertainty and partial (or no) reversibility originated from the economics literature and has since attracted the interest of the applied mathematics community. (See [16, 13, 11, 27, 1, 7, 32, 34, 35, 12, 20] and the references therein.) Mathematical analysis of such control problems has evolved considerably from the initial heuristics to the more sophisticated and standard stochastic control approach, and from the very special case to cases with general payoff functions. (See [22, 24, 25, 18, 19, 29, 14, 15, 10, 3, 4, 8, 9].) Most recently, Merhi and Zervos [31] analyzed this problem in great generality and provided explicit solutions for the *special case* where the payoff is of Cobb–Douglas type. Their method is to directly solve the HJB equations for the value function, assuming certain regularity conditions known as the smooth fit principle.

---

\*Received by the editors March 14, 2007; accepted for publication (in revised form) July 23, 2008; published electronically January 7, 2009.

<http://www.siam.org/journals/sicon/47-6/68533.html>

<sup>†</sup>Department of Industrial Engineering and Operations Research, University of California at Berkeley, Berkeley, CA 94720-1777 (xinguo@ieor.berkeley.edu).

<sup>‡</sup>School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853-3801 (pit3@cornell.edu).



Indeed, this smooth fit principle is critical in most of the works involving explicit solutions for optimal stopping, optimal switching, and singular control problems. However, for many control problems in real options, queuing, and wireless communications (see [30, 5, 23, 6]), there is no regularity for *either the value function or the boundaries*. In fact, one can have very simple examples where neither the value function nor the boundary between the action and continuation regions is continuous. (See Examples 5.3 and 5.4 in section 5.)

In spite of the alternative and powerful viscosity solution approach for the issue of regularity for the value function, explicit characterization for structure of the optimal policy and the value function still relies heavily on the smooth fit principle or continuity of the boundary. In fact, the smooth fit principle is sometimes exploited in the applied literature even without the standard verification theorem argument. However, when there is no continuity in the boundary or no connectedness in the interior of the continuation region, it may be hard and even incorrect to directly apply the traditional “guess the boundary and verify” approach. (See again Examples 5.3, 5.4, and 5.5 and the discussion afterwards.)

Therefore, two fundamental mathematical issues remain: (1) sufficient and necessary conditions for regularity properties for *both* the value function *and* the boundaries, and (2) characterization for the value function and for the action and continuation regions when these regularity conditions fail. Understanding these issues is especially important in cases where only numerical solutions are available, and for which the assumption on the degree of the smoothness could be *wrong*. (See also the discussions in section 5.2.)

This paper addresses the above two issues, in particular the second one, via the study of a class of singular control problems. Our work combines the techniques of [21] and [28]. The former established a fundamental connection between singular controls and switching controls, and the latter used the viscosity solution approach to solve optimal switching problems. We establish *both necessary and sufficient conditions* on the differentiability of the value function and on the smooth fit principle. These conditions lead to a derivative-based characterization of the investment, disinvestment, and continuation regions even for nonsmooth value functions. In fact, for the case *when the payoff function is not smooth*, this paper is the first to rigorously characterize the action and continuation regions, and to explicitly construct both the optimal policy and the value function. We emphasize that the payoff function in our paper  $H(\cdot)$  is *any concave function* of the capacity and may be *neither monotonic nor differentiable*. This includes the special cases investigated by [20, 31, 21]. In particular, when  $H$  satisfies the well-known Inada conditions (i.e., continuously differentiable, strictly increasing, strictly concave, with  $H(0) = 0$ ,  $H'(0^+) = \infty$ ,  $H'(\infty) = 0$ ), our results show that the boundaries between regions are indeed continuous and strictly increasing, as postulated and exploited in previous works (see [16, 13, 27, 1, 32, 34, 31, 2]). Also note that our method can be applied to more general (diffusion) processes for the price dynamics, other than the geometric Brownian motion assumed for explicitness in this paper. Finally, the construction between the functional form of the boundaries and the payoff function itself is also novel, as the value function and the boundaries may be neither smooth nor strictly monotonic, as in the existing literature.

Another work relevant to this paper is [2], which provides a great deal of economic insight into the singular control problem. However, [2] handles only payoff functions satisfying the Inada conditions. Moreover, its derivation of the value function is dependent on these assumptions and seems valid only when the boundaries are of the

very particular form illustrated in [2, Figure 1]. (See Example 5.5 and the discussion afterwards in section 5.2.)

*Outline.* The control problem is formally stated with preliminary analysis in section 2. Details of the derivation and solution are in section 3. The analysis of the regularity of the value function and the region characterization is in section 4. Examples are provided in section 5, including cases for which the value function is not differentiable, the optimal controlled process not continuous, the boundaries of the action regions not smooth, and the interior of the continuation region not simply connected.

## 2. Mathematical problem and preliminary analysis.

**2.1. Problem.** Let  $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$  be a filtered probability space, and assume a given bounded interval  $[a, b] \subset (-\infty, \infty)$ . Consider the following problem.

*Fundamental problem.*

$$(1) \quad V(x, y) := \sup_{(\xi^+, \xi^-) \in \mathcal{A}'_y} J(x, y; \xi^+, \xi^-),$$

with

$$J(x, y; \xi^+, \xi^-) := \mathbb{E} \left[ \int_0^\infty e^{-\rho t} H(Y_t) X_t^x dt - \int_0^\infty e^{-\rho t} K_1 d\xi_t^+ - \int_0^\infty e^{-\rho t} K_0 d\xi_t^- \right]$$

subject to

$$\begin{aligned} Y_t &:= y + \xi_t^+ - \xi_t^-, \quad y \in [a, b], \\ dX_t^x &:= \mu X_t^x dt + \sqrt{2\sigma} X_t^x dW_t, \quad X_0 := x > 0, \\ H : [a, b] &\rightarrow \mathbb{R} \text{ is concave with } H(y) = H(a) + \int_a^y h(z) dz, \\ K_1 + K_0 &> 0, \mu < \rho, \text{ and (without loss of generality) } K_1 > 0. \end{aligned}$$

The supremum is taken over all strategies  $(\xi^+, \xi^-) \in \mathcal{A}'_y$ , where

$$\begin{aligned} \mathcal{A}'_y &:= \left\{ (\xi^+, \xi^-) : \xi^\pm \text{ are left continuous, nondecreasing processes, } \xi_0^\pm = 0; \right. \\ &\quad y + \xi_t^+ - \xi_t^- \in [a, b]; \\ &\quad \left. \mathbb{E} \left[ \int_0^\infty e^{-\rho t} d\xi_t^+ + \int_0^\infty e^{-\rho t} d\xi_t^- \right] < \infty \right\}. \end{aligned}$$

This is a continuous time formulation of the aforementioned risk management problem. The capacity level  $Y$  is a controlled process represented by  $(\xi_t^+)_{t \geq 0}$  and  $(\xi_t^-)_{t \geq 0}$ , which are  $\mathbb{F}$ -adapted, nondecreasing càglàd processes and, respectively, stand for the cumulative capacity expansion and reduction by time  $t$ ; the market price  $X$  is modeled by a geometric Brownian motion; the rate of resource extraction is modeled by the function  $H(Y)$ ;  $K_0$  is the cost of capacity reduction with  $K_0 < 0$  representing a partial recovery of the initial investment;  $K_1$  is the cost of capacity increase. The goal of the company is to maximize its long-term profit with a payoff function that depends on both the resource extraction rate and the market price, with a form of  $H(Y)X$ .

*Remark 2.1.* We make a few remarks on the formulation here.

- First, the assumption of  $K_1 > 0$  is without loss of generality. Indeed, if  $K_1 \leq 0$ , then one considers the control problem on  $[0, b - a]$  for  $b - Y_t$  instead of  $Y_t$ .
- Second, the constraint that  $K_0 + K_1 > 0$  rules out the simple arbitrage opportunity and the condition  $\mu > \rho$  is necessary for the finiteness of the value function.
- Third, since  $h$  is clearly nonincreasing from the concavity of  $H$ , one can choose its left or right continuous versions without changing  $H$  or the value function of the control problem  $V$ .
- Finally, using a simple Ito's analysis for semimartingales, one can easily see that this formulation has several equivalent extensions. For example, one can replace  $H(Y_t)X_t$  by  $H(Y_t)(X_t) - C_0Y_t - C_1 \int_0^t Y_s ds$  to take into account possible running cost  $C_0$  and cumulating cost  $C_1$ ; one can also substitute  $H(Y_t)X_t$  with  $H(Y_t)X_t^\lambda$ .

To be consistent with the literature in (ir)reversible investment, the running payoff function in this paper is assumed to depend on the resource extraction rate and the market price in the form of  $H(Y)X$  (equivalent to  $H(Y)X^\lambda$ ), and we focus on this simple and most standard version of singular control problem (1).

**2.2. Preliminary.** Throughout the paper, we define  $m < 0 < 1 < n$  to be the roots of  $\sigma^2 x^2 + (\mu - \sigma^2)x - \rho = 0$ , so that

$$(2) \quad m, n = \frac{-(\mu - \sigma^2) \pm \sqrt{(\mu - \sigma^2)^2 + 4\sigma^2\rho}}{2\sigma^2}.$$

We also observe the identity  $\rho = -\sigma^2 mn$  and define the useful quantity  $\eta > 0$ :

$$(3) \quad \eta := \frac{1}{\rho - \mu} = \frac{-mn}{(n-1)(1-m)\rho} = \frac{1}{\sigma^2(n-1)(1-m)}.$$

Next, let  $R(x, y) := J(x, y; 0, 0)$  be the no-action expected payoff. Then,

$$(4) \quad R(x, y) := \mathbb{E} \left[ \int_0^\infty e^{-\rho t} H(y) X_t^x dt \right] = \eta H(y)x,$$

$$(5) \quad r(x, y) := R_y(x, y) = \mathbb{E} \left[ \int_0^\infty e^{-\rho t} h(y) X_t^x dt \right] = \eta h(y)x.$$

Moreover,  $|J(x, y; \xi^+, \xi^-)| < \infty$  for all  $(\xi^+, \xi^-) \in \mathcal{A}'_y$  from the boundedness of  $H$ . In fact, we have the following.

**PROPOSITION 2.2** (finiteness of value function).  $V(x, y) \leq \eta Mx + K_1(b - a)$ , where  $M = \sup_{y \in [a, b]} |H(y)| < \infty$ .

*Proof.* Let  $x > 0$  and  $y \in [a, b]$  be given. Since  $\rho > \mu$ , we have

$$\mathbb{E} \left[ \int_0^\infty e^{-\rho t} [H(Y_t)X_t^x] dt \right] \leq \mathbb{E} \left[ \int_0^\infty e^{-\rho t} [MX_t^x] dt \right] \leq \eta Mx.$$

Note that for any given  $(\xi^+, \xi^-) \in \mathcal{A}'_y$ , we have  $a - y \leq \xi_t^+ - \xi_t^- \leq b - y$ . From integration by parts, for any  $t > 0$ ,

$$(6) \quad - \int_{[0, T)} e^{-\rho t} d\xi_t^+ \leq - \int_{[0, T)} e^{-\rho t} d\xi_t^- + (y - a),$$

which, together with  $K_1 + K_0 > 0$  and  $K_1 > 0$ , implies

$$\begin{aligned} \mathbb{E} \left[ -K_1 \int_0^\infty e^{-\rho t} d\xi_t^+ - K_0 \int_0^\infty e^{-\rho t} d\xi_t^- \right] &\leq K_1(y - a) - (K_1 + K_0) \mathbb{E} \left[ \int_0^\infty e^{-\rho t} d\xi_t^- \right] \\ &\leq K_1(b - a). \end{aligned}$$

Since these bounds are independent of the control, we have

$$V(x, y) \leq \eta Mx + K_1(b - a) < \infty. \quad \square$$

**3. Deriving value function and optimal control.** Our solution approach is built on the general correspondence between singular controls and switching controls developed in [21]. For the reader's convenience, we recall here the most relevant concepts.

### 3.1. Key concepts.

**DEFINITION 3.1.** A switching control  $\alpha = (\tau_n, \kappa_n)_{n \geq 0}$  consists of an increasing sequence of stopping times  $(\tau_n)_{n \geq 0}$  and a sequence of new regime values  $(\kappa_n)_{n \geq 0}$  that are assumed immediately after each stopping time.

**DEFINITION 3.2.** A switching control  $\alpha = (\tau_n, \kappa_n)_{n \geq 0}$  is called admissible if the following hold almost surely:  $\tau_0 = 0$ ;  $\tau_{n+1} > \tau_n$  for  $n \geq 1$ ,  $\tau_n \rightarrow \infty$ ; and for all  $n \geq 0$ ,  $\kappa_n \in \{0, 1\}$  is  $\mathcal{F}_{\tau_n}$  measurable, with  $\kappa_n = \kappa_0$  for even  $n$  and  $\kappa_n = 1 - \kappa_0$  for odd  $n$ .

**PROPOSITION 3.3.** There is a one-to-one correspondence between admissible switching controls and the regime indicator function  $I_t(\omega)$ , which is an  $\mathbb{F}$ -adapted càglàd process of finite variation, so that  $I_t(\omega) : \Omega \times [0, \infty) \rightarrow \{0, 1\}$ , with

$$(7) \quad I_t := \sum_{n=0}^{\infty} \kappa_n 1_{\{\tau_n < t \leq \tau_{n+1}\}}, \quad I_0 = \kappa_0.$$

**DEFINITION 3.4.** Let  $y \in \bar{\mathcal{I}}$  be given, and for each  $z \in \mathcal{I}$ , let  $\alpha(z) = (\tau_n(z), \kappa_n(z))_{n \geq 0}$  be a switching control. The collection  $(\alpha(z))_{z \in \mathcal{I}}$  is consistent if

$$(8) \quad \alpha(z) \text{ is admissible for Lebesgue-almost every } z \in \mathcal{I},$$

$$(9) \quad I_0(z) := \kappa_0(z) = 1_{\{z \leq y\}} \text{ for Lebesgue-almost every } z \in \mathcal{I},$$

and for all  $t < \infty$ ,

$$(10) \quad \int_{\mathcal{I}} (I_t^+(z) + I_t^-(z)) dz < \infty, \text{ almost surely, and}$$

$$(11) \quad I_t(z) \text{ is decreasing in } z \text{ for } \mathbb{P} \otimes dz\text{-almost every } (\omega, z).$$

Here  $I_t^+(z)$  and  $I_t^-(z)$  are defined by

$$I_t^+ := \sum_{n > 0, \kappa_n = 1}^{\infty} 1_{\{\tau_n < t\}}, \quad I_0^+ = 0, \quad \text{and} \quad I_t^- := \sum_{n > 0, \kappa_n = 0}^{\infty} 1_{\{\tau_n < t\}}, \quad I_0^- = 0.$$

**LEMMA 3.5** (from switching controls to singular controls). Given  $y \in \bar{\mathcal{I}}$  and a consistent collection of switching controls  $(\alpha(z))_{z \in \mathcal{I}}$ , define two processes  $\xi^+$  and  $\xi^-$  by setting  $\xi_0^+ = 0$ ,  $\xi_0^- = 0$ , and for  $t > 0$ ,  $\xi_t^+ := \int_{\mathcal{I}} I_t^+(z) dz$ ,  $\xi_t^- := \int_{\mathcal{I}} I_t^-(z) dz$ . Then

1. the pair  $(\xi^+, \xi^-) \in \mathcal{A}_y$  is an admissible singular control;

2. up to indistinguishability,

$$Y_t = y + \int_y^\infty I_t(z) 1_{\{z \in \mathcal{I}\}} dz + \int_{-\infty}^y (I_t(z) - 1) 1_{\{z \in \mathcal{I}\}} dz; \quad \text{and}$$

3. for all  $t$ , we almost surely have

$$Y_t = \text{ess sup}\{z \in \mathcal{I} : I_t(z) = 1\} = \text{ess inf}\{z \in \mathcal{I} : I_t(z) = 0\},$$

where  $\text{ess sup } \emptyset := \inf \mathcal{I}$  and  $\text{ess inf } \emptyset := \sup \mathcal{I}$ .

DEFINITION 3.6. A singular control  $(\xi^+, \xi^-)$  is integrable if

$$(12) \quad \mathbb{E} \left[ \int_0^\infty e^{-\rho t} |H(Y_t) X_t^x| dt + \int_{[0, \infty)} e^{-\rho t} |K_1| d\xi_t^+ + \int_{[0, \infty)} e^{-\rho t} |K_0| d\xi_t^- \right] < \infty.$$

**3.2. Solving singular control via the switching problem.** Our derivation of the solution relies on the following connection between the value function of the singular control problem and that of a switching control problem [21, Theorem 3.7].

LEMMA 3.7. The value function in problem (1) is given by

$$(13) \quad V(x, y) = \eta H(a)x + \int_a^y v_1(x, z) dz + \int_y^b v_0(x, z) dz,$$

where  $v_0$  and  $v_1$  are solutions to the following optimal switching problems:

$$(14) \quad v_k(x, z) := \sup_{\substack{\alpha \in \mathcal{B} \\ \kappa_0 = k}} \mathbb{E} \left[ \int_0^\infty e^{-\rho t} [h(z) X_t^x] I_t dt - \sum_{n=1}^\infty e^{-\rho \tau_n} K_{\kappa_n} \right],$$

provided that the subsequent switching control is consistent and the resulting singular control is integrable. Here,  $\alpha = (\tau_n, \kappa_n)_{n \geq 0}$  is an admissible switching control,  $\mathcal{B}$  is the subset of admissible switching controls  $\alpha = (\tau_n, \kappa_n)_{n \geq 0}$  such that  $\mathbb{E}[\sum_{n=1}^\infty e^{-\rho \tau_n}] < \infty$ , and  $I_t$  is the regime indicator function for any given  $\alpha \in \mathcal{B}$ .

In light of Lemma 3.7, our derivation goes as follows: First, we shall solve the corresponding optimal switching problems; we shall then check that the corresponding collection of switching controls is consistent, which implies that it corresponds to an admissible singular control; and finally, we shall establish the existence of the corresponding integrable optimal singular control  $(\hat{\xi}^+, \hat{\xi}^-) \in \mathcal{A}'_y$  and derive the corresponding value function.

**3.2.1. Step 1: Solving the optimal switching problem.** In this section, we shall solve the switching problem (14),

$$v_k(x, z) := \sup_{\substack{\alpha \in \mathcal{B} \\ \kappa_0 = k}} \mathbb{E} \left[ \int_0^\infty e^{-\rho t} [h(z) X_t^x] I_t dt - \sum_{n=1}^\infty e^{-\rho \tau_n} K_{\kappa_n} \right].$$

First, according to [33, Theorem 1.4.1] and [28, Lemma 3.2], in addition to  $X$  being a geometric Brownian motion, we easily see the following.

PROPOSITION 3.8. For fixed  $z \in [a, b]$  and  $k \in \{0, 1\}$ ,  $v_k(x, z)$  is  $C^1$  in  $x$ . Moreover, for every  $x > 0$ ,  $|\frac{\partial}{\partial x} v_k(x, z)| \leq \eta |h(z)|$ .

Next, by modifying the argument in [28, Theorem 3.1] for  $h \geq 0$  to the case of  $h < 0$ , we obtain the following.

PROPOSITION 3.9.  $v_0$  and  $v_1$  are the unique viscosity solutions with linear growth condition to the following system of variational inequalities:

$$(15) \quad \min \{-\mathcal{L}v_0(x, z), v_0(x, z) - v_1(x, z) + K_1\} = 0,$$

$$(16) \quad \min \{-\mathcal{L}v_1(x, z) - h(z)x, v_1(x, z) - v_0(x, z) + K_0\} = 0,$$

with boundary conditions  $v_0(0^+, z) = 0$  and  $v_1(0^+, z) = \max\{-K_0, 0\}$ . Here  $\mathcal{L}$  is the generator of the diffusion  $X^x$ , killed at rate  $\rho$ , given by  $\mathcal{L}u(x, z) = \sigma^2 u_{xx}(x, z) + \mu u_x(x, z) - \rho u(x, z)$ .

*Solution to the optimal switching problem.* Finally, we explicitly solve  $v_0$  and  $v_1$  based on [28, Theorem 4.2]. We see the following cases.

*Case I:  $K_0 \geq 0$ .* For each  $z \in (a, b)$ , the switching regions are described in terms of  $F(z)$  and  $G(z)$ , which take values in  $(0, \infty]$ .

First, for each  $z \in (a, b)$  such that  $h(z) = 0$ , it is never optimal to switch, since  $K_0 \geq 0$  and  $K_1 > 0$ , and so we take  $F(z) = \infty = G(z)$ . For this case,  $v_0(x, z) = 0 = v_1(x, z)$ .

Second, for  $z$  such that  $h(z) > 0$ , we have  $G(z) < \infty$  and it is optimal to switch from regime 0 to regime 1 (to invest in the project at level  $z$ ) when  $X_t^x \in [G(z), \infty)$ . Since  $K_0 \geq 0$ , it is never optimal to switch from regime 1 to regime 0 (i.e.,  $F(z) = \infty$ ). Furthermore, we have

$$\begin{aligned} v_0(x, z) &= \begin{cases} A(z)x^n, & x < G(z), \\ \eta h(z)x - K_1, & x \geq G(z), \end{cases} \\ v_1(x, z) &= \eta h(z)x. \end{aligned}$$

Since  $v_0$  is  $C^1$  at  $G(z)$ , we get

$$\begin{cases} A(z)G(z)^n &= \eta h(z)G(z) - K_1, \\ nA(z)G(z)^{n-1} &= \eta h(z). \end{cases}$$

That is,

$$\begin{cases} G(z) &= \nu h(z)^{-1}, \\ A(z) &= \frac{K_1}{(n-1)} G(z)^{-n} = \frac{K_1}{(n-1)} \nu^{-n} h(z)^n, \end{cases}$$

where  $\nu = K_1 \sigma^2 n(1 - m)$ .

Finally, when  $h(z) < 0$ , it is optimal to switch from regime 1 to regime 0 (disinvest at level  $z$ ) when  $X_t^x \in [F(z), \infty)$ . Since  $K_1 > 0$ , it is never optimal to switch from regime 0 to regime 1 (i.e.,  $G(z) = \infty$ ). The derivation of the value function proceeds analogously to the derivation for the case of  $h(z) > 0$ .

*Case II:  $K_0 < 0$ .* First of all, for each  $z \in (a, b)$  such that  $h(z) \leq 0$ , it is always optimal to disinvest because  $K_0 < 0$ . That is,  $F(z) = \infty = G(z)$ . In this case, clearly  $v_0(x, z) = 0$  and  $v_1(x, z) = -K_0$ .

Next, for each  $z \in (a, b)$  such that  $h(z) > 0$ , it is optimal to switch from regime 0 to regime 1 (to invest in the project at level  $z$ ) when  $X_t^x \in [G(z), \infty)$ , and to switch from regime 1 to regime 0 (disinvest at level  $z$ ) when  $X_t^x \in (0, F(z)]$ , where  $0 < F(z) < G(z) < \infty$ .

Moreover,  $v_0$  and  $v_1$  are given by

$$\begin{aligned} v_0(x, z) &= \begin{cases} A(z)x^n, & x < G(z), \\ B(z)x^m + \eta x h(z) - K_1, & x \geq G(z), \end{cases} \\ v_1(x, z) &= \begin{cases} A(z)x^n - K_0, & x \leq F(z), \\ B(z)x^m + \eta x h(z), & x > F(z). \end{cases} \end{aligned}$$

Smoothness of  $v(x, z)$  at  $x = G(z)$  and  $x = F(z)$  from Proposition 3.8 leads to

$$(17) \quad \begin{cases} A(z)G(z)^n &= B(z)G(z)^m + \eta G(z)h(z) - K_1, \\ nA(z)G(z)^{n-1} &= mB(z)G(z)^{m-1} + \eta h(z), \\ A(z)F(z)^n &= B(z)F(z)^m + \eta F(z)h(z) + K_0, \\ nA(z)F(z)^{n-1} &= mB(z)F(z)^{m-1} + \eta h(z). \end{cases}$$

Eliminating  $A(z)$  and  $B(z)$  from (17) yields

$$(18) \quad \begin{cases} K_1 G(z)^{-m} + K_0 F(z)^{-m} &= \frac{-m}{(1-m)\rho} h(z) (G(z)^{1-m} - F(z)^{1-m}), \\ K_1 G(z)^{-n} + K_0 F(z)^{-n} &= \frac{n}{(n-1)\rho} h(z) (G(z)^{1-n} - F(z)^{1-n}). \end{cases}$$

Since the viscosity solutions to the variational inequalities are unique and  $C^1$  according to Proposition 3.9, for every  $z$  there is a unique solution  $F(z) < G(z)$  to (18). Let  $\kappa(z) = F(z)h(z)$ ,  $\nu(z) = G(z)h(z)$ ; then the following system of equations for  $\kappa(z)$  and  $\nu(z)$  is guaranteed to have a unique solution for each  $z$ :

$$\begin{cases} K_1 \nu(z)^{-m} + K_0 \kappa(z)^{-m} &= \frac{-m}{(1-m)\rho} (\nu(z)^{1-m} - \kappa(z)^{1-m}), \\ K_1 \nu(z)^{-n} + K_0 \kappa(z)^{-n} &= \frac{n}{(n-1)\rho} (\nu(z)^{1-n} - \kappa(z)^{1-n}). \end{cases}$$

Moreover, these equations depend on  $z$  only through  $\nu(z)$  and  $\kappa(z)$ , implying that there exist unique constants  $\kappa, \nu$  such that  $\kappa(z) \equiv \kappa$  and  $\nu(z) \equiv \nu$  for all  $z$ . Hence  $F(z) = \kappa h(z)^{-1}$ ,  $G(z) = \nu h(z)^{-1}$ , with  $\kappa < \nu$  being the unique solutions to

$$\begin{cases} \frac{1}{1-m} [\nu^{1-m} - \kappa^{1-m}] &= -\frac{\rho}{m} [K_1 \nu^{-m} + K_0 \kappa^{-m}], \\ \frac{1}{n-1} [\nu^{1-n} - \kappa^{1-n}] &= \frac{\rho}{n} [K_1 \nu^{-n} + K_0 \kappa^{-n}]. \end{cases}$$

Given  $F(z)$  and  $G(z)$ ,  $A(z)$  and  $B(z)$  are solved from (17),

$$\begin{cases} B(z) &= -\frac{G(z)^{-m}}{n-m} \left( \frac{G(z)h(z)}{\sigma^2(1-m)} - nK_1 \right) = -\frac{F(z)^{-m}}{n-m} \left( \frac{F(z)h(z)}{\sigma^2(1-m)} + nK_0 \right), \\ A(z) &= \frac{G(z)^{-n}}{n-m} \left( \frac{G(z)h(z)}{\sigma^2(n-1)} + mK_1 \right) = \frac{F(z)^{-n}}{n-m} \left( \frac{F(z)h(z)}{\sigma^2(n-1)} - mK_0 \right). \end{cases}$$

**3.2.2. Step 2: Corresponding switching controls.** Given the solution to the optimal switching problems, it is clear that the optimal switching control for any level  $z \in (a, b)$  is given by the following.

*Case I.* For  $z \in (a, b)$  and  $x > 0$ , let  $F$  and  $G$  be as given in Case I of Theorem 3.14. The switching control  $\hat{\alpha}_k(x, z) = (\hat{\tau}_n(x, z), \hat{\kappa}_n(z))_{n \geq 0}$ , starting from  $\hat{\tau}_0(x, z) = 0$  and  $\hat{\kappa}_0(z) = k$  is given as the following:

- for  $n \geq 1$  if  $k = 0$ ,  $\hat{\tau}_1(x, z) = \inf\{t > 0 : X_t^x \in [G(z), \infty)\}$ , and for  $n \geq 2$ ,  $\hat{\tau}_n(z) = \infty$ ;
- for  $n \geq 1$  if  $k = 1$ ,  $\hat{\tau}_1(x, z) = \inf\{t > 0 : X_t^x \in [F(z), \infty)\}$ , and for  $n \geq 2$ ,  $\hat{\tau}_n(z) = \infty$ .

*Case II.* For  $z \in (a, b)$  and  $x > 0$ ,  $F$  and  $G$  are as given in Case II of Theorem 3.14. The switching control  $\hat{\alpha}_k(x, z) = (\hat{\tau}_n(x, z), \hat{\kappa}_n(z))_{n \geq 0}$ , starting from  $\hat{\tau}_0(x, z) = 0$  and  $\hat{\kappa}_0(z) = k$  is given as the following:

- for  $n \geq 1$  if  $\hat{\kappa}_{n-1}(z) = 0$ ,  $\hat{\tau}_n(x, z) = \inf\{t > \tau_{n-1} : X_t^x \in [G(z), \infty)\}$ ,  $\hat{\kappa}_n(z) = 1$ ;
- for  $n \geq 1$  if  $\hat{\kappa}_{n-1}(z) = 1$ ,  $\hat{\tau}_n(x, z) = \inf\{t > \tau_{n-1} : X_t^x \in (0, F(z)]\}$ ,  $\hat{\kappa}_n(z) = 0$ .

**3.2.3. Step 3: Consistency of the switching controls.** Now, define the collection of admissible switching controls  $(\hat{\alpha}(x, z))_{z \in (a, b)}$  so that  $\hat{\alpha}(x, z) = \hat{\alpha}_0(x, z)$  for  $z > y$  and  $\hat{\alpha}(x, z) = \hat{\alpha}_1(x, z)$  for  $z \leq y$ . Then we have the following.

PROPOSITION 3.10. *The collection of switching controls  $(\hat{\alpha}(x, z))_{z \in (a, b)}$  is consistent.*

To prove the consistency, the following monotonicity properties of  $F$  and  $G$  are essential:  $F$  is nonincreasing and  $G$  is nondecreasing in Case I, and  $F$  is nondecreasing and  $G$  is nonincreasing in Case II.

To start, for each  $z \in (a, b)$ , denote  $\hat{I}_t(x, z)$  to be the regime indicator function of the optimal switching control  $\hat{\alpha}(x, z)$ . That is,  $\hat{I}_t(x, z) = \sum_{n=0}^{\infty} \hat{\kappa}_n(z) 1_{\{\hat{\tau}_n(x, z) < t \leq \hat{\tau}_{n+1}(x, z)\}}$ . Then the consistency follows from the following lemmas.

LEMMA 3.11. *For every  $x > 0$  and  $t > 0$ ,  $\hat{I}_t(x, \cdot)$  is nonincreasing.*

*Proof.* For simplicity, we omit the dependence on  $x$  from the notation.

- *Case I.* Fix  $x > 0$  and  $t > 0$ . Let  $w < z$  be given and suppose that  $\hat{I}_t(z) = 1$ . In the event that  $t \leq \hat{\tau}_1(z)$ , we have  $w < z \leq y$  and hence  $F(w) \geq F(z)$  since  $F$  is nonincreasing. So by definition,  $\hat{\tau}_1(w) \geq \hat{\tau}_1(z) \geq t$ . Thus,  $\hat{I}_t(w) = 1$  for  $w \leq y$ .

Now in the event that  $t > \hat{\tau}_1(z)$ ,  $\hat{I}_t(z) = 1$  implies that for some  $s < t$ ,  $X_s^x \in [G(z), \infty)$ , i.e.,  $\sup\{s \leq t : X_s^x \geq G(z)\} \geq G(z)$ . However, since  $G$  is nondecreasing,  $G(z) \geq G(w)$ . Hence  $\sup\{s \leq t : X_s^x \geq G(w)\} \geq G(w)$  and  $\hat{I}_t(w) = 1$ .

Since  $\hat{I}_t(z) = 1$  implies that  $\hat{I}_t(w) = 1$  for any  $w < z$ ,  $\hat{I}_t(x, \cdot)$  is nonincreasing.

- *Case II.* Fix  $x > 0$  and  $t > 0$ . Let  $w < z$  be given and suppose that  $\hat{I}_t(z) = 1$ . In the event that  $t \leq \hat{\tau}_1(z)$ , we have  $w < z \leq y$  and hence  $F(w) \leq F(z)$ . So by definition,  $\hat{\tau}_1(w) \geq \hat{\tau}_1(z) \geq t$ . Thus,  $\hat{I}_t(w) = 1$  for  $w \leq y$ .

Now in the event that  $t > \hat{\tau}_1(z)$ ,  $\hat{I}_t(z) = 1$  implies that for some  $s < t$ ,  $X_s^x \in [G(z), \infty)$  and also that  $X^x$  must have been in the set  $[G(z), \infty)$  more recently than in  $[0, F(z)]$ , i.e.,

$$\sup\{s \leq t : X_s^x \in [G(z), \infty)\} > \sup\{s \leq t : X_s^x \in (0, F(z)]\}.$$

However, since  $[G(z), \infty) \subset [G(w), \infty)$  and  $(0, F(w)] \subset (0, F(z)]$  for  $w < z$ , this implies

$$\begin{aligned} \sup\{s \leq t : X_s^x \in [G(w), \infty)\} &\geq \sup\{s \leq t : X_s^x \in [G(z), \infty)\} \\ &> \sup\{s \leq t : X_s^x \in (0, F(z)]\} \\ &\geq \sup\{s \leq t : X_s^x \in (0, F(w)]\}. \end{aligned}$$

Hence  $X^x$  was in  $[G(w), \infty)$  more recently than in  $(0, F(w)]$ , meaning that  $\hat{I}_t(w) = 1$ .

Since  $\hat{I}_t(z) = 1$  implies that  $\hat{I}_t(w) = 1$  for any  $w < z$ ,  $\hat{I}_t(x, \cdot)$  is nonincreasing.  $\square$

LEMMA 3.12. *For every  $x > 0$ ,  $t > 0$ ,  $\int_a^b (\hat{I}_t^+(x, z) + \hat{I}_t^-(x, z)) dz < \infty$  almost surely.*

*Proof.*

- *Case I.* This is easy to prove by recalling that  $\hat{I}_t^+(x, z) + \hat{I}_t^-(x, z)$  represents the number of switches at level  $z$  up to time  $t$ . Since there is at most one switch at each level  $z$ ,  $\hat{I}_t^+(x, z) + \hat{I}_t^-(x, z) \leq 1$ . Hence  $\int_a^b (\hat{I}_t^+(x, z) + \hat{I}_t^-(x, z)) dz \leq b - a < \infty$ .
- *Case II.* Since  $[a, b]$  is bounded, it suffices to show that for all  $(x, t)$ ,  $\hat{I}_t^+(x, z) + \hat{I}_t^-(x, z)$  is almost surely bounded in  $z$ . Let  $x > 0$  and  $t > 0$  be given.



Recall that  $\hat{I}_t^+(x, z) + \hat{I}_t^-(x, z)$  represents the number of switches at level  $z$  up to time  $t$ . When  $h(z) \leq 0$ , there is exactly one switch. When  $h(z) > 0$ ,  $0 < F(z) < G(z) < \infty$ ,  $G(z) = \nu h(z)^{-1}$ , and  $F(z) = \kappa h(z)^{-1}$ . Note that after the first switch, each subsequent switch requires that  $X^x$  move from  $(0, F(z))$  to  $[G(z), \infty)$  or vice versa.

Alternatively,  $\log(X^x)$  must move from  $(-\infty, \log(F(z)))$  to  $[\log(G(z)), \infty)$ , traveling a minimum distance of  $\log(G(z)) - \log(F(z)) = \log(\nu) - \log(\kappa) > 0$  for each switch. In particular, this quantity is independent of  $z$ .

Since  $\log(X^x)$  is a Brownian motion with drift, its sample paths are almost surely uniformly continuous on  $[0, t]$ . Thus, for almost all  $\omega \in \Omega$ , there exists some  $\delta(\omega) > 0$  such that for any  $x > 0$  and all  $s, r \in [0, t]$ , with  $|s - r| < \delta(\omega)$ ,

$$|\log(X_s^x(\omega)) - \log(X_r^x(\omega))| < \log(\nu) - \log(\kappa) = \log(G(z)) - \log(F(z)).$$

Hence, for any level  $z \in [a, b]$ , there is at least  $\delta(\omega)$  amount of time in between any two switches (after the first one). Hence there can be at most  $1 + \frac{t}{\delta(\omega)}$  switches at level  $z$  in  $[0, t]$ . Thus

$$\hat{I}_t^+(x, z) + \hat{I}_t^-(x, z) \leq 1 + \frac{t}{\delta} < \infty \quad \text{almost surely.} \quad \square$$

**3.2.4. Step 4: Corresponding optimal singular control.** It remains to check the integrability of the corresponding singular control, which is obvious from the following proposition due to [31].

**PROPOSITION 3.13.** *For any  $y \in [a, b]$  and any pair  $(\xi^+, \xi^-)$  of left continuous, nondecreasing processes, with  $\xi_0^\pm = 0$  and  $y + \xi_t^+ - \xi_t^- \in [a, b]$  for all  $t$ , either*

- A.  $(\xi^+, \xi^-) \in \mathcal{A}'_y$ , or
- B. *there exists an  $\mathbb{F}$ -adapted process  $Z$  such that  $U. \leq Z$ . almost surely,  $\mathbb{E}[|Z_T|] < \infty$  for all  $T \geq 0$ , and  $\limsup_{T \rightarrow \infty} \mathbb{E}[Z_T] = -\infty$ , where*

$$U_T(y, \xi^+, \xi^-) := \int_0^T e^{-\rho t} [H(Y_t) X_t^x] dt - K_1 \int_{[0, T)} e^{-\rho t} d\xi_t^+ - K_0 \int_{[0, T)} e^{-\rho t} d\xi_t^-.$$

Therefore, we conclude that there exists a corresponding integrable singular control  $(\hat{\xi}^+, \hat{\xi}^-) \in \mathcal{A}'_y$ , and define  $\hat{Y}_t = y + \hat{\xi}_t^+ - \hat{\xi}_t^-$ .

**3.2.5. Solution: Value function and optimal control.** In short, we summarize the solution for problem (1) as follows.

**THEOREM 3.14 (value function).** *The value function  $V(x, y)$  is given by*

$$(19) \quad V(x, y) = \eta H(a)x + \int_a^y v_1(x, z) dz + \int_y^b v_0(x, z) dz,$$

where  $v_0$  and  $v_1$  are given explicitly based on  $K_0$  as follows.

Case I:  $(K_0 \geq 0)$ .

- 1. For each  $z \in (a, b)$  such that  $h(z) = 0$ ,  $v_0(x, z) = v_1(x, z) = 0$ .
- 2. For each  $z \in (a, b)$  such that  $h(z) > 0$ ,

$$\begin{cases} v_0(x, z) &= \begin{cases} A(z)x^n, & x < G(z), \\ \eta h(z)x - K_1, & x \geq G(z), \end{cases} \\ v_1(x, z) &= \eta h(z)x, \end{cases}$$

where  $G(z) = \nu h(z)^{-1}$ , and  $A(z) = \frac{K_1}{(n-1)} \left(\frac{h(z)}{\nu}\right)^n$ , with  $\nu = K_1 \sigma^2 n(1 - m)$ .

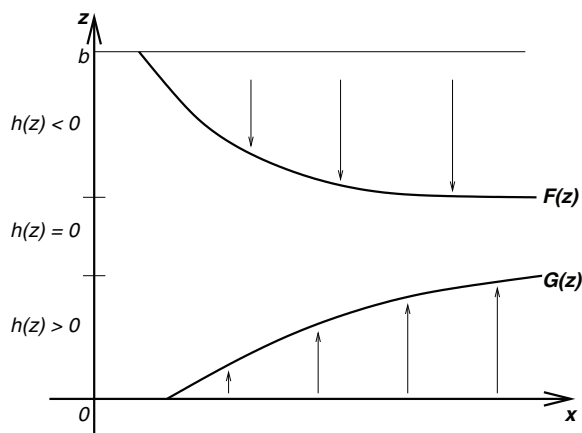
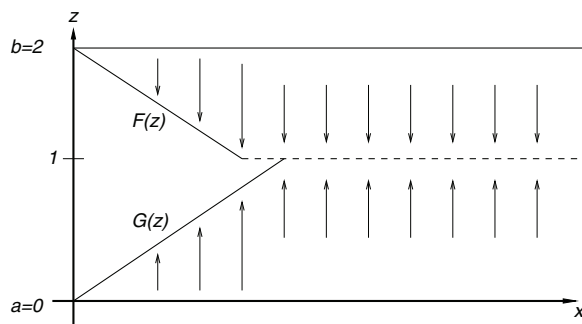


FIG. 1. Illustration of Case I of Theorem 3.14 when boundaries are smooth.

FIG. 2. Illustration of Case I of Theorem 3.14 when boundaries are NOT smooth:  $h(z) = c/z$  for  $z < 1$  and  $h(z) = -d/(2-z)$  for  $z > 1$  with  $K_0/d < K_1/c$ .

3. For each  $z \in (a, b)$  such that  $h(z) < 0$ ,

$$\begin{cases} v_0(x, z) = 0, \\ v_1(x, z) = \begin{cases} B(z)x^n + \eta h(z)x, & x < F(z), \\ -K_0, & x \geq F(z), \end{cases} \end{cases}$$

where  $F(z) = -\frac{\kappa}{h(z)}$ , and  $B(z) = \frac{K_0}{(n-1)}\kappa^{-n}(-\frac{h(z)}{\kappa})^n$ , with  $\kappa = K_0\sigma^2n(1-m)$ .

See Figures 1 and 2.

Case II: ( $K_0 < 0$ ).

1. For each  $z \in (a, b)$  such that  $h(z) \leq 0$ ,  $v_0(x, z) = 0$ ,  $v_1(x, z) = -K_0$ .

2. For each  $z \in (a, b)$  such that  $h(z) > 0$ ,

$$(20) \quad v_0(x, z) = \begin{cases} A(z)x^n, & x < G(z), \\ B(z)x^m + \eta h(z)x - K_1, & x \geq G(z), \end{cases}$$

$$(21) \quad v_1(x, z) = \begin{cases} A(z)x^n - K_0, & x \leq F(z), \\ B(z)x^m + \eta h(z)x, & x > F(z). \end{cases}$$

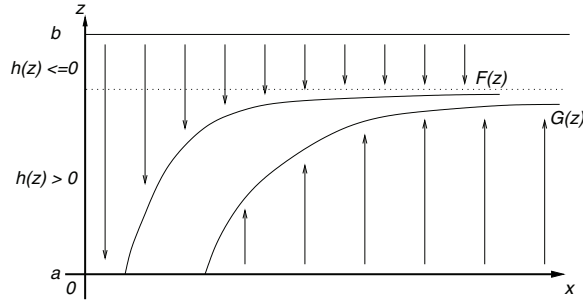


FIG. 3. Illustration of Case II of Theorem 3.14 when boundaries are smooth.

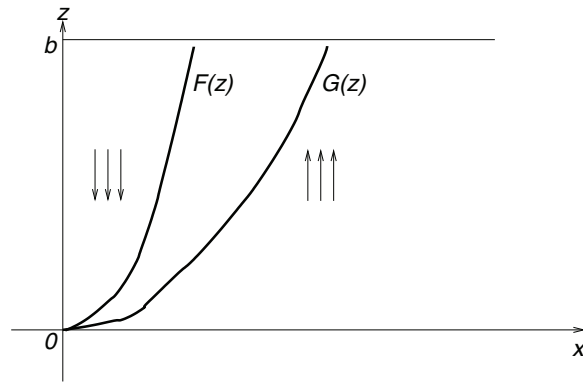


FIG. 4. Example of Case II of Theorem 3.14 when boundaries are smooth.

Here

$$(22) \quad A(z) = \frac{h(z)^n}{(n-m)\nu^n} \left( \frac{\nu}{\sigma^2(n-1)} + mK_1 \right) = \frac{h(z)^n}{(n-m)\kappa^n} \left( \frac{\kappa}{\sigma^2(n-1)} - mK_0 \right),$$

$$(23) \quad B(z) = \frac{-h(z)^m}{(n-m)\nu^m} \left( \frac{\nu}{\sigma^2(1-m)} - nK_1 \right) = \frac{-h(z)^m}{(n-m)\kappa^m} \left( \frac{\kappa}{\sigma^2(1-m)} + nK_0 \right).$$

The functions  $F$  and  $G$  are nondecreasing with

$$(24) \quad F(z) = \frac{\kappa}{h(z)} \quad \text{and} \quad G(z) = \frac{\nu}{h(z)},$$

where  $\kappa < \nu$  are the unique solutions to

$$(25) \quad \frac{1}{1-m} [\nu^{1-m} - \kappa^{1-m}] = -\frac{\rho}{m} [K_1\nu^{-m} + K_0\kappa^{-m}],$$

$$(26) \quad \frac{1}{n-1} [\nu^{1-n} - \kappa^{1-n}] = \frac{\rho}{n} [K_1\nu^{-n} + K_0\kappa^{-n}].$$

See Figures 3 and 4.

**THEOREM 3.15** (optimal control). *The optimal singular control  $(\hat{\xi}^+, \hat{\xi}^-) \in \mathcal{A}'_y$  exists. For each  $z \in (a, b)$ , the optimal control is described in terms of  $F(z)$  and  $G(z)$  from Theorem 3.14 such that the following hold.*

- (Case I,  $K_0 \geq 0$ ): For  $z$  such that  $h(z) > 0$ , it is optimal to invest in the project past level  $z$  when  $X_t^x \in [G(z), \infty)$ , and never disinvest. When  $h(z) < 0$ , it is optimal to disinvest below level  $z$  when  $X_t^x \in [F(z), \infty)$ , and it is never optimal to invest. When  $h(z) = 0$ , it is optimal to neither invest nor disinvest (i.e.,  $F(z) = \infty = G(z)$ ).
- (Case II,  $K_0 < 0$ ): For  $z$  such that  $h(z) > 0$ , it is optimal to invest in the project past level  $z$  when  $X_t^x \in [G(z), \infty)$ , and to disinvest below level  $z$  when  $X_t^x \in (0, F(z)]$ . For  $z$  such that  $h(z) \leq 0$ , it is always optimal to disinvest.

**3.3. Optimally controlled process.** Having obtained the value function and the optimal control policy, now we can describe explicitly the optimally controlled process.

First, from Lemma 3.5 we clearly have the following.

LEMMA 3.16. *Given  $(x, y) \in (0, \infty) \times [a, b]$ , the optimally controlled process  $\hat{Y}$  is indistinguishable from  $\sup\{z \in (a, b) : \hat{I}_t(x, z) = 1\} = \inf\{z \in (a, b) : \hat{I}_t(x, z) = 0\}$ .*

Next we see the following.

LEMMA 3.17. *Let  $S \leq T$  be nonnegative random variables. Then with probability one,*

- $\hat{Y}$  is nondecreasing on  $(S, T]$  for  $(X^x, \hat{Y}) \in (\mathcal{S}_0)^c$  on  $(S, T)$ ;
- $\hat{Y}$  is nonincreasing on  $(S, T]$  for  $(X^x, \hat{Y}) \in (\mathcal{S}_1)^c$  on  $(S, T)$ ;
- $\hat{Y}$  is constant on  $(S, T]$  for  $(X^x, \hat{Y}) \in \mathcal{C}$  on  $(S, T)$ .

Consequently, with probability one,  $(X_t^x, \hat{Y}_t) \in \mathcal{C}$  for all  $t > 0$  and  $d\hat{Y}_t$  is supported on  $\partial\mathcal{C}$ .

*Proof.* We shall prove only the first claim. (The second one follows by a similar argument, and the last one is immediate from the definition of  $\mathcal{C}$  and the first two.) Take any  $x > 0$ . If  $(X^x, \hat{Y}) \in (\mathcal{S}_0)^c$  on  $(S, T)$ , then in light of Lemma 3.16 and the fact that  $h$  has at most countably many discontinuities, clearly it suffices to show that for any  $z \in (a, b)$  such that  $h$  is continuous at  $z$ ,  $\hat{I}_t(x, z)$  is almost surely nondecreasing on  $(S, T]$ .

Given  $z \in (a, b)$  where  $h$  is continuous, fix  $t > 0$  and consider the event that  $t \in (S, T)$  and  $\hat{I}_t(x, z) = 1$ . On this event  $\hat{Y}_t \geq z$  almost surely. Furthermore, for any  $s \in [t, T)$ , we have  $(X_s^x, \hat{Y}_s) \in \mathcal{C}$ , and hence  $X_s^x > F(\hat{Y}_s^-) \geq F(z^-) = F(z)$ , since  $z$  is a continuity point of  $h$ . This implies that there is no switching to regime 0 at level  $z$ , and hence with probability one,  $\hat{I}_s(x, z) = 1$  for all  $s \in [t, T)$ . By the left continuity of  $\hat{I}$ , this implies  $\hat{I}_T(x, z) = 1$  as well. Since  $\hat{I}_t(x, z) \in \{0, 1\}$ , this implies that  $\hat{I}_t(x, z)$  is indeed nondecreasing on  $(S, T]$ .  $\square$

Now, we have the next theorem.

THEOREM 3.18 (optimally controlled process). *The resulting optimal control process  $\hat{Y}_t$  is given by the following.*

Case I (up to indistinguishability). For  $t > 0$ ,

- if  $h(y^+) > 0$ , then  $\hat{Y}_t = \max\{G^\rightarrow(M_t), y\}$ ;
- if  $h(y^+) = 0$  or  $h(y^-) = 0$ , then  $\hat{Y}_t = y$ ;
- if  $h(y^-) < 0$ , then  $\hat{Y}_t = \min\{F^\rightarrow(M_t), y\}$ .

Here  $M_t = \max\{X_s^x : s \in [0, t]\}$ , and  $F^\rightarrow$  and  $G^\rightarrow$  are, respectively, the left continuous inverses of  $F$  (nonincreasing) and  $G$  (nondecreasing) such that

$$\begin{aligned} F^\rightarrow(x) &= \inf\{z \in (a, b) : F(z) < x\} = \sup\{z \in (a, b) : F(z) \geq x\}, \\ G^\rightarrow(x) &= \inf\{z \in (a, b) : G(z) \geq x\} = \sup\{z \in (a, b) : G(z) < x\}, \end{aligned}$$

with  $\inf \emptyset = b$  and  $\sup \emptyset = a$ .

Case II (up to indistinguishability). For  $t > 0$ ,

$$(27) \quad \hat{Y}_t = \begin{cases} G^\rightarrow(M_t^0) \vee y & \text{on } \{t \leq S_1\}, \\ F^\leftarrow(m_t^n) \wedge \hat{Y}_{S_n} & \text{on } \{S_n < t \leq T_n\}, \\ G^\rightarrow(M_t^n) \vee \hat{Y}_{T_n} & \text{on } \{T_n < t \leq S_{n+1}\}, \end{cases}$$

and  $\lim_{n \rightarrow \infty} S_n = \infty = \lim_{n \rightarrow \infty} T_n$  almost surely. Here  $F^\leftarrow(x)$  and  $G^\rightarrow(x)$  are, respectively, the right continuous inverse of  $F$  and the left continuous inverse of  $G$ , both of which are nondecreasing and are given by

$$\begin{aligned} F^\leftarrow(x) &= \inf\{z \in (a, b) : F(z) > x\} = \sup\{z \in (a, b) : F(z) \leq x\}, \\ G^\rightarrow(x) &= \inf\{z \in (a, b) : G(z) \geq x\} = \sup\{z \in (a, b) : G(z) < x\}, \end{aligned}$$

with  $\inf \emptyset = b$  and  $\sup \emptyset = a$ . Moreover, the stopping times  $(S_n)$  and  $(T_n)$  are given by

$$\begin{aligned} S_1 &= \inf\{t > 0 : (X_t^x, \hat{Y}_t) \in \mathcal{S}_0\}, & T_1 &= \inf\{t > S_1 : (X_t^x, \hat{Y}_t) \in \mathcal{S}_1\}, \\ S_n &= \inf\{t > T_{n-1} : (X_t^x, \hat{Y}_t) \in \mathcal{S}_0\}, & T_n &= \inf\{t > S_n : (X_t^x, \hat{Y}_t) \in \mathcal{S}_1\}. \end{aligned}$$

Lastly, the processes  $M_t^n$ ,  $m_t^n$  are defined by  $M_t^0 = \max\{X_t^x : 0 \leq s \leq t\}$  and

$$m_t^n = \min\{X_t^x : S_n \leq s \leq t\} 1_{\{S_n \leq t\}}, \quad M_t^n = \max\{X_t^x : T_n \leq s \leq t\} 1_{\{T_n \leq t\}}.$$

*Proof of Theorem 3.18.*

Case I. Recall the optimal switching controls for Case I. Suppose  $0 < h(y^+)$ ; then  $0 < h(y)$  since  $h$  is nonincreasing, and thus  $F(z) = \infty$  and  $G(z) < \infty$ . Let  $t > 0$  be fixed and observe that  $\hat{I}_t(z) \equiv 1$  for all  $z \leq y$  and for  $z > y$ ,  $\hat{I}_t(z) = 1_{\{\hat{\tau}_1(z) < t\}}$ . So  $\hat{I}_t(z) = 1$  if and only if  $z \leq y$  or  $t > \hat{\tau}_1(z)$ . Almost surely,  $t > \hat{\tau}_1(z)$  is equivalent to  $M_t > G(z)$ . Hence

$$\begin{aligned} \hat{Y}_t &= \sup\{z \in (a, b) : I_t(z) = 1\} = y \vee \sup\{z \in (a, b) : t > \hat{\tau}_1(z)\} \\ &= y \vee \sup\{z \in (a, b) : G(z) < M_t\} = \max\{G^\rightarrow(M_t), y\}. \end{aligned}$$

Now, since  $M$  is increasing,  $\max\{G^\rightarrow(M_t), y\}$  is also left continuous, and thus they are indistinguishable.

A similar argument proves the result for  $h(y-) < 0$ .

Suppose  $h(y+) = 0$  or  $h(y-) = 0$ . Then for all  $z > y$ , we have  $h(z) \leq 0$ , and hence it is never optimal to switch to regime 1. Since  $\hat{I}_t(0) = 0$ , this is true for all  $t$  and  $\hat{I}_t(z) \equiv 0$ . Similarly, for all  $z \leq y$ ,  $h(z) \leq 0$  and so  $\hat{I}_t(z) \equiv 1$ . Thus  $\hat{Y}_t = y$  for all  $t$ .

Case II. First, we show that  $\lim_{n \rightarrow \infty} S_n = \infty = \lim_{n \rightarrow \infty} T_n$  almost surely. Let  $\tilde{S}_n = \sup\{t < T_n : (X_t^x, \hat{Y}_t) \in \mathcal{S}_0\}$  be the last exit time of the process  $(X^x, \hat{Y})$  from  $\mathcal{S}_0$  before  $T_n$ . Then  $S_n \leq \tilde{S}_n \leq T_n$ , and  $(X_t^x, \hat{Y}_t) \in \mathcal{C}$  on  $(\tilde{S}_n, T_n)$ . By Lemma 3.17,  $\hat{Y}$  is constant on  $(\tilde{S}_n, T_n]$ . Thus, in between  $\tilde{S}_n$  and  $T_n$ , the process  $(X_t^x, \hat{Y}_{T_n})$  must travel between  $\mathcal{S}_0$  and  $\mathcal{S}_1$ . This means that between  $\tilde{S}_n$  and  $T_n$ ,  $\log(X^x)$  must travel between  $\log(F(\hat{Y}_{T_n}^-))$  and  $\log(G(\hat{Y}_{T_n}^+))$ .

Meanwhile, we have

$$\begin{aligned} \log(G(\hat{Y}_{T_n}^+)) - \log(F(\hat{Y}_{T_n}^-)) &= \log(\nu) - \log(h(\hat{Y}_{T_n}^+)) - \log(\kappa) + \log(h(\hat{Y}_{T_n}^-)) \\ &\geq \log(\nu) - \log(\kappa) > 0. \end{aligned}$$

Since this quantity is positive, and independent of  $n$ , and  $\log(X^x)$  is a Brownian motion, there exists a positive random variable  $\epsilon > 0$  such that  $\epsilon \leq T_n - \hat{S}_n \leq T_n - S_n \leq S_{n+1} - S_n$ . Hence  $\lim_{n \rightarrow \infty} S_n = \infty$  almost surely. Since  $T_n \geq S_n$  for all  $n$ ,  $\lim_{n \rightarrow \infty} T_n = \infty$  almost surely as well.

Next, fix  $t > 0$  and note that that almost surely  $t \in (T_n, S_{n+1}]$  or  $t \in (S_n, T_n]$ , for some  $n$ , where  $T_0 = 0$ . We consider the case that  $t \in (T_n, S_{n+1}]$  for some  $n \geq 0$ . The proof for the case  $t \in (S_n, T_n]$  is similar.

Note that  $(X^x, \hat{Y}) \in (\mathcal{S}_0)^c$  on  $(\hat{S}_n, S_{n+1})$ , and hence by Lemma 3.17,  $\hat{I}_s(x, z)$  is nondecreasing on  $[T_n, S_{n+1}] \subset (\hat{S}_n, S_{n+1})$  for all  $z \in (a, b)$  such that  $h$  is continuous at  $z$ .

Thus, in the event that  $t \in (T_n, S_{n+1}]$ , we know that  $\hat{I}_{T_n}(z) = 1$  for all  $z < \hat{Y}_{T_n}$  and  $\hat{I}_{T_n}(z) = 0$  for all  $z > \hat{Y}_t$ . Since  $\hat{I}$  is nondecreasing on  $[T_n, S_{n+1}]$ , this means that  $\hat{I}_t(x, z) = 1$  if and only if  $z < \hat{Y}_{T_n}$  or if  $X_s^x \geq G(z)$  for some  $s \in [T_n, t)$ . The latter condition is almost surely equivalent to  $G(z) < M_t^n$ . Thus, by Lemma 3.16, in the event that  $t \in (T_n, S_{n+1}]$ , we almost surely have

$$\begin{aligned} \hat{Y}_t &= \sup\{z \in (a, b) : I_t(z) = 1\} = \hat{Y}_{T_n} \vee \sup\{z \in (a, b) : G(z) < M_t^n\} \\ &= G^{-}(M_t^n) \vee \hat{Y}_{T_n}. \end{aligned}$$

A similar argument shows that in the event that  $t \in (S_n, T_n]$ , we almost surely have  $\hat{Y}_t = F^{-}(m_t^n) \wedge \hat{Y}_{S_n}$ . Hence, we have proved that for each  $t$ , the statement in (27) holds almost surely. Moreover, since  $M^n$  is increasing and  $G^{-}$  is left continuous,  $G^{-}(M_t^n)$  is left continuous in  $t$ . Similarly, since  $m^n$  is decreasing and  $F^{-}$  is right continuous,  $F^{-}(m_t^n)$  is left continuous in  $t$ . Thus, the right-hand side of (27) is left continuous in  $t$  and hence indistinguishable from  $\hat{Y}$ .  $\square$

**4. Regularity, smooth fit, and region characterization.** In this section, we shall establish necessary and sufficient conditions for the smooth fit principle by exploiting both the structure of the payoff function and the explicit solution of the value function. This analysis leads to the characterization for both the continuation and action regions.

#### 4.1. Regularity and smooth fit.

**THEOREM 4.1** (sufficient conditions).  *$V(x, y)$  is  $C^1$  in  $x$  for all  $(x, y) \in (0, \infty) \times [a, b]$ , and*

$$\frac{\partial}{\partial x} V(x, y) = \eta H(a) + \int_a^y \frac{\partial}{\partial x} v_1(x, z) dz + \int_y^b \frac{\partial}{\partial x} v_0(x, z) dz.$$

*Moreover, if  $H$  is  $C^1$  on an open interval  $\mathcal{J} \subset [a, b]$ , then  $V(x, y)$  is  $C^1$  in  $y$  on  $(0, \infty) \times \mathcal{J}$ ; that is,  $V(x, y)$  is  $C^{1,1}$  on  $(0, \infty) \times \mathcal{J}$ .*

*Proof.* First, by the representation of  $V(x, y)$  in (19), it suffices to check that for a fixed  $y \in [a, b]$ ,  $u'(x) = \int_a^y \frac{\partial}{\partial x} v_1(x, z) dz$  for all  $x > 0$ , where  $u(x) = \int_a^y v_1(x, z) dz$ .

Note that  $\int_a^y |v_1(x, z)| dz < \infty$ , and  $|\frac{\partial}{\partial x} v_1(x, z)|$  is locally bounded by a constant factor of  $h(z)$  by Proposition 3.8. Moreover, for every  $\delta > 0$  such that  $x - \delta > 0$ , there

exists a constant  $C$  such that

$$\int_a^y \int_{-\delta}^{\delta} \left| \frac{\partial}{\partial x} v_1(x + \theta, z) \right| d\theta dz \leq \int_a^y \int_{-\delta}^{\delta} C h(z) d\theta dz = 2\delta C [H(y) - H(a)] < \infty.$$

Hence, by the dominated convergence theorem,  $v$  is continuous, and by [17, Theorem A.9.1],  $u'(x) = \int_a^y \frac{\partial}{\partial x} v_1(x, z) dz$  for all  $x > 0$ .

Furthermore, suppose that  $H(y)$  is  $C^1$  in an open interval  $\mathcal{J} \subset [a, b]$ . Then for  $x > 0$  and  $y \in \mathcal{J}$ ,

$$\begin{aligned} \lim_{z \rightarrow y} \mathbb{E} \left[ \int_0^\infty |e^{-\rho t} h(z)(X_t^x) - e^{-\rho t} h(y) X_t^x| dt \right] &= \lim_{z \rightarrow y} \mathbb{E} \left[ \int_0^\infty e^{-\rho t} X_t^x dt \right] |h(z) - h(y)| \\ &= \eta x \lim_{z \rightarrow y} |h(z) - h(y)| = 0. \end{aligned}$$

So  $v_k(x, \cdot)$  is continuous at  $y$ , and hence  $V(x, y)$  is  $C^1$  in  $y$  for all  $(x, y) \in (0, \infty) \times \mathcal{J}$ . (See also Theorem 3.13 in [21].)  $\square$

To study the necessary conditions for the continuous differentiability of the value function on  $y$ , we start by defining  $d(x, y) = V_{y^+}(x, y) - V_{y^-}(x, y)$ .

First, note that  $v_1(x, \cdot) - v_0(x, \cdot) = \mathbb{E} \left[ \int_0^\infty e^{-\rho t} h(z) X_t^x dt \right]$ . Since  $h(y)$  is nonincreasing and hence  $\mathbb{E} \left[ \int_0^\tau e^{-\rho t} h(z) X_t^x dt \right]$  is nonincreasing in  $z$  for any stopping time  $\tau$ , we have the following.

**LEMMA 4.2.** *For  $x > 0$ ,  $v_1(x, \cdot) - v_0(x, \cdot)$  is decreasing. Therefore,  $d(x, \cdot)$  has only countably many discontinuities.*

This lemma, coupled with the variational inequalities in (15) and (16), leads to the following.

**PROPOSITION 4.3.**  *$V(x, y)$  is both left and right differentiable in  $y$ , with  $V_{y^+}$  and  $V_{y^-}$  decreasing in  $y$  and  $-K_0 \leq V_{y^+}(x, y) \leq V_{y^-}(x, y) \leq K_1$ . Thus,  $d(x, y) \leq 0$ .*

Note that the above results on regularity are based on the general properties of the payoff function  $H$  and on the relation between the value function  $V(x, y)$  of singular control problem (1) with the value functions  $v_k(x, z)$  of the corresponding optimal switching problems.

In the following, we exploit the explicit solutions of  $v_k(x, y)$  to establish further regularity properties of  $V(x, y)$  with respect to  $y$ .

**PROPOSITION 4.4.** *The left and right derivatives  $V_{y^-}(x, y)$  and  $V_{y^+}(x, y)$  are  $C^1$  in  $x$ . That is,  $d(x, y)$  is  $C^1$  in  $x$ .*

*Proof.* We provide the proof for  $V_{y^+}(x, y)$  in Case II with  $h(y^+) > 0$ , and other cases can be verified by similar arguments. Clearly, it suffices to verify that  $V_{y^+}(x, y)$  is continuous and differentiable (with zero derivative) at  $x = F(y^+)$  and  $x = G(y^+)$ .

In Case II,  $F$  and  $G$  are nondecreasing, and so taking limits of the difference between  $v_0$  and  $v_1$  in (21) and (20) gives

$$(28) \quad V_{y^+}(x, y) = \begin{cases} -K_0, & x \leq F(y^+), \\ B(y^+)x^m - A(y^+)x^n + \eta h(y^+)x, & F(y^+) < x \leq G(y^+), \\ K_1, & x > G(y^+), \end{cases}$$

$$(29) \quad V_{y^-}(x, y) = \begin{cases} -K_0, & x < F(y^-), \\ B(y^-)x^m - A(y^-)x^n + \eta h(y^-)x, & F(y^-) \leq x < G(y^-), \\ K_1, & x \geq G(y^-). \end{cases}$$

By the continuity of  $v_1$  and  $v_0$  in (20), we have

$$\begin{aligned} \lim_{x \downarrow G(y^+)} V_{y^+}(x, y) &= K_1 \\ \lim_{x \uparrow G(y^+)} V_{y^+}(x, y) &= B(y^+)G(y^+)^m - A(y^+)G(y^+)^n + \eta h(y^+)G(y^+) \\ &= \lim_{z \downarrow y} [B(z)G(z)^m - A(z)G(z)^n + \eta h(z)G(z)] \\ &= \lim_{z \downarrow y} [v_1(G(z), z) - v_0(G(z), z)] = K_1. \end{aligned}$$

Hence  $V_{y^+}(x, y)$  is continuous at  $G(y^+)$ .

Moreover, by the continuous differentiability of  $v_1$  and  $v_0$  in (20) and (21), we have

$$\begin{aligned} \lim_{h \downarrow 0} \frac{V_{y^+}(G(y^+) + h, y) - V_{y^+}(G(y^+), y)}{h} &= \lim_{h \downarrow 0} \frac{K_1 - K_1}{h} = 0 \\ \lim_{h \uparrow 0} \frac{V_{y^+}(G(y^+) + h, y) - V_{y^+}(G(y^+), y)}{h} &= mB(y^+)G(y^+)^{m-1} - nA(y^+)G(y^+)^{n-1} + \eta h(y^+) \\ &= \lim_{z \downarrow y} [mB(z)G(z)^{m-1} - nA(z)G(z)^{n-1} + \eta h(z)] \\ &= \lim_{z \downarrow y} \frac{\partial}{\partial x} [v_1(G(z), z) - v_0(G(z), z)] = 0. \end{aligned}$$

Hence  $V_{y^+}(x, y)$  is  $C^1$  at  $G(y^+)$  (and similarly at  $F(y^+)$ ).  $\square$

**THEOREM 4.5** (necessary and sufficient conditions for smooth fit).  *$V(x, y)$  is continuously differentiable in  $x$  for all  $(x, y) \in (0, \infty) \times [a, b]$ .  $V(x, y)$  is differentiable in  $y$  at the point  $(x, y)$  if and only if*

$$(x, y) \in \{(x, y) \in (0, \infty) \times (a, b) : H \text{ is differentiable at } y\} \cup \mathcal{S}_0 \cup \mathcal{S}_1,$$

where  $\mathcal{S}_0$  and  $\mathcal{S}_1$  are given as in (30). Alternatively, it is not differentiable in  $y$  at the point  $(x, y)$  if and only if

$$(x, y) \in \{(x, y) \in (0, \infty) \times (a, b) : H \text{ is not differentiable at } y\} \cap \mathcal{C}.$$

This theorem follows naturally from the following lemma and the proposition.

**LEMMA 4.6.** *If  $h$  is continuous at  $y$ , then for all  $x > 0$ ,  $d(x, y) = 0$ .*

**PROPOSITION 4.7.** *If  $h$  is not continuous at  $y$ , then in Case I,  $d(x, y) = 0$  for  $x \geq \min\{F(y^-), G(y^+)\}$  and  $d(x, y) < 0$  for  $x < \min\{F(y^-), G(y^+)\}$ . In Case II,  $d(x, y) = 0$  for  $x \leq F(y^-)$  and  $x \geq G(y^+)$  and  $d(x, y) < 0$  for  $x \in (F(y^-), G(y^+))$ .*

*Proof of Proposition 4.7.* Suppose that there exists  $y \in (a, b)$  where  $h$  is not continuous. We shall prove the result in Case II when  $h(y^+) > 0$ , and other cases can be verified by similar arguments.

First, since  $h$  is nonincreasing,  $\lim_{z \downarrow y} h(z) < \lim_{z \uparrow y} h(z)$ . This also implies that the switching boundaries  $F(z) = \kappa h(z)^{-1}$  and  $G(z) = \nu h(z)^{-1}$  are discontinuous at  $y$ . Clearly, by (28) and (29),  $d(x, y) = 0$  for  $x < F(y^-)$  and for  $x > G(y^+)$ . By the continuity of  $d$ , this is also true of  $x = F(y^-)$  and  $x = G(y^+)$ .



Next, without loss of generality, assume that  $h$  and hence  $G$  is right continuous. Then, pick  $x$  such that  $G(z^-) \leq x < G(y) = G(y^+)$ . Since  $x < G(y)$ , then  $v_1(x, y) - v_0(x, y) < K_1$  from the HJB equations (15) and (16). Furthermore, by Lemma 4.2,  $v_1(x, z) - v_0(x, z) \leq v_1(x, y) - v_0(x, y) < K_1$  for all  $z > y$ . Hence

$$\begin{aligned} V_{y+}(x, y) &= \lim_{z \downarrow y} v_1(x, z) - v_0(x, z) \leq v_1(x, y) - v_0(x, y) < K_1, \text{ and} \\ V_{y-}(x, y) &= \lim_{z \uparrow y} v_1(x, z) - v_0(x, z) = K_1, \end{aligned}$$

where the last equality follows from the fact that  $x \geq G(z)$  for all  $z < y$ . Thus,  $d(x, y) < 0$  for all  $x \in [G(y^-), G(y^+)]$ . A similar argument proves that in addition to the above,  $d(x, y) < 0$  for all  $x \in (F(y^-), F(y^+)]$ .

Finally, let  $x_0 \in (F(y^+), G(y^-))$  be given. We know that  $d(F(y^-), y) = 0 = d(G(y^+), y)$ ,  $d(x, y) \leq 0$  and that  $d$  is  $C^1$  in  $x$ . Suppose  $d(x_0, y) = 0$ , implying that  $x_0$  is a local maximum, and hence  $d_x(x_0, y) = 0$ . Furthermore, by the mean value theorem, there must be two points  $x_1 \in (F(y^-), x_0)$  and  $x_2 \in (x_0, G(y^+))$  such that  $d_x(x_1, y) = 0 = d_x(x_2, y)$ . In fact, since  $d(x, y) < 0$  for all  $x \in (F(y^-), F(y^+)]$  and  $x \in [G(y^-), G(y^+))$ , we must have  $x_1 \in (F(y^+), x_0)$  and  $x_2 \in (x_0, G(y^-))$ .

Let  $f(x) = x^{-(m-1)}d_x(x, y)$ . Since  $x_0, x_1, x_2 > 0$  and  $0 = d_x(x_0, y) = d_x(x_1, y) = d_x(x_2, y)$ , we must also have  $0 = f(x_0) = f(x_1) = f(x_2)$ . However, by (28) and (29), for  $x \in (F(y^+), G(y^-))$ ,  $d(x, y) = \Delta B(y)x^m - \Delta A(y)x^n + \eta\Delta h(y)x$ , with  $\Delta B(y) = B(y^+) - B(y^-)$ ,  $\Delta A(y) = A(y^+) - A(y^-)$ , and  $\Delta h(y) = h(y^+) - h(y^-)$ . So by differentiating, we have that for  $x \in (F(y^+), G(y^-))$ ,  $f(x) = x^{-(m-1)}d_x(x, y) = m\Delta B(y) - n\Delta A(y)x^{n-m} + \eta\Delta h(y)x^{1-m}$ .

Now,  $f$  is  $C^1$  on  $(F(y^+), G(y^-))$ ; hence by the mean value theorem again, there must be two points,  $\hat{x}_1 \in (x_1, x_0) \subset (F(y^+), G(y^-))$  and  $\hat{x}_2 \in (x_0, x_2) \subset (F(y^+), G(y^-))$  such that  $f_x(\hat{x}_1) = 0 = f_x(\hat{x}_2)$ . Thus  $f_x$  must have at least two positive roots. Differentiating again, we have, for  $x \in (F(y^+), G(y^-))$ ,

$$\begin{aligned} f_x(x) &= -n(n-m)\Delta A(y)x^{n-m-1} + (1-m)\eta\Delta h(y)x^{-m} \\ &= x^{-m}((1-m)\eta\Delta h(y) - n(n-m)\Delta A(y)x^{n-1}). \end{aligned}$$

Thus,  $f_x(x)$  can have at most one positive root, a contradiction. Thus  $d(x_0, y) < 0$ . Since  $x_0 \in (F(y^+), G(y^-))$  was arbitrary,  $d(x, y) < 0$  for all  $x \in (F(y^+), G(y^-))$ .  $\square$

Finally, we can explicitly compute  $V_{xy}$  and  $V_{yx}$  from the derivatives of  $v_k(x, y)$ .

**THEOREM 4.8.** *If  $V_y(x, \hat{y})$  exists in a neighborhood of  $\hat{x}$ , then  $V_{xy}$  and  $V_{yx}$  exist at  $(\hat{x}, \hat{y})$ , with  $V_{xy}(\hat{x}, \hat{y}) = V_{yx}(\hat{x}, \hat{y}) = \frac{\partial}{\partial x}[v_1(\hat{x}, \hat{y}) - v_0(\hat{x}, \hat{y})]$ .*

*Proof.* The existence of  $V_{yx}$  at  $(\hat{x}, \hat{y})$  is clear with  $V_{yx}(\hat{x}, \hat{y}) = \frac{\partial}{\partial x}[v_1(\hat{x}, \hat{y}) - v_0(\hat{x}, \hat{y})]$ . Moreover, by Theorem 4.5, the existence of  $V_y(\hat{x}, y)$  for all  $y$  in a neighborhood of  $\hat{y}$  means that either  $\hat{y}$  is a continuity point of  $h$ , or  $(\hat{x}, \hat{y})$  is in the interior of  $S_0 \cup S_1$ .

If  $\hat{y}$  is a continuity point of  $h$ , by the representation of  $V_x$  in Theorem 4.1, it is sufficient to show that  $u_1(y) := \frac{\partial}{\partial x}v_1(x, y)$  and  $u_0(y) := \frac{\partial}{\partial x}v_0(x, y)$  are continuous at  $\hat{y}$ .

We prove that  $u_0(y)$  is continuous at  $\hat{y}$  for Case II. (Similar arguments apply to other cases.) In this case,  $v_0$  is  $C^1$  in  $x$ , and from (20),

$$u_0(y) = \frac{\partial}{\partial x}v_0(x, z) = \begin{cases} nA(z)x^{n-1}, & x < G(z), \\ mB(z)x^{m-1} + \eta h(z), & x \geq G(z), \end{cases}$$

where  $nA(z)G(z)^{n-1} = mB(z)G(z)^{m-1} + \eta h(z)$ . Since  $h$  is continuous at  $\hat{y}$ , the continuity of  $A$ ,  $B$ , and  $G$  follows by their representation in Theorem 3.14, and hence the continuity of  $u_0(y)$  at  $\hat{y}$  follows from its expression.

If  $(\hat{x}, \hat{y})$  is in the interior of  $\mathcal{S}_1$ , then the explicit forms in Theorem 3.14 imply that for all  $(x, y)$  in a neighborhood of  $(\hat{x}, \hat{y})$ , we have  $\frac{\partial}{\partial x}v_0(x, y) = \frac{\partial}{\partial x}v_1(x, y)$ , and the limits in  $y$  from both the left and the right exist. Thus, by the representation in Theorem 4.1, the left and right derivatives of  $V_x(\hat{x}, \hat{y})$  exist and are given by

$$\begin{aligned} V_{xy^+}(\hat{x}, \hat{y}) &= \lim_{y \downarrow \hat{y}} \frac{\partial}{\partial x}v_1(\hat{x}, y) - \lim_{y \downarrow \hat{y}} \frac{\partial}{\partial x}v_0(\hat{x}, y) = \lim_{y \downarrow \hat{y}} \left( \frac{\partial}{\partial x}v_1(\hat{x}, y) - \frac{\partial}{\partial x}v_0(\hat{x}, y) \right) = 0, \\ V_{xy^-}(\hat{x}, \hat{y}) &= \lim_{y \uparrow \hat{y}} \frac{\partial}{\partial x}v_1(\hat{x}, y) - \lim_{y \uparrow \hat{y}} \frac{\partial}{\partial x}v_0(\hat{x}, y) = \lim_{y \uparrow \hat{y}} \left( \frac{\partial}{\partial x}v_1(\hat{x}, y) - \frac{\partial}{\partial x}v_0(\hat{x}, y) \right) = 0. \end{aligned}$$

Thus,  $V_{xy}$  exists, since the left and right derivatives are equal. Furthermore, it is easy to verify that for  $(\hat{x}, \hat{y})$  in the interior of  $\mathcal{S}_1$ ,  $V_{yx}(\hat{x}, \hat{y}) = \frac{\partial}{\partial x}[v_1(\hat{x}, \hat{y}) - v_0(\hat{x}, \hat{y})] = 0$ . A similar argument applies to  $(\hat{x}, \hat{y})$  in the interior of  $\mathcal{S}_0$ , thereby proving the claim.  $\square$

**COROLLARY 4.9.** *If  $H$  is  $C^1$  on an open interval  $\mathcal{J} \subset [a, b]$ , then  $V_{yx}$  and  $V_{xy}$  exist and are continuous with  $V_{xy}(x, y) = V_{yx}(x, y) = \frac{\partial}{\partial x}[v_1(x, y) - v_0(x, y)]$  on  $(0, \infty) \times \mathcal{J}$ .*

#### 4.2. Region characterization.

**THEOREM 4.10** (region characterization). *Under the optimal singular control  $(\hat{\xi}^+, \hat{\xi}^-) \in \mathcal{A}'_y$ , define the corresponding investment  $(\mathcal{S}_1)$ , disinvestment  $(\mathcal{S}_0)$ , and continuation  $(\mathcal{C})$  regions by*

$$(30) \quad \begin{cases} \mathcal{S}_0 &:= \begin{cases} \{(x, z) \in (0, \infty) \times [a, b] : x \geq \lim_{w \uparrow z} F(w)\} & \text{if } K_0 \geq 0 \text{ (Case I),} \\ \{(x, z) \in (0, \infty) \times [a, b] : x \leq \lim_{w \uparrow z} F(w)\} & \text{if } K_0 < 0 \text{ (Case II),} \end{cases} \\ \mathcal{S}_1 &:= \{(x, z) \in (0, \infty) \times [a, b] : x \geq \lim_{w \downarrow z} G(w)\}, \\ \mathcal{C} &:= (0, \infty) \times [a, b] \setminus (\mathcal{S}_0 \cup \mathcal{S}_1). \end{cases}$$

*Then, the action and continuation regions can be characterized as*

$$(31) \quad \begin{cases} \mathcal{S}_0 &= \{(x, y) \in (0, \infty) \times [a, b] : V_y(x, y) = -K_0\}, \\ \mathcal{S}_1 &= \{(x, y) \in (0, \infty) \times [a, b] : V_y(x, y) = K_1\}, \\ \mathcal{C} &= \{(x, y) \in (0, \infty) \times [a, b] : V_{y^-}(x, y) > -K_0, V_{y^+}(x, y) < K_1\}. \end{cases}$$

*Proof of Theorem 4.10.* Recall that  $V_{y^-}$  and  $V_{y^+}$  exist by Proposition 4.3 and that  $-K_0 \leq V_{y^+} \leq V_{y^-} \leq K_1$ .

Thus,  $V_y(x, y) = -K_0$  if and only if  $V_{y^-}(x, y) = -K_0$ , and from the expression for  $V_{y^-}$  in (29), we have that  $V_{y^-}(x, y) = -K_0$  for  $x < F(y^-)$  (in Case II). However, by the continuity of  $V_{y^-}$  in Proposition 4.4, we get  $V_{y^-}(x, y) = -K_0$  if and only if  $x \leq F(y^-)$ , which is true if and only if  $(x, y) \in \mathcal{S}_0$  by (30). Thus,  $V_y(x, y) = -K_0$  if and only if  $(x, y) \in \mathcal{S}_0$ .

The same argument applied to  $V_{y^+}(x, y) = K_1$  shows that  $V_y(x, y) = K_1$  if and only if  $(x, y) \in \mathcal{S}_1$ . Lastly, the claim for  $\mathcal{C}$  follows since it is the complement of  $\mathcal{S}_0 \cup \mathcal{S}_1$ .

A similar argument also applies in Case I.

**5. Examples and discussions.** By now, it is clear from our analysis that without sufficient smoothness of the payoff function, the value function may be non-differentiable and the boundaries may be nonsmooth or not strictly monotonic. Moreover, when the payoff function  $H$  is not continuously differentiable, the interior of  $\mathcal{C}$

may not be simply connected. Note, however, that the regions  $\mathcal{S}_0$ ,  $\mathcal{S}_1$ , and  $\mathcal{C}$  are mutually disjoint and simply connected by the monotonicity of  $F$  and  $G$ .

We elaborate on these points with some concrete examples.

**5.1. Examples.** Taking parameters  $\kappa, \nu, h$  as defined in the main results in section 3, we fix here  $[a, b] = [0, 2]$ ,  $K_0 < 0$ , and  $0 < \beta < 1$ . Recall that since  $K_0 < 0$ ,  $F(z) = \kappa h(z)^{-1}$  and  $G(z) = \nu h(z)^{-1}$ .

*Example 5.1* (the boundaries are  $C^1$  but NOT strictly increasing because  $H$  is not strictly concave).

$$H(z) = \begin{cases} z, & z \leq 1, \\ \arctan(z-1) + 1, & z > 1. \end{cases}$$

$$h(z) = \begin{cases} 1, & z \leq 1, \\ \frac{1}{1+(z-1)^2}, & z > 1. \end{cases}$$

See Figure 5.

*Example 5.2* ( $F, G$  are only  $C^0$  because  $H$  is not  $C^2$ ).

$$H(z) = \begin{cases} z, & z \leq 1, \\ \frac{z^\beta - 1}{\beta} + 1, & z > 1, \end{cases}$$

$$h(z) = \begin{cases} 1, & z \leq 1, \\ z^{\beta-1}, & z > 1. \end{cases}$$

See Figure 6.

*Example 5.3* (the value function is NOT  $C^{1,1}$ ;  $F, G$  are NOT continuous because  $H$  is not  $C^1$ ).

$$H(z) = \begin{cases} z, & z \leq 1, \\ \frac{2\kappa}{(\kappa+\nu)} \frac{z^\beta - 1}{\beta} + 1, & z > 1, \end{cases}$$

$$h(z) = \begin{cases} 1, & z \leq 1, \\ \frac{2\kappa}{\kappa+\nu} z^{\beta-1}, & z > 1. \end{cases}$$

See Figure 7.

*Example 5.4* (interior of continuation region NOT connected).

$$H(z) = \begin{cases} z, & z \leq 1, \\ \frac{\kappa}{2\nu} \frac{z^\beta - 1}{\beta} + 1, & z > 1, \end{cases}$$

$$h(z) = \begin{cases} 1, & z \leq 1, \\ \frac{\kappa}{2\nu} z^{\beta-1}, & z > 1. \end{cases}$$

See Figure 8.

Note that Examples 5.2–5.4 all have payoff functions of the form

$$H(z) = \begin{cases} z, & z \leq 1, \\ \phi \frac{z^\beta - 1}{\beta} + 1, & z > 1, \end{cases}$$

for some constant  $\phi$ . To ensure the concavity of  $H$ , we must have  $\phi \in [0, 1]$ . When  $\phi = 1$ , we recover Example 5.1 and Figure 6. For  $\frac{\kappa}{\nu} < \phi < 1$ , the regions are described by Figure 7. Lastly, for  $0 < \phi \leq \frac{\kappa}{\nu}$ , the interior of the continuation region is not connected, as in Figure 8.

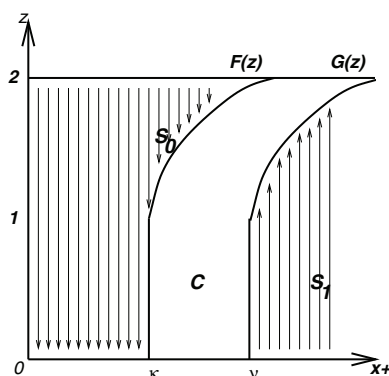


FIG. 5.  $F, G$  are  $C^1$  but NOT strictly increasing because  $H$  is NOT strictly concave.

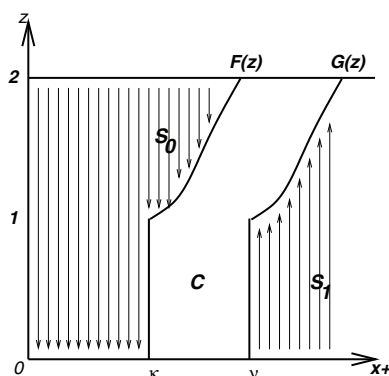


FIG. 6.  $F, G$  are only  $C^0$  because  $H$  is not  $C^2$ .

**5.2. Discussion.** The above examples demonstrate how the regularity assumptions typically assumed by the traditional HJB approach may fail.

First, according to that approach, the value function  $V(x, y)$  would satisfy some (quasi-)variational inequalities so that

$$\max\{\sigma^2 x^2 V_{xx}(x, y) + bxV_x(x, y) - rV(x, y) + H(y)x, V_y(x, y) - K_1, -V_y(x, y) - K_0\} = 0.$$

In general, while searching for a solution, one would assume a priori smoothness for the value function and the boundary. For example, in [2] and [31],  $V$  is derived from the class of  $C^{2,1}$ . However, Example 5.3 shows that although the HJB variational inequality may still hold, one should search for a solution in a larger class of functions, such as in  $C^{1,0}$ .

Furthermore, Example 5.3 shows that in general, one may not have the smoothness of the boundary, as the boundaries  $F$  and  $G$  are not necessarily continuous or not even strictly increasing. Indeed, in this example,  $F$  and  $G$  are inversely proportional to  $h$ , which may be neither.

Finally, we compare our results and method with those in [2].

*Example 5.5* (general case of [2] for geometric Brownian motion). Let  $x > 0$  and  $y \in [a, b]$ , with  $K_0 < 0$  and  $h > 0$  on  $[a, b]$ . Then  $F$  and  $G$  are nondecreasing and

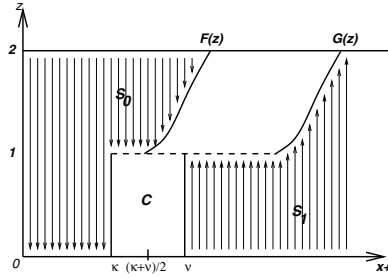
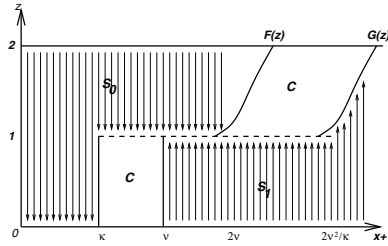
FIG. 7. Value function NOT  $C^{1,1}$ ;  $F, G$  NOT continuous because  $H$  is not  $C^1$ .

FIG. 8. Interior of continuation region NOT connected.

given by (24). Define

$$y_0(x) = G^{\leftarrow}(x) \wedge b = \sup\{z : G(z) \leq x\} = \sup\{z : h(z) \leq (x/\nu)^{-1}\} \wedge b,$$

$$y_1(x) = F^{\rightarrow}(x) \vee a = \inf\{z : F(z) \geq x\} = \inf\{z : h(z) \geq (x/\kappa)^{-1}\} \vee a.$$

Then  $y_0(x) \leq y_1(x)$ , and

- $x \leq F(z)$  for  $z > y_1(x)$ ;
- $F(z) < x < G(z)$  for  $y_0(x) < z < y_1(x)$ ;
- $G(z) \leq x$  for  $z < y_0(x)$ .

When, in addition,  $H$  satisfies the Inada conditions, this example generalizes those in [2] when  $X$  is a geometric Brownian motion. Compared to the very special form appearing in [2], our results show that, in order to compute the value function, integration of  $v_k(x, z)$  is necessary, which we reduce to three possible cases as follows, depending on whether  $(x, y)$  is in  $\mathcal{S}_0$ ,  $\mathcal{S}_1$ , or  $\mathcal{C}$ .

1.  $(x, y) \in \mathcal{S}_0$ : Then  $y_1 \leq y$  and

$$V(x, y) = \eta H(y_1)x + x^m \int_a^{y_1} B(z)dz + x^n \int_{y_1}^b A(z)dz - K_0(y - y_1).$$

2.  $(x, y) \in \mathcal{C}$ : Then  $y_0 < y < y_1$  and

$$V(x, y) = \eta H(y)x + x^m \int_a^y B(z)dz + x^n \int_y^b A(z)dz.$$

3.  $(x, y) \in \mathcal{S}_1$ : Then  $y \leq y_0$  and

$$V(x, y) = \eta H(y_0)x + x^m \int_a^{y_0} B(z)dz + x^n \int_{y_0}^b A(z)dz - K_1(y_0 - y),$$

where  $A$  and  $B$  are as given by (22)–(23).

**Acknowledgments.** The authors thank the associate editor and the two anonymous referees for their constructive and detailed remarks, which led to a substantial improvement of the paper.

## REFERENCES

- [1] A. B. ABEL AND J. C. EBERLY, *An exact solution for the investment and value of a firm facing uncertainty, adjustment costs, and irreversibility*, J. Econom. Dynam. Control, 21 (1997), pp. 831–852.
- [2] L. H. ALVAREZ, *A General Theory of Optimal Capacity Accumulation under Price Uncertainty and Costly Reversibility*, Working Paper, Helsinki Center of Economic Research, Helsinki, Finland, 2006.
- [3] L. H. R. ALVAREZ, *Singular stochastic control in the presence of a state-dependent yield structure*, Stochastic Process. Appl., 86 (2000), pp. 323–343.
- [4] L. H. R. ALVAREZ, *Singular stochastic control, linear diffusions, and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.
- [5] D. ASSAF, *Estimating the state of a noisy continuous time Markov chain when dynamic sampling is feasible*, Ann. Appl. Probab., 7 (1997), pp. 822–836.
- [6] B. ATA, J. M. HARRISON, AND L. A. SHEPP, *Drift rate control of a Brownian processing system*, Ann. Appl. Probab., 15 (2005), pp. 1145–1160.
- [7] F. M. BALDURSSON AND I. KARATZAS, *Irreversible investment and industry equilibrium*, Finance Stoch., 1 (1997), pp. 69–89.
- [8] P. BANK, *Optimal control under a dynamic fuel constraint*, SIAM J. Control Optim., 44 (2005), pp. 1529–1541.
- [9] F. BOETIUS, *Bounded variation singular stochastic control and Dynkin game*, SIAM J. Control Optim., 44 (2005), pp. 1289–1321.
- [10] F. BOETIUS AND M. KOHLMANN, *Connections between optimal stopping and singular stochastic control*, Stochastic Process. Appl., 77 (1998), pp. 253–281.
- [11] K. A. BREKKE AND B. ØKSENDAL, *Optimal switching in an economic activity under uncertainty*, SIAM J. Control Optim., 32 (1994), pp. 1021–1036.
- [12] M. B. CHIAROLLA AND U. G. HAUSSMANN, *Explicit solution of a stochastic, irreversible investment problem and its moving threshold*, Math. Oper. Res., 30 (2005), pp. 91–108.
- [13] M. H. A. DAVIS, M. A. H. DEMPSTER, S. P. SETHI, AND D. VERMES, *Optimal capacity expansion under uncertainty*, Adv. in Appl. Probab., 19 (1987), pp. 156–176.
- [14] M. H. A. DAVIS AND M. ZERVOS, *A problem of singular stochastic control with discretionary stopping*, Ann. Appl. Probab., 4 (1994), pp. 226–240.
- [15] M. H. A. DAVIS AND M. ZERVOS, *A pair of explicitly solvable singular stochastic control problems*, Appl. Math. Optim., 38 (1998), pp. 327–352.
- [16] A. K. DIXIT AND R. S. PINDYCK, *Investment under Uncertainty*, Princeton University Press, Princeton, NJ, 1994.
- [17] R. DURRETT, *Probability: Theory and Examples*, 2nd ed., Duxbury Press, Belmont, CA, 1996.
- [18] N. EL KAROUI AND I. KARATZAS, *Probabilistic aspects of finite-fuel, reflected follower problems*, Acta Appl. Math., 11 (1988), pp. 223–258.
- [19] N. EL KAROUI AND I. KARATZAS, *Integration of the optimal risk in a stopping problem with absorption*, in Séminaire de Probabilités, XXIII, Lecture Notes in Math. 1372, Springer, Berlin, 1989, pp. 405–420.
- [20] X. GUO AND H. PHAM, *Optimal partially reversible investment with entry decision and general production function*, Stochastic Process. Appl., 115 (2005), pp. 705–736.
- [21] X. GUO AND P. TOMECEK, *Connections between singular control and optimal switching*, SIAM J. Control Optim., 47 (2008), pp. 421–443.
- [22] J. M. HARRISON AND M. I. TAKSAR, *Instantaneous control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 439–453.
- [23] J. M. HARRISON AND J. A. VAN MIEGHEM, *Dynamic control of Brownian networks: State space collapse and equivalent workload formulations*, Ann. Appl. Probab., 7 (1997), pp. 747–771.
- [24] I. KARATZAS, *Probabilistic aspects of finite-fuel stochastic control*, Proc. Nat. Acad. Sci. U.S.A., 82 (1985), pp. 5579–5581.
- [25] I. KARATZAS AND S. E. SHREVE, *Connections between optimal stopping and singular stochastic control. II. Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–451.
- [26] I. KARATZAS AND S. E. SHREVE, *Brownian Motion and Stochastic Calculus*, Grad. Texts Math. 113, Springer-Verlag, New York, 1988.

- [27] T. Ø. KOBILA, *A class of solvable stochastic investment problems involving singular controls*, Stochastics Stochastics Rep., 43 (1993), pp. 29–63.
- [28] V. LY VATH AND H. PHAM, *Explicit solution to an optimal switching problem in the two-regime case*, SIAM J. Control Optim., 46 (2007), pp. 395–426.
- [29] J. MA, *On the principle of smooth fit for a class of singular stochastic control problems for diffusions*, SIAM J. Control Optim., 30 (1992), pp. 975–999.
- [30] L. F. MARTINS, S. E. SHREVE, AND H. M. SONER, *Heavy traffic convergence of a controlled, multiclass queueing system*, SIAM J. Control Optim., 34 (1996), pp. 2133–2171.
- [31] A. MERHI AND M. ZERVOS, *A model for reversible investment capacity expansion*, SIAM J. Control Optim., 46 (2007), pp. 839–876.
- [32] A. ØKSENDAL, *Irreversible investment problems*, Finance Stoch., 4 (2000), pp. 223–250.
- [33] H. PHAM, *On the smooth-fit property for one-dimensional optimal switching problem*. in Séminaire de Probabilités XL, Lecture Notes in Math. 1899, Springer, Berlin, 2007, pp. 187–199.
- [34] J. A. SCHEINKMAN AND T. ZARIPHPOULOU, *Optimal environmental management in the presence of irreversibilities. Intertemporal equilibrium theory: Indeterminacy, bifurcations, and stability*, J. Econom. Theory, 96 (2001), pp. 180–207.
- [35] H. WANG, *Capacity expansion with exponential jump diffusion processes*, Stoch. Stoch. Rep., 75 (2003), pp. 259–274.

## STABILIZABILITY AND STABILITY ROBUSTNESS OF STATE DERIVATIVE FEEDBACK CONTROLLERS\*

WIM MICHIELS<sup>†</sup>, TOMÁŠ VYHLÍDAL<sup>‡</sup>, HENRI HUIJBERTS<sup>§</sup>, AND HENK NIJMEIJER<sup>¶</sup>

**Abstract.** We study the stabilizability of a linear controllable system using state derivative feedback control. As a special feature the stabilized system may be fragile, in the sense that arbitrarily small modeling and implementation errors may destroy the asymptotic stability. First, we discuss the pole placement problem and illustrate the fragility of stability with examples of a different nature. We also define a notion of stability, called  $p$ -stability, which explicitly takes into account the effect of small modeling and implementation errors. Next, we investigate the effect on the fragility of including a low-pass filter in the control loop. Finally, we completely characterize the stabilizability and  $p$ -stabilizability of linear controllable systems using state derivative feedback. In the stabilizability characterization the odd number limitation, well known in the context of the stabilization of unstable periodic orbits using Pyragas-type time-delayed feedback, plays a crucial role.

**Key words.** stabilizability, robustness, feedback, systems theory, delay

**AMS subject classifications.** 93B35, 93D09, 93D15

**DOI.** 10.1137/070697136

**1. Introduction.** We analyze the stabilizability and stabilization of the linear or linearized system

$$(1) \quad \dot{x}(t) = Ax(t) + Bu(t),$$

where  $x(t) \in \mathbb{R}^n$  is the system state vector and  $u(t) \in \mathbb{R}^m$  is the system input at time  $t$ , using state derivative feedback,

$$(2) \quad u(t) = K_d \dot{x}(t), \quad K_d \in \mathbb{R}^{n \times m}.$$

In this stabilizability study we explicitly take into account the effects of arbitrarily small modeling, approximation, and implementation errors, which may, for instance, be caused by feedback latency, unmodeled sensor and actuator dynamics, and approximations of the derivatives.

From a practical point of view, the motivation for using state derivative feedback (2) instead of conventional state feedback,  $u(t) = K_s x(t)$ , comes from applications

---

\*Received by the editors July 13, 2007; accepted for publication (in revised form) August 12, 2008; published electronically January 7, 2009. This paper presents results of the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture, and of the Center of Excellence on Optimization in Engineering of the K.U.Leuven. Research was supported by the Ministry of Education of the Czech Republic under project 1M0567.

<http://www.siam.org/journals/sicon/47-6/69713.html>

<sup>†</sup>Department of Computer Science, Katholieke Universiteit Leuven, 3001 Heverlee, Belgium (Wim.Michiels@cs.kuleuven.be). This work was partly done while the first author was at the Department of Mechanical Engineering of the Eindhoven University of Technology. He is a postdoctoral fellow of the Fund for Scientific Research, Flanders (Belgium).

<sup>‡</sup>Centre for Applied Cybernetics, Department of Instrumentation and Control, Czech Technical University, Prague, Czech Republic (Tomas.Vyhldal@fs.cvut.cz).

<sup>§</sup>School of Engineering and Materials Science, Queen Mary, University of London, London E1 4NS, United Kingdom (H.J.C.Huijberts@qmul.ac.uk).

<sup>¶</sup>Department of Mechanical Engineering, Eindhoven University of Technology, 5600 HB Eindhoven, The Netherlands (h.nijmeijer@tue.nl).



where accelerometers are used for measuring the system's motion. Typical applications are in vibration control of mechanical systems where the state variables are positions and velocities, while the accelerations, which are the sensed variables, are directly used for feedback; see [2, 3, 1]. In some applications, including vibration suppression, the fact that control law (2) keeps the steady state solutions of the uncontrolled system invariant might be considered as a positive feature.

From a theoretical point of view the stabilizability study of system (1) with control law (2) is a challenging problem if robustness aspects are taken into account. As will be addressed later in the text, the synthesis of the derivative feedback can be accomplished using modified pole placement methods under a controllability assumption. However, the achieved closed-loop dynamics may be *fragile* in the sense that the stability of the controlled system may lack robustness against *arbitrarily small* modeling and implementation errors. In other words, although the nominal system is asymptotically stable, some stability margins may be equal to zero. The main goals of the paper consist of studying this fragility problem and making a *complete* characterization of the stabilizability of (1) with (2) in the presence of any type of small modeling and implementation errors.

Other problems where a fragility of stability has been observed can be found in the literature. More specifically, a lack of robustness of stability against small feedback delays or against small delay changes has been observed for boundary controlled (hyperbolic) partial differential equations; see, e.g., [5, 6, 9, 12, 18, 22]; for feedback controlled descriptor systems see [11], and for neutral functional differential equations see [13, 14]. In [23] a model for gradient play dynamics is discussed where the discretization of a derivative may induce instability. In [15, 17] it was shown that the discretization of distributed delays in control laws, arising in the context of finite spectrum assignment of time-delay systems, may destabilize the system, even if the discretization stepsize is arbitrarily small.

The structure of the paper is as follows: In section 2 the pole placement problem for system (1) using control law (2) is first discussed. Next, a unified framework for studying the effect of small modeling and implementation errors is presented and a practical notion of stability is introduced. In section 3 two examples of a different nature are presented which illustrate that closed-loop stability may not be robust against small modeling and implementation errors. The observations on the instability mechanisms lead us to section 4, where we investigate to what extent the inclusion of a low-pass filter may solve the robustness problems. Finally, in sections 5 and 6 a full characterization of the stabilizability of (1) using (2) is given, with and without the effect of small modeling and implementation errors being taken into account. It will be shown that in the former case an important role is played by a condition that boils down to the so-called odd number limitation, well known in the context of the stabilization of unstable periodic solutions using Pyragas-type time-delayed controllers [21] and in the context of delay difference feedback [10]. Some concluding remarks end the paper.

The following notation and definitions will be adopted:  $\mathbb{C}$  ( $\mathbb{C}^+$ ,  $\mathbb{C}^-$ ) is the set of complex numbers (with strictly positive and strictly negative real parts), and  $j = \sqrt{-1}$ . For  $\lambda \in \mathbb{C}$ ,  $\bar{\lambda}$ ,  $\Re(\lambda)$ , and  $\Im(\lambda)$  define the complex conjugate, the real part, and the imaginary part of  $\lambda$ . For  $\Omega \subset \mathbb{C}$ ,  $\partial\Omega$  denotes the boundary of  $\Omega$  and  $\text{clos}(\Omega)$  its closure.  $\mathbb{R}$  ( $\mathbb{R}^+$ ,  $\mathbb{R}^-$ ) denotes the set of real numbers (larger than or equal to zero, smaller than or equal to zero).  $\mathbb{N}$  is the set of natural numbers and is assumed to include zero.  $\mathbb{Z}$  is the set of integers. For an operator or matrix  $A$ ,  $\sigma(A)$  and  $\rho(A)$

denote its spectrum and spectral radius, respectively. Throughout the paper we make the following assumption.

*Assumption 1.1.* The pair  $(A, B)$  is controllable.

## 2. Prerequisites.

**2.1. Pole placement using state derivative feedback.** Note that the closed-loop system (1)–(2) may be rewritten as

$$(3) \quad (I - BK_d)\dot{x}(t) = Ax(t).$$

Thus, the closed-loop system is well-posed if  $BK_d$  has no eigenvalue equal to one. The characteristic function of the closed-loop system is then given by

$$(4) \quad H_0(\lambda) := \det(\lambda I - A - BK_d\lambda).$$

We first give some preliminary results considering the stability of the closed-loop system.

**PROPOSITION 2.1.** *If  $\det(A) = 0$ , then the closed-loop system (1)–(2) has a zero characteristic root for all values of  $K_d$ .*

Conditions for arbitrary pole placement by state derivative feedback (2) are given in [1] as follows.

**PROPOSITION 2.2.** *If the pair  $(A, B)$  is controllable, then all characteristic roots of the closed-loop system can be assigned at arbitrarily positions in  $\mathbb{C} \setminus \{0\}$  using the control law (2) if and only if  $\det(A) \neq 0$ .*

The following result from [1] establishes relations with pole assignment using state feedback.

**PROPOSITION 2.3.** *Assume that  $\det(A) \neq 0$ . With the control law (2), the closed-loop system has the same characteristic roots as with the state feedback*

$$(5) \quad u(t) = K_s x(t)$$

if

$$K_d = K_s(A + BK_s)^{-1}.$$

Given the fact that algorithms for designing the controller (5) are widely available, this is an important result. However, algorithms have also been developed for a direct design of the state derivative feedback gain  $K_d$ . A general pole placement technique for state derivative feedback was proposed in [1] for single-input delay-free systems and in [2] for multiple-input systems. The same authors proposed a linear quadratic regulator for computing state derivative feedback in [3]. The application of acceleration feedback to vibration suppression problems has been discussed at length in [20] and [7].

*Remark 2.4.* If  $\det(-A) < 0$ , i.e., if the system (1) has an odd number of characteristic roots in  $\mathbb{C}^+$ , stabilization implies that an *odd* number of unstable roots need to be shifted to the left half plane, while a root cannot cross the imaginary axis at zero. This may sound counterintuitive but is always possible: if  $K_d$  is increased from zero to the stabilizing value, then some characteristic roots move to the right half plane via infinity, where for the critical value of  $K_d$  the system is not well-posed. For an example of this, one can consider the system

$$\dot{x}(t) = x(t) + u(t), \quad u(t) = k_d \dot{x}(t), \quad u, x \in \mathbb{R},$$

for which clearly the open-loop system has an odd number of characteristic roots in  $\mathbb{C}^+$ . Further, the closed-loop characteristic root  $\lambda$  satisfies

$$\lambda = \frac{1}{1 - k_d}, \quad k_d \neq 1.$$

It is straightforwardly seen that the closed-loop system is stable for  $k_d > 1$  and that  $\lambda \rightarrow \infty$  if  $k_d \uparrow 1$ .

**2.2. Framework for robustness analysis.** As we shall illustrate in the next section, the use of state derivative feedback may introduce some fragility in the sense that *arbitrarily small* modeling and implementation errors, e.g., arbitrarily small delays in the feedback loop, may change the system's behavior significantly and may even render an asymptotically stable nominal system unstable. To create a unifying framework to study this fragility problem, we assume that the modeling and implementation errors are such that the characteristic function of the actual system is given by

$$(6) \quad H(\lambda; p) := \det(\lambda I - \tilde{A}(p) - \tilde{B}(p) G_1(\lambda; p) K_d \lambda G_2(\lambda; p)),$$

where  $p \in (\mathbb{R}^+)^{n_p}$  denotes some parameters and the functions

$$(7) \quad \begin{aligned} \tilde{A} : (\mathbb{R}^+)^{n_p} &\rightarrow \mathbb{R}^{n \times n}, & p &\mapsto \tilde{A}(p), \\ \tilde{B} : (\mathbb{R}^+)^{n_p} &\rightarrow \mathbb{R}^{n \times m}, & p &\mapsto \tilde{B}(p), \\ G_1 : \mathbb{C} \times (\mathbb{R}^+)^{n_p} &\rightarrow \mathbb{C}^{m \times m}, & (\lambda, p) &\mapsto G_1(\lambda; p), \\ G_2 : \mathbb{C} \times (\mathbb{R}^+)^{n_p} &\rightarrow \mathbb{C}^{n \times n}, & (\lambda, p) &\mapsto G_2(\lambda; p), \end{aligned}$$

satisfy the following assumption.

*Assumption 2.5.*

1. The functions  $p \mapsto \tilde{A}(p)$  and  $p \mapsto \tilde{B}(p)$  are continuous;
2.  $\lim_{p \rightarrow 0} \tilde{A}(p) = A$ ;  $\lim_{p \rightarrow 0} \tilde{B}(p) = B$ ;
3. for every  $p \in (\mathbb{R}^+)^{n_p}$  and  $i = 1, 2$ , the functions  $G_i(\cdot; p)$  are meromorphic; for every  $\lambda \in \mathbb{C}$ , the functions  $G_i(\lambda; \cdot)$  are continuous;
4.  $G_i(\lambda; 0) = I$  for all  $\lambda \in \mathbb{C}$  and  $i = 1, 2$ ;
5. for every compact set  $\Omega \subset \mathbb{C}$ , we have

$$(8) \quad \lim_{p \rightarrow 0} \max_{\lambda \in \Omega} \|G_i(\lambda; p) - I\| = 0, \quad i = 1, 2;$$

6. there exist constants  $M, N, P > 0$  such that for all  $\lambda \in \mathbb{C}$  with  $\Re(\lambda) \geq -N$  and for all  $p \in (\mathbb{R}^+)^{n_p}$  with  $\|p\| \leq P$ ,

$$\|G_i(\lambda; p)\| \leq M, \quad i = 1, 2.$$

In Figure 1 a block diagram is drawn of the controlled system in the presence of modeling and implementation errors. For  $p = 0$ , the function (6) reduces to  $H_0$  in (4). It is clear that  $\tilde{A}(p)$  and  $\tilde{B}(p)$  model uncertainty on  $A$  and  $B$ . To motivate the inclusion of  $G_1$  and  $G_2$  in the control loop we present some examples of a different nature, which all satisfy Assumption 2.5.

- *Feedback delays.* The case where

$$\begin{aligned} n_p &= n + m, \quad p = (\tau_{u_1}, \dots, \tau_{u_n}, \tau_{x_1}, \dots, \tau_{x_n}), \\ G_1(\lambda; p) &= \text{diag}(e^{-\lambda \tau_{u_1}}, \dots, e^{-\lambda \tau_{u_m}}), \quad G_2(\lambda, p) = \text{diag}(e^{-\lambda \tau_{x_1}}, \dots, e^{-\lambda \tau_{x_n}}) \end{aligned}$$

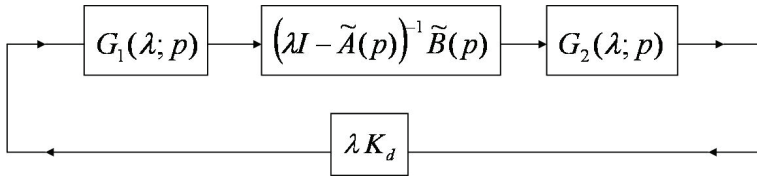


FIG. 1. Feedback interpretation of the controlled system.

corresponds to the case where one assumes a time delay  $\tau_{u_k}$  in the  $k$ th component of the input  $u$  and a time delay  $\tau_{x_l}$  in the measurement of the  $l$ th component of  $\dot{x}$ , where  $1 \leq k \leq m$ ,  $1 \leq l \leq n$ .

- *Numerical computation of derivatives.* If, for example,

$$(9) \quad n_p = 1, \quad G_1(\lambda; p) = I, \quad G_2(\lambda; p) = g_2(\lambda; p)I,$$

with

$$g_2(\lambda; p) = \begin{cases} \frac{1-e^{-\lambda p}}{\lambda p}, & \lambda p \neq 0, \\ 1, & \lambda p = 0, \end{cases}$$

then we get

$$\lambda K_d G_2(\lambda; p) = K_d \frac{1 - e^{-\lambda p}}{p}.$$

Thus, this corresponds to the situation where  $\dot{x}(t)$  in control law (2) is not measured directly, but computed on-line from the measurements of  $x(t)$  using the finite-difference formula

$$(10) \quad \dot{x}(t) \approx \frac{x(t) - x(t-p)}{p}.$$

- *Unmodeled dynamics.*  $G_1$  and  $G_2$  may, for instance, model neglected actuator and sensor dynamics.

From a fragility point of view we are interested in the relationship between the stability properties for  $p = 0$  and those for small  $p \neq 0$ . For this, we first introduce the following practical notion of closed-loop stability.

**DEFINITION 2.6.** *The closed-loop system formed by the feedback interconnection of system (1) with output  $x$  and a controller with transfer function  $-C(\lambda)$  is  $p$ -stable if its null solution is asymptotically stable, and for every set of functions (7) satisfying Assumption 2.5, there is a constant  $\hat{p} > 0$  such that the zeros of*

$$\det(\lambda I - \tilde{A}(p) - \tilde{B}(p)G_1(\lambda; p)C(\lambda)G_2(\lambda; p))$$

are in  $\mathbb{C}^-$  for all  $p \in (\mathbb{R}^+)^{n_p}$  with  $\|p\| < \hat{p}$ .

In this way, requiring that (1), stabilized with (2), be robust against small modeling and implementation errors can be rephrased as requiring the  $p$ -stability of (1)–(2). Note that this concept of stability is closely related to the concept of  $w$ -stability defined in [8].

To conclude this section, we note that, for an arbitrary set of functions (7) that satisfy Assumption 2.5, the requirement that the roots of  $H_0$  in (4) be in  $\mathbb{C}^-$ , along

with Assumption 2.5, is in general not sufficient to guarantee that the zeros of  $H$  in (6) are in  $\mathbb{C}^-$  for sufficiently small  $p$ . However, an (approximate) matching of  $n$  zeros always takes place, as shown by the following result.

**PROPOSITION 2.7.** *Assume that  $\det(I - BK_d) \neq 0$  and let  $\mu_i \in \mathbb{C}$ ,  $1 \leq i \leq n$ , be the eigenvalues of  $(I - BK_d)^{-1}A$ . There exist a number  $\hat{p} > 0$  and  $n$  continuous functions*

$$\hat{\lambda}_i : [0, \hat{p}]^{n_p} \rightarrow \mathbb{C}, \quad p \mapsto \hat{\lambda}_i(p), \quad 1 \leq i \leq n,$$

which satisfy  $H(\hat{\lambda}_i(p); p) \equiv 0$  and

$$(11) \quad \lim_{p \rightarrow 0} \hat{\lambda}_i(p) = \mu_i, \quad 1 \leq i \leq n.$$

*Proof.* Let  $\mathcal{B} \subset \mathbb{C}$  be an open disk which contains all zeros of  $H_0$ . The function  $H(\lambda; p)$  uniformly converges on compact subsets of  $\mathbb{C}$  to  $H_0(\lambda)$  as  $p \rightarrow 0$ . This implies the existence of a number  $\hat{p}$  such that

$$\max_{\lambda \in \partial \mathcal{B}} |H_0(\lambda) - H(\lambda; p)| < \min_{\lambda \in \partial \mathcal{B}} |H_0(\lambda)| \quad \text{for all } p \in [0, \hat{p}]^{n_p}.$$

By Rouché's theorem one concludes that  $H(\lambda; p)$  and  $H_0(\lambda)$  both have  $n$  zeros in  $\mathcal{B}$  if  $p \in [0, \hat{p}]^{n_p}$ . The existence of the functions  $\hat{\lambda}_i$ ,  $1 \leq i \leq n$ , follows from the combination of this result with the continuity of the individual zeros of  $H$  with respect to (w.r.t.)  $p$ . The assertion (11) can be shown in a similar way by letting  $\mathcal{B}$  be a disk with arbitrarily small radius centered at a zero of  $H_0$ , and taking the same steps.  $\square$

In what follows, we focus on the behavior of the other zeros which may be introduced by the implementation or approximation.

**3. Sensitivity of stability to arbitrarily small modeling and implementation errors.** With two case studies we first show that even if all zeros of (4) are in  $\mathbb{C}^-$ , then (6) may have zeros in  $\mathbb{C}^+$  for *arbitrarily small* values of  $p$ . The first case study, inspired by [23], concerns a numerical approximation of the derivatives in the control law, and the second concerns the effect of a neglected time delay. In both cases the eigenvalue distribution of  $BK_d$  determines the position of the characteristic roots introduced by the approximation.

**3.1. Approximation of derivatives.** With the approximation (10) of the derivatives in control law (2), or, equivalently, with the transfer functions (9), the characteristic function of the closed-loop system becomes

$$I_1(\lambda; p) := \det \left( \lambda I - A - BK_d \frac{1 - e^{-\lambda p}}{p} \right).$$

We then have the following result.

**PROPOSITION 3.1.** *If  $BK_d$  has at least one eigenvalue which does not belong to  $\text{clos}(S)$ , where*

$$(12) \quad S := \left\{ \lambda \in \mathbb{C} : \Im(\lambda) \in (-\pi, \pi) \text{ and } \Re(\lambda) < \begin{cases} \Im(\lambda) \cotan(\Im(\lambda)), & \Im(\lambda) \in (-\pi, 0) \cup (0, \pi), \\ 1, & \Im(\lambda) = 0 \end{cases} \right\},$$

then there exist numbers  $\hat{p} > 0$ ,  $c > 0$  and a function  $\hat{\lambda} : (0, \hat{p}) \rightarrow \mathbb{C}$ ,  $p \mapsto \hat{\lambda}(p)$ , such that

$$I_1(\hat{\lambda}(p); p) = 0 \quad \text{and} \quad \Re(\hat{\lambda}(p)) > \frac{c}{p}$$

for all  $p \in (0, \hat{p})$ .

If all eigenvalues of  $BK_d$  belong to the set  $S$ , then there exist numbers  $\hat{p} > 0$  and  $c > 0$  such that for all  $0 < p \leq \hat{p}$ , the function  $I_1(\lambda; p)$  has exactly  $n$  zeros in the half plane  $\{\lambda \in \mathbb{C} : \Re(\lambda) > -\frac{c}{p}\}$ .

*Proof.* Let

$$G(\lambda; p) := p^n I_1\left(\frac{\lambda}{p}; p\right) = \det(\lambda I - Ap - BK_d(1 - e^{-\lambda})),$$

and let

$$\tilde{G}(\lambda) = \det(\lambda I - BK_d(1 - e^{-\lambda})).$$

It is clear that

$$(13) \quad \tilde{G}(\lambda) = 0 \iff \lambda - \lambda_i(1 - e^{-\lambda}) = 0, \quad i = 1, \dots, n,$$

where  $\lambda_i$ ,  $i = 1, \dots, m$ , are the eigenvalues of  $BK_d$ . In what follows we distinguish between two cases.

*Case 1:*  $\exists k \in \{1, \dots, n\} : \lambda_k \notin \text{clos}(S)$ . Following from expression (13) and Lemma A.1 in the appendix,  $\tilde{G}$  has a zero in  $\mathbb{C}^+$ . By the uniform convergence of  $G(\cdot; p)$  to  $\tilde{G}$  on compact sets as  $p \rightarrow 0$  and Rouché's theorem, there exist constants  $c_1 > 0$  and  $\hat{p}_1 > 0$  such that for all  $p \in (0, \hat{p}_1)$ ,  $G(\cdot; p)$  has a zero in the right half plane  $\{\lambda \in \mathbb{C} : \Re(\lambda) > c_1\}$ . Hence by (15),  $I_1(\cdot; p)$  has a zero in the half plane  $\{\lambda \in \mathbb{C} : \lambda > c_1/p\}$ .

*Case 2:*  $\lambda_i \in S$ ,  $1 \leq i \leq n$ . By expression (13) and Lemma A.1,  $\tilde{G}$  has a zero at the origin with multiplicity  $n$ , while all other zeros are confined to the open left half plane. Since the zeros of  $G(\cdot; p)$  belong to the set

$$(14) \quad \left\{ \lambda \in \mathbb{C} : |\lambda| \leq \|Ap\| + \|BK_d\|(1 + e^{-\Re(\lambda)}) \right\}$$

and  $G(\cdot; p)$  uniformly converges to  $\tilde{G}$  on compact sets as  $p \rightarrow 0$ , an application of Rouché's theorem allows us to conclude the existence of constants  $c > 0$  and  $\hat{p} > 0$  such that for all  $p \in (0, \hat{p})$ ,

1.  $G(\cdot; p)$  has exactly  $n$  zeros in the half plane  $\{\lambda \in \mathbb{C} : \Re(\lambda) > -c\}$ ;
2. all other zeros of  $G(\cdot; p)$  are in the half plane  $\{\lambda \in \mathbb{C} : \Re(\lambda) < -c\}$ .

This is equivalent to the assertion of the proposition when taking into account the equivalence

$$(15) \quad G(\lambda; p, 0) = 0 \iff I_1(\lambda p; p) = 0. \quad \square$$

*Remark 3.2.* In [23] the special case is treated where  $A$  and  $BK_d$  are multiples of each other, which stems from a gradient play dynamics application.

**COROLLARY 3.3.** *If  $BK_d$  has an eigenvalue outside  $\text{clos}(S)$  and control law (2) is stabilizing, then the closed-loop system is not  $p$ -stable. Furthermore,  $I_1(\lambda; p)$  has zeros in  $\mathbb{C}^+$  for arbitrarily small values of  $p$ .*

**3.2. Feedback delay.** We assume the presence of an unmodeled feedback delay  $p$  in all input channels, that is,

$$G_1(\lambda; p) = e^{-\lambda p} I, \quad G_2(\lambda; p) = I.$$

Then the characteristic function becomes

$$I_2(\lambda; p) := \det(\lambda I - A - BK_d \lambda e^{-\lambda p}).$$

As shown in the following result, the eigenvalue distribution of  $BK_d$  determines the behavior of the zeros thus introduced.

**PROPOSITION 3.4.** *If  $\rho(BK_d) > 1$ , there exist numbers  $\hat{p} > 0$ ,  $c > 0$  and a function  $\hat{\lambda}: (0, \hat{p}) \rightarrow \mathbb{C}$ ,  $p \mapsto \hat{\lambda}(p)$ , such that*

$$I_2(\hat{\lambda}(p); p) = 0 \quad \text{and} \quad \Re(\hat{\lambda}(p)) > \frac{c}{p}$$

for all  $p \in (0, \hat{p})$ .

*If  $\rho(BK_d) < 1$ , there exist numbers  $\hat{p} > 0$  and  $c > 0$  such that for all  $0 < p \leq \hat{p}$ , the function  $I_2(\lambda; p)$  has exactly  $n$  zeros in the half plane  $\{\lambda \in \mathbb{C} : \Re(\lambda) > -\frac{c}{p}\}$ .*

*Proof.* If  $\rho(BK_d) > 1$ , it follows from the theory developed in [4, 16] that for all  $p > 0$ , there exists a sequence of complex numbers  $\{\lambda_\nu\}_{\nu \geq 0}$  such that

$$I_2(\lambda_\nu; p) = 0, \quad \nu \geq 1, \\ \lim_{\nu \rightarrow \infty} \Im(\lambda_\nu) = +\infty, \quad \lim_{\nu \rightarrow \infty} \Re(\lambda_\nu) = \frac{\log(\rho(BK_d))}{p}.$$

By letting  $c = (\log(\rho(BK_d)))/2$  the statement of the proposition follows.

If  $\rho(BK_d) < 1$ , there exists a  $c > 0$  such that  $\rho(BK_d e^c) < 1$ . Next, we let  $\lambda_0$  be a zero of  $I_2(\cdot; p)$  satisfying  $\Re(\lambda) > -c/p$ . As the matrix  $(I - BK_d e^{-\lambda p})$  is invertible if  $\Re(\lambda) > -c/p$ , we get

$$\det(\lambda_0 I - (I - BK_d e^{-\lambda_0 p})^{-1} A) = 0.$$

This implies

$$(16) \quad |\lambda_0| \leq \max_{\Re(\lambda) \geq -c/p} \|(I - BK_d e^{-\lambda p})^{-1} A\| \leq M,$$

where

$$M := \max_{\Re(\lambda) \geq -c} \|(I - BK_d e^{-\lambda})^{-1} A\|.$$

Consequently, all zeros of  $I_2(\cdot; p)$  in the half plane  $\{\lambda \in \mathbb{C} : \Re(\lambda) \geq -c/p\}$  also lie in the disk  $\{\lambda \in \mathbb{C} : \|\lambda\| \leq M\}$ . Combining this result with Proposition 2.7 yields the assertion to be proven.  $\square$

**COROLLARY 3.5.** *If  $BK_d$  has an eigenvalue outside the unit disk and control law (2) is stabilizing, then the closed-loop system is not practically stable. Furthermore,  $I_2(\lambda; p)$  has zeros in  $\mathbb{C}^+$  for arbitrarily small values of  $p$ .*

**4. Filtered state derivative feedback.** Using Assumption 2.5 and Rouché-type arguments, all but  $n$  zeros of  $H(\lambda; p)$  move off to infinity as the parameter  $p$  tends to zero. An obstruction to  $p$ -stability occurs if some of these zeros move off to infinity *without* leaving the closed right half plane. Such situations are illustrated

with Propositions 3.1 and 3.4. A natural way to prevent the presence of zeros with a large modulus in the right half plane consists of including a *low-pass filter* in the control scheme. Note that such an approach has already been successfully applied to the discretization of distributed delay controllers in the context of finite spectrum assignment, where similar robustness problems occur [15, 17].

When applying a first order filter to the control law (2), the controller becomes

$$(17) \quad T\dot{u}(t) + u(t) = K_d\dot{x}(t),$$

where  $T = 1/\omega_f$  is the time constant of the filter, and  $\omega_f$  is its cutoff frequency.

The feedback system that consists of (1) and (17) is given by

$$(18) \quad \dot{z}(t) = \begin{bmatrix} A & B \\ 0 & -\frac{1}{T}I \end{bmatrix} z(t) + \begin{bmatrix} 0 & 0 \\ \frac{1}{T}K_d & 0 \end{bmatrix} \dot{z}(t),$$

where  $z(t) = [x(t)^T \ u(t)^T]^T$ . This system can be rewritten as

$$(19) \quad \dot{z}(t) = \begin{bmatrix} A & B \\ \frac{1}{T}K_dA & \frac{1}{T}(K_dB - I) \end{bmatrix} z(t).$$

Let  $J_0(\lambda; T)$  be the characteristic function of (18)–(19), that is,

$$(20) \quad J_0(\lambda; T) := \det \left( \lambda I - \begin{bmatrix} A & B \\ \frac{1}{T}K_dA & \frac{1}{T}(K_dB - I) \end{bmatrix} \right).$$

We perform the subsequent analysis in two steps. First, we discuss the effect of the introduction of the filter on the stability of the *nominal system*. Next, we investigate the effect of small approximation and implementation errors. We end the section with a brief discussion of the results.

**4.1. Filter design.** Intuitively, one might expect that the introduction of a filter with a sufficiently high cutoff frequency (i.e.,  $T$  is sufficiently small) has little influence on the dynamic behavior of the nominal system. However, this is not always the case, as shown by the following result.

**PROPOSITION 4.1.** *Assume that  $\det(I - BK_d) \neq 0$  and let  $\xi_i \in \mathbb{C}$ ,  $1 \leq i \leq n$ , be the eigenvalues of  $(I - BK_d)^{-1}A$ . There exist a number  $\hat{T}$  and  $n$  continuous functions*

$$\hat{\eta}_i : (0, \hat{T}) \rightarrow \mathbb{C}, \quad T \mapsto \hat{\eta}_i(T), \quad 1 \leq i \leq n,$$

*which satisfy  $J(\hat{\eta}_i(T); T) \equiv 0$  and*

$$(21) \quad \lim_{T \rightarrow 0+} \hat{\eta}_i(T) = \xi_i, \quad 1 \leq i \leq n.$$

*If  $(K_dB - I)$  is Hurwitz, then there exist numbers  $\hat{c} > 0$ ,  $\hat{T} > 0$  such that for all  $T \in (0, \hat{T})$ , the function  $J_0(\lambda; T)$  has exactly  $n$  zeros in the half plane*

$$\left\{ \lambda \in \mathbb{C} : \Re(\lambda) > -\frac{\hat{c}}{T} \right\}.$$

*If  $(K_dB - I)$  has an eigenvalue in the open right half plane, then there exist numbers  $\hat{c} > 0$  and  $\hat{T} > 0$  such that for all  $T \in (0, \hat{T})$ , the function  $J_0(\lambda; T)$  has a zero in the half plane*

$$\left\{ \lambda \in \mathbb{C} : \Re(\lambda) > \frac{\hat{c}}{T} \right\}.$$



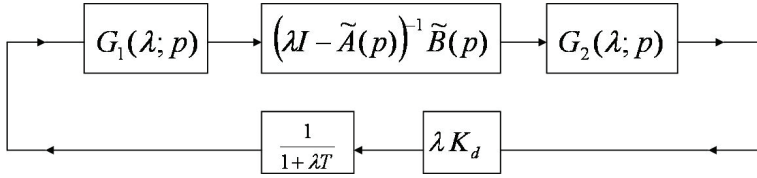


FIG. 2. Feedback interpretation of the controlled system with a low-pass filter included in the feedback.

*Proof.* The proof follows the same steps as the proofs of Propositions 2.7 and 3.1, and it is therefore not developed in detail. We restrict ourselves to some consideration on the large eigenvalues introduced by the filter. From the normalized characteristic function

$$T^{n+m} J_0 \left( \frac{\mu}{T}; T \right) = \det \left( \mu I - \begin{bmatrix} AT & BT \\ K_d A & (K_d B - I) \end{bmatrix} \right),$$

where the argument of  $\det(\cdot)$  becomes triangular as  $T \rightarrow 0+$ , it is apparent that for small values of  $T$ , the eigenvalues of  $(K_d B - I)$  determine the large eigenvalues of the closed-loop system.  $\square$

From Proposition 4.1 one concludes that the presence of the filter may induce an additional stability constraint, namely, the Hurwitz stability of the matrix  $(K_d B - I)$ .

*Remark 4.2.* The above results can also be interpreted in terms of robustness of the unfiltered feedback system (1)–(2). If the nominal system (1)–(2) is asymptotically stable but  $(K_d B - I)$  is not Hurwitz, then the stability of the nominal system is not robust against arbitrarily small modeling error described by

$$(22) \quad n_p = 1; \quad G_1(\lambda; p) = I; \quad G_2(\lambda; p) = \frac{1}{1 + p\lambda}.$$

Note that (22) satisfies Assumption 2.5. In this way the above analysis can be seen as another illustration of the fragility, already illustrated in the previous section with two examples of a different nature.

**4.2. Effect of small modeling and implementation errors.** We consider the controller (17) and assume that  $T$  and  $K_d$  are such that the null solution of (18) is asymptotically stable.

In the presence of small modeling and approximation errors as described in section 2.1 the characteristic function (20) becomes

$$(23) \quad J(\lambda; T, p) := \det \left( \lambda I - \begin{bmatrix} \tilde{A}(p) & \tilde{B}(p) \\ \frac{1}{T} G_1(\lambda; p) K_d G_2(\lambda; p) \tilde{A}(p) & \frac{1}{T} (G_1(\lambda; p) K_d G_2(\lambda; p) \tilde{B}(p) - I) \end{bmatrix} \right),$$

where  $\tilde{A}, \tilde{B}, G_1, G_2$  satisfy Assumption 2.5. The corresponding block diagram of the closed-loop system is displayed in Figure 2.

Due to the low-pass filter in the control loop, the closed-loop system, described by (1) and (17), is  $p$ -stable in the sense of Definition 2.6, as can be concluded from the following proposition.

PROPOSITION 4.3. Assume that  $T > 0$  is fixed. Let  $\xi_i \in \mathbb{C}$ ,  $1 \leq i \leq n + m$ , be the zeros of  $J_0(\lambda; T)$ . There exist a number  $\hat{p}$  and  $n + m$  continuous functions

$$\hat{\eta}_i : [0, \hat{p})^{n_p} \rightarrow \mathbb{C}, \quad p \mapsto \hat{\eta}_i(p), \quad 1 \leq i \leq n + m,$$

which satisfy  $J(\hat{\eta}_i(p); T, p) \equiv 0$  and

$$(24) \quad \lim_{p \rightarrow 0} \hat{\eta}_i(p) = \xi_i, \quad 1 \leq i \leq n + m.$$

Furthermore, there exist numbers  $\tilde{p} > 0$  and  $c > 0$  such that for all  $p \in (\mathbb{R}^+)^{n_p}$  with  $\|p\| \leq \tilde{p}$ , the functions  $J(\lambda; T, p)$  and  $J_0(\lambda; T)$  have the same number of zeros in the half plane

$$(25) \quad \{\lambda \in \mathbb{C} : \Re(\lambda) \geq -c\}.$$

*Proof.* The proof of the first assertion is similar to the proof of Proposition 2.7, and is omitted. For the second assertion, choose numbers  $N$  and  $P$  according to item 6 of Assumption 2.5. Choose  $c \in (0, N)$  such that  $J_0$  has no zero with real part equal to  $c$ . Let  $\lambda_0$  be a zero of  $J(\cdot; T, p_0)$ , where  $\|p_0\| < P$ . If  $\Re(\lambda_0) \geq -c$ , then we get from (23) that

$$\lambda_0 \in \sigma \left( \begin{bmatrix} \tilde{A}(p) & \tilde{B}(p) \\ \frac{1}{T}G_1(\lambda_0; p)K_dG_2(\lambda_0; p)\tilde{A}(p) & \frac{1}{T}(G_1(\lambda_0; p)K_dG_2(\lambda_0; p)\tilde{B}(p) - I) \end{bmatrix} \right).$$

As  $\rho(\cdot) \leq \|\cdot\|$ , this implies that  $|\lambda_0| \leq R$ , where

$$(26) \quad R = \sup \left\{ \left\| \begin{bmatrix} \tilde{A}(p) & \tilde{B}(p) \\ \frac{1}{T}G_1(\lambda; p)K_dG_2(\lambda; p)\tilde{A}(p) & \frac{1}{T}(G_1(\lambda; p)K_dG_2(\lambda; p)\tilde{B}(p) - I) \end{bmatrix} \right\| : \lambda \in \mathbb{C}, p \in \mathbb{R}^{n_p}, \Re(\lambda) \geq -c, \|p\| \leq P \right\}.$$

Note that by Assumption 2.5 the right-hand side of (26) is finite. We conclude that if  $\|p\| \leq P$ , then all zeros of  $J(\cdot; T, p)$  in the half plane (25) are confined to the compact set

$$\Omega := \{\lambda \in \mathbb{C} : \Re(\lambda) \geq -c, |\lambda| \leq R\}.$$

Next, from Assumption 2.5 it follows that  $J(\lambda; T, p)$  converges to  $J_0(\lambda; T)$  uniformly on compact sets as  $p \rightarrow 0$ . Hence, we can choose  $\tilde{p} \in (0, P)$  such that for all  $p \in \mathbb{R}^{n_p}$  with  $\|p\| \leq \tilde{p}$  the following estimate holds:

$$\max_{\lambda \in \partial\Omega} |J_0(\lambda; T) - J(\lambda; T, p)| \leq \min_{\lambda \in \partial\Omega} |J_0(\lambda; T)|.$$

The second assertion of the proposition then follows from an application of Rouché's theorem.  $\square$

**4.3. Discussion.** In section 3 we have shown by construction that (unfiltered) state derivative feedback may lead to a fragile closed-loop system in the sense that the closed-loop system is stable but not  $p$ -stable. This was illustrated for a numerical

approximation of derivatives by a finite difference formula and for a small delay in the feedback loop. In both cases the eigenvalues of the matrix  $BK_d$  determine the robustness of stability. However, even if the resulting conditions on the gain  $K_d$  are satisfied, stability may still lack robustness against other types of perturbations satisfying Assumption 2.5.

In this section we have shown that if the filtered derivative feedback control law (17) is stabilizing, then the closed-loop system is  $p$ -stable. However, the existence of a stabilizing filtered derivative feedback for the nominal system may again impose restrictions on the gain, as expressed in Proposition 4.1. Summarizing, we have the following result.

**THEOREM 4.4.** *Assume that the control law (2) asymptotically stabilizes system (1).*

*If the matrix  $(BK_d - I)$  is Hurwitz, then the filtered control law (17) is stabilizing for small values of  $T$  and results in a  $p$ -stable closed-loop system.*

*If  $(BK_d - I)$  has an eigenvalue in  $\mathbb{C}^+$ , then the closed-loop system with control law (2) is not  $p$ -stable. Furthermore, the filtered control law (17) is not stabilizing for small values of  $T$ .*

*Proof.* The proof follows from Proposition 4.1, Remark 4.2, and Proposition 4.3.  $\square$

Further refinements will be made in the next section, where relations between stabilizing values of  $K_d$  and the eigenvalue distribution of  $(BK_d - I)$  will be taken into account.

**5. Conditions for  $p$ -stabilizability.** In this section we discuss the  $p$ -stabilizability problem and the design of stabilizing state derivative controllers of the form (2) or (17). As we shall see, the condition  $\det(-A) > 0$ , which is satisfied if  $A$  has no zero eigenvalue and an even number of eigenvalues in the closed right half plane, will play a crucial role.

Throughout this section we assume that  $(A, B)$  is controllable and that  $A$  is cyclic. Remarks on the noncyclic case will be made in the next section.

We start by stating a technical lemma.

**LEMMA 5.1.** *Assume that  $\det(-A) < 0$ . If the control law (2) is stabilizing, then the matrix  $BK_d$  has a real eigenvalue larger than one.*

*Proof.* The proof is based on a continuation argument. If we consider the feedback  $u = kK_d\dot{x}(t)$ , where  $k \in [0, 1]$  is a real parameter, then the closed-loop system becomes

$$(27) \quad (I - kBK_d)\dot{x}(t) = Ax(t).$$

Since  $\det(-A) < 0$ , for  $k = 0$  the system has an odd number of characteristic roots in the open right half plane. On the other hand, for  $k = 1$  it has an even number of characteristic roots in the open right half plane (zero) as it is assumed that (2) is stabilizing. Because the characteristic roots appear in complex conjugate pairs and depend continuously on  $k$ , there must be a value of  $k = \tilde{k} \in (0, 1)$  for which there is a characteristic root either at zero or “at infinity.” The former is not possible because a zero characteristic root is invariant w.r.t. changes of  $k$  and would contradict the stability for  $k = 1$ . The latter implies that  $\det(I - \tilde{k}BK_d) = 0$ . Consequently,  $\tilde{k}BK_d$  has an eigenvalues equal to one, or, equivalently,  $BK_d$  has a real eigenvalue equal to  $1/\tilde{k} > 1$ .  $\square$

The two following theorems are direct corollaries.

THEOREM 5.2. *If  $\det(-A) < 0$  and control law (2) is stabilizing, then the closed-loop system is not  $p$ -stable.*

*Proof.* This result can be shown in three different ways. As  $BK_d$  has an eigenvalue larger than one, an approximation of the derivative with a finite difference scheme (Proposition 3.1), a small feedback delay (Proposition 3.4), as well as a neglected first order lag (Proposition 4.1, Remark 4.2) destroy stability.  $\square$

THEOREM 5.3. *If  $\det(-A) < 0$ , then the system cannot be stabilized with a control law of the form (17). Moreover, every dynamic control law of the form*

$$(28) \quad \dot{\zeta}(t) = A_f \zeta(t) + B_f \dot{x}(t), \quad u(t) = C_f \zeta(t),$$

*with  $A_f$  Hurwitz, results in an unstable closed-loop system.*

*Proof.* The first assertion is a consequence of Theorem 4.4 and Lemma 5.1. The proof of the second assertion is by contradiction and employs a continuation argument. Assume that control law (28) is stabilizing. With the parameterized control law

$$\dot{\zeta}(t) = A_f \zeta(t) + k B_f \dot{x}(t), \quad u(t) = C_f \zeta(t),$$

with parameter  $k \in [0, 1]$ , the closed-loop system becomes

$$(29) \quad \left( I - \begin{bmatrix} 0 & 0 \\ k B_f & 0 \end{bmatrix} \right) \dot{\xi}(t) = \begin{bmatrix} A & B C_f \\ 0 & A_f \end{bmatrix} \xi(t),$$

where  $\xi(t) = [x(t)^T \quad \zeta^T(t)]^T$ . For  $k = 0$  the system has an odd number of characteristic roots in the open right half plane as  $\det(-A) \det(-A_f) < 0$ , while for  $k = 1$  it has an even number of characteristic roots in the open right half plane (zero) as it is assumed that (28) is stabilizing. Because the characteristic roots appear in complex conjugate pairs and depend continuously on  $k$ , there must be a value of  $k = \tilde{k} \in (0, 1)$  for which there is a characteristic root either at zero or “at infinity.” The former is not possible because a characteristic root at zero is invariant w.r.t. changes of  $k$ , which contradicts the stability for  $k = 1$ ; the latter is not possible because the matrix

$$I - \begin{bmatrix} 0 & 0 \\ k B_f & 0 \end{bmatrix}$$

is invertible for all values of  $k$ .  $\square$

Remark 5.4. This result is expected from the stabilization mechanism outlined in Remark 2.4 and the fact that a filter prevents characteristic roots with a large modulus.

Next, we consider the case where  $\det(-A) > 0$ . We first have the following lemma.

LEMMA 5.5. *If  $\det(-A) > 0$ , then there always exists a stabilizing control law of the form (2) for which all eigenvalues of  $BK_d$  are zero.*

*Proof.* The existence of such a control law is shown by construction in the proof of Theorem 2 of [10]. To make this paper self contained, we outline the main steps.

There exist a transformation of the state,  $z = T_x x$ , and of the input,  $w = T_u u$ , that put system (1) in the controller canonical form

$$(30) \quad \dot{z}(t) = A_c z(t) + B_c w(t),$$

with

$$A_c = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & & & 0 & 1 \\ -a_n & \cdots & \cdots & -a_2 & -a_1 \end{bmatrix}, \quad B_c = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix} \left| \begin{array}{c} \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{array} \right| \cdots \left| \begin{array}{c} \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{array} \right| \begin{bmatrix} \tilde{B}_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{bmatrix}.$$

The control law

$$(31) \quad w(t) = \tilde{K}_d \dot{z}(t),$$

where

$$\tilde{K}_d = \begin{bmatrix} -k_{n-1} & \cdots & -k_1 & 0 \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix},$$

then results in a closed-loop system with characteristic equation

$$\lambda^n + \sum_{l=1}^{n-1} (a_l + k_l) \lambda^{n-l} + a_n = 0.$$

As  $a_n = \det(-A) > 0$ , there always exist stabilizing values of  $k_1, \dots, k_{n-1}$ . Furthermore, all eigenvalues of  $B_c \tilde{K}_d$  are equal to zero. In the original coordinates the control law (31) becomes

$$u(t) = T_u^{-1} \tilde{K}_d T_x \dot{x}(t) := K_d \dot{x}(t). \quad \square$$

By combining Lemma 5.5 with Theorem 4.4 we arrive at the following result.

**THEOREM 5.6.** *If  $\det(-A) > 0$ , then there always exists a stabilizing controller of the form (17) for which the closed-loop system is  $p$ -stable.*

The obtained stabilizability conditions are summarized in Table 1.

TABLE 1  
*p-stabilizability of the controllable system (1) using state derivative feedback. A is assumed cyclic.*

	$\det(-A) < 0$	$\det(-A) = 0$	$\det(-A) > 0$
stabilizable	yes	no	yes
$p$ -stabilizable	no (neither with (2) nor with (17) and (28))	no	yes (with (17))

**6. Remarks on the noncyclic case.** If the cyclic index<sup>1</sup> of  $A$  is  $k > 1$ , then there always exist transformations of the state,  $z = T_x x$ , and the input,  $w = T_u u$ , which transform (1) into

$$(32) \quad \dot{z}(t) = A_c z(t) + B_c w(t),$$

<sup>1</sup>The maximum of the geometric multiplicities of its eigenvalues.

where

$$A_c = \begin{bmatrix} A_{11} & & 0 \\ & \ddots & \\ 0 & & A_{kk} \end{bmatrix}, \quad B_c = \left[ \begin{array}{ccc|ccc} B_{11} & \cdots & B_{1k} & \cdots & B_{1m} \\ & & \vdots & & \vdots \\ 0 & & B_{kk} & \cdots & B_{km} \end{array} \right],$$

with  $A_{ii}$  cyclic and the pair  $(A_{ii}, B_{ii})$  controllable for each  $i = 1, \dots, k$ .

As their proofs do not depend on the cyclic index of  $A$ , Theorems 5.2 and 5.3 remain valid.

**THEOREM 6.1.** *If  $\det(-A) < 0$  and the control law (2) is stabilizing, then the closed-loop system is not  $p$ -stable. Furthermore, every dynamic control law of the form (28), with  $A_f$  Hurwitz, results in an unstable closed-loop system.*

Based on a decomposition into  $k$  subproblems induced by the canonical form (32), and an application of Theorem 5.6 to each subproblem, we arrive at the following result.

**THEOREM 6.2.** *If  $\det(-A_{ii}) > 0$ ,  $1 \leq i \leq k$ , then there always exists a stabilizing controller of the form (17) for which the closed-loop system is  $p$ -stable.*

When comparing Theorem 6.1 with Theorem 6.2, the stabilizability analysis is complete if one can make an assertion for the case where

$$(33) \qquad \det(-A) > 0, \quad \exists i \in \{1, \dots, k\} : \det(-A_{ii}) < 0.$$

This is still an open problem. The following example indicates that in such case a lack of practical stability may occur.

*Example 6.3.* The system

$$\begin{cases} \dot{x}_1(t) = x_1(t) + u_1(t), \\ \dot{x}_2(t) = x_2(t) + u_2(t) \end{cases}$$

is of the form (32) with  $A_c = I, B_c = I$ , and  $A_{11} = A_{22} = 1$ . Clearly, we have  $\det(-A_c) > 0$ , but  $\det(-A_{11}) = \det(-A_{22}) < 0$ .

The control law

$$u(t) = K_d \dot{x}(t), \quad K_d \in \mathbb{R}^{2 \times 2},$$

is stabilizing if and only if the eigenvalues of the matrix  $K_d$  have real part larger than one. By both Propositions 3.1 and 3.4, the closed-loop system is not  $p$ -stable.

If we apply the filtered control law

$$T\dot{u}(t) + u(t) = K_d \dot{x}(t), \quad T > 0, \quad K_d \in \mathbb{R}^{2 \times 2},$$

then the characteristic equation of the closed-loop is given by

$$\det \left( \lambda^2 I + \lambda \left( \frac{1}{T} (I - K_d) - I \right) - \frac{1}{T} I \right) = 0.$$

As this equation has a zero in  $\mathbb{C}^+$  for all values of  $T$  and  $K_d$ , stabilization is not possible.

**7. Concluding remarks.** The possible lack of stability robustness of the stabilized system (1)–(2) against arbitrarily small modeling and implementation errors, which was illustrated with several example case studies, led us to the practical notion

of  $p$ -stability. The stabilizability and  $p$ -stabilizability of system (1) with the (filtered) control law (2) was characterized.

For the generic case where  $A$  is cyclic, a *complete* characterization of stabilizability is described in Table 1. Surprisingly, if  $\det(-A) < 0$ , which is satisfied if  $A$  has an odd number of eigenvalues in the open right half plane, the system *is* stabilizable, but *not*  $p$ -stabilizable, using neither static feedback nor dynamic feedback. This shows that the so-called odd number condition,  $\det(-A) < 0$ , well known in the context of Pyragas-type time-delayed feedback controllers, is also a *fundamental obstruction* to stabilizability in this context, yet only appears at a second stage, where robustness aspects are considered.

For the case where  $A$  is noncyclic a full characterization of stabilizability and  $p$ -stabilizability was made, except when (33) holds, that is, in the canonical form there are subsystems which satisfy the odd number condition, but the whole systems does not. Example 6.3 illustrates that again a lack of robustness of stability may occur and suggests that the necessary and sufficient  $p$ -stabilizability condition in the noncyclic case is given by  $\det(-A_{ii}) > 0$  for all  $i = 1, \dots, k$ . Whether or not this is true is an open problem and a topic of further research.

Motivated by the large number of applications of sampled-data systems, which appear, for instance, in network controlled systems, it is also worthwhile to investigate to what extent the controller construction, observed phenomena, and robustness issues discussed in the paper carry over to the hybrid case where the controller is discrete and difference based. In the overview article [15], where one investigates the effect on stability of approximating distributed delays in a class of control laws for time-delay systems, it is shown that also with pure discrete approximations of control laws fragility problems may occur, and may lead to phenomena which are different from those observed for continuous approximations.

#### Appendix. A technical lemma.

LEMMA A.1. *Consider the equation*

$$(34) \quad \lambda - \mu(1 - e^{-\lambda}) = 0,$$

where  $\mu \in S$ , with  $S$  given by (12). Then all roots of (34) are in  $\mathbb{C}^-$ , except for a root at the origin with multiplicity one.

*Proof.* We distinguish two cases as below.

Case 1:  $\mu \notin \mathbb{R}$ . We first characterize the zeros of the auxiliary function

$$h(\lambda; \nu) := \lambda - \mu(1 - e^{-\lambda\nu})$$

as a function of the parameter  $\nu \geq 0$ . For  $\nu = 0$ , the function  $h$  has only one zero at the origin, which is invariant w.r.t. changes of  $\nu$ . If  $\nu$  is increased from zero, then the number of zeros in the closed right half plane can increase only if zeros cross the imaginary axis. To see this, note that  $h(\lambda; \nu) = 0$  implies

$$|\lambda| = |\mu| |1 - e^{-\lambda\nu}|$$

and

$$|\lambda| \leq |\mu| \left(1 + e^{-\Re(\lambda)\nu}\right).$$

Consequently, whatever the value of  $\nu \geq 0$ , the zeros of  $h(\cdot, \nu)$  in the closed right half plane (if any) have modulus smaller than or equal to  $2|\mu|$ . Thus by varying  $\nu$  the

number of zeros of  $h$  in the right half plane cannot be changed by zeros coming from infinity, only by zeros crossing the imaginary axis.

A zero at the origin with multiplicity larger than one cannot occur since

$$h'(0; \nu) = 1 - \mu\nu \neq 0.$$

If  $h$  has a zero  $j\omega$ ,  $\omega \in \mathbb{R} \setminus \{0\}$ , for some value of  $\nu$ , then we have

$$j\omega = \mu(1 - e^{-j\omega\nu}).$$

Solving this equation yields

$$(35) \quad \omega = \omega^* := 2\Im(\mu), \quad \nu = \nu_k^* := \frac{\angle(\mu) + \pi k}{\Im(\mu)}, \quad k \geq 1.$$

Next, we look at the crossing direction of a zero on the imaginary axis w.r.t. the parameter  $\nu$ . Taking the derivative of  $h(\lambda(\nu); \nu) = 0$  w.r.t.  $\nu$  at  $(\lambda, \nu) = (j\omega^*, \nu_k^*)$  yields

$$\lambda'(\nu^*) = \frac{j\omega^* \mu e^{-j\omega^* \nu^*}}{1 - \nu \mu e^{-j\omega^* \nu}},$$

from which we get

$$\begin{aligned} \Re(\lambda'(\nu^*)^{-1}) &= \Re\left(\frac{1}{j\omega^* \mu e^{-j\omega^* \nu^*}}\right) = \Re\left(\frac{1}{j\omega^* (\mu - j\omega^*)}\right) \\ &= \Re\left((2\Im(\mu)(\Im(\mu) + j\Re(\mu)))^{-1}\right) > 0 \end{aligned}$$

and

$$(36) \quad \Re(\lambda'(\nu^*)) > 0.$$

From (35) and (36) we can conclude that

1. if  $\nu \in [0, \nu_1^*)$ , then all zeros of  $h$  are in  $\mathbb{C}^-$ , except for a zero at the origin with multiplicity one;
2. if  $\nu > \nu_1^*$ , then  $h$  has zeros in  $\mathbb{C}^+$ .

The assertion of the proposition is straightforward when taking into account that  $\mu \in S$  ( $\mu \notin \text{clos}(S)$ ) is equivalent to  $\nu_1^* > 1$  ( $\nu_1^* < 1$ ).

*Case 2:*  $\mu \in \mathbb{R}$ . The equation  $h = 0$  is the characteristic equation of the system  $\dot{x}(t) = \mu x(t) - \mu x(t - \tau)$ . The stability of this system has been analyzed in, e.g., [19, 24], from which the statements of the proposition follow.  $\square$

## REFERENCES

- [1] T. ABDELAZIZ AND M. VALÁŠEK, *Pole-placement for SISO linear systems by state-derivative feedback*, IEE Proc. Control Theory Appl., 151 (2004), pp. 377–385.
- [2] T. ABDELAZIZ AND M. VALÁŠEK, *Direct algorithm for pole placement by state-derivative feedback for multi-input linear systems—nonsingular case*, Kybernetika, 41 (2005), pp. 637–660.
- [3] T. ABDELAZIZ AND M. VALÁŠEK, *State derivative feedback by LQR for linear time-invariant system*, in Proceedings of the 16th IFAC World Congress, Elsevier, 2005.
- [4] C. AVELLAR AND J. HALE, *On the zeros of exponential polynomials*, Math. Anal. Appl., 73 (1980), pp. 434–452.
- [5] R. DATKO, *Not all feedback stabilized hyperbolic systems are robust with respect to small time delays in their feedbacks*, SIAM J. Control Optim., 26 (1988), pp. 697–713.



- [6] R. DATKO AND Y. YOU, *Some second-order vibrating systems cannot tolerate small time delays in their damping*, J. Optim. Theory Appl., 70 (1991), pp. 521–537.
- [7] S. DYKE, B. SPENCER, P. QUAST, M. SAIN, D. KASPARI, AND T. SOONG, *Acceleration feedback control of MDOF structures*, J. Engrg. Mech., 122 (1996), pp. 907–917.
- [8] T. GEORGIOU AND M. SMITH, *w-Stability of feedback systems*, Systems Control Lett., 13 (1989), pp. 217–277.
- [9] J. HALE, *Effects of delays on dynamics*, in Topological Methods in Differential Equations and Inclusions, A. Granas, M. Frigon, and G. Sabidussi, eds., Kluwer Academic, Dordrecht, The Netherlands, 1995, pp. 191–238.
- [10] H. KOKAME, K. HIRATA, K. KONISHI, AND T. MORI, *Difference feedback can stabilize uncertain steady states*, IEEE Trans. Automat. Control, 46 (2001), pp. 1908–1913.
- [11] H. LOGEMANN, *Destabilizing effects of small time-delays on feedback-controlled descriptor systems*, Linear Algebra Appl., 272 (1998), pp. 131–153.
- [12] H. LOGEMANN AND R. REBARBER, *The effect of small time-delays on the closed-loop stability of boundary control systems*, Math. Control Signals Systems, 9 (1996), pp. 123–151.
- [13] H. LOGEMANN AND S. TOWNLEY, *The effect of small delays in the feedback loop on the stability of neutral systems*, Systems Control Lett., 27 (1996), pp. 267–274.
- [14] W. MICHIELS, K. ENGELBORGH, D. ROOSE, AND D. DOCHAIN, *Sensitivity to infinitesimal delays in neutral equations*, SIAM J. Control Optim., 40 (2001), pp. 1134–1158.
- [15] W. MICHIELS, S. MONDIÉ, D. ROOSE, AND M. DAMBRINE, *The effect of approximating distributed delay control laws on stability*, in Advances in Time-Delay Systems, Lecture Notes in Comput. Sci. Engrg., Springer-Verlag, Berlin, 2004, pp. 207–225.
- [16] W. MICHIELS AND T. VYHLÍDAL, *An eigenvalue based approach for the stabilization of linear time-delay systems of neutral type*, Automatica, 41 (2005), pp. 991–998.
- [17] S. MONDIÉ AND W. MICHIELS, *Finite spectrum assignment of unstable time-delay systems with a safe implementation*, IEEE Trans. Automat. Control, 48 (2003), pp. 2207–2212.
- [18] Ö. MORGÜL, *On the stabilization and stability robustness against small delays of some damped wave equations*, IEEE Trans. Automat. Control, 40 (1995), pp. 1626–1630.
- [19] S.-I. NICULESCU, *Delay Effects on Stability. A Robust Control Approach*, Lecture Notes in Control Inform. Sci. 269, Springer-Verlag, London, 2001.
- [20] N. OLGAC, H. ELMALI, M. HOSEK, AND M. RENZULLI, *Active vibration control of distributed systems using delayed resonator with acceleration feedback*, Trans. ASME J. Dynam. Systems Measurement Control, 119 (1997), pp. 380–389.
- [21] K. PYRAGAS, *Continuous control of chaos by self-controlling feedback*, Phys. Lett. A, 170 (1992), pp. 421–428.
- [22] R. REBARBER AND S. TOWNLEY, *Robustness with respect to delays for exponential stability of distributed parameter systems*, SIAM J. Control Optim., 37 (1998), pp. 230–244.
- [23] R. SIPAHI, G. ARSLAN, AND S.-I. NICULESCU, *Some remarks on control strategies for continuous gradient play dynamics*, in Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, CA, 2006.
- [24] G. STÉPÁN, *Retarded Dynamical Systems: Stability and Characteristic Function*, Res. Notes in Math. 210, Longman Scientific, London, 1989.

## MEASURE-VALUED SOLUTIONS FOR A DIFFERENTIAL GAME RELATED TO FISH HARVESTING\*

ALBERTO BRESSAN<sup>†</sup> AND WEN SHEN<sup>†</sup>

**Abstract.** In this paper we consider a model for the harvesting of marine resources, described by an elliptic equation. Since the cost functionals have sublinear growth with respect to the pointwise intensity of fishing effort, optimal solutions are in general measure-valued. For the control problem in one space dimension, we prove the existence of optimal strategies. Uniqueness is established within a class of measures with small total mass. We also study the differential game, modeling the presence of several competing fishing companies, and prove the existence of a Nash equilibrium solution. This is obtained as a fixed point of a continuous transformation in a space of positive Radon measures.

**Key words.** differential games, optimal control, measured-valued solutions, fish harvest

**AMS subject classifications.** 34B15, 34B18, 49J20, 49N25, 49N70, 49N90, 91A10

**DOI.** 10.1137/07071007X

**1. Introduction.** This paper is concerned with a noncooperative differential game modeling the harvesting of marine resources. The general setting of the problem will be discussed for an  $N$ -dimensional domain. The main results, however, will be proved in the one-dimensional case. Consider a bounded domain  $\Omega \subset \mathbb{R}^N$  with smooth boundary. In practice,  $\Omega \subset \mathbb{R}^2$  will describe the region occupied by a lake or a sea. Denote by  $\phi = \phi(t, x)$  the density of fish at time  $t$  at the point  $x \in \Omega$ . In absence of fishing activity, assume that the fish population evolves according to the parabolic equations with source term

$$(1.1) \quad \phi_t = \Delta \phi + g(x, \phi), \quad x \in \Omega,$$

with Neumann boundary conditions

$$(1.2) \quad \nabla \phi \cdot \mathbf{n} = 0, \quad x \in \partial\Omega.$$

A natural choice for  $g$  is

$$(1.3) \quad g(x, \phi) = \alpha (h(x) - \phi) \phi.$$

Here the constant  $\alpha > 0$  is a growth rate, while  $h(x)$  denotes the maximum fish population that can be supported by the habitat at  $x$ . Let  $u_i = u_i(t, x)$  be the intensity of harvesting conducted by the  $i$ th fishing company, at time  $t$  at the location  $x \in \Omega$ . In the presence of this fishing activity, the fish population evolves according to

$$(1.4) \quad \phi_t = \Delta \phi + g(x, \phi) - \sum_{i=1}^m \phi u_i(t, x).$$

---

\*Received by the editors December 4, 2007; accepted for publication (in revised form) August 15, 2008; published electronically January 7, 2009. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

<http://www.siam.org/journals/sicon/47-6/71007.html>

<sup>†</sup>Department of Mathematics, Penn State University, University Park, PA 16802 (bressan@math.psu.edu, shen\_w@math.psu.edu).

Throughout this paper, we consider steady state solutions. These satisfy the elliptic equation

$$(1.5) \quad \Delta\phi + g(x, \phi) = \sum_{i=1}^m \phi u_i(x),$$

together with the Neumann boundary conditions (1.2). The strategies  $u_i(x)$  are now assumed to be independent of time. We are interested in optimal fishing strategies for the various companies. At a steady fishing rate, the  $i$ th company will sustain a cost

$$(1.6) \quad \int_{\Omega} c_i(x) u_i(x) dx.$$

It is reasonable to assume that different companies will incur different costs  $c_i$  for fishing at a given location  $x$ . This is because their home bases may be located at coastal cities with different distances from  $x$ . In addition, there may be restrictions to the fishing activities of companies of various countries. For example, if there is an international treaty that does not allow the  $k$ th company to harvest fish in a subregion  $\Omega' \subset \Omega$ , this will be modeled by setting  $c_k(x) = \infty$  for  $x \in \Omega'$ . If a subregion  $\Omega_0$  is set aside as a marine park where no fishing is permitted, then  $c_i(x) = \infty$  for all  $x \in \Omega_0$ ,  $i = 1, \dots, m$ . The profit for the  $i$ th company will be proportional to the total fish caught:

$$(1.7) \quad p_i = \int_{\Omega} P \cdot \phi(x) u_i(x) dx.$$

Here  $P$  denotes the unit price of fish on the market. By a variable rescaling, it is not restrictive to take  $P = 1$ . The total payoff for the  $i$ th company is therefore

$$(1.8) \quad J_i = \int_{\Omega} (\phi(x) - c_i(x)) u_i(x) dx.$$

The function  $u_i = u_i(x)$  describes the strategy of the  $i$ th company. It is reasonable to assume that it satisfies the constraints

$$(1.9) \quad u_i(x) \geq 0, \quad \int_{\Omega} u_i(x) dx \leq M_i.$$

Here  $M_i$  denotes the maximum amount of fishing within the capabilities of the  $i$ th company. In practice, this may depend on the number of fishermen and on the size of fishing boats available.

*Remark 1.* The cost functionals given at (1.6) are linear with respect to (w.r.t.)  $u_i$ . More general, nonlinear cost functionals could take the form

$$(1.10) \quad \Psi_i \left( \int_{\Omega} c_i(x) u_i(x) dx \right),$$

where  $\Psi_i$  is a convex function, with  $\Psi_i(0) = 0$ ,  $\Psi'_i(0) > 0$ . This accounts for the fact that, as the total amount of harvesting increases, the cost increases superlinearly. On the other hand, a nonlinear cost functional of the form

$$(1.11) \quad G(u) \doteq \int_{\Omega} \psi(u(x)) dx,$$

with  $\psi'(s) \rightarrow \infty$  as  $s \rightarrow \infty$ , is not realistic. For example, assume  $\Omega \subset \mathbb{R}^2$  and choose any point  $\bar{x} \in \Omega$ . Consider the two strategies

$$u(x) \equiv \frac{1}{\text{meas}(\Omega)}, \quad u^{(r)}(x) = \begin{cases} 1/(\pi r^2) & \text{if } |x - \bar{x}| < r, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that

$$\int_{\Omega} u(x) dx = \int_{\Omega} u^{(r)}(x) dx = 1.$$

For the functional (1.10), fishing over the entire domain or on a small part of it yields the same cost. However, for the cost functional  $G$ , as we reduce the fishing area more and more, the cost tends to infinity. Indeed

$$\lim_{r \rightarrow 0} \int_{\Omega} \varphi(u^{(r)}(x)) dx = \infty.$$

This feature is not supported by practical experience.

*Remark 2.* We also observe that integral constraints such as (1.9) are meaningful, while pointwise constraints of the form

$$(1.12) \quad |u_i(x)| \leq M$$

are not. Indeed, (1.12) models the presence of a “parking meter,” so that a fishing boat is allowed to stay in one place up to a maximum amount of time, then it must move elsewhere.

The above remarks indicate that in a realistic model, the cost of harvesting grows at most linearly w.r.t. the pointwise intensity  $u(x)$ . This has important implications for the corresponding optimization problem. Indeed, the existence of optimal strategies can now be obtained not in  $\mathbf{L}^1(\Omega)$  but in the space  $\mathcal{M}_+(\bar{\Omega})$  of nonnegative Radon measures supported on the closure of the domain  $\Omega$ .

This fact is actually consistent with practical experience. If an open subset  $\Omega' \subset \Omega$  is set aside as a marine reserve, the most profitable place to catch fish is right along the boundary of the reserve, i.e., on  $\Omega \cap \partial\Omega'$ . In an optimal strategy, a positive amount of harvesting thus takes place on a set of measure zero. This strategy is described by a measure  $\mu$ , singular w.r.t. Lebesgue measure on  $\Omega$ . We stress that the study of optimal strategies within a space of Radon measures is a major difference between the present paper and earlier literature on the subject [1, 7, 8, 9, 10, 11, 12, 13].

The remainder of the paper is organized as follows. In section 2 we review the definition of Nash equilibrium solution of the differential game. All subsequent analysis deals with the case of one space dimension. Results are expected to be qualitatively similar in more space dimensions. However, there are two features of one-dimensional problems which play a key role in our proofs:

- (i) The Green kernels of second order operators are Lipschitz continuous.
- (ii) Given a two-point boundary value problem on the interval  $[0, R]$  with a source involving a positive measure  $\mu$ , as in [6] one can introduce a new variable  $s$  such that  $s(x) \doteq x + \mu([0, x])$ . Rewriting the differential equation in terms of  $s$  as independent variable, the singularities are removed and classical ODE theory applies.

For given measures  $\mu_i$ , in section 3 we study the existence and uniqueness of strictly positive solutions to the boundary value problem (1.5), (1.2). The existence

and the uniqueness of measure-valued solutions to the optimal control problems for the various players are studied in sections 4 and 5, respectively. Our uniqueness result is obtained by proving the strict concavity of the payoff functional (1.8), as long as the measures  $\mu_i$  remain small. Finally, in section 6 we prove the existence of a Nash equilibrium solution to the differential game. This is obtained as a fixed point of a continuous transformation, in a space of measures.

**2. Nash equilibrium solutions.** From now on, we consider fishing strategies described by positive Radon measures  $\mu_i \in \mathcal{M}_+(\bar{\Omega})$ . The fish population will reach an equilibrium state  $\phi = \phi(x)$ , determined by the elliptic problem with measure-valued sources

$$(2.1) \quad \Delta\phi + g(x, \phi) = \phi \cdot \sum_{i=1}^m \mu_i,$$

and Neumann boundary conditions

$$(2.2) \quad \nabla\phi \cdot \mathbf{n} = 0, \quad x \in \partial\Omega.$$

For the analysis of elliptic equations with measure-valued source terms we refer to [2, 3, 4, 5].

Each company seeks a strategy  $\mu_i$  in order to maximize its payoff

$$(2.3) \quad J_i \doteq \int_{\Omega} \phi d\mu_i(x) - \Psi_i \left( \int_{\Omega} c_i d\mu_i \right),$$

subject to the constraints

$$(2.4) \quad \mu_i \geq 0, \quad \mu_i(\bar{\Omega}) \leq M_i.$$

**DEFINITION 1.** *We say that the  $(1+m)$ -tuple  $(\phi, \mu_1^*, \dots, \mu_m^*)$  is a Nash equilibrium solution of the noncooperative differential game (2.1)–(2.3) if the function  $\phi$  provides a solution to the elliptic boundary value problem (2.1)–(2.2) and, for each  $i = 1, \dots, m$ , the measure  $\mu = \mu_i^*$  achieves the maximum payoff for the optimal control problem*

$$(2.5) \quad \text{maximize: } J_i(\mu) \doteq \int_{\Omega} \phi d\mu(x) - \Psi_i \left( \int_{\Omega} c_i d\mu \right),$$

subject to

$$(2.6) \quad \Delta\phi(x) + g(x, \phi) - \phi \cdot \sum_{j \neq i} \mu_j^*(x) = \phi\mu,$$

with boundary conditions (2.2), and with the constraints

$$(2.7) \quad \mu \geq 0, \quad \mu(\bar{\Omega}) \leq M_i.$$

**3. Positive solutions of the boundary value problem.** From now on we focus on the case of one space dimension, so that our domain is a bounded interval:  $\Omega \doteq ]0, R[$ . In this section we study the uniqueness of nonnegative solutions to the boundary value problem

$$(3.1) \quad \phi'' + g(x, \phi) - \phi \cdot \mu = 0, \quad \phi'(0) = \phi'(R) = 0,$$

where primes denote derivatives w.r.t. the space variable  $x$ , while  $\mu \in \mathcal{M}(\bar{\Omega})$  is a positive Radon measure supported on the closed interval  $\bar{\Omega} = [0, R]$ . On the function  $g$  we impose the following assumptions:

(A1) One has  $g(x, \phi) = f(x, \phi) \phi$ , where the function  $f = f(x, \phi)$  is continuous w.r.t. both variables and twice continuously differentiable w.r.t.  $\phi$ . Moreover, for some continuous function  $h = h(x)$  one has

$$f(x, 0) > 0, \quad f_\phi(x, \phi) < 0, \quad f(x, h(x)) = 0 \quad \text{for all } x \in [0, R], \quad \phi \geq 0.$$

Here and in what follows, by  $f_\phi, f_{\phi\phi}$  we denote, respectively, the first and the second partial derivative of  $f$  w.r.t.  $\phi$ .

Since in (3.1)  $\mu$  can be a measure, a precise concept of solution should first be given.

DEFINITION 2. *By a solution of (3.1) we mean a Lipschitz continuous map  $x \mapsto \phi(x)$  such that*

(i) *the map  $x \mapsto \phi'(x)$  has bounded variation and satisfies*

$$(3.2) \quad \lim_{x \rightarrow 0+} \phi'(x) = \phi(0) \mu(\{0\}), \quad \lim_{x \rightarrow R-} \phi'(x) = -\phi(R) \mu(\{R\});$$

(ii) *for every test function  $\eta \in C_c^1([0, R])$  one has*

$$(3.3) \quad \int_0^R \left\{ -\phi' \eta' + g(x, \phi) \eta \right\} dx - \int_0^R \phi \eta d\mu = 0.$$

We notice that  $\phi \equiv 0$  is always a solution. The next two lemmas are concerned with the existence and uniqueness of strictly positive solutions. In the case where  $\mu$  has a positive density w.r.t. Lebesgue measure, these results are entirely standard. However, the case where  $\mu$  contains point masses must be handled with some care.

LEMMA 1. *Let the assumptions (A1) hold.*

(i) *Let  $\phi$  be a nonnegative solution to (3.1) with  $\phi(y) > 0$  for some  $y \in [0, R]$ . Then there exists  $\delta > 0$  such that*

$$(3.4) \quad \phi(x) \geq \delta \quad \text{for all } x \in [0, R].$$

(ii) *The problem (3.1) can have at most one nontrivial positive solution.*

*Proof.* 1. Consider the set

$$(3.5) \quad \mathcal{D} \doteq \{x \in ]0, R[; \quad \phi'(x) \text{ exists, } \mu(\{x\}) = 0\}.$$

Observe that

$$\text{meas}([0, R] \setminus \mathcal{D}) = 0.$$

Construct a standard mollifier  $\varphi : \mathbb{R} \mapsto [0, 1]$ , with

$$\varphi \in C_c^\infty, \quad \int_{-1}^1 \varphi(x) dx = 1, \quad \varphi(x) = 0 \quad \text{if } |x| \geq 1,$$

and set  $\varphi_\varepsilon(x) \doteq \varepsilon^{-1} \varphi(\varepsilon^{-1} x)$ . Given a couple of points  $x_1, x_2 \in \mathcal{D}$ , consider the test functions  $\eta_\varepsilon \doteq \chi_{[x_1, x_2]} * \varphi_\varepsilon$ , so that

$$\eta_\varepsilon(x) \doteq \int_{x_1}^{x_2} \varphi_\varepsilon(x - y) dy.$$

Inserting these test functions in (3.3) and letting  $\varepsilon \rightarrow 0$ , for every couple of points  $x_1, x_2$  such that  $\mu(\{x_1\}) = \mu(\{x_2\}) = 0$  we obtain

$$(3.6) \quad \phi'(x_2) - \phi'(x_1) = \int_{x_1}^{x_2} \phi \, d\mu - \int_{x_1}^{x_2} f(x, \phi) \phi \, dx.$$

In particular, for every point  $x \in ]0, R[$ , we have

$$(3.7) \quad \phi'(x+) - \phi'(x-) = \mu(\{x\}) \phi(x).$$

2. Assume  $\phi(\bar{x}) = 0$ , for some  $\bar{x} \in [0, R]$ . Recalling that  $\phi$  is nonnegative, we claim that

$$(3.8) \quad \phi'(\bar{x}) = 0$$

as well. Indeed, if  $\bar{x} = 0$  or  $\bar{x} = R$ , the boundary conditions (3.2) yield  $\phi'(0+) = 0$  or  $\phi'(R-) = 0$ , respectively. On the other hand, if  $0 < x < R$ , as  $x \rightarrow \bar{x}$  by (3.7) the left and right limits of  $\phi'(x)$  coincide. Therefore  $\phi'(\bar{x})$  exists. If this derivative were  $\neq 0$ , the function  $\phi$  would take values with opposite signs in a neighborhood of  $\bar{x}$ , against the nonnegativity assumption. Hence (3.8) must hold.

3. For notational convenience, define  $\psi(x) \doteq \phi'(x)$ . Introduce the variable

$$(3.9) \quad s \doteq x - \bar{x} + \mu([\bar{x}, x]).$$

The map  $x \mapsto s(x)$  is strictly increasing. It admits a nondecreasing inverse  $s \mapsto x(s)$  which is continuous with Lipschitz constant one. Namely,

$$(3.10) \quad \theta(s) \doteq \frac{d}{ds} x(s) \in [0, 1]$$

for almost every  $s$ . Observe that (3.9)–(3.10) formally yield  $ds = dx + d\mu = \theta ds + (1 - \theta)ds$ . To take care of points where  $\mu$  has a point mass, and  $\phi'$  thus has a jump, we set

$$s^+(s) \doteq \max \{ \sigma; \ x(\sigma) = x(s) \}, \quad s^-(s) \doteq \min \{ \sigma; \ x(\sigma) = x(s) \}$$

and define

$$\phi(s) \doteq \phi(x(s)), \quad \psi(s) \doteq \frac{s - s^-}{s^+ - s^-} \cdot \phi'(x(s) +) + \frac{s^+ - s}{s^+ - s^-} \cdot \phi'(x(s) -).$$

We observe that these maps provide a solution to the system of ODEs

$$(3.11) \quad \begin{cases} \frac{d}{ds} \phi(s) = \theta(s) \cdot \psi(s), \\ \frac{d}{ds} \psi(s) = (1 - \theta(s)) \cdot \phi(s) - \theta(s) \cdot f(x(s), \phi(s)) \phi(s), \end{cases}$$

with initial data

$$(3.12) \quad \phi(0) = \psi(0) = 0.$$

The right-hand side of (3.11) is bounded, Lipschitz continuous w.r.t.  $\phi, \psi$ , and measurable w.r.t.  $s$ . Therefore the only solution with initial data (3.12) is  $\phi(s) = \psi(s) = 0$  for every  $s$ .

4. Reverting to the original variables, in the previous step we have shown that if  $\phi(\bar{x}) = 0$  for some  $\bar{x} \in [0, R]$ , then  $\phi(x) \equiv 0$ . Equivalently, if  $\phi(y) > 0$  for some  $y$ , then  $\phi(x) > 0$  for all  $x \in [0, R]$ . Since the interval  $[0, R]$  is compact, this proves part (i) of the lemma.

5. To prove claim (ii) concerning uniqueness, let  $\phi_1, \phi_2$  be any two strictly positive solutions. Assume that  $\phi_1(y) < \phi_2(y)$  for some  $y \in [0, R]$ . Define

$$(3.13) \quad \lambda \doteq \max_{x \in [0, R]} \frac{\phi_2(x)}{\phi_1(x)} > 1.$$

We then have

$$(3.14) \quad \lambda \phi_1(x) \geq \phi_2(x) \quad \text{for all } x \in [0, R],$$

while

$$(3.15) \quad \lambda \phi_1(\bar{x}) = \phi_2(\bar{x})$$

for at least one point  $\bar{x} \in [0, R]$ . Define  $\varphi \doteq \lambda \phi_1 - \phi_2$ . This implies

$$(3.16) \quad \varphi(\bar{x}) = 0, \quad \varphi(x) \geq 0 \quad \text{for all } x \in [0, R].$$

Moreover,

$$(3.17) \quad \varphi'' + f(x, \phi_1(x)) \varphi - \mu \varphi = \left[ f(x, \phi_2(x)) - f(x, \phi_1(x)) \right] \phi_2(x).$$

Notice that, in a neighborhood of  $\bar{x}$ , the right-hand side of (3.17) is strictly negative, because  $f$  is strictly decreasing w.r.t.  $\phi$ . Using the auxiliary variable  $s$  as in (3.9), and setting  $\psi \doteq \varphi' = d\varphi/dx$ , we can represent  $\varphi$  as the solution to the Cauchy problem

$$(3.18) \quad \begin{cases} \frac{d}{ds} \varphi(s) = \theta(s) \cdot \psi(s), \\ \frac{d}{ds} \psi(s) = (1 - \theta(s)) \cdot \varphi(s) - \theta(s) \cdot f(x(s), \phi_1) \varphi + \theta(s) \kappa(s). \end{cases}$$

Here

$$\kappa(s) \doteq \left[ f(x(s), \phi_2(x(s))) - f(x(s), \phi_1(x(s))) \right] \phi_2(x(s)) < 0$$

as  $x(s)$  ranges in a neighborhood of the point  $\bar{x}$ .

6. Two cases should be considered. If  $0 < \bar{x} < R$ , since  $\varphi \geq 0$  we have

$$(3.19) \quad \lambda \phi_1'(\bar{x}-) - \phi_2'(\bar{x}-) \leq 0 \leq \lambda \phi_1'(\bar{x}+) - \phi_2'(\bar{x}+).$$

From the identities

$$\phi_2'(\bar{x}+) - \phi_2'(\bar{x}-) = \mu(\{\bar{x}\}) \cdot \phi_2(\bar{x}) = \mu(\{\bar{x}\}) \cdot \lambda \phi_1(\bar{x}) = \lambda \phi_1'(\bar{x}+) - \lambda \phi_1'(\bar{x}-),$$

it follows that the left- and right-hand sides of (3.19) are equal, and hence they both vanish. By the parametrization (3.9), at  $s = 0$  we have

$$(3.20) \quad \varphi(0) = \lambda \phi_1(\bar{x}) - \phi_2(\bar{x}) = 0, \quad \psi(0) = \lambda \phi_1'(\bar{x}) - \phi_2'(\bar{x}) \geq 0.$$



Since  $\kappa(s) < 0$  for  $s \in ]0, s_0]$ , with  $s_0 > 0$  small, from equations (3.18) and the initial conditions (3.20) it follows that  $\varphi(s) < 0$  for  $s > 0$  sufficiently small. This yields a contradiction with (3.16).

7. Finally, we consider the case where  $\bar{x} = 0$ . According to the boundary conditions (3.2), the identity  $\phi_2(0) = \lambda\phi_1(0)$  implies  $\phi'_2(0+) = \lambda\phi'_1(0+)$ . Therefore, for  $s = 0$  equations (3.18) should again be solved with the boundary conditions (3.20). As before, we conclude that  $\varphi(s) < 0$  for  $s > 0$  small, reaching a contradiction. The case where  $\bar{x} = R$  is entirely analogous.  $\square$

In the following, by a *subsolution* of problem (3.1) we mean a Lipschitz continuous function  $\phi$  which satisfies the following two conditions.

(i) The map  $x \mapsto \phi'(x)$  has bounded variation and satisfies

$$(3.21) \quad \lim_{x \rightarrow 0+} \phi'(x) \geq \phi(0) \mu(\{0\}), \quad \lim_{x \rightarrow R-} \phi'(x) \leq -\phi(R) \mu(\{R\}).$$

(ii) For every nonnegative test function  $\eta \in C_c^1([0, R])$  one has

$$(3.22) \quad \int_0^R \left\{ -\phi' \eta' + f(x, \phi) \phi \eta \right\} dx - \int_0^R \phi \eta d\mu \geq 0.$$

Repeating the previous analysis at (3.5)–(3.7), we see that  $\phi$  is a subsolution if and only if the boundary condition (3.21) holds and, moreover, for every  $x_1, x_2 \in ]0, R[$  such that  $\mu(\{x_1\}) = \mu(\{x_2\}) = 0$ , we have

$$(3.23) \quad \phi'(x_2) - \phi'(x_1) \geq \int_{x_1}^{x_2} \phi d\mu - \int_{x_1}^{x_2} f(x, \phi) \phi dx.$$

Supersolutions can be defined in an entirely similar way, reversing the inequalities in (3.21)–(3.22). The next result provides a necessary and sufficient condition for the existence of a strictly positive solution. Since it is an obvious extension of the corresponding result valid when  $\mu$  is absolutely continuous, we only sketch the main arguments of the proof.

LEMMA 2. *Let assumptions (A1) hold. Then the following are equivalent:*

(i) *The linear eigenvalue problem*

$$(3.24) \quad \phi'' + f(x, 0)\phi - \phi\mu = \lambda\phi, \quad \phi'(0) = \phi'(R) = 0$$

*has a strictly positive solution for some  $\lambda > 0$ .*

(ii) *The nonlinear problem (3.1) has a strictly positive solution.*

*Proof.* 1. Assume that (i) holds. Observe that the constant function

$$\phi^+(x) \equiv h_{\max} \doteq \max_{x \in [0, R]} h(x)$$

is a supersolution of (3.1). Indeed  $f(x, h_{\max}) \leq 0$  for all  $x \in [0, R]$ . On the other hand, let  $\phi_\lambda$  be a positive eigenfunction, corresponding to an eigenvalue  $\lambda > 0$ . Then, for all  $\varepsilon > 0$  sufficiently small, by the continuity of  $f$  we have

$$\varepsilon\phi_\lambda'' + f(x, \varepsilon\phi_\lambda(x)) \varepsilon\phi_\lambda - \varepsilon\phi_\lambda \mu \geq \varepsilon\phi_\lambda'' + [f(x, 0) - \lambda] \varepsilon\phi_\lambda - \varepsilon\phi_\lambda \mu = 0.$$

Hence the function  $\phi^-(x) \doteq \varepsilon\phi_\lambda(x)$  is a strictly positive subsolution of (3.1). By possibly reducing the size of  $\varepsilon$  we can assume that  $\phi^-(x) \leq h_{\max} = \phi^+(x)$ . Having constructed an upper and a lower solution, we conclude that there exists a solution  $\phi$

of (3.1) such that  $\phi^-(x) \leq \phi(x) \leq \phi^+(x)$  for all  $x \in [0, R]$ . Following a well-established technique, this solution  $\phi$  can be defined as the supremum of all subsolutions  $\leq \phi^+$ .

2. Conversely, assume that (3.1) admits a strictly positive solution  $\phi$ . In analogy with (3.9), define the new space variable

$$(3.25) \quad s = x + \mu([0, x])$$

and set  $\theta(s) = dx(s)/ds \in [0, 1]$ . We need to show that there exists  $\lambda > 0$  and a solution to the system

$$(3.26) \quad \begin{cases} \frac{d}{ds}\phi_\lambda(s) = \theta(s) \cdot \psi_\lambda(s), \\ \frac{d}{ds}\psi_\lambda(s) = (1 - \theta(s)) \cdot \phi_\lambda(s) - \theta(s) \cdot [f(x(s), 0) - \lambda] \phi_\lambda(s), \end{cases}$$

with  $\phi_\lambda > 0$  and with boundary conditions

$$(3.27) \quad \phi_\lambda(0) = 1, \quad \psi_\lambda(0) = \psi_\lambda(S) = 0.$$

Here  $S \doteq R + \mu([0, R])$ . Introducing the quotient function

$$(3.28) \quad \Phi_\lambda(s) \doteq \frac{\psi_\lambda(s)}{\phi_\lambda(s)},$$

from (3.26)–(3.27) we find that  $\Phi_\lambda$  must satisfy

$$(3.29) \quad \frac{d}{ds}\Phi_\lambda(s) = (1 - \theta(s)) - \theta(s)[f(x(s), 0) - \lambda] - \theta(s)\Phi_\lambda^2(s) \quad \text{for a.e. } s \in [0, S],$$

$$(3.30) \quad \Phi_\lambda(0) = 0,$$

together with the terminal condition  $\Phi_\lambda(S) = 0$ . For each  $\lambda \geq 0$ , consider the solution  $\Phi_\lambda$  of the Cauchy problem (3.29)–(3.30). As  $\lambda$  increases, we clearly have

$$\lim_{\lambda \rightarrow +\infty} \Phi_\lambda(S) = +\infty.$$

On the other hand, when  $\lambda = 0$  we have  $\Phi_0(S) < 0$ . Indeed, by assumption there exists a solution  $(\phi, \psi)$  of (3.11) with

$$\psi(0) = \psi(S) = 0, \quad \phi(s) > 0 \quad \text{for all } s \in [0, S].$$

Hence the function

$$\Phi(s) \doteq \psi(s)/\phi(s)$$

is well defined and satisfies

$$(3.31) \quad \frac{d}{ds}\Phi(s) = (1 - \theta(s)) - \theta(s)f(x(s), \phi(s)) - \theta(s)\Phi^2(s) \quad \text{for a.e. } s \in [0, S],$$

$$(3.32) \quad \Phi(0) = \Phi(S) = 0.$$

Recalling that  $f(x(s), \phi(s)) < f(x(s), 0)$  for every  $s \in [0, S]$ , by a standard comparison argument for first order ODEs we conclude that

$$\Phi_0(s) \leq \Phi(s) \quad \text{for all } s \in [0, S].$$

Moreover, at the terminal point we have the strict inequality  $\Phi_0(S) < \Phi(S)$ . Since the value  $\Phi_\lambda(S)$  depends continuously on  $\lambda$ , there exists a particular value  $\lambda^* > 0$  for which  $\Phi_{\lambda^*}(S) = 0$ . Returning to the original variables, this yields the desired solution to the linear boundary value problem (3.24).  $\square$

By the previous analysis, the nonlinear boundary value problem (3.1) has a strictly positive solution if and only if for  $\lambda = 0$  the solution  $\Phi_0$  of the Cauchy problem (3.29)–(3.30) satisfies  $\Phi_0(S) < 0$ . This yields the following.

**COROLLARY 1.** *A sufficient condition for problem (3.1) to have a strictly positive solution is that*

$$(3.33) \quad \mu([0, R]) < \int_0^R f(x, 0) dx.$$

Indeed, according to the proof of Lemma 2, a strictly positive solution exists if and only if the solution  $\Phi_0$  of the Cauchy problem

$$(3.34) \quad \frac{d}{ds} \Phi_0(s) = (1 - \theta(s)) - \theta(s) f(x(s), 0) - \theta(s) \Phi_0^2(s), \quad \Phi_0(0) = 0,$$

satisfies  $\Phi_0(S) < 0$ . Integrating (3.34) one obtains

$$\Phi_0(S) \leq \int_0^S (1 - \theta(s)) ds - \int_0^S \theta(s) f(x(s), 0) ds = \mu([0, R]) - \int_0^R f(x, 0) dx.$$

This establishes (3.33).

**4. The optimal control problem.** In this section we analyze the optimal control problem for one fishing company. This is formulated as an optimization problem within a space of nonnegative Radon measures for an elliptic PDE with Neumann boundary data. In one space dimension it takes the form

$$(4.1) \quad \text{maximize:} \quad J(\mu) = \int_0^R \phi(x) d\mu(x) - \Psi \left( \int_0^R c(x) d\mu(x) \right)$$

subject to

$$(4.2) \quad \phi''(x) + f(x, \phi) \phi - \phi \cdot \nu = \phi \cdot \mu, \quad x \in ]0, R[,$$

$$(4.3) \quad \phi'(0) = \phi'(R) = 0.$$

Here the measure  $\mu$  describes the fishing intensity for the particular company under consideration. This will be an element of the space  $\mathcal{M}([0, R])$  of Radon measures supported on the compact interval  $[0, R]$ . The measure  $\mu$  must be suitably chosen in order to maximize the profit  $J(\mu)$ . On the other hand,  $\nu$  is a given Radon measure, accounting for the combined intensity of fishing of all other companies. This situation is relevant in the study of differential games. Of course, if only one fishing company is active, we would simply take  $\nu \equiv 0$ .

On the cost functional  $J$  we make the following assumptions:

(A2) The cost function  $c : [0, R] \mapsto [c_0, \infty]$  is bounded from below by some constant  $c_0 > 0$  and is lower semicontinuous. The function  $\Psi$  is nondecreasing, convex, and lower semicontinuous and satisfies

$$(4.4) \quad \Psi(0) = 0, \quad \Psi'(0) = 1.$$

Finally,  $\nu$  is a nonnegative Radon measure on  $[0, R]$ .

Notice that the case  $c(x) = +\infty$  for every  $x$  in an open subset  $\Omega' \subset \Omega$  is allowed. Similarly, there may exist a value  $s_0 > 0$  such that  $\Psi(s) = +\infty$  for all  $s > s_0$ . Some preliminary observations are listed below.

*Remark 3.* For every nonnegative measures  $\mu, \nu \geq 0$ , the solution of (4.2)–(4.3) satisfies

$$(4.5) \quad 0 \leq \phi(x) \leq \max_{x \in [0, R]} h(x),$$

where  $h$  is the function introduced in assumption (A1).

*Remark 4.* Using the fact that the pointwise maximum of two solutions is a subsolution, we obtain the uniqueness of the maximal solution. Throughout the following, for a given control measure  $\mu$ , we shall always refer to this maximal solution. According to Lemma 1, this maximal solution is either identically zero or uniformly positive on the entire domain  $[0, R]$ .

*Remark 5.* Given two measures  $\mu, \tilde{\mu}$ , we write  $\tilde{\mu} \prec \mu$  if

$$\tilde{\mu}(A) \leq \mu(A)$$

for every open set  $A$ . In this case, the corresponding maximal solutions of (4.2)–(4.3) satisfy

$$(4.6) \quad \tilde{\phi}(x) \geq \phi(x), \quad x \in [0, R].$$

Indeed, the function  $\tilde{\phi}$  is then a supersolution of (4.2)–(4.3).

*Remark 6.* Let  $\phi_0$  be the equilibrium population of the fish if our specific company does no harvesting at all, so that

$$(4.7) \quad \phi_0''(x) + f(x, \phi_0)\phi_0 - \phi_0 \cdot \nu = 0, \quad x \in \Omega.$$

Then the optimal control strategy is  $\mu^* \equiv 0$  if and only if

$$(4.8) \quad \phi_0(x) \leq c(x) \quad \text{for all } x \in [0, R].$$

*Remark 7.* In the spatially homogeneous case  $c(x) \equiv \bar{c}$ ,  $d\nu = \bar{\nu} dx$ ,  $f(x, \phi) = (\bar{h} - \phi)$ , we expect to find a optimal solution  $\mu$  having constant density  $\bar{u}$  w.r.t. Lebesgue measure, so that  $d\mu = \bar{u} dx$ . In this case, the solution of (4.2)–(4.3) is explicitly computed:

$$(4.9) \quad \phi \equiv \bar{h} - \bar{\nu} - \bar{u};$$

otherwise  $\phi \equiv 0$  if the right-hand side of (4.9) is negative. The optimal value of  $\bar{u}$  is found by maximizing the scalar function

$$J(\bar{u}) = R(\bar{h} - \bar{\nu} - \bar{u}) - \Psi(R\bar{c}\bar{u}).$$

The main result of this section is the existence of an optimal strategy  $\mu^*$ , within the space of Radon measures. As we show in a subsequent section, this more general formulation cannot be avoided. Indeed, when the cost function  $c(\cdot)$  is discontinuous, the optimization problem may have no classical solution. The optimal strategy  $\mu^*$  is typically a measure which contains point masses and does not admit a density  $u^* \in \mathbf{L}^1([0, R])$  w.r.t. Lebesgue measure.

THEOREM 1. *Let assumptions (A1)–(A2) hold. Then the maximization problem (4.1)–(4.3) admits an optimal solution  $\mu^* \in \mathcal{M}_+([0, R])$  within the family of nonnegative Radon measures on  $[0, R]$ .*

*Proof.* 1. Let  $(\mu_n)_{n \geq 1}$  be a maximizing sequence, so that  $\mu_n \in \mathcal{M}_+([0, R])$  for all  $n \geq 1$  and

$$(4.10) \quad \lim_{n \rightarrow \infty} J(\mu_n) = \sup_{\mu} J(\mu).$$

Call  $\phi_n$  the corresponding maximal solution of (4.2)–(4.3). Notice that it is not restrictive to assume that the support of  $\mu_n$  is contained in the set

$$\Omega_n \doteq \{x \in [0, R]; \phi_n(x) \geq c_0\}.$$

Indeed, in the opposite case we can define a new measure  $\tilde{\mu}_n \prec \mu_n$  by setting

$$\tilde{\mu}_n(A) \doteq \mu_n(A \cap \Omega_n).$$

As in Remark 5, the corresponding solutions satisfy  $\tilde{\phi}_n \geq \phi_n$ , and hence  $J(\tilde{\mu}_n) \geq J(\mu_n)$ .

2. From (4.2)–(4.3) it now follows that

$$(4.11) \quad c_0 \mu_n([0, R]) \leq \int_0^R \phi_n d\mu_n \leq \int_0^R f(x, \phi_n) \phi_n dx \leq \int_0^R f(x, 0) h_{max} dx.$$

This proves that the sequence  $\mu_n$  is uniformly bounded.

3. By compactness, we can now select a subsequence, still called  $(\mu_n)$ , which converges weakly to a measure  $\mu^* \in \mathcal{M}_+([0, R])$ . Since the functions  $\phi_n$  are uniformly Lipschitz continuous and bounded on  $[0, R]$ , we have the uniform convergence  $\phi_n \rightarrow \phi^*$ , where  $\phi^*$  provides a solution to (4.2)–(4.3) with  $\mu = \mu^*$ . We now have

$$(4.12) \quad \lim_{n \rightarrow \infty} \int_0^R \phi_n d\mu_n = \int_0^R \phi^* d\mu^*.$$

Moreover, the lower semicontinuity of the functions  $c$  and  $\Psi$  implies

$$(4.13) \quad \begin{aligned} \int_0^R c d\mu^* &\leq \liminf_{n \rightarrow \infty} \int_0^R c d\mu_n, \\ \Psi \left( \int_0^R c d\mu^* \right) &\leq \liminf_{n \rightarrow \infty} \Psi \left( \int_0^R c d\mu_n \right). \end{aligned}$$

Together, (4.12)–(4.13) yield

$$\lim_{n \rightarrow \infty} J(\mu_n) \leq J(\mu^*).$$

Hence the strategy  $\mu^*$  is optimal.  $\square$

**5. Uniqueness of optimal solutions.** This section is concerned with the uniqueness of optimal solutions. We will show that, within a class of measures with small total mass, the optimal strategy is unique. This will follow from the strict concavity of the payoff functional  $J$ . Throughout the following, we denote by

$$(5.1) \quad \|\mu\| \doteq \sup \left\{ \int_0^R \varphi d\mu; \|\varphi\|_{C^0} \leq 1 \right\}$$

the total mass of a measure  $\mu$ . Notice that if  $\mu$  is nonnegative, one simply has  $\|\mu\| = \mu([0, R])$ . Given two distinct positive measures  $\mu, \tilde{\mu} \in \mathcal{M}_+([0, R])$ , define

$$\sigma \doteq \frac{\tilde{\mu} - \mu}{\|\tilde{\mu} - \mu\|}.$$

Of course,  $\sigma$  is a Radon measure, not necessarily positive, with unit norm.

For  $\epsilon \geq 0$  small, let  $\phi^\epsilon$  be the corresponding solution of boundary value problem (4.2)–(4.3), with  $\mu$  replaced by  $\mu^\epsilon \doteq \mu + \epsilon\sigma$ . We seek conditions which ensure that the scalar map

$$(5.2) \qquad \epsilon \mapsto J(\mu + \epsilon\sigma)$$

is strictly concave. It is clear that the map

$$\epsilon \mapsto \int_0^R c \, d\mu^\epsilon = \int_0^R c \, d\mu + \epsilon \int_0^R c \, d\sigma$$

is affine. Since we are assuming that the function  $s \mapsto \Psi(s)$  is convex, the same is true of the map

$$\epsilon \mapsto J^-(\mu^\epsilon) \doteq \Psi\left(\int_0^R c \, d\mu^\epsilon\right).$$

It thus suffices to study the concavity of the map

$$(5.3) \qquad \epsilon \mapsto J^+(\mu^\epsilon) \doteq \int_0^R \phi^\epsilon \, d\mu^\epsilon$$

for  $\epsilon \geq 0$  small. In the following analysis, we assume

$$(5.4) \qquad \|\nu\|, \|\mu\| << 1,$$

so that the corresponding solution of (4.2)–(4.3) will be close to the solution  $\psi_0$  of

$$(5.5) \qquad \psi_0'' + f(x, \psi_0) \psi_0 = 0, \qquad \psi_0'(0) = \psi_0'(R) = 0,$$

where no fishing occurs.

We begin with a formal analysis. Assume that at  $\epsilon \approx 0$ , the map  $\epsilon \mapsto \phi^\epsilon$  admits an asymptotic expansion

$$(5.6) \qquad \phi^\epsilon = \phi_0 + \epsilon\phi_1 + \epsilon^2\phi_2 + o(\epsilon^2).$$

Observe that each function  $\phi^\epsilon$  satisfies

$$(5.7) \qquad (\phi^\epsilon)'' + f(x, \phi^\epsilon)\phi^\epsilon = \phi^\epsilon(\nu + \mu + \epsilon\sigma),$$

with Neumann boundary conditions (4.3). Inserting (5.6) in (5.7) and expanding in powers of  $\epsilon$  we see that the functions  $\phi_0, \phi_1, \phi_2 : [0, R] \mapsto \mathbb{R}$  should provide solutions to the boundary value problems

$$(5.8) \quad \left\{ \begin{array}{l} \phi_0'' + f(x, \phi_0)\phi_0 = \phi_0(\nu + \mu), \\ \phi_1'' + f_\phi(x, \phi_0)\phi_0\phi_1 + f(x, \phi_0)\phi_1 = \phi_1(\nu + \mu) + \phi_0\sigma, \\ \phi_2'' + \frac{1}{2}f_{\phi\phi}(x, \phi_0)\phi_0\phi_1^2 + f_\phi(x, \phi_0)(\phi_0\phi_2 + \phi_1^2) + f(x, \phi_0)\phi_2 = \phi_2(\nu + \mu) + \phi_1\sigma, \end{array} \right.$$

with Neumann boundary conditions

$$(5.9) \quad \phi'_0(0) = \phi'_1(0) = \phi'_2(0) = 0, \quad \phi'_0(R) = \phi'_1(R) = \phi'_2(R) = 0.$$

Computing the second derivative at  $\epsilon = 0$ , we find

$$(5.10) \quad \begin{aligned} \frac{1}{2} \frac{d^2}{d\epsilon^2} J^+(\mu^\epsilon) &= \frac{1}{2} \frac{d^2}{d\epsilon^2} \int_0^R \phi^\epsilon d\mu^\epsilon \\ &= \frac{1}{2} \frac{d^2}{d\epsilon^2} \int_0^R (\phi_0 + \epsilon \phi_1 + \epsilon^2 \phi_2) d(\mu + \epsilon \sigma) \\ &= \int_0^R \phi_1 d\sigma + \int_0^R \phi_2 d\mu. \end{aligned}$$

We need to study the sign of the right-hand side of (5.10) and check if it is strictly negative. From the first two equations in (5.8) it follows that

$$(5.11) \quad \nu + \mu = \frac{\phi''_0}{\phi_0} + f(x, \phi_0),$$

$$(5.12) \quad \sigma = \frac{\phi''_1}{\phi_0} + f_\phi(x, \phi_0) \phi_1 - \phi_1 \frac{\phi''_0}{\phi_0^2}.$$

Integrating by parts and using the boundary conditions (5.9), we obtain

$$(5.13) \quad \begin{aligned} \int_0^R \phi_1 d\sigma &= \int_0^R \left( \frac{\phi''_1}{\phi_0} \phi_1 + f_\phi(x, \phi_0) \phi_1^2 - \phi_1^2 \frac{\phi''_0}{\phi_0^2} \right) dx \\ &= \int_0^R \left[ -\phi'_1 \left( \frac{\phi_1}{\phi_0} \right)' + f_\phi(x, \phi_0) \phi_1^2 + \left( \frac{\phi_1^2}{\phi_0^2} \right)' \phi'_0 \right] dx \\ &= \int_0^R \left[ -\frac{1}{\phi_0} (\phi'_1)^2 + 3 \frac{\phi'_0}{\phi_0^2} \phi'_1 \phi_1 + \left( f_\phi(x, \phi_0) - 2 \frac{(\phi'_0)^2}{\phi_0^3} \right) \phi_1^2 \right] dx. \end{aligned}$$

Since we always assume (5.4), in (5.8)–(5.9) we expect  $\phi_0 \approx \psi_0$ , where  $\psi_0$  is the solution of (5.5). On the other hand, the function  $\phi_1$  can be essentially arbitrary. In order to control the sign of the right-hand side of (5.13), on the Hilbert–Sobolev space  $H^1([0, R])$ , we consider the homogeneous quadratic functional

$$(5.14) \quad Q(v) = \int_0^R \left[ \frac{1}{\psi_0} (v')^2 - 3 \frac{\psi'_0}{\psi_0^2} v'v + \left( 2 \frac{(\psi'_0)^2}{\psi_0^3} - f_\phi(x, \psi_0) \right) v^2 \right] dx.$$

By an elementary inequality we see that if there exists  $\delta_0 > 0$  such that

$$(5.15) \quad f_\phi(x, \psi_0) + \frac{1}{4} \frac{[\psi'_0(x)]^2}{\psi_0^3(x)} \leq -\delta_0 < 0 \quad \text{for all } x \in [0, R],$$

then the functional  $Q$  is strictly positive definite. Namely,

$$(5.16) \quad Q(v) \geq 2\delta_1 \|v\|_{H^1}^2$$

for some constant  $\delta_1 > 0$  and all  $v \in H^1$ . For  $\nu, \mu \approx 0$  one has  $\phi_0 \approx \psi_0$ . Therefore the right-hand side of (5.13) will still be strictly negative definite. Concerning the second integral on the right-hand side of (5.10), we expect that it can be rendered arbitrarily

small by choosing the measure  $\mu$  small enough. These preliminary computations motivate the following.

**THEOREM 2.** *Let assumptions (A1)–(A2) hold. Moreover, assume that (5.15) holds, so that the quadratic functional  $Q$  in (5.14) is strictly positive definite. Then there exists a constant  $\delta^\sharp > 0$  such that, for any given measure  $\nu \in \mathcal{M}_+$ , the optimization problem (4.1)–(4.3) can have at most one solution within the set of measures  $\mu \in \mathcal{M}_+([0, R])$  which satisfy the additional condition*

$$(5.17) \qquad \qquad \qquad \|\mu\| + \|\nu\| \leq \delta^\sharp.$$

*In particular, if  $\nu = 0$  and*

$$(5.18) \qquad \qquad \qquad \frac{\Psi(c_0\delta^\sharp)}{\delta^\sharp} > h_{max},$$

*then the optimization problem has exactly one solution  $\mu \in \mathcal{M}_+$ .*

*Proof.* Assume that there exist two distinct optimal solutions  $\mu_1, \mu_2$ . To obtain a contradiction, we will show that the map

$$(5.19) \qquad \qquad \qquad \epsilon \mapsto J(\epsilon\mu_2 + (1 - \epsilon)\mu_1)$$

is strictly concave. This will be achieved in several steps.

1. Fix any  $\bar{\epsilon} \in [0, 1]$  and define

$$\mu \doteq \bar{\epsilon}\mu_2 + (1 - \bar{\epsilon})\mu_1, \qquad \sigma \doteq \frac{\mu_2 - \mu_1}{\|\mu_2 - \mu_1\|}.$$

We consider the map  $s \mapsto J(\mu + s\sigma)$  and check that, at  $s = 0$ ,

$$\frac{d^2}{ds^2} J(\mu + s\sigma) < 0.$$

Indeed, let  $\phi_0, \phi_1, \phi_2$  be the solutions of (5.8)–(5.9). Notice that, for  $\|\nu + \mu\|$  sufficiently small, these functions are uniquely defined. Call  $\phi^s$  the solution of (4.2)–(4.3) corresponding to the measure  $\mu + s\sigma$ . We then have

$$\lim_{s \rightarrow 0} \frac{1}{s^2} \|\phi^s - \phi_0 - s\phi_1 - s^2\phi_2\|_{C^0} = 0.$$

Hence

$$(5.20) \qquad \qquad \qquad \left. \frac{d^2}{ds^2} J(\mu + s\sigma) \right|_{s=0} = \int_0^R \phi_1 \, d\sigma + \int_0^R \phi_2 \, d\mu.$$

2. We now rewrite the third equation in (5.8) as

$$(5.21) \qquad \qquad \qquad \phi_2'' + \left( f_\phi(x, \phi_0)\phi_0 + f(x, \phi_0) - \nu - \mu \right) \phi_2 = \phi_1\sigma - \left( \frac{1}{2} f_{\phi\phi}(x, \phi_0)\phi_0 + f_\phi(x, \phi_0) \right) \phi_1^2.$$

This is a linear, nonhomogeneous equation for  $\phi_2$ . Recalling (5.12), its solution can be written in the form

$$(5.22) \qquad \qquad \qquad \begin{aligned} \phi_2(x) = \int_0^R K(x, y) \cdot & \left[ \left( \frac{\phi_1''}{\phi_0} + f_\phi(y, \phi_0)\phi_1 - \phi_1 \frac{\phi_0''}{\phi_0^2} \right) \phi_1 \right. \\ & \left. - \left( \frac{1}{2} f_{\phi\phi}(y, \phi_0)\phi_0 + f_\phi(y, \phi_0) \right) \phi_1^2 \right] dy. \end{aligned}$$



Here  $K(x, y)$  is a Green kernel and all terms inside the square brackets are evaluated at the point  $y$ . Integrating by parts we obtain

$$\begin{aligned}
 \phi_2(x) &= \int_0^R K(x, y) \cdot \left[ \phi_1'' \frac{\phi_1}{\phi_0} - \phi_0'' \frac{\phi_1^2}{\phi_0^2} - \frac{1}{2} f_{\phi\phi}(y, \phi_0) \phi_0 \phi_1^2 \right] dy \\
 &= - \int_0^R \left( K_y \frac{\phi_1}{\phi_0} + K \frac{\phi_1'}{\phi_0} - K \frac{\phi_1 \phi_0'}{\phi_0^2} \right) \phi_1' dy \\
 &\quad + \int_0^R \left( K_y \frac{\phi_1^2}{\phi_0^2} + K \frac{2\phi_1 \phi_1'}{\phi_0^2} - 2K \frac{\phi_1^2 \phi_0'}{\phi_0^3} \right) \phi_0' dy \\
 &\quad - \int_0^R \frac{1}{2} K f_{\phi\phi}(y, \phi_0) \phi_0 \phi_1^2 dy.
 \end{aligned}
 \tag{5.23}$$

We now observe that  $\phi_0$  is uniformly bounded and Lipschitz continuous, and bounded away from zero. Similarly, the kernel  $K$  is uniformly bounded and Lipschitz continuous w.r.t. both variables  $x, y$ . From (5.21) we thus derive an estimate of the form

$$\|\phi_2\|_{C^0} \leq C_2 \int_0^R [(\phi_1')^2 + \phi_1^2] dx
 \tag{5.24}$$

for some constant  $C_2$ .

3. Since the quadratic form  $Q$  in (5.14) by assumption is strictly positive definite, by continuity we can assume that, for all  $\|\mu\| \leq 2\delta^\sharp$ , the corresponding quadratic form is also strictly positive definite, namely,

$$\begin{aligned}
 Q^{\nu+\mu}(v) &\doteq \int_0^R \left( \frac{1}{\phi_0} (v')^2 - 3 \frac{\phi_0'}{\phi_0^2} v' v + \left( 2 \frac{(\phi_0')^2}{\phi_0^3} - f_\phi(x, \phi_0) \right) v^2 \right) dx \\
 &\geq \varepsilon_0 \int_0^R [(v')^2 + v^2] dx
 \end{aligned}
 \tag{5.25}$$

for some constant  $\varepsilon_0 > 0$ . Here  $\phi_0$  is the solution of the first equation in (5.8). Of course, as  $\nu + \mu \rightarrow 0$ , we have  $\phi_0 \rightarrow \psi_0$ .

Going back to the expression (5.10) for the second derivative, we obtain the estimate

$$\begin{aligned}
 \frac{d^2}{ds^2} J(\mu + s\sigma) \Big|_{s=0} &= \int_0^R \phi_1 d\sigma + \int_0^R \phi_2 d\mu \\
 &\leq -Q^{\nu+\mu}(\phi_1) + \|\phi_2\|_{C^0} \|\mu\| \\
 &\leq (-\varepsilon_0 + C_2 \|\mu\|) \int_0^R [(\phi_1')^2 + \phi_1^2] dx.
 \end{aligned}
 \tag{5.26}$$

If  $\|\mu\| < \varepsilon_0/C_2$ , then either  $\phi_1 \equiv 0$  or the left-hand side in (5.26) is strictly negative. We conclude that, at  $\epsilon = \bar{\epsilon}$ , either the first derivative  $dJ(\mu^\epsilon)/d\epsilon$  vanishes or else the second derivative is strictly negative. Since  $\bar{\epsilon} \in [0, 1]$  was arbitrary, this shows that there can be at most one optimal solution within the set of measures  $\mu$  such that  $\|\nu + \mu\| \leq \delta^\sharp$ .

4. To prove the last statement, assume that  $\nu = 0$  and (5.16) holds. In this case, we claim that any optimal solution  $\mu$  satisfies  $\|\mu\| \leq \delta^\sharp$ . Indeed, if  $\|\mu\| > \delta^\sharp$ , observing that  $\phi(x) \leq h_{max}$  and  $c(x) \geq c_0$  for all  $x \in [0, R]$ , and using the convexity

of the function  $\Psi$ , we compute

$$(5.27) \quad \int_0^R \phi(x) d\mu - \Psi \left( \int_0^R c(x) d\mu \right) \leq h_{\max} \|\mu\| - \Psi(c_0 \|\mu\|) \\ \leq h_{\max} \|\mu\| - \Psi(c_0 \delta^\#) \cdot \frac{\|\mu\|}{\delta^\#} < 0.$$

By (5.27) the measure  $\mu$  achieves a negative payoff and is not optimal, being worse than the zero measure, which achieves a null payoff.  $\square$

**6. Solutions to the differential game.** The aim of this section is to establish the existence of measure-valued, Nash equilibrium solutions to a differential game with  $m$  players. Denoting by  $\mu_i$  the intensity of fishing by the  $i$ th company, the density of fish population will satisfy

$$(6.1) \quad \phi'' + f(x, \phi)\phi = \phi \cdot \sum_{i=1}^m \mu_i, \quad \phi'(0) = \phi'(R) = 0.$$

We always assume that the function  $f$  satisfies assumptions (A1). The goal of the  $i$ th player is to maximize his payoff

$$(6.2) \quad J_i \doteq \int_0^R \phi(x) d\mu_i dx - \Psi \left( \int_0^R c_i(x) d\mu_i(x) \right)$$

among all nonnegative measures  $\mu_i \in \mathcal{M}_+$  on the closed interval  $[0, R]$ , subject to the constraint

$$(6.3) \quad \|\mu_i\| = \mu_i([0, R]) \leq \delta_i.$$

Notice that by (6.3) we impose an upper bound on the total amount of fishing activity carried out by the  $i$ th company. In practice, this limitation is due to the finite number of boats, fishermen, and working hours.

**DEFINITION 3.** *By a Nash equilibrium solution we mean an  $m$ -tuple of nonnegative Radon measures  $\mu = (\mu_1, \dots, \mu_m)$  such that, for each  $i = 1, \dots, m$ , the following holds. Setting  $\nu \doteq \sum_{j \neq i} \mu_j$ , the measure  $\sigma = \mu_i$  provides a solution to the optimization problem*

$$(6.4) \quad \text{maximize:} \quad J_i(\sigma) = \int_0^R \phi d\sigma - \Psi_i \left( \int_0^R c_i(x) d\sigma \right),$$

*subject to the bound (6.3) and with*

$$(6.5) \quad \phi''(x) + f(x, \phi)\phi - \phi \cdot \nu = \phi \cdot \sigma, \quad \phi'(0) = \phi'(R) = 0.$$

On the cost functionals  $J_i$  we make the following assumptions:

(A2) <sub>$i$</sub>  The cost function  $c_i : [0, R] \mapsto [c_0, \infty]$  is lower semicontinuous and strictly positive. Moreover, the function  $\Psi_i$  is convex and lower semicontinuous and satisfies

$$(6.6) \quad \Psi_i(0) = 0, \quad \Psi'_i(0) = 1.$$

THEOREM 3. *Let assumptions (A1) hold, together with (A2)<sub>i</sub>, for every  $i = 1, \dots, m$ . Moreover, assume that the solution  $\psi_0$  of (5.5) satisfies (5.15). Then there exists  $\delta > 0$  such that if*

$$(6.7) \quad \sum_{i=1}^m \delta_i \leq \delta,$$

*then the differential game (6.1)–(6.3) admits a Nash equilibrium solution.*

*Proof.* 1. By the analysis in section 4, for a given measure  $\nu$ , the optimization problem (6.3)–(6.5) for the  $i$ th player has at least one solution. The presence of the additional constraint (6.3) actually simplifies the proof, because the upper bound on  $\|\mu_i\|$  is now explicitly assumed. If

$$\|\mu_i\| + \|\nu\| = \|\mu_i\| + \sum_{j \neq i} \|\mu_j\| \leq \sum_{j=1}^m \delta_i$$

is sufficiently small, by Theorem 2 this optimal solution  $\mu_i$  is unique. We can thus write  $\mu_i = \mathcal{T}_i(\nu)$ , for a suitable transformation  $\mathcal{T}_i$  in the space of positive Radon measures.

2. Consider the compact set  $\mathcal{K}$  consisting of  $m$ -tuples of nonnegative measures  $\mu = (\mu_1, \dots, \mu_m)$  such that

$$\|\mu_i\| \leq \delta_i \quad i = 1, \dots, m.$$

Define the transformation  $\mathcal{T} : \mathcal{K} \mapsto \mathcal{K}$  by setting

$$\mathcal{T}(\mu) = (\mathcal{T}_1(\nu_1), \dots, \mathcal{T}_m(\nu_m)).$$

Here  $\mathcal{T}_i(\nu_i)$  is the measure providing the unique optimal solution to the problem (6.3)–(6.5), with  $\nu = \nu_i \doteq \sum_{j \neq i} \mu_j$ .

By the previous analysis, the map  $\mathcal{T}$  is a well-defined transformation of  $\mathcal{K}$  into itself. We claim that  $\mathcal{T}$  is continuous w.r.t. the weak convergence of measures. Indeed, consider a sequence of  $m$ -tuples of Radon measures  $(\mu_{1,n}, \dots, \mu_{m,n})_{n \geq 1}$ , and assume the weak convergence  $\mu_{i,n} \rightharpoonup \mu_i$  as  $n \rightarrow \infty$  for each  $i = 1, \dots, m$ . Of course, this implies the weak convergence

$$(6.8) \quad \nu_{i,n} \doteq \sum_{j \neq i} \mu_{j,n} \rightharpoonup \nu_i \doteq \sum_{j \neq i} \mu_j.$$

For each  $n \geq 1$ , let  $\sigma_{i,n} = \mathcal{T}_i(\nu_{i,n})$  be the unique measure that optimizes the corresponding problem (6.3)–(6.5), with  $\nu = \nu_{i,n}$ . Moreover, let  $\sigma_i$  be the measure which provides the optimal solution of (6.3)–(6.5) when  $\nu = \nu_i$ . We claim that the weak convergence  $\sigma_{i,n} \rightharpoonup \sigma_i$  holds. By a compactness argument, by possibly taking a subsequence we can assume that  $\sigma_{i,n} \rightharpoonup \sigma_{i,\infty}$  for some Radon measure  $\sigma_{i,\infty}$  with  $\|\sigma_{i,\infty}\| \leq \delta_i$ . To prove our claim, it suffices to show that  $\sigma_{i,\infty}$  provides an optimal solution to the problem of (6.3)–(6.5) when  $\nu = \nu_i$ . Indeed, the uniqueness result proved in section 5 will then imply  $\sigma_i = \sigma_{i,\infty}$ .

By the Ascoli–Arzelà theorem, we can assume the convergence  $\phi_n \rightarrow \phi_\infty$  of the corresponding solutions of (6.5), uniformly on  $[0, R]$ . Observe that  $\phi_\infty$  provides the solution of (6.5), with  $\nu = \nu_i$  and  $\sigma = \sigma_{i,\infty}$ . In particular,

$$(6.9) \quad \lim_{n \rightarrow \infty} \int_0^R \phi_n d\sigma_{i,n} = \int_0^R \phi_\infty d\sigma_{i,\infty}.$$

Using the assumptions of lower semicontinuity, we obtain

$$(6.10) \quad \Psi \left( \int_0^R c_i(x) d\sigma_{i,\infty} \right) \leq \liminf_{n \rightarrow \infty} \Psi \left( \int_0^R c_i(x) d\sigma_{i,n} \right).$$

Now let  $\hat{\phi}_n$  be the solution of (6.5) with  $\nu = \nu_{i,n}$  and  $\sigma = \sigma_i$ . Observe that  $\hat{\phi}_n \rightarrow \phi_\infty$  uniformly on  $[0, R]$ . Moreover, when  $\nu = \nu_{i,n}$  the measure  $\sigma_{i,n}$  performs better than  $\sigma_i$ . Therefore

$$\begin{aligned} & \int_0^R \phi_\infty d\sigma_{i,\infty} - \Psi \left( \int_0^R c_i(x) d\sigma_{i,\infty} \right) \\ & \geq \limsup_{n \rightarrow \infty} \left[ \int_0^R \phi_n d\sigma_{i,n} - \Psi \left( \int_0^R c_i(x) d\sigma_{i,n} \right) \right] \\ & \geq \lim_{n \rightarrow \infty} \left[ \int_0^R \hat{\phi}_n d\sigma_i - \Psi \left( \int_0^R c_i(x) d\sigma_i \right) \right] \\ & = \int_0^R \phi_\infty d\sigma_i - \Psi \left( \int_0^R c_i(x) d\sigma_i \right). \end{aligned}$$

The above inequalities show that the strategy  $\sigma_{i,\infty}$  achieves a payoff at least as good as  $\sigma_i$ . Hence it is optimal. By uniqueness, we conclude that  $\sigma_{i,\infty} = \sigma_i$ , as claimed. This establishes the continuity of the transformation  $\mathcal{T}$ , w.r.t. the topology of weak convergence of measures.

3. We now observe that  $\mathcal{T}$  is a continuous map from the compact, convex set  $\mathcal{K}$  into itself. By Schauder's fixed point theorem, it admits at least one fixed point  $\mu^* = (\mu_1^*, \dots, \mu_m^*)$ . By definition, this provides the required Nash equilibrium solution to the differential game.  $\square$

**7. Concluding remarks.** In this paper, we observed that a natural formulation of the optimal harvesting problem involves cost functionals with sublinear growth. As a consequence, the optimal strategies can be measure-valued. We expect that this optimal measure  $\mu^*$  will indeed be singular w.r.t. Lebesgue measure when the cost function  $c = c(x)$  is discontinuous. Results in this direction should be obtained by deriving necessary conditions for optimality.

Another issue that deserves further investigation is the range of validity of the uniqueness result. Here we established uniqueness of optimal strategies within a class of measures with small total mass. This result seems far from optimal. For applications, it would be useful to also cover the case of large-size control measures  $\mu^*$ . From our analysis, uniqueness appears to be related to the smallness of the gradient  $\nabla\phi$  of the solution. This might yield other forms of the uniqueness result. It would also be interesting to study the case of loss of uniqueness, looking for counterexamples in the case where  $\nabla\phi$  has large oscillations.

## REFERENCES

- [1] O. ARINO AND J. A. MONTERO, *Optimal control of a nonlinear elliptic population system*, Proc. Edinburgh Math. Soc., 43 (2000), pp. 225–241.
- [2] L. BOCCARDO, *Elliptic and parabolic differential problems with measure data*, Boll. Un. Mat. Ital. A (7) , 11 (1997), pp. 439–461 (in Italian).

- [3] L. BOCCARDO AND T. GALLOUËT, *Nonlinear elliptic and parabolic equations involving measure data*, J. Funct. Anal., 87 (1989), pp. 149–169.
- [4] L. BOCCARDO AND T. GALLOUËT, *Nonlinear elliptic equations with right hand side measures*, Comm. Partial Differential Equations, 17 (1992), pp. 641–655.
- [5] L. BOCCARDO, T. GALLOUËT, AND L. ORSINA, *Existence and uniqueness of entropy solutions for nonlinear elliptic equations with measure data*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 13 (1996), pp. 539–551.
- [6] A. BRESSAN AND F. RAMPAZZO, *On differential systems with vector-valued impulsive controls*, Boll. Un. Mat. Ital. B (7), 2 (1988), pp. 641–656.
- [7] A. CAÑADA, J. L. GÁMEZ, AND J. A. MONTERO, *Study of an optimal control problem for diffusive nonlinear elliptic equations of logistic type*, SIAM J. Control Optim., 36 (1998), pp. 1171–1189.
- [8] M. DELGADO, J. A. MONTERO, AND A. SUÁREZ, *Optimal control for the degenerate elliptic logistic equation*, Appl. Math. Optim., 45 (2002), pp. 325–345.
- [9] M. DELGADO, J. A. MONTERO, AND A. SUÁREZ, *Study of the optimal harvesting control and the optimality system for an elliptic problem*, SIAM J. Control Optim., 42 (2003), pp. 1559–1577.
- [10] S. M. LENHART AND J. A. MONTERO, *Optimal control of harvesting in a parabolic system modeling two subpopulations*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1129–1141.
- [11] S. LENHART AND J. T. WORKMAN, *Optimal Control Applied to Biological Models*, Chapman & Hall/CRC, Boca Raton, FL, 2007.
- [12] M. G. NEUBERT, *Marine reserves and optimal harvesting*, Ecology Lett., 6 (2003), pp. 843–849.
- [13] T. ROUBICEK, *Noncooperative games with elliptic systems*, in Optimal Control of Partial Differential Equations, Internat. Ser. Numer. Math. 133, Birkhäuser, Basel, 1999, pp. 245–255.

## LINEAR QUADRATIC DIFFERENTIAL GAMES: CLOSED LOOP SADDLE POINTS\*

MICHEL C. DELFOUR<sup>†</sup> AND OLIVIER DELLO SBARBA<sup>†</sup>

**Abstract.** The object of this paper is to revisit the results of Bernhard [*J. Optim. Theory Appl.*, 27 (1979), pp. 51–69] on two-person zero-sum linear quadratic differential games and generalize them to utility functions without positivity assumptions on the matrices acting on the state variable and to linear dynamics with bounded measurable data matrices. Our paper specializes to *state feedback* via Lebesgue measurable *affine closed loop strategies* with possible non- $L^2$ -integrable singularities. After sharpening the recent results of Delfour [*SIAM J. Control Optim.*, 46 (2007), pp. 750–774] on the characterization of the open loop lower and upper values of the game, it first deals with  $L^2$ -integrable closed loop strategies and then with the larger family of strategies that may have non- $L^2$ -integrable singularities. A new conceptually meaningful and mathematically precise definition of a closed loop saddle point is introduced to simultaneously handle state feedbacks of the  $L^2$  type and smooth locally bounded ones, except at most in the neighborhood of finitely many instants of time. A necessary and sufficient condition is that the free end problem be *normalizable almost everywhere*. This relaxation of the classical notion allows singularities in the feedback law at an infinite number of instants, including accumulation points that are not isolated. A complete classification of closed loop saddle points is given in terms of the convexity/concavity properties of the utility function, and connections are given with the open loop lower value, upper value, and value of the game.

**Key words.** linear quadratic differential game, two person, zero sum, saddle point, value of a game, Riccati differential equation, open loop and closed loop strategies, integrable singularities

**AMS subject classifications.** 91A05, 91A23, 49N70, 91A25

**DOI.** 10.1137/070696593

**1. Introduction.** The object of this paper is to revisit the pioneering work of Bernhard [2, 3] on two-person zero-sum linear quadratic differential games and generalize it to utility functions without positivity assumptions on the matrices acting on the state variable and to linear dynamics with bounded measurable data matrices. Our paper specializes to *state feedback* via Lebesgue measurable *affine closed loop strategies* with possible non- $L^2$ -integrable singularities. After sharpening the recent results of Delfour [5] on the characterization of the open loop lower and upper values of the game in section 2, it first deals with  $L^2$ -integrable closed loop strategies and then with the larger family of strategies that may have non- $L^2$ -integrable singularities.

In section 3 several equivalent necessary and sufficient conditions are given for the existence of a closed loop saddle point with respect to  $L^2$ -integrable affine closed loop strategies, for instance, the *normality* of the problem; the existence of an  $H^1(0, T)$  solution to the associated matrix Riccati differential equation. It was shown in [5] that the existence of a solution to the coupled state-adjoint state system is a necessary condition for the existence of a finite open loop lower value, upper value, or value of the game, and that the difference essentially depends on the convexity of the utility function with respect to the control of the minimizing player and on its concavity with respect to the control of the maximizing player. This condition is also necessary

---

\*Received by the editors July 9, 2007; accepted for publication (in revised form) August 23, 2008; published electronically January 7, 2009. This research has been supported by a discovery grant of the National Sciences and Engineering Research Council of Canada.

<http://www.siam.org/journals/sicon/47-6/69659.html>

<sup>†</sup>Centre de Recherches Mathématiques and Département de Mathématiques et de Statistique, Université de Montréal, P.O. Box 6128, Centre-ville Station, Montréal QC, H3C 3J7, Canada (delfour@crm.umontreal.ca, olivier.dello.sbarba@umontreal.ca).

for the existence of a closed loop saddle point. It leads to a complete classification in terms of the convexity/concavity properties of the utility function.

Section 4 deals with two delicate issues. The first is the very definition of a closed loop saddle point in the presence of closed loop strategies with non- $L^2$ -integrable singularities. As was pointed out in [2, p. 68 and Remark 5.1] such strategies may lead to conflicting terms that simultaneously blow up in the utility function. Under the positivity assumptions, one may possibly get around this problem by setting the utility function equal to  $\pm\infty$ , but we do not have them here. So we had to introduce a new conceptually meaningful and mathematically precise definition (cf. Definition 4.5). It states that the original problem can be transformed via feedback in such a way that the resulting problem has an open loop saddle point at  $(0,0)$ . The second related issue is to specify the class of affine closed loop strategies (cf. Definition 4.3) in such a way that we can simultaneously handle, in the same framework,  $L^2$ -integrable closed loop strategies and smooth locally bounded ones, except at most in the neighborhood of finitely many instants of time as in [2].

It turns out that the classical definition of a closed loop saddle point (cf. Definition 3.2) can be an *undeterminate* or a *degenerate* one when either the open loop lower or upper value of the game is not finite (cf. Theorems 4.4 and 4.5). For instance, Berkovitz's equivalence [1] (see our Lemma 3.2) may not apply, as shown in Example 4.2. The proper point of view is that of Definition 4.1, which states that the two closed loop strategies cannot be chosen independently. They must be linked through the *admissibility condition* of Definition 4.3. This subtle difference fundamentally changes the nature of the problem and makes it different from the classical theory of saddle points with respect to two independent sets. We show that the slight relaxation of the definition of *normalizability* of the free end problem in the sense of [2, Definition 3.2] from isolated instants to a set of instants of zero measure is a necessary and sufficient condition for the existence of a closed loop saddle point. This relaxation of the classical notion allows singularities in the feedback law at an infinite number of instants, including accumulation points that are not isolated. This condition is also used to make sense of solutions with singularities to the matrix Riccati differential equation.

In section 4.7, we show that under the convexity-concavity condition, Definitions 3.2 and 4.5 of closed loop saddle points coincide and that closed loop strategies with non- $L^2$ -integrable singularities are useless. These singularities naturally occur when either the open loop lower or upper value of the game is not finite. We complete the classification of closed loop saddle points in section 4.8 along with conditions expressed in terms of the convexity/concavity properties of the utility function. We conclude in section 4.9 with an example of a nonnormalizable problem with finite open loop lower value that can be achieved by state feedback via a solution of the matrix Riccati differential equation.

## 2. Definitions, notation, and main results.

**2.1. System, utility function, values of the game.** Given a finite dimensional Euclidean space  $\mathbf{R}^d$  of dimension  $d \geq 1$ , the *norm* and *inner product* will be denoted by  $|x|$  and  $x \cdot y$ , respectively, and irrespective of the dimension  $d$  of the space. Given  $T > 0$ , the norm and inner product in  $L^2(0, T; \mathbf{R}^n)$  will be denoted  $\|f\|$  and  $(f, g)$ . The norm in the Sobolev space  $H^1(0, T; \mathbf{R}^n)$  will be written  $\|f\|_{H^1}$ .

Consider the following two-player zero-sum game over the finite time interval

$[0, T]$  characterized by the quadratic *utility function*

$$(2.1) \quad C_{x_0}(u, v) \stackrel{\text{def}}{=} Fx(T) \cdot x(T) + \int_0^T Q(t)x(t) \cdot x(t) + |u(t)|^2 - |v(t)|^2 dt,$$

where  $x$  is the solution of the linear differential system

$$(2.2) \quad x'(t) = A(t)x(t) + B_1(t)u(t) + B_2(t)v(t) \quad \text{a.e. in } [0, T], \quad x(0) = x_0,$$

$x_0 \in \mathbf{R}^n$  is the *initial state* at time  $t = 0$ ,  $u \in L^2(0, T; \mathbf{R}^m)$ ,  $m \geq 1$ , is the strategy of the first player, and  $v \in L^2(0, T; \mathbf{R}^k)$ ,  $k \geq 1$ , is the strategy of the second player. We assume that  $F$  is an  $n \times n$ -matrix and that  $A$ ,  $B_1$ ,  $B_2$ , and  $Q$  are matrix-functions of appropriate orders that are measurable and bounded almost everywhere in  $[0, T]$ . Moreover,  $Q(t)$  and  $F$  are symmetrical. It will be convenient to use the following compact notation and drop the a.e. in  $[0, T]$  wherever no confusion arises:

$$(2.3) \quad C_{x_0}(u, v) = Fx(T) \cdot x(T) + \int_0^T Qx \cdot x + |u|^2 - |v|^2 dt,$$

$$(2.4) \quad x' = Ax + B_1u + B_2v \quad \text{in } [0, T], \quad x(0) = x_0.$$

The above assumptions on  $F$ ,  $A$ ,  $B_1$ ,  $B_2$ , and  $Q$  will be used throughout this paper. The transpose of a matrix  $M$  will be denoted  $M^\top$ , the inverse of its transpose  $M^{-\top}$ , and  $R(t)$  will denote the matrix  $B_1(t)B_1(t)^\top - B_2(t)B_2(t)^\top$ .

DEFINITION 2.1. Let  $x_0$  be an initial state in  $\mathbf{R}^n$  at time  $t = 0$ .

(i) The game is said to achieve its open loop lower value (resp., upper value) if

$$(2.5) \quad v^-(x_0) \stackrel{\text{def}}{=} \sup_{v \in L^2(0, T; \mathbf{R}^k)} \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v)$$

$$(2.6) \quad \left( \text{resp., } v^+(x_0) \stackrel{\text{def}}{=} \inf_{u \in L^2(0, T; \mathbf{R}^m)} \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_{x_0}(u, v) \right)$$

is finite. By definition,  $v^-(x_0) \leq v^+(x_0)$ .

(ii) The game is said to achieve its open loop value if its open loop lower value  $v^-(x_0)$  and upper value  $v^+(x_0)$  are achieved and  $v^-(x_0) = v^+(x_0)$ . The open loop value of the game will be denoted by  $v(x_0)$ .

(iii) A pair  $(\bar{u}, \bar{v})$  in  $L^2(0, T; \mathbf{R}^m) \times L^2(0, T; \mathbf{R}^k)$  is an open loop saddle point of  $C_{x_0}(u, v)$  in  $L^2(0, T; \mathbf{R}^m) \times L^2(0, T; \mathbf{R}^k)$  if for all  $u$  in  $L^2(0, T; \mathbf{R}^m)$  and all  $v$  in  $L^2(0, T; \mathbf{R}^k)$ ,

$$(2.7) \quad C_{x_0}(\bar{u}, v) \leq C_{x_0}(\bar{u}, \bar{v}) \leq C_{x_0}(u, \bar{v}).$$

DEFINITION 2.2. Associate with  $x_0 \in \mathbf{R}^n$  the sets

$$(2.8) \quad V(x_0) \stackrel{\text{def}}{=} \left\{ v \in L^2(0, T; \mathbf{R}^k) : \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v) > -\infty \right\},$$

$$(2.9) \quad U(x_0) \stackrel{\text{def}}{=} \left\{ u \in L^2(0, T; \mathbf{R}^m) : \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_{x_0}(u, v) < +\infty \right\}.$$



**2.2. Properties of the utility function.** Recall from [5] that the utility function  $C_{x_0}(u, v)$  is infinitely differentiable and that its Hessian of second order derivatives is independent of  $(u, v)$ . Indeed,<sup>1</sup>

$$(2.10) \quad \frac{1}{2}dC_{x_0}(u, v; \bar{u}, \bar{v}) = Fx(T) \cdot \bar{y}(T) + (Qx, \bar{y}) + (u, \bar{u}) - (v, \bar{v}),$$

where  $x$  is the solution of (2.4) and  $\bar{y}$  is the solution of

$$(2.11) \quad \bar{y}' = A\bar{y} + B_1\bar{u} + B_2\bar{v}, \quad \bar{y}(0) = 0.$$

It is customary to introduce the *adjoint system*

$$(2.12) \quad p' + A^\top p + Qx = 0, \quad p(T) = Fx(T)$$

and rewrite expression (2.10) for the gradient in the form

$$(2.13) \quad \frac{1}{2}dC_{x_0}(u, v; \bar{u}, \bar{v}) = (B_1^\top p + u, \bar{u}) + (B_2^\top p - v, \bar{v}).$$

Hence  $dC_{x_0}(\hat{u}, \hat{v}; \bar{u}, \bar{v}) = 0$  for all  $\bar{u}$  and  $\bar{v}$  if and only if the *coupled system*

$$(2.14) \quad \begin{cases} \hat{x}' = A\hat{x} - R\hat{p}, & \hat{x}(0) = x_0, \\ \hat{p}' + A^\top \hat{p} + Q\hat{x} = 0, & \hat{p}(T) = F\hat{x}(T) \end{cases}$$

has a solution  $(\hat{x}, \hat{p})$  in  $H^1(0, T; \mathbf{R}^n)^2$  with  $(\hat{u}, \hat{v}) = (-B_1^\top \hat{p}, B_2^\top \hat{p})$ .

As expected, the Hessian is independent of  $(u, v)$ ,

$$(2.15) \quad \frac{1}{2}d^2C_{x_0}(u, v; \bar{u}, \bar{v}; \hat{u}, \hat{v}) = F\tilde{y}(T) \cdot \bar{y}(T) + (Q\tilde{y}, \bar{y}) + (\tilde{u}, \bar{u}) - (\tilde{v}, \bar{v}),$$

where  $\bar{y}$  is the solution of (2.11) and  $\tilde{y}$  is the solution of

$$(2.16) \quad \tilde{y}' = A\tilde{y} + B_1\tilde{u} + B_2\tilde{v}, \quad \tilde{y}(0) = 0.$$

In particular, for all  $x_0, u, v, \bar{u}$ , and  $\bar{v}$ ,  $d^2C_{x_0}(u, v; \bar{u}, \bar{v}; \bar{u}, \bar{v}) = 2C_0(\bar{u}, \bar{v})$ .

**2.3. Games with finite open loop lower or upper values.** We recall and sharpen the results of [5, Thms. 2.2, 2.3, and 2.4] when the open loop lower or upper value of the game is finite for a given initial state  $x_0$ . In each case, the global assumption of finiteness for *all* initial state  $x_0 \in \mathbf{R}^n$  yields the *uniqueness* of solution  $(x, p)$  of the coupled system (2.14) (cf. [5, Thms. 2.6, 2.7, and 2.8]).

**THEOREM 2.1.** *The following conditions are equivalent.*

(i) *There exist  $\hat{u}$  in  $L^2(0, T; \mathbf{R}^m)$  and  $\hat{v}$  in  $L^2(0, T; \mathbf{R}^k)$  such that*

$$(2.17) \quad C_{x_0}(\hat{u}, \hat{v}) = \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, \hat{v}) = \sup_{v \in L^2(0, T; \mathbf{R}^k)} \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v).$$

(ii) *The open loop lower value  $v^-(x_0)$  of the game is finite.*

<sup>1</sup>Given a real function  $f$  defined on a Banach space  $B$ , the *first directional semiderivative* at  $x$  in the direction  $v$  (when it exists) is defined as  $df(x; v) = \lim_{t \searrow 0} (f(x + tv) - f(x))/t$ . When the map  $v \mapsto df(x; v) : B \rightarrow \mathbf{R}$  is linear and continuous, it defines the *gradient*  $\nabla f(x)$  as an element of the dual  $B^*$  of  $B$ . The *second order bidirectional derivative* at  $x$  in the directions  $(v, w)$  (when it exists) is defined as  $d^2f(x; v, w) = \lim_{t \searrow 0} (df(x + tw; v) - df(x; v))/t$ . When the map  $(v, w) \mapsto d^2f(x; v, w) : B \times B \rightarrow \mathbf{R}$  is bilinear and continuous, it defines the *Hessian operator*  $Hf(x)$  as a continuous linear operator from  $B$  to  $B^*$ .

(iii) *There exists a solution in  $H^1(0, T; \mathbf{R}^n)^2$  of the coupled system*

$$(2.18) \quad \begin{cases} x' = Ax - Rp, & x(0) = x_0, \\ p' + A^\top p + Qx = 0, & p(T) = Fx(T), \end{cases}$$

*and the following identities are verified:*

$$(2.19) \quad \sup_{v \in V(0)} \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_0(u, v) = \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_0(u, 0) = C_0(0, 0).$$

*Under such conditions, the optimal controls and the open loop lower value are given by the expressions*

$$(2.20) \quad \hat{u} = -B_1^\top p, \quad \hat{v} = B_2^\top p, \quad \text{and} \quad C_{x_0}(\hat{u}, \hat{v}) = p(0) \cdot x_0.$$

*Remark 2.1.* The additional condition  $B_2^\top p \in V(x_0)$  that appeared in [5, Thms. 2.2 and 2.6] is redundant. To see that, recall that the last identity (2.19) is equivalent to the convexity of the mapping  $u \mapsto C_{x_0}(u, v)$ . By [5, Thm. 3.1] the convexity plus a solution of the coupled system (2.18) yields that

$$\inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, B_2^\top p) = C_{x_0}(-B_1^\top p, B_2^\top p) > -\infty \quad \Rightarrow \quad B_2^\top p \in V(x_0).$$

Similarly, the additional condition  $-B_1^\top p \in U(x_0)$  that appeared in [5, Thms. 2.3 and 2.7] is also redundant.

**THEOREM 2.2.** *The following conditions are equivalent.*

(i) *There exist  $\hat{u}$  in  $L^2(0, T; \mathbf{R}^m)$  and  $\hat{v}$  in  $L^2(0, T; \mathbf{R}^k)$  such that*

$$(2.21) \quad C_{x_0}(\hat{u}, \hat{v}) = \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_{x_0}(\hat{u}, v) = \inf_{u \in L^2(0, T; \mathbf{R}^m)} \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_{x_0}(u, v).$$

(ii) *The open loop upper value  $v^+(x_0)$  of the game is finite.*

(iii) *There exists a solution  $(x, p) \in H^1(0, T; \mathbf{R}^n)^2$  of the coupled system (2.18), and the following identities are verified:*

$$(2.22) \quad \inf_{u \in U(0)} \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_0(u, v) = \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_0(0, v) = C_0(0, 0).$$

*Under such conditions, the optimal controls and the open loop upper value are given by expressions (2.20).*

For the case when  $v^-(x_0)$  and  $v^+(x_0)$  are both finite, Zhang [9] proved that they are equal. So by combining this with the previous theorems, we get the complete picture.

**THEOREM 2.3.** *The following conditions are equivalent.*

(i) *There exist  $\hat{u}$  in  $L^2(0, T; \mathbf{R}^m)$  and  $\hat{v}$  in  $L^2(0, T; \mathbf{R}^k)$  such that*

$$(2.23) \quad \begin{aligned} C_{x_0}(\hat{u}, \hat{v}) &= \inf_{u \in L^2(0, T; \mathbf{R}^m)} \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_{x_0}(u, v) \\ &= \sup_{v \in L^2(0, T; \mathbf{R}^k)} \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v) \end{aligned}$$

*(that is,  $C_{x_0}(u, v)$  has a saddle point).*

(ii) *The open loop value  $v(x_0)$  of the game is finite.*

- (iii) *There exists a solution  $(x, p) \in H^1(0, T; \mathbf{R}^n)^2$  of the coupled system (2.18), and the following identities are verified:*

$$(2.24) \quad \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_0(u, 0) = C_0(0, 0) = \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_0(0, v).$$

- (iv) *The lower and upper open loop values,  $v^-(x_0)$  and  $v^+(x_0)$ , are finite. Under such conditions, the optimal controls and the value are given by expressions (2.20) and  $v(x_0) = v^-(x_0) = v^+(x_0)$ .*

There are six possible cases according to the fact that  $v^-(x_0)$  and  $v^+(x_0)$  are finite,  $-\infty$ , or  $+\infty$ . If either  $v^-(x_0)$  or  $v^+(x_0)$  is finite, there is a solution  $(x, p) \in H^1(0, T; \mathbf{R}^n)^2$  of the coupled system (2.18), and we have the following three cases:

$$\begin{aligned} \text{(a)} \quad & \left| \begin{array}{l} v^-(x_0) \text{ finite} \\ v^+(x_0) \text{ finite} \end{array} \right| \left| \begin{array}{l} \inf_u C_0(u, 0) = C_0(0, 0) \\ \sup_v C_0(0, v) = C_0(0, 0) \end{array} \right| \left| \begin{array}{l} u \mapsto C_0(u, 0) \text{ convex} \\ v \mapsto C_0(0, v) \text{ concave} \end{array} \right| \\ \text{(b)} \quad & \left| \begin{array}{l} v^-(x_0) \text{ finite} \\ v^+(x_0) = +\infty \end{array} \right| \left| \begin{array}{l} \inf_u C_0(u, 0) = C_0(0, 0) \\ \sup_v \inf_u C_0(u, v) = C_0(0, 0) \\ \sup_v C_0(0, v) > C_0(0, 0) \end{array} \right| \left| \begin{array}{l} u \mapsto C_0(u, 0) \text{ convex} \\ v \mapsto \inf_u C_0(u, v) \text{ concave} \\ v \mapsto C_0(0, v) \text{ not concave} \end{array} \right| \\ \text{(c)} \quad & \left| \begin{array}{l} v^-(x_0) = -\infty \\ v^+(x_0) \text{ finite} \end{array} \right| \left| \begin{array}{l} \inf_u C_0(u, 0) < C_0(0, 0) \\ \inf_u \sup_v C_0(u, v) = C_0(0, 0) \\ \sup_v C_0(0, v) = C_0(0, 0) \end{array} \right| \left| \begin{array}{l} u \mapsto C_0(u, 0) \text{ not convex} \\ u \mapsto \sup_v C_0(u, v) \text{ convex} \\ v \mapsto C_0(0, v) \text{ concave} \end{array} \right| \end{aligned}$$

There are three more cases as follows that can occur even if the coupled system (2.18) has a solution  $(x, p) \in H^1(0, T; \mathbf{R}^n)^2$  and neither  $v^-(x_0)$  nor  $v^+(x_0)$  is finite:

$$\begin{aligned} \text{(d)} \quad & \left| \begin{array}{l} v^-(x_0) = -\infty \\ v^+(x_0) = +\infty \end{array} \right| \left| \begin{array}{l} \inf_u C_0(u, 0) < C_0(0, 0) \\ \sup_v C_0(0, v) > C_0(0, 0) \end{array} \right| \left| \begin{array}{l} u \mapsto C_0(u, 0) \text{ not convex} \\ v \mapsto C_0(0, v) \text{ not concave} \end{array} \right| \\ \text{(e)} \quad & \left| \begin{array}{l} v^-(x_0) = +\infty \\ v^+(x_0) = +\infty \end{array} \right| \left| \begin{array}{l} \inf_u C_0(u, 0) = C_0(0, 0) \\ \sup_v \inf_u C_0(u, v) > C_0(0, 0) \\ \sup_v C_0(0, v) > C_0(0, 0) \end{array} \right| \left| \begin{array}{l} u \mapsto C_0(u, 0) \text{ convex} \\ v \mapsto \inf_u C_0(u, v) \text{ not concave} \\ v \mapsto C_0(0, v) \text{ not concave} \end{array} \right| \\ \text{(f)} \quad & \left| \begin{array}{l} v^-(x_0) = -\infty \\ v^+(x_0) = -\infty \end{array} \right| \left| \begin{array}{l} \inf_u \sup_v C_0(u, v) < C_0(0, 0) \\ \inf_u C_0(u, 0) < C_0(0, 0) \\ \sup_v C_0(0, v) = C_0(0, 0) \end{array} \right| \left| \begin{array}{l} u \mapsto C_0(u, 0) \text{ not convex} \\ u \mapsto \sup_v C_0(u, v) \text{ not convex} \\ v \mapsto C_0(0, v) \text{ concave} \end{array} \right| \end{aligned}$$

Case (d) can occur by combining a system of type (b) with a system of type (c) and a utility function equal to the sum of the two utility functions. Case (e) occurs for the following system and utility function:

$$(2.25) \quad \begin{aligned} x'(t) &= tu(t) + t^3v(t) \text{ in } [0, 2], \quad x(0) = x_0, \\ C_{x_0}(u, v) &= \frac{3}{8}x(2) \cdot x(2) + \int_0^2 |u(t)|^2 - |v(t)|^2 dt. \end{aligned}$$

Finally, by duality, case (f) can also occur.

**3.  $L^2$ -integrable closed loop strategies.** We generalize classical results to  $L^2$ -integrable affine closed loop feedback strategies for general  $F$  and  $Q(t)$  under the assumptions of section 2.1 on the matrix functions  $A$ ,  $B_1$ ,  $B_2$ ,  $Q$ , and  $F$ . We also give a classification of the possible cases in terms of the open loop properties of lower value, upper value, and value of the game and the convexity/concavity of the utility function.

### 3.1. Definitions and main results.

DEFINITION 3.1 ( $L^2$ -integrable affine closed loop strategies).

$$\Phi \stackrel{\text{def}}{=} \left\{ \phi : [0, T] \times \mathbf{R}^n \rightarrow \mathbf{R}^m \left| \begin{array}{l} \text{such that } x \mapsto \phi(t, x) \text{ is affine and} \\ t \mapsto \phi(t, x) \text{ belongs to } L^2(0, T; \mathbf{R}^m) \end{array} \right. \right\}.$$

$$\Psi \stackrel{\text{def}}{=} \left\{ \psi : [0, T] \times \mathbf{R}^n \rightarrow \mathbf{R}^k \left| \begin{array}{l} \text{such that } x \mapsto \psi(t, x) \text{ is affine and} \\ t \mapsto \psi(t, x) \text{ belongs to } L^2(0, T; \mathbf{R}^k) \end{array} \right. \right\}.$$

We say that  $\phi$  or  $\psi$  is linear if  $\phi(t, x)$  or  $\psi(t, x)$  is linear in  $x$ .

Remark 3.1. To each  $\phi \in \Phi$  (resp.,  $\psi \in \Psi$ ) we can associate an  $L^2(0, T; \mathbf{R}^m)$ -vector function  $u$  and an  $m \times n$ -matrix  $L^2$ -function  $U$  such that  $\phi(t, x) = u(t) + U(t)x$  (resp., an  $L^2(0, T; \mathbf{R}^k)$ -vector function  $v$  and a  $k \times n$ -matrix  $L^2$ -function  $V$  such that  $\psi(t, x) = v(t) + V(t)x$ ). The matrix functions  $U$  and  $V$  may have singularities, but they are globally  $L^2$ -integrable. As a result, the fundamental matrix associated with the  $L^2$ -matrix function  $A + B_1U + B_2V$  will be invertible everywhere in  $[0, T]$ . Therefore for all  $\phi \in \Phi$  and  $\psi \in \Psi$ , the closed loop system

$$(3.1) \quad \begin{aligned} x' &= Ax + B_1\phi(x) + B_2\psi(x), \quad x(0) = x_0, \\ x' &= (A + B_1U + B_2V)x + B_1u + B_2v, \quad x(0) = x_0 \end{aligned}$$

has a unique solution in  $H^1(0, T; \mathbf{R}^n)$  for all  $x_0 \in \mathbf{R}^n$ . This means that all pairs  $(\phi, \psi) \in \Phi \times \Psi$  are *admissible*, and, a fortiori, all pairs of the form  $(\phi, v)$  or  $(u, \psi)$  are admissible for all  $u \in L^2(0, T; \mathbf{R}^m)$  and  $v \in L^2(0, T; \mathbf{R}^k)$ .

DEFINITION 3.2.

- (i) Given  $x_0 \in \mathbf{R}^n$ , we say that  $(\phi^*, \psi^*) \in \Phi \times \Psi$  is an  $L^2$ -integrable closed loop saddle point of  $C_{x_0}(\phi, \psi)$  in  $\Phi \times \Psi$  if for all  $\phi \in \Phi$  and  $\psi \in \Psi$ ,

$$(3.2) \quad C_{x_0}(\phi^*, \psi) \leq C_{x_0}(\phi^*, \psi^*) \leq C_{x_0}(\phi, \psi^*).$$

- (ii) We say that  $(\phi^*, \psi^*) \in \Phi \times \Psi$  is an  $L^2$ -integrable global closed loop saddle point of  $C_{x_0}(\phi, \psi)$  in  $\Phi \times \Psi$  if for all  $x_0 \in \mathbf{R}^n$  and for all  $\phi \in \Phi$  and  $\psi \in \Psi$  the inequalities (3.2) are verified.

By definition,  $C_{x_0}(\phi^*, \psi^*)$  is finite. Thus the saddle point is “nondegenerate” in the sense of [2]. The “global version” is better adapted to closed loop strategies. The interest in a closed loop strategy associated with a single initial state is rather limited.

Given any two pairs  $(\phi_1, \psi_1)$  and  $(\phi_2, \psi_2)$  achieving an  $L^2$ -integrable closed saddle point, the mixed pairs  $(\phi_1, \psi_2)$  and  $(\phi_2, \psi_1)$  are admissible and also achieve an  $L^2$ -integrable closed loop saddle point. Hence the value of the closed loop saddle point is unique (cf. [1]).

LEMMA 3.1. Given  $x_0 \in \mathbf{R}^n$ , for all pairs  $(\phi_1^*, \psi_1^*) \in \Phi \times \Psi$  and  $(\phi_2^*, \psi_2^*) \in \Phi \times \Psi$  verifying (3.2),  $C_{x_0}(\phi_1^*, \psi_1^*) = C_{x_0}(\phi_2^*, \psi_2^*)$ .

We quote Berkovitz’s equivalence lemma [1].

LEMMA 3.2. Given  $x_0 \in \mathbf{R}^n$ , the following statements are equivalent:

- (i)  $(\phi^*, \psi^*) \in \Phi \times \Psi$  is an  $L^2$ -integrable closed loop saddle point of  $C_{x_0}(\phi, \psi)$ ;  
 (ii) there exists a pair  $(\phi^*, \psi^*) \in \Phi \times \Psi$  such that for all  $u \in L^2(0, T; \mathbf{R}^m)$  and all  $v \in L^2(0, T; \mathbf{R}^k)$ ,

$$(3.3) \quad C_{x_0}(\phi^*, v) \leq C_{x_0}(\phi^*, \psi^*) \leq C_{x_0}(u, \psi^*).$$

THEOREM 3.1. *The following statements are equivalent.*

- (i)  $(\phi^*, \psi^*) \in \Phi \times \Psi$  is an  $L^2$ -integrable global closed loop saddle point of  $C_{x_0}(\phi, \psi)$ .  
 (ii) There exists  $(\phi^*, \psi^*) \in \Phi \times \Psi$  such that for all  $x_0 \in \mathbf{R}^n$ ,  $u \in L^2(0, T; \mathbf{R}^m)$ , and  $v \in L^2(0, T; \mathbf{R}^k)$ ,

$$(3.4) \quad C_{x_0}(\phi^*, \psi^* + v) \leq C_{x_0}(\phi^*, \psi^*) \leq C_{x_0}(\phi^* + u, \psi^*).$$

- (iii) For all  $x_0 \in \mathbf{R}^n$  there exist a unique solution  $(\hat{x}, \hat{p}) \in H^1(0, T; \mathbf{R}^n)^2$  of

$$(3.5) \quad \begin{cases} \hat{x}' = A\hat{x} - R\hat{p}, & \hat{x}(0) = x_0, \\ \hat{p}' + A^\top \hat{p} + Q\hat{x} = 0, & \hat{p}(T) = F\hat{x}(T) \end{cases}$$

and  $L^2$ -integrable matrices  $U_*$  and  $V_*$  of appropriate orders such that for all  $x_0 \in \mathbf{R}^n$ ,

$$(3.6) \quad \hat{u} = -B_1^\top \hat{p} = U_* \hat{x}, \quad \hat{v} = B_2^\top \hat{p} = V_* \hat{x}.$$

- (iv) (Invariant embedding). For all initial times  $s \in [0, T]$ , there exists a unique  $H^1(s, T)$  solution of the coupled matrix system

$$(3.7) \quad \begin{cases} \hat{X}'_s = A\hat{X}_s - R\hat{\Lambda}_s, & \hat{X}_s(s) = I, \\ \hat{\Lambda}'_s + A^\top \hat{\Lambda}_s + Q\hat{X}_s = 0, & \hat{\Lambda}_s(T) = F\hat{X}_s(T). \end{cases}$$

By convention, set  $\hat{X}_T(T) = I$  and  $\hat{\Lambda}_T(T) = F$ .

- (v) (Normality).  $\det X(t) \neq 0$  everywhere in  $[0, T]$ , where  $(X, \Lambda)$  is the  $H^1(0, T)$  solution of the matrix differential system

$$(3.8) \quad \begin{cases} X' = AX - R\Lambda, & X(T) = I, \\ \Lambda' + A^\top \Lambda + QX = 0, & \Lambda(T) = F. \end{cases}$$

- (vi) There exists a symmetrical solution  $P$  with elements in  $H^1(0, T)$  to the matrix Riccati differential equation

$$(3.9) \quad P' + PA + A^\top P - PRP + Q = 0, \quad P(T) = F.$$

In particular  $C_{x_0}(\phi^*, \psi^*) = P(0)x_0 \cdot x_0$ , and the closed loop strategies are given by

$$(3.10) \quad \phi^*(t, x) = -B_1^\top(t)P(t)x = U_*(t)x \text{ and } \psi^*(t, x) = B_2^\top(t)P(t)x = V_*(t)x.$$

*Proof.* (i)  $\Rightarrow$  (ii). Let  $\hat{x}$  be the trajectory corresponding to the pair  $(\phi^*, \psi^*)$ , and denote by  $(\hat{u}, \hat{v}) = (\phi^*(x), \psi^*(x))$  the corresponding control pair. Let  $U_*(t)$  and  $V_*(t)$  be the respective matrices and  $u_*(t)$  and  $v_*(t)$  be the respective vectors such that  $\phi^*(t, x) = U_*(t)x + u_*(t)$  and  $\psi^*(t, x) = V_*(t)x + v_*(t)$ . Then

$$(3.11) \quad \hat{x}' = (A + B_1U_* + B_2V_*)\hat{x} + B_1u_* + B_2v_*, \quad \hat{x}(0) = x_0.$$

For all  $u \in L^2(0, T; \mathbf{R}^m)$  and  $v \in L^2(0, T; \mathbf{R}^k)$ , the pair  $(\phi^* + u, \psi^* + v) \in \Phi \times \Psi$  and

$$(3.12) \quad C_{x_0}(\phi^*, \psi^* + v) \leq C_{x_0}(\phi^*, \psi^*) \leq C_{x_0}(\phi^* + u, \psi^*).$$

(ii)  $\Rightarrow$  (iii). Introduce the notation  $c_{x_0}(u, v)$  for the utility function  $C_{x_0}(\phi^* + u, \psi^* + v)$ :

$$c_{x_0}(u, v) \stackrel{\text{def}}{=} Fx(T) \cdot x(T) + \int_0^T Qx \cdot x + |U_*x + u_* + u|^2 - |V_*x + v_* + v|^2 dt,$$

and denote by  $x$  the solution of the corresponding state equation

$$(3.13) \quad x' = (A + B_1U_* + B_2V_*)x + B_1(u_* + u) + B_2(v_* + v), \quad x(0) = x_0.$$

Then the closed loop saddle point inequalities (3.12) become open loop saddle point inequalities for system (3.13) and the new quadratic utility function  $c_{x_0}(u, v)$  satisfies the saddle point condition

$$(3.14) \quad \forall u \in L^2(0, T; \mathbf{R}^m) \text{ and } v \in L^2(0, T; \mathbf{R}^k), \quad c_{x_0}(0, v) \leq c_{x_0}(0, 0) \leq c_{x_0}(u, 0),$$

and the pair  $(0, 0)$  achieves that saddle point. By [5, Lemma 3.1]  $c_{x_0}(u, v)$  is convex-concave and  $dc_{x_0}(0, 0; u, v) = 0$  for all  $u$  and  $v$ . In particular, the coupled system

$$(3.15) \quad \begin{cases} \hat{x}' = (A + B_1U_* + B_2V_*)\hat{x} + B_1u_* + B_2v_*, & \hat{x}(0) = x_0, \\ \hat{p}' + (A + B_1U_* + B_2V_*)^\top \hat{p} + Q\hat{x} + U_*^\top (U_*\hat{x} + u_*) - V_*^\top (V_*\hat{x} + v_*) = 0, \\ \hat{p}(T) = F\hat{x}(T), \end{cases}$$

$$0 = -B_1^\top \hat{p} - (U_*\hat{x} + u_*) \text{ and } 0 = B_2^\top \hat{p} - (V_*\hat{x} + v_*)$$

has a solution  $(\hat{x}, \hat{p})$  in  $H^1(0, T; \mathbf{R}^n)^2$ . After substitution, it can be rewritten as

$$(3.16) \quad \begin{aligned} \hat{x}' &= A\hat{x} - R\hat{p}, & \hat{x}(0) &= x_0, \\ \hat{p}' + A^\top \hat{p} + Q\hat{x} &= 0, & \hat{p}(T) &= F\hat{x}(T). \end{aligned}$$

By assumption this is true for all  $x_0 \in \mathbf{R}^n$ . But, when system (3.16) has a solution for all  $x_0$ , its solution is unique (cf, [5, section 2.6, pp. 760–761]). As a result, for  $x_0 = 0$ , we have  $(\hat{x}, \hat{p}) = (0, 0)$ , and from identities (3.15),

$$0 = -B_1^\top \hat{p} - (U_*\hat{x} + u_*) \text{ and } 0 = B_2^\top \hat{p} - (V_*\hat{x} + v_*) \Rightarrow u_* = 0 \text{ and } v_* = 0.$$

Hence the feedback controls are of the form  $\hat{u} = U_*\hat{x} = -B_1^\top \hat{p}$  and  $\hat{v} = V_*\hat{x} = B_2^\top \hat{p}$ .

(iii)  $\Rightarrow$  (iv). By assumption for all  $x_0 \in \mathbf{R}^n$ , the coupled system (3.16) has a unique solution  $(\hat{x}, \hat{p})$ . By linearity of the solution of system (3.5) with respect to  $x_0$ , there exist  $H^1(0, T)$ -matrices  $(\hat{X}, \hat{\Lambda})$  solution of the matrix system

$$(3.17) \quad \begin{aligned} \hat{X}' &= A\hat{X} - R\hat{\Lambda}, & \hat{X}(0) &= I, \\ \hat{\Lambda}' + A^\top \hat{\Lambda} + Q\hat{X} &= 0, & \hat{\Lambda}(T) &= F\hat{X}(T). \end{aligned}$$

But the conditions  $U_*\hat{x} = -B_1^\top \hat{p}$  and  $V_*\hat{x} = B_2^\top \hat{p}$  for all  $x_0$  imply that  $U_*\hat{X} = -B_1^\top \hat{\Lambda}$  and  $V_*\hat{X} = B_2^\top \hat{\Lambda}$ , and  $\hat{X}$  is also the unique solution of the equation

$$(3.18) \quad \hat{X}' = (A + B_1U_* + B_2V_*)\hat{X}, \quad \hat{X}(0) = I.$$

Since the elements of the matrix function  $A + B_1U_* + B_2V_*$  are  $L^2$ -functions, the associated fundamental matrix solution  $\Phi(t, s)$  is invertible,  $\hat{X}(t)x_0 = \hat{x}(t) = \Phi(t, 0)x_0$ , and, a fortiori,  $\hat{X}(t) = \Phi(t, 0)$  is invertible in  $[0, T]$ . In particular, for all  $s \in [0, T[$ ,  $(\hat{X}_s(t), \hat{\Lambda}_s(t)) = (\hat{X}(t)\hat{X}(s)^{-1}, \hat{\Lambda}(t)\hat{X}(s)^{-1})$  is a solution of system (3.7).

(iv)  $\Rightarrow$  (v). First, observe that  $\hat{X}_s(T)$  is invertible for all  $s \in [0, T]$ . For  $s < T$ , let  $h \neq 0$  be such that  $\hat{X}_s(T)h = 0$ . The pair  $(x_s(t), p_s(t)) = (\hat{X}_s(t)h, \hat{\Lambda}_s(t)h)$  is a solution of the system  $x'_s = Ax_s - Rp_s$ ,  $x_s(s) = h$ , and  $p'_s + A^\top p_s + Qx_s = 0$ ,  $p_s(T) = Fx_s(T)$ , with  $(x_s(T), p_s(T)) = (0, 0)$ . Hence  $(x_s, p_s) = (0, 0)$  and  $h = x_s(s) = 0$ , a contradiction. For all  $0 \leq s \leq t \leq T$ ,  $\hat{X}_s(T) = \hat{X}_t(T)\hat{X}_s(t)$  and  $\det \hat{X}_s(t) \neq 0$ . Defining the matrix functions  $(X(t), \Lambda(t)) = (\hat{X}_0(t)\hat{X}_0(T)^{-1}, \hat{\Lambda}_0(t)\hat{X}_0(T)^{-1})$ , they are a solution of the matrix differential system (3.7), and necessarily,  $\det X(t) \neq 0$  everywhere in  $[0, T]$ .

(v)  $\Rightarrow$  (vi). Since  $X(t)$  is invertible for all  $t \in [0, T]$ , then  $P(t) = \Lambda(t)X(t)^{-1}$  is a matrix of  $H^1(0, T)$  functions. Moreover,  $P$  is symmetrical since  $\Lambda^\top X$  is by computing the derivative of  $\Lambda^\top X - X^\top \Lambda$ . Finally,  $P$  is a solution of the Riccati matrix differential equation (3.9). By uniqueness of the solution  $(x, p)$  of (3.5),  $(x(t), p(t)) = (\hat{X}_0(t)x_0, \hat{\Lambda}_0(t)x_0) = (X(t)\hat{X}_0(T)x_0, \Lambda(t)\hat{X}_0(T)x_0)$ . Hence  $p(t) = P(t)x(t)$ ,  $\hat{u}(t) = -B_1^\top p(t) = -B_1^\top P(t)x(t)$ , and  $\hat{v}(t) = B_2^\top p(t) = B_2^\top P(t)x(t)$ .

(vi)  $\Rightarrow$  (i). Let  $x \in H^1(0, T; \mathbf{R}^n)$  be the solution of

$$(3.19) \quad x' = Ax + B_1u + B_2v, \quad x(0) = x_0,$$

and let  $P$  be an  $H^1(0, T)$  solution of the matrix Riccati differential equation (3.9). By the classical argument of Bernhard [2], we get

$$C_{x_0}(u, v) = P(0)x_0 \cdot x_0 + \int_0^T |u + B_1^\top Px|^2 - |v - B_2^\top Px|^2 dt.$$

Choose the closed loop linear strategies  $\phi^*(t, x) = -B_1^\top(t)P(t)x$  and  $\psi^*(t, x) = B_2^\top(t)P(t)x$ . Then for all  $v \in L^2(0, T; \mathbf{R}^k)$  and all  $u \in L^2(0, T; \mathbf{R}^m)$ ,

$$\begin{aligned} C_{x_0}(\phi^*, \psi^*) &= P(0)x_0 \cdot x_0, \\ C_{x_0}(u, \psi^*) &= P(0)x_0 \cdot x_0 + \int_0^T |u + B_1^\top Px|^2 dt \geq P(0)x_0 \cdot x_0 = C_{x_0}(\phi^*, \psi^*), \\ C_{x_0}(\phi^*, v) &= P(0)x_0 \cdot x_0 - \int_0^T |v - B_2^\top Px|^2 dt \leq P(0)x_0 \cdot x_0 = C_{x_0}(\phi^*, \psi^*). \end{aligned}$$

By Lemma 3.2(ii), the linear pair  $(\phi^*, \psi^*)$  is a global closed loop saddle point. Finally, the pair of closed loop strategies  $\phi^*(t, x) = -B_1^\top(t)P(t)x = U_*(t)x$  and  $\psi^*(t, x) = B_2^\top(t)P(t)x = V_*(t)x$  yields the global closed loop saddle point  $P(0)x_0 \cdot x_0$ .  $\square$

**3.2. Classification of  $L^2$ -integrable closed loop saddle points.** One of the necessary conditions for the existence of a closed loop saddle point is the existence of a solution to the coupled system in  $(\hat{x}, \hat{p})$  that turns out to also be a necessary condition for the finiteness of the open loop lower value, upper value, or value of the game. The difference essentially depends on the convexity of the utility function with respect to  $u$  and on its concavity with respect to  $v$  that yields to the following classification.

**THEOREM 3.2.** *Assume that  $(\phi^*, \psi^*) \in \Phi \times \Psi$  is an  $L^2$ -integrable closed loop saddle point of  $C_{x_0}(\phi, \psi)$ .*

(a)  $v(x_0)$  is finite if and only if  $C_{x_0}(u, v)$  is convex in  $u$  and concave in  $v$ .

- (b)  $v^-(x_0)$  is finite and  $v^+(x_0) = +\infty$  if and only if  $C_{x_0}(u, v)$  is convex in  $u$  and not concave in  $v$ .
- (c)  $v^+(x_0)$  is finite and  $v^-(x_0) = -\infty$  if and only if  $C_{x_0}(u, v)$  is concave in  $v$  and not convex in  $u$ .
- (d)  $v^-(x_0) = -\infty$  and  $v^+(x_0) = +\infty$  if and only if  $C_{x_0}(u, v)$  is not convex in  $u$  and not concave in  $v$ .
- (e)  $v^-(x_0) = v^+(x_0) = +\infty$  cannot occur.
- (f)  $v^-(x_0) = v^+(x_0) = -\infty$  cannot occur.

In the first three cases,  $C_{x_0}(\phi^*, \psi^*)$  is equal to  $v(x_0)$ ,  $v^-(x_0)$ , and  $v^+(x_0)$ , respectively.

*Remark 3.2.* Case (d) can occur. An example can be constructed by using a first system of type (b) and a second system of type (c) without interconnection, with utility function equal to the sum of the two utility functions.

*Remark 3.3.* In Bernhard [2], the utility function was convex in  $u$  since  $F \geq 0$  and  $Q(t) \geq 0$ . Only cases (a) and (b) can occur, and case (e) is a degenerate one.

We need the following lemma.

LEMMA 3.3.

- (i) For all  $v \in L^2(0, T; \mathbf{R}^k)$ ,

$$(3.20) \quad \inf_{\phi \in \Phi} C_{x_0}(\phi, v) = \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v),$$

$$(3.21) \quad \sup_{\psi \in \Psi} \inf_{\phi \in \Phi} C_{x_0}(\phi, \psi) \geq \sup_{v \in L^2(0, T; \mathbf{R}^k)} \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v).$$

- (ii) For all  $u \in L^2(0, T; \mathbf{R}^m)$ ,

$$(3.22) \quad \sup_{\psi \in \Psi} C_{x_0}(u, \psi) = \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_{x_0}(u, v),$$

$$(3.23) \quad \inf_{\phi \in \Phi} \sup_{\psi \in \Psi} C_{x_0}(\phi, \psi) \leq \inf_{u \in L^2(0, T; \mathbf{R}^m)} \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_{x_0}(u, v).$$

*Proof of Lemma 3.3.* We need only prove (i). Since  $L^2(0, T; \mathbf{R}^m) \subset \Phi$ ,

$$\inf_{\phi \in \Phi} C_{x_0}(\phi, v) \leq \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v).$$

Conversely, given the pair  $(\phi, v)$ , let  $x \in H^1(0, T; \mathbf{R}^n)$  be the solution of the system

$$x' = Ax + B_1\phi(x) + B_2v, \quad x(0) = x_0,$$

and let  $u = \phi(x) \in L^2(0, T; \mathbf{R}^m)$ . This implies that

$$\begin{aligned} C_{x_0}(\phi, v) &= C_{x_0}(u, v) \geq \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v) \\ \Rightarrow \inf_{\phi \in \Phi} C_{x_0}(\phi, v) &\geq \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v) \Rightarrow \inf_{\phi \in \Phi} C_{x_0}(\phi, v) = \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v). \end{aligned}$$

The second inequality follows from the fact that  $L^2(0, T; \mathbf{R}^k) \subset \Psi$ .  $\square$

*Proof of Theorem 3.2.* From inequality (3.21),  $v^-(x_0) \leq C_{x_0}(\phi^*, \psi^*) < +\infty$  and case (e) cannot occur. Similarly, from inequality (3.23),  $v^+(x_0) \geq C_{x_0}(\phi^*, \psi^*) > -\infty$  and case (f) cannot occur. Therefore we are left with the first four cases.



(b) From the first part of the proof of Theorem 3.1, system (3.16) has a solution, and identities (3.15) are verified:

$$(3.24) \quad \begin{cases} \hat{x}' = A\hat{x} - R\hat{p}, & \hat{x}(0) = x_0, \\ \hat{p}' + A^\top \hat{p} + Q\hat{x} = 0, & \hat{p}(T) = F\hat{x}(T), \end{cases}$$

$$(3.25) \quad 0 = -B_1^\top \hat{p} - (U_* \hat{x} + u_*) \text{ and } 0 = B_2^\top \hat{p} - (V_* \hat{x} + v_*).$$

Using the controls  $(\hat{u}, \hat{v}) = (U_* \hat{x} + u_*, V_* \hat{x} + v_*) = (-B_1^\top \hat{p}, B_2^\top \hat{p})$ , the above system can be rewritten as

$$\begin{cases} \hat{x}' = A\hat{x} - B_1 B_1^\top \hat{p} + B_2 \hat{v}, & \hat{x}(0) = x_0, & \hat{u} = -B_1^\top \hat{p}, \\ \hat{p}' + A^\top \hat{p} + Q\hat{x} = 0, & \hat{p}(T) = F\hat{x}(T). \end{cases}$$

If  $C_{x_0}(u, v)$  is convex in  $u$ , this implies that  $\hat{u}$  is a minimizer of  $C_{x_0}(u, \hat{v})$  over  $u$  (cf., for instance, [5, Thm 3.1]). Therefore

$$\sup_{v \in L^2(0, T; \mathbf{R}^k)} \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v) \geq \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, \hat{v}) = C_{x_0}(\hat{u}, \hat{v}).$$

But, by construction of  $(\hat{u}, \hat{v})$ ,  $C_{x_0}(\phi^*, \psi^*) = C_{x_0}(\hat{u}, \hat{v})$ . Combining the above inequality with inequality (3.21) in Lemma 3.3, we get

$$v^-(x_0) = \sup_{v \in L^2(0, T; \mathbf{R}^k)} \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, v) = \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, \hat{v}) = C_{x_0}(\hat{u}, \hat{v})$$

and hence the finiteness of  $v^-(x_0)$ . If, in addition,  $C_{x_0}(u, v)$  was concave in  $v$ , then by [5, Thms. 2.5 and 2.4] the value of the game, and hence  $v^+(x_0)$ , would be finite, and this contradicts our assumption. Conversely, if  $v^-(x_0)$  is finite, then the mapping  $u \mapsto C_{x_0}(u, v)$  is convex ([5, Thm. 2.2(iii), last part of identity (2.35) and Remark 2.2]). If  $v^+(x_0)$  was also finite, then by [5, Thms. 2.5 and 2.4(iii)]  $C_{x_0}(u, v)$  would be concave in  $v$  in contradiction with our assumption.

The proof of (c) is dual to the proof of (b). The proof of (a) is similar to the proof of (b) and (c). Case (d) is the complement of all the other cases, so it can only occur when  $C_{x_0}(u, v)$  is neither convex in  $u$  nor concave in  $v$ .  $\square$

*Remark 3.4.* If the problem is normal and  $F \geq 0$  and  $Q(t) \geq 0$ , then  $v^-(x_0)$  is finite and  $v^-(x_0) \geq 0$  for all  $x_0 \in \mathbf{R}^n$ , and necessarily  $P(0)x_0 \cdot x_0 \geq 0$  for all  $x_0 \in \mathbf{R}^n$ .

#### 4. Not necessarily $L^2$ -integrable affine closed loop strategies.

**4.1. The curse of singularities.** We now extend the definitions and results of the previous section to Lebesgue measurable feedback matrices with singularities that are not necessarily  $L^2$ -integrable in any of its neighborhood. In order to accommodate such strategies, we first enlarge the sets of strategies  $\Phi$  and  $\Psi$ .

**DEFINITION 4.1.** *The class of affine closed loop strategies is defined as follows:*

$$\begin{aligned} \tilde{\Phi} &\stackrel{\text{def}}{=} \left\{ \phi : [0, T] \times \mathbf{R}^n \rightarrow \mathbf{R}^m \left| \begin{array}{l} \text{such that } x \mapsto \phi(t, x) \text{ is affine,} \\ t \mapsto \phi(t, x) \text{ is Lebesgue measurable, and} \\ t \mapsto \phi(t, 0) \text{ belongs to } L^2(0, T; \mathbf{R}^m) \end{array} \right. \right\}, \\ \tilde{\Psi} &\stackrel{\text{def}}{=} \left\{ \psi : [0, T] \times \mathbf{R}^n \rightarrow \mathbf{R}^k \left| \begin{array}{l} \text{such that } x \mapsto \psi(t, x) \text{ is affine,} \\ t \mapsto \psi(t, x) \text{ is Lebesgue measurable, and} \\ t \mapsto \psi(t, 0) \text{ belongs to } L^2(0, T; \mathbf{R}^k) \end{array} \right. \right\}. \end{aligned}$$

We say that  $\phi$  (resp.,  $\psi$ ) is a linear closed loop strategy if  $\phi$  (resp.,  $\psi$ ) is linear in  $x$ .

To any  $(\phi, \psi) \in \tilde{\Phi} \times \tilde{\Psi}$  are associated measurable matrix functions  $U(t)$  and  $V(t)$  and  $L^2$ -vector functions  $u$  and  $v$  such that  $\phi(t, x) = U(t)x + u(t)$  and  $\psi(t, x) = V(t)x + v(t)$ . At that level of generality, an *admissibility condition* on the pair  $(\phi, \psi) \in \tilde{\Phi} \times \tilde{\Psi}$  is required to make sense of a solution of the underlying differential equation. How should the admissible affine closed loop strategies be chosen in order to preserve the assumption of *normalizability* of [2] that makes sense of a non- $H^1(0, T)$ -solution  $P$  to the matrix Riccati differential equation? It is clear that the choice of the space of solutions of the matrix Riccati differential equation and the specification of the families  $\Phi$  and  $\Psi$  are closely related.

It has been known that the solution of the scalar Riccati differential equation can exhibit singularities that are not *movable branch points*, at least when the coefficients are smooth functions of  $t$  (cf. Ince [6, section 12.51, p. 293]). Another interesting property is that “the general solution of the Riccati equation is expressible rationally in terms of any three distinct particular solutions, and also that the anharmonic ratio of any four solutions is constant. It also shows that the general solution is a rational function of the constant of integration (cf. Ince [6, section 2.15, p. 23 and section 12.51, p. 294]).”

This result was extended to the  $n \times n$  ( $n \geq 2$ ) solution of the matrix Riccati differential equation by Sorine and Winternitz [8], but with five particular solutions in the general case and four in the *symplectic case* that corresponds to our assumptions on the data matrices. They also give some thought to the space of solutions:<sup>2</sup> “For smooth coefficients  $A$ ,  $B$ ,  $C$ , and  $D$  in the MRE (6) the solution space consists of meromorphic matrices: the matrix elements may have first-order poles, the positions of which depends on the initial conditions. In other words, the MRE (6) has the Painlevé property [6]: the solutions have no moving critical points, i.e., no branch points or essential singularities, the positions of which depend on the initial conditions (cf. [8, pp. 271–272]).”

**4.2. Bernhard’s conditions [2] in the free end case.** In the free end case with  $F \geq 0$  and  $Q(t) \geq 0$ , the necessary and sufficient condition of Bernhard [2, Thm. 3.1] for the existence of a nondegenerate closed loop saddle point in the sense of [2, Definition 2.3 and Remark 5.1] reduces to the following three properties:

- (ii)  $X(t)$  is invertible except possibly at isolated points in  $[0, T]$ , where  $(X, \Lambda)$  is the unique  $H^1(0, T)$  matrix solution of

$$(4.1) \quad \begin{cases} X' = AX - R\Lambda, & X(T) = I, \\ \Lambda' + A^\top \Lambda + QX = 0, & \Lambda(T) = F; \end{cases}$$

- (iii)  $x_0 \in \text{Im } X(0)$ ;

- (iv) for all  $t \in [0, T]$ ,  $P(t) \geq 0$ ,

where  $P$  is defined in terms of  $\Lambda$  and the pseudoinverse of  $X$  as follows:

$$(4.2) \quad P(t) = \Lambda(t) X(t)^\dagger, \quad X(t)^\dagger \stackrel{\text{def}}{=} \begin{cases} [X(t)^\top X(t)]^{-1} X(t)^\top & \text{if } X(t) \neq 0, \\ \text{arbitrary} & \text{if } X(t) = 0, \end{cases}$$

and  $[X(t)^\top X(t)]^{-1}$  is the inverse of  $X(t)^\top X(t)$  as a matrix from  $\text{Im } X(t)^\top$  onto itself.

<sup>2</sup>Sorine and Winternitz [8] consider solutions  $W$  of the general matrix Riccati differential equation  $W' = A + WB + CW + WDW$ ,  $W(T) = W_0$ .

Condition (ii) defines the matrix function  $P(t)$  almost everywhere in  $[0, T]$  and gives meaning to a solution of the Riccati differential equation via the solution  $(\Lambda, X)$  of system (4.1). The positivity of  $F$  and  $Q(t)$  makes the utility function  $C_{x_0}(u, v)$  convex in  $u$  and this leads to the positivity of  $P(t)$  (cf. Remark 3.4). Our objective is to relax those positivity assumptions as in Theorem 3.1. Without them some of the competitive terms in the utility function may simultaneously blow up, making it difficult to set the utility function equal to  $\pm\infty$  (cf. [2, p. 68 and Remark 5.1]). Moreover, non- $L^2$ -integrable singularities in the closed loop strategies invalidate the equivalence (ii) of Lemma 3.2 when either the open loop lower or upper value of the game is not finite. So the very definition of a closed loop saddle point has to be properly revisited, and the family of pairs of admissible strategies is no longer  $\Phi \times \Psi$  but a subspace  $S$  of an enlarged product space  $\tilde{\Phi} \times \tilde{\Psi}$  containing  $\Phi \times \Psi$ .

**4.3. Normalizability and its consequences.** Given the matrices  $A$ ,  $B_1$ ,  $B_2$ ,  $Q$ , and  $F$  verifying the conditions of section 2.1, system (4.1) always has a unique solution  $(X, \Lambda)$  with elements in  $H^1(0, T)$ . Introduce the notation

$$(4.3) \quad Z \stackrel{\text{def}}{=} \{s \in [0, T] : \det X(s) = 0\}.$$

By definition,  $T \notin Z$  since  $X(T) = I$ . We now extend the definition of [2] (property (ii) in section 4.2) to a larger class of systems with bounded measurable coefficients.

**DEFINITION 4.2.** *The problem (2.1)–(2.2) is normalizable if  $\det X(t) \neq 0$  almost everywhere in  $[0, T]$ , where  $(X, \Lambda)$  is the  $H^1(0, T)$ -solution of the matrix differential system (4.1).*

Here  $Z$  is possibly infinite with accumulation points that are not isolated, as can be seen from the following example.

*Example 4.1.* Consider an extension of the example from [2, Example 5.1, p. 67]:

$$(4.4) \quad x'(t) = B_1(t)u(t) + B_2(t)v(t) \text{ a.e. in } [0, 2], \quad x(0) = x_0,$$

$$(4.5) \quad C_{x_0}(u, v) = \frac{1}{2}|x(2)|^2 + \int_0^2 |u(t)|^2 - |v(t)|^2 dt,$$

where

$$(4.6) \quad B_1(t) \stackrel{\text{def}}{=} \begin{cases} 2-t, & 1 < t \leq 2, \\ 2^{\frac{n}{2}+1} \left( \frac{1}{2^n} - t \right), & \frac{1}{2^{n+1}} < t \leq \frac{1}{2^n}, \quad n \geq 0, \end{cases}$$

$$(4.7) \quad B_2(t) \stackrel{\text{def}}{=} \begin{cases} t, & 1 < t \leq 2, \\ 2^{\frac{n}{2}+1} \left( t - \frac{1}{2^{n+1}} \right), & \frac{1}{2^{n+1}} < t \leq \frac{1}{2^n}, \quad n \geq 0. \end{cases}$$

It is readily seen that both  $B_1$  and  $B_2$  are measurable and bounded. Here  $A = 0$ ,  $F = 1/2$ ,  $Q = 0$ , and  $R = B_1 B_1^* - B_2 B_2^*$ ,

$$(4.8) \quad R(t) = \begin{cases} 4(1-t), & 1 < t \leq 2, \\ \left( \frac{3}{2^n} - 4t \right), & \frac{1}{2^{n+1}} < t \leq \frac{1}{2^n}, \quad n \geq 0. \end{cases}$$

The solution of system (4.1) is given by the expressions

$$(4.9) \quad \begin{aligned} X(t) &= \begin{cases} (1-t)^2, & 1 < t \leq 2, \\ \left(t - \frac{1}{2^n}\right) \left(t - \frac{1}{2^{n+1}}\right), & \frac{1}{2^{n+1}} < t \leq \frac{1}{2^n}, \quad n \geq 0, \\ 0, & t = 0, \end{cases} \\ \Lambda(t) &= \frac{1}{2}. \end{aligned}$$

Here  $Z = \{1/2^n : n \geq 0\} \cup \{0\}$  has an infinite number of isolated points plus the accumulation point 0 that is not isolated since  $1/2^n \rightarrow 0$ .

Thus Definition 4.2 is a natural extension of the normalizability in the sense of [2] to systems with bounded measurable coefficients.

LEMMA 4.1. *If problem (2.1)–(2.2) is normalizable in the sense of Bernhard [2], then  $Z$  contains a finite number of instants in  $[0, T[$ .*

*Proof.* If the compact set  $Z$  has an infinite number of points, then there exists a sequence of distinct points  $\{t_n\}$  in  $Z$  and an accumulation point  $t_0 \in Z$ ,  $t_n \neq t_0$ , such that  $t_n \rightarrow t_0$ . But this is impossible since each point of  $Z$  is an isolated point in  $[0, T]$ : there is an open interval  $(a, b)$  such that  $t_0 \in (a, b)$  and  $(a, b) \cap [0, T] \setminus \{t_0\} \subset [0, T] \setminus Z$ .  $\square$

Remark 4.1. Since  $Z$  has zero measure we know that it does not contain non-trivial intervals. One interesting issue that was raised by the referee is whether the set  $Z$  is countable or not. It is countable in our example, but can it be uncountable like the Cantor set? However, all the results of this paper are independent of that open issue.

The normalizability property relies on the fact that the state equation can be solved backward in finite dimension. In general, this would not be true for infinite dimensional evolution systems. Yet, in finite dimension, normalizability turns out to be equivalent to a weaker form of *invariant embedding with respect to the initial time* that would be more natural in infinite dimension. Denote by  $Z'$  the set of all initial times  $s \in [0, T[$  such that the matrix differential system

$$(4.10) \quad \begin{cases} \hat{X}'_s = A\hat{X}_s - R\hat{\Lambda}_s, & \hat{X}_s(s) = I, \\ \hat{\Lambda}'_s + A^\top \hat{\Lambda}_s + Q\hat{X}_s = 0, & \hat{\Lambda}_s(T) = F\hat{X}_s(T) \end{cases}$$

has a solution  $(\hat{X}_s, \hat{\Lambda}_s)$  with elements in  $H^1(s, T)$ .

LEMMA 4.2.

(i)  $Z = Z'$ .

(ii) For  $s \in [0, T[ \setminus Z'$ , the matrix differential system (4.10) has a unique solution  $(\hat{X}_s, \hat{\Lambda}_s)$  with elements in  $H^1(s, T)$  for all  $t \in [s, T[ \setminus Z'$ ,  $\det \hat{X}_s(t) \neq 0$ , and  $P(s) = \Lambda(s)X(s)^{-1} = \hat{\Lambda}_s(s)$ .

By convention we set  $\hat{X}_T(T) = I$  and  $\hat{\Lambda}_T(T) = F$ .

Remark 4.2. From part (i) of Lemma 4.2, normalizability is equivalent to invariant embedding with respect to almost all initial times. This equivalence should be compared with (iv) in Theorem 3.1. It says that the decoupling operator  $P(s)$  can be defined almost everywhere as  $\hat{\Lambda}_s(s)$  and that the invariant embedding with respect to almost all initial times can still be done as in [5]. This property was observed for the Riccati differential equation associated with Helmholtz equation<sup>3</sup> of waveguides.

<sup>3</sup>The author would like to thank Jacques Henry (INRIA) for bringing this work to his attention.

Due to a resonance phenomenon, the invariant embedding cannot be done at an at most countable set of initial times. This material can be found in the Ph.D. thesis of Champagne [4], where a sup-inf formulation is also introduced.

*Proof.* (i) For  $s \in [0, T[ \setminus Z$ , it is easy to check that the pair of matrices  $(\hat{X}_s(t), \hat{\Lambda}_s(t)) = (X(t)X(s)^{-1}, \Lambda(t)X(s)^{-1})$  is an  $H^1(s, T)$ -solution of system (4.10). Hence,  $[0, T[ \setminus Z \subset [0, T[ \setminus Z'$  and  $Z' \subset Z$ . Conversely, for all  $s \in [0, T[ \setminus Z'$ ,  $\hat{X}_s(T)$  is invertible. Indeed, if there exists  $h \neq 0$  such that  $\hat{X}_s(T)h = 0$ , then the pair  $(x, p) = (\hat{X}_s h, \hat{\Lambda}_s h)$  would be a solution of the system  $x' = Ax - Rp$ ,  $x(T) = 0$ ,  $p' + A^\top p + Qx = 0$ ,  $p(T) = Fx(T) = 0$ . Hence  $(x, p) = (0, 0)$  and  $0 = x(s) = \hat{X}_s(s)h = h \neq 0$ , a contradiction. Define for  $s \in [0, T[ \setminus Z'$  the new matrices  $(X_s(t), \Lambda_s(t)) = (\hat{X}_s(t)\hat{X}_s(T)^{-1}, \hat{\Lambda}_s(t)\hat{X}_s(T)^{-1})$ . They are a solution of system (4.1) in  $[s, T]$ . Hence  $X_s$  is the restriction of  $X$  to  $[s, T]$ ,  $X(t) = \hat{X}_s(t)\hat{X}_s(T)^{-1}$  on  $[s, T]$ ,  $X(s) = \hat{X}_s(T)^{-1}$ . Since  $\hat{X}_s(T)$  is invertible,  $X(s)$  is invertible,  $s \in [0, T[ \setminus Z$ ,  $[0, T[ \setminus Z' \subset [0, T[ \setminus Z$ , and  $Z \subset Z'$ . Therefore  $Z' = Z$ .

(ii) To prove that the solution of system (4.10) is unique for each  $s \in [0, T[ \setminus Z'$ , consider an arbitrary solution  $(\hat{X}_s, \hat{\Lambda}_s)$ . Define  $(\overline{X}_s(t), \overline{\Lambda}_s(t)) = (\hat{X}_s(t)X(s), \hat{\Lambda}_s(t)X(s))$ . It is readily seen that they are a solution of the system

$$\begin{cases} \overline{X}_s' = A\overline{X}_s - R\overline{\Lambda}_s, & \overline{X}_s(T) = \hat{X}_s(T)X(s), \\ \overline{\Lambda}_s' + A^\top \overline{\Lambda}_s + Q\overline{X}_s = 0, & \overline{\Lambda}_s(T) = F\hat{X}_s(T)X(s). \end{cases}$$

This matrix linear system with final conditions has the unique  $H^1$ -solution  $\overline{X}_s(t) = X(t)\hat{X}_s(T)X(s)$  and  $\overline{\Lambda}_s(t) = \Lambda(t)\hat{X}_s(T)X(s)$ . But we have shown in part (i) that  $\hat{X}_s(T)$  is invertible and that  $X(s) = \hat{X}_s(T)^{-1}$ . Therefore

$$(\hat{X}_s(t), \hat{\Lambda}_s(t)) = (X(t)X(s)^{-1}, \Lambda(t)X(s)^{-1})$$

is unique,  $P(s) = \Lambda(s)X(s)^{-1} = \hat{\Lambda}_s(s)$ , and  $\hat{X}_s(t)$  is invertible for all  $t \in [s, T] \setminus Z'$ .  $\square$

Starting with Definition 4.2 of normalizability, we now proceed in a constructive way to identify the appropriate definition of a closed loop saddle point in the presence of non- $L^2$ -integrable singularities in the closed loop strategies. The normalizability property gives a precise meaning to the *closed loop system* and to a solution of the matrix Riccati differential equation.

LEMMA 4.3. *Assume that the problem (2.1)–(2.2) is normalizable. Then*

- (i)  $P(s) = \Lambda(s)X(s)^{-1}$  is uniquely defined and symmetrical for all  $s \in [0, T] \setminus Z$ ,  $P$  verifies the matrix Riccati differential equation

$$P' + PA + A^\top P - PRP + Q = 0, \quad P(T) = F,$$

in  $[0, T] \setminus Z$ , and  $PX$  is the unique  $H^1(0, T)$ -solution of the matrix equation

$$(PX)' + A^\top(PX) + QX = 0, \quad (PX)(T) = F.$$

- (ii)  $X$  is an  $H^1(0, T)$  solution of the closed loop matrix equation

$$(4.11) \quad X' = (A - RP)X, \quad X(T) = I,$$

such that  $\det X(t) \neq 0$  in  $[0, T] \setminus Z$ , and  $-B_1^\top PX = -B_1^\top \Lambda$  and  $B_2^\top PX = B_2^\top \Lambda$  belong to  $L^2(0, T)$ .

- (iii) for all  $s \in [0, T] \setminus Z$ ,  $\hat{X}_s$  is an  $H^1(s, T)$ -solution of the closed loop matrix differential equation

$$(4.12) \quad \hat{X}'_s = (A - RP)\hat{X}_s, \quad \hat{X}_s(s) = I,$$

such that  $\det \hat{X}_s(t) \neq 0$  in  $[s, T] \setminus Z$ , and  $P\hat{X}_s$  is the unique  $H^1(s, T)$ -solution of the matrix equation

$$(P\hat{X}_s)' + A^\top(P\hat{X}_s) + Q\hat{X}_s = 0, \quad (P\hat{X}_s)(T) = F\hat{X}_s(T).$$

*Proof.* (i) It is easy to verify that the derivative of the matrix function  $X^\top \Lambda - \Lambda^\top X$  is null and that  $(X^\top \Lambda - \Lambda^\top X)(T) = 0$ . Hence  $X^\top \Lambda = \Lambda^\top X$  and  $P = \Lambda X^{-1} = (\Lambda X^{-1})^\top = X^{-\top} \Lambda^\top = P^\top$  almost everywhere in  $[0, T]$ . The second part follows by definition of  $P$  and the identity  $\Lambda = PX$ .

(ii) The second statement follows from the fact that, by definition of  $P$ ,  $X' = AX - R\Lambda = AX - R\Lambda X^{-1}X = (A - RP)X$ .

(iii) Equation (4.12) for  $\hat{X}_s$  follows from identity  $\hat{\Lambda}_s(t) = \hat{\Lambda}_t(t)\hat{X}_s(t) = P(t)\hat{X}_s(t)$ .  $\square$

**4.4. Admissible closed loop affine strategies.** The closed loop strategies  $(\phi, \psi) \in \tilde{\Phi} \times \tilde{\Psi}$  of Definition 4.1 with possible non- $L^2$ -integrable singularities need to be linked through an admissibility condition  $(\phi, \psi) \in S$  to make sense of a solution of the underlying differential equation. It fundamentally changes the nature of the problem. This subtle difference prevents the use of the nice classical results of the theory of saddle points with respect to two fixed independent sets  $\Phi$  and  $\Psi$ . For instance, two pairs  $(\phi_1, \psi_1) \in S$  and  $(\phi_2, \psi_2) \in S$  cannot be mixed:  $(\phi_1, \psi_2)$  and  $(\phi_2, \psi_1)$  need not belong to  $S$  as shown in Example 4.2 for the pairs  $(\phi^*, \psi^*) \in S$  and  $(0, 0) \in S$ . In particular, property (ii) of Berkovitz's equivalence, Lemma 3.2, is not verified for  $u = 0$  or  $v = 0$ .

*Example 4.2.* Consider the example from [2, Example 5.1, p. 67]:

$$(4.13) \quad x'(t) = (2 - t)u(t) + tv(t) \text{ a.e. in } [0, 2], \quad x(0) = x_0,$$

$$(4.14) \quad C_{x_0}(u, v) = \frac{1}{2}|x(2)|^2 + \int_0^2 |u(t)|^2 - |v(t)|^2 dt.$$

Here  $A = 0$ ,  $B_1(t) = 2 - t$ ,  $B_2(t) = t$ ,  $F = 1/2$ ,  $Q = 0$ , and  $R = B_1B_1^* - B_2B_2^* = 4(1 - t)$ . The Riccati equation reduces to

$$P' - 4(1 - t)P^2 = 0, \quad P(2) = \frac{1}{2} \Rightarrow P(t) = \frac{1}{2(t - 1)^2}.$$

Its solution is positive and blows up at  $t = 1$ . The open loop lower value of the game is  $v^-(x_0) = (x_0)^2/2$  and the open loop upper value of the game is  $v^+(x_0) = +\infty$  for all  $x_0 \in \mathbf{R}$ . The closed loop strategies have a singularity in  $t = 1$ :

$$(4.15) \quad \phi^*(t, x) = -\frac{2 - t}{2(t - 1)^2}x \quad \text{and} \quad \psi^*(t, x) = \frac{t}{2(t - 1)^2}x.$$

Yet the state  $x$  is an  $H^1(0, 2)$ -function and the controls  $\hat{u}$  and  $\hat{v}$  are  $L^2$ -functions:

$$(4.16) \quad x(t) = x_0(t - 1)^2, \quad \hat{u}(t) = -(2 - t)\frac{1}{2}x_0, \quad \text{and} \quad \hat{v}(t) = t\frac{1}{2}x_0.$$

Moreover,  $X(t) = (t - 1)^2$ .

In general for the pair  $(\phi^*, v)$ , both the resulting  $L^2$ -norms of the state  $x$  and the control  $u = \phi^*(x)$  will blow up even when  $v = 0$ :

$$(4.17) \quad x'(t) = -\frac{(2-t)^2}{2(t-1)^2}x(t) + tv(t) \text{ in } [0, 2], \quad x(0) = x_0,$$

$$(4.18) \quad \text{for } v = 0, \quad x(t) = e^{\frac{1}{2}[\frac{1}{t-1} - (t-1)]}|t-1|x_0,$$

where  $x(1^-) = 0$  and  $x(1^+) = \infty$ . So the equivalent condition (ii) of Berkovitz's Lemma 3.2 is not verified.

DEFINITION 4.3. *We say that the pair  $(\phi, \psi) \in \tilde{\Phi} \times \tilde{\Psi}$  belongs to  $S$  or simply that  $(\phi, \psi)$  is an admissible pair if the associated matrix differential equation*

$$(4.19) \quad X' = (A + B_1U + B_2V)X, \quad X(T) = I,$$

has an  $H^1(0, T)$ -solution such that  $\det X(t) \neq 0$  almost everywhere in  $[0, T]$ ,  $UX$  and  $VX$  are  $L^2(0, T)$  matrices,  $|X^{-1}|u \in L^2(0, T; \mathbf{R}^m)$ , and  $|X^{-1}|v \in L^2(0, T; \mathbf{R}^k)$ .

Remark 4.3. Since  $A$ ,  $B_1$ , and  $B_2$  are  $L^\infty$ -matrix functions, it implies that the feedback matrix functions  $U$  and  $V$  will have properties similar to the ones of  $X'X^{-1}$ . As a consequence, given an admissible pair  $(\phi, \psi) \in S$  and  $y_0 \in \mathbf{R}^n$ ,  $x(t) = X(t)y_0$  is a solution in  $H^1(0, T; \mathbf{R}^n)$  of the equation

$$(4.20) \quad x' = [A + B_1U + B_2V]x, \quad x(0) = X(0)y_0,$$

(or  $x(T) = y_0$ ) such that

$$u = Ux = UXy_0 \in L^2(0, T; \mathbf{R}^m) \quad \text{and} \quad v = Vx = VXy_0 \in L^2(0, T; \mathbf{R}^k).$$

However, this solution of system (4.20) is not unique. Indeed, if  $x$  is a solution of system (4.20),  $t_i \in Z$ , and  $x_i$  in  $H^1(0, T; \mathbf{R}^n)$  is given by

$$x_i(t) = \begin{cases} 0, & 0 \leq t < t_i \\ X(t)z_i, & t_i \leq t \leq T \end{cases} \quad \forall z_i \in \ker X(t_i),$$

then  $X(t)y_0 + x_i(t)$  is also a solution of system (4.20).

The admissibility condition amounts to a change in the state variable via the associated transformation  $X(t)$ .

LEMMA 4.4. *Assume that  $(\phi, \psi) \in S$ ,  $\phi(t, x) = U(t)x + u_0(t)$ , and  $\psi(t, x) = V(t)x + v_0(t)$ . Let  $X$  be a solution of system (4.19) such that  $\det X(t) \neq 0$  almost everywhere in  $[0, T]$ .*

- (i) *For all  $y_0 \in \mathbf{R}^n$ ,  $u \in L^2(0, T; \mathbf{R}^m)$  such that  $|X^{-1}|u \in L^2(0, T; \mathbf{R}^m)$ , and  $v \in L^2(0, T; \mathbf{R}^k)$  such that  $|X^{-1}|v \in L^2(0, T; \mathbf{R}^k)$ , the system*

$$(4.21) \quad y' = X^{-1}(B_1u + B_2v), \quad y(0) = y_0,$$

*has a unique solution in  $H^1(0, T; \mathbf{R}^n)$ ,  $x \stackrel{\text{def}}{=} Xy$  is the unique solution in*

$$(4.22) \quad H_X^1(0, T; \mathbf{R}^n) \stackrel{\text{def}}{=} \left\{ x \in H^1(0, T; \mathbf{R}^n) : \begin{array}{l} \exists y \in H^1(0, T; \mathbf{R}^n) \\ \text{such that } x = Xy \end{array} \right\}$$

of the system

$$(4.23) \quad x' = [A + B_1U + B_2V]x + B_1u + B_2v, \quad x(0) = X(0)y_0,$$

up to a function of the form  $X(t)z_0$  for some  $z_0 \in \ker X(0)$ , and all the solutions of (4.23) in  $H_X^1(0, T; \mathbf{R}^n)$  are given by the expression

$$(4.24) \quad x(t) = X(t) \left[ y_0 + z_0 + \int_0^t X^{-1}(B_1u + B_2v) ds \right] \quad \forall z_0 \in \ker X(0).$$

- (ii) The subspaces  $\mathcal{U} = \{u \in L^2(0, T; \mathbf{R}^m) : |X^{-1}|u \in L^2(0, T; \mathbf{R}^m)\}$  and  $\mathcal{V} = \{v \in L^2(0, T; \mathbf{R}^k) : |X^{-1}|v \in L^2(0, T; \mathbf{R}^k)\}$  are dense in  $L^2(0, T; \mathbf{R}^m)$  and  $L^2(0, T; \mathbf{R}^k)$ , respectively.

*Proof.* (i) By assumption on  $u$  and  $v$ , the right-hand side of (4.21) belongs to  $L^2(0, T; \mathbf{R}^n)$ , and its unique solution  $y$  belongs to  $H^1(0, T; \mathbf{R}^n)$ . By direct computation of the derivative of  $x = Xy$ , it is easy to check that  $x$  is a solution of system (4.23). Consider two solutions  $x_1$  and  $x_2$  of system (4.23). The difference  $z = x_2 - x_1$  is a solution in  $H_X^1(0, T; \mathbf{R}^n)$  of

$$z' = [A + B_1U + B_2V]z, \quad z(0) = 0.$$

Since  $z \in H_X^1(0, T; \mathbf{R}^n)$ , there exists  $y \in H^1(0, T; \mathbf{R}^n)$  such that  $z = Xy$ . The function  $y = X^{-1}z$  is the solution in  $H^1(0, T)$  of

$$y' = 0 \text{ in } (0, T), \quad y(T) = z(T).$$

Therefore  $y(t) = z(T)$  on  $[0, T]$ . This implies that  $z(t) = X(t)z(T)$  in  $[0, T]$ ,  $0 = z(0) = X(0)z(T)$ , and  $z(T) \in \ker X(0)$ . Therefore two solutions of system (4.23) can differ only by  $X(t)z_0$  for some  $z_0 \in \ker X(0)$ .

- (ii) Define  $w_n(t) = 1$  if  $|X(t)| \geq 1/n$  and  $0$  if  $|X(t)| < 1/n$ . For any  $u \in L^2(0, T; \mathbf{R}^m)$ , the sequence  $\{u_n = u w_n\} \subset \mathcal{U}$  converges to  $u$  in  $L^2(0, T; \mathbf{R}^m)$  by the Lebesgue dominated convergence theorem.  $\square$

As for normalizability, Definition 4.3 is equivalent to the following definition.

**DEFINITION 4.4.** We say that the pair  $(\phi, \psi) \in \tilde{\Phi} \times \tilde{\Psi}$  belongs to  $S$  or simply that  $(\phi, \psi)$  is an admissible pair if, for almost all  $s \in [0, T]$ , the matrix differential equation

$$(4.25) \quad X'_s = (A + B_1U + B_2V)X_s, \quad X_s(s) = I,$$

has an  $H^1(s, T)$ -solution,  $UX_s$  and  $VX_s$  are  $L^2(s, T)$  matrices,  $|X_s^{-1}|u \in L^2(s, T; \mathbf{R}^m)$ , and  $|X_s^{-1}|v \in L^2(s, T; \mathbf{R}^k)$ .

#### 4.5. Necessary and sufficient conditions for normalizability.

**LEMMA 4.5.** Assume that problem (2.1)–(2.2) is normalizable. Let  $(X, \Lambda)$  be the solution of system (4.1),  $Z = \{t \in [0, T] : \det X(t) = 0\}$ , and let  $P$  be defined by (4.2).

- (i)  $X$  is a solution in  $H^1(0, T)$  of the matrix equation (4.11),

$$X' = (A - RP)X, \quad X(T) = I,$$

such that  $\det X(t) \neq 0$  in  $[0, T] \setminus Z$ ,

$$(4.26) \quad \begin{aligned} U_*X &= -B_1^\top PX = -B_1^\top \Lambda \in L^2(0, T; \mathbf{R}^n)^n, \\ V_*X &= B_2^\top PX = B_2^\top \Lambda \in L^2(0, T; \mathbf{R}^n)^n \end{aligned}$$



for the matrices

$$(4.27) \quad U_*(t) \stackrel{\text{def}}{=} -B_1^\top(t)P(t) \text{ and } V_*(t) \stackrel{\text{def}}{=} B_2^\top(t)P(t)$$

and the pair  $(\phi^*, \psi^*) \in S$ , where  $(\phi^*(t, x), \psi^*(t, x)) = (U_*(t)x, V_*(t)x)$ .

- (ii) Given  $x_0 \in \text{Im } X(0)$ ,  $u \in L^2(0, T; \mathbf{R}^m)$  such that  $|X^{-1}|u \in L^2(0, T; \mathbf{R}^m)$ , and  $v \in L^2(0, T; \mathbf{R}^k)$  such that  $|X^{-1}|v \in L^2(0, T; \mathbf{R}^k)$ , the system

$$(4.28) \quad x' = [A - RP]x + B_1u + B_2v, \quad x(0) = x_0,$$

has a solution  $x \in H_X^1(0, T; \mathbf{R}^n)$  unique up to an element  $X(t)z_0$  for some  $z_0 \in \ker X(0)$ . Moreover, for any solution  $x \in H_X^1(0, T; \mathbf{R}^n)^n$  of (4.28),

$$(4.29) \quad C_{x_0}(\phi^*(x) + u, \psi^*(x) + v) = \Lambda(0)y_0 \cdot x_0 + \int_0^T |u|^2 - |v|^2 dt,$$

$C_{x_0}(\phi^*, \psi^*) = \Lambda(0)y_0 \cdot x_0$  is independent of the choice of  $y_0$  such that  $X(0)y_0 = x_0$ , and  $C_{x_0}(\phi^*(x) + u, \psi^*(x) + v)$  is independent of the choice of the solution  $x$  in  $H_X^1(0, T; \mathbf{R}^n)$  to system (4.28).

- (iii) For all  $x_0 \in \text{Im } X(0)$ ,  $u \in L^2(0, T; \mathbf{R}^m)$  such that  $|X^{-1}|u \in L^2(0, T; \mathbf{R}^m)$ , and  $v \in L^2(0, T; \mathbf{R}^k)$  such that  $|X^{-1}|v \in L^2(0, T; \mathbf{R}^k)$ ,

$$(4.30) \quad C_{x_0}(\phi^*, \psi^* + v) \leq C_{x_0}(\phi^*, \psi^*) \leq C_{x_0}(\phi^* + u, \psi^*).$$

*Remark 4.4.* In view of Lemma 4.4(ii), inequalities (4.30) are verified on dense subsets of  $L^2(0, T; \mathbf{R}^m)$  and  $L^2(0, T; \mathbf{R}^k)$ . They define an open loop saddle point in  $(0, 0) \in \mathcal{U} \times \mathcal{V}$  after a change of the state variable via the transformation  $X(t)$ .

*Proof.* (i) This follows from Lemma 4.3(ii).

(ii) Let  $x$  be a solution of system (4.28) in  $H_X^1(0, T; \mathbf{R}^n)$  and  $y_0$  be such that  $X(0)y_0 = x_0$ . By Lemma 4.4(i), the solutions  $x$  of system (4.28) in  $H_X^1(0, T; \mathbf{R}^n)$  are given by (4.24),

$$x(t) = X(t)y(t), \quad y(t) = y_0 + z_0 + \int_0^t X^{-1}(B_1u + B_2v) ds \quad \forall z_0 \in \ker X(0),$$

and, by definition of  $P(t)$ ,  $Px = \Lambda y$ . Hence  $Px$  is an  $H^1(0, T; \mathbf{R}^n)$ -function. Differentiate the inner product  $\Lambda y \cdot x$  as follows:

$$\begin{aligned} \frac{d}{dt} \Lambda y \cdot x &= \Lambda' y \cdot x + \Lambda y' \cdot x + \Lambda y \cdot x' \\ &= -[A^\top \Lambda + QX]y \cdot x + \Lambda X^{-1}(B_1u + B_2v) \cdot x + \Lambda y \cdot [(A - RP)x + B_1u + B_2v] \\ &= -[Qx \cdot x + | - B_1^\top Px + u|^2 + |B_2^\top Px + v|^2] + |u|^2 - |v|^2 \end{aligned}$$

and

$$\begin{aligned} C_{x_0}(\phi^* + u, \psi^* + v) &= \Lambda(0)(y_0 + z_0) \cdot x_0 + \int_0^T |u|^2 - |v|^2 dt \\ &\Rightarrow C_{x_0}(\phi^*, \psi^*) = \Lambda(0)(y_0 + z_0) \cdot x_0. \end{aligned}$$

But this last expression is dependent only on  $x_0$ . Assume that there exist  $y_0^1$  and  $y_0^2$  such that  $X(0)y_0^1 = x_0 = X(0)y_0^2$ . Let  $z_0^1$  and  $z_0^2$  be the respective elements of  $\ker X(0)$  associated with  $y_0^1$  and  $y_0^2$ . Then  $y_0^2 + z_0^2 - (y_0^1 - z_0^1) \in \ker X(0)$ . By the symmetry in

Lemma 4.3,  $X^\top(0)\Lambda(0)(y_0^2 + z_0^2 - (y_0^1 - z_0^1)) = \Lambda(0)^\top X(0)(y_0^2 + z_0^2 - (y_0^1 - z_0^1)) = 0$ . Hence

$$\begin{aligned} & \Lambda(0)(y_0^2 + z_0^2) \cdot x_0 - \Lambda(0)(y_0^1 + z_0^1) \cdot x_0 \\ &= \Lambda(0)(y_0^2 + z_0^2 - (y_0^1 - z_0^1)) \cdot x_0 = X(0)^\top \Lambda(0)(y_0^2 + z_0^2 - (y_0^1 - z_0^1)) \cdot y_0^1 = 0. \end{aligned}$$

The value of  $C_{x_0}(\phi^*, \psi^*)$  depends only on  $x_0$ .  $C_{x_0}(\phi^* + u, \psi^* + v)$  depends only on  $x_0$ ,  $u$ , and  $v$  and is independent of the choice of a solution  $x$  in  $H_X^1(0, T; \mathbf{R}^n)$  to system (4.28).

(iii) It is readily seen from part (ii) that

$$\forall u \in \mathcal{U}, \forall v \in \mathcal{V}, \quad C_{x_0}(\phi^*, \psi^* + v) \leq C_{x_0}(\phi^*, \psi^*) \leq C_{x_0}(\phi^* + u, \psi^*),$$

and  $(\hat{u}, \hat{v}) = (0, 0)$  is an open loop saddle point.  $\square$

We are now ready to prove the following result that sheds light on the choice of a definition of the closed loop saddle point in the presence of closed loop strategies with non- $L^2$ -integrable singularities.

THEOREM 4.1. *The following statements are equivalent.*

- (i) *Problem (2.1)–(2.2) is normalizable in the sense of Definition 4.2.*
- (ii) *There exists a pair of closed loop strategies  $(\phi^*, \psi^*) \in S$  that for all  $x_0 \in \text{Im } X(0)$ , all  $u \in L^2(0, T; \mathbf{R}^m)$  such that  $|X^{-1}|u \in L^2(0, T; \mathbf{R}^m)$ , and all  $v \in L^2(0, T; \mathbf{R}^k)$  such that  $|X^{-1}|v \in L^2(0, T; \mathbf{R}^k)$ ,*

$$(4.31) \quad C_{x_0}(\phi^*, \psi^* + v) \leq C_{x_0}(\phi^*, \psi^*) \leq C_{x_0}(\phi^* + u, \psi^*).$$

- (iii) *There exists a pair of linear closed loop strategies  $(\phi^*, \psi^*) \in S$  (that is,  $\psi^*(t, x) = V_*(t)x$  and  $\phi^*(t, x) = U_*(t)x$ ) such that for all  $x_0 \in \text{Im } X(0)$ , there exists a solution  $(\hat{x}, \hat{p}) \in H_X^1(0, T; \mathbf{R}^n) \times H^1(0, T; \mathbf{R}^n)$  of*

$$(4.32) \quad \begin{cases} \hat{x}' = A\hat{x} - R\hat{p}, & \hat{x}(0) = x_0, \\ \hat{p}' + A^\top \hat{p} + Q\hat{x} = 0, & \hat{p}(T) = F\hat{x}(T), \end{cases}$$

and

$$(4.33) \quad -B_1^\top \hat{p} = U_* \hat{x}, \quad B_2^\top \hat{p} = V_* \hat{x}.$$

For all  $x_0 \in \text{Im } X(0)$  the feedback strategies associated with  $C_{x_0}$  are given by

$$(4.34) \quad \phi^*(t, x) = -B_1^\top(t)P(t)x \quad \text{and} \quad \psi^*(t, x) = B_2^\top(t)P(t)x,$$

where  $P$  is defined by (4.2), and the value of the closed loop saddle point by

$$(4.35) \quad C_{x_0}(\phi^*, \psi^*) = \Lambda(0)y_0 \cdot x_0$$

for some  $y_0 \in \mathbf{R}^n$  such that  $x_0 = X(0)y_0$ , and this value is independent of the choice of  $y_0$  such that  $x_0 = X(0)y_0$ .

*Proof.* (i)  $\Rightarrow$  (ii). This follows by Lemma 4.5.

(ii)  $\Rightarrow$  (iii). Denote by  $U_*$  and  $V_*$  and  $u_*$  and  $v_*$  the matrix and vector functions associated with the pair  $(\phi^*, \psi^*) \in S$ . By Definition 4.3 and Lemma 4.4, there exists a solution  $\bar{X}$  in  $H_X^1(0, T)$  to the matrix differential equation (4.19),

$$(4.36) \quad \bar{X}' = (A + B_1 U_* + B_2 V_*) \bar{X}, \quad \bar{X}(T) = I,$$

such that  $\det \bar{X}(t) \neq 0$  almost everywhere in  $[0, T]$ ,  $U_* \bar{X}$ , and  $V_* \bar{X}$  in  $L^2(0, T)$ ;  $u_* \in L^2(0, T; \mathbf{R}^m)$  such that  $|\bar{X}^{-1}|u_* \in L^2(0, T; \mathbf{R}^k)$ , and  $v_* \in L^2(0, T; \mathbf{R}^m)$  such that  $|\bar{X}^{-1}|v_* \in L^2(0, T; \mathbf{R}^k)$ .

We know that for all  $x_0 \in \text{Im } \bar{X}(0)$ ,  $(\hat{u}, \hat{v}) = (0, 0)$  is an open loop saddle point of  $C_{x_0}(\phi^*(x) + u, \psi^*(x) + v)$ , where  $x$  is a solution in  $H_{\bar{X}}^1(0, T; \mathbf{R}^n)$  of the state equation

$$x' = (A + B_1 U_* + B_2 V_*)x + B_1(u_* + u) + B_2(v_* + v), \quad x(0) = x_0.$$

To get around the nonuniqueness of solution, we start from the other state equation,

$$y' = \bar{X}^{-1}(B_1(u_* + u) + B_2(v_* + v)), \quad y(0) = y_0,$$

for all  $y_0 \in \mathbf{R}^n$  and  $u \in L^2(0, T; \mathbf{R}^m)$  such that  $|\bar{X}^{-1}|u \in L^2(0, T; \mathbf{R}^k)$ , and  $v \in L^2(0, T; \mathbf{R}^m)$  such that  $|\bar{X}^{-1}|v \in L^2(0, T; \mathbf{R}^k)$ . By construction,  $x(t) = \bar{X}(t)y(t)$ , and we just substitute in the utility function to get it in terms of  $y$  rather than  $x$ :  $c_{y_0}(u, v) = C_{\bar{X}(0)y_0}(\phi^*(\bar{X}y) + u, \psi^*(\bar{X}y) + v)$ . This will take care of all  $x_0 \in \text{Im } \bar{X}(0)$ . From the closed loop saddle point inequalities,

$$(4.37) \quad c_{y_0}(0, v) \leq c_{y_0}(0, 0) \leq c_{y_0}(u, 0)$$

for all  $u \in L^2(0, T; \mathbf{R}^m)$  such that  $|\bar{X}^{-1}|u \in L^2(0, T; \mathbf{R}^k)$  and  $v \in L^2(0, T; \mathbf{R}^k)$  such that  $|\bar{X}^{-1}|v \in L^2(0, T; \mathbf{R}^k)$ , the pair  $(0, 0)$  achieves an open loop saddle point. By [5, Lemma 3.1],  $c_{y_0}(u, v)$  is convex-concave and  $dc_{y_0}(0, 0; u, v) = 0$  for all  $u$  and  $v$ . A direct computation yields

$$(4.38) \quad \begin{aligned} \frac{1}{2}dc_{y_0}(0, 0; u, v) &= F\hat{y}(T) \cdot z(T) + \int_0^T Q\bar{X}\hat{y} \cdot \bar{X}z + (U_*\bar{X}\hat{y} + u_*) \cdot (U_*\bar{X}z + u) \\ &\quad - (V_*\bar{X}\hat{y} + v_*) \cdot (V_*\bar{X}z + v) dt, \\ \hat{y}' &= \bar{X}^{-1}(B_1 u_* + B_2 v_*), \quad \hat{y}(0) = y_0, \quad z' = \bar{X}^{-1}(B_1 u + B_2 v), \quad z(0) = 0. \end{aligned}$$

By introducing the solution  $\pi \in H^1(0, T; \mathbf{R}^n)$  of the adjoint equation

$$(4.39) \quad \begin{aligned} \pi' + \bar{X}^\top Q\bar{X}\hat{y} + \bar{X}^\top U_*^\top (U_*\bar{X}\hat{y} + u_*) - \bar{X}^\top V_*^\top (V_*\bar{X}\hat{y} + v_*) &= 0, \\ \pi(T) &= F\hat{y}(T), \end{aligned}$$

we get

$$0 = \frac{1}{2}dc_{y_0}(0, 0; u, v) = \int_0^T (B_1^\top \bar{X}^{-\top} \pi + U_* \bar{X} \hat{y} + u_*) \cdot u + (B_2^\top \bar{X}^{-\top} \pi - V_* \bar{X} \hat{y} - v_*) \cdot v dt.$$

By the density of  $\mathcal{U}$  in  $L^2(0, T; \mathbf{R}^m)$  and  $\mathcal{V}$  in  $L^2(0, T; \mathbf{R}^k)$  in Lemma 4.4(ii),

$$(4.40) \quad \begin{aligned} B_1^\top \bar{X}^{-\top} \pi + U_* \bar{X} \hat{y} + u_* &= 0 \text{ and } B_2^\top \bar{X}^{-\top} \pi - V_* \bar{X} \hat{y} - v_* = 0 \\ \Rightarrow -B_1^\top \bar{X}^{-\top} \pi &\in L^2(0, T; \mathbf{R}^m) \text{ and } B_2^\top \bar{X}^{-\top} \pi \in L^2(0, T; \mathbf{R}^k). \end{aligned}$$

Using the above identities (4.40) to eliminate  $u^*$  and  $v^*$ , we see that the pair  $(\hat{y}, \pi)$  is a solution in  $H^1(0, T; \mathbf{R}^n) \times H^1(0, T; \mathbf{R}^n)$  of the linear system

$$(4.41) \quad \begin{cases} \pi' + \bar{X}^\top Q\bar{X}\hat{y} - \bar{X}^\top (U_*^\top B_1^\top + V_*^\top B_2^\top) \bar{X}^{-\top} \pi = 0, & \pi(T) = F\hat{y}(T), \\ \hat{y}' = \bar{X}^{-1} [-R\bar{X}^{-\top} \pi - (B_1 U_* + B_2 V_*) \bar{X} \hat{y}], & \hat{y}(0) = y_0. \end{cases}$$

However, the solution of the equation for  $\pi$  is not unique as an element of  $H^1(0, T; \mathbf{R}^n)$ . Let  $\hat{p} \in H^1(0, T; \mathbf{R}^n)$  be the solution of the equation

$$(4.42) \quad \hat{p}' + A^\top \hat{p} + Q\bar{X}\hat{y} = 0, \quad \hat{p}(T) = F\hat{y}(T),$$

and consider the  $H^1(0, T; \mathbf{R}^n)$ -function  $\bar{X}^\top \hat{p}$ . A direct computation using equation (4.36) for  $\bar{X}$  shows that  $\bar{X}^\top \hat{p}$  is also a solution of the first equation of (4.41). So we can choose a solution  $\pi$  in the space

$$(4.43) \quad H_{\bar{X}^\top}^1(0, T; \mathbf{R}^n) \stackrel{\text{def}}{=} \left\{ q \in H^1(0, T; \mathbf{R}^n) : \begin{array}{l} \exists r \in H^1(0, T; \mathbf{R}^n) \\ \text{such that } q = \bar{X}^\top r \end{array} \right\}.$$

Moreover, by using  $\pi = \bar{X}^\top \hat{p}$  and by introducing the variable  $\hat{x} \stackrel{\text{def}}{=} \bar{X} \hat{y}$  in  $H_{\bar{X}}^1(0, T; \mathbf{R}^n)$ , we finally get that, for all  $y_0 \in \mathbf{R}^n$ , the pair  $(\hat{x}, \hat{p})$  is a solution in  $H_{\bar{X}}^1(0, T; \mathbf{R}^n) \times H^1(0, T; \mathbf{R}^n)$  of the system

$$(4.44) \quad \begin{cases} \hat{x}' = A\hat{x} - R\hat{p}, & \hat{x}(0) = X(0)y_0, \\ \hat{p}' + A^\top \hat{p} + Q\hat{x} = 0, & \hat{p}(T) = F\hat{x}(T), \end{cases}$$

and from identity (4.40),

$$(4.45) \quad B_1^\top \hat{p} + U_* \hat{x} + u_* = 0, \quad B_2^\top \hat{p} - V_* \hat{x} - v_* = 0.$$

By linearity of system (4.44), the pair  $(2\hat{x}, 2\hat{p})$  is also a solution for the initial condition  $2y_0$ , and necessarily,

$$(4.46) \quad 2(B_1^\top \hat{p} + U_* \hat{x}) + u_* = 0, \quad 2(B_2^\top \hat{p} - V_* \hat{x}) - v_* = 0.$$

The last two sets of identities yield  $u_* = 0$ ,  $v_* = 0$ , and

$$(4.47) \quad -B_1^\top \hat{p} = U_* \hat{x}, \quad B_2^\top \hat{p} = V_* \hat{x}.$$

From this we conclude that the strategies  $(\phi^*, \psi^*) \in S$  are linear since  $\psi^*(t, x) = V_*(t)x$  and  $\phi^*(t, x) = U_*(t)x$ .

(iii)  $\Rightarrow$  (i). Let  $\bar{X}$  be an  $H^1(0, T)$ -solution of the matrix equation  $\bar{X}' = (A + B_1 U_* + B_2 V_*)\bar{X}$ ,  $\bar{X}(T) = I$ . By substituting identities (4.33) into the first equation of (4.32), we get for all  $x_0 \in \text{Im } \bar{X}(0)$ ,  $\hat{x}' = (A + B_1 U_* + B_2 V_*)\hat{x}$ ,  $\hat{x}(0) = x_0$ . By Lemma 4.4(i), the solutions of this equation in  $H_{\bar{X}}^1(0, T; \mathbf{R}^n)$  are  $\bar{X}(t)y_0$  for all  $y_0 \in \mathbf{R}^n$  such that  $x_0 = X(0)y_0$  and, in particular,  $\hat{x}(T) = y_0$ . As a consequence, for all  $y_0 \in \mathbf{R}^n$  the coupled system

$$(4.48) \quad \begin{cases} \hat{x}' = A\hat{x} - R\hat{p}, & \hat{x}(T) = y_0, \\ \hat{p}' + A^\top \hat{p} + Q\hat{x} = 0, & \hat{p}(T) = Fy_0 \end{cases}$$

has a solution  $(\hat{x}, \hat{p}) \in H_{\bar{X}}^1(0, T; \mathbf{R}^n) \times H^1(0, T; \mathbf{R}^n)$  such that  $\hat{x}(t) = \bar{X}(t)y_0$ . But the solution of system (4.48) with final conditions is unique. By linearity, there exists a unique  $H^1(0, T)$ -solution  $(X, \Lambda)$  to the matrix system

$$(4.49) \quad \begin{cases} X' = AX - R\Lambda, & X(T) = I, \\ \Lambda' + A^\top \Lambda + QX = 0, & \Lambda(T) = F, \end{cases}$$

such that  $\hat{x} = Xy_0$  and  $\hat{p} = \Lambda y_0$ . By uniqueness, for all  $y_0 \in \mathbf{R}^n$ ,  $\hat{x}(t) = \bar{X}(t)y_0$ . Hence  $X(t) = \bar{X}(t)$  and  $\det X(t) \neq 0$  for almost all  $t$  in  $[0, T]$ . This shows that the system is normalizable and completes the proof.  $\square$

**4.6. Definition of a closed loop saddle point.** In view of Theorem 4.1, we adopt the following definition that says that the original problem can be changed via feedback in such a way that the resulting problem has an open loop saddle point at  $(0, 0)$ . It will still be referred to as a *closed loop saddle point*, yet it is more a notion of *structural saddle point*.

DEFINITION 4.5.

- (i) Given  $x_0 \in \mathbf{R}^n$ ,  $(\phi^*, \psi^*) \in S$  is a closed loop saddle point of  $C_{x_0}$  if there exists a solution in  $H_X^1(0, T; \mathbf{R}^n)$  to the state equation

$$(4.50) \quad \hat{x}' = (A + B_1 U_* + B_2 V_*)\hat{x} + B_1 u_* + B_2 v_*, \quad \hat{x}(0) = x_0,$$

and for all solutions  $\hat{x} \in H_X^1(0, T; \mathbf{R}^n)$  of (4.50), all  $u \in L^2(0, T; \mathbf{R}^m)$  such that  $|X^{-1}|u \in L^2(0, T; \mathbf{R}^m)$ , all  $v \in L^2(0, T; \mathbf{R}^k)$  such that  $|X^{-1}|v \in L^2(0, T; \mathbf{R}^k)$ , and all solutions  $x_u, x_v \in H_X^1(0, T; \mathbf{R}^n)$  of the state equations

$$(4.51) \quad x'_u = (A + B_1 U_* + B_2 V_*)x_u + B_1(u_* + u) + B_2 v_*, \quad x_u(0) = x_0,$$

$$(4.52) \quad x'_v = (A + B_1 U_* + B_2 V_*)x_v + B_1 u_* + B_2(v_* + v), \quad x_v(0) = x_0,$$

the following inequalities are verified:

$$(4.53) \quad C_{x_0}(\phi^*(x_v), \psi^*(x_v) + v) \leq C_{x_0}(\phi^*(\hat{x}), \psi^*(\hat{x})) \leq C_{x_0}(\phi^*(x_u) + u, \psi^*(x_u)).$$

- (ii) We say that  $(\phi^*, \psi^*) \in S$  is an  $X(0)$ -global closed loop saddle point of  $C_{x_0}$  if for all  $x_0 \in \text{Im } X(0)$  the inequalities (4.53) are verified.

*Remark 4.5.* By definition of  $H_X^1(0, T; \mathbf{R}^n)$ , there exists  $y \in H^1(0, T; \mathbf{R}^n)$  such that  $\hat{x} = Xy$ ,  $x_0 = \hat{x}(0) = X(0)y(0)$ , and the existence of a solution to the state equation (4.50) implies that  $x_0 \in \text{Im } X(0)$ .

*Remark 4.6.* If  $(\phi^*, \psi^*) \in \Phi \times \Psi$  is an  $L^2$ -integrable global closed loop saddle point in the sense of Definition 3.2(ii), it is an  $X(0)$ -global closed loop saddle point by the equivalence of parts (i) and (ii) of Theorem 3.1.

**4.7. Closed loop saddle points when  $v(x_0)$  is finite.** We first study closed loop saddle points when  $v(x_0)$  is finite.

THEOREM 4.2. Given  $x_0 \in \mathbf{R}^n$ , the following statements are equivalent.

- (i)  $C_{x_0}$  has a closed loop saddle point  $(\phi^*, \psi^*) \in S$ , and  $C_{x_0}(u, v)$  is convex in  $u$  and concave in  $v$ .
- (ii)  $C_{x_0}$  has a closed loop saddle point  $(\phi^*, \psi^*) \in \Phi \times \Psi$ , and  $C_{x_0}(u, v)$  is convex in  $u$  and concave in  $v$  (case (a) of Theorem 3.2).
- (iii)  $C_{x_0}$  has an open loop saddle point.

*Proof.* (i)  $\Rightarrow$  (iii). By Remark 4.5,  $x_0 \in \text{Im } X(0)$ : there exists  $y_0$  such that  $x_0 = X(0)y_0$ . From part (ii) of the proof of Theorem 4.1, there exists a solution  $(\hat{x}, \hat{p})$  in  $H^1(0, T; \mathbf{R}^n)^2$  to the coupled system (4.48) with  $\hat{x}(0) = X(0)y_0$ . Since  $C_{x_0}(u, v)$  is convex in  $u$  and concave in  $v$ ,  $C_{x_0}$  has an open loop saddle point by [5, Thm. 2.4].

(iii)  $\Rightarrow$  (ii). Denote by  $(\hat{u}, \hat{v})$  the pair achieving the open loop saddle point. By [5, Thm. 2.4],  $C_{x_0}(u, v)$  is convex in  $u$  and concave in  $v$ . By definition of  $\Phi \times \Psi$ ,  $(\hat{u}, \hat{v}) \in \Phi \times \Psi$ . By inequalities (3.20) and (3.21) in Lemma 3.3,

$$\begin{aligned} \inf_{\phi \in \Phi} C_{x_0}(\phi, \hat{v}) &= \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, \hat{v}) = C_{x_0}(\hat{u}, \hat{v}), \\ \sup_{\psi \in \Psi} \inf_{\phi \in \Phi} C_{x_0}(\phi, \psi) &\geq v^-(x_0) \geq \inf_{\phi \in \Phi} C_{x_0}(\phi, \hat{v}) = C_{x_0}(\hat{u}, \hat{v}). \end{aligned}$$

By inequalities (3.22) and (3.23) in Lemma 3.3,

$$\begin{aligned} \sup_{\psi \in \Psi} C_{x_0}(\hat{u}, \psi) &= \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_{x_0}(\hat{u}, v) = C_{x_0}(\hat{u}, \hat{v}), \\ \inf_{\phi \in \Phi} \sup_{\psi \in \Psi} C_{x_0}(\phi, \psi) &\leq v^+(x_0) \leq \sup_{\psi \in \Psi} C_{x_0}(\hat{u}, \psi) = C_{x_0}(\hat{u}, \hat{v}). \end{aligned}$$

Hence, there exists  $(\hat{u}, \hat{v}) \in \Phi \times \Psi$  such that for all  $\phi \in \Phi$  and all  $\psi \in \Psi$ ,

$$C_{x_0}(\hat{u}, \psi) \leq C_{x_0}(\hat{u}, \hat{v}) \leq C_{x_0}(\phi, \hat{v}).$$

By Lemma 3.2,  $(\hat{u}, \hat{v})$  is a closed loop saddle point of  $C_{x_0}$  in  $\Phi \times \Psi$ .

(ii)  $\Rightarrow$  (i). By assumption,  $C_{x_0}$  is convex-concave. As in Remark 4.6, if  $(\phi^*, \psi^*) \in \Phi \times \Psi$  is an  $L^2$ -integrable closed loop saddle point in the sense of Definition 3.2(i), it is a closed loop saddle point in the sense of Definition 4.5(i). Hence this proof follows from the proof of (i)  $\Rightarrow$  (ii) of Theorem 3.1.  $\square$

We also have a global version of the previous theorem (case (a)).

**THEOREM 4.3.** *The following statements are equivalent.*

- (i)  $C_{x_0}$  has an  $X(0)$ -global closed loop saddle point  $(\phi^*, \psi^*) \in S$ ,  $X(0) = \mathbf{R}^n$ , and  $C_{x_0}(u, v)$  is convex in  $u$  and concave in  $v$ .
- (ii)  $C_{x_0}$  has a global closed loop saddle point  $(\phi^*, \psi^*) \in \Phi \times \Psi$  and  $C_{x_0}(u, v)$  is convex in  $u$  and concave in  $v$ .
- (iii) For all  $x_0 \in \mathbf{R}^n$ ,  $C_{x_0}$  has an open loop saddle point.

*Proof.* (i)  $\Rightarrow$  (iii). This follows from the equivalence of (i) and (iii) in Theorem 4.2.

(iii)  $\Rightarrow$  (ii). By [5, Thm. 2.4],  $C_{x_0}(u, v)$  is convex in  $u$  and concave in  $v$ . By [5, Thm. 2.9], there exists a unique symmetrical solution to the Riccati equation with elements in  $H^1(0, T)$ . By Theorem 3.1,  $C_{x_0}$  has a global closed loop saddle point  $(\phi^*, \psi^*) \in \Phi \times \Psi$ .

(ii)  $\Rightarrow$  (i). By Theorem 3.1(v),  $X(t)$  is invertible for all  $[0, T]$ ,  $\text{Im } X(0) = \mathbf{R}^n$ , and the problem is normalizable. By Theorem 4.1,  $C_{x_0}$  has an  $X(0)$ -global closed loop saddle point in  $S$  and  $X(0) = \mathbf{R}^n$ .  $\square$

**4.8. Closed loop saddle points when either  $v^-(x_0)$  or  $v^+(x_0)$  is not finite.** From Theorem 4.2, when the value of the game  $v(x_0)$  is finite, the closed loop strategy in  $S$  can be trivially chosen as  $(\phi^*(t, x), \psi^*(t, x)) = (\hat{u}(t), \hat{v}(t))$ ; from Theorem 4.3, when the value of the game  $v(x_0)$  is finite for all  $x_0 \in \mathbf{R}^n$ , the global closed loop strategy is equal to the  $L^2$ -integrable closed loop strategy  $(\phi^*, \psi^*) = (-B_1^\top P \hat{x}, B_2^\top P \hat{x}) \in \Phi \times \Psi$ , where  $P$  is the  $H^1(0, T)$  solution of the Riccati differential equation. The conclusion is that *closed loop strategies with non- $L^2$ -integrable singularities will only occur when either  $v^-(x_0)$  or  $v^+(x_0)$  is not finite.*

The new Definition 4.5 of a closed loop saddle point was introduced to accommodate closed loop strategies with non- $L^2$ -integrable singularities, but its relation to the open loop values is not as straightforward as in the  $L^2$ -integrable case of Theorem 3.2. Yet, a complete classification is obtained for the six cases along the lines of Theorem 3.2 in terms of the  $u$ -convexity and  $v$ -concavity of the utility function.

**THEOREM 4.4.** *Assume that  $C_{x_0}$  has a closed loop saddle point  $(\phi^*, \psi^*) \in S$ . Denote by  $(\hat{u}, \hat{v}) = (\phi^*, \psi^*)$  the associated controls.*

- (i) (case (b)).  $v^-(x_0)$  finite and  $v^+(x_0) = +\infty$  if and only if  $C_{x_0}(u, v)$  is convex in  $u$  and not concave in  $v$ , and  $v \mapsto \inf_v C_{x_0}(u, v)$  is concave. In that case,

$$(4.54) \quad v^-(x_0) = \inf_{u \in L^2(0, T; \mathbf{R}^n)} C_{x_0}(u, \hat{v}) = C_{x_0}(\hat{u}, \hat{v}) = C_{x_0}(\phi^*, \psi^*).$$

- (ii) (case (e)).  $v^-(x_0) = v^+(x_0) = +\infty$  if and only if  $C_{x_0}(u, v)$  is convex in  $u$  and not concave in  $v$ , and  $v \mapsto \inf_v C_{x_0}(u, v)$  is not concave. In that case,

$$\sup_{\psi \in \Psi} \inf_{\phi \in \Phi} C_{x_0}(\phi, \psi) = \inf_{\phi \in \Phi} \sup_{\psi \in \Psi} C_{x_0}(\phi, \psi) = +\infty.$$

*Proof.* From the proof of Theorem 4.1, there exists a solution  $(\hat{x}, \hat{p})$  in  $H^1(0, T; \mathbf{R}^n)^2$  to the coupled system (4.48) ( $x_0 \in \text{Im } X(0)$ ). Since  $C_{x_0}(u, v)$  is convex in  $u$ , we have

$$v^-(x_0) \geq \inf_{u \in L^2(0, T; \mathbf{R}^m)} C_{x_0}(u, \hat{v}) = C_{x_0}(\hat{u}, \hat{v}) > -\infty$$

for the pair  $(\hat{u}, \hat{v}) = (-B_1^\top \hat{p}, B_2^\top \hat{p})$ . Moreover, since  $C_{x_0}(u, v)$  is not concave in  $v$ , then  $v^+(x_0) = +\infty$ . As a result, only two cases can occur:  $v^-(x_0)$  finite and  $v^+(x_0) = +\infty$  or  $v^-(x_0) = v^+(x_0) = +\infty$ . The property in (ii) follows from Lemma 3.3.  $\square$

*Remark 4.7.* Contrarily to the  $L^2$ -integrable closed loop saddle point, case (e) can occur, as can be seen from the system and utility function (2.25).

Since the cases (c) and (f) are dual of cases (b) and (e), we have the dual result.

**THEOREM 4.5.** Assume that  $C_{x_0}$  has a closed loop saddle point  $(\phi^*, \psi^*) \in S$ . Denote by  $(\hat{u}, \hat{v}) = (\phi^*, \psi^*)$  the associated controls.

- (i) (case (c)).  $v^+(x_0)$  finite and  $v^-(x_0) = -\infty$  if and only if  $C_{x_0}(u, v)$  is concave in  $v$  and not convex in  $u$  and  $u \mapsto \sup_v C_{x_0}(u, v)$  is convex. In that case

$$(4.55) \quad v^+(x_0) = \sup_{v \in L^2(0, T; \mathbf{R}^k)} C_{x_0}(\hat{u}, v) = C_{x_0}(\hat{u}, \hat{v}) = C_{x_0}(\phi^*, \psi^*).$$

- (ii) (case (f)).  $v^-(x_0) = v^+(x_0) = -\infty$  if and only if  $C_{x_0}(u, v)$  is concave in  $v$  and not convex in  $u$  and  $u \mapsto \sup_v C_{x_0}(u, v)$  is not convex. In that case

$$\sup_{\psi \in \Psi} \inf_{\phi \in \Phi} C_{x_0}(\phi, \psi) = \inf_{\phi \in \Phi} \sup_{\psi \in \Psi} C_{x_0}(\phi, \psi) = -\infty.$$

*Remark 4.8.* Part (ii) of Theorems 4.4 and 4.5 justifies the terminology “degenerate” for cases (e) and (f), since what would be the candidate for a closed loop saddle point identity in  $\Phi \times \Psi$  is, respectively, equal to  $+\infty$  and  $-\infty$ . This is to be compared with [2, Definition 2.3 and Theorem 3.1]. If the problem is normalizable, then the conclusions of Theorems 4.4 and 4.5 hold for all  $x_0 \in \text{Im } X(0)$  (cf. Theorem 4.1).

Finally, by complementarity, we have the last case (d) of Theorem 3.2.

**THEOREM 4.6** (case (d)). Assume that  $C_{x_0}$  has a closed loop saddle point  $(\phi^*, \psi^*) \in S$ . Then  $v^-(x_0) = -\infty$  and  $v^+(x_0) = +\infty$  if and only if  $C_{x_0}(u, v)$  is not convex in  $u$  and not concave in  $v$ .

*Remark 4.9.* Case (d) can definitely occur, as all the other cases. Again, in the work of Bernhard [2], the utility function was convex in  $u$  since  $F \geq 0$  and  $Q(t) \geq 0$ . Hence, only cases (a) and (b) could occur, and case (e) is still a degenerate one in the sense of his definition.

**4.9. An example of a problem that is not normalizable.** We conclude with an example where  $\det X(t) = 0$  on an interval of nonzero length. Yet there are solutions to the matrix Riccati differential equation, and the open loop lower value of the game is finite for all initial conditions. The associated strategies can be obtained by feedback. So, the condition  $\det X(t) \neq 0$  almost everywhere in  $[0, T]$  might not be the most general one, and Definitions 4.3 and 4.5 might be further relaxed or generalized.

*Example 4.3.* Consider the dynamics and utility function in the time interval  $[0, 3]$ ,

$$(4.56) \quad x'(t) = b_1(t)u(t) + b_2(t)v(t), \text{ a.e. in } [0, 3], \quad x(0) = x_0,$$

$$(4.57) \quad C_{x_0}(u, v) = \frac{1}{2}|x(3)|^2 + \int_0^3 |u(t)|^2 - |v(t)|^2 dt,$$

where

$$(4.58) \quad b_1(t) = \begin{cases} 2-t, & 0 \leq t < 1, \\ 0, & 1 \leq t < 2, \\ 3-t, & 2 \leq t \leq 3, \end{cases} \quad b_2(t) = \begin{cases} t, & 0 \leq t < 1, \\ 0, & 1 \leq t < 2, \\ t-1, & 2 \leq t \leq 3. \end{cases}$$

Here  $A = 0$ ,  $B_1(t) = b_1(t)$ ,  $B_2(t) = b_2(t)$ ,  $F = 1/2$ ,  $Q = 0$ , and

$$(4.59) \quad R(t) = b_1(t)^2 - b_2(t)^2 = \begin{cases} 4(1-t), & 0 \leq t < 1, \\ 0, & 1 \leq t < 2, \\ 4(2-t), & 2 \leq t \leq 3. \end{cases}$$

We show that  $v^-(x_0) = (x_0)^2/2$  and  $v^+(x_0) = +\infty$ . For the open loop lower value of the game, the minimization with respect to  $u$  has a unique solution for all  $(x_0, v)$  since the utility function  $u \mapsto C_{x_0}(u, v)$  is convex and bounded below by  $-\|v\|_{L^2}^2$ . The minimizer is completely characterized by the coupled system

$$\begin{cases} x' = -b_1^2 p + b_2 v \text{ a.e. in } [0, 3], & x(0) = x_0, & \hat{u} = -b_1 p, \\ p' = 0 \text{ a.e. in } [0, 3], & p(3) = \frac{1}{2}x(3). \end{cases}$$

From this,

$$J_{x_0}^-(v) \stackrel{\text{def}}{=} \inf_{u \in L^2(0,2;\mathbf{R})} C_{x_0}(u, v) = C_{x_0}(\hat{u}, v) = \frac{1}{4} \left[ x_0 + \int_0^3 b_2 v ds \right]^2 - \int_0^3 |v|^2 dt.$$

It is readily seen that  $J_{x_0}^-$  is concave in  $v$  and that the supremum with respect to  $v$  of  $J_{x_0}^-(v)$  exists. Indeed, from the first order condition, for all  $v$ ,

$$\frac{1}{2} dJ_{x_0}^-(\hat{v}; v) = \frac{1}{4} \left[ x_0 + \int_0^3 b_2 \hat{v} ds \right] \int_0^3 b_2 v ds - \int_0^3 \hat{v} v(t) dt = 0,$$

there is a unique stationary point  $\hat{v}(t) = b_2(t)x_0/2$ , the Hessian is negative,

$$\frac{1}{2} d^2 J_{x_0}^-(\hat{v}; v; v) = \frac{1}{4} \left[ \int_0^3 b_2 v ds \right]^2 - \int_0^3 |v|^2 dt \leq -\frac{1}{3} \int_0^3 |v|^2 dt,$$

and the open loop lower value of the game is  $v^-(x_0) = J_{x_0}^-(\hat{v}) = (x_0)^2/2$ . Moreover,

$$(4.60) \quad \hat{u}(t) = \begin{cases} -(2-t)\frac{1}{2}x_0, & 0 \leq t < 1, \\ 0, & 1 \leq t < 2, \\ -(3-t)\frac{1}{2}x_0, & 2 \leq t \leq 3, \end{cases} \quad \hat{v}(t) = \begin{cases} t\frac{1}{2}x_0, & 0 \leq t < 1, \\ 0, & 1 \leq t < 2, \\ (t-1)\frac{1}{2}x_0, & 2 \leq t \leq 3. \end{cases}$$



However, the open loop upper value of the game is  $v^+(x_0) = +\infty$  for all  $x_0 \in \mathbf{R}$ . Indeed, pick the sequence of controls  $\{v_n\}$ ,  $n \geq 1$ ,  $v_n(t) = 0$  in  $[0, 2]$  and  $v_n(t) = n$  in  $[2, 3]$ . The corresponding sequence of states at time  $t = 3$  is

$$x_n(3) = x_0 + \int_0^3 b_1 u \, dt + n \int_2^3 (t-1) \, dt = \left[ x_0 + \int_0^3 b_1 u \, dt \right] + \frac{3}{2}n.$$

Denote by  $X$  the square bracket that does not depend on  $n$ . Then

$$\begin{aligned} C_{x_0}(u, v_n) &= \frac{1}{2} \left| X + \frac{3}{2}n \right|^2 + \int_0^3 |u|^2 \, dt - \int_2^3 n^2 \, dt \\ &= \frac{1}{8}n^2 + \frac{3}{2}nX + \frac{X^2}{2} + \int_0^3 |u|^2 \, dt \rightarrow +\infty \text{ as } n \rightarrow +\infty. \end{aligned}$$

Thus for all  $x_0 \in \mathbf{R}$  and  $u \in L^2(0, T; \mathbf{R})$ ,

$$\sup_{v \in L^2(0, T; \mathbf{R})} C_{x_0}(u, v) = +\infty \Rightarrow v^+(x_0) = +\infty.$$

Therefore,  $C_{x_0}(u, v)$  has no open loop saddle point. For all  $x_0$ , the coupled system

$$(4.61) \quad \hat{x}' = -R\hat{p}, \quad \hat{x}(0) = x_0 \quad \text{and} \quad \hat{p} = 0, \quad \hat{p} = \frac{1}{2}\hat{x}(3), \quad \hat{u} = -b_1\hat{p}, \quad \text{and} \quad \hat{v} = b_2\hat{p}$$

has a unique solution in  $H^1(0, 3)$ . The unique solution of system (4.1),

$$(4.62) \quad \begin{cases} X' = AX - R\Lambda, & X(T) = I, \\ \Lambda' + A^\top \Lambda + QX = 0, & \Lambda(T) = F, \end{cases}$$

is given by

$$(4.63) \quad \begin{aligned} X(t) &= \begin{cases} (t-1)^2, & 0 \leq t < 1 \\ 0, & 1 \leq t < 2 \\ (t-2)^2, & 2 \leq t \leq 3 \end{cases}, & P_c(t) &= \begin{cases} \frac{1}{2(t-1)^2}, & 0 \leq t < 1 \\ c \text{ (arbitrary)}, & 1 \leq t < 2 \\ \frac{1}{2(t-2)^2}, & 2 \leq t \leq 3 \end{cases}. \\ \Lambda(t) &= 1/2, \end{aligned}$$

The problem is not normalizable since  $X(t) = 0$  in  $[1, 2]$ . Yet, the associated optimal strategies are feedback strategies of the usual forms,  $\hat{u} = -b_1 P_c x$  and  $\hat{v} = b_2 P_c x$ , and the linear closed loop strategies are

$$\phi^*(t, x) = U_c(t)x = -b_1(t)P_c(t)x \quad \text{and} \quad \psi^*(t, x) = V_c(t)x = b_2(t)P_c(t)x.$$

The function  $X$  is also a solution of (4.19) in Definition 4.3,

$$X' = (b_1 U_c + b_2 V_c)X, \quad X(3) = I.$$

If we adopt the convention that a function  $u \in L^2(0, 3)$  such that  $|X^{-1}|u \in L^2(0, 3)$  implies that  $u = 0$  on  $Z = \{t \in [0, 3] : X(t) = 0\} = [1, 2]$  and adopt the same for the function  $v$ , it can be shown that  $dc_{x_0}(0, 0, : u, v) = 0$  for all  $u$  and  $v$ . However, we have not been able to prove the convexity-concavity of  $c_{x_0}(u, v)$  to conclude that the problem has a closed loop saddle point in the sense of Definitions 4.3 and 4.5.

Another issue is the meaning of a solution to the Riccati equation  $P' - RP^2 = 0$ , with final value  $P(3) = 1/2$  and a discontinuous function  $R$ . What is the effect of a discontinuity in  $R(t)$ ? For instance the following solutions are continuous in  $t = 1$ :

$$P(t) = \begin{cases} \bar{c}/(1 + 2\bar{c}(t-1)^2), & 0 \leq t < 1 \\ \bar{c}, & 1 \leq t < 2 \\ 1/(2(t-2)^2), & 2 \leq t \leq 3 \end{cases}$$

for some arbitrary constant  $\bar{c} \in \mathbf{R}$ . The singularity at  $t_2 = 2$  is independent of  $\bar{c}$ . For  $\bar{c} > -1/2$  it is the only singularity. For  $\bar{c} \leq -1/2$ , there is a second singularity at  $t_1 = 1 - \sqrt{1/(-2\bar{c})}$  in the interval  $[0, 1)$ . We also have the solutions  $P_c$  of (4.63) with a singularity in  $t_1 = 1$ . Therefore the solution of the Riccati equation is not unique.

#### REFERENCES

- [1] I. BERKOVITZ, *Lectures on differential games*, in *Differential Games and Related Topics*, H. W. Kuhn and G. P. Szego, eds., North-Holland, Amsterdam, Holland, 1971, pp. 3–45.
- [2] P. BERNHARD, *Linear-quadratic, two-person, zero-sum differential games: Necessary and sufficient conditions*, *J. Optim. Theory Appl.*, 27 (1979), pp. 51–69.
- [3] P. BERNHARD, *Technical comment to: "Linear-quadratic two-person zero-sum differential games: Necessary and sufficient conditions"* [*J. Optim. Theory Appl.*, 27 (1979), no. 1, pp. 51–69], *J. Optim. Theory Appl.*, 31 (1980), pp. 283–284.
- [4] I. CHAMPAGNE, *Méthodes de factorisation des équations aux dérivées partielles*, Thèse de doctorat, École Polytechnique, Paris, France; also available as INRIA Report TU-1125, Le Chesnay, France, 2004.
- [5] M. C. DELFOUR, *Linear quadratic differential games: Saddle point and Riccati differential equation*, *SIAM J. Control Optim.*, 46 (2007), pp. 750–774.
- [6] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1944.
- [7] W. T. REID, *Riccati Differential Equations*, Academic Press, New York, 1972.
- [8] M. SORINE AND P. WINTERNITZ, *Superposition laws for solutions of differential matrix Riccati equations arising in control theory*, *IEEE Trans. Automat. Control*, 30 (1985), pp. 266–272.
- [9] P. ZHANG, *Some results on two-person zero-sum linear quadratic differential games*, *SIAM J. Control Optim.*, 43 (2005), pp. 2157–2165.

# A UNIQUENESS RESULT FOR $p$ -MONOTONE VISCOSITY SOLUTIONS OF HAMILTON–JACOBI EQUATIONS IN BOUNDED DOMAINS\*

MARTIN V. DAY<sup>†</sup>

**Abstract.** We consider a class of Hamilton–Jacobi equations  $H(x, Du(x)) = 0$  with no  $u$ -dependence and with continuity properties consistent with recent applications in queueing theory. Continuous viscosity solutions are considered in a compact polyhedral domain, with oblique derivative (Neumann-type) boundary conditions. Comparison and uniqueness results are presented, which use monotonicity of  $H(x, p)$  in the  $p$  variable for values of  $p$  in the appropriate sub- and superdifferential sets of the solution  $u(x)$ . Several examples illustrate the results.

**Key words.** viscosity solution, Hamilton–Jacobi equation, uniqueness

**AMS subject classifications.** 49L25, 35D99, 35F30

**DOI.** 10.1137/070700401

**1. Introduction.** The theory of viscosity solutions to first order partial differential equations provides a satisfying approach to Hamilton–Jacobi equations for many types of optimal control problems and differential games. Bardi and Capuzzo-Dolcetta [4] give an extensive introduction to the basic theory and its application to a variety of optimization problems. At the heart of the theory are the fundamental comparison and uniqueness results, which identify the optimal value function as the unique viscosity solution of the appropriate Hamilton–Jacobi equation. Those comparison and uniqueness results generally depend on some monotonicity property of the Hamiltonian  $H$ . For instance, in the case of discounted infinite horizon problems, the Hamilton–Jacobi equation includes a term  $\lambda u$  ( $\lambda > 0$  being the discount rate). This provides monotonicity in  $u$  which is the key to the proof of the typical comparison result, such as [4, Theorem II.3.1].

In this paper we consider problems of the form

$$H(x, Du(x)) = 0,$$

in which the Hamiltonian  $H(x, p)$  has no  $u$ -dependence. It is well known that without some additional property, solutions may be nonunique. (See Example 6 in section 5, for instance.) Ishii [18] provides an approach which assumes convexity of  $p \mapsto H(x, p)$  and the existence of a special smooth subsolution  $\varphi$ . (See also [4, section II.5.3].) The idea is to perturb a given subsolution by a (small) convex combination with  $\varphi$  to obtain a “strict” subsolution. A basic comparison result (very like our Lemma 2) then implies the desired inequality. An elementary example is the eikonal equation

$$H(x, p) = |p| - h(x),$$

where  $h$  is continuous and strictly positive on the spatial domain  $\Omega$ . This category of problems can also be treated using the transformation of Kružkov. This can be applied generally when there is a strictly positive lower bound for the running cost  $L$

---

\*Received by the editors August 17, 2007; accepted for publication (in revised form) August 27, 2008; published electronically January 7, 2009.

<http://www.siam.org/journals/sicon/47-6/70040.html>

<sup>†</sup>Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123 (day@math.vt.edu).

of (19) below. See Bardi and Soravia [5] and the references in [18]. Our  $p$ -monotone approach is also applicable to such problems; see Example 2 below.

Another approach is that of Camilli and Siconolfi [7]. They are interested in equations of the form

$$H(x, p) - f(x) = 0$$

and seek to identify maximal subsolutions. (In some control problems this is the standard characterization of the desired viscosity solution; see Soravia [21].) They obtain a definitive characterization of maximal subsolutions in terms of a special singular solution property. Their approach is rather technical, using convexity of the sets  $\{p : H(x, p) - f(x) \leq 0\}$  and a special topology in  $\Omega$  associated with them. Among their few simple hypotheses on the Hamiltonian is the assumption that  $t \mapsto H(x, tp)$  is strictly increasing in  $t \in [0, 1]$  for all  $p$ . We note that this is essentially the  $p$ -monotone property that we exploit below. We would comment that our results also provide a simple sufficient condition for a viscosity solution to be the maximal subsolution, namely that it be a  $p$ -monotone supersolution.

We are motivated by a growing body of work using control problems and differential games for asymptotic analysis of queueing networks. These problems often involve oblique-derivative boundary conditions on some part of  $\partial\Omega$ . (Although only Dirichlet conditions were considered in [18] and [7], presumably generalizations are possible.) These examples typically do not have the convexity needed for either the approach of [18] or [7]; see Examples 4 and 5 in section 5. However, the literature does contain some uniqueness results for certain problems of this type. The germ of our  $p$ -monotone argument can be found in the proof of Theorem 5 of Atar, Dupuis, and Schwartz [2] (see their equation (37)). Although it is not a viscosity solution result, the structured verification theorem of Day [12] uses a “positive storage condition” which is related to  $p$ -monotonicity (as we will see in Example 5). The essential feature underlying these results is monotonicity of  $t \mapsto H(x, tp)$ , not necessarily for all  $p$  but just for those  $p = \zeta \in D^\pm u(x)$  that are not accounted for by the boundary conditions. Our intent here is to develop comparison and uniqueness results based on this property for problems with oblique-derivative boundary conditions, such as are typical in queueing applications. This class of problems also motivates our regularity hypotheses on  $H$ .

There are a few other comparison results in the literature which employ properties of the  $p$ -dependence of  $H$ . For instance, the development in Crandall, Ishii, and Lions [8] assumes that a special test function  $\mu(x)$  exists for which  $\lambda \mapsto H(x, p + \lambda D\mu(x)) - H(x, p)$  satisfies a certain lower bound; see their (H2). We note that such a hypothesis is entirely a property of the Hamiltonian and depends on the existence of  $\mu(x)$ . In general our notion of a  $p$ -monotone solution depends on the specific solution  $u(x)$ , not solely on  $H$ .

In section 2 we pose the specific type of boundary value problem we will address, using oblique-derivative conditions on the boundary of a compact polyhedral domain. Section 3 presents a basic comparison result (Lemma 2) for sub- and supersolutions to a pair of “strictly separated” equations. (That strict separation generally implies a comparison result is well known; see Crandall, Ishii, and Lions [9].) The  $p$ -monotone results are then developed in section 4. Our main result (Theorem 4) implies that when a  $p$ -monotone solution exists it is the unique viscosity solution—the “complete solution” in the terminology of [4]. We conclude by looking at several examples in section 5.

**2. Preliminaries and hypotheses.** We consider a domain  $\Omega$  which is assumed to be a compact convex polyhedron in  $\mathbb{R}^n$ , defined by a finite collection of  $m$  linear constraints,

$$(1) \quad \Omega = \{x \in \mathbb{R}^n : n_i \cdot x \geq c_i \text{ for each } i = 1, \dots, m\}.$$

The  $n_i$  are unit vectors (inward normals) and the  $c_i$  are constants. For  $x \in \partial\Omega$  (the boundary of  $\Omega$ ) we define the set of active constraints as

$$I(x) = \{i : n_i \cdot x = c_i\}$$

and take  $I(x) = \emptyset$  for  $x \in \Omega^\circ$  (the interior of  $\Omega$ ). We consider a closed subset  $\mathcal{T} \subseteq \Omega$  on which Dirichlet data will be prescribed. This could be part of the boundary, but that it not necessary. Values for  $u$  are prescribed on  $\mathcal{T}$  by a continuous function  $g : \mathcal{T} \rightarrow \mathbb{R}$ ,

$$(2) \quad u(x) = g(x), \quad x \in \mathcal{T}.$$

It will be convenient to use the notation

$$\Omega_{\delta, \mathcal{T}} = \{x \in \Omega : \text{dist}(x, \mathcal{T}) > \delta\}$$

to refer to the part of  $\Omega$  at least  $\delta > 0$  away from  $\mathcal{T}$ . (We allow  $\mathcal{T} = \emptyset$ , in which case  $\Omega_{\delta, \mathcal{T}} = \Omega$ .) On the rest of the boundary,  $\partial\Omega \setminus \mathcal{T}$ , we want to require oblique-derivative boundary conditions using a collection of vectors  $d_i$ ,  $i = 1, \dots, m$ ,

$$(3) \quad -d_i \cdot Du(x) = 0, \quad i \in I(x).$$

In  $\Omega \setminus \mathcal{T}$  itself we consider a Hamilton–Jacobi equation,

$$(4) \quad H(x, Du(x)) = 0.$$

If  $\mathcal{T} = \partial\Omega$ , we have a standard Dirichlet problem. If  $\mathcal{T} = \emptyset$  we have a typical Neumann-type problem. In general the problem is a mixture of these two types.

**2.1. Continuity hypothesis on the Hamiltonian.** Appropriate continuity hypotheses for the Hamiltonian  $H$  are important. The examples we have in mind use a Hamiltonian of the form (19) below, with  $f = f(a, b)$  independent of state and running cost  $L = h(x) + \ell(a, b)$  with separate state and player components. This leads to a Hamiltonian of separated form,  $H(x, p) = H_0(p) - h(x)$ . But all we really need are continuity hypotheses consistent with that. We assume there exist  $m : [0, \infty) \rightarrow [0, \infty)$  with  $m(0) = 0$  and continuous at 0, and  $M : [0, \infty)^2 \rightarrow [0, \infty)$  with  $M(0, R) = 0$  and  $M(\cdot, R)$  continuous at 0 for each  $R < \infty$ , such that for all  $x, y \in \Omega$  and  $p, q \in \mathbb{R}^d$  with  $|p|, |q| \leq R$ , we have

$$(5) \quad |H(x, p) - H(y, q)| \leq m(|x - y|) + M(|p - q|, R).$$

**2.2. Technical hypotheses on  $\Omega$  and  $d_i$ .** The oblique-derivative boundary conditions (3) are closely associated with the Skorokhod problem for  $\Omega$ ; see Dupuis and Ishii [14]. Control problems for systems including a Skorokhod problem in their dynamics are common in queueing theory and lead to Hamilton–Jacobi equations with boundary conditions (3); see Lions [19], Dupuis and Ishii [15], and Day [11]. Although the Skorokhod problem does not appear in our results below, hypotheses from [14] regarding  $\Omega$  and the  $d_i$  of the boundary conditions are important ingredients for the proof of Lemma 2. For that purpose we assume the following.

• *B-hypothesis* [14, Assumption 2.1]. There exists a compact, convex  $B \subseteq \mathbb{R}^n$  with  $0 \in B^\circ$  and the following property: If  $z \in \partial B$  and  $|z \cdot n_i| < 1$ , then  $\nu \cdot d_i = 0$  for all unit outward normals to  $B$  at  $z$ . ( $\nu$  is an outward normal to  $B$  at  $z$  if  $\nu \cdot (z - x) \geq 0$  for all  $x \in B$ .) For further discussion of this hypothesis and an illustrative figure, see Dupuis and Ramanan [17].

• *Coercivity hypothesis*. For each  $x \in \partial\Omega$ , and any  $a_i \in \mathbb{R}$ ,

$$(6) \quad \left( \sum_{i \in I(x)} a_i d_i \right) \cdot \left( \sum_{i \in I(x)} a_i n_i \right) \geq 0,$$

with equality only if  $a_i = 0$  for all  $i \in I(x)$ . It is shown in Day [10] that this, together with the *B-hypothesis*, implies [14, Assumption 3.1] concerning the existence of a discrete projection map. Moreover, it implies that, for each  $x \in \partial\Omega$ , the  $d_i$ ,  $i \in I(x)$ , are linearly independent, which is needed for Lemma 1 below. We might have assumed [14, Assumption 3.1] along with this linear independence property, but (6) is a convenient sufficient condition for both and is easy to verify in examples, since it reduces to checking positive definiteness of a small number of matrices.

These hypotheses provide the following technical result, which will be needed for the proof of Lemma 2.

LEMMA 1. *Assume the B-hypothesis and the coercivity hypothesis.*

(a) *There exists a  $C^1$  function  $\mu : \Omega \rightarrow [0, 1]$  with the property that  $d_i \cdot D\mu(x) < 0$  whenever  $x \in \partial\Omega$  and  $i \in I(x)$ .*

(b) *There exists a  $C^1$  function  $\xi : \mathbb{R}^n \rightarrow [0, \infty)$  with the properties that*

(i)  *$\xi^{1/2}$  is a norm on  $\mathbb{R}^d$ , and*

(ii) *for any  $x \in \mathbb{R}^n$  and  $i = 1, \dots, m$ ,  $x \cdot n_i \geq 0$  [ $\leq 0$ ] implies  $d_i \cdot D\xi(x) \geq 0$  [ $\leq 0$ ].*

*Proof.* Part (a) is Lemma 3.2 of Dupuis and Ishii [15]. Their hypothesis (B.6) follows from the independence of  $d_i$ ,  $i \in I(x)$ , pointed out above. The other hypotheses are simple to check in our setting.

Part (b) follows from arguments given in Atar and Dupuis [1], which we outline. (See their remark on page 1109.) First, it is shown that the property of  $B$  is equivalent to an extended property, namely that if  $z \in \partial B$  and  $\nu$  is an outward normal to  $B$  at  $z$ , then

$$z \cdot n_i \geq -1 [\leq 1] \text{ implies } d_i \cdot \nu \geq 0 [\leq 0].$$

(Although [1] only considers  $\Omega = \mathbb{R}_+^n$ , the extension argument based on Dupuis and Ramanan [17] applies in general.) Next, given that the set  $B$  exists, it is argued that  $B$  can be assumed symmetric with a smooth boundary, in the sense that the unit outward normal  $\nu(x)$  is uniquely determined and continuous as a function of  $x \in \partial B$ . Such a  $B$  determines a (smooth) norm on  $\mathbb{R}^n$ , defined by

$$\|x\|_B = \inf\{r > 0 : x \in rB\}.$$

$B$  is the closed unit ball with respect to  $\|\cdot\|_B$ . We define  $\xi(x) = \|x\|_B^2$ . It follows that  $\xi$  is  $C^1$ , and for a given  $x$ ,

$$D\xi(x) = b\|x\|_B \nu,$$

where  $b = b(x) > 0$  is a scalar function and  $\nu$  the unit outward normal to  $B$  at  $z = x/\|x\|_B \in \partial B$ . Therefore if  $x \cdot n_i \geq 0$ , then  $-1 < 0 \leq z \cdot n_i$ , so that the extended

property of  $B$  above implies  $d_i \cdot \nu \geq 0$ , which in turn implies  $d_i \cdot D\xi(x) \geq 0$ . The other case is proven analogously, or by appeal to symmetry.  $\square$

As a consequence of (a), observe that there exists a constant  $\mu_0 > 0$  such that

$$(7) \quad \mu_0 < -d_i \cdot D\mu(x) \text{ for all } x \in \partial\Omega, i \in I(x).$$

**2.3. Viscosity solutions.** In the proof of Lemma 2 we will use the generalization of (3) to

$$(8) \quad C - d_i \cdot Du(x) = 0, i \in I(x),$$

where  $C$  is a constant. We want to state carefully what it means to be a viscosity sub- or supersolution of (4) with boundary conditions (8) on  $\Omega \setminus \mathcal{T}$ . Note that the definitions will not refer to (2) on  $\mathcal{T}$ ; we prefer to express that separately by referring to “subsolutions with  $u(x) \leq g(x)$  on  $\mathcal{T}$ ” as needed.

We will consider only continuous functions  $u : \Omega \rightarrow \mathbb{R}$  as possible solutions. For  $x \in \Omega$  the superdifferential set  $D^+u(x)$  consists of those  $\zeta \in \mathbb{R}^n$  which occur as the value  $\zeta = D\phi(x)$  for some  $C^1$  function  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  with the property that  $u(x) - \phi(x) \geq u(y) - \phi(y)$  for all  $y \in \Omega$  sufficiently close to  $x$ . For the correct viscosity-sense understanding of (8) it is important to note that  $x$  is a local maximum of  $u - \phi$  *only relative to*  $\Omega$ . For  $x \in \partial\Omega$  this means that even if  $u$  is smooth,  $D^+u(x)$  can contain many  $\zeta$  other than  $Du(x)$  itself. (See Lemma 7 in section 5.) Similarly,  $D^-v(x)$  consists of  $\zeta$  arising as  $\zeta = D\phi(x)$  for some  $C^1$  function  $\phi(x)$  such that  $v - \phi$  has a local minimum at  $x$  relative to  $\Omega$ . The function  $u(x) \in C(\Omega)$  is called a *subsolution* of

$$(9) \quad H(x, Du(x)) = 0 \text{ on } \Omega \setminus \mathcal{T} \text{ with } C - d_i \cdot Du(x) = 0 \text{ on } \partial\Omega \setminus \mathcal{T}$$

provided the following hold for all  $\zeta \in D^+u(x)$ :

- (i) if  $x \in \Omega^\circ \setminus \mathcal{T}$ , then  $H(x, \zeta) \leq 0$ ;
- (ii) if  $x \in \partial\Omega \setminus \mathcal{T}$ , then either  $H(x, \zeta) \leq 0$  or  $C - d_i \cdot \zeta \leq 0$  for some  $i \in I(x)$ .

In other words, at boundary points only one of the inequalities  $H(x, \zeta) \leq 0$ ,  $C - d_i \cdot \zeta \leq 0$  ( $i \in I(x)$ ) needs to hold. This is the, now standard, viscosity formulation of first order equations with “Neumann-type” boundary conditions (see Barles and Lions [6]), generalized to consider different boundary conditions  $C - d_i \cdot Du(x) = 0$  on the different planar faces of  $\partial\Omega$ . We can express this subsolution definition succinctly by writing

$$(10) \quad H(x, \zeta) \wedge \min_{i \in I(x)} (C - d_i \cdot \zeta) \leq 0 \text{ for all } x \in \Omega \setminus \mathcal{T} \text{ and } \zeta \in D^+u(x)$$

and using the convention that  $\min_{i \in I(x)} = +\infty$  if  $I(x) = \emptyset$ . The definition of a *supersolution* is obtained by reversing all the inequalities in (i) and (ii) and considering  $\zeta \in D^-u(x)$  instead. We would replace (10) by  $H(x, \zeta) \vee \max_{i \in I(x)} (C - d_i \cdot \zeta) \geq 0$ .

**3. A basic comparison result for strictly separated equations.** The task of this section is to establish a basic comparison result for oblique-derivative boundary conditions (3) analogous to that of Ishii [18, Lemma 1]. The comparison argument of Atar, Dupuis, and Schwartz [2] is close to ours and is the source of our approach to handling the boundary conditions. The use of a norm such as  $\xi$  in the function  $\Phi_\epsilon$  of the proof below originated in Dupuis, Ishii, and Soner [16].

LEMMA 2. Assume that  $g : \mathcal{T} \rightarrow \mathbb{R}$  is continuous and that  $u, v \in C(\Omega)$  with  $u \leq g \leq v$  on  $\mathcal{T}$  are such that

(a)  $u$  is a subsolution of  $H(x, Du(x)) + \eta_+(x) = 0$  on  $\Omega \setminus \mathcal{T}$ , with  $-d_i \cdot Du(x) = 0$  on  $\partial\Omega \setminus \mathcal{T}$ ; and

(b)  $v$  is a supersolution of  $H(x, Dv(x)) - \eta_-(x) = 0$  on  $\Omega \setminus \mathcal{T}$ , with  $-d_i \cdot Dv(x) = 0$  on  $\partial\Omega \setminus \mathcal{T}$ ,

where  $\eta_{\pm} : \Omega \rightarrow \mathbb{R}$  have the property that for each  $\delta > 0$ ,

$$(11) \quad \inf_{x \in \Omega_{\delta, \mathcal{T}}} \eta_+(x) + \inf_{x \in \Omega_{\delta, \mathcal{T}}} \eta_-(x) > 0.$$

Then  $u(x) \leq v(x)$  for all  $x \in \Omega$ .

We will say that the  $u$  and  $v$  of this lemma are viscosity sub- and supersolutions to a *strictly separated* pair of equations. Note that because of (11) this notion of strict separation depends on the choice of  $\mathcal{T}$ . Also observe that we have made no regularity assumption on the  $\eta_{\pm}$ . The inequality (11) is all the proof needs. An alternate hypothesis would be to assume that  $\inf_{\Omega_{\delta, \mathcal{T}}} [\eta_+(x) + \eta_-(x)] > 0$  along with continuity of (one of) the  $\eta_{\pm}$ .

*Proof.* Let  $0 < c_{\epsilon} < 1$  be a family of constants with  $c_{\epsilon} \rightarrow 0$  as  $\epsilon \downarrow 0$ . Near the end of the proof we will be more specific about how  $c_{\epsilon}$  should be chosen, but that detail is not needed yet. Given  $\epsilon > 0$ , define

$$u_{\epsilon}(x) = u(x) - c_{\epsilon}\mu(x), \quad v_{\epsilon}(x) = v(x) + c_{\epsilon}\mu(x),$$

where  $\mu(x)$  is as in Lemma 1 above. It follows that  $\zeta_{\epsilon} \in D^+u_{\epsilon}(x)$  iff  $\zeta = \zeta_{\epsilon} + c_{\epsilon}D\mu(x) \in D^+u(x)$ . Notice that

$$-d_i \cdot \zeta = -d_i \cdot (\zeta_{\epsilon} + c_{\epsilon}D\mu(x)) \geq -d_i \cdot \zeta_{\epsilon} + c_{\epsilon}\mu_0,$$

where  $\mu_0$  is as in (7). Therefore, the subsolution hypothesis of (a) implies that for all  $\zeta \in D^+u(x)$ ,

$$[H(x, \zeta_{\epsilon} + c_{\epsilon}D\mu(x)) + \eta_+(x)] \wedge \min_{i \in I(x)} (c_{\epsilon}\mu_0 - d_i \cdot \zeta_{\epsilon}) \leq 0.$$

In other words,  $u_{\epsilon}$  is a subsolution of

$$(12) \quad H(x, Du_{\epsilon}(x) + c_{\epsilon}D\mu(x)) + \eta_+(x) = 0 \text{ on } \Omega \setminus \mathcal{T} \text{ with } c_{\epsilon}\mu_0 - d_i \cdot Du_{\epsilon}(x) = 0 \text{ on } \partial\Omega \setminus \mathcal{T}.$$

Similarly,  $v_{\epsilon}$  is a supersolution of

$$(13) \quad H(x, Dv_{\epsilon}(x) - c_{\epsilon}D\mu(x)) - \eta_-(x) = 0 \text{ on } \Omega \setminus \mathcal{T} \text{ with } -c_{\epsilon}\mu_0 - d_i \cdot Dv_{\epsilon}(x) = 0 \text{ on } \partial\Omega \setminus \mathcal{T}.$$

Now suppose that  $\sup_{\Omega}(u(x) - v(x)) > 0$ . Then because  $\mu(x)$  is bounded and  $c_{\epsilon} \rightarrow 0$ , there is a positive constant  $\rho$  so that for all sufficiently small  $\epsilon > 0$ ,

$$(14) \quad 0 < \rho < \sup_{\Omega}[u_{\epsilon}(x) - v_{\epsilon}(x)].$$

We now give a version of the usual argument leading to a contradiction. Define

$$\Phi_{\epsilon}(x, y) = u_{\epsilon}(x) - v_{\epsilon}(y) - \epsilon^{-1}\xi(x - y),$$

where  $\xi(\cdot)$  is as in Lemma 1, and let  $(x_{\epsilon}, y_{\epsilon}) \in \Omega \times \Omega$  be a maximizing pair for  $\Phi_{\epsilon}$ . By comparison to  $x = y$ , we have

$$(15) \quad \Phi_{\epsilon}(x_{\epsilon}, y_{\epsilon}) \geq \rho.$$



From  $\Phi_\epsilon(x_\epsilon, x_\epsilon) \leq \Phi_\epsilon(x_\epsilon, y_\epsilon)$  it follows that

$$(16) \quad \epsilon^{-1}\xi(x_\epsilon - y_\epsilon) \leq v_\epsilon(x_\epsilon) - v_\epsilon(y_\epsilon).$$

Since  $v$  and  $\mu$  are bounded, and  $0 < c_\epsilon < 1$ , it follows that  $v_\epsilon$  is bounded (independent of  $\epsilon$ ). We deduce that  $\xi(x_\epsilon - y_\epsilon) = O(\epsilon)$ . Since all norms on  $\mathbb{R}^n$  are equivalent, we have

$$(17) \quad \|x_\epsilon - y_\epsilon\| = O(\epsilon^{1/2}).$$

Next, we claim that none of the limit points of  $x_\epsilon$  (as  $\epsilon \downarrow 0$ ) can be in  $\mathcal{T}$ . Indeed, if (along a sequence of  $\epsilon \downarrow 0$ ) we had  $x_\epsilon \rightarrow z \in \mathcal{T}$ , then by (17)  $y_\epsilon \rightarrow z$  as well. It follows that

$$\lim_\epsilon [u_\epsilon(x_\epsilon) - v_\epsilon(y_\epsilon)] \leq g(z) - 0\mu(z) - [g(z) + 0\mu(z)] = 0.$$

Since  $v$  and  $\mu$  are continuous,  $v_\epsilon$  is equicontinuous with respect to  $\epsilon$ . This, together with (16), implies that  $\epsilon^{-1}\xi(x_\epsilon - y_\epsilon) \rightarrow 0$ . Therefore  $\Phi_\epsilon(x_\epsilon, y_\epsilon) \rightarrow 0$ , contrary to (15), and this proves our claim. The claim means that there exists  $\delta > 0$  so that  $x_\epsilon, y_\epsilon \in \Omega_{\delta, \mathcal{T}}$  for all sufficiently small  $\epsilon$ . By hypothesis (11), there exists  $\eta_0 > 0$  so that

$$\eta_0 \leq \eta_+(x_\epsilon) + \eta_-(y_\epsilon),$$

for all  $\epsilon > 0$  sufficiently small.

Now  $u_\epsilon(x) - [v_\epsilon(y_\epsilon) + \epsilon^{-1}\xi(x - y_\epsilon)]$  is maximized at  $x = x_\epsilon$ . Therefore  $\zeta_\epsilon \doteq \epsilon^{-1}D\xi(x_\epsilon - y_\epsilon) \in D^+u_\epsilon(x_\epsilon)$ . Since  $D\xi$  is continuous and  $\Omega$  is compact, it follows that

$$\zeta_\epsilon = O(\epsilon^{-1}).$$

If it were the case that  $x_\epsilon \in \partial\Omega$ , then by definition of  $\Omega$  we would have  $n_i \cdot (x_\epsilon - y_\epsilon) \leq 0$  for all  $i \in I(x_\epsilon)$ . By property (ii) of  $\xi$  in Lemma 1, it follows that  $d_i \cdot \zeta_\epsilon \leq 0$  for all  $i \in I(x_\epsilon)$ . Therefore,

$$c_\epsilon\mu_0 - d_i \cdot \zeta_\epsilon \geq c_\epsilon\mu_0 > 0.$$

Since we know  $x_\epsilon \notin \mathcal{T}$ , (12) implies that

$$H(x_\epsilon, \zeta_\epsilon + c_\epsilon D\mu(x_\epsilon)) + \eta_+(x_\epsilon) \leq 0.$$

Arguing in the same way, from the fact that  $y = y_\epsilon$  maximizes

$$v_\epsilon(y) - [u(x_\epsilon) - \epsilon^{-1}\xi(x_\epsilon - y)],$$

we are led to the conclusion that

$$H(y_\epsilon, \zeta_\epsilon - c_\epsilon D\mu(y_\epsilon)) - \eta_-(y_\epsilon) \geq 0.$$

Therefore,

$$0 < \eta_0 \leq \eta_+(x_\epsilon) + \eta_-(y_\epsilon) \leq H(y_\epsilon, \zeta_\epsilon - c_\epsilon D\mu(y_\epsilon)) - H(x_\epsilon, \zeta_\epsilon + c_\epsilon D\mu(x_\epsilon)).$$

Now we know that for some constant  $K$  (independent of  $\epsilon > 0$ ),  $|\zeta_\epsilon \pm c_\epsilon D\mu| \leq \epsilon^{-1}K$ . Our continuity hypotheses on  $H(x, p)$  imply that the right-hand side of the above expression is bounded above by

$$m(|x_\epsilon - y_\epsilon|) + M(2c_\epsilon|\mu|, \epsilon^{-1}K).$$

The first term converges to 0 because  $|x_\epsilon - y_\epsilon| \rightarrow 0$ . We can choose  $c_\epsilon \downarrow 0$  so that the second term  $\rightarrow 0$  as well. For such choices we have a contradiction to the positive lower bound  $\eta_0$ . This contradiction implies that  $\sup_\Omega [u(x) - v(x)] \leq 0$ , concluding the proof.  $\square$

**4.  $p$ -monotone uniqueness.** We want to use monotonicity properties of  $H$  in the  $p$  variable to produce the additional  $\eta_{\pm}(x)$  terms needed for application of Lemma 2. Intuitively, we want to use a property such as

$$H(x, s\zeta) < H(x, \zeta) \text{ for } 0 < s < 1 \quad \text{and} \quad H(x, \zeta) < H(x, s\zeta) \text{ for } 1 < s.$$

However, this is considerably stronger than needed for the proof. For the subsolution case,  $0 < s < 1$ , we don't really need  $H(x, s\zeta) < H(x, \zeta)$ , only  $H(x, s\zeta) < 0$ , but holding uniformly on compacts disjoint from  $\mathcal{T}$ . We express this as

$$H(x, s\zeta) + \eta_s(x) \leq 0$$

for some function  $\eta_s(x)$  which is uniformly positive on each  $\Omega_{\delta, \mathcal{T}}$ . Moreover, we only need these properties for those  $\zeta \in D^+u(x)$  such that the inequality (10) is not satisfied by virtue of the  $-d_i \cdot \zeta$  terms. This can be stated succinctly by saying that  $u(x)$  is a subsolution of

$$H(x, sDu(x)) + \eta_s(x) = 0 \text{ on } \Omega \setminus \mathcal{T} \text{ with } -d_i \cdot Du(x) = 0 \text{ on } \partial\Omega \setminus \mathcal{T},$$

which is what we need to invoke Lemma 2. The following definition is based on this weakened monotonicity requirement.

DEFINITION 3. A viscosity subsolution  $u(x)$  of

$$(18) \quad H(x, Du(x)) = 0 \text{ on } \Omega \setminus \mathcal{T} \text{ with } -d_i \cdot Du(x) = 0 \text{ on } \partial\Omega \setminus \mathcal{T}$$

is called  $p$ -monotone if, for some  $\delta_0 > 0$  and each  $1 - \delta_0 < s < 1$ , there exists a function  $\eta_s : \Omega \rightarrow [0, \infty)$  with  $\inf_{\Omega_{\delta, \mathcal{T}}} \eta_s > 0$  for each  $\delta > 0$ , so that  $u(x)$  is a subsolution of

$$H(x, sDu(x)) + \eta_s(x) = 0 \text{ on } \Omega \setminus \mathcal{T} \text{ with } -d_i \cdot Du(x) = 0 \text{ on } \partial\Omega \setminus \mathcal{T}.$$

A viscosity supersolution  $v(x)$  of (18) is called  $p$ -monotone if, for some  $\delta_0 > 0$  and each  $1 < s < 1 + \delta_0$ , there exists a function  $\eta_s : \Omega \rightarrow [0, \infty)$  with  $\inf_{\Omega_{\delta, \mathcal{T}}} \eta_s > 0$  for each  $\delta > 0$ , so that  $v(x)$  is a supersolution of

$$H(x, sDv(x)) - \eta_s(x) = 0 \text{ on } \Omega \setminus \mathcal{T} \text{ with } -d_i \cdot Dv(x) = 0 \text{ on } \partial\Omega \setminus \mathcal{T}.$$

A viscosity solution which is both a  $p$ -monotone subsolution and a  $p$ -monotone supersolution is called a  $p$ -monotone solution.

We observe that  $p$ -monotonicity concerns  $s\zeta$  for  $s < 1$  in the case of a subsolution, but  $1 < s$  for a supersolution. It is possible for a viscosity solution to have the  $p$ -monotone property in the supersolution sense but not the subsolution sense. This would be a viscosity solution and a  $p$ -monotone supersolution, but not a  $p$ -monotone solution.

We are now ready for our main theorem. The basic idea is that if  $u(x)$  is a subsolution, then  $p$ -monotonicity will imply that  $su(x) + (s-1)c$  is a "strict" subsolution. (The constant term  $(s-1)c$  is to insure that  $su(x) + (s-1)c \leq g$  in case  $g(x) < 0$ . The fact that  $H$  has no  $u$ -dependence allows us to add such constants with impunity.) We then appeal to Lemma 2 and let  $s \uparrow 1$ .

THEOREM 4. Suppose  $u$  is a  $p$ -monotone subsolution of (9) with  $u \leq g$  on  $\mathcal{T}$ , and  $v$  is (any) supersolution with  $g \leq v$  on  $\mathcal{T}$ . Then  $u(x) \leq v(x)$  for all  $x \in \Omega$ . Likewise if  $u$  is (any) subsolution and  $v$  is a  $p$ -monotone supersolution with  $u \leq g \leq v$  on  $\mathcal{T}$ ,

then  $u(x) \leq v(x)$  for all  $x \in \Omega$ . If (9) has a  $p$ -monotone solution  $v$ , then  $v$  is the complete solution (i.e., it is the unique solution, the maximal subsolution, and the minimal supersolution).

COROLLARY 5. A viscosity solution which is a  $p$ -monotone supersolution is the maximal subsolution.

*Proof.* We focus on the  $p$ -monotone subsolution case. Let

$$-c = \min_{\mathcal{T}} g(x).$$

On  $\mathcal{T}$  we have  $u(x) + c \leq g(x) + c$ . Since  $0 \leq g(x) + c$ , it follows that (for any  $0 < s < 1$ )  $s(u(x) + c) \leq g(x) + c$  on  $\mathcal{T}$ . This is equivalent to

$$u_s(x) \doteq su(x) + (s - 1)c \leq g(x), \quad x \in \mathcal{T}.$$

Now  $\zeta_s \in D^+u_s(x)$  iff  $\zeta_s = s\zeta$  for some  $\zeta \in D^+u(x)$ . If  $x \in \partial\Omega \setminus \mathcal{T}$  and  $-d_i \cdot \zeta_s > 0$  for all  $i \in I(x)$ , then  $-d_i \cdot \zeta > 0$  for all  $i \in I(x)$ , so by the  $p$ -monotone subsolution property for  $u(x)$ ,

$$H(x, \zeta_s) + \eta_s(x) = H(x, s\zeta) + \eta_s(x) \leq 0.$$

The same inequality holds for  $x \in \Omega^\circ$ . We conclude that  $u_s$  is a viscosity subsolution of

$$H(x, Du_s(x)) + \eta_s(x) = 0 \text{ on } \Omega \setminus \mathcal{T} \text{ with } -d_i \cdot Du_s(x) = 0 \text{ on } \partial\Omega \setminus \mathcal{T}.$$

We can now apply Lemma 2 to  $u_s$  and  $v$ , using  $\eta_+(x) = \eta_s(x)$  for  $u_s$  and  $\eta_-(x) \equiv 0$  for  $v$ . The lemma implies that  $u_s(x) \leq v(x)$  all  $x \in \Omega$  as follows: for all  $1 - \delta_0 < s < 1$ ,

$$su(x) + (s - 1)c \leq v(x).$$

Letting  $s \uparrow 1$  implies  $u(x) \leq v(x)$ , as claimed. The supersolution case (using  $s \downarrow 1$ ) is analogous. The rest of the assertions of the theorem and corollary are now elementary.  $\square$

In general the  $p$ -monotone property may depend on the specific solution, since the definition only concerns  $\zeta \in D^\pm u(x)$ . However, for some Hamiltonians all (sub- or super-) solutions will be  $p$ -monotone. We consider in particular Hamiltonians associated with a running cost  $L(x, a, b)$ ,

$$(19) \quad H(x, p) = \inf_{b \in \mathcal{B}} \sup_{a \in \mathcal{A}} \{-p \cdot f(x, a, b) - L(x, a, b)\},$$

still assuming the continuity hypotheses of section 2.1 above. The next lemma shows that uniform positivity of the running cost provides a simple sufficient condition for all solutions to have the  $p$ -monotone property. (The argument is embedded in the proof of [2, Theorem 5].) When the lemma applies, Theorem 4 becomes a simple comparison and uniqueness theorem for *all* viscosity solutions.

LEMMA 6. Suppose that  $H(x, p)$  is given by (19), and that there exists a function  $\sigma : \Omega \rightarrow [0, \infty)$  with the property that  $0 < \inf_{\Omega_{\delta, \mathcal{T}}} \sigma(x)$  for each  $\delta > 0$  and for which

$$\sigma(x) \leq L(x, a, b)$$

for all  $a \in \mathcal{A}$ ,  $b \in \mathcal{B}$ ,  $x \in \Omega$ . Then every subsolution and every supersolution of (9) is  $p$ -monotone.

Note that since  $\sigma(x)$  is allowed to vanish on  $\mathcal{T}$ , the choice of  $\mathcal{T}$  may affect the applicability of the lemma.

*Proof.* Suppose that  $0 < s < 1$  and consider any  $\zeta \in \mathbb{R}^n$ . We have

$$\begin{aligned} -s\zeta \cdot f(a, b) - L(x, a, b) &= s[-\zeta \cdot f(a, b) - L(x, a, b)] - (1-s)L(x, a, b) \\ &\leq s[-\zeta \cdot f(a, b) - L(x, a, b)] - (1-s)\sigma(x). \end{aligned}$$

Taking  $\inf_{b \in \mathcal{B}} \sup_{a \in \mathcal{A}}$  yields  $H(x, s\zeta) \leq sH(x, \zeta) - (1-s)\sigma(x)$ . Let  $\eta_s(x) = (1-s)\sigma(x)$ . We have

$$H(x, s\zeta) + \eta_s(x) \leq sH(x, \zeta),$$

holding for all  $\zeta$ . It follows from this that any subsolution is a  $p$ -monotone subsolution.

The supersolution argument is analogous by using  $1 < s$ ,  $\eta_s(x) = (s-1)\sigma(x)$ , with the appropriate inequalities reversed.  $\square$

**5. Examples.** We now discuss several examples, most of which are taken from existing literature, which illustrate the applicability and limitations of the above results. In all the examples, the Hamiltonian has the form  $H(x, p) = H_0(p) - h(x)$ , for which the hypotheses (5) are easy to verify. We omit those details, as well as the confirmations of the  $B$ -hypothesis and the coercivity hypothesis.

Numerous optimal control or differential game problems have been posed for “fluid limits” of queueing networks. The most common domain for these examples is the nonnegative orthant  $\Omega = \mathbb{R}_+^d$ . Being unbounded, this is outside the scope of our results above. Our first example makes the point that our main result, Theorem 4, can fail in unbounded domains.

*Example 1.* In Day [11] an example in two dimensions was considered for the Hamiltonian

$$(20) \quad H(x, p) = \frac{1}{2}\|p\|^2 - \frac{1}{2}\|x\|^2.$$

This arises as in (19) using

$$(21) \quad L(x, a, b) = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|a\|^2, \quad f(x, a, b) = a,$$

with  $\mathcal{A} = \mathbb{R}^2$ . ( $\mathcal{B}$  is irrelevant.) With  $\mathcal{T} = \{(0, 0)\}$  and  $\sigma(x) = \frac{1}{2}\|x\|^2$  we see that Lemma 6 applies, and therefore *all* viscosity solutions are  $p$ -monotone. The equation, however, was considered in the *unbounded* half-space  $\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 \leq 1\}$ , using  $d = (-1, 0)$  ( $= -\gamma(x)$  in the notation of [11]) for the boundary condition on  $\partial\Omega$ , and taking  $g(0, 0) = 0$ . If Theorem 4 were valid for unbounded domains, solutions would be unique. However, in [11] it was shown that both  $v(x) = \frac{1}{2}x_1^2 \pm \frac{1}{2}x_2^2$  are viscosity solutions.

The rest of our examples will use compact  $\Omega$  as hypothesized. Examples 2–4 illustrate the applicability of Lemma 6.

*Example 2.* The “eikonal” equation

$$|Du(x)| - h(x) = 0, \quad u(x) = g(x) \text{ on } \partial\Omega,$$

with  $h(x) > 0$  on  $\Omega$  was cited above in reference to the approach of Ishii [18]. We simply observe that  $H(x, p) = |p| - h(x)$  is obtained from (19) using  $f(x, a, b) = a$ ,  $a \in \mathcal{A} = \{a : |a| \leq 1\}$ , and  $L(x, a, b) = h(x)$ . ( $\mathcal{B}$  is irrelevant.) Lemma 6 applies, so

that all solutions are  $p$ -monotone, and Theorem 4 provides the usual comparison and uniqueness results for this Hamiltonian on bounded domains, for any choice of  $\mathcal{T}$ .

*Example 3.* The doctoral dissertation of Menendez [20] considers an example using dynamics of the form

$$(22) \quad f(x, a, b) = \lambda - Ga$$

in a bounded rectangle  $\Omega$  in two dimensions. The running cost  $L(x, a, b) = \frac{1}{2}|x|^2 + 1.1$  is strictly positive. The control set  $\mathcal{A}$  is compact and there is no dependence on  $b$ . This problem again falls within the scope of Lemma 6 (regardless of  $\mathcal{T}$ ), so that Theorem 4 applies to all viscosity solutions. Although [20] does not employ viscosity solution techniques, our results above show that they would be a viable alternative approach.

*Example 4.* A rather different problem is considered by Atar, Dupuis, and Schwartz [2]. Here a differential game is studied which provides an asymptotic description of a risk-sensitive stochastic control problem. In the stochastic control problem, reaching the target set  $\mathcal{T}$  ( $\partial_o G$  in their notation) is viewed as an event to be avoided, so the control attempts to maximize the time until this occurs. This becomes the maximizing player in the limiting game. The minimizing player emerges from the asymptotic analysis as the limiting representation of the random fluctuations.

The problem fits our format in the case that all the arrival parameters  $\lambda_i$  are positive. (If some  $\lambda_i = 0$ , then different boundary conditions are to be used on some parts of  $\partial\Omega \setminus \mathcal{T}$ .) We recast their problem in our notation.  $\Omega$  (their  $G$ ) is the rectangle  $\times_1^d [0, z_i]$  in  $\mathbb{R}^d$ .  $\mathcal{T}$  consists of the portion of the boundary where  $x_i = z_i$  for one or more coordinates. The  $d_i$  are the  $-\tilde{v}_i$  (below) for the respective faces  $\partial_i \Omega = \{x : x_i = 0\}$ . The maximizing player chooses the control  $b = (u_1, \dots, u_d)$  in a compact polygon  $\mathcal{B}$ . The minimizing player chooses a vector of rate perturbation factors  $a = (\alpha_i^\lambda, \alpha_i^\mu : i = 1, \dots, d)$ , with  $a \in \mathcal{A} = [0, \infty)^{2d}$ . The state dynamics are

$$f(x, a, b) = \sum_i \lambda_i \alpha_i^\lambda e_i + \sum_i u_i \mu_i \alpha_i^\mu \tilde{v}_i,$$

where  $e_i$  are the standard unit vectors in  $\mathbb{R}^d$ , and  $\tilde{v}_i$  are the *service event vectors*,  $\tilde{v}_i = e_{i'} - e_i$ , where  $i \rightarrow i'$  indicates the routing sequence in the network. The running cost is

$$L(x, a, b) = c + \sum_i \lambda_i \ell(\alpha_i^\lambda) + \sum_i u_i \mu_i \ell(\alpha_i^\mu),$$

$$\text{where } \ell(\alpha) = \alpha \log(\alpha) - \alpha + 1.$$

Here  $c > 0$  is a positive constant,  $\lambda_i > 0$ , and  $\mu_i \geq 0$ , so  $L(x, a, b) \geq c$ . Thus the hypotheses of Lemma 6 are satisfied once again, so that Theorem 4 applies to all viscosity solutions.

Our last two examples are beyond the scope of Lemma 6, and the details are more involved. The following lemma will assist us in checking the boundary conditions for (locally) smooth solutions.

LEMMA 7. Assume the coercivity condition (6). Suppose  $x \in \partial\Omega$  and  $u$  is continuously differentiable in a neighborhood of  $x$ .

(a)  $\zeta \in D^+u(x)$  iff  $\zeta = Du(x) + \sum_{i \in I(x)} \beta_i n_i$  for some choice of  $\beta_i \geq 0$ . Analogously,  $\zeta \in D^-u(x)$  iff  $\zeta = Du(x) - \sum_{i \in I(x)} \beta_i n_i$  for some  $\beta_i \geq 0$ .

(b) If  $-d_i \cdot Du(x) \leq 0$  for all  $i \in I(x)$ , then the viscosity subsolution property with boundary conditions holds as follows: for all  $\zeta \in D^+u(x)$ ,

$$H(x, \zeta) \wedge \min_{i \in I(x)} (-d_i \cdot \zeta) \leq 0.$$

Analogously, if  $-d_i \cdot Du(x) \geq 0$  for all  $i \in I(x)$ , then for all  $\zeta \in D^-u(x)$ ,

$$H(x, \zeta) \vee \max_{i \in I(x)} (-d_i \cdot \zeta) \geq 0.$$

*Proof.* The proof of (a) is the first paragraph of the proof of [12, Theorem 2.1]. For (b), suppose that  $-d_i \cdot Du(x) \leq 0$  for all  $i \in I(x)$  and consider any  $\zeta \in D^+u(x)$ . By (a) we know that  $\zeta = Du(x) + \sum_{i \in I(x)} \beta_i n_i$  with  $\beta_i \geq 0$ . We can assume some  $\beta_i > 0$  for some  $i \in I(x)$ ; else  $-d_i \cdot \zeta = -d_i \cdot Du(x) \leq 0$  follows directly. Observe that

$$\sum_{i \in I(x)} \beta_i d_i \cdot \zeta = \left( \sum_{i \in I(x)} \beta_i d_i \cdot Du(x) \right) + \left( \sum_{i \in I(x)} \beta_i d_i \right) \cdot \left( \sum_{i \in I(x)} \beta_i n_i \right).$$

By hypothesis, the first term on the right side is nonnegative. The last term is positive by the coercivity hypothesis and our assumption that  $\beta_i > 0$  for some  $i$ . Therefore the left side is positive. This implies that  $d_i \cdot \zeta > 0$  for some  $i \in I(x)$ . Consequently,

$$H(x, \zeta) \wedge \min_{i \in I(x)} (-d_i \cdot \zeta) \leq 0,$$

regardless of the value of  $H(x, \zeta)$ . The supersolution case in (b) is argued analogously.  $\square$

*Example 5.* The recent papers [3], [13], and [12] of Day and others explore a robust control approach to fluid queueing models, using state dynamics of the form

$$f(x, a, b) = b - Ga,$$

a compact control space  $\mathcal{A}$ , and opposing quadratic costs for the state and “disturbance”  $b \in \mathcal{B} = \mathbb{R}^n$  as follows:

$$(23) \quad L(x, a, b) = \frac{1}{2} \|x\|^2 - \frac{1}{2} \|b\|^2.$$

The resulting Hamiltonian is

$$(24) \quad \begin{aligned} H(x, p) &= \sup_{a \in \mathcal{A}} p \cdot Ga - \frac{1}{2} \|x\|^2 - \frac{1}{2} \|p\|^2 \\ &= \frac{1}{2} \sup_{a \in \mathcal{A}} (\|Ga\|^2 - \|p - Ga\|^2 - \|x\|^2). \end{aligned}$$

Since  $\|Ga\|$ ,  $a \in \mathcal{A}$  is bounded, we see from the second form that  $H(x, p) \geq 0$  implies a bound on  $\|x\|$ . Thus these problems are reasonable to consider *only* in bounded domains  $\Omega$ . The examples in the literature consider a bounded polygon  $\Omega$  consisting of  $x \in \mathbb{R}^d$  with  $x_i \geq 0$  and  $\eta \cdot x \leq c$  for a particular vector  $\eta$ . In [3] and [13] the boundary  $\eta \cdot x = c$  is omitted from  $\Omega$  and in its place an admissibility condition is imposed on controls, which prohibits the state from approaching this missing boundary. (See the “minimum performance criterion” and its discussion in section 2.4 of [13].) In [12] all of  $\partial\Omega$  is included, consistent with our formulation. Section 6 of [12] considers a

specific example of the type considered here. We will need to take advantage of certain explicit calculations, which would be cumbersome for that example. Instead, we will consider a simple instance of the example(s) of [3, sections 1–3], modified to include all of  $\partial\Omega$  in accordance with our hypotheses here.

We let  $G$  be the  $2 \times 2$  identity matrix. (In [3] this corresponds to  $s_i = \gamma = 1$ .) The control set is  $\mathcal{A} = \{(a_1, a_2) : 0 \leq a_i, a_1 + a_2 = 1\}$ . The Hamiltonian (24) simplifies to

$$(25) \quad H(x, p) = \max(p_1, p_2) - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|p\|^2.$$

We consider the planar domain

$$\Omega = \{x \in \mathbb{R}_+^2 : x_1 + x_2 \leq r\}$$

for  $r < 1$ . If  $r = 1$ , then our  $\Omega$  would be (the closure of) the domain considered in [3]. With  $r < 1$  the domain here is slightly smaller. This reduction of the domain is important for the  $p$ -monotone property. We identify the faces and respective normal vectors as follows:

$$\begin{aligned} \partial_1\Omega &= \{x \in \Omega : x_1 = 0\}, & n_1 &= (1, 0), \\ \partial_2\Omega &= \{x \in \Omega : x_2 = 0\}, & n_2 &= (0, 1), \\ \partial_3\Omega &= \{x \in \Omega : x_1 + x_2 = r\}, & n_3 &= (-1/\sqrt{2}, -1/\sqrt{2}). \end{aligned}$$

We take  $d_i = n_i$  for all the faces. The target set will be the origin,  $\mathcal{T} = \{(0, 0)\}$ , with  $g(0, 0) = 0$ .

The constructions of [3] produce a  $C^1$  solution to  $H(x, Du(x)) = 0$ . We will see that this is a  $p$ -monotone solution, even though Lemma 6 does not apply. We will first indicate briefly how the viscosity solution properties are verified, and then turn our attention to  $p$ -monotonicity. The solution is symmetric about the diagonal  $x_1 = x_2$ . We confine our discussion to the lower-right half of  $\Omega$ :  $0 \leq x_2 \leq x_1 \leq r$ . The analysis on the other half follows by symmetry.

In the subregion  $0 \leq x_2 \leq x_1 \leq r$  the solution is most conveniently described in terms of the orthogonal basis,

$$\mu = (1/2, -1/2), \quad \eta = (1/2, 1/2).$$

(In the notation of [3, page 335],  $\mu = \mu_1 = \eta_{\{1\}} - \eta_{\{1,2\}}$  and  $\eta = \mu_2 = \eta_{\{1,2\}}$ .) The gradient  $Du(x)$  is related to  $x$  in terms of parameters  $0 \leq t_1 \leq t_2 \leq \pi/2$  by the expressions

$$(26) \quad x = \sin(t_1)\mu + \sin(t_2)\eta, \quad Du(x) = [1 - \cos(t_1)]\mu + [1 - \cos(t_2)]\eta.$$

The parameters can be eliminated to obtain the explicit expressions for  $0 \leq x_2 \leq x_1 \leq r$ ,

$$\begin{aligned} u(x) &= x_1 - \frac{1}{4} \left( \sqrt{1 - (x_1 - x_2)^2} (x_1 - x_2) + \sin^{-1}(x_1 - x_2) \right. \\ &\quad \left. + (x_1 + x_2) \sqrt{1 - (x_1 + x_2)^2} + \sin^{-1}(x_1 + x_2) \right), \\ \frac{\partial u}{\partial x_1} &= 1 - \frac{1}{2} \left( \sqrt{1 - (x_1 - x_2)^2} + \sqrt{1 - (x_1 + x_2)^2} \right), \\ \frac{\partial u}{\partial x_2} &= \frac{1}{2} \left( \sqrt{1 - (x_1 - x_2)^2} - \sqrt{1 - (x_1 + x_2)^2} \right). \end{aligned}$$

The parametric representation is more convenient for most purposes. For instance, observe that for  $p = Du(x)$ ,  $\max(p_1, p_2) = p_1$  is equivalent to  $p \cdot \mu = \frac{1}{2}[1 - \cos(t_1)] \geq 0$ , which does hold. Therefore,

$$(27) \quad H(x, p) = p_1 - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|p\|^2 = p \cdot (\mu + \eta) - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|p\|^2.$$

It is now straightforward to evaluate this, using the orthogonality of  $\mu$  and  $\eta$  to confirm that  $H(x, Du(x)) = 0$ . The explicit formulae provide the easiest way to check that

$$(28) \quad \partial u / \partial x_i \geq 0 \text{ for both } i,$$

because  $x_2 \leq x_1$ , and

$$(29) \quad \partial u / \partial x_2 = 0 \text{ when } x_2 = 0.$$

By Lemma 7(b), (29) implies that the viscosity boundary conditions are satisfied on  $\partial_2\Omega$ . On  $\partial_3\Omega$  we have from (28) that  $-d_3 \cdot Du(x) \geq 0$ , so that the supersolution boundary condition is satisfied there, as well as at the corner  $(r, 0)$ .

The subsolution property on  $\partial_3\Omega$  and at the corner takes more careful examination. For these  $x$  we need to identify the  $\zeta \in D^+u(x)$ , for which  $-d_i \cdot Du(x) > 0$  for all  $i \in I(x)$ , and for these we need to check that  $H(x, \zeta) \leq 0$  holds. Consider the corner  $x = (r, 0)$  specifically. From the explicit formulas,  $Du(x) = (1 - \sqrt{1 - r^2}, 0)$ . By Lemma 7, the  $\zeta \in D^+u(x)$  are

$$\zeta = Du(x) + \beta_2 n_2 + \beta_3 n_3, \quad \beta_i \geq 0.$$

One finds that the  $\zeta$  with  $\beta_i \geq 0$  and  $-d_i \cdot \zeta > 0$  comprise the triangle in the  $\zeta$ -plane with vertices  $(0, 0)$ ,  $Du(x) = [1 - \sqrt{1 - r^2}](1, 0)$ , and  $[1 - \sqrt{1 - r^2}](\frac{1}{2}, -\frac{1}{2})$ . For future reference, notice that all such  $\zeta$  satisfy

$$(30) \quad \|\zeta\| \leq \|Du(x)\|.$$

What we need at the moment is that  $\zeta_1 \leq 0 \leq \zeta_2$ , so that just as in (27),

$$H(x, \zeta) = \zeta \cdot (\mu + \eta) - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|\zeta\|^2.$$

For the particular  $\zeta$  identified above, this works out to be

$$H(x, \zeta) = \frac{1}{2} \left( -\beta_2^2 + \sqrt{2}\beta_3\beta_2 - \beta_3 \left( \beta_3 + \sqrt{2}\sqrt{1 - r^2} \right) \right),$$

from which one may verify that  $H(x, \zeta) \leq 0$  for all  $\beta_i \geq 0$ . This confirms the viscosity subsolution property at the corner.

For  $x \in \partial_3\Omega$  with  $x_2 \leq x_1 < r$  the calculations are similar but simpler. The  $\zeta \in D^+u(x)$  with  $-d_3 \cdot \zeta > 0$  are  $\zeta = Du(x) + \beta_3 n_3$  with

$$(31) \quad 0 \leq \sqrt{2}\beta_3 < 1 - \cos(t_2).$$

Since  $n_3 = -\sqrt{2}\eta$ , we have

$$(32) \quad \zeta = [1 - \cos(t_1)]\mu + [1 - \cos(t_2) - \sqrt{2}\beta_3] \eta.$$



Notice that (32) implies that (30) again holds. Since  $\mu \cdot \zeta = \mu \cdot Du(x) \geq 0$ , we can again work out that

$$\begin{aligned} H(x, \zeta) &= \zeta \cdot (\mu + \eta) - \frac{1}{2} \|x\|^2 - \frac{1}{2} \|\zeta\|^2 \\ &= \frac{-1}{2} \beta_3 \left[ \sqrt{2} \cos(t_1) + \beta_3 \right] \leq 0 \text{ since } \beta_3 \geq 0. \end{aligned}$$

This completes the verification that  $u(x)$  is a viscosity solution to our problem.

We now consider  $p$ -monotonicity. Note that due to the  $-\|b\|^2$  term, there is no finite lower bound for the  $L$  of (23). Thus Lemma 6 does *not* apply. Even so, we will see that  $u(x)$  is a  $p$ -monotone solution. Observe that for any  $s > 0$ , we have  $\max(s\zeta_1, s\zeta_2) = s \max(\zeta_1, \zeta_2)$ . As a consequence we have the following identity:

$$(33) \quad H(x, s\zeta) = sH(x, \zeta) + (s-1) \left[ \frac{1}{2} \|x\|^2 - \frac{s}{2} \|\zeta\|^2 \right].$$

Consider the supersolution  $p$ -monotonicity property first. As observed above,  $-d_i \cdot Du(x) \geq 0$  on all boundary faces, so that only the interior points are involved in the  $p$ -monotonicity supersolution property. Since  $H(x, Du(x)) = 0$ , we see from (33) that  $p$ -monotonicity requires that

$$(34) \quad 0 < \frac{1}{2} \|x\|^2 - \frac{s}{2} \|Du(x)\|^2$$

for  $x \neq 0$  and  $s \approx 1$ . For  $s = 1$  this is the *positive storage condition* (see [3, (2.25)] and [12, (33)]), which was important for the verification results obtained in those papers. Here we are interested in  $1 < s$ . The parametric representation of  $Du(x)$  allows us to check (34) directly as follows:

$$\frac{1}{2} \|x\|^2 - \frac{s}{2} \|Du(x)\|^2 = \frac{1}{4} [\sin(t_1)^2 - s(1 - \cos(t_1))^2] + \frac{1}{4} [\sin(t_2)^2 - s(1 - \cos(t_2))^2].$$

Now  $0 \leq t_t \leq t_2$  and  $\frac{1}{2} \sin(t_2) = \eta \cdot x \leq \frac{r}{2}$ . Thus  $t_i \leq \sin^{-1}(r) < \frac{\pi}{2}$ , since  $r < 1$ . It is elementary to check that there exists  $\delta_0 > 0$ , so that

$$\sin(t)^2 - s(1 - \cos(t))^2 > 0$$

for all  $0 \leq t \leq \sin^{-1}(r)$  and all  $0 < s < 1 + \delta_0$ . This implies that (34) holds, and so  $u(x)$  is indeed a  $p$ -monotone supersolution, using

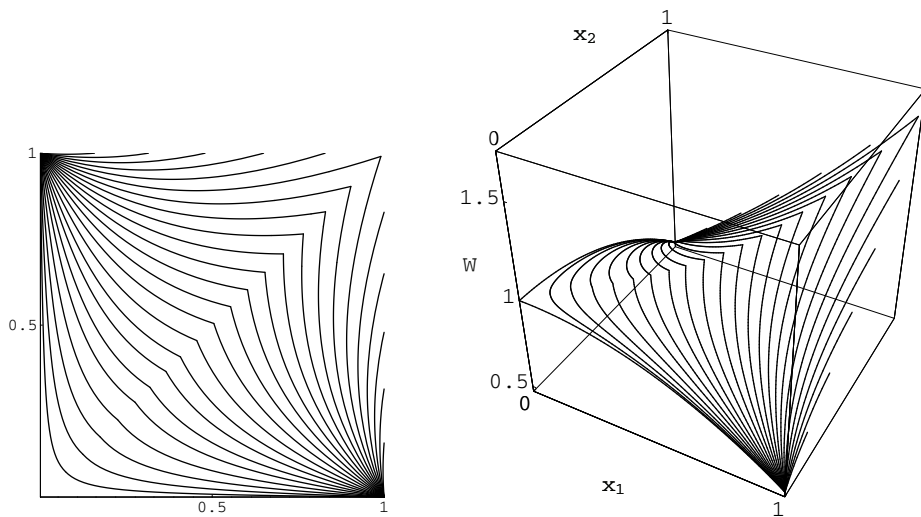
$$\eta_s(x) = (s-1) \left[ \frac{1}{2} \|x\|^2 - \frac{s}{2} \|Du(x)\|^2 \right].$$

Finally, consider the subsolution  $p$ -monotone property. Based on (33), for  $s-1 < 0$ , we need to know that, for  $\zeta \in D^+u(x)$  with  $-d_i \cdot \zeta > 0$  all  $i \in I(x)$ ,

$$0 < \frac{1}{2} \|x\|^2 - \frac{s}{2} \|\zeta\|^2.$$

But we observed in (30) above that for all such  $\zeta$ ,  $\|\zeta\| \leq \|Du(x)\|$ , and so

$$\frac{1}{2} \|x\|^2 - \frac{s}{2} \|\zeta\|^2 \geq \frac{1}{2} \|x\|^2 - \frac{s}{2} \|Du(x)\|^2.$$

FIG. 1.  $w(x)$ .

Thus we can again use

$$\eta_s(x) = (s-1) \left[ \frac{1}{2} \|x\|^2 - \frac{s}{2} \|Du(x)\|^2 \right],$$

which is strictly positive on  $\Omega \setminus \mathcal{T}$ , as shown above.

In summary,  $u(x)$  is a  $p$ -monotone viscosity solution and hence the complete solution of our problem.

Finally, we offer a new example which exhibits nonuniqueness of solutions when no  $p$ -monotone solution exists, but for which comparisons based on  $p$ -monotonicity properties are still possible.

*Example 6.* We return to the Hamiltonian (20) but consider the cube

$$\Omega = \{(x_1, x_2) : 0 \leq x_i \leq 1\}.$$

We number the boundary faces as

$$\begin{aligned} \partial_1 \Omega &= \{x \in \Omega : x_1 = 0\}, & \partial_2 \Omega &= \{x \in \Omega : x_2 = 0\}, \\ \partial_3 \Omega &= \{x \in \Omega : x_1 = 1\}, & \partial_4 \Omega &= \{x \in \Omega : x_2 = 1\}. \end{aligned}$$

The normals are  $n_1 = (1, 0)$ ,  $n_2 = (0, 1)$ ,  $n_3 = (-1, 0)$ , and  $n_4 = (0, -1)$ , and we take  $d_i = n_i$  for all faces. Consider the target set consisting of the two off-diagonal corners,  $\mathcal{T} = \{(1, 0), (0, 1)\}$ , taking  $g = \frac{1}{2}$  at both corners.

It is elementary to check that  $v(x) = \frac{1}{2}(x_1^2 + x_2^2)$  is a classical solution of  $0 = H(x, Dv(x))$  in the interior of  $\Omega$ . A second solution  $w$  is illustrated in Figure 1. It is symmetric about the diagonal line  $\Gamma = \{x \in \Omega : x_1 = x_2\}$ , but is nondifferentiable on  $\Gamma$  (and at the corners in  $\mathcal{T}$ ). In the upper left triangle,  $0 \leq x_1 \leq x_2 \leq 1$ , it is constructed from the family of characteristics (illustrated in the left pane of Figure 1),

$$\begin{aligned} \dot{x} &= H_p(x, p) = p; & x(0) &= (0, 1), \\ \dot{p} &= -H_x(x, p) = x; & p(0) &= (\cos(\theta), -\sin(\theta)), \quad 0 \leq \theta \leq \pi/2, \\ \dot{w} &= p \cdot \dot{x}; & w(0) &= 1/2 = g(x(0)), \end{aligned}$$

and extended by symmetry across  $\Gamma$ . It turns out that *both*  $v$  and  $w$  are viscosity solutions of  $H(x, Du(x)) = 0$  in  $\Omega \setminus \mathcal{T}$  with  $-d_i \cdot Du(x) = 0$  on  $\partial\Omega \setminus \mathcal{T}$  and  $u = g$  on  $\mathcal{T}$ . Moreover,  $v$  satisfies the oblique derivative boundary conditions at the points of  $\mathcal{T}$  as well. The verification of these assertions is similar to that of the previous example; we omit it for brevity.

In light of Theorem 4, neither  $v$  nor  $w$  can be a  $p$ -monotone solution. Lemma 6 does not apply here since  $x = (0, 0)$  does not belong to  $\mathcal{T}$  and  $L$  of (21) has no positive lower bound at this point. In fact, neither  $v$  nor  $w$  is a  $p$ -monotone solution in either the sub- or supersolution sense. In order for  $v$  to be a  $p$ -monotone subsolution we would need, for  $0 < s < 1$ , a function  $\eta_s(x) > 0$  (off  $\mathcal{T}$ ) with

$$H(x, sDv(x)) \leq -\eta_s(x), \quad x \in \Omega^\circ.$$

Now

$$H(x, sp) = s^2 \frac{1}{2} \|p\|^2 - \frac{1}{2} \|x\|^2 = s^2 H(x, p) + \frac{s^2 - 1}{2} \|x\|^2.$$

Since  $H(x, Dv(x)) = 0$ ,

$$H(x, sDv(x)) = \frac{s^2 - 1}{2} \|x\|^2.$$

Thus we would need  $\frac{s^2-1}{2} \|x\|^2 \leq -\eta_s(x)$  to be *uniformly* negative in a neighborhood of  $(0, 0)$ . This is clearly not possible. The same argument applies to  $w$  if we keep  $x$  off the diagonal. For the supersolution case, we would need  $\frac{s^2-1}{2} \|x\|^2 \geq \eta_s(x)$  to be uniformly positive in a neighborhood of  $(0, 0)$ , which is likewise impossible.

In Figure 2 we have plotted both solutions  $v$  and  $w$ . It is apparent that  $v \leq w$ . This can be deduced from Theorem 4 by considering the enlarged target set  $\mathcal{T}' = \{(0, 0), (0, 1), (1, 0)\}$ . Now Lemma 6 *does* apply; both  $v$  and  $w$  are  $p$ -monotone for this  $\mathcal{T}'$ . If we take  $g(0, 0) = 0 = v(0, 0)$ , then  $v$  is the complete solution of the

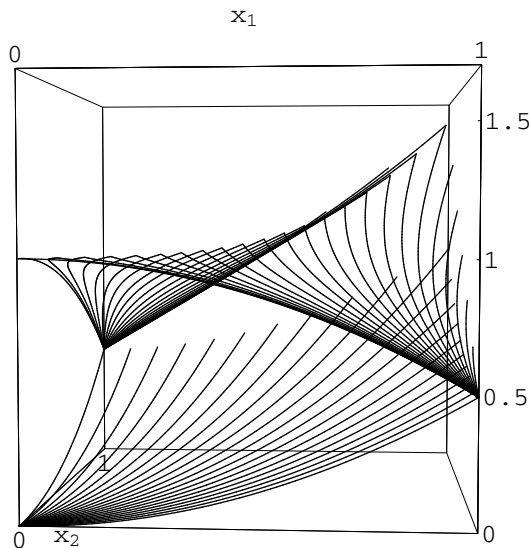


FIG. 2.  $v(x) \leq w(x)$ .

problem, but since  $w(0,0) = 1 > g(0,0)$ ,  $w$  is only a supersolution. Thus  $v \leq w$  follows from the comparison theorem. If instead we take  $g(0,0) = 1$ , then  $w$  is the complete solution. We obtain a supersolution by adding a constant to  $v$  as follows:  $\tilde{v}(x) = 1 + v(x)$ . In that case,  $\tilde{v} \geq g$  on  $\mathcal{T}$ , so that the comparison theorem implies  $v + 1 \geq w$ .

Also consider the target set  $\mathcal{T}'' = \{(0,0)\}$  consisting of the origin alone, with  $g(0,0) = 0$ . As above, Lemma 6 applies, so that  $v$  is the complete solution. According to Theorem 4, there can be no other viscosity solutions. Adding a constant,  $w - 1$  conforms to  $g$  at the origin. But investigation of the corners  $(0,1)$  and  $(1,0)$  shows that the supersolution condition fails there (details omitted). It is, however, a subsolution, which implies  $w - 1 \leq v$ , as we already deduced above.

## REFERENCES

- [1] R. ATAR AND P. DUPUIS, *A differential game with constrained dynamics and viscosity solutions of a related HJB equation*, Nonlinear Anal., 51 (2002), pp. 1105–1130.
- [2] R. ATAR, P. DUPUIS, AND A. SCHWARTZ, *An escape time criterion for queueing networks: Asymptotic risk-sensitive control via differential games*, Math. Oper. Res., 28 (2003), pp. 801–835.
- [3] J. A. BALL, M. V. DAY, AND P. KACHROO, *Robust feedback control of a single server queueing system*, Math. Control Signals Systems, 12 (1999), pp. 307–345.
- [4] M. BARDI AND I. CAPPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser, Boston, 1997.
- [5] M. BARDI AND P. SORAVIA, *Hamilton–Jacobi equations with singular boundary conditions on a free boundary and applications to differential games*, Trans. Amer. Math. Soc., 325 (1991), pp. 205–229.
- [6] G. BARLES AND P. L. LIONS, *Fully nonlinear Neumann type boundary conditions for first-order Hamilton–Jacobi equations*, Nonlinear Anal., 16 (1991), pp. 143–153.
- [7] F. CAMILLI AND A. SICONOLFI, *Maximal subsolutions for a class of degenerate Hamilton–Jacobi problems* Indiana Univ. Math. J., 48 (1999), pp. 1111–1132.
- [8] M. G. CRANDALL, H. ISHII, AND P. L. LIONS, *Uniqueness of viscosity solutions of Hamilton–Jacobi equations revisited*, J. Math. Soc. Japan, 39 (1987), pp. 581–595.
- [9] M. G. CRANDALL, H. ISHII, AND P. L. LIONS, *User’s guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.
- [10] M. V. DAY, *On the velocity projection map for polyhedral Skorokhod problems*, Appl. Math. E-Notes, 5 (2005), pp. 52–59 (electronic).
- [11] M. V. DAY, *On Neumann type boundary conditions for Hamilton–Jacobi equations in smooth domains*, Appl. Math. Optim., 53 (2006), pp. 359–381.
- [12] M. V. DAY, *Boundary-influenced robust controls: Two network examples*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 662–698.
- [13] M. V. DAY, J. HALL, J. MENENDEZ, D. POTTER, AND I. ROTHSTEIN, *Robust optimal service analysis of single-server re-entrant queues*, Comput. Optim. Appl., 22 (2002), pp. 261–302.
- [14] P. DUPUIS AND H. ISHII, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics Stochastics Rep., 35 (1991), pp. 31–62.
- [15] P. DUPUIS AND H. ISHII, *On oblique derivative problems for fully nonlinear second-order elliptic PDE’s on domains with corners* Hokkaido Math. J., 20 (1991), pp. 135–164.
- [16] P. DUPUIS, H. ISHII, AND H. M. SONER, *A viscosity solution approach to the asymptotic analysis of queueing systems*, Ann. Probab., 18 (1990), pp. 226–255.
- [17] P. DUPUIS AND K. RAMANAN, *Convex duality and the Skorokhod problem*, I and II, Probab. Theory Related Fields, 115 (1999), pp. 153–195, 197–236.
- [18] H. ISHII, *A simple, direct proof of uniqueness for solutions of Hamilton–Jacobi equations of Eikonal type*, Proc. Amer. Math. Soc., 100 (1987), pp. 247–251.
- [19] P. L. LIONS, *Neumann type boundary conditions for Hamilton–Jacobi equations*, Duke Math. J., 52 (1985), pp. 793–820.
- [20] J. MENENDEZ, *Computational Methods for Control of Queueing Models in Bounded Domains*, Ph.D. dissertation, Virginia Tech, Blacksburg, VA, 2007.
- [21] P. SORAVIA,  *$\mathcal{H}_\infty$  control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.

## DELAY-INTEGRAL-QUADRATIC CONSTRAINTS AND ABSOLUTE STABILITY OF TIME-PERIODIC FEEDBACK SYSTEMS\*

D. A. ALTSHULLER†

**Abstract.** The paper considers the problem of absolute stability of systems with time-periodic nonlinear blocks. The approach presented in this paper is based on the so-called quadratic criterion and relies on integral-quadratic inequalities (constraints), which also involve time delays. The results are given in the frequency domain. The paper also includes geometric interpretation and numerical treatment of the new stability criteria.

**Key words.** absolute stability, frequency domain, nonlinear control systems, time-periodic systems

**AMS subject classifications.** 93D09, 93D10

**DOI.** 10.1137/070702758

**1. Introduction.** Consider a system consisting of a linear block represented by an integral equation

$$(1.1) \quad \sigma(t) = \alpha(t) + K_0 \xi(t) + \int_0^t K(t-s) \xi(s) ds$$

and a nonlinear block represented by the equation

$$(1.2) \quad \xi(t) = \varphi(\sigma(t), t).$$

Here  $\sigma(t) \in \mathbf{R}^m$ ;  $\xi(t) \in \mathbf{R}^p$ ; and  $K_0$  and  $K(t)$  are matrices of appropriate dimensions. It will always be assumed that  $\varphi(\sigma, t+T) = \varphi(\sigma, t)$ . Additional assumptions concerning this function, hereafter called the nonlinearity, will be stated in the hypotheses of the theorems.

Clearly, this system of equations can be combined into a single nonlinear Volterra integral equation. Existence theory for the equations of this type is well established under very broad assumptions about the functions involved (see, for example, [11]). Therefore, we are not going to concern ourselves with this aspect of the problem.

In this paper we are going to address the problem of absolute stability of the system (1.1)–(1.2). Roughly speaking, this problem is concerned with finding conditions involving only the linear block such that the system is stable for all nonlinearities belonging to a certain class.

Most of the research of the stability problem for the systems of this type has been confined to the case when the nonlinearity is a scalar function not depending explicitly on time. When the nonlinearity is monotone with respect to the state variable  $\sigma$  and its slope does not exceed a constant  $\mu$ , the most general results have the so-called multiplier form: The system is absolutely stable if for some constant  $\varepsilon > 0$  and for all  $\omega \in (-\infty, +\infty)$

$$(1.3) \quad \operatorname{Re}\{\mu^{-1} + W(i\omega)\}Z(i\omega) > \varepsilon > 0.$$

---

\*Received by the editors September 13, 2007; accepted for publication (in revised form) August 27, 2008; published electronically January 7, 2009. Some results from this paper were presented (in a different form) at the IEEE 2002 Automatic Control Conference [1], the IEEE 2002 Conference on Decision and Control [2], and the IFAC 2004 Symposium on System, Structure, and Control [4].

<http://www.siam.org/journals/sicon/47-6/70275.html>

†Crane Aerospace and Electronics, 3000 Winona Ave., Burbank, CA 91510 (altshuller@ieee.org).

Here  $W(i\omega)$  is the frequency response of the linear block defined using the Fourier transform of its kernel:

$$(1.4) \quad W(i\omega) = -K_0 - \tilde{K}(i\omega) = -K_0 - \int_0^\infty K(t)e^{-i\omega t} dt.$$

The function  $Z(i\omega)$  is called the Zames–Falb multiplier after the classic paper by Zames and Falb [24], in which the most general form for  $Z(i\omega)$  has been given. The stability criteria written in this form have a convenient geometric interpretation [9].

For nonstationary systems, the problem of absolute stability was studied by Pyatnitsky and Molchanov [12, 13, 14, 16]. In [16] Pyatnitsky reduces the problem to the investigation of certain piecewise-linear systems and to variational problems. In [12, 13, 14] he and Molchanov focus on Lyapunov functions. Unfortunately, their criteria, unlike condition (1.3), are not easy to verify.

The objective of this paper is to obtain frequency-domain absolute stability criteria for systems with time-periodic nonlinearities. The early results for such systems go back to the book by Narendra and Taylor [15] and, for linear systems, to the two-volume classic by Yakubovich and Starzhinskii [23].

The method used in this paper is based on the so-called quadratic constraints. This method was used by Yakubovich [19] to prove some absolute stability criteria for systems similar in form to (1.1)–(1.2) but with stationary nonlinearities. The central concept of the method is to establish that the input and output signals of the nonlinear part of the system satisfy a certain set of integral-quadratic inequalities (constraints). These constraints may involve the values of these signals explicitly in time domain or, alternatively, their Fourier transforms. In the latter case, thoroughly investigated in the renown paper by Megretski and Rantzer [10] and further developed by Yakubovich [20, 21, 22], these constraints are said to be in the frequency domain.

The constraints used in this paper are in the time domain, but they also involve the values of the input and output signals at some earlier points in time (delays). Since they also involve integration over time, they are called delay-integral-quadratic constraints. They are, in fact, a special case of the constraints in frequency domain, but tend to arise more naturally from the properties of the input and output signals. For systems with stationary nonlinearities this method can be used [5] to prove the classical results of Zames and Falb [24].

In this paper the method will be extended to the case when the nonlinearities are periodic in time, and this periodicity, along with some other conditions, will play a crucial part in formulating the constraints.

The outline of the paper is as follows. First, for the sake of making the paper self-contained, we review the concept of the quadratic criterion of absolute stability (subsection 2.1). Then, in subsection 2.2, we shall prove two integral inequalities. The constraints will then follow naturally from these inequalities. The results will first be stated and proved for slope-restricted nonlinearities, thus extending the criteria of Zames and Falb to the case of time-periodic nonlinearities (subsection 3.1). We shall also give a geometric interpretation and numerical implementation of these new results (subsection 3.2). Next, in section 4, we consider the so-called quasimonotone nonlinearities (for the stationary case the corresponding result was proved by Barabanov [6]). In section 5 we shall prove a criterion for a class of MIMO linear time-periodic systems, which includes as a special case a SISO result by Yakubovich [22]. Finally, in section 6, we are going to obtain a criterion for a certain parametric class of nonlinearities, once again extending to the time-periodic case the result of Barabanov [6].

**2. Preliminaries.** In this section we present some background information on the so-called quadratic criterion of absolute stability. We also prove the integral inequalities needed for application of this criterion to the systems under consideration.

**2.1. Quadratic criterion for absolute stability.** For the sake of making this paper self-contained, we are going to restate the notation, definitions, and some lemmas (without proof) from the earlier paper [5].

Throughout the paper it will always be assumed that the linear block (1.1) satisfies the regularity conditions which we now define.

DEFINITION 2.1. *The linear block (1.1) satisfies the regularity conditions if*

$$(2.1) \quad |\alpha(\cdot)| \in L^2(0, +\infty) \cap L^\infty(0, +\infty)$$

and there exist positive constants  $C$  and  $\beta$ , such that

$$(2.2) \quad |K(t)| \leq Ce^{-\beta t}.$$

In this paper we are going to consider only locally square-integrable signals. In other words, it will be assumed that  $|\sigma(\cdot)|, |\xi(\cdot)| \in L^2_{loc}[0, +\infty]$ , where  $L^2_{loc}[0, +\infty]$  denotes the set of functions that are in  $L^2[0, t_0]$  for any  $t_0 > 0$ . Clearly, if the linear block satisfies the regularity conditions, then local square-integrability of the input signal implies local square-integrability of the output. Existence of such signals follows from the existence theorems for nonlinear integral equations.

For the purposes of this subsection, we let the nonlinear block of the system be represented instead of (1.2) by a more general form

$$(2.3) \quad [\sigma(\cdot), \xi(\cdot)] \in \mathcal{N}.$$

The set  $\mathcal{N}$  will be defined as follows: For any  $\sigma(\cdot) \in L^2_{loc}, \xi(\cdot) \in L^2_{loc}$  there exists a sequence  $t_k \rightarrow +\infty$ , possibly dependent on the functions  $\sigma(\cdot)$  and  $\xi(\cdot)$ , such that

$$(2.4) \quad \int_0^{t_k} \mathcal{F}_j(\sigma(t), \xi(t), \sigma(t-\tau), \xi(t-\tau))dt \geq 0 \quad \forall \tau \in \mathcal{T}, j = 1, 2, \dots, N.$$

Here  $\mathcal{T}$  is a countable subset of nonnegative real numbers and  $\mathcal{F}_j(\sigma_1, \xi_1, \sigma_2, \xi_2)$  are given quadratic forms. If they depend on neither  $\sigma_2$  nor  $\xi_2$ , then condition (2.4) reduces to a set of ordinary integral-quadratic constraints in the time domain [19].

Let us define the following sets:

$$\mathcal{Z} = \{z(\cdot) = [\sigma(\cdot), \xi(\cdot)] : \sigma(t) \in \mathbb{R}^p, \xi(t) \in \mathbb{R}^m, |z(\cdot)| \in L^2_{loc}[0, +\infty]\},$$

$$\mathcal{Z}_s = \{z(\cdot) \in \mathcal{Z} : |z(\cdot)| \in L^2[0, +\infty]\},$$

$$\mathcal{L} = \{z(\cdot) \in \mathcal{Z} : \text{equation (1.1) holds}\},$$

$$\mathcal{L}_s = \mathcal{L} \cap \mathcal{Z}_s,$$

$$\mathcal{M} = \{z(\cdot) \in \mathcal{Z} : \text{equation (2.4) holds for some } t_k \rightarrow +\infty\}.$$

The elements of the set  $\mathcal{Z}$  are called processes, and the elements of the set  $\mathcal{Z}_s$  are called stable processes. We also need the set  $\mathcal{M}^\infty \subset \mathcal{Z}_s$  of stable processes satisfying

$$(2.5) \quad \int_0^\infty \mathcal{F}_j(\sigma(t), \xi(t), \sigma(t-\tau), \xi(t-\tau))dt \geq 0 \quad \forall \tau \in \mathcal{T}, j = 1, 2, \dots, N.$$

The convergence of the integrals in (2.5) follows from the fact that the processes in the integrands are stable, i.e., globally square-integrable, and the integrands are quadratic forms.

The absolute stability, which is the main subject of this paper, is defined as follows.

**DEFINITION 2.2.** *The system defined by (1.1)–(2.3) is called absolutely stable if all processes  $z(\cdot) \in \mathcal{L} \cap \mathcal{N}$  are stable and there exists a constant  $\lambda$ , same for all processes, such that for all processes  $\|z(\cdot)\|^2 \leq \lambda \|\alpha(\cdot)\|^2$ .*

Here  $\|\cdot\|$  denotes the usual Euclidean norm in  $L^2[0, +\infty]$ . Note that existence of solutions of system (1.1)–(2.3) implies that  $\mathcal{L} \cap \mathcal{N} \neq \emptyset$ .

Each of the quadratic forms  $\mathcal{F}_j$  is extended to a Hermitian form  $\mathcal{F}_j(\tilde{\sigma}_1, \tilde{\sigma}_2, \tilde{\xi}_1, \tilde{\xi}_2)$ , where  $\tilde{\sigma}_1, \tilde{\sigma}_2 \in \mathbf{C}^p$  and  $\tilde{\xi}_1, \tilde{\xi}_2 \in \mathbf{C}^m$ . The Hermitian matrix  $\Pi_j(i\omega, \tau)$  is defined for  $\tau \in \mathcal{T}$  by the equation

$$(2.6) \quad \tilde{\xi}^* \Pi_j(i\omega, \tau) \tilde{\xi} = \mathcal{F}_j(-W(i\omega)\tilde{\xi}, \tilde{\xi}, -W(i\omega)\tilde{\xi}e^{-i\omega\tau}, \tilde{\xi}e^{-i\omega\tau}).$$

Here  $W(i\omega)$  is the frequency response of the linear block defined by (1.4).

The frequency condition (FC) is stated as follows: There exists a collection of  $N$  sequences  $\{\theta_n\}$  such that for some  $\varepsilon > 0$  we have

$$(2.7) \quad \sum_{j=1}^N \sum_{n=1}^{\infty} \Pi_j(i\omega, \tau_n) \theta_{nj} \leq -\varepsilon I_m, \quad \tau_n \in \mathcal{T}.$$

Here  $\theta_{nj}$  is the  $n$ th term of the  $j$ th sequence.

The FC in the paper [5] is stated differently: The Stieltjes integral is used instead of the infinite series. The reason is that in [5] the delay  $\tau$  can take any value. In this paper we need only the discrete set of values  $\tau_n = nT$ . The FC is adapted accordingly for our purposes.

We also need the notion of minimal stability. In order to define it we first need to introduce the stable continuations of processes.

**DEFINITION 2.3.** *A stable continuation of a process  $z(\cdot)$  in  $\mathcal{M}^\infty$  is a sequence of processes  $z_k(\cdot) \in \mathcal{L}_s \cap \mathcal{M}^\infty$  such that  $z_k(\cdot) = z(\cdot)$  for  $0 \leq t \leq t_k$  with  $t_k \rightarrow \infty$ .*

**DEFINITION 2.4.** *The system defined by (1.1), (2.3) is called minimally stable if every process  $z(\cdot) \in \mathcal{L} \cap \mathcal{N}$  has a stable continuation in  $\mathcal{M}^\infty$ .*

We are now in position to state the quadratic criterion for absolute stability, on which all the results of the paper will ultimately be based. It is an immediate consequence of Theorem 1 in [5].

**LEMMA 2.5.** *Suppose that the FC (2.7) is satisfied and system (1.1)–(2.3) is minimally stable. Then this system is absolutely stable.*

This lemma asserts that the minimal stability of a system is an essential ingredient of the absolute stability. Usually it is easy to verify by explicitly constructing the stable continuation required by Definition 2.4. To this end, it is often sufficient to construct a bounded continuation of a process, a concept which we now define.

**DEFINITION 2.6.** *A bounded continuation of a process  $z(\cdot)$  in  $\mathcal{M}$  is a sequence of processes  $z_k(\cdot) \in \mathcal{L} \cap \mathcal{M}$  such that  $|z_k(\cdot)| \in L^\infty[0, \infty]$  and  $z_k(\cdot) = z(\cdot)$  for  $0 \leq t \leq t_k$  with  $t_k \rightarrow \infty$ .*

The following lemma will be used throughout the paper to prove absolute stability of various systems under consideration. It is an immediate consequence of Lemma 2.5 and [5, Lemma 2].

**LEMMA 2.7.** *Suppose that the linear block (1.1) satisfies the regularity conditions (Definition 2.1) and the FC (2.7) holds. Then any bounded continuation of any process*



$z(\cdot) \in \mathcal{L}$  in  $\mathcal{M}$  is a stable continuation of  $z(\cdot)$  in  $\mathcal{M}^\infty$ . Furthermore, if every process  $z(\cdot) \in \mathcal{L} \cap \mathcal{N}$  has a bounded continuation in  $\mathcal{M}$ , then the system is absolutely stable.

The reader is referred to [5] for proofs and further details.

Lemma 2.7 asserts that the absolute stability of a system can be proved by showing that every process has a bounded continuation and that the appropriate FC is satisfied. The FC, in turn, will be derived from the constraints, which will naturally follow from the integral inequalities proved in the next subsection.

**2.2. Integral inequalities.** In this subsection we are going to prove two integral inequalities used later to derive the constraints needed for application of the quadratic criterion.

First, for a given function  $g(x, t)$ , continuous in each argument, we define

$$(2.8) \quad G(x, t) = \int_0^x g(x, t) dx.$$

The following lemma will be used in later sections. It extends the result of [18] to functions of two variables.

**LEMMA 2.8.** *Suppose that a function  $g(x, t)$ , continuous in each argument, is periodic in  $t$  with a period  $T$  and there exists a function  $H(u, v)$ , such that the following inequality holds for all  $x, y$ , and  $t$ :*

$$(2.9) \quad G(y, t) - G(x, t) + (x - y)g(x, t) \geq -H(x, g(x, t)) - H(y, g(y, t)).$$

*Suppose further that  $G(x, t) \geq 0$  for all  $x$  and  $t$ .*

*Then for any real number  $a$  and any measurable function  $x(t)$ , such that  $x(t) \equiv 0$  for  $t < 0$ , the following inequality holds:*

$$(2.10) \quad \int_0^a \{g(x(t), t)[x(t) - x(t - T)] + H(x(t), g(x(t), t))\} dt \\ + \int_0^a H(x(t - T), g(x(t - T), t)) dt \geq 0.$$

*If, in addition, the function  $g(x, t)$  is odd in  $x$ , the following inequality holds for any real number  $a$  and any measurable function  $x(t)$ , such that  $x(t) \equiv 0$  for  $t < 0$ :*

$$(2.11) \quad \int_0^a \{g(x(t), t)[x(t) + x(t - T)] + H(x(t), g(x(t), t))\} dt \\ + \int_0^a H(x(t - T), g(x(t - T), t)) dt \geq 0$$

*Proof.* Since the function  $g(x, t)$  is periodic in  $t$  with the period  $T$ , we have  $G(y, t - T) = G(y, t)$ . Using this we can replace inequality (2.9) with

$$(2.12) \quad G(y, t - T) - G(x, t) + (x - y)g(x, t) + H(x, g(x, t)) + H(y, g(y, t)) \geq 0.$$

Now we prove that inequality (2.12) implies inequality (2.10). In order to do that, we add to and subtract from the left-hand side of (2.10) the expression

$$\int_0^a [G(x(t - T), t - T) - G(x(t), t)] dt.$$

This yields

$$\begin{aligned}
 & \int_0^a \{g(x(t), t)[x(t) - x(t - T)] + H(x(t), g(x(t), t)) + H(x(t - T), g(x(t - T), t))\} dt \\
 &= \int_0^a \{g(x(t), t)[x(t) - x(t - T)] + H(x(t), g(x(t), t)) + H(x(t - T), g(x(t - T), t))\} dt \\
 & \quad + \int_0^a [G(x(t - T), t - T) - G(x(t), t)] dt \\
 & \quad + \int_0^a [G(x(t), t) - G(x(t - T), t - T)] dt \\
 &= \int_0^a \{G(x(t - T), t - T) - G(x(t), t) + g(x(t), t)[x(t) - x(t - T)]\} dt \\
 & \quad + \int_0^a [H(x(t), g(x(t), t)) + H(x(t - T), g(x(t - T), t))] dt \\
 & \quad + \int_{a-T}^a G(x(t), t) dt.
 \end{aligned}$$

The last integral is nonnegative because of the assumption that  $G(x, t) \geq 0$  for all  $x$  and  $t$ . By setting  $x = x(t)$ ,  $y = x(t - T)$ , and applying inequality (2.12) we conclude that the sum of the first two integrals is also nonnegative. Hence, the entire expression is nonnegative. This completes the proof of inequality (2.10).

Suppose now that the function  $g(x, t)$  is odd in  $x$ . Then the function  $G(x, t)$  is even in  $x$ , and we can replace  $G(y, t)$  with  $G(-y, t)$  in inequality (2.9). Inequality (2.11) can now be proved by repeating the above arguments, which completes the proof of the lemma.  $\square$

It is worth mentioning that this lemma can be proved if the assumption of the periodicity of the function  $g(x, t)$  in  $t$  is replaced with a weaker condition  $G(y, t - T) \geq G(y, t)$  for some number  $T$ . This fact can be used to derive stability conditions for some other classes of time-dependent nonlinearities [3].

**3. SISO systems with slope restricted nonlinearities.** In this section we are going to consider SISO systems (i.e.,  $p = m = 1$ ) with the nonlinearity satisfying the slope restriction inequality:

$$(3.1) \quad 0 \leq \frac{\varphi(\sigma_2, t) - \varphi(\sigma_1, t)}{\sigma_2 - \sigma_1} \leq \mu < \infty, \quad \varphi(0, t) \equiv 0.$$

Systems in which the nonlinearity does not depend explicitly on time have been well studied. The most general result is due to Zames and Falb [24]. However, for systems in which the nonlinearity is not stationary, the best known result is the celebrated circle criterion.

The method of delay-integral-quadratic constraints makes it possible to prove a stability criterion, similar in form to the result established by Zames and Falb for stationary systems.

**3.1. Stability multipliers.** In this subsection we are going to state and prove two results for this type of system.

The first result concerns the systems that satisfy the sector condition and the assumptions stated in the introduction.

THEOREM 3.1. Assume the following:

- I. The linear block (1.1) satisfies the regularity condition (Definition 2.1).
- II. The function  $\varphi(\sigma, t)$ , continuous in each variable, satisfies condition (3.1) and  $\varphi(\sigma, t + T) = \varphi(\sigma, t)$ .
- III. There exists a series  $\sum_{n=-\infty}^{n=+\infty} \vartheta_n < 1$  with nonnegative terms such that for all real values of  $\omega$

$$(3.2) \quad \operatorname{Re} \left\{ \left[ \mu^{-1} + W(i\omega) \right] \left[ 1 - \sum_{n=-\infty}^{\infty} \vartheta_n e^{-i\omega n T} \right] \right\} \geq \varepsilon > 0.$$

Then system (1.1)–(1.2) is absolutely stable (Definition 2.2).

*Proof.* First, define the following quadratic form of the variables  $\xi_1$  and  $\sigma_1$ :

$$(3.3) \quad \mathcal{F}_1(\sigma_1, \xi_1) = \xi_1(\sigma_1 - \mu^{-1}\xi_1).$$

Condition (3.1) implies that

$$(3.4) \quad \mathcal{F}_1(\sigma_1(t), \xi_1(t)) \geq 0.$$

Define the quadratic forms

$$\begin{aligned} \mathcal{F}_2(\sigma_1, \xi_1, \sigma_2, \xi_2) &= (\xi_1 - \xi_2)(\sigma_1 - \mu^{-1}\xi_1), \\ \mathcal{F}_3(\sigma_1, \xi_1, \sigma_2, \xi_2) &= (\sigma_1 - \mu^{-1}\xi_1 - \sigma_2 + \mu^{-1}\xi_2)\xi_1. \end{aligned}$$

Condition (3.1) and Lemma 2.8 with  $H(u, v) \equiv 0$  together imply that the following inequality holds for any  $\tau = nT$  and any  $t_k > 0$ :

$$(3.5) \quad \int_0^{t_k} \mathcal{F}_j(\sigma(t), \xi(t), \sigma(t - \tau), \xi(t - \tau)) dt \geq 0, \quad j = 2, 3.$$

Therefore, the FC (2.7) takes the following form: There exist sequences  $\theta_{1n}$ ,  $\theta_{2n}$ , and  $\theta_{3n}$ , all with nonnegative terms, such that for all real values of  $\omega$

$$(3.6) \quad \operatorname{Re} \left\{ \left[ \mu^{-1} + W(i\omega) \right] \left[ \Theta - \sum_{n=0}^{\infty} \theta_{2n} e^{i\omega n T} - \sum_{n=0}^{\infty} \theta_{3n} e^{-i\omega n T} \right] \right\} \geq \varepsilon > 0.$$

Here  $\Theta = \sum_{n=1}^{\infty} (\theta_{1n} + \theta_{2n} + \theta_{3n})$ .

It is easy to see that this condition is equivalent to condition III if we set  $\vartheta_n = \theta_{2n}/\Theta$  for  $n > 0$  and  $\vartheta_n = \theta_{3n}/\Theta$  for  $n < 0$ .

In order to complete the proof and use Lemma 2.7 we must show that every process has a bounded continuation. Let  $z(\cdot) = [\sigma(\cdot), \xi(\cdot)]$  be an arbitrary process, such that  $\xi(t) = \varphi(\sigma(t), t)$ . Let  $m_k$  be a number such that  $|\sigma(t)| \leq m_k$  for almost all  $0 \leq t \leq t_k$ . Define the following functions:

$$(3.7) \quad \varphi_k(\sigma, t) = \begin{cases} \varphi(-m_k, t) & \text{if } \sigma < -m_k, \\ \varphi(\sigma, t) & \text{if } |\sigma| \leq m_k, \\ \varphi(m_k, t) & \text{if } \sigma > m_k. \end{cases}$$

Consider a process  $z_k[\sigma_k(\cdot), \xi_k(\cdot)]$ , for which  $\xi_k(t) = \varphi_k(\sigma_k(t), t)$ . This process satisfies the constraints (3.4) and (3.5) and is, therefore, a bounded continuation of the process  $z(\cdot) = [\sigma(\cdot), \xi(\cdot)]$ . The proof is complete.  $\square$

If, in addition, the nonlinearity is odd, it is possible to weaken the condition that all terms of the series  $\sum_{n=-\infty}^{+\infty} \vartheta_n$  must be nonnegative.

**THEOREM 3.2.** *Suppose that conditions I and II of Theorem 3.1 are met and, in addition, the function  $\varphi(\sigma, t)$  is odd in  $\sigma$ . Suppose further that there exists an absolutely convergent series  $\sum_{n=-\infty}^{+\infty} \vartheta_n < 1$  such that for all real values of  $\omega$  condition (3.6) holds.*

*Then system (1.1)–(1.2) is absolutely stable (Definition 2.2).*

*Proof.* The proof proceeds along the same steps as that of Theorem 3.1. We can define the same quadratic forms  $\mathcal{F}_1$ ,  $\mathcal{F}_2$ , and  $\mathcal{F}_3$ . Inequalities (3.4) and (3.5) hold.

Furthermore, we define the following two quadratic forms:

$$\begin{aligned}\mathcal{F}_4(\sigma_1, \xi_1, \sigma_2, \xi_2) &= (\xi_1 + \xi_2)(\sigma_1 - \mu^{-1}\xi_1), \\ \mathcal{F}_5(\sigma_1, \xi_1, \sigma_2, \xi_2) &= (\sigma_1 - \mu^{-1}\xi_1 + \sigma_2 - \mu^{-1}\xi_2)\xi_1.\end{aligned}$$

Since the function  $\varphi(\sigma, t)$  is odd in  $\sigma$ , Lemma 2.8 with  $H(u, v) \equiv 0$ , together with the condition (3.1), implies that inequality (3.5) holds for  $j = 4$  and  $j = 5$  for any  $\tau = nT$ .

Therefore, the FC takes the following form: There exist sequences  $\theta_{1n}, \theta_{2n}, \dots, \theta_{5n}$ , all with nonnegative terms, such that for all real values of  $\omega$

$$(3.8) \quad \operatorname{Re} \{ [\mu^{-1} + W(i\omega)] Z(i\omega) \} \geq \varepsilon > 0.$$

Here

$$\begin{aligned}Z(i\omega) &= \Theta - \sum_{n=0}^{\infty} \theta_{2n} e^{i\omega nT} - \sum_{n=0}^{\infty} \theta_{3n} e^{-i\omega nT} + \sum_{n=0}^{\infty} \theta_{4n} e^{i\omega nT} + \sum_{n=0}^{\infty} \theta_{5n} e^{-i\omega nT}, \\ \Theta &= \sum_{n=1}^{\infty} (\theta_{1n} + \theta_{2n} + \theta_{3n} + \theta_{4n} + \theta_{5n}).\end{aligned}$$

To show that this is equivalent to the FC of Theorem 3.2, define  $\vartheta_n = \theta_{2n} - \theta_{4n}/\Theta$  for  $n > 0$  and  $\vartheta_n = \theta_{3n} - \theta_{5n}/\Theta$  for  $n < 0$ .

Existence of the bounded continuations for every process is verified in exactly the same way as in the proof of Theorem 3.1. The functions, defined by (3.7), are odd in  $\sigma$  if the nonlinearity is odd in  $\sigma$ . Therefore, the system is absolutely stable by Lemma 2.7.  $\square$

**3.2. Geometric interpretation and numerical examples.** With a slight abuse of notation, the FC (3.2) in Theorems 3.1 and 3.2 can be rewritten in the form

$$(3.9) \quad [\mu^{-1} + \operatorname{Re} W(i\omega)] \operatorname{Re} Z(i\omega) - \operatorname{Im} W(i\omega) \operatorname{Im} Z(i\omega) > 0,$$

where

$$(3.10) \quad Z(i\omega) = 1 - \sum_{n=-\infty}^{\infty} \vartheta_n e^{-i\omega nT}$$

and the sequence  $\vartheta_n$  satisfies the conditions of either Theorem 3.1 or Theorem 3.2, depending on the context.

Let us define the following two functions:

$$\Phi(\omega) = \frac{\mu^{-1} + \operatorname{Re}W(i\omega)}{\operatorname{Im}W(i\omega)}, \quad \Psi(\omega) = \frac{\operatorname{Im}Z(i\omega)}{\operatorname{Re}Z(i\omega)}.$$

For the sake of simplicity let us assume that  $W(i\omega)$  is a proper rational function. Then it is easy to show [9] that the graph of the function  $\Phi(\omega)$  consists of branches with asymptotes. The ends of the branches point either to  $+\infty$  (called stalactites) or to  $-\infty$  (called stalagmites). The FC (3.2) holds if a function  $\Psi(\omega)$  can be found such that its graph separates the stalactites from the stalagmites. For the circle criterion,  $\Psi(\omega) \equiv 0$ , i.e., the stalactites must be separated from the stalagmites by the abscissa axis.

For the FC (3.2) we have

$$(3.11) \quad \Psi(\omega) = \frac{\sum_{n=-\infty}^{\infty} \vartheta_n \sin \omega n T}{\sum_{n=-\infty}^{\infty} \vartheta_n \cos \omega n T - 1}.$$

Furthermore,  $\sum_{n=-\infty}^{\infty} \vartheta_n < 1$  and there is an additional requirement that  $\vartheta_n \geq 0$  unless the nonlinearity is odd in  $\sigma$ .

It is often easier to use this approach if the infinite series in (3.11) are replaced with trigonometric polynomials as illustrated in the following numerical examples.

Consider a system with the linear block defined by the transfer function:

$$(3.12) \quad W(s) = 0.04 + \frac{s^2}{[(s + 0.5)^2 + 0.81][(s + 0.5)^2 + 1.21]}.$$

Let  $\mu = 20$ . For the function  $\Psi(\omega)$  we choose  $\vartheta_1 = 0.1$ ,  $\vartheta_2 = 0.5$ , and  $\vartheta_3 = 0.2$ . Figure 3.1 shows the plot for  $T = 0.25\pi$ , allowing us to conclude that the system is absolutely stable. The same is true for the case with  $T = 0.3\pi$  (Figure 3.2) and all the intermediate values of  $T$ .

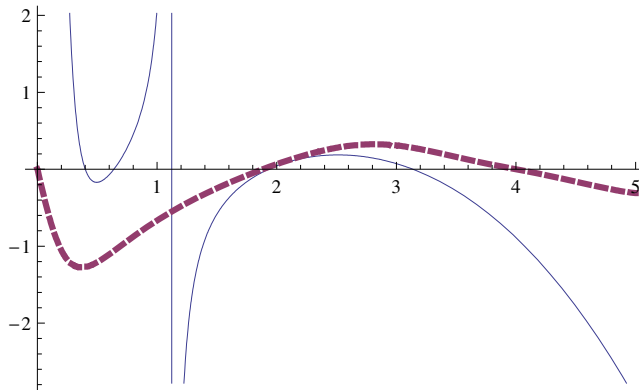


FIG. 3.1. Functions  $\Phi(\omega)$  (solid line) and  $\Psi(\omega)$  (broken line) for  $T = 0.25\pi$ .

Now consider the case with an odd nonlinearity. This means we may use negative values for the coefficients in (3.11). Set  $\vartheta_1 = -0.25$ ,  $\vartheta_2 = 0.5$ , and  $\vartheta_3 = 0.2$ . Figure 3.3 shows the plot for the case of  $T = 0.21\pi$ . Hence, the system is absolutely stable for this value of the period. The same is, once again, true for the case with  $T = 0.3\pi$  (Figure 3.4) and all the intermediate values of  $T$ . Notice how including the additional

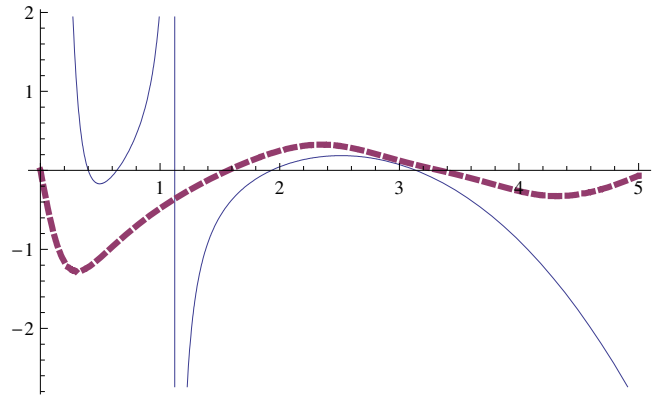


FIG. 3.2. Functions  $\Phi(\omega)$  (solid line) and  $\Psi(\omega)$  (broken line) for  $T = 0.3\pi$ .

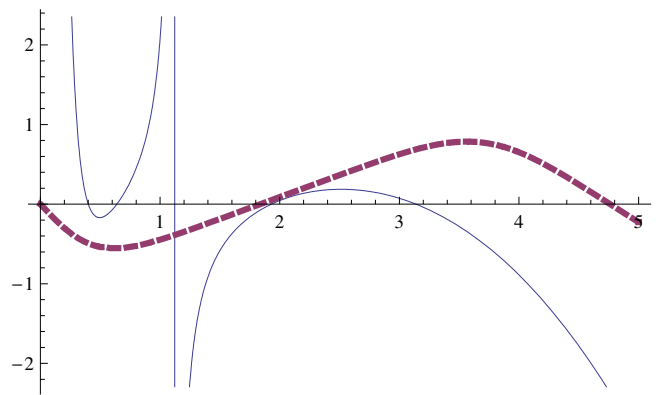


FIG. 3.3. Functions  $\Phi(\omega)$  (solid line) and  $\Psi(\omega)$  (broken line) for  $T = 0.21\pi$  with odd nonlinearity.

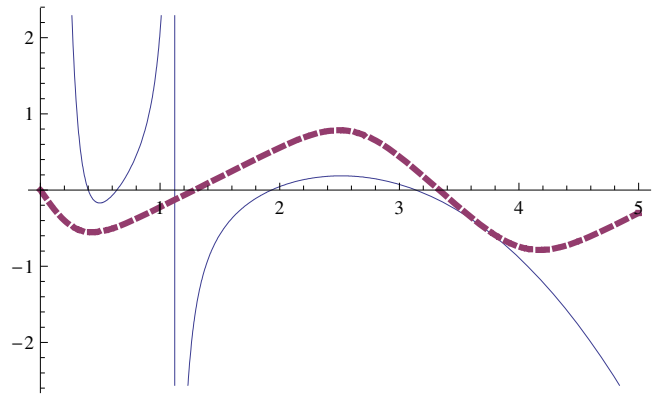


FIG. 3.4. Functions  $\Phi(\omega)$  (solid line) and  $\Psi(\omega)$  (broken line) for  $T = 0.3\pi$  with odd nonlinearity.

requirement that the nonlinearity be odd (and thus relaxing the requirement that the coefficients in (3.11) must be nonnegative) allowed us to widen the range of periods, for which absolute stability of the system can be proved using this criterion.

It is worth noting that since the branches of the curve  $\Phi(\omega)$  intersect the abscissa axis, the circle criterion is not applicable.

**4. SISO systems with quasimonotone sector nonlinearities.** In this section we are going to replace the slope restriction inequality (3.1) with a weaker sector condition:

$$(4.1) \quad 0 \leq \frac{\varphi(\sigma, t)}{\sigma} \leq \mu < \infty, \quad \varphi(0, t) \equiv 0.$$

The requirement that the nonlinearity be nondecreasing in  $\sigma$  is relaxed to a weaker condition that it be quasimonotone. This concept is introduced by the following definition, paraphrased from [6].

**DEFINITION 4.1.** *A function  $g(x)$  is called quasimonotone if there exists a positive-semidefinite quadratic form  $B(r, w)$  such that for all  $x$  and  $y$*

$$(4.2) \quad G(y) - G(x) + (x - y)g(x) \geq -B(x, g(x)) - B(y, g(y)),$$

where

$$G(x) = \int_0^x g(x) dx.$$

The form  $B(r, w)$  is called a defining form of the quasimonotone function  $g(x)$ .

Clearly, if  $B(r, w) \equiv 0$ , condition (4.2) reduces to the concavity condition for the function  $G(x)$ , i.e., to the condition that the function  $g(x)$  is nondecreasing.

For this type of nonlinearity we have the following stability criterion.

**THEOREM 4.2.** *Assume the following:*

- I. *The linear block (1.1) satisfies the regularity conditions (Definition 2.1).*
- II. *The function  $\varphi(\sigma, t)$  is continuous in each argument, satisfies the sector condition (4.1), is quasimonotone in  $\sigma$  with a defining form  $B(r, w)$ , and  $\varphi(\sigma, t + T) = \varphi(\sigma, t)$ .*
- III. *There exists a series  $\sum_{n=-\infty}^{\infty} \vartheta_n < 1$  with nonnegative terms such that for all real values of  $\omega$*

$$(4.3) \quad \operatorname{Re} \left\{ [\mu^{-1} + W(i\omega)] \left[ 1 - \sum_{n=-\infty}^{n=+\infty} \vartheta_n e^{-i\omega n T} \right] - 2B(W(i\omega), 1) \right\} \geq \varepsilon > 0.$$

*Then system (1.1)–(1.2) is absolutely stable (Definition 2.2).*

*Proof.* The proof of this theorem proceeds along the same steps as the results in the previous section. First, define the following quadratic form of the variables  $\xi$  and  $\sigma$ :

$$(4.4) \quad \mathcal{F}_1(\sigma_1, \xi_1) = \xi_1(\sigma_1 - \mu^{-1}\xi_1).$$

Condition (4.1) implies that

$$(4.5) \quad \mathcal{F}_1(\sigma_1(t), \xi_1(t)) \geq 0.$$

Next, define the following two quadratic forms:

$$\mathcal{F}_2(\sigma_1, \xi_1, \sigma_2, \xi_2) = (\xi_1 - \xi_2)(\sigma_1 - \mu^{-1}\xi_1) + B(\sigma_1, \xi_1) + B(\sigma_2, \xi_2),$$

$$\mathcal{F}_3(\sigma_1, \xi_1, \sigma_2, \xi_2) = (\sigma_1 - \mu^{-1}\xi_1 - \sigma_2 + \mu^{-1}\xi_2)\xi_1 + B(\sigma_1, \xi_1) + B(\sigma_2, \xi_2).$$

In order to apply Lemma 2.8 we set  $x(t) = \mu\sigma(t) + \varphi(\sigma(t), t)$  and  $H(u, v) = B(\sigma_0(u, t), v)$ , where  $\sigma_0(u, t)$  is a root of the equation  $u = \mu\sigma_0 + \varphi(\sigma_0, t)$ . We obtain for any  $\tau = nT$  and any  $t_k > 0$

$$(4.6) \quad \int_0^{t_k} \mathcal{F}_j(\sigma(t), \xi(t), \sigma(t - \tau), \xi(t - \tau)) dt \geq 0, \quad j = 2, 3.$$

Therefore, the FC (2.7) takes the following form: There exist sequences  $\theta_{1n}$ ,  $\theta_{2n}$ , and  $\theta_{3n}$ , all with nonnegative terms, such that for all real values of  $\omega$

$$(4.7) \quad \operatorname{Re} \{ [\mu^{-1} + W(i\omega)] Z(i\omega) - 2B(W(i\omega), 1) \} \geq \varepsilon > 0.$$

Here

$$Z(i\omega) = \Theta - \sum_{n=0}^{\infty} \theta_{2n} e^{i\omega nT} - \sum_{n=0}^{\infty} \theta_{3n} e^{-i\omega nT},$$

$$\Theta = \sum_{n=1}^{\infty} (\theta_{1n} + \theta_{2n} + \theta_{3n}).$$

It is easy to see that this condition is equivalent to condition III if we set  $\vartheta_n = \theta_{2n}/\Theta$  for  $n > 0$  and  $\vartheta_n = \theta_{3n}/\Theta$  for  $n < 0$ .

In order to show that every process has a bounded continuation so that we can use Lemma 2.7, we use the same method as in the proof of Theorem 3.1. Suppose that  $t_k \rightarrow \infty$  is an arbitrary sequence. Let  $m_k = \max_{t \in [0, t_k]} |\sigma(t)|$  and let  $s_k$  be a real number, such that for all  $t$  and all  $\sigma \in [0, m_k]$  the following inequality holds:

$$(4.8) \quad |\varphi(s_k, t)| \geq |\varphi(\sigma, t)|.$$

Define the functions  $\varphi_k(\sigma, t)$  as follows:

$$(4.9) \quad \varphi_k(\sigma, t) = \begin{cases} \varphi(-s_k, t) & \text{if } \sigma < -m_k, \\ \varphi(\sigma, t) & \text{if } |\sigma| \leq m_k, \\ \varphi(s_k, t) & \text{if } \sigma > m_k. \end{cases}$$

Note that these functions may have discontinuities of the first kind in the argument  $\sigma$ , while the nonlinearity  $\varphi(\sigma, t)$  is assumed to be continuous in each argument. However, this is not a concern since the functions  $\varphi_k(\sigma, t)$  can be approximated by continuous functions with a level of precision sufficient for the relevant inequalities to hold.

Now we need to prove that the functions  $\varphi_k(\sigma, t)$  are quasimonotone in  $\sigma$  with the same defining form  $B(r, w)$  as  $\varphi(\sigma, t)$ . This means proving that for all  $u$  and  $v$  the following inequality holds:

$$(4.10) \quad \int_u^v \varphi_k(\sigma, t) d\sigma + (u - v)\varphi_k(u, t) \geq -B(u, \varphi_k(u, t)) - B(v, \varphi_k(v, t)).$$

For the sake of certainty suppose that  $0 < u < v$ . The arguments in other cases are similar. If  $u < v \leq m_k$ , then  $\varphi_k(\sigma, t) = \varphi(\sigma, t)$  on the entire interval of integration, and inequality (4.10) clearly holds. If  $m_k < u < v$ , the left-hand side of inequality (4.10) vanishes, while the right-hand side is nonpositive, and, therefore, it holds.



Consider the case when  $u < m_k < v$ . We have

$$\begin{aligned}
 & \int_u^v \varphi_k(\sigma, t) d\sigma + (u - v)\varphi_k(u, t) \\
 &= \int_u^{m_k} \varphi_k(\sigma, t) d\sigma + \int_{m_k}^v \varphi_k(\sigma, t) d\sigma + (u - m_k)\varphi_k(u, t) + (m_k - v)\varphi_k(u, t) \\
 &= \int_u^{m_k} \varphi_k(\sigma, t) d\sigma + (u - m_k)\varphi_k(u, t) + \int_{m_k}^v \varphi_k(\sigma, t) d\sigma + (m_k - v)\varphi_k(u, t) \\
 &\geq -B(u, \varphi_k(u, t)) - B(m_k, \varphi_k(m_k, t)) + (v - m_k)[\varphi(s_k, t) - \varphi(u, t)] \\
 &> -B(u, \varphi_k(u, t)) - B(m_k, \varphi_k(m_k, t)) \\
 &> -B(u, \varphi_k(u, t)) - B(v, \varphi_k(v, t)).
 \end{aligned}$$

Having proved that the functions  $\varphi_k(\sigma, t)$  are quasimonotone in  $\sigma$  with the same defining form as  $\varphi(\sigma, t)$ , we can now conclude that every process has a bounded continuation, which completes the proof of this theorem.  $\square$

Just as in the case of slope-restricted nonlinearities, the result can be strengthened if an additional requirement that the nonlinearity be odd in  $\sigma$  is imposed.

**THEOREM 4.3.** *Suppose that conditions I and II of Theorem 4.2 are met and, in addition, the function  $\varphi(\sigma, t)$  is odd in  $\sigma$ . Suppose further that there exists an absolutely convergent series  $\sum_{n=-\infty}^{\infty} \theta_n < 1$  such that for all real values of  $\omega$  condition (4.3) holds.*

*Then system (1.1)–(1.2) is absolutely stable (Definition 2.2).*

*Proof.* The proof proceeds along the same steps as the other stability criteria. We can define the same quadratic forms  $\mathcal{F}_1$ ,  $\mathcal{F}_2$ , and  $\mathcal{F}_3$  as in the proof of Theorem 4.2. Inequalities (4.5) and (4.6) hold.

Further, we define the following two quadratic forms:

$$\begin{aligned}
 \mathcal{F}_4(\sigma_1, \xi_1, \sigma_2, \xi_2) &= (\xi_1 + \xi_2)(\sigma_1 - \mu^{-1}\xi_1) + B(\sigma_1, \xi_1) + B(\sigma_2, \xi_2), \\
 \mathcal{F}_5(\sigma_1, \xi_1, \sigma_2, \xi_2) &= (\sigma_1 - \mu^{-1}\xi_1 + \sigma_2 - \mu^{-1}\xi_2)\xi_1 + B(\sigma_1, \xi_1) + B(\sigma_2, \xi_2).
 \end{aligned}$$

Since the function  $\varphi(\sigma, t)$  is odd in  $\sigma$ , Lemma 2.8 implies that inequality (4.6) holds for  $j = 4$  and  $j = 5$  for any  $\tau = nT$ .

Therefore, the FC takes the following form: There exist sequences  $\theta_{1n}, \theta_{2n}, \dots, \theta_{5n}$ , all with nonnegative terms, such that for all real values of  $\omega$  inequality (4.7) holds with

$$\begin{aligned}
 Z(i\omega) &= \Theta - \sum_{n=0}^{\infty} \theta_{2n} e^{i\omega nT} - \sum_{n=0}^{\infty} \theta_{3n} e^{-i\omega nT} + \sum_{n=0}^{\infty} \theta_{4n} e^{i\omega nT} + \sum_{n=0}^{\infty} \theta_{5n} e^{-i\omega nT}, \\
 \Theta &= \sum_{n=1}^{\infty} (\theta_{1n} + \theta_{2n} + \theta_{3n} + \theta_{4n} + \theta_{5n}).
 \end{aligned}$$

To show that this is equivalent to the frequency condition of the theorem, define, just as before,  $\vartheta_n = \theta_{2n} - \theta_{4n}/\Theta$  for  $n > 0$  and  $\vartheta_n = \theta_{3n} - \theta_{5n}/\Theta$  for  $n < 0$ .

In order to exhibit bounded continuation for every process, we use the same functions  $\varphi_k(\sigma, t)$  defined by (4.9). They are odd in  $\sigma$ , and the proof can be completed by the same argument as that of Theorem 3.2.  $\square$

It is important to note here that the frequency conditions in these two theorems do not have the multiplier form. Therefore, the geometric interpretation from subsection 3.2 cannot be used directly. Numerical implementation of this result should be considered an open problem.

**5. Linear periodic MIMO systems.** In this section we turn our attention to linear MIMO systems. The function  $\varphi(\sigma, t)$  is assumed to have the form

$$(5.1) \quad \varphi(\sigma, t) = P(t)\sigma(t).$$

Here  $P(t)$  is a nonsingular  $m \times m$  matrix,  $P(t+T) = P(t)$ . Furthermore, it will be assumed that this matrix satisfies the MIMO analogue of the sector condition; i.e., for all  $t$  the following inequality holds:

$$(5.2) \quad \text{Sym} P^{-1}(t) - \mu^{-1} I_m \geq 0.$$

Here, for a real matrix  $X$ ,  $\text{Sym} X = (X + X^*)/2$ .

We also impose the following requirement: There exist constant matrices  $Q$  and  $S$ , such that for all values of  $t$

$$(5.3) \quad P^*(t)Q - SP(t) \equiv 0.$$

Clearly, (5.3) is fulfilled by setting  $P = Q = 0$ , which, as we shall see, yields the MIMO analogue of the circle criterion. Aside from this trivial possibility, this condition may at first glance appear rather restrictive. However, it is satisfied if  $P(t)$  is symmetric. It also holds if  $P(t) = GH(t)$ , where  $G$  is a constant nonsingular matrix and  $H(t)$  is either symmetric (set  $Q = S = G^{-1}$ ) or orthogonal (set  $Q = G$  and  $S = G^{-1}$ ).

The determination of existence of the required matrices  $G$  and  $H(t)$  can be made as follows. Recall that any matrix can be represented as a product of an orthogonal and a symmetric matrix. Thus we can always write  $P(t) = U(t)H(t)$ , where  $H^2(t) = P^*(t)P(t)$  and  $U(t) = P(t)H^{-1}(t)$ . If the matrix  $U(t)$  turns out to be constant, we set  $G = U(t)$ .

Let us illustrate this concept with a simple example. Let

$$P(t) = \begin{bmatrix} 1 & \sin t \\ \cos t & 1 \end{bmatrix}.$$

Performing the above computation, we find

$$U(t) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Therefore, in this case there exists the desired representation for the matrix  $P(t)$ . Another characterization of matrices satisfying (5.3) is given by the following statement.

**PROPOSITION 5.1.** *Suppose that there exists a constant matrix  $Q$ , such that*

$$(5.4) \quad [P'(t)P^{-1}(t)]^*Q \equiv QP^{-1}(t)P'(t).$$

*Then the matrix  $S = P^*(t)QP^{-1}(t)$  is constant.*

This proposition is proved by differentiating the matrix  $S$ . Condition (5.4) implies that its derivative is identically equal to zero.

Therefore, a determination of whether a given matrix  $P(t)$  satisfies the condition (5.3) can be made by attempting to solve (5.4). If a constant solution can be found, the question is answered affirmatively. The equation of this type is discussed in the book by Gantmacher [8].

Solving (5.4) for the matrix  $P(t)$  in the above example, we obtain the following two relationships for the components of the matrix  $Q$ :  $q_{11} = q_{22} \tan t$  and  $q_{12} = q_{21} - q_{22} \sec t$ . Set  $q_{11} = q_{22} = 0$  and  $q_{12} = q_{21} = 1$  to obtain the desired constant matrix  $Q$ .

The stability criterion for systems of this type is stated as follows.

THEOREM 5.2. *Assume the following:*

- I. *The linear block (1.1) satisfies the regularity conditions (Definition 2.1).*
- II. *The function  $\varphi(\sigma, t)$  is defined by  $\varphi(\sigma, t) = P(t)\sigma(t)$  with the nonsingular  $m \times m$  matrix  $P(t)$  satisfying the following conditions for all values of  $t$ :  $P(t+T) = P(t)$ ,  $\text{Sym}P^{-1}(t) - \mu^{-1}I_m \geq 0$ , and there exist constant matrices  $Q$  and  $S$ , such that for all values of  $t$  condition (5.4) is satisfied.*
- III. *There exists an odd periodic function  $q(\omega)$  with the period  $2\pi/T$  such that for all real values of  $\omega$  the following inequality holds:*

$$(5.5) \quad \text{Re}\{\mu^{-1}I_m + W(i\omega) + iq(\omega)[QW(i\omega) - W^*(i\omega)S]\} \geq \varepsilon I_m > 0.$$

*Then system (1.1)–(1.2) is absolutely stable (Definition 2.2).*

*Proof.* As before, define the quadratic form

$$(5.6) \quad \mathcal{F}_1(\sigma_1, \xi_1) = \xi_1^*(\sigma_1 - \mu^{-1}\xi_1).$$

Condition (5.2) implies that

$$(5.7) \quad \mathcal{F}_1(\sigma_1, \xi_1) \geq 0.$$

Next, define the quadratic forms

$$\begin{aligned} \mathcal{F}_2(\sigma_1, \xi_1, \sigma_2, \xi_2) &= \xi_1^*Q\sigma_2 - \sigma_2^*S\xi_2, \\ \mathcal{F}_3(\sigma_1, \xi_1, \sigma_2, \xi_2) &= \sigma_2^*S\xi_1 - \xi_2^*Q\sigma_1. \end{aligned}$$

Now we show that condition (5.3) implies

$$(5.8) \quad \mathcal{F}_j(\sigma(t), \xi(t), \sigma(t-nT), \xi(t-nT)) \equiv 0, \quad j = 2, 3.$$

Indeed, for  $j = 2$  we have

$$\begin{aligned} & \xi^*(t)Q\sigma(t-nT) - \sigma^*(t)S\xi(t-nT) \\ &= \sigma^*(t)P^*(t)Q\sigma(t-nT) - \sigma^*(t)SP(t-nT)\sigma(t-nT) \\ &= \sigma^*(t)[P^*(t)Q - SP(t-nT)]\sigma(t-nT) \\ &\equiv 0. \end{aligned}$$

The proof for  $j = 3$  is similar.

After some algebraic manipulations the FC takes the following form: There exists a sequence  $\theta_n$  such that for all real values of  $\omega$

$$(5.9) \quad \text{Re} \left\{ \mu^{-1}I_m + W(i\omega) + [QW(i\omega) - W^*(i\omega)S] \sum_{n=0}^{\infty} \theta_n \sin \omega nT \right\} \geq \varepsilon I_m > 0.$$

The series in this inequality is a Fourier series of an odd periodic function with period  $2\pi/T$ . Denoting this function by  $q(\omega)$ , we conclude that this FC is the same as condition III of the theorem.

Minimal stability can be verified by the same procedure as that in the paper [22]. By Lemma 2.5 the system is absolutely stable.  $\square$

It can be easily seen that if we set  $Q = S = 0$ , we obtain a MIMO analogue of the circle criterion.

If the matrix  $P(t)$  is symmetric (which also includes the case of  $m = 1$ ), we set  $Q = S = I_m$ . If, in addition, the matrix  $W(s)$  is Hermitian, then inequality (5.5) simplifies to

$$(5.10) \quad \operatorname{Re}\{\mu^{-1}I_m + W(i\omega)[1 + iq(\omega)]\} \geq \varepsilon I_m > 0.$$

This is a MIMO analogue of the result of Yakubovich [22].

**6. SISO systems with parametric class of nonlinearities.** In this section we are, again, going to consider the SISO systems satisfying the sector condition (4.1). In addition, it will be assumed that for all values of  $t$

$$(6.1) \quad \begin{aligned} |\sigma_1(t)\varphi(\sigma_2(t), t) - \sigma_2(t)\varphi(\sigma_1(t), t)| &\leq F(\sigma_1(t), \varphi(\sigma_1(t), t)) \\ &+ F(\sigma_2(t), \varphi(\sigma_2(t), t)). \end{aligned}$$

Here  $F(u, v)$  is a real quadratic form. Of particular interest is the case when  $F(u, v) = \gamma v(u - \mu^{-1}v)$ . Then if  $\gamma = 0$ , inequality (6.1) holds if and only if  $\varphi(\sigma, t)$  is a linear function in  $\sigma$ . For infinite value of  $\gamma$ , this inequality holds for all functions satisfying the sector condition (4.1). The stability criterion, however, can be proved for the general case.

**THEOREM 6.1.** *Assume the following:*

- I. *The linear block (1.1) satisfies the regularity conditions (Definition 2.1).*
- II. *The function  $\varphi(\sigma, t)$  is periodic in  $t$  with a period  $T$ , continuous in each argument, satisfies the sector condition (4.1) and inequality (6.1) for some real quadratic form  $F(u, v)$ .*
- III. *There exists an odd periodic function  $q(\omega)$  with a period  $2\pi/T$  such that for all real values of  $\omega$  the following inequality holds:*

$$(6.2) \quad \operatorname{Re}[\mu^{-1} + W(i\omega) - F(-W(i\omega), 1) + iq(\omega)W(i\omega)] \geq \varepsilon > 0.$$

*Then system (1.1)–(1.2) is absolutely stable (Definition 2.2).*

*Proof.* Define the quadratic form

$$(6.3) \quad \mathcal{F}_1(\sigma_1, \xi_1) = \xi_1(\sigma_1 - \mu^{-1}\xi_1).$$

Condition (4.1) implies that

$$(6.4) \quad \mathcal{F}_1(\sigma_1(t), \xi_1(t)) \geq 0.$$

Next, define the following two quadratic forms:

$$\begin{aligned} \mathcal{F}_2(\sigma_1, \xi_1, \sigma_2, \xi_2) &= F(\sigma_1, \xi_1) + F(\sigma_2, \xi_2) + \sigma_1\xi_2 - \sigma_2\xi_1, \\ \mathcal{F}_3(\sigma_1, \xi_1, \sigma_2, \xi_2) &= F(\sigma_1, \xi_1) + F(\sigma_2, \xi_2) + \sigma_2\xi_1 - \sigma_1\xi_2. \end{aligned}$$

Condition (6.1) implies

$$(6.5) \quad \mathcal{F}_j(\sigma(t), \xi(t), \sigma(t - nT), \xi(t - nT)) \equiv 0, \quad j = 2, 3.$$

The FC now takes the following form: There exists a sequence  $\theta_n$  such that for all real values of  $\omega$

$$\operatorname{Re} \left\{ [\mu^{-1} + W(i\omega) - F(-W(i\omega), 1)] \sum_{n=0}^{\infty} \theta_n + iW(i\omega) \sum_{n=0}^{\infty} \theta_n \sin \omega nT \right\} \geq \varepsilon > 0.$$

Just as in the previous section, this inequality includes a Fourier series of an odd periodic function with period  $2\pi/T$ . Denoting this function by  $q(\omega)$  we conclude that this FC is the same as condition III of the theorem.

Existence of a bounded continuation for every process is verified by introducing the same functions  $\varphi_m(\sigma, t)$  as in the proof of Theorem 3.1. It is easy to check that they satisfy both the sector condition and inequality (6.1). By Lemma 2.7 the system is absolutely stable.  $\square$

In the special case of  $F(u, v) = \gamma v(u - \mu^{-1}v)$  the FC (6.2) can, after some algebraic manipulations, be rewritten in the form

$$(6.6) \quad \operatorname{Re} \{ [\mu^{-1} + W(i\omega)][1 + \gamma + iq(\omega)] \} \geq \varepsilon > 0.$$

Several points are worth noting about this condition. First, it has a multiplier form, similar to the conditions in subsection 3.1, and is, therefore, amenable to the geometric interpretation and numerical procedure described in subsection 3.2.

If  $\gamma = 0$ , then the system reduces to the linear case described in section 5 with  $m = 1$ . The FC (6.6) then reduces to the form in Yakubovich's paper [22]. It is easy to see that if a multiplier exists for  $\gamma = 0$ , one can also be found for any  $\gamma > 0$ . On the other hand, as  $\gamma \rightarrow \infty$ , this inequality tends pointwise to the circle criterion, as the class of nonlinearities satisfying condition (6.1) expands to include all the functions satisfying the sector condition (3.1).

**7. Conclusions.** We have considered several types of systems with time-periodic feedback and, using the method of delay-integral-quadratic constraints, proved absolute stability criteria for each of them. Analogous results have previously been known for stationary systems.

It is of interest to note that some of the results have a multiplier form, similar to the corresponding results for stationary systems. The difference lies in the fact that for stationary systems the multipliers can be represented in the form of Fourier–Stieltjes integrals. For time-periodic systems these integrals are replaced with Fourier series.

This poses a significant difficulty in numerical implementation of the results. For stationary systems there usually exists a matrix realization of the multipliers. This makes it possible to express the frequency-domain conditions as linear matrix inequalities [7], which can be solved numerically. For time-periodic systems and the resulting multipliers such matrix realization may not exist. Therefore, numerical interpretation of these results is presently an open problem.

One possible approach is based on the fact that, just as in case of stationary systems, the multipliers can be interpreted geometrically. However, the interpretation differs in that the curve, separating stalactites from stalagmites, must be periodic in  $\omega$  with a period of  $2\pi/T$ . This is a significant restriction, compared to stationary systems. Nevertheless, numerical interpretation of the results is still possible as illustrated in subsection 3.2.

Future research may be directed at finding an algorithm for constructing multipliers, analogous to the one described for stationary systems by Safonov and Wyetznar [17], that would improve the applicability of the absolute stability criteria.

## REFERENCES

- [1] D. A. ALTSHULLER, *Zames-Falb multipliers for systems with time-periodic nonlinearities*, in Proceedings of the 2002 American Control Conference, Anchorage, AK, 2002, pp. 68–73.
- [2] D. A. ALTSHULLER, *A generalization of the frequency domain stability criterion to a wider class of systems*, in Proceedings of the 2002 IEEE 2002 Conference on Decision and Control, Las Vegas, NV, 2002, pp. 2335–2339.
- [3] D. A. ALTSHULLER, *Stability multipliers for systems with nonstationary nonlinearities*, Vestnik St. Petersburg State Univ., (2003), pp. 3–12 (in Russian).
- [4] D. A. ALTSHULLER, *Frequency-domain stability criteria for two classes of systems with time-periodic feedback*, in Proceedings of the 2004 Symposium on Systems, Structure, and Control, Oaxaca, Mexico, 2004, pp. 164–169.
- [5] D. A. ALTSHULLER, A. V. PROSKURNIKOV, AND V. A. YAKUBOVICH, *Frequency-domain criteria for dichotomy and absolute stability for integral equations with quadratic constraints involving delays*, Dokl. Math., 70 (2004), pp. 998–1002.
- [6] N. E. BARABANOV, *State space extension method in the theory of absolute stability*, IEEE Trans. Automat. Control, 45 (2000), pp. 2335–2339.
- [7] S. BOYD, L. EL GHAOU, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.
- [8] F. R. GANTMACHER, *Matrix Theory*, Vol. 1, Chelsea Publishing, Providence, RI, 2000.
- [9] A. V. LIPATOV, *Stability of continuous systems with one nonlinearity*, Dokl. Akad. Nauk SSSR, 267 (1981), pp. 1069–1072 (in Russian).
- [10] A. MEGRETSKI AND A. RANTZER, *System analysis via integral quadratic constraints*, IEEE Trans. Automat. Control, 42 (1997), pp. 818–830.
- [11] R. K. MILLER, *Nonlinear Volterra Integral Equations*, W.A. Benjamin, Menlo Park, CA, 1971.
- [12] A. P. MOLCHANOV AND E. S. PYATNITSKY, *Lyapunov functions defining the necessary and sufficient conditions for absolute stability of nonlinear control systems I*, in E. S. Pyatnitsky. Selected Scholarly Works, Vol. 1, Fizmatlit, Moscow, 2004, pp. 219–235 (in Russian).
- [13] A. P. MOLCHANOV, AND E. S. PYATNITSKY, *Lyapunov functions defining the necessary and sufficient conditions for absolute stability of nonlinear control systems II*, in E. S. Pyatnitsky. Selected Scholarly Works, Vol. 1, Fizmatlit, Moscow, 2004, pp. 236–251 (in Russian).
- [14] A. P. MOLCHANOV, AND E. S. PYATNITSKY, *Lyapunov functions defining the necessary and sufficient conditions for absolute stability of nonlinear control systems III*, in E. S. Pyatnitsky. Selected Scholarly Works, Vol. 1, Fizmatlit, Moscow, 2004, pp. 252–268 (in Russian).
- [15] K. S. NARENDRA AND J. H. TAYLOR, *Frequency Domain Criteria for Absolute Stability*, Academic Press, New York, 1973.
- [16] E. S. PYATNITSKY, *Absolute stability of nonstationary nonlinear systems*, in E. S. Pyatnitsky. Selected Scholarly Works, Vol. 1, Fizmatlit, Moscow, 2004, pp. 191–206 (in Russian).
- [17] M. G. SAFONOV AND G. WYETZNER, *Computer-aided analysis renders Popov criterion obsolete*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 1128–1131.
- [18] J. C. WILLEMS AND M. GRUBER, *Comments on "Combined frequency-time stability criterion for automatic control systems,"* IEEE Trans. Automat. Control, AC-12 (1967), pp. 217–219.
- [19] V. A. YAKUBOVICH, *Frequency conditions for stability of solutions of automatic control integral equations*, Vestnik Leningrad Univ., (1967), pp. 109–125 (in Russian).
- [20] V. A. YAKUBOVICH, *Quadratic criterion for absolute stability*, Dokl. Math., 58 (1998), pp. 169–171.
- [21] V. A. YAKUBOVICH, *Necessity in quadratic criterion for absolute stability*, Internat. J. Robust Nonlinear Control, 10 (2000), pp. 200–209.
- [22] V. A. YAKUBOVICH, *Popov's method and its subsequent development*, European J. Control, 8 (2002), pp. 889–907.
- [23] V. A. YAKUBOVICH AND V. M. STARZHINSKII, *Linear Differential Equations with Periodic Coefficients*, Vols. 1 and 2, Wiley, New York, 1975.
- [24] G. ZAMES AND P. L. FALB, *Stability conditions for systems with monotone and slope-restricted nonlinearities*, SIAM J. Control, 6 (1968), pp. 89–108.

## PARTIAL STABILITY FOR A CLASS OF NONLINEAR SYSTEMS\*

EDUARDO F. COSTA<sup>†</sup> AND ALESSANDRO ASTOLFI<sup>‡</sup>

**Abstract.** This paper studies a nonlinear, discrete-time matrix system arising in the stability analysis of Kalman filters. These systems present an internal coupling between the state components that gives rise to complex dynamic behavior. The problem of partial stability, which requires that a specific component of the state of the system converge exponentially, is studied and solved. The convergent state component is strongly linked with the behavior of Kalman filters, since it can be used to provide bounds for the error covariance matrix under uncertainties in the noise measurements. We exploit the special features of the system—mainly the connections with linear systems—to obtain an algebraic test for partial stability. Finally, motivated by applications in which polynomial divergence of the estimates is acceptable, we study and solve a partial semistability problem.

**Key words.** stability, nonlinear systems, matrix analysis, Kalman filter stability

**AMS subject classifications.** 93D99, 70K20, 93B99, 15A99, 93E11

**DOI.** 10.1137/070708421

**1. Introduction.** Partial stability (PS) refers to the class of problems that deal with stability of some state components with respect to (w.r.t.) those same components or w.r.t. all state components, or even w.r.t. some (nonfixed) components. It has been studied for linear and nonlinear systems from different perspectives [2, 6, 9, 10, 14], and vector Lyapunov functions constitute one of the first and most common tools in the literature, which can be traced back to the fifties [12]. PS arises naturally in many applications, e.g., in situations where some of the variables are not important as far as the operation and performance are concerned, or are not essential at all (in a sense related to model order reduction problems); see [14] for a quite complete assessment of specific problems in PS literature.

In this paper we deal with a discrete-time matrix system arising in the analysis of the error covariance matrix  $V_k$  of Kalman filters under noise measurement uncertainties; see [3]. Therein it is shown that, assuming the filter gains are calculated taking into account an initial error covariance  $\Sigma$  that differs from the actual one  $V_0$ , if the “incorrect” calculated error covariances are bounded, then for each  $0 \leq \zeta < 1$  there exist  $\tau > 0$  and  $M = M' \geq 0$ , providing the bounds

$$(1 - \tau)Z_k - \zeta^{-k}M \leq V_k \leq (1 + \tau)Z_k + \zeta^{-k}M, \quad k \geq 0,$$

for the actual error  $V_k$ , where  $Z_k$  is a component of the state  $(Z_k, X_k)$  of the system

---

\*Received by the editors November 16, 2007; accepted for publication August 28, 2008; published electronically January 7, 2009. This work was supported in part by FAPESP grants 06/02004-0 and 06/04210-6 and the EPSRC research grant EP/E057438, “Nonlinear Observation Theory with Applications to Markov Jump Systems.” A preliminary form of this paper appeared in Proceedings of the 17th IFAC World Congress 2008 [5].

<http://www.siam.org/journals/sicon/47-6/70842.html>

<sup>†</sup>Departamento de Matemática Aplicada e Estatística, Universidade de São Paulo, C.P. 668, 13560-970, São Carlos, SP, Brazil; currently visiting the Electrical and Electronic Engineering Department, Imperial College London, SW7 2AZ, London, UK (efcosta@icmc.usp.br).

<sup>‡</sup>Electrical and Electronic Engineering Department, Imperial College London, SW7 2AZ, London, UK, and Dipartimento di Informatica, Sistemi e Produzione, Università di Roma “Tor Vergata,” 00133 Roma, Italy (a.astolfi@imperial.ac.uk).

described by the equations

$$(1) \quad \Theta : \begin{cases} Z_{k+1} = H_k A Z_k A' H'_k, \\ X_{k+1} = A X_k A', \quad k \geq 0, \\ (Z_0, X_0) = (H_0 V_0 H'_0, \Sigma), \end{cases}$$

where  $Z_k$ ,  $X_k$ ,  $V_0$ , and  $\Sigma$  are symmetric positive semidefinite matrices and the square matrix  $A$  is assumed to be known.  $H_k$ ,  $k \geq 0$ , stands for the orthogonal projection onto the null space of  $X_k$ . For instance, when  $\ker\{\Sigma\} \subset \ker\{V_0\}$  (e.g., when  $\Sigma = V_0$  or  $\Sigma > 0$ ), we have  $H_0 V_0 H'_0 = 0$ , yielding  $Z_k = 0$ ,  $k \geq 0$ . We refer to  $H_k$  as the coupling projections, since it couples the dynamics of the  $Z$ -component with the trajectory of the  $X$ -component.

From the standpoint of the Kalman filter application, it is important to characterize, in terms of  $A$  and  $\Sigma$ , the existence, for each  $V_0$ , of  $0 \leq \beta < 1$  and  $\bar{Z}$  such that  $Z_k \leq \beta^k \bar{Z}$ ,  $k \geq 0$ , or, for each  $V_0$  and  $0 \leq \zeta < 1$ , of  $\bar{Z}$  such that  $\zeta^k Z_k \leq \bar{Z}$ , meaning that the  $Z$ -component cannot diverge exponentially. Moreover, since  $X_0 = \Sigma$  is fixed, we are interested only in the behavior of the  $Z$ -component w.r.t.  $V_0$ . We refer to these problems as PS and partial semistability<sup>1</sup> (PSS), respectively. PSS is relevant in situations where polynomial divergence of the Kalman estimates is acceptable.

Approaches for PS and PSS of general nonlinear systems can be employed, in principle. However, they are too general to yield an easy-to-test algebraic condition. The available results for PS of linear systems do not apply directly to  $\Theta$ , and it is worth mentioning that it is inappropriate to deal with the problem assuming that  $H_k$  are general projections, not connected with  $X$ , *in order to retrieve linearity*; in fact, in such a modified setting,  $Z_k$  can diverge exponentially, whereas  $A$  is stable<sup>2</sup> ( $A$  stable implies PS; see Remark 2). There is no available result specialized to system  $\Theta$ .

This paper exploits the special features of  $\Theta$ , for instance, the  $X$ -component obeys a linear difference equation and the coupling is via orthogonal projections; see other interesting properties in Proposition 1. We make use of a sequence of transformations  $W_k$ ,  $k \geq 0$ , playing the role of time-variant bases that allow one to characterize the convergence of the null space of  $X_k$ , as in Lemmas 6 and 7, and allow for adequate evaluations for the coupling projections in Lemma 9. Based on these evaluations, we show that  $\Theta$  is PSS if and only if the structural relation

$$(2) \quad \ker\{J\Sigma J^{-1}\} \cap \mathcal{J} = \{0\}$$

holds, where  $J$  is a similarity transformation such that  $JAJ^{-1}$  is in Jordan form and  $\mathcal{J}$  stands for the unstable subspace<sup>3</sup> of  $JAJ^{-1}$ . Recalling from linear systems theory that  $(A, \Sigma)$  semistabilizable can be interpreted as requiring that  $\Sigma$  excites the unstable space of  $A$ , the interpretation of (2) is that  $\Sigma$  has to “completely excite” the unstable space of  $A$ . Regarding PS, a similar condition holds, where  $\mathcal{J}$  is replaced with  $\mathcal{J}_S^\perp$  and  $\mathcal{J}_S$  is the stable space of  $A$ . These conditions are compared with classical notions of stabilizability and semistabilizability of  $(A, \Sigma)$ ; see Remark 2. Moreover, the conditions can be employed for “stabilization,” for instance, to obtain a  $\Sigma$  that provides PS or PSS; see Remark 3.

Apart from inherent theoretical significance, the derived conditions pave the way for obtaining sharp conditions for stability and semistability of Kalman filters [3, 4],

<sup>1</sup>Following the terminology of [1].

<sup>2</sup>For example, set  $A = \frac{3}{4} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$  and  $H_k = V_0 = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ ,  $k \geq 0$ .  $A$  is stable and  $Z_k$  diverge.

<sup>3</sup>Please see section 2 for definitions.



which is a highly important issue, since existing conditions are conservative as they are either necessary or sufficient, or rely on additional assumptions, such as the existence of limiting stationary filters; see [7, 11, 13, 15].

The paper is organized as follows. Section 2 presents definitions and preliminary results. Section 3 introduces the sequence of transformations that allow us to derive a simple structure for  $A$  and to simplify the evaluation of the projections  $H_k$ . These results allow us to obtain testable conditions for PS and PSS in section 4. Finally, section 5 provides some conclusions.

**2. Definitions and preliminary results.** Let  $\mathbb{R}^n$  denote the  $n$ th dimensional Euclidean space. Let  $\mathbb{D}$  (respectively,  $\bar{\mathbb{D}}$ ) be the open (closed) unit disk. Let  $e_i$ ,  $i = 1, \dots, n$ , be the canonical basis of  $\mathbb{R}^n$ .  $[v_1, \dots, v_m]$  stands for the vector space spanned by  $v_1, \dots, v_m \in \mathbb{R}^n$ . For vector subspaces  $\mathcal{E}$  and  $\mathcal{F}$ ,  $\mathcal{E} \perp \mathcal{F}$  means that  $\mathcal{E}$  and  $\mathcal{F}$  are orthogonal,  $\mathcal{E}^\perp$  is such that  $\mathcal{E}^\perp \perp \mathcal{E}$ ,  $\mathcal{E} \oplus \mathcal{F}$  is the direct sum of  $\mathcal{E}$  and  $\mathcal{F}$ , and  $\mathcal{E} \ominus \mathcal{F} = \mathcal{E} \cap \mathcal{F}^\perp$ . Let  $\mathcal{R}^{r,s}$  (respectively,  $\mathcal{R}^r$ ) represent the normed linear space formed by all  $r \times s$  real matrices (respectively,  $r \times r$ ) and  $\mathcal{R}^{r*}$  ( $\mathcal{R}^{r0}$ ) the cone  $\{U \in \mathcal{R}^r : U = U'\}$  (the closed convex cone  $\{U \in \mathcal{R}^r : U = U' \geq 0\}$ ), where  $U'$  denotes the transpose of  $U$ . For  $U \in \mathcal{R}^n$ ,  $\lambda_i(U)$ ,  $i = 1, \dots, n$ , stands for an eigenvalue of  $U$ .  $\lambda_i(U)$  is referred to as a semistable (respectively, stable) eigenvalue when it lies in  $\bar{\mathbb{D}}$  ( $\mathbb{D}$ ). The associated eigenvector  $v \in \mathbb{R}^n$  is semistable (stable); otherwise it is unstable. The space spanned by all stable eigenvectors is referred to as the stable subspace of  $U$ , and similarly for semistable and unstable semispaces.

Regarding the system  $\Theta$  and its state trajectory  $(Z_k, X_k)$ ,  $k \geq 0$ , we employ the notation  $Z_k(V_0)$  and  $X_k(\Sigma)$  to emphasize the dependence on  $V_0$  and  $\Sigma$ ; when the dependence of  $Z_k$  on  $\Sigma$  (indirectly via the coupling projections) is relevant, we employ the notation  $Z_k(V_0, \Sigma)$ . The coupling projections  $H_k$  give rise to the nonlinearities of  $\Theta$ ; for example, with  $A = I$  and  $V_0 \neq 0$  we have  $Z_k(V_0, I) + Z_k(V_0, 0) = V_0 \neq 0 = Z_k(2V_0, I)$ ,  $k \geq 0$ . Some useful features of system  $\Theta$  are presented in what follows.

**PROPOSITION 1.** *Consider system  $\Theta$  and, for  $U \in \mathcal{R}^{n0}$ , let  $U^*$  stand for the pseudoinverse of  $U$ . The following statements hold:*

- (i)  $H_k = I - X_k^* X_k = I - (\zeta X_k)^*(\zeta X_k)$ ,  $k \geq 0$ ,  $\zeta \neq 0$ .
- (ii)  $(Z_k(\zeta U), X_k(\zeta \Sigma)) = (\zeta Z_k(U), \zeta X_k(\Sigma))$ ,  $k \geq 0$ ,  $\zeta \geq 0$ .
- (iii) If  $U_1 \geq U_0$ , then  $Z_k(U_1) \geq Z_k(U_0)$ ,  $k \geq 0$ .
- (iv) If  $\Sigma_1 \geq \Sigma_0$ , then  $X_k(\Sigma_1) \geq X_k(\Sigma_0)$  and  $Z_k(V_0, \Sigma_1) \leq Z_k(V_0, \Sigma_0)$ ,  $k \geq 0$ .

The stability notions considered in this paper are as follows.

**DEFINITION 1.** *Consider system  $\Theta$ . We say that  $(A, \Sigma)$  is partially semistable (PSS) if, for each  $0 \leq \zeta < 1$  and  $V \in \mathcal{R}^{n0}$ , there exists  $\bar{Z} \in \mathcal{R}^{n0}$  such that  $\zeta^k Z_k(V) \leq \bar{Z}$ ,  $k \geq 0$ . We say that  $(A, \Sigma)$  is partially stable (PS) if, for each  $V \in \mathcal{R}^{n0}$ , there exists  $0 \leq \beta < 1$  and  $\bar{Z} \in \mathcal{R}^{n0}$  such that  $Z_k \leq \beta^k \bar{Z}$ ,  $k \geq 0$ .*

*Example 1.* Consider the system  $\Theta$  with

$$(3) \quad A = \begin{bmatrix} d & 1 \\ 0 & d \end{bmatrix}, \quad \Sigma = \sigma \sigma'.$$

Set  $d = -1$  and  $\sigma = [0 \ 0]'$ . From (1) we have that  $X_k = 0$ ,  $k \geq 0$ . Direct inspection of PSS and PS via Definition 1 is virtually impossible, as it involves exhaustive searches for  $\bar{Z}$ , for each  $V$  and  $\zeta$ . Moreover, there is no evidence on how to modify the parameters (e.g.,  $\sigma$ ) in order to achieve PSS or PS. In Example 5 we shall see that  $(A, \Sigma)$  is PSS (and not PS), despite the fact that  $Z_k$  can diverge with polynomial rate;

for instance, with  $V = vv'$  and  $v = e_2$ , (1) yields

$$Z_k(V) = A^k V A'^k = \begin{bmatrix} k^2 & -k \\ -k & 1 \end{bmatrix}.$$

LEMMA 1. *Consider system  $\Theta$ . The following statements hold:*

- (i)  *$(A, \Sigma)$  is PSS if and only if, for each  $\rho > 1$ , there exists  $\bar{Z} \in \mathcal{R}^{n_0}$  such that  $Z_k(I) \leq \rho^k \bar{Z}$ ,  $k \geq 0$ .*
- (ii)  *$(A, \Sigma)$  is PS if and only if there exist  $\bar{Z} \in \mathcal{R}^{n_0}$  and  $0 \leq \gamma < 1$  such that  $Z_k(I) \leq \gamma^k \bar{Z}$ ,  $k \geq 0$ .*

*Proof of (i).* (Necessity.) It follows by setting  $V = I$  in Definition 1.

(Sufficiency.) For each  $V \in \mathcal{R}^{n_0}$  we can pick  $\kappa > 0$  such that  $\kappa V \leq I$  and Proposition 1 (iii) yields  $Z_k(\kappa V) \leq Z_k(I) \leq \rho^k \bar{Z}$ , which leads to  $\rho^{-k} Z_k(V) \leq \kappa^{-1} \bar{Z}$ .

*Proof of (ii).* It is similar to the proof of (i) and is not presented.  $\square$

Consider now the linear time-varying system related to the dynamics of the  $Z$ -component of  $\Theta$ , defined by

$$(4) \quad \Theta_Z : \begin{cases} z_{k+1} = H_k A z_k, & k \geq 0, \\ z_0 = H_0 z, \end{cases}$$

where  $z_k \in \mathbb{R}^n$  is the state and  $z \in \mathbb{R}^n$ . Not surprisingly, PSS and PS of  $(A, \Sigma)$  are strongly connected to semistability and stability of  $\Theta_Z$ , as stated in the following lemma.

LEMMA 2. *Consider systems  $\Theta$  and  $\Theta_Z$ . The following statements hold:*

- (i)  *$(A, \Sigma)$  is PSS if and only if, for each  $z \in \mathbb{R}^n$  and  $0 \leq \zeta < 1$ , there exist  $\alpha \geq 0$  and  $0 \leq \beta < 1$  such that  $\|\zeta^k z_k\| \leq \alpha \beta^k$ .*
- (ii)  *$(A, \Sigma)$  is PS if and only if for each  $z \in \mathbb{R}^n$  there exist  $\alpha \geq 0$  and  $0 \leq \beta < 1$  such that  $\|z_k\| \leq \alpha \beta^k$ .*

*Proof of (i).* We employ the notation  $z_k(z)$ ,  $\alpha_z$ , and  $\beta_z$  to emphasize the dependence on  $z$ .

(Necessity.) Let  $\iota > 1$  and, for each  $0 \leq \zeta < 1$ , let  $\rho = \zeta^{-2}\iota$ . From Lemma 1, there exists  $\bar{Z}$  such that  $Z_k(I) \leq (\zeta^{-2}\iota)^k \bar{Z}$ . Equivalently,

$$(5) \quad \zeta^{2k} Z_k(I) \leq \iota^k \bar{Z}.$$

Consider  $z$  such that  $\|z\| \leq 1$ . Note that  $Z_0(I) = H_0 H'_0 \geq H_0 z z' H'_0 = z_0 z'_0$  and that employing (1) and (4) recursively yields  $Z_k \geq z_k z'_k$ ,  $k \geq 0$ . Then (5) leads to  $\zeta^{2k} z_k z'_k \leq \zeta^{2k} Z_k \leq \iota^k \bar{Z}$ , and taking the trace we obtain

$$(6) \quad \|\zeta^k z_k\|^2 \leq \iota^k \text{tr}(\bar{Z}), \quad \|z\| \leq 1.$$

Now consider  $\|z\| > 1$ . Note from (4) that  $z_k(\|z\|^{-1}z) = \|z\|^{-1}z_k(z)$ , which allows us to employ (6) to evaluate

$$(7) \quad \|\zeta^k z_k\|^2 = \|\zeta^k z_k(\|z\|^{-1}z)\|^2 \|z\|^2 \leq \iota^k \text{tr}(\bar{Z}) \|z\|^2.$$

From (6) and (7) we have that, for each  $z \in \mathbb{R}^n$ ,  $\|\zeta^k z_k\|^2 \leq \iota^k \text{tr}(\bar{Z}) \max(1, \|z\|^2)$ .

(Sufficiency.) For each  $\gamma > 1$ , let  $\zeta = \sqrt{\gamma^{-1}}$  and note that for each  $z$ , by hypothesis,  $z_k(z) z_k(z)' \leq \zeta^{-2k} \beta_z^{2k} \alpha_z^2 I = \gamma^k \beta_z^{2k} \alpha_z^2 I$ . Then, for each  $z = e_i$ ,  $i = 1, \dots, n$ , we can write

$$(8) \quad z_k(e_i) z'_k(e_i) \leq \gamma^k \beta_{e_i}^{2k} \alpha_{e_i}^2 I.$$

Since  $H_k$  is an orthogonal projection, employing (1) we obtain  $Z_0(I) = H_0 H'_0 \leq I = z_0(e_1)z'_0(e_1) + \cdots + z_0(e_n)z'_0(e_n)$ , and it is simple to check by induction that

$$(9) \quad Z_k(I) \leq z_k(e_1)z'_k(e_1) + \cdots + z_k(e_n)z'_k(e_n).$$

Equations (8) and (9) lead to

$$(10) \quad Z_k(I) \leq \gamma^k \beta_{e_i}^{2k} \alpha_{e_1}^2 I + \cdots + \gamma^k \beta_{e_i}^{2k} \alpha_{e_n}^2 I \leq \gamma^k \bar{\beta}^{2k} n \bar{\alpha}^2 I,$$

where  $\bar{\alpha} = \max(\alpha_{e_1}, \dots, \alpha_{e_n})$  and  $\bar{\beta} = \max(\beta_{e_1}, \dots, \beta_{e_n})$ . Since  $\beta_{e_i} < 1$  and  $\bar{\beta} < 1$ , (10) leads to  $Z_k(I) \leq \gamma^k (n \bar{\alpha}^2 I)$  and Lemma 1 completes the proof.

*Proof of (ii).* It is similar to the proof of (i), replacing  $\iota > 1$ ,  $0 \leq \zeta < 1$ , and  $\gamma > 1$  with  $0 \leq \iota < 1$ ,  $\zeta = 1$ , and  $\gamma = 1$ , respectively.  $\square$

Similarly to the sequence  $z_k$  connected with the  $Z$ -component of  $\Theta$ , we introduce a vector sequence related to  $X$ , as follows. Consider the solution  $X_k = A^k \Sigma A^{k'}$  for the  $X$ -component. Introduce the rank-one decomposition

$$(11) \quad \Sigma = \sigma_1 \sigma'_1 + \cdots + \sigma_{r_\Sigma} \sigma'_{r_\Sigma},$$

where  $r_\Sigma$  stands for the rank of  $\Sigma$ , and the linear system defined by

$$\Theta_X : x_k(\sigma) = A^k \sigma.$$

It is simple to check that

$$X_k = x_k(\sigma_1)x_k(\sigma_1)' + \cdots + x_k(\sigma_{r_\Sigma})x_k(\sigma_{r_\Sigma})'$$

and  $H_k$  is the orthogonal projection onto  $[x_k(\sigma_1), \dots, x_k(\sigma_{r_\Sigma})]^\perp$ .

**3. Evaluations for the coupling projections.** The spaces spanned by the trajectory  $x_k = A^k \sigma$  play an important role in this paper, because they drive the projection  $H_k$ . We now present certain characterizations for convergence of these spaces. Note that, taking into account the original basis, there may be no convergence for  $[x_k]$ ; see Examples 2 and 4. In this paper we employ the bases introduced as follows, related to Jordan forms [8], in view of the fact that they lead to a simpler characterization for  $[x_k]$ , and despite the drawback of an inherent time dependence.

**PROPOSITION 2.** *For each  $A \in \mathcal{R}^n$  there is a sequence of transformations  $W_k$ ,  $k \geq 0$ , such that  $A = W_{k+1}^{-1} \bar{A} W_k$  and  $A^k = W_k^{-1} \bar{A}^k W_0$ ,  $k \geq 0$ , with*

$$\bar{A} = \text{diag}(\mathcal{A}(\eta_1), \dots, \mathcal{A}(\eta_j)),$$

where  $\mathcal{A}(\eta_i)$ ,  $0 \leq i \leq j \leq n$ , is an upper triangular Jordan block with eigenvalue  $\eta_i$ , and  $\eta_i$  is a real nonnegative number, corresponding to certain eigenvalues  $\lambda_\ell(A)$ ,  $0 \leq \ell \leq n$ , with  $|\lambda_\ell(A)| = \eta_i$ , ordered in such a manner that  $\eta_i \geq \eta_j$  whenever  $i \geq j$ . Moreover, there exists  $\kappa$ ,  $0 \leq \kappa < 1$ , such that  $(1 - \kappa) \leq \|W_k\| \leq (1 + \kappa)$ ,  $k \geq 0$ .

The bases of Proposition 2 are employed throughout the paper, hence we introduce the following notation. Unless otherwise stated, for any  $V \in \mathcal{R}^{n,r}$  and  $v \in \mathbb{R}^n$ , we define  $\bar{V} \in \mathcal{R}^{n,r}$  and  $\bar{v} \in \mathbb{R}^n$  as  $\bar{V} = W_0 V$  and  $\bar{v} = W_0 v$ . For instance, we denote  $W_0 \sigma$  simply by  $\bar{\sigma}$ . The matrix  $A$  associated with the transformation  $W_0$  is usually clear from the context; otherwise we employ the explicit notation  $W_0(A)$ . For  $\sigma, z \in \mathbb{R}^n$ , define  $\bar{z}_k, \bar{x}_k \in \mathbb{R}^n$ ,  $k \geq 0$ , by

$$(12) \quad \bar{z}_{k+1} = (\bar{H}_k \bar{A}) \bar{z}_k, \quad k \geq 1, \quad \bar{z}_0 = \bar{H}_0 \bar{z}, \quad \bar{x}_k(\sigma) = \bar{A}^k \bar{\sigma}, \quad k \geq 0,$$

where

$$(13) \quad \bar{H}_k = W_k H_k W_k^{-1}.$$

*Example 2.* Consider

$$(14) \quad A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \bar{A} = I, W_{2\ell} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, W_{2\ell+1} = I, \ell \geq 0.$$

It is simple to check that the statements of Proposition 2 are satisfied. For  $\sigma = \begin{bmatrix} 1 & 1 \end{bmatrix}'$ ,  $x_k = A^k \sigma$  is in the form  $x_{2\ell} = \begin{bmatrix} 1 & 1 \end{bmatrix}'$  and  $x_{2\ell+1} = \begin{bmatrix} 1 & -1 \end{bmatrix}'$ ,  $\ell \geq 0$ ; hence the spaces spanned by  $[x_k]$  do not converge in any sense as  $k \rightarrow \infty$ . On the other hand,  $\bar{x}_k = \bar{A}^k \bar{\sigma} = W_0 \sigma = \begin{bmatrix} 1 & -1 \end{bmatrix}'$ ,  $k \geq 0$ . The matrix  $A$  is in Jordan form, making clear that the Jordan form is not convenient for the characterization of convergence of  $[x_k]$  that we seek. Note that  $\bar{A}$  is also in Jordan form, but is not similar to  $A$ .

Convergence of state trajectories is preserved, as stated in the next result.

LEMMA 3. *The following statements hold:*

- (i)  $z_k = W_{k-1}^{-1} \bar{z}_k$ ,  $x_k(\sigma) = W_k \bar{x}_k(\sigma)$ ,  $k \geq 0$ .
- (ii) *There exists  $\kappa$ ,  $0 \leq \kappa < 1$ , such that  $(1 + \kappa)^{-1} \|\bar{z}_k\| \leq \|z_k\| \leq (1 - \kappa)^{-1} \|\bar{z}_k\|$  and  $(1 - \kappa) \|\bar{x}_k(\sigma)\| \leq \|x_k(\sigma)\| \leq (1 + \kappa) \|\bar{x}_k(\sigma)\|$ ,  $k \geq 0$ .*

*Proof of (i).* We have from Proposition 2 that  $A = W_k^{-1} \bar{A} W_{k-1}$ ,  $k \geq 1$ ; substituting this equality and (13) in (4) yields

$$\begin{aligned} z_k &= (H_{k-1} A) \cdots (H_1 A) (H_0 v) = (W_{k-1}^{-1} \bar{H}_{k-1} W_{k-1} W_{k-1}^{-1} \bar{A} W_{k-2}) \cdots (W_0^{-1} \bar{H}_0 W_0 v) \\ &= W_{k-1}^{-1} (\bar{H}_{k-1} \bar{A}) \cdots (\bar{H}_1 \bar{A}) \bar{H}_0 (W_0 v) = W_{k-1}^{-1} \bar{z}_k. \end{aligned}$$

The proof for the second statement of (i) is analogous.

*Proof of (ii).* Accordingly to Proposition 2, there is a  $0 \leq \kappa < 1$  such that  $\|W_k\| \leq (1 + \kappa)$ ,  $k \geq 0$ , and assertion (i) allows us to write  $\|z_k\| \leq \|W_{k-1}^{-1}\| \|\bar{z}_k\| \leq (1 - \kappa)^{-1} \|\bar{z}_k\|$ . A similar evaluation yields  $\|z_k\| \geq (1 + \kappa)^{-1} \|\bar{z}_k\|$ . The proof for the second statement of (ii) is analogous.  $\square$

PROPOSITION 3. *Consider system  $\Theta_Z$  and the system  $\Theta_{\zeta Z}$  that arises by replacing the matrix  $A$  with  $(\zeta A)$ ,  $\zeta \geq 0$ , and let  $z_{\zeta,k}$  be the corresponding trajectory. Then  $\zeta^k z_k = z_{\zeta,k}$  and, similarly,  $\zeta^k \bar{z}_k = \bar{z}_{\zeta,k}$ .*

*Proof.* The first statement follows from Proposition 1 (i)–(ii). Moreover, the first statement and Lemma 3 (i) yield  $\zeta^k W_{k-1}^{-1}(A) \bar{z}_k = W_{k-1}^{-1}(\zeta A) \bar{z}_{\zeta,k}$ ; hence the second statement follows from the fact that one can always set  $W_k(\zeta A) = W_k(A)$  (even for  $\zeta = 0$ ).  $\square$

Consider now  $\bar{A}$  and let  $e_1, \dots, e_{q_u}$  and  $e_{q_u+1}, \dots, e_{q_s}$  be the associated unstable eigenvectors and semistable eigenvectors, respectively. Introduce the subspaces

$$(15) \quad \mathcal{U} = [e_1, \dots, e_{q_u}], \quad \mathcal{S} = [e_1, \dots, e_{q_s}].$$

Of course,  $\mathcal{U}^\perp = [e_{q_u+1}, \dots, e_n]$  and  $\mathcal{S}^\perp = [e_{q_s+1}, \dots, e_n]$ . The block structure of  $\bar{A}$  in Proposition 2 allows for the next invariance results.

LEMMA 4.  $\mathcal{U}, \mathcal{U}^\perp, \mathcal{S}$ , and  $\mathcal{S}^\perp$  are  $\bar{A}$ -invariant.

Note that  $\bar{A}$  is in Jordan form [8], leading to several links with available results for Jordan forms. For example, there are invariance results similar to the ones of Lemma 4; see, e.g., [1]. Another useful connection is as follows. Let  $J$  be the similarity matrix for which  $JAJ^{-1}$  is the Jordan form of  $A$  and let  $\mathcal{J}$  stand for the vector subspace spanned by the unstable eigenvectors of  $JAJ^{-1}$ . Then, for each  $\sigma \in \mathbb{R}^n$ , the

projection of  $J\sigma$  onto  $\mathcal{J}$  is zero if and only if the projection of  $\bar{\sigma} = W_0\sigma$  onto  $\mathcal{U}$  is zero, yielding the following result, given without proof, which is useful for representing the main results in terms of Jordan forms.

LEMMA 5.  $\ker\{W_0\Sigma W_0'\} \cap \mathcal{U} = \{0\}$  if and only if  $\ker\{J\Sigma J'\} \cap \mathcal{J} = \{0\}$ .

The spaces spanned by  $x_k$  may not converge in any sense (see Example 2), which implies that there may be no convergence for the projections  $H_k$ . However, the convenient structure of  $\bar{A}$  provides that  $[\bar{x}_k]$  always converge in a certain sense, allowing us to derive approximation results for  $\bar{H}$ . In order to make the convergence notion precise, we define, for the (nontrivial) vector subspaces  $\mathcal{U}$  and  $\mathcal{V}$ , the quantity

$$(16) \quad \theta_{\mathcal{V}}(\mathcal{U}) = \max_{v \in \mathcal{V}, v \neq 0} \min_{u \in \mathcal{U}, u \neq 0} 1 - (\|u\| \|v\|)^{-1} u'v.$$

Note from the structure of  $\bar{A}$  that if  $\sigma \in \mathbb{R}^n$  is such that  $\eta$  is the largest eigenvalue for which  $\sigma'\nu \neq 0$ , where  $\nu$  is an eigenvector associated with  $\eta$ , and assuming  $\eta$  unique (i.e., no other eigenvalue of  $\bar{A}$  equals  $\eta$ ), then there are  $\pi \geq 0$  and  $0 \leq \rho < 1$  such that  $\theta_{\mathcal{S}_\eta}([x_k(\sigma)]) \leq \pi\rho^k$ , where  $\mathcal{S}_\eta$  is the space spanned by the eigenvectors associated with  $\eta$ . This signifies that  $x_k(\sigma)$  and  $\mathcal{S}_\eta$  “align” with exponential rate. Moreover, there is a  $\varphi > 0$  such that  $\theta_{\mathbb{R}^n \ominus \mathcal{S}_\eta}([x_k(\sigma)]) \geq \varphi$  for a sufficiently large  $k$ . One can explore the convenient block structure of  $\bar{A}$  to obtain the more general characterization given in Lemmas 6 and 7 without proof; recall that  $\bar{\sigma} = W_0\sigma$  and  $\bar{x}_k(\sigma) = \bar{A}^k\bar{\sigma} = W_k A^k \sigma$ .

LEMMA 6. Consider  $\sigma_j \in \mathbb{R}^n$ ,  $j = 1, \dots, m$ , and assume  $\mathcal{S}$  is nontrivial. If  $\ker\{\bar{\sigma}_1\bar{\sigma}'_1 + \dots + \bar{\sigma}_m\bar{\sigma}'_m\} \cap \mathcal{S} = \{0\}$ , then there exist  $\pi \geq 0$  and  $0 \leq \rho < 1$  such that

$$\theta_{\mathcal{S}}([\bar{x}_k(\sigma_1), \dots, \bar{x}_k(\sigma_m)]) \leq \pi\rho^k.$$

Conversely to Lemma 6, if  $\sigma_j$  does not “completely excite” the subspace  $\mathcal{S}$ , then the space spanned by  $\bar{x}_k(\sigma_1)$  does not “align” with  $\mathcal{S}$ . It is convenient for later reference to formalize this in terms of  $\mathcal{U}$  rather than  $\mathcal{S}$ .

LEMMA 7. Consider  $\sigma_j \in \mathbb{R}^n$ ,  $\sigma_j \neq 0$ ,  $j = 1, \dots, m$ . If  $\ker\{\bar{\sigma}_1\bar{\sigma}'_1 + \dots + \bar{\sigma}_m\bar{\sigma}'_m\} \cap \mathcal{U} \neq \{0\}$ , then there exist  $\varphi > 0$  and a subspace  $\mathcal{F}$  spanned by eigenvectors of  $\bar{A}$  such that  $\mathcal{U} \ominus \mathcal{F}$  is nontrivial and

$$\begin{aligned} \theta_{\mathcal{F}}([\bar{x}_k(\sigma_1), \dots, \bar{x}_k(\sigma_m)]) &\leq o(k), \quad k \geq 0, \\ \theta_{\mathcal{U} \ominus \mathcal{F}}([\bar{x}_k(\sigma_1), \dots, \bar{x}_k(\sigma_m)]) &\geq \varphi, \quad k \geq 1, \end{aligned}$$

where  $o(\cdot)$  is a nonnegative-valued, strictly decreasing function.

Example 3. Consider the system  $\Theta$  of Example 1 with  $\sigma = e_2$ . The statements of Proposition 2 hold with  $\bar{A} = A$  and  $W_k = I$ ,  $k \geq 0$ , whenever  $d \geq 0$ , or

$$\bar{A} = -A, \quad W_{2\ell} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad W_{2\ell+1} = -W_{2\ell}, \quad \ell \geq 0, \quad d < 0.$$

Consider  $d$  such that  $|d| > 1$ . Clearly,  $\bar{\sigma} = \sigma$  and  $\mathcal{U} = \mathbb{R}^n$  satisfy the hypothesis of Lemma 7. As  $k \rightarrow \infty$ ,  $\bar{x}_k(\sigma)$  “aligns” with  $\mathcal{F} = [e_1]$ . Figure 1 illustrates the behavior of  $\theta$ ; note from the detail presented in Figure 1 (ii), for  $d = 10$ , that the convergence is slower in the interval  $100 \leq t \leq 200$  than in the interval  $0 \leq t \leq 100$ , suggesting it is not exponential. The case when  $|d| \leq 1$  is not addressed by Lemmas 6 or 7 (note that  $\mathcal{U}$  is trivial; for  $|d| < 1$ ,  $\mathcal{S}$  is trivial and, for  $|d| = 1$ ,  $\sigma$  does not completely excite  $\mathcal{S} = \mathbb{R}^n$ ).

The coupling projections  $\bar{H}$  differ from  $H$  as they are not orthogonal, because of the “distortion” introduced by the new bases (see Example 4), but they are similar

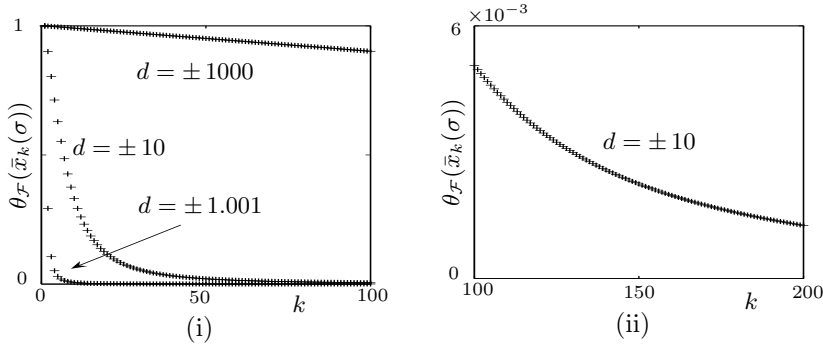


FIG. 1. Behavior of  $\theta$  for system  $\Theta$  of Example 3.

to  $H$  in the sense that  $\bar{H}\bar{v} = 0$  whenever  $Hv = 0$ . We shall need the following related result.

LEMMA 8. Consider the rank-one decomposition (11) for  $\Sigma$ . The projections  $\bar{H}_k$ ,  $k \geq 0$ , are such that  $\bar{H}_k v = 0$  for  $v \in [\bar{x}_k(\sigma_1), \dots, \bar{x}_k(\sigma_{r_\Sigma})]$ .

Proof. Note that  $\bar{H}_k v = \bar{H}_k(\pi_1 \bar{x}_k(\sigma_1) + \dots + \pi_{r_\Sigma} \bar{x}_k(\sigma_{r_\Sigma}))$  for certain scalars  $\pi_j$ ,  $j = 0, \dots, r_\Sigma$ , and, for each term of this sum, one can employ Lemma 3 (i) to evaluate  $\bar{H}_k \bar{x}_k(\sigma_j) = W_k H_k W_k^{-1} W_k x_k(\sigma_j) = W_k H_k x_k(\sigma_j) = 0$ ,  $j = 0, \dots, r_\Sigma$ , since  $H_k x_k(\sigma_j) = 0$  by definition of  $H_k$ .  $\square$

As  $\bar{x}_k(\sigma_j)$ ,  $0 \leq j \leq r_\Sigma$ , aligns with  $\mathcal{S}$  ( $\mathcal{F}$ , respectively) as stated in Lemma 6 (Lemma 7, respectively), we have that the projections  $\bar{H}_k$  onto  $[\bar{x}(\sigma_1), \dots, \bar{x}(\sigma_{r_\Sigma})]$  “tend to align” with the orthogonal projection onto  $\mathcal{S}^\perp$  ( $\mathcal{F}^\perp$ , respectively), which allows us to obtain the approximation results that will be useful for section 4. We present these results in the next lemma, in which  $S$ ,  $T$ , and  $U$  denote the orthogonal projections onto  $\mathcal{S}^\perp$ ,  $\mathcal{F}^\perp$ , and  $\mathcal{U} \ominus \mathcal{F}$ , respectively.

LEMMA 9. If  $\ker\{W_0 \Sigma W_0'\} \cap \mathcal{S} = \{0\}$  and  $\ker\{W_0 \Sigma W_0'\} \cap \mathcal{S}^\perp = \mathcal{S}^\perp$ , then there exist  $\pi \geq 0$ ,  $0 \leq \rho < 1$ , such that, for  $k \geq 0$ ,

$$(i) \quad \|S(I - \bar{H}_k)v\| \leq \pi \rho^k \|v\| \text{ and } \|\bar{H}_k(I - S)v\| \leq \pi \rho^k \|v\|.$$

If  $\ker\{W_0 \Sigma W_0'\} \cap \mathcal{U} \neq \{0\}$ , then there exist  $\delta, \lambda > 0$  and a nonnegative-valued, strictly decreasing function  $o(\cdot)$  such that, for  $k \geq 0$ ,

$$(ii) \quad \|T(I - \bar{H}_k)v\| \leq o(k) \|v\| \text{ and } \|\bar{H}_k(I - T)v\| \leq o(k) \|v\|;$$

$$(iii) \quad \|(U\bar{A})^{k+1}Uv\| \geq (1 + \delta)\|Uv\|;$$

$$(iv) \quad T\bar{A}Uv = U\bar{A}Uv;$$

$$(v) \quad \|\bar{H}_k Uv\| \leq \lambda \|Uv\|.$$

Proof. (i). Lemma 6 leads to the result, provided  $\mathcal{S}$  is nontrivial; for trivial  $\mathcal{S}$  it is simple to check that  $S = H_k = I$  and the result holds with  $\pi = 0$ . (ii) Lemma 8 can be employed when  $\Sigma \neq 0$ . The case with trivial  $\Sigma$  leads to  $T = H_k = I$  and  $o(k) = 0$ ,  $k \geq 0$ . (iii) It follows from the fact that  $U$  is the projection onto  $\mathcal{U} \ominus \mathcal{F}$ , which is spanned by unstable eigenvectors of  $\bar{A}$ ; moreover,  $(1 + \delta)$  equals the minimal of these eigenvalues. (iv)  $\mathcal{U} \ominus \mathcal{F}$  is not necessarily  $\bar{A}$ -invariant in general, but one can easily check from the structure of  $\bar{A}$  that, for  $w \in \mathcal{U} \ominus \mathcal{F}$ ,  $\bar{A}w \in \mathcal{U}$ , in such a manner that the component of  $\bar{A}w$  in  $\mathcal{U}^\perp$  is zero and  $T\bar{A}w = U\bar{A}w$ . (v) It follows from the facts that  $\bar{H}_k = W_k H_k W_k^{-1}$  and  $(1 - \kappa) \leq \|W_k\| \leq (1 + \kappa)$  for some  $0 \leq \kappa < 1$ , as in Lemma 2.  $\square$

Remark 1 (condition  $\ker\{W_0 \Sigma_1 W_0'\} \cap \mathcal{S}^\perp = \mathcal{S}^\perp$  in Lemma 9). In order to show that  $\ker\{W_0 \Sigma W_0'\} \cap \mathcal{S} = \{0\}$  is a sufficient condition for PSS, we may initially consider a “modified”  $\Sigma_1$  such that  $\ker\{W_0 \Sigma_1 W_0'\} \cap \mathcal{S}^\perp = \mathcal{S}^\perp$  and then employ the inequality

$Z_k(V_0, \Sigma) \leq Z_k(V_0, \Sigma_1)$  of Proposition 1 (iv) to extend the result to the original  $\Sigma$ ; see the proof of Theorem 1. Since  $\Sigma_1$  does not excite  $\mathcal{S}^\perp$  and excites all  $\mathcal{S}$ , there is no need to consider excited and nonexcited subspaces of  $\mathcal{S}^\perp$  and  $\mathcal{S}$  (as opposed, e.g., to  $\mathcal{U} \cap \mathcal{F}$  and  $\mathcal{U} \cap \mathcal{F}^\perp$ ). However, the above inequality is not suitable for dealing with the necessary condition for PS. That is why we study the projection  $S$  in (i) of Lemma 9, and  $T$  and  $U$  in (ii) of the same lemma.

*Example 4.* Consider the systems  $\Theta$  and  $\Theta_X$  with

$$(17) \quad A = \begin{bmatrix} 1 & -0.1 & 0 \\ 0.2 & 1 & 0 \\ 0 & 0 & 0.99 \end{bmatrix}, \quad \Sigma = \sigma\sigma', \quad \sigma = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Set  $W_k = \tilde{A}^{-k}$ , with

$$\tilde{A} \approx \begin{bmatrix} 0.9901 & -0.099 & 0 \\ 0.1980 & 0.9901 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \bar{A} = \begin{bmatrix} \sqrt{1.02} & 0 & 0 \\ 0 & \sqrt{1.02} & 0 \\ 0 & 0 & 0.99 \end{bmatrix}.$$

Figure 2 (i) illustrates how simple the behavior of  $\bar{x}_k$  is when compared to  $x_k$ . The oscillatory behavior of  $x_k$  prevents convergence of  $\theta_V(x_k)$  for any fixed  $V$ . The hypothesis of Lemma 7 holds with  $\ker\{W_0\Sigma W_0'\} \cap \mathcal{U} = [e_2]$  and, in fact, for  $\mathcal{F} = [e_1]$ ,  $\theta_{\mathcal{F}^\perp \cap \mathcal{U}}(\bar{x}_k(\sigma)) \geq 1$ , and  $\theta_{\mathcal{F}}(\bar{x}_k(\sigma)) \leq 0.3 \times 0.96^k$ . We have checked that  $\bar{H}_k \bar{x}_k(\sigma) = 0$ ,  $k \geq 0$ , confirming Lemma 8. The statements of Lemma 9 (ii)–(iv) hold with  $\delta = 0.001$ ,  $\lambda = 1.031$ , and  $o(k) = 0.985^k$ ; Figure 2 (ii) and (iii) illustrate the behavior of the quantities  $\|T(I - \bar{H}_k)v\|$  and  $\|\bar{H}_k Uv\|$  for  $v = (\sqrt{3}/3) [1 \ 1 \ 1]'$ . Note that  $\|\bar{H}_k Uv\|$  presents an oscillation due to the fact that  $\bar{H}_k$  are *nonorthogonal* projections.

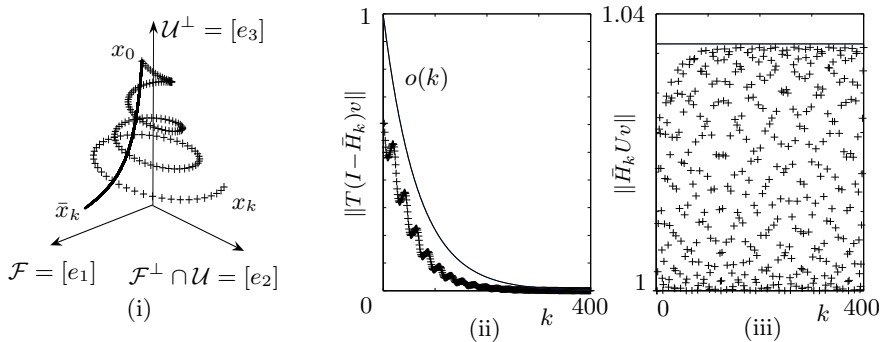


FIG. 2. Simulation results for system  $\Theta_X$  of Example 4. (i) State trajectory  $\bar{x}_k(\sigma)$ . (ii) and (iii) The quantities of Lemma 9 (ii) and (iv).

An important feature of the case with  $\ker\{W_0\Sigma W_0'\} \cap \mathcal{U} \neq \{0\}$  is that  $\text{Im}(\bar{H}) \cap \mathcal{U} \neq \{0\}$ , which follows from the fact that  $\bar{H}$  cannot “cover”  $\mathcal{U} \ominus \mathcal{F}$  as stated in Lemma 7. This fact, together with the structure of invariant spaces presented in Lemma 4, allows us to pick an initial condition  $\bar{z}$  for which the associated  $\bar{z}_k$  has a nontrivial projection onto  $\mathcal{U} \ominus \mathcal{F}$ , as stated in the next proposition, the proof of which is omitted.

**PROPOSITION 4.** *If  $\ker\{W_0\Sigma W_0'\} \cap \mathcal{U} \neq \{0\}$ , then there exists  $\bar{z} \in \mathcal{S}$  such that  $U\bar{z}_k \neq 0$ ,  $k \geq 0$ .*

**4. Testable condition for PS and PSS.** This section presents, initially, a sufficient condition for PS, with an extension to PSS. Then a necessary condition for PSS is presented and extended to PS. Finally, the results are gathered together in Theorem 1.

#### 4.1. Sufficient conditions.

LEMMA 10. Consider  $W_0$  as in Proposition 2, the subspace  $\mathcal{S}$  as in (15), and  $\bar{z}_k$  as in (12). If  $\ker\{W_0\Sigma W'_0\} \cap \mathcal{S} = \{0\}$  and  $\ker\{W_0\Sigma W'_0\} \cap \mathcal{S}^\perp = \mathcal{S}^\perp$ , then for each  $\bar{z}$  there exist  $\chi \geq 0$  and  $0 \leq \beta < 1$  such that  $\|\bar{z}_k\| \leq \chi\beta^k$ .

*Proof.* For ease of notation, in this proof we write  $\bar{z}$ ,  $\bar{A}$ , and  $\bar{H}$  as  $z$ ,  $A$ , and  $H$ , respectively; for  $\ell \geq 0$ ,  $w_{1,\ell}, w_{2,\ell}, w_{3,\ell}$  stand for vectors with  $\|w_{j,\ell}\| \leq 1$ . Recall the orthogonal projection  $S$  used in Lemma 9. From Lemma 4 we have that both  $\mathcal{S}^\perp$  and  $\mathcal{S}$  are  $A$ -invariant, in such a manner that  $ASz_{k+\ell} \in \mathcal{S}^\perp$  and  $A(I-S)z_{k+\ell} \in \mathcal{S}$ ,  $k, \ell \geq 0$ . Moreover,  $\ker\{W_0\Sigma W'_0\} \cap \mathcal{S} = \{0\}$ , and hence the conditions of Lemma 9 (i) hold, allowing us to evaluate, for  $k, \ell \geq 0$ ,

$$\begin{aligned}
 SH_{k+\ell+1}(ASz_{k+\ell}) &= S(ASz_{k+\ell}) + \pi\rho^{k+\ell+1}\|ASz_{k+\ell}\|w_{1,\ell} \\
 &= ASz_{k+\ell} + \pi\rho^{k+\ell+1}\|ASz_{k+\ell}\|w_{1,\ell}, \\
 (18) \quad H_{k+\ell+1}A(I-S)z_{k+\ell} &= H_{k+\ell+1}(I-S)(A(I-S)z_{k+\ell}) \\
 &= \pi\rho^{k+\ell+1}\|A(I-S)z_{k+\ell}\|w_{2,\ell},
 \end{aligned}$$

where  $\pi, \rho$  are as in Lemma 9. Now we shall show inductively that

$$\begin{aligned}
 (19) \quad z_{k+\ell+1} &= A^{\ell+1}Sz_k + (I-S)H_{k+\ell+1}A^{\ell+1}Sz_k \\
 &\quad + 2\pi\|A\|^{\ell+1}\|z_k\|(\rho^{k+1} + \dots + \rho^{k+\ell+1})w_{3,\ell}, \quad \ell \geq 0.
 \end{aligned}$$

For  $\ell = 0$ , from (4) and (18) we have that

$$\begin{aligned}
 z_{k+1} &= H_{k+1}Az_k = H_{k+1}ASz_k + H_{k+1}A(I-S)z_k \\
 &= SH_{k+1}ASz_k + (I-S)H_{k+1}ASz_k + H_{k+1}A(I-S)z_k \\
 &= ASz_k + (I-S)H_{k+1}ASz_k \\
 &\quad + \pi\rho^{k+\ell+1}(\|ASz_k\|w_{1,0} + \|A(I-S)z_k\|w_{2,0}) \\
 &= ASz_k + (I-S)H_{k+1}ASz_k + \pi\rho^{k+\ell+1}(2\|A\|\|z_k\|w_{3,0}),
 \end{aligned}$$

and assuming (19) holds for  $\ell - 1$ , similarly to the above we evaluate from (4)

$$\begin{aligned}
 z_{k+\ell+1} &= H_{k+\ell+1}Az_{k+\ell} \\
 &= H_{k+\ell+1}A^{\ell+1}Sz_k \\
 &\quad + H_{k+\ell+1}A(I-S)H_{k+\ell}A^\ell Sz_k \\
 &\quad + H_{k+\ell+1}A(2\pi\|A\|^\ell\|z_k\|(\rho^{k+1} + \dots + \rho^{k+\ell})w_{3,\ell-1}) \\
 &= SH_{k+\ell+1}A^{\ell+1}Sz_k + (I-S)H_{k+\ell+1}A^{\ell+1}Sz_k \\
 &\quad + H_{k+\ell+1}A(I-S)H_{k+\ell}A^\ell Sz_k \\
 &\quad + H_{k+\ell+1}A(2\pi\|A\|^\ell\|z_k\|(\rho^{k+1} + \dots + \rho^{k+\ell})w_{3,\ell-1})
 \end{aligned}$$



and, from (18),

$$\begin{aligned}
 z_{k+\ell+1} &= A^{\ell+1}Sz_k + \pi\rho^{k+\ell+1}\|A^{\ell+1}Sz_k\|w_{1,k+\ell+1} \\
 &\quad + (I - S)H_{k+\ell+1}A^{\ell+1}Sz_k \\
 &\quad + \pi\rho^{k+\ell+1}\|A(I - S)H_{k+\ell}A^\ell Sz_k\|w_{2,k+\ell+1} \\
 &\quad + H_{k+\ell+1}A(2\pi\|A\|^\ell\|z_k\|(\rho^{k+1} + \dots + \rho^{k+\ell})w_{3,\ell-1}) \\
 &= A^{\ell+1}Sz_k + (I - S)H_{k+\ell+1}A^{\ell+1}Sz_k \\
 &\quad + 2\pi\|A\|^{\ell+1}\|z_k\|(\rho^{k+1} + \dots + \rho^{k+\ell} + \pi\rho^{k+\ell+1})w_{3,\ell},
 \end{aligned}$$

completing the inductive proof of (19). Then we can write, for  $k, \ell \geq 0$ ,

$$\begin{aligned}
 \|z_{k+\ell+1}\| &\leq \|A^{\ell+1}Sz_k\| + \|(I - S)H_{k+\ell+1}A^{\ell+1}Sz_k\| \\
 (20) \quad &\quad + 2\pi\|A\|^{\ell+1}\|z_k\|(\rho^{k+1} + \dots + \rho^{k+\ell+1}) \\
 &\leq 2\|A^{\ell+1}Sz_k\| + 2\pi\|A\|^{\ell+1}\|z_k\|(\rho^{k+1} + \dots + \rho^{k+\ell+1}).
 \end{aligned}$$

Now consider the term  $A^\ell Sz_k$ ,  $\ell \geq 1$ . Since  $\mathcal{S}^\perp$  is  $A$ -invariant and corresponds to the subspace spanned by eigenvectors associated to eigenvalues (strictly) inside the unit disk, one has that  $\|A^\ell Sz_k\| \leq \eta\gamma^\ell\|Sz_k\| \leq \eta\gamma^\ell\|z_k\|$  for some scalars  $\eta \geq 0$  and  $0 \leq \gamma < 1$ . Then we set

$$\ell_0 : \eta\gamma^{\ell_0} \leq 1/4,$$

and from (20) with  $\ell = \ell_0 - 1$  we obtain

$$\begin{aligned}
 \|z_{k+\ell_0}\| &\leq (2\eta\gamma^{\ell_0} + 2\pi\|A\|^{\ell_0}(\rho^{k+1} + \dots + \rho^{k+\ell_0}))\|z_k\| \\
 (21) \quad &\leq (1/2 + 2\pi\|A\|^{\ell_0}(\rho^{k+1} + \dots + \rho^{k+\ell_0}))\|z_k\|, \quad k \geq 0.
 \end{aligned}$$

Now we set  $k_0$  such that  $2\pi\|A\|^{\ell_0}(\rho^{k_0+1} + \dots + \rho^{k_0+\ell_0}) < 1/2$ . From (21) with  $k = k_0$ , we obtain  $\|z_{k_0+\ell_0}\| \leq \bar{\beta}\|z_{k_0}\|$ , where  $\bar{\beta} = (1/2 + 2\pi\|A\|^{\ell_0}(\rho^{k_0+1} + \dots + \rho^{k_0+\ell_0})) < 1$ ; similarly, from (21) with  $k = k_0 + m\ell_0$ ,  $m \geq 0$ , we obtain

$$\|z_{k_0+m\ell_0+\ell_0}\| \leq \bar{\beta}\|z_{k_0+m\ell_0}\| \leq \bar{\beta}^2\|z_{k_0+(m-1)\ell_0}\| \leq \dots \leq \bar{\beta}^{m+1}\|z_{k_0}\|.$$

Finally, we have that each  $k \geq k_0$  can be written in the form  $k = k_0 + m\ell_0 + r$  for some  $0 \leq r < \ell_0$  and  $m$  with  $(k - k_0)/\ell_0 \leq m \leq (k - k_0)/\ell_0 + 1$ , leading to

$$\begin{aligned}
 \|z_k\| &\leq \|A\|^r\|z_{k_0+m\ell_0}\| \\
 &\leq \|A\|^{\ell_0}\bar{\beta}^m\|z_{k_0}\| \leq \|A\|^{\ell_0}\|z_{k_0}\|\bar{\beta}^{-1}(\pi^{1/\ell_0})^k, \quad k \geq k_0,
 \end{aligned}$$

and since  $\|z_k\| \leq \|A\|^k\|z_0\|$ ,  $k < k_0$ , it is a simple matter to check that we can set  $\beta = \bar{\beta}^{1/t_0} < 1$  and find  $\chi \geq 0$  for which  $\|z_k\| \leq \chi\beta^k$ ,  $k \geq 0$ .  $\square$

Lemma 10 can be easily extended to the context of semistability of the system  $\Theta_Z$  by employing  $\xi < 1$  as a scaling factor that “converts”  $\mathcal{U}$  associated with the matrix  $\bar{A}$  into  $\mathcal{S}_\xi$  associated with  $\xi\bar{A}$ .

**COROLLARY 1.** *Consider the system  $\Theta_Z$ ,  $W_0$  as in Proposition 2 and  $\mathcal{U}$  as in (15). If  $\ker\{W_0\Sigma W_0'\} \cap \mathcal{U} = \{0\}$  and  $\ker\{W_0\Sigma W_0'\} \cap \mathcal{U}^\perp = \mathcal{U}^\perp$ , then for each  $z$  and  $0 \leq \zeta < 1$  there exist  $\alpha \geq 0$  and  $0 \leq \beta < 1$  such that  $\|\zeta^k z_k\| \leq \alpha\beta^k$ .*

*Proof.* Let  $\Theta_Z$ ,  $\mathcal{S}$ , and  $\mathcal{U}$  correspond to the matrix  $A$  and, for  $0 \leq \xi < 1$ , let  $\Theta_{\xi Z}$ ,  $\mathcal{S}_\xi$ , and  $\mathcal{U}_\xi$  correspond to the matrix  $\xi A$ . Note that the eigenvalues of  $\bar{A}$  lying in the unit disk are shifted to eigenvalues of  $\xi \bar{A}$  inside the disk, yielding  $\mathcal{U} \supset \mathcal{S}_\xi$  for a general  $0 \leq \xi < 1$ . Let  $\xi$  be sufficiently close to one, in such a manner that  $\xi \geq \zeta$  and  $\mathcal{U} = \mathcal{S}_\xi$ . This leads to  $(\ker\{W_0 \Sigma W'_0\} \cap \mathcal{S}_\xi) = (\ker\{W_0 \Sigma W'_0\} \cap \mathcal{U}) = \{0\}$ ; furthermore,  $\mathcal{S}_\xi^\perp = \mathcal{U}^\perp$  yields  $(\ker\{W_0 \Sigma W'_0\} \cap \mathcal{S}_\xi^\perp) = (\ker\{W_0 \Sigma W'_0\} \cap \mathcal{U}^\perp) = \mathcal{U}^\perp = \mathcal{S}_\xi^\perp$ . Employing the result of Lemma 10 for system  $\Theta_{\xi Z}$  yields that there exist  $\chi \geq 0$  and  $0 \leq \beta < 1$  for which

$$(22) \quad \|\bar{z}_{\xi,k}\| \leq \chi \beta^k.$$

From Proposition 3 we get that  $\bar{z}_{\xi,k} = \xi^k \bar{z}_k$ , allowing us to obtain from (22)

$$(23) \quad \|\xi^k \bar{z}_k\| = \|\bar{z}_{\xi,k}\| \leq \chi \beta^k,$$

and Lemma 3 (ii) provides  $\|\xi^k z_k\| \leq (1 - \kappa)^{-1} \|\xi^k \bar{z}_k\| \leq (1 - \kappa)^{-1} \chi \beta^k$ . Recalling that  $\zeta \leq \xi$ , we have  $\|\zeta^k z_k\| \leq \|\xi^k z_k\| \leq (1 - \kappa)^{-1} \chi \beta^k$ .  $\square$

**4.2. Necessary conditions.** Conversely to Corollary 1, if  $\Sigma$  does not completely excite  $\mathcal{U}$ , then exponential divergence takes place. It is convenient, for later reference, to formalize the result as follows.

LEMMA 11. *Consider the system  $\Theta_Z$ ,  $W_0$  as in Proposition 2 and  $\mathcal{U}$  as in (15). If  $\ker\{W_0 \Sigma W'_0\} \cap \mathcal{U} \neq \{0\}$ , then there exist  $z \in \mathbb{R}^n$  and  $0 \leq \zeta < 1$  such that for all  $\chi \geq 0$  and  $0 \leq \psi < 1$ ,  $\|\zeta^k z_k\| > \chi \psi^k$  for some  $k \geq 0$ .*

*Proof.* We start setting  $\zeta < 1$  sufficiently close to one, in such a manner that  $\zeta(1 + \delta) > 1$ , where  $\delta$  is as in Lemma 9, and, simultaneously,  $\lambda_i(\zeta A) \notin \bar{\mathbb{D}}$  if and only if  $\lambda_i(A) \notin \bar{\mathbb{D}}$ ,  $0 \leq i \leq n$  (the unstable space of  $\zeta A$  equals the unstable space of  $A$ ). For ease of notation, in what follows we write  $\bar{z}$ ,  $\bar{A}$ , and  $\bar{H}$  as  $z$ ,  $A$ , and  $H$ , respectively; for  $\ell \geq 0$ ,  $w_{1,\ell}$ ,  $w_{2,\ell}$ ,  $w_{3,\ell}$  stand for vectors with  $\|w_{j,\ell}\| \leq 1$ . We shall need an evaluation that is analogous to (19) of Lemma 10. In fact, (19) involves projections onto  $\mathcal{S}^\perp$  and  $\mathcal{S}$  via  $S$  and  $I - S$ , respectively, and now we consider projections onto  $\mathcal{U} \ominus \mathcal{F}$ ,  $\mathcal{F}$ , and  $\mathcal{U}^\perp \ominus \mathcal{F}$  via  $U$ ,  $(I - T)$ , and  $(I - U)T$ , respectively. Using Lemma 9 (ii) and (iii) yields

$$\begin{aligned} z_{k+\ell+1} &= (TA)^{\ell+1} U z_k + A^{\ell+1} (I - U) T z_k \\ &\quad + (I - T) H_{k+\ell+1} (TA)^{\ell+1} U z_k \\ &\quad + (I - T) H_{k+\ell+1} A^{\ell+1} (I - U) T z_k \\ &\quad + 4 \|A\|^{\ell+1} \|z_k\| (o(k+1) + \dots + o(k+\ell+1)) w_{3,\ell}, \quad \ell \geq 0, \end{aligned}$$

and, since Lemma 9 (v) provides  $(TA)^{\ell+1} U = (UA)^{\ell+1} U$ , this can be written as

$$\begin{aligned} z_{k+\ell+1} &= (UA)^{\ell+1} U z_k + A^{\ell+1} (I - U) T z_k \\ &\quad + (I - T) H_{k+\ell+1} (UA)^{\ell+1} U z_k \\ &\quad + (I - T) H_{k+\ell+1} A^{\ell+1} (I - U) T z_k \\ &\quad + 4 \|A\|^{\ell+1} \|z_k\| (o(k+1) + \dots + o(k+\ell+1)) w_{3,\ell}, \quad \ell \geq 0. \end{aligned} \tag{24}$$

Similarly to (18), we have from Lemma 9 (ii)

$$\begin{aligned} TH_{k+\ell+1} (UA)^{\ell+1} U z_k &= T(UA)^{\ell+1} U z_k + o(k+\ell+1) \|(UA)^{\ell+1} U z_k\| w_{1,\ell} \\ &= (UA)^{\ell+1} U z_k + o(k+\ell+1) \|(UA)^{\ell+1} U z_k\| w_{1,\ell}, \end{aligned}$$

which can be substituted in the third term on the right-hand side of (24), leading to

$$\begin{aligned}
 z_{k+\ell+1} &= H_{k+\ell+1}(UA)^{\ell+1}Uz_k - o(k+\ell+1)\|(UA)^{\ell+1}Uz_k\|w_{1,\ell} \\
 &\quad + A^{\ell+1}(I-U)Tz_k + (I-T)H_{k+\ell+1}A^{\ell+1}(I-U)Tz_k \\
 &\quad + 4\|A\|^{\ell+1}\|z_k\|(o(k+1) + \cdots + o(k+\ell+1))w_{3,\ell} \\
 (25) \quad &= H_{k+\ell+1}(UA)^{\ell+1}Uz_k \\
 &\quad + A^{\ell+1}(I-U)Tz_k + (I-T)H_{k+\ell+1}A^{\ell+1}(I-U)Tz_k \\
 &\quad + 5\|A\|^{\ell+1}\|z_k\|(o(k+1) + \cdots + o(k+\ell+1))w_{4,\ell}, \quad \ell \geq 0.
 \end{aligned}$$

Regarding the second and third terms on the right-hand side of (25), recall that  $(I-U)Tz_z \in (\mathcal{U}^\perp \ominus \mathcal{F}) \subset \mathcal{U}^\perp$ , yielding that  $A^k(I-U)Tz_z$  may present polynomial divergence, as  $k \rightarrow \infty$ ; hence we can write for each  $\gamma > 1$ , and in particular for  $\gamma$  such that  $\gamma\zeta < 1$ ,

$$\begin{aligned}
 &\|A^{\ell+1}(I-U)Tz_k + (I-T)H_{k+\ell+1}A^{\ell+1}(I-U)Tz_k\| \\
 (26) \quad &\leq \|A^{\ell+1}(I-U)Tz_k\| + \|(I-T)\| \|H_{k+\ell+1}\| \|A^{\ell+1}(I-U)Tz_k\| \\
 &= 2\|A^{\ell+1}(I-U)Tz_k\| \leq 2\eta\gamma^{\ell+1}\|z_k\|
 \end{aligned}$$

for some  $\eta \geq 0$ . Note that (25), (26), and Lemma 9 (v) lead to

$$\begin{aligned}
 \|z_{k+\ell+1}\| &\leq \|H_{k+\ell+1}(UA)^{\ell+1}Uz_k\| + 2\eta\gamma^{\ell+1}\|z_k\| \\
 &\quad + 5\|A\|^{\ell+1}\|z_k\|(o(k+1) + \cdots + o(k+\ell+1)) \\
 (27) \quad &\leq \lambda\|(UA)^{\ell+1}Uz_k\| + 2\eta\gamma^{\ell+1}\|z_k\| \\
 &\quad + 5\|A\|^{\ell+1}\|z_k\|(o(k+1) + \cdots + o(k+\ell+1)).
 \end{aligned}$$

On the other hand, premultiplying both sides of (24) by  $U$  and employing the fact that, for  $v \in \mathbb{R}^n$ ,  $(I-T)v \in \mathcal{F}$ , yielding  $(I-T)v \perp (\mathcal{F}^\perp \cap \mathcal{U})$  and hence  $U(I-T) = 0$ , evaluations similar to the above ones provide

$$\begin{aligned}
 \|Uz_{k+\ell+1}\| &= \|(UA)^{\ell+1}Uz_k + UA^{\ell+1}(I-U)Tz_k \\
 &\quad + 4\|A\|^{\ell+1}\|z_k\|(o(k+1) + \cdots + o(k+\ell+1))Uw_{3,\ell}\| \\
 (28) \quad &\geq \|(UA)^{\ell+1}Uz_k\| - \eta\gamma^{\ell+1}\|z_k\| \\
 &\quad - 4\|A\|^{\ell+1}\|z_k\|(o(k+1) + \cdots + o(k+\ell+1)).
 \end{aligned}$$

By substituting (28) in (27) we get that

$$\begin{aligned}
 \|z_{k+\ell+1}\| &\leq \lambda(\|Uz_{k+\ell+1}\| + \eta\gamma^{\ell+1}\|z_k\| \\
 &\quad + 4\|A\|^{\ell+1}\|z_k\|(o(k+1) + \cdots + o(k+\ell+1))) \\
 &\quad + 2\eta\gamma^{\ell+1}\|z_k\| + 5\|A\|^{\ell+1}\|z_k\|(o(k+1) + \cdots + o(k+\ell+1)) \\
 &= \lambda\|Uz_{k+\ell+1}\| + (2+\lambda)\eta\gamma^{\ell+1}\|z_k\| \\
 &\quad + (5+4\lambda)\|A\|^{\ell+1}\|z_k\|(o(k+1) + \cdots + o(k+\ell+1)), \quad \ell \geq 0,
 \end{aligned}$$

or equivalently, for  $\ell \geq 0$ ,

$$(29) \quad \begin{aligned} \|Uz_{k+\ell+1}\| &\geq \lambda^{-1}\|z_{k+\ell+1}\| - \lambda^{-1}(2+\lambda)\eta\gamma^{\ell+1}\|z_k\| \\ &\quad - \lambda^{-1}(5+4\lambda)\|A\|^{\ell+1}\|z_k\|(o(k+1) + \dots + o(k+\ell+1)). \end{aligned}$$

Now we employ the fact that  $\mathcal{U} \ominus \mathcal{F}^\perp$  is associated with unstable eigenvalues of  $A$ . We proceed similarly to the above, employing (24) with  $k$  replaced by  $k+\ell+1$  and  $\ell$  replaced by  $m-1$ , and Lemma 9 (ii)–(v), to evaluate

$$\begin{aligned} \|z_{k+\ell+m+1}\| &\geq \|(UA)^m Uz_{k+\ell+1} + (I-T)H_{k+\ell+m+1}(UA)^m Uz_{k+\ell+1} \\ &\quad + (I-T)H_{k+\ell+m+1}A^m(I-U)Tz_{k+\ell+1}\| - \|A^m(I-U)Tz_{k+\ell+1}\| \\ &\quad - 4\|A\|^m\|z_k\|(o(k+\ell+2) + \dots + o(k+\ell+m+1)). \end{aligned}$$

Recalling that  $(I-T)$  and  $U$  are orthogonal projections and employing Lemma 9 (iii) and an evaluation similar to (26) (with the same  $\gamma$  as in (26), such that  $\gamma\zeta < 1$ ), the above inequality leads to

$$\begin{aligned} \|z_{k+\ell+m+1}\| &\geq \|(UA)^m Uz_{k+\ell+1}\| - \|A^m(I-U)Tz_{k+\ell+1}\| \\ &\quad - 4\alpha\|A\|^m\|z_{k+\ell+1}\|(\beta^{k+\ell+2} + \dots + \beta^{k+\ell+m+1}) \\ &\geq (1+\delta)^m\|Uz_{k+\ell+1}\| - \eta\gamma^m\|z_{k+\ell+1}\| \\ &\quad - 4\|A\|^m\|z_{k+\ell+1}\|(o(k+\ell+2) + \dots + o(k+\ell+m+1)) \end{aligned}$$

and, from (29),

$$\begin{aligned} \|z_{k+\ell+m+1}\| &\geq (1+\delta)^m\lambda^{-1}\left(\|z_{k+\ell+1}\| - (2+\lambda)\eta\gamma^{\ell+1}\|z_k\| \right. \\ &\quad \left. - (5+4\lambda)\|A\|^{\ell+1}(o(k+1) + \dots + o(k+\ell+1))\|z_k\| \right) \\ &\quad - \eta\gamma^m\|z_{k+\ell+1}\| \\ &\quad - 4\|A\|^m(o(k+\ell+2) + \dots + o(k+\ell+m+1))\|z_{k+\ell+1}\|. \end{aligned}$$

Multiplying both sides of the above equation by  $\zeta^{k+\ell+m+1}$  and setting  $\xi = \gamma\zeta$  and  $z_{\zeta,k} = \zeta^k z_k$ ,  $\ell \geq 0$ , lead to

$$(30) \quad \begin{aligned} \|z_{\zeta,k+\ell+m+1}\| &\geq \zeta^m(1+\delta)^m\lambda^{-1}\left(\|z_{\zeta,k+\ell+1}\| - (2+\lambda)\eta\xi^{\ell+1}\|z_{\zeta,k}\| \right. \\ &\quad \left. - (5+4\lambda)\zeta^{\ell+1}\|A\|^{\ell+1}(o(k+1) + \dots + o(k+\ell+1))\|z_{\zeta,k}\| \right) \\ &\quad - \eta\xi^m\|z_{\zeta,k+\ell+1}\| \\ &\quad - 4\zeta^m\|A\|^m(o(k+\ell+2) + \dots + o(k+\ell+m+1))\|z_{\zeta,k+\ell+1}\|. \end{aligned}$$

Recall that  $\zeta(1+\delta) > 1$ ,  $\xi = \gamma\zeta < 1$ , and  $o(\cdot)$  is a decreasing function, allowing for the following settings. For an arbitrary  $\bar{m} > 0$ , let  $m$  be such that

$$(31) \quad \zeta^m(1+\delta)^m\lambda^{-1} - \eta\xi^m > 6\bar{m}.$$

Set  $\ell = \max(\ell_1, \ell_2)$ , where  $\ell_1$  is such that

$$(32) \quad \zeta^m(1+\delta)^m\lambda^{-1}(2+\lambda)\eta\xi^{\ell_1+1} < (1/2)\bar{m}$$

and  $\ell_2$  is such that  $4\zeta^m\|A\|^m(o(\ell_2+2)+\cdots+o(\ell_2+m+1))<3\bar{m}$ , yielding

$$(33) \quad 4\zeta^m\|A\|^m(o(k+\ell+2)+\cdots+o(k+\ell+m+1))<3\bar{m}, \quad k \geq 0.$$

Finally, let  $k$  be such that

$$(34) \quad \zeta^m(1+\delta)^m\lambda^{-1}(5+4\lambda)\zeta^{\ell+1}\|A\|^{\ell+1}(o(k+1)+\cdots+o(k+\ell+1))<(1/2)\bar{m}.$$

Substituting (31)–(34) in (30) provides

$$\begin{aligned} \|z_{\zeta,k+\ell+m+1}\| &\geq 6\bar{m}\|z_{\zeta,k+\ell+1}\| - (1/2)\bar{m}\|z_{\zeta,k}\| - (1/2)\bar{m}\|z_{\zeta,k}\| - 3\bar{m}\|z_{\zeta,k+\ell+1}\| \\ &= \bar{m}\|z_{\zeta,k+\ell+1}\| + 2\bar{m}\|z_{\zeta,k+\ell+1}\| - \bar{m}\|z_{\zeta,k}\|. \end{aligned}$$

Using this inequality in a recursive fashion, substituting  $k$  with  $k+qm$ ,  $q \geq 0$ , we obtain

$$(35) \quad \|z_{\zeta,k+\ell+1+(q+1)m}\| \geq \bar{m}^q\|z_{\zeta,k+\ell+1+m}\| + \sum_{j=0}^q (2\bar{m})^j\|z_{\zeta,k+\ell+1+jm}\| - \sum_{j=0}^q \bar{m}^j\|z_{\zeta,k+jm}\|.$$

The second term on the right-hand side of (35) dominates the third one, leading to exponential divergence; in particular, we can write for  $q \geq q_0$  for some  $q_0$  that  $\|z_{\zeta,k+\ell+1+(q+1)m}\| \geq \|z_{\zeta,k+\ell+1+m}\|$ , leading to

$$\zeta^{k+\ell+1+(q+1)m}\|z_{k+\ell+1+(q+1)m}\| \geq \zeta^{k+\ell+1+m}\|z_{k+\ell+1+m}\|$$

or, equivalently,  $\|z_{k+\ell+1+(q+1)m}\| \geq \zeta^{-q}\|z_{k+\ell+1+m}\|$ ,  $q \geq q_0$ . It is important to mention that we can pick an initial condition  $z$  for which  $z_{k+\ell+1+m} \neq 0$ ; see Proposition 4.  $\square$

Similarly to the proof of Corollary 1, we employ a “scaling factor”  $\xi < 1$  that converts  $\mathcal{S}$  associated with  $\bar{A}$  into  $\mathcal{U}_{\xi^{-1}}$  associated with  $\xi^{-1}\bar{A}$ , and extend the result of Lemma 11 to the context of PS. It is worth mentioning that Lemma 11 applied to matrix  $\xi^{-1}\bar{A}$  provides  $\|\zeta^k\xi^{-k}\bar{z}_k\| > \chi\psi^k$  (in analogy with (23)), or, equivalently,  $\|\bar{z}_k\| > \chi(\psi\xi^{-1})^k$ , and we can set both  $\zeta$  and  $\xi$  arbitrarily close to one, to obtain the next result, the proof of which is not presented.

**COROLLARY 2.** *Consider the system  $\Theta_Z$ ,  $W_0$  as in Proposition 2 and  $\mathcal{S}$  as in (15). If  $\ker\{W_0\Sigma W_0'\} \cap \mathcal{S} \neq \{0\}$ , then there exist  $z \in \mathbb{R}^n$  such that for all  $\chi \geq 0$  and  $0 \leq \psi < 1$ ,  $\|z_k\| > \chi\psi^k$  for some  $k \geq 0$ .*

### 4.3. Main result.

**THEOREM 1.** *Consider the system  $\Theta$ . Let  $J$  represent the similarity transformation for which  $JAJ^{-1}$  is in Jordan form and let  $\mathcal{J}$  and  $\mathcal{J}_{\mathcal{S}}$  stand for the unstable space and the stable space of  $JAJ^{-1}$ , respectively.  $(A, \Sigma)$  is PSS if and only if*

$$(36) \quad \ker\{J\Sigma J'\} \cap \mathcal{J} = \{0\}.$$

*$(A, \Sigma)$  is PS if and only if*

$$(37) \quad \ker\{J\Sigma J'\} \cap \mathcal{J}_{\mathcal{S}}^{\perp} = \{0\}.$$

*Proof.* Regarding the first statement, consider  $W_0$  as in Proposition 2 and  $\mathcal{U}$  as in (15). Condition (36) holds if and only if

$$(38) \quad \ker\{W_0\Sigma W_0'\} \cap \mathcal{U} = \{0\};$$

see Lemma 5. Assume that  $\ker\{W_0\Sigma W_0'\} \cap \mathcal{S}^\perp = \mathcal{S}^\perp$ . It follows from Corollary 1 and Lemma 11 that (38) is a necessary and sufficient condition for the existence, for each  $z \in \mathbb{R}^n$  and  $0 \leq \zeta < 1$ , of  $\alpha \geq 0$  and  $0 \leq \beta < 1$  such that  $\|\zeta^k z_k\| \leq \alpha\beta^k$ . Lemma 2 extends the result to PSS of  $(A, \Sigma)$ . Now we consider the case when  $\ker\{W_0\Sigma W_0'\} \cap \mathcal{S}^\perp \neq \mathcal{S}^\perp$ , that is,  $\Sigma$  also excites  $\mathcal{S}^\perp$ . In this situation, we can write  $\Sigma = \Sigma_1 + \Sigma_2$  with  $\Sigma_1, \Sigma_2 \in \mathcal{R}^{n0}$  and  $\ker\{W_0\Sigma_1 W_0'\} \cap \mathcal{S}^\perp = \mathcal{S}^\perp$ , to conclude that  $(A, \Sigma_1)$  is PSS, that is, for each  $0 \leq \zeta < 1$  and  $V \in \mathcal{R}^{n0}$ , there exists  $\bar{Z} \in \mathcal{R}^{n0}$  for which

$$(39) \quad Z_k(V, \Sigma_1) \leq \zeta^{-k} \bar{Z}.$$

Finally, since  $\Sigma - \Sigma_1 = \Sigma_2 \geq 0$ , Proposition 1 (iv) yields  $Z_k(V, \Sigma) \leq Z_k(V, \Sigma_1)$  and (39) provides  $Z_k(V, \Sigma) \leq \zeta^{-k} \bar{Z}$ , i.e.,  $(A, \Sigma)$  is PSS. The proof for the second statement follows from Corollary 2 and Lemma 10 in a similar manner as above.  $\square$

*Remark 2.* Either  $\Sigma > 0$  or semistable  $A$  imply  $(A, \Sigma)$  is PSS, which implies that  $(A, \Sigma)$  is semistabilizable. Indeed,  $\Sigma > 0$  provides  $\ker\{J\Sigma J'\} = \{0\}$  and semistable  $A$  yields  $\mathcal{J} = \{0\}$ , and in both cases (36) holds. Regarding the second implication,  $(A, \Sigma)$  not semistabilizable means that  $\Sigma$  does not excite an “entire” unstable mode of  $A$ , and (36) does not hold. PSS is not comparable to stabilizability of  $(A, \Sigma)$ ; indeed, Example 6 illustrates the situation when  $(A, \Sigma)$  is stabilizable but  $\Theta$  is not PSS, whereas system  $\Theta$  with  $A = 1$  and  $\Sigma = 0$  illustrates the opposite situation. Similarly, stable  $A$  imply that  $(A, \Sigma)$  is PS, which implies that  $(A, \Sigma)$  is stabilizable.

*Example 5.* Consider the system  $\Theta$  of Example 3 with  $|d| > 1$ . One has that  $J = I$ ,  $\ker\{J\Sigma J'\} = [e_1]$ , and  $\mathcal{J} = \mathcal{J}_S^\perp = \mathbb{R}^n$ . Neither (36) nor (37) holds, and from Theorem 1 we have that  $(A, \Sigma)$  is not PS or PSS. For  $|d| < 1$ ,  $\mathcal{J} = \mathcal{J}_S^\perp = \{0\}$ , and the system is PS and PSS. For  $|d| = 1$ ,  $\mathcal{J} = \{0\}$  and  $\mathcal{J}_S^\perp = \mathbb{R}^n$ , and hence only (36) holds and the system is “strictly” PSS; the same holds if  $\sigma = e_2$  is replaced with  $\sigma_2 = 0$ , as in Example 1.

*Example 6.* Consider the system  $\Theta$  of Example 4. It is simple to check that (36) is not satisfied and, according to Theorem 1,  $(A, \Sigma)$  is not PSS. See Figure 3 (i) for the behavior of the  $Z$ -component of the state trajectory. Note that  $(A, \Sigma)$  is stabilizable but the system is not PSS. Now, replace  $A$  with  $1.02^{-1/2}A$ , which leads to  $\lambda_i(A) \in \mathbb{D}$  and  $\mathcal{J} = \{0\}$ . In this situation, only (36) holds. Note via Figure 3 (ii) that the  $Z$ -component presents an oscillatory behavior, which is compatible with PSS.

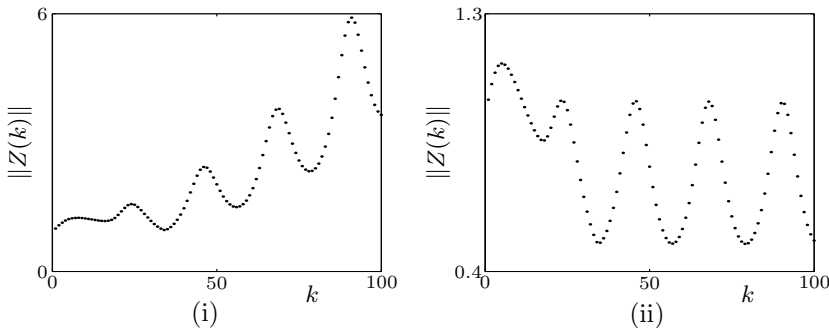


FIG. 3. Behavior of the  $Z$ -component for the setups of Example 6.

*Remark 3.* The conditions of Theorem 1 can be employed in the problem of “stabilization” of the system  $\Theta$  via a suitable choice of  $\Sigma$ . For instance,  $\Sigma = \sigma_1\sigma_1' + \dots + \sigma_r\sigma_r'$  with  $r = \dim(\mathcal{J})$  is such that  $(A, \Sigma)$  is PSS if and only if  $\sigma_j$ ,  $0 \leq j \leq r$ ,

are linearly independent vectors with nontrivial projections onto  $\mathcal{J}$ , and a similar condition holds for PS. As illustration, for the system  $\Theta$  of Example 5, if we set  $\sigma_1 = e_1$  and  $\sigma_2 = e_2$ , we obtain both PS and PSS; the same is valid for the system of Example 6 (of course, now  $e_1, e_2 \in \mathbb{R}^3$  and  $\Sigma$  is rank deficient).

**5. Concluding remarks.** In this paper we have explored the structure of the system  $\Theta$  in (1), with special attention to the relations among the initial condition  $\Sigma$  of the  $X$ -component, its dynamics (governed by  $A$ ), and the coupling with the  $Z$ -component via the orthogonal projection  $H$ . We obtain the structural, testable condition (36) for PSS, with the interpretation that  $\Sigma$  has to completely excite the unstable modes of  $A$ . Similarly, the condition (37) for PS requires that  $\Sigma$  excite all modes of  $A$  except the stable ones. These interpretations are particularly meaningful in the scenario of Kalman filtering for linear time-invariant systems, meaning that the noise in the initial condition excites the unstable or the “semiunstable” dynamics of the plant; indeed, the derived conditions are essential to obtain, as discussed in [4], necessary and sufficient conditions for avoiding *actual* exponential divergence of estimates and bounded error estimates, under incorrect noise measurements, which is a significant result, taking into account the conservativeness of existing conditions. The results can also be employed in the problem of stabilization of the filter via a suitable choice of noise model; see Remark 3.

## REFERENCES

- [1] H. ABOU-KANDIL, G. FREILING, V. IONESCU, AND G. JANK, *Matrix Riccati Equations in Control and Systems Theory*, Birkhäuser, Basel, 2003.
- [2] V. CHELLABOINA AND W. M. HADDAD, *A unification between partial stability and stability theory for time-varying systems*, IEEE Control Systems Mag., 22 (2002), pp. 66–75.
- [3] E. F. COSTA AND A. ASTOLFI, *A necessary and sufficient condition for semi-stability of the recursive Kalman filter*, in Proceedings of the 2008 American Control Conference, Seattle, WA, 2008, pp. 1280–1285.
- [4] E. F. COSTA AND A. ASTOLFI, *On the stability of the recursive Kalman filter for linear, time-invariant systems*, in Proceedings of the 2008 American Control Conference, Seattle, WA, 2008, pp. 1286–1291.
- [5] E. F. COSTA AND A. ASTOLFI, *Partial semi-stability for a class of non-linear systems*, in Proceedings of the 17th IFAC World Congress 2008, Seoul, South Korea, 2008, pp. 1442–1447.
- [6] T. E. DJAFERIS, *Partial stability preserving maps and stabilization*, in Proceedings of the 42nd IEEE Conference on Decision and Control, 2006, pp. 2490–2495.
- [7] R. J. FITZGERALD, *Divergence of the Kalman filter*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 736–747.
- [8] P. R. HALMOS, *Finite-Dimensional Vector Spaces*, 2nd ed., Springer, London, 1974.
- [9] A. P. MOLCHANOV, A. N. MICHEL, AND Y. SUN, *Converse theorems of the principal Lyapunov results for partial stability of general dynamical systems on metric spaces*, in Proceedings of the 42nd IEEE Conference on Decision and Control, 2003, pp. 5085–5090.
- [10] S. G. NERSESOV AND W. M. HADDAD, *On the stability and control of nonlinear dynamical systems via vector Lyapunov functions*, IEEE Trans. Automat. Control, 51 (2006), pp. 203–215.
- [11] C. F. PRICE, *An analysis of divergence problem in the Kalman filter*, IEEE Trans. Automat. Control, 13 (1968), pp. 699–702.
- [12] V. V. RUMYANTSEV, *On the stability of motion with respect to part of the variables*, Vestnik Moscov. Univ. Ser. Mat. Mekh. Fiz. Astron. Khim., 4 (1957), pp. 9–16.
- [13] S. SANGSUK-IAM AND T. E. BULLOCK, *Analysis of discrete-time Kalman filtering under incorrect noise covariances*, IEEE Trans. Automat. Control, 35 (1990), pp. 1304–1309.
- [14] V. I. VOROTNIKOV, *Partial Stability and Control*, Birkhäuser, Boston, 1998.
- [15] J. L. WILLEMS AND F. M. CALLIER, *Divergence of the stationary Kalman filter for correct and for incorrect noise variances*, IMA J. Math. Control Inform., 9 (1992), pp. 47–54.

## THE CONTROL VARIATIONAL APPROACH FOR DIFFERENTIAL SYSTEMS\*

JÜRGEN SPREKELS<sup>†</sup> AND DAN TIBA<sup>‡</sup>

**Abstract.** In this work a new approach to generalized Naghdi shell and curved rod models is discussed. The method is based on optimal control theory and has a wide range of applications. Some abstract variants are also indicated.

**Key words.** curved rods, shells, variational methods, generalized Naghdi model, optimal control

**AMS subject classifications.** 49N90, 74K10, 74K25, 49S05

**DOI.** 10.1137/070710561

**1. Introduction.** The control variational method is a new variational approach for the solution of differential equations. It is based on the minimization of the usual energy via optimal control arguments. The method was introduced in a sequence of papers by Arnăutu et al. [1] and Sprekels and Tiba [8], [10] and has many variants. Due to its relationship to the energy functional, we consider that this approach is a generalization of the classical variational argument. We mention that already Glowinski and Pironneau [4], [3] noticed that for the biharmonic equation optimal control theory may be advantageously applied in the numerical approximation. The recent monograph of Neittaanmäki, Sprekels, and Tiba [7] includes a detailed discussion of the control variational approach in the case of Kirchhoff–Love arches and for simplified plate models. The method is very flexible and allows one to obtain new results, on a theoretical as well as a numerical level. We briefly recall here the example of the linear elasticity system, following Sprekels and Tiba [12]. If  $\Omega \subset \mathbb{R}^3$  is a bounded domain representing the elastic body, fixed on  $\Gamma \subset \partial\Omega$ , and if

$$V(\Omega) = \{\bar{v} \in H^1(\Omega)^3; \bar{v}|_{\Gamma} = 0\},$$

then the weak formulation of the isotropic linear elasticity system is

$$(1.1) \quad \int_{\Omega} [\lambda e_{pp}(\bar{u}) e_{qq}(\bar{v}) + 2\mu e_{ij}(\bar{u}) e_{ij}(\bar{v})] dx = \int_{\Omega} f_i v_i dx \quad \forall \bar{v} \in V(\Omega).$$

Here,  $\bar{u} = (u_1, u_2, u_3) \in V(\Omega)$  is the displacement vector,  $\bar{f} = (f_1, f_2, f_3) \in L^2(\Omega)^3$  is the load vector,  $\lambda \geq 0$ ,  $\mu > 0$  are the Lamé coefficients of the elastic material,  $e_{ij}(\cdot)$ ,  $i, j = \overline{1, 3}$ , denote the symmetrized gradients, and the summation convention is used.

---

\*Received by the editors December 11, 2007; accepted for publication (in revised form) September 5, 2008; published electronically January 7, 2009.  
<http://www.siam.org/journals/sicon/47-6/71056.html>

<sup>†</sup>Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin, Germany (sprekels@wias-berlin.de).

<sup>‡</sup>Institute of Mathematics, Romanian Academy, P.O. Box 1–764, 014700 Bucharest, Romania (dan.tiba@imar.ro). The work of this author was supported by grant 2-CEX06-11-18/2006 of ANCS, Romania.



The optimal control problem

$$(P) \quad \text{Min} \left\{ \int_{\Omega} \left\{ \mu |w|_{\mathbb{R}^9}^2 + \lambda |\operatorname{div} u|^2 + \mu \left[ \left( \frac{\partial u_1}{\partial x_1} \right)^2 + \left( \frac{\partial u_2}{\partial x_2} \right)^2 + \left( \frac{\partial u_3}{\partial x_3} \right)^2 \right] \right. \right. \\ \left. \left. + 2\mu \left( \frac{\partial u_1}{\partial x_2} \frac{\partial u_2}{\partial x_1} + \frac{\partial u_1}{\partial x_3} \frac{\partial u_3}{\partial x_1} + \frac{\partial u_2}{\partial x_3} \frac{\partial u_3}{\partial x_2} \right) \right\} dx \right\}$$

for all  $w \in L^2(\Omega)^9$  is associated to (1.1), subject to the state equation

$$(1.2) \quad \int_{\Omega} \nabla \bar{u} : \nabla \bar{v} dx = \int_{\Omega} w : \nabla \bar{v} dx + \frac{1}{\mu} \int_{\Omega} \bar{f} \cdot \bar{v} dx \quad \forall \bar{v} \in V(\Omega).$$

Here,  $\nabla \bar{u}$ ,  $\nabla \bar{v}$  denote the Jacobian matrices, and

$$\nabla \bar{u} : \nabla \bar{v} = \sum_{i,j=1}^3 \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j}, \quad f \cdot v = \sum_{i=1}^3 f_i v_i.$$

Problem (P) is an unconstrained optimal control problem whose state system (1.2) consists of three uncoupled Poisson equations for  $\{u_i\}_{i=1,3}$ . Therefore, problem (P) is particularly easy to solve and, according to Sprekels and Tiba [12], its optimal state  $u^*$  is the unique solution to (1.1).

In this paper, we demonstrate similar properties in the application of the control variational method to shell and curved rod models of generalized Naghdi type (see sections 2 and 3, respectively). The last section discusses some abstract variants of the method. Further examples and applications may be found in the preprints of Sprekels and Tiba [13] and Tiba [14], on which this paper is based.

**2. Naghdi shell models.** We first introduce the generalized Naghdi model for thin elastic shells, following Sprekels and Tiba [11]. Let  $\omega \in \mathbb{R}^2$  be a bounded Lipschitzian domain,  $\varepsilon > 0$  “small,” and  $\Omega = \omega \times ]-\varepsilon, \varepsilon[$ . We assume that  $\partial\omega = \bar{\gamma}_0 \cup \bar{\gamma}_1$ ,  $\gamma_0 \cap \gamma_1 = \emptyset$ , and we denote

$$\Gamma_0 = \gamma_0 \times ]-\varepsilon, \varepsilon[, \quad \Gamma_1 = \partial\Omega \setminus \Gamma_0, \quad V(\omega) = \{\bar{v} = (v_1, v_2, v_3) \in H^1(\omega)^3; \bar{v}|_{\gamma_0} = 0\}.$$

Let  $p : \omega \rightarrow \mathbb{R}$  be piecewise in  $C^2(\bar{\omega})$ , and let  $\bar{n} : \omega \rightarrow \mathbb{R}^3$ ,  $\bar{n} = (n_1, n_2, n_3)$ , be the unit normal vector to the graph of  $p$  in  $\mathbb{R}^3$ . We denote by  $\bar{\pi} = (\pi_1, \pi_2, \pi_3) = (x_1, x_2, p(x_1, x_2))$  this graph, which represents the midsurface of the shell. That is,  $\bar{n}$  is given by

$$(2.1) \quad \bar{n} = \frac{\frac{\partial \bar{\pi}}{\partial x_1} \wedge \frac{\partial \bar{\pi}}{\partial x_2}}{\left| \frac{\partial \bar{\pi}}{\partial x_1} \wedge \frac{\partial \bar{\pi}}{\partial x_2} \right|_{\mathbb{R}^3}} = \left( -\frac{p_1}{\sqrt{1 + p_1^2 + p_2^2}}, -\frac{p_2}{\sqrt{1 + p_1^2 + p_2^2}}, \frac{1}{\sqrt{1 + p_1^2 + p_2^2}} \right),$$

where  $p_1 = \frac{\partial p}{\partial x_1}$ ,  $p_2 = \frac{\partial p}{\partial x_2}$ .

We define the transformation  $F : \Omega \rightarrow F(\Omega) = \hat{\Omega} \subset \mathbb{R}^3$  by

$$(2.2) \quad F(\bar{x}) = F(x_1, x_2, x_3) = \bar{\pi}(x_1, x_2) + x_3 \bar{n}(x_1, x_2).$$

If  $\varepsilon > 0$  is small enough, it is one-to-one (see Ciarlet [2, Thm. 3.1-1]). Denote by  $J = \nabla F$  the Jacobian of  $F$ , and let  $(h_{ij}(\bar{x}))_{i,j=1,3} = J(\bar{x})^{-1}$ . Starting from the linear elasticity system (1.1), and using the assumption that the displacement has the form

$$(2.3) \quad \hat{y}(\hat{x}) = \bar{u}(x_1, x_2) + x_3 \bar{r}(x_1, x_2) \quad \forall \hat{x} \in \hat{\Omega}, \quad \bar{x} = F^{-1}(\hat{x}) \in \Omega,$$

one can deduce the generalized Naghdi model

$$\begin{aligned}
 (2.4) \quad & B([\bar{u}, \bar{r}], [\bar{\mu}, \bar{\rho}]) \\
 &= \lambda \int_{\Omega} \left\{ \sum_{i=1}^3 \left[ \left( \frac{\partial u_i}{\partial x_1} + x_3 \frac{\partial r_i}{\partial x_1} \right) h_{1i} + \left( \frac{\partial u_i}{\partial x_2} + x_3 \frac{\partial r_i}{\partial x_2} \right) h_{2i} + r_i h_{3i} \right] \right\} \\
 &\quad \cdot \left\{ \sum_{j=1}^3 \left[ \left( \frac{\partial \mu_j}{\partial x_1} + x_3 \frac{\partial \rho_j}{\partial x_1} \right) h_{1j} + \left( \frac{\partial \mu_j}{\partial x_2} + x_3 \frac{\partial \rho_j}{\partial x_2} \right) h_{2j} + \rho_j h_{3j} \right] \right\} \\
 &\quad \cdot |\det J(\bar{x})| d\bar{x} \\
 &\quad + 2\mu \int_{\Omega} \sum_{i=1}^3 \left[ \left( \frac{\partial u_i}{\partial x_1} + x_3 \frac{\partial r_i}{\partial x_1} \right) h_{1i} + \left( \frac{\partial u_i}{\partial x_2} + x_3 \frac{\partial r_i}{\partial x_2} \right) h_{2i} + r_i h_{3i} \right] \\
 &\quad \cdot \left[ \left( \frac{\partial \mu_i}{\partial x_1} + x_3 \frac{\partial \rho_i}{\partial x_1} \right) h_{1i} + \left( \frac{\partial \mu_i}{\partial x_2} + x_3 \frac{\partial \rho_i}{\partial x_2} \right) h_{2i} + \rho_i h_{3i} \right] |\det J(\bar{x})| d\bar{x} \\
 &\quad + \mu \int_{\Omega} \sum_{i < j} \left\{ \left[ \left( \frac{\partial u_i}{\partial x_1} + x_3 \frac{\partial r_i}{\partial x_1} \right) h_{1j} + \left( \frac{\partial u_i}{\partial x_2} + x_3 \frac{\partial r_i}{\partial x_2} \right) h_{2j} + r_i h_{3j} \right. \right. \\
 &\quad \left. \left. + \left( \frac{\partial u_j}{\partial x_i} + x_3 \frac{\partial r_j}{\partial x_1} \right) h_{1i} + \left( \frac{\partial u_j}{\partial x_2} + x_3 \frac{\partial r_j}{\partial x_2} \right) h_{2i} + r_j h_{3i} \right] \right. \\
 &\quad \cdot \left[ \left( \frac{\partial \mu_i}{\partial x_1} + x_3 \frac{\partial \rho_i}{\partial x_1} \right) h_{1j} + \left( \frac{\partial \mu_i}{\partial x_2} + x_3 \frac{\partial \rho_i}{\partial x_2} \right) h_{2j} + \rho_i h_{3j} \right. \\
 &\quad \left. \left. + \left( \frac{\partial \mu_j}{\partial x_1} + x_3 \frac{\partial \rho_j}{\partial x_1} \right) h_{1i} + \left( \frac{\partial \mu_j}{\partial x_2} + x_3 \frac{\partial \rho_j}{\partial x_2} \right) h_{2i} + \rho_j h_{3i} \right] \right\} \\
 &\quad \cdot |\det J(\bar{x})| d\bar{x} \\
 &= \sum_{i=1}^3 \left\{ \int_{\Omega} f_i(\mu_i + x_3 \rho_i) d\bar{x} + \int_{\Gamma_1} h_i(\mu_i + x_3 \rho_i) d\sigma \right\}, \\
 &\quad \forall \bar{\mu} = (\mu_1, \mu_2, \mu_3), \quad \forall \bar{\rho} = (\rho_1, \rho_2, \rho_3) \in V(\omega).
 \end{aligned}$$

Above,  $\bar{f} = (f_1, f_2, f_3)$  represents the body forces and  $\bar{h} = (h_1, h_2, h_3)$  the surface tractions. The shell is partially clamped along  $\Gamma_0$ . The existence of a unique solution  $\bar{u}, \bar{r} \in V(\omega)$  follows from the Lax–Milgram lemma and Korn’s inequality; see Sprekels and Tiba [11].

As in the case of (1.1), (1.2), we show that it is possible to solve directly the generalized Naghdi shell model (2.4) via a control problem governed by a finite number of independent Poisson equations. Usual gradient methods may be applied for its solution. Our choice is motivated by its simplicity; other choices are also possible, as the following sections will make clear.

We associate with (2.4) the following unconstrained control problem:

$$(2.5) \quad \min_{w \in L^2(\omega)^{12}} \left\{ L(w) = \frac{1}{2} B([\bar{u}, \bar{r}], [\bar{u}, \bar{r}]) + \frac{1}{2} \sum_{i=1}^3 \int_{\omega} [|w_1^i|_{\mathbb{R}^2}^2 + |w_2^i|_{\mathbb{R}^2}^2] dx_1 dx_2 \right. \\ \left. - \frac{1}{2} \sum_{i=1}^3 \int_{\omega} [|\nabla u_i|_{\mathbb{R}^2}^2 + |\nabla r_i|_{\mathbb{R}^2}^2] dx_1 dx_2 \right\},$$

subject to

$$(2.6) \quad \sum_{i=1}^3 \int_{\omega} [\nabla u_i \cdot \nabla \phi_i + \nabla r_i \cdot \nabla \psi_i] dx_1 dx_2 \\ = \sum_{i=1}^3 \int_{\omega} [w_1^i \cdot \nabla \phi_i + w_2^i \cdot \nabla \psi_i] dx_1 dx_2 \\ + \sum_{i=1}^3 \left\{ \int_{\Omega} f_i(\phi_i + x_3 \psi_i) d\bar{x} + \int_{\Gamma_1} h_i(\phi_i + x_3 \psi_i) d\tau \right\} \quad \forall \phi, \psi \in V(\omega).$$

**THEOREM 2.1.** *The gradient of the cost (2.5) in the point  $w = [w_1, w_2] \in L^2(\omega)^{12}$  has the form*

$$(2.7) \quad \langle \nabla L(w_1, w_2), [v_1, v_2] \rangle = \sum_{i=1}^3 \int_{\omega} [\langle w_1^i + \nabla p_i, \nabla \mu_i \rangle_{\mathbb{R}^2} + \langle w_2^i + \nabla q_i, \nabla \rho_i \rangle_{\mathbb{R}^2}] dx_1 dx_2$$

for  $v_1^i = \nabla \mu_i$ ,  $v_2^i = \nabla \rho_i$ ,  $i = \overline{1, 3}$ , and any  $[\bar{\mu}, \bar{\rho}] \in V(\omega)^2$ .

Here,  $\langle \cdot, \cdot \rangle$  is the scalar product in  $L^2(\omega)^{12}$ , and  $[p, q] \in V(\omega)^2$  satisfy the adjoint equation

$$(2.8) \quad \sum_{i=1}^3 \int_{\omega} [\langle \nabla p_i, \nabla \phi_i \rangle_{\mathbb{R}^2} + \langle \nabla q_i, \nabla \psi_i \rangle_{\mathbb{R}^2}] dx_1 dx_2 \\ = \mathcal{B}([\bar{u}, \bar{r}], [\phi, \psi]) - \sum_{i=1}^3 \int_{\omega} [\langle \nabla u_i, \nabla \phi_i \rangle_{\mathbb{R}^2} + \langle \nabla r_i, \nabla \psi_i \rangle_{\mathbb{R}^2}] dx_1 dx_2 \quad \forall \phi, \psi \in V(\omega).$$

If  $[u^*, r^*] \in V(\omega)^2$  denotes the optimal state corresponding to the optimal control  $[w_1^*, w_2^*] \in L^2(\omega)^{12}$  for the problem (2.6), (2.5), then  $[u^*, r^*]$  is the unique solution to (2.4).

*Proof.* We take variations around the optimal pair, given by any  $[\bar{\mu}, \bar{\rho}] \in V(\omega)^2$  and  $v_1^i = \nabla \mu_i$ ,  $v_2^i = \nabla \rho_i$ ,  $i = \overline{1, 3}$ , for the state, respectively, the control. Clearly,  $[u^*, r^*] + \lambda[\bar{\mu}, \bar{\rho}]$  and  $w^* + \lambda[v_1, v_2]$ ,  $\lambda \in \mathbb{R}$ , satisfy (2.6) and give an admissible pair for the control problem (2.5). Then the optimality of  $w^*$  yields that

$$(2.9) \quad 0 = \lim_{\lambda \rightarrow 0} \frac{L(w^* + \lambda[v_1, v_2]) - L(w^*)}{\lambda} \\ = \mathcal{B}([u^*, r^*], [\bar{\mu}, \bar{\rho}]) + \sum_{i=1}^3 \int_{\omega} [\langle (w_1^*)^i, v_1^i \rangle_{\mathbb{R}^2} + \langle (w_2^*)^i, v_2^i \rangle_{\mathbb{R}^2}] \\ - \sum_{i=1}^3 \int_{\omega} [\langle \nabla u_i^*, \nabla \mu_i \rangle_{\mathbb{R}^2} + \langle \nabla r_i^*, \nabla \rho_i \rangle_{\mathbb{R}^2}] \\ = \mathcal{B}([u^*, r^*], [\bar{\mu}, \bar{\rho}]) - \sum_{i=1}^3 \left\{ \int_{\omega} f_i(\mu_i + x_3 \rho_i) d\bar{x} + \int_{\Gamma_1} h_i(\mu_i + x_3 \rho_i) d\sigma \right\},$$

by the state equation (2.5) with  $\bar{\mu}, \bar{\rho}$  as test functions, and by fixing  $v_1^i = \nabla \mu_i$ ,  $v_2^i = \nabla \rho_i$ ,  $i = \overline{1, 3}$ . The first and last terms in (2.9) prove the last statement of the theorem.

The first equality in (2.9) and the adjoint system (2.8) give

$$\begin{aligned} \langle \nabla L(w), [v_1, v_2] \rangle &= \sum_{i=1}^3 \int_{\omega} [\langle w_1^i, v_1^i \rangle_{\mathbb{R}^2} + \langle w_2^i, v_2^i \rangle_{\mathbb{R}^2}] \, dx_1 \, dx_2 \\ &\quad + \sum_{i=1}^3 \int_{\omega} [\langle \nabla p_i, \nabla \mu_i \rangle_{\mathbb{R}^2} + \langle \nabla q_i, \nabla \rho_i \rangle_{\mathbb{R}^2}] \, dx_1 \, dx_2, \end{aligned}$$

which shows (2.7).  $\square$

*Remark.* The optimality system characterizing the solution of (2.6), (2.5) is given by (2.6), (2.8), and by the Pontryagin maximum principle

$$\sum_{i=1}^3 \int_{\omega} [\langle (w_1^*)^i + \nabla p_i^*, \nabla \mu_i \rangle_{\mathbb{R}^2} + \langle (w_2^*)^i + \nabla q_i^*, \nabla \rho_i \rangle_{\mathbb{R}^2}] \, dx_1 \, dx_2 = 0,$$

where  $[p_i^*, q_i^*]_{i=\overline{1,3}}$  are computed by (2.7), with  $\bar{u}^*, \bar{r}^*$  on the right-hand side. Both (2.6) and (2.8) are equivalent to a system of six Poisson equations.

*Remark.* If state constraints are added to the control problem (2.6), (2.5), then the optimal state will satisfy a variational inequality associated with (2.4). In Arnăutu et al. [1], a similar situation was discussed in the case of a simplified plate model.

**PROPOSITION 2.2.** *The optimal control problem (2.6), (2.5) has a unique optimal pair  $[u^*, r^*] \in V(\omega)^2$ ,  $[w_1^*, w_2^*] \in L^2(\omega)^{12}$ .*

*Proof.* We introduce the auxiliary mappings

$$(2.10) \quad \tau_i = \nabla u_i - w_1^i \in L^2(\omega)^2, \quad i = \overline{1, 3},$$

$$(2.11) \quad \theta_i = \nabla r_i - w_2^i \in L^2(\omega)^2, \quad i = \overline{1, 3}.$$

Then

$$(2.12) \quad |w_1^i|_{\mathbb{R}^2}^2 = |\nabla u_i|_{\mathbb{R}^2}^2 + |\tau_i|_{\mathbb{R}^2}^2 - 2 \langle \nabla u_i, \tau_i \rangle_{\mathbb{R}^2}, \quad i = \overline{1, 3},$$

and similarly for  $w_2^i$ . Relations (2.10)–(2.12) show that the cost functional (2.5) may be rewritten in the form

$$\begin{aligned} (2.13) \quad L(w) &= \frac{1}{2} \mathcal{B}([\bar{u}, \bar{r}], [\bar{u}, \bar{r}]) + \frac{1}{2} \sum_{i=1}^3 \int_{\omega} [|w_1^i|_{\mathbb{R}^2}^2 + |w_2^i|_{\mathbb{R}^2}^2] \, dx_1 \, dx_2 \\ &\quad - \frac{1}{2} \sum_{i=1}^3 \int_{\omega} [|\nabla u_i|_{\mathbb{R}^2}^2 + |\nabla r_i|_{\mathbb{R}^2}^2] \, dx_1 \, dx_2 \\ &= \frac{1}{2} \mathcal{B}([\bar{u}, \bar{r}], [\bar{u}, \bar{r}]) + \frac{1}{2} \sum_{i=1}^3 \int_{\omega} [|\tau_i|_{\mathbb{R}^2}^2 + |\theta_i|_{\mathbb{R}^2}^2] \, dx_1 \, dx_2 \\ &\quad - \sum_{i=1}^3 \int_{\omega} [\langle \nabla u_i, \tau_i \rangle_{\mathbb{R}^2} + \langle \nabla r_i, \theta_i \rangle_{\mathbb{R}^2}] \, dx_1 \, dx_2 \\ &= \frac{1}{2} \mathcal{B}([\bar{u}, \bar{r}], [\bar{u}, \bar{r}]) + \frac{1}{2} \sum_{i=1}^3 \int_{\omega} [|\tau_i|_{\mathbb{R}^2}^2 + |\theta_i|_{\mathbb{R}^2}^2] \, dx_1 \, dx_2 \\ &\quad - \sum_{i=1}^3 \left\{ \int_{\Omega} f_i(u_i + x_3 r_i) \, d\bar{x} + \int_{\Gamma_1} h_i(u_i + x_3 r_i) \, d\sigma \right\}, \end{aligned}$$

where we also use (2.6) with the test functions  $\phi_i = u_i$ ,  $\psi_i = r_i$ ,  $i = \overline{1, 3}$ .

If  $\{w_1^n, w_2^n\} \in L^2(\omega)^{12}$  is a minimizing sequence for the problem (2.6), (2.5), and  $\{\bar{u}_n, \bar{r}_n\} \in V(\omega)^2$  are the corresponding states, then the values of the associated cost are bounded from above by a constant, and (2.13) shows that  $\{\bar{\tau}^n\}, \{\bar{\theta}^n\}$  are bounded in  $L^2(\omega)^6$  and  $\{\bar{u}_n\}, \{\bar{r}_n\}$  are bounded in  $V(\omega)$ . Here, we also used that  $\mathcal{B}(\cdot, \cdot)$  is coercive in  $V(\omega)^2$  (cf. Sprekels and Tiba [11]) and that the linear terms in (2.13) are dominated by  $\mathcal{B}(\cdot, \cdot)$ .

Denoting by  $\tilde{\tau}, \tilde{\theta}, \tilde{u}, \tilde{r}, \tilde{w}_1, \tilde{w}_2$  the weak limits of the above quantities, on a suitable subsequence, we can pass to the limit in (2.6), (2.11), (2.10) to obtain similar relations between the limit points. In particular,  $\{\tilde{w}_1, \tilde{w}_2\} \in L^2(\omega)^{12}$  and  $\{\tilde{u}, \tilde{r}\} \in V(\omega)^2$  define an admissible pair for the control problem (2.6), (2.5). By (2.13), and owing to the weak lower semicontinuity of convex functionals, we can infer that

$$(2.14) \quad \begin{aligned} & \liminf_{n \rightarrow \infty} L([\bar{u}_n, \bar{r}_n], [w_1^n, w_2^n]) \\ & \geq \frac{1}{2} \mathcal{B}([\tilde{u}, \tilde{r}], [\tilde{u}, \tilde{r}]) + \frac{1}{2} \sum_{i=1}^3 \int_{\omega} \left[ |\tilde{\tau}_i|_{\mathbb{R}^2}^2 + |\tilde{\theta}_i|_{\mathbb{R}^2}^2 \right] dx_1 dx_2 \\ & \quad - \sum_{i=1}^3 \left\{ \int_{\Omega} f_i(\tilde{u}_i + x_3 \tilde{r}_i) d\bar{x} + \int_{\Gamma_1} h_i(\tilde{u}_i + x_3 \tilde{r}_i) d\sigma \right\}. \end{aligned}$$

By virtue of (2.14) and (2.13) (used in the converse sense), we deduce the optimality of  $[\tilde{w}_1, \tilde{w}_2]$ , which we redenote by  $[w_1^*, w_2^*]$ . Uniqueness is a consequence of the strict convexity of  $\mathcal{B}(\cdot, \cdot)$  and of (2.13).  $\square$

*Remark.* The form (2.13) of the cost functional also shows the connections with the energy minimization corresponding to the Naghdi shell model, as mentioned in the introduction.

**3. Curved rods.** The application to curved rod models exploits essentially the structure of the coefficients of the differential equations. We first indicate some details in this respect. To this end, denote by  $\bar{\theta} \in W^{2,\infty}(0, L)^3$ ,  $L > 0$ , the parametrization of a three-dimensional Jordan curve, which will represent the line of centroids of the curved rod. Let  $\omega \subset \mathbb{R}^2$  be some bounded Lipschitzian domain, not necessarily simply connected.

If  $\bar{t}, \bar{n}, \bar{b}$  denote the local orthonormal frame associated at each point  $x_3 \in [0, L]$  with the curve  $\bar{\theta}$ , we consider the geometric transformation

$$(3.1) \quad \begin{aligned} F : \Omega = \omega \times ]0, L[ & \rightarrow F(\Omega) = \hat{\Omega} \subset \mathbb{R}^3, \\ F(\bar{x}) = F(x_1, x_2, x_3) & = \bar{\theta}(x_3) + x_1 \bar{n}(x_3) + x_2 \bar{b}(x_3) \\ & \quad \forall (x_1, x_2) \in \Omega, \forall x_3 \in ]0, L[. \end{aligned}$$

In the following, the Jacobian of  $F$ ,  $J = \nabla F$ , and its inverse,  $J(\bar{x})^{-1} = (h_{ij}(\bar{x}))_{i,j=\overline{1,3}}$ , are needed. Elementary computations give

$$(3.2) \quad J(\bar{x})^{-1} = \begin{pmatrix} n_1 - \frac{c t_1 x_2}{\det J(\bar{x})} & n_2 - \frac{c t_2 x_2}{\det J(\bar{x})} & n_3 - \frac{c t_3 x_2}{\det J(\bar{x})} \\ b_1 + \frac{c t_1 x_1}{\det J(\bar{x})} & b_2 + \frac{c t_2 x_1}{\det J(\bar{x})} & b_3 + \frac{c t_3 x_1}{\det J(\bar{x})} \\ \frac{t_1}{\det J(\bar{x})} & \frac{t_2}{\det J(\bar{x})} & \frac{t_3}{\det J(\bar{x})} \end{pmatrix},$$

$$(3.3) \quad \det J(\bar{x}) = 1 - \beta x_1 - a x_2.$$

The coefficients  $a, \beta, c \in L^\infty(0, L)$  appearing in (3.2), (3.3) are similar to torsion and curvature known from classical differential geometry and may be obtained from the “equations of motion”

$$(3.4) \quad \begin{aligned} \bar{t}'(x_3) &= a(x_3) \bar{b}(x_3) + \beta(x_3) \bar{n}(x_3), \\ \bar{b}'(x_3) &= -a(x_3) \bar{t}(x_3) + c(x_3) \bar{n}(x_3), \\ \bar{n}'(x_3) &= -\beta(x_3) \bar{t}(x_3) - c(x_3) \bar{b}(x_3). \end{aligned}$$

We note that  $\bar{t}, \bar{n}, \bar{b}$  are not necessarily obtained as the Frenet or the Darboux local frames associated with  $\bar{\theta}$ . In Neittaanmäki, Sprekels, and Tiba [7, Chap. 6] such a local frame was constructed under the mere assumption that  $\bar{\theta} \in C^1[0, L]^3$ .

We also assume that the selection of axes in  $\omega \subset \mathbb{R}^2$  is such that

$$(3.5) \quad 0 = \int_{\omega} x_1 dx_1 dx_2 = \int_{\omega} x_2 dx_1 dx_2 = \int_{\omega} x_1 x_2 dx_1 dx_2,$$

which is usual in the literature on curved rods; cf. Murat and Sili [6]. If the diameter of  $\omega$  is sufficiently small, then  $|x_1|, |x_2|$  are small for  $(x_1, x_2) \in \omega$ , and since  $\beta, a \in L^\infty(0, L)$ , relation (3.3) shows that we may assume

$$(3.6) \quad \det J(\bar{x}) \geq c_0 > 0 \quad \forall \bar{x} \in \Omega.$$

Then it is known that  $F : \Omega \rightarrow \hat{\Omega}$  is a one-to-one transformation (see Ciarlet [2, Thm. 3.1.–1]), which justifies the definition (3.1) of the curved rod.

Starting from the linear elasticity system (1.1), and using the condition that the displacement of  $\hat{x} \in \hat{\Omega}$  has the form

$$(3.7) \quad \bar{y}(\hat{x}) = \bar{\tau}(x_3) + x_1 \bar{N}(x_3) + x_2 \bar{B}(x_3),$$

with  $\bar{x} = (x_1, x_2, x_3) = F^{-1}(\hat{x}) \in \Omega$ , we obtain the following model for the curved rod:

$$(3.8) \quad \begin{aligned} \mathcal{B}(\bar{y}, \bar{v}) &= \tilde{\lambda} \int_{\Omega} \sum_{i,j=1}^3 \left[ N_i(x_3) h_{1i}(\bar{x}) + B_i(x_3) h_{2i}(\bar{x}) + (\tau'_i(x_3) + x_1 N'_i(x_3) \right. \\ &\quad \left. + x_2 B'_i(x_3)) h_{3i}(\bar{x}) \right] \cdot \left[ M_j(x_3) h_{1j}(\bar{x}) + D_j(x_3) h_{2j}(\bar{x}) \right. \\ &\quad \left. + (\mu'_j(x_3) + x_1 M'_j(x_3) + x_2 D'_j(x_3)) h_{3j}(\bar{x}) \right] |\det J(\bar{x})| d\bar{x} \\ &\quad + \tilde{\mu} \int_{\Omega} \sum_{i < j} \left[ N_i(x_3) h_{1j}(\bar{x}) + B_i(x_3) h_{2j}(\bar{x}) + (\tau'_i(x_3) + x_1 N'_i(x_3) \right. \\ &\quad \left. + x_2 B'_i(x_3)) h_{3j}(\bar{x}) + N_j(x_3) h_{1i}(\bar{x}) + B_j(x_3) h_{2i}(\bar{x}) \right. \\ &\quad \left. + (\tau'_j(x_3) + x_1 N'_j(x_3) + x_2 B'_j(x_3)) h_{3i}(\bar{x}) \right] \\ &\quad \cdot \left[ M_i(x_3) h_{1j}(\bar{x}) + D_i(x_3) h_{2j}(\bar{x}) + (\mu'_i(x_3) + x_1 M'_i(x_3) \right. \\ &\quad \left. + x_2 D'_i(x_3)) h_{3j}(\bar{x}) + M_j(x_3) h_{1i}(\bar{x}) + D_j(x_3) h_{2i}(\bar{x}) \right. \\ &\quad \left. + (\mu'_j(x_3) + x_1 M'_j(x_3) + x_2 D'_j(x_3)) h_{3i}(\bar{x}) \right] |\det J(\bar{x})| d\bar{x} \end{aligned}$$

$$\begin{aligned}
& + 2 \tilde{\mu} \int_{\Omega} \sum_{i=1}^3 \left[ N_i(x_3) h_{1i}(\bar{x}) + B_i(x_3) h_{2i}(\bar{x}) + (\tau'_i(x_3) + x_1 N'_i(x_3) \right. \\
& \quad \left. + x_2 B'_i(x_3)) h_{3i}(\bar{x}) \right] \cdot \left[ M_i(x_3) h_{1i}(\bar{x}) + D_i(x_3) h_{2i}(\bar{x}) \right. \\
& \quad \left. + (\mu'_i(x_3) + x_1 M'_i(x_3) + x_2 D'_i(x_3)) h_{3i}(\bar{x}) \right] |\det J(\bar{x})| d\bar{x} \\
& = \sum_{\ell=1}^3 \int_{\Omega} f_{\ell}(\bar{x}) (\mu_{\ell}(x_3) + x_1 M_{\ell}(x_3) + x_2 D_{\ell}(x_3)) |\det J(\bar{x})| d\bar{x}
\end{aligned}$$

for any test functions  $\bar{\mu} = (\mu_1, \mu_2, \mu_3)$ ,  $\bar{M} = (M_1, M_2, M_3)$ ,  $\bar{D} = (D_1, D_2, D_3) \in H_0^1(0, L)^3$ . We denote by  $\bar{v} = (\bar{\mu}, \bar{M}, \bar{D}) \in H_0^1(0, L)^9$  and  $\bar{y} = (\tau_1, \tau_2, \tau_3, N_1, N_2, N_3, B_1, B_2, B_3) \in H_0^1(0, L)^9$  the vector of the unknowns. The bilateral null conditions, given by the choice of the space  $H_0^1(0, L)$ , correspond to the clamped curved rod.

The condition (3.7) and the choice of the test functions in a similar form are of the same type as (2.3), and that is why the model (3.8) was called a generalized Naghdi model for curved rods in [5] where it was studied. It consists of a system of nine ordinary differential equations for nine unknown functions.

It should be clear that  $\bar{\tau}$  describes the translation of the points on the line of centroids, and the vectors  $\bar{N} + \bar{n}$ ,  $\bar{B} + \bar{b}$  reflect the deformation of the orthogonal frame in the cross section (which remains plane but not necessarily orthogonal to the tangent of the deformed line of centroids, i.e., to  $\bar{\theta}' + \bar{\tau}'$ ). This allows for shear and for length or volume changes after the deformation. The vector  $\bar{f} = (f_1, f_2, f_3) \in L^2(0, L)^3$  represents the body forces acting on the curved rod, and  $\bar{\lambda} \geq 0$ ,  $\bar{\mu} > 0$  are the Lamé coefficients that characterize the elastic material. Note that the model is valid only for small deformations and for thin rods. As a special type of argument concerning the validity of such models and their stability with respect to iterative approaches, we quote [7, Chap. 6.1.2] (for Kirchhoff–Love arches) and [7, Chap. 6.2.3] (for three-dimensional curved rods). Especially relevant in this respect are [7, Ex. 6.1.11 and 6.2.13], where some experiments perfectly matching the physical interpretation are discussed for these models.

We associate with (3.8) the following optimal control problem with the basic properties (to be argued in what follows) that it solves (3.8) and has a very simple structure:

$$\begin{aligned}
(3.9) \quad \min \bigg\{ & \tilde{\lambda} \int_{\Omega} \sum_{i,j=1}^3 U_{ii}(\bar{x}) U_{jj}(\bar{x}) |\det J(\bar{x})| d\bar{x} \\
& + \tilde{\mu} \int_{\Omega} \sum_{i < j} [U_{ij}(\bar{x}) + U_{ji}(\bar{x})]^2 |\det J(\bar{x})| d\bar{x} \\
& + 2 \tilde{\mu} \int_{\Omega} \sum_{i=1}^3 U_{ii}^2(\bar{x}) |\det J(\bar{x})| d\bar{x} - 2 \sum_{i=1}^3 \int_{\Omega} f_i(\bar{x}) [\tau_i(x_3) + x_1 N_i(x_3) \\
& \quad + x_2 B_i(x_3)] |\det J(\bar{x})| d\bar{x} \bigg\},
\end{aligned}$$

subject to the state system

$$\begin{aligned}
(3.10) \quad & N_i(x_3) h_{1j}(\bar{x}) + B_i(x_3) h_{2j}(\bar{x}) + [\tau'_i(x_3) + x_1 N'_i(x_3) \\
& \quad + x_2 B'_i(x_3)] h_{3j}(\bar{x}) = U_{ij}(\bar{x}) \text{ in } \Omega,
\end{aligned}$$

$$(3.11) \quad N_i(0) = B_i(0) = \tau_i(0) = 0, \quad i = \overline{1, 3},$$

and to the control constraints

$$(3.12) \quad U = \{U_{ij}\}_{i,j=\overline{1,3}} \in \mathcal{V} \subset L^2(\Omega)^9.$$

Here,  $\mathcal{V} \neq \emptyset$  is the closed linear subspace in  $L^2(\Omega)^9$  generated from all functions in  $L^2(0, L)$  having zero mean value, used on the “position” of  $\tau'_i, N'_i, B'_i$ ,  $i = \overline{1, 3}$ , in (3.10). Clearly,  $N_i, B_i$  can be immediately computed using (3.11) and simple integration, which gives  $U_{ij}$  on the right-hand side of (3.10) and spans  $\mathcal{V}$ .

Note that in (3.10),  $(x_1, x_2) \in \omega$  appears as a parameter. The constraint (3.12) ensures that (3.10) has a unique solution and that

$$(3.13) \quad \tau_i(L) = N_i(L) = B_i(L) = 0, \quad i = \overline{1, 3}.$$

One could impose (3.13) instead of (3.12), but our choice is motivated by its explicit character with respect to the control unknown.

**THEOREM 3.1.** *The optimal control problem (3.9)–(3.12) has a unique optimal “pair”  $U^* = \{U_{ij}^*\} \in \mathcal{V}$ ,  $[\tau_i^*, B_i^*, N_i^*]_{i=\overline{1,3}} \in H_0^1(0, L)^9$ .*

*Proof.* Let  $L(U)$  denote the cost functional (3.9). Then there are  $\alpha > 0$ ,  $\bar{c} > 0$  such that

$$(3.14)$$

$$\alpha L(U) \geq \int_{\Omega} \sum_{i < j} [U_{ij}(\bar{x}) + U_{ji}(\bar{x})]^2 d\bar{x} + \int_{\Omega} \sum_{i=1}^3 U_{ii}^2(\bar{x}) - \bar{c} \sum_{i=1}^3 \int_0^L [\tau_i^2 + N_i^2 + B_i^2]^{\frac{1}{2}} dx_3,$$

where (3.3) and simple binomial inequalities have been used. Using (3.10) and (3.14), and again some binomial inequalities, we obtain (where we denote  $z_i = \tau'_i + x_1 N'_i + x_2 B'_i$ )

$$\begin{aligned} \alpha L(U) &\geq \int_{\Omega} \sum_{i < j} [N_i h_{1j} + B_i h_{2j} + z_i h_{3j} + N_j h_{1i} + B_j h_{2i} + z_j h_{3i}]^2 d\bar{x} \\ (3.15) \quad &+ \int_{\Omega} \sum_{i=1}^3 [N_i h_{1i} + B_i h_{2i} + z_i h_{3i}]^2 d\bar{x} - \bar{c} \sum_{i=1}^3 \int_0^L [\tau_i^2 + N_i^2 + B_i^2]^{\frac{1}{2}} dx_3 \\ &\geq \int_{\Omega} \sum_{i < j} (z_i h_{3j} + z_j h_{3i})^2 d\bar{x} + \int_{\Omega} \sum_{i=1}^3 z_i^2 h_{3i}^2 d\bar{x} - \hat{c} \sum_{i=1}^3 \int_0^L [N_i^2 + B_i^2] dx_3 \\ &\quad - \bar{c} \sum_{i=1}^3 \int_0^L [\tau_i^2 + N_i^2 + B_i^2]^{\frac{1}{2}} dx_3. \end{aligned}$$

Above, in the structure of the constant  $\hat{c} > 0$ , we also used that  $h_{ij} \in L^\infty(\Omega)$ ,  $i, j = \overline{1, 3}$ . In (3.15), we applied the following identities:

$$(3.16)$$

$$\frac{1}{2} \sum_{i < j} (z_i h_{3j} + z_j h_{3i})^2 + \frac{3}{2} \sum_{i=1}^3 z_i^2 h_{3i}^2 = \frac{1}{2} \sum_{i=1}^3 z_i^2 \sum_{j=1}^3 h_{3j}^2 + \frac{1}{2} \sum_{i < j} (z_i h_{3i} + z_j h_{3j})^2,$$



$$\begin{aligned}
 (3.17) \quad & \int_{\Omega} z_i^2 d\bar{x} = \int_{\Omega} [\tau'_i + x_1 N'_i + x_2 B'_i]^2 d\bar{x} \\
 &= \int_0^L \int_{\omega} (\tau'_i)^2 d\bar{x} + 2 \int_0^L \tau'_i N'_i dx_3 \int_{\omega} x_1 dx_1 dx_2 + 2 \int_0^L \tau'_i B'_i dx_3 \int_{\omega} x_2 dx_1 dx_2 \\
 &\quad + \int_0^L (N'_i)^2 dx_3 \int_{\omega} x_1^2 dx_1 dx_2 + \int_0^L (B'_i)^2 dx_3 \int_{\omega} x_2^2 dx_1 dx_2 \\
 &\quad + 2 \int_0^L N'_i B'_i dx_3 \int_{\omega} x_1 x_2 dx_1 dx_2 \geq \tilde{c} \int_0^L [(\tau'_i)^2 + (N'_i)^2 + (B'_i)^2] dx_3,
 \end{aligned}$$

with  $\tilde{c} := \min\{\text{meas}(\omega), \int_{\omega} x_1^2 dx_1 dx_2, \int_{\omega} x_2^2 dx_1 dx_2\} > 0$ ,

$$(3.18) \quad \sum_{j=1}^3 h_{3j}^2 = \frac{1}{\det J(\bar{x})} \sum_{j=1}^3 t_j^2 = \frac{1}{\det J(\bar{x})} \geq \check{c} > 0.$$

Relations (3.17), (3.18) are consequences of (3.5), respectively, of (3.2), (3.3), and of the assumption  $\bar{\theta} \in W^{2,\infty}(0, L)^3$ . From (3.15)–(3.18), we can infer that

$$\begin{aligned}
 (3.19) \quad & \alpha L(U) + \bar{c} \sum_{i=1}^3 \int_0^L [\tau_i^2 + N_i^2 + B_i^2]^{\frac{1}{2}} dx_3 \\
 & \geq \hat{c} \sum_{i=1}^3 \left[ |\tau_i|_{H_0^1(0,L)}^2 + |N_i|_{H_0^1(0,L)}^2 + |B_i|_{H_0^1(0,L)}^2 \right] \\
 & \quad - \hat{C} \sum_{i=1}^3 \left[ |N_i|_{L^2(0,L)}^2 + |B_i|_{L^2(0,L)}^2 \right]^2,
 \end{aligned}$$

where  $\hat{c}, \hat{C}, \bar{c}$  are positive constants that do not depend on  $\tau_i, N_i, B_i, U_{ij}$ ,  $i, j = \overline{1, 3}$ .  $\square$

LEMMA 3.2. *There is some  $\delta > 0$  such that*

$$\begin{aligned}
 (3.20) \quad & \alpha L(U) + \bar{c} \sum_{i=1}^3 \int_0^L [\tau_i^2 + N_i^2 + B_i^2]^{\frac{1}{2}} dx_3 \\
 & \geq \delta \sum_{i=1}^3 \left[ |\tau_i|_{H_0^1(0,L)}^2 + |N_i|_{H_0^1(0,L)}^2 + |B_i|_{H_0^1(0,L)}^2 \right]
 \end{aligned}$$

for all  $\tau_i, N_i, B_i \in H_0^1(0, L)$ ,  $i = \overline{1, 3}$ , obtained by (3.10) from some  $\{U_{ij}\}_{i,j=\overline{1,3}} \in \mathcal{V}$ .

*Proof.* Owing to (3.14), we have

$$\alpha L(U) + \bar{c} \sum_{i=1}^3 \int_0^L [\tau_i^2 + N_i^2 + B_i^2]^{\frac{1}{2}} dx_3 \geq 0.$$

Assume that (3.20) is false. Then for any  $\varepsilon > 0$  there are  $U^\varepsilon = \{U_{ij}^\varepsilon\}_{i,j=\overline{1,3}} \in \mathcal{V}$  and

$\{\tau_i^\varepsilon, N_i^\varepsilon, B_i^\varepsilon\}_{i=\overline{1,3}} \neq 0$ , associated with  $U_\varepsilon$  by (3.10), such that

$$\begin{aligned}
 (3.21) \quad & \varepsilon \sum_{i=1}^3 \left[ |\tau_i^\varepsilon|_{H_0^1(0,L)}^2 + |N_i^\varepsilon|_{H_0^1(0,L)}^2 + |B_i^\varepsilon|_{H_0^1(0,L)}^2 \right] \\
 & \geq \alpha L(U^\varepsilon) + \bar{c} \sum_{i=1}^3 \int_0^L [(\tau_i^\varepsilon)^2 + (N_i^\varepsilon)^2 + (B_i^\varepsilon)^2]^{\frac{1}{2}} dx_3 \\
 & \geq \int_\Omega \sum_{i < j} [U_{ij}^\varepsilon(\bar{x}) + U_{ji}^\varepsilon(\bar{x})]^2 d\bar{x} + \int_\Omega \sum_{i=1}^3 (U_{ii}^\varepsilon)^2 d\bar{x} \geq 0.
 \end{aligned}$$

Notice that we may assume that

$$\sum_{i=1}^3 \left[ |\tau_i^\varepsilon|_{H_0^1(0,L)}^2 + |N_i^\varepsilon|_{H_0^1(0,L)}^2 + |B_i^\varepsilon|_{H_0^1(0,L)}^2 \right] = 1$$

by scaling with the square root of this factor (if it differs from unity) in (3.10), and in the first and last terms of (3.21). Then we may assume without loss of generality that

$$\tau_i^\varepsilon \rightarrow \tau_i, \quad N_i^\varepsilon \rightarrow N_i, \quad B_i^\varepsilon \rightarrow B_i, \quad i = \overline{1,3},$$

weakly in  $H_0^1(0, L)$  and strongly in  $L^2(0, L)$ . By virtue of (3.10), we also see that  $U_{ij}^\varepsilon \rightarrow U_{ij}$  weakly in  $L^2(\Omega)$ . All the above convergences are valid for some subsequence. Also,  $\{\tau_i, N_i, B_i\}_{i=\overline{1,3}}$ , together with  $\{U_{ij}\}_{i,j=\overline{1,3}}$ , satisfy (3.10), while  $\{U_{ij}\} \in \mathcal{V}$ .

Passing to the limit in (3.21), we find that

$$(3.22) \quad \int_\Omega \sum_{i < j} [U_{ij}(\bar{x}) + U_{ji}(\bar{x})]^2 dx + \int_\Omega \sum_{i=1}^3 (U_{ii})^2 d\bar{x} = 0.$$

From (3.22) it follows that

$$(3.23) \quad N_i h_{1i} + B_i h_{2i} + z_i h_{3i} = 0, \quad i = \overline{1,3},$$

$$(3.24) \quad N_i h_{1j} + B_i h_{2j} + z_i h_{3j} + N_j h_{1i} + B_j h_{2i} + z_j h_{3i} = 0, \quad i \neq j.$$

Now fix  $j = j_0$  in (3.24), and let  $i_1, i_2$  be the two possible choices of indices  $i$  satisfying the condition in (3.24). We multiply (3.23), written for  $i = i_1$  or  $i = i_2$ , by  $h_{3j_0}$ , and subtract the result from (3.24) multiplied by  $h_{3i_1}$  or  $h_{3i_2}$ , respectively. Adding the results to (3.23) written for  $i = j$ , and multiplied by  $h_{3j_0}$ , we find that

$$(3.25) \quad z_{j_0} \sum_{i=1}^3 h_{3i}^2 = \tilde{\Gamma}_{j_0}(\bar{N}, \bar{B}),$$

where  $\tilde{\Gamma}_{j_0}$  is some linear expression of  $\bar{N}, \bar{B}$ . By (3.2), we obtain from (3.25) that

$$(3.26) \quad \tau'_i + x_1 N'_i + x_2 B'_i = \Gamma_i(\bar{N}, \bar{B}), \quad i = \overline{1,3},$$

where, again,  $\Gamma_i$  is linear in  $\bar{N}, \bar{B}$ .

Giving  $(x_1, x_2) \in \omega$  several independent values and taking into account (3.11), we conclude that all limit points  $\{\tau_i, N_i, B_i\}_{i=\overline{1,3}}$  are identically zero in their domains of definition. A similar conclusion follows for  $\{U_{ij}\}_{i,j=\overline{1,3}}$  by (3.10).

We combine the first inequality in (3.21) with (3.19):

$$\begin{aligned}
 (3.27) \quad & \varepsilon \sum_{i=1}^3 \left[ |\tau_i^\varepsilon|_{H_0^1(0,L)}^2 + |N_i^\varepsilon|_{H_0^1(0,L)}^2 + |B_i^\varepsilon|_{H_0^1(0,L)}^2 \right] \\
 & \geq \hat{c} \sum_{i=1}^3 \left[ |\tau_i^\varepsilon|_{H_0^1(0,L)}^2 + |N_i^\varepsilon|_{H_0^1(0,L)}^2 + |B_i^\varepsilon|_{H_0^1(0,L)}^2 \right] \\
 & \quad - \hat{C} \sum_{i=1}^3 \left[ |N_i^\varepsilon|_{L^2(0,L)}^2 + |B_i^\varepsilon|_{L^2(0,L)}^2 \right].
 \end{aligned}$$

Using the above scaling argument and its conclusion in (3.27), we arrive at the contradiction

$$0 \geq \hat{c} > 0,$$

since  $|N_i^\varepsilon|_{L^2(0,L)}^2 + |B_i^\varepsilon|_{L^2(0,L)}^2 \rightarrow 0$ . This completes the proof of (3.20).  $\square$

*Proof of Theorem 3.1 (continued).* Let  $\{U_{ij}^n\}_{i,j=\overline{1,3}}$ ,  $\{\tau_i^n, N_i^n, B_i^n\}_{i,j=\overline{1,3}}$  be a minimizing sequence in  $\mathcal{V} \times H_0^1(0,L)^9$  for the problem (3.9)–(3.12). Clearly,  $L(U^n)$  is majorized from above by a constant, and inequality (3.20) shows that  $\{\tau_i^n, N_i^n, B_i^n\}_{i=\overline{1,3}}$  is bounded in  $H_0^1(0,L)^9$ . Consequently, by (3.10),  $\{U_{ij}^n\}_{i,j=\overline{1,3}}$  are bounded in  $L^2(\Omega)^9$ . Let  $\{\tau_i^*, N_i^*, B_i^*\}_{i=\overline{1,3}}$  and  $\{U_{ij}^*\}_{i,j=\overline{1,3}}$  denote, respectively, their weak limits, on a subsequence. Then  $\{U_{ij}^*\}_{i,j=\overline{1,3}} \in \mathcal{V}$ , which is a closed linear space.

One can pass to the limit in (3.10) to see that  $\{\tau_i^*, N_i^*, B_i^*\}_{i=\overline{1,3}}$  and  $\{U_{ij}^*\}_{i,j=\overline{1,3}}$  form an admissible couple, and then, using the weak lower semicontinuity of cost functional (3.9), to conclude their optimality for the problem (3.9)–(3.12).

The uniqueness is an automatic consequence of the next result and of (3.10).  $\square$

**THEOREM 3.3.** *The optimal state  $\{\tau_i^*, N_i^*, B_i^*\}_{i=\overline{1,3}}$  is the unique solution to the system (3.8) that governs the generalized Naghdi model for curved rods.*

*Proof.* For any  $\{V_{ij}\}_{i,j=\overline{1,3}} \in \mathcal{V}$ , we define the system in variations by

$$\begin{aligned}
 (3.28) \quad & M_i(x_3) h_{1j}(\bar{x}) + D_i(x_3) h_{2j}(\bar{x}) + [\mu_i'(x_3) + x_1 M_i'(x_3) \\
 & + x_2 D_i'(x_3)] h_{3j}(\bar{x}) = V_{ij}(\bar{x}), \quad i, j = \overline{1,3},
 \end{aligned}$$

$$(3.29) \quad M_i(0) = D_i(0) = \mu_i(0) = 0, \quad i = \overline{1,3}.$$

Next, we perform admissible variations around the optimal pair of the following form:

$$\{\tau_i^*, N_i^*, B_i^*\}_{i=\overline{1,3}} + \lambda \{\mu_i, M_i, D_i\}_{i=\overline{1,3}}; \quad \{U_{ij}^*\}_{i,j=\overline{1,3}} + \lambda \{V_{ij}\}_{i,j=\overline{1,3}}.$$

By (3.28), (3.29), and (3.10)–(3.12), this couple is admissible for any  $\lambda \in \mathbb{R}$ . Subtracting the corresponding cost and the optimal cost, dividing by  $\lambda > 0$  or  $\lambda < 0$ , respectively, and taking the limits as  $\lambda \rightarrow 0$ , the minimum property of  $\{U_{ij}^*\}_{i,j=\overline{1,3}}$

yields the associated Euler equation,

(3.30)

$$\begin{aligned} 0 = & \tilde{\lambda} \int_{\Omega} \sum_{i,j=1}^3 [U_{ii}^* V_{jj} + U_{jj}^* V_{ii}] \det J(\bar{x}) d\bar{x} \\ & + 2\tilde{\mu} \int_{\Omega} \sum_{i < j} [U_{ij}^* + U_{ji}^*] [V_{ij} + V_{ji}] \det J(\bar{x}) d\bar{x} + 4\tilde{\mu} \int_{\Omega} \sum_{i=1}^3 U_{ii}^* V_{ii} \det J(\bar{x}) d\bar{x} \\ & - 2 \int_{\Omega} \sum_{i=1}^3 f_i [\mu_i + x_1 M_i + x_2 D_i] \det J(\bar{x}) d\bar{x}. \end{aligned}$$

If the  $V_{ij}$  are replaced as in (3.28), and if the  $U_{ij}^*$  are replaced as in (3.10), then a simple computation shows that (3.30) becomes (3.8), which is known to have a unique solution. This ends the proof.  $\square$

**PROPOSITION 3.4.** *If  $\{U_{ij}^*\}$  is known, then  $\{\tau_i^*, N_i^*, B_i^*\}_{i=\overline{1,3}}$  can be computed explicitly.*

*Proof.* Starting from (3.2), one can check the following orthogonality-type relations:

$$(3.31) \quad \sum_{j=1}^3 h_{1j} b_j = 0, \quad \sum_{j=1}^3 h_{3j} b_j = 0, \quad \sum_{j=1}^3 h_{2j} b_j = 1,$$

$$(3.32) \quad \sum_{j=1}^3 h_{1j} n_j = 1, \quad \sum_{j=1}^3 h_{2j} n_j = 0, \quad \sum_{j=1}^3 h_{3j} n_j = 0,$$

$$(3.33) \quad \sum_{j=1}^3 h_{1j} t_j = -\frac{c x_2}{\det J(\bar{x})}, \quad \sum_{j=1}^3 h_{2j} t_j = \frac{c x_1}{\det J(\bar{x})}, \quad \sum_{j=1}^3 h_{3j} t_j = \frac{1}{\det J(\bar{x})}.$$

Consequently, multiplying (3.10) by  $n_j$  (respectively, by  $b_j, t_j$ ) and adding for  $j = \overline{1,3}$ , the relations (3.31)–(3.33) yield that

$$(3.34) \quad B_i^* = \sum_{j=1}^3 U_{ij}^* b_j, \quad i = \overline{1,3},$$

$$(3.35) \quad N_i^* = \sum_{j=1}^3 U_{ij}^* n_j, \quad i = \overline{1,3},$$

$$\begin{aligned} (3.36) \quad (\tau_i^*)' + x_1 (N_i^*)' + x_2 (B_i^*)' = & \sum_{j=1}^3 U_{ij}^* t_j \det J(\bar{x}) + \sum_{j=1}^3 U_{ij}^* n_j c x_2 \\ & - \sum_{j=1}^3 U_{ij}^* b_j c x_1, \quad i = \overline{1,3}. \end{aligned}$$

Thus, integrating over  $[0, x_3]$  in (3.36) and subtracting (3.34), (3.35), we also obtain an explicit formula for  $\{\tau_i^*\}_{i=\overline{1,3}}$ , which completes the proof.  $\square$

*Remark.* Relations (3.34)–(3.36) may be extended to any admissible couple  $\{\tau_i, N_i, B_i\}_{i=\overline{1,3}}, \{U_{ij}\}_{i,j=\overline{1,3}}$ .

*Remark.* Let us denote by  $\Lambda_i(U_{ij})$  the right-hand side of (3.36). Then we can perform the following substitution in (3.9):

$$\begin{aligned}
 (3.37) \quad & \sum_{i=1}^3 \int_{\Omega} f_i [\tau_i + x_1 N_i + x_2 B_i] \det J \, d\bar{x} \\
 &= - \sum_{i=1}^3 \int_{\Omega} [\tau'_i + x_1 N'_i + x_2 B'_i] \int_0^{x_3} f_i(x_1, x_2, \rho) \det J(x_1, x_2, \rho) \, d\rho \, d\bar{x} \\
 &= - \sum_{i=1}^3 \int_{\Omega} \Lambda_i(U_{ij}) \int_0^{x_3} f_i(x_1, x_2, \rho) \det J(x_1, x_2, \rho) \, d\rho \, d\bar{x}.
 \end{aligned}$$

In this way, the optimal control problem (3.9)–(3.12) can be transformed into a mathematical programming problem defined on  $\mathcal{V} \subset L^2(\Omega)^9$ , since the state does not appear anymore in the cost functional. However, in order to recover the solution to (3.8), one has to solve (3.10) or use Proposition 3.4.

**PROPOSITION 3.5.** *The directional derivative at the point  $\{U_{ij}\}_{i,j=\overline{1,3}} \in \mathcal{V}$  and in the direction  $\{V_{ij}\}_{i,j=\overline{1,3}}$  is given by*

$$\begin{aligned}
 (3.38) \quad & \langle \nabla L(U_{ij}), \{V_{ij}\} \rangle = \tilde{\lambda} \int_{\Omega} \sum_{i,j=1}^3 [U_{ii} V_{jj} + U_{jj} V_{ii}] \det J(\bar{x}) \, d\bar{x} \\
 &+ 2\tilde{\mu} \int_{\Omega} \sum_{i < j} [U_{ij} + U_{ji}] [V_{ij} + V_{ji}] \det J(\bar{x}) \, d\bar{x} + 4\tilde{\mu} \int_{\Omega} \sum_{i=1}^3 U_{ii} V_{ii} \det J(\bar{x}) \, d\bar{x} \\
 &- 2 \int_{\Omega} \sum_{i=1}^3 \Lambda_i(V_{ij}) \int_0^{x_3} f_i(x_1, x_2, \rho) \det J(x_1, x_2, \rho) \, d\rho \, d\bar{x}.
 \end{aligned}$$

Here,  $\langle \cdot, \cdot \rangle$  is the scalar product in  $L^2(\Omega)^9$ .

*Proof.* It is similar to the deduction of the Euler equation (3.30). The last integral may be rewritten as in (3.37).  $\square$

*Remark.* By (3.38), one can solve (3.9)–(3.12) by standard gradient with projection methods.

**4. Abstract variants.** In this section, we briefly discuss two abstract variants of the control variational method. To this end, let  $V \subset H$  be two separable Hilbert spaces with dense and continuous embedding, endowed with scalar products  $(\cdot, \cdot)_V$  and  $(\cdot, \cdot)_H$ , respectively. Let  $V^*$  be the dual of  $V$  ( $H$  is identified with its own dual space), and let  $A_1, A_2 : V \rightarrow V^*$  be linear, continuous, symmetric operators, with  $A_1$  positively definite.

We consider the equation

$$(4.1) \quad (A_1 + A_2)y = f \in H,$$

and we assume the existence of a unique solution  $y \in V$  to (4.1).

The first optimal control problem associated with (4.1) that we take into account is

$$(4.2) \quad \min \left\{ \frac{1}{2} |w|_H^2 + \frac{1}{2} (A_2 y, y)_{V^* \times V} \right\},$$

subject to  $w \in H$  (no constraints) and

$$(4.3) \quad A_1^{1/2} y = g + w,$$

where  $g \in V$  is the unique solution of  $A_1^{1/2} g = f$ , and where  $A_1^{1/2} : V \rightarrow H$  is the square root of  $A_1$ , defined by the relation

$$(A_1^{1/2} y, A_1^{1/2} v)_H = (A_1 y, v)_{V^* \times V} \quad \forall y, v \in V.$$

Clearly,  $A_1^{1/2}$  is symmetric and positive definite, and (4.3) has a unique solution  $y \in V$  for any  $w \in H$ .

We denote by  $[w^*, y^*] \in H \times V$  an optimal pair for the unconstrained control problem (4.2), (4.3), which is assumed to exist. Now, take arbitrary variations of the form

$$y^* + \lambda z, \quad \lambda \in \mathbb{R}, \quad z \in V, \quad \text{and} \quad w^* + \lambda v, \quad v = A_1^{1/2} z \in H.$$

Then, as in (3.30), one may establish the Euler equation associated with (4.2), (4.3):

$$(4.4) \quad (w^*, v)_H + (A_2 y^*, z)_{V^* \times V} = 0$$

for any  $v \in H$ ,  $z \in V$  as above.

In (4.4), we replace  $v$  and  $w^*$ , and we infer that

$$(4.5) \quad \begin{aligned} 0 &= \left( A_1^{1/2} y^* - g, A_1^{1/2} z \right)_H + (A_2 y^*, z)_{V^* \times V} \\ &= (A_1 y^*, z)_{V^* \times V} + (A_2 y^*, z)_{V^* \times V} - \left( A_1^{-1/2} f, A_1^{1/2} z \right)_H \\ &= ((A_1 + A_2) y^*, z)_{V^* \times V} - (f, z)_H \quad \forall z \in V. \end{aligned}$$

Clearly, (4.5) shows that  $y^*$  is a solution to (4.1).

*Remark.* If also  $A_2$  is positive definite, then the control problem (4.2), (4.3) is coercive and strictly convex, and the existence of the optimal pair  $[w^*, y^*]$  is standard.

We consider now a second optimal control problem that may be associated with (4.1):

$$(4.6) \quad \min \left\{ (u, y)_{V^* \times V} - 3(u, \tilde{g})_{V^* \times V} + (A_2 y, y)_{V^* \times V} \right\},$$

subject to  $u \in V^*$  (again no constraints) and

$$(4.7) \quad A_1 y = u - f.$$

Here,  $\tilde{g} \in V$  is the unique solution to  $A_1 \tilde{g} = f$ , which exists under the assumptions imposed on  $A_1$ . The equation in variations associated with (4.7) is

$$(4.8) \quad A_1 \xi = \omega$$

for any  $\omega \in V^*$  (equivalently, for any  $\xi \in V$ , since  $A_1 : V \rightarrow V^*$  is an isomorphism). The Euler equation for the control problem (4.6), (4.7) is

$$(4.9) \quad 0 = (u^*, \xi)_{V^* \times V} + (\omega, y^*)_{V^* \times V} - 3(\omega, \tilde{g})_{V^* \times V} + 2(A_2 y^*, \xi)_{V^* \times V}$$

for any  $[\omega, \xi]$  satisfying (4.8). In (4.9),  $[u^*, y^*] \in V^* \times V$  denotes an optimal pair of (4.6), (4.7), which is assumed to exist. The argument for establishing (4.9) is the same as for (3.30). Combining (4.7)–(4.9), we obtain that

$$\begin{aligned} (4.10) \quad 0 &= (u^*, \xi)_{V^* \times V} + (A_1 \xi, y^*)_{V^* \times V} - 3(A_1 \xi, \bar{g})_{V^* \times V} + 2(A_2 y^*, \xi)_{V^* \times V} \\ &= (u^*, \xi)_{V^* \times V} + (A_1 y^*, \xi)_{V^* \times V} - 3(f, \xi)_{V^* \times V} + 2(A_2 y^*, \xi)_{V^* \times V} \\ &= 2(A_1 y^*, \xi) - 2(f, \xi)_{V^* \times V} + 2(A_2 y^*, \xi)_{V^* \times V} \end{aligned}$$

for any  $\xi \in V$ . Clearly, (4.10) shows that  $y^*$  satisfies (4.1). We thus have proved the following proposition.

**PROPOSITION 4.1.** *Any optimal state of either (4.2), (4.3) or (4.6), (4.7) is a solution to (4.1).*

*Remark.* The control variational method replaces (4.1) by (4.3) (alternatively by (4.7)). Both (4.3) and (4.7) involve just the inversion of  $A_1$ , and no invertibility properties are required for  $A_2$ . That is, one may choose  $A_1$  as the “nice and simple” part of (4.1) and solve (4.1) by inverting  $A_1$  (or  $A_1^{1/2}$ ) several times.

*Remark.* One could apply one of the above abstract approaches directly in section 2 or 3. However, we preferred simple and “adapted” direct approaches in the previous sections, in order to underline the flexibility of the control variational method. Further examples and applications concerning multiscale problems, hyperbolic equations, and singular and degenerate systems are briefly indicated in Tiba [14].

**Acknowledgment.** The authors thank the anonymous reviewers for suggestions that led to an improvement of the presentation.

#### REFERENCES

- [1] V. ARNĂUTU, H. LANGMACH, J. SPREKELS, AND D. TIBA, *On the approximation and the optimization of plates*, Numer. Funct. Anal. Optim., 21 (2000), pp. 337–354.
- [2] PH. CIARLET, *Mathematical Elasticity. Vol. III. Theory of Shells*, North-Holland, Amsterdam, 2000.
- [3] R. GLOWINSKI AND O. PIRONNEAU, *Sur la résolution numérique du problème de Dirichlet pour l'opérateur biharmonique par une méthode quasi-directe*, C. R. Acad. Sci. Paris Sér. A-B, 282 (1976), pp. A223–A226.
- [4] R. GLOWINSKI AND O. PIRONNEAU, *Sur une méthode quasi-directe pour l'opérateur biharmonique et ses applications à la résolution des équations de Navier–Stokes*, Ann. Sci. Math. Québec, 1 (1977), pp. 231–245.
- [5] A. IGNAT, J. SPREKELS, AND D. TIBA, *A model of a general elastic curved rod*, Math. Methods Appl. Sci., 25 (2002), pp. 835–854.
- [6] F. MURAT AND A. SILI, *Comportement asymptotique des solutions du système de l'élasticité linéarisée anisotrope hétérogène dans des cylindres minces*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 179–184.
- [7] P. NEITTAANMÄKI, J. SPREKELS, AND D. TIBA, *Optimization of Elliptic Systems. Theory and Applications*, Springer, New York, 2006.
- [8] J. SPREKELS AND D. TIBA, *On the approximation and optimization of fourth order elliptic problems*, in Optimal Control of Partial Differential Equations (Chemnitz, 1998), Internat. Ser. Numer. Math. 133, Birkhäuser, Basel, 1999, pp. 277–286.
- [9] J. SPREKELS AND D. TIBA, *Sur les arches lipschitziennes*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 179–184.
- [10] J. SPREKELS AND D. TIBA, *Control variational methods for differential equations*, in Optimal Control of Complex Structures (Oberwolfach, 2000), Internat. Ser. Numer. Math. 139, Birkhäuser, Basel, 2002, pp. 245–257.
- [11] J. SPREKELS AND D. TIBA, *An analytic approach to a generalized Naghdi shell model*, Adv. Math. Sci. Appl., 12 (2002), pp. 175–190.

- [12] J. SPREKELS AND D. TIBA, *Optimal design of mechanical structures*, in Control Theory of Partial Differential Equations, Lecture Notes Pure Appl. Math. 242, Chapman & Hall/CRC, Boca Raton, FL, 2005, pp. 259–271.
- [13] J. SPREKELS AND D. TIBA, *An Optimal Control Approach to Curved Rods*, Preprint 1209, Weierstrass Institute, Berlin, 2007.
- [14] D. TIBA, *Applications of the Control Variational Method*, Preprint 4, Institute of Mathematics of the Romanian Academy, Bucharest, 2007.



## POSITIVE FORMS AND STABILITY OF LINEAR TIME-DELAY SYSTEMS\*

MATTHEW M. PEET<sup>†</sup>, ANTONIS PAPACHRISTODOULOU<sup>‡</sup>, AND SANJAY LALL<sup>§</sup>

**Abstract.** We consider the problem of constructing Lyapunov functions for linear differential equations with delays. For such systems it is known that exponential stability implies the existence of a positive Lyapunov function which is quadratic on the space of continuous functions. We give an explicit parameterization of a sequence of finite-dimensional subsets of the cone of positive Lyapunov functions using positive semidefinite matrices. This allows stability analysis of linear time-delay systems to be formulated as a semidefinite program.

**Key words.** Lyapunov stability, delay systems, semidefinite programming

**AMS subject classifications.** 34D20, 90C22, 37F10

**DOI.** 10.1137/070706999

**1. Summary of the paper.** In this paper we present an approach to the parameterization of Lyapunov functions for infinite-dimensional systems. In particular, we consider linear time-delay systems. These are systems which can be represented in the form

$$\dot{x}(t) = \sum_{i=0}^k A_i x(t - h_i),$$

where  $x(t) \in \mathbb{R}^n$ . In the simplest case we are given the delays  $h_0, \dots, h_k$  and the matrices  $A_0, \dots, A_k$  and we would like to determine whether the system is stable. For such systems it is known that if the system is stable, then there exists a Lyapunov function of the form

$$V(\phi) = \int_{-h}^0 \begin{bmatrix} \phi(0) \\ \phi(s) \end{bmatrix}^T M(s) \begin{bmatrix} \phi(0) \\ \phi(s) \end{bmatrix} ds + \int_{-h}^0 \int_{-h}^0 \phi(s)^T N(s, t) \phi(t) ds dt,$$

where  $h = \max\{h_0, \dots, h_k\}$  and  $M$  and  $N$  are piecewise continuous matrix-valued functions. Here  $\phi : [-h, 0] \rightarrow \mathbb{R}^n$  is an element of the state space, which in this case is the space of continuous functions mapping  $[-h, 0]$  to  $\mathbb{R}^n$ . The function  $V$  is thus a quadratic form on the state space. The derivative is also such a quadratic form, and the matrix-valued functions which define it depend linearly on  $M$  and  $N$ .

In this paper we develop an approach which uses semidefinite programming to construct piecewise continuous functions  $M$  and  $N$  such that the function  $V$  is positive and its derivative is negative. Roughly speaking, our contributions are as follows.

---

\*Received by the editors October 31, 2007; accepted for publication (in revised form) August 5, 2008; published electronically January 9, 2009.

<http://www.siam.org/journals/sicon/47-6/70699.html>

<sup>†</sup>Department of Mechanical, Materials, and Aerospace Engineering, Illinois Institute of Technology, Chicago, IL 60616 (mpeet@iit.edu).

<sup>‡</sup>Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK (antonis@eng.ox.ac.uk).

<sup>§</sup>Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305-4035 (lall@stanford.edu).

*Spacing functions.* In Theorem 5, we show that for any piecewise continuous function  $M$

$$V_1(\phi) = \int_{-h}^0 \begin{bmatrix} \phi(0) \\ \phi(s) \end{bmatrix}^T M(s) \begin{bmatrix} \phi(0) \\ \phi(s) \end{bmatrix} ds$$

is positive for all  $\phi$  if and only if there exists a piecewise continuous matrix-valued function  $T$  such that

$$M(t) + \begin{bmatrix} T(t) & 0 \\ 0 & 0 \end{bmatrix} \geq 0 \quad \text{for all } t,$$

$$\int_{-h}^0 T(t) dt = 0.$$

That is, we convert positivity of the integral to *pointwise* positivity of  $M$ . If we assume that  $M$  is polynomial, then pointwise positivity may be easily enforced, and in the case of positivity on the real line this is equivalent to a sum-of-squares constraint. The assumption that  $M$  is polynomial has recently been shown to be nonconservative. The constraint that  $T$  integrates to zero is a simple linear constraint on the coefficients of  $T$ . Notice that the sufficient condition that  $M(s)$  be pointwise nonnegative is conservative, and as the equivalence above shows it is easy to generate examples where  $V_1$  is nonnegative even though  $M(s)$  is not pointwise nonnegative.

*Positive polynomial kernels.* We give a sum-of-squares characterization of positive polynomial kernels. We consider the quadratic form

$$\int_{-h}^0 \int_{-h}^0 \phi(s)^T N(s, t) \phi(t) ds dt.$$

In Theorem 7, we show that the quadratic form is positive if and only if there exists a positive semidefinite matrix  $Q \succeq 0$  such that  $N(s, t) = Z(s)^T Q Z(t)$ , where  $Z$  is a vector of monomials. This condition allows us to test positivity of a polynomial kernel using semidefinite programming and implies the existence of a sum-of-squares-type representation. Note that pointwise positivity of  $N$  is not sufficient for positivity of the functional. The condition that the derivative of the Lyapunov function be negative is similarly enforced.

This paper is organized as follows. We begin with a discussion of the relevant history and prior work. We then give a definition of the system for which we will prove stability and a presentation of the class of Lyapunov functions we will construct. Following this section, we present our first result giving pointwise condition for positivity of the functional. Next, we give a parameterization of piecewise polynomial matrices. We use sum-of-squares techniques to parameterize matrices which are pointwise nonnegative, and we present a new result on the parameterization of kernel matrices which define positive quadratic forms. We then return to the problem of linear time-delay systems to document the derivative of the Lyapunov function along trajectories of the system. Finally, we use this derivative to define a semidefinite program for proving stability of linear time-delay systems.

**2. Background and prior work.** The use of Lyapunov functions on an infinite-dimensional space to analyze differential equations with delay originates with the work of [10]. For linear systems, quadratic Lyapunov functions were first considered by [19].

The book of [4] presents many useful results in this area, and further references may be found there as well as in [5, 9] and [13].

The idea of using semidefinite programming and sum of squares to solve polynomial optimization problems has many sources. Well-known examples include [11, 12] and [16]. Work on this problem continues actively, with recent results to be found in, e.g., [6] and [1].

An early example of using sum-of-squares polynomials together with semidefinite programming to construct polynomial Lyapunov functions for nonlinear ordinary differential equations can be found in [15]. Substantial work has been done in this area, with contributions from, e.g., [23] and [2].

The idea of constructing Lyapunov functions for linear time-delay systems using semidefinite programming is not original or new. A typical approach has been to examine subclasses of functions  $M$  and  $N$ , e.g., constant matrices, piecewise linear functions, etc. In practice, some of these approaches have been shown to be highly accurate.

The motivation for this paper is not exclusively the stability of linear time-delay systems. It is our hope, by investigating properties of a positive form known to be necessary and sufficient for stability of infinite-dimensional systems, and by using polynomial methods, that our results will also be useful in analysis of nonlinear and partial differential systems.

Building on the results of this paper, a treatment of nonquadratic Lyapunov functions for nonlinear time-delay systems appears in [14].

**2.1. Notation.** Let  $\mathbb{N}$  denote the set of nonnegative integers. Let  $\mathbb{S}^n$  be the set of  $n \times n$  real symmetric matrices, and for  $X \in \mathbb{S}^n$  we write  $X \succeq 0$  to mean that  $X$  is positive semidefinite. For two matrices  $A, B$ , we denote the Kronecker product by  $A \otimes B$ . For  $X$  any Banach space and  $I \subset \mathbb{R}$  any interval, let  $\Omega(I, X)$  be the space of all functions

$$\Omega(I, X) = \{f : I \rightarrow X\}$$

and let  $C(I, X)$  be the Banach space of bounded continuous functions

$$C(I, X) = \{f : I \rightarrow X \mid f \text{ is continuous and bounded}\}$$

equipped with the norm

$$\|f\| = \sup_{t \in I} \|f(t)\|_X.$$

We will omit the range space when it is clear from the context; for example we write  $C[a, b]$  to mean  $C([a, b], X)$ . A function is called  $C^n(I, X)$  if the  $i$ th derivative exists and is a continuous function for  $i = 0, \dots, n$ . A function  $f \in C[a, b]$  is called piecewise continuous if there exists a finite number of points  $a < h_1 < \dots < h_k < b$  such that  $f$  is continuous at all  $x \in [a, b] \setminus \{h_1, \dots, h_k\}$  and its right- and left-hand limits exist at  $\{h_1, \dots, h_k\}$ .

Define also the projection  $F_t : \Omega[-h, \infty) \rightarrow \Omega[-h, 0]$  for  $t \geq 0$  and  $h > 0$  by

$$(F_t x)(s) = x(t + s) \quad \text{for all } s \in [-h, 0].$$

We follow the usual convention and denote  $F_t x$  by  $x_t$ .

**3. Linear time-delay systems.** Suppose  $0 = h_0 < h_1 < \cdots < h_k = h$  and  $A_0, \dots, A_k \in \mathbb{R}^{n \times n}$ . We consider linear differential equations with delay, of the form

$$(1) \quad \dot{x}(t) = \sum_{i=0}^k A_i x(t - h_i) \quad \text{for all } t \geq 0,$$

where the trajectory  $x : [-h, \infty) \rightarrow \mathbb{R}^n$ . The boundary conditions are specified by a given function  $\phi : [-h, 0] \rightarrow \mathbb{R}^n$  and the constraint

$$(2) \quad x(t) = \phi(t) \quad \text{for all } t \in [-h, 0].$$

If  $\phi \in C[-h, 0]$ , then there exists a unique function  $x$  satisfying (1) and (2). The system is called *exponentially stable* if there exist  $\sigma > 0$  and  $a \in \mathbb{R}$  such that for every initial condition  $\phi \in C[-h, 0]$  the corresponding solution  $x$  satisfies

$$\|x(t)\| \leq ae^{-\sigma t} \|\phi\| \quad \text{for all } t \geq 0.$$

We write the solution as an explicit function of the initial conditions using the map  $G : C[-h, 0] \rightarrow \Omega[-h, \infty)$ , defined by

$$(G\phi)(t) = x(t) \quad \text{for all } t \geq -h,$$

where  $x$  is the unique solution of (1) and (2) corresponding to initial condition  $\phi$ . Also for  $s \geq 0$  define the *flow map*  $\Gamma_s : C[-h, 0] \rightarrow C[-h, 0]$  by

$$\Gamma_s \phi = F_s G \phi,$$

which maps the state of the system  $x_t$  to the state at a later time  $x_{t+s} = \Gamma_s x_t$ .

**3.1. Lyapunov functions.** Lyapunov theory for infinite-dimensional systems closely parallels that for finite-dimensional systems. The difference is that the state space is now a function space, and therefore Lyapunov functions are actually functions of functions. For linear time-delay systems, the Lyapunov functions we define here are functions of segments of the trajectory and, in particular, of the state  $x_t$ . For a given  $V : C[-h, 0] \rightarrow \mathbb{R}$ , we use the standard notion of the *Lie derivative* or derivative of a function on a vector field. The Lie derivative of  $V$  is defined by the flow map,  $\Gamma$ , as

$$\dot{V}(\phi) = \limsup_{r \rightarrow 0^+} \frac{1}{r} (V(\Gamma_r \phi) - V(\phi)).$$

In keeping with tradition, we will use the notation  $\dot{V}$  to denote the Lie derivative. We will consider the set  $X$  of quadratic functions, where  $V \in X$  if there exist piecewise continuous functions  $M : [-h, 0] \rightarrow \mathbb{S}^{2n}$  and  $N : [-h, 0] \times [-h, 0] \rightarrow \mathbb{R}^{n \times n}$  such that

$$(3) \quad V(\phi) = \int_{-h}^0 \begin{bmatrix} \phi(0) \\ \phi(s) \end{bmatrix}^T M(s) \begin{bmatrix} \phi(0) \\ \phi(s) \end{bmatrix} ds + \int_{-h}^0 \int_{-h}^0 \phi(s)^T N(s, t) \phi(t) ds dt.$$

The following result shows that for linear systems with delay the system is exponentially stable if and only if there exists a quadratic Lyapunov function. Define the sets  $H = \{-h_0, \dots, -h_k\}$  and  $H^c = [-h, 0] \setminus H$ .

**THEOREM 1.** *The linear system defined by (1) and (2) is exponentially stable if and only if there exists a Lie-differentiable function  $V \in X$  and  $\varepsilon > 0$  such that for all  $\phi \in C[-h, 0]$*

$$(4) \quad \begin{aligned} V(\phi) &\geq \varepsilon \|\phi(0)\|^2, \\ \dot{V}(\phi) &\leq -\varepsilon \|\phi(0)\|^2. \end{aligned}$$

Further  $V \in X$  may be chosen such that the corresponding functions  $M$  and  $N$  of (3) have the following smoothness property:  $M(s)$  and  $N(s, t)$  are continuous on  $s, t \in [-h, 0] \setminus \{-h_0, \dots, -h_k\}$ .

*Proof.* See [4] or [7] for a recent proof.  $\square$

The consequence of this theorem is that stability of a linear time-delay system is equivalent to the existence of a Lyapunov function which decreases along segments of the trajectory. In the next few sections, we will give results which will allow us to better understand the positivity of the function.

**4. Positivity of integral forms.** The goal of this section is to present results which enable us to characterize functions  $V \in X$  which satisfy the positivity conditions in (4) and have the form

$$V(y) = \int_{-h}^0 \begin{bmatrix} y(0) \\ y(t) \end{bmatrix}^T M(t) \begin{bmatrix} y(0) \\ y(t) \end{bmatrix} dt.$$

Before stating the main result in Theorem 5, we give a few necessary lemmas.

**LEMMA 2.** *Suppose  $f: [-h, 0] \rightarrow \mathbb{R}$  is piecewise continuous. Then the following are equivalent:*

- (i)  $\int_{-h}^0 f(t) dt \geq 0$ .
- (ii) *There exists a function  $g: [-h, 0] \rightarrow \mathbb{R}$  which is piecewise continuous and satisfies*

$$f(t) + g(t) \geq 0 \quad \text{for all } t,$$

$$\int_{-h}^0 g(t) dt = 0.$$

*Proof.* The direction (ii)  $\implies$  (i) is immediate. To show the other direction, suppose (i) holds, and let  $g$  be

$$g(t) = -f(t) + \frac{1}{h} \int_{-h}^0 f(s) ds \quad \text{for all } t.$$

Then  $g$  satisfies (ii).  $\square$

The next lemma shows that minimizing over continuous functions is as good as minimizing over piecewise continuous functions.

**LEMMA 3.** *Suppose  $H = \{-h_0, \dots, -h_k\}$ , and let  $H^c = [-h, 0] \setminus H$ . Let  $f: [-h, 0] \times \mathbb{R}^n \rightarrow \mathbb{R}$  be continuous on  $H^c \times \mathbb{R}^n$ , and suppose there exists a bounded function  $z: [-h, 0] \rightarrow \mathbb{R}$ , continuous on  $H^c$ , such that for all  $t \in [-h, 0]$*

$$f(t, z(t)) = \inf_x f(t, x).$$

Further suppose for each bounded set  $X \subset \mathbb{R}^n$  the set

$$\{f(t, x) \mid x \in X, t \in [-h, 0]\}$$

is bounded. Then

$$(5) \quad \inf_{y \in C[-h,0]} \int_{-h}^0 f(t, y(t)) dt = \int_{-h}^0 \inf_x f(t, x) dt.$$

*Proof.* Let

$$K = \int_{-h}^0 \inf_x f(t, x) dt.$$

It is easy to see that

$$\inf_{y \in C[-h,0]} \int_{-h}^0 f(t, y(t)) dt \geq K$$

since, if not, then there would exist some continuous function  $y$  and some interval on which

$$f(t, y(t)) < \inf_x f(t, x),$$

which is clearly impossible.

We now show that the left-hand side of (5) is also less than or equal to  $K$  and hence equals  $K$ . We need to show that for any  $\varepsilon > 0$  there exists  $y \in C[-h, 0]$  such that

$$\int_{-h}^0 f(t, y(t)) dt < K + \varepsilon.$$

To do this, for each  $n \in \mathbb{N}$  define the set  $H_n \subset \mathbb{R}$  by

$$H_n = \bigcup_{i=1}^{k-1} (h_i - \alpha/n, h_i + \alpha/n)$$

and choose  $\alpha > 0$  sufficiently small so that  $H_1 \subset (-h, 0)$ . Let  $z$  be as in the hypothesis of the lemma, and pick  $M$  and  $R$  so that

$$M > \sup_{t \in [-h, 0]} \|z(t)\|,$$

$$R = \sup\{|f(t, x)| \mid t \in [-h, 0], \|x\| \leq M\}.$$

For each  $n$  choose a continuous function  $x_n : [-h, 0] \rightarrow \mathbb{R}^n$  such that  $x_n(t) = z(t)$  for all  $t \notin H_n$  and

$$\sup_{t \in [-h, 0]} \|x_n(t)\| < M.$$

This is possible, for example, by linear interpolation. Now we have for the continuous function  $x_n$

$$\begin{aligned} \int_{-h}^0 f(t, x_n(t)) dt &= K + \int_{-h}^0 (f(t, x_n(t)) - f(t, z(t))) dt \\ &= K + \int_{H_n} (f(t, x_n(t)) - f(t, z(t))) dt \\ &\leq K + 4R\alpha(k-1)/n. \end{aligned}$$

This proves the desired result.  $\square$

The following lemma states that when the  $\arg \min_z f(t, z)$  is piecewise continuous in  $t$  we have the desired result.

LEMMA 4. Suppose  $f: [-h, 0] \times \mathbb{R}^n \rightarrow \mathbb{R}$  and the hypotheses of Lemma 3 hold. Then the following are equivalent:

(i) For all  $y \in C[-h, 0]$

$$\int_{-h}^0 f(t, y(t)) dt \geq 0.$$

(ii) There exists  $g: [-h, 0] \rightarrow \mathbb{R}$  which is piecewise continuous and satisfies

$$f(t, z) + g(t) \geq 0 \quad \text{for all } t, z,$$

$$\int_{-h}^0 g(t) dt = 0.$$

*Proof.* Again we need only show that (i) implies (ii). Suppose (i) holds; then

$$\inf_{y \in C[-h, 0]} \int_{-h}^0 f(t, y(t)) dt \geq 0,$$

and hence by Lemma 3 we have

$$\int_{-h}^0 r(t) dt \geq 0,$$

where  $r: [-h, 0] \rightarrow \mathbb{R}^n$  is given by

$$r(t) = \inf_x f(t, x) \quad \text{for all } t.$$

The function  $r$  is continuous on  $H^c$  since  $f$  is continuous on  $H^c \times \mathbb{R}^n$ . Hence, by Lemma 2, there exists  $g$  such that condition (ii) holds, as desired.  $\square$

We now specialize the result of Lemma 4 to the case of quadratic functions. It is shown that in this case, under certain conditions, the  $\arg \min_z f(t, z)$  is piecewise continuous.

THEOREM 5. Suppose  $M: [-h, 0] \rightarrow \mathbb{S}^{m+n}$  is piecewise continuous, and there exists  $\varepsilon > 0$  such that for all  $t \in [-h, 0]$  we have

$$M_{22}(t) \geq \varepsilon I,$$

where  $M$  is partitioned as

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

with  $M_{22}: [-h, 0] \rightarrow \mathbb{S}^n$ . Then the following are equivalent:

(i) For all  $x \in \mathbb{R}^m$  and continuous  $y: [-h, 0] \rightarrow \mathbb{R}^n$

$$(6) \quad \int_{-h}^0 \begin{bmatrix} x \\ y(t) \end{bmatrix}^T M(t) \begin{bmatrix} x \\ y(t) \end{bmatrix} dt \geq 0.$$

- (ii) *There exists a function  $T : [-h, 0] \rightarrow \mathbb{S}^m$  which is piecewise continuous and satisfies*

$$M(t) + \begin{bmatrix} T(t) & 0 \\ 0 & 0 \end{bmatrix} \geq 0 \quad \text{for all } t \in [-h, 0],$$

$$\int_{-h}^0 T(t) dt = 0.$$

*Proof.* Again we need only show that (i) implies (ii). Suppose  $x \in \mathbb{R}^n$ , and define

$$f(t, z) = \begin{bmatrix} x \\ z \end{bmatrix}^T M(t) \begin{bmatrix} x \\ z \end{bmatrix} \quad \text{for all } t, z.$$

Since by the hypothesis  $M_{22}$  has a lower bound, it is invertible for all  $t$  and its inverse is piecewise continuous. Therefore  $z(t) = -M_{22}(t)^{-1}M_{21}(t)x$  is the unique minimizer of  $f(t, z)$  with respect to  $z$ . By the hypothesis (i), we have that for all  $y \in C[-h, 0]$

$$\int_{-h}^0 f(t, y(t)) dt \geq 0.$$

Hence by Lemma 4 there exists a function  $g$  such that

$$g(t) + f(t, z) \geq 0 \quad \text{for all } t, z,$$

$$(7) \quad \int_{-h}^0 g(t) dt = 0.$$

The proof of Lemma 2 gives one such function as

$$g(t) = -f(t, z(t)) + \frac{1}{h} \int_{-h}^0 f(s, z(s)) dt.$$

We have

$$f(t, z(t)) = x^T (M_{11}(t) - M_{12}(t)M_{22}^{-1}(t)M_{21}(t))x,$$

and therefore  $g(t)$  is a quadratic function of  $x$ , say  $g(t) = x^T T(t)x$ , and  $T : [-h, 0] \rightarrow \mathbb{S}^m$  is continuous on  $H^c$ . Then (7) implies

$$x^T T(t)x + \begin{bmatrix} x \\ z \end{bmatrix}^T M(t) \begin{bmatrix} x \\ z \end{bmatrix} \geq 0 \quad \text{for all } t, z, x,$$

as required.  $\square$

Notice that the strict positivity assumption on  $M_{22}$  in Theorem 5 is implied by the existence of an  $\epsilon > 0$  such that

$$V(x) \geq \epsilon \|x\|_2^2,$$

where  $\|\cdot\|_2$  denotes the  $L_2$ -norm.

We have now shown that the convex cone of functions  $M$  such that the first term of (3) is nonnegative is exactly equal to the sum of the cone of pointwise nonnegative



functions and the linear space of functions whose integral is zero. The key benefit of this is that it is easy to parameterize the latter class of functions, and in particular when  $M$  is a polynomial these constraints are semidefinite representable constraints on the coefficients of  $M$ . Note that in (6) the vectors  $x$  and  $y$  are allowed to vary independently, whereas (3) requires that  $x = y(0)$ . It is, however, straightforward to show using the technique in the proof of Lemma 3 that this additional constraint does not change the result.

**5. Polynomial matrices and kernels.** In this paper we use piecewise polynomial matrices as a conveniently parameterized class of functions to represent the functions  $M$  and  $N$  defining the Lyapunov function (3) and its derivative. Theorem 5 has reduced nonnegativity of the first term of (3) to pointwise nonnegativity of a piecewise polynomial matrix in one variable.

We first make some definitions which we will use in this paper; some details on polynomial matrices may be found in [21] and [8]. We consider polynomials in  $n$  variables. As is standard, for  $\alpha \in \mathbb{N}^n$  define the monomial in  $n$  variables  $x^\alpha$  by  $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ . We say  $M$  is a real polynomial matrix in  $n$  variables if for some finite set  $W \subset \mathbb{N}^n$  we have

$$M(x) = \sum_{\alpha \in W} A_\alpha x^\alpha,$$

where  $A_\alpha$  is a real matrix for each  $\alpha \in W$ . A convenient representation of polynomial matrices is as a quadratic function of monomials. Suppose  $z$  is a vector of monomials in the variables  $x$ , such as

$$z(x) = \begin{bmatrix} 1 \\ x_1 \\ x_1 x_2^2 \\ x_3^4 \end{bmatrix}.$$

For convenience, assume the length of  $z$  is  $d + 1$ . Let  $Q \in \mathbb{S}^{n(d+1)}$  be a symmetric matrix. Then the function  $M$  defined by

$$(8) \quad M(x) = (I_n \otimes z(x))^T Q (I_n \otimes z(x))$$

is an  $n \times n$  symmetric polynomial matrix, and every real symmetric polynomial matrix may be represented in this way for some monomial vector  $z$ . If we partition  $Q$  as

$$Q = \begin{bmatrix} Q_{11} & \cdots & Q_{1n} \\ \vdots & & \vdots \\ Q_{n1} & \cdots & Q_{nn} \end{bmatrix},$$

where each  $Q_{ij} \in \mathbb{R}^{(d+1) \times (d+1)}$ , then the  $i, j$  entry of  $M$  is

$$M_{ij}(x) = z(x)^T Q_{ij} z(x).$$

Given a polynomial matrix  $M$ , it is called a *sum of squares* if there exist a vector of monomials  $z$  and a positive semidefinite matrix  $Q$  such that (8) holds. In this case,

$$M(x) \succeq 0 \quad \text{for all } x,$$

and therefore the existence of such a  $Q$  is a sufficient condition for the polynomial  $M$  to be globally pointwise positive semidefinite. A matrix polynomial in one variable is pointwise nonnegative semidefinite on the real line if and only if it is a sum of squares; see [3]. Given a matrix polynomial  $M(x)$ , we can test whether it is a sum of squares by testing whether there is a matrix  $Q$  such that

$$(9) \quad M(x) = (I_n \otimes z(x))^T Q (I_n \otimes z(x)),$$

$$Q \succeq 0,$$

where  $z$  is the vector of all monomials with degree half the degree of  $M$ . Equation (9) is interpreted as equality of polynomials, and equating their coefficients gives a finite set of linear constraints on the matrix  $Q$ . Therefore to find such a  $Q$  we need to find a positive semidefinite matrix subject to linear constraints, and this is therefore testable via semidefinite programming. See [25] for background on semidefinite programming.

**5.1. Piecewise polynomial matrices.** Define the intervals

$$H_i = \begin{cases} [-h_1, 0] & \text{if } i = 1, \\ [-h_i, -h_{i-1}) & \text{if } i = 2, \dots, k. \end{cases}$$

A matrix-valued function  $M : [-h, 0] \rightarrow \mathbb{S}^n$  is called a *piecewise polynomial matrix* if for each  $i = 1, \dots, k$  the function  $M$  restricted to the interval  $H_i$  is a polynomial matrix. We represent such piecewise polynomial matrices as follows. Define the vector of indicator functions  $g : [-h, 0] \rightarrow \mathbb{R}^k$  by

$$g_i(t) = \begin{cases} 1 & \text{if } t \in H_i, \\ 0 & \text{otherwise} \end{cases}$$

for all  $i = 1, \dots, k$  and all  $t \in [-h, 0]$ . Let  $z(t)$  be the vector of monomials

$$z(t) = \begin{bmatrix} 1 \\ t \\ t^2 \\ \vdots \\ t^d \end{bmatrix}$$

and for convenience also define the function  $Z_{n,d} : [-h, 0] \rightarrow \mathbb{R}^{nk(d+1) \times n}$  by

$$Z_{n,d}(t) = g(t) \otimes I_n \otimes z(t).$$

Then it is straightforward to show that  $M$  is a piecewise matrix polynomial if and only if there exist matrices  $Q_i \in \mathbb{S}^{n(d+1)}$  for  $i = 1, \dots, k$  such that

$$(10) \quad M(t) = Z_{n,d}(t)^T \text{diag}(Q_1, \dots, Q_k) Z_{n,d}(t).$$

The function  $M$  is pointwise positive semidefinite, i.e.,

$$M(t) \succeq 0 \quad \text{for all } t \in [-h, 0],$$

if there exist positive semidefinite matrices  $Q_i$  satisfying (10). We refer to such functions as *piecewise sum-of-squares matrices*, and we define the set of such functions

$$\Sigma_{n,d} = \{ Z_{n,d}^T(t) Q Z_{n,d}(t) \mid Q = \text{diag}(Q_1, \dots, Q_k), Q_i \in \mathbb{S}^{n(d+1)}, Q_i \succeq 0 \}.$$

If we are given a function  $M : [-h, 0] \rightarrow \mathbb{S}^n$  which is piecewise polynomial and want to know whether it is piecewise sum of squares, then this is computationally checkable using semidefinite programming. Naturally, the number of variables involved in this task scales as  $kn^2(d+1)^2$  when the degree of  $M$  is  $2d$ .

**5.2. Piecewise polynomial kernels.** We consider functions  $N$  of two variables  $s, t$  which we will use as a kernel in the quadratic form

$$(11) \quad \int_{-h}^0 \int_{-h}^0 \phi(s)^T N(s, t) \phi(t) ds dt$$

which appears in the Lyapunov function (3). A polynomial in two variables is referred to as a *binary* polynomial. A function  $N : [-h, 0] \times [-h, 0] \rightarrow \mathbb{S}^n$  is called a *binary piecewise polynomial matrix* if for each  $i, j \in \{1, \dots, k\}$  the function  $N$  restricted to the set  $H_i \times H_j$  is a binary polynomial matrix. It is straightforward to show that  $N$  is a symmetric binary piecewise polynomial matrix if and only if there exists a matrix  $Q \in \mathbb{S}^{nk(d+1)}$  such that

$$N(s, t) = Z_{n,d}^T(s) Q Z_{n,d}(t).$$

Here  $d$  is the degree of  $N$ , and recall that

$$Z_{n,d}(t) = g(t) \otimes I_n \otimes z(t).$$

We now proceed to characterize the binary piecewise polynomial matrices  $N$  for which the quadratic form (11) is nonnegative for all  $\phi \in C([-h, 0], \mathbb{R}^n)$ . We first state the following lemma.

LEMMA 6. *Suppose  $z$  is the vector of monomials*

$$z(t) = \begin{bmatrix} 1 \\ t \\ t^2 \\ \vdots \\ t^d \end{bmatrix}$$

*and the linear map  $A : C[0, 1] \rightarrow \mathbb{R}^{d+1}$  is given by*

$$A\phi = \int_0^1 z(t)\phi(t) dt.$$

*Then  $\text{rank } A = d + 1$ .*

*Proof.* Suppose for the sake of a contradiction that  $\text{rank } A < d + 1$ . Then  $\text{range } A$  is a strict subset of  $\mathbb{R}^{d+1}$ , and hence there exists a nonzero vector  $q \in \mathbb{R}^{d+1}$  such that  $q \perp \text{range } A$ . This means

$$\int_0^1 q^T z(t)\phi(t) dt = 0$$

for all  $\phi \in C[0, 1]$ . Since  $q^T z$  and  $\phi$  are continuous functions, define the function  $v : [0, 1] \rightarrow \mathbb{R}$  by

$$v(t) = \int_0^t q^T z(s) ds \quad \text{for all } t \in [0, 1].$$

Since  $v$  is absolutely continuous, we have for every  $\phi \in C[0, 1]$  that

$$\begin{aligned} \int_0^1 \phi(t) dv(t) &= \int_0^1 q^T z(t) \phi(t) dt \\ &= 0, \end{aligned}$$

where the integral on the left-hand side of the above equation is the Stieltjes integral. The function  $v$  is also of bounded variation, since its derivative is bounded. The Riesz representation theorem [20] implies that if  $v$  is of bounded variation and

$$\int_0^1 \phi(t) dv(t) = 0$$

for all  $\phi \in C[0, 1]$ , then  $v$  is constant on an everywhere dense subset of  $(0, 1)$ . Since  $v$  is continuous, we have that  $v$  is constant, and therefore  $q^T z(t) = 0$  for all  $t$ . Since  $q^T z$  is a polynomial, this contradicts the statement that  $q \neq 0$ .  $\square$

We now state the positivity result.

**THEOREM 7.** *Suppose  $N$  is a symmetric binary piecewise polynomial matrix of degree  $d$ . Then*

$$(12) \quad \int_{-h}^0 \int_{-h}^0 \phi(s)^T N(s, t) \phi(t) ds dt \geq 0$$

for all  $\phi \in C([-h, 0], \mathbb{R}^n)$  if and only if there exists  $Q \in \mathbb{S}^{nk(d+1)}$  such that

$$N(s, t) = Z_{n,d}^T(s) Q Z_{n,d}(t),$$

$$Q \succeq 0.$$

*Proof.* We need only show the *only if* direction. Suppose  $N$  is a symmetric binary piecewise polynomial matrix. Let  $d$  be the degree of  $N$ . Then there exists a symmetric matrix  $Q$  such that

$$N(s, t) = Z_{n,d}^T(s) Q Z_{n,d}(t).$$

Now suppose that the inequality (12) is satisfied for all continuous functions  $\phi$ . We will show that every such  $Q$  is positive semidefinite. To see this, define the linear map  $J : C([-h, 0], \mathbb{R}^n) \rightarrow \mathbb{R}^{nk(d+1)}$  by

$$J\phi = \int_{-h}^0 (g(t) \otimes I_n \otimes z(t)) \phi(t) dt.$$

Then

$$\int_{-h}^0 \int_{-h}^0 \phi(s)^T N(s, t) \phi(t) ds dt = (J\phi)^T Q (J\phi).$$

The result we desire holds if  $\text{rank } J = nk(d+1)$ , since in this case  $\text{range } J = \mathbb{R}^{nk(d+1)}$ . If  $Q$  has a negative eigenvalue with corresponding eigenvector  $q$ , then there exists  $\phi$  such that  $q = J\phi$  so that the quadratic form will be negative, contradicting the hypothesis.

To see that  $\text{rank } J = nk(d+1)$ , define for each  $i = 1, \dots, k$  the linear map  $L_i : C[H_i] \rightarrow \mathbb{R}^n$  by

$$L_i \phi = \int_{H_i} z(t) \phi(t) dt.$$

If we choose coordinates for  $\phi$  such that

$$\phi = \begin{bmatrix} \phi|_{H_1} \\ \phi|_{H_2} \\ \vdots \\ \phi|_{H_k} \end{bmatrix},$$

where  $\phi|_{H_j}$  is the restriction of  $\phi$  to the interval  $H_j$ , then we have in these coordinates that  $J$  is

$$J = \text{diag}(L_1, \dots, L_k) \otimes I_n.$$

Further, by Lemma 6 the maps  $L_i$  each satisfy  $\text{rank } L_i = d+1$ . Therefore  $\text{rank } J = nk(d+1)$ , as desired.  $\square$

The following corollary gives a tighter degree bound on the representation of  $N$ .

**COROLLARY 8.** *Let  $N$  be a binary piecewise polynomial matrix of degree  $2d$  which is positive in the sense of (12); then there exists  $Q \in \mathbb{S}^{nk(d+1)}$  such that*

$$N(s, t) = Z_{n,d}^T(s) Q Z_{n,d}(t),$$

$$Q \succeq 0.$$

*Proof.* The binary representation used in Theorem 7 had the form

$$N(s, t) = Z_{n,2d}^T(s) P Z_{n,2d}(t),$$

$$P \succeq 0,$$

where  $P \in \mathbb{S}^{nk(2d+1)}$ . However, in any such representation, it is clear that  $P_{ij,ij} = 0$  for  $i = d+2, \dots, 2d+1$  and  $j = 1, \dots, kn$ . Therefore, since  $P \succeq 0$ , these rows and columns are 0 and can be removed. Define  $Q$  to be the reduction of  $P$ .  $Z_{n,d}$  is the corresponding reduction of  $Z_{n,2d}$ . Then  $Q \in \mathbb{S}^{nk(d+1)}$ ,  $Q \succeq 0$ , and

$$N(s, t) = Z_{n,2d}^T(s) P Z_{n,2d}(t) = Z_{n,d}^T(s) Q Z_{n,d}(t). \quad \square$$

For convenience, we define the set of symmetric binary piecewise polynomial matrices which define positive quadratic forms by

$$\Gamma_{n,d} = \{ Z_{n,d}^T(s) Q Z_{n,d}(t) \mid Q \in \mathbb{S}^{nk(d+1)}, Q \succeq 0 \}.$$

As for  $\Sigma_{n,d}$ , if we are given a binary piecewise polynomial matrix  $N : [-h, 0] \times [-h, 0] \rightarrow \mathbb{S}^n$  of degree  $2d$  and want to know whether it defines a positive quadratic form, then this is computationally checkable using semidefinite programming. The number of variables involved in this task scales as  $(nk)^2(d+1)^2$ .

**6. Derivatives of the Lyapunov function.** In this section we will take the opportunity to define the relationship between the functions  $M$  and  $N$ , which define the Lyapunov function  $V$ , and the functions  $D$  and  $E$ , which define the derivative of the Lyapunov function along trajectories of the system. As the results are well known, we will not give detailed derivations for these derivatives. More expository explanations can be found in, e.g., [4] or [13].

**6.1. Single delay case.** We first present the single delay case, as it will illustrate the formulation in the more complicated case of several delays. Suppose that  $V \in X$  is given by (3), where  $M : [-h, 0] \rightarrow \mathbb{S}^{2n}$  and  $N : [-h, 0] \times [-h, 0] \rightarrow \mathbb{R}^{n \times n}$ . Since there is only one delay, if the system is exponentially stable, then there always exists a Lyapunov function of this form with continuous functions  $M$  and  $N$ . Then the Lie derivative of  $V$  is

$$\dot{V}(\phi) = \int_{-h}^0 \begin{bmatrix} \phi(0) \\ \phi(-h) \\ \phi(s) \end{bmatrix}^T D(s) \begin{bmatrix} \phi(0) \\ \phi(-h) \\ \phi(s) \end{bmatrix} ds + \int_{-h}^0 \int_{-h}^0 \phi(s)^T E(s, t) \phi(t) ds dt.$$

Partition  $D$  and  $M$  as

$$M(t) = \begin{bmatrix} M_{11} & M_{12}(t) \\ M_{21}(t) & M_{22}(t) \end{bmatrix}, \quad D(t) = \begin{bmatrix} D_{11} & D_{12}(t) \\ D_{21}(t) & D_{22}(t) \end{bmatrix}$$

so that  $M_{11} \in \mathbb{S}^n$  and  $D_{11} \in \mathbb{S}^{2n}$ . Without loss of generality we have assumed that  $M_{11}$  and  $D_{11}$  are constant. The functions  $D$  and  $E$  are linearly related to  $M$  and  $N$  by

$$\begin{aligned} D_{11} &= \begin{bmatrix} A_0^T M_{11} + M_{11} A_0 & M_{11} A_1 \\ A_1^T M_{11} & 0 \end{bmatrix} \\ &\quad + \frac{1}{h} \begin{bmatrix} M_{12}(0) + M_{21}(0) & -M_{12}(-h) \\ -M_{21}(-h) & 0 \end{bmatrix} \\ &\quad + \frac{1}{h} \begin{bmatrix} M_{22}(0) & 0 \\ 0 & -M_{22}(-h) \end{bmatrix}, \\ D_{12}(t) &= \begin{bmatrix} A_0^T M_{12}(t) - \dot{M}_{12}(t) + N(0, t) \\ A_1^T M_{12}(t) - N(-h, t) \end{bmatrix}, \\ D_{22}(t) &= -\dot{M}_{22}(t), \\ E(s, t) &= \frac{\partial N(s, t)}{\partial s} + \frac{\partial N(s, t)}{\partial t}. \end{aligned}$$

**6.2. Multiple-delay case.** Recall that we define the intervals

$$H_i = \begin{cases} [-h_1, 0] & \text{if } i = 1, \\ [-h_i, -h_{i-1}) & \text{if } i = 2, \dots, k. \end{cases}$$

We first give the complete class of functions which define the Lyapunov function,  $V$ , and its derivative:

$$\begin{aligned}
 Y_1 = \Big\{ & M : [-h, 0] \rightarrow \mathbb{S}^{2n} \mid \\
 & M_{11}(t) \text{ is constant} && \text{for all } t \in [-h, 0], \\
 & M \text{ is } C^1 \text{ on } H_i && \text{for all } i = 1, \dots, k \Big\}, \\
 Y_2 = \Big\{ & N : [-h, 0] \times [-h, 0] \rightarrow \mathbb{S}^n \mid \\
 & N(s, t) = N(t, s)^T && \text{for all } s, t \in [-h, 0], \\
 & N \text{ is } C^1 \text{ on } H_i \times H_j && \text{for all } i, j = 1, \dots, k \Big\},
 \end{aligned}$$

and, for its derivative, define

$$\begin{aligned}
 Z_1 = \Big\{ & D : [-h, 0] \rightarrow \mathbb{S}^{(k+2)n} \mid \\
 & D_{ij}(t) \text{ is constant} && \text{for all } t \in [-h, 0] \\
 & && \text{for } i, j = 1, \dots, 3, \\
 & D \text{ is } C^0 \text{ on } H_i && \text{for all } i = 1, \dots, k \Big\}, \\
 Z_2 = \Big\{ & E : [-h, 0] \times [-h, 0] \rightarrow \mathbb{S}^n \mid \\
 & E(s, t) = E(t, s)^T && \text{for all } s, t \in [-h, 0], \\
 & E \text{ is } C^0 \text{ on } H_i \times H_j && \text{for all } i, j = 1, \dots, k \Big\}.
 \end{aligned}$$

Here  $M \in Y_1$  is partitioned according to

$$(13) \quad M(t) = \begin{bmatrix} M_{11} & M_{12}(t) \\ M_{21}(t) & M_{22}(t) \end{bmatrix},$$

where  $M_{11} \in \mathbb{S}^n$  and  $D \in Z_1$  are partitioned according to

$$(14) \quad D(t) = \begin{bmatrix} D_{11} & D_{12} & D_{13} & D_{14}(t) \\ D_{21} & D_{22} & D_{23} & D_{24}(t) \\ D_{31} & D_{32} & D_{33} & D_{34}(t) \\ D_{41}(t) & D_{42}(t) & D_{43}(t) & D_{44}(t) \end{bmatrix},$$

where  $D_{11}, D_{33}, D_{44} \in \mathbb{S}^n$  and  $D_{22} \in \mathbb{S}^{(k-1)n}$ . Let  $Y = Y_1 \times Y_2$  and  $Z = Z_1 \times Z_2$ . Notice that if  $M \in Y_1$ , then  $M$  need not be continuous at  $h_i$  for  $1 \leq i \leq k-1$ ; however, we require it to be right continuous at these points. We also define the derivative  $\dot{M}(t)$  at these points to be the right-hand derivative of  $M$ . We define the continuity and derivatives of functions in  $Y_2, Z_1$ , and  $Z_2$  similarly.

We define the jump values of  $M$  and  $N$  at the discontinuities as follows:

$$\Delta M(h_i) = \lim_{t \rightarrow (-h_i)_+} M(t) - \lim_{t \rightarrow (-h_i)_-} M(t)$$

for each  $i = 1, \dots, k - 1$ , and similarly we define

$$\Delta N(h_i, t) = \lim_{s \rightarrow (-h_i)_+} N(s, t) - \lim_{s \rightarrow (-h_i)_-} N(s, t).$$

The derivative of a Lyapunov function can be defined as a linear map  $Y \mapsto Z$ . This is made explicit in the following definition.

DEFINITION 9. Define the map  $L : Y \rightarrow Z$  by  $(D, E) = L(M, N)$  if for all  $t, s \in [-h, 0]$  we have

$$\begin{aligned} D_{11} &= A_0^T M_{11} + M_{11} A_0 \\ &\quad + \frac{1}{h} (M_{12}(0) + M_{21}(0) + M_{22}(0)), \\ D_{12} &= [M_{11} A_1 \quad \cdots \quad M_{11} A_{k-1}] \\ &\quad - [\Delta M_{12}(h_1) \quad \cdots \quad \Delta M_{12}(h_{k-1})], \\ D_{13} &= \frac{1}{h} (M_{11} A_k - M_{12}(-h)), \\ D_{22} &= \frac{1}{h} \operatorname{diag}(-\Delta M_{22}(h_1), \dots, -\Delta M_{22}(h_{k-1})), \\ D_{23} &= 0, \\ D_{33} &= -\frac{1}{h} M_{22}(-h), \\ D_{14}(t) &= N(0, t) + A_0^T M_{12}(t) - \dot{M}_{12}(t), \\ D_{24}(t) &= \begin{bmatrix} \Delta N(-h_1, t) + A_1^T M_{12}(t) \\ \vdots \\ \Delta N(-h_{k-1}, t) + A_{k-1}^T M_{12}(t) \end{bmatrix}, \\ D_{34}(t) &= A_k^T M_{12}(t) - N(-h, t), \\ D_{44}(t) &= -\dot{M}_{22}(t) \end{aligned}$$

and

$$E(s, t) = \frac{\partial N(s, t)}{\partial s} + \frac{\partial N(s, t)}{\partial t}.$$

Here  $M$  is partitioned as in (13),  $D$  is partitioned as in (14), and the remaining entries are defined by symmetry.

The map  $L$  is the Lie derivative operator applied to the set of functions specified by (3); this is stated precisely below. Notice that this implies that  $L$  is a linear map.



LEMMA 10. Suppose  $M \in Y_1$  and  $N \in Y_2$  and  $V$  is given by (3). Let  $(D, E) = L(M, N)$ . Then the Lie derivative of  $V$  on the vector field of (1) is given by

$$(15) \quad \dot{V}(\phi) = \int_{-h}^0 \begin{bmatrix} \phi(-h_0) \\ \vdots \\ \phi(-h_k) \\ \phi(s) \end{bmatrix}^T D(s) \begin{bmatrix} \phi(-h_0) \\ \vdots \\ \phi(-h_k) \\ \phi(s) \end{bmatrix} ds + \int_{-h}^0 \int_{-h}^0 \phi(s)^T E(s, t) \phi(t) ds dt.$$

**7. Stability conditions.** We can now use the results of the paper and the linear map from Definition 9 to give stability conditions.

THEOREM 11. Suppose there exist  $d \in \mathbb{N}$  and piecewise matrix polynomials  $M, T, N, D, U, E$  such that

$$\begin{aligned} M + \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix} &\in \Sigma_{2n,d}, \\ -D + \begin{bmatrix} U & 0 \\ 0 & 0 \end{bmatrix} &\in \Sigma_{(k+2)n,d}, \\ N &\in \Gamma_{n,d}, \\ -E &\in \Gamma_{n,d}, \\ (D, E) &= L(M, N), \\ \int_{-h}^0 T(s) ds &= 0, \\ \int_{-h}^0 U(s) ds &= 0, \\ M_{11} &\succ 0, \\ D_{11} &\prec 0. \end{aligned}$$

Then the system defined by (1) and (2) is exponentially stable.

*Proof.* Assume  $M, T, N, D, U, E$  satisfy the above conditions, and define the function  $V$  by (3). Then Lemma 10 implies that  $\dot{V}$  is given by (15). The function  $V$  is the sum of two terms, each of which is nonnegative. The first is nonnegative by Theorem 5, and the second is nonnegative since  $N \in \Gamma_{n,d}$ . The same is true for  $\dot{V}$ . The strict positivity conditions of equations (4) hold since  $M_{11} \succ 0$  and  $-D_{11} \succ 0$ , and Theorem 1 then implies stability.  $\square$

The feasibility conditions specified in Theorem 11 are semidefinite representable. In particular the condition that a piecewise polynomial matrix lie in  $\Sigma$  is a set of linear and positive semidefinite constraints on its coefficients. Similarly, the condition that  $T$  and  $U$  integrate to zero is simply a linear equality constraint on its coefficients. Standard semidefinite programming codes may therefore be used to efficiently find such piecewise polynomial matrices. Most such codes will also return a dual certificate of infeasibility if no such polynomials exist.

As in the Lyapunov analysis of nonlinear systems using sum-of-squares polynomials, the set of candidate Lyapunov functions is parameterized by the degree  $d$ . This allows one to search first over polynomials of low degree and increase the degree if that search fails.

There are various natural extensions of this result. The first is to the case of uncertain systems, where we would like to prove stability for all matrices  $A_i$  in some given semialgebraic set. This is possible by extending Theorem 11 to allow Lyapunov functions which depend polynomially on unknown parameters. A similar approach may be used to check stability for systems with uncertain delays. Additionally, stability of systems with distributed delays defined by polynomial kernels can be verified. It is also straightforward to extend the class of Lyapunov functions, since it is not necessary that each piece of the piecewise sums-of-squares functions be nonnegative on the whole real line. To do this, one can use techniques for parameterizing polynomials nonnegative on an interval; for example, every polynomial  $p(x) = f(x) - (x-1)(x-2)g(x)$  where  $f$  and  $g$  are sums of squares is nonnegative on the interval  $[1, 2]$ .

**8. Numerical examples.** In this section we present the results of some example computations using the approach described above. The computations were performed using MATLAB software, together with the SOSTOOLS [18] toolbox and SeDuMi [22] code for solving semidefinite programming problems.

**8.1. Illustration.** Consider the process of proving stability using the results of this paper. The following system has well-known stability properties:

$$(16) \quad \dot{x}(t) = -x(t-1).$$

A MATLAB implementation of the algorithm in this paper has been developed and is available online, along with several tools for polynomial matrix manipulation [17]. This implementation returns the following Lyapunov function for system (16). For symmetric matrices, subdiagonal elements are suppressed:

$$V(x) = \int_{-1}^0 \begin{bmatrix} x(0) \\ x(s) \end{bmatrix}^T M(s) \begin{bmatrix} x(0) \\ x(s) \end{bmatrix} ds + \int_{-1}^0 \int_{-1}^0 x(s)^T R(s, t) x(t) ds dt,$$

where

$$M(s) = \begin{bmatrix} 27.3 & -16.8 + 2.74s \\ & 24.3 + 8.53s \end{bmatrix}$$

and

$$R(s, t) = 9.08.$$

Positivity is proven using the function

$$t(s) = -.915 + 1.83s$$

and the sum-of-squares functions

$$Q(s) = \begin{bmatrix} 13 & -3.3 \\ & 12.2 \end{bmatrix} \geq 0$$

and

$$V(s) = Z(s)^T L Z(s),$$

where

$$Z(s) = \begin{bmatrix} 1 & s & 0 & 0 \\ 0 & 0 & 1 & s \end{bmatrix}^T$$

and

$$L = \begin{bmatrix} 28.215 & 5.585 & -16.8 & -1.973 \\ & 13 & 1.413 & -3.3 \\ & & 24.3 & 10.365 \\ & & & 12.2 \end{bmatrix} \geq 0.$$

This is because  $-s(s+1) \geq 0$  for  $s \in [-1, 0]$  and

$$M(s) + \begin{bmatrix} t(s) & 0 \\ 0 & 0 \end{bmatrix} = -s(s+1)Q(s) + V(s).$$

Furthermore,

$$R(s) = 9.08 \geq 0.$$

Therefore, by Theorems 11 and 5, the Lyapunov function is positive.

The derivative of the function is given by

$$\dot{V}(x) = \int_{-1}^0 \begin{bmatrix} x(0) \\ x(-1) \\ x(s) \end{bmatrix}^T D(s) \begin{bmatrix} x(0) \\ x(-1) \\ x(s) \end{bmatrix} ds,$$

where

$$-D(s) = \begin{bmatrix} 9.3 & 7.76 & -6.34 \\ & 15.77 & -7.72 + 2.74s \\ & & 8.53 \end{bmatrix}.$$

Negativity of the function is proven using the function

$$U(s) = \begin{bmatrix} .0055 + .011s & -.272 - .544s \\ & -.458 - .916s \end{bmatrix},$$

where

$$\int_{-1}^0 U(s) ds = 0,$$

and the sum-of-squares functions

$$X(s) = \begin{bmatrix} 8.86 & 1.90 & -3.23 \\ & 11.54 & -3.71 \\ & & 7.74 \end{bmatrix}$$

and

$$Y(s) = Z(s)^T L Z(s),$$

where

$$Z(s) = \begin{bmatrix} 1 & s & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & s & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & s \end{bmatrix}^T$$

and

$$L = \begin{bmatrix} 9.3 & 4.43 & 7.76 & .61896 & -6.34 & -1.0371 \\ & 8.86 & 1.281 & 1.9 & -2.1929 & -3.23 \\ & & 15.77 & 5.77 & -7.72 & .01171 \\ & & & 11.54 & -.98171 & -3.71 \\ & & & & 8.53 & 3.87 \\ & & & & & 7.74 \end{bmatrix} \geq 0.$$

Negativity follows since

$$-D(s) + \begin{bmatrix} U(s) & 0 \\ 0 & 0 \end{bmatrix} = -s(s + 1)X(s) + Y(s).$$

Therefore, by Theorem 11, the derivative of the Lyapunov function is negative. Stability follows by Theorem 1.

**8.2. A single delay.** We consider the following instance of a system with a single delay:

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -2 & 0.1 \end{bmatrix} x(t) + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} x(t - h).$$

For a given  $h$ , we use semidefinite programming to search for a Lyapunov function of degree  $d$  that proves stability. Using a bisection search over  $h$ , we determine the maximum and minimum  $h$  for which the system may be shown to be stable. These are shown below.

$d$	$h_{\min}$	$h_{\max}$
1	.10017	1.6249
2	.10017	1.7172
3	.10017	1.71785

When the degree  $d = 3$ , the bounds  $h_{\min}$  and  $h_{\max}$  are tight [4]. For comparison, we include here the bounds obtained by Gu, Kharitonov, and Chen [4] using a piecewise linear Lyapunov function with  $n$  segments.

$n$	$h_{\min}$	$h_{\max}$
1	.1006	1.4272
2	.1003	1.6921
3	.1003	1.7161

**8.3. Multiple delays.** Consider the system with two delays below:

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -1 & \frac{1}{10} \end{bmatrix} x(t) + \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix} x(t - \frac{h}{2}) + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} x(t - h).$$

As above, using a bisection search over  $h$ , we prove stability for the range of  $h$  below.

$d$	$h_{\min}$	$h_{\max}$
1	.20247	1.354
2	.20247	1.3722

Here for degree  $d = 2$ , the bounds obtained are tight. Again we include here bounds obtained by Gu, Kharitonov, and Chen [4] using a piecewise linear Lyapunov function with  $n$  segments.

$n$	$h_{\min}$	$h_{\max}$
1	.204	1.35
2	.203	1.372

**9. Summary.** In this paper we developed an approach for computing Lyapunov functions for linear systems with delay. In general this is a difficult computational problem, and certain specific classes of this problem are known to be NP-hard [24]. However, the set of Lyapunov functions is convex, and this enables us to effectively test feasibility of a subset of this set. Specifically, we parameterize a convex set of positive quadratic functions using the set of polynomials as a basis, and the main results here are Theorems 5 and 7. Combining these results with the well-known approach using sum-of-squares polynomials allows one to use standard semidefinite programming software to compute Lyapunov functions. This gives a nested sequence of computable sufficient conditions for stability of linear delay systems, indexed by the degree of the polynomial. In principle this enables searching over increasing degrees to find a Lyapunov function, although further work is needed to enhance existing semidefinite programming codes to make this more efficient in practice.

It is possible that Theorems 5 and 7 are applicable more widely, specifically to stability analysis of nonlinear and partial differential systems, as well as to controller synthesis. One specific extension that is possible is analysis of delay systems with uncertain parameters, for which sufficient conditions for existence of a Lyapunov function may be given using convex relaxations. It is also possible to analyze stability of nonlinear delay systems, in the case that the dynamics are defined by polynomial delay differential equations. Preliminary work has also been done on stability analysis of certain types of partial differential equations. Further extensions to allow synthesis of stabilizing controllers are of interest and may be possible.

## REFERENCES

- [1] G. CHESI, *On the gap between positive polynomials and SOS of polynomials*, IEEE Trans. Automat. Control, 52 (2007), pp. 1066–1072.
- [2] G. CHESI, A. TESI, A. VICINO, AND R. GENESIO, *On convexification of some minimum distance problems*, in Proceedings of the 5th European Control Conference, Karlsruhe, Germany, 1999.
- [3] M. D. CHOI, T. Y. LAM, AND B. REZNICK, *Real zeros of positive semidefinite forms I*, Math. Z., 171 (1980), pp. 1–26.
- [4] K. GU, V. L. KHARITONOV, AND J. CHEN, *Stability of Time-Delay Systems*, Birkhäuser, Boston, 2003.

- [5] J. K. HALE AND S. M. VERDUYN LUNEL, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.
- [6] D. HENRION AND A. GARULLI, EDS., *Positive Polynomials in Control*, Lecture Notes in Control and Inform. Sci. 312, Springer-Verlag, New York, 2005.
- [7] V. L. KHARITONOV AND D. HINRICHSSEN, *Exponential estimates for time delay systems*, Systems Control Lett., 53 (2004), pp. 395–405.
- [8] M. KOJIMA, *Sums of Squares Relaxations of Polynomial Semidefinite Programs*, Research Report B-397, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo, Japan, 2003.
- [9] V. KOLMANOVSKII AND A. MYSHKIS, *Introduction to the Theory and Applications of Functional Differential Equations*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.
- [10] N. N. KRASOVSKII, *Stability of Motion*, Stanford University Press, Palo Alto, CA, 1963.
- [11] J. B. LASSERRE, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
- [12] Y. NESTEROV, *Squared functional systems and optimization problems*, in High Performance Optimization, Appl. Optim. 33, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 405–440.
- [13] S.-I. NICULESCU, *Delay Effects on Stability: A Robust Control Approach*, Lecture Notes in Control and Inform. Sci. 269, Springer-Verlag, New York, 2001.
- [14] A. PAPACHRISTODOULOU, M. M. PEET, AND S. LALL, *Stability analysis of nonlinear time-delay systems*, IEEE Trans. Automat. Control, to appear.
- [15] P. PARRILO, *On a decomposition of multivariable forms via LMI methods*, in Proceedings of the American Control Conference, 2000, pp. 322–326.
- [16] P. A. PARRILO, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 2000.
- [17] M. PEET, *Web Site for Matthew M. Peet*, <http://mmae.iit.edu/~mpeet>, 2008.
- [18] S. PRAJNA, A. PAPACHRISTODOULOU, AND P. A. PARRILO, *Introducing SOSTOOLS: A general purpose sum of squares programming solver*, in Proceedings of the IEEE Conference on Decision and Control, 2002.
- [19] I. M. REPIN, *Quadratic Liapunov functionals for systems with delay*, J. Appl. Math. Mech., 29 (1965), pp. 669–672.
- [20] F. RIESZ AND B. SZ-NAGY, *Functional Analysis*, Dover, New York, 1990.
- [21] C. W. SCHERER AND C. W. J. HOL, *Asymptotically exact relaxations for robust LMI problems based on matrix-valued sum-of-squares*, in Proceedings of the International Symposium on Mathematical Theory of Networks and Systems (MTNS), 2004.
- [22] J. F. STURM, *Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11/12 (1999), pp. 625–653.
- [23] W. TAN, *Nonlinear Control Analysis and Synthesis Using Sum-of-Squares Programming*, Ph.D. thesis, University of California, Berkeley, CA, 2006.
- [24] O. TOKER AND H. OZBAY, *Complexity issues in robust stability of linear delay-differential systems*, Math. Control Signals Systems, 9 (1996), pp. 386–400.
- [25] L. VANDENBERGHE AND S. BOYD, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

## STABILITY AND ASYMPTOTIC OPTIMALITY OF GENERALIZED MAXWEIGHT POLICIES\*

SEAN MEYN†

**Abstract.** It is shown that stability of the celebrated MaxWeight or back pressure policies is a consequence of the following interpretation: either policy is myopic with respect to a surrogate value function of a very special form, in which the “marginal disutility” at a buffer vanishes for a vanishingly small buffer population. This observation motivates the *h*-MaxWeight policy, defined for a wide class of functions *h*. These policies share many of the attractive properties of the MaxWeight policy as follows: (i) Arrival rate data is not required in the policy. (ii) Under a variety of general conditions, the policy is stabilizing when *h* is a perturbation of a monotone linear function, a monotone quadratic, or a monotone Lyapunov function for the fluid model. (iii) A perturbation of the relative value function for a workload relaxation gives rise to a myopic policy that is approximately average-cost optimal in heavy traffic, with logarithmic regret. The first results are obtained for a general Markovian network model. Asymptotic optimality is established for a general Markovian scheduling model with a single bottleneck, and with homogeneous servers.

**Key words.** queueing networks, routing, scheduling, optimal control

**AMS subject classifications.** Primary, 90B35, 68M20, 90B15; Secondary, 93E20, 60J20

**DOI.** 10.1137/06067746X

**1. Introduction.** While it is popular to cite the curse of dimensionality when discussing optimization of stochastic networks, there are many classes of effective policies that are easily implemented, require limited information, and have other attractive properties. A well-known example is the MaxWeight policy of Tassiulas and Ephremides [53]. This policy can be interpreted as a myopic policy for the associated fluid model with respect to a quadratic function,

$$(1) \quad h(x) = \frac{1}{2}x^T D x, \quad x \in \mathbb{R}^\ell,$$

with  $D > 0$  a diagonal matrix. Stability theory for this and similar classes of policies has been extended in multiple directions over the past 15 years [20, 52, 48, 17, 13], and in particular these policies are known to be approximately optimal in heavy traffic under certain conditions on the network; see [55, 49, 33] and the recent comprehensive results by Dai and Lin [14].

These results are fragile: Diagonal quadratics are one of a very few function classes for which the myopic policy is known to be stabilizing for general classes of network models. In contrast, stability of the fluid model under a myopic policy is virtually universal [10, 7, 35].

It is important to find broader classes of stabilizing policies for complex networks. It is known that the MaxWeight policy can perform poorly since it makes use of so little information [50].

---

\*Received by the editors December 11, 2006; accepted for publication (in revised form) September 1, 2008; published electronically January 9, 2009. Financial support from the National Science Foundation (grant ECS-0523620) and from DARPA (under the ITMANET program RK 2006-07284) is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

<http://www.siam.org/journals/sicon/47-6/67746.html>

†Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (meyn@uiuc.edu).

To explain the gap between the stochastic and deterministic models, we consider some simple, well-known examples. The stochastic model favored in this paper is the controlled random walk (CRW) model in which the queue length process  $\mathbf{Q}$  evolves on  $\mathbb{Z}_+^\ell$  in discrete time according to the recursion,

$$(2) \quad Q(t+1) = Q(t) + B(t+1)U(t) + A(t+1), \quad t \geq 0, \quad Q(0) = x.$$

The allocation sequence  $\mathbf{U}$  evolves on  $\mathbb{Z}_+^{\ell_u}$  for some integer  $\ell_u$ ; the arrival sequence  $\mathbf{A}$  is  $\ell$ -dimensional, and  $\mathbf{B}$  is an  $\ell \times \ell_u$  matrix sequence; and each sequence has integer-valued entries.

The allocation sequence  $\mathbf{U}$  is subject to both integral and linear constraints of the form  $U(t) \in \mathbf{U}_\diamond$ ,  $t \geq 0$ , where

$$(3) \quad \mathbf{U}_\diamond := \{u \in \{0, 1\}^{\ell_u} : Cu \leq \mathbf{1}\}.$$

The  $\ell_m \times \ell_u$  matrix  $C$  is called the constituency matrix: Each of the rows of  $C$  corresponds to a “resource” in the network. Its entries are assumed to be binary.

The CRW model is a generalization of the network model of Lippman obtained via uniformization [31]. Versions of this model appear throughout the communications and operations research literature, and in particular appear in the paper [53] that introduced the MaxWeight policy.

The fluid model  $\mathbf{q} = \{q(t) : t \geq 0\}$  satisfies the linear equations

$$(4) \quad q(t) = x + Bz(t) + \alpha t, \quad t \geq 0,$$

where  $x \in \mathbb{R}_+^\ell$  is the initial condition,  $B$  and  $\alpha$  are the mean values of  $B(t)$  and  $A(t)$ , respectively, and  $\mathbf{z}$  is the cumulative allocation process evolving on  $\mathbb{R}_+^{\ell_u}$ . The constraints on  $\mathbf{z}$  are analogous to those on  $\mathbf{U}$ . We let  $\mathbf{U}$  denote the convex hull  $\mathbf{U} := \text{conv}(\mathbf{U}_\diamond)$  and assume that for each  $0 \leq t_0 < t_1$ ,

$$\frac{z(t_1) - z(t_0)}{t_1 - t_0} \in \mathbf{U}.$$

The fluid model (4) is also expressed as the ODE model,

$$(5) \quad \frac{d^+}{dt} q(t) = B\zeta(t) + \alpha, \quad t \geq 0,$$

where  $\zeta(t) \in \mathbf{U}$  denotes the allocation rate vector at time  $t$ , and the “+” denotes right derivative.

Suppose that  $h: \mathbb{R}_+^\ell \rightarrow \mathbb{R}_+$  is any  $C^1$  function that vanishes only at the origin. The *h-myopic policy* is defined for the fluid and stochastic models by the respective feedback laws,

$$(6) \quad \phi^F(x) = \arg \min_{\zeta \in \mathbf{U}(x)} \langle \nabla h(x), B\zeta + \alpha \rangle,$$

$$(7) \quad \phi^D(x) = \arg \min_{u \in \mathbf{U}_\diamond(x)} \mathbb{E}[h(Q(t+1)) - h(Q(t)) \mid Q(t) = x, U(t) = u],$$

where the “(x)” is used to capture boundary constraints,

$$(8) \quad \begin{aligned} \mathbf{U}_\diamond(x) &= \{u \in \mathbf{U} : (B(t)u)_i \geq 0 \text{ a.s. when } x_i = 0\}, \\ \mathbf{U}(x) &= \{u \in \mathbf{U} : v_i := (Bu + \alpha)_i \geq 0 \text{ when } x_i = 0\}. \end{aligned}$$



The MaxWeight policy coincides with the  $h$ -myopic policy for the fluid model when  $h$  is equal to the quadratic (1). This motivates an alternative policy for the stochastic model, the  $h$ -MaxWeight policy,

$$(9) \quad \phi^{\text{MW}}(x) = \arg \min_{u \in U_o(x)} \langle \nabla h(x), Bu + \alpha \rangle, \quad x \in \mathbb{Z}_+^\ell.$$

Note that  $\phi^{\text{MW}} = \phi^{\text{D}}$  when  $h$  is a linear function of  $x$ .

The policy (6) is stabilizing for the fluid model under mild assumptions on the function  $h$  (see [10] and [7, Thm. 12.5] for linear functions and [35, Proposition 11] for a smooth norm on  $\mathbb{R}^\ell$ ). The proof is based on establishing that the “drift” defined by

$$(10) \quad \frac{d^+}{dt} h(q(t)) = \left\langle \nabla h(q(t)), \frac{d^+}{dt} q(t) \right\rangle, \quad t \geq 0,$$

is strictly negative when  $q(t) \neq 0$ .

The  $h$ -myopic policy (7) for the stochastic model may or may not be stabilizing, depending upon the particular network and the structure of the function  $h$ . One difficulty is that the corresponding drift for the stochastic model,

$$(11) \quad \mathbb{E}[h(Q(t+1)) - h(Q(t)) \mid Q(t) = x, U(t) = u],$$

can be positive for certain values of  $x$  on the boundary of the state space. This important distinction between the two models is illustrated in the following two examples.

*Instability in the model of Rybko and Stolyar.* Consider the model of Kumar, Seidman, Rybko, and Stolyar shown in Figure 1 [27, 47]. A typical choice for  $h$  is a cost function  $c$ , and a typical cost function in network applications is the  $\ell_1$  norm,  $c(x) = |x| = \sum x_i$ . With  $h = c$ , the myopic policy for the fluid model gives priority to the exit buffers if no machine is starved of work. Suppose that the parameters satisfy

$$(12) \quad \mu_1 > \mu_2 \text{ and } \mu_3 > \mu_4.$$

If, for example,  $x_1 > 0$  and  $x_4 > 0$ , yet  $x_2 = x_3 = 0$ , we then have

$$\phi_1^{\text{F}}(x) = \mu_2 \mu_1^{-1}, \quad \phi_4^{\text{F}}(x) = 1 - \phi_1^{\text{F}}(x).$$

The  $h$ -myopic policy for the stochastic model is very different: The optimization (7) defines  $\phi_4^{\text{F}}(x) = 1$  if  $x_4 \geq 1$ , and  $\phi_2^{\text{F}}(x) = 1$  if  $x_2 \geq 1$ . This is precisely the policy found to be destabilizing in [47].

*Work stoppage under a myopic policy.* The  $h$ -myopic policy may be entirely irrational. Consider the pair of queues in tandem illustrated in Figure 2. Suppose that a linear cost function is given  $c(x) = c_1 x_1 + c_2 x_2$ , with  $c_2 > c_1$ . The  $h$ -myopic policy for the fluid model with  $h = c$  is nonidling at Station 2, while at Station 1,

$$(13) \quad \phi_1^{\text{F}}(x) = \begin{cases} 0 & \text{if } x_2 > 0, \\ \min(1, \mu_2 \mu_1^{-1}) & \text{if } x_2 = 0, x_1 > 0. \end{cases}$$

The  $h$ -myopic policy is pathwise optimal when  $\mu_1 \geq \mu_2$ .

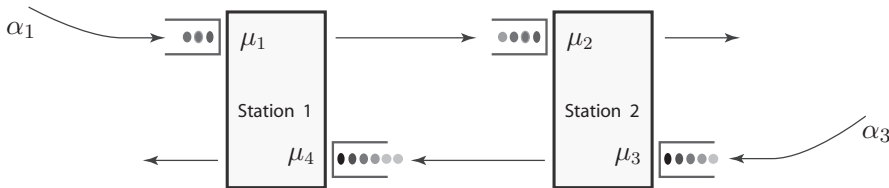


FIG. 1. The Kumar–Seidman–Rybko–Stolyar model.

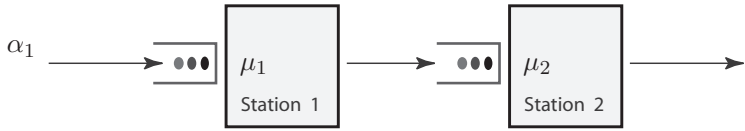


FIG. 2. *Tandem queues.*

For a CRW model defined consistently with the fluid model, we have for  $x \in \mathbb{Z}_+^2$ ,

$$\begin{aligned} \phi^{\text{MW}}(x) &= \phi^{\text{D}}(x) = \arg \min_{u \in \mathcal{U}_\diamond(x)} \mathbb{E}[c(Q(t+1)) \mid Q(t) = x, U(t) = u] \\ &= \arg \min_{u \in \mathcal{U}_\diamond(x)} (c_1(\alpha_1 - \mu_1 u_1) + c_2(\mu_1 u_1 - \mu_2 u_2)). \end{aligned}$$

At Station 2 this policy is nonidling, while at Station 1,

$$\phi_1^{\text{MW}}(x) = \arg \min_{u \in \mathcal{U}_\diamond(x)} ((c_2 - c_1)\mu_1 u_1).$$

That is, Station 1 is *always idle* under our assumption that  $c_2 > c_1$ !

Instability is a consequence of *additional constraints in the stochastic model*. The choices are limited in the CRW model, so from certain states on the boundary it is not possible to find an allocation  $u$  such that the drift in (11) is negative. Why then is this possible when  $h$  is a diagonal quadratic?

We show in this paper that the key property that is required is the derivative condition,

$$(14) \quad \frac{\partial}{\partial x_j} h(x) = 0 \quad \text{when } x_j = 0.$$

For a quadratic we have  $\nabla h(x) = Dx$ , and hence (14) does hold when  $D$  is diagonal. With  $h$  interpreted as an approximate value function, the derivative  $\frac{\partial}{\partial x_j} h(x)$  represents the “marginal disutility” of an additional increment of inventory at buffer  $j$ . If this marginal disutility is zero, then it is reasonable to shift inventory to this buffer when possible. Thus starvation of resources is avoided, which is the cause of instability in these two examples.

In this paper we make these informal observations precise. Moreover, to obtain a wide class of policies we describe a perturbation technique used to modify a given function so that (14) holds. Suppose that  $c$  is a norm on  $\mathbb{R}^\ell$ , such as  $c(x) = \sum |x_i|$ , and that  $h_0: \mathbb{R}^\ell \rightarrow \mathbb{R}_+$  is any  $C^1$  function that satisfies the dynamic programming *inequality* for the fluid model,

$$(15) \quad \min_{u \in \mathcal{U}(x)} \langle \nabla h_0(x), Bu + \alpha \rangle \leq -c(x), \quad x \in \mathbb{R}_+^\ell.$$

With  $\phi^{\text{F}}$  defined in (6) using  $h_0$ , and  $v := B\phi^{\text{F}}(x) + \alpha$ , the bound (15) is equivalent to the functional inequality  $\langle \nabla h_0, v \rangle \leq -c$ .

For example, if  $\|\cdot\|_h$  is any norm on  $\mathbb{R}^\ell$  that is monotone and  $C^1$  on  $\mathbb{R}_+^\ell$ , then (15) holds with  $h_0(x) = \frac{1}{2}\|x\|_h^2$  and  $c(\cdot) = \varepsilon_0\|\cdot\|_h$  for some  $\varepsilon_0 > 0$ . Another solution to the dynamic programming inequality is the quadratic (1), in which  $D$  is not necessarily diagonal but satisfies  $D_{ij} \geq 0$  and  $D_{ii} > 0$  for each  $i, j$ . In some cases a fluid value function is piecewise quadratic,  $C^1$ , and satisfies (15) with equality. An example is contained in section 2.2.2.

A perturbation of  $h_0$  is obtained through a change of variables: For fixed  $\theta \geq 1$ , we denote

$$(16) \quad \tilde{x}_i := x_i + \theta(e^{-x_i/\theta} - 1) \quad \text{for any } i \text{ and } x$$

and let  $\tilde{x}$  denote the corresponding vector  $\tilde{x} := (\tilde{x}_1, \dots, \tilde{x}_\ell)^\top \in \mathbb{R}_+^\ell$ . The function  $h$  is then defined by

$$(17) \quad h(x) = h_0(\tilde{x}), \quad x \in \mathbb{R}_+^\ell.$$

An application of the chain rule shows that (14) holds. The first main result of this paper is based on this observation.

**THEOREM 1.1.** *Consider the model (2) satisfying the following conditions:*

(i) *The independent and identically distributed (i.i.d.) process  $(\mathbf{A}, \mathbf{B})$  has integer entries, and a finite second moment.*

(ii)  *$B_{ij}(t) \geq -1$  for each  $i, j$ , and  $t$ , and for each  $j \in \{1, \dots, \ell_u\}$  there exists a unique value  $i_j \in \{1, \dots, \ell\}$  satisfying*

$$(18) \quad B_{ij}(t) \geq 0 \quad \text{a.s. } i \neq i_j.$$

(iii) *The function  $h_0: \mathbb{R}_+^\ell \rightarrow \mathbb{R}_+$  satisfies the following:*

(a) *Smoothness: The gradient  $\nabla h_0$  is Lipschitz continuous.*

(b) *Monotonicity:  $\nabla h_0(x) \in \mathbb{R}_+^\ell$  for  $x \in \mathbb{R}_+^\ell$ .*

(c) *The dynamic programming inequality (15) holds, with  $c$  a norm on  $\mathbb{R}^\ell$ .*

*Then, there exists  $\theta_0 < \infty$  and  $\bar{\eta}_h < \infty$  such that for any  $\theta \geq \theta_0$ , the following bound holds under the  $h$ -MaxWeight policy:*

$$(19) \quad \mathbb{E}[h(Q(t+1)) - h(Q(t)) \mid Q(t) = x] \leq -\frac{1}{2}c(x) + \frac{1}{2}\bar{\eta}_h.$$

*Consequently,*

$$(20) \quad n^{-1}\mathbb{E}\left[\sum_{t=0}^{n-1} c(Q(t)) \mid Q(t) = x\right] \leq 2n^{-1}h(x) + \bar{\eta}_h, \quad n \geq 1, \quad x \in \mathbb{Z}_+^\ell.$$

*Proof.* This is based on results obtained in section 2.2.2: Combining the bounds obtained in Lemmas 2.10 and 2.11 gives, under the  $h$ -MaxWeight policy, for each  $x \in \mathbb{Z}_+^\ell$ ,

$$\mathbb{E}[h(Q(t+1)) - h(Q(t)) \mid Q(t) = x] \leq -c(x) + k_{2,10} \log(1 + \|x\|) + k_{2,11}(1 + \theta^{-1}\|x\|),$$

where the constants are independent of  $\theta$ . Choosing  $\theta > 2k_{2,11}$ , we obtain the bound (19).

Equation (20) then follows from the comparison theorem, Theorem 2.2.  $\square$

Assumption (ii) implies that the matrix  $-B(t)$  is *Leontief* with probability one for each  $t$ , and that its expectation  $-B = -\mathbb{E}[B(t)]$  is also Leontief. Bramson and Williams [6] call a network *unitary* if assumption (ii) holds and, in addition, the rows of  $C$  are orthogonal (interpreted as the absence of “simultaneous resource possession”). Relaxations of assumption (ii) that imply stability of the MaxWeight policy are contained in [13, 14] (called *maximum pressure policies* in these papers).

For networks that are unitary, the  $h$ -MaxWeight policy has a simple representation in terms of the *generalized Klimov indices*,

$$(21) \quad \Theta_j(x) := -\sum_i B_{ij} \frac{\partial}{\partial x_i} h(x), \quad x \in \mathbb{Z}_+^\ell, \quad j \in \{1, \dots, \ell_u\}.$$

For a unitary model, for any  $j$  we denote by  $s(j) \in \{1, \dots, \ell_m\}$  the unique value of  $s$  satisfying  $C_{s,j} = 1$ .

PROPOSITION 1.2. *Suppose that the CRW model is unitary. Suppose, moreover, that  $h$  is  $C^1$ , monotone, and satisfies the boundary conditions (14). Then, the  $h$ -MaxWeight policy can be described as follows: For each  $s \in \{1, \dots, \ell_m\}$  and  $x \in \mathbb{Z}_+^\ell$ , denote  $\Theta_s^*(x) := \max\{\Theta_j(x) : s(j) = s\}$ . If  $\Theta_s^* < 0$ , then  $U_j(t) = 0$  whenever  $s(j) = s$ . Otherwise, priority is giving to buffers that achieve the maximum,*

$$\sum \{U_j(t) : s(j) = s, \Theta_j(x) = \Theta_s^*\} = 1.$$

*Proof.* For a unitary model, the optimization (9) decouples into  $\ell_m$  separate optimization problems, each with a single linear constraint obtained from the respective row of  $C$ . The proof is completed on noting that  $-\Theta_j$  is the coefficient of  $u_j$  in the objective function of (9).  $\square$

A drawback to Theorem 1.1 is that stability holds only for  $\theta > 0$  sufficiently large. Section 2.3 considers the alternative change of variables  $\tilde{x}_i := x_i \log(1 + x_i/\theta)$ ,  $i = 1, \dots, \ell$ . Theorem 2.14 shows that the resulting  $h$ -MaxWeight policy is stabilizing, for any fixed  $\theta > 0$ , in the sense that a version of (19) holds.

The inequality (19) is a *Lyapunov drift condition* of the form developed in [41, 18] and is also similar to the bounds used in [11, 4, 28, 26, 44] to obtain performance bounds for networks. Under natural assumptions on the model, the bound (19) implies that the controlled network is geometrically ergodic, so that the mean  $\mathbb{E}[c(Q(t))]$  converges to its steady-state value geometrically fast from each initial condition [41, 43, 40, 26, 35, 19]. Proposition 2.9 below contains sufficient conditions for geometric ergodicity for a particular version of the  $h$ -MaxWeight policy.

In section 3 we move to an asymptotic, heavy-traffic setting to obtain finer performance bounds. Dai and Lin's recent paper [14] contains a comprehensive survey on the theory of networks in heavy traffic. While the results in section 3 use language and some results from the heavy-traffic literature, the goals and conclusions are very different from those of [14] or any other papers from this literature.

A heavy-traffic analysis is based on the construction of a one-dimensional parameterized family of networks with increasing load. Let  $\kappa > 0$  denote the parameter, and assume that the load increases to one as  $\kappa \rightarrow \infty$ . Letting  $Q^\kappa(t; x)$  denote the queue-length process for the  $\kappa$ th network at time  $t$  with initial condition  $x$ , a “central limit” scaling is applied,

$$(22) \quad Q^{\kappa, \kappa}(t; x) = \frac{1}{\kappa} Q^\kappa(\kappa^2 t; \kappa x).$$

This is defined for all  $t \in \mathbb{R}_+$  via linear interpolation.

In virtually all of the asymptotic results contained in the literature it is assumed that there is a single bottleneck in heavy traffic or, more generally, *complete resource pooling* [2, 49, 33, 1, 14]. Let  $\xi \in \mathbb{R}^\ell$  denote the corresponding workload vector, and assume its entries are nonnegative. Then the workload process  $\widehat{W}^\kappa(t; x) = \xi^T Q^\kappa(t; x)$  evolves on  $\mathbb{R}_+$  and can be compared to a minimal workload process  $\widehat{W}^\kappa(t; x)$ . Section 3 restricts to a simplified setting in which a realization of the minimal process evolves as a simple queue,

$$(23) \quad \widehat{W}^\kappa(t+1) = \widehat{W}^\kappa(t) - S_1(t+1)\mathbb{1}\{\widehat{W}^\kappa(t) \geq 1\} + L_1(t+1), \quad t \geq 0,$$

where  $(S_1, L_1)$  is i.i.d. on  $\mathbb{Z}_+^2$ , and  $S_1$  is Bernoulli (see (83)). The load is given by  $\rho_\bullet = \mathbb{E}[L_1(t)]/\mathbb{E}[S_1(t)]$ .

For a convex cost function  $c: \mathbb{R}_+^\ell \rightarrow \mathbb{R}_+$ , the *effective cost*  $\bar{c}: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is the value of the convex program,

$$(24) \quad \begin{aligned} \bar{c}(w) &= \min_x c(x) \\ \text{s.t.} \quad &\xi^\top x = w, \quad x \in \mathbb{R}_+^\ell. \end{aligned}$$

A sequence of policies is declared to have *heavy-traffic asymptotical optimality* (HTAO) if the following properties are verified.

**State space collapse.** For each  $t$ , the scaled queue-length process has asymptotically minimal cost, subject to its workload, in the sense that

$$\lim_{\kappa \rightarrow \infty} (c(Q^{\kappa, \kappa}(t; x)) - \bar{c}(W^{\kappa, \kappa}(t; x))) = 0,$$

where the convergence is in probability. This is called state space collapse because typically it implies that the queue-length processes converge to a one-dimensional subspace,

$$(25) \quad \lim_{\kappa \rightarrow \infty} \|Q^{\kappa, \kappa}(t; x) - \mathcal{X}^*(Q^{\kappa, \kappa}(t; x))\| = 0,$$

where  $\mathcal{X}^*$  denotes the projection. The vector  $\mathcal{X}^*(Q)$  is known as the *effective state* corresponding to  $Q$  (see (84)).

**Asymptotic minimality.** The scaled workload process  $W^{\kappa, \kappa}(t; x)$  and the scaled minimal workload process  $\widehat{W}^{\kappa, \kappa}(t; x)$ , defined as in (22), each converges in distribution to a reflected Brownian motion (RBM) with common drift and covariance.

A uniform version of HTAO is formulated in [36]. The paper considers multiclass networks with multiple bottlenecks and renewal inputs. Under the assumption that the effective cost is a monotone function of  $w$ , among other assumptions, the following uniform asymptotic bounds are obtained for a proposed policy: For any other policy, letting  $Q^{\kappa'}$  denote the resulting state process,

$$(26) \quad \begin{aligned} \frac{1}{T} \int_0^T c(Q^{\kappa}(t; x)) dt &\leq \frac{1}{T} \int_0^T c(Q^{\kappa'}(t; x)) dt + O(\log((1 - \rho_\bullet)^{-1})), \\ 0 \leq T &\leq \frac{1}{(1 - \rho_\bullet)^3}. \end{aligned}$$

The policies considered in [36] are based on those of [32], which are generalizations of the policy introduced in [2] for a particular example.

HTAO for MaxWeight and certain generalizations is established in the aforementioned papers [55, 49, 33, 14]. However, state space collapse for the MaxWeight policy is obtained with respect to an implicitly defined cost function [49]. In [14] an approximation is obtained: For a given linear cost function, the projection  $\mathcal{X}^*$  is of the form

$$\mathcal{X}^*(x) = (\xi^\top x) \frac{c_{i^*}}{\xi_{i^*}} \mathbf{1}^{i^*},$$

where  $i^* \in \arg \min \{c_i / \xi_i\}$ , and  $\mathbf{1}^i$  denotes the  $i$ th basis element in  $\mathbb{R}^\ell$ . For each  $\varepsilon > 0$ , the authors construct a version of the MaxWeight policy satisfying

$$(27) \quad \lim_{\kappa \rightarrow \infty} \mathbb{P}\{\|Q^{\kappa, \kappa}(t; x) - \mathcal{X}^*(Q^{\kappa, \kappa}(t; x))\| > \varepsilon \|Q^{\kappa, \kappa}(t; x)\|\} = 0.$$

The present paper is concerned with steady-state performance. The average cost is denoted

$$(28) \quad \eta = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[c(Q(t; x))],$$

where the limit is independent of  $x$  under the assumptions imposed in the main results. Note that HTAO as formulated above does not imply that the performance measured by average cost is approximately optimal, or even that  $\eta$  is finite. HTAO *suggests* a bound of the form

$$(29) \quad \eta \leq \hat{\eta}^* + o(\hat{\eta}^*),$$

where  $\hat{\eta}^* = \mathbb{E}[\bar{c}(\widehat{W}^\kappa(t))]$ ,  $\bar{c}$  denotes the effective cost, and the expectation is in steady state. This heuristic has been established rigorously only in special cases, based on the assumption of weak convergence of the scaled workload processes.

Suppose that the scaled workload processes converge in distribution to an RBM,

$$\widehat{W}^{\kappa, \kappa}(t; x) \xrightarrow{w} \widehat{W}^\infty(t; x), \quad \kappa \rightarrow \infty.$$

If, moreover, the scaled steady-state means are convergent,  $\mathbb{E}[\bar{c}(\widehat{W}^{\kappa, \kappa}(t; x))] \rightarrow \hat{\eta}^\infty = \mathbb{E}[\bar{c}(\widehat{W}^\infty(t))]$ ,  $\kappa \rightarrow \infty$ , then we can then reinterpret (29) as the limit

$$(30) \quad \lim_{\kappa \rightarrow \infty} \mathbb{E}[c(Q^{\kappa, \kappa}(t; x))] = \hat{\eta}^\infty,$$

where the expectations are all in steady state. The approximation (30) has been established for the single queue in the pioneering work of Kingman [24, 25], and for generalized Jackson networks by Gamarnik and Zeevi [19, Corollary 2, p. 73]. Stolyar conjectures in [49] that the invariant distributions for  $Q^{\kappa, \kappa}$  will converge weakly under the MaxWeight policy, provided there is complete resource pooling. This, together with uniform integrability, would imply the limit (30).

The limit (27) also suggests an asymptotic bound of the form

$$\eta \leq \hat{\eta}^* + O(\varepsilon \hat{\eta}^*),$$

but no such result is established in [14] or elsewhere.

Section 3 treats HTAO in an average-cost sense. The main result establishes *logarithmic regret*: For some fixed constant  $k_0 < \infty$ , independent of load,

$$(31) \quad \hat{\eta}^* \leq \eta \leq \hat{\eta}^* + k_0 \log(\hat{\eta}^*).$$

This is a significant refinement of (29) since  $\hat{\eta}^* \rightarrow \infty$  as the network load approaches unity. HTAO of the form (31) was obtained for the first time in [37] in several examples based on Lyapunov techniques similar to those used in this paper. The main idea is to take the optimal value function for the fluid model and introduce a penalty function to account for possible starvation when the state reaches the boundary of  $\mathbb{R}_+^\ell$ . In each example considered, a single policy is proposed that is independent of network load and is based on a switching curve with logarithmic growth. For example, for the tandem queues, the policy has the form

$$(32) \quad \text{serve buffer 1 if } x_2 \leq \beta \log(1 + x_1/\beta),$$

where  $\beta > 0$  is a sufficiently large constant.

The approach used in section 3 is similar: The function  $h_0$  is chosen as an approximation to a fluid value function. Rather than a penalty function, the change of variables (17) is used to construct a stabilizing  $h$ -MaxWeight policy and, under stronger conditions, HTAO with logarithmic regret.

The development is greatly simplified by imposing further structure on the model. The following assumptions are imposed in Theorem 3.1, the main result of section 3 that establishes logarithmic regret.

(HTAO 1). The network is described by a scheduling model with deterministic routing: For each  $i \in \{1, \dots, \ell\}$ , after processing at buffer  $i$  a customer either enters some buffer  $i_+ \in \{1, \dots, \ell\}$  or exits the system. The *routing matrix*  $R$  is the  $\ell \times \ell$  matrix defined for  $i, j \in \{1, \dots, \ell\}$  as  $R_{ij} = \mathbb{1}_{j=i_+}$ . The routing matrix satisfies  $R^\ell = 0_{\ell \times \ell}$ ; this ensures that each customer receives at most  $\ell$  services during its lifetime in the network. The CRW scheduling model is described by the recursion

$$(33) \quad Q(t+1) = Q(t) + \sum_{i=1}^{\ell} M_i(t+1)[-1^i + 1^{i_+}]U_i(t) + A(t+1), \quad Q(0) = x,$$

where  $M_i$  is a Bernoulli sequence for each  $i$ . The vector  $1^{i_+}$  is taken to be zero if customers exit the system following service at buffer  $i$ .

(HTAO 2). The network is of a generalized “Kelly type” in which service statistics are determined by the station, not the buffer. For each  $s \in \{1, \dots, \ell_m\}$ , Station  $s$  is said to be *homogeneous* if the random variables  $\{M_j(t) : s(j) = s\}$  are all identical. It is assumed that the network is homogeneous, meaning that each station satisfies this condition.

(HTAO 3). The cost function is linear.

(HTAO 4). There is a single bottleneck in heavy traffic, and the convex program (24) that defines the effective cost is assumed to possess a unique maximizer for each  $w \in \mathbb{R}_+$ .

We believe that many of these assumptions are nonessential. Potential extensions are surveyed in section 4.

The CRW scheduling model (33) is of the form (2) with

$$B(t) = -[I - R^T]M(t), \quad t \geq 1.$$

The Leontief condition (18) follows from the assumptions on  $M$  and  $R$ , where  $i_j = j$  for each  $j \in \{1, \dots, \ell\}$ . The vector load  $\rho \in \mathbb{R}_+^{\ell_m}$  is given by

$$(34) \quad \rho = CM^{-1}[I - R^T]^{-1}\alpha,$$

where  $M$  is a diagonal matrix, equal to the common mean of  $M(t)$ . The network load is the maximum,  $\rho_\bullet = \max_{1 \leq i \leq \ell} \rho_i$ .

Homogeneity is used to construct a one-dimensional workload relaxation satisfying the recursion (23). A one-dimensional workload process  $W$  corresponding to the most heavily loaded station is compared to its (minimal) relaxation  $\widehat{W}^*$ . For each  $t \geq 0$ , the lower bound holds with probability one,

$$(35) \quad W(t) \geq \widehat{W}^*(t), \quad t \geq 0,$$

under *any* policy for  $Q$ . This simplifies a proof of logarithmic regret since a lower bound on performance is explicit,  $\widehat{\eta}^* = \mathbb{E}[\widehat{c}(\widehat{W}^*(t))]$ .

A relaxation of the form (23) was introduced in the thesis of Laws [30] to obtain performance bounds (see also [29, 45]). Relaxations of this form and multidimensional extensions for the purposes of control synthesis and performance approximation are the subject of [9, 37, 23, 39].

Two significant contributions in the present paper are worth highlighting as follows:

(i) In each example considered in [37] the policy was explicitly constructed based on switching curves of the form (32). This requires considerable intuition for more complex models. In our two main results, Theorem 1.1 and Theorem 3.1, the policy is derived from the given value function via the minimization (9).

(ii) This is the first paper to give a completely general approach to HTAO for average cost, and in particular bounds of the form (31) for a general class of models.

The remainder of the paper is organized as follows. Section 2 concerns stability of  $h$ -MaxWeight policies. Asymptotic optimality is treated in section 3; Theorem 3.1 establishes a bound of the form (31) for a family of models with increasing load. Section 4 contains conclusions and possible extensions.

**2. MaxWeight policies.** In this section we consider the general CRW model (2) under the assumptions of Theorem 1.1: The sequence  $\{A(t), B(t)\}$  is i.i.d. with a finite second moment, and (18) holds for  $B(t)$ . This is a very general model, detailed as follows:

(i) Controlled routing from buffer  $i$  is modeled by allowing  $i_j = i$  for more than one  $j \in \{1, \dots, \ell_u\}$ . Then,  $U_j(t) = 1$  indicates that a customer at buffer  $i$  is routed to buffer  $i_j^+$ .

(ii) It is straightforward to model a flexible server, as found in the processor sharing models of [21, 22]; optimal policies for the fluid and CRW models are described in [35, p. 760] for a simple example.

(iii) Assembly-disassembly systems can be modeled. For example, following service completion at buffer  $i_j$ , a customer can spawn several new jobs. This is modeled by defining positive values of  $B_{ij}(t)$  for several values of  $i \neq i_j$ .

(iv) On interpreting an arrival as an increment of demand, the CRW model (2) can be used to model inventory systems. In this setting, an entry of  $U(t)$  can model an order for new raw material [8].

All of the policies considered in this paper are stationary and deterministic: For the CRW model it is assumed that there is a function  $\phi: \mathbb{Z}_+^\ell \rightarrow \mathcal{U}_\diamond$  such that

$$(36) \quad U(t) = \phi(Q(t)), \quad t \geq 0.$$

Hence  $\mathbf{Q}$  is a time-homogeneous Markov chain on  $\mathbb{Z}_+^\ell$ , with transition matrix denoted  $P$ . That is, for each  $x, y \in \mathbb{Z}_+^\ell$  and  $t \geq 0$ ,

$$P(x, y) = \mathbb{P}\{Q(t+1) = y \mid Q(t) = x\} = \mathbb{P}\{x + B(1)\phi(x) + A(1) = y\}.$$

For any function  $g: \mathbb{R}^\ell \rightarrow \mathbb{R}$ , the *generator*  $\mathcal{D}$  is defined as the difference operator,

$$\mathcal{D}g(x) := \mathbb{E}[g(Q(t+1)) - g(Q(t)) \mid Q(t) = x] = \sum_{y \in \mathbb{Z}_+^\ell} P(x, y)[g(y) - g(x)].$$

The analysis of the  $h$ -MaxWeight policy is based on bounds on the drift  $\mathcal{D}h$  for the stochastic model. The construction of these bounds is based on an analysis of the corresponding drift  $\langle B\phi^{\text{MW}}(x) + \alpha, \nabla h(x) \rangle$  for the fluid model. The first step is an



application of the mean value theorem, which implies the following representation for any  $Q(t) \in \mathbb{Z}_+^\ell$  and any  $t \geq 0$ :

$$(37) \quad \begin{aligned} h(Q(t+1)) - h(Q(t)) &= \langle \nabla h(\bar{Q}), \Delta(t+1) \rangle \\ &= \langle \nabla h(Q(t)), \Delta(t+1) \rangle \\ &\quad + \langle \nabla h(\bar{Q}) - \nabla h(Q(t)), \Delta(t+1) \rangle, \end{aligned}$$

where  $\Delta(t+1) := Q(t+1) - Q(t)$ , and  $\bar{Q} \in \mathbb{R}_+^\ell$  lies on the line connecting  $Q(t+1)$  and  $Q(t)$ . Consequently,

$$(38) \quad \mathcal{D}h(x) = \langle \nabla h(x), v \rangle + b_h,$$

where

$$(39) \quad v(x) = \mathbb{E}[\Delta(t+1) \mid Q(t) = x], \quad b_h(x) = \mathbb{E}[\langle \nabla h(\bar{Q}) - \nabla h(Q(t)), \Delta(t+1) \rangle \mid Q(t) = x].$$

To deduce stability based on (38) we obtain a bound on  $\langle \nabla h(x), v \rangle$  under the given policy. We then show that the second term  $b_h(x)$  is relatively small in magnitude.

We begin with a review of some Lyapunov theory for Markov and fluid models.

**2.1. Stochastic stability.** When does a stabilizing policy exist for a network model? How do we test for stability? To answer these questions we consider first the fluid model.

Denote the velocity set for the fluid model by

$$(40) \quad \mathbf{V} := \{v = B\zeta + \alpha : \zeta \in \mathbf{U}\}.$$

In the general setting of this section, the “load condition”  $\rho_\bullet < 1$  translates into the following:

$$(41) \quad \text{The origin is an interior point of } \mathbf{V}.$$

If (41) holds, then there exists  $\varepsilon > 0$  such that the vector  $v^x = -\varepsilon x/|x|$  lies in  $\mathbf{V}$  for each  $x \in \mathbb{R}_+^\ell$ . Setting  $\zeta = v^x$  from the initial condition  $q(0) = x$ , we have  $q(t) = q(0) - (\varepsilon/|x|)t$  for  $0 \leq t \leq |x|/\varepsilon$ . For a given policy for the fluid model, such as (7), we consider the value function,

$$(42) \quad J(x) = \int_0^\infty c(q(t; x)) dt.$$

We let  $J^*$  denote the minimum of (42) over all policies. Setting  $h = J$  in (6), we obtain the drift inequality,

$$\min_{u \in \mathbf{U}(x)} \langle \nabla J(x), Bu + \alpha \rangle \leq -c(x), \quad x \in \mathbb{R}_+^\ell,$$

and this inequality is an equality when  $J = J^*$ . We thereby obtain another class of functions that satisfy the dynamic programming inequality (15). The fluid value function is typically  $C^1$  in workload models [38].

We now turn to the CRW model. The general form of the Lyapunov condition considered here is condition (V3), or the special case known as Foster’s criterion [41]. All forms involve bounds on the generator applied to a function  $V: \mathbb{Z}_+^\ell \rightarrow \mathbb{R}_+$ .

(i) The controlled network satisfies *Foster's criterion* if there is a constant  $b < \infty$  and a finite set  $S \subset \mathbb{Z}_+^\ell$  such that

$$(V2) \quad \mathcal{D}V(x) \leq -1 + b\mathbb{1}_S(x), \quad x \in \mathbb{Z}_+^\ell.$$

(ii) The network satisfies condition (V3) if for a function  $f: \mathbb{Z}_+^\ell \rightarrow [1, \infty)$ ,

$$(V3) \quad \mathcal{D}V(x) \leq -f(x) + b\mathbb{1}_S(x), \quad x \in \mathbb{Z}_+^\ell,$$

where again  $b < \infty$  and  $S \subset \mathbb{Z}_+^\ell$  is a finite set.

We have the following simple but useful result relating the policies  $\phi^{\text{MW}}$  and  $\phi^{\text{D}}$ .

**PROPOSITION 2.1.** *Suppose that (V3) holds under the  $h$ -MaxWeight policy with  $V$  a constant multiple of  $h$ . Then the same bound holds for the  $h$ -myopic policy.*

*Proof.* If  $V = kh$  for some  $k < \infty$ , then we have under the  $h$ -myopic policy,

$$\begin{aligned} P_{\phi^{\text{D}}}V(x) &= k \arg \min_{u \in \mathcal{U}_\circ(x)} \mathbb{E}[h(Q(t+1)) \mid Q(t) = x, U(t) = u] \\ &\leq k\mathbb{E}[h(Q(t+1)) \mid Q(t) = x, U(t) = \phi^{\text{MW}}(x)] \\ &= P_{\phi^{\text{MW}}}V(x) \leq V(x) - f(x) + b\mathbb{1}_S(x). \quad \square \end{aligned}$$

The most common approach to establishing (V3) is to construct a function  $h: \mathbb{Z}_+^\ell \rightarrow (0, \infty)$  and a constant  $\bar{\eta} < \infty$  that solve the *Poisson inequality*,

$$(43) \quad \mathcal{D}h \leq -c + \bar{\eta}.$$

If  $c$  has bounded sublevel sets (e.g.,  $c$  defines a norm on  $\mathbb{R}^\ell$ ), then this implies (V3) with  $V = h$ ,  $f = 1 + c/2$ ,  $b = \bar{\eta} + 1$ , and  $S$  is the sublevel set  $S = \{x : c(x) \leq 2(\bar{\eta} + 1)\}$ . The comparison theorem implies that the steady-state mean of  $c$  is bounded by  $\bar{\eta}$  when (43) holds. Theorem 2.2 is also the most common approach to obtaining bounds on expectations involving stopping times. For a proof see [41].

**THEOREM 2.2** (comparison theorem). *Suppose that the nonnegative functions  $V, f, g$  satisfy the bound,*

$$(44) \quad \mathcal{D}V \leq -f + g.$$

*Then for each  $x \in \mathbb{Z}_+^\ell$  and any stopping time  $\tau$ , we have*

$$\mathbb{E}_x \left[ \sum_{t=0}^{\tau-1} f(Q(t)) \right] \leq V(x) + \mathbb{E}_x \left[ \sum_{t=0}^{\tau-1} g(Q(t)) \right].$$

The average cost is finite under (V3), provided  $f$  dominates the cost function. The following result follows from Theorems 14.0.1 and 17.0.1 of [41]. A sufficient condition for 0-irreducibility is given in Proposition 3.2.

**THEOREM 2.3.** *Consider the CRW model (2) controlled using a stationary policy. Suppose that there exists a solution to (V3) satisfying  $k_0^{-1}\|x\| \leq c(x) \leq k_0 f(x)$  for some  $k_0 < \infty$  and all  $x \in \mathbb{Z}_+^\ell$ . Suppose, moreover, that the controlled network is 0-irreducible: For each  $x \in \mathbb{Z}_+^\ell$ ,*

$$(45) \quad \sum_{t=0}^{\infty} \mathbb{P}\{Q(t) = \mathbf{0} \mid Q(0) = x\} > 0,$$

*and  $P(\mathbf{0}, \mathbf{0}) = \mathbb{P}\{A(t) = \mathbf{0}\} > 0$ . Then the following hold:*

(i)  $\mathbf{Q}$  is an aperiodic Markov chain with unique invariant measure  $\pi$ . The average cost defined in (28) is finite, is independent of initial condition, and coincides with the mean with respect to  $\pi$ :

$$\eta = \pi(c) := \sum \pi(x)c(x).$$

(ii) The law of large numbers holds: For each initial condition,

$$\eta(n) := n^{-1} \sum_{t=0}^{n-1} c(Q(t)) \rightarrow \eta, \quad n \rightarrow \infty \text{ a.s.}$$

(iii) The mean ergodic theorem holds: For each initial condition,

$$\mathbb{E}[c(Q(t))] \rightarrow \eta, \quad t \rightarrow \infty.$$

The simplest example is the CRW model for the single server queue defined by the recursion

$$(46) \quad Q(t+1) = Q(t) - S(t+1)U(t) + A(t+1), \quad t \in \mathbb{Z}_+,$$

with given initial condition  $Q(0) = x \in \mathbb{Z}_+$ . A solution to Poisson's inequality (43) is obtained with  $c(x) \equiv x$  and  $V = J^*$ , where the fluid value function is quadratic,

$$(47) \quad J^*(x) = \frac{1}{2} \frac{x^2}{\mu - \alpha}.$$

Theorem 2.4 establishes formulae for the steady-state mean as well as the associated solution to Poisson's equation with  $c$  the identity function ( $c(x) = x$  for  $x \in \mathbb{R}_+$ ),

$$(48) \quad \mathcal{D}h(x) = -x + \eta, \quad x \in \mathbb{Z}_+.$$

The formula (50) for the steady-state mean may be viewed as an analogue of the celebrated Pollaczek-Khintchine formula for the M/G/1 queue. The proof is based on refinements of the comparison theorem applied to the function  $V = J^*$  [37, 39].

THEOREM 2.4. Consider the CRW queueing model (46) satisfying  $\rho = \alpha/\mu < 1$ , and define

$$(49) \quad m^2 = \mathbb{E}[(S(1) - A(1))^2], \quad m_A^2 = \mathbb{E}[A(1)^2], \quad \sigma^2 = \rho m^2 + (1 - \rho)m_A^2.$$

Then

(i) there is a unique invariant probability measure  $\pi$  on  $\mathbb{Z}_+$ , with steady-state mean

$$(50) \quad \eta := \mathbb{E}_\pi[Q(0)] = \frac{1}{2} \frac{\sigma^2}{\mu - \alpha};$$

(ii) a solution to Poisson's equation (48) is the quadratic

$$(51) \quad h(x) = J^*(x) + \frac{1}{2} \mu^{-1} \left( \frac{m^2 - m_A^2}{\mu - \alpha} \right) x, \quad x \in \mathbb{Z}_+.$$

The MaxWeight policy is defined by (9) with  $h$  a quadratic,

$$(52) \quad \phi^{\text{MW}}(x) = \arg \min_{u \in \mathcal{U}_\diamond(x)} \langle Dx, Bu + \alpha \rangle, \quad x \in \mathbb{Z}_+^\ell,$$

where  $D > 0$  is a diagonal matrix. Proposition 2.5 implies that  $\phi^{\text{MW}}$  coincides with the  $h$ -myopic policy for the fluid model. This result is a special case of Proposition 2.8 that follows.

**PROPOSITION 2.5.** *Suppose that assumptions (i) and (ii) of Theorem 1.1 hold. Then, for each  $x \in \mathbb{Z}_+^\ell$ , the allocation  $\phi^{\text{MW}}(x)$  defined by the MaxWeight policy can be expressed as a solution to the linear program,*

$$(53) \quad \begin{aligned} \phi^{\text{MW}}(x) = \quad & \arg \max \quad x^\top D(I - R^\top)Mu \\ & \text{s.t.} \quad u \in \mathcal{U}. \end{aligned}$$

Proposition 2.5 easily leads to a proof of stability of the MaxWeight policy. Theorem 2.6 is essentially contained in earlier work [53, 13]. We present the short proof since the same ideas are used to prove Theorem 1.1 and generalizations that follow.

**THEOREM 2.6.** *Suppose that  $\rho_\bullet < 1$ , and that assumptions (i) and (ii) of Theorem 1.1 hold. Then, for any diagonal matrix  $D > 0$ , the network controlled under the MaxWeight policy has a solution to Poisson's inequality (43) with  $V = h$  the quadratic defined in (1), and  $c(x) \equiv \varepsilon_0|x|$  for some  $\varepsilon_0 > 0$ .*

*Proof.* Since (41) holds when  $\rho_\bullet < 1$ , there exists  $\varepsilon > 0$  such that the vector  $v$  with coefficients  $v_i = -\varepsilon$ ,  $i \geq 1$ , lies in  $\mathcal{V}$ . By definition there exists  $u \in \mathcal{U}$  such that  $Bu + \alpha = v$ , so that by Proposition 2.5,

$$\langle B\phi^{\text{MW}}(x) + \alpha, \nabla h(x) \rangle = \langle B\phi^{\text{MW}}(x) + \alpha, Dx \rangle \leq v^\top Dx = -\varepsilon \sum D_{ii}x_i \leq -\varepsilon_0|x|,$$

with  $\varepsilon_0 = \varepsilon(\min_i D_{ii})$ . We thus arrive at a version of the Poisson inequality,

$$\mathcal{D}_{\text{MW}}h(x) := \mathbf{E}_{\text{MW}}[h(Q(t+1)) - h(Q(t)) \mid Q(t) = x] \leq -\varepsilon_0|x| + \bar{\eta}_D,$$

with

$$\bar{\eta}_D := \frac{1}{2} \max_{x' \in \mathbb{Z}_+^\ell, u \in \mathcal{U}_\diamond(x)} \mathbf{E}[(Q(t+1) - Q(t))^\top D(Q(t+1) - Q(t)) \mid Q(t) = x', U(t) = u]. \quad \square$$

**2.2. Perturbed functions.** We now analyze the drift  $\mathcal{D}h$  represented in (38) to establish stability of the  $h$ -MaxWeight policy. We return to the general CRW model (2).

An application of the chain rule of differentiation shows the following.

**PROPOSITION 2.7.** *For any  $C^1$  function  $h_0$ , the function  $h$  defined in (17) satisfies the derivative conditions (14). We have the explicit representations as follows:*

(i) *The first derivative is given by*

$$(54) \quad \nabla h(x) = [I - M_\theta] \nabla h_0(\tilde{x}),$$

where

$$(55) \quad M_\theta = M_\theta(x) = \text{diag}(e^{-x_i/\theta}), \quad x \in \mathbb{R}_+^\ell.$$

(ii) *If  $h_0$  is  $C^2$ , then the Hessian of  $h$  is*

$$(56) \quad \nabla^2 h(x) = [I - M_\theta] \nabla^2 h_0(\tilde{x}) [I - M_\theta] + \theta^{-1} \text{diag}(M_\theta \nabla h_0(\tilde{x})).$$

Hence  $h$  is convex provided  $h_0$  is both convex and monotone.

A key step in the proof of Theorem 1.1 is to generalize Proposition 2.5.

PROPOSITION 2.8. *Suppose that assumptions (i) and (ii) of Theorem 1.1 hold, and that  $h$  is any  $C^1$  monotone function satisfying the derivative conditions (14). Then, for each  $x \in \mathbb{Z}_+^\ell$ , the allocation  $\phi^{\text{MW}}(x)$  defined by the  $h$ -MaxWeight policy can be expressed as a solution to the linear program,*

$$(57) \quad \begin{aligned} \phi^{\text{MW}}(x) = \quad & \arg \min \quad \langle Bu, \nabla h(x) \rangle \\ & \text{s.t.} \quad u \in \mathbf{U}. \end{aligned}$$

*Proof.* The proof requires that we demonstrate that the minimum in (9) can be relaxed to a minimum over all of  $\mathbf{U}$ .

Recall the definition of the generalized Klimov indices in (21), and the interpretation that  $-\Theta_j$  is the coefficient of  $u_j$  in the objective function of (9). Monotonicity of  $h$  implies that each partial derivative of  $h$  is nonnegative. This assumption, combined with (18), implies that

$$(58) \quad \Theta_j(x) \leq 0 \text{ whenever } x_{i_j} = 0.$$

It then follows that the optimizer  $u^*$  of the linear program (57) satisfies, without loss of generality,  $u_j^* = 0$  whenever  $x_{i_j} = 0$ . This shows that  $u^* \in \mathbf{U}(x)$  for  $x \in \mathbb{Z}_+^\ell$ , which proves (57).

To show that  $u^*$  can be chosen in  $\mathbf{U}_\diamond$  we argue that optimizers of linear programs can be chosen from among the extreme points in the constraint region. The extreme points for this linear program are contained in  $\mathbf{U}_\diamond$  because of the definition  $\mathbf{U} := \text{conv}(\mathbf{U}_\diamond)$ .  $\square$

Consider the special case in which  $h_0$  is linear.

**2.2.1. Perturbed linear function.** Suppose that  $h_0(x) = c^\top x$ , where the vector  $c \in \mathbb{R}_+^\ell$  has nonzero coefficients, so that the function  $h$  can be expressed as

$$(59) \quad h(x) = \sum_{i=1}^{\ell} c_i \tilde{x}_i, \quad x \in \mathbb{R}_+^\ell.$$

An application of Proposition 2.7 shows that the derivative condition (14) holds, and that the first and second derivatives are given by

$$(60) \quad \nabla h(x) = [I - M_\theta]c, \quad \nabla^2 h(x) = \theta^{-1} \text{diag}(M_\theta c).$$

Hence the function  $h$  is monotone and strictly convex.

We show in Proposition 2.9 that the  $h$ -MaxWeight policy is stabilizing provided  $\theta \geq 1$  is suitably large. Although the dynamic programming inequality (15) fails for the linear function  $h_0$ , it does hold for the function  $k_0 h_0^2$ , where  $k_0$  is a sufficiently large constant, and  $c(x) = c^\top x$ . Hence Proposition 2.9(ii) could be deduced from Theorem 1.1.

PROPOSITION 2.9. *Suppose that assumptions (i) and (ii) of Theorem 1.1 hold, along with the stabilizability condition (41). Then, there exists  $\theta_0 > 0$  such that the following hold under the  $h$ -MaxWeight policy with  $h_0$  linear, provided  $\theta \geq \theta_0$ :*

(i) *The controlled network satisfies Foster's criterion. The function  $V$  in (V2) can be taken as a constant multiple of  $h$ .*

(ii) Condition (V3) holds: There exists  $\varepsilon_2 > 0$ ,  $b_2 < \infty$ , and a finite set  $S$  satisfying

$$\mathcal{D}V \leq -f + b_2 \mathbb{1}_S$$

with  $V = 1 + \frac{1}{2}h^2$ ,  $f = 1 + \varepsilon_2 h$ .

(iii) Suppose that for some  $\varepsilon > 0$  the arrival process satisfies  $\mathbb{E}[\exp(\varepsilon \|A(t)\|)] < \infty$ . Then condition (V4) holds: For some  $\varepsilon_e > 0$ ,  $\delta_e > 0$ ,  $b_e < \infty$ , and a finite set  $S$ ,

$$\mathcal{D}V \leq -\delta_e V + b_e \mathbb{1}_S$$

with  $V = \exp(\varepsilon_e h)$ . Hence  $\mathbf{Q}$  is geometrically ergodic provided (45) holds [41].

*Proof.* We apply the second-order mean value theorem to obtain

$$(61) \quad \begin{aligned} h(Q(t+1)) - h(Q(t)) &= \langle \nabla h(Q(t)), \Delta(t+1) \rangle \\ &\quad + \frac{1}{2} \Delta(t+1)^\top [\nabla^2 h(\bar{Q})] \Delta(t+1), \end{aligned}$$

where again  $\bar{Q} \in \mathbb{R}_+^\ell$  lies on the line connecting  $Q(t+1)$  and  $Q(t)$ . This implies the identity (38) with  $b_h$  redefined as

$$b_h(x) = \frac{1}{2} \mathbb{E}[\Delta(t+1)^\top \nabla^2 h(\bar{Q}) \Delta(t+1) \mid Q(t) = x].$$

The expression for the second derivative in (60) then gives

$$\mathcal{D}h(x) = \langle \nabla h(x), v \rangle + \theta^{-1} b_\Delta,$$

where

$$b_\Delta = \frac{1}{2} \|c\| \sup_{x' \in \mathbb{Z}_+^\ell, u \in \mathcal{U}_\phi(x')} \mathbb{E}[\|\Delta(t+1)\|^2 \mid Q(t) = x', U(t) = u] < \infty.$$

We now obtain an upper bound on  $\langle \nabla h(x), v \rangle$  under the  $h$ -MaxWeight policy. The expression for the first derivative in (60) implies the bound

$$\frac{\partial}{\partial x_i} h(x) = c_i (1 - e^{-x_i/\theta}) \geq \underline{c} (1 - e^{-x_i/\theta}), \quad 1 \leq i \leq \ell,$$

with  $\underline{c} := \min_j c_j$ . Exactly as in the proof of Theorem 2.6, we can consider arbitrary  $v \in \mathbf{V}$  to obtain bounds on the value of (57). This is justified by Proposition 2.8. The stabilizability condition (41) implies that there exists  $\varepsilon > 0$  such that the vector with components  $v_i = -\varepsilon$ ,  $1 \leq i \leq \ell$ , lies in  $\mathbf{V}$  for each  $x \in \mathbb{R}_+^\ell$ . By definition, there exists  $u \in \mathcal{U}$  satisfying  $Bu + \alpha = v$ . Consequently, under the  $h$ -MaxWeight policy,

$$\mathcal{D}h(x) \leq -\varepsilon \underline{c} \max_i (1 - e^{-x_i/\theta}) + \theta^{-1} b_\Delta.$$

Suppose that  $|x| \geq \ell\theta$ . Then  $x_i \geq \theta$  for at least one  $i$ , and we obtain the bound

$$(62) \quad \mathcal{D}h(x) \leq -\frac{1}{2} \varepsilon \underline{c} + \theta^{-1} b_\Delta \quad \text{if } |x| \geq \ell\theta.$$

The right-hand side is negative provided  $\theta > 2b_\Delta/(\varepsilon \underline{c})$ . Fixing  $\theta$  satisfying this bound, we obtain the desired solution to (V2) with  $V = 2(\varepsilon \underline{c})^{-1} h$ , and  $S = \{x : |x| < \ell\theta\}$ . This establishes (i).

To establish (ii) we begin with the identity

$$\frac{1}{2}[h(Q(t+1))]^2 - \frac{1}{2}[h(Q(t))]^2 = h(Q(t))(h(Q(t+1)) - h(Q(t))) + \frac{1}{2}[h(Q(t+1)) - h(Q(t))]^2.$$

On taking conditional expectations of both sides, we obtain  $\mathcal{D}V(x) = h(x)[\mathcal{D}h(x)] + b_{\Delta 2}(x)$ , where

$$b_{\Delta 2} = \frac{1}{2} \sup_{x' \in \mathbb{Z}_+^{\ell}, u \in \mathcal{U}_{\diamond}(x')} \mathbb{E}[[h(Q(t+1)) - h(Q(t))]^2 \mid Q(t) = x', U(t) = u] < \infty.$$

Applying (i), we obtain a version of the Poisson inequality (43) with this  $V$ , which implies that (V3) also holds.

Part (iii) follows from (i) combined with [41, Theorem 16.3.1] (see also [35, Theorem 4]).  $\square$

In the next example we find that the  $h$ -MaxWeight policy considered in Proposition 2.9(i) mirrors the discounted-cost optimal policy.

*Tandem queues: Emergence of a threshold policy.* Suppose that we replace the linear cost function used in (13) with the convex cost function  $h: \mathbb{R}_+^2 \rightarrow \mathbb{R}_+$  defined in (59). Figure 3 shows a plot of the sublevel sets of this function.

The  $h$ -MaxWeight policy defined in (9) minimizes the inner product,

$$\langle Bu + \alpha, \nabla h(x) \rangle = -\mu_1 u_1 c_1 (1 - e^{-x_1/\theta}) + (\mu_1 u_1 c_1 - \mu_2 u_2 c_2)(1 - e^{-x_2/\theta}) + \langle \alpha, \nabla h(x) \rangle.$$

The  $h$ -MaxWeight policy is thus nonidling at Station 2, and at Station 1 the policy can be expressed as a switching curve,

$$\phi_1^{\text{MW}}(x) = \mathbb{1}\{-c_1(1 - e^{-x_1/\theta}) + c_2(1 - e^{-x_2/\theta}) \leq 0\}, \quad x_1 \geq 1.$$

For small values of  $x_1$ , a first-order Taylor series gives the approximation  $\phi_1^{\text{MW}}(x) \approx \mathbb{1}\{x_2 \leq (c_1/c_2)x_1\}$ . For large  $x_1$  there are three cases to consider, depending on the relative sizes of  $c_1$  and  $c_2$ . If  $c_1 = c_2$ , then the approximation is equality,  $\phi_1^{\text{MW}}(x) = \mathbb{1}\{x_2 \leq x_1\}$  for all  $x$ . If  $c_1 > c_2$ , then Station 1 does not idle for large  $x_1$ , exactly as in the  $c$ -myopic policy. The  $h$ -MaxWeight policy is most interesting when  $c_2 > c_1$ . In this case, for  $x_1 \gg \theta$  the policy can be approximated by a threshold policy,  $\phi_1^{\text{MW}}(x) \approx \mathbb{1}\{x_2 \leq \bar{q}_2\}$ , where the threshold  $\bar{q}_2$  is the solution to the equation  $c_2(1 - e^{-\bar{q}_2/\theta}) = c_1$ . That is,

$$(63) \quad \bar{q}_2 = \theta \left| \log \left( 1 - \frac{c_1}{c_2} \right) \right|.$$

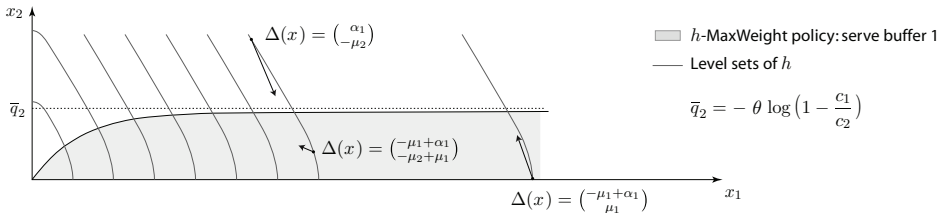


FIG. 3. The perturbed cost function defined in (17) with  $h_0$  linear on  $\mathbb{R}^2$  satisfies  $\min_u \langle \nabla h(x), Bu + \alpha \rangle < 0$  for each nonzero  $x \in \mathbb{R}_+^2$ . This geometry is illustrated in this figure using the tandem queues. The contour plots shown are the level sets  $\{x : h(x) = r\}$  for  $r = 1, 2, \dots$ .

Figure 3 illustrates  $\phi^{\text{MW}}$  when  $c_1 = 1$ ,  $c_2 = 3$ , and  $\theta = 10$ . The asymptote (63) is  $\bar{q}_2 = 10 \log(3/2) \approx 4$  in this special case.

For comparison consider the discounted-cost optimal policy, minimizing

$$\sum_{t=0}^{\infty} (1 + \gamma)^{-t-1} \mathbb{E}[c(Q(t)) \mid Q(t) = x],$$

for a given  $\gamma > 0$ . Letting  $h_\gamma^*(x)$  denote the minimizing value, the optimal policy is expressed as the  $h_\gamma^*$ -myopic policy, and the discounted-cost dynamic programming equation holds:

$$\gamma h_\gamma^*(x) = c(x) + \min_{u \in U_\circ(x)} \mathbb{E}[h_\gamma^*(Q(t+1)) - h_\gamma^*(Q(t)) \mid Q(t) = x, U(t) = u].$$

Consider the CRW model described by the recursion

$$Q(t+1) = Q(t) + (-\mathbf{1}^1 + \mathbf{1}^2)U_1(t)M_1(t) - \mathbf{1}^2U_2(t)M_2(t) + \mathbf{1}^1A_1(t+1), \quad t \geq 0,$$

in which the statistics of  $\Phi(t) := (M_1(t), M_2(t), A_1(t))^T$ ,  $t \geq 1$ , are consistent with a model obtained through uniformization:  $\Phi$  is i.i.d., with marginal distribution defined by

$$(64) \quad \mathbb{P}\{\Phi(t) = \mathbf{1}^1\} = \mu_1, \quad \mathbb{P}\{\Phi(t) = \mathbf{1}^2\} = \mu_2, \quad \mathbb{P}\{\Phi(t) = \mathbf{1}^3\} = \alpha_1,$$

with  $\mu_1 + \mu_2 + \alpha_1 = 1$ . We take  $c_1 = 1$ ,  $c_2 = 3$ ,  $\rho_1 = 9/11$ , and  $\rho_2 = 9/10$ .

For any finite  $\gamma$ , the following approximation holds:

$$\lim_{r \rightarrow \infty} r^{-1} h_\gamma^*([rx]) = \gamma^{-1} c(x), \quad x \in \mathbb{R}_+^2,$$

where  $[\cdot]$  denotes the componentwise integer part of a vector. Hence for large  $x$  far from the boundary, the optimal policy coincides with the  $c$ -myopic policy. In fact, it can be shown that the optimal policy is approximated by a static threshold, as seen in the two examples shown in Figure 4 (see [39, Example 10.6.1]). Hence the optimal policy is similar in form to the  $h$ -myopic policy illustrated in Figure 3.

**2.2.2. Perturbed value function.** We now consider a function  $h_0$  that serves as a Lyapunov function for the fluid model. Our goal is to complete the proof of Theorem 1.1, which amounts to establishing (V3) in the form of the Poisson inequality (19) with  $V = 2h$ .

Before proving the theorem, we present an example to illustrate the structure of the  $h$ -MaxWeight policy with  $h_0 = J^*$ , the optimal fluid value function defined below (42). In simple examples it is approximated by a switching curve with logarithmic growth.

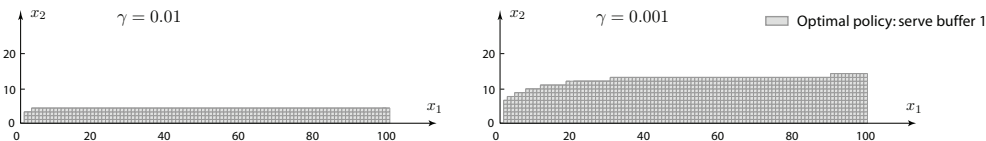


FIG. 4. Discounted-cost optimal policy for the tandem queues with cost parameters  $(c_1, c_2) = (1, 3)$ . The load parameters are  $\rho_1 = 9/10$  and  $\rho_2 = 9/11$ , and the linear cost is defined by  $c_1 = 1$ ,  $c_2 = 3$ . On the left  $\gamma = 0.01$  and on the right  $\gamma = 0.001$ .



*Tandem queues: Translation of the optimal policy.* If  $\rho_1 < \rho_2$  and  $c_1 < c_2$ , then the fluid value function is purely quadratic:

$$(65) \quad J^*(x) = \frac{1}{2} \frac{c_1}{\mu_2 - \alpha_1} (x_1 + x_2)^2 + \frac{1}{2} \frac{c_2 - c_1}{\mu_2} x_2^2, \quad x \in \mathbb{R}_+^2.$$

Letting  $h_0 = J^*$ , the dynamic programming inequality (15) is satisfied with equality:

$$\min_{u \in \mathcal{U}(x)} \langle \nabla h_0(x), Bu + \alpha \rangle = -c(x) \quad x \in \mathbb{R}_+^\ell.$$

The derivative conditions (14) fail, so we do not know if the  $h_0$ -MaxWeight policy is stabilizing for the CRW model.

To compute the  $h$ -MaxWeight policy, we write (65) as

$$h_0(x) = J^*(x) = \frac{1}{2} d_1 (x_1 + x_2)^2 + \frac{1}{2} d_2 x_2^2, \quad x \in \mathbb{R}_+^2,$$

so that the gradient of  $h(x) = h_0(\tilde{x})$  can be expressed as

$$\nabla h(x) = [I - M_\theta] \nabla h_0(\tilde{x}) = \begin{pmatrix} d_1(\tilde{x}_1 + \tilde{x}_2)(1 - e^{-x_1/\theta}) \\ (d_1(\tilde{x}_1 + \tilde{x}_2) + d_2\tilde{x}_2)(1 - e^{-x_2/\theta}) \end{pmatrix}.$$

Writing  $Bu + \alpha = (-\mu_1 u_1 + \alpha_1, \mu_1 u_1 + \mu_2 u_2)^\top$ , we obtain for any  $x \in \mathbb{Z}_+^\ell$ ,  $u \in \mathcal{U}(x)$ ,

$$\begin{aligned} \langle \nabla h(x), Bu + \alpha \rangle &= \mu_1 u_1 [d_1(e^{-x_1/\theta} - e^{-x_2/\theta})(\tilde{x}_1 + \tilde{x}_2) + d_2(1 - e^{-x_2/\theta})\tilde{x}_2] \\ &\quad - \mu_2 u_2 [d_1(1 - e^{-x_2/\theta})(\tilde{x}_1 + \tilde{x}_2) + d_2(1 - e^{-x_2/\theta})\tilde{x}_2] \\ &\quad + \alpha_1 d_1(1 - e^{-x_1/\theta})(\tilde{x}_1 + \tilde{x}_2). \end{aligned}$$

Minimizing over  $u$  we see that the policy is nonidling at Station 2. At Station 1 we have  $u_1 = 1$  if and only if  $x_1 \geq 1$  and the coefficient of  $u_1$  is nonpositive. That is, the policy at Station 1 is defined by the switching curve described by the equation

$$(66) \quad d_1(e^{-x_1/\theta} - e^{-x_2/\theta})(\tilde{x}_1 + \tilde{x}_2) + d_2(1 - e^{-x_2/\theta})\tilde{x}_2 = 0.$$

When  $x_1$  is large, we obtain the approximation

$$(67) \quad x_2 \approx s(x_1) := \theta \log \left( 1 + \frac{d_1}{d_2} x_1 \right),$$

where, by (65),

$$\frac{d_1}{d_2} = \left( \frac{c_2}{c_1} - 1 \right)^{-1} \frac{1}{1 - \rho_2}.$$

This is an approximation to (66) in the sense that for all sufficiently large  $x_1$  there is a unique  $x_2$  such that  $(x_1, x_2)$  solve (66), and the ratio  $x_2/s(x_1)$  tends to unity as  $x_1 \rightarrow \infty$ .

A policy defined by a switching curve  $s(x_1)$  of the form given in (67) is similar to the policy introduced in [37] to obtain HTAO (see (32) and the surrounding discussion).

Consider now the average-cost optimal policy for the CRW model, with statistics defined in (64) and linear cost with  $(c_1, c_2) = (1, 3)$ . It is known that the average-cost optimal policy exists, and that it is  $h^*$ -myopic with respect to the relative value

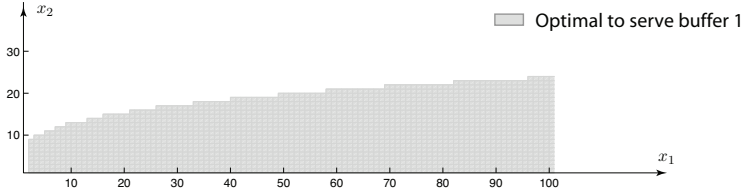


FIG. 5. Average-cost optimal policy for the tandem queues with  $c_1 = 1$ ,  $c_2 = 3$ ,  $\rho_1 = 9/11$ , and  $\rho_2 = 9/10$ .

function (see [5] and [39, Chapter 9]). Moreover, Theorem 7.2 of [34] implies that the following approximation holds:

$$\lim_{r \rightarrow \infty} r^{-2} h^*([rx]) = J^*(x), \quad x \in \mathbb{R}_+^2.$$

The average-cost optimal policy for the CRW model is shown in Figure 5 with  $\rho_1 = 9/11 < \rho_2$ . This policy can be represented by a switching curve  $s$  that is concave and unbounded in  $x_1$ , similar to (67).

The proof of Theorem 1.1 is organized in the following two lemmas.

LEMMA 2.10. *Under the assumptions of Theorem 1.1 we have, under the  $h$ -MaxWeight policy, for some constant  $k_{2.10}$ ,*

$$\langle \nabla h(x), v^{\text{MW}} \rangle \leq -c(x) + k_{2.10} \log(1 + \|x\|), \quad x \in \mathbb{Z}_+^\ell,$$

where  $v^{\text{MW}} = B\phi^{\text{MW}}(x) + \alpha$ .

*Proof.* Fix a constant  $\beta_- \geq \theta$ , and define

$$s_-(r) = \beta_- \log(1 + r/\beta_-), \quad r \geq 0.$$

To prove the lemma we compare  $v^{\text{MW}}$  with another velocity vector  $v \in \mathbb{V}$ , subject to the following constraints:

$$(68) \quad v_i \geq 0 \quad \text{whenever} \quad x_i < s_-(\|x\|), \quad i = 1, \dots, \ell.$$

The minimum of  $\langle \nabla h(x), v \rangle$  over  $v$  satisfying these constraints provides a bound under the  $h$ -MaxWeight policy. Proposition 2.8 is critical here so that we can ignore lattice constraints as we search for bounds on this inner product.

The purpose of (68) is to obtain the following bound:

$$(69) \quad -e^{-x_i/\theta} v_i \leq |v_i| (1 + \|x\|/\beta_-)^{-\beta_-/\theta}, \quad i = 1, \dots, \ell.$$

Since  $h_0$  is assumed monotone we have  $\nabla h_0: \mathbb{R}_+^\ell \rightarrow \mathbb{R}_+^\ell$ , and applying (54) we obtain

$$\langle \nabla h(x), v \rangle \leq \langle \nabla h_0(\tilde{x}), v \rangle + \|v\| \|\nabla h_0(\tilde{x})\| (1 + \|x\|/\beta_-)^{-\beta_-/\theta}.$$

Since  $\nabla h_0$  is also Lipschitz and  $\beta_- \geq \theta$ , this gives, for some constant  $k_0$ ,

$$(70) \quad \langle \nabla h(x), v \rangle \leq \langle \nabla h_0(\tilde{x}), v \rangle + k_0.$$

To bound (70) we shift  $\tilde{x}$  as follows: Let  $\tilde{x}^- \in \mathbb{Z}_+^\ell$  denote the vector with components

$$\tilde{x}_i^- = \lfloor (\tilde{x}_i - s_-(\|x\|))_+ \rfloor, \quad i = 1, \dots, \ell,$$

where  $\lfloor \cdot \rfloor$  denotes the integer part. In view of (15), there exists  $u \in \mathcal{U}(x)$  such that with  $v = Bu + \alpha$ ,

$$\langle \nabla h_0(\tilde{x}^-), v \rangle \leq -c(\tilde{x}^-).$$

Moreover, we have  $\tilde{x}_i^- = 0$  whenever the constraint on  $x_i$  in (68) is active. Since  $u \in \mathcal{U}(x)$ , this implies that the vector  $v = Bu + \alpha$  satisfies  $v_i \geq 0$ . That is,  $v$  satisfies the constraint (68).

Using this  $v$  in (70) gives

$$\begin{aligned} \langle \nabla h(x), v \rangle &\leq \langle \nabla h_0(\tilde{x}^-), v \rangle + \langle \nabla h_0(\tilde{x}) - \nabla h_0(\tilde{x}^-), v \rangle + k_0 \\ &\leq -c(\tilde{x}^-) + \|v\| \|\nabla h_0(\tilde{x}) - \nabla h_0(\tilde{x}^-)\| + k_0 \\ &\leq -c(x) + |c(x) - c(\tilde{x}^-)| + \|v\| \|\nabla h_0(\tilde{x}) - \nabla h_0(\tilde{x}^-)\| + k_0. \end{aligned}$$

This completes the proof since  $c$  and  $\nabla h_0$  are Lipschitz.  $\square$

LEMMA 2.11. *Under the assumptions of Theorem 1.1 we have, under the h-MaxWeight policy, for some constant  $k_{2.11}$ ,*

$$\mathcal{D}h(x) \leq \langle \nabla h(x), v^{\text{MW}} \rangle + k_{2.11}(1 + \theta^{-1}\|x\|), \quad x \in \mathbb{Z}_+^\ell,$$

where  $v^{\text{MW}} = B\phi^{\text{MW}}(x) + \alpha$ .

*Proof.* The first-order mean value theorem (37) results in the representation (38) for  $\mathcal{D}h$  with  $b_h$  defined in (39). The Cauchy–Schwarz inequality gives,

$$(71) \quad b_h(x) \leq \mathbb{E}[\|\nabla h(\bar{Q}) - \nabla h(Q(t))\|^2 \mid Q(t) = x]^{\frac{1}{2}} \mathbb{E}[\|\Delta(t+1)\|^2 \mid Q(t) = x]^{\frac{1}{2}}.$$

It remains to bound the right-hand side.

Given  $Q(t) = x$ , an application of Proposition 2.7 gives

$$\nabla h(\bar{Q}) - \nabla h(x) = [I - M_\theta(\bar{Q})](\nabla h_0(\tilde{\bar{Q}}) - \nabla h_0(\tilde{x})) + [M_\theta(x) - M_\theta(\bar{Q})]\nabla h_0(\tilde{x}),$$

where  $\tilde{\bar{Q}}$  is obtained from  $\bar{Q}$  via the pointwise transformation (16). The first expectation on the right-hand side of (71) is bounded through an application of the triangle inequality,

$$\begin{aligned} &\mathbb{E}[\|\nabla h(\bar{Q}) - \nabla h(Q(t))\|^2 \mid Q(t) = x]^{\frac{1}{2}} \\ &\leq \mathbb{E}[\|[I - M_\theta(\bar{Q})](\nabla h_0(\tilde{\bar{Q}}) - \nabla h_0(\tilde{x}))\|^2 \mid Q(t) = x]^{\frac{1}{2}} \\ (72) \quad &+ \mathbb{E}[\|[M_\theta(x) - M_\theta(\bar{Q})]\nabla h_0(\tilde{x})\|^2 \mid Q(t) = x]^{\frac{1}{2}}. \end{aligned}$$

To bound the first term on the right-hand side of (72) we apply the Lipschitz condition on  $h_0$ : For some constant  $k_0$ ,

$$\|\nabla h_0(\tilde{\bar{Q}}) - \nabla h_0(\tilde{x})\| \leq k_0 \|\tilde{\bar{Q}} - \tilde{x}\| \leq \|\Delta(t+1)\|.$$

Hence the first term is bounded over  $x$ .

The second term is bounded using the mean value theorem. The  $i$ th diagonal element of  $[M_\theta(x) - M_\theta(\bar{Q})]$  admits the bound

$$\begin{aligned} |e^{-x_i/\theta} - e^{-\bar{Q}_i/\theta}| &= e^{-x_i/\theta} |1 - e^{-(\bar{Q}_i - x_i)/\theta}| \\ &\leq e^{-x_i/\theta} (1 - e^{-\bar{\Delta}_i/\theta}) \mathbb{1}\{\bar{Q}_i > x_i\} \\ &\quad + e^{-x_i/\theta} (e^{\ell_u/\theta} - 1) \mathbb{1}\{\bar{Q}_i < x_i\}, \end{aligned}$$

where  $\bar{\Delta}_i = A_i(1) + \sum_j |B_{ij}(1)|$ , and we have used the fact that  $\sum_j B_{ij}(1) \geq -\ell_u$  under (18). The right-hand side can be bounded through a second application of the mean value theorem, giving

$$|e^{-x_i/\theta} - e^{-\bar{Q}_i/\theta}| \leq e^{-x_i/\theta} (e^{\ell_u/\theta} - e^{-\bar{\Delta}_i/\theta}) \leq \theta^{-1} e^{\ell_u/\theta} (\ell_u + \bar{\Delta}_i).$$

The Lipschitz condition on  $\nabla h_0$  and second moment conditions on  $(\mathbf{A}, \mathbf{B})$  then imply that for some  $k_3 < \infty$ ,

$$\begin{aligned} \mathbb{E}[\| [M_\theta(x) - M_\theta(\bar{Q})] \nabla h_0(\tilde{x}) \|^2 \mid Q(t) = x]^\frac{1}{2} &\leq \theta^{-1} e^{\ell_u/\theta} (\sqrt{\ell} \ell_u + \mathbb{E}[\|\bar{\Delta}\|^2]^\frac{1}{2}) \|\nabla h_0(\tilde{x})\| \\ &\leq k_3 \theta^{-1} (1 + \|x\|). \end{aligned}$$

This combined with (19), (71), and (72) completes the proof.  $\square$

**2.3. Universally stabilizing policies.** The policies described in the previous sections are stabilizing, provided the parameter  $\theta > 0$  is chosen sufficiently large. With a different change of variables we can construct a family of policies that are stabilizing regardless of the parameter.

In this subsection only we redefine  $\tilde{x}$  as

$$(73) \quad \tilde{x}_i := x_i \log(1 + x_i/\theta),$$

where  $\theta > 0$  is fixed but arbitrary. Theorem 1.1 and Proposition 2.9 can be generalized using this new definition of  $\tilde{x}$ . First we require derivative formulae.

**PROPOSITION 2.12.** *For any  $C^1$  function  $h_0$ , the function  $h$  defined in (17) with  $\tilde{x}$  defined componentwise in (73) satisfies the derivative conditions (14) as follows:*

(i) *The first derivative is given by*

$$(74) \quad \nabla h(x) = L_\theta \nabla h_0(\tilde{x}),$$

where

$$(75) \quad L_\theta(x) = \text{diag}(x_i/(\theta + x_i) + \log(1 + x_i/\theta)), \quad x \in \mathbb{R}_+^\ell.$$

(ii) *If  $h_0$  is  $C^2$ , then the Hessian of  $h$  is*

$$(76) \quad \nabla^2 h(x) = L_\theta \nabla^2 h_0(\tilde{x}) L_\theta + \theta^{-1} \text{diag}(N_\theta \nabla h_0(\tilde{x})),$$

where  $N_\theta(x) = \text{diag}(\theta/(\theta + x_i) + \theta^2/(\theta + x_i)^2)$ . In particular,  $h$  is convex provided  $h_0$  is both convex and monotone.

Proposition 2.13 establishes stability of the  $h$ -MaxWeight policy when  $h_0$  is linear. We omit the proof since it is similar to the proof of Proposition 2.9(i), applying Proposition 2.12 instead of Proposition 2.7.

**PROPOSITION 2.13.** *Suppose that assumptions (i) and (ii) of Theorem 1.1 hold, along with the stabilizability condition (41). Then, under the  $h$ -MaxWeight policy with  $h_0$  linear, and  $\tilde{x}$  defined in (73) with  $\theta > 0$ , there exists  $\varepsilon = \varepsilon(\theta) > 0$  such that condition (V3) holds with  $f(x) = 1 + \varepsilon \log(1 + \|x\|)$ .*

We now consider a version of Theorem 1.1. It is necessary to strengthen the  $L_2$  condition on the arrival process to a second moment bound on  $\tilde{\mathbf{A}}$ ,

$$(77) \quad \mathbb{E}[\|\tilde{\mathbf{A}}(t)\|^2] := \sum_{i=1}^{\ell} \mathbb{E}[(A_i(t) \log(1 + A_i(t)/\theta))^2] < \infty.$$

Moreover, it appears that the monotonicity assumption must be strengthened to

$$(78) \quad \frac{\partial}{\partial x_i} h_0(x) \geq \varepsilon_{78} x_i, \quad x \in \mathbb{R}_+^\ell, \quad i = 1, \dots, \ell,$$

where  $\varepsilon_{78} > 0$  is constant. For example, (78) holds for a quadratic (1) in which  $D_{ij} \geq 0$  and  $D_{ii} > 0$  for each  $i, j$ .

**THEOREM 2.14.** *Consider the  $h$ -MaxWeight policy in which the function  $h$  is based on the change of variables  $\tilde{x}$  defined in (73). The constant  $\theta > 0$  is fixed but arbitrary. The network model (2) and function  $h_0$  satisfy all of the assumptions of Theorem 1.1. In addition, the arrival process satisfies (77), and  $h_0$  satisfies the uniform bound (78) with  $\varepsilon_{78} > 0$ . Then, there exists  $\delta_h > 0$  and  $\bar{\eta}_h < \infty$  such that the following version of the Poisson inequality (43) holds under the  $h$ -MaxWeight policy:*

$$(79) \quad \mathcal{D}h(x) = \mathbb{E}[h(Q(t+1)) - h(Q(t)) \mid Q(t) = x] \leq -\delta_h \|x\| (\log(1 + \|x\|))^2 + \bar{\eta}_h.$$

*Proof.* We first obtain an extension of Lemma 2.10: There exists  $\delta_h^0 > 0$  such that

$$(80) \quad \langle \nabla h(x), v^{\text{MW}} \rangle \leq -\delta_h^0 \|x\| (\log(1 + \|x\|))^2, \quad x \in \mathbb{Z}_+^\ell.$$

As in the proof of Proposition 2.6, we have the bound  $\langle \nabla h(x), v^{\text{MW}} \rangle \leq \langle \nabla h(x), v \rangle$  with  $v_i = -\varepsilon$ ,  $i \geq 1$ , and  $\varepsilon > 0$  chosen so that  $v \in \mathbb{V}$ . Hence,

$$\langle \nabla h(x), v^{\text{MW}} \rangle \leq -\varepsilon \sum_{i=1}^{\ell} \frac{\partial}{\partial x_i} h(x) = -\varepsilon \sum_{i=1}^{\ell} (x_i/(\theta + x_i) + \log(1 + x_i/\theta)) \frac{\partial}{\partial x_i} h_0(\tilde{x}).$$

This combined with the bound (78) gives

$$\langle \nabla h(x), v^{\text{MW}} \rangle \leq -\varepsilon \varepsilon_{78} \sum_{i=1}^{\ell} \log(1 + x_i/\theta) \tilde{x}_i.$$

From the definition of  $\tilde{x}$  we obtain (80) for some  $\delta_h^0 > 0$ .

Based on (80) we complete the proof using an extension of Lemma 2.11. Combining (71) and (80) gives

$$(81) \quad \begin{aligned} \mathcal{D}h(x) &\leq -\delta_h^0 \|x\| (\log(1 + \|x\|))^2 \\ &\quad + \mathbb{E}[\|\nabla h(\bar{Q}) - \nabla h(Q(t))\|^2 \mid Q(t) = x]^{\frac{1}{2}} \mathbb{E}[\|\Delta(t+1)\|^2 \mid Q(t) = x]^{\frac{1}{2}}. \end{aligned}$$

Following the arguments used in the proof of Lemma 2.11, and using Proposition 2.12 in place of Proposition 2.7, we obtain the bound, for some constant  $k_0$ ,

$$\mathbb{E}[\|\nabla h(\bar{Q}) - \nabla h(Q(t))\|^2 \mid Q(t) = x]^{\frac{1}{2}} \leq k_0 \|x\| (\log(1 + \|x\|)).$$

This together with (81) completes the proof of (79), where the constant  $\delta_h$  can be chosen arbitrarily in the open interval  $(0, \delta_h^0)$ .  $\square$

**3. Relaxations and heavy traffic.** We now establish HTAO under the  $h$ -MaxWeight policy for a specifically chosen function  $h$ , and under further restrictions on the network model.

Throughout this section we consider the homogeneous scheduling model (33) subject to the following conventions. The load parameters defined in (34) are expressed as

$$\rho_i = \lambda_i / \mu_i, \quad 1 \leq i \leq \ell_m,$$

where  $\mu_i$  is the common mean of  $\{M_j(t) : s(j) = i\}$ , and  $\lambda_i$  is the  $i$ th component of  $C[I - R^T]^{-1}\alpha$ . It is assumed throughout this section that  $\rho_1 = \max_{1 \leq i \leq \ell} \rho_i$ , so that  $\rho_\bullet = \rho_1$ . This can be achieved by choice of indices. We let  $\xi \in \mathbb{Z}_+^\ell$  denote the first column of the  $\ell \times \ell_m$  matrix  $[I - R]^{-1}C^T$ , so that  $\xi^T \alpha = \lambda_1$ .

Homogeneity implies that the random variables  $\{M_j(t) : s(j) = 1\}$  are all identical; we let  $S_1(t)$  denote their common value, and we let  $L_1(t) = \xi^T A(t)$ . The one-dimensional workload process  $W(t) = \langle \xi, Q(t) \rangle$  evolves as

$$(82) \quad W(t+1) = W(t) - S_1(t+1) + S_1(t+1)\mathcal{L}(t) + L_1(t+1), \quad t \geq 0,$$

where  $\mathcal{L}(t) := 1 - [CU(t)]_1$ .

The one-dimensional relaxation is defined on the same probability space with  $Q$  and evolves as a controlled random walk analogous to (82):

$$(83) \quad \widehat{W}(t+1) = \widehat{W}(t) - S_1(t+1) + S_1(t+1)\widehat{\mathcal{L}}(t) + L_1(t+1), \quad t \geq 0.$$

The idleness process  $\{\widehat{\mathcal{L}}(t)\}$  is assumed nonnegative and adapted to  $\{\widehat{W}(t), S_1(t), L_1(t)\}$ . The relaxation is denoted  $\widehat{W}^*$  when controlled using the nonidling policy,  $\widehat{\mathcal{L}}^*(t) = \mathbb{1}\{\widehat{W}^*(t) \geq 1\}$ . In this case we have (35), provided each process has the common initialization  $\widehat{W}^*(0) = W(0) = \langle \xi, Q(0) \rangle$ , which we assume henceforth.

For a convex cost function  $c: \mathbb{R}_+^\ell \rightarrow \mathbb{R}_+$  the effective cost  $\bar{c}: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is defined in (24). For each  $w \in \mathbb{R}_+$  an *effective state*  $\mathcal{X}^*(w)$  is defined to be any vector  $x^* \in \mathbb{R}_+^\ell$  that achieves the minimum in (24):

$$(84) \quad \mathcal{X}^*(w) = \arg \min_{x \in \mathbb{R}_+^\ell} \left( c(x) : \xi^T x = w \right).$$

It follows from the definitions that the following bound holds for all  $t$ :

$$(85) \quad c(Q(t)) \geq \bar{c}(W(t)) \geq \bar{c}(\widehat{W}^*(t)).$$

**3.1. Starvation and relaxations.** It is helpful to consider a workload relaxation to see how starvation arises under a myopic policy.

**3.1.1. Linear cost function.** If  $c(x) = c^T x$ ,  $x \in \mathbb{R}_+^\ell$ , then the effective state in a one-dimensional relaxation is given by

$$\mathcal{X}^*(w) = \left( \frac{1}{\xi_{i^*}} \mathbf{1}^{i^*} \right) w, \quad w \in \mathbb{R}_+,$$

where the index  $i^*$  is any solution to  $c_{i^*}/\xi_{i^*} = \min_{1 \leq i \leq \ell} (c_i/\xi_i)$ . The effective cost is given by the linear function  $\bar{c}(w) = c(\mathcal{X}^*(w)) = (c_{i^*}/\xi_{i^*})w$ ,  $w \in \mathbb{R}_+$ .

This underlines the conflict that arises frequently in network optimization: Optimization of an idealized model dictates zero inventory at various stations, while in a more realistic model, adopting a “zero-inventory policy” results in starvation of resources.

**3.1.2. Quadratic cost function.** If  $c: \mathbb{R}_+^\ell \rightarrow \mathbb{R}_+$  is quadratic, of the form  $c(x) = \frac{1}{2}x^\top D x$ ,  $x \in \mathbb{R}_+^\ell$  for a symmetric matrix  $D$ , then the effective state is again linear in the workload value in any one-dimensional relaxation.

For the scheduling model considered in this section the workload vector  $\xi$  has nonnegative entries. Suppose that  $D^{-1}$  also has nonnegative entries. In this case we have the explicit expression

$$\mathcal{X}^*(w) = \left( (\xi^\top D^{-1} \xi)^{-1} D^{-1} \xi \right) w, \quad w \in \mathbb{R}_+,$$

and the effective cost is the one-dimensional quadratic

$$\bar{c}(w) = \frac{1}{2} (\xi^\top D^{-1} \xi)^{-1} w^2, \quad w \in \mathbb{R}_+.$$

For example, if  $D > 0$  is diagonal, and  $\xi$  has strictly positive entries, then the effective state  $\mathcal{X}^*(w)$  has strictly positive entries for any  $w > 0$ . In conclusion, the conflict observed for linear cost functions does not arise when using a quadratic function satisfying these conditions.

**3.2. Logarithmic regret.** We now construct a policy satisfying (31) with  $c$  a linear cost function. The policy is defined as the  $h$ -MaxWeight policy for a specific function  $h$ .

We saw in section 3.1.1 that the effective state  $x^* = \mathcal{X}^*(w)$  can be constructed so that  $x_i^* = 0$  for all but one  $i \in \{1, \dots, \ell\}$ . By choice of indices we assume that  $x_i^* = 0$  for  $i \geq 2$  and any  $w \in \mathbb{R}_+$ . Consequently, the effective cost is given by

$$(86) \quad \bar{c}(w) = c(\mathcal{X}^*(w)) = \frac{c_1}{\xi_1} w, \quad w \in \mathbb{R}_+.$$

We assume, moreover, that the solution to (24) is unique, which amounts to the following strict bound:

$$(87) \quad \frac{c_i}{\xi_i} > \frac{c_1}{\xi_1}, \quad i = 2, \dots, \ell.$$

Theorem 2.4 implies the following formula for the steady-state cost for the relaxation:

$$(88) \quad \hat{\eta}^* = \frac{1}{2} \frac{\sigma_\kappa^2}{\mu_1 - \lambda_1} \frac{c_1}{\xi_1},$$

where  $\sigma_\kappa^2 := \rho_\bullet \mathbb{E}[(S_1(1) - L_1(1))^2] + (1 - \rho_\bullet) \mathbb{E}[(L_1(1))^2]$ . From (85) we evidently have  $\hat{\eta}^* \leq \eta^*$ . To establish (31) we require a bound in the reverse direction.

We now introduce a family of network models, parameterized by a scalar  $\kappa \in [1, \infty)$  that represents load. It is assumed that, for some fixed  $\kappa_0 \geq 1$ ,  $\bar{\rho} < 1$ , we have

$$(89) \quad \begin{aligned} \rho_\bullet &:= \rho_1 = 1 - 1/\kappa && \text{for each } \kappa \in [\kappa_0, \infty) \\ \text{and } \rho_i &\leq \bar{\rho} && \text{for each } \kappa \text{ and } i \geq 2. \end{aligned}$$

The one-dimensional workload process  $\mathbf{W}$  is given in (82), where we suppress the dependency of  $\{\mathbf{Q}, \mathbf{W}, \mathbf{S}, \mathbf{L}\}$  on the parameter  $\kappa$  to simplify notation.

The fluid model for the one-dimensional workload process is expressed as  $\frac{d^+}{dt} w(t) = -(\mu_1 - \lambda_1) + \iota(t)$  with  $\iota(t) \geq 0$ . Given the cost function  $\bar{c}$  given in (86), the fluid value function is given by

$$(90) \quad \hat{J}^*(w) = \frac{1}{2} \frac{c_1}{\xi_1} \frac{w^2}{\mu_1}, \quad w \geq 0.$$

The solution to Poisson's equation (51) is the sum of  $\widehat{J}^*$  and a linear function of  $w$ . We take the function  $h_0$  used to define the  $h$ -MaxWeight policy as a different perturbation of  $\widehat{J}^*$ : Fix a positive constant  $b > 0$ , and define

$$(91) \quad h_0(x) = \widehat{J}^*(\xi^\top x) + \frac{1}{2}b(c(x) - \bar{c}(\xi^\top x))^2.$$

This function is monotone since  $c(x) \geq \bar{c}(\xi^\top x)$  for each  $x$ . We then take  $h(x) = h_0(\tilde{x})$ , or

$$(92) \quad h(x) := \widehat{J}^*(\tilde{w}) + \frac{1}{2}b(c(\tilde{x}) - \bar{c}(\tilde{w}))^2, \quad x \in \mathbb{R}_+^\ell,$$

where  $\tilde{x}$  is the  $\ell$ -dimensional vector with components given in (16), and  $\tilde{w} := \sum \xi_i \tilde{x}_i$ .

For each  $\kappa$  we denote by  $\eta = \eta(\kappa)$  the steady-state cost under the policy for the CRW model, and  $\widehat{\eta}^*$  the optimal average cost for the one-dimensional relaxation. Applying (89), the representation (88) becomes

$$(93) \quad \widehat{\eta}^* = \frac{1}{2} \frac{\sigma_\kappa^2}{\mu_1} \frac{c_1}{\xi_1} \kappa.$$

Note that  $\eta$  and  $\widehat{\eta}^*$  are unbounded as the network load  $\rho_\bullet$  approaches unity, and  $\widehat{\eta}^*$  is of order  $\kappa$ . Hence (31) implies the bounds

$$(94) \quad \eta^* \leq \eta \leq \eta^* + k_{94} \log(\kappa), \quad \kappa \geq \kappa_0.$$

**THEOREM 3.1.** *Suppose that the following hold for the parameterized family of networks:*

(HTAO 1). *For each  $\kappa$ , the network is a CRW scheduling model with deterministic routing.*

(HTAO 2). *The network is homogeneous for each  $\kappa$ . Moreover,*

(a) *the random variables  $\{A^\kappa(t), S_s^\kappa(t) : t \geq 1, \kappa \in [1, \infty], s \geq 1\}$  are defined on a common probability space and are monotone in  $\kappa$ : For each  $s \in \{1, \dots, \ell_m\}$ ,*

$$S_s^\kappa(t) \downarrow S_s^\infty(t), \quad A^\kappa(t) \uparrow A^\infty(t), \quad \kappa \rightarrow \infty \text{ a.s. } t \geq 1.$$

(b) *the distribution of  $S_s^\kappa(t)$  is Bernoulli for each  $s$ , and  $\mathbb{E}[\|A^\infty(t)\|^2] < \infty$ .*

(c) *for some  $\varepsilon > 0$  independent of  $\kappa$  and  $s$ ,*

$$(95) \quad \mathbb{P}\{S_s^\kappa(t) = 1 \text{ and } A^\kappa(t) = 0\} \geq \varepsilon.$$

(HTAO 3). *Linear cost.*

(HTAO 4). *The load parameters satisfy (89). Hence  $\rho^\kappa \rightarrow \rho^\infty$  as  $\kappa \rightarrow \infty$ , where  $\rho_1^\infty = 1$  and  $\rho_i^\infty < 1$  for  $i \geq 2$ .*

*Moreover, the effective state is unique: Equation (87) holds.*

*Then, there exists  $\theta_0 > 0$  and  $b_0 > 0$  such that the following conclusions hold for the  $h$ -MaxWeight policy with  $h$  defined in (92), with  $\theta \geq \theta_0$  and  $b \geq b_0$ : There exists  $\kappa_0 > 0$  such that the controlled network is ergodic, in the sense that it is 0-irreducible and (V3) holds, for each  $\kappa \geq \kappa_0$ . Moreover, the family of controlled networks satisfies the bound (94) for some fixed  $k_{94} < \infty$ . That is, the policy has HTAO with logarithmic regret.*

The proof is based on the construction of a Lyapunov function  $V: \mathbb{Z}_+^\ell \rightarrow \mathbb{R}_+$  satisfying a refinement of (V3),

$$(96) \quad \mathbb{E}_x[V(Q(1))] \leq V(x) - c(x) + \widehat{\eta}^* + \mathcal{E}(x), \quad x \in \mathbb{Z}_+^\ell,$$

where the error  $\mathcal{E}$  has at most logarithmic growth,



$$(97) \quad \mathcal{E}(x) = k_{97} \left( \log(\kappa + c(x)) + \kappa / (1 + (\xi^T x)^2) \right), \quad x \in \mathbb{R}_+^\ell, \quad \kappa \geq 1.$$

This is achieved using the same steps used in section 2.2 to establish stability of the  $h$ -MaxWeight policy: First, we obtain a bound on the term  $\langle \nabla h(x), v \rangle$  appearing in (38). We then decompose the term  $b_h(x)$  into a bounded term, and a term whose mean is equal to  $\hat{\eta}^*$ .

**3.3. Proof of Theorem 3.1.** Throughout this section we assume that the assumptions of Theorem 3.1 hold. We let  $\mathcal{E}$  denote the function of  $x$  and  $\kappa$  given in (97). The constant  $k_{97}$  may differ in each appearance.

We first establish irreducibility.

**PROPOSITION 3.2.** *Under the assumptions of Theorem 3.1 the policy  $\phi^{\text{MW}}$  is 0-irreducible for each  $\kappa < \infty$ .*

*Proof.* Monotonicity of  $h$  implies that the  $h$ -MaxWeight policy is “weakly non-idling”: For any  $t$ ,

$$\sum_{i=1}^{\ell} U(t) \geq 1 \quad \text{whenever } Q(t) \neq \mathbf{0}.$$

Combining this with (95) we can conclude that for some  $\delta > 0$ , and any nonzero  $x \in \mathbf{X}_\diamond$ ,

$$\mathbb{P}\{\text{A service is completed at time } t \text{ and } A(t) = \mathbf{0} \mid Q(t-1) = x\} \geq \delta, \quad t \geq 1.$$

Since each customer in the network requires service at most  $\ell$  times, it follows that  $P^T(x, \mathbf{0}) \geq \delta^T$  for  $T = \ell|x|$ . This establishes 0-irreducibility.

Aperiodicity also follows from (95) since  $P(\mathbf{0}, \mathbf{0}) = \mathbb{P}\{A(t) = \mathbf{0}\} > 0$ .  $\square$

Recall that the proof of Theorem 1.1 was based on Lemmas 2.10 and 2.11. The following two propositions are refinements of these results.

**PROPOSITION 3.3.** *Under the  $h$ -MaxWeight policy, we have for some  $\varepsilon_{3.3} > 0$ , independent of  $b$  and  $x$ ,*

$$(98) \quad \langle \nabla h(x), v^{\text{MW}} \rangle \leq -\bar{c}(w) - \varepsilon \left( b|c(x) - \bar{c}(w)| + \|M_\theta(x)\xi\|\kappa w \right) + \mathcal{E}(x),$$

where  $w = \xi^T x$  and  $v^{\text{MW}} = B\phi^{\text{MW}}(x) + \alpha$ .

**PROPOSITION 3.4.** *Under the  $h$ -MaxWeight policy, we have for some  $k_{3.4} < \infty$ , independent of  $b$ ,  $\kappa$ , and  $x$ ,*

$$(99) \quad \mathcal{D}h(x) \leq \langle \nabla h(x), v^{\text{MW}} \rangle + \hat{\eta}^*(x) + k_{3.4} \theta^{-1} \left( b|c(x) - \bar{c}(w)| + \|M_\theta(x)\xi\|\kappa w \right) + \mathcal{E}(x),$$

where  $v^{\text{MW}} = B\phi^{\text{MW}}(x) + \alpha$  and

$$(100) \quad \hat{\eta}^*(x) := \frac{1}{2} \frac{\kappa}{\mu_1} \frac{c_1}{\xi_1} \sigma_\kappa^2(x), \quad \text{with} \quad \sigma_\kappa^2(x) := \mathbb{E}[(\xi^T \Delta(t+1))^2 \mid Q(t) = x].$$

We begin with the proof of Proposition 3.3. Note that

$$(101) \quad c(x) - \bar{c}(w) = \sum_{i=2}^{\ell} \left( c_i - (c_1/\xi_1)\xi_i \right) x_i,$$

which is nonnegative by (87). That is, we have  $|c(x) - \bar{c}(w)| = c(x) - \bar{c}(w)$ . To prove the proposition we first apply Proposition 2.7 to obtain a representation for the gradient of  $h$ ,

$$(102) \quad \nabla h(x) = \nabla \hat{J}^*(\tilde{w})[I - M_\theta]\xi + b(c(\tilde{x}) - \bar{c}(\tilde{w}))[I - M_\theta] \left( c - \frac{c_1}{\xi_1} \xi \right),$$

with

$$\nabla \hat{J}^*(w) = \frac{c_1}{\xi_1} \frac{w}{\mu_1 - \lambda_1} = \frac{c_1}{\xi_1} \frac{\kappa}{\mu_1} w, \quad w \geq 0.$$

A representation for the drift easily follows.

LEMMA 3.5. *For any  $v \in \mathbf{V}$ ,  $x \in \mathbb{R}_+^\ell$ , we have*

$$(103) \quad \begin{aligned} \langle \nabla h(x), v \rangle &= \nabla \hat{J}^*(\tilde{w}) \langle \xi, v \rangle - \nabla \hat{J}^*(\tilde{w}) \sum_{i=1}^{\ell} e^{-x_i/\theta} \xi_i v_i \\ &\quad + [c(\tilde{x}) - \bar{c}(\tilde{w})] \sum_{i=2}^{\ell} (1 - e^{-x_i/\theta}) b_i v_i \end{aligned}$$

with  $b_i = b(c_i - (c_1/\xi_1)\xi_i)$ ,  $i \geq 2$ .

Fix constants  $\beta_+ > \beta_- > 0$ , and define for  $r \geq 0$ ,

$$(104) \quad s_-(r) = \beta_- \log(1 + r/\beta_-), \quad s_+(r) = \beta_+ \log(1 + r/\beta_+).$$

It is assumed throughout that  $\beta_- \geq 3\theta^{-1}$ .

Similar to the proof of Lemma 2.10, to bound (103) we impose the following constraints on the velocity vector  $v$ :

$$(105) \quad v_i \geq 0 \quad \text{if } x_i \leq s_-(\xi^T x) - \beta_- \log(|\xi|) \text{ and } \xi_i > 0,$$

where  $|\xi| := \sum \xi_i$ . The minimum of  $\langle \nabla h(x), v \rangle$  over  $v$  satisfying (105) provides a bound under the  $h$ -MaxWeight policy. The following two results imply that (105) is feasible for a policy that is nonidling at Station 1.

LEMMA 3.6. *For each  $x \in \mathbb{R}_+^\ell$  we have  $\max\{x_i : \xi_i > 0\} \geq s_-(\xi^T x) - \beta_- \log(|\xi|)$ .*

*Proof.* Letting  $x^\infty = \max\{x_i : \xi_i > 0\}$ , we have  $\xi^T x \leq x^\infty |\xi|$ , and hence by concavity of the logarithm, with  $w = \xi^T x$ ,

$$\begin{aligned} s_-(w) &= \beta_- \log(1 + w/\beta_-) \leq \beta_- \log(1 + x^\infty |\xi|/\beta_-) \\ &\leq \beta_- \left( \log(|\xi|) + (1 + x^\infty |\xi|/\beta_- - |\xi|)/|\xi| \right). \end{aligned}$$

The right-hand side is bounded above by  $\beta_- \log(|\xi|) + x^\infty$  since  $|\xi| \geq 1$ . This gives the desired bound.  $\square$

We now establish a set of feasible values for  $v$ .

LEMMA 3.7. *There exists  $\kappa_v \geq 1$  and  $\varepsilon_v > 0$  such that for each  $\kappa \geq \kappa_v$  we have*

$$\{v : \|v\| \leq \varepsilon_v \quad \text{and} \quad \langle \xi, v \rangle \geq -(\mu_1 - \lambda_1)\} \subset \mathbf{V}.$$

*Proof.* The velocity space  $\mathbf{V}^\kappa$  is a polyhedron for each  $\kappa$ , and as  $\kappa \rightarrow \infty$  these sets converge to a polyhedron whose interior is nonempty, with a single face meeting the origin given by  $\{\xi^T v = 0\}$ .  $\square$

*Proof of Proposition 3.2.* To avoid trivialities we assume that  $\ell \geq 3$  (no less than three buffers), and that  $\xi_i > 0$  for at least three values of  $i$ .

The drift obtained with any  $v \in \mathbf{V}$  provides an upper bound on the drift obtained using the  $h$ -MaxWeight policy. We take  $v \in \mathbf{V}$  of the specific form  $v = v^* + v^0$  with  $v^* = -(\mu_1 - \lambda_1)\xi_1^{-1}\mathbf{1}^1$ , and  $v^0$  orthogonal to  $\xi$  so that  $\xi^T v = \xi^T v^* = -(\mu_1 - \lambda_1)$ . As in the proof of Lemma 2.10, we apply Proposition 2.8 to relax lattice constraints and boundary constraints in our construction of  $v$ .

Lemma 3.7 implies that we can find  $\varepsilon_{3.7} > 0$  such that the following inclusion holds for each  $\kappa \geq \kappa_v$ :

$$\{v = v^* + v^0 \in \mathbb{R}^\ell : \xi^T v^0 = 0 \text{ and } |v_i^0| \leq \varepsilon_{3.7} \text{ for each } i \geq 1\} \subset \mathbf{V}^\kappa.$$

With this value of  $\varepsilon_{3.7}$  fixed, we set  $v_i^0 = -\varepsilon_{3.7}$  for each  $i$  satisfying  $\xi_i = 0$ , and  $v_i^0 = 0$  for all but (at most) three indices satisfying  $\xi_i > 0$ . In Cases (ii) and (iii) below there are just two nonnull indices, denoted  $i_\ominus$  and  $i_\oplus$ , with  $v_{i_\ominus}^0 < 0$  and  $v_{i_\oplus}^0 > 0$ .

To complete the specification of  $v^0$  we introduce the index sets

$$\begin{aligned} I_\oplus &= \{i \geq 1 : \xi_i > 0, x_i \leq s_-(\xi^T x) - \beta_- \log(|\xi|)\}, \\ I_\ominus &= \{i \geq 2 : \xi_i > 0, x_i > s_+(\xi^T x)\}. \end{aligned}$$

The choice of  $v^0$  depends upon these sets as follows:

- (i) If  $I_\oplus = \emptyset$  and  $I_\ominus = \emptyset$ , then  $v_i^0 = 0$  for each  $i$  satisfying  $\xi_i > 0$ .
- (ii) If  $I_\oplus = \emptyset$  and  $I_\ominus \neq \emptyset$ , then we take  $i_\ominus \in I_\ominus$  arbitrary with  $v_{i_\ominus}^0 \xi_{i_\ominus} = -\varepsilon_{3.7}$ , and set  $v_1^0 \xi_1 = \varepsilon_{3.7}$ .
- (iii) If  $I_\oplus \neq \emptyset$  and  $I_\ominus = \emptyset$ , then we take

$$i_\oplus \in \arg \min\{x_i : i \geq 2, \xi_i > 0\}, \quad i_\ominus = 1,$$

and define  $v_{i_\oplus}^0 \xi_{i_\oplus} = \varepsilon_{3.7}^2$ ,  $v_{i_\ominus}^0 \xi_{i_\ominus} = -\varepsilon_{3.7}^2$ , and  $v_i^0 = 0$  for all other  $i$  satisfying  $\xi_i > 0$ .

- (iv) If  $I_\oplus \neq \emptyset$  and  $I_\ominus \neq \emptyset$ , then  $i_\ominus \in \arg \max\{x_i : i \geq 2, \xi_i > 0\}$ . To determine  $i_\oplus$  there are two subcases to consider.

- (a) If  $I_\oplus = \{1\}$ , then  $v_{i_\ominus}^0 \xi_{i_\ominus} = -\varepsilon_{3.7}$ , and  $i_\oplus = 1$  with  $v_1^0 \xi_1 = \varepsilon_{3.7}$ .
- (b) Otherwise,  $i_\oplus \in \arg \min\{x_i : i \geq 2, \xi_i > 0\}$ , and we take

$$v_{i_\oplus}^0 \xi_{i_\oplus} = \varepsilon_{3.7}^2, \quad v_{i_\ominus}^0 \xi_{i_\ominus} = -\varepsilon_{3.7}, \quad \text{and} \quad v_1^0 \xi_1 = \varepsilon_{3.7} - \varepsilon_{3.7}^2,$$

where again  $v_i = 0$  for all other  $i$  satisfying  $\xi_i > 0$ .

The added complexity in cases (iii) and (iv) is due to the positive drift induced by  $v_{i_\oplus}$ . By imposing the constraint that this is of order  $\varepsilon_{3.7}^2$  rather than  $\varepsilon_{3.7}$ , we can maintain a negative overall drift.

This choice of  $v$  satisfies (105). Moreover, under the assumption that  $\beta_- \geq 3\theta^{-1}$  and  $\beta_+ > \beta_-$ , we have for some constants  $k_{106}$ ,  $\varepsilon_{106}$ , and all  $x$ ,

$$\begin{aligned} |\nabla \hat{J}^*(\tilde{w}) e^{-x_i/\theta} \xi_i v_i| &\leq k_{106} \kappa (1 + w^2)^{-1} \leq \mathcal{E}(x), & i \notin I_\oplus, \\ (106) \quad -\nabla \hat{J}^*(\tilde{w}) e^{-x_{i_\oplus}/\theta} \xi_{i_\oplus} v_{i_\oplus} &\leq -\varepsilon_{106} \|M_\theta(x) \xi\| \kappa w & \text{if } I_\oplus \neq \emptyset, \\ \|M_\theta(x) \xi\| \kappa w &\leq \mathcal{E}(x) & \text{if } I_\oplus = \emptyset. \end{aligned}$$

Combining the bounds in (106) with Lemma 3.5 we obtain

$$\begin{aligned} \langle \nabla h(x), v \rangle &\leq -(\mu_1 - \lambda_1) \nabla \hat{J}^*(\tilde{w}) - \varepsilon_{106} \|M_\theta(x) \xi\| \kappa w + \mathcal{E}(x) \\ &\quad + [c(\tilde{x}) - \bar{c}(\tilde{w})] \sum_{i=2}^{\ell} (1 - e^{-x_i/\theta}) b_i v_i. \end{aligned}$$

To complete the proof we argue that the following bound holds: For some  $\varepsilon_{107} > 0$ ,

$$(107) \quad [c(\tilde{x}) - \bar{c}(\tilde{w})] \sum_{i=2}^{\ell} (1 - e^{-x_i/\theta}) b_i v_i \leq -\varepsilon_{107} b [c(\tilde{x}) - \bar{c}(\tilde{w})] + \mathcal{E}(x).$$

It is here that we require the fact that at most one value of  $v_i$  is positive for  $i \geq 2$ , and that for all such  $i$  we have the bound  $v_i \xi_i \leq \varepsilon_{3.7}^2$ .

If  $x_i > s_+(\xi^\top x)$  for some  $i \geq 2$  (not necessarily satisfying  $\xi_i > 0$ ), then  $v_i = -\varepsilon_{3.7}$  for some  $i \geq 2$ . In fact, with  $i_- \in \arg \max\{x_i : i \geq 2\}$  we have  $v_{i_-} = -\varepsilon_{3.7}$ , and from (101) we obtain  $c(x) - \bar{c}(w) \leq |c|x_{i_-}$ . Consequently,

$$\sum_{i=2}^{\ell} (1 - e^{-x_i/\theta}) b_i v_i \leq -(b_{i_-} \varepsilon_{3.7} - b_{i_{\oplus}} \varepsilon_{3.7}^2) + \varepsilon_{3.7} b_{i_-} e^{-[c(x) - \bar{c}(w)]/(|c|\theta)}.$$

The bound (107) follows for  $\varepsilon_{3.7} > 0$  sufficiently small: Fix  $\varepsilon_{3.7} < \min_{i,j \geq 2} b_i/b_j$  and set  $\varepsilon_{107} = \min_{i,j \geq 2} (b_i \varepsilon_{3.7} - b_j \varepsilon_{3.7}^2)/b$ . Note that the positive term  $b_{i_{\oplus}} \varepsilon_{3.7}^2$  is absent in cases (i) and (ii), so that we are considering the worst case in which  $I_{\oplus} \neq \emptyset$ .

If  $x_i \leq s_+(\xi^\top x)$  for each  $i \geq 2$ , then it may be impossible to guarantee the negative drift  $v_i = -\varepsilon_{3.7}$  for any  $i \geq 2$ . But this is irrelevant since in this case,

$$[c(\tilde{x}) - \bar{c}(\tilde{w})] \leq \mathcal{E}(x),$$

so that (107) follows trivially.  $\square$

*Proof of Proposition 3.3.* We begin with a representation of the form (38) based on a second-order mean value theorem of the form (61). We write  $h_0(x) = \frac{1}{2} x^\top H_0 x$ , with

$$H_0 = \frac{\kappa}{\mu_1} \frac{c_1}{\xi_1} \xi \xi^\top + b \left( c - \left( \frac{c_1}{\xi_1} \right) \xi \right) \left( c - \left( \frac{c_1}{\xi_1} \right) \xi \right)^\top.$$

Based on this expression combined with the mean value theorem, we obtain

$$(108) \quad \begin{aligned} \mathcal{D}h(x) &= \langle \nabla h(x), v \rangle + \frac{1}{2} \mathbb{E}[\Delta(t+1)^\top H_0 \Delta(t+1) \mid Q(t) = x] \\ &\quad + \frac{1}{2} \mathbb{E}[\Delta(t+1)^\top (\nabla^2 h(\bar{Q}) - H_0) \Delta(t+1) \mid Q(t) = x]. \end{aligned}$$

We also have, by definition of  $\hat{\eta}^*(x)$  in (100),

$$(109) \quad \begin{aligned} &\mathbb{E}[\Delta(t+1)^\top H_0 \Delta(t+1) \mid Q(t) = x] \\ &= \hat{\eta}^*(x) + b \mathbb{E}[(c - (c_1/\xi_1)\xi)^\top \Delta(t+1)]^2 \mid Q(t) = x. \end{aligned}$$

We apply Proposition 2.7 to bound the final term in (108):

$$(110) \quad \nabla^2 h(x) - H_0 = -[M_\theta H_0 + M_\theta H_0] + M_\theta H_0 M_\theta + \theta^{-1} \text{diag}(M_\theta \nabla h_0(\tilde{x})).$$

We have for any  $\Delta \in \mathbb{R}^\ell$ ,

$$\begin{aligned} &\Delta^\top [-[M_\theta H_0 + M_\theta H_0] + M_\theta H_0 M_\theta] \Delta \\ &= \kappa c_1 / (\mu_1 \xi_1) [-2(\Delta^\top \xi)(\Delta^\top M_\theta \xi) + (\Delta^\top M_\theta \xi)^2] + O(1) \\ &= \kappa c_1 / (\mu_1 \xi_1) [-(\Delta^\top M_\theta \xi) + 2(\Delta^\top M_\theta \xi)((\Delta^\top M_\theta \xi) - (\Delta^\top \xi))] + O(1) \\ &\leq \kappa c_1 / (\mu_1 \xi_1) [-(\Delta^\top M_\theta \xi) + 2\|\Delta\|^2 \|M_\theta \xi\| \|(I - M_\theta)\xi\|] + O(1), \end{aligned}$$

where terms that are independent of  $\kappa$  are suppressed using the “big O” notation. Applying the mean value theorem as in the proof of Lemma 2.11, we obtain the crude bound,  $\|(I - M_\theta)\xi\| \leq \theta^{-1}w$ , and hence for some  $k_0 < \infty$ ,

$$-[M_\theta H_0 + M_\theta H_0] + M_\theta H_0 M_\theta \leq k_0(\theta^{-1}\|M_\theta \xi\|\kappa w + 1)I.$$

Also, for a possibly larger constant  $k_0$ ,

$$\begin{aligned} \|M_\theta \nabla h_0(\tilde{x})\| &= \|M_\theta(\kappa\mu_1^{-1}\bar{c}(\tilde{w})\xi + b(c(\tilde{x}) - \bar{c}(\tilde{w}))(c - (c_1/\xi_1)\xi))\| \\ &\leq k_0(\|M_\theta \xi\|\kappa w + b|c(\tilde{x}) - \bar{c}(\tilde{w})|). \end{aligned}$$

Consequently, for some  $k_0 < \infty$ ,

$$\nabla^2 h(x) - H_0 \leq k_0\theta^{-1}(\kappa\|M_\theta \xi\|w + b|c(\tilde{x}) - \bar{c}(\tilde{w})|)I + k_0I.$$

This combined with (108) and (109) completes the proof.  $\square$

*Proof of Theorem 3.1.* Following Propositions 3.3 and 3.4, the proof of the theorem amounts to establishing the drift (96) for a function  $V$  derived from  $h$ . We define

$$V(x) = h(x) + \frac{1}{2} \frac{c_1}{\xi_1} \mu^{-1} \left( \frac{m^2 - m_L^2}{\mu - \alpha} \right) \xi^\top x, \quad x \in \mathbb{Z}_+^\ell,$$

where  $m^2 := \mathbb{E}[(S_1(1) - L_1(1))^2]$  and  $m_L^2 \mathbb{E}[(L_1(1))^2]$ . That is, we are re-introducing the linear term appearing in the solution to Poisson’s equation for the relaxation. Based on the definitions of  $\hat{\eta}^*$  and  $\hat{\eta}^*(x)$  in (88) and (100), we obtain the following identity for any policy:

$$\mathbb{E} \left[ \frac{1}{2} \frac{c_1}{\xi_1} \mu^{-1} \left( \frac{m^2 - m_L^2}{\mu - \alpha} \right) (W(t+1) - W(t)) \mid Q(t) = x \right] = \hat{\eta}^* - \hat{\eta}^*(x).$$

Hence the function  $V$  does satisfy (96).

This bound implies that (V3) holds, so that  $\pi(c)$  is finite for any finite  $\kappa$ . An application of the comparison theorem gives

$$\pi(c) \leq \hat{\eta}^* + \pi(\mathcal{E}).$$

From the form of  $\mathcal{E}$  given in (97), it follows that  $\pi(c)$  is bounded by a constant times  $\kappa$ . In fact, by the bound above, (97), and Jensen’s inequality, we obtain

$$\pi(c) \leq \hat{\eta}^* + k_{97} \log(\kappa + \pi(c)) + k_{97} \kappa \mathbb{E}_\pi[(1 + (\xi^\top Q(t))^2)^{-1}]$$

so that for a possibly larger constant,

$$\pi(c) \leq \hat{\eta}^* + k_{97} \log(\kappa) + k_{97} \kappa \mathbb{E}_\pi[(1 + (\xi^\top Q(t))^2)^{-1}].$$

Moreover, applying (35) we obtain,

$$\pi(c) \leq \hat{\eta}^* + k_{97} \log(\kappa) + k_{97} \kappa \mathbb{E}[(1 + (\widehat{W}^*(t))^2)^{-1}],$$

where the expectation is taken for the steady-state relaxation. Lemma A.2 of [37] implies that  $\kappa \mathbb{E}[(1 + (\widehat{W}^*(t))^2)^{-1}]$  is uniformly bounded in  $\kappa$ , so this final bound completes the proof.  $\square$

**4. Extensions, conjectures, and conclusions.** The generalized MaxWeight policies proposed in this paper can be designed to capture all of the desirable features observed in Tassiulus' original policy. Depending upon the structure of  $h$ , the policy can be designed to depend only on local information, as in the standard algorithm, or it can utilize more information, if available.

There remain many questions.

*Statistics.* Generalizations of Theorems 1.1 and 2.14 to network models with renewal arrivals and service are straightforward by applying fluid limit techniques [12, 15, 13]. It would be of interest to develop alternative methods to cope with these more complex models to obtain sharp performance estimates in heavy traffic. In particular, to date logarithmic regret has been established only for homogeneous Markovian models. To relax the homogeneity (Kelly-type) assumption, the workload process should be constructed in units of time rather than inventory as done here. It would be much more interesting to find a counterexample within the class of renewal models with finite second moment, though this is not likely to exist.

Of greater practical importance is the issue of memory: Long-range dependent models remain a frontier. It may be possible to extend fluid limit techniques to establish stability. New techniques are required to obtain performance bounds.

*Information.* Design of the function  $h_0$  requires stability considerations; e.g., monotonicity has been imposed as a blanket assumption in each of the main results. A second consideration is the amount of information required for implementation.

Consider the special case of the quadratic function (1). Monotonicity holds provided  $D_{ij} \geq 0$  for each  $i, j$ , and the dynamic programming inequality holds if in addition  $D_{ii} > 0$  for each  $i$ . In typical scheduling and routing models the MaxWeight policy in which  $D$  is diagonal requires only local information [53]. The main result of this paper shows that this can be relaxed through the introduction of a state transformation. If the matrix  $D$  that defines  $h_0$  is sparse, then the amount of information required for implementation of the  $h$ -MaxWeight policy will be limited.

Suppose that  $D$  is a band matrix with width  $n_0 \geq 1$  ( $D_{ij} = 0$  for each  $i, j$  satisfying  $|i - j| > n_0$ ). For the same scheduling and routing models considered in earlier work, the resulting  $h$ -MaxWeight policy will require “ $(n_0 + 1)$ -hop” local information. For example, for a network consisting of a sequence of  $N$  queues in tandem, for each integer  $i \in [n_0, N - n_0 - 1]$  the policy at Station  $i$  will require queue length information at buffers  $\{i \pm k, |k| \leq n_0\}$  and buffer  $i + n_0 + 1$ .

Another important class of quadratics is those with low rank. Suppose that for linearly independent vectors  $\{d^i : i = 1, \dots, n_1\} \subset \mathbb{R}_+^\ell$  we have

$$h_0(x) = \frac{1}{2} \sum_{i=1}^{n_1} d^i d^{i^\top}.$$

For example, the function (91) is of this form with  $n_1 = 2$ . This matrix is not banded in general, but the  $h$ -MaxWeight policy will again require limited information. For scheduling and routing models the required information is the same “one-hop” data required in the usual MaxWeight algorithm, along with values of the inner products  $\{d^{i^\top} Q(t) : i = 1, \dots, n_1\}$ . For small values of  $n_1$  this information might be distributed through message passing.

*Performance bounds for universally stabilizing policies.* The  $h$ -MaxWeight policy considered in Theorem 2.14 using the change of variables (73) is universally stabilizing, but its performance is not understood. It will be interesting to evaluate its

performance in the setting of section 3. It will not yield logarithmic regret in general under the assumptions of Theorem 3.1.

CONJECTURE 1. *Under the assumptions of Theorem 3.1, with  $\tilde{x}$  redefined via (73) and the  $L_2$  bound (77) satisfied for  $A^\infty(t)$ , the  $h$ -MaxWeight policy will result in asymptotic minimality of workload, with logarithmic regret, in the sense that*

$$\mathbb{E}[W(t; x)] \leq \mathbb{E}[\widehat{W}^*(t; x)] + k_1 \log(\kappa)$$

for some  $k_1$ , each  $\kappa$  sufficiently large, and each  $t$  an initial condition.

It is likely that minimality is not difficult to establish. It is also likely that state space collapse (25) will hold in some form, perhaps with  $\mathcal{X}^*$  defined with respect to the perturbed cost function,

$$\mathcal{X}^*(w) = \arg \min_{x \in \mathbb{R}_+^\ell} (c(\tilde{x}) : \xi^\top x = w).$$

*Linearity and logarithmic regret.* If  $c$  is an arbitrary norm on  $\mathbb{R}^\ell$ , then the average cost for the relaxation becomes

$$\widehat{\eta}^* = \frac{1}{2} \overline{c}(1) \frac{\sigma_\kappa^2}{\mu_1} \kappa,$$

where the definition of the effective cost is given in (24). This generalizes (93) since  $\overline{c}(1) = c_1/\xi_1$  under the assumptions of Theorem 3.1. A stability proof is simplified if the function  $h_0$  in (91) used to define the  $h$ -MaxWeight policy is redefined via

$$h_0(x) = \widehat{J}^*(\xi^\top x) + \frac{1}{2} b \|x - x^*\|^2.$$

CONJECTURE 2. *The conclusions of Theorem 3.1 remain valid when  $c$  is a general norm,  $h_0$  redefined as above, and with the remaining assumptions of the theorem maintained.*

*Multiple bottlenecks.* The generalization of Theorem 3.1 to multiple bottlenecks is a significant open problem. This is difficult because we do not have an explicit representation for the relative value function for the relaxation, and we do not know the optimal policy when the effective cost is not monotone.

To formulate a conjecture we again restrict to the homogeneous scheduling model with linear cost. If there are  $n$  bottlenecks in heavy traffic, we consider an  $n$ -dimensional relaxation, as formulated in [9]. Analysis of the usual MaxWeight policy based on a workload relaxation of dimension  $n \geq 2$  is contained in [39, Chapter 6].

The workload relaxation evolves in a positive cone  $W \subset \mathbb{R}^n$  according to a multi-dimensional recursion of the form (83). Let  $\widehat{\eta}^*$  denote the optimal average cost, and suppose that  $\hat{h}_* : W \rightarrow \mathbb{R}$  solves the average-cost optimality equations for the relaxation, perhaps approximately: For some finite constant  $k_3$  and all  $\kappa$ ,

$$\min \mathbb{E}[\hat{h}_*(\widehat{W}(t+1)) \mid \widehat{W}(t) = w, \mathcal{L}(t) = \iota] \leq \hat{h}_*(w) - \overline{c}(w) + \widehat{\eta}^* + k_3 \log(\kappa),$$

where  $\mathcal{L}(t)$  denotes the idleness at time  $t$ , and the minimum is over all  $\iota \geq 0$  such that  $\widehat{W}(t+1) \in W$  with probability one. The integer constraints are relaxed so that  $\hat{h}_*$  is defined on all of  $W$ . Based on the relative value function for the relaxation, the function  $h_0$  in (91) is redefined via

$$(111) \quad h_0(x) := \hat{h}_*(\xi^\top x) + \frac{1}{2} b [c(x) - \overline{c}(\xi^\top x)]^2, \quad x \in \mathbb{R}_+^\ell.$$

The function  $\hat{h}_*$  will depend upon  $\kappa$  in a parameterized model, but the constant  $b$  will be fixed as in Theorem 3.1.

CONJECTURE 3. *Theorem 3.1 can be extended to the case where there are precisely  $n$  bottlenecks as  $\kappa \rightarrow \infty$ , based on the  $h$ -MaxWeight policy with  $h_0$  as given in (111).*

If true, this provides a valuable tool for constructing an effective policy in a complex network setting.

A final topic of current research is to create a bridge between the concepts developed in this paper, and recent methodology that has emerged in the machine learning literature. Given a parameterized family of functions  $\{h_\alpha : \alpha \in \mathbb{R}^d\}$ , we seek the value of  $\alpha$  such that the  $h_\alpha$ -MaxWeight policy has the best performance in this class. There are a variety of methods for finding an optimizer based on simulation [3, 51, 54, 16, 46]. It is hoped that specialized algorithms can be constructed for networks based on the techniques introduced here.

## REFERENCES

- [1] B. ATA AND S. KUMAR, *Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies*, Ann. Appl. Probab., 15 (2005), pp. 331–391.
- [2] S. L. BELL AND R. J. WILLIAMS, *Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: Asymptotic optimality of a continuous review threshold policy*, Ann. Appl. Probab., 11 (2001), pp. 608–649.
- [3] D. BERTSEKAS AND J. N. TSITSIKLIS, *Neuro-Dynamic Programming*, Atena Scientific, Cambridge, MA, 1996.
- [4] D. BERTSIMAS, D. GAMARNIK, AND J. N. TSITSIKLIS, *Stability conditions for multiclass fluid queueing networks*, IEEE Trans. Automat. Control, 41 (1996), pp. 1618–1631.
- [5] V. S. BORKAR, *Convex analytic methods in Markov decision processes*, in Handbook of Markov Decision Processes, Internat. Ser. Oper. Res. Management Sci. 40, Kluwer, Boston, MA, 2002, pp. 347–375.
- [6] M. BRAMSON AND R. J. WILLIAMS, *Two workload properties for Brownian networks*, Queueing Systems Theory Appl., 45 (2003), pp. 191–221.
- [7] H. CHEN AND D. D. YAO, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, Stochastic Modelling and Applied Probability, Springer-Verlag, New York, 2001.
- [8] M. CHEN, R. DUBRAWski, AND S. P. MEYN, *Management of demand-driven production systems*, IEEE Trans. Automat. Control, 49 (2004), pp. 686–698.
- [9] M. CHEN, C. PANDIT, AND S. P. MEYN, *In search of sensitivity in network optimization*, Queueing Systems Theory Appl., 44 (2003), pp. 313–363.
- [10] D. P. CONNORS, G. FEIGIN, AND D. YAO, *Scheduling semiconductor lines using a fluid network model*, IEEE Trans. on Robotics and Automation, 10 (1994), pp. 88–98.
- [11] D. BERTSIMAS, I. PASCHALIDIS, AND J. N. TSITSIKLIS, *Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance*, Ann. Appl. Probab., 4 (1994), pp. 43–75.
- [12] J. G. DAI, *On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models*, Ann. Appl. Probab., 5 (1995), pp. 49–77.
- [13] J. G. DAI AND W. LIN, *Maximum pressure policies in stochastic processing networks*, Oper. Res., 53 (2005), pp. 197–218.
- [14] J. G. DAI AND W. LIN, *Asymptotic optimality of maximum pressure policies in stochastic processing networks*, Ann. Appl. Probab., to appear.
- [15] J. G. DAI AND S. P. MEYN, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Trans. Automat. Control, 40 (1995), pp. 1889–1904.
- [16] D. P. DE FARIAS AND B. V. ROY, *The linear programming approach to approximate dynamic programming*, Oper. Res., 51 (2003), pp. 850–865.
- [17] A. ERYILMAZ, R. SRIKANT, AND J. R. PERKINS, *Stable scheduling policies for fading wireless channels*, IEEE/ACM Trans. Networks, 13 (2005), pp. 411–424.



- [18] G. FAYOLLE, V. A. MALYSHEV, AND M. V. MEN'SHIKOV, *Topics in the constructive theory of countable Markov chains*, Cambridge University Press, Cambridge, UK, 1995.
- [19] D. GAMARNIK AND A. ZEEVI, *Validity of heavy traffic steady-state approximations in generalized Jackson networks*, *Adv. Appl. Probab.*, 16 (2006), pp. 56–90.
- [20] L. GEORGIADIS, W. SZPANKOWSKI, AND L. TASSIULAS, *A scheduling policy with maximal stability region for ring networks with spatial reuse*, *Queueing Systems Theory Appl.*, 19 (1995), pp. 131–148.
- [21] J. M. HARRISON, *Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete-review policies*, *Ann. Appl. Probab.*, 8 (1998), pp. 822–848.
- [22] J. M. HARRISON AND M. J. LÓPEZ, *Heavy traffic resource pooling in parallel-server systems*, *Queueing Systems Theory Appl.*, 33 (1999), pp. 339–368.
- [23] S. G. HENDERSON, S. P. MEYN, AND V. B. TADIĆ, *Performance evaluation and policy selection in multiclass networks*, *Discrete Event Dyn. Syst.*, 13 (2003), pp. 149–189. Special issue on learning, optimization and decision making.
- [24] J. F. C. KINGMAN, *The single server queue in heavy traffic*, *Proc. Cambridge Philos. Soc.*, 57 (1961), pp. 902–904.
- [25] J. F. C. KINGMAN, *On queues in heavy traffic*, *J. Roy. Statist. Soc. Ser. B*, 24 (1962), pp. 383–392.
- [26] P. R. KUMAR AND S. P. MEYN, *Duality and linear programs for stability and performance analysis queueing networks and scheduling policies*, *IEEE Trans. Automat. Control*, 41 (1996), pp. 4–17.
- [27] P. R. KUMAR AND T. I. SEIDMAN, *Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems*, *IEEE Trans. Automat. Control*, AC-35 (1990), pp. 289–298.
- [28] S. KUMAR AND P. R. KUMAR, *Performance bounds for queueing networks and scheduling policies*, *IEEE Trans. Automat. Control*, AC-39 (1994), pp. 1600–1611.
- [29] C. N. LAWS AND G. M. LOUTH, *Dynamic scheduling of a four-station queueing network*, *Probab. Engng. Inform. Sci.*, 4 (1990), pp. 131–156.
- [30] N. LAWS, *Dynamic Routing in Queueing Networks*, Ph.D. thesis, Cambridge University, Cambridge, UK, 1990.
- [31] S. LIPPMAN, *Applying a new device in the optimization of exponential queueing systems*, *Oper. Res.*, 23 (1975), pp. 687–710.
- [32] C. MAGLARAS, *Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality*, *Ann. Appl. Probab.*, 10 (2000), pp. 897–929.
- [33] A. MANDELBAUM AND A. L. STOLYAR, *Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized  $c\mu$ -rule*, *Oper. Res.*, 52 (2004), pp. 836–855.
- [34] S. P. MEYN, *The policy iteration algorithm for average reward Markov decision processes with general state space*, *IEEE Trans. Automat. Control*, 42 (1997), pp. 1663–1680.
- [35] S. P. MEYN, *Sequencing and routing in multiclass queueing networks part I: Feedback regulation*, *SIAM J. Control Optim.*, 40 (2001), pp. 741–776.
- [36] S. P. MEYN, *Sequencing and routing in multiclass queueing networks part II: Workload relaxations*, *SIAM J. Control Optim.*, 42 (2003), pp. 178–217.
- [37] S. P. MEYN, *Dynamic safety-stocks for asymptotic optimality in stochastic networks*, *Queueing Systems Theory Appl.*, 50 (2005), pp. 255–297.
- [38] S. P. MEYN, *Workload models for stochastic networks: Value functions and performance evaluation*, *IEEE Trans. Automat. Control*, 50 (2005), pp. 1106–1122.
- [39] S. P. MEYN, *Control Techniques for Complex Networks*, Cambridge University Press, Cambridge, UK, 2007.
- [40] S. P. MEYN AND D. G. DOWN, *Stability of generalized Jackson networks*, *Ann. Appl. Probab.*, 4 (1994), pp. 124–148.
- [41] S. P. MEYN AND R. L. TWEEDIE, *Markov Chains and Stochastic Stability*, 2nd ed., Springer-Verlag, London, 1993. Available online at <http://black.csl.uiuc.edu/~meyn/pages/book.html>.
- [42] S. P. MEYN AND R. L. TWEEDIE, *Markov Chains and Stochastic Stability*, 3rd ed., Cambridge University Press, Cambridge, UK, to appear.
- [43] S. P. MEYN AND R. L. TWEEDIE, *Computable bounds for convergence rates of Markov chains*, *Ann. Appl. Probab.*, 4 (1994), pp. 981–1011.
- [44] J. R. MORRISON AND P. R. KUMAR, *New linear program performance bounds for queueing networks*, *J. Optim. Theory Appl.*, 100 (1999), pp. 575–597.
- [45] J. OU AND L. M. WEIN, *Performance bounds for scheduling queueing networks*, *Ann. Appl. Probab.*, 2 (1992), pp. 460–480.

- [46] B. V. ROY, *Neuro-dynamic programming: Overview and recent trends*, in Markov Decision Processes: Models, Methods, Directions, and Open Problems, E. Feinberg and A. Shwartz, eds., Kluwer, Dordrecht, 2001, pp. 43–82.
- [47] A. N. RYBKO AND A. L. STOLYAR, *On the ergodicity of random processes that describe the functioning of open queueing networks*, Problemy Peredachi Informatsii, 28 (1992), pp. 3–26.
- [48] S. SHAKKOTTAI, R. SRIKANT, AND A. L. STOLYAR, *Pathwise optimality of the exponential scheduling rule for wireless channels*, Adv. Appl. Probab., 36 (2004), pp. 1021–1045.
- [49] A. L. STOLYAR, *Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic*, Adv. Appl. Probab., 14 (2004), pp. 1–53.
- [50] V. SUBRAMANIAN AND D. LEITH, *Draining time based scheduling algorithm*, in Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, LA, 2007, pp. 1162–1167.
- [51] R. SUTTON AND A. BARTO, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998 Available online at <http://www.cs.ualberta.ca/Esutton/book/ebook/the-book.html>
- [52] L. TASSIULAS, *Adaptive back-pressure congestion control based on local information*, IEEE Trans. Automat. Control, 40 (1995), pp. 236–250.
- [53] L. TASSIULAS AND A. EPHREMIDES, *Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks*, IEEE Trans. Automat. Control, 37 (1992), pp. 1936–1948.
- [54] J. N. TSITSIKLIS AND B. V. ROY, *An analysis of temporal-difference learning with function approximation*, IEEE Trans. Automat. Control, 42 (1997), pp. 674–690.
- [55] J. A. VAN MIEGHEM, *Dynamic scheduling with convex delay costs: The generalized c- $\mu$  rule*, Ann. Appl. Probab., 5 (1995), pp. 809–833.